

UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE ESTUDIOS ESTADÍSTICOS



TESIS DOCTORAL

**Explainability and business sense in machine learning models
for credit risk assesment**

**Explicabilidad y sentido de negocio en modelos de aprendizaje
automático para evaluación de riesgo de crédito**

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Miller Janny Ariza Garzón

Directores

Javier Arroyo Gallardo
María Jesús Segovia Vargas
Juana María del Mar Camacho Miñano

Madrid

UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE ESTUDIOS ESTADÍSTICOS



TESIS DOCTORAL

**EXPLAINABILITY AND BUSINESS SENSE IN
MACHINE LEARNING MODELS FOR CREDIT RISK
ASSESSMENT**

**EXPLICABILIDAD Y SENTIDO DE NEGOCIO EN
MODELOS DE APRENDIZAJE AUTOMÁTICO PARA
LA EVALUACIÓN DE RIESGO DE CRÉDITO**

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Miller Janny Ariza Garzón

DIRECTORES

Javier Arroyo Gallardo

María Jesús Segovia Vargas

Juana María del Mar Camacho Miñano

UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE ESTUDIOS ESTADÍSTICOS



TESIS DOCTORAL

**EXPLAINABILITY AND BUSINESS SENSE IN
MACHINE LEARNING MODELS FOR CREDIT RISK
ASSESSMENT**

**EXPLICABILIDAD Y SENTIDO DE NEGOCIO EN
MODELOS DE APRENDIZAJE AUTOMÁTICO PARA
LA EVALUACIÓN DE RIESGO DE CRÉDITO**

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Miller Janny Ariza Garzón

PROGRAMA

DOCTORADO EN ANÁLISIS DE DATOS (DATA SCIENCE)

DIRECTORES

Javier Arroyo Gallardo

María Jesús Segovia Vargas

Juana María del Mar Camacho Miñano

Madrid, octubre de 2023

Agradecimientos

En primer lugar, deseo expresar mi sincero agradecimiento a Dios por brindarme la oportunidad de conocer a seres excepcionales, tanto en el ámbito académico como en el humano. Estos distinguidos individuos se destacan por su integridad y excelencia profesional, así como por poseer cualidades humanas extraordinarias. Entre ellos, quiero destacar al profesor Javier Arroyo Gallardo, quien ha fungido como mi guía y tutor, así como uno de mis directores. A él le debo un profundo agradecimiento por su comprensión, compromiso, disposición, esfuerzo, buen lenguaje y dedicación a lo largo de todo mi proceso de formación doctoral. Sin su valiosa orientación y apoyo, difícilmente habría podido culminar este proyecto. El profesor Javier me ha escuchado, acompañado y respaldado en los momentos de dificultad y tensión que he experimentado durante estos últimos años.

Asimismo, quiero reconocer a la profesora María Jesús Segovia Vargas, quien también ejerce como directora, por su constante respaldo, motivación, confianza, espiritualidad, así como por sus amables, precisos y detallados comentarios, los cuales han estado siempre orientados hacia el logro de cada una de las tareas y metas propuestas en este proceso doctoral. Además, agradezco su preocupación por generar oportunidades de crecimiento tanto académico como profesional en mi vida. De igual manera, deseo expresar mi gratitud hacia la profesora María del Mar Camacho Miñano, cuyos sabios consejos y palabras de aliento, así como sus minuciosos comentarios y correcciones, han sido de un valor incalculable. En las ocasiones en que se presentó la oportunidad, trabajamos de la mano para asegurar la consecución de mis objetivos académicos.

No puedo omitir mencionar a la profesora María Concepción Latorre, quien me abrió la posibilidad de acceder a esta formación y enriquecimiento personal.

Por último, pero no menos importante, debo agradecer a mi familia por su confianza, compañía, paciencia y apoyo incondicional a lo largo de este intenso periodo de mi vida. A mi esposa, Julieth, quiero expresarle un agradecimiento especial por decidir, con amor,

respaldar este sueño, renunciando a sus propias prioridades y ajustando el curso de sus planes y proyectos.

Resumen

El cambiante panorama de la modelización del riesgo de crédito, debido al avance tecnológico y la proliferante generación de plataformas con alternativas de financiación, alternas a las instituciones tradicionales, ha planteado cuestiones críticas en relación con la eficacia del aprendizaje automático para satisfacer las exigencias de los reguladores y de los usuarios, en particular en el mercado de préstamos P2P. En este contexto, se abordan algunos vacíos identificados en la literatura en relación con la modelización, su aplicabilidad orientada al negocio y la interpretabilidad de los resultados del aprendizaje automático. Proponemos estrategias alternativas de modelización, flujos de trabajo, incorporando predictibilidad sentido de negocio y explicabilidad. Se demuestra que este enfoque no sólo mejora la precisión de las predicciones frente a alternativas tradicionales, sino que permite identificar las variables determinantes, y a su vez, perfiles de riesgo o segmentos de mayor rentabilidad. No sólo se discierne la importancia de las variables, sino también las relaciones no lineales, los cambios de dispersión y las rupturas estructurales, características difíciles de captar con métodos convencionales. Además, identificamos las ventajas financieras de incorporar los beneficios en el proceso de modelización. El uso de herramientas de explicabilidad aclara los cambios en la importancia y la forma de dependencia de las variables en comparación con los métodos convencionales, que a menudo pasan por alto los costes y los beneficios de la correcta o incorrecta clasificación. Esto arroja luz sobre los factores que impulsan la concesión rentable de préstamos a pequeñas empresas en el contexto del mercado P2P, distintos de los que influyen en la clasificación tradicional del riesgo de crédito. A su vez, pudimos extraer algunas limitaciones, especialmente en este segmento, caracterizado por un mayor riesgo. De esta manera, aunque este trabajo se centra en la modelización del riesgo crediticio en el mercado P2P, empleando el aprendizaje automático con cualidades de explicabilidad y sentido de negocio para dar mayor transparencia e incrementar la probabilidad de uso de estas herramientas para al proceso de otorgamiento de créditos, nuestro enfoque es extrapolable a

otros ámbitos. Abordamos con rigor un problema de clasificación con clases desbalanceadas, se aprecia cómo introducirle costes y beneficios en el aprendizaje, se estudia el aporte de las variables en contextos complejos con herramientas de explicabilidad, lo que es extrapolable a cualquier problema de clasificación, pudiéndose obtener resultados reseñables no solamente en dominios financieros.

Palabras clave

Modelización de la gestión del riesgo, Modelización del riesgo crediticio, Modelización de beneficios e inversiones, Préstamos P2P, Aprendizaje automático, Explicabilidad, Interpretabilidad, Valores de Shapley, Valores SHAP, Modelos de refuerzo, Modelo de refuerzo de gradiente extremo (XGBoost), Regresión logística, Préstamos a pequeñas empresas, Modelos sensibles a los costes, Aprendizaje sensible a los beneficios, Inteligencia artificial explicable (XAI), Explicabilidad post-hoc, Análisis bibliométrico.

Abstract

The changing landscape of credit risk modeling, due to technological advancement and the proliferating generation of platforms with alternative financing alternatives to traditional institutions, has raised critical questions regarding the effectiveness of machine learning in meeting the demands of regulators and users, particularly in the P2P lending market. In this context, we address some gaps identified in the literature regarding modeling, its business-oriented applicability, and the interpretability of machine learning results. We propose alternative modeling strategies workflows, incorporating business sense predictability and explainability. We show that this approach not only improves the accuracy of predictions compared to traditional alternatives but also allows the identification of the critical variables and, in turn, risk profiles or segments of higher profitability. It not only discerns the importance of variables but also nonlinear relationships, dispersion changes and structural breaks, characteristics that are difficult to capture with conventional methods. In addition, we identify the financial advantages of incorporating profits into the modeling process. Explainability tools clarify changes in the importance and form of dependence of variables compared to conventional methods, which often overlook the costs and benefits of correct or incorrect classification. In doing so, it sheds light on the factors that drive profitable small business lending in the context of the P2P market, distinct from those influencing traditional credit risk classification. At the same time, we were able to extract some limitations, especially in this segment, characterized by higher risk. Thus, although this work focuses on modeling credit risk in the P2P market using machine learning with qualities of explainability and business sense to provide greater transparency and increase the likelihood of using these tools for the credit granting process, our approach is extrapolated to other areas. We rigorously address a classification problem with unbalanced classes and show how to introduce costs and benefits in learning; we study the contribution of variables in complex contexts with explainability tools, which can be transferable to any classification problem,

being able to obtain remarkable results not only in financial domains.

Keywords

Risk management modeling, Credit risk modeling, Profit and investment modeling, P2P lending, Machine learning, Explainability, Interpretability, Shapley values, SHAP values, Boosting models, Extreme gradient boosting (XGBoost), Logistic regression, Small business lending, Cost-sensitive models, Profit-sensitive learning, Explainable artificial intelligence (XAI), Post-hoc explainability, Bibliometric analysis.

Índice general

Resumen	I
Abstract	III
Lista de figuras	VIII
Lista de tablas	IX
1. Introducción	1
1.1. El aprendizaje automático y el modelado de riesgo en los préstamos peer-to-peer	1
1.2. Objetivos	7
1.3. Explicabilidad y sentido de negocio en el modelado del riesgo de crédito . . .	10
1.4. Estructura del documento	14
2. Artículo. Risk-return modelling in the P2P lending market: trends, gaps, recommendations and future directions	16
2.1. Introduction	18
2.2. Research Methodology	21
2.2.1. Research tools	21
2.2.2. Search and screening of the relevant papers	24
2.3. Results	26
2.3.1. Bibliometric analysis	26
2.3.2. Systematic analysis	40
2.4. Discussion	62

2.5. Research gaps, recommendations, and future directions	64
Annex. Table of variables and specific model of the selected papers	70
3. Artículo. Explainability of a machine learning granting scoring model in peer-to-peer lending	74
3.1. Introduction	76
3.2. Machine learning and credit risk modeling in P2P lending	79
3.3. Explainability of predictions in machine learning	82
3.3.1. SHAP values	84
3.3.2. Explainability elements in logistic regression	88
3.4. Experimental setting	90
3.4.1. Data preparation	90
3.4.2. Model adjustment	93
3.4.3. Model evaluation	96
3.4.4. Model explanation	96
3.5. Results	97
3.5.1. Exploratory statistical analysis	97
3.5.2. Model evaluation	100
3.5.3. Model explanation	101
3.6. Conclusions	111
4. Artículo. Profit-sensitive machine learning classification with explanations in credit risk: The case of small businesses in peer-to-peer lending	113
4.1. Introduction	115
4.2. Cost-sensitive methods in loan evaluation	119
4.2.1. Classification errors and cost-sensitive methods	119
4.2.2. Review of cost-sensitive and profit-sensitive methods for loan evaluation	122
4.2.3. Contextualizing our profit-sensitive proposal	127

4.3.	A proposal of profit-sensitive measures	128
4.3.1.	A profit-sensitive confusion matrix	128
4.3.2.	Profit-sensitive measures	131
4.4.	A profit-sensitive machine learning approach with explanations	132
4.4.1.	Direct profit-sensitive optimization of the XGBoost algorithm	133
4.4.2.	Indirect profit-sensitive approaches	134
4.4.3.	A proposal of profit-sensitive XGBoost in the estimation and decision process	137
4.4.4.	Model Explanation	137
4.5.	Experimental setting	139
4.5.1.	Dataset and model variables	139
4.5.2.	Estimation of the profit functions	142
4.5.3.	Models considered	144
4.5.4.	Model adjustment	146
4.5.5.	Performance metrics	146
4.6.	Results	148
4.6.1.	Preliminary descriptive and inferential analysis	148
4.6.2.	Performance evaluation	151
4.6.3.	Profit-sensitive explainability	155
4.7.	Conclusions	160
5.	Discusión y Conclusiones	165
5.1.	En relación con el diseño de los experimentos	166
5.2.	Principales Contribuciones	169
5.3.	Oportunidades y trabajo futuro	172
5.4.	Conclusiones finales	174
	Bibliografía	177

Lista de figuras

2.1. Distribution of primary papers selected from 2010-2019 and up until April 2020	27
2.2. Map of the final sample of paper participation by country, and contributions between countries according to coauthorship	30
2.3. Distribution of the number of lending platforms in the P2P market by country	31
2.4. Main authors' citation graph	32
2.5. Graph of the co-occurrence and clusters for the concepts present in titles and abstracts	33
2.6. Graph of the co-occurrence and time trend for the concepts present in titles and abstracts	34
2.7. Graph of co-occurrence and time trend for concepts present in keywords . . .	39
3.1. Graphical representation of SHAP values	86
3.2. Steps of the model development and assessment	91
3.3. Feature importance for each considered model	102
3.4. Aggregated relative importance by model	103
3.5. Dependence graphs of quantitative variables.	108
3.6. Dependence graphs of categorical variables.	110
4.1. Aggregated relative importance by XGBoost model	155
4.2. Dependence graphs of quantitative variables.	157
4.3. Dependence graphs of categorical variables.	159

Lista de tablas

2.1. Measures of contribution per author	28
2.2. Publication venues of the P2P lending market papers	29
2.3. Descriptive summary of clusters in the co-occurrence network based on abstracts and titles	35
2.4. Datasets	41
2.5. Software used	43
2.6. Business problem	44
2.7. Paradigm used (93 papers, the five papers that did not use datasets were excluded)	50
2.8. Methodological aspects (93 papers, the five papers that did not use datasets were excluded)	53
2.9. Other modeling aspects	60
3.1. Description of the explanatory variables according to the information in Kaggle	93
3.2. Exploratory statistical analysis of quantitative variables vs target variable . .	97
3.3. Relative frequency distribution and default rate by category of the categorical variables	98
3.4. Performance on training and test samples	100
3.5. Performance comparison for the test sample	100
3.6. Performance measures by class for the test sample	101
3.7. Explainability and inference logistic regression	105
4.1. Confusion matrix with miss-classification costs and profits	129

4.2. Description of the input variables	141
4.3. Description of variables	142
4.4. Modeling approaches used in the paper according to the method considered, the learning optimization and decision functions used and the rule used for decision making.	145
4.5. Descriptive Analysis. Quantitative variables	148
4.6. Descriptive Analysis. Categorical variables	149
4.7. Financial description of loans	150
4.8. Expected profit and score-based classification measures in validation	151
4.9. Profit and classification measures for the decisions in validation	152
4.10. Profit values of the misclassification matrix in validation	154

Capítulo 1

Introducción

1.1. El aprendizaje automático y el modelado de riesgo en los préstamos peer-to-peer

En un mundo en evolución e innovación acelerada, mucho más globalizado e interconectado, desarrollado tecnológicamente, con nuevas demandas y desafíos, surgen nuevas alternativas financieras, nichos más personalizados que buscan suplir las necesidades emergentes de los mercados. En este contexto, surgen las “fintech” (Financial innovations enabled by digital technology, “Innovaciones financieras facilitadas por la tecnología digital”, en español) como propuesta para ofrecer servicios y productos financieros mejorados, a la medida de los agentes, mediante el uso y aprovechamiento de los avances tecnológicos. Específicamente, han surgido desarrollos en servicios de crédito, depósitos y captación de capital; servicios de pagos, compensación y liquidación, incluidas las monedas digitales o criptomonedas; servicios de gestión de inversiones y su negociación; y seguros y reaseguros ([Navaretti et al., 2018](#)).

Como propósito, las fintech generan formas más económicas de contratación financiera, abrevian la intermediación y aminoran sus costos, amplían la cobertura, reducen los tiempos de respuesta, y mejoran el bienestar de los consumidores ([Chen et al., 2019b](#); [Fuster et al., 2019](#); [Navaretti et al., 2018](#); [Thakor, 2020](#)). Aparecen como actores del mercado para suplir problemas de escasa oferta que no se suplen con las fuentes tradicionales, evitando costos

por intermediación ([Feyen et al., 2021](#)). A su vez, reducen la exclusión financiera. [Demir et al. \(2022\)](#), por ejemplo, muestran evidencia de como las fintech son facilitadoras clave de la inclusión financiera y de la reducción de la desigualdad de los ingresos.

En este contexto, surge el mercado de préstamos entre pares, peer-to-peer (P2P), a través de fintechs, con una innovación que permite a las personas y las empresas acceder a financiamiento y realizar transacciones directamente entre sí, eliminando la necesidad de intermediarios financieros tradicionales. Concretamente, emergen fintech apoyadas en plataformas tecnológicas, cobran una comisión por el servicio de poner en contacto a prestamistas y prestatarios, y permiten a los prestatarios obtener créditos directamente de los prestamistas a tipos más bajos y con mayor accesibilidad que las alternativas crediticias convencionales. Los prestamistas, por su parte, pueden obtener una rentabilidad competente frente a otros productos financieros ([Emekter et al., 2015](#)).

El auge de este nuevo mercado, igual que sucede con las fintech en general, está asociado a factores como la reducción de los costes de intermediación, escasa regulación ([Basha et al., 2021](#)), cubrimiento de necesidades de financiación que no se alcanzan a cubrir con el sistema tradicional, y la evolución de la tecnología, lo que permite personalizar la experiencia y mejorar el servicio de sus actores. El mercado P2P emerge no solo como alternativa de financiación, sino como opción de inversión, incentivando una mayor inclusión financiera de personas y empresas excluidas de los sistemas habituales ([Claessens et al., 2018](#)).

Las plataformas de préstamos entre particulares (P2P) surgen también como vías innovadoras para que las pequeñas empresas accedan al crédito ([Kgoroadira et al., 2019](#); [Segal, 2015](#)). Vemos como, en economías locales, estas plataformas facilitan la concesión de préstamos a empresas que se ajustan a los Objetivos de Desarrollo Sostenible (ODS), contribuyendo a la sostenibilidad económica y al desarrollo social ([Hudaefi, 2020](#)). Por tal razón, el mercado P2P se considera una alternativa no competitiva y complementaria de la banca convencional [Milne & Parboteeah \(2016\)](#), un enfoque en el que están trabajando diferentes gobiernos y organismos reguladores de todo el mundo([ROFIEG, Expert Group](#)

on [Regulatory Obstacles to Financial Innovation](#), 2019). En resumen, este nuevo mercado complementa los sistemas tradicionales de financiación e inversión, apoya el crecimiento económico y la competencia en el sector financiero, y coadyuva al desarrollo de las economías a nivel local y global.

Sin embargo, al igual que sucede con las fuentes típicas de financiación e inversión, estas alternativas se ven expuestas a una serie de riesgos, tradicionales y nuevos, y estudiarlos es fundamental para lograr una gestión financiera responsable, prudente y eficaz. Administrar adecuadamente sus riesgos, contribuye a proteger el capital, orientar las decisiones de inversión, desarrollar y cumplir con el sistema normativo y regulatorio, gestionar eficientemente el crédito y anticiparse ante situaciones financieras desfavorables. En un entorno caracterizado por la inestabilidad económica, la gestión del riesgo financiero es una tarea fundamental para orientar las decisiones financieras y atenuar los efectos adversos de los riesgos financieros. Además, contribuye a que los individuos, las compañías y las entidades comprendan, gestionen y tomen, con conocimiento y medida, decisiones sobre sus operaciones financieras. Un sistema de control y gestión de riesgos, mejora la ventaja competitiva de las organizaciones, reduce las pérdidas potenciales de eventos inciertos y permite explotar las nuevas oportunidades de negocio, lo que garantiza la supervivencia y sostenibilidad de estas organizaciones ([de Araújo Lima et al., 2020](#); [Elahi, 2013](#); [Radner & Shepp, 1996](#); [Simons, 2019](#)). Las entidades y compañías con un desarrollado plan de gestión de riesgos están mejor preparadas para hacer frente a los retos complejos que surgen en la actualidad ([Grace et al., 2015](#); [Saeidi et al., 2019](#); [Tekathen & Dechow, 2013](#)).

Dentro de los riesgos financieros que hay que gestionar, hay uno que sobresale en P2P con mayor naturalidad por su objeto de negocio, el riesgo de crédito, entendido como el riesgo de que un prestatario no pueda cumplir con sus obligaciones financieras a tiempo, no efectúe los pagos requeridos, y entre en mora ([Capasso et al., 2020](#); [Chen et al., 2016b](#)). La evaluación de este riesgo se asocia con una tarea de clasificación binaria en la que se busca predecir el incumplimiento de las obligaciones crediticias ([Chen et al., 2016b](#)), y se

considera más ponderante en P2P por varias razones; presenta mayores niveles de asimetría de información ([Serrano-Cinca & Gutierrez-Nieto, 2016](#)), selección adversa, y riesgo moral ([Cummins et al., 2019](#)). Son los prestamistas quienes principalmente asumen el riesgo ([Serrano-Cinca & Gutierrez-Nieto, 2016](#)). Además, este es un mercado que aún carece de suficiente regulación ([Gao et al., 2020](#)).

Por esto, la ciencia de datos y las herramientas de aprendizaje automático, se han presentado como complemento de la modelación estándar, para enfrentar estos desafíos y potenciar el desarrollo de estas nuevas alternativas de financiación ([Bazarbash, 2019](#); [Financial Stability Board, 2017](#); [Giudici et al., 2020](#)). Estas herramientas ofrecen mayor precisión, permiten la inclusión en los modelos y algoritmos de diferentes fuentes de información, de variado tipo (estructurado y no estructurado), con volúmenes de información mucho más desafiantes, y con necesidades de respuesta con mayor oportunidad, recogiendo patrones de no linealidad y complejidad; atributos que los modelos tradicionales no siempre pueden capturar.

Las técnicas de aprendizaje automático, como núcleo de las herramientas de inteligencia artificial (AI), están transformando y revolucionando nuestra forma de abordar la gestión de riesgos en general, la toma de decisiones, y por supuesto, la gestión del riesgo financiero, siendo crédito uno de los campos en donde mayores aplicaciones hay ([Aziz & Dowling, 2019](#)).

Entiéndase aprendizaje automático, o machine learning por su popularidad y uso frecuente en inglés, como una rama de la ciencia de datos que utiliza la experiencia recogida en datos para tomar decisiones, generalizar patrones detectables, y crear reglas futuras de decisión desde los ejemplos dados. Esta área estudia teorías, algoritmos y aplicaciones de sistemas que intentan aprender como lo hacen los humanos ([Sugiyama, 2015](#)). [Ij \(2018\)](#) lo diferencia de los métodos estadísticos; aunque ambos en principio se usan para la predicción, los modelos estadísticos contienen como propósito la inferencia, que se consigue mediante la creación y el ajuste de uno o varios modelos de probabilidad específicos para cada proyecto. Los métodos de aprendizaje automático, por su parte, se centran en la predicción utilizando algoritmos de aprendizaje para encontrar patrones en datos a menudo complejos y difíciles

de manejar. Además, estas herramientas hacen suposiciones mínimas sobre los sistemas generadores de datos, y pueden ser eficaces incluso cuando los datos se recogen sin un diseño experimental cuidadosamente controlado y en presencia de interacciones no lineales complicadas. Como lo menciona el mismo [Ij \(2018\)](#), a pesar de los convincentes resultados de las predicciones de estos modelos, la falta de un modelo explícito puede dificultar la relación directa de las soluciones y relaciones subyacentes con los conocimientos existentes del campo de aplicación, temática en la que ahondaremos más adelante.

Dentro del machine learning, existe un conjunto de herramientas conocidas como supervisadas, que se da cuando en el aprendizaje, existe como referente una variable “dependiente” cuantitativa numérica o categórica a cuyos valores se les llama etiquetas. Se utilizan datos de entrenamiento etiquetados y una colección de ejemplos de entrenamiento recolectados en variables de entrada, para predecir luego las etiquetas de nuevos ejemplos de datos. Una de las tareas supervisadas más comunes es la “clasificación”, la que se da cuando la variable objetivo referente es categórica, y la “regresión” que ajusta los datos cuando la variable objetivo es cuantitativa ([Sarker, 2021](#)).

En particular, en P2P, estas nuevas herramientas de machine learning se han venido utilizando para diferentes problemáticas de la gestión del riesgo de crédito. Dentro de éstas se encuentra, por ejemplo, la verificación de identidad, la validación de información, evaluación de la experiencia financiera previa, el otorgamiento, el seguimiento, la venta cruzada, el proceso de cobranzas y la evaluación de fraude. Muchos son los algoritmos que se han venido estudiando para estas problemáticas. Destacan alternativas basadas en árboles de decisión (tree-based) como las presentadas en [Bastani et al. \(2019\)](#); [Ding et al. \(2017\)](#); [Misheva et al. \(2018\)](#); [Wang et al. \(2021\)](#); [Xia et al. \(2017a\)](#), usadas con frecuencia principalmente por su fácil interpretación y aplicación. También se han utilizado técnicas consideradas cajas negras, por carecer de explicabilidad e interpretabilidad, entre las que se encuentran propuestas usando máquinas de vectores de soporte (SVM) como las desarrolladas por [Cho et al. \(2019\)](#); [Ding et al. \(2017\)](#); [Duan \(2019\)](#); [Xia et al. \(2018\)](#); [Ye et al. \(2018\)](#); [Zhu](#)

et al. (2019), métodos que combinan predictores, ya sean simples o complejos, como los desarrollados por Boiko Ferreira et al. (2017); Nguyen Truong et al. (2019); Niu et al. (2019); Xia (2019), fundamentalmente basados en árboles de decisión (tree-based ensemble methods), y propuestas de Deep Learning y redes neuronales artificiales (RNA), como los trabajos de Byanjankar et al. (2015); Van-Sang et al. (2019); Wang et al. (2019); Zhao (2015).

También existen desarrollos con modelos Bayesianos (Guo et al., 2016; Jiang et al., 2018), métodos borrosos Ji et al. (2020), entre otras opciones.

Una de las tareas y fases más relevantes en la gestión del riesgo crediticio, que también subyace en el mercado P2P, es el otorgamiento de crédito. Este tópico corresponde con una de las problemáticas más estudiadas mediante aprendizaje automático. Para estudiarla, como tarea de clasificación, una de sus principales variables objetivo es el incumplimiento o default de una obligación crediticia. Varias de las propuestas desarrolladas para predecir esta variable con aprendizaje automático evidencian altos niveles de desempeño o predicción, con indicadores de clasificación superiores a los métodos tradicionales, esencialmente por su capacidad para identificar patrones con mayor complejidad (Leo et al., 2019). Estos resultados evidencian el potencial de su aplicación, sin embargo, como ya se mencionó, su uso viene también acompañado con nuevos desafíos y riesgos que limitan su explotación (Alonso Robisco & Carbo Martinez, 2022; Dupont et al., 2020; ROFIEG, Expert Group on Regulatory Obstacles to Financial Innovation, 2019).

Dentro de los más resaltables destacamos, la demanda de la interpretabilidad y transparencia. Como se mencionó, varios de los métodos de aprendizaje automático con mejores resultados suelen considerarse cajas negras por sus escasas posibilidades de interpretación. A su vez, las entidades reguladoras exigen elementos complementarios de validación e interpretabilidad para que estas nuevas posibilidades de modelización generen confianza y aumenten su uso (Basel Committee on Banking Supervision, 2017). Las entidades reguladoras y los usuarios finales necesitan de modelos interpretables y, por ello, suelen confiar

en los modelos estándar (Leo et al., 2019). Como se menciona en [ROFIEG, Expert Group on Regulatory Obstacles to Financial Innovation \(2019\)](#), la explicabilidad e interpretabilidad de la inteligencia artificial y de los modelos de aprendizaje automático, son elementos necesarios para proteger a los consumidores y las empresas, para facilitar los procesos de regulación, y para cumplir las expectativas de supervisión. Los modelos usados para la gestión de riesgo de crédito en el mercado P2P, en todas las etapas, incluyendo el otorgamiento, deben ofrecer a inversores y prestatarios explicaciones de los procesos de decisión y de las políticas de crédito, y a su vez, responder a las demandas de interpretación y seguimiento de los reguladores. Las plataformas utilizadas para desarrollar el mercado P2P requieren herramientas de medición del riesgo de crédito, no solo con altos niveles de desempeño y precisión en la predicción, sino también, como se ha reiterado, con interpretabilidad, componente demandado por las entidades reguladoras y de supervisión ([Board, 2017](#); [Carr & Bailey, 2018](#)).

A su vez, implícitamente, todas las propuestas alternativas de aprendizaje automático exigen un conocimiento del contexto, de la teoría, del dominio, de las variables que intervienen, de las limitaciones y de las posibilidades de producción de estos modelos, que garanticen su aplicabilidad, a lo que llamaremos, sentido de negocio.

Por lo anterior, nuestro trabajo busca evaluar la aplicabilidad de los modelos de aprendizaje automático como alternativa para la gestión del riesgo de crédito, específicamente, para el otorgamiento, en función de tres cualidades: la predictibilidad, comúnmente considerada, la interpretabilidad y el sentido de negocio, escasamente incorporadas y estudiadas en los modelos. Nuestra propuesta se evaluara en el contexto del mercado P2P, pero es trasladable como estrategia a cualquier campo que demande estos tres componentes.

1.2. Objetivos

Para lograrlo, nos planteamos cuatro objetivos asociados con tres artículos de investigación en el contexto del mercado señalado:

- Identificar las principales tendencias de investigación, necesidades, problemas u oportunidades relacionadas con la aplicación real de los modelos de aprendizaje automático en la gestión del riesgo y de la rentabilidad.
- Comparar los modelos de riesgo de crédito de otorgamiento tradicionales con los basados en aprendizaje automático en términos de rendimiento (precisión de la predicción) y de interpretabilidad.
- Desarrollar una propuesta de aprendizaje automático para la gestión del riesgo de crédito en la etapa de otorgamiento que incorpore el sentido de negocio a través de la inclusión de los costes y beneficios derivados de la correcta o incorrecta clasificación de las instancias.
- Comparar la propuesta de aprendizaje automático estándar con la propuesta del objetivo anterior en función de rendimiento a nivel financiero (en términos de rentabilidad) y de clasificación, así como del papel que juegan las variables en cada uno de ellos a través de las herramientas de explicabilidad.

A continuación explicamos cómo concretamos cada uno de estos objetivos.

Para el primero, llevamos a cabo una revisión de literatura. Realizamos un análisis bibliométrico y sistemático de los principales artículos publicados en el base de datos Web of Science, en el contexto del mercado de préstamos P2P, poniendo un foco en la gestión del riesgo, la rentabilidad, el sentido de negocio, y en el componente de explicabilidad. Para el análisis bibliométrico realizamos un análisis de redes de citación y co-ocurrencia con base en los desarrollos de [Van Eck & Waltman \(2010\)](#) y [Aria, M. and Cuccurullo \(2017\)](#) y sus softwares libres para la visualización y análisis de datos bibliográficos.

Para el segundo, desarrollamos una aplicación de modelado del riesgo de crédito de otorgamiento en el mercado P2P, comparando el rendimiento del modelo clásico de regresión logística frente al de varios de los modelos de aprendizaje automático más populares, incluyendo el modelo XGBoost ([Chen & Guestrin, 2016](#)), considerado de los más potentes como

se señala en [Boiko Ferreira et al. \(2017\)](#); [Chang et al. \(2022\)](#); [Li et al. \(2018a\)](#); [Xia et al. \(2017a\)](#). En el planteamiento de estos modelos prestamos especial atención al sentido de negocio mediante una correcta delimitación del problema y de las variables que intervienen. A su vez, comparamos el nivel explicabilidad de los dos enfoques, a través de una herramienta post-hoc que se basa en el uso de valores de Shapley ([Shapley, 1953](#)), desarrollada por [Lundberg & Lee \(2017\)](#), conocida como valores de SHAP. Más adelante proporcionaremos más detalles de esta herramienta.

A fin de concretar el tercero, hemos desarrollado una propuesta para el otorgamiento de crédito en el mercado P2P extendiendo el modelo XGBoost para que incorpore el sentido de negocio en varios puntos de la modelización, concretamente valorando el impacto financiero de los errores y aciertos de la clasificación. En concreto, introducimos esta información de beneficios en tres puntos del modelado: la estimación de la función de aprendizaje del algoritmo, la optimización de hiperparámetros y la función de decisión.

Para el cuarto objetivo, comparamos la propuesta con sentido de negocio del objetivo anterior con la propuesta clásica, tanto en términos de predicción como de negocio y rentabilidad financiera. Además, analizamos como explotan estos modelos las variables a través de las herramientas de explicabilidad. En concreto, identificamos las diferencias de importancia relativa, relación de dependencia, no linealidad y análisis del negocio, asociados con las variables incluidas. En la aplicación de este modelo nos hemos focalizado en los préstamos a pequeñas empresas del mercado P2P, un segmento más desafiante y menos estudiado en la literatura ([Nowak et al., 2018](#)). Las pequeñas empresas tienen más dificultades para acceder a alternativas de financiación que las grandes empresas, y el mercado P2P se convierte en una alternativa de financiación. Sin embargo, este es un segmento que se considera de alto riesgo debido a su opacidad informativa ([Mills & McCarthy, 2016](#)) y a que los que buscan estas alternativas, pueden ser prestatarios que han quedado por fuera de las fuentes tradicionales por evaluaciones de riesgo ([Nowak et al., 2018](#)).

1.3. Explicabilidad y sentido de negocio en el modelado del riesgo de crédito

Como lo hemos mencionado, adicional a la cualidad de predictibilidad, asegurada en la mayoría de las propuestas de modelación de la literatura, incorporamos y evaluamos dos elementos transversales que buscan incrementar la confianza y la aplicabilidad de las herramientas de aprendizaje automático en la gestión del riesgo de crédito en el mercado P2P; la explicabilidad y el sentido de negocio. Explicabilidad, claramente reclamada en la literatura ([Carvalho et al., 2019](#); [Doshi-Velez & Kim, 2017](#); [Lipton, 2018](#); [Miller, 2019](#)), y el sentido de negocio, resaltado también por algunos autores ([Bahnsen et al., 2014](#); [Verbraken et al., 2014](#); [Xia et al., 2017a](#)). Consideramos que ambos elementos son cualidades de un constructo mayor, la transparencia.

Explicabilidad la entendemos como la cualidad de comprender los principales determinantes de una decisión basada en un modelo ([Joseph, 2019](#)), o como la característica activa de un modelo que denota cualquier acción o procedimiento adoptado por una herramienta con la intención de aclarar o detallar sus funciones internas ([Barredo Arrieta et al., 2020](#)). Las explicaciones permiten identificar posibles sesgos o problemas en los datos y garantizar que los algoritmos funcionen como se espera ([Gilpin et al., 2018](#)). Según [Doshi-Velez & Kim \(2017\)](#), esta cualidad ayuda a garantizar la imparcialidad en la toma de decisiones, a detectar y corregir los sesgos, a resaltar posibles perturbaciones adversas que podrían cambiar la predicción y a asegurar que sólo las variables significativas, con soporte teórico, infieran en el resultado, garantizando causalidad, una verdadera relación subyacente en el razonamiento del modelo.

Cabe mencionar que explicabilidad como concepto se ha venido transformando. Es un constructo dinámico que tiende a intercambiarse con el concepto de interpretabilidad. Para varios autores están estrechamente relacionados y los usan intercambiabilmente ([Joseph, 2019](#); [Miller, 2023](#); [Molnar, 2021](#); [Silva et al., 2018](#)). Autores como [Ali et al. \(2023\)](#); [Barredo Arrieta et al. \(2020\)](#); [Doshi-Velez & Kim \(2017\)](#); [Rudin \(2019\)](#) los separan asociando

interpretabilidad como una cualidad intrínseca y pasiva de los modelos, mientras que explicabilidad lo consideran una cualidad activa, una acción, herramienta o interfaz, que permite clarificar las decisiones de un modelo o su funcionamiento. No obstante, al definir el concepto de interpretabilidad, o detallar su uso, requieren y asocian la acción de explicar. Es por esto, que todas estas acciones y cualidades se recogen en un área de investigación que ha venido creciendo en años recientes, el XAI (eXplainable Artificial Intelligence), una entidad no cerrada que recoge todos los propósitos asociados con hacer de los modelos soportados en inteligencia artificial, explicables, interpretables, transparentes y fiables, sin perder su cualidades de precisión en la predicción. Dada la complejidad de los dos conceptos, y la intercambiabilidad de su uso, nosotros también utilizaremos explicabilidad e interpretabilidad indistintamente en nuestro trabajo.

Es así que, la explicabilidad se puede dar en diferentes etapas, dimensiones y en distintos niveles de la modelación, todos elementos necesarios para garantizar su confianza y aplicación. [Ali et al. \(2023\)](#), habla, por ejemplo, de explicabilidad en cuatro ejes; en los datos, en el modelo, post-hoc (después de su estimación) y en la evaluación misma de las explicaciones. [Barredo Arrieta et al. \(2020\)](#), de otro lado, define la explicabilidad en función de la audiencia, identificando necesidades diferentes dependiendo del nivel de conocimiento y contexto de cada actor y usuario de un modelo: los expertos que dominan la teoría del campo de aplicación, administradores, los científicos de datos que construyen el modelo, los desarrolladores y los reguladores. [Hassija et al. \(2023\)](#) habla de dos tipos de interpretabilidad, la perceptiva que unifica todas las explicaciones e interpretaciones que son bien percibidas por los seres humanos, generalmente por pruebas visuales, y las que proviene directamente de estructuras matemáticas, unificando todas las explicaciones e interpretabilidades que se revelan de la estructura o de capas más profundas de los modelos. [Rudin \(2019\)](#) por su parte, promueve la búsqueda de modelos intrínsecamente interpretables y menciona cualidades como la monotonía de las variables, la causalidad y las restricciones estructurales asociadas con el conocimiento del dominio en donde se desarrolla el modelo, lo que llamamos, sentido de

negocio en el contexto de nuestro trabajo. [Molnar \(2021\)](#) y [Lundberg & Lee \(2017\)](#) por otro lado, hablan de interpretabilidad a nivel global y local, el primero, explicando las principales relaciones subyacentes de los modelos, y el segundo, asociado con explicaciones individuales de las decisiones. Estos autores identifican a su vez elementos de interpretabilidad como la importancia relativa de las variables y las relaciones de dependencia entre las variables de entrada y salida de los modelos. Esto muestra que XAI es un área multidimensional de investigación en pleno desarrollo.

En particular, el componente de explicabilidad en las propuestas de aprendizaje automático que proponemos, lo adicionamos utilizando los desarrollos de [Lundberg & Lee \(2017\)](#), [Lundberg et al. \(2018a\)](#) y [Lundberg et al. \(2019\)](#), quienes proponen un marco unificado para la interpretación y la explicación de predicciones llamados valores de SHAP (SHapley Additive exPlanations), basado en agregaciones de los valores de Shapley, con el apoyo de Local Interpretable Model-Agnostic Explanations (LIME) ([Ribeiro et al., 2016](#)) y otros métodos. Estos valores estiman la atribución de las variables mediante un modelo sustituto agnóstico post-hoc, incrementando los niveles de transparencia, ofreciendo elementos de interpretabilidad y explicabilidad a nivel global y local. Globalmente, estiman cuánto y cómo contribuye cada variable, positiva o negativamente, a la variable objetivo, y localmente, explican la manera como una observación dada se asigna como perteneciente a una categoría de predicción en función de las contribuciones de las variables. Con estas herramientas, aportamos capacidad explicativa a modelos complejos de diversa naturaleza, e identificamos cómo participan los factores de riesgo individualmente por obligación y a nivel agregado en los modelos.

Por su parte, el sentido de negocio, como elemento necesario para garantizar la aplicación de los modelos, lo entendemos como la comprensión del dominio, el estudio de su aplicabilidad, el conocimiento de los elementos de contexto, de la problemática estudiada y de las variables tenidas en cuenta, en función, no solo del riesgo, sino de la rentabilidad. Además de una evaluación de la aplicabilidad, de la puesta en producción. Es decir, un modelo se

considera que posee la cualidad de sentido de negocio si al menos: está bien construido de acuerdo con la problemática y teoría en contexto, y contiene una valoración económica de sus resultados. Existen otros atributos que aportan al componente de sentido de negocio, como la evaluación de la aplicabilidad y puesta en producción de los modelos o el desarrollo de herramientas de seguimiento, que no abordamos en nuestra propuesta.

Muchos de los artículos se enfocan en presentar una metodología innovadora con buenos niveles de precisión en la predicción, sin reflexionar lo suficiente sobre estos elementos. Por esto, no solo hemos revisado e identificado las variables que son realmente conocidas en el momento del otorgamiento de crédito, las que se pueden usar para predecir el incumplimiento, sino que en una de las propuestas, hemos incorporado el componente de rentabilidad junto con la minimización riesgo en un solo método a través de una categoría de algoritmos de aprendizaje automático llamados cost-sensitive o sensibles a los costos. Estos métodos, ponderan de forma diferencial los posibles resultados de la predicción en comparación con las categorías reales (verdaderos positivos, falsos positivos, etc.) derivados de la matriz de clasificación errónea (misclassification) o de confusión. Con esto, podemos asignar un coste (o beneficio), económico o no, a cada resultado de la matriz. De este modo, no sólo abordamos el problema del desequilibrio de clases, crucial en el riesgo de crédito cuando la clase de incumplimiento es minoritaria, sino que también nos permite incorporar mayor sentido de negocio a los métodos de clasificación, mediante la incorporación del componente de beneficio financiero o rentabilidad, potenciando la probabilidad de su aplicación en contextos reales. En el mercado P2P, son pocas las aplicaciones de alternativas sensibles a los costos que encontramos en la literatura. Algunas las hallamos en [Boiko Ferreira et al. \(2017\)](#); [Cho et al. \(2019\)](#); [Malekipirbazari & Aksakalli \(2015\)](#); [Wang et al. \(2018a\)](#); [Xia et al. \(2017a\)](#); [Ye et al. \(2018\)](#), en su mayoría, asociadas con una matriz de costos que no incluye valores financieros.

En nuestro caso, hemos incluido los beneficios en tres fases de la modelización: indirectamente, en la búsqueda de hiperparámetros, utilizando como función objetivo una función de

beneficios anualizados esperados derivado de la matriz de clasificación errónea, directamente, en el algoritmo de aprendizaje utilizando el gradiente y la hessiana de una función objetivo similar, e indirectamente, utilizando diferentes métodos de predicción de la categoría default (decisión) a partir de la probabilidad estimada. En esta última, mediante un proceso de optimización de los beneficios para un umbral apropiado ([Herasymovych et al., 2019](#); [Verbraken et al., 2014](#)), y a través del método de Bayes Minimum Risk classifier (BMR) ([Bahnsen et al., 2014](#)). Como contribución, resaltamos la posibilidad de incorporar en un solo proceso de modelización, dos de los objetivos de negocio a controlar en el otorgamiento de créditos, riesgo y rentabilidad. Riesgo al predecir la probabilidad de default, y rentabilidad al encontrar tal probabilidad maximizando una función de beneficios anualizados medios directa e indirectamente.

Al incorporar la explicabilidad y el sentido de negocio esperamos que el enfoque de aprendizaje automático no sólo genere mejores resultados en términos de desempeño, sino que nos permita entender la complejidad del problema, comprender el rol de las variables, la forma como intervienen, reconocer las oportunidades, en comparación con los métodos tradicionales, estándar, o insensibles a los costos.

1.4. Estructura del documento

En resumen, nuestro trabajo se presenta en cinco capítulos, el primero, esta introducción y puesta en contexto.

El capítulo 2 incluye el artículo de tendencias, brechas, recomendaciones y direcciones futuras identificadas en la modelización de riesgo-rentabilidad en el mercado de préstamos P2P. En este, a través de dos análisis, uno bibliométrico y otro sistemático, revisamos la literatura académica publicada durante la última década, identificamos las principales tendencias de investigación y encontramos brechas potenciales que limitan el uso de las propuestas desarrolladas. Identificamos los artículos más influyentes y la evolución de los tópicos más relevantes. Categorizamos los artículos según elementos metodológicos y aspec-

tos del contexto. Ponemos foco en las cualidades de predictibilidad, explicabilidad y sentido de negocio, y proporcionamos algunas recomendaciones y direcciones de investigación futuras en este campo.

El capítulo 3 incluye un artículo que describe la aplicación de un modelo de aprendizaje automático para el otorgamiento de crédito en el mercado P2P, al que se le incluye elementos de explicabilidad. En este, evaluamos el conocido modelo de regresión logística y comparamos su predictibilidad con varios algoritmos de aprendizaje automático. Además, comparamos su explicabilidad mediante valores de SHAP, basados en los valores de Shapley. Evaluamos la importancia relativa, la dependencia, cambio estructural y no linealidad de las variables incluidas. Mostramos algunas oportunidades que se abren con la aplicación de los modelos de aprendizaje automático al incluir la cualidad de explicabilidad.

En el capítulo 4, tomando como objeto de investigación parte de los hallazgos del capítulo 3, presentamos un artículo con una propuesta de aprendizaje automático sensible a los beneficios (profit-sensitive) para el otorgamiento de crédito de pequeñas empresas en el mercado P2P, con elementos también de explicabilidad. En la propuesta sensible a los beneficios incluimos no solo los costos financieros, sino la rentabilidad anualizada por obligación. Además, utilizamos herramientas de explicabilidad para estudiar los cambios de modelo entre los enfoques sensibles e insensibles a los beneficios, y evaluamos si se obtienen mejores resultados de negocio al incluir la rentabilidad como objetivo complementario al riesgo.

Y por último, el capítulo 5, que incluye las conclusiones, los principales hallazgos, la discusión, oportunidades y futuras líneas de investigación.

Es importante mencionar que, esta tesis, se ha construido por artículos, los cuales se han incluido en su idioma de publicación original (inglés) en cada uno de los capítulos correspondientes.

Capítulo 2

Artículo. Risk-return modelling in the P2P lending market: trends, gaps, recommendations and future directions

Este trabajo se ha publicado en la revista indexada en Web of Science con Journal Citation Reports (JCR) Electronic Commerce Research and Applications en el volumen número 49 de septiembre-octubre de 2021, p. 101079. (<https://doi.org/10.1016/j.elerap.2021.101079>)

Tiene como coautores a María Del Mar Camacho Miñano, María-Jesús Segovia y Javier Arroyo.

El trabajo ha sido apoyado por los proyectos europeo H2020 de coordinación en el marco del convenio de subvención No 825215, COST (European Cooperation in Science and Technology) con referencia COST Action 19130 y por el proyecto de investigación Santander-UCM de la convocatoria de 2019, con referencia PR87/19-22586.

Abstract

Peer-to-peer (P2P) lending is a market with significant growth in recent years. We review the academic literature published during the last decade on P2P lending to identify the main research trends and find potential gaps that limit stakeholders' use of research proposals. We perform both a bibliometric and systematic analysis. The bibliometric analysis will identify the most influential papers and the relationship and evolution of the main topics. In the systematic analysis, we categorized the documents according to methodological elements and business aspects. Remarkably, many proposals include artificial intelligence or machine learning algorithms. However, many of them lack a proper understanding of the application context, the definition of potential variables in a business framework, explainability, etc. Such elements should be recognized as essential elements to exploit their benefits. In this respect, we provide some recommendations and show future research directions.

Index terms— Risk management modeling, credit risk modeling, profit and investment modeling, P2P lending, machine learning, literature review, bibliometric analysis

2.1. Introduction

The peer-to-peer (P2P) lending market is based on technological platforms that charge a fee for the service of connecting lenders and borrowers. Borrowers can obtain credit directly from lenders at lower rates and with greater accessibility than conventional credit alternatives, while lenders can obtain higher returns than with other financial products (Emekter et al., 2015). The P2P lending market emerges as an alternative for investment and financing that breaks down some of the intermediation barriers offered by traditional banking. This market can be considered a collaborative economy agreement (Serrano-Cinca et al., 2015) that even allows the financial inclusion of people excluded from traditional systems (Claessens et al., 2018). It is considered a complementary and non-competitive market to conventional banking (Milne & Parboteeah, 2016), an approach that different governments and regulatory bodies worldwide are working on (ROFIEG, Expert Group on Regulatory Obstacles to Financial Innovation, 2019).

The rise of this new market is also associated with factors such as the reduction of intermediation costs, the incipient regulation customized to the business, and the evolution of big data and artificial intelligence tools that complement traditional modeling, allowing the experience to be personalized and the service to be improved (Financial Stability Board, 2017; Giudici et al., 2020). Claessens et al. (2018) differentiate leverage factors by country, market, and regulatory framework and detail some challenges of the new credit markets. Some of the most important risks are information asymmetries (Serrano-Cinca & Gutierrez-Nieto, 2016), adverse selection, and moral hazard (Cummins et al., 2019), and lenders bear credit risk on most platforms (Serrano-Cinca & Gutierrez-Nieto, 2016).

However, the P2P lending market also has inherent risks, for example, in China, the lack of regulation of the P2P market caused a surge in fraud and illegal activity. As a result, this market suffered a regression in China. As Gao et al. (2020) mentioned, even though the P2P market is based on technological innovation, it remains a financial problem that demands proper credit and financial risk management. The collapse of the Chinese P2P

lending market contributed to making visible the need for efficient regulatory policies that are customized to the challenges of the new markets, and, based on competent technicians, also highlighted aspects in [ROFIEG, Expert Group on Regulatory Obstacles to Financial Innovation \(2019\)](#). As it is a unique market, there is a need for customized risk management and regulation; however, it should be unified for the different countries involved, to truly protect investors and the entire market in a globalized context.

Quantitative modeling is considered an important activity for risk management. Nevertheless, it is also vital to maximize investors' profits and provide quality of service in a technological and globalized environment. Its use can be improved by recognizing factors such as appropriate business knowledge and understanding the associated risk factors, risk stages, and different techniques customized to the specific problems in its financial management. It is also essential to establish a modeling structure that ensures unbiasedness and efficiency and concerns predictive accuracy measures and what they mean in a business context ([Xia et al., 2017a](#)). It is equally important to have sufficient data, take advantage of various sources with validation and appropriate treatment ([Giudici et al., 2020](#); [ROFIEG, Expert Group on Regulatory Obstacles to Financial Innovation, 2019](#)), and know a range of powerful prediction techniques with explainability ([Ariza-Garzón et al., 2020](#); [Bussmann et al., 2020](#); [ROFIEG, Expert Group on Regulatory Obstacles to Financial Innovation, 2019](#)), among other aspects. These provide elements to guarantee its use, control, and trust from regulators and users.

Bearing all these ideas in mind and given the topic's relevance, we aim to find main research trends in the literature associated with the different quantitative methodologies used in managing the risk-return pair in the P2P lending market, which includes several of the aspects described above. In other words, our goal is to know what topics have already been studied and addressed in the P2P lending market and what are the biggest challenges and limitations in need of further study. To address these questions, we carry out a bibliometric and traditional systematic analysis of 104 papers from the Web of Science (WoS)

bibliographic database, which corresponds to papers on the topic published between 2010 and the first quarter of 2020 (April 15, 2020).

The bibliometric analysis will show the main references, trends, and different developments in the last decade. Our paper has a different approach from prior papers published, such as [Bachmann et al. \(2011\)](#), which is focused mainly on the description of existing platforms, the identification of the determinants associated with the proper management of financing sources, and the obtention of returns on investment during the second half of the first decade of this century. In our case, a systematic analysis complements the bibliometric analysis by analyzing several in-depth aspects of risk and return management in this market. The criteria used are the business focus of the proposals, the methodological contribution, the definition of the target variables, the type of modeling paradigm used, the quantitative contribution, and the description of the modeling stages. To conclude, we propose some recommendations and identify research gaps, which we recommend be investigated and deepened for adequate risk and profit management and the healthy development of the P2P lending market in our countries.

To achieve the purposes described, we organize the rest of the paper as follows. Section [2.2](#) presents the research methodology, with the definition of our research questions and the selection criteria of the papers studied. In section [2.3](#), we show the results. First, the bibliometric analysis includes general statistics of the worldwide P2P market, publications by region, indicators of cooperation among authors, and citation and co-occurrence analysis through graphs and cluster analysis of concepts and methodological elements, among other aspects. In the second part of the section, we show the results of a systematic analysis, where we include descriptive criteria, such as the datasets and software used. Furthermore, we characterize and classify the papers by some important aspects of the P2P business: the business problem involved, the data modeling paradigm used, methodological aspects highlighted in each study, and some modeling aspects required by market regulators and supervisors. In section [2.4](#), we discuss the results. Finally, in section [2.5](#), we summarize the

main conclusions and offer some recommendations for future work.

2.2. Research Methodology

The goal of this study is to provide an overview of the current research on the P2P lending market. To conduct it, we propose three research questions:

1. Research question 1 (RQ1): How has the literature on the P2P lending market evolved in the last ten years? More specifically, what are the countries that are researching it? Which venues are publishing this research? What are the most influential papers?
2. Research question 2 (RQ2): What are the main contents developed in the P2P lending market in the last decade? Datasets? Business problems? Methodology? Data modeling paradigm used? Current major research trends?
3. Research question 3 (RQ3): What are the main research gaps and future research directions in the P2P lending market?

To answer the research questions, we use two tools: a bibliometric analysis and a systematic literature review. More precisely, we use bibliometric analysis to answer RQ1. Part of the bibliometric analysis also helps us to answer RQ2. However, to deepen this question, we carry out a systematic literature review, where we carefully analyze and categorize each paper. To answer RQ3, we integrated the results of the two types of analysis, looking for underdeveloped topics and emerging aspects.

2.2.1. Research tools

In the following subsections, we present methodological details that support the bibliometric analysis with VOSviewer and how we have used them in our research. We also detail the criteria used to review and classify the papers in the systematic analysis.

Bibliometric analysis This methodology provides an overview of the research area, identifying the most relevant documents, topics, and trends associated with managing both risk and profitability in the P2P market. We explore citation and co-occurrence with network analysis using VOSviewer ([Van Eck & Waltman, 2010](#)), a software tool for analyzing and visualizing freely accessible bibliographic data on www.vosviewer.com.

In the first case, the network represents citations (links) between documents (nodes). We use the VOSviewer definition of citation link, where a link is created between two articles when one article cites the other. VOSviewer represents them as undirected links; however, we can generally infer the direction of the link from the year of publication of both nodes and the colour of the nodes and links.

In the co-occurrence network, nodes represent topics, and there is a link between topics when they appear in the same document, or more precisely, in their titles, abstracts, or keywords. We only account for the presence of the terms, regardless of the number of times the terms may appear.

For this network, VOSviewer performs NLP to identify the most relevant noun phrases, known as terms. It starts by tagging verbs, nouns, adjectives, etc. Then, it uses a linguistic filter that identifies the most relevant noun phrases. The filter is based on the distribution of co-occurrences and the randomness of their presence in the text through the Kullback-Leibler distance. The distribution of each term is compared with the overall distribution of co-occurrences over noun phrases. VOSviewer, therefore, calculates a relevance score for each noun phrase. Noun phrases have a high relevance score if they co-occur mainly with a limited set of other noun phrases and a low relevance score if their co-occurrences are close to randomness. For more detail on the process, see [Van Eck & Waltman \(2011\)](#).

The visualization of the co-occurrence network and the cluster analysis are based on the similarity matrix from the set of terms, where each cell represents the association strength or proximity index between two nodes (terms). We manually merged synonyms or closely related terms (e.g., decision tree also includes decision tree classifier, decision tree technique,

or decision-tree approach; artificial neural network includes terms such as backpropagation algorithm, bp neural network, bp neural network method, bp neural network model, bp neural network model, and mlp). We also discarded terms that were too general or could have an ambiguous meaning out of context (e.g., P2P lending, credit, market, efficiency, and evidence, among others).

For both the citation and co-occurrence networks, the final layout of the nodes in a two-dimensional space, VOSviewer uses a technique similar to multidimensional scaling. The mapping and clustering process maximization of a modularity function are unified in a single approach (Waltman et al., 2010). More precisely, node proximity implies a stronger relationship between the items, either by citation or co-occurrence. In addition, the node size is proportional to the number of citations in the citation network, and the number of occurrences in the co-occurrence network. For more details on VOSviewer, the interested reader can refer to its manual (Van Eck & Waltman, 2020).

Furthermore, we use two types of representations from VOSviewer:

1. Clusters by color on a network built using measures of similarity associated with the connection or relationship between the observed units, which can be papers or terms. The links represent the connection or the relationship between two units.
2. A temporal trend on a network. The same network described in the previous representation, but with a color degradation from green to yellow that represents how recent the unit's appearance is from oldest to most recent.

Finally, we use bibliometrix, a library on R developed by (Aria, M. and Cuccurullo, 2017), to analyze authorship and coauthorship statistics.

Systematic literature review To complement the bibliometric analysis, we carry out a systematic review of the literature where we manually analyze several aspects of the proposed approaches, focusing on risk and profit management. In particular, we consider the following aspects related to the modeling process:

1. Data sets: legal person (individuals or companies) considered and the country or region.
2. Software tools.
3. The business problem (e.g., default classification/default probability, profit scoring model, fraud, LGD, etc.)
4. The data modeling paradigm.
5. Methodological aspects (e.g., performance evaluation, the inclusion of new variables, class imbalance treatment, reject inference).
6. Other modeling aspects specifically associated with classification models that we consider crucial to ensuring the correct applicability and replicability of the results include cross-validation strategy, hyperparameter tuning, the use of logistic regression as a reference model, the statistical comparison of classifiers' performance, and explainability.

We reviewed the documents and classified them according to these criteria. The annex shows a table with some details related to the variables and the specific type of model used in each document.

2.2.2. Search and screening of the relevant papers

We used Web of Science (WoS), the main scientific journal database ([Birkle et al., 2020](#); [Gong et al., 2019](#)), to search for papers relevant to our literature review. To construct the query, we identified the following key aspects that the retrieved papers should include:

1. The main topic: peer to peer lending market.
2. Modeling methods and algorithms used.
3. Terms associated with risk management.

4. Terms focusing on the default or profit generated from the obligation loan.

For each aspect, the authors agreed on potential keywords, including synonyms. The final keywords used are the logical conjunction of the following:

1. TOPIC: (("peer to peer market." OR "P2P Platforms." OR ".online Peer-to-Peer credit") OR ("peer-to-peer credit") OR ("P2P credit") OR ("peer to peer credit") OR (".online Peer-to-Peer lending") OR ("peer-to-peer lending") OR ("P2P lending") OR ("peer to peer lending") OR ("social lending"))
2. TOPIC: (statistical analysis OR survival OR learning schemes OR supervised models OR classification models OR learning techniques OR model OR machine learning OR logistic OR cox OR method OR algorithm OR prediction)
3. TOPIC: (credit risk OR risk assessment OR risk OR risk assessment OR financial risk)
4. TOPIC: (loan evaluation OR loan default OR probability of default OR lenders' profitability OR profit OR creditworthiness OR creditworthiness rates OR scoring OR pay loan OR pay back the loan OR default OR status OR borrower status)

We also refine by years of publication between 2010 and the first quarter of 2020 (April 15, 2020). The main reason for selecting this range is that the prior literature review on a similar topic was [Bachmann et al. \(2011\)](#), who considered the first decade of the century. Our study aims to provide an updated review focusing on the most recent research trends, and we consider more criteria for analysis. We manually ensured that all the retrieved papers were relevant to our aims.

We discard two papers from the 106 documents retrieved by the query since their object of study is not associated with the P2P lending market. One focuses on the real estate market, and the other focuses on credit risk in general.

As a result, for the bibliometric analysis, we consider 104 documents published in conferences and journals. All of them are supposed to follow high-quality standards because they have been revised by blind peer review (Clarivate, 2021). We decide not to exclude conference papers because we consider that they may reflect emerging AI and ML trends, especially in computer science conferences Yli-Huumo et al. (2016). The analysis uses authors, citations, titles, abstracts, and keywords in English.

We considered the papers from the bibliometric analysis and removed non-English papers and those without full-text availability for the systematic review (Bae, 2018; Liu et al., 2019; Park & Choi, 2019; Soo H., 2016; Sungbok, 2018; Wu et al., 2018). We obtain a final sample of 98 papers.

2.3. Results

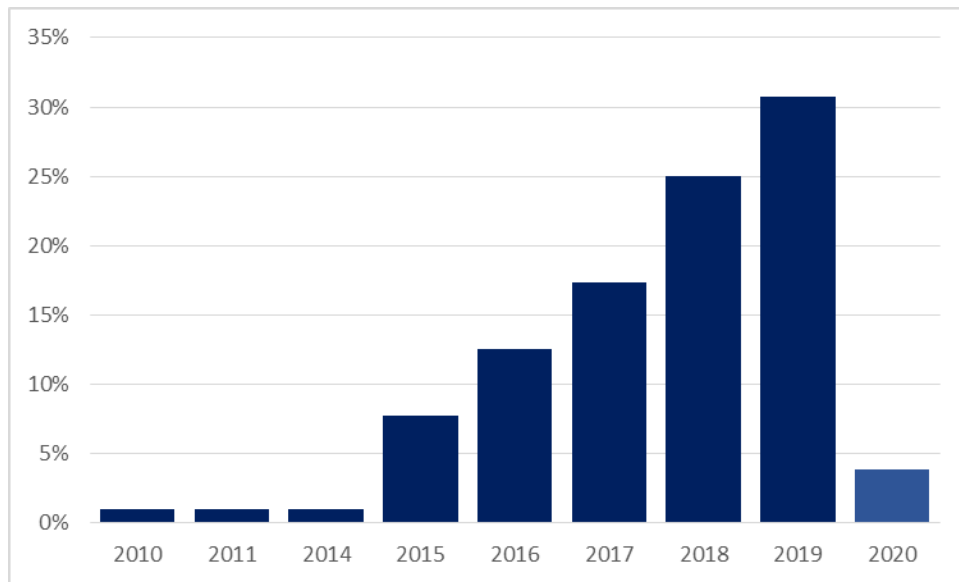
In this section, we present the results of the bibliometric and systematic analyses.

2.3.1. Bibliometric analysis

Analysis of the publication year, authors, source, and geographic distribution

Figure 2.1 shows the distribution of the selected papers by publication year. Given this trend, we can conclude that the P2P lending market is gaining attention (almost one-third of the papers were published in 2019). Since complete information has not been provided for 2020, it is difficult to assess whether it follows the trend from previous years.

Figura 2.1: Distribution of primary papers selected from 2010-2019 and up until April 2020



Source: Prepared by the authors using Clarivate Web of Science data

Table 2.1 of authoring statistics uses the bibliometrix library in R ([Aria, M. and Cuccurullo, 2017](#)). There are 12 papers with only one author. According to the data, on average, each author wrote 37% of a document (documents/authors), which is equivalent to 2.7 authors per document (authors/documents). In reality, there were 3.25 authors on average per document (authors appearance/documents). The difference is that the first metric is calculated with the total number of authors compared to the total number of documents. The second calculates the average number of authors per document. The collaboration index is almost 3, which means three co-authors per article, calculated on the multiauthored article set (total authors of multi-authored articles/total multiauthored articles), according to the indicator presented in [Koseoglu \(2016\)](#). Thus, we can conclude that there is frequent collaboration in this research area.

Tabla 2.1: Measures of contribution per author

Index	Value
Single-authored documents	12
Documents per author	0.37
Authors per Document	2.72
Co-Authors per Documents	3.25
Collaboration Index	2.96

Source: Prepared by the authors using bibliometrix and Web of Science data

Table 2.2 shows the publication venues of the retrieved papers. We can see that 44 papers (42 % of our sample) are published in top journals, that is, Q1 and Q2 journals in their respective WoS categories. The journals that appear more frequently are IEEE Access, an open-access interdisciplinary journal that aims for fast dissemination, and Electronic Commerce Research and Applications, a journal whose scope perfectly fits the topic of P2P lending. In addition, we can find journals from mathematics (including statistics and operational research), economics (from finance and electronic commerce), computer science (from the areas of artificial intelligence and information systems), and multidisciplinary areas. This shows that the problem can be addressed from different perspectives.

Tabla 2.2: Publication venues of the P2P lending market papers

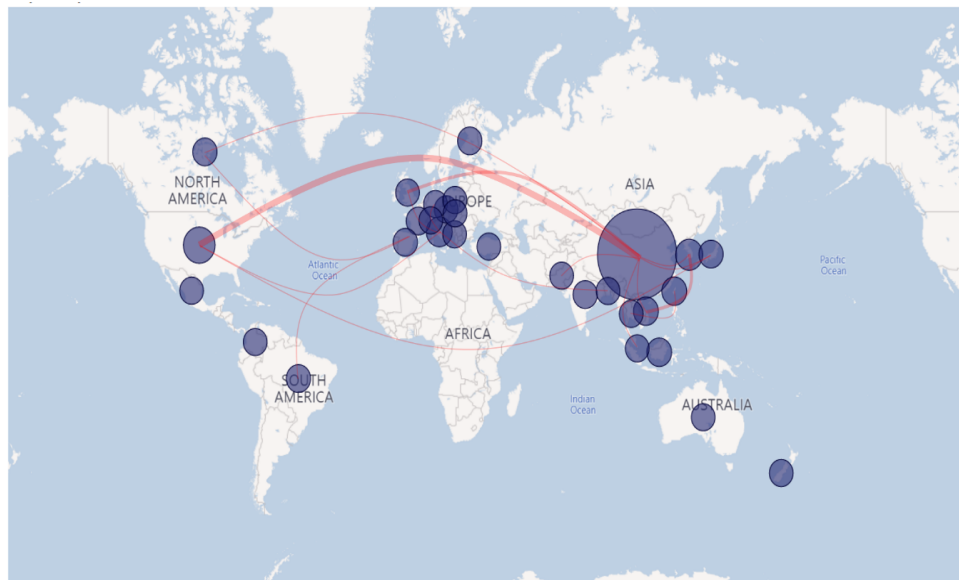
Documents Classification	Journals Information	Number
Q1 journals*	IEEE Access	6
	Electronic Commerce Research and Applications	5
	Journal of Management Information Systems	3
	Mathematics	2
	Expert Systems with Applications	2
	Decision Support Systems	2
	Journal of the Royal Statistical Society Series A-Statistics in Society	1
	Journal of Marketing Research	1
	European Journal of Operational Research	1
	World Wide Web: Internet and Web Information Systems	1
	Journal of The Franklin Institute-Engineering and Applied Mathematics	1
	Journal of Banking and Finance	1
	Finance Research Letters	1
	Journal of Computational and Applied Mathematics	1
	Engineering Applications of Artificial Intelligence	1
	Information Systems Frontiers	1
Q2 journals*	Electronic Commerce Research	3
	Physica A-Statistical Mechanics and its Applications	3
	North American Journal of Economics and Finance	1
	Sustainability	1
	Plos One	1
	Expert Systems	1
	Quality Engineering	1
	International Journal of Electronic Commerce	1
	Annals of Operations Research	1
	Journal of Forecasting	1
Other journals	-	24
Proceedings		36

*Quartiles according to the web of science 2019 Journal Impact Factor . We considered the best quartile if the journal was classified in more than one WoS category.

Source: Prepared by the authors using data from Clarivate Web of Science

Figure 2.2 shows the geographical distribution of authorship and coauthorship. Interestingly, authors from China participated in most of the papers (approximately 94.2%), followed by the USA (14.4%), Italy, and other Eastern countries. It is also noteworthy that the most relevant intercountry collaborations by the number of published articles are between the USA and China, followed by the UK and China, with a smaller proportion of intercountry contributions between authors from other Eastern countries.

Figura 2.2: Map of the final sample of paper participation by country, and contributions between countries according to coauthorship



Source: Prepared by the authors using Clarivate Web of Science data and bibliometrix

This result is unsurprising since P2P had tremendous growth in China over the last decade. However, at the time of this analysis, the market of this country has deteriorated due to the lack of an adequate regulatory framework. This market presented a significant increase in illegality and fraud, with few risk management practices. This overshadowed the entities shown to have a robust offer and efficient risk management, forcing the intervention in and closure of many platforms.

Figure 2.3 shows the number of P2P platforms worldwide, but the figure was not available for China. The top ten countries in order of the highest number of platforms are the USA,

the UK, Indonesia, Germany, Mexico, Switzerland, Latvia, Estonia, Spain, and South Korea.

Figura 2.3: Distribution of the number of lending platforms in the P2P market by country



Source: Compiled by the authors based on data from <https://P2Pmarketdata.com/>, accessed April 17, 2020

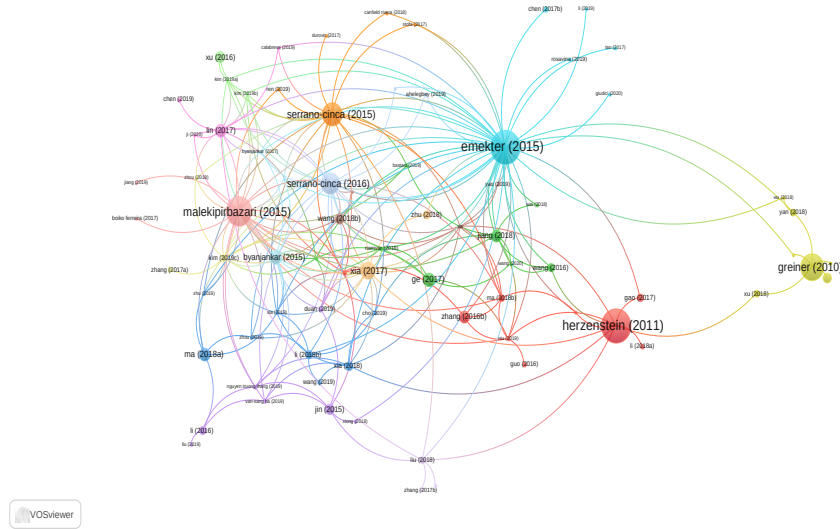
Citation analysis

In this section, we use citation analysis to identify the papers with the greatest number of links, which means the greatest number of citations. Figure 4 shows the citation network. Each circle represents a paper. The larger circles represent papers that have more citations. Colors are associated with each article. This map allows us to visualize the citation structure and identify the most relevant scientific community documents. We identify the most relevant papers to answer RQ1.

Consequently, according to Figure 2.4, the main articles are Emekter et al. (2015); Greiner & Wang (2010); Herzenstein et al. (2011); Malekipirbazari & Aksakalli (2015); Serrano-Cinca & Gutierrez-Nieto (2016); Serrano-Cinca et al. (2015); Xia et al. (2017a), with Emekter et al. (2015) being the most cited publication. However, the relevance of the papers has to be leveraged by the time they have been published because recent papers have fewer

opportunities to be cited. Thus, we believe that papers such as [Serrano-Cinca & Gutierrez-Nieto \(2016\)](#); [Xia et al. \(2017a\)](#) will have a greater impact in the coming years. Both papers deal with risk management models with a business focus, making them very suitable for the industry.

Figura 2.4: Main authors' citation graph



Source: Prepared by the authors using VOSviewer based on the Web of Science data

The most cited papers are mostly devoted to identifying the key factors in credit risk models in P2P lending. They mainly use statistical and econometric models. They propose cross-sectional models and some survival models. These models try to predict endogenous variables such as the credit contract's approval, default prediction, performance, or profit on a credit obligation. However, works such as [Malekipirbazari & Aksakalli \(2015\)](#); [Xia et al. \(2017a\)](#) try to improve credit risk management prediction and obtain profits by incorporating machine learning techniques, mainly decision tree methods.

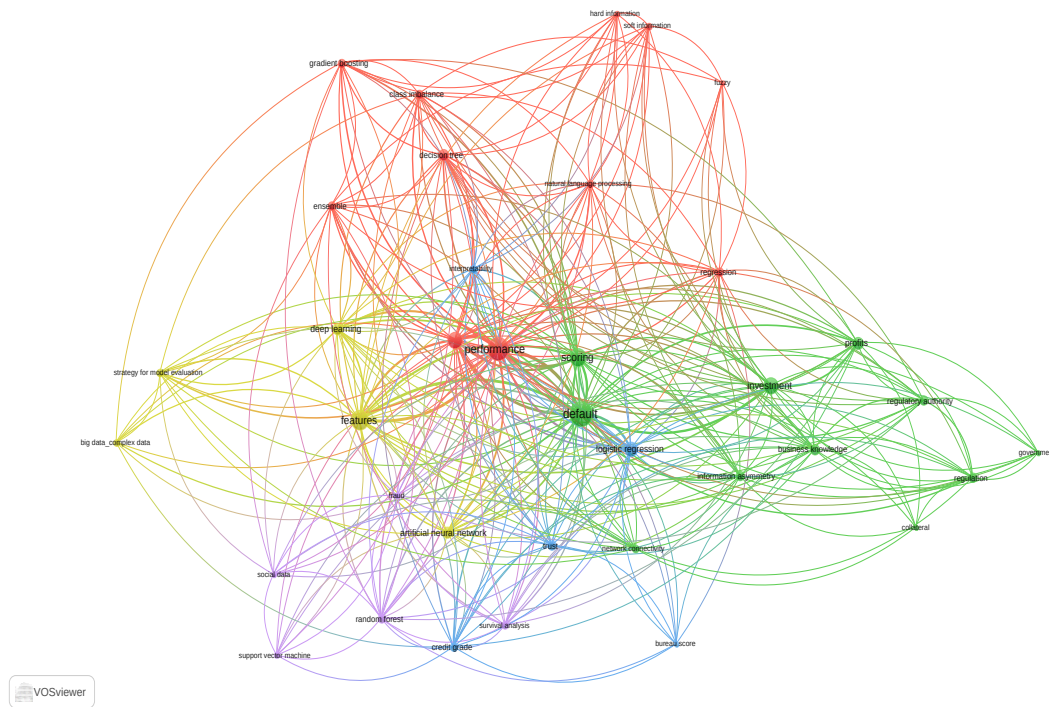
Co-occurrence analysis of abstract and titles

In this section, we identify the main topics associated with risk and profit management in P2P lending using a co-occurrence network constructed with the textual data in the

abstracts and the titles of publications. We also carry out a cluster analysis that will identify communities on the networks. We use VOSviewer for such purposes, which also provides mapping and visualization using multivariate methods. The resulting plots show as many labels and nodes as possible, prioritizing those most relevant nodes over the less critical ones.

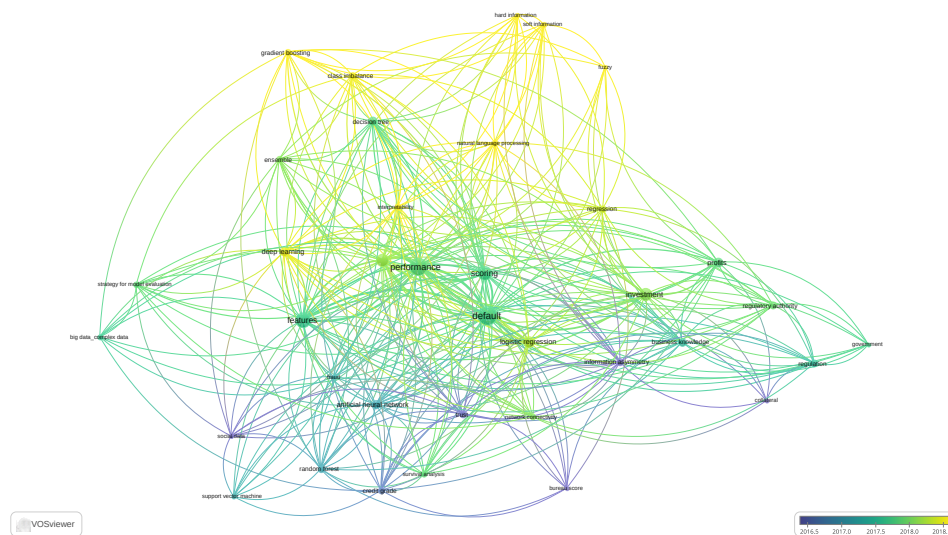
Figures 5 and 6 show the co-occurrence networks. The nearness of the nodes represents how closely the terms are related, while the size represents the number of publications in which each term appears. The color of the nodes in Figure 5 corresponds to the cluster to which each term has been assigned, while in Figure 6, lighter colors are associated with more recent publications, and darker colors are associated with older publications.

Figura 2.5: Graph of the co-occurrence and clusters for the concepts present in titles and abstracts



Source: Prepared by the authors using VOSviewer based on the Web of Science data

Figura 2.6: Graph of the co-occurrence and time trend for the concepts present in titles and abstracts



Source: Prepared by the authors using VOSviewer based on the Web of Science data

Tabla 2.3: Descriptive summary of clusters in the co-occurrence network based on abstracts and titles

Name of Cluster	Label	Occur	Avg. Oc- cur	Avg. Pub. Year	Avg. Pub. Year	Avg. Links	Avg. Links
Business and regulation (green)	default	60	16.5	2017.7	2017.5	36	19.8
	scoring	32		2017.7		33	
	investment	25		2018		26	
	profits	15		2017.8		19	
	business knowledge	11		2017.5		23	
	information asymmetry	9		2016.4		18	
	regulatory authority	9		2017.9		17	
	regulation	9		2017.4		15	
	network connectivity	6		2018		15	
	collateral	3		2016.3		7	
	government	3		2017.7		9	
Performance evaluation, ensemble models, and soft data (red)	performance	50	12.36	2017.8	2018.4	35	19.7
	classifier	26		2018.1		33	
	decision tree	12		2017.8		21	
	ensemble	8		2018		18	
	gradient boosting	8		2018.5		17	
	regression	7		2018.3		20	
	class imbalance	7		2018.4		18	
	natural language proces- sing	6		2018.7		18	
	soft information	5		2019.2		14	
	hard information	4		2018.5		13	
	fuzzy	3		2019		10	
Features and neural network models (yellow)	features	37	14.8	2017.6	2017.7	31	20
	deep learning	13		2018.4		24	
	artificial neural network	13		2017.2		22	
	strategy for model evalua- tion	6		2017.8		13	
	big data_complex data	5		2017.6		10	
Logistic regression and interpretability (blue)	logistic regression	21	9.2	2018.2	2017.3	26	18.8
	trust	9		2016.6		19	
	credit grade	7		2016.7		20	
	interpretability	6		2018.7		19	
	bureau score	3		2016.3		10	
Fraud, survival and other models (purple)	Random forest	7	4.8	2017.1	2017.2	20	15
	survival analysis	6		2017.8		16	
	fraud	4		2017.3		15	
	support vector machine	4		2017.3		11	
	social data	3		2016.3		13	

Source: Prepared by the authors using VOSviewer based on the Web of Science data

In addition, Table 2.3 presents a descriptive summary of the resulting clusters. The publication year variable was not considered in the clustering process, so it is shown in the table to complement the description of the clusters. The first column (*Name of Cluster*) represents the name we assigned to each cluster: "Business and regulation"(green color), "Performance evaluation, ensemble models and new sources of information"(red color), "Features and neural networks models"(yellow), "Logistic regression and interpretability"(blue) and "Fraud, survival, and other models"(purple). The names try to summarize the terms that belong to each cluster. In addition, we ordered the clusters in the table by relevance according to the frequency and centrality defined by the number of links. The second column symbolizes the labels of the terms grouped in each cluster. The third column (*Occur*) represents the number of papers in which each term appears. The fourth one (*Avg. Occur*) shows the average number of documents in which each term appears. The fifth (*Avg. pub. Year*) indicates the average year of publication of the term considering the articles that appear within. The sixth column (*Avg. Avg. pub. Year*) indicates the average year of publication for all the terms of the cluster. The seventh represents the number of node links (the links represent co-occurrence). Finally, the eighth (*Avg. Links*) shows the average number of links for each term within a cluster.

Given Table 2.3 and Figures 2.5 and 2.6, we will characterize the clusters below. The first cluster, "Business and regulation"(green color), includes terms associated with modeling objectives from the business point of view: "default"modeling (60 occurrences), "scoring"(32 occurrences), investment"(25 occurrences), and "profits"(15 occurrences) models. These terms are also characterized by their high centrality according to the number of links with 36, 33, 25, and 15 links. As evidenced in some papers, the modeling of risk or profit management must include solid and comprehensive regulatory guidelines (regulation,regulatory authority, and "government") to support the sustainability and development of the P2P lending market. The appearance of terms related to business and regulation highlights the need to study and consider this framework for action in developing modeling proposals

for risk and profit management. Oddly enough, the term "network connectivity" appears in this cluster because some papers use statistics derived from network analysis as risk determinants in recent years (Avg. Pub. Year of 2018). These network statistics somehow consider latent factors determined by the relationship among users, and that may represent a structure of contagion, which aligns with the regulatory perspectives. Another term appears as collateral, "less central (7 links), frequent (3 occurrences), and studied less recently (Avg. pub. Year 2016.3), to which more attention should be devoted as an element of hedging and risk management in P2P business.

A second cluster, both numerous and diverse, is the one we call "Performance evaluation, ensemble models, and soft data" (red color). This cluster is mainly represented by terms associated with classification as a modeling objective (the term "classifier" has 26 occurrences and 33 links) and with model evaluation focused on "performance" (50 occurrences and 35 links) through different metrics (see also Figure 2.5). Classification is one of the main topics in the literature, and the evaluation criteria in these papers typically focus on performance metrics. However, these papers usually neglect other aspects that would boost their application by regulators and users, as mentioned later. Some of these papers use "ensemble" models (8 occurrences and 18 links) or "gradient boosting" (8 occurrences and 17 links) models to improve performance, and "decision tree" methods (12 occurrences and 21 links) are typically used as benchmarks. The term "class imbalance" also appears because it is one of the main problems in credit risk. Finally, we can find recently treated terms such as "natural language processing" (Avg. pub. Year of 2018.7), "soft information" (Avg. pub. Year of 2019.2), "hard information" (Avg. pub. Year of 2018.5) or "fuzzy" (Avg. pub. Year of 2018.5). They account for new sources of information or new ways of representing information in the models.

We named the third cluster "Features and Neural Networks models" (yellow color). We mainly found terms related to "features" (37 occurrences and 31 links) and "big data_complex data" (five occurrences and ten links), both related to the exploration of new sources of information or new ways to include the information. Such venues seem to gain

strength together with modeling alternatives such as "artificial neural networks"(13 occurrences and 22 links). In recent years, "deep learning"(13 occurrences and 24 links) methodologies have been used by researchers (Avg. pub. Year of 2018.4) as an alternative and valuable tool when considering the varied "features and sources of complex information from "big data_complex data". The studies proposed in this line have highlighted the "strategy for model evaluation"(six occurrences and 13 links) as an element to be considered when evaluating and comparing models.

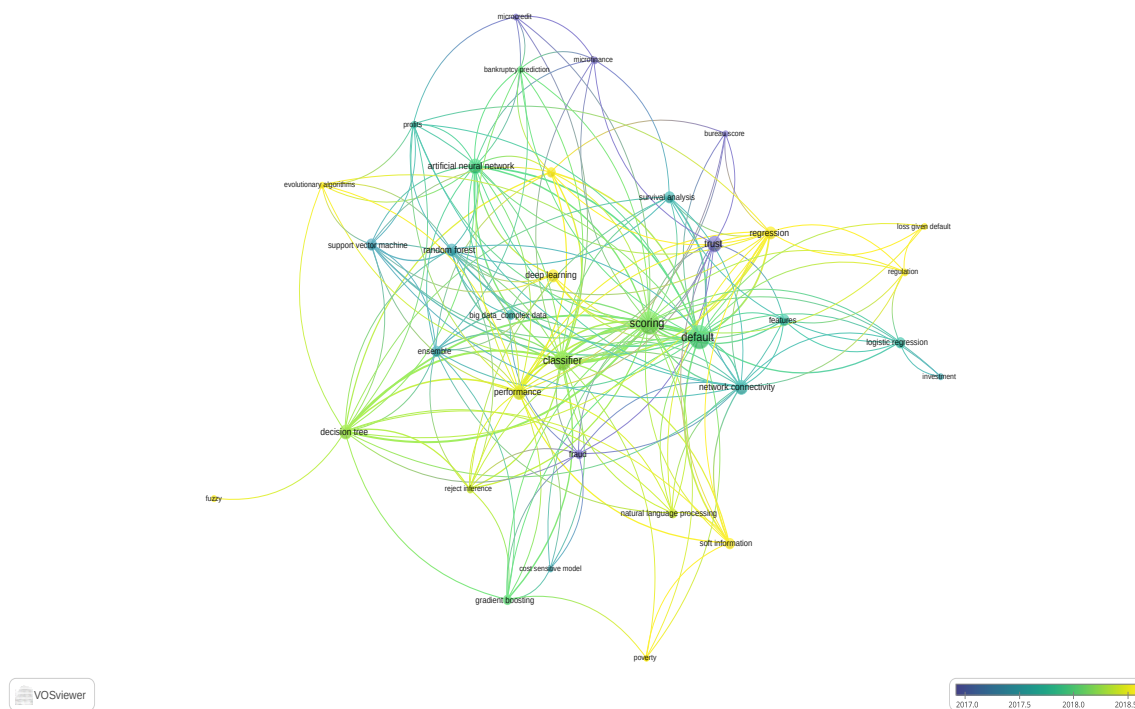
The fourth cluster, "Logistic regression and interpretability"(blue), includes elements of classic risk management such as "logistic regression"(21 occurrences and 26 links). "Logistic regression" is one of the central topics in the connectivity network because it is typically used as a benchmark model in many publications (see Figure 2.5). "Bureau score"(three occurrences and ten links) and "credit grade"(seven occurrences and ten links) appear with minor occurrences. While "credit grade" is used as an alternative modeling variable, "bureau score" is one of the relevant variables in most credit-granting models. When there is financial experience, it describes the behavior of the credit history of loan applicants. This cluster is associated with "interpretability"(six occurrences and 19 links). It is a rare and recently studied term (Avg. pub. Year of 2018.7) but is deeply associated with the "trust"(nine occurrences and 19 links) term. Likely, these features are greatly valued by regulators who have precise "interpretability" and "trust", which is why "logistic regression" continues to be a reference model, despite finding better performances in other modeling proposals.

The last cluster (purple color) presents the lowest frequency (Avg. Occur of 4.8) and centrality (Avg. Links of 15), and deals with terms not seen in recent years (Avg. Avg. pub. Year of 2017.2). We call this cluster "fraud, survival, and other models", and it includes some well-known machine learning methods such as random forest and "support vector machine". It also includes papers that consider "social data" information. They also propose "fraud" models (4 occurrences), which is a topic that barely worked, given the importance it has in the P2P market (see the collapse of the Chinese P2P market already mentioned).

"Survival models." also appeared a few times (six occurrences and 16 links). It is a classic alternative to default probability models and behavior models, which are widely used in expected loss estimation, thanks to longitudinal evaluation. Options of "survival models." are worth studying and deepening them from machine learning, for the possibility of aligning with the requirements of the leading international regulators, Basel, or IFRS.

Co-occurrence analysis of keywords

Figura 2.7: Graph of co-occurrence and time trend for concepts present in keywords



Source: Prepared by the authors using VOSviewer based on the Web of Science data

This section briefly comments on the co-occurrence network created using keywords (see Figure 2.7). The network is similar to that using the terms from the abstracts and titles. However, we can find new terms such as "microfinance" and "microcredit". This indicates the association of the P2P market with the supply of alternative financial services for low-

income people and small businesses, leveraging financial inclusion. We also find that the use of cost-sensitive models stands out in publications in mid-2017 as an alternative that includes business elements in credit risk management. One of the relevant components of the expected loss calculation, the "loss given default"(LGD), appeared less frequently around the year 2018, drawing attention to the need for more in-depth research on this topic for the P2P market. We can also see the evolution of the use of machine learning methods over time: support vector machine and random forest were mostly used around 2017. They later became gradient boosting, and more recently, the use of deep learning. "Logistic regression."also appears, as it is typically used as a benchmark method. Finally, the issue of reject inference also appears in the last two years. It could be an important element to incorporate into the new modeling alternatives to strengthen their credibility.

2.3.2. Systematic analysis

In this section, we carry out a systematic review of the literature on P2P lending. We analyze several aspects of the papers to answer RQ2. We have considered 98 English-written papers, as explained in section 2.2.2.

Dataset and software

We start our systematic review by analyzing the data set used to assess risk and profit management strategies in P2P lending. We identify the dataset by the name of the company and its country. Table 2.4 shows the frequency of the data sets in the literature. The Lending Club (LC) data set is the most popular. LC is one of the platforms with the highest volume of users in the United States. The data were used in 35 of the research papers analyzed, and approximately 35 % of the papers analyzed. Most likely, this is due to the quality, variety, and quantity of the data. Kaggle, the popular data science website, also makes it possible to work with data from many people from different countries. As a result, it is a data set studied from different perspectives and with standard and novel methodological strategies.

Tabla 2.4: Datasets

Data sets (Observation Unit)	Country / Region	Number	Authors/year
Lending Club (Individuals)	United States	30	Bastani et al. (2019); Boiko Ferreira et al. (2017); Cai and Zhang (2020); Calabrese et al. (2019); Cho et al. (2019); Duan (2019); Durovic (2017); Emekter et al. (2015); Goumieroux and Lu (2019); Jin and Zhu (2015); A. Kim and Cho (2019); Kim and Cho (2019a); Kim and Cho (2019b); Kumar et al. (2016); Ma, Sha, et al. (2018); Malekipirbazari and Aksakalli (2015); Namvar et al. (2018); Rodrigues et al. (2018); Serrano-Cinca and Gutierrez-Nieto (2016); Serrano-Cinca et al. (2015); Stofa (2017); Wang, Kou, et al. (2018); Wan et al. (2019); Wang et al. (2020); Wei et al. (2018); Xia et al. (2019); Ye et al. (2018); Zang et al. (2015); Zhou et al. (2018); Zhu et al. (2019)
Undefined platforms (Individuals and enterprises)	China	16	Chen et al. (2016); Guo et al. (2016); Jiang et al. (2019); Li, Ding, et al. (2018); Li et al. (2020); Niu et al. (2019); Ma, Zhao, et al. (2018); Wang, Jiang, et al. (2018); Wang et al. (2019); Xia and Li (2016); Xu and Zhang (2017); Yan et al. (2016); Yan et al. (2017); Yang et al. (2019); Yuan et al. (2018); Zhou et al. (2019)
Ppdai (Individuals)	China	6	Chen et al. (2019); Xu et al. (2015); Zhang, Wang, et al. (2016); Zhang, Jia, et al. (2016); Zhang, Geng, et al. (2017); Zhao (2015)
Renrendai (Individuals)	China	5	Gao et al. (2017); Li, Hao, et al. (2018); Liu et al. (2018); Tao et al. (2017); Yao et al. (2019)
Prosper (Individuals)	United States	5	Greiner & Wang (2010); Herzenstein et al. (2011); Ren et al. (2019); Tan et al. (2017); Wang et al. (2016)
No use data sets	-	5	Liu and Yan (2016); Pokorna and Sponer (2016); Pur et al. (2014); Wang (2018); Xiong (2018)
Lending Club, We.com (Individuals)	China and United States	3	Xia et al. (2017); Xia et al. (2018); Xia (2019)
Jinan Hengxin Micro-Investment Advisory Co., Ltd. (Individuals)	China	2	Zhang, Wang, Chen, Shang, et al. (2017); Zhang, Wang, Chen, Zhao, et al. (2017)
Paipaidai (Individuals)	China	2	Xinmin et al. (2019); Zhu (2018)
Two undefined institutions and Lending Club (Individuals)	Australia, Germany, and the United States	2	Nguyen Truong et al. (2019); Van-Sang et al. (2019)
Modefinance a ECAI (SMEs)	Europe	2	Giudici et al. (2020); Hadji-Misheva et al. (2018)
Bondora (Individuals)	Europe, Estonia, Finland, and Spain	2	Byanjankar et al. (2015); Byanjankar (2017)
Wangdaizhijia P2P online loan industry portal (Platforms)	China	2	Fu et al. (2019); Ge et al. (2017)
No Information	-	2	Ji et al. (2020); Li et al. (2016)
HNL-Home of Network Loan (p2p Lending Intermediaries)	China	1	Li et al. (2019)
LendingMarket (Individuals)	China	1	Xu and Chau (2018)
MyLending (Individuals)	China	1	Xu et al. (2016)
Paipai (Individuals)	China	1	Chen (2017)
Eloan (Individuals)	China	1	Jiang et al. (2018)
Three undefined institutions (Individuals)	Australia, Germany, and Brazil	1	Ding et al. (2017)
Yooli (Individuals)	China	1	Lin et al. (2017)
European External Credit Assessment Institution-ECAI (SMEs)	Europe	1	Ahelegbey et al. (2019)
Bandung (SMEs)	Indonesia	1	Rosavina et al. (2019)
Undefined platforms (Individuals)	Indonesia	1	Amalia et al. (2019)
Undefined platform (Individuals)	Mexico	1	Canfield (2018)
Funding Circle (Individuals)	United Kingdom	1	Pierrakis (2019)
KIVA non-profit organization (Individuals)	United States	1	Uddin et al. (2018)
Data simulated	-	1	Lee et al. (2017)

Source: Compiled by the authors

Apart from the Lending Club case, there is a strong tendency to use data from China. This may be caused by the growth of the Chinese market, the availability of the information, and the concern of an important group of researchers who seek to provide answers to the regulatory problems of this market. In the literature, the most commonly used datasets are from [ppdai.com](#), [renrendai.com](#), and [we.com](#).

As we saw in Figure 2.3, the relevant data set is closely related to the development of P2P markets by country, globally. The data was primarily from China and the United States, followed by continental Europe, the United Kingdom, Mexico, and, in a lesser proportion, even Southeast Asia.

It is important to mention that most of the studies have been carried out on loans to individuals. In a few cases, the P2P business lending market has been selected as the object of study, particularly for SMEs. In particular, the papers by [Ahelegbey et al. \(2019\)](#); [Giudici et al. \(2020\)](#); [Misheva et al. \(2018\)](#) use information from European External Credit Assessment Institutions (ECAIs), such as modeFinance, an agency specializing in companies and banks, and a fintech credit rating evaluation in Europe.

Table 2.5 focuses on the software used. *R* and *Python* are the most popular programming environments used. In Python, *Scikit-learn* (machine learning library) and *Keras* (neural networks library) stand out for their frequency of use. Other software such as *TensorFlow* for research involving deep learning, usually combined with *Python*, is also used. Software packages such as *SPSS*, *STATA*, and *WEKA* are also employed, but to a lesser extent (the first two are mainly for traditional statistical and econometric models). Interestingly, only 47 papers report the software used.

Tabla 2.5: Software used

Software	Number	Authors/Year
Python	10	Boiko Ferreira et al. (2017); Cho et al. (2019); Ding et al. (2017); Li et al. (2019); Li et al. (2020); Li, Ding, et al. (2018); Namvar et al. (2018); Rodrigues et al. (2018); Xia (2019); Xia et al. (2017)
R	9	Byanjankar (2017); Byanjankar et al. (2015); Calabrese et al. (2019); Chen et al. (2019); Giudici et al. (2020); Jiang et al. (2019); Wan et al. (2019); Wang, Jiang, et al. (2018); Xia et al. (2019)
SPSS	7	Emekter et al. (2015); Greiner and Wang (2010); Jin and Zhu (2015); Serrano-Cinca and Gutierrez-Nieto (2016); Yan et al. (2016); Yan et al. (2017); Xu and Chau (2018)
Python, TensorFlow	5	Bastani et al. (2019); Fu et al. (2019); Nguyen Truong et al. (2019); Van-Sang et al. (2019); Wang et al. (2019)
MATLAB	2	Yang et al. (2019); Zang et al. (2015)
STATA	3	Canfield (2018); Chen et al. (2016); Chen (2017)
Weka	3	Cai and Zhang (2020); Wang et al. (2016); Xu et al. (2016)
Python, SPSS, TensorFlow	1	Duan (2019)
Python, STATA	1	Niu et al. (2019)
R, Weka	1	Malekipirbazari and Aksakalli (2015)
R, Stanford CoreNLP, SentiStrength	1	Wang et al. (2020)
R, MATLAB, Portfolio Safeguard (PSG)	1	Wei et al. (2018)
R, Bazhuayu web-crawler software	1	Yao et al. (2019)
LISREL	1	Amalia et al. (2019)
Weka, Liblinear, LibSVM	1	Guo et al. (2016)

Source: Compiled by the authors

Business problem

This section proposes several categories associated with different problems with risk management, profit, and market knowledge. It is worth mentioning that a paper can be found in several categories because it can analyze and propose solutions made up of various models dealing with different aspects of the market. The defined categories are listed in Table 2.6.

Tabla 2.6: Business problem

Business problem	Number	Authors/Year
Default classification/Default probability	70	Ahelegbey et al. (2019); Bastani et al. (2019); Boiko Ferreira et al. (2017); Byanjankar et al. (2015); Byanjankar (2017); Cai and Zhang (2020); Calabrese et al. (2019); Canfield (2018); Chen (2017); Chen et al. (2019); Chen et al. (2016); Cho et al. (2019); Ding et al. (2017); Duan (2019); Durovic (2017); Emekter et al. (2015); Gao et al. (2017); Ge et al. (2017); Giudici et al. (2020); Gourieroux and Lu (2019); Guo et al. (2016); Herzenstein et al. (2011); Jiang et al. (2018); Jiang et al. (2019); Jin and Zhu (2015); A. Kim and Cho (2019); Kim and Cho (2019a); Kim and Cho (2019b); Kumar et al. (2016); Li et al. (2016); Li, Ding, et al. (2018); Li et al. (2020); Lin et al. (2017); Liu et al. (2018); Ma, Sha, et al. (2018); Ma, Zhao, et al. (2018); Malekipirbazari and Aksakalli (2015); Hadji-Misheva et al. (2018); Namvar et al. (2018); Nguyen Truong et al. (2019); Niu et al. (2019); Rodrigues et al. (2018); Serrano-Cinca et al. (2015); Serrano-Cinca and Gutierrez-Nieto (2016); Stofa (2017); Tan et al. (2017); Tao et al. (2017); Uddin et al. (2018); Van-Sang et al. (2019); Wang et al. (2016); Wang et al. (2019); Wang et al. (2020); Wang, Jiang, et al. (2018); Wang, Kou, et al. (2018); Xia (2019); Xia et al. (2017); Xia et al. (2018); Xia et al. (2019); Xinmin et al. (2019); Xu and Chau (2018); Yang et al. (2019); Ye et al. (2018); Yuan et al. (2018); Zang et al. (2015); Zhang, Jia, et al. (2016); Zhang, Wang, Chen, Shang, et al. (2017); Zhang, Wang, Chen, Zhao et al., (2017); Zhou et al. (2019); Zhu (2018); Zhu et al. (2019)
Risk and investment/Lending decision/Asset allocation	25	Amalia et al. (2019); Chen et al. (2016); Fu et al. (2019); Gao et al. (2017); Greiner and Wang (2010); Herzenstein et al. (2011); Ji et al. (2020); Lee et al. (2017); Li, Hao, et al. (2018); Pierrakis (2019); Ren et al. (2019); Rosavina et al. (2019); Tan et al. (2017); Tao et al. (2017); Wei et al. (2018); Xinmin et al. (2019); Xu and Chau (2018); Xu and Zhang (2017); Yan et al. (2016); Yan et al. (2017); Yao et al. (2019); Zhang, Geng, et al. (2017); Zhang, Wang, et al. (2016); Zhao (2015); Zhu (2018)
Profit Model/profit scoring model	6	Bastani et al. (2019); Cho et al. (2019); Serrano-Cinca and Gutierrez-Nieto (2016); Tan et al. (2017); Xia and Li (2016); Ye et al. (2018)
Market knowledge: Risk and regulation (Descriptive, theoretical, and qualitative analysis)	5	Liu and Yan (2016); Pokorna and Sponer (2016); Pur et al. (2014), Wang (2018); Xiong (2018)
Fraud	3	Li et al. (2019); Xu et al. (2015); Xu et al. (2016)
LGD	3	Gourieroux and Lu (2019); Xia et al. (2017); Zhou et al. (2018)
Prepayment	1	Wan et al. (2019)

Source: Compiled by the authors

The models that try to estimate the default and approximate the probability of this event, the “default classification/default probability” category, are the most frequent, with 70 papers, more than two-thirds of the studies analyzed. Many of them propose new methodological approaches, new variables, or combinations of various types of information. They typically exhibit remarkable results in classification and performance indicators (see Annex at the end of the paper for more details about each proposal). However, we draw attention to some limitations that we have found in many of the studies. First, although many of

them describe alternatives for managing credit risk and are even intended to be used in the lending process, few draw attention to selecting variables for this purpose. More precisely, many proposals include input variables that are indeed the output of the risk assessment (e.g., grade or levels of risk estimated by the platforms or interest rates, among others). As a result, these proposals could overestimate or underestimate the performance indicators, and they cannot be used because they do not suit a real-life lending assessment process. Second, few papers specify the type of models they propose for risk management (application model, behavioral model, collection model, among others). We believe that it is essential to make such a distinction since this determines the moment of application in risk management, the potential variables to be used, and the modeling methodology.

Third, there are many ways in which the default event is defined, but some of the studies do not even detail this aspect when it is crucial for interpreting the results. For example, having a model that flags defaults older than 30 days past-due in one year is different from a model that defines default as an unpaid obligation during the life of the credit. This kind of distinction must be considered when designing proposals, and in turn, it must be accompanied by a good understanding of the business and regulatory needs. As an example, some papers define default as charged-off transactions analyzed against fully paid transactions at the end of the credit life, including [Cho et al. \(2019\)](#); [Kim & Cho \(2019a,c,d\)](#); [Rodrigues et al. \(2018\)](#); [Serrano-Cinca & Gutierrez-Nieto \(2016\)](#); [Stofa \(2017\)](#); [Ye et al. \(2018\)](#). Other definitions can be found in the Annex 2.5.

All these facts limit the application due to a lack of uniformity and consistency. These two characteristics are objectives sought by regulatory bodies, especially for transnational growing markets such as P2P lending. From the joint work between academics, those who propose models for risk management, regulators, and platforms, tailored solutions can be presented, which generate greater confidence and probability of application. This synergy would ensure a better understanding of the business's nature and structure, which is reflected in better modeling, risk management, and service, enhancing the sustainability and

development of the P2P lending market.

We also want to highlight that most papers are focused on classification, which is valuable for lending processes. However, the robust estimation of the probability of default (PD) needs to be studied in more depth. This aspect is crucial in the application and lending processes; it is a component in the classification process and assists in the rate allocation and provisioning processes to better manage the risk, investment, and regulatory processes. The above implies calibration methods, incorporating forward-looking macroeconomic factors, and searching for other factors, which are recognized as entering the modeling processes, as shown in the list of independent variables in Annex 2.5.

Another line of research seeks to model investor's rate of return, identify the determinants, or offer sets of better investment alternatives through the "profit model/profit scoring" model alternatives (see Table 2.6). Within this category, there are studies, such as [Serrano-Cinca & Gutierrez-Nieto \(2016\)](#); [Xia & Li \(2016\)](#) or [\(Ye et al., 2018\)](#), that propose a profit score combining random forest (RF) and genetic algorithms (AG) (see the Annex for details of the objective, independent variables, and the methodologies used). Other papers in this category that also propose default prediction models incorporate these results in profit scoring models. These are the cases of [Bastani et al. \(2019\)](#) that propose a two-stage model using deep learning, or [Tan et al. \(2017\)](#) that establishes a recommendation model, maximizing the total profit of investors, incorporating a novel variable, and the time value of money prediction (TVM). Additionally, [Cho et al. \(2019\)](#), with their proposal for an instance-based entropy fuzzy support vector machine (IEFSVM) and several methods for treating imbalance, use the less-risky transactions' return on credit operations as a final criterion for investment decisions.

Another important field of research is related to the investment, recommendation, or decision models, which we have grouped under the category of "Risk and investment/Lending decision/Asset allocation" (see Table 2.6). Here, papers such as those by [Zhang et al. \(2016b\)](#) and [Zhang et al. \(2017a\)](#) use different collaborative filtering algorithms. [Ren et al. \(2019\)](#)

raise the problem of investment decisions under Markowitz’s theory, taking into account variables such as the number of days required for an application to be fully funded. [Xu & Zhang \(2017\)](#) propose an investment index per platform based on an analytic hierarchy process. [Ji et al. \(2020\)](#) proposed risk ranking through fuzzy methodologies and the interactive multicriteria decision-making method (TODIM). See the Annex for further details.

Moreover, we include [Lee et al. \(2017\)](#), who evaluate the impact of collateral on P2P lending as a tool to generate trust and a more efficient operation based on risk and investment. [Tan et al. \(2017\)](#) also propose a recommender system, allowing investors to choose the most profitable and least risky borrowers. In this category, we have also included the papers by [Herzenstein et al. \(2011\)](#), [Greiner & Wang \(2010\)](#), and [Chen et al. \(2016a\)](#). They study the determinants of variables, such as the percentage of the requested amount that is funded by investors, along with the final interest rate, or the variation between the maximum rate payable versus the assigned rate (see the Annex for details). In this way, they shed light on the applicant factors used by investors to make reliable investment decisions. For this same purpose, [Herzenstein et al. \(2011\)](#) study credit performance, including narrative variables, combined with demographic variables and loan characteristics. [Chen et al. \(2016a\)](#) propose mixed regression and survival models to investigate possible gender discrimination in China’s P2P credit loan market. They find empirical evidence that supports their hypothesis.

In the same category, the papers by [Tao et al. \(2017\)](#), [Zhu \(2018\)](#), [Xu & Chau \(2018\)](#), and [Xinmin et al. \(2019\)](#) that model both the prediction of default and whether the credit requested can be funded under traditional statistical methodologies. They also propose new variables related to information disclosure, guarantees, and the safety promise offered by the platforms. Concretely, [Xu & Chau \(2018\)](#)’s paper includes soft information, among other inputs, while [Xinmin et al. \(2019\)](#) focus on the impact of the variable for successful borrowing times. [Yao et al. \(2019\)](#) evaluate the influence of the description of the loan’s purpose with text mining techniques. Finally, [Wei et al. \(2018\)](#) use alternative transformations on the variables and an alternative objective function (buffered AUC) to understand the loan

approval processes. See the Annex for more details on each proposal.

In this broad category, we can find studies that try to characterize and explain the market. For example, [Yan et al. \(2016\)](#) and [Yan et al. \(2017\)](#) study the determinants of the number of investors per platform in China and identify variables that positively impact social capital, risk management, and operating duration. In addition to studying the determinants of the 30 days past-due default rate, [Gao et al. \(2017\)](#) studies the determinants of the percentage of bad debt in the total amount of a loan, including variables from a forward-looking and backward-looking segment, the latter to identify variables from previous experience with the platforms. On the other hand, [Li et al. \(2018b\)](#), using connectivity networks and regression models, analyze the systemic risk from topological variables derived from the network generated by the P2P market. [Zhao \(2015\)](#) simulates, using neural network models, expert risk assessments on credit granted. [Pierrakis \(2019\)](#) analyzes a survey of lenders from a platform in the United Kingdom using principal component analysis (PCA) to study their investment criteria and their motivation to invest. [Rosavina et al. \(2019\)](#) uses qualitative methods to explore the determinants of borrowing for SMEs in Indonesia. Additionally, they evaluate some aspects, including collateral, requirements, process duration, interest rate, costs, and profits.

We also include works such as [Fu et al. \(2019\)](#), which help us understand the dynamics of the P2P market for efficient decision making. They characterize the feelings of investors' comments using convolutional neural networks for sentence classification (TextCNN) and a long- and short-term memory neural network (LSTM). In addition, they predict an index for the daily trading volume of the Chinese platforms from a temporal series of changes that describes the dynamics of this variable and others.

Other lines of work that appear less frequency include relevant fields to guarantee risk management, sustainability, and market efficiency. They include "Fraud," the estimation of Loss Given Default or "LGD," and prepayment as a target variable. These areas may be research opportunities to strengthen the development of the P2P market. Another category

that has been explored very little is what we have called “Market knowledge: Risk and Regulation,” which presents descriptive, theoretical, contextual, and qualitative elements of the P2P markets, including benefits, actors, risks, and market failures. These works do not use data but focus on how P2P lending markets can ensure their development, continuity, and stability as an alternative for investment and financing. They suggest several elements of regulation and supervision, another essential and necessary line for this market to develop healthily.

Data modeling paradigm

In Table 2.7, we segment the methodological paradigms that underlie the quantitative tools used, bearing in mind again that a paper can use more than one paradigm depending on the objective of the research. Thus, it can appear in several categories.

Tabla 2.7: Paradigm used (93 papers, the five papers that did not use datasets were excluded)

Main paradigm used	Number	Authors/Year
Statistics and econometrics models	47	Ahelegbey et al. (2019); Amalia et al. (2019); Boiko Ferreira et al. (2017); Byanjankar (2017); Cai and Zhang (2020); Calabrese et al. (2019); Canfield (2018); Chen et al. (2016); Chen (2017); Chen et al. (2019); Cho et al. (2019); Ding et al. (2017); Durovic (2017); Emekter et al. (2015); Gao et al. (2017); Ge et al. (2017); Gourieroux and Lu (2019); Greiner and Wang (2010); Herzenstein et al. (2011); Jiang et al. (2018); Jiang et al. (2019); Li et al. (2016); Li et al. (2020); Li, Ding, et al. (2018); Li, Hao, et al. (2018); Lin et al. (2017); Liu et al. (2018); Hadji-Misheva et al. (2018); Namvar et al. (2018); Pierrakis (2019); Serrano-Cinca et al. (2015); Serrano-Cinca and Gutierrez-Nieto (2016); Stofa (2017); Tan et al. (2017); Tao et al. (2017); Wan et al. (2019); Wang et al. (2016); Wei et al. (2018); Xia and Li (2016); Xinmin et al. (2019); Xu and Chau (2018); Xu and Zhang (2017); Yan et al. (2016); Yan et al. (2017); Yao et al. (2019); Zhou et al. (2018); Zhu (2018)
Tree-based models	37	Bastani et al. (2019); Boiko Ferreira et al. (2017); Cai and Zhang (2020); Ding et al. (2017); Guo et al. (2016); Jin and Zhu (2015); Jiang et al. (2019); A. Kim and Cho (2019); Kim and Cho (2019a); Kim and Cho (2019b); Kumar et al. (2016); Li, Ding, et al. (2018); Li et al. (2020); Ma, Zhao, et al. (2018); Malekipirbazari and Aksakalli (2015); Ma, Sha, et al. (2018); Hadji-Misheva et al. (2018); Namvar et al. (2018); Nguyen Truong et al. (2019); Niu et al. (2019); Rodrigues et al. (2018); Tan et al. (2017); Van-Sang et al. (2019); Wang et al. (2016); Wang et al. (2020); Wang, Kou, et al. (2018); Wang, Jiang, et al. (2018); Xia et al. (2017); Xia et al. (2018); Xia et al. (2019); Xia (2019); Xu et al. (2015); Xu et al. (2016); Ye et al. (2018); Zhang, Jia, et al. (2016); Zhou et al. (2019); Zhu et al. (2019)
Support Vector Machine	25	Bastani et al. (2019); Cho et al. (2019); Ding et al. (2017); Duan (2019); Fu et al. (2019); Guo et al. (2016); Jiang et al. (2018); Jin and Zhu (2015); A. Kim and Cho (2019); Kim and Cho (2019a); Kim and Cho (2019b); Li et al. (2020); Malekipirbazari and Aksakalli (2015); Nguyen Truong et al. (2019); Rodrigues et al. (2018); Van-Sang et al. (2019); Wang, Kou, et al. (2018); Xia et al. (2018); Xia (2019); Xu et al. (2015); Xu et al. (2016); Yang et al. (2019); Ye et al. (2018); Zhou et al. (2019); Zhu et al. (2019)
Use ensemble models	22	Bastani et al. (2019); Boiko Ferreira et al. (2017); Cho et al. (2019); Jiang et al. (2019); Kim and Cho (2019a); Kim and Cho (2019b); Ma, Sha, et al. (2018); Ma, Zhao, et al. (2018); Malekipirbazari and Aksakalli (2015); Nguyen Truong et al. (2019); Niu et al. (2019); Rodrigues et al. (2018); Tan et al. (2017); Van-Sang et al. (2019); Wang, Jiang, et al. (2018); Wang et al. (2020); Xia et al. (2017); Xia et al. (2018); Xia et al. (2019); Xia (2019); Ye et al. (2018); Zhu et al. (2019)
Deep Learning	11	Bastani et al. (2019); Duan (2019); Fu et al. (2019); Kim and Cho (2019a); Kim and Cho (2019b); Li, Ding, et al. (2018); Li et al. (2020); Nguyen Truong et al. (2019); Van-Sang et al. (2019); Wang et al. (2019); Zhang, Wang, Chen, Shang, et al. (2017)
Artificial Neural Networks	7	Byanjankar et al. (2015); Jin and Zhu (2015); Xu et al. (2015); Zang et al. (2015); Yuan et al. (2018); Zhang, Wang, Chen, Zhao, et al. (2017); Zhao (2015)
Bayesian-probabilistic models	6	Boiko Ferreira et al. (2017); Guo et al. (2016); Jiang et al. (2018); Rodrigues et al. (2018); Wang et al. (2016); Wang, Kou, et al. (2018)
Recommendation algorithms	4	Ren et al. (2019); Tan et al. (2017); Zhang, Wang, et al. (2016); Zhang, Geng, et al. (2017)
Propose heterogenous ensemble	6	Guo et al. (2016); Zhou et al. (2019); A. Kim & Cho (2019); Li, Ding, et al. (2018); Li, Ding, et al. (2018); Namvar et al. (2018)
Fuzzy methods	2	Cho et al. (2019); Ji et al. (2020)
Qualitative models	2	Rosavina et al. (2019); Uddin et al. (2018)
Propose homogenous ensemble	1	Ding et al. (2017)

Source: Compiled by the authors

The largest category is “Statistics and econometrics models,” with 47 papers, almost all of the papers analyzed. These papers typically use explanatory models to identify the type of

relationship and measure the impact of the determinants of the dependent variable (default, fraud, investment, return, etc.). We believe that it is the most widely used precisely because of its capacity for interpretation, the ease of comparison with business theory and knowledge, the ease of estimation and implementation, and the greater generalized understanding of researchers, users, and regulators.

The next category is “Tree-based models,” with 37 papers. Decision trees have been used throughout the last decades as an alternative for managing credit risk, mainly because of their easy interpretation and application. However, in recent years, their application has increased due to advances in tree-based ensemble methods, which obtain better classification performance than standard decision trees, although explainability is sacrificed. The main advances have come from homogeneous ensemble techniques, such as bagging and boosting, random forest (RF), and extreme gradient boosting (XGBoost is the most commonly used (see the Annex 2.5)). The developments of the boosting technique have also brought the application of algorithms such as adaptive boosting (AdaBoost) and light gradient boosting machine (LightGBM), among other alternatives. In the tree-based category, 22 papers use such tree-based ensemble methods.

In the category of “Propose heterogeneous ensemble,” six papers propose their own heterogeneous ensemble. As in homogenous ensemble techniques, they aim to increase the learning ability of algorithms, reduce the variability of classifiers or predictive models, generalize performance, and guarantee scalability. They are also presented as an alternative to the problem of class imbalance. In other words, these methods are offered as an alternative to reduce the over adjustment of individual options and to produce superior predictive results and a better capacity for generalization, which are visible in data sets with the class imbalance problem. On the other hand, only one paper, [Ding et al. \(2017\)](#), “proposes homogenous ensemble” based on the clustering processes.

Another widely used method is the “support vector machine (SVM),” with 25 papers. However, most papers present SVM as an alternative model for comparison (see Annex for

more details of the models used in each paper). In the case of artificial neural networks (ANNs), seven papers use them, mainly backpropagation algorithms. However, the more sophisticated “deep learning” methods are being more commonly used (11). These factors account for the great development and applicability of these methods. The most popular techniques are long short-term memory (LSTM) techniques, wide deep learning (WDP), deep neural networks based on multilayer perceptrons, and convolutional neural networks (CNNs), among other options.

Finally, smaller categories include methods such as “Bayesian-probabilistic models,” with six papers, mainly with the Naive Bayes (NB) algorithm, and “recommendation algorithms” with four papers focusing on the choice of profitable investment alternatives as a tool for platforms and investors in borrowing. “Fuzzy methods” (2) and “qualitative models” (2) are less frequent.

Methodological aspects

In this section, we focus on the methodological aspects of the proposed models. We have identified ten nonexclusive categories that are shown in Table [2.8](#).

Tabla 2.8: Methodological aspects (93 papers, the five papers that did not use datasets were excluded)

Methodological aspects	Number	Authors/Year
Performance/evaluation	44	Bastani et al. (2019); Boiko Ferreira et al. (2017); Byanjankar et al. (2015); Byanjankar (2017); Cai and Zhang (2020); Calabrese et al. (2019); Chen (2017); Chen et al. (2019); Cho et al. (2019); Duan (2019); Jiang et al. (2019); Jin and Zhu (2015); A. Kim and Cho (2019); Kim and Cho (2019a); Kim and Cho (2019b); Kumar et al. (2016); Li, Ding, et al. (2018); Li et al. (2019); Li et al. (2020); Liu et al. (2018); Ma, Sha, et al. (2018); Malekipirbazari and Aksakalli (2015); Nguyen Truong et al. (2019); Niu et al. (2019); Rodrigues et al. (2018); Uddin et al. (2018); Van-Sang et al. (2019); Wang et al. (2019); Wang et al. (2020); Wang, Jiang, et al. (2018); Wang, Kou, et al. (2018); Xia et al. (2017); Xia et al. (2018); Xia (2019); Xia et al. (2019); Xu et al. (2015); Yang et al. (2019); Ye et al. (2018); Yuan et al. (2018); Zang et al. (2015); Zhang, Wang, Chen, Shang, et al. (2017); Zhang, Wang, Chen, Zhao, et al. (2017); Zhao (2015); Zhu et al. (2019)
Explanation	37	Amalia et al. (2019); Byanjankar (2017); Calabrese et al. (2019); Canfield (2018); Chen (2017); Chen et al. (2019); Chen et al. (2016); Durovic (2017); Emekter et al. (2015); Gao et al. (2017); Ge et al. (2017); Giudici et al. (2020); Gourieroux and Lu (2019); Greiner and Wang (2010); Herzenstein et al. (2011); Jiang et al. (2018); Lee et al. (2017); Li, Hao, et al. (2018); Lin et al. (2017); Liu et al. (2018); Hadji-Misheva et al. (2018); Niu et al. (2019); Serrano-Cinca et al. (2015); Serrano-Cinca and Gutierrez-Nieto (2016); Stofa (2017); Tao et al. (2017); Wan et al. (2019); Xia and Li (2016); Xinmin et al. (2019); Xu and Chau (2018); Yan et al. (2016); Yan et al. (2017); Yang et al. (2019); Yao et al. (2019); Zhang, Jia, et al. (2016); Zhou et al. (2018); Zhu (2018)
Inclusion of new variables	23	Ahelegbey et al. (2019); Fu et al. (2019); Gao et al. (2017); Ge et al. (2017); Giudici et al. (2020); Guo et al. (2016); Herzenstein et al. (2011); Jiang et al. (2018); Li et al. (2019); Li, Hao, et al. (2018); Ma, Zhao, et al. (2018); Hadji-Misheva et al. (2018); Niu et al. (2019); Wang et al. (2016); Wang et al. (2019); Wang et al. (2020); Wei et al. (2018); Xia et al. (2019); Xu et al. (2015); Xu and Chau (2018); Yao et al. (2019); Zhang, Jia, et al. (2016); Zhu (2018)
Class imbalance treatment	9	Bastani et al. (2019); Boiko Ferreira et al. (2017); Cho et al. (2019); Ding et al. (2017); Malekipirbazari and Aksakalli (2015); Wang, Kou, et al. (2018); Xia et al. (2017); Ye et al. (2018); Zhu et al. (2019)
Feature selection	9	Cai and Zhang (2020); Jiang et al. (2019); Jin and Zhu (2015); Kim and Cho (2019a); Kim and Cho (2019b); Nguyen Truong et al. (2019); Van-Sang et al. (2019); Xu et al. (2015); Zhu et al. (2019)
Rank by risk or profitability	6	Ji et al. (2020); Ren et al. (2019); Tan et al. (2017); Xu and Zhang (2017); Zhang, Wang, et al. (2016); Zhang, Geng, et al. (2017)
Descriptive analysis	3	Durovic (2017); Pierrakis (2019); Rosavina et al. (2019)
Reject inference	2	Xia et al. (2018); Xia (2019)
Outliers treatment	1	Li et al. (2016)

Source: Compiled by the authors

The largest category is “Performance/evaluation” because 44 publications address this aspect. Those articles focus on default classification and use performance measures as the main evaluation criterion. The most frequent measures are accuracy (ACC), even if the data are typically imbalanced, the area under the ROC curve (AUC), precision, and recall. On the other hand, measures such as the F1 score, the harmonic mean between recall and precision, Kolmogorov-Smirnov (KS), or the H-measure present a lower frequency.

Another category we want to draw attention to is “Reject inference,” which is a modeling component not very frequent in the literature. The articles in this category propose to reject inference techniques to correct the sampling bias generated by modeling only with the set of accepted loans, even though the models’ use is established on the whole set of

loan applications. This bias can affect the predictability of the models. The input feature distributions and the label proportion of the target variables could be very different if the data set would include the rejected applications. For this purpose, [Xia et al. \(2018\)](#) and [Xia \(2019\)](#) combined classifier gradient boosting decision tree (GBDT) techniques with semisupervised methods and outlier detection methods, respectively (see Annex for details).

The category of “Class imbalance treatment” includes the works that tackle the problem of unbalanced classes typical in P2P lending, where the percentage of default, fraud, or accepted applications is the minority class. This may be problematic since the algorithms mainly learn from the majority class. Additionally, many performance metrics assign the same cost to the classification errors, which is far from the market needs ([Xia et al., 2017a](#)). Different strategies can alleviate this problem, such as resampling methods to balance classes, the adjustments of misclassification costs or the definition of cost functions to be minimized (either directly in the learning processes or after it), the calibration of the cutoff value of the models employing cost-sensitive models, etc. More precisely, [Bastani et al. \(2019\)](#) and [Zhu et al. \(2019\)](#) use resampling techniques such as under sampling, oversampling, or synthetic minority oversampling techniques (SMOTEs). Regarding cost-sensitive models, [Xia et al. \(2017a\)](#) provides an excellent description of this line of research and proposes to adjust the estimation process and learning of the extreme gradient boosting algorithm (XGBoost), taking into consideration the costs of misclassification. Similarly, [Wang et al. \(2018a\)](#), [Ye et al. \(2018\)](#), [Boiko Ferreira et al. \(2017\)](#), and [Cho et al. \(2019\)](#) use and compare cost-sensitive strategies with resampling strategies. However, [Cho et al. \(2019\)](#) additionally propose and contrast an instance-based entropy fuzzy support vector machine model (IEFSVM) for imbalanced datasets. They select the least risky registers to compose a portfolio that seeks high expected returns based on an investment model. [Ye et al. \(2018\)](#), in turn, compare their cost-sensitive approach to standard methods based on profit indicators. [Ding et al. \(2017\)](#) proposes an ensemble using a novel technique of under sampling supported by clustering techniques. Likewise, [Malekipirbazari & Aksakalli \(2015\)](#), on their

random forest (RF) proposal, incorporates a cost-weighted matrix that increases the cost of erroneous classification associated with borrowers with poor default behavior through a cost-sensitive meta-algorithm incorporated in the WEKA software (see the Annex for more details).

On the other hand, another group of papers is focused on identifying determinants, validating relationship hypotheses, and understanding the types of relationships of different potential variables with the target variable. For example, default, fraud, approval, interest rate decrease, and the number of investors per platform, among others (see Annex for the list of target variables used). We have classified these studies in Table 8 under the “Explanation” category. This aspect is demanded by regulators, supervisors, and users of machine learning methods. It is strongly associated with transparency and confidence in credit market risk and profit management, particularly in P2P lending. We found that 31 out of 37 used classic statistical and econometric methods, such as linear regression models and logit, probit, or survival models. Such models include the explanatory component through the interpretation of coefficients and testing of inferential hypotheses about them.

Among the proposals that try to emphasize the explanation component, we highlight [Jiang et al. \(2018\)](#) and [Yao et al. \(2019\)](#) for their methodological innovations. The authors, using text mining tools such as latent Dirichlet allocation (LDA), among other techniques, seek to incorporate and validate new soft information features through LR. LR is the most commonly used model in papers that predict default through classification models (see Annex for more details). Another important paper is [Ge et al. \(2017\)](#), which analyses the effect of social media information on default prediction. In addition, [Chen et al. \(2019a\)](#) makes a proposal with logistic quantile regression (LQR) for the same objective variable and with explanatory elements. [Amalia et al. \(2019\)](#) uses the structural equation model (SEM) to causally understand the aspects that generate confidence in the P2P market. [Li et al. \(2018b\)](#) and [Misheva et al. \(2018\)](#) validate and incorporate the determinants of credit risk associated with the P2P market agents’ connectivity network topology into regression

models. In line with them, [Calabrese et al. \(2019\)](#) propose a flexible bivariate regression model advisable for class imbalance events and recognizing the dependence on borrowers' behavior on platforms and credit bureau background (see the Annex [2.5](#) for details).

The category of "Feature selection" has nine papers that emphasize this modeling aspect (see Table [2.8](#)). These studies use methodologies to select the best set of features, seeking to improve the classification performance. Nevertheless, they typically ignore the explanation in the causal relationship with the dependent variable. This aspect of machine learning is a primary subject of research today.

In the category of "Inclusion of new variables," we can find works incorporating new sources and types of information that prove to be decisive in risk and profit management models in the P2P market. The main papers are written by [Herzenstein et al. \(2011\)](#), [Wang et al. \(2016\)](#), [Jiang et al. \(2018\)](#), and [Yao et al. \(2019\)](#), who use text mining algorithms within natural language processing (NLP) technologies. They include narrative features derived from aspects such as economic hardship, and qualities such as hardworking, successful, moral, and religious. Additionally, they add variables that are derived from aspects related to deception, subjectivity, sentiment, readability, personality, etc. or soft features related to assets, income, work, family, agriculture, length, capital turnover, investment and entrepreneurship, business expansion, and the ambiguity of purpose, among others (see the Annex [2.5](#)). These variables are mainly taken from loan titles, textual data generated from statements describing the purpose of the loan, and all descriptive text paragraphs related to each application. [Wang et al. \(2020\)](#) proposes a soft factor mining method in terms of the distribution of the kinds of semantics expressed in the descriptive loan text. They also evaluate the inclusion of these features by performance compared to linguistic and stylistic soft factors. In turn, [Xia et al. \(2019\)](#) complements demographic and financial information with narrative data and includes soft information related to loan description, borrowers' character, and variables obtained from a clustering procedure on soft information. [Li et al. \(2019\)](#), with advanced natural language processing (NLP) techniques, evaluated the risk

of fraud associated with platforms in China. They use variables derived from management team members’ working experience, educational background, and composition. Text mining techniques are used on new information sources, and the explanation and prediction components in risk models are improved, benefiting management in the P2P market.

We can also find works that deal with users’ internet and telecommunication information. [Wang et al. \(2019\)](#) uses information on borrowers’ online operation behavior on P2P lending websites, use variables such as registration records, login records, click records, browse records, authentication records, etc. [Guo et al. \(2016\)](#) focuses on the evaluation of capacity, character, and conditions within the five Cs of credit risk management, excluding capital and collateral. They also include information obtained from the internet, heterogeneous social data, mainly demographic information derived from social networks, content generation, and the structure of each user’s social network through the demographic, tweet, and network feature components. [Ge et al. \(2017\)](#) and [Zhang et al. \(2016a\)](#) evaluate the role of social media information in the prediction of default.

Similarly, [Niu et al. \(2019\)](#) and [Xu et al. \(2016\)](#) study the role of social network information. This last paper also evaluates information derived from herding manipulation to predict fraud. They find that herding behavior increases the likelihood that lenders will invest in listings that have already received bids from others. In addition, [Ma et al. \(2018a\)](#) include patterns derived from mobile phone use (phone calls, text messages, and data traffic), demographics, mobility patterns, telecommunication patterns, app usage patterns, and telecommunication records to predict loan default.

Another trend of the “inclusion of new variables” consists of the use of network topology information, a field that has gained attention in recent years. [Li et al. \(2018b\)](#), [Misheva et al. \(2018\)](#), and [Giudici et al. \(2020\)](#) include variables derived from the market connectivity network topology for systematic risk assessment in the first paper and credit risk in the latter two. Similarly, [Ahelegbey et al. \(2019\)](#), with inference derived from a latent factor model on financial indicators, distinguishes between connected and unconnected business

communities. This distinction improves the predictive performance of scoring models in P2P lending for SMEs.

Other works study the impact of particular features of P2P lending platforms. [Zhu \(2018\)](#) studies how a platform’s promise of security as guaranteed support to back a loan presents a level of negative relevance in the models, bringing problems of moral risk and adverse selection and affecting the probability of loan financing and default. They find that loans with security promises have a higher quoted amount, a lower interest rate, higher ratings, more investors, a higher funding success rate, but a higher default rate than unsecured loans. [Xu & Chau \(2018\)](#) examines the impact of communication between lenders and borrowers on financing outcomes and loan performance, collecting variables related to information disclosure, social influence, and the quantity and quality of information exchanged, among other characteristics (see the Annex 2.5). Finally, in this category, [Fu et al. \(2019\)](#) characterize investors’ comments’ sentiments. From a time series of changes describing the dynamics of this variable, they predict an index of the daily trading volume of Chinese platforms. They take comments from the first authorized P2P information platform in China and one of the largest portals in the P2P industry.

On the other hand, [Gao et al. \(2017\)](#) empirically validate the use of forward-looking and backward-looking mechanisms. The first mechanism with credit indicators and the second with variables describe borrowers’ experiences on the platform. These variables (Annex 2.5) are complemented with information on loans and titles (e.g., length). On the other hand, [Wei et al. \(2018\)](#), trying to capture the nonlinearity of characteristics, uses a cubic spline regression transformation, combining their proposal with a classification optimization procedure through buffered AUC (bAUC). However, we consider that these performance-enhancing transformations must also be evaluated in light of the explainability of the model. We include this article in the “Inclusion of new variables” group due to the transformation proposed to test the variables differently from the traditional way.

Other modeling aspects aspects

This section highlights some elements associated with classification models, mainly for default and fraud estimation in P2P lending risk management. In particular, elements that provide confidence to users and, especially, to regulators and supervisors, who must understand them to regulate and supervise effectively ([ROFIEG, Expert Group on Regulatory Obstacles to Financial Innovation, 2019](#)). In particular, we highlight “Cross-validation,” “Hyperparameter tuning,” “Use of logistic regression,” “Statistical comparison of the performance of classifiers,” and “Explainability.” All these elements reveal how the models were estimated, how well they perform, and how they work.

In this line, Table [2.9](#) presents these elements. We analyzed the papers that proposed classification models, 74 out of 98 papers (75 % of the total). It is worth mentioning that 49 of them included machine learning methods.

Tabla 2.9: Other modeling aspects

Component	Number	Authors/Year
Use of logistic regression	49	Ahelegbey et al. (2019); Boiko Ferreira et al. (2017); Byanjankar et al. (2015); Byanjankar (2017); Cai and Zhang (2020); Calabrese et al. (2019); Canfield (2018); Chen (2017); Chen et al. (2019); Ding et al. (2017); Duan (2019); Jiang et al. (2019); Jiang et al. (2018); Ge et al. (2017); Giudici et al. (2020); Gourieroux and Lu (2019); Guo et al. (2016); Li, Ding, et al. (2018); Li et al. (2020); Lin et al. (2017); Liu et al. (2018); Ma, Zhao, et al. (2018); Malekipirbazari and Aksakalli (2015); Hadji-Misheva et al. (2018); Namvar et al. (2018); Nguyen Truong et al. (2019); Niu et al. (2019); Rodrigues et al. (2018); Wang et al. (2016); Serrano-Cinca and Gutierrez-Nieto (2016); Serrano-Cinca et al. (2015); Stofa (2017); Tan et al. (2017); Tao et al. (2017); Van-Sang et al. (2019); Wang, Kou, et al. (2018); Wei et al. (2018); Xinmin et al. (2019); Xu and Chau (2018); Xia et al. (2017); Xia et al. (2018); Xia et al. (2019); Xia (2019); Yao et al. (2019); Ye et al. (2018); Wang, Jiang, et al. (2018); Wang et al. (2020); Zhang, Jia, et al. (2016); Zhu et al. (2019)
Explainability	38	Ahelegbey et al. (2019); Byanjankar et al. (2015); Byanjankar (2017); Calabrese et al. (2019); Canfield (2018); Chen (2017); Chen et al. (2019); Durovic (2017); Emekter et al. (2015); Gao et al. (2017); Ge et al. (2017); Giudici et al. (2020); Gourieroux and Lu (2019); Guo et al. (2016); Herzenstein et al. (2011); Jin and Zhu (2015); Jiang et al. (2018); Li, Ding, et al. (2018); Lin et al. (2017); Li et al. (2020); Liu et al. (2018); Ma, Sha, et al. (2018); Ma, Zhao, et al. (2018); Hadji-Misheva et al. (2018); Niu et al. (2019); Serrano-Cinca and Gutierrez-Nieto (2016); Serrano-Cinca et al. (2015); Stofa (2017); Tao et al. (2017); Uddin et al. (2018); Wan et al. (2019); Xinmin et al. (2019); Xu and Chau (2018); Xia et al. (2017); Yang et al. (2019); Yao et al. (2019); Zhu (2018); Zhang, Jia, et al. (2016)
Cross-validation	36	Bastani et al. (2019); Boiko Ferreira et al. (2017); Byanjankar (2017); Cai and Zhang (2020); Cho et al. (2019); Guo et al. (2016); Ding et al. (2017); Jiang et al. (2019); Jiang et al. (2018); Jin and Zhu (2015); Kim and Cho (2019a); Kim and Cho (2019b); Kumar et al. (2016); Li, Ding et al., (2018); Li et al. (2019); Li et al. (2020); Malekipirbazari and Aksakalli (2015); Namvar et al. (2018); Nguyen Truong et al. (2019); Niu et al. (2019); Rodrigues et al. (2018); Van-Sang et al. (2019); Wang et al. (2016); Wang, Jiang et al., (2018); Wang et al. (2020); Wei et al. (2018); Xia et al. (2019); Xia (2019); Xu et al. (2016); Yang et al. (2019); Ye et al. (2018); Zhang, Jia, et al. (2016); Zhang, Wang, Chen, Shang, et al. (2017); Zhang, Wang, Chen, Zhao, et al. (2017); Zhou et al. (2019); Zhu (2018)
Hyperparameter tuning	20	Boiko Ferreira et al. (2017); Cho et al. (2019); Jiang et al. (2019); Kim and Cho (2019a); Kim and Cho (2019b); Li, Ding, et al. (2018); Li et al. (2020); Ma, Zhao, et al. (2018); Malekipirbazari and Aksakalli (2015); Niu et al. (2019); Rodrigues et al. (2018); Wang et al. (2019); Wang et al. (2020); Xia et al. (2017); Xia et al. (2018); Xia et al. (2019); Xia (2019); Yang et al. (2019); Ye et al. (2018); Zhou et al. (2019)
Statistical comparison of the performance of classifiers	10	Cho et al. (2019); Jiang et al. (2019); Kim and Cho (2019a); Ma, Sha, et al., (2018); Niu et al. (2019); Wang, Jiang, et al. (2018); Wang et al. (2020); Xia et al. (2017); Xia et al. (2018); Xia (2019)

Source: Compiled by the authors

Several studies use mechanisms to evaluate the efficiency of the proposals in their evaluation strategy (55 papers), with “cross-validation” being the most commonly used (36 papers). Other studies use other alternatives, including the mere separation of training, validation, and test sets (or only training and testing) or the replication and evaluation of these segmentations randomly, a given number of times.

The use of “hyperparameter tuning” is also analyzed. This process shows the purpose of estimating optimal, or at least better, models. Machine learning methods have different parameters that control the learning process, and it is essential to ensure their performance and stability. However, only 20 papers mention “hyperparameter tuning” processes. Among the most commonly used methods are searching on grids, Bayesian optimization methods,

and genetic algorithms.

We considered it important to highlight the studies that “use logistic regression,” either directly or as a benchmark. Remarkably, 49 papers “Use logistic regression” as the main model or as the benchmark methodology in the P2P lending market. Thus, logistic regression seems to still be the reference model for P2P lending loan evaluation proposals, as it is for regulators, banks, and financial institutions in traditional products.

In turn, in the category “Statistical comparison of the performance of classifiers,” we highlight the works that perform statistical inference as a decision criterion against the modeling alternatives’ performance. Only ten papers use a statistical significance test to compare the performance of the competing alternatives. The tests used are paired-t, Friedman, DeLong, or Shapiro-Wilk.

Finally, in the “Explainability” category, we consider studies that include elements that provide explainability to their proposals. Such elements could be, for example, dependence analysis, sensitivity or statistical causality, analysis of coefficients or parameters, inference of parameters, analysis of cases, interpretation of signs, evaluation of monotonicity, nonlinearities, changes of structure, and feature importance analysis, among other aspects. In particular, we counted 38 papers that included the “Explainability” component. However, most of them (26 papers) do so for statistical or econometric models. More precisely, they do so through coefficient estimations and inferential tests. Few papers include explanatory elements in machine learning (14 papers), and such papers mostly (8 papers) only study feature importance, ignoring aspects such as monotonicity, the nonlinearity of relationships, and structural changes, which could be considered through approaches such as those developed in recent years with techniques of local and global explicability and interpretability (Molnar, 2021). As a result, most machine learning models lack a critical aspect demanded for managing risk in the P2P lending market, which would help to extend their adoption.

2.4. Discussion

The P2P market has been growing hand in hand with further technological developments, and risk and profit management have become more sophisticated. Although this has brought opportunities for investment and financing for lenders and borrowers, it has also generated difficulties in regulation and supervision. Governing bodies and p2p platforms must adapt quickly to new market demands for risk and profit management to ensure participants' financial health and the market's sustainability. For this purpose, they must cover financial elements and technological and methodological elements, where machine learning and data science tools are strategic if used efficiently.

Current technologies and, in particular, machine learning, offer the potential for broader, deeper, and faster analysis of large data sets, incorporating diverse sources of information that may be relevant for risk assessment and profitability management of companies and individuals, much more so in a developing market such as P2P lending ([ROFIEG, Expert Group on Regulatory Obstacles to Financial Innovation, 2019](#)). The development of machine learning and new data science technologies helps provide more suitable and adapted products. For lenders, these solutions can help protect against credit risk and fraud, reduce the cost of the credit evaluation process, encourage product development, distribution, and monitoring, and enable better consumer-customer interaction ([ROFIEG, Expert Group on Regulatory Obstacles to Financial Innovation, 2019](#)). Therefore, given the growth of P2P lending and the increasing adoption of new technologies such as machine learning for efficient risk management and profit assessment, we considered it relevant to review the academic literature. It helped us identify the main research trends and research and development opportunities that contribute to the P2P lending market's healthy development. We also found some limitations in many of the studies that we consider relevant to emphasize to be avoided in future works.

Related to RQ1, we found that among the most relevant works by citation are [Emekter et al. \(2015\)](#); [Greiner & Wang \(2010\)](#); [Herzenstein et al. \(2011\)](#); [Serrano-Cinca & Gutierrez-](#)

Nieto (2016); Serrano-Cinca et al. (2015), and Xia et al. (2017a), with Emekter et al. (2015) being the most cited publication. These papers are mostly devoted to identifying the key factors in credit risk models in P2P lending. They mainly use statistical and econometric models. Additionally, the papers by Serrano-Cinca & Gutierrez-Nieto (2016) and Xia et al. (2017a) have had a great impact. They present proposals for risk management models with a business focus, which we believe make them very suitable for the industry.

Additionally, we identify that China and the United States have more experience in the P2P market. However, the eastern market shows deterioration and reduction of the market due to fraud and regulatory issues. Most likely, the most significant number of publications come from eastern countries and China in particular. The most frequent cooperation between researchers is between the two countries mentioned. Consequently, we can conclude that the countries with the largest market also have researched the topic more intensely; however, we cannot establish a causal relationship.

In RQ2, we asked about the main trends and contents in the literature. The most used data set is that from Lending Club, a company from the US. This is probably because it provides comprehensive information, including variables for both the application and the behavior of the obligations. We can see a strong tendency to use data from the Chinese market that may be due to market growth and its subsequent problems, which make it suitable for research. Hence, by country, data mostly come from China and the United States, but we found other datasets from Mexico, Southeast Asia, and the United Kingdom, where P2P lending is also developing.

The software is mainly free, open-source Python or R libraries. Its use is frequently associated with free access and the possibility of enabling replicability and the reutilization of previous developments of other researchers' developments with few restrictions, if any. We have identified that many publications focus on credit risk classification models using new machine learning methods at the business level. These models typically improve the performance of traditional alternatives, which eventually favor investors, platforms, and con-

sumers. In recent papers, we can find new trends in the use of machine learning, such as deep learning, decision tree ensembles (mainly random forest and gradient boosting algorithms), and heterogeneous ensembles. However, despite substantial methodological progress, few efforts have been made to add interpretability and explainability to the resulting models, which may hinder their adoption due to regulatory requirements and industry needs.

On the other hand, a significant part of the publications use statistical and econometric models, probably due to their interpretability, integration with business theory and knowledge, ease of estimation and implementation, and generalized understanding of researchers, users, and user regulators. It is worth mentioning that logistic regression is typically used as a benchmark or base model.

We also found another trend: complementing traditional information with variables from other sources or using soft information and NLP for credit risk and investment models and profits. For example, many papers explore the use of new input variables derived from new information sources such as social networks (connections, consultations, photographs, messages, and videos), connectivity networks, the use of telecommunication devices, georeferencing, the textual analysis of the documents that accompany the applications, etc. Such variables help to improve the performance indicators.

2.5. Research gaps, recommendations, and future directions

In this section, we answer RQ3, where we asked about research gaps and future research directions. As we have seen, new proposals based on machine learning and artificial intelligence have become more relevant. Some papers include aspects that others do not, but few cover most of the elements that we consider to ensure their effectiveness, efficiency, and transparency.

First, we reflect on the sources of information, which are increasingly diverse and require complex handling. The potential predictors of risk and profit are more varied and complex

in their treatment, including soft data, variables from NLP, connectivity networks, other sources of big data, etc. Such new data typically improves the levels of predictability and the performance of the models. However, in many cases in academic research, models exploit these sources without a conscious study on their use in the market context. As researchers, we should strive to generate proposals without discrimination and implicit biases associated with selecting and treating new sources of information. Possible biases should be reflected upon and investigated in light of the ethics, culture, and laws of the financial context in which the p2p market develops.

With respect to the population under study related to the treated business problem, it is interesting to note that very few studies focus on P2P loans for businesses and companies, most of which have focused on loans for individuals. Having businesses and companies as the object of study changes the treatment of the sources of information, and the evaluation criteria should be more financial. This line of research deserves a more in-depth analysis.

Additionally, we found that some other business problems are understudied, such as fraud, debt collection, credit provision (LGD), third-party collateral warranties, etc. All of them are necessary for the healthy development of the market.

Surprisingly, we found that many papers lack a proper specification of the business problem. For example, few papers clarify whether the proposed model is for granting or behavior. This lack prevents the use of these models by regulators and end-users. We have also found problems in the default definition; in some papers, there is even no mention of how this event is defined, which could cause an underestimation of the risk or an overestimation of the profit.

We believe that it is essential to clearly define the model's objective in the business context since this establishes the evaluation criteria and meets the demands of interpretability and transparency. It is different from having a granting model, where the main objective is the classification, to a default probability model where the interest is in provisioning the expected losses associated with the investment. Alternatively, if the aim is to construct sco-

ring, this will define the access and pricing rate assigned to a liability. We believe that this distinction is not made in detail in many cases and would certainly allow researchers to provide solutions more in line with market demand.

In turn, papers should have an adequate selection of the independent variables according to the model's purpose. This aspect has been overlooked in some research proposals where the inclusion of variables was guided merely by improving the prediction performance. For example, for granting loans, it makes no sense to use decision variables such as interest rate or the level of risk estimated by the companies because they are closely linked to the event being analyzed, generating biases in the results and limiting their use. The variables included must consider the operability of the business. These facts limit the application due to a lack of uniformity and consistency, which are objectives sought by regulatory bodies, especially for transnational growing markets such as P2P lending. The appropriate choice of variables for the models is a requirement that must be complied by research proposals.

Regarding the software used, we believe that we must continue to take advantage of the benefits of open-access software, such as Python and R. They have proven effective in effectively integrating different information sources, but more importantly, they enable access to other researchers' developments, with few limitations, contributing to the democratization of knowledge in general and the healthy development of the P2P lending market.

Regarding the methodological aspects, we will highlight those present in some of the works reviewed and offer the potential for the market's healthy development, for example, the proposal of cost-sensitive credit risk models that help mitigate the class imbalance problem and provide insightful error measures for the business. From our perspective, cost-sensitive models for P2P lending are a research line that deserves to be further developed.

We also call attention to the correction of some biases that researchers have considered in a few studies. One of them is the method of rejecting inference as an alternative to the bias that arises from not knowing the rejected applications. Many works are built on data already filtered by the data providers, affecting the models' actual performance and limiting

their use.

Reject inference is a problem of estimating a counterfactual that should be studied. Another bias, perhaps to a lesser extent, arises in models that aim to estimate the probability of the event's probability, which deserves evaluation criteria different from those used in classification. They deserve calibration methods of the estimated probability, which academics have rarely studied in light of the business contexts. On the other hand, we would also like to draw attention to the use of variable transformations in the models, again with a business sense that guarantees explainability and transparency, such as categorical variables. In this case, transformations using dummy variables, the weight of evidence (WOE), standardizations, and normalizations are used indistinctly in the works, improving performance but limiting their interpretation and application. We believe that variable transformation in future proposals needs more careful consideration to ensure the proposals' usefulness.

We would also like to mention that the evaluation process and strategy should not only seek to guarantee performance in terms of classification or prediction. Research proposals should also incorporate other aspects already considered by regulators to evaluate models for traditional product markets: representativeness, replicability, scalability, and temporal stability. We believe that there is no standard minimum criteria evaluation that could guarantee the suitability and application of the proposals that recognize the business context, the type of problem, the regulatory framework, and the sources of information. The evaluation process of risk and profit modeling proposals is one of the topics in the P2P lending market that merits further research.

The explainability component also deserves further attention since understanding the models' decisions should be implicit in the modeling processes. Given the extended use of machine learning methods and the great need for model explainability and transparency required by regulators, supervisors, financial institutions, and governments, we believe that more research on this topic is needed. Most machine learning proposals focus on performance and neglect interpretability. The models with explainable purposes are still supported in

traditional methodologies, econometrics, and statistical tools. The papers that use machine learning models rarely include this aspect, and if they do, they only include an analysis of feature importance. Thus, the interpretability components are still a challenge and a necessity for machine learning proposals. As future research lines, there is a need to generate frameworks that ensure the ethical application of new solutions, particularly credit, whose central elements rely on the explainability of decisions ([ROFIEG, Expert Group on Regulatory Obstacles to Financial Innovation, 2019](#)). In P2P lending, we are starting to see some efforts along this research line, as shown in [Ariza-Garzón et al. \(2020\)](#) and [Bussmann et al. \(2020\)](#), both published after our literature review. At a broader level, there is great interest in reflecting on the component of explainability in modeling and offering alternative explanations for new machine learning models, as shown in [Carvalho et al. \(2019\)](#), [Rudin \(2019\)](#), and [Barredo Arrieta et al. \(2020\)](#).

Finally, we highlight another line of research that includes articles that help understand the market, deficiencies in regulation, and economic, political, and financial approaches. Most of these studies are based on qualitative research, with theoretical and contextual elements, seeking to identify the shortcomings and opportunities that help the development, continuity, and stability of the P2P market as an alternative for investment, inclusion, and financing. We believe this is a research line that deserves further attention and needs to be complemented with quantitative research. Regarding the limitations of this study, we point out that although the review of articles in the main journals has been both extensive and intensive, the design established could have caused some studies on the subject in other publications to be overlooked, for example, due to the limitation of the database selected or the query used. Assertions on findings are, therefore, only based on articles retrieved.

In summary, P2P lending is a flourishing market, and its research is also buoyant and active. Nevertheless, it has essential shortcomings and opportunities that must be addressed and studied to ensure its service, development, and sustainability. We hope that this study will serve to draw attention to some of these.

Additionally, we consider it vital that academics, regulators, and fintech work together to develop a framework for the risk management and profitability of financial products in the p2p lending market. We believe that joint work solutions could be presented that are much more tailored to the P2P context, generating a greater probability of use and confidence in these solutions. This synergy would ensure a better understanding of the business's nature and structure, which would lead to better modeling, risk management, and service, enhancing the sustainability and development of the P2P lending market. In this way, the research will help develop the market for the digitalization of finance, financial coverage and inclusion, and the offer of better products and services to benefit individuals, households, and companies.

Annex. Table of variables and specific model of the selected papers.

Year	Dependent variables	Independent variables	Specific Model	Authors/Year
2010	(1) % funded; (2) % reduction interest rate	financial, loan, soft (social capital, listing quality: description-image)	MLR	Greiner and Wang (2010)
2011	(1) % funded; (2) % reduction interest rate; (3) performance (4 categories: paid ahead of schedule and in full, paid as scheduled, payments between one and four months late, Default)	demographic, loan, narrative variables (trustworthy, economic hardship, hardworking, successful, moral, and religious)	MR, MLR	Herzenstein et al. (2011)
2015	Default	demographic, loan, financial, employment, credit	ANN, LR	Byanjankar et al. (2015)
			S-CR, LR	Serrano-Cinca et al. (2015)
	Default; fully paid	loan, financial, credit demographic, loan, financial, employment, credit	S-CR, LR RF, KNN, SVM, LR, FICO, Lcgrade	Emekter et al. (2015) Malekipirbazari and Aksakalli (2015)
	Default: with outstanding repayment records	loan, financial, credit	SVM, RBF, MLP, DT-CART, DT-CHAID	Jin and Zhu (2015)
	Expert assessment of the risk	loan, credit, employment, financial	BP-ANN	Zang et al. (2015)
	Fraud	demographic, loan behavior, credit loan application, e-commerce, text description, social network	BP-ANN DT, SVM, ANN	Zhao (2015) Xu et al. (2015)
2016	(1) % funded; (2) final interest rate; (3) Default: 60 + days past due	demographic, loan, credit, experience lending website, interest rate offered, duration (list-expire or loan), role (borrower or lender)	HMR, LR, S-CR	Chen et al. (2016)
	(1) IRR; (2) Default: Charged off, fully paid	demographic, loan, financial, employment, credit	MLR, LR, DT	Serrano-Cinca and Gutierrez-Nieto (2016)
	# investors	financial, credit, social capital, risk management (mortgage collateral or third-party guarantee), listing information, operation duration	ELM-MLR	Yan et al. (2016)
	Default	heterogeneous social data: demographic, tweet, network, high-level (features derived from different classifiers)	GBDT, RF, BAG, NB, L1-LR, SVM; with LDA	Guo et al. (2016)
		loan, credit, soft information (deception, subjectivity, sentiment, readability, personality and mode of thought)	LR, DT-C4.5, NB, MLP, RF	Wang et al. (2016)
	Default; fully paid	demographic, loan, credit historic, social media	DT, LR, ANN	Zhang et al. (2016a)
	Default: 16 + days past due, charged off; Current, Fully Paid, In grace period	loan	RF, DT, BAG	Kumar et al. (2016)
	Default: overdue	demographic, credit, loan, platform variables (successful loan number, failed loan number, membership score, prestige, forum currency, contribution, group)	OC	Li et al. (2016)
	Fraud	loan, learning, past performance, social networking, herding manipulation	RF, SVM	Xu et al. (2016)
	repayment rate	loan application, loan, demographic, platform	MLR	Xia and Li (2016)
	top-N recommendation	Invest or not, proportion express bid amount user invest in a project, categorical proportion amount variable	UCF, ICF, LDA	Zhang et al. (2016b)
2017	(1) Funded; (2) interest rate; (3) Default	demographic, loan, credit, financial, third-party credit guarantee, offline authentication	LR, TOBIT, PROBIT	Tao et al. (2017)
	# investors	financial, credit, social capital, risk management (mortgage collateral or third-party guarantee), listing information, operation duration	ELM-MLR	Yan et al. (2016)
	Default	demographic, financial, loan, credit NI	LR ECSC-L2-LR, ECSC-DT, ECSC-SVM, ECVWC-L2-LR, ECVWC-DT, ECVWC-SVM	Chen (2017) Ding et al. (2017)
		demographic, loan, financial, employment, credit	cs-XGBoost, LR, RF, cs-LR-T, cs-RF-T, cs-LR-sm, cs-RF-sm	Xia et al. (2017)
	Default: 120 + days past due	demographic, loan, campaign, image, social media	LR, PSM, IV, Dif-Dif	Ge et al. (2017)
	Default: 120 + days past due-charged off	credit, loan, financial, verification, number of derogatory public records	LR	Stofa (2017)
	Default: 150 + days past due-charged off	NI	cs-GNB, cs-DT, cs-LR, ru-GNB, sm-GNB, sm-DT, ru-DT, ru-LR, sm-LR, AdaBoost, BAG, RF	Boiko Ferreira et al. (2017)
	Default: 30 + days past due; Bad debt rate	Forward-looking credit indicators (demographic, financial, employment), Backward-looking credit indicators (Successful borrowing times, borrowing amount), loan, borrowing description and titles (lengthened)	PROBIT, TOBIT. Three-stage dynamic game to analyze the default behavior of borrowers	Gao et al. (2017)
	Default: three consecutive payments past due	demographic, loan, verification, employment, financial	S-DA(KM), S-CR, S-ANN, S-LR	Byanjankar (2017)
	Default: charged off	loan	S-DA(KM)	Durovic (2017)
	Default: overdue	demographic, employment, financial, credit	FNT, IBPNN, RBF, ANN LSTM, IBPNN, RBF, ANN	Zhang et al. (2017c) Zhang et al. (2017b)
	Degree by node	demographic, loan, financial, employment, credit profit (lenders and borrowers), collateral	LR BBN	Lin et al. (2017) Lee et al. (2017)

(continued on next page)

Annex. Table (continued). Variables and specific model of the selected papers.

Year	Dependent variables	Independent variables	Specific Model	Authors/Year
	Rank by risk or profitability top-N recommendation	financial, management, operation, guarantee Default, time value of money, profits	AHP-Index TCVM (profit: term, r, default(RF, GBDT, KNN, NB LR, DT); TVM)	Xu and Zhang (2017) Tan et al. (2017)
		Invest or not, proportion express bid amount user's invest in a project, categorical proportion amount variable	UCF, ICF, LDA	Zhang et al. (2017a)
2018	(1) Funded; (2) Default	demographic, loan, credit historic, online seller, student, safety promise, number of bidders	L1-LR	Zhu (2018)
	(1) Funded; (2) interest rate; (3) % Funded; (4) # Bids; (5) Default	demographic, loan, Soft Information (communication, interaction, Lender Comment- Social Influence, Borrower Response-Info Quality)	MLR, LR, S-CR	Xu and Chau (2018)
	5 credit grades	demographic, loan, financial	BP-ANN	Yuan et al. (2018)
	7 credit grades	loan, financial, credit	csDT, csNB, csLR, csSVM with 3 cost matrices	Wang et al. (2018a)
	Approved; declined	financial, loan, credit historic, employment	LR	Wei et al. (2018)
	Default	demographic, loan, credit, network behavior, third- party, social network	E(XGBoost, DNN, LR), XGBoost, DNN, LR	Li et al. (2018a)
		demographic, financial, loan	LR, S-DA	Canfield (2018)
		demographic, loan, credit, financial	LR, PROBIT	Liu et al. (2018)
		demographic, loan, financial, credit	CPL-ELightGBM, CPL-ELF, CPL-ELR, CPL-SVM, E-SVM, E-RF, E-LightGBM, S3VM, Ex, Ag, LR, SVM, RF	Xia et al. (2018)
		demographic, loan, reputational risk, Macro (sectorial risk, country risk)	QM-EBRR	Uddin et al. (2018)
		demographics, mobility patterns, phone usage data: (phone calls, text messages, and data traffic), telecommunication patterns, App usage patterns and telecommunication records	AdaBoost, RF, LR	Ma et al. (2018a)
		financial ratios + degree and closeness centrality measures	CN-LR, CN-DT, LR, DT	Hadji-Misheva et al. (2018)
	Default; fully paid	loan, financial, loan behavior	LightGBM, XGBoost	Ma et al. (2018b)
		demographic, loan, financial, soft features (asset, income, work, family, agriculture, and length)	RF, LR, NB, SVM; with LDA	Jiang et al. (2018)
	Default: 120 + days past due-charged off	demographic, loan, financial, credit	RFoGAPS, RF, SVM, DT, KNN, LN, Actual profit, LR	Ye et al. (2018)
	Default: 150 + days past due-charged off	demographic, loan, financial	E-FIC(GBDT, AdaBoost, LR), E-MV (GBDT, AdaBoost, LR), E-OWAo (GBDT, AdaBoost, LR), E-OWAp (GBDT, AdaBoost, LR), GBDT, AdaBoost, LR	Namvar et al. (2018)
	Default: 90 + days past due	demographic, loan, financial, credit	S-EMRF, MCM, S-CR, LR	Wang et al. (2018b)
	Default: charged off, fully paid	loan, financial	GBDT, SVM, LR, KNN, NB, RF	Rodrigues et al. (2018)
	LGD	loan, credit, financial	MLR	Zhou et al. (2018)
	network systemic risk: DVtf	network variables	CN, MLR	Li et al. (2018b)
2019	Grade $\geq$ A: 1, Grade $<$ A: -1	demographic, employment, financial, loan, credit, institutional guarantee	GA-SVM	Yang et al. (2019)
	(1,3) Default; (2) Rejected	demographic, loan, employment, credit	LD, LR, SVM, ANN, KNN, RF with RBM-FS	Nguyen Truong et al. (2019)
	(1) % negative sentiment comments-(2) trading volume index-(3) Weekday	% negative sentiment comments, TVI (trading volume index-ND), Weekday	Text-CNN-LSTM, Text-CNN-VAR, Text-CNN-DNN, Text-CNN-MLR, Text- CNN-RF, Text-CNN-SVM	Fu et al. (2019)
	(1) default-charged off-(2) 1 + public record bankruptcies	Thau-kendall (copula), loan, financial, credit, spatial variables	BivGEV, BivProbit, LR	Calabrese et al. (2019)
	(1) Default; (2) IRR	demographic, loan, financial, credit	DP, WL, WDP, RF, GB, SVM with IHT, BC, ru, so, sm	Bastani et al. (2019)
	(1) Default; (2) Rejected	demographic, loan, employment, credit	LD, LR, SVM, ANN, KNN, RF with RBM-FS	Van-Sang et al. (2019)
	(1) Default; (2) Successful	demographic, loan, verification and proof variables, successful borrowing times,	LR, Probit	Xinmin et al. (2019)
	(1) Potential defaulted; (2) Default	loan, credit, financial, employment, region information	LR, RF, SVM; S3VM, OD-LightGBM	Xia (2019)
	Default	demographic, financial, soft (loan description, character, clustering result based soft information)	CatBoost, LR, DT, BNN, RF, GBDT, XGBoost	Xia et al. (2019)
		demographic, registration, loan, social network financial ratios	LightGBM, RF, AdaBoost and LR CN(LF)-LR	Niu et al. (2019) Ahelegbey et al. (2019)
		NI	E(GBDT, XGBoost, LightGBM), NN, LR, RF, SVM, KNN, AdaBoost	Zhou et al. (2019)
		soft information (online operation behavior: data of borrower's online operation behavior on P2P lending website), credit	AM-LSTM, BOA-XGBoost, LSTM, BLSTM, BLSTM-Meanpool.	Wang et al. (2019)
	Default: 16 + days past due, charged off, In Grace period; Current, Fully Paid, Issued	loan, verification, application type	RF, DT, SVM, LR with sm	Zhu et al. (2019)
	Default: 90 + days past due	demographic, loan, employment, financial, credit	Latency and Incidence with MCM: RF- TDH, RF-Cox, LR-Cox, LR-TDH and	Jiang et al. (2019)

(continued on next page)

Annex. Table (continued). Variables and specific model of the selected papers.

Year	Dependent variables	Independent variables	Specific Model	Authors/Year
2020	Default: charged off, fully paid	demographic, loan, credit demographic, loan, employment, financial, credit	RF, LR, DT, B-LR, B-DT, BAG-LR, BAG-DT E(LP, TSVM), LP, TSVM, DT IEFSVM, cs-AdaBoost, cs-RF, EE, ru-B, w-ELM, cs-XGBoost, EFSVM	Kim and Cho (2019a), Cho et al. (2019)
	Default: 120 + days past due, 120- days past due, 0 safe loans	demographic, loan, financial	DNN, LR, LD, DT, SVM, RBF, MLP, AdaBoost	Duan (2019)
	Default: charged off, fully paid	demographic, loan, employment, financial, credit	CNN: (CNN, Inception, ResNet, DenseNet, Inception-ResNet); MLP, SVM, KNN, DT, RF	Kim and Cho (2019a),
			DP-CNN, CNN: (CNN, Inception, ResNet, DenseNet, Inception-ResNet); MLP, SVM, KNN, DT, RF	Kim and Cho (2019b),
	Default: with outstanding repayment records	demographic, loan, platform authentication, regulation change of the government	LQR	Chen et al. (2019)
	Hazard exposed (absconded with ill-gotten gains, difficult withdrawing, out of business and investigated by Economic Crime Investigation Police), Normal.	Text information: Management team members' working experience, educational background, and composition	MUN-LETCLA, Doc2vec, LDA, Dependency, Syntactic, Keyword	Li et al. (2019)
	fully repay the loan in advance	demographic, loan, verification, financial, employment, credit	S-CR	Wan et al. (2019)
	Funded	demographic, loan, credit, financial, employment, platform audition, soft information (capital turnover, investment and entrepreneurship, expanding business, ambiguous purpose, house decoration, household expenses, Daily consumption, Car purchase)	LDA(VSM)-LR	Yao et al. (2019)
	Number of loans by grade. top-N recommendation	return-risk trade-off (probability of appearance of loans in the future and maximum number of days the investor is willing to wait to invest their funds)	MO	Ren and Malik (2019)
	Total loss ratio (LGD, balance at default)	demographic, loan, employment, financial	SPM-LIR	Gourieroux and Lu (2019)
	Lending Intention, Platform Trust	Protection Policies, Platform Trust	SEM	Amalia et al. (2019)
	Survey: demographic, investment background, investment behavior, investment criteria, motivation to lend variables		DA, PCA	Pierrakis (2019)
	Semi structured interviews (nonprobability sampling): Collateral, requirement, procedure, online process, fast process, interest rate, credit scoring, profits, costs variables		QM, DA	Rosavina et al. (2019)
	Default	demographic, loan, financial, employment, credit	DT-J48, LR	Cai and Zhang (2020)
		demographic, network behavior, third party information, social network, loan transaction time, loan financial ratios, degree(number of neighbors of a node), strength(average distance of a node for its neighbors) and PageRank (importance of a node in a network by assigning relative scores to all nodes in the network)	E(XGBoost, DNN, LR), XGBoost, DNN, LR, AdaBoost, GBDT, RF, DT, KNN, SVM LR	Li et al. (2020) Giudici et al. (2020)
	Default: 120 + days past due	demographic, loan, credit, soft information (semantic vs. linguistic and stylistic factors)	LR, L1-LR, RF, XGBoost with Semantic	Wang et al. (2020)
	Rank by risk or profitability	demographic, loan, employment, financial, credit	Soft Factor Mining Method, LDA DHPF-TODIM	Ji et al. (2020)

Source: Compiled by the authors

Legend: AdaBoost: Adaptive Boosting, Ag: Augmentation, AHP: Analytic Hierarchy Process, AM: Attention Mechanism, ANN: Artificial Neural Network, Ap: Actual profit, B: Boosting, BAG: Bagging method, BBN: Bianconi-Barabasi network model, BC: Balance Cascade, BivGEV: Bivariate Generalized Extreme Value regression, BivProbit: Bivariate Probit regression, BLSTM: Bidirectional LSTM network, BNN: Bagging Neural Network, BOA: Bayesian hyperparameter optimization, BP: Back Propagation, C4.5: C4.5 decision tree algorithm, CatBoost: Unbiased boosting with categorical features, CN: Connective Network model, CNN: Convolutional Neural Network, CPLE: Contrastive Pessimistic Likelihood Estimation, CR: Cox

Regression, cs: cost sensitive, DA: Descriptive Analysis, DenseNet: Dense Convolutional Network, DHPF-TODIM: Dual Hesitant Pythagorean Fuzzy number + TODIM approach ((an acronym in Portuguese of interactive and multicriteria decision making), Dif-Dif: Difference-in-Differences, DNN: Deep Neural Network, DP: Deep Learning, DST: Dempster–Shafer theory, DT: Decision Tree, E: Ensemble, EBRR: Expert Based Risk Rating, ECSC: Ensemble Classification based on Supervised Clustering, ECVWC: Ensemble Classification based on Variable Weighting Clustering, EE: Easy Ensemble, EFSVM: Entropy Fuzzy Support Vector Machine, ELM: Elaboration Likelihood Model, EMRF: survival Ensemble Mixture RF, Ex: Extrapolation, FIC: Fuzzy Integral Combination, FICO: Fair Isaac Corporation Score, FNT: Flexible Neural Tree, FS: Feature selection, GA: Genetic Algorithm, GB: Gradient Boosting regression, GBDT: Gradient Boosting Decision Tree, GNB: Gaussian Naive Bayes, HMR: Hierarchical Multiple Regression, IBPNN: Improved BP Neural Network, ICF: Item based Collaborative Filtering, IEF SVM: Instance-based Entropy Fuzzy Support Vector Machine, IHT: Instance Hardness Threshold, IV: Instrumental Variables, KM: Kaplan-Meier method, KNN: K-Nearest Neighbor, L1: Lasso (L1 norm) Regularization, L2: Ridge (L2 norm) Regularization, LC: Lending Club, LD: Linear Discriminant Analysis, LDA: Latent Dirichlet Allocation, LF: Latent factor model, LightGBM: Light Gradient Boosting Machine, LIR: semiparametric estimator Least Impulse Response, LP: Label Propagation, LQR: Logistic Quantile Regression, LR: Logistic Regression, LSTM: Long Short-Term Memory, MCM: Mixture Cure Model, MLP: Multilayer Perceptron , MLR: Multiple Linear Regression, MO: Markowitz optimization model, MR: Multinomial Regression, MUN-LETCLA: Multiple NLP (Natural Language Processing) Integrated Learning Text Classifier Model, MV: Majority voting, NB: Naive Bayes, NR: Nuclear Regression model, OC: Outliers Cluster-based method, OD: Outlier Detection, OWAo: Optimistic OWA (ordered weighting averaging), OWA_p: Pessimistic OWA (ordered weighting averaging), PCA: Principle Component Analysis, PSM: Propensity Score Matching, QM: Qualitative Model, RBF: Radial Basis Function, RBM: Restricted Boltzmann Machines, ResNet: Residual Neural network, RF: Random Forest, RFoGAPS: Random Forest optimized using a genetic algorithm with profit score, ru: random under sampling, S: Survival, S3VM: semi supervised variants of SVM, SEM: Structural Equation Model, sm: smote (Synthetic Minority Oversampling Technique), so: random oversampling, SPM: Semi-Parametric Transformation model, SVM: Support Vector Machine , T: Threshold, TCVM: Total Capital Value Maximization, TDH: Time-Dependent Hazards, TSVM: Transductive Support Vector Machine, TVM: Time Value of Money, UCF: User based Collaborative Filtering, VAR: Vector Autoregressive Model, VSM: Vector Space Model, WDP: Wide and Deep Learning, WL: Wide Learning, XGBoost: Extreme Gradient Boosting. NI: No Information.

Capítulo 3

Artículo. Explainability of a machine learning granting scoring model in peer-to-peer lending

Este trabajo se ha publicado en la revista indexada en Web of Science con Journal Citation Reports (JCR) IEEEAccess en el volumen número 8 de abril de 2020, con páginas en el intervalo 64873-64890. (<https://doi.org/10.1109/ACCESS.2020.2984412>)

Tiene como coautores a Javier Arroyo, Antonio Caparrini y María-Jesús Segovia.

El trabajo ha sido apoyado por los proyectos europeo H2020 de coordinación en el marco del convenio de subvención No 825215, y por el proyecto de investigación Santander-UCM de la convocatoria de 2019, con referencia PR87/19-22586.

Abstract

Peer-to-peer (P2P) lending demands effective and explainable credit risk models. Typical machine learning algorithms offer high prediction performance, but most of them lack explanatory power. However, this deficiency can be solved with the help of the explainability tools proposed in the last few years, such as the SHAP values. In this work, we assess the well-known logistic regression model and several machine learning algorithms for granting scoring in P2P lending. The comparison reveals that the machine learning alternative is superior in terms of not only classification performance but also explainability. More precisely, the SHAP values reveal that machine learning algorithms can reflect dispersion, nonlinearity and structural breaks in the relationships between each feature and the target variable. Our results demonstrate that it is possible to have machine learning credit scoring models be both accurate and transparent. Such models provide the trust that the industry, regulators and end-users demand in P2P lending and may lead to a wider adoption of machine learning in this and other risk assessment applications where explainability is required.

Index terms— Credit risk, P2P lending, explainability, Shapley values, boosting, logistic regression

3.1. Introduction

Credit risk analysis typically relies on statistical models such as logistic regression, probit regression, discriminant analysis and Cox survival models, among others (Leo et al., 2019) and (Serrano-Cinca & Gutierrez-Nieto, 2016). These methods offer good performance, are easy to understand, and do not pose computational problems (Andriosopoulos et al., 2019). On the other hand, machine learning alternatives frequently offer a better predictive performance because they can identify more complex risk patterns (Leo et al., 2019). Nonetheless, most machine learning methods are typically black boxes with little or no chance of being interpreted.

As a result, banks are proceeding with caution in the adoption of machine learning for credit risk modeling. Furthermore, the requirements of regulatory entities are usually associated with traditional models and do not correspond to the challenges associated with these new alternatives (Babel et al., 2019). The application of machine learning models requires complimentary validation elements, and study and analysis of new biases or interpretability, among other aspects, even if they are not explicitly detailed (Basel Committee on Banking Supervision, 2017).

Interpretability and transparency are essential for the different models of credit administration processes, such as granting, behavior and collection processes, or fraud detection. They are demanded to guarantee a fair, regulated and monitored credit delivery process. Regulatory entities and end users require interpretable models and hence typically rely on models such as logistic regression (Leo et al., 2019).

Technological development has also brought about new credit products such as the peer-to-peer (P2P) lending market (Milne & Parboteeah, 2016). P2P lending works as loans between individuals, borrowers and investors, connected through technological platforms. P2P lending eliminates the intermediation of traditional institutions, and consequently, the information asymmetry is much more marked than in traditional banking. As a result, measuring credit risk in an interpretable manner is even more challenging in P2P lending

than with traditional products.

Intermediation platforms offer scoring models to mitigate risk, protect investors and maintain financial stability (Ahelegbey et al., 2019; Li et al., 2018a; Serrano-Cinca & Gutierrez-Nieto, 2016). Scoring models are also used to rate credit claimants and, on some platforms, establish differential rates by level of risk, associated with differential returns for investors (Claessens et al., 2018).

The scoring model P2P market should offer explanations to administrators, investors, and borrowers, and support the definition of credit policies, among other aspects. Intermediation platforms require risk measurement tools with not only high levels of accuracy but also interpretability, which is a feature also requested by regulatory entities (Board, 2017; Carr & Bailey, 2018).

As different authors emphasize (Florez-Lopez & Ramon-Jeronimo, 2015; Sun et al., 2014), the development of understandable and explainable models is one of the most important research topics in the prediction of financial default, especially in areas where technological developments generate openness to different financial credit products, as is the case with the disintermediated P2P lending market, which is still in development and in the process of regulation.

In this work, we assess two credit scoring approaches for P2P lending in terms of performance and explainability. More precisely, we compare machine learning algorithms with logistic regression, which is a well-established technique in credit risk. For the machine learning alternatives, we apply a decision tree (Breiman et al., 1984), a bagging classification approach such as random forest (Breiman, 2001), and XGBoost (Chen & Guestrin, 2016), which is a gradient boosting classifier. We use genetic algorithms to search for an appropriate hyperparameter combination for each machine learning algorithm, in the same line as other works (Yang et al., 2019).

We use data from Lending Club, which have been extensively used in other works (Emekter et al., 2015; Malekipirbazari & Aksakalli, 2015; Xia et al., 2017a), but with no particular

emphasis on explainability, except for some cases where part of their objectives are explanatory and use naturally interpretable alternatives such as regression models or decision trees (Serrano-Cinca & Gutierrez-Nieto, 2016; Serrano-Cinca et al., 2015). We use the data to develop granting model, and consequently, we only consider the variables in the credit application.

Regarding the performance comparison, we use statistical inference to determine which approach performs better. We also include a detailed analysis of classification metrics for each class (default and nondefault), which is not as frequent as it should be in the literature.

Regarding the explainability comparison, we use the Shapley values (Shapley, 1953), as implemented in Lundberg et al. (2019, 2018a); Lundberg & Lee (2017). These works present a unified framework for interpreting predictions known as SHAP (SHapley Additive exPlanations) based on aggregations of Shapley values with the support of Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016) and other methods. SHAP values increase model transparency by offering interpretability at the global and local levels. Globally, they estimate how much each variable contributes, either positively or negatively, to the target variable. Locally they explain why a given observation is assigned as belonging to a class and the contributions of the variables. We extend the notion of SHAP values to logistic regression. We estimate the values for both approaches and analyze them with the help of graphical tools. It is worth mentioning that we also adjust the estimates of SHAP values for categorical variables to better account for the interdependence of each category.

Our results show that the machine learning approach not only obtains better results in terms of performance, but also sheds light on the complexity of the problem at hand, which is often obscured by linear approaches. More precisely, the SHAP values reveal that machine learning approaches can detect complex nonlinear relationships, including dispersion and structural breaks, that cannot be reflected by logistic regression.

The rest of this document is organized as follows. In Section 3.2, the relevant literature on machine learning and credit risk modeling in P2P lending is reviewed. Section 3.3 presents

the elements and concepts of explainability of predictions in machine learning considered in this paper. In Section 3.4, the data set, models and methods adopted are presented. Section 3.5 presents the empirical results of the analyses, including explainability elements. Section 3.6 concludes by summarizing the findings and giving some future research directions.

3.2. Machine learning and credit risk modeling in P2P lending

In the P2P lending literature, the granting and behavior models typically rely on machine learning methods to obtain better predictions. For example, [Malekipirbazari & Aksakalli \(2015\)](#) use random forests and compare their performance against that of k-NN, SVM and logistic regression. Artificial neural networks are used in the works by [Zhang et al. \(2017c\)](#), [Zhang et al. \(2017b\)](#), [Yuan et al. \(2018\)](#), and [Duan \(2019\)](#). The latter compares the performance of an artificial neural network approach with that of alternatives such as logistic regression, linear discriminant analysis, decision trees and support vector machines. Machine learning methods can be found together with survival models in the articles by [Wang et al. \(2018b\)](#) and [Jiang et al. \(2019\)](#).

Machine learning has been also used to address important aspects at a predictive level such as the problem of class imbalance [Boiko Ferreira et al. \(2017\)](#), [Namvar et al. \(2018b\)](#) in the optimization of hyperparameters, for example, using genetic algorithms [Yang et al. \(2019\)](#) and in feature selection [Van-Sang et al. \(2019\)](#) or the improvement of bias reflected in the prediction [Xia \(2019\)](#); [Xia et al. \(2018\)](#).

However, it is very rare to find studies that use machine learning and include some type of explainability or interpretability analysis. In this regard, [Jin & Zhu \(2015\)](#) propose artificial neural networks, decision trees and SVM to predict the probability of default and only include, as an element of interpretability, a relative importance analysis of the variables for the methodologies of artificial neural networks and decision trees. Similarly, [Li et al. \(2018a\)](#) use an ensemble of techniques such as XGBoost, deep neural networks

and logistic regression and compare both individual methods and different ensembles. In their comparison, they categorize their proposals according to interpretability and indicate that the best prediction technique, which is the ensemble of the three methods, is poor at interpretation. In addition, for the XGBoost technique they also include a feature importance analysis. The most interesting precedent is perhaps the research by [Ma et al. \(2018b\)](#). It includes novel variables such as telephone usage patterns to improve the predictive ability of the probability of default using AdaBoost, random forests and logistic regression. This study not only performs an analysis of feature importance of the machine learning models but also evaluates the logistic regression monotonicity through the sign of the coefficients.

On the other hand, interpretability can be found in the works that use artificial intelligence (or other innovative methods) to include new variables in a logistic regression model. The coefficients of the new variables are analyzed in terms of interpretability, and the relevance is verified using statistical inference. This is the case of the work by [Yao et al. \(2019\)](#). A logistic regression model is proposed with variables derived from text mining applied to the purpose of a loan. Similarly, [Ahelegbey et al. \(2019\)](#) use latent factors models and connectivity networks to generate segments, on which they estimate different logistic regression models for a set of small and medium-sized enterprises. However, in both cases, the model is a statistical one and its interpretation is similar to that for statistical models.

In summary, in P2P lending it is unusual to find granting or behavior credit models with machine learning prediction techniques that include interpretative or explanatory elements. If there are, the only component that frequently appears is that of feature importance for the techniques based on decision trees. Although several research works have tried to overcome the lack of interpretability with these feature importance measures, they fall short of the goal of model understanding ([Florez-Lopez & Ramon-Jeronimo, 2015](#)).

Something similar occurs with the proposals to predict or explain the return or benefit received by the lenders in the P2P market, among which are profit scoring models. While interpretive components naturally appear in classical econometric models based on regression,

in machine learning proposals, if they are considered, it is only through feature importance. See, for example, the works by [Serrano-Cinca & Gutierrez-Nieto \(2016\)](#), [Xia et al. \(2017a\)](#), [Ye et al. \(2018\)](#), [Bastani et al. \(2019\)](#) and [Cho et al. \(2019\)](#).

Likewise, in credit risk modeling of traditional products ([Chen et al., 2014](#); [Lahsasna et al., 2010](#); [Lin et al., 2012](#); [Teles et al., 2020](#)), decision trees are often used, as they provide both a nonlinear mapping and an interpretable model.

Rule extraction is another machine learning tool used in this area because it is both effective and interpretable. Interestingly, different authors have used indicators to evaluate the interpretability of credit risk rule-based approaches. For example, [Florez-Lopez & Ramon-Jeronimo \(2015\)](#) evaluate a novel decision tree ensemble approach using the number of rules, the number of features, a measurement of the distinguishability on variable partitions, etc. Similarly, the fuzzy rule-based credit classification method proposed by [Gorzalczany & Rudzinski \(2016\)](#) is measured in terms of not only accuracy but also interpretability. They use an indicator that measures the fuzzy rule complexity and also the numbers of rules, features, fuzzy sets describing the attributes, etc. Along the same line, [Hayashi & Oishi \(2018\)](#) propose a rule extraction method for credit scoring and compare it against other rule extraction methods using the two dimensions of accuracy and interpretability (number of rules). The indicators used for transparent rule-based methods mostly measure the complexity of the set of rules extracted. Our contribution is vastly different, as we try to add transparency and interpretability to machine learning models. We do this with the help of surrogate models to determine which variables contribute more to the classification and in which direction.

This is an important need because the academic literature on machine learning methods usually lacks an interpretability analysis, e.g., the work with artificial neural networks in [Abdou & Pointon \(2011\)](#). This lack is noted in surveys about the use of machine learning in credit risk ([Lessmann et al., 2015](#)) or even other forms of risk in finance ([Leo et al., 2019](#)). [Andriosopoulos et al. \(2019\)](#) note that as the analytical models for credit risk analysis become more complex, their understandability becomes an important issue, particularly

from a supervisory point of view. However, private companies are turning their attention to interpretability as evidences the Explainable Machine Learning Challenge sponsored by FICO, a well-known credit scoring company (Chen et al., 2018; Flint et al., 2018; Holter et al., 2020).

The conclusion we draw from the literature review is that machine learning approaches are typically used in P2P lending credit risk models. However, the studies usually do not address the interpretability of the models in depth. This lack also occurs in the area of credit risk for traditional products and in other financial areas. Nevertheless, it is often recognized that interpretability is of critical importance and that regulatory entities and end-users demand these aspects as essential elements (Carvalho et al., 2019).

This work aims at filling this gap. To this end, we use the SHAP values for explaining machine learning models in the context of credit risk in P2P lending. SHAP values have been successfully used in different contexts, including the medical domain (Lundberg et al., 2018b). In the following section, we introduce the key concepts of explainability in artificial intelligence and machine learning, as well as the techniques most commonly used for the explainability of prediction and in particular the SHAP values.

3.3. Explainability of predictions in machine learning

In this section, we introduce the SHAP values that will be used below to explain the credit risk models. However, in order to make the article-self contained, we will briefly present some key concepts and works of machine learning explainability that will help us to frame the SHAP values better.

Recently, several theoretical works have proposed interpretability and explainability as essential aspects of machine learning methods (Doshi-Velez & Kim, 2017; Kim et al., 2016; Miller, 2019). Carvalho et al. (2019) review interpretability from an ontological and epistemological perspective. They evidence the growing interest in the approach by surveying the scientific events dedicated to it. Lipton (2018) reflects on the importance and complexity

of the interpretation component in modeling and highlights the demand and need to have predictive and interpretable models and, in turn, refine the conception of interpretability in machine learning models.

From a practical perspective, [Molnar \(2021\)](#) defines and establishes some elements of interpretability and explainability and details how several techniques can be applied for this purpose. In turn, in a much more specific line, he presents three properties to take into account in interpretable models: linearity, monotonicity, and interaction.

- Linearity refers to the linear association between a variable and the target variable.
- Monotonicity indicates that the relationship between a specific input and the target outcome always follows the same direction throughout the entire domain of characteristics.
- Interaction is the ability to naturally include interactions between features to predict the target variable.

The above aspects and others, such as the relative importance of the variables included in the prediction models, can be evaluated via some methods (post hoc). A broad subset corresponds to agnostic methods of explanation, which are useful for interpreting what a model does, utterly independent of the technique used to create the model. Among the most popular methods are partial dependence plots ([Friedman, 2001](#)), local effects graphs ([Apley & Zhu, 2016](#)), which are an alternative to partial dependence plots with a lower computational cost, and variable importance ([Fisher et al., 2018](#)), which offers a set of measures to assess the importance and dependence of the variables in a model, or Individual Conditional Expectation (ICE) plots ([Goldstein et al., 2015](#)) which highlight the variation in the fitted values across the range of a covariate. As it can be seen, many explainability methods use graphical representations to ease the interpretation.

A more sophisticated approach is offered by the LIME methodology ([Ribeiro et al., 2016](#)), which uses local surrogate models to explain the individual predictions of machine learning

models. It aims to understand why the machine model made a particular prediction under the assumption of local linearity. The LIME methodology includes the use of permutations of data samples in local linear regression approximations to observe the resulting impact on the output.

Lundberg et al. (2019, 2018a); Lundberg & Lee (2017) propose SHAP values to give interpretation capacity to complex models of diverse nature. The SHAP values combine, among others, ideas from the LIME methodology and the Shapley values (Shapley, 1953), which explain the value of the predictions, assuming that each feature in a model is a player in a game in which the prediction is the payment. SHAP values presents an efficient solution to the computational challenge demanded by explanation models, taking into account all possible orders of the variables in the evaluation of feature importance. Additionally, they allow one to obtain a global interpretation based on aggregations of Shapley values. The next section explains how they work.

3.3.1. SHAP values

SHAP values are based on the definition of Shapley values (Shapley, 1953), which explain the prediction through the marginal contribution of each feature. The feature values of an instance of the data set behave as actors in a coalition, and Shapley values allow a fair distribution of the payoff, in this case the prediction, according to their contribution.

To set the SHAP values Lundberg & Lee (2017) start by generalizing the explanation models as a class called additive feature attributions methods that includes, among others the linear model, LIME (Ribeiro et al., 2016), DeepLIFT (Deep Learning Important Features)(Shrikumar et al., 2017) and classic Shapley value estimation.

The additive feature attribution methods class is based on an explanation model g defined as an interpretable approximation of the original f prediction model. The explanation model can be written as a linear combination of binary variables.

$$g(z') = \phi_0 + \sum_{i=1}^M \phi_i z'_i, \quad (3.1)$$

where M is the number of input features, $\phi_i \in \mathbb{R}$ represents feature attribution values of feature i and $z' \in \{0, 1\}^M$ with z'_i representing a feature being observed ($z'_i = 1$) or unknown ($z'_i = 0$). In game theory terms, z' represent the coalition vector and M the maximum coalition size.

A relevant characteristic of the additive feature attribution methods class is that there is a single unique solution with three desirable properties: local accuracy, missingness and consistency. The local accuracy indicates that the result of the explanation model matches the original prediction model that you want to locally explain. Missingness asserts that missing features have no importance. Finally, consistency establishes that if a model changes such that the contribution of some feature increases or remains the same regardless of the other entries, the allocation of that input should not decrease.

Therefore, derived from the first property, the attribution values capture the difference between the output for prediction $f(x)$ and the expected model output based on a single input x . This can be represented as

$$f(x) - \phi_0(f, x) = \sum_{i=1}^M \phi_i(f, x), \quad (3.2)$$

with $\phi_0(f, x) = E[f(z)] = E[f(x)]$ being the expected value of the model over the training data set and $\phi_i(f, x)$ a numerical value that represents the impact of characteristic i in the prediction of model f given input x .

To fulfill this property and the others, SHAP values $\phi_i(f, x)$ are calculated based on the idea of Shapley values to attribute ϕ_i values to each feature as in (3.3):

$$\phi_i(f, x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)], \quad (3.3)$$

where N is the set of all input features, S is a coalition of features representing the set of nonzero indexes in z' associated with the present features, and $f_x(S) = E(f(x)|x_S)$ corresponds to the expected value of the function conditioned on a subset S of the input features.

Figure 3.1 shows a graphical representation of this idea, where $f(x) = \sum_{i=0}^M \phi_i$. SHAP values explain the prediction f function as a sum of the effects ϕ_i of each feature introduced into a conditional expectation. The figure illustrates a single ordering of the features and how the contribution of each feature is introduced, one at a time, into a conditional expectation function of the f output model. In nonlinear models or when the input features are not independent, the SHAP values result from averaging all possible orderings (Lundberg & Lee, 2017).

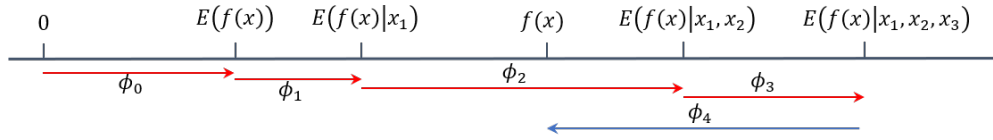


Figure 3.1: Graphical representation of SHAP values

Source: Similar to that in Lundberg et al. (2018a)

The efficient estimation of $E(f(x)|x_S)$ and the complexity of evaluating all the possibilities derived from Equation (3.3) pose a computational challenge. Lundberg et al. propose different algorithms depending on the typology or structure of the machine learning model to estimate SHAP values. For efficient computation in tree-based models, they propose Tree SHAP Lundberg et al. (2019, 2018a), taking advantage of the additivity property of the Shapley values and the hierarchy of the trees.

For categorical variables, Shapley values are computed for each category under the assumption of independence. Thus, the nonpresence of each of the categories affects the value of the attribution based on the prediction of the presence of the category. In other words, there is an aggregate effect that must be contemplated, and it is not considered in the proposals by Lundberg et al. to the best of our knowledge. We propose a straightforward approximation to recalculate Shapley values for categorical values as follows:

$$\phi_{jk^*i} = \begin{cases} \sum_k \phi_{jki} & \text{if } x_{jk} = k^* \\ 0 & \text{else} \end{cases} \quad (3.4)$$

The Shapley values estimated under the assumption of independence are added for each option of the categorical variables, including the Shapley value that represents that the i th instance does not present the k^* option in categorical variable j , while ϕ_{jki} represents the estimated Shapley values under independence for each category k in instance i ; we compute ϕ_{jk^*i} , which represents the Shapley value of the j th categorical variable for category k^* in instance i .

A remarkable feature of SHAP values is the visualization tools developed to better appreciate their insights, including the SHAP summary plot and the dependence plot [Lundberg et al. \(2018a\)](#).

The SHAP summary plot summarizes the individual attributions and allows us to appreciate feature importance and monotonicity. Feature importance is calculated by taking the average of the absolute values per feature across the data.

$$I_j = \frac{1}{n} \sum_{i=1}^n |\phi_{ji}| \quad (3.5)$$

In the SHAP summary plot, features are first sorted by their global impact (4.14). Then dots representing the SHAP values ϕ_{ji} are plotted horizontally and stacked vertically when they run out of space. Each dot is colored by the feature value, from low to high.

The relative feature importance can be estimated as the share of the importance of variable j over the aggregate importance for all the variables:

$$\frac{I_j}{\sum_m I_m} \quad (3.6)$$

The dependence plots shows the effect that a feature has on the predictions made by the model. Dependence plots are scatter plots in which each dot represents the feature value and the SHAP value of each individual $\{(x_{ji}, \phi_{ji})\}_{i=1}^n$. This plot makes it possible to appreciate not only monotonicity if it exists but also heteroscedasticity and the shape of the relationships.

It is important to remark that the estimated SHAP values can be represented in units of log-odds ratios, which we adopt. Assuming additivity of the importance of characteristics

in this space is natural, as this occurs with the link function in the logistic regression model (Lundberg et al., 2018). This will enable us to compare the explainability of both models in a graphical manner.

Because of the richness of SHAP values, other plots can be used to facilitate the explainability and interpretability of a model's prediction.¹

Since SHAP values are finally an estimate of the Shapley values, the terms will be used interchangeably in the rest of the document.

3.3.2. Explainability elements in logistic regression

In this section, we review some of the explainability elements of the well-known logistic regression model and propose a straightforward extension of the Shapley values for logistic regression to enable the comparison with the machine learning approach.

Let Y_i be the default variable in obligation i with respect to the variables $x_{1i}, \dots, x_{ki}$ that describe the features of a borrower. In a logistic regression, the probability of default is denoted

$$P(Y = 1|x_{1i}, \dots, x_{ki}) = F(x_{1i}, \dots, x_{ki}) \quad (3.7)$$

with F being the link function, which, in this case, is the logistic distribution, so

$$P(Y = 1|x_{1i}, \dots, x_{ki}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)} \quad (3.8)$$

If

$$\text{logit}(x) = \log\left(\frac{x}{1-x}\right), \quad (3.9)$$

then

$$\text{logit}(P(Y = 1|x_{1i}, \dots, x_{ki})) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad (3.10)$$

and $\beta_0, \beta_1, \dots, \beta_k$ can be estimated through the maximum likelihood method.

¹https://github.com/slundberg/shap/blob/master/notebooks/plots/decision_plot.ipynb

The interpretation of the logistic model will be based on the coefficients, the estimated marginal effects and the odds ratios. For a variable j , these are $\hat{\beta}_j$, the odds ratio obtained from $\exp(\hat{\beta}_j)$, and the estimated marginal effect,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \frac{\partial P(Y = 1|x_1, \dots, x_j, \dots, x_k)}{\partial x_j} &\approx \\ \frac{1}{n} \sum_{i=1}^n \hat{P}(Y_i = 1|x_{1i}, \dots, x_{ki})[1 - \hat{P}(Y = 1|x_{1i}, \dots, x_{ki})] &\hat{\beta}_j. \end{aligned} \quad (3.11)$$

The betas are related to changes in the logistic link function, where positive values are associated with a higher probability of the event under evaluation, default, and negative values with a reduction in this probability.

For categorical variables, their interpretation depends on the category defined as the basis for the construction of dummy variables. If the coefficient is positive, then there is an increase in the average logit equivalent to the value of the coefficient compared to the base category, which implies a greater probability. These values are considered relevant depending on their respective inference.

As the logistic response function is essentially nonlinear, it is not possible to directly interpret the betas as marginal effects versus probability, which is why the estimated marginal effects are also presented.

An odds ratio is a ratio of probabilities, a measure of the association between features and the default. It is determined by the ratio of the probability of default given a particular exposure to any of the features of borrower to the probability that the result will occur in the absence of that exposure. It evaluates whether an obligation with the risk factor or feature is more or less likely than an obligation without that risk factor for the result of interest, default.

It follows that an odds ratio with a value of 1 indicates that the presence of the risk factor is not associated with the probability of occurrence of default and will be determined by a beta close to zero.

Finally, given Equation (3.2), the Shapley values can be extended to logistic regression

models. The contribution of the i -th variable to the f prediction function, i.e., the *logit* function in the logistic regression model, is determined in Molnar (2021) as

$$\phi_i(f, x_i) = \beta_i x_i - E(\beta_i x_i) \quad (3.12)$$

So,

$$\begin{aligned} \sum_{i=1}^k \phi_i(f, x_i) = \\ \beta_0 + \sum_{i=1}^k \beta_i x_i - E\left(\beta_0 + \sum_{i=1}^k \beta_i x_i\right) = f(x) - E(f(x)) \end{aligned} \quad (3.13)$$

Thus, we estimate the $\phi_i(f, x_i)$ values for the logit prediction function through $\hat{\beta}_i(x_i - \bar{x}_i)$.

For categorical variables, we propose to correct the estimates of the Shapley values using the aggregation approach shown in Equation (3.4).

3.4. Experimental setting

For our experiment on credit risk modeling at the granting stage in P2P lending, we use a public data set from Lending Club (LC). We compare both the performance and the explainability of the several machine learning classifiers and the logistic regression model (LR), which is widely used for the construction of credit risk rating models and preferred by regulators and end users for, among other aspects, the advantages it offers in estimation, inference, implementation and interpretation. Figure 3.2 shows the steps followed in our experiment. More details on each step are given below.

3.4.1. Data preparation

Lending Club offers loans through a technological platform for various personal finance purposes and is today one of the companies that dominate the US P2P lending market. The considered data set is publicly available in Kaggle² and corresponds to all the loans issued by

²<https://www.kaggle.com/wordsforthewise/lending-club>

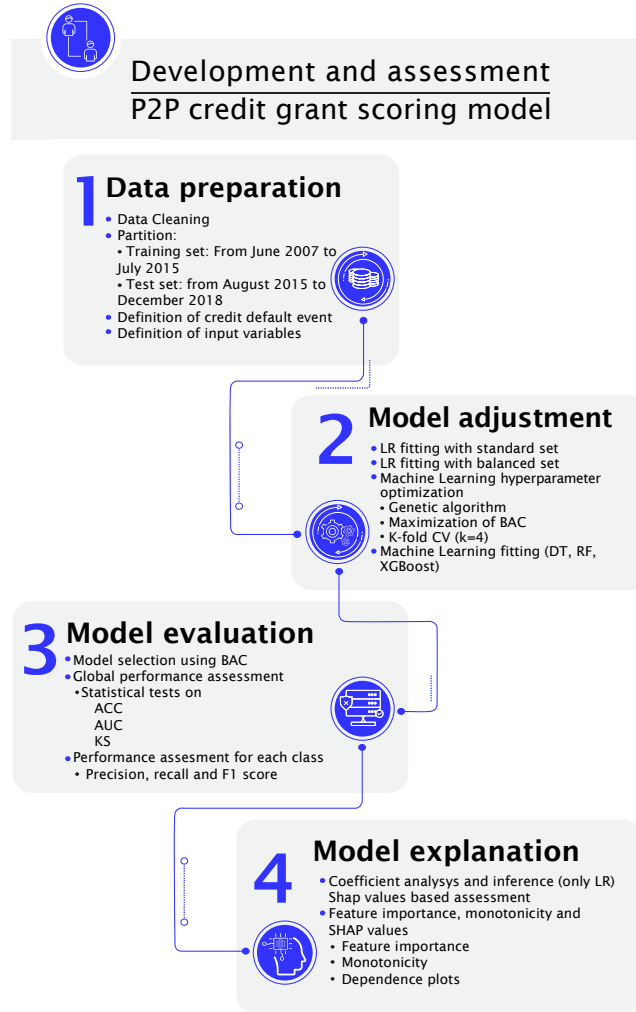


Figura 3.2: Steps of the model development and assessment

Lending Club between 2007 and 2018. Loans are described by 75 features, including credit scores, number of finance inquiries, address (including zip code and state), and collections, among other features.

Since we are building a model for granting credit, we create our target variable based on the final resolution of the credit: the default category corresponds to the event charged off and the nondefault category to the event fully paid. We do not take into account other values in the loan status variable since this variable represents the state of the loan at the end of the considered time window. Thus, there is no certainty about the stability of default

or not default of the obligation.

As a result, our data set consists of 1,347,681 records or obligations (approximately 60 % of the available data set). We also clean the resulting data set for completeness and consistency (less than 1 % of our data set was filtered out). The default rate in the final data set is close to 20 %.

The explanatory variables that we use correspond only to the information available at the time of the application. Variables such as the interest rate, grade or subgrade are generated by the company as a result of a credit risk assessment process, so they should not be taken into account in risk models to predict the default in granting of credit.

Furthermore, we construct some variables to aggregate or simplify the information described by several variables. The credit experience with LC (*experience_c*) is constructed as a binary variable that indicates whether the borrower is new for the entity. This variable is constructed from the credit date of the previous obligation in LC and the credit date of the current obligation; if the difference between dates is positive, it is not considered as a new experience with LC. The debt variable (*dti_n*) is estimated as the ratio calculated using the co-borrowers' total payments on the total debt obligations divided by the co-borrowers' combined monthly income. In the original data set, information from the Fair Isaac Corporation credit bureau (FICO) is given by two values that define the credit scoring interval of the applicant to the bureau. Our FICO variable (*fico_n*) is estimated as the average of these two values. Table 3.1 lists the input variables that we use in our models, 5 categorical and 4 quantitative.

We can make an assumption about some relationships between input and target variables according to our experience and the literature [Serrano-Cinca & Gutierrez-Nieto \(2016\)](#); [Yao et al. \(2019\)](#) on a different data set. For example, we expect a positive relationship between monthly indebtedness (*dti_n*), the amount of the loan requested (*loan_amnt*) and the probability of default, and an inverse relationship of this probability with the bureau score (*fico_n*) and the annual income(*revenue*). We will validate such assumptions in Section

3.5.3.

Finally, we divide the data set into two time windows, the first with obligations generated until July 2015 (657,602 records), with which the training process is carried out. The loans generated from August 2015 until December 2018 (690,079 records) are used as the test set. With this time-aware training-test split, we intend to evaluate the proposed models in a manner as realistic as the data allow because the information available in the Kaggle data set does not specify when the final state of the loan is reached.

Tabla 3.1: Description of the explanatory variables according to the information in Kaggle

<i>Variable</i>	<i>Description</i>
Categorical variables	
<i>emp_length</i>	Employment length. Current employment time in years categorized by LC into 12 categories, including the no information category.
<i>experience_c</i>	Previous credit experience with LC (binary).
<i>purpose</i>	Purpose of the loan provided by the borrower. It has 14 possible values: car, credit_card, debt_consolidation, educational, home_improvement, house, major_purchase, medical, moving, other, renewable_energy, small_business, vacation, wedding.
<i>home_ownership</i>	Home ownership status provided by the borrower during the registration process. Categories defined by the entity: Mortgage, rent, own, other (other, none and any).
<i>addr_state</i>	State in the US provided by the borrower in the loan application.
Quantitative variables	
<i>revenue</i>	Yearly income self-reported in the registration process.
<i>dti_n</i>	Debt ratio for the group of applicants for obligations excluding mortgages. Monthly information. Income self-reported.
<i>loan_amnt</i>	Amount of credit requested by the borrower.
<i>fico_n</i>	Credit bureau score. Defined between 300 and 850, reported by Fair Isaac Corporation as a summary risk measure based on historical credit information reported at the time of application.

3.4.2. Model adjustment

Logistic regression

Given the low default rate of close to 20 %, two proposals for logistic regression models are presented. One estimates the model with the observed training data (we call it LR), and the second model adjusts the training data set with a sampling method for the class

imbalance problem (we call it LR.BS, where BS stands for balanced set).

The sampling method used in the second case is a hybrid between undersampling and oversampling (Guo et al., 2017), also used in (Boiko Ferreira et al., 2017; Namvar et al., 2018b), which consists of equating the minority class proportion, default, with the nondefault randomly selected cases from the majority class and generating random records from the minority class.

Machine learning algorithms

The machine learning algorithms considered are decision tree, random forest and XGBoost. These methods have already been used with good results for the evaluation of credit risk in P2P lending (Bastani et al., 2019; Boiko Ferreira et al., 2017; Jiang et al., 2019; Li et al., 2018a; Serrano-Cinca & Gutierrez-Nieto, 2016; Xia, 2019). Still, they were limited in terms of interpretation, except for the decision tree model.

Decision tree (DT) is one of the most widely used machine learning models for credit risk because it combines interpretability and predictive capability (Serrano-Cinca & Gutierrez-Nieto, 2016). It produces a hierarchical tree representation to visualize classification, where nodes represent rules over features and leaves represent categories. However, the resulting tree is highly dependent on the training data set. We will use DT as a reference for more sophisticated machine learning.

Random forest (RF) (Breiman, 2001) is a popular ensemble technique that combines a large number of independent decision trees estimated on random data sets. A prediction from random forest corresponds to an aggregation of the independent predictions of each of the single trees via averaging, or voting for the classification case.

The method of eXtreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016) is a popular implementation of the gradient tree boosting approach. Boosting is a reinforcement algorithm that adds iterations of the model in a sequential process by adjusting the weights of the weak learners (trees), minimizing the error iteration after iteration. Each subsequent tree aims to reduce the errors of the previous tree. This reduces model bias and generally

improves accuracy. In particular, XGBoost is an advanced gradient boosting model that avoids overfitting by weighing the decrease of the objective function and the complexity of the model.

For DT and RF we used the Python implementations from scikit-learn (Pedregosa et al., 2011), while for XGBoost, we used the implementation in Python.³

The machine learning models we propose have several parameters that can be fine tuned. We carried out hyperparameter optimization using a genetic algorithm (GA).⁴

Our GA optimizes the inverse of the balanced accuracy (BAC) loss function. We use BAC to increase the importance of the error in the minority class (default). More precisely, we use the fitness value in a stratified k-fold cross-validation setting with $k = 4$ using the training data set previously described, where stratification helps preserve the class imbalance in the same proportion in each fold.

The parameters considered in the optimization process depend on the machine learning algorithm and its implementation. However, they relate either to the learning capacity of the algorithm or to how to avoid overfitting. For example, in the case of RF and XGB, we include the number of trees, the maximum depth of the trees, the subsample ratio used to grow each tree, etc. For DT, we use the maximum depth, the number of samples required to split an internal node, or to be at a leaf node. Furthermore, in all three cases, we have set the parameters to account for the class imbalanced.

The settings of the genetic algorithm search are:

- Epochs: 20
- Initial population: 50
- Mating method: two-point crossover
- Selection: Select the best individual among four randomly chosen individuals, 100 times

³https://xgboost.readthedocs.io/en/latest/python/python_intro.html

⁴The Python library DEAP, <http://deap.readthedocs.io/en/master/>

- Mutation: Mutate an individual by replacing attributes, with probability 0.35 by a number uniformly drawn between the decided lower and upper bounds of the attribute.

3.4.3. Model evaluation

We measure the performance of the proposed models using several well-known measures frequently used in credit risk: the classification accuracy rate (ACC), area under the receiver operating characteristic curve (AUC) and Kolmogorov-Smirnov statistic (KS).

We include the balanced accuracy (BAC), which measures the average accuracy obtained from both the minority and majority classes, default and nondefault. This measure emphasizes the importance of the error in the minority class (default), which better represents risk.

Additionally, we assess whether the differences found between the performances in the prediction of the proposed models are considered statistically important. We use McNemar test for the difference between ACCs (McNemar, 1947), DeLong test for the case of AUCs (Sun & Xu, 2014), and Krzanowski-Hand test to compare KSs (Krzanowski & Hand, 2011).

Furthermore, we analyze the performance of each class using precision, recall and the F1 measure.

For each class, precision measures the fraction of correct predictions among the predictions of a class. In contrast, recall measures the fraction of instances of a class correctly retrieved by the classifier. Finally, F1 measure is the harmonic mean of precision and recall and summarizes both measures.

3.4.4. Model explanation

We compare the explanatory power of the logistic regression and the machine learning models using the SHAP values. Such comparison is one of the main contributions of this work. For the sake of brevity, we will only show the explainability of the best machine learning model, which has been XGBoost.

We analyze the relative importance of the variables and the monotonicity and dependence between the independent variables and the target one. These concepts are shown with the help of graphical tools that allow us to identify nonlinearities, structural and distributional changes, and atypical values.

It is important to remark that we recalculate the SHAP values for categorical variables to better account for the interdependence among categories (see Equation 3.4). We show the resulting values as dependence plots to identify different levels of risk by category as well as atypical behavior.

3.5. Results

3.5.1. Exploratory statistical analysis

Tabla 3.2: Exploratory statistical analysis of quantitative variables vs target variable

Variable	All		Default		Non default		KS D-Test
	mean	sd	mean	sd	mean	sd	
<i>fico_n</i>	698.16	31.85	689.83	25.95	700.24	32.83	0.14***
<i>loan_amnt</i>	14408.23	8715.34	15547.02	8813.70	14123.91	8667.31	0.08***
<i>revenue</i>	77369.68	70362.96	71698.23	65906.69	78785.66	71361.97	0.06***
<i>dti_n</i>	18.30	11.15	20.20	11.82	17.83	10.93	0.11***

*** significant at the 0.01 level

As a preliminary analysis, Table 3.3 and Table 3.2 report summary statistics for the whole data set. These results help in understanding the data set and contextualizing the explainability analysis in Section 3.5.3.

The tables include statistical inferences to assess the potential of the variables considered to predict the risk of default and validate the preliminary assumptions laid out in Section 3.4.1. For quantitative variables, we use the Kolmogorov-Smirnov test to compare the empirical probability distributions of the loans in default with those not in default, indicating whether a significant difference is observed. For categorical variables, we use the chi-square test to assess whether there is a significant association between the independent categorical

Tabla 3.3: Relative frequency distribution and default rate by category of the categorical variables

Variable	Category	Rel. Freq.	Default Rate	Chi ² -Test	Variable	Category	Rel. Freq.	Default Rate	Chi ² -Test
emp_length	<1 year	8.05 %	20.54 %	2843.16***	addr_state ^a	AL	1.23 %	23.64 %	3398.94***
	1 year	6.59 %	20.59 %			AZ	2.43 %	19.65 %	
	2 years	9.06 %	19.82 %			CO	2.21 %	15.54 %	
	3 years	8.00 %	19.99 %			CT	1.47 %	17.39 %	
	4 years	5.99 %	19.76 %			FL	7.11 %	21.50 %	
	5 years	6.26 %	19.62 %			GA	3.23 %	18.43 %	
	6 years	4.67 %	19.38 %			IL	3.85 %	18.11 %	
	7 years	4.43 %	19.51 %			IN	1.61 %	21.44 %	
	8 years	4.51 %	19.95 %			KY	0.95 %	21.01 %	
	9 years	3.79 %	19.91 %			LA	1.15 %	23.17 %	
	10+ years	32.85 %	18.80 %			CA	14.60 %	19.63 %	
	NI	5.80 %	26.96 %			MA	2.31 %	19.08 %	
home_ownership	RENT	39.74 %	23.23 %	6733.25***	MD	2.32 %	21.34 %		
	MORTGAGE	49.46 %	17.23 %		MI	2.62 %	20.31 %		
	OWN	10.76 %	20.63 %		MN	1.78 %	19.76 %		
	OTHER	0.01 %	20.88 %		MO	1.58 %	21.34 %		
	ANY	0.02 %	19.58 %		MS	0.49 %	26.12 %		
	NONE	0.00 %	16.33 %		NC	2.81 %	20.79 %		
purpose	car	1.09 %	14.70 %	4180.47***	NJ	3.60 %	21.12 %		
	credit_card	21.93 %	16.93 %		NM	0.55 %	21.37 %		
	debt_consolidation	57.97 %	21.15 %		NV	1.50 %	21.98 %		
	educational	0.03 %	20.80 %		NY	8.17 %	22.05 %		
	home_improvement	6.51 %	17.76 %		OH	3.26 %	20.52 %		
	house	0.54 %	21.91 %		OK	0.91 %	23.46 %		
	major_purchase	2.19 %	18.60 %		OR	1.22 %	14.41 %		
	medical	1.16 %	21.84 %		PA	3.39 %	20.82 %		
	moving	0.71 %	23.41 %		SC	1.19 %	16.28 %		
	other	5.81 %	21.08 %		TN	1.51 %	21.42 %		
	renewable_energy	0.07 %	23.72 %		TX	8.18 %	19.84 %		
	small_business	1.16 %	29.86 %		UT	0.75 %	17.09 %		
	vacation	0.67 %	19.19 %		VA	2.83 %	19.95 %		
	wedding	0.17 %	12.43 %		WA	2.17 %	15.79 %		
experience_c	with financial experience	99.99 %	19.98 %	0.99	WI	1.32 %	18.37 %		
	without financial experience	0.01 %	12.00 %		WV	0.36 %	15.53 %		

***significant at the 0.01 level.

^a for the sake of brevity, we only show some of the most representative states.

variable and the dependent variable.

Table 3.2 reports the descriptive statistics and the test results for the quantitative variables versus the target variable. According to the test, all quantitative variables show important differences in the observed distribution, which supports the use of these variables for risk models. As expected, lower FICO (*fico_n*), lower income (*revenue*), higher requested amount (*loan_amnt*) and higher indebtedness (*dti_n*) seem to be associated with higher risk. The average values descriptively support this proposition.

Table 3.3 presents the descriptive statistics and the test results for the categorical variables. According to the tests, all variables are associated with the target variable, except for the credit experience in LC variable (*experience_c*), which does not seem to be an essential determinant of default risk at the granting stage. We analyze them in detail below.

For the employment length variable (*emp_length*), the proportion of default in the category of more than ten years represents the lowest default proportion (18.8 %). The highest percentages are given for options of shorter length, one or less than one year, with values close to 21 %, showing an expected ordering of this variable with default. It is worth mentioning that there is a percentage of obligations that do not report this value, and their risk proportion is high (27 %).

For the home ownership variable (*home_ownership*), the category that evidences higher risk is rent (23.3 % default), while for the purpose variable (*purpose*), the small business category presents the most considerable risk (approximately 30 %).

Regarding the credit experience in the LC variable (*experience_c*), similar variables are typically included in P2P credit risk works (Bastani et al., 2019), even if some of them find them not relevant (Serrano-Cinca & Gutierrez-Nieto, 2016; Serrano-Cinca et al., 2015). Despite the negative test result, we decided to consider it because the machine learning algorithms could be capable of recognizing nonlinearities and more complex association structures. In addition, we want to observe through SHAP values the degree of association of this variable. Logistic regression, although it allows some of these elements to be seen

inferentially, is limited when the data exhibit complex underlying structures.

3.5.2. Model evaluation

Tabla 3.4: Performance on training and test samples

Model	Training				Test			
	BAC	ACC	AUC	KS	BAC	ACC	AUC	KS
LR	50.1	81.9	65.4	22.1	50.2	78.1	66.6	23.9
LR.BS	61.0	61.0	65.4	22.1	61.9	60.5	66.6	23.8
DT	61.5	58.5	66.1	22.2	60.5	60.8	64.7	18.0
RF	62.7	62.7	67.8	26.1	61.4	64.4	66.3	24.0
XGB	62.8	62.6	68.0	27.2	62.4	63.6	67.4	26.4

Tabla 3.5: Performance comparison for the test sample

Differences	ACC ^a	AUC ^b	KS ^c
XGB-LR	-14.54***	0.81***	2.51***
XGB-LR.BS	3.11***	0.84***	2.58***
XGB-DT	2.82***	2.71***	8.49***
XGB-RF	-0.78***	1.06***	2.36***

***significant at the 0.01 level.

^a McNemar test, ^b DeLong test and ^c Krzanowski-Hand test.

Table 3.4 shows the performance results of the methods considered. According to the (unbalanced) accuracy measure the LR is much better than the others. However this is due to the imbalanced nature of the data set. An examination of the performance of LR on the default and nondefault classes in Table 3.6 reveals that is a bad risk model: the LR classifier identifies less than 1 % of the defaulted loans (recall 0.7 %), and the success rate in labeling nondefault loans is much lower than that for the other models (approximately 78 % versus 85 %). Accuracy is not suitable to compare the classifiers of unbalanced sets.

If we analyze the other measures in Table 3.4, XGB outperforms the other models according to the balanced accuracy used to adjust the models, but also according to the KS and AUC that assess the ability of a classifier to differentiate between the two classes. Table 3.5

shows that the differences in the global metrics are statistically significant, which reinforces the idea that XGB is globally better than the other alternatives. It is worth mentioning that the results show that DT performs worse than the LR.BS in this data set.

Tabla 3.6: Performance measures by class for the test sample

Model	Default			Non default		
	Precision	Recall	F1	Precision	Recall	F1
LR	44.4	0.7	1.3	78.3	99.8	87.7
LR.BS	30.7	64.4	41.6	85.7	59.4	70.2
DT	30.1	60.1	40.1	84.6	61.0	70.9
RF	32.0	56.0	40.7	84.5	66.7	74.5
XGB	32.1	60.2	41.9	85.3	64.5	73.5

Table 3.6 shows that according to the F1 measure, the XGB model is better than the others for the default class. The RF has the second best F1 measure (after the LR) for the non default class due to a high recall. However, it performs worse in terms of precision of the nondefault class, which is not convenient for a grant model. In any case, this table illustrates the complexity of the classification problem at hand and the subtlety of the different behaviors that the classifiers exhibit.

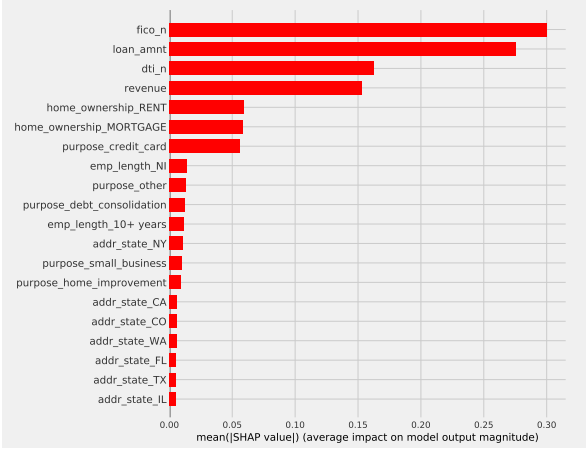
According to these results, we can say that the XGB globally performs better than the other methods considered.

3.5.3. Model explanation

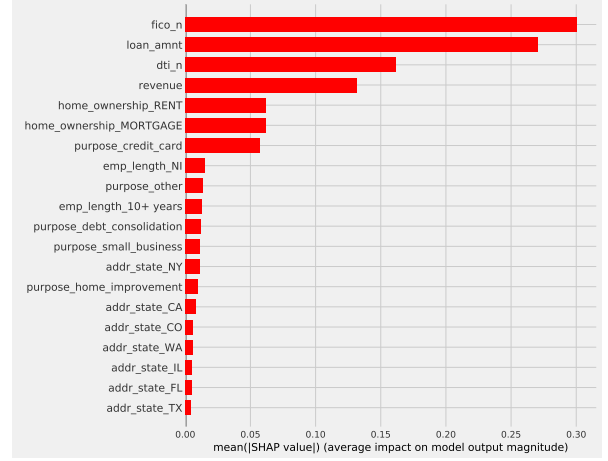
In this section, we compare the explainability capabilities of both the XGBoost and the logistic regression using Shapley values.

Feature importance, monotonicity and Shapley values

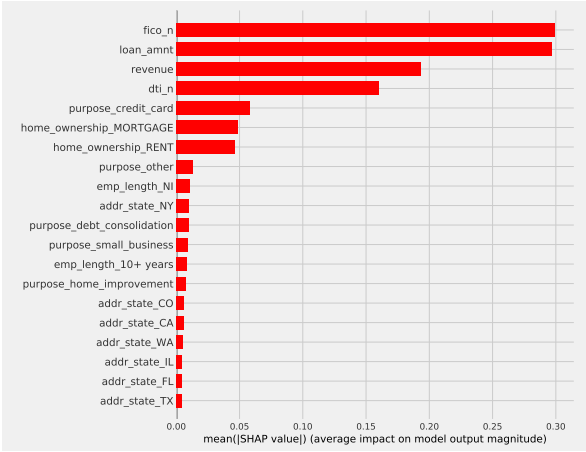
Figures 3.3 and 3.4 show the feature importance of the two logistic regressions and the XGBoost approach. In Figure 3.3, we can see the importance of the different values of the categorical variables, while in Figure 3.4, we can see the aggregated importance of the categorical variable.



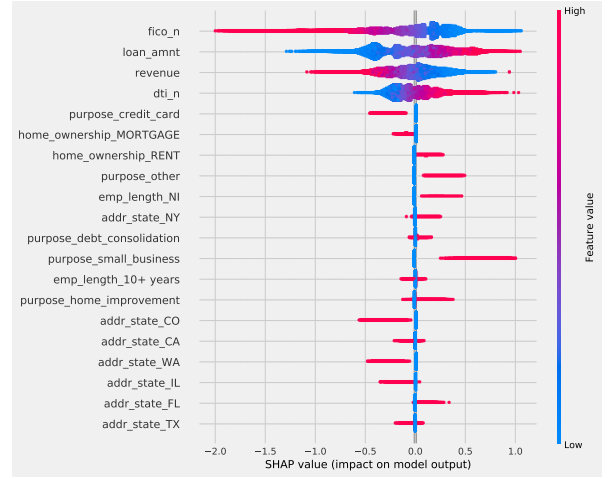
(a) Feature importance for LR



(b) Feature importance for LR-BS



(c) Feature importance for XGBoost

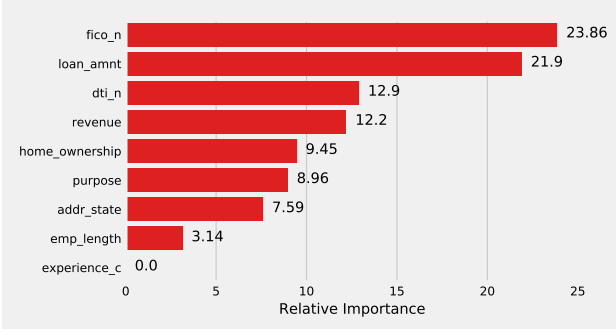


(d) Dependence and monotonicity of features for XGBoost

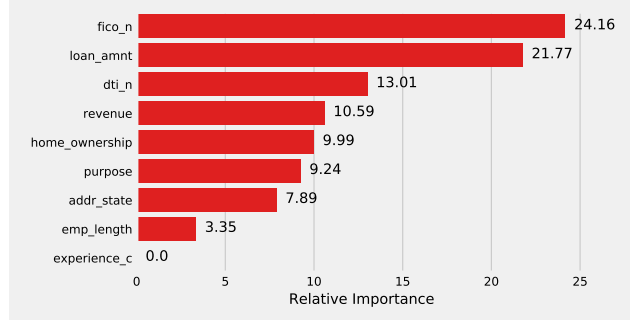
Figure 3.3: Feature importance for each considered model

According to Figure 3.4, the importance ranking in both LR and LR.BS is exactly the same, even if the importance values slightly vary. In the case of XGB, the ranking is mostly the same, but the positions of the *revenue* and *dti_n* variables and of the *home_ownership* and *purpose* variables are swapped.

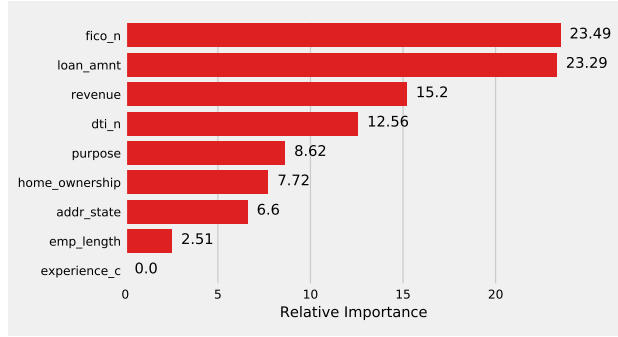
Interestingly, quantitative variables are more important than qualitative ones (roughly 70 % vs 30 %). The FICO credit score (*fico_n*) is the most important variable in all the cases, followed by *loan_amnt*. For the categorical variables, in the logistic regression mo-



(a) Aggregated relative importance LR



(b) Aggregated relative importance LR.BS



(c) Aggregated relative importance XGB

Figure 3.4: Aggregated relative importance by model

models, *home_ownership* is the most relevant one, while in XGBoost, *purpose* is. In Figure 3.3, we can see the most influential categories, which are RENT and MORTGAGE for *home_ownership* and credit_card from *purpose*. Again, the order in XGBoost is different from that provided by the logistic regression models, but the three approaches show that these three categories are notably more important than the rest.

Unsurprisingly, the feature importance ranking from XGBoost slightly differs from that from logistic regression. The usefulness of the variables for classification is mostly similar, even if the methods exploit them in different manners.

It is worth mentioning that in Figures 3.3a, 3.3b and 3.3c, the order of importance of the values of categorical variables is affected by their frequency. In this sense, the ranks compare categories and variables, although they are not exactly the same, and each category has a frequency that affects its importance. However, we include them in the ranking to gauge

the contribution of each category to the prediction model. In contrast, in Figure 3.4, we perform a more robust estimation of the importance of variables based on Shapley values, considering each categorical variable by aggregating categorical values.

Beyond feature importance, we show the Shapley values of each feature for XGBoost in Figure 3.3d. In this figure, the X axis represents the Shapley values, where positive values mean higher default probability and negative values correspond to lower default probability, while the Y axis has the features sorted by feature importance; finally, the dots are colored using a gradient that ranges from blue (lower feature values) to red (higher feature values). Thus, if a feature presents a change on the horizontal axis from blue to red the Shapley value increases, then there exists a monotonically increasing relationship with respect to the risk of default. This is the case for the *loan_amnt* and *dti_n* features. In contrast, the relationship is monotonically decreasing for the case of *fico_n* and *revenue*. For *revenue*, we can find an outlying red dot that breaks the monotonicity and that represents a case where extremely high revenues increased the risk of default. Interestingly, the careful analysis of these detailed plots can help identify cases with particular behavior that may require further inspection to avoid model overfitting.

For some features, we can observe how the color barely changes in a part of the observed range of Shapley values, which means that for the same feature value, there is a range of different attributions in the observed individuals, i.e., the same feature value does not have the same impact on all the individuals, depending on the values of the rest of the features. This is an interesting aspect that cannot be appreciated in logistic regression models. This phenomenon can be clearly appreciated in the categorical values, each of them exhibiting different ranges. For example, the value home improvement as the *purpose* of the loan can both decrease and increase the default probability, while the *purpose* small business always increases the default probability, but its impact greatly varies.

In logistic regression models, we can analyze the coefficients, marginal effects and odds to draw similar conclusions. However, we cannot observe differences at the observation level

Tabla 3.7: Explainability and inference logistic regression

Variables	Coefficients		Marginal effects		Odds	
	LR	LR-BS	LR	LR-BS	LR	LR-BS
Intercept	-0.741	0.450			0.477	1.568
<i>fico_n</i>	-0.013***	-0.013***	-0.002***	-0.003***	0.987	0.987
<i>loan_amnt</i>	4.08E-05***	4.00E-05***	5.61E-05***	1.00E-05***	1.000	1.000
<i>revenue</i>	-4.94E-06***	-4.24E-06***	-6.80E-06***	-1.06E-05***	0.999	0.999
<i>dti_n</i>	0.024***	0.024***	0.003***	0.006***	1.025	1.025
(<i>emp_length</i>)1 year	-0.027	-0.031**	-0.004	-0.008**	0.974	0.969
(<i>emp_length</i>)10+ years	-0.056***	-0.062***	-0.008***	-0.016***	0.945	0.940
(<i>emp_length</i>)2 years	-0.058***	-0.061***	-0.008***	-0.015***	0.944	0.941
(<i>emp_length</i>)3 years	-0.044***	-0.04***	-0.006**	-0.01**	0.957	0.961
(<i>emp_length</i>)4 years	-0.045**	-0.055***	-0.006**	-0.014***	0.956	0.947
(<i>emp_length</i>)5 years	-0.047***	-0.039***	-0.006**	-0.010**	0.954	0.962
(<i>emp_length</i>)6 years	-0.012	-0.023	-0.002	-0.006	0.988	0.977
(<i>emp_length</i>)7 years	-0.009	-0.010	-0.001	-0.002	0.991	0.990
(<i>emp_length</i>)8 years	0.019	0.010	0.003	0.003	1.019	1.010
(<i>emp_length</i>)9 years	0.014	0.028*	0.002	0.007	1.014	1.028
(<i>emp_length</i>)NI	0.261***	0.281***	0.039***	0.070***	1.298	1.324
(<i>purpose</i>)credit_card	-0.246***	-0.266***	-0.032***	-0.066***	0.782	0.767
(<i>purpose</i>)debt_consolidation	0.023	0.007	0.003	0.002	1.023	1.007
(<i>purpose</i>)educational	0.624***	0.767***	0.104***	0.183***	1.866	2.154
(<i>purpose</i>)home_improvement	0.164***	0.152***	0.024***	0.038***	1.178	1.164
(<i>purpose</i>)house	0.274***	0.217***	0.041***	0.054***	1.315	1.243
(<i>purpose</i>)major_purchase	0.153***	0.138***	0.022***	0.034***	1.165	1.148
(<i>purpose</i>)medical	0.315***	0.319***	0.048***	0.079***	1.370	1.376
(<i>purpose</i>)moving	0.38***	0.374***	0.059***	0.093***	1.463	1.454
(<i>purpose</i>)other	0.267***	0.257***	0.04***	0.064***	1.306	1.293
(<i>purpose</i>)renewable_energy	0.422***	0.487***	0.067**	0.120***	1.525	1.628
(<i>purpose</i>)small_business	0.785***	0.848***	0.136***	0.201***	2.193	2.334
(<i>purpose</i>)vacation	0.213***	0.187***	0.031***	0.047***	1.238	1.206
(<i>purpose</i>)wedding	-0.144*	-0.181***	-0.019**	-0.045***	0.866	0.835
(<i>home_ownership</i>)MORTGAGE	7.107	7.194	0.870	0.946	1.22E03	1.33E03
(<i>home_ownership</i>)NONE	7.188	7.120	0.831***	0.501***	1.32E03	1.24E03
(<i>home_ownership</i>)OTHER	7.605	7.971	0.833***	0.502***	2.01E03	2.89E03
(<i>home_ownership</i>)OWN	7.215	7.316	0.902**	0.668	1.36E03	1.50E03
(<i>home_ownership</i>)RENT	7.370	7.470	0.930	0.947	1.59E03	1.76E03
<i>experience_c</i>	0.010	0.242	0.001	0.060	1.010	1.274
(<i>addr_state</i>)AL	0.319***	0.335***	0.049***	0.083***	1.376	1.398
(<i>addr_state</i>)AR	0.303***	0.368***	0.046***	0.091***	1.353	1.445
(<i>addr_state</i>)AZ	0.100	0.114**	0.014	0.028**	1.105	1.120
(<i>addr_state</i>)CA	0.065	0.034	0.009	0.009	1.067	1.035
(<i>addr_state</i>)CO	-0.160**	-0.176***	-0.021**	-0.044***	0.852	0.838
(<i>addr_state</i>)CT	0.051	0.055	0.007	0.014	1.052	1.056
(<i>addr_state</i>)DC	-0.322***	-0.299***	-0.04***	-0.074***	0.725	0.742
(<i>addr_state</i>)DE	0.155*	0.082	0.022	0.020	1.167	1.085
(<i>addr_state</i>)FL	0.180***	0.160***	0.026**	0.040**	1.197	1.173
(<i>addr_state</i>)GA	0.024	-0.005	0.003	-0.001	1.025	0.995
(<i>addr_state</i>)HI	0.014	0.004	0.002	0.001	1.014	1.004
(<i>addr_state</i>)IA	0.376	0.951*	0.059	0.222**	1.457	2.589
(<i>addr_state</i>)ID	-0.395	-1.283	-0.047	-0.282	0.674	0.277
(<i>addr_state</i>)IL	-0.018	-0.04	-0.003	-0.01	0.982	0.961
(<i>addr_state</i>)IN	0.210***	0.192***	0.031**	0.048***	1.234	1.211
(<i>addr_state</i>)KS	-0.105	-0.117**	-0.014	-0.029**	0.900	0.890
(<i>addr_state</i>)KY	0.202***	0.209***	0.03**	0.052***	1.224	1.232
(<i>addr_state</i>)LA	0.284***	0.274***	0.043***	0.068***	1.328	1.315
(<i>addr_state</i>)MA	0.096	0.070	0.014	0.018	1.101	1.073
(<i>addr_state</i>)MD	0.184***	0.146***	0.027**	0.036**	1.202	1.157
(<i>addr_state</i>)ME	-6.762	-7.100	-0.165***	-0.497***	0.001	0.001
(<i>addr_state</i>)MI	0.122*	0.118**	0.017	0.029**	1.130	1.125
(<i>addr_state</i>)MN	0.131*	0.102*	0.019	0.025	1.140	1.107
(<i>addr_state</i>)MO	0.211***	0.198***	0.031**	0.049***	1.234	1.219
(<i>addr_state</i>)MS	0.356***	0.347***	0.055***	0.086***	1.428	1.415
(<i>addr_state</i>)MT	-0.069	-0.137**	-0.009	-0.034**	0.933	0.872
(<i>addr_state</i>)NC	0.170**	0.172***	0.025**	0.043***	1.186	1.188
(<i>addr_state</i>)ND	0.960	0.835	0.175	0.198	2.612	2.305
(<i>addr_state</i>)NE	0.903***	0.744***	0.163**	0.178***	2.467	2.103
(<i>addr_state</i>)NH	-0.268***	-0.267***	-0.034***	-0.066***	0.765	0.765
(<i>addr_state</i>)NJ	0.193***	0.175***	0.028**	0.044***	1.213	1.191
(<i>addr_state</i>)NM	0.183**	0.131**	0.027**	0.033**	1.200	1.140
(<i>addr_state</i>)NV	0.249***	0.257***	0.037***	0.064***	1.282	1.293
(<i>addr_state</i>)NY	0.228***	0.218***	0.033***	0.054***	1.256	1.244
(<i>addr_state</i>)OH	0.181***	0.149***	0.026**	0.037**	1.198	1.161
(<i>addr_state</i>)OK	0.300***	0.321***	0.045***	0.08***	1.350	1.379
(<i>addr_state</i>)OR	-0.203***	-0.191***	-0.026**	-0.048***	0.816	0.826
(<i>addr_state</i>)PA	0.179***	0.148***	0.026**	0.037**	1.196	1.159
(<i>addr_state</i>)RI	0.061	0.018	0.009	0.004	1.063	1.018
(<i>addr_state</i>)SC	-0.097	-0.142**	-0.013	-0.036**	0.908	0.867
(<i>addr_state</i>)SD	0.168*	0.162**	0.024	0.04**	1.183	1.176
(<i>addr_state</i>)TN	0.233***	0.232***	0.034**	0.058***	1.262	1.261
(<i>addr_state</i>)TX	0.040	0.031	0.006	0.008	1.041	1.031
(<i>addr_state</i>)UT	0.028	0.054	0.004	0.013	1.028	1.055
(<i>addr_state</i>)VA	0.144**	0.126**	0.021**	0.032**	1.155	1.135
(<i>addr_state</i>)VT	-0.307***	-0.358***	-0.038**	-0.089***	0.736	0.699
(<i>addr_state</i>)WA	-0.145**	-0.157***	-0.019**	-0.039**	0.865	0.855
(<i>addr_state</i>)WI	0.015	0.024	0.002	0.006	1.015	1.024
(<i>addr_state</i>)WV	-0.099	-0.051	-0.013	-0.013	0.906	0.950
(<i>addr_state</i>)WY	-0.169*	-0.120	-0.022	-0.030	0.844	0.887

***significant at the 0.01 level, **significant at the 0.05 level and *significant at the 0.1.

because values are aggregated for the whole model. According to Table 3.7, which reports the logistic regression models, *fico_n* and *revenue* have negative coefficients and odds below 1, which represents an inverse relationship with default probability. The converse is true for *dti_n* and *loan_amnt*.

Regarding categorical variables, we analyze interesting values of some of them. For example, the coefficients of the *purpose* of credit card are -0.246 and -0.26 in LR and LR.BS, respectively. Their average marginal effects are at levels of -0.03 and -0.06, and their odds values are 0.78 and 0.76. Thus, both are associated with lower risk, compared to the base category.

In the case of having as a purpose, *purpose*, investment in small businesses, the coefficients are 0.78 and 0.85, there are significant average marginal effects of 0.14 and 0.20, and there are high odds values of 2.19 and 2.33. These values ratify the increase in the default risk when having this credit purpose. However, the *purpose* of debt consolidation, although it presents positive coefficients and odds higher than 1, is not considered significant. According to Figure 3.3d, the results found through the logistic regression for this variable are similar to those of XGboost.

Regarding the variable *emp_length*, the coefficients of having employment for more than ten years are negative in the logistic regression proposals, -0.056 and -0.062, with odds less than 1, which are significant, suggesting lower risk against the base case of having *emp_length* less than one year of employment. However, in XGBoost, the relationship is not so clear. Not having information, NI, is a risk factor, with odds higher than 1. In this case, this conclusion can also be drawn from the Shapley values of XGBoost in Figure 3.3d.

As mentioned earlier, previous credit experience, *experience_c*, was not important in the machine learning proposal or in the logistic regression models. The statistical significance analysis of the logistic regression models confirms its low predictive power (see Table 3.7), which can be evidenced in the statistical analysis of coefficients, marginal effects and odds ratios. In the XGBoost proposal (see Figure 3.4c), this variable has a relative contribution

depending on the Shapley values, which are close to zero.

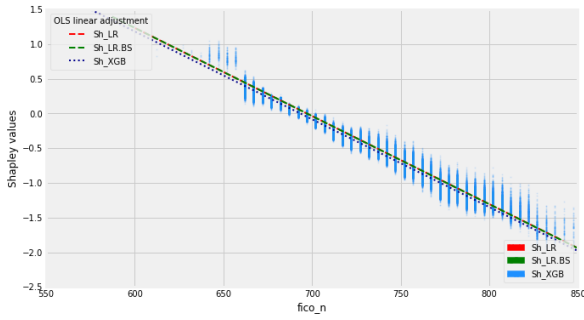
As a result, the conclusions in terms of feature importance and relationship with the target variable that can be drawn from XGBoost with the help of the Shapley values are mostly similar to those derived from the logistic regression models. However, Shapley values in nonlinear models such as XGBoost can help us discover more complex relationships than those that can be inferred with linear models such as logistic regression. This is particularly true for the dependence plots, as we will see in the next section.

Dependence plots for quantitative variables

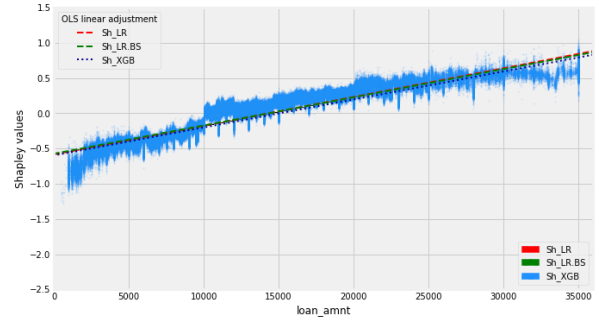
Figure 3.5 shows the dependence plots for the quantitative variables according to the Shapley values. In these plots, we include the lines that represent the dependence for the logistic regression models. However, it can be observed that machine learning alternatives represent dependence in a more comprehensive manner, including nonlinearity and heteroscedasticity.

For example, in the FICO variable (*fico_n*) in Figure 3.5a, XGBoost finds a decreasing aggregate monotonic relationship, similar to the logistic regression alternatives, but also identifies structural changes for the different values of this variable and different variability across the range of the variable. Values less than or equal to 660 points in FICO scoring (*fico_n*) show higher levels of risks than those predicted by the general trend. Score values between 660 and 715 show little dispersion and little contribution to the prediction, so other features probably come into play to determine risk. However, from 715 to 800, another segment is observed, with the default risk decreasing as the score increases. The segment is homoscedastic although with greater dispersion than that in other segments. Two subtle structural changes of the general trend can be seen at 800 and 825 points. Although scores in these segments decrease the risk of default, in each of the segments, the precise score observed does not seem to affect the impact, i.e., increasing the score in each segment does not decrease the default risk.

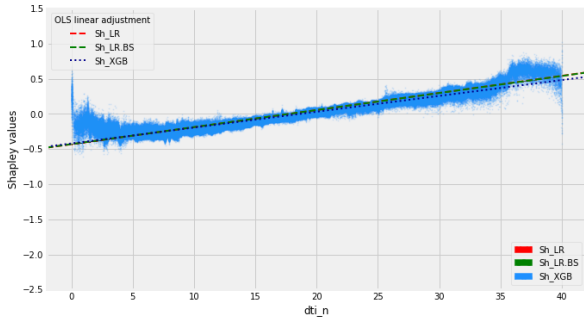
In Figure 3.5b, we can see the dependence plot for the *loan_amnt* variable, which also



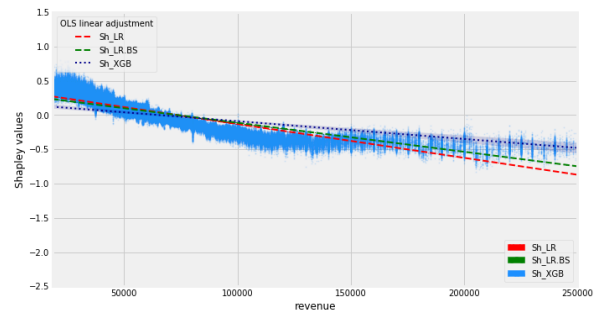
(a) *fico_n* dependence graph



(b) *loan_amnt* dependence graph



(c) *dti_n* dependence graph



(d) *revenue* dependence graph

Figure 3.5: Dependence graphs of quantitative variables.

evidences structural changes in the impact on the dependent variable. For example, a clear break can be observed at 10,000 USD, where changes in the level and the trend can be observed. The level change denotes an overall increase in the risk profile, even if the trend moderates the slope. Such changes may be associated with entity policies not mentioned in the data set information. In general terms, considering the whole range, the risk increases as the loan amount increases, and a saturation effect can be observed for values greater than approximately 275,000. Again the linear trends inferred by the logistic regression models

hide subtle but important aspects.

Figure 3.5c shows that variable *dti_n* also exhibits nonlinear behavior. Nonlinearity and heteroscedasticity can be observed in the extremes of the range.

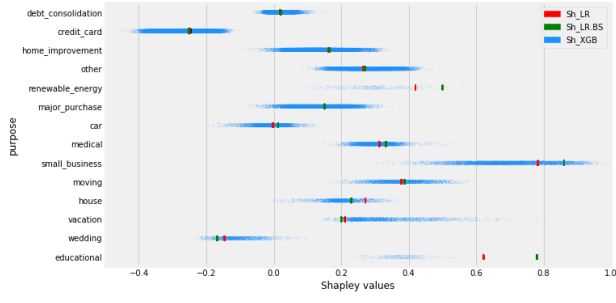
Finally, the *revenue* variable shown in Figure 3.5d exhibits a sophisticated behavior. First, the trend from the aggregated impact that can be estimated for XGBoost is notably different than those from the logistic regression alternatives, although all three are negative. If we look at the disaggregated impact for XGBoost, we can observe that for values below 110,000 USD, as *revenue* increases the risk of default is reduced. However, for *revenue* values over 110,000 USD, the observed slope is zero, even if some strange phenomena can be observed. This fact evidences the need to complement the risk analysis with other variables. As already mentioned, this change in the trend can hardly be noticed with the logistic regression.

Dependence plots for categorical variables

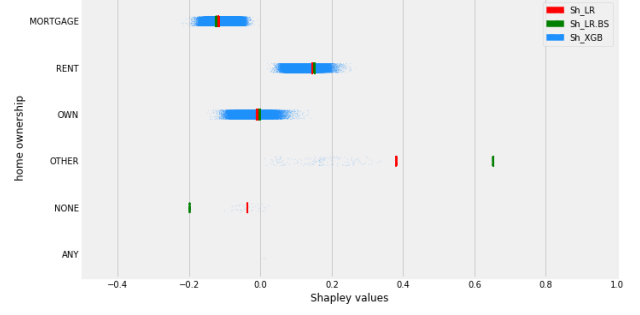
According to the estimated feature importance (see Figure 3.4), the categorical variables have less impact on the dependent variable than the quantitative ones. However, some specific categories contribute in an important way to the prediction, as shown in in Figure 3.3, for example, the value credit card for the variable *purpose* or the values MORTGAGE or RENT for *home_ownership*. In this section, we examine in detail the categorical values with the help of dependence plots. We show the Shapley values of each category recalculated to aggregate the effect of the nonpresence of the rest of the categories, as shown in Equation 3.4.

Figure 3.6a shows the impact of the categories of the *purpose* variable. Interestingly, some values such as debt consolidation and car, have almost no impact on the risk of default. However, credit card option is the value that implies the lowest risk, while small businesses is the one that implies the highest.

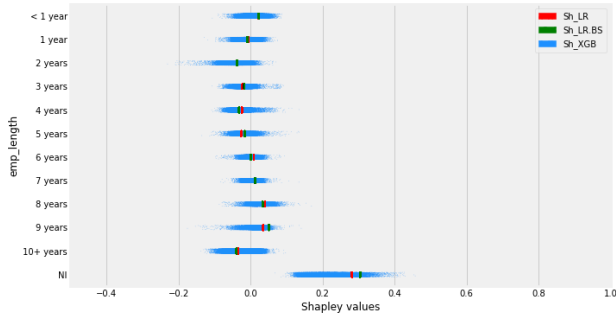
For the variable *home_ownership*, none of the categories have a great impact on the risk of default, as can be seen in Figure 3.6b. However, the impacts of MORTGAGE and RENT



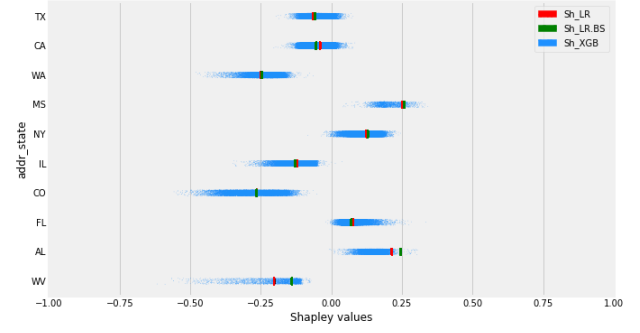
(a) *purpose* dependence graph



(b) *home_ownership* dependence graph



(c) *emp_length* dependence graph



(d) *addr_state* dependence graph. Some geographical states

Figure 3.6: Dependence graphs of categorical variables.

are not negligible, the first as an option that decreases the level of risk and the second the opposite. For the case of the *emp_length* variable, most categories have a small impact on the default risk, with the exception of the value of NI, which denotes not having information, which clearly increases the default risk level; see Figure 3.6c.

Finally, Figure 3.6d shows the dependence plot for some values of the *addr_state* variable. None of them have a great impact, especially TX (Texas) and CA (California). However, CO (Colorado) decreases the risk of default, while Mississippi (MS) increases it.

If we look at the Shapley values of the logistic regression derived from the calculation on the coefficients, we can see that they are within the range of the Shapley values obtained for the machine learning model XGBoost. Typically, these values are located in a central position. However, we can find categories for which the values are located at one of the extremes, which shows that logistic regression and XGBoost exploit the variables in different ways. See, for example, the value vacation for the variable *purpose*, where the logistic regression values are located in the lower bound of the XGBoost observed range, that is, XGBoost considers that this value can be associated with higher levels of risk than those predicted by logistic regression models. However, exactly the opposite occurs for the MS and AL values for the variable *addr_state*, where XGBoost generally considers these values less risky than the logistic regression models.

3.6. Conclusions

This article has shown that credit risk machine learning approaches may outperform statistical approaches, such as logistic regression, in terms of not only classification performance but also explainability.

The graphical representation of the SHAP values makes it possible to evidence aspects of the relationship between each feature variable and the target one that are overlooked by linear approaches. Such aspects include curved relationships, structural breaks, heteroscedasticity and outlying behavior.

Furthermore, in our approach we have adopted what we consider should be standard practice in machine learning credit models, namely:

- Hyperparameter optimization to find suitable configurations of the algorithms.
- Explicitly dealing with the imbalanced nature of the data (either by a resampling strategy or by setting weighting schemes in the machine learning method).
- Comparing classifier performance using statistical inference.
- Assessing the classification performance for each class.

Our article is relevant for credit risk modeling in general, where reliable and transparent models are required by the regulators and industry. Furthermore, it is timely given the growing interest in explainable machine learning models for credit risk that we have addressed in our review of the literature. The good performance of the tree boosting classifier is in line with other results in the literature ([Li et al., 2018a](#); [Munkhdalai et al., 2019](#); [Xia et al., 2017b](#)), but our article shows that the better results come from a better description of the relationships among the variables.

In this regard, it is worth mentioning that some authors are proposing new methods to estimate the Shapley values to better account for dependence ([Aas et al., 2019](#)). Our article has shown how to adjust the SHAP values for the categories of categorical variables to better account for the dependence among categories. Improvement of the theory of the Shapley values to better account for dependence and include other aspects, such causality and inferential tools ([Joseph, 2019](#)), will lead to a wider adoption of machine learning models in credit risk modeling and other domains.

Capítulo 4

Artículo. Profit-sensitive machine learning classification with explanations in credit risk: The case of small businesses in peer-to-peer lending

Este trabajo se encuentra bajo revisión en la revista indexada en Web of Science con Journal Citation Reports (JCR) Electronic Commerce Research and Applications.

Tiene como coautores a Javier Arroyo, María-Jesús Segovia y Antonio Caparrini.

Este trabajo ha sido apoyado por los proyectos COST (European Cooperation in Science and Technology) con referencia COST Action 19130 y por la subvención del Ministerio Español de Ciencia e Innovación de referencia PID2020- 115700RB-I00.

Abstract

We propose a profit-sensitive approach for credit risk modelling in p2p lending for small businesses, one of the most financially complex segments. In particular, we go beyond traditional and cost-sensitive approaches by including both the financial costs and profits. We introduce the profit information at three points of the modeling: the estimation of the learning function of the classification algorithm, the hyperparameter optimization, and the decision function, and we apply it to the popular XGBoost algorithm. Additionally, we use explainability tools to study the model changes between the profit-sensitive and profit-insensitive approaches. Our approach can obtain higher profitability than profit-insensitive approaches by reducing the acceptance rate. The explanatory tools reveal the key features of the granted loans. Our proposal can be extended to other loan markets or to other classification problems as long as the cells of the misclassification matrix can have an economic value.

Index terms— Credit risk, P2P lending, Small business loans, Cost-sensitive models, Profit-sensitive learning, Extreme gradient boosting, Explainability, Shapley values

4.1. Introduction

Recently, financing and investment alternatives have emerged, leveraged by technological development and the latest financial crises. One of these alternatives is peer-to-peer or P2P lending, where lenders and borrowers can connect directly without standard intermediaries such as banks or traditional lending companies. Consequently, credit consumers have a new way to borrow, and investors make money by lending it.

In P2P lending, technological platforms act as intermediaries that obtain money for fees from borrowers and investors for connecting them, managing information and providing the means to carry out the lending operation. This results in a more streamlined process involving less paperwork, reducing intermediation's burden and operational costs [Guo et al. \(2016\)](#). Also, borrowers can obtain credit at lower interest rates than traditional credit alternatives, and investors receive profits that typically equal or exceed other investment options ([Emekter et al., 2015](#)). In turn, P2P lending gives credit access to people typically excluded from traditional products and financial services ([Maskara et al., 2021](#)). In the words of [Milne & Parboteeah \(2016\)](#), P2P is considered a complementary and non-competitive market with conventional banking.

Nevertheless, P2P lending comes with some critical risks mainly derived from a more intensified information asymmetry ([Serrano-Cinca & Gutierrez-Nieto, 2016](#)), a moral hazard that must be managed ([Cummins et al., 2019](#)). The problem is intensified because, in most platforms, the lenders are the ones that bear this credit risk ([Serrano-Cinca & Gutierrez-Nieto, 2016](#)). Furthermore, P2P lending is affected by events that systemically affect the financial markets. A relevant example could be the COVID-19 pandemic in 2020. Although in the beginning, it brought an increase in default rates, bringing a reduction in profits or losses for lenders, substantially reducing supply, in the last year, it has experienced a recovery that has positioned the P2P market as one of the lending and investment alternatives. However, such positioning is associated with risk factors such as falling income, rising unemployment and the tightening of lending requirements by traditional entities, demanding, even more,

a definition of profitable strategies that adjust to the risk experienced by investors in the P2P market ([Invest In Club, 2021](#)).

Due to these risks, P2P lending platforms offer risk measurement services, including advisory and accompaniment for lenders and borrowers, diversification of loan portfolios, education and training services, building payment agreements, improving identity verification processes, and, as expected, increasing interest rates on new loans. However, other platforms move away from these purposes and focus on attracting borrowers by adjusting only to interest rates, encouraging the granting of risky and expensive loans, and increasing their probability of default, leading to their collapse and failure ([Klein et al., 2023](#)). As a result, robust risk management is needed in the P2P market. As said by [Gao et al. \(2020\)](#), P2P lending still presents a traditional financial problem of non-payment that requires proper credit and financial risk management; and as stated by regulators, it also requires novel, transparent and tailored regulation policies ([ROFIEG, Expert Group on Regulatory Obstacles to Financial Innovation, 2019](#)).

In addition, the P2P market is an alternative for a niche needing funding sources, such as small businesses. Small businesses have more difficulty accessing financing alternatives than larger businesses ([Mach et al., 2014](#); [Sharma et al., 2023](#)). Reducing the access gap between the demand and supply of financial products and services is part of the Sustainable Development Goals (SDGs) of the United Nations ([Carpentier & Braun, 2020](#); [Liguori & Bendickson, 2020](#)). At the same time, the small business niche is a high-risk segment due to its opacity of information ([Mills & McCarthy, 2016](#)) that usually attracts higher-risk borrowers without access to traditional sources ([Nowak et al., 2018](#)). The aforementioned makes conventional risk models unsuitable ([Chang et al., 2022](#)).

To address all these needs, new approaches to credit risk model management are emerging and among them, those that use machine learning stand out. According to [Ariza-Garzón et al. \(2021\)](#), machine learning methods are used for default prediction in P2P lending, offering better predictive performance than traditional econometric or statistical methods.

For example, Support Vector Machines (Bastani et al., 2019; Rodrigues et al., 2018; Xia, 2019), classic artificial neural networks (Byanjankar et al., 2015; Lyócsa et al., 2022; Zang et al., 2015; Zhao, 2015), or deep learning (Duan, 2019; Fu et al., 2019; Kim & Cho, 2019c; Li et al., 2020; Van-Sang et al., 2019). Regarding tree-based models, the classic decision trees are still used (Misheva et al., 2018; Serrano-Cinca & Gutierrez-Nieto, 2016), but ensemble approaches are gaining attention (Jiang et al., 2018; Ma et al., 2018a). According to (Ariza-Garzón et al., 2021), those derived from homogeneous boosting techniques have gained attention in recent years, and especially XGBoost Chen & Guestrin (2016), which has shown a better performance than other alternatives (Ariza-Garzón et al., 2020; Boiko Ferreira et al., 2017; Chang et al., 2022; Li et al., 2018a; Xia et al., 2017a). Thus, we will base our proposal on XGBoost, given its popularity and good performance.

Besides the use of machine learning, new credit risk management also includes the use of cost-sensitive models. These models differently weigh the possible outcomes of the prediction (true positives, false positives, etc.) by assigning a cost to each outcome. In doing so, they not only deal with the class imbalance problem, which is crucial in credit risk where the default class is a minority, but they also serve to incorporate a business sense into the classification methods. As a result, cost-sensitive models provide credit-granting decisions based on the classification costs, facilitating their use in real life. In particular, we found a few cost-sensitive alternatives for the P2P lending market (Boiko Ferreira et al., 2017; Cho et al., 2019; Malekipirbazari & Aksakalli, 2015; Wang et al., 2018a; Xia et al., 2017a; Ye et al., 2018). Although these models are often referred to as cost-sensitive alternatives and, in most cases, only include misclassification costs, we would like to highlight the inclusion of profits, making a more business-sensitive proposition. In any case, costs can be viewed as negative profits, and our proposal will incorporate the associated profits to deal with the class imbalance problem and have a business-oriented credit-granting method. In this way, we will include the financial costs (-profits) of a bad classification and the profits of a correct default prediction. To this approach, we will call the profit-sensitive model.

More specifically, in our proposal, we will indirectly and directly explore the inclusion of profits (including costs). In the first case, we incorporate the profit as an expected value function to be maximized in the hyperparameter optimization of the machine learning algorithm. In the direct algorithm, we include them in the XGBoost learning through the gradient and the Hessian of an objective loss function that, in our case, corresponds to the additive inverse of the expected profit function. We will evaluate both alternatives using a measure sensitive to the expected profit, a probabilistic profit-sensitive measure.

Furthermore, we will also analyze the performance, considering the profit results once the prediction is made. In other words, the credit-granting decision for each obligation will be based on the probability learned by the classification algorithm and a decision method. To do so, we will explore using several decision methods to define the prediction. The decision methods are the search of the optimal decision threshold ([Herasymovych et al., 2019](#); [Verbraken et al., 2014](#)) and the binary Bayes Minimum Risk classifier (BMR) ([Bahnsen et al., 2014](#)) on the estimated probability.

Finally, using powerful machine learning modelling alternatives, even incorporating a business sense, is insufficient for credit granting. There is also a significant demand for explanatory and interpretative models by the industry because they bring transparency and control, which is required by regulators ([ROFIEG, Expert Group on Regulatory Obstacles to Financial Innovation, 2019](#)) and is also welcome by users. As noted in the literature review on P2P lending by [Ariza-Garzón et al. \(2021\)](#), few machine learning approaches provide interpretability or explainability.

One recently developed tool that helps provide transparency and explainability to machine learning classification models is the SHAP values (SHapley Additive exPlanations) ([Lundberg et al., 2020](#); [Lundberg & Lee, 2017](#)). They provide explanatory capacity to complex models of diverse nature at both local and global levels. In the case of credit risk models, they help to determine how risk factors participate individually by obligation and at an aggregated level in the models and help to explain the models ([Ariza-Garzón et al.,](#)

2020; Bussmann et al., 2020). We will extend the SHAP values to be used in a profit-sensitive classification context and compare them with the traditional SHAP values from the classification algorithm without considering the profits. In this way, we aim to understand what factors drive profitable loans and if they differ from those driving the traditional classification problem in credit risk.

This work will propose and validate a credit risk model based on a state-of-the-art machine learning approach, including profit-sensitive learning, going beyond the proposed cost-sensitive models. In particular, we will evaluate the effect of including profit information in the XGB algorithm and interpret the resulting model using explainability tools. For the empirical evaluation of our proposal, we will use the dataset from Lending Club, which has been extensively used in other works in recent years (Bastani et al., 2019; Boiko Ferreira et al., 2017; Calabrese et al., 2019; Cho et al., 2019; Duan, 2019; Durovic, 2017; Xia et al., 2017a). We will focus on the segment of small businesses, which is more challenging and understudied in the literature (Nowak et al., 2018).

The rest of this document is organized as follows. Section 4.2 introduces the family of cost-sensitive measures and models and reviews the main works in P2P lending. Our proposal of profit-sensitive measures derived from the misclassification matrix is in Section 4.3, a profit-sensitive machine learning approach with explainability in Section 4.4, and experimental setting in Section 4.5. Section 4.6 presents the empirical results of the analyses, including explainability elements. Section 4.7 summarises the main findings.

4.2. Cost-sensitive methods in loan evaluation

4.2.1. Classification errors and cost-sensitive methods

Default probability estimation models often focus on performance measures based on the binary classification of successes and failures. In doing so, they wrongly assume that misclassification costs have the same weight or value and disregard the class imbalance problem typical in credit risk. Although performance measures, such as Balanced Accuracy (BACC)

or the F1 score, mitigate the class imbalance, they neglect the economic or financial effect that miss-classification may generate. For example, investing in a loan that falls through can lead to both loan principal and interest losses, which is different from the value of the lost opportunity to invest in a loan that will be repaid (Bahnsen et al., 2015).

In P2P lending, it is essential to incorporate the financial effects of misclassification errors because this market is exposed to higher levels of risk than traditional businesses. Lenders assume Such risk in a more important proportion, causing market instability. Nevertheless, according to Ariza-Garzón et al. (2021), most models for the P2P market focus on estimating the probability of default (PD) or predicting default, and they are still evaluated with traditional performance measures that are either unaware of or insensitive to the costs of miss-classification. Thus, although risk problems remain latent, it is clear that a business-oriented definition of the loss function and its incorporation into the training and evaluation processes aligns the modelling with economic objectives, allowing for better risk management and contributing to business stability.

Standard machine learning and statistical methods typically use the so-called 0-1 loss function, assigning 0 to a correctly classified instance and 1 to an incorrect one. Since the 0-1 loss function uses the same cost associated with a miss-classification for all classes considered, it is highly susceptible to skewed class distributions (Fernández et al., 2018) and a non-real business sense valuation of the associated costs.

In credit risk modelling, we typically have a binary classification with two categories, $y_i = 1$ Default, and $y_i = 0$ Non-default. In a cost-sensitive approach, the 2×2 confusion matrix summarizes the costs of each classification decision for each instance. The matrix entries are given by the costs associated with miss-classification $C(k, l)$ for each instance, where k represents the predicted label $\hat{y}_i$ and l the actual label y_i . Thus, let $C(k, l)$ denote the miss-classification costs of classifying a record with real label $y_i = l$ predicted as $\hat{y}_i = k$, with $k, l = 0, 1$.

In a typical cost-sensitive approach, the only cost considered is the cost of a false negative,

$C(0,1)$, which is the cost of classifying a bad loan as non-default because it represents a financial cost for the company. On the other hand, $C(1,0) = C(0,0) = C(1,1) = 0$, as these outcomes do not suppose a cost strictly, even if they impact the financial result. However, in standard machine learning and statistical approaches, the loss function typically considers $C(1,1) = C(0,0) = 0$ and $C(1,0) = C(0,1) = 1$.

From a business perspective, the resulting model has a greater business sense if it includes other costs, such as opportunity costs and even profits. In this case, the resulting approaches are typically called profit-sensitive. Specifically, $C(1,0)$ is the cost of miss-classifying a good loan as a default (false positive, FP), and $C(0,1)$ is the cost of classifying a bad loan as non-default (false negative, FN). $C(1,1)$ corresponds to the profit of achieving a true positive (TP), that is, correctly predicting a default, and $C(0,0)$ the profit of obtaining a true negative (TN), that is, correctly predicting a good loan.

If we use costs and profits to measure the performance of a model, it is also natural to include such financial values to learn how to classify with a business sense. The set of techniques that use this approach is known as cost-sensitive methods (or profit-sensitive methods if they include profits). While those that assume the same miss-classification costs for all cases are known as cost-insensitive methods (Ling & Sheng, 2010).

In the literature (Fernández et al., 2018; Ling & Sheng, 2010), cost-sensitive methods can be classified into direct and indirect. The direct approach introduces the miss-classification costs into a classifier’s training and learning procedure. The indirect approach (Xia et al., 2017a), also known as meta-learning methods, where cost-insensitive classifiers become cost-sensitive by preprocessing (e.g., sampling or resampling) the training data or by postprocessing by modifying the outputs of a classifier during the classification phase (e.g. by changing the threshold or cutoff in a probabilistic model to classify as default or non-default category).

Other taxonomies of algorithms dealing with the class imbalance and Cost-Sensitive Methods are proposed by He & Garcia (2009) and Petrides & Verbeke (2022). In the first

work, the authors consider a category where methods incorporate cost-sensitive functions directly into the classification algorithms. Another category includes an expected cost function in the decision criteria for the best modelling alternative. The second work categorizes the cost-sensitive models considering the misclassification costs and defines categories of pre-training, training and post-training methods. It includes a category based on sampling in the pre-training stage, where the over-sample or under-sample techniques are determined by the ratio between false negative and false positive costs and is called sampling-based. In our work, we include this category in the indirect methods.

Also, (Petrides & Verbeke, 2022) draws attention to a category where records are weighted by their costs and used with any algorithm. For us, it corresponds with a direct method. These costs can also be used in ensemble methods, which the authors call cost-sensitive ensemble methods. They also mention the Cost-Sensitive Post-training methods, which include the definition of a threshold for the probability, which is estimated based on a process of minimization of an expected function of classification costs, which we call and use as an indirect method.

4.2.2. Review of cost-sensitive and profit-sensitive methods for loan evaluation

As mentioned by Xia et al. (2017a) and later corroborated by (Ariza-Garzón et al., 2021), cost-sensitive loan evaluation models have received little attention in the P2P lending context and even in traditional loan evaluation, and when used, most of them are indirect methods. However, if we broaden our scope and look at the financial field, we can find cost-sensitive methods in credit card fraud (Akila & Reddy, 2018; Almhaithawi et al., 2020; Correa Bahnsen et al.; Nami & Shajari, 2018), credit card business (Zhang et al., 2020) or retail credit (Oreski & Oreski, 2018).

In this section, we will review the most relevant cost-sensitive approaches related to loan evaluation proposed recently, complementing the review by Xia et al. (2017a) for the

previous years. We will focus on direct cost-sensitive approaches and indirect ones based on postprocessing. We will not delve into proposals that have focused on indirect methods by preprocessing, considered by [Bastani et al. \(2019\)](#), [Duan \(2019\)](#), or [Zhu et al. \(2019\)](#), because our proposal does not consider such methods.

For each approach, we will analyze the objective function and the definition of the costs and profits in the confusion matrix, and we will also note how they measure performance. We will not review the works by [Cho et al. \(2019\)](#) or [Calabrese et al. \(2019\)](#), because, while they deal with the class imbalance problem, they do not use a function that depends on the confusion matrix.

Finally, it is important to remark that most of the reviewed methods are applied in P2P lending. The proposals rarely come from traditional markets, probably due to the increasing interest in P2P lending, its need for accurate risk assessment, and data availability.

Regarding indirect methods, [Malekipirbazari & Aksakalli \(2015\)](#) consider a random forest (RF) as a base model and incorporates a cost-weighted matrix that increases the cost of miss-classification associated with bad borrowers through a cost-sensitive function using WEKA. The study includes accuracy (ACC) and the area under the curve (AUC), among other traditional measures for performance evaluation.

[Wang et al. \(2018a\)](#) compare the incorporation of different non-financial cost matrices indirectly on different classification methods such as C4.5 Decision Tree (DT C4.5), Naive Bayes (NB), Logistic regression (LR) and Support Vector Machine (SVM). For the comparison, they use non-financial cost measures plus other typical performance measures such as ACC.

[Herasymovych et al. \(2019\)](#) propose a dynamic reinforcement learning approach to optimize the acceptance threshold of a credit score. They maximize an expected profit function similar to the one proposed by [Verbraken et al. \(2014\)](#), which modifies the threshold of the classifiers to achieve incremental financial profit, taking into account Loss Given Default (LGD), the interest rate, and other financial components across the entire loan data set. In

particular, they develop a dynamic reinforcement learning system that adapts the threshold in response to the changing data distribution to maximize profits. From our point of view, they develop a dynamic approach of an indirect thresholding method. They use as an objective function an expected classification profit per loan with respect to a threshold value and assume $C(1, 0) = C(1, 1) = 0$. Their evaluation is performed using the convergence of the optimization method, along with AUC and total profits. This work incorporates the profits of each cell, aggregates them in the function and optimizes them considering a maximization problem; thus, it can be considered a profit-sensitive model.

[Wang et al. \(2021\)](#) estimate the different credit risk grades for a data set from Lending Club. They use a cost-sensitive matrix with financial elements like LGD, returns interest rate and the probability of default associated with the different risk grades and transitions. They apply an indirect cost-sensitive approach to DT C4.5, RF, LR and SVM methods. They use ACC and financial measures like total cost, average cost, and cost-saving rate (decreased cost after using cost-sensitive) to measure performance. Although the ACC is low in the cost-sensitive proposals, their results show that costs are reduced, obtaining a more financially efficient method. This work highlights the importance and benefits of using financial information to construct the cost matrix.

On the other hand, regarding direct methods, [Bahnsen et al. \(2015\)](#) develop a direct cost-sensitive decision tree that introduces cost-sensitive elements in the impurity measure and the pruning criterion. They apply their method in three contexts. First, in direct marketing for classifying customers more likely to accept an offer to open a bank long-term deposit account. Second, in credit card fraud, and lastly, the one with the most interest for this work, in loan evaluation. In particular, they assume $C(k, l) = 0$ when $k = l$ and incorporate financial elements such as LGD for the case of false negatives. They also consider the opportunity cost determined by the loss in profit (e.g. the interest rate) by rejecting what would have been a good loan, netting the expected profit that the lenders can obtain by using the money in other loans, for the case of false positives. The evaluation uses traditional

performance measures such as ACC or the F1-Score but includes a financial measure called cost savings. Their proposal is also compared with other alternatives containing indirect cost-sensitive methods such as BMR ([Bahnsen et al., 2014](#)), showing the profits of business-oriented decision-making. It is worth mentioning that, although its proposal includes the value of profits in some cells, it is still oriented towards cost minimization; thus, it still can be considered a cost-sensitive approach.

[Maldonado et al. \(2017\)](#) addresses the class imbalance problem but incorporates alternative information acquisition costs into the classification problem. They predict loan default using Support Vector Machines (SVM), including these costs in the objective function within the optimization process, and using constraints that aid in the feature selection process. The proposal can be considered direct since the costs are included in the algorithm optimization. However, it does not deal with the costs in the classification matrix.

[Xia et al. \(2017a\)](#) proposes a cost-sensitive XGBoost model. Their proposal mainly focuses on the costs $C(k, l)$ with $k \neq l$, assuming 0 when $k = l$. They compare their proposal against different indirect cost-sensitive alternatives using the AUC and the annualized return rate (ARR). Their model is used to construct Markowitz-type P2P lending investment portfolios, and they find the most profitable credits at the lowest risk. It is worth mentioning that although this is a direct cost-sensitive proposal based on the XGBoost model as ours, it does not incorporate financial costs and benefits into the misclassification matrix. In this proposal, the inclusion of cost information in the classification matrix is made by means of an alternative cost function whose gradient and Hessian are already developed. Specifically, they use the structure of a cost-insensitive loss function and convert it into a cost-sensitive one through one of the alternative functions developed by [Masnadi-Shirazi & Vasconcelos \(2011\)](#) called cost-sensitive logistic loss. The results show that a high accuracy based on the AUC measure does not necessarily bring high profits measured through ARR.

[Ye et al. \(2018\)](#) consider for each obligation not only the costs but also the profits derived from the confusion matrix results and incorporate them directly in the objective function.

Thus, they present a profit-sensitive model. Specifically, they use genetic algorithms to establish the optimal weights that maximize the resulting profit function in a Random Forest voting process. Having the profits and not only the costs as an objective function is one of the few cost-sensitive works incorporating information for $C(k, l)$, when $k = l$. The paper evaluates the performance of their proposal against several traditional methodologies using ACC and AUC. They also used as a performance measure and objective function, a profit function that mainly incorporates the principal and interest of each loan.

In turn, [Rao et al. \(2020\)](#) proposes a direct cost-sensitive RF model called Syncretic Cost-sensitive Random Forest (SCSRF). They consider several aspects in their cost function, such as the sampling distribution of the risk categories by feature and the sampling distribution of the features in the different random samples of the algorithm, and assume a factor for the class imbalance problem. However, none of these aspects is purely financial. In addition, within the random forest, they use a weighted voting system associated with the degree of accuracy of each decision tree. The evaluation of the proposal focuses on traditional performance measures like AUC and ACC, among others, and reveals that the SCSRF obtain better results than other RF cost-sensitive approaches and classical LR or DT methods.

[Pes & Lai \(2021\)](#) proposes direct and indirect cost-sensitive alternatives to deal with class imbalance. It compares different options for including a differential cost matrix by type of classification error in the learning algorithm. Also, they include these costs misclassification matrix on the probability threshold after a feature selection process. In the pre-modelling stage, the approach sensitizes the selection of variables to costs. As evaluation criteria, they use the MCC, G-mean and F1. In particular, the indirect approach with a previous feature selection process proves to be the alternative with the best performance, highlighting the importance of taking costs into account in class prediction.

Thus, of the above, few studies include information associated with the financial costs associated with default risk, either directly in the learning process or indirectly in post-processing alternatives. At the same time, proposals that include information on profits are

even scarcer.

4.2.3. Contextualizing our profit-sensitive proposal

In our work, we propose a profit-sensitive approach that considers the profits and losses of each cell of the confusion matrix and a profit maximization problem. Specifically, we incorporate expected profits from loans granted and losses from loans granted that eventually default. It is also possible to include opportunity costs, as [Bahnsen et al. \(2015\)](#) did; however, we consider that including these elements demands very coarse assumptions that we cannot verify with our data. These costs correspond to the losses associated with the profits from rejected good loans, including the expected profit lenders could have made by using the money in other loans. However, our proposal makes it possible to include all the elements of the confusion matrix if it is demanded.

We aim to recognize both profits and costs or losses realistically. Although we are not considering operational costs, the defined costs and profits represent the key elements of credit risk and aim to provide more realistic and integral lending decisions.

We will use as a base model the XGBoost algorithm, which is one of the most popular and powerful machine learning classification models [Ariza-Garzón et al. \(2020\)](#); [Chang et al. \(2022\)](#). As explained below, we will include the profit information, both directly and indirectly.

The direct approach includes an expected profit function in the learning process of the XGBoost algorithm. Our approach extends the one proposed by [Xia et al. \(2017a\)](#) for the same algorithm. We incorporate the profits and cost in the misclassification matrix and use the resulting matrix in the objective function to guide the learning of the XGBoost, while they incorporate the costs by means of a specific cost-sensitive function. Furthermore, we estimate the costs and profits in another way. Specifically, we include the actual financial costs and profits estimated from the available Lending Club data by obligation in the hits and misses of the misclassification matrix.

In the direct process, our proposal includes an Expected Maximum Profit (EMP) function based on the one proposed by Verbraken et al. (2014). As they did, we aim to optimize a profit function by considering the average loss-given default (LGD), the annualized return on investment (ROI) on the loans and other financial elements.

As a result, our algorithm provides a default probability. Then, we use it indirectly to define a decision indicator for granting new credits. For such purpose, we consider three alternatives: set the threshold as 0.5, use a thresholding process to optimize the threshold (Herasymovych et al., 2019; Verbraken et al., 2014), and a method that consists of a lending profit-sensitive decision-making method based on a binary Bayes Minimum Risk classifier (BMR) as proposed by Bahnsen et al. (2014); Bahnsen et al. (2015). Herasymovych et al. (2019) use a thresholding process dynamically. They develop a reinforcement learning system that constantly adapts the threshold in response to distributional changes in the data to maximize the company’s profits.

As a result, our proposal will provide decisions both directly and indirectly informed by financial results. We will evaluate the performance of our proposal with traditional measures and use profit measures based on the profits and costs of our confusion matrix.

4.3. A proposal of profit-sensitive measures

This section presents a profit-sensitive confusing matrix, specifying the loss and profits in each cell. Then, we introduce the profit functions that can be derived from it, which will be used in the modelling process or for evaluation purposes.

4.3.1. A profit-sensitive confusion matrix

Our profit-sensitive function includes the profit for each lending decision once the outcome is observed. Specifically, we include the profits and the losses derived from the main risk factors associated with the grant process. The confusion matrix entries will be denoted by $\pi(k, l)$ as a representation of the profits; regardless, they can be positive or negative. Table

4.1 shows the resulting confusion matrix.

Tabla 4.1: Confusion matrix with miss-classification costs and profits

	Actual positive $y_i = 1$	Actual negative $y_i = 0$
Predicted positive $\hat{y}_i = 1$	$\pi_{TP_i} = \pi(1, 1)$	$\pi_{FP_i} = \pi(1, 0)$
Predicted negative $\hat{y}_i = 0$	$\pi_{FN_i} = \pi(0, 1)$	$\pi_{TN_i} = \pi(0, 0)$

In particular, $\pi_{TP_i} = \pi(1, 1)$ represents the financial outcome that corresponds with a true positive (TP), that is, correctly predicting a default, which implies avoiding a financial loss. On its part, $\pi_{TN_i} = \pi(0, 0)$ corresponds with the profit of obtaining a true negative (TN), that is, correctly predicting a good loan which represents a successful investment decision. On the other side, $\pi_{FP_i} = \pi(1, 0)$ represents the opportunity cost of miss-classifying a good loan as a Default (false positive, FP), while $\pi_{FN_i} = \pi(0, 1)$ is the loss of classifying a bad loan as Non-default (false negative, FN).

Different approximations for estimating the financial values in the confusion matrix might exist. They mostly depend on the availability of information associated with the business or market analyzed in the data at hand. The more information you have about the financial implications of each obligation (that is, its actual costs, losses and profits), the more accurate the estimation of the values of each cell of the matrix and, consequently, the more business sense will have the granting model and the more helpful it will be to take business decisions.

Each platform or entity should be able to assess the values of the matrix and use profit-sensitive alternatives such as the ones we propose. Below, we will review how these values can be approached or have been approached in the literature and how we will represent them. A more detailed specification of these values in our paper will be given in the next section once we have introduced our data set and the information it includes.

For the case of a correctly predicted default, π_{TP_i} , we assume that the profit is zero, as

most works did in the past (Bahnsen et al., 2015; Herasymovych et al., 2019; Verbraken et al., 2014); i.e., there is no profit from the correct identification of a default obligation. Alternatively, Ye et al. (2018) considers that, in that case, the classifier avoids the loss of principal for the lender; thus, it treats this element as a potential return. Furthermore, it could also be considered that the money is reinvested, but this case would require making further assumptions on the expected loss or profit.

Regarding π_{FP_i} , it may represent what the entity stopped earning by not granting a loan that would have been returned, but it should take into account that this loss is reduced by what can be obtained in average profit by putting the money in other borrowers (Bahnsen et al., 2015). This approach requires an assumption to value this component adequately. These assumptions would remove us from a realistic estimate of our study problem. For this reason, we will consider it zero, as other studies have done (Herasymovych et al., 2019; Verbraken et al., 2014).

For π_{FN_i} , we can estimate the loss the lender can achieve by considering the expected loss from the obligation falling into default. For doing that, we can include components that estimate the risk of potential losses based on Basel guidelines, with critical risk parameters like the probability of default (PD), LGD (Bahnsen et al., 2015; Verbraken et al., 2014) and exposure at default (EAD), which are used to obtain Expected Loss as the product $PD \cdot LGD \cdot EAD$. The P2P lending market inherently has high PD and LGD relative to the traditional lending context due to information asymmetry, non-professional lenders and uncollateralized lending (Xia et al., 2017a). If we had the information available, we could even incorporate International Financial Reporting Standards (IFRS) in the profit and loss assessment, improving the estimate for this matrix cell.

Finally, for π_{TN_i} , we can associate this value with the expected annualized ROI of having placed an obligation correctly (Verbraken et al., 2014). For example, Ye et al. (2018) determine this element as the product of principal and interest rate.

4.3.2. Profit-sensitive measures

The profit-sensitive measures used in the modelling process, decision process and evaluation of the model must be consistent with the values defined in the confusion matrix. However, at some point, they could differ as the objective at each point or the information required demands other interests.

For example, if our classification method provides the default probability estimated by the models F_i , we could use a profit function incorporating the probabilistic information. While in other parts of the process, we need to use profit measures that reflect the impact of the classification decision ($\hat{y}$). Below we make this distinction for presenting a set of profit-sensitive measures based on the confusion matrix in Table 4.1.

Expected profit-sensitive measures

It is possible to propose probabilistic profit measures that measure the expected profits for the classification methods that yield a probability, such as the XGBoost. In particular, we can associate F_i , representing $F_i(x_i)$, with the adjusted distribution function associated with default event per loan, obtained as a function of the feature vector x_i of the i -th loan or credit obligation and being y_i its outcome (default or not). In particular, F_i could follow a sigmoid function as in logistic regression. In such case, we can define the expected profit (EP) function per loan application as in Eq. (4.1):

$$\pi_i^E(F_i) = y_i (F_i \pi_{TP_i} + (1 - F_i) \pi_{FN_i}) + (1 - y_i) (F_i \pi_{FP_i} + (1 - F_i) \pi_{TN_i}) \quad (4.1)$$

Using the previous definition, we can estimate the Expected Total Profit (ETP), similar to that proposed by Verbraken et al. (2014), as in Eq. (4.2):

$$\pi_T^E(F) = \sum_{i=1}^N \pi_i^E(F_i) = \sum_{i=1}^N (y_i (F_i \pi_{TP_i} + (1 - F_i) \pi_{FN_i}) + (1 - y_i) (F_i \pi_{FP_i} + (1 - F_i) \pi_{TN_i})) \quad (4.2)$$

It is more convenient to consider the expected average profit instead of the expected total profit as it has a clearer business interpretation. Thus, we propose in Eq. (4.3) the

average expected profit rate per obligation:

$$R\pi_T^E = \frac{\pi_T^E(F)}{N} \quad (4.3)$$

Profit-sensitive measures based on the predictions

For determining the predicted outcome $\hat{y}_i$, we will use some decision functions based on thresholding and Bayes Minimum Risk methods (BMR), as we will explain in the 4.4.2 section. Once we have predicted the outcome $\hat{y}_i$, we can define profit functions that measure the financial impact of the classification decisions. We can use such functions to assess the proposed methods' financial performance and guide their modelling process.

In particular, We can define the obligation profit as in Eq. (4.4).

$$\pi_i(\hat{y}_i) = y_i (\hat{y}_i \pi_{TP_i} + (1 - \hat{y}_i) \pi_{FN_i}) + (1 - y_i) (\hat{y}_i \pi_{FP_i} + (1 - \hat{y}_i) \pi_{TN_i}). \quad (4.4)$$

As for the case of the expected profit-sensitive measures, we can define a total profit function expressed in the currency used in the data set as in Eq. (4.5):

$$\pi_T = \sum_{i=1}^N \pi_i(\hat{y}_i). \quad (4.5)$$

Again, to provide a clearer business interpretation, we can define the average profit rate per obligation considering the prediction as in Eq. (4.6):

$$R\pi_T = \frac{\pi_T}{N} \quad (4.6)$$

4.4. A profit-sensitive machine learning approach with explanations

We identify three points where it is possible to incorporate the profit-sensitive confusion matrix into the modeling and the decision processes: model learning/estimation, model tuning, and decision process.

By the learning/estimation function, we refer to the one used by the method to model the relationship between the observed values and the output. By model tuning, we refer to the

process of hyperparameter configuration that optimizes a given function, but it could also encompass other tasks such as feature selection or feature engineering. Finally, by decision process, we refer to the fact that some machine learning and statistical methods yield a probability as an output, and such probability needs to be transformed into a classification decision; this process can be done with the help of a profit-sensitive decision method.

We will call a method direct profit-sensitive (or π -sensitive) if it includes a profit function in the machine learning method’s learning function or the statistical method’s estimation process. At the same time, we will refer to indirect profit-sensitive approaches as those that include a profit function for model tuning or for producing the model’s decision. A model that does not include a profit function at any point will be called profit-insensitive.

In addition, we will also use profit-sensitive measures to evaluate the resulting model and yield profit-sensitive model explanations. In the first case, we will assess how annualized profitable the resulting model is, while in the second one, we will determine the contribution of each variable to the profit result.

It is important to remark that all the profit-sensitive measures used throughout the process should be based on the same profit-sensitive confusion matrix. Still, the measures used at each point could differ since each point of the modelling process may have different needs.

In the following subsections, we will present how we will incorporate profit-sensitive measures, both directly (to estimate an XGBoost model) and indirectly (to decide on granting a loan), how we evaluate its financial performance and how to use explainability tools, particularly the SHAP values, with the proposed profit-sensitive measures.

4.4.1. Direct profit-sensitive optimization of the XGBoost algorithm

This section explains how to introduce a profit function as a learning function of the eXtreme Gradient Boosting (XGBoost) method proposed by [Chen & Guestrin \(2016\)](#). To do

that, we take advantage of the fact that XGBoost is generalizable to different loss functions. However, it demands the learning function to be continuously differentiable, as it happens with other boosting methods and some neural network models. Thus, the profit function defined must comply with this requirement.

We use the average expected profit function in Eq. (4.3). The learning or estimation method will determine the function values of F that maximize the profit, that is, $\operatorname{argmax}_F R\pi_T^E(F)$. Since machine learning algorithms typically minimize a loss function, we will consider as loss function our profit function in negative, i.e., $-R\pi_T^E(F)$.

In the Python implementation of the XGBoost model used,¹ the inclusion of a different loss function just requires the definition of the gradient and the Hessian of each instance for the optimization process. This implementation uses a diagonal approximation of the Hessian matrix. It incorporates this approximation into a second-order Taylor expansion of the objective loss function, optimised using Newton's method, considering a diagonal approximation of the Hessian matrix makes the optimization process more efficient, reduces the computational load, and gains convergence speed without neglecting the precision in searching for optimal values.

For our loss function $-R\pi_T^E(F)$, the gradient and the Hessian for each instance are defined as $g(x_i)$ and $h(x_i)$, with $g(x_i)$ as in Eq. (4.7) and $h(x_i)$ as in Eq. (4.8), respectively.

$$g(x_i) = F_i(x_i) (F_i(x_i) - 1) (y_i (\pi_{TP_i} - \pi_{FN_i}) + (1 - y_i) (\pi_{FP_i} - \pi_{TN_i})) \quad (4.7)$$

$$h(x_i) = F_i(x_i) (F_i(x_i) - 1) (1 - 2F_i(x_i)) (y_i (\pi_{TP_i} - \pi_{FN_i}) + (1 - y_i) (\pi_{FP_i} - \pi_{TN_i})) \quad (4.8)$$

4.4.2. Indirect profit-sensitive approaches

In this section, we explain how to incorporate profit functions outside the learning process of the algorithms.

¹https://xgboost.readthedocs.io/en/latest/python/python_intro.html

Model tuning

We can use a profit function as the guiding criteria of a model tuning process for the method considered. In machine learning, testing different hyperparameter configurations and even a different set of input variables is typical to decide which model setting is best according to a given criterion. The hyperparameter search can be carried out using, for example, a Grid Search, a Random Search, Genetic Algorithms or Bayesian Optimization. In our case, we will use the Bayesian optimization method with the support of the open-source package of Python *hyperopt*, which is based on [Bergstra et al. \(2013\)](#).

For model tuning, we can use any profit function. However, we will use a probabilistic one because our subsequent decision process will use the probabilities derived from the model adjusted with the best parameter set found.

In particular, in our experiment, we will use the annualized average expected profit rate per obligation $R\pi_T^E$ as in Eq. (4.3), as a criterion to select the best set of hyperparameters, regardless of whether or not the learning process incorporates the standard learning function or a profit-sensitive one.

We chose the annualized average expected profit rate because it is the same measure used in the direct estimation. It also aligns with the measure used in the decision function, which measures observed rather than expected returns.

In short, the chosen objective function allows us to estimate the probabilities of default that guarantee a higher expected net profit per financial liability which is consistent with our aim in the decision process of obtaining a higher average net profit.

Decision function

With respect to the inclusion of the profit function in the decision rules to define the prediction of $\hat{y}_i$, there are different ways to do that. Specifically, in this work, we will consider two of them: thresholding and Bayes Minimum Risk (BMR).

Regarding the thresholding process, many works use the 0.5 thresholds on F_i by default

and pay no attention to setting an adequate threshold value to improve the performance of the models. However, approaches such as those from [Verbraken et al. \(2014\)](#) and [Bahnsen et al. \(2014\)](#) define a decision rule based on an acceptance threshold or cut-off t on F_i that minimizes a loss function (L), i.e., $\operatorname{argmin}_t L$, as we did. Thus, the decision rule for defining the prediction $\hat{y}_i$ is set by a sign function such as $\operatorname{sign}(F_i - t)$.

As our work aims to maximize the profits and not the classification performance, we will use $\operatorname{argmax}_t \pi(t)$, being $\pi(t)$ a function that recognizes the profits derived from a correct or incorrect classification for each of the cells found in the matrix in Table 4.1. We will determine the threshold that maximizes the profit in the training data and evaluate it in the validation sets.

We could consider any profit function based on the default categories' final prediction $\hat{y}$, and in our case, we use the function $R\pi_T$ as in Eq. (4.6) that represents the net profit per financial liability and provides a clear financial interpretation.

Regarding BMR, we obtain $\hat{y}_i$ as proposed by [Bahnsen et al. \(2014\)](#) and [Bahnsen et al. \(2015\)](#). More precisely, as a function of risk R , as in Eq. (4.9).

$$\hat{y}_i = 1 \text{ if } R(\hat{y}_i = 1 \mid x_i) < R(\hat{y}_i = 0 \mid x_i). \quad (4.9)$$

In particular, given a profit-sense function $R\pi$, we define it as in Eq. (4.10).

$$\hat{y}_i = 1 \text{ if } R\pi(\hat{y}_i = 1 \mid x_i) > R\pi(\hat{y}_i = 0 \mid x_i), \quad (4.10)$$

where

$$R\pi(\hat{y}_i = 1 \mid x_i) = \pi_{TP_i} F_i + \pi_{FP_i} (1 - F_i), \text{ and} \quad (4.11)$$

$$R\pi(\hat{y}_i = 0 \mid x_i) = \pi_{TN_i} (1 - F_i) + \pi_{FN_i} F_i. \quad (4.12)$$

To realistically apply this strategy, given that we will not have all the matrix values from the application, we have estimated the misclassification values from the average data of the defaulted and non-defaulted obligations. For example, for π_{FN_i} , we used the average LGD of the defaulted records and the mean term to annualize the profit. However, it could be

possible to estimate these matrix values in other ways, resulting in different applications of this prediction method.

It is important to remark that the indirect approaches can also be applied to other methods, such as logistic regression, regardless of whether or not the learning or estimation process incorporates the profit function. This is precisely what we will do in our experiment.

4.4.3. A proposal of profit-sensitive XGBoost in the estimation and decision process

Our proposal takes as a basis the XGBoost that uses as a learning function the expected profit function in Eq. 4.3, that is, looking for the best average profit rate. We guide the hyperparameter search, maximizing the same expected profit function. For the best hyperparameters, the XGBoost trained provides a probability for a given loan. Then, we use decision rules based on the probability within a profit function, the average profit rate in Eq. 4.6, as a decision criterion of seeking to have better financial performance in rate versus classical methods or cost-insensitive methods converted into cost-sensitive through profit-based decision rules.

Thus, our approach includes profit-sensitive in three points. For the case of boosting methods, Masnadi-Shirazi & Vasconcelos (2011) says that although cost-insensitive approaches can be made indirectly cost-sensitive through threshold mapping, taking into account a cost matrix, the achieved result does not lead to optimal cost-sensitive detectors.

As explained in Section 4.5, we will compare our approach with other less sophisticated ones that only consider profit information in one or two points. In this way, we will evaluate whether the direct approach is better.

4.4.4. Model Explanation

To provide explainability to the proposed models, identify the drivers of the default and the drivers of the default weighted by the costs and profits derived from the misclassification

errors;

In this work, we will use the SHAP values ? initially proposed in the work of [Lundberg & Lee \(2017\)](#) to provide explainability to the proposed models. We will use them to identify the drivers of the default and the drivers of the profit or, more precisely, the drivers of the default weighted by the costs and profits derived from the misclassification errors. We will rely on the open-source Python *shap* package <https://github.com/shap/shap> for this.

This approach allows us to estimate the contribution of the variables in a model’s prediction, to know the variables’ relative importance, to identify the monotonicity and dependence relationships, and even to identify nonlinearity relationships derived exclusively from machine learning proposals.

The SHAP values generally express the difference between the prediction result in $f(x)$ and an expected model result based on a single input x , as in Eq. (4.13):

$$f(x) - \phi_0(f, x) = \sum_{i=1}^M \phi_i(f, x) \quad (4.13)$$

where M represent the number of input features, $\phi_i(f, x)$ a numerical value representing the impact of feature i on the model prediction f given input x and $\phi_0(f, x)$ the expected value of the prediction model on the training data set. The SHAP values explain the f prediction function as a sum of each input feature’s ϕ_i effects.

The SHAP values are mainly based on the definition of Shapley values developed in game theory, specifically in cooperative games ([Shapley, 1953](#)). In general, they aim to explain the prediction through the marginal contribution of each feature. The feature values of an instance of the data set behave like actors in a coalition. The Shapley values allow an equitable distribution of the payoff. In this case, the prediction, according to their contribution, i.e., the SHAP values $\phi_i(f, x)$ are calculated on the idea of Shapley values to attribute ϕ_i values to each feature.

This methodology not only generalizes Shapley’s values but also uses other additive feature attribution methods, such as Local Interpretable Model-Agnostic Explanations (LIME)

by Ribeiro et al. (2016) and Learning Important FeaTures (DeepLIFT) by Shrikumar et al. (2017), among others.

For the SHAP values of categorical values, we consider the adjustment proposed by Ariza-Garzón et al. (2020) that considers the dependence relationships between the categories of the variables. This affects the relative importance of the variables, which will be calculated from the mean of the absolute values per feature across the data according to Eq. (4.14):

$$I_j = \frac{1}{n} \sum_{i=1}^n |\phi_{ji}| \quad (4.14)$$

So, the relative importance of the characteristic will be estimated as the share of the importance of variable j over the aggregate importance for all variables.

The SHAP values will be used to observe the changes in the feature importance and the dependency relationships that can be observed when including profit information in the XGBoost approach and to identify the key feature values of the resulting model.

4.5. Experimental setting

4.5.1. Dataset and model variables

We will use the public data set from Lending Club,² which includes loans issued by the company between 2007 and 2018. The Lending Club (LC) data set is widely used in credit risk publications and the most used when addressing the P2P market (Ariza-Garzón et al., 2021).

It is important to note that the Lending Club dataset is of loans granted. This affects the experiment in several ways. One is that the dataset is biased toward a priori less risky loans since many of the a priori riskier loans were likely not granted. This makes it more difficult for a granting model that only uses the information available at the time of application, such as the one we propose, to determine which loans defaulted. Another is that financial

²<https://www.kaggle.com/wordsforthewise/lending-club>

performance information is available for all loans in the dataset, which allows us to use realistic data to propose a profit-based model, such as the one we propose in this paper.

Since we aim to develop a credit granting model, we create a target variable based on the credit’s final resolution. For this purpose, we use the loan status variable, representing the loan state at the end of the data set time window. Our target variable has two values representing the loan’s final state: the default category corresponds to a loan eventually charged off, and the non-default one represents fully paid loans. Our data set excludes loans with other values in the loan status variable since we cannot determine their final state.

As previously mentioned, many credit companies and banks consider small business credit as a specific product that deserves particular treatment, such as collecting specific information to assess the risk, granting, and eventual conditions. In the LC data set, the loans of the *small business* category represent around 1 % of the loans (15,575 loans) and present the highest default rate, 29.86 %, being the default rate of 19.86 % in the whole data set. Other works have already drawn attention to the credit risk of this segment ([Ariza-Garzón et al., 2020](#); [Havrylchyk & Verdier, 2018](#); [Polena & Regner, 2018](#); [Serrano-Cinca et al., 2015](#)) and some of them even have focused their analysis and modeling efforts on it ([Nowak et al., 2018](#)).

As we explained, we also consider that this type of credit needs more careful risk management and will focus our analysis on it. We recognize that its segment would need to be better characterized financially, including the type of investment made and its projected indicators. Nevertheless, we do not have such variables.

Tabla 4.2: Description of the input variables

<i>Variable</i>	<i>Description</i>
Categorical variables	
<i>emp_length</i>	Employment length. LC represents the current employment time in years into 12 categories, including the no information category.
<i>home_ownership_n</i>	The homeownership status provided by the borrower during the registration process. Categories defined by the entity: mortgage, rent, own, other (which includes other, none and any).
Quantitative variables	
<i>revenue</i>	Yearly income of the applicant (and co-borrowers, if any) self-reported in the registration process.
<i>dti_n</i>	Debt ratio per obligation. Debts are provided by the credit bureau (mortgage debts are excluded). The indicator is monthly. The income is the self-reported monthly income of the applicant (and co-borrowers, if any).
<i>loan_amnt</i>	Amount of credit requested by the applicant.
<i>fico_n</i>	Credit bureau score of the applicant. Defined between 300 and 850, reported by Fair Isaac Corporation as a summary risk measure based on historical credit information reported at the time of application.

Table 4.2 shows our models' categorical and quantitative variables. These variables are taken directly from the credit applications. Below, we explain the decisions and transformations we took for some variables.

In the *home_ownership_n* variable (see Table 4.2), we merged the categories 'other', 'none' and 'any' in the category 'other' since we did not differentiate these options with clarity and they exhibit similar behavior for the percentage of default. The other options of this categorical variable remain the same. We consider the employment length (*emp_length* in Table 4.2) as another categorical variable since it includes the category of no information and the category of more than ten years in the current employment.

Regarding the Fair Isaac Corporation credit bureau (FICO) variable (*fico_n* in Table 4.2), in the original data set, the information from FICO is given by two values that define the credit scoring interval of the applicant to the bureau, that is, a minimum and maximum range of limits to which the borrower's FICO belongs at loan origination. However, our FICO variable (*fico_n*) is estimated as the average of these two values to have a single

indicator of the creditworthiness of potential borrowers.

For the case of the debt variable (dtn), it is estimated from the original dataset variables as the ratio calculated from the total debts of the co-borrowers over the total debt obligation divided by the combined monthly income of the co-borrowers.

4.5.2. Estimation of the profit functions

Tabla 4.3: Description of variables

<i>Variable</i>	<i>Description</i>
<i>term</i>	The number of payments on the loan. Values are in months and can be either 36 or 60.
<i>funded_amnt</i>	The total amount committed to that loan.
<i>total_payment</i>	Payments received to date for total amount funded.
<i>total_rec_prncp</i>	Principal received to date.
<i>recoveries</i>	Post charge off gross recovery.
<i>collection_recovery_fee</i>	Post charge off collection fee.

Table 4.3 lists the financial variables used to calculate the values of our profit-based confusion matrix. The information available in the data set conditions how such values are estimated. It is worth clarifying that we have not included operative costs related to the loan management and granting process, use of the platforms, evaluation of models, or study of loan applications. The values we propose for the cells of the confusion matrix are explained below.

The value π_{FP_i} is the value of not granting a loan that will be returned, and we consider it zero. Alternatively, it could be the amount that the entity did not earn with borrower i minus what would be achieved on average by placing the money with other customers, i.e., the lender loses the benefit of what would have been a good customer. In that case, the expected loss would be netted by recognizing the expected profit of placing the money with other borrowers, similar, for example, to what [Bahnsen et al. \(2014\)](#) proposes. Furthermore, we do not have data to estimate these values for the small-business purpose adequately. For this reason, we assume that the money will be reinvested in new obligations and that, in

our case, it would not represent any loss. Similarly, as mentioned previously, we assume that π_{TP_i} is zero, avoiding a financial loss for an individual lender.

For π_{TN_i} , we define it as annualized recoveredROI (ROI_rec_i) of having classified obligation i correctly as in Eq. (4.15)

$$\pi_{TN_i} = (ROI_rec_i + 1)^{\frac{12}{term_i}} - 1, \quad (4.15)$$

where ROI_rec_i is as in Eq. (4.16)

$$ROI_rec_i = \frac{total_pymnt_i}{funded_amnt_i} - 1, \quad (4.16)$$

where $total_pymnt_i$ includes the principal recovered, interest and late fees received to date.

For π_{FN_i} case, we use an estimation of annualized profit achieved by a lender, taking into account the estimation of expected loss arising from the obligation falling into default as in Eq. (4.17):

$$\pi_{FN_i} = \left(\frac{(1 - LGD_i)(funded_amnt_i - total_rec_prncp) + total_pymnt_i}{funded_amnt_i} \right)^{\frac{12}{term_i}} - 1. \quad (4.17)$$

Based on Basel guidelines, the π_{FN_i} value includes components that estimate potential loss risk. In particular, critical risk parameters, such as the probability of default (PD), LGD and exposure at default (EAD), are used to obtain the Expected Loss ($PDLGDEAD$) in credit risk management. In our case, the value of EAD corresponds mainly to the outstanding loan amount of the obligation at the time of our evaluation.

For the case of LGD , according to Zhou et al. (2018), there are mainly three approaches for its estimation: workout LGD, market LGD and Implied Market LGD. For our data, we follow the ideas of Papoukova & Hajek (2019) and use a workout LGD estimation as in Eq. (4.18):

$$LGD_i = 1 - \frac{recoveries_i - collection_recovery_fee_i}{outstanding_loan_amount_i} \quad (4.18)$$

where $outstanding_loan_amount$ is estimated as $(funded_amnt - total_rec_prncp)$.

4.5.3. Models considered

In this section, we describe the different approaches we will use for modeling the problem of granting loans, including profit-sensitive elements. As explained in Sections 4.4.1 and 4.4.2, we can introduce such elements at three points of the modelling stage, namely:

1. The profit function used in the learning of the algorithm, $R\pi_T^E$ as in Eq. 4.3 from which we obtain the gradient and hessian.
2. The profit function to guide the model tuning, in our case, using Bayesian optimization, $R\pi_T^E$ as in Eq. 4.3.
3. The function used for making the decision on the credit granting. In this case, we will consider two alternatives:
 - BMR, where we compare the risks-profits associated with predicting an obligation as Default or as not-Default, as in Eq. 4.10.
 - Thresholding, where we find the threshold that maximizes the function $R\pi_T$ as in Eq. 4.6.

Our main proposal includes profit functions at the three points of the modeling stage using the XGBoost algorithm (*XGB*). In this way, we evaluate the performance that direct profit-sensitive methods can generate in conjunction with indirect techniques for model-tuning and decision-making.

We will confront this result with the ones obtained by the simplest approach of not considering profit elements at any stage. In particular, as non-profit sensitive alternatives, we will consider:

1. The standard learning function used by the method considered, parametrized whenever possible, to deal with the unbalanced nature of our data set.

2. The balanced accuracy measure (BACC) guides the hyperparameter optimization. This measure averages the accuracy obtained from both the minority and majority classes.
3. The use of the standard threshold or cut-off $t = 0,5$.

We will also consider other intermediate approaches that consider including the profit functions only at some of the points of the modelling stage.

Tabla 4.4: Modeling approaches used in the paper according to the method considered, the learning optimization and decision functions used and the rule used for decision making.

Row	Abbreviation	Method	Learning Function	Optimization Function	Decision Function	Decision Rule
1	XGB-HO-0.5	XGB	Standard	BACC	None	$t = 0,5$
2	XGB-HO- πt	XGB	Standard	BACC	$R\pi_T$	t
3	XGB-HO- π BMR	XGB	Standard	BACC	$R\pi_T$	BMR
4	XGB- π HO-0.5	XGB	Standard	$R\pi_T^E$	None	$t = 0,5$
5	XGB- π HO- πt	XGB	Standard	$R\pi_T^E$	$R\pi_T$	t
6	XGB- π HO- π BMR	XGB	Standard	$R\pi_T^E$	$R\pi_T$	BMR
7	π XGB- π HO-0.5	XGB	$R\pi_T^E$	$R\pi_T^E$	None	$t = 0,5$
8	π XGB- π HO- πt	XGB	$R\pi_T^E$	$R\pi_T^E$	$R\pi_T$	t
9	π XGB- π HO- π BMR	XGB	$R\pi_T^E$	$R\pi_T^E$	$R\pi_T$	BMR

For the sake of clarity, Table 4.4 shows all the modeling approaches that we will compare in our experiment. The table also shows which function is used at each stage: learning, optimization and decision. We will explain the abbreviations used to refer to them. Each abbreviation consists of three parts separated by '-'. The first is the classification method (XGBoost named as XGB for all the cases), the second is the optimization approach (Bayesian optimization or BO for all the cases), and the third is the decision method. For the decision stage, we will represent the thresholding method by t and the binary Bayes Minimum Risk classifier (Bahnsen et al., 2014) by BMR, while the standard threshold is denoted by 0,5. The inclusion of π in the abbreviation represents that the profit function is considered at that stage, while its omission means that no profit function is used. For example, the approaches that include all the profit-sense approaches proposed in this paper can be

found in rows 8 and 9, identified by the abbreviations $\pi\text{XGB-}\pi\text{GA-t}$ and $\pi\text{XGB-}\pi\text{GA-BMR}$, respectively.

4.5.4. Model adjustment

We used a stratified evaluation k-fold cross-validation with $k = 4$ on the training set to avoid over-fitting in the machine learning models. We use the k-fold to search for the best hyperparameter combination using Bayesian Optimization, provided by the Python open-source package *hyperopt* (<http://hyperopt.github.io/hyperopt/>). We define the initial space search for the optimisation assuming that the hyperparameters are uniformly distributed.

The hyperparameters considered for the XGBoost are the number of trees ($n_estimators$), the maximum tree depth (max_depth), the subsample ratio of the training instances in every boosting iteration ($subsample$), the proportion of features when constructing each tree ($colsample_bytree$), the proportion of features considered by level ($colsample_bylevel$), the learning rate (eta), and the regularization hyperparameters ($alpha$, $lambda$ and $gamma$). Furthermore, for the profit-insensitive and the indirect profit-sensitive algorithms, we also optimized the $scale_pos_weight$ parameter to deal with the class-imbalance problem.

For the XGBoost we will use the Python implementation of XGBoost (<https://xgboost.ai/>) proposed by Chen & Guestrin (2016). We also use the Python library Scikit-learn (<https://scikit-learn.org/stable/>) (Pedregosa et al., 2011).

4.5.5. Performance metrics

Regarding performance, the different approaches will be evaluated both in financial and classification terms. Among the financial measures, we will use the profit functions included in the learning as $R\pi_T^E$ defined in Eq.(4.3), which is a probabilistic measure that represents the expected return, and a criterion that measures the financial impact of the decision taken as $R\pi_T$ defined in Eq.(4.6). We will also provide the loan acceptance rate (LAR) that

measures each method's percentage of loans granted.

As classification metrics, we will show measures such as the aforementioned AUC and Kolmogorov-Smirnov (KS) and metrics on the observed classification such as the aforementioned Balanced Accuracy (*BACC*).

We will also use inferential methods to evaluate whether the profit-sensitive methods generate better financial performance than the traditional alternatives. We want to assess whether profit models are selected in the set of models with better statistical performance per profit. For doing so, we will use the model confidence set (MCS) methodology proposed by Hansen et al. (2011) and used in other credit risk works (Lyócsa et al., 2022). The MCS methodology selects the best models based on a given significance level (α). The higher α , the lower the confidence level, and the more models tend to be selected in the top set of models. We will use inferential criteria through a non-parametric test for multiple comparisons and paired samples to compare alternatives. The algorithm is a model selection process that consists of performing a hypothesis testing process that starts with all models and eliminates options as long as the null hypothesis is rejected. The process stops when the null hypothesis is not rejected or when only one model option remains. We will use the MCS library of R <http://cran.r-project.org/web/packages/MCS/index.html> developed by Bernardi & Catania (2018).

In our case, we define the loss function as the additive inverse of the profit per model and per obligation i ($l_{i,m} = -R_{\pi_{i,m}}$), similar to what Lyócsa et al. (2022) did. Thus, a higher profit translates into a lower loss. The difference between losses of the m and n models is given by $d_{i,mn} = l_{i,m} - l_{i,n}$, with m, n models under evaluation. Under the null hypothesis $E(d_{i,mn}) = 0$ for all m, n models under evaluation.

We will use T_{max} , see Hansen et al. (2011) for an explanation, as a test statistic. For the inferential process and selection of the best set of models, the method uses a bootstrapping procedure, which, in our case, will consider 3,000 bootstrap samples.

4.6. Results

4.6.1. Preliminary descriptive and inferential analysis

In this section, we analyze the data set: the independent variables used, those that characterize the financial aspects of the obligations and their impact on the profit function, and those used to predict the outcome. These results will help us better contextualise the discussion of the results in the following sections.

Tabla 4.5: Descriptive Analysis. Quantitative variables

Variable	Statistic	Non-Default	Default	Total
revenue	Mean	\$ 93,443.10	\$ 90,129.74	\$ 92,453.88
	Median	\$ 75,000.00	\$ 74,000.00	\$ 75,000.00
	SD	\$ 78,535.02	\$ 97,266.65	\$ 84,573.41
	KS D-Test	0.04***		
dti_n	Mean	14.10	15.71	14.58
	Median	13.06	14.88	13.60
	SD	8.74	9.57	9.03
	KS D-Test	0.07***		
fico_n	Mean	708.54	698.69	705.60
	Median	702.00	692.00	697.00
	SD	38.61	32.18	37.08
	KS D-Test	0.10***		
loan_amnt	Mean	\$ 15,032.92	\$ 17,039.00	\$ 15,631.85
	Median	\$ 12,150.00	\$ 15,000.00	\$ 13,750.00
	SD	\$ 9,702.54	\$ 9,722.70	\$ 9,751.56
	KS D-Test	0.10***		

*** significant at the 0.01 level

Tabla 4.6: Descriptive Analysis. Categorical variables

Variable	Category	Count	Rel. Frec.	Default Rate	CHI Test
home_ownership_n	MORTGAGE	6,923	44.4 %	26.5 %	71.648***
	OTHER	20	0.1 %	25.0 %	
	OWN	1,674	10.7 %	30.0 %	
	RENT	6,958	44.7 %	33.1 %	
emp_length	<1 year	1,354	8.7 %	31.4 %	25.338***
	1 year	1,203	7.7 %	33.2 %	
	2 years	1,707	11.0 %	29.7 %	
	3 years	1,545	9.9 %	30.4 %	
	4 years	1,179	7.6 %	32.2 %	
	5 years	1,140	7.3 %	28.9 %	
	6 years	779	5.0 %	28.0 %	
	7 years	750	4.8 %	32.5 %	
	8 years	713	4.6 %	29.0 %	
	9 years	513	3.3 %	27.9 %	
	10+ years	4,254	27.3 %	27.9 %	
	NI	438	2.8 %	32.0 %	

*** significant at the 0.01 level

Tables 4.5 and 4.6 present a descriptive summary of the data set we are using, including some inferential tests to identify the potential of the chosen variables to predict default. In particular, we use the Kolmogorov-Smirnov test for quantitative variables to compare the empirical cumulative probability distributions between the Default and Non-default records. On the other hand, for categorical variables, we use the chi-square test to evaluate the dependence of each of these variables on the target variable (Default and Non-default).

Table 4.5 shows that the behavior of quantitative variables differs significantly for default and non-default obligations. Defaulted obligations are characterized by lower *revenue*, higher indebtedness (*dti_n*), lower FICO (*fico_n*) and higher requested amount (*loan_amnt*). These results are consistent with what is expected for credit risk.

Table 4.6 shows the result for qualitative variables where we find a significant dependence

between the target variable and the variables that describe the homeownership status (*home_ownership_n*) and the employment length (*emp_length*). In the *home_ownership_n* variable, the *RENT* and *OWN* categories show the highest risk (33.1 % and 30 % default rates, respectively). In *emp_length* variable, the categories that denote a shorter length or no information are among those with higher risk, all over 30 %. In general, we see that employment length is not perfectly ordered with the default rate, finding, for example, significant default rates also in 7 years of employment length (32.5 %), which supports the inclusion of this variable in the model through one-hot encoding.

Tabla 4.7: Financial description of loans

Variable	Statistic	Non-Default	Default	Total
Loans	Count	10925	4650	15575
	Rate	70.144 %	29.856 %	100.000 %
Annualized profit rate	Mean	0.05395	-0.21065	-0.02505
	Median	0.05411	-0.19699	0.03559
	SD	0.02935	0.15859	0.15091

Table 4.7 shows the financial characterization of the loans. The table shows that the small business segment has a high default rate, and the annualized profit rate of the defaulted loans exhibits a mean loss of 0.21065 (the loans have terms of 36 or 60 months). The total distribution is also slightly skewed to the left with a negative mean and positive median, suggesting the possibility of obtaining a small profit from selecting some of the obligations.

The mean annualized profit rate is -0.02505, with high volatility, and the default is high. All these aspects reinforce the idea of the high risk of the small business segment. These results demand a risk prediction model from granting but show the complexity of achieving it, as we expected from the literature review of this financial market segment. Our profit-sensitive approaches will validate whether it is possible to increase the naive profit, find positive values and, thus, reduce risk for lenders.

4.6.2. Performance evaluation

In this section, we report the results of the different approaches in the k-fold cross-validation setting. The tables report the average of the measures across the four validation folds together with their respective standard deviation (denoted as sd).

Tabla 4.8: Expected profit and score-based classification measures in validation

Model	$R\pi_T^E$	sd	AUC	sd	KS	sd
Naive by granting all loans	-0.02505	0.00122	-	-	-	-
Naive default rate	-0.00748	0.00037	-	-	-	-
XGB-HO	-0.01196	0.00091	0.5595	0.0070	0.0946	0.0047
XGB- π HO	-0.00361	0.00035	0.6151	0.0069	0.1785	0.0119
π XGB- π HO	0.00059	0.00019	0.5848	0.0173	0.1312	0.0265

Table 4.8 shows the performance of the estimated probability function for the XGBoost alternatives. At this point, we only have three alternatives: the profit-insensitive one, the one that uses indirect learning of the profits (in the hyperparameter optimization) and the one that uses both direct and indirect learning of the profits (in the learning function of the XGBoost and the hyperparameter optimization). The only alternative that obtains a positive expected profit rate is the π XGB- π HO, which uses the profit directly and indirectly, while the worst is the profit-insensitive alternative (XGB-HO). Furthermore, all the proposed methods perform much better than the naive method that grants all the loans. However, if we look at the naive method that considers the observed default rate (29.86 %), it outperforms the profit-insensitive alternative. This result shows the potential to improve profitability by using a reasonable estimate of a probability associated with default, as it can balance both risk and profitability in a unified model. The better this estimate is, the greater the chances of obtaining good expected profitability results, which enhances potential investments in small business loan purposes and helps identify better loans.

Regarding the classification measures (AUC and KS), the worst method is, again, the profit-insensitive alternative. It suggests that trying to guide the learning using profit information can be useful for more accurately estimating the probabilities of the classification

function.

Tabla 4.9: Profit and classification measures for the decisions in validation

Model	$R\pi_T$	sd	LAR	sd	DR	sd	$BACC$	sd
XGB-HO-0.5	-0.01195	0.00116	0.60411	0.0086	0.27271	0.00377	0.53727	0.00554
XGB-HO-t	-0.01754	0.00177	0.79904	0.01515	0.28423	0.00293	0.52737	0.00598
XGB-HO-BMR	-0.00716	0.00040	0.43634	0.00767	0.25634	0.00150	0.54399	0.00222
XGB- π HO-0.5	0.00041*	0.00038	0.06388	0.00350	0.13505	0.01960	0.52484	0.00217
XGB- π HO-t	-0.00081	0.00085	0.24597	0.02013	0.19118	0.01154	0.56282	0.00600
XGB- π HO-BMR	0.00021*	0.00009	0.01393	0.00303	0.10929	0.05830	0.50625	0.00222
π XGB- π HO-0.5	0.00077*	0.00026	0.08706	0.01652	0.16941	0.00916	0.52665	0.00421
π XGB- π HO-t	0.00051*	0.00040	0.10684	0.02056	0.17699	0.01871	0.53048	0.00443
π XGB- π HO-BMR	0.00069*	0.00032	0.05849	0.00889	0.16097	0.01686	0.51916	0.00352

Indicates model that belongs to the set of top models for a given α level (0.10), with MCS procedure.

Table 4.9 shows the performance of the models once the classification decision is made. We have three alternatives for each of the methods considered in the previous step: profit-insensitive (using the 0.5 threshold) and profit-sensitive (thresholding and BMR). The business measures include the average profits of the prediction options $R\pi_T$ the loan acceptance rate (LAR) and the default rate (DR), while as classification measures, we show the balanced accuracy $BACC$.

First, we can see that all the methods outperform the mean annualized profit rate of the dataset, which was -0.02505. Thus, learning using either profit-sensitive or -insensitive approaches is beneficial in financial terms. However, a closer look reveals that the approaches with profit-sensitive learning generally perform best. In particular, using the MCS procedure, we identified the methods that yield the highest annualized average profits and we can see that all the methods selected use profit-sensitive learning; the best ones include the profit directly in the learning, that is, the π XGB- π HO with the three decision-making strategies. Next, two of the indirect-learning options, XGB- π HO-0.5 and XGB- π HO-BMR, are also included in the model confidence set (MCS). In either of the best methods, the LAR rates are low (ranging from 24.6 % to 1.4 %). Thus, these methods have a selective behaviour that results in higher returns. If we look at the balanced accuracy, we can see that the three

methods with the highest value do not belong to the selection by the MCS procedure, which denotes no clear correlation between classification performance (using a balanced measure) and profitability, as already reported in other works (Verbraken et al., 2014; Wang et al., 2021; Xia et al., 2017a).

Another remarkable conclusion is that the methods that include the profit in the learning function and in the hyperparameter optimization are better than those that only include it in the hyperparameter optimization for each of the three decision-making methods. Similarly, the methods that include the profit in the hyperparameter optimization are better than the profit-insensitive ones for each of the three decision-making methods. This suggests that including profit information for guiding the learning, both directly or indirectly, is good for maximizing profits. However, if we look at the decision-making alternatives, the results are not conclusive. The profit-insensitive decision-making strategy, considering a 0.5 threshold, is the best strategy for the π XGB- π HO and XGB- π HO cases. In contrast, the BMR is the best decision-making strategy for the profit-insensitive case (XGB-HO). The thresholding method cannot produce better results in validation than the 0.5 threshold. As a result, our results do not support that profit-sensitive decision-making contributes in financial terms, at least under the three decision-making proposals studied.

We can see that the XGB-HO methods have the highest acceptance rates (all of them over 40 %) but at the same time the highest default rates (all of them over 25 %, while the default rate in the whole dataset, as shown in Table 4.7, was close to 30 %). The rest of the methods, which use profit for learning directly or indirectly, obtain much lower default rates, but also much lower acceptance rates. It means that, in this case, including profit in the learning process produces classifiers that cut risk by reducing the default rate, which in turn is achieved by applying more restrictive filtering of the loans exploiting their financial information.

Tabla 4.10: Profit values of the misclassification matrix in validation

Model	π_{TN}	sd	π_{FN}	sd
XGB-HO-0.5	0.05179	0.00018	-0.21066	0.00511
XGB-HO-t	0.05291	0.00028	-0.21054	0.00637
XGB-HO-BMR	0.05065	0.00047	-0.21098	0.00372
XGB- π HO-0.5	0.03982	0.00087	-0.20536	0.01468
XGB- π HO-t	0.04578	0.00050	-0.21017	0.00626
XGB- π HO-BMR	0.03394	0.00308	-0.15038	0.03983
π XGB- π HO-0.5	0.05168	0.00157	-0.19842	0.01890
π XGB- π HO-t	0.05088	0.00203	-0.20516	0.01698
π XGB- π HO-BMR	0.05111	0.00174	-0.19281	0.03462

Finally, in Table 4.10, we look at the values of our profit-sensitive misclassification matrix that can take a value different to zero according to our assumptions to define the profit function in Section 4.5.2. Those values are the profit obtained by the true negatives (π_{TN}), i.e., the correctly identified non-default loans, and the loss obtained by the false negatives (π_{FN}), i.e. the accepted loans that will default. As a reference, we will use the mean annualized profit rates of the non-defaulted and defaulted loans in the whole dataset (see Table 4.7), which are 0.05395 and -0.21065, respectively. We can see that in the methods with profit-insensitive learning (i.e. XGB-HO) the values of the mean profit and the mean loss are very similar to those in the whole dataset. It means that the classification method cannot discriminate the loans in financial terms efficiently. However, the methods that use indirect profit-based learning (i.e. XGB- π HO) are able to obtain smaller losses on average at the expense of cutting the average profits (values below 0.05). Interestingly, the methods that present better financial behaviour are those that use both direct and indirect profit-based learning (i.e. π XGB- π HO) because they can reduce the losses on average without greatly reducing the average profits (all of them over 0.05). Thus, we can see that introducing profit information in the learning process produces classifiers that simultaneously look at cutting risks and obtaining profits. We will deepen into the behavior of the three approaches in the next section.

4.6.3. Profit-sensitive explainability

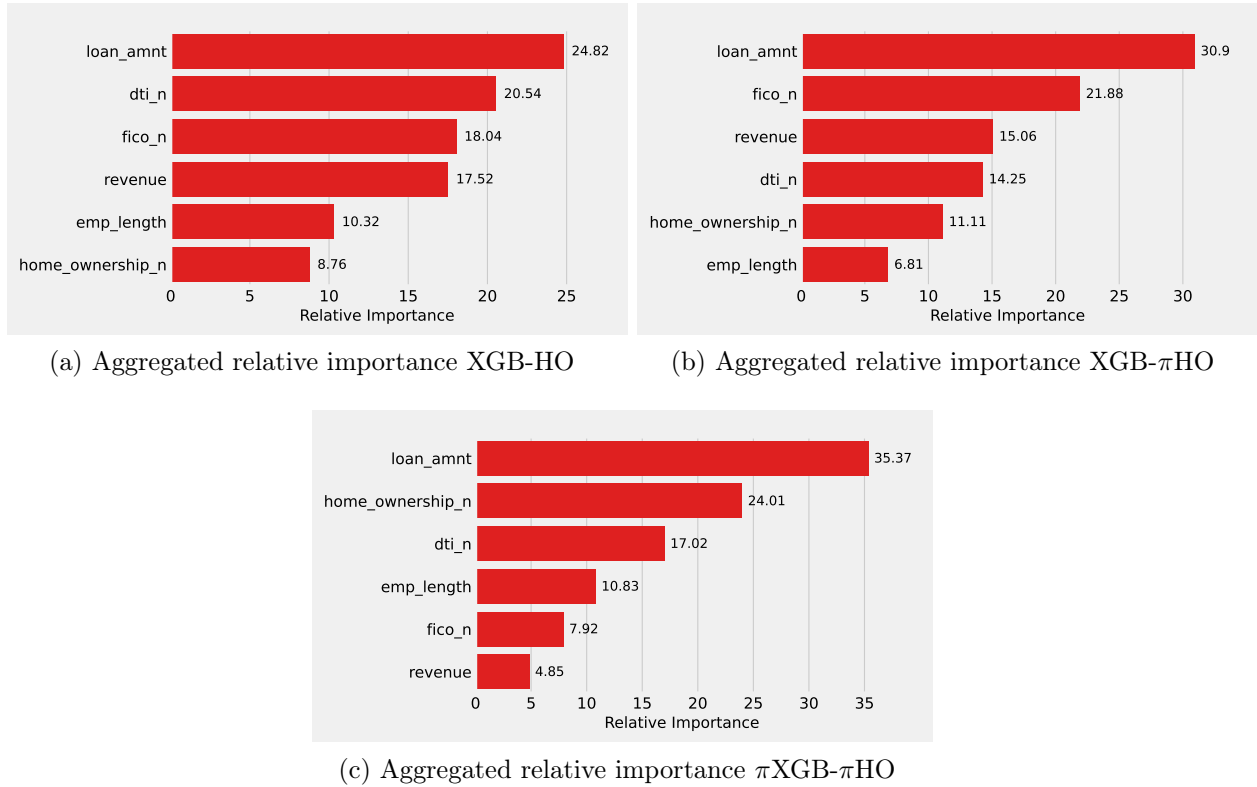


Figura 4.1: Aggregated relative importance by XGBoost model

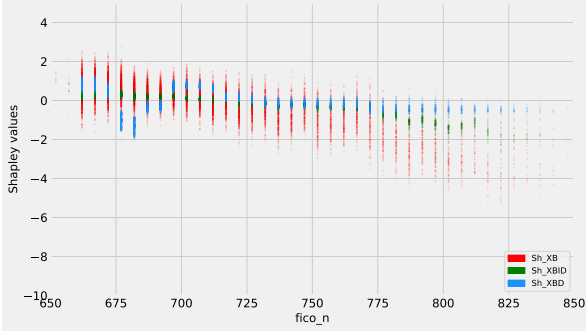
In this section, we will analyze the three models' explainability. Figure 4.1 shows the relative feature importance.

For the profit-insensitive model in Fig. 4.1a, we see that quantitative variables are the most important, followed by the length of employment *emp_length* and *home_ownership_n*. It is important to highlight that the order of the variables for this model is completely different from that reported by Ariza-Garzón et al. (2020) for the same model and the whole Lending Club dataset. It suggests that the segment of small businesses demand their own credit risk models.

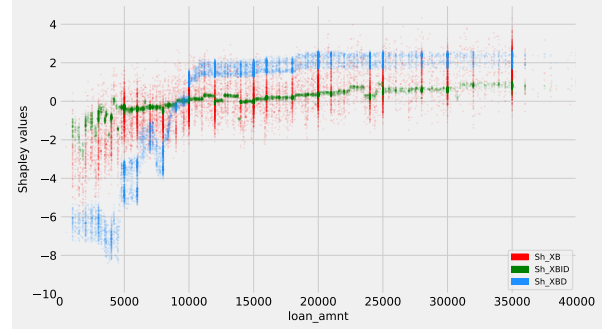
In the case of the indirect profit-sensitive model in Figure 4.1b, the most important variables are again the quantitative ones, but the *loan_amnt* is followed by the FICO (*fico_n*)

variable, and together account for more of the 50 % of the relative importance. At the same time, the *dti_n* variable loses prominence compared to the profit-insensitive option. On the qualitative side, the variable *home_ownership_n* takes on greater importance, with a share of over 10 %, while *emp_length* significantly reduces its share.

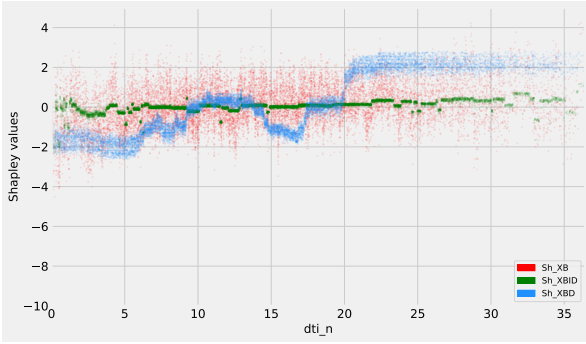
Finally, Figure 4.1c shows the variable importance of the direct and indirect profit-sensitive model. In this case, the loan amount variable remains the most relevant variable, with even higher relative importance, which means it is even more important for detecting profits than defaults. It is followed by the *home_ownership_n* variable, which stands out from the other variables. It means that its role is more important in a profit-sensitive model than in a model that just predicts default. The *dti_n* indebtedness maintains its prominence, and the *emp_length* increases it. The FICO (*fico_n*) and *revenue* variables lose strength, which means that they are more relevant for predicting the default rather than for seeking profits. FICO (*fico_n*) recognizes prior financial experience but does not guarantee higher average profit. Similarly, for the case of the *revenue* variable. Furthermore, revenues in the small-business segment combine business and applicant revenues and are self-reported. This variable does not allow for a precise risk prediction in combination with the expected profit per obligation. This last variable would be more relevant if this small business market were to specialize, recognizing the specific income of the business alternative with an adequate verification process.



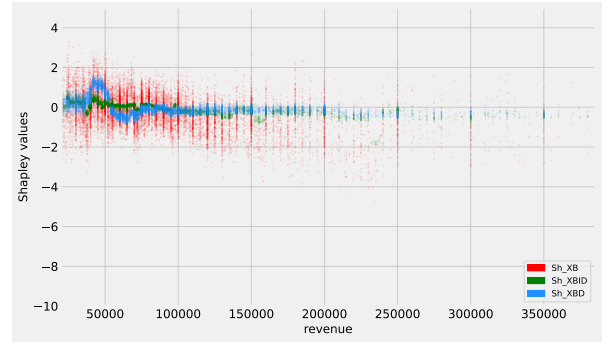
(a) *fico_n* dependence graph



(b) *loan_amnt* dependence graph



(c) *dti_n* dependence graph



(d) *revenue* dependence graph

Figure 4.2: Dependence graphs of quantitative variables.

In Figure 4.2, we see the dependence graphs of the quantitative variables for the three models (the insensitive approach in red, the indirect approach in green and the direct and indirect approach in blue). We can appreciate that the dependency relationships of the variables are different for the three models. Interestingly, the dispersion of the profit-insensitive model is higher for the four variables. It means that when predicting default, each loan can contribute differently for the same value of a given value. However, when predicting default informed by profit, the variation of the contribution is reduced. It means that considering

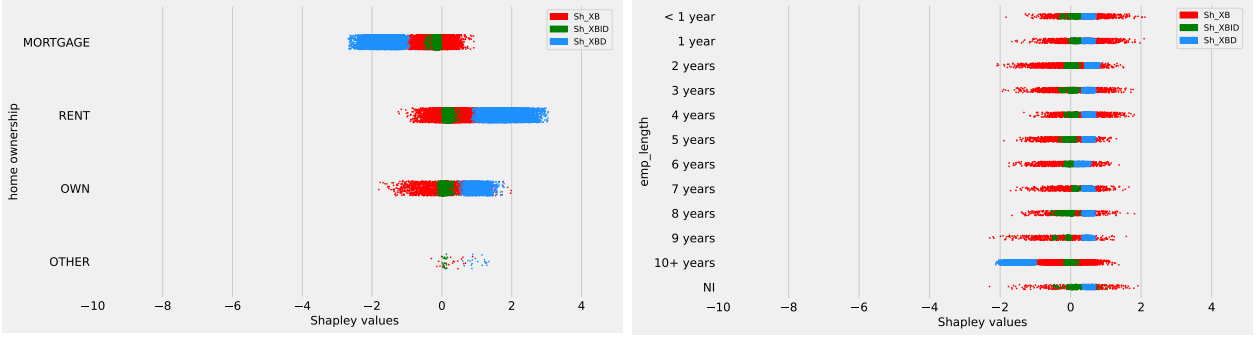
profit reduces the uncertainty of the variable's role in the model.

If we look at the FICO variable (*fico_n*) in Figure 4.2a, we see, for example, that its highest values contribute much more to detecting non-default in the profit-insensitive model (red dots) than in the ones that include profit, which is closer to zero. If we look at the model that considers profit both directly and indirectly. We see a break on 680, as if detecting a niche of profitable non-defaulted loans at slightly lower values is possible. At values above 750, we also identify profitable non-defaulted loans.

Loan amount, shown in Figure 4.2b was the most relevant variable for all the models. We can see how having a direct profit-sensitive proposal substantially changes the structure and linearity of the variable's relationship for default prediction. Although the higher the loan amount, the higher the risk, we can see that the model that considers profit both directly and indirectly has a more marked relationship and takes more extreme values. This model considers that the loans below \$5,000 are extremely interesting (SHAP values around -6), and those over \$10,000 are the ones the model tries to avoid (SHAP values around 2). This behavior suggests a market opportunity for microloans.

The variable *dtn* is shown in Figure 4.2c. Again, this variable's role and contribution are more marked for the model that considers both profit directly and indirectly. However, the relationship shows several breaks as if identifying niches of different behavior. Loans with values of *dtn* below 6% are the ones more interesting (SHAP values around -2), while those over 20 are the ones that should be avoided (SHAP values around 2).

In Figure 4.2d we see the dependency of the revenue variable. The profit-insensitive model allowed for more extreme values and hinted at an inverse relationship (the higher the revenues, the less probability of default). However, such a relationship is less clear for the profit-sensitive models, where we also see how the SHAP values tend to concentrate around zero, corroborating the loss of relevance of this variable.



(a) *home_ownership* dependence graph

(b) *emp_length* dependence graph

Figure 4.3: Dependence graphs of categorical variables.

The dependency relationship of the categorical variables is shown in Figure 4.3. Interestingly, the model that considers profit information both directly and indirectly determines more clearly the contribution of each category to the prediction, as it can assign to each category SHAP values that do not include zero. It can be seen in the *home_ownership_n* variable in Figure 4.3a, where this model considers MORTGAGE as the lowest risk category and RENT and OWN as the riskier categories. However, the role of the categories is much less marked in the other models. It reveals that this variable can offer more valuable information when looking for profit. In particular, it could be the case that when you have a mortgage debt, you tend to take better care of the budget, and your decisions are more measured.

The case of the *emp_length* variable in Figure 4.3b is similar. The model that considers the profits both directly and indirectly can assign more significant values to the categories and is much less dispersed than the profit-insensitive alternative. In particular, it considers that all the categories are similar in risk, except for the case of those loans from people who held their employment for more than 10 years, a niche that the model considers that reduces the default risk weighted by profit information.

The explainability tools have helped us find that profit-sensitive ones use the input variables differently than those profit-insensitive ones. In addition, we have seen that in this case, profit information serves to determine the variable contribution more precisely, that is, with less variability. Furthermore, for the model that considers profit both directly and indirectly, the explainability tools have helped us to determine a clear profile of the most interesting loans in terms of default weighted by profitability for the small business segment. These loans should be below \$5,000, and the borrower should have a DTI below 5 %, a mortgage, and hold his or her employment for more than ten years, while the FICO score and the revenues are much less relevant.

4.7. Conclusions

We have proposed a machine learning model for credit risk that uses profit information at three modeling steps: in the learning function, guiding the hyperparameter optimization, and making the granting (i.e. classification) decision. We have applied it as a granting model for the small business loans of the Lending Club dataset, which is a particularly complex and risky market segment. Finally, we have analyzed how the role of the variables differ when profit information is considered with the help of explainability tools.

In particular, our profit-based approach goes beyond the cost-sensitive scope by including costs and profits in the misclassification matrix. Moreover, we include the matrix in the three aforementioned steps using the popular machine-learning algorithm XGBoost. For incorporating profits in the learning function, we defined the profit-based gradient and the Hessian for each instance using the cell values in the matrix. For model tuning, we used Bayesian optimization to find the hyperparameters that maximized the $R\pi_T^E$ as in Eq. 4.3. We used both thresholding and Bayes Minimum Risk (BMR) for the decision method. In the first case, we aim to maximize the $R\pi_T$, as in Eq. 4.6, while in the BMR, we compare the risks-profits associated with predicting an obligation as default or as not-default, as in Eq. 4.10.

Our results show that including profit information in the learning function and optimizing the hyperparameters translate into better returns. However, the profit-based approaches for analysing the classification decision did not obtain more profits than using the standard method (i.e., the 0.5 threshold). This result is surprising and might be due to the complexity of the small business segment. It remains as future research considering other competing methods, e.g. (Herasymovych et al., 2019; Verbraken et al., 2014; Xia et al., 2017a), in other datasets to find further evidence about profit-based decision methods.

We also observed that profitability and classification performance do not necessarily correlate, as was observed in previous works (Verbraken et al., 2014; Wang et al., 2021; Xia et al., 2017a). Our results extend this notion as they show that the models obtained by maximizing the classification performance and maximizing profit are different in terms of how they exploit the information of the variables. In particular, the explainability analysis revealed that including profit information in the model changes the variables' role. This was observed in the different relative importance assigned by the models to each feature, but also in changes in the functional relationships of the quantitative variables with the target variable or in the different roles that played some categories of the categorical variables. Furthermore, we also observed that in the profit-based models, the contribution of a variable for a given value has less variance, which means that including the profit consideration reduced the uncertainty of the variable's role present in the profit-insensitive models.

These facts support the idea that credit risk modeling must include profit information to be successfully applied in real life and improve risk management. We have proposed a combined methodology, which shows the advantages of including risk and profit in a single modeling process, unlike other papers that do it separately with a model for each target (Lyócsa et al., 2022; Serrano-Cinca & Gutierrez-Nieto, 2016) or in stages, with a model for predicting default and then selecting the most profitable obligations, such as (Bastani et al., 2019; Xia et al., 2017a), among others. As can be seen, there are several alternatives to manage the risk-profit dichotomy. Therefore, it is necessary to investigate them further to

find the best ways to tackle this problem in lending models.

Going deeper into the results of our experiment from the business perspective, we observed that the best returns corresponded with models with low acceptance rates, ranging from 11 % to 1 % (i.e. models that grant very few loans). This result suggests that obtaining profits in the small-business segment implies filtering out most of the loans, granting only those with the lowest expectations of default weighted by profit. At this point, it is important to remember that the loans in the Lending Club dataset were granted; that is, the loans in the dataset were a priori less risky than those not granted (and not part of the dataset). These aspects make it more challenging to build a granting model, as we have seen in the low classification rates obtained, and make it more surprising the result of the profitable models are those with extremely low acceptance rates.

Furthermore, explainability tools also helped us identify a prototypical loan profile with low risk weighted by profitability for the model that considers profit directly and indirectly. In particular, the best opportunities to grant credit for small business purposes are mainly found in low indebtedness below 6 % and loan amounts below \$5,000, which greatly reduce the risk.

Regarding home ownership, the mortgage category is the only one that reduces the risk, as it happens with more than ten years of employment. The income and FICO variables are not very relevant to our profit-based model. In the case of income, probably a verified net income of the business (instead of the applicant's self-reported income) would produce a different result. We can see that the prototypical loan, according to our model, is very specific.

All these aspects reinforce the idea of the complexity not only of the risk-profit dichotomy but also of the particular segment where we worked: the small business segment in p2p lending. In the Lending Club dataset, this segment has a high default rate (close to 30 %) and high potential losses (a mean annualized profit rate of -0.025 with a high variability). When small businesses use P2P lending, they look to an alternative market to cover their

financial needs not satisfied by traditional markets due to risk conditions. Our methods have proven useful for selecting the most promising financial obligations regarding risk and profit in this setting. However, the average return and loan acceptance rates are very low. This fact reinforces the idea that the intervention of the state or investors who wish to support this type of proposition is needed. Small businesses tend to have low returns below the interest rates charged by traditional financing alternatives, increasing their risk of default. Thus, offering financing tools to support the productive segment of small businesses becomes vital. While p2p lending can be a financing alternative for some small businesses, the segment demands its own regulatory and leverage conditions and tailor-made application forms asking for specific information about the business object of the loan. Of course, the segment would also benefit from using profit-based approaches like the ones proposed in this paper. In this way, it will be possible to offer small businesses better financing alternatives with fairer rates and, in turn, serve as an investment and support scenario for those who wish to leverage new small business ventures. All this would contribute to mitigating the gap between the demand and supply of financial products and services for small businesses, which is one of the implicit objectives of the United Nations SDGs.

Finally, from a classification perspective, as [Sun et al. \(2009\)](#) points out for class-imbalance problems, financial default is a concept with such a degree of complexity that the patterns of each class in the feature space may overlap, and discriminatory rules become difficult to induce. However, our work has shown that it is possible to successfully combine the search for risk minimization and high profitability in one single model, resulting in better financial and risk management. Furthermore, our work demonstrates that explainability tools can help to understand better the main drivers of risk-weighted by profit, appreciate the changes with respect to profit-insensitive alternatives and find good profiles of loans with business sense. Although we worked in the case of small businesses in P2P lending, the approach presented in our paper can be easily reproduced in other loan markets or in any classification problem where misclassification errors have an economic impact, helping for

better risk management by accounting for the financial impact of the decision.

Capítulo 5

Discusión y Conclusiones

Deseo empezar este apartado comentando que las ideas seminales de este trabajo han estado motivadas por las reflexiones que obtuve de mi participación como investigador y capacitador en el proyecto Europeo Fin-Tech Risk Management H2020, derivado de mi proceso de formación doctoral. Este proyecto estuvo compuesto por Universidades y centros de investigación de toda Europa, FinTechs, asociaciones, reguladores y supervisores, en donde representábamos a España a través de la Universidad Complutense de Madrid, en temas de Big Data Analytics, Artificial Intelligence y Blockchain.

Tuvimos comunicación directa con los reguladores del país (Banco de España, la Comisión Nacional del Mercado de Valores y la Dirección General de Seguros y Fondos de Pensiones), en donde la demanda de las cualidades de explicabilidad y el sentido de negocio en los modelos de gestión del riesgo de crédito se hizo explícita, principalmente al usar herramientas de aprendizaje automático. Por lo tanto, los objetos de estudio de nuestra propuesta surgen como ejemplo de la interacción entre fintechs, reguladores y académicos. Los primeros con conocimiento de las nuevas oportunidades de negocio basadas en el desarrollo tecnológico, los segundos con experiencia y conocimiento sobre las oportunidades y limitaciones asociadas con la normativa, y los últimos, con conocimiento detallado, reflexionado, profundo y consciente de las oportunidades y limitaciones de los modelos de gestión de riesgos que pueden aplicarse en los contextos de innovación tecnológica y financiera. Este

manera de construir es poco frecuente y visible en los artículos de investigación estudiados, como lo identificamos y reflexionamos en el capítulo 2.

Hemos podido ver que reguladores y supervisores reconocen el potencial de las herramientas de aprendizaje automático, pero aún no les genera la suficiente confianza por su falta de transparencia y explicabilidad en sus versiones más complejas, por lo que se inclinan por mantener los modelos tradicionales.

En la misma línea, la revisión de literatura de los artículos sobre el mercado P2P que mostramos en el capítulo 2, no solo sirvió para identificar las principales tendencias de investigación en la gestión del riesgo, sino que también pudimos descubrir pocos esfuerzos por añadir interpretabilidad y explicabilidad a los modelos propuestos, lo que dificultaba su adopción. No había una preocupación evidente por responder a las necesidades de transparencia e interpretabilidad que demandan los reguladores y la industria. El foco de las propuestas se centraba principalmente en la obtención de indicadores de rendimiento en la predicción, y escasamente se analizaba la importancia relativa de las variables. Del contexto y la problemática, apenas se presentaban las variables sin analizar su aplicabilidad y uso en contextos reales, afectando el sentido de negocio de las propuestas. La falta de inclusión de estos elementos reducía los niveles de transparencia y confiabilidad.

En particular, en el diseño de los experimentos de nuestras propuestas, abordamos estos componentes con rigor, como veremos en los siguientes apartados.

5.1. En relación con el diseño de los experimentos

Abordamos el diseño de los experimentos atendiendo a aspectos como el conjunto de datos, la definición de la variable objetivo y las variables de entrada, modelos utilizados, elementos de evaluación como la estrategia y las medidas de desempeño, así como los métodos de búsqueda de hiperparámetros, entre otros aspectos. También, resaltamos el componente de explicabilidad en los modelos propuestos, como parte del diseño. De estos elementos daremos detalles a continuación.

Para nuestros artículos aplicados (capítulos 3 y 4), utilizamos como conjunto de datos, registros reales del problema de otorgamiento de crédito de una de las empresas que impulsó el desarrollo del mercado de P2P en Estados Unidos y en el mundo, Lending Club. Como se vio en la revisión de literatura, capítulo 2, es el más usado en investigaciones de este mercado P2P y se considera que tiene una calidad suficiente para propuestas de modelización. Para el segundo artículo, capítulo 3, hemos utilizado todos los préstamos disponibles en el conjunto de datos. Y para el tercero, capítulo 4, nos hemos concentrando únicamente en los préstamos a empresas pequeñas, uno de los segmentos con mayor necesidad de financiación, más complejo y de mayor riesgo por su escasez de información y asimetrías de la misma.

Nuestras propuestas empíricas corresponden con modelos para la concesión de créditos, por lo que definimos nuestra variable objetivo, categórica, en función de la resolución final del crédito, con las opciones impago o no de la obligación. La categoría de impago corresponde al evento castigado o declarado como pérdida, y la categoría de no impago al evento totalmente pagado. Esta variable representa el estado del crédito al final de la ventana temporal considerada, por lo que solo utilizamos registros con esta información.

Como variables de entrada, solo utilizamos información disponible en el momento de la solicitud. Variables como el tipo de interés, el grado de riesgo o el subgrado de riesgo son generados por las entidades como resultado de un proceso de evaluación de riesgo previo, y no deben tenerse en cuenta en los modelos para predecir el impago en la concesión de créditos desde el otorgamiento. Son variables muy relacionadas con la variable objetivo, su inclusión genera sesgos dándoles importancia errónea a estas variables, y no genera modelos realmente aplicables. Aunque hacer esta selección previa de variables limita las posibilidades de entrada de información, garantiza construir con mayor sentido de negocio y el uso potencial, real, de estas propuestas.

Nuestras propuestas de modelización se han basado en técnicas de aprendizaje automático frecuentemente usadas en problemas de otorgamiento, en las que el método XGBoost suele ser preferido por sus altos indicadores de desempeño. En artículo del capítulo 3 utili-

zamos este método junto con otras alternativas de aprendizaje automático en comparación con la regresión logística, preferido por reguladores y usuarios por su simplicidad, facilidad de comprensión y transparencia. En artículo del capítulo 4, dado los buenos resultados del modelo XGBoost derivados de nuestro artículo previo, comparamos su poder predictivo frente a alternativas que involucran el sentido de negocio en el modelo.

En términos predictivos hemos realizado la comparación en función de métricas bien conocidas y frecuentemente usadas en riesgo de crédito como la exactitud (accuracy en inglés), el área bajo la curva (o AUC en sus siglas en inglés), el indicador de Kolmogorov-Smirnov o el indicador F1 sobre la clase de interés, aunque también incorporamos métricas por clase como la precisión y exhaustividad (precision y recall en inglés) para la clases de préstamos impagados y pagados. Además, hemos utilizado la medida de exactitud balanceada (balanced accuracy en inglés), que mide la exactitud media obtenida tanto de las clases mayoritarias como de las minoritarias cuando hay desequilibrio de clases, equiparando el peso al aumentar la importancia de la clase minoritaria (incumplimiento), la cual está más asociada con el riesgo. Se pretende así, evitar el sesgo mostrado por la exactitud. Para nuestro último artículo hemos incluido métricas financieras con base en la rentabilidad media anualizada por obligación, y también indicadores relevantes para el negocio como son el número de préstamos concedidos y la tasa de impagos sobre los préstamos concedidos.

Respecto al proceso de búsqueda de hiperparámetros, por lo que hemos utilizado dos herramientas: algoritmos genéticos para el primer artículo aplicado (capítulo 3) y optimización bayesiana para el segundo, (capítulo 4). Como métrica objetivo utilizamos la exactitud balanceada en el primero, y en el segundo, utilizamos una que está basada en los beneficios esperados anualizados con base en la información financiera por obligación.

La búsqueda de hiperparámetros la realizamos utilizando una estrategia de validación cruzada (k-folds) estratificada, con el fin de evaluar la estabilidad y eficiencia de la predicción, y preservar la distribución de las categorías de la variable objetivo en cada selección. Además, para la comparación también usamos esta estrategia de evaluación incorporando pruebas

estadísticas inferenciales, componente que aunque necesario e informativo, no se acostumbra a incluir en las publicaciones en este campo, como se destacó en el artículo 1, (capítulo 2).

Respecto al problema de desequilibrio de clases, en el capítulo 3 se aborda para el caso del modelo base de regresión logística mediante remuestreo (undersampling y oversampling), y se incorpora un parámetro que ayuda a ponderar los errores de la clase minoritaria en los modelos de aprendizaje automático. Para el segundo, capítulo 4, se incorpora información financiera en la valoración de los errores y aciertos de clasificación, más detalles a continuación.

Para complementar y tener una propuesta integral de flujo de trabajo con modelos de aprendizaje automático, hemos incorporado la cualidad de explicabilidad mediante valores de SHAP, como ya se mencionó, utilizando herramientas visuales que permiten analizar la dependencia entre variables, su monotonidad, no linealidad y cambio estructural, además de identificar segmentos de mayor riesgo y de oportunidad en términos de negocio. De nuestra parte, ampliamos la noción de valores SHAP a la regresión logística, y ajustamos su cálculo para variables de entrada categóricas, con esto, se puede dar mejor cuenta de la dependencia entre categorías. Hemos comparado en el primer artículo aplicado la explicabilidad proporcionada por estas herramientas sobre el mejor modelo probado XGboost, con el tradicional de regresión logística. Aunque en la regresión logística podemos analizar parte de estos componentes a través de sus coeficientes, la oportunidad relativa o razón de ventajas (odds ratio en ingles) y los efectos marginales, no podemos observar diferencias a nivel de observación y de no linealidad que si proporcionan las herramientas post-hoc como los valores de SHAP. Es decir, con estas herramientas, aportamos capacidad explicativa a modelos simples y complejos de diversa naturaleza.

5.2. Principales Contribuciones

Adicional a los hallazgos y necesidades que hemos identificado en el capítulo 2, y que hemos resumido en la introducción de este apartado de conclusiones, 5, identificamos otras

aportaciones que resumimos a continuación.

A nivel teórico la principal aportación es la presentación de un modelo que incluye los beneficios, el sentido de negocio, en tres puntos del proceso; la definición de hiperparámetros, la función de aprendizaje y la decisión otorgamiento. Se comparó el desempeño de estas alternativas, utilizando como base el modelo XGBoost. Principalmente, incluimos una función objetivo sensible a los beneficios dentro del aprendizaje del modelo con base en el gradiente y la hessiana derivadas para cada instancia, en función de los valores financieros de las celdas de la matriz de clasificación. Nos basamos en una propuesta sensible a los beneficios, reconociendo los valores económicos de pérdida o beneficio de cada una de las celdas de la matriz de clasificación correcta y errónea de la predicción, también llamada matriz de confusión o de misclassification. Para la decisión, con base en las probabilidades de las propuestas, se compararon las técnicas de umbralización (thresholding) y Bayes Minimum Risk (BMR).

Desde el punto de vista predictivo, nuestro primer artículo aplicado, capítulo 3, muestra que el método XGBoost no solo es mejor que la regresión logística, sino que es estadísticamente mejor que otras alternativas. Además, como se muestra en nuestro segundo artículo aplicado, capítulo 4, si a este método se le incorpora el negocio a través de los beneficios directamente en la función de aprendizaje, genera mejores resultados financieros que al incluirlos en cualquier otra fase de la modelización, o que no incluirlos, en lo que llamamos modelos insensible a los beneficios. Cabe mencionar que la búsqueda de beneficios no se asocia necesariamente con una mejor clasificación, como se puede observar en los resultados del capítulo 4, ratificando lo encontrado en trabajos previos, pero sí permite la selección de las obligaciones más rentables. Se distinguen las bondades de nuestra propuesta; controlamos el riesgo mediante el modelo de predicción de incumplimiento, y garantizamos mejores niveles de rentabilidad al incluir la función de beneficios en los procesos de modelización.

Es interesante mencionar que las herramientas de explicabilidad nos permiten ver cambios sutiles y gruesos en nuestras propuestas, mostrando que el uso que se hace de las

variables es diferencial. Se aprecia que clasificar con sentido de negocio es un problema que puede ser muy distinto al de hacerlo solo buscando indicadores de clasificación. Variables como los ingresos, el endeudamiento, el score de buró, el tipo de vivienda aportan y presentan relaciones complejas con el riesgo, o con el riesgo ponderado por rentabilidad, de una manera diferente a como lo hacen en los modelos tradicionales, o a los modelos insensible a los beneficios, proporcionando así más elementos de negocio para tomar las decisiones de otorgamiento y de inversión.

La información sobre los beneficios sirve para determinar la contribución de las variables con mayor precisión, es decir, con menos variabilidad. Es decir, la consideración del beneficio reduce la incertidumbre del papel de la variable en los modelos. Se puede ver con las herramientas de explicabilidad, que considerar el beneficio nos ha ayudado a determinar un perfil claro de los préstamos más interesantes en términos de impago ponderado por la rentabilidad, específicamente, para el segmento de las pequeñas empresas.

Evidenciamos así, las bondades financieras de la inclusión de los beneficios en los procesos de modelización, y sus limitaciones en términos de rentabilidad en el contexto de los préstamos a pequeñas empresas, un segmento de mayor riesgo. Desde el punto de vista de la clasificación, el problema de predicción del impago conlleva problemas de desequilibrio de clases, y en contextos como los préstamos a pequeñas empresas se halla tal grado de complejidad que los patrones de cada clase en el espacio de características pueden solaparse, y resulta difícil inducir reglas discriminatorias. Nuestro trabajo ha demostrado que es posible combinar con éxito la búsqueda de la minimización del riesgo de crédito inducido por rentabilidad en un único modelo de clasificación, lo que se traduce en una mejor gestión financiera y del riesgo. Además, con las herramientas de explicabilidad, logramos comprender mejor los principales impulsores del riesgo de crédito ponderado por beneficios, apreciar los cambios con respecto a alternativas insensibles a los beneficios y encontrar mejores perfiles de préstamos con sentido empresarial.

5.3. Oportunidades y trabajo futuro

A continuación describiremos algunas oportunidades de investigación que se derivan de nuestro trabajo.

Para empezar hablaremos del conjunto de datos y la ventana de observación utilizada. Vemos oportunidad en poder ampliar la ventana de observación y explorar nuevos conjuntos de datos, también usados en la literatura, ver el capítulo 2. De esta manera, podemos ratificar la aplicabilidad de nuestras propuestas, extrayendo y comparando la importancia y forma funcional de las variables incluidas, como ingresos, endeudamiento, monto solicitado, tipo de vivienda, antigüedad del empleo, experiencia financiera previa, y otras variables relevantes. Por supuesto, aunque esta tesis utiliza un conjunto de préstamos P2P, la propuesta puede aplicarse a préstamos otorgados por entidades financieras tradicionales.

Con relación al problema financiero, adicional al otorgamiento de crédito, podemos extrapolar nuestra propuesta a otros problemas poco estudiados en el mercado P2P, como el fraude, el cobro de deudas, la estimación del LGD, la predicción de la rentabilidad, entre otros, todos ellos necesarios para el sano desarrollo del mercado.

Desde el punto de vista metodológico, podemos estudiar la inclusión del componente de explicabilidad y sentido de negocio con propuestas sensibles a los beneficios, profit-sensitive, pero, estudiando y controlando por sesgos presentes como la inferencia de rechazo (reject inference), es decir, el sesgo que supone no conocer y tratar con las solicitudes rechazadas en la modelación. Muchos trabajos, como el nuestro, basan los resultados en datos ya filtrados (créditos otorgados con un análisis previo de riesgo), lo que puede afectar la medición real del rendimiento de los modelos y limitar su uso. La inferencia del rechazo es un problema de estimación de contrafactual que podría proporcionar elementos interesantes el controlarlo e incluirlo.

Otra oportunidad se encuentra en el contexto de desequilibrio de clases, en donde se dificulta la obtención precisa de probabilidades para los métodos de aprendizaje automático, lo que demanda de técnicas de calibración para obtener mejores procesos de decisión. Es un

tópico poco estudiado en contextos empresariales y vale la pena incorporarlo en propuestas sensibles a los beneficios que dependen de la probabilidad, como la nuestra, para ver su impacto en las métricas de desempeño y en la decisión del otorgamiento de créditos.

De otro lado, nosotros hemos propuesto una metodología que muestra las ventajas de incluir riesgo y beneficio en un único proceso de modelización. Sin embargo, existen otros trabajos que abordan estos dos componentes por separado, con un modelo para cada objetivo (Lyócsa et al., 2022; Serrano-Cinca & Gutierrez-Nieto, 2016), o por etapas, con un modelo de predicción de impago y posterior selección de las obligaciones más rentables (Bastani et al., 2019; Xia et al., 2017a). Lo anterior muestra que existen varias alternativas para gestionar la dicotomía riesgo-rentabilidad. Esto evidencia que es necesario investigar más a fondo las formas más aconsejables de abordar este dúo financiero en los modelos de préstamos mediante estudios comparativos.

Con respecto al componente de explicabilidad, existe un gran interés por seguir estudiándolo como componente en la modelización, y ofrecer técnicas de explicaciones de alternativas para las nuevas propuestas. Los modelos de aprendizaje automático exploran y reconocen relaciones subyacentes mucho más complejas que los métodos tradicionales, y áreas de investigación como el XAI tienen como propósito incorporarle explicabilidad, interpretabilidad y transparencia a la inteligencia artificial, estudiando los procesos y métodos que permiten a los usuarios humanos comprender y confiar en los resultados y productos creados por los algoritmos. En el contexto de modelos de gestión del riesgo de crédito, aunque claramente XAI es un área de estudio relevante y pertinente que ha ganado fuerza en la última década, aún merece dedicación por parte de investigadores, usuarios de los modelos de aprendizaje automático, y de los reguladores. Es necesario llegar a acuerdos conceptuales, construir algunas taxonomías transversales que unifiquen el lenguaje, líneas de estudio definidas, un marco que unifique estándares, requisitos y criterios, y mayor desarrollo. Atributos necesarios para incorporar formalmente la cualidad de explicabilidad como un requerimiento exigido para los modelos por lo entes reguladores.

En específico, para herramientas como los valores de SHAP, se pueden aplicar nuevos desarrollos que complementan su aporte. Por ejemplo, existen desarrollo en contextos de dependencia de las variables ([Aas et al., 2021](#)), estudiando relaciones de causalidad ([Albini et al., 2022](#); [Chou et al., 2022](#)) e incorporando inferencia estadística ([Joseph, 2019](#)). También, se puede explorar la inclusión de análisis de sensibilidad e impulso respuesta, como se puede realizar a través de los métodos tradicionales y sus coeficientes.

Precisamente, el concepto de causalidad es uno de los que ha tomado mayor relevancia con el desarrollo del XAI, algunos investigadores han argumentado recientemente que, para que un modelo de aprendizaje automático alcance el nivel de explicabilidad humana, debe proporcionar explicaciones humanas causalmente comprensibles ([Chou et al., 2022](#)), aunque esta es una necesidad que exhiben también los modelos tradicionales. Podemos encontrar desarrollos en el campo de casualidad del aprendizaje automático y de la inteligencia artificial en ([Janzing et al., 2020](#); [Rawal et al., 2023](#); [Schölkopf, 2022](#)).

De otro lado, encontramos corrientes que rechazan las herramientas post-hoc y prefieren la búsqueda de modelos intrínsecamente interpretables ([Rudin, 2019](#); [Rudin et al., 2022](#)). Estos últimos también pueden estudiarse y compararse con aproximaciones como las que propone esta tesis doctoral.

5.4. Conclusiones finales

Las carencias detectadas en la literatura en cuanto a definición del experimento con sentido de negocio y de interpretabilidad, generaban dudas sobre si el aprendizaje automático era capaz de satisfacer las demandas de los reguladores y usuarios. Sin embargo, nuestros resultados muestran que este enfoque de modelización no sólo puede generar mejores resultados en términos de desempeño en la predicción, sino que también arrojan luces sobre la complejidad del problema. Se detecta no solo la contribución relativa de las variables, sino también relaciones no lineales, cambios en la dispersión y rupturas estructurales, que no pueden reflejarse fácilmente con la regresión logística o con otros métodos tradicionales.

También, evidenciamos las bondades financieras de la inclusión de los beneficios en los procesos de modelización, junto con sus limitaciones en términos de rentabilidad en el contexto de los préstamos a pequeñas empresas, un segmento de mayor riesgo.

Las herramientas de explicabilidad nos permiten ver cómo cambia el rol de las variables frente a los métodos tradicionales, insensibles a los costos y a los beneficios. Se puede comprender qué factores impulsan los préstamos rentables a pequeñas empresas, y cómo difieren de los que impulsan el problema de clasificación tradicional en el riesgo de crédito.

El uso del aprendizaje automático y de las nuevas tecnologías de ciencia de datos ayuda a ofrecer productos más adecuados y adaptados a las necesidades de los actores de mercados como el P2P. Por ejemplo, para los prestamistas estas soluciones les ayudan a protegerse contra el riesgo crediticio y el fraude, reducir el costo de los procesos de evaluación crediticia, focalizar la distribución y el monitoreo de productos y permitir una mejor interacción de los agentes del mercado. Con estas soluciones los prestatarios tienen acceso a productos financieros, con menores costos, tramites, y con tiempos de respuesta más oportunos.

El aprendizaje automático ayuda con la gestión del riesgo asociado con el otorgamiento de crédito en mercados complejos, ávidos de herramientas que impulsen su desarrollo, como lo es, el mercado de préstamos p2p para pequeñas empresas, tal como lo evidenciamos en una de nuestros artículos.

Aunque nuestra propuesta se asocia con préstamos en el caso de pequeñas empresas o con el problema de otorgamiento de crédito en el mercado P2P, el enfoque presentado puede reproducirse fácilmente en otros mercados de préstamos o en cualquier problema de predicción categórica en el que los errores de clasificación tengan un impacto económico, y deseemos incluirlos para mejorar la decisión de los procesos de modelización. Lo que finalmente presentamos es un flujo de trabajo integral para problemas de clasificación que tenga en cuenta la predicción, el sentido en contexto y su interpretabilidad, coadyuvando a su transparencia, confiabilidad y aplicación. En general, se aborda con rigor un problema de clasificación con clases desbalanceadas, se aprecia cómo introducirle costes y beneficios

en el aprendizaje, se estudia el aporte de las variables en contextos complejos, lo que es extrapolable a cualquier problema de clasificación, pudiéndose obtener resultados reseñables no solamente en dominios financieros.

Al margen de la importancia que tuvo la interacción con los reguladores para esta tesis, consideramos de vital importancia que se abran más espacios de trabajo conjunto entre estos actores, fundamentales en la construcción de conocimiento, potenciando el desarrollo de un marco diferencial para la gestión del riesgo y de la rentabilidad de los nuevos productos y de las alternativas de financiación basados en el creciente avance tecnológico. Creemos que esta estrategia permite presentar soluciones mucho más adaptadas al contexto, generando una mayor transparencia, confianza y probabilidad de uso. Esta sinergia garantizaría una mejor comprensión de la naturaleza y la estructura del negocio, lo que conduciría a una mejor modelización, gestión de riesgos y servicio, promoviendo su sostenibilidad. De este modo, la investigación ayudará a desarrollar un mercado apoyado en la digitalización de las finanzas, con mayor transparencia, explicabilidad, cobertura e inclusión financiera, con la oferta de mejores productos y servicios en beneficio de particulares, hogares y empresas.

Bibliografía

- Aas, K., Jullum, M., & Løland, A. (2019). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *arXiv e-prints*, (p. arXiv:1903.10464).
- URL <https://arxiv.org/abs/1903.10464>
- Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298, 103502.
- URL <https://www.sciencedirect.com/science/article/pii/S0004370221000539>
- Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting, Finance and Management*, 18(2-3), 59–88.
- Ahelegbey, D. F., Giudici, P., & Hadji-Misheva, B. (2019). Latent factor models for credit scoring in P2P systems. *Physica a-Statistical Mechanics and Its Applications*, 522, 112–121.
- Akila, S., & Reddy, S. U. (2018). Cost-sensitive Risk Induced Bayesian Inference Bagging (RIBIB) for credit card fraud detection. *Journal of Computational Science*, 27, 247–254.
- Albini, E., Long, J., Dervovic, D., & Magazzeni, D. (2022). Counterfactual shapley additive explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, (pp. 1054–1070).
- Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable

- artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99, 101805.
- Almhaithawi, D., Jafar, A., & Aljnidi, M. (2020). Example-dependent cost-sensitive credit cards fraud detection using SMOTE and Bayes minimum risk. *SN Applied Sciences*, 2(9).
- Alonso Robisco, A., & Carbo Martinez, J. M. (2022). Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction. *Financial Innovation*, 8(1), 70.
- Amalia, N., Dalimunthe, Z., & Triono, R. A. (2019). *The Effect of Lender's Protection on Online Peer-to-Peer Lending in Indonesia*.
- Andriosopoulos, D., Doumpos, M., Pardalos, P. M., & Zopounidis, C. (2019). Computational approaches and data analytics in financial services: A literature review. *Journal of the Operational Research Society*, 70(10), 1581–1599.
- Apley, D. W., & Zhu, J. (2016). Visualizing the effects of predictor variables in black box supervised learning models. *arXiv e-prints*, (p. arXiv:1612.08468).
URL <https://ui.adsabs.harvard.edu/abs/2016arXiv161208468A>
- Aria, M. and Cuccurullo, C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics. Elsevier.*, 11(4), 959–975.
- Ariza-Garzón, M. J., Arroyo, J., Caparrini, A., & Segovia-Vargas, M. (2020). Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending. *IEEE Access*, 8, 64873–64890.
- Ariza-Garzón, M.-J., Camacho-Miñano, M.-D.-M., Segovia-Vargas, M.-J., & Arroyo, J. (2021). Risk-return modelling in the p2p lending market: Trends, gaps, recommendations and future directions. *Electronic Commerce Research and Applications*, 49, 101079.

- Aziz, S., & Dowling, M. (2019). Machine learning and AI for risk management. *Disrupting finance: FinTech and strategy in the 21st century*, (pp. 33–50).
- Babel, B., Buehler, K., Pivonka, A., Richardson, B., & Waldron, D. (2019). Derisking machine learning and artificial intelligence. *McKinsey Company*, (p. 7).
 URL <https://www.mckinsey.com/business-functions/risk/our-insights/derisking-machine-learning-and-artificial-intelligence>
- Bachmann, A., Becker, A., Buerckner, D., Hilker, M., Kock, F., Lehmann, M., Tiburtius, P., & Funk, B. (2011). Online peer-to-peer lending-a literature review. *Journal of Internet Banking and Commerce*, 16(2), 1–18.
- Bae, J. K. (2018). A Study on the Determinant Factors of P2P Loans and Activation Factors of P2P Lending Market TT - P2p. *Logos Management Review*, 16(2), 21–36.
- Bahnsen, A. C., Aouada, D., & Ottersten, B. (2014). Example-dependent cost-sensitive logistic regression for credit scoring. In *2014 13th International Conference on Machine Learning and Applications*, (pp. 263–269).
- Bahnsen, A. C., Villegas, S., Aouada, D., & Ottersten, B. (2015). Example-dependent cost-sensitive decision trees. *Expert Systems with Applications*, 42(19), 6609 – 6619.
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115.
 URL <http://www.sciencedirect.com/science/article/pii/S1566253519308103>
- Basel Committee on Banking Supervision (2017). *Basel III: Finalising post-crisis reforms*.
 URL <https://www.bis.org/bcbs/publ/d424.pdf>

- Basha, S. A., Elgammal, M. M., & Abuzayed, B. M. (2021). Online peer-to-peer lending: A review of the literature. *Electronic Commerce Research and Applications*, 48, 101069.
- Bastani, K., Asgari, E., & Namavari, H. (2019). Wide and deep learning for peer-to-peer lending. *Expert Systems with Applications*, 134, 209–224.
- Bazarbash, M. (2019). *Fintech in financial inclusion: Machine learning applications in assessing credit risk*. International Monetary Fund.
 URL <https://www.imf.org/en/Publications/WP/Issues/2019/05/17/FinTech-in-Financial-Inclusion-Machine-Learning-Applications-in-Assessing-Credit-Risk-46883>
- Bergstra, J., Yamins, D., & Cox, D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, (pp. 115–123). PMLR.
- Bernardi, M., & Catania, L. (2018). The model confidence set package for R. *International Journal of Computational Economics and Econometrics*, 8(2), 144–158.
- Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*, 1(1), 363–376.
 URL https://doi.org/10.1162/qss_a_00018
- Board, F. S. (2017). Artificial intelligence and machine learning in financial services. Market developments and financial stability implications. *Financial Stability Board*, (p. 45).
 URL <https://www.fsb.org/wp-content/uploads/P011117.pdf>
- Boiko Ferreira, L. E., Barddal, J. P., Enembreck, F., & Gomes, H. M. (2017). *Improving Credit Risk Prediction in Online Peer-to-Peer (P2P) Lending Using Imbalanced Learning Techniques*, (pp. 175–181). Proceedings-International Conference on Tools With Artificial Intelligence.

- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth Statistics/Probabilitys. Chapman and Hall/CRC; Edición: 1.
- Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in fintech risk management. *Frontiers in Artificial Intelligence*, 3, 26.
- Byanjankar, A. (2017). *Predicting Credit Risk in Peer-to-Peer Lending with Survival Analysis*.
- Byanjankar, A., Heikkila, M., & Mezei, J. (2015). Predicting Credit Risk in Peer-to-Peer Lending: A Neural Network Approach. In *2015 IEEE Symposium Series on Computational Intelligence*, (pp. 719–725).
- Cai, S., & Zhang, J. (2020). Exploration of credit risk of P2P platform based on data mining technology. *Journal of Computational and Applied Mathematics*, 372.
- Calabrese, R., Osmetti, S. A., & Zanin, L. (2019). A joint scoring model for peer-to-peer and traditional lending: a bivariate model with copula dependence. *Journal of the Royal Statistical Society Series a-Statistics in Society*, 182(4), 1163–1188.
- Canfield Rivera, C. E. (2018). Determinants of default in P2P lending: The mexican case. *Independent Journal of Management Production*, 9(1), 1–24.
- Capasso, G., Gianfrate, G., & Spinelli, M. (2020). Climate change and credit risk. *Journal of Cleaner Production*, 266, 121634.
- Carpentier, C. L., & Braun, H. (2020). Agenda 2030 for sustainable development: A powerful global framework. *Journal of the International Council for Small Business*, 1(1), 14–23.

- Carr, B., & Bailey, N. (2018). Explainability in predictive modeling. *Machine Learning Thematic Series, Institute of International Finance, I*, 28.
- URL https://www.iif.com/portals/0/Files/private/32370132_machine_learning_explainability_nov_2018.pdf
- Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8).
- Chang, A.-H., Yang, L.-K., Tsaih, R.-H., & Lin, S.-K. (2022). Machine learning and artificial neural networks to construct p2p lending credit-scoring model: A case using lending club data. *Quantitative Finance and Economics*, 6(2), 303–325.
- Chen, B., Zeng, W., & Lin, Y. (2014). Applications of Artificial Intelligence Technologies in Credit Scoring: a Survey of Literature. In *2014 10th International Conference on Natural Computation (ICNC)*, Proceedings International Conference on Natural Computation, (pp. 658–664).
- Chen, C., Dong, M. C., Liu, N., & Sriboonchitta, S. (2019a). Inferences of default risk and borrower characteristics on P2P lending. *North American Journal of Economics and Finance*, 50.
- Chen, C., Lin, K., Rudin, C., Shaposhnik, Y., Wang, S., & Wang, T. (2018). An interpretable model with globally consistent explanations for credit risk. *arXiv e-prints*, (p. arXiv:1811.12615).
- URL <https://ui.adsabs.harvard.edu/abs/2018arXiv181112615C>
- Chen, D., Li, X., & Lai, F. (2016a). Gender discrimination in online peer-to-peer credit lending: evidence from a lending platform in China. *Electronic Commerce Research*, 17(4), 553–583.
- Chen, M. A., Wu, Q., & Yang, B. (2019b). How valuable is fintech innovation? *The Review of Financial Studies*, 32(5), 2062–2106.

- Chen, N., Ribeiro, B., & Chen, A. (2016b). Financial credit risk assessment: a recent review. *Artificial Intelligence Review*, 45, 1–23.
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (pp. 785–794). New York, NY, USA: ACM.
- Chen, Y. (2017). Research on the Credit Risk Assessment of Chinese Online Peer-to-peer Lending Borrower on Logistic Regression Model. In *3rd Asian Pacific Conference on Energy, Environment and Sustainable Development*, DEStech Transactions on Environment Energy and Earth Sciences, (pp. 216–221).
- Cho, P., Chang, W., & Song, J. W. (2019). Application of instance-based entropy fuzzy support vector machine in peer-to-peer lending investment decision. *IEEE Access*, 7, 16925–16939.
- Chou, Y.-L., Moreira, C., Bruza, P., Ouyang, C., & Jorge, J. (2022). Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, 81, 59–83.
- Claessens, S., Frost, J., Turner, G., & Zhu, F. (2018). Fintech credit markets around the world: size, drivers and policy issues. *BIS Quarterly Review*.
URL https://www.bis.org/publ/qtrpdf/r_qt1809e.pdf
- Clarivate (2021). Web of science journal evaluation process and selection criteria.
URL <https://clarivate.com/webofsciencegroup/journal-evaluation-process-and-selection-criteria/>
- Correa Bahnsen, A., Villegas, S., Aouada, D., & Ottersten, B. (????). Fraud Detection by Stacking Cost-Sensitive Decision Trees. In *Security Science and Technology*, (pp. 251–266). World Scientific.

- Cummins, M., Lynn, T., Mac an Bhaird, C., & Rosati, P. (2019). Addressing Information Asymmetries in Online Peer-to-Peer Lending. In *Disrupting Finance*.
- de Araújo Lima, P. F., Crema, M., & Verbano, C. (2020). Risk management in smes: A systematic literature review and future directions. *European Management Journal*, 38(1), 78–94.
- Demir, A., Pesqué-Cela, V., Altunbas, Y., & Murinde, V. (2022). Fintech, financial inclusion and income inequality: a quantile regression approach. *The European Journal of Finance*, 28(1), 86–107.
- Ding, H., Zhang, P., Lu, T., Gu, H., & Gu, N. (2017). *Credit Scoring Using Ensemble Classification Based on Variable Weighting Clustering*.
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv e-prints*, (p. arXiv:1702.08608).
URL <https://ui.adsabs.harvard.edu/abs/2017arXiv170208608D>
- Duan, J. (2019). Financial system modeling using deep neural networks (DNNs) for effective risk assessment and prediction. *Journal of the Franklin Institute-Engineering and Applied Mathematics*, 356(8), 4716–4731.
- Dupont, L., Fliche, O., & Yang, S. (2020). Governance of artificial intelligence in finance. *ACPR Banque De France*.
URL <https://acpr.banque-france.fr/en/governance-artificial-intelligence-finance>
- Durovic, A. (2017). Estimating probability of default on peer to peer market - survival analysis approach. *Journal of Central Banking Theory and Practice*, 6(2), 149–167.
- Elahi, E. (2013). Risk management: the next source of competitive advantage. *Foresight*, 15(2), 117–131.

- Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online peer-to-peer (P2P) lending. *Applied Economics*, 47(1), 54–70.
- Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets*, vol. 10. Springer.
- Feyen, E., Frost, J., Gambacorta, L., Natarajan, H., Saal, M., et al. (2021). Fintech and the digital transformation of financial services: Implications for market structure and public policy. *BIS Papers*.
- URL <https://www.bis.org/publ/bppdf/bispap117.htm>
- Financial Stability Board (2017). Artificial intelligence and machine learning in financial services. Market developments and financial stability implications. Tech. rep.
- URL [internal-pdf://0180047707/P011117.pdfhttps://www.fsb.org/wp-content/uploads/P011117.pdf](https://www.fsb.org/wp-content/uploads/P011117.pdf)
- Fisher, A., Rudin, C., & Dominici, F. (2018). All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously. *arXiv e-prints*, (p. arXiv:1801.01489).
- URL <https://ui.adsabs.harvard.edu/abs/2018arXiv180101489F>
- Flint, A., Nourian, A., & Koister, J. (2018). xAI Toolkit: Practical, Explainable Machine Learning. Tech. rep., FICO Decisions.
- URL https://www.fico.com/sites/default/files/2018-06/FICO_xAI_Toolkit-Practical_Explainable_Machine_Learning_4547WP_EN.pdf
- Florez-Lopez, R., & Ramon-Jeronimo, J. M. (2015). Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. a correlated-adjusted decision forest proposal. *Expert Systems with Applications*, 42(13), 5737–5753.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.

- Fu, X., Zhang, S., Chen, J., Ouyang, T., & Wu, J. (2019). A Sentiment-Aware Trading Volume Prediction Model for P2P Market Using LSTM. *IEEE Access*, 7, 81934–81944.
- Fuster, A., Plosser, M., Schnabl, P., & Vickery, J. (2019). The Role of Technology in Mortgage Lending. *The Review of Financial Studies*, 32(5), 1854–1899.
URL <https://doi.org/10.1093/rfs/hhz018>
- Gao, Y., Sun, J., & Zhou, Q. (2017). Forward looking vs backward looking An empirical study on the effectiveness of credit evaluation system in China’s online P2P lending market. *China Finance Review International*, 7(2), 228–248.
- Gao, Y., Yu, S., Chen, M., & Shiue, Y. (2020). A 2020 perspective on “The performance of the P2P finance industry in China”. *Electronic Commerce Research and Applications*.
- Ge, R., Feng, J., Gu, B., & Zhang, P. (2017). Predicting and Deterring Default with Social Media Information in Peer-to-Peer Lending. *Journal of Management Information Systems*, 34(2), 401–424.
- Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An approach to evaluating interpretability of machine learning. *arXiv pre-print arXiv:1806.00069*, (p. 118).
URL <https://arxiv.org/abs/1806.00069>
- Giudici, P., Hadji-Misheva, B., & Spelta, A. (2020). Network based credit risk models. *Quality Engineering*, 32(2), 199–211.
- Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1), 44–65.
- Gong, R., Xue, J., Zhao, L., Zolotova, O., Ji, X., & Xu, Y. (2019). A bibliometric analysis

- of green supply chain management based on the Web of Science (WOS) platform. *Sustainability*, 11(12), 3459.
- Gorzalczany, M. B., & Rudzinski, F. (2016). A multi-objective genetic optimization for fast, fuzzy rule-based credit classification with balanced accuracy and interpretability. *Applied Soft Computing*, 40, 206–220.
- Gourieroux, C., & Lu, Y. (2019). Least impulse response estimator for stress test exercises. *Journal of Banking & Finance*, 103, 62–77.
- Grace, M. F., Leverty, J. T., Phillips, R. D., & Shimpi, P. (2015). The value of investing in enterprise risk management. *Journal of Risk and Insurance*, 82(2), 289–316.
- Greiner, M. E., & Wang, H. (2010). Building Consumer-to-Consumer Trust in E-Finance Marketplaces: An Empirical Analysis. *International Journal of Electronic Commerce*, 15(2), 105–136.
- Guo, G., Zhu, F., Chen, E., Liu, Q., Wu, L., & Guan, C. (2016). From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring. *ACM Transactions on the Web*, 10(4).
- Guo, H. X., Li, Y. J., Shang, J., Gu, M. Y., Huang, Y. Y., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73, 220–239.
- Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2), 453–497.
- Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A. (2023). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, (pp. 1–30).

- Havrylchyk, O., & Verdier, M. (2018). The financial intermediation role of the p2p lending platforms. *Comparative Economic Studies*, 60(1), 115–130.
- Hayashi, Y., & Oishi, T. (2018). High accuracy-priority rule extraction for reconciling accuracy and interpretability in credit scoring. *New Generation Computing*, 36(4), 393–418.
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.
- Herasymovych, M., Märka, K., & Lukason, O. (2019). Using reinforcement learning to optimize the acceptance threshold of a credit scoring model. *Applied Soft Computing*, 84, 105697.
- Herzenstein, M., Sonenshein, S., & Dholakia, U. M. (2011). Tell Me a Good Story and I May Lend You Money: The Role of Narratives in Peer-to-Peer Lending Decisions. *Journal of Marketing Research*, 48, S138–S149.
- Holter, S., Gomez, O., & Bertini, E. (2020). FICO Explainable Machine Learning Challenge Creating visual explanations to black-box machine learning models. Tech. rep.
URL http://www.ml-explainer.com/static/images/FICO_paper.pdf
- Hudaefi, F. A. (2020). How does islamic fintech promote the SDGs? Qualitative evidence from Indonesia. *Qualitative Research in Financial Markets*, 12(4), 353–366.
- Ij, H. (2018). Statistics versus machine learning. *Nat Methods*, 15(4), 233.
- Invest In Club (2021). Peer to peer lending europe: Covid-19 impact on the market. Online; accessed 12 August 2021.
URL <https://crowdfunding-platforms.com/peer-to-peer-lending-europe-all-about-the-market>

- Janzing, D., Minorics, L., & Blöbaum, P. (2020). Feature relevance quantification in explainable AI: A causal problem. In *International Conference on artificial intelligence and statistics*, (pp. 2907–2916). PMLR.
- Ji, X., Yu, L., & Fu, J. (2020). Evaluating Personal Default Risk in P2P Lending Platform: Based on Dual Hesitant Pythagorean Fuzzy TODIM Approach. *Mathematics*, 8(1), 1–8. URL <http://dx.doi.org/10.3390/math8010008>
- Jiang, C., Wang, Z., & Zhao, H. (2019). A prediction-driven mixture cure model and its application in credit scoring. *European Journal of Operational Research*, 277(1), 20–31.
- Jiang, C. Q., Wang, Z., Wang, R. Y., & Ding, Y. (2018). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Annals of Operations Research*, 266(1-2), 511–529.
- Jin, Y., & Zhu, Y. (2015). *A data-driven approach to predict default risk of loan for online Peer-to-Peer (P2P) lending*, (pp. 609–613). International Conference on Communication Systems and Network Technologies.
- Joseph, A. (2019). Shapley regressions: A framework for statistical inference on machine learning models. Tech. Rep. 784. URL <https://www.bankofengland.co.uk/-/media/boe/files/working-paper/2019/shapley-regressions-a-framework-for-statistical-inference-on-machine-learning-models.pdf>
- Kgoroadira, R., Burke, A., & van Stel, A. (2019). Small business online loan crowdfunding: who gets funded and what determines the rate of interest? *Small Business Economics*, 52, 67–87.
- Kim, A., & Cho, S.-B. (2019a). An ensemble semi-supervised learning method for predicting defaults in social lending. *Engineering Applications of Artificial Intelligence*, 81, 193–199.

- Kim, A., & Cho, S.-B. (2019b). An ensemble semi-supervised learning method for predicting defaults in social lending. *Engineering Applications of Artificial Intelligence*, 81, 193–199.
- Kim, B., Khanna, R., & Koyejo, O. (2016). Examples are not Enough, Learn to Criticize! Criticism for interpretability. *Advances in Neural Information Processing Systems*, 29.
- Kim, J.-Y., & Cho, S.-B. (2019c). Predicting repayment of borrows in peer-to-peer social lending with deep dense convolutional network. *Expert Systems*, 36(4).
- Kim, J.-Y., & Cho, S.-B. (2019d). Towards repayment prediction in peer-to-peer social lending using deep learning. *Mathematics*, 7(11), 1041.
- Klein, G., Shtudiner, Z., & Zwillling, M. (2023). Why do peer-to-peer (P2P) lending platforms fail? The gap between P2P lenders’ preferences and the platforms’ intentions. *Electronic Commerce Research*, 23(2), 709–738.
- Koseoglu, M. A. (2016). Growth and structure of authorship and co-authorship network in the strategic management realm: Evidence from the Strategic Management Journal. *BRQ Business Research Quarterly*, 19 (3), 153–170.
- Krzanowski, W. J., & Hand, D. J. (2011). Testing the difference between two kolmogorov-smirnov values in the context of receiver operating characteristic curves. *Journal of Applied Statistics*, 38(3), 437–450.
- Kumar, V. L., Natarajan, S., Keerthana, S., Chinmayi, K. M., Lakshmi, N., & Ieee (2016). *Credit Risk Analysis in Peer-to-Peer Lending System*.
- Lahtasna, A., Ainon, R. N., & Teh, Y. W. (2010). Credit Scoring Models Using Soft Computing Methods: A Survey. *International Arab Journal of Information Technology*, 7(2), 115–123.
- Lee, Y.-W., Chen, S.-h., Yu, T., & Ieee (2017). Analysis of the Impact of Collateral on

- Peer-to-Peer Lending. In *2017 Ieee/Sice International Symposium on System Integration*, IEEE/SICE International Symposium on System Integration, (pp. 77–82).
- Leo, M., Sharma, S., & Maddulety, K. (2019). Machine learning in banking risk management: A literature review. *Risks*, 7(1).
- Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136.
- Li, H., Zhang, Y., Zhang, N., & Jia, H. (2016). Detecting the abnormal lenders from P2P lending data. In H. Lee, Y. Shi, J. Lee, F. Cordova, I. Dzitac, G. Kou, & J. Li (Eds.) *Promoting Business Analytics and Quantitative Management of Technology: 4th International Conference on Information Technology and Quantitative Management*, vol. 91 of *Procedia Computer Science*, (pp. 357–361).
- Li, L., Feng, Y., Lv, Y., Cong, X., Fu, X., & Qi, J. (2019). Automatically Detecting Peer-to-Peer Lending Intermediary Risk-Top Management Team Profile Textual Features Perspective. *IEEE Access*, 7, 72551–72560.
- Li, W., Ding, S., Chen, Y., & Yang, S. (2018a). Heterogeneous ensemble for default prediction of peer-to-peer lending in china. *IEEE Access*, 6, 54396–54406.
- Li, W., Ding, S., Wang, H., Chen, Y., & Yang, S. (2020). Heterogeneous ensemble learning with feature engineering for default prediction in peer-to-peer lending in China. *World Wide Web-Internet and Web Information Systems*, 23(1), 23–45.
- Li, Y., Hao, A., Zhang, X., & Xiong, X. (2018b). Network topology and systemic risk in Peer-to-Peer lending market. *Physica a-Statistical Mechanics and Its Applications*, 508, 118–130.

- Liguori, E., & Bendickson, J. S. (2020). Rising to the challenge: Entrepreneurship ecosystems and sdg success. *Journal of the International Council for Small Business*, 1(3-4), 118–125.
- Lin, W.-Y., Hu, Y.-H., & Tsai, C.-F. (2012). Machine Learning in Financial Crisis Prediction: A Survey. *IEEE Transactions on Systems Man and Cybernetics Part C-Applications And Reviews*, 42(4), 421–436.
- Lin, X. C., Li, X. L., & Zheng, Z. (2017). Evaluating borrower’s default risk in peer-to-peer lending: evidence from a lending platform in china. *Applied Economics*, 49(35), 3538–3545.
- Ling, C. X., & Sheng, V. S. (2010). *Cost-Sensitive Learning*, (pp. 231–235). Boston, MA: Springer US.
- Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43.
- Liu, C., & Yan, J. (2016). Researches on Risks and Precautions of Chinese P2P Lending. In M. Kuek, & R. Zhao (Eds.) *Proceedings of the 23rd International Business Annual Conference*, (pp. 731–736).
- Liu, H., Zhou, S., & Yang, W. (2019). Research on Intelligent Internet Financial Investment Model. In R. Su (Ed.) *2019 International Conference on Image and Video Processing, and Artificial Intelligence*, vol. 11321 of *Proceedings of SPIE*, (p. 113211P).
- Liu, Y., Zhou, Q., Zhao, X., & Wang, Y. (2018). Can Listing Information Indicate Borrower Credit Risk in Online Peer-to-Peer Lending? *Emerging Markets Finance and Trade*, 54(13), 2982–2994.
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2019). Explainable ai for trees: From local

explanations to global understanding. *arXiv preprint arXiv:1905.04610*.

URL <https://arxiv.org/abs/1905.04610>

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, 2(1), 56–67.

Lundberg, S. M., Erion, G. G., & Lee, S. (2018a). Consistent individualized feature attribution for tree ensembles. *CoRR*, *abs/1802.03888*.

URL <http://arxiv.org/abs/1802.03888>

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.) *Advances in Neural Information Processing Systems 30*, (pp. 4765–4774). Curran Associates, Inc.

URL <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K. W., Newman, S. F., Kim, J., & Lee, S. I. (2018b). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10), 749–760.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., & Lee, S.-I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10), 749–760.

Lyócsa, Š., Vašaničová, P., Hadji Misheva, B., & Vateha, M. D. (2022). Default or profit scoring credit systems? Evidence from European and US peer-to-peer lending markets. *Financial Innovation*, 8(1), 1–21.

- Ma, L., Zhao, X., Zhou, Z., & Liu, Y. (2018a). A new aspect on P2P online lending default prediction using meta-level phone usage data in China. *Decision Support Systems*, 111, 60–71.
- Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018b). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, 31, 24–39.
- Mach, T., Carter, C., & Slattery, C. (2014). Peer-to-peer lending to small businesses. Finance and Economics Discussion Series 2014-10, Board of Governors of the Federal Reserve System (U.S.).
URL <https://www.federalreserve.gov/pubs/feds/2014/201410/201410pap.pdf>
- Maldonado, S., Pérez, J., & Bravo, C. (2017). Cost-based feature selection for support vector machines: An application in credit scoring. *European Journal of Operational Research*, 261(2), 656–665.
- Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*, 42(10), 4621–4631.
- Maskara, P. K., Kuvvet, E., & Chen, G. (2021). The role of P2P platforms in enhancing financial inclusion in the United States: An analysis of peer-to-peer lending across the rural–urban divide. *Financial Management*, 50(3), 747–774.
- Masnadi-Shirazi, H., & Vasconcelos, N. (2011). Cost-sensitive boosting. *IEEE transactions on pattern analysis and machine intelligence*, 33(2), 294–309.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157.

- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38.
- Miller, T. (2023). Explainable AI is dead, long live explainable AI! Hypothesis-driven decision support. *arXiv preprint arXiv:2302.12389*.
URL <https://arxiv.org/abs/2302.12389>
- Mills, K., & McCarthy, B. (2016). The state of small business lending: Innovation and technology and the implications for regulation. *Harvard Business School Entrepreneurial Management Working Paper*, (17-042), 17–042.
- Milne, A., & Parboteeah, P. (2016). The business models and economics of peer-to-peer lending. *Centre for European Policy Studies*, 17, 36.
- Misheva, B. H., Giudici, P., Pediroda, V., & Ieee (2018). Network-based models to improve credit scoring accuracy. In *2018 Ieee 5th International Conference on Data Science and Advanced Analytics*, Proceedings of the International Conference on Data Science and Advanced Analytics, (pp. 623–630).
- Molnar, C. (2021). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*.
URL <https://christophm.github.io/interpretable-ml-book/>
- Munkhdalai, L., Munkhdalai, T., Namsrai, O. E., Lee, J. Y., & Ryu, K. H. (2019). An empirical comparison of machine-learning methods on bank client credit assessments. *Sustainability*, 11(3), 699.
- Nami, S., & Shajari, M. (2018). Cost-sensitive payment card fraud detection based on dynamic random forest and k-nearest neighbors. *Expert Systems with Applications*, 110, 381–392.

- Namvar, A., Naderpour, M., & Ieee (2018a). Handling Uncertainty in Social Lending Credit Risk Prediction with a Choquet Fuzzy Integral Model. In *2018 Ieee International Conference on Fuzzy Systems*, IEEE International Conference on Fuzzy Systems.
- Namvar, A., Siami, M., Rabhi, F., & Naderpour, M. (2018b). Credit risk prediction in an imbalanced social lending environment. *International Journal of Computational Intelligence Systems*, 11(1), 925–935.
- Navaretti, G. B., Calzolari, G., Mansilla-Fernandez, J. M., & Pozzolo, A. F. (2018). Fintech and banking. friends or foes? Tech. rep., European Economy – Banks, Regulation, and the Real Sector Review of Economic Conditions in Italy magazine. Year 3–Issue 2.
URL https://european-economy.eu/wp-content/uploads/2018/01/EE_2.2017-2.pdf
- Nguyen Truong, T., Khuat Thanh, S., Ngo Thi Thu, T., Nguyen Ha, N., & Tran Manh, D. (2019). Improve Risk Prediction in Online Lending (P2P) Using Feature Selection and Deep Learning. *International Journal of Computer Science and Network Security*, 19(11), 216–222.
- Niu, B., Ren, J., & Li, X. (2019). Credit scoring using machine learning by combing social network information: Evidence from peer-to-peer lending. *Information*, 10(12), 397.
- Nowak, A., Ross, A., & Yench, C. (2018). Small business borrowing and peer-to-peer lending: Evidence from lending club. *Contemporary Economic Policy*, 36(2), 318–336.
- Oreski, S., & Oreski, G. (2018). Cost-sensitive learning from imbalanced datasets for retail credit risk assessment. *TEM Journal*, 7(1), 59–73.
- Papouskova, M., & Hajek, P. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision Support Systems*, 118, 33 – 45.

- Park, S., & Choi, D. (2019). A study on P2P lending deadline prediction model based on machine learning. *Journal of KIISE*, 46(2), 174–183.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pes, B., & Lai, G. (2021). Cost-sensitive learning strategies for high-dimensional and imbalanced data: A comparative study. *PeerJ Computer Science*, 7, e832.
- Petrides, G., & Verbeke, W. (2022). Cost-sensitive ensemble learning: a unifying framework. *Data Mining and Knowledge Discovery*, 36(1), 1–28.
- Pierrakis, Y. (2019). Peer-to-peer lending to businesses: Investors’ characteristics, investment criteria and motivation. *International Journal of Entrepreneurship and Innovation*, 20(4), 239–251.
- Pokorna, M., & Sponer, M. (2016). Social lending and its risks. In S. Kapounek, & V. Krutikova (Eds.) *19th International Conference Enterprise and Competitive Environment 2016*, vol. 220 of *Procedia Social and Behavioral Sciences*, (pp. 330–337).
- Polena, M., & Regner, T. (2018). Determinants of borrowers’ default in p2p lending under consideration of the loan risk class. *Games*, 9(4).
- Pur, S., Huesig, S., Mann, H.-G., & Schmidhammer, C. (2014). How to analyze the disruptive potential of business model innovation in two-sided markets?: The case of peer-to-peer lending marketplaces in Germany. In D. F. Kocaoglu, T. R. Anderson, T. U. Daim, D. C. Kozanoglu, K. Niwa, & G. Perman (Eds.) *2014 Portland International Conference on Management of Engineering Technology*, Portland International Conference on Management of Engineering and Technology, (pp. 693–709).

- Radner, R., & Shepp, L. (1996). Risk vs. profit potential: A model for corporate strategy. *Journal of economic dynamics and Control*, 20(8), 1373–1393.
- Rao, C., Liu, M., Goh, M., & Wen, J. (2020). 2-stage modified random forest model for credit risk assessment of P2P network lending to “Three Rurals” borrowers. *Applied Soft Computing*, 95, 106570.
- Rawal, A., Raglin, A., Sadler, B. M., & Rawat, D. B. (2023). Explainability and causality for robust, fair, and trustworthy artificial reasoning. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications V*, vol. 12538, (pp. 493–500). SPIE.
- Ren, K., Malik, A., & Acm (2019). *Investment recommendation system for low-liquidity online peer to peer lending (P2PL) marketplaces*.
- Ribeiro, M. T., Singh, S., Guestrin, C., & Assoc Comp, M. (2016). Why should I trust you? Explaining the predictions of any classifier. *KDD’16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 1135–1144).
- Rodrigues, D. S., Brasil, A. R. A., Costa, M. B., Komati, K. S., & Pinto, L. A. (2018). A Comparative Analysis of Loan Requests Classification Algorithms in a Peer-to-Peer Lending Platform. In *Proceedings of the 14th Brazilian Symposium on Information Systems*, (pp. 325–332).
- ROFIEG, Expert Group on Regulatory Obstacles to Financial Innovation (2019). Thirty recommendations on regulation, innovation and finance. Tech. Rep. December, European Commission, Brussels-Belgium.
- URL https://ec.europa.eu/info/files/191113-report-expert-group-regulatory-obstacles-financial-innovation_en
- Rosavina, M., Rahadi, R. A., Kitri, M. L., Nuraeni, S., & Mayangsari, L. (2019). P2P lending

- adoption by SMEs in Indonesia. *Qualitative Research in Financial Markets*, 11(2), 260–279.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. URL <https://doi.org/10.1038/s42256-019-0048-x>
- Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16, 1–85.
- Saeidi, P., Saeidi, S. P., Sofian, S., Saeidi, S. P., Nilashi, M., & Mardani, A. (2019). The impact of enterprise risk management on competitive advantage by moderating role of information technology. *Computer Standards & Interfaces*, 63, 67–82.
- Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3), 160.
- Schölkopf, B. (2022). Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, (pp. 765–804).
- Segal, M. (2015). Peer-to-peer lending: A financing alternative for small businesses. *Issue Brief*, 10.
- Serrano-Cinca, C., & Gutierrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decision Support Systems*, 89, 113–122.
- Serrano-Cinca, C., Gutierrez-Nieto, B., & Lopez-Palacios, L. (2015). Determinants of Default in P2P Lending. *PLoS ONE*, 10(10).
- Shapley, L. S. (1953). A value for n-person games. *Contributions to the theory of games*

- (*Annals of mathematics studies*), H. W. Kuhn A. W. Tucker (Eds.). Princeton University Press, 2(28), 307–317.
- Sharma, S. K., Ilavarasan, P. V., & Karanasios, S. (2023). Small businesses and fintech: a systematic review and future directions. *Electronic Commerce Research*, (pp. 1–41).
- Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *arXiv e-prints*, (p. 1704.02685).
URL <https://arxiv.org/abs/1704.02685>
- Silva, W., Fernandes, K., Cardoso, M., & Cardoso, J. (2018). Towards complementary explanations using deep neural networks. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, vol. 11038 of *Lecture Notes in Computer Science*, (pp. 133–140). Granada, Spain: Springer International Publishing.
- Simons, R. (2019). The role of management control systems in creating competitive advantage: new perspectives. In *Management Control Theory*, (pp. 173–194). Routledge.
- Soo H., A. (2016). FinTech supporting Government’s Policy, its Implementing Measures and Legal Institution in UK- focused on the Payment Service Industry. *kangwon Law Review*, 49, 179–219.
- Stofa, T. (2017). Analysis of Repayment Failures in P2P Lending. *CEFE 2nd Central European Conference in Finance and Economics*, (pp. 773–781).
- Sugiyama, M. (2015). *Introduction to statistical machine learning*. Morgan Kaufmann.
- Sun, J., Li, H., Huang, Q.-H., & He, K.-Y. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems*, 57, 41–56.
- Sun, X., & Xu, W. (2014). Fast implementation of delong’s algorithm for comparing the

- areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21(11), 1389–1393.
- Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04), 687–719.
- Sungbok, L. (2018). Study on the Financial Intermediary Role of P2P Lending Platform - P2P. *Journal of Money Finance*, 32(2), 21–62.
- Tan, Y., Zheng, X., Zhu, M., Wang, C., Zhu, Z., & Yu, L. (2017). Investment Recommendation with Total Capital Value Maximization in Online P2P Lending. In O. Hussain, L. Jiang, X. Fei, C. W. Lan, & K. M. Chao (Eds.) *2017 Ieee 14th International Conference on E-Business Engineering (ICEBE)*, (pp. 159–165).
- Tao, Q., Dong, Y., & Lin, Z. (2017). Who can get money? Evidence from the Chinese peer-to-peer lending platform. *Information Systems Frontiers*, 19(3), 425–441.
- Tekathen, M., & Dechow, N. (2013). Enterprise risk management and continuous re-alignment in the pursuit of accountability: A german case. *Management Accounting Research*, 24(2), 100–121.
- Teles, G., Rodrigues, J. J. P. C., Saleem, K., Kozlov, S., & Rabelo, R. A. L. (2020). Machine learning and decision support system on credit scoring. *Neural Computing & Applications*, 32, 9809–9826.
- Thakor, A. V. (2020). Fintech and banking: What do we know? *Journal of Financial Intermediation*, 41, 100833.
URL <https://www.sciencedirect.com/science/article/pii/S104295731930049X>
- Uddin, M. J., Vizzari, G., Bandini, S., & Imam, M. O. (2018). A case-based reasoning

- approach to rate microcredit borrower risk in online Kiva P2P lending model. *Data Technologies and Applications*, 52(1), 58–83.
- Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84 (2), 523–538.
- Van Eck, N. J., & Waltman, L. (2011). Text mining and visualization using VOSviewer. *arXiv preprint arXiv:1109.2058*.
- Van Eck, N. J., & Waltman, L. (2020). *VOSviewer Manual*.
URL https://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.16.pdf
- Van-Sang, H., Dang-Nhac, L., Choi, G. S., Ha-Nam, N., & Yoon, B. (2019). Improving Credit Risk Prediction in Online Peer-to-Peer (P2P) Lending Using Feature selection with Deep learning. *2019 21st International Conference on Advanced Communication Technology*, 6(1), 20–31.
- Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research*, 238(2), 505 – 513.
- Waltman, L., Van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, 4(4), 629–635.
URL <https://www.sciencedirect.com/science/article/pii/S1751157710000660>
- Wan, J., Zhang, H., Zhu, X., Sun, X., & Li, G. (2019). Research on influencing factors of P2P network loan prepayment risk based on Cox proportional hazards. In E. HerreraViedma, Y. Shi, D. Berg, J. Tien, F. J. Cabrerizo, & J. Li (Eds.) *7th International Conference on Information Technology and Quantitative Management*, vol. 162 of *Procedia Computer Science*, (pp. 842–848).

- Wang, C., Han, D., Liu, Q., & Luo, S. (2019). A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism lstm. *IEEE Access*, 7, 2161–2168.
- Wang, H., Kou, G., & Peng, Y. (2018a). Cost-sensitive Classifiers in Credit Rating A Comparative Study on P2P Lending. *2018 7th International Conference on Computers Communications and Control*, (pp. 210–213).
- Wang, H., Kou, G., & Peng, Y. (2021). Multi-class misclassification cost matrix for credit ratings in peer-to-peer lending. *Journal of the Operational Research Society*, 72(4), 923–934.
- Wang, L. (2018). Supervision of Peer-to-Peer Lending in China. In J. Liu, & K. L. Teves (Eds.) *Proceedings of the 2018 2nd International Conference on Education, Economics and Management Research*, vol. 182 of *Advances in Social Science Education and Humanities Research*, (pp. 291–293).
- Wang, S., Qi, Y., Fu, B., & Liu, H. (2016). Credit Risk Evaluation Based on Text Analysis. *International Journal of Cognitive Informatics and Natural Intelligence*, 10(1), 1–11.
- Wang, Z., Jiang, C., Ding, Y., Lyu, X., & Liu, Y. (2018b). A novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending. *Electronic Commerce Research and Applications*, 27, 74–82.
- Wang, Z., Jiang, C., Zhao, H., & Ding, Y. (2020). Mining Semantic Soft Factors for Credit Risk Evaluation in Peer-to-Peer Lending. *Journal of Management Information Systems*, 37(1), 282–308.
- Wei, X., Gotoh, J.-y., & Uryasev, S. (2018). Peer-To-Peer Lending: Classification in the Loan Application Process. *Risks*, 6(4).
- Wu, C., Zhang, D., & Wang, Y. (2018). Evaluating the risk performance of online peer-to-peer lending platforms in China. *Journal of Risk Model Validation*, 12(2), 63–87.

- Xia, L., & Li, J.-f. (2016). Analysis on Credit Risk Assessment of P2P. In E. Qi, J. Shen, & R. Dou (Eds.) *Proceedings of the 22nd International Conference on Industrial Engineering and Engineering Management: Core Theory and Applications of Industrial Engineering*, (pp. 907–914).
- Xia, Y. (2019). A Novel Reject Inference Model Using Outlier Detection and Gradient Boosting Technique in Peer-to-Peer Lending. *IEEE Access*, 7, 92893–92907.
- Xia, Y., He, L., Li, Y., Liu, N., & Ding, Y. (2019). Predicting loan default in peer-to-peer lending using narrative data. *Journal of Forecasting*, 39(2), 260–280.
- Xia, Y., Liu, C., & Liu, N. (2017a). Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. *Electronic Commerce Research and Applications*, 24, 30–49.
- Xia, Y., Yang, X., & Zhang, Y. (2018). A rejection inference technique based on contrastive pessimistic likelihood estimation for P2P lending. *Electronic Commerce Research and Applications*, 30, 111–124.
- Xia, Y. F., Liu, C. Z., Li, Y. Y., & Liu, N. N. (2017b). A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, 78, 225–241.
- Xinmin, W., Peng, H., Akram, U., Yan, M., & Attiq, S. (2019). The effect of successful borrowing times on behavior of investors: An empirical investigation of the P2P online lending market. *Human Systems Management*, 38(4), 385–393.
- Xiong, J. (2018). Risk identification and monitoring model of online P2P lending. In J. Liu, & K. L. Teves (Eds.) *Proceedings of the 2018 2nd International Conference on Education, Economics and Management Research*, vol. 182 of *Advances in Social Science Education and Humanities Research*, (pp. 360–363).

- Xu, J., & Chau, M. (2018). Cheap Talk? The Impact of Lender-Borrower Communication on Peer-to-Peer Lending Outcomes. *Journal of Management Information Systems*, 35(1), 53–85.
- Xu, J., Chen, D., & Chau, M. (2016). Identifying Features for Detecting Fraudulent Loan Requests on P2P Platforms. In L. Zhou, L. Kaati, W. Mao, & G. A. Wang (Eds.) *2016 IEEE International Conference on Intelligence and Security Informatics: Cybersecurity and Big Data*, (pp. 79–84). IEEE.
- Xu, J. J., Lu, Y., & Chau, M. (2015). P2P Lending Fraud Detection: A Big Data Approach. In M. Chau, G. A. Wang, & H. Chen (Eds.) *Intelligence and Security Informatics, Paisi 2015*, vol. 9074 of *Lecture Notes in Computer Science*, (pp. 71–81).
- Xu, L., & Zhang, Y. (2017). A Credit Rating Model for Online P2P Lending Based on Analytic Hierarchy Process. In J. Xu, A. Hajiyeve, S. Nickel, & M. Gen (Eds.) *Proceedings of the Tenth International Conference on Management Science and Engineering Management*, vol. 502 of *Advances in Intelligent Systems and Computing*, (pp. 537–549).
- Yan, Y., Lv, Z., & Hu, B. (2016). Building Investor Trust in the P2P Lending Platform with a Focus on Chinese P2P Lending Platforms. In *2016 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI)*, (pp. 470–474). IEEE.
- Yan, Y., Lv, Z., & Hu, B. (2017). Building investor trust in the P2P lending platform with a focus on Chinese P2P lending platforms. *Electronic Commerce Research*, 18(2), 203–224.
- Yang, X., Fan, W., Wang, L., Yang, S., & Wang, W. (2019). Risk control of online P2P lending in China based on health investment. *Ekoloji*, 28(107), 2013–2022.
- Yao, J., Chen, J., Wei, J., Chen, Y., & Yang, S. (2019). The relationship between soft information in loan titles and online peer-to-peer lending: EVidence from RenRenDai platform. *Electronic Commerce Research*, 19(1), 111–129.

- Ye, X., Dong, L.-A., & Ma, D. (2018). Loan evaluation in P2P lending based on Random Forest optimized by genetic algorithm with profit score. *Electronic Commerce Research and Applications*, 32, 23–36.
- Yli-Huumo, J., Ko, D., Choi, S., Park, S., & Smolander, K. (2016). Where is current research on blockchain technology? A systematic review. *PLoS ONE*, 11(10), e0163477.
- Yuan, Z. N., Wang, Z. H., & Xu, H. (2018). *Credit Risk Assessment of Peer-to-Peer Lending Borrower Utilizing BP Neural Network*, vol. 6 of *Lecture Notes on Data Engineering and Communications Technologies*, (pp. 22–33).
- Zang, D., Qi, M., & Fu, Y. (2015). The credit risk assessment of P2P lending based on BP Neural Network. *Industrial Engineering and Management Science*, 2014, 91–94.
- Zhang, W., Xu, W., Hao, H., & Zhu, D. (2020). Cost-sensitive multiple-instance learning method with dynamic transactional data for personal credit scoring. *Expert Systems with Applications*, 157.
- Zhang, Y., Geng, X., & Jia, H. (2017a). The Scoring Matrix Generation Method and Recommendation algorithm in P2P Lending. In R. Bahsoon, & Z. Chen (Eds.) *2017 13th Ieee World Congress on Services*, IEEE World Congress on Services, (pp. 86–89).
- Zhang, Y., Jia, H., Diao, Y., Hai, M., & Li, H. (2016a). Research on Credit Scoring by fusing social media information in Online Peer-to-Peer Lending. In H. Lee, Y. Shi, J. Lee, F. Cordova, I. Dzitac, G. Kou, & J. Li (Eds.) *Promoting Business Analytics and Quantitative Management of Technology: 4th International Conference on Information Technology and Quantitative Management*, vol. 91 of *Procedia Computer Science*, (pp. 168–174).
- Zhang, Y., Wang, D., Chen, Y., Shang, H., & Tian, Q. (2017b). *Credit Risk Assessment Based on Long Short-Term Memory Model*, vol. 10362 of *Lecture Notes in Computer Science*, (pp. 700–712).

- Zhang, Y., Wang, D., Chen, Y., Zhao, Y., Shao, P., & Meng, Q. (2017c). *Credit Risk Assessment Based on Flexible Neural Tree Model*, vol. 10261 of *Lecture Notes in Computer Science*, (pp. 215–222).
- Zhang, Y., Wang, X., Qian, Y., & Jia, H. (2016b). The research of recommendation algorithms in P2P lending. In *International Conference on Mechanical, Control, Electric, Mechatronics, Information and Computer (MCEMIC 2016)*, (pp. 241–247).
- Zhao, J. (2015). Research on Mathematical Model P2P Online Credit Risk Evaluation Based on Data Processing. In J. Wang, & Y. Qin (Eds.) *Proceedings of the 2015 International Conference on Education Technology, Management and Humanities Science*, vol. 27 of *Advances in Social Science Education and Humanities Research*, (pp. 897–900).
- Zhou, G., Zhang, Y., & Luo, S. (2018). P2P Network Lending, Loss Given Default and Credit Risks. *Sustainability*, 10(4), 1010.
- Zhou, J., Li, W., Wang, J., Ding, S., & Xia, C. (2019). Default prediction in P2P lending from high-dimensional data based on machine learning. *Physica a-Statistical Mechanics and Its Applications*, 534.
- Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. 162, 503–513.
- Zhu, Z. (2018). Safety promise, moral hazard and financial supervision: Evidence from peer-to-peer lending. *Finance Research Letters*, 27, 1–5.