Computational Economics
https://doi.org/10.1007/s10614-025-10962-9

Machine Learning XAI for Early Loan Default Prediction

Leticia Monje[1] · **Ramón Alberto Carrasco**[2] · **Manuel Sánchez-Montañés**[3]

Accepted: 13 April 2025

Abstract
Early default prediction with predictive models is of crucial importance for financial institutions, Fintech or Peer to Peer (P2P) lending platforms, as it allows them to effectively mitigate the potential risks associated with customer or debtor defaults, anticipating before this becomes a major problem. This proactive approach serves to avoid the consequent impact on provisions and, subsequently, on the institution's capital. On the other hand, advanced predictive models are often less interpretable than traditional models such as probit (Abdou & Pointon, 2011) and logistic regression (Bolton, 2009; Liu et al. 2024). Due to this lower explainability, our goal was to develop a methodology that allows building an advanced predictive model together with a linguistically interpretable explanation useful for decision making from large volumes of data. For this purpose, our case study was the loan dataset of Lending Club, the largest P2P lending platform in the world. As a result, we obtained a model based on the eXtreme Gradient Boosting (XGBoost) together with its linguistic interpretation using a surrogate model and the 2-tuple fuzzy linguistic model Monje et al., (*Mathematics* 10:1428, 2022). This model allows us to identify five risk categories (very low, low, medium, high and very high).

Keywords Peer to peer lending · Machine learning · Early default risk · XAI · Fuzzy

✉ Leticia Monje
lmonje01@ucm.es

Ramón Alberto Carrasco
ramoncar@ucm.es

Manuel Sánchez-Montañés
manuel.smontanes@uam.es

[1] Faculty of Statistics, Complutense University Puerta de Hierro, 28040 Madrid, Spain

[2] Marketing Department, Faculty of Statistics, Complutense University Puerta de Hierro, 28040 Madrid, Spain

[3] Computer Science Department, Universidad Autónoma de Madrid, 28049 Madrid, Spain

Published online: 03 June 2025

1 Introduction

In the aftermath of the 2008 crisis, and due in part to the general public's declining confidence in the stability of banking institutions, a lending model emerged that, although it might seem modern, is based on old lending practices. For this reason, Lending Club was launched in May 2007, a lending platform between individuals and investors without the intermediation of a traditional financial institution. Based on the crowdlending technique, it offered P2P loans, with terms ranging from three to five years and amounts up to $35,000 (Namvar, 2013). It is thus about person-to-person lending: people with money (investors), based on credit information, lend money through a platform to anonymous people who need it (borrower) (Wang et al., 2015). In this context, P2P and bank lending have some differences e.g. intermediation: in a bank, the loan is done through a financial institution, while in P2P loans, it is done through an online platform directly between the borrower and the lender. Another difference is that the application process in a bank is long and needs a lot of documentation, while for P2P loans it is simpler and faster. In addition, P2P loans have higher interest rates than bank loans due to the perceived risk of the lenders. Finally, bank loans are less flexible because banks have to comply with set admission policies. In contrast, P2P and bank loans share similarities in loan usage, credit evaluation and periodic payments (De Roure et al., 2016; Wang et al., 2015).

Currently, the P2P lending model has evolved, as investors, such as Fintechs, can apply their money to a portion of a loan or split it into several, without the need for an investor to lend money to a specific client (Chishti, 2016). Thus, these new investors are using new ways of analyzing risk that include Big Data, Machine Learning techniques and the incorporation in the analysis of unstructured information with the aim of assessing risk more accurately. Consequently, these new technologies are making a difference in the financial environment, especially for credit risk management. The reason is that they make it possible to process a large amount of data in a short time, while predicting defaults more accurately, which is crucial for the sector. Thus, credit risk analysis is crucial, because it is the first barrier to defend solvency, but also to provide liquidity adequately and without inefficiencies. In this sense, identifying signs of early deterioration in customers is crucial, since preventive measures can be taken at that time to avoid further deterioration of the customer. This is done by monitoring early delinquency, which is considered when the default is more than 30 days old, allowing entities to take preventive measures and seek mutually viable solutions before significant delinquency occurs.

The use of machine learning techniques in the financial system is becoming increasingly widespread for the early detection of defaults. This type of default can be modeled by companies and does not legally require white-box models. Therefore, sophisticated black-box models such as Neural Networks and XGBoost can be used (Adadi & Berrada, 2018), but they are not interpretable. However, it is crucial to subsequently apply eXplainable Artificial Intelligence (XAI) techniques to these models in order to be understood and audited (Heng & Subramanian, 2022).

For all these reasons, our objective is to develop a predictive and linguistically interpretable model, useful for decision making using large volumes of data, that

predicts early default. To validate the goodness of this model, we have applied it to a real case, the dataset of P2P loans granted in the period 2007-2020Q3 from the Lending Club platform. To achieve our goal, we achieved explainability of the developed algorithm using Machine Learning techniques (Gunning et al., 2019).

The rest of the paper is structured as follows: Sect. 2 presents the state of the art and compares relevant works found that are related to our proposal; Sect. 3 presents the basics of Decision Trees, Fuzzy models and XAI, on which our proposal is based; Sect. 4 presents the model for predicting early loan default in lending, and the method to obtain its interpretation; in Sect. 5 we discuss the results obtained; finally, in Sect. 6, we show the conclusions and future work.

2 Related Work

Since delinquency is the biggest problem a financial institution can have, our paper focuses on the prediction of early default. Being able to foresee early default allows preventive measures such as debt restructuring before default, thus avoiding accounting defaults. This makes provisioning unnecessary and would not reduce available capital. P2P lending markets have grown a lot in recent years, due to the fact that they involve private-to-private lending on an online platform (Lenz, 2016). If a borrower is unable to repay, lenders must bear the credit risk when default occurs Therefore, early default prediction is crucial. Traditional approaches, mainly probabilistic financial models such as credit scorecards, use a linear model such as a logistic regression (Dreiseitl and Ohno-Machado, 2002). To improve default prediction, more sophisticated Machine Learning models are used. Furthermore, it is important that these models are easily interpretable in order to validate them and check for possible biases, so Explainable Artificial Intelligence (XAI) plays an important role. Model interpretability (Arrieta et al., 2020) is crucial for transparency and traceability, especially in default prediction (Leo et al., 2019).

The bibliometric analysis is divided into two parts: studies that seek to predict default, and studies that interpret in depth the “black box” models that predict default. The literature search was conducted using combinations of the following keywords:"forecast","prediction*","credit","loan","mortgage","lending","machine learning","deep learning","P2P","peer to peer","XAI","explainable artificial intelligence".

In the study (S. H. Chen et al., 2019a, 2019b), a Machine Learning model with feature selection is proposed and it is shown that credit risk prediction for P2P lending can be improved using logistic regression in addition to appropriate feature selection. Subsampling was used to balance the dataset. Another similar study (Zhou et al., 2019) proposes the use of gradient boosting decision trees (GBDT), XGBoost, and the light gradient boosting machine (LightGBM). In the same line, (Ma et al., 2018) and (Ko et al., 2022) also propose LightGBM.

In (Li et al., 2018), a multi-round ensemble learning model based on heterogeneous ensembles was designed for default risk prediction. In this model, XGBoost is initially used for ensemble learning, and the XGBoost, deep neural network (DNN)

and logistic regression are considered as heterogeneous individual learners which are subject to linear weighting. In (Yang et al., 2022) a new Sparrow Search Algorithm (SSA)-CatBoost model is proposed, which combines SSA and CatBoost to improve prediction and classification. Comparing other known Machine Learning models and the CatBoost model, the SSA-CatBoost model has the best classification and prediction accuracy. In (Song et al., 2020) it is proposed a multi-view ensemble learning method based on distance-to-model and adaptive clustering (DM-ACME) which uses gradient boosting. This method produces a set of diverse ensembles of decision trees. In (Zhu et al., 2019) it is proposed a loan default prediction model based on Random Forest (RF) trained using the Synthetic Minority Oversampling Technique (SMOTE) resampling method to cope with the class imbalance problem.

On the other hand, in (T. Chen, 2021) the authors deal with the class imbalance problem using the XGBoost model with three resampling methods (SMOTE, NearMiss and random selection). In (Niu et al., 2020), an ensemble resampling based on data distribution (REMDD) is performed, in which using the subsampling method based on majority class data distribution (UMCDD) deals with l the class imbalance problem improving the classification of REMMD. In (Li et al., 2020a, 2020b) it is proposed a model that adds missing values to the model for self-training improving its performance. This paper addresses the class imbalance problem by proposing a new model based on heterogeneous ensemble learning. In (Li, 2022), a predictive model based on deep learning is proposed for online loan default prediction. The Back-Propagation Neural Network (BPNN) model is compared with a Support Vector Machine (SVM) and a Regression Model, with BPNN model having the highest accuracy. In this line, (Kim & Cho, 2019) presents another deep learning model, a convolutional neural network (CNN), for repayment prediction in P2P social lending.

P2P loan data is generated in real time, but many of the loans are pending repayment decisions because the term has not yet expired. In (J. Y. Kim & Cho, 2022) it is considered that labeled unexpired data improves the prediction performance of the model. For this purpose, they propose a joint classifier composed of several Convolutional Neural Network (CNN) models including GoogLeNet, ResNet, and DenseNet, and the additional data labeled by Dempster-Shafer fusion. In (Kriebel and Stitz, 2022) the authors propose a method to improve default risk prediction using the extraction of keywords and short text fragments from user comments. This work shows that simpler methods, such as average embedding neural networks help to obtain better predictability compared to more complex methods, such as Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT (RoBERTa).

In (Fu et al., 2020), it is proposed to extract keywords from investors' comments and then use a model based on bidirectional long short-term memory (BiLSTM). In (Lee et al., 2021), a graph model based on nonlinear graph convolutional networks (GCNs) is proposed. Three types of information about borrowers (loan information, credit history information and soft information), are used.

On the other hand, in (Xu et al., 2021) the factors affecting repayment are analyzed. Four machine learning methods (RF, XGBoost, Gradient Boosting and Neural Network) were applied to predict the important factors affecting repayment. The

results showed that borrowers who have passed only video, cell phone, job, residence, or education level verification are more likely to default, while those who have passed asset identity verification are less likely to default.

In the paper (Arora & Kaur, 2020) it is introduced Bolasso (Bootstrap-Lasso), a technique designed to select consistent and relevant features. The pre-selected features generated by Bolasso are applied to various classification algorithms, such as RF, SVM, Naïve Bayes (NB), and K-Nearest Neighbors (K-NN) to test their predictive accuracy. It is observed that Bolasso-enabled RF (BS-RF) provides the best results for credit risk assessment in that paper.

On the other hand, in (Cho et al., 2019) an investment decision model in the P2P lending market is proposed that consists of classifying fully paid loans using the Instance-based Entropy Fuzzy Support Vector Machine method (IEFSVM), which is a modified version of the Entropy-based Fuzzy Support Vector Machine method (EFSVM). By applying the model to the loan dataset, the loans that are expected to be fully paid are determined. To predict loan default (Stevens et al., 2020) used the XGBoost model and the explainability of the model was extracted by post hoc explanations using the SHapley Additive exPlanations (SHAP) method. In (Moscato et al., 2021) a comparative study of some of the most widely used scoring models for credit risk prediction was performed. They used different techniques to deal with the class imbalance problem and finally, the three best approaches were evaluated in terms of their explainability using different XAI tools.

Therefore, in this work we consider the XGBoost model, a model widely used in the cited literature. Table 1 shows the works specifically related to default prediction and the explainability of the models used. The table also includes our proposal.

3 Methodology

The dataset includes 2,925,493 loans and 141 variables with information about each client. First, an exploratory analysis of the data was performed. By analyzing all the variables and their different values, leaky variables that included information after the loan was formalized and those that did not provide relevant information were removed. In addition, variables with more than 40% missing values were removed. Once the dataset was cleaned the variables were reduced to 57 variables (Polena and Regner, 2018) we split the dataset into 70% train and 30% test and used different Machine Learning algorithms to construct models to predict defaults.

The Machine Learning algorithms investigated included Logistic Regression, Decision Tree, Random Forest (RF), Balanced Bag classifier, XGBoost, K-NN, Gaussian NaÏve Bayes (Gaussian NB), Neural Networks, and AdaBoost.

Due to class imbalance (He and Garcia, 2009; Alam et al., 2020), where 20.20% of loans experience early defaults, we applied methodologies such as SMOTE, Random Undersampling and Random Oversampling (Costello and Lee, 2019). The fundamental consideration in our model selection process was the minimization of false negatives and accurately identifying positive cases. This emphasis is due to the fact that early detection of defaults serves to avoid their deterioration and the consequent impact on the financial institution's capital.

Table 1 Relevant works on default risk prediction and XAI

Ref	Datasets	Fundamentals	Application	Interpretability
Chen et al., 2019a, 2019b	P2P loans	Logistic Regression (LR)	Default risk prediction	Black box, non-interpretability
Zhou et al., 2019	P2P loans	Several Machine Learning methods, GBDT, XGBoost, LightGBM	Default risk prediction	Black box, non-interpretability
Ma et al., 2018	P2P loans in Asia	LightGBM	Default risk prediction	Black box, non-interpretability
Ko et al., 2022	P2P loans	LightGBM	Mitigate the risks of default and asymmetric information	
Li et al., 2018	P2P loans in China	Several Machine Learning methods: XGBoost, DNN, LR	Default risk prediction. Comparison with traditional machine learning models and ensemble learning models	
Yang et al., 2022	Credit rating	SSA-CatBoost	Default risk prediction and classification	
Song et al., 2020	P2P loans	DM-ACME	Default risk prediction	
Zhu et al., 2019	P2P loans	RF-SMOTE	Default risk prediction	
T. Chen, 2021	P2P loans	SMOTE, NearMiss	Default risk prediction. Imbalanced problem	
Niu et al., 2020	P2P loans	REMDD, UMCDD	Default risk prediction. Imbalanced problem	
Li et al., 2020a, 2020b	P2P loans in China	Several Machine Learning methods: XGBoost, DNN, LR	Default risk prediction. Imbalanced problem	
Li, 2022	P2P loans	NN	Default risk prediction	
Kim & Cho, 2019	P2P loans	CNN	Default risk prediction	
Kim & Cho, 2022	P2P loans	CNN	Ensemble classifier for the repayment prediction in social lending	
Kriebel and Stitz, 2022	P2P loans	BERT, RoBERTa	Default risk prediction using short pieces of user-generated text	
Fu et al., 2020	P2P loans	BiLSTM	Extract keywords from investor comments to predict the default risk	
Lee et al., 2021	P2P loans	GCN	Default risk prediction	
Xu et al., 2021	P2P loans in China	Several Machine Learning methods: RF, XGBT, GBM, NN	Default risk prediction using data from renrendai.com	
Arora & Kaur, 2020	P2P loans	Several Machine Learning methods: RF, SVM, NB, K-NN BS-RF	Default risk prediction	
Cho et al., 2019	P2P loans	IEFSVM	Default risk prediction	

Table 1 (continued)

Ref	Datasets	Fundamentals	Application	Interpretability
Stevens et al., 2020	P2P loans	SHAP	Default risk prediction and explainability	SHAP
Moscato et al., 2021	P2P loans	Several Machine Learning methods: LR, RF, MLP	Default risk prediction and explainability	SHAP, LIME, Anchors, BEEF and LORE
Our proposal	P2P loans	XGBoost	Default risk prediction and explainability	2-Tuple Fuzzy Linguistic Surrogate Trees and Rules

Therefore, the selected model prioritizes high recall, high F1 score, and robust AUC (Area under the Receiver Operating Characteristic curve) (Krzanowski and Hand, 2009).. With these fundamentals in mind, the following section delves exclusively into the intricacies of the XGBoost classifier model (Li et al., 2020a, 2020b; Chen & Guestrin, 2016) complemented with optimized parameters, recognized as the most promising default prediction model.

XGBoost is a so-called black-box model, and the problem with such models is that it is very difficult to intuitively explain how it has reached its conclusion (in our case, predicting whether a customer defaults or not). However, model interpretability (Arrieta et al., 2020) is crucial for transparency and traceability, especially in default prediction (Leo et al., 2019). Therefore, interpretation using XAI techniques is crucial. In this sense, from a business perspective, the XAI technique of surrogate models can provide us an intuitive interpretation of the model. (Molnar, 2020).

The inputs to the surrogate model have been a subset of variables that provide key information about the financial situation, credit history and loan details:

-Term_60_months: This variable indicates the term of the loan in months. Specifically, in this case, it refers to loans with a term of 60 months (5 years). It can be used to evaluate the duration of the financial commitment associated with the loan.
- Application_type: It describes the type of application submitted for the loan. It can have two main values:"Individual"if the application is submitted by a single person, and"Joint"if the application is joint, i.e., submitted by more than one person.
- Delinq_2 yrs: It represents the number of late payments in the last two years. This variable is an indicator of payment history and may influence the assessment of the applicant's credit risk.
-Inq_last_6 mths: It indicates the number of credit history inquiries made in the last 6 months. A higher number of inquiries may suggest greater financial activity. For a better understanding of the rules obtained by the surrogate model, we represent each of the continuous variables with a linguistic variable based on a set of linguistic values (Very Low, Low, Medium, High, Very High). Therefore, we propose the use of fuzzy linguistic variables to improve the interpretability of these variables, which were originally proposed by Zadeh (Zadeh, 1975). This type of variable represents linguistic concepts, based on fuzzy sets, which are commonly used by humans. Specifically, we propose the use of the 2-tuple fuzzy linguistic model, which allows for higher accuracy in the representation and computation of these fuzzy concepts.

The process is shown in Fig. 1.

We will now explain the proposed process based on three models: XGBoost, the 2-tuple fuzzy linguistic model, and the surrogate tree and rules.

3.1 Extreme Gradient Boosting Classifier

XGBoost is an optimized and distributed gradient boosting library for constructing Machine Learning models under the Gradient Boosting framework (Chen &

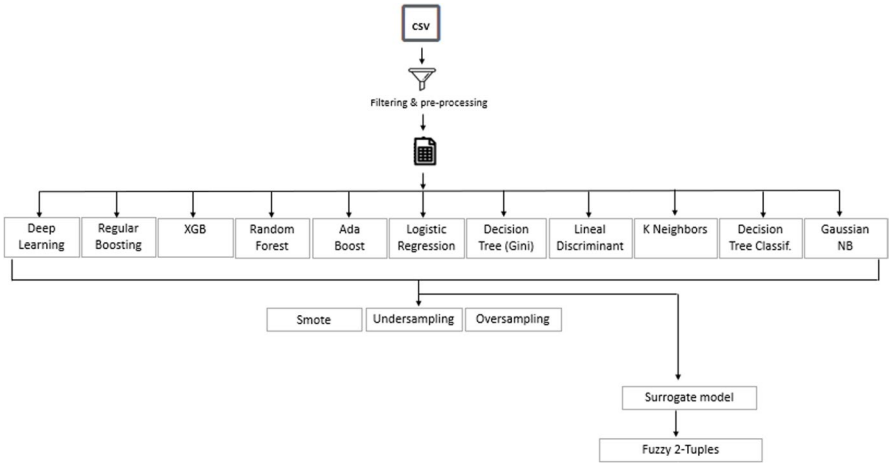

Fig. 1 Machine learning and XAI processes

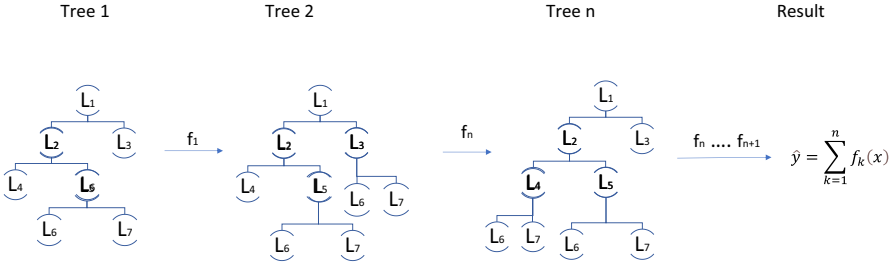

Fig. 2 Prediction with XGBoost model: each of the n trees produces a prediction f, which is combined with the others to produce the final prediction. L1, L2, etc. represent the nodes of a particular tree

Guestrin, 2016). Due to the quality of the models that can be built with XGBoost, it has become a widely used tool. In addition, it is flexible and fast to train, works with both classification and regression problems and with large volumes of data. The algorithm is a supervised learning technique based on decision trees. It consists of an assembly of sequential trees as shown in Fig. 2, where the subsequent tree learns the result of the previous trees by correcting the error produced by them, until the error can no longer be corrected.

If the new model performs better, it is used as a basis for further modifications. If, on the other hand, it performs worse, the training algorithm returns to the previous best model and modifies it differently.

3.2 The 2-Tuple Fuzzy Linguistic Model

The 2-tuple model proposed in (Herrera & Martínez, 2000) is a versatile model used in different areas (Marín Díaz et al., 2021; Bueno et al., 2022a, 2022b; Carrasco et al., 2015; Monje et al., 2022) since it allows a more precise representation of the

fuzzy linguistic terms of a linguistic variable without losing linguistic interpretability. For this reason, it is a widely used model in different areas.

This model represents the information as a pair of values (s_i,α_i), where s_i∈S and $\alpha_i \in [-0.5,0.5)$

Definition 1. The symbolic translation of a linguistic term is defined as the difference between the value β∈ [0,g] obtained from a symbolic operation and the index of the closest linguistic term si in S. The result of this difference is a number in the [− 0.5, 0.5) interval.

To perform transformations between two-tuple linguistic values and numeric values defined in the granularity interval, as well as to perform computational processes on 2-tuple linguistic values, the two-tuple linguistic model defines the following pair of functions:

Definition 2. Let $S = s_0, \ldots, s_g$ a set of linguistic terms, $\langle S\rangle = S \times [-0.5,0.5)$ and $\beta \in [0, g]$ a value representing the result of a symbolic operation. Then, the linguistic 2-tuple expressing information equivalent to β is obtained using the following function:

$$\begin{gathered}\Delta_S : [0, g] \to \langle S\rangle \\ \Delta_S(\beta) = (s_i, \alpha_i), with\{i = round(\beta)\alpha = \beta - i, \alpha\epsilon[-0.5,0.5),\end{gathered} \tag{1}$$

where *round*(•) is the visual rounding operator, s_i is the label with index closest to β and α is the value of the symbolic translation.

Thus, a value in the interval [0, g] always identified with a 2-tuple linguistic value in $\langle S\rangle$.

Definition 3. Let $S = s_0, \ldots, s_g$ a set of linguistic terms and $(s_i, \alpha_i)\epsilon\langle S\rangle = S \times [-0.5,0.5)$. The numerical value in the granularity interval $[0, g]$ representing the linguistic value 2-tuple (s_i, α_i) is obtained using the function:

$$\begin{gathered}\Delta_S^{-1} : \langle S\rangle \to [0, g] \\ \Delta_S^{-1}(s_i, \alpha_i) = i + \alpha = \beta\end{gathered} \tag{2}$$

Along with the representation model seen above, we will define 2-tuple linguistic value comparison operators. Given two 2-tuple linguistic values (*sk*,1) and (*sl*,α2) representing quantities of information:

- If $k < l$, then (s_k, α_1) is less than (s_l, α_2).
- If $k = l$, then

  1) If $\alpha_1 = \alpha_2$, then (s_k, α_1) and (s_l, α_2) represent the same information.
  2) If $\alpha_1 < \alpha_2$, then (s_k, α_1) is less than (s_l, α_2).
  3) If $\alpha_1 > \alpha_2$, then (s_k, α_1) is greater than (s_l, α_2).

Default is the variable to be predicted, so in our study it is the dependent variable.

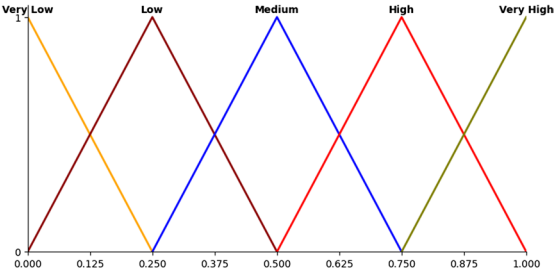

Fig. 3 Definition of set S for the variable "probability of early default"

As will be seen in Sect. 4, we will represent the probability of default with the 2-tuple model to improve the linguistic interpretability of this probability obtained with the surrogate model.

For this purpose, we define,

$$S = \{s_0, ..., s_T\}$$

with:

$$T = 4 : s_o = VeryLow = VL,$$

$$s_1 = Low = L,$$

$$s_2 = Medium = M,$$

$$s_3 = High = H,$$

$$s_4 = VeryHigh = VH$$

with the definition shown in Fig. 3. Thus, for example, if in a given context the probability of early default is (Medium, −0.2), we can interpret that it is well above "Very Low" and well below "Very High", and is also below "Medium".

3.3 Surrogate Tree and Rules

The so-called black box algorithms are those that are not intuitively interpretable by humans, since they make decisions that we cannot understand (Minh et al., 2022). In this sense, artificial intelligence (AI) has made great advances in this type of algorithms in recent years.

As a result, the area of eXplainable Artificial Intelligence (XAI) has emerged, where there are two main approaches:

- Global approach: methods that try to explain in a generic way the decisions of black box models.

- Local approach: methods that try to explain particular decisions of black box models.

The main XAI techniques are shown in Table 2.

We use a global model to understand the prediction of early default on loan applications. Because we are estimating the probability of default, which is a continuous variable, we use an interpretable model based on regression decision trees. To do so, we will use the following algorithm (Molnar, 2020):

1. Select the data set X used to train the black-box model.
2. Obtain the probabilities of default $p_{blackbox}$ for the dataset X estimated by the black-box (non-interpretable) model.
3. Train a regression decision tree on the dataset X with the aim of predicting $p_{blackbox}$.
4. Obtain the rules equivalent to the decision tree.
5. Measure how well the surrogate model replicates the predictions of the black-box model.
6. Fuzzify continuous numerical variables using the 2-tuple fuzzy linguistic model (Eq. 1).
7. Interpret the surrogate fuzzy linguistic model.

For step 6, we will use the coefficient of determination R^2, which is a standard measure of how well the regression tree explains the predictions of the black-box model we are trying to explain.

To interpret black-box models, we usually use surrogate models that are decision trees, with the final tree expressed as a simple set of rules. In this sense, there are several works focused on the extraction of rules (Bologna, 2019; Keneni et al., 2019; Singh et al., 2019). The fuzzy version of these rules uses fuzzy sets that allow to model fuzzy concepts in a way closer to human conception. Therefore, it is an interpretable model (Fernández et al., 2019). An example of a fuzzy rule is: IF (term_ 60 months $=0$) and ΔS (application_type ≤ 1) and (delinq_2 yrs ≤ 1) and (inq_last_6 mths ≤ 0) and (bc_util $\leq$ Medium) and (acc_open_past_24 mths > 6) then probability = Low.

In step 7, the most appropriate fuzzy linguistic value of the target variable is obtained. The fuzzification process converts the variable value into a linguistic variable label and may result in information loss. By using the 2-tuple model, there is no information loss and hence an accurate value can be represented as a fuzzy value using the linguistic label values and symbolic translation. Although fuzzy rules have been applied to prediction problems and other areas (Viji et al., 2020), as mentioned above, we have not found application of the 2-tuple model to our interpretability problem.

4 Proposed Model

To achieve our objectives, we will rely on the phases of Knowledge Discovery in Databases (KDD) proposed in (Fayyad, et al., 1996). This framework divides the overall process into a series of ordered phases, which involve understanding the domain and

Table 2 Main features of different XAI techniques (Elshawi et al., 2019; Bueno et al., 2022a)

Interpretation		Technique	Advantages	Disadvantages
Local	Global			
x		Shapley Value Explanations	Two axioms efficiency, symmetry its result is a single allocation of resources. Strong game theory	Extremely computationally expensive
x		Local Surrogate Models	Intuitive explanation	Can give different explanations to nearby points
	x	Global Surrogate Model	Flexibility and ease of assessing how well it approximates the black box model	Trustworthiness. The alternative model must approximate the black box model to be reliable
	x	Feature Interaction	It takes into account the interaction between traits	Interaction is computationally expensive
	x	Individual Conditional Expectation	Easy to understand and intuitive to interpret	Assumption of distribution completely independent between features
	x	Partial Dependence Plot	Easy interpretation	Assumption of distribution completely independent between features
	x	Feature Importance	Iteractions between features are considered	Unclear whether it can be used with training or test datasets

defining the scope of application, data extraction, preprocessing the data (including selection, cleaning and transformation), and interpretation/evaluation of the results obtained. This method is shown in Fig. 4 and its steps are explained below.

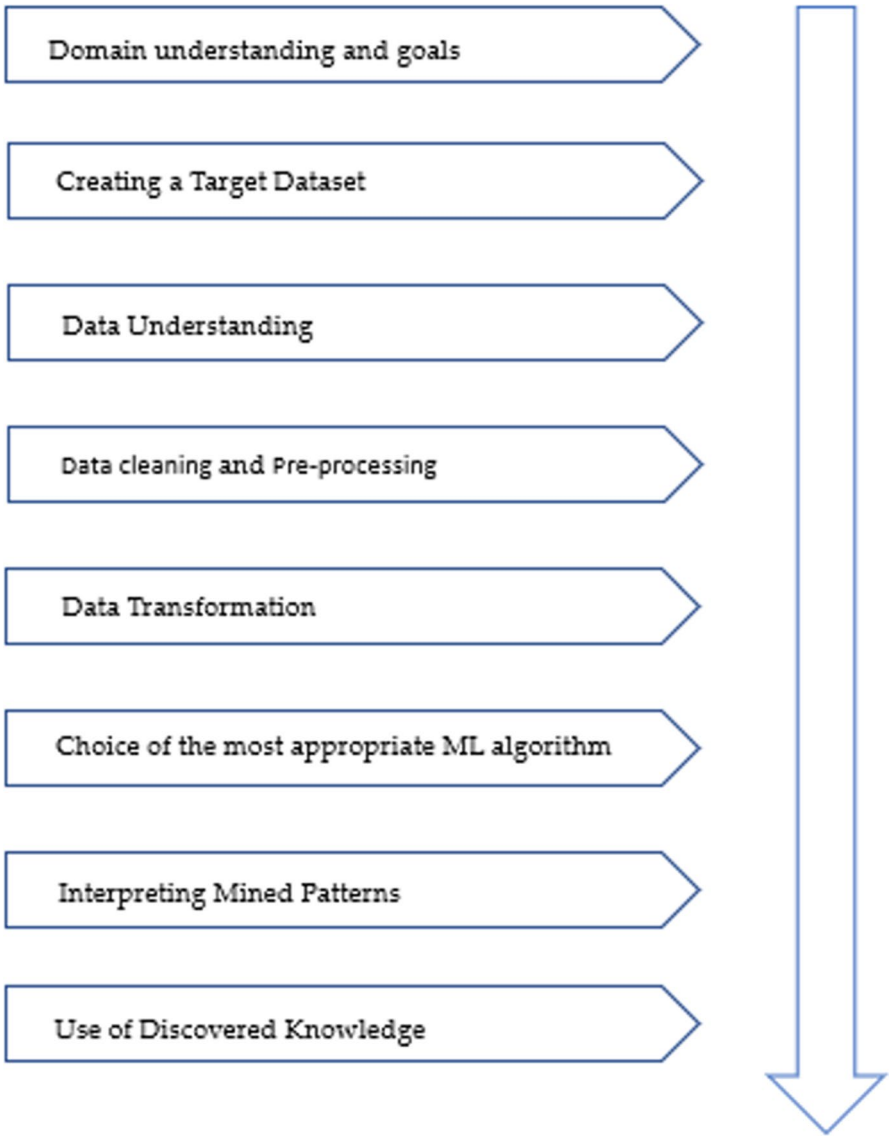

Fig. 4 Proposed strategy

4.1 Domain understanding and goals

There are currently two types of approaches for default prediction:

- White-box models: These models have transparent and easily interpretable internal processes, but they cannot capture complex relationships in the data. Therefore, they have lower accuracy compared to black-box models (Fendi et al., 2017).
- Black-box models: These are very complex models that can handle large amounts of data and nonlinear relationships between variables. Their complexity allows them to achieve high levels of accuracy in difficult predictive tasks, prioritizing high precision and predictive performance over interpretability. However, their lack of transparency presents significant challenges for banking regulators in terms of effective supervision, risk assessment, and bias detection (Rudin, 2019).

Therefore, the objective was to develop a model capable of accurately predicting early default, as it is vitally important for financial institutions to avoid incurring losses and to understand the underlying determinants of such predictions. This enables institutions to make informed lending decisions based on predicted outcomes In turn, regulators are better equipped to monitor, assess and manage financial risks effectively, as accuracy and interpretability help promote sounder financial practices.

Our approach aims to combine the high accuracy of a black-box model with the explainability of a white-box model, allowing the firm to effectively manage the risk of early default. Understanding the underlying drivers of these defaults is equally crucial, as it enables firms to make informed decisions on loan approvals based on customer patterns. This, in turn, provides regulators with enhanced capabilities to monitor, assess and manage financial risks more effectively, thereby promoting sounder financial practices. Table 3 summarizes the applicability of each approach.

Within the scope of this study, our focus is on the analysis of P2P lending spanning the period from 2007 to the third quarter of 2020.

4.2 Cleaning the Dataset

Data on peer-to-peer lending was provided in the form of a flat file in CSV format. This file contained detailed information about the loan applicants. The variables in the file contained data relevant to the loan application and subsequent instances. Data processing was performed using the Python programming language. Finally, the variables selected were those that did not contain any post-loan approval information, and at the same time provided relevant information for predicting early default.

Table 3 Applicability of each approach to early default prediction

Approach	Advantages	Disadvantages	Reference
	Increased interpretability and transparency: This is valuable for financial analysts, regulators, and other stakeholders who need to understand the decision analysts, regulators, and other stakeholders who need to understand the decision-making process to justify or take corrective action For the regulator it also allows for clear explanations on an individual or global level and to be able to compare methodology with other entities	Lower predictive capacity generates a lower quality credit portfolio, which puts the solvency and profitability of the financial institution at risk	
White box		Reduced ability to work with large masses of data: As the number of variables or the complexity of the number of variables or the complexity of the data, white-box models can become less efficient and therefore impact profitability Oversimplification of data and to a loss of significant information by resulting in reduced effectiveness	(Ko et al., 2022
	Greater precision: this translates into savings in provisions and consequently less impact on capital. Consequently, less impact on capital	Lack of interpretability: Regulators and financial institutions often require decisions to be explainable and auditable	
Black box	Handling of large volumes of data: allows learning from more diverse examples and better generalizing and therefore greater accuracy		Malekipirbazari & Aksakalli, 2015
	Adaptation and self-learning: they can automatically adapt to changing data and continuously improve their accuracy		
Black box with XAI	High predictive capacity: Increase profitability by avoiding subsequent impacts on provisions and, consequently, on capital. Explainable preventive measures can be taken to minimise financial losses by intervening and offering customised solutions to clients based on criteria, such as debt restructuring		Our proposal
	High interpretability: Meets the requirements of regulators in terms of requiring decisions to be explainable and auditable		
	Handling large volumes of data: better predictions and more informed decision-making		

4.3 Data Understanding

Data collection and review is the focus of the second phase of the KDD process. We focus on the analysis of early defaults and for this we will only consider loans for which we know how they have ended (fully paid and charged off, and those that have been 31–120 days late). The variable"Delinq_2 yrs"corresponds to the count of delinquency incidences extending beyond 30 days in the borrower's credit history over the previous 2 years. The numerical value assigned to this variable varies depending on the count of early defaults experienced by the client, so it is a variable with information from the past.

The dependent variable is a constructed binary variable. The value 1 represents loans that at some point had a default of more than 30 days, such as charged-off loans, loans overdue 31–120, and those whose loan status is fully paid but their payment term was longer than initially agreed. The value 0 represents fully paid loans with a payment term equal to or less than initially agreed.

Therefore, out of a total of 806,161 loans, we have a total of 20.20% of operations with early default (171,066), as shown in Fig. 5.

4.4 Data Cleaning and Preprocessing

This is the third stage of the KDD process, which focuses on data cleaning and preprocessing. To do this, the first step was to eliminate variables that provide information on events that occur after loan formalization, and variables that are not relevant to our study. Next, we removed the variables with more than 40% missing values.

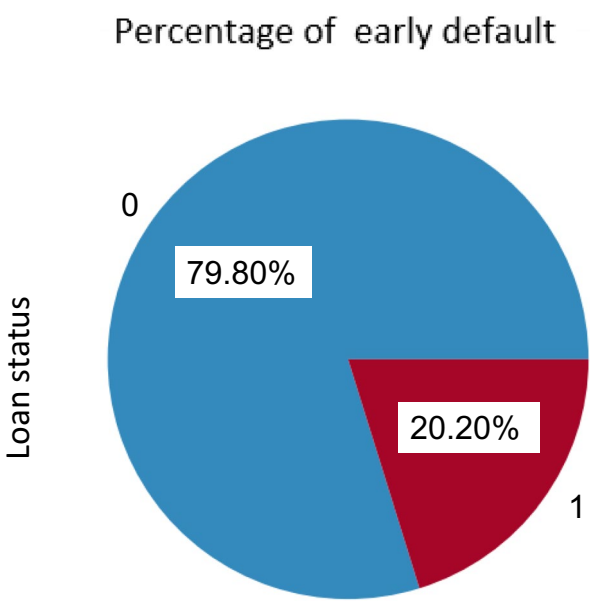

Fig. 5 Percentage of early default

4.5 Data Transformation

This is the fourth stage of the KDD process, which focuses on transforming the data so that the algorithms can work properly with it. To do this, categorical variables were converted to numerical variables (see Fig. 6) using the following methods:

- One Hot Encoding: for each categorical variable W (e.g. verification_status, purpose, etc.), a column is created for each different value that the variable can take. The column W_Z for categorical variable W is 1 or 0 depending on whether the value of W is $_Z$ or not, respectively.
- Label encoding: in this step the labels are converted to a numeric format such as grade: A:1, B:2, C:3, D:4, E:5, F:6, G:7 subgrade A1:1, A2:2, A3:3 A4:4, B1:5, B2:6, B3:7, B4:8, C1:9, C2:10, C3:11 C4:12 …, emp_length: 10 + years: 10, 9 years: 9, 8 years: 8, 7 years: 7, 6 years: 6, 5 years: 5,4 years: 4, 3 years: 3, 2 years: 2, 1 year: 1,’< 1 year: 0.

4.6 Choice of the Most Appropriate ML Algorithm

At this stage of the KDD process, we will proceed to build the predictive system. To do this, we have developed different models in order to select the most predictive

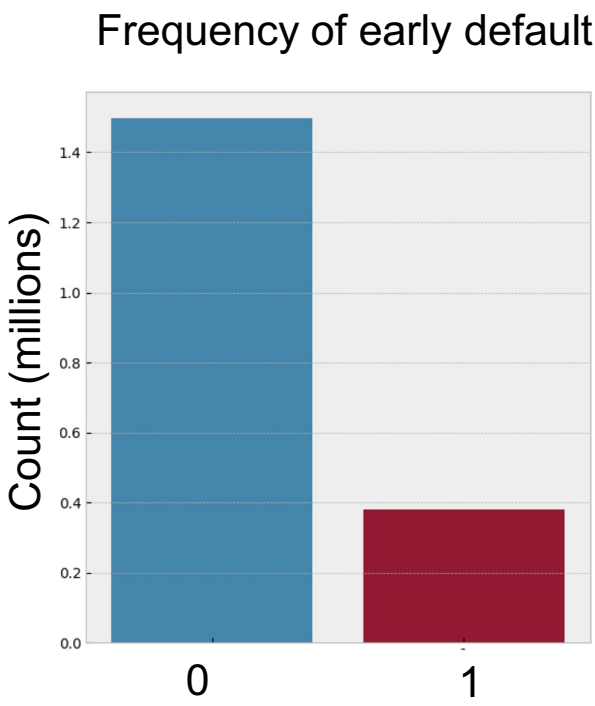

Fig. 6 Frequency of early default

one. To choose the final model, we have evaluated the performance of each model and optimized the hyperparameters using Grid Search, managing to find the optimal combination of hyperparameters that maximize the performance of the model. Finally, the selected model was XGBoost. We have tried to improve the performance of the model using class balancing techniques without success.

The rationale for choosing the XGBoost model for early loan default prediction was as follows:

Performance. Higher AUC and accuracy: XGBoost has the highest AUC value (0.731) among all the models evaluated, indicating a better ability to discriminate between classes. In addition, XGBoost also has the highest accuracy (0.795) among all the models evaluated. Having a model with the highest possible AUC and accuracy is crucial for the financial institution. On the one hand, a higher AUC directly translates into better lending and credit risk management decisions. On the other hand, predicting defaults more accurately allows the financial institution to reduce losses and optimize its risk mitigation strategies.

XGBoost's optimized hyperparameters and explanatory techniques such as surrogate modeling provide clearer insights into the variables contributing to the predictions than other black box models such as DL, which is important for decision making in the context of default probability. These features are critical for an application as critical as early prediction of loan defaults, where each incremental improvement in accuracy can have a significant impact on the financial institution's profitability and risk management.The model comparison is shown in Table 4.

- The optimized hyperparameters of our final XGBoost model are:Minimum sum of instance weight (Hessian): 5
- Maximum depth of a tree: 6
- Output probability:'binary:logistic'
- Gamma: 10. This parameter controls the minimum loss reduction required to perform an additional partition at a leaf node of the tree. The algorithm will be more conservative the higher the value of gamma.

Figure 7 shows the ROC curve in test of the constructed model. *ROC curves for continuous data*. CRC Press).

4.7 Interpreting Mined Patterns

This phase is essential to close the cycle of the KDD process and to convert the results into valuable and actionable information. For this purpose, we have chosen a surrogate model considered a white box as an interpretive model that explains the predictions of the Machine Learning model we built with XGBoost, which is a black –box model. The surrogate model aims to provide a clearer understanding through rules of the relationships between input variables and model predictions, clearly representing the logical decisions from a business perspective.

Table 4 Area Under the Curve (AUC) of the different models considered

	Without grid search		With grid search		Smote		Oversampling		Undersampling	
Models	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
XGBoost	0.730	0. 794	0. 731	0. 795	0. 729	0. 793	0. 662	0. 720	0. 663	0. 744
Ada Boost Classifier	0.715	0. 793	0. 721	0. 791	0. 699	0. 780	0. 7147	0. 642	0. 714	0. 645
Logistic Regression	0.714	0. 792	0. 715	0. 798	0. 679	0. 752	0. 642	0. 598	0. 710	0. 658
Random Forest Classfier	0.711	0. 793	0. 714	0. 791	0. 710	0. 788	0. 713	0. 785	0. 714	0. 651
Decision Tree Classifier	0.694	0. 788	0. 703	0. 787	0. 692	0. 779	0. 551	0. 697	0. 572	0. 574
GaussianNB	0.687	0. 308	0. 687	0. 272	0. 643	0. 401	0. 657	0. 409	0. 658	0. 536
Balanced Bagging Classier	0. 657	0. 714	0. 713	0. 651	0. 668	0. 711	0. 600	0. 766	0. 665	0. 536
KNeighbors Classsifer	0. 642	0. 768	0. 672	0. 785	0. 569	0. 588	0. 551	0. 569	0. 568	0. 548
Deep Learning	0. 542	0. 790	0. 725	0. 793	0. 704	0. 778	0. 563	0. 783	0. 620	0. 215

The inputs to the surrogate model have been a subset of variables that provide key information about the financial situation, credit history and loan details (Term_60_months, Application_type, Delinq_2 yrs, Inq_last_6 mths).

Next, rules from the surrogated model were extracted. In this context, each node and branch of the tree is transformed into conditional rules that are based on the inherent characteristics of the dataset. The combination of nodes and branches allowed for the construction of more complex business-sense rules that span multiple conditions, providing clarity on the fundamental patterns and relationships captured by the model.

The rules we have obtained from the surrogate model are shown in Table 5.

As discussed above in Sect. 3.2, we can represent the continuous numerical variables with their corresponding 2-tuple value if we want to improve interpretability. First, we perform a transformation as a previous and necessary step for the transformation to a 2-tuple value (see Fig. 3). In this sense, we use the rank percent function, which returns the relative percentile rank of each of the predictions within the group, thus transforming the variables to represent them in the interval [0, 1]. The results of these transformations are shown in Table 5.

Once the linguistic representation of the continuous numerical variables had been obtained through a fuzzification process, the values of these variables were replaced in the corresponding rules. It should be noted that the representation of the original variables in linguistic form, completely exact and without loss of information, was possible through the 2-tuple model. Following this process, we obtained the rules shown in Table 6.

S is defined as Fig. 3

Δ_S is defined as Eq. 1.

4.8 Use of Discovered Knowledge

In Sect. 4.6 we obtained an XGBoost model with AUC = 0.73 in the prediction of early default in P2P lending, as can be seen in Fig. 7. In addition, as seen in Sect. 4.7, the model can be interpreted from the business point of view by a set of

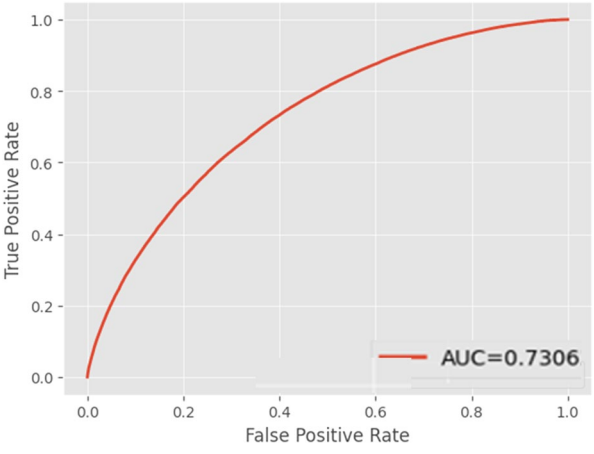

Fig. 7 Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) in the test set

Table 5 Surrogate rules obtained for the XGBoost model

Rule id	Antecedent (IF)	Consequent (THEN)
1	(term_ 60 months = 0) and (application_type ≤ 1) and (delinq_2 yrs ≤ 1) and (inq_last_6 mths ≤ 0) and (bc_util ≤ 67.45) and (acc_open_past_24 mths ≤ 6)	prediction = 0.00
2	(term_ 60 months = 0) and (application_type ≤ 1) and (delinq_2 yrs ≤ 1) and (inq_last_6 mths ≤ 0) and (bc_util > 67.45)	prediction = 0.125
3	(term_ 60 months = 0) and (application_type ≤ 1) and (delinq_2 yrs ≤ 1) and (inq_last_6 mths ≤ 0) and (bc_util ≤ 67.45) and (acc_open_past_24 mths > 6)	prediction = 0.250
4	(term_ 60 months = 0) and (application_type ≤ 1) and (delinq_2 yrs ≤ 1) and (inq_last_6 mths > 0) and (acc_open_past_24 mths ≤ 6)	prediction = 0.375
5	(term_ 60 months = 0) and (application_type ≤ 1) and (delinq_2 yrs > 1)	prediction = 0.5
6	(term_ 60 months = 0) and (application_type ≤ 1) and (delinq_2 yrs ≤ 1) and (inq_last_6 mths > 0) and (acc_open_past_24 mths > 6)	prediction = 0.625
7	(term_ 60 months = 1) and (grade ≤ 4 and (application_type ≤ 1) and (inq_last_6 mths ≤ 0)	prediction = 0.750
8	(term_ 60 months = 1) and (grade ≤ 4) and (application_type ≤ 1) and (inq_last_6 mths > 0)	prediction = 0.875
9	(term_ 60 months = 1) and (grade > 4) and (application_type > 1)	prediction = 1.0

Table 6 AAA2-tuple representation of the predicted target variable and surrogate rules obtained for the XGBoost model with 2-tuple linguistic representation

Rule id	Antecedent (IF)	Consequent (THEN)
1	Δ_S (term_ 60 months = 0) and Δ_S (application_type $\leq$ 1) and Δ_S (delinq_2 yrs $\leq$ 1) and Δ_S (inq_last_6 mths $\leq$ 0) and Δ_S (bc_util $\leq$ Medium) and Δ_S (acc_open_past_24 mths $\leq$ 6)	Δ_S (prediction) = (Very Low)
2	Δ_S (term_ 60 months = 0) and Δ_S (application_type $\leq$ 1) and Δ_S (delinq_2 yrs $\leq$ 1) and Δ_S (inq_last_6 mths $\leq$ 0) and Δ_S (bc_util > Medium)	Δ_S (prediction) = (Very Low,0.125)
3	Δ_S (term_ 60 months = 0) and Δ_S (application_type $\leq$ 1) and Δ_S (delinq_2 yrs $\leq$ 1) and Δ_S (inq_last_6 mths $\leq$ 0) and Δ_S (bc_util $\leq$ Medium) and Δ_S (acc_open_past_24 mths > 6)	Δ_S (prediction) = (Low)
4	Δ_S (term_ 60 months = 0) and Δ_S (application_type $\leq$ 1) and Δ_S (delinq_2 yrs $\leq$ 1) and Δ_S (inq_last_6 mths > 0) and Δ_S (acc_open_past_24 mths $\leq$ 6)	Δ_S (prediction) = (Medium,−0.125)
5	Δ_S (term_ 60 months = 0) and Δ_S (application_type $\leq$ 1) and Δ_S (delinq_2 yrs > 1)	Δ_S (prediction) = (Medium)
6	Δ_S (term_ 60 months = 0) and Δ_S (application_type $\leq$ 1) and Δ_S (delinq_2 yrs $\leq$ 1) and Δ_S (inq_last_6 mths > 0) and Δ_S (acc_open_past_24 mths > 6)	Δ_S (prediction) = (Medium, + 0.125)
7	Δ_S (term_ 60 months = 1) and Δ_S (grade $\leq$ 4 and Δ_S (application_type $\leq$ 1) and Δ_S (inq_last_6 mths $\leq$ 0)	Δ_S (prediction) = (High)
8	Δ_S (term_ 60 months = 1) and Δ_S (grade $\leq$ 4) and Δ_S (application_type $\leq$ 1) and Δ_S (inq_last_6 mths > 0)	Δ_S (prediction) = (Very High,−0.125)
9	Δ_S (term_ 60 months = 1) and Δ_S (grade > 4) and Δ_S (application_type > 1)	Δ_S (prediction) = (Very High)

9 rules (R^2 in training = 0.806, R^2 in test = 0.808). Therefore, we obtain an easy-to-understand explanation for the predictions of the black-box model.

The models classified loan applications into early default risk categories. In this way, the admission process could be automated, with Very Low and Low risk applications being automatically approved, and High or Very High risk applications being automatically denied. The results obtained could be used to make strategic decisions, such as adjusting lending policies to avoid subsequent impacts on provisions and capital, complying with regulatory requirements and launching marketing campaigns aimed at clients with a low probability of early default by offering them preferential conditions.

Therefore, by implementing our strategy, the financial institution would achieve greater transparency in its decision-making processes, improve its relationship with the supervisor, significantly reduce the default rate by improving its financial stability, improve efficiency in the evaluation of loan applications, and increase customer satisfaction by offering them financial products more suited to their needs and/or under better conditions.

5 Discussion and Results

In this section we develop the extracted rules and compare our method with the more classical approaches discussed in Sect. 4.1.

There is a clear business interpretation of the rules extracted by the surrogate model.

In rule 1, the loan requested is for a single person and the term of the loan is three years, so it is short term and the risk of exposure of the entity to default is short, which makes the risk of early default lower. The client has not had a delay of + 30 days, or at most one, in his/her payments in the last 2 years. In addition, the credit utilization ratio is less than or equal to the "average", the number of inquiries to the client's credit history in the last six months is zero so he/she has not requested any product in that period, and the number of accounts opened in the last 24 months is less than or equal to 6, so the client has not requested excessive credit in the form of loans, credit cards or other forms of financing. All this confirms that the use of credit is controlled and therefore the probability of early default of the loan is very low.

In Rule 2 the conditions remain the same, but the credit utilization ratio is higher than"Medium", which implies that the use of available credit in relation to total credit is high and implies poor credit management. Therefore, the probability of early default increases to Very Low at + 0.125.

In Rule 3, the conditions of Rule 1 are maintained, except if the client has more than six open accounts in the last 24 months. An excessive number of open accounts may indicate a higher credit risk. This suggests that the person is looking for credit in a short period of time, and this implies a sign of financial difficulties. Therefore, the probability of default is now"Low".

Rule 4: the loan is requested by an individual and for a term of three years, there is no more than 30 days overdue in the last two years, there are six or fewer open accounts, but there are inquiries in a short period of time. This could indicate a customer who has been looking for credit from different entities and what it

may suggest is that he/she is having financial difficulties. Therefore the probability of early default would increase to"Medium, −0.125".

Rule 5: the number of defaults of more than 30 days in the last two years is greater than 1, so the probability of early default rises to"Medium".

Rule 6: the client has more than six open accounts in the last 24 months and there are inquiries in a short period of time. This denotes that the client has requested credit in a short period of time because he/she may have financial difficulties. Therefore, the probability of early default rises to "Medium, + 0.125".

Rule 7. Although the client who requested the loan has had no risk inquiries in the last six months and has a good rating (A-B-C-D), the term of the requested loan is five years. This entails a higher probability of default because the entity is exposed to a possible default for a longer period of time. Therefore, the probability of early default rises to “High”.

Rule 8. The conditions of rule 7 remain the same, but the customer has had inquiries in the last six months. This could indicate that he/she has been seeking credit, which could be interpreted as a potential risk of over-indebtedness, so the probability of default would rise to below very high:"Very High, −0.125".

Rule 9. The requested loan is for 5 years and the number of applicants is 2, their rating is high risk (E–F-G), so the probability of early default is “Very High”.

We emphasize that the 2-tuple model allowed us, on the one hand, a good linguistic interpretability of the XGBoost model. On the other hand, it gave us precision in the interpretation of the rules, allowing us to separate the default prediction into Very Low, Low, Medium, High, and Very High, but with different symbolic translation, so that we can distinguish that some cases have a lower than others.

In Fig. 8, we present a comparison of the proposed approach with the classical methods discussed in Sect. 4.1. For this, we use models developed with optimized

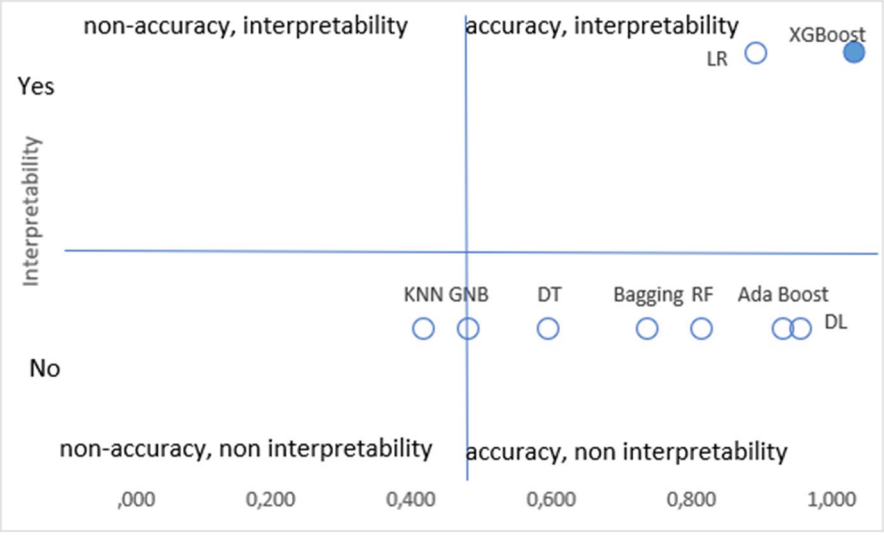

Fig. 8 Comparison of models’ accuracy and interpretability

hyperparameters using Grid Search (Table 4). The X-axis represents the AUC ranking, where we ranked the models using the Rank Function based on their AUC, and the Y-axis indicates whether the models are interpretable or not.

Our approach strikes a balance between accuracy and transparency, delivering superior performance without compromising interpretability, making it an advanced and understandable solution for informed decision-making.

6 Conclusions and Future Work

The problem raised in this article is one of the most important for financial institutions, Fintech or P2P lending platforms. It focuses on extracting linguistic interpretability from a black-box model that estimates having probability of early default on loans. This allows preventive measures to be taken and viable solutions to be sought for both parties before delinquency occurs. For this purpose, we have used a model that cannot be interpreted by humans, such as XGBoost, considered a black-box. Interpretability is crucial for the transparency, traceability and understanding of each and every decision taken in granting credit, as well as for regulators to be convinced of the quality of the models.

We have found works in the literature aimed at predicting defaults, and works that seek to make these models interpretable. To the best of our knowledge, there is no work specifically oriented to the problem of predictability and interpretability of early defaults in P2P loans such as the one presented here.

In this study, we have built an XGBoost algorithm that capable of predicting early loan default, and we have empirically tested it on a large volume dataset with loans granted between 2007 and 2020Q3. Finally, a black-box model that is interpretable thanks to the surrogate model and the 2-tuple fuzzy linguistic model has been constructed. With the 2-tuple fuzzy model, the linguistic interpretability of the generated XAI model was improved without losing accuracy. We have not found in the literature the joint use of both approaches.

From an economic point of view, the interpretability of early default in models that cannot be interpreted by humans makes it possible for the financial institution to detect early default before it occurs. This allows preventive measures to be taken, such as extending the loan term or granting a grace period, seeking viable solutions for both parties before delinquency occurs. On the other hand, it is possible to explain to the regulator the prediction obtained by the model, thus being able to use black-models that are more precise, which means savings in provisions and an improvement in the capital.

Finally, this methodology can be applied in future work to explain default in other non-P2P loans.

Abbreviations *AI*: Artificial Intelligence; *BERT*: Bidirectional Encoder Representations from Transformers; *BEEF*: Backward Elimination Explanation Finder; *BiLSTM*: Bidirectional Long Short-Term memory; *BPNN*: BackPropagation Neural Network; *BS-RF*: Bolasso-enabled Random Forest; *CNN*: Convolutional Neural Network; *CNN2D*: 2-Dimensional Convolutional Neural Network; *DM-ACME*: Distributed Machine Learning—ACME; *DNN*: Deep Neural Network; *EFSVM*: Entropy Fuzzy Support Vector Machine; *GBDT*: Gradient-Boosted Decision Tree; *GCN*: Graph Convolutional Network;

IEFSVM: Instance Entropy Fuzzy Support Vector Method; *KDD*: Knowledge Discovery in Databases; *K-NN*: K-Nearest Neighbors; *LightGBM*: Light Gradient-Boosting Machine; *LIME*: Local Interpretable Model-agnostic Explanations; *NB*: Naïve Bayes; *NN*: Neural Network; *P2P*: Peer to Peer; *REMDD*: Resampling Ensemble Model based on Data Distribution; *RF*: Random Forest; *RoBERTa*: Robustly Optimized Bert; *SHAP*: SHapley Additive exPlanations; *SMOTE*: Synthetic Minority Oversampling Technique; *SSA*: Sparrow Search Algorithm; *SVM*: Support Vector Machines; *UMCDD*: Undersampling Based on Majority Class Data Distribution; *XAI*: EXplainable Artificial Intelligence; *XGBoost*: EXtreme Gradient Boosting; *XGBT*: EXtreme Gradient Boosting Tree

Acknowledgements This work has been supported by the grant PID2022-139297OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU; European Project KA220-HED—Cooperation partnerships in higher education in the Erasmus+ Program in 2022 (Ref: 2022-1-IT02-KA220-HED-000090206); M. Sánchez-Montañés has been supported by grants PID2021-122347NB-I00 and PID2024-155923NB-I00 (MCIN/AEI and ERDF - "A way of making Europe").

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data Availability The datasets analyzed during the current study are available in the Kaggle repository https://www.kaggle.com/datasets/ethon0426/lending-club-20072020q1.

References

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. https://doi.org/10.1002/isaf.325

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S., ... & Khushi, M. (2020). An investigation of credit card default prediction in the imbalanced datasets. *IEEE Access,* 8, 201173–201198. https://doi.org/10.1109/ACCESS.2020.3033784.

Arora, N., & Kaur, P. D. (2020). A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. *Applied Soft Computing, 86*, 105936. https://doi.org/10.1016/j.asoc.2019.105936

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Bologna, G. (2019). A simple convolutional neural network with rule extraction. *Applied Sciences, 9*, 2411. https://doi.org/10.3390/app9122411

Bolton, C. (2009). Logistic regression and its application in credit scoring. Dissertation, University of Pretoria.

Bueno, I., Carrasco, R. A., Porcel, C., & Herrera-Viedma, E. (2022a). Profiling clients in the tourism sector using fuzzy linguistic models based on 2-tuples. *Procedia Comput. Sci., 199*, 718–724. https://doi.org/10.1016/j.procs.2022.01.089

Bueno, I., Carrasco, R. A., Ureña, R., & Herrera-Viedma, E. (2022b). A business context aware decision-making approach for selecting the most appropriate sentiment analysis technique in e-marketing situations. *Information Sciences, 589*, 300–320. https://doi.org/10.1016/j.ins.2021.12.080

Carrasco, R. A., Blasco, M. F., & Herrera-Viedma, E. (2015). A 2-tuple fuzzy linguistic RFM model and its implementation. *Procedia Computer Science, 55*, 1340–1347.

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* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Chen, S. F., Chakraborty, G., & Li, L. H. (2019). Feature selection on credit risk prediction for peer-to-peer lending. In New Frontiers in Artificial Intelligence: JSAI-isAI 2018 Workshops, JURISIN, AI-Biz, SKL, LENLS, IDAA, Yokohama, Japan, November 12-14, 2018, Revised Selected Papers (pp. 5-18). Springer International Publishing, https://doi.org/10.1007/978-3-030-31605-1_1.

Chen, S., Wang, Q., & Liu, S. (2019). Credit risk prediction in peer-to-peer lending with ensemble learning framework. In 2019 Chinese Control And Decision Conference (CCDC) (pp. 4373–4377). IEEE, https://doi.org/10.1109/CCDC.2019.8832412.

Chen, T. (2021). Credit Default Risk Prediction of Lenders with Resampling Methods. In 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI) (pp. 123–127). IEEE, https://doi.org/10.1109/MLBDBI54094.2021.00032.

Chishti, S. (2016). How peer to peer lending and crowdfunding drive the fintech revolution in the UK. Banking beyond banks and money: A Guide to Banking Services in the Twenty-First Century, 55–68, https://doi.org/10.1007/978-3-319-42448-4_4.

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. https://doi.org/10.1109/ACCESS.2019.2896474

Costello, F. J., & Lee, K. C. (2019). Exploring the performance of synthetic minority over-sampling technique (SMOTE) to predict good borrowers in P2P lending. *Journal of Digital Convergence, 17*(9), 71–78. https://doi.org/10.14400/JDC.2019.17.9.071

De Roure, C., Pelizzon, L., & Tasca, P. (2016). How does P2P lending fit into the consumer credit market?, https://doi.org/10.2139/ssrn.2756191.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics, 35*(5–6), 352–359.

Elshawi, R., Al-Mallah, M. H., & Sakr, S. (2019). On the interpretability of machine learning-based model for predicting hypertension. *BMC Medical Informatics and Decision Making, 19*, 146. https://doi.org/10.1186/s12911-019-0874-0

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39*(11), 27–34.

Fendi, U., Sawalha, I., Shamieh, J., & Jaara, O. O. (2017). Early Warning Indicators for Monitoring Non Performing Loans in Jordanian Banking System. *International Journal of Business and Social Science, 8*(6), 104–114.

Fernandez, A., Herrera, F., Cordon, O., del Jesus, M. J., & Marcelloni, F. (2019). Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to? *IEEE Computational Intelligence Magazine, 14*(1), 69–81. https://doi.org/10.1109/MCI.2018.2881645

Fu, X., Ouyang, T., Chen, J., & Luo, X. (2020). Listening to the investors: A novel framework for online lending default prediction using deep learning neural networks. *Information Processing & Management, 57*(4), 102236. https://doi.org/10.1016/j.ipm.2020.102236

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI-Explainable artificial intelligence. *Science Robotics, 4*(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Heng, Y. S., & Subramanian, P. (2022). A Systematic Review of Machine Learning and Explainable Artificial Intelligence (XAI) in Credit Risk Modelling. In Proceedings of the Future Technologies Conference (FTC) 2022, Volume 1 (pp. 596–614). Cham: Springer International Publishing], https://doi.org/10.1007/978-3-031-18461-1_39.

Herrera, F., & Martínez, L. (2000). A 2-tuple fuzzy linguistic representation model for computing with words. *IEEE Transactions on Fuzzy Systems, 8*(6), 746–752. https://doi.org/10.1109/91.890332

Keneni, B. M., Kaur, D., Al Bataineh, A., Devabhaktuni, V. K., Javaid, A. Y., Zaientz, J. D., & Marinier, R. P. (2019). Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles. *IEEE Access, 7*, 17001–17016. https://doi.org/10.1109/ACCESS.2019.2893141

Kim, J. Y., & Cho, S. B. (2019). Towards repayment prediction in peer-to-peer social lending using deep learning. *Mathematics, 7*(11), 1041. https://doi.org/10.3390/math7111041

Kim, J. Y., & Cho, S. B. (2022). Ensemble of diverse deep neural networks with pseudo-labels for repayment prediction in social lending. *Science Progress, 105*(3), 00368504221124004. https://doi.org/10.1177/00368504221124004

Ko, P. C., Lin, P. C., Do, H. T., & Huang, Y. F. (2022). P2P lending default prediction based on AI and statistical models. *Entropy, 24*(6), 801. https://doi.org/10.3390/e24060801

Kriebel, J., & Stitz, L. (2022). Credit default prediction from user-generated text in peer-to-peer lending using deep learning. *European Journal of Operational Research, 302*(1), 309–323. https://doi.org/10.1016/j.ejor.2021.12.024

Krzanowski, W. J., & Hand, D. J. (2009). *ROC curves for continuous data*. Crc Press.

Lee, J. W., Lee, W. K., & Sohn, S. Y. (2021). Graph convolutional network-based credit default prediction utilizing three types of virtual distances among borrowers. *Expert Systems with Applications, 168*, 114411. https://doi.org/10.1016/j.eswa.2020.114411

Lenz, R. (2016). Peer-to-peer lending: Opportunities and risks. *European Journal of Risk Regulation, 7*(4), 688–700. https://doi.org/10.1017/S1867299X00010126

Leo, M., Sharma, S., & Maddulety, K. (2019). Machine learning in banking risk management: A literature review. *Risks, 7*(1), 29. https://doi.org/10.3390/risks7010029

Li, B. (2022). Online loan default prediction model based on deep learning neural network. *Computational Intelligence and Neuroscience, 2022*, 4276253. https://doi.org/10.1155/2022/4276253

Li, W., Ding, S., Chen, Y., & Yang, S. (2018). Heterogeneous ensemble for default prediction of peer-to-peer lending in China. *Ieee Access, 6*, 54396–54406. https://doi.org/10.1109/ACCESS.2018.2810864

Li, H., Cao, Y., Li, S., Zhao, J., & Sun, Y. (2020a). XGBoost model and its application to personal credit evaluation. *IEEE Intelligent Systems, 35*(3), 52–61. https://doi.org/10.1109/MIS.2020.2972533

Li, W., Ding, S., Wang, H., Chen, Y., & Yang, S. (2020b). Heterogeneous ensemble learning with feature engineering for default prediction in peer-to-peer lending in China. *World Wide Web, 23*, 23–45. https://doi.org/10.1007/s11280-019-00676-y

Liu, X., Wang, H., Zhang, K., Lin, K., Shi, Q., & Zeng, F. (2024). Credit Default of P2P Online Loans Based on Logistic Regression Model Under Factor Space Theory Risk Prediction Research. In *International Conference on Intelligent Information Processing* (pp. 410–424). Cham: Springer Nature Switzerland.

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018). 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. https://doi.org/10.1016/j.elerap.2018.08.002

Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications, 42*(10), 4621–4631. https://doi.org/10.1016/j.eswa.2015.02.001

Marín Díaz, G., Carrasco, R. A., & Gómez, D. (2021). RFID: A Fuzzy Linguistic Model to Manage Customers from the Perspective of Their Interactions with the Contact Center. *Mathematics, 9*, 2362. https://doi.org/10.3390/math9192362

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review,* 1–66, https://doi.org/10.1007/s10462-021-10088-y.

Molnar, C. (2020) Interpretable machine learning. Available online: https://christophm.github.io/interpretableml-book/. Accessed 11 May 2024.

Monje, L., Carrasco, R. A., Rosado, C., & Sánchez-Montañés, M. (2022). Deep Learning XAI for Bus Passenger Forecasting: A Use Case in Spain. *Mathematics, 10*, 1428. https://doi.org/10.3390/math10091428

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications, 165*, 113986. https://doi.org/10.1016/j.eswa.2020.113986

Namvar, E. (2013). An introduction to peer-to-peer loans as investments. *Journal of Investment Management First Quarter*. https://doi.org/10.2139/ssrn.2227181

Niu, K., Zhang, Z., Liu, Y., & Li, R. (2020). Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending. *Information Sciences, 536*, 120–134. https://doi.org/10.1016/j.ins.2020.05.040

Polena, M., & Regner, T. (2018). Determinants of borrowers' default in P2P lending under consideration of the loan risk class. *Games, 9*(4), 82.

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. https://doi.org/10.1038/s42256-019-0048-x

Singh, N., Singh, P., & Bhagat, D. (2019). A rule extraction approach from support vector machines for diagnosing hypertension among diabetics. *Expert Systems with Applications, 130*, 188–205. https://doi.org/10.1016/j.eswa.2019.04.029

Song, Y., Wang, Y., Ye, X., Wang, D., Yin, Y., & Wang, Y. (2020). Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in P2P lending. *Information Sciences, 525*, 182–204. https://doi.org/10.1016/j.ins.2020.03.027

Stevens, A., Deruyck, P., Van Veldhoven, Z., & Vanthienen, J. (2020). Explainability and fairness in machine learning: Improve fair end-to-end lending for kiva. In 2020 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 1241–1248). IEEE, https://doi.org/10.1109/SSCI47803.2020.9308371.

Viji, C., Raja, J. B., Ponmagal, R. S., Suganthi, S. T., Parthasarathi, P., & Pandiyan, S. (2020). Efficient fuzzy based K-nearest neighbour technique for web services classification. *Microprocessors and Microsystems, 76*, 103097. https://doi.org/10.1016/j.micpro.2020.103097

Wang, H., Chen, K., Zhu, W., & Song, Z. (2015). A process model on P2P lending. *Financial Innovation, 1*(1), 1–8. https://doi.org/10.1186/s40854-015-0002-9

Xu, J., Lu, Z., & Xie, Y. (2021). Loan default prediction of Chinese P2P market: A machine learning methodology. *Scientific Reports, 11*(1), 18759. https://doi.org/10.1038/s41598-021-98361-6

Yang, R., Wang, P., & Qi, J. (2022). A novel SSA-CatBoost machine learning model for credit rating. *Journal of Intelligent & Fuzzy Systems, 44*(2), 2269–2284. https://doi.org/10.3233/JIFS-221652

Zadeh, L. A. (1975). The concept of a linguistic variable and its applications to approximate reasoning Pt I. *Information Sciences, 8*, 199–249. https://doi.org/10.1016/0020-0255(75)90036-5

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*, 122370. https://doi.org/10.1016/j.physa.2019.122370

Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science, 162*, 503–513. https://doi.org/10.1016/j.procs.2019.12.017