FACULTAD DE ESTUDIOS ESTADÍSTICOS

MÁSTER EN MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS

Curso 2017/2018

Trabajo de Fin de Máster

***TITULO: MODELO DE CLASIFICACIÓN PARA INVERSIÓN EN PRÉSTAMOS* DE BONDORA**

***Alumno*: Verónica Company Blasco**

***Tutor*: Lorenzo Escot Mangas**

Noviembre de 2018

UNIVERSIDAD COMPLUTENSE
MADRID

Índice

1. Planteamiento del problema

Desde la quiebra de Lehman Brothers en septiembre de 2008 y la crisis económica mundial que la siguió, los mercados financieros han restringido sustancialmente el crédito, a la vez que la rentabilidad de los productos de ahorro más conservadores ha caído hasta tipos de interés del 0%.

Muchos particulares y pequeñas empresas han sido expulsados del sistema financiero, tanto en el rol de inversores como en el de prestatarios. En este contexto económico se empiezan a escuchar nuevos vocablos como crowdfunding y crowdlending, surgidos de la economía colaborativa. Según define el colectivo Sharing España, miembro de la Asociación Española de la Economía Digital, en el informe *Los modelos colaborativos y bajo demanda en plataformas digitales*, la economía colaborativa es la economía compuesta por "aquellos modelos de producción, consumo o financiación que se basan en la intermediación entre la oferta y la demanda generada en relaciones entre iguales (P2P o B2B) o de particular a profesional a través de plataformas digitales que no prestan el servicio subyacente, generando un aprovechamiento eficiente y sostenible de los bienes y recursos ya existentes e infrautilizados, permitiendo utilizar, compartir, intercambiar o invertir los recursos o bienes, pudiendo existir o no una contraprestación entre los usuarios".

Es en el año 2009, al calor de estos nuevos modelos económicos, cuando nace en Estonia Bondora, una plataforma digital que ofrece préstamos al consumo no bancarios en tres países europeos: Estonia, Finlandia y España; con inversores que proceden de más de 40 países. Aunque en sus inicios la plataforma sí era una plataforma de economía colaborativa, pues simplemente ponía en contacto a demandantes de financiación y a financiadores en una relación *peer-to-peer*, desde el año 2014, cuando amplió capital y entraron a formar parte de la plataforma fondos de inversión como Global Founders Capital, el modelo de negocio cambió al de economía de acceso.

Según define Sharing España, forman parte de una economía de acceso "aquellos modelos de consumo en los cuales una empresa, con fines comerciales, pone a disposición de un conjunto de usuarios unos bienes para su uso temporal, adaptándose al tiempo de uso efectivo que requieren dichos usuarios y flexibilizando la localización espacial de los mismos". La diferencia entre economía colaborativa y economía de acceso radica en que, en la economía de acceso la plataforma sí presta el servicio subyacente y los usuarios no suelen tener contacto entre sí para efectuar la transacción.

1.1 Qué es Bondora

Bondora es un proveedor de créditos no bancarios cien por cien digital especializado en ofrecer préstamos no garantizados (de mayor riesgo), de importe medio (de 500 a 10.000 euros) y con amortizaciones a corto y medio plazo (de 3 a 60 meses). Su cliente objetivo son consumidores de rentas medias y bajas a los que los bancos o prestamistas convencionales no pueden atender por restricciones regulatorias, de balance o técnicas. A sus clientes-prestatarios les ofrece "una buena experiencia de usuario, con ofertas personalizadas y tarifas y costes justos y transparentes", según citan en su web. También

les ofrece servicios propios como el *B-Secure*, un servicio que permite modificar el calendario de pagos.

Los préstamos son financiados por múltiples inversores, mediante la compra de una participación o la totalidad de un préstamo. El inversor recibe el principal de su participación más los intereses correspondientes. A los inversores Bondora les ofrece una avanzada analítica de crédito que pueden utilizar para invertir en préstamos utilizando el *Portfolio Manager*, una aplicación online con la que el inversor únicamente decide el nivel de riesgo que quiere asumir y la herramienta elige por él los préstamos. Pero Bondora también pone a disposición del inversor la información de todos los préstamos en cartera actualizada diariamente, para que el inversor pueda ser autónomo y tomar sus propias decisiones sobre en qué préstamos invertir y cuánto dinero invierte en cada préstamo.

El atractivo de la plataforma está en que cualquier persona con ahorros puede invertir en los préstamos de Bondora, ya que el inversor decide cuánto invierte en cada préstamo y la inversión mínima por préstamo son 5 euros. Además, al ser una plataforma única que opera en varios países: puede invertir en préstamos de distintos países con un solo click y diversificar el riesgo.

Otra ventaja de la plataforma es que ella misma crea dos mercados: el mercado primario, en el que la plataforma pone a disposición de los inversores los préstamos que ella ha financiado a clientes, y el mercado secundario, en el que los inversores pueden vender sus participaciones en los préstamos a otros inversores.

1.2 Cómo gana dinero Bondora

La plataforma obtiene sus ganancias de la financiación y servicios a los préstamos de consumo:

- Cobra honorarios a los prestatarios por el contrato al amortizar totalmente el principal
- Cobra mensualmente a los prestatarios por la gestión de su cuenta durante todo el periodo de amortización del préstamo
- Cobra tasas de recobro de deuda a los prestatarios cuyos préstamos han pasado por el proceso de recobro, o se deduce el importe de las tasas del flujo de efectivo recuperado
- Cobra mensualmente a los prestatarios que han contratado el servicio B-Secure, que permite flexibilizar las condiciones de pago
- Ingresos por los intereses devengados de los préstamos no vendidos en el mercado secundario

Si se analizan las distintas formas que tiene la plataforma de ingresar dinero, se comprueba que solamente dos de las cinco están alineadas con los intereses de un inversor que, en lugar de contratar un depósito, invierte en Bondora: los honorarios de la amortización total del principal y los intereses devengados.

Las otras tres fuentes de ingresos podrían hacer que se dieran problemas de agencia entre la plataforma y el inversor, ya que a la plataforma podría interesarle que los préstamos se refinanciaran, porque así cobraría más tiempo las tasas por cuenta y por

B-Secure. Incluso puede que a la plataforma le interesara que los préstamos hicieran impagos que sabe que va a recuperar con el proceso de recobro, ya que así ingresaría más dinero por tasas. Para el inversor estas circunstancias no serían buenas porque va a ver el capital de su inversión "retenido" por más tiempo del que él quería cuando invirtió en los préstamos.

1.3 Ciclo de vida de un préstamo en Bondora

El procedimiento de solicitud de un préstamo es el siguiente:

1. El prestatario rellena la solicitud del préstamo en la web. La solicitud es un cuestionario en el que el solicitante tiene que facilitar información personal y de contacto, sociodemográfica, laboral, económica, información de otros préstamos y tiene que adjuntar documentos que respalden la información aportada en el cuestionario.
2. Bondora verifica la información aportada por el solicitante mediante la documentación aportada como fotocopia del DNI, nómina, extractos bancarios… Además, recopila información adicional del solicitante disponible en burós de crédito, bases de datos locales, registro civil y de la propiedad, y también recopila información del comportamiento del solicitante de redes sociales y otros proveedores.
3. La solicitud pasa el proceso de detección de fraude: se realiza la verificación de la identidad del solicitante así como su información de contacto (domicilio, teléfono y email) y se realiza la verificación de la cuenta bancaria.
4. Bondora evalúa el riesgo de la solicitud. Se califica el riesgo de la solicitud en base a los datos recopilados y a los modelos de scoring de Bondora. La calificación de la solicitud va de AA (la mejor) a F (la peor), incluyendo además la calificación HR para las solicitudes de alto riesgo. La calificación se otorga en función de la pérdida esperada a un año vista de la solicitud.
5. Bondora calcula el precio (interés) del préstamo de la solicitud en función de la calificación de riesgo, el calendario de flujos estimado y el retorno estimado.
6. El solicitante firma el contrato del préstamo y la domiciliación SEPA.
7. Se emite el préstamo. Sólo el 10% de las solicitudes de préstamos son aprobadas por Bondora y llegan a emitirse.
8. Bondora ofrece en el mercado primario el préstamo. El mercado primario es la venta de un préstamo de Bondora a sus inversores
9. Los inversores compran total o parcialmente el préstamo
10. Se producen los flujos monetarios
11. Un inversor puede vender su préstamo o su participación en un préstamo en el mercado secundario (mercado entre inversores de Bondora)

1.3.1. ScoreCard de Bondora e interés del préstamo

En el modelo de scoring de Bondora las variables que más pesan, según indican en su web, son las variables totalmente verificadas y, muy especialmente, las variables de comportamiento que obtienen de proveedores de confianza. Una de estas variables es, por ejemplo, el pago de los recibos de telefonía. El rating se basa en la pérdida esperada, no sólo en la probabilidad de impago. También estima la probabilidad de recuperación del préstamo. Esto hace que dos préstamos con la misma probabilidad de impago

tengan calificaciones distintas, incluso que dos préstamos del mismo prestatario tengan calificaciones diferentes, en función de la probabilidad de recobro.

La calificación estima la pérdida esperada a un año vista; según definen, la proporción de interés bruto que el préstamo no recibe debido a las pérdidas del préstamo.

En la siguiente tabla se recogen las pérdidas esperadas mínima y máxima para cada calificación, publicada en su web:

Tabla 1. Pérdida esperada por calificación del rating de Bondora

Risk Rating	Min EL%	Max EL%
AA	0.0%	2.0%
A	2.0%	3.0%
B	3.0%	5.5%
C	5.5%	9.0%
D	9.0%	13.0%
E	13.0%	18.0%
F	18.0%	25.0%
HR	25.0%	>25.0%

En la fase de determinación del tipo de interés del préstamo se utiliza la estimación de pérdida esperada a un año para generar la curva de flujos esperada para la vida del préstamo. El cálculo se realiza utilizando curvas prepago, curvas de impago, incluyendo la estimación de impagos no recuperados, y las curvas de recobro que se estiman para el préstamo. Los datos para la estimación de curvas se recopilaron durante el periodo de contracción económica de la crisis, por lo que están realizadas para un escenario económico adverso. Estas curvas se incorporan al modelo de estimación de flujos para poder ofrecer al prestatario el tipo de interés más ajustado posible, de manera que genere una tasa interna de rendimiento igual al rendimiento esperado en un grupo estático de préstamos con las mismas características que el préstamo de la solicitud.

1.3.2 Solicitantes de préstamos de Bondora

Para poder acceder a un préstamo de Bondora el solicitante tiene que cumplir los siguientes requisitos:

- Ser residente en Estonia, Finlandia o España y tener, al menos, 18 años
- Tener unos ingresos superiores a 300€ si reside en Estonia, 600€ si reside en España y 1.000€ si reside en Finlandia
- Tener un buen historial crediticio en el que no figuren atrasos o incumplimiento en los pagos, insolvencia o algún procedimiento de ejecución de embargo
- No tener antecedentes de conductas de ludopatía
- La suma de sus préstamos con Bondora no puede superar los 11.000€

Los costes que tiene la plataforma para los demandantes de financiación son:

- Comisión de apertura: 5.95% del principal
- Comisión anual de gestión de cuenta: 4%
- Servicio B-Secure (si se contrata): 10€ al mes
- Costes del recobro de deudas

1.3.3 Inversores en préstamos de Bondora

Para un inversor europeo es muy sencillo empezar a invertir en Bondora; únicamente hay que hacerse una cuenta como inversor en la web, rellenar el formulario de identificación, seleccionar la estrategia y traspasar fondos a la cuenta de Bondora.

Los inversores de fuera de la Unión Europea tienen que ser inversores acreditados.

Cuando un préstamo en la cartera de un inversor tiene un impago, el inversor deja de percibir el principal y los intereses del préstamo, aunque no tiene que preocuparse por el recobro de ese impago, ya que es la plataforma la que realiza ese servicio. No obstante, los costes de la recuperación sí se trasladan al inversor, pues se los deducen del importe recuperado. En estos gastos se incluyen los honorarios a las agencias de recobro, honorarios de abogados, tasas administrativas, tasas de los tribunales, etc. Si la deuda no se recupera, no se hacen deducciones.

Si el inversor no quiere esperar al proceso de recuperación del préstamo, puede intentar venderlo en el mercado secundario.

1.4 Proceso de impago y recobro de Bondora

Cuando un prestatario deja de pagar una cuota del préstamo, entra en el siguiente circuito:

Tabla 2. Acciones que realiza Bondora tras un impago

De 1 a 7 días	Se le envían diariamente al prestatario emails, sms y cartas notificadas diariamente. Además, recibe llamadas automatizadas pidiéndole que page la deuda o que se ponga en contacto con Bondora para encontrar una solución
De 8 a 60 días	Se pasa la gestión del cobro a agencias de recuperación de deuda (la propia de Bondora, una externa o ambas en paralelo) para la recuperación de la deuda
De 60 a 75 días	La gestión del cobro vuelve a Bondora, que notifica vía email al prestatario que se le van a cobrar intereses de demora y se van a emprender acciones legales. Se suele otorgar un periodo de gracia de 15 días al prestatario para realizar el pago.
De 75 a 200 días	Si han pasado más de 75 días de sin recibir pagos y cuando el importe de la deuda supera a dos mensualidades del préstamo, Bondora clasifica el préstamo como impagado. Bondora inicia la demanda civil en el caso de Finlandia, ya que el proceso está automatizado y se dicta sentencia en un plazo de 4 meses. En el caso de Estonia y España el proceso es más largo, puede durar hasta 12 meses, por lo que la gestión del cobro se traspasa a agencias locales de recobro. Se publican los detalles del prestatario en los burós de crédito.
Más de 200 días	En el caso de Finlandia, el agente judicial gestiona la liquidación de las deudas del prestatario tras el proceso judicial. En el caso de Estonia y España, si la agencia local de recobro no ha tenido éxito y si se ven posibilidades de recuperación de la deuda por el procedimiento civil, se inicia la demanda civil.

Por tanto, Bondora clasifica un préstamo como impagado cuando han pasado más de 74 días desde el último pago y el importe que se adeuda es igual o superior a dos mensualidades del préstamo.

1.5 Objetivo del trabajo

Bondora es una plataforma online en la que cualquier persona puede invertir en préstamos, convirtiéndose en una alternativa a los productos de ahorro tradicionales que ofrece mayores rentabilidades. Los préstamos en los que se puede invertir mediante

la plataforma no están garantizados, por lo que el inversor está asumiendo mayor riesgo que en los préstamos garantizados. Esto, unido al hecho de que los demandantes de financiación que acuden a la plataforma han quedado fuera del mercado de financiación tradicional (bancos y entidades de crédito ya les han rechazado como clientes) hace que sea imprescindible un modelo de valoración del riesgo de cada solicitante.

Bondora tiene su propio modelo de valoración del riesgo que pone a disposición de los inversores, para que elijan en que préstamos invertir. Dicho modelo es un rating que va desde la AA (riesgo de impago muy bajo) a HR (alto riesgo de impago). No obstante, visto el modelo de negocio de Bondora y de dónde obtiene sus ingresos (principalmente de las tasas y comisiones que cobra a los prestatarios por mantenimiento de cuenta, servicios adicionales y comisiones por mora), pueden darse problemas de agencia entre los intereses de Bondora como compañía (maximizar su beneficio) y los objetivos de los inversores (maximizar la rentabilidad de su inversión) que hace interesante que cada inversor se construya su propio modelo con su propia definición de préstamo malo.

Dado que Bondora lo que hace es vender los préstamos a sus inversores total o parcialmente (sí invierte en los préstamos que no se venden completamente a los inversores) transfiere la mayoría del riesgo por impago a sus inversores. Además, los costes de la recuperación de los impagados los repercute al prestatario y al inversor. Si se tiene en cuenta que, como empresa, quiere maximizar sus ingresos, quizá su modelo de valoración del riesgo no esté alineado con los objetivos del inversor.

Por tanto, el objetivo de este trabajo es, por un lado, testear si el modelo de valoración de la plataforma, el rating, está alineado con los intereses del inversor y, por otro lado, construir un modelo de clasificación, utilizando toda la información que la plataforma pone a disposición de los inversores, que permita al inversor escoger en qué préstamos invertir en función de sus propios objetivos.

Es un problema clásico de clasificación de prestatarios.

2. Metodología

La realización del trabajo está basada en la metodología SEMMA, metodología desarrollada por SAS Institute y cuyo nombre es el acrónimo de sus cinco fases en inglés: *Sample, Explore, Modify, Model* y *Assess*. A continuación, se resumen las tareas ejecutadas en cada una de las fases:

- **Muestra** (*Sample*). La primera fase consiste en seleccionar una muestra representativa de la población sobre la que se va a realizar el estudio del problema. En el caso particular de este trabajo se conoce la información de toda la población y la población total está muy acotada, lo que permite que los sistemas informáticos puedan tratar toda la información. No obstante, de la muestra resultante descrita en el apartado 3 de este trabajo, siempre se utilizará un 70% de los datos para entrenar los modelos y se reservará un 30% para la validación.
- **Exploración de datos** (*Explore*). Consiste en el estudio de los datos, tanto estadística como gráficamente, para comprender mejor el conjunto de datos y extraer información que ayude a la resolución del problema. Además, va a detectar valores faltantes así como anómalos, en cuyo caso habrá que decidir si se excluyen de la muestra o si se tratan. También ser realizará el análisis univariante y multivariante, para ver qué variables será necesario transformar y si hay necesidad de crear nuevas variables, así como se estudiará la relación de las variables independientes con la variable objetivo.
- **Manipulación de variables** (*Modify*). Consiste en la transformación de variables y tramificación de las mismas, en aras de mejorar la eficiencia de los modelos. También se valorará si estas variables transformadas entran a formar parte del modelo en función del estadístico Valor de la Información y el Estadístico de Gini.
- **Modelización** (*Model*). Consiste en la modelización de la relación de las variables independientes con la variable objetivo con el fin de clasificar a los sujetos de la población como buenos o malos. Se modelizarán la regresión logística y los siguientes algoritmos con el método de validación cruzada:
    - Regresión logística
    - Redes neuronales
    - Bagging
    - Random Forest
    - Gradient Boosting
- **Valoración** (*Assess*). Mediante el análisis de bondad de los distintos modelos, se va a elegir cuál es el que mejor clasifica a los prestatarios como buenos o malos, basándose en la menor tasa de error.

No se va a realizar inferencia de denegados porque todos los préstamos que Bondora pone a disposición del inversor ya están concedidos, es decir, en la población no hay denegados. El inversor conoce el comportamiento de todos los préstamos de la plataforma, independientemente de que haya invertido en ellos o no, ya que Bondora expone esta información de manera abierta a cualquier inversor.

2.1 Marco teórico sobre las técnicas de clasificación empleadas[1]

2.1.1 Regresión logística binaria

Es una técnica estadística que estudia la relación causal entre las variables independientes y la variable objetivo, que tiene que ser binaria. La técnica modela la probabilidad de que ocurra el evento sobre la de que no ocurra en función de las variables independientes. Se basa en los Odds Ratios, medida de asociación que proporciona un estimador, basado en un intervalo de confianza, para las relaciones entre variables binarias, cuantificando el cambio en la probabilidad de que la variable estimada pertenezca a una u otra categoría. Es necesario recurrir a métodos iterativos de optimización ya que no existe una fórmula explícita que permita obtener los parámetros que maximizan la verosimilitud.

2.1.2 Redes neuronales

La red neuronal es un algoritmo que intenta imitar el funcionamiento de las neuronas en el cerebro humano; cada neurona procesa y combina estímulos de las otras neuronas con las que está conectada. Con la experiencia, se va formando el mecanismo de memoria del cerebro.

La red neuronal está formada por nodos conectados entre sí, organizados en grupos llamados *capas*. Cada capa se conecta con la siguiente mediante la función de combinación que otorga pesos, que representan la interacción entre los nodos de las capas. Posteriormente se aplica la función de activación, generalmente no lineal, que impone el límite que se debe sobrepasar antes de propagarse a otro nodo.

La redes neuronales aprenden y se forman a sí mismas, luego no necesitan ser programadas de forma explícita. Esto hace que requieran una gran cantidad de datos para poder "entrenar" a la red, que gane en experiencia, y que sea robusta para encontrar la combinación de parámetros que mejor clasificarán la variable.

2.1.3 Bagging

El Bagging, o Agregación de Bootstrap, es un algoritmo de aprendizaje automático que combina la salida de varios árboles mediante el promediado de modelos con la finalidad de mejorar la estabilidad y precisión de los algoritmos de aprendizaje automático. El proceso del algoritmo consiste en crear *n* muestras de los datos originales, posteriormente se crean *m* modelos predictivos para cada muestra y, finalmente, se construye un único modelo predictivo a partir del promedio de los modelos anteriores.

2.1.4 Random Forest

Es un método de aprendizaje conjunto para la clasificación, basado en la construcción de múltiples árboles de decisión en el entrenamiento que incorporan la aleatoriedad en las variables utilizadas para segmentar cada nodo de cada árbol, seleccionando la mejor variable para la partición del nodo. Es una modificación del Bagging que construye una amplia colección de árboles no correlacionados, que luego promedia, que corrigen el sobreajuste de los árboles de decisión a los datos de entrenamiento.

[1] En este apartado, por razones de espacio, solo se hará referencia a los principales elementos de cada técnica. Para un análisis más detallado véase Han, J., Pei, J. & Kamber, M. (2011). *Data mining: concepts and techniques.* Elseveier.

2.2.5 Gradient Boosting

Es un algoritmo de aprendizaje automático utilizado para problemas de clasificación basado en árboles de decisión. El algoritmo se entrena mediante la actualización de los pesos de las observaciones pertenecientes a la clase del evento a través de la optimización en dirección descendente (gradiente negativo) de una función de error determinada. El algoritmo consigue dar mayor relevancia en cada iteración a las observaciones mal clasificadas en iteraciones anteriores.

2.2.6 Validación cruzada

Es una técnica utilizada para evaluar la bondad de ajuste de un modelo que pretende garantizar la independencia entre la partición de datos de la muestra utilizada para el entrenamiento y la partición de datos utilizada para la validación. El método consiste en realizar *n* particiones de la muestra en datos de entrenamiento y datos de validación y, posteriormente, calcular la media aritmética obtenida de las medidas de evaluación sobre las distintas particiones.

3. Selección de la muestra

3.1 Fuente de datos

Los datos son proporcionados por la plataforma de dos maneras:

- Informes públicos: Son las BBDD que la plataforma pone a disposición de cualquier usuario de forma pública en su web https://www.bondora.com/es/public-reports. Hay seis conjuntos de datos:
    - **Loan dataset**: el conjunto diario de datos de los préstamos. Se excluyen los datos protegidos por las leyes de protección de datos. Este conjunto de datos se actualiza cada día.
    - **Historic payments**: datos de todos los pagos recibidos en euros, categorizados por fecha, identificador del préstamo y tipo de pago.
    - **Secondary market transactions history**: datos de todos los préstamos que se han colocado en el mercado secundario de Bondora.
    - **Loan schedules**: datos del calendario de pagos en euros, pasados y futuros, de todos los préstamos categorizados por fecha, identificador del préstamo y tipo de pago.
    - **Debt events**: datos de todos los eventos de recobro de préstamos que han entrado en mora y se han incluido en este proceso.
    - **Portfolio CashFlow, PnL Statement and Balance Sheet**: conjunto de datos de flujos del portfolio total de Bondora, Pérdidas y Ganancias y Balance a cierre de mes. Se actualiza de forma mensual.
- Acceso a los datos a través de API. Hay que tener cuenta en Bondora y estar registrado como usuario de la API. Está indicada para descargas de datos automáticas.

Dado que el trabajo se va a abordar desde el punto de vista de un pequeño inversor con capital limitado que no va a tener necesidad de valorar el riesgo de los préstamos cada día, se utilizará la descarga manual de los datos. Si el objetivo del trabajo fuera ejecutar

el modelo de scoring a diario entonces se utilizarían los datos de la API con un proceso de descarga automático.

El conjunto de datos objeto de estudio es **Loandataset**. Contiene 112 variables y cuenta con más de 60.000 registros. Las variables se pueden agrupar en:

- Variables relacionadas con la solicitud del préstamo (12). Son todas identificativas de las características de la solicitud de préstamo.
- Variables relacionadas con el solicitante (39):
    - 2 variables de tipo identificativo
    - 16 variables de tipo social
    - 17 variables de tipo económico
    - 4 variables de tipo calificación del solicitante (calificaciones externas)
- Variables relacionadas con el préstamo (49):
    - 11 variables identificativas del préstamo
    - 15 variables relacionadas con los pagos de las cuotas del préstamo
    - 14 variables relacionadas con el impago del préstamo
    - 6 variables relacionadas con los recobros del préstamo
    - 3 variables relacionadas con la reestructuración del préstamo
    - 2 variables relacionadas con los fallidos del préstamo
    - 9 variables relacionadas con la calificación que da Bondora al préstamo

La descripción detallada de las variables del set de datos se encuentra en el Anexo I.

3.2 Definición de préstamo malo

Al pequeño inversor/ahorrador le interesa invertir el dinero en los préstamos que no van a realizar impagos ni de principal ni de interés en toda la vida del préstamo. Para ello se va a analizar el comportamiento de los préstamos de Bondora con el objetivo de tener una idea de en qué momento se producen los impagos y qué impagos son difíciles de recuperar.

La variable que nos indica cuando se ha realizado un impago del principal de un préstamo es ***DebtOccuredOn***. La variable ***ActiveLateCategory*** es la categorización de ***DebtOccuredOn***. La variable que nos indica cuándo se ha realizado un impago de los intereses es ***DebtOccuredOnForSecondary***, que no tiene categorización en la BBDD.

Si se compara la categorización de los préstamos que actualmente tienen un impago de principal frente a la peor categorización de impago de principal de los préstamos (su peor impago histórico) se ve que, de los préstamos que alguna vez estuvieron entre 31 y 60 días impagados el 42% están actualmente al corriente de pago, el 32% siguen en ese estado de impago y el 25% tienen un impago inferior a los 31 días. Es para los préstamos que alguna vez estuvieron entre 91 y 120 días en mora en los que observamos que, aunque el porcentaje de los que actualmente están al corriente de pago es similar al de las categorías anteriores, el porcentaje de los que siguen teniendo una mora de entre 91 y 120 días es muy significativo, pues alcanza casi el 60%.

Se observa que la tasa de impagos de la cartera de préstamos es bastante mala. Los préstamos con impagos del principal suponen el 45%.

Tabla 3. Comparativa impago actual vs peor impago

	WorseLateCategory									
ActiveLateCategory	Al corriente	1-7	8-15	16-30	31-60	61-90	91-120	121-150	151-180	180+
Al corriente	100%	60%	47%	41%	42%	39%	41%	31%	36%	28%
1-7		40%	21%	18%	10%	2%	0.1%	0.1%		0.01%
8-15			32%	8%	5%	2%	0.2%		0.1%	0.01%
16-30				33%	10%	3%	0.1%			0.00%
31-60					32%	3%	0.1%	0.6%		0.03%
61-90						50%	0.1%			
91-120							58%	0.3%		
121-150								68%		
151-180									64%	0.00%
180+										72%

Si se considera, además del impago del principal, el impago de los intereses, el número de préstamos al corriente de pago de la cartera disminuye hasta el 50%. Se observa también que, de los préstamos que plataforma considera como Al corriente de pago, una parte tiene impagos en los intereses, lo que reduce la rentabilidad de esos préstamos.

Tabla4. Comparativa impago actual vs peor impago considerando impago de intereses

	WorseLateCategory																			
Active Late Category	Al corriente		1-7		8-15		16-30		31-60		61-90		91-120		121-150		151-180		180+	
	sin imp	imp ppal /int	sin imp	imp ppal /int	sin imp	imp ppal /int	sin imp	imp ppal /int	sin imp	imp ppal /int	sin imp	imp ppal /int	sin imp	imp ppal /int	sin imp	imp ppal /int	sin imp	imp ppal /int	sin imp	imp ppal /int
Al corriente	96%	4%	50%	10%	35%	12%	28%	12%	32%	11%	34%	5%	40%	1%	29%	1%	34%	2%	25%	4%
1-7				40%		21%		18%		10%		2%		0.08%		0.09%				0.01%
8-15						32%		8%		5%		2%		0.24%				0.10%		0.01%
16-30								33%		10%		3%		0.08%						0.00%
31-60										32%		3%		0.08%		1%				0.03%
61-90												50%		0.08%						
91-120														58%		0.26%				
121-150																68%				
151-180																		64%		0.00%
180+																				72%

En ambas tablas se observa que, de los préstamos cuyo peor impago ha sido inferior a 60 días, sólo la tercera parte se mantienen actualmente en su peor impago, el resto o se ha recuperado o vuelven a tener un impago de categoría menor. En cambio, para los préstamos cuyo peor impago ha sido superior a 60 días, se observa que más de la mitad sigue actualmente teniendo el impago de la peor categoría, ya que poco más de la tercera parte se recuperan. Se podría decir que el puno de no retorno en términos de

mora para los préstamos de Bondora se encuentra en los 60 días, es decir, que aquellos préstamos que tienen un impago superior a 60 días es difícil que se recuperen.

Se define como préstamo malo aquél que a día de hoy tiene un impago de principal o intereses.

Se considera un préstamo bueno a aquél que no ha tenido ningún impago ni de principal ni de intereses. No hay préstamos indeterminados.

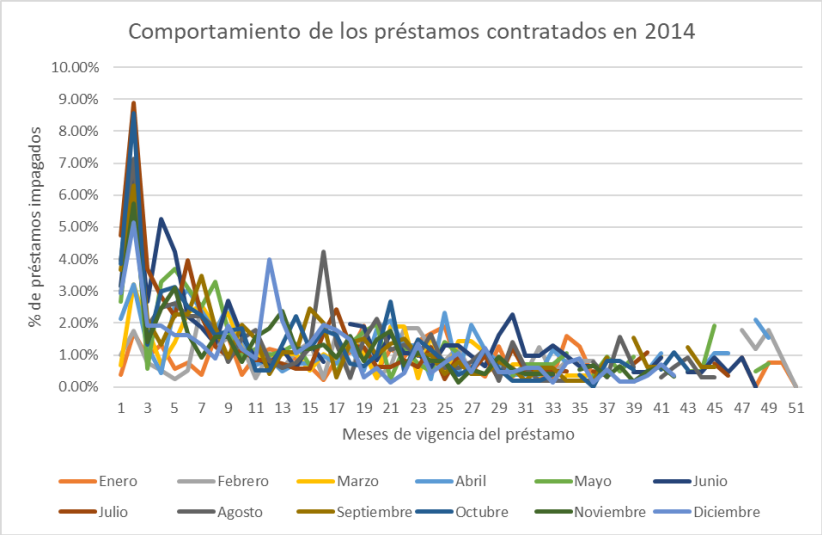
Se crea una nueva variable que se añade a **Loandataset**, *Impago*, que tendrá valor 0 si el préstamo es bueno y valor 1 si el préstamo es malo.

3.3 Definición ventanas temporales

Los modelos de scoring se basan en la idea de que los comportamientos pasados predicen los comportamientos futuros. Para realizar el análisis, se recogen los datos de los préstamos contratados durante una ventana temporal determinada y, posteriormente, se estudia su comportamiento durante otra ventana temporal específica para determinar si son buenos o malos.

Para determinar estas ventanas es interesante conocer en qué momento de la vida de los préstamos se producen, mayoritariamente, los impagos. Se ha construido un gráfico que muestra, para cada mes de contratación de un préstamo, la tasa de impago en función de los meses de su vigencia, considerando como préstamo impagado aquél que tiene la variable ***ActiveLateCategory*** con valores superiores a los 60 días. Por ejemplo, para el año 2014, en esta tabla vemos que las tasas más altas de impagos se producen en los primeros meses de vida de los préstamos:

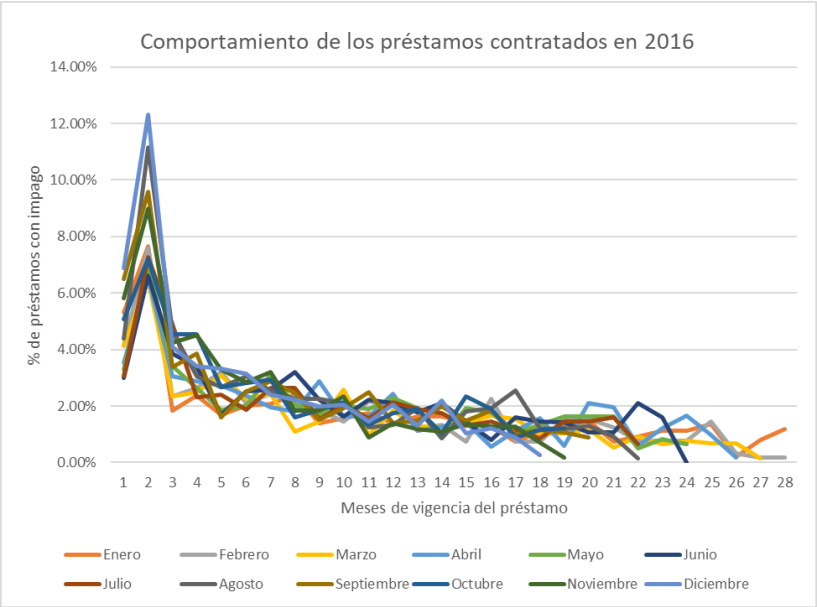
Gráfico 1. Comportamiento de los impagos de los préstamos contratados en 2014



Los préstamos se impagan desde el primer día. Si obtenemos la misma tabla de comportamiento de la mora para préstamos más recientes, por ejemplo, los contratados en 2016, los resultados se repiten: se producen más impagos al inicio de la vida de los

préstamos. Es decir, los prestatarios de Bondora son muy malos; impagan desde el primer momento.

Gráfico 2. Comportamiento de los impagos de los préstamos contratados en 2016



Por tanto, la ventana temporal específica en la que se va a estudiar el comportamiento de los préstamos es 12 meses, ya que en este intervalo de tiempo es en el que se producen la mayoría de los impagos. La ventana de la muestra va a ser todos los préstamos contratados antes del 1 de abril de 2017 y la ventana para el estudio de comportamiento de los préstamos va a ser del 1 de abril de 2017 al 30 de junio de 2017.

El detalle de los inputs para lo gráficos de esta sección se puede encontrar en el Anexo II.

4.Exploración y tratamiento de datos

4.1 Variables a descartar de la muestra

4.1.1 Variables que no se conocerán en el momento en que el préstamo sale a la venta en el mercado primario

La BBDD tiene entre las variables algunas relativas al comportamiento del préstamo que no se conocerán en el momento en que el préstamo sale al mercado primario, como por ejemplo los pagos de principal realizados, fecha de impago, recuperaciones... Se eliminan estas variables. El detalle de las variables excluidas por este motivo se puede encontrar en el Anexo III.

4.1.2 Datos faltantes

Si se analizan los estadísticos básicos de las variables, se observa que algunas de ellas presentan un número elevado de valores faltantes, como las variables cualitativas ***CreditScoreEeMini***, ***CreditScoreEsEquifaxRisk***, ***CreditScoreEsMicroL*** y ***CreditScoreFiAsiakasTietoRiskGra***. La hipótesis es que sean burós de crédito locales y que únicamente estén disponibles para un país. Al obtener las concurrencias de cada variable por país, vemos que se confirma la hipótesis:

Tabla 5. Numero de préstamos valorados por cada buró de crédito

País	Año	CreditScore EeMini	CreditScore EsEquifaxRisk	CreditScore EsMicroL	CreditScore FiAsiakasTietoRiskGra
EE	2012	13	0	0	0
EE	2013	97	0	0	0
EE	2014	3437	0	0	0
EE	2015	3429	0	0	0
EE	2016	6022	0	0	0
EE	2017	11109	0	6939	0
EE	2018	7755	0	7755	0
ES	2013	0	61	61	0
ES	2014	0	1608	1608	0
ES	2015	0	2187	2187	0
ES	2016	0	2642	2642	0
ES	2017	0	4172	4174	0
ES	2018	0	1076	1076	0
FI	2013	0	0	0	143
FI	2014	0	0	0	1352
FI	2015	0	0	0	2424
FI	2016	0	0	0	1848
FI	2017	0	0	2000	2595
FI	2018	0	0	2304	2304

El único score común a los tres países es el ***CreditScoreEsMicroL***, que desde 2017 también califica los préstamos de Estonia y Finlandia. Dado que se ha establecido que

los préstamos contratados a partir de abril del 2017 serán los de la muestra test, quedando por tanto excluidos de la muestra de entrenamiento para realizar el modelo de scoring, estas tres variables hacen que haya que plantearse la posibilidad de hacer un modelo de scoring diferenciado para cada país de los solicitantes. En realidad, el modelo de valoración que proporciona Bondora (variable ***Rating***) ya hace un scoring para cada país.

Otras variables con muchos *missing* son las variables ***Rating_V0***, ***Rating_V1*** y ***Rating_V2***. Estas variables se corresponden con la calificación que Bondora da a los solicitantes. La plataforma ha realizado cuatro versiones del modelo de calificación. Si se analiza el número de préstamos valorados por cada versión, se observa que los préstamos anteriores a 2013 no eran calificados por la plataforma y que todos los préstamos posteriores a 2013 están calificados por alguna de las versiones. Dado que la última versión del modelo de calificación es la variable ***Rating***, se pueden descartar las demás variables de la muestra.

Tabla 6. Número de préstamos valorados por cada modelo de Rating de Bondora

Año	Rating_V0	Rating_V1	Rating_V2	Rating	EL_V0	EL_V1	NumPrestamos
2009	0	0	0	0	0	0	665
2010	0	0	0	0	0	0	1157
2011	0	0	0	0	0	0	451
2012	0	13	0	13	0	13	454
2013	2388	121	2510	2509	2388	121	2510
2014	2181	5261	7161	7442	2181	5261	7455
2015	0	7526	8042	8043	0	7527	8046
2016	0	0	7427	10512	0	0	10514
2017	0	0	0	17931	0	0	17933
2018	0	0	0	11135	0	0	11135

Lo mismo sucede con las variables ***EL_V0*** y ***EL_V1***, que también se descartan.

4.1.3 Variables con valores anómalos

Las variables ***UseOfLoan***, ***MaritalStatus***, ***EmploymentStatus***, y ***OccupationArea*** presentan, para casi la mitad de los préstamos el valor *-1*, valor que no concuerda con ninguna de las categorías que se describen para las variables en la web del set de datos. Además, las cuatro variables presentan este valor de manera simultánea. Analizando los datos se ve que este hecho empieza a darse desde junio de 2017, es decir, que desde junio de 2017 no se están informando correctamente estos cuatro campos para ningún préstamo. Esto sugiere que Bondora ha dejado de recopilar la información de estas variables, seguramente porque no sean variables decisivas en su modelo de scoring.

Tabla 7. Préstamos con las variables UseOfLoan, MaritalStatus, EmploymentStatus, *y* OccupationArea *con valor -1*

Año	Mes	NumUseOfLoan	NumMaritalStatus	NumEmploymentStatus	NumOccupationArea	NumPrestamos
2017	1	0	0	0	0	1141
2017	2	0	0	0	0	840
2017	3	0	0	0	0	1120
2017	4	0	0	0	0	1055
2017	5	3	3	3	3	1152
2017	6	1333	1333	1333	1333	1344
2017	7	1727	1727	1727	1727	1727
2017	8	2061	2061	2061	2061	2061
2017	9	1976	1976	1976	1976	1976
2017	10	1842	1842	1842	1842	1842
2017	11	1879	1879	1879	1879	1879
2017	12	1794	1794	1794	1794	1794
2018	1	1883	1883	1883	1883	1883
2018	2	1540	1540	1540	1540	1540
2018	3	1591	1591	1591	1591	1591
2018	4	1701	1701	1701	1701	1701
2018	5	1729	1729	1729	1729	1729
2018	6	1841	1841	1841	1841	1841
2018	7	837	837	837	837	837

Aunque las variables sí están bien informadas para los préstamos anteriores a junio de 2017, no tiene sentido incluirlas en la muestra que se va a utilizar para construir el modelo ya que, al aplicar el modelo al conjunto de datos de la performance y a las nuevas solicitudes de préstamos no tendremos estas variables bien informadas. Por tanto, se eliminan estas variables.

La variable ***WorkExperience*** presenta un número elevadísimo de datos faltantes, el 40%. De nuevo se comprueba que los *missings* se dan de forma generalizada desde junio de 2017 y que, además, se ven afectados el mismo número de préstamos que en el caso anterior de las variables con valor *-1*. Se obtiene la misma conclusión que en el caso anterior y, por tanto, también se va a eliminar esta variable de la muestra.

Surge la duda de si hay más variables relacionadas con el solicitante que salgan del modelo de valoración de Bondora en su última versión, por lo que se comprueba los valores distintos de las variables relativas al solicitante para los préstamos cuya solicitud se ha realizado con posterioridad al 1 de junio de 2017. Se encuentra que las siguientes variables tienen un único valor para todos los préstamos desde esta fecha, lo que sugiere que también han sido eliminadas del modelo de valoración de Bondora y que no se está recogiendo esta información en las solicitudes, por lo que también se eliminarán del conjunto de datos muestral:

Tabla 8. Variables que presentan un único valor desde junio 2017

Variable	Valor
MaritalStatus	-1
NrOfDependants	
EmploymentStatus	-1
EmploymentPosition	
WorkExperience	
OccupationArea	-1
IncomeFromPrincipalEmployer	0
IncomeFromPension	0
IncomeFromFamilyAllowance	0
IncomeFromSocialWelfare	0
IncomeFromLeavePay	0
IncomeFromChildSupport	0
IncomeOther	0
DebtToIncome	0
FreeCash	0
UseOfLoan	-1

4.1.4 Registros a eliminar de la muestra

La variable ***Age*** tiene, para 53 préstamos, valores inferiores a 18 cuando una de las condiciones que exige la propia plataforma para acceder a los préstamos es ser mayor de 18 años. Claramente son valores informados incorrectamente, ya que tienen valor 0, 1 y 2. Los 53 préstamos son anteriores a 2013. Se eliminan de la muestra ya que se cuenta con registros suficientes de préstamos buenos y malos.

Las variables ***ExpectedLoss*** y ***ExpectedReturn*** sólo se están calculando desde enero de 2013, que es cuando se implanta la versión 1 (la segunda) de su modelo de valoración. Los datos empiezan a tener una calidad suficiente en 2014, además que es el momento en el que la plataforma adopta el modelo actual: la plataforma decide qué préstamos financia en base a su scoring y posteriormente vende estos préstamos a los inversores. Por tanto, para la muestra de entrenamiento que se va a utilizar para la generación del modelo de valoración se incluirán los datos de los préstamos contratados entre el 1 de enero de 2014 y el 31 de marzo de 2017.

La muestra contiene la información de 29.062 préstamos; los préstamos aprobados por Bondora entre el 1 de enero de 2014 y el 31 de marzo de 2017. De ellos el 58.87% han realizado algún impago de principal o intereses y sólo el 41.13% no ha realizado ningún impago.

4.2 Análisis univariante

4.2.1 Variables de la solicitud

Se realiza una primera exploración de las variables de la muestra. Se empieza por las variables relacionadas con la solicitud del préstamo. Estas indican que los préstamos se solicitan principalmente entre semana, siendo el lunes el día que más solicitudes recibe la plataforma, mayoritariamente entre las 10 y las 17h.

Gráfico 3. Histograma ApplicationSignedWeekday

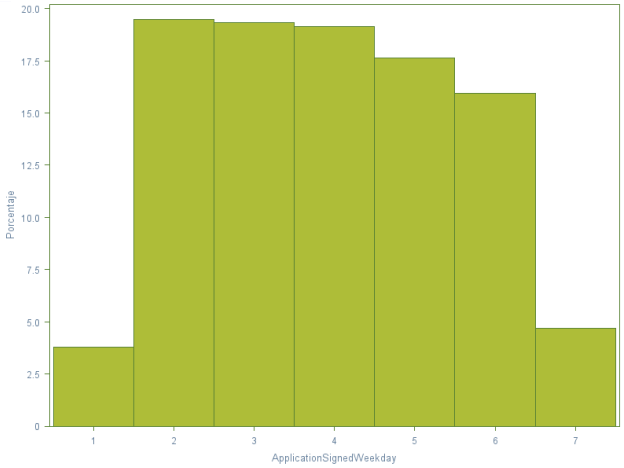


Gráfico 4. Histograma ApplicationSignedHour

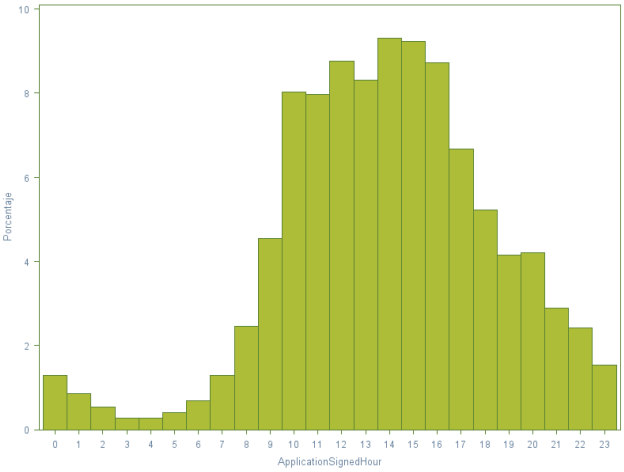


Gráfico 5. Histograma AppliedAmount

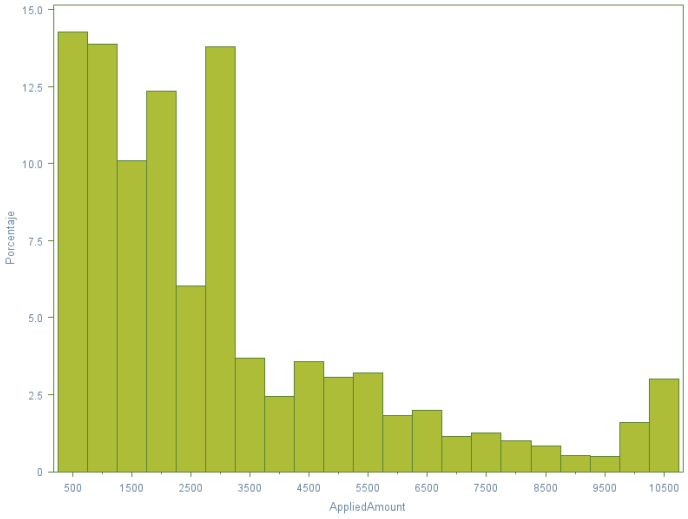
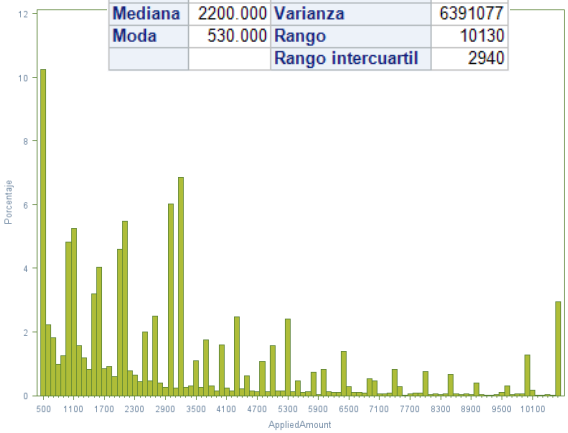


Gráfico 6. Detalle histograma AppliedAmount

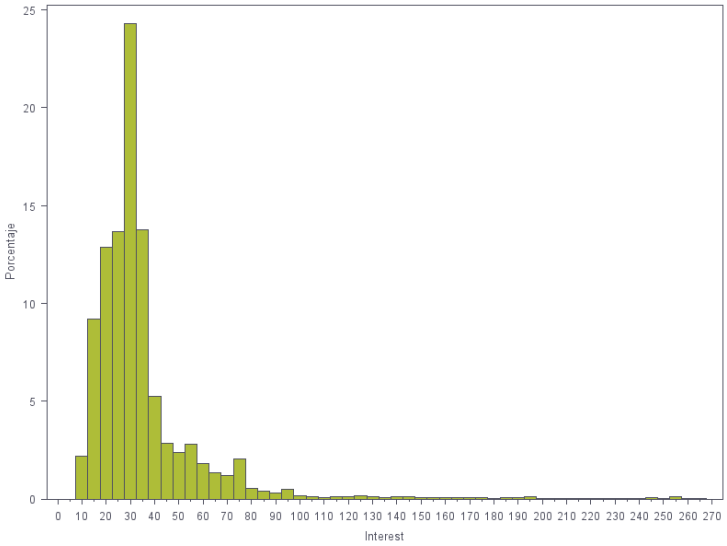
Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	3045.745	Desviación std	2528
Mediana	2200.000	Varianza	6391077
Moda	530.000	Rango	10130
		Rango intercuartil	2940



El préstamo medio solicitado es de 3.046€, aunque el importe más solicitado es de 530€ y la moda son 2.200€. La distribución del importe solicitado no nos dice gran cosa, más allá de que se solicitan más préstamos de menor importe y hay ciertos importes como los millares que se solicitan más.

Con respecto al tipo de interés, se observa que el tipo medio es el 31%, el mínimo es el 7.65% y el máximo es del 263%. Hay 691 préstamos que tienen un tipo de interés superior al 100% y, sorprendentemente, no todos tienen impago. Pese a que son un número pequeño, no es un dato erróneo, por lo que se van a mantener. Es variable candidata a la transformación logarítmica.

Gráfico 7. Histograma y tablas de estadísticos Interest



Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	35.72513	Desviación std	26.20072
Mediana	30.64000	Varianza	686.47791
Moda	31.00000	Rango	256.01000
		Rango intercuartil	13.76000

Observaciones extremas			
Inferior		Superior	
Valor	Observación	Valor	Observación
7.62	8645	262.90	23944
7.91	19906	263.59	15951
7.96	7512	263.59	20294
8.08	23607	263.63	9451
8.08	8288	263.63	9491

El tipo de interés de los préstamos personales de Bondora es realmente alto si se compara con el mercado, hecho que puede ser muy atrayente para los inversores. No obstante, será interesante compararlo con el retorno esperado que estima Bondora; puede que entonces los préstamos ya no sean tan golosos como inversión.

4.2.2 Variables del solicitante

Se va a analizar ahora las variables relacionadas con el solicitante. Lo primero que encontramos es que el 67.33% de los solicitantes son nuevos clientes de Bondora.

Tabla 9. Tabla de frecuencias de NewCreditCustomer

NewCreditCustomer	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
False	9506	32.67	9506	32.67
True	19593	67.33	29099	100.00

Si se analiza la edad de los solicitantes, se ve que la edad media está en los 38 años, la mediana en los 37 y la moda en los 30. Los prestatarios más longevos contaban con 70 años en el momento de solicitud del préstamo. Asimismo, se sabe que más del 50% de los solicitantes son hombres (0), poco más del 40% son mujeres (1) y el resto son de género indeterminado (2). La variable ***Age***, por su distribución, es candidata para una trasformación de raíz cuadrada.

Gráfico 8. Histograma Age

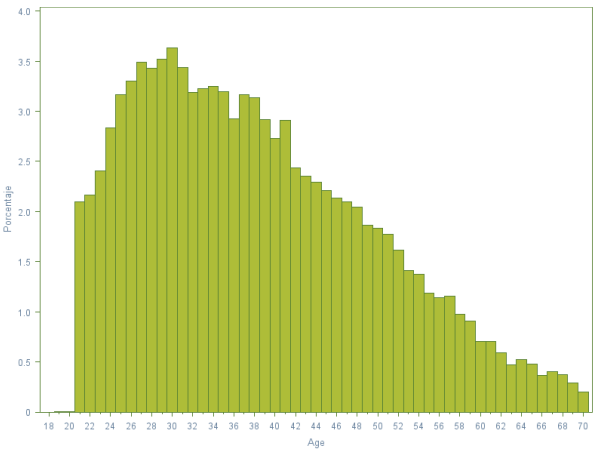
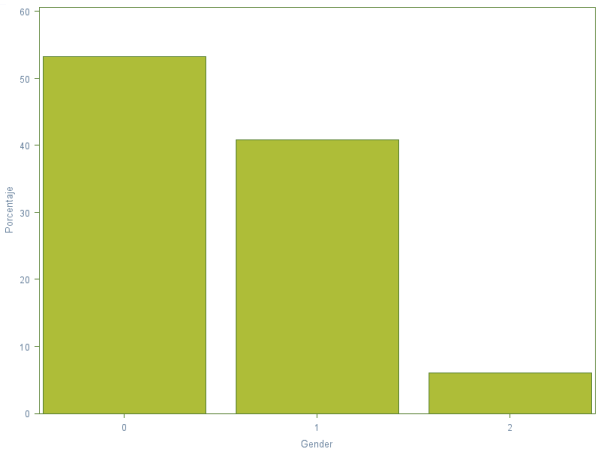


Gráfico 9. Histograma Gender



Con respecto al nivel educativo de los solicitantes, el grupo mayoritario es el de los solicitantes con estudios de secundaria (4), seguido por el de solicitantes con estudios superiores (5) y, a poca distancia, el de solicitantes con formación profesional (3). Con respecto al idioma que utilizaron en la solicitud, es curiosos que entre un 8% y 9% utilizaran el ruso, cuando los préstamos únicamente se comercializan para Estonia, Finlandia y España. El estonio (1) se ha utilizado en el 46% de las solicitudes, el castellano (6) en el 24% y el finés (4) en el 21%. Los porcentajes para la variable ***LanguageCode*** concuerdan con los porcentajes del país de los solicitantes:

Gráfico 10. Frecuencias Education

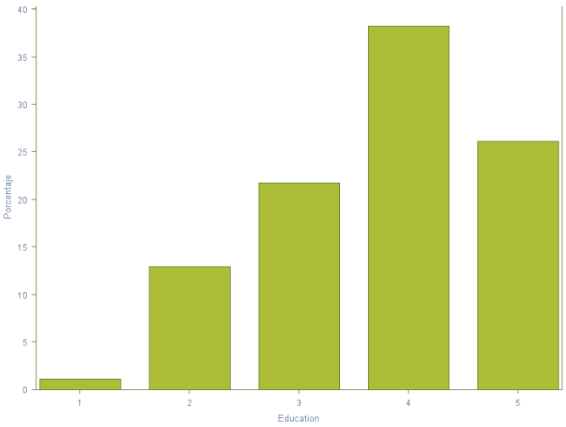
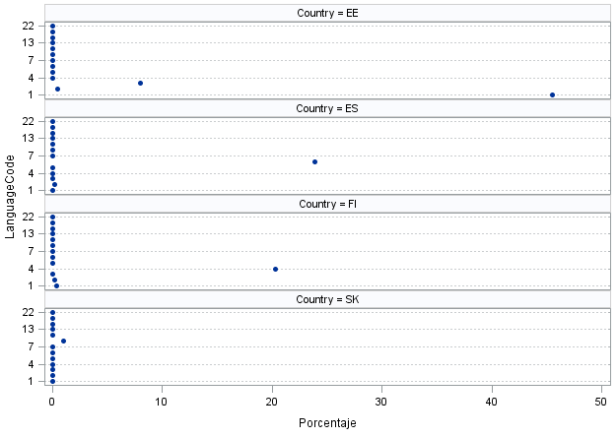
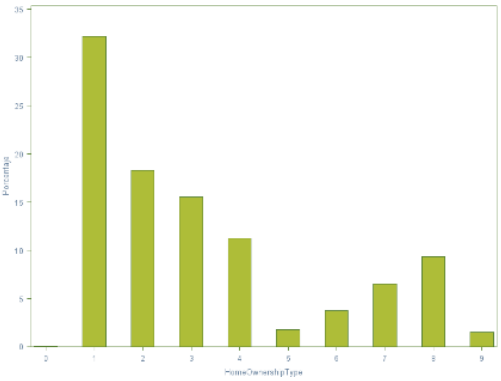


Gráfico 11. Relación País-Idioma



Con respecto al tipo de residencia, la mayoría son propietarios de su residencia, seguidos por el grupo de los que viven en casa de los padres. En tercer lugar, se encuentran los que viven de alquiler en un inmueble amueblado, mientras que los que viven en un inmueble sin amueblar ocupan la cuarta posición. Les siguen los hipotecados, en quinto lugar.

Gráfico 11. Frecuencias HomeOwnershipType



Si se tiene en cuenta su situación laboral, la mayoría de los solicitantes lleva más de 5 años en su puesto de trabajo actual. En esta variable se observa que dos categorías son en realidad la misma (más de 5 años en el puesto actual), luego hay que fusionarlas en una única en la muestra:

Tabla 10. Frecuencias de EmploymentDurationCurrentEmployer

EmploymentDurationCurrentEmployer				
EmploymentDurationCurrentEmploye	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
MoreThan5Years	11215	38.79	11215	38.79
TrialPeriod	485	1.68	11700	40.47
UpTo1Year	5329	18.43	17029	58.91
UpTo2Years	4046	14.00	21075	72.90
UpTo3Years	3290	11.38	24365	84.28
UpTo4Years	2297	7.95	26662	92.23
UpTo5Years	2247	7.77	28909	100.00
Total de valores ausentes = 190				

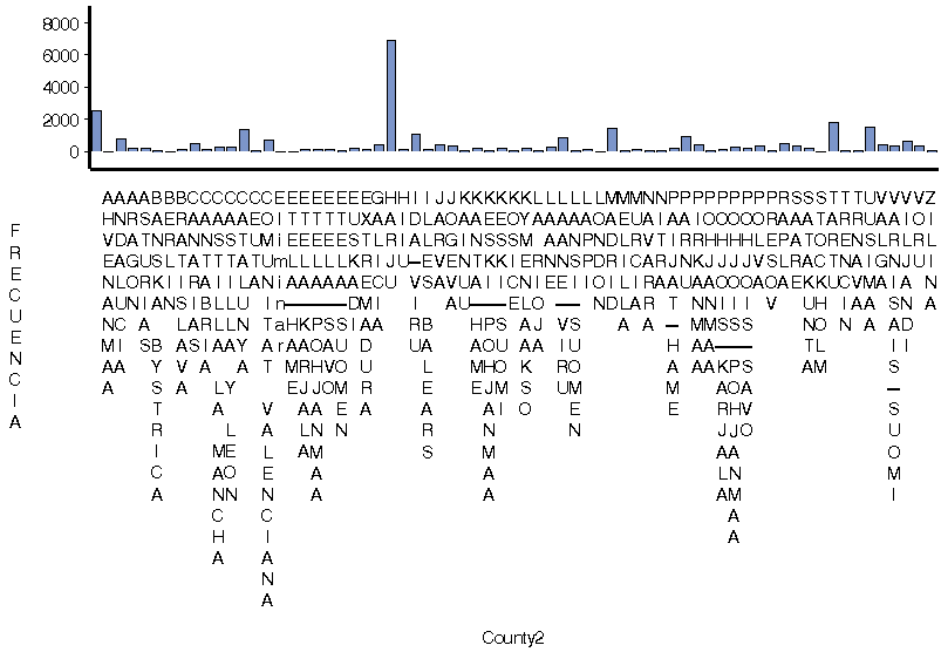
El origen de los prestatarios es otro dato a tener en cuenta. Como ya se ha comentado antes, los solicitantes son, mayoritariamente de Estonia (54%), seguidos de los españoles (24%) y los finlandeses (21%). La base de datos cuenta con dos grados más de granularidad en cuanto al origen: ***County*** (que se interpretará como comunidad autónoma para España y región para Finlandia y Estonia) y ***City***. Ambas variables son campos de texto libre, por lo que están aprovisionadas desastrosamente. Para la variable ***County*** se va a hacer el esfuerzo de armonizar los nombres para tener una variable informativa más, pero en el caso de ***City*** no se va a hacer, en parte porque implica ciudades de otros países con lenguas no latinas bastante difíciles de interpretar.

Tabla 11. Ejemplo aprovisionamiento campo County

County	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
BARCELONA	3	0.01	3	0.01
ETELÃ„-KARJALA	4	0.01	7	0.03
HARJUMAA	1	0.00	8	0.03
POHJOIS-KARJALA	2	0.01	10	0.04
POHJOIS-POHJANMAA	1	0.00	11	0.04
PÃ„RNUMAA	1	0.00	12	0.04
- HARJU MAAKOND -	2	0.01	14	0.05
050-4413197	1	0.00	15	0.06
1	229	0.85	244	0.91
A CORUNA	12	0.04	256	0.95
A CORUÃ‘A	54	0.20	310	1.15
A CORUÃ‡A	1	0.00	311	1.16
A Coruna	1	0.00	312	1.16
A CoruÃ±a	2	0.01	314	1.17
A coruÃ±a	1	0.00	315	1.17
AALICANTE	1	0.00	316	1.17
ADSTURIAS	3	0.01	319	1.19
AHVENAMAA	1	0.00	320	1.19
AKAA	4	0.01	324	1.20

Para armonizar la variable ***County*** en ***County2*** se ha asignado a cada valor de ***County*** el valor correspondiente a las regiones de Finlandia, Estonia y Eslovaquia y las Comunidades Autónomas de España. Realizar este proceso no ha sido nada glamuroso, se ha realizado buscando en Google los campos y buceando hasta encontrar la región. Ha sido un trabajo tedioso (aunque se ha aprendido mucho de geografía Estonia, Finlandesa y Eslovaca…) pero el objetivo es ver si esta variable bien aprovisionada entraría en el modelo. De ser así, el aprovisionamiento de esta variable de manera semiautomática por parte de la plataforma sería una mejora.

Gráfico 12. Frecuencias County2



Se observa que la región con mayores prestatarios es Harju, la región donde se encuentra la capital de Estonia, seguida de Tartu, también región de Estonia, si obviamos los valores *missing*. En Finlandia la región con más prestatarios es Uusimaa, región en la que se encuentra Helsinki, la capital del país y que es la tercera por frecuencia. A esta le siguen Madrid y Catalunya, seguidas de cerca por Andalucía y la Comunitat Valenciana.

Los solicitantes de financiación de Bondora tienen unos ingresos mensuales medios de 1.392€, aunque la moda son 1.000€. Se observa que excepcionalmente hay algún caso de salario mensual que supera los 7.000€ y, aunque para algunos casos los ingresos están verificados, para otros no, por lo que se eliminarán de la muestra los registros con ***IncomeTotal*** superior a 7.000€ que no estén verificados.

Gráfico 13. Histograma IncomeTotal

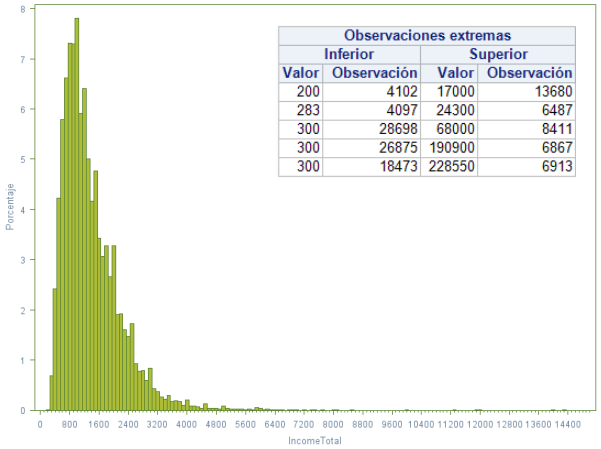
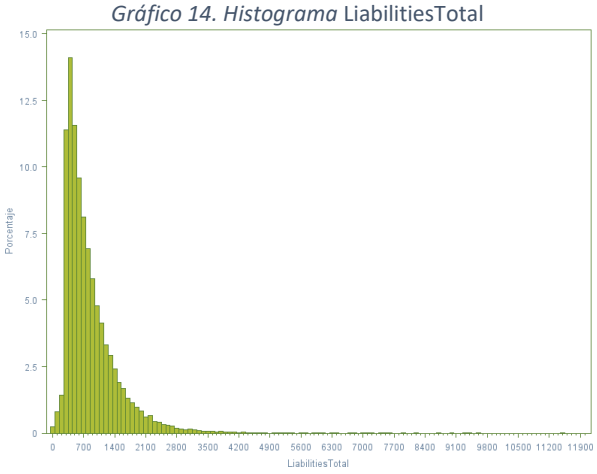


Gráfico 14. Histograma LiabilitiesTotal



Los solicitantes de Bondora tienen unos pasivos mensuales medios de 844€, aunque la mediana son 660€ y la moda 250€. También hay valores excepcionales que superan los 5.000€ de pasivo mensual. Se les va a dar el mismo tratamiento que a la variable ***IncomeTotal***.

Los prestatarios de Bondora, tienen más pasivos, de media 4 más, pero sólo una tercera parte ha tenido previamente otro préstamo en cualquier entidad de crédito. Ambas variables (***ExistingLiabilities*** y ***NrOfPreviousLoansBeforeLoan***) son candidatas a la transformación logarítmica. De nuevo, el histograma del importe de los préstamos

Gráfico 15. Histograma ExistingLiabilities

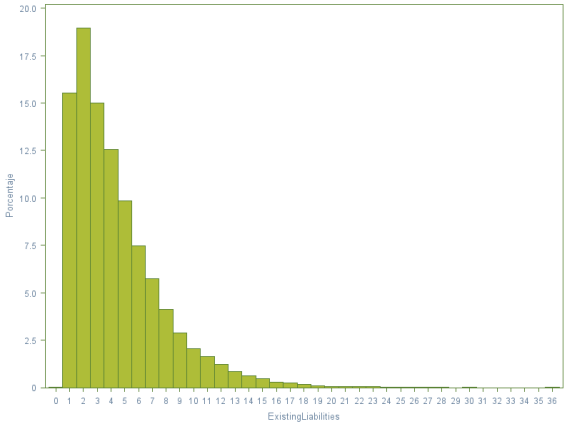
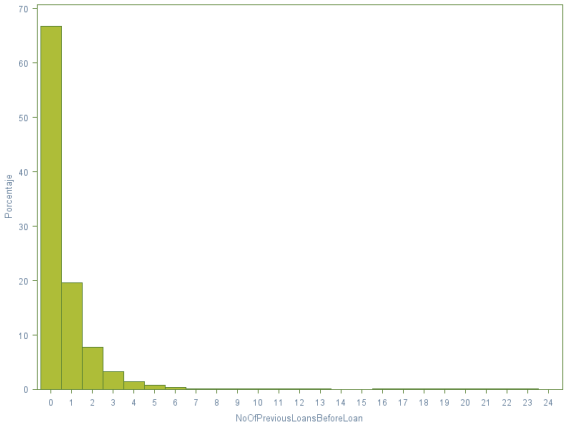


Gráfico 16. Histograma NrOfPreviousLoansBeforeLoan



anteriores tampoco dice mucho, más allá de que hay ciertos importes que se solicitan más.

Por último, si se analiza el comportamiento de los prestatarios en sus préstamos anteriores se ve que se ha ido pagando parte de los préstamos, pero lo cierto es que la variable ***PreviousRepaymentsBeforeLoan*** no dice demasiado, como ***AmountOfPreviousLoansBeforeLoan***. Quizá sería más interesante crear una variable nueva que sea el cociente de ambas y que informe qué porcentaje de los préstamos anteriores se ha pagado ya:

$$PorcRepaymentsPreviousLoans = \frac{PreviousRepaymentsBeforeLoan}{AmountOfPreviousLoansBeforeLoan}$$

Gráfico 17. Histograma AmountOfPreviousLoansBeforeLoan

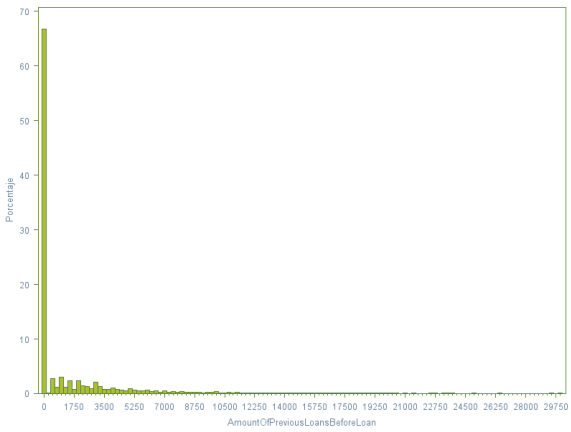
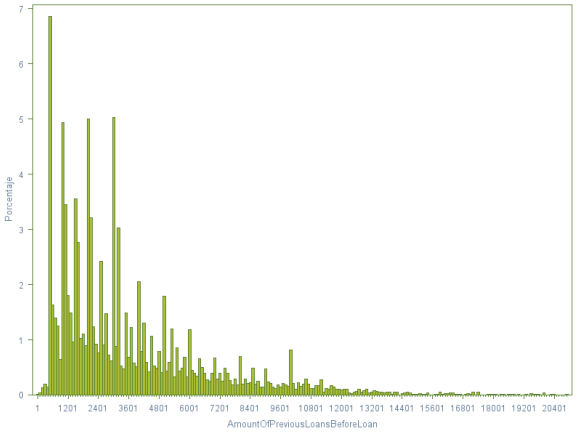


Gráfico 18. Histograma AmounOfPreviousLoansBeroreLoan sin los valores 0.



Lo mismo sucede con las amortizaciones anticipadas. La variable del importe amortizado por sí misma dice poco, es más informativa si dice qué importe del principal de los préstamos anteriores se ha pagado anticipadamente. Por lo que también parece interesante crear una nueva variable que recoja esta información:

$$PorcEarlyRepaymentsPreviousLoans = \frac{PreviousEarlyRepaymentsBeforeLoan}{AmountOfPreviousLoansBeforeLoan}$$

Los comportamientos futuros de los solicitantes intentan predecirlos los ratings de crédito. La plataforma tiene la información de cuatro ratings de crédito. Estos muestran que, en principio, los prestatarios estonios de la plataforma son buenos; el 73% tiene la mejor calificación que indica que no han tenido problemas de pago anteriores. Los

Gráfico 19. Frecuencias CreditScoreEeMini

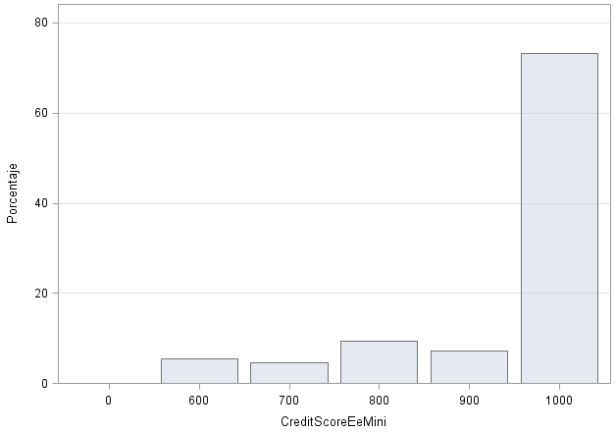
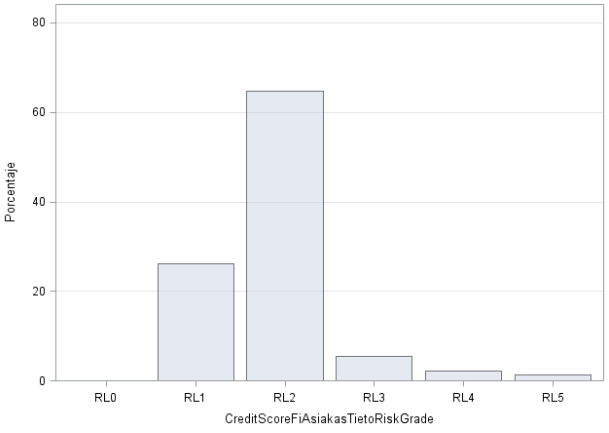


Gráfico 20. Frecuencias CreditScoreFiAsiakasTietoRiskGrade



prestatarios finlandeses, según el rating, parece que tampoco van a hacer grandes impagos, pues el 90% de ellos tiene riesgo bajo o muy bajo.

Los prestatarios españoles son los que salen peor parados en las previsiones; de los dos ratings para españoles, el de Equifax dice que más del 80% de los solicitantes españoles tienen una probabilidad de incumplimiento alta, mientras que el de MicroL (rating específico para prestatarios de alto riesgo) dice que el 60% de los solicitantes se encuentra entre las menores probabilidades de impago, dentro de que son prestatarios de alto riesgo.

Gráfico 21. Frecuencias CreditScoreEsEquifaxRisk

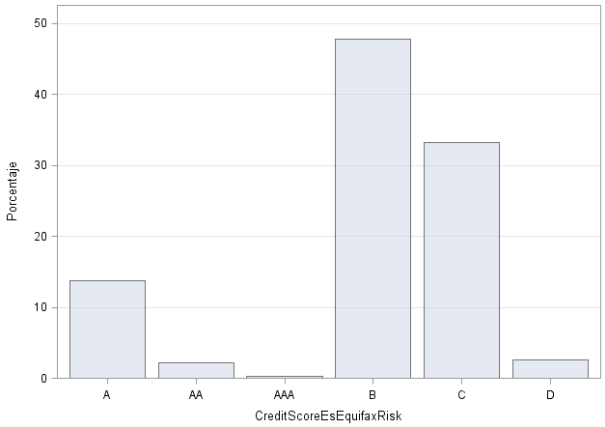
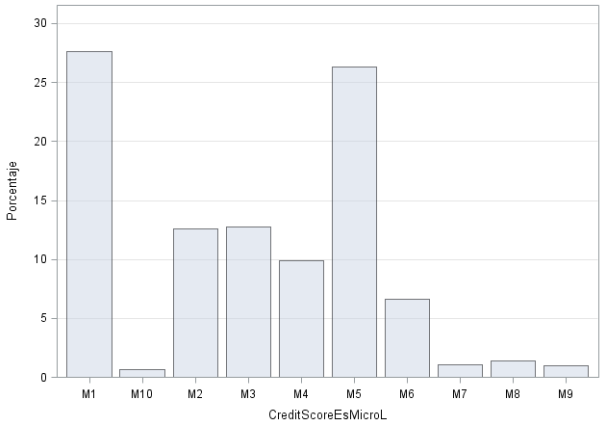


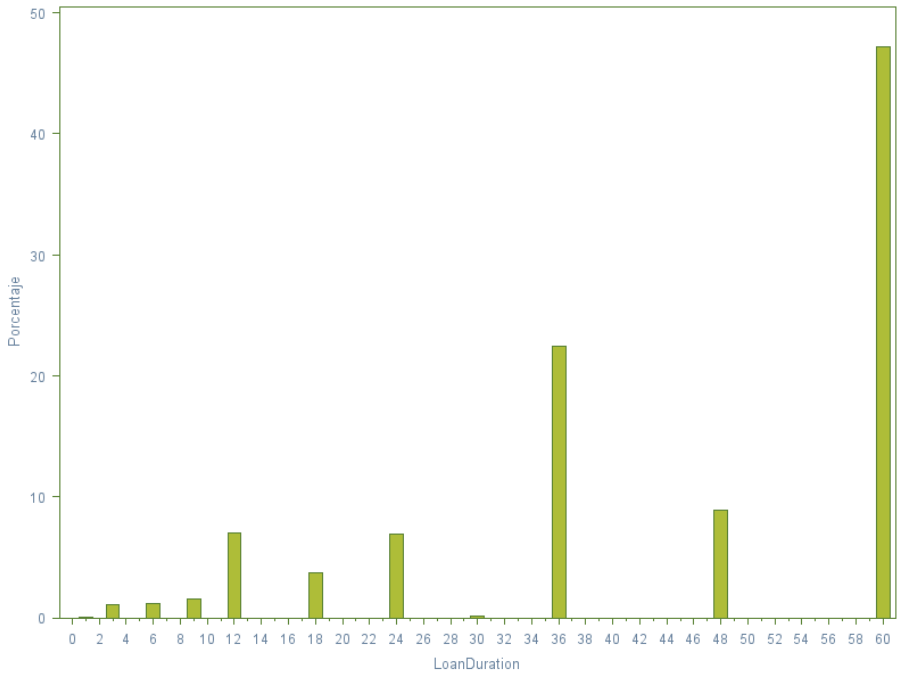
Gráfico 22. Frecuencias CreditScoreEsMircoL



4.2.3 Variables del préstamo

Los préstamos solicitados en Bondora tienen una duración media de 44 meses, pero casi un 50% se solicitan al plazo máximo; 60 meses. El siguiente plazo más popular es el de 36 meses. El plazo para el que menos solicitudes se reciben es el de 30 meses.

Gráfico 23. Frecuencias LoanDuration



La cuota media de los préstamos es de 136€, pero el 60% de los préstamos paga cuotas inferiores a la media. Se observan algunos *outliers* de más de 1.500€ de cuota mensual. Con respecto al día de pago, los más populares son el día 10 y el 15 de mes, seguidos por el día 1. El 75% de los préstamos pagan la cuota antes del día 17 y no hay ninguno que la pague los últimos dos días de mes.

Gráfico 24. Histograma MonthlyPayment

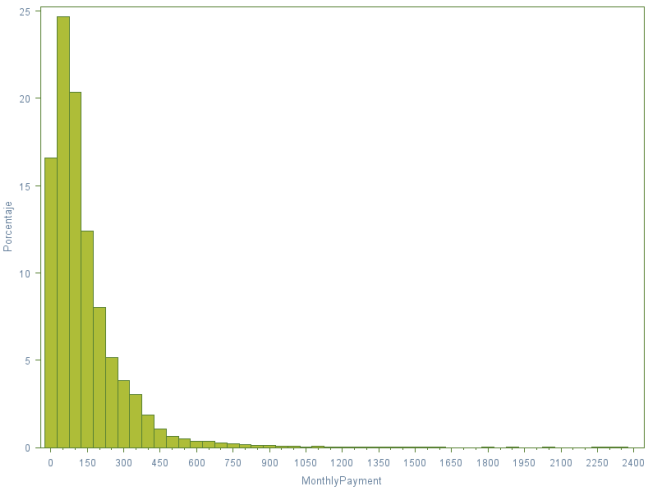
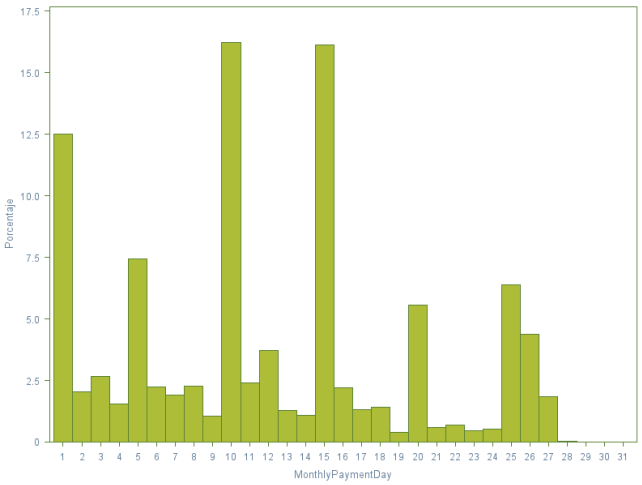


Gráfico 25. Histograma MonthlyPaymentDay



Si se analiza la calificación que da la plataforma a los préstamos que concede, se ve que el 25% de las solicitudes que obtienen financiación están calificadas de alto riesgo. En cambio, sólo un 5.3% de las solicitudes tiene buena calificación (AA y A). La probabilidad de impago de los préstamos que facilita Bondora es, de media, el 23%, pero la moda está en el 17%. Es decir, que la mitad de los préstamos tienen una probabilidad de impago superior al 17%.

Gráfico 26. Frecuencias Rating

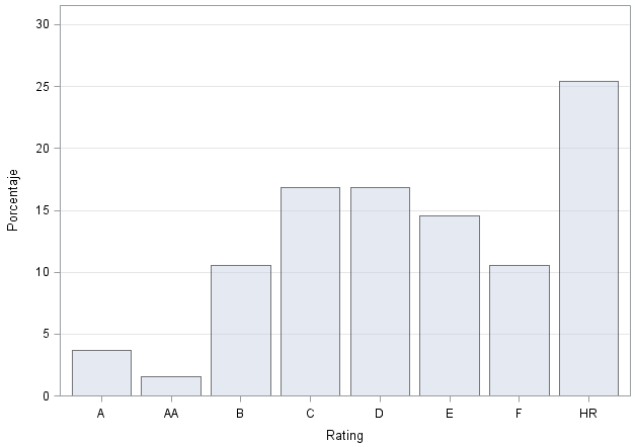
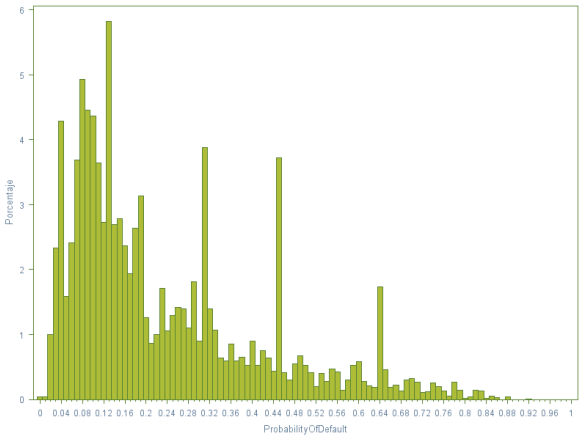


Gráfico 27. Histograma ProbabilityOfDefault



La ratio de pérdida en caso de incumplimiento (LGD) en los préstamos de Bondora es bastante alta, pues la mayoría de los préstamos tienen una LGD de entre el 60% y el 90%. La pérdida esperada (EL) en cambio es un poco más optimista; la media se sitúa en el 19% y la mitad de los préstamos están por debajo del 13% de pérdida esperada. No obstante, se repiten los mismos picos en valores concretos en los que hay más

préstamos con esa pérdida esperada, derivada de los picos en la probabilidad de impago, ya que:

$$EL = PD \times LGD \times EAD$$

$$Pérdida\ Esperada\ (EL) = Probabilidad\ de\ Impago\ (PD) \times Ratio\ Pérdida\ en\ caso\ de\ Incumplimiento\ (LGD) \times Exposición\ en\ el\ momento\ de\ impago\ (EAD)$$

Gráfico 28. Histograma LossGivenDefault

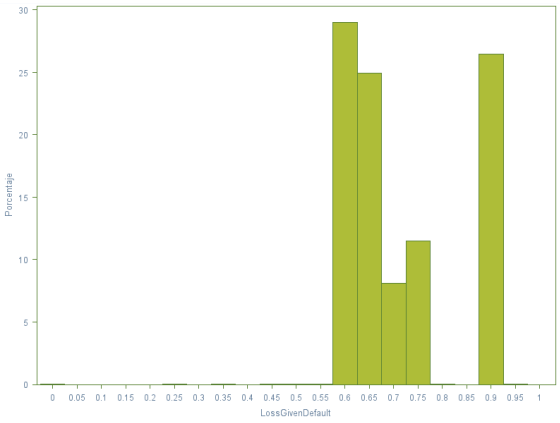
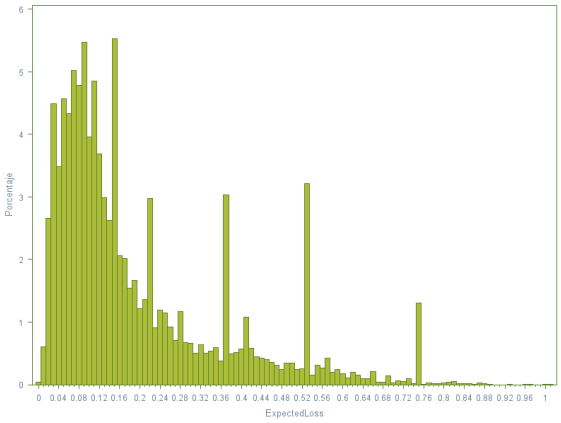
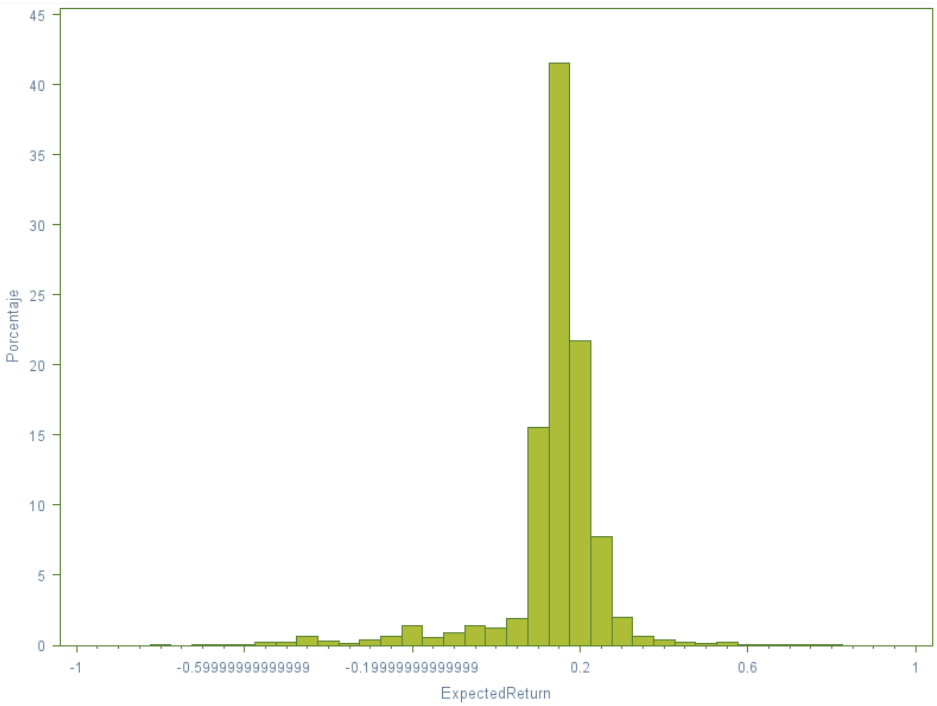


Gráfico 29. Hisograma ExpectedLoss



Por último, el histograma del retorno esperado no es tan alentador para el inversor como prometía el de intereses. Si el tipo de interés medio está en el 35%, el retorno esperado medio que estima la plataforma es del 14%. Además, hay toda una cola a la izquierda del gráfico que indica que hay en torno a un 8% de los préstamos para los que se esperan pérdidas.

Gráfico 30. Histograma ExpectedReturn



4.3 Análisis multivariante

4.3.1 Variables continuas

Si se analiza la correlación entre las variables continuas, se ve que las variables identificativas de la solicitud del préstamo así como las variables de fecha de la solicitud, firma del préstamo y subasta están totalmente correlacionadas. Esto el lógico ya que todas forman parte del proceso de alta del préstamo que se da en un periodo como máximo de 72 horas.

Tabla 12. Correlación entre las variables relacionadas con la solicitud del préstamo

Coeficientes de correlación Pearson, N = 29094 Prob > |r| suponiendo H0: Rho=0							
	LoanNumber	ListedOnUTC	BiddingStartedOn	LoanApplicationStartedDate	LoanDate	FirstPaymentDate	ModelVersion
LoanNumber	1.00000	0.99379	0.99379	0.99385	0.99391	0.99365	0.84980
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
ListedOnUTC	0.99379	1.00000	1.00000	0.99998	0.99998	0.99964	0.88449
	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001
BiddingStartedOn	0.99379	1.00000	1.00000	0.99998	0.99998	0.99964	0.88451
	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001
LoanApplicationStartedDate	0.99385	0.99998	0.99998	1.00000	0.99996	0.99962	0.88453
	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001
LoanDate	0.99391	0.99998	0.99998	0.99996	1.00000	0.99966	0.88372
	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001
FirstPaymentDate	0.99365	0.99964	0.99964	0.99962	0.99966	1.00000	0.88295
	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001
ModelVersion	0.84980	0.88449	0.88451	0.88453	0.88372	0.88295	1.00000
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	

Se observa que la correlación entre ***LoanDate*** y ***FirstPaymentDate*** es totalmente lineal y que la correlación entre ***LoanDate*** y ***LoanNumber*** es cuasilineal, esto es así porque el ***LoanNumber*** se asigna a todas las solicitudes de préstamo independientemente de que Bondora la financie o no, y las que no financia no están en nuestra base de datos.

Gráfico 31. Correlación entre LoanDate *y* FirstPaymentDate

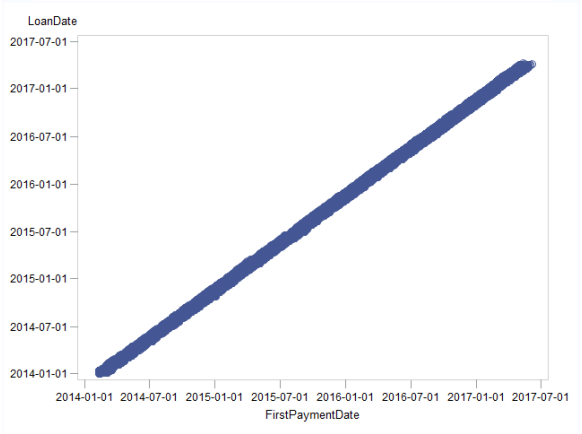


Gráfico 32. Correlación entre LoanDate *y* LoanNumber

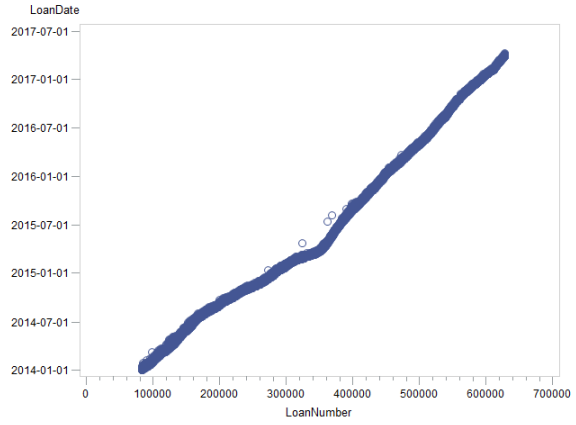
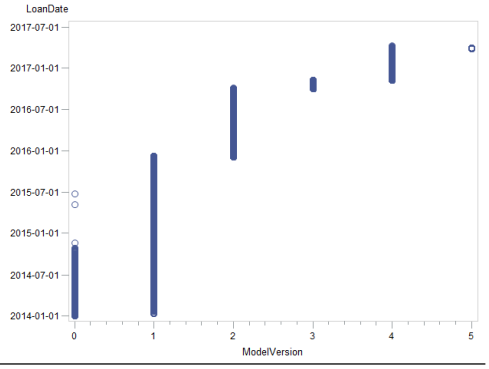


Gráfico 33. Correlación entre LoanDate *y* Model Version

La otra variable que presenta una fuerte correlación es ***ModelVersion***, lo cual también es lógico, ya que según la fecha de solicitud Bondora ha utilizado el modelo de valoración vigente en esa fecha.

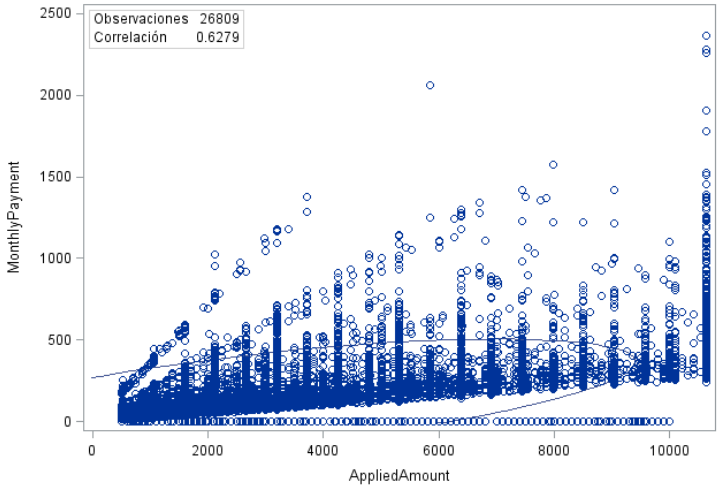


Por tanto, de todas estas variables únicamente se van a mantener ***LoanData*** y ***ModelVersion***, para evitar problemas de multicolinealidad.

Asimismo, otras variables continuas que aparentemente podrían tener relación, como la edad y los ingresos (si se supone que, a mayor edad, mayor experiencia profesional, mayor salario y, por tanto, mayores ingresos) no presentan correlación en los datos de Bondora.

Lógicamente, entre el importe solicitado y la cuota a pagar del préstamo hay cierta correlación positiva, más cuando los plazos disponibles para devolver el préstamo están preestablecidos:

Gráfico 34. Correlación entre AppliedAmount *y* MonthlyPayment

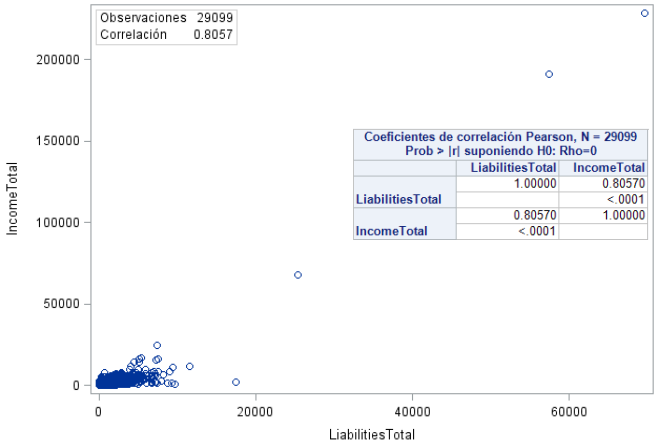


Coeficientes de correlación Pearson, N = 26809
Prob > |r| suponiendo H0: Rho=0

	AppliedAmount	MonthlyPayment
AppliedAmount	1.00000	0.62786
		<.0001
MonthlyPayment	0.62786	1.00000
	<.0001	

También están correlacionados positivamente los ingresos totales de los solicitantes y sus pasivos totales. Además, si se construye una nueva variable que sea el cociente entre los pasivos totales y los ingresos totales, a la que se llama ***PorFreeCash***, para eliminar la multicolinealidad de estas dos variables, la nueva variable también está correlacionada positivamente con el número de pasivos del solicitante.

Gráfico 35. Correlación entre LiabilititesTotal *e* IncomeTotal



Coeficientes de correlación Pearson, N = 29099
Prob > |r| suponiendo H0: Rho=0

	LiabilitiesTotal	IncomeTotal
LiabilitiesTotal	1.00000	0.80570
		<.0001
IncomeTotal	0.80570	1.00000
	<.0001	

Gráfico 36. Detalle de correlación entre LiabilitiesTotal *e* IncomeTotal *si eliminamos las observaciones extremas*

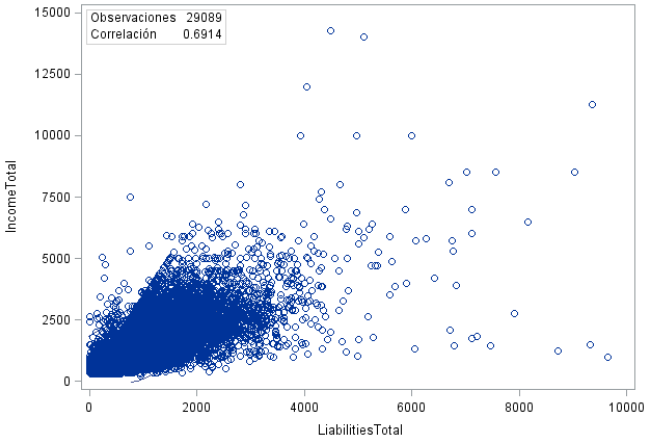
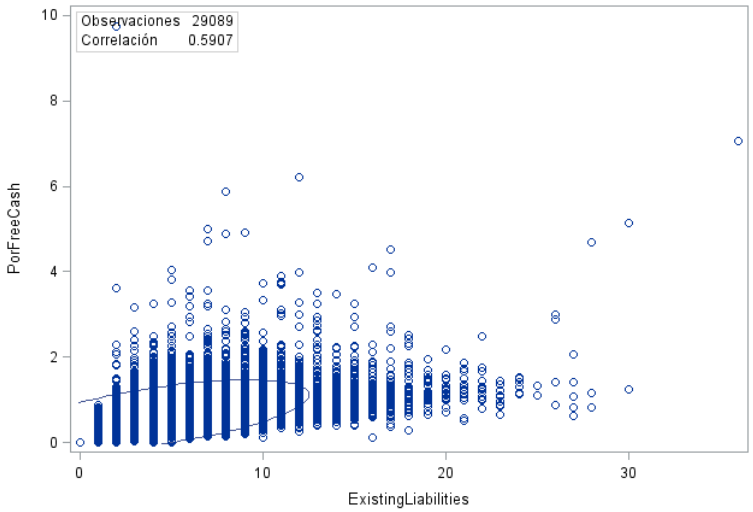


Gráfico 37. Correlación entre ExistingLiabilities *y* PorFreeCash



Si se analizan las correlaciones entre el tipo de interés del préstamo y la probabilidad de impago, la pérdida esperada y el retorno esperado, se ve que existe correlación positiva, aunque no es lineal. Además, por los gráficos parece que hubiera distintas relaciones entre ellas, es decir, se sabe que la plataforma ha utilizado distintos modelos para calcular estas métricas y parece que los distintos modelos se ven en los gráficos de dispersión:

Gráfico 38. Correlación entre Interest *y* ProbabilityOfDefault

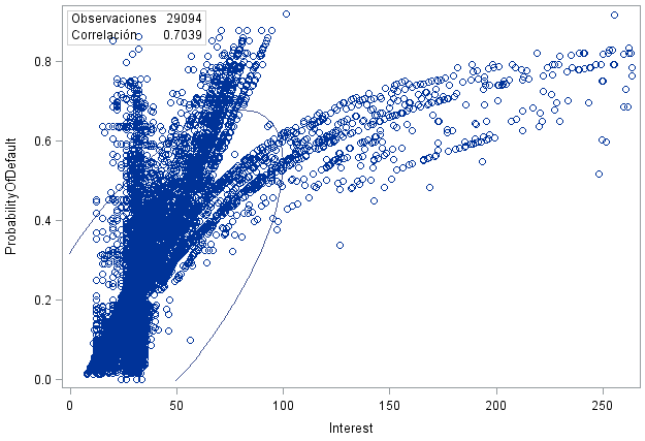


Gráfico 39. Correlación entre Interest *y* ExpectedReturn

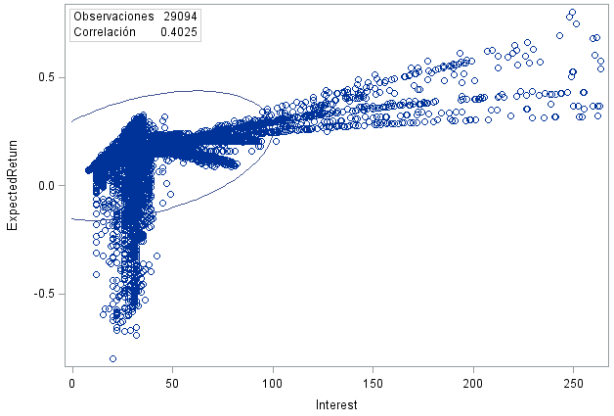
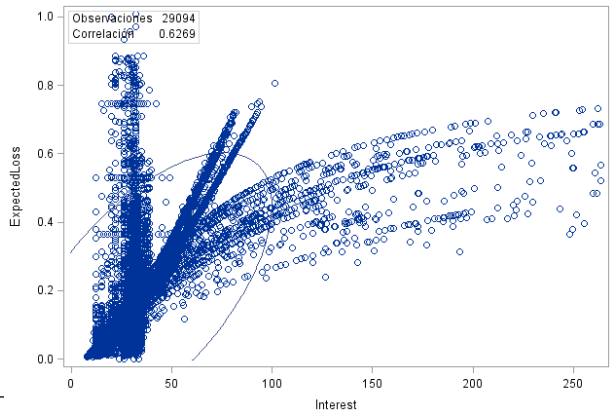
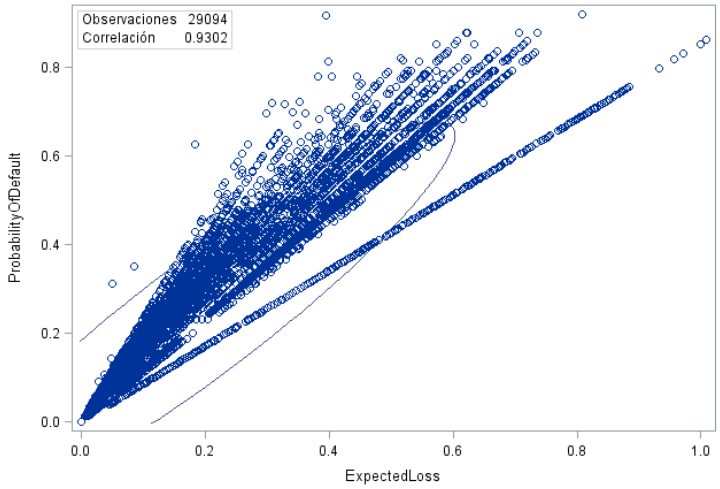


Gráfico 40. Correlación entre Interest *y* ExpectedLoss



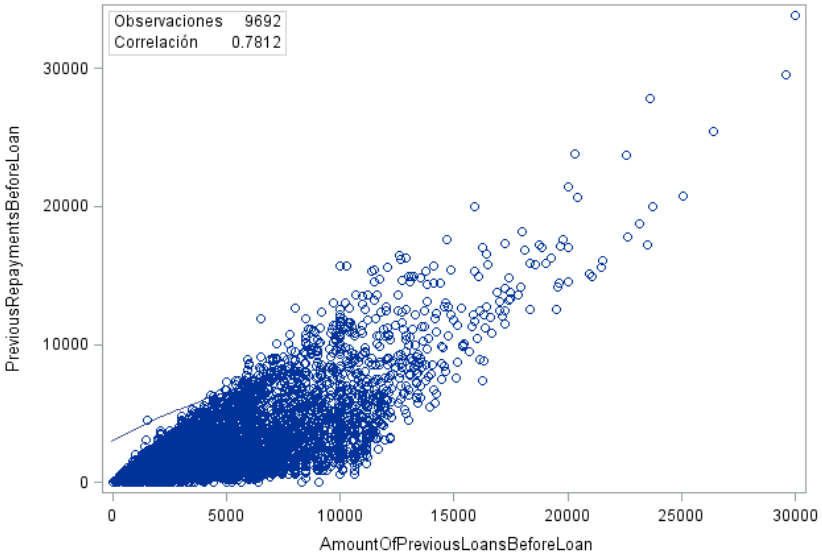
Ya se ha dicho anteriormente que la variable ***ExpectedLoss*** es función de las variables ***ProbabilityOfDefault*** y ***LossGivenDefault***. Tal y como se observa en el gráfico de dispersión hay una correlación lineal clara, por lo que se elimina ***ExpectedLoss*** de la muestra para eliminar problemas de multicolinealidad:

Gráfico 41. Correlación entre ExpectedLoss *y* ProbabilityOfDefault



Otras variables que también presentan correlaciones son las referidas a los préstamos anteriores a la solicitud de los prestatarios. En particular, el principal de los préstamos y el principal ya amortizado.

Gráfico 42. Correlación entre AmountOfPreviousLoansBeforeLoan *y* PreviousRepaymentsBeforeLoan



Para mitigar el efecto de la multicolinealidad se va a crear la nueva variable

$$PorcRepaymentsPreviousLoans = \frac{PreviousRepaymentsBeforeLoan}{AmountOfPreviousLoansBeforeLoan}$$

que es el porcentaje ya amortizado de los préstamos.

Aunque las variables ***PreviousEarlyRepaymentsBeforeLoan*** y ***AmountOfPreviousLoansBeforeLoan*** no presentan apenas correlación, puede ser interesante crear una nueva variable siguiendo el modelo de la anterior, que informe sobre el porcentaje de los préstamos que se ha amortizado anticipadamente:

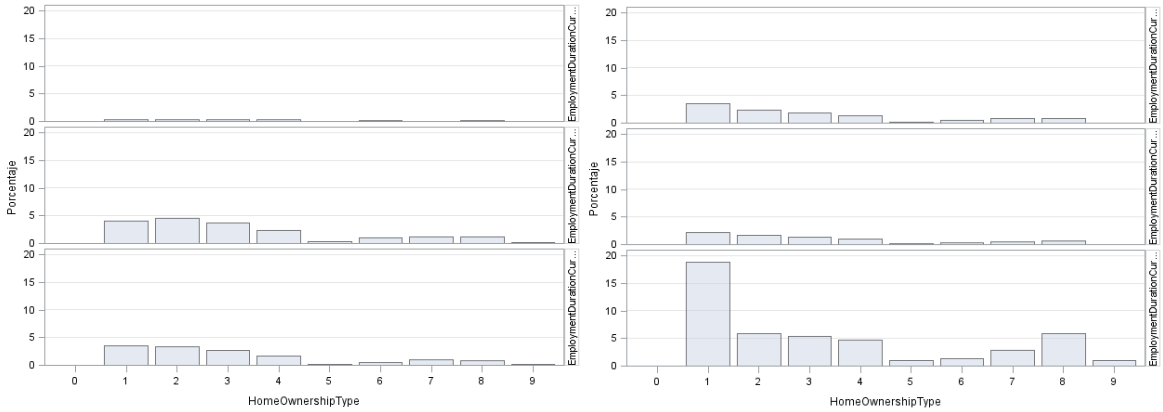
$$PorcEarlyRepaymentsPreviousLoans = \frac{PreviousEarlyRepaymentsBeforeLoan}{AmountOfPreviousLoansBeforeLoan}$$

4.3.2 Variables categóricas

Se realiza una comparativa de la distribución de los datos en las variables categóricas para intentar desentrañar mayor conocimiento sobre los prestatarios. Se intenta ver si los solicitantes de préstamos en Bondora cumplen con las características estándar de asociación entre las variables o si tienen alguna peculiaridad.

Para empezar se analiza la relación entre el tipo de vivienda y la duración del empleo actual. Se observa que, cuando la duración es menor de dos años, el tipo de vivienda más común entre los solicitantes es la casa de los padres (2), seguidos por los que viven de alquiler (3 y 4). En cuanto la duración del empleo supera los dos años, empieza a aumentar el número de solicitantes con la vivienda en propiedad (1, no hipotecada), aunque el tipo de vivienda favorita de los solicitantes sigue siendo el alquiler y la vivienda hipotecada (8) todavía queda muy por debajo de la casa de los padres. La tendencia creciente de la vivienda en propiedad se incrementa hasta que la duración del empleo es mayor de cinco años, donde ya casi el 20% de los prestatarios tiene la vivienda en propiedad, el número de prestatarios que viven de alquiler es la mitad de los que tienen la vivienda en propiedad y el número de solicitantes que viven en casa de los padres o que tienen una hipoteca es el 5%.

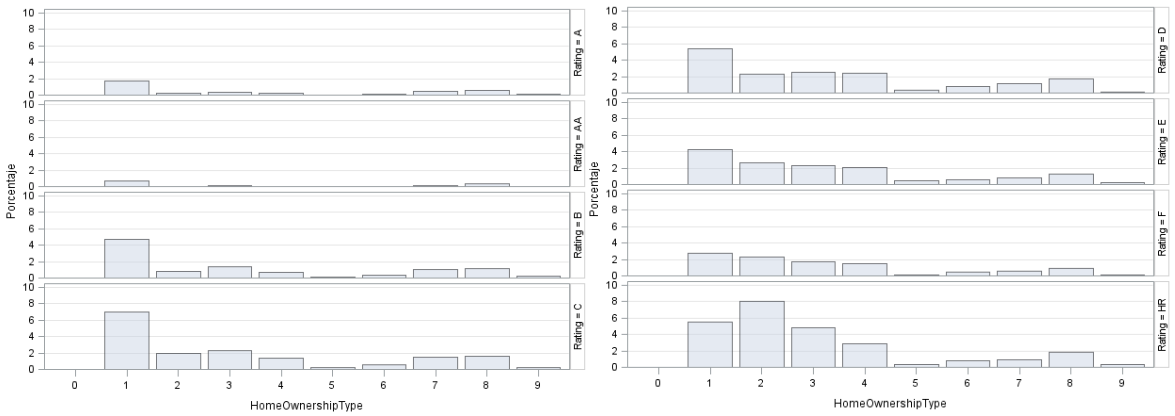
Gráfico 43. Frecuencias de HomeOwnershipType *por* EmploymentDurationCurrentEmployer



Surge la pregunta de qué tipo de relación tendrá el tipo de vivienda con el rating que otorga la plataforma. Se observa que, para los ratings de la AA a la C (de muy bueno a medio-malo) el tipo de vivienda mayoritario es la vivienda en propiedad. Para los ratings del D al F la diferencia entre prestatarios con la vivienda en propiedad y los que la tienen en alquiler se reducen, llegando a superar la vivienda en alquiler a la vivienda en propiedad. Lo curioso es que para la calificación de alto riesgo, el tipo de vivienda mayoritario de los prestatarios es la vivienda de los padres, incluso por encima de la

vivienda en alquiler. Considerando que vivir en casa de los padres ahorra un montón de gastos…

Gráfico 44. Frecuencias de HomeOwnershipType *por* Rating



Se quiere ver cuál es la edad en función del tipo de vivienda. No sorprenden los resultados: los que tienen vivienda en propiedad o hipotecada son mayoritariamente de mediana edad en adelante; los que viven de alquiler o compartiendo piso son mayoritariamente jóvenes y, a partir de los 30 años, van descendiendo; el porcentaje de los que viven en vivienda social se mantiene más o menos estable en todas las edades.

Gráfico 45.1. Frecuencias de Age *por* HomeOwnershipType *para las categorías 1 (vivienda en propiedad), 2 (vivienda en casa de los padres) y 3 (vivienda alquilada amueblada)*

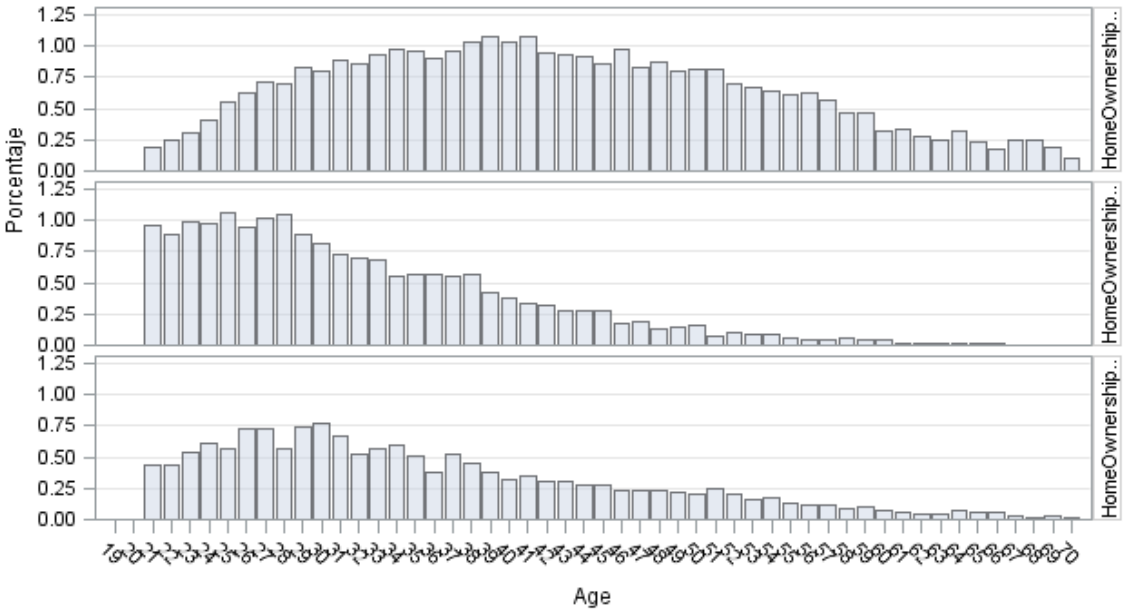


Gráfico 45.2. Frecuencias de Age *por* HomeOwnershipType *para las categorías 4 (vivienda alquilada sin amueblar), 5 (vivienda social), 6 (coarrendamiento) y 7 (copropiedad)*

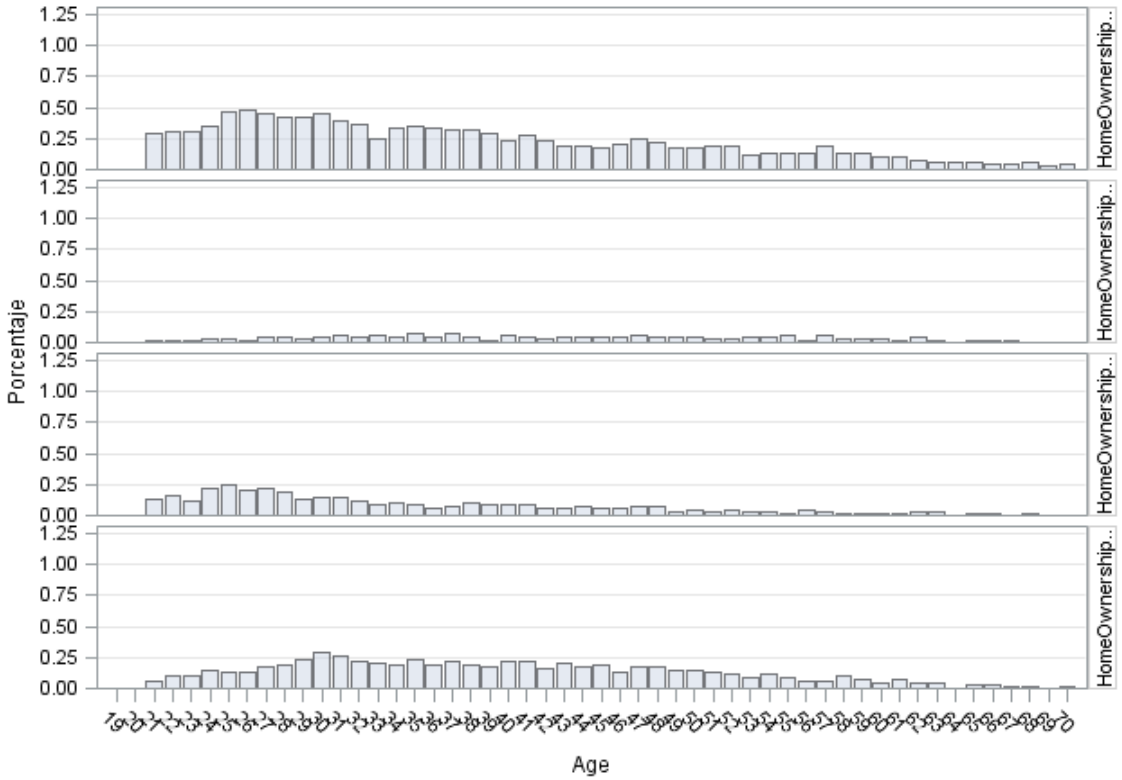
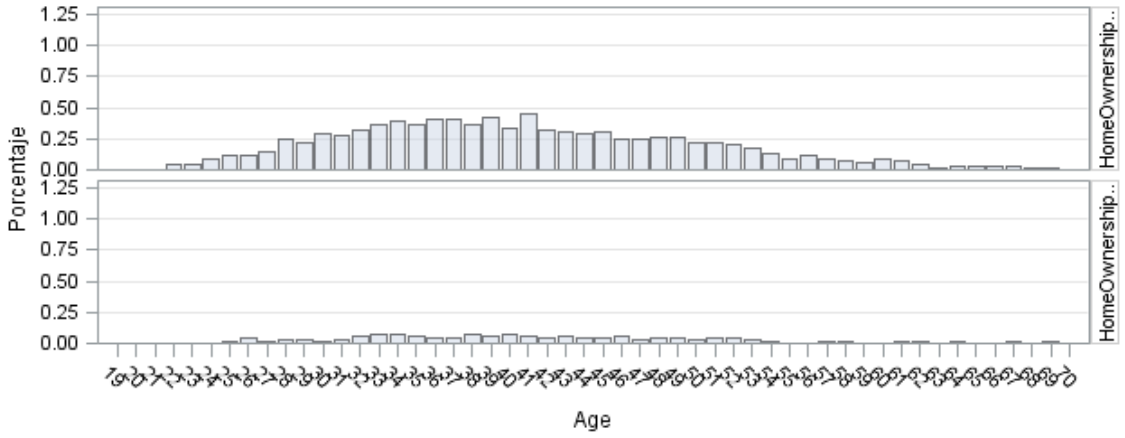
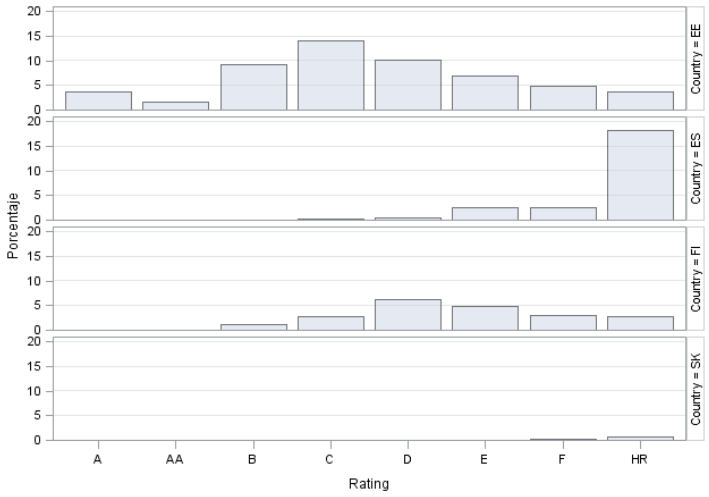


Gráfico 45.3. Frecuencias de Age *por* HomeOwnershipType *para las categorías 8 (vivienda con hipoteca) y 9 (propiedad con gravamen)*



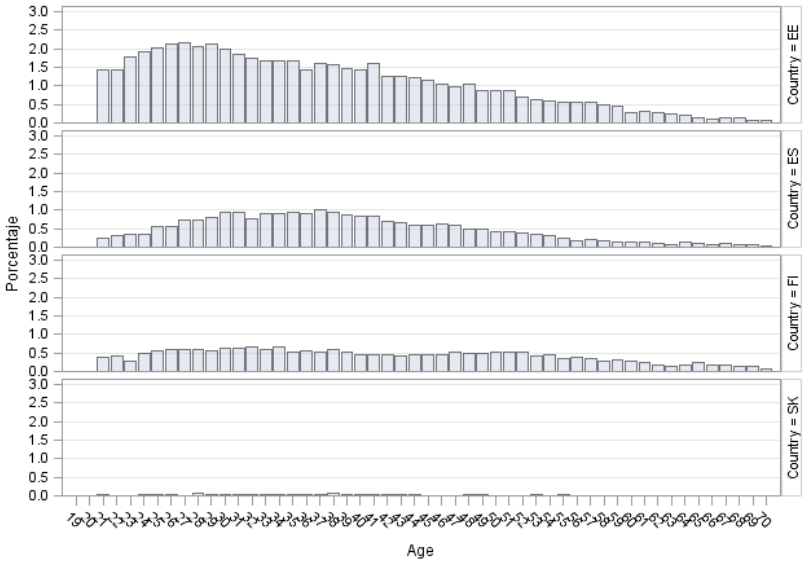
Si se analiza la relación entre el Rating y el país del solicitante se comprueba que, mientras que en Estonia y Finlandia la mayoría de los prestatarios tienen una calificación media (B, C, D), en España la mayoría están calificados como de alto riesgo. Si se sacan las distribuciones de rating por región se ve que siguen el mismo patrón que la distribución de rating por país del país de la región.

Gráfico 46. Frecuencias de Rating *por* Country



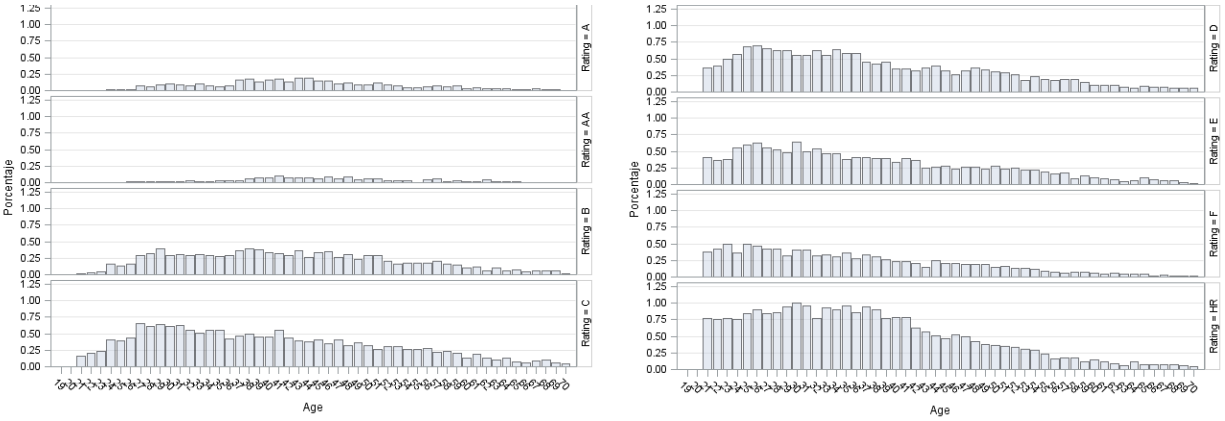
De nuevo se quiere ver la edad de los prestatarios, ahora en función del país. Se observa que Estonia aporta el mayor número de prestatarios jóvenes. En comparación, España aporta más prestatarios de mediana edad y Finlandia aporta más o menos el mismo porcentaje de prestatarios en todas las edades.

Gráfico 47. Frecuencias de Age *por* Country



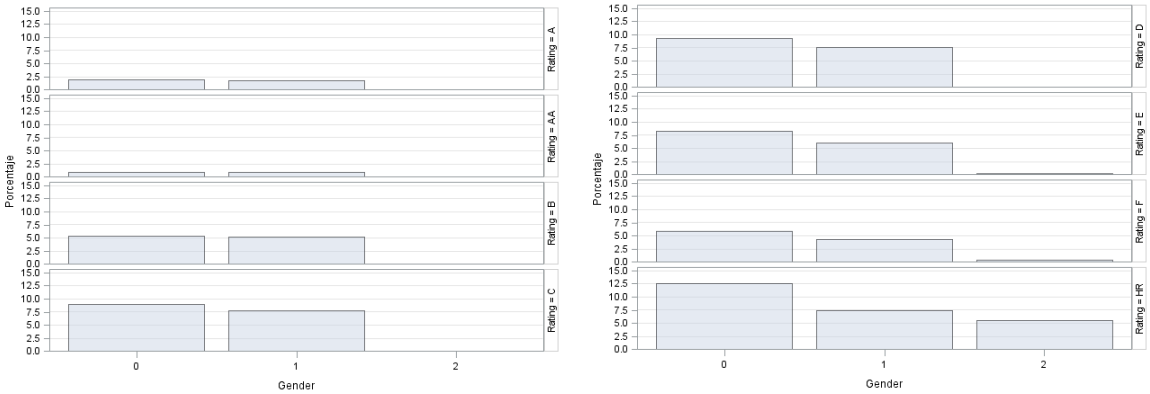
Si se compara la edad por rating se observa que, a medida que el rating va siendo peor, el eje de simetría de la distribución de la edad se va moviendo hacia la izquierda, lo que quiere decir que cuanto peor es el rating se encuentra en él a prestatarios de menor edad. En el caso del rating HR, los prestatarios jóvenes y de mediana edad son numerosos.

Gráfico 48. Frecuencias de Age *por* Rating



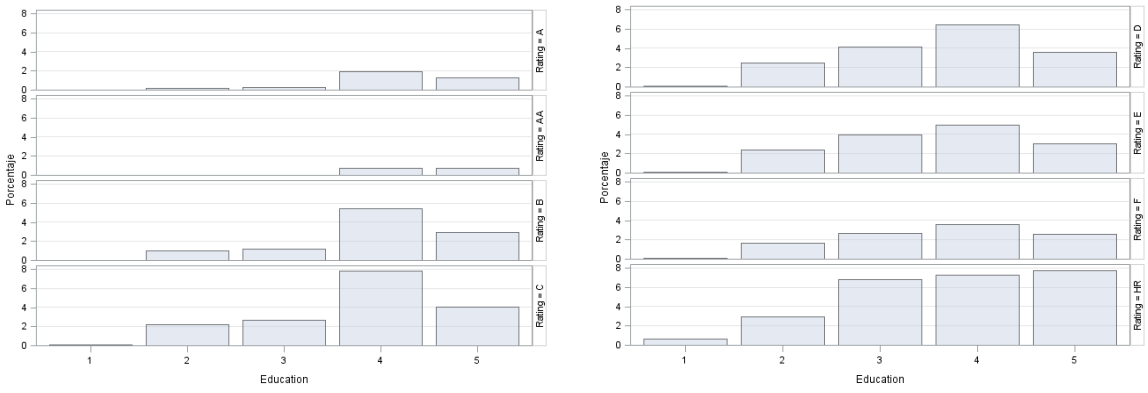
Si se analiza el rating por sexo se observa que, para ratings buenos, la proporción de hombres y mujeres más o menos es la misma pero, a medida que el rating va empeorando, la proporción de hombres aumenta sobre la de mujeres y aparecen los solicitantes de género indefinido. Para la calificación de alto riesgo, el porcentaje de prestatarias mujeres es casi la mitad que el de los hombres, y el porcentaje de las personas de género indefinido se aproxima mucho al de las mujeres.

Gráfico 49. Frecuencias de Gender *por* Rating



Al comparar el rating por el nivel educativo se observa que, para los ratings de la AA a la F los porcentajes de los distintos niveles educativos se mantienen, más o menos, como en el histograma de la muestra. En cambio, para el rating HR, el de alto riesgo, se observa que el grupo de solicitantes más numerosos es el de prestatarios con estudios superiores. Sorprende que el mayor número de prestatarios con alto riesgo sea el de aquellos con estudios universitarios, ya que, en general, por su formación deberían

Gráfico 50. Frecuencias de Education *por* Rating



tener ingresos más altos y mejores empleos aunque, por otro lado, se sabe que la crisis ha afectado mucho a los jóvenes con estudios universitarios.

Si se estudia la edad para los prestatarios con estudios superiores, se ve que no son el grupo con más jóvenes, pues entre los prestatarios con estudios de secundaria hay muchos más jóvenes. Tampoco lo explica la inestabilidad en el empleo, pues entre los prestatarios con estudios superiores la estabilidad es del mismo orden que entre los prestatarios con estudios secundarios.

Gráfico 51. Frecuencias de Age *por* Education *para las categorías 4 (estudios secundarios) y 5 (estudios superiores) de* Education

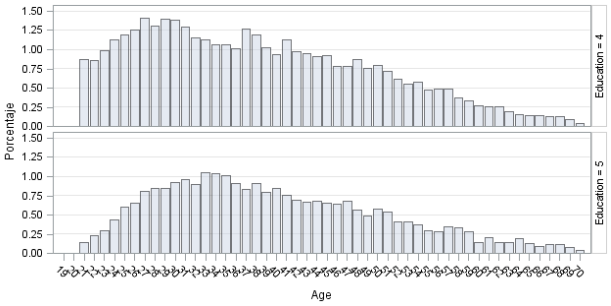
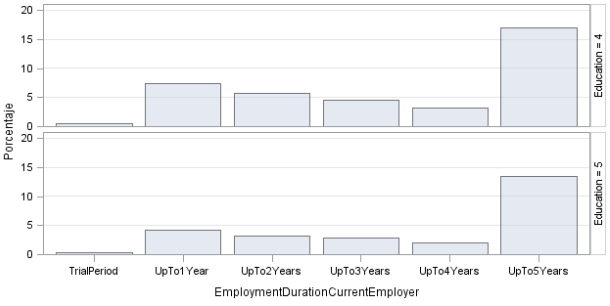


Gráfico 52. Frecuencias de EmploymentDurantionCurrentEmployer *por* Education *para las categorías 4 (estudios secundarios) y 5 (estudios superiores) de* Education



Lo que sí explica que entre los prestatarios de mayor riesgo la mayoría tengan estudios universitarios es la aportación de prestatarios que hace España. Ya se ha visto que la gran mayoría de los prestatarios españoles tienen riesgo alto. De los españoles, el 38% tiene estudios universitarios y, de ellos, el 71% tiene rating HR; es decir, que el 27% de los prestatarios españoles tiene estudios universitarios y son prestatarios de alto riesgo. La aportación de Estonia y Finlandia a los prestatarios de alto riesgo es baja en general y, todavía es más baja para los estudios superiores.

Gráfico 53. Frecuencias de Education *por* Rating de los prestatarios de Estonia *para las categorías malas de* Rating

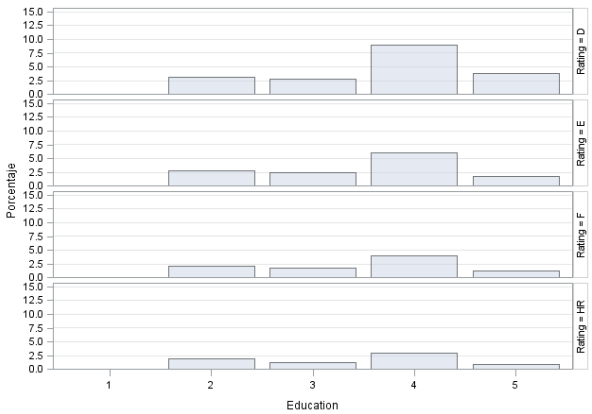


Gráfico 54. Frecuencias de Education *por* Rating de los prestatarios de Finlandia *para las categorías malas de* Rating

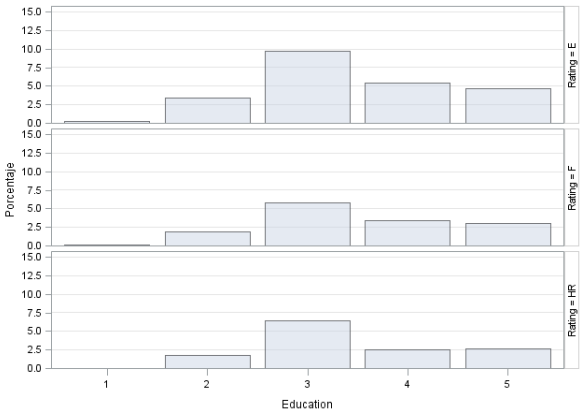
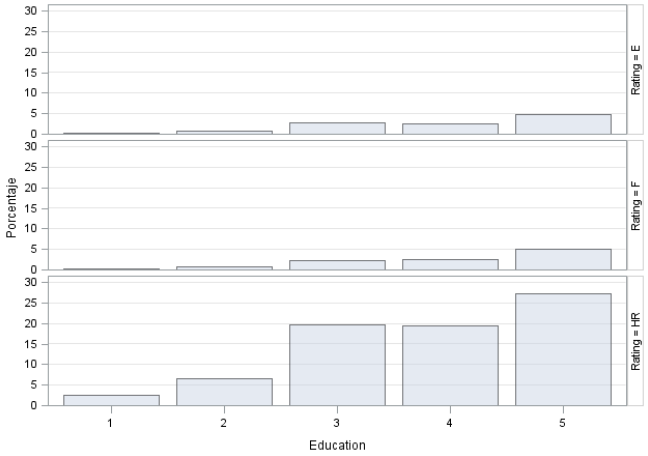


Gráfico 55. Frecuencias de Education *por* Rating de los prestatarios de España *para las categorías malas de* Rating



5.Modificación de variables

5.1 Creación de nuevas variables

Se resume en este apartado las nuevas variables que se crean y cuya justificación se ha explicado en el punto 3 de este trabajo.

En primer lugar, se ha añadido a la muestra la variable objetivo ***Impago***, variable binaria que tomará valor 0 si el préstamo no tiene actualmente impagos de principal o intereses y que tomará valor 1 si el préstamo tiene actualmente algún impago de principal, interés o ambos.

Se crea también la nueva variable ***County2*** que armoniza las regiones, reduciendo sus categorías.

Se crean las variables ***PorcRepaymentsPreviousLoans*** y ***PorcEarlyRepaymentsPreviousLoans*** como una mejor forma de recoger la información de los préstamos previos del prestatario.

Se crea la variable ***PorFreeCash*** como cociente de las siguientes variables para poder eliminar ambas de la muestra y evitar posibles problemas de multicolinealidad debido a la correlación que mantienen:

$$PorFreeCash = \frac{IncomeTotal - LiabilitiesTotal}{IncomeTotal}$$

5.2 Transformación de variables

En el análisis univariante se han visto los histogramas de las variables continuas y, como se ha comentado, algunas de ellas son candidatas a la transformación para intentar suavizar su distribución y que se asemeje más a la distribución normal.

Para ello se han creado nuevas variables que, posteriormente incluiremos en la selección de variables para ver si entran en el modelo. Las variables que se han creado como transformación de variables continuas son:

$$\boldsymbol{TransAge} = \sqrt[2]{Age}$$

$$\boldsymbol{TransIncomeTotal} = \log IncomeTotal$$

$$\boldsymbol{TransLiabilitiesTotal} = \log LiabilitiesTotal$$

$$\boldsymbol{TransInterest} = \log Interest$$

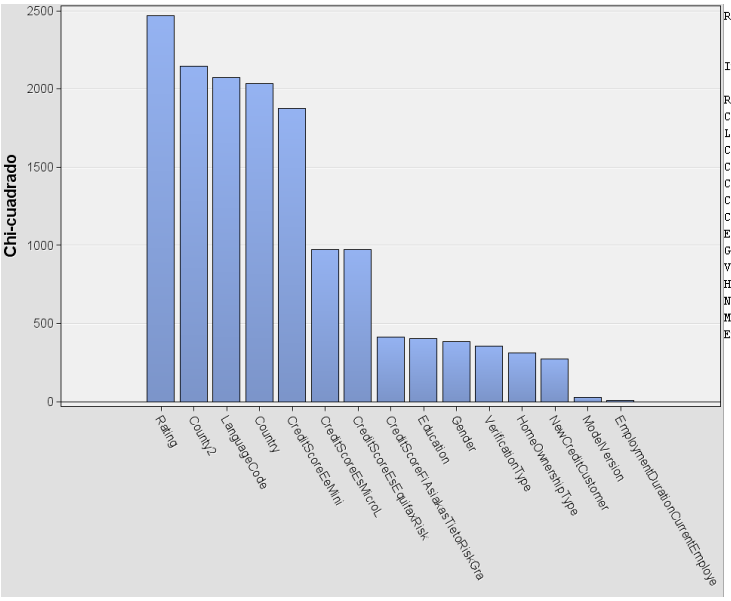
$$\boldsymbol{TransExistingLiabilites} = \log ExistingLiabilites$$

$$\boldsymbol{TransNoOfPreviousLoansBeforeLoan} = \log NoOfPreviousLoansBeforeLoan$$

5.3 Relación variables independientes con la variable objetivo

Se utiliza el nodo de *Explorador de estadísticos* de SAS Miner para ver la relación de las variables continuas y categóricas con la variable objetivo. Para las variables categóricas, el test de la Chi-Cuadrado indica que las variables categóricas más relevantes son ***Rating***, ***County2***, ***LanguageCode***, ***Country*** y ***CreditScoreEeMini***.

Gráfico 56. Variables categóricas seleccionadas por el test de la Chi-cuadrado

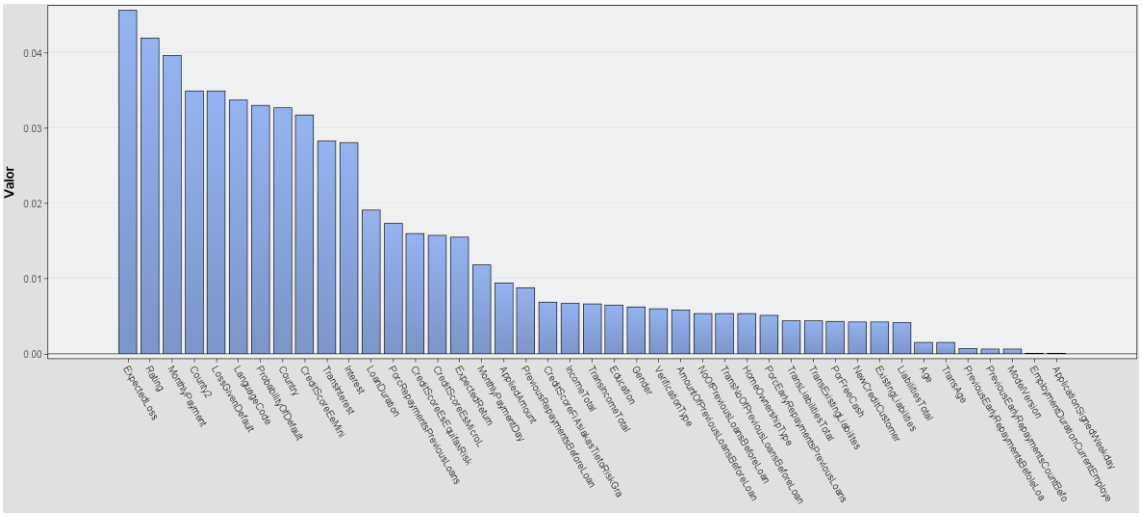


Rol de los datos=TRAIN Objetivo=Impago

Input	Chi-cuadrado	Df	Prob
Rating	2469.1175	8	<.0001
County2	2146.4665	67	<.0001
LanguageCode	2073.9163	12	<.0001
Country	2035.0779	3	<.0001
CreditScoreEeMini	1873.0083	6	<.0001
CreditScoreEsMicroL	974.8486	10	<.0001
CreditScoreEsEquifaxRisk	972.7378	6	<.0001
CreditScoreFiAsiakasTietoRiskGra	414.7891	6	<.0001
Education	402.4522	4	<.0001
Gender	383.5082	2	<.0001
VerificationType	354.3060	3	<.0001
HomeOwnershipType	313.2605	9	<.0001
NewCreditCustomer	274.9637	1	<.0001
ModelVersion	28.8248	6	<.0001
EmploymentDurationCurrentEmploye	8.8492	6	0.1822

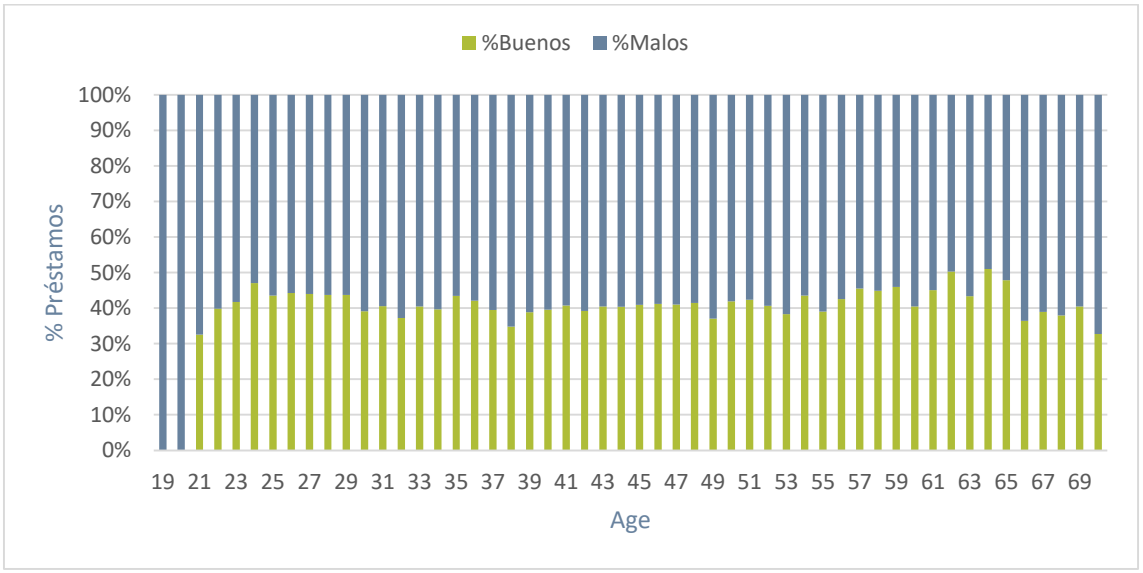
Si se analiza la importancia de las variables por el Valor de la variable, se observa que de las variables continuas las que tendrían mayor relación con la variable objetivo son ***ExpectedReturn***, ***MonthlyPayment***, ***LossGivenDefault*** y ***ProbabilityOfDefault***.

Gráfico 57. Valor de las variables de la muestra



Se calculan las tablas de frecuencias con respecto a la variable ***Impago*** y se calculan las frecuencias en porcentaje por categoría. Se observa que el impago no está relacionado con la edad, pues excepto para 19 y 20 años, el porcentaje de préstamos impagados es, más o menos, el mismo para todas las edades.

Gráfico 58. Frecuencia de préstamos buenos y malos por Edad



Sí se observa, en cambio, que hay más préstamos impagados entre los clientes nuevos, casi un 10% más que en los antiguos clientes. También se observa que verificar los ingresos de los solicitantes no es suficiente para garantizar que el prestatario vaya a pagar pues, aunque entre aquellos solicitantes cuyos gastos e ingresos se verifican, los que impagan son el 50%. Entre los que solamente se verifican los ingresos o no se verifican, impagan más del 60%.

Gráfico 59. Frecuencia de préstamos buenos y malos por NewCreditCustomer

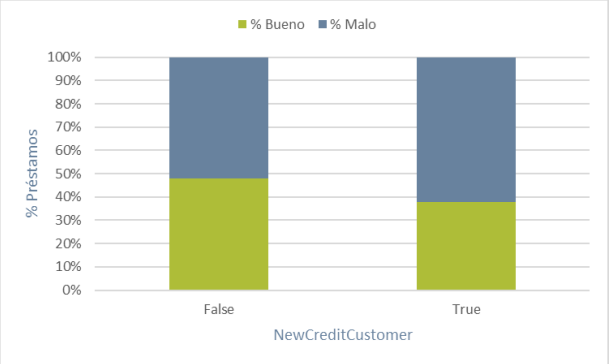
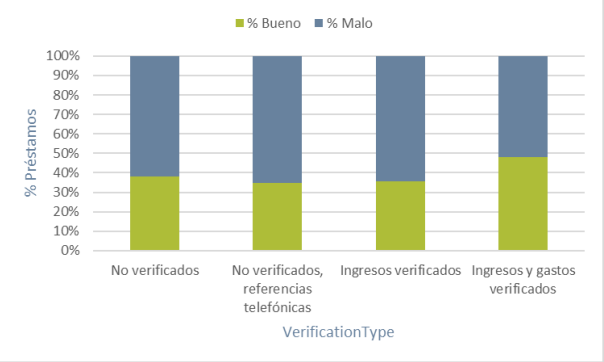


Gráfico 60. Frecuencia de préstamos buenos y malos por VerificationType



Por género, las mujeres son un poco mejores pagadoras que los hombres, pero los que tienen una tasa altísima de impago son las personas de género indeterminado. Si se pone el foco en la educación de los prestatarios, los que cometen menos impagos son aquellos con estudios de secundaria y con estudios superiores.

Gráfico 61. Frecuencia de préstamos buenos y malos por Gender

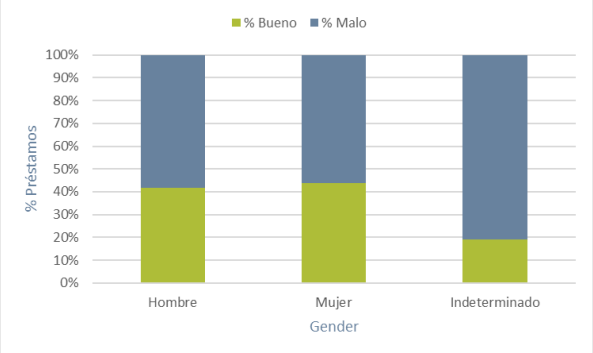
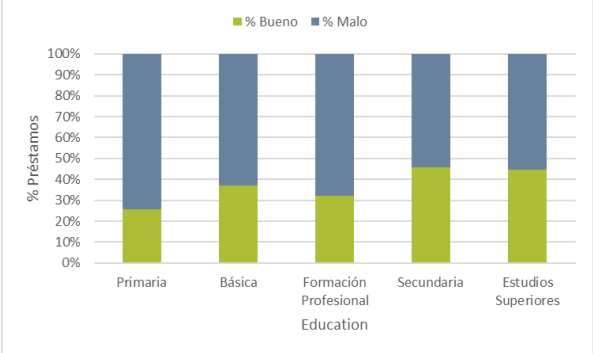


Gráfico 62. Frecuencia de préstamos buenos y malos por Education



La estabilidad en el empleo parece que no está relacionada con el impago salvo para los que no informan este campo, que sí tienen una tasa de impago significativamente mayor a la del resto de prestatarios. La vivienda, en cambio, parece que sí es indicativa del impago, pues aquellos que no tienen la vivienda en propiedad tienen tasas más altas de impago que los que son propietarios o tienen una hipoteca sobre su vivienda.

Gráfico 63. Frecuencia de préstamos buenos y malos por EmploymentDurationCurrentEmploye

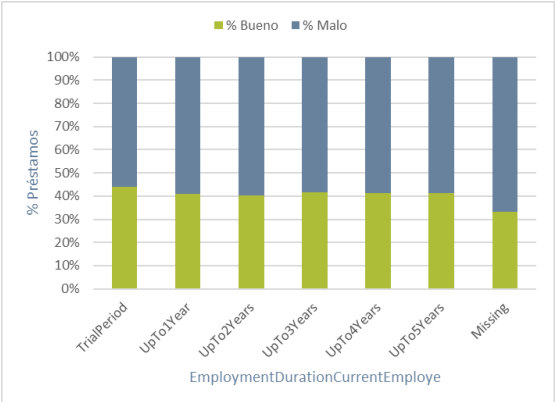
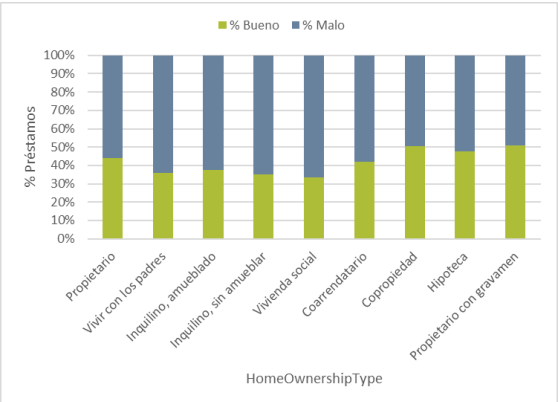


Gráfico 64. Frecuencia de préstamos buenos y malos por HomeOwnershipType



Por último, el origen y el idioma del prestatario sí indican qué solicitantes harán más impagos: los solicitantes de Estonia son los que tienen menor tasa de impago, mientras que los de España tienen la tasa de impago más alta.

Gráfico 65. Frecuencia de préstamos buenos y malos por Country

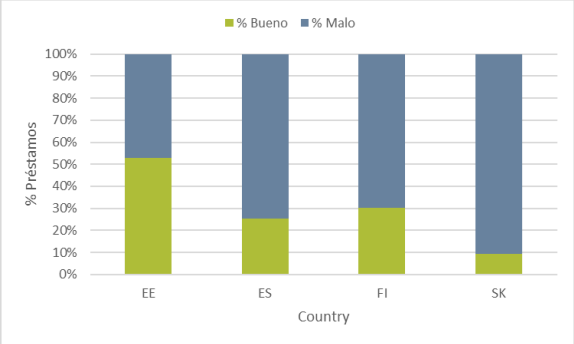
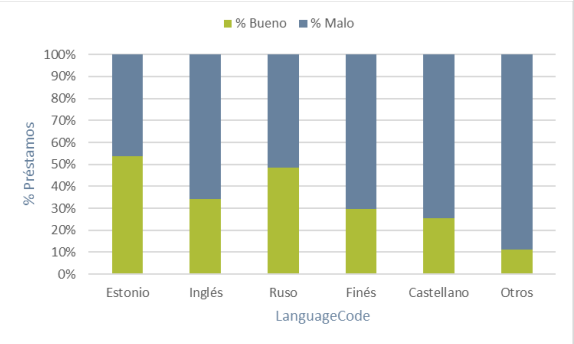


Gráfico 66. Frecuencia de préstamos buenos y malos por LanguageCode



Las variables continuas están poco correlacionas con la variable ***Impago***. De ellas, las que tienen mayor correlación positiva son la duración del préstamo, la pérdida esperada, la probabilidad de impago y el tipo de interés. La variable que tiene mayor correlación negativa con ***Impago*** es el porcentaje de amortizaciones de los préstamos anteriores. Todas estas correlaciones son lógicas ya que, a mayor pérdida esperada, mayor riesgo de impago y, por tanto, el préstamo es más caro. A la vez, cuanta menos deuda queda pendiente de los préstamos anteriores significa que ha ido pagando sus préstamos, lo que podría indicar un buen comportamiento como prestatario. Y además tiene más dinero disponible para pagar nuevas cuotas de préstamos.

Tabla 13. Variables continuas más correlacionadas con Impago

Coeficientes de correlación Pearson, N = 9678 Prob > |r| suponiendo H0: Rho=0						
	Impago	PorcRepaymentsPreviousLoans	TransInterest	LoanDuration	ExpectedLoss	ProbabilityOfDefault
Impago	1.00000	-0.23892	0.17973	0.20576	0.19444	0.18626
		<.0001	<.0001	<.0001	<.0001	<.0001

5.4 Agrupación de variables

5.4.1 Variables continuas

Se va a tramificar las variables continuas para poder ver mejor las relaciones entre las variables y la variable objetivo y obtener mayor conocimiento del comportamiento de los solicitantes.

Se utiliza SAS Miner para realizar la agrupación de una forma sencilla. Este nodo de SAS Miner permite hacer una selección de las variables continuas así como una tramificación de las mismas. Cada variable se separa en atributos a los que se le asignan puntos basándose en un análisis estadístico que tiene en cuenta, entre otros, el poder predictivo de las características de cada grupo así como la correlación entre ellas.

Los estadísticos que facilita SAS Miner para facilitar la decisión del agrupamiento son tres:

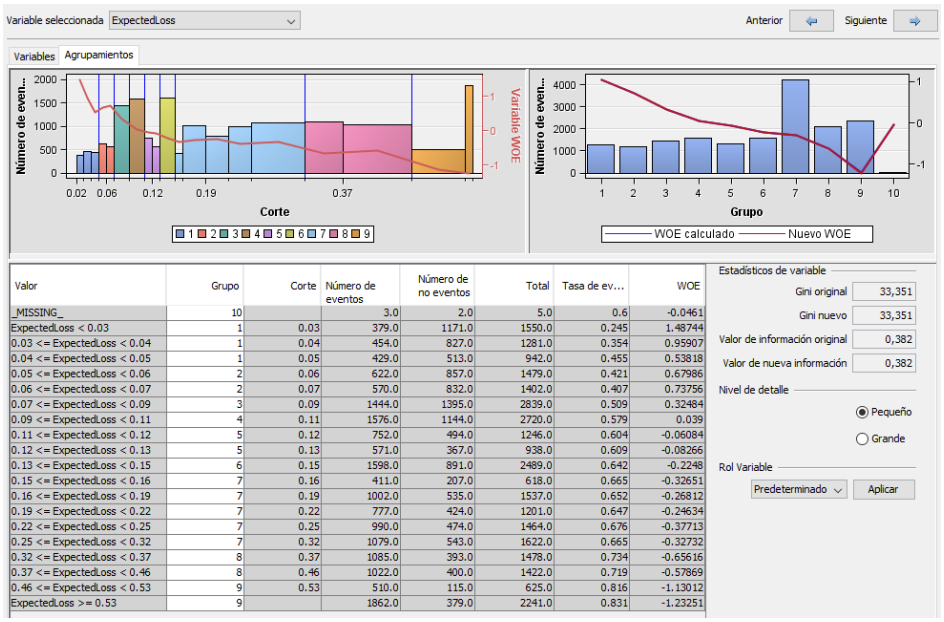
- **Weight Of Evidence (WOE)**. Es una medida de la diferencia entre la proporción buenos y malos en cada atributo. Mide la fuerza de cada grupo para separar

buenos y malos. Cuanto mayor sea la diferencia en el WOE de grupos contiguos, mayor es la capacidad predictiva de ese atributo.

- **Valor de la información (IV)**. Estadístico que ayuda a la determinación del número de tramos y que también se utiliza para estimar si una variable discrimina entre buenos y malos. Cuanto mayor sea el IV mejor, teniendo en cuenta estos parámetros:
    - IV ≤ 0.02 → Variable no predictiva
    - 0.02 ≤ IV ≤ 0.1 → Variable con poder predictivo débil
    - 0.1 ≤ IV ≤ 0.3 → Variable con predictividad media
    - 0.3 ≤ IV ≤ 0.5 → Variable con poder predictivo fuerte
    - 0.5 ≤ IV → Demasiado bueno para ser cierto. Variable sospechosa
- **Estadístico de Gini.** Estadístico que utilizado a nivel variable sirve para verificar el poder discriminante de la variable. Toma valores entre 0 y 100%. En general, se siguen los siguientes parámetros:
    - Gini ≤ 5% → Se descarta la variable en el modelo multivariante
    - 5% ≤ Gini ≤ 15% → Poder de predicción bajo. SAS Miner las descarta del modelo por defecto
    - 15% ≤ Gini → Poder predictivo alto

La agrupación que propone SAS basándose en el estadístico de Gini y el valor de la información para la variable ***ExpectedLoss*** con 10 grupos que es el máximo que se ha parametrizado en el nodo es:

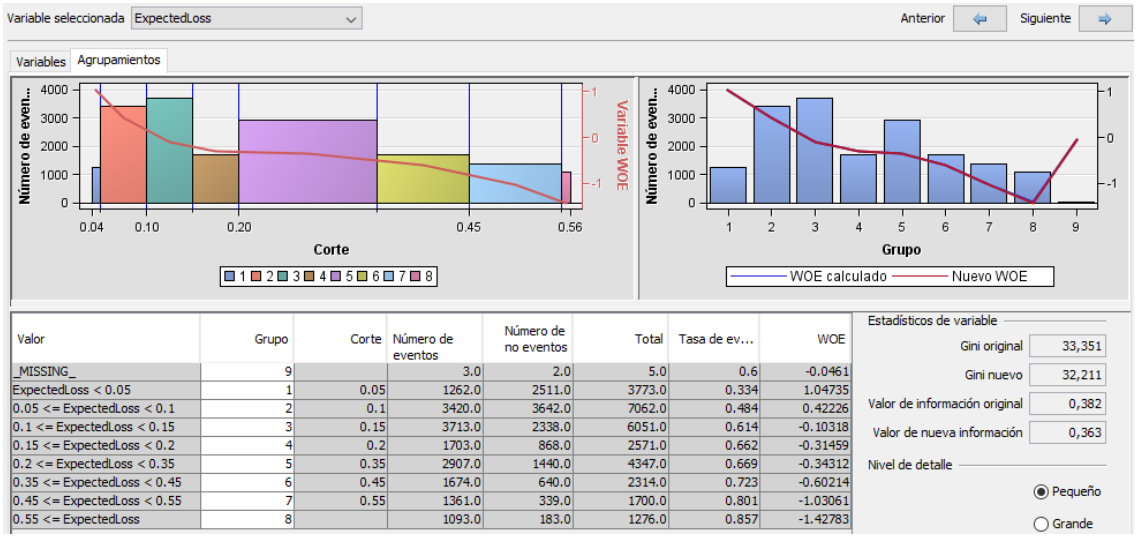
Figura 1. Agrupación de ExpectedLoss *propuesta por SAS Miner*



Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	10		3.0	2.0	5.0	0.6	-0.0461
ExpectedLoss < 0.03	1	0.03	379.0	1171.0	1550.0	0.245	1.48744
0.03 <= ExpectedLoss < 0.04	1	0.04	454.0	827.0	1281.0	0.354	0.95907
0.04 <= ExpectedLoss < 0.05	1	0.05	429.0	513.0	942.0	0.455	0.53818
0.05 <= ExpectedLoss < 0.06	2	0.06	622.0	857.0	1479.0	0.421	0.67986
0.06 <= ExpectedLoss < 0.07	2	0.07	570.0	832.0	1402.0	0.407	0.73756
0.07 <= ExpectedLoss < 0.09	3	0.09	1444.0	1395.0	2839.0	0.509	0.32484
0.09 <= ExpectedLoss < 0.11	4	0.11	1576.0	1144.0	2720.0	0.579	0.039
0.11 <= ExpectedLoss < 0.12	5	0.12	752.0	494.0	1246.0	0.604	-0.06084
0.12 <= ExpectedLoss < 0.13	5	0.13	571.0	367.0	938.0	0.609	-0.08266
0.13 <= ExpectedLoss < 0.15	6	0.15	1598.0	891.0	2489.0	0.642	-0.2248
0.15 <= ExpectedLoss < 0.16	7	0.16	411.0	207.0	618.0	0.665	-0.32651
0.16 <= ExpectedLoss < 0.19	7	0.19	1002.0	535.0	1537.0	0.652	-0.26812
0.19 <= ExpectedLoss < 0.22	7	0.22	777.0	424.0	1201.0	0.647	-0.24634
0.22 <= ExpectedLoss < 0.25	7	0.25	990.0	474.0	1464.0	0.676	-0.37713
0.25 <= ExpectedLoss < 0.32	7	0.32	1079.0	543.0	1622.0	0.665	-0.32732
0.32 <= ExpectedLoss < 0.37	8	0.37	1085.0	393.0	1478.0	0.734	-0.65616
0.37 <= ExpectedLoss < 0.46	8	0.46	1022.0	400.0	1422.0	0.719	-0.57869
0.46 <= ExpectedLoss < 0.53	9	0.53	510.0	115.0	625.0	0.816	-1.13012
ExpectedLoss >= 0.53	9		1862.0	379.0	2241.0	0.831	-1.23251

Se reagrupan los segmentos para que no haya tantos tramos y la pérdida esperada sea más fácil de interpretar en el modelo de scoring, intentando que el índice de Gini y el Valor de la información no emporen demasiado. De esta manera se reagrupan los segmentos como sigue:

Figura 2. Agrupación final de ExpectedLoss



Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	9		3.0	2.0	5.0	0.6	-0.0461
ExpectedLoss < 0.05	1	0.05	1262.0	2511.0	3773.0	0.334	1.04735
0.05 <= ExpectedLoss < 0.1	2	0.1	3420.0	3642.0	7062.0	0.484	0.42226
0.1 <= ExpectedLoss < 0.15	3	0.15	3713.0	2338.0	6051.0	0.614	-0.10318
0.15 <= ExpectedLoss < 0.2	4	0.2	1703.0	868.0	2571.0	0.662	-0.31459
0.2 <= ExpectedLoss < 0.35	5	0.35	2907.0	1440.0	4347.0	0.669	-0.34312
0.35 <= ExpectedLoss < 0.45	6	0.45	1674.0	640.0	2314.0	0.723	-0.60214
0.45 <= ExpectedLoss < 0.55	7	0.55	1361.0	339.0	1700.0	0.801	-1.03061
0.55 <= ExpectedLoss	8		1093.0	183.0	1276.0	0.857	-1.42783

Estadísticos de variable

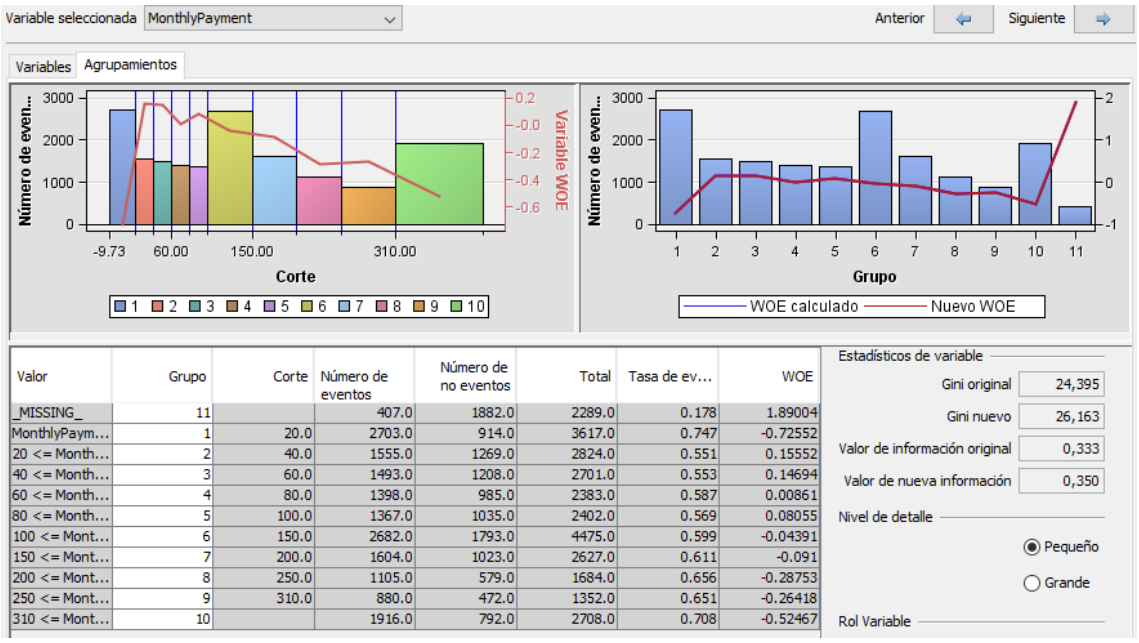
Gini original	33,351
Gini nuevo	32,211
Valor de información original	0,382
Valor de nueva información	0,363

Nivel de detalle: Pequeño / Grande

En este caso, el agrupamiento final hace bajar tanto el estadístico de Gini (Gini nuevo) como el Valor de la información.

Del mismo modo se tramifica la variable ***MonthlyPayment*** en los 11 grupos que se detallan, consiguiendo mejorar tanto el índice de Gini como el VI sobre la agrupación inicial que propone SAS:

Figura 3. Agrupación final MonthlyPayment



Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	11		407.0	1882.0	2289.0	0.178	1.89004
MonthlyPaym...	1	20.0	2703.0	914.0	3617.0	0.747	-0.72552
20 <= Month...	2	40.0	1555.0	1269.0	2824.0	0.551	0.15552
40 <= Month...	3	60.0	1493.0	1208.0	2701.0	0.553	0.14694
60 <= Month...	4	80.0	1398.0	985.0	2383.0	0.587	0.00861
80 <= Month...	5	100.0	1367.0	1035.0	2402.0	0.569	0.08055
100 <= Mont...	6	150.0	2682.0	1793.0	4475.0	0.599	-0.04391
150 <= Mont...	7	200.0	1604.0	1023.0	2627.0	0.611	-0.091
200 <= Mont...	8	250.0	1105.0	579.0	1684.0	0.656	-0.28753
250 <= Mont...	9	310.0	880.0	472.0	1352.0	0.651	-0.26418
310 <= Mont...	10		1916.0	792.0	2708.0	0.708	-0.52467

Estadísticos de variable

Gini original	24,395
Gini nuevo	26,163
Valor de información original	0,333
Valor de nueva información	0,350

Nivel de detalle: Pequeño / Grande

Rol Variable

Aquí la nueva agrupación hace aumentar tanto el índice de Gini como el Valor de la Información.

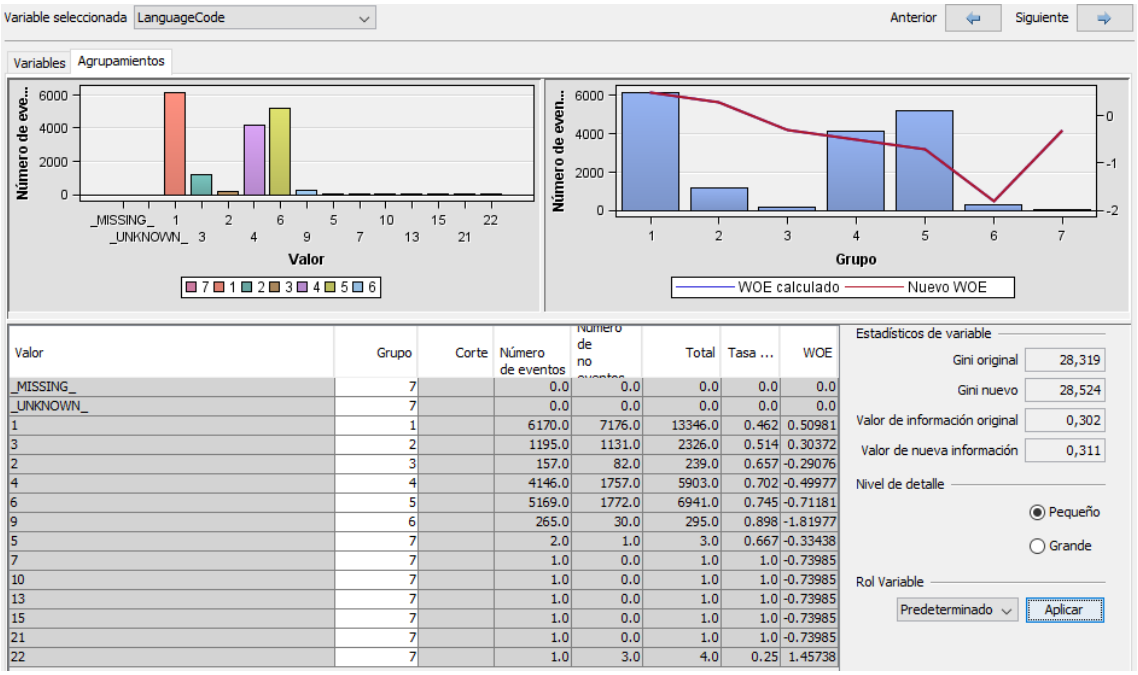
En todas las agrupaciones se ha considerado a los valores missing como una nueva categoría. Se ha hecho así porque el valor missing es un atributo diferenciador de los prestatarios. Por ejemplo, en el caso de las variables ***AmountOfPreviousLoans*** o ***PorRepaymentsPreviousLoans*** el valor missing indica que el prestatario no ha tenido préstamos anteriormente.

Se han tramificado las variables continuas ***ExpectedLoss***, ***MonthlyPayment***, ***ProbabilityOfDefault***, ***Interest***, ***AppliedAmount***, ***AmountOfPreviousLoans***, ***PorFreeCash***, ***ExpectedReturn,*** ***PorcRepaymentsPreviousLoans*** y ***PorcEarlyRepaymentsPreviousLoans***. El detalle de cómo ha quedado la categorización se puede consultar en el Anexo IV.

5.4.2 Variables categóricas

En el caso de las variables categóricas, únicamente se ha reagrupado la variable ***LanguageCode***, porque había muchas categorías con muy pocas observaciones. La nueva categorización ha consistido en dejar los cinco idiomas más utilizados (Estonio, Finés, Castellano, Ruso e Inglés) y agrupar los demás bajo la categoría Otros.

Figura 11. Agrupación final LanguageCode



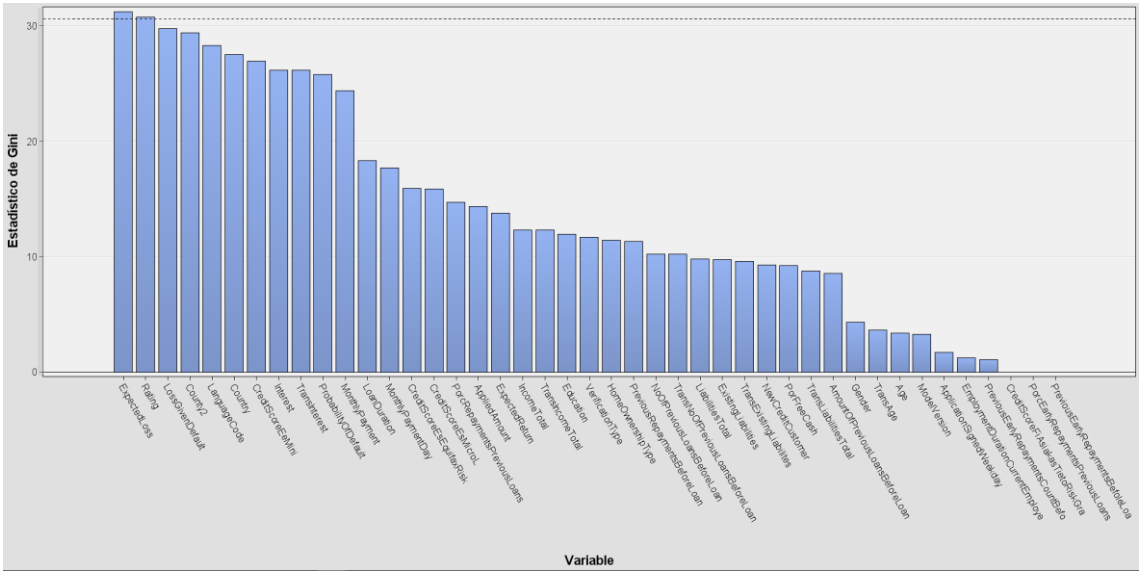
Valor	Grupo	Corte	Número de eventos	Número de no	Total	Tasa ...	WOE
MISSING	7		0.0	0.0	0.0	0.0	0.0
UNKNOWN	7		0.0	0.0	0.0	0.0	0.0
1	1		6170.0	7176.0	13346.0	0.462	0.50981
3	2		1195.0	1131.0	2326.0	0.514	0.30372
2	3		157.0	82.0	239.0	0.657	-0.29076
4	4		4146.0	1757.0	5903.0	0.702	-0.49977
6	5		5169.0	1772.0	6941.0	0.745	-0.71181
9	6		265.0	30.0	295.0	0.898	-1.81977
5	7		2.0	1.0	3.0	0.667	-0.33438
7	7		1.0	0.0	1.0	1.0	-0.73985
10	7		1.0	0.0	1.0	1.0	-0.73985
13	7		1.0	0.0	1.0	1.0	-0.73985
15	7		1.0	0.0	1.0	1.0	-0.73985
21	7		1.0	0.0	1.0	1.0	-0.73985
22	7		1.0	3.0	4.0	0.25	1.45738

La reagrupación no aporta una gran mejora como variable discriminante, pero puede hacer que funcionen mejor los algoritmos al simplificarla. Se ha valorado la opción de reagrupar la variable HomeOwnershipType porque hay categorías muy similares como, por ejemplo, *Vivienda alquilada amueblada*, *Vivienda alquilada sin amueblar* y *Coarrendamiento*, pero tras probar y ver que el reagrupamiento no mejoraba el poder discriminante de la variable, se ha decidido dejar la categorización original.

5.5 Selección de variables

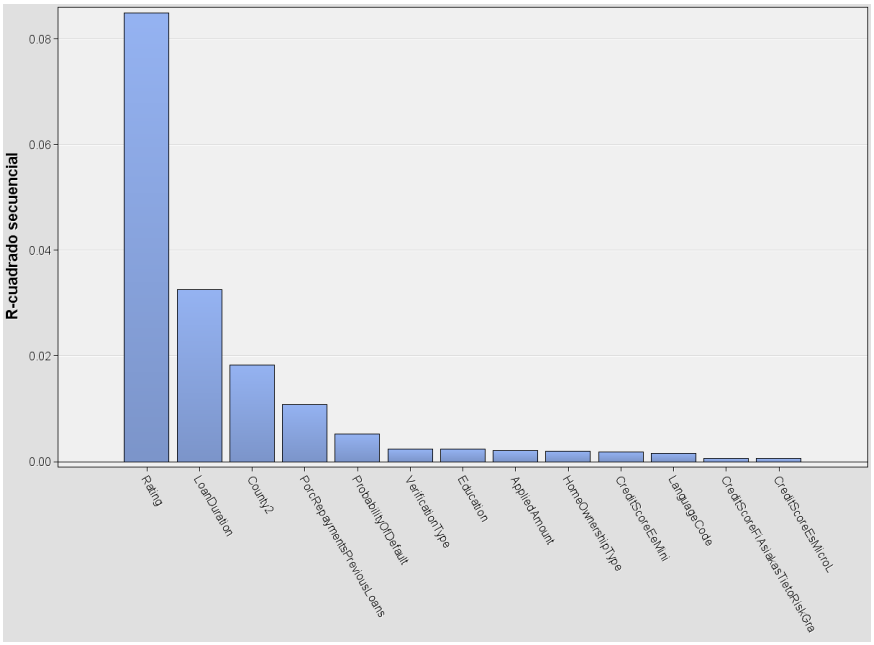
En una primera selección de variables, basada en el estadístico de Gini, se ve que la única variable transformada que entraría en el modelo sería ***TransInterest*** y no mejora a la variable original. Por lo tanto, se descartan las variables transformadas de la muestra.

Gráfico 67. Estadístico de Gini para la selección de variables



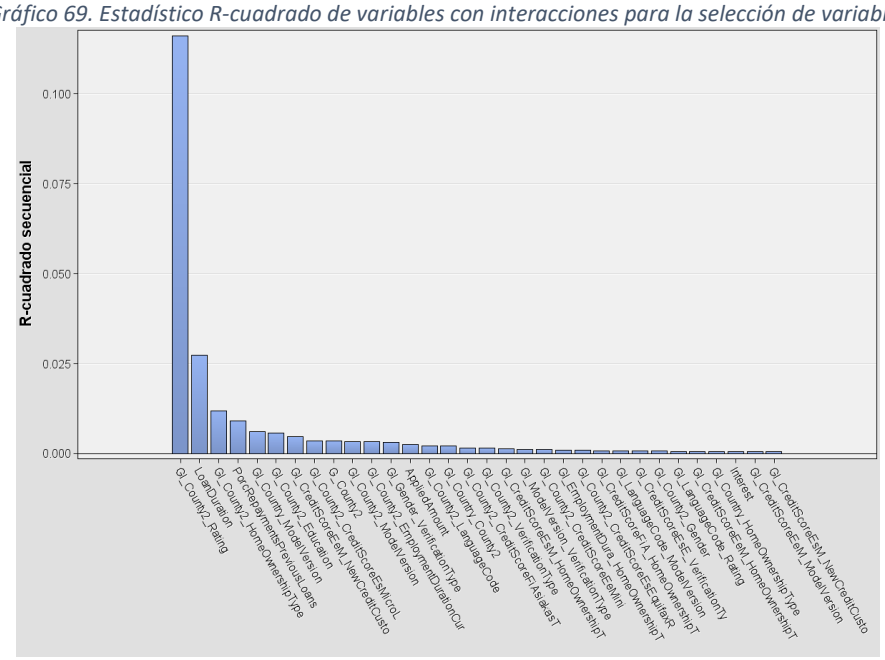
Si se tiene en cuenta el R-cuadrado para ver los efectos en el modelo, se observa que la variable más importante es ***Rating***, seguida de ***LoanDuration*** y de ***County2***. Parece pues importante hacer un correcto aprovisionamiento de la región del prestatario.

Gráfico 68. Estadístico R-cuadrado para la selección de variables



Si se introducen interacciones de las variables, los efectos cambian. El estadístico R-cuadrado clarifica los efectos: siguen siendo las mismas variables las que tienen mayor impacto en el modelo pero ahora el ***Rating*** se ve matizado por la región del prestatario y la región, ***County2***, se ve matizada por el tipo de residencia.

Gráfico 69. Estadístico R-cuadrado de variables con interacciones para la selección de variables



6. Modelización

Para la modelización se va a utilizar la regresión logística y los algoritmos Redes Neuronales, Bagging, Random Forest y Gradient Boosting utilizando el método de la validación cruzada.

Como el objetivo del trabajo es comprobar si las variables calculadas que aporta la plataforma Bondora (***Rating***, ***LossGivenDefault***, ***ExpectedReturn***, ***ProbabilityOfDefault***) son explicativas en el modelo para la clasificación de impagos bajo la definición de préstamo malo expuesta, se crean cuatro sets de datos de la muestra que se utilizarán en la modelización, dos que incluirán estas variables calculadas por la plataforma y dos que no las incluirán:

- TODASORIG – Set que incluye las variables de calificación de Bondora, incluyendo las de burós externos, sin la tramificación descrita en el apartado 5 del trabajo. Los modelos de este set de datos se identifican por terminar en TS
- TODASTRAM – Set que incluye las variables de calificación de la plataforma, con las variables tramificadas según se ha descrito en el apartado 5 del trabajo. Los modelos de este set de datos se identifican por terminar en TT
- CONOCIDASORIG – Set que no incluye las variables de calificación de Bondora, es decir, solo incluye las variables que proporciona el solicitante, sin tramificar. Los modelos de este set de datos se identifican por terminar en CS
- CONOCIDASTRAM – Set que no incluye las variables de calificación de la plataforma, tramificadas. Los modelos de este set de datos se identifican por terminar en CT.

De esta manera se podrá comprobar si el mejor modelo incluye o no estas variables. Si las incluye, podremos concluir que son explicativas para la definición de préstamo malo dada en el apartado 3.2.

Se recuerda que la muestra contiene la información de 29.062 préstamos; los préstamos aprobados por Bondora entre el 1 de enero de 2014 y el 31 de marzo de 2017. De ellos el 58.87% han realizado algún impago de principal o intereses y sólo el 41.13% no ha realizado ningún impago.

6.1 Regresión logística

Como ya se ha comentado anteriormente, en el caso de la regresión logística hay que hacer un procedimiento de descubrimiento iterativo para determinar cuáles son las variables más explicativas del modelo. Para ello se van a modelizar varias regresiones logísticas por varios métodos.

En primer lugar, se realiza una regresión logística paso a paso, sin interacciones en las variables, para los cuatro sets de datos con el objetivo de tener una primera idea de qué variables entrarían en el modelo. Cabe comentar que es necesario eliminar de las variables independientes las *CreditScore* que facilita la plataforma, ya que tienen un número elevadísimo de missing y la regresión logística no se ejecuta. El resultado de las variables que entran en los modelos son:

Tabla 13. Variables seleccionadas por el procedimiento PROCLOGISTIC

Modelo	Variables continuas	Variables categóricas
REG03TT	Age LoanDuration MonthlyPaymentDay	VerificationType Country Education County2 TMonthlyPayment TInterest TPorcRepaymentsPreviousLoans TAppliedAmount TAmountOfPreviousLoansBeforeLoan TPorFreeCash TProbabilityOfDefault
REG03CT	ExistingLiabilities LoanDuration MonthlyPaymentDay	County2 Education HomeOwnershipType TAmountOfPreviousLoansBeforeLoan TAppliedAmount TInterest TLanguageCode TMonthlyPayment TPorFreeCash TPorcRepaymentsPreviousLoans VerificationType
REG03TS	AppliedAmount ExistingLiabilities ExpectedReturn LoanDuration LossGivenDefault MonthlyPaymentDay PorcRepaymentsPreviousLoans ProbabilityOfDefault	Country Education Rating
REG03CS	AppliedAmount Interest LoanDuration MonthlyPayment MonthlyPaymentDay PorFreeCash PorcRepaymentsPreviousLoans	County2 Education LanguageCode NewCreditCustomer VerificationType

Se observa que la variable ***County2*** entra en los modelos, luego su correcto aprovisionamiento parece necesario. En cambio, la variable ***Rating***, variable de calificación de la plataforma no entra en los dos modelos para todas las variables, únicamente entra en aquél en el que no se tramifican las variables.

Se crean interacciones entre las variables con ayuda de la macro *%interactodo* (autor, Javier Portela) que también las ordena en función del estadístico F. El resultado son 600 iteraciones.

Se incluyen las iteraciones en la regresión logística *stepwise* para los modelos, pero el modelo es tan grande que el ordenador se queda sin memoria y no puede procesar la regresión logística. Dado que SAS Miner también permite hacer interacciones, se ejecuta la regresión logística en este software. De nuevo, para los sets de datos con variables tramificadas, el software no es capaz de ejecutar la regresión por falta de memoria. Sí consigue ejecutarse para los sets de datos sin tramificar. Éstos son los dos modelos con interacciones:

Tabla 14. Variables seleccionadas por el procedimiento PROCLOGISTIC con interacción de variables

Modelo	Variables continuas	Variables categóricas
REG04TS	Age AppliedAmount ExistingLiabilities ExpectedReturn LoanDuration LossGivenDefault MonthlyPaymentDay PorcRepaymentsPreviousLoans ProbabilityOfDefault County2*EmploymentDurationCurrentEmploye	Education LanguageCode Rating
REG04CS	AppliedAmount Interest LoanDuration MonthlyPayment MonthlyPaymentDay PorFreeCash PorcRepaymentsPreviousLoans County2*EmploymentDurationCurrentEmploye	Education NewCreditCustomer VerificationType

La interacción que entra en ambos modelos es la resultante de la región del prestatario y la antigüedad en su puesto de trabajo actual.

Para seguir descubriendo nuevos modelos, se han calculado regresiones con el método *stepwise* para trescientas semillas utilizando la macro *%randomselect* (autor, Javier Portela). La macro proporciona la salida con los modelos más frecuentes para estas semillas:

Tabla 15. Modelos más frecuentes regresión logística stepwise para TODASTRAM

	efecto	Cantidad de frecuencia	Porcentaje de frecuencia total
1	Intercept Age LoanDuration MonthlyPaymentDay LossGivenDefault NoOfPreviousLoansBef	257	85.382059801
2	Intercept Age LoanDuration ExistingLiabilities MonthlyPaymentDay LossGivenDefault NoOfPreviousLoansBef	44	14.617940199

Tabla 16. Modelos más frecuentes regresión logística stepwise para CONOCIDASTRAM

	efecto	Cantidad de frecuencia	Porcentaje de frecuencia total
1	Intercept Age LoanDuration ExistingLiabilities NoOfPreviousLoansBef	226	75.083056478
2	Intercept LoanDuration ExistingLiabilities NoOfPreviousLoansBef	62	20.598006645
3	Intercept Age LoanDuration ExistingLiabilities MonthlyPaymentDay NoOfPreviousLoansBef	11	3.6544850498
4	Intercept LoanDuration ExistingLiabilities MonthlyPaymentDay NoOfPreviousLoansBef	2	0.6644518272

Tabla 17. Modelos más frecuentes regresión logística stepwise para TODASORIG

	efecto	Cantidad de frecuencia	Porcentaje de frecuencia total
1	Intercept LoanDuration ExistingLiabilities MonthlyPaymentDay LossGivenDefault NoOfPreviousLoansBef MonthlyPayment ExpectedLoss ProbabilityOfDefault PorcRepaymentsPrevio AppliedAmount ExpectedReturn	100	33.222591362
2	Intercept LoanDuration MonthlyPaymentDay LossGivenDefault MonthlyPayment ExpectedLoss ProbabilityOfDefault PorcRepaymentsPrevio AppliedAmount PorFreeCash ExpectedReturn	75	24.916943522
3	Intercept LoanDuration ExistingLiabilities MonthlyPaymentDay LossGivenDefault MonthlyPayment ExpectedLoss ProbabilityOfDefault PorcRepaymentsPrevio AppliedAmount ExpectedReturn	62	20.598006645

Tabla 18. *Modelos más frecuentes regresión logística stepwise para CONOCIDASORIG*

	efecto	Cantidad de frecuencia	Porcentaje de frecuencia total
1	Intercept LoanDuration ExistingLiabilities MonthlyPaymentDay NoOfPreviousLoansBef MonthlyPayment Interest PorcRepaymentsPrevio AppliedAmount	94	31.22923588
2	Intercept LoanDuration MonthlyPaymentDay MonthlyPayment Interest PorcRepaymentsPrevio AppliedAmount PorFreeCash	88	29.235880399
3	Intercept LoanDuration ExistingLiabilities MonthlyPaymentDay MonthlyPayment Interest PorcRepaymentsPrevio AppliedAmount	85	28.239202658

Se seleccionan los modelos salidos de la regresión logística y de los modelos más frecuentes para, mediante el uso de la validación cruzada, comprobar cuál es el mejor de ellos, entendiendo por mejor modelo aquél con una menor tasa de fallos. Se utilizan 100 semillas para cada modelo y cada partición tiene 70% de datos de entrenamiento y 30% para validación.

Tabla 19. Modelos de regresión logística seleccionados para la comparación de modelos

Modelo	Origen Modelo
REG01TT	1r modelo más frecuente según randomselect
REG02TT	2o modelo más frecuente según randomselect
REG03TT	Regresión logística stepwise
REG01CT	1r modelo más frecuente según randomselect
REG02CT	2o modelo más frecuente según randomselect
REG03CT	Regresión logística stepwise
REG01TS	1r modelo más frecuente según randomselect
REG02TS	2o modelo más frecuente según randomselect
REG03TS	Regresión logística stepwise
REG04TS	Regresión logística stepwise, con interacciones
REG01CS	1r modelo más frecuente según randomselect
REG02CS	2o modelo más frecuente según randomselect

REG03CS	Regresión logística stepwise
REG04CS	Regresión logística stepwise, con interacciones

Gráfico 70. Distribución de media por modelo para los modelos de regresión logística

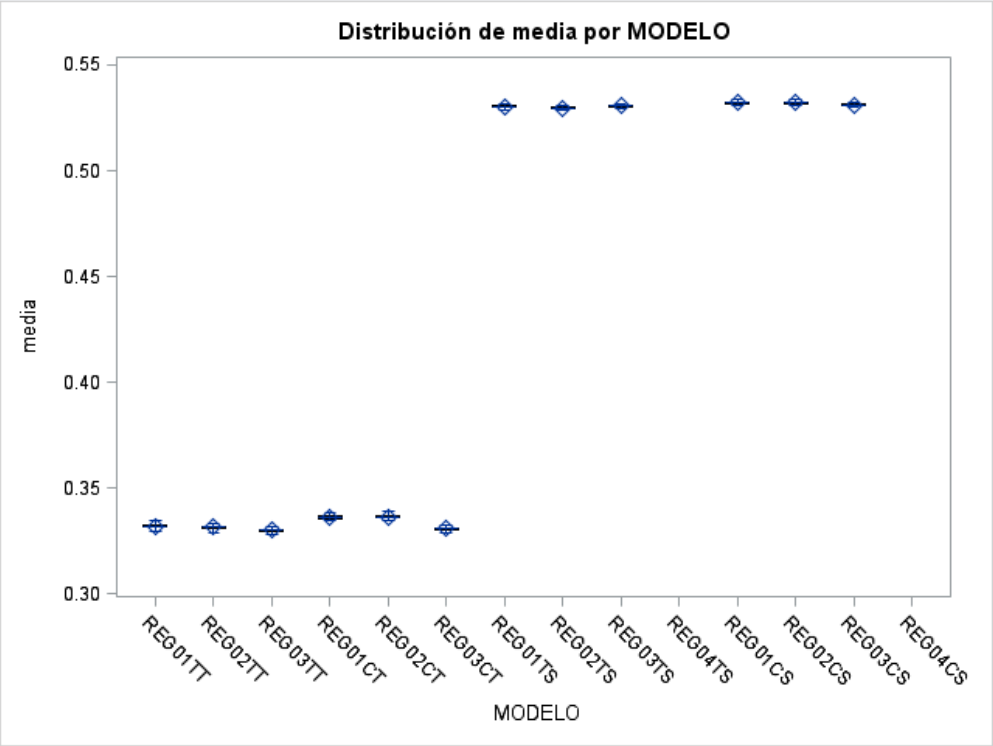
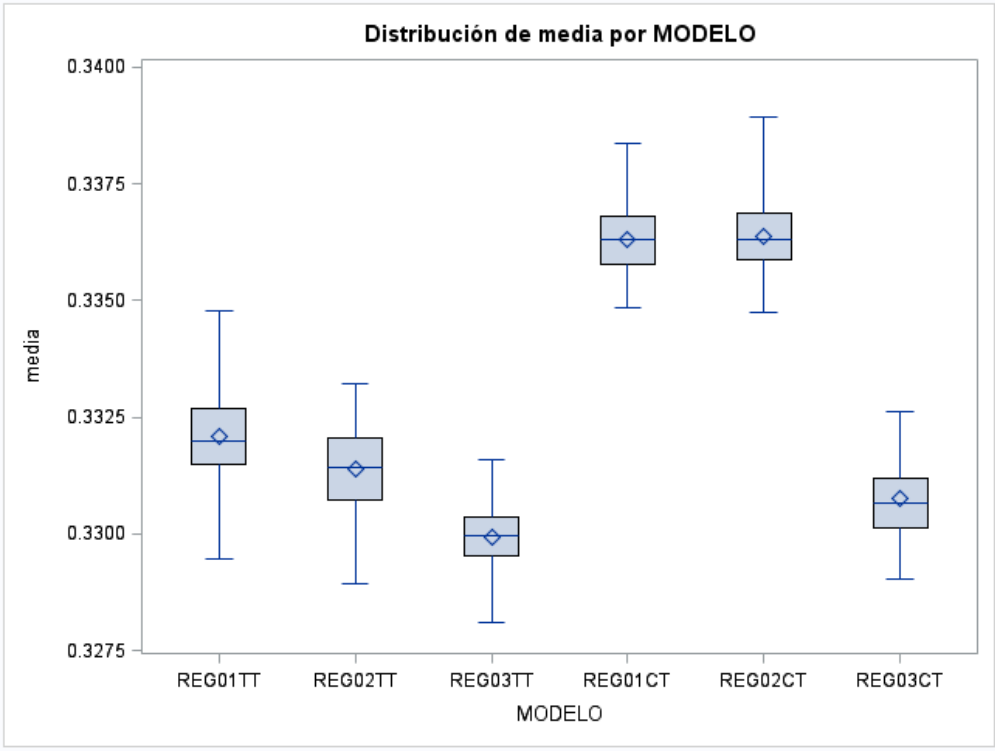


Gráfico 71. Detalle de la distribución de media por modelo para los mejores modelos de regresión logística



Se preseleccionan los modelos REG03TT y REG03CT para la comparación con modelos surgidos de otras técnicas.

Se observa que las regresiones que funcionan mejor son las que utilizan las variables tramificadas. Ambos modelos clasifican los impagados en función de:

- nivel educativo y región del solicitante
- ingresos y gastos están verificados y el ingreso disponible del solicitante
- cuota y duración del préstamo, importe e intereses y el día de pago
- importe de préstamos anteriores y porcentaje de estos préstamos ya pagado

La diferencia entre ambos está en que, mientras que REG03TT clasifica también en función de la edad y la probabilidad de impago calculada por Bondora, el modelo REG03CT clasifica también en función del tipo de vivienda, el idioma y los pasivos del solicitante.

6.2 Redes neuronales

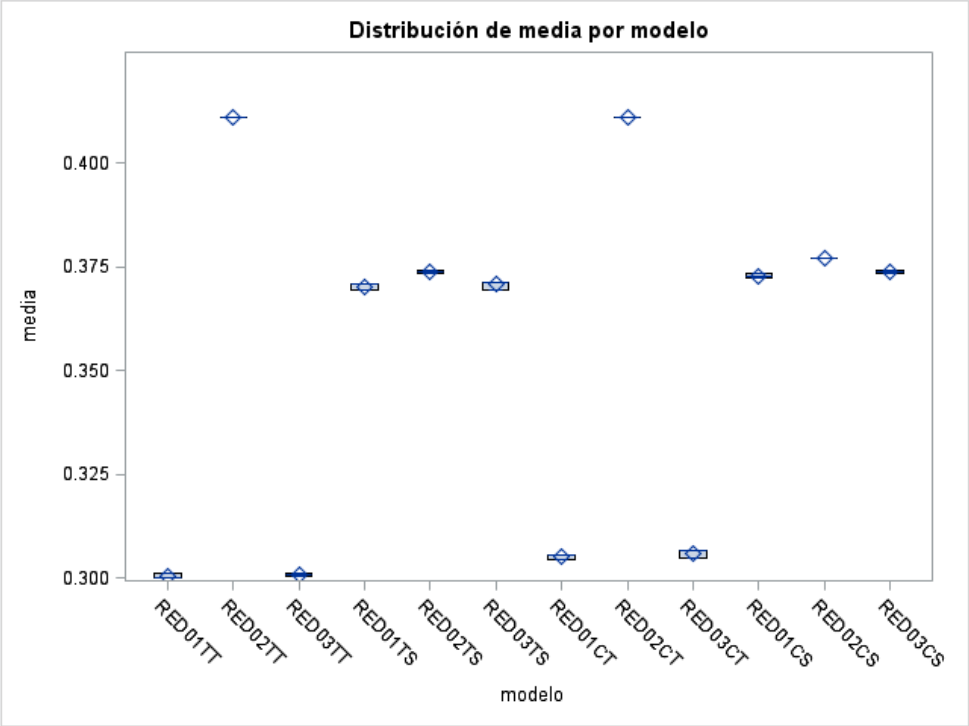
Para las redes neuronales se han seleccionado los siguientes modelos para ser comparados, tras probar previamente con distintos métodos para el entrenamiento de los mínimos cuadrados y con distintos nodos:

Tabla 20. Modelos de redes seleccionados para la comparación de modelos

Modelo	Nº Nodos	Función Activación	Método
RED01TT	10	tanh	bprop
RED02TT	10	log	bprop
RED03TT	10	linh	bprop
RED01TS	10	tanh	bprop
RED02TS	10	log	bprop
RED03TS	10	linh	bprop
RED01CT	10	tanh	bprop
RED02CT	10	log	bprop
RED03CT	10	linh	bprop
RED01CS	10	tanh	bprop
RED02CS	10	log	bprop
RED03CS	10	linh	bprop

Tras entrenar las redes y utilizar validación cruzada, las mejores redes en función de la tasa de error son las redes RED01TT y RED03TT. En el caso de las redes neuronales los mejores resultados se dan en modelos que utilizan las variables tramificadas. La función de activación que mejor funciona es la tangencial.

Gráfico 72. Distribución de media por modelo para los modelos de redes neuronales



6.3 Bagging

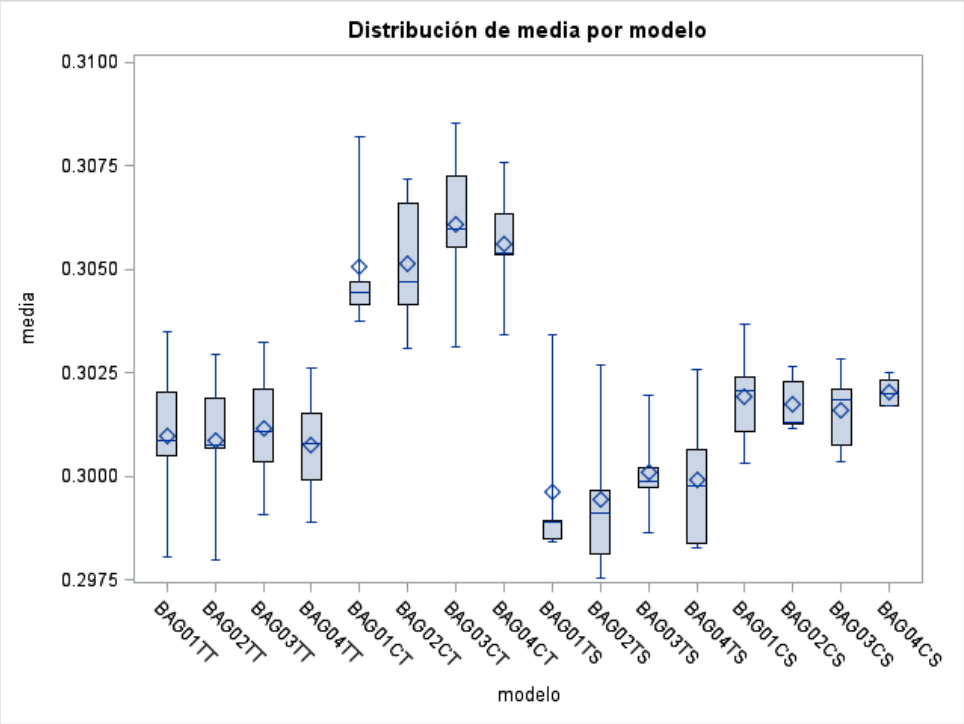
Para los modelos de Baggin se han seleccionado modelos que generan un número distinto de árboles para promediar el modelo y cuyos árboles tiene distinto número de hojas. Siempre tienen dos ramas, ya que estamos modelizando una variable dependiente binaria.

Tabla 21. Modelos de bagging seleccionados para la comparación de modelos

Modelo	% Muestra	Nº Hojas	Nº Máx Árboles
BAG01TT	70	25	20
BAG02TT	70	25	40
BAG03TT	70	15	20
BAG04TT	70	15	40
BAG01CT	70	25	20
BAG02CT	70	25	40
BAG03CT	70	15	20
BAG04CT	70	15	40
BAG01TS	70	25	20
BAG02TS	70	25	40
BAG03TS	70	15	20
BAG04TS	70	15	40
BAG01CS	70	25	20
BAG02CS	70	25	40
BAG03CS	70	15	20
BAG04CS	70	15	40

Tras comparar los modelos, se observa que los que tienen la media de error más baja son los que utilizan todas las variables sin tramificar. No obstante, son modelos que tienen bastante dispersión con respecto a la media.

Gráfico 73. Distribución de media por modelo para los modelos de bagging



En cambio, los modelos que utilizan solamente las variables relativas al prestatario, aunque tienen la media un poquito más alta, son modelos con menor dispersión, más robustos. Por tanto, para compararlos con los modelos de otras técnicas se van a seleccionar los modelos BAG02TS, BAG02CS y BAG04CS.

6.4 Random Forest

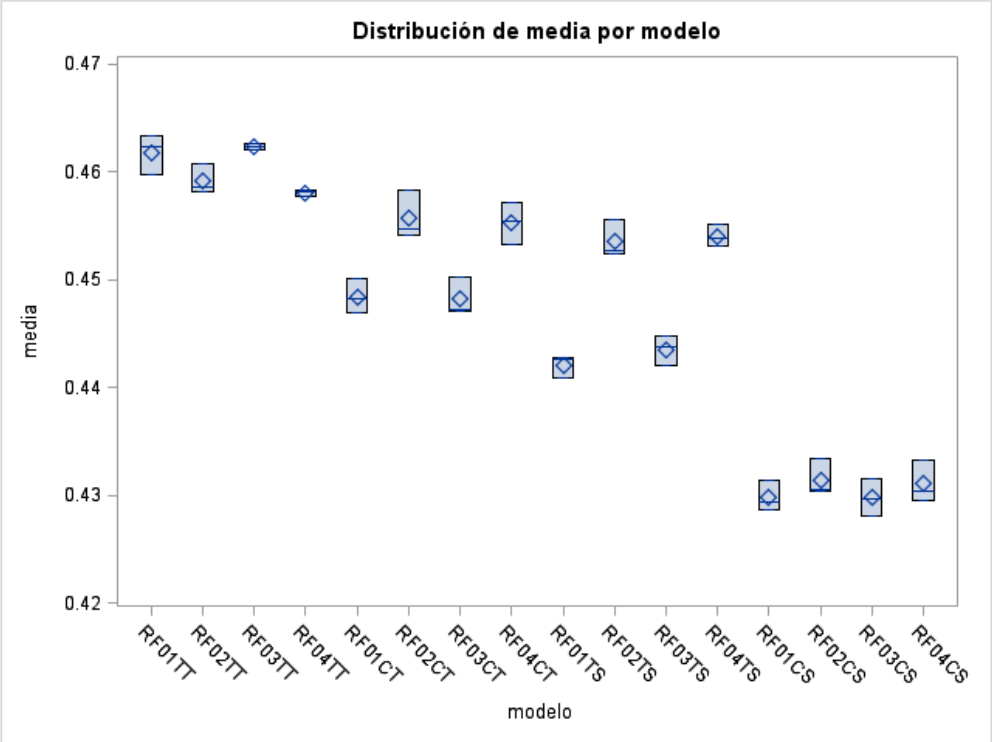
Se han seleccionado los siguientes modelos, que generan distinto número máximo de árboles sobre los que promediar los resultados, tienen distinto número de hojas y para todos, tras varias pruebas, se ha decidido que el número máximo de variables que permitan sea 20. Todos los árboles generados tienen dos ramas.

Al comparar los modelos, se observa que para el set de datos que mejor funcionan es para CONOCIDASORIG, aunque ya se ve que los resultados de media de tasa de fallos son bastante peores que para las técnicas anteriores. Esto indica que puede que los modelos de bagging están sobreajustados, ya que la diferencia en la media de la tasa de error es de diez puntos porcentuales. Para la comparación con otras técnicas se seleccionan los modelos RF01CS y RF03CS.

Tabla 22. Modelos de random forest seleccionados para la comparación de modelos

Modelo	% Muestra	Nº Variables	Nº Hojas	Nº Máx Árboles
RF01TT	70	20	5	20
RF02TT	70	20	15	20
RF03TT	70	20	5	40
RF04TT	70	20	15	40
RF01CT	70	20	5	20
RF02CT	70	20	15	20
RF03CT	70	20	5	40
RF04CT	70	20	15	40
RF01TS	70	20	5	20
RF02TS	70	20	15	20
RF03TS	70	20	5	40
RF04TS	70	20	15	40
RF01CS	70	20	5	20
RF02CS	70	20	15	20
RF03CS	70	20	5	40
RF04CS	70	20	15	40

Gráfico 74. Distribución de media por modelo para los modelos de random forest



6.5 Gradient Boosting

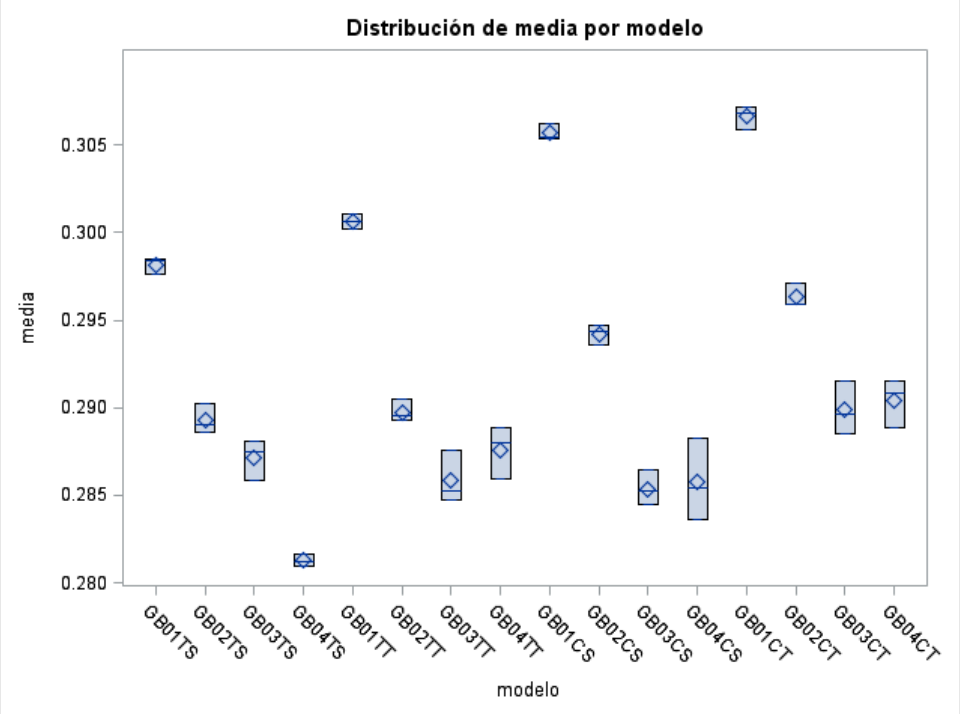
Dado que el método consiste en repetir la construcción de árboles de clasificación modificando las predicciones iniciales en cada iteración para minimizar los residuos en sentido decreciente, cuantas más iteraciones haga el modelo, menor será el residuo,

tendiendo a cero. Se parametrizan los modelos con distintas tasas de aprendizaje, número final de hojas y número de iteraciones en aras de reducir el sobreajuste. Los modelos seleccionados para la comparación de modelos por validación cruzada son:

Tabla 23. Modelos de gradient boosting seleccionados para la comparación de modelos

Modelo	Tasa Aprendizaje	Nº Iteraciones	Profundidad
GB01TS	0,01	100	5
GB02TS	0,01	100	15
GB03TS	0,05	300	5
GB04TS	0,05	300	15
GB01TT	0,01	100	5
GB02TT	0,01	100	15
GB03TT	0,05	300	5
GB04TT	0,05	300	15
GB01CS	0,01	100	5
GB02CS	0,01	100	15
GB03CS	0,05	300	5
GB04CS	0,05	300	15
GB01CT	0,01	100	5
GB02CT	0,01	100	15
GB03CT	0,05	300	5
GB04CT	0,05	300	15

Gráfico 75. Distribución de media por modelo para los modelos de gradient boosting



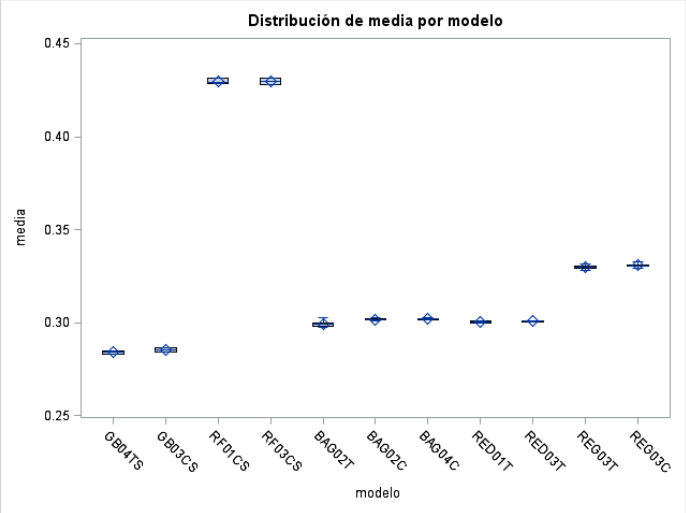
Los modelos con la media de tasa de fallos más baja son GB04TS y GB03CS, ambos con 300 iteraciones aunque cada uno con una profundidad distinta.

7. Valoración de los modelos

7.1 Comparación de los mejores modelos

Se comparan por validación cruzada los modelos seleccionados para cada técnica.

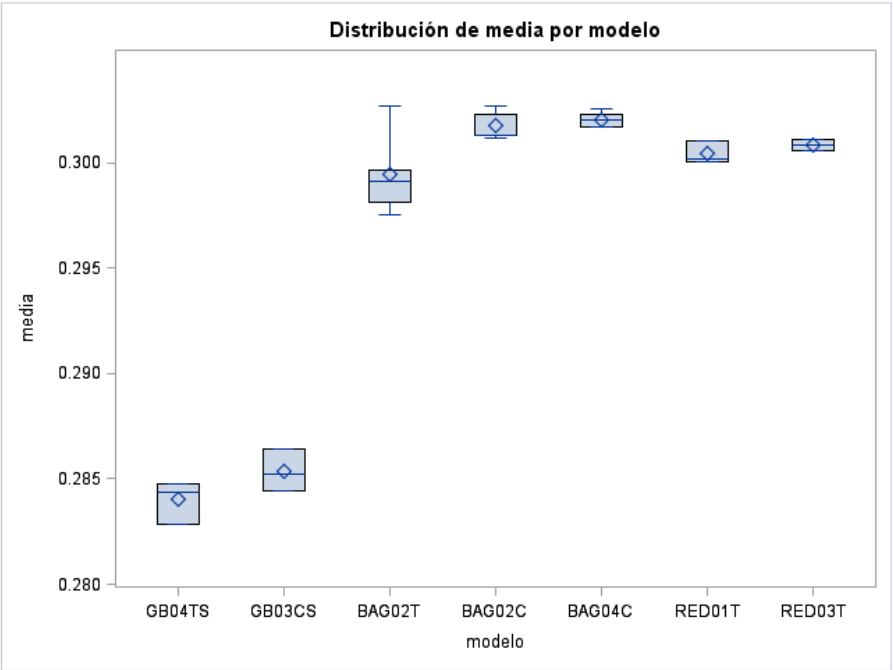
Gráfico 76. Distribución de media por modelo para los mejores modelos



Se observa que los modelos con la media de la tasa de fallos más alta son los de Random Forest, a mucha distancia de los siguientes peores, que son las Regresiones Logísticas. Entre las Redes Neuronales y el Bagging apenas hay diferencia y los que presentan una mejor tasa de fallos son los modelos de Gradient Boosting.

Los modelos que mejor clasifican tienen preferencia por las variables sin tramificar, pero las diferencias en la tasa de fallos no son grandes entre utilizar todas las variables o solo las informadas por el prestatario.

Gráfico 77. Detalle de la distribución de media por modelo para los mejores modelos

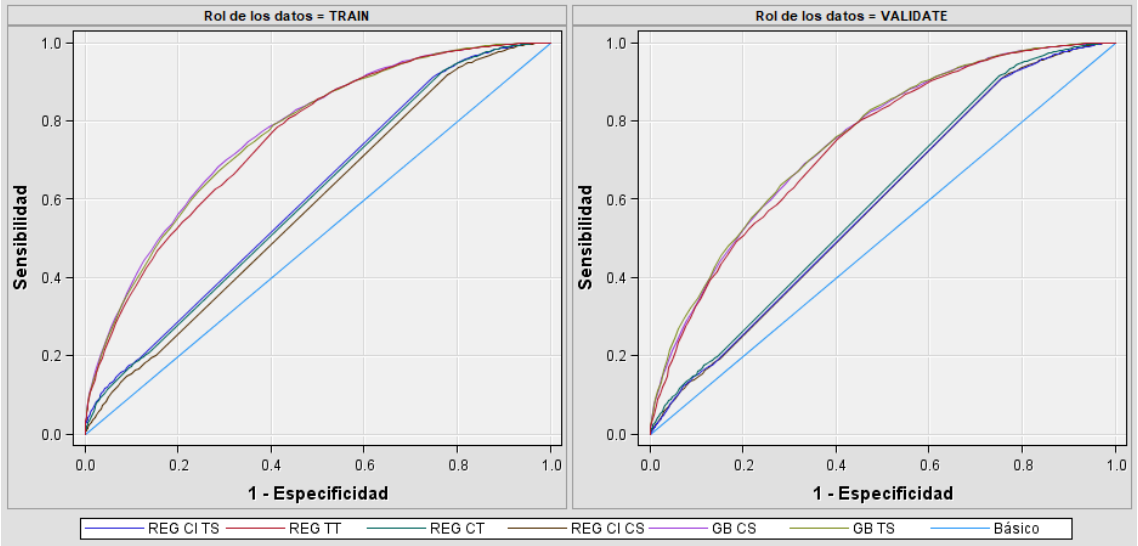


El mejor modelo según la comparación por validación cruzada es el Gradient Boosting para todas las variables sin interacciones.

7.2 Bondad del modelo seleccionado

Generamos la curva ROC para el mejor, así como para el segundo mejor modelo, GB03CS y las regresiones logísticas seleccionadas anteriormente. Esta curva es una representación de la relación entre la Sensibilidad (tasa de verdaderos positivos, es decir, observaciones que clasificarían la sucesión del evento correctamente) y la Especificidad (tasa de falsos positivos, es decir, observaciones que clasificarían erróneamente la sucesión del evento). Cuanto más a la izquierda se sitúa la curva, mejor clasifica el modelo ya que, para una tasa de falsos positivos dada, mayor es la tasa de verdaderos positivos.

Figura 12. Curvas ROC para datos de entrenamiento y validación



Con respecto a la bondad del modelo, en el gráfico se ve claramente que compiten los dos modelos de Gradient Boosting junto con la regresión logística para todas las variables tramificadas.

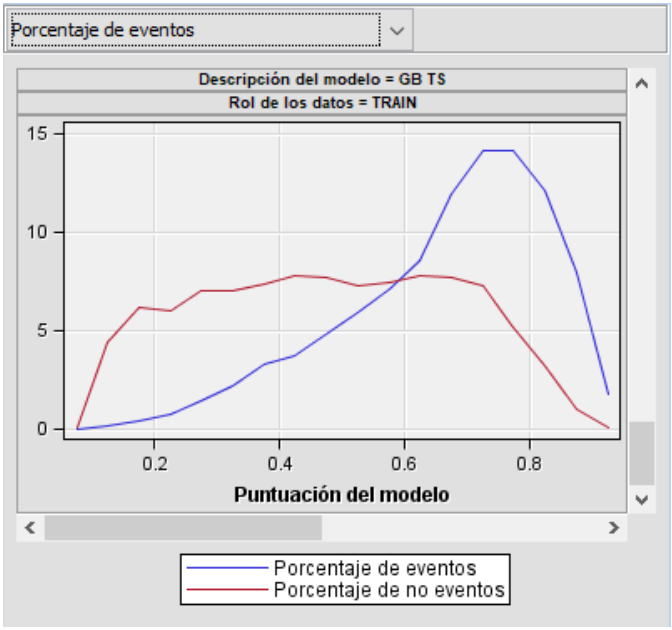
Si se calcula el área bajo la curva ROC, AUC, se ve que el área es mayor para el modelo GB04TS, el seleccionado también como mejor modelo por el método de la validación cruzada.

Tabla 24. Área bajo la curva para los modelos comparados

Modelo seleccionado	Nodo predecesor	Nodo del modelo	Descripción del modelo	Variable objetivo	Etiqueta objetivo	Criterio de selección: Valid: Misclassification Rate	Entrenar: índice Roc
Y	Boost	Boost	GB TS	Impago		0.296823	0.765
	Boost2	Boost2	GB CS	Impago		0.300264	0.769
	Reg4	Reg4	REG TT	Impago		0.304851	0.755
	Reg3	Reg3	REG CT	Impago		0.359445	0.602
	Reg	Reg	REG CI CS	Impago		0.365982	0.583
	Reg5	Reg5	REG CI TS	Impago		0.367244	0.609

Un AUC del 0.765 indica que el modelo es aceptable. Si se analiza cómo el modelo separa la tasa de eventos de la tasa de no eventos, se ve que no las separa demasiado bien, ya que tanto a un lado como a otro de la intersección de ambas curvas quedan muchas observaciones de evento/no evento.

Figura 13. Curvas de tasa de evento para el mejor modelo



Por último, se clasifican los prestatarios de la muestra de datos test utilizando el modelo de Gradient Boosting para todas las variables sin tramificar. El modelo clasifica a el 30% de los prestatarios como buenos y al 70% como malos.

Rol de los datos=SCORE Tipo de salida=CLASSIFICATION

Variable	Valor numérico	Valor formateado	Número de ocurrencias	Porcentaje
I_Impago	.	0	1064	29.9887
I_Impago	.	1	2484	70.0113

Se compara el valor de clasificación otorgado por el modelo de Gradient Boosting a los préstamos de la muestra test con el valor de la variable ***Impago*** del dataset **LoanData** para estos mismos préstamos, haciendo el cruce por la variable ***LoanId***. Se obtiene que, del 30% de los préstamos clasificados como negativos (no evento) y que son los que interesan al inversor, pues son los que no realizan impagos, un tercio finalmente sí hará algún impago.

Tabla 25. Tasa de fallo del mejor modelo sobre la muestra test

Precisión	Núm Observaciones	Porcentaje	% Fallo o Acierto
FN	405	11.41%	**35.57%**
FP	857	24.15%	
VN	659	18.57%	**64.43%**
VP	1627	45.86%	

El modelo de Gradient Boosting clasifica bien en un 64.43% de las veces.

La plataforma no proporciona un modelo de clasificación, sino un rating. Se podría asimilar el rating de Bondora a un modelo de clasificación si se considera que los ratings AA y A clasifican el no evento (el pago en nuestro caso) y el resto de ratings clasifican el evento (el impago). Éste sería el método que podría seguir el inversionista para tomar la decisión de en qué préstamos invertir. Si se aplica este modelo de clasificación a los datos test la tasa de fallos es la siguiente:

Tabla 26. Tasa de fallo de considerar los préstamos con rating AA y A como no evento en los datos test

Precisión	Num Observaciones	Porcentaje	% Fallo o Acierto
FN	126	3.55%	**40.70%**
FP	1318	37.15%	
VN	198	5.58%	**59.30%**
VP	1906	53.72%	

Luego el modelo de Gradient Boosting mejora ligeramente la tasa de fallos además de ampliar el abanico de préstamos en los que puede invertir el inversor de Bondora.

8. Conclusiones

El objetivo del trabajo era, por un lado, comprobar si el modelo de valoración del riesgo que utiliza la plataforma Bondora está alineado con los intereses del inversor, dado el modelo de negocio de la plataforma y, por otro, construir un modelo de clasificación que permita al inversor elegir en qué préstamos invertir acotando el riesgo de impago.

La plataforma valora el riesgo de cada prestatario en función de la probabilidad de impago, la pérdida esperada y la probabilidad de recuperación del préstamo, otorgando a cada prestatario una calificación que va de la AA (la mejor) a HR (la peor). Esta calificación es la única herramienta de que dispone el inversor para decidir en qué préstamos invertir. Para comprobar si este modelo de valoración del riesgo es congruente con los intereses del inversor, se ha modelizado el modelo de clasificación de prestatarios con dos distintos sets de datos: unos que incluyen las variables de calificación que otorga la plataforma a los prestatarios (***Rating***, ***LossGivenDefault***, ***ExpectedReturn***, ***ProbabilityOfDefault***) y otros que no las incluyen, de manera que puedan competir entre ellos para ver cuál clasifica mejor a los prestatarios en función de la definición de préstamo malo dada y qué variables entran en el modelo.

Se comprueba que el mejor modelo, en función de la media de la tasa de fallos, incluye las variables de calificación que Bondora ofrece. La inclusión de las variables de calificación no supone una gran mejora en los modelos, pues el segundo mejor modelo no las utiliza, pero en ningún caso su inclusión empeora los modelos de clasificación para la definición de préstamo malo que se ha dado.

Se puede concluir, entonces, que el rating que la plataforma pone a disposición de los inversores sí está alineado con el objetivo de los inversores; detectar aquellos prestatarios que van a realizar impago para no invertir en esos préstamos.

No obstante, el inversor, además de la valoración del riesgo, necesita saber en qué préstamos invertir. Por un lado, a priori, el modelo de clasificación que podría construirse el inversor con la información que facilita la plataforma sería invertir sólo en aquellos préstamos con rating bueno (rating AA y A), considerando que el resto de préstamos van a realizar impago en algún momento. Por otro lado, el resultado de la modelización de este trabajo indica que el mejor modelo de clasificación de prestatarios es el algoritmo Gradient Boosting de 15 nodos de profundidad, una tasa de aprendizaje del 0.05 y 300 iteraciones, y que éste utiliza las variables de calificación de Bondora.

Si se compara la tasa de fallos del modelo construido sobre el rating con la tasa de fallos del modelo de Gradient Boosting construido en este trabajo, se observa que el segundo mejora al primero. Por tanto, aunque la variable ***Rating*** que facilita la plataforma ayuda a discriminar los préstamos buenos de los malos, los inversores pueden tomar una mejor decisión si construyen su propio modelo para decidir en qué préstamos invertir.

Durante el proceso de análisis de variables y construcción de modelos se ha detectado que hay variables que, de estar aprovisionadas correctamente, entrarían a formar parte de los modelos, como es el caso de la variable ***County***. Actualmente es un campo de texto libre que, de ser un campo que se pudiera seleccionar de una forma estandarizada podría mejorar el modelo de calificación de la plataforma.

Por último, una advertencia a los inversores despistados: la rentabilidad que promete el tipo de interés de los préstamos dista mucho de la rentabilidad esperada. Como una evolución a este trabajo se podría crear un modelo para construir una cartera de inversión en los préstamos calificados como no evento que maximice la rentabilidad del inversor.

Bibliografía

Boyes, W. J.; Hoffman, D. L. y Low, A. S. (1989): «*An econometric analysis of the bank credit scoring problem*», Journal of Econometrics, vol. 41, págs. 3-14.

Centro de ayuda de Bondora. https://support.bondora.com/hc/en-us

Han, J.; Pei, J. y Kamber, M. (2011). *Data mining: concepts and techniques*. Elseveier.

Hastie, T.; Tibshirani, R. y Friedman, J. (2009). *The elements of statistical learning*. New York. NY: Springer.

James, G.; Witten, D.; Hastie, T. y Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York: Springer.

Michie,D.; Spiegelhalter, D.J. y Taylor, C.C. (1994). *Machine Learning, Neural and Statistical Clasification*

Portela, Javier(2006). *Manual de Programación en SAS*. Ediciones Fiec

Rodriguez Martin, Sara (2017). *Los modelos colaborativos y bajo demanda en plataformas digitales. https://www.adigital.org/media/plataformas-colaborativas.pdf*

Roopam Upadhyay. Blog YOU CANalytics. http://ucanalytics.com/blogs/

Siddiqi, Naeem (2006). *Credit Risk Scorecards Developing and Implementing Intelligent Credit Scoring*. https://support.sas.com/content/dam/SAS/support/en/books/cedit-risk-scorecards/59376_excerpt.pdf

ANEXOS

Anexo I. Descripción de las variables de Loandata.

Variable	Tipo	Subtipo	Descripción	Tipo de dato	Tipo de variable
ModelVersion	Préstamo	Calificación	La versión del modelo de Calificación utilizado para emitir la Calificación de Bondora	Número	Categórica
Rating	Préstamo	Calificación	Calificación de Bondora emitida por el modelo de Calificación	Número	Continua
Rating_V0	Préstamo	Calificación	Calificación de Bondora emitida por la versión 0 del modelo de Calificación	Número	Continua
Rating_V1	Préstamo	Calificación	Calificación Bondora emitida por la versión 1 del modelo de Calificación	Número	Continua
Rating_V2	Préstamo	Calificación	Calificación Bondora emitida por la versión 2 del modelo de Calificación	Número	Continua
ExpectedLoss	Préstamo	Calificación	Pérdida esperada calculada por el modelo de calificación actual	Número	Continua
ExpectedReturn	Préstamo	Calificación	Rendimiento esperado calculado por el modelo de calificación actual	Número	Continua
EL_V0	Préstamo	Calificación	Pérdida esperada calculada por la versión especificada del modelo de Calificación	Número	Continua
EL_V1	Préstamo	Calificación	Pérdida esperada calculada por la versión especificada del modelo de Calificación	Número	Continua
CurrentDebtDaysPrimary	Préstamo	Default	Número de días de deuda del principal del préstamo en el momento de la actualización del conjunto de datos	Número	Continua
CurrentDebtDaysSecondary	Préstamo	Default	Número de días de deuda de los intereses del préstamo	Número	Continua
LossGivenDefault	Préstamo	Default	Proporciona el porcentaje de exposición pendiente en el momento del incumplimiento que es probable que un inversor pierda si un préstamo realmente incumple. Esto significa la proporción de fondos perdidos para el inversor después de toda la recuperación esperada y la contabilización del valor temporal del dinero recuperado. En general, el parámetro LGD se debe estimar en función de las recuperaciones históricas. Sin embargo, en mercados nuevos donde la experiencia limitada no nos permite una pérdida más precisa dadas las estimaciones predeterminadas, se supone una LGD del 90%.	Número	Continua
ProbabilityOfDefault	Préstamo	Default	Probabilidad de impago, se refiere a la probabilidad de impago del préstamo dentro del horizonte de un año	Número	Continua
DefaultDate	Préstamo	Default	La fecha en que el préstamo entró en estado de impago y se inició el proceso de cobro. Si el impago ya se ha recuperado, aparece vacía.	Fecha	Continua
PlannedPrincipalPostDefault	Préstamo	Default	La cantidad de capital que se planificó para recibir después de que ocurriera el incumplimiento	Número	Continua
PlannedInterestPostDefault	Préstamo	Default	La cantidad de interés que se planificó para recibir después de que ocurriera el incumplimiento	Número	Continua
EAD1	Préstamo	Default	Exposición en caso de impago, principal pendiente	Número	Continua
EAD2	Préstamo	Default	Exposición en caso de impago, importe del préstamo menos todos los pagos previos al momento de impago	Número	Continua

ActiveLateCategory	Préstamo	Default	Cuando un préstamo está en Deuda Principal, se clasificará por Días de Deuda del Principal: 1-7 8-15 16-30 31-60 61-90 91-120 121-150 151-180 180+	Texto	Categórica
WorseLateCategory	Préstamo	Default	Muestra el último período más largo de días en que el préstamo estaba en deuda principal: 1-7 8-15 16-30 31-60 61-90 91-120 121-150 151-180 180+	Texto	Categórica
GracePeriodStart	Préstamo	Default	Fecha del comienzo del período de gracia	Fecha	Continua
GracePeriodEnd	Préstamo	Default	Fecha del final del período de gracia	Fecha	Continua
ActiveLateLastPaymentCategory	Préstamo	Default	Muestra cuántos días han pasado desde el último pago y categorizado si está vencido: 1-7 8-15 16-30 31-60 61-90 91-120 121-150 151-180 180+	Texto	Categórica
PrincipalWriteOffs	Préstamo	Fallido	Principal llevado a pérdidas, no recuperable	Número	Continua
InterestAndPenaltyWriteOffs	Préstamo	Fallido	Intereses y penalizaciones llevados a pérdidas, no recuperables	Número	Continua
LoanNumber	Préstamo	Identificativo	Número de préstamo único que aparece en el sistema Bondora	Texto / Número	Categórica
LoanDate	Préstamo	Identificativo	Fecha en que se emitió el préstamo	Fecha	Continua
ContractEndDate	Préstamo	Identificativo	Fecha en que terminó el contrato de préstamo	Fecha	Continua
MaturityDate_Original	Préstamo	Identificativo	Fecha de vencimiento del préstamo de acuerdo con el calendario original del préstamo	Fecha	Continua
Amount	Préstamo	Identificativo	Importe que el prestatario recibió en el mercado primario. Este es el saldo principal de su compra en el mercado secundario	Número	Continua
LoanDuration	Préstamo	Identificativo	Plazo del préstamo (en meses).	Número	Continua

UseOfLoan	Préstamo	Identificativo	Describe el uso del préstamo: 0 Consolidación préstamo 1 Inmuebles 2 Reformas en la vivienda 3 Negocio 4 Educación 5 Viajes 6 Vehiculos 7 Otros 8 Salud 101 Financiación para el capital de trabajo 102 Adquisición de equipos de maquinaria 103 Renovación de inmuebles 104 Financiación de cuentas por cobrar 105 Aquisición de medios de transporte 106 Financiación de construcciones 107 Compra de existencias 108 Adquisición de inmuebles 109 Garantización de obligaciones (provisiones) 110 Otros Todos los códigos en formato 1XX son para préstamos comerciales que no son compatibles desde octubre 2012	Número	Categórica
DebtOccuredOn	Préstamo	Default	La fecha en que se impagó el principal	Fecha	Continua
DebtOccuredOnForSecondary	Préstamo	Default	La fecha en que se impagaron los intereses	Fecha	Continua
Status	Préstamo	Identificativo	El estado actual de la solicitud de préstamo Current Late Repaid	Texto	Categórica
MaturityDate_Last	Préstamo	Identificativo	Fecha de vencimiento del préstamo en la fecha de generación de informes (en el momento actual)	Fecha	Continua
FirstPaymentDate	Préstamo	Pagos	Fecha del primer pago de acuerdo al calendario inicial del préstamo	Fecha	Continua
MonthlyPayment	Préstamo	Pagos	Monto estimado que el prestatario debe pagar cada mes	Número	Continua
MonthlyPaymentDay	Préstamo	Pagos	El día del mes programado para los pagos de los préstamos. La fecha real se ajusta para fines de semana y festivos (por ejemplo, si es 10 o Domingo entonces el pago se realizará el día 11 de ese mes)	Número	Continua
ActiveScheduleFirstPaymentReache	Préstamo	Pagos	Si se ha alcanzado la primera fecha de pago de acuerdo con el calendario activo	Texto	Binaria
LastPaymentOn	Préstamo	Pagos	La fecha del último pago actual recibido del prestatario	Fecha	Continua
PrincipalPaymentsMade	Préstamo	Pagos	Pagos del principal hechos por el prestatario. Incluye el principal recuperado mediante proceso de recobro	Número	Continua
InterestAndPenaltyPaymentsMade	Préstamo	Pagos	Pagos de intereses y penalizaciones hecho por el prestatario	Número	Continua
PrincipalBalance	Préstamo	Pagos	Principal que todavía debe ser pagado por el prestatario	Número	Continua
InterestAndPenaltyBalance	Préstamo	Pagos	Intereses y penalizaciones que todavía debe ser pagado por el prestatario	Número	Continua
NextPaymentDate	Préstamo	Pagos	Según calendario, la próxima fecha para que el prestatario realice su pago	Fecha	Continua
NextPaymentNr	Préstamo	Pagos	De acuerdo al calendario, el número del próximo pago	Número	Continua

NrOfScheduledPayments	Préstamo	Pagos	According to schedule the count of scheduled payments	Número	Continua
PlannedPrincipalTillDate	Préstamo	Pagos	De acuerdo con el cronograma activo, la cantidad de capital que la inversión debería haber recibido	Número	Continua
PlannedInterestTillDate	Préstamo	Pagos	De acuerdo con el cronograma activo, la cantidad de interés que debería haber recibido la inversión	Número	Continua
PrincipalOverdueBySchedule	Préstamo	Pagos	De acuerdo con el cronograma actual, el principal vencido	Número	Continua
PrincipalRecovery	Préstamo	Recobro	El principal que se recuperó debido al proceso de cobranza de los préstamos de la deuda	Número	Continua
InterestRecovery	Préstamo	Recobro	Interés recuperado debido al proceso de cobranza de préstamos de la deuda	Número	Continua
RecoveryStage	Préstamo	Recobro	Etapa actual según el modelo de recuperación: 1 Cobro 2 Recuperación 3 Cancelación - llevado a pérdidas	Número	Categórica
StageActiveSince	Préstamo	Recobro	Cuánto tiempo ha estado activa la etapa de recuperación actual	Número	Continua
PrincipalDebtServicingCost	Préstamo	Recobro	Coste del servicio relacionado con la recuperación de la deuda en función del capital de la inversión	Número	Continua
InterestAndPenaltyDebtServicingCost	Préstamo	Recobro	Coste del servicio relacionado con la recuperación de la deuda en función del interés y las penalidades de la inversión	Número	Continua
RefinanceLiabilities	Préstamo	Reestructuración	Montante total de los pasivos después de la refinanciación	Número	Continua
Restructured	Préstamo	Reestructuración	Indica si la fecha de vencimiento original del préstamo se ha aumentado en más de 60 días: -True -False	Texto	Binaria
ReScheduledOn	Préstamo	Reestructuración	La fecha en que se asignó un nuevo calendario de pago al prestatario	Fecha	Continua
LanguageCode	Solicitante	Sociales	1 Estonio 2 Inglés 3 Ruso 4 Finlandés 5 Alemán 6 Español 9 Eslovaco	Número	Categórica
Age	Solicitante	Sociales	Edad del prestatario en años (del solicitante)	Número	Continua
DateOfBirth	Solicitante	Sociales	La fecha del nacimiento del prestatario	Fecha	Categórica
Gender	Solicitante	Sociales	0 Hombre 1 Mujer 2 Indefinido	Número	Categórica
Country	Solicitante	Sociales	País de residencia del prestatario	Texto	Categórica
County	Solicitante	Sociales	Región del prestatario	Texto	Categórica
City	Solicitante	Sociales	Ciudad del prestatario	Texto	Categórica
Education	Solicitante	Sociales	1 Educación primaria 2 Educación básica 3 Formación profesional 4 Educación secundaria 5 Educación superior	Número	Categórica
MaritalStatus	Solicitante	Sociales	1 Casado/a 2 Pareja de hecho 3 Soltero/a 4 Divorciado/a 5 Viudo/a	Número	Categórica
NrOfDependants	Solicitante	Sociales	Número de niños u otros dependientes	Número	Categórica

EmploymentStatus	Solicitante	Sociales	1 Desempleado 2 Empleados parcialmente 3 Totalmente empleado 4 Autónomos 5 Emprendedor 6 Jubilados	Número	Categórica
EmploymentDurationCurrentEmploye	Solicitante	Sociales	Tiempo de trabajo con el empleador actual: TrialPeriod UpTo1Year UpTo2Years UpTo3Years UpTo4Years UpTo5Years MoreThan5Years Retiree Others	Texto	Categórica
EmploymentPosition	Solicitante	Sociales	Posición con el empleador actual	Texto	Categórica
WorkExperience	Solicitante	Sociales	Experiencia laboral total: LessThan2Years 2To5Years 5To10Years 10To15Years 15To25Years MoreThan25Years	Texto	Categórica
OccupationArea	Solicitante	Sociales	1 Otros 2 Minería 3 Procesamiento 4 Energía 5 Utilidades 6 Construcción 7 al por menor y al por mayor 8 Transporte y almacenamiento 9 Hostelería y restauración 10 Información y telecomunicaciones 11 Finanzas y seguros 12 inmobiliaria 13 Investigación 14 Administrativo 15 Servicio Civil y militar 16 Educación 17 Salud y asistencia social 18 Arte y entretenimiento 19 Agricultura, silvicultura y pesca	Texto	Categórica
HomeOwnershipType	Solicitante	Sociales	0 Personas sin Hogar 1 Propietario 2 Vivir con los padres 3 Inquilino, propiedad pre-amueblado 4 Inquilino, propiedad sin amueblar 5 Vivienda social 6 coarrendatario 7 Copropiedad 8 Hipoteca 9 Propietario con gravamen	Texto	Categórica
IncomeFromPrincipalEmployer	Solicitante	Económicos	Salario	Número	Continua
IncomeFromPension	Solicitante	Económicos	Pensión	Número	Continua
IncomeFromFamilyAllowance	Solicitante	Económicos	Subsidio familiar	Número	Continua
IncomeFromSocialWelfare	Solicitante	Económicos	Subsidio bienestar social	Número	Continua

IncomeFromLeavePay	Solicitante	Económicos	Ingresos del prestatario por licencia de paternidad	Número	Continua
IncomeFromChildSupport	Solicitante	Económicos	Ingreso de la manutención de menores	Número	Continua
IncomeOther	Solicitante	Económicos	Otros ingresos	Número	Continua
IncomeTotal	Solicitante	Económicos	Total de ingresos	Número	Continua
ExistingLiabilities	Solicitante	Económicos	Número de pasivos existentes de prestatario	Número	Continua
LiabilitiesTotal	Solicitante	Económicos	Total pasivos mensuales	Número	Continua
DebtToIncome	Solicitante	Económicos	Proporción del ingreso bruto mensual del prestatario que se destina al pago de préstamos	Número	Continua
FreeCash	Solicitante	Económicos	Ingreso disponible después de pagar los pasivos mensuales	Número	Continua
NoOfPreviousLoansBeforeLoan	Solicitante	Económicos	Número de préstamos anteriores	Número	Continua
AmountOfPreviousLoansBeforeLoan	Solicitante	Económicos	Valor de los préstamos anteriores	Número	Continua
PreviousRepaymentsBeforeLoan	Solicitante	Económicos	Cuánto pagó el prestatario de sus préstamos anteriores antes del préstamo	Número	Continua
PreviousEarlyRepaymentsBeforeLoa	Solicitante	Económicos	Cuánto pagó el prestatario como amortizaciones anticipadas de sus préstamos anteriores	Número	Continua
PreviousEarlyRepaymentsCountBefo	Solicitante	Económicos	Cuántas amortizaciones anticipadas ha realizado el prestatario de sus préstamos anteriores	Número	Continua
UserName	Solicitante	Identificativo	Nombre de usuario del cliente solicitante de un préstamo que da Bondora	Texto	Categórica
NewCreditCustomer	Solicitante	Identificativo	Muestra si el cliente tiene historia crediticia previa en Bondora: False - El cliente tiene al menos 3 meses de historial crediticio en Bondora True - El cliente no tiene historial crediticio	Texto	Binaria
CreditScoreEsMicroL	Solicitante	Calificación	Una puntuación que está específicamente diseñada para clasificar los riesgos de los prestatarios de alto riesgo (definidos por Equifax como prestatarios que no tienen acceso a préstamos bancarios); una medida de la probabilidad de incumplimiento un mes antes; la puntuación se da en una escala de 10 grados, de la mejor puntuación a la peor: M1, M2, M3, M4, M5, M6, M7, M8, M9, M10.	Texto	Categórica
CreditScoreEsEquifaxRisk	Solicitante	Calificación	Puntuación genérica para los solicitantes de préstamos que no tienen operaciones activas pendientes en ASNEF; una medida de la probabilidad de incumplimiento un año más adelante; la puntuación se da en una escala de 6 grados: AAA ("Muy bajo") AA ("Bajo") A ("Promedio") B ("Promedio alto") C ("Alto") D ("Muy alto")	Texto	Categórica
CreditScoreFiAsiakasTietoRiskGrade	Solicitante	Calificación	Modelo de puntuación crediticia para Asiakastieto finlandés: RL1 Riesgo muy bajo 01-20 RL2 Riesgo bajo 21-40 RL3 Riesgo promedio 41-60 RL4 Riesgo grande 61-80 RL5 Riesgo enorme 81-100	Texto	Categórica

CreditScoreEeMini	Solicitante	Calificación	1000 No hay problemas de pagos anteriores 900 problemas de pagos terminaron hace 24-36 meses 800 problemas de pagos terminaron hace 12-24 meses 700 problemas de pagos terminaron hace 6-12 meses 600 problemas de pago terminados hace menos de 6 meses 500 problemas de pago activos	Número	Categórica
LoanId	Solicitud		ID único asignado a las solicitudes de préstamo	Texto	Categórica
ListedOnUTC	Solicitud		Fecha en que apareció la solicitud de préstamo en el mercado primario	Fecha	Continua
BiddingStartedOn	Solicitud		Inicio de la puja	Fecha	Continua
BidsPortfolioManager	Solicitud		El monto de las ofertas de inversión realizadas por los administradores de carteras	Número	Continua
BidsApi	Solicitud		La cantidad de ofertas de inversión realizadas a través de Api	Número	Continua
BidsManual	Solicitud		La cantidad de ofertas de inversión hechas manualmente	Número	Continua
LoanApplicationStartedDate	Solicitud		Fecha en la que se inició la solicitud del préstamo	Fecha	Continua
ApplicationSignedHour	Solicitud		Hora de la firma de la solicitud de préstamo	Número	Continua
ApplicationSignedWeekday	Solicitud		Día de la semana de la firma de la solicitud de préstamo	Número	Categórica
VerificationType	Solicitud		Método utilizado para la verificación de los datos de solicitud de préstamo: 1 Ingresos no verificados 2 Ingresos no verificados, con referencias telefónicas cruzadas 3 Ingresos verificados 4 Ingresos y gastos verificados	Número	Categórica
AppliedAmount	Solicitud		Importe solicitado por el prestatario originalmente	Número	Continua
Interest	Solicitud		Tipo de interés máximo aceptado en la solicitud de préstamo	Número	Continua
ReportAsOfEOD			Fecha de generación del informe	Fecha	Continua

Anexo II. Tabla comportamiento de la mora en 2014 y 2016

2014												
Tiempo en vigor del préstamo (en meses)	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
1	0.38%	1.00%	0.69%	2.14%	2.67%	3.16%	4.75%	3.95%	3.67%	3.86%	2.93%	2.94%
2	1.70%	1.75%	**3.20%**	**3.21%**	**6.67%**	**6.99%**	**8.90%**	**7.14%**	**6.30%**	**8.55%**	**5.73%**	**5.14%**
3	0.95%	0.75%	1.83%	1.28%	0.57%	2.66%	3.68%	1.91%	2.10%	1.33%	1.40%	1.91%
4	1.34%	0.51%	0.70%	0.43%	3.29%	5.25%	2.84%	2.48%	1.33%	3.00%	2.45%	1.91%
5	0.57%	0.26%	1.40%	2.60%	3.68%	4.23%	2.21%	2.61%	2.27%	3.12%	3.09%	1.62%
6	0.76%	0.51%	2.33%	3.04%	3.10%	2.20%	3.95%	2.22%	2.27%	2.50%	1.67%	1.62%
7	0.39%	**2.56%**	2.56%	2.39%	2.52%	1.86%	2.21%	2.22%	3.47%	2.25%	0.91%	1.33%
8	1.57%	1.28%	1.86%	1.74%	3.29%	1.35%	1.26%	1.83%	1.87%	1.62%	1.56%	0.89%
9	1.57%	1.53%	2.33%	2.60%	1.74%	2.71%	0.95%	0.78%	0.93%	1.75%	1.56%	1.92%
10	0.40%	1.31%	1.19%	1.56%	1.00%	1.59%	1.17%	1.50%	1.94%	1.83%	0.78%	1.19%
11	0.99%	0.26%	0.95%	0.67%	1.20%	0.88%	0.84%	1.78%	1.52%	0.52%	1.57%	0.89%
12	1.19%	1.05%	0.95%	0.89%	1.00%	0.71%	0.84%	0.82%	0.41%	0.52%	1.83%	4.00%
13	1.08%	0.56%	1.03%	0.49%	1.12%	1.18%	0.75%	0.60%	1.08%	1.33%	2.38%	2.09%
14	0.87%	1.12%	1.03%	0.74%	1.34%	0.59%	0.56%	0.75%	1.08%	2.22%	1.19%	1.04%
15	0.65%	1.40%	0.51%	0.74%	0.67%	1.38%	0.56%	1.21%	2.46%	1.19%	1.19%	1.34%
16	0.22%	0.28%	1.03%	0.98%		0.79%	1.68%	4.22%	2.00%	1.78%	1.32%	1.94%
17	0.87%	1.68%	0.77%	0.74%	0.45%		2.43%	1.36%	0.31%	1.63%	0.93%	1.79%
18	1.08%	0.28%	1.54%	1.23%	1.34%	1.96%	1.31%	0.30%	1.38%	0.74%	1.59%	1.49%
19	**1.91%**	1.84%	1.08%	0.52%	1.72%	1.89%	1.24%	1.48%	1.51%	0.67%	0.80%	0.30%
20	0.48%	0.31%	0.27%	1.82%	1.97%	0.84%	0.62%	2.14%	1.01%	1.00%	1.47%	0.60%
21	1.20%	0.61%	1.89%	2.08%	0.25%	1.68%	0.62%	1.15%	1.34%	2.66%	1.74%	0.15%
22	0.96%	1.84%	1.89%	1.30%	1.23%	1.05%	0.83%	1.32%	1.51%	0.50%	0.67%	0.45%
23	1.44%	1.84%	0.27%	1.30%	0.74%	1.05%	0.62%	0.99%	1.01%	1.50%	1.07%	1.35%
24	1.67%	1.23%	1.62%	0.26%	0.49%	0.63%	1.24%	1.65%	1.01%	1.16%	0.80%	0.45%
25	1.90%	0.82%	0.36%	2.32%	1.41%	1.30%	0.24%	0.60%	0.40%	0.78%	0.82%	0.75%
26	0.63%	1.23%	1.43%	0.39%	1.06%	1.30%	0.97%	0.60%	0.80%	0.39%	0.14%	1.06%
27	0.63%	0.41%	1.43%	1.93%		0.97%	0.48%	0.80%	0.40%	0.58%	0.55%	0.45%
28	0.32%	0.41%	1.08%	1.16%	1.41%	0.65%	0.97%	1.20%	0.40%	1.16%	0.41%	1.21%
29	1.27%	0.82%		0.77%		1.62%	0.24%	0.20%	0.80%	0.58%	0.96%	0.45%
30	0.32%	0.41%	0.72%	0.39%	0.35%	2.27%	1.21%	1.40%	0.60%	0.19%	0.55%	0.45%
31	0.63%	0.41%	0.72%	0.39%	0.71%	0.97%	0.48%	0.60%	0.20%	0.19%	0.41%	0.60%
32	0.32%	1.23%	0.36%	0.39%	0.71%	0.97%	0.48%	0.20%	0.60%	0.19%	0.41%	0.60%
33	0.32%	0.41%		1.16%	0.71%	1.30%	0.48%	0.20%	0.60%	0.39%	0.41%	0.15%
34	1.58%	0.82%	0.36%	0.77%	1.06%	0.97%	0.48%	0.20%	0.20%			0.75%
35	1.27%	0.82%	0.36%			0.65%			0.20%	0.39%	0.55%	0.90%
36	0.32%	0.82%	0.36%	0.39%	0.35%	0.65%	0.48%	0.80%	0.20%	0.00%	0.68%	0.15%
37	0.93%				0.96%		0.36%	0.31%	0.93%	0.82%	0.33%	0.53%
38		0.59%			0.48%			1.56%		0.82%	0.66%	0.18%
39					0.96%	0.46%	0.73%	0.62%	1.54%	0.54%	0.17%	0.18%

40			0.48%	0.52%		0.46%	1.09%		0.62%		0.50%	0.36%
41	0.93%		0.95%	1.05%	0.48%	0.93%		0.31%	0.62%	0.54%	0.66%	0.71%
42								0.62%		1.09%	0.33%	0.36%
43			0.48%		0.48%	0.46%	0.36%	0.93%	1.23%	0.54%		
44	1.40%			0.52%	0.48%	0.46%		0.31%	0.62%			
45			0.48%	1.05%	1.92%	0.93%	0.73%	0.31%	0.62%			
46				1.05%		0.46%	0.36%					
47	0.93%	1.78%				0.93%						
48	0.00%	1.18%	0.48%	2.09%	0.48%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
49	0.76%	1.77%		1.54%	0.74%							
50	0.76%	0.88%										
60	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

2016												
Tiempo en vigor del préstamo (en meses)	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
1	5.32%	4.14%	4.11%	3.53%	3.30%	3.00%	3.04%	4.40%	6.49%	5.07%	5.80%	6.88%
2	**7.64%**	**7.59%**	**6.46%**	**6.57%**	**6.97%**	**6.61%**	**7.28%**	**11.37%**	**9.60%**	**7.21%**	**9.00%**	**12.32%**
3	1.83%	2.34%	2.35%	3.04%	3.42%	3.85%	4.89%	4.67%	3.37%	4.54%	4.24%	4.08%
4	2.36%	2.65%	2.50%	2.86%	2.63%	3.37%	2.30%	3.05%	3.86%	4.54%	4.51%	3.40%
5	1.68%	3.21%	3.10%	2.74%	1.88%	1.69%	2.41%	2.68%	1.60%	2.67%	3.29%	3.32%
6	2.02%	2.09%	2.26%	2.37%	2.00%	2.53%	1.86%	3.05%	2.53%	2.80%	2.86%	3.15%
7	2.07%	2.42%	2.42%	1.95%	3.12%	2.58%	2.65%	2.54%	2.86%	2.95%	3.21%	2.40%
8	2.41%	2.28%	1.09%	1.82%	2.08%	3.19%	2.65%	2.26%	2.45%	1.61%	1.82%	2.23%
9	1.38%	1.85%	1.45%	2.86%	1.69%	2.21%	1.55%	2.26%	1.50%	1.88%	1.82%	1.97%
10	1.57%	1.45%	2.58%	1.62%	2.02%	1.61%	2.02%	2.10%	1.94%	2.15%	2.35%	2.06%
11	1.75%	2.17%	0.98%	1.62%	1.89%	2.24%	1.57%	1.24%	2.49%	1.34%	0.87%	1.46%
12	1.40%	2.03%	1.84%	2.43%	2.29%	2.11%	2.13%	1.33%	1.38%	1.74%	1.39%	2.06%
13	1.63%	1.20%	1.27%	1.13%	1.93%	1.77%	1.88%	1.94%	1.49%	1.90%	1.17%	1.31%
14	1.63%	1.35%	1.27%	1.28%	0.89%	2.09%	1.74%	0.85%	1.99%	1.02%	1.08%	2.18%
15	1.45%	0.75%	1.40%	1.42%	1.93%	1.45%	1.30%	1.82%	1.49%	2.34%	1.35%	1.05%
16	1.27%	2.24%	1.65%	0.57%	1.78%	0.81%	1.45%	1.94%	1.82%	1.90%	1.26%	1.22%
17	0.72%	0.75%	1.53%	1.13%	1.04%	1.61%	1.16%	2.54%	1.00%	0.88%	1.26%	0.87%
18	1.09%	0.75%	1.02%	1.56%	1.33%	1.45%	0.87%	1.21%	1.16%	1.17%	0.72%	0.26%
19	1.30%	1.54%	1.04%	0.60%	1.63%	1.41%	1.44%	1.17%	1.05%	1.23%	0.18%	
20	1.49%	1.54%	1.17%	2.10%	1.63%	1.06%	1.44%	1.30%	0.88%			
21	0.74%	1.23%	0.52%	1.95%	1.63%	1.06%	1.60%	0.91%				
22	0.93%	0.77%	0.91%	0.60%	0.49%	2.12%	0.64%	0.13%				
23	1.12%	0.77%	0.65%	1.20%	0.81%	1.59%						
24	1.12%	0.77%	0.78%	1.65%	0.65%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
25	1.38%	1.46%	0.68%	0.97%								
26	0.20%	0.32%	0.68%	0.16%								
27	0.79%	0.16%	0.14%									
28	1.18%	0.16%										

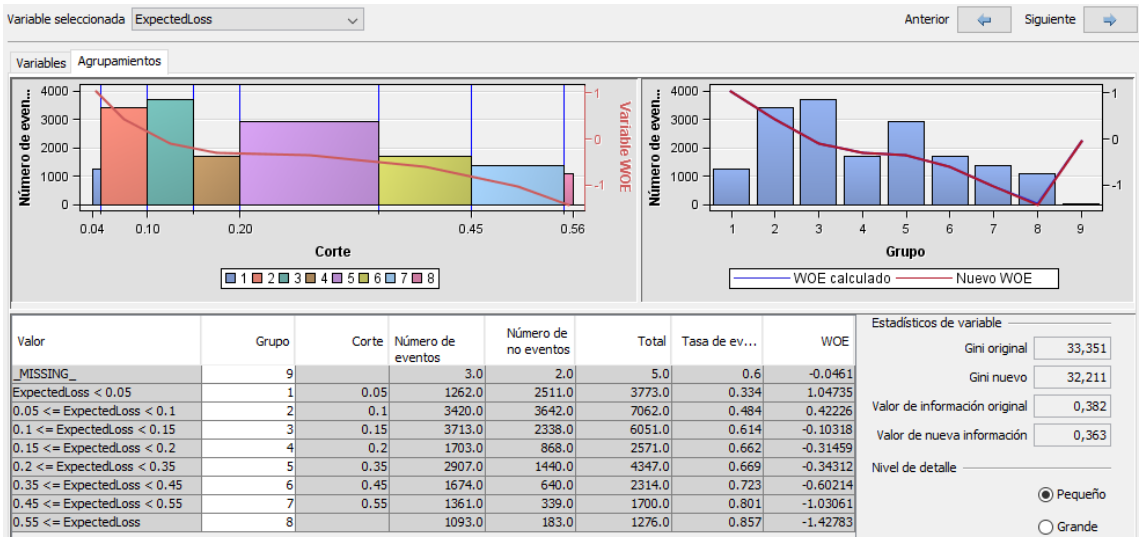
Anexo III. Variables excluidas de la muestra de entrenamiento junto con el motivo de exclusión.

Variable descartada	Motivo
Amount	No se conoce en el momento de la subasta
ActiveLateCategory	No se conoce en el momento de la subasta
ActiveLateLastPaymentCategory	No se conoce en el momento de la subasta
ActiveScheduleFirstPaymentReache	No se conoce en el momento de la subasta
BidsApi	No se conoce en el momento de la subasta
BidsManual	No se conoce en el momento de la subasta
BidsPortfolioManager	No se conoce en el momento de la subasta
City	Imposible homogeneizar
ContractEndDate	Variable identificativa que no aporta información
CurrentDebtDaysPrimary	No se conoce en el momento de la subasta
CurrentDebtDaysSecondary	No se conoce en el momento de la subasta
DebtOccuredOn	No se conoce en el momento de la subasta
DebtOccuredOnForSecondary	No se conoce en el momento de la subasta
DebtToIncome	Dato que ha dejado de aprovisionarse
DefaultDate	No se conoce en el momento de la subasta
EAD1	No se conoce en el momento de la subasta
EAD2	No se conoce en el momento de la subasta
EL_V0	Metrica desfasada, que no utiliza en la actualidad
EL_V1	Metrica desfasada, que no utiliza en la actualidad
EmploymentPosition	Dato que ha dejado de aprovisionarse
EmploymentStatus	Dato que ha dejado de aprovisionarse
FirstPaymentDay	Variable identificativa que no aporta información
FreeCash	Dato que ha dejado de aprovisionarse
GracePeriodStart	No se conoce en el momento de la subasta
GracePeriodEnd	No se conoce en el momento de la subasta
IncomeFromChildSupport	Dato que ha dejado de aprovisionarse
IncomeFromFamilyAllowance	Dato que ha dejado de aprovisionarse
IncomeFromLeavePay	Dato que ha dejado de aprovisionarse
IncomeFromPension	Dato que ha dejado de aprovisionarse
IncomeFromPrincipalEmployer	Dato que ha dejado de aprovisionarse
IncomeFromSocialWelfare	Dato que ha dejado de aprovisionarse
IncomeOther	Dato que ha dejado de aprovisionarse
InterestAndPenalgyBalance	No se conoce en el momento de la subasta
InterestAndPenaltyDebtServicingC	No se conoce en el momento de la subasta
InterestAndPenaltyPaymentsMade	No se conoce en el momento de la subasta
InterestAndPenaltyWriteOffs	No se conoce en el momento de la subasta
InterestRecovery	No se conoce en el momento de la subasta
LastPaymentOn	No se conoce en el momento de la subasta
LoanDate	Variable identificativa que no aporta información

MaritalStatus	Dato que ha dejado de aprovisionarse
MaturityDate_Original	Variable identificativa que no aporta información
MaturityDate_Last	No se conoce en el momento de la subasta
NextPaymentDate	No se conoce en el momento de la subasta
NextPaymentNr	No se conoce en el momento de la subasta
NrOfDependants	Dato que ha dejado de aprovisionarse
NrOfScheduledPayments	No se conoce en el momento de la subasta
OccupationArea	Dato que ha dejado de aprovisionarse
PlannedInterestPostDefault	No se conoce en el momento de la subasta
PlannedInterestTillDate	No se conoce en el momento de la subasta
PlannedPrincipalPostDefault	No se conoce en el momento de la subasta
PlannedPrincipalTillDate	No se conoce en el momento de la subasta
PrincipalBalance	No se conoce en el momento de la subasta
PrincipalDebtServicingCost	No se conoce en el momento de la subasta
PrincipalOverdueBySchedule	No se conoce en el momento de la subasta
PrincipalPaymentsMade	No se conoce en el momento de la subasta
PrincipalRecovery	No se conoce en el momento de la subasta
PrincipalWriteOffs	No se conoce en el momento de la subasta
Rating_V0	Metrica desfasada, que no utiliza en la actualidad
Rating_V1	Metrica desfasada, que no utiliza en la actualidad
Rating_V2	Metrica desfasada, que no utiliza en la actualidad
RecoveryStage	No se conoce en el momento de la subasta
RefinanceLiabilities	No se conoce en el momento de la subasta
ReportAsOfEOD	Indica el día de la descarga de la información
ReScheduledOn	No se conoce en el momento de la subasta
Restructured	No se conoce en el momento de la subasta
StageActiveSince	No se conoce en el momento de la subasta
Status	No se conoce en el momento de la subasta
UseOfLoan	Dato que ha dejado de aprovisionarse
UserName	Variable identificativa que no aporta información
WorkExperience	Dato que ha dejado de aprovisionarse
WorseLateCategory	No se conoce en el momento de la subasta

Anexo IV. Detalle de la tramificación de las variables continuas

Figura 1. Agrupación final ExpectedLoss



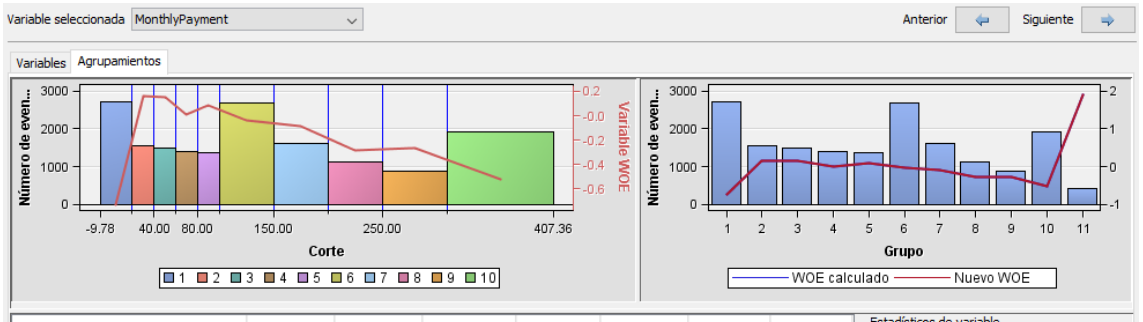
Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	9		3.0	2.0	5.0	0.6	-0.0461
ExpectedLoss < 0.05	1	0.05	1262.0	2511.0	3773.0	0.334	1.04735
0.05 <= ExpectedLoss < 0.1	2	0.1	3420.0	3642.0	7062.0	0.484	0.42226
0.1 <= ExpectedLoss < 0.15	3	0.15	3713.0	2338.0	6051.0	0.614	-0.10318
0.15 <= ExpectedLoss < 0.2	4	0.2	1703.0	868.0	2571.0	0.662	-0.31459
0.2 <= ExpectedLoss < 0.35	5	0.35	2907.0	1440.0	4347.0	0.669	-0.34312
0.35 <= ExpectedLoss < 0.45	6	0.45	1674.0	640.0	2314.0	0.723	-0.60214
0.45 <= ExpectedLoss < 0.55	7	0.55	1361.0	339.0	1700.0	0.801	-1.03061
0.55 <= ExpectedLoss	8		1093.0	183.0	1276.0	0.857	-1.42783

Estadísticos de variable

Gini original	33,351
Gini nuevo	32,211
Valor de información original	0,382
Valor de nueva información	0,363

Nivel de detalle: Pequeño / Grande

Figura 2. Agrupación final MonthlyPayment



Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de e...	WOE
MISSING	11		407.0	1883.0	2290.0	0.178	1.89117
MonthlyPayment < 20	1	20.0	2703.0	914.0	3617.0	0.747	-0.72492
20 <= MonthlyPayment < 40	2	40.0	1556.0	1269.0	2825.0	0.551	0.15547
40 <= MonthlyPayment < 60	3	60.0	1493.0	1208.0	2701.0	0.553	0.14754
60 <= MonthlyPayment < 80	4	80.0	1401.0	985.0	2386.0	0.587	0.00706
80 <= MonthlyPayment < 100	5	100.0	1367.0	1035.0	2402.0	0.569	0.08115
100 <= MonthlyPayment < 150	6	150.0	2686.0	1797.0	4483.0	0.599	-0.04257
150 <= MonthlyPayment < 200	7	200.0	1606.0	1024.0	2630.0	0.611	-0.09067
200 <= MonthlyPayment < 250	8	250.0	1107.0	581.0	1688.0	0.656	-0.2853
250 <= MonthlyPayment < 310	9	310.0	885.0	472.0	1357.0	0.652	-0.26925
310 <= MonthlyPayment	10		1925.0	795.0	2720.0	0.708	-0.52498

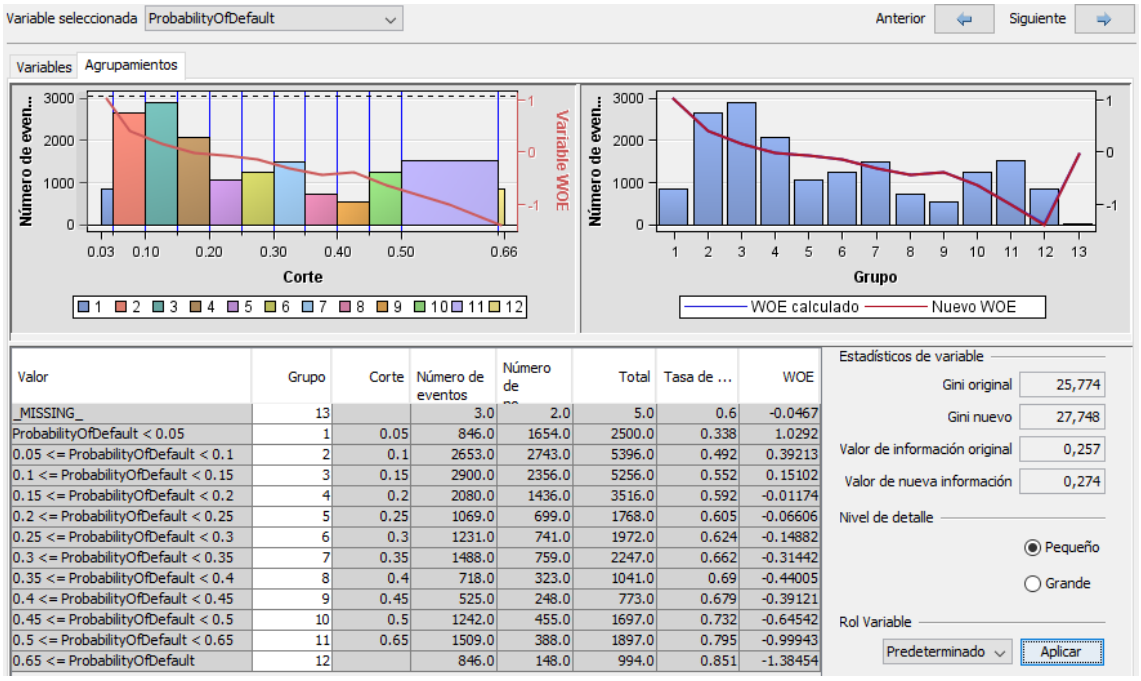
Estadísticos de variable

Gini original	25,549
Gini nuevo	26,162
Valor de información original	0,339
Valor de nueva información	0,350

Nivel de detalle: Pequeño / Grande

Rol Variable

Figura 3. Agrupación final ProbabilityOfDefault



Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ...	WOE
MISSING	13		3.0	2.0	5.0	0.6	-0.0467
ProbabilityOfDefault < 0.05	1	0.05	846.0	1654.0	2500.0	0.338	1.0292
0.05 <= ProbabilityOfDefault < 0.1	2	0.1	2653.0	2743.0	5396.0	0.492	0.39213
0.1 <= ProbabilityOfDefault < 0.15	3	0.15	2900.0	2356.0	5256.0	0.552	0.15102
0.15 <= ProbabilityOfDefault < 0.2	4	0.2	2080.0	1436.0	3516.0	0.592	-0.01174
0.2 <= ProbabilityOfDefault < 0.25	5	0.25	1069.0	699.0	1768.0	0.605	-0.06606
0.25 <= ProbabilityOfDefault < 0.3	6	0.3	1231.0	741.0	1972.0	0.624	-0.14882
0.3 <= ProbabilityOfDefault < 0.35	7	0.35	1488.0	759.0	2247.0	0.662	-0.31442
0.35 <= ProbabilityOfDefault < 0.4	8	0.4	718.0	323.0	1041.0	0.69	-0.44005
0.4 <= ProbabilityOfDefault < 0.45	9	0.45	525.0	248.0	773.0	0.679	-0.39121
0.45 <= ProbabilityOfDefault < 0.5	10	0.5	1242.0	455.0	1697.0	0.732	-0.64542
0.5 <= ProbabilityOfDefault < 0.65	11	0.65	1509.0	388.0	1897.0	0.795	-0.99943
0.65 <= ProbabilityOfDefault	12		846.0	148.0	994.0	0.851	-1.38454

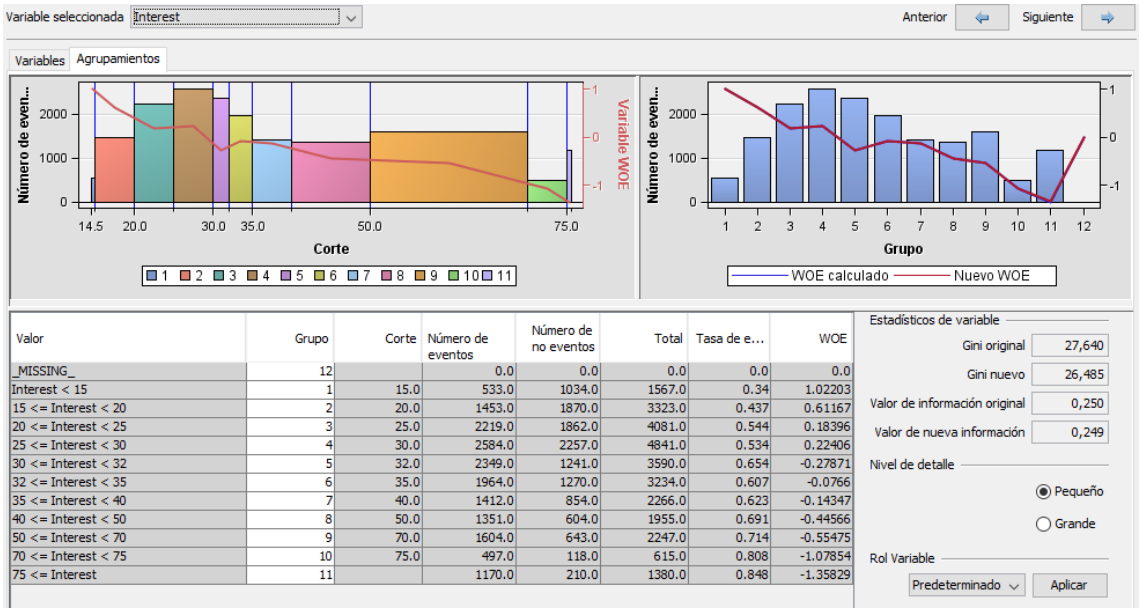
Estadísticos de variable

Gini original	25,774
Gini nuevo	27,748
Valor de información original	0,257
Valor de nueva información	0,274

Nivel de detalle: Pequeño / Grande

Rol Variable: Predeterminado — Aplicar

Figura 4. Agrupación final Interest



Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de e...	WOE
MISSING	12		0.0	0.0	0.0	0.0	0.0
Interest < 15	1	15.0	533.0	1034.0	1567.0	0.34	1.02203
15 <= Interest < 20	2	20.0	1453.0	1870.0	3323.0	0.437	0.61167
20 <= Interest < 25	3	25.0	2219.0	1862.0	4081.0	0.544	0.18396
25 <= Interest < 30	4	30.0	2584.0	2257.0	4841.0	0.534	0.22406
30 <= Interest < 32	5	32.0	2349.0	1241.0	3590.0	0.654	-0.27871
32 <= Interest < 35	6	35.0	1964.0	1270.0	3234.0	0.607	-0.0766
35 <= Interest < 40	7	40.0	1412.0	854.0	2266.0	0.623	-0.14347
40 <= Interest < 50	8	50.0	1351.0	604.0	1955.0	0.691	-0.44566
50 <= Interest < 70	9	70.0	1604.0	643.0	2247.0	0.714	-0.55475
70 <= Interest < 75	10	75.0	497.0	118.0	615.0	0.808	-1.07854
75 <= Interest	11		1170.0	210.0	1380.0	0.848	-1.35829

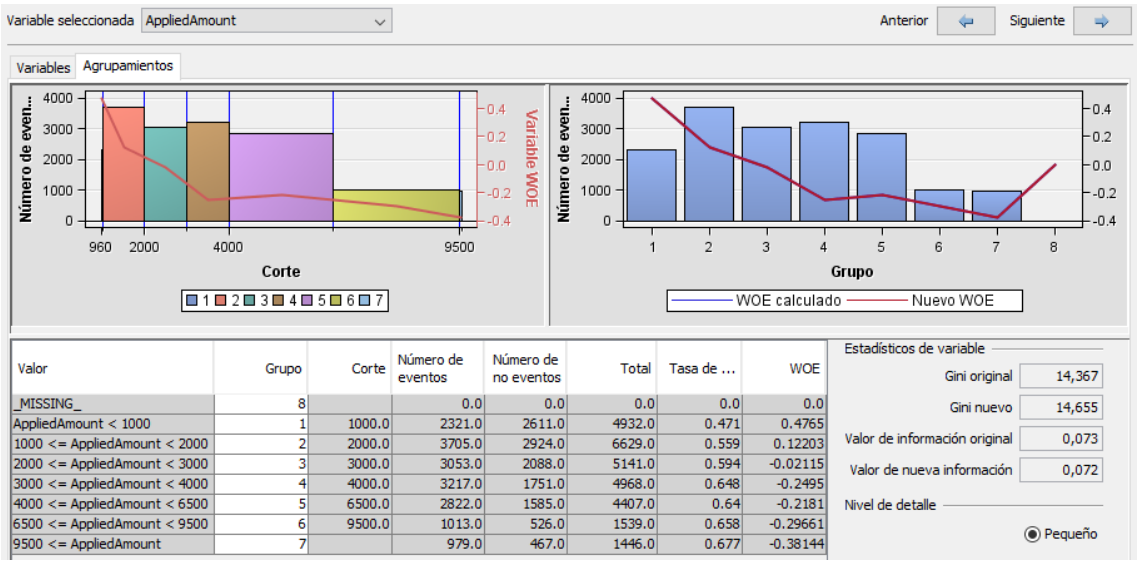
Estadísticos de variable

Gini original	27,640
Gini nuevo	26,485
Valor de información original	0,250
Valor de nueva información	0,249

Nivel de detalle: Pequeño / Grande

Rol Variable: Predeterminado — Aplicar

Figura 5. Agrupación final AppliedAmount



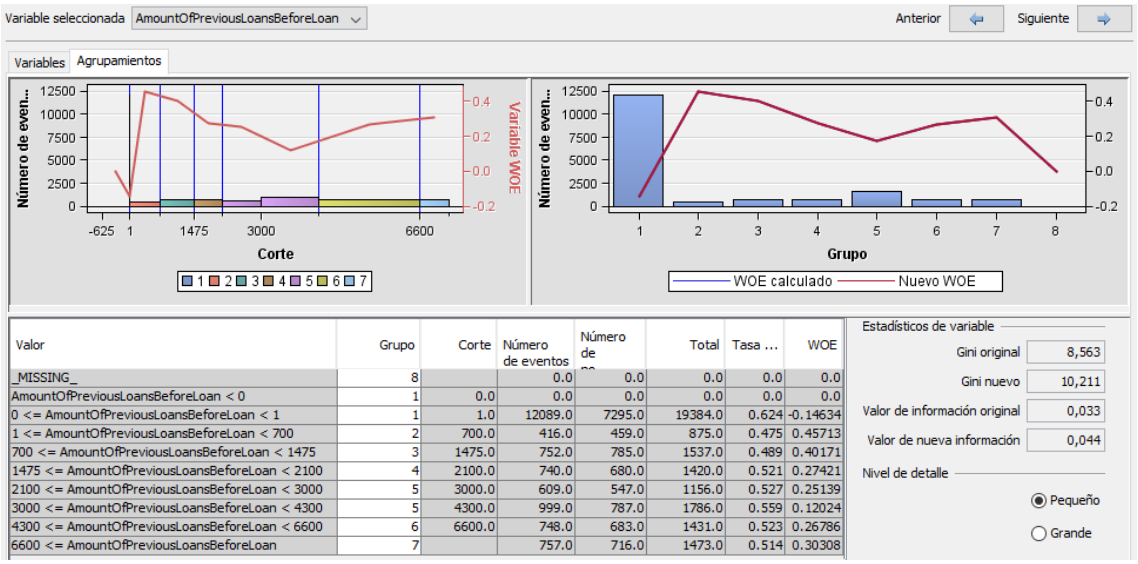
Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ...	WOE
MISSING	8		0.0	0.0	0.0	0.0	0.0
AppliedAmount < 1000	1	1000.0	2321.0	2611.0	4932.0	0.471	0.4765
1000 <= AppliedAmount < 2000	2	2000.0	3705.0	2924.0	6629.0	0.559	0.12203
2000 <= AppliedAmount < 3000	3	3000.0	3053.0	2088.0	5141.0	0.594	-0.02115
3000 <= AppliedAmount < 4000	4	4000.0	3217.0	1751.0	4968.0	0.648	-0.2495
4000 <= AppliedAmount < 6500	5	6500.0	2822.0	1585.0	4407.0	0.64	-0.2181
6500 <= AppliedAmount < 9500	6	9500.0	1013.0	526.0	1539.0	0.658	-0.29661
9500 <= AppliedAmount	7		979.0	467.0	1446.0	0.677	-0.38144

Estadísticos de variable

Gini original	14,367
Gini nuevo	14,655
Valor de información original	0,073
Valor de nueva información	0,072

Nivel de detalle: Pequeño

Figura 6. Agrupación final AmountOfPreviousLoansBeforeLoan



Valor	Grupo	Corte	Número de eventos	Número de	Total	Tasa ...	WOE
MISSING	8		0.0	0.0	0.0	0.0	0.0
AmountOfPreviousLoansBeforeLoan < 0	1	0.0	0.0	0.0	0.0	0.0	0.0
0 <= AmountOfPreviousLoansBeforeLoan < 1	1	1.0	12089.0	7295.0	19384.0	0.624	-0.14634
1 <= AmountOfPreviousLoansBeforeLoan < 700	2	700.0	416.0	459.0	875.0	0.475	0.45713
700 <= AmountOfPreviousLoansBeforeLoan < 1475	3	1475.0	752.0	785.0	1537.0	0.489	0.40171
1475 <= AmountOfPreviousLoansBeforeLoan < 2100	4	2100.0	740.0	680.0	1420.0	0.521	0.27421
2100 <= AmountOfPreviousLoansBeforeLoan < 3000	5	3000.0	609.0	547.0	1156.0	0.527	0.25139
3000 <= AmountOfPreviousLoansBeforeLoan < 4300	5	4300.0	999.0	787.0	1786.0	0.559	0.12024
4300 <= AmountOfPreviousLoansBeforeLoan < 6600	6	6600.0	748.0	683.0	1431.0	0.523	0.26786
6600 <= AmountOfPreviousLoansBeforeLoan	7		757.0	716.0	1473.0	0.514	0.30308

Estadísticos de variable

Gini original	8,563
Gini nuevo	10,211
Valor de información original	0,033
Valor de nueva información	0,044

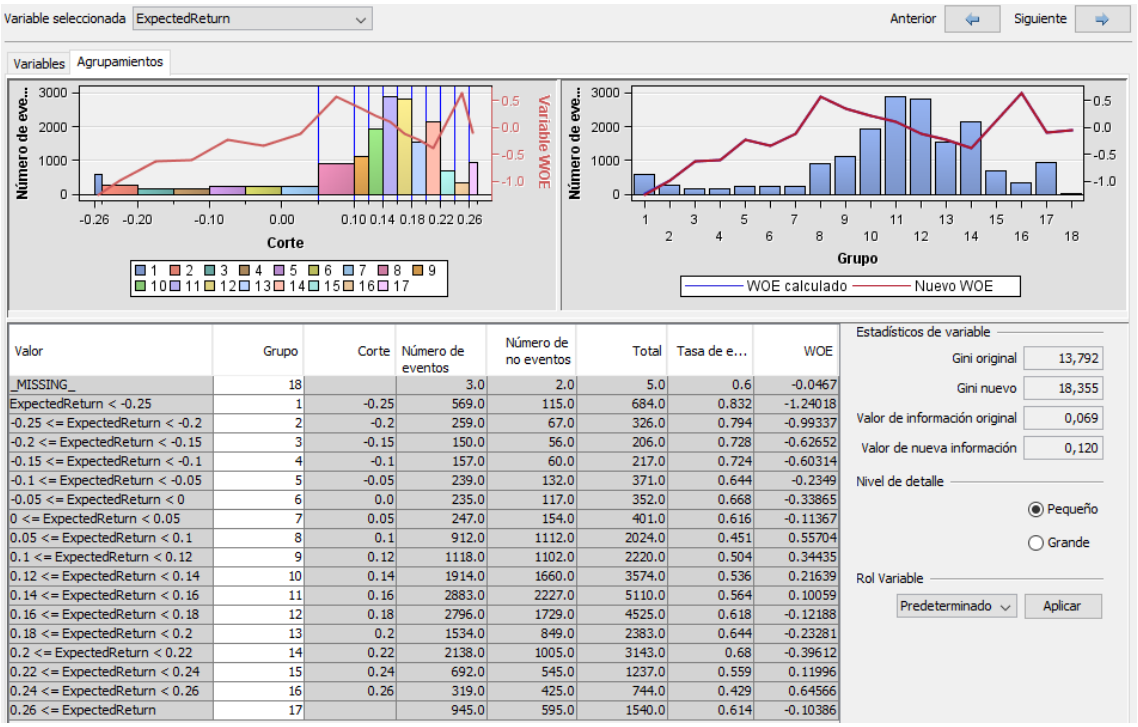
Nivel de detalle: Pequeño / Grande

Figura 7. Agrupación final PorFreeCash

Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	13		0.0	0.0	0.0	0.0	0.0
PorFreeCash < 0.3	1	0.3	808.0	483.0	1291.0	0.626	-0.15578
0.3 <= PorFreeCash < 0.35	2	0.35	2175.0	1147.0	3322.0	0.655	-0.28111
0.35 <= PorFreeCash < 0.4	3	0.4	1394.0	842.0	2236.0	0.623	-0.14539
0.4 <= PorFreeCash < 0.45	4	0.45	1395.0	946.0	2341.0	0.596	-0.02964
0.45 <= PorFreeCash < 0.5	5	0.5	1286.0	958.0	2244.0	0.573	0.06432
0.5 <= PorFreeCash < 0.6	6	0.6	2538.0	2051.0	4589.0	0.553	0.14572
0.6 <= PorFreeCash < 0.7	7	0.7	2051.0	1739.0	3790.0	0.541	0.19375
0.7 <= PorFreeCash < 0.8	8	0.8	1519.0	1285.0	2804.0	0.542	0.19147
0.8 <= PorFreeCash < 0.9	9	0.9	1085.0	851.0	1936.0	0.56	0.11584
0.9 <= PorFreeCash < 1	10	1.0	841.0	558.0	1399.0	0.601	-0.05147
1 <= PorFreeCash < 1.2	11	1.2	982.0	618.0	1600.0	0.614	-0.10434
1.2 <= PorFreeCash	12		1036.0	474.0	1510.0	0.686	-0.42315

Estadísticos de variable
Gini original 9,224
Gini nuevo 10,176
Valor de información original 0,030
Valor de nueva información 0,034
Nivel de detalle
Pequeño
Grande
Rol Variable
Predeterminado
Aplicar

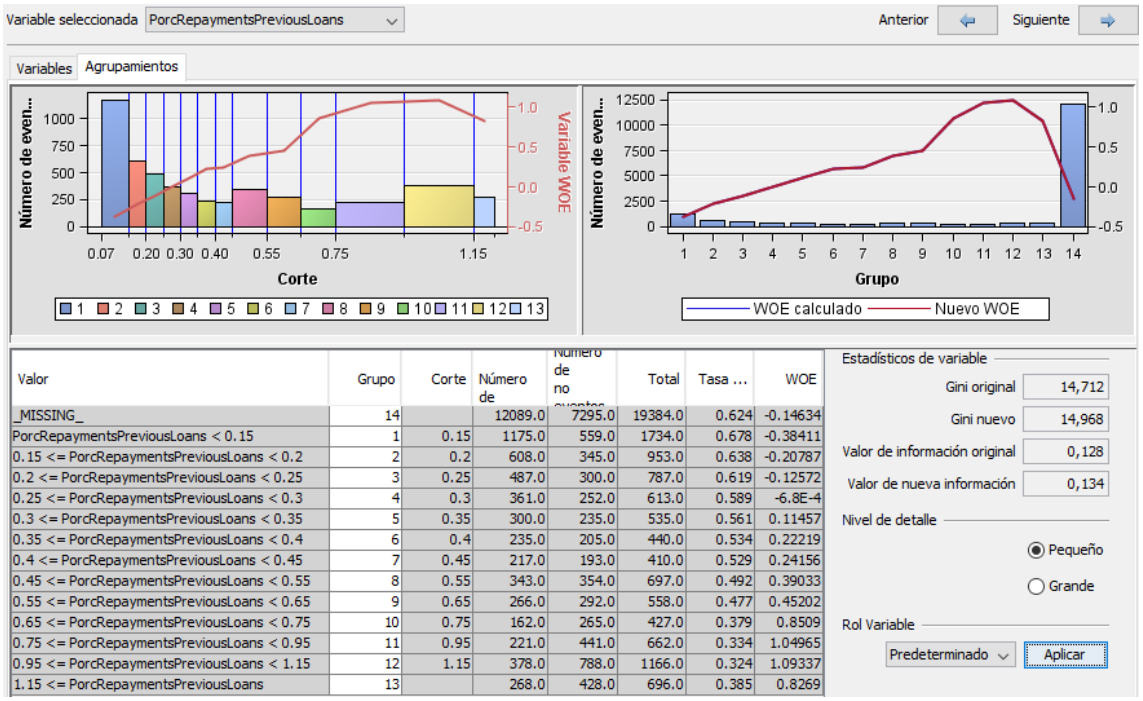
Figura 8. Agrupación final ExpectedReturn



Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de e...	WOE
MISSING	18		3.0	2.0	5.0	0.6	-0.0467
ExpectedReturn < -0.25	1	-0.25	569.0	115.0	684.0	0.832	-1.24018
-0.25 <= ExpectedReturn < -0.2	2	-0.2	259.0	67.0	326.0	0.794	-0.99337
-0.2 <= ExpectedReturn < -0.15	3	-0.15	150.0	56.0	206.0	0.728	-0.62652
-0.15 <= ExpectedReturn < -0.1	4	-0.1	157.0	60.0	217.0	0.724	-0.60314
-0.1 <= ExpectedReturn < -0.05	5	-0.05	239.0	132.0	371.0	0.644	-0.2349
-0.05 <= ExpectedReturn < 0	6	0.0	235.0	117.0	352.0	0.668	-0.33865
0 <= ExpectedReturn < 0.05	7	0.05	247.0	154.0	401.0	0.616	-0.11367
0.05 <= ExpectedReturn < 0.1	8	0.1	912.0	1112.0	2024.0	0.451	0.55704
0.1 <= ExpectedReturn < 0.12	9	0.12	1118.0	1102.0	2220.0	0.504	0.34435
0.12 <= ExpectedReturn < 0.14	10	0.14	1914.0	1660.0	3574.0	0.536	0.21639
0.14 <= ExpectedReturn < 0.16	11	0.16	2883.0	2227.0	5110.0	0.564	0.10059
0.16 <= ExpectedReturn < 0.18	12	0.18	2796.0	1729.0	4525.0	0.618	-0.12188
0.18 <= ExpectedReturn < 0.2	13	0.2	1534.0	849.0	2383.0	0.644	-0.23281
0.2 <= ExpectedReturn < 0.22	14	0.22	2138.0	1005.0	3143.0	0.68	-0.39612
0.22 <= ExpectedReturn < 0.24	15	0.24	692.0	545.0	1237.0	0.559	0.11996
0.24 <= ExpectedReturn < 0.26	16	0.26	319.0	425.0	744.0	0.429	0.64566
0.26 <= ExpectedReturn	17		945.0	595.0	1540.0	0.614	-0.10386

Estadísticos de variable
Gini original 13,792
Gini nuevo 18,355
Valor de información original 0,069
Valor de nueva información 0,120
Nivel de detalle
Pequeño
Grande
Rol Variable
Predeterminado
Aplicar

Figura 9. Agrupación final PorcRepaymentsPreviousLoans



Valor	Grupo	Corte	Número de	Número de no cuentas	Total	Tasa ...	WOE
MISSING	14		12089.0	7295.0	19384.0	0.624	-0.14634
PorcRepaymentsPreviousLoans < 0.15	1	0.15	1175.0	559.0	1734.0	0.678	-0.38411
0.15 <= PorcRepaymentsPreviousLoans < 0.2	2	0.2	608.0	345.0	953.0	0.638	-0.20787
0.2 <= PorcRepaymentsPreviousLoans < 0.25	3	0.25	487.0	300.0	787.0	0.619	-0.12572
0.25 <= PorcRepaymentsPreviousLoans < 0.3	4	0.3	361.0	252.0	613.0	0.589	-6.8E-4
0.3 <= PorcRepaymentsPreviousLoans < 0.35	5	0.35	300.0	235.0	535.0	0.561	0.11457
0.35 <= PorcRepaymentsPreviousLoans < 0.4	6	0.4	235.0	205.0	440.0	0.534	0.22219
0.4 <= PorcRepaymentsPreviousLoans < 0.45	7	0.45	217.0	193.0	410.0	0.529	0.24156
0.45 <= PorcRepaymentsPreviousLoans < 0.55	8	0.55	343.0	354.0	697.0	0.492	0.39033
0.55 <= PorcRepaymentsPreviousLoans < 0.65	9	0.65	266.0	292.0	558.0	0.477	0.45202
0.65 <= PorcRepaymentsPreviousLoans < 0.75	10	0.75	162.0	265.0	427.0	0.379	0.8509
0.75 <= PorcRepaymentsPreviousLoans < 0.95	11	0.95	221.0	441.0	662.0	0.334	1.04965
0.95 <= PorcRepaymentsPreviousLoans < 1.15	12	1.15	378.0	788.0	1166.0	0.324	1.09337
1.15 <= PorcRepaymentsPreviousLoans	13		268.0	428.0	696.0	0.385	0.8269

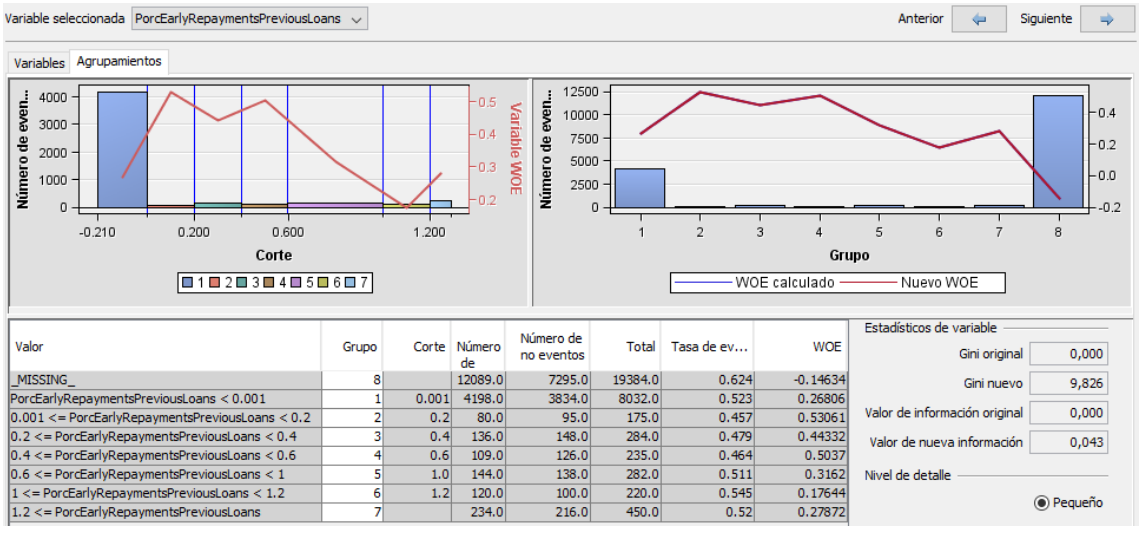
Estadísticos de variable

Gini original	14,712
Gini nuevo	14,968
Valor de información original	0,128
Valor de nueva información	0,134

Nivel de detalle: Pequeño / Grande

Rol Variable: Predeterminado — Aplicar

Figura 10. Agrupación final PorcEarlyRepaymentsPreviousLoans



Valor	Grupo	Corte	Número de	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	8		12089.0	7295.0	19384.0	0.624	-0.14634
PorcEarlyRepaymentsPreviousLoans < 0.001	1	0.001	4198.0	3834.0	8032.0	0.523	0.26806
0.001 <= PorcEarlyRepaymentsPreviousLoans < 0.2	2	0.2	80.0	95.0	175.0	0.457	0.53061
0.2 <= PorcEarlyRepaymentsPreviousLoans < 0.4	3	0.4	136.0	148.0	284.0	0.479	0.44332
0.4 <= PorcEarlyRepaymentsPreviousLoans < 0.6	4	0.6	109.0	126.0	235.0	0.464	0.5037
0.6 <= PorcEarlyRepaymentsPreviousLoans < 1	5	1.0	144.0	138.0	282.0	0.511	0.3162
1 <= PorcEarlyRepaymentsPreviousLoans < 1.2	6	1.2	120.0	100.0	220.0	0.545	0.17644
1.2 <= PorcEarlyRepaymentsPreviousLoans	7		234.0	216.0	450.0	0.52	0.27872

Estadísticos de variable

Gini original	0,000
Gini nuevo	9,826
Valor de información original	0,000
Valor de nueva información	0,043

Nivel de detalle: Pequeño

Anexo V. Código creación datasets

```
/****** Creación sets de datos test *******/

/*Se utiliza para la creación de los sets de datos de las dos muestras, entrenamiento y tesT.
Se cambia el nombre de los archivos y la fecha de selección de los préstamos.

/*1*/
/*Al set de datos LoanData se le añade la definición de préstamo bueno y malo
en la variable Impago. Se eliminarían los posibles préstamos indeterminados,
pero por la definición, no hay*/


data tfm.loandata_v1;
set tfm.loandata;
if ActiveLateCategory='' and DebtOccuredOn='' and DebtOccuredOnForSecondary='' then Impago=0;
else if ActiveLateCategory='' and DebtOccuredOn ='' and DebtOccuredOnForSecondary^='' then Impago=1;
```

```
else if ActiveLateCategory^='' or DebtOccuredOn^='' then Impago=1;
else delete;
run;

/*2*/
/*Se selecciona para la muestra la ventana temporal definida:
Los préstamos contratados desde el 1 de abril de 2017
al 30 de junio de 2017*/
data tfm.testmuestra1;
set tfm.loandata_v1;
where datepart(LoanApplicationStartedDate)>='01APR17'd and
datepart(LoanApplicationStartedDate)<='30JUN17'd;
run;

/*3*/
/* Se hace la limpieza de datos. Se eliminan de la muestra:
- las variables no conocidas en el momento de solicitud del préstamo ( indicadas en anexo 1)
- las variables que se descartan del modelo actualmente
- la variable de fecha de la extracción
- la variable City porque es imposible de homogeneizar
- préstamos cuyo solicitante es menor de 18 : Age <18*/

data tfm.testmuestra2;
set tfm.testmuestra1;
drop Amount
        ActiveLateCategory
        ActiveLateLastPaymentCategory
        ActiveScheduleFirstPaymentReache
        BidsPortfolioManager
        BidsApi
        BidsManual
        City
        ContractEndDate
        CurrentDebtDaysPrimary
        CurrentDebtDaysSecondary
        DateOfBirth
        DebtOccuredOn
        DebtOccuredOnForSecondary
        DebtToIncome
        DefaultDate
        EAD1
        EAD2
        EL_V0
        EL_V1
        EmploymentPosition
        EmploymentStatus
        FirstPaymentDate
        FreeCash
        GracePeriodStart
        GracePeriodEnd
        IncomeFromChildSupport
        IncomeFromFamilyAllowance
        IncomeFromLeavePay
        IncomeFromPension
        IncomeFromPrincipalEmployer
        IncomeFromSocialWelfare
        IncomeOther
        InterestAndPenaltyBalance
        InterestAndPenaltyDebtServicingC
        InterestAndPenaltyPaymentsMade
        InterestAndPenaltyWriteOffs
        InterestRecovery
        LastPaymentOn
        LoanDate
        MaritalStatus
        MaturityDate_Last
        MaturityDate_Original
        NextPaymentDate
        NextPaymentNr
        NrOfDependants
        NrOfScheduledPayments
        OccupationArea
        PlannedInterestPostDefault
        PlannedInterestTillDate
        PlannedPrincipalPostDefault
        PlannedPrincipalTillDate
        PrincipalBalance
        PrincipalDebtServicingCost
        PrincipalOverdueBySchedule
        PrincipalPaymentsMade
        PrincipalRecovery
        PrincipalWriteOffs
        Rating_V0
        Rating_V1
        Rating_V2
        RecoveryStage
        RefinanceLiabilities
        ReportAsOfEOD
```

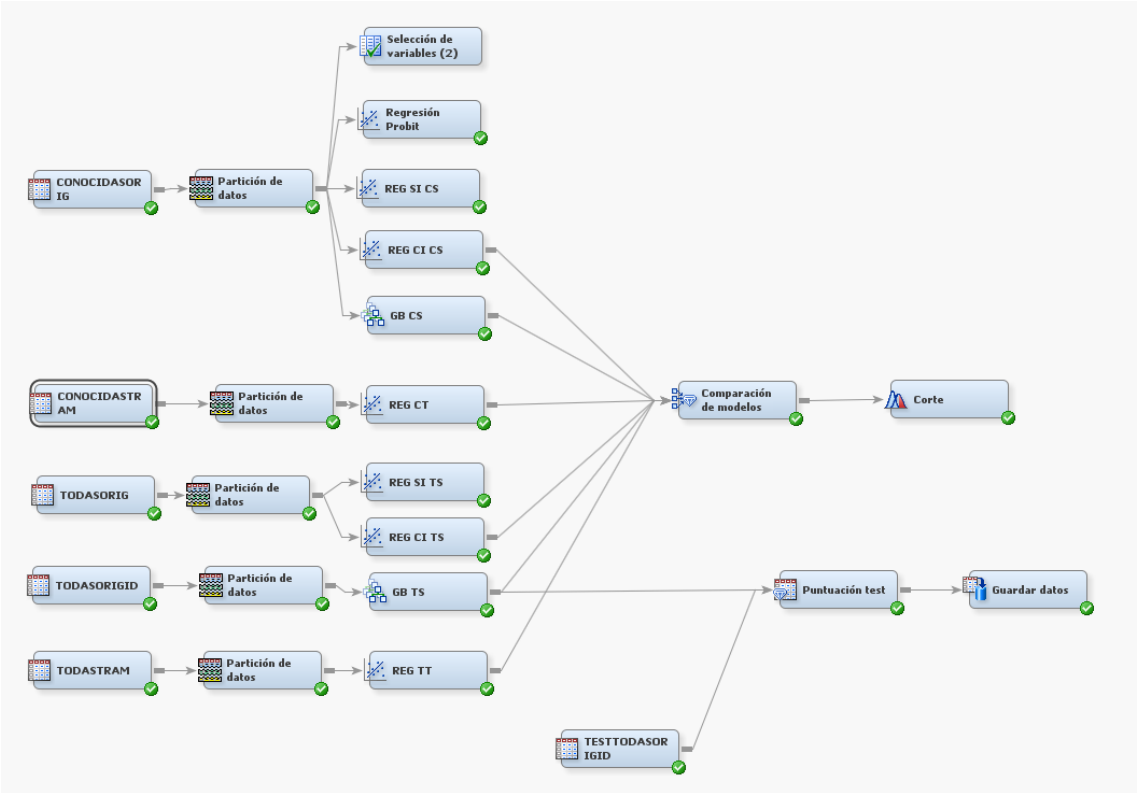
```
        ReScheduledOn
        Restructured
        StageActiveSince
        Status
        UseOfLoan
        WorkExperience
        WorseLateCategory;
if Age<18 then delete;
run;

/*4*/
/*Se crean nuevas variables con las transformaciones y se crean las nuevas variables porcentajes*/
data tfm.testmuestra3;
set tfm.testmuestra2;
TransAge=sqrt(Age);
TransIncomeTotal=log(IncomeTotal);
TransLiabilitiesTotal=log(LiabilitiesTotal);
TransInterest=log(Interest);
TransExistingLiabilites=log(ExistingLiabilities);
TransNoOfPreviousLoansBeforeLoan=log(NoOfPreviousLoansBeforeLoan);
PorFreeCash=LiabilitiesTotal/IncomeTotal;
PorcRepaymentsPreviousLoans=PreviousRepaymentsBeforeLoan/AmountOfPreviousLoansBeforeLoan;
PorcEarlyRepaymentsPreviousLoans=PreviousEarlyRepaymentsBefoleLoa/AmountOfPreviousLoansBeforeLoan;
run;

/*5*/
/*Se tratan los datos de County que ya se han homogeneizado en las provincias/comunidades autónomas
de cada país. Se homogeneizan las categorías de EmploymentDurationCurrentEmployee. Se eliminan
los outlier de IncomeTotal y LiabilitiesTotal que no están verificados*/
proc sql;
create table tfm.testmuestra4 as
select distinct a.*, b.County2
from tfm.testmuestra3 a
left join tfm.regiones b
on a.County=b.County;
quit;
data tfm.testmuestra4;
set tfm.testmuestra4;
if EmploymentDurationCurrentEmploye='MoreThan5Years' then
EmploymentDurationCurrentEmploye='UpTo5Years';
if IncomeTotal > 7000 and VerificationType not in (3,4) then delete;
if LiabilitiesTotal > 5000 and VerificationType^=4 then delete;
run;


/*6*/
/* Tramificación/agrupación de variables continuas y categóricas*/
proc sql;
create table tfm.testmuestra5 as
select *,
        case when MonthlyPayment is null then 'missing'
        when MonthlyPayment <20 then '<20'
        when MonthlyPayment >=20 and MonthlyPayment <40 then '20-40'
        when MonthlyPayment >=40 and MonthlyPayment <60 then '40-60'
        when MonthlyPayment >=60 and MonthlyPayment <80 then '60-80'
        when MonthlyPayment >=80 and MonthlyPayment <100 then '80-100'
        when MonthlyPayment >=100 and MonthlyPayment <150 then '100-150'
        when MonthlyPayment >=150 and MonthlyPayment <200 then '150-200'
        when MonthlyPayment >=200 and MonthlyPayment <250 then '200-250'
        when MonthlyPayment >=250 and MonthlyPayment <310 then '250-310'
        else '>310'
        end as TMonthlyPayment,
        case when ExpectedLoss is null then 'missing'
        when ExpectedLoss <0.05 then '<0.05'
        when ExpectedLoss >=0.05 and ExpectedLoss <0.1 then '0.05-0.10'
        when ExpectedLoss >=0.1 and ExpectedLoss <0.15 then '0.10-0.15'
        when ExpectedLoss >=0.15 and ExpectedLoss <0.20 then '0.15-0.20'
        when ExpectedLoss >=0.20 and ExpectedLoss <0.25 then '0.20-0.25'
        when ExpectedLoss >=0.20 and ExpectedLoss <0.35 then '0.25-0.35'
        when ExpectedLoss >=0.35 and ExpectedLoss <0.45 then '0.35-0.45'
        when ExpectedLoss >=0.45 and ExpectedLoss <0.55 then '0.45-0.55'
        else '>0.55'
        end as TExpectedLoss,
        case when ProbabilityOfDefault is null then 'missing'
        when ProbabilityOfDefault <0.05 then '<0.05'
        when ProbabilityOfDefault >=0.05 and ProbabilityOfDefault <0.10 then '0.05-0.10'
        when ProbabilityOfDefault >=0.10 and ProbabilityOfDefault <0.15 then '0.10-0.15'
        when ProbabilityOfDefault >=0.15 and ProbabilityOfDefault <0.20 then '0.15-0.20'
        when ProbabilityOfDefault >=0.20 and ProbabilityOfDefault <0.25 then '0.20-0.25'
        when ProbabilityOfDefault >=0.25 and ProbabilityOfDefault <0.30 then '0.25-0.30'
        when ProbabilityOfDefault >=0.30 and ProbabilityOfDefault <0.35 then '0.30-0.35'
        when ProbabilityOfDefault >=0.35 and ProbabilityOfDefault <0.40 then '0.35-0.40'
        when ProbabilityOfDefault >=0.40 and ProbabilityOfDefault <0.45 then '0.40-0.45'
        when ProbabilityOfDefault >=0.45 and ProbabilityOfDefault <0.50 then '0.45-0.50'
        when ProbabilityOfDefault >=0.50 and ProbabilityOfDefault <0.65 then '0.50-0.65'
        else '>0.65'
        end as TProbabilityOfDefault,
        case when LanguageCode is null then 'missing'
```

```
        when LanguageCode=1 then 'Estonio'
        when LanguageCode=2 then 'Ingles'
        when LanguageCode=3 then 'Ruso'
        when LanguageCode=4 then 'Fines'
        when LanguageCode=6 then 'Castellano'
        when LanguageCode=9 then 'Eslovaco'
        else 'Otros'
        end as TLanguageCode,
        case when Interest is null then 'missing'
        when Interest <15 then '<15'
        when Interest >=15 and Interest <20 then '15-20'
        when Interest >=20 and Interest <25 then '20-25'
        when Interest >=25 and Interest <30 then '25-30'
        when Interest >=30 and Interest <35 then '30-35'
        when Interest >=35 and Interest <40 then '35-40'
        when Interest >=40 and Interest <50 then '40-50'
        when Interest >=50 and Interest <60 then '50-60'
        when Interest >=60 and Interest <75 then '60-75'
        else '>75'
        end as TInterest,
        case when PorcRepaymentsPreviousLoans is null and NoOfPreviousLoansBeforeLoan=0 then 'Sin
prestamos anteriores'
        when PorcRepaymentsPreviousLoans is null and NoOfPreviousLoansBeforeLoan^=0 then 'missing'
        when PorcRepaymentsPreviousLoans =0 and NoOfPreviousLoansBeforeLoan=0 then 'Sin prestamos
anteriores'
        when PorcRepaymentsPreviousLoans =0 and NoOfPreviousLoansBeforeLoan^=0 then 'Sin pagos
previos'
        when PorcRepaymentsPreviousLoans >0 and PorcRepaymentsPreviousLoans <0.15 then '0-0.15'
        when PorcRepaymentsPreviousLoans >=0.15 and PorcRepaymentsPreviousLoans <0.20 then '0.15-
0.20'
        when PorcRepaymentsPreviousLoans >=0.20 and PorcRepaymentsPreviousLoans <0.25 then '0.20-
0.25'
        when PorcRepaymentsPreviousLoans >=0.25 and PorcRepaymentsPreviousLoans <0.30 then '0.25-
0.30'
        when PorcRepaymentsPreviousLoans >=0.30 and PorcRepaymentsPreviousLoans <0.35 then '0.30-
0.35'
        when PorcRepaymentsPreviousLoans >=0.35 and PorcRepaymentsPreviousLoans <0.40 then '0.35-
0.40'
        when PorcRepaymentsPreviousLoans >=0.40 and PorcRepaymentsPreviousLoans <0.45 then '0.40-
0.45'
        when PorcRepaymentsPreviousLoans >=0.45 and PorcRepaymentsPreviousLoans <0.55 then '0.45-
0.55'
        when PorcRepaymentsPreviousLoans >=0.55 and PorcRepaymentsPreviousLoans <0.65 then '0.55-
0.65'
        when PorcRepaymentsPreviousLoans >=0.65 and PorcRepaymentsPreviousLoans <0.75 then '0.65-
0.75'
        when PorcRepaymentsPreviousLoans >=0.75 and PorcRepaymentsPreviousLoans <0.95 then '0.75-
0.95'
        when PorcRepaymentsPreviousLoans >=0.95 and PorcRepaymentsPreviousLoans <1.15 then '0.95-
1.15'
        else '>1.15'
        end as TPorcRepaymentsPreviousLoans,
        case when AppliedAmount <1000 then '<1000'
        when AppliedAmount >1000 and AppliedAmount <2000 then '1000-2000'
        when AppliedAmount >2000 and AppliedAmount <3000 then '2000-3000'
        when AppliedAmount >3000 and AppliedAmount <4000 then '3000-4000'
        when AppliedAmount >4000 and AppliedAmount <6500 then '4000-6500'
        when AppliedAmount >6500 and AppliedAmount <9500 then '6500-9500'
        else '>9500'
        end as TAppliedAmount,
        case when IncomeTotal <500 then '<500'
        when IncomeTotal >=500 and  IncomeTotal <800 then '500-800'
        when IncomeTotal >=800 and  IncomeTotal <1000 then '800-1000'
        when IncomeTotal >=1000 and  IncomeTotal <1350 then '1000-1350'
        when IncomeTotal >=1350 and  IncomeTotal <1600 then '1350-1600'
        when IncomeTotal >=1600 and  IncomeTotal <1900 then '1600-1900'
        when IncomeTotal >=1900 and  IncomeTotal <2100 then '1900-2100'
        when IncomeTotal >=2100 and  IncomeTotal <2400 then '2100-2400'
        when IncomeTotal >=2400 and  IncomeTotal <2800 then '2400-2800'
        else '>2800'
        end as TIncomeTotal,
        case when LiabilitiesTotal =0 then 'Sin pasivos'
        when LiabilitiesTotal >0 and LiabilitiesTotal <250 then '<250'
        when LiabilitiesTotal >=250 and  LiabilitiesTotal <300 then '250-300'
        when LiabilitiesTotal >=300 and  LiabilitiesTotal <350 then '300-350'
        when LiabilitiesTotal >=350 and  LiabilitiesTotal <450 then '350-450'
        when LiabilitiesTotal >=450 and  LiabilitiesTotal <550 then '450-550'
        when LiabilitiesTotal >=550 and  LiabilitiesTotal <650 then '550-650'
        when LiabilitiesTotal >=650 and  LiabilitiesTotal <750 then '650-750'
        when LiabilitiesTotal >=750 and  LiabilitiesTotal <850 then '750-850'
        when LiabilitiesTotal >=850 and  LiabilitiesTotal <950 then '850-950'
        when LiabilitiesTotal >=950 and  LiabilitiesTotal <1150 then '950-1150'
        when LiabilitiesTotal >=1150 and  LiabilitiesTotal <1500 then '1150-1500'
        when LiabilitiesTotal >=1500 and  LiabilitiesTotal <2000 then '1500-2000'
        else '>2000'
        end as TLiabilitiesTotal,
        case when AmountOfPreviousLoansBeforeLoan =0 then 'Sin prestamos anteriores'
        when AmountOfPreviousLoansBeforeLoan >0 and AmountOfPreviousLoansBeforeLoan <700 then '<700'
```

```
        when AmountOfPreviousLoansBeforeLoan >=700 and AmountOfPreviousLoansBeforeLoan <1475 then
'700-1475'
        when AmountOfPreviousLoansBeforeLoan >=1475 and AmountOfPreviousLoansBeforeLoan <2100 then
'1475-2100'
        when AmountOfPreviousLoansBeforeLoan >=2100 and AmountOfPreviousLoansBeforeLoan <3000 then
'2100-3000'
        when AmountOfPreviousLoansBeforeLoan >=3000 and AmountOfPreviousLoansBeforeLoan <4300 then
'3000-4300'
        when AmountOfPreviousLoansBeforeLoan >=4300 and AmountOfPreviousLoansBeforeLoan <6600 then
'4300-6600'
        else '>6600'
        end as TAmountOfPreviousLoansBeforeLoan,
        case when PorFreeCash <0.30 then '<0.30'
        when PorFreeCash >=0.30 and  PorFreeCash <0.35 then '0.30-0.35'
        when PorFreeCash >=0.35 and  PorFreeCash <0.40 then '0.35-0.40'
        when PorFreeCash >=0.40 and  PorFreeCash <0.45 then '0.40-0.45'
        when PorFreeCash >=0.45 and  PorFreeCash <0.50 then '0.45-0.50'
        when PorFreeCash >=0.50 and  PorFreeCash <0.60 then '0.50-0.60'
        when PorFreeCash >=0.60 and  PorFreeCash <0.70 then '0.60-0.70'
        when PorFreeCash >=0.70 and  PorFreeCash <0.80 then '0.70-0.80'
        when PorFreeCash >=0.80 and  PorFreeCash <0.90 then '0.80-0.90'
        when PorFreeCash >=0.90 and  PorFreeCash <1 then '0.90-1'
        when PorFreeCash >=1 and  PorFreeCash <1.2 then '1-1.2'
        else '>1.2'
        end as TPorFreeCash,
        case when PorcEarlyRepaymentsPreviousLoans is null and NoOfPreviousLoansBeforeLoan=0 then
'Sin prestamos anteriores'
        when PorcEarlyRepaymentsPreviousLoans is null and NoOfPreviousLoansBeforeLoan^=0 then
'missing'
        when PorcEarlyRepaymentsPreviousLoans =0 and NoOfPreviousLoansBeforeLoan=0 then 'Sin
prestamos anteriores'
        when PorcEarlyRepaymentsPreviousLoans =0 and NoOfPreviousLoansBeforeLoan^=0 then 'Sin
amortizaciones anticipadas'
        when PorcEarlyRepaymentsPreviousLoans >0 and NoOfPreviousLoansBeforeLoan<0.2 then '<0.2'
        when PorcEarlyRepaymentsPreviousLoans >0.2 and NoOfPreviousLoansBeforeLoan<0.4 then '<0.2-
0.4'
        when PorcEarlyRepaymentsPreviousLoans >0.4 and NoOfPreviousLoansBeforeLoan<0.6 then '<0.4-
0.6'
        when PorcEarlyRepaymentsPreviousLoans >0.6 and NoOfPreviousLoansBeforeLoan<1 then '<0.6-1'
        when PorcEarlyRepaymentsPreviousLoans >1 and NoOfPreviousLoansBeforeLoan<1.2 then '<1-1.2'
        else '>1.2'
        end as TPorcEarlyRepaymentsPreviousL,
        case when ExpectedReturn is null then 'missing'
        when ExpectedReturn <-0.25 then '-0.25'
        when ExpectedReturn>=-0.25 and ExpectedReturn <-0.2 then '-0.25-(-0.20)'
        when ExpectedReturn>=-0.20 and ExpectedReturn <-0.15 then '-0.20-(-0.15)'
        when ExpectedReturn>=-0.15 and ExpectedReturn <-0.1 then '-0.15-(-0.10)'
        when ExpectedReturn>=-0.10 and ExpectedReturn <-0.05 then '-0.10-(-0.05)'
        when ExpectedReturn>=-0.05 and ExpectedReturn <0 then '-0.05-0'
        when ExpectedReturn>=0 and ExpectedReturn <0.05 then '0-0.05'
        when ExpectedReturn>=0.05 and ExpectedReturn <0.1 then '0.05-0.10'
        when ExpectedReturn>=0.10 and ExpectedReturn <0.12 then '0.10-0.12'
        when ExpectedReturn>=0.12 and ExpectedReturn <0.14 then '0.12-0.14'
        when ExpectedReturn>=0.14 and ExpectedReturn <0.16 then '0.14-0.16'
        when ExpectedReturn>=0.16 and ExpectedReturn <0.18 then '0.16-0.18'
        when ExpectedReturn>=0.18 and ExpectedReturn <0.20 then '0.18-0.20'
        when ExpectedReturn>=0.20 and ExpectedReturn <0.22 then '0.20-0.22'
        when ExpectedReturn>=0.22 and ExpectedReturn <0.24 then '0.22-0.24'
        when ExpectedReturn>=0.24 and ExpectedReturn <0.26 then '0.24-0.26'
        else '>26'
        end as TExpectedReturn
from tfm.testmuestra4;


/*Set con todas la variables sin tramificar y eliminamos las transformadas
porque por selección de variables no van a entrar en el modelo. También eliminamos las
vaeriables que son componentes de las variables porcentaje*/

/*El evento es el impago*/
data tfm.testtodasorig;
set tfm.testmuestra4;
drop LoanId LoanNumber ListedOnUTC BiddingStartedOn UserName
LoanApplicationStartedDate ApplicationSignedHour
County TransAge TransIncomeTotal TransLiabilitiesTotal TransInterest TransExistingLiabilites
TransNoOfPreviousLoansBeforeLoan
AmountOfPreviousLoansBeforeLoan PreviousRepaymentsBeforeLoan PreviousEarlyRepaymentsBefoleLoa
IncomeTotal LiabilitiesTotal Income Impago;
run;


/*Set con sin las variables de calificación que proporciona Bondora, sin tramificar y eliminamos las
transformadas
porque por selección de variables no van a entrar en el modelo. También eliminamos las
vaeriables que son componentes de las variables porcentaje*/
data tfm.testconocidasorig;
set tfm.testmuestra4;
drop LoanId LoanNumber ListedOnUTC BiddingStartedOn UserName
LoanApplicationStartedDate ApplicationSignedHour
```

```
County TransAge TransIncomeTotal TransLiabilitiesTotal TransInterest TransExistingLiabilites
TransNoOfPreviousLoansBeforeLoan
AmountOfPreviousLoansBeforeLoan PreviousRepaymentsBeforeLoan PreviousEarlyRepaymentsBefoleLoa
IncomeTotal LiabilitiesTotal
ExpectedLoss LossGivenDefault ExpectedReturn ProbabilityOfDefault ModelVersion Rating
CreditScoreEsMicroL
CreditScoreEsEquifaxRisk CreditScoreFiAsiakasTietoRiskGra CreditScoreEeMini Income Impago;
run;

/*Set con todas la variables tramificadas y eliminamos las transformadas
porque por selección de variables no van a entrar en el modelo. También eliminamos las
vaeriables que son componentes de las variables porcentaje*/

data tfm.testtodastram;
set tfm.testmuestra5;
drop LoanId LoanNumber ListedOnUTC BiddingStartedOn UserName
LoanApplicationStartedDate ApplicationSignedHour
County TransAge TransIncomeTotal TransLiabilitiesTotal TransInterest TransExistingLiabilites
TransNoOfPreviousLoansBeforeLoan
AmountOfPreviousLoansBeforeLoan PreviousRepaymentsBeforeLoan PreviousEarlyRepaymentsBefoleLoa
IncomeTotal LiabilitiesTotal
MonthlyPayment ExpectedLoss ExpectedReturn ProbabilityOfDefault LanguageCode Interest
PorcRepaymentsPreviousLoans
AppliedAmount AmountOfPreviousLoansBeforeLoan PorFreeCash PorcEarlyRepaymentsPreviousLoans Income
Impago;
run;
```

Anexo VI. Diagrama SAS Miner para valoración de modelo y puntuación de datos test



Anexo VII. Generación tablas tasa de fallos

```
/******** Evaluación Tarjeta Puntuación *********/

/*Juntamos el valor original de Impago y la puntuación que le da sas*/
proc sql;
create table tfm.comparapuntuacion as
select a.LoanId, a.Impago, b.EM_CLASSIFICATION
```

```
from tfm.testtodasorigconimpagoid a
left join tfm.tarjetapuntuacion_score b
on a.LoanId=b.LoanId;
quit;

data tfm.comparapuntuacion;
set tfm.comparapuntuacion;
impago_cat=put(Impago,z1.);
run;

data tfm.comparapuntuacion;
set tfm.comparapuntuacion;
drop impago;
run;

data tfm.comparapuntuacion;
set tfm.comparapuntuacion;
rename impago_cat=impago;
run;



/*Comparamos cuántos ha acertado*/
proc sql;
create table tfm.aciertos as
select *,
        case when (Impago= '1' and EM_CLASSIFICATION='1') or (Impago='0' and EM_CLASSIFICATION='0')
then 1
        else 0
        end as aciertos,
        case when Impago='1' and EM_CLASSIFICATION='1' then 'VP'
        when Impago='1' and EM_CLASSIFICATION='0' then 'FN'
        when Impago='0' and EM_CLASSIFICATION='1' then 'FP'
        when Impago='0' and EM_CLASSIFICATION='0' then 'VN'
        end as precision
from tfm.comparapuntuacion;
quit;

proc sql;
create table tfm.tasas as
select precision, count(LoanId) as numobservaciones
from tfm.aciertos
group by precision;
quit;

proc sql;
create table tfm.tasas2 as
select aciertos, count(LoanId) as numobservaciones
from tfm.aciertos
group by aciertos;
quit;

/********** Comparación con Rating ***********/
proc sql;
create table tfm.compararating as
select LoanId, rating, impago
from tfm.testtodasorigconimpagoid;
quit;
proc sql;
create table tfm.compararating as
select LoanId, rating, impago,
        case when rating='AAA' oR rating='AA' or rating ='A' then 0
        else 1
        end as clasificacion
from tfm.compararating;
proc sql;
create table tfm.aciertosrating as
select *,
        case when (Impago= 1 and clasificacion=1) or (Impago=0 and clasificacion=0) then 1
        else 0
        end as aciertos,
        case when Impago=1 and clasificacion=1 then 'VP'
        when Impago=1 and clasificacion=0 then 'FN'
        when Impago=0 and clasificacion=1 then 'FP'
        when Impago=0 and clasificacion=0 then 'VN'
        end as precision
from tfm.compararating;
quit;
proc sql;
create table tfm.tasasrating as
select precision, count(LoanId) as numobservaciones
from tfm.aciertosrating
group by precision;
quit;

proc sql;
create table tfm.tasasrating2 as
select aciertos, count(LoanId) as numobservaciones
```

```
from tfm.aciertosrating
group by aciertos;
quit;
```