Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL DE ESPECIALIZACIÓN

PAGOS DE UNA COMPAÑÍA DE CRÉDITOS Y LA REDUCCIÓN DE COSTOS

La búsqueda de una predicción a través de modelos de machine learning.

AUTOR: MERCEDES ARTESE

[DICIEMBRE 2019]

Resumen

El problema que se pretende trabajar es cuáles serían las implicancias que tendría una mala toma de decisiones en una empresa de crédito asociativo. Los préstamos resuelven una cantidad importante de situaciones de la economía personal, familiar y empresarial de estos tiempos. Un mal uso del crédito puede generar al usuario un problema económico como, por ejemplo: contraer una deuda superior a nuestra capacidad de pago.

Las diferentes formas de créditos tienen costos asociados, en algunos casos altos. Estos costos se derivan de la realización determinadas tareas operativas para decidir otorgar o no los recursos, y componen el Costo Financiero Total, que incluye una tasa de interés aplicada, gastos de análisis crediticio de quien pide el préstamo, un valor por la contratación de seguros obligatorios como ser seguro de vida o incendio, costos de cuenta creada para pago de cuotas y gastos operacionales. Como a menudo estos costos suelen ser muy altos, es necesario implementar mejoras continuas en los procesos de tomas de decisiones, que ayuden a ser más eficientes a un menor costo.

Entonces, el problema que se aborda en este trabajo es cómo obtener una predicción adecuada de la tasa de cumplimiento de pagos en una compañía de créditos asociativos que permita reducir los costos a partir del uso de machine learning, para esto se plantean como objetivo general analizar el valor de la tasa de cumplimiento de pagos que permita minimizar los costos en los créditos de organizaciones prestamistas, para lo cual será importante en primer lugar describir la necesidad de la autogestión para organizaciones recuperadas, luego desarrollar un modelo de machine learning que permita predecir la tasa de cumplimiento de pagos, y por último analizar la manera de implementar la predicción para deducir la tasa de cumplimiento de pago en los créditos.

Para poder cumplir con estos objetivos, se toma la base de datos “Lending Club Loan Data” de la plataforma Kaggle. Este archivo contiene datos completos de préstamos para todos los préstamos emitidos hasta el 2007- 2015, incluido el estado actual del préstamo (actual, tardío, pagado en su totalidad, etc.) y la información de pago más reciente. Las características adicionales incluyen puntajes de crédito, cantidad de consultas financieras, direcciones que incluyen código postal y estado, y colecciones, entre otras.

A partir de esta información, se procede a realizar modificaciones en esta base original para mejorarla y obtener una predicción lo más exacta posible. Estas modificaciones implican tanto la creación como así la modificación, eliminación, imputación de determinadas columnas, que se detallan al final de este trabajo en el Anexo I. Por otra parte, esta base extraída originalmente, contiene datos que no se encuentran balanceados, estos es la proporción de las variables de decisión son muy dispares. Por precisarse que estas diferencias se acerquen, se procede a balancearlos hasta alcanzar valores equitativos con herramientas de RapidMiner.

Estructura

Introducción

Las agencias de crédito fueron transformando la información en modelos matemáticos de probabilidad y riesgo, alejándose de la individualidad. En base a esto, se analiza la importancia que tiene en la toma de decisiones la consistencia e integración de los datos, como así la contextualización en la recolección de los mismos.

A la hora de evaluar el costo de un crédito, se considera que uno de los criterios más importantes es la tasa de interés efectiva anual que le cobra la entidad. Sin embargo, hay otros costos en los que incurre en la solicitud de un préstamo, o cuando incumple las cuotas, o cuando paga anticipadamente los saldos de capital, y cuyo monto puede ser tan relevante que incluso induciría al cliente a cambiar de entidad prestamista. Se considera que el mayor componente de los costos de este tipo de organizaciones está determinado por la tasa de incumplimiento de pago de sus prestatarios. Debido a esto, se realiza un análisis para segmentar la cartera de clientes con el fin de determinar la aceptación o no del crédito para disminuir los costos.

El riesgo de crédito es la posibilidad de sufrir una pérdida como consecuencia de un impago por parte de nuestra contrapartida en una operación financiera. Supone además, una variación en los resultados financieros posteriores a la quiebra de una compañía, tanto de activos financieros o carteras de inversión. Por esto, es también una forma de medir la probabilidad que tiene un moroso frente a su acreedor de cumplir con sus obligaciones de pago, durante la vida del activo o durante el plazo establecido en el contrato. A diferencia de riesgo de mercado en el cual la distribución de riesgo para las partes es simétrica, en el riesgo de crédito la distribución es asimétrica negativa, es decir, que no se distribuye de forma homogénea el riesgo para cada una de las partes.

La idea es realizar entonces, un análisis de un probable valor de la tasa de cumplimiento de pagos de créditos para poder evaluar la reducción en los costos de este tipo de organizaciones. Para esto es importante realizar una descripción de la necesidad de la autogestión para organizaciones recuperadas y evaluar qué tipo de organizaciones se incluyen dentro de esta clasificación y por qué. Luego, es preciso desarrollar un modelo de

machine learning que permita predecir la tasa en cuestión, a partir de variables determinantes de la misma.

Por último, es interesante analizar diferentes alternativas de implementación de la predicción realizada, de manera de darle funciones que resulten operativas para las organizaciones implicadas. Una de estas alternativas es el cálculo de la pérdida esperada crediticia según la normativa NIIF 9, siendo que hasta su implementación las normas contables basaban el reconocimiento de las pérdidas a partir de la obtención de una evidencia de la materialización de ésta. Esta nueva metodología de valuación de la pérdida propone tener en cuenta no sólo las pérdidas derivadas del incumplimiento incurrido, sino que también considera la posible pérdida a futuro, lo cual da a la organización una herramienta más para poder componer las previsiones necesarias para hacer frente a las pérdidas que aún no se están percibiendo.

Para cumplir el objetivo entonces, este trabajo se estructura de la siguiente manera: en un apartado se describirán las organización recuperadas, sus características y sus necesidades de autofinanciación, como así la intervención de los grandes volúmenes de datos en la gestión de este tipo de compañías y otras problemáticas asociadas al análisis y gestión de grades volúmenes de datos; en otro apartado se realizará el desarrollo de un modelo de machine learning que permita predecir la tasa de cumplimiento de pagos, y en un último apartado se dará una probable implementación de la predicción que de utilidad al trabajo realizado para las compañías de créditos.

Apartado 1: Organizaciones recuperadas y grandes volúmenes de datos

1.1 Organizaciones recuperadas

A partir de diferentes crisis, en las cuales muchas organizaciones fueron afectadas, surgen las organizaciones "recuperadas". Como consecuencia de esto, nacen organizaciones sin fines de lucro que trabajan brindando asistencia técnica y financiera a colectivos de trabajadores auto gestionados. En la Argentina, se presentó una nueva línea de préstamos destinados a grupos asociativos auto gestionados que realicen actividades productivas o de comercialización de bienes y servicios. A estos fines, se conformó "La Base" a finales del año 2004, en un contexto en el que las empresas recuperadas crecieron exponencialmente y que, además, enfrentaban la dificultad de obtener recursos financieros para capital de trabajo. Esta organización, entregaba préstamos, con el objetivo de fortalecer los procesos de autogestión productiva: "En función de este objetivo quienes forman parte de La Base definen su trabajo desde las finanzas para la autogestión, en el marco de la heterogeneidad contenida en el concepto de finanzas solidarias o alternativas." (Pinto, S., & Litman, L., 2019, pág. 7)

Días Coelho sostiene que "las finanzas son la ciencia que trata la utilización del dinero, su costo, su rendimiento, protección y control, captación y reciclaje de sus distintos productos". Pero en la práctica, este sistema deja afuera al sector informal de la economía. Durante el auge del sistema capitalista, muchos países crecieron de manera exponencial, pero en la misma medida lo hizo la desigualdad. De esta manera muchos sectores al no poder cumplir con las formalidades requeridas para acceder a un crédito, se vieron desplazados del beneficio de obtener uno. Verbeke sostiene que las microfinanzas surgen como una respuesta posible a la existencia de mercados financieros incompletos mediante la prestación de servicios financieros dirigidos a proyectos o microemprendimientos que generalmente están excluidos del sistema bancario formal (Verbeke, 2007, pág. 194).

Dada la magnitud de los problemas, los doctores Luis del Castillo Sánchez junto con José Manuel Pozo Rodríguez concluyeron que "resulta utópico pensar que el Microcrédito sea la solución a la pobreza, al desempleo y a las condiciones de vida y salud de las personas más desfavorecidas. Sin embargo, el hecho de constituir una alternativa al sistema financiero tradicional, el desarrollo de instituciones que otorgan microcréditos y el impulso de los

proyectos locales, lo convierte en un instrumento complementario de los necesarios programas de desarrollo de alcance nacional y estratégico".

Se toma una base de datos de una compañía estadounidense que permite a sus prestatarios crear listas de préstamos en su página web aportando datos de ellos mismos y de los préstamos que les gustaría obtener. Todos los préstamos son destinados a particulares no asegurados que pueden rondar entre 1.000 y 35.000 dólares. En función de la puntuación de crédito del prestatario, el historial de crédito, de la cantidad solicitada y de la proporción entre ingresos y gastos del prestatario, esta compañía determina si el prestatario es confiable y, en caso de serlo, asigna a los préstamos que son aprobados un grado de crédito que determina la tasas y el tipo de interés a pagar. El plazo de préstamo habitualmente es unos tres años; también está disponible el periodo de cinco años con un tipo de interés mayor y tasas adicionales.

Cuando fue fundada inicialmente esta organización, se posicionó como un servicio de red social que permitiese fijar bases para que existiese la oportunidad de encontrar grupos afines, basándose en la teoría de que los solicitantes de préstamos serían más reconocidos para los prestamistas si existiesen afinidades y una vinculación social. Desarrollaron un algoritmo llamado LendingMatch para identificar factores de relación social como la ubicación geográfica, o la vinculación educativa o profesional, y permitir su contacto a través de la red social. Esto implica que tanto los inversores como los prestatarios son personas y no entidades financieras.

Los inversores pueden buscar y definir los préstamos en la página web de la compañía y seleccionar los préstamos en los que desea invertir sobre la base de la información aportada sobre el solicitante, la cantidad del préstamo, el nivel de interés, y el propósito del préstamo. Los préstamos sólo pueden ser seleccionados conforme a los tipos de interés definidos por la compañía, pero los inversores pueden decidir qué cantidad destinar a cada solicitud. Los inversores obtienen dinero en función del tipo de interés (cuyas tasas van del 6,03% al 24,89%) dependiendo del grado de crédito asociado al préstamo. Por otro lado, la compañía obtiene dinero de aplicar a los solicitantes una cuota inicial y a los inversores unos gastos de administración. Una vez emitidos los préstamos, esta compañía adquiere los préstamos al banco emisor pasan a ser responsabilidad de la compañía, prometiéndole al prestamista los pagos recibidos del beneficiario del préstamo, menos las tasas de servicio, mientras los créditos figuran como carentes de garantía, es decir que existe el riesgo de perder todo o

parte de la inversión si la compañía se volviera insolvente o declarase la bancarrota, incluso si el receptor del préstamo siguiese pagando

Muchas veces, las microfinanzas y los microcréditos fueron destinados a brindar un conjunto de servicios financieros enfocados a la atención de las personas o microempresas, como asi pymes que no tienen acceso al crédito bancario, como estos casos que se plantean. Se trata entonces de destinatarios con ausencia de garantías reales, cierto grado de informalidad que les impide cumplir con los requisitos bancarios, baja dotación de sus activos o un número reducido de empleados. Sin embargo, a partir de la sanción de la Ley de Microcréditos las microfinanzas se encuentran enfocadas al fortalecimiento de la economía solidaria. Segun Heller, "Una empresa pertenece a la economía social si su actividad productiva se basa en técnicas de organización específicas. Estas técnicas se fundamentan en los principios de solidaridad y participación (que normalmente responden a la norma un hombre-un voto entre sus miembros, sean estos productores, usuarios o consumidores, así como en los valores de autonomía y de ciudadanía)"
Hasta el año 2005, los microcréditos en la Argentina eran otorgados solamente por las instituciones de financiamiento a microempresas (IMF): sociedades anónimas, cooperativas, asociaciones civiles, y otros del sector privado.

La experiencia de las empresas recuperadas en Argentina a lo largo de la última década demuestra que los trabajadores lograron organizar la producción y la comercialización por sus propios medios, sin la necesidad de contar con la presencia de CEOs. La recuperación de fábricas por parte de los trabajadores no forma parte de un proyecto político, sino que fue una estrategia de supervivencia en un contexto de extrema desocupación. Las prácticas concretas que asumieron estas experiencias colectivas y la construcción de herramientas organizativas autogestionarias son producto de un contexto político, económico, social y cultural que sirvió de marco habilitante para el despliegue de experiencia alternativas.

Según el artículo de Gavriela Roffinelli, Vanesa Ciolli y Sergio Papi "este fenómeno ha sido atractivo en el ámbito de las ciencias sociales; numerosas investigaciones se iniciaron con el fin de conocer y reflexionar sobre estas experiencias. La mayor parte de los trabajos de investigación describen la cantidad de empresas, el número de trabajadores involucrados, las ramas de producción en que se desarrollan, las formas de organización asumidas, los procesos de toma de decisiones, las transformaciones en el imaginario

colectivo. Sin embargo, carecen de un análisis que explique las tensiones que su inserción (subordinada) en las relaciones sociales capitalistas les plantean para su desarrollo tanto en el nivel de las unidades productivas como en el nivel de la construcción de nuevas relaciones sociales en la producción de bienes y servicios orientados a satisfacer las necesidades humanas" (Gavriela Roffinelli, Vanesa Ciolli y Sergio Papi, 2013 , pág.17).

Es factible entonces considerar que el desarrollo de las empresas recuperadas se ha visto condicionado por la concentración monopólica de las ramas productivas en las que se desempeñan y por la ambigüedad de las acciones estatales.

1.2 Intervención de los grandes volúmenes de datos en la gestión de compañías recuperadas

Los informes de crédito al consumo se han extendido mucho más allá de las instituciones de crédito. En los EE. UU., país de origen de la base analizada, las organizaciones consultan informes crediticios cuando toman decisiones sobre si otorgar a las personas acceso a una amplia gama de recursos económicos, incluyendo viviendas de alquiler, seguros, servicios públicos y, lo más controvertido, empleos (Shepard, 2012; Traub, 2013).

Sin embargo, la forma en que los profesionales de contratación dan sentido a los informes de crédito sigue siendo un misterio (Bryan y Palmer, 2012). A diferencia de los puntajes de crédito, los informes de crédito son documentos largos con datos detallados sobre cuentas de tarjetas de crédito individuales, hipotecas, préstamos estudiantiles, deudas médicas, y sentencias judiciales relacionadas con dinero, entre otros.

Si bien los sociólogos económicos y los especialistas en vigilancia han llamado la atención sobre el alcance cada vez mayor de los registros de crédito en la vida cotidiana (Marron, 2009; Fourcade y Healy, 2013; Roderick, 2014; Rona-Tas, 2017), este trabajo generalmente supone que los puntajes de crédito basados en el riesgo de las empresas financieras, se trasladen a ámbitos no crediticios. Esto concuerda con la creciente dependencia de las organizaciones en las métricas para tomar decisiones (Espeland y Sauder, 2007; Espeland y Stevens, 2008; Timmermans y Epstein, 2010), aunque la falta de evidencia clara sobre si el historial crediticio predice matemáticamente el comportamiento en el lugar de trabajo complica esta posibilidad. (Barbara Kiviat, 2017, pág. 1)

Por el contrario, la literatura que demuestra el significado social y moral irreprimible de la vida económica, abre la posibilidad de que profesionales realicen interpretaciones más tradicionales del historial crediticio.

Los modelos matemáticos de probabilidad y riesgo, reemplazaron los juicios morales en la medida que quienes usaban la historia crediticia aprendieron a transformar la calificación de riesgo individual en grupos de riesgo que permiten predecir el cumplimiento en los pagos, y a partir de esto los prestamistas solo se preocupaban en si su score era alto o bajo.

Hoy en día los profesionales de marketing pueden deducir mucha información tangible e intangible sobre cualquier consumidor de todo el mercado, a partir del uso de tecnología de Big Data como Prizm y Esri, que les permite una investigación más a fondo.

1.3 Otros problemas asociados a la gestión de datos

Por otro lado, se plantea el desafío de mantener la discriminación fuera de la imagen. Usando los llamados "Modelos de Propensión", se toma una población y se determina la probabilidad de que estén en el mercado para un nuevo automóvil. Se estudian estilos de vida y datos demográficos de los individuos de interés y se ordenan, teniendo así una pieza más de inteligencia para usar. Se tiene información de al menos el 95% de la población estadounidense, y el modelado se está volviendo mucho más sofisticado con una precisión mejorada y resultados más rápidos, además el modelado y la creación de nueva inteligencia se está haciendo en tiempo real y no es estático, pudiendo hacer y usar predicciones de una forma más acelerada.

Para los consumidores, el Big Data es una herramienta que impulsa la publicidad más personalizada y efectiva, ofreciendo anuncios y ofertas más relevantes e interesantes para cada persona en comparación con los anuncios aleatorios que de otra manera se mostrarían en línea. Pero esa granularidad puede causar repercusiones significativas en cuanto al potencial de discriminación planteándose si el Big Data significa una herramienta para la inclusión o exclusión.

Una de las dificultades viene de la mano de que la frecuencia de almacenamiento a nivel algorítmico y que se encuentre patentado. La abogada de la ACLU, Rachel Goodman dice:

“Sabemos lo suficiente sobre lo que existe y la forma en que el mundo ha funcionado históricamente para realmente sospechar que la discriminación está ocurriendo. Una y otra vez en la historia, hemos visto cómo los préstamos de crédito e hipotecarios y otros tipos de préstamos terminan siendo distribuidos de manera desigual”.

Los legisladores han promulgado regulaciones para tratar de mantener las cosas justas, especialmente en materia de vivienda y servicios financieros. La ley estadounidense contra la discriminación también prohíbe el uso de ciertos datos conocidos como “clases protegidas” como medio de discriminación. Aunque también hay casos en los que determinados productos que pueden considerarse discriminatorios por estar destinados a personas con diferencias socioeconómicas, aportan valor a la persona. Es cuando los especialistas en mercado utilizan la etnicidad para mantener las ofertas atractivas fuera del alcance de ciertos segmentos de la población que se convierte en discriminación.

Otra cuestión son los códigos postales que pueden ser tan útiles para los comercializadores porque son representantes o representantes cercanos de otros factores y ofrecen más información que la mera ubicación, como por ejemplo la raza. Al confiar en un proxy como el código postal, se puede llegar a discriminar racialmente hasta incluso de manera involuntaria. Y eliminar el proxy del modelo no siempre es una opción. En bienes raíces, por ejemplo, la ubicación es un factor demasiado esencial para eliminar de consideración. Es aún más complicado en línea donde las reglas del juego de discriminación son menos evidentes.

No tiene que ser así. Algunos consideran que Internet es el lugar ideal para eliminar muchos de los sesgos que afectan a las interacciones humanas fuera de línea, pudiendo ser utilizada sin importar la raza, el origen étnico u otras categorías protegidas para identificar a las personas como buenas perspectivas en función de su actividad y comportamiento en línea.

Pero eso no es lo que pasa la mayor parte del tiempo. Se han creado en Estados Unidos instituciones para mejorar las prácticas de vivienda y en la banca. Se considera que aquellas entidades que regulan deben considerar cómo la ley existente les permite exigir esa transparencia en línea, especialmente en marketing. La transparencia es particularmente importante cuando se trata de algoritmos de marketing, dice ella.

Se plantea, además, que la tecnología puede ayudar si los algoritmos de Machine Learning que trabajan con Big Data resultan en discriminación racial, entonces otros algoritmos pueden medir el efecto de la discriminación. Una vez que se ha medido el efecto, entonces la sociedad y el gobierno pueden decidir si la discriminación es intencional o no, y qué tipo de compensación o acción correctiva se puede tomar.

Mientras tanto, existen compañías que invirtieron significativamente en herramientas que brindan a sus clientes información sobre dónde, cuándo y a quiénes sirven sus máquinas los anuncios para que tengan la oportunidad de dar su opinión sobre lo que consideren que tiene sentido. Otras compañías limitan cómo se pueden usar sus datos, negándose a vender datos a determinadas empresas, por ejemplo. Internamente, programas de capacitación sensibiliza a los empleados sobre los problemas involucrados y las evaluaciones periódicas de impacto sobre la privacidad, ayudan a que la gente se detenga y piensen en lo que podría generarle a nivel personal utilizar sus datos de otra forma.

Apartado 2: Desarrollo de un modelo de machine learning que permita predecir la tasa de cumplimiento de pagos.

2.1 Introducción a modelos de predicción

El llamado "Machine Learning", es el estudio de algoritmos que tienen la capacidad de aprender de datos sin ser específicamente programados para ello. También es conocido con otros nombres menos comerciales, como aprendizaje estadístico o minería de datos, y con otros más comerciales, como inteligencia artificial.

Hoy en día, las tecnologías de Machine Learning han dado un gran salto en el mundo de las organizaciones, pudiendo ser utilizadas para conseguir ventajas competitivas al alcance de cualquier organización pequeña.

Esta innovación tecnológica se divide en aprendizaje supervisado o con intervención humana, y no supervisado que no requiere la intervención humana.
En este trabajo, se utilizarán herramientas de aprendizaje supervisado, consistente en hacer predicciones a futuro basadas en comportamientos o características que se han visto en datos ya almacenados, es decir: el histórico de datos. Esto, además, permite buscar patrones en datos históricos relacionando todos campos con un campo objetivo o "label".

El análisis predictivo inicia principalmente con las características o patrones que tienen las variables objetivo. Para la predicción realizada en este trabajo, se toma una base limpia y ordenada, y se abre en RapidMiner con el operador "Read Excel". Se cambia el rol a "loan_status" como "label", es decir, la variable a predecir, y a "id" como "id" para que no sea tenida en cuenta para la predicción, y luego el resto de las columnas dummies propias y las creadas por RapidMiner servirán para predecir.

La potencialidad de este set de datos con predicciones, viene de la mano de la posibilidad de construir un algoritmo de Machine Learning para predecir a las personas que podrían incumplir con sus préstamos, a ser utilizado por Lending Club para analizar la pérdida esperada de los préstamos a partir de la determinación de una tasa de incumplimiento de pagos. Luego, se puede tomar este trabajo para predecir durante determinada cantidad futura

de años hasta la fecha y ver qué cantidad real de incumplimiento se registró y luego utilizar esos datos para realizar nuevas predicciones o incluso para entrenar el modelo nuevamente y mejorar su precisión.

Para poder realizar una predicción, es preciso contar con una base de datos de entrenamiento sobre la cual se apliquen los modelos de prueba, y una base de práctica donde se aplicará el mejor modelo obtenido. A estos fines, se utiliza el operador **Cross Validation** que es un operador anidado. Tiene dos subprocesos: un subproceso de capacitación y un subproceso de prueba. El subproceso de entrenamiento se utiliza para entrenar un modelo. El modelo entrenado se aplica luego en el subproceso de Pruebas. El rendimiento del modelo se mide durante la fase de prueba y se solicitarán los criterios "Accuracy" y el área bajo la curva ROC (AUC ROC). En el operador se utilizarán 10 folds y el sampling type será estratificado.

2.2 Modelos Utilizados

El primer modelo ultilizado fue el de **Random Forest**, el cual solo usa árboles y lo que hace es tomar además de muestras de registros, muestras de las columnas. La ventaja de este modelo es que soporta cualquier base de datos, aunque haya datos perdidos. Se hace la prueba con 100 árboles bajo criterios "Accuracy", "Gain Ratio" e "Information Gain".

En segundo lugar, se realizará una prueba con el modelo **Stacking**, modelo que hace regresión logística de los mejores datos. Se le da más peso a los datos que más aportan, y menos a los que menos me aportan. Se toma el conjunto de datos y toma las variables de a una pero no se sabe si hay o no colinealidad. Como funciona muy bien en modelos de regresión, se utilizarán varios modelos básicos, como k-NN, Deep Learning, Decision Tree, Random Forest y Bagging. Dentro del proceso, del lado derecho se utiliza "Generalized Linear Model" que es un algoritmo que ajusta modelos lineales generalizados a los datos al maximizar la probabilidad de registro, usando un cálculo de ajuste del modelo paralelo, muy rápido y sumamente adaptable a los modelos con un número limitado de predictores con coeficientes distintos de cero. El proceso que hace este modelo es tomar la muestra entera y se la pasa a los diferentes modelos. Antes de hacerlo correr, se activa "keep all attributes".

El tercer modelo que se prueba es un árbol de decisión, modelo llamado **Decission Tree**, al cual Se le solicita que estime 5 árboles de opciones diferentes en las que las palabras de esta

base pueden llegar a alguno de ambos resultados, utilizando como criterio “Accuracy”, “Gain Ratio” e “Information Gain”.

A continuación, el cuarto modelo probado es **Induction Rules**. Este operador funciona comenzando con las clases menos prevalentes, el algoritmo crece iterativamente y reduce las reglas hasta que no queden ejemplos positivos o la tasa de error sea superior al 50%. En la fase de crecimiento, para cada regla se agregan condiciones ávidamente a la regla hasta que sea perfecta (es decir, 100% precisa). El procedimiento prueba todos los valores posibles de cada atributo y selecciona la condición con mayor ganancia de información. La poda en los árboles de decisión es una técnica en la que se eliminan los nodos de hojas que no contribuyen al poder discriminativo del árbol de decisión. Esto se hace para convertir un árbol sobre-específico o sobre-ajustado a una forma más general para mejorar su poder predictivo en conjuntos de datos invisibles. Un concepto similar de poda implica en los conjuntos de reglas.

A continuación, se realiza una prueba de **Regresión Logística**, cuyo operador es una versión simplificada del operador de modelo lineal generalizado. Para realizar una Regresión logística, el parámetro Familia se establece automáticamente en binomial, y el parámetro de enlace en logit. Solo los parámetros más importantes se pueden ajustar para que este operador proporcione una regresión logística fácil de usar. La implementación de Regresión logística puede manejar los datos de entrenamiento con la etiqueta binominal (o polinominal de 2 clases), y los atributos nominales y numéricos de las características. El operador inicia un clúster H2 O local de 1 nodo y ejecuta el algoritmo en él. Aunque utiliza un nodo, la ejecución es paralela.

Y, por último, se ensaya una **Regresión Logística SMV**, que es un método de aprendizaje que se puede utilizar tanto para regresión como para la clasificación, y proporciona un algoritmo rápido y buenos resultado para muchas tareas de aprendizaje. Admite varios tipos de kernel, incluyendo punto, radial, polinomial, neural, anova, entre muchas otras.

De esta forma, se obtiene una predicción superadora para la variable a predecir de la base de datos utilizada.

2.3. Resultados de los Modelos

Los resultados de las distintas combinaciones de estos modelos aplicados, las referencias de cada uno y sus parámetros se encuentran explicados y con sus resultados en un cuadro al final de este trabajo en el Apéndice II. Lo que se puede obtener son diferentes combinaciones de valores de área bajo la curva ROC y de exactitud de predicción, encontrándose que, de todos los modelos ensayados, el que mejor aproxima una predicción es un Decision Tree con Decision, con una profundidad de 10 ramificaciones y en el cual para verificar la exactitud se utilizará el criterio Information Gain. La profundidad de un árbol va a variar dependiendo del tamaño y características de la base que se esté entrenando, y se usa para restringir la profundidad del árbol de decisión, cuya medida máxima no pone límite en la profundidad del árbol. En este caso, tendrá 10 nodos de profundidad.

Los resultados obtenidos dan un valor de un 100% de exactitud en la predicción y un micro average o promedio de desvío de un 0,1%

Se puede apreciar que aquellos modelos con los que mejor se aproxima la predicción que se pretende realizar, es el modelo de Árboles de Decisión con criterio del área bajo la curva ROC (AUC ROC).

La curva ROC es un gráfico en el que se observan todos los pares sensibilidad-especificidad resultantes de la variación continua de los puntos de corte en todo el rango de resultados observados. En el eje y de coordenadas se sitúa la sensibilidad o fracción de verdaderos positivos, mientras que en el eje x se sitúa la fracción de falsos positivos o 1-especificidad. La idea de la curva ROC es medir que combinación se tiene ponderando más cuanto más arriba están los positivos. Cuantos más errores se tienen más arriba, peor es el modelo. Lo ideal es obtener un modelo que tenga muchos más positivos en la parte superior y todos los negativos posibles en la parte inferior. El área máxima, y por ende el mejor caso de curva ROC es 1, y el peor caso es 0. Se trata de una curva y no una función porque no puede tener dos imágenes para un valor de “x”. Lo que se hace es graficar Verdaderos Positivos contra Falsos Positivos. Una de las características que tiene esto es que se puede saber por probabilidad si se ordenó bien una clasificación. Se puede observar también una línea azul, llamada **Línea de Azar**, que se forma tomando variables al azar y

da se obtiene una línea muy próxima a la línea media. La idea es que la curva ROC este bien despegada de esa línea, que este lejos del azar, ya que el objetivo es maximizar el área bajo la curva (AUC).

Se construyó entonces de manera exitosa un algoritmo para predecir a las personas que podrían incumplir con sus préstamos, que podría ser utilizado por Lending Club para analizar sus préstamos. Además, se pueden llegar a utilizar otras técnicas o algoritmos para mejorar el poder de predicción.

La principal dificultad que esta base presentó es la gran diferencia entre quienes cumplieron con sus pagos y quienes no durante el transcurso de los 8 años de datos.
Se puede tomar este trabajo para predecir durante determinada cantidad futura de años hasta la fecha y ver qué cantidad real de incumplimiento se registró y luego utilizar esos datos para realizar nuevas predicciones o incluso para entrenar el modelo nuevamente y mejorar su precisión.

Además, a futuro, se podrían a aplicar otros modelos como “Bagging”, o realizar predicciones utilizando programas como R o Python y comparar resultados.

Apartado 3. Implementación de la predicción el cálculo de pérdida esperada crediticia.

Es posible utilizar la predicción realizada, para facilitar la implementación del modelo de pérdidas crediticias esperadas que el BCRA pondrá en vigencia a partir del año 2020, establecido por las Normas Internacionales de Información Financiera (NIIF) y que alcanza a todos los instrumentos financieros contemplados bajo la norma anterior (NIC 39).

La idea de esta nueva normativa, cuyo disparador fue la crisis de 2008, es que estas entidades aporten al Banco Central información acerca de cómo reconocerán sus pérdidas a futuro y ya no sólo las incurridas u observables a la fecha de estimación, lo cual a la vez implica que cada compañía deberá disponer de un modelo de calculo que estime las pérdidas crediticias esperadas y no incurridas aun, para lo que será necesario y fundamental definir el concepto de incumplimiento de pago (default).

Es muy importante destacar que la principal complejidad de este nuevo modelo se encuentra en la determinación de la probabilidad de default o incumplimiento (PD) en todas las unidades de negocio de la empresa afectada.

Por otro lado, la recopilación y el análisis de datos para hallar el método óptimo de cálculo de la mencionada PD requerirán tiempo, por lo cual NIIF 9 plantea diferentes posibilidades que pueden ajustarse a varios activos o instrumentos afectados, lo cual implica una cifra que es ajustable en el tiempo y que depende de la realización de monitoreo constantes y automatizado de la tasa de mora.

3.1. Premisas y Limitaciones

La metodología de cálculo que debe aplicarse para la estimación de provisiones se basa en el concepto de pérdida esperada, el cual tiene una visión prospectiva del riesgo de crédito al cual se expone una cartera de créditos.
En este sentido, en esta sección se detallan premisas adoptadas y limitaciones que deben ser consideradas en los ejercicios de implementación de metodologías de cálculo de provisiones por NIIF 9.

Respecto de las limitaciones, es posible que la compañía no cuente con estimaciones internas de probabilidades de default para determinados segmentos, ni con modelos internos que permitan estimar el parámetro de riesgo Loss Given Default. En estos casos, se sugieren metodologías alternativas para la obtención de PD's y LGD's, de modo tal de reflejar de manera adecuada la medición del riesgo de crédito.

Se utilizan varias premisas para el desarrollo de este modelo. En primer lugar, la metodología que se aplica para el cálculo de provisiones está basada en el concepto de pérdida esperada, el cual, a diferencia de la pérdida incurrida, utilizada por NIC 39, tiene una visión prospectiva del riesgo de crédito al cual se expone una cartera de créditos.

Por otro lado, los criterios establecidos para los eslabones de riesgo de la cartera de crédito fueron definidos por la compañía, los cuales se encuentran basados en una serie de indicadores, dentro de los cuales pueden usarse días de atraso, variación en las calificaciones obtenidas de los modelos de clasificación de clientes, cantidad de veces que la deuda fue reestructurada, etc., y que permiten medir el deterioro de las operaciones a través del tiempo.

Y, por último, para aquellos segmentos que no son alcanzados por modelos internos de clasificación de clientes, se sugieren metodologías alternativas para el cálculo del parámetro de riesgo probabilidad de default.

1.2 Marco Metodológico

A diferencia de la normativa predecesora, NIC 39, que basaba la estimación de previsiones en base a la pérdida incurrida, el cálculo por NIIF 9 se basa en la estimación de las pérdidas esperadas por incumplimiento crediticio. De este modo, se considera tanto el valor económico de las pérdidas que ocurrieron antes de la fecha de reporte, como así también las que se espera que ocurran en el futuro. Es entonces que para las exposiciones que no se encuentran en situación de incumplimiento, se descuenta la pérdida desde el momento esperado de default al momento de la evaluación.

El valor económico de pérdida al momento de incumplimiento se estima a partir del producto de los parámetros de riesgo: Probabilidad de Default (PD), Loss Given Default(LGD) y Exposición Efectiva de Pérdida (EAD). Luego, la fórmula empleada para el cálculo de previsiones está representada en la siguiente expresión:

$$Pérdida\ Esperada\ NIIF\ 9\ =\ PDPIT\ *\ LGD\ *\ EAD\ *\ Desc_{MDE}-MA$$

Dónde $PDPIT$ es la probabilidad de que la contraparte deje de cumplir con sus obligaciones contractuales en entre la fecha de reporte ("momento actual") y una fecha posterior ("momento de default"); la LGD es el **porcentaje de la exposición** en default que finalmente se pierde en caso de incumplimiento. Equivale al complemento de la tasa de recuperación, es decir: el importe recuperado sobre la EAD y que contempla, entre otros factores, el efecto de las garantías. La EAD es el **monto expuesto** al momento de concretarse el incumplimiento, es decir, representa el monto máximo que la compañía podría perder en una operación en caso de incumplimiento de la contraparte y asumiendo que la recuperación de las posibles garantías que afectan a la misma fuese nula. Por lo tanto, la EAD es el saldo expuesto al momento de materializarse el incumplimiento. Finalmente, la tasa $Desc_{MDE-MA}$ es el descuento financiero del monto expuesto por el plazo que transcurre entre un momento de default futuro y el momento actual, utilizando la tasa activa del crédito.

La pérdida esperada crediticia se mide generalmente en base al riesgo de impago durante uno o dos horizontes de tiempo diferentes, dependiendo de si hubo evidencia de aumento exponencial del riesgo de crédito del prestatario desde que la exposición fue reconocida por primera vez. La previsión por la pérdida para las exposiciones que no se hayan incrementado de manera importante en el riesgo de crédito se basan en las pérdidas esperadas a 12 meses,

mientras que la previsión para las exposiciones que hayan sufrido un incremento importante en el riesgo de crédito se basan en las pérdidas llamadas "Life Time" o durante el tiempo de vida de la operación.

Tanto la PD, como también la EAD, el LGD y el efecto del descuento, reflejan la vida esperada o el período de exposición, siendo que la entidad financiera calcula cada uno de esos componentes para una serie de intervalos de tiempo durante el período de exposición, el cual puede ser mensual, trimestral o anualmente, por ejemplo, y los suma para derivar luego la pérdida esperada crediticia para toda la vida del instrumento.

Se calcula la EAD como:

$$EAD = Saldo + CCF * Max[(Límite - Saldo),0]$$

Donde:

- Saldo: Monto adeudado.
- Límite: Saldo disponible para operar
- CCF: Factor de conversión crediticia

Se le asignarán los factores de conversión a las exposiciones pertenecientes a productos como tarjetas de créditos y acuerdos en cuentas corrientes. Respecto de este factor, la normativa limita la estimación de futuras aceleraciones en el uso de este tipo de productos, su límite de crédito convenido en el contrato.

Además, la norma hace mención a líneas de crédito automáticamente renovables, como tarjetas de crédito y sobregiros, las cuales pueden ser retiradas contractualmente con sólo un día de notificación por el prestamista. Pero, en la práctica los prestamistas continúan prolongando el crédito por periodos mayores y solo podrán retirar el servicio después de que el riesgo crediticio del prestatario se incremente. Esto último puede influir negativamente en el cálculo de las pérdidas esperadas, ya que se notifica tarde esta información.

Estos instrumentos generalmente se caracterizan por no tener una condición fijada o estructura de reembolso y, habitualmente tienen un periodo de cancelación contractual corto; por la capacidad contractual para cancelar el contrato no se hace cumplir en la gestión normal del día a día del instrumento financiero y el contrato puede solo cancelarse cuando la entidad pasa a ser consciente de un incremento en el riesgo crediticio a nivel del servicio; y por gesionarse sobre una base colectiva.

1.3 Algunas definiciones de interés

En primer lugar, resulta interesante definir claramente a lo que se llama "Incumplimiento", a lo cual, la normativa NIIF 9 establece que cuando se define incumplimiento para determinar el riesgo de que ocurra un incumplimiento, cada entidad aplicará una definición de incumplimiento congruente con la definición utilizada a efectos de gestión del riesgo crediticio interno para el instrumento financiero relevante y considerará indicadores cualitativos tales como pactos financieros, cuando sea apropiado. Y agrega "sin embargo, hay una presunción refutable de que un incumplimiento no ocurrirá después de que un activo financiero esté en mora 90 días, a menos que una entidad tenga información razonable y sustentable que un criterio de incumplimiento más aislado es más apropiado (…)", y por esto se ha determinado que el incumplimiento para la cartera de créditos se materializa cuando los activos estén con al menos 90 días de mora.

En segunda instancia, la normativa NIIF 9 sugiere que el deterioro se reconozca en **tres etapas**, dentro de las cuales estarán en stage 1 aquellor activos cuya calidad crediticia no se ha deteriorado significativamente desde el reconocimiento inicial y en los cuales no hay indicios de incremento significativo en el riesgo crediticio; en stage 2 activos en los cuales se verifica un incremento significativo en el riesgo y, por ende, con empeoramiento significativo de su calidad crediticia, pero todavía sin evidencia objetiva de evento de deterioro, y por último en el stage 3 se encuentran activos con evidencia de deterioro a la fecha de reporte

Para esto, el Banco Central de la República Argentina tiene establecidas situaciones de deudores, basándose en días de atraso, que se requiere sean considerados a la hora de establecer los stages, esto es: Situacion Normal para aquellas operaciones cuyo atraso en el pago que no supere los 31 días, riesgo bajo cuando el atraso en el pago es de más de 31 y

hasta 90 días desde el vencimiento, riesgo medio para operaciones con atraso en el pago de más de 90 y hasta 180 días, riesgo alto si el atraso en el pago de más de 180 días hasta un año. Una vez superadas esas instancias, se entra dentro de lo que son cuentas irrecuperables, y se puede distinguir entre irrecuperables con atrasos superiores a un año e irrecuperables por disposición técnica cuando se trate de una deuda con una exentidad.

Como se mencionó, la norma NIIF 9, requiere que las previsiones por pérdida reflejen:

- Pérdidas esperadas a 12 meses sobre la cartera de créditos, es decir aquellas pérdidas esperadas que resulten de los eventos de incumplimiento del crédito que sean posibles dentro de los 12 meses siguientes a la fecha de presentación de reporte
- Pérdidas esperadas durante toda la vida sobre la cartera de crédito, es decir pérdidas esperadas que resulten de todos los posibles eventos de incumplimiento durante la vida de la operación.

A fines prácticos, la norma NIIF 9 permite que la entidad asuma que el riesgo de crédito de una operación no se ha incrementado de manera significativa, si se determina que el riesgo de crédito es “bajo” a la fecha de presentación de reporte, considerándose riesgo “bajo” si hay un riesgo bajo de incumplimiento, esto ocurre cuando el prestatario tiene una capacidad fuerte para satisfacer sus obligaciones de flujos de efectivo contractuales en el corto plazo y cambios adversos en las condiciones económicas y de negocio en el largo plazo no necesariamente impactarían en una reducción de la capacidad del prestatario para cumplir sus obligaciones relacionadas con los flujos de efectivo contractuales.

De esta manera, quedan definidos los tres stages de evaluación del riesgo crediticio mencionados anteriormente, y dentro de los cuales se sugiere que los créditos sean clasificados, según la clasificación de riesgos regulatorias o a cualquier otro indicador que demuestre evidencia de un aumento significativo del riesgo de una operación o cliente, como puede ser, por ejemplo, por medio del aumento de la probabilidad de default del individuo o empresa.

Respecto de la definición de **Pérdida** Esperada, la norma NIIF 9 define las pérdidas crediticias esperadas para los compromisos de préstamo sin utilizar:

"Para compromisos de préstamo sin utilizar, una pérdida crediticia es el valor presente de la diferencia entre flujos de efectivo contractuales que se deben a la entidad si el tenedor del compromiso de préstamo dispone del préstamo (EAD); y los flujos de efectivo que la entidad espera recibir si dispone del préstamo, es decir EAD * (1 – PD * LGD)^n

De aquí surge la expresión:

$$Pérdida\ Esperada\ Crediticia = EAD - EAD * (1 - PD * LGD)$$

Luego:

$$Pérdida\ Esperada\ Crediticia = EAD * [1 - (1 - PD * LGD)] = EAD * PD * LGD$$

Una vez obtenido el cálculo de la pérdida esperada, se puede desarrollar un modelo Forward Looking que va a arrojar un valor de pérdida esperada a un plazo de tiempo "t" de cada operación, obtenido a partir de condicionar las PD en cada instante, esto es, para el primer período, la probabilidad de default va a ser la PD calculada:

$$PD_1 = PD$$

Para el período siguiente, se considera la probabilidad de que esta operación no haya entrado en default el período anterior, pero sí lo haga en este período, es decir:

$$PD_2 = (1 - PD) * PD$$

Y esto se realizará para los "t" períodos que dure la operación.

Una vez realizado este cálculo, se actualizará el valor de pérdida esperada obtenido a una tasa efectiva de origen, por la cantidad de períodos "t" que se haya calculado la misma. Esto hará bajar el valor bruto obtenido originalmente, pero es el valor real que tiene al día de valuación. Para la aplicación de este descuento financiero, se suele considerar la distribución uniforme de default y se considera actualización a un período "t + ½".

Lo interesante de la propuesta que se hace es poder automatizar el calculo que la PD y de esta manera poder utilizar un único código de R o SQL por ejemplo, que sirva para generalizar este cálculo y el de todas las variables, además de poder automatizarse un cálculo de sensibilidad. Actualmente R ofrece un amplio conjunto de librería para el desarrollo de software de riesgo, pudiendo integrarse con una inmensa cantidad de tipos de base de datos del mercado tanto públicas como privadas permitiendo construir aplicaciones donde se tenga controlado en todo momento la gestión del dato.

Conclusión

Se pudo visualizar la dificultad que tiene hoy en día el manejo de grandes volúmenes de datos, derivado de cuestiones de consistencia e integralidad de los mismos, producción de bases representativas, normalizadas y sin errores. Además de sumarse a estos inconvenientes, el problema de la dificultad que trae la obtención de datos en sí misma, causado por normativas de protección de datos personales principalmente.

Luego de analizarse la necesidad que tienen muchas compañías recuperadas de conseguir fondos para poder auto gestionar la producción, pero sin contar con elementos que reflejen solidez en cuanto a riesgo crediticio, se ha creado un modelo de aprendizaje automático que permitió obtener un valor predictivo de la probabilidad de incumplimiento de pago de los créditos que éstos solicitarían.

Es posible entonces, a partir de una base de datos lo suficientemente grande y con consistencia e integralidad de datos, obtener un valor de tasa que de información concreta acerca del comportamiento de cada cliente, dentro de cada operación.
Además, se pueden llegar a utilizar otras técnicas o algoritmos para mejorar el poder de predicción.
La principal dificultad que esta base presentó es la gran diferencia entre quienes cumplieron con sus pagos y quienes no durante el transcurso de los 8 años de datos.

Se destaca, además, la gran utilidad que esta predicción tiene para el cálculo de variables de gran importancia para la gestión y valuación de organizaciones, como es el caso del cálculo de montos de pérdidas esperadas, como también puede ser los valores de garantías y recupero crediticio, o como puede ser en términos de organización los costos de créditos.

Además, pueden aplicarse otras herramientas de aprendizaje automático para realizar diferentes tipos de consultas dentro de la base que puedan facilitar la obtención de determinados listados de datos filtrados para poder analizar bases intermedias que pudieran precisarse en algún proceso de utilización de la predicción utilizada.

Referencias bibliográficas

- Calderón, M. C. L. (2001). Los microcréditos: un nuevo instrumento de financiación para luchar contra la pobreza. *Revista de economía mundial*, *5*.

- Civil, D. H. I. A. Experiencias de incidencia en política pública de organizaciones auto-gestionadas: desde la autonomía como proyecto y hacia la democracia como régimen de sentido.
- Del Castillo Sánchez, L., & Rodríguez, J. M. P. (2019). Desarrollo Local y Microcrédito. *Revista Economía y Desarrollo (Impresa)*, *138*(2).
- Estrella, G., & Iván, Á. (2019). Validación del Procedimiento de Colocación de Operaciones de Microcrédito en la Cooperativa de Ahorro y Crédito Policia Nacional. (Master's thesis).
- Freyre, H. F. (2019). Finanzas solidarias y la teoría de los microcréditos.
- Gillis, T. B., & Spiess, J. L. (2019). Big Data and Discrimination. The University of Chicago Law Review, 86(2), 459-488. 2006
- Kiviat, B. (2017). The art of deciding with data: evidence from how employers translate credit reports into hiring decisions. *Socio-Economic Review*.
- Nieto, B. G. (2003). Microcrédito y desarrollo local. *Acciones e investigaciones sociales*, (18), 115-128.
- Noyes, K. (2015). Will big data help end discrimination—or make it worse? Fortune.
- Pinto, S., & Litman, L. (2019). Informe: Una aproximación cuantitativa a las cooperativas de trabajo desde las finanzas para la autogestión. Observatorio Social sobre Empresas Recuperadas y Autogestionadas, (14 (2019)).
- Verbeke, C. O. (2007). Las finanzas y la Economía Social. Experiencias Argentinas. Buenos Aires: Altamira.
- Heller, C. (2006). Rol de la economía social para un nuevo modelo de país. Revista del Instituto de la Cooperación. Nº 169, Buenos Aires.
- IFRS Foundation (2014). Normativa NIIF 9 Instrumentos Financieros.

Anexo I

En este anexo se describen las distintas variables de la base de datos original y su significado:

Variables	Descripción de la variable
acc_now_delinq	Número de cuentas en las cuales el prestatario no es moroso
acc_open_past_24mths	Número de operaciones abiertas en los últimos 24 meses.
acceptD	Fecha en la que el prestatario acepta la oferta
addr_state	El estado proporcionado por el prestatario en la solicitud de préstamo.
all_util	Saldo límite de crédito en todas las operaciones.
annual_inc	Ingreso anual informado por el prestatario durante el registro.
annual_inc_joint	Ingreso anual combinado autoinformado proporcionado por los co-prestatarios durante el registro
application_type	Indica si el préstamo es una solicitud individual o una solicitud conjunta con dos co-prestatarios
avg_cur_bal	Saldo promedio actual de todas las cuentas
bc_open_to_buy	Total abierto para comprar en tarjetas bancarias giratorias.
bc_util	Relación entre el saldo actual total y el límite de crédito / crédito alto para todas las cuentas de tarjetas bancarias
chargeoff_within_12_mths	Número de cancelaciones dentro de 12 meses
collection_recovery_fee	Cargo posterior por cobro
collections_12_mths_ex_med	Número de colecciones en 12 meses excluyendo colecciones médicas
creditPullD	La fecha en que Loan Club obtuvo el crédito para este préstamo.
debt_settlement_flag	Señala si el prestatario, que ha cancelado o no, está trabajando con una compañía de liquidación de deudas.
debt_settlement_flag_date	La fecha más reciente en que se ha establecido el indicador de liquidación de deudas
deferral_term	Cantidad de meses que se espera que el prestatario pague menos que el monto de pago mensual contractual debido a un plan por dificultades económicas
delinq2Yrs	La cantidad de incidencias de morosidad de más de 30 días en el archivo de crédito del prestatario durante los últimos 2 años
delinqAmnt	Monto adeudado adeudado por las cuentas en las que el prestatario está ahora en mora.
Desc	Descripción del préstamo proporcionado por el prestatario

disbursement_method	Método por el cual el prestatario recibe su préstamo. Los valores posibles son: CASH, DIRECT_PAY
Dti	Relación calculada utilizando los pagos mensuales totales de la deuda del prestatario sobre el total de las obligaciones de la deuda, excluyendo la hipoteca y el préstamo LC solicitado, dividida por el ingreso mensual autoinformado del prestatario.
dti_joint	Relación calculada utilizando los pagos mensuales totales de los co-prestatarios sobre el total de las obligaciones de la deuda, excluyendo las hipotecas y el préstamo LC solicitado, dividida por el ingreso mensual combinado de los co-prestatarios.
earliest_cr_line	Fecha en que se abrió la primera línea de crédito del prestatario
effective_int_rate	Tasa de interés efectiva es igual a la tasa de interés en una Nota reducida por la estimación de Lending Club del impacto de los intereses no cobrados antes del cobro.
emp_length	Duración del empleo en años. Los valores posibles están entre 0 y 10, donde 0 significa menos de un año y 10 significa diez o más años.
emp_title	Título del trabajo proporcionado por el Prestatario al solicitar el préstamo. *
empLength	Duración del empleo en años. Los valores posibles están entre 0 y 10, donde 0 significa menos de un año y 10 significa diez o más años.
expD	Fecha en que expirará el listado
expDefaultRate	Tasa de incumplimiento esperada del préstamo.
fico_range_high	Rango del límite superior al que pertenece el FICO del prestatario al originarse el préstamo.
fico_range_low	Rango del límite inferior al que pertenece el FICO del prestatario al originarse el préstamo.
funded_amnt	Monto total comprometido con ese préstamo en ese momento.
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
Grade	Grado de préstamo asignado por Loan Club
hardship_amount	El pago de intereses que el prestatario se ha comprometido a realizar cada mes mientras está en un plan de dificultades
hardship_dpd	Días de cuenta vencidos a partir de la fecha de inicio del plan de dificultades
hardship_end_date	La fecha de finalización del período de plan de dificultades
hardship_flag	Señala si el prestatario está o no en un plan de dificultades
hardship_last_payment_amount	El último monto de pago a partir de la fecha de inicio del plan de dificultades

hardship_length	Cantidad de meses que el prestatario realizará pagos más pequeños de lo que normalmente está obligado debido a un plan de dificultades
hardship_loan_status	Estado del préstamo a partir de la fecha de inicio del plan de dificultades
hardship_payoff_balance_amount	El monto del saldo de liquidación a partir de la fecha de inicio del plan de dificultades
hardship_reason	Describe la razón por la que se ofreció el plan de dificultades
hardship_start_date	La fecha de inicio del período de plan de dificultades
hardship_status	La descripción del plan de dificultades está activo, pendiente, cancelado, completado o interrumpido
hardship_type	Describe la oferta del plan de adversidades
home_ownership	Estado de propiedad de la vivienda proporcionado por el prestatario durante el registro. Nuestros valores son: ALQUILER, PROPIO, HIPOTECA, OTROS.
Id	Identificación única asignada de Loan Club para la lista de préstamos.
il_util	Relación entre el saldo actual total y el límite de crédito / crédito alto en todas las cuentas de instalación
ils_exp_d	Fecha de vencimiento de la plataforma
initial_list_status	Estado de listado inicial del préstamo. Los valores posibles son - W, F
inq_fi	Número de consultas de finanzas personales
inq_last_12m	Número de consultas de crédito en los últimos 12 meses
inq_last_6Mths	Número de consultas en los últimos 6 meses (excluyendo las consultas de automóviles e hipotecas)
Installment	Pago mensual adeudado por el prestatario si el préstamo se origina
intRate	Tasa de interés sobre el préstamo
isIncV	Indica si el ingreso fue verificado por Loan Club, no verificado o si la fuente de ingreso fue verificada
issue_d	El mes en que se financió el préstamo.
last_credit_pull_d	El mes más reciente que Loan Club obtuvo crédito para este préstamo
last_fico_range_high	Límite superior del rango al que pertenece el último FICO del prestatario pertenece.
last_fico_range_low	Límite inferior del rango al que pertenece el último FICO del prestatario pertenece.
last_pymnt_amnt	Last total payment amount received
last_pymnt_d	Last month payment was received
listD	Fecha en que la solicitud del prestatario fue listada en la plataforma.
loan_amnt	Importe indicado del préstamo solicitado por el prestatario. Si en algún momento, el departamento de crédito reduce el monto del préstamo, se reflejará en este valor.

loan_status	Estado actual del préstamo
max_bal_bc	Saldo actual máximo adeudado en todas las cuentas revolventes
memberId	Identificación única asignada de Loan Club para el miembro prestatario
mo_sin_old_il_acct	Meses desde la apertura de la cuenta bancaria más antigua
mo_sin_old_rev_tl_op	Meses desde que se abrió la cuenta giratoria más antigua
mo_sin_rcnt_rev_tl_op	Meses desde la apertura de la cuenta revolvente más reciente
mo_sin_rcnt_tl	Meses desde que se abrió la cuenta más reciente
mortAcc	Número de cuentas hipotecarias
Msa	Área estadística metropolitana del prestatario
mths_since_last_delinq	Número de meses desde la última morosidad del prestatario.
mths_since_last_major_derog	Meses desde la última calificación de 90 días o peor
mths_since_last_record	Número de meses desde el último registro público
mths_since_most_recentInq	Meses desde la última consulta
mths_since_oldest_il_open	Meses desde la apertura de la cuenta bancaria más antigua
mths_since_rcnt_il	Meses desde la apertura de las cuentas a plazos más recientes
mths_since_recent_bc	Meses desde que se abrió la cuenta Bankcard más reciente
mths_since_recent_bc_dlq	Meses desde la última morosidad de tarjetas bancarias
mths_since_recent_inq	Meses desde la última consulta
mths_since_recent_revol_delinq	Meses desde la última morosidad rotatoria
mthsSinceRecentLoanDelinq	Meses desde la última morosidad de las finanzas personales
next_pymnt_d	Próxima fecha de pago programada
num_accts_ever_120_pd	Número de cuentas cada 120 o más días vencidos
num_actv_bc_tl	Número de cuentas bancarias actualmente activas
num_actv_rev_tl	Número de operaciones giratorias actualmente activas
num_bc_sats	Número de cuentas bancarias satisfactorias
num_bc_tl	Número de cuentas bancarias
num_il_tl	Número de cuentas a plazos
num_op_rev_tl	Número de cuentas rotativas abiertas
num_rev_accts	Número de cuentas rotativas
num_rev_tl_bal_gt_0	Número de operaciones revolventes con saldo mayor a 0
num_sats	Número de cuentas satisfactorias
num_tl_120dpd_2m	Número de cuentas actualmente 120 días vencidos (actualizado en los últimos 2 meses)
num_tl_30dpd	Número de cuentas actualmente vencidas a 30 días (actualizadas en los últimos 2 meses)

num_tl_90g_dpd_24m	Número de cuentas con 90 o más días de vencimiento en los últimos 24 meses
num_tl_op_past_12m	Número de cuentas abiertas en los últimos 12 meses
open_acc	Número de líneas de crédito abiertas en el archivo de crédito del prestatario
open_acc_6m	Número de operaciones abiertas en los últimos 6 meses
open_act_il	Número de operaciones de pago actualmente activas
open_il_12m	Número de cuentas a plazos abiertas en los últimos 12 meses
open_il_24m	Número de cuentas a plazos abiertas en los últimos 24 meses
open_rv_12m	Número de operaciones rotativas abiertas en los últimos 12 meses
open_rv_24m	Número de operaciones revolventes abiertas en los últimos 24 meses
orig_projected_additional_accrued_interest	Monto de interés adicional original proyectado que se acumulará para el plan de pago por dificultades económicas dado a partir de la Fecha de inicio de dificultades. Este campo será nulo si el prestatario ha roto su plan de pago por dificultades económicas.
out_prncp	Principal pendiente restante por importe total financiado
out_prncp_inv	Principal pendiente restante por parte del monto total financiado por los inversionistas
payment_plan_start_date	Día en que vence el primer pago del plan por dificultades. Por ejemplo, si un prestatario tiene un período de plan de dificultades de 3 meses, la fecha de inicio es el comienzo del período de tres meses en el que el prestatario puede realizar pagos de solo intereses.
pct_tl_nvr_dlq	Porcentaje de operaciones nunca en mora
percent_bc_gt_75	Porcentaje de todas las cuentas de tarjetas bancarias mayores al 75% del límite
policy_code	Código de política públicamente disponible = 1 Código de política de nuevos productos no disponible públicamente = 2
pub_rec	Número de registros públicos derogatorios
pub_rec_bankruptcies	Número de quiebras de registros públicos
pubRec	Número de registros públicos derogatorios
Purpose	Categoría proporcionada por el prestatario para la solicitud de préstamo.
pymnt_plan	Indica si se ha establecido un plan de pago para el préstamo
Recoveries	Cargo posterior a la recuperación bruta
reviewStatus	Estado del préstamo durante el período de cotización. Valores: APROBADO, NO_APPROBADO.
reviewStatusD	Fecha en que la solicitud de préstamo fue revisada por Loan Club

revol_bal	Balance rotatorio de crédito total
revol_bal_joint	Suma del saldo de crédito renovable de los co-prestatarios, neto de saldos duplicados
revol_util	Tasa de utilización de la línea rotativa, o la cantidad de crédito que el prestatario está utilizando en relación con todo el crédito rotatorio disponible.
sec_app_chargeoff_within_12_mths	Número de cancelaciones dentro de los últimos 12 meses en el momento de la solicitud para el solicitante secundario
sec_app_collections_12_mths_ex_med	Número de colecciones dentro de los últimos 12 meses, excluyendo las colecciones médicas al momento de la solicitud para el solicitante secundario
sec_app_earliest_cr_line	Primera línea de crédito en el momento de la solicitud para el solicitante secundario
sec_app_fico_range_high	Rango FICO (bajo) para el solicitante secundario
sec_app_fico_range_low	Rango FICO (alto) para el solicitante secundario
sec_app_inq_last_6mths	Consultas de crédito en los últimos 6 meses al momento de la solicitud para el solicitante secundario
sec_app_mort_acc	Número de cuentas hipotecarias al momento de la solicitud para el solicitante secundario
sec_app_mths_since_last_major_derog	Meses desde la última calificación de 90 días o peor al momento de la solicitud para el solicitante secundario
sec_app_num_rev_accts	Número de cuentas rotativas en el momento de la solicitud para el solicitante secundario
sec_app_open_acc	Número de operaciones abiertas en el momento de la solicitud para el solicitante secundario
sec_app_open_act_il	Número de operaciones a plazos actualmente activas en el momento de la solicitud para el solicitante secundario
sec_app_revol_util	Relación entre el saldo actual total y el límite de crédito / crédito alto para todas las cuentas revolventes
serviceFeeRate	Tarifa de servicio pagada por el inversionista para este préstamo.
settlement_amount	El monto del préstamo que el prestatario ha acordado para pagar
settlement_date	La fecha en que el prestatario acuerda el plan de liquidación.
settlement_percentage	El monto de la liquidación como un porcentaje del saldo de liquidación del préstamo
settlement_status	El estado del plan de liquidación del prestatario. Los valores posibles son: COMPLETO, ACTIVO, ROTO, CANCELADO, DENEGADO, BORRADOR
settlement_term	La cantidad de meses que el prestatario estará en el plan de liquidación
sub_grade	Subgrado de préstamo asignado
tax_liens	Número de gravámenes fiscales
Term	Número de pagos del préstamo. Los valores están en meses y pueden ser 36 o 60.

Title	Título del préstamo proporcionado por el prestatario
tot_coll_amt	Montos totales de cobro jamás adeudados
tot_cur_bal	Saldo total actual de todas las cuentas
tot_hi_cred_lim	Total alto crédito / límite de crédito
total_acc	Número total de líneas de crédito actualmente en el archivo de crédito del prestatario
total_bal_ex_mort	Saldo de crédito total excluyendo hipoteca
total_bal_il	Saldo total actual de todas las cuentas a plazos
total_bc_limit	Total límite superior de crédito tarjetas bancarias / límite de crédito
total_cu_tl	Número de operaciones financieras
total_il_high_credit_limit	Cuota total límite susperior de crédito / límite de crédito
total_pymnt	Pagos recibidos hasta la fecha por el monto total financiado
total_pymnt_inv	Pagos recibidos hasta la fecha por parte del monto total financiado por los inversionistas
total_rec_int	Intereses recibidos hasta la fecha
total_rec_late_fee	Cargos atrasados recibidos hasta la fecha
total_rec_prncp	Principal recibido hasta la fecha
total_rev_hi_lim	Límite superior de crédito rotatorio total / límite de crédito
totalAcc	Número total de líneas de crédito actualmente en el archivo de crédito del prestatario
url	URL para la página de Loan Club con datos de listado.
verification_status	Indica si el ingreso fue verificado por Loan Club, no verificado o si la fuente de ingreso fue verificada
verified_status_joint	Indica si el ingreso conjunto de los co-prestatarios fue verificado por LC, no verificado o si la fuente de ingreso fue verificada
zip_code	Primeros 3 números del código postal proporcionado por el prestatario en la solicitud de préstamo

Apéndice I: Preparación de la base de datos

A continuación, se detallan las modificaciones realizadas sobre la base de datos original a fines de mejorarla y obtener una predicción lo más exacta posible:

1. Columnas modificadas: "loan_status", se borran todas las filas menos aquellas que contienen "Fully paid", "Default" y "Charged Off". Luego se fusionan 'Charged off' y 'Default' juntos, lo que significa que cualquier persona que caiga en esta categoría incumplirá con su préstamo. Luego de esto, la base resultante que tiene en la columna de estado del préstamo solamente "Fully Paid" y "Default" con 425.834 datos. A la columna "term" se quita el texto para dejar solo datos numéricos.

2. Columnas creadas: tiempo transcurrido al mes de mayo de 2019 "days_dince_earliest_cr_line" "days_since_issue", "days_to_next_pymnt", "days_since_last_pymnt", "days_since_last_credit_pull", "days_since_sec_app_earliest_cr_line", "days_from_hardship_start", "days_to_hardship_end", "days_from_payment_plan_start", "days_ of_settlement".

3. Dummies (0;1): Hardship_flag, pymnt_plan, emp_length , hardship_loan_status, hardship_reason

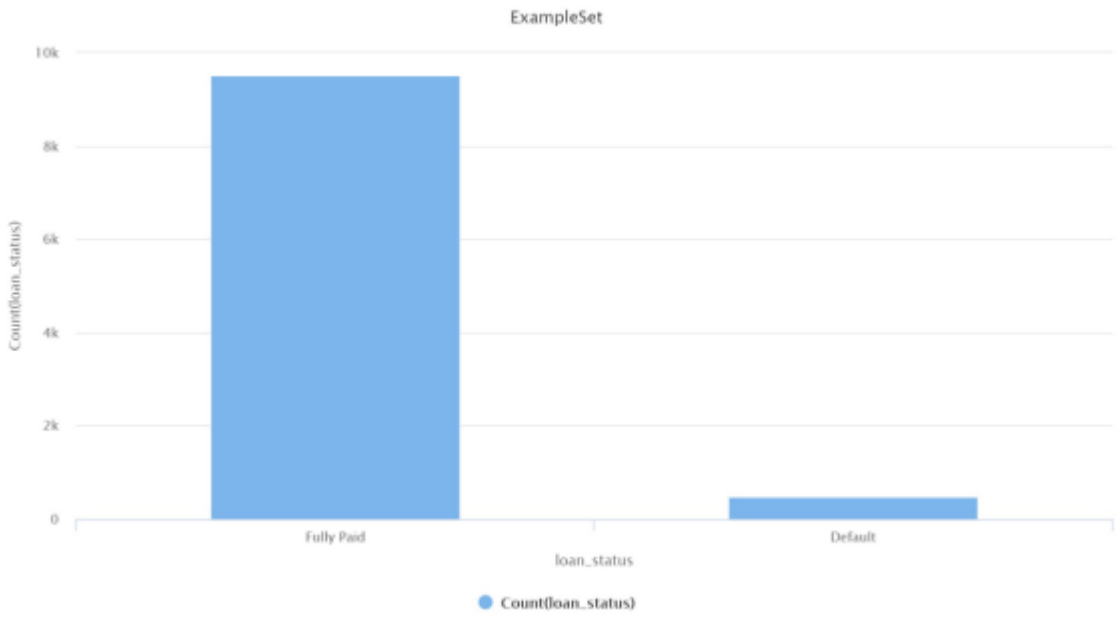
4. Columnas eliminadas: "zip_code","desc", "URL" "member_id","hardship_type", "issue_d", "next_pymnt_d","last_pymnt_d", "last_credit_pull_d", "sec_app_earliest_cr_line", "hardship_start_date", "hardship_end_date", "payment_plan_start_date", "settlement_date"

5. Columnas Imputadas: "All_util", "Annual_inc_joint", "Avg_cur_val", "Bc_open_to_buy", "Bc_util", "Days_of_settlement", "Dti", "Dti_joint", "Il_util", "Mo_sin_old_il_act", "Mths_since_last_delinq", "Mths_since_last_major_derog", ths_since_last_record", "Mths_since_rcnt_il", "Mths_since_recent_bc", "Mths_since_recent_bc_dlq", "Mths_since_recent_inq", "Mths_since_recent_revol_delinq", "Num_tl_120dpd_2m", "Percent_bc_gt_75",

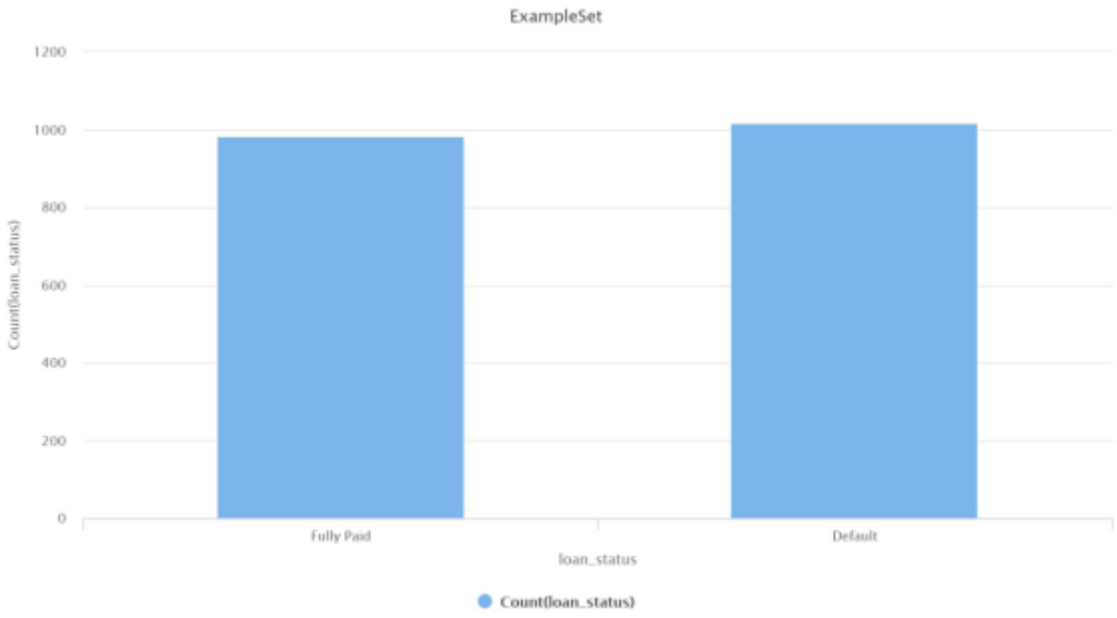
"Revol_bal_joint", "Revol_util", "Sec_app_chargeoff_within_12_mths", "Sec_app_collectiones_12_mths_ex_med", "Sec_app_inq_last_6mths", "Sec_app_mort_acc", "Sec_app_mth_since_last_major_derog", "Sec_app_num_rev_accts", "Sec_app_open_acc", "Sec_app_open_act_il", "Sec_app_revol_util".

Para mejorar la base, se toma una muestra de 100.000 datos y se ingresa en RapidMiner con el operador "Read Excel", al cual en primera instancia no se le harán cambios en los roles de las variables, pero se solicitará que se imputen los datos perdidos a partir del operador "Impute Missing Values". Siendo que este operador solo puede completar datos de tipo numéricos, es preciso transformar los datos nominales en numéricos con el operador "Nominal to Numerical". Este último genera nuevas columnas de aquellas con datos nominales vacíos y las completa con pruebas lógicas.

Pero esta base, reducida y ahora completa, se encuentra desbalanceada, siendo que se tienen 4,85% de "Default" y un 95,15% de "Fully Paid":



El problema que trae aparejado es que la mayoría de los algoritmos trabajan mejor cuando la base esta balanceada ya que la mayoría tratan de maximizar la exactitud y cuando se tiene una clase mayoritaria como en este caso, se predice la clase mayoritaria. Esto se resuelve de varias maneras, en este caso se utiliza el "sobremuestreo" también en RapidMiner de la siguiente forma: en primer lugar, se recurre al operador Generate Weight (Stratification) que, a partir del peso total de cada fila, partirá ese peso según la clase, y le asignará un peso a cada clase. Luego se combina con el operador Sample (Bootstrapping), que hace reemplazos tomando en cuenta esos pesos. Este operador toma la clase minoritaria y repite datos para agrandarla, entonces repite muchas veces lo que está en la clase minoritaria que, si bien duplica datos, hace que la clase se balancee. Toma muestras basándose en los pesos como probabilidad. La base obtenida muestra la siguiente relación:



Donde puede apreciarse que la proporción de cada clase es similar, siendo 49,10% Fully Paid y 50,85% Default, por lo cual se puede asegurar que la base se encuentra balanceada.

Apéndice II: Modelos utilizados

En este apéndice se describen los distintos modelos utilizados y los criterios correspondientes a cada caso:

Nombre	Modelo / Criterios
RF 1	Random Forest Árboles= 100. Maximal Deph= 10 Criterio= Accuracy
RF 2	Random Forest Árboles= 100. Maximal Deph= 10 Criterio= Information Gain
RF 3	Random Forest Árboles= 100. Maximal Deph= 10 Criterio= Gain Ratio
DT 1	Decision Tree Maximal Deph= 10 Criterio= Accuracy
DT 2	Decision Tree Maximal Deph= 10 Criterio= Information Gain
DT 3	Decision Tree Maximal Deph= 10 Criterio= Gain Ratio
Stacking 1	Stacking Random Forest. Árboles= 100 Maximal Deph= 10 Criterio= Gain Ratio Decision Tree Criterio= Gain Ratio k-NN k= 5
Stacking 2	Stacking Random Forest. Árboles= 100 Maximal Deph= 10 Criterio= Information Gain Decision Tree Criterio= Information Gain k-NN k= 5

Stacking 3	Stacking Random Forest. Árboles= 100 Maximal Deph= 10 Criterio= Accuracy Decision Tree Criterio= Accuracy k-NN k= 5
Stacking 4	Stacking Random Forest. Árboles= 100 Maximal Deph= 10 Criterio= Information Gain Decision Tree Criterio= Information Gain Deep Learning Criterio= Tanh Hidden Layer Sizes= 50/50
Stacking 5	Stacking Random Forest. Árboles= 100 Maximal Deph= 10 Criterio= Information Gain Decision Tree Criterio= Information Gain Deep Learning Criterio= Tanh Hidden Layer Sizes= 100/100
Regresión Logística	Logistic Regression solver= AUTO standardize add intercept compute p-values remove collinear columns missing values habdling MeanImputation max iterations= 0 max runtime seconds= 0

RL SMV	Logistic Regression (SMV) Kernel Type= dot Kernel cache= 200 C= 1.0 Convergence Epsilon= 0.001 max iterations= 100.000
Rule Induction	Rule Induction Criterio= Information Gain Sample Ratio= 0.9 Pureness= 0.9 Minimal Prune Benefit= 0.25

Resultados de los Modelos

Modelo	Resultado Accuracy	Resultado AUC ROC
RF 1	accuracy: 95.67% +/- 0.49% (micro average: 95.67%)	AUC: 0.936 +/- 0.057 (micro average: 0.936) (positive class: Default)
RF 2	accuracy: 97.00% +/- 1.67% (micro average: 97.00%)	AUC: 0.987 +/- 0.011 (micro average: 0.987) (positive class: Default)
RF 3	accuracy: 98.40% +/- 0.69% (micro average: 98.40%)	AUC: 0.995 +/- 0.003 (micro average: 0.995) (positive class: Default)
DT 1	accuracy: 99.38% +/- 0.27% (micro average: 99.38%)	AUC: 0.959 +/- 0.007 (micro average: 0.959) (positive class: Default)
DT 2	accuracy: 99.99% +/- 0.01% (micro average: 99.99%)	**AUC: 1.000 +/- 0.001 (micro average: 1.000) (positive class: Default)**
DT 3	accuracy: 99.37% +/- 0.40% (micro average: 99.37%)	AUC: 0.956 +/- 0.006 (micro average: 0.956) (positive class: Default)
Stacking 1	accuracy: 97.81% +/- 0.22% (micro average: 97.81%)	AUC: 0.997 +/- 0.001 (micro average: 0.997) (positive class: Default)
Stacking 2	accuracy: 97.88% +/- 0.19% (micro average: 97.88%)	AUC: 0.998 +/- 0.001 (micro average: 0.998) (positive class: Default)

Stacking 3	accuracy: 97.78% +/- 0.20% (micro average: 97.78%)	AUC: 0.997 +/- 0.001 (micro average: 0.997) (positive class: Default)
Stacking 4	accuracy: 97.79% +/- 0.20% (micro average: 97.79%)	AUC: 0.997 +/- 0.001 (micro average: 0.997) (positive class: Default)
Stacking 5	accuracy: 97.91% +/- 0.20% (micro average: 97.91%)	AUC: 0.997 +/- 0.001 (micro average: 0.997) (positive class: Default)
Regresión Logística	accuracy: 99.76% +/- 0.16% (micro average: 99.76%)	AUC: 0.985 +/- 0.005 (micro average: 0.985) (positive class: Default)
RL SMV	accuracy: 99.65% +/- 0.08% (micro average: 99.65%)	AUC: 0.998 +/- 0.001 (micro average: 0.998) (positive class: Default)
Rule Induction	accuracy: 95.33% +/- 0.01% (micro average: 95.33%)	AUC: 0.500 +/- 0.000 (micro average: 0.500) (positive class: Default)