Universidad del Desarrollo
Facultad de Ingeniería

EVITEMOS EL CASTIGO

Modelo de predicción de operaciones de crédito con posible default financiero

POR: CRISTÓBAL EULUFI MALBRÁN

Proyecto de grado presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESOR GUÍA:

DR. LEO FERRES

Diciembre 2020

SANTIAGO

AGRADECIMIENTO

A mi señora Rocío y a mi hija Renata por todo el apoyo y la paciencia en este gran desafío que me impuse. También a mi papá Renato y mi hermano Matías quienes me ayudaron con la revisión del documento.

TABLA DE CONTENIDO

Resumen

Las empresas del rubro financiero obtienen sus ingresos en base a los intereses percibidos de los créditos otorgados a sus clientes, por consiguiente, cuando un cliente no cancela las cuotas de su crédito, la institución asume la pérdida del capital prestado (castigo), impactando negativamente en sus utilidades.

Coopeuch, institución financiera cooperativa, minimiza el riesgo de castigo de los créditos mediante la modalidad de descuento por planilla, sin embargo, a pesar de esta metodología, tiene pérdidas generadas por el no pago de los créditos equivalente al 30% de su margen financiero, lo cual afecta en el remanente entregado a sus socios.

El objetivo de este paper es comparar modelos de marchine learning y ver cuál de ellos predice de mejor manera que operación de crédito se castigará para generar estrategias de cobranza y reducir las pérdidas por no pago.

Se estudiaron tres modelos puntuales, regresión logística, random forest y extreme gradiente boosting (XGBoost), modelos más utilizados en estudios similares realizados por otros autores, donde se pensaba que el modelo XGBoost tendría mejor performance mediante el algoritmo de descenso de gradiente.

El modelo que mejor resultado entrega al estudio fue Random Forest, con un accuracy de 95% y un ROC-AUC de 98,9%.

1. Introducción

El rubro financiero es un mercado donde existe gran cantidad de participantes: se destacan principalmente los bancos, pero en la actualidad también es posible identificar otras instituciones como cajas de compensación, cooperativas de ahorro y crédito, etc. Estas últimas se han posicionado de manera relevante en el mercado de créditos de consumo. Todos estos actores buscan a través de la oferta de créditos incrementar sus ingresos con el objetivo de maximizar sus utilidades.

Las empresas financieras, generan sus ingresos en base a los intereses percibidos principalmente de los créditos de consumo otorgados a cada cliente, por lo cual, entre más créditos cursen mensualmente, mayor serán los ingresos que tendrán las respectivas instituciones.

Sin embargo, todas las instituciones financieras del mercado buscan el mismo objetivo: maximizar sus ingresos otorgando créditos de consumo a las personas que los necesitan. Es por esta razón que cada participante del mercado analiza los posibles prospectos (clientes) para ofrecerle las mejores condiciones crediticias, de tal forma que estos, adquieran un nuevo crédito de consumo con ellos o trasladen sus créditos que tienen en otras instituciones (compra de cartera).

En la actualidad la oferta crediticia es amplia y las personas tienen facilidad de endeudamiento, lo que genera en un número no menor de casos la imposibilidad de pago de las obligaciones financieras, repercutiendo en el castigo de los créditos

Para una empresa del rubro financiero, un crédito castigado es cuando un cliente no paga sus obligaciones financieras (valor de la cuota pactada) por un período de 180 días consecutivos (6 cuotas correlativas). Una vez cumplido este plazo, la institución

financiera asume la pérdida contable del capital prestado y los intereses por cobrar, afectando directamente las utilidades en el estado de resultado.

En relación con lo anterior es que las instituciones financieras han creado áreas especializadas donde evalúan el riesgo de cada uno de los clientes, analizando diferentes parámetros que les permitan determinar cuál de ellos cumple con las normativas internas para ser sujeto a crédito. Sin embargo, posterior al otorgamiento del crédito, las condiciones del cliente (laborales o de endeudamiento) pueden cambiar y así no poder cumplir con sus obligaciones de pago.

Coopeuch, institución financiera cooperativa, con más de 50 años de experiencia en el mercado financiero, se ha destacado en el apoyo de sus socios, pertenecientes principalmente a los grupos socioeconómicos C3, D y E, siendo estos dos últimos grupos son de riesgo para bancos dada la alta probabilidad de no pago de sus créditos a causa de la baja renta que estos perciben.

Este riesgo es minimizado por Coopeuch en base a la estructura de descuento por planilla de las cuotas mensuales de los créditos. Sin embargo, a pesar de este formato de negocio, algunos créditos no son pagados por los socios y la cooperativa tiene que asumir la pérdida del capital prestado, lo que impacta directamente en las utilidades que para el año 2019 fue aproximadamente MM$56.500, lo que representa un 30% del margen financiero y un 90% de las utilidades de ese período.

El poder reducir el monto de créditos castigados anualmente, beneficiaría a los socios de Coopeuch, ya que las cooperativas una vez al año reparten entre sus socios las utilidades obtenidas el período anterior, lo que generaría un aumento del remanente de cada socio.

Este paper propone una evaluación de tres diferentes modelos de Machine Learning para encontrar cual permite predecir qué clientes, en base a comportamientos de pago, deuda global, etc., no pagará su crédito de consumo de tal forma de generar estrategias de cobranza y ofertas atractivas para la contención del crédito. Los modelos que se compararon fueron: Regresión logística, Random Forest y Extreme Gradient Boosting (XGBoost).

Para la implementación de los modelos identificados anteriormente, se hizo uso de base de datos con todos los créditos vigentes con más de 30 días de mora y su comportamiento mensual desde octubre de 2018 a diciembre de 2019 identificando cuál de ellos fue castigada. Dentro de la información que se tiene, es posible identificar el monto del crédito al curse, las cuotas pactadas, la edad del socio, tasa de interés, cuotas pagadas, etc.

2. Trabajo Relacionado

El riesgo de default crediticio y la contención de la cartera vencida son temas de gran importancia en las empresas del rubro financiero, razón por la cual han sido estudiados ampliamente en diferentes publicaciones relacionadas con instituciones financieras. Como menciona Pazmiño en su estudio [1], el tener un manejo de la cartera vencida influye en la rentabilidad de una institución financiera. Así mismo, Vera [2] indica que es necesario un modelo de gestión adecuado para el cobro de la cartera vencida para optimizar el desarrollo de las actividades que realizan las personas que trabajan en un banco. Al igual que los autores mencionados anteriormente, Armijo y Oña [3], señalan en su estudio que un buen modelo de gestión permitiría incrementar los porcentajes de la recuperación de la cartera vencida, mientras que Caiza [4] agrega que un buen modelo de cobranza ayudaría a disminuir los índices de la cartera vencida.

Si bien, la elaboración de modelos de gestión para la contención de la cartera vencida es de real importancia, Alfaro, Gallardo y Stein [5] hacen referencia a los principales factores por los cuales los créditos de consumo e hipotecarios no son pagados, dentro de los cuales se identifica la edad del jefe de hogar, el número de personas que aportan con ingresos al hogar y el nivel educacional.

No obstante a lo anterior y en base a la disponibilidad de grandes volúmenes de información y capacidades de procesamiento de datos que actualmente permiten las herramientas informáticas, se han generado nuevas investigaciones de cómo predecir el incumplimiento financiero. Biron y Medina [6], comparan ocho algoritmos de clasificación para predecir el no pago de los créditos en el sistema financiero chileno, dentro de los modelos utilizados se identifican regresión logística, clasificador bayesiando, support vector machine, gradient boosting, entre otros, donde el algoritmo de gradient boosting machines y un ensamble heterogéneo lograron identificar de mejor

manera la variable objetivo con respecto al resto de los modelos estudiados. Siguiendo esta línea, I-Chen Yeh y Che-hui Lien [7] en su estudio, utilizan modelos de regresión logística, K-Nearest Neighbor, entre otros, para predecir el no pago de los créditos. Del mismo modo, Gurý y Gurný [8], realizan la comparación de tres modelos para predecir la probabilidad de incumplimiento de pagos de los créditos, donde el modelo de regresión logística es el más apropiado para determinar la probabilidad de incumplimiento en los bancos de Estados Unidos.

Sin embargo, no solo se han realizado comparaciones de modelos que predicen la probabilidad del incumplimiento. Celas y Cuencas [9] generan un modelo de machine learning en base a una regresión logística que predice el riesgo de no pago de los créditos. También, Trujillo [10] hace uso de técnicas de machine learning para la predicción del incumplimiento de las obligaciones financieras, tales como regresión logística, support vector machine, Naive Bayes, etc, donde el algoritmo Adaboost es el más regular, en base a las métricas estudiadas, sin embargo, el modelo Random Forest obtiene mejor ROC-AUC.

3. Hipótesis y Objetivos

3.1. Hipótesis

El modelo Extreme Gradient Boosting (XGBoost), gracias a su algoritmo de optimización de descenso de gradiente, permite identifica de mejor manera la variable objetivo (**Operación_Castigada)** y entregar el mejor resultado al indicador ROC-AUC con respecto a los otros modelos en estudio.

3.2. Objetivo General

Esta investigación tiene como objetivo desarrollar un modelo predictivo mediante herramientas de machine learning el cual nos permita predecir si el cliente cometerá default financiero, de tal forma de poder desarrollar estrategias de contención específicas, con anticipación, de forma de minimizar pérdidas financieras y reducir de costos.

3.3. Objetivos Específicos

Para el desarrollo de este paper se establecieron los siguientes objetivos específicos:

- Seleccionar modelos de machine learnig a estudiar.
- Realizar análisis exploratorio de los datos (EDA).
- Generar features de acuerdo con las necesidades de los datos.
- Entrenar tres modelos de machine learning.
- Comparar rendimiento de acuerdo con el indicador ROC-AUC.

4. Metodología

En la elaboración de este paper y con el objetivo de abordar cada una de las etapas, de tal manera de obtener el resultado deseado al final del estudio, se desarrollaron los siguientes pasos.

- **Elección de los modelos a comparar,** en base a los resultados obtenidos en la revisión bibliográfica y proyectos anteriores se definen los modelos a utilizar.
- **Descripción de los datos,** análisis inicial de dataset, donde se revisa la información general que se tiene a disposición.
- **Análisis exploratorio de datos (EDA) y Feature Engineering,** revisión estadística de los datos y la relación que existe entre ellos, además se depuran las variables con outliers y se crean nuevas features.
- **Resultados,** presentación de los resultados obtenidos de la implementación de cada uno de los modelos.
- **Conclusiones,** entrega de los resultados del proyecto, dando respuesta a la hipótesis y al objetivo del estudio.

5. Elección de los Modelos

Para este proyecto se seleccionaron tres modelos de machine learning los cuales nos predicen que cliente cometerá default. La elección de estos se realizó en base a los algoritmos más utilizados en los proyectos analizados en la revisión bibliográfica como también en un algoritmo utilizado en un proyecto personal realizado para la empresa, el cual entregó resultados aceptables para datos similares utilizados para este proyecto.

En primera instancia se seleccionó el modelo de regresión logística, modelo más popular para proyectos de scoring bancario, el cual es utilizado por varios autores para la predicción del default financiero.

El segundo modelo seleccionado fue utilizado en un proyecto de similares características para Coopeuch entregando buenos resultados, este modelo es Random Forest. Además, este modelo arroja buenos resultados en el estudio de Trujillo [10].

Finalmente, el último modelo que se utilizó fue Extreme Gradient Boosting (XGBoost), este modelo se eligió dado que al igual que Random Forest utiliza árboles de decisión para obtener los resultados, además este algoritmo utiliza el método de descenso de gradiente, método utilizado por Biron y Medina [6] para su estudio y que determina que Gradient Boosting Machine otorga una predicción más precisa.

En la siguiente sección se presentan de forma resumida los 3 modelos de machine learning que se seleccionaron para comparar sus performances.

5.1. Regresión logística

La regresión logística, es una técnica de aprendizaje automático, utilizada para calcular la probabilidad del resultado de una variable binaria “0” y “1”. Este modelo mide la relación entre una variable dependiente en función de una o más variables independientes o predictoras mediante una función logarítmica o sigmoide.

$$\text{función sigmoide} = \sigma(x) = \frac{1}{1 + e^{-x}}$$

Donde se puede ver que para los valores de x (predictor) positivos y grandes el valor de e^{-x} tiende a cero, lo que da como resultado de la función 1, ahora para x (predictores) negativos y grandes, el valor tiende al infinito por lo que la función sigmoide es 0.

5.2. Random Forest

Random Forest es un modelo de aprendizaje automático supervisado donde se generan “n” árboles de decisión en paralelo, cada uno se entrena con una muestra de los datos levemente distinta para luego mediante la técnica de bagging, combinar los resultados de los “n” árboles generados, de esta forma se reduce la varianza de los resultados, con el objetivo de obtener mayor precisión del resultado buscado.

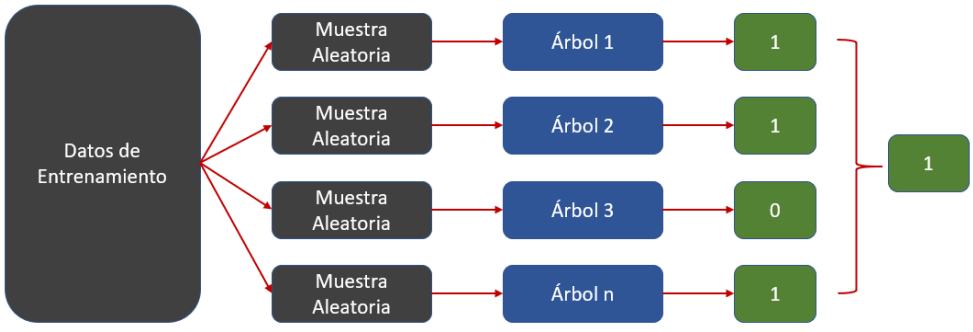

Figura 1: Esquema funcionamiento modelo Random Forest.

Un árbol de decisión está conformado por un nodo raíz, el cual contiene la muestra de datos disponible, este nodo se divide en función de las variables más importantes del set de datos, de esta forma se crean nuevos nodos, que nuevamente de dividen en función de las variables más importantes del flujo, finalmente el árbol termina de generarse cuando el proceso llega a un nodo termina u hoja, el cual entrega la clasificación buscada.

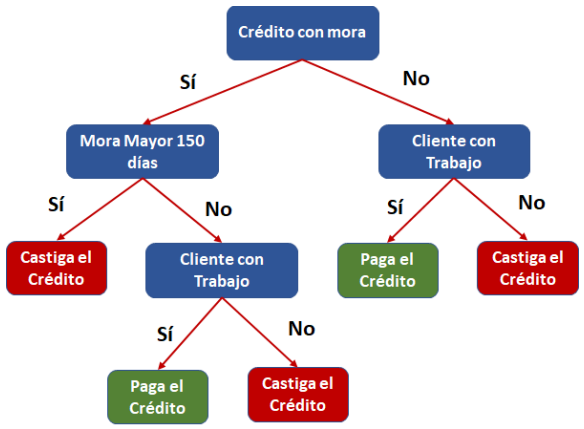

Figura 2: Esquema funcionamiento árbol de decisión.

5.3. Extreme Gradient Boosting (XGBoost)

XGBoost, al igual que Random Forest, es un algoritmo de aprendizaje supervisado que trabaja con múltiples árboles de decisión, sin embargo, los árboles generados por este modelo son elaborados secuencialmente (Boosting) de tal forma que para cada nueva iteración se aprenda del error de los árboles anteriores y de esta forma corregir el error, hasta que ya no se pueda realizar ninguna corrección más, este método se conoce como "gradiente descendente". De esta forma se pueden obtener mejores resultados predictivos y mayor estabilidad del modelo.

Se inicia el proceso con un árbol A_0 para predecir la variable objetivo "y", al resultado obtenido se asocia con un residual ($y - A_0$), de esta forma se obtiene un nuevo árbol H_1

el que ajusta el error del proceso anterior y posteriormente se combina con A_0 para obtener nuevo un nuevo árbol A_1. Este proceso itera hasta que el error cuadrático medio sea minimizado al máximo.

6. Descripción de los Datos

Los datos utilizados son registros de operaciones de crédito de consumo con 31 o más días de mora, desde octubre 2018 a diciembre de 2019, donde se ha eliminado información que permita identificar a la persona afectada. En total se tiene un dataset con 254.519 filas y 34 features.

Se han seleccionado solo registros de créditos con mora mayor o igual a 31 días con el propósito de poder determinar cuáles de estos créditos tienen mayor probabilidad de cometer default financiero, con el propósito de generar estrategias de contención anticipadas y así reducir pérdidas económicas.

En el anexo N° 1 se encuentra la descripción de cada una de las variables que conforman el dataset.

Como ya sabemos cuántas feature tiene nuestro dataset y cuantos registros lo conforma, continuamos con un análisis más detallado de este, donde en promedio se tienen 16.968 registros mensuales, cuya distribución es bastante homogénea comparando los meses.

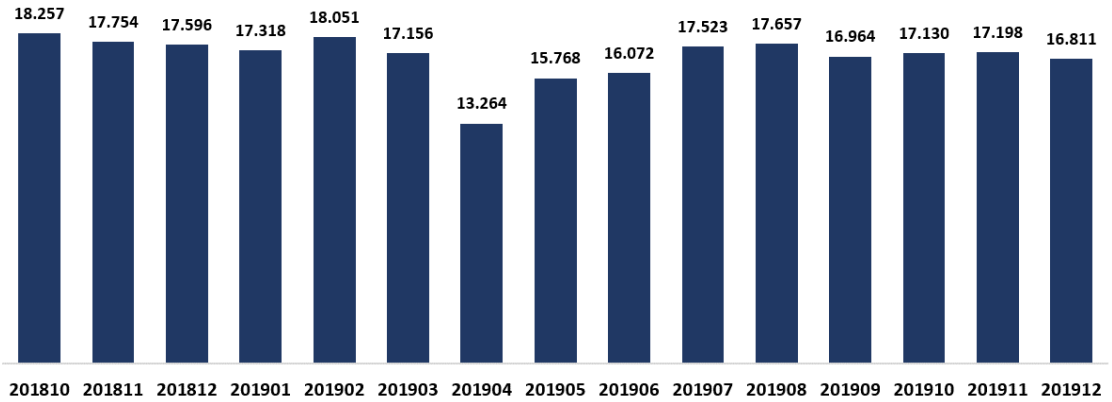

Figura 3: Distribución mensual de los datos.

Al revisar el número total de registros por mes, observamos que durante el mes de abril de 2019 la cantidad de operaciones de crédito en mora disminuye con respecto al total operaciones de los otros meses. Esta situación se debe a que la empresa, al ser una cooperativa, las utilidades obtenidas durante el ejercicio del año anterior son distribuida entre sus socios, existiendo una restricción, que consiste en que la proporción de las utilidades que le corresponde a un cliente con operaciones de crédito en mora se utilizan para pagar la deuda morosa y dejar las operaciones de crédito vigentes y así de reducir el riesgo de default durante este período.

Teniendo identificada la distribución de los registros y aclarado el comportamiento de los datos para el mes de abril, observamos con más detalle nuestra variable objetivo, **operación_castigada**, donde observamos que el 48% de las operaciones de crédito se castiga, variable que buscamos predecir.

Anteriormente fue posible revisar el comportamiento mensual de los datos y cuantos registros entraban en default, sin embargo, no todos los meses se mantienen el mismo promedio de castigo, debido a la distribución del remanente que hace que las operaciones dejen de estar en mora. En la gráfica a continuación se muestra el comportamiento mensual de los créditos que entran en default, donde se aprecia que es un sistema estable en la distribución de las variables.

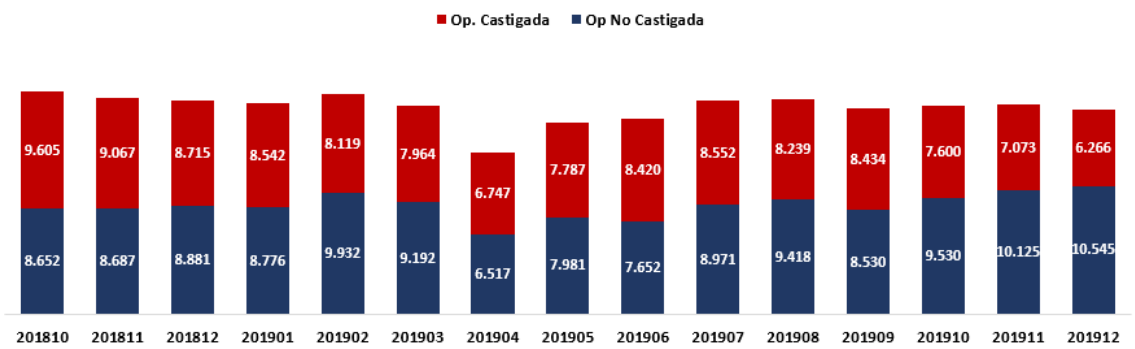

Figura 4: Distribución mensual de la variable objetivo.

Dentro de la información obtenida de la revisión del dataset, se destaca que aproximadamente el 62% de las operaciones de crédito con mora mayor a 31 días son del sexo femenino, donde principalmente los créditos solicitados son mediante la modalidad descuento por planilla, 71% de las operaciones, proporción similar en los hombres, con un 69% de los créditos cursados a través de esta modalidad.

Con relación a lo anterior, donde el descuento por planilla es la opción dominante a la hora de solicitar los créditos, ya que de esta forma se minimiza el riesgo de no pago de las obligaciones financieras por parte de los clientes, observamos que el 49% de los créditos pertenecen a instituciones públicas, tales como municipalidades, servicios de salud, educación entre otros y un 21% de ellos pertenecen a clientes que se desempeñan en empresas del sector privado.

Si bien Coopeuch atiende a clientes de todos los estratos socioeconómicos, el público objetivo son personas que no tienen posibilidad de solicitar financiamiento para sus proyectos en bancos dado las bajas rentas que estos poseen, donde la renta promedio asciende a $492.000. Lo bajo de la renta se debe a que el 67% de los clientes poseen estudios técnicos o inferiores.

Dado que la empresa tiene presencia a nivel nacional, es que el 71% de los registros de créditos morosos fueron cursados fuera de la región metropolitana, concentrándose en mayor proporción en la zona sur del país, territorio comprendido entre la región del Bio Bío y Magallanes, con un 32% del total de casos.

7. Análisis Exploratorio de Datos (EDA) y Feature Engineering

En el análisis exploratorio se realizó una descripción estadística de cada una de las variables numérica, donde se profundizó la revisión cuando se detectaban comportamientos anómalos en los datos.

Se inició con un análisis estadístico de las siguientes variables: **Id_Mes, Operacion_Renegociada, Op_Reliquidada, Operacion_Castigada, Campaña_Contencion, Campaña_gestionada, Campaña_Cerrada, Campaña_Gestion_exitosa, Dias_mora**, donde logramos ver su comportamiento, promedio y descartamos la existencia de anomalías en la cual aplicar un análisis más profundo.

La Variable **Id_Mes** corresponde a un registro tipo año mes para identificar el mes de la toma de los registros de las operaciones en mora, por lo que no requiere mayor análisis.

No obstante, se revisó el comportamiento de la variable **Dias_mora**, variable que nos indica que tan cerca de ser castigado está el crédito, ya que cuando llega a los 180 días de mora, se genera el default financiero.

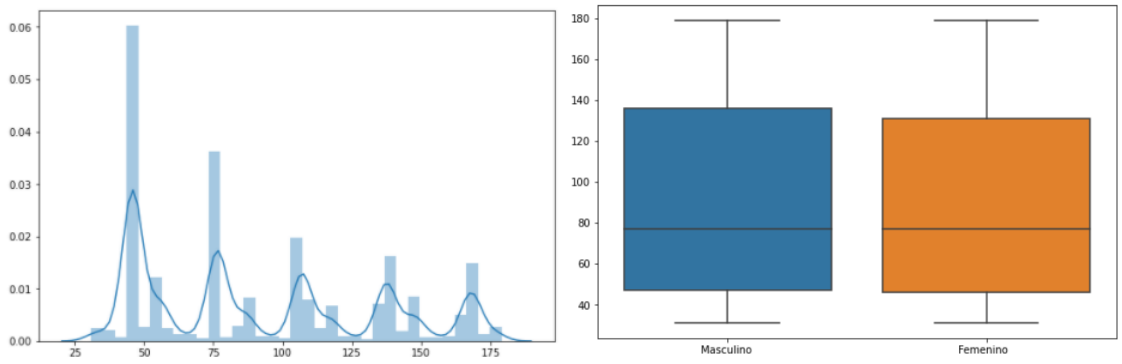

Figura 5: Histograma y Boxplot feature Días_mora.

Al revisar la figura 4, se pudo apreciar que el 50% de los créditos en mora para ambos géneros ronda los 80 días. También logramos identificar la existencia de algunos picos en determinados días de mora. Al agrupar esta información en tramos, de acuerdo con la clasificación de mora utilizada por la empresa, identificamos que existe una mayor concentración de registros en el tramo "B-" la cual contiene créditos morosos entre 31 y 60 días, seguido por el tramo "C" con un rango de días de mora entre 61 y 90, finalmente el tramo más riesgoso en el "D-" donde contiene operaciones con más de 151 días de mora y son aquellas que están próximas a entrar en default.

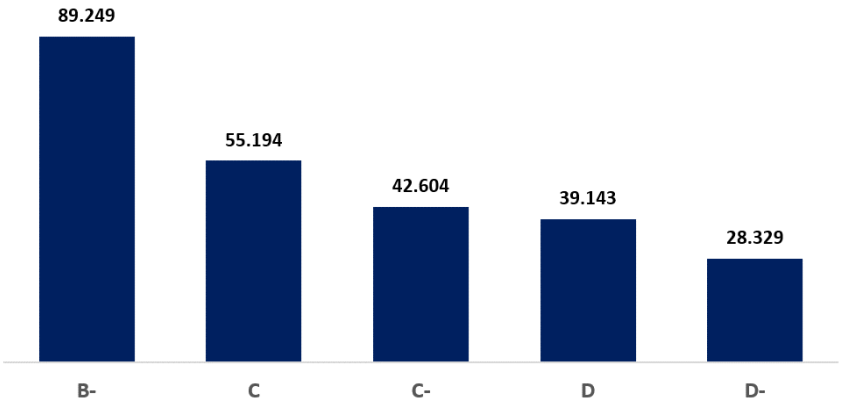

Figura 6: Distribución días de mora por tramo.

Continuando con la revisión de los datos, se analizó el comportamiento de la variable **Edad_socio**, donde observamos que ésta feature no presenta problemas de outliers, ya que la edad mínima para cursar un crédito es de 18 años y los casos con clientes con 91 años fueron cursados hace un tiempo y se encuentran próximos a terminar de pagar sus créditos. Además, identificamos que la edad promedio de los socios es de 42,6 años.

Ahora bien, a través del histograma generado en base a la variable **Edad_socio**, es posible observar a simple vista que esta variable no tiene una distribución normal, ya que posee

una asimetría positiva. Además, se desprende de la figura 6, que las personas que cometen default son aquellos de menor edad, inferior al promedio de 42 años.

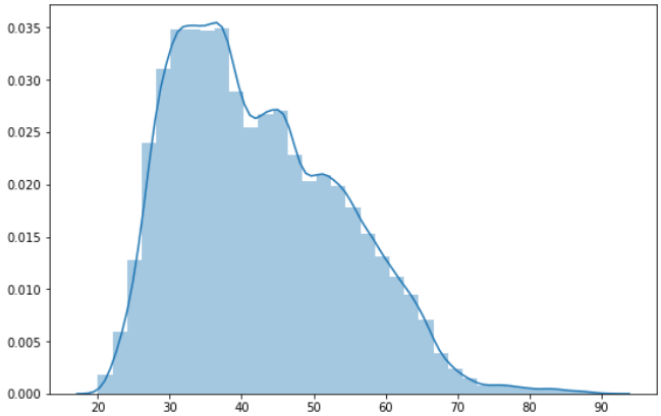

Figura 7: Histograma de feature Edad_Socio.

Al revisar la variable **Antiguedad_laboral** se pudo observar que presenta valores atípicos, y al analizar los registros con mayor detalle se identificó que el 55% del total de los datos posee cero años de antigüedad, además se identificaron registros con antigüedad negativa y otros con un máximo de 1.927 años de trabajo, por lo cual no usaremos esta feature ya que no aporta al objetivo por tener datos poco confiables.

Siguiendo con el análisis, se revisó la variable **Cuotas_Pactadas** donde se encontraron algunos outliers, tales como operaciones de crédito con menos de 3 cuotas. Con esta condición se identificaron 10 registros, estos son eliminados ya que por política de la empresa el plazo mínimo de un crédito es de 3 cuotas.

Al analizar el otro extremo de la curva, generada por la variable **Cuotas_Pactadas,** se identificaron 88 registros con plazo superior a las 120 cuotas, plazo que no concuerda con la política de la empresa, la que indica que el plazo máximo de un crédito es de 84 cuotas, existiendo algunas excepciones que permiten que en casos muy puntuales lleguen a las 120.

Al profundizar con los casos antes descritos, nos percatamos que para los créditos con plazos superiores a 120 cuotas la edad del cliente, supera los 60 años, además existen clientes con baja renta y plazos muy largos esto indica la existencia de datos erróneos, por lo cual se procedió a eliminar estos registros del dataset.

Por otro lado, en el histograma asociado a la variable **Cuotas_Pactadas** se ven pick de curses en operaciones de 36, 48 y 60 cuotas, plazos más recurrentes a la hora de solicitar un crédito.

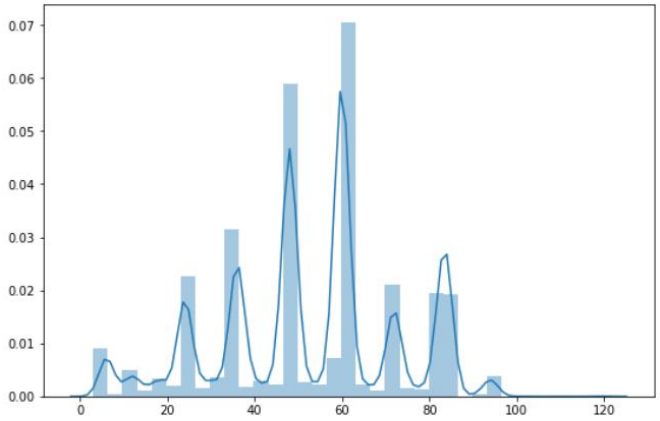

Figura 8: Histograma feature cuotas_pactadas.

Al mirar el comportamiento de la variable **Cuotas_Canceladas,** ésta no presenta datos anómalos. En la revisión nos percatamos que el promedio los clientes cancelan 19 cuotas de su crédito antes de entrar en mora y cometer default. Cabe destacar que un cliente puede solicitar un crédito y nunca pagar una cuota, por lo que es factible y no representa anomalía que existan registros con cero cuotas canceladas.

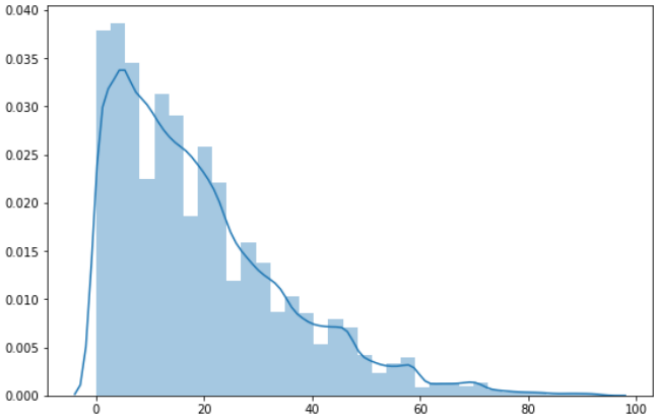

Figura 9: Histograma feature cuotas_canceladas.

En el caso de la variable **Renta_liquida,** se aprecia que en promedio la renta de los clientes de Coopeuch es de aproximadamente $492.000 valor que concuerda con la renta del público objetivo de los segmentos donde está enfocada la empresa, sin embargo, se identificaron 43 registros con renta mayor a los $10.00.000, ya que Coopeuch no solo atiende a clientes de los segmentos “D” y “E”, sino que atiende, en menor número, clientes de segmentos bancarizados. Al revisar los casos indicados anteriormente, nos percatamos que existe una incongruencia entre la deuda interna, deuda externa y nivel educacional de los clientes en cuestión, además al revisar cada uno de forma individual identificamos que corresponden a 9 clientes diferentes, por lo cual estos registros ensucian nuestros datos, por lo que son eliminados del estudio. Estos 43 casos equivalen a menos un 0.02% del total de los datos por lo que los resultados no se verán afectados.

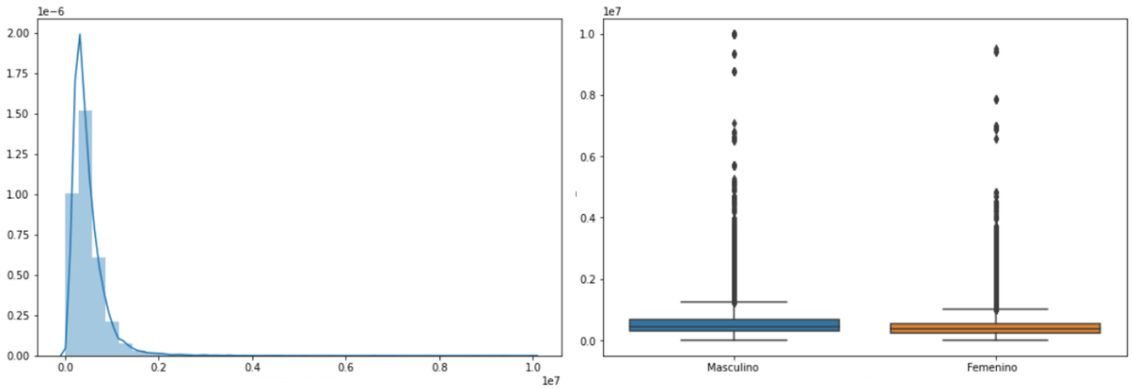

Figura 10: Histograma y Boxplot feature renta_liquida.

En un segundo análisis de la variable y al revisar tanto el histograma como los boxplot generados en base a la variable **Renta_liquida,** nos percatamos que aún persisten outliers ya que el 75% de los clientes tienen renta líquida inferior a $600.000, en este caso se generó una nueva feature llamada **Tramo_Renta,** la cual permitió visualizar de mejor manera el comportamiento de la renta de los clientes, la que se concentra a la izquierda del gráfico.

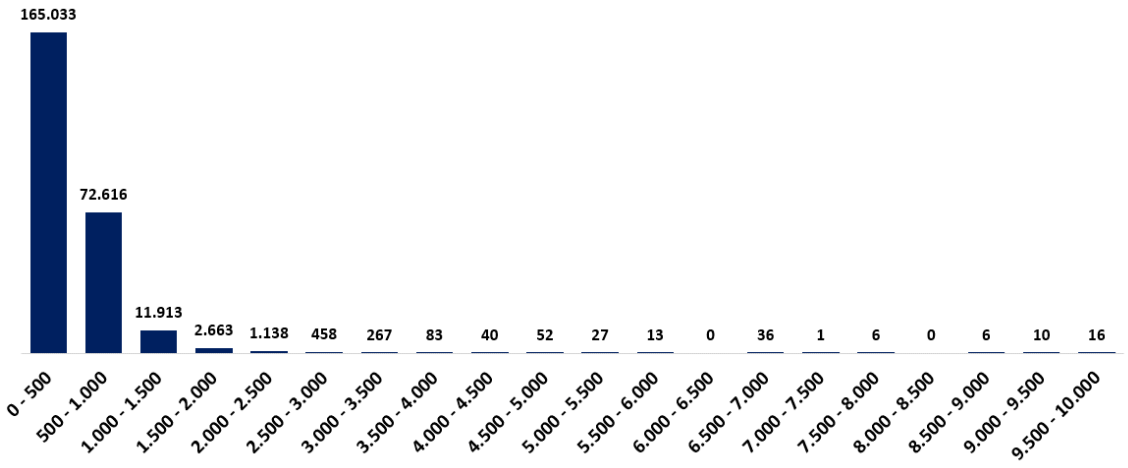

Figura 11: Distribución feature tramo_renta (M$).

La **Deuda_Ext_Consumo** también requirió de una revisión individual, ya que se identificaron casos atípicos, tal como se ha comentado anteriormente, la renta promedio de los clientes es de aprox. $490.000 por lo cual la deuda externa no puede ser muy elevada, ya que gran parte de los clientes, con este nivel de renta, no están bancarizados y por ende no puede acceder a créditos en cualquier institución financiera. En esta situación encontramos 523 registros con deuda externa superior a los $60.000.000, los cuales no fueron considerados en el estudio, ya que no concuerdan con los niveles de renta de los clientes objetivos de la Cooperativa.

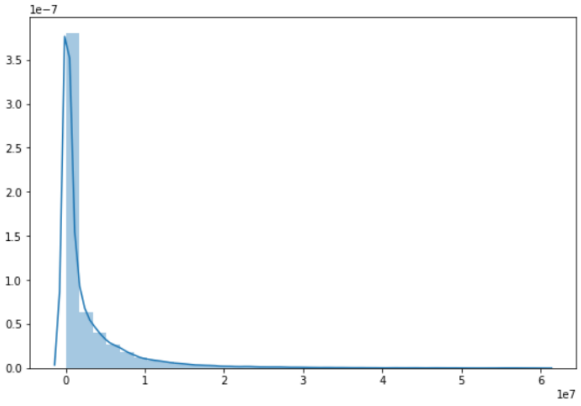

Figura 12: Histograma feature deuda_ext_consumo.

Al revisar la información entregada por el histograma, nos percatamos que tiene un comportamiento similar a la feature **Renta_liquida,** por lo cual se implementó el mismo tratamiento para eliminar los outliers, el cual fue crear una nueva variable llamada **Tramo_Deuda_ext**, la que nos permitió ver el comportamiento de los clientes e identificar que la mayor parte de ellos tiene una deuda externa inferior a $1.000.000.

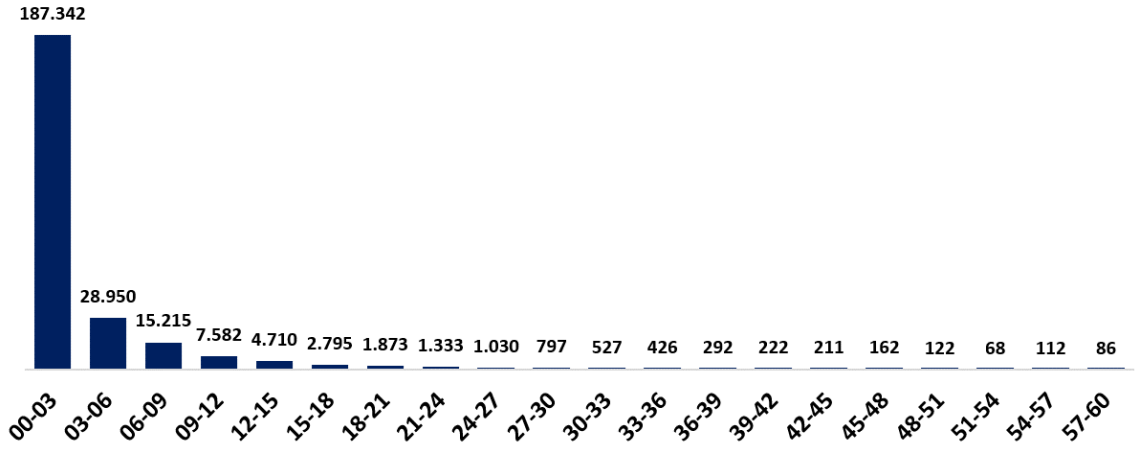

Figura 13: Distribución feature tramo_deuda_ext (MM$).

Continuando con nuestra revisión de los datos, encontramos que la variable **Saldo_Credito,** tiene comportamiento similar a las variables descritas anteriormente, donde el saldo promedio (deuda promedio de los créditos de los clientes) es de

aproximadamente $3.710.000 en base a esta información es que no se consideraron para los análisis posteriores los registros de crédito con saldo superior a los $20.000.000 ya que, al ser montos elevados, se realizan campañas de contención focalizadas.

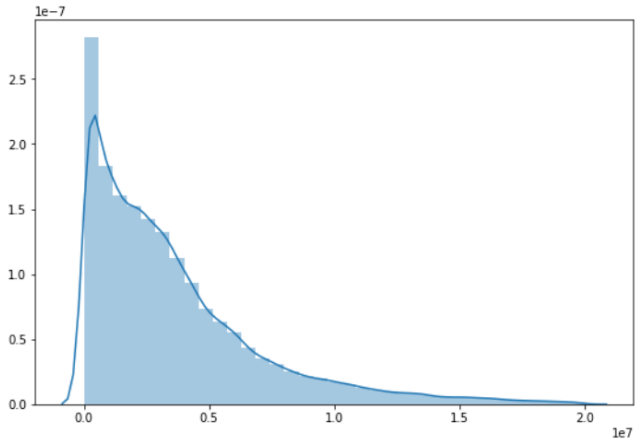

Figura 14: Histograma feature saldo_credito.

El histograma de la variable **Saldo_Credito,** nos indica que la deuda de los clientes se concentra por debajo de los $5.000.000, sin embargo, existe una gran dispersión en los datos de clientes con deudas sobre los $10.000.000, con el objetivo de mejorar la precisión de los resultados, se creó una nueva variable llamada **Tramo_Saldo_Credito**, donde se agrupan los registros en 10 tramos.

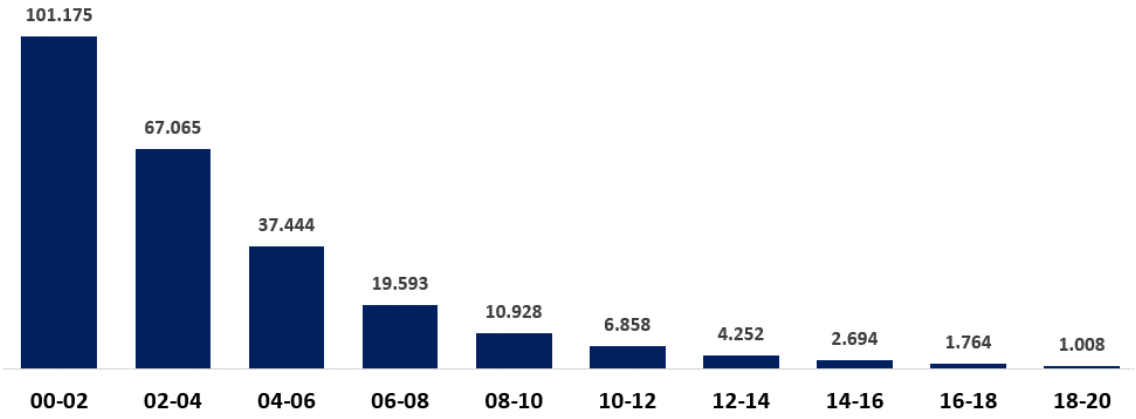

Figura 15: Distribución feature tramo_saldo_credito (MM$).

Al analizar la variable **Monto_Total_Credito,** observamos inmediatamente que tiene similar comportamiento con la variable **Saldo_Credito** ya que estas poseen correlación, esto se debe a que la variable en análisis corresponde al monto del crédito otorgado al cliente y la anterior corresponde al monto que adeuda el cliente posterior el pago de las cuotas.

Dado que las variables indicadas anteriormente tienen comportamientos similares, las operaciones de monto mayor a los $20.000.000 no son consideradas en el resto del análisis, ya que para ser cursadas requieren que sean aprobadas por un comité de gerentes y cuando entran en mora, son sujeto a campañas focalizadas.

El tratamiento para esta variable es similar a las descritas anteriormente y se crea nueva feature con el objetivo de optimizar el dataset para cuando se apliquen los modelos.

La Variable **Tasa_Interes** presenta algunos parámetros anormales, ya que se observaron registros con tasa de 6,3%, la cual es excesiva para un crédito de consumo, dado que la tasa para este tipo de créditos se encuentra en rangos inferiores 3,5%. Los registros con tasa superior a 3,5% son descartados ya que la tasa es inversamente proporcional al monto del crédito y para el caso de los registros con tasas superiores al 3,5% los créditos superan los $2.700.000 monto donde las tasas comienzan a ser más competitivas con el resto del mercado, por ende, más bajas.

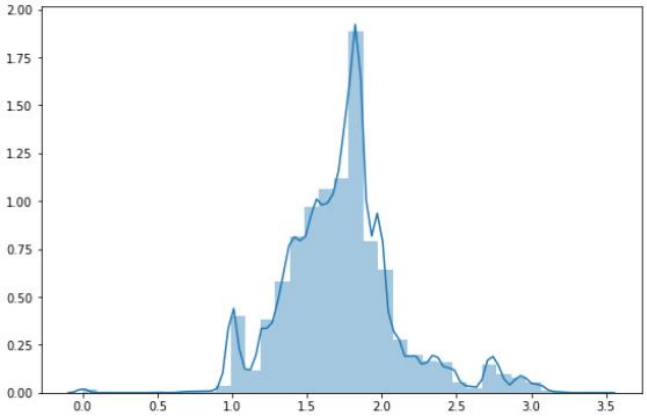

Figura 16: Histograma tasa_interes.

Gracias a la información que entrega el histograma de la variable **Tasa_Interes,** se observó que esta variable tiene un comportamiento normal y el resto de los registros se encuentra dentro de parámetros aceptables, ya que, en ocasiones especiales, se han cursado créditos a tasa 0%.

Con respecto a la variable **valor_cuota,** ésta no requiere de mayor análisis, dado que posee una asimetría positiva y con un valor promedio de la cuota de aprox. $130.000 monto que guarda relación con las rentas de los clientes, cabe destacar que el 50% de los clientes tiene una cuota inferior al promedio.

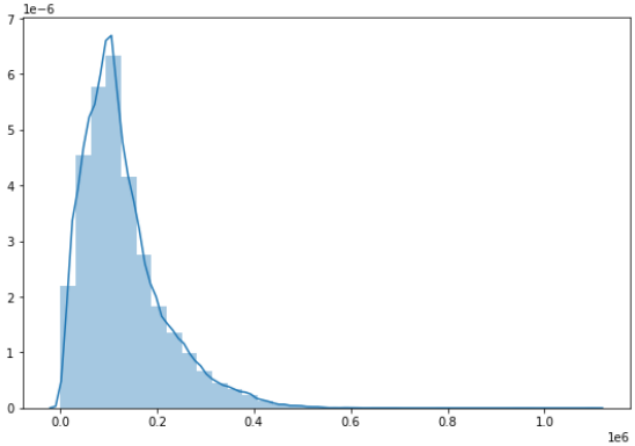

Figura 17: Histograma feature valor_cuota.

Los outliers de esta feature no generan confusión en la revisión, ya que al revisar los casos identificamos que las cuotas de alto monto corresponden a créditos de monto mayor de clientes de rentas altas.

Las variables **Op_Mora** y **Oper_Vigente** no presentan valores atípicos, ya que ambas están en rangos aceptables (entre 1 y 8 operaciones), lo que indica que un cliente puede tener hasta 8 operaciones de crédito en paralelo con Coopeuch y los 8 créditos los tiene en mora. Esta situación no se da con frecuencia, ya que en estos casos se propone al cliente repactar toda su deuda en uno o dos créditos paralelos. En base a esta información, estas variables no requieren de mayor análisis ya que están dentro de los parámetros permitidos.

La variable **Deuda_Interna_Consumo**, nos indica que los clientes de Coopeuch tienen una deuda promedio de $4.730.000 con esta institución, donde nos percatamos que existen valores que se escapan del total de registros, tales como deudas por sobre los $15.000.000 y como se ha comentado anteriormente, los clientes son principalmente de los grupos socioeconómicos D y E, donde para ellos es imposible tener deudas de este nivel.

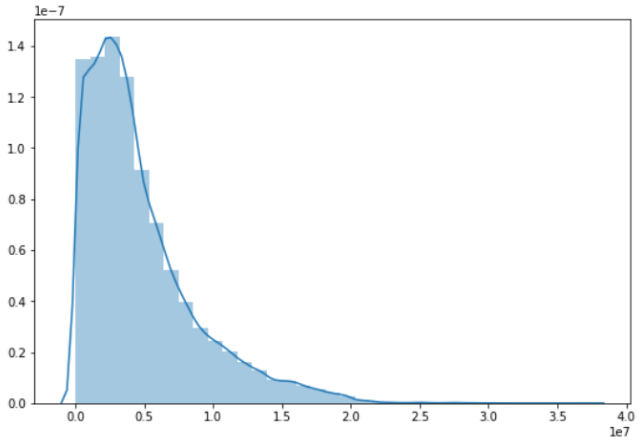

Figura 18: Histograma feature deuda_interna_consumo.

Se generó un histograma para ver el comportamiento de la curva de deuda de los clientes, de esta visualización se pudo observar que el 75% de los clientes de la muestra tienen una deuda interna igual o menor a los $6.400.000 y dado que la empresa también atiende a clientes de nivel socioeconómico más altos, se generó un nueva feature la cual se nombró **Tramo_Deuda_Interna**, la que agrupa a los clientes en tramos para facilitar la implementación posterior de los modelos a estudiar.

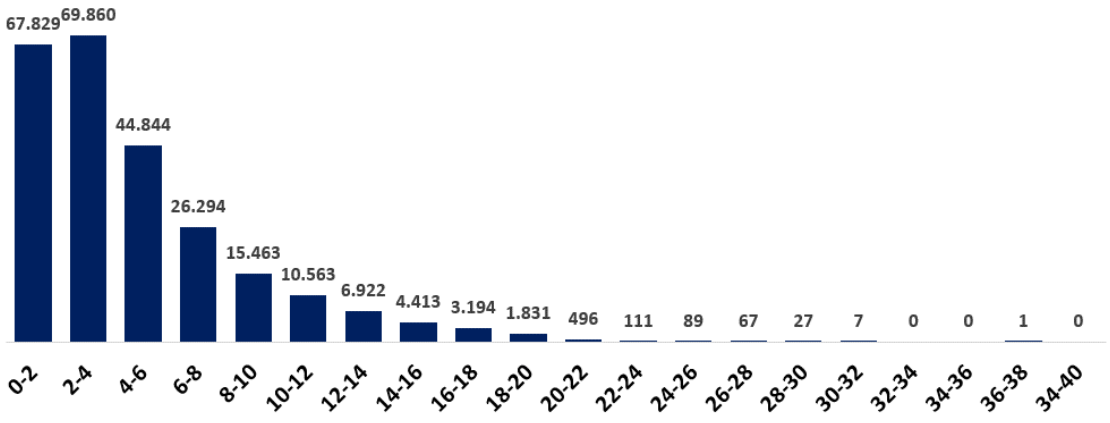

Figura 19: Distribución feature tramo_deuda_interna (MM$).

De esta forma es más fácil apreciar el comportamiento de la curva y en qué tramos se concentran la mayor cantidad de clientes de acuerdo con la deuda total que tienen con la institución financiera.

Finalmente, para dar término al proceso de análisis estadístico de los datos, se revisó la variable **Deuda_Ext_Hipoteca**, la cual se descarta de la base ya que al realizar la exploración nos encontramos que más del 75% de los registros tienen valor cero y el mayor monto encontrado supera los $270.000.000.

Ya terminada la revisión individual de las variables, analizamos la correlación de cada una de las variables, donde destacaron algunos casos que requirió una revisión más profunda.

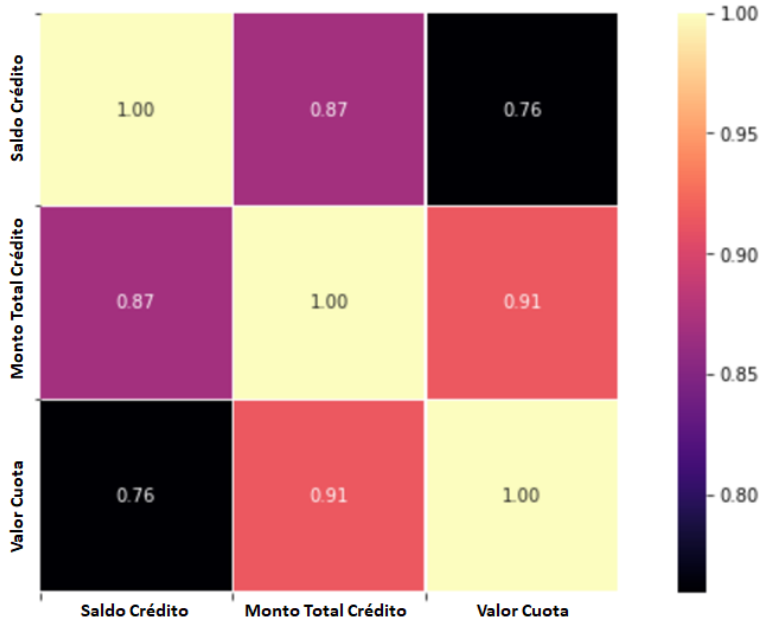

Figura 20: Matriz correlaciones features saldo_credito, monto_total_credito y valor_cuota.

En esta situación, observamos en primera instancia que existe una alta correlación entre las variables **Saldo_Credito**, **Monto_Total_credito** y **Valor_Cuota**, esto se debe a que la primer es el remanente del crédito que aún está pendiente por pagar, la segunda corresponde al monto total del crédito solicitado por el cliente y el valor de la cuota va a ser más alto mientras más alto sea el crédito solicitado.

Al existir una alta correlación entre las variables **Saldo_Credito** y **Monto_Total_Credito**, se crea una nueva feature la cual corresponde al ratio del saldo del crédito sobre el monto total, esta variable es llamada **Ratio_Pago_Credito**.

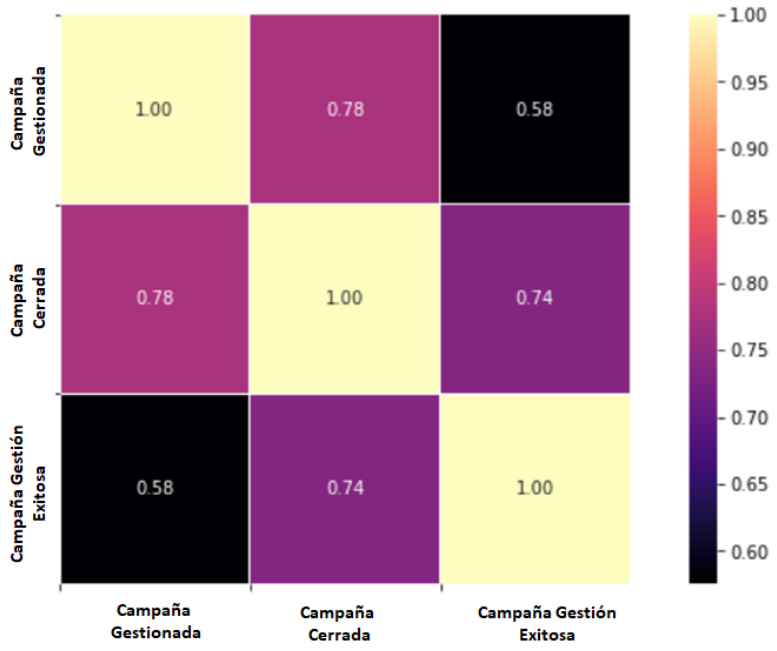

Figura 21: Matriz correlaciones features gestión campañas.

Otro grupo de variables con alta correlación son las features asociadas a las campañas de contención, las cuales son **Campaña_Gestionada, Campaña_Cerrada** y **Campaña_Gestion_Exitosa**. Donde las dos primeras variables se correlacionan entre sí, ya que para cerrar la campaña tiene que haber sido gestionada, la gestión del ejecutivo pudo haber generado que el cliente pagará la cuota morosa, como también puede que el cliente haya hecho caso omiso a la información entregada por el ejecutivo y continuará sin pagar sus obligaciones financieras.

En base a la información entregada por la matriz de correlaciones, eliminaremos la variable **Campaña_Cerrada**, ya que para cerrarla tiene que ser gestionada, independiente del resultado de la gestión, por lo cual se eliminará de la base.

Finalmente, nuestra matriz de correlaciones de nuestras variables se muestra de acuerdo con la siguiente figura.

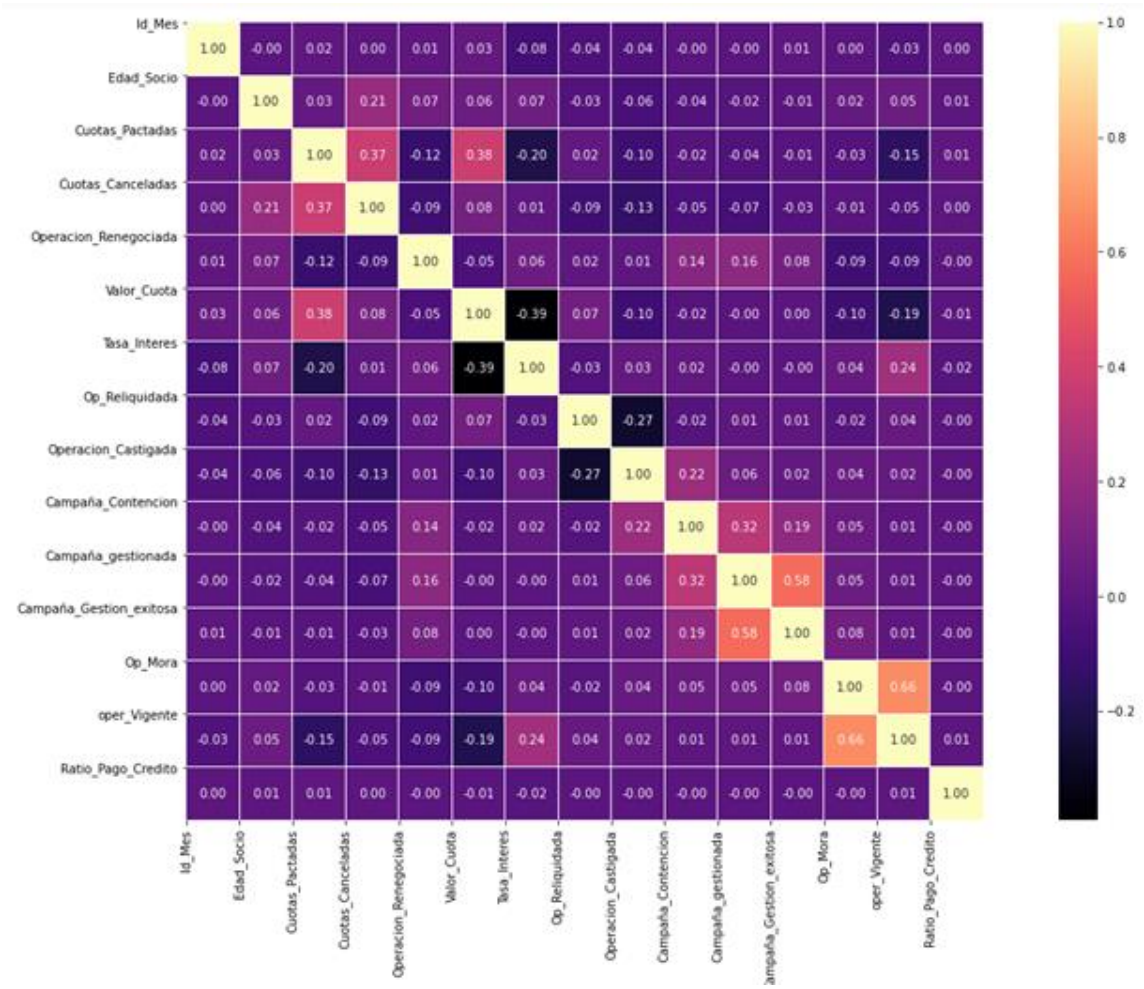

Figura 22: Matriz correlaciones features dataset.

8. Resultados

Ya terminada la revisión de los datos y creadas las nuevas variables requeridas, contamos con 252.011 registros y 29 features, donde se procedió a dividir nuestros dataset en dos grupos, train (70%) y test (30%), esto con el objetivo de entrenar los modelos sujetos a estudio, para posteriormente, con el set de datos test, validar la efectividad de cada uno de los modelos.

Ya con los datos divididos se entrenaron cada uno de los modelos y aplicamos el método de cross-validations, con el objetivo de validar que los resultados obtenidos son prueba de un buen entrenamiento del modelo no de un sobreajuste de estos.

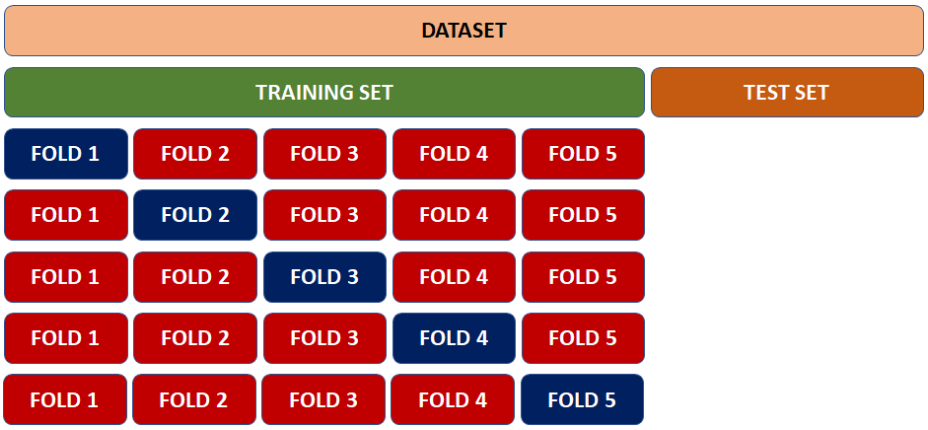

Figura 23: Esquema funcionamiento cross-validations.

8.1. Regresión logística

Para implementar el modelo de regresión logística, se configuran algunos de los hiperparámetros, con el objetivo de obtener el mejor resultado posible.

```
LogisticRegression (tol=0.00001, C=0.8, fit_intercept=True, class_weight='balanced',solver='liblinear', max_iter=100)
```

Figura 24: Hiperpárametros modelo regresión logística.

Una vez teniendo los datos preparados para la implementación del modelo se entrena el algoritmo de regresión logística, el cual no nos entrega resultados muy alentadores, ya que obtenemos una accuracy de 72% con una precisión de 70%. Esto nos indica que el modelo no es capaz de identificar la feature objetivo.

Otro de los resultados que calculamos y que nos permiten comparar la eficiencia del modelo es el indicador ROC-AUC, el cual fue de 80,1% y el RMSE de 0,28.

	0	1
0	28.316	11.280
1	9.671	26.337

	Precision	Recall	F1-Score	Support
0	0,75	0,72	0,73	39.596
1	0,70	0,73	0,72	36.008
Accuracy			0,72	75.604
Macro Avg	0,72	0,72	0,72	75.604
Weighted Avg	0,72	0,72	0,72	75.604

Tabla 1: Matriz de confusión y métricas de desempeño regresión logística.

8.2. Random Forest

Del mismo modo que para el modelo anterior, se definen los siguientes hiperpárametros para optimizar el resultado del algoritmo Random Forest.

```
RandomForestClassifier(n_estimators=150,criterion='gini',random_state=1,class_weight='balanced',n_jobs=1)
```

Figura 25: Hiperpárametros modelo random forest.

Ya teniendo los resultados del modelo de regresión logística, aplicamos los datos al modelo de Random Forest, donde a diferencia del algoritmo anterior, los resultados fueron mucho más alentadores, ya que nos arroja un accuracy de 95% y una precisión de 94%, logrando identificar la clase objetivo y predecir con mayor certeza cuáles operaciones serán castigadas.

Para este modelo también se calcularon los indicadores de ROC-AUC, el cual fue de 98,9% y un RMSE de 0,05.

	0	1
0	37.565	2.031
1	2.064	33.944

	Precision	Recall	F1-Score	Support
0	0,95	0,95	0,95	39.596
1	0,94	0,94	0,94	36.008
Accuracy			0,95	75.604
Macro Avg	0,95	0,95	0,95	75.604
Weighted Avg	0,95	0,95	0,95	75.604

Tabla 2: Matriz de confusión y métricas de desempeño Random Forest.

8.3. Extreme Gradient Boosting (XGBoost)

Finalmente se implementa el algoritmo XGBoost, donde los resultados, donde los hiperparámetros utilizados para en entrenamiento del modelo fueron:

```
XGBClassifier(n_estimators = 100,learning_rate = 0.1)
```

Figura 26: Hiperpárametros modelo Extreme gradient boosting.

al igual que la regresión logística, no fueron bastante acertados ya que con un accuracy de 78% y una precisión de 77% no entrega indicadores de generar una predicción aceptable para poder dar una solución al problema que se desea resolver.

Los indicadores ROC-AUC y RMSE para el algoritmo de Extreme Gradient Boosting fueron 86,5% y 0,22 respectivamente.

	0	1
0	31.326	8.270
1	8.729	27.279

	Precision	Recall	F1-Score	Support
0	0,78	0,79	0,79	39.596
1	0,77	0,76	0,76	36.008
Accuracy			0,78	75.604
Macro Avg	0,77	0,77	0,77	75.604
Weighted Avg	0,78	0,78	0,78	75.604

Tabla 3: Matriz de confusión y métricas de desempeño extreme gradient boosting.

8.4. Comparativo de resultados

Ya con los resultados de los tres modelos sujetos a estudio generamos una tabla que nos permitió consolidar todos los resultados en un solo lugar para poder revisar y determinar cuál de ellos tuvo mejor rendimiento.

Model	True Positive	False Positive	True Negative	False Negative	ROC-AUC	RMSE	Accuracy (Test)
Logistic Regression	28.316	11.280	26.337	9.671	80,1%	0,28	72,3%
XGBoost	31.326	8.270	27.279	8.729	86,5%	0,22	77,5%
Random Forest	37.565	2.031	33.944	2.064	98,9%	0,05	94,6%

Tabla 4: Matriz de resultados modelos.

Cuando comparamos los resultados de los tres modelos y observamos la gráfica a primera vista identificamos el modelo Random Forest es aquel que reacciona de mejor manera a los datos, superando a los otros modelos en el indicador ROC-AUC.

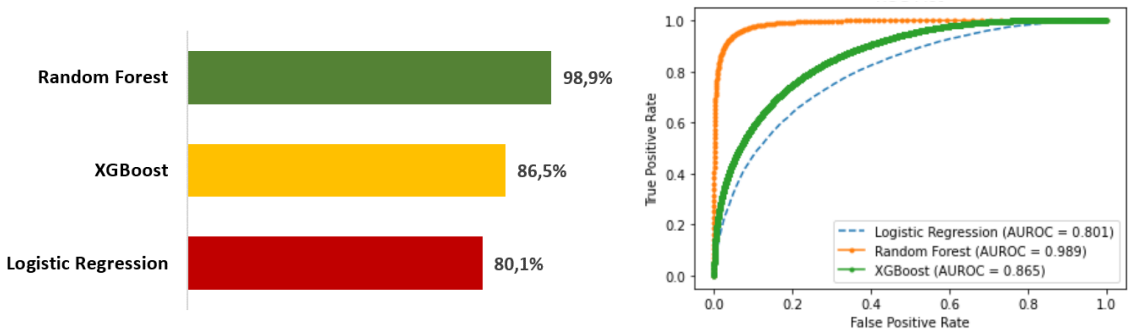

Figura 27: Gráfico comparativo ROC-AUC.

Del mismo modo, el modelo random forest tiene un menor RMSE, lo que demuestra que entrega mejores resultados que los otros modelos.

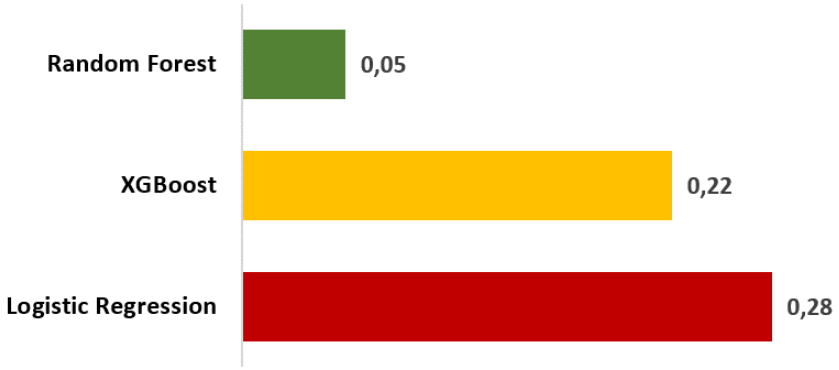

Figura 28: Gráfico comparativo RMSE.

También observamos que del mismo modo que en las gráficas anteriores, el modelo Random Forest tiene un mayor accuracy que los otros dos algoritmos estudiados.

Figura 29: Gráfico comparativo Accuracy (test).

Finalmente, y de acuerdo con la gráfica, el modelo regresión logística, es aquel que reaccionó de la peor forma con respecto a los modelos estudiados con una mayor cantidad de falsos positivos y falsos negativos.

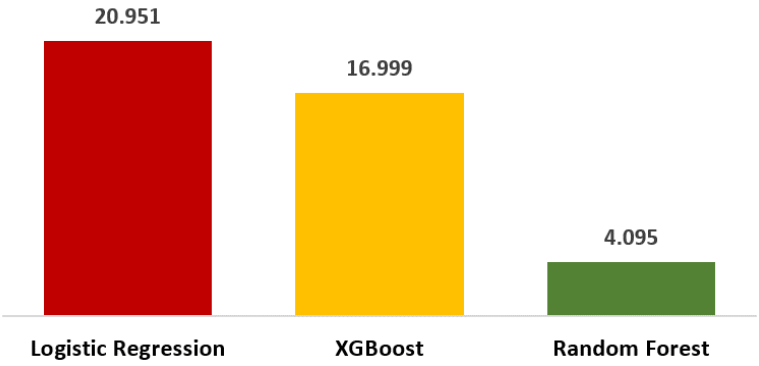

Figura 30: Gráfico comparativo observaciones falsas.

9. Conclusiones

El método de descuento por planilla es una forma de minimizar el riesgo de default financiero que utiliza la cooperativa, con el cual acerca el financiamiento a personas con bajas renta. Sin embargo, a pesar de esta metodología, existe un número importante de créditos que se encuentran en mora.

Este paper busca, mediante la comparación de los rendimientos de tres modelos de machine learning, cuál de ellos predicen de mejor manera el default financiero de los clientes de Coopeuch. Las métricas utilizadas para realizar esta comparación fueron el indicador ROC-AUC, RMSE y el accuracy y el modelo con mejores resultados de los indicadores indicados anteriormente fue Random Forest.

Random forest obtiene un indicador ROC-AUC de 98,9%, destacándose con respecto a los otros 2 modelos en estudio, esto se debe a que este algoritmo, al realizar variados modelos “débiles” de bajo sesgo los cuales al ser combinados mediante el método de bagging, obtiene un resultado de bajo sesgo y baja varianza, donde la predicción del modelo es mucho más precisa, en cambio para caso el modelo Extreme Gradient Boosting este no obtiene los resultados esperados, dado que cada árbol se entrena secuencialmente ajustando los resultados del árbol anterior para obtener la mejor predicción, para esto se requiere configurar de forma óptima sus hiperparámetros y seleccionar las features más significativas.

Con los resultados obtenidos rechazamos la hipótesis, la cual indicaba que el modelo Extreme Gradient Boosting reaccionaba mejor a los datos y entregaría una mejor predicción.

Este estudio puede ser mejorado optimizando la configuración de los hiperparámetros de cada uno de los modelos estudiados y seleccionando las features de mayor importancia, para determinar si el indicador ROC-AUC mejora en cada uno de los modelos.

Este trabajo puede ser utilizado e implementado para poder desarrollar estrategias de cobranza focalizadas en los clientes que puedan cometer default financiero y de esta forma poder minimizar los indicadores de riesgo e incrementar el remanente de la empresa.

Bibliografía

1. Pazmiño Real, D. (2011). *La cartera vencida y su incidencia en la rentabilidad del Banco Nacional De Fomento sucursal Ambato durante el periodo comprendido de enero a diciembre del 2009*. https://repositorio.pucesa.edu.ec/handle/123456789/696. Visitada el 14-07-2020.
2. Vera Andrade, E. (2013). *Gestión de crédito y cobranza para prevenir y recuperar la cartera vencida del banco Pichincha de la ciudad de Guayaquil en el periodo 2011*. https://repositorio.uide.edu.ec/handle/37000/1560. Visitada el 17-08-2020.
3. Armijos Loaiza, A. D., & Oña Muñoz, J. C. (2015). *Modelo de gestión de crédito y cobranza para recuperar cartera vencida en la Cooperativa de ahorro y crédito San Miguel de Los Bancos y sus tres agencias que la integran*. http://dspace.ups.edu.ec/handle/123456789/10202. Visitada el 17-08-2020.
4. Caiza Chango, C. (2015). *Modelo de gestión de cobranza para disminuir la cartera vencida en la cooperativa de ahorro y crédito Pakarymuy Ltda agencia Pelileo*. https://1library.co/document/q5w17n3q-modelo-gestion-cobranza-disminuir-cooperativa-credito-pakarymuy-pelileo.html. Visitada el 24-06-2020.
5. Alfaro, R., Gallardo, N., & Stein, R. (2012). *The determinants of household debt default*. Revista de análisis económico, *27*(1), 57-70. https://doi.org/10.4067/S0718-88702012000100003. Visitada el 14-07-2020.
6. Biron, M. & Medina, V. (2018). *Comparación de algoritmos de clasificación para el incumplimiento crediticio. Aplicación al sistema bancario chileno*.

https://www.cmfchile.cl/portal/publicaciones/610/articles-29852_doc_pdf.pdf. Visitada el 13-08-2020.

7. Yeh, I.-C., & Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2, Part 1), 2473-2480. https://doi.org/10.1016/j.eswa.2007.12.020. Visitada el 13-08-2020.
8. Gurný, P., & Gurny, M. (2013). Comparison of Credit Scoring Models on Probability of Default Estimation for US Banks. *Prague Economic Papers*, *22*, 163-181. https://doi.org/10.18267/j.pep.446. Visitada el 13-08-2020.
9. Cela, G., & Cuenca, J. P. (2019). *Propuesta de modelo de machine learning para la evaluación de riesgo de crédito utilizando algoritmos de predicción para la Cooperativa de Ahorro y Crédito La Merced Ltda.* https://www.researchgate.net/publication/337480778_Propuesta_de_modelo_de_machine_learning_para_la_evaluacion_de_riesgo_de_credito_utilizando_algoritmos_de_prediccion_para_la_Cooperativa_de_Ahorro_y_Credito_La. Visitada el 17-08-2020.
10. Trujillo Fernández, D. (2017, junio). *Aplicación de metodologías Machine Learning en la gestión de riesgo de crédito* [Info:eu-repo/semantics/bachelorThesis]. E.T.S. de Ingenieros Informáticos (UPM). http://oa.upm.es/47087/. Visitada el 17-08-2020.

11. Cavallo Parra, D. (2002). *La lucha por evitar el default y la devaluación*. La Nación. https://www.lanacion.com.ar/economia/la-lucha-por-evitar-el-default-y-la-devaluacion-nid392083/. Visitada el 24-06-2020.
12. Sagner T, A. (2012). *El influjo de cartera vencida como medida de riesgo de crédito: análisis y aplicación al caso de chile*. Revista de análisis económico, *27*(1), 27-53. https://doi.org/10.4067/S0718-88702012000100002. Visitada el 17-08-2020.
13. Li, S., Wang, M., & He, J. (2013, mayo 12). *Prediction of Banking Systemic Risk Based on Support Vector Machine* [Research Article]. Mathematical Problems in Engineering; Hindawi. https://doi.org/10.1155/2013/136030. Visitada el 17-08-2020.
14. Rendón Morán, Karen. (2018). *Modelo de gestión de cobranzas para recuperación de cartera vencida en Tablicon S. A.* http://repositorio.ug.edu.ec/handle/redug/30581. Visitada el 24-06-2020.
15. Ossandón, J. (2012). *Destapando la Caja Negra: Sociologías de los Créditos de Consumo en Chile* [MPRA Paper]. https://mpra.ub.uni-muenchen.de/42181/. Visitada el 24-06-2020.

Anexo # 1

Nombre Variable	Definición de la Variable
ID_Mes	Código que identifica el mes de la toma de los registros.
Cod_Convenio	Código alfanumérico que identifica la entidad donde se desempeña el cliente.
Grupo_Producto	Nombre de clasificación del tipo de crédito cursado.
Ambito	Sector de la empresa donde trabaja el socio, público, privado, pago directo.
Fecha_Apertura	Fecha de curse del crédito.
Tramo_Mora	Clasificación de la operación en base a los días de mora.
Edad_Socio	Edad del Cliente al mes de registro.
Sexo_Socio	Sexo Cliente.
Fecha_Nacimiento	Fecha Nacimiento Cliente.
Antiguedad_Laboral	Antigüedad en la empresa donde se desempeña el cliente.
Cuotas_Pactadas	Plazo en meses del crédito de consumo.
Cuotas_Canceladas	Cuotas canceladas al momento del registro mensual.
Dias_Mora	Días de atraso en el pago de la cuota vigente.
Operación_Renegociada	Indicador si la operación tuvo alguna modificación en las condiciones por falta de capacidad de pago.
Saldo_Credito	Saldo vigente del crédito al momento del registro.
Monto_Total_Credito	Monto total del crédito cursado.
Oficina	Sucursal donde fue cursado el crédito.
Regional	Gerencia Zonal donde fue otorgado el crédito.
Tipo_Oficina	Tipo de sucursal que cursó el crédito (Comercial, Comercial Liviana, Venta).
Valor_Cuota	Valor mensual acordado para pago del crédito.
Tasa_Interes	Precio al cual se cursa el crédito de consumo.

Operación_Reliquidada	Indicador si la operación en mora fue pagada completamente cursando una nueva operación de crédito.
Operación_Castigada	Indicador de castigo de la operación (1 – Castigada 0- No castigada).
Campaña_Contencion	Asignación de registro de cobranza a un ejecutivo para su gestión.
Campaña_Gestionada	Caso de campaña gestionado por el ejecutivo, es decir, el ejecutivo se contacta con el cliente para informar su mora.
Campaña_Cerrada	Campaña Gestionada sin éxito y cliente no contactado.
Campaña_Gestion_Exitosa	Contacto efectivo con el cliente donde se informa su mora y que proceda a regularizar su caso.
Op_mora	Número total de operaciones con mora del cliente al momento del registro.
Op_Vigentes	Número total de operaciones vigentes del cliente al momento del registro.
Deuda_Interna_Consumo	Deuda interna del cliente. Endeudamiento total en la cooperativa.
Renta_Liquida	Remuneración líquida del socio al momento de cursar de crédito.
Nivel_Educacional	Último nivel de educación cursado por el cliente.
Deuda_Ext_Consumo	Deuda de crédito de consumo del socio declarada en el sistema financiero (declarada por todos los participantes de la industria), información provista por la Superintendencia de bancos e instituciones financieras.
Deuda_Ext_Hipoteca	Deuda de crédito hipotecario del socio declarada en el sistema financiero (declarada por todos los participantes de la industria), información provista por la Superintendencia de Bancos e Instituciones Financieras.