Universidad del Desarrollo
Facultad de Ingeniería

MODELO DE PROPENSIÓN A LA TOMA DE CRÉDITOS DE CONSUMO EN UNA EMPRESA DEL SECTOR FINANCIERO

POR: JOSÉ IGNACIO AYALA MESA

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:

Sr. PABLO REINOSO

AGOSTO 2020

SANTIAGO

Modelo de propensión a la toma de créditos de consumo en una empresa del sector financiero

José Ignacio Ayala Mesa

Agosto 2020

Abstract

Una tendencia a nivel mundial en el sector financiero y que también se está dando en Chile, es la utilización de diversas técnicas que permiten sacar provecho al creciente volumen de información que día a día están acumulando las empresas. En este paper, buscamos aplicar específicamente, técnicas de machine learning para desarrollar un modelo que permita predecir si un cliente tomará o no un crédito de consumo durante los próximos tres meses. La base de clientes trabajada fue un desarrollo conjunto con una cooperativa de ahorro y crédito del sector financiero, con presencia a nivel nacional, con quienes se construyó un dataset de 265.332 clientes de un período de 20 meses, donde identificamos diferentes variables y evaluamos cómo estas inciden en el comportamiento crediticio de cada persona. Entre las variables de mayor importancia aparecen el nivel de ingresos de cada cliente y sus niveles de endeudamiento. El segmento de mayores ingresos duplica su nivel de endeudamiento promedio por cliente con respecto al segundo segmento de mayor ingreso. Otra variable de importancia es la que permite discriminar entre empresas del sector público y del sector privado, donde detectamos diferencias al momento en que un cliente toma un crédito.

Con la aplicación del algoritmo de clasificación seleccionado, pudimos discriminar con una probabilidad de hasta un 90% entre clientes que toman y no toman un crédito de consumo. Los resultados que obtuvimos se explican principalmente por las diferentes aperturas hechas al modelo, junto con los ajustes de hiperparámetros aplicados, lo cual nos permitió escalar la precisión en cada una de las iteraciones del modelo desarrollado.

Keywords: XGBoost, boosting, classification algorithms, credit loan

1 Introducción

Actualmente una gran cantidad de empresas están comenzando a darle una real importancia a la toma de decisiones basada en los datos que se pueden obtener de diversas fuentes[1]. En el sector financiero, también es un tema que

se ha empezado a utilizar y que día a día toma mayor fuerza cuando se requiere evaluar importantes definiciones que afectan el core del negocio.

Un aspecto muy relevante y que va en directa relación con los clientes de todas las instituciones financieras, es el estudio de cómo estos se comportan en el mercado, específicamente en la elección de los productos de crédito que ofrece la industria. Dicho comportamiento puede ser asociado a distintas características que tenga cada persona en cuanto a la información financiera tanto pública, como propia de una empresa del sector financiero.

Reconocer e individualizar esas necesidades de crédito que cada cliente puede tener, es el desafío que abordamos en nuestro proyecto, a través de la aplicación de un modelo de aprendizaje automático.

En particular, nuestro trabajo tiene como objetivo principal, el desarrollo de un modelo de propensión al crédito de consumo en una cooperativa de ahorro y crédito del país, hecho que permitirá evaluar la efectividad de este modelo, al seleccionar a las personas que tienen las mayores propensiones a adquirir un crédito en los próximos tres meses.

La cooperativa de ahorro y crédito con la que desarrollamos el proyecto, está inmersa en la industria financiera, siendo hoy un actor relevante en la banca minorista a nivel nacional, con un fuerte foco de sus operaciones en el crecimiento de sus activos asociados a los préstamos de créditos de consumo. Un aumento de sus colocaciones de créditos a personas naturales y por ende, el aumento del margen financiero asociado a este tipo de clientes, ha sido y seguirá siendo parte de las definiciones estratégicas que tiene esta empresa. Asociado a esta definición de negocio, la cooperativa ha tenido un importante desarrollo en temas relacionados con la explotación de datos que puede capturar de sus clientes, mediante herramientas de CRM y, por esta misma razón, todo proyecto o idea asociada al análisis y búsqueda de insights que aporten al negocio, está entre las tareas que quieren seguir potenciando como compañía en constante crecimiento.

La cooperativa trabaja con asignaciones de clientes pre aprobados (campañas de crédito) en los diferentes canales de venta disponibles (plataforma, ventas en terreno, ventas online), pero no cuenta con un modelo de propensión a la toma de créditos de consumo, que seleccione los casos con las mayores probabilidades de éxito. Dichas asignaciones de campañas se hacen solamente discriminando los casos por cada canal de venta (un cliente asignado a un solo canal de venta / ejecutivo), con el fin de no llamar por diferentes vías a cada persona que se pretende contactar. Para el desarrollo de nuestro modelo de propensión, el proyecto se basa en un dataset elaborado con información financiera pública con la que cuentan todas las instituciones financieras supervisadas por la CMF (Comisión para el Mercado Financiero, ex SBIF)[2] e información propia de los clientes de la compañía. Aquí, el crédito de consumo y toda la data asociada a

este producto es una parte medular para el desarrollo del proyecto.

Dado lo anterior y revisando el detalle del producto consumo, existen diferentes políticas de admisión de clientes, para los cuales existen dos grandes clasificaciones: créditos nuevos y créditos reliquidados. Estos últimos corresponden a créditos que, al ser otorgados, prepagan una operación anterior que el cliente tenía vigente con la cooperativa, además de otorgársele un monto extra de libre disposición. Para la construcción del modelo se trabajó con la cartera de créditos reliquidados, por dos motivos:

1. Los créditos reliquidados alcanzaron el 48% del total de ventas de consumo de la cooperativa. Durante el año 2019, se vendieron a nivel país, 403.000 millones de pesos en créditos de consumo reliquidados.

2. Los créditos reliquidados se otorgan siempre a clientes antiguos de la cooperativa, por lo tanto, para este grupo se conoce en mayor profundidad su historial y su comportamiento crediticio. Se cuenta con más historia de cómo han evolucionado los clientes en el tiempo, por ejemplo, en la capacidad de pago de sus deudas adquiridas.

En línea con las políticas de admisión, las campañas de pre aprobaciones mensuales separan a los clientes entre los que toman un crédito por primera vez y los que toman créditos reliquidados. Estos dos grupos siempre tienen distintas ofertas, cada uno con distintos niveles en montos, tasas y plazos. Al cliente antiguo que toma un crédito con reliquidación, se le puede estudiar en mayor profundidad y su comportamiento es de mayor conocimiento por parte de la empresa. La consecuencia principal de esto, es que las ofertas a clientes antiguos que reliquidan un crédito siempre son más atractivas en tasas y montos, que las ofertas a clientes nuevos. Un ejecutivo generalmente busca créditos de mayor monto, para lograr de manera más rápida sus objetivos comerciales del mes (cumplimientos de metas de venta). Dado esto, definir la oferta de pre aprobados a clientes antiguos se vuelve una decisión de gran importancia, debido a que tiene fuertes impactos en los resultados personales de cada ejecutivo (las rentas variables que recibe) y en los resultados de la cooperativa, impactando de manera directa la línea de margen financiero.

2 Trabajos Relacionados

Diversos trabajos asociados a técnicas de analítica y machine learning se han llevado a cabo en el ámbito financiero con el fin de mejorar diversas inquietudes que siempre han tenido las instituciones de este sector, tanto en Chile, como en el resto del mundo. A nivel nacional, uno de los casos de estudio asociado a la banca es el realizado por la Superintendencia de Bancos e Instituciones Financieras el año 2018[3] (actual CMF). Este trabajo plantea la aplicación y comparación de diversos modelos de clasificación para determinar

la probabilidad de default, es decir, la probabilidad de que un cliente tenga un incumplimiento en el pago de sus deudas. Lo que desarrolla la SBIF difiere con nuestro proyecto en que se aplica específicamente a los créditos comerciales y no de consumo, además de que la variable objetivo también es distinta a lo que nosotros buscamos. Lo relevante y que está en línea con nuestro desarrollo es la utilización en común de algunas de las variables independientes que fueron de relevancia en nuestro modelo (ingresos de los clientes, ventas, deudas de cada producto).

Los autores Munkhdalai, Namsrai, Lee y Ryu [4] (2019) publican en la revista académica Sustainability del MDPI, otro estudio en el cual se aplican modelos de clasificación en el sector financiero para determinar puntajes crediticios de los clientes que inciden directamente en las pérdidas por riesgo de crédito. Este estudio tiene dos puntos interesantes: el primero es que compara los resultados obtenidos con los modelos de clasificación versus el criterio de una persona experta en la determinación del puntaje de crédito. Aquí, los resultados arrojaron que la aplicación de algoritmos de aprendizaje automático habrían permitido generar menores niveles de pérdidas en comparación con el sistema tradicional de evaluación crediticia. El segundo punto es que en este estudio utilizan diversos algoritmos, pero los dos con mejores resultados fueron las redes neuronales y el algoritmo XGBoost. Este último, refuerza el hecho de que es el modelo que elegimos en nuestro proyecto para la determinación de un modelo de propensión a la toma de créditos de consumo.

Otro trabajo publicado en MDPI (2019) por los autores Niu, Ren y Li [5], referente a los scores de incumplimiento crediticio, pero con la particularidad de que realiza un estudio comparativo para entender si la información publicada en redes sociales, ayuda a mejorar la capacidad predictiva en las instituciones financieras. Este trabajo aplica los algoritmos de clasificación: random forest, AdaBoost y LightGBM. estos dos últimos siempre relacionados y compitiendo con el algoritmo XGBoost aplicado en nuestro proyecto. Los resultados que obtienen indican que la información en redes sociales puede ser muy útil a la hora de mejorar la débil información que en ocasiones se tiene de las personas que solicitan un crédito.

El trabajo realizado por Alexandru Coser [6], publicado en el ECECSR Journal (2020) aborda con algoritmos de clasificación, otro tema que es de gran importancia en la banca. Este estudio hace referencia a la predicción de la fuga de clientes de una institución financiera, junto con el estudio de las principales características que tienen los clientes que abandonan un banco. Aquí también se hace uso de modelos de clasificación para estimar la probabilidad de que un cliente se convierta en fuga, utilizando el área bajo la curva (AUC) para evaluar el rendimiento de los dos modelos aplicados.

Trabajos de predicción de fugas también se aplican en otras industrias, tal como se hizo en la publicación del autor Mishachandar (2017), donde realiza un análisis al comportamiento de los clientes en la industria de las telecomunica-

ciones [7] a través del algoritmo de clasificación Naïve Bayes, para finalmente reconocer a aquellos con mayores probabilidades de fuga y así aplicar técnicas de retención de clientes.

3 Dataset

La cooperativa para la cual desarrollamos este trabajo tiene presencia a nivel nacional, con un total de 83 oficinas administradas por nueve gerencias regionales (Norte Grande, Norte Chico, Quinta, Centro Sur, Octava, Sur, Austral, Oriente y Poniente). El dataset trabajado se focalizó en la regional Norte Chico, la cual consta a día de hoy, con ocho oficinas comerciales. El motivo de utilizar solamente la información de esta gerencia regional pasa por mantener un proyecto acotado en cuanto a detalle de información y tiempos de desarrollo, que sirva como piloto para evaluar el modelo que estamos planteando. En un trabajo futuro dentro de la cooperativa, tenemos acordado abordar otras gerencias regionales, lo que implica necesariamente evaluar nuevas variables y adaptar el modelo a las distintas realidades de cada zona del país. Por ejemplo, la zona norte de Chile aborda clientes que difieren en sus rubros con respecto a los del sur del país, donde también hay diferencias en cuanto a morosidad, endeudamiento, niveles de cesantía, carga financiera, que implican una adaptación de nuestro trabajo para cada zona.

La tabla 1 muestra las oficinas de la gerencia regional y la cantidad de clientes que hoy son atendidos en la zona.

Tabla 1 Clientes de la cooperativa versus habitantes para la Gerencia Regional Norte Chico

Oficina	N° de Clientes	Habitantes	% Cobertura (clientes / Hbtes.)
La Serena	17.365	195.382	8,9%
Copiapó	15.246	175.172	8,7%
Ovalle	11.182	111.272	10,0%
Vallenar	7.796	51.917	15,0%
Coquimbo	5.640	204.068	2,8%
Illapel	2.223	39.910	5,6%
Diego de Almagro	1.174	13.925	8,4%
Los Vilos	440	21.382	2,1%

El dataset construido cuenta con 20 meses de historia, desde Julio 2017 hasta febrero 2019, con un total de 265.332 registros. Actualmente, esta regional tiene 16.266 clientes con crédito, es decir, clientes que actualmente tienen una deuda de consumo con la empresa.

En cuanto a la cantidad de variables, el dataset cuenta con 43 columnas,

donde hay 9 variables categóricas y 34 variables numéricas, que están relacionadas principalmente a productos de crédito y productos de pasivos (ahorros) que tiene cada cliente, más la tenencia de productos de medios de pago (tarjetas de crédito y/o cuentas vista). Entre estas variables también se cuentan las que están asociadas a la información pública del sistema financiero y que corresponde a las deudas que cada persona tiene en el sistema.

Mayor detalle de todas las variables que incorporamos en el dataset de nuestro proyecto, puede ser revisado en anexo 1.

4 Metodología

Para la obtención de un modelo de machine learning que prediga si un cliente tomará o no un crédito de consumo, establecimos las etapas secuenciales para cumplir este desafío. Nuestra metodología pretende ser una guía para entender de manera práctica el cómo avanzar paso a paso a la hora de construir un modelo de predicción que pueda ser utilizado en la realidad de la empresa. Las etapas de la metodología son las siguientes:

- Análisis Exploratorio de Datos (EDA)
- Feature Engineering
- Análisis y Resultados
  - Aplicación inicial del algoritmo XGBoost
  - Resampling de los datos
  - Feature Selection
  - Ajuste de hiperparámetros
  - Aperturas del modelo (segmentación)

A continuación explicamos cada uno de estos puntos en detalle.

5 Análisis Exploratorio de Datos (EDA)

Como foco de análisis entre las distintas variables, tenemos la variable objetivo (de nombre Ind-Venta+3M), la cual toma solamente dos valores: 1 o 0. Esta variable tendrá un valor de 1 cuando el cliente registre la toma de un crédito de consumo durante los siguientes 3 meses. En caso contrario, la variable toma el valor 0 (cero). La distribución de la variable objetivo se puede ver en la figura 1.

El alcance de nuestro proyecto con respecto a la predicción de la variable objetivo, solamente aborda si el cliente toma o no toma un crédito de consumo.

Aquí no será relevante el monto del crédito de consumo pre aprobado, porque dicha labor recae principalmente en las áreas de riesgo de crédito de la cooperativa, no es un tema en el que nosotros tengamos la facultad de incidir de manera independiente.

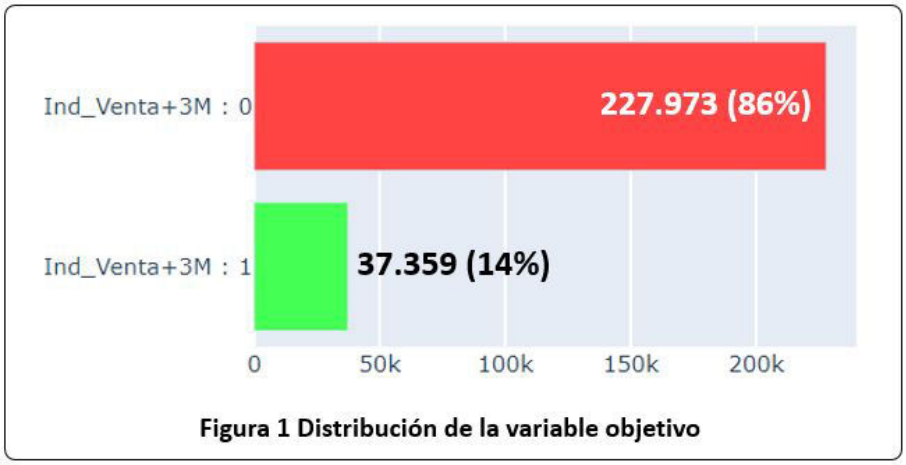


Figura 1 Distribución de la variable objetivo

La variable objetivo tiene un grado de desbalanceo entre los valores positivos (personas que toman un crédito en los siguientes tres meses) versus las personas que no toman un crédito. Para el primer caso, se tienen 37.359 clientes, que equivalen al 14% del total de datos, mientras que, para el segundo caso, los clientes que no toman un crédito son 227.973 y equivalen al 86% de los datos. Este punto es de relevancia al momento de generar un modelo de propensión. En un punto posterior se verá el balanceo del dataset con el objetivo de lograr un mejor resultado en la clasificación que entregue el modelo.

Un punto interesante y que también se debe analizar, tiene relación con las diferencias entre géneros que existen en el dataset construido. La figura 2, lo evidencia.

El dataset muestra que la cantidad de mujeres es un 140% mayor a la cantidad de hombres (187.379 vs 77.953), situación muy distinta a la realidad de las dos regionales a las que atiende la cooperativa (Figura 3). El porcentaje de hombres versus mujeres de la 3a y 4a región tiene una mínima variación positiva en favor del género femenino (51% vs 49%), sin embargo, los clientes de la cooperativa tienen una distribución muy distinta, donde anivel total, el 71% de ellos son del género femenino.

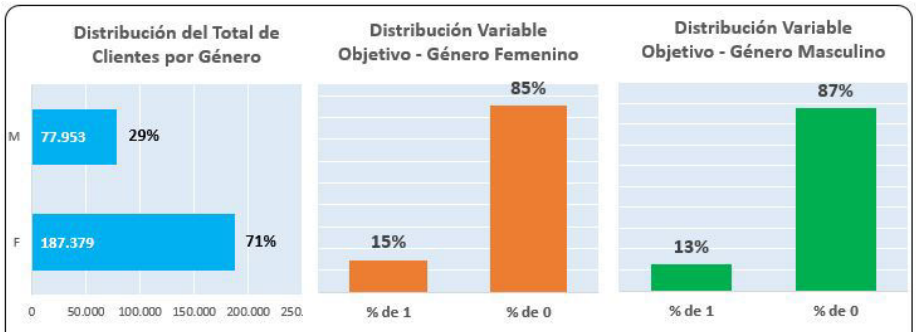


Figura 2 Distribución de clientes por género / Distribución variable objetivo por género

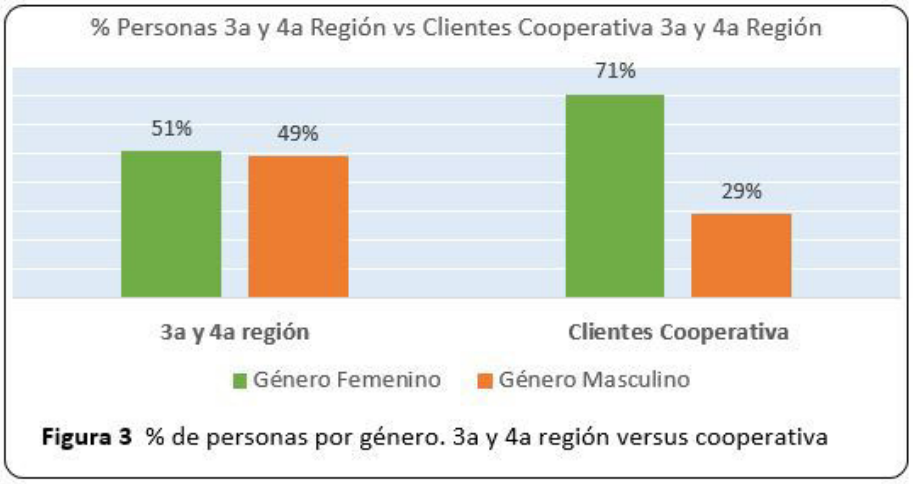


Figura 3 % de personas por género. 3a y 4a región versus cooperativa

La figura 2 también indica que se produce el efecto de que el género femenino registra un mayor porcentaje de toma de créditos que el género masculino, con un 15% versus un 13% respectivamente.

Para entender de mejor manera cómo se comporta la variable objetivo, se analizan a continuación otras aperturas en base a las distintas características que se incorporaron en el dataset.

Los 4 segmentos que actualmente tiene la cooperativa clasifican a los clientes en base a dos variables: edad y renta. Tanto el segmento Base, como el Emergente tienen la renta de cada cliente como la principal variable de discriminación. En cambio, para los segmentos adulto joven y adulto medio, se clasifican prin-

cipalmente por la variable edad.

Importante destacar que en el segmento Base están los clientes con los menores niveles de sueldo. Aquí también están contenidos, todos aquellos que trabajan jornadas menores a las 45 horas semanales y por ende, tiene rentas bajo los valores de salario mínimo.

Al revisar la apertura de la variable objetivo en los distintos segmentos que maneja la empresa, podemos entender que el segmento "Emergente" es el de menor porcentaje de toma de créditos de consumo con respecto al total de clientes del segmento (figura 4).

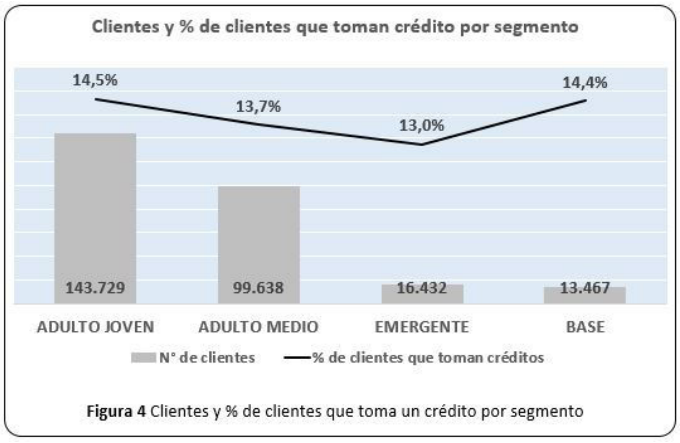


Figura 4 Clientes y % de clientes que toma un crédito por segmento

El menor porcentaje de toma de créditos del segmento emergente puede estar directamente relacionado con los niveles de ingreso de los clientes que componen dicho segmento. La figura 5 muestra la renta líquida promedio de cada segmento, donde se puede apreciar claramente la fuerte diferencia del segmento emergente con el resto de los grupos.

Un mayor nivel de ingresos como el que registra el segmento emergente, está directamente relacionado con el acceso que tienen dichos clientes a la banca tradicional. Su mayor capacidad de endeudamiento los convierte ya, en un grupo interesante a los cuales se les entrega una oferta tradicional de productos financieros en el mercado, situación contraria a la que tienen los segmentos base, adulto medio y adulto joven de la cooperativa. En línea con lo anterior, las diferencias de sueldo, inciden directamente en la capacidad de endeudamiento que tiene cada segmento. La figura 6 muestra la diferencia que se da entre los

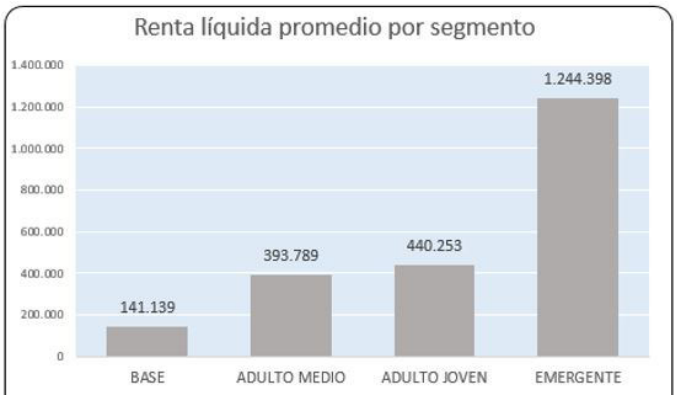


Figura 5 Renta líquida promedio de socios de la cooperativa

clientes emergentes con respecto al resto de los segmentos en cuanto a montos de los préstamos promedio que tiene cada segmento(segmento emergente duplica al siguiente segmento en deuda de consumo en la cooperativa).

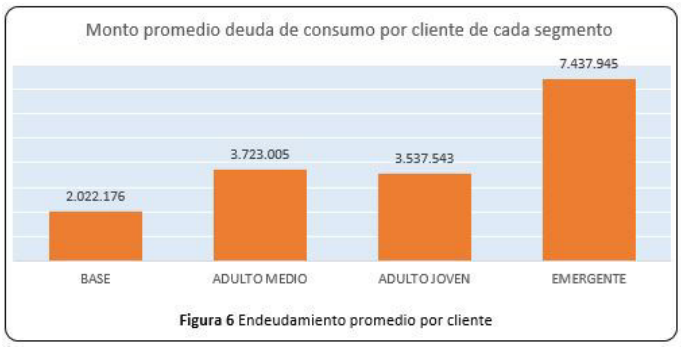


Figura 6 Endeudamiento promedio por cliente

En línea con la renta y el endeudamiento del segmento emergente y su capacidad de adquirir deuda en el sistema financiero tradicional (bancos), se puede indicar que la variable "Share of Wallet" permite reconocer las diferencias de renta entre los clientes menos bancarizados (deuda del cliente en la cooperativa) y los más bancarizados (deuda predominante en la competencia).

El "Share of Wallet" "COMPETENCIA" indica que el cliente de la cooperativa tiene un mayor nivel de endeudamiento en otras instituciones financieras controladas por la CMF, mientras que el "Share of Wallet" "COOPERATIVA" implica que el cliente tiene una mayor deuda de consumo dentro de la cooperativa. Las diferencias más acentuadas se registran nuevamente en el segmento

Renta Líquida Promedio

	COMPETENCIA	COOPERATIVA
EMERGENTE	1.606.155	1.117.406
ADULTO JOVEN	515.638	425.767
ADULTO MEDIO	489.719	379.187
BASE	158.858	152.908

Tabla 2 Share of Wallet en renta líquida

emergente, donde los clientes con su "share of wallet" en la competencia (con la mayor parte de su deuda fuera de la cooperativa) tienen una renta líquida que es un 44% más alta que los clientes con su "share of wallet" en la cooperativa (su mayor nivel de deuda está en la cooperativa). Estos clientes que tienen su mayor nivel de endeudamiento fuera de la cooperativa y que tienen los mayores sueldos entre el total de clientes de esta institución, son personas que tienen un mayor nivel de bancarización, no tienen a la cooperativa como su principal institución financiera y esto puede ser también motivo para que la toma de créditos sea levemente menor en este segmento.

Para terminar la revisión de la variable objetivo, no se puede obviar el tipo de empresas a los que atiende la cooperativa. El core del negocio financiero de la empresa está en los créditos de consumo con descuento por planilla. Para esto, la cooperativa logra acuerdos con las empresas a cuyos funcionarios se les otorgan productos de crédito, de ahorro y medios de pago. Una vez que se vende un crédito con esta modalidad de recaudación por planilla, la empresa con la que la cooperativa establece el acuerdo, descuenta las cuotas a sus funcionarios y realiza el pago de la deuda a la cooperativa. Así, entre los mayores beneficios de esta modalidad de recaudación, está el mantener un riesgo de crédito controlado y que permite a la cooperativa entregar apoyo financiero a una parte importante de la población que no tiene acceso a la banca.

Una distinción importante en estas empresas en convenio tiene que ver con su pertenencia al sector público o privado. Aquí, también se producen diferencias importantes que deben ser consideradas a la hora de construir modelos de propensión. La figura 7 evidencia que, en el sector público, la proporción de clientes que toma un crédito versus el total de clientes, es similar en todos los segmentos, salvo en adultos jóvenes, donde tiene un aumento por sobre el 1%

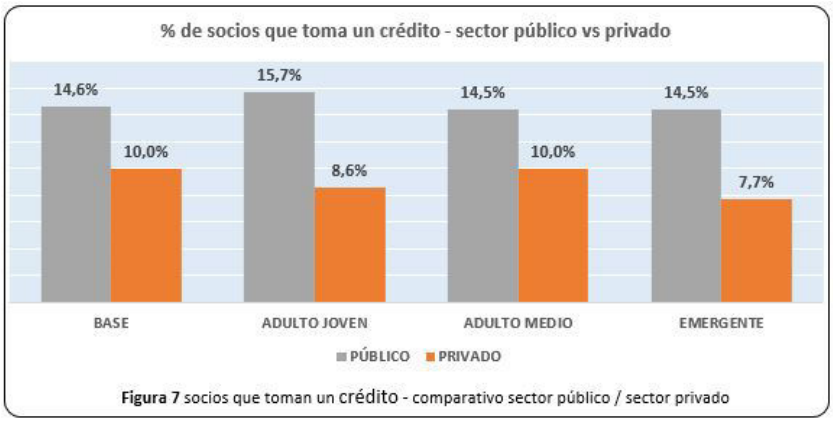


Figura 7 socios que toman un crédito - comparativo sector público / sector privado

Por el contrario, en el sector privado, tanto el segmento emergente, como el segmento adulto joven, registran fuertes caídas en el porcentaje de clientes que toman un crédito de consumo, llegando al 7,7% y 8,6% del total, respectivamente.

6 Feature Engineering

El punto de mayor relevancia aquí es todo lo que se trabajó con la incorporación y creación de variables para agregar al dataset, con el fin de mejorar la precisión del modelo. Inicialmente, se incorporó la variable de la deuda en el sistema financiero que tiene cada cliente de la cooperativa y a partir de esta se crearon las siguientes variables: variación de la deuda en el sistema financiero de 3 meses hacia atrás, 6 meses hacia atrás y 12 meses hacia atrás.

A partir de las cuotas pactadas de cada cliente en sus créditos de consumo y la cantidad de cuotas ya pagadas de su deuda, se crea la variable "Avance". Esta variable corresponde al cociente entre cuotas pagadas y cuotas pactadas y se deja expresada como un entero. La utilización de esta variable se sustenta en el hecho de que existe un importante volumen de créditos (40 por ciento) que son reliquidados por los clientes dentro de las primeras 12 cuotas de avance en el pago. La figura 8 muestra una alta concentración de la variable entre el 0% y el 20% en el avance del pago de las cuotas de la cartera de créditos.

Una siguiente variable creada e incorporada al dataset es la variable SaldoRenta, que corresponde al cociente entre el saldo de consumo de cada cliente y la renta líquida que cada uno de ellos tiene. Esta variable nos da una medida aproximada de a cuántas veces su nivel de renta equivale la deuda de consumo que tiene el cliente con la empresa.

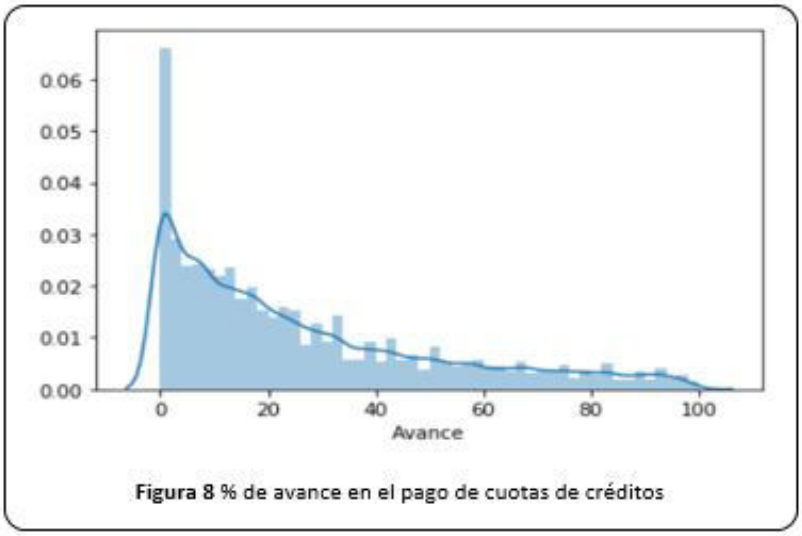


Figura 8 % de avance en el pago de cuotas de créditos

Creamos también la variable Reliq3A, la cual corresponde a la cuenta de la cantidad de veces que el socio tuvo reliquidaciones de créditos durante un período de 3 años. Esta variable permite tener una referencia con respecto a los clientes que constantemente reliquidan sus créditos con la cooperativa (reliquidar es cuando un cliente pide un nuevo crédito que prepaga la deuda previa y adicionalmente vuelve a pedir un monto extra de endeudamiento).

Otra incorporación de variable corresponde a la que llamamos PMT. Esta tiene relación con el nivel de endeudamiento máximo que puede tener cada cliente. Esto influye directamente en la toma de crédito que puede querer un cliente, limitándolo en base a su nivel de endeudamiento.

Por último, una variable de gran relevancia incorporada, tiene relación con el nivel de endeudamiento máximo que cada convenio le permite tener a sus funcionarios al endeudarse con la cooperativa. Este endeudamiento de cada cliente incide en el monto del crédito que puede solicitar cada persona y en su capacidad para volver a reliquidar el crédito que tiene con la compañía. Las principales diferencias entre los niveles de endeudamiento máximo que se dan entre los diferentes convenios que tiene la empresa, hacen referencia al tipo de empresa en convenio con la que se trabaja o el sector económico a la que pertenece.

7 Análisis y Resultados

Para el dataset construido, aplicaremos el algoritmo XGBoost con el fin de obtener un modelo de machine learning adecuado para la predicción de toma de créditos de consumo en los clientes que tiene la cooperativa.

Otros métodos utilizados en trabajos similares en la industria financiera, es el que llevó a cabo la actual CMF (ex-SBIF, Superintendencia de Bancos e Instituciones Financieras), donde desarrollan un trabajo comparativo de algoritmos de clasificación. La diferencia de este trabajo con el que nosotros planteamos, es que su objetivo es detectar el incumplimiento crediticio de los clientes del sistema financiero nacional, además de estar enfocado en la cartera de créditos comerciales otorgados a personas naturales. Nuestro trabajo aborda los créditos de consumo para personas naturales (la diferencia entre estos tipos de crédito radica principalmente en el uso que se le da a cada tipo de préstamo, un crédito comercial se utiliza principalmente para financiar proyectos o emprendimientos de una persona, pero asociados a su personalidad jurídica).

El trabajo realizado por la CMF toma 8 algoritmos distintos de clasificación, siendo uno de ellos el modelo de gradient boosting machines (GBM), similar al algoritmo XGBoost. El propio autor del algoritmo XGBoost (Tianqi Chen, [8]) explica la diferencia, indicando que "XGBoost utiliza una formalización del modelo más regularizada para controlar el overfitting, lo cual mejora el rendimiento". Adicionalmente, el data scientist Vikesh Singh Baghel [9] explica de la siguiente manera la diferencia entre GBM y XGBoost: "Para calcular la tasa de paso o avance del algoritmo, GBM lo hace en dos etapas, sin embargo, XGBoost ejecuta esto en una sola etapa, donde además añade regularización en la función de pérdida para contrarrestar el overfitting".

Este algoritmo utiliza una sumatoria de muchos clasificadores débiles, en particular árboles de decisión, que se procesan de manera secuencial y se asocian a esta función de pérdida en uno de los hiperparámetros del modelo. Cada vez que XGBoost genere un árbol, este buscará minimizar la función de pérdida que se especifique (la función más común que se utiliza es la raíz del error cuadrático medio, RMSE). Esta sumatoria de clasificadores débiles, termina formando un clasificador fuerte, donde el objetivo es minimizar el RMSE a través de una adecuada tasa de aprendizaje (el parámetro indicado).

El motivo de la selección de este algoritmo tiene que ver con dos temas: velocidad de ejecución para dataset acotados como es el caso de nuestro proyecto y rendimiento del modelo. La velocidad de ejecución se debe principalmente a que solamente trabaja con datos numéricos, donde siempre debe haber una preparación previa con todas las variables object que se puedan tener en el dataset. Una vez que se dejan todas las variables como numéricas, la rapidez de procesamiento es superior a gran parte de los algoritmos de clasificación.

En tanto, el rendimiento del modelo se puede justificar, al ver que es uno de los algoritmos que domina, a día de hoy, las diversas competencias de machine learning con modelos de clasificación. Adicionalmente, el algoritmo XGBoost logra un gran rendimiento porque tiene un amplio rango de hiperparámetros que permiten hacer ajustes al modelo y así mejorar sus resultados.

Una muestra de lo comentado, lo menciona el autor del artículo "XGBoost Algorithm: Long May She Reign!", Vishal Morde [10], donde se mencionan la paralelización, la "poda de árboles" y la optimización de hardware como las variables que otorgan a XGBoost una excelente velocidad y eficiencia en el uso de los recursos. Adicionalmente, este autor comenta las mejoras de este algoritmo para aumentar la eficiencia. Por último, el artículo referenciado muestra un comparativo donde el primer lugar, tanto en la predicción (curva AUC), como en el tiempo de entrenamiento, los tiene el algoritmo XGBoost.

7.1 Aplicación Inicial del Algoritmo XGBoost

La aplicación inicial de este algoritmo de boosting tiene por objetivo evaluar las predicciones de entrada que se obtienen, utilizando la totalidad de las variables consideradas en etapas previas de este trabajo. Para el entrenamiento del modelo se definió un corte del 80% del dataset, en tanto, para el testeo del modelo, se deja el restante 20%. Una vez instanciado y ajustado el modelo a los datos, se evalúan dos ítems: la precisión del modelo y el error cuadrático medio.

Medición	Modelo Inicial
Accuracy (train)	85,9%
Accuracy (test)	85,9%
RMSE	37,5%
AUC	0,63

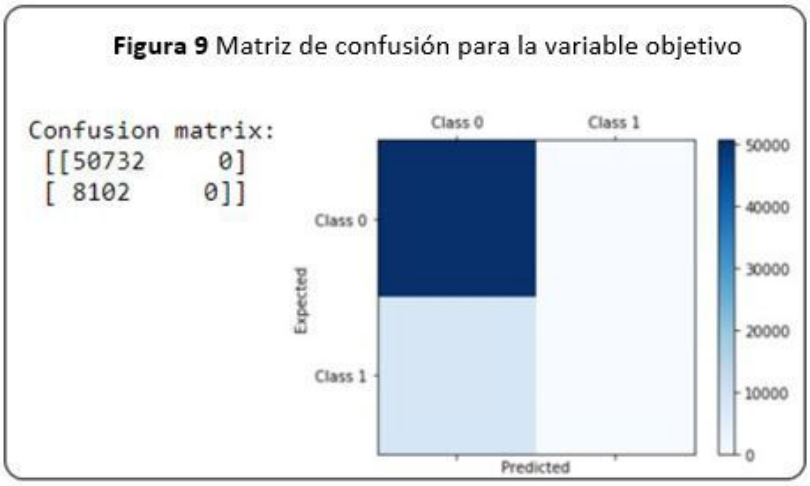
Tabla 3 Precisión y error del modelo

La tabla 3 muestra a simple vista que la aplicación del algoritmo sobre el dataset trabajado arroja un muy buen resultado en términos de precisión, tanto en el set de entrenamiento, como en el set de testeo. La mínima diferencia entre las dos precisiones que se muestran implica que no estamos en presencia

de sobreajuste. Sin embargo, el mayor inconveniente de esta aplicación radica en el desbalanceo del dataset. Como se indicó previamente, para la variable objetivo, que toma valores 1 o 0 (cero), existen 227.973 clientes que tienen la clase 0 (cero). Para la clase 1, hay 37.359 clientes. Dada esta situación, el algoritmo de clasificación XGBoost predice con mucha mayor frecuencia la clase más común, sin realizar análisis alguno a las variables involucradas en el modelo y, por consiguiente, arrojar un accuracy de 85,9% en la data de entrenamiento y testeo.

Dado lo anterior, una buena herramienta para analizar cómo fue la predicción del algoritmo, es la matriz de confusión, que muestra las predicciones correctas e incorrectas para cada clase.

Figura 9 Matriz de confusión para la variable objetivo



La figura 9 muestra que 50.732 clases 0 (cero) fueron predichas correctamente y 8.102 clases 1 fueron predichas erróneamente como clase 0.

Por último, con estos resultados obtenidos con el modelo "por defecto", se obtiene la curva ROC, que nos permite evaluar el rendimiento del modelo aquí planteado, en todos sus umbrales de clasificación. Esta curva representa la tasa de verdaderos positivos versus la tasa de falsos positivos.

El área bajo la curva (AUC), nos permite evaluar el rendimiento agregado del modelo en todos los umbrales de clasificación posibles y nos entrega la probabilidad de que el modelo planteado clasifique un valor positivo (clase 1) más alto que un valor negativo (clase 0). Con un valor AUC inicial de 0,63 se puede indicar que hay un 63% de probabilidad de que el modelo distinga entre la clase positiva (clase 1), es decir, que un cliente tome un crédito de consumo y la negativa (clase 0), donde el cliente no toma un crédito de consumo.

7.2 Resampling de los datos

Dado el desbalanceo que se tiene en el dataset, se utiliza la manera más simple de balancear un dataset; eliminar entradas de la clase 0 (cero) para equilibrar el número de entradas de la clase 1. El corte de la clase cero se hizo inicialmente en 72.000 entradas. Con esto, el dataset considera las 72.000 clases cero y 37.359 clases 1. A nivel total, el dataset queda con 109.359 entradas.

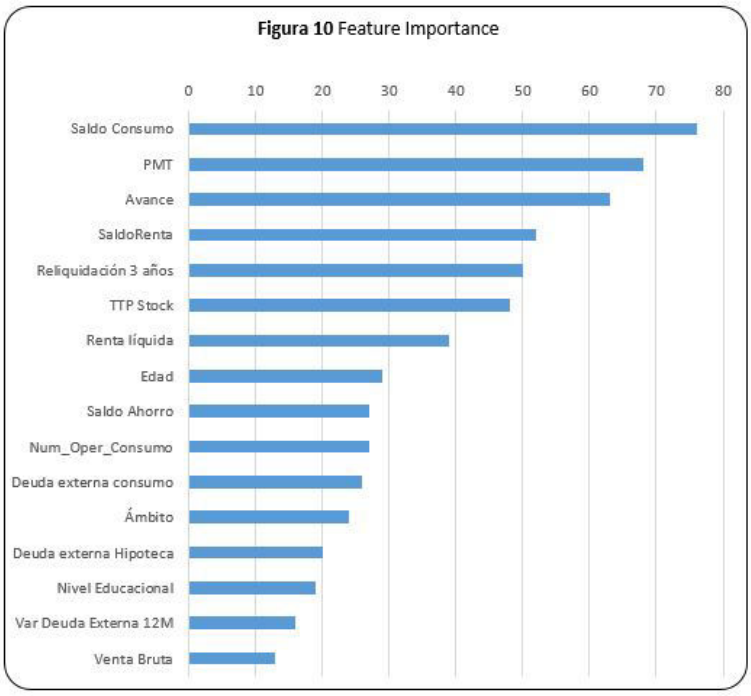
Medición	Modelo Inicial	Con Under Sampling
Accuracy (train)	85,9%	66,5%
Accuracy (test)	85,9%	66,5%
RMSE	37,5%	57,8%
AUC	0,63	0,63

Tabla 4 Precisión y error del modelo

Como lo indica la tabla 4, al hacer un under sampling al dataset, la precisión del modelo baja considerablemente, a un valor de 66,5% para entrenamiento y testeo respectivamente. En tanto, el error cuadrático medio sube a un 57,8% y la curva AUC, a un 0,63. Aún hay espacio para mejorar el modelo.

7.3 Feature Selection

Un beneficio de la librería del algoritmo XGBoost es que permite graficar las variables independientes del dataset en un ranking de mayor a menor importancia.



Para acotar las variables, se tomaron las que tenían puntajes sobre 50, ya que la siguiente en orden de importancia, bajaba su puntuación a 32. Dado lo anterior, las 6 variables independientes que se utilizaron para volver a generar el modelo son: Saldo de Consumo, Avance, Edad, TTP Stock (tasa del stock de créditos), Reliquidaciones de los últimos tres años y Renta Líquida. Con estos datos, se vuelve a ejecutar el algoritmo XGBoost. Como se puede ver en la tabla 5, al seleccionar las 6 variables de mayor importancia, el modelo tiene una merma en sus resultados, en base al accuracy y al error cuadrático medio. Sin embargo, lo relevante en esta etapa es que, con solamente 6 variables independientes, el modelo arroja prácticamente los mismos resultados con respecto al modelo que contaba con las 42 variables.

Medición	Modelo Inicial	Con Under Sampling	Con Feature Selection
Accuracy (train)	85,9%	66,5%	66,3%
Accuracy (test)	85,9%	66,5%	66,3%
RMSE	37,5%	57,8%	58,0%
AUC	0,63	0,63	0,61

Tabla 5 Precisión y error del modelo

7.4 Ajuste de hiper parámetros

Luego de balancear el dataset y seleccionar las variables de mayor importancia, se puede comenzar con la etapa de ajustar los hiper parámetros del algoritmo XGBoost.

- Max depth: que define el tamaño de los árboles de decisión. Con poca profundidad se logra una performance pobre, ya que dichos árboles capturan pocos detalles del problema. Con mucha profundidad, capturan mucho detalle y aparece el sobreajuste en los datos de entrenamiento y por lo tanto, limitando la capacidad del modelo para lograr buenas predicciones con datos nuevos.
- Learning rate: este hiper parámetro puede ser ajustado para controlar la ponderación de los nuevos árboles que se van agregando al modelo.
- N estimators: corresponde al número de árboles del modelo.

Estos parámetros principales, fueron ajustados de manera aleatoria, para ir viendo cómo los cambios de estos valores afectan el desempeño del modelo.

```
# Ajuste de hiperparámetros
cv_params = {'max_depth': [8], 'min_child_weight': [1]}
ind_params = {'learning_rate': 0.02, 'n_estimators': 1200, 'seed':0, 'subsample': 0.8, 'colsample_bytree': 0.8,
              'objective': 'reg:logistic'}
model = GridSearchCV(xgb.XGBClassifier(**ind_params), cv_params, scoring = 'accuracy', cv = 3, n_jobs = -1)
```

Figura 11 Hiperparámetros iniciales del algoritmo XGBoost

Como primera iteración, se comenzaron a ajustar manualmente los principales hiperparámetros. Con los valores de la figura 11, se obtienen los siguientes resultados (tabla 6).

Medición	Modelo Inicial	Con Under Sampling	Con Feature Selection	Hiper-parámetros
Accuracy (train)	85,9%	66,5%	66,3%	75,0%
Accuracy (test)	85,9%	66,5%	66,3%	68,9%
RMSE	37,5%	57,8%	58,0%	55,7%
AUC	0,63	0,63	0,61	0,72

Tabla 6 Precisión y error del modelo

El parámetro de mayor sensibilidad al momento de ajustar, es el hiperparámetro MaxDepth. Con el aumento de este por sobre 5, se produce sobre ajuste. Con los de la figura 11, se genera un sobre ajuste, ejemplificado con el cambio que tiene la precisión del modelo entre los datos de entrenamiento y los datos de testeo, donde se pasa de un 75% a un 68,9% y un error de 55,7%.

El área bajo la curva, nos indica un valor de 0,72, con una importante alza respecto del 0,61 obtenido en la versión previa (figura 12).

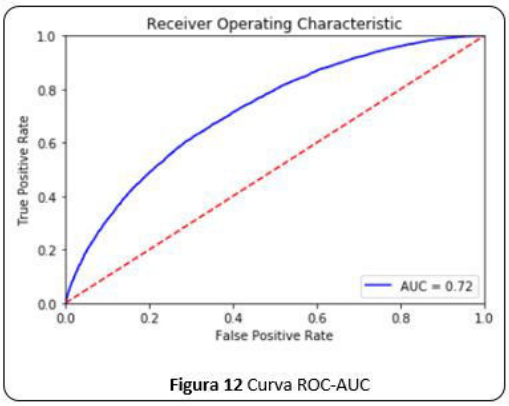


Figura 12 Curva ROC-AUC

Una segunda ejecución del modelo, modificando solamente el hiperparámetro

"n estimators" a 5000 y que se refiere a la cantidad de árboles de decisión que utilizará el algoritmo, si bien nos aumenta en forma importante los tiempos de procesamiento de la data, nos permite una mejora importante en la precisión del modelo. Aquí el cambio se refleja en la curva AUC, con un valor del 0,81.

Medición	Modelo Inicial	Con Under Sampling	Con Feature Selection	Hiper-parámetros	Hiper-parámetros
Accuracy (train)	85,9%	66,6%	66,5%	75,7%	90,6%
Accuracy (test)	85,9%	66,5%	66,5%	69,3%	74,0%
RMSE	37,5%	57,8%	57,9%	55,4%	50,9%
AUC	0,64	0,64	0,61	0,73	0,81

Tabla 7 Precisión y error del modelo

7.5 Aperturas del Modelo (segmentación)

Un tema en el que inmediatamente se marcan diferencias, tiene que ver con los segmentos de renta en los que la empresa clasifica a sus clientes. Al día de hoy, existen 4 segmentos distintos: Base, Adulto Joven, Adulto Medio y Emergente. Para este último, se dan dos hechos relevantes. Primero, que el nivel de renta líquida promedio de los clientes de este segmento, es un 182% mayor que el segundo segmento en nivel de renta (ver figura 13). Segundo, que los niveles de endeudamiento que alcanzan los clientes Emergentes, al menos duplican al resto de los segmentos.

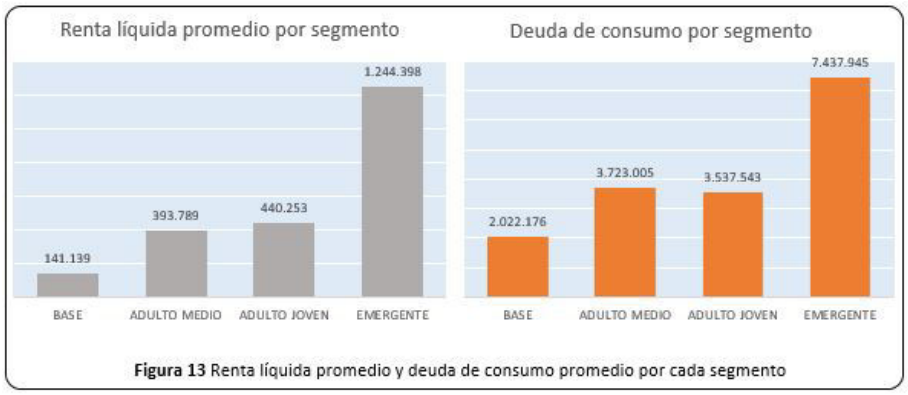


Figura 13 Renta líquida promedio y deuda de consumo promedio por cada segmento

En base a esta información y a partir del último modelo trabajado con todos los segmentos, se generan dos nuevos modelos de clasificación, separando el segmento Emergente, del resto de los segmentos. Dado lo anterior, los hiperparámetros ajustados para ambos modelos son los siguientes:

```
# Ajuste de hiperparámetros
cv_params = {'max_depth': [8], 'min_child_weight': [1]}
ind_params = {'learning_rate': 0.02, 'n_estimators': 5000, 'seed':0, 'subsample': 0.8, 'colsample_bytree': 0.8,
              'objective': 'reg:logistic'}
model = GridSearchCV(xgb.XGBClassifier(**ind_params), cv_params, scoring = 'accuracy', cv = 3, n_jobs = -1)
```

Figura 14 Hiperparámetros ajustados del algoritmo XGBoost

Lo interesante de esta apertura, es el incremento en la precisión del modelo, tanto para el segmento Emergente, como para el resto de los segmentos. A continuación, en la tabla 8, se detallan las mediciones comparativas de estos dos nuevos modelos diferenciando segmentos, versus el último modelo trabajado con todos los segmentos de la empresa que previamente revisamos en la tabla 7.

Medición	Hiper-parámetros	Emergente	Resto Segmentos
Accuracy (train)	90,6%	99,0%	95,6%
Accuracy (test)	74,0%	84,8%	76,0%
RMSE	50,9%	39,0%	48,9%
AUC	0,81	0,89	0,84

Tabla 8 Precisión y error del modelo

El modelo para el segmento Emergente pasa de un 0,81 a un 0,89 en la curva AUC, mientras que el modelo para el resto de los segmentos, pasa a un 0,84. La curva ROC-AUC nos está indicando que los modelos "Emergente" y "Resto Segmentos" permiten discriminar con probabilidades del 89% y 84% respectivamente, entre los clientes que tomarán un crédito de consumo y los que no lo tomarán.

Importante destacar que para el modelo "emergente" y el modelo "resto de segmentos", hubo variaciones en la selección de las variables de mayor importancia para cada modelo, con respecto al original revisado en etapas previas.

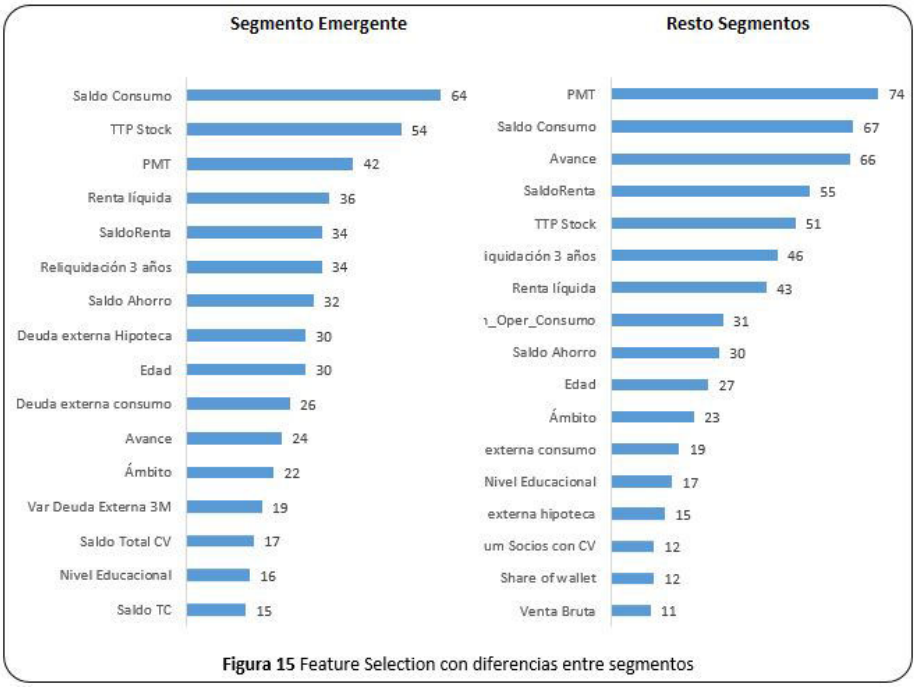


Figura 15 Feature Selection con diferencias entre segmentos

Como segunda apertura del dataset original, se trabajan los modelos para los convenios del sector público y del sector privado, también comparando a ambos con el modelo inicial que considera la totalidad del dataset, pero con la selección de variables más relevantes y el ajuste de sus hiperparámetros. Con esto, revisamos si hay un upgrade de rendimiento al hacer la discriminación entre empresas públicas y privadas. Los resultados que se obtienen y se muestran en la tabla 9, son similares a la apertura que discriminó el segmento emergente del resto de los segmentos. El modelo que considera solamente el universo de convenios del sector público, registra un upgrade de un 2% en su precisión, respecto del modelo con el dataset completo. En tanto, el modelo que se aplica solo a convenios del sector privado, muestra una mejora del 9% en su precisión, logrando una probabilidad en la curva AUC del 0,9.

Medición	Hiper-parámetros	Sector Público	Sector Privado
Accuracy (train)	90,6%	95,8%	99,0%
Accuracy (test)	74,0%	75,6%	85,0%
RMSE	50,9%	49,3%	39,3%
AUC	0,81	0,83	0,90

Tabla 9 Precisión y error del modelo

8 Conclusiones

Para lograr buenos resultados en la aplicación de un modelo de clasificación, es determinante conocer y analizar la información con la que se cuenta, lograr entenderla y a partir de esa experiencia, poder tomar decisiones que permitan aplicar adecuadamente dicho modelo. Contar con personas o equipos de diversa experiencia se torna fundamental al momento de enriquecer la información inicial que se tiene, hecho que puede incidir muy positivamente en mejorar aún más los resultados de un modelo. En el caso particular de este proyecto aplicado en una compañía, fue de vital importancia el apoyo y orientación de personas que tenían conocimientos relevantes con respecto a algunas variables que se debían considerar. Tomando esto, el modelo mejoró sus resultados.

La evidencia del análisis exploratorio del dataset, permitió definir la generación de variantes de modelos de clasificación que mejoran los resultados con respecto al modelo inicial que consideraba la totalidad del dataset.

El algoritmo XGBoost presenta múltiples y sensibles hiperparámetros, destacando tres de ellos, que permiten incrementar la precisión del modelo con cada iteración. Esto es un punto determinante y donde toman gran relevancia tareas como la selección de variables de mayor importancia o la adición de nuevas variables. Los incrementales de rendimiento del modelo fueron mayores con los ajustes de hiperparámetros que con otros ajustes.

El desarrollo de este modelo puede seguir mejorando en la medida que se sigan estudiando nuevas incorporaciones de variables relevantes que impacten directo en la propensión a la toma de créditos de consumo y/o se continúe con la apertura del modelo en base a la identificación de diversas variables que se detecten. La interacción con personas y equipos tanto de áreas comerciales, como de riesgos, pueden decantar en diversos modelos que se ajusten a todos los tipos de cliente que tiene la compañía.

Referencias

1.- Fuentes.
1.1.- Revista Gerencia. La información como habilitadora del negocio
https://bit.ly/35Ge60d
1.2.- Reportaje BBVA
https://bbva.info/2VZ9YEN

2.- Comisión para el Mercado Financiero
https://www.cmfchile.cl

3.- Comparación de algoritmos de clasificación para el incumplimiento crediticio. Aplicación al sistema bancario chileno
https://www.sbif.cl/sbifweb/servlet/ConozcaSBIF?indice=C.D.AidContenido=17772

4.- An Empirical Comparison of Machine-Learning. Methods on Bank Client Credit Assessments
https://www.mdpi.com/2071-1050/11/3/699

5.- Credit Scoring Using Machine Learning by Combing Social Network Information: Evidence from Peer-to-Peer Lending
https://doi.org/10.3390/info10120397

6.- Propensity to Churn in Banking
https://bit.ly/2Ygk5XI

7.- Predicting customer churn using targeted proactive retention
https://bit.ly/3kWDqaf

8.- https://www.quora.com/What-is-the-difference-between-the-R-gbm-gradient-boosting-machine-and-xgboost-extreme-gradient-boosting/answer/Tianqi-Chen-1

9.- https://medium.com/analytics-vidhya/math-behind-gbm-and-xgboost-d00e8536b7de

10.- https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

Anexo 1 - detalle variables

Variable	Detalle
ID_MES	indicador del mes del stock de clientes
ENTY_ID	variable que anonimiza el rut de cliente
ID_CONVENIO	código de convenio al que pertenece cada cliente
OFICINA	nombre de la oficina a la que pertenece el cliente
SEXO	género del cliente
EDAD	edad del cliente
RENTA_LIQUIDA	sueldo líquido de cada cliente
SHARE_WALL_CONS	indica si deuda SBIF es mayor en cooperativa o en otra institución
SEGM_RENTA	segmento del cliente
RANGO_EDAD	tramo de edad para clasificar a cliente
AMBITO	indica si el empleador del cliente es del sector público o privado
SOCIO_CON_PRODUCT	indica si el cliente tiene productos de crédito
SOCIOS	indica si cliente es socio de la cooperativa
SALDO_CONSUMO	deuda de crédito de consumo del cliente
NUM_SOCIOS_CONSUMO	sirve para contar los clientes que tienen deuda de consumo
TTP_STOCK	tasa promedio de la cartera de consumo
MTO_VTA_NETA	monto neto del préstamo, sin considerar el monto reliquidado
MTO_VTA_BRUTA	suma del monto neto y el monto reliquidado
NUM_OPER_VTA_TOT	indica si cliente tiene al menos un crédito de consumo
TTP_VTA_TOT	tasa de venta de los créditos de consumo
NUM_PREPAGOS_TOT	indicador de la cantidad de prepagos que ha tenido un cliente
MTO_PREPAGOS_TOT	monto total de prepagos hecho por cliente
NUM_CASTIGOS_TOT	indicador de castigo (morosidad de 180 días) del cliente
N_SOCIOS_HIPO	indicador de que el cliente tiene o no un crédito hipotecario
SALDO_HIPO	deuda hipotecaria que el cliente tiene en la cooperativa
POSEE_TC	indicador de que el cliente posee o no posee tarjeta de crédito
NUM_TC_ACT	indicador de número de tarjetas de crédito activas del cliente
SALDO_TC	saldo total de tarjetas utilizadas por cliente
N_SOCIOS_AHORRO	indicador de que cliente tiene o no tiene ahorro en la cooperativa
SALDO_AHORRO	saldo total ahorrado por cliente en la cooperativa
N_SOCIO_DAP	indicador de que cliente tiene o no depósitos a plazo
SALDO_DAP	saldo total en depósitos a plazo de cliente en la cooperativa
N_SOCIOS_CV	indicador de que cliente tiene o no tiene cuenta vista en la cooperativa
SALDO_TOTA_CV	saldo total en la cuenta vista del cliente
NUM_CV_CON_USO	indicador de que la cuenta vista del cliente tiene uso
NUM_TC_USO	indicador de que la tarjeta de credito del cliente tiene uso
NIVEL_EDUCACIONAL	nivel educacional del cliente, desde básica a universitaria
DEUDA_EXT_CONSUMO	deuda del cliente en créditos de consumo en todo el sistema financiero
DEUDA_EXT_HIPOTECA	deuda del cliente en créditos hipotecario en todo el sistema financiero
VAR_DEUDA_EXT_3M	variación de la deuda en sistema financiero tres meses previos
VAR_DEUDA_EXT_6M	variación de la deuda en sistema financiero seis meses previos
VAR_DEUDA_EXT_12M	variación de la deuda en sistema financiero doce meses previos
SaldoRenta	Cociente entre la deuda de consumo del cliente y su renta líquida
Reliq3A	cantidad de reliquidaciones que tuvo un cliente durante tres años
Ind_Venta+3M	variable objetivo, indica si cliente toma o no toma un crédito de consumo