PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

ELABORACIÓN DE UN MODELO DE PROPENSIÓN AL CRÉDITO DE CONSUMO CAPAZ DE DISCRIMINAR A NIVEL DE INDIVIDUO UTILIZANDO SOLAMENTE INFORMACIÓN FINANCIERA DE CARÁCTER PÚBLICA

ALAN GROSS SÁNCHEZ

Tesis para optar al grado de
Magíster en Ciencias de la Ingeniería

Profesor Supervisor:
ALEJANDRO MAC CAWLEY

Santiago de Chile, (abril, 2018)

© 2018, Alan Gross Sánchez

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

ELABORACIÓN DE UN MODELO DE PROPENSIÓN AL CRÉDITO DE CONSUMO CAPAZ DE DISCRIMINAR A NIVEL DE INDIVIDUO UTILIZANDO SOLAMENTE INFORMACIÓN FINANCIERA DE CARÁCTER PÚBLICA

ALAN GROSS SANCHEZ

Tesis presentada a la Comisión integrada por los profesores:

ALEJANDRO MAC CAWLEY

TOMÁS REYES

MAURICIO VARAS

GONZALO YÁÑEZ

Para completar las exigencias del grado de
Magíster en Ciencias de la Ingeniería

Santiago de Chile, (abril, 2018)

A mi novia, familia y equipo de trabajo que me apoyaron durante todo el trayecto.

AGRADECIMIENTOS

Quisiera agradecer a todo el equipo de Inteligencia de Negocios por todas las facilidades brindadas para poder llevar a cabo este proyecto, a mi novia y a mi familia por el constante e incondicional apoyo en los momentos más difíciles y especialmente a Gustavo Del Real y Alejandro Mac Cawley por la enorme paciencia y el continuo *feedback* a lo largo del proceso.

ÍNDICE GENERAL

Pág.

DEDICATORIA ii

AGRADECIMIENTOS iii

ÍNDICE DE TABLAS vi

ÍNDICE DE FIGURAS vii

RESUMEN viii

ABSTRACT ix

1. Introducción 1

2. Revisión Bibliográfica 77

3. Modelo Conceptual 1313
   - 3.1 Contexto 13
   - 3.2 Interacción entre variables 15
   - 3.3 Componente temporal 16

4. Metodología 19
   - 4.1 Limpieza de datos 19
   - 4.2 Elaboración y optimización de la variable de respuesta 20
   - 4.3 Generación de características 223
   - 4.4 Segmentación 26
   - 4.5 Selección de características 27
   - 4.6 Ajustes preliminares para la selección del modelo predictivo 29
   - 4.7 Selección del modelo predictivo 31
   - 4.8 Flujograma 33

5. Resultados 334
   - 5.1 Limpieza de datos 334
   - 5.2 Elaboración y optimización de la variable de respuesta 35

5.3 Generación de características 40
5.4 Segmentación 41
5.5 Selección de características 41
5.6 Selección del modelo predictivo 42
5.7 Discusión de resultados 46

6. Validación del Modelo 49

7. Conclusiones 51

BIBLIOGRAFIA 53

A N E X O S 556

Anexo A: Resumen descriptivo de cada campo de la base de datos de la SBIF 57

Anexo B: Variables creadas 59

Anexo C: Parámetros de inicialización de modelos. 65

ÍNDICE DE TABLAS

Pág.

Tabla 5.1.1: Volumen de datos tras aplicar filtros 35
Tabla 5.2.1: Resultados de la optimización del parámetro α 38
Tabla 5.6.1: Valores de precisión para el primer decil 42
Tabla 5.6.2: Tiempo de ejecución de los modelos 42
Tabla 5.6.3: Cantidad óptima de variables.....44
Tabla 5.6.4: Nivel de confianza (KS) para ambos modelos en ambos segmentos……..44
Tabla 5.6.5: Porcentaje de concordancia entre modelos decil a decil…….45
Tabla 5.6.6: Análisis LIFT por segmento……46
Tabla 6.1.1: Respuestas del experimento de validación…….....49
Tabla 6.1.2: Ventas del experimento de validación…….50

ÍNDICE DE FIGURAS

Pág.

Figura 1.1.1: Monto adeudado por tipo de cartera 2
Figura 1.1.2: Número de deudores por tipo de cartera.....3
Figura 1.1.3: Variación del número de clientes y monto adeudado por tipo de cartera para los últimos tres años 4
Figura 4.3.1: Cálculo de la edad a partir del RUT 25
Figura 4.3.2: Información disponible para predecir.....26
Figura 4.8.1: Flujograma desde la data cruda hasta la segmentación 33
Figura 4.8.2: Flujograma desde la selección de variables hasta la selección del modelo…33
Figura 5.2.1: Distribución de montos de créditos de consumo…….. .36
Figura 5.2.2: Nivel de deuda y diferencia con el mes anterior para un sujeto en particular……...37
Figura 5.2.3: Curva de porcentaje de falsos para distintos valores del parámetro….....39
Figura 5.2.4: Explicación gráfica del cálculo del Jump……...40

RESUMEN

El objetivo general del trabajo, fue desarrollar y validar un modelo conceptual capaz de detectar las necesidades crediticias futuras de las personas del sistema financiero y así cuantificar la propensión a créditos de consumo, utilizando únicamente información financiera de carácter pública. Para llevar a cabo el estudio, se procedió a generar la variable dependiente como una combinación de las ventas de créditos de consumo del banco con las ventas presuntas en otros bancos. Para esto fue necesario generar un modelo capaz de determinar cuándo una persona tomó un crédito de consumo en otra institución. Posteriormente, se ejecutó el proceso de *feature generation* y *feature selection*, los cuales consistieron en generar variables de todo tipo a partir de la información proveniente de la SBIF, para luego seleccionar sólo aquellas que fuesen relevantes para el modelo. Finalmente, sobre un set que se dividió en 70% *training* y 30% *testing*, se hizo competir los siguientes modelos de aprendizaje automático: regresión logística, SVM, bosques aleatorios, redes neuronales y modelos de *boosting*. Los mejores resultados se obtuvieron con XGBoost, obteniendo un LIFT de 19,3% para el primer decil, en 10 minutos y con un KS de 0,39 en el caso de NumJump = 1 y 21% en el primer decil de LIFT en 15 minutos con un KS de 0,25 para el caso de NumJump > 1. Al poner en producción el modelo, se obtuvo una efectividad del 2,86% en ventas sobre respuestas, una cifra más de cuatro veces mayor a lo acostumbrado por el banco en campañas sobre prospectos (0,69%).

Palabras Claves: *score* de propensión, crédito de consumo, información pública, *big data*, *machine learning*, XGBoost.

ABSTRACT

The main objective of this work was to develop and validate a conceptual model capable of detecting the future credit needs of people in the financial system and thus quantify the personal loans propensity, using only public financial information. To carry out the study, the dependent variable was generated as a combination of the bank's sales of personal loans added to presumed sales at other banks. For this, it was necessary to generate a model capable of determining when a person took a loan in another institution. Subsequently, a process of feature generation and feature selection was executed. It consisted in generate variables of all kinds from the data coming from the SBIF, and then selecting only those that were relevant for the model. Finally, on a set that was divided into 70% for training and 30% for testing, the following models of machine learning were compared: logistic regression, SVM, random forests, neural networks and boosting models. The best results were obtained with XGBoost, with a LIFT of 19.3% for the first decile, in 10 minutes and with a KS of 0.39 in the case of NumJump = 1, and 21% in the first decile of LIFT in 15 minutes with a KS of 0.25 for the case of NumJump> 1. When putting the model into production, an effectiveness of 2.86% was obtained on sales over answers, a number more than four times higher than usual in campaigns on prospects (0.69%).

Keywords: propensity score, personal loans, public information, big data, machine learning, XGBoost.

1. INTRODUCCIÓN

Empresas de todo el mundo han incorporado modelos de propensión para determinar si un cliente es proclive a comportarse de la manera que la empresa desea a partir de un estímulo específico. Las compañías pertenecientes al rubro financiero no se han quedado atrás y han adoptado modelos de propensión para estudiar las características y comportamientos colectivos de sus clientes y no clientes y así definir cuáles de ellos serán más propensos a aceptar o rechazar una oferta.

El sistema financiero chileno está compuesto por dos grandes grupos: emisores bancarios y emisores no bancarios. El primero es una conglomeración compuesta por 20 bancos, donde 14 de ellos caen en la categoría de "Bancos Establecidos en Chile", cinco se definen como "Sucursales de Bancos Extranjeros" y uno corresponde a un banco estatal. Por otra parte están los emisores no bancarios, donde se encuentran las cooperativas, casas comerciales, cajas de compensación, entre otros. Todas estas entidades son reguladas por la Superintendencia de Bancos e Instituciones Financieras (Superintendencia de Bancos e Instituciones Financieras [SBIF], 2017).

Dentro del mercado de los emisores bancarios la deuda crediticia se divide en tres grupos: deuda comercial, deuda hipotecaria y deuda de consumo. La deuda comercial es aquella que contraen las entidades con giro comercial con fines de inversión. Actualmente en el sistema financiero hay 1.206.305 deudores dentro de esta categoría, quienes adeudan 80,1 billones de pesos. La cartera hipotecaria se compone de aquellas entidades que contraen deuda con el fin de adquirir un bien raíz, para este caso, el sistema financiero cuenta con 1.044.887 deudores por un monto total de 41,8 billones de pesos. Finalmente, el mercado de la deuda de consumo es aquel en donde las entidades contraen deudas con un fin distinto a los casos anteriormente descritos. Esta cartera se compone por 3.519.079 deudores, quienes suman 17,9 billones de pesos. Tras consolidar las cifras para todos los grupos mencionados, el sistema financiero chileno completo se

compone de un total de 5.770.271 entidades financieramente activas quienes adeudan 139,9 billones de pesos.

En las siguientes figuras se aprecia de manera gráfica tanto el monto adeudado (en billones de pesos), como el número de deudores por tipo de cartera:

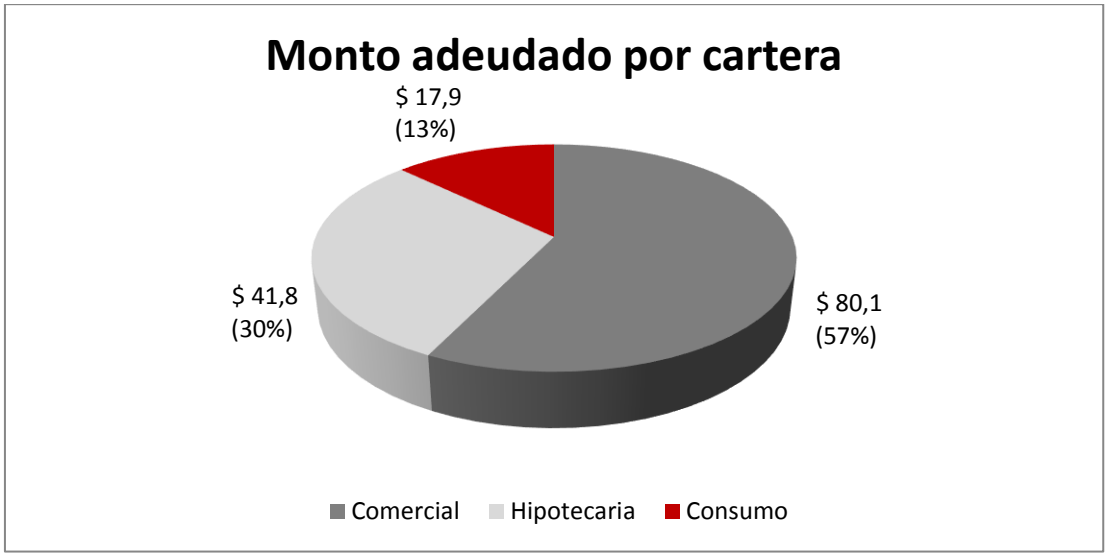

Figura 1.1.1: Monto adeudado por tipo de cartera

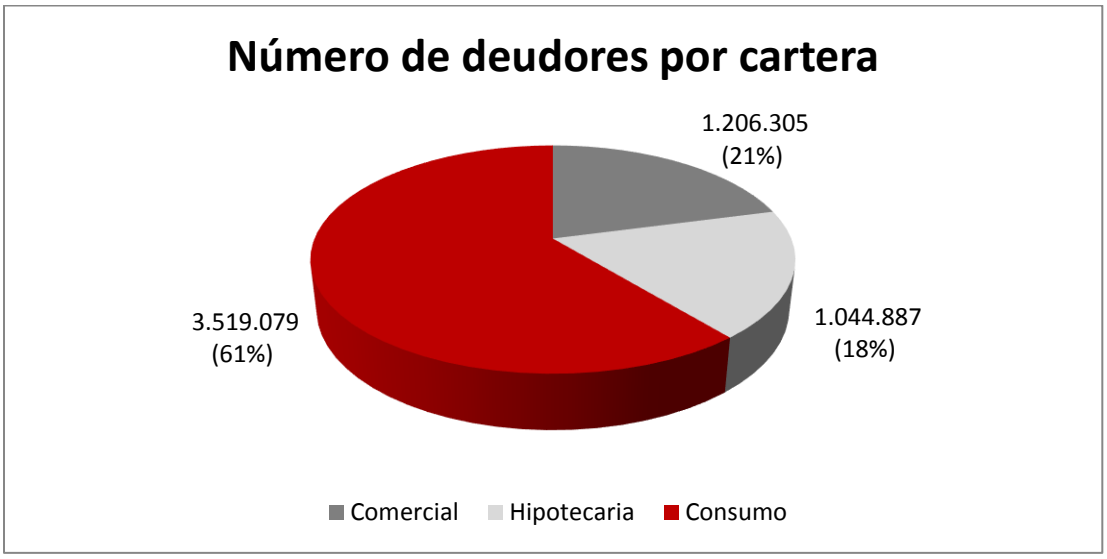

Figura 1.1.2: Número de deudores por tipo de cartera

En los últimos tres años el sistema financiero ha experimentado un crecimiento tanto en el número de clientes como en el monto adeudado. La cartera de los emisores bancarios ha aumentado un 4,4% en cuanto a número de clientes y un 26,1% en cuanto al monto adeudado. Para el caso de la deuda comercial, este porcentaje de crecimiento es parejo, ya que en el mismo periodo experimentó un crecimiento del 19,4% en el número de clientes y de 19,3% en el monto. En el caso de la cartera hipotecaria el porcentaje de clientes sólo aumentó un 7,3%, mientras que el monto adeudado sufrió un cambio mayor, aumentando un 43,1%. Finalmente para el caso de la cartera de consumo, se da un comportamiento atípico en comparación con las demás, ya que el número de clientes disminuyó un 0,7%, mientras que el monto adeudado aumentó 22,4%. Esto se explica según Camus (2015), porque la estrategia de los bancos ha cambiado y se está enfocando en los segmentos más altos de la población. En la siguiente figura se presenta el crecimiento previamente mencionado de manera gráfica:

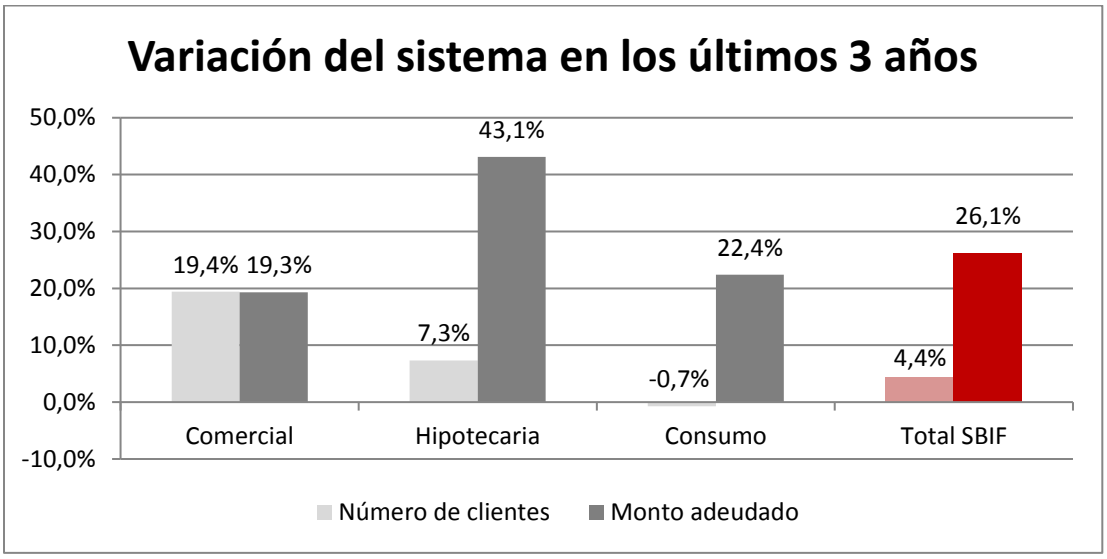

Figura 1.1.3: Variación del número de clientes y monto adeudado por tipo de cartera para los últimos tres años

El objetivo del presente trabajo es el de desarrollar y validar un modelo conceptual capaz de detectar las necesidades crediticias futuras de los clientes del sistema financiero y así cuantificar la propensión a créditos de consumo. Con esto se busca mejorar la efectividad de las actuales campañas del banco con el que se trabajará, mediante una correcta asignación de clientes a las distintas fuerzas de ventas, para que se contacte sólo a aquellas personas que sean más propensas a adquirir un crédito de consumo en los próximos tres meses. Esta efectividad se puede medir de dos formas: manteniendo fijo el número de personas contactadas y comparar el número de llamadas exitosas del modelo actual versus el propuesto en este trabajo, o bien, manteniendo fijo el número de llamados exitosos y comparar el número de personas que debieron ser contactadas para lograr el objetivo. En este caso se utilizará la primera medida, es decir, contactar al mismo número de personas y esperar un aumento en el número de llamadas exitosas a modo de aumentar las ventas de la empresa.

Siguiendo en la línea del objetivo descrito en el párrafo anterior, las hipótesis que se buscan comprobar en este trabajo son, en primer lugar, si es posible determinar la propensión crediticia de una persona —a nivel de individuo— utilizando sólo información financiera de carácter pública, y en segundo lugar, que el uso de esta metodología arrojará mejores resultados que el modelo actual, el cual usa tanto información pública como información propia del banco, pero trabaja sobre agrupaciones de clientes construidas en base a su renta estimada.

Para lograr este objetivo se desarrollarán modelos matemáticos y estadísticos que en conjunto con herramientas de *Big Data* e Inteligencia de Negocios permitirán generar un conjunto de características que describan en su totalidad el comportamiento financiero de la muestra de estudio a través del tiempo, seleccionar cuáles de estas características son las más relevantes al momento de perfilar a los clientes de la muestra, y finalmente elaborar un modelo basado en estas características que permita predecir la probabilidad de compra de un crédito de consumo en los próximos tres meses.

La contribución de este trabajo radica en tres puntos principales. Primero, se analizarán y establecerán las variables que más afectan a la propensión crediticia, en base a información de carácter pública proveniente de la SBIF; segundo, que en vez de estudiar a los clientes como un colectivo, se estudiará a cada persona como una entidad individual, con un comportamiento financiero único que lo distingue de las demás personas; y tercero, en la creación de un modelo de propensión a nivel de individuo que utilice las variables antes mencionadas para generar un puntaje que permita identificar a aquellas personas que tengan propensión a necesitar un crédito de consumo.

Respecto al alcance del proyecto, para este estudio sólo se considera la propensión al producto crédito de consumo. La información de otros tipos de créditos, como comerciales o hipotecarios, será utilizada, pero siempre en función de mejorar la calibración del modelo y nunca intentando predecir la propensión a esos productos. El

monto a ofertar tampoco es considerado como parte de este proyecto, dejando explícitamente abierta la inquietud de profundizar en este tema. Finalmente, el alcance de este trabajo abarca únicamente a personas naturales, dejando fuera micro, pequeñas, medianas y grandes empresas que, si bien se encuentran presentes en la base de datos de la SBIF, no están dentro del foco del proyecto.

Este trabajo se estructura en cinco secciones. En la primera sección se recorrerá la literatura para establecer el estado del arte en el que se encuentra la industria bancaria respecto a modelos de propensión y selección de clientes. En la segunda sección se establecerá el modelo conceptual sobre el cual se basa la investigación, dejando claro el contexto, la información disponible, la interacción entre las variables y la importancia de la componente temporal sobre estas mismas. En la tercera sección se expondrá la metodología utilizada, comenzando por la limpieza de la data, pasando por la creación y selección de variables para terminar con la elección del modelo. En la cuarta sección se expondrán los resultados obtenidos tras aplicar la metodología previamente descrita. Finalmente, en la quinta sección se discutirán y analizarán los resultados obtenidos para formular las principales conclusiones.

2. REVISIÓN BIBLIOGRÁFICA

Numerosas son las publicaciones relacionadas al rubro bancario e instituciones financieras donde se ha planteado la utilización de *scores* —calculados utilizando distintas herramientas de la rama de inteligencia de negocios y *Big Data*— para generar sistemas de apoyo a la toma de decisiones, al momento de enfrentarse a problemas de *pricing* (Phillips, 2013), propensión a tomar una oferta, predicción de no pago o *default*, entre otros (Tsai y Wu, 2008). Sin embargo, la mayoría de los autores se centran en resolver la problemática matemática o algorítmica inherente al problema, más que en la problemática de negocio que existe detrás de predecir el comportamiento crediticio de las personas o de calcular puntajes de propensión al crédito (Moro, Cortez y Rita, 2014). Hand y Henley (1997) plantean que esto tiene total sentido, debido a que los acuerdos de confidencialidad que existen entre empresas e investigadores indican que las compañías que facilitan los datos están dispuestas a divulgar información de carácter científica, pero no comercial.

Sin perjuicio de lo anterior, las metodologías utilizadas por los distintos autores comparten la cualidad de ser altamente replicables, ya que se sustentan en modelos estructurados de minería de datos o aprendizaje automático, logrando llegar a establecer complejos modelos lógicos matemáticos que terminan en resultados exitosos en situaciones análogas a la que se quiere tratar.

En el ámbito de la salud, Murdoch y Detsky (2013) sugieren que aprovechar la recolección de datos de pacientes y profesionales podría ser una forma importante de mejorar la calidad y la eficiencia de la prestación de servicios de salud. Sin embargo, gran parte de este conjunto de datos actualmente se percibe como un subproducto de la prestación de atención de salud, en lugar de un activo central para mejorar su eficiencia. Los autores postulan que el uso de los datos amplía considerablemente la capacidad de generar y difundir nuevos conocimientos, al mismo tiempo que permiten transformar la

atención en el área de la salud mediante la entrega de información directa a los pacientes y ofreciendo prácticas médicas personalizadas, haciendo que los pacientes desempeñen un rol más activo en su proceso de atención médica. Este estudio se alinea perfectamente y sirve de introducción al concepto que se quiere tratar en este trabajo, ya que refiere a la visualización de los datos como un activo estratégico de las empresas, que permite mejorar la eficiencia de sus procesos. En este caso lo que se busca es mejorar la precisión de las campañas comerciales del banco, cambiando las actuales listas de clientes a contactar por unas que incluyan sólo a aquellos clientes con mayor índice de propensión y así aprovechar de mejor manera los esfuerzos de los ejecutivos para contactar clientes.

Cabe destacar que el problema de seleccionar los *n* mejores clientes de una lista, o en este caso, seleccionar aquellos clientes que sean más propensos a adquirir un producto dentro de una base de personas, es considerado un problema NP-Duro (Talla Nobibon, Leus & Spieksma, 2011). No obstante, dentro del mundo de la inteligencia de negocios, Lau, Chow y Liu (2004) describen el potencial de las técnicas de la minería de datos basada en características personales de los consumidores para seleccionar clientes y asignarlos en campañas de marketing de un banco en Hong Kong. Lo rescatable de este estudio es el uso, por parte de los autores, de características propias de las personas en conjunto con información bancaria para crear la base de variables discriminatorias, sin embargo, la deficiencia recae en que no se validó ningún modelo impulsado por los datos.

En este mismo ámbito, Moro, Laureano y Cortez (2012) exploraron modelos impulsados por datos para modelar el éxito del *telemarketing* bancario. Sin embargo, sólo se lograron buenos modelos al utilizar atributos que se conocen durante la ejecución de la llamada, como por ejemplo la duración de esta misma. Este estudio no deja de ser relevante; a pesar de que no proporciona información para predecir el futuro, sí arroja importantes *insights* que pueden utilizarse para explicar el pasado. Si bien no está dentro

del alcance de este estudio explicar el pasado o crear un manual de mejores prácticas en función de esto mismo, sí se rescata el modelo conceptual utilizado, que busca crear modelos impulsados por datos para ser aplicados dentro de las diferentes instituciones.

En lo que respecta al uso de herramientas y aplicaciones de *Big Data*, Martens y Provost (2011) identificaron un grupo objetivo de clientes para un banco belga, a partir de una pseudo red social compuesta por las transacciones entre clientes, y probaron que para dos ofertas de distintos productos la tasa de adopción de estos difiere dependiendo de la red en la que se encuentren, demostrando que las probabilidades aumentan al tratarse de un producto que se encuentre próximo en la red. Este estudio es interesante, dado que considera como variable principal la relación y cercanía entre entidades financieras. No obstante, a pesar de encontrarse en el estado del arte, este tipo de relaciones no se consideran en el alcance de este trabajo, ya que para esto requiere incluir información transaccional del banco y la idea de este trabajo es generar un modelo que se alimente solamente de información pública.

En lo que respecta a la unión entre considerar el foco en el negocio que propone la inteligencia de negocios y las herramientas analíticas que provee el área de la *Big Data*, nacen estudios como el de Hossein Javaheri (2008), en donde se trata una problemática bastante más cercana a la propuesta en este trabajo. Aquí se analizó cómo una campaña de marketing en medios masivos de comunicación podría afectar la compra de un nuevo producto bancario. Los datos fueron recogidos de un banco en Irán con un total de 22.427 clientes durante el período de seis meses en que se realizó la campaña. Se asumió que todos los clientes que compraron el producto (7%) fueron influenciados por la campaña de marketing y se realizó un estudio para determinar qué características los hacían propensos a aceptar la oferta presentada. Los datos históricos permitieron la extracción de un total de 85 características relacionadas con la recencia, frecuencia y monto de la operación (RFM), por una parte, y edad del cliente, por otra. Se usaron 26 de estos atributos para alimentar una tarea de clasificación binaria modelada mediante

un algoritmo de SVM utilizando 2/3 de los clientes seleccionados al azar para el entrenamiento y 1/3 para las pruebas. La precisión de clasificación alcanzada fue de 81% a través de un análisis LIFT, es decir, el modelo clasificó correctamente al 81% de los casos ubicados dentro del decil de mayor puntaje o probabilidad. Sobre el análisis LIFT se hablará en mayor detalle en la Metodología (sección 4.7). Este modelo podría seleccionar el 79% de los respondedores positivos con sólo el 40% de los clientes. Este estudio brinda un sólido marco conceptual sobre el cual establecer bases, ya que considera la generación y selección de variables a partir de atributos de la población, indica claramente que las variables predominantes son aquellas pertenecientes al conjunto RFM (recencia, frecuencia y monto) y la edad, genera una división de la muestra para distinguir datos de entrenamiento y de prueba y finalmente fija un mecanismo para responder al problema de cuántas personas contactar para lograr la precisión deseada. Lo que no se considera en este estudio por la naturaleza del problema que intenta resolver, es la estacionalidad. Al tratarse de una campaña aislada, el autor no tiene que lidiar con el componente temporal y por eso basta con tomar seis meses de profundidad histórica, a diferencia de esta investigación, donde la variable tiempo debe ser considerada para poder asignar un *score* de propensión a cada cliente.

Por último, otro ejemplo donde se ataca un problema similar al que se busca estudiar en este trabajo, es el que exponen Moro, Cortez y Rita (2014), donde se propone un enfoque de minería de datos para predecir el éxito de las llamadas de *telemarketing* en la venta de depósitos bancarios a largo plazo. Se abordó un banco minorista portugués con datos recogidos de 2008 a 2013, incluyendo así los efectos de la reciente crisis financiera. Se analizó, mediante herramientas de *Big Data*, un conjunto de 150 características relacionadas con clientes bancarios, productos y atributos socioeconómicos. Se exploró una selección de características semiautomáticas en la fase de modelado que permitió seleccionar un conjunto reducido de 22 características. También se comparó cuatro modelos de minería de datos: regresión logística, árboles de decisión, redes neuronales y máquinas de soporte vectorial. Utilizando dos métricas;

área bajo la curva ROC (AUC) y área bajo la curva acumulativa LIFT (ALIFT), los cuatro modelos fueron probados en un conjunto de evaluación. Las redes neuronales presentaron los mejores resultados (AUC = 0.8 y ALIFT = 0.7) permitiendo llegar al 79% de los suscriptores, seleccionando al 50% de los clientes mejor clasificados. Además, se aplicaron dos métodos de extracción de conocimiento, un análisis de sensibilidad y un árbol de decisión al modelo de redes neuronales y se revelaron varios atributos clave. Tal extracción de conocimiento confirmó que el modelo obtenido era creíble y valioso para los gerentes de campañas de *telemarketing*. Los aspectos más importantes de este trabajo y donde radica el valor a extraer, son el uso de una profundidad histórica de datos (que permite aislar los efectos de la estacionalidad), la utilización de algoritmos semiautomáticos para la selección de variables, la comparación de distintos modelos mediante métricas estándar y el uso de métodos de extracción de conocimiento para abrir y darles un sentido humano a estas cajas negras que representan los modelos basados en redes neuronales. La deficiencia que posee el estudio mencionado y que se pretende incorporar en este trabajo, es que no se considera a las personas como entes individuales, sino que como una masa de clientes sobre la cual se aplican los modelos.

Para este trabajo se pretende utilizar un modelo conceptual que se base en los modelos descritos en los últimos dos párrafos, añadiendo ciertas modificaciones que permitan hacer que el modelo funcione a nivel de individuo en vez de a nivel de población y que considere la estacionalidad de la demanda. En síntesis, se tomará un enfoque de minería de datos para generar características relacionadas a la recencia, frecuencia y monto de las operaciones, junto con la edad de los clientes, como en el estudio de Hossein Javaheri (2008), pero esta vez a nivel de individuo y utilizando la data de los últimos cinco años. Se seleccionarán algunas de estas características de manera semiautomática y se harán competir distintos modelos predictivos para ver cuál de ellos es el que entrega mejores resultados como proponen Moro, Cortez y Rita (2014). La forma de comparar entre los distintos modelos será mediante el análisis del gráfico LIFT que cada uno de

ellos genere. Finalmente, el gráfico LIFT del modelo ganador permitirá establecer el punto el corte a través del score que discriminará entre una persona propensa y una que no lo es.

3. MODELO CONCEPTUAL

3.1 Contexto

Para llevar a cabo el estudio se utilizará la información presente en la base de datos que proporciona mensualmente la Superintendencia de Bancos e Instituciones Financieras (SBIF), institución encargada de "supervisar las empresas bancarias así como de otras entidades, en resguardo de los depositantes u otros acreedores y del interés público y su misión es velar por el buen funcionamiento del sistema financiero" (Parrado, 2017).

Una vez al mes, la SBIF recopila en una base de datos única, de manera agregada, la información financiera perteneciente a cada uno de los bancos de la plaza, se estructura de manera ordenada, se estandariza y se envía de vuelta a las instituciones financieras con el fin de que todas cuenten con el mismo nivel de información al momento de evaluar la situación crediticia de sus potenciales clientes.

Como es de suponer, recopilar la información de todos los bancos, depurarla, estructurarla y estandarizarla, toma tiempo. De hecho, el proceso completo desde que la Superintendencia solicita la información a los bancos e instituciones financieras, hasta que entrega de vuelta la base consolidada, toma aproximadamente cuarenta días. Sobre este desfase se hablará en la sección 4.3, ya que agrega una complicación extra a la predicción.

La base de datos de la SBIF posee 41 campos de información financiera para cada entidad que haya emitido una deuda en algún momento. Esta información se encuentra agregada por mes y por RUT conteniendo la información de todos los bancos de Chile. Dentro de los campos previamente mencionados, hay campos que

sólo tienen valores para empresas, campos que sólo tienen valores para personas naturales, campos que son agregaciones de otros, entre otros.

En particular, para el caso de las personas naturales son pocos los campos que realmente entregan información relevante sobre el estado financiero de las personas. Entre ellos están los siguientes:

- Deuda directa vigente: corresponde a la suma de todas las deudas directas de una persona en un mes determinado.
- Deuda indirecta vigente: corresponde a la suma de todas las deudas indirectas de una persona en un mes determinado.
- Deuda crédito consumo: deuda directa que corresponde a la suma de todas las deudas de una persona correspondientes a créditos de consumo, tarjetas de crédito, líneas de crédito, líneas de sobregiro y productos similares, en todos los bancos en un mes determinado.
- Instituciones con deuda: corresponde a la cantidad de instituciones en las cual el deudor posee deudas de consumo en un mes específico.
- Cupo disponible: corresponde a la suma de los montos no utilizados en las distintas tarjetas, líneas de crédito o sobregiro que una persona pueda tener en sus distintos bancos en un mes específico.
- Deuda hipotecaria: deuda directa que corresponde a toda la deuda hipotecaria de una persona, ya sea para viviendas o con fines generales, en todos los bancos para un mes específico.
- Deuda comercial: deuda directa que corresponde a toda la deuda comercial que una persona pueda poseer. Esta considera préstamos para negocios comerciales, personas con giro, créditos estudiantiles y productos similares donde el propósito del préstamo sea considerado como una inversión.
- Instituciones con deuda comercial: refleja la cantidad de instituciones en donde el deudor registra préstamos comerciales.

Las variables antes mencionadas se pueden agrupar de la siguiente manera: campos relacionados a deuda de consumo, deuda hipotecaria, deuda comercial y deuda indirecta. Estas son relevantes, porque permiten aproximarse al comportamiento de las personas en el sistema financiero (Moro, Cortez & Rita 2014) y porque hacen posible extraer información como la frecuencia con la que una persona se endeuda, los montos por los que decide endeudarse, como distribuye su deuda dentro de las distintas instituciones financieras e incluso predecir de cierta manera el ciclo de vida de la persona (Lau, Chow & Liu, 2004). En definitiva, con estas cuatro agrupaciones de información, se puede perfilar al cliente financieramente hablando.

3.2 Interacción entre variables

Estas distintas variables interactúan entre sí de diversas maneras. Por ejemplo, manteniendo fija la cantidad de instituciones con deuda, y asumiendo que la persona no ha adquirido nuevos créditos de consumo, la deuda de consumo está inversamente relacionada con el cupo disponible, es decir, una es función de la otra, a medida que aumenta la deuda, disminuye el cupo disponible y vice versa. Este tipo de movimientos se explican por el uso de las tarjetas y líneas de crédito. Ahora bien, si la cantidad de instituciones con deuda aumenta y el cupo disponible no lo hace, podría presumirse que la persona adquirió una nueva deuda en otra institución. Al revés, si la cantidad de instituciones aumenta y el cupo disponible también lo hace, se podría presumir que la persona aperturó productos en otro banco. Otro ejemplo es si la cantidad de instituciones disminuye, pero la deuda no lo hace, podría significar que la persona refinanció y consolidó sus deudas en una menor cantidad de instituciones. Si las instituciones disminuyen y la deuda también, entonces podríamos estar hablando del caso en que la persona terminó de pagar una deuda. Como se puede apreciar, sólo tomando estas tres variables y todos sus posibles valores (aumentar, mantenerse constante o disminuir) se puede

formar 27 combinaciones distintas, donde cada una de ellas posee una explicación financiera diferente.

La deuda indirecta, por su parte, corresponde a la suma de las deudas que una persona contrae de manera colateral. Un ejemplo de este caso sería el de actuar como aval en la compra de una propiedad. Este campo por sí sólo no se relaciona de manera directa con los previamente mencionados, sin embargo, puede entregar información que indique que la persona posee una sólida base financiera, dado que de otro modo no hubiese sido aceptado como aval en la transacción.

Finalmente, hay variables que no se encuentran detalladas de manera explícita, pero que no son difíciles de estimar, como por ejemplo: la edad de las personas. Esta variable es relevante, ya que en la literatura se argumenta que los niveles de deuda se hacen más difíciles de estimar en personas jóvenes por falta de historia pasada (Brown, Garino, Taylor & Price, 2005). Si a esto se le suman los componentes de la psicología de los jóvenes, como el exceso de confianza o el sentimiento de tener todo bajo control, se puede concluir que el comportamiento crediticio de este grupo etario presentará comportamientos más erráticos que en el caso de una persona de mayor edad (Norvilitis, Szablicki & Wilson, 2003). Para estimar la edad se utiliza el RUT. El RUT es un número único de carácter correlativo que se le asigna a cada chileno al momento de nacer, por lo que con un ajuste lineal o cuadrático entre este número correlativo y la edad para un grupo de personas conocidas, es posible determinar de manera aproximada la edad de cualquier persona nacida en el país. Este tema se retomará en la sección 4.3.

3.3 Componente temporal

No obstante, el valor de todas las variables explicadas y relacionadas en los párrafos anteriores, no radica en su estado actual, sino que en el desarrollo en el tiempo de las mismas (Volkov, Benoit & Van den Poel, 2017). Es por esto que

estas deben ser estudiadas individuo a individuo para poder extraer su forma de actuar y poder entender su comportamiento dentro del sistema financiero. De esta manera será posible predecir, en base a su pasado, cuál es su futuro más probable.

De aquí nace la pregunta de cómo relacionar la componente temporal al perfil financiero de las personas. Para esto, lo que muchos autores hacen a lo largo de la literatura (Jiang & Tuzhilin, 2006), es ocupar medidas como la tendencia, el nivel actual, máximos, mínimos y todo tipo de indicadores, para distintos intervalos de tiempo, ya sean seis meses, un año, dos años o incluso más. En definitiva descomponen una variable en muchas otras variables que agrupadas pueden brindar un buen indicio de cómo varió la componente temporal para poder llegar al estado actual. Lo que se intenta es entender cómo el pasado afecta al futuro, sólo a través del presente.

Los campos presentes en la base de datos de la SBIF se pueden categorizar en tres tipos distintos:

a) Campos referentes al nivel de deuda de la persona.
b) Campos referentes a la cantidad de instituciones en las que la persona posee deuda.
c) Campos que en primera instancia no aportan información, pero que a partir de los cuales se pueden derivar variables relevantes, como el caso del RUT y la edad.

Para el primer caso es necesario descomponer la variable en tantas formas como sea posible extraer información, ya sea utilizando tendencias, mínimos, máximos, medianas o valor acumulado a la fecha, por nombrar algunos ejemplos, para distintos periodos de tiempo. No obstante, este proceso puede hacer crecer mucho el número de columnas sobre las cuales trabajan los algoritmos y hacer más ineficientes los modelos de decisión, por lo que muchos autores recomiendan usar

modelos de selección de variables (Moro & Laureano, 2011) para determinar de todas las variables creadas, cuáles son las que realmente aportan valor y así eliminar aquellas que sólo entorpecen los cálculos (Guyon & Elisseeff, 2003). En el segundo caso, lo que interesa es reconocer momentos en los que haya existido un alza o una baja en el número de instituciones, por lo que se propone crear marcadores que dejen en evidencia estos movimientos para distintos periodos de tiempo. Finalmente, en el tercer caso, por tratarse de campos derivados, hay que tratar cada uno por separado dependiendo de qué es lo que simbolizan. En el caso de la edad de una persona, esta ya contiene el pasado en sí misma, por lo que no es necesario realizar ninguna descomposición y basta con el valor puntual.

Finalmente, al cruzar la interacción entre las variables de la base de datos de la SBIF, la relevancia de la componente temporal explicada en el párrafo anterior y lo avances realizados por otros autores expuestos en la revisión bibliográfica, es sugerente pensar que la función de propensión buscada que prueba la primera hipótesis ("es posible determinar la propensión crediticia de una persona, a nivel de individuo, utilizando sólo información financiera de carácter pública") debiese ser de la forma:

$$Score\ de\ propensión = f(R(v,i,t), F(v,i,t), M(v,i,t), E(i,t), \dots)$$

Donde las funciones $R(v,i,t)$, $F(v,i,t)$ y $M(v,i,t)$ corresponden a la generalización de los conceptos utilizados por Hossein Javaheri (2008) de recencia, frecuencia y características monetarias de la variable v para el individuo i, en el instante t, mientras que la función $E(i,t)$ corresponde a la edad del individuo i en el instante t. Adicionalmente, la forma de dicha función dependerá del modelo que pruebe ser el más preciso dentro de los que se harán competir, como fue en el caso de la investigación realizada por Moro, Cortez y Rita (2014).

4. METODOLOGÍA

4.1 Limpieza de datos

Al poseer la información histórica de cada entidad que ha emitido deuda en Chile con un nivel de detalle mensual, la base de datos de la SBIF posee un volumen difícil de manejar. Si se reduce la carga histórica y sólo se considera la información presente desde noviembre del año 2012 en adelante, fecha desde la cual el banco cuenta con información estructurada y de fácil acceso, ésta posee más de 430 millones de registros correspondientes a más de 10.3 millones de personas y otras entidades. Para cada entidad, cada mes, se cuenta con 43 columnas de información, donde 2 de ellas corresponden a la llave primaria de la tabla (RUT, Fecha) y las otras 41 corresponden a información financiera.

Dado que la SBIF lleva un registro de todas las entidades que han contraído una deuda en el país y el modelo se limita a personas, será necesario realizar ciertos filtros a las filas de la tabla para poder excluir a las empresas y otras entidades que no se consideran dentro del alcance del estudio, dejando sólo a las personas naturales.

En esta misma línea, para hacer más precisa la selección de las personas a las que el modelo quiere apuntar, sólo se considerará a aquellas personas que hayan tenido al menos una deuda de cualquier tipo dentro de los últimos seis meses, a modo de dejar fuera a todas aquellas personas que entraron en el sistema bancario por un motivo puntual, pero que no han seguido siendo parte activa de este.

Finalmente, dado que la idea central del modelo es crear un *score* de propensión al crédito para que el banco pueda elegir de mejor manera a sus clientes, se excluirá a todas aquellas personas que no cumplan con las políticas más duras de riesgo. De esta manera, sólo se trabajará con aquellas personas que cumplan con todos los

filtros de riesgo y adicionalmente, con aquellas que si bien no pasan todos los filtros, si pasarán los más duros. Tras correr todos los filtros previamente mencionados, se formará la primera selección del público objetivo.

Con la base ya de un tamaño más abordable y conteniendo sólo a aquellas personas de interés para el modelo, se procederá a estudiar cada una de las columnas de la base de datos en profundidad. El primer análisis consistirá en ver cuáles son los posibles valores que pueden tomar cada uno de los campos y con qué frecuencia aparecen. Se eliminarán todas aquellas columnas en donde un valor se repita en más del 95% de los casos, por considerarse columnas no discriminatorias. Otros análisis a las columnas de la base se llevarán a cabo más adelante, en la sección 4.5.

4.2 Elaboración y optimización de la variable de respuesta

Para poder entrenar el modelo se requiere definir cuál será la variable dependiente. En este caso, y por consideraciones de negocios, se establece que dicha variable será una variable binaria que tomará valor 1 si el cliente tomó un crédito de consumo en este u otro banco en alguno de los tres meses a partir del mes en el que se recibe la información de la SBIF, o bien 0 en otros casos. Es decir, la variable dependiente se puede expresar como:

$$Response = MAX(Trigger_t, Trigger_{t+1}, Trigger_{t+2})$$

Donde $Trigger_t$ toma valor 1 si en el mes t se realizó una operación de compra de crédito de consumo en este u otro banco y 0 en otros casos.

Es aquí donde aparece la primera complicación, dado que la información sobre la deuda de consumo que entrega la SBIF mensualmente a todos los bancos, viene agregada a nivel de persona, y considerando que en Chile una persona promedio

posee deudas en tres bancos al mismo tiempo, es imposible saber a ciencia cierta cuánto de esta deuda corresponde a qué producto de consumo (créditos, tarjetas o líneas de crédito) o a qué institución. A modo de ejemplo, una persona podría haber usado sus tarjetas de crédito u otros productos de diferentes bancos y endeudarse por un monto equivalente o incluso mayor al mínimo que exigen las instituciones para tomar un crédito de consumo. Este monto podría ser confundido a simple vista como un crédito de consumo en la información que provee la SBIF, pero en realidad puede corresponder a cualquier otra cosa.

Dado que sólo es posible tener certeza de las operaciones que fueron realizadas en este banco en específico, es necesario desarrollar un método que permita identificar cuándo una persona tomó un crédito de consumo en otra institución, a modo de disminuir los falsos negativos. Para poder llevar esta tarea a cabo, será necesario acudir a la información propia del banco.

A la base de datos extraída de la fase de *data cleansing* se le aplicarán tres filtros más en el siguiente orden:

Seleccionar solamente a las personas que:

a) Alguna vez hayan sido clientes del banco.
b) Hayan tenido sólo una institución con deuda en todo su historial.
c) Hayan comprado un crédito de consumo en el banco al menos una vez.

El fin de aplicar estos filtros es poder estudiar cómo se mueve la variable deuda de consumo en el momento en que una persona toma un crédito dentro del banco, sin tener el ruido asociado a que el cliente pudiese tener deudas en más de una institución, para luego salir en búsqueda de este mismo movimiento en toda la base de datos de la SBIF y así poder identificar cuándo una persona cursó una operación en alguna otra institución.

Con esta nueva información, el problema de elaborar la variable de respuesta $Response$ se puede subdividir en dos etapas: primero, determinar si el cliente tomó un crédito de consumo en este banco utilizando la información directamente desde los registros de ventas; y segundo, determinar si el cliente tomó un crédito de consumo en otro banco mediante el método previamente descrito. En definitiva, la variable $Response$ se puede describir de la siguiente manera:

$$Response = MAX(OriginalResponse, Jump)$$

Donde $OriginalResponse$ se define como:

$$OriginalResponse = MAX(Trigger_t, Trigger_{t+1}, Trigger_{t+2})$$

Donde $Trigger_t$ toma valor 1 si en el mes t se realizó una operación de compra de crédito de consumo en este banco y 0 en otros casos. Y de manera análoga, $Jump$ se define como:

$$Jump = MAX(Jump_t, Jump_{t+1}, Jump_{t+2})$$

Donde $Jump_t$ toma valor 1 si en el mes t se realizó una operación de compra de crédito de consumo en otro banco y 0 en otros casos.

Con esta nueva variable $Response$, se soluciona el problema de los falsos negativos inherentes a sólo ver los registros de ventas del banco y por sobre todo, el problema relacionado a no poder captar la información de las personas que no son actualmente clientes. Adicionalmente, se trata parcialmente el problema en donde, dado que $Jump$ es una variable estimada y posee un error asociado, existen casos en los que este no considera como compras de créditos de consumo

operaciones que realmente lo fueron, lo que entrega robustez a la variable, al menos en los casos de falsos positivos dentro de esta institución.

Con la variable dependiente correctamente definida se procederá a seleccionar de la muestra original —proveniente de la fase de Limpieza de datos— 45 mil personas para entrenar el modelo. Estas contarán con la condición de alguna vez haber sido clientes del banco y alguna vez haber comprado un crédito de consumo en esta institución. De esta manera se podrá trabajar con personas que han cursado operaciones de las cuales se posee un total nivel de información, y hacer un *over sample* de la muestra de interés, es decir, aumentar el número de casos donde la característica deseada —personas que hayan tomado un crédito de consumo— se haga presente.

4.3 Generación de características

A partir de los campos seleccionados tras la fase de Limpieza de datos, sumado a la profundidad histórica de estos mismos, se generarán distintos tipos de variables con el fin de representar el pasado financiero de cada una de las personas a través de una foto del presente (Domingos, 2012), tal como se explica previamente en el Modelo Conceptual. Para esto se utilizarán diversas técnicas que buscan hacer evidente tanto la interacción entre variables, como la importancia de la componente temporal. Entre las técnicas más básicas están las de incluir en la foto del presente, el valor puntual de la variable en fechas pasadas, la creación de indicadores que muestren la presencia de una característica deseada en el tiempo y también el cálculo de mínimos, medianas, promedios y máximos para distintas profundidades históricas.

El segundo tipo de variables creadas corresponden a la suma, resta, multiplicación o división de los campos originales de la base de datos de la SBIF. Esto, a modo de hacer evidente relaciones e interacciones entre variables que posean un

significado desde el punto de vista del negocio. Ejemplos de estas variables son la suma de la deuda de consumo con el cupo disponible, donde se aprecia, de manera aproximada, el cupo total de la persona. Otro ejemplo es el del nivel de deuda dividido por el número de instituciones, donde se aprecia, también de manera aproximada, el nivel de deuda promedio por institución. Para todas estas variables se realizará el mismo procedimiento mencionado en el párrafo anterior a modo de incorporar información del pasado en la foto del presente.

El tercer tipo de variables creadas corresponde a variables referidas a modelos de RFM, en dónde se utilizará el campo *Response* para calcular de distintas maneras la recencia, frecuencia y monto con el que las personas toman un crédito de consumo. Estas variables son relevantes, ya que permiten, según los estudios de Hossein Javaheri (2008) y Moro, Cortez y Rita (2014), describir de mejor manera el comportamiento de las personas al incorporar de lleno la componente temporal en el análisis.

El cuarto tipo de variables que se utilizará es el de las variables derivadas a partir de modelos simples, como lo es calcular la edad de una persona a partir de su RUT. En este caso en particular lo que se hará será ajustar una regresión polinomial de segundo grado a un grupo de 148.529 personas cuyas edades y RUTs son conocidos. Gracias a esta información, será posible estimar de manera aproximada el ciclo de vida en que se encuentran las distintas personas. En la figura 4.3.1 se puede observar el modelo generado.

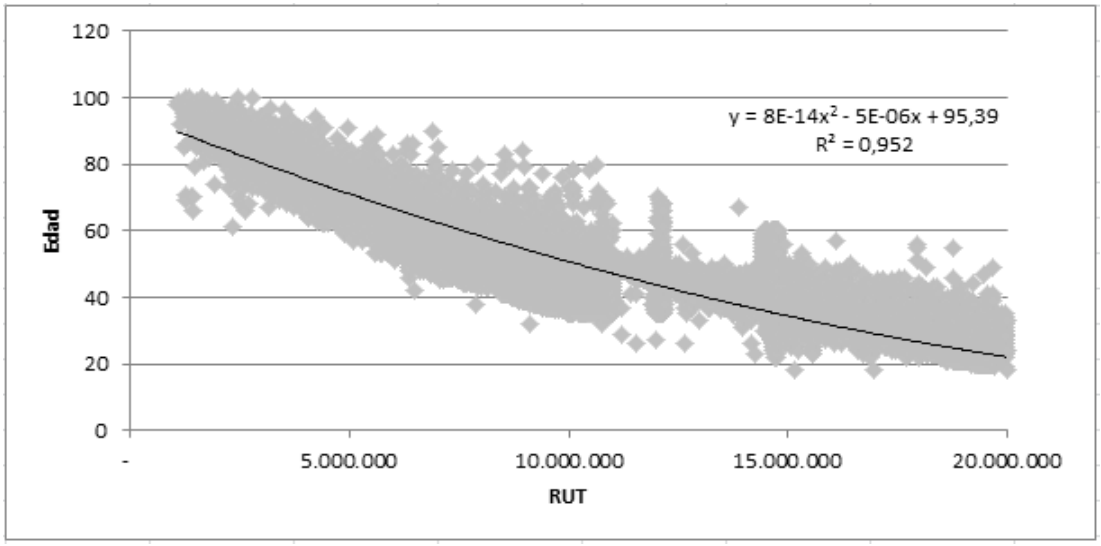

Figura 4.3.1: Cálculo de la edad a partir del RUT

Finalmente, se combinarán de distintas maneras todos los tipos de variables descritas en los cuatro párrafos anteriores, a modo de hacer evidente cualquier relación que pudiese tener sentido desde el punto de vista del negocio.

Las variables generadas utilizarán información desde noviembre de 2012 (fecha a partir de la cual se encuentra estructurada la data del banco) hasta diciembre de 2016, para predecir la venta de un crédito de consumo en la ventana comprendida entre febrero de 2017 y abril del mismo año.

El hecho de que no se considere enero de 2017 para el análisis se explica por el desfase temporal de 40 días que tiene la data de la SBIF, como fue mencionado al principio de este documento. La siguiente imagen (figura 4.3.2) explica de manera gráfica la información que se utilizará para construir las variables (zona de color gris), la zona del pasado sobre la que no se posee información (de color blanco) y la ventana temporal donde se pretende predecir la compra de un crédito de consumo (de color rojo). Adicionalmente, la línea punteada de color negro indica

la fecha en la que llega la información de la SBIF del último mes disponible y por lo tanto el momento en el que se ejecutará el modelo.

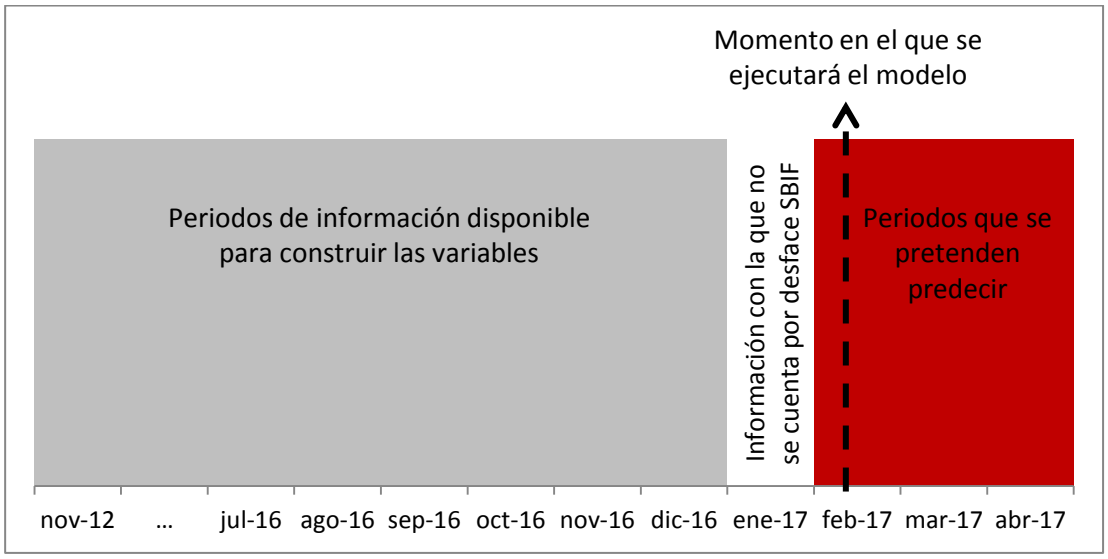

Figura 4.3.2: Información disponible para predecir

4.4 Segmentación

Las variables generadas en la etapa previa, formarán dos grupos claramente diferenciados. Este es el caso de los clientes que hasta la fecha hayan tenido como máximo un *jump* versus aquellos que hasta el mismo momento hayan tenido más de uno.

Generar esta distinción es clave, dado que las variables que aplican para el primer grupo, no necesariamente aplicarán para el segundo. Ejemplos evidentes de estos casos son las variables que miden comportamientos entre *jumps* —como las de los modelos RFM— en donde el primer grupo no posee información alguna, mientras que para el segundo grupo, este tipo de variables entregarán información muy relevante que permitirá entender de manera más amplia el comportamiento

histórico de las personas. Por este motivo, desde este punto en adelante, se separará la muestra en personas que a la fecha sólo hayan tenido un *jump* (NumJump = 1) y personas que a la fecha hayan tenido más de un *jump* (NumJump > 1).

Realizar esta segmentación permitirá aprovechar al máximo la información que cada grupo pueda otorgar, dado que del segundo grupo se puede capturar más información y con mayor nivel de detalle, por lo que se desarrollarán modelos distintos para cada uno de los segmentos.

4.5 Selección de características

La primera selección de variables se llevará a cabo utilizando el análisis de correlación de Pearson de R (R Core Team, 2014) para eliminar todas aquellas variables que tengan un coeficiente de correlación mayor a 90% con alguna otra por considerarse colineales, procurando dejar una variable de cada par correlacionado para no perder dicha información.

No obstante, no todas estas variables serán utilizadas por el modelo final, por lo que queda comprobar cuáles de estas variables son realmente relevantes para el desarrollo del mismo. Para cada uno de los segmentos previamente diferenciados (NumJump = 1 y NumJump > 1) se utilizará tres métodos seleccionadores de características: *Recursive Feature Elimination* (RFE), Boruta y XGBoost.

RFE se basa en la idea de iterativamente construir modelos, escoger las variables que mejor o peor desempeño muestran respecto a la precisión objetivo, dejarlas de lado y volver a generar el modelo con los campos restantes (Granitto, Furlanello, Biasioli & Gasperi, 2006). En este caso los modelos fueron construidos en base a la técnica de bosques aleatorios (Liaw & Wiener, 2002). Este proceso termina una

vez que se han agotado todas las columnas de la base de datos y luego se procede a rankear las variables según el momento en el que fueron eliminadas.

Boruta se basa en la misma idea que constituye las bases del clasificador de los bosques aleatorios, es decir, que mediante la adición de aleatoriedad al sistema y la recopilación de resultados del conjunto de muestras aleatorizadas, se puede reducir el impacto engañoso de las fluctuaciones y las correlaciones aleatorias. De esta manera, el método proporciona una visión más clara de cuáles son los atributos realmente importantes (Kursa & Rudnicki, 2010).

XGBoost es una implementación del método *Gradient Boosted Decision Trees* diseñado para correr de manera rápida y mejorar el rendimiento computacional del algoritmo, razón que explicaría el que domine la tabla de posiciones en las últimas competencias de aprendizaje automático aplicado Kaggle. El método consiste en construir un modelo de manera similar a como lo hacen otros métodos de *boosting* y luego generalizarlo permitiendo optimizar una función de pérdida diferenciable arbitraria (Friedman, 2001). Para esto, crea modelos de predicción en forma de ensambles de modelos de clasificación débil —por lo general árboles de decisión— los que clasifican la muestra secuencialmente obteniendo errores menores al 50%. Así, se consigue un clasificador robusto a partir de clasificadores débiles. En este caso se optimizaron tres funciones por separado: *Accuracy*, *Precision* y ROC.

Para cada una de las técnicas mencionadas previamente, se seleccionará el top 100 de las variables más influyentes, se cruzarán los resultados para las tres técnicas y se seleccionarán los valores únicos transversales a las tres respuestas. Posteriormente estas variables serán rankeadas promediando el ranking en el que aparecieron en sus listas originales. En caso de que una variable aparezca en una lista y en otra no, se considerará que su ranking es el último, es decir 100. De esta

manera, se obtendrá una lista en donde las primeras variables serán las que según los tres métodos seleccionadores de variables son las más influyentes y las últimas serán las que probablemente menos influyan en la predicción. Esta pre selección servirá para hacer más eficiente la selección del modelo.

4.6 Ajustes preliminares para la selección del modelo predictivo

Se utilizará una selección de cinco algoritmos predictivos de aprendizaje automático: regresiones logísticas, bosques aleatorios, máquina de soporte vectorial, XGBoost y redes neuronales (Delen, Sharda & Kumar, 2007). Estos cumplen con la característica de ser computacionalmente implementables o haber probado su efectividad en varios proyectos o competencias, como la competencia Kaggle de aprendizaje automático (Friedman, Hastie & Tibshirani, 2001).

Las regresiones logísticas, son una opción popular que opera una transformación logística no lineal suave sobre un modelo de regresión múltiple y permite la estimación de probabilidades de clase. Debido a la combinación lineal aditiva de sus variables independientes, el modelo es fácil de interpretar. Sin embargo, el modelo es bastante rígido y no puede modelar relaciones no lineales complejas (Venables & Ripley, 2002).

Los bosques aleatorios son una combinación de árboles predictores, donde cada árbol es una estructura de bifurcación que representa un conjunto de reglas en una forma jerárquica. Cada iteración depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. Esta representación se puede traducir en un conjunto de reglas del estilo “si..., entonces...”, las que permiten entender más fácilmente la mecánica que hay detrás de la decisión (Breiman, Friedman, Stone & Olshen, 1984).

La máquina de soporte vectorial (SVM, por sus siglas en inglés) transforma un *input* m-dimensional en un espacio n-dimensional mediante el uso de un mapeo no lineal. Luego, encuentra el mejor hiperplano de separación lineal, relacionado a un conjunto de puntos de vectores de soporte, para separar dos clases en los espacios más amplios posibles (SVM) (Cortes & Vapnik, 1995).

XGBoost fue explicado en la sección anterior. Sin embargo, este algoritmo cuenta con dos modalidades: selección de variables y modelo predictivo. En esta ocasión se utilizará la modalidad de modelo predictivo (Friedman, 2001).

Finalmente están las redes neuronales (NN, por sus siglas en inglés). El perceptrón multicapa es la arquitectura de redes neuronales más popular. Se adoptó un perceptrón multicapa con n capas ocultas de H nodos ocultos y un nodo de salida. El hiperparámetro H establece el modelo de complejidad de aprendizaje. Una NN con un valor de H = 0 es equivalente al modelo regresión lineal, mientras que un valor alto H permite que la NN aprenda relaciones no lineales complejas. Dado que se utiliza la función logística, el nodo de salida produce automáticamente una estimación de probabilidad ([0, 1]). La solución final de NN depende de la elección de los pesos iniciales (Haykin, 2009).

Los modelos de regresiones logísticas y XGBoost serán programados en el lenguaje estadístico R, mientras que los Bosques aleatorios, SVM y Redes neuronales serán programados en el lenguaje de programación Python debido a las exigencias computacionales que presentan. En el caso de las redes neuronales se utilizará Keras con Tensorflow como *backend* para probar distintas arquitecturas con diferentes cantidades de capas ocultas. Todos estos modelos serán ejecutados en un computador de cuatro núcleos con procesador Intel Core i5-6300U CPU @2.40GHz, 16GB de RAM y sistema operativo Windows 7 de 64 bits.

Hasta este punto aún no se ha decidido cuál de los cinco modelos mencionados previamente se utilizará para predecir, y dado que cada modelo utiliza distintos parámetros para calibrarse, se procederá a atacar este problema previo a la fase de seleccionar el modelo. En primer lugar se seleccionarán los parámetros que optimicen cada modelo para luego seleccionar el mejor modelo y la cantidad de variables óptimas que maximicen los resultados de la predicción.

Para entrenar a los distintos modelos y determinar cuáles son los parámetros que los optimizan, se dividirá la muestra en una sección de entrenamiento (70% de los datos) y una de validación (30% restante). Se utilizará la lista de variables calculadas en la pre selección y se intentará predecir la probabilidad de compra de un crédito de consumo. Para comprobar si la predicción fue correcta, se utilizará el vector de respuesta *Response* calculado como se explica en la sección 4.2, que considera tanto la información de los créditos vendidos en este banco, como la estimación realizada para saber si una persona compró un crédito en alguna otra institución a través de la ecuación del *Jump*.

La selección de los parámetros que optimizan a cada uno de los modelos se llevará a cabo de manera automática utilizando *Grid Search*, del paquete Caret en R o bien Scikit-Learn de Python dependiendo del caso. Los distintos parámetros utilizados para ejecutar los modelos se encuentran detallados en el Anexo 3.

4.7 Selección del modelo predictivo

Para poder comparar distintos modelos se adoptó el criterio de análisis LIFT. Esta metodología consiste en ordenar los registros desde el mayor hasta el menor puntaje y luego dividir la muestra en 10 grupos con la misma cantidad de registros, donde el primer decil posee los registros con el mayor puntaje y el décimo, aquellos con el menor puntaje. Estos 10 grupos se ubican en el eje x, quedando a la izquierda el grupo más propenso y hacia la derecha los menos. En el eje y, por otra

parte, se cuantifica el porcentaje de respuestas positivas reales que obtuvo cada grupo a partir de datos conocidos. De esta manera, se entiende que un modelo ideal generaría un gráfico LIFT con forma de “L”, es decir, entre los grupos más propensos se encuentran la mayor cantidad de respuestas positivas, mientras que un modelo perfectamente aleatorio debería entregar como resultado una recta horizontal, ya que todos los grupos debieran tener el mismo porcentaje de respuestas positivas (Coppock, 2002).

Al área acumulada bajo esta curva se le conoce como ALIFT, y se utiliza para determinar con cuántos deciles se debe trabajar para lograr abarcar un cierto porcentaje de las respuestas positivas. Como es de suponer, a medida que se incorporan nuevos deciles, aquellos de menor *score*, la cantidad de respuestas positivas crece, pero lo hace a tasas decrecientes hasta llegar a un punto en el que se abarca toda la población, alcanzando un 100% de respuestas positivas.

Siguiendo en esta línea, se procederá a correr cada uno de los modelos entrenados en la sección anterior —cuyos parámetros fueron optimizados— sobre el set de validación (30% de la muestra que fue excluida para el entrenamiento) para generar un puntaje que indique la probabilidad o propensión de una persona a adquirir un crédito de consumo en la ventana de observación. A partir de este valor, se generará el gráfico LIFT y se comparará la precisión dentro del primer grupo entre los distintos modelos.

Adicionalmente, se probará variar el número de variables a utilizar y se medirá el tiempo de ejecución de cada una de las opciones, variables que también son relevantes al momento de hacer la selección final el modelo.

4.8 Flujograma

A modo de explicar de manera simple y gráfica la metodología a utilizar, se proponen los siguientes flujogramas. En la figura 4.8.1 se muestra el flujo de la información desde que se toma la base de datos proveniente de la SBIF hasta que se genera la segmentación propuesta. En la figura 4.8.2, se muestra el flujo que cada una de estas muestras segmentadas deberá seguir hasta llegar a la selección final del modelo.

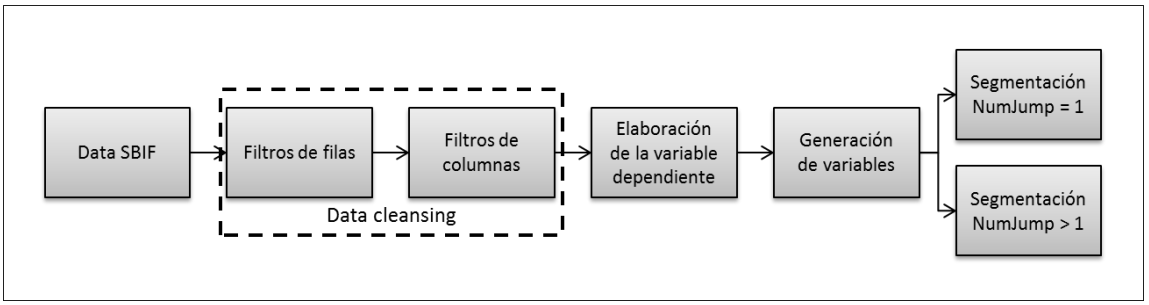

Figura 4.8.1: Flujograma desde la data cruda hasta la segmentación

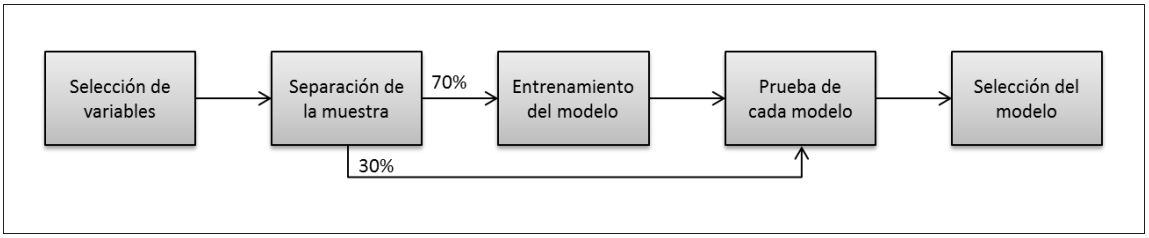

Figura 4.8.2: Flujograma desde la selección de variables hasta la selección del modelo

5. RESULTADOS

5.1 Limpieza de datos

Tras aplicar los filtros para excluir a empresas y otras entidades que están fuera del alcance del estudio, se logró disminuir la base de 430 millones de registros a sólo 130 millones. Concordantemente, la cantidad de RUTs distintos dentro de la base disminuyó de 10.3 a 2.6 millones. Adicionalmente, al excluir a aquellas personas que llevan inactivas los últimos seis meses en el sistema financiero, se redujo la base de 130 a 106 millones de registros y el número de RUTs bajó de 2.6 a 2 millones. Finalmente al excluir aquellas personas que no cumplen los filtros de riesgo se redujo la base de 106 a 85 millones de registros y la cantidad de RUTs descendió de 2 a 1.6 millones de personas.

Al correr el filtro sobre las columnas de la base, el resultado fue que algunas variables poseían un sólo valor posible y muchas otras tenían un sólo valor que se repetía en la mayoría de las ocasiones. Con esto, se redujo la cantidad de columnas de 43 a 10, donde 2 de ellas corresponden a la llave primaria (RUT y fecha) y 8 corresponden a información financiera de la persona (deuda directa vigente, deuda de consumo, cupo disponible, instituciones con deuda, deuda comercial, instituciones con deuda comercial, deuda hipotecaria y deuda indirecta vigente).

En la Tabla 5.1.1 se puede apreciar cómo cada paso aportó a la reducción del volumen de los datos.

Tabla 1Tabla 5.1.1: Volumen de datos tras aplicar filtros

	Base Completa	**Filtro 1**	**Filtro 2**	**Filtro 3**	**Filtro 4**
Filas	430.000.000	130.000.000	106.000.000	85.000.000	85.000.000
Columnas	43	43	43	43	10
Tamaño de la tabla en millones de celdas	**18.490**	**5.590**	**4.558**	**3.655**	**850**

5.2 Elaboración y optimización de la variable de respuesta

Para poder estudiar el comportamiento de la variable deuda crédito de consumo proveniente de la información que provee la SBIF, y así identificar cuando una persona toma una operación en otra institución, se aislaron 3.245 casos de personas que cumplen con los tres filtros propuestos en la metodología, a saber, que hayan sido clientes del banco, que sólo hayan tenido deudas en una institución y que alguna vez hayan comprado un crédito de consumo. Para todos estos RUTs se extrajo la información de su deuda con una profundidad histórica máxima de 50 meses (152.793 registros).

La primera hipótesis fue que si la diferencia entre un mes y otro en la deuda de consumo era mayor que el monto mínimo factible, entonces es porque se había cursado una operación. Sin embargo, se concluyó que no se puede considerar el monto mínimo que exigen los bancos para tomar un crédito de consumo como gatillador de una compra, dado que la variable deuda de consumo también incluye movimientos de tarjeta y línea de crédito, los que pueden alcanzar montos superiores al monto mínimo de un crédito de consumo.

La segunda hipótesis surgió tras estudiar el histograma de las diferencias en la deuda de consumo para los meses en los que las personas aparecían con una operación de consumo registrada, como aparece en la figura 5.2.1. De aquí se obtuvo que el 80% de las operaciones se cursan por montos mayores o iguales a tres veces el monto mínimo, por lo que se propuso dicho monto como gatillador de una operación. Sin embargo, esta opción fue descartada, dado que las variaciones en las deudas de personas de alto patrimonio eran demasiado altas y en las de bajo patrimonio demasiado bajas, por lo que en el primer caso se generaban saltos constantemente, mientras que en el segundo no se gatillaban nunca, razón por la cual se descartó también la posibilidad de utilizar un número fijo que definiera la compra de un crédito de consumo.

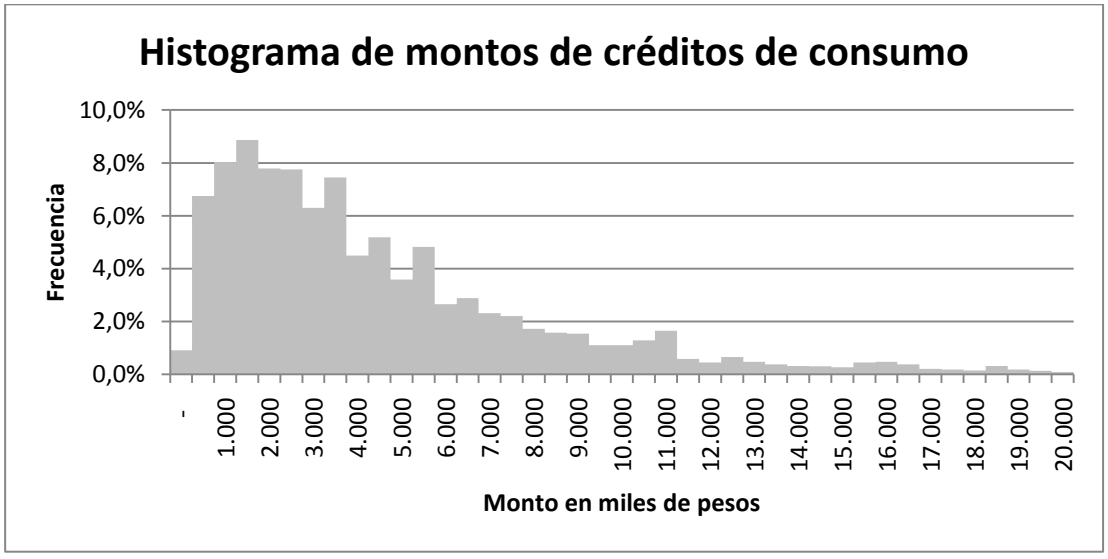

Figura 5.2.1: Distribución de montos de créditos de consumo

Finalmente, siguiendo las ideas propuestas por Nun et al. (2015) quienes estudian y caracterizan estrellas de manera individual utilizando el espectro de luz que estas emiten mediante series de tiempo, surge la hipótesis de encontrar algún tipo de patrón de comportamiento en las diferencias en el nivel de deuda de crédito de

consumo mes a mes. Con esto, lo que se busca es comparar a cada persona consigo misma y con sus niveles de deuda en el tiempo, en vez de tratar a los individuos como miembros de un clúster, donde se asume que todos los miembros poseen el mismo comportamiento. En la figura 5.2.2 se aprecia en color rojo la curva que grafica el nivel de deuda de consumo de una persona en particular en el tiempo. En color gris, se aprecia la diferencia en el nivel de deuda de consumo entre el mes actual y el mes anterior.

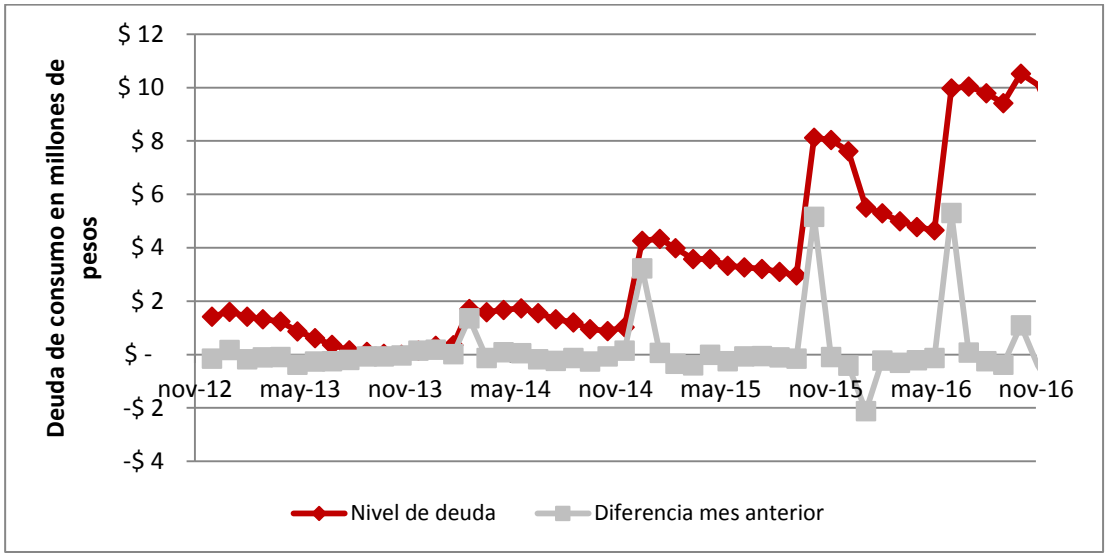

Figura 5.2.2: Nivel de deuda y diferencia con el mes anterior para un sujeto en particular

Tras evaluar distintas curvas, estudiar sus comportamientos y considerando que en las series de tiempo la data no se genera independientemente, que la dispersión varía en el tiempo y que por lo general tienen una tendencia o comportamiento cíclico (Falk, 2012), se propuso la siguiente fórmula:

$$Jump = \begin{cases} 1 & si\ DiffDeuda > AVG(DiffDeuda) + \alpha * STDEV(DiffDeuda) \\ 0 & en\ otros\ casos \end{cases}$$

Donde $DiffDeuda$ representa la diferencia entre el nivel de deuda de consumo del mes actual menos el nivel de deuda del mes anterior y α es un parámetro a determinar. En definitiva, lo que se busca es determinar —a nivel de individuo— cuál es el rango normal de deuda en el que se mueve la persona.

Se iteró dentro de los 3.245 casos para encontrar el parámetro que minimiza la cantidad de falsos positivos más falsos negativos en comparación a los datos reales, a modo de disminuir el error asociado a este nuevo *jump*. El resultado de esta optimización fue = 2,3 con un error asociado de 1,53%. Los resultados numéricos tras esta iteración se pueden apreciar en la tabla 5.2.3 y la curva que se genera al graficar la suma de falsos positivos más falsos negativos, en la figura 5.2.4.

Tabla 5.2.1: Resultados de la optimización del parámetro α

Valores del parámetro	Verdaderos positivos	Verdaderos negativos	Falsos positivos	Falsos negativos	Porcentaje de falsos
2,00	3.881	146.407	952	1.462	1,58%
2,05	3.843	146.454	905	1.500	1,58%
2,10	3.820	146.498	861	1.523	1,56%
2,15	3.788	146.546	813	1.555	1,55%
2,20	3.759	146.593	766	1.584	1,54%
2,25	3.723	146.629	730	1.620	1,54%
2,30	**3.692**	**146.669**	**690**	**1.651**	**1,53%**
2,35	3.662	146.698	661	1.681	1,53%
2,40	3.636	146.721	638	1.707	1,54%
2,45	3.601	146.747	612	1.742	1,55%
2,50	3.559	146.769	590	1.784	1,55%

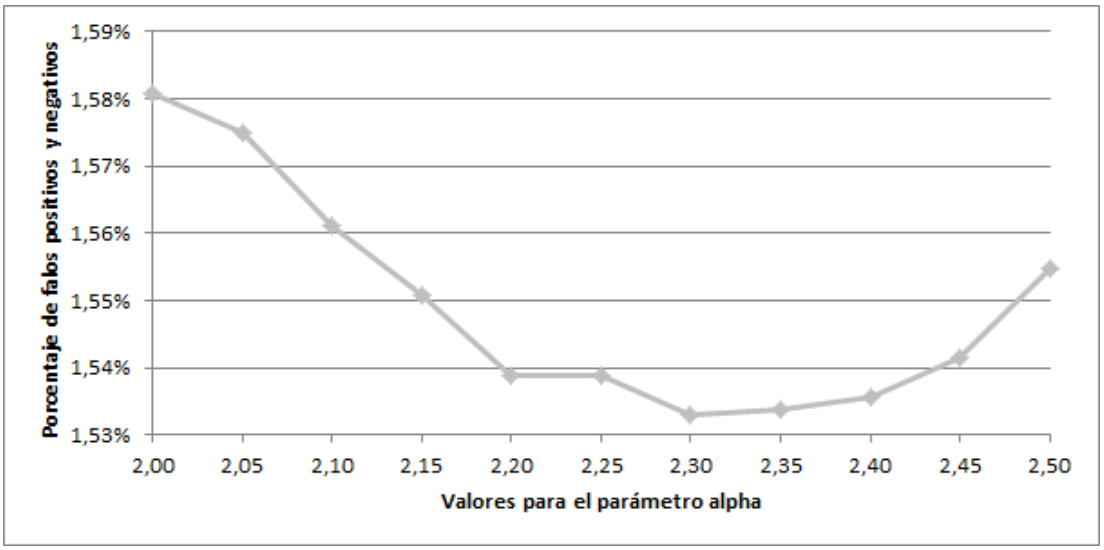

Figura 5.2.3: Curva de porcentaje de falsos para distintos valores del parámetro

Para asegurar la optimalidad y estabilidad en el tiempo de este parámetro, se varió la ventana de observación con la que se calculó. Al desfasar la ventana en un mes se obtuvieron 3.260 RUTs (156.439 registros), un coeficiente de 2,35 y nuevamente error del 1,53%. Al realizar más pruebas se concluye que el error se mantiene constante en torno al 1,5%, mientras que el parámetro α varía levemente dependiendo de la muestra, pero siempre en las vecindades del 2,3, por lo que se tomó este valor para realizar todos los cálculos. En síntesis, la ecuación para calcular si una persona tomó o no un crédito en otro banco se resume en:

$$Jump = \begin{cases} 1 & si\ DiffDeuda > AVG(DiffDeuda) + 2{,}3 * STDEV(DiffDeuda) \\ 0 & en\ otros\ casos \end{cases}$$

A modo de hacer más tangible la ecuación obtenida tras el proceso de optimización, al aplicar esta metodología sobre el ejemplo señalado previamente en la figura 5.2.2, se obtiene la figura 5.2.5, donde la línea verde representa el promedio de las diferencias en la deuda de consumo (primer término de la ecuación del *Jump*) y la línea azul representa la banda generada al sumar 2,3 veces

la desviación estándar al promedio (primer más segundo término de la ecuación del *Jump*). Es decir, cada vez que la línea gris, que representa la diferencia en la deuda de consumo entre el mes actual y el mes anterior, sobrepase la línea azul, indica que el cliente tomó un crédito de consumo.

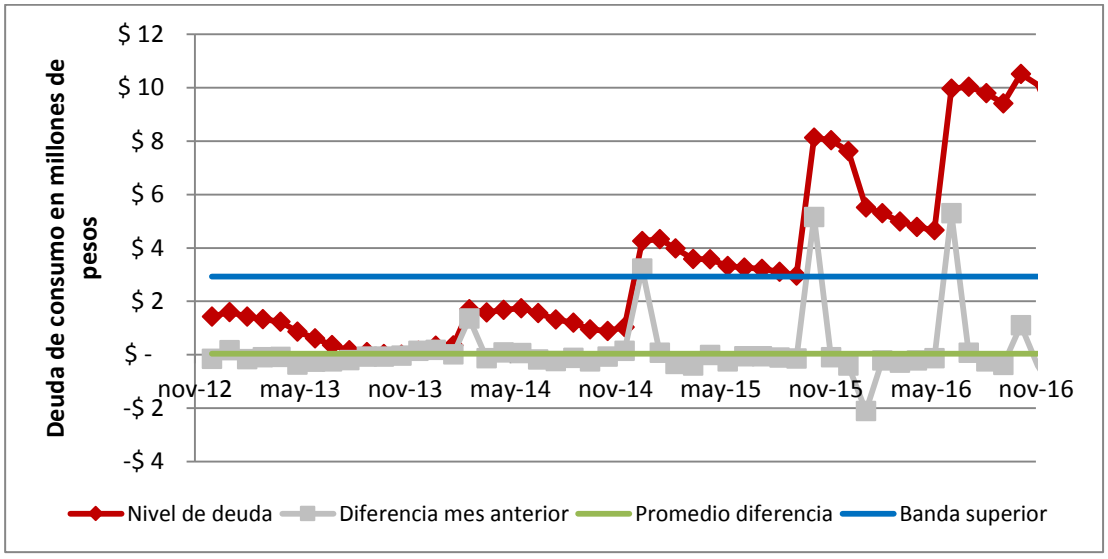

Figura 5.2.4: Explicación gráfica del cálculo del *Jump*

5.3 Generación de características

Tras incorporar valores pasados de los 10 campos originalmente extraídos de la base de datos de la SBIF a la foto financiera del presente de los sujetos, más información de mínimos, medianas, promedios, máximos, indicadores de distintos tipos, variables calculadas mediante modelos de RFM, edad y combinaciones varias de todo lo anterior, el resultado final ascendió a 531 variables distintas, las que se encuentran disponibles en los anexos.

5.4 Segmentación

La muestra original que se planteó en la metodología para entrenar el modelo, es decir, las 45 mil personas que pasen los filtros de la fase de Limpieza de datos, hayan sido clientes del banco y hayan cursado alguna vez una operación de consumo en esta institución, contiene exactamente 45.024 RUTs con una tasa de respuesta del 12.27%. Tras dividir la muestra en los dos grupos mencionados en la metodología, el grupo que sólo ha tenido un *jump* a la fecha quedó con 12.098 RUTs (26.88% del total) y una tasa de respuesta de 10.28%, mientras que el grupo que ha tenido más de un *jump*, quedó compuesto por 32.926 RUTs (73.12% del total) con una tasa de respuesta de 13.0%.

5.5 Selección de características

Dado que a partir de los 10 campos originales de la base de datos de la SBIF se generaron 531 campos con información, es lógico pensar que existe cierto grado de correlación entre ellos. Tras aplicar el análisis de correlación de Pearson propuesto en la metodología, fue posible eliminar 320 columnas para el caso de NumJump = 1, quedando 211 variables en la base de datos; mientras que para el caso de NumJump > 1, se eliminaron 315 columnas, quedando 216 campos con información.

Posteriormente, al ejecutar los modelos de selección de variables —mencionados en la sección 4.5— sobre el conjunto de campos resultantes de la fase anterior, seleccionar el top 100 de las variables más influyentes para cada modelo, cruzar todos los resultados y seleccionar los valores únicos presentes en las tres respuestas, se redujo de 211 a 161 columnas para el caso de NumJump = 1 y de 216 a 169 para el caso de NumJump > 1.

Esta primera selección sirve para acotar y hacer más eficiente las opciones de búsqueda de los modelos predictivos, gracias a que los algoritmos deberán correr sobre 161 o 169 campos (dependiendo del caso) y no sobre la población total de 531 campos generados. Cabe destacar que estas selecciones de más de 160 variables no son las definitivas, sino que generan el marco dentro del cual estará el conjunto final de variables a seleccionar por los modelos predictivos a modo de mejorar la precisión.

5.6 Selección del modelo predictivo

Para llegar al modelo final se fueron descartando opciones por distintos motivos. En primer lugar, se comparó el resultado del análisis LIFT para todos los modelos. Cuatro de los cinco modelos estudiados, entregaron resultados cercanos al 18% de precisión evaluando el primer decil del segmento NumJump = 1 y 20% en el segmento NumJump > 1, a excepción del modelo de regresión logística, el cual no superó el 16%, razón por la cual fue descartado. En la tabla 5.6.1 se registran los valores precisión para el primer grupo en el gráfico LIFT en cada uno de los modelos.

Tabla 5.6.1: Valores de precisión para el primer decil

Modelo	**Segmento NumJump =1**	**Segmento NumJump > 1**
Regresión logística	13,8%	15,4%
Bosques aleatorios	18,9%	20,4%
SVM	17,7%	20,8%
XGBoost	19,2%	21,0%
Redes neuronales	19,1%	21,2%

En el caso de los modelos de Bosques aleatorios y SVM, también fueron descartados, pero esta vez no por su mal desempeño, sino que por el largo tiempo de ejecución computacional que requieren. En la tabla 5.6.2 los tiempos de ejecución de los modelos.

Tabla 5.6.2: Tiempo de ejecución de los modelos

Modelo	**Segmento NumJump =1**	**Segmento NumJump > 1**
Regresión logística	< 5 min	< 5 min
Bosques aleatorios	~ 36 hrs	NA
SVM	~ 24 hrs	NA
XGBoost	~ 10 min	~ 15 min
Redes neuronales	~ 10 min	~ 15 min

No obstante, como XGBoost y las redes neuronales entregaron resultados parecidos en tiempos similares, se profundizó en ambos modelos. En primer lugar, se probó reducir el número de variables para ver dónde se optimiza el resultado. Específicamente se probó con los siguientes números de variables: 10, 15, 16, … , 25, 30, 50, 100, siempre seleccionando los campos desde la lista previamente ordenada desde la más a la menos influyente. La cantidad de variables que optimiza los resultados fueron los siguientes:

Tabla 5.6.3: Cantidad óptima de variables

Modelo	**Segmento NumJump =1**	**Segmento NumJump > 1**
XGBoost	23	24
Redes neuronales	25	25

El número de variables obtenido en ambos casos está dentro del rango de los resultados conseguidos por Moro, Cortez y Rita (2014): 22 características; y Hossein Javaheri (2008): 26 características, lo que valida el resultado. Sin embargo, esta modificación no alteró significativamente la diferencia en el porcentaje de precisión del primer grupo del gráfico de LIFT entre ambos modelos, por lo que se realizó la prueba de bondad de ajuste Kolmogorov-Smirnov (KS) a modo de discriminar y seleccionar al modelo ganador. La tabla 5.6.4 muestra los resultados del *test* KS para ambos modelos aplicados en ambos segmentos. Como es posible apreciar, XGBoost obtiene un mejor resultado en ambos segmentos.

Tabla 5.6.4: Nivel de confianza (KS) para ambos modelos en ambos segmentos

Modelo	**Segmento NumJump =1**	**Segmento NumJump > 1**
XGBoost	0,39	0,25
Redes neuronales	0,19	0,14

Al contrastar cada uno de los individuos seleccionados en los distintos deciles por ambos métodos y en cada uno de los segmentos, se aprecia una mayor concordancia en los primeros y últimos deciles que en los deciles centrales. Es decir, los mayores grados de concordancia se dan en los extremos y la mayor variabilidad en el centro. En la tabla 5.6.5 se aprecia el porcentaje de concordancia por decil entre ambos modelos para cada uno de los segmentos.

Tabla 5.6.5: Porcentaje de concordancia entre modelos decil a decil

Decil	Segmento NumJump =1	Segmento NumJump > 1
1	80,4%	83,0%
2	75,4%	78,8%
3	71,8%	75,3%
4	70,1%	69,0%
5	66,2%	67,1%
6	60,7%	61,9%
7	68,7%	68,0%
8	75,0%	74,0%
9	81,2%	80,1%
10	88,6%	87,4%

Dado que ambos modelos obtuvieron resultados similares en cuanto a precisión, se ejecutaron en tiempos parecidos, y para los extremos se obtuvo un alto nivel de concordancia, se optó por XGBoost como el modelo ganador, ya que este superó en el indicador KS a las redes neuronales en ambos segmentos. A continuación, se muestran los resultados del análisis LIFT, detallando la precisión obtenida por decil para cada uno de los segmentos.

Tabla 5.6.6: Análisis LIFT por segmento

Decil	Precisión por Modelo XGBoost NumJump =1	Precisión aleatoria NumJump = 1	Precisión por Modelo XGBoost NumJump >1	Precisión aleatoria NumJump > 1
1	19,3%	10,3%	21,0%	13,0%
2	12,9%	10,3%	17,4%	13,0%
3	11,3%	10,3%	15,4%	13,0%
4	11,3%	10,3%	14,0%	13,0%
5	9,1%	10,3%	12,0%	13,0%
6	8,8%	10,3%	11,9%	13,0%
7	8,0%	10,3%	10,8%	13,0%
8	8,3%	10,3%	11,1%	13,0%
9	7,4%	10,3%	9,6%	13,0%
10	6,4%	10,3%	6,7%	13,0%

Como es posible notar en la tabla anterior, para ambos segmentos la probabilidad de que una persona entre el primer y el cuarto decil tome un crédito de consumo es mayor a la de elegir una persona al azar. Análogamente, la probabilidad de que una persona entre el quinto y el décimo decil tome un crédito de consumo es menor a la de elegir una persona aleatoriamente, lo que indica que efectivamente el modelo está discriminando.

5.7 Discusión de resultados

A modo de darle un significado más tangible a los resultados obtenidos, se realizó una comparación entre estos y los resultados que obtuvieron dos de los autores mencionados en la revisión bibliográfica.

En el caso del segmento NumJump = 1, al seleccionar al 40% de población mejor rankeada se logra abarcar al 53,3% de la población con respuestas positivas, mientras que para el caso del segmento NumJump > 1, al seleccionar el 40% de la población se abarca al 52,1% de la población con respuestas positivas. Al comparar estos resultados con los obtenidos por Hossein Javaheri (2008), quien logra seleccionar el 79% de las respuestas positivas con sólo el 40% de los clientes, resulta en una efectividad baja.

Si se repite la misma lógica, pero esta vez aplicada sobre el 50% de los clientes, en el caso de NumJump = 1 se puede separar al 62,2% de las respuestas positivas, mientras que en el caso de NumJump > 1, el 61,4%. Al hacer la misma comparación, pero esta vez con los resultados obtenidos por Moro, Cortez y Rita (2014), quienes fueron capaces de aislar al 79% de las respuestas positivas con el 50% de los clientes, también resulta en una baja efectividad.

En síntesis, tras comparar la efectividad obtenida por este modelo versus la que obtuvieron los autores de las dos publicaciones mencionadas previamente, resulta baja a primera vista. La diferencia del 17% en efectividad que se tiene con Moro, Cortez y Rita (2014) podría explicarse en gran medida debido a dos factores. En primer lugar, en que este modelo para discriminar, sólo utiliza información pública proveniente de una fuente: la SBIF; mientras que ellos se nutrieron de diversas fuentes de información pública como también de toda la información privada de dicha institución para generar la clasificación. En segundo lugar, hay que considerar la dificultad que agrega el hecho de que la información venga agregada por cliente y no por producto. En el caso descrito, los autores contaban con la información específica sobre depósitos a plazo, mientras que en el caso estudiado, la información viene agregada a nivel de deuda de consumo, lo que considera créditos de consumo, tarjetas de crédito, líneas de créditos, entre otros. Por otra parte, la diferencia en efectividad con Hossein Javaheri (2008) se puede explicar

considerando que, adicionalmente a lo mencionado para los otros autores, existe la diferencia de que él buscaba predecir el éxito de un campaña en particular, específica y delimitada, y no la venta de un producto, que puede ser comprado por necesidad del cliente, por promoción del banco o por cualquier otro factor, lo que reduce el problema en muchas dimensiones.

6. VALIDACIÓN DEL MODELO

Para validar el modelo propuesto, se realizó un *testing* en el que seleccionaron dos muestras. Para la primera, se eligieron 80.000 no clientes que el modelo haya etiquetado como pertenecientes a los primeros cinco deciles. Para la segunda muestra, se tomaron 3.000 no clientes que estuviesen exclusivamente en el décimo decil. A ambas muestras se les envió un mensaje de texto (SMS) idéntico indicando que el banco posee una oferta para ellos y que en caso de contestar "Sí" al SMS, serían contactados.

Al cabo de un mes desde que se enviaran los SMS se analizaron los resultados. Para el caso de la muestra propensa, de los 80.000 SMS enviados sólo 56.085 fueron efectivamente recibidos. De estos, se 1.540 personas respondieron, donde 771 respuestas se clasificaron como positivas (tasa de respuestas positivas sobre SMS efectivamente entregados: 1,37%). Por otra parte, para el caso de la muestra poco propensa, de los 3.000 SMS enviados, solo 2.005 fueron efectivamente recibidos. De estos, 21 personas respondieron, donde 12 de ellas se clasificaron como positivas (tasa de respuestas positivas sobre SMS efectivamente entregados: 0,59%).

Tabla 6.1.1: Respuestas del experimento de validación

Muestra	**Enviados**	**Recibidos**	**Respuestas totales**	**Respuestas positivas**	**Respuestas positivas por SMS recibido**
Propensa	80.000	56.085	1.540	771	1,37%
Poco propensa	3.000	2.005	21	12	0,59%

Si bien para ambos casos hubo respuestas positivas, al momento de contactar de vuelta a las personas que respondieron "Sí" —mediante un llamado telefónico realizado por los ejecutivos de la fuerza de venta— para efectuar la venta del crédito de consumo, la muestra seleccionada como propensa obtuvo un total de 44 operaciones cursadas,

mientras que la muestra poco propensa no cursó ninguna sola operación. A modo de contextualizar, 44 ventas sobre un total de 1.540 respuestas (positivas o negativas) entrega una efectividad del 2,86% sobre lo respondido, cifra que supera con creces al 0,69% histórico del banco en efectividad de campañas sobre no clientes.

Tabla 6.1.2: Ventas del experimento de validación

Muestra	**Respuestas totales**	**Respuestas positivas**	**Operaciones cursadas**	**Venta por contacto efectivo**
Propensa	1.540	771	44	2,86%
Poco propensa	21	12	0	0%

De esta manera se valida que si bien el modelo no tiene los niveles de efectividad que lograron otros autores al resolver problemas relativamente similares, este sí sirve para discriminar entre personas que tomarán o no un crédito de consumo y de esta manera hacer más eficiente el proceso de venta.

7. CONCLUSIONES

En la industria bancaria, así como en otras industrias, la constante presión por aumentar ingresos y disminuir costos hace que optimizar la selección de clientes sea un aspecto clave. Bajo este contexto es que se desarrolló un sistema de apoyo a la toma de decisiones, basado en información proveniente de la SBIF, que estima la probabilidad a nivel individual de que una persona adquiera un crédito de consumo en los próximos tres meses.

Siguiendo esta línea, el objetivo general del trabajo fue desarrollar y validar un modelo conceptual capaz de detectar las necesidades crediticias futuras de las personas del sistema financiero y así cuantificar la propensión a créditos de consumo. Para esto se plantearon dos hipótesis. En primer lugar, que es posible determinar la propensión crediticia de una persona a nivel de individuo utilizando sólo información financiera de carácter pública, y en segundo lugar, que esta metodología arrojará mejores resultados que el modelo actual.

Para lograr el objetivo se dio un énfasis particular a la detección de las operaciones de compra de créditos de consumo en la SBIF. Para esto se desarrolló el modelo de *Jumps*, el cual es capaz de identificar dichas operaciones con un error del 1,5%, valor considerablemente menor al 9,8% del modelo actual que opera bajo montos fijos por tramo de renta en vez de operar de manera dinámica a nivel de individuo.

Para poder predecir si ocurrirá una compra en los próximos tres meses, se hizo competir cinco modelos predictivos de aprendizaje automático: regresiones logísticas, bosques aleatorios, máquina de soporte vectorial, XGBoost y redes neuronales. Para compararlos se usaron tres métricas: su precisión en el gráfico LIFT, el tiempo de ejecución y la prueba de bondad de ajuste Kolmogorov-Smirnov. Los mejores resultados se obtuvieron con XGBoost, obteniendo un LIFT de 19,3% para el primer decil, en 10 minutos y con

un KS de 0,39 en el caso de NumJump = 1 y 21% en el primer decil de LIFT en 15 minutos con un KS de 0,25 para el caso de NumJump > 1.

La efectividad del modelo es menor a la obtenida por otros autores en experiencias anteriores, 53% versus 79% al comparar con Hossein Javaheri (2008) y 62% versus 79% al comparar con Moro, Cortez y Rita (2014), ambos bajos sus respectivas métricas. No obstante al poner en producción el modelo, se obtuvo una efectividad del 2,86% en ventas sobre respuestas, una cifra más de cuatro veces mayor a lo acostumbrado por el banco en campañas sobre prospectos (0,69%).

De esta manera, se comprueban ambas hipótesis, dado que es posible determinar la propensión crediticia de una persona a nivel de individuo utilizando sólo información financiera de carácter pública, y además, esta metodología arrojó mejores resultados que el modelo actual. Esta conclusión valida también el modelo conceptual planteado, ya que fue posible generar un modelo de propensión utilizando la interacción entre variables y la componente temporal para generar campos de RFM, los que son la base de la función predictora.

Queda latente la inquietud de mejorar la efectividad del modelo. Para esto se propone la incorporación de nueva data de carácter público, pero no necesariamente de la SBIF, como lo es, por ejemplo, información del Banco Central, del Instituto Nacional de Estadísticas, de distintas empresas de encuestas, entre otros. Siempre con el objetivo de que sirva para generar nuevas variables y así calibrar más aún los resultados obtenidos.

Finalmente, quedan propuestos futuros estudios que busquen ampliar el espectro de los modelos a distintos productos, como lo son los créditos comerciales o los créditos hipotecarios; distintos públicos objetivos, como lo son las micro, pequeñas, medianas y grandes empresas y también ampliar las funcionalidades, como lo sería por ejemplo, determinar el monto óptimo a ofertar a cada uno de los individuos seleccionados.

BIBLIOGRAFIA

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Belmont, Wadsworth.

Brown, S., Garino, G., Taylor, K., & Price, S. W. (2005). Debt and financial expectations: An individual -and household- level analysis. *Economic Inquiry*, *43*(1), 100–120. https://doi.org/10.1093/ei/cbi008

Camus, J. (2015). Acceso al crédito en la realidad actual de Chile. *Diario Financiero.* https://www.df.cl/noticias/opinion/columnistas/acceso-al-credito-en-la-realidad-actual-de-chile/2015-06-11/172926.html [Acceso Septiembre 23, 2017].

Coppock, D. S. (2002). Why lift? Data modeling and mining. *Information Management*. http://www.information-management.com/news/5329-1.html [Acceso Octubre 1, 2017].

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

Delen, D., Sharda, R., & Kumar, P. (2007). Movie forecast Guru: A Web-based DSS for Hollywood managers. *Decision Support Systems*, *43*(4), 1151–1170. https://doi.org/10.1016/j.dss.2005.07.005

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78-87. https://doi.org/10.1145/2347736.2347755

Falk, M., Marohn, F., Michel, R., Hofmann, D., Macke, M., Spachmann, C., Englert, S. (2012). A First Course on Time Series Analysis: Examples with SAS. Chair of Statistics, University of Würzburg. http://nbn-resolving.org/urn:nbn:de:bvb:20-opus-72617

Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189-1232. http://www.jstor.org/stable/2699986

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. New York, Springer series in statistics.

Granitto, P., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, *83*(2), 83–90. https://doi.org/10.1016/j.chemolab.2006.01.007

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157-1182. http://www.jmlr.org/papers/v3/guyon03a.html

Hand, D., & Henley, W. (1997). Statistical Classification Methods in Consumer Credit Scoring: a Review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *160*(3), 523-541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

Haykin, S. S. (2009). Neural networks and learning machines. Pearson Prentice Hall, 2009.

Hossein Javaheri, S. (2008). Response modeling in direct marketing : a data mining based approach for target selection (Dissertation). http://urn.kb.se/resolve?urn=urn:nbn:se:ltu:diva-50999

Jiang, T., & Tuzhilin, A. (2006). Segmenting customers from population to individuals: Does 1-to-1 keep your customers forever? *IEEE Transactions on Knowledge and Data Engineering*, *18*(10), 1297–1311. https://doi.org/10.1109/TKDE.2006.164

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal Of Statistical Software*, *36*(11), 1–13.

Lau, K., Chow, H., & Liu, C. (2004). A database approach to cross selling in the banking industry: Practices, strategies and challenges. *Journal of Database Marketing & Customer Strategy Management*, *11*(3), 216–234. https://doi.org/10.1057/palgrave.dbm.3240222

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

Martens, D., & Provost, F. (2011). Pseudo-social network targeting from consumer transaction data. *NYU Working Paper, No.* CEDER-11-05. http://archive.nyu.edu/handle/2451/31253

Moro, S., & Laureano, R. M. S. (2011). Using Data Mining for Bank Direct Marketing: An application of the CRISP-DM methodology. In *Proceedings of the 2011 European Simulation and Modelling Conference*, 117–121. http://hdl.handle.net/1822/14838

Moro, S., Laureano, R., & Cortez, P. (2012). Enhancing bank direct marketing through data mining. In *Proceedings of the 41th European Marketing Academy Conference [EMAC]*. European Marketing Academy. http://hdl.handle.net/1822/21409

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31. https://doi.org/10.1016/j.dss.2014.03.001

Murdoch, T. B., & Detsky, A. S. (2013). The Inevitable Application of Big Data to Health Care. *JAMA*, *309*(13), 1351- 1352. https://doi.org/10.1001/jama.2013.393

Norvilitis, J. M., Szablicki, P. B., & Wilson, S. D. (2003). Factors Influencing Levels of Credit-Card Debt in College Students. *Journal of Applied Social Psychology*, *33*(5), 935–947. https://doi.org/10.1111/j.1559-1816.2003.tb01932.x

Nun, I., Protopapas, P., Sim, B., Zhu, M., Dave, R., Castro, N., & Pichara, K. (2015). FATS: Feature Analysis for Time Series. In *arXiv preprint arXiv:1506.00010*. http://arxiv.org/abs/1506.00010

Parrado, E. (s.f) Carta de Presentación. *SBIF.cl.* http://www.sbif.cl/sbifweb/servlet/ConozcaSBIF?indice=7.5.1.1&idContenido=10001 [Acceso Julio 17, 2017].

Phillips, R. (2013). Optimizing prices for consumer credit. *Journal of Revenue and Pricing Management*, *12*(4), 360–377. https://doi.org/10.1057/rpm.2013.9

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Superintendencia de Bancos e Instituciones Financieras de Chile [SBIF] (2017). Sistema Financiero de Chile. *SBIF.cl.* http://www.sbif.cl/sbifweb/servlet/ConozcaSBIF?indice=7.5.1.1&idContenido=477 [Acceso Octubre 1, 2017].

Talla Nobibon, F., Leus, R., & Spieksma, F. (2011). Optimization models for targeted offers in direct marketing: Exact and heuristic algorithms. *European Journal of Operational Research*, *210*(3), 670–683. https://doi.org/10.1016/j.ejor.2010.10.019

Tsai, C., & Wu, J. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, *34*(4), 2639–2649. https://doi.org/10.1016/j.eswa.2007.05.019

Venables, W.N. & Ripley, B.D, (2002), *Modern Applied Statistics with S*. New York, Springer-Verlag.

Volkov, A., Benoit, D. F., & Van den Poel, D. (2017). Incorporating sequential information in bankruptcy prediction with predictors based on Markov for discrimination. *Decision Support Systems*, *98*, 59–68. https://doi.org/10.1016/j.dss.2017.04.008

A N E X O S

ANEXO A: RESUMEN DESCRIPTIVO DE CADA CAMPO DE LA BASE DE DATOS DE LA SBIF

- **AnoMes**: año y mes de consulta en formato "201601".
- **Banco**: binaria que indica si el cliente está bancarizado o no. Como sólo interesa estudiar los clientes con deuda, se tomarán los clientes con "Banco" = 1.
- **RutNumerico**: número de identificación nacional. Entero largo 10.
- **Dígito**: corresponde al dígito verificador del RUT. Puede tomar valores del 0 al 9 y K.
- **RazonSocial**: corresponde al nombre asociado a un RUT.
- **M_DeudaDirectaVigente**: entero positivo. 5% de los valores son 0, ninguno de los demás supera el 1%.
- **M_DeudaDiretaMorosa**: entero positivo. 92% de los valores son 0, ninguno de los demás supera el 0,08%.
- **M_DeudaDirectaVencida**: entero positivo. 95% de los valores son 0, ninguno de los demás supera el 0,02%.
- **M_DeudaDirectaiFinanciera**: todos los valores son 0.
- **M_DeudaDirectaOperacionesPactadas**: entero positivo. Sólo hay dos valores que no son 0.
- **M_DeudaIndirectaVigente**: entero positivo. 95% de los valores son 0, ninguno de los demás supera el 0,03%.
- **M_DeudaIndirectaVencida**: entero positivo. 99,5% de los valores son 0, ninguno de los demás supera el 0,0009%.
- **M_DeudaComercial**: entero positivo. 80% de los valores son 0, ninguno de los demás supera el 0,06%.
- **M_DeudaCréditoConsumo**: entero positivo. 18% de los valores son 0, ninguno de los demás supera el 0,3%.
- **N_InstitucionesConDeuda**: entero que toma valores entre 0 y 12, concentrándose el 99% en el rango 0 a 4.
- **M_CreditoHipotecario**: entero positivo. 80% de los valores son 0, ninguno de los demás supera el 0,002%.
- **M_CastigosDirectos**: entero positivo. 98% de los valores son 0, ninguno de los demás supera el 0,002%.
- **M_CastigosIndirectos**: entero positivo. 99,9% de los valores son 0, ninguno de los demás supera el 0,0001%.
- **M_CupoLíneaCrédito**: entero positivo. 28% de los valores son 0, ninguno de los demás supera el 1,4%.
- **M_DeudaComercialVigente**: entero positivo. 93,7% de los valores son 0, ninguno de los demás supera el 1,1%.
- **M_DeudaComercialVencida**: entero positivo. 95,9% de los valores son 0, ninguno de los demás supera el 0,04%.
- **M_DeudaCréditoComerciales**: entero positivo. 80,7% de los valores son 0, ninguno de los demás supera el 18,1%.
- **M_DeudaLeasing**: entero positivo. 99,9% de los valores son 0, ninguno de los demás supera el 0,00003%.
- **M_DeudaMorosaLeasing**: entero positivo. 99,9% de los valores son 0, ninguno de los demás supera el 0,00008%.
- **M_CredDiralDia**: todos los valores son 0.

- **M_CredDirlmp30**: todos los valores son 0.
- **M_CredDirlmp90**: todos los valores son 0.
- **M_OpeFin**: todos los valores son 0.
- **M_InstDeuAdq**: todos los valores son 0.
- **M_CredIndalDia30**: todos los valores son 0.
- **M_CredIndImp**: todos los valores son 0.
- **M_CredCom**: todos los valores son 0.
- **M_CredCon**: todos los valores son 0.
- **M_NInstCredCon**: todos los valores son 0.
- **M_CredViv**: todos los valores son 0.
- **M_CredDirImp3A**: todos los valores son 0.
- **M_CredIndImp3A**: todos los valores son 0.
- **M_MtoLineaCred**: todos los valores son 0.
- **M_CredCont**: todos los valores son 0
- **M_CredImp90**: todos los valores son 0.
- **M_NCredComer**: todos los valores son 0.
- **M_CredLeasDia**: todos los valores son 0.
- **M_CredLeasImp**: todos los valores son 0.

ANEXO B: VARIABLES CREADAS

1 **Edad**

2 **NumJumps:**
Número de saltos en la ventana de observación.

3 **AvgJumpDuration:**
Promedio de meses entre saltos.

4 **JumpDuration_between_Last2_and_Last1:**
El número de meses entre el último salto y el penúltimo salto.

5 **Last2JumpAmount**
El cambio en deudacreditoconsumo en el penúltimo salto.

6 **LastJump_close_to_500**
El cambio en deudacreditoconsumo dividido por 500 en el último salto

7 **Last1JumpAmountB**
El cambio en deudacreditoconsumo en el último salto.

8 **avg_TotalChangeAfterTheJump_v_TheJump**
Calcula el cambio total en deudacreditoconsumo comenzando desde el salto hasta el siguiente salto y se divide en el NumJumps menos 1.

9 **TotalChangeAfterLast1Jump**
Calcula el cambio total en deudacreditoconsumo desde el último salto hasta el final de la ventana de observación.

10 **TotalChangeAfterLast1Jump_v_TheJump**
Calcula el cambio total en deudacreditoconsumo comenzando desde el último salto hasta el final de la ventana de observación dividido por Last1JumpAmountB.

11 **TotalNumJumpsUp**
Número total de saltos que sigue una tendencia de aumento en deudacreditoconsumo.

12 **PropJumpsUp**
TotalNumJumpsUp dividido por NumJumps.

13 **Recencia**
El número de meses entre el último salto al final de la ventana de observación.

14 - 20 **diff_deudacreditoconsumo_sumXm_on_last_jump**
El cambio total en deudacreditoconsumo antes del último salto, dentro de los últimos 3, 6, 9, 12, 24, 36, 48 meses.

21 - 27 **diff_deudacreditoconsumo_Xm_max_on_last_jump**
El máximo diff_deudacreditoconsumo antes del último salto, dentro de los últimos 3, 6, 9, 12, 24, 36, 48 meses.

28 - 34 **diff_deudacreditoconsumo_Xm_min_on_last_jump**
El mínimo de diff_deudacreditoconsumo antes del último salto, dentro de los últimos 3, 6, 9, 12, 24, 36, 48 meses.

35 - 41 **diff_deudacreditoconsumo_Xm_mean_on_last_jump**
El promedio diff_deudacreditoconsumo antes del último salto, dentro del último salto más reciente 3, 6, 9, 12, 24, 36, 48 meses.

42 - 44 **diff_deudacreditoconsumo_sumXm_v_3m_on_last_jump**
diff_deudacreditoconsumo_sumXm_on_last_jump dividido por diff_deudacreditoconsumo_sum3m_on_last_jump para 6, 9 y 12 meses.

45 **diff_deudacreditoconsumo_sum12m_v_6m_on_last_jump**

46 **diff_deudacreditoconsumo_sum24m_v_12m_on_last_jump**

47 - 53	**diff_cupolineacredito_sumXm_on_last_jump** El cambio total en cupolineacredito antes los 3, 6, 9, 12, 24, 36, 48 meses antes del último salto.
54 - 60	**diff_cupolineacredito_Xm_max_on_last_jump** El máximo diff_cupolineacredito antes del último salto, dentro de los últimos 3, 6, 9, 12, 24, 36, 48 meses.
61 - 67	**diff_cupolineacredito_Xm_min_on_last_jump** El mínimo diff_cupolineacredito antes del último salto, en los últimos 3, 6, 9, 12, 24, 36, 48 meses.
68 - 74	**diff_cupolineacredito_Xm_mean_on_last_jump** El promedio de diff_cupolineacredito antes del último salto, dentro de los últimos 3, 6, 9, 12, 24, 36, 48 meses.
75 - 77	**diff_cupolineacredito_sumXm_v_3m_on_last_jump** diff_cupolineacredito_sumXm_on_last_jump dividido por diff_cupolineacredito_sum3m_on_last_jump para 6, 9 y 12 meses.
78	**diff_cupolineacredito_sum12m_v_6m_on_last_jump**
79	**diff_cupolineacredito_sum24m_v_12m_on_last_jump**
80 - 83	**diff_deudacreditoconsumo_sumXm_after_last_jump** La suma de diff_deudacreditoconsumo para los 3, 6, 9 y 12 meses después del último salto.
84 - 87	**diff_deudacreditoconsumo_sumXm_after_last_jump_perc** La suma de diff_deudacreditoconsumo para los 3, 6, 9 y 12 meses después del último salto dividido por la cantidad del último salto.
88 - 91	**diff_deudacreditoconsumo_Xm_max_after_last_jump** El máximo diff_deudacreditoconsumo después del último salto, dentro de los 3, 6, 9 y 12 meses próximos al salto.
92 - 95	**diff_deudacreditoconsumo_Xm_min_after_last_jump** El mínimo diff_deudacreditoconsumo después del último salto, dentro de los 3, 6, 9 y 12 meses próximos al salto.
96 - 99	**diff_deudacreditoconsumo_Xm_mean_after_last_jump** La media diff_deudacreditoconsumo después del último salto, dentro de los 3, 6, 9 y 12 meses próximos al salto.
100 - 102	**diff_deudacreditoconsumo_sumXm_v_3m_after_last_jump** La suma de diff_deudacreditoconsumo para los 6, 9 y 12 meses después del último salto dividido en la misma suma para los 3 meses después del salto.
103	**diff_deudacreditoconsumo_sum12m_v_6m_after_last_jump**
104 - 107	**diff_cupolineacredito_sumXm_after_last_jump** La suma de las diferencias en el cupolineacredito los 3, 6, 9 y 12 meses después del último salto.
108 - 111	**diff_cupolineacredito_Xm_max_after_last_jump** El máximo diff_cupolineacredito en los 3, 6, 9 y 12 meses después del último salto.
112 - 115	**diff_cupolineacredito_Xm_min_after_last_jump** El mínimo diff_cupolineacredito en los 3, 6, 9 y 12 meses después del último salto.
116 - 119	**diff_cupolineacredito_Xm_mean_after_last_jump** El promedio de diff_cupolineacredito en los 3, 6, 9 y 12 meses después del último salto.

120 - 122	**diff_cupolineacredito_sumXm_v_3m_after_last_jump** La suma de diff_cupolineacredito para los 6, 9 y 12 meses después del último salto dividido en la misma suma para los 3 meses después del salto.
123	**diff_cupolineacredito_sum12m_v_6m_after_last_jump**
124 - 126	**m_CupoTotal_Xavg** Promedio anual de CupoTotal que es la suma de cupolineacredito y deudacreditoconsumo para los últimos 3 años.
127	**m_CupoTotal_(X)-(X-1)avg** Diferencia entre los promedios de CupoTotal para los últimos dos años.
128 - 130	**m_cupolineacredito_Xavg** Promedio anual de m_cupolineacredito para los últimos 3 años.
131	**m_cupolineacredito_(X)-(X-1)avg** Diferencia entre los promedios de m_cupolineacredito para los últimos dos años.
132 - 134	**n_institucionescondeuda_Xavg** Promedio anual de n_institucionescondeuda para los últimos 3 años.
135	**n_institucionescondeuda_(X)-(X-1)avg** Diferencia entre los promedios de n_institucionescondeuda para los últimos dos años.
136 - 138	**m_creditohipotecario_Xavg** Promedio anual de m_creditohipotecario para los últimos 3 años.
139	**m_creditohipotecario_(X)-(X-1)avg** Diferencia entre los promedios de m_creditohipotecario para los últimos dos años.
140 - 142	**m_deudacreditoconsumo_Xavg** Promedio anual de deudacreditoconsumo para los últimos 3 años.
143	**m_deudacreditoconsumo_(X)-(X-1)avg** Diferencia entre los promedios de m_deudacreditoconsumo para los últimos dos años.
144 - 146	**m_deudacreditoconsumo_Xmed** La mediana de deudacreditoconsumo para cada uno de los últimos 3 años.
147	**m_deudacreditoconsumo_(X)-(X-1)med** Diferencia entre las medianas de m_deudacreditoconsumo para los últimos dos años.
148 - 150	**diff_m_deudacreditoconsumo_Xavg** Promedio anual de diff_deudacreditoconsumo para cada uno de los últimos 3 años.
151	**diff_m_deudacreditoconsumo_(X)-(X-1)avg** Diferencia entre los promedios de diff_m_deudacreditoconsumo de los últimos dos años.
152 -154	**diff_m_deudacreditoconsumo_Xmed** Mediana de diff_deudacreditoconsumo para cada uno de los últimos 3 años.
155	**diff_m_deudacreditoconsumo_(X)-(X-1)med** Diferencia entre las medianas de diff_m_deudacreditoconsumo de los últimos dos años.
156 - 158	**diff_m_deudacreditoconsumo_Xsd** Desviación estándar de diff_deudacreditoconsumo para cada uno de los últimos 3 años.

159	**diff_m_deudacreditoconsumo_(X)-(X-1)sd** Diferencia entre las desviaciones estándar de diff_m_deudacreditoconsumo de los últimos dos años.
160 - 171	**deudacreditoconsumo_X** deudacreditoconsumo, 1 mes antes, 2 meses antes, ..., 12 meses antes.
172 - 178	**diff_deudacreditoconsumo_sumXm** El cambio total en deudacreditoconsumo de los últimos 3, 6, 9, 12, ..., 48 meses.
179 - 185	**diff_deudacreditoconsumo_maxXm** El máximo de diff_deudacreditoconsumo de los últimos 3, 6, 9, 12, ..., 48 meses.
186 - 192	**diff_deudacreditoconsumo_minXm** El mínimo de diff_deudacreditoconsumo de los últimos 3, 6, 9, 12, ..., 48 meses.
193 - 194	**diff_deudacreditoconsumo_medXm** La mediana diff_deudacreditoconsumo de los últimos 3, 6, 9, 12, ..., 48 meses.
200 - 206	**diff_deudacreditoconsumo_sdXm** Desviación estándar de diff_deudacreditoconsumo de los últimos 3, 6, 9, 12, ..., 48 meses.
207 - 213	**diff_deudacreditoconsumo_meanXm** La media de diff_deudacreditoconsumo de los últimos 3, 6, 9, 12, ..., 48 meses.
214 - 220	**deudacreditoconsumo_sumXm** La suma de deudacreditoconsumo de los últimos 3, 6, 9, 12, ..., 48 meses.
221 - 227	**deudacreditoconsumo_maxXm** El máximo deudacreditoconsumo de los últimos 3, 6, 9, 12, ..., 48 meses.
228 - 234	**deudacreditoconsumo_minXm** El mínimo deudacreditoconsumo de los últimos 3, 6, 9, 12, ..., 48 meses.
235 - 241	**deudacreditoconsumo_meanXm** La media deudacreditoconsumo de los últimos 3, 6, 9, 12, ..., 48 meses.
242 - 248	**diff_CupoTotal_sumXm** El cambio total en diff_CupoTotal más reciente 3, 6, 9, 12, ..., 48 meses.
249 - 255	**diff_CupoTotal_maxXm** El máximo diff_CupoTotal de los últimos 3, 6, 9, 12, ..., 48 meses.
256 - 262	**diff_CupoTotal_minXm** El mínimo diff_CupoTotal de los últimos 3, 6, 9, 12, ..., 48 meses.
263 - 269	**diff_CupoTotal_medXm** La mediana diff_CupoTotal de los últimos 3, 6, 9, 12, ..., 48 meses.
270 - 276	**diff_CupoTotal_sdXm** La desviación estándar diff_CupoTotal de los últimos 3, 6, 9, 12, ..., 48 meses.
277 - 283	**diff_CupoTotal_meanXm** La media diff_CupoTotal de los últimos 3, 6, 9, 12, ..., 48 meses.
284 - 287	**deudacreditoconsumo_sumXm_v_1m** La suma de deudacreditoconsumo de los últimos 3, 6, 9 y 12 meses dividido en deudacreditoconsumo del último mes.
288 - 291	**deudacreditoconsumo_sumXm_d_1m** La suma de deudacreditoconsumo de los últimos 3, 6, 9 y 12 meses menos deudacreditoconsumo del último mes.
292 - 294	**deudacreditoconsumo_sum6m_v_3m** La suma de deudacreditoconsumo de los últimos 6, 9 y 12 meses dividido en la suma de deudacreditoconsumo de los últimos 3 meses.

295 **deudacreditoconsumo_sum12m_v_6m**
La suma de deudacreditoconsumo de los últimos 12 meses dividido en la suma de deudacreditoconsumo de los últimos 6 meses.

296 **deudacreditoconsumo_sum24m_v_12m**
La suma de deudacreditoconsumo de los últimos 24 meses dividido en la suma de deudacreditoconsumo de los últimos 12 meses.

297 - 349 **160 a 171, 214 a 241 y 284 a 296 normalizadas por número de instituciones (mismo nombre, pero con sufijo "_ninst")**

350 - 352 **direct_v_indirect_Xavg**
Promedio anual de la deuda directa dividida por la deuda indirecta para los últimos 3 años.

353 - 355 **direct_d_indirect_Xavg**
Promedio anual de la deuda directa menos la deuda indirecta para los últimos 3 años.

356 - 358 **consumer_utilization_Xavg**
Promedio anual de deudacreditoconsumo / deudadirectavigente para los últimos 3 años.

359 - 452 **160 a 241 y 284 a 296 pero para cupolineacredito y diff_cupolineacredito.**

453 - 455 **n_institucionescondeuda_Xmax**
El máximo de n_institucionescondeuda durante el año para cada uno de los últimos 3 años.

456 **n_institucionescondeuda_(X)-(X-1)max**
Diferencia entre n_institucionescondeuda del último y penúltimo año.

457 - 461 **debt_group_both_Xmavg**
El porcentaje de meses durante los últimos 3, 6, 9, 12, 24 meses donde el cliente pertenece a ambos grupos de deuda ('Directa' e 'Indirecta').

462 - 466 **debt_group_direct_only_Xmavg**
El porcentaje de meses durante los últimos 3, 6, 9, 12, 24 meses donde el cliente pertenece al grupos de deuda 'Directa'.

467 - 471 **debt_group_neither_Xmavg**
El porcentaje de meses durante los últimos 3, 6, 9, 12, 24 meses donde el cliente no pertenece a ninguno de los grupos de deuda.

472 - 476 **debt_group_indirect_only_Xmavg**
El porcentaje de meses durante los últimos 3, 6, 9, 12, 24 meses donde el cliente pertenece sólo al grupo de deuda indirecta.

477 - 481 **debt_type_all_Xmavg**
El porcentaje de meses durante los últimos 3, 6, 9, 12, 24 meses donde el cliente posee todo tipo de deuda ('Consumo', 'Hipotecaria' y 'Comercial').

482 - 486 **debt_type_cs_and_m_Xmavg**
El porcentaje de meses durante los últimos 3, 6, 9, 12, 24 meses donde el cliente posee deuda de "Consumo" e "Hipotecaria".

487 - 491 **debt_type_cm_and_m_Xmavg**
El porcentaje de meses durante los últimos 3, 6, 9, 12, 24 meses donde el cliente posee deuda "Comercial" e "Hipotecaria".

492 - 496 **debt_type_cm_and_cs_Xmavg**
El porcentaje de meses durante los últimos 3, 6, 9, 12, 24 meses donde el cliente posee deuda "Comercial" y de "Consumo".

497 - 501 **debt_type_cs_Xmavg**

El porcentaje de meses durante los últimos 3, 6, 9, 12, 24 meses donde el cliente posee sólo deuda de ‘Consumo’.

502 - 506 **debt_type_m_Xmavg**

El porcentaje de meses durante los últimos 3, 6, 9, 12, 24 meses donde el cliente posee sólo deuda ‘Hipotecaria’.

507 - 511 **debt_type_cm_Xmavg**

El porcentaje de meses durante los últimos 3, 6, 9, 12, 24 meses donde el cliente posee sólo deuda ‘Comercial’.

512 - 516 **debt_type_none_Xmavg**

El porcentaje de meses durante los últimos 3, 6, 9, 12, 24 meses donde el cliente no posee ninguno de los tipos.

517 - 521 **diff_deudacreditoconsumo_flag_Xmavg**

El porcentaje de meses durante los últimos 3, 6, 9, 12, 24 meses donde el diff_deudacreditoconsumo del cliente es positivo.

522 - 526 **m_creditohipotecario_flag_Xmavg**

El porcentaje de meses durante los últimos 3, 6, 9, 12, 24 meses donde el creditohipotecario del cliente es positivo.

527 - 531 **m_cupolineacredito_flag_Xmavg**

El porcentaje de meses durante los últimos 3, 6, 9, 12, 24 meses donde el cupolineacredito del cliente es positivo.

ANEXO C: PARÁMETROS DE INICIALIZACIÓN DE MODELOS.

1. Regresión logística:
    1. alpha = (0, 0.1, … , 0.6) donde alpha = 1 corresponde a la regresión Lasso y 0 a Ridge
    2. lambda = (0, 0.03, … , 0.3) que controla de manera general la fuerza de la penalidad.

2. Bosques aleatorios:
    1. max_depth: máxima profundidad del árbol [3, 5].
    2. max_features: número de variables a considerar para generar la próxima división [‘auto’, 10].
    3. min_sample_split: mínimo de muestras requeridas para dividir un nodo [3, 5, 10].
    4. min_samples_leaf: número mínimo de registros para ser considerado un nodo hoja [3, 5, 10].
    5. criterion: función para medir la calidad de la división. Los criterios disponibles son “gini” para la impureza y “entropy” para la ganancia de información [‘gini’, ’entropy’].

3. Máquina de soporte vectorial (SVM):
    1. Se calibró usando la escala de Platt

4. XGBoost:
    1. Eta: contracción de tamaño de paso utilizada en la actualización para evitar el sobreajuste (0.01, 0.05, 0.1).
    2. max_depth: máxima profundidad del árbol. Incrementar este valor hace el modelo más complejo y aumenta las probabilidades de *overfitting* (4, 6, 8).
    3. nrounds: número de rondas de *boosting* (50, 100, 200).

5. Redes neuronales:
   La arquitectura de una capa no fue capaz de discriminar y siempre predijo la misma clase. La de dos capas tuvo un mal desempeño en comparación con los otros modelos, por lo que se añadió una tercera capa de *Dropout* para evitar el sobreajuste y en esta ocasión sí se obtuvo buenos resultados. La adición de más capas no mejoró los resultados, por lo que se seleccionó la arquitectura de tres capas.