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

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

TRABAJO FINAL DE ESPECIALIZACIÓN

PREDICCIÓN DEL RIESGO DE DEFAULT DE CRÉDITOS DE HOGAR

UN ANÁLISIS DESDE MACHINE LEARNING INCORPORANDO VARIABLES ALTERNATIVAS

AUTORA: MARTINA BAJO

DICIEMBRE 2019

Estructura

Introducción

La revolución tecnológica en materia de datos e inteligencia artificial de los últimos años impulsa la transformación digital de las organizaciones. En este proceso, las entidades que manejan grandes volúmenes de información, cuentan con una oportunidad competitiva para la toma de decisiones sustentadas por datos. Este es el caso de bancos y entidades financieras, los cuales poseen registros de todas las transacciones, movimientos y operaciones de sus clientes, que son utilizados por varios sectores de las organizaciones con distintos fines.

Una de las aplicaciones comúnmente utilizadas, ligada al otorgamiento de créditos son los modelos de score: un sistema automatizado el cual parte de un set de información del cliente que se procesa por un algoritmo previamente definido con el fin de calificar al cliente en términos de su cumplimiento y capacidad de pago. En algunos casos, los resultados de los modelos pueden ser más bien informativos o para gestión interna, pero en otros son de carácter decisivo. Es por esto que la definición del input necesario para ese algoritmo junto con las reglas de decisión que tenga inmersas son cruciales, ya que condicionarán el otorgamiento de créditos.

Actualmente los modelos de score para la estimación del riesgo de default en créditos de hogar, habitualmente utilizan como input principal información crediticia histórica de los solicitantes de crédito bajo la premisa que el comportamiento pasado es suficiente para predecir la probabilidad de impago futura. Sin embargo, estos modelos son restrictivos para personas con poco historial crediticio o fuera del sistema bancario. Es por esto que Home Credit Group[1] pone a disposición una amplia variedad de datos, (no estrictamente financieros) con el fin de que la incorporación de éstos a un modelo de Machine Learning permita que personas con capacidad de pago no queden excluidas del sistema crediticio.

De aquí surge la siguiente problemática a resolver en este trabajo:

¿Cómo mejorar el acceso al crédito de personas con poco historial crediticio y bancario pero con capacidad de pago, mediante un modelo de Machine Learning?

1 Home Credit Group: Institución financiera no-bancaria de origen Checo. Opera en 10 países europeos y está orientada a ofrecer préstamos a personas con escaso historial crediticio.

Para resolver esta problemática se buscará proponer un modelo predictivo eficaz incluyendo variables alternativas de manera tal que personas con pocos registros bancarios y financieros pero con capacidad de pago puedan acceder a un crédito de hogar. Para lograrlo se abordará este trabajo desde tres puntos:

- ✓ Analizar, depurar y validar la información provista, crear nuevas variables que se consideren pertinentes.
- ✓ Reducir la dimensionalidad de los datos, ya que se está trabajando con grandes volúmenes de información
- ✓ Encontrar un modelo que incluya variables Validar que este modelo tenga impacto en el potencial otorgamiento de créditos, especialmente en personas con capacidad de pago que hubieran sido rechazadas en un modelo con variables exclusivamente financieras.

En este proceso se buscará comprobar la siguiente hipótesis:

La incorporación de variables alternativas a un modelo de riesgo de default, mejora el rendimiento del modelo en términos de la precisión en la predicción en un 5%.

Para verificar esta hipótesis se estructurará el trabajo de la siguiente manera:

En el primer apartado, se definirán los conceptos claves que permitirán comprender este trabajo, y se presentará el set de datos a trabajar. En el segundo apartado, se mostrarán dos técnicas utilizadas para la reducción de dimensionalidad de datos. En el tercer apartado se mostrarán los modelos elegidos, comparando los resultados obtenidos con un modelo con variables financieras y otro con variables financieras y alternativas. Se tomará como parámetro el resultado de un modelo con variables tradicionales a fin de no arribar a un algoritmo que genere antiselección ni que provoque mayor riesgo para la entidad Por último, se obtendrán conclusiones como cierre en el apartado final.

1. Sistema crediticio y Machine Learning

En este apartado se abordarán algunos conceptos iniciales necesarios para poder comprender el enfoque del trabajo. En primer lugar, se contextualizará este trabajo en un marco teórico, que conducirá a la definición de los términos clave de este trabajo. En segundo lugar, se reflexionará acerca de la importancia de incluir variables alternativas a un modelo de scoring. Finalmente, se procederá a hacer una presentación general de la información a utilizar.

1.1. Scoring bancario en créditos de hogar

En los últimos años, bancos y entidades financieras han desarrollado sistemas sofisticados para modelar el riesgo a los que se enfrentan en sus negocios. Este incentivo proviene no sólo de la intención de gestionar, cuantificar y desagregar el tipo y magnitud de riesgo que poseen, sino que también está impulsado por las reglamentaciones que impone el marco regulatorio actual. Sin ahondar profundamente en este tema, es pertinente hacer una breve descripción del acuerdo de Basilea[2] ya que juega un rol fundamental en este contexto porque en él se plantea la probabilidad de default que es el objeto de estudio de este trabajo.

El acuerdo de Basilea incentiva la medición formal de riesgos, con el objetivo de construir una base sólida para la regulación prudente del capital, la supervisión y la disciplina de mercado, así como perfeccionar la gestión del riesgo y la estabilidad financiera (Basle Comite on Bank Supervision). Uno de los tres pilares de esta regulación está orientado a la estimación del requerimiento mínimo de capital cuyo componente principal es el riesgo de crédito. La pérdida esperada (Expected Loss) por este riesgo se calcula como:

$$EL = PD \; x \; LGD \; x \; EAD$$

PD=Probabilidad de default

LGD= severidad / pérdida dado el incumplimiento

EAD= exposición en el momento del incumplimiento

[2] Acuerdo de supervisión bancaria o recomendaciones sobre regulación bancaria emitidos por el Comité de Basilea de Supervisión Bancaria.

La PD se define como la probabilidad de que el deudor no abone la totalidad de sus obligaciones crediticias, o bien el deudor se encuentre en situación de mora durante más de una determinada cantidad de días con respecto a cualquier obligación crediticia frente al grupo bancario (Bluhm, C., Overbeck, L., & Wagner, C). Según el acuerdo deberá estimarse la perdida esperada para cada una de las operaciones de la entidad.

En el marco de este trabajo la operación relevante son los créditos hipotecarios para adquisición de viviendas. Los créditos de hogar son un producto financiero a través del cual una entidad permite disponer de la cantidad necesaria para comprar una vivienda. Las entidades de crédito exigen una garantía antes de conceder un préstamo. En el caso de los hipotecarios, el titular del préstamo pone de garantía (hipoteca) el propio inmueble, que pasará a la entidad financiera en caso de impago.

Más allá de las garantías, las entidades deben asegurarse la capacidad de pago de sus clientes. Para evaluar el riesgo de crédito a nivel individual, se realizan técnicas de score, las cuales se definen como métodos estadísticos utilizados para clasificar a los solicitantes de crédito, o incluso a quienes ya son clientes de la entidad evaluadora, entre las clases de riesgo 'bueno' y 'malo'" (Hand y Henley)

A través de un algoritmo se calcula la PD de cada deudor y se le asigna una calificación o score (una PD reescalada) la cual es un indicador del nivel de riesgo del cliente. A este score se le determina un corte el cual se utilizará como criterio para otorgar o denegar un préstamo. El input y el algoritmo que definan, serán determinantes.

En este trabajo se busca calcular dicha la probabilidad de default a través de un modelo de Machine Learning, con un enfoque alternativo en términos de los datos que se utilizan como input. De esta manera, se busca ampliar la perspectiva con la que se aborda esta temática bajo la premisa de que el costo de la mala estimación de la probabilidad de default puede ser muy alto (Khandani, A. E., Kim, A. J., & Lo, A. W.): desde el punto de vista del negocio una mala clasificación (sobreestimación de capacidad de pago) implica una pérdida económica, desde el punto de vista del cliente (subestimación de su capacidad de pago), implica el rechazo de la solicitud de un crédito y en consecuencia, el impedimento de adquirir un hogar.

Es por esto que se partirá de un set de variables que habitualmente se emplean para scoring vinculadas a comportamiento frente a productos financieros solicitados a Home Credit o

cualquier otra entidad. Luego, se incluirán otro tipo de variables tales como: características de la vivienda actual donde reside el solicitante, tipo de contacto que registró, datos relevados al momento de aplicación, datos del cliente, bienes que posee, entre otros.
La enumeración y los detalles de las variables se darán a conocer en el último subapartado de este capítulo.

1.2. Importancia de la inclusión de otro tipo de variables

El scoring crediticio es una de las principales barreras que debe sortear una persona que desea solicitar un crédito hipotecario para adquirir una vivienda. Para determinar este algoritmo se ha encontrado que la información crediticia histórica de cada cliente es suficientemente predictivo, medido en términos del nivel de precisión exigido en el modelo, es decir; se encuentra que esa información ajusta bien porque predice con exactitud que un cliente entre o no en default. Este análisis ocasiona que los resultados se orienten sobre un subconjunto de clientes, que son más rentables y que posiblemente tengan más respaldo financiero.

Es por esto, que puede implicar que un segmento de la población, quede sucesivamente excluido por un factor o hecho particular que haya marcado su historial crediticio perpetuando su condición y aún acentuándola. ¿Qué sucedería con un cliente que tuvo un problema grave de salud que le impidió cumplir con sus obligaciones en el pasado? ¿y si un solicitante recientemente se incorporó al sistema bancario formal y no hay tiene registros pasados para ser evaluados? O incluso, si un barrio se asocia con una determinada característica que está ligada a un mal comportamiento, ¿no se lo estaría "condenando" y excluyéndolo del sistema acentuando una condición marginal?

Esta problemática que se presenta en muchos sectores en la era de Big Data puede tratarse desde distintos enfoques (Big Data and Discrimination Talia B. Gillis† & Jann L. Spiess):

I. Desde el input: se tiene principal interés en la elección del tipo de información que se incluye en el modelo bajo la premisa que los datos que se incluyen son los que pueden causar discriminación (ej. raza, religión, color).

II. Desde el algoritmo: se busca focalizar en las decisiones que se toman en el armado de los modelos y en la supervisión de estados parciales que arroja el mismo, bajo la premisa de que un algoritmo en sí no discrimina, sino que influyen determinadas consideraciones y omisiones que puedan ocurrir en el proceso.

III. Desde los resultados: se fundamenta que la discriminación puede ocasionarse en la incorrecta interpretación de resultados obtenidos. Se propone también efectuar un test a posteriori para comprobar que no haya discriminación en los resultados.

Si bien se tomarán algunas recomendaciones u observaciones del segundo y tercer enfoque, se trabajará principalmente sobre el primero. Se considera que la problemática planteada puede ser abordada directamente incrementando el espectro de variables que se utilizan con el fin de que agreguen otro tipo de información para lograr la inclusión financiera de clientes con capacidad de pago.

Asimismo, se analizará si es pertinente excluir algunas variables que puedan ser consideradas discriminatorias; con respecto a este punto, cabe destacar que según los autores Gillis & Spiess debe examinarse con detenimiento cada variable, ya que pueden ser discriminatorias de forma indirecta; por ej. la religión es un dato sensible que suele excluirse, pero el dato del barrio donde vive el cliente puede estar altamente correlacionado con personas determinada religión, y de esta manera con esa variable se podría estar efectuando una selección indeseada. En este sentido, tanto las variables "religión" como "barrio" deberían ser excluidas del modelo. Este criterio puede contener un cierto grado de subjetividad inmersa, ya que en un extremo muchos de los datos pueden considerarse discriminatorios; quedarán sujetos al criterio del analista. Habiendo enmarcado teóricamente la temática, se procede a caracterizar los datos con los que se desarrollará este trabajo.

1.3. Variables financieras vs. Alternativas

El set de datos a utilizar fue provisto por la entidad "Home Credit" a través de "Kaggle", un sitio web orientado a la comunidad de analistas, científicos de datos y machine learners con el espíritu de promover e incentivar el aprendizaje y uso de técnicas de machine learning a través de competencias (habitualmente premiadas). Para cumplir con la privacidad de datos, el sitio exige algunas condiciones; en el caso de esta competencia, está vinculada con la

anonimizarían de datos, es evitar que una persona pueda ser identificada por los datos publicados. Es importante destacar que Home Credit es una institución financiera no bancaria que se caracteriza por otorgar créditos a personas que habitualmente serían rechazadas en otras entidades. Se orienta a personas con escaso historial crediticio, justamente por incluir otro tipo de información. Con el espíritu de mejorar la inclusión crediticia pone a disposición el set de datos.

Este set de datos, con el enfoque de este trabajo se va a dividir en dos grupos: uno de ellos vinculado información histórica de otros productos financieros que haya tenido el cliente, dentro o fuera de Home Credit. Cabe destacar que se trata de cualquier producto, ya sea préstamos de bajo o alto monto, tarjetas de crédito, etc. El segundo grupo, que es sobre el que más se trabajará, contiene todo tipo de información que no responda a la descripción del primer grupo. Se mencionan a continuación algunas temáticas de información con ejemplos: información sobre el cliente (edad sexo, estudios alcanzados, etc.), sobre el trabajo y sus condiciones laborales (antigüedad, tipo de organización, etc.), sobre la documentación presentada (cantidad de documentos presentada), sobre la aplicación (día de la semana y fecha a la que aplicó al crédito por primera vez, período de tiempo transcurrido entre la registración y la aplicación, etc.) entre otros.

2. Gestión de información provista por Home Credit

En este capítulo se tratará la gestión y el tratamiento de los datos, con el fin de preparar el input a utilizar para entrenar los modelos predictivos. Sera abordado con dos apartados organizados de la siguiente manera:

El primer apartado tratará acerca de la estructura y características de la base de datos y el tratamiento de la información; depuración, limpieza, agrupación de datos, creación de variables. En el segundo apartado se mostrarán las técnicas utilizadas para la reducción de la dimensional dad del set de datos: análisis factorial y de componentes principales.

2.1 Estructura, diseño y tratamiento de datos

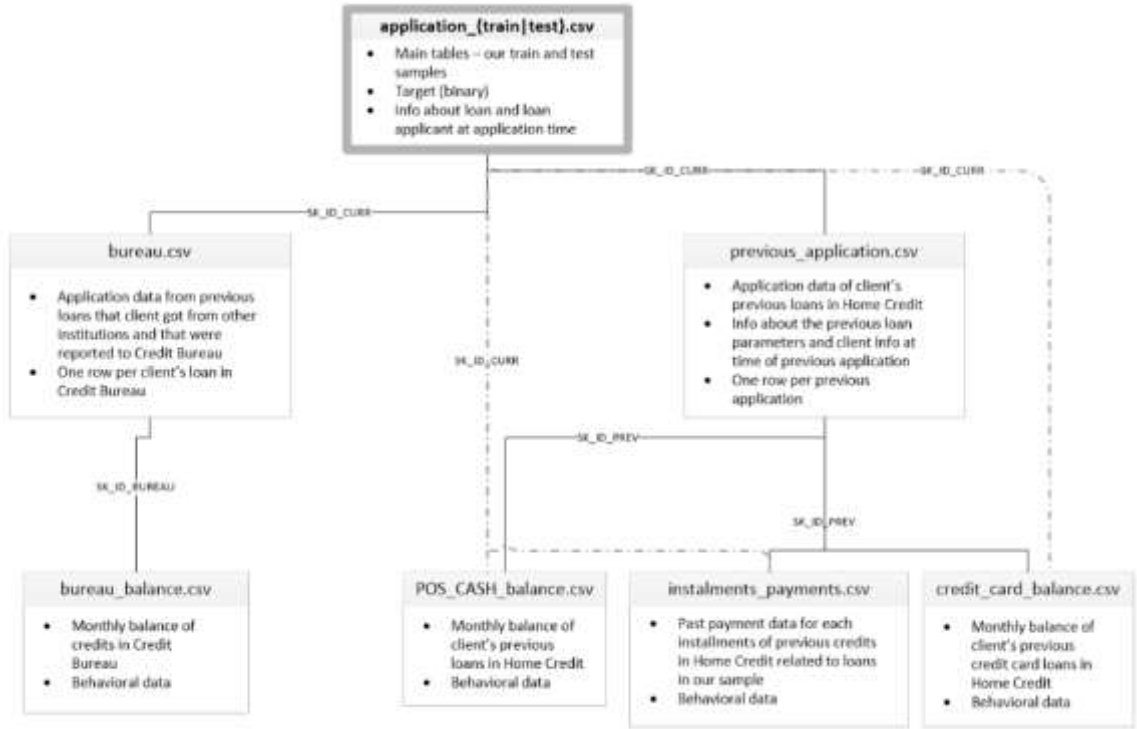
Para un correcto procesamiento y entendimiento de datos es necesario identificar a qué sistema de datos pertenece, cuál es su estructura y como tratarlos. En este apartado se definirán y clasificarán los datos provistos.

El set de datos responde a un sistema DBMS relacional. Estos sistemas se basan en tablas o relaciones de cada **entidad** que representen; puede ser una persona, cosa o hecho sobre las que se guarde información. Las entidades están organizada por **atributos** (columnas) que indican la característica o propiedad que se quiere representar, y **ocurrencias** (filas) que es cada instancia de esa entidad dentro del atributo al que pertenezca. La información puede estructurase en filas y columnas gracias a que el tipo de datos que se trabaja en estos sistemas es **estructurado** es decir, tienen una estructura interna identificable en términos del formato, longitud y tipo de información que contienen. Cada una de las relaciones, se vinculan a través de una campo o combinación de campos denominados **clave**, los cuales deberán permitir identificar unívocamente la ocurrencia de la entidad a la que se refiere (E. Chinkes). Habitualmente los sistemas DBMS utilizan un lenguaje estructurado de sentencias “SQL” para consultar y manipular los datos. Existen varios software que operan con este lenguaje, tales como: MySQL, SQLServer, Mariadb, Access, entre otros. El software utilizado para tratar las bases de datos para este trabajo es SQL-Server.

A continuación se expone el diseño conceptual de la bases de datos: cada cuadro indica una entidad, con una breve descripción de la información que contiene, las líneas indican relación entre tablas y el campo clave mediante el cual vincularlas. Cada variable que contengan las

identidades son atributos, y cada valor que tome ese atributo será una ocurrencia: por ejemplo, el atributo "Target" que se intenta predecir tendrá ocurrencias de "1 o 0" según si el cliente entró en default o no:

Imagen 1: Diseño conceptual de la base de datos



Fuente: Diseño extraído del sitio web Kaggle provisto por Home Credit, 2019

La tabla *application_Train* es la tabla principal, la cual contiene información de todos los créditos de hogar solicitados dentro de un año e información vinculada a éstos. En esta entidad se encuentran todas variables alternativas, exceptuando 3 atributos que corresponden a un score crediticio. La tabla *previous_application* y sus dependencias, contienen información histórica de todos los producto que hayan sido previamente solicitados por un cliente que actualmente tiene un crédito en Home Credit en esta misma entidad financiera (independiente si fue otorgado o no): está vinculado a la tabla principal por el código del crédito actual "SK_ID_CURR". Esta relación tiene **cardinalidad** "de uno a muchos", indicando la cantidad de ocurrencias que se conectan en las relaciones: implica que para cada crédito actual puede haber más de un producto solicitado previamente.

La tabla bureau y su dependencia contienen información de los producto que hayan sido solicitados por un cliente que actualmente tiene un crédito en Home Credit a cualquier entidad financiera del país. Esta información se obtiene gracias al Bureau, el cual contiene información crediticia y financiera de todas las entidades, y se comparte colaborativamente a fines de poseer más información comportamental de clientes. Esta relación también tiene cardinalidad "de uno a muchos".

Para poder proceder al modelado se requiere concentrar toda la información en una sola tabla cuya clave sea el código de crédito actual; a ese nivel se incorpora información del cliente, su comportamiento anterior y si entró en default o no. Es por esto que deben sumarizarse con determinados criterios definidos, la información histórica que corresponde a distintas aplicaciones previas del cliente, ya sea de bureau o propia de Home Credit para asignarlas a una crédito actual. De esta manera se buscará resumir en una única ocurrencia y varios atributos, el comportamiento histórico del cliente, asociado al crédito tiene actualmente. Con respecto a la tabla principal, se analizó la información provista para poder crear nuevas variables que se consideren pertinentes, detectar Outliers, corregir ausencia de datos, etc.
En el apartado, se encuentra el cuadro "número 1" con un resumen en el que se muestra el input final luego del tratamiento descripto hasta aquí de las variables financieras.

Con respecto a las variables alternativas (provistas en la tabla principal), cabe mencionar que se crearon variables adicionales utilizando información de la tabla, por ejemplo a partir del día de semana en la que se aplicó al crédito se crea una marca binario de *fin de semana.* Asimismo, con el fin de simplificar aquellas variables categóricas que contenían varios niveles, se crearon grupos según su *bad rate*[3]; por ejemplo, de los 58 tipos de organizaciones en las que puede estar empleado el cliente que solicita un préstamo, se los simplificó en 10 grupos, según su bad rate y tipo de organización. Por último, se analizaron Outliers, valores nulos y duplicidades; el principal inconveniente encontrado, surgió en torno a un subconjunto de variables en el cual se describen características de la vivienda del cliente, la cual está incompleta. Es por esto que se trabajará de forma independiente con este subset, el

[3] Bad rate: "tasa de malos" utilizada en variables categóricas. Representa la proporción de casos malos sobre el total. En este caso indica la proporción de clientes que entró en default.

cual contiene 54 atributos. En el cuadro número 2 del apéndice, se muestra un resumen de las variables alternativas (excluido el subset).

2.2 Reducción de dimensionalidad: PCA y Análisis factorial

En este apartado se buscará adentrar en el análisis exploratorio de los datos con el fin de comprender el tipo y la calidad de información. Debido a que se parte de una gran cantidad de atributos se considera necesario analizar la significatividad de éstos para poder efectuar una preselección de información ya que la potencia y eficiencia de algunos modelos a evaluar puede disminuir cuando se incorporan variables en exceso o cuando no son representativa del target a analizar.

Asimismo, cabe destacar que si bien se busca incluir variables alternativas al modelo, debe garantizarse que esta incorporación agregue valor y precisión y no perjudique la performance. De esta manera, se buscará comprobar que dichas variables agregan información y sean relevantes. Para este análisis se utilizará el método de componentes principales y análisis factorial. Se trata de técnicas de aprendizaje no – supervisado, es decir que no se busca establecer una relación entre una variable respuesta y otras explicativas sino que se utiliza para extraer información de conjunto de datos. Estas técnicas facilitan la visualización de datos, al mismo tiempo que son útiles para la reducción de dimensionalidad de información; así si se evaluarán cómo están compuestos los componentes principales en cada método, buscando catalogar el tipo de información que contengan y validando que en los primeros factores se encuentran variables provenientes de variables alternativas.

Los modelos y técnicas efectuados, se desarrollan en R un software estadístico libre y colaborativo. A continuación se exponen resultados encontrados en cada técnica, los scripts utilizados para arribar a estos valores, pueden encontrarse en el apéndice.

Se abordará en primer instancia el análisis de PCA o componentes principales, método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información, ya que busca encontrar una combinación lineal (normalizada) de las variables originales de un set de datos. Es por esto que para poder efectuar este análisis es requerimiento incluir un gran número de variables numéricas que

estén correlacionadas entre sí. En el caso bajo estudio, se cuenta con más de 70 variables que se validó mediante una matriz de correlación que están vinculadas entre sí. Debido a que los atributos están expresados en distintas magnitudes se las centró y estandarizó. A continuación se exponen los resultados arribados luego de aplicar el método de PCA, se muestran primeros 20 componentes:

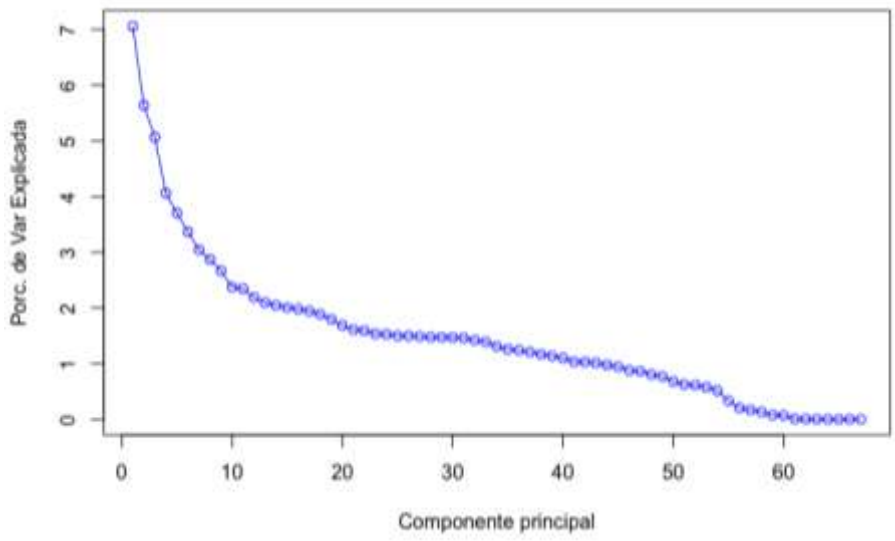
Tabla 1: Proporción de varianza explicada por cada CP

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	2,17499	1,94333	1,84248	1,6492	1,57556	1,50216	1,42846	1,38717	1,33718	1,26101
Proportion of Variance	0,07061	0,05637	0,05067	0,0406	0,03705	0,03368	0,03046	0,02872	0,02669	0,02373
Cumulative Proportion	0,07061	0,12697	0,17764	0,2182	0,25529	0,28897	0,31942	0,34814	0,37483	0,39856
	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20
Standard deviation	1,25347	1,21082	1,18476	1,17094	1,15931	1,1517	1,14087	1,1242	1,09633	1,06295
Proportion of Variance	0,02345	0,02188	0,02095	0,02046	0,02006	0,0198	0,01943	0,01886	0,01794	0,01686
Cumulative Proportion	0,42201	0,44389	0,46484	0,48531	0,50537	0,5252	0,54459	0,56346	0,5814	0,59826

Fuente: elaboración propia con datos de Home Credit, 2019

Se podrán obtener tantas componentes principales como variables disponibles; el orden de los mismos se corresponde con la variabilidad que expliquen sobre la varianza total; el primero es el que mayor varianza recoja, la segunda debe recoger la máxima variabilidad no recogida por la primera, y así sucesivamente, eligiendo un número que recoja un porcentaje suficiente de varianza total. Se expone en el siguiente gráfico la proporción de varianza explicada por cada componente:

Gráfico 1: Proporción de varianza explicada por cantidad de Componentes

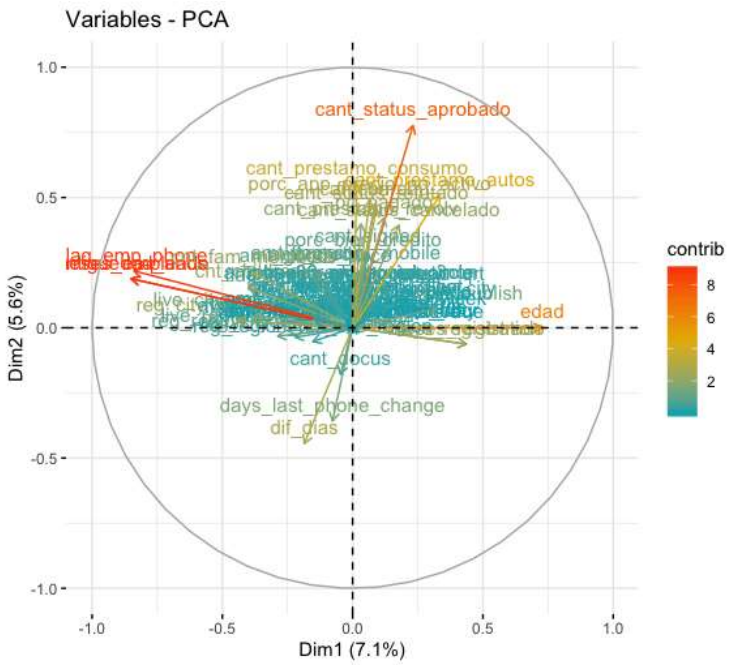


Fuente: elaboración propia en Software R con datos de Home Credit, 2019

En esta aplicación, se busca encontrar las variables más significativas. Es por esto que se debe ahondar en cada componente para entender qué infamación contienen. Se tomarán como ejemplos los primeros dos componentes, para mostrar el análisis que se efectuó con cada uno de ellos.

En el siguiente gráfico se representan las dos primeras componentes; el eje x representa a la primera componente, y el eje y representa a la segunda. La ubicación de cada variable, indicado con el direccionamiento del argumento adyacente refleja la contribución de la misma a cada dimensión. Las variables podrán tomar valores entre -1 y 1, de acuerdo a la importancia que tengan en cada eje, es decir en cada dimensión:

Gráfico 2: Primeras dos dimensiones - CPA



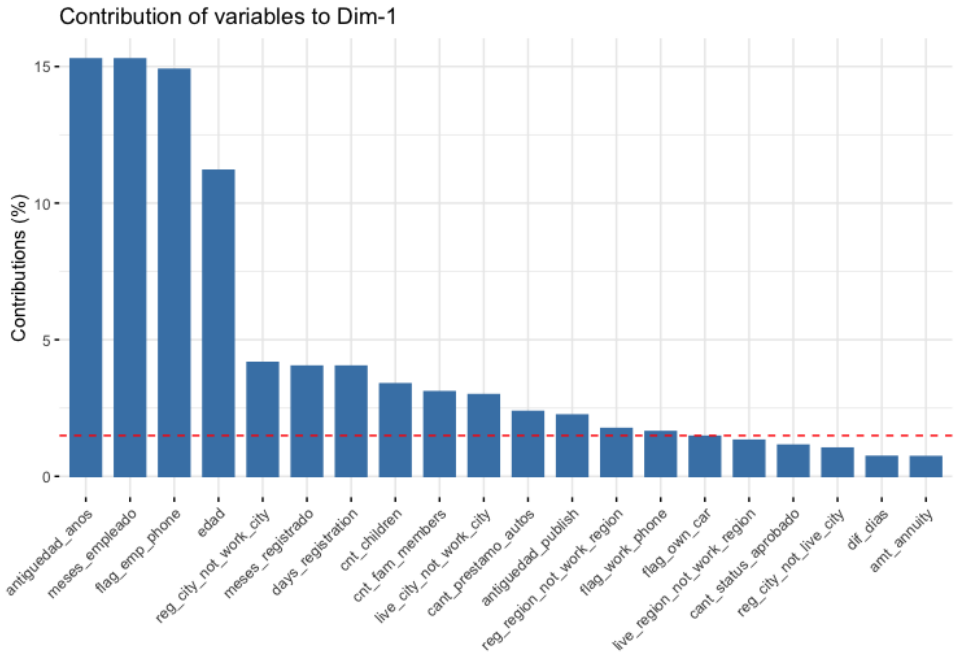
Fuente: elaboración propia en Software R con datos de Home Credit, 2019

A modo de ejemplo, la variable "edad" contribuye en gran medida a la variabilidad de la varianza de la primera dimensión o componente (por estar ubicada a lo largo de su eje) y prácticamente no se considera para la segunda dimensión (por estar situado en 0). Caso

contrario ocurre con el atributo "cant_status_aporbado". La escala de colores en el nombre de cada variable, acompaña la intensidad de la contribución.

Para clarificar la información provista del gráfico anterior, se expone el siguiente gráfico el cual ordena de forma descendente las variables más relevantes para la primera componente:

Gráfico 3: Contribución de variables principales en la primera dimensión



Fuente: elaboración propia en Software R con datos de Home Credit, 2019

Como podía esperarse, los atributos que más variabilidad explican son los que en el gráfico circular están pintados en naranja, cercanos al eje X. Observando ahora el histograma, se encuentra que por el tipo de variables lo encabezan "antigüedad_años (laboral)" "meses_empleado" y "marca teléfono trabajo", puede concluirse que la primera dimensión engloba las *características laborales* de los solicitantes de crédito. Asimismo, se concluye que estos atributos son de los más significativos por pertenecer a la primera componente. Este análisis se replicó para cada una de las componentes restantes, registrando qué representan y sus variables más relevantes.

Se considera pertinente aplicar al mismo conjunto de datos la técnica de Análisis Factorial. Si bien son metodologías similares, la principal diferencia es que el análisis de componentes

principales (ACP) analiza la varianza total del conjunto de variables observadas y , con base en ellas, trata de determinar las dimensiones básicas (o cantidad de "componentes") que las definen. En el análisis factorial (AF) el estudio de las interrelaciones entre las variables se restringe, a la varianza común (o covarianza), es decir, a la búsqueda de un número reducido de "factores" que expresen lo que es común al conjunto de variables observadas. A continuación se exponen los primeros 20 factores obtenidos para el Análisis Factorial:

Tabla 2: Análisis Factorial - aporte de cada factor

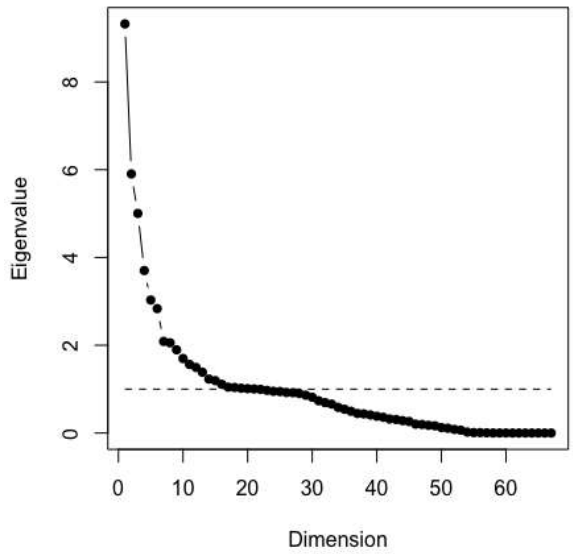
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10
SS loadings	4,299	3,323	2,775	2,297	2,182	1,988	1,93	1,901	1,665	1,658
Proportion Var	0,064	0,05	0,041	0,034	0,033	0,03	0,029	0,028	0,025	0,025
Cumulative Var	0,064	0,114	0,155	0,189	0,222	0,252	0,281	0,309	0,334	0,358
	Factor 11	Factor 12	Factor 13	Factor 14	Factor 15	Factor 16	Factor 17	Factor 18	Factor 19	Factor 20
SS loadings	1,545	1,527	1,313	1,137	1,123	1,119	1,047	0,892	0,788	0,78
Proportion Var	0,023	0,023	0,02	0,017	0,017	0,017	0,016	0,013	0,012	0,012
Cumulative Var	0,382	0,404	0,424	0,441	0,458	0,474	0,49	0,503	0,515	0,527

Fuente: elaboración propia en Software R con datos de Home Credit, 2019

En el cuadro se puede observar que los factores están ordenados según la proporción de varianza que logren representar. El primer factor, explica el 6% de la varianza. El segundo el 5%, y así sucesivamente. Con los primeros 20 factores, se obtiene el 50% de la misma; si bien puede resultar algo escaso, cabe recordar que se incorporaron al análisis más de 150 variables, es por esto que se espera encontrar muchos factores significativos.

De forma análoga se grafican los autovalores, vinculados a la varianza explicada por cada factor. Se considera que el punto de inflexión de la curva, marca el número óptimo de factores a considerar:

Gráfico 4: Análisis Factorial - aporte de cada factor

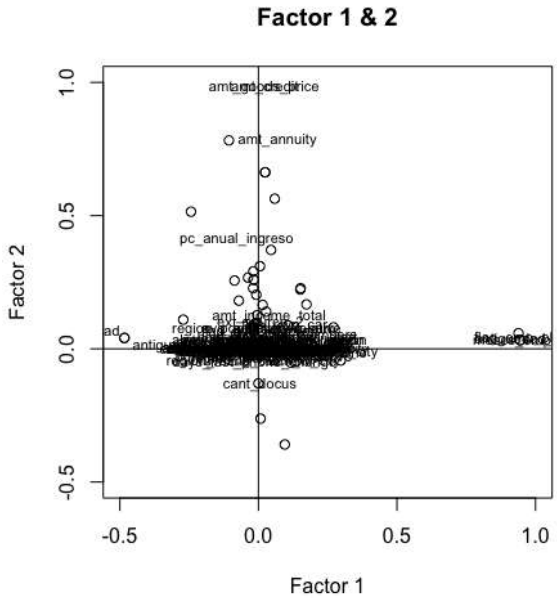


Fuente: elaboración propia en Software R con datos de Home Credit, 2019

En este caso, el número óptimo de dimensiones sería 17.

Al igual que en PCA, se pueden observar cómo se compone cada grupo, en un gráfico de dos ejes que representen a los dos primeros componentes. Se espera que los resultados sean equivalentes:

Gráfico 5: Dependencias de los dos principales factores

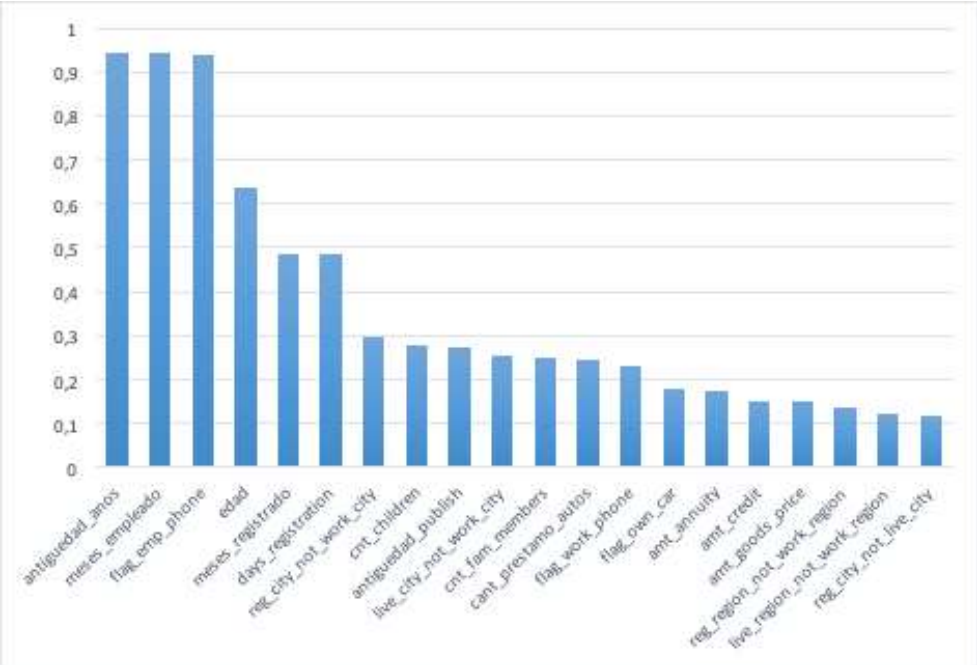


Fuente: elaboración propia en Software R con datos de Home Credit, 2019

Si se compara este gráfico con el obtenido en PCA, se observarán similitudes en términos de las variables explicativas del factor. Como era de esperarse, los primeros dos factores y dimensiones, compartes atributos como edad, o anualidades.

De forma análoga, para clarificar lo expuesto en gráfico anterior se enumeran los principales componentes del primer factor en el siguiente gráfico, ordenado según el porcentaje de varianza que expliquen:

Gráfico 6: Principales componentes del primer factor

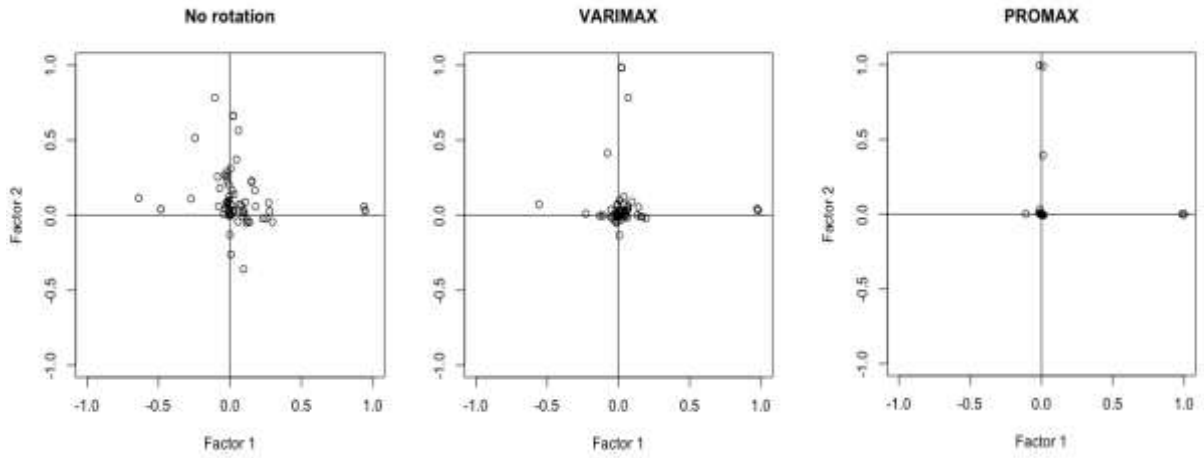


Fuente: elaboración propia en Software R con datos de Home Credit, 2019

Como era de esperarse, se encuentran nuevamente similitudes con los resultados obtenidos en componentes principales, donde las principales variables están vinculadas a las condiciones laborales del cliente.

En relación al gráfico "Factor 1 y 2" es interesante mencionar que es posible formular "rotaciones" en los ejes, con el fin de facilitar la interpretación de resultados. Es decir, es una forma de reexpresar los loadings de los factores, y por lo tanto impacta en la expresión gráfica de los factores. Las opciones de rotación de ejes pueden ser: sin rotación, varimax(ortogonal) y promax(oblicua). En los siguientes gráficos se muestra el mismo set de datos, expresados de la distinta manera. Se excluyen las etiquetas para limpiar el gráfico (el primer gráfico es análogo al anteriormente mostrado como "Factor 1 y 2"):

Gráfico 7: Efectos de la rotación de factores



Fuente: elaboración propia en Software R con datos de Home Credit, 2019

Los puntos observados continúan existiendo en todos los gráficos. Las diferencias se explican por la rotación, que impacta en el lugar que se ubiquen los puntos, y de esta manera en la interpretación que se efectúe de los resultados.

De la misma manera, se continúa con el resto de los factores, para comprender y comparar los resultados obtenidos con ambas técnicas.

3. Modelos predictivos para la estimación de la probabilidad de default

En este capítulo, se buscará exhibir resultados a los que se arribaron a partir de los modelos creados. Cabe destacar que se entrenaron distintos algoritmos para la predicción de la probabilidad default, pero se mostrarán aquellos que arrojaron una mayor AUC – área bajo la curva ROC. Se elige esta métrica como criterio para elegir el modelo, ya que es un buen parámetro para distinguir resultados de clasificación. Es decir, en el contexto de este trabajo, el AUC indica la probabilidad de que ante dos clientes, uno que entra en default y otro que no, el algoritmo los clasifique de forma correcta.

Los entrenamientos se efectuaron en el software R Studio, previo dividir la base en tres muestras: entrenamiento, testeo y validación. Esta separación es fundamental para poder verificar que el modelo ajusta correctamente. Dicho de otra forma, se evita un sobreajuste lo cual significa que la predicción es correcta sólo para ese subset de datos, porque el algoritmo "aprendió" sus características de manera que no sería capaz de predecir, sino de replicar los mismos resultados. Para evitar este efecto indeseable, se efectúa un proceso de cross – validation el cual consiste en tomar las bases de entrenamiento y testeo, particionados en n cortes y se itera n veces entrenando el modelo con todos los cortes y excepto uno, que utiliza para validar el entrenamiento. Luego de que el proceso anterior finalizó, se utiliza la muestra de validación y se compara la predicción del modelo con los targets reales de este conjunto. De esta manera, en caso que la métrica utilizada se mantenga relativamente similar al momento de cross-validation, se comprueba que el modelo no sobreajuste, y esa métrica es representativa del algoritmo.

Se entrenaron distintos algoritmos para la predicción de la probabilidad de default. En cada caso, se entrenaba para ambos modelos; variables financieras y otro para variables financieras y alternativas. El algoritmo que mayor precisión arrojó fue el de XGBoost. Este modelo, parte de árboles de decisiones ero potenciando los resultados de estos, debido al procesamiento secuencial de los datos con una función de pérdida o coste, la cual, minimiza el error iteración tras iteración, haciéndolo de esta manera, un pronosticador fuerte.

Este algoritmo, cuenta con la ventaja de tener capacidad de lidiar con distintos tipos de información, e incluso soporta irregularidades o valores perdidos en la información. Sin embargo, cuenta con la particularidad de contar con varios hiperparámetros, que es crucial

sean bien configurados para obtener la máxima precisión posible en las predicciones. Para poder obtener la configuración óptima, se recurrió a un método de optimización el cual requiere que se indiquen una serie de posibles valores para cada hiperparámetro, y el algoritmo buscará dentro de los valores otorgados la combinación óptima. Esta configuración podrá observarse en el código de R utilizado en el apéndice.

A continuación se mostrarán los resultados del modelo calibrado con variables financieras, en el segundo apartado se mostrarán resultados obtenidos utilizando variables financieras y alternativas. Por último, se contrastarán resultados obtenidos en los dos apartados previos.

3.1.Modelos predictivos con variables financieras

El modelo de XGBoost con variables financieras arrojó un AUC de 75.07%, luego de iterar 107 veces. El resultado del modelo, contemplando las 20 variables más importantes se observa en el siguiente cuadro:

Tabla 3: Principales variables financieras- XGBoost

Feature	Gain	Cover	Frequency	Importance
ext_source_3	0,274	0,158	0,101	0,274
ext_source_2	0,269	0,142	0,072	0,269
ext_source_1	0,084	0,081	0,067	0,084
pc_pagado	0,040	0,040	0,052	0,040
avg_anualidades_activos	0,022	0,032	0,034	0,022
avg_amt_credit_sum_debt	0,021	0,029	0,041	0,021
avg_amt_credit_sum	0,021	0,030	0,047	0,021
avg_amt_drawings_current	0,020	0,028	0,025	0,020
avg_amt_total_receivable	0,020	0,031	0,039	0,020
fl_closed	0,018	0,028	0,031	0,018
max_cnt_inst_future	0,016	0,026	0,026	0,016
cant_status_refutado	0,015	0,028	0,013	0,015
amt_payment	0,015	0,016	0,029	0,015
sum_dias_atraso	0,014	0,019	0,026	0,014
days_entry_payment	0,013	0,011	0,028	0,013
ct7	0,011	0,020	0,010	0,011
avg_aplicacion_credito	0,011	0,008	0,018	0,011
avg_amt_credit_max_overdue	0,009	0,021	0,025	0,009
avg_credito_no_activo	0,009	0,014	0,015	0,009
fl_active	0,008	0,020	0,023	0,008

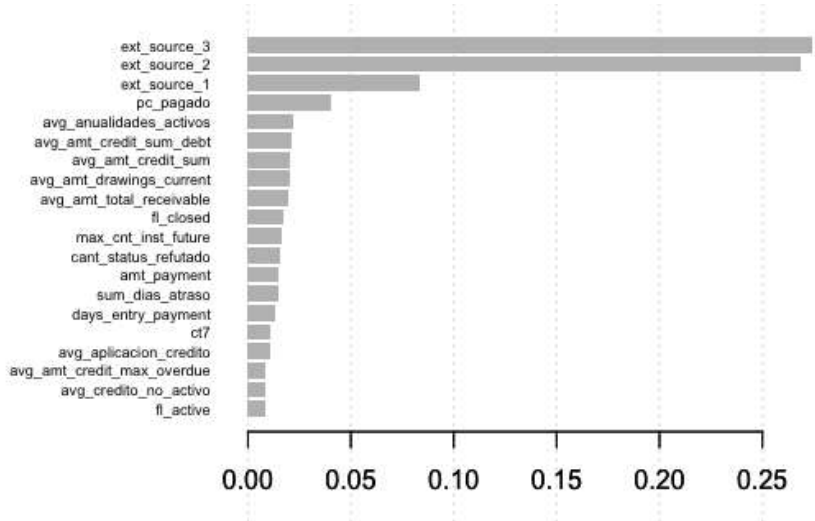
Fuente: elaboración propia con resultados obtenidos modelo XGBoost, 2019

Donde “Gain” o ganancia indica la contribución relativa de la característica correspondiente al modelo, calculada tomando la contribución de cada característica para cada árbol en el modelo. Un valor más alto de esta métrica en comparación con otra característica implica

que es más importante para generar una predicción. “Cover” o cobertura significa el número relativo de observaciones relacionadas con esta característica, considerando la cantidad de veces que se utilza para decidir la apertura del nodo en cada árbol del modelo final. “Frequency” o frecuencia representa el número relativo de veces que ocurre una característica particular en los árboles del modelo.

En “Importance” o importancia se define la medida más relevante para establecer el orden de las variables, en este caso es con ganancia. De forma gráfica pueden observarse la importancia de las variables obtenidas ordenadas de forma en el siguiente cuadro:

Gráfico 8: Importancia de variables financieras

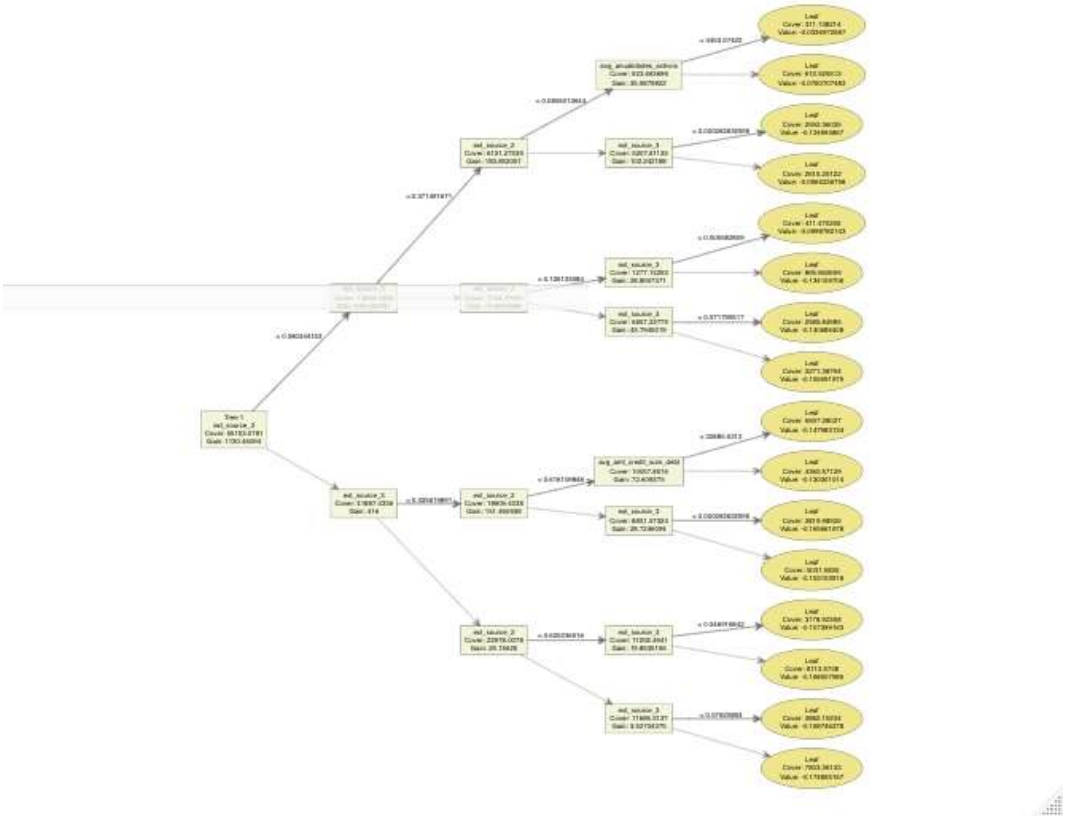


Fuente: elaboración propia en Software R con datos de Home Credit, 2019

Puede observarse que la mayor significatividad viene dada por las fuentes externas que catalogan el comportamiento del cliente. Las primeras tres variables responden a estas características, seguido de datos vinculados a las característica del préstamo, monto solicitado, etc.

Para clarificar el modo en el que opera el algoritmo, se muestra uno de los árboles (con una profundidad simplificada, con el fin de interpretar los resultados) que utiliza el modelo. Es decir, el algoritmo itrera sucesivas veces, armando en cada iteración árboles de distintas características, con el fin de catalogar o “aprender” en cada corrida un subconjunto distinto de los datos, para poder amalgamarlos al final. Uno de estos árboles luce como el siguiente:

Gráfico 9: Caso de ejemplo – árbol utilizado



Fuente: elaboración propia en Software R con datos de Home Credit, 2019

3.2. Incorporación de variables alternativas

Se procede en este apartado a mostrar resultados análogos al modelo anterior, pero considerando en el input variables alternativas y financieras de forma conjunta.

El modelo de XGBoost con variables financieras arrojó un AUC de 77.6%, luego de iterar 107 veces. El resultado del modelo, contemplando las 40 variables más importantes se observa en el siguiente cuadro (en negrita marcadas las “alternativas”:

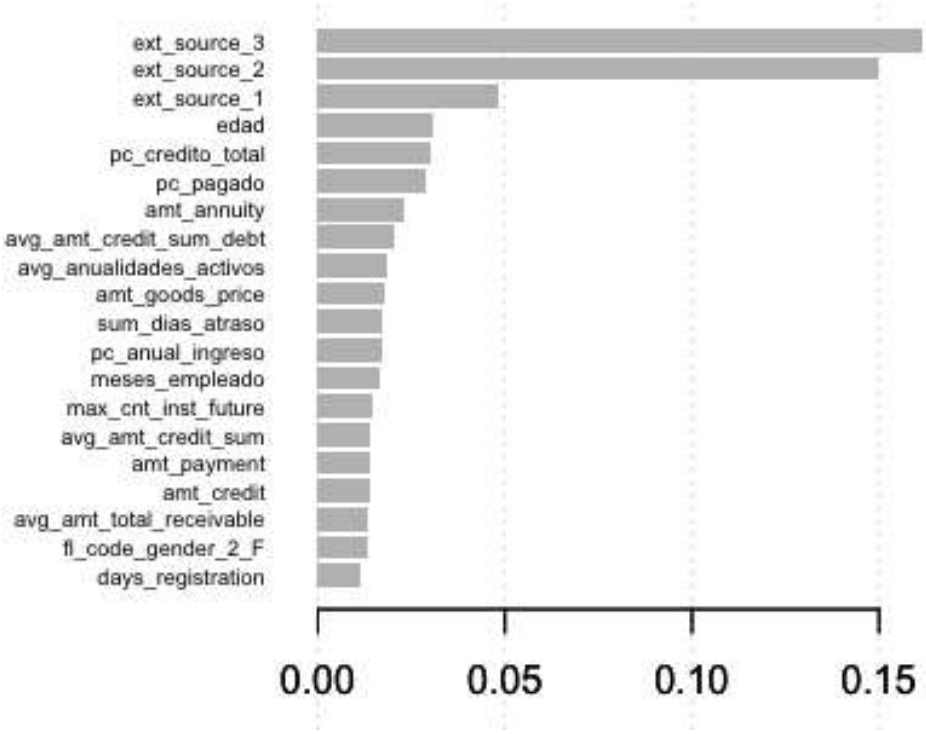
Tabla 4: Principales variables financieras y alternativas- XGBoost

Feature	Gain	Cover	Frequency	Importance
ext_source_3	0,161	0,078	0,046	0,161
ext_source_2	0,150	0,076	0,040	0,150
ext_source_1	0,048	0,047	0,033	0,048
edad	0,030	0,035	0,027	0,030
pc_credito_total	0,030	0,025	0,028	0,030
pc_pagado	0,029	0,025	0,036	0,029
amt_annuity	0,023	0,024	0,032	0,023
avg_amt_credit_sum_debt	0,020	0,020	0,025	0,020
avg_anualidades_activos	0,019	0,025	0,026	0,019
amt_goods_price	0,018	0,019	0,021	0,018
sum_dias_atraso	0,017	0,011	0,024	0,017
pc_anual_ingreso	0,017	0,015	0,030	0,017
meses_empleado	0,016	0,014	0,020	0,016
max_cnt_inst_future	0,015	0,013	0,018	0,015
avg_amt_credit_sum	0,014	0,018	0,022	0,014
amt_payment	0,014	0,013	0,019	0,014
amt_credit	0,014	0,018	0,022	0,014
avg_amt_total_receivable	0,013	0,014	0,013	0,013
fl_code_gender_2_F	0,013	0,013	0,012	0,013
days_registration	0,012	0,011	0,023	0,012
days_last_phone_change	0,011	0,009	0,022	0,011
antiguedad_publish	0,011	0,013	0,020	0,011
fl_active	0,011	0,017	0,012	0,011
avg_credito_activo	0,011	0,015	0,020	0,011
max_cnt_instal	0,010	0,010	0,016	0,010
avg_amt_drawings_current	0,010	0,017	0,011	0,010
days_entry_payment	0,010	0,008	0,016	0,010
fl_name_education_type_2_1	0,010	0,007	0,005	0,010
days_instalment	0,010	0,010	0,017	0,010
amt_instalment	0,009	0,008	0,016	0,009
region_population_relative	0,009	0,015	0,017	0,009
cant_status_refutado	0,008	0,010	0,008	0,008
fl_closed	0,008	0,009	0,011	0,008
avg_amt_credit_max_overdue	0,008	0,010	0,011	0,008
totalarea_mode	0,008	0,007	0,015	0,008
avg_aplicacion_credito	0,007	0,006	0,013	0,007
avg_credito_no_activo	0,007	0,006	0,010	0,007
own_car_age	0,006	0,007	0,007	0,006
cant_docus	0,006	0,007	0,006	0,006
ct7	0,006	0,009	0,002	0,006

Fuente: elaboración propia con resultados obtenidos modelo XGBoost, 2019

La variable alternativa "edad" es la primera que figura en cuarto lugar, y el resto de las variables aparecen después de la posición número 12 con importancia relativa baja. De forma gráfica pueden observarse la importancia de las variables obtenidas ordenadas de forma en el siguiente cuadro:

Gráfico 10: Importancia de variables financieras y alternativas



Fuente: elaboración propia en Software R con datos de Home Credit, 2019

A continuación se interpretará el impacto de estos resultados en el contexto de un banco, para la gestión de información y la toma de decisiones.

3.3.Comparación de resultados y modelos obtenidos

Para poder proceder a la toma de decisiones basadas en datos, es necesario expresar o traducir los resultados obtenidos de forma que se vinculen con la realidad que se está modelando, y los valores que buscan predecirse.

Para esto, habitualmente en bancos se suele utilizarse una representación de la distribución de la cartera de clientes para impactar resultados y gestionar el riesgo. Se exponen resultados en cuantiles de clientes teniendo en cuenta la probabilidad de default estimada en el modelo, es decir, se ordenan de forma descendente los clientes por su PD y se los agrupa en k grupos, en este caso deciles, como se muestra a continuación:

Tabla 5: Distribución de cartera – Predicción con modelo variables financieras

Decil	Cant. clientes	Targets (malos)	Predicción (malos)	PD Real	PD Estimada
10	3.923	1.043	1.050	26,6%	26,8%
9	3.867	560	550	14,5%	14,2%
8	3.972	385	403	9,7%	10,1%
7	3.921	319	304	8,1%	7,8%
6	3.921	244	241	6,2%	6,1%
5	3.937	175	193	4,4%	4,9%
4	3.904	130	153	3,3%	3,9%
3	4.008	107	123	2,7%	3,1%
2	3.839	83	88	2,2%	2,3%
1	3.915	50	57	1,3%	1,5%
TOTAL	39.207	3.096	3.162	7,9%	8,1%

Fuente: elaboración propia con resultados obtenidos, 2019

Tabla 6: Distribución de cartera – Predicción con modelo variables financieras + alternativas

Decil	Cant. clientes	Targets (malos)	Predicción (malos)	PD Real	PD Estimada
10	3.921	1.116	1.094	28,5%	27,9%
9	3.995	602	581	15,1%	14,5%
8	3.846	411	391	10,7%	10,2%
7	3.845	288	292	7,5%	7,6%
6	3.996	194	235	4,9%	5,9%
5	3.986	170	183	4,3%	4,6%
4	3.856	110	139	2,9%	3,6%
3	3.920	85	110	2,2%	2,8%
2	3.921	81	83	2,1%	2,1%
1	3.921	39	51	1,0%	1,3%
	39.207	3.096	3.159	7,9%	8,1%

Fuente: elaboración propia con resultados obtenidos, 2019

Que un cliente figure en el primer decil, implica que es "malo" en términos de la probabilidad de default esperada, y por el contrario, los clientes que se ubiquen en el último decil son los más deseables por el banco, por estimarse con menor probabilidad de default. Cuanto más distintos sean los deciles (mayor diferencia de pd promedio) implica que el modelo clasifica

mejor a los clientes porque los separa correctamente. Ambas tablas reflejan los resultados de misma muestra

Para poder comparar los resultados de forma más clara, se procede a acumular los valores de los deciles de tablas previas, es decir, el último decil contiene la totalidad de la cartera, ya que se van sumarizando los resultados. Se comparan la probabilidad estimada por cada modelo, y la real obtenida (ordenado según la estimación) en la siguiente tabla:

Tabla 7: Distribución de cartera – Comparación de predicciones

		MODELO VBES FINANCIERAS		MODELO VBES FINANC + ALT	
Deciles Acumulados	Observaciones	PD Real	PD Estimada	PD Real	PD Estimada
100%	39207	7,90%	8,07%	7,90%	8,04%
90%	35284	5,82%	5,99%	5,52%	5,68%
80%	31417	4,75%	4,97%	4,23%	4,43%
70%	27445	4,04%	4,22%	3,32%	3,63%
60%	23524	3,35%	3,64%	2,78%	3,03%
50%	19603	2,78%	3,13%	2,26%	2,54%
40%	15666	2,36%	2,69%	1,87%	2,12%
30%	11762	2,04%	2,28%	1,68%	1,74%
20%	7754	1,72%	1,87%	1,34%	1,40%
10%	3915	1,28%	1,45%	0,84%	1,02%

Fuente: elaboración propia con resultados obtenidos, 2019

En este cuadro se evidencia el impacto de la mejora en el AUC. Los bancos estableces criterios para definir las políticas que se tomen para el otorgamiento de créditos. Habitualmente, se define un corte de PD hasta el cual el banco está dispuesto a otorgar créditos; en este caso a un mismo corte de PD, por ejemplo 1.7% con el primer modelo se estarían otorgando préstamos hasta el 20% de la cartera con el segundo hasta el 30% de la cartera incluso con una pd inferior.

En el caso que se defina un criterio con respecto a porcentaje de otorgamiento de créditos, se encuentra que la pd promedio real esperada para cada decil es mayor en el modelo de variables financieras que en el de financieras + alternativas. Por ejemplo, con un corte del 30 % de otorgamiento (hasta el tercer decil) con el primer modelo se obtendría una pd esperada de 2.04%, mientras con el segundo modelo de 1,68%. El banco estaría obteniendo una pérdida esperada menor gracias a que el modelo clasifica mejor.

Cabe mencionar que la pd se utiliza no sólo para estimar este comportamiento que definirá la potencial pérdida por incumplimiento, sino que como se mencionó en el primer apartado es un componente para el requerimiento de reservas mínimas requeridas, por lo tanto esta probabilidad impacta tanto en resultados económicos como en resultados de gestión.

Conclusión

En este trabajo se buscó comprobar que las variables alternativas son significativas y relevantes en términos de la precisión que agregan a un modelo de Machine Learning. Como se mostró en el apartado anterior, el criterio de medición elegido "AUC", aumentó de 75.07% a 77.6%, dos puntos porcentuales que para la entidad financiera se traduce en: una reducción significativa de la probabilidad de default en la cartera a la que se le otorguen los préstamo si decide mantener el mismo porcentaje de otorgamiento de créditos, o bien puede incrementar la cantidad de créditos a otorgar si toma como decisión mantener una determinad probabilidad de default que soporte financieramente.

Desde la perspectiva del cliente, estos resultados también son beneficiosos, ya que puede permitir la inclusión crediticia de aquellos clientes cuya probabilidad de default se vea modificada en esta incorporación de información adicional. Es evidente que al mantenerse prácticamente la misma probabilidad promedio esperada en los dos modelos, habrá un subconjunto de la muestra que se verá beneficiada por estas variables y otro subconjunto perjudicada, pero la intención que se busca es conocer de manera más precisa el comportamiento esperado de los clientes sin castigar a un grupo una condición determinada.

Se considera que los resultados obtenidos producto de la apertura en el tipo de input es el mayor aporte de este trabajo, ya que como se mencionó en el primer apartado, la discriminación suele ser una problemática que se presenta de forma recurrente en el mundo de Machine Learning, y es importante dimensionar que una omisión o sesgo a la hora de considerar un set de datos iniciales, computar un algoritmo, o incluso la interpretación del mismo, puede causar exclusión o perpetuar una condición desfavorable. Este criterio podría transferirse a la mayoría de los proyectos y datasets, ya que se trata simplemente de un enfoque o perspectiva para abordar un análisis.

Por último, cabe mencionar que hay una diferencia de 3 puntos porcentuales entre la precisión obtenida en este trabajo y la obtenida por el ganador de la competencia de Kaggle; es por esto que existen aún posibilidades de mejora en el modelo, lo cual representa una oportunidad de futura línea de investigación. En línea con este punto, personalmente esperaba encontrar que las variables alternativas tuviesen una mayor importancia relativa a la que se encontró: en el modelo final elegido, la primer variable alternativa es "edad", la cual figuraba en quinto puesto, y el resto de las variables alternativas recién toman posición

a partir del 12ª puesto, con una importancia relativa baja. Es por esto, que es posible continuar trabajando con este dataset con el fin de entrenar otros algoritmos, crear nuevas variables o intentar otras técnicas que permitan continuar mejorando la inclusión crediticia.

Referencias bibliográficas

- ✓ Basle Committee on Banking Supervision, & Bank for International Settlements. (1999). Credit risk modelling: Current practices and applications. The Committee.
- ✓ Bluhm, C., Overbeck, L., & Wagner, C. (2016). Introduction to credit risk modeling. Chapman and Hall/CRC.
- ✓ Gillis, T. B., & Spiess, J. L. (2019). Big Data and Discrimination. The University of Chicago Law Review, 86(2), 459-488.
- ✓ Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11), 2767-2787.
- ✓ Panjer, H. H. (2006). Operational Risk.: Modeling Analytics (Vol. 620). John Wiley & Sons.
- ✓ Peng, R. D. (2016). R programming for data science (p. 471). Leanpub.
- ✓ Signes, A. T. (2018). La gobernanza colectiva de la protección de datos en las relaciones laborales:" big data", creación de perfiles, decisiones empresariales automatizadas y los derechos colectivos. Revista de derecho social, (84), 69-88.
- ✓ Chinkes, E. (2001). Modelado de sistemas de información. *Buenos Aires: Economizarte*.

Apéndice

Tabla 1 Apéndice – Descripción de variables financieras a incorporar en el modelo

FUENTE	ATRIBUTO	TIPO DE DATO	CRITERIO DE SUMARIZACION	DESCRIPCIÓN DEL ATRIBUTO
application_train	EXT_SOURCE_1	FLOAT	-	Normalized score from external data source
application_train	EXT_SOURCE_2	FLOAT	-	Normalized score from external data source
application_train	EXT_SOURCE_3	FLOAT	-	Normalized score from external data source
bureau	SK_BUREAU_ID	INT	CONTEO	Recoded ID of previous Credit Bureau credit related to our loan.
bureau	CREDIT_ACTIVE	CATEGORICO	CONTEO POR CATEGORIA	Status of the Credit Bureau (CB) reported credits
bureau	CREDIT_CURRENCY	CATEGORICO	CONTEO POR CATEGORIA	Recoded currency of the Credit Bureau credit
bureau	AMT_CREDIT_MAX_OVERDUE	INT	MAX	Maximal amount overdue on the Credit Bureau credit so far.
bureau	CNT_CREDIT_PROLONG	INT	CONTEO	How many times was the Credit Bureau credit prolonged
bureau	AMT_CREDIT_SUM	FLOAT	PROMEDIO	Current credit amount for the Credit Bureau credit
bureau	AMT_CREDIT_SUM_DEBT	FLOAT	PROMEDIO	Current debt on Credit Bureau credit
bureau	AMT_CREDIT_SUM_LIMIT	FLOAT	PROMEDIO	Current credit limit of credit card reported in Credit Bureau
bureau	AMT_CREDIT_SUM_OVERDUE	FLOAT	PROMEDIO	Current amount overdue on Credit Bureau credit
bureau	CREDIT_TYPE	CATEGORICO	CONTEO POR CATEGORIA	Type of Credit Bureau credit (Car, cash,...)
bureau	AMT_ANNUITY	FLOAT	PROMEDIO	Annuity of the Credit Bureau credit
POS_CASH_balance	SK_ID_PREV	INT	-	ID of previous credit in Home credit related to loan in our sample.
POS_CASH_balance	SK_ID_CURR	INT	-	ID of loan in our sample
POS_CASH_balance	MONTHS_BALANCE	INT	ULTIMO	Month of balance relative to application date
POS_CASH_balance	CNT_INSTALMENT	INT	ULTIMO	Term of previous credit (can change over time)
POS_CASH_balance	CNT_INSTALMENT_FUTURE	INT	ULTIMO	Installments left to pay on the previous credit
POS_CASH_balance	NAME_CONTRACT_STATUS	CATEGORICO	CONTEO POR CATEGORIA	Contract status during the month
POS_CASH_balance	SK_DPD	INT	MAX	DPD (days past due) during the month of previous credit
POS_CASH_balance	SK_DPD_DEF	INT	MAX	DPD during the month with tolerance of the previous credit
credit_card_balance	SK_ID_PREV	INT	-	ID of previous credit in Home credit related to loan in our sample.
credit_card_balance	SK_ID_CURR	INT	-	ID of loan in our sample
credit_card_balance	MONTHS_BALANCE	INT	ULTIMO	Month of balance relative to application date
credit_card_balance	AMT_BALANCE	FLOAT	PROMEDIO	Balance during the month of previous credit
credit_card_balance	AMT_CREDIT_LIMIT_ACTUAL	FLOAT	PROMEDIO	Credit card limit during the month of the previous credit
credit_card_balance	AMT_DRAWINGS_ATM_CURRENT	FLOAT	PROMEDIO	Amount drawing at ATM during the month of the previous credit
credit_card_balance	AMT_DRAWINGS_CURRENT	FLOAT	PROMEDIO	Amount drawing during the month of the previous credit
credit_card_balance	AMT_DRAWINGS_OTHER_CURRENT	FLOAT	PROMEDIO	Amount of other drawings during the month of the previous credit
credit_card_balance	AMT_DRAWINGS_POS_CURRENT	FLOAT	PROMEDIO	Amount drawing or buying goods during the month of the previous credit
credit_card_balance	AMT_RECIVABLE	FLOAT	PROMEDIO	Amount receivable on the previous credit
credit_card_balance	CNT_DRAWINGS_CURRENT	INT	ULTIMO	Number of drawings during this month on the previous credit
previous_application	NAME_CONTRACT_TYPE	CATEGORICO	CONTEO POR CATEGORIA	Contract product type of the previous application
previous_application	AMT_ANNUITY	FLOAT	PROMEDIO	Annuity of previous application
previous_application	AMT_APPLICATION	FLOAT	PROMEDIO	For how much credit did client ask on the previous application
previous_application	AMT_CREDIT	FLOAT	PROMEDIO	Final credit amount on the previous application.
previous_application	AMT_DOWN_PAYMENT	FLOAT	PROMEDIO	Down payment on the previous application
previous_application	AMT_GOODS_PRICE	FLOAT	PROMEDIO	Goods price of good that client asked for (if applicable) on the previous application
previous_application	WEEKDAY_APPR_PROCESS_START	INT	MODA	On which day of the week did the client apply for previous application
previous_application	HOUR_APPR_PROCESS_START	TIME	-	Approximately at what day hour did the client apply for the previous application
previous_application	NAME_CONTRACT_STATUS	CATEGORICO	CONTEO POR CATEGORIA	Contract status (approved, cancelled, ...) of previous application
previous_application	CODE_REJECT_REASON	CATEGORICO	CONTEO POR CATEGORIA	Why was the previous application rejected
previous_application	NAME_CLIENT_TYPE	CATEGORICO	CONTEO POR CATEGORIA	Was the client old or new client when applying for the previous application
installments_payments	SK_ID_PREV	INT	-	ID of previous credit in Home credit related to loan in our sample.
installments_payments	SK_ID_CURR	INT	-	ID of loan in our sample
installments_payments	NUM_INSTALMENT_VERSION	INT	MAX	Version of installment calendar (0 is for credit card) of previous credit.
installments_payments	NUM_INSTALMENT_NUMBER	INT	MAX	On which installment we observe payment
installments_payments	DAYS_INSTALMENT	INT	SUMA	When the installment of previous credit was supposed to be paid.
installments_payments	DAYS_ENTRY_PAYMENT	INT	SUMA	When was the installments of previous credit paid actually.
installments_payments	AMT_INSTALMENT	FLOAT	SUMA	What was the prescribed installment amount of previous credit on this installment
installments_payments	AMT_PAYMENT	FLOAT	SUMA	What the client actually paid on previous credit on this installment

Fuente: elaboración propia en Software R con datos de Home Credit, 2019

Nota: para sumarizar la información se evalúa el tipo de dato que contiene y a partir de este, se elige un criterio que se estima más representativo para resumir información (puede ser conteo, suma, promedio, máximo, mínimo, último, primero o varios de éstos). Cabe destacar que en algunos casos se resume información de más de 50 productos solicitados, es por esto que se excluyeron del análisis algunas variables debido a que carecían de sentido al tener que sumarizarlas.

Tabla 2 Apéndice – Descripción de variables financieras a incorporar en el modelo

ATRIBUTO	TIPO DE DATO	DESCRIPCIÓN DEL ATRIBUTO
SK_ID_CURR	CLAVE	ID of loan in our sample
TARGET	A PREDECIR	Target variable 1: default, 0 not-default
NAME_CONTRACT_TYPE	CAT	Identification if loan is cash or revolving
CODE_GENDER	CAT - BINARIO	Gender of the client
FLAG_OWN_CAR	CAT - BINARIO	Flag if the client owns a car
FLAG_OWN_REALTY	CAT - BINARIO	Flag if client owns a house or flat
CNT_CHILDREN	INT	Number of children the client has
AMT_INCOME_TOTAL	FLOAT	Income of the client
AMT_CREDIT	FLOAT	Credit amount of the loan
AMT_ANNUITY	FLOAT	Loan annuity
AMT_GOODS_PRICE	FLOAT	For consumer loans it is the price of the goods for which the loan is given
NAME_TYPE_SUITE	CAT	Who was accompanying client when he was applying for the loan
NAME_INCOME_TYPE	CAT	Clients income type (businessman, working, maternity leave,Ö)
NAME_EDUCATION_TYPE	CAT	Level of highest education the client achieved
NAME_FAMILY_STATUS	CAT	Family status of the client
NAME_HOUSING_TYPE	CAT	What is the housing situation of the client
REGION_POPULATION_RELATIVE	FLOAT	Normalized population of region where client lives
DAYS_BIRTH	INT	Client's age in days at the time of application
DAYS_EMPLOYED	INT	How many days before the application the person started current employment
DAYS_REGISTRATION	INT	How many days before the application did client change his registration
DAYS_ID_PUBLISH	INT	How many days before the application did client change the identity documenT
OWN_CAR_AGE	CAT	Age of client's car
FLAG_MOBIL	CAT - BINARIO	Did client provide mobile phone (1=YES, 0=NO)
FLAG_EMP_PHONE	CAT - BINARIO	Did client provide work phone (1=YES, 0=NO)
FLAG_WORK_PHONE	CAT - BINARIO	Did client provide home phone (1=YES, 0=NO)
FLAG_CONT_MOBILE	CAT - BINARIO	Was mobile phone reachable (1=YES, 0=NO)
FLAG_PHONE	CAT - BINARIO	Did client provide home phone (1=YES, 0=NO)
FLAG_EMAIL	CAT - BINARIO	Did client provide email (1=YES, 0=NO)
OCCUPATION_TYPE	CAT - BINARIO	What kind of occupation does the client have
CNT_FAM_MEMBERS	INT	How many family members does client have
REGION_RATING_CLIENT	CAT - ORDINAL	Our rating of the region where client lives (1,2,3)
REGION_RATING_CLIENT_W_CITY	CAT - ORDINAL	Our rating of the region where client lives with taking city into account (1,2,3)
WEEKDAY_APPR_PROCESS_START	CAT	On which day of the week did the client apply for the loan
HOUR_APPR_PROCESS_START	TIME	Approximately at what hour did the client apply for the loan
REG_REGION_NOT_LIVE_REGION	CAT - BINARIO	Flag if client's permanent address does not match contact address
REG_REGION_NOT_WORK_REGION	CAT - BINARIO	Flag if client's permanent address does not match work address
LIVE_REGION_NOT_WORK_REGION	CAT - BINARIO	Flag if client's contact address does not match work address
REG_CITY_NOT_LIVE_CITY	CAT - BINARIO	Flag if client's permanent address does not match contact address
REG_CITY_NOT_WORK_CITY	CAT - BINARIO	Flag if client's permanent address does not match work address
LIVE_CITY_NOT_WORK_CITY	CAT - BINARIO	Flag if client's contact address does not match work address
ORGANIZATION_TYPE	CAT	Type of organization where client works
OBS_30_CNT_SOCIAL_CIRCLE	INT	How many observation 30 DPD (days past due) default
DEF_30_CNT_SOCIAL_CIRCLE	INT	How many observation defaulted on 30 DPD (days past due)
OBS_60_CNT_SOCIAL_CIRCLE	INT	How many observation observable 60 DPD (days past due) default
DEF_60_CNT_SOCIAL_CIRCLE	INT	How many observation defaulted on 60 (days past due) DPD
DAYS_LAST_PHONE_CHANGE	INT	How many days before application did client change phone
FLAG_DOCUMENTS (20 ATRIBUTOS)	INT	Did client provide document 2.. document 21
AMT_REQ_CREDIT_BUREAU_HOUR	INT	Number of enquiries to Credit Bureau about the client one hour before application
AMT_REQ_CREDIT_BUREAU_DAY	INT	Number of enquiries to Credit Bureau about the client one day before application
AMT_REQ_CREDIT_BUREAU_WEEK	INT	Number of enquiries to Credit Bureau about the client one week before application
AMT_REQ_CREDIT_BUREAU_MON	INT	Number of enquiries to Credit Bureau about the client one month before applicatior
AMT_REQ_CREDIT_BUREAU_QRT	INT	Number of enquiries to Credit Bureau about the client 3 month before application
AMT_REQ_CREDIT_BUREAU_YEAR	INT	Number of enquiries to Credit Bureau about the client one day year

Fuente: elaboración propia en Software R con datos de Home Credit, 2019