Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Negocios y Administración Pública

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

TRABAJO FINAL DE ESPECIALIZACIÓN

Abandono de clientes en el sector bancario

Análisis exploratorio y predictivo a través de minería de datos.

AUTOR: DENIS TROSMAN
MENTOR: ROBERTO ABALDE

DICIEMBRE 2021

Resumen

Las instituciones financieras sufren año tras año la partida de sus clientes tanto en términos monetarios como en términos de competencia con sus pares. Por este motivo, resulta pertinente para estas tener en claro qué enfoques y sobre qué tipos de usuarios deben orientar sus campañas de retención de clientes. Este trabajo genera tanto un análisis exploratorio como predictivo de los casos de abandono en un banco comercial. Particularmente, en términos de descontinuación de uso de tarjetas de crédito ofrecidas por este. A partir de una base de datos de alrededor de 10.000 usuarios, se utilizaron métodos de análisis multivariante de datos para reducir la dimensionalidad de la base de datos y englobar tanto atributos explicativos como clientes de características similares. Así mismo, se generaron seis modelos predictivos de casos de abandono, otorgando el mejor resultado en términos de AUC aquellos que trabajan bajo los algoritmos GradientBoosting y XGBoosting, de 0,94.

Palabras clave: abandono de clientes – minería de datos – análisis exploratorio – modelo predictivo – análisis de conglomerados –componentes principales

Abstract

Financial institutions suffer year after year from customer attrition both in monetary and competition terms against their peers. For this reason, it is relevant for them to have a clear sight on what approaches and which types of users they should focus their customer retention campaigns on. This study generates both an exploratory and predictive analysis of customer attrition of a particular commercial bank. Particularly, in terms of discontinuing credit cards usage. Using a database of around 10,000 bank customers, multivariate data analysis methods were used to reduce the dimensionality of the database and encompass both explanatory attributes and customers with similar characteristics. Likewise, six predictive models were generated to predict customers abandonment. Those working under Gradient Boosting and XGBoosting algorithms showed the greatest performances in terms of AUC, at 0.94.

Keywords: Customer's attrition – data mining – exploratory analysis – predictive model – cluster analysis– principal components

Índice

Introducción

Las empresas e instituciones pierden grandes cantidades de dinero tanto directa e indirectamente cuando sus clientes o suscriptores dejan de acceder a sus productos o servicios. Analizando y entendiendo correctamente esta métrica de abandono de clientes, estas instituciones pueden minimizar la pérdida de ingresos proveniente de estas cancelaciones. En tiempos como los actuales, donde la competencia es extraordinariamente alta en cuanto a oferta de servicios financieros, esta problemática resulta de suma importancia para instituciones financieras como lo son, por ejemplo, los bancos comerciales. Está en manos de estos últimos tomar las medidas correctas y eficientes para evitar la existencia de abandono de clientes en su cartera.

Ya sea en el sector bancario o en cualquier empresa u organización cuyo objetivo sea maximizar su número de clientes activos, resulta importante determinar tanto los motivos por los cuales clientes dejan de consumir servicios ofrecidos como así también las tendencias que puedan observarse en las características específicas de estos. Minimizando este número de abandono de clientes, las compañías pueden maximizar sus beneficios. Estos casos pueden estar impulsados tanto por distintas dinámicas internas de las propias instituciones, empresas u organizaciones ofertantes de los productos y servicios, como también por variables exógenas a estas como lo son aquellas características demográficas o económicas de los clientes.

Algunos de los motivos por los cuales detectar los posibles casos de abandono ayudan a un banco comercial pueden ser, por ejemplo, que los costos de marketing para adquirir nuevos clientes son altos, por lo que retener clientes es necesario para cubrir los costos hundidos. O que la retención de clientes sirve para poder calcular el valor de vida de estos. Otro beneficio de este análisis es que tiene una relación directa con poder expandir la base de clientes de los bancos, como también identificar que acciones tomadas por los bancos están efectivamente mejorando esta medida de abandono y cuales están teniendo un impacto negativo. *"All these considerations are typically included in the concept of customer relationship management (CRM), which is a business strategy that modifies the processes management of a company. The proper use of CRM allows a company to improve its revenues, ensuring the customers' satisfaction (i.e. improving customer retention)"* (Ling y Yen, 2001).

De la misma manera, el estudio de análisis de casos de abandono es una parte importante en el proceso de *customer relationship management.* Contiene información sobre aspectos relevantes que puedan llegar a impactar en el comportamiento de los clientes, como pueden ser el precio, la calidad de servicio, la reputación de la organización, la eficiencia de las campañas de retención, y más. (Coltman, 2007).

Para reducir el número de casos de abandono dentro de una institución– sea financiera o de cualquier otro tipo– existen análisis descriptivos y predictivos de datos. A partir del correcto modelaje que explique que variables impactan en mayor peso a estos casos de abandono, se puede tanto entender como predecir que suscriptores necesitan un cambio de estrategia o un refuerzo de atención para minimizar el potencial costo de estos yéndose a una institución competidora. Las instituciones hacen uso de distintos métodos estadísticos y *data mining*, como lo son el análisis de conglomerados, de componentes principales, el análisis discriminante, el análisis de correspondencia o la generación de árboles de decisión, para poder deducir y pronosticar los casos de abandono de clientes de un caso de estudio como lo es un banco comercial, e idear posibles estrategias para maximizar la retención de los clientes. Los datos de los bancos están generalmente estructurados en un gran *data warehouse*, o *data lake.* No es fácil aplicar técnicas de *data mining* a estos directamente, ya que además de tener una gran cantidad de datos que ralentizan el análisis, también causan ruido y generan información relevante. Por estos motivos, las instituciones deben hacer un muestreo de sus datos, y realizar un preprocesamiento de estos.

El propósito de este trabajo es realizar un análisis exhaustivo de estos casos en los cuales los clientes dejan de estar suscriptos a servicios, especialmente aquellos ofrecidos por instituciones bancarias, y poder así idear estrategias que sean exitosas a la hora de maximizar la retención de clientes. Se buscará entender cuáles son los motivos por los cuales los clientes dejan de consumir determinados servicios y como pueden prevenirse este abandono de clientes en base a las características y comportamientos específicos de este grupo.

Dado este objetivo, se hará uso de herramientas de estadística descriptiva y predictiva para realizar un modelo de clasificación que pueda predecir correctamente los casos de abandono de clientes en el sector bancario y, a partir de este, poder desarrollar posibles estrategias de retención de clientes. Para determinar cuáles variables sirven para describir el proceso de desuso de servicios ofrecidos por un caso de estudio de una institución bancaria, y que enfoque puede tomarse para minimizar el número de clientes que dejan de consumir los servicios ofrecidos, también se buscará elaborar un modelo que permita predecir de la mejor manera posible a potenciales casos de abandono futuros. Resulta importante describir las variables que mejor explican el comportamiento de los clientes que han renunciado al servicio ofrecido, tanto para poder enfocarse en la mejora o modificación de estas, como para poder hacer predicciones más eficientes.

A la hora de analizar qué variables demográficas o financieras son determinantes para reconocer posibles casos de abandono, se comenzará con una evaluación general de los datos y un análisis exploratorio. Este será fundamental para poder generar un análisis descriptivo y observar las distribuciones dentro de los grupos de clientes que abandonaron los servicios bancarios y los que no. Así mismo, se volverá a este objetivo luego de elegir el mejor modelo de predicción de casos de abandono de clientes. Esto es así ya que, al hacer uso de modelos predictivos, una parte de sus resultados pueden ser los pesos de las variables en el momento de generar la predicción.

Para encontrar el mejor modelo explicativo de la base de datos en cuestión, se realizará un análisis de conglomerados (clustering), un análisis de correspondencia y un análisis de componentes principales (PCA). Finalmente, para poder generar el mejor modelo de predicción de abandono de clientes, se utilizarán los modelos de árboles de decisión y el ensamble *Random Forest*, como también otros modelos clasificatorios como la regresión logística, *XGB* o *Gradient Boosting*.

Análisis exploratorio del abandono de clientes bancarios

Como primer paso de este trabajo, se procederá a realizar un análisis descriptivo exhaustivo de la importancia del tema en la actualidad, como también de la base de datos que fue escogida para generar el análisis y entrenar los modelos predictivos.

1.1. Importancia de retención de clientes en el sector bancario

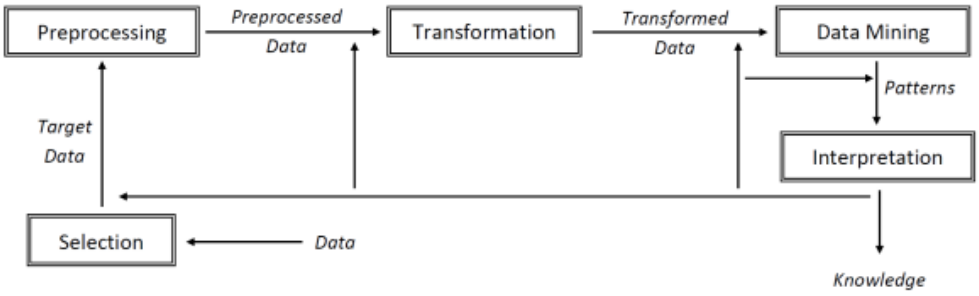
La utilización de métodos de minería de datos en el sector bancario sigue evolucionando y es de suma importancia a la hora de medir las distintas métricas que puedan definirse como claves para hacer un seguimiento de la performance de cada institución. Recolectar y analizar datos da una ventaja competitiva que los negocios e instituciones ya reconocen hace largos años. Para esto, los modelos estadísticos descriptivos y predictivos resultan claves, ayudando a determinar patrones y tendencias tanto de los clientes en general como del rubro al que pertenecen.

"*Customer Retention is an increasingly pressing issue in today's ever competitive commercial arena*", (Kaur, et. al 2013). Como remarca Kaur, una de estas métricas que es de suma importancia, en particular para las instituciones financieras, es la de los clientes que abandonan sus servicios. Estos usuarios que han finalizado la suscripción de un servicio funcionan como una medida tanto de duración de clientes como de observación de posibles métodos para evitarlos. Dado que generalmente un caso de abandono significa una huida hacia una institución denominada como competencia, la importancia del entendimiento de esta métrica resulta determinante. Algunos estudios han confirmado que el costo de adquirir un nuevo cliente puede llegar a costar hasta cinco veces más el costo de satisfacer y mantener clientes existentes (Lemmens, et. al, 2013). La retención de estos clientes tiene un gran impacto en cada institución financiera, y entender el verdadero valor de un posible caso de cancelación ayuda a las relaciones de manejo de relación de clientes (CRM) de estas.

Los casos de abandono no solo son importantes para los bancos y organizaciones por la pérdida futura de ganancia, sino también por los fallidos intentos que estos realizan en términos de campañas de marketing o atención al cliente que resultan inefectivas y se traducen en un costo. El incorrecto *targeting* de usuarios en campañas de retención puede también traducirse en el cliente cansándose de las promociones que considera innecesarias, lo cual al mismo tiempo puede derivar en una cancelación del servicio o al cliente ignorando futuras ofertas. Por este motivo, es clave para los bancos entender sobre qué tipo de clientes deben focalizar sus campañas buscando la retención de estos, como también saber que casos pueden considerarse irremediables para saber en qué grupo no hay que invertir capacidades o tiempo.

Como deja notar la Figura 1, para extraer un verdadero *insight* de una base de datos hay distintos pasos que las instituciones deben realizar. Primero, hacer una selección de los datos que poseen y preprocesarlas. A este *set* de datos se le deben luego generar las transformaciones necesarias para facilitar lo que posteriormente será la fase de *data mining*, como puede ser, por ejemplo, la corrección de formatos o selección de variables.

Figura 1. Descubrimiento de conocimiento en un esquema de base de datos (KDD).



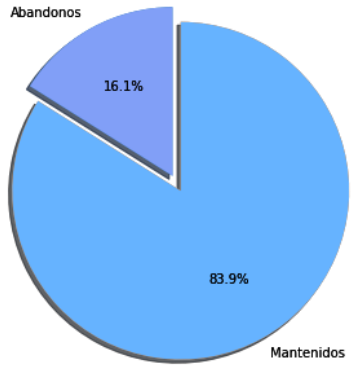
Fuente: Avon (2015)

Finalizada la etapa de *data mining*, pueden generarse patrones que sean interpretables y se traduzcan en conocimiento.

1.2. Descripción de la base de datos

En este trabajo se utilizará un set de datos con información de 10,127 clientes actuales y pasados de un banco comercial. Como se puede ver en el Gráfico 1, un 16,1% de los clientes en esta base de datos son determinados *attrited customers*, o, en otras palabras, casos de abandono. El resto, son clientes existentes. En este caso particular, la variable a considerarse como abandono es el desuso de tarjetas de crédito.

Gráfico 1. Proporcion de clientes según abandono de banco



Fuente: Elaboración propia en base a Kaggle

Como variables explicativas, se tomarán en cuenta variables demográficas como lo son la edad, el sexo, la educación y el ingreso de los clientes, y variables financieras internas del banco comercial como la categoría de los productos, el tiempo de vinculación, el límite de crédito y el total de transacciones, entre otras.

Tabla 1. Definición de variables.

Variables	**Definición**
Attrition	Variable binaria definitoria de casos de abandono
Variables demográficas	Sexo, Edad, Educación, Estado civil, Nivel de ingresos
Card category	Tipo de tarjeta (Blue, Silver, Gold, Platinium)
Months on book	Tiempo de relación con el banco
Total relationship count	Número de productos
Meses inactivos	Cantidad de meses inactivos en los últimos 12 meses
Contactos	Numero de contactos en los últimos 12 meses
Límite de crédito	Límite de gasto en tarjeta de crédito
Revolving balance	Porción del gasto de tarjeta de crédito que no es pagado al final del ciclo
Open-to-buy	Diferencia entre límite de crédito y balance
Detalle de transacciones	Cambio del volumen de transacciones, Volumen de transacciones, Total de transacciones, Cambio en el total de transacciones
Average card utilization ratio	Porcentaje del crédito disponible que no es pagado al final del ciclo

Fuente: Elaboración propia en base a Kaggle.

En cuanto a la limpieza de la base de datos utilizada, se nota la existencia de valores nulos o desconocidos para algunas de las variables:

Tabla 2. Cantidad de valores nulos

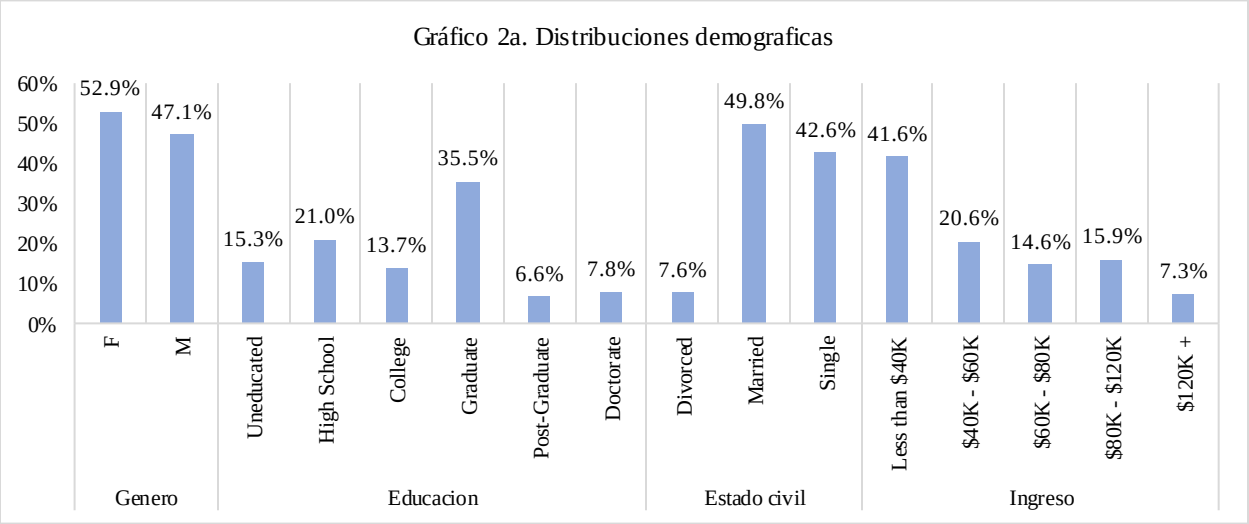
Variable	N° nulos
Education_Level	1,519 (15,00%)
Marital_Status	749 (7,40%)
Income_Category	1,112 (10,98%)

Fuente: Elaboración propia en base a Kaggle.

Para resolver este problema de datos nulos, se utilizó el algoritmo KNN (vecinos cercanos) para imputar datos faltantes. Se ha utilizado la librería *missingpy* [1]y su función *KNNImputer* en Python. Para cumplir con este objetivo, se han transformado las variables categóricas, como lo son por ejemplo el nivel educativo (*Uneducated, High-School, College, Graduate, Post-Graduate, Doctorate)*, para que cada posible valor de estas esté contemplado de forma numérica y sea posible estandarizar la base de datos a partir de la función StandardScaler de *sklearn.*[2] Se escogió el modelo KNN con dos vecinos y el método de distancia para realizar la imputación, y esta misma fue generada a partir de *batchs* para que funcione más eficientemente el modelo.

En cuanto a la presencia de *outliers*, se ha optado por no eliminar observaciones de la base de datos ya que el número de estas no puede considerarse lo suficientemente alto como para hacerlo. Sin embargo, se nota a partir del método *Z-Score*, cual básicamente resume la distancia de los valores medios en una medida de desviaciones estándar, que se eliminarían un total de 814 observaciones tomando como *outliers* aquellos con un score absoluto mayor o igual a 3.

[1] https://pypi.org/project/missingpy/
[2] https://scikit-learn.org/stable/



Fuente: Elaboración propia en base a Kaggle.

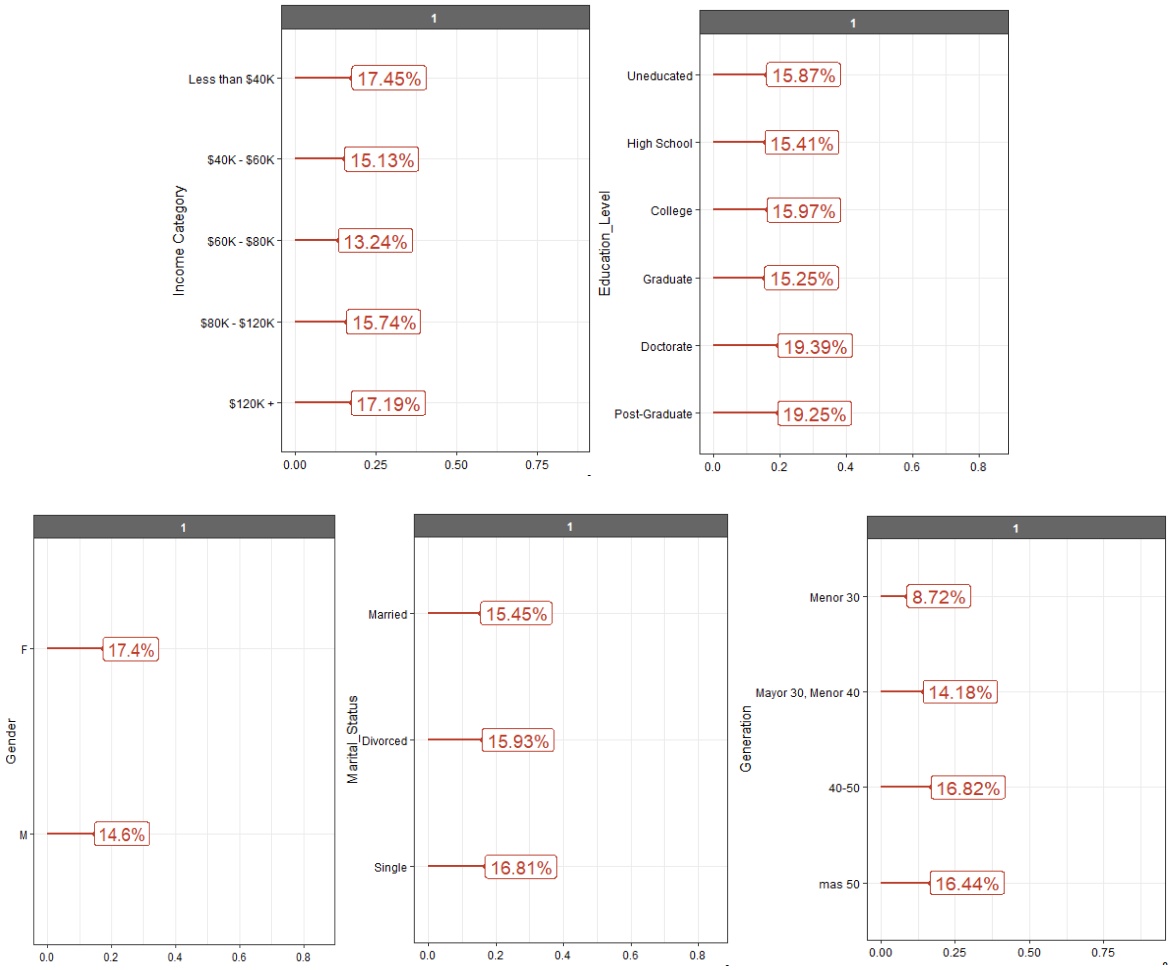
Ahora bien, una vez imputados los valores faltantes, se observaron algunas características de los usuarios en la base de datos. Como deja notar el Gráfico 2a, un 53% de las observaciones son mujeres, y un 47% hombres. El nivel de educación más común entre los clientes bancarios es el de *graduate*, en un 36%. Así mismo, un 50% de los clientes están casados. Por último, se observa que el nivel de ingresos más visto es el de menor a $40 mil dólares anuales, llegando a un 42% de las observaciones de la base de datos.

1.3. Características particulares de clientes que han abandonado el servicio

Al observar los aspectos particulares en el grupo de clientes que han abandonado el servicio ofrecido del banco, se pueden notar diferencias con respecto a aquellos que siguen estando suscriptos. Como enseñan los gráficos 2b a 2f, aquellos clientes con ingresos medios–U$S 60,000 a U$S 80,000– tienen el menor porcentaje de casos de abandono (13%), comparado con un 17% entre aquellos de menor y mayor ingreso. Por otro lado, clientes de sexo femenino tienen dos puntos porcentuales mayores casos de abandono que aquellos de sexo masculino.

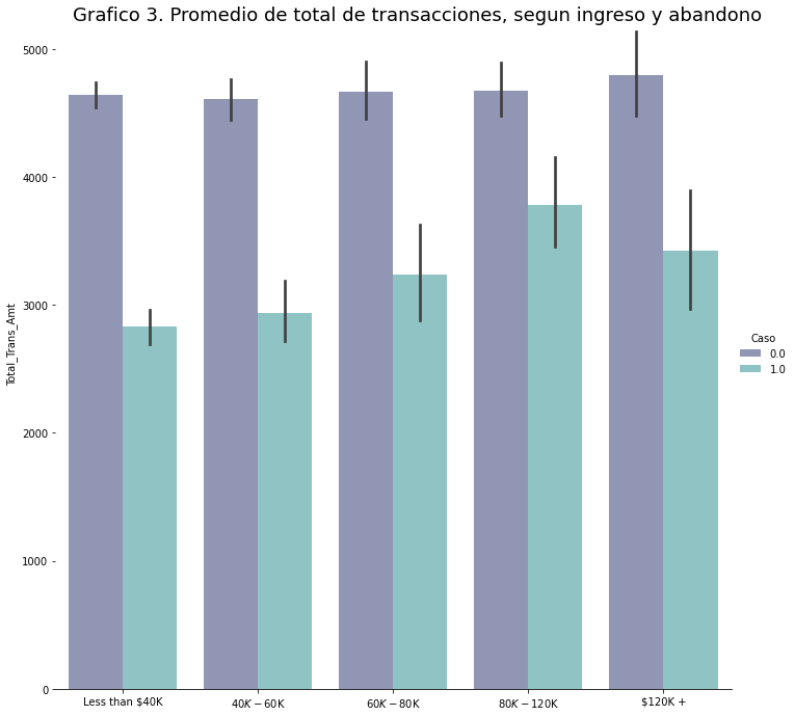
Los menores de 30 años también cuentan con un bajo porcentaje de casos de abandono (9%), comparado con un 17 y 16 por ciento entre aquellos de entre 40 y 50 años y mayores de 50 años, respectivamente. Así mismo, los clientes con mayores niveles de educación fueron los que más dejaron de consumir los servicios ofrecidos por el banco, llegando a un 19%.

Gráficos 2 b-f. Casos de abandono según variable demográfica.



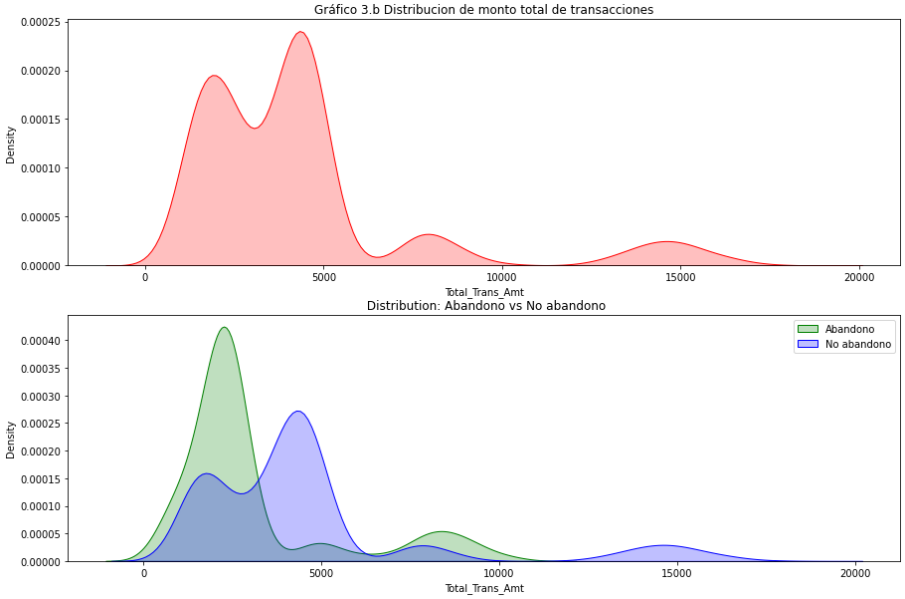
Fuente: elaboración propia en base a Kaggle.

Por otro lado, como se puede notar en los gráficos 3 y 3.b, los clientes que no han sido casos de abandono presentan un mayor total de transacciones en promedio, como podría esperarse. Es decir, una menor cantidad de dinero transaccionado puede ser una variable de interés a la hora de focalizar y predecir que clientes tienen mayor probabilidad de abandonar los servicios. Esto sucede tanto para aquellos clientes de bajos ingresos como de altos.



Fuente: elaboración propia en base a Kaggle.

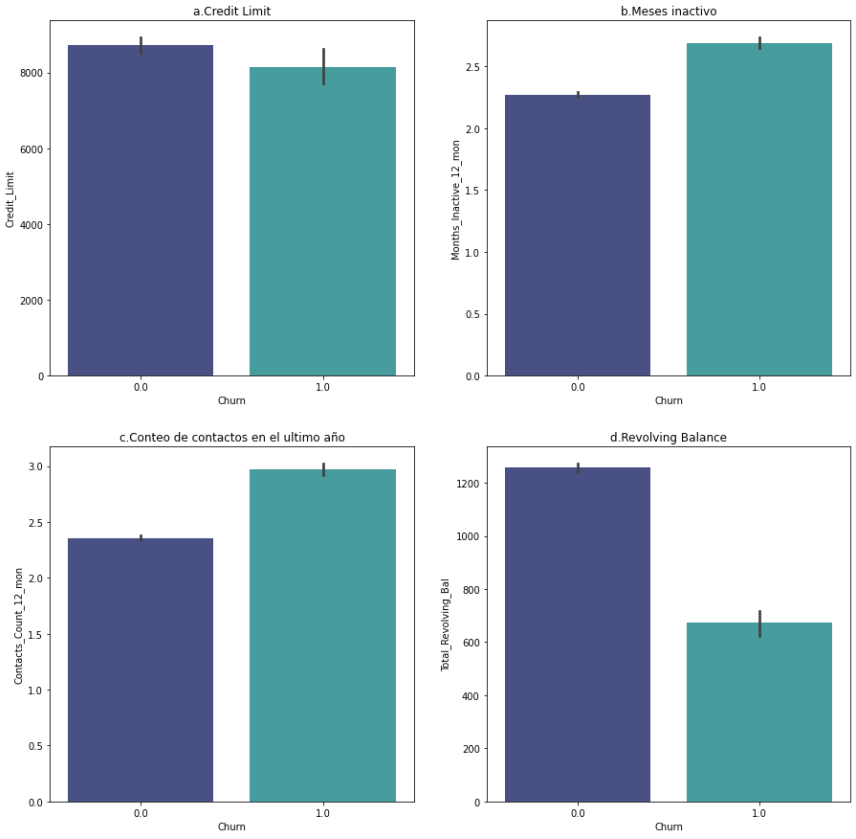
Como enseña el Grafico 3b, la distribución del monto total de transacciones está concentrada en un valor bajo entre los casos de abandono (verde), mientras que la distribución de aquellos que siguen consumiendo el producto bancario (azul) contiene mayor peso en valores más altos.



Fuente: elaboración propia en base a Kaggle.

Al adentrarse en las diferencias entre estos dos tipos de clientes según los valores de variables de uso de los servicios bancarios, se observan diferencias entre los valores promedios de, por ejemplo, la cantidad de meses inactivos. El Gráfico 4.b deja ver que aquellos casos de abandono llevaban una mayor cantidad de meses inactivos– 3, en promedio– que aquellos clientes vigentes– 2,4, en promedio. Resulta importante notar como los clientes que luego abandonaron los servicios también tenían un mayor promedio de conteo de contactos que se les han realizado en los últimos 12 meses (Gráfico 4.c). Esto puede ser una medida del esfuerzo realizado por el banco para evitar que pase lo sucedido.

Graficos 4. a,b,c y d. Promedios de variables financieras



Fuente: elaboración propia en base a Kaggle.

Otra diferencia que puede notarse es aquella relacionada al límite de crédito y *revolving balance*. Los clientes vigentes tienen mayores valores de ambos, en promedio, que los casos de abandono. Mientras que los clientes vigentes mostraron un promedio de $8,727 de límite de crédito, los casos de abandono uno de $8,136. En cuanto al *revolving balance*, uno de $1,257 contra uno de $673,

respectivamente. Esto puede indicar como un mayor uso de la tarjeta de crédito, a pesar de dejar un mayor monto impago a fin de mes, resulta una variable diferenciadora para saber a qué grupo enfocar las campañas de retención.

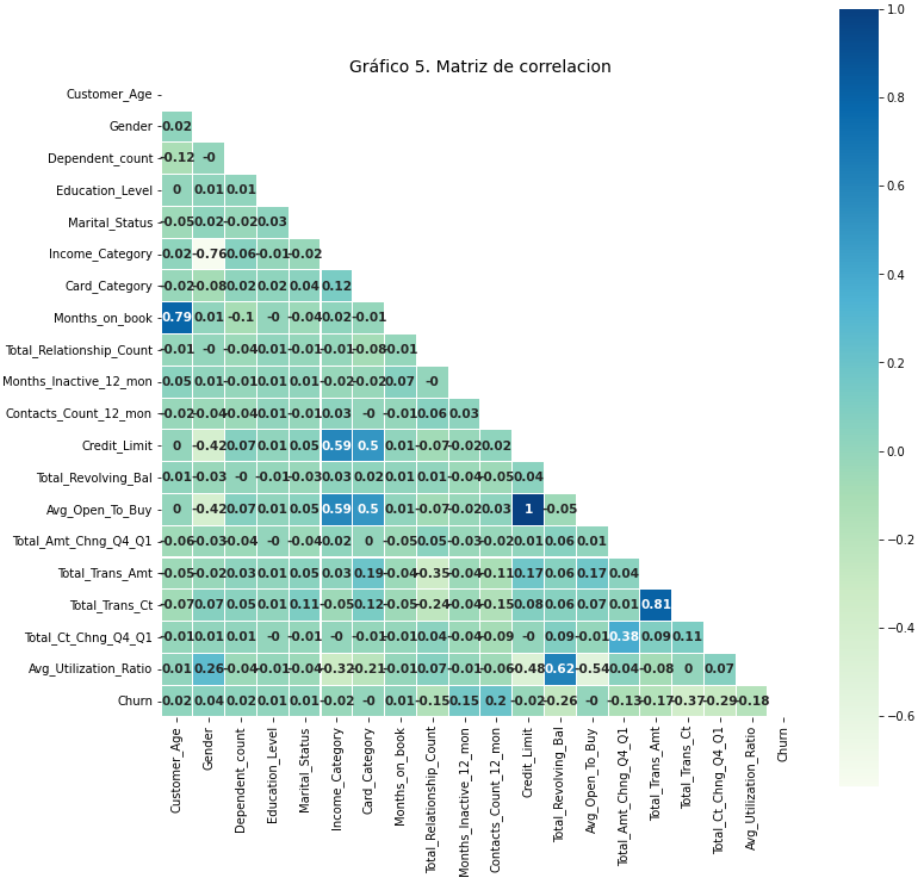
Métodos de análisis de datos multivariantes

Como se ha comentado previamente, se cuenta con una gran cantidad de trabajos que analizan casos de abandono con el objetivo de hacer un análisis lo más exhaustivo y efectivo posible. Para este análisis en cuestión, se tendrá en cuenta el análisis de *clusters*, el análisis de componentes principales¸ el análisis de correspondencia y modelos predictivos como lo son los de árboles de decisión, también en concepto de ensambles.

2.1. Reducción de dimensionalidad a partir de PCA

> "*En muchas ocasiones el investigador se enfrenta a situaciones en las que, para analizar un fenómeno, dispone información de muchas variables que están correlacionadas entre si en mayor o menor grado. Estas correlaciones son como un velo que impiden evaluar adecuadamente el papel que juega cada variable en el fenómeno estudiado. El análisis de componentes principales (PCA) permite pasar a un nuevo conjunto de variables- componentes principales- que gozan de la ventaja de estar incorrelacionadas entre sí y que, además, pueden ordenarse de acuerdo con la información que llevan incorporada*."
>
> - Úriel y Aldás (2017)

El método de *principal component analysis (PCA)* puede ser aplicado a la hora de trabajar con variables métricas que están correlacionadas entre sí. Su objetivo fundamental es condensar la información original en un conjunto más pequeño de variables, llamadas factores o componentes principales, con la menor pérdida de información posible. Para lograr esta reducción de dimensionalidad, el PCA crea combinaciones lineales de las variables originales que son ortogonales entre si (no correlacionadas). Como se observa en el Gráfico 5, se cuentan con variables correlacionadas, ya sea en un menor o mayor nivel.



Fuente: elaboración propia en base a Kaggle.

Las componentes principales son creadas de modo que resulten en una combinación lineal de las variables originales y que la varianza sea máxima, sujeta a la restricción que la suma de los pesos o coeficientes de la combinación lineal al cuadrado sea igual a 1:

$$(1) \quad Z_{1i} = u_{11} X_{1i} + u_{12} X_{2i} + \cdots + u_{1p} X_{pi}$$

Ahora bien, la manera en la que puede ayudar este método de análisis multivariado a la hora de realizar un análisis exploratorio o predictivo de casos de abandono de bancos es tal en que permite reducir la dimensionalidad de las variables explicativas manteniendo la mayor cantidad de varianza posible. Esto puede ayudar a mejorar la performance de los modelos predictivos ya que facilitaría el entrenamiento al tomar menos variables para generar las clasificaciones a las observaciones. Así mismo, generaría un menor nivel de *overfitting* reduciendo variables que realmente no tienen un gran peso en el análisis.

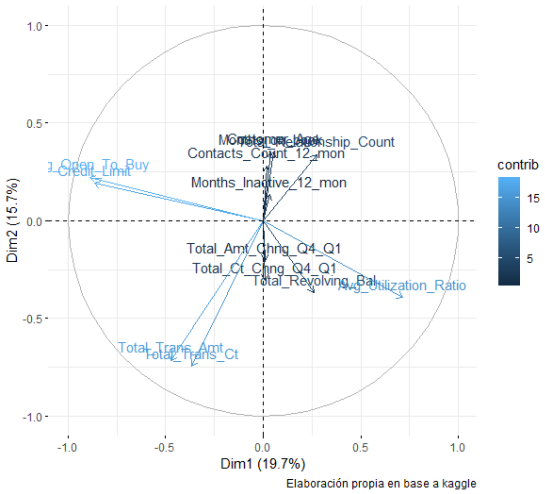
Para este caso particular, previo a utilizar este método se han eliminado las variables categóricas, previamente convertidas en numéricas, ya que el PCA funciona sobre variables continuas (y no tan bien con discretas). Haciendo uso de las librerías *stats*[3] y *factoextra*[4] del programa RStudio, se han calculado los componentes principales, tomando la matriz de correlaciones en vez de covarianzas debido a las diferencias en las métricas de medición de las distintas variables explicativas.

A través de la realización de un *scree plot*, se ha visto que no hay ningún componente que explique una gran cantidad de la varianza de la base de datos, sino qué está dividida equitativamente entre ellos. Luego del quinto componente principal es que se explica el 70% de la varianza. También pueden escogerse estos primeros cinco componentes por el criterio del autovalor mayor a 1. La lógica detrás de este último es que "Si una componente no es capaz de explicar más información que una variable, no va a facilitar la reducción de datos, es decir, facilitar la interpretabilidad de la información" (Úriel y Aldás, 2017). Por lo tanto, solo estos serán considerados en el análisis.

Para entender que significan estos cinco componentes principales, puede visualizarse la contribución de cada variable a cada componente. Cuanto mayor sea la tal llamada carga, más influye en cada componente. La Tabla 3 y el Gráfico 6 enseñan estas, y se puede notar como el primer componente– el cual explica la mayor cantidad de varianza, alrededor de 20% – está influenciado por las variables de límite de crédito y el promedio de la diferencia de este y el balance actual de la cuenta de cada cliente (*Average Open to Buy*), mientras que el segundo con las cantidades y montos totales de transacciones.

[3] https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/princomp
[4] https://cran.r-project.org/web/packages/factoextra/index.html

Gráfico 6. Relación entre primeros componentes y variables



Elaboración propia en base a kaggle

Por otro lado, el tercer componente expresa una relación con las variables que tienen un tinte temporal, como lo son la cantidad de meses suscriptos al servicio y la edad de los clientes. Las variables de variación de cantidad y montos totales de transacciones explican al cuarto componente, mientras que el quinto y último es influenciado mayormente por el *revolving balance*.

Tabla 3. Cargas de variables en componentes principales.

Variables	Componente 1	Componente 2	Componente 3	Componente 4	Componente 5
Customer_Age	-	0.248	0.648	-	-
Months_on_book	-	0.246	0.649	-	-
Total_Relationship_Count	0.173	0.237	-0.161	0.281	-
Months_Inactive_12_mon	-	-	-	-0.1	-
Contacts_Count_12_mon	-	0.197	-0.118	-	-
Credit_Limit	-0.541	0.133	-	0.27	-0.282
Total_Revolving_Bal	0.163	-0.258	0.137	0.375	-0.621
Avg_Open_To_Buy	-0.556	0.156	-	0.236	-0.226
Total_Amt_Chng_Q4_Q1	-	-0.144	-	0.529	0.41
Total_Trans_Amt	-0.295	-0.5	0.201	-0.168	-
Total_Trans_Ct	-0.228	-0.523	0.189	-0.178	-
Total_Ct_Chng_Q4_Q1	-	-0.213	-	0.517	0.405
Avg_Utilization_Ratio	0.448	-0.275	0.109	0.157	-0.35

Elaboración propia en base a kaggle

Para hacer uso de estas nuevas variables o componentes, el método le asigna un valor a cada variable llamado puntaje (*score*). Observando estos puntajes respecto a cada componente para cada cliente, el banco comercial puede llevar a cabo una distinción sobre que variables tienen mayor peso sobre ellos, y así enfocar correctamente sus ofertas. De esta manera, en el tercer Apartado,

donde se realizarán los modelos de predicción, se comparan los modelos originales con los modelos que únicamente tomen a estos cinco componentes generados.

2.2. Comportamiento de clientes a través de *clusters*

"El análisis de conglomerados (*clusters*) tiene por objeto agrupar elementos en grupos homogéneos en función de las similitudes o similaridades entre ellos.", (Peña, 2002). Como explica Peña, mientras en PCA se intenta reducir el número de variables explicativas, en el análisis de conglomerados se intentará agrupar a las observaciones en base a estos componentes descubiertos en base a sus valores a lo largo de distintas variables.

En este caso, se hará este análisis de conglomerados para determinar los grupos que puedan llegar a formarse dentro de los casos abandono. Los grupos serán homogéneos respecto a las variables utilizadas para caracterizarlos, y al mismo tiempo serán lo más diferente posible entre sí. Para esto, se utilizará como medida de similitud a la distancia euclídea:

$$(2) \quad D_{ij} = \sqrt{\sum_{p=1}^{k} \left(X_{ip-X_{jp}}\right)^2}$$

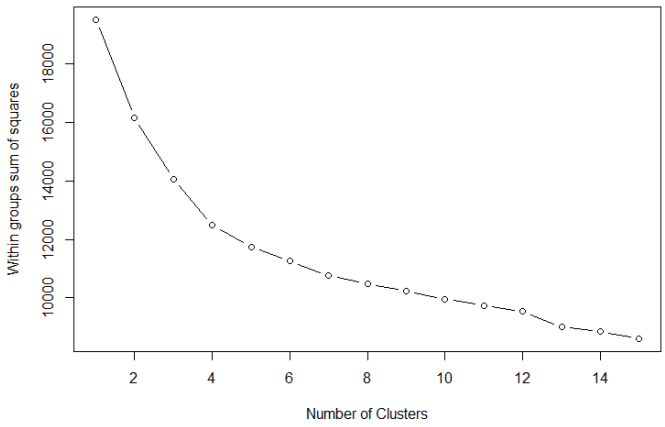
Ahora bien, se cuentan con dos métodos para formar agrupaciones; el jerárquico, y el no jerárquico. En este trabajo se analizará únicamente al segundo, haciendo uso del algoritmo *K-means*. Cada método tiene una manera distinta de determinar el número de *clusters* óptimo:

- *Ward*: minimiza la varianza "dentro" de los *clusters*. $D_{ij} = \left\|x_i - y_j\right\|^2$
- Vecino lejano: la distancia entre dos *clusters* es la máxima distancia entre dos puntos. $\mathrm{Dij} = \max \mathrm{x} \in \mathrm{Ci}, \mathrm{y} \in \mathrm{Cj}\ \mathrm{d}(\mathrm{x},\mathrm{y})$
- Vecino cercano: la distancia entre dos *clusters* es la mínima distancia entre dos puntos. $\mathrm{Dij} = \min \mathrm{x} \in \mathrm{Ci}, \mathrm{y} \in \mathrm{Cj}\ \mathrm{d}(\mathrm{x},\mathrm{y})$
- *Average*: la distancia entre *clusters* es el promedio de las distancias entre dos puntos. $\mathrm{Dij} = \mathrm{suma}\ \mathrm{x} \in \mathrm{Ci}, \mathrm{y} \in \mathrm{Cj}\ \frac{\mathrm{d}(\mathrm{x},\mathrm{y})}{ni \times nj}$

- *K-means*: este método es uno de "reubicación". Cumpliendo una serie de pasos, crea centroides iniciales en base al número de *clusters* que se le indique, y luego asocia a cada observación con el centroide más cercano recalculando cada centroide hasta llegar a converger.

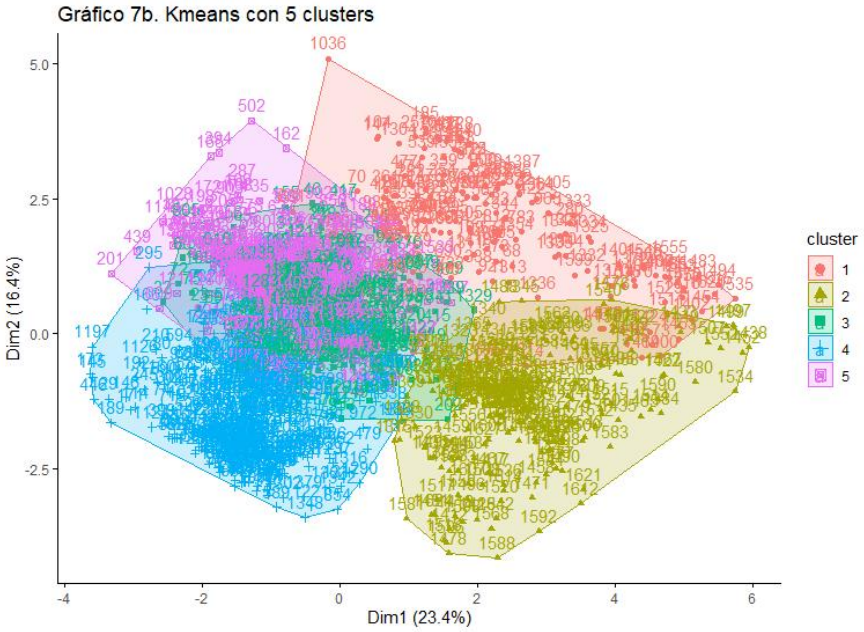
Luego de agrupar a los casos de abandono, estandarizar los datos y dejar de lado a las variables demográficas categóricas para que funcione correctamente el algoritmo, se escogió un numero de cinco *clusters* a analizar, elección basada en el Gráfico 7.

Gráfico 7. Número de clusters segun Kmeans



Elaboración propia en base a Kaggle.

Aunque el Gráfico 7b muestre superposición entre las distintas agrupaciones al reflejar los agrupamientos del *K-means* en un espectro de los primeros dos componentes, este análisis es útil para notar las diferencias entre estos grupos formados.



Elaboración propia en base a Kaggle.

La Tabla 4 muestras las diferencias. El primer grupo contiene un mayor límite de crédito y monto disponible de compra (cuales están altamente correlacionados) que el resto, mientras que el *Cluster* 2 cuenta con los mayores números en cuanto a cantidad y monto de transacciones, como también variación de estos en el último trimestre. El tercer grupo puede ser identificado como aquel que tiene mayor antigüedad en el banco, como también la mayor cantidad de meses inactivo. Así mismo, tienen el menor límite de crédito y cantidad de productos. Siguiendo con el cuarto *cluster*, se nota que estos tienen el mayor *revolving balance*, cual, recordando, era la cantidad de crédito que terminaba impago a fin de periodo. Así mismo, contienen el mayor ratio de utilización, cual muestra que tanto crédito usan según la cantidad disponible. Por último, el *cluster* 5 puede ser descripto como aquel con los menores números en cuanto a totales y cantidad de transacciones, y también menor antigüedad en el banco.

Tabla 4. Promedios por cluster

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Months_on_book	0.16	-0.25	0.58	-0.03	-0.42
Total_Relationship_Count	-0.17	-0.17	-0.18	0.06	0.25
Months_Inactive_12_mon	0.01	-0.06	0.54	0.01	-0.44
Contacts_Count_12_mon	0.09	-0.04	-0.18	-0.06	0.18
Credit_Limit	2.44	0.28	-0.40	-0.43	-0.31
Total_Revolving_Bal	-0.02	0.05	-0.52	1.43	-0.55
Avg_Open_To_Buy	2.43	0.28	-0.34	-0.57	-0.26
Total_Amt_Chng_Q4_Q1	0.06	0.89	0.37	-0.21	-0.59
Total_Trans_Amt	0.19	2.15	-0.37	-0.37	-0.44
Total_Trans_Ct	0.16	1.71	-0.34	-0.25	-0.34
Total_Ct_Chng_Q4_Q1	-0.03	1.12	0.04	-0.13	-0.43
Avg_Utilization_Ratio	-0.53	-0.27	-0.43	1.66	-0.48

Elaboración propia en base a Kaggle.

Estos análisis pueden ser útiles para el banco comercial para poder evaluar los distintos grupos de clientes que han dejado de consumir sus servicios– en este caso, tarjetas de crédito. Este método de análisis multivariado permite una segmentación tal que pueda ayudar a notar qué es lo que le falta al cliente, y hacer enfoque en esto en las campañas de retención. Si un cliente pertenece al *Cluster 1*, por ejemplo, se debería analizar si las ofertas hacia este deberían estar focalizadas en un mayor límite de crédito, o al contrario, en productos que estos no tengan actualmente. Aquellos del *Cluster 2*, al contar con un gran número de transacciones tanto en monto como en cantidad, es posible que su salida se haya debido a un motivo distinto a la operabilidad de sus pagos. Por esto, por ejemplo, podrían ser ofertados mejores tasas de interés para retener su dinero antes de ser transaccionado.

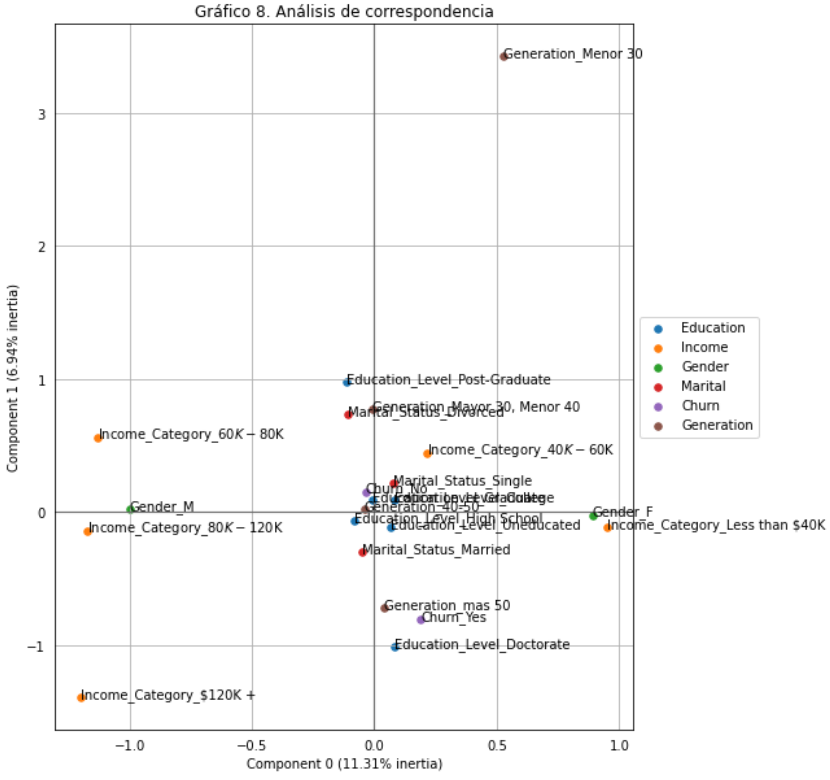
2.3. Análisis de correspondencia

Otro método utilizado en esta base es el Análisis de Correspondencias. El objetivo es usar esta técnica descriptiva para representar tablas de contingencia y analizar la frecuencia de las variables cualitativas. Esto permite resumir la información representándola en un espacio o dimensión menor, ayudando a observar de manera sencilla la relación entre casos de abandono y variables demográficas. Para comenzar con el método se parte de una tabla de frecuencias que, en este caso,

representa variables tales como el género, nivel de ingresos, estado civil, generación y educación, cruzando por la variable abandono. A partir de allí se calculan los perfiles de las filas y columnas (total de casos), y la masa que corresponde al peso que representa la cantidad de observaciones. Por otra parte, se calcula la inercia que mide la dispersión de los perfiles en el espacio. El objetivo es que la nueva representación mantenga las diferencias χ^2 relativas entre ellos lo más inalteradas posibles.

"En el análisis de correspondencias, el mapa mostrará las distancias entre los distintos niveles de dos variables no métricas, por lo que suele decir que el análisis de correspondencias sirve para visualizar tablas de contingencia", (Úriel y Aldas, 2017). De esta manera, podemos ver de forma gráfica (ver Gráfico 8) las distancias relativas de distintas variables demográficas a los casos de abandono. Por ejemplo, se nota que la variable Generation_Mas_50, cual como dice el nombre, engloba a las observaciones con edad mayor a 50 años, presenta una cercanía a Churn_Yes, variable de caso de abandono positivo.

Gráfico 8. Análisis de correspondencia



Elaboración propia en base a Kaggle.

Similarmente, aquellos con la más alta educación (doctorado) y aquellos con los más altos ingresos (mayores a U$S 120.000) también presentan cercanía al punto Churn_Yes. Lo mismo sucede, aunque en menor medida, con los clientes casados. Por otro lado, aquellas demográficas más cercanas al punto Churn_No son, por ejemplo, los clientes de estado civil solteros, aquellos con estudios universitarios (*College*) o con menor educación, de edad 40 a 50 años, y de ingresos más bajos.

Modelos predictivos de abandono de clientes bancarios

"The easiest way to make churn predictions is to observe customers' behavior and to create, with the help of experience, some rules that classify a customer as churner. For example, a bank could label as churner a user that has not made transactions for a long time and that has a low account balance. However, all these rules are created without a scientific method, using only experience and intuition, so the results may be below the expectations. A powerful method is required to make forecasts more reliable than those based just on experience. An effective alternative is Data Mining, which is an automatic discovery process of interesting information applied to large data stored in appropriate repositories as databases, data warehouses or files. One task is to analyze historical data in order to create a mathematical model (which conceptually represents the real world under investigation) and, later, to make predictions on recent data based on this model. Therefore, it perfectly fits our objective of understanding customers data to make churn prediction."

- Avon, 2015

Una gran cantidad de trabajos han sido realizados con el objetivo de predecir de la mejor manera qué variables explican en mayor proporción la decisión de un cliente de abandonar un banco. Para esto, diversos modelos estadísticos han sido utilizados. Prasad (2012), por ejemplo, utiliza un modelo de clasificación y regresión de árboles (CART). Kaur, et al, (2013) hace uso de los modelos Naive Bayes, arboles de decision y *support vector machine*. Autores como Zoric (2016) utilizan modelos de redes neuronales artificiales a la hora de predecir abandonos en un banco comercial.

Otros estudios como el de Chiang, et al (2003) obtuvo resultados de predicción de casos de abandono utilizando un modelo de patrones secuenciales, por ejemplo. Zhao, et al (2005) realizó un modelo SVM (*support vector machine*) para explorar también la tasa de abandono de clientes, y Lemmens (2003) el uso de árboles de clasificación, también para observar los potenciales del *data mining* para estos casos de estudio. Como estos, existen diversos casos que intentan explicar de la mejor manera las variables con mayor peso a la hora de detectar clientes que dejan de consumir servicios en el sector bancario.

Por último, también cabe destacar que los modelos de predicción de casos de abandono o *churns* también pueden tener un aspecto de series temporales. Dado que lo importante para los bancos e instituciones es predecir que clientes tienen mayor probabilidad de abandonarlos, la performance de una predicción solo puede verse luego de un tiempo determinado, el cual dependerá de cómo decida medirlo cada uno. Esto no significa que un modelado en una base de datos de corte transversal no sea lo suficientemente completa, ya que pueden ser útiles para entender los comportamientos de los clientes que ya han abandonado los servicios ofrecidos, que probablemente se vean repetidos en clientes que aún no lo han hecho.

3.1. Desarrollo de modelos de aprendizaje automático

El problema de predicción a enfrentar en la base de datos en cuestión es uno de clasificación binaria– abandono o no abandono. Como se vio en los primeros apartados, únicamente se cuenta con un 16% de la muestra considerados como clientes que han abandonado los servicios del banco. Esto genera un problema llamado *class imbalance*. Esto significa que el número total de una clase de datos es bastante menor que el de otra clase. En otras palabras, la base de datos muestra una subrepresentación de casos de abandono.

Esto puede traducirse en un problema a la hora de generar modelos de predicción, ya que los modelos de *machine-learning* trabajan mejor cuando las distribuciones de las clases de datos son lo más parecidas posibles. Cuando el problema de clasificación contiene una mala distribución de la tal llamada *minority class*– en este ejemplo, los casos de abandono–, la performance del modelo

puede no ser lo suficientemente alta, otorgando errores del lado de la mala clasificación, o modelos de alta *accuracy* y bajo *recall*.

Para esto, se utilizó el algoritmo de *oversampling* conocido como SMOTE, de la librería *imblearn*, o Synthetic Minority Oversampling Technique.[5] En pocos pasos, SMOTE funciona de la siguiente manera: en principio, identifica un punto de la clase minoritaria. Luego, identifica sus vecinos más cercanos aleatoriamente. Finalmente, crea una observación sintética que se encuentre entre estos puntos en el espacio que los conecta.[6] Es importante recalcar que SMOTE debe ser aplicado sobre los datos de entrenamiento, y no sobre todo el *dataset*. Solo debe servir para que el modelo sea mejor entrenado. Las diferencias entre los resultados de los modelos que toman en cuenta este *oversampling* y aquellos que no serán mencionados en el próximo subapartado.

Así como se nota la existencia del algoritmo SMOTE, también existen otras técnicas para solucionar el desbalance de clases. Las más comunes, *oversampling* y *undersampling*. Estos son comúnmente generados aleatoriamente. La diferencia entre estos dos es que, mientras el primero duplica observaciones de la clase minoritaria– en nuestro ejemplo, casos de abandono–, el *undersampling* elimina observaciones de la clase mayoritaria. Se ha escogido el algoritmo SMOTE sobre estos por distintos motivos. Al reducir el número de observaciones en la base de datos, el método de *undersampling* se traduce en una pérdida general de información. Dado el numero reducido de clientes en la base de datos utilizada, este no fue tomado como una opción. Por otro lado, el método de *oversampling* es considerado como una forma *naive* de agrandar la clase minoritaria, debido a que no aumenta la variedad del set de entrenamiento, sino que simplemente duplica los casos existentes. Mientras que SMOTE debe ser utilizado con cuidado, este no solo agranda el set de entrenamiento, sino que también le aumenta la variedad de las observaciones. Al hacer esto en base a sus características en cuanto a los atributos en cuestión, permite englobar casos de abandono que, debido a la pequeña muestra obtenida, pueden no estar siendo considerados.

[5] https://towardsdatascience.com/class-imbalance-smote-borderline-smote-adasyn-6e36c78d804
[6] Un problema que puede surgir de SMOTE es que este elija puntos que sean outliers, y luego los multiplique artificialmente cuando cree sus observaciones sintéticas.

Para resolver este problema de clasificación se tomaron en cuenta distintos modelos. Uno de los modelos de aprendizaje más conocidos es el de **Decision Trees**. Estos son fáciles de interpretar ya que ayudan a revelar las variables más importantes a la hora de generar la predicción. El modelo generado forma nodos que representan una prueba para cada variable particular. El resultado de esta prueba divide a los registros en subgrupos, creando nuevos nodos consecuentemente del a misma manera. Una vez que un criterio es formado, el nodo deja de dividirse y se genera una "hoja". Entonces, cada camino desde la "raíz" hacia cada hoja representa una regla de decisión. Este modelo es considerado útil ya que funciona para conceptualizar relaciones no lineales entre distintas variables. Sin embargo, este modelo de árboles de decisión es considerado como volátil, ya que sufre ante modificaciones o actualizaciones en las bases de datos. En otras palabras, es probable que en algunos casos haya *overfitting*. (Witten, et. al 2016)

Similarmente, se realizarán predicciones utilizando el modelo **Random Forest.** Este es un modelo de ensamble en el que diferentes clasificadores generan un mejor modelo como un conjunto. Utiliza la técnica de *bagging*, la cual sirve para crear diferentes particiones de datos de entrenamiento. Esto soluciona en gran parte el mayor problema de los árboles de decisión, ya que como se mencionó previamente, estos últimos son volátiles y de alta dependencia del set de datos que se utilice para su entrenamiento. "El nombre *Bagging* proviene de la abreviatura de *Bootstrap AGGregatING*. Como su nombre lo indica, los dos ingredientes clave de *Bagging* son *bootstrap* y combinación. Por lo general, esto implica el uso de un solo algoritmo de aprendizaje automático, casi siempre un árbol de decisiones, el cual es entrenado utilizando diferentes conjuntos de datos de entrenamiento. Así se generan N arboles completamente expandidos y las predicciones que realizan cada uno de los miembros del conjunto se combinan usando alguna estadística simple: votaciones, promedios, promedios ponderados, etc. La clave del método es la forma en que se genera cada conjunto de datos para entrenar cada modelo. Aquí, cada modelo obtiene su propio conjunto de datos utilizando el método de *bootstrap*. Esto quiere decir que una fila dada puede estar presente en un conjunto de datos cero, una o múltiples veces", (Santiago, 2021).

Al generar distintos sets de entrenamiento para el mismo modelo, Random Forest logra reducir la volatilidad y mejorar la performance predictiva. Para esto, genera un sistema de votos en el cual

cada árbol "presenta" sus reglas, y aquellas que aparezcan en mayor cantidad son las escogidas para el modelo final. De esta manera, reduce la probabilidad de *overfitting* y la varianza entre los resultados. Una de las desventajas, sin embargo, es el mayor tiempo computacional relativo a otros modelos que no precisen de ensambles.

Otra técnica por utilizar es la de **Boosting**. Este es un algoritmo de aprendizaje automático que ayuda a mejorar las performances de los modelos. Como *bagging*, utiliza el sistema de votación para clasificaciones, combinando distintos modelos del mismo tipo. Mientras que en *bagging* cada modelo es construido por separado, en *boosting* estos modelos son preparados teniendo en cuenta la performance de sus pares. "*Boosting encourages new models to become experts for instances handled incorrectly by earlier ones. A final difference is that boosting weights a model's contribution by its performance rather than giving equal weight to all models"*. (Witten, et. al 2016). Siguiendo a Witten, así como la técnica les aplica una ponderación a los modelos por su buena performance, en el mismo sentido los penaliza por su error de clasificación. Un ejemplo de este tipo de modelos puede ser el de **AdaBoosting**, el de **GradientBoosting** o **XGBoost**.

La **regresión logística** también es de suma utilidad a la hora de generar modelos clasificatorios. A diferencia de una regresión lineal, la regresión logística funciona correctamente para modelos que estudian resultados binarios. Esta funciona interpretando la probabilidad de obtener un resultado igual a 1 (tomando a 1 como el caso positivo, como por ejemplo "abandono"), prediciendo si la probabilidad es mayor a la de un umbral. La manera de transformar los datos para su mapeo es la función de Sigmoid:

$$(3) \qquad Sigmoid\,(z) = \frac{1}{1 + e^{-z}}$$

La regresión logística es usada en *data mining* debido a su facilidad de realización y su capacidad de obtener relaciones no lineales en los datos, gracias a esta transformación.

Como estos modelos brevemente descriptos, también existen una gran cantidad de otros con el fin de predecir problemas de clasificación. Así mismo, estos modelos pueden mejorar su performance a través de una optimización de sus parámetros, o *hyperparameters tuning*.

3.2. Resultados generales

Para evaluar los modelos, se hará uso de la métrica *Area Under the Curve*, mayormente conocida como AUC. Es únicamente aplicable a problemas de clasificación binarios, ya que tiene en cuenta casos "positivos" o "negativos". Antes de indagar en la definición de esta métrica, es necesario explicar otros conceptos como lo son la *Sensitivity* y *Specificity*. Obsérvese la siguiente matriz:

Figura 2

		Clase predicha	
		Positivo	Negativo
Clase real	Positivo	True Positives	False Negatives
	Negativo	False Positives	True Negatives

La *Sensitivity* o *Recall* nos muestra la proporción de clases positivas que fueron correctamente asignadas:

$$(4) \quad Sensitivity = \frac{TP}{TP + FN}$$

En este ejemplo, de todos los casos de abandono, la proporción que ha sido predicha correctamente. Por otro lado, la *Specificity* nos enseña la proporción de clases negativas correctamente predichas:

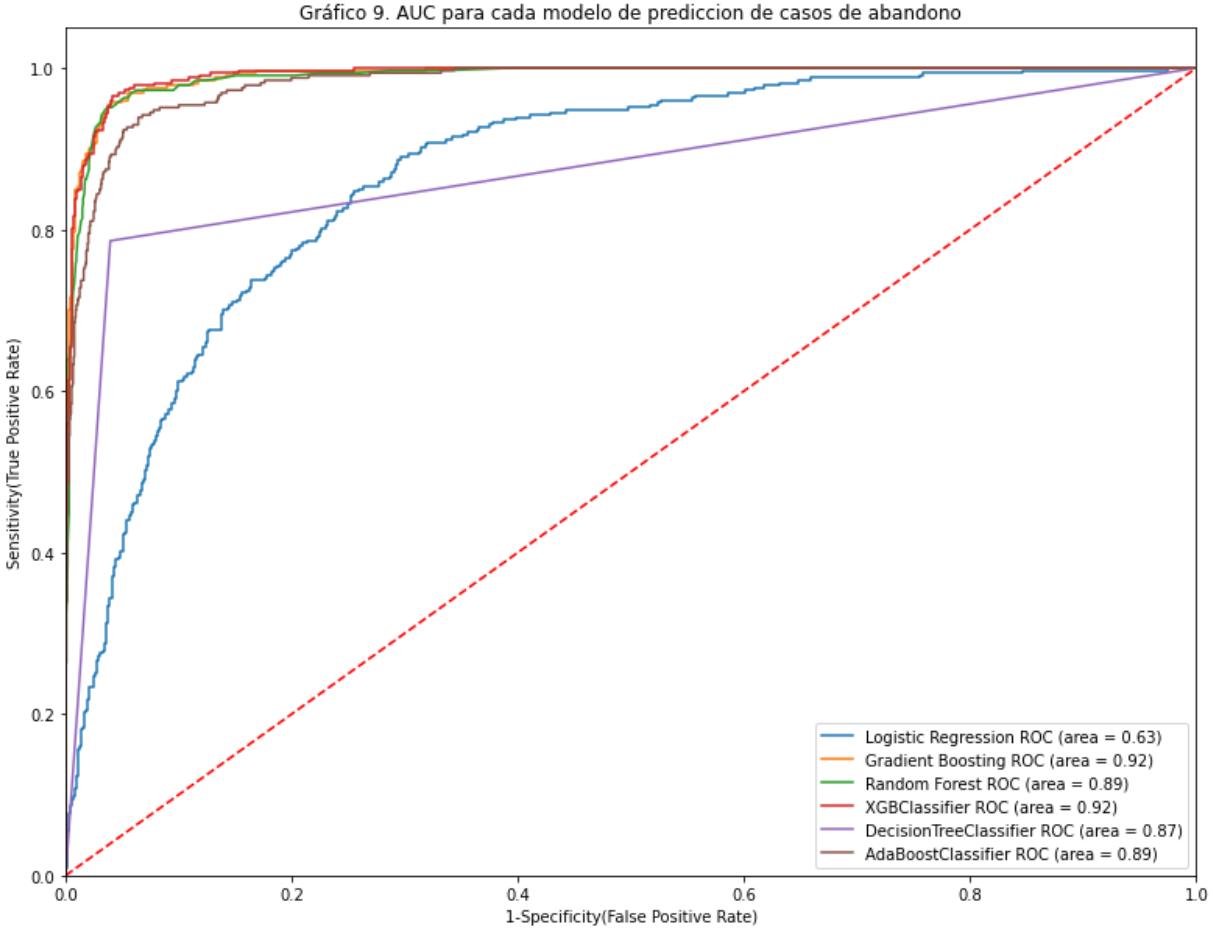
$$(5) \quad Specificity = \frac{TN}{TN + FP}$$

De estas dos se puede obtener una tercera métrica, llamada *False Positive Rate*, conocida también como $1 - Specificity$. Esta dice la proporción de casos negativos que fue incorrectamente asignada por el modelo. En este caso, serían aquellos casos que no fueron de abandono pero que si han sido pronosticados como tales.

Ahora bien, la curva ROC-AUC enseña el grado de *Sensitivity* y $1 - Specificity$ en las predicciones. En otras palabras, qué tan bien discrimina el modelo de clasificación entre las dos clases. Muestra cuántas clasificaciones positivas correctas se pueden obtener a medida que permite más y más falsos positivos.

Definida esta métrica, pueden empezar a compararse los resultados de los modelos realizados. Como enseña el Gráfico 9, se han generado distintos modelos de clasificación. El de mejor performance en términos de esta métrica es al de *Gradient Boosting Classifier* y al de *XGBoost Classifier,* con un AUC de 0,92. El modelo con menor AUC fue el de Regresión Logistica (0,63). *Random Forest* entregó un mejor puntaje que el de Árboles de decisión, siendo 0,89 y 0,87, respectivamente. Por último, el modelo *AdaBoost* otorgó un AUC de 0,89.

Gráfico 9. AUC para cada modelo de prediccion de casos de abandono

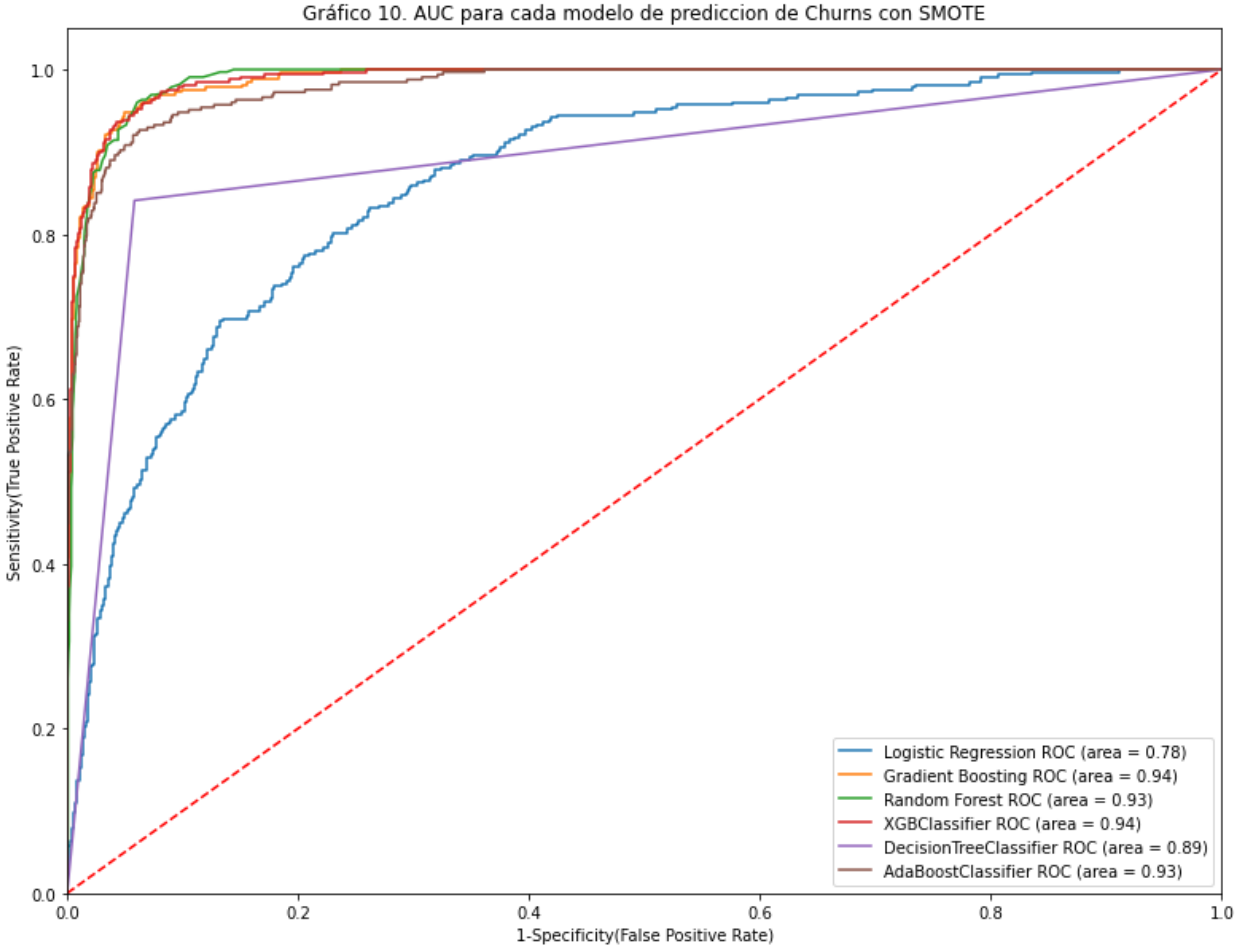


Fuente: Elaboración propia en base a Kaggle.

Como se comentó previamente, el problema de la base de datos utilizada es que presenta un gran desbalance. Por este motivo, se puso en función a la herramienta SMOTE para generar un *oversampling* de la clase minoritaria (abandonos). Este algoritmo fue aplicado únicamente a la partición de entrenamiento. El entrenamiento (80% de los datos) paso de contener 8.101 a 13.602 observaciones. En este set nuevo de entrenamiento, el balance de casos positivos y negativos es el mismo (50% ambos).

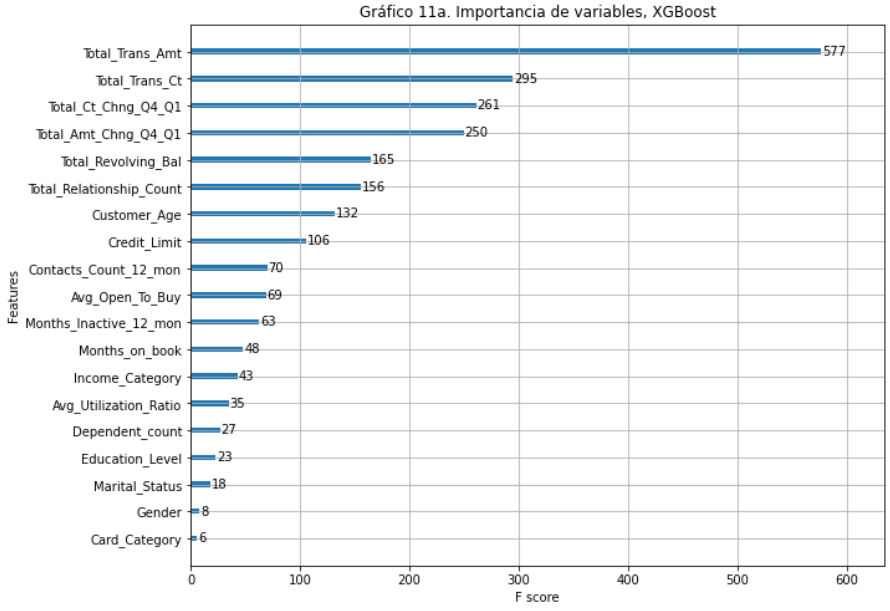
El Gráfico 10 deja notar las mejoras en cuanto a AUC de todos los modelos realizados. El mayor salto de vio en el modelo de Regresión Logística, pasando a un área bajo la curva ROC de 0,78. En cuanto al resto de los modelos, se puede notar por ejemplo el salto del AUC del modelo *Random Forest* de 0,89 a 0,93. El modelo *AdaBoost Classifier* mejoró de 0,89 a 0,93, como también lo han hecho los modelos de *Gradient Boosting* y *XGBoost*, pasando de 0,92 a 0,94 ambos.

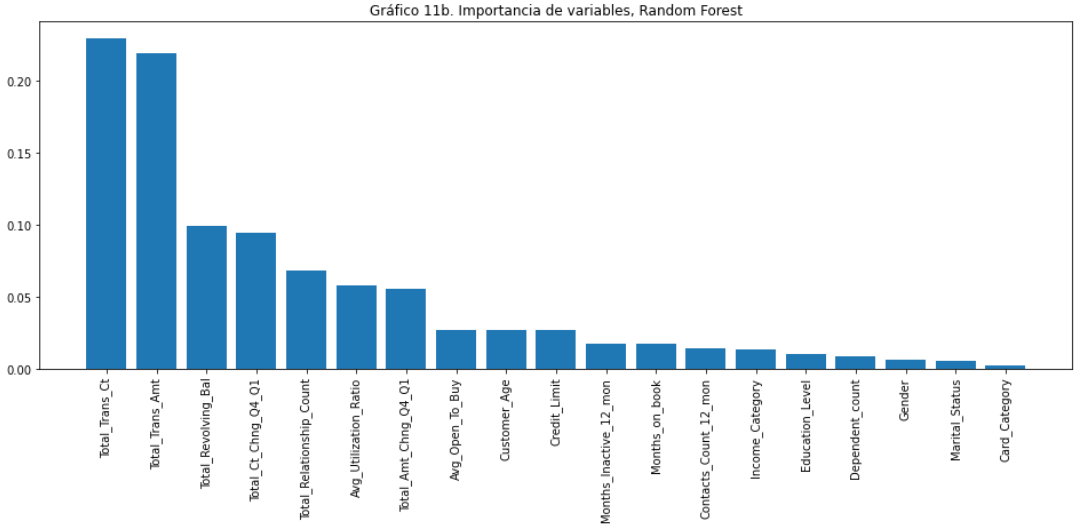
Gráfico 10. AUC para cada modelo de prediccion de Churns con SMOTE



Fuente: Elaboración propia en base a Kaggle.

Como también fue mencionado previamente, una ventaja de estos modelos de clasificación que utilizan arboles de decisión como fundamentación es que permiten observar las importancias relativas de cada *feature* o variable considerada. Los Gráficos 11a y 11b dejan ver como las variables *Total_Trans_Amt* y *Total_Trans_Ct* son las de mayor relevancia a la hora de generar las predicciones de abandono. La primera es la de mayor peso en el modelo *XGBoost*, mientras que la segunda la de mayor peso en el modelo *Random Forest.* El total de monto impago a fin de mes (*Total Revolving Balance)* tiene una mayor importancia en el modelo *Random Forest*, mientras que el modelo *XGBoost* presta más atención a los cambios en las cantidades y montos de transacciones realizadas.





Fuente: Elaboración propia en base a Kaggle.

Otra ventaja del modelo *XGBoost* es que nos permite ver las probabilidades explicitas de ser casos de abandono para cada una de las observaciones. La Tabla 5 fue construida tomando un promedio para cada una de las combinaciones demográficas– en este caso, Educación, Género y Estado Civil. De esta manera, se puede notar que la combinación demográfica Femenino-Post-Graduate-Casado tiene la mayor probabilidad de ser un caso de abandono (0,28), mientras que los clientes de sexo Masculino, sin educación y divorciados, los de menor probabilidad (0,10).

Tabla 5. Probabilidades de abandono según combinaciones demográficas. Modelo XGB.

Combinación demográfica	Probabilidad	Combinación demográfica	Probabilidad
F-Post-Graduate-Married	0.28	F-College-Single	0.19
M-Doctorate-Single	0.26	M-Doctorate-Married	0.19
M-Doctorate-Divorced	0.25	F-Uneducated-Divorced	0.19
F-Post-Graduate-Single	0.23	F-High School-Single	0.19
M-College-Divorced	0.22	M-Post-Graduate-Single	0.19
M-Post-Graduate-Married	0.22	M-High School-Married	0.18
M-Graduate-Divorced	0.21	F-College-Married	0.18
F-Post-Graduate-Divorced	0.21	M-Uneducated-Single	0.18
M-College-Single	0.20	M-High School-Single	0.18
F-High School-Married	0.20	F-Graduate-Divorced	0.18
F-Doctorate-Married	0.20	F-High School-Divorced	0.17
F-Doctorate-Single	0.20	M-Post-Graduate-Divorced	0.17
F-Graduate-Married	0.20	M-Graduate-Single	0.17
F-Uneducated-Single	0.20	M-Graduate-Married	0.16
F-Graduate-Single	0.20	M-High School-Divorced	0.16
F-Doctorate-Divorced	0.19	F-Uneducated-Married	0.15
M-Uneducated-Married	0.19	F-College-Divorced	0.12
M-College-Married	0.19	M-Uneducated-Divorced	0.10

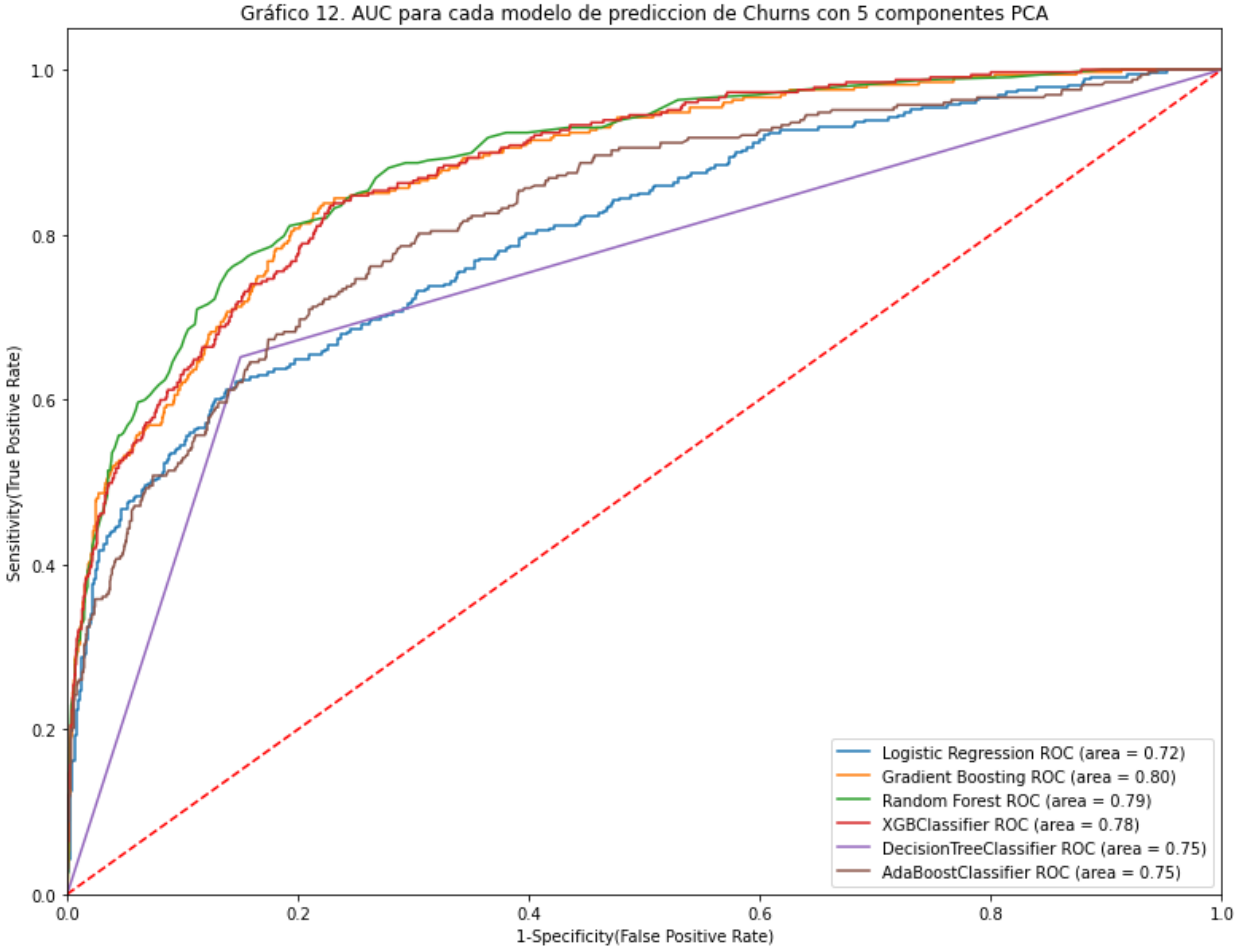
Fuente: Elaboración propia en base a Kaggle.

Este enfoque demográfico puede ser de gran utilidad para los bancos comerciales, ya que suma su valor agregado en términos de características particulares de sus clientes. Ahora no solo cuentan con información acerca de qué motivos o variables pueden afectar a los clientes, sino qué tipo de clientes son más proclives a finalizar los servicios con ellos. Esto, a la hora de decidir cómo y sobre quién enfocar las campañas de retención, se puede traducir en un gran ahorro de tiempo y recursos, y en una mejora en términos de mantenimiento de clientes.

Volviendo al análisis de componentes principales, se pueden utilizar los componentes obtenidos para reemplazar las variables originales. De esta manera, la dimensión se vería reducida y podría ayudar a la predicción del modelo. Sin embargo, como muestra el Gráfico 12, los puntajes de los modelados en cuanto a AUC se ven reducidos. El modelo con mayor área bajo la curva ROC sigue siendo el de *Gradient Boosting* (0,80), seguido por *Random Forest* (0,79) y *XGB* (0,78).

Posiblemente, esta reducción de performance se debe a que la correlación entre las variables de la base de datos no es tan grande como para englobar una gran proporción de la varianza total. En otra base de datos con una varianza menos repartida entre una menor cantidad de componentes, el resultado puede llegar a ser el opuesto.

Gráfico 12. AUC para cada modelo de prediccion de Churns con 5 componentes PCA



Fuente: Elaboración propia en base a Kaggle.

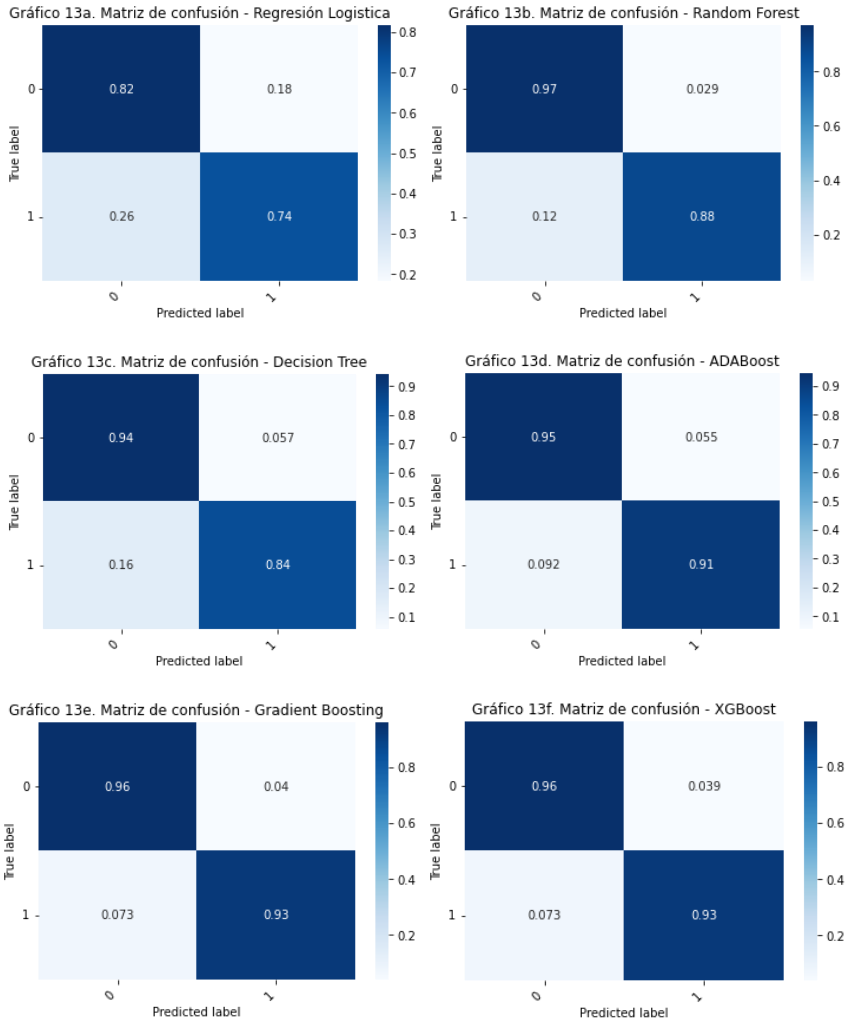
Por último, también se han analizado los resultados de estos modelos de clasificación previamente estandarizando la base de datos. Sin embargo, no se notó diferencias con los resultados sin esta transformación.

3.3. Elección del mejor modelo predictivo

> "*Según el teorema de no free lunch (NFL), no existe un algoritmo universal que funcione bien para cualquier conjunto de datos y para cualquier problema. Esto quiere decir que nuestros modelos aprenden en el contexto de suposiciones que hacemos sobre los datos. Si las suposiciones son correctas, entonces nuestros modelos pueden hacer mejor sentido de los datos. Si nuestras suposiciones son incorrectas, en el mejor de los casos tendremos un modelo que genera predicciones con un grado de incertidumbre grande; en el peor de los casos, tendremos un modelo muy confidente generando las predicciones incorrectas*"
>
> -Santiago (2021)

Para adentrarse en los resultados de cada modelo de predicción, una manera de hacerlo es observando las matrices de confusión. Estas permiten ver el nivel de predicción para cada clase particular. Como dejan ver los Gráficos 13a a 13f, para todos los modelos la predicción de la clase abandono (representada con un 1) es peor que para aquellos clientes vigentes. Esto sucede debido a la gran sobrerrepresentación de clientes del segundo tipo en la base de datos.

El modelo de regresión logística, por ejemplo, solo es eficiente en el 74% de los casos de abandono. Es decir, hay un 26% de casos reales de abandono que no son predichos por el modelo. Por el otro lado, este porcentaje de valores correctos mal predichos se reduce a un 18% para los casos de clientes que no dejan de consumir las tarjetas de crédito del banco. Siguiendo con el modelo *Decision Tree*, este valor de casos de abandono correctamente predichos salta a un 84%, y a los clientes vigentes en un 94%. Sin embargo, este no es tan alto como el observado, por ejemplo, en el modelo *Random Forest,* cual llega a un 88%, y "no abandono" a un 97%.



Fuente: Elaboración propia en base a Kaggle.

Esta medida de correctas predicciones de los casos positivos (abandono) también es conocida como *Recall* – medida en que el modelo identifica *True Positives.* En este caso, es de las medidas más importantes de evaluación de modelos. No resulta por igual a un banco comercial predecir correctamente a un cliente vigente que predecir a un cliente que va a abandonar sus servicios. Si identifica como abandono a un cliente que no iba a abandonar, el problema sería que realiza una política de retención sobre un cliente que no la necesita realmente. Pero si no predice correctamente a un caso de abandono, entonces no aplicaría una política de retención a un cliente que sí la necesita.

Siguiendo con estas métricas, el modelo *ADAboost* supera la barrera del 90% en la proporción de correctas predicciones de casos de abandono, y presenta un 95% de eficacia en predicción de clase vigente. Por otro lado, los dos mejores modelos en cuanto a AUC (*Gradient Boosting* y *XGBoost),* predicen un 93% de los casos de abandonos y 96% de clientes vigentes correctamente, ambos por igual.

La precisión de un modelo, similarmente, muestra el ratio de todos los *True Positives* sobre todos los positivos. En este caso, resulta más importante la medida de *Recall.* Sin embargo, si el problema fuera de otro tipo, como por ejemplo decidir si se le debiera otorgar un préstamo a un cliente, resulta útil observar una alta precisión de los modelos, ya que perder un cliente por declinarle un préstamo es una pérdida de ingresos para los bancos comerciales.

Un último paso por realizar para analizar qué modelo es el mejor para predecir los casos de abandono para este problema específico, es observar cómo se comportan los resultados cuando se cambian los sets de entrenamientos utilizados. Para esto podemos realizar un análisis de *Cross Validation,* o validación cruzada. Esta consiste en realizar múltiples particiones de los datos para luego estimar una métrica de cada una de ellas. Finalmente, se promedian los resultados. Para este caso, se utilizará *K-Fold Cross Validation*. De esta manera, generando cinco particiones distintas para generar los datos de entrenamiento, se puede evaluar las diferencias de performance obtenidas según estas.

Para el modelo que utiliza *Logistic Regression,* se nota un AUC promedio de 0,83. Sin embargo, el desvío estándar de estos resultados es tan grande como 0,13. Lo que habla de un posible *overfitting*. Para *Random Forest,* el promedio fue de 0,87, con un desvío de 0,11. Para el modelo *XGB,* el AUC promedio de las cinco iteraciones fue de 0,88, con un desvio de 0,10. Por último, para los modelos *Decision Tree, ADAboosting* y *GradientBoosting,* el promedio de AUC fue de 0,81, 0,87 y 0,87, respectivamente. Y el desvío, de 0,13, 0,11 y 0,11, respectivamente.

Observando estas diferencias y puntajes entre y de los modelos, se advierte que todos estos están sujetos a sufrir un *overfitting.* Esto puede verse en el alto desvío estándar de los resultados extraídos de *Cross Validating,* y puede deberse a la baja cantidad de observaciones con la que

cuenta la base de datos sobre la cual se construyen los modelos. Empero, se considera a los modelos *Gradient Boosting* y *XGB* como aquellos que obtuvieron los mayores puntajes en términos de AUC, *Recall* y desviación estándar de *Cross Validation*.

Conclusión

El objetivo de este trabajo ha sido el de realizar un análisis exhaustivo de los casos de abandono en un banco comercial particular. Luego de mencionar la importancia que tiene este tipo de análisis para estas instituciones financieras, se ha hecho uso de distintos métodos de análisis multivariado de datos para poder obtener perspectivas tanto de que tipo de clientes son más proclives a abandonar los productos bancarios como de que tipo de variables resultan importantes para poder predecir estos casos.

En el primer apartado, se ha presentado la base de datos en cuestión, sumado a una descripción general de la literatura sobre el tema y de la importancia de los tal llamados *churns*– casos de abandono– en el sector bancario. Se han expuesto tanto los beneficios como los costos de estos generando o no este tipo de análisis dentro de sus instituciones.

En el segundo apartado, se han aplicado diversos tipos de análisis estadísticos multivariantes para poder observar comportamientos de este tipo de clientes que no puedan ser notados a simple vista. Se comenzó con el análisis de componente principales, cual ayuda a descomponer las variables explicativas de los modelos en componentes que no tienen correlación entre sí, logrando reducir la dimensionalidad de la base de datos, manteniendo en lo posible la mayor cantidad de varianza general. De esta manera, se consiguió un agrupamiento de variables en componentes que logren explicar alrededor del 70% de la varianza total, y se han expuesto las descripciones y relaciones de estos componentes con las variables originales. Luego, un análisis de *clusters* ha ayudado a separar a los clientes que han abandonado el banco en cinco grupos a través del algoritmo *k-Means*, con diferencias en valores promedios de distintas variables. Por último, un análisis de correspondencias ha otorgado un análisis adicional de las características demográficas de los clientes, permitiendo notar cuales de estos grupos demográficos tienen mayor cercanía al abandono.

Finalmente, en el tercer apartado se construyeron distintos modelos predictivos de aprendizaje automático, de carácter clasificatorio, para predecir de la mejor manera a clientes que hayan abandonado la institución financiera. Dentro de los mejores modelos obtenidos, se nota la presencia de *XGB* y *Gradient Boosting,* obteniendo ambos un AUC de 0,94. Estos fueron analizados luego con una validación cruzada y un análisis de métricas tales como *recall* y precisión, advirtiendo que los modelos predictivos generados puedan estar sujetos a un *overfitting*, probablemente debido al bajo tamaño de la base de datos, y en especial de la clase "abandono".

Como resultado del trabajo y contraste de las hipótesis y objetivos planteados, se notó tanto la presencia de diferencias entre las características particulares del conjunto de clientes como entre las distintas variables explicativas en cuanto a especificar el comportamiento de los casos de abandono. Por ejemplo, se destacó la cercanía de aquellos clientes con alto nivel educativo, aquellos clientes de edad mayor a 50 años, y aquellos clientes casados y de alto nivel de ingresos a descontinuar la relación con el banco. Así mismo, los modelos predictivos realizados indicaron que las variables que mayor peso tienen en las reglas de decisión para clasificar a clientes como casos de abandono o no, son las de los montos y cantidades de transacciones realizadas, como también las variaciones de estos dos a lo largo del tiempo. Otras variables con un alto peso son las del balance total a fin de mes (*Total Revolving Balance)* y el total de productos utilizados. Esto va en orden al análisis exploratorio realizado a comienzo del trabajo, cual enseñó las diferencias entre los casos de abandono y los clientes actuales según distintas variables financieras.

Como propuesta a las instituciones financieras que sufran de este problema, se sugiere generar análisis constantes tanto de las variables demográficas de sus clientes como de las variables financieras de estos mismos. Agrupando a su cartera de clientes según las combinaciones demográficas y observando como las variables financieras de estos van variando a lo largo del tiempo, los bancos comerciales pueden tener un enfoque personalizado de sus clientes y así saber dónde enfocar las campañas de retención.

Cabe destacar que, en la realidad, la métrica de performance de modelos va a deber tener en cuenta la variable temporal. Los modelos predecirían la probabilidad de que un cliente abandone los productos ofrecidos en determinado plazo– a definir por cada investigador. De esta manera, una predicción de abandono puede considerarse correcta, por ejemplo, si el cliente abandona el banco en un plazo menor a tres meses. Mientras tanto, estas instituciones pueden realizar distintas campañas de retención sobre aquellos que hayan sido clasificados con una alta probabilidad de abandonarlos, para tomar como medida de performance un estudio en diferencias, o del estilo A/B *testing*, que logre identificar si el éxito de campañas de retención se debe o no a los clientes que fueron afectados por esta o no.

Referencias bibliográficas

-Athanassopoulos, A.D. (2000). Customer Satisfaction Cues To Support Market Segmentation And Explain Switching Behavior. Journal of Business Research, Volume 47, Issue 3, pp. 191- 207.

-Aurélie Lemmens & Sunil Gupta, Managing churn to maximize profits, 2013.

-Avon, V. (2015). Machine learning techniques for customer churn prediction in banking environments. Everis Italia S.p.a.

- Au W., Chan C.C., Yao X.: A Novel evolutionary data mining algorithm with applications to churn prediction. IEEE Trans. on evolutionary comp. 7 (2003) 532–545

-Berry, M.J.A. and Linoff, G., Mastering Data Mining: The Art and Science of Customer Relationship Management, Wiley Computer Publishing, New York, NY, 2000.

-Chiang, D., Wang, Y., Lee, S., & Lin, C. (2003). Goal-oriented sequential pattern for network banking churn analysis. Expert Systems with Applications, 25(3), 293–302.

- Coltman, Tim. (2007). Why build a customer relationship management capability?. The Journal of Strategic Information Systems. 16. 301-320. 10.1016/j.jsis.2007.05.001.

-Garland R.: Investigating indicators of customer profitability in personal retail banking. Proc. of the Third Annual Hawaii Int. Conf. on Business (2003) 18–21

-Kaur, M., Singh, K., & Sharma, N. (2013). Data Mining as a tool to Predict the Churn Behaviour among Indian bank customers. International Journal on Recent and Innovation Trends in Computing and Communication, 1(9), 720-725.

- Lemmens, A., & Croux, C. (2003). Bagging and boosting classification trees to predict churn. DTEW Research Report 0361

-Ling R. & Yen D. C., Customer relationship management: An analysis framework and implementation strategies, 2001.

-Peña, D. (2002). Análisis de datos multivariantes.

-Poel, Van Den, D. & Lariviere, B. (2003). Customer Attrition Analysis For Financial Services Using Proportional Hazard Models. European Journal of Operational Research, Vol. 157, Issue 1, pp. 196-217.

-Prasad, Devi U. (2012). Prediction Of Churn Behavior Of Bank Customers Using Data Mining Tools. Indian Journal of Marketing. Volume 42, Issue 9, September 2012.

-Santiago, F. (2021). "Implementación de modelos de aprendizaje automático". E72.1.02.

-Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl Jr, K. C. (2017). Data mining for business analytics: concepts, techniques, and applications in R. John Wiley & Sons.

-Uriel, E. y Aldás, J. (2017). Análisis multivariante aplicado con R. 2°a edición.

-Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & DATA, M. (2005). Practical machine learning tools and techniques. In DATA MINING (Vol. 2, p. 4).

-Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. Expert Systems with Applications, 36(3), 5445-5449.

- Zhao, Y., Li, B., Li, X., Liu, W., & Ren, S. (2005). Customer churn prediction using improved one-class support vector machine. Lecture Notes in Computer Science, 3584, 300–306.

-Zoric, Alisa Bilal, (2016), Predicting customer churn in banking industry using neural networks, Interdisciplinary Description of Complex Systems - scientific journal, 14, issue 2, p. 116-124.

Apéndices

Link al cuaderno Google Collaboratory: aquí

Anexo – reporte del mentor

El presente trabajo plantea el problema de implementar estrategias exitosas de retención sobre los clientes más propensos a abandonar una empresa. Considero que dicho problema es muy relevante en el contexto de las empresas bancarias.

Además, considero que el objetivo general propuesto de evaluar la posibilidad de identificar los clientes bancarios con alta probabilidad de abandono por medio de modelos de minería de datos es coherente con el problema planteado. Y dicho objetivo es consistente con la hipótesis de que los datos demográficos, transaccionales y comerciales de los clientes permiten identificar a los clientes con mayor riesgo de abandono.

Finalmente, considero que el presente trabajo alcanza el objetivo propuesto al presentar un detallado análisis de los datos y un correcto desarrollo del modelo de predicción.