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

Minería de datos en CRM. Modelos de predicción de abandono de clientes bancarios

AUTOR: CANDELA BEYREUTHER
MENTOR: ROBERTO ABALDE

DICIEMBRE 2021

Resumen

Los clientes son el activo más importante para cualquier organización y, bajo el paradigma de marketing actual, generar relaciones duraderas y rentables con los mismos es fundamental. La competencia en el sector de servicios financieros se ha revolucionado e intensificado en las últimas décadas, producto de la innovación digital, y sus organizaciones cada vez más ponen el énfasis en el cuidado de sus clientes. El uso de las técnicas de minería de datos le permite a una organización aprender de las experiencias pasadas, encontrar patrones de comportamiento entre sus clientes e identificar aquellos que son propensos a culminar su relación con la organización.

Dada esta problemática, el presente trabajo tiene por objetivo evaluar la posibilidad de identificar los clientes de una institución bancaria con una alta probabilidad de abandono por medio de modelos de minería de datos de clasificación. La metodología empleada se basa en el proceso CRISP-DM bajo la cual se desarrollan seis modelos de aprendizaje supervisado que son evaluados con el criterio AUC-ROC. Los resultados muestran que es posible construir un sistema de predicción del abandono de clientes, lo que permite a la organización mejorar las acciones de retención y comprender las causas del abandono.

Palabras clave: CRM, abandono de clientes, minería de datos, modelos de predicción, clientes bancarios.

Índice

Introducción

Los clientes son el activo más importante para las organizaciones y, por ello, la gestión de las relaciones con los clientes es el eje central del paradigma actual de marketing de relaciones. En este escenario, la satisfacción y retención de los clientes se vuelve un asunto de gran relevancia para la gestión de los negocios y supone importantes ahorros para las organizaciones dado que el costo de conseguir nuevos clientes es ampliamente superior al costo de retener a los clientes existentes. La tasa de pérdida o tasa de cancelación de clientes es una métrica que mide el porcentaje de abandono de los clientes en un determinado período de tiempo. En contraposición, la tasa de retención de clientes es una métrica que mide el porcentaje de los clientes que se han conservado en un determinado período de tiempo.

El sector de servicios financieros y bancarios ha experimentado una gran transformación en los últimos años, impulsada principalmente por la innovación digital. Los cambios experimentados han dado lugar a nuevos modelos de negocio y a nuevos participantes compitiendo por los clientes de la industria. Hoy en día, una persona enfrenta ilimitadas opciones en lo que se refiere a dónde colocar su dinero, gestionar sus pagos o pedir un préstamo y le resulta cada vez más fácil el movimiento de sus activos entre diferentes prestadores. Estos factores, en conjunto, intensifican la competencia y significan que el abandono de clientes es un asunto de gran relevancia para las organizaciones del sector. Para los bancos, la construcción de relaciones rentables y duraderas con sus clientes ha adquirido mucha importancia en las últimas décadas.

Dada esta problemática, el presente trabajo se centra en el caso de una institución bancaria que enfrenta una tasa de pérdida de clientes del 20% sin conocer las causas del abandono. En este contexto, el interrogante que se plantea es ¿cómo se pueden identificar los clientes con mayor propensión a abandonar los servicios del banco de modo de poder implementar estrategias de retención sobre ellos? Para responder al interrogante planteado, el objetivo general del presente trabajo es evaluar la posibilidad de identificar los clientes bancarios con una alta probabilidad de abandono por medio de modelos de minería de datos de clasificación.

Para alcanzar este objetivo, en un primer apartado, se introduce el cambio del paradigma del marketing a la CRM y se hace foco en la importancia que reviste la fidelización de clientes. Asimismo, se estudia la transformación del sector de servicios bancarios y financieros de los últimos años y el problema del abandono de clientes en

dicha industria. Por último, se destaca el aporte de la minería de datos a la CRM, y a la retención de clientes en particular, y se introducen los modelos predictivos de clasificación.

Luego, en un segundo apartado, se presenta la metodología CRISP-DM para proyectos de minería de datos y se identifica y describe la fuente de datos. Además, se realiza un análisis exploratorio y descriptivo de las variables contenidas en la base de datos de los clientes de la organización bajo estudio. En un último subapartado, se preparan los datos para el modelado predictivo.

En un tercer apartado se describen con mayor detalle los modelos de clasificación que luego son desarrollados para clasificar a los clientes bancarios y obtener su probabilidad de abandono. En este apartado, adicionalmente, se selecciona una métrica para evaluar los modelos predictivos, se evalúan los modelos desarrollados y se elige el modelo que mejor clasifica a los clientes de acuerdo con su probabilidad de abandono. En último lugar, se analizan las características de los modelos para evaluar las principales variables que determinan el abandono.

El apartado final del trabajo contiene su conclusión, en la que se resumen los principales hallazgos y resultados obtenidos. Además, se presentan recomendaciones para futuros estudios.

Minería de datos y CRM: el caso de los servicios financieros

Para comenzar el presente trabajo, resulta pertinente realizar una conceptualización del problema del abandono de clientes para la CRM e introducir el aporte de la minería de datos a este problema. Para ello, en un primer subapartado, se presenta la CRM y su énfasis en la conservación de clientes. En un segundo subapartado, se hace foco en el sector de servicios financieros y bancarios y en el impacto del problema del abandono de clientes para las organizaciones que lo componen. En un tercer y último subapartado, se define la minería de datos y se presenta el aporte de los modelos predictivos de clasificación al problema de la retención de clientes.

1.1 CRM y retención de clientes

Los clientes son el activo estratégico más importante para cualquier organización (Vicente, 2009) y las organizaciones que tienen éxito en la actualidad comparten la característica de estar fuertemente orientadas a sus clientes (Kotler y Armstrong, 2008).

De hecho, en las últimas décadas, el paradigma del marketing ha migrado del tradicional marketing de transacciones a un nuevo marketing de relaciones (Vicente, 2009) centrado en la CRM, o gestión de relaciones con el cliente[1]. Para Kotler y Armstrong (2008), la CRM se ha vuelto el eje más importante del marketing moderno y la definen como "el proceso global de construir y mantener relaciones rentables con los clientes mediante la entrega de un valor superior y una mayor satisfacción" (p. 15).

Hoy en día, las organizaciones se valen de la gestión de las relaciones con los clientes para crear con ellos lazos que sean duraderos y rentables (Kotler y Armstrong, 2008). Bajo este paradigma, la satisfacción y fidelización de los clientes se vuelve esencial para la administración del negocio. Perder un cliente significa una pérdida de su CLV o valor de vida del cliente, el cual se define como el valor actual de los ingresos que se obtendrán del cliente a lo largo de toda la relación comercial con él.

Las empresas ponen el énfasis en la conservación de los clientes principalmente debido a que, en la mayoría de las industrias maduras, adquirir clientes no es una tarea fácil (Tsiptsis y Chorianopoulos, 2009). Y no sólo resulta dificultoso para las organizaciones reemplazar a los clientes perdidos con clientes de la competencia sino que generalmente es más costoso atraer a un nuevo cliente que retener a uno actual. Al respecto, Kotler y Armstrong (2008) señalan que puede costar de cinco a diez veces más conseguir un nuevo cliente que mantener satisfecho a uno actual y Vicente (2009) resalta que las empresas "pierden a más de la mitad de sus clientes cada cinco años" (p. 743) y que los costos de captar nuevos clientes crecen, en el mismo tiempo, de manera exponencial. Este hecho tiene un impacto directo en los costos operativos y presupuestos de marketing de una compañía.

Por otro lado, construir relaciones duraderas con los clientes es más rentable para las compañías. Verbeke et al. (2012) señalan que los clientes de largo plazo generan beneficios más altos para las empresas, suelen ser menos sensibles a las acciones de marketing de la competencia y que presentan una probabilidad mayor de atraer nuevos clientes a través de recomendaciones boca a boca. Los autores agregan que una pequeña mejora en la retención de clientes puede llevar a una mejora sustancial en las ganancias de la compañía.

[1] Esta transición al marketing de relaciones encuentra su fundamento en el hecho de que, en muchas industrias, los mercados se encuentran saturados y que las empresas enfrentan competidores cada vez más sofisticados (Kotler y Armstrong, 2008) y clientes cada vez más exigentes (Guadarrama Tavira y Rosales Estrada, 2015).

Resulta importante, entonces, definir qué es y cómo se mide el abandono y la retención de clientes. Los clientes que abandonan una organización son aquellos que dejan de adquirir sus productos o servicios o que cesan, de forma voluntaria o involuntaria, su relación contractual con la misma. La tasa de abandono o tasa de cancelación de clientes (conocida también como Churn Rate) se puede obtener de la siguiente manera:

$$Tasa\ de\ abandono = \frac{Clientes\ que\ han\ abandonado\ la\ compañía\ en\ el\ período}{Clientes\ al\ inicio\ del\ período} x\ 100$$

El período al que hace referencia la fórmula puede ser mensual, semestral, anual o el período que la organización determine conveniente para realizar la evaluación. Este tiempo suele variar dependiendo, por ejemplo, si la relación cliente-proveedor se materializa contractualmente o no. De manera análoga se obtiene la tasa de retención de clientes:

$$Tasa\ de\ retención = \frac{Clientes\ que\ permanecen\ con\ la\ compañía\ al\ final\ del\ período}{Clientes\ al\ inicio\ del\ período} x\ 100$$

Gold (2020) resalta que la tasa de abandono y la tasa de retención son dos caras de una misma moneda y que reducir el abandono es equivalente a aumentar la retención y viceversa.

1.2. El problema del abandono de clientes en el sector bancario

La globalización y la desregulación de los mercados están cambiando rápidamente el escenario competitivo de muchos sectores económicos (García et al., 2017). La aparición de nuevos competidores y tecnologías genera un aumento en la competencia y, con él, una creciente preocupación entre las compañías proveedoras de servicios por crear lazos más fuertes con sus consumidores (García et al., 2017).

En la industria financiera se ha iniciado una profunda transformación en los últimos años. Las tecnologías digitales han cambiado la gestión de pagos, de préstamos y de capitales – un proceso que la pandemia por el COVID-19 ha acelerado (Feyen et al., 2021). En efecto, el aumento en la adopción de herramientas digitales causado por la pandemia ha enfatizado la necesidad del sector de servicios bancarios y financieros de mejorar sus capacidades digitales o correr el riesgo de perder clientes en manos de sus

competidores, incluyendo nuevos participantes y compañías tecnológicas (Tapestry Networks, 2021).

La innovación digital ha traído grandes avances en la conectividad de los sistemas, en las capacidades de procesamiento computacional y sus costos, y en la creación y uso de datos (Feyen et al., 2021). Estos desarrollos han dado lugar a nuevos modelos de negocio y nuevos jugadores en el sector de servicios financieros y bancarios. Las denominadas *tech giants*, gigantes tecnológicos como Google, Apple o Amazon, ya están ofreciendo soluciones bancarias (PwC, 2020), a la par que el nuevo y especializado segmento *FinTech* ha desagregado la industria financiera, apoyándose en la tecnología digital para acceder a mercados de nicho (Feyen et al., 2021)[2].

Este escenario resulta en ilimitadas opciones ofrecidas a los consumidores cuando se trata de dónde colocar su dinero, canalizar sus pagos y/o solicitar capital. Adicionalmente, la tecnología facilita, más que nunca, el movimiento de fondos y activos entre diferentes proveedores. En conjunto, esta simplicidad y la abundancia de opciones hacen que sea cada vez más fácil para los consumidores cambiar de banco o finalizar una relación comercial. Además, cada vez es más usual que una persona posea cuentas en diferentes bancos o diferentes productos en cada uno de ellos (Glady et al., 2009). Es decir que los consumidores pueden mudar una porción de su cartera de productos a una empresa de la competencia.

La posibilidad de perder clientes rentables en manos de competidores agresivos ha llevado a las compañías a poner el foco en la preservación de los clientes existentes y no tanto en la captura de clientes nuevos (García et al., 2017). En el sector de servicios financieros, el abandono de clientes preocupa principalmente a las empresas que mantienen con sus clientes relaciones contractuales no vinculantes, como las empresas de tarjetas de créditos y bancos, en las que son comunes las tasas de abandono de hasta un 30% (Kaemingk, 2018). En el caso de los bancos, existen algunos productos para los cuales la relación contractual puede ser vinculante pero, incluso en esos casos, Kaemingk (2018) señala que las tasas de abandono podrían llegar hasta un 7%.

El incremento en la competencia, tanto por parte de los grandes bancos que mejoran su oferta digital como de los nuevos participantes de la industria, significa que la experiencia del consumidor se ha vuelto un campo de batalla cada vez más importante

[2] El término *FinTech* refiere a las tecnologías digitales que poseen el potencial de transformar la provisión de servicios financieros, modificando y propiciando el desarrollo de nuevos modelos de negocios, aplicaciones, procesos y productos (Feyen et al., 2021).

para las instituciones financieras (Kaemingk, 2018). Los consumidores cada día son más conscientes de las opciones ofrecidas por los competidores y, como resultado, sus expectativas son cada vez mayores (García et al, 2017). En este contexto, la satisfacción de los clientes se vuelve un asunto crucial. Kaemingk (2018) resalta que, después de todo, la experiencia del consumidor impulsa la lealtad de los clientes y reduce el abandono, lo cual es fundamental en el sector de servicios financieros, en el que el CLV es una métrica clave.

El objetivo final es, entonces, retener y reforzar las relaciones comerciales con los clientes rentables, para lo cual las organizaciones deben construir mecanismos fuertes de prevención del abandono (García et al, 2017). En este sentido, Kaemingk (2018) resalta un estudio conducido por el Qualtrics XM Institute sobre clientes que han abandonado instituciones bancarias, el 56% de los cuales indicó que el banco podría haber cambiado su opinión. Este estudio refleja claramente la importancia de las estrategias de fidelización en el sector bancario.

1.3. Minería de datos para predicción de abandono de clientes

La CRM se fundamenta en la gestión de la información de los clientes y en la detección de los puntos de contacto con ellos, de modo de aprovecharlos para mejorar la fidelidad y rentabilidad (Vicente, 2009). Para Vicente (2009), se vuelve crítico identificar a cada uno de los clientes, lo cual implica contar con la mayor cantidad de datos posibles sobre ellos. En este sentido, para el marketing de relaciones obtener información de los clientes es tanto o más importante que transmitirles información (Pinto, 1997), lo cual se ha posibilitado gracias a la gran explosión tecnológica de las últimas décadas y al desarrollo de grandes bases de datos.

Esta nueva realidad corre el eje del estudio y análisis de los mercados y pone en el centro al conocimiento profundo y extenso de cada uno de los clientes que posibilita la información provista por las bases de datos (Vicente, 2009). Sin embargo, las bases de clientes se componen de diferentes personas, con necesidades y comportamientos diferentes y con diferentes potenciales que deben ser gestionados adecuadamente (Tsiptsis y Chorianopoulos, 2009). Por ello, se debe tener en cuenta que apuntar a una personalización de las relaciones en un mundo masivo implica necesariamente generar nuevas modalidades de procesamiento y análisis de los datos disponibles (Vicente, 2009).

Existen tres funciones que resultan básicas a la hora de obtener el mayor provecho de un sistema de gestión de la relación con el cliente. Ellas son:

1) **Medir** los resultados de los esfuerzos aplicados al marketing, utilizando como patrón la rentabilidad de los clientes
2) **Predecir** mediante el *data mining* el comportamiento de los clientes y aprender de las experiencias pasadas
3) **Actuar** utilizando los medios más idóneos para llegar a los clientes con las ofertas más ajustadas a sus deseos (Vicente, 2009, p. 746).

Centrándose en el segundo punto, el autor plantea que la minería de datos es un componente crítico del marketing de relaciones.

La minería de datos se enmarca en lo que Evans (2017) llama Analítica de Negocios y que es definida por el autor como el uso de datos, tecnologías de la información, análisis estadísticos, métodos cuantitativos y modelos matemáticos y computacionales para ayudar a la gestión a obtener un conocimiento más profundo de las operaciones de su negocio y tomar mejores decisiones. En este marco, la minería de datos se constituye en un proceso orientado al entendimiento y descubrimiento de patrones entre los datos que generen algún tipo de ventaja, por ejemplo, económica o competitiva (Witten et al., 2017). De esta manera, la minería de datos convierte los datos en conocimiento y en información accionable (Tsiptsis y Chorianopoulos, 2009).

En relación al problema del abandono de clientes, una de las preguntas que intentan responder las organizaciones refiere a cuáles son los clientes que tienen una alta probabilidad de abandonar la empresa, para lo cual pueden hacer uso de la analítica predictiva y las técnicas de minería de datos y aprendizaje automático. La analítica predictiva busca predecir el futuro por medio del estudio detallado de información histórica, detectando patrones y relaciones en los datos y extrapolando estas relaciones hacia adelante en el tiempo (Evans, 2017). Valiéndose del conocimiento que aporta la analítica predictiva, los tomadores de decisiones de una organización pueden tomar acciones que permitan, en muchos casos, modificar ese futuro y mejorar la competitividad de la organización.

Así, la minería de datos puede ayudar a la retención de clientes ya que permite una identificación temprana de clientes valiosos con mayor propensión a abandonar la organización, de modo de poder implementar a tiempo estrategias enfocadas de fidelización sobre ellos (Tsiptsis y Chorianopoulos, 2009). Guadarrama Tavira y Rosales Estrada (2015) denotan, al respecto, que los clientes que han abandonado la empresa aportan mucha información y que las organizaciones deben aprender de ellos. Los autores

agregan que estudiar estos clientes permite conocer los motivos de su retirada, corregir las deficiencias y evitar que se vayan los clientes actuales.

Con el fin de identificar a los clientes con probabilidad de abandono, las organizaciones pueden aplicar modelos predictivos de clasificación. El objetivo de estos modelos es clasificar un determinado resultado en dos o más categorías, basándose en diferentes atributos (Evans, 2017). Generalmente, una organización posee una base de datos con diferentes atributos sobre sus clientes que pueden constituirse como atributos predictores de otra variable de interés que indique si el cliente ha abandonado o no la organización. En los modelos de clasificación, los datos para los que la clasificación es conocida son usados para crear reglas, que luego son aplicadas a otros datos para los cuales la clasificación es desconocida (Shmueli et al., 2018).

Las probabilidades de abandono que se obtienen de los modelos de clasificación pueden ser usadas por las compañías para dirigir sus campañas de fidelización. Y es que la construcción de modelos de predicción de abandono es principalmente relevante desde el punto de vista de la retención (Glady et al., 2009). No obstante, no todos los clientes revisten el mismo interés para una empresa, ni todos los segmentos son rentables o centrales a su estrategia. Por ello, los resultados de los modelos predictivos pueden ser combinados con el valor presente y potencial de los clientes para priorizar las estrategias de retención (Tsiptsis y Chorianopoulos, 2009).

Metodología y datos utilizados para la predicción del abandono

Introducida la preocupación por el abandono o retención de clientes en el sector bancario y la contribución de los modelos de minería de datos de clasificación, en este segundo apartado se presenta la metodología y datos empleados. En un primer subapartado, se realiza una descripción del modelo CRISP-DM para proyectos de minería de datos. En un segundo subapartado, se identifica y se provee una descripción de la base de datos de clientes bancarios. En tercer lugar, se efectúa un análisis exploratorio sobre los datos para obtener una primera aproximación a su distribución, su interés para el objetivo del trabajo y a las posibles relaciones entre ellos. Por último, en el cuarto subapartado, se preparan los datos para la posterior construcción de modelos de predicción.

2.1. Metodología CRISP-DM para proyectos de minería de datos

La metodología empleada para abordar el objetivo definido sigue el modelo CRISP-DM, que son las siglas de *Cross-Industry Standard Process for Data Mining*. Bajo este modelo, el ciclo de vida de un proyecto de minería de datos es un proceso iterativo que comienza con el entendimiento del negocio y finaliza con la implementación de un modelo.

La primera fase, entonces, comienza con la comprensión de los objetivos del negocio y la decisión de si la minería de datos puede ayudar a conseguirlos (Witten et al., 2017). Se deben poder responder preguntas como ¿cómo se van a usar los resultados del proyecto? o ¿quiénes se verán afectados por los mismos? (Shmueli et al., 2018).

En una segunda fase, los datos son obtenidos y estudiados para juzgar si los mismos son adecuados para continuar con el proceso (Witten et al., 2017). La obtención de datos puede implicar también la integración de datos de diferentes fuentes o bases (Shmueli et al., 2018). En esta etapa, además, se efectúa un análisis descriptivo y exploratorio de los datos para comenzar a obtener conocimiento de los mismos que potencie y retroalimente el entendimiento sobre el negocio. Los datos se exploran para estudiar si sus valores se encuentran dentro de rangos esperables, si presentan valores extremos o valores perdidos y para verificar consistencia en la definición de variables, unidades de medida y períodos de tiempo, por ejemplo (Shmueli et al., 2018).

La tercera fase es la preparación de datos e involucra el preprocesamiento de los datos para que los algoritmos de minería de datos puedan producir modelos (Witten et al., 2017). En este momento, los datos se limpian y se preparan para ser usados en la etapa posterior de aplicación de modelos. Típicamente, en esta fase, se crean también nuevas variables a partir de las existentes que puedan mejorar la capacidad predictiva de los modelos. Este proceso es comúnmente denominado ingeniería de predictores.

Definido el problema de minería de datos a abordar, en la cuarta fase, se desarrollan los algoritmos de aprendizaje automático. La preparación de datos y el modelado se retroalimentan y producen iteraciones ya que generalmente los resultados obtenidos de los modelos generan conocimiento que afecta la elección de técnicas de preprocesamiento (Witten et al., 2017). La naturaleza iterativa de la etapa de modelado se debe, asimismo, a que normalmente no hay un único modelo que solucione el problema en cuestión. Como regla general se suelen probar diferentes modelos y diferentes variantes de los mismos, ajustando los valores de sus parámetros (Shmueli et al., 2018).

En la quinta y anteúltima fase, los modelos son evaluados en relación a sus capacidades predictivas y el mejor modelo es elegido. Esta elección involucra una interpretación de los resultados de los modelos, una evaluación de diferentes métricas de rendimiento y un análisis de los errores que los modelos cometen. Además, al elegir un modelo, previo a avanzar a la etapa final, su capacidad predictiva se reevalúa sobre un nuevo conjunto de datos y sus resultados son evaluados por un experto del negocio para verificar que se alineen con el conocimiento del negocio (Verbeke et al., 2012).

En caso de hallarse un modelo lo suficientemente bueno, la sexta y última fase implica su implementación dentro de las operaciones de la organización. El presente trabajo excluye de su alcance la implementación de un modelo de minería de datos y abarca desde la comprensión de los objetivos y necesidades del negocio, en este caso la institución bancaria, hasta la evaluación de diferentes modelos para seleccionar aquél que clasifique mejor a los clientes según su probabilidad de abandono.

2.2. Fuente y descripción de la base de datos

Los datos que se utilizan para el desarrollo del trabajo son datos secundarios obtenidos de la plataforma *Kaggle*[3]. El conjunto de datos posee datos demográficos y comerciales de 10.000 clientes de una institución bancaria, de los cuales 2.037 han cerrados sus cuentas con la entidad. En particular, el conjunto de datos contiene los siguientes atributos sobre los clientes:

1. **RowNumber**: Se trata de un atributo que indica el número de fila.
2. **CustomerId**: Se trata de un atributo identificador y representa el código único de identificación del cliente.
3. **Surname**: Apellido del cliente.
4. **Geography**: Variable categórica que indica el país de residencia del cliente. El conjunto contiene datos de clientes de Francia, España y Alemania.
5. **Gender**: Atributo binario que denota el género del cliente (masculino o femenino).
6. **Age**: Edad del cliente.
7. **Tenure**: Número de años que el cliente ha sido cliente del banco o antigüedad de la relación comercial en años.

[3] Recuperados de: https://www.kaggle.com/mathchi/churn-for-bank-customers

8. **Balance**: Balance de cuenta bancaria del cliente, expresado en unidades monetarias.
9. **CreditScore**: Puntos de calificación crediticia del cliente.
10. **NumOfProducts**: Cantidad de productos que el cliente ha adquirido del banco.
11. **HasCrCard**: Variable binaria que indica si el cliente posee una tarjeta de crédito de la institución bancaria.
12. **IsActiveMember**: Variable binaria que indica si el cliente es un cliente activo del banco[4].
13. **EstimatedSalary**: Salario estimado del cliente, expresado en unidades monetarias.
14. **Exited**: La variable objetivo "Exited" es de tipo binaria e indica si el cliente ha abandonado o no la compañía. Esta variable toma valor 1 en el caso en que el cliente ha abandonado a la empresa y valor 0 en caso contrario (el cliente mantiene su relación comercial con el banco).

Es importante remarcar que la definición del abandono depende de la organización, industria y/o segmento de clientes. Hay pérdida de clientes de manera total y parcial y abandono voluntario e involuntario. En lo que respecta a la organización bajo estudio, se define que un cliente ha abandonado el banco en el caso en que voluntariamente haya cerrado todas sus cuentas y productos con la compañía.

2.3. Análisis exploratorio sobre las variables predictoras y variable objetivo

Identificadas la variable objetivo y las variables predictoras, se procede con el análisis exploratorio y descriptivo de los atributos de la base de datos. El análisis exploratorio permite ahondar en la descripción de las variables, examinar la interrelación entre ellas, identificar subconjuntos de instancias que revistan interés y desarrollar una primera noción de las posibles asociaciones, tanto entre los predictores así como entre los predictores y la variable objetivo (Larose y Larose, 2014). Larose y Larose (2014) proponen explorar los datos pero sin perder de vista el objetivo general propuesto. Siguiendo esta lógica propuesta por los autores, se realiza un análisis exploratorio sobre los atributos, buscando ahondar en su relación con la variable a predecir.

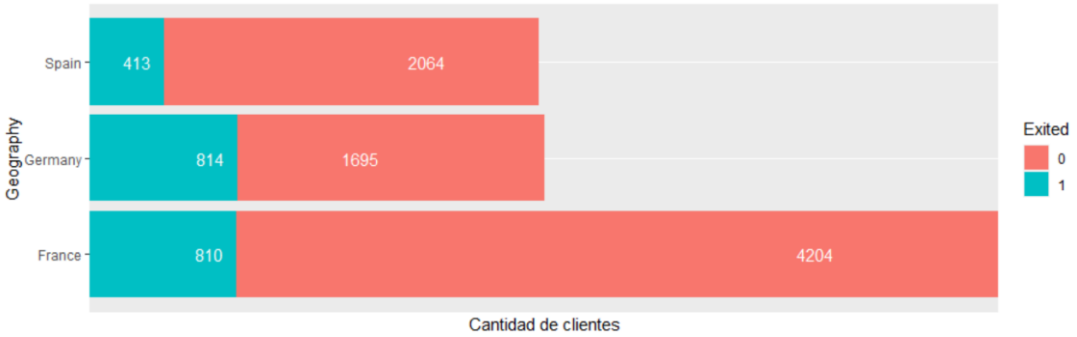
Como ha sido destacado en el subapartado anterior, la variable a predecir es la variable Exited que indica si el cliente ha culminado su relación comercial con el banco

[4] Nótese que no se posee información acerca de la determinación de la clasificación de los clientes para esta variable.

(tomando valor 1 en este caso). Del total de clientes bajo estudio, 2.037 han abandonado la compañía, lo que representa una tasa de abandono del 20,37%.

Continuando el análisis del conjunto de datos con el país de residencia de los clientes bancarios, en la Figura 1 se puede ver que un poco más del 50% de los clientes del banco provienen de Francia, encontrándose el resto distribuidos en similar proporción entre España y Alemania. Sin embargo, si se evalúa la distribución por país de los clientes que han abandonado la organización, se observa una mayor tasa de abandono en Alemania. Específicamente, un 32% de los clientes provenientes de ese país ha abandonado el banco mientras que dicha cifra se ubica en torno al 16% para los clientes de Francia y al 17% para clientes provenientes de España.

Figura 1 - Cantidad de clientes por país de residencia y proporción de abandono



Fuente: Elaboración propia en R Studio.

La proporción de abandono también es dispar si se considera el género del cliente. Al analizar la distribución de la variable se observa que 5.457 clientes son de género masculino y los 4.543 restantes de género femenino, siendo este género el que presenta una mayor cantidad de clientes que han abandonado. Puntualmente, la tasa de cancelación de clientes para el género femenino es del 25% y del 16% para el masculino.

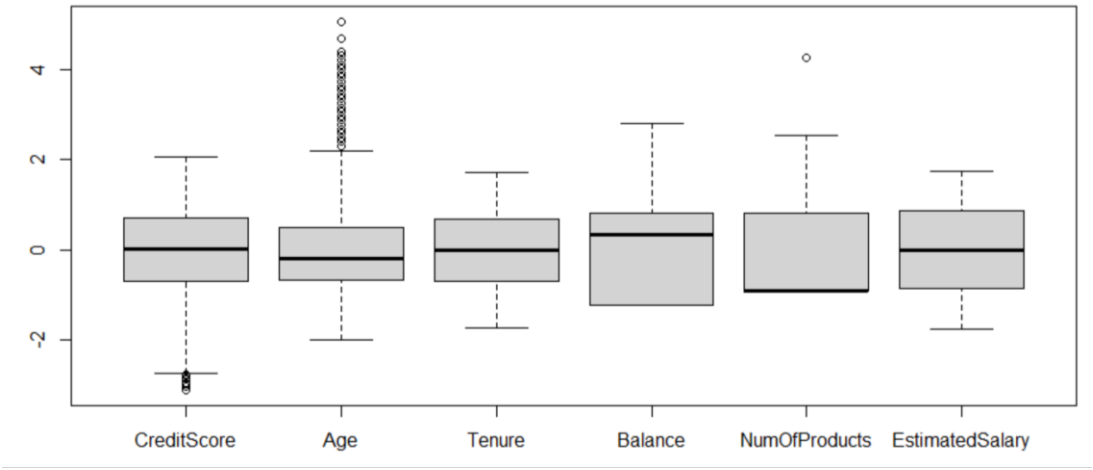
El conjunto de datos de clientes posee, además de la variable objetivo, otras dos variables binarias que indican si el cliente posee una tarjeta de crédito dentro de los productos adquiridos al banco y si es un cliente activo. Observando la distribución de la primera de ellas, se puede notar que aproximadamente el 70% de los clientes posee una tarjeta de crédito de la institución. La proporción de clientes que han abandonado la organización es muy similar entre ambos grupos y se ubica en torno a la tasa de abandono general del 20%.

Concluyendo el análisis exploratorio sobre las variables categóricas, se analiza la variable IsActiveMember. Se observa una cantidad similar de clientes activos y clientes no activos pero una proporción desigual en la propensión al abandono de uno y otro

grupo. Mientras que la tasa de abandono de los clientes activos es del 14%, la tasa de abandono de los clientes no activos prácticamente la duplica en valor y se ubica en un 27%.

Volviendo ahora la atención sobre las variables predictoras cuantitativas, se visualiza, en una primera instancia, la distribución de los datos correspondientes a estas seis variables por medio del diagrama de cajas que se muestra en la Figura 2. El análisis de las variables cuantitativas se complementa, asimismo, con una evaluación de sus medidas de posición y de dispersión, las cuales se muestran en la Tabla 1.

Figura 2 - Diagrama de cajas de los atributos Age, Tenure, CreditScore, Balance, NumOfProducts y EstimatedSalary[5]



Fuente: Elaboración propia en R Studio.

[5] Nótese que para una mejor comparación de las distribuciones de las variables, en esta figura, las mismas se encuentran estandarizadas, de modo de llevarlas a igual escala. Para estandarizar se resta la media y se divide por el desvío estándar de modo tal que todas las variables posean media 0 y varianza 1.

Tabla 1 - Medidas de posición y de dispersión de Age, Tenure, Balance, CreditScore, EstimatedSalary y NumOfProducts

	Age	Tenure	Balance	Credit Score	Estimated Salary	Num Of Products
Mínimo	18	0	0	350	11,58	1
Cuartil 1	32	3	0	584	51.002,11	1
Media	38,92	5,01	76.485,89	650,52	100.090,24	1,53
Mediana	37	5	97.198,54	652	100.193.91	1
Cuartil 3	44	7	127.644,24	718	149.388,24	2
Máximo	92	10	250.898,09	850	199.992,48	4
Desvío estándar	10,48	2,89	62.397,40	96,65	57.510,49	0,58

Fuente: Elaboración propia con R Studio.

Del estudio de la variable Age, que denota la edad de los clientes, se concluye que la misma presenta una alta dispersión, la cual se evidencia claramente en el diagrama de cajas. Además, se evidencia una amplia diferencia entre el tercer cuartil y el valor máximo de la variable. En el siguiente histograma se puede apreciar la distribución de la edad de los clientes con mayor detalle y se incorpora la proporción de clientes que han abandonado para evaluar si la edad de los mismos podría ser un buen predictor de su probabilidad de abandono.

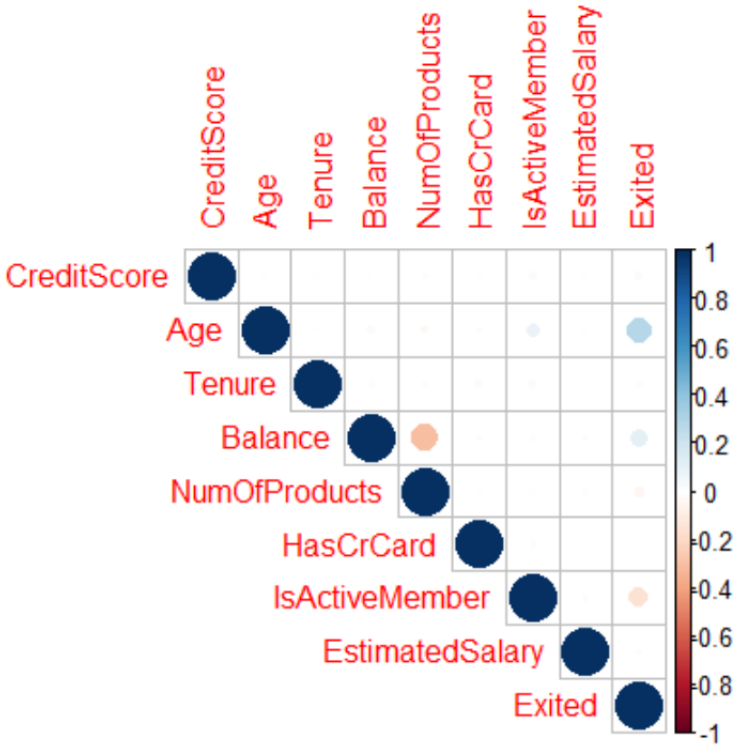
Figura 3 - Histograma de Age según abandono de clientes

Fuente: Elaboración propia en R Studio.

Como se puede notar en la Figura 3, la variable Age presenta una concentración de los datos en torno a la media y una asimetría positiva. Sin embargo, si se observa la

distribución de edades de los clientes que han abandonado la compañía, se aprecia una mayor concentración de las mismas en torno al tercer cuartil, lo que podría indicar cierta relación entre la variable Age y Exited. De hecho, la edad media de los clientes que han abandonado la compañía es superior a la media de la variable y se ubica en 44,84 años. Adicionalmente, si se observa la matriz de correlación entre las variables se puede ver que la variable Exited presenta una correlación positiva con la variable Age, principalmente, y, en menor medida, con la variable IsActiveMember y Balance.

Figura 4 - Matriz de Correlación



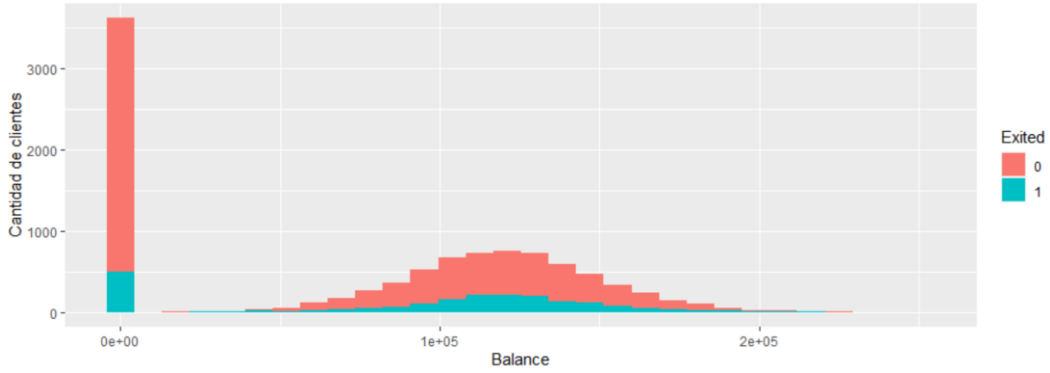
Fuente: Elaboración propia en R Studio.

Como se puede ver en la matriz, las variables predictoras no muestran una correlación alta entre ellas, con excepción de las variables Balance y NumOfProducts. El estudio de la correlación entre las variables predictoras es especialmente relevante para determinar la necesidad de reducir la dimensionalidad del conjunto de datos para la etapa de desarrollo de modelos. En este caso, el hecho de que las variables no presenten correlaciones significativas entre ellas significa que no es necesario aplicar técnicas de reducción de la dimensión.

En lo que respecta a la antigüedad de la relación comercial de los clientes con el banco, medida en años, la misma toma valores entre 0 y 10. Tanto para los clientes que han abandonado como para los que no, la variable Tenure presenta una distribución simétrica y la media coincide con la mediana.

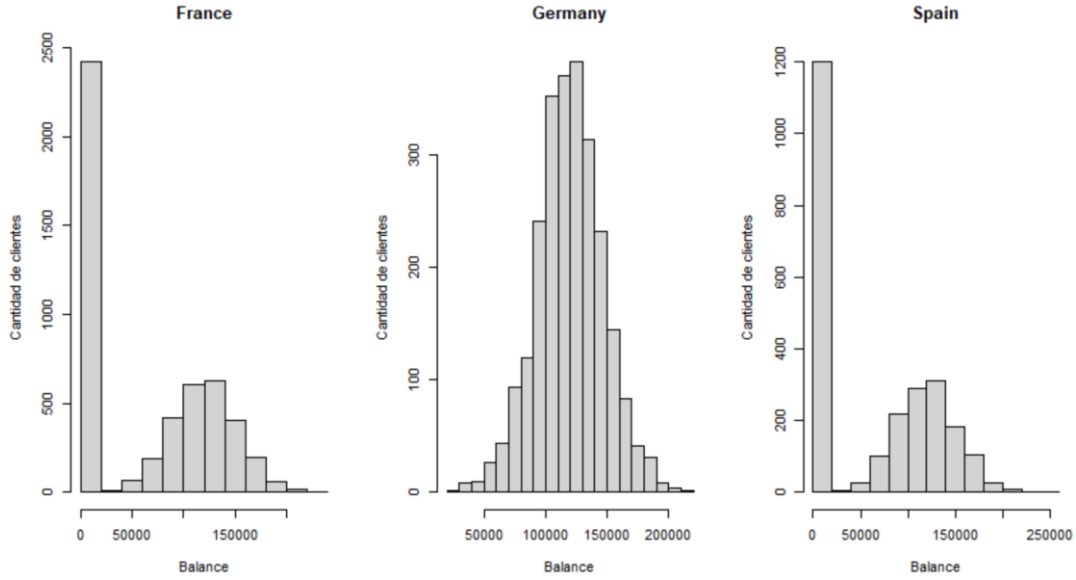
La variable Balance, por su parte, presenta una distribución particular, principalmente dada por el hecho de que el 36% de los clientes posee un saldo de cuenta bancaria de 0 unidades monetarias, como se puede apreciar en la Figura 5. Por otro lado, cabe resaltar que esta variable presenta una distribución dispar al analizarla por país de residencia de los clientes, como se puede ver en la Figura 6. Mientras que la variable presenta una distribución similar en Francia y España, la distribución de la variable para los clientes de Alemania dista mucho de la distribución general de la variable y se encuentra concentrada simétricamente en torno a la media.

Figura 5 - Histograma de Balance según abandono de clientes



Fuente: Elaboración propia en R Studio.

Figura 6 - Histograma de Balance por país



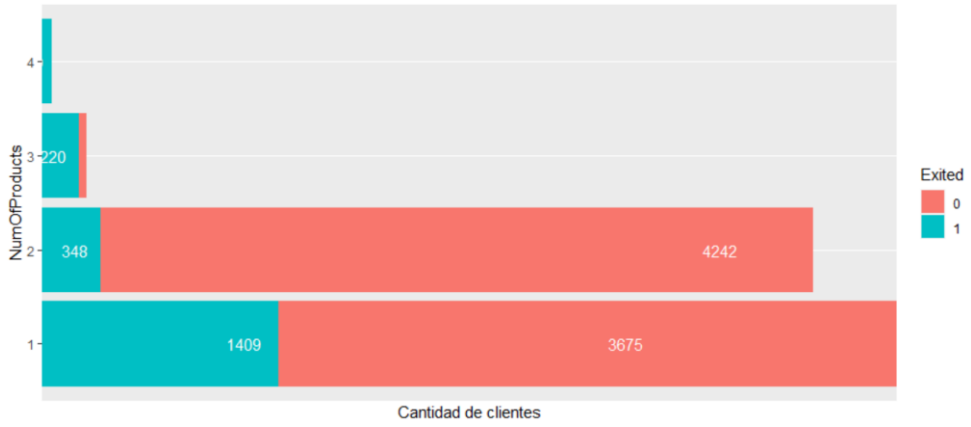
Fuente: Elaboración propia en R Studio.

La distribución desigual de la variable Balance por país y el hecho de que Alemania presenta una mayor tasa de abandono de clientes que el resto de los países

podría explicar el hecho de que el balance promedio es superior en los clientes que han culminado su relación comercial con el banco. Precisamente, mientras que el balance promedio de los clientes que mantienen su vínculo comercial es de 72.745 unidades monetarias, el correspondiente a los clientes que sí han abandonado la compañía es de 91.109 unidades monetarias.

Virando el análisis exploratorio hacia la variable NumOfProducts, que indica la cantidad de productos que el cliente ha adquirido, la misma presenta una asimetría positiva ya que la variable toma valores entre 1 y 4 pero el 97% de los clientes ha adquirido tan solo uno o dos productos. Del estudio de esta variable, resulta llamativo el hecho de que el 100% de los clientes que adquirieron 4 productos y el 83% de los clientes que adquirieron 3, ha finalizado su vínculo comercial con el banco.

Figura 7 - Cantidad de clientes según cantidad de productos adquiridos y proporción de abandono



Fuente: Elaboración propia en R Studio.

La calificación crediticia de los clientes toma valores entre 350 y 850 siendo la media 650 y presentando en su distribución una leve asimetría hacia la izquierda o negativa. En un primer análisis, no se observan diferencias sustanciales entre la calificación crediticia de los clientes que han abandonado la empresa y la de los que no lo han hecho. Por último, el salario estimado también sigue una distribución similar en ambos grupos, ubicándose la media cerca de las 100.000 unidades monetarias y presentando una distribución prácticamente simétrica respecto de la misma.

En conclusión, del análisis exploratorio de las variables contenidas en el set de datos se obtiene una primera aproximación a su rango y distribución, así como a su relación con la variable objetivo Exited. Se advierte que las variables predictoras que presentan, a priori, cierta relación con la variable objetivo son Geography, Gender, IsActiveMember, Age, Balance y NumOfProducts. Es esperable, entonces, que los

algoritmos de minería de datos que se utilicen para la predicción del abandono de los clientes incorporen estas variables para efectuar la predicción. De cualquier manera, para la construcción de modelos se consideran todas las variables dado que podrían existir ciertas asociaciones entre ellas, para determinados conjuntos de clientes, que no son fácilmente identificables en un primer análisis.

2.4. Preparación de datos para el modelado predictivo

La preparación de datos implica su limpieza y transformación para la posterior aplicación de modelos. Cabe resaltar que del análisis exploratorio de los datos de clientes se observa que no poseen valores faltantes ni valores atípicos significativos por lo que no resultan necesarias las tareas de gestión de casos de ese tipo. Las tareas de preparación de datos que se llevan a cabo se describen a continuación.

En primer lugar, se eliminan ciertas variables que no poseen valor predictivo y que no tienen relevancia para el problema de negocio bajo tratamiento. Estas variables eliminadas del conjunto son RowNumber, que indica el número de fila de la base de datos, y Surname, que representa el apellido del cliente y no debería tener influencia alguna en la decisión de abandono. La variable CustomerId se retiene como identificador de los clientes pero no se considera en la construcción de los algoritmos.

En segundo lugar, dado que la mayoría de los modelos de clasificación no pueden procesar variables categóricas, se efectúa una numerización binaria de los atributos Gender y Geography. En el caso de la variable Gender se reemplaza el valor Male por 0 y Female por 1. Para la variable Geography, que toma tres valores diferentes, se reemplaza al atributo nominal por tres atributos numéricos binarios.

En tercer lugar, debido a que algunos modelos son sensibles a la escala de los predictores, se estandarizan los valores de las variables numéricas[6]. Estas variables son CreditScore, Age, Tenure, Balance, NumOfProducts y Estimated Salary. Para estandarizar, a cada dato se le resta la media de la variable y se lo divide por el desvío estándar.

En cuarto y último lugar, se aumenta la dimensionalidad del conjunto de datos a partir de la creación de nuevos atributos. Como fue señalado, en la etapa de preparación de datos es usual la creación de variables derivadas de las existentes que puedan mejorar la capacidad de predicción de los modelos, lo que se evalúa en la etapa posterior de

[6] El modelo del Vecino Más Cercano, por ejemplo, computa la distancia euclideana entre registros en función de los valores de los atributos de entrada y, por ello, necesita igualar sus escalas.

modelado. En función del análisis exploratorio de los datos, se estima conveniente la incorporación de tres nuevos atributos.

Dado que la variable Age presenta la mayor correlación con la variable a predecir y que la antigüedad del cliente debería, en parte, responder a la edad, se crea un primer atributo que indica la proporción entre la antigüedad de la relación comercial con el cliente y su edad[7]. Segundamente, se observa en el análisis de los datos que el abandono parece aumentar con la edad de los clientes. Una posible explicación a este fenómeno la constituye el hecho de que los clientes más jóvenes generalmente no tienen un historial crediticio extenso que les permita obtener una buena calificación crediticia y, por ende, mayores y mejores ofertas de instituciones bancarias competidoras. Para evaluar esta hipótesis, se incorpora una variable que calcula el coeficiente entre la calificación crediticia y la edad del cliente. Por último, se incorpora un ratio entre el salario estimado del cliente y su edad ya que se supone cierta relación entre la edad, la antigüedad laboral y el salario obtenido.

Resultaría interesante, además, considerar el ratio entre el balance de cuenta del cliente y su salario estimado, lo que podría dar una mejor idea del uso de la cuenta bancaria. Clientes que presenten un balance bajo en relación a su salario podrían estar canalizando sus depósitos en otras entidades. No obstante, esta variable no se incorpora debido a que no se posee información sobre la metodología empleada para estimar el salario de los clientes y se observan instancias con salarios estimados muy bajos que resultarían en valores atípicos significativos de construirse el coeficiente.

En la Tabla 2 se resumen los atributos construidos, su forma de cálculo a partir de las variables predictoras existentes y sus medidas de posición y dispersión.

[7] No debería esperarse, por ejemplo, que un cliente de 20 años de edad posea una antigüedad con una institución bancaria de 10 años. Se puede suponer que conforme aumenta la edad, mayor es la probabilidad de que el cliente posea una antigüedad mayor.

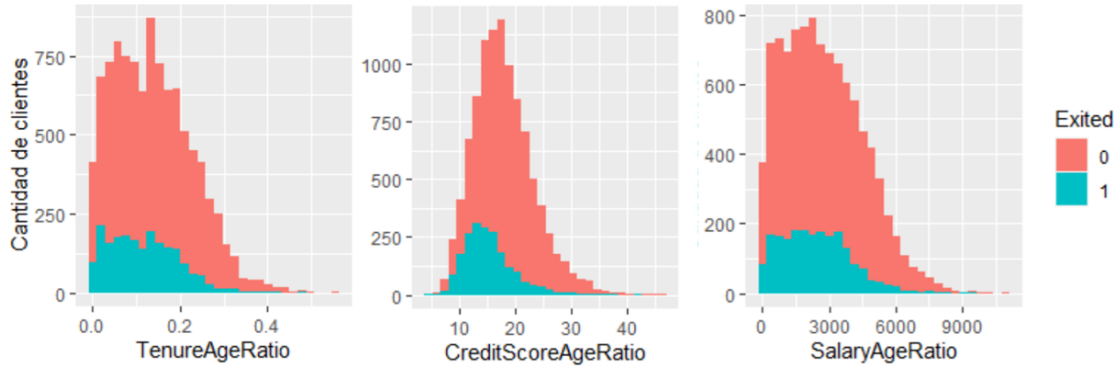
Tabla 2 - Variables derivadas de atributos predictores

	TenureAge Ratio	**CreditScoreAge Ratio**	**SalaryAge Ratio**
Fórmula	Tenure/Age	CreditScore/Age	EstimatedSalary/Age
Mínimo	0,00	4,85	0,25
Cuartil 1	0,06	14,08	1.315,89
Media	0,13	17,87	2.751,42
Mediana	0,12	17,28	2.535,83
Cuartil 3	0,20	20,96	3.979,23
Máximo	0,55	46,88	10.962,34
Desvío estándar	0,08	5,37	1.786,66

Fuente: Elaboración propia con R Studio.

En los histogramas de la Figura 8, se puede observar la distribución de las nuevas variables y se puede notar que las tres variables presentan una asimetría positiva debido a que presentan valores máximos alejados de la media. En el caso de CreditScoreAgeRatio, se puede notar que los clientes que abandonaron la organización se encuentran más concentrados en torno al Cuartil 1. Y efectivamente estos clientes poseen un valor para el coeficiente CreditScoreAgeRatio menor a la media total y menor a la media de los clientes que no abandonaron. En los otros dos ratios se da esta misma situación aunque en menor proporción.

Figura 8 - Histograma de TenureAgeRatio, CreditScoreAgeRatio y SalaryAgeRatio según abandono de clientes



Fuente: Elaboración propia en R Studio.

Los valores de estos nuevos atributos también se estandarizan en preparación para la aplicación de modelos que se desarrolla en el apartado siguiente.

Modelos de predicción de abandono de clientes bancarios

En línea con la metodología propuesta, en el presente apartado se procede con las fases de construcción y evaluación de modelos de predicción de abandono de clientes. A tal fin, en un primer subapartado, se describen los modelos de clasificación que son desarrollados para identificar a los clientes bancarios con mayor probabilidad de abandono. En el segundo subapartado, se estudia el valor para el negocio de los modelos de predicción y se selecciona una medida de evaluación que permita comparar los clasificadores. En un tercer subapartado, se detalla el proceso de construcción de modelos y se resume su evaluación. Además, se elige el método que mejor clasifica a los clientes de acuerdo con su probabilidad de abandono. En el último subapartado, se analizan las principales variables predictoras del abandono.

3.1. Métodos predictivos de clasificación

Dada la naturaleza binaria de la variable a predecir, para el problema bajo estudio, se propone la aplicación de algoritmos de aprendizaje supervisado que permitan obtener resultados binarios. Algunos de los clasificadores generan un resultado binario directo y otros producen una probabilidad de pertenencia a una y otra clase (Glady et al., 2009) y, usando un valor de corte sobre estas probabilidades, asignan predicciones a cada registro (Shmueli et al., 2018). Como ya fue asentado, en este caso, para cada registro del conjunto de datos se posee una variable de interés, “Exited”, que indica si el cliente ha abandonado o no la institución bancaria, y el resto de las variables son posibles atributos predictores.

Se desarrollan y evalúan diferentes modelos predictivos de clasificación, en los que los datos con una clasificación previamente asignada son usados para crear reglas que permitan predecir la clasificación desconocida de nuevos datos. Se aplican diferentes métodos de predicción dado que el rendimiento de los mismos puede variar según las características del conjunto de datos, los tipos de patrones que existan entre ellos, el cumplimiento de los supuestos del método así como el objetivo de análisis (Shmueli et al., 2018).

El primer método de predicción que se emplea es el método del **Vecino Más Cercano**, o K-NN por sus siglas en inglés. Es un algoritmo que busca encontrar *k* registros en una base de datos que sean similares al que se busca clasificar (Evans, 2017). Estos “vecinos” son usados para la clasificación del nuevo registro, al que se le asigna la clase predominante entre sus registros vecinos (Shmueli et al., 2018). Como la clase del registro

objetivo se establece mediante votación, a *k* usualmente se le asigna un número impar para un problema de dos categorías. Una cuestión central para este método es el cómputo de la distancia entre instancias, basado en los valores de los atributos predictores. La medida de distancia más popular es la distancia euclideana (Shmueli et al., 2018) y es la que se emplea en el presente trabajo.

El segundo método que se aplica es el **Árbol de Decisión**. El árbol de decisión es un modelo que se basa en separar los registros en diferentes subgrupos a partir de las variables predictoras, creando con estas separaciones reglas de clasificación (Shmueli et al., 2018). El método realiza sucesivas divisiones de los datos, eligiendo el predictor más relevante en cada una de ellas (Tsiptsis y Chorianopoulos, 2009). Los dos conceptos clave detrás de este modelo son las particiones recurrentes, para construir el árbol, y la poda, para luego recortarlo (Shmueli et al., 2018). El objetivo es la producción de subconjuntos de datos homogéneos en relación a la variable a predecir y heterogéneos en relación al resto de los subconjuntos. Uno de los principales beneficios de este modelo es que produce resultados transparentes y fácilmente interpretables.

En tercer lugar se desarrolla el modelo **Naïve Bayes**. Es un clasificador estadístico basado en el Teorema de Bayes que calcula la probabilidad de que los registros pertenezcan a cada una de las posibles clases, dado un conjunto de predictores (Shmueli et al., 2018). Es decir que el modelo considera los valores de las variables predictoras y obtiene la probabilidad condicionada de que la instancia pertenezca a una u otra clase.

El cuarto método que se emplea es la **Regresión Logística**, que extiende las ideas de la regresión lineal al caso en que la variable objetivo es categórica (Shmueli et al., 2018). Este modelo también busca predecir la probabilidad de que la variable dependiente pertenezca a una u otra clase, basándose en los valores de las variables independientes (Evans, 2017). Para ello, el modelo toma la función *logit* de la variable objetivo y la expresa como una función lineal de las variables predictoras. En la regresión logística intervienen dos pasos: en el primero se obtiene la probabilidad de pertenencia del registro a cada clase; en el segundo, se emplea un valor de corte sobre estas probabilidades con el objetivo de clasificar al registro en una clase.

Otro método que se evalúa es el de **Redes Neuronales**. Las redes neuronales artificiales son sistemas formados por unidades de cómputo, las neuronas artificiales o nodos, conectadas entre sí y capaces de aprender y generalizar relaciones desconocidas entre entradas y salidas a partir de instancias reales. Funcionan a través de diferentes capas de neuronas conectadas entre sí. La capa de entrada contienen las variables predictoras

que se introducen al modelo. La capa de nodos de salida contiene el resultado relativo a la variable a predecir. Las capas intermedias, u "ocultas", son las responsables de asignar diferentes ponderaciones o pesos a los valores de las variables del modelo y transformarlos en una salida, ya sea para la siguiente capa o la salida final de la red. Las redes neuronales utilizan un procedimiento iterativo en el que los resultados predichos son evaluados en función de los resultados observados y los errores encontrados son usados para ajustar las ponderaciones iniciales. La principal desventaja de este modelo es que es considerado una "caja negra" en el sentido de que no permite obtener una explicación de sus predicciones (Tsiptsis y Chorianopoulos, 2009).

Por último, se aplica un ensamble. Un ensamble combina múltiples modelos de aprendizaje supervisado en un "súper-modelo" (Shmueli et al., 2018). Se aplica un ensamble de modelos para evaluar si la capacidad predictiva del mismo es superior. En este caso, se elige la construcción de un **Random Forest** que constituye un ensamble de árboles de decisión. Este método divide el conjunto de datos en diferentes muestras aleatorias con reemplazo y, con ellas, entrena diferentes árboles clasificadores, pudiendo variar también las variables predictoras utilizadas en cada uno de ellos. Para la clasificación final combina las predicciones de todos los árboles del "bosque" obtenido.

3.2. Métrica de evaluación de modelos

Para poder comparar los resultados de los modelos resulta necesario establecer una medida que permita evaluarlos. La mayoría de las métricas de evaluación de modelos de clasificación derivan de la matriz de confusión que resume las predicciones correctas e incorrectas realizadas por los clasificadores. Las filas y columnas de la matriz de confusión corresponden con las clases observadas y las clases predichas, tal como se puede observar en la Figura 9. Las dos celdas diagonales de la matriz de confusión representan el número de clasificaciones correctas en las que la clase predicha coincide con la clase observada. Las celdas que se encuentran fuera de la diagonal de la matriz indican el total de clasificaciones erróneas.

Figura 9 - Matriz de Confusión

	Predicción	
Exited	**0**	**1**
0	Predicción correcta: Verdadero Negativo (VN)	Predicción incorrecta: Falso Positivo (FP)
1	Predicción incorrecta: Falso Negativo (FN)	Predicción Correcta: Verdadero Positivo (VP)

Fuente: Elaboración propia.

Al evaluar modelos de predicción es importante considerar su medida de evaluación en función del valor para el negocio. Para ello, resulta útil expresar el valor que cada resultado de la matriz de confusión representa para el problema de la retención de clientes. En el caso de los verdaderos negativos, clientes que tienen baja probabilidad de abandono y así lo predice el modelo, el banco no llevará adelante ninguna acción por lo que el resultado no aporta beneficios ni costos para el negocio[8]. Los falsos positivos representan clientes que no tienen intención alguna de terminar la relación comercial pero que el modelo predice que sí lo harán. En estos casos, la empresa destinará recursos para contactar al cliente y/o para ofrecerle algún incentivo para quedarse, lo que representa un costo que no trae ningún beneficio asociado ya que el cliente no tenía intenciones de abandonar. Los falsos negativos son clientes insatisfechos y no detectados por los modelos de clasificación, resultando en una pérdida del cliente y de su CLV. Los verdaderos positivos representan clientes con alta probabilidad de abandono, identificados como tales por el modelo, a los que la empresa puede incluir en sus campañas de fidelización. En este caso, la organización incurre en un costo de retención del cliente pero conserva su CLV[9].

Métodos completos de evaluación de modelos de predicción de abandono centrados en el beneficio para el negocio pueden encontrarse en Neslin et al. (2006) y en Verbeke

[8] No se considera, en este punto, el costo fijo de desarrollar e implementar el modelo de predicción del abandono que debería distribuirse entre todos los resultados.

[9] En esta aproximación al valor para el negocio de los resultados del modelo se realizan varios supuestos tales como que el CLV es mayor que el costo de la retención, que la empresa incluirá en sus campañas de retención a todos los clientes con probabilidad de abandono identificados por los modelos, que el costo de las campañas es igual para todos los clientes y que todos los clientes objetivo de las campañas serán retenidos. En la práctica, muchos de estos supuestos generalmente no se cumplen. Usualmente, no hay campañas de retención adecuadas para todos los segmentos de clientes por lo que las empresas consideran otros factores en la decisión de qué clientes contactar (Gold, 2020). Además, las estrategias de retención suelen enfocarse en una menor porción de clientes con una alta probabilidad de abandono y alto CLV (García et al., 2017).

et al. (2012), en los que se calcula el beneficio máximo que se puede obtener de las predicciones incluyendo una fracción óptima de clientes en las campañas de retención. Desafortunadamente, no se posee para el desarrollo del presente trabajo información sobre la estructura de costos de la institución bancaria ni del valor que representan los diferentes clientes, por lo que no es posible implementar medidas de evaluación basadas en una maximización del beneficio. Sin embargo, a los efectos de determinar la medida de evaluación de los clasificadores resulta útil tener presente el valor para el negocio de las diferentes clasificaciones correctas e incorrectas.

Dado que la variable a predecir presenta clases desbalanceadas (los clientes que han abandonado representan una porción menor del conjunto de datos), la exactitud general del modelo no constituye una medida relevante. Esto se debe a que, en casos de clases desbalanceadas, los resultados verdaderos negativos son más comunes que los verdaderos positivos y de menor valor para el negocio. Por la naturaleza desigual en la importancia de las clases y resultados del modelo, cobran relevancia las medidas de sensibilidad y especificidad, las cuales se obtienen de la siguiente manera:

$$Sensibilidad = \frac{VP}{VP + FN}$$

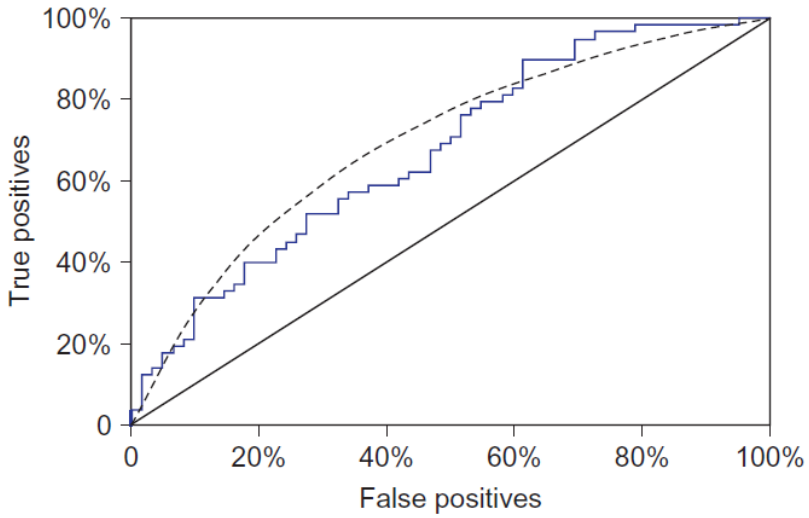
$$Especificidad = \frac{VP}{VP + FP}$$

La sensibilidad representa la probabilidad de éxito del modelo en reconocer un registro positivo como tal (Keramati et al., 2016). En este caso, representa el porcentaje de clientes con una alta probabilidad de abandono que la empresa puede identificar e incluir en sus campañas de retención. La especificidad es la probabilidad de que un registro predicho como clase positiva sea un verdadero positivo. En el problema bajo estudio, una alta especificidad significa que los recursos que la organización destina para sus estrategias de fidelización se dirigen a los clientes correctos.

Por lo expuesto, se emplea el criterio del "área bajo la curva ROC" para la evaluación de modelos. Las curvas ROC evalúan la capacidad predictiva de los clasificadores al graficar en el eje vertical el ratio de verdaderos positivos contra el ratio de falsos positivos en el eje horizontal (Witten et al., 2017). El ratio de verdaderos positivos es la sensibilidad del modelo. El de falsos positivos representa el porcentaje de falsos positivos sobre el total de registros pertenecientes a la clase negativa y es equivalente a $(1 - especificidad)$. Así, la curva ROC muestra las clasificaciones positivas correctas que el

modelo puede obtener a medida que permite más y más falsos positivos. En la Figura 10 se puede observar un ejemplo de curva ROC.

Figura 10 – Ejemplo de curva ROC



Fuente: Witten et al., 2017.

El rendimiento de un modelo es mayor cuanto más se acerca su curva ROC resultante a la esquina superior izquierda. La línea base está representada por la diagonal y representa el rendimiento de una clasificación aleatoria. La medida de evaluación que resume una curva ROC es el "área bajo la curva ROC", también conocida como AUC o AUC-ROC, que toma valores desde 0,5 (clasificación aleatoria o línea base) hasta 1 (perfecta discriminación entre clases). Una interpretación intuitiva del AUC es que estima la probabilidad de que una instancia positiva seleccionada al azar sea correctamente clasificada o asignada una probabilidad mayor de pertenecer a la clase positiva que una instancia negativa seleccionada al azar (Verbeke et al., 2012). En este caso, que la probabilidad de abandono que se asigna a un cliente que abandonó sea mayor que la que se asigna a uno que no lo hizo.

3.3. Desarrollo y evaluación de modelos de predicción de abandono de clientes

Definida la medida de evaluación de los modelos de predicción, se procede a su desarrollo que comienza con el entrenamiento de los modelos. Para que un modelo de clasificación pueda asociar patrones en los datos con clases específicas, es necesario que el modelo se entrene o aprenda de casos en los que la clase es conocida. Para ello, el conjunto de datos que se posee se divide en tres subconjuntos, el mayor de los cuales se destina al entrenamiento de los modelos.

Una vez que un modelo identifica las relaciones entre las variables predictoras y la variable objetivo a partir del conjunto de datos de entrenamiento, se evalúa su capacidad de predicción sobre un segundo subconjunto de los datos, desconocido para el modelo, llamado conjunto de prueba. Este conjunto de prueba se usa para comparar resultados de diferentes modelos o de diferentes versiones de un mismo modelo, variando sus parámetros, para poder seleccionar aquellos modelos y parámetros que mejoran la capacidad de predicción. La tercera partición del conjunto de datos, o conjunto de validación, se utiliza para evaluar el rendimiento de los mejores modelos elegidos sobre nuevas instancias o nuevos clientes, en este caso.

Para evitar que los modelos sobreajusten a los datos de entrenamiento y obtengan resultados pobres sobre los datos desconocidos se aplica en este trabajo una validación cruzada. La validación cruzada es un procedimiento que implica partir el conjunto de datos en *k* subconjuntos o "*folds*". El modelo, entonces, es entrenado *k* veces y, en cada una de ellas, uno de los *folds* es utilizado como conjunto de prueba y el resto como conjunto de entrenamiento (Shmueli et al., 2018). La evaluación del modelo resulta del promedio de las evaluaciones de los *k* modelos entrenados (Keramati et al., 2016). En este trabajo se destina un 80% de los datos para el entrenamiento y prueba de los modelos, aplicando validación cruzada de 10 *folds*, y un 20% de los datos para la validación. En todos los casos el muestreo es estratificado para garantizar igual proporción de la clase objetivo en todas las muestras.

El entrenamiento de modelos se desarrolla sobre la herramienta RapidMiner Studio[10]. Para todos los modelos, se considera para el entrenamiento el conjunto de variables originales de datos de clientes bancarios y luego se introducen de a uno los atributos derivados en la etapa de preparación de datos. En el modelo final se consideran los atributos derivados que mejoran el rendimiento del modelo, lo que resulta de comparar el AUC obtenido con y sin ellos. Únicamente en el caso de la regresión logística, la variable creada como el coeficiente entre la calificación crediticia de los clientes y su edad (CreditScoreAgeRatio) mejora la predicción y es considerada en el modelo final. En el resto de los modelos ninguno de los atributos derivados resulta en una mejor evaluación, por lo que se construyen los modelos finales con las variables originales del conjunto de datos.

[10] RapidMiner Studio versión 9.9.002. Educational Edition registered to Candela Beyreuther. Copyright © 2001-2021 RapidMiner GmbH

Como parte del desarrollo y evaluación de modelos se realiza una evaluación no sólo de los diferentes modelos sino de cada uno de los modelos bajo diferentes parámetros. Haciendo uso de la función *Optimize Parameters (Grid)* de RapidMiner Studio se realiza una optimización de parámetros de cada uno de los modelos, en la que se prueban diferentes combinaciones de los parámetros, alterando sus valores dentro de ciertos rangos, para encontrar aquél conjunto que mejore el resultado. Los valores óptimos de los parámetros, resultantes de este proceso, se utilizan para el entrenamiento de los modelos finales. En el Apéndice I se incluye un detalle de los parámetros incluidos en la optimización, los valores de cada uno de ellos, el valor asignado por defecto por RapidMiner Studio y el mejor valor encontrado.

Tabla 3 - Evaluación y comparación de modelos con parámetros por defecto y parámetros óptimos

Modelo	**Parámetros por defecto**		**Parámetros óptimos**	
	AUC-ROC entrenamiento	AUC-ROC validación	AUC-ROC entrenamiento	AUC-ROC validación
Árbol de decisión	0,786	0,801	0,840	0,838
Vecino más cercano	0,790	0,794	0,822	0,835
Regresión Logística	0,771	0,777	0,771	0,777
Naïve Bayes	0,781	0,802	0,781	0,802
Red Neuronal	0,848	0,854	0,846	0,855
Random Forest	0,839	0,853	0,863	0,873

Fuente: Elaboración propia en base a resultados obtenidos en RapidMiner Studio.

En la Tabla 3 se resume la primera evaluación de los modelos comparando el área bajo la curva ROC obtenida de ellos, tanto con parámetros por defecto como con parámetros óptimos. En el caso del Árbol de Decisión, Vecino Más Cercano y Random Forest, se puede notar cómo la optimización de parámetros mejora los resultados obtenidos de la predicción de los modelos. En el caso de la Regresión Logística, Naïve Bayes y Red Neuronal, los parámetros óptimos no distan de los parámetros asignados por defecto por lo que la evaluación en uno y otro caso es muy similar.

La evaluación de los modelos se realiza también comparando el AUC resultante del conjunto de entrenamiento y prueba y el obtenido sobre el conjunto de validación. Como fue resaltado, la capacidad predictiva de los modelos no debe ser solamente

evaluada sobre datos conocidos para el modelo sino en relación a su capacidad para predecir datos nuevos. Después de todo, el principal objetivo de estos modelos es predecir la probabilidad de abandono de clientes que todavía no lo han hecho para que la organización pueda actuar a tiempo. Se compara el AUC obtenido de los modelos sobre el conjunto de entrenamiento y sobre el conjunto de validación para detectar un posible sobreajuste. Es esperable que el AUC del modelo sobre los datos de validación sea inferior al obtenido con los datos de entrenamiento (ya que el modelo aprendió de estos datos) pero una diferencia muy grande entre ambos indicaría un sobreajuste del modelo a los datos de entrenamiento.

En función a los resultados obtenidos y expuestos en la Tabla 3 se puede concluir que no hay un sobreajuste de los modelos a los datos de entrenamiento ya que, en todos los casos, no se observan discrepancias significativas entre el AUC obtenido del conjunto de entrenamiento y prueba y el obtenido aplicando el modelo al conjunto de datos de validación. El modelo que menor capacidad predictiva presenta, bajo mejores parámetros, es la Regresión Logística con un AUC de 0,78. Los modelos que mejor rendimiento obtienen son el Árbol de Decisión, la Red Neuronal y el Random Forest, que encabeza la lista con un AUC de 0,87.

Para continuar con la evaluación de los modelos y poder elegir un modelo a implementar en la organización para la predicción de abandono de clientes, se evalúan los tres mejores modelos, bajo sus parámetros óptimos, aplicando diferentes estrategias de muestreo de los datos. Como fue notado, los modelos son primeramente evaluados aplicando un muestreo estratificado. No obstante, en un problema de predicción de abandono de clientes con clases desbalanceadas, los modelos de clasificación podrían experimentar dificultades en distinguir qué clientes tienen mayor probabilidad de abandono (Verbeke et al., 2012). Por ello, para balancear las clases, se recurre a diferentes estrategias de muestreo, como el sobremuestreo y el submuestreo, con el objetivo de evaluar si posibilitan una mejor capacidad predictiva de los modelos.

Tanto el sobremuestreo como el submuestreo buscan generar igual distribución de clases en la variable objetivo para mejorar el entrenamiento de los modelos. El sobremuestreo de la clase minoritaria significa que instancias de ella son duplicadas para igualar en cantidad las instancias de la clase mayoritaria. Así, el sobremuestreo no agrega nuevos datos al conjunto sino que aumenta la disponibilidad de algunos de ellos. El submuestreo, por su parte, es una técnica que balancea los datos dejando la clase minoritaria intacta y tomando una muestra de la clase mayoritaria. Con esta técnica se

pierden datos ya que el submuestreo no considera todas las instancias de la clase mayoritaria.

Es importante resaltar que al aplicar técnicas de submuestreo y sobremuestreo, la distribución de la variable objetivo es modificada sobre el conjunto de entrenamiento mientras que en el conjunto de validación se mantiene la proporción original. Esto resulta de la importancia de evaluar los clasificadores sobre un conjunto de datos que sea representativo del que recibirá el modelo una vez implementado. En la Tabla 4 se presenta el AUC obtenido del Árbol de Decisión, Red Neuronal y Random Forest bajo los tres tipos de muestreo.

Tabla 4 – Evaluación de mejores modelos bajo diferentes técnicas de muestreo

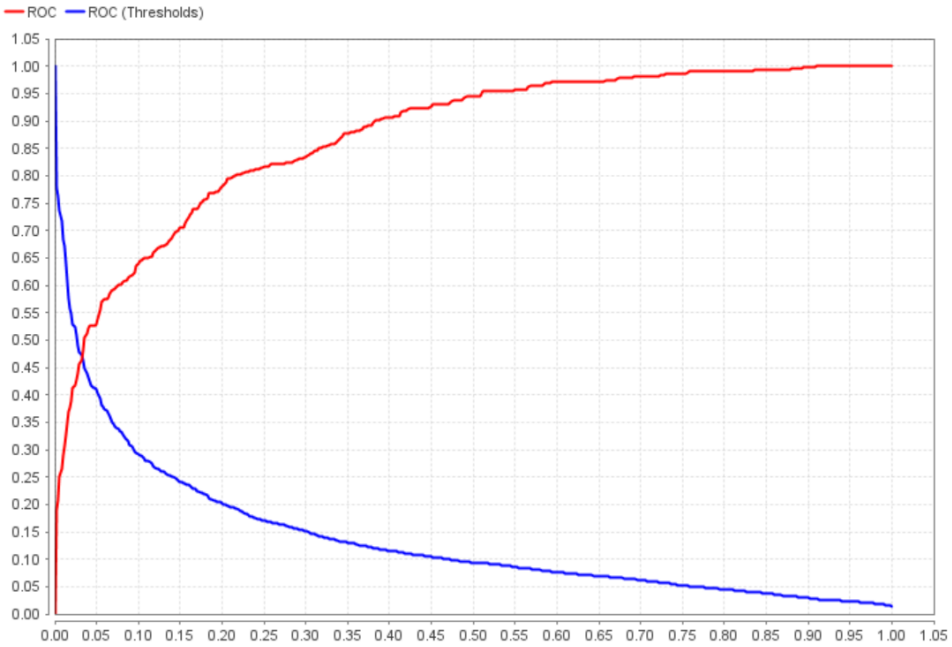
Modelo	**AUC-ROC entrenamiento**			**AUC-ROC validación**		
	Muestreo estratificado	Sobre-muestreo	Sub-muestro	Muestreo estratificado	Sobre-muestreo	Sub-muestreo
Árbol de Decisión	0,840	0,877	0,830	0,838	0,823	0,832
Red Neuronal	0,846	0,695	0,841	0,854	0,700	0,833
Random Forest	0,863	0,926	0,860	0,873	0,866	0,865

Fuente: Elaboración propia en base a resultados obtenidos en RapidMiner Studio.

En el caso del Árbol de Decisión y Random Forest, se puede notar que, al aplicar sobremuestreo, aumenta el sobreajuste a los datos de entrenamiento ya que el AUC obtenido de ellos se incrementa notablemente mientras que el AUC cae al aplicar los modelos al conjunto de datos de validación. Los resultados obtenidos de la Red Neuronal con sobremuestreo resultan muy pobres en relación al muestreo estratificado debido a que, bajo esta técnica, la red produce predicciones, en su mayoría, correspondientes a la clase positiva. El submuestro, por su lado, no produce sobreajuste pero arroja peores resultados, tanto en el entrenamiento como en la validación, para los tres modelos.

En resumen, el modelo que resulta en la mejor predicción del abandono de clientes es el ensamble Random Forest, bajo parámetros óptimos, considerando para la predicción las variables originales del conjunto de datos y aplicando un muestreo estratificado para la partición del conjunto de datos y validación cruzada. Este modelo presenta la curva ROC que se muestra en la Figura 11 y obtiene un área bajo la curva ROC de 0,873. La matriz de confusión obtenida por el modelo sobre el conjunto de validación se presenta en la Tabla 5.

Figura 11 - Curva ROC de Random Forest



Fuente: Elaboración propia en RapidMiner Studio.

Tabla 5 - Matriz de confusión de Random Forest sobre conjunto de validación

	Predicción	
Exited	**0**	**1**
0	1551	42
1	232	175

Fuente: Elaboración propia en base a resultados obtenidos en RapidMiner Studio.

El AUC obtenido por el modelo implica una mejora del 74,6% respecto del escenario base de ausencia de modelo o predicción aleatoria. Adicionalmente, el modelo obtiene una especificidad de 80,65% lo que significa que el 80% de los clientes que predice tienen una alta probabilidad de abandono efectivamente son clientes pertenecientes a la clase positiva. Como fue señalado, una alta especificidad significa una mejor asignación de los recursos que la organización destina para retener a los clientes. La sensibilidad obtenida por el modelo es del 43% y representa el porcentaje de clientes propensos al abandono que el modelo logra identificar correctamente. De esta manera, si la organización tuviera éxito en sus estrategias de fidelización podría correctamente identificar y retener a estos clientes, logrando reducir la tasa de abandono al 11,6%.

3.4. Principales predictores del abandono de clientes bancarios

Una vez seleccionado el mejor método de clasificación, se propone estudiar sus características y analizar si, además de la predicción, permite obtener un conocimiento de las causas de la pérdida de clientes o los motivos que generan mayor propensión de abandono. Para ello, se evalúa el peso que el modelo otorga a las diferentes variables para identificar los principales predictores del abandono. Este análisis es de utilidad para la comprensión de los factores más relevantes en la satisfacción de los clientes y para el diseño de estrategias de retención efectivas.

Tabla 6 - Importancia de variables en Random Forest

Variable	Peso	Peso acumulado
CreditScore	0,21	0,21
Age	0,19	0,40
EstimatedSalary	0,15	0,55
Balance	0,14	0,69
Tenure	0,10	0,80
NumOfProducts	0,06	0,86
Geography = France	0,03	0,89
Gender	0,03	0,92
IsActiveMember	0,03	0,94
HasCrCard	0,02	0,96
Geography = Spain	0,02	0,98
Geography = Germany	0,02	1,00

Fuente: Elaboración propia en base a resultados obtenidos en RapidMiner Studio.

Del análisis de los pesos de las variables para el modelo Random Forest, expuestos en la Tabla 6, una primera apreciación la constituye el hecho de que 5 de las 12 variables consideradas representan el 80% del peso total. Estos principales predictores son la calificación crediticia del cliente, su edad, su salario estimado, su balance de cuenta y la antigüedad de su relación comercial con el banco: variables, en su mayoría, comerciales y transaccionales.

Resulta interesante notar que la calificación crediticia toma el lugar de mayor relevancia para la predicción del abandono cuando, inicialmente, en un análisis exploratorio del conjunto de datos, no se observaba una relación entre esta variable y el abandono. Esto podría significar que clientes con una mejor calificación obtienen

mayores y/o mejores ofertas de empresas competidoras. La edad, por su parte, es la variable más correlacionada con el abandono y efectivamente toma un peso significativo en la predicción del modelo. Como fue señalado, los clientes que abandonan tienen una edad promedio mayor que los que no lo hacen. La tercera variable, por orden de importancia, es el salario estimado del cliente. En este caso, dada la relevancia de esta variable para la predicción del abandono, la organización podría reevaluar su metodología de estimación o buscar obtener el dato del ingreso que efectivamente obtienen sus clientes. Otra variable que presenta un peso similar al salario estimado es Balance, que es una de las tres variables que presenta correlación con la variable objetivo. Tenure también se encuentra en el grupo de principales predictores a pesar que no se observa una diferencia significativa en la antigüedad promedio de los clientes que abandonaron y los que no lo hicieron.

Otro hallazgo relevante lo constituye el hecho de que la mayoría de las variables que presentan, en un primer análisis, cierta relación con el abandono de clientes no toman un peso relevante en el modelo de clasificación. Específicamente, en un inicio se relaciona al abandono con el país de residencia de los clientes, habiendo una proporción mayor de clientes pertenecientes a la clase positiva de Alemania, aunque esta variable es la que menor peso presenta en el modelo. Tampoco se observa un peso significativo para el modelo de las variables Gender, IsActiveMember y NumOfProducts, cuando el análisis exploratorio sugiere lo contrario.

Conclusión

En este trabajo se abordó la problemática del abandono de clientes en una empresa del sector de servicios financieros y bancarios y se buscó responder el interrogante ¿cómo se pueden identificar los clientes con mayor propensión a abandonar los servicios del banco de modo de poder implementar estrategias de retención sobre ellos? Para ello, el objetivo planteado era evaluar la posibilidad de identificar los clientes bancarios con una alta probabilidad de abandono por medio de modelos de minería de datos de clasificación.

Para ello, en un primer apartado, se introdujo la importancia de la retención de clientes bajo el paradigma de la CRM, el problema del abandono de clientes en empresas del sector bancario y el aporte de la minería de datos que posibilita la identificación temprana de clientes con alta probabilidad de abandono. En un segundo apartado, se presentó la metodología CRISP-DM que se aplicó a lo largo del trabajo. Además, se

realizó un análisis exploratorio y descriptivo de los datos de clientes bancarios y se prepararon los datos para el modelado. En el tercer apartado, se describieron los modelos de clasificación propuestos, se eligió una medida de evaluación de los modelos y éstos fueron desarrollados y evaluados aplicando optimización de parámetros y diferentes técnicas de muestreo. Por último, se seleccionó el modelo que mejor clasifica a los clientes bancarios de acuerdo a su probabilidad de abandono y se analizó la importancia relativa de las variables predictoras.

El principal resultado obtenido es que efectivamente se alcanza el objetivo propuesto en el trabajo ya que, con los datos disponibles de los clientes bancarios, es posible construir un modelo de predicción que distinga a los clientes según su probabilidad de abandono. De los seis modelos desarrollados, el modelo que mejor clasifica a los clientes bancarios según su probabilidad de abandono es Random Forest, el cual obtiene un área bajo la curva ROC de 0,87. La implementación de este modelo significa que la organización puede identificar aquellos clientes insatisfechos y con mayor propensión a prescindir de los productos y servicios del banco, de modo de diseñar estrategias de fidelización orientadas a ellos. La combinación del conocimiento del negocio y los resultados del modelo de minería de datos implican que la organización puede optimizar la gestión de la relación con sus clientes y ganar importantes ventajas competitivas.

Como fue resaltado, el abandono y la retención son dos caras de una misma moneda y, por ello, los resultados obtenidos de un modelo de predicción del abandono son relevantes desde el punto de vista de la retención. El resultado obtenido de Random Forest implica que la organización puede realizar una eficiente gestión de los recursos destinados a las campañas de retención y que, de resultar exitosa en ellas, puede bajar la tasa de abandono de clientes hasta un 11,6%.

Por otro lado, el modelo predictivo permitió obtener una idea de las principales variables que inciden en el abandono, siendo ellas la calificación crediticia del cliente, su edad, su salario estimado, su balance de cuenta y la antigüedad de su relación comercial con el banco. Esta información puede ser usada por la empresa para mejorar su relación con los clientes, optimizar sus procesos así como mejorar y personalizar su oferta de productos.

Una posible ampliación de este trabajo implicaría enriquecer la base de datos de clientes bancarios de modo de contar con más variables relacionadas al consumo de productos y características socio-económicas de los clientes. Además, resultaría

interesante contar con más datos sobre las características del abandono de clientes, tales como cuándo ocurrió, en qué condiciones, por qué canales, etc. Como próximos pasos en trabajos futuros, también sería deseable incluir dentro de la evaluación de modelos un análisis específico de los errores cometidos por ellos. Además, si se contara con más datos de la organización bajo estudio se podría efectuar una evaluación de modelos centrada en la función de valor o beneficio que los modelos significan para la empresa. Otra posible extensión del presente trabajo implica avanzar en la selección de grupos de clientes para incluir en campañas de retención, combinando el conocimiento del negocio con el aporte de los modelos predictivos y un análisis del valor de los clientes.

Referencias bibliográficas

Evans, J. R. (2017). *Business Analytics. Methods, Models and Decis*ions (2nd ed.). Pearson Education Limited.

Feyen, E., Frost, J., Gambacorta, L., Natarajan, H. & Saal, M. (2021). Fintech and the digital transformation of financial services: implications for market structure and public policy. *BIS Papers*, No 117.

García, D. L., Nebot, A. & Vellido, A. (2017). Intelligent data analysis approaches to churn as a business problem: a survey. *Knowledge and Information Systems*, 51, 719–774. https://doi.org/10.1007/s10115-016-0995-z

Glady, N., Baesens, B. & Croux, C. (2009). Modeling churn using customer lifetime value. *European Journal of Operational Research*, 197, 402–411. https://doi.org/10.1016/j.ejor.2008.06.027

Gold, C. (2020). *Fighting Churn with Data: the science and strategy of customer retention*. Manning Publications Co.

Guadarrama Tavira, E., & Rosales Estrada, E. M. (2015). Marketing relacional: valor, satisfacción, lealtad y retención del cliente. Análisis y reflexión teórica. *Ciencia y Sociedad*, 40(2), 307-340.

Kaemingk, D. (2018, August 29). Reducing customer churn for banks and financial institutions. *Qualtrics*. https://www.qualtrics.com/blog/customer-churn-banking/

Keramati, A., Ghaneei, H. & Mirmohammadi, S. M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation*, 2(10), 1-13. https://doi.org/10.1186/s40854-016-0029-6

Kotler, P., & Armstrong, G. (2008). *Principios de Marketing*. Pearson Educación, S.A.

Larose, D. T., & Larose, C. D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining* (2nd ed.). John Wiley & Sons.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J. & Mason, C. H. (2006). Defection detection: measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, 43(2), 204-211.

Pinto, St. K. (1997). Marketing de relación o la transformación de la función de marketing. *Harvard Deusto Business Review*, 4(2), 32-40.

PwC (2020). Banking & Fintech, 2° edición. PwC. https://www.pwc.com.ar/es/publicaciones/assets/fintech-y-bancos.pdf

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R. & Lichtendahl Jr., K. C. (2018). Data Mining for Business Analytics: Concepts, Techniques, and Applicacions in R. John Wiley & Sons.

Tapestry Networks (2021, February 18). How technology is driving competitive advantage in financial services. EY. https://www.ey.com/en_gl/financial-services/how-technology-is-driving-competitive-advantage-in-fs

Tsiptsis, K., & Chorianopoulos, A. (2009). *Data mining techniques in CRM: Inside customer segmentation*. John Wiley & Sons.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J. & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: a profit driven data mining approach. *European Journal of Operational Research*, 218, 211–229.

Vicente, M. A. (2009). *Marketing y competitividad. Nuevos enfoques para nuevas realidades*. Buenos Aires: Prentice Hall - Pearson Education.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data Mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann.

Apéndices

Apéndice I – Optimización de parámetros de los modelos

En la Tabla 7 se incluye, para cada modelo, el listado de los parámetros incluidos en la optimización de parámetros. Para cada parámetro se informan los valores probados, el valor asignado por defecto por RapidMiner Studio y el mejor valor del parámetro según el resultado de la optimización. En el caso en que el parámetro representa una variable binaria o categórica, se incluyen en la optimización todos los posibles valores del parámetro. En el caso en que el parámetro es una variable cuantitativa, se definen los valores o rangos del parámetro en función de las características y tamaño del conjunto de datos y la capacidad de procesamiento de la herramienta.

Tabla 7 - Optimización de parámetros de los modelos

Modelo	Parámetro	Valores del parámetro	Valor por defecto	Mejor valor del parámetro
Arbol de Decisión	Criterion	gain_ratio information_gain gini_index accuracy	Gain_ratio	Information_gain
	Maximal Depth	5, 10, 15, 20, 25, 30	10	10
	Minimal leaf size	2, 4, 6, 8, 10, 12, 14, 16, 18, 20	2	8
	Apply prunning	True, False	True	True
	Apply preprunning	True, False	True	True
Vecino Más Cercano	k	1, 3, 5, 7, 9, 11, 13, 15, 17, 19	5	15
	Weighted vote	True, False	True	True
Regresión Logística	Solver	auto, irlsm, l_bfgs, coordinate_descent_naive, coordinate_descent	Auto	IRLSM
	Standardize	True, False	True	True
	Add intercept	True, False	True	True

Naïve Bayes	Laplace correction	True, False	True	True
Red Neuronal	Training cycles	100, 200, 300, 400, 500, 600, 700, 800, 900, 1000	200	700
	Learning rate	0.01, 0.02, 0.03, 0.05, 0.1	0.01	0.01
	Shuffle	True, False	True	True
	Normalize	True, False	True	True
Random Forest	Number of trees	50, 100, 200, 500	100	500
	Maximal depth	5, 10, 15	10	10
	Criterion	gain_ratio information_gain gini_index accuracy	Gain_ratio	Information_gain
	Apply prunning	True, False	False	False
	Voting strategy	Confidence vote, majority vote	Confidence vote	Confidence vote

Fuente: Elaboración propia en base a resultados obtenidos en RapidMiner Studio.

Anexo – Reporte del mentor

El presente trabajo plantea el problema de implementar estrategias exitosas de retención sobre los clientes más propensos a abandonar un banco. 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.