Master in Management + Analytics

Aplicación de modelos predictivos para cálculo de probabilidad de churn en importante entidad bancaria

Autoría: *Mayor Lupo, Juana*
Año académico: *2023*

UNIVERSIDAD TORCUATO DI TELLA

MASTER IN MANAGEMENT + ANALYTICS

APLICACIÓN DE MODELOS PREDICTIVOS PARA CÁLCULO DE PROBABILIDAD DE CHURN EN IMPORTANTE ENTIDAD BANCARIA

TESIS

Mayor Lupo, Juana

Mayo 2023

Tutor: Martos Venturini, Gabriel

Abstract

The purpose of this document is focused on describing how we've been able to measure the customer churn rate of an important bank by the creation of a predictive model.

From the customer experience sector, we observe the importance of having this valuable information not only to improve our actions towards customers satisfaction but also, to share it with managers to support their decision-making processes. Ensuring their efforts are data-driven rather than assumptions, is key.

In addition, by estimating the probability of turn over per client we will be able to implement more accurate short-term retention programs. This analysis shows that retaining a current customer is more cost-effective than acquiring a new one and that the chances of success increases when managing during the first few months of your customer's life cycle.

The first step in this work was building a robust database using specialized tools and techniques. A features engineering process was carried out to enhance data quality and relevance, allowing the incorporation of highly important variables for predictive analysis. These actions sat the foundation for developing accurate and reliable probability models.

Using non-supervised learning techniques, customers were properly segmented. The outcome groups obtained from the k-means algorithm are then compared to the ad-hoc segments defined by the bank.

Finally, the estimates of the churn probabilities and the results from the clustering model are combined together to define new strategies for customer retention.

Resumen

En el siguiente documento se expondrá como se llevó a cabo el primer modelo predictivo para el cálculo de la tasa de abandono que se realizó en importante entidad bancaria.

Desde el sector de experiencia del cliente de esta entidad, observamos la inminente necesidad de contar con esta información no solo para gestionar mejor las distintas acciones hacia los clientes desde el punto de vista de su satisfacción sino también para reportar a las distintas gerencias involucradas y que sus esfuerzos estén basados en datos y no en meras suposiciones u opiniones sobre la potencial fuga de los clientes.

Además, se demostrará la importancia de la estimación de la probabilidad de abandono por cliente a la hora de implementar un programa de retención a corto plazo. Se comprenderá que resulta más rentable retener a un cliente existente que adquirir uno nuevo, y se resaltará la importancia de gestionarlo durante los primeros meses de su ciclo de vida como cliente del banco.

Los primeros pasos se enfocaron en la construcción de una base de datos robusta, utilizando herramientas y técnicas especializadas. Se llevó a cabo un proceso de ingeniería de atributos para mejorar la calidad y relevancia de los datos, lo que permitió incorporar variables de gran importancia para el análisis predictivo. Estas acciones sentaron las bases para el desarrollo de modelos precisos y confiables en el estudio de la probabilidad de abandono de los clientes.

Utilizando técnicas de aprendizaje no supervisado se segmentaron a los clientes de forma conveniente. Los grupos obtenidos mediante el algoritmo de *k-means* son contrastados con los segmentos definidos de manera ad-hoc por el Banco.

Finalmente se combinan las estimaciones de las probabilidades de fuga y los resultados obtenidos del modelo de *clustering* a la hora de analizar diferentes estrategias de retención de clientes.

Índice

Índice de Tablas

Índice de Figuras

1. Introducción

1.1. Contexto

Dado el contexto actual de necesidad constante de entender el comportamiento efímero de los clientes, una gran entidad financiera (de ahora en más "el banco"), busca comenzar a utilizar sus fuentes de información para entender el comportamiento de los mismos.

El banco, actualmente utiliza sus datos[1], principalmente, para entender lo que ya sucedió, se observa el pasado y se toman decisiones en base a eso. Existen pocos análisis predictivos que ayuden a entender comportamientos o a tomar decisiones adelantándonos a los sucesos. Por este motivo resulta interesante utilizar los datos disponibles para entrenar modelos predictivos que aporten y mejoren procesos ya implementados.

La preocupación de la gran mayoría de las gerencias del banco ante este fenómeno, se debe a que durante el 2021 se notó un leve aumento en la tasa de abandono de los clientes dejando en evidencia que los esfuerzos que se estaban realizando no estaban dando resultado. Pasando de 8.34% entre diciembre 2019 y diciembre 2020 a un 9.09% de diciembre 2020 a diciembre 2021.

Si lo observamos de manera trimestral, durante el 4to trimestre 2020 la tasa de abandono fue de 2.16%, luego, en el 1er trimestre de 2021, pasó a 3.04%, siendo el promedio de los próximos tres trimestres del 2021 2.70%.

Para analizar estos valores de manera crítica, es necesario cuantificar la pérdida, es decir, cuántos pesos se están perdiendo en promedio por cada cliente que abandona, y preguntarnos si es rentable la retención de todos o habría que segmentarlos. A lo largo de este documento se abordó simultáneamente tanto las estimaciones de las probabilidades de fuga de cada cliente como así también una estimación del valor potencial de ellos, con la intención de cuantificar el impacto económico de diferentes políticas de retención.

Además de lo expuesto, es oportuno comprender el contexto desde el punto de vista de los datos disponibles. Al ponernos en contacto con estos, se encontraron inconvenientes en algunas variables que resultaban importantes para el análisis, por ejemplo, aquellas obtenidas de las encuestas de experiencia del cliente como también otras que tenían que ver con los productos disponibles de cada cliente y el uso de los canales. Las mismas fueron trabajadas y analizadas para que puedan o no ser utilizadas en los modelos propuestos.

1.2. Problema

Se plantea como problema la necesidad de saber por qué los clientes abandonan el banco, dando de baja sus productos o dejándolos de utilizar. Entender estos motivos para que

[1] Los datos utilizados por los diferentes equipos se refieren a la información histórica del cliente, que incluye su comportamiento y uso de los diversos canales proporcionados por el banco, así como datos relacionados con la utilización de los diferentes productos que se ofrecen.

luego se pueda implementar una estrategia estructurada de retención de clientes de manera integral.

Este suele ser uno de los grandes desafíos en muchas de las gerencias del banco, desde Experiencia del Cliente, por ejemplo, buscan saber si la satisfacción con respecto al banco influye en esta decisión. Desde Marketing, observan si es por la cantidad de productos que posee o utiliza con más frecuencia o desde la gerencia de Riesgo, por ejemplo, analizan si la mora es un buen indicador.

La retención de clientes es un elemento importante de la estrategia bancaria en el entorno cada vez más competitivo de hoy. Retener a un cliente es más económico que buscar nuevos. Según Reichheld, (1996), citado en Cohen, D. A., Gan, C., Hwa, A., & Chong, E. Y., (2006), plantea que los clientes que mantienen una relación a largo plazo con una entidad financiera suelen sentir una mayor confianza en ella, lo que se traduce en una mayor disposición a adquirir más productos. Además, cuando estos clientes están satisfechos, pueden compartir su experiencia positiva con otras personas, generando así una recomendación boca a boca muy valiosa para la entidad (Reichheld y Kenny, 1990, citado en Cohen, D. A., Gan, C., Hwa, A., & Chong, E. Y., 2006). Si los criterios y procesos de retención de clientes no están bien administrados, los clientes pueden abandonar sus bancos, sin importar cuánto se esfuercen las entidades para retenerlos.

1.3. Objetivo

A través de un modelo predictivo, estimar la probabilidad de *churn* [2]de los clientes con la mayor efectividad posible. Para ello, se estudian varios modelos, permitiendo seleccionar el que mejor predice según los datos disponibles, utilizando distintas métricas que reflejen adecuadamente las necesidades del banco hoy en día.

Como la aplicación práctica es un punto de gran relevancia para el banco, una vez que contemos con la probabilidad de abandono por cliente, es posible plantear dos casos de uso, por un lado, crear un programa de retención de clientes de corto plazo, por ejemplo, con el lanzamiento de una promoción o beneficio de la manera más eficiente preguntándonos estratégicamente a quienes debemos apuntar o tratar de llegar, esto es clave para retener aquellos clientes que tenemos actualmente. El segundo caso de uso, se basa en desarrollar una herramienta que permita identificar patrones y tendencias en el comportamiento de los clientes, lo que puede ayudar a la organización a mejorar la satisfacción y a aumentar su fidelidad, este es un proceso de largo o mediano plazo. Con los resultados obtenidos se buscará entender el comportamiento de los clientes, sus necesidades, y los motivos que provocan su deserción que ayuden a tomar decisiones en base a datos y no suposiciones y lograr así, desarrollar una estrategia de retención de los mismos a largo plazo.

En resumen, el análisis de costo-beneficio de una política de retención, sumado al entendimiento de los motivos principales por los cuales los clientes abandonan, que se llevarán adelante en el desarrollo de esta tesis, resulta esencial para mantener una ventaja competitiva en el mercado y ofrecer una experiencia excepcional al cliente. Los beneficios del banco dependen de cuántos

[2] La "probabilidad de *churn*" o "*churn*" es la estimación de la posibilidad de que un cliente abandone el banco o detenga la relación con él.

clientes se adquieran y el nivel de retención de estos, por lo que se crea la necesidad de saber qué es probable que provoque el abandono. (Velu, A., 2021).

2. Datos

Para estimar los modelos predictivos expuestos en la **sección 3** de este documento utilizaré la base de clientes del banco, que por un tema de integridad de datos hará referencia a diciembre 2021, evaluando que sucedió con ellos hasta agosto 2022.

Es decir, tomando como base los clientes a diciembre 2021, se clasificarán los mismos según lo sucedido hasta agosto 2022, con el fin de clasificarlos como ACTIVOS, DESACTIVADOS y BAJAS. Esta clasificación nos servirá para luego entrenar los distintos modelos en donde la variable a predecir será la fuga.

Es necesario entonces, definir los siguientes conceptos:

- "Clientes activos" serán aquellos que durante el período de enero-agosto 2022 no hicieron *churn*.
- "Clientes dados de baja" o "Clientes desactivados" son quienes no utilizaron los productos del banco por 180 días, es decir, son los que sí hicieron *churn*.

Para elegir la base de datos a utilizar, evalué el *Churn Rate* (tasa de abandono) general del banco utilizando datos históricos. Existen tres grandes grupos de clientes, quienes acreditan haberes, jubilados y luego el resto de los clientes que se los integra en uno llamado "Clientela General". El *Churn Rate* fue de 9,09% durante todo 2021 tomando todos los clientes, es decir los tres grupos, pero si miramos más en detalle, podemos observar el porcentaje para cada grupo en particular:

- Acreditación de Haberes tuvieron un *Churn* Rate de 1,29%
- Los Jubilados obtuvieron una tasa de 3,35%,
- Por último, Clientela General de 17,32%.

Por los valores expuestos y porque a nivel de análisis es interesante saber los motivos y trabajar sobre ellos, se optó por evaluar a los clientes que se encuentran dentro del último grupo.

El grupo de interés consta de 435.938 observaciones y 60 variables predictoras. A lo largo de esta tesis uno de los grandes desafíos fue realizar modificaciones en muchas de las variables para que estas puedan ser utilizadas de la mejor manera posible.

Según Bilal Zorić, A. (2016), en su paper Predicting customer churn in banking industry using neural networks, una vez definido el problema comercial a resolver, la segunda fase de un proyecto de análisis de datos, es la de recopilación y preparación de datos. Durante esta etapa se transforman de tal manera, que puedan ser utilizados por los modelos, realizándose una limpieza, que es el proceso de detección y corrección, o eliminación de registros incorrectos, inexactos o irrelevantes. También aclara que es probable que las tareas de preparación de datos se realicen varias veces hasta lograr la mejor base con los recursos que contamos. Esta fase puede tomar hasta el 80 % de

todo el tiempo de análisis. Y afirma que la calidad de los datos es un desafío importante para lograr que el modelo prediga de la mejor manera posible.

A continuación, se enumeran y describen las variables disponibles en la base de datos.

2.1. Variables originales de la base de clientes disponible

Variables Categóricas correspondientes a cada cliente a diciembre 2021:

Tabla 1. Variables Categóricas

Variable	Tipo	Descripción
Dependencia	Categórica	Indica la sucursal en donde los productos del cliente estan radicados.
HomeBanking	Categórica	Indica si el cliente tiene o no Homebanking.
Sexo	Categórica	Se indica si el clientes tiene sexo femenino o masculino.
Localidad	Categórica	Lugar de residencia del cliente.
CodigoPostal	Categórica	Código según el lugar de residencia del cliente.
Arquetipos	Categórica	Indica la segmentación que se realiza internamente a los clientes.
OperaConOtroBanco	Categórica	Indica, según lo declaro por el cliente en las encuestas si opera o no con otros bancos ademas del banco en cuestión.
GrupoNPS	Categórica	Grupo al cual pertenece según la nota de la variable NPSPromedio. Podria ser Promotor si califico 9 o 10, pasivo si puso 8 o 7 y detractor si es 6 o menos.
Encuestado	Categórica	Variable que indica si el cliente tuvo la posibilidad de ser encuestado en algun momento o no todavia.
BaseRelacionalQ1a	Categórica	Varible que expone los tópicos en los cuales se estructura el comentario que el cliente escribió en la encuesta. Existen 4 variables: Tópicos positivos, negativos, neutros y mixtos.

Variables Numéricas correspondientes a cada cliente a diciembre 2021:

Tabla 2. Variables Numéricas

Variable	Tipo	Descripción
Edad	Numérica	Indica la cantidad de años que tiene el cliente.
Antigüedad	Numérica	Indica la cantidad de tiempo que pasó entre que el cliente comenzó su relación con el banco y diciembre 2021.
NPSPromedio	Categórica	Promedio de la nota por cliente ante la pregunta de recomendación realizada en la encuesta.
CantTotalTCAdic	Numérica	Cantidad de tarjetas de crédito adicionales. Se unifican las tres marcas MASTER, VISA, CABAL.
CantidadPres	Numérica	Cantidad de préstamos por cliente. Existe una variable para cada tipo de préstamo: Hipotecario, Hipotecario UVA, Personales, Personales UVA, Prendarios, Retención de Haberes, Empresa, VADEEBS.
CantTotalTC	Numérica	Cantidad de tarjetas de crédito. Se unifican las tres marcas MASTER, VISA, CABAL.
CantidadCA	Numérica	Cantidad de Cajas de ahorro. Cuento con la variable en pesos y otra en dólares.
CantidadCtaCtePesos	Numérica	Cantidad de cuentas corriente en pesos.
CantConsumoTD	Numérica	Cantidad de consumos realizados en tarjeta de débito.
CantPFCanalesElectronicos	Numérica	Cantidad de Plazos Fijos obtenidos por canales electrónicos.

Variable	Tipo	Descripción
CantidadPlazoFijo	Numérica	Cantidad de Plazos Fijos. Está la variable en pesos y la de dólares.
CantTotalSeguros	Numérica	Cantidad de seguros contratados. Se unifica en una sola variable los distintos tipos de seguros: VIDA, ATM, Patrimonial.
CantCajadeSeguridad	Numérica	Cantidad de Cajas de seguridad utilizadas por cliente.
SaldoCajaAhorro	Numérica	Saldo a fin de mes de Cajas de ahorro. Cuento con la variable en pesos y otra en dólares.
SaldoPres	Numérica	Saldos a fin de mes de préstamos por cliente. Existe una variable para cada tipo de préstamo: Hipotecario, Hipotecario UVA, Personales, Personales UVA, Prendarios, Retención de Haberes, Empresa, VADEEBS.
SaldoCajaHaberes	Numérica	Saldo a fin de mes de cuenta sueldo si lo tuviera.
SaldoCtaCte	Numérica	Saldo a fin de mes de las cuentas corriente. Cuento con una variable en pesos y otra en dólares.
SaldoPromedio	Numérica	Saldos promedio del mes. Existe esta variable para cajas de ahorro en pesos, en dólares y cuenta corriente acreedor y deudor.
ImportePlazoFijo	Numérica	Importe invertido en plazos fijos. Variable en pesos como en dólares.
ConsumoTC	Numérica	Monto total de consumos del mes en tarjetas de crédito. Existe una variable por cada marca: VISA, MASTER, CABAL. Tambien hay una variable agregada que suma las tres variables en una.
ConsumoTD$	Numérica	Monto total de consumos del mes en tarjetas de débito.
SaldoCajaHaberesDebitos	Numérica	Saldos de los débitos de la cuenta sueldo si la tuviera.
NumeroIdentificación	Numérica	Número arbitrario utilizado para identificar al cliente.

2.2. Bases de datos accesorias

Para completar la base de clientes nombrada anteriormente se utilizaron otras fuentes de datos que permitieron agregar variables al modelo. Dichas fuentes fueron importantes para la gestión, y en ciertos casos permitieron mejorar la calidad de la información preexistente.

Las variables agregadas se enumeran a continuación:

2.2.1. Variable “LabelFinal”

Como se nombró en la **sección 2**, la variable a predecir “LabelFinal” fue generada con bases que exponen la variación de estado de un cliente de un mes a otro. Además de contener las variables que tiene la base de clientes tanto de un mes como del mes anterior, nos informa si el cliente fue dado de baja, dado de alta o mantuvo su situación. Gracias a la variable clave “NumeroIdentificación” se pudieron cruzar las bases y contar con la variable “CambioEstado” en la Base de Clientes.

Para generar la variable “LabelFinal”, se evaluó la situación de cada cliente en cada mes desde enero 2022 hasta agosto 2022, tomando como referencia los clientes de diciembre 2021.

La variable "CambioEstado" de la base de datos "Variación estado cliente" de cada uno de los meses, fue incorporada a la Base de Clientes, tal como se explicó anteriormente, para entrenar los modelos predictivos. Una vez que se agregaron estas variables, se analizó la situación de cada cliente hasta agosto de 2022. Por ejemplo, si un cliente estaba activo en enero de 2022, se dio de baja por desactivación en febrero y volvió a estar activo en julio, manteniendo la misma situación en agosto, se considera que este cliente está activo en la base de clientes a utilizar.

Los clientes del banco poseen características que, por su naturaleza, conllevan cambios de estado de un mes a otro, ya sea de activo a inactivo o viceversa. Por esta razón, se consideraron 8 meses para analizar con precisión lo que realmente ocurrió con cada uno de ellos.

La base de variación de estado contiene las siguientes variables:

Tabla 3. Base "Variación estado cliente"

Variables mes 1	Variables mes 2	Variables generales
TIPO CLIENTE	TOTAL CLIENTES Anterior.TIPO CLIENTE	CambioEstado
AdicTCCabal	TOTAL CLIENTES Anterior.TipoDocumento	Dep. unificada
AdicTCMaster	TOTAL CLIENTES Anterior.NumeroDocumento	Edad
AdicTCVisa	TOTAL CLIENTES Anterior.Activo/No Activo	FechaVinculacion
CantCADolares	TOTAL CLIENTES Anterior.GRUPO CLIENTE	Mype con Actividad Comercial
CantCAPesos	TOTAL CLIENTES Anterior.Dependencia	Alta Masiva Policía
CantCtaCtePesos	TOTAL CLIENTES Anterior.AdicTCMaster	FechaSaldos
CantidadPlazoFijoDolares	TOTAL CLIENTES Anterior.AdicTCVisa	FechaUltMovCA
CantidadPlazoFijoPesos	TOTAL CLIENTES Anterior.CantCADolares	CONDICIÓN NUEVA
CantCajadeSeguridad	TOTAL CLIENTES Anterior.CantCAPesos	ISOLSegATM
CantidadPresEmpresa	TOTAL CLIENTES Anterior.CantCtaCtePesos	ISOLSegPATRIMON
CantidadPresHipotec	TOTAL CLIENTES Anterior.CantidadPlazoFijoDolares	ISOLSegVIDA
CantPresHipotecUVAs	TOTAL CLIENTES Anterior.CantidadPlazoFijoPesos	
CantidadPresPers	TOTAL CLIENTES Anterior.CantCajadeSeguridad	
CantidadPresPersUVAs	TOTAL CLIENTES Anterior.CantidadPresEmpresa	
CantidadPresPrend	TOTAL CLIENTES Anterior.CantidadPresHipotec	
CantidadPresRetHab	TOTAL CLIENTES Anterior.CantPresHipotecUVAs	
CantidadVADEEBS	TOTAL CLIENTES Anterior.CantidadPresPers	
TCCabal	TOTAL CLIENTES Anterior.CantidadPresPersUVAs	
TCMaster	TOTAL CLIENTES Anterior.CantidadPresPrend	
TCVisa	TOTAL CLIENTES Anterior.CantidadPresRetHab	
TOTAL PRODUCTOS	TOTAL CLIENTES Anterior.CantidadVADEEBS	
Dependencia	TOTAL CLIENTES Anterior.TCCabal	
TipoDocumento	TOTAL CLIENTES Anterior.TCMaster	
NumeroDocumento	TOTAL CLIENTES Anterior.TCVisa	
Activo/No Activo	TOTAL CLIENTES Anterior.TOTAL PRODUCTOS	
GRUPO CLIENTE		

2.2.2. Variable "Homebanking"

Esta ya existía dentro de las variables de la Base de Clientes. Nos dice "Si", si el cliente usó el Homebanking en el último mes (diciembre de 2021) o "No" si no lo usó, pero al analizarla, se descubrió que la misma estaba incompleta, con muchos valores nulos.

Luego de una búsqueda exhaustiva de datos, conseguí que me brinden desde la Gerencia de Canales Alternativos, una base de todos aquellos usuarios de Homebanking, pero este no podía tomarse en su totalidad ya que incluía usuarios que no lo utilizaban hacía meses. Desde el punto de vista de Marketing y Experiencia del Cliente, consideramos que para que un usuario se considere activo, mínimo tendría que usarlo una vez por mes, por ese motivo, de la base de usuarios solo nos quedamos con aquellos que hayan tenido como última fecha de ingreso algún día de diciembre 2021.

De esta manera, mejoramos la variable existente. El cruce nuevamente se realizó gracias a la variable clave "NumeroIdentificación".

La base de usuarios de homebanking contiene las siguientes variables:

Tabla 4. Base "Usuarios Homebanking"

Source.Name.1	Email
Source.Name.2	Teléfono
Banco	Calle
Usuario	Número
Cuenta	Localidad
DNI	Código Postal
Apellido	Fecha último ingreso
Nombre	Fecha Archivo

2.2.3.Variable "PromedioSaldo4q" y "CantidadPF4q"

Las variables de saldos mensuales, promedios y las cantidades de las distintas cuentas y/o productos exponen una foto del mes en cuestión, en este caso, como se explicó, muestra los valores que el cliente tenia durante diciembre 2021. Sin embargo, no son del todo representativas esas variables, ya que por una cuestión de *timing* o *clearing* los saldos pueden estar distribuidos de manera distinta mes a mes o por cómo se realicen los distintos pagos o cobros podrían llegar a cambiar. Mirar un solo mes puede estar perdiendo información relevante sobre el verdadero comportamiento de los clientes.

Por ese motivo, se armaron variables que traigan el promedio de los saldos o de las cantidades del último trimestre del 2021 (4q). De esta manera, se captan ciertas estacionalidades o desfasajes que puedan existir logrando así información más certera de los clientes.

Las variables incorporadas son:

Tabla 5. Variables agregadas cuarto trimestre

Variable	Tipo	Descripción
PromedioSaldoJubi4q	Numérica	Promedio del saldo de caja de ahorro jubilados del 4q 2021.
PromedioSaldoPromedioCAPesos4q	Numérica	Promedio de los saldos promedio del mes de CA Pesos durante el 4q 2021.
CantidadPFUSD4q	Numérica	Promedio de la cantidad de plazos fijos en dólares del último q del 2021.
PromedioSaldoHaberes4q	Numérica	Promedio del saldo de la cuenta sueldo del 4q 2021.
PromedioSaldoPromedioCAUSD4q	Numérica	Promedio de los saldos promedio del mes de CA USD durante el 4q 2021.
CantidadPFPesos4q	Numérica	Promedio de la cantidad de plazos fijos en pesos del último q del 2021.
PromedioSaldoCtaCtePesos4q	Numérica	Promedio del saldo de las cuentas corrientes del 4q 2021.

2.2.4.Variable “Grupo NPS”, “NPS Promedio” y “BaseRelacionalQ1a”

Uno de los principales hitos a tratar durante el desarrollo de la tesis, es encontrar o descartar cierta relación entre la variable NPS y el abandono de los clientes. El NPS es un indicador comúnmente utilizando en Experiencia del Cliente, que indica que tanto nos recomiendan. El NPS se calcula en base a las respuestas obtenidas a la pregunta de qué tan probable es que nuestros clientes nos recomienden a un amigo/colega/familiar. Se consideran Detractores a todos aquellos que responden del 0 al 6, Pasivos a los que responden 7 u 8 y Promotores quienes califican con 9 o 10. El NPS es la diferencia de la proporción de promotores menos la proporción de detractores.

Tal como exponen los gráficos de la **sección 2.3**, la relación entre el NPS y el abandono de los clientes pareciera ser un tanto débil, sin embargo, descartar el NPS para el análisis de *churn* no es la mejor opción ya que una vez que tengamos las causas por las cuales el cliente abandona, debemos diseñar el programa de retención de clientes y uno de los puntos importantes a tener en cuenta para ello, será justamente buscar aumentar el NPS. El NPS es una de las mejores herramientas que se suelen utilizar para mejorar o entender distintas medidas. El desafío está en no tomar únicamente el promedio de la nota que responden los clientes encuestados sino buscar armar una variable que capte otras cuestiones.

Hennig-Thurau, T., & Klee, A. (1997) como conclusión de su análisis sobre la relación entre la satisfacción y la retención de clientes ha presentado un modelo conceptual que postula que la relación entre satisfacción y retención de clientes está moderada por la calidad de la relación y debe interpretarse como no lineal. La calidad de la relación se ha introducido como una variable tridimensional que incorpora la percepción de calidad relacionada con el producto o servicio del cliente, la confianza del cliente y su compromiso de relación.
Por otro lado, Richards, (1996), Jones y Sasser, (1995), citados en Cohen, D. A., Gan, C., Hwa, A., & Chong, E. Y., (2006) comentan que la satisfacción del cliente ha sido ampliamente reconocida como un factor crítico en la determinación de la retención del cliente en una organización, según estudios previos. Sin embargo, las organizaciones deben comprender cómo mantener a sus clientes, incluso si aparentan estar satisfechos. Se sugiere que los clientes insatisfechos pueden optar por no abandonar, porque no esperan recibir un mejor servicio en otro lugar. O también, los clientes satisfechos pueden buscar otros bancos porque creen que podrían recibir un mejor servicio en otro lugar. Sin embargo, mantener a los clientes también depende de una serie de otros factores. Estos incluyen una gama más amplia de opciones de productos, mayor conveniencia, mejores precios y mayores ingresos.

Más allá de estos comentarios, a los fines prácticos del desarrollo de esta tesis, se tomó la variable NPS Promedio y se debatió si realmente es correcto lo que se observa en el análisis preliminar.

Sin embargo, nos encontramos con un problema al momento de tomar las encuestas para ser incorporadas en la base de clientes que se utilizó luego para entrenar los modelos predictivos, la tasa de respuesta de la misma es muy baja (entre el 7% y el 10%), contienen una gran cantidad de valores nulos.

Se imputaron los datos faltantes utilizando una media condicional, es decir que, completamos el valor de NPS con el promedio de las notas de aquellos que tenían mismo Arquetipo, misma cantidad de plazos fijos en pesos durante el cuarto trimestre de 2021 y si uso o no Homebanking durante diciembre 2021 (utilizando la variable descripta en el **punto 2.2.2**). Para

todos aquellos casos (181) que no coincidan con estas características fue reemplazado con el promedio general de la variable NPS original.

Cabe aclarar que para armar la variable NPS se tomaron todas las encuestas que van desde el 01/01/2020 al 31/12/2021 y para todos aquellos que durante ese periodo hayan respondido más de una vez se les realizó el promedio de las notas. Es decir, a cada cliente se le asignó la nota del 0 al 10 con la cual calificó al banco o si fue más de una su promedio.

El motivo del periodo tomado de encuestas se debe a que el envío de la misma, que incluye la pregunta para el armado del indicador, se realiza una vez al mes y solo a 25000 clientes activos o inactivos de manera aleatoria, de los cuales obtenemos entre un 7% y un 10% de respuestas, por lo cual es muy difícil encontrar respuestas si tomamos un periodo más corto y también al tomar un periodo de más de dos años captamos cierta estacionalidad en las respuestas que podría afectar a la nota obtenida.

La variable "Grupo NPS" categoriza según la nota que tiene cada cliente en la variable "NPS Promedio" (completa) en promotores, pasivos y detractores.

Con respecto a las variables llamadas "BaseRelacionalQ1a" que hacen referencia al análisis de sentimiento sobre las respuestas a la pregunta de texto abierto "¿Qué deberíamos mejorar para que nos puntúes con un 10?" que se le hace a todos aquellos que no responden 10 a la pregunta de recomendación, asignarles un valor y completarlas bajo algún criterio llevaría un trabajo muy arduo y quizás corresponda realizar otro documento especializado en esto, pero desde el punto de vista teórico y relacionado con lo expuesto para la variable NPS Promedio en los párrafos anteriores, distintos estudios realizados, entre ellos "The fallacy of the net promoter score: Customer loyalty predictive model" de Zaki, M., Kandeil, D., Neely, A., & McColl-Kennedy, J. R. (2016). explican que, aunque la medida NPS se puede utilizar como un indicador de lealtad, no ofrece una explicación de la causa raíz o las causas de una puntuación baja.

Además, esta medida generalmente se toma al final del camino del cliente, lo que potencialmente oculta los problemas subyacentes de preocupación, que forman la base para identificar mejoras. Confiar únicamente en una métrica simple de un solo cliente es arriesgado y, por lo tanto, se alienta a las empresas a adoptar un enfoque multidimensional más matizado para predecir mejor el comportamiento del cliente utilizando los comentarios que escriben.

Zaki, M., Kandeil, D., Neely, A., & McColl-Kennedy, J. R. (2016) también exponen cuatro ítems distintos que se deberían tener en cuenta a la hora de entender la satisfacción del cliente. En primer lugar, comprender la falta de fiabilidad del NPS o satisfacción general como única medida de lealtad, es decir la nota que coloca el cliente ante la pregunta de NPS.

En segundo lugar, el marco propuesto debe integrar múltiples fuentes de datos de clientes, se debe incluir datos demográficos, de comportamiento y de actitud de los clientes al evaluar la lealtad de este. La combinación de múltiples fuentes de datos respalda la crítica de usar una sola métrica de lealtad. Hacer un juicio sobre la lealtad del cliente sin tener en cuenta los datos de comportamiento es engañoso.

Tercero, un modelo de análisis predictivo que utilice técnicas de big data para predecir la lealtad del cliente e identifique a los clientes que ya no realizan negocios con la organización.

En cuarto lugar, se debe ampliar el enfoque de minería de textos lingüísticos (*Text Analytics*) para determinar el estado de la queja y las emociones de cada cliente mediante la

retroalimentación textual de minería de texto que divide a los clientes en grupos de quejas, neutrales o satisfechos. La integración de los comentarios textuales contribuye a establecer métodos múltiples sistemáticos utilizando un enfoque de big data para capturar y analizar los datos de los clientes de diferentes fuentes.

2.2.5.Variable “CantidadReclamos”

Si un cliente realiza uno o más reclamos sobre algún producto es probable que el mismo esté disconforme con el banco y al largo o corto plazo pueda dejar de operar con la misma.

Sin embargo, este planteo se expone desde la subjetividad, y por ello, se incorpora al modelo la cantidad de reclamos realizados por cada cliente para luego observar si dicha variable afecta o no al abandono.

Para obtener esta variable se cruzaron la base de clientes con la base de reclamos realizados durante todo el año 2020 y 2021, quedando como *feature* la sumatoria del total de reclamos por cada cliente. El cruce, como los anteriores, se realizó con la clave “NumeroIdentificación”.

Las variables de la base de reclamos es la siguiente:

Tabla 6. Base “Reclamos realizados”

NumeroIdentificación	Tipo_Operacion	Prod - subpr - motivo
Documento	Producto	STI y resto
ApellidoNombre	Subproducto	Días
Mail	Motivo	Año
Fecha_Ingreso	Fecha_Derivacion	Nro mes
Sucursal	Fecha_Finalizacion	Status
Suc_Descripcion	Inconveniente	Resuelto96hs
Estado	Sucursal_Origen	Canal de origen

2.2.6.Variable “CantidadIdasSucursal”

Un punto importante a la hora de determinar si un cliente está próximo a desactivarse es observar la contactabilidad que tiene ese cliente con el banco. Más allá que haya usado o no el Homebanking, que es uno de los principales canales, es interesante ver la relación entre el abandono o desactivación de los clientes con la cantidad de veces que fue a la sucursal en los últimos 12 meses.

Por este motivo, y utilizando la base de turneros, es decir, los datos diarios de quienes realizan el *check-in* en alguna sucursal se pudo obtener la cantidad de veces que cada persona fue a la sucursal a lo largo de todo 2021.

Las variables de la base de turneros son:

Tabla 7. Base “Turneros”

Fecha ingreso	Sucursal turnero
Hora ingreso	Nro Doc
Hora finalización	Tipo de cliente
Acción	Espera [min]
Trámite	Atención [min]
Mail AD sin @	Fecha - DNI
Puesto	Prioridad Cita
Perfil de puesto	Nombre de sala de espera

2.2.7.Variable “Rentabilidad”

Hoy en día, no contamos con una medida de rentabilidad por clientes individuo (clientes analizados en este documento), únicamente se tiene rentabilidad a nivel de banca, productos, sucursales, entre otros. Sin embargo, considero que tener una variable en el modelo que represente, de alguna manera, la rentabilidad de cada uno de los clientes, no solo aportaría al modelo en sí, sino que también ayudaría a analizar los resultados desde el punto de vista económico. Interesa saber, por ejemplo, cuanto perdemos por un cliente que se va, entre otras cuestiones relevantes.

Luego de investigar sobre como poder hacerlo con los datos con los que el banco cuenta, y con los informes realizados en la Gerencia de Planeamiento y Control de Gestión, creé la variable a partir de las rentabilidades de los productos que corresponden a la “banca individuos”.

El objetivo era buscar las rentabilidades unitarias de cada producto de manera general, tomando la rentabilidad total y la cantidad de productos que tiene el banco en cartera.

En primer lugar, se buscaron las rentabilidades totales de cada producto, pero al poder estar afectada por cierta estacionalidad opté por tomar la rentabilidad de cada producto durante los últimos 3 meses del año 2021 y utilizar su promedio para el análisis. Lo mismo se realizó con la cartera, se buscó la cantidad de los últimos tres meses y se tomó el promedio.

Se pudo conseguir las rentabilidades y la cartera de los siguientes productos:

- Caja de ahorro: unificado pesos y dólares.
- Cuenta Corriente: unificado pesos y dólares.
- Cajas de seguridad.
- Plazo Fijo: unificado pesos y dólares.
- Préstamos Hipotecarios Pesos.
- Préstamos Hipotecarios UVAs.
- Préstamos Prendarios.
- Préstamos Personales Pesos.
- Préstamos Personales UVAs.
- Préstamo Retención de Haberes.
- Tarjetas de Crédito.
- Seguros: unificado por tipo.

Con esta información, se obtuvo la medida de rentabilidad unitaria de cada producto y se utilizó esta información más la cantidad de cada producto por cliente para obtener su rentabilidad.

$$\boldsymbol{rentabilidad\ por\ cliente} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{X_i} * \boldsymbol{R_i} \qquad (1)$$

i: representa cada producto.

X: cantidad de producto i del cliente.

R: rentabilidad unitaria del producto i.

A modo de ejemplo, para el caso de las tarjetas de crédito, se observa que la rentabilidad promedio del producto durante el cuarto trimestre de 2021 fue de $-393.583.145,96. Por otro lado, la cartera promedio de dicho producto (que representa la cantidad total de tarjetas de crédito colocadas por el banco) durante el mismo período fue de 689.493,67. Esto nos lleva a una **rentabilidad unitaria** (R_i) de $-570,83 ($-393.583.145,96/689.493,67) para el producto tarjetas de crédito.

A continuación, la cantidad de tarjetas de crédito que cada cliente posee (X_i) se registra en la variable "CantTotalTC" en la Base de Clientes. Por lo tanto, la rentabilidad de un cliente en relación al producto se obtiene multiplicando la rentabilidad unitaria del producto por la cantidad de tarjetas de crédito que el cliente tenga. Este procedimiento se aplica de manera consistente para todos los productos, y al sumar cada una de estas rentabilidades, se obtiene la rentabilidad total que el cliente aporta al banco.

Como se mencionó anteriormente, este enfoque de cálculo es simplificado y no refleja completamente el valor real que un cliente aporta al banco. Sin embargo, en términos relativos, sigue siendo una medida valiosa.

2.3. Análisis preliminar de los datos: Observaciones principales.

Una vez que se obtuvo la base a utilizar, y ya incorporada la variable a predecir, según se explicó arriba, realicé un análisis preliminar de los datos, donde se pudieron observar distintas falencias y situaciones que se abordaron oportunamente luego.

2.3.1. Distribución de los datos

En primer lugar, se observó la distribución de clases de la variable a predecir: LabelFinal.

Figura 1. Distribución Base de Clientes

Label	Recuento	%
Churn	149976	34,40%
No Churn	286053	65,60%
Total	**436029**	**100,00%**

Se realizaron distintos gráficos para observar como LabelFinal se relacionaban con algunas variables, como, por ejemplo, Arquetipos y Sexo.

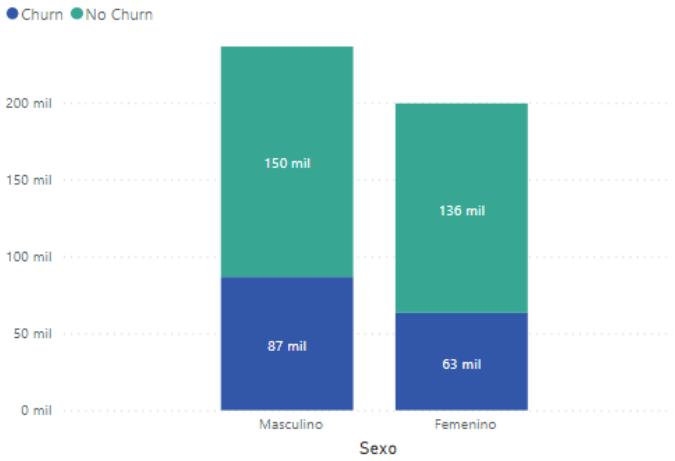


Figura 2. Relación Label - Sexo

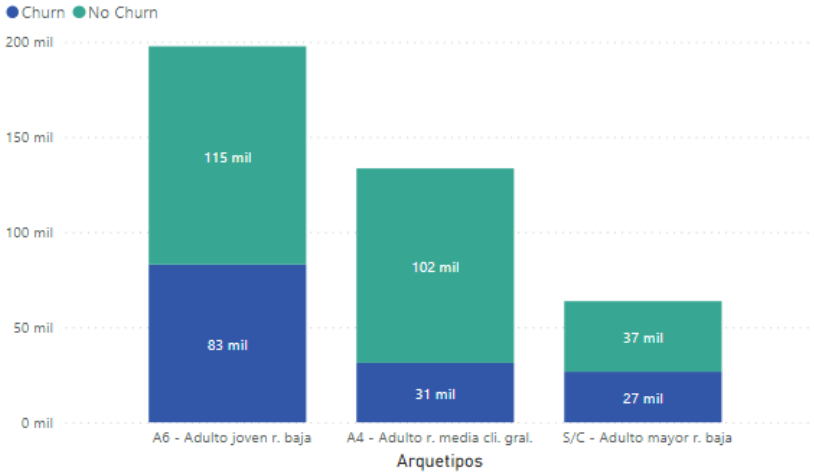


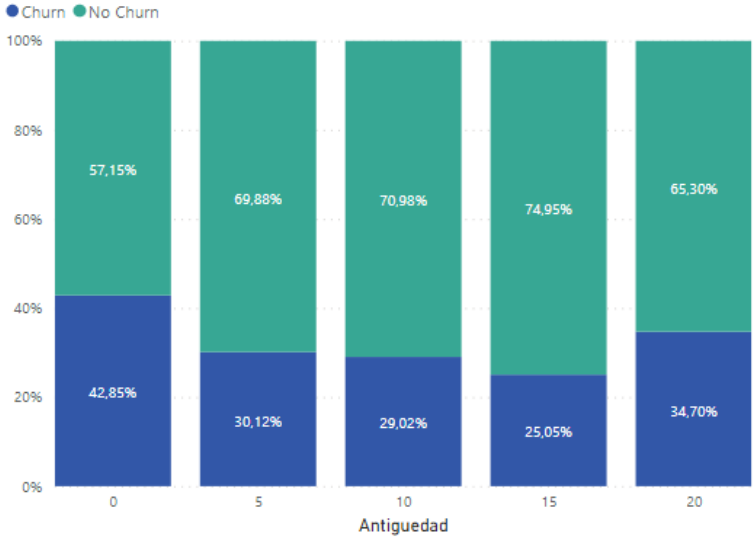
Figura 3. Relación Label - Arquetipo

En la **figura 2**, más allá de que hay mayor cantidad de clientes hombres respecto a mujeres, la incidencia del sexo en el *churn* es muy baja, ya que el 37% de los hombres abandonan mientras que las mujeres lo hace un 31%, es decir, que en un primero momento no se observa una relación entre estas variables.

En cambio, en la **figura 3**, si podemos encontrar diferencias en la incidencia a la fuga según los distintos arquetipos. “A6 – Adulto joven r. baja” y “S/C – Adulto mayor r. baja” en ambos casos, la diferencia entre *churn* y no *churn* es de aproximadamente 16 puntos porcentuales, la cantidad de clientes que no hacen *churn* alcanzan casi el 60% mientras que hacen *churn* un 40%. En cambio, en el caso de, “A4 – Adulto r. media cli. Gral.” un 76% de los clientes no abandonan mientras que solo el 24% si lo hace. Desde la experiencia comercial, esto se debe a que aquellos clientes jóvenes suelen perseguir beneficios o posibilidades de créditos como también la facilidad para operar, que sea de la manera más cómoda y rápida posible, ante alguna situación desafortunada el cliente cambia rápidamente de banco. Con respecto a los adultos mayores, en este caso, los clientes no suelen irse de manera voluntaria, pero existe un % de deserción por deceso que suele ser alto en las estadísticas. Por último, los adultos renta media son aquellos que menos abandonan el banco. Un estudio interno realizado sobre los Arquetipos revela que los adultos con ingresos medios son los que más frecuentan la sucursal bancaria, llaman al centro de atención telefónica y utilizan las redes sociales, mostrando una actitud activa hacia el banco. Además, son aquellos que, en la mayoría de los casos, cuentan con un mayor número de productos contratados en comparación con otros segmentos. Cabe aclarar, que todos los clientes, más allá de su arquetipo y posible motivo de fuga se conservan dentro de la Base de Clientes para entrenar los distintos modelos predictivos.

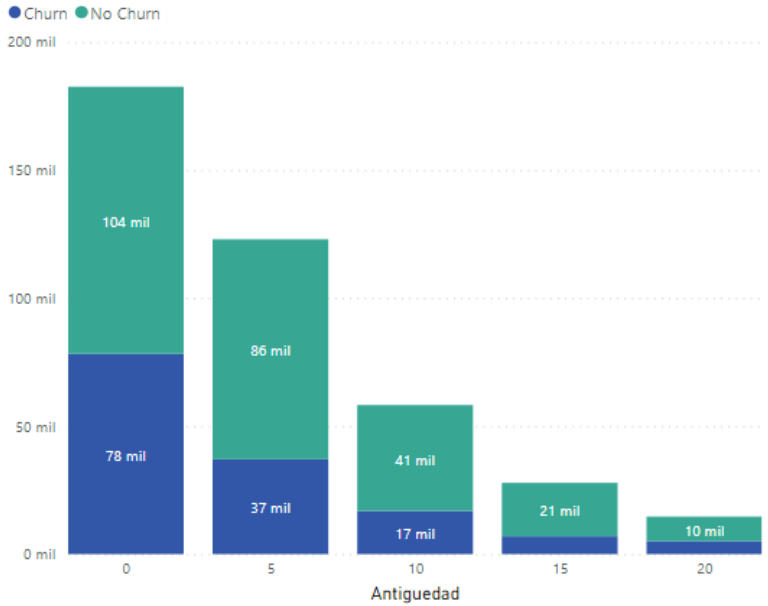
Dentro del análisis parece correcto mostrar cómo se comportan los clientes según su antigüedad, en esta ocasión busqué la comparación si el cliente había hecho o no *churn*:

Figura 4. Relación Label - Antigüedad



En el primer gráfico podemos observar cómo se distribuyen los *churn* y no *churn* según los años de antigüedad que tienen con el banco tomando como 100% el rango de antigüedad de cada bloque y en el segundo tomando la cantidad de clientes en cada rango de antigüedad.

Figura 5. Relación Label – Antigüedad v2



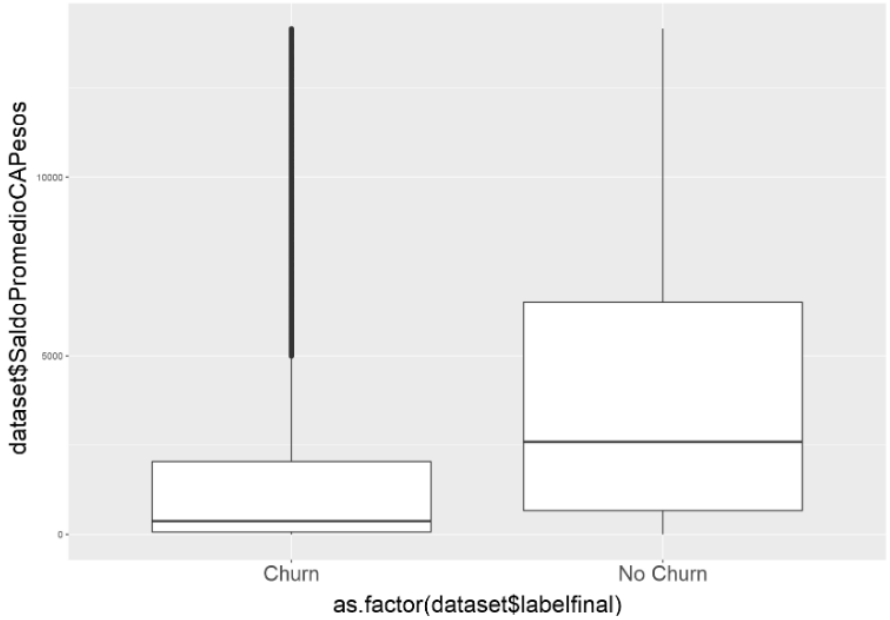
En ambos gráficos, **figura 4** y **figura 5,** podemos ver que la incidencia es mucho más alta durante el primer año de vida como cliente del banco.

Estos datos despiertan una alerta, ya que confirma la necesidad urgente de un programa de retención de clientes en el corto plazo, buscando conservar los nuevos clientes, que probablemente hayan sido captados por alguna campaña específica y nunca lograron fidelizarse.

Una de las variables claves para entender el comportamiento de los clientes es el saldo promedio de la caja de ahorro que deja durante un mes, ya que no solo el banco cuenta con ese dinero de manera diaria, sino que también demuestra que el cliente lo utiliza como su banco principal ya que deja parte de sus ingresos en las cuentas.

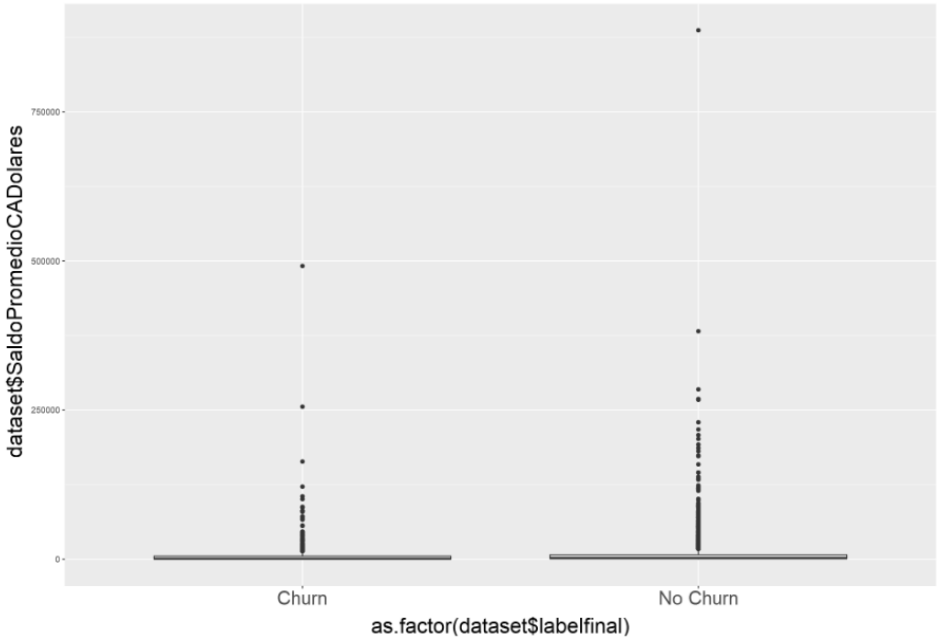
Observamos la distribución de esas variables (tanto pesos como dólares) según si son clientes *churn* o no *churn*:

Figura 6. Relación Label – Saldo Promedio CA Pesos



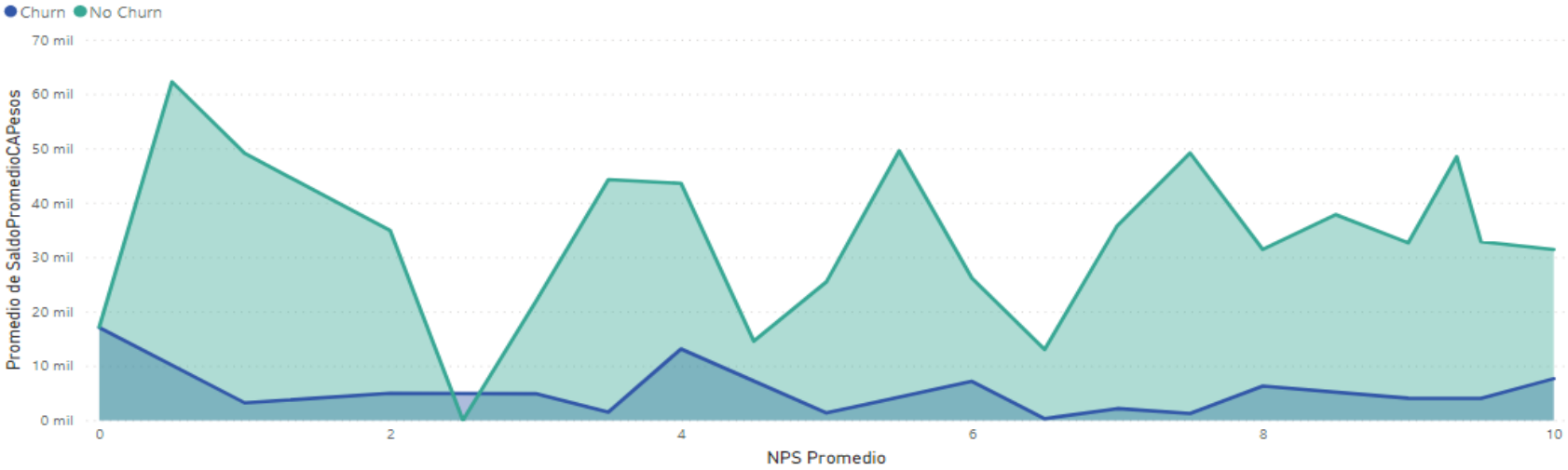
En el caso de pesos, se observa como la mediana del saldo promedio es más alta para aquellos clientes que no hicieron *churn* respecto a los que sí hicieron. Para el caso de dólares, la diferencia es más chica sumado a que los valores *outliers* distorsionan los resultados.

Figura 7. Relación Label – Saldo Promedio CA Dólares



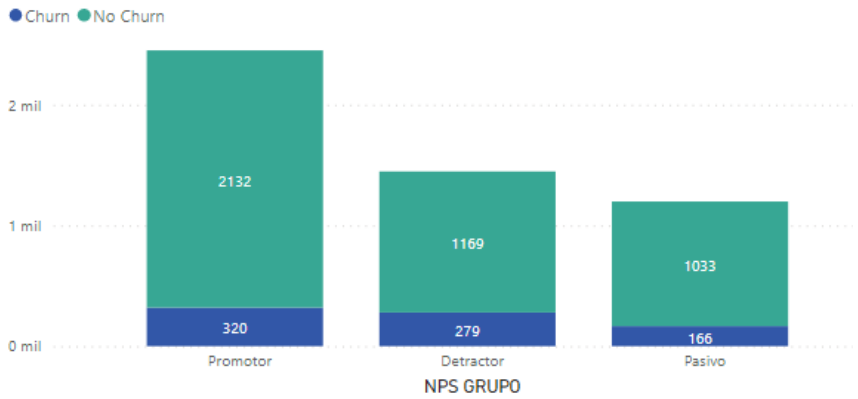
También se analizó la variable a predecir con respecto al NPS (para este análisis se utilizó la variable original, sin las modificaciones explicadas), en un primer gráfico trate de encontrar alguna relación entre la nota que los clientes pusieron con su saldo promedio en pesos, pero como se observa la relación no parece ser tan directa, es decir no se puede confirmar que los que más saldo dejan lo hacen porque están satisfechos con el banco.

Figura 8. Relación Label – NPS Promedio – Saldo Promedio CA Pesos



Lo mismo sucede si observamos la distribución según la nota:

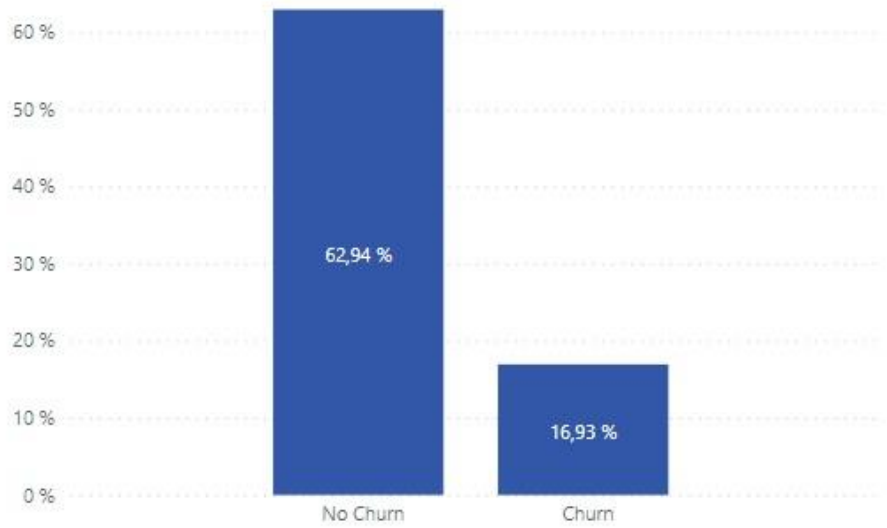
Figura 9. Relación Label – NPS Grupo



De los clientes que hicieron *churn*, existen más promotores que detractores con relación a la cantidad nominal. Aunque la relación en las tres categorías es parecida, más del 80% de los clientes no hicieron *churn* y aproximadamente el 20% si hizo. De esta manera, se confirma la existencia de una relación débil entre el abandono y el NPS, siendo insuficiente para poder explicar las bajas, sosteniendo lo explicado en el **punto 2.2.4.**

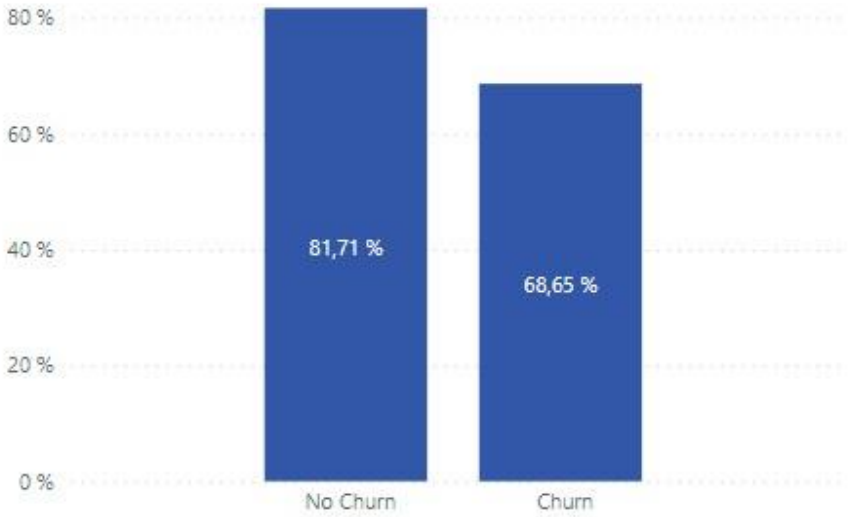
Se observa cómo influye tener tarjetas de crédito y plazos fijos al abandono de los clientes:

Figura 10. Relación Label – Porcentaje de clientes con tarjetas de crédito en cada grupo



En el caso de tarjetas de crédito, se observa que los No *Churn* tienen 62,94% de probabilidad de tener tarjetas mientras que el *Churn* solo el 16,93%.

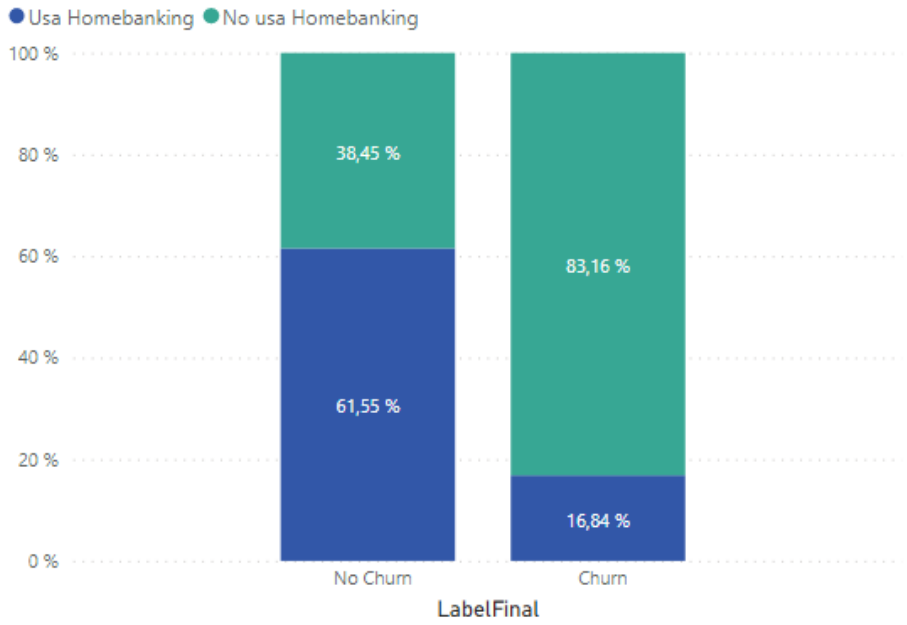
Figura 11. Relación Label – Porcentaje de clientes con plazos fijos en cada grupo



En el caso de Plazo Fijo, es más parejo pasando de 68,65% en el caso de los *Churn* a 81,71% para No *Churn*.

A este análisis previo para entender las variables, también resultó interesante agregar la relación entre el abandono de clientes y el uso de Homebanking. Como se explicó, que el cliente utilice el homebanking te permite entender si al menos este realizó alguna transacción con el banco.

Figura 12. Relación Label – Porcentaje de uso de Homebanking



Como se puede ver, de los clientes que hicieron *churn* solo el 16,84% utilizaba Homebanking, mientras que de los que no hicieron *churn* el 61,55% lo hace. Más allá de si esta variable ayude a predecir al modelo que luego se implementará, es importante entender sus relaciones para gestionar sobre aquellos que no lo utilizan.

Mismo análisis se realizó con la cantidad de reclamos[3], donde se buscó entender cuántos reclamos aproximadamente realizó un cliente que abandonó el banco versus un cliente que sigue activo. Sin embargo, se observa que mientras que los clientes que abandonaron hicieron en promedio 0,03 reclamos (menos de 1 por cliente), los clientes no *churn* hicieron 0,11. Evidentemente no es una variable que implique a la hora que un cliente decida cambiar de banco para operar.

Figura 13. Relación Label – Promedio de reclamos realizados por grupo

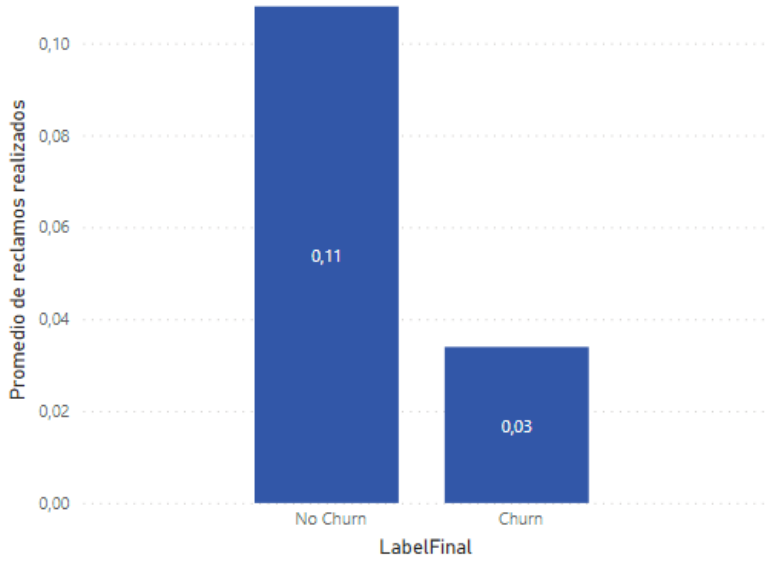
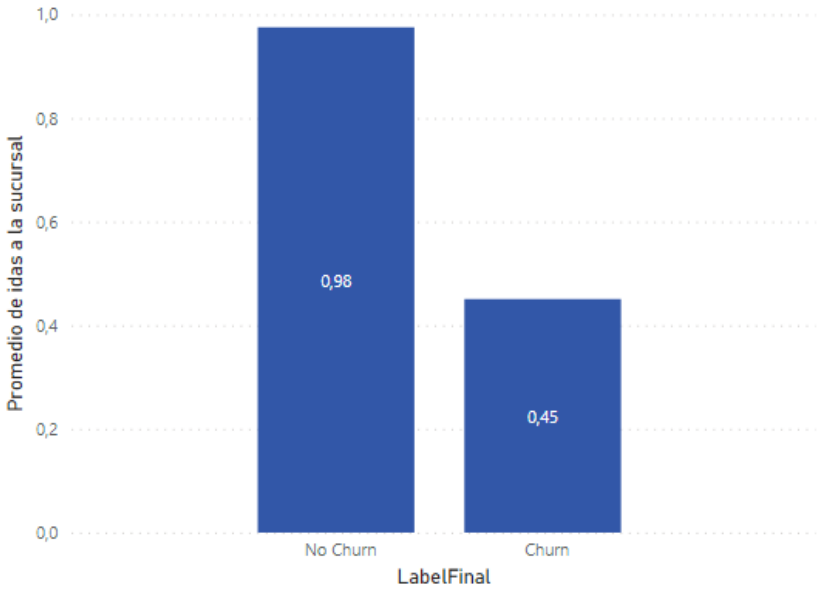


Figura 14. Relación Label – Promedio de idas a la sucursal por grupo



[3] Los reclamos incluyen aquellas quejas o denuncias sobre inconvenientes que tuvieron con algún producto o servicio del banco por factores internos al mismo.

Los clientes que siguen operando con el banco fueron 0,98 veces a las distintas sucursales en promedio, mientras que lo que no lo hacen fueron la mitad de las veces.

De estos últimos tres gráficos podemos intuir que la contactabilidad del cliente, ya sea a través de Homebanking realizando alguna transacción o consulta, si realiza algún reclamo a través del Centro de atención telefónica o bien si va a la Sucursal de forma presencial es un buen indicador que el cliente está interesado en mantener la relación con el banco.

JM Valdez Mendia & JJA Flores-Cuautle (2022) en su artículo “Hacia la experiencia de hiperpersonalización del cliente: un enfoque basado en datos” resalta la importancia de los modelos comerciales omnicanal en la actualidad, los cuales integran tanto puntos de contacto físicos como digitales para interactuar con los clientes. En este contexto, una estrategia de hiperpersonalización se basa en la capacidad de la organización para recopilar y utilizar los datos de los clientes con el fin de crear experiencias personalizadas. Al implementar un plan organizacional de hiperpersonalización, se logran dos objetivos principales: ofrecer experiencias personalizadas y aumentar el número de clientes que reciben dichas experiencias.

Además, es fundamental comprender que cada interacción entre los diferentes puntos de contacto de la organización y sus clientes genera datos. Estos puntos de contacto representan una oportunidad valiosa para que las empresas registren información de los clientes, la cual puede generar valor para ellos.

3. Metodología

Durante el desarrollo del modelo predictivo para estimar la probabilidad de *churn* de los clientes, se realizó una prueba inicial utilizando la base de datos que contenía todas las variables mencionadas en las **secciones 2.1 y 2.2**. Esta prueba se conoce como modelo *benchmark*[4] y se discute en la **sección 3.2**. Posteriormente, se evaluaron los cambios realizados y se comparó con otros modelos predictivos más sofisticados para analizar su impacto en el rendimiento.

Para ajustar el modelo, se realizaron imputaciones de datos faltantes y se corrigieron los valores extremos y atípicos de acuerdo con la metodología detallada en la **sección 3.3**.

Luego se realizaron diferentes técnicas de ingeniería de atributos para pasar por último al planteo de Modelos Predictivos con regularización y modelos de ensamble que capten las distintas relaciones entre las variables según las técnicas correspondientes.

Aunque se siguió el proceso general para el conjunto completo de clientes, se observa cierta variabilidad en la base en términos de la cantidad de productos que cada cliente posee. Por esta razón, se llevó a cabo un análisis no solo de la base de clientes en su totalidad (con los filtros mencionados anteriormente), sino también se probó entrenar distintos modelos, primero, para aquellos clientes con 1 o 2 productos, y luego para aquellos con 3 o más productos. Esto se hizo con el objetivo de comparar el rendimiento entre ambos grupos y determinar si sería beneficioso plantear un modelo predictivo para un grupo de personas con características más homogéneas, o si no es necesario hacer esta distinción.

A continuación, se explica el detalle de cada uno de los pasos que se llevaron a cabo.

3.1. Descripción de las Herramientas de Software

Para llevar adelante el desarrollo de esta tesis se utilizó R-Studio, aplicándose distintas librerías para lograr no solo la implementación de modelos predictivos que proporcionen la *probabilidad de churn*, objetivo principal del documento, sino también se utilizaron distintas librerías y funciones para realizar muchas de las técnicas aplicadas en el arduo proceso de Ingeniería de atributos, explicado en **sección 3.4.**

Algunas de las librerías, paquetes y funciones utilizados fueron:

- ***Librerías "pROC" y "ROCR":*** utilizadas para calcular y graficar la curva ROC de los distintos modelos implementados.
- ***Librería "dplyr":*** utilizada en distintas partes del código de R completo, principalmente para agilizar modificaciones en los datos gracias al conjunto de herramientas para manipulación de datos que proporciona.
- ***Librería "glmnet":*** utilizada para implementar las distintas funciones para entrenar modelos de regresión logística. Según los hiperparámetros utilizados en cada una de ellas, los modelos podían ser con regularización o no, *Lasso, ridge* o *elastic net.*

[4] El Modelo *Benchmark* se refiere al primer modelo utilizado como referencia para posteriormente compararlo con los demás modelos implementados.

- ***Librería "corrplot":*** Permitió realizar gráficos de correlación entre las distintas variables, indispensable para entender los datos y luego tomar decisiones de modelado.
- ***Librería "randomForest":*** Fue necesaria para entrenar no solo el modelo *Random Forest* para obtener la *probabilidad de churn* sino también que al modificar los hiperparámetros de la función se pudo observar la performance en modelos *Bagging*.
- ***Librerías "scatterplot3d", "ggplot2", "gridExtra", "rpart.plot":*** Dichas librerías brindaron herramientas necesarias para realizar muchos de los gráficos utilizados, no solo para ser incorporados en el documento de tesis sino también para comprender los datos, su relación, su comportamiento y tomar decisiones de modelado en base a eso.
- ***Paquete "caret":*** contiene herramientas que agilizan el proceso de desarrollo de modelos predictivos. Permite la separación de datos, en base de *train*[5] *y test*[6], por ejemplo.
- ***Función "Kmeans":*** algoritmo de clasificación no supervisado utilizado para agrupar a los clientes del banco en k grupos según distintas características.
- ***Función "Prcomp (PCA)":*** permite simplificar la complejidad del dataset con muchas dimensiones, como es el caso de la Base de Clientes del banco, pero conservando la mayor parte de su información.

3.2. Medidas de performance en problemas de clasificación

Las métricas que se utilizaron para realizar las comparaciones de performance de los modelos son PRECISION, RECALL, F1-SCORE, TASA DE ERROR y CURVA ROC.

Se calculan a partir de la matriz de confusión que se obtiene una vez evaluado el modelo entrenado en los datos de testeo:

		Predicted class		Total instances
		+	−	
Actual class	+	TP	FN	P
	−	FP	TN	N

Donde:

TP ⟶ El modelo predice positivos cuando lo son.

TN ⟶ El modelo predice negativo cuando lo son.

FP ⟶ El modelo predice positivos cuando no lo son.

FN ⟶ El modelo predice negativo cuando no lo son.

[5] Los datos de entrenamiento o "*Train*" son los datos que se usaron para entrenar los modelo.
[6] Los datos "*Test*" son del mismo tipo que los que forman el *Train Set* pero que no se han empleado para entrenar el modelo.

Precision ⟶ **TP/(TP+FP)**

Recall ⟶ **TP/(TP+FN)**

Para un modelo estimado, en general, al variar el umbral de clasificación, tanto la *Precision* como *Recall* puede mejorar a cambio de una disminución de la otra. Para lograr el *trade off* entre ambas, se debe explorar entre muchos modelos en busca de aquel que mejor explote la relación entre los datos a la hora de predecir la fuga.

F1 - score: (2*prec*rec) / (prec + rec) ⟶ Esta medida trata de ponderar de manera conjunta precision y recall. Tiende a estar cerca del mínimo entre ambas.

Estas dependen del umbral utilizado para calcular la matriz de confusión. Según el umbral que pongamos vamos a obtener distintos valores de *recall*, *precision* y *F1 – score*. Por este motivo, resulta interesante observar medidas que tratan de independizarse del umbral que se define, probando en todos los puntos de corte posibles, una medida popular se deduce en la Curva ROC (AUC).

Curva ROC (AUC) ⟶ Se toma el área bajo la curva ROC (AUC) como un indicador de la capacidad de predicción del clasificador. Evalúa la relación entre el Ratio de Verdaderos Positivos (TP/P) y el Ratio de Falsos Positivos (FP/N). En el mejor de los casos tendrá un área igual a 1 y en el peor de los casos será de 0,5.

Estas medidas de performance son para modelos de clasificación y se decidió calcular varias de ellas, por lo que representan y porque luego permitirán entender en que enfocarnos como institución y apuntar a mejorar, según la necesidad del momento.

A modo de ejemplo, imaginamos distintas situaciones donde se utilizarían cada una de estas medidas:

- Si nuestra prioridad es no gastar innecesariamente plata y enviar una campaña de préstamo a tasa cero solo a clientes para los que estemos seguros que servirá, es decir, que el cliente es probable que cambie de banco, buscamos que nuestro umbral de decisión sea alto y exigente. De esta manera se les enviará la campaña a menos clientes pero que con seguridad se podrían ir del banco y se debe accionar sobre ellos.

- Si nuestro objetivo es aumentar el stock de clientes, lo que se busca es captar la mayor cantidad de personas que se puedan dar de baja, para realizar acciones de retención a los mismos. En este caso, se busca bajar el umbral de decisión contemplando a todos lo que, si se van a dar de baja, pero también a otros que su probabilidad de abandono es baja.

La métrica de interés depende en gran medida del problema en cuestión. Es una buena práctica utilizar una única métrica para guiar las decisiones. Para el desarrollo de la tesis se exponen cada una de ellas para cada modelo entrenado ya que luego se utilizarán según las distintas necesidades, pero se realizarán esfuerzos para mejorar la Curva ROC (AUC).

3.3. Modelo *Benchmark*

3.3.1. Pre – procesamiento de los datos

Al analizar la base, se observó que 20 variables contenían valores faltantes (NA) y otras tantas poseen datos atípicos. La mayoría de ellas relacionadas con los saldos promedio de los distintos productos o cantidad de los mismos, como se comentó en la **sección 3** hay muchos clientes que solo tienen un producto con el banco y esto genera que ellos no aparezcan en ciertas bases que nutren a la base general.

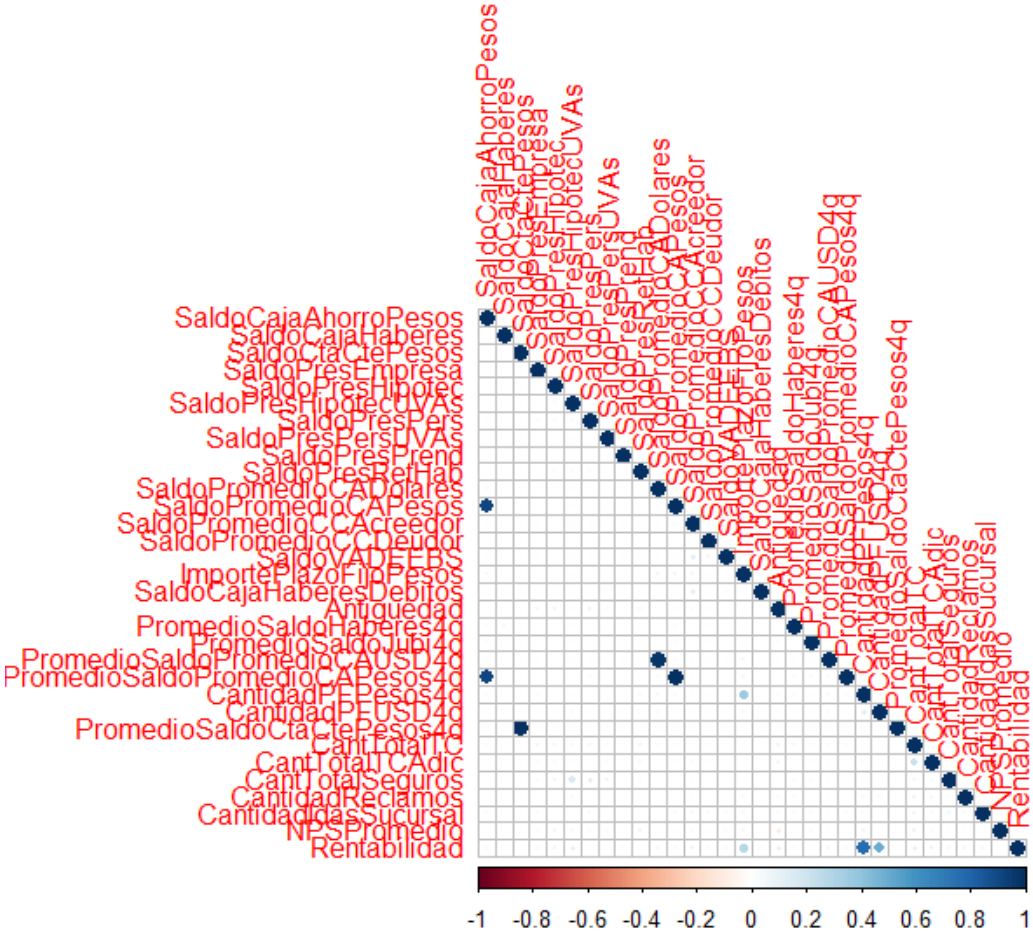
Como se expuso en **punto 2.2.4**, esto mismo sucede con las variables que contienen las respuestas de la encuesta Relacional obtenidas desde Experiencia del Cliente.

En la **sección 3.3** se detalla el tratamiento que se realizó para cada uno de estos valores faltantes o atípicos.

Otro punto importante a tener en cuenta a la hora de armar un modelo predictivo es observar la correlación entre las variables independientes.

Quedándonos con las variables numéricas sin valores faltantes, obtenemos:

Figura 15. Correlación de variables



Cuando existe correlación entre las variables independientes podemos estar teniendo un problema de multicolinealidad.

Las variables independientes deberían ser eso, independientes. Y esto se debe a que, si el grado de correlación entre las variables independientes es alto, no podremos aislar la relación entre cada variable independiente y la variable dependiente (respuesta).

Es decir, cuando las variables independientes están muy correlacionadas los cambios en una variable están asociados con cambios en otra variable y, por lo tanto, los coeficientes de regresión del modelo ya no van a medir el efecto de una variable independiente sobre la respuesta manteniendo constante, o sin variar, el resto de predictores.

Para corregir la multicolinealidad de los datos existen algunas opciones que se pueden utilizar:

- Eliminar algunas de las variables independientes altamente correlacionadas.
- Combinar linealmente las variables independientes, por ejemplo, realizar un PCA para crear nuevos predictores independientes y volver a ajustar el modelo de regresión con ellos.

- Realizar un análisis diseñado para variables altamente correlacionadas, por ejemplo, la regresión de mínimos cuadrados parciales.
- Realizar una regresión que pueda manejar la multicolinealidad, por ejemplo, *LASSO* y la regresión de Ridge.

Para el desarrollo de la tesis se decidió dejar aquellas variables correlacionadas, y buscar solucionar este problema implementando PCA **(punto 3.4.3. Análisis de Componentes Principales – PCA)** cuando se trabajó en ingeniería de atributos y también entrenando un modelo de Regresión *Lasso* con regularización **(sección 3.5. Regresión con Regularización: *Lasso*)**.

Posiblemente el modelo *Benchmark* utilizado se verá afectado ya que, al ser una regresión logística sin regularización, corremos el riesgo que no sean tratadas como corresponde.

3.3.2. Implementación de *modelo Benchmark*

Una vez separada la base en *Train* y *Test* de manera aleatoria, se corrió un primer modelo de Regresión Logística sin regularización.

El modelo de Regresión Logística propone:

$$P\left(Y = 1|X_1, \dots, X_p\right) = \frac{e^{\,\boldsymbol{\beta_0}+\sum_{i=1}^{p}\boldsymbol{\beta_i X_i}}}{1 + e^{\,\boldsymbol{\beta_0}+\sum_{i=1}^{p}\boldsymbol{\beta_i X_i}}} \qquad (2)$$

Es decir, modelamos la probabilidad de fuga de forma no lineal por medio de los parámetros β = (β0, . . ., βp).

Dada $S_n = \{(\mathbf{x}1, y1), \cdots, (\mathbf{x}n, yn)\}$, aprendemos los parámetros del modelo resolviendo:

$$\text{maximize}_{b_0,\mathbf{b}} \sum_{i:y_i=1} \log\left(\frac{e^{\,\boldsymbol{\beta_0}+x_i^T b}}{1+e^{\,\boldsymbol{\beta_0}+x_i^T b}}\right) + \sum_{i:y_i=0} \log\left(\frac{1}{1+e^{\,\boldsymbol{\beta_0}+x_i^T b}}\right) \quad (3)$$

Lo que se busca maximizar es una log-Verosimilitud: probabilidad de observar una muestra como Sn para cada conjunto de valores de los parámetros del modelo.

Cuando trabajamos con un modelo de regresión logística tiene pros y contras, por un lado, podemos afirmar que es un modelo interpretable, no solo podés predecir sino también interpretar los parámetros. Son fáciles de escalar en contextos de big data y sirven para problemas de clasificación ya que devuelve outputs probabilísticos. Sin embargo, es un modelo poco robusto, y no suele funcionar bien cuando existen datos atípicos.

Por este motivo, este modelo únicamente se utiliza a modo de *benchmark*.

Se evaluó la performance en *Train* y luego en *Test* de las distintas métricas antes nombradas en la **sección 3.1. Medidas de performance en problemas de clasificación**, obteniendo los siguientes resultados en *Test*:

Tabla 8. Performance Modelo *Benchmark*

Métrica	Performance
Precision	94,726%
Recall	83,417%
F1 - score	88,712%
Tasa de Error	15,780%
Curva ROC	89,775%

3.4. Imputación de Datos

3.4.1. Valores faltantes

Se llevaron a cabo dos procesos distintos para aquellas variables que contenían valores faltantes. Para todas aquellas variables donde más del 80% de los datos eran NA se optó por eliminarlas del modelo ya que no aportaban a la predicción del abandono de clientes. Las variables que tomaron este destino fueron:

Tabla 9. Variables con más del 80% de los datos vacíos

SaldoCtaCteDolares
BaseRelacionalQ1aPositive
BaseRelacionalQ1aNeutral
BaseRelacionalQ1aNegative
BaseRelacionalQ1aMixed
OperaConOtroBanco

El resto de las variables, que son mayormente variables que indican cantidad de productos o saldos, se completaron con ceros ya que al estar vacíos significaba que no se había encontrado valor para ese cliente en la base de ese producto, es decir no lo posee. Estas variables son:

Tabla 10. Variables con menos del 80% de los datos vacíos

CantCADolares	ImportePlazoFijoDolares
CantCAPesos	PromedioSaldoHaberes4q
CantidadPresHipotec	PromedioSaldoJubi4q
CantPresHipotecUVAs	PromedioSaldoPromedioCAUSD4q
CantidadPresPrend	PromedioSaldoPromedioCAPesos4q
SaldoCajaAhorroDolares	CantidadPFUSD4q
CantPFCanalesElectronicos	PromedioSaldoCtaCtePesos4q

3.4.2. Datos *Outliers*

Para todas aquellas variables que contenían alguno de sus valores *outliers* analizadas según lo expuesto en el **punto 3.2.1** se realizaron cambios correspondientes:

- *SaldoCtaCtePesos:* Los valores mayores al máximo valor con gran cantidad de clientes fue reemplazado por el promedio de la variable.
- *PromedioSaldoCtaCtePesos4q:* Los valores mayores al máximo valor con gran cantidad de clientes fue reemplazado por el promedio de la variable.
- *CantidadIdasSucursal:* Algunos clientes tenían más cantidad de idas a la sucursal que los días del año, estos casos fueron reemplazados por el promedio de la variable.

3.4.3. Prueba en modelo *Benchmark*

Al realizar los cambios enunciados en los **puntos 3.3.1. y 3.3.2** se volvió a correr el modelo *benchmark*, justamente para poder observar como afectaron a la performance de un modelo base los cambios implementados.

El procedimiento para la implementación del modelo *benchmark* es el mismo que el expuesto en el **punto 3.2.2** y los resultados en *Test* fueron los siguientes:

Tabla 11. Performance Modelo *Benchmark* con tratamiento

Métrica	**Performance**
Precision	94,726%
Recall	83,415%
F1 - score	88,712%
Tasa de Error	15,786%
Curva ROC	89,782%

Como se puede observar, el tratamiento de *outliers* no afectó considerablemente a la performance del modelo *benchmark*, sin embargo, para la implementación de los siguientes modelos consideramos importante que estos estén correctamente modificados.

3.5. Ingeniería de Atributos

No importa de dónde obtengamos la información, siempre se destaca la importancia de generar atributos “buenos” o valiosos para entrenar el modelo predictivo que tengamos. Antes de utilizar dicho modelo para entrenar o validar, es crucial realizar un trabajo en los datos y en los diferentes atributos del conjunto de datos. Lo que el algoritmo pueda aprender dependerá en gran medida de los atributos que dispone.

Cuando hablamos de datos "buenos", nos referimos a aquellos que no solo representan los aspectos más relevantes del conjunto de datos, sino que también cumplen con los supuestos

del modelo que queremos utilizar. A veces, aunque tengamos un modelo poco sofisticado, como uno lineal, podemos lograr que sea útil mediante la ingeniería de atributos.

A continuación, se enumeran las modificaciones / incorporaciones que se realizaron durante el proceso de Ingeniería de atributos:

3.5.1.Nuevos ratios y variables

Se generó, en primer lugar, la variable "CantidadTotalProductos", representa la sumatoria de la cantidad de productos que posee cada cliente. Es importante esta variable porque al ser una base heterogénea en cantidad de productos permite reducir este dato en una sola.

Con el mismo objetivo, se generan variables categóricas que serán 1 cuando la cantidad o el saldo es mayor a 0 o 0 si es 0 el saldo o la cantidad que posee un cliente del producto.

Se realizó este tratamiento para aquellas variables de saldos y cantidades que suman el 4to trimestre del 2021, ya que son las variables más representativas que toman los datos de forma integral:

Tabla 12. Variables agregadas

PromedioSaldoHaberes4q
PromedioSaldoJubi4q
PromedioSaldoPromedioCAUSD4q
PromedioSaldoPromedioCAPesos4q
CantidadPFPesos4q
CantidadPFUSD4q
PromedioSaldoCtaCtePesos4q

También se hizo con la cantidad total de tarjetas de crédito y sus adicionales (CantidadTotalTCyAdic).

Se crea una variable similar pero que va a contener un 1 si la cantidad de productos (variable: "CantidadTotalProductos") es mayor a 2 y cero si es menor. Dentro de los distintos clientes que podemos obtener en la base utilizada, la cantidad de productos que posee cada uno, es un buen indicador para ver qué tipo de cliente es.

Resumiendo, se agregaron al dataset, que se va a utilizar para entrenar el modelo, las siguientes variables:

Tabla 13. Nuevas variables categóricas

CantidadTotalProductos	CatePromedioSaldoPromedioCAUSD4q
CantidadTotalTCyAdic	CatePromedioSaldoPromedioCAPesos4q
CateCantidadTotalTCyAdic	CateCantidadPFPesos4q
CatePromedioSaldoHaberes4q	CateCantidadPFUSD4q
CatePromedioSaldoJubi4q	CatePromedioSaldoCtaCtePesos4q
CateCantidadTotalProductos	

3.5.2. Segmentación de clientes – *kmeans*

Más allá que en la base contamos con el dato de Arquetipo, que es una segmentación que se realizó desde la Gerencia de Experiencia del Cliente para identificarlos en grupos y así poder entender sus preferencias, sus usos, su satisfacción, entre otros puntos, parece oportuno aprovechar esta tesis para realizar una segmentación a través del algoritmo *k-means*, dejando de lado la subjetividad que puede llegar a tener el armado del Arquetipo de forma no automatizada.

Los algoritmos de aprendizaje no supervisado, como es el caso de *k-means*, solo utiliza los valores conocidos (variables independientes) y el objetivo es descubrir patrones interesantes detrás de los datos.

El aprendizaje no supervisado es más desafiante que el supervisado, es más subjetivo a la hora de tomar decisiones, principalmente porque no se dispone de una métrica clara a optimizar. Muchas veces (como en este caso) forma parte de la etapa del preprocesamiento de los datos.

Según argumentos teóricos que James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013) exponen en An introduction to statistical learning (Vol. 6, Capítulo 10), *Clustering* se refiere a una serie de técnicas cuyo objetivo es encontrar subgrupos o "*clusters*" en un conjunto de datos.

La idea es particionar los datos de manera que:

1- Las observaciones que pertenecen a un grupo sean similares entre ellas;
2- Las observaciones de grupos distintos sean distintas entre ellas.

K-means clustering es uno de los tantos algoritmos de *clustering* y busca dividir al conjunto de datos en K subconjuntos distintos sin solapamiento. Se debe fijar el valor de K antes de correr el algoritmo.

Los *cluster* deben cumplir las siguientes condiciones:

1- Todas las observaciones deben estar en algún *cluster*.
2- Ninguna observación debe pertenecer a más de un *cluster*.

K-means asume que una buena asignación es aquella que dado un valor K minimiza lo más posible la variabilidad intra *cluster (within-cluster variation).*

Entonces si W(Cj) es una medida que indica cuanto las observaciones de un *cluster* j difieren entre sí. El problema se puede escribir como:

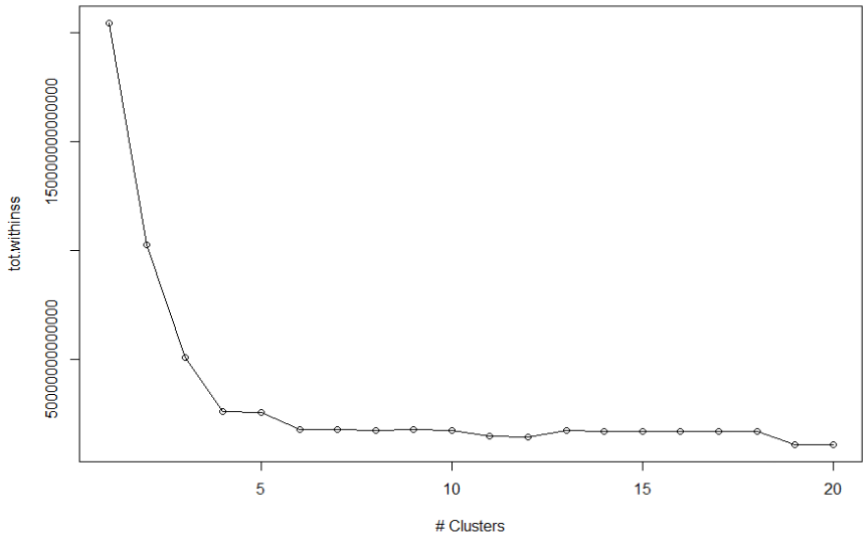
$$\mathbf{minimize}_{C_1,\ldots,C_K}\left\{\sum_{k=1}^{K} W(C_k)\right\} \qquad (4)$$

Por cuestiones prácticas y para mejor interpretación de resultados, este algoritmo solo se implementó utilizando cinco *features*: Edad, Antigüedad, CantidadIdasSucursal, Rentabilidad y CantidadTotalProductos.

Antes de entrenarlo, escalamos las variables ya que tenían distinta unidad de medida.

Primero, se observó cómo evoluciona la función objetivo a medida que aumenta K, corriendo modelos de *K-means* con K que vayan del 1 al 20, se elegirá como valor de K aquel en donde aparezca el "codo" ya que ahí es donde el error deja de bajar tan pronunciadamente, no vale la pena seguir aumentando.

Figura 16. Gráfico de codo.



Se asignaron las observaciones en 4 clusters quedando distribuidas de esta manera:

Tabla 14. Distribución de las observaciones en *clusters*

1	2	3	4
40759	302937	91221	1017

Para interpretar los clusters se observaron los promedios:

Tabla 15. Interpretación de *clusters*

Cluster	Edad	Antiguedad	CantidadIdasSucursal	Rentabilidad	CantidadTotalProductos
Cluster 1	53,91	6,34	43,73	11200,3	2,38
Cluster 2	41,92	3,77	0,66	3005,57	1,98
Cluster 3	48,71	7,58	1,28	45855,83	5,11
Cluster 4	62,60	13,82	0,57	5234,13	2,00

- *Cluster 1* incluye aquellos con edad y antigüedad media, que suelen ir bastante a las sucursales, tienen una rentabilidad media/alta y por ende una cantidad de productos también media/alta.
- *Cluster 2* son clientes relativamente nuevos, jóvenes, que no suelen ir a las sucursales y que como tienen pocos productos su rentabilidad es baja.
- *Cluster 3* incluye clientes de edad media/alta que van a la sucursal y que tienen una rentabilidad alta.

- Por último, el *Cluster 4* son personas mayores, clientes con muchos años de antigüedad, que ya no suelen ir a las sucursales y que la cantidad de productos que tienen y su rentabilidad es baja.

Luego del análisis realizado esta variable se incluyó dentro del dataset que se va a utilizar.

3.5.3. Análisis de Componentes Principales – PCA

Otra técnica de aprendizaje no supervisado es el análisis de componentes principales (PCA) que también va a ser utilizada, en este caso, como parte del proceso de ingeniería de atributos.

Gil Martínez, C. (2018, junio) en su trabajo Análisis de componentes principales (PCA) indica que una de las aplicaciones de PCA es la reducción de dimensionalidad (variables), perdiendo la menor cantidad de información (varianza) posible: cuando contamos con un gran número de variables cuantitativas posiblemente correlacionadas (indicativo de existencia de información redundante), PCA permite reducirlas a un número menor de variables transformadas (componentes principales) que expliquen gran parte de la variabilidad en los datos.

Cada dimensión o componente principal generada por PCA será una combinación lineal de las variables originales, y serán además no correlacionadas entre sí.

Tal como hicimos con el algoritmo *k-means*, se aplicó PCA solo a algunas variables numéricas, se buscó incorporar aquellas que estaban correlacionadas, estas variables también se escalaron para aplicarles PCA.

Las variables incorporadas fueron:

Tabla 16. Variables utilizadas en PCA

CantidadPFPesos4q	Rentabilidad
CantidadPlazoFijoDolares	SaldoCajaAhorroPesos
PromedioSaldoPromedioCAUSD4q	SaldoPromedioCAPesos
PromedioSaldoPromedioCAPesos4q	PromedioSaldoCtaCtePesos4q
CantidadPlazoFijoPesos	SaldoPromedioCADolares
CantidadPFUSD4q	

La autora también expone que el objetivo es identificar las combinaciones lineales que mejor representan las variables (X1,...,Xp). Sean (Z1,Z2,...,ZM) M<p combinaciones lineales de las p variables originales, es decir:

$$Zm = \sum_{j=1}^{p} \phi_{jm} \boldsymbol{X_j} \qquad (5)$$

donde $\phi_{1m}, \phi_{2m}, \ldots, \phi_{pm}$ son las cargas o *loadings* de los componentes principales (por ejemplo, ϕ_{11} correspondería al primer loading de la primera componente principal). Los loadings dan idea sobre qué peso tiene cada variable en cada componente. Cada vector de loadings, de longitud igual a p, define además la dirección en el espacio sobre el cual la varianza de los datos es mayor.

La combinación lineal se normaliza para no inflar la varianza, por lo que la suma de cuadrados de los *loadings* se iguala a 1.

La primera componente principal (Z1) es aquella cuya dirección refleja o contiene la mayor variabilidad en los datos (por lo que esta componente será la que más información contenga). Este vector define la línea lo más próxima posible a los datos y que minimiza la suma de las distancias perpendiculares entre cada dato y la línea representada por la componente (usando como medida de cercanía el promedio de la distancia euclídea al cuadrado):

$$\boldsymbol{Z_{i1}} = \phi_{11}\boldsymbol{X_{i1}} + \phi_{21}\boldsymbol{X_{i2}} + \cdots + \phi_{p1}\boldsymbol{X_{ip}} \quad (6)$$

donde ϕ_{11} corresponde al primer *loading* de la primera componente principal.

En otras palabras, el vector de *loadings* de la primera componente principal resuelve el problema de optimización:

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2 \right\} \text{ subject to } \sum_{j=1}^{p}\phi_{j1}^2 = 1 \quad (7)$$

La segunda componente principal (Z_2) será una combinación lineal de las variables, que recoja la segunda dirección con mayor varianza de los datos, pero que no esté correlacionada con Z1. Esta condición es equivalente a decir que la dirección de Z_2 (vector ϕ_2) ha de ser perpendicular u ortogonal respecto a Z_1 (vector ϕ_1).

La varianza total presente en los datos se define matemáticamente como:

$$\sum_{j=1}^{p}\text{Var}(X_j) = \sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n}x_{ij}^2 \quad (8)$$

mientras que la varianza explicada por la m-ésima componente principal se corresponde con

$$\frac{1}{n}\sum_{i=1}^{n} z_{im}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jm}x_{ij}\right)^2 \quad (9)$$

Aplicando lo expuesto a las variables que se eligieron se observan cuales son aquellas PCA que absorben el mayor porcentaje de la varianza del modelo.

Figura 17. Gráfico de PCA – Explicación de varianza

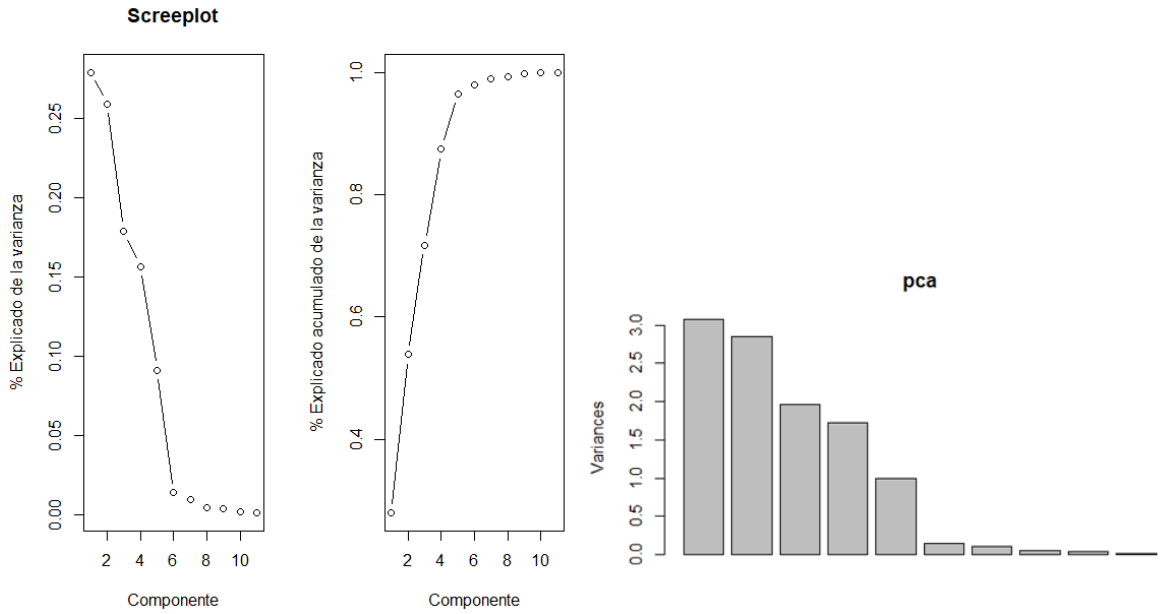


Tabla 17. Varianza explicada por las variables PCA

0,2791968	0,2591791	0,1789511	0,1564971	0,091058	0,0139594	0,0097423	0,004534	0,0037714	0,0020156	0,0010951

Tanto en el gráfico de "codo" como en el de barras y también observando los valores de varianza explicada, muestran que es óptimo elegir las primeras cinco PCA explicando el 95% de la varianza del modelo.

Estas variables fueron incorporadas al dataset total utilizado para entrenar los modelos.

3.6. Regresión con Regularización: *Lasso*

Realizado el análisis planteado hasta el momento y con el estudio de los distintos *features* incorporados, comienza el proceso de entrenamiento de modelos más complejos y sofisticados que se espera que logren mejorar la performance obtenida con respecto al modelo *Benchmark* expuesto en la **sección 3.2**. Tenerlo nos ayudará a comparar resultados y determinar cuál es el mejor modelo o cuál es la mejor forma de utilizar los datos.

En este paso, vamos a probar un modelo de Regresión Logística con regularización, en este caso utilizamos *Lasso.*

También aquí James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013) en su libro An introduction to statistical learning (Vol. 6, Capítulo 6), explican que cuando p >> 0, los modelos de regresión logística sin regularización suelen presentar mucha variabilidad. Ante perturbaciones en los datos de *Train* se producen cambios considerables en los parámetros estimados en el modelo.

Para controlar estos cambios, se debe introducir una restricción de presupuesto sobre los parámetros del modelo que controla la variabilidad (hiperparámetro $\boldsymbol{\lambda}$). Se busca, de cierta forma, penalizar la log-verosimilitud que se intenta maximizar con un término que cuantifica la complejidad del modelo logístico.

$$\text{maximize}_{b_0,\mathbf{b}}\left\{\ell(b_0, \mathbf{b} \parallel \mathbf{S_n}) - \boldsymbol{\lambda}((\mathbf{1}-\boldsymbol{\alpha}) \parallel \mathbf{b} \parallel_2^2 + \boldsymbol{\alpha} \parallel \mathbf{b} \parallel_1\right\} \qquad (10)$$

Permite la eliminación de *features* con menor relevancia, colaborando a eliminar el *trade off* entre sesgo y varianza.

Además de las bondades generales de los modelos de regresión, permite eliminar la posibilidad de *overfitting.*

Lo importante en este modelo es la elección del valor óptimo del hiperparámetro $\boldsymbol{\lambda}$ ya que:

- Cuando tenemos un $\boldsymbol{\lambda}$ muy pequeña, habrá mucha varianza y poco sesgo, tendremos riesgo de *overfitting* o sobreajuste.
- Ante $\boldsymbol{\lambda}$ muy grandes, los parámetros estimados se van haciendo cero y el aporte de cada *feature* a la variable de respuesta es cada vez menor. En este caso, tendremos poca varianza, pero mucho sesgo. Corriendo riesgo de *underfitting* o subajuste.

3.6.1.Valores arbitrarios de $\boldsymbol{\lambda}$

En primer lugar, se probó un modelo de Regresión Logística con regularización *Lasso* utilizando un parámetro $\boldsymbol{\lambda}$ arbitrario, se realizó el mismo proceso utilizado para el modelo *Benchmark* **(punto 3.2.2. Implementación de modelo *Benchmark*)** pero entrenando el modelo con valores de $\boldsymbol{\lambda}$ distintos de cero.

Probamos el modelo eligiendo $\boldsymbol{\lambda}$ = 0,008, primero con los datos de *Train* y luego en *Test* obteniendo los siguientes resultados:

Tabla 18. Performance Regresión Logística con regularización arbitraria

Métrica	Performance
Precision	94,646%
Recall	83,252%
F1 - score	88,583%
Tasa de Error	15,970%
Curva ROC	89,380%

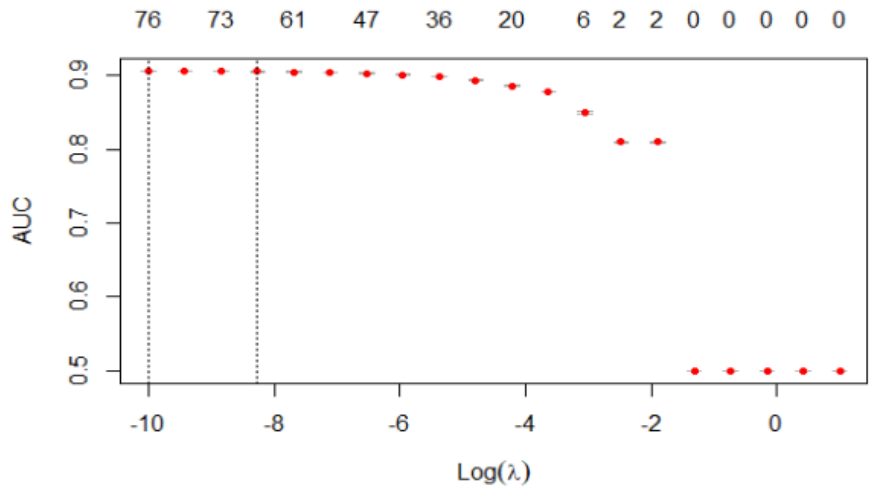
3.6.2. Lambda vía *K-fold cross-validation*

Sin embargo, elegir el parámetro de manera arbitraria puede no ser la mejor idea a la hora de entrenar y predecir en base a modelos predictivos.

Buscamos ajustar λ a través de *K-fold cross-validation*. Para este caso, usaremos K = 5.

Utilizando una grilla con 20 valores distintos de λ, graficamos como varia la curva ROC en función de los distintos valores.

Figura 18. AUC según distintos valores de λ



Se selecciona el mejor λ obtenido, es decir el λ que minimiza el CV-error.

Luego, se realizan las predicciones sobre el conjunto de *Test* con el mejor λ y evaluamos los resultados:

Tabla 19. Performance Regresión Logística con regularización CV

Métrica	Performance
Precision	94,623%
Recall	83,670%
F1 - score	88,880%
Tasa de Error	15,618%
Curva ROC	90,544%

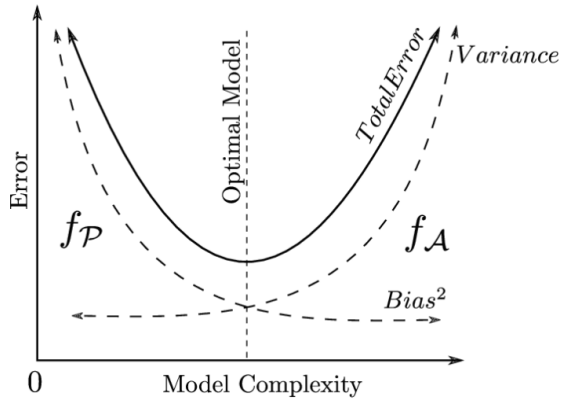
3.7. Modelos de Ensamble

Existen modelos de "agregación" (o ensamble) que logran reducir la variabilidad de los modelos antes expuestos.

El esquema general del aprendizaje ensamblado plantea que con el mismo *Train Set* se construyen varios modelos de árbol, se entrenan, para luego combinar las predicciones de cada uno.

Al entrenar muchos modelos y combinarlos se logra reducir la varianza y por ende reducir el error, es decir la capacidad predictiva aumenta. El desafío está en calibrar la complejidad del modelo que se van ensamblando, mediante la elección de los hiperparámetros.

Figura 19. Complejidad del modelo vs. error



Para poder implementar los distintos modelos que siguen a continuación se tuvo que utilizar la estrategia de *One Hot Encoding* (OHE) ya que solo admiten entrenar modelos a partir de matrices.

A continuación, se presentan los diferentes modelos implementados, junto con los principales fundamentos teóricos que James, G., Witten, D., Hastie, T. y Tibshirani, R. (2013) plantean en su libro "An introduction to statistical learning" (Vol. 6, Capítulo 8).

3.7.1. Modelo *Bagging*

Una vez realizado el OHE del dataframe, separamos la misma en bases de *Train* y *Test*.

Se entrenó el modelo *Bagging* a través de la librería *RandomForest*. Se busca formar B datasets de *train*, remuestreando sobre las filas del dataset – Remuestreo *Bootstrap* – y para cada B se estima un modelo de árbol.

Una de las desventajas del ensamble *Bagging* es que una re-muestra *bootstrap* NO es equivalente a obtener información nueva de la población (no vale el supuesto de independencia) ya que existen observaciones duplicadas dentro y entre las re-muestras. Por este motivo, los modelos que fiteamos estarán fuertemente correlacionados.

En este caso, cuando queremos hacer predicciones, se realiza el promedio de las predicciones de los distintos modelos:

$$\boldsymbol{Ensamble} = \frac{\mathbf{1}}{\boldsymbol{B}} \sum_{\boldsymbol{b=1}}^{\boldsymbol{B}} \hat{\boldsymbol{f}}^{\boldsymbol{b}}_{\boldsymbol{boot}}(\boldsymbol{X}) \qquad (11)$$

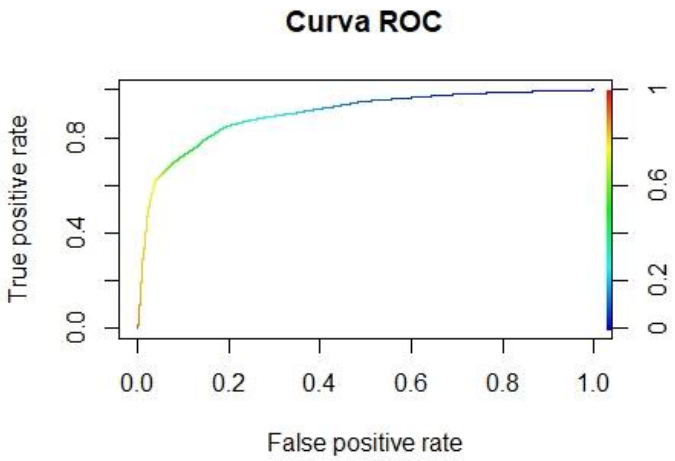
La cantidad de árboles B no produce *overfitting*, se debe elegir un B >> 0 hasta que se aplane el error empírico estimado.

Sin embargo, cuando miramos la complejidad de los árboles, la cantidad de nodos, si puede producir *over o underfitting*, debemos buscar la cantidad de nodos óptima para no sobreajustar el modelo.

Para este primer modelo de ensamble, se eligieron parámetros de manera arbitraria, utilizando B = 500 y 10 nodos.

Observamos la performance en *Test* del modelo a través de la Curva ROC. Alcanzando un área bajo la curva de 80,79%.

Figura 20. Curva ROC – Modelo *Bagging*



3.7.2. *Random Forest* – hiperparámetros arbitrarios

Para salvar la correlación entre las re-muestras del modelo *Bagging* se implementa un modelo *Random Forest*, que plantea un re-muestreo doble: por un lado, las observaciones y por otro lado los *features*.

Se aleatorizan a la hora de correr los modelos para que la varianza sea lo más chica posible.

Al igual que en *Bagging* se recomienda utilizar B >> 0, sin embargo, en este caso también se debe tener cuidado con los hiperparámetros que se eligen.

Corremos riesgo de *underfitting*, en contextos de muchos *features* irrelevantes al elegir m (cantidad de *features* a elegir por modelo) pequeño y corremos riesgo de *overfitting* cuando los árboles pueden crecer sin restricciones.

Antes de llevar a cabo el modelo final de esta tesis, probamos como nos va con un modelo *Random Forest* utilizando hiperparametros arbitrarios.

Los hiperparámetros elegidos son:

- *mtry* = 50 – número de variables candidatas para corte,
- *ntree*=1000 – número de árboles bootstap,
- *sample* = 0.5*floor(*Train*) – tamaño de cada re-muestra,
- *maxnodes* = 70 – número máximo de nodos terminales en cada árbol,
- *nodesize* = 500 – cantidad mínima de observaciones nodo terminal,

Por último, evaluamos su performance en *Test*:

Tabla 20. Performance Modelo *Random Forest* con hiperparámetros arbitrarios

Métrica	**Performance**
Tasa de Error	15,830%
Curva ROC	89,440%

3.7.3. *Random Forest* – hiperparámetros OOB

Como modelo final elegimos un correr un algoritmo *Random Forest* pero seleccionando los parámetros a través de OOB *(out-of-bag)*, es decir realizando una sintonía fina de los hiperparametros del modelo, quedándonos con aquellos que mejor performance tengan.

Al mismo tiempo que re-muestreamos y promediamos modelos, se estima indirectamente el error fuera de la muestra de *Train*.

Cuando se realizan las muestras *bootstap*, los tamaños de las mismas son iguales, pero como son con reposición puede que no incluyan todas las observaciones del *Train Set*, estas son las que se encontraran en OOB.

Con estas observaciones que no forman parte se computan errores de predicción que luego se promedia convenientemente.

Para llevarlo a cabo a los fines prácticos, se crea una grilla de hiperparametros, es decir para cada uno de ellos, se crea una fila con una secuencia de números y luego se entrena el modelo con cada una de las combinaciones posibles de hiperparametros elegidos.

Los valores utilizados fueron:

Tabla 21. Grilla de hiperparámetros

Hiperparámetros	**Valores**
M	45,50,55,60,65,85,100
maxnode	45,55,60,70,100,110
nodesize	500,600,700,800,900,1000,3000,5000

Se obtienen así los mejores hiperparámetros para luego re-entrenar el modelo con los seleccionados, ellos son:

Tabla 22. Mejores hiperparámetros

Hiperparámetros	**Valores**
M	65
maxnode	110
nodesize	1000

Utilizamos esto para predecir en *Test* y obtenemos los siguientes resultados:

Tabla 23. Performance Modelo *Random Forest* con OOB.

Métrica	**Performance**
Tasa de Error	12,440%
Curva ROC	90,930%

4. Resultados

4.1. Comparación de resultados entre modelos

A continuación, se expone un cuadro comparativo con las distintas performances de los modelos, como se explicó en la **sección 3.1. Medidas de performance en problemas de clasificación**, ante este tipo de situaciones corresponde observar y destinar esfuerzos en métricas que no dependan del umbral, como es **Curva ROC**, por este motivo, vamos a realizar las comparaciones en base a esta métrica de performance.

Tabla 24. Performances integral

Modelo Benchmark - Regresión Logística sin regularización.	
AUC máxima	89,78%
Modelo Benchmark - Regresión Logística sin regularización.	
Tratamiento a valores nulos y *outliers*	✓
AUC máxima	89,78%
Regresión Logística con regularización Lasso - Hiperparámetros arbitrarios.	
Tratamiento a valores nulos y *outliers*	✓
Ingeniería de atributos	✓
AUC máxima	89,38%
Regresión Logística con regularización Lasso - K-fold cross-validation	
Tratamiento a valores nulos y *outliers*	✓
Ingeniería de atributos	✓
K-fold cross-validation para elección de hiperparámetros	✓
AUC máxima	90,54%

Modelo Bagging - Hiperparámetros arbitrarios	
Tratamiento a valores nulos y *outliers*	✓
Ingeniería de atributos	✓
AUC máxima	80,79%
Random Forest - Hiperparámetros arbitrarios.	
Tratamiento a valores nulos y *outliers*	✓
Ingeniería de atributos	✓
AUC máxima	89,44%
Random Forest - Hiperparámetros OOB.	
Tratamiento a valores nulos y *outliers*	✓
Ingeniería de atributos	✓
Selección de mejores hiperparámetros a través de OOB	✓
AUC máxima	90,93%

Como se puede observar la mejor performance la obtiene el modelo de *Random Forest* con OOB para la elección de hiperparámetros, sin embargo se puede ver como la performance de Regresión Logística con regularización *Lasso* , eligiendo el hiperparámetro λ a través de ***k-fold cross-validation*****,** es muy similar.

Los modelos que fueron entrenados con hiperparámetros con criterio de elección siempre son la mejor opción para este tipo de problemas, por este motivo la importancia de realizar una correcta selección.

4.2. Importancia de atributos

Un punto importante a la hora de correr un modelo predictivo, es que luego el mismo pueda ser interpretable, es decir, tratar de entender que variables fueron las más importante para determinar si un cliente, en este caso, abandona o no el banco.

A los fines prácticos, es necesario conocer estas variables, saber cuáles de ellas son gestionables o bien entender el comportamiento de los clientes que están por hacer *churn*.

Para el caso de Regresión Logística con regularización *Lasso*, quedándonos con el modelo donde λ, fue elegido con *k-fold cross-validation*, mediante la función *summary* se obtuvieron todas aquellas variables activas, es decir, las que están incluidas en el modelo y cuyo valor estimado de beta es distinto de cero. Las demás variables no se muestran porque su beta estimada es cero. Luego, se tomaron aquellas variables cuyo valor absoluto de beta es más alto, lo que indica su mayor importancia.

Las 20 variables más importantes para este modelo son:

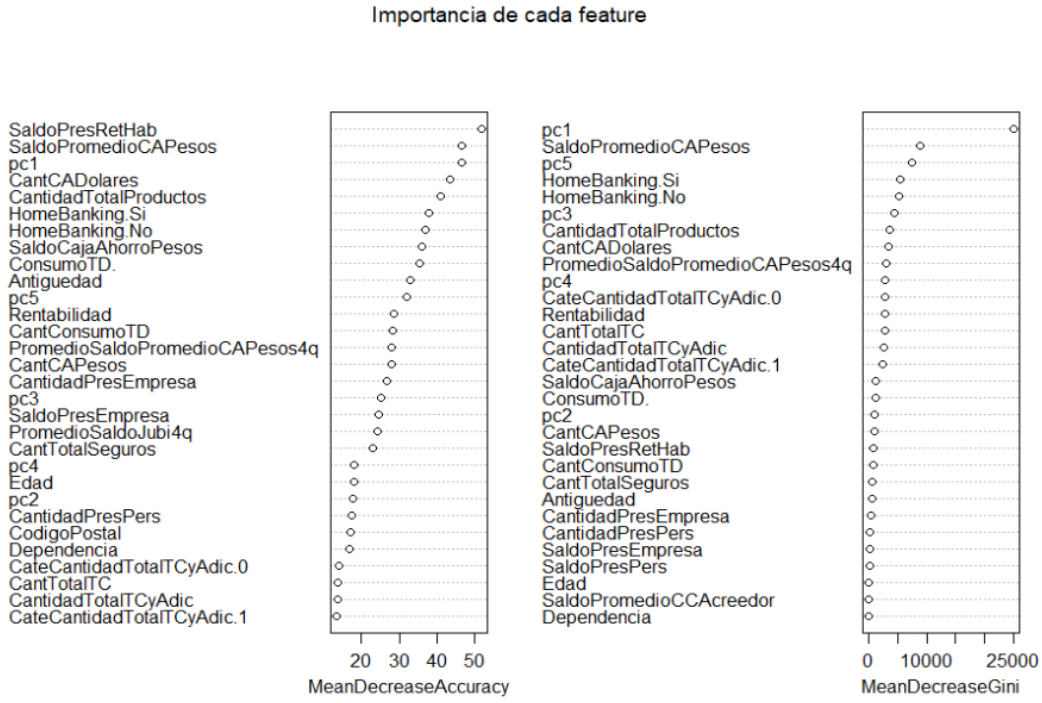
Tabla 25. Importancia de atributos RL

Variable	Asignación	beta
"CantidadPresHipotec"	1	-3,319694247
"CantidadVADEEBS"	1	-3,160608017
"CateCantidadTotalTCyAdic1"	1	-2,730295808
"CantPresHipotecUVAs"	1	-2,678933849
"CantCajadeSeguridad"	1	-2,396193186
"CateCantidadPFUSD4q1"	1	-2,100366566
"CantidadPresPrend"	1	-1,902076243
"HomeBankingSi"	1	-1,713209578
"CantidadPlazoFijoPesos"	1	-1,609246943
"CantidadPresPers"	1	-1,519587196
"CateCantidadPFPesos4q1"	1	-1,425835602
"CantidadPresRetHab"	1	-1,357755969
"CantTotalSeguros"	1	-1,238516419
"CatePromedioSaldoJubi4q1"	1	1,041089311
"CantidadPresPersUVAs"	1	-0,949549799
"CateCantidadTotalProductos1"	1	0,613982769
"cluster3"	1	0,561845961
"CantCADolares"	1	0,556600872
"CantPFCanalesElectronicos"	1	0,550082945
"CatePromedioSaldoHaberes4q1"	1	-0,546920504

Para el caso de los modelos de ensamble, es más complejo saber cuál es la importancia de las variables ya que los árboles son “cajas negras”, se necesita trasparentar los modelos calculando la importancia del *feature* j-ésimo como el promedio de la importancia de este *feature* a lo largo del ensamble. Esto se utilizó para el modelo *Random Forest* con OOB.

Al correr la función en *RStudio VarImpPlot* obtenemos la siguiente salida:

Figura 21. Importancia de atributos RF



Como se puede observar, las variables más importantes de ambos modelos no coinciden en su totalidad, esto se debe principalmente a que son modelos (como explicamos en las secciones anteriores) que utilizan métodos distintos para el cálculo de *probabilidad de churn*. Los modelos *Random Forest* capturan relaciones no lineales de las variables, mientras que los modelos de regresión con regularización *Lasso* capturan relaciones lineales.

Sin embargo, se observa como en ambos casos muchas de las variables que se fueron agregando a lo largo del trabajo, tanto aquellas que se incluyeron en el proceso de Ingeniería de Atributos, como por ejemplo, las variables categóricas de cantidad y saldos de productos, las PCA agregadas, cantidad total de productos, rentabilidad, etc. como también aquellas variables que fueron modificadas antes de comenzar a modelar, como son las variables que promedian los saldos o cantidades del cuarto trimestre o la variable mejorada de homebanking, entre otras, son tomadas por el modelo para predecir y son las que mejor lo hacen. Con estos resultados volvemos a constatar la relevancia del proceso de procesamiento de datos para la implementación de modelos predictivos.

4.3. Performance dataset homogéneo vs. Única dataset heterogéneo.

Realizado todo el trabajo antes expuesto, y observando la heterogeneidad de la base con la que se cuenta para el cálculo de predicción, resulta interesante observar los resultados antes obtenidos, pero dividiendo la base en dos, es decir entrenar dos modelos, tal como se hizo hasta ahora, pero con dos datasets distintos.

Un buen criterio para hacerlo es observando la cantidad de productos que tiene cada cliente, es decir, quedándonos con una base para todos aquellos clientes que tienen tres o más productos con el banco y por separado con todos aquellos que tienen menos de tres productos.

Obtenemos dos bases con las siguientes dimensiones:

Tabla 26. Tabla de proporciones (+2)

Con 3 o más productos	
Churn	***No Churn***
14,91%	85,09%

Tabla 27. Tabla de proporciones (-3)

Clientes con 2 o menos productos	
Churn	***No Churn***
44,86%	55,14%

Como se puede ver, para el caso de "Con 3 o más productos" las clases quedaron desbalanceadas, es decir, de los datos de la Base de Clientes con dicho filtro, el 14,91% hizo *churn,* mientras que el 85,09% no hizo. Esto afecta a las predicciones del modelo si no se subsana antes de entrenarlo. Para que el modelo logre capturar aquellos fenómenos que sean extraños u ocurran poco se llevó a cabo el proceso de ***Oversampling*** que se basa en re-muestrear sobre la clase minoritaria. Se elige aleatoriamente filas de la clase minoritaria y las duplica.

Una vez hecho esto entrenamos modelos de Regresión Logística con regularización y *Random Forest* con OOB tal como se realizó para el total de la base.

También para cada base se entrenó nuevamente el algoritmo *k-means* ya que la distribución de los *clusters* debía ser recalculada por el cambio de observaciones.

Lo mismo se realizó con PCA, se calculó nuevamente el método para ambas bases.

Los resultados que se obtuvieron fueron los siguientes:

Tabla 28. Comparación de performance

	Clientes con 3 o más productos	***Clientes con 2 o menos productos***	***Todos los clientes***
Regresión Logística con regularización Lasso - K-fold cross-validation			
Tratamiento a valores nulos y *outliers*	✓	✓	✓
Balanceo de clases	✓		
Ingeniería de atributos	✓	✓	✓
K-fold cross-validation para elección de hiperparámetros	✓	✓	✓
AUC máxima	85,04%	91,12%	90,54%
Random Forest - hiperparámetros OOB.			
Tratamiento a valores nulos y *outliers*	✓	✓	✓
Balanceo de clases	✓		
Ingeniería de atributos	✓	✓	✓
Selección de mejores hiperparámetros a través de OOB	✓	✓	✓
AUC máxima	87,36%	92,80%	90,93%

Como se puede ver si nuestro objetivo sería analizar a los clientes con pocos productos, tomarlos por separado y analizar sus predicciones es la mejor opción ya que la performance en ambos modelos mejora del total, sin embargo, cuando observamos las métricas de aquellos clientes que tienen más productos, la performance disminuye respecto al general.

Para este caso, se analizarán las predicciones que obtuvimos de entrenar el modelo con el total de clientes.

5. Discusión final y Análisis de aplicación

Como conclusión de los distintos análisis y desarrollos que se realizaron a lo largo de la tesis, se consideró llevar adelante un análisis descriptivo de los resultados obtenidos, y también, destinar un lugar al planteo de casos de uso en donde utilizar la probabilidad de abandono de los clientes es sumamente importante y permite a las distintas gerencias del banco tomar decisiones y destinar sus esfuerzos de la mejor manera posible.

Evaluados cada uno de los modelos y observando sus diferentes performances, se determinó que el que mejor predice el abandono de los clientes del banco es *Random Forest* con elección de hiperparámetros a través de OOB y esta es la que se usará en lo que sigue.

Para realizar el análisis correspondiente se usará la probabilidad de abandono de cada uno de los clientes. Se agruparon los clientes según su probabilidad de abandono y se observaron solo algunas variables relevantes para cada grupo, obteniéndose los siguientes resultados:

Tabla 29. Predicción *churn* por producto

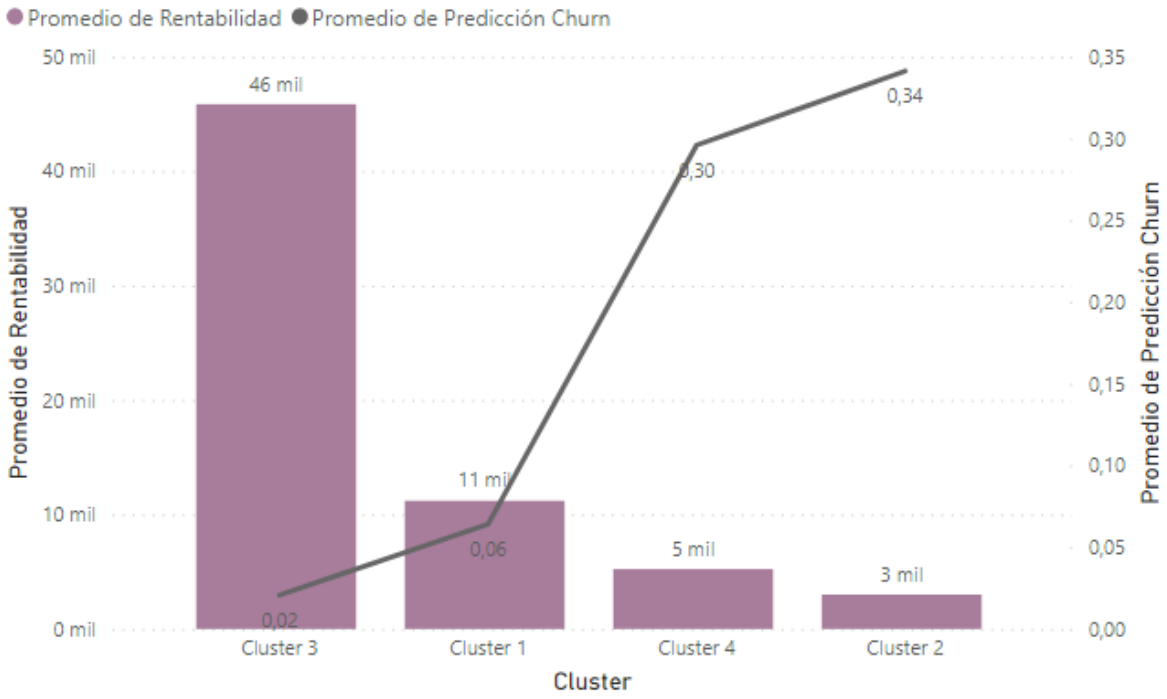
Predicción Churn Promedio	***Saldo Promedio CA Pesos 4q***	***Cantidad PF Pesos 4q***	***Cantidad PF USD 4q***	***Cantidad Total TC***
[0,0, 0,3]	27821,45	87520	16924	196811
[0,3, 0,6]	9088,42	292	99	8551
[0,6, 0,9]	8671,71	25	9	42
[0,9, 1,0]	254,73	0	0	0

Tanto en cantidad de Plazo Fijo, como en Tarjetas de Crédito aquellos con alta probabilidad de irse no tienen productos, tal como era de esperarse. Sin embargo, se puede deducir que existe una posibilidad de colocación de productos en aquellos que tienen una probabilidad baja de *churn* (segundo rango 0,3 a 0,6) y tienen una baja cantidad de productos respecto a los que están en el primer rango (0,0 a 0,3).

En lo referido a los saldos, se visualizan resultados parecidos con respecto a las cantidades, aunque el saldo promedio que dejan en las cajas de ahorro pesos, durante un trimestre, aquellos que están en el segundo y tercer rango, son similares. Es decir, que no implementar un plan de retención de clientes sobre aquellos que tienen más de un 60% de probabilidad de abandonar (probabilidad entre 0,6 y 0,9) evitando así el posible *churn* de este grupo, podría provocar una baja en los saldos promedios de las cajas de ahorro (saldos diarios con los que cuenta el banco para posibles inversiones) significativa.

Entender la relación de la rentabilidad de cada cliente, y la predicción de *churn* resulta útil para dimensionar las posibles pérdidas:

Figura 22. Rentabilidad vs. Promedio de Predicción de *Churn* por *Clusters*



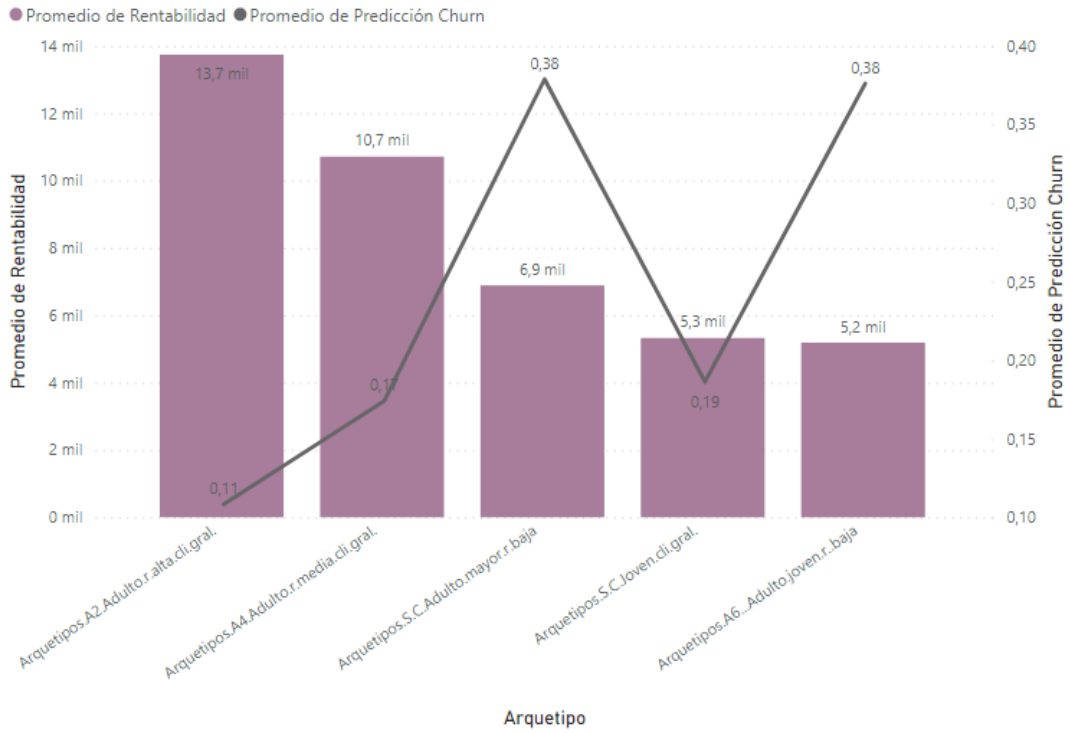
En este caso, además lo dividimos según los *clusters* que anteriormente agregamos gracias al algoritmo *K-means* en el **punto 3.4.2**.

Aquellos *clusters* con mayor rentabilidad tienen menos probabilidad de abandono, respecto a los que tienen menor rentabilidad.

Esto nos indica que, si se realizan esfuerzos sobre aquellos clientes con pocos productos en la cartera del banco, conseguiríamos retenerlos y por consecuencia aumentar la rentabilidad total.

Como se comentó anteriormente, dentro de la base de clientes existe una segmentación realizada de manera no automatizada utilizando ciertos criterios (edad, ingresos y tipo de cliente) llamada Arquetipo, observamos si su relación con la Rentabilidad y las predicciones son similares a los *clusters* obtenidos de manera automatizada.

Figura 23. Rentabilidad vs. Promedio de Predicción de *Churn* por Arquetipo



Aquellos clientes de altos ingresos (r.alta) son los que mayor rentabilidad tienen y los de ingresos bajos (r.baja) los que menos tienen. También coincide la idea de que aquellos con mayor rentabilidad son justamente los que menor probabilidad de abandonar corren.

Hoy en día, la segmentación llamada “Arquetipos”, aún no es utilizada por las gerencias del banco para llevar adelante gestiones de campañas, únicamente para análisis descriptivos, sin embargo, desde mi punto de vista, se pueden encontrar soluciones más eficaces si así se hiciera.

Por ejemplo, de la **Figura 23**, podemos deducir que, si bien es correcto mantener satisfechos a los clientes con altos ingresos y rentabilidad, son los mejores clientes que tiene el banco probablemente, es muy poco probable que estos cambien de banco, por ende, destinar esfuerzos para seguir reteniéndolos en un principio no valdría la pena.

Uno de los casos de uso de la probabilidad de abandono por cliente, es desarrollar un programa de retención de clientes de corto plazo, por ejemplo, lanzando campañas de promociones, donde además de enfocarnos en cada cliente podríamos observar que impacto monetario tendría en cada uno de los Arquetipos.

A continuación, se expondrá un ejemplo práctico de cómo llevar adelante este tipo de casos y observar cómo puede aportar lo desarrollado en este documento hasta ahora.

Lo que se busca es reducir la tasa de abandono en un 2% para el próximo trimestre. Como vimos en la **sección 2. Datos**, el banco en cuestión cuenta con 435938 clientes, cuya tasa de abandono es de 17,32% anual. Para cumplir con el objetivo planea ofrecer una reducción de 0,5% como tasa preferencial para préstamos de hasta 50.000 pesos para el 50% de los clientes con alta tasa de abandono (entre 0.9 y 1). Se espera que la promoción retenga al 40% de los clientes tomados.

Los datos necesarios para el cálculo son:

- **Costo de adquisición de nuevos clientes:** servirá además para comparar si este supera o no el costo de la campaña para cada cliente. Para calcularlo, se investigó sobre el presupuesto y los objetivos del banco. Según el presupuesto, el costo total destinado a la adquisición de nuevos clientes durante el 2022 es de $622.080.641,00. La cantidad de clientes en diciembre 2021 fue de 409938 y el plan estratégico determinó que la variación en cantidad de clientes dentro de clientela general sea de 26000 clientes. Teniendo en cuanta la tasa de abandono anual, la cantidad de altas que se deberían lograr para el cumplimiento es de 62.894. Entonces el costo de adquirir un nuevo cliente es de 9.890,87 (622.080.6421,00/62894).
- **Costo total de la promoción:** Si contamos con préstamos de hasta $50.000 a una tasa anual de 100% y ofrecemos una reducción de 0,5%, en vez de cobrar $50.000 anuales cobraríamos $47500, obteniendo una pérdida de $2500. 102.181 son los clientes con mayor tasa de abandono, si nos quedamos con el 50%, decimos que los clientes susceptibles a la misma son 51409, por ende, el costo total de la promoción será de $128.522.500,00 ($2500*51409).
- **Costo por cliente retenido:** Este será de $6.250 - $128.522.500/20563,6 ya que el 40% es esperable retener

Si comparamos este costo con el costo de adquirir un nuevo cliente ($9.890,87), el banco puede concluir que la promoción es rentable, ya que el costo por cliente retenido es significativamente menor que el costo de adquirir un nuevo cliente.

Como se aclaró más arriba, la clasificación de los clientes por Arquetipos no es utilizada más que para análisis descriptivos por ese motivo este caso no podemos estudiarlo por Arquetipo. Sin embargo, una parte importante de este análisis de costo – beneficio es conocer el ROI o retorno de inversión de la promoción. Al contar con la rentabilidad promedio de cada Arquetipo podemos calcular el ROI que obtendríamos si aplicamos la promoción a cada uno de ellos y entender, en qué casos resulta más rentable llevarlo a cabo.

Tabla 30. Cálculo del ROI por Arquetipo

Arquetipo	Clientes susceptibles de entrar en la promoción	Rentabilidad promedio por cada cliente retenido	Clientes que se esperan retener (40% de los susceptibles)	Costo de la promoción por cliente	ROI
Adulto r. alta	428	$ 14.895,93	171	$ 2.500	238,3%
Adulto r. media	8.058	$ 12.467,12	3.223	$ 2.500	199,5%
Adulto joven r. baja	32.074	$ 7.434,79	12.829	$ 2.500	119,0%
Adulto mayor r. baja	10.523	$ 10.191,77	4.209	$ 2.500	163,1%
Joven SC	328	$ 6.076,19	131	$ 2.500	97,2%

Más allá que aplicar el programa de retención a los clientes con altos ingresos da un retorno de la inversión alto, la cantidad de clientes que se esperan recuperar es baja con respecto a otros arquetipos. Esto implica que haya poca variedad de clientes, dejando aquellos con menores recursos fuera y corriendo riesgos de abandono en el largo plazo ante posibles cambios en sus características.

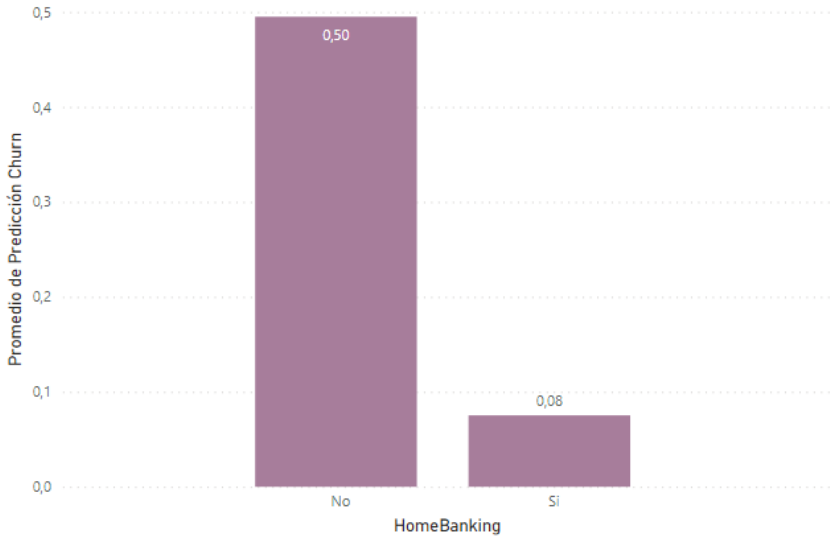
Resulta interesante observar, aquellos clientes adultos con ingresos medios y bajos cuyo ROI es 199,5% y 163,1% respectivamente. Realizar este tipo de acciones podría traer grandes resultados y evitar tener la necesidad de adquirir nuevos clientes.

Por último, observamos la relación entre las predicciones de abandono con el uso de homebanking y también con las notas de NPS de cada cliente. Ambas variables fueron analizadas antes de correr los modelos y como era de esperar, se observaron relaciones similares.

Que un cliente no use o mejor dicho que deje de usar el homebanking es una alerta de que su probabilidad de fuga es alta. Probablemente uno de los motivos principales para dejar de operar a través del homebanking es la intención que el cliente tiene de abandonar el banco.

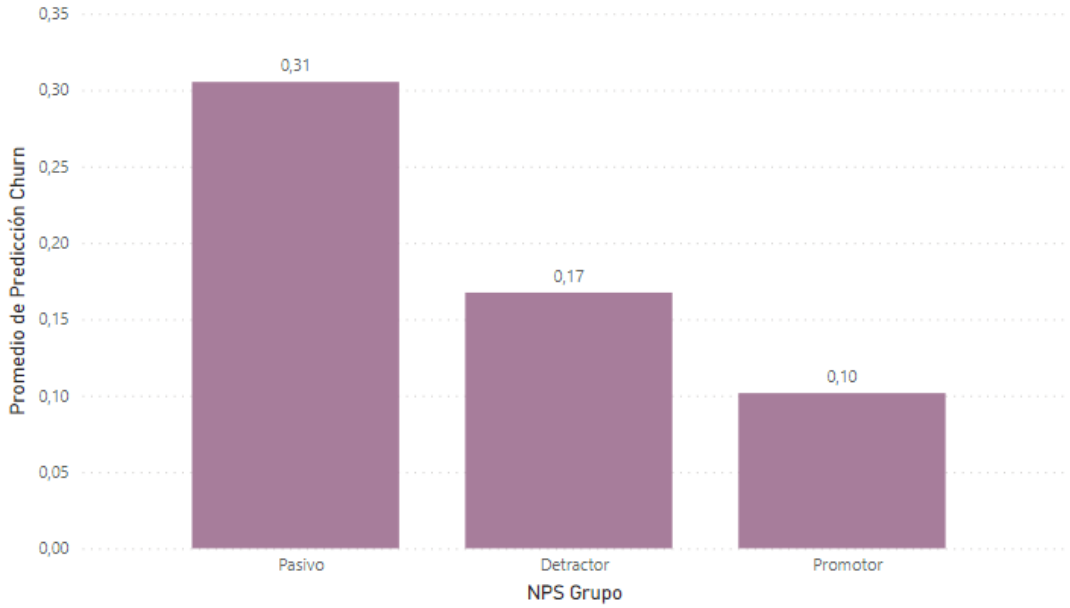
Observar aquellos clientes que no hayan utilizado el homebanking en los últimos 30 días y comprender su situación, entendiendo sus últimos movimientos, si intento comunicarse por algún medio, si cuenta con productos o bien si existen productos que se adapten a sus necesidades o preferencias, es clave para luego aplicar un programa de retención como el planteado anteriormente, buscando su reactivación antes de que tomen la decisión final de abandonar.

Figura 24. Promedio de Predicción por uso de Homebanking



Aquellos clientes Pasivos (que califican con 7 u 8) son los que mayor probabilidad de abandono tienen, los siguen los detractores y luego los promotores. Tal como se describió, la variable NPS no explica correctamente la tasa de abandono, debemos entender los factores que si influyen el abandono para luego buscar un aumento en el NPS.

Figura 25. Promedio de Predicción de *Churn* por NPS Grupo



Finalmente podemos afirmar que la probabilidad de abandono brinda información valiosa para anticiparse a posibles pérdidas de clientes y tomar medidas proactivas. Al segmentar a los clientes según su probabilidad de abandono, se pueden diseñar estrategias personalizadas de retención y desarrollar campañas específicas dirigidas a los segmentos de mayor riesgo de fuga.

Esto no solo contribuye al área del marketing, sino también a la gestión de productos y servicios, donde se pueden enfocar los esfuerzos en mejorar la oferta para retener a los clientes más propensos a abandonar. En resumen, los análisis descriptivos de resultados y los casos de uso han demostrado que la probabilidad de abandono es una herramienta estratégica para las gerencias del banco, brindando información valiosa que les permite tomar decisiones informadas y asignar recursos de manera eficiente en la lucha contra la fuga de clientes.

6. Referencias

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. (2013). An introduction to statistical learning. Vol. 6. New York: springer.

Cohen, D. A., Gan, C., Hwa, A., & Chong, E. Y. (2006). Customer satisfaction: a study of bank customer retention in New Zealand.

Velu, A. (2021). Customer *Churn* Management Using Predictive Modeling–A Machine Learning Approach. Journal of Emerging Technologies and Innovative Research, 8(4).

Bilal Zorić, A. (2016). Predicting customer *churn* in banking industry using neural networks. Interdisciplinary Description of Complex Systems: INDECS, 14(2), 116-124

Hennig-Thurau, T., & Klee, A. (1997). The impact of customer satisfaction and relationship quality on customer retention: A critical reassessment and model development. Psychology & marketing, 14(8), 737-764.

JM Valdez Mendia & JJA Flores-Cuautle (2022) Hacia la experiencia de hiperpersonalización del cliente: un enfoque basado en datos, Cogent Business & Management, 9:1, 2041384, DOI:10.1080/23311975.2022.2041384.

Rahman, F., & Devanbu, P. (2013, May). How, and why, process metrics are better. In 2013 35th International Conference on Software Engineering (ICSE) (pp. 432-441). IEEE.

Zaki, M., Kandeil, D., Neely, A., & McColl-Kennedy, J. R. (2016). The fallacy of the net promoter score: Customer loyalty predictive model. Cambridge Service Alliance, 10, 1-25

Alice Zheng and Amanda Casari (2018). Feature Engineering for Machine Learning. Principles and Techniques for Data Scientists. O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472

Lindsay I Smith (2002). A tutorial on Principal Components Analysis

Abdi, Hervé, and Lynne J. Williams. (2010). “Principal Component Analysis.” John Wiley and Sons, Inc. WIREs Comp Stat 2: 433–59.

Kevin Dunn (2023). *Process improvement using data.*

Tan, P. N., Steinbach, M., & Kumar, V. (2005). Introduction to Data Mining. Pearson.

Alpaydin, E. (2014). Introduction to Machine Learning (sección 5.7). MIT Press.