**ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL**

**FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICAS**

**PROYECTO DE TITULACIÓN**

PREVIO A LA OBTENCIÓN DEL TÍTULO DE:

**“MAGÍSTER EN ESTADÍSTICA APLICADA”**

**TEMA:**

# Predicción del abandono de tarjetahabiente aplicado en una institución financiera ecuatoriana

**AUTOR:**

YELTSIN ALEXANDER CASTRO LOAIZA

Guayaquil - Ecuador

2022

# Resumen

En el presente trabajo de titulación, se aplica métodos estadísticos para construir un modelo de predicción de la propensión de no uso (churn) de tarjeta de crédito. Haciendo uso de la información financiera de un banco ecuatoriano, se explotan las siguientes dimensiones de información disponible: información de la tarjeta de crédito, información de la central de riesgo, información sociodemográfica de los clientes y el registro de su actividad transaccional. Para la selección de las principales características se considerada no sólo la correlación que presentan para explicar el evento sino también el sentido económico. Los algoritmos utilizados en el presente estudio incluyen la regresión logística y el árbol de decisiones dado que han demostrado ser herramientas de clasificación maduras y estables, en comparación con el resto de los métodos.

Los resultados indican que el bosque aleatorio presenta un mejor ajuste entregando al negocio la capacidad de elaborar estrategias personalizadas para retener a los clientes.

Palabras clases: Fuga de clientes, retención, banco, Machine Learning.

# Abstract

In the present work, statistical methods are applied to build a predictive model of the propensity of credit card churn. Using the financial information of an Ecuadorian bank, the following dimensions of available information are exploited: credit card information, credit bureau information, sociodemographic information of customers and the record of their transactional activity. For the selection of the main characteristics, not only the correlation they present to explain the event but also the economic sense was considered. The algorithms used in the present study include logistic regression and decision tree since they have proven to be mature and stable classification tools, compared to the rest of the methods.

The results indicate that the random forest presents a better fit giving the business the ability to develop customized strategies to retain customers.

Keyword: Churn, retention, bank, Machin Learning.

# DEDICATORIA

Para Nivia, Angel, Alejandro, Wilson y Ariana.

# AGRADECIMIENTO

Agradezco a mi familia que ha sabido entender el tiempo que dedico a cumplir mis sueños y a las personas que han depositado su confianza en mis ideas. Mención especial a los profesores que aportaron con su experiencia y a la Ph.D Andrea Garcia Angulo por su gran aporte al presente trabajo.

# DECLARACIÓN EXPRESA

La responsabilidad por los hechos y doctrinas expuestas en este Proyecto de Titulación me corresponde exclusivamente y ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría. El patrimonio intelectual del mismo corresponde exclusivamente a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Titulación referido.


Yeltsin Alexander Castro Loaiza

# TRIBUNAL DE GRADUACIÓN

Firmado electrónicamente por:
**OMAR HONORIO RUIZ BARZOLA**

Ph. D. Omar Honorio Ruiz Barzola
PRESIDENTE

Firmado electrónicamente por:
ANDREA CRISTINA GARCIA ANGULO

Ph. D. Andrea Cristina Garcia Angulo
TUTOR

Firmado electrónicamente por:
**FRANCISCO ANTONIO MOREIRA VILLEGAS**

M.Sc. Francisco Antonio Moreira Villegas
DOCENTE EVALUADOR

# ABREVIATURAS O SIGLAS

A continuación, se expresan las siglas con sus respectivas definiciones:

- RFM: Recencia, frecuencia y monetaria.
- CRM: Customer Relationship Management, o Gestión de las relaciones con clientes.

# TABLA DE CONTENIDO

# LISTADO DE FIGURAS

# LISTADO DE TABLAS

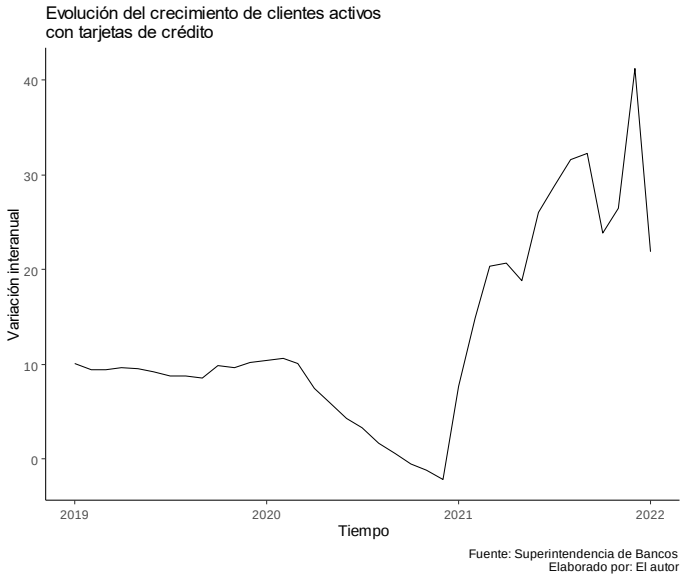
# CAPÍTULO 1

## 1. INTRODUCCIÓN

La realidad empresarial que afrontan todas las instituciones financieras está inmersa en un clima empresarial de alta competitividad donde los bancos deben retener clientes y a su vez intentan incrementar su participación de mercado, mediante la captura de nuevos clientes. Nie, Rowe, Zhang, Tian, & Shi (2011) desarrollan un análisis integral de cómo aplicando técnicas estadísticas se logra predecir el abandono del servicio de tarjeta de crédito. Entre sus principales hallazgos se aprecia que cuando se incrementa la tasa de retención de clientes en un 5% aumenta las ganancias de un banco hasta en un 85%. Además, Verbeke, Marters, Mous, & Baesens (2011) encuentran que captar nuevos clientes cuesta más en cualquiera empresa que retener los que ya posee, donde probablemente, ésta acción genere más rentabilidad.

El número de tarjeta habientes ha aumentado de manera considerable en Ecuador en los últimos años. Tomando información publicada por la Superintendencia de Bancos del Ecuador[1], se aprecia que desde enero del 2018 hasta diciembre del 2021 hubo un incremento de más de 1.6 millones de tarjetas de crédito activas. En la gráfica 1.1 se muestra el crecimiento interanual registrado entre el 2019 y el 2021, en el 2019 el crecimiento promedio fue de 9.43%, lo que refleja un comportamiento diferente a lo apreciado en el 2020 (año de mayor impacto de la pandemia de COVID-2019 en la economía ecuatoriana) donde hasta se alcanzan crecientes negativos, la variación promedio reflejada en el 2020 fue de 4.18%, mientras que en el 2021 se aprecia un crecimiento considerable con respecto al 2020, cerrando diciembre del 2021 con un incremento del 41% (1,398,432) con respecto al 2020.

[1] En el siguiente link se encuentra disponible la información publicada por la Superintendencia de Banco: https://estadisticas.superbancos.gob.ec/portalestadistico/portalestudios/




***Gráfica 1.1:*** *Evolución del crecimiento interanual de la cantidad de tarjetas activas en el Ecuador.*

Dado el incremento registrado en colocación de tarjetas de crédito y el comportamiento del mercado lleva a creer que la competencia entre varios bancos es de competencia por acaparar nuevos clientes.
Adicionalmente, se esperaría que los tarjeta habientes transformen su lealtad de un banco a otro por múltiples atribuciones, como la disponibilidad tecnológica en sus plataformas, personal del banco amable, bajas tasas de interés, beneficios con plazos de gracia, proximidad de la ubicación geográfica, etc. Por lo tanto, existe la necesidad de desarrollar soluciones analíticas que permitan predecir qué clientes son propensos a abandonar el servicio en función de los datos demográficos, transaccionales, comportamiento de pago de obligaciones en el sistema financiero e información del entorno macroeconómico.

El fenómeno de la rotación de clientes no es sólo un problema de las instituciones financieras, es muy frecuente en otras actividades económicas como la industria de servicios, telecomunicaciones, entretenimiento digital, etc. Bolton (1998) detecta que las empresas deben ser proactivas en conocer de manera anticipada los clientes que dejen de percibir los niveles de satisfacción además que los sistemas de recolección de contacto son las principales fuentes de información para detectar el fin de la relación comercial.

Bolton & Bramlett (2000) plantea que teóricamente las organizaciones deben utilizar sus datos para segmentar a los clientes dado su comportamiento de compra para ofrecer una mejor experiencia en los servicios, en lugar de simplemente realizar campañas con sus características sociodemográficas. Luego con la investigación realizada por Burez & Van den Poel (2007) se comprueba que, utilizando técnicas estadísticas para predecir el abandono de clientes, la empresa de telecomunicaciones disminuye su nivel de abandono de manera considerable.

Considerando que la detección temprana debe estar incluida en la planificación estratégica de toda organización dado que el proceso actual de retención gira entorno al estado transaccional de la tarjeta habiente, se impulsa el presente proyecto para entregar insumos para efectuar acciones específicas de retención, lo que se traduce en ganancias. Para esto se utilizarán diferentes modelos clasificación para predecir

la probabilidad de que un cliente deje de utilizar los servicios otorgados por el producto de cartera.

### 1.1. Descripción del problema

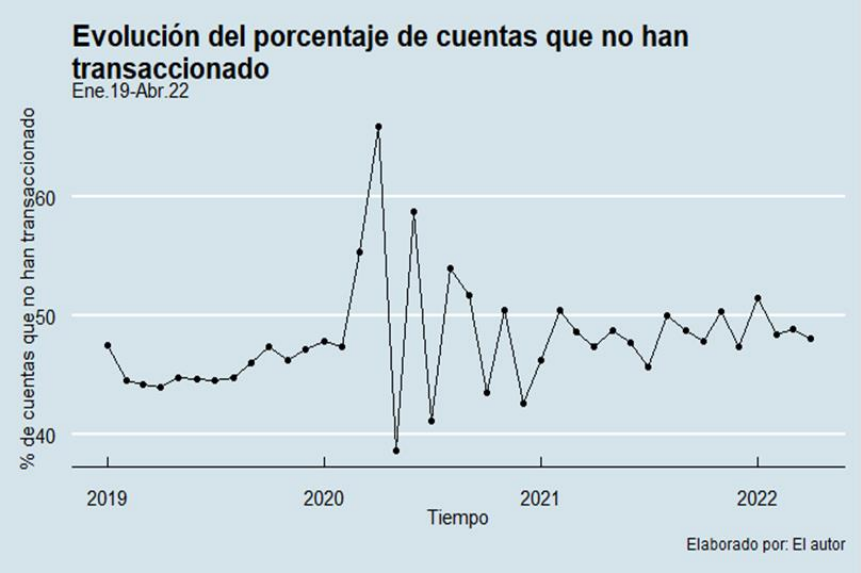
En la literatura, la fuga de clientes se conoce con el término de "churn", mismo que hace referencia a la cantidad de clientes que abandonan el producto o servicios brindado por una determinada empresa. Larivière & Van den Poel (2004) manifiesta que existen tres tipos de fuentes de fuga de clientes. La primera fuente de fuga se denomina abandono total y se da cuando el cliente anula todas las relaciones comerciales con la entidad; la segunda, se la conoce como parcial, se efectúa cuando cancela una determinada cantidad de productos y el sobrante lo realiza con la competencia; la última forma de cancelar se da cuando se inactiva por un largo periodo de tiempo la utilización de los servicios o productos con la entidad. Por lo tanto, el abandono de un cliente puede ser predicho con la información que reposa en los sistemas de la empresa.

Según Konukal (2018), para predecir el abandono de un cliente se lo puede realizar por medio de su comportamiento transaccional previo a que suceda el evento, mientras que, Lessmann, & Verstraeten (2017) afirman que la predicción de abandono o fuga de clientes se asigna con su comportamiento transaccional histórico. De igual manera, Ballings & Dirk (2012) descubren que la experiencia del cliente con la institución emisora de la tarjeta de crédito reduce la probabilidad de abandono.

Con los diferentes enfoques existentes en la literatura, se plantea la definición de clientes de tarjeta de crédito con estatus de abandono, para esto se hace uso del sistema de "Recency, Frequency and Monetary" (RFM) para tarjetahabientes. El RFM, permite conocer un alto nivel del comportamiento del cliente con la tarjeta, para este caso se dará mayor participación a la recencia y al valor que tiene el RFM. La recencia indica hace cuantos meses el cliente ha utilizado su producto, mientras que el RFM nos permite saber si es un cliente rentable o no para la institución financiera.

Actualmente, 30% de los tarjeta habientes de la institución estudiada registran su último consumo en más de 6 meses, mientras que por parte de la categoría del RFM el 37% de los clientes tiene un estatus de perdidos, es decir, son clientes que han transaccionado en más de 6 meses, poseen un bajo nivel de cantidad transaccionadas y los montos transaccionados se encuentran por debajo del percentil 25.

Se ha detectado que en promedio el 48% de las cuentas no han utilizado sus tarjetas de crédito a pesar de estar activos. En la gráfica 1.2 se muestra la evolución del porcentaje de cuentas entre el 2019 y 2022, donde se aprecia un crecimiento sostenible desde el 2019 hasta el 2020 pero en la mayoría de los meses del 2020, la estabilidad de la serie cambia de manera rotunda alcanzando su máximo valor de no uso en abril del 2020 (66%). Desde el 2021 hasta la actualidad se aprecia que la variabilidad de la serie ha venido disminuyendo, pero refleja tener valores superiores a los registrados prepandemia.




***Gráfica 1.2:*** Evolución del porcentaje de cuentas que no han transaccionado.

## 1.2. Objetivos

### 1.2.1. Objetivo General

Clasificar a los clientes según la probabilidad de abandono del servicio de tarjeta de crédito, basándose en la predicción de modelos estadísticos para la gestión y monitoreo de productos de colocación.

### 1.2.2. Objetivos Específicos

- Determinar la probabilidad que un cliente abandone la tarjeta de crédito dado su comportamiento transaccional histórico, variables sociodemográficas, nivel de ingresos, asociación con el banco y nivel de riesgo en el sistema financiero.

- Identificar las variables relevantes que permitan construir un modelo de clasificación óptimo, obtenido de esta manera resultados estables.

- Encontrar oportunidades de mejora en el proceso de retención de clientes para facilitar la acción de los departamentos involucrados.

## 1.3. Alcance

El presente trabajo de titulación se enfocará en clasificar a los clientes de la institución financiera ecuatoriana con alta propensión a abandonar las tarjetas de crédito dado el nivel de uso. Para esto se utilizarán datos de las transacciones realizadas, comportamiento de pago de obligaciones, variables sociodemográficas, experiencia con otros productos de colocación y nivel de ingresos, considerando información de los últimos 5 años para poder contemplar el escenario de la pandemia en la solución analítica.

# CAPÍTULO 2

## 2. MARCO TEÓRICO

En este capítulo se abordará la importancia que tiene el presente tema de investigación para la institución financiera, dado que refleja la conexión que existe entre la institución y el cliente. Por lo que resulta importante conocer el comportamiento transaccional, hábitos de compra, recurrencias de compras, necesidades, perfiles sociodemográficos y cómo utilizando técnicas estadísticas se obtiene las consideraciones antes detalladas aportando información para que las decisiones sean fundamentadas en evidencia estadística. Además de detallar soluciones aplicadas a tarjeta de crédito, se brindará información de investigaciones aplicadas a otras industrias y revisión de estudios aplicados en Ecuador.

Cuando se trata de clientes de tarjeta de crédito, se debe pensar en el ciclo de vida del cliente y buscar la forma que la relación sea duradera. Para que esto suceda debe existir un trabajo colaborativo entre la experiencia de los colaboradores y las soluciones tecnológicas que permitan trasladar el conocimiento adquirido de los clientes a soluciones analíticas basadas en la información disponible en la organización, logrando ser más eficientes y eficaces en la divulgación de campañas y beneficios que poseen. (Botelho & Frederico Damian, 2010)

El universo de las tarjetas de crédito se considera cambiante. Qi, et (2009) manifiestan que las instituciones emisoras deben actuar de manera proactiva ante las modificaciones del entorno. Entre los retos más relevantes que afrontan las instituciones se presenta la facilidad con la que el cliente tienden a irse a otra institución o entidad, dado que puede darse una compra de cartera (ofrecimiento de tasa de interés baja), extensión en plazos, cuotas de gracia, potenciales condonaciones, entre otras, haciendo que los clientes siempre estén expuestos a considerar mejores ofertas. Como resultado de esta práctica, las empresas han desarrollado sistemas para gestionar la fuga de clientes mediante el uso de técnicas estadísticas buscan los clientes más propensos en desertar.

Existen muchas definiciones de fuga de clientes, tanto así que en la literatura se conoce como modelos de churn, mismo que varía según su sector. Por ejemplo, empresas tecnológicas lo definen cuando un usuario no ha iniciado sesión en un tiempo en su página web en un periodo considerable o ha dejado de usar el producto o ha finalizado el contrato adquirido con la empresa. Por lo que, Gady, Baesens, & Croux (2009) definen que un cliente está propenso a terminar su relación con la empresa cuando el Valor de Vida del cliente también conocido como Customer Lifetime Value (por sus siglas en inglés, CLV) se reduce de manera drástica durante un periodo de tiempo prolongado. Dado que algunas empresas se enfocan en clientes que según su CLV generan altos márgenes de ingresos y generan estrategias comerciales enfocado a este segmento de clientes, en este caso resulta relevante el cálculo de la probabilidad de abandono de los clientes de alto valor. Coussement, Lessmann, & Verstraeten (2017) consideran que la predicción de abandono resume de manera cuantitativa los cambios en su comportamiento histórico y el parentesco a

perfiles que han abandonado los servicios, logrando diferencias los clientes fidelizados.

Surge la importancia de detectar las características que afectan al abandono y que a su vez permitan la construcción de modelos estadísticos para ayudar a las empresas a crear ofertas adecuadas. Estas ofertas deben ser canalizadas por medio de las diferentes plataformas tecnológicas que las empresas manejen para gestionar los clientes, entre ellas se encuentra el conocido Customer Relationship Management, por sus siglas en inglés (CRM), evitando el desgaste de los clientes y construyendo relaciones sólidas con sus clientes, permitiendo entender la interacción que existe entre el negocio y los clientes. Ultsch (2002) encuentra que el CRM permite gestionar la relación con los clientes para conocer los vínculos entre el cliente y el negocio, haciendo uso de técnicas de aprendizaje estadístico se plantea encontrar información que permita adquirir conocimientos acerca de los clientes.

Por otro lado, se manifiesta que las tecnologías de almacenamiento de datos han permitido capturar información relevante para modelar la deserción de los clientes, en el caso financiero se encuentra que el volumen de transacciones y los montos consumidos son variables relevantes para predecir dicho evento (Qian, Jiang, & Tsui, 2006). Se conoce que las entidades que emiten tarjetas de crédito poseen información muy relevante sobre el uso de dichos productos y la información de los clientes, los campos que se han utilizado para resolver el problema de abandono son la cantidad de uso del cliente, montos consumidos, comercios donde compra, localidad de compra (local o extranjero), cuotas diferidas de las compras, forma de pago, entre otras que permiten conocer el comportamiento transaccional.

Multiples estudios que aplican minería de datos para estudiar el churn de clientes, consideran que no es usual que los clientes tomen la decisión de abandonar un producto de manera esporádica. Chen & Bose (2009) consideran que es relevante monitorear los cambios en comportaiento transaccional, dado que los clientes incrementan su riesgo a dejar la institución cuando empieza a disminuir su monto y frecuencia. Dejando a consideración que el no uso de puede considerarse como señal de incoformidad o de no necesidad.

La implementación de mejoras en la gestión de retención de clientes han generado las siguientes ganancias económicas:

1. Mejorar la usabilidad de la terjeta de crédito y recomendar el producto de boca a otras personas.
2. Disminuir la probabilidad de futura fuga dado el incremento a la lealtad por las campañas de retención.
3. La retención de los clientes reduce los costos al largo plazado dado que se registra información de su comportameinto transaccional.

Al conocer la importancia de los clientes de retención de los clientes se debe considerar los clientes que deben ser retenidos y los que no. En este sentido se debe hacer énfasis en los clientes que otorgan mayores beneficios para la empresa, por lo que empezar a gestionar el portafolio de clientes como activos estratégicos es clave para conseguir el éxito de la institución, permitiendo alcanzar una ventaja competetiva sostenible en el tiempo. (Valenzuela, Madariaga, & Blasco, 2007)

El evento de deserción de los clientes ha sido abordado con herramientas estadísticas. Keramati, Ghaneei, & Mohammad Mirmohammadi (2016) se basan en tecnologías existentes para recopilar información de las bases de datos de un banco para implementar minería de datos. La implementación de estás técnicas es para la extracción de conocimiento sobre las características relevantes de los clientes, utilizando árboles de decisiones, otorgando insumos necesarios para que los gerentes puedan identificar a los clientes propensos a abandonar la institución y puedan establecer estrategías de retención de los clientes. Demirberk (2021) mediante la mezcla de técnicas de estadísticas, Support Vector Machine con Optimización Bayesiana, desarrolla un algoritmo capaz de predecir la deserción de los clientes de tarjetas de crédito proponiendo a los departamentos de marketing soluciones técnicas que permiten enfocar sus esfuerzos para retener clientes.

Mutanen (2006) aplica regresión logística para implementar la gestión de la tasa de retención de clientes mejorando la relación entre la empresa y el cliente. Sin embargo, se encontró con la dificultad de tiempo dado que los perfiles de los clientes se encuentran en movimiento, por lo que resulta difícil tener un modelo estándar para predecir el evento. Bilal Zoric (2016) implementa métodos de míneria de datos y redes neuronales para predecir la rotación de los clientes utilizando las variables de ingresos, estatus laboral, aspectos sociodemográficos y la cantidad de productos que posee con el banco, encontrando que a medida que el cliente posee más de dos productos es menos propenso a no utilizar la tarjeta de crédito.

Al momento de realizar esta investigación no se ha encontrado trabajos relacionados a resolver la fuga de clientes aplicado a bancos ecuatorianos, pero sí a otras instituciones financieras, como Orellana Salcedo & Quezada Pico (2017) que aplica a una cooperativa de ahorro un estudio de técnicas cuantitativa y cualitativa para predecir posibles fugas o cambios de instituciones de los clientes para poder solventar estrategias de retención. Mediante el estudio cualitativo, levantan información necesaria para examinar el proceso de retención, misma información que sirve para crear un modelo de regresión logística para encontrar la relación entre las variables que influyen para identificar a los clientes más propensos a dejar de ser clientes de la institución financiera. Mientras Bohórquez, Torys, & Paredes Aguirre (2020) generan una solución analítica para anticiparse a la fuga de clientes de una administradora de fondos ecuatoriana, ante la competencia del mercado en el que se desenvuelve y buscan solucionar dicho problema con datos, para los cuales aplican árbol de decisiones, bosque aleatorio y regresión logística.

Queda demostrado en el presente capítulo que el problema de churn de clientes es ampliamente estudiado en múltiples sectores empresariales para las cuales se ha planteado diferentes soluciones que han reflejado mejoras en los niveles de retención, rentabilidad generada por los clientes y toma de decisiones estratégicas del negocio para mejorar la relación entre los clientes y las empresas. Esto motiva a aplicar dichas investigaciones sobre la presente entidad financiera con su producto de tarjeta de crédito, como se desarrollará en el siguiente capítulo.

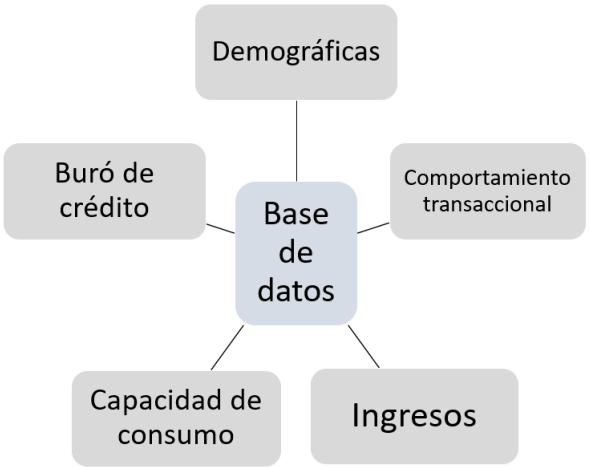
# CAPÍTULO 3

## 3. METODOLOGÍA

Para poder solventar la necesidad de que el negocio utilice su información para anticiparse a la conducta de los tarjetahabientes. En el presente capítulo se aborda la descripción de la base de datos a utilizar, estadísticas descriptivas de las variables predictoras, comportamiento de la variable dependiente, variables relevantes dado métricas estadísticas, el proceso de modelamiento y factores relevantes que permitirán escoger el mejor modelo.

### 3.1. Datos

Los datos implementados en el presente trabajo provienen de los registros administrativos de un banco privado en Ecuador. La construcción de la base de datos fue inspirada en los documentos revisados en el capítulo previo, en la siguiente gráfica se muestra a un alto nivel las dimensiones que cumple la base de datos:




***Gráfica 3.1:*** *Descripción de las dimensiones de la base de datos.*

La base de datos se encuentra constituida por las siguientes 5 dimensiones: demográficas, contemplan la edad, nivel de estudios, estado civil entre otras variables que caracterizan al tarjetahabiente; comportamiento transaccional, almacena las transacciones realizadas hasta que el cliente dejo de utilizar su tarjeta, dicha información se utiliza en forma monetaria y cantidad transaccional; el buró de crédito, registra la información del comportamiento de consumo y hábitos de pago en el sistema financiero; ingresos, son los valores percibidos por los tarjeta habientes dado su dependencia laboral; finalmente la capacidad de consumo que caracteriza los niveles de ingresos y gastos registrados por el cliente.

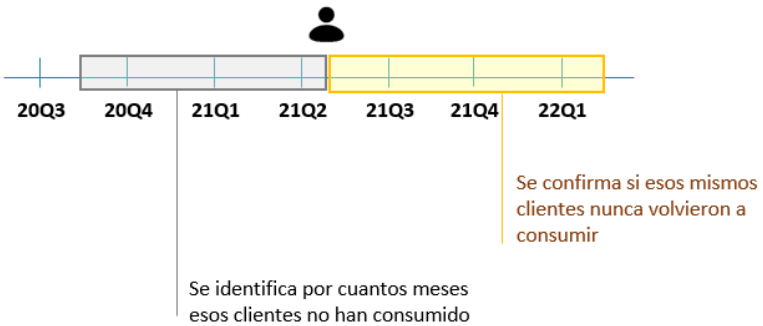
### 3.2. Variable dependiente

En la sección de descripción del problema se narra la forma en cómo se consolidó la variable dependiente. Para poder solventar el objetivo principal de la presente investigación, se detecta en el comportamiento de los tarjetahabientes aquellos que dejen de transaccional.

Dichos tarjetahabientes al no ser gestionados incurren en costos de activación por medio de campañas de marketing entre otros. Para definir la variable dependiente se obtuvo información histórica del comportamiento de las tarjetas habientes (disponible en la Gráfica 1.2) para lo cual se observa que en promedio los clientes que no han utilizado su tarjeta son cerca del 50%. A pesar de que las condiciones le permiten transaccionar no lo han hecho.

Además, se observa que el comportamiento transaccional cambia antes, durante y después de pandemia. Con la finalidad de evitar el sesgo de las estimaciones al momento de modelar se excluye del análisis los meses de pandemia, se cataloga de manera mensual a los clientes que están activos y que han dejado de consumir hace más de 6 meses.

Para determinar el evento a explicar se realizó un análisis por tramos de tiempo del cual se considera un punto de corte en un determinado periodo de tiempo y se observa en la ventana de tiempo posterior. Dicha idea se ve plasmada en la gráfica 3.2, donde en la sección de tiempo de la izquierda se identifica los clientes que han consumido hasta el periodo de tiempo que se corta el rango de tiempo (donde se ubica el icono de usuario) y los que se encuentran hacia atrás son los que han dejado de utilizar la tarjeta hace un periodo de tiempo. Estos clientes que han dejado de consumir (la sección de la izquierda) se busca si en el periodo de la derecha han vuelto a consumir (sección de la derecha).




***Gráfica 3.2:*** *Esquema de detección de meses que los clientes dejan de consumir.*

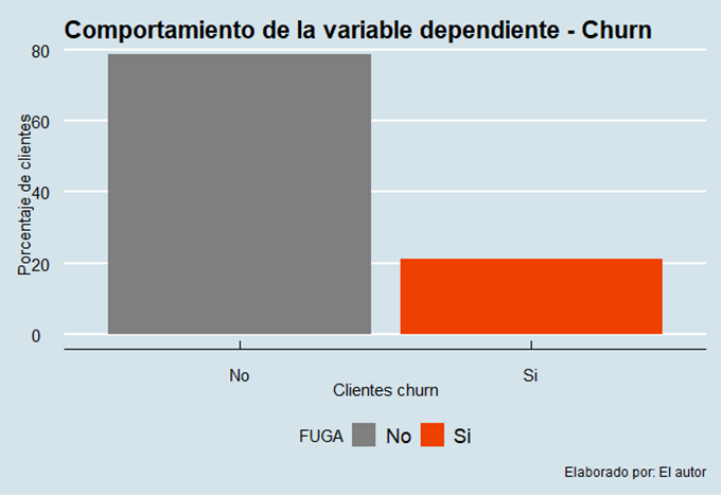
En la tabla 1 se presenta un proceso iterativo en 6 periodos diferentes del esquema antes mencionado para detectar el periodo de tiempo de no retorno. Los tarjetahabientes que dejan de consumir en al menos 5 meses, más de la mitad de dichos clientes vuelven a consumir.

| | Han vuelto a consumir? | | |
|---|---|---|---|
| Tiempo que no dejaron de consumir | NO Promedio | NO D.Estandar | SI Promedio |
| 01. Si consume | 4% | 0.02 | 96% |
| 02. Dejo de consumir hace 1 mes | 18% | 0.06 | 82% |
| 03. Dejo de consumir hace 2 meses | 28% | 0.08 | 72% |
| 04. Dejo de consumir hace 3 meses | 38% | 0.08 | 62% |
| 05. Dejo de consumir hace 4 meses | 47% | 0.07 | 53% |
| 06. Dejo de consumir hace 5 meses | 52% | 0.09 | 48% |
| 07. Dejo de consumir hace 6 meses | 59% | 0.07 | 41% |
| 08. Dejo de consumir hace 7 meses | 63% | 0.08 | 37% |
| 09. Dejo de consumir hace 8 meses | 68% | 0.09 | 32% |
| 10. Dejo de consumir hace 9 meses | 72% | 0.07 | 28% |
| 11. Dejo de consumir hace 10 meses | 76% | 0.08 | 24% |
| 13. Dejo de consumir hace 11 meses | 78% | 0.06 | 22% |
| 14. Dejo de consumir hace 12 meses | 81% | 0.06 | 19% |
| 15. Mas de 12 meses | 91% | 0.03 | 9% |

***Tabla 3.1:*** *Resultado del proceso iterado para detectar el punto de no retorno de los clientes que fuga*

Una vez que se determina que los clientes que dejan de consumir en al menos 6 meses son los que tienen una alta probabilidad de no volver a utilizar su tarjeta. El presente trabajo se enfocará en predecir a los clientes que sean propensos a dejar de recibir los servicios brindados por una tarjeta de crédito de un banco privado en Ecuador. Considerando lo asociado a la literatura, el objetivo principal de este trabajo es la anticipación de clientes que caigan en la categoría de perdidos, es decir, que sean predichos con anticipación para activar diferentes departamentos que puedan hacer su respectiva gestión para buscar retenerlos.

En la siguiente gráfica se muestra la distribución que tiene la variable dependiente dentro de la base de datos:




***Gráfica 3.3:*** *Comportamiento de la variable dependiente.*

Se encuentra que al periodo de análisis el 22% de los clientes de la base de datos son considerados en estado de churn o fuga dado que han dejado de transaccionar hace más de 6 meses.

### 3.3. Variables predictoras

La idea de estructurar la base de datos gira entorno a lo plasmado por Siddiqui (2017) donde construye modelos de predicción de no pago utilizando modelos dinámicos con bases de datos de corte transversal y sus variables cuantitativas son rezagadas. Al realizar dicha metodología se logra romper la estacionalidad de los eventos; es decir, controla los meses que tienen una alta presencia de clientes que no pagan sus obligaciones, mitigando el riesgo de tener problemas de endogeneidad[2].

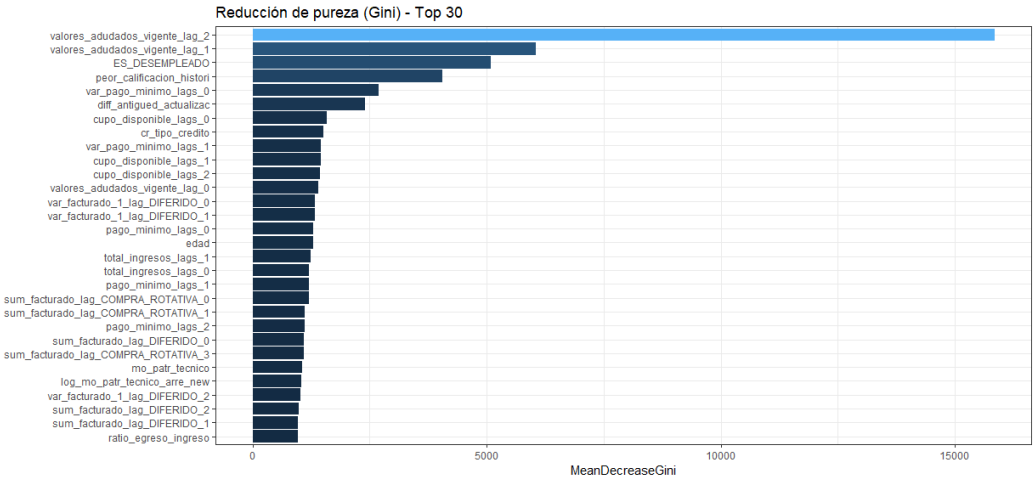
La base de datos posee alrededor de 400 variables, de las cuales 383 son cuantitativas y el complemento cualitativas. Por motivos de confidencialidad no se coloca el listo total de variables. Dado la alta dimensionalidad de la base para modelar, se aplicó dos formas de seleccionar variables: la primera, es utilizando métodos de selección de variables y la segunda, por conocimiento del giro del negocio.

Dado que la base de datos posee una alta cantidad de variables predictoras, se aplicó el análisis de componentes principales dado que Jolliffe (2002) lo utiliza para reducir la alta dimensionalidad de las bases de datos dado que en cada componente se almacena la información más relevante de cada conjunto de variables dado que a mayor cantidad de información se relaciona con mayor variabilidad, bajo esta primicia Oquendo (2021) aplica el análisis de componentes para reducir la cantidad de campos disponibles en la base de datos, lo cual se intentó aplicar en el presente trabajo pero no se obtuvo buen ajuste en el modelo.

### 3.4. Estadísticas descriptivas

Utilizando la métrica de impureza de Gini, dado que nos indica la capacidad de clasificar erróneamente una observación para lo cual Menze, y otros (2009) aplican esta definición para encontrar las variables que son malas clasificadoras dentro de la base de datos. Haciendo uso de las herramientas disponibles en el paquete estadístico caret, se extrae la reducción de pureza de Gini, que consiste en encontrar los predictores de mayor relevancia de la base de datos para clasificar de manera correcta. En la gráfica 3.4 se muestra el top 30 de las variables más relevantes:

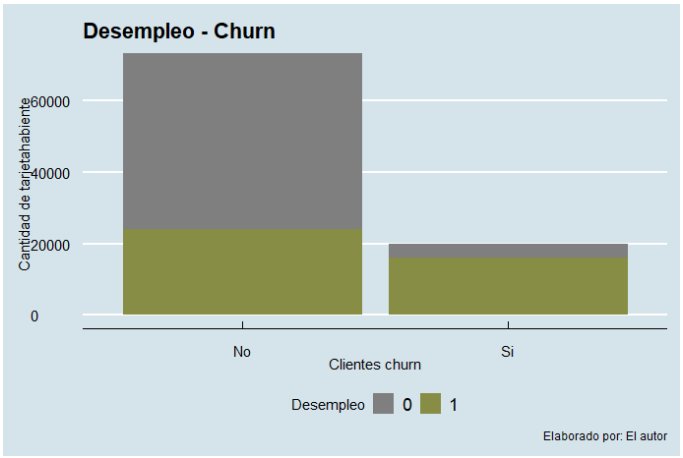
[2] En el siguiente link se encuentra la recopilación de estudios que tratan el trabajo de la endogeneidad: https://www.sciencedirect.com/topics/social-sciences/endogeneity-problem




***Gráfica 3.4:*** *Top 30 de variables con reducción de pureza (Gini)*

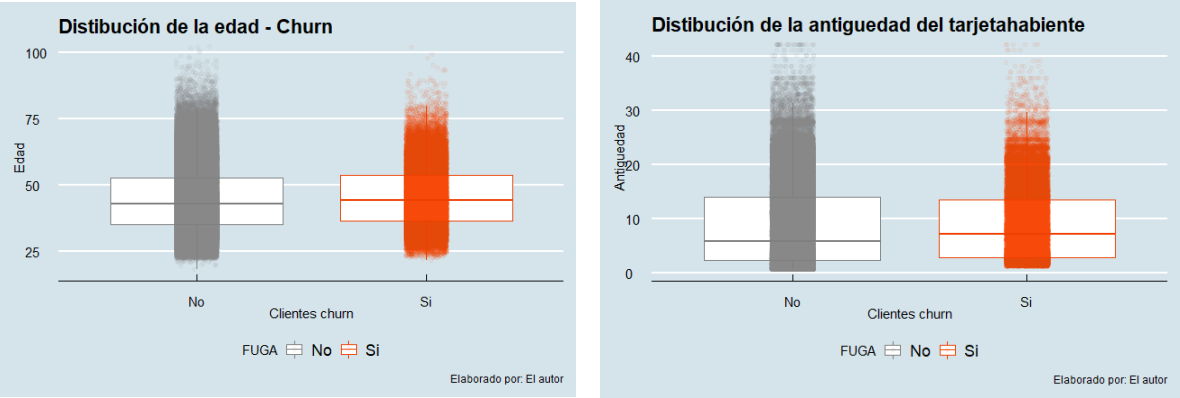
La gráfica 3.4 demuestra que las variables provenientes del buró de crédito: "valores adeudados en el momento de que ocurre el evento" hasta 2 rezagos menos son relevantes para clasificar los clientes que van a dejar de utilizar la tarjeta de crédito además de la peor calificación histórica. Por otra parte, se registran variables del nivel de ingresos y el total de empleo obtenidos hasta hace dos meses y finalmente se aprecia que el comportamiento de pago y de consumo son relevantes para predecir si un tarjetahabiente dejará de utilizar la tarjeta.

Dado que el nivel de desempleo se encuentra presente entre las variables de interés, se examina la participación de la variable dependiente. A continuación, se presenta el comportamiento de ambas variables, se aprecia que gran parte de los tarjetahabientes no están desempleados adicionalmente en su mayoría no han dejado de consumir. Mientras que, el 90% de los que han dejado de consumir registran que estaban desempleados en ese momento.




***Gráfica 3.5:*** *Estado de los tarjetahabientes con respecto a la fuga.*

Al mantener reuniones constantes con el giro del negocio para descubrir qué variables son las que consideran relevantes para explicar que un cliente deje o no de utilizar su tarjeta de crédito, se menciona que la edad y la antigüedad que mantiene relación con la institución son relevantes. Para comprobar dicha hipótesis se crea las siguientes gráficas que reflejan la distribución de dichas variables:




a) Edad b) Antigüedad

***Gráfica 3.6:*** *Distribución de la edad y la antigüedad contrastada con si el cliente comete o no fuga.*

En la gráfica 8 se evidencia que la edad y la antigüedad de manera visual no podría ser un buen predictor dado que la distribución reflejada en los boxplot es indiferente. El rango intercuartílico entre los clientes que dejan o no de utilizar la tarjeta de crédito con respecto a la edad y a la antigüedad son semejantes. Por lo que, la hipótesis planteada por el negocio refleja que no presenta un cambio brusco entre las categorías.

## 3.5. Modelado

Para el caso del modelo analítico se va a explorar algoritmos de clasificación como son Arboles de Clasificación, Bosques Aleatorias, Regresión Logística y Regularizaciones sobre la regresión logística buscando con esto obtener el modelo que mejor se ajusta a los datos y que sea consistente con sus resultados. Para medir la consistencia de los resultados, se ha buscado acoplarse al esquema "Out Of Sample" dado que Ballings & Van den Poel (2018) aplica en un proyecto de fuga donde plantean utilizar modelos dinámicos para mejorar el poder predictivo y las conjeturas realizadas por los modelos fuera de muestra.

La literatura de fuga de cliente gira en torno a dos técnicas analíticas que son la regresión logística y los árboles de decisiones (Neslin, Gupta, & Kamakura (2006) y Risselada, Verhoef, & Bijmolt (2003)). Se encuentra que en el ajuste de los resultados de los algoritmos depende de la normalización de los datos, cantidad de variable cualitativas y el tamaño de la muestra de entrenamiento. Se conoce que los árboles de clasificación son utilizados principalmente para problemas de clasificación y como variables predictoras ingresan campos numéricos y categóricos, dividiendo de esta forma los predictores en zonas separadas que no pueden ser sobrepuestas.

La aplicación de los Bosques Aleatorios (o cocidos como Random Forest) son propuestos por Breiman (2001) para mejorar los indicadores de predicción y mitigar el sobreajuste producido al aplicar Árboles de Clasificación dado que aplica la técnica de remuestreo denominada Boostrap que permite inferir la distribución a partir de repeticiones sujetas a la misma muestra.

El Bosque Alteatorio y regresión logística utilizando la regularización de Lasso son ampliamente aplicados en marketing como Lariviere & Van den Poel (2005) donde encuentran que el comportamiento pasado de los clientes influye de manera indirecta a la deserción de los clientes de una empresa de retail dado que, a medida que el cliente transacciona de manera más consecutiva reduce la probabilidad de que deje de consumir. Demostrando que los modelos se acoplan de manera significativa a los datos otorgando resultados confiables al negocio, incrementando la rentabilidad generada por cada cliente.

Dado que la base de datos será segmentada en 3 partes la primera consiste en separar datos hasta noviembre del 2021 para modelar, los datos que van desde diciembre del 2021 hasta mayo del 2022 servirán para probar el comportamiento del modelo fuera de muestra. Para la base de modelar será separada en entrenamiento y prueba, las misma que tendrán serán divididas de manera aleatoria en 80% y 20%, respectivamente. Donde en la base de prueba se va a seleccionar el mejor modelo dado la métrica de AUC.

# CAPÍTULO 4

## 4. RESULTADOS

En este capítulo se plasmará los resultados obtenidos aplicando la metodología planteada en el capítulo previo y la elección del mejor modelo.

### 4.1. Resultados de los modelos predictivos

Se probaron cinco diferentes métodos estadísticos de clasificación para responder a la problemática principal del presente trabajo. En primer lugar, iniciemos con los resultados obtenidos en el modelo de Regresión Logística, la capacidad de predecir de manera correcta es del 76% dado su nivel de Accuracy, alcanzando dichos valores con una Sensibilidad del 83% y Especificidad del 69%. (Anexo 2)

El modelo de regresión logística cuenta con 69 variables de las cuales 53 son estadísticamente significativas. De las variables con significancia estadística, a medida que incrementan la edad, los montos consumidos en compras resumidas de manera contemporánea, consumos diferidos con uno y dos rezagos, cantidad de transacciones realizadas en compras rotativas con un rezago, cupo disponible, cantidad de hijos menores, montos adeudados en el sistema financiero y antigüedad con la institución financiera, incrementan la probabilidad de que un cliente deje de utilizar la tarjeta de crédito mientras que, a medida que incrementan montos diferidos, cantidad de meses que se realizaron avances y nivel de patrimonio disminuyen la probabilidad de que suceda el evento estudiado.

Con la finalidad de eliminar las variables que no aportan de manera significativa a la predicción se utiliza el método de Regularización de Lasso para obtener las variables que permiten optimizar el error estimado en el modelo y la metodología de Stepwise en sentido de backward para eliminar las variables que no aportan de manera estadística. El ajuste medido por la métrica del Accuracy varía de manera sustancial con respecto a su estado original, el modelo de Stepwise reporta el valor del 76% en Accuracy. Dichos modelos cumplen con los supuestos de normalidad en sus residuos mientras que el modelo normal no. Por lo que, se recomienda en caso de escoger modelos que minimizan la varianza se escoja el modelo de Stepwise dado que los resultados se van a mantener en el tiempo.

Por otro lado, el modelo de Árbol de Decisiones logró obtener un ajuste en la base de prueba del 83% en Accuracy, 83% de Especificidad y 82% de Sensibilidad. Las variables relevantes que tienen mayor importancia al momento de clasificar son: montos adeudados, situación laboral, cantidad de empleos, ingresos percibidos por dependencia laboral, valor del mínimo a pagar, consumos realizados en rotativo, antigüedad con el banco, variaciones en el pago del mínimo, variación de consumos diferidos y la disponibilidad del cupo.

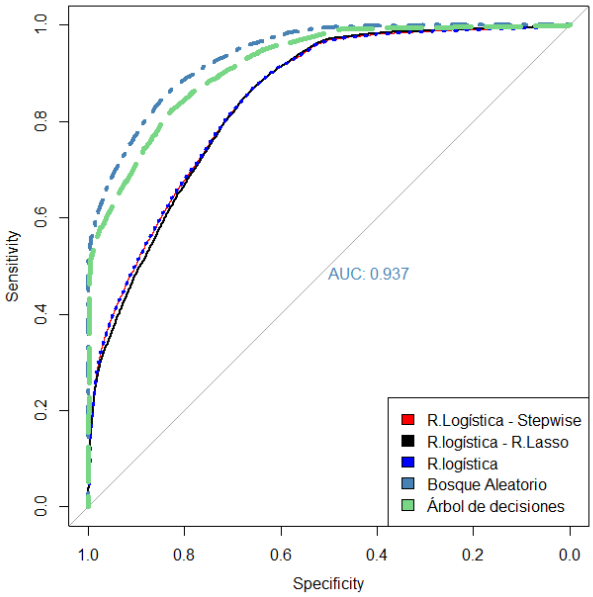
La intuición que se rescata de los modelos vistos hasta el momento son la interpretabilidad de estos y cómo ayudan al negocio. Según Varian (2014), la

regresión logística permite al negocio conocer las variables relevantes que incrementan o disminuye la probabilidad estimada mientras que, los resultados de los árboles de decisiones permiten entregar al negocio perfiles o reglas basadas en datos. Al combinar los resultados, el negocio podrá conocer el comportamiento de sus tarjetahabientes y poder tomar decisiones informadas.
Finalmente, los resultados del Bosque Aleatorio obtienen un mayor Accuracy en comparación con el resto de los modelos, el valor alcanzado es de 85%. El nivel de Especificidad es del 85% y Sensibilidad del 84%.

Las salidas del software sobre el ajuste en la base de datos de los 5 modelos se proporcionan en la sección de anexos.

### 4.2. Elección del mejor modelo

Para escoger el mejor modelo se utilizó la métrica AUC de los 5 modelos. En la gráfica 9 se muestra la comparativa entre modelos, el Bosque Aleatorio alcanza el valor más alto en el dataset de prueba del 94%.




***Gráfica 4.1:*** *Resultado de la curva ROC.*

Dado que el modelo se plantea para desplegar estrategias de retención sobre los clientes activos y que transaccionan. Se expone los modelos a la muestra que comprende a noviembre del 2021 hasta junio del 2022 para aplicarlo a un esquema de Out Of Sample. En la tabla 2 se muestra los valores en dicho escenario. Se evidencia la prueba de estabilidad del modelo seleccionado para desplegar la solución analítica clasifica de manera correcta el 70% de los casos.

| Modelo | Ajuste (Out of Sample) |
|---|---|
| Random Forest | 70% |
| Árbol de Decisiones | 68% |
| R. Logística con regularización Lasso | 65% |
| R. Logística con Stepwise | 65% |
| R. Logística | 63% |

***Tabla 4.1:*** Ajuste fuera de muestra.

En el Anexo 1 se registra que con la base de datos se obtuvo 20 componentes con valores propios mayores a 1, las variables que componen dichos componentes fueron probados en un modelo sin tener buen ajuste lo cual se debe a que en primer lugar que los componentes estén compuestos por variables con gran variabilidad no significa que sean buenos predictores. La reducción de dimensiones según lo trabajado por Urdinez & Caterina (2020) concuerda que con esta metodología se reduce dimensiones de la base de datos pero se enfatiza en que las variables deben tratar del mismo tema dado que los componentes girarán en torno al tema tratado.

# CAPÍTULO 5

## 5. CONCLUSIONES Y RECOMENDACIONES

Uno de los desafíos más importantes que enfrentan las instituciones financieras es la retención de los clientes en sus productos de activos y pasivos. La solución planteada del presente proyecto se enfoca en clientes activos de tarjeta de crédito, prevenir la salida de los clientes se vuelve crucial. Más aun con los factores como la creciente competencia, y clientes más exigentes que desean tener mayor calidad a un menor precio. La fuga de clientes, denominado como churn, afecta a un grupo cada vez más grande de la industria.

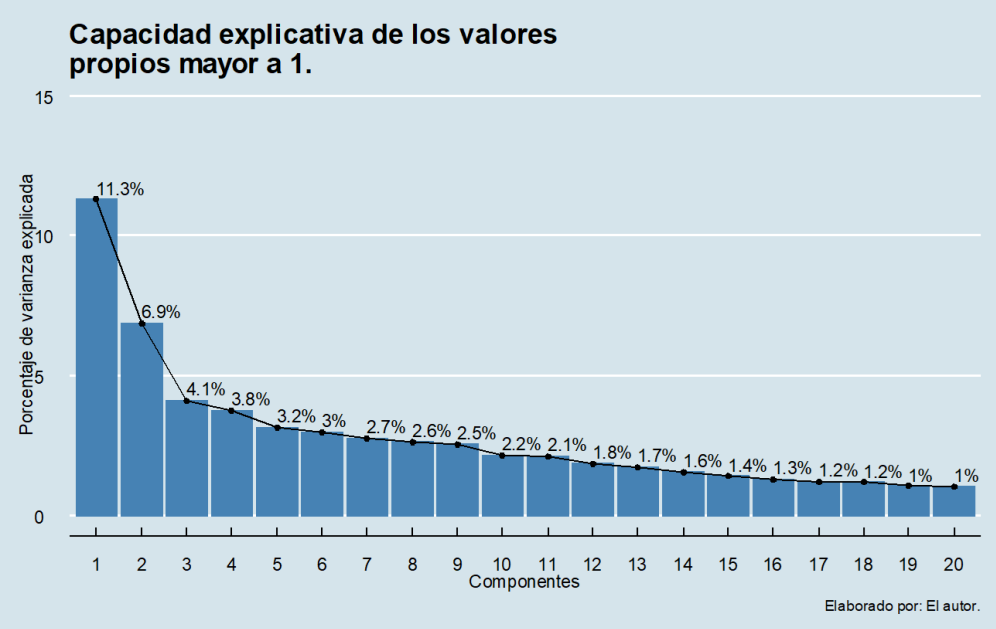
La relevancia del presente proyecto gira entorno a los costos de que atraer nuevos clientes es mucho más alto que mantener los clientes existentes. Es importante monitorear y prevenir la deserción de los clientes, para lo cual es presente proyecto propone una solución analítica que permite anticiparse a que suceda el evento, corrigiendo de esta forma los problemas de mediano y largo plazo como, por ejemplo, los clientes leales a largo plazo son embajadores de la organización y de las promociones realizadas.

El análisis de fuga de clientes de tarjeta de crédito gira al entorno de identificar clientes que dejen de transaccionar dado que la inactividad de clientes resta oportunidades al negocio de experimentar con otros clientes. El análisis determina la probabilidad de que un determinado cliente deje de utilizar los productos. Las variables que resultaron relevantes para construir el modelo probabilístico son las provenientes de las 5 dimensiones planteadas, las de mayor impacto son: cliente desempleado y valores adeudados en el sistema financiero. La relevancia que toma el estado de su situación laboral corresponde a que la institución financiera evalúa la fuente proveniente de ingresos, lo que se conecta con la variable de endeudamiento dado que se observa que tanto puede hacer frente con sus ingresos dependientes a las obligaciones adquiridas.

La probabilidad permite al banco de manera proactiva identificar clientes con suficiente probabilidad para generar alertas a los equipos correspondientes para plantear diferentes estrategias de retención. El éxito de la presente solución analítica se medirá por la disminución de clientes que dejan de utilizar su tarjeta de crédito, sujeto a la efectividad de las estrategias planteadas para romper el comportamiento previsto por el modelo.
Considerando que el modelo de predicción de fuga de clientes hace bien su trabajo hasta 3 meses después de la predicción según Neslin, Gupta, & Kamakura (2006) para lo cual se alertará al negocio con la base de clientes para su gestión de 3 meses y se revisará el desempeño en 2 corridas.
Las estrategias que se planteen deben ser evaluadas para poder establecer la reciprocidad de los clientes con la campaña y los objetivos planteados por el negocio.

## Anexos

**Anexo 1:** Componentes principales con valores propios mayores a 1.




**Anexo 2:** Resultados del ajuste en dataset de test – Regresión Logística.

```
Confusion Matrix and Statistics

          Reference
Prediction   NO   SI
        NO 6902 1709
        SI 3041 8234

               Accuracy : 0.7611
                 95% CI : (0.7551, 0.7671)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5223

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6942
            Specificity : 0.8281
         Pos Pred Value : 0.8015
         Neg Pred Value : 0.7303
             Prevalence : 0.5000
         Detection Rate : 0.3471
   Detection Prevalence : 0.4330
      Balanced Accuracy : 0.7611

       'Positive' Class : NO
```

**Anexo 3:** Resultados del ajuste en dataset de test – Regresión Logística con regularización de Lasso.

```
Confusion Matrix and Statistics

          Reference
Prediction   NO   SI
        NO 6861 1700
        SI 3082 8243

               Accuracy : 0.7595
                 95% CI : (0.7535, 0.7655)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5191

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6900
            Specificity : 0.8290
         Pos Pred Value : 0.8014
         Neg Pred Value : 0.7279
             Prevalence : 0.5000
         Detection Rate : 0.3450
   Detection Prevalence : 0.4305
      Balanced Accuracy : 0.7595

       'Positive' Class : NO
```

**Anexo 3:** Resultados del ajuste en dataset de test – Regresión Logística con Stepwise.

```
Confusion Matrix and Statistics

          Reference
Prediction   NO   SI
        NO 6906 1713
        SI 3037 8230

               Accuracy : 0.7611
                 95% CI : (0.7551, 0.7671)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5223

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6946
            Specificity : 0.8277
         Pos Pred Value : 0.8013
         Neg Pred Value : 0.7305
             Prevalence : 0.5000
         Detection Rate : 0.3473
   Detection Prevalence : 0.4334
      Balanced Accuracy : 0.7611

       'Positive' Class : NO
```

**Anexo 4:** Resultados del ajuste en dataset de test – Árbol de Decisiones.

```
Confusion Matrix and Statistics

          Reference
Prediction   NO   SI
        NO 8223 1766
        SI 1720 8177

               Accuracy : 0.8247
                 95% CI : (0.8193, 0.83)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6494

 Mcnemar's Test P-Value : 0.446

            Sensitivity : 0.8270
            Specificity : 0.8224
         Pos Pred Value : 0.8232
         Neg Pred Value : 0.8262
             Prevalence : 0.5000
         Detection Rate : 0.4135
   Detection Prevalence : 0.5023
      Balanced Accuracy : 0.8247

       'Positive' Class : NO
```

**Anexo 4:** Resultados del ajuste en dataset de test – Bosque Aleatorio.

```
Confusion Matrix and Statistics

          Reference
Prediction   NO   SI
        NO 8440 1566
        SI 1503 8377

               Accuracy : 0.8457
                 95% CI : (0.8406, 0.8507)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6913

 Mcnemar's Test P-Value : 0.2631

            Sensitivity : 0.8488
            Specificity : 0.8425
         Pos Pred Value : 0.8435
         Neg Pred Value : 0.8479
             Prevalence : 0.5000
         Detection Rate : 0.4244
   Detection Prevalence : 0.5032
      Balanced Accuracy : 0.8457

       'Positive' Class : NO
```

# Bibliografía


Ballings, M., & Van den Poel, D. (2012). Customer event history for churn prediction: How long is long enough? *CRM Ugent*, 13517-13522.

Ballings, M., & Van den Poel, D. (2018). Customer event history for churn prediction: how long is long enough? *Ghent University Academic Bbibliography*.

Bilal Zoric, A. (2016). Predicting customer churn in banking industry using neural networks. *Interdisciplinary Description of Complex Systems*, 116-124. Opgehaald van https://ideas.repec.org/a/zna/indecs/v14y2016i2p116-124.html#:~:text=In%20this%20paper%2C%20we%20used,those%20customers%20are%20worth%20retaining.

Bohórquez, M., Torys, J., & Paredes Aguirre, M. (2020). Modelos de predicción de deserción de clientes para una adminsitradora de fondos ecuatoriana. *Revista ESPOL*, 1-15.

Bolton, R. (1998). A Dynamic Model of the Duration of the Customer's Relationship with a Continuous Service Provider: The Role of Satisfaction. *Marketing Science*, 45-65. Opgehaald van https://pubsonline.informs.org/doi/abs/10.1287/mksc.17.1.45

Bolton, R., Kannan, P., & Bramlett, M. (2000). Implications of loyalty program membership and service experiences for customer retention and value. *Journal of the Academy of Marketing Science*, 95-108. Opgehaald van https://link.springer.com/article/10.1177/0092070300281009

Botelho, D., & Frederico Damian, T. (2010). Modelado de probabilidad de churn. *Revista de Administracao de Empresas*. Opgehaald van https://doi.org/10.1590/S0034-75902010000400005

Breiman, L. (2001). Random Forests. *Machine Learing*, 5-32.

Burez, J., & Van den Poel, D. (2007). CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications*, 277-288. Opgehaald van https://www.sciencedirect.com/science/article/abs/pii/S0957417405003374

Chen, X., & Bose, I. (2009). Hybrid Models Using Unsupervised Clustering for Prediction of Customer Churn. *Journal of Organizational Computing & Electronic Commerce*, 133-151. Opgehaald van https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.149.3287&rep=rep1&type=pdf

Coussement, K., Lessman, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems*, 27-36. Opgehaald van https://research-repository.uwa.edu.au/en/publications/a-comparative-analysis-of-data-preparation-algorithms-for-custome

Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems*, 27-36. Opgehaald van https://doi.org/10.1016/j.dss.2016.11.007

Demirberk, K. (2021). Predicting Credit Card Customer Churn Using Support Vector Machine Based on Bayesian Optimization. *Mathematics Subject Classification*, 827-836. doi:10.31801/cfsuasmas.899206

Gady, N., Baesens, B., & Croux, C. (2009). Modeling churn using customer lifetime value. *European Journal of Operations Research*, 402-411. Opgehaald van https://doi.org/10.1016/j.ejor.2008.06.027

Jolliffe, T. (2002). Principal Component Analysis. *Encyclopedia of Statistics in Behavioral Science*.

Keramati, A., Ghaneei, H., & Mohammad Mirmohammadi, S. (2016). Developing a prediction model for customer churn from electronic bank services using data mining. *Financial Innovation*, 1-16. doi:10.1186/s40854-016-0029-6

Keramati, A., Ghaneel, H., & Mohammad Mirmohammadi, S. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation*, 1-16. Opgehaald van http://dx.doi.org/10.1186/s40854-016-0029-6

Konukal, S. (2018). CREDIT CARD CHURN PREDICTION WITH MACHINE LEARNING ALGORITHMS. *MEF University*, 1-40. Opgehaald van https://openaccess.mef.edu.tr/xmlui/bitstream/handle/20.500.11779/1183/SerapKonuksal.pdf?sequence=6&isAllowed=y

Larivière, B., & Van den Poel, D. (2004). Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services. *Expert Systems with Applications*, 277-285.

Lariviere, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, 472-484. Opgehaald van https://www.sciencedirect.com/science/article/abs/pii/S0957417405000965

Menze, B., Kelm, M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*. doi:https://doi.org/10.1186/1471-2105-10-213

Mutanen, T. (2006). Customer churn analysis – a case study. *Independent Research Project in Applied Mathematics*, 3-18. Opgehaald van http://salserver.org.aalto.fi/vanhat_sivut/Opinnot/Mat-2.4108/pdf-files/emut06.pdf

Neslin, S., Gupta, S., & Kamakura, W. (2006). Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. *Journal of Marketing Research*, 204-211. Opgehaald van https://journals.sagepub.com/doi/abs/10.1509/jmkr.43.2.204

Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, 15273-15285.

Oquendo, F. (2021). Determinantes de la Rentabilidad en Cooperativas de ahorro y crédito en Ecuador. Un análisis mediante Machine Learning. *Facultad en Ciencias Económicas y Empresariales*, 20-22.

Orellana Salcedo, G., & Quezada Pico, J. (2017). Desarrollo de un modelo matemático experimental que permite determinar la predicción de fugas de clientes en el sector de las cooperativas de la industria financiera. *Universidad Internacional del Ecuador - Facultad de Ciencias Administrativas y Económicas*. Opgehaald van https://repositorio.uide.edu.ec/bitstream/37000/1949/1/T-UIDE-1467.pdf

Qi, J., Zhang, L., Liu, Y., Li, L., Zhou, Y., Shen Yao, . . . Li, H. (2009). ADTreesLogit model for customer churn prediction. *Annals of Operations Research*. Opgehaald van https://doi.org/10.1007/s10479-008-0400-8

Qian, Z., Jiang, W., & Tsui, K.-L. (2006). Churn detection via customer profile modelling. *International Journal of Production Research*, 2913-2933. Opgehaald van https://doi.org/10.1080/00207540600632240

Risselada, H., Verhoef, P., & Bijmolt, T. (2003). Staying Power of Churn Prediction Models. *Journal of Interative Marketing*, 198-208.

Siddiqui, N. (2017). *Intelligent Credit Scoring* (Vol. II). New Jersey: John Wiley & Sons.

Ultsch, A. (2002). Emergent self-organising feature maps used for prediction and prevention of churn in mobile phone markets. *Journal of Targeting, Measurement and Analysis for Marketing*, 314-324. Opgehaald van https://link.springer.com/content/pdf/10.1057%252Fpalgrave.jt.5740056.pdf

Urdinez, F., & Caterina, L. (2020). Analisis de Componentes Principales. In F. Urdinez, & A. Cruz, *Analizar Datos Políticos* (pp. 203-245).

Valenzuela, L., Madariaga, J., & Blasco, M. (2007). Customer Value Orientation and the New Metrics of Marketing: Review and Analysis. *Panorama Socioeconómico*, 70-75. Opgehaald van https://www.redalyc.org/pdf/399/39903407.pdf

Varian, H. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 3-28. Opgehaald van https://www.aeaweb.org/articles?id=10.1257/jep.28.2.3

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, 2354-2364. Opgehaald van https://www.sciencedirect.com/science/article/abs/pii/S0957417410008067