Escuela de Negocios. Master in Management + Analytics

Predicción de fuga de usuarios para una Fintech argentina

Autoría: Jumbo Narváez, Mishell Carolina
Año: *2024*

UNIVERSIDAD TORCUATO DI TELLA

MASTER IN MANAGEMENT + ANALYTICS

PREDICCIÓN DE FUGA DE USUARIOS PARA UNA FINTECH ARGENTINA

TESIS

Mishell Carolina Jumbo Narváez

mayo 2024

Tutor: Martín Ezequiel Masci

Resumen

Con el crecimiento y la aceleración de los productos y servicios ofrecidos por las *Fintech*, las empresas se enfrentan a una mayor competencia y el riesgo de perder usuarios. Esto impulsa a las *Fintech* a implementar estrategias de retención para atraer y mantener usuarios existentes, además de captar nuevos usuarios.

El objetivo de esta tesis es desarrollar e implementar un modelo de aprendizaje automático para identificar usuarios con alta probabilidad de dejar de utilizar los servicios de una *Fintech*, y utilizar esta información para diseñar y probar una estrategia efectiva de retención de usuarios. Este trabajo proporciona una metodología práctica para comprender el valor de los modelos de aprendizaje automático y su impacto financiero en un negocio.

Como resultado de este estudio, se logró desarrollar una estrategia que permitió a la *Fintech* incrementar los ingresos en un 35%. Estos resultados no solo beneficiarán a la empresa en cuestión, sino que la metodología también podría ser aplicada a otras empresas de la industria que enfrenten desafíos similares, como la fuga de usuarios.

Abstract

With the growth and acceleration of products and services offered by *Fintech*, companies face increased competition and the risk of losing users. This drives *Fintech* companies to implement retention strategies to attract and maintain existing users, in addition to attracting new users.

The objective of this thesis is to develop and implement a *machine learning* model to identify users with a high probability of stopping using the services of a *Fintech*, and use this information to design and test an effective user retention strategy. This work provides a practical methodology for understanding the value of *machine learning* models and their financial impact on a business.

As a result of this study, a strategy was developed that allowed the *Fintech* to increase its sales by 35%. These results will not only benefit the company in question, but the methodology could also be applied to other companies in the industry that face similar challenges, such as user *churn*.

Índice

Índice de tablas

Índice de figuras

1. Introducción

1.1. Contexto

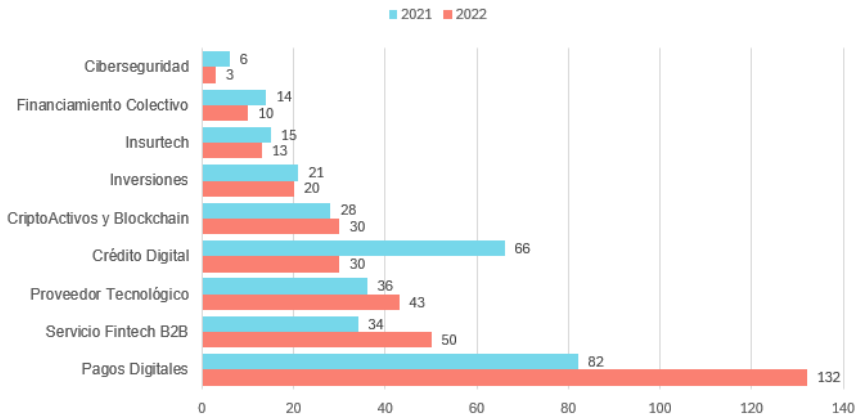
1.1.1. Proveedores de servicios de pago en Argentina

La relación de la tecnología en el ámbito financiero ha dado lugar al fenómeno "*Fintech*" que, según Maestre (2022), "*Fintech* es un sector integrado por empresas que utilizan la tecnología para mejorar o automatizar servicios y procesos financieros". Esto indica que estas empresas ofrecen una variada gama de productos o servicios financieros de manera accesible, mediante plataformas tecnológicas, tanto a los consumidores como a las empresas y a las instituciones financieras (Picón Montero & Vásquez Silva, 2022, p.31).

Según el informe de Finnovista; Banco Interamericano de Desarrollo; BID Invest. (2022, p.14), se indica que entre 2017 y 2021, el número de emprendimientos *Fintech* aumentó de 703 a 2.482 en América Latina. Es decir, que en "el periodo comprendido entre 2017 y 2021, el crecimiento anual promedio ha sido del 37% (equivalente al 253% desde la primera recolección de datos en 2017 hasta diciembre de 2021)". Esto indica que se está apostando cada vez más por transformar el ecosistema de finanzas digitales. Esta industria abarca nueve verticales de negocio, donde cada categoría se dedica únicamente a un servicio concreto, dando respuesta a necesidades financieras y digitales muy específicas.

Figura 1

Número de emprendimientos, por líneas de negocios Fintech en América Latina



Nota. Datos al cierre de cada periodo y estimación a diciembre de 2022.

Con base a la Figura 1 proporcionada por la Cámara Argentina *Fintech* (2022, p.4) se observa que la línea de negocio de pagos digitales tiene una participación significativa con 132 empresas en Argentina. Esta categoría está relacionada con el sistema de pagos y cobros, donde se incluyen billeteras digitales, servicios de procesamiento de pagos, agregadores, pasarelas de pagos, y empresas de remesas internacionales, entre otros, según informa la Agencia Argentina de Inversiones y Comercio Internacional (2023, p.10).

Una pasarela de pagos, también conocida como *gateway* de pagos, es un sistema que intercambia información de una transacción entre el comprador, el vendedor y las instituciones financieras, siempre con un procesador de pagos como intermediario. Esto permite a las empresas, tanto físicas como en línea, aceptar, procesar y gestionar diferentes métodos de pago, como tarjetas de crédito, tarjetas de débito y monederos digitales, de manera segura y eficiente. Esta plataforma actúa como intermediario y cobra una comisión por cada transacción procesada (Stripe, 2023).

Estas pasarelas de pago son desarrolladas por los Proveedores de Servicios de Pago (PSPs), que, según el Banco Central de la República Argentina (2023), incluyen un total de 141 empresas. Un PSP es una empresa *Fintech* que actúa como intermediario clave en el procesamiento de transacciones financieras, facilitando que las empresas acepten y gestionen pagos electrónicos. Los PSPs conectan a los comerciantes con el sistema financiero global, asegurando que las transacciones se realicen de manera segura y sin problemas. Además, los PSPs ofrecer servicios adicionales como gestión de riesgos, informes, remesas de fondos y protección contra fraudes, convirtiéndose en componentes esenciales del ecosistema de pagos (Muscillo, Vitale, & Peters, 2020).

1.1.2. Modelos de predicción para la predicción de churn

En un mercado altamente competitivo, los consumidores tienen la capacidad de cambiar de proveedores en su búsqueda de servicios de mayor calidad y eficiencia. Por ello, es importante que las empresas conozcan la tasa de abandono (*churn rate*), ya que permite medir el grado de retención de los usuarios y saber cuántos han dejado de serlo (Pozo, 2020).

Al consultar la literatura relacionada con esta temática, encontramos el artículo de Gutiérrez González (2020) sobre las "Técnicas de *machine learning* en el análisis del *churn rate*", donde indica que las técnicas de *data mining* con mejor resultado basándonos en la precisión media para resolver este tema son los árboles de decisión, seguido de las regresiones.

Otro factor interesante en este estudio es que la precisión de los árboles de decisión y las regresiones disminuye cuanto mayor es la base de datos estudiada. Por lo tanto, podría ser beneficioso aplicar redes neuronales y *support vector machine* en tamaños de muestra más grandes. Estos modelos son conocidos por su capacidad para manejar grandes volúmenes de datos y capturar relaciones no lineales complejas, lo que podría mejorar la precisión y la capacidad predictiva en conjuntos de datos más extensos.

Mientras tanto, en el artículo sobre "Predicción de *Churn* de Seguros con *LightGBM*" de Tralice (2019), se realiza un estudio en el sector de seguros utilizando el modelo *LightGBM*, donde se experimenta la implementación óptima de la ventana de tiempo. Dado que al aumentar la distancia entre los meses con los que se entrenó el modelo y el mes que se quiere predecir, puede existir un concepto *drift*. Concluye que entrenar con datos más actuales mejora el poder de predicción, y, por lo tanto, es fundamental tratar de manejar información más reciente.

Un artículo realizado dentro de la industria *Fintech* por Sierchuk (2022), titulado "Una estrategia de retención integradora que utiliza algoritmos de *Machine Learning* con el objetivo de eficientizar el uso del presupuesto de Marketing", realiza un

experimento con varios modelos, donde el modelo *LightGBM* dio mejores resultados después de realizar *random search* con *cross-validation* para la prueba de hiperparámetros. Sin embargo, manifiesta que los resultados pueden ser mejores aplicando cambios en la base de datos mediante ingeniería de atributos.

En referencia a trabajos prácticos previos sobre la predicción de la fuga de usuarios en empresas, destaca el artículo de Segura (2022) titulado "Desarrollo de un modelo de predicción de fuga de clientes y diseño de experimento para la aplicación de estrategias de fidelización en factoring". En este estudio, se llevó a cabo una simulación financiera del experimento a corto plazo y una simulación de la implementación de la estrategia a mediano plazo. Se compararon las utilidades y la cantidad de clientes acumulados para diversos niveles de fuga y retención.

Los resultados indicaron que, a corto plazo, la inversión condujo a una disminución del 11% en las utilidades generadas por el experimento, aunque se observó un aumento del 6% en la cantidad de clientes. Complementando estos hallazgos, la simulación a 12 meses reveló que tanto las utilidades como la cantidad de clientes acumulados experimentaron un aumento del 3%.

1.1.3. Modelo de negocio de la Fintech

La *Fintech* en estudio, fundada en 2020, ha crecido significativamente desde sus inicios en medio de la pandemia, emergiendo como un PSP dedicado a ofrecer soluciones tecnológicas innovadoras y en pasarelas de pago en argentina. Este trabajo se enfoca en un cliente específico de la *Fintech*, una empresa automotriz que administra planes de ahorro.

Un plan de ahorro es un sistema colaborativo que permite a los usuarios comprar un auto 0km mediante el pago de cuotas mensuales. Los usuarios se agrupan y, una vez alcanzado el número necesario de suscriptores, comienzan a pagar cuotas mensuales. Estos planes suelen tener una duración de 84 cuotas (7 años), y cada cuota se calcula dividiendo el precio actualizado del auto por el número total de meses del plan, más los gastos de administración y seguro de vida. Los autos se adjudican

mensualmente por dos métodos: sorteo, un proceso aleatorio que selecciona ganadores de manera equitativa, o licitación, donde los usuarios ofertan una cantidad de dinero para adelantar la entrega del vehículo, siendo asignado al mejor postor.

La *Fintech* ha desarrollado una plataforma exclusiva para este cliente, donde los usuarios pueden gestionar sus planes de ahorro. La plataforma permite el pago de cuotas y otros conceptos o documentos utilizando diversas alternativas de métodos de pago, incluyendo tarjetas de crédito y débito, pagos en efectivo y tres modalidades gestionadas por la *Fintech*: débito inmediato (DEBIN), transferencia a través de una cuenta virtual uniforme (CVU) y débito en cuenta automático (DC). La *Fintech* cobra una comisión por cada transacción exitosa.

Diariamente, la *Fintech* procesa las transacciones recaudadas, consolidando y preparando la información para informar al cliente sobre los pagos recibidos de los usuarios. Este proceso incluye la rendición de lotes, que es la transferencia de los fondos acumulados en las cuentas recaudadoras gestionadas por la *Fintech*.

1.2. Problema

Desde su inicio, este proveedor de servicios ha estado ofreciendo servicios de procesamiento de pagos a través de tres *gateways* distintos. Tras un análisis se ha observado que la mayoría de las transacciones procesadas y registradas corresponden al método de DEBIN, con un 69% de las transacciones totales, seguido de un 16% que utiliza CVU, y un 15% que se realiza mediante DC. Todas estas operaciones se llevan a cabo dentro de su plataforma, y se cobra una comisión basada en un porcentaje por el tipo de método de pago utilizado, lo cual ha sido un componente fundamental de su modelo de negocio ya que representa sus ingresos o ventas.

Sin embargo, a partir de marzo de 2023, el panorama ha experimentado algunos cambios. La empresa automotriz ha optado por ampliar sus opciones de métodos de cobro, colaborando con otros entes financieros, donde el proveedor de servicios de pagos se encarga únicamente de consolidar e informar los pagos

recibidos por estos, sin imponer ninguna comisión. Con la consecuencia de que los usuarios que utilizan los servicios de la *Fintech* tienen la oportunidad de optar por métodos de pago alternativos.

Una de las métricas más utilizadas por las empresas es conocer la retención de usuarios. Según Zendesk (2022) “es la capacidad que tiene una empresa para mantener la estabilidad de su cartera de clientes a través de la satisfacción y la calidad del producto o servicio.” Una de las herramientas analíticas útiles para entender la retención de usuarios a lo largo del tiempo es el análisis de cohortes[1], que se basa en dividir a los usuarios por cohortes e interpretar cuál es el tiempo de vida útil de cada una de ellas (Saltos, 2022).

Figura 2

Matriz de cohortes Fintech 2023

Meses de Vida	0	1	2	3	4	5	6	7	8	Promedio
ene-23	100,00%	60,33%	55,93%	44,36%	41,55%	39,10%	37,88%	35,68%	32,71%	49,73%
feb-23	100,00%	47,90%	37,53%	35,84%	33,12%	32,26%	30,18%	27,14%	26,99%	41,22%
mar-23	100,00%	40,68%	39,01%	36,34%	34,12%	31,88%	28,57%	27,84%		42,31%
abr-23	100,00%	46,10%	39,10%	37,25%	34,28%	31,17%	30,42%			45,47%
may-23	100,00%	57,61%	55,36%	50,87%	47,23%	46,63%				59,62%
jun-23	100,00%	58,77%	52,45%	46,59%	44,09%					60,38%
jul-23	100,00%	54,72%	45,39%	42,54%						60,66%
ago-23	100,00%	47,08%	41,75%							62,94%
sep-23	100,00%	51,70%								75,85%
Promedio	100,00%	51,65%	45,82%	41,97%	39,07%	36,21%	31,76%	30,22%	29,85%	45,17%
Variación		-48,35%	-11,30%	-8,39%	-6,92%	-7,31%	-12,28%	-4,86%	-1,22%	

En la figura 2, se presenta una matriz de retención de usuarios de la *Fintech* en el año 2023. Las filas corresponden a cada cohorte mensual, abarcando desde enero hasta septiembre de 2023, mientras que las columnas representan los meses siguientes a la fecha de creación de cada cohorte. El mes 0 representa el mes de creación de cada cohorte, el mes 1 es el siguiente, y así sucesivamente. Los valores en la tabla indican el porcentaje de usuarios de cada cohorte que permanecieron utilizando los métodos de pago de la *Fintech* en meses específicos. Por ejemplo, podemos observar que el 47.08% de los clientes que realizaron su primer pago a través de la plataforma de la *Fintech* en el mes de agosto continuaron utilizando los

[1] El término cohorte hace referencia a grupos de usuarios que fueron adquiridos por primera vez en el mismo periodo de tiempo.

servicios de pago de la *Fintech* en el mes 1, es decir, el mes de septiembre (mes subsiguiente) a su primera transacción.

Un dato relevante es que, en promedio, solo se retiene un 51.65% de los usuarios en el mes siguiente a su primera transacción mediante los métodos de pago DEBIN, CVU o DC, lo que representa la mayor pérdida de usuarios que no vuelven a utilizar estos servicios. Esto conlleva a pérdidas significativas para la *Fintech*, ya que no genera ingresos. En un mercado altamente competitivo y con numerosas opciones disponibles, es esencial realizar un análisis detallado y aplicar herramientas y estrategias efectivas para contrarrestar esta disminución en la retención de usuarios.

1.3. Objetivo

De acuerdo con lo mencionado anteriormente, el Proveedor de Servicios de Pagos (PSP) se encuentra en un mercado altamente competitivo, lo que requiere la exploración de estrategias comerciales y la retención efectiva de los usuarios actuales para mantener una posición sólida en el mercado. Como se evidenció en el punto anterior la *Fintech* en el mes siguiente que realizar la primera transacción se pierde un 48.35% de usuarios. Es decir, estos usuarios no vuelven a utilizar métodos de pagos que se ofrecen en la plataforma y deciden pagar por otros métodos de pago distintos de las cuales la *Fintech* no cobra ninguna comisión por esa transaccionalidad.

Para mitigar esta disminución en la pérdida de usuarios, se aprovechará la información contenida en la base de datos del PSP de los usuarios y sus respectivos historiales de transacciones a lo largo del tiempo. La finalidad principal es permitir que un modelo de aprendizaje automático identifique patrones de comportamiento y pronostique qué usuarios tienen una alta probabilidad de cambiar su método de pago, diferentes a DEBIN, CVU o DC. Este proceso involucra la aplicación de tres modelos de aprendizaje automático distintos: regresión logística, árboles de decisión y XGBoost. Estos modelos se someterán a una posterior comparación para determinar cuál de ellos ofrece un rendimiento superior en la precisión de las estimaciones.

Con el propósito de validar el éxito del modelo elegido, se decidió retener una parte de los datos en nuestro conjunto de información, específicamente los registros más recientes correspondientes al último mes disponible. Esta selección responderá a la intención de evaluar el desempeño y la precisión del modelo en un entorno real, permitiendo verificar si los usuarios identificados como aquellos con alta probabilidad de cambiar su método de pago efectivamente dejarán de pagar por DEBIN, CVU o DC.

Con este estudio se busca recomendar a la *Fintech* establecer estrategias de retención de usuarios que en primera instancia será utilizar la mensajería SMS - Bot de WhatsApp como herramienta para promover y destacar los métodos de pago ofrecidos por la *Fintech*. Es esencial indicar que esta estrategia no se aplicará a todos los usuarios debido a su costo asociado. Por lo tanto, es importante la validación del modelo ya que se centrará en los usuarios identificados con una alta probabilidad de cambiar su método de pago a uno que no genera comisiones para el PSP.

Se llevará a cabo una comparación entre los ingresos que se podría ganar y los usuarios operativos a lo largo del tiempo que la *Fintech* podría obtener mediante la aplicación de la estrategia de envío de mensajes. De esta manera, se calculará el porcentaje de incremento en los ingresos y, al mismo tiempo, mantener o mejorar el indicador de retención de usuarios. Es decir, el presente trabajo busca cómo predecir la fuga de usuarios para una *Fintech* argentina, con el objetivo de evitar la disminución en los niveles de ingresos y mantener una posición sólida en el mercado.

2. Datos

Una vez identificado el modelo de negocio, el problema y el objetivo, en este capítulo se revisó la información de la *Fintech*, este consta de dos etapas distintas. En la primera etapa, se describirá el proceso de obtención y comprensión de los datos. Mientras que, en la segunda etapa, se llevará a cabo una revisión detallada de la distribución de las variables, con una explicación del tratamiento aplicado a los valores atípicos y a los datos faltantes. Además, se realizará una reorganización de la

información, junto con la creación de nuevas variables con el propósito de descubrir características demográficas e identificar comportamientos de los usuarios para poder entender el problema de la fuga de usuarios de la *Fintech*.

2.1. Obtención y compresión de datos

Los datos fueron provistos por la *Fintech* y contienen registros desde el inicio de las operaciones que corresponde de abril 2020 hasta el diciembre 2023. De acuerdo con el objetivo del trabajo, que implica desarrollar un modelo para predecir la fuga de usuarios usando técnicas de aprendizaje automático, con el fin de identificar a aquellos usuarios con alta probabilidad de abandonar o dejar de utilizar los servicios de pago ofrecidos por la *Fintech*, es fundamental entender las fuentes de datos y el significado de cada uno de los campos que contiene los archivos.

El primer archivo contiene información de la tabla llamada “User”, que almacena detalles sobre los usuarios que se dan de alta en un plan de ahorro en la automotriz con el fin de que puedan ser activados y utilizados por los usuarios dentro de la plataforma de la *Fintech*. Esta tabla abarca un total de 608.863 observaciones, distribuidas en 78 variables. Esta incluye datos importantes y detalles clave, como la fecha de nacimiento, la fecha de registro, la dirección de correo electrónico, el tipo de persona, el CUIT[2], el nombre del titular del plan, el estado civil, el género y la provincia de residencia, entre otras.

Aunque la plataforma tiene un número determinado de usuarios registrados, algunos no han ingresado para realizar gestiones de pago. Para utilizar la plataforma, es necesario generar una contraseña por primera vez, pero hay usuarios que no han completado este procedimiento y se consideran no activados. Lo considero importante mencionar porque estos usuarios no tendrán ningún registro en la tabla de operaciones o transacciones.

[2] CUIT refiere a clave única de identificación tributaria que es un código con el que la AFIP identifica a trabajadores autónomos, comercios y empresas.

Mientras que el segundo archivo contiene información de la tabla "Transacciones" la cual abarca todas las actividades operativas de pago llevadas a cabo por el usuario dentro de la plataforma diseñada por la *Fintech*. Esta registra 5.187.793 observaciones con 69 variables, en ella encontramos información como el tipo de *gateway* o método de pago utilizado por el usuario, el estado de la transacción (acreditada, pendiente, cancelada), fecha de creación de la transacción, fecha de acreditación, la fecha de rendición del pago al cliente, monto cobrado, monto de la comisión cobrada para la *Fintech*, el CUIT del usuario que realizó el pago y el tipo de documento pagado, entre otros. En esta misma tabla, se identificó que existen 192.054 *CUIT* distintos que utilizaron los tres métodos de pago ofrecidos por la *Fintech*, es decir que en promedio existen 27 transacciones por usuario. Además, esta tabla incluye operaciones que no pudieron completarse y que, por lo tanto, no fueron procesadas con éxito.

Según el Apéndice A, se detalla la cantidad de datos no nulos por variable de las dos tablas. Se observa que diversas variables presentan valores vacíos que podrían no contribuir al análisis y al posterior modelo predictivo. Por esta razón, se opta por seleccionar las variables más relevantes en términos demográficos y que proporcionen una mayor cantidad de información.

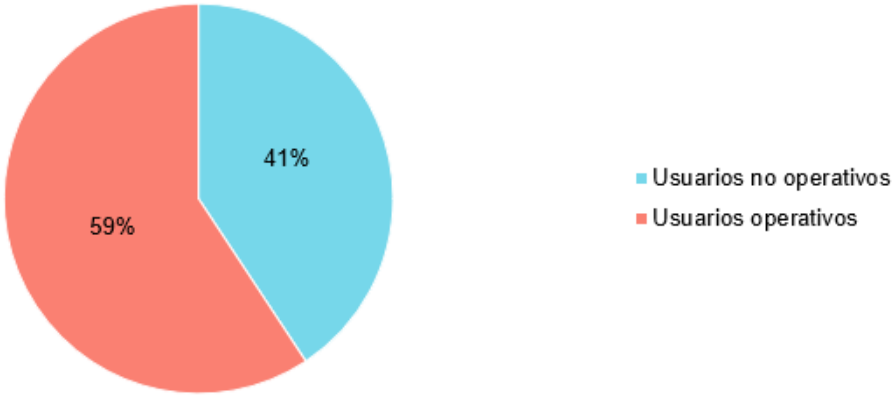
2.2. Análisis exploratorio

Como punto de partida, se analizó la distribución porcentual del estado de los usuarios en la plataforma de la *Fintech*. La Figura 3 muestra que el 41% de los usuarios intentaron realizar pagos, pero no lograron completar ninguna transacción exitosa. Para el análisis, se incluyeron solo los usuarios operativos, definidos como aquellos que tuvieron al menos un pago exitoso o una transacción acreditada correctamente, además de posibles intentos fallidos. Los usuarios no operativos,

aquellos que solo tienen intentos fallidos, fueron excluidos del análisis. Excluyendo al 41% de usuarios no operativos, el análisis se reduce a 119.465 usuarios.

Figura 3

Distribución porcentual de estado de usuarios en la plataforma



La primera variable analizada es la cantidad de planes asociados a cada usuario. Según se aprecia en la Figura 4, un 38% de los usuarios tienen registrados dos planes de autos, un 35% solo un plan, y un 9% tienen tres planes. Además, hay usuarios con más de cuatro planes, lo que sugiere que son empresas o concesionarios. Por otro lado, se identificó que un 10% de los CUIT registrados no cuentan con información de alta en la tabla de usuarios, indicando que el pago se realizó a través de una cuenta bancaria diferente al titular del plan.

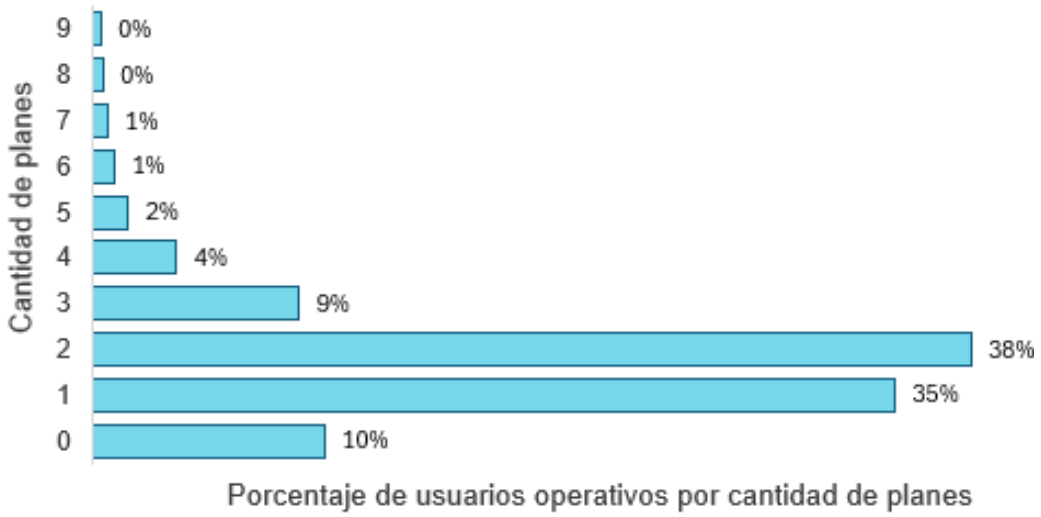
Los usuarios con un solo plan se han suscrito para la compra de un auto. Aquellos con dos o tres planes generalmente lo hacen para uso personal y familiar, o con fines empresariales o de inversión. Los usuarios con más de cuatro planes suelen ser empresas comerciales, como concesionarias, compañías de alquiler de vehículos, empresas de servicios, instituciones educativas, agencias de turismo y viajes, empresas de seguridad y grandes corporaciones que requieren una flota de vehículos para sus operaciones.

Este análisis brinda una visión detallada sobre la distribución de los planes asociados a cada usuario, permitiendo determinar la frecuencia de los pagos en relación con la cantidad de planes suscritos. Se destaca que los usuarios con múltiples

planes muestran un mayor compromiso con la plataforma, lo que se traduce en una mayor cantidad de operaciones.

Figura 4

Distribución porcentual de cantidad de planes de usuarios operativos

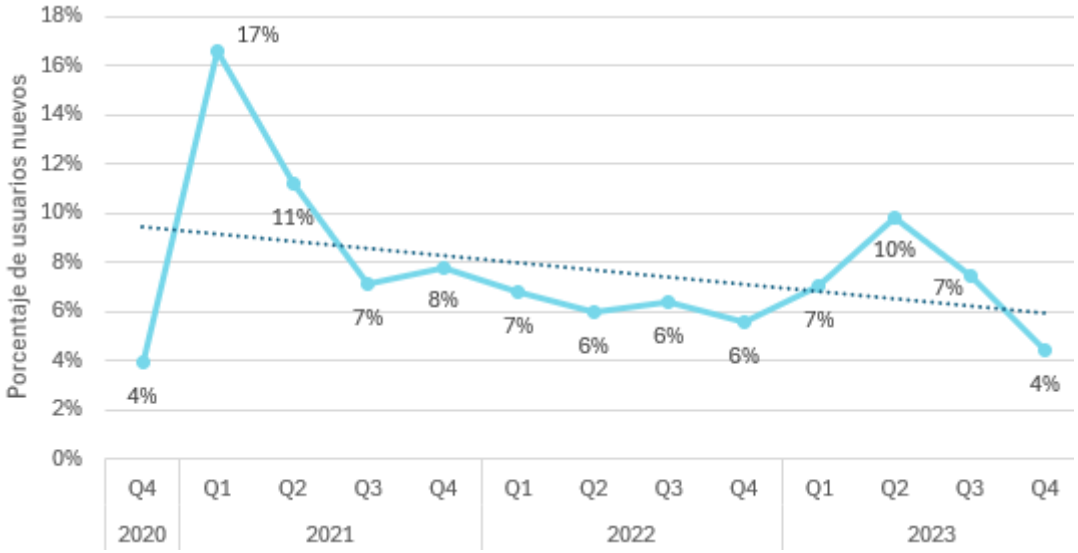


Se analizó la cantidad de usuarios nuevos que realizaron su primera transacción según el mes y año. En el primer trimestre de 2021, se observó que el 17% del total de usuarios nuevos comenzaron a transaccionar en ese período, como se muestra en la Figura 5. Este trimestre marcó el momento en que la *Fintech* atrajo la mayor cantidad de usuarios nuevos, influenciado por la pandemia del COVID-19.

En el segundo trimestre de 2023, se identificó un ligero aumento de usuarios nuevos, representando el 10% del total. Sin embargo, a partir de esa fecha, se observa una tendencia a la baja en el periodo analizado. Este hallazgo refuerza la necesidad de implementar estrategias para prevenir la fuga de usuarios y mantener a los usuarios operativos en la *Fintech*, dado que no se está evidenciando un incremento significativo de nuevos usuarios.

Figura 5

Distribución porcentual de usuarios nuevos por trimestre



En el ámbito de la información demográfica de los usuarios que han realizado transacciones con la *Fintech*, se destaca que, en cuanto al tipo de persona, un 78% corresponde a personas naturales, un 2% a entidades jurídicas, y un 20% se desconoce debido a la falta de actualización de información en la plataforma, según se muestra en la Figura 6. En lo que respecta al estado civil, el 56% de los usuarios no tienen registrada esta información. De aquellos para los que se dispone de información, el 23% son solteros y el 15% están casados como se evidencia en la Figura 7.

Figura 6

Distribución porcentual de usuarios operativos por tipo de persona

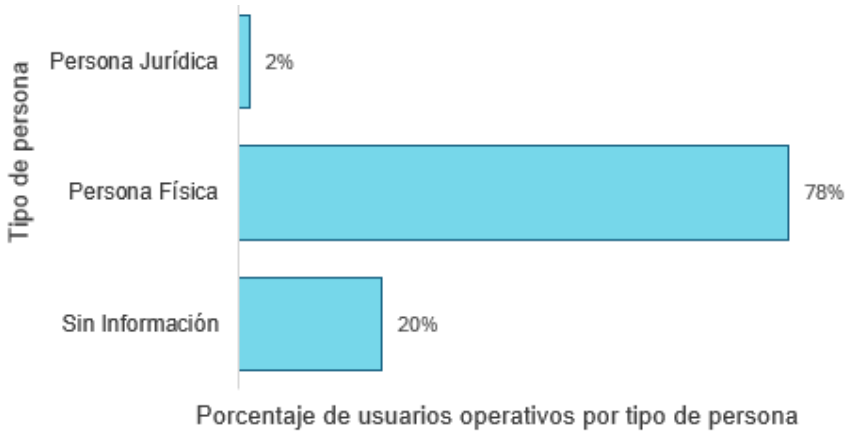
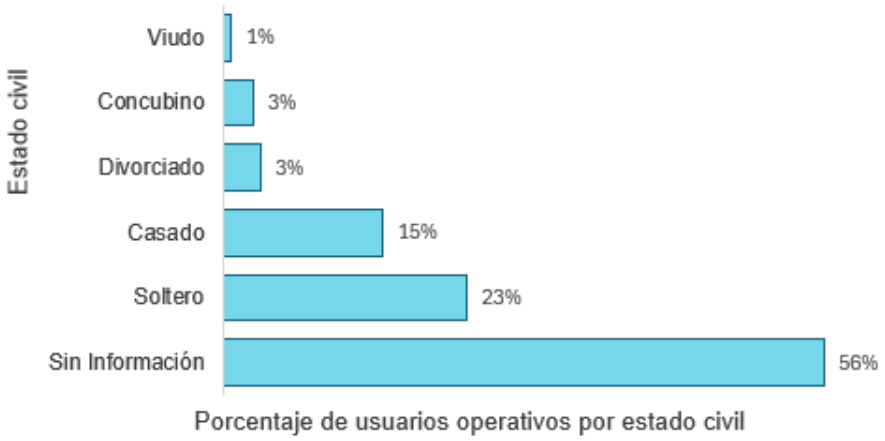


Figura 7

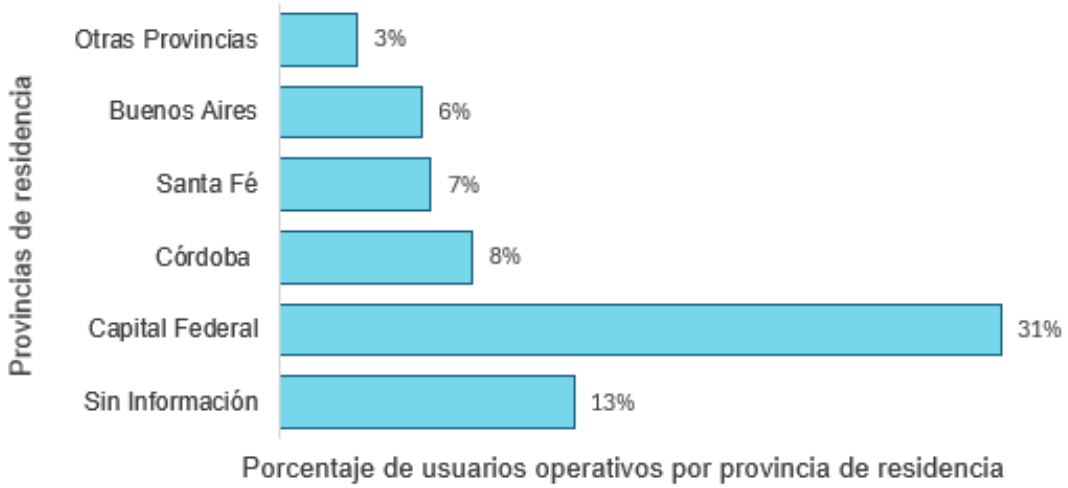
Distribución porcentual de usuarios operativos por estado civil



En cuanto a la ubicación geográfica de los usuarios que se detalla en la Figura 8, se observa que el 31% son residentes de Capital Federal, seguidos por un 8% en Córdoba, un 7% en Santa Fe y un 6% en Buenos Aires. Cabe destacar que un 13% no cuenta con información detallada sobre su ubicación geográfica.

Figura 8

Distribución porcentual de usuarios operativos por provincia



Nota: La categoría 'Otras provincias' incluye aquellas cuya participación individual representa menos del 3% del total de usuarios. En conjunto, estas provincias representan el 35% de los usuarios.

2.2.1. Tratamiento de outliers y missing values

En el apartado anterior, se detalló de manera descriptiva las distribuciones de las variables demográficas de los usuarios operativos contenidos en la base de datos. Sin embargo, es importante abordar situaciones que requieren atención para la construcción del modelo de predicción de fuga. Estos desafíos incluyen la presencia de errores o valores anómalos en la base de datos de operaciones realizadas por los usuarios, que podrían clasificarse como *outliers* o *missing values*.

2.2.2.1. Outliers

Un *outlier* es una observación atípica y extrema en una muestra estadística o serie temporal de datos que puede afectar potencialmente a la estimación de los parámetros. En el análisis de los *outliers*, se tomarán variables numéricas que reflejan el comportamiento de las operaciones de los usuarios, es decir información que contiene la tabla de transacciones. En esta ocasión se consideran dos variables, en primer lugar, el importe cobrado (expresado en pesos argentinos) del usuario y, por otro lado, el valor de la comisión que genera la *Fintech* por la transacción exitosa. El objetivo es identificar posibles transacciones que puedan estar sesgadas debido a la presencia de observaciones con valores anormales o atípicos, que se detallan a continuación:

Tabla 1

Medidas estadísticas en variables cuantitativas

Estadístico	Importe Cobrado	Importe Comisión
Mínimo	$0	$0
Mediana	$32.270	$0
Promedio	$156.175	$692
Máximo	$270.000.000	$999.000
Desviación Estándar	$2.089.551	$9.365
Tercer Cuartil	$85.463	$151

Nota: Montos expresados en pesos argentinos

En la Tabla 1, se presentan las medidas estadísticas de las dos variables. Se observa que no se registran valores negativos, cumpliendo con la coherencia de los datos. Además, se identifica que los montos varían desde $0 hasta $270.000.000 para el monto recaudado y hasta $999.0000 para las comisiones. Es evidente una alta desviación estándar en ambas variables, indicativa de una significativa variabilidad en los montos cobrados por los usuarios y en las comisiones generadas. Aunque se reconoce la presencia de *outliers*, se ha decidido no eliminar observaciones, ya que se tiene la intención de emplear modelos de aprendizaje automático capaces de gestionar eficientemente dichos valores atípicos.

2.2.2.2. Missing values

Los *Missing Values* son los datos o valores faltantes ocurren cuando no se almacena un valor de datos para la variable en una observación. En este análisis las variables que contienen mayor porcentaje de *missing values* se detallan en la tabla 2.

Tabla 2

Valores faltantes de variables

Variables	Cantidad	Porcentaje
Fecha rendición	3.203.728	77%
Lote rendición	3.197.800	77%
Tipo de comprobante	84	0%

La fecha de rendición y el lote de rendición se completan únicamente cuando una transacción se finaliza exitosamente, es decir, cuando el usuario realiza un pago exitoso a través de la plataforma de la *Fintech*. La fecha de rendición señala el momento en que se informa al cliente sobre el pago recaudado y se liquida o transfiere el monto recaudado al usuario pagador, incluyendo la comisión correspondiente para la *Fintech*. Por otro lado, el lote de rendición es un código que permite identificar el conjunto de pagos recaudados en un intervalo de tiempo específico.

A pesar de su relevancia, las variables de fecha de rendición y lote de rendición presentan valores faltantes en el 77% de la base de datos. Esta ausencia de datos se

justifica por la naturaleza de la transacción, ya que se encuentran en un estado no finalizado. Es decir, únicamente el 23% de las observaciones corresponden a transacciones exitosas en la tabla de operaciones.

Respecto a la variable "tipo de comprobante", se identifican 84 transacciones sin clasificación del tipo de documento que el usuario intenta pagar. Sin embargo, se prevé abordar esta falta de información utilizando los datos detallados de la variable "concepto", la cual proporciona información más específica sobre el pago y permitirá completar los datos faltantes en esta categoría.

Además, se detectaron 120 transacciones en las que no se identifica al usuario correspondiente, dado que el campo CUIT no se ajustaba a la estructura esperada para este tipo de campo. Por lo tanto, se optó por utilizar únicamente los usuarios que tuvieran un CUIT válido, es decir, aquellos que contaran con la cantidad de 11 dígitos. En consecuencia, se optó por eliminar estas transacciones de la base de datos, dado que representan un 0.0028% y no afectan de manera significativa la integridad de la base. Con este análisis la cantidad de usuarios para análisis son **119.448.**

2.2.2. Reestructuración de la base de datos

Como se mencionó anteriormente, es fundamental indicar que la tabla de transacciones registra múltiples operaciones realizadas por cada usuario a lo largo del tiempo, lo que genera múltiples registros por usuario, cada uno asociado a una característica de pago específica. Debido a esta estructura de datos, se ha tomado la decisión de consolidar toda esta información en una única base de datos agrupado por usuario.

Para llevar a cabo esta unión, se comenzó por unir la información de las operaciones, aplicando una consolidación por usuario en función de la cantidad de operaciones. Estas fueron segmentadas por su resultado, ya sea cantidad de transacciones exitosas, intentos o canceladas por los usuarios. Adicionalmente, se recopiló información sobre la cantidad de operaciones exitosas por *gateway*, tipo de documento de pago, la primera y última fecha de pago, la identificación de

operaciones exitosas realizadas durante fines de semana o entre semana, la franja horaria, número de operaciones exitosas pagadas en la primera quincena y segunda quincena y cantidad de pagos realizados por año.

En cuanto a los importes, con el objetivo de mitigar la influencia de la inflación en la moneda argentina, se optó por convertirlos a dólares estadounidenses, utilizando la tasa de cambio conocida como dólar contado con liquidación. A partir de esta información, se obtuvo para cada usuario la suma total de transacciones, así como el promedio del importe cobrado y la comisión cancelada por cada transacción.

Luego, se procedió a unificar los datos básicos o demográficos de la tabla de usuarios. Para cada usuario, se incorporaron detalles como la cantidad de planes asociados, fecha de alta, tipo de persona, estado civil, fecha de nacimiento, sexo y provincia.

Durante este proceso, se observó que algunos usuarios realizaban transacciones a través de la *Fintech*, pero no contaban con un registro de alta, lo que resultaba en la falta de información demográfica para dichos usuarios. Por lo tanto, se tomó la decisión de eliminar estos casos de la base de datos. Este ajuste afectó al 8% de los registros, un porcentaje considerado no significativo para el análisis, dado que los usuarios sin registro de alta no aportan información demográfica, lo que limita su contribución al objetivo del estudio. Además, al eliminar estos casos, se garantiza la coherencia y la calidad de los datos utilizados.

Posteriormente, se reevaluaron los valores faltantes en los datos recién incorporados y se identificó que algunas variables, como estado civil, fecha de nacimiento y sexo, presentaban una proporción de valores faltantes superior al 43%. En vista de esta situación, se optó por eliminar dichas variables, resultando en una base final que cuenta con **109.769 usuarios y con 41 variables.**

La base de datos creada mediante la consolidación de la información de las tablas de "user" y "transacciones" no incluye una variable que indique si un usuario se ha fugado o no. En otras palabras, esta información inicial no es suficiente para

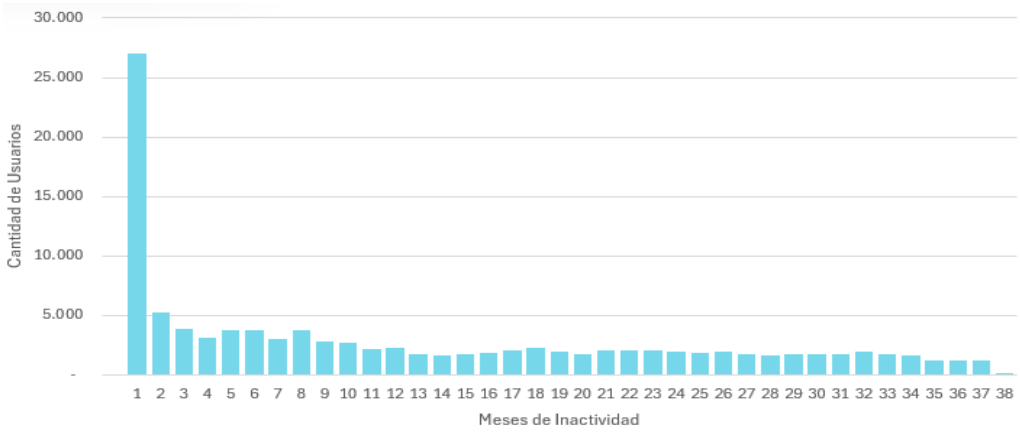
desarrollar un modelo de predicción de fuga de clasificación. Por esta razón, se ha utilizado esta información como un insumo para la construcción de una nueva variable, la cual está específicamente diseñada para ser aplicada en el modelo de predicción de la fuga de usuarios.

La construcción de la base de datos final sigue una metodología que, en una primera instancia, implementa un análisis de cohortes, dividiendo a los usuarios en 34 grupos distintos, agrupados según el año y mes de su primera transacción, desde noviembre de 2020 hasta noviembre de 2023. Es importante destacar que no se considerará el mes de diciembre de 2023 en este análisis, ya que los usuarios que realizan su primera transacción en diciembre no se podrá determinar su comportamiento en el siguiente mes. Por lo tanto, se excluyen del análisis los usuarios que transaccionan por primera vez en diciembre de 2023 y las cohortes se reducen a un total de **33 cohortes y con 107.623 usuarios.**

Después, se decidió elegir la última fecha en la que el usuario realizó un pago exitoso a través de los métodos de pagos ofrecido por la *Fintech* para calcular el período de tiempo transcurrido desde su última transacción hasta la fecha de estudio. Este análisis se presenta en un histograma que se muestra en la Figura 9. Se observa que el intervalo de meses de inactividad operativo es de 1 a 38 meses.

Figura 9

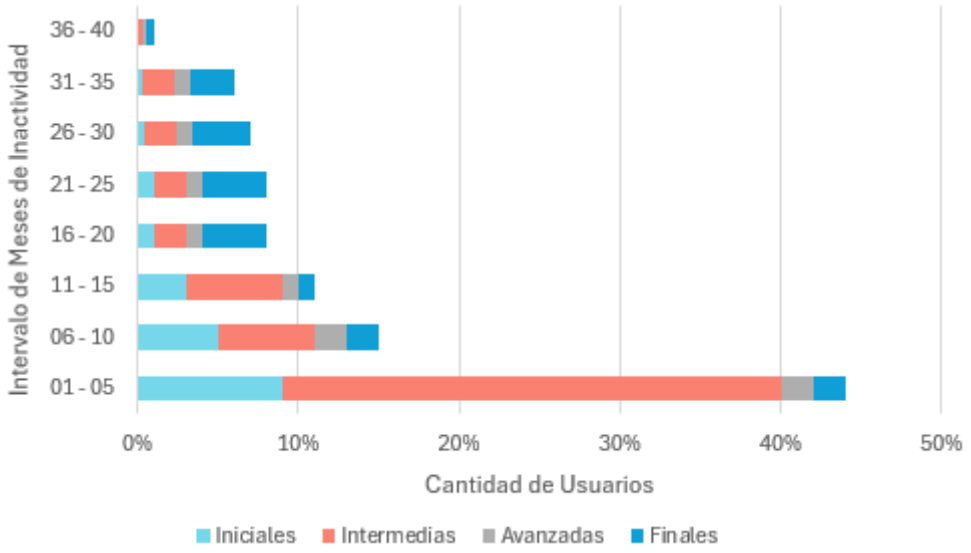
Histograma de meses de inactividad de usuarios operativos desde su última transacción registrada



Posteriormente, se realizó un análisis del último número de cuota pagado por cada usuario, clasificando las cuotas en cuatro etapas: inicial (cuotas 1 a 21), intermedia (cuotas 22 a 42), avanzada (cuotas 43 a 63) y final (cuotas 64 a 84). En la Figura 10 se muestra que, a partir del mes 20, se registran usuarios en las etapas avanzadas y finales de los planes.

Figura 10

Participación porcentual de cantidad de usuarios por meses de inactividad de usuarios operativos desde su última transacción registrada segmentado por etapa de plan.



Este grupo de usuarios con un periodo de inactividad prolongado superior a 16 meses generalmente se encuentra en las etapas finales de su plan de ahorro de autos, representando el 30% de los usuarios. Por lo tanto, se decidió no considerar a aquellos usuarios con más de 15 meses de inactividad en el análisis. Esta elección se justifica porque estos usuarios suelen estar en la etapa final de sus planes, lo cual podría distorsionar el análisis.

Además, para la predicción de fuga de usuarios, es crucial que los usuarios sean lo más actuales posible. Por ello, se ha decidido considerar únicamente a los usuarios con un máximo de 15 meses de inactividad. Esta decisión reduce la base de datos a 66.879 usuarios, excluyendo aquellos que no han registrado un pago de cuota

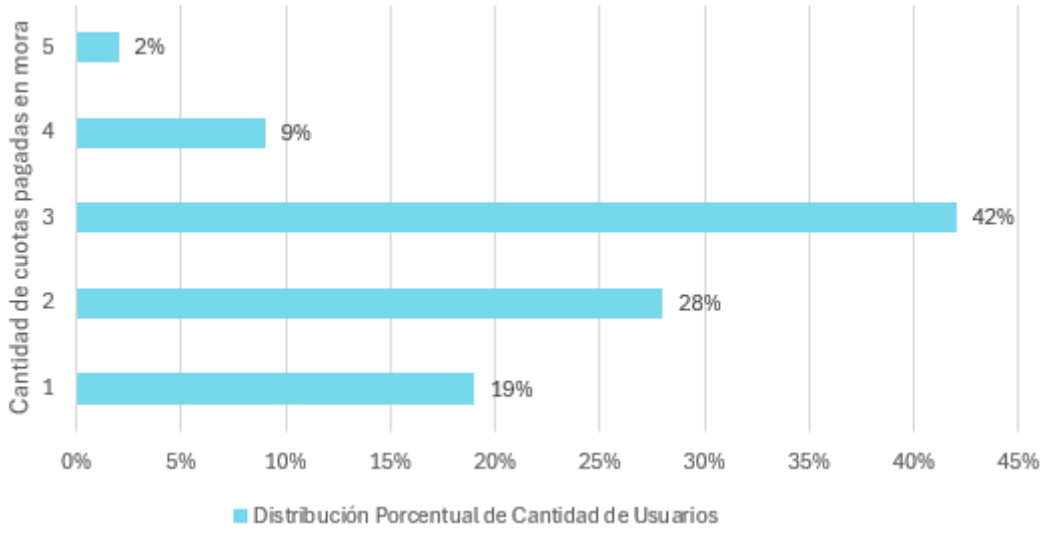
a través de los métodos de pago ofrecidos por la *Fintech* en ese periodo y usuarios que han pagado la última cuota (84).

Dentro de los contratos de suscripción de un plan de ahorro de auto, se menciona que, si un usuario acumula cinco cuotas impagas, el contrato se rescindirá automáticamente. Con esta premisa, se considera que los usuarios con 1 a 5 meses de inactividad, equivalentes a 5 cuotas, están dentro del periodo límite para cancelar sus cuotas. Los usuarios con más de 5 cuotas impagas son indicativos de haber cambiado su método de pago. La falta de información adicional sobre los planes nos impide saber si estos usuarios continúan pagando por otros métodos o si se dieron de baja del plan. Por tal razón, estos usuarios se identificarán como usuarios no retenidos.

De estos usuarios, se identificó que el 66% de los usuarios pagaron sus cuotas dentro de la fecha de vencimiento, el 27% pagaron con intereses, y un 7% realizaron pagos anticipados de cuotas. Entre los usuarios que pagaron con intereses, el 42% liquidó hasta tres cuotas acumuladas, mientras que el 28% pagó dos cuotas, como se observa en la Figura 11.

Figura 11

Distribución porcentual de cantidad de usuarios por cantidad de cuotas pagadas en mora



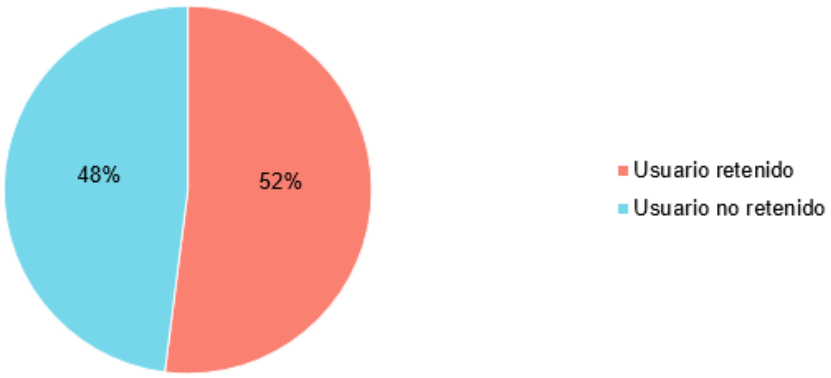
Dado que los usuarios estudiados están vinculados al pago de planes de autos con la expectativa de realizar pagos mensuales, se ha decidido clasificar como "no retenidos" a aquellos cuyo lapso desde su última transacción exitosa supera los 90 días. Este criterio se fundamenta en los análisis previos, que muestran que los usuarios tienen un límite de hasta 5 cuotas impagas y que la mayoría de los usuarios pagan cuotas acumuladas de hasta 3 cuotas.

Se realizó un análisis adicional en el que se seleccionaron los usuarios que se dieron de alta al menos un año antes de la última fecha de recolección de datos y se calculó su período de inactividad. Identificamos a aquellos que realizaron pagos después de un período de inactividad de más de 90 días. Los resultados mostraron que aproximadamente el 85% de estos usuarios realizaron pagos después de superar los 90 días de inactividad. Esto indica que, aunque algunos usuarios pagan sus cuotas con retraso e intereses adicionales, la mayoría lo hace dentro de un margen de 3 cuotas acumuladas, como se observó anteriormente.

Basándonos en lo expuesto anteriormente, hemos llegado a la conclusión de que el 52% de los usuarios en la base de datos son clasificados como "usuarios no retenidos", como se detalla en la Figura 12. Esta proporción señala un equilibrio en la base de datos, ya que se observa una similitud en la cantidad de observaciones en cada clase, lo cual resulta beneficioso para el proceso de modelado.

Figura 12

Distribución porcentual de usuarios por estado de retención



2.2.3. Ingeniería de atributos o construcción de nuevas variables

Al suscribirse a un plan de auto, los usuarios deben acceder a la plataforma para gestionar el pago mensual de su plan. Este proceso incluye varias operaciones manuales, como iniciar sesión en la plataforma, seleccionar el método de pago y completar la transacción según el método de pago elegido. Los ratios de transacciones exitosas/acreditadas y canceladas reflejan la facilidad o dificultad que enfrentan los usuarios en estos procesos. Por ejemplo, un alto ratio de transacciones exitosas indica una experiencia de usuario positiva, mientras que un alto ratio de transacciones fallidas o canceladas sugiere posibles problemas en la plataforma que podrían afectar la retención de usuarios.

Por tal razón, se procedió a la creación de variables para proporcionar una representación operativa del comportamiento de los usuarios. Entre estas variables destacan los ratios calculados de transacciones exitosas en relación con el total de transacciones por usuario, el ratio de transacciones por intentos de pago y el ratio por cancelaciones o transacciones abandonadas. Estas métricas son fundamentales para evaluar la experiencia del usuario con respecto a la plataforma diseñada por la *Fintech*.

Asimismo, se incorporaron dimensiones temporales, como la proporción de usuarios que efectuaron pagos en la primera o segunda quincena, junto con la antigüedad del usuario y el tiempo transcurrido desde su alta hasta la primera fecha de pago. Adicionalmente, se generaron indicadores financieros como el importe transaccionado en dólares por plan de usuario, ratios de transacciones pagadas por *gateway* y el promedio del ratio de Comisión, que evalúa la relación entre el total transaccionado y la comisión generada para la *Fintech*.

Después de llevar a cabo la depuración de la información, se procedió al análisis de la correlación entre las variables explicativas, y los resultados se presentan en la Figura 13. En este análisis, se destacaron las variables principales para una visualización más clara. Se observaron relaciones notables, especialmente con la

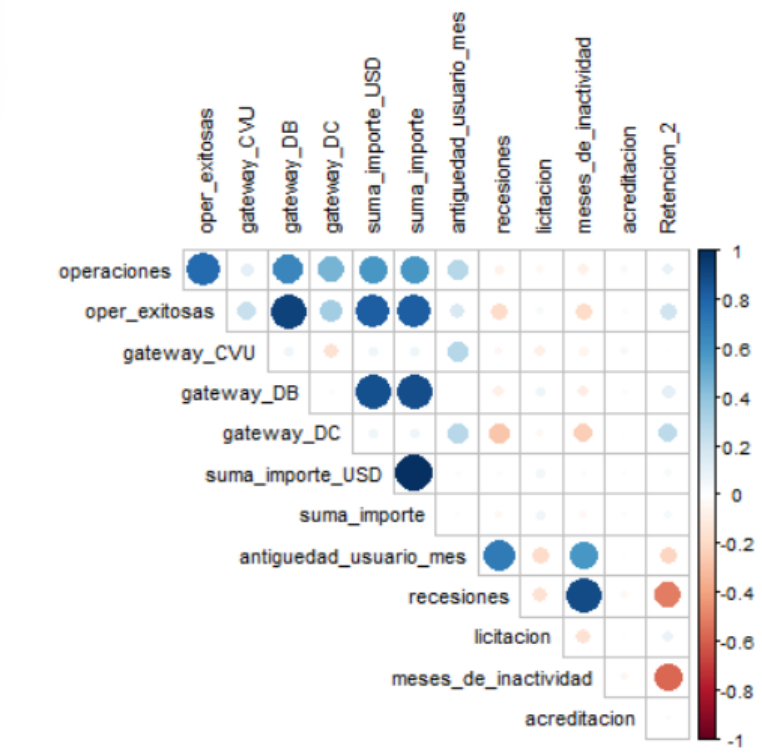
variable dependiente, que en este caso es el estado de retención, donde un usuario no retenido se representa con el valor 0.

A partir de este análisis, se puede concluir que existen correlaciones positivas significativas con las operaciones exitosas, especialmente cuando se trata de realizar pagos con DEBIN y el monto transaccionado. Un aumento en estas variables se asocia con usuarios retenidos.

Como era de esperar, se observó que variables como la cantidad de meses sin pagar y los meses de inactividad muestran una correlación positiva con la categorización de usuarios no retenidos. En otras palabras, un aumento en estos indicadores está vinculado a la probabilidad de que un usuario no sea retenido.

Figura 13

Matriz de correlación de variables



2.2.4. Análisis de usuarios no retenido con variables explicativas

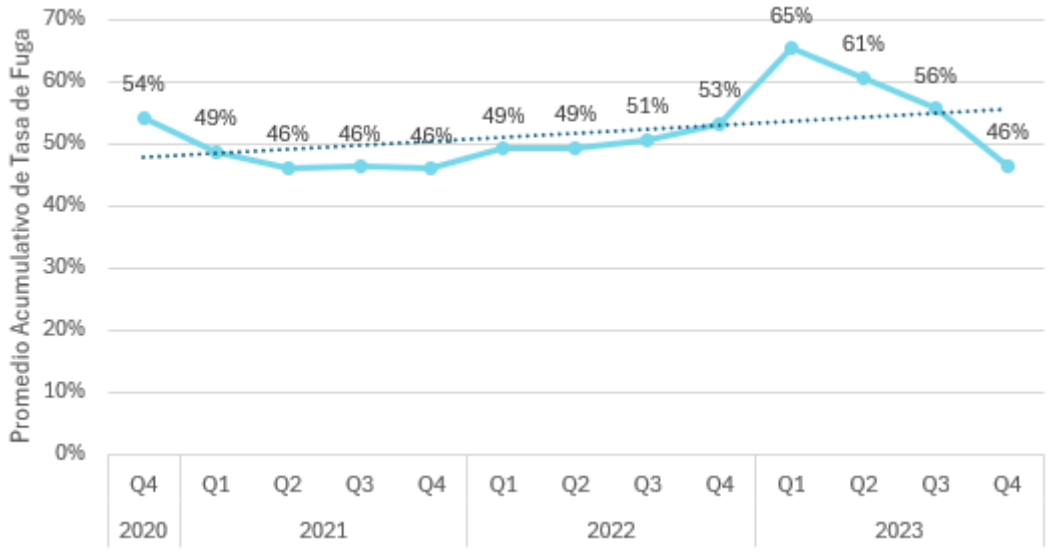
Es crucial llevar a cabo un análisis que relacione las variables de la base de datos con la variable dependiente, que en este caso es la fuga de usuarios. Para comenzar, es fundamental visualizar la evolución de la tasa de fuga trimestral a lo

largo del período analizado. En la Figura 14 se presenta la tasa acumulativa de fuga hasta el punto de corte. Inicialmente, en el primer trimestre de 2020, la *Fintech* enfrentó una tasa de fuga del 54%, lo que representa a los usuarios que no pagaron después de 90 días de su última transacción. Sin embargo, desde entonces, esta tendencia ha ido en aumento, aunque en el último trimestre se observa una disminución en la tasa de fuga, llegando al 46%.

En el primer trimestre de 2023, se observa un aumento significativo en la tasa de fuga, alcanzando el 65%. Esta tasa acumulativa incluye a los usuarios desde el inicio hasta el primer trimestre de 2023. Esta tendencia se atribuye a la expansión de opciones para el pago mensual de cuotas proporcionadas por la empresa automotriz. La introducción de métodos de pago alternativos, distintos a los ofrecidos por la *Fintech*, implica que esta última no actuará como intermediario en la cobranza, lo que conlleva una pérdida de ingresos para la *Fintech*.

Figura 14

Evolución de la tasa de fuga acumulada trimestral



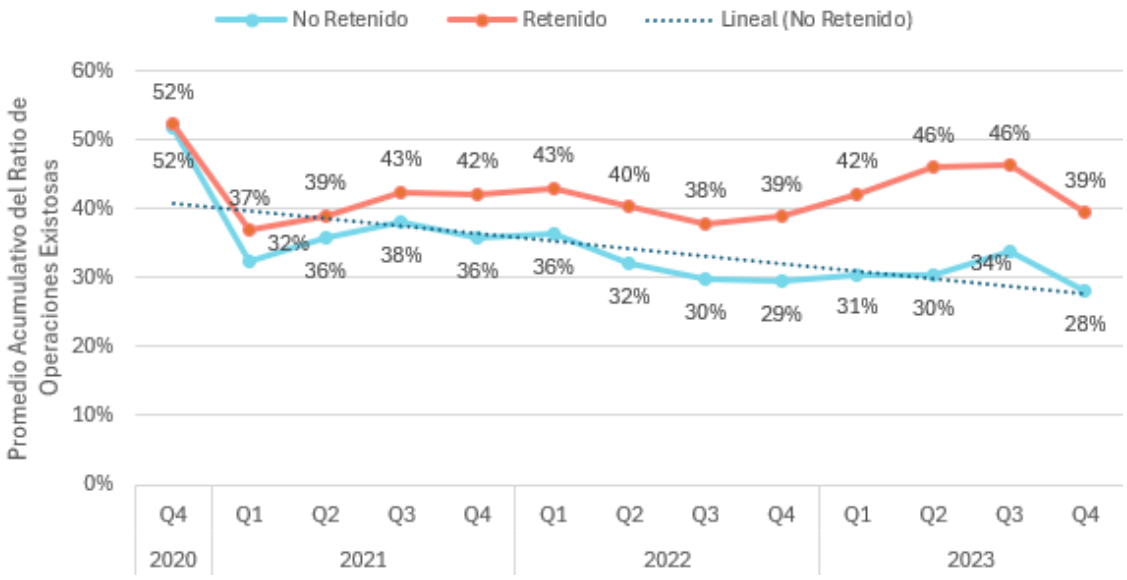
Posteriormente, se llevó a cabo un análisis sobre cómo influye el ratio de operaciones exitosas en la retención de usuarios, según la Figura 15. En este contexto, se observa que a medida que el ratio de éxito disminuye, la proporción de

usuarios no retenidos tiende a aumentar. En contraste, a medida que el ratio de éxito se incrementa, se evidencia una mayor proporción de usuarios retenidos.

Este hallazgo indica que la experiencia del usuario en la plataforma puede tener un impacto significativo en su decisión de dejar de utilizarla. Un ratio más elevado de operaciones exitosas podría estar asociado con una experiencia positiva del usuario, contribuyendo a la retención, mientras que un bajo ratio podría sugerir inconvenientes o insatisfacciones que llevan a la fuga de usuarios.

Figura 15

Evolución del promedio del ratio de operaciones exitosas mensual



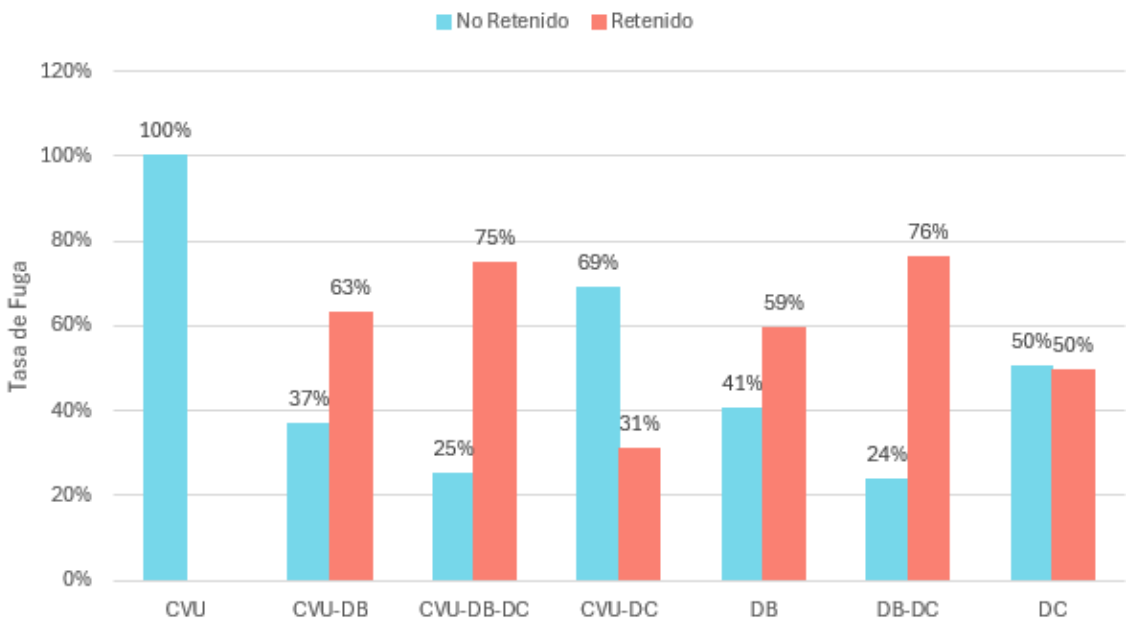
Otro aspecto relevante para el análisis es la segmentación de los usuarios retenidos según el método de pago utilizado según se muestra en la Figura 16. Se indica que el 100% de usuarios que pagaron exclusivamente mediante CVU se consideran como usuarios no retenidos. Este resultado tiene sentido, dado que la *Fintech* dio de baja este servicio en marzo, lo que evidencia que existen usuarios se fugaron. No obstante, es notable que parte de estos usuarios que inicialmente usaron CVU optaron por migrar a otros métodos de pago, como DB y DC.

Adicionalmente, se observa que el 41% de los usuarios que realizan transacciones únicamente con DB son considerados como no retenidos. En contraste, entre los usuarios que combinan los métodos de pago DB y DC, solo el 24% se clasifica como

no retenidos. Este patrón sugiere que la elección del método de pago puede influir en la retención de los usuarios.

Figura 16

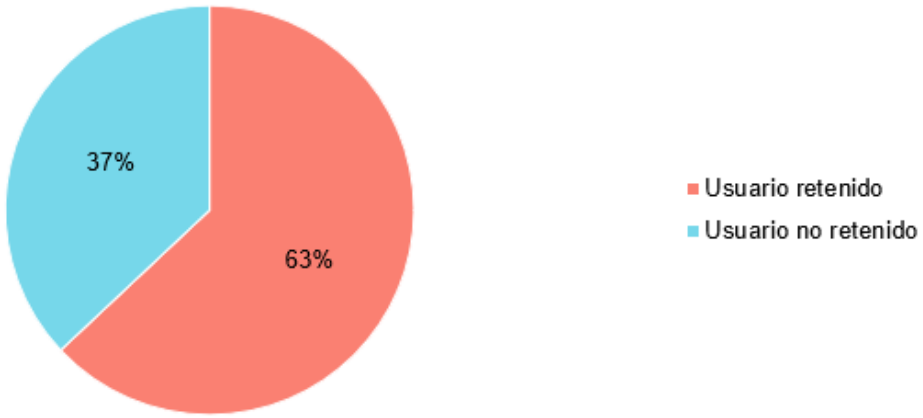
Tasa de fuga segmentada por método de pago utilizado por el usuario.



Por último, el equipo de soporte al cliente facilitó la información de los usuarios que presentaron reclamos debido a dificultades en el pago a través de los métodos de planes ofrecidos por la *Fintech*, donde se revela que un 37% de ellos fueron categorizados como "usuarios no retenidos", como se observa en la Figura 17. Este hallazgo sugiere que los problemas en la prestación del servicio pueden tener un impacto significativo en la probabilidad de que un usuario vuelva a utilizar los servicios de la *Fintech*.

Figura 17

Distribución porcentual del estado de retención por los usuarios que realizaron un reclamo de servicio.



3. Metodología

Después de analizar detalladamente el comportamiento de los usuarios en diversas variables explicativas, este capítulo se centra en el desarrollo de tres técnicas de aprendizaje automático para la predicción de fuga de usuarios. Además, se detallarán y establecerán las métricas de desempeño que se utilizarán para evaluar la eficacia de los modelos. También se explorarán procesos para optimizar los hiperparámetros y mejorar el rendimiento general del modelo.

Los modelos utilizados para predecir la fuga de usuarios de la *Fintech* para el presente trabajo son la regresión logística, arboles de decisión y XGBoost. En la evaluación de estos modelos, se empleará un conjunto uniforme de variables con el objetivo de probar y comparar diversas configuraciones. A continuación, detallo las principales variables que serán consideradas:

Variable Dependiente:

- Estado de retención de usuarios (Retenido / No Retenido)

Variables Independientes:

Variable	Tipo	Descripción
Iduser	Numérica	Código único del usuario, asignado de manera cronológica

*Gateway*_CVU	Numérica	Cantidad de operaciones exitosas por CVU
*Gateway*_DB	Numérica	Cantidad de operaciones exitosas por DB
*Gateway*_DC	Numérica	Cantidad de operaciones exitosas por DC
MétodoPago	Categórica	Combinaciones de *gateways* utilizados por el usuario
ratio_oper_exitosas	Numérica	Ratio de operaciones exitosas del usuario, calculado como la división entre la cantidad de operaciones exitosas y el total de operaciones registradas del usuario
TipoPersona	Categórica	Razón social del usuario (Física o Jurídica)
Provincia	Categórica	Provincia de residencia del usuario
Grupo_edad	Categórica	Intervalos de edad en los que se encuentra el usuario (18-35, 36-50, 51-65, 66-90)
Monto USD	Numérica	Monto acumulado de las operaciones exitosas, expresado en dólares estadounidenses
desc_planes	Categórica	Categoría del usuario de acuerdo con la cantidad de planes
Antigüedad_usuarios_mes	Numérica	Cantidad de meses de antigüedad del usuario
Pagos_año	Numérica	Cantidad de operaciones exitosas por año
typcbte_alicuota	Numérica	Cantidad de operaciones exitosas de alícuota (porción fija de los gastos de administración del plan de ahorro)
typcbte_cuotas	Numérica	Cantidad de operaciones exitosas de cuotas (pagos mensuales que los usuarios realizan como parte del plan de ahorro)
typcbte_anticipo_cuotas	Numérica	Cantidad de operaciones exitosas de

		anticipo de cuotas (pagos adelantados de las cuotas mensuales del plan de ahorro)
typcbte_cuotas_mora	Numérica	Cantidad de operaciones exitosas de cuotas_mora (pagos realizados para ponerse al día con cuotas atrasadas)
typcbte_licitacion	Numérica	Cantidad de operaciones exitosas de licitación (pagos realizados para adelantar la adjudicación del auto mediante subasta)
Licitación	Binaria	Valor 1 si el usuario realizó una licitación, 0 si no
cuotas_mora	Binaria	Valor 1 si el usuario realizó pagos en mora, 0 si no
Acreditación	Binaria	Valor 1 si el usuario realizó una acreditación errónea y presentó un reclamo, 0 si no
R_A	Numérica	Ratio entre la cantidad de cuotas impagas y la cantidad de meses de antigüedad del usuario

Es fundamental, antes de ejecutar los modelos de clasificación, dividir la base de datos en dos grupos con atributos comparables para validar de forma efectiva el rendimiento de los modelos propuestos. En este proceso, se selecciona el 80% del conjunto de datos para ser utilizado como conjunto de entrenamiento de los modelos, mientras que se reserva el 20% restante para llevar a cabo el testeo.

Con el fin de preservar la temporalidad de los datos y asegurar que las observaciones en el conjunto de validación reflejen de manera precisa las condiciones más recientes, se establece la premisa de que los datos de testeo representen la información más actualizada. Esta consideración es esencial para garantizar que las predicciones se realicen en un entorno que refleje de manera adecuada las condiciones actuales.

Se realizó una verificación adicional, donde se buscó garantizar que la división de los datos mantenga la misma distribución de observaciones etiquetadas como usuarios retenidos y no retenidos en ambos conjuntos, como se detalla en la Tabla 3. Este enfoque evita desbalances que podrían afectar los resultados, permitiendo así una evaluación más precisa del rendimiento de los modelos en ambas clases.

Tabla 3

Cantidad porcentual de observaciones para entrenamiento y testeo

Variable	Estado de Retención	% de Estado de Retención	Cant Obs
Entrenamiento	Retenido	52%	27.953
Entrenamiento	No Retenido	48%	25.550
Testeo	Retenido	52%	6.988
Testeo	No Retenido	48%	6.388

3.1. Técnicas de *machine learning* en la predicción de fuga

3.1.1. Regresión logística

La regresión logística binaria se emplea para determinar la relación entre una variable dependiente dicotómica y una o más variables independientes o explicativas, que pueden ser tanto cualitativas como cuantitativas. Su propósito es obtener una estimación ajustada de la probabilidad de que ocurra un evento basándose en estas variables independientes (Pérez, Pino, Ballester, & Moreno, 2010).

La variable dependiente es categórica y presenta dos categorías posibles, comúnmente etiquetadas como 0 y 1. En el contexto de la predicción de fuga de usuarios, esta técnica se emplea para modelar la relación entre un conjunto de variables explicativas y la probabilidad de que un usuario sea clasificado como retenido (1) o no retenido (0) (Hosmer, Lemeshow, & Sturdivant, 2013).

La regresión logística utiliza una función logística, también conocida como función sigmoide, para transformar la combinación lineal de las variables explicativas en una probabilidad entre 0 y 1. La función sigmoide se define como:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

Donde:

$P(Y = 1)$, es la probabilidad de que la variable dependiente sea 1

e, es la base del logaritmo natural

$\beta_0, \beta_1, \beta_n$, son los coeficientes del modelo

x_1, x_2, x_n son las variables explicativas

La aplicación de la regresión logística arroja resultados que se destacan por su interpretación intuitiva y explicativa, facilitando la comprensión de cómo cada variable influye en la probabilidad de que un usuario sea catalogado como retenido o no retenido. Este atributo resulta particularmente valioso en la toma de decisiones informadas. Además, su eficiencia al trabajar con variables explicativas binarias y categóricas la convierte en una herramienta sumamente versátil. Esta técnica no se limita únicamente a modelar relaciones entre variables, sino que también demuestra habilidad para manejar situaciones en las que las categorías no son linealmente separables, mejorando así su utilidad en diversos escenarios (Tabachnick & Fidell, 2019).

De acuerdo con Ortuño (2022), se enlistan los supuestos que se deben verificar para aplicar un modelo de regresión logística. Estos incluyen la necesidad de que la variable a predecir sea binaria, la linealidad en la relación entre el logit o log-odds de la variable respuesta y cada variable predictora (verificado únicamente para las variables numéricas continuas), la independencia entre las observaciones, y la ausencia de multicolinealidad entre las variables predictoras.

Para asegurar de que los supuestos de la regresión logística se cumplieran en el trabajo, se realizó varias verificaciones. Se examinó gráficamente la relación entre el logit de la variable respuesta y cada variable predictora continua mediante gráficos de dispersión y de tendencia. Se reviso la estructura de los datos para garantizar que no provinieran de mediciones repetidas del mismo individuo o estuvieran relacionadas de alguna manera. Además, se realizó pruebas de correlación para evaluar la presencia de multicolinealidad entre las variables predictoras.

3.1.2. Árboles de decisión

Un árbol de decisión es un modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente. El modelo en sí mismo comprende una serie de decisiones lógicas, similares a un diagrama de flujo, con nodos de decisión que indican una decisión sobre un atributo. Estos se dividen en ramas que indican las elecciones de la decisión. El árbol termina con nodos de hoja o *leaf nodes* (también conocidos como nodos terminales) que denotan el resultado de seguir una combinación de decisiones (Ferrero, R., 2020).

Según Quinlan (1986), para construir un árbol de decisión, se utiliza un criterio de impureza para evaluar la calidad de las divisiones en cada nodo. Este criterio mide qué tan "puras" son las divisiones en función de las clases de la variable respuesta. Los dos criterios de impureza más comunes son el índice Gini y la entropía.

El índice Gini mide la impureza de un conjunto de datos y se define como:

$$Gini(t) = 1 - \Sigma_{i=1}^{c} p_i^2$$

Donde t es el nodo, c es el número de clases, y p_i es la proporción de instancias de la clase i en el nodo t.

La entropía, por otro lado, se define como:

$$Entropía\ (t) = -\sum_{i=1}^{c} p_i log_2(p_i)$$

Donde p_i es la proporción de instancias de la clase i en el nodo t.

En este trabajo, utilizaremos el índice Gini como criterio de impureza para la construcción del árbol de decisión. Este índice nos permitirá evaluar la calidad de las divisiones y optimizar el modelo para la predicción de la fuga de usuarios.

Los árboles de decisión ofrecen varias ventajas en un modelo de clasificación. En primer lugar, son interpretables y fácilmente visualizables. Además, permiten la identificación de las características más relevantes para la toma de decisiones. Breiman (1984) señalan que los árboles pueden manejar datos categóricos y numéricos sin la necesidad de preprocesamiento adicional. Además, son robustos frente a datos ruidosos y pueden capturar patrones no lineales en los datos.

Sin embargo, presentan desventajas. Su simplicidad puede traducirse en resultados menos efectivos en comparación con otros modelos más complejos, y su falta de robustez puede ser un problema. Generar árboles muy profundos puede hacer que pequeños cambios en el conjunto de datos resulten en cambios significativos en la salida del modelo, lo que puede llevar a problemas de sobreajuste.

3.1.3. XGBoost

XGBoost, desarrollado por Chen y Guestrin (2016), es una potente extensión de árboles de decisión que ha ganado gran popularidad en tareas de predicción y clasificación. Este modelo combina la simplicidad de los árboles de decisión con la capacidad de mejorar el rendimiento mediante técnicas avanzadas de regularización y optimización.

La fórmula básica para la predicción en XGBoost se define como:

$$\hat{y}_i = \Sigma_{k=1}^{K} f_k(x_i)$$

Donde $\hat{y}_i$ es la predicción para la instancia i. K es el número de árboles, y f_k es la función del K-ésimo árbol. El modelo busca minimizar una función de pérdida regularizada, que incluye términos de pérdida y penalización para evitar sobreajuste.

XGBoost ofrece varias ventajas significativas. Por un lado, es altamente eficiente y escalable, lo que permite manejar grandes conjuntos de datos. Además, aborda el problema de sobreajuste mediante técnicas como la poda de árboles y la penalización de complejidad, logrando modelos más robustos. XGBoost también destaca por su capacidad para manejar características faltantes sin necesidad de imputación previa.

XGBoost no está exento de desafíos. La configuración óptima de hiperparámetros puede requerir un ajuste cuidadoso, y el modelo podría ser más propenso al sobreajuste si no se controla adecuadamente. Sin embargo, su flexibilidad y rendimiento general lo convierten en una elección popular en una variedad de aplicaciones de aprendizaje automático (Ramírez, 2024).

3.2. Evaluación de modelos

3.2.1. Métrica de desempeño de modelos

Para evaluar el desempeño de los modelos de clasificación utilizados en este trabajo, se emplean varias métricas clave. La exactitud (*Accuracy*) se define como la proporción de predicciones correctas realizadas por el modelo sobre el total de casos evaluados, calculada como la suma de verdaderos positivos y verdaderos negativos dividida por el total de casos. En el contexto de este trabajo, los verdaderos positivos se refieren a los usuarios que predijimos correctamente que no volverían y efectivamente no volvieron, mientras que los falsos negativos son aquellos usuarios que predijimos que volverían, pero en realidad no lo hicieron. La precisión (*Precision*) mide la proporción de verdaderos positivos entre todos los casos que el modelo ha clasificado como positivos, es decir, usuarios que predijimos que no volverían y realmente no volvieron, sobre el total de usuarios predichos como que no volverían (Según Torres 2023).

Adicionalmente, la recuperación o sensibilidad (*Recall*) se calcula como la proporción de verdaderos positivos sobre todos los casos que realmente son positivos,

proporcionando una medida de la capacidad del modelo para identificar correctamente los usuarios que no volverán. La especificidad evalúa la proporción de verdaderos negativos, aquellos que correctamente predijimos que volverían, sobre el total de casos que realmente deberían ser negativos. Por otro lado, la tasa de falsos positivos (también conocida como tasa de falsa alarma) se calcula como la proporción de falsos positivos sobre la suma de falsos positivos y verdaderos negativos, indicando la frecuencia con la que el modelo identifica incorrectamente un usuario como que no volverá cuando en realidad volverá.

Para visualizar y comparar el rendimiento general de los modelos, se utiliza la Curva ROC (Receiver Operating Characteristic), que es una representación gráfica de la relación entre los falsos positivos y los verdaderos positivos a través de diferentes umbrales de clasificación.

El umbral de corte en los modelos de clasificación binaria es esencial para determinar cómo se clasifican los usuarios entre los que retornan y los que no. Aunque comúnmente se utiliza un umbral predeterminado de 0.5, este valor puede ajustarse para optimizar el rendimiento del modelo. La elección del umbral impacta directamente en métricas clave como *precision* y *recall*. Por ejemplo, un umbral más bajo puede mejorar la sensibilidad del modelo para detectar usuarios que no retornan, pero a expensas de una menor precisión. Por otro lado, un umbral más alto puede mejorar la precisión, pero podría perderse en la detección de usuarios no retornados (James, Witten, Hastie, & Tibshirani, 2021).

Determinar el umbral óptimo implica evaluar las preferencias del negocio y las implicaciones de costos asociadas con los falsos positivos y falsos negativos. Es crucial realizar análisis para comparar cómo varía el rendimiento del modelo con diferentes umbrales de corte. Esto garantiza que se seleccione el modelo más adecuado según los objetivos específicos de *precision* y *recall*. En la práctica, esta elección estratégica del umbral no solo mejora la capacidad del modelo para predecir

la retención de usuarios, sino que también optimiza su utilidad y efectividad en escenarios empresariales reales.

La AUC (Área Bajo la Curva) de esta curva es un indicador crucial, se calcula como el área bajo la curva ROC ya que mide la capacidad predictiva del modelo, siendo una métrica estándar en la evaluación de modelos de clasificación. Un AUC más alto indica un mejor rendimiento del modelo en términos de su capacidad para distinguir entre usuarios que volverán y aquellos que no lo harán.

3.2.2. Optimización de hiperparámetros

La optimización de hiperparámetros es un componente esencial en el desarrollo de modelos de predicción de fuga de usuarios, donde se busca mejorar la capacidad del modelo para identificar y anticipar la pérdida de los usuarios. Este proceso implica ajustar los parámetros del modelo para optimizar su rendimiento, lo que se traduce en una mayor precisión en la identificación de usuarios propensos a dejar de utilizar la plataforma (Chauhan 2020).

Esto se refiere a la búsqueda de la combinación más efectiva de valores para los parámetros de un modelo. En el contexto de la predicción de fuga, esto implica encontrar la configuración óptima que maximice la capacidad del modelo para detectar patrones relevantes relacionados con la fuga de usuarios. La importancia de este proceso radica en su impacto directo en la eficacia del modelo, permitiendo una adaptación más precisa a las características específicas del conjunto de datos y mejorando la capacidad predictiva del modelo (Carrasco, R. A., Bueno, I., & Montero, J.-M. 2023).

Hiperparámetros en modelos de árboles de decisión:

En modelos de árboles de decisión, la optimización se centra en parámetros clave como *maxdepth*, *minsplit*, y *minbucket*. El parámetro *maxdepth* determina la profundidad máxima del árbol, controlando su complejidad. *Minsplit* y *minbucket* regulan la creación de nodos dividiendo el árbol. Ajustar estos hiperparámetros

permite encontrar un equilibrio entre la capacidad del modelo para adaptarse a los datos y evitar el sobreajuste.

Hiperparámetros en XGBoost:

En el caso de XGBoost, una técnica popular es la búsqueda en rejilla, donde se define un conjunto de valores posibles para los hiperparámetros en un espacio predeterminado. Para este trabajo, se utiliza el programa R-Studio, el cual ofrece funciones como "*Caret*" y "*random_grid*" para explorar eficientemente estas combinaciones, lo que permite seleccionar valores que maximizan las métricas de evaluación. Ajustar la tasa de aprendizaje, la profundidad máxima del árbol, y otros parámetros específicos de XGBoost puede potenciar su capacidad predictiva en la identificación de la fuga de usuarios. Este enfoque brinda flexibilidad y precisión al adaptar el modelo a las complejidades específicas del conjunto de datos.

4. Resultados

Una vez establecidas las técnicas de aprendizaje automático y definidas las métricas de desempeño de los modelos, este capítulo presenta los resultados obtenidos al aplicar dichos modelos a la base de datos construida. A partir de estos resultados, se seleccionará el mejor modelo y se llevará a cabo un análisis de la importancia de las variables. Posteriormente, se realizará un experimento para evaluar el impacto financiero de este modelo predictivo para el modelo de negocio de la *Fintech*.

4.1. Elección de modelo de clasificación

Como se mencionó en el capítulo anterior, en este trabajo se entrenaron tres modelos diferentes utilizando técnicas avanzadas de aprendizaje automático, los cuales fueron optimizados mediante ajustes de hiperparámetros. Los resultados

obtenidos de este proceso revelaron aspectos significativos sobre el comportamiento de los usuarios en la plataforma de la *Fintech*.

Tabla 4

Resultados de modelos de clasificación

Modelo	AUC	*Accuracy*	*Precision*	*Recall*
Regresión logística	0.9639430857	0.9013905502	0.9059143635	0.9052661706
Arboles de decisión	0.9179199856	0.8756728469	0.9167318829	0.8381511162
Árboles de decisión con optimización de hiperparámetros	0.968409706	0.9028857656	0.923477743	0.8876645678
XGBoost	0.9863536637	0.8774671053	0.9942759015	0.8129168129
XGBoost con optimización de hiperparámetros	0.9811289902	0.9186602871	0.9464796795	0.9025655022

Se comenzó el entrenamiento con el modelo de regresión logística, seleccionado por su simplicidad y la configuración predeterminada de sus hiperparámetros. Este modelo, al asumir que los datos son linealmente separables, tiende a tener un rendimiento inferior en problemas de clasificación en comparación con los modelos de árboles de decisión más complejos. Como se evidencia en la tabla 4, la métrica de AUC registrada fue de 0.9639. Posteriormente, se procedió a implementar un modelo de árboles de decisión, manteniendo una metodología de entrenamiento y testeo similar. En este caso, la métrica de AUC alcanzó un valor de 0.9179.

Para mejorar el rendimiento del modelo, se realizaron intentos de ajuste de hiperparámetros, explorando diversas combinaciones que involucraron modificaciones en parámetros como *maxdepth*, *minsplit* y *minbucket*. Tras ejecutar nuevamente el modelo con estas nuevas configuraciones, se logró obtener una mejora en la métrica

de AUC, alcanzando un valor de 0.9684. Este resultado reflejó una leve superación en el rendimiento respecto al modelo de regresión inicialmente evaluado.

Finalmente, se procedió a entrenar un modelo de XGBoost, utilizando la misma base de datos, pero previamente aplicando el proceso de *one hot encoding*[3] para manejar las variables categóricas. Este método, ampliamente utilizado en el tratamiento de variables categóricas, permitió mejorar la capacidad predictiva del modelo. Tras la ejecución, se obtuvo un valor de AUC de 0.9863, indicando un rendimiento prometedor. Además, se realizaron intentos de optimización de los hiperparámetros del modelo XGBoost, lo que resultó en un desempeño superior a todos los modelos anteriores.

Como se ha evidenciado en los modelos detallados anteriormente, los valores de AUC son notablemente elevados, lo que refleja un rendimiento sólido de los modelos. Sin embargo, con el fin de mitigar el riesgo de sobreajuste sobre el conjunto de validación, se optó por realizar una validación cruzada utilizando un nuevo conjunto de datos que presentaba las mismas variables y estructuras, pero que no habían sido utilizadas previamente en el entrenamiento ni testeo. En este experimento, nos centramos específicamente en los dos modelos más prometedores. Como se muestra en la Tabla 5, si bien el AUC experimentó una ligera disminución, las métricas resultantes aún se mantienen en niveles altos.

Tabla 5

Resultados de modelos de clasificación con validación cruzada

Modelo	AUC	*Accuracy*	*Precision*	*Recall*
Árboles de decisión con optimización de hiperparámetros	0.9290732	0.9292763	0.9312018	0.9336005
XGBoost con optimización de hiperparámetros	0.9879417	0.7978469	0.7211896	0.9994276

[3] Crea una columna binaria para cada nivel de categoría y devuelve una matriz con estas representaciones binarias.

Al comparar estos dos modelos bajo la validación cruzada, el XGBoost con Hiperparámetro muestra un rendimiento excepcional, con un AUC de 0.9879. Esta métrica resalta la capacidad del modelo para distinguir entre las clases, lo que sugiere una alta precisión en sus predicciones. Sin embargo, al examinar las métricas de *precision* y *recall*, se observa que, aunque la precisión es relativamente alta (0.7212), el *recall* alcanza un valor extremadamente elevado de 0.9994. Esto indica que el modelo XGBoost tiene una capacidad para identificar los verdaderos positivos, pero esto se logra a expensas de un aumento en los falsos positivos.

En contraste, el modelo de árboles de decisión con hiperparámetros ajustados exhibe un AUC ligeramente más bajo, alcanzando 0.9291. Sin embargo, las métricas de *precision* y *recall* son más equilibradas en comparación con el modelo XGBoost, registrando valores de 0.9312 y 0.9336 respectivamente. Esto sugiere que el modelo es más conservador en sus predicciones, lo que resulta en una menor tasa de falsos positivos, aunque también implica una menor tasa de verdaderos positivos.

Con base en estos resultados, hemos optado por seleccionar el modelo de XGBoost con hiperparámetros y validación cruzada debido a su destacado rendimiento. Sin embargo, para equilibrar las métricas de *precision* y *recall*, se implementarán mejoras considerando los costos y beneficios asociados. Estas mejoras se detallarán en el próximo capítulo de análisis financiero, donde se evaluará el impacto económico

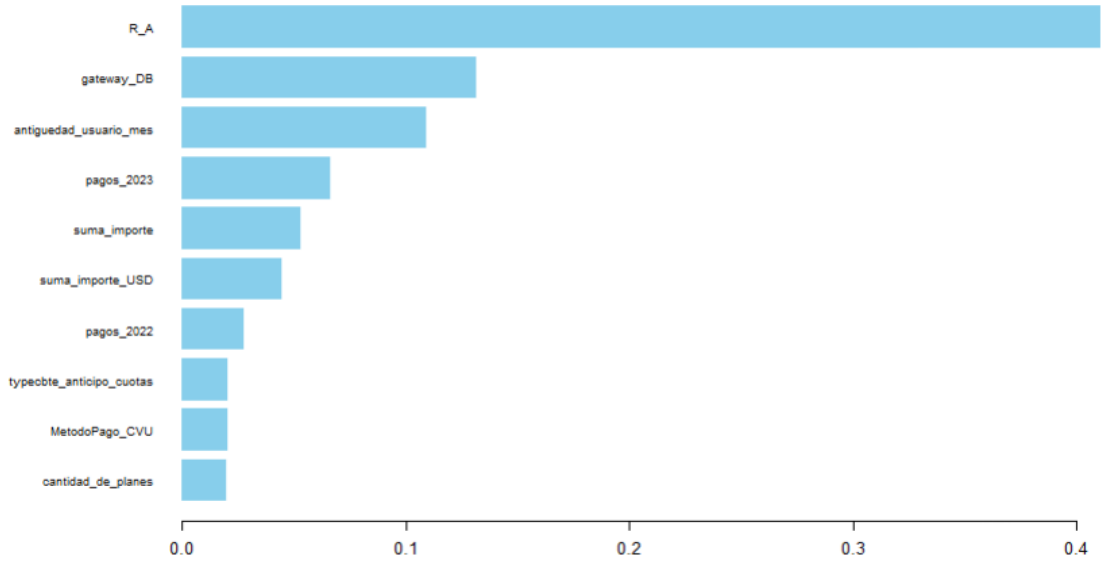
4.2. Análisis de la importancia de variables

Dado que hemos seleccionado el modelo XGBoost debido a su mejor rendimiento, ahora podemos examinar la importancia de cada variable durante su ejecución para clasificar. Esta importancia se relaciona con la capacidad explicativa de las variables sobre la varianza de las observaciones. En otras palabras, las variables más importantes tienen un mayor poder para distinguir entre los distintos comportamientos de los usuarios, ya sean retenidos o no retenidos. La Figura 18 presenta un *ranking* de

las diez variables más relevantes para el modelo, lo que nos proporciona información valiosa sobre los factores que influyen significativamente en la retención de usuarios.

Figura 18

Importancia de variables



Entre las variables más destacadas en términos de importancia se encuentran R_A y *gateway*_DB, así como la antigüedad del usuario. Específicamente, la variable R_A sobresale por representar el ratio entre la cantidad de cuotas que un usuario ha dejado de pagar y su antigüedad, lo cual proporciona una perspectiva clara del estado financiero del usuario. La variable *gateway*_DB ocupa el segundo lugar en importancia, indicando que la frecuencia de pagos realizados por DEBIN mediante este método puede influir en el comportamiento del usuario, sugiriendo una mayor recurrencia en ciertos casos.

En tercer lugar, en términos de importancia se encuentra la variable que indica la antigüedad del usuario, la cual está relacionada con la frecuencia de pagos realizados en el año 2023.

Además, otra variable relevante es el método de pago por CVU, lo cual tiene sentido en el contexto de la fuga de usuarios, especialmente cuando dicho servicio se dio de baja. Aunque incluir CVU como variable predictora puede parecer controvertido,

su significancia estadística dentro del modelo de predicción justifica su inclusión. A pesar de que CVU podría no ser relevante en el futuro, actualmente contribuye a un modelo más robusto y preciso. Esta inclusión permite capturar mejor las tendencias y comportamientos actuales, proporcionando una base sólida para desarrollar estrategias de retención más efectivas.

La importancia de estas variables se alinea con la lógica del negocio estudiado en relación con la fuga de usuarios. Por lo tanto, se interpreta como una confirmación de que el modelo está capturando de manera efectiva los comportamientos de negocio presentes en los datos, lo que fortalece su validez y utilidad para la predicción de la retención de usuarios.

4.3. Diseño de experimento

Con el objetivo de implementar los resultados obtenidos del modelo, es crucial considerar el contexto de la *Fintech*, actualmente se encuentra en un período de inicio-crecimiento y aún no ha implementado estrategias específicas de retención de usuarios. Por lo tanto, se planifica llevar a cabo un experimento para evaluar el impacto que el modelo puede tener en el negocio de la *Fintech*, específicamente en términos financieros. Este experimento será fundamental para entender cómo el modelo puede contribuir a mejorar la retención de usuarios y, generar un impacto positivo en la rentabilidad de la *Fintech*.

4.3.1. Estrategia propuesta de retención de usuarios

La estrategia diseñada para mejorar la retención de usuarios en la *Fintech* consiste en la implementación de una campaña mediante el uso del SMS-Bot de WhatsApp. Este enfoque se centra en resaltar y promover los diversos métodos de pago disponibles en la plataforma, con el fin de incentivar a los usuarios a utilizar estos servicios. El primer contacto se realizará dentro de los primeros 7 días, con el objetivo de comunicar los beneficios y ventajas de utilizar la plataforma para realizar sus pagos mensuales.

Esta estrategia se enfocará en usuarios específicos identificados con una alta probabilidad de no retención, evitando así costos innecesarios asociados con el envío indiscriminado de mensajes. El principal objetivo de esta iniciativa es evitar la disminución en los niveles de ingresos y, al mismo tiempo, mantener o mejorar los indicadores de retención de usuarios.

4.3.2. Métricas de éxito

En el contexto de este experimento, el éxito se define en función de la prevención de la disminución de los ingresos y la retención de usuarios. La principal métrica de éxito se centrará en comparar los ingresos, calculando el porcentaje de incremento en las ventas. Esta métrica se considera crucial, ya que nuestro objetivo principal es asegurar que los ingresos se mantengan o aumenten, en lugar de disminuir.

Además, se utilizará el porcentaje de crecimiento de los usuarios como una métrica de control. El objetivo aquí es mantener y posiblemente mejorar esta métrica, ya que incrementar la cantidad de usuarios influye en aumentar la tasa de retención en el largo plazo de la *Fintech*. Comparar esta métrica antes y después de la implementación de la estrategia permitirá evaluar la efectividad del modelo de predicción.

4.3.3. Construcción de hipótesis

Para construir la hipótesis, es fundamental formular una pregunta que refleje el objetivo del experimento. Dado que nuestra métrica de éxito se centra en el incremento de los ingresos de la *Fintech*, la pregunta clave es si podemos generar un aumento en los ingresos al detectar patrones que identifiquen a usuarios propensos a convertirse en usuarios no retenidos.

Por lo tanto, la pregunta formulada es la siguiente: ¿Podemos generar un incremento en los ingresos al implementar una estrategia de retención que considere

la probabilidad de un usuario de convertirse en no retenido, según un modelo de *Machine Learning*?

Basándonos en esta pregunta, la hipótesis propuesta para el experimento sería la siguiente: "Diseñar e implementar una estrategia de retención que considere la probabilidad de que un usuario se convierta en no retenido mediante un modelo de *Machine Learning* resultará en un incremento en los ingresos de la *Fintech*".

4.3.4. Desarrollo del experimento

Para comenzar, se optó por trabajar con los usuarios operativos, ya que estos tienen la capacidad de elegir entre mantenerse o cambiar los métodos de pago ofrecidos por la *Fintech*. Luego, se procedió a segmentar estos datos en cuatro categorías según los resultados obtenidos mediante el modelo predictivo: verdaderos retenidos, falsos retenidos, falsos no retenidos y verdaderos no retenidos.

Posteriormente, se elaboró una matriz de confusión que detalla la cantidad de usuarios y el monto estimado a cobrar por cada segmento. Con esta información, se calcularon los ingresos esperadas si se hubiera implementado la estrategia y la cantidad de usuarios que podrían haber sido retenidos.

Finalmente, se exploraron métodos para mejorar las métricas de *precision* y *recall* con el propósito de elevar la métrica de éxito y optimizar el rendimiento de la estrategia de retención.

4.3.5. Análisis financiero

Después de definir las métricas del experimento, se procedió a realizar un análisis financiero que abordaría la implementación del modelo de predicción. El propósito de esta sección es explorar diversos escenarios y evaluar el éxito obtenido en el experimento. De esta manera, buscamos obtener una estimación del costo-beneficio asociado a la aplicación de un modelo de predicción de la fuga de usuarios.

A continuación, en la Tabla 6 se presenta una matriz de confusión donde se describen los cuatro segmentos en los que se dividió a los usuarios y su impacto

financieramente para la *Fintech*. En este contexto, un valor de 1 representa a un usuario retenido, es decir, aquellos que decidieron permanecer y no cambiar su método de pago ofrecido por la *Fintech*. Por otro lado, un valor de 0 indica un usuario no retenido, que son aquellos que optaron por cambiar su método de pago ofrecido por la *Fintech*.

Tabla 6

Interpretación para la Fintech de la matriz de confusión

		Realidad	
		1	0
Predicciones	1	(TP) Verdaderos Retenidos Los que efectivamente no se fugaron.	(FP) Falsos Retenidos Los que el modelo predijo que se quedaban, pero se fueron. La *Fintech* pierde ingresos
	0	(FN) Falsos No Retenidos Los que el modelo predijo que se iban, pero se quedaron. La *Fintech* pierde el costo de retención	(TN) Verdaderos No Retenidos Los que efectivamente se fugaron pero el modelo lo puede detectar. Se puede implementar una estrategia de retención.

4.3.6. Datos utilizados y supuestos

Los datos utilizados provienen de la misma base de datos empleada para evaluar los modelos de clasificación. Sin embargo, se trabajó con una muestra de 6.000 usuarios, agregando columnas relacionadas con el porcentaje de comisión que la *Fintech* cobra por cada transacción realizada por el usuario.

En cuanto a los supuestos, se estableció el costo asociado a contactar a los usuarios que recibirán el mensaje de la propuesta. Se determinó que el costo de retención para un usuario que efectivamente deja de utilizar los métodos de pago ofrecidos por *Fintech*, y que no fue detectado por el modelo, es de $10. Por otro lado, para aquellos usuarios detectados por el modelo pero que optaron por permanecer, el costo se reduce a $2. Además, se estimó un costo de $6 para los usuarios

identificados erróneamente como propensos a irse pero que, en realidad, deciden quedarse.

Adicionalmente, se asumió que la eficiencia de la estrategia, medida como la proporción de usuarios incentivados a pagar por los métodos de pago de la *Fintech* tras recibir el mensaje, es del 70%. Estos supuestos son fundamentales para evaluar el impacto financiero de la implementación de la estrategia de retención propuesta.

En resumen, se consideraron los siguientes costos para los diferentes conjuntos, como se muestra en la Tabla 7:

Tabla 7

Costos asociados a la estrategia de retención

		Realidad	
		1	0
Predicciones	1	(TP) Verdaderos Retenidos Costo: $ 0	(FP) Falsos Retenidos Costo: $10
	0	(FN) Falsos No Retenidos Costo: $6	(TN) Verdaderos No Retenidos Costo: $2

4.3.7. Simulación de la estrategia de retención

Para simular el impacto financiero de la estrategia de retención, se utilizó la información de la matriz de confusión, que proporciona la cantidad de usuarios en cada segmento y la suma estimada de los ingresos. Los ingresos sin estrategia (SE) corresponden a las comisiones reales obtenidas por la *Fintech*, derivadas del procesamiento de pagos a través de los métodos CVU, DB, y DC de los usuarios que transaccionan mediante estos medios. Por otro lado, los ingresos con estrategia (CE) se calcularon utilizando la siguiente fórmula:

$$Ventas\ CE = TP + FN + FP + TN$$

Donde:

$TP = Ingresos_{[1][1]}$

$FN = Ingresos_{[0][1]} - (Costos\ retención * Cantidad\ de\ Usuarios_{[0][1]}$

$FP = -(Costos\ retención * Cantidad\ de\ Usuarios_{[1][0]}$

$$TP = Ingresos_{[1][1]} - \left(Costos\ retención * Cantidad\ de\ Usuarios_{[1][1]}\right) * \%\ Efectividad\ de\ la\ estrategía$$

Mientras que la cantidad de usuarios sin estrategia (SE), se calculó de acuerdo con la cantidad de usuarios que realmente pagaron por la plataforma de la *Fintech*, mientras que usuarios con estrategia (CE), mediante la siguiente formula:

$$Usuarios\ CE = Usuarios_{[1][1]} + Usuarios_{[0][1]} + \left(Usuarios_{[0][0]} * \%\ Efectividad\ estrategia\right)$$

De acuerdo con las fórmulas descritas anteriormente, se llevó a cabo una comparación detallada que examina tanto la diferencia en montos y unidades como el porcentaje de aumento generado por la implementación de la estrategia de retención.

Tabla 8

Comparación de los ingresos y usuarios con la implementación de la estrategia

Ingresos SE	Ingresos CE	Diferencia	%	Usuarios SE	Usuarios CE	Diferencia	%
$29.083	$34.373	$5.290	▲18%	3.146	4.337	1.191	▲ 38%

Según los resultados, se identifica que, con la implementación de la estrategia, la *Fintech* experimentaría un incremento del 18% en sus ingresos por comisiones, mientras que, en términos de usuarios, se observa un aumento del 38% en el número de usuarios retenidos. La diferencia en los porcentajes de crecimiento (18% en ingresos y 38% en usuarios retenidos) se debe a que los ingresos dependen del número de transacciones exitosas, mientras que el aumento de usuarios retenidos refleja una mayor cantidad de usuarios que continúan utilizando los servicios de la *Fintech* para procesar sus pagos. Este comportamiento resalta el potencial beneficio de dirigir los mensajes a usuarios con alta probabilidad de fuga, optimizando así los esfuerzos de retención.

Considerando los costos asociados, la estrategia de retención tendría un costo de $14.944, en contraste con el escenario donde se enviarían mensajes a todos los usuarios, con un costo de $46.376. Esta diferencia substancial subraya la importancia de una estrategia selectiva y focalizada, maximizando así los ingresos y optimizando los recursos disponibles.

Con el objetivo de mejorar las métricas del modelo de predicción y obtener resultados más robustos, se decidió modificar el umbral de decisión del mejor modelo. Se ajustó el umbral de 0.5 a 0.7 para hacer que el modelo sea más restrictivo al predecir la clase de usuarios que realmente cambiarán su método de pago, es decir, aquellos que abandonarán la plataforma. Esta modificación tiene como objetivo aumentar la precisión del modelo, ya que será más cauteloso al clasificar este tipo de usuarios. Sin embargo, es probable que esto también reduzca el *recall* y aumente los costos de retención.

Tras realizar esta modificación del umbral y utilizando las mismas condiciones del experimento anterior, se obtuvieron los siguientes resultados:

Tabla 9

Comparación de ingresos y usuarios con la implementación de la estrategia mejorada la métrica de precision y recall

Mes	Ingresos SE	Ingresos CE	Diferencia	%	Usuarios SE	Usuarios CE	Diferencia	%
1	$29.083	$39.428	$10.346	▲35%	3.146	4.490	1.191	▲ 43%

Con estos resultados, se observa un aumento del 17% en los ingresos en comparación con el experimento anterior. Esto indica que el modelo de predicción ha logrado cumplir el objetivo establecido de incrementar los ingresos en un 35%, según la métrica de éxito definida. Además, se evidencia un incremento en la cantidad de usuarios retenidos.

5. Conclusiones

5.1. Logros alcanzados en el proyecto

En este trabajo, se implementó un modelo de aprendizaje automático para identificar usuarios con alta probabilidad de dejar de utilizar los servicios ofrecidos por la *Fintech*, con el objetivo de crear una estrategia de retención de usuarios. Los datos de usuarios y su actividad operativa fueron proporcionados por la *Fintech*, lo que permitió construir una base de datos con los campos necesarios para realizar un análisis exploratorio de las variables y descubrir patrones o comportamientos de los usuarios.

Posteriormente, se desarrollaron modelos de aprendizaje automático, incluyendo regresión logística, árboles de decisión y XGBoost. Tras comparar el rendimiento de los modelos, se determinó que el modelo XGBoost ajustado con hiperparámetros fue el más efectivo, y se llevó a cabo una validación cruzada para garantizar su robustez, donde arrojo un AUC de 0.9879.

Finalmente, se diseñó un experimento para evaluar la eficacia del modelo en la práctica. Se utilizó una muestra de usuarios con probabilidades proporcionadas por el modelo, y se estableció un umbral de decisión para dirigir la estrategia de retención hacia aquellos usuarios más propensos a dejar de utilizar los servicios. Se estimó que esta estrategia podría aumentar los ingresos de la *Fintech* en un 35% y retener al 43% de los usuarios. En consecuencia, se destaca la importancia de utilizar las probabilidades proporcionadas por el modelo para incrementar los ingresos y retener a los usuarios.

5.2. Limitaciones y futuras posibles mejoras

Como limitaciones del presente trabajo, se identificó la presencia de campos de información de usuarios incompletos o desactualizados. Aunque los resultados del modelo fueron prometedores, se reconoce que mejorar la integridad y actualización de la base de datos sería beneficioso para el análisis exploratorio y el desarrollo de modelos futuros, ya que podrían descubrirse nuevos patrones o comportamientos significativos en los datos.

En cuanto a las mejoras del modelo, es importante tener en cuenta el costo computacional asociado. El proceso se realizó en una computadora de un estudiante, y se encontraron dificultades al llevar a cabo la ingeniería de atributos en una base de datos de aproximadamente 5 millones de registros, lo que llevó a buscar alternativas como por ejemplo consolidar información por usuario, agrupar métricas, construir variables binarias para optimizar las consultas y cálculos y lograr cumplir con los objetivos establecidos.

Además, en relación con la implementación de la estrategia de retención, se sugiere considerar una segunda etapa en la cual los mensajes sean personalizados para cada usuario. Aunque la tasa de retención del 70% utilizada en el análisis fue una asunción basada en la literatura y no un hecho empírico derivado de un experimento inicial, la personalización de los mensajes podría aumentar la eficacia del envío. Esta medida se adaptaría mejor a las necesidades y comportamientos individuales de los usuarios, lo que potencialmente aumentaría la tasa de retención.

Referencias

Agencia Argentina de Inversiones y Comercio Internacional. (2023). *Informe sectorial para inversores internacionales: Tecnología / Fintech* [Archivo PDF]. https://www.inversionycomercio.ar/pdf/sectores/tecno /AAICI_TecnoFIN.pdf

Banco Central de la República Argentina. (2023). *Sistema financiero Entidades no financieras Registro de proveedores de servicios de pago*.https://www.bcra.gob.ar/SistemasFinancierosYdePagos/Proveedores-servicios-de-pago-ofrecen-cuentas-de-pago.asp

Breiman. (2001). *Statistical modeling: The two cultures*. https://doi.org/10.1214/ss/1009213726

Cámara Argentina Fintech. (2022). *Evolución del Empleo Fintech 2022: Ecosistema Argentino*. [Archivo PDF]. https://camarafintech.org/wp-content/uploads/2022/09/Informe-Empleo-Fintech-2022-Camara-Argentina-Fintech.pdf

Carrasco, R. A., Bueno, I., & Montero, J.-M. (2023). *Boosting y el algoritmo XGBoost.* https://cdr-book.github.io/cap-boosting-xgboost.html#:~:text=29.4.1%20Hiperpar%C3%A1metros%20del%20modelo,tanto%2C%20a%20una%20mayor%20precisi%C3%B3n.

Chauhan, N. (2020). *Optimización De Hiper Parámetros Para Modelos De Aprendizaje Automático.* https://www.datasource.ai/es/data-science-articles/optimizacion-de-hiper-parametros-para-modelos-de-aprendizaje-automatico

Díaz González, L., Covarrubias, D., & Sistachs Vega, V. (2015). *Selección de modelos en regresión logística binaria bajo el paradigma bayesiano.* https://elibro.net/ereader/siduncu/91386

Ferrero, R., 2020 *Los árboles de decisión son uno de los algoritmos más utilizados para la toma de decisiones en Machine*

Learning.https://www.maximaformacion.es/blog-dat/que-son-los-arboles-dedecision-y-para-que-sirven/

Finnovista; Banco Interamericano de Desarrollo; BID Invest. (2022). *Fintech en América Latina y el Caribe: un ecosistema consolidado para la recup eración*. http://dx.doi.org/10.18235/0004202

Gutiérrez González, D. (2020). *Técnicas de machine learning en el análisis del churn rate*. https://repositorio.unican.es/xmlui/handle/10902/19075

Maestre, R. (2022). *Qué es fintech y por qué es el futuro de las finanzas.* http s://www.iebschool.com/blog/que-es-fintech-finanzas/

Muscillo, M., Vitale, I., & Peters, N. (2020). *Economía: Definiciones de Fintech para comprender mejor el ecosistema.* https://netnews.com.ar/nota/2778-Definiciones-de-Fintech-para-comprender-mejor -el-ecosistema

Ortuño, M., 2022 Un modelo de regresión logística para el análisis de los aspectos que influyen en la anulación de pólizas de seguros de automóviles. https://masteres.ugr.es/estadistica-aplicada/sites/master/moea/public/inline-files/TFM_ORTU%C3%91O_ROIG_MARIA.pdf

Pérez, R. Pino, G. Ballester, D. & Moreno, R. (2010). *Modelo de regresión logística para estimar la dependencia según la escala de Lawton y Brody*. DOI: 10.1016/j.semerg.2010.03.004

Picón Montero, P. A., & Vásquez Silva, D. (julio de 2022). *Análisis de las princ ipales variables Fintech y su impacto en el Sistema Financiero Tradiciona l Colombiano.* [Archivo PDF]. https://digitk.areandina.edu.co/bitstream/handle/areandina/5027/Trabajo%20de%20grado.pdf?sequence=1&isAllowed=y

Pozo, J. (2020). Fidelización : ¿ Qué es el churn rate?. https://elviajedelcliente.com/churn-rate/

Ramírez, J (2024). *Domina los Hiperparámetros de XGBoost: Optimizando el Rendimiento de Entrenamiento y Ejecución en SageMaker (con Ejemplos en Python)*

Read the docs. (2023). *Introduction to Boosted Trees.* https://xgboost.readthedocs.io/en/stable/tutorials/model.html

Saltos, J. (2022). *¿Qué es un análisis de cohortes y cómo hacer uno?* https://jorgesantos.co/analisis-de-cohortes/

Segura, J. (2022). *Desarrollo de un modelo de predicción de fuga de clientes y diseño de experimento para la aplicación de estrategias de fidelización en factoring.* https://repositorio.uchile.cl/ha ndle/2250/ 186972

Sierchuk, S. (2022). *Predicción de Churn en Fintech*. https://repositorio.utdt.edu/handle/20.500.13 098/11874

Stripe. (2023). *Aspectos básicos de las pasarelas de pagos: Qué son y qué función desempeñan en el procesamiento de los pagos.* https://stripe.com/es/resources/more/payment-gateways-101

Torres, L. (2023). Curva ROC y AUC en Python. https://www.themachinelearners.com/curva-roc-vs-prec-recall/

Tralice, F. (2019). *Predicción de Churn de Seguros con LightGBM.* https://repositorio.utdt.edu/handle/20.500.13098/11242

Zendesk. (2022). *¿Qué es la retención de clientes y cómo aumenta las gananci as?* https://www.zendesk.com.mx/blog/que-es-retencion-de-clientes/

Apéndice A. Revisión de distribución de variables

Figura 19

Cantidad de datos no nulos por variable en tabla “Transacciones”

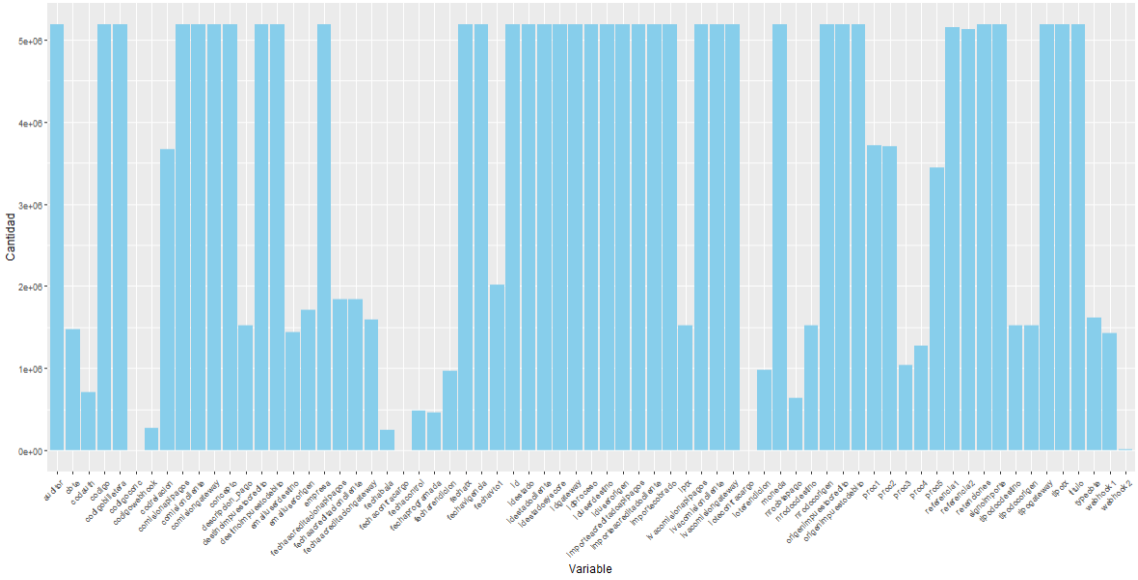
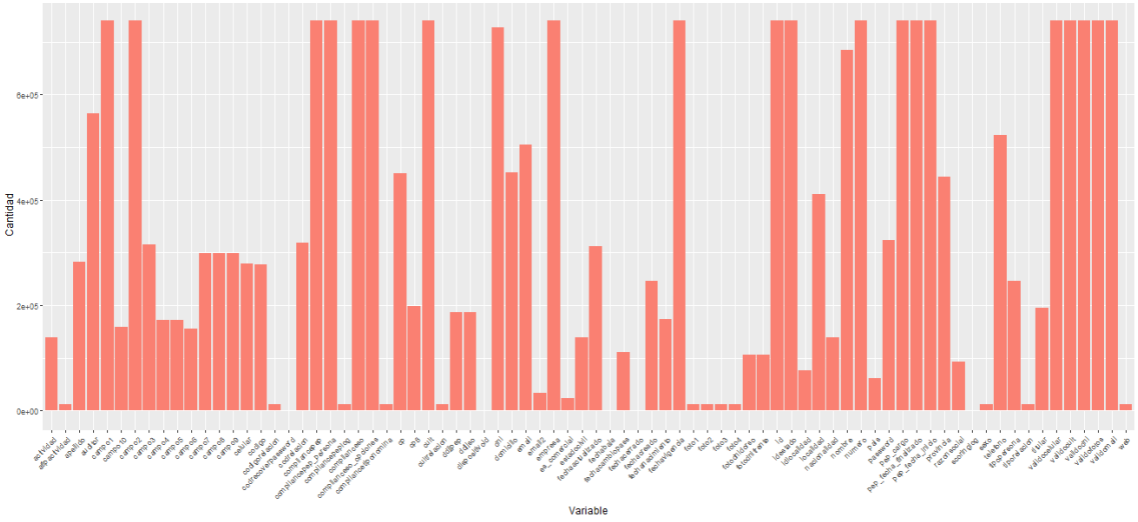


Figura 20

Cantidad de datos no nulos por variable en tabla “User”



Apéndice B. Curva ROC

Figura 21

Curva ROC del mejor modelo de clasificación XGBoost con optimización de hiperparámetros