Modelos predictivos para la segmentación de clientes de un banco neo digital

Alumno: Lucas Solari Fredericksen

Universidad: Universidad Adolfo Ibáñez

Carrera: Ingeniería Civil Industrial

Empresa y área: Banco Ripley-Chek Área Comercial

Profesor: Gonzalo Anriquez Gundián

Fecha: 2023-24

Resumen Ejecutivo

El área comercial del banco neo digital Chek está encargada de retener y captar clientes a través de distintas estrategias y/o campañas. Esto implica la promoción de los servicios y productos que ofrece la empresa, la adquisición de nuevos clientes, gestionar la relación con los clientes actuales e impulsar la participación de los clientes en los productos financieros ofrecidos. Dentro de los objetivos anuales propuestos por la empresa, es lograr una cantidad determinada de stock de clientes del último producto lanzado (a comienzos de Julio), la tarjeta de crédito. Para conseguir este objetivo, se hace necesario, una sinergia y comunicación, ya sea, entre miembros del equipo o con otras áreas, para recopilar información, coordinar comunicaciones, generar campañas comerciales, recibir la base de usuarios con oferta de tarjeta de crédito, etc. Los canales de comunicación por los que se comunican los diferentes beneficios, y se gestiona la relación con los usuarios son notificaciones push, WhatsApp, mail, SMS y banners dentro de la aplicación, al coexistir distintos productos que ofrece Chek, es necesario estar en constante coordinación con el equipo con el fin de no agobiar a los clientes de la empresa.

Dado esto, para medir el desempeño de las campañas comerciales, se utilizan distintas métricas, tales como, tasa de apertura de los canales de comunicación, cantidad de captaciones de crédito, cantidad de “clientes ready” (usuarios que ingresaron todos sus datos, verificaron su identidad, ampliaron su cupo y están dentro de la base de clientes con oferta de tarjeta de crédito Chek) , tasa de conversión de tarjetas de crédito y el interés mostrado por los usuarios (clics a los banners dentro de la aplicación).

Para contextualizar la situación inicial del proyecto, entre los meses de Julio y Septiembre, el cumplimiento de la meta para las captaciones de tarjeta de crédito fue extremadamente baja, alcanzando tan solo 50 de 4.250 captaciones de clientes de tarjeta de crédito esperadas (alrededor de 1% de la meta). Además, los canales de comunicación, en específico, las notificaciones push, tienen una tasa de apertura de menos del 1% y la tasa de conversión de tarjetas de crédito promediaba el 2%.

Dado lo anterior, es posible apreciar que no se están cumpliendo con las expectativas de las metas planteadas por la empresa para la tarjeta de crédito, y esto es debido a que las distintas comunicaciones que se realizan por parte del equipo comercial son genéricas y no son segmentadas, se distingue, tan solo, por

usuarios que tienen oferta de tarjeta de crédito y usuarios que no cuentan con oferta de tarjeta de crédito. Dicho esto, y en discusión con el equipo, se planteó el objetivo de generar 100 captaciones de tarjeta de crédito (contribuir con un 10% de la “nueva meta” propuesta de llegar a un stock de 1000 usuarios con tarjeta de crédito para fines de Diciembre).

Para lograr este objetivo, se analizó e investigó diferentes soluciones, con el fin de encontrar el óptimo. De esta forma, se desarrolló un modelo predictivo de clasificación, utilizando el algoritmo de regresión logística, con el fin de encontrar aquellos usuarios más propensos a captar la tarjeta de crédito Chek, utilizando las métricas de cross-validation (accuracy, precisión, f1 score y recall para medir la efectividad de este. Es necesario decir, que previamente se hizo un modelo de balanceo con el fin de lograr que el modelo de clasificación tenga una mayor precisión y eficiencia, debido a que al haber una clase minoritaria (captados), no lograba identificarlos como tal, y los clasificaba como la clase mayoritaria (no captados). Para luego, llevar a cabo una segmentación de usuarios, a través de un modelo de agrupamiento, utilizando el algoritmo K- Means y el método del codo para elegir la cantidad de clústeres, permitiendo la ejecución de campañas comerciales personalizadas y adaptadas al perfil de estos usuarios. Dando como resultado 3 clústeres, a los cuales se analizaron sus características, para generar así 5 campañas comerciales, según los “departamentos” más concurridos en tiendas Ripley o ripley.com (página web y tienda online de Ripley), siendo estos “PERFUMERIA”, “JUGUETERIA”,” DECORACION “, “VESTUARIO HOMBRE”, “VESTUARIO MUJER”.

Dentro de los resultados obtenidos por las campañas comerciales generados por el proyecto, se debe destacar el aumento en 2 puntos porcentuales, de 2% a 4% promedio de la tasa de conversión de tarjetas de crédito y el aumento de la tasa de apertura en las notificaciones push, promediando un 6% de apertura. Además, se logró aumentar en un 8% la tasa de interés (clics en banners relacionados a las campañas comerciales de tarjeta de crédito). De la misma forma, se logró superar la meta esperada de 100 captaciones de usuarios de tarjeta de crédito, obteniendo 120 captaciones en total. Dado los resultados de este conjunto de objetivos específicos, se pudo lograr el objetivo general de mejorar el rendimiento y aumentar la cantidad de conversiones en 100 unidades de la tarjeta de crédito del banco neo-digital Chek mediante una estrategia de segmentación de usuarios efectiva.

Abstract

The commercial area of the neo digital bank Chek oversees retaining and attracting clients through different strategies and/or campaigns. This involves promoting the services and products offered by the company, acquiring new customers, managing the relationship with current customers, and driving customer participation in the financial products offered. Among the annual objectives proposed by the company is to achieve a certain amount of customer stock for the latest product launched (at the beginning of July), the credit card. To achieve this objective, constructive collaboration and communication are necessary, either between team members or with other areas, to collect information, coordinate communications, generate commercial campaigns, receive the user base with a credit card offer, etc. The communication channels through which the different benefits are communicated, and the relationship with users is managed, are push notifications, WhatsApp, email, SMS, and banners within the application. As assorted products offered by Chek coexist, it is necessary to be constantly Coordination with the team in order not to overwhelm the company's clients.

Given this, to measure the performance of commercial campaigns, different metrics are used, such as the opening rate of communication channels, the number of credit acquisitions, the number of “ready customers” (users who entered all their data, verified their identity, they expanded their quota and are within the customer base with Chek credit card offer, credit card conversion rate and the interest shown by users (clicks on the banners within the application).

To contextualize the initial situation of the project, between the months of July and September, the fulfillment of the goal for credit card acquisitions was extremely low, reaching only 50 of the 4,250 expected credit card customer acquisitions (around 1 % of goal). Additionally, communication channels, specifically push notifications, have an open rate of less than 1% and the credit card conversion rate averaged 2%.

Given the above, it is possible to see that the expectations of the goals set by the company for the credit card are not being met, and this is because the different communications made by the commercial team are generic and are not segmented, it is only distinguished by users who have a credit card offer and users who do not have a credit card offer. Having said this, and in discussion with the team, the objective of generating 100 credit card acquisitions was set (contributing

with 10% of the proposed “new goal” of reaching a stock of 1000 credit card users for the purposes of December).

To achieve this objective, different solutions were analyzed and investigated to find the optimal one. In this way, a predictive classification model was developed, using the logistic regression algorithm, to find those users most likely to acquire the Chek credit card, using the cross-validation metrics (accuracy, precision, f1 score and recall measuring its effectiveness. It is necessary to say that a balancing model was previously made to ensure that the classification model has greater precision and efficiency, due to the fact that there is a minority class (captured), could not identify them as such, and classified them as the majority class (not captured). Then, conduct a segmentation of users, through a grouping model, using the K-Means algorithm and the elbow method to choose the number of clusters, allowing the execution of personalized commercial campaigns adapted to the profile of these users. Resulting in 3 clusters, whose characteristics were analyzed, to generate 5 commercial campaigns, according to the busiest “departments” in stores Ripley or ripley.com (Ripley's website and online store), these being “PERFUMERY”, “TOY STORE”, “DECORATION”, “MEN'S CLOTHING”, “WOMEN'S CLOTHING”.

Among the results obtained by the commercial campaigns generated by the project, it is worth highlighting the increase of 2 percentage points, from 2% to 4% average, in the credit card conversion rate and the increase in the opening rate in the push notifications, averaging 6% opening. In addition, the interest rate was increased by 8% (clicks on banners related to credit card commercial campaigns). In the same way, the expected goal of 100 credit card user acquisitions was exceeded, obtaining 120 acquisitions in total. Given the results of this set of specific objectives, the overall objective of improving performance and increasing the number of conversions in 100 units of the neo-digital bank Chek credit card through an effective user segmentation strategy.

Índice

Introducción
Contexto

Chek es una empresa fundada en 2019, está ligada al Banco Ripley, y se posiciona como una de las empresas líderes del rubro de los bancos neo-digitales. Los bancos neo-digitales son instituciones financieras, que operan exclusivamente en línea (en el caso de Chek a través de una aplicación móvil), sin sucursales físicas (a diferencia de la banca tradicional). Tienen la capacidad de ofrecer servicios bancarios a través de aplicaciones móviles, sitios web o plataformas en línea, de forma fácil e intuitiva esto permite que sean más flexibles y ágiles en comparación a la banca tradicional. Pueden ofrecer distintos productos, tales como, cuentas corrientes, cuentas de ahorro, tarjetas de débito y crédito, súper avances, avances, créditos de consumo, entre otros. Además de ofrecer distintos servicios, por ejemplo, pagos de cuentas, cargas para las tarjetas del sistema de transporte público (Tarjeta BIP), transferencias al extranjero, etc.

El área comercial de la empresa está encargada de retener y captar clientes a través de distintas estrategias y/o campañas. Esto implica la promoción de los servicios y productos que ofrece la empresa, la adquisición de nuevos clientes, gestionar la relación con los clientes actuales e impulsar la participación de los clientes en los productos financieros ofrecidos. Por otra parte, el área comercial también se dedica a analizar la competencia, identificar oportunidades de mercado y mejorar continuamente la oferta de servicios y productos para satisfacer las distintas necesidades de los clientes de la empresa. A grandes rasgos, el área comercial de la empresa es la responsable de la estrategia de

crecimiento y de garantizar el cumplimiento de las metas u objetivos propuestos por la empresa.

Dado lo anterior, la empresa Chek ofrece distintos productos financieros, tales como, créditos de consumo, avances, super avances y tarjetas de créditos. Además, dispone de los servicios de carga tarjeta BIP, transferencias al extranjero, tarjeta de débito, transferencias a terceros y pago con QR.

Por otro lado, los 3 pilares fundamentales en que se basa este rubro son las etapas de adquisición, retención y monetización. La primera etapa se enfoca en captar nuevos usuarios hacia la empresa con el objetivo de aumentar la visibilidad y atraer clientes que podrían estar interesados en los productos que se ofrecen. La etapa de retención se da una vez que los usuarios ya han sido adquiridos, y consta de gestionar la relación con los clientes para fomentar la lealtad y el compromiso de estos con la empresa. Por último, la etapa de monetización hace referencia a cómo la empresa traduce la base de clientes formada en los pasos anteriores en ingresos para rentabilizar el negocio.

Dado el contexto anterior, Chek se encuentra entre la etapa de retención y monetización. Dispone de una base total de aproximadamente 1.500.000 usuarios los cuales son divididos en 3 grupos, estos son, tier 1 (usuarios que se inscribieron en la app a través de su número de celular, tienen un cupo máximo de $20.000 CLP), tier 2 (usuarios que ingresaron todos sus datos a la app pero aún no validan su identidad a través de la cédula de identidad, tienen cupo máximo de $100.000 CLP) y por último, tier 3 (usuarios que ingresaron todos sus datos, validaron su identidad a través de la cédula de identidad, tiene cupo máximo de $1.000.000 CLP).

Problemática

Actualmente, Chek presenta 5 canales de comunicación para interactuar con los usuarios, estos son, SMS, mail, WhatsApp, banners dentro de la aplicación y notificaciones push (notificaciones que llegan a los celulares) y de esta forma enviar las distintas campañas comerciales y ofertas de productos. El problema que se desarrollará en este proyecto es el de la ineficacia de la segmentación de

las comunicaciones de las campañas comerciales, ya que, estas son genéricas para los distintos usuarios de la empresa, se distingue, tan solo, por usuarios que tienen oferta de tarjeta de crédito y usuarios que no cuentan con oferta de tarjeta de crédito. Debido a esto las tasas de apertura y conversiones no han sido las esperadas, y por consecuencia, no se ha cumplido con las expectativas de metas financieras propuestas por la empresa.

Cabe destacar, que Chek no cuenta con un departamento de análisis de riesgo financiero propio, por lo que el departamento de análisis de riesgo financiero de Banco Ripley es el que debe enviar mensualmente una base de datos con aquellos usuarios con oferta de productos financieros. En base a esto, las campañas comerciales se segmentan en “usuarios con oferta de productos financieros” y “usuarios sin oferta de productos financieros”.

Además, es importante recalcar que los banners son visibles tan solo para aquellos usuarios que sean tier 3, tengan una versión actualizada de la aplicación móvil y estén dentro de la base de datos de los usuarios con ofertas enviadas por el departamento de riesgo financiero perteneciente a Banco Ripley. Siendo estos los únicos usuarios que tienen la posibilidad de cursar alguno de los productos ofrecidos por la empresa.

A continuación, se muestran tablas con las ventas y captaciones, contra las metas esperadas para el mes. Además, se muestra un funnel de Septiembre que demuestra las personas que entran a la aplicación, cuantas de estas tienen oferta de tarjeta de crédito, cuántas cumplen con los requisitos para captar la tarjeta de crédito (“clientes ready”), cuantos de estos mostraron interés (dieron clic en algún banner dentro de la app relacionado a tarjeta de crédito), y por último cuantos de los usuarios que mostraron interés captaron la tarjeta.

Imagen 1: Funnel Tarjeta de Crédito Septiembre 2023

2023	Julio	Agosto	Septiembre
Captaciones	14	16	20
Meta captaciones	50	1.500	2.700
Cumplimiento captaciones	28%	1%	1%

Tabla 1: Cumplimiento captaciones tarjeta de crédito 2023

En la tabla 2, es posible ver el cumplimiento de las metas para el primer trimestre desde el lanzamiento de la tarjeta de crédito, claramente, están bajo las expectativas de la empresa (alrededor de 1% de cumplimiento).

Objetivos

En congruencia con el problema y el contexto descrito anteriormente, se hace necesario definir los alcances de este proyecto. Es importante recalcar, que realizar una segmentación de usuarios eficiente para las comunicaciones de las campañas comerciales, implican un impacto directo en los resultados de la empresa. Por esto es por lo que es necesario realizar un análisis exhaustivo de los objetivos y los efectos o consecuencias que estos podrían tener.

Objetivo General

Dado que la tarjeta de crédito es el producto actualmente más “débil” de la empresa Chek en cuanto al cumplimiento de metas. El objetivo general, de este proyecto es mejorar el rendimiento y aumentar la cantidad de captaciones en 100 unidades de la tarjeta de crédito del banco neo-digital Chek para el periodo de actividad de las campañas comerciales generadas (20 de Noviembre hasta 24 de Diciembre), mediante una estrategia de segmentación de usuarios efectiva que permita mejorar los resultados de la tarjeta de crédito, y de esta forma contribuir con el cumplimiento de las metas comerciales.

Objetivos Específicos

Por otra parte, haciendo uso de la metodología SMART para la definición de los objetivos:

Específico: Existen 3 objetivos específicos, los cuales impactan directamente la rentabilidad de la empresa.

- En primer lugar, aumentar la tasa de conversión de la tarjeta de crédito, la situación inicial (primer trimestre de lanzamiento de la tarjeta de crédito JULIO-AGOSTO-SEP) promedia una tasa de conversión de 2%, se espera al menos subir en 1% esta tasa, respecto al interés generado.

- Por otro lado, se espera aumentar la tasa de apertura del canal de comunicación, notificación push. El canal de comunicación más débil son las comunicaciones push, el cual actualmente tiene una tasa de apertura promedio menor al 1%. Para fines de este proyecto se espera una tasa de apertura promedio del 5% para este canal.

- Y, por último, aumentar la cantidad de clientes interesados, estos son usuarios que hayan dado clic en alguno de los banners que hay dentro de la aplicación. Estos objetivos afectan directamente en el cumplimiento de las metas comerciales y el objetivo general del proyecto. En la situación inicial (primer trimestre de lanzamiento de la tarjeta de crédito JULIO-

AGOSTO-SEP), el interés promedio generado por los banners promedia un 17%. Se espera un aumento de 5 puntos porcentuales.

Medible: Las métricas con las que se medirá el impacto del proyecto sobre los resultados de Chek son las siguientes.

- En primer lugar, la cantidad de captaciones de tarjetas de créditos atribuibles a una de las campañas comerciales realizadas.

- En segundo lugar, la tasa de apertura de las distintas comunicaciones que se utilicen para las campañas comerciales realizadas, enfocándose principalmente notificaciones push.

- Además, se medirá la tasa de conversión de la tarjeta de crédito, es decir, cuantos de los usuarios con oferta que han mostrado interés concretan la captación

- Y, por último, la cantidad porcentual de clientes que muestran interés (usuarios que hacen clic en algún banner relacionado a tarjeta de crédito dentro de la aplicación).

Alcanzable: Este proyecto es realizable con el presupuesto y recursos dispuesto por la empresa Chek, tales como, VPN, Big Query (Google Cloud), Python, licencia de Microsoft Office, Firebase, entre otros.

Relevante: Este proyecto es relevante para la empresa, ya que impactará directamente en el alcance de las metas comerciales de Chek y contribuirá con el objetivo estratégico de aumentar la rentabilidad de los productos financieros y brindar un mejor servicio a los clientes.

Temporal: Este proyecto pretende implementarse en el presente semestre, teniendo como cierres mediados de Noviembre, realizando revisiones regularmente para hacer futuras correcciones y lograr resultados óptimos.

Estado del Arte

La evolución tecnológica y la disponibilidad de los datos ha impactado significativamente en el área comercial de las empresas, tanto en las campañas comerciales como en el análisis de datos que facilitan los usuarios. La segmentación de usuarios desempeña un papel fundamental al momento de realizar las distintas comunicaciones de las campañas comerciales, generando ofertas más personalizadas y aumentando el nivel de interés de los clientes sobre los productos ofrecidos. Dentro de las segmentaciones de usuarios se pueden clasificar las siguientes:

Segmentación Demográfica: Se segmenta por edad, género, ubicación, grupo socioeconómico, estilo de vida, etc.

Segmentación Comportamental: Se segmenta según el análisis del comportamiento del usuario, tal como, el historial de compras, interacciones con las distintas comunicaciones (mail, notificación push, WhatsApp, etc.), navegación por la web, etc.

Segmentación geográfica: Este tipo de segmentación agrupa los usuarios según su localización. Utiliza variables como idioma, cultura, tipo de población (rural o urbana), idioma, etc.

Segmentación psicográfica: Segmenta según creencias, personalidad, intereses, etc.

Segmentación predictiva: Utiliza algoritmos de aprendizaje automático para predecir el comportamiento de los usuarios, y de esta forma, generar campañas más personalizadas y aumentar la efectividad de estas.

Dado que la segmentación predictiva es la más completa, y para fines de este

proyecto con los recursos y técnicas disponibles son adecuadas para aplicarla, se profundizará en este tipo de segmentación. El análisis predictivo examina un conjunto de datos para interpretarlos, detectar patrones, y obtener predicciones sobre estos. A continuación, se presentan algunos modelos predictivos.

Modelo de clasificación: Este modelo predice si es que los datos pertenecen a una clase. Se crea una clasificación que permite analizar eficientemente los tipos de datos. Es el más sencillo de los modelos, y utiliza respuestas binarias.

Modelo de agrupación: Este modelo asigna una variable en distintos grupos, basándose en los atributos que comparten. Es útil para identificar características y comportamientos que comparten determinados grupos dentro de la base de datos.

Modelo de pronóstico: Este modelo utiliza los datos históricos para predecir métricas de valor, de esta forma, estima el valor numérico de nueva información en base a la información antigua. Se usa mayormente para proyecciones de demanda, proyecciones de inventario, etc.

Modelo de valores atípico: Este modelo se basa en la anomalía de los datos, ya sea porque son atípicos por sí mismos o lo son en comparación con otros datos de su mismo grupo. Se utiliza para detectar fraudes, artículos defectuosos, irregularidades, etc.

Modelo de serie temporal: Este modelo utiliza datos de un período para desarrollar una métrica que utiliza para proyectar lo que sucederá a futuro. Es usado para comprender como una métrica se desarrollará a lo largo del tiempo.

Para fines de este proyecto, se profundizará en los modelos predictivos de agrupación y clasificación, en el ámbito de las ciencias de datos, es común que en los proyectos se agrupe información similar con el objetivo de analizarla con mayor eficiencia, a esto se le denomina “clustering”, esta es una técnica basada en identificar características comunes, patrones, tendencias y relaciones entre datos para lograr agruparlos por categorías y segmentos. Esto trae beneficios tales como una mejor comprensión de los datos, mejorar la calidad de los servicios y/o productos ofrecidos, aumentar la eficiencia de las campañas comerciales, reducir costos, entre otros.

A continuación, se detallarán distintos tipos de algoritmos clustering que pueden llegar a ser útiles para la implementación del proyecto.

En primer lugar, el algoritmo “K-Means” es uno de los más utilizados, se basa en centroides y es el algoritmo de aprendizaje no supervisado más simple.
K-Means minimiza la varianza de los datos dentro de un grupo, funciona mejor con una cantidad reducida de datos, ya que es iterativo, por lo que mientras más datos haya en el conjunto mayor será el tiempo de demora.

Por otro lado, el algoritmo “DBScan” es un tipo de agrupamiento basado en la densidad, separando “regiones” por áreas de baja densidad para detectar valores atípicos entre conjuntos de alta densidad. Utiliza 2 parámetros para definir los grupos “minPts” (el número mínimo de puntos de datos que deben agruparse para que un área se considere de alta densidad) y “eps” (la distancia utilizada para determinar si un punto de datos está en la misma área que otros puntos de datos).

También, el algoritmo de “Mezcla Gaussiana”, este modelo calcula la probabilidad de que un punto de datos pertenezca a una distribución Gaussiana específica y ese es el grupo en el que se ubicará.

Otro algoritmo es “BIRCH” (Balance Iterative Reducing and Clustering using Hierarchies), este divide datos en resúmenes que contienen información sobre la distribución de los datos, los cuales se agrupan en lugar de los puntos de datos originales. La desventaja de este algoritmo es que solo puede usarse con datos numéricos, debiendo realizar transformaciones para aquellos datos categóricos.

Por otro lado, el algoritmo de agrupamiento por “Propagación de Afinidad” funciona diferente a todos los anteriormente nombrados, en este caso, los datos se comunican entre sí para saber qué tan similares son y de esta forma ir agrupando los datos. Tiene una gran ventaja, ya que no es necesario definir la cantidad de clústeres esperados.

Además, existe el algoritmo OPTICS (Ordering Points to Identify the Clustering Structure), al igual que el algoritmo DBScan es un algoritmo basado en densidad, con la diferencia de que este encuentra agrupaciones significativas en datos que varían de densidad, lo cual facilita la detección de diferentes grupos de densidad.

Por último, el algoritmo de “Jerarquía Aglomerativa”, es usado para agrupar datos en función de sus similitudes. Agrupa de arriba hacia abajo, donde cada dato se asigna a su grupo, luego de cada iteración los grupos similares se unen hasta que cada dato forma parte de un “grupo raíz”. Es útil para encontrar pequeños clústeres. El resultado es parecido a un dendrograma, facilitando la visualización de los grupos.

En términos de modelos predictivos de clasificación, se describirán los más comunes.

En primer lugar, los árboles de decisión, los cuales son algoritmos de aprendizaje automático supervisado utilizados para predecir y clasificar. Construyen un modelo que divide los datos en distintas ramas, cada uno representando una categoría o resultado. Estos modelos son útiles para problemas con múltiples variables y resultados posibles.

Por otro lado, el algoritmo de máquinas de vectores de soporte (SVM), es una técnica de aprendizaje automático supervisado utilizada para la clasificación y regresión. Construyen un modelo que separa los datos en diferentes clases haciendo uso de un hiperplano.

Además, existe el algoritmo de K-Vecinos más cercanos (KNN), el cual se basa en la similitud entre los puntos de datos. Utiliza los “K” puntos de datos más cercanos para clasificar un nuevo punto de datos.

Por último, el modelo de clasificación de regresión logística, el cual aplica una clasificación binaria para predecir el resultado. Es utilizado en problemas en que la respuesta es “sí” o “no”, 0 o 1, “captado” o “no captado”, “verdadero” o “falso”, etc.

En el rubro bancario, están presentes los modelos predictivos en distintas áreas del negocio. Se utilizan para la gestión de clientes, gestión del riesgo, soporte a las operaciones, entre otros.

Por ejemplo, un caso de éxito es el de Banco Itaú Argentina, el cual implementó un modelo predictivo para optimar sus campañas de ventas cruzadas de

productos (cross-selling). El banco argentino necesitaba incrementar la tasa de respuesta de sus campañas de ventas, con el fin de aumentar el flujo de ingresos y su participación en el mercado.

Dado esto, desarrollo modelos predictivo para seleccionar a los usuarios con mayor probabilidad de aceptar una oferta. Como resultados, Banco Itaú Argentina, mejoró la comprensión de los clientes, y, por tanto, la precisión de sus campañas, aumentando los ingresos obtenidos de la cartera de clientes existentes en un 40%. Además, la utilización de estas técnicas de optimización incrementó el margen de contribución de los clientes en un 60%.

Otro caso de éxito es el del Banco Columbia, el banco enfrentaba un problema de morosidad por parte de los clientes, por lo que desarrolló un modelo predictivo que le permitió automatizar la originación y gestión de sus productos, acotando el riesgo y maximizando el valor de los clientes. Como resultados, Banco Columbia, redujo los índices de morosidad en un 20%, incrementó la velocidad en el procesamiento de solicitudes, disminuyó los costos administrativos y obtuvo una mayor precisión en el proceso de evaluación, suscripción y supervisión de los créditos.

Dicho esto, es posible apreciar que estos casos presentan 2 soluciones de empresas líderes a nivel regional en la industria bancaria, evidenciando casos de éxitos mediante la implementación de modelos predictivos.

Evaluación de Costos

En términos de costos, es necesario definir cómo actúan los diferentes canales de comunicación con los usuarios, cuáles son sus gastos y la disponibilidad del presupuesto de la empresa para implementar la campaña comercial ideada por el proyecto.

En primer lugar, los canales de comunicación de WhatsApp, notificaciones push y mailing tienen un costo anual. En el caso de WhatsApp es a través de la plataforma “MasivApp”, para las notificaciones push se utiliza la plataforma “Firebase” y para el mailing se utiliza la plataforma “JIRA”, cada uno con planes de “Enterprise”. Para el caso de los SMS, es el canal

de comunicación más costoso, ya que por cada usuario al que se le envía un mensaje de texto hay un cobro de $2 CLP.

Por otro lado, cada canal de comunicación tiene un límite de envío de mensajería. En el caso de las notificaciones push son ilimitadas y pueden enviarse a la cantidad de usuarios que sea necesario, siempre y cuando tengan descargada la aplicación. Para WhatsApp, el envío de comunicaciones es limitado (3 veces a la semana máximo) y la cantidad de usuarios es limitada (10.000 usuarios por base de datos). Para el mailing, no hay límite de envío, pero la cantidad de usuarios es limitada (500.000 usuarios por base de datos). Las notificaciones push no tienen un límite de envíos ni de usuarios, sin embargo, la cantidad de usuarios que se cargan por base debe ser prudente (alrededor de 500.000 usuarios como tope), ya que una base excesivamente grande puede tener problemas de carga y perdida de datos. Por último, en el caso de los SMS la cantidad de envío es limitada (2 veces por semana) y la cantidad de usuarios es ilimitada. Los planes contratados de las plataformas contratadas por la empresa son de "Enterprise", y si bien, el proyecto no tiene un impacto en estos, se detallan a continuación. Para la plataforma JIRA el costo es de aproximadamente $40.000 CLP mensual por empleado que usa la plataforma (Chek dispone de 60 empleados aproximadamente) dando un costo anual de aproximadamente $30.000.000 CLP. Firebase, tiene un costo anual de $15.000.000 aproximadamente. Por último, MasivApp tiene un costo de $10.000.000 anuales aproximadamente.

Dicho lo anterior, es necesario estar en constante comunicación y coordinación con el resto del equipo del área comercial, dado que además de que cada canal de comunicación tiene un límite de envíos (como se expuso anteriormente), el exceso de envíos de comunicaciones comerciales por parte de la empresa puede producir consecuencias negativas debido al hostigamiento, tales como, que el usuario desinstale la aplicación móvil (actualmente el costo de "enroll[1]"es de aproximadamente "$15.000CLP", por lo cual esto conlleva una pérdida a la empresa), caer en mensajería como SPAM, entre otros casos.

Por último, cabe recalcar que cada campaña comercial puede conllevar un gasto adicional, dado que se pueden realizar incentivos a los usuarios para lograr captaciones, aumento de tier, que el usuario actualice la aplicación móvil, entre otros.

[1] Usuarios que descargan la aplicación y obtienen su tarjeta Chek

Canal de comunicación	Costo Anual (Plataforma)	Limite de usuarios	Límite de envíos (semanal)
WhatsApp	$10.000.000	10.000	3
Mailing	$30.000.000	500.000	Ilimitado
Notificación Push	$15.000.000	Ilimitado	Ilimitado

Tabla 2: Costos por plataforma

Evaluación de Riesgos

Para la implementación del proyecto, es necesario tener en consideración los posibles riesgos y mitigaciones que pueda conllevar. Para esto, se utilizará una matriz con el impacto y la probabilidad de ocurrencia, con el objetivo de hacer una medición del riesgo. Asimismo, se asignará un puntaje y un nivel de clasificación según los criterios anteriormente descritos.

Bajo	1 a 4
Medio	5 a 9
Alto	10 a 16

Probabilidad	Impacto			
	Leve (1)	Moderado (2)	Grave (3)	Crítico
Poco probable	1	2	3	4
Medianamente Probable	2	4	6	8
Probable	3	6	9	12
Altamente Probable	4	8	12	16

Riesgo	Probabilidad	Impacto	Valoración	Clasificación	Mitigación
Caídas de las plataformas de los canales de comunicación.	Poco Probable	Moderado	2	Bajo	Utilizar otro canal de comunicación disponible, si el problema se da regularmente, buscar alternativas de plataforma.
Resistencia al cambio.	Medianamente probable	Moderado	4	Bajo	Demostrar el impacto de la solución propuesta en los resultados de la empresa.
Desgaste del canal de comunicación.	Medianamente probable	Grave	6	Medio	Adaptar estrategias de comunicacón efectivas.
Saturación del canal de comunicación.	Medianamente probable	Grave	6	Medio	Diversificar los canales de comunicación utilizados y/o explorar nuevas alternativas.
Ineficacia de la segmentación de usuarios.	Medianamente probable	Grave	6	Medio	Refinar continuamente los criterios de segmentación basandose en los datos que se recopilen y la retroalimentacion que entreguen los usuarios.

Tabla 3, 4 y 5: Matrices Evaluación de Riesgos

Elección de Solución

Para efectos de este proyecto, en términos de segmentación, se optó por la segmentación predictiva, dado que abarca en plenitud las características de los usuarios, la base de datos recopilada contiene los datos necesarios, y además en la etapa académica se aprendieron distintos métodos para poder realizar este tipo de segmentación.

A continuación, se presenta una matriz que pondera los distintos modelos predictivos en una escala del tipo Likert de 1 a 5 puntos, siendo 1 muy bajo y 5 muy alto. Para la asignación de estos puntajes y las ponderaciones se discutió con el equipo del área comercial para considerar los factores más relevantes. Dado esto, se determinaron los siguientes criterios, puntajes y ponderaciones.

Criterio	Alineación con el proyecto	Efectividad e Impacto	Recursos Disponibles	Facilidad de implementacion	
Tipo/Ponderación	25%	25%	25%	25%	Total
Modelo de clasificacion	5	5	5	4	4.75
Modelo de regresión	3	2	4	3	3
Modelo de agrupación	5	5	4	5	4.75
Modelo de pronóstico	3	2	4	3	3
Modelo de valores atípicos	2	1	4	3	2.5
Modelo de serie temporal	1	2	3	3	2.25

Tabla 6: Asignación de puntaje y ponderación de soluciones.

Metodología

Para este proyecto se utilizará la Metodología Fundamental para la Ciencia de Datos hecha por el científico de datos de la organización IBM Analytics, John B.

Rollins. Esta metodología consta de 10 etapas las cuales serán descritas a continuación.

Etapa 1: Comprensión del negocio

Es la etapa inicial de todos los proyectos. Se hace necesario definir el problema, los objetivos del proyecto y requisitos de la solución desde la perspectiva de la empresa.

Etapa 2: Enfoque analítico

Esta etapa implica expresar el problema anteriormente definido, y desde la perspectiva de los recursos técnicos identificar las técnicas más adecuadas para obtener el resultado deseado.

Etapa 3: Requisitos de datos

Para llevar a cabo el proyecto es necesario definir los requisitos en términos de contenidos de datos, formatos, representaciones, etc.

Etapa 4: Recopilación de datos

Esta etapa requiere identificar y reunir los recursos de datos disponibles. En caso de existir “lagunas” en la recopilación de datos, es necesario revisar los requisitos de datos y recopilar nuevos datos.

Etapa 5: Comprensión de datos

Luego de la recopilación de datos, se procede a hacer un análisis exploratorio de estos, utilizando estadística descriptiva y técnicas de visualización para comprender el contenido de los datos y evaluar su calidad. Si es que se encuentran vacíos es necesario volver a la etapa de recopilación de datos.

Etapa 6: Preparación de datos

En esta etapa ocurren actividades tales como, la limpieza de datos (eliminar duplicados, datos nulos, dar un formato adecuado, etc.), combinar datos de las distintas fuentes, transformar los datos, etc.

Etapa 7: Modelado

Esta etapa utiliza el conjunto de datos preparados anteriormente y se enfoca en desarrollar el modelo según el enfoque analítico definido en la etapa 2.

Etapa 8: Evaluación

Antes de la implementación del modelo, se realiza una evaluación con el objetivo de comprender su calidad y comprobar que aborda la problemática de manera óptima.

Etapa 9: Implementación

Cuando el modelo ya ha sido desarrollado satisfactoriamente y aprobado por la empresa, se implementa en un entorno de pruebas, limitadamente hasta que se haya evaluado completamente.

Etapa 10: Retroalimentación

Al recopilar los resultados del modelo implementado, se obtendrá una retroalimentación sobre el rendimiento de este y el impacto que generó en el entorno que se implementó. Es importante recalcar que en esta etapa se pueden hacer ajustes según la retroalimentación recibida

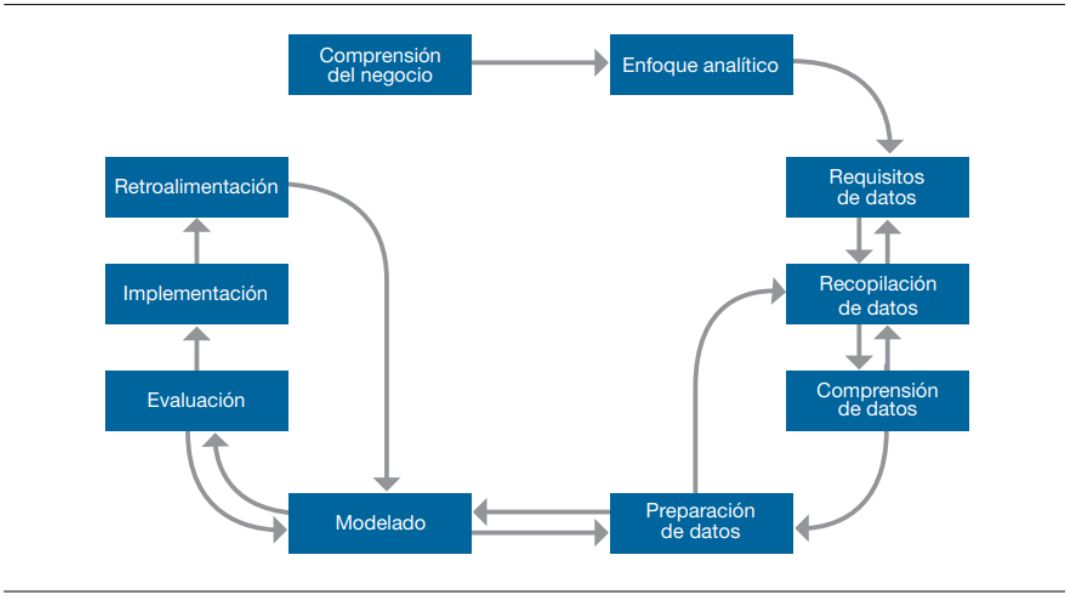


Imagen 2: Metodología Fundamental para Ciencia de Datos (IBM Analytics)

Medidas de desempeño

Las medidas o métricas de desempeño son fundamentales para la medición cuantitativa del proceso y el impacto en los resultados esperados una vez concluya el proyecto. Dado esto, a continuación, se definen las siguientes

medidas de desempeño, se explicará qué es lo que mide y cómo se calcula.

Para el objetivo general, se medirán las captaciones de tarjetas de crédito que sean atribuibles a las comunicaciones de las campañas comerciales. Esto se logrará mediante el seguimiento del ID del usuario, haciendo el cruce entre las bases de captaciones y la base de la campaña comercial dependiendo del canal de comunicación.

Para el primer objetivo específico, se medirá la tasa de conversión de la tarjeta de crédito, mediante la siguiente formula:

$$\%Tasa\ de\ conversión = \frac{N°\ de\ captaciones}{N°\ de\ clics\ banner} * 100$$

Por otro lado, se harán mediciones de la tasa de apertura de las distintas campañas comerciales dependiendo del canal de comunicación. Esto se hará mediante la siguiente fórmula:

$$\%Tasa\ de\ Apertura = \frac{N°\ de\ elementos\ \ abiertos}{N°\ de\ elementos\ enviados} * 100$$

Por último, se hará la medición de la cantidad porcentual de clientes que muestran interés (usuarios que hacen clic en algún banner relacionado a tarjeta de crédito dentro de la aplicación). Y el aumento porcentual respecto a la situación inicial. Se mide con la siguiente fórmula:

$$\%Tasa\ de\ interés(clic\ en\ banner) = \frac{N°\ de\ clics\ en\ banner}{N°\ de\ clientes\ ready\ que\ hicieron\ login} * 100$$

Planificación

A continuación, se presentará una carta Gantt con la respectiva planificación del proyecto y una tabla con los recursos que se usarán. Cabe destacar que para la realizar este proyecto, es necesario estar conectado a la VPN corporativa de Banco Ripley, con el fin de obtener los datos necesarios para el análisis.

N° Semana	36 Sem	37 Sem	38 Sem	39 Sem	40 Sem	41 Sem	42 Sem	43 Sem	44 Sem	45 Sem	46 Sem
Mes	septiembre	septiembre	septiembre	septiembre	octubre	octubre	octubre	octubre	octubre	noviembre	noviembre
Proyecto	4-sep	11-sep	18-sep	25-sep	2-oct	9-oct	16-oct	23-oct	30-oct	6-nov	13-nov
Definición y Levantamiento del problema											
Recopilación de Datos											
Limpieza y preprocesamiento de datos											
Analisis Exploratorio de Datos											
Estudio de Estado del Arte											
Segmentación de Usuarios											
Evaluación de Costos											
Evaluación de Riesgos											
Desarrollo de Estrategia de Campaña											
Implementación de Campaña BETA											
Evaluacion de Resultados											
Implementación de Campaña con correcciones											
Evaluación de Resultados											
Revisión y cierre del proyecto											

Tabla 8: Carta Gantt del proyecto

Plataformas
Jupyter
Dbeaver
Microsoft Office
Big Query (Google Cloud)

Lenguaje de Programacion
Python
SQL
Microsoft Excel

Tabla 9 y 10: Resumen lenguajes de programación y plataformas

Desarrollo

En términos de desarrollo y en congruencia con la metodología presentada anteriormente, la implementación del proyecto se ejecutó en 10 etapas. La primera etapa que se enfrenta es la de comprensión del negocio, donde se conocen los dolores de la empresa, se definen los objetivos y requisitos del proyecto, se profundiza en el funcionamiento del rubro, etc. Dado esto, se pudo deducir que el principal dolor que estaba pasando el banco neo digital Chek es la poca cantidad de captaciones obtenidas por la tarjeta de crédito, en comparación con las expectativas generadas prelanzamiento de este producto, por lo que se definió la problemática, los objetivos y los requisitos en base a este dolor. Luego de esta etapa, se pasa al enfoque analítico, donde se evaluaron las capacidades técnicas y recursos requeridos para abordar el proyecto según el problema identificado. Dentro de

algunos de estos son VPN, Big Query, Jupyter. Utilizando lenguajes como SQL o Python. Además de las distintas bases de datos que se recopilaron para la ejecución de este proyecto. Luego, en la etapa de requisitos de datos, se especifican los formatos, la información requerida para llevar a cabo el proyecto, entre otros. Dando paso a la etapa de recopilación de datos, etapa en la cual se realiza una búsqueda exhaustiva e identificación de bases de datos disponibles por la empresa, dentro de estas, la base de todos los clientes de Chek , la base de usuarios con oferta de tarjeta de crédito, la base de transacciones de los últimos 6 meses de los usuarios Chek y la base de transacciones en Ripley y ripley.com (tienda online y página web de Ripley) de usuarios con oferta de tarjeta de crédito Chek. Se evalúan posibles lagunas, datos erróneos, etc. Para esta etapa fue necesario estar en constante comunicación con el equipo de business intelligence, donde a través de la plataforma JIRA, se deben ir haciendo los pedidos correspondientes para que ellos lo clasifiquen según prioridad.

A continuación de la recopilación de datos, se procede a comprender estos.
En esta etapa, se realizó un análisis exploratorio de los datos aplicando técnicas de estadística descriptiva y visualización para comprender la calidad y la información que entregan los datos. al analizar las bases de datos, hacer los cruces entre estas, las limpiezas respectivas, etc.

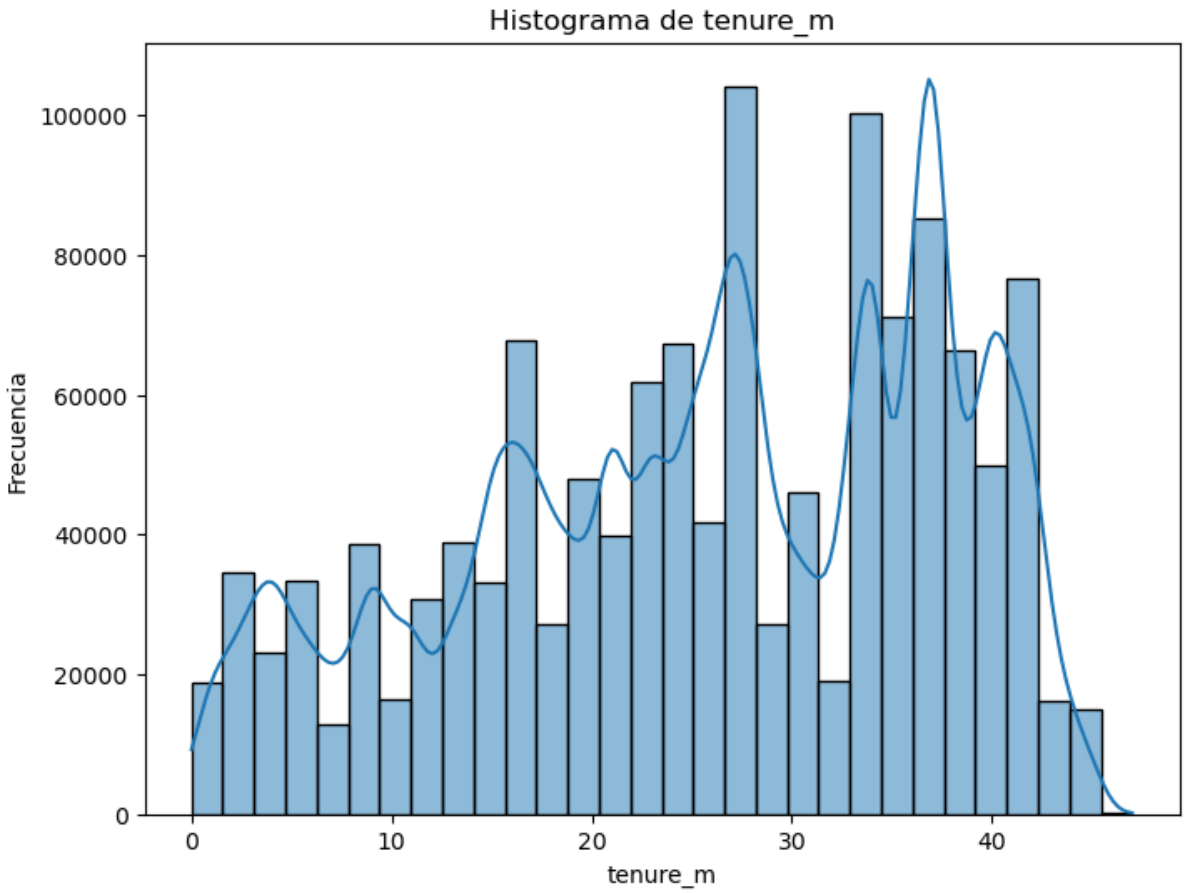


Imagen 6: Ejemplo EDA, histograma variable tenure (desde hace cuantos meses tiene instalada el usuario la aplicación Chek)

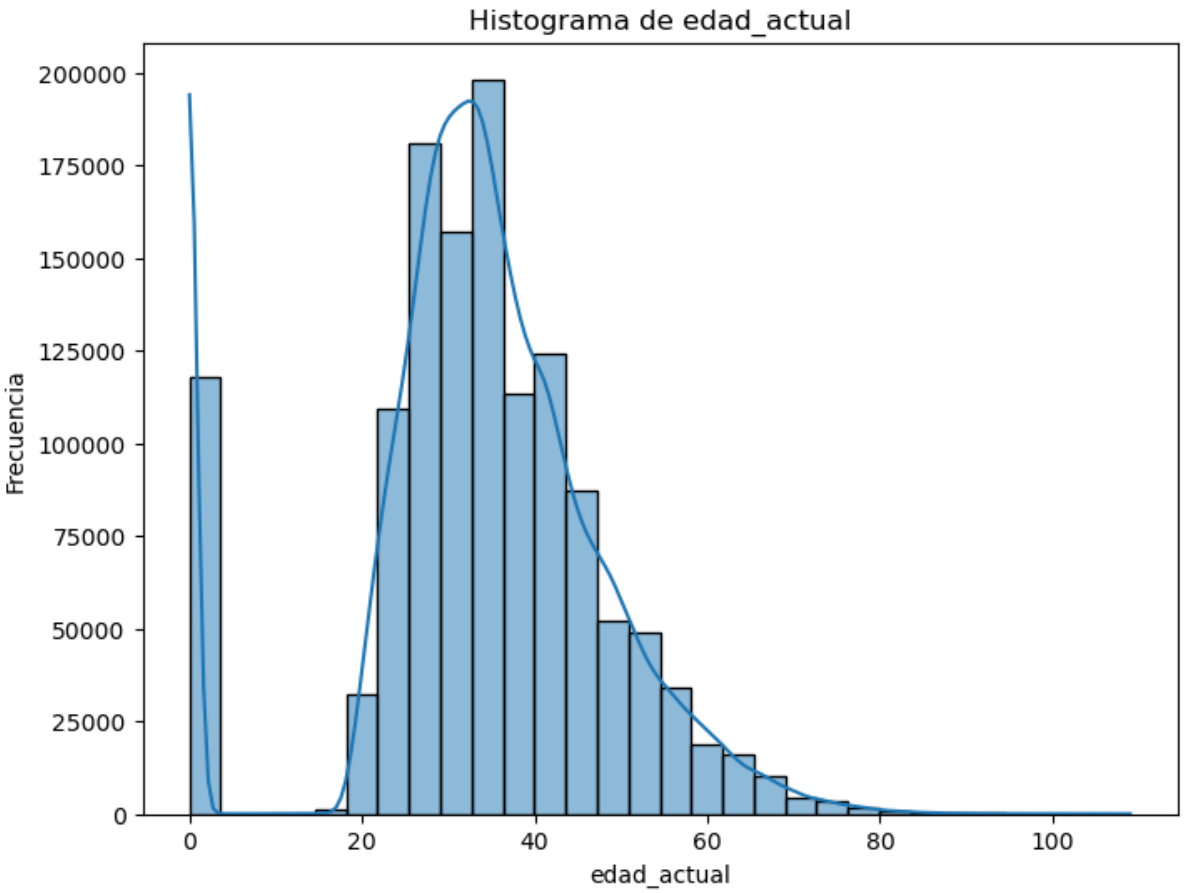


Imagen 7: Ejemplo EDA, histograma con la edad actual del usuario

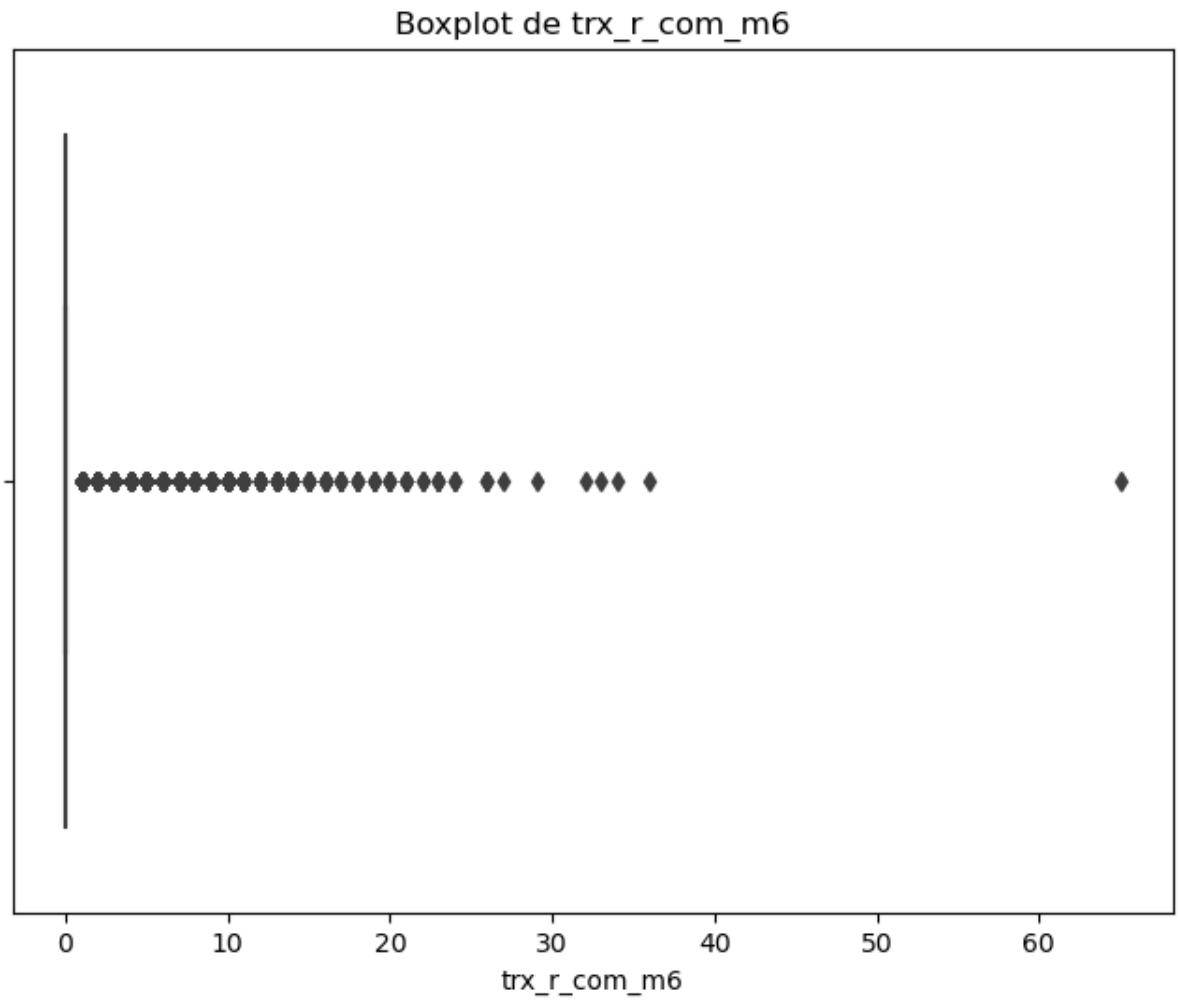


Imagen 3: Ejemplo EDA, Boxplot transacciones en ripley.com los últimos 6 meses

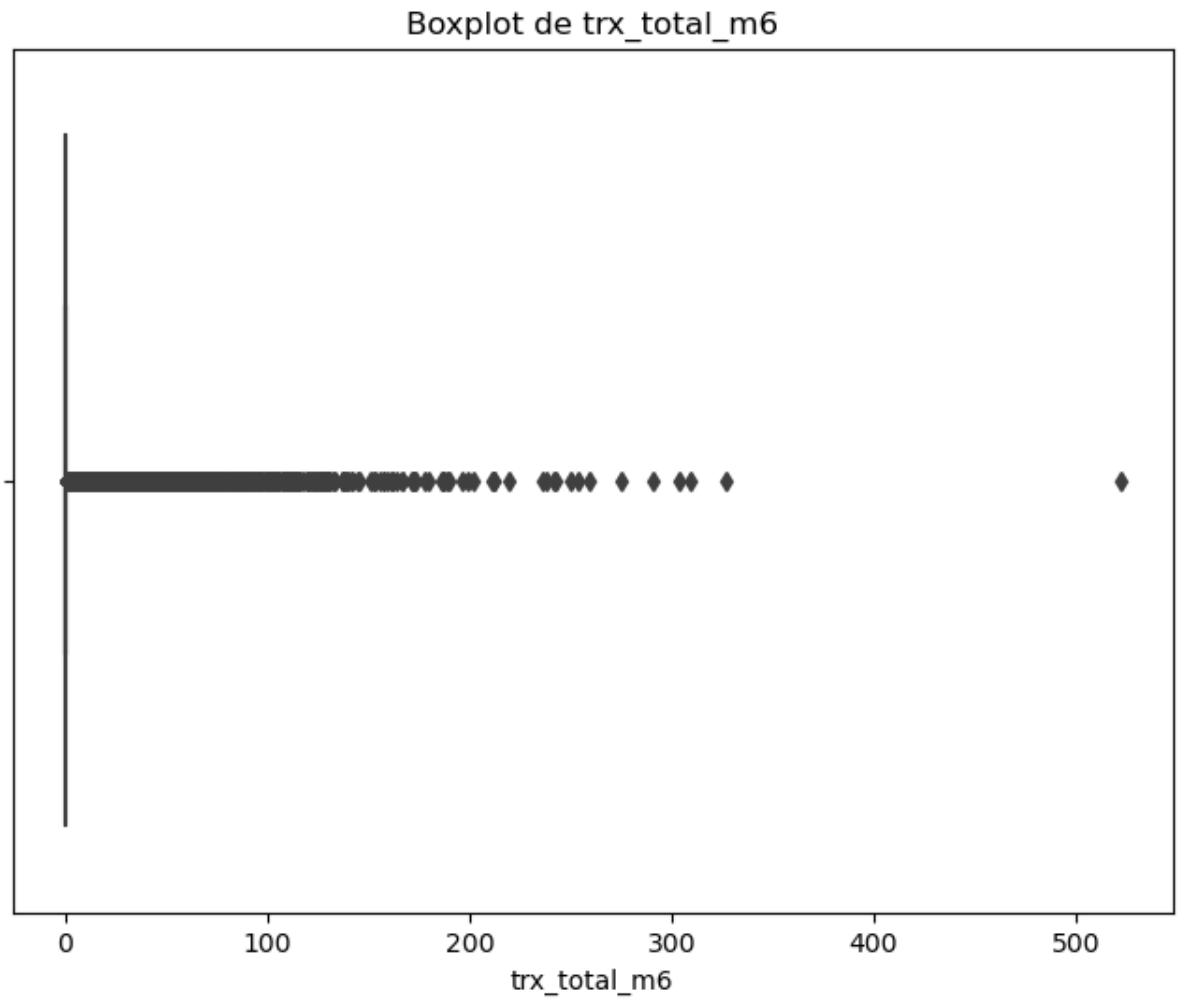


Imagen 4: Ejemplo EDA, Boxplot transacciones totales en los últimos 6 meses

Además, se analizó exhaustivamente las bases de datos obtenidas, es decir, la base de datos de todos los usuarios de Chek, la base de usuarios con oferta de tarjeta de crédito, la base de transacciones de clientes Chek y la base de transacciones en Ripley y ripley.com de usuarios con oferta, observando la distribución de sus variables, la relevancia de la información que entregan, la relación entre estas, entre otros descubrimientos. Por ejemplo, dentro de algunas conclusiones que se pueden sacar tras haber superado estas etapas, se definieron las variables categóricas y numéricas, la distribución de los rangos de edad, la distribución de los grupos socioeconómicos, la cantidad de hombres y mujeres dentro de la base, los rubros más consumidos por los usuarios, la cantidad de transacciones efectuadas por los

usuarios con la tarjeta Chek, el tiempo desde que descargaron la aplicación los usuarios, entre otros. Los principales hallazgos se encuentran retratados visualmente a continuación.

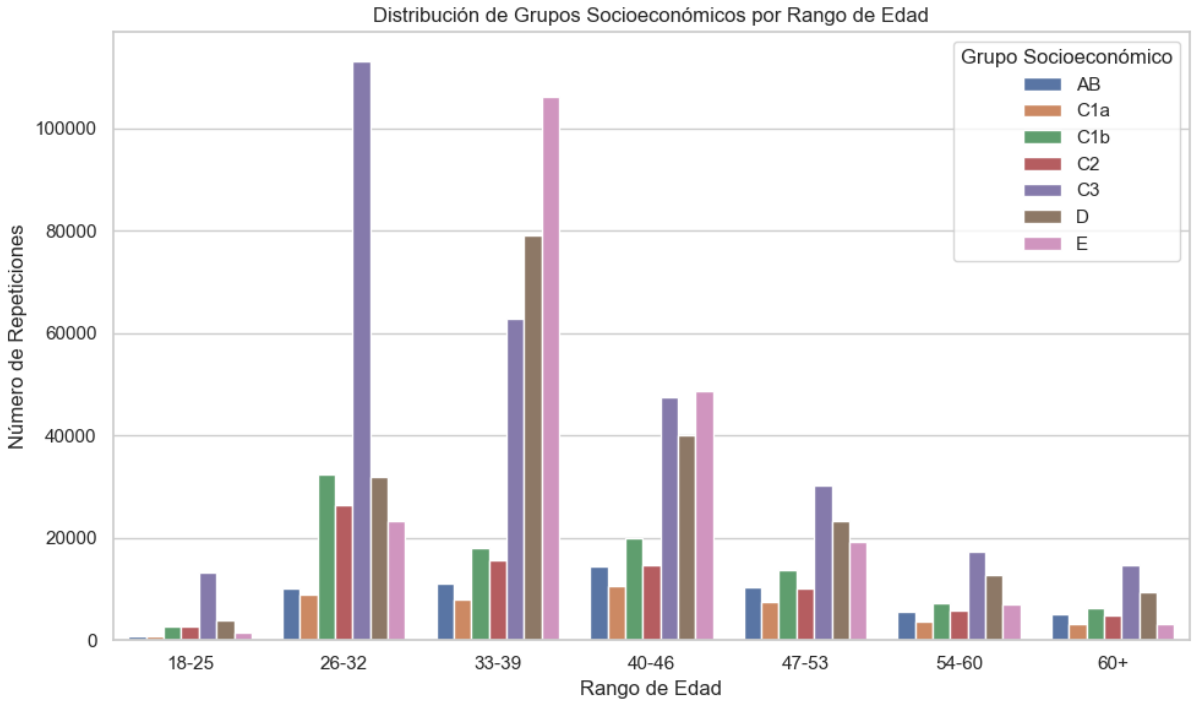


Imagen 5: Distribución de Grupos socioeconómicos por Rango de edad

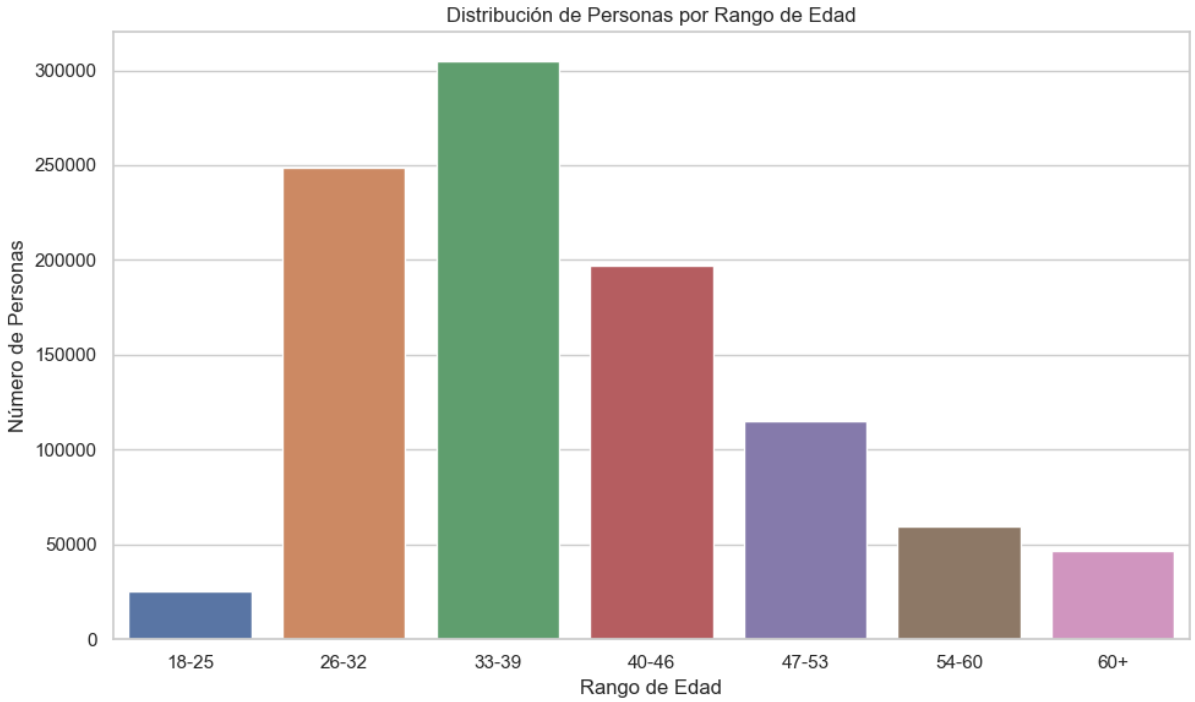


Imagen 6: Distribución de Personas por Rango de edad

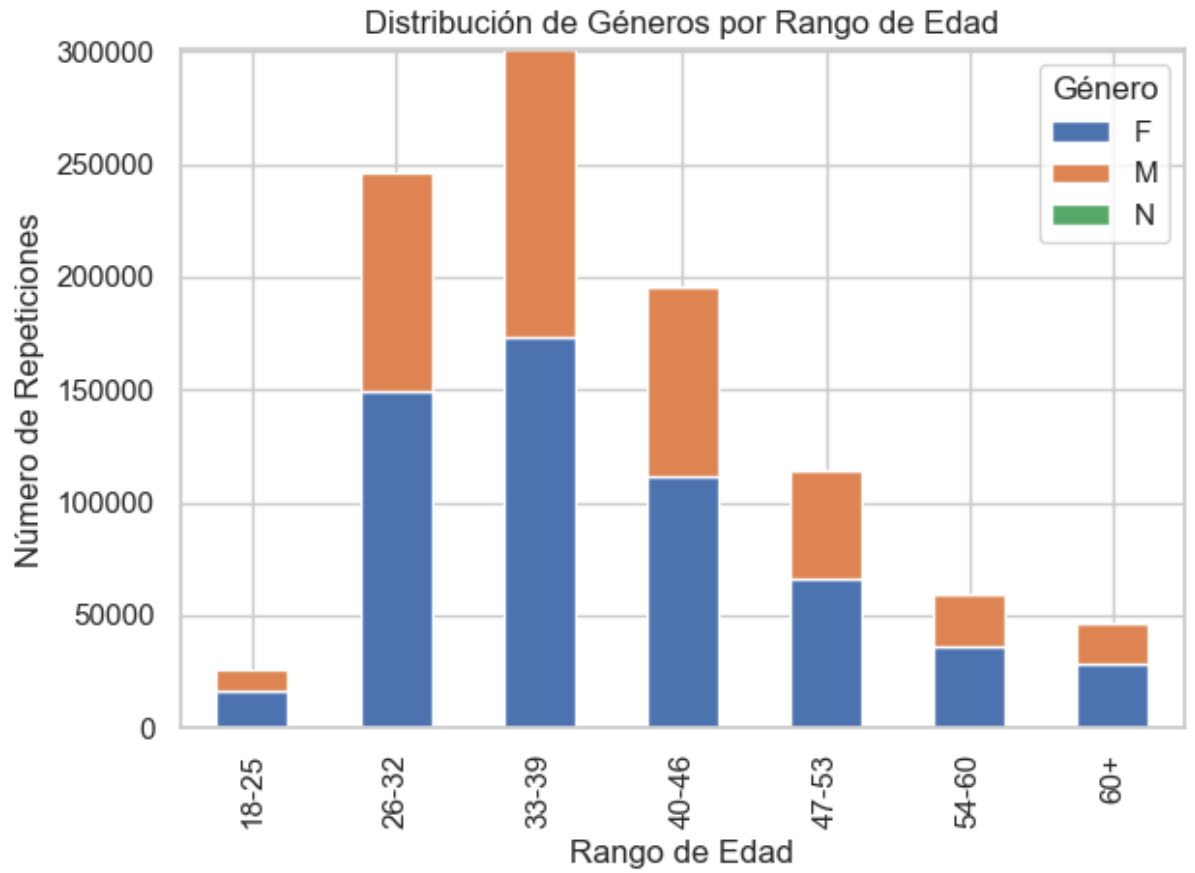


Imagen 7: Distribución de géneros por rango de edad

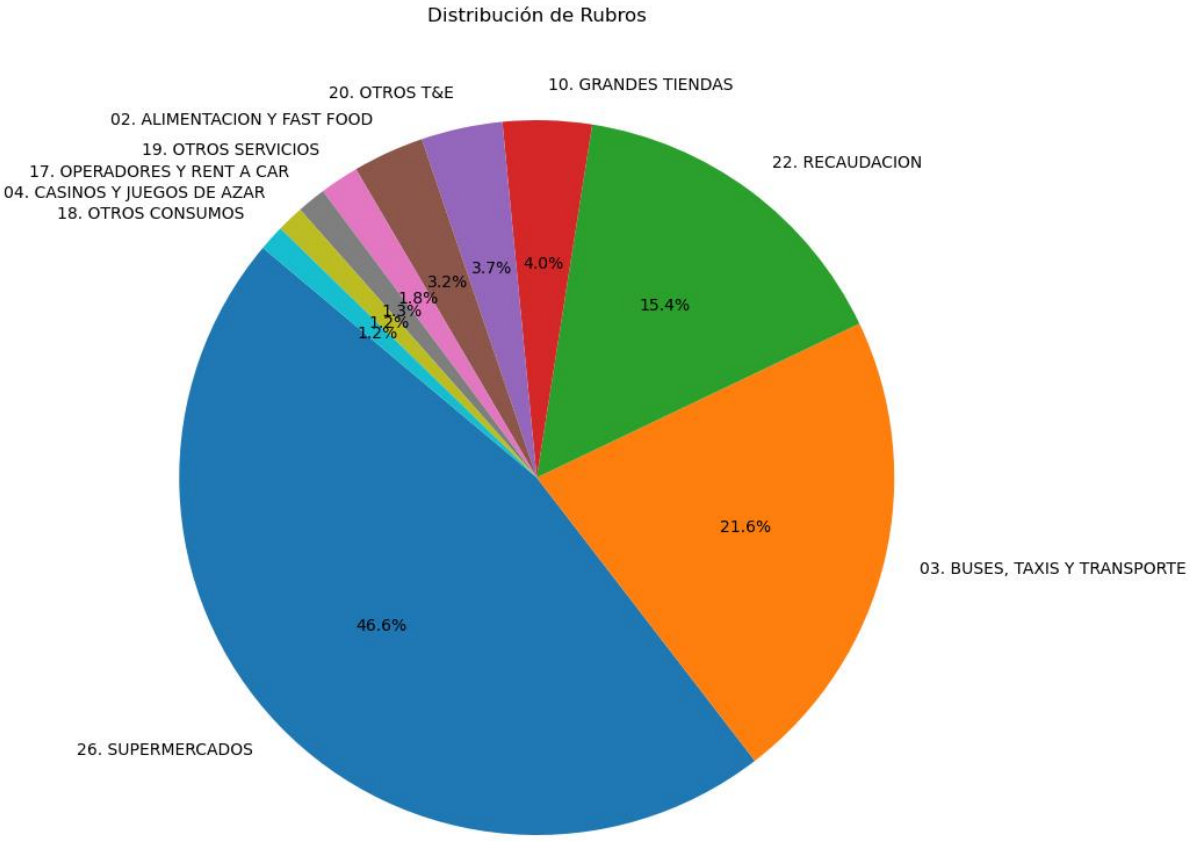


Imagen 8: Distribución de Rubros concurridos por los usuarios de Chek

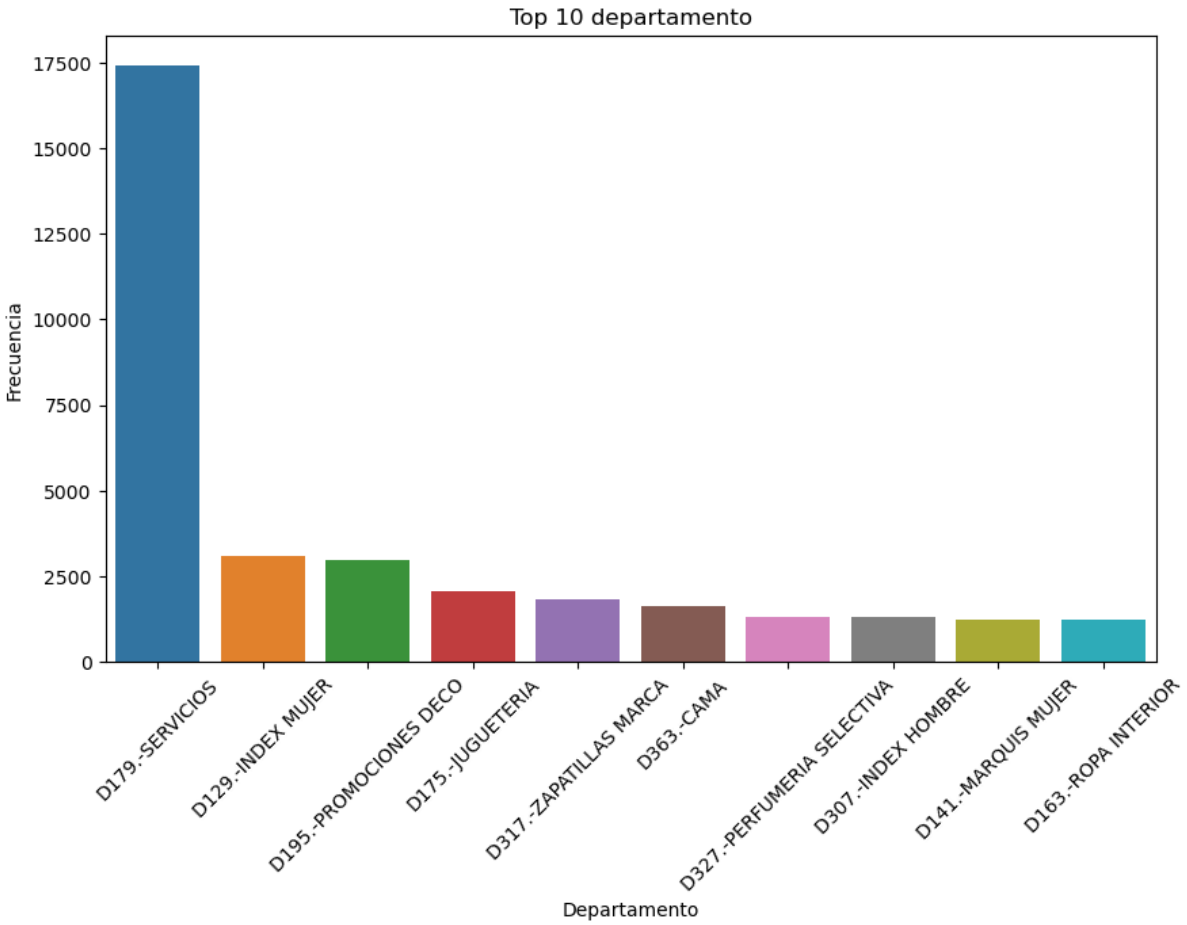


Imagen 9: Distribución de Departamentos de Ripley más concurridos por los usuarios de Chek

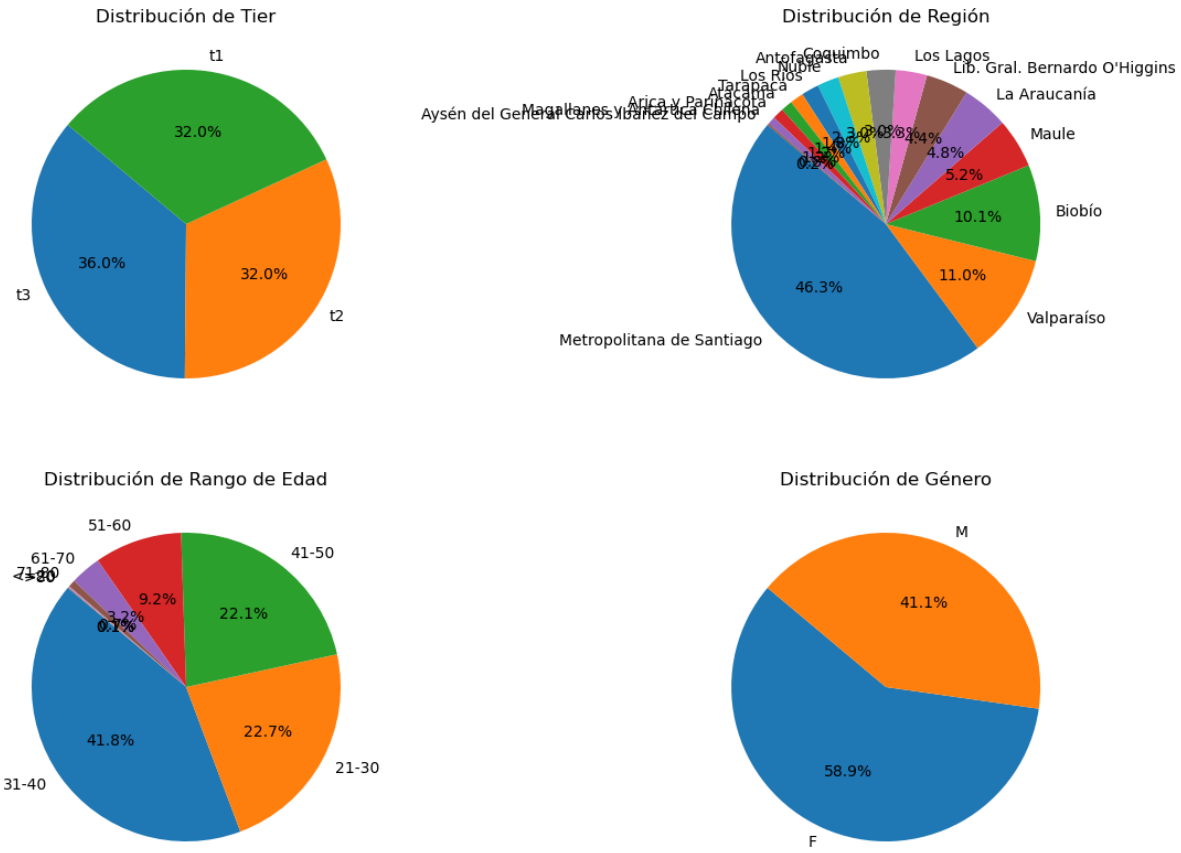


Imagen 10: Distribución de género, rango de edad, región y tier.

Tras comprender las distintas bases de datos, anteriormente nombradas, se procedió a prepararlas para la etapa de modelado, se eliminaron datos duplicados, nulos, se estandariza el formato, se eliminan incongruencias de la base de datos. Además, se combinan las diferentes bases y se transforman los datos para crear una base de datos coherente y preparada para el modelamiento.

En primera instancia para la etapa de modelado, se debió utilizar el método de “información mutua”, el cual es un algoritmo de selección de variables, especializado en modelos predictivos de clasificación, este evalúa la relación entre las variables y la variable objetivo (la variable binaria “captados”, 0 si no capta la tarjeta de crédito y 1 si capta la tarjeta de crédito). Los resultados de este algoritmo se muestran en la tabla 24, donde podemos ver que hay una congruencia con las variables como, edad, tenure (hace cuantos meses tiene la aplicación Chek el usuario), ingreso promedio, región, entre otras.

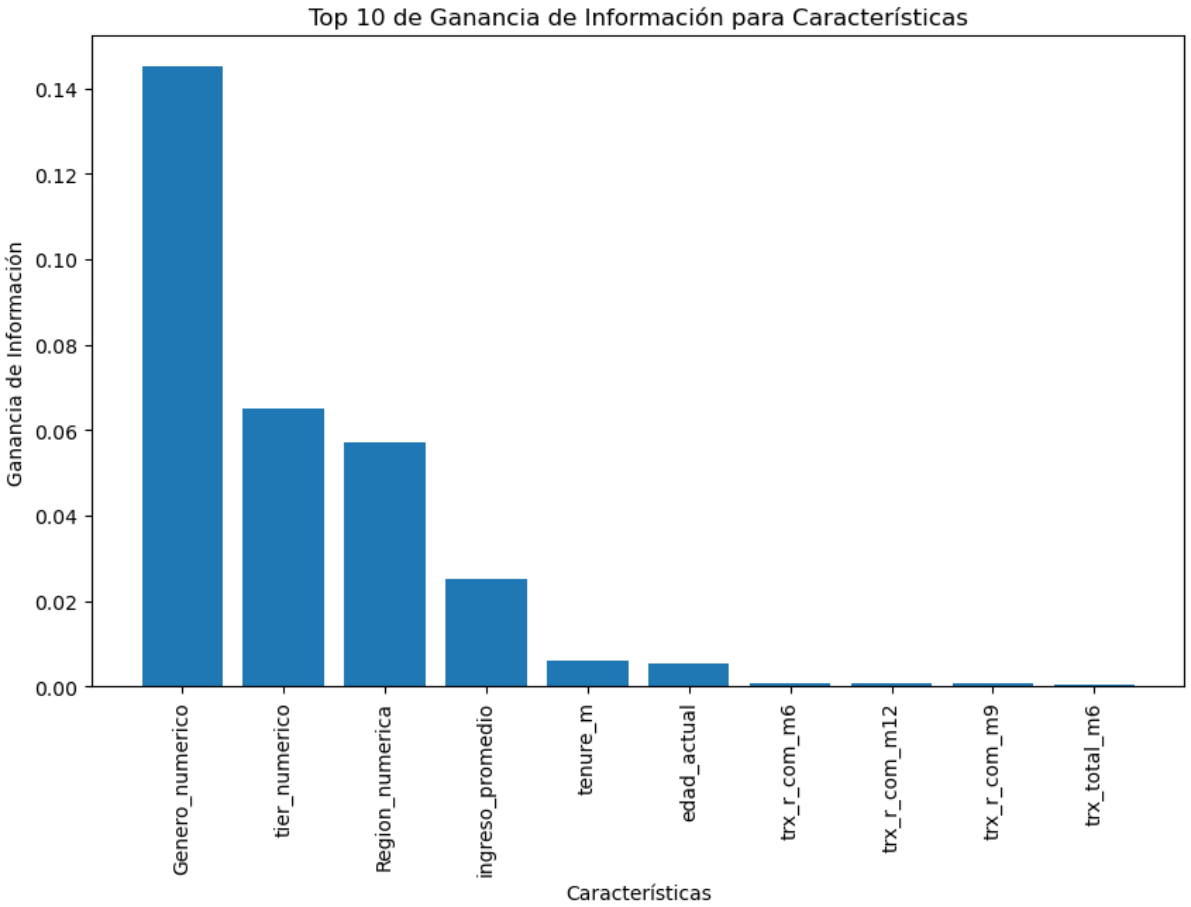


Imagen 11: Selección de variables

Tras esto, se procedió a ejecutar exhaustivamente distintos algoritmos de modelos predictivos de clasificación, tales como árboles de decisión, regresión logística, máquinas de vectores de soporte (SVM) y K-Vecinos (K-NN) con el fin de lograr los resultados óptimos, siendo el algoritmo de regresión logística el idóneo para resolver este problema. Dado los primeros resultados (según las métricas accuracy, precisión, f1-score y recall, no era idóneo), se deduce que se debe hacer uso de un modelo predictivo desbalanceado.

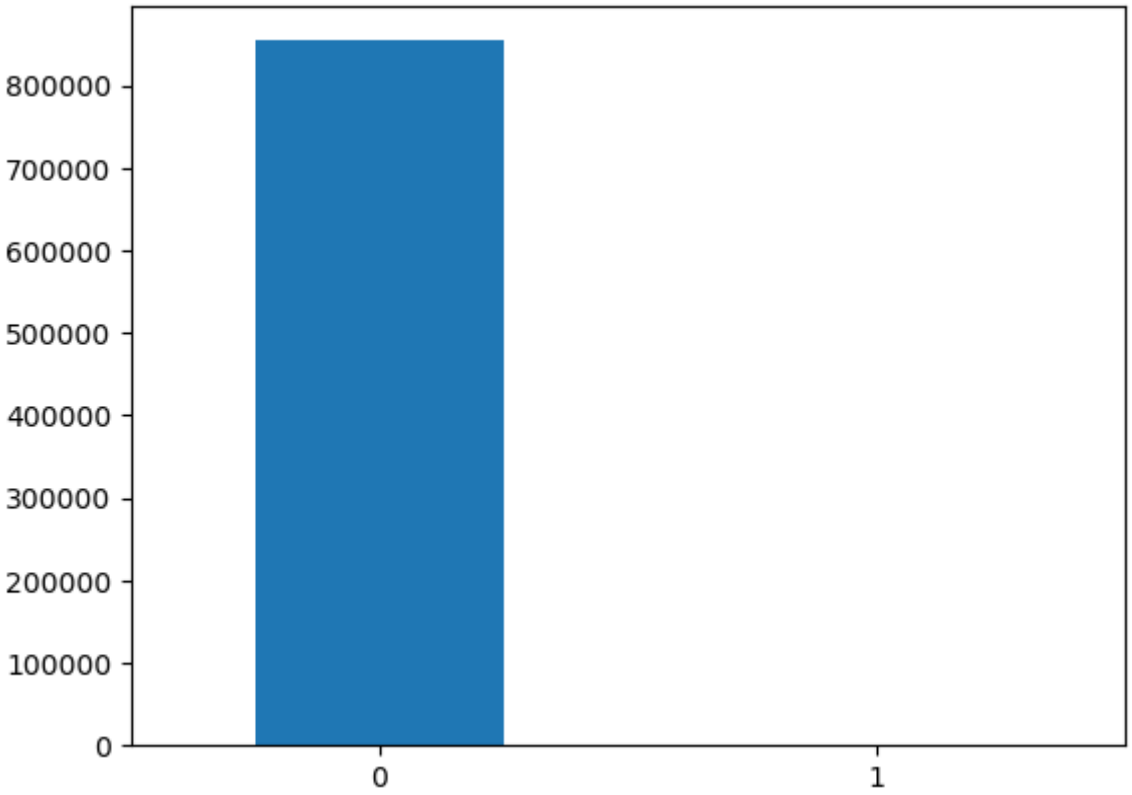


Imagen 12: Gráfico de captados (1) vs No captados (0)

	precision	recall	f1-score
0	1.00	1.00	1.00
1	0.00	0.00	0.00
accuracy			1.00

Imagen 13: Resultados modelo de clasificación pre-balanceo

Dado esto, se aborda el modelo desbalanceado, aplicando distintos algoritmos, tales como, penalización para compensar, resampling (oversampling y subsampling, SMOTE y Balanced Bagging Classifier, siendo este último el que dio mejores resultados. Es necesario decir, que se utilizaron las métricas de precisión, recall, f1 score y cross validation (accuracy) para medir los resultados de los modelos de clasificación, aunque esta última no sea tan efectiva para modelos desbalanceados como en este caso. Estas medidas se interpretan de la siguiente forma alta precision y alto recall significa que el modelo predice perfectamente las clases (minoritarias y mayoritarias), alta precision y bajo recall significa que el modelo no detecta la clase muy bien pero cuando lo hace es

altamente confiable, baja precision y alto recall significa que el modelo detecta bien las clases pero es poco confiable en la predicción, y por ultimo baja precision y bajo recall significa que el modelo no logra clasificar las clases correctamente. A continuación, se muestra una tabla con los resultados obtenidos por los modelos desbalanceados.

	precision	recall	f1-score
0	1.00	0.85	0.92
1	0.42	0.87	0.56
accuracy			0.85

Imagen 14: Resultados modelo de clasificación post balance

Algoritmo/Métrica	Variable	Precision	Recall	F1-Score	Accuracy
Balanced BaggingClassifier	0	1	0.95	0.92	0.85
	1	0.42	0.87	0.56	
Penalización para compensar	0	1	0.98	0.99	0.78
	1	0.07	0.93	0.13	
Subsampling	0	1	0.84	0.92	0.75
	1	0.01	0.93	0.02	
Oversampling	0	1	0.99	0.99	0.79
	1	0.14	0.89	0.24	
SMOTE	0	1	1	1	0.71
	1	0.28	0.85	0.32	

Tabla 11: Resultados de los distintos modelos de clasificación post balance; 0 es no capta y 1 es capta tarjeta de crédito.

Debido a este descubrimiento, se logra mejorar considerablemente el resultado de las métricas anteriormente descritas, por lo que el modelo se encuentra disponible para la primera etapa de implementación (piloto), donde se generó una campaña genérica (que se comunicó a través de una notificación push) para el 20% de la base de datos de aquellos usuarios que fueron clasificados como “propensos a conseguir la tarjeta de crédito “con el objetivo de hacer mejoras y evaluar el rendimiento del modelo.

Tras conseguir los resultados del piloto y recibir el feedback y/o recomendaciones, se procede a ejecutar algoritmos de modelos de agrupamiento

con el fin de segmentar estas bases, y de esta forma, segmentar las campañas comerciales. Tras varios intentos y verificación de resultados, el óptimo se encontró utilizando el método de KMeans, el que dio 3 clústeres (utilizando el método del codo), y verificando con el coeficiente de Silhoutte, el cual dio un resultado de 0,78 (78%), por lo que los clústeres estarían bien definidos. Dado esto, se generaron banners, notificaciones push, envíos de WhatsApp y mail para los distintos “clústeres”, cada canal de comunicación con un mensaje personalizado para cada grupo objetivo. A continuación, un gráfico que demuestra el monto y la frecuencia de compra de los clústeres en tiendas Ripley y ripley.com. Además de las características para cada clúster.

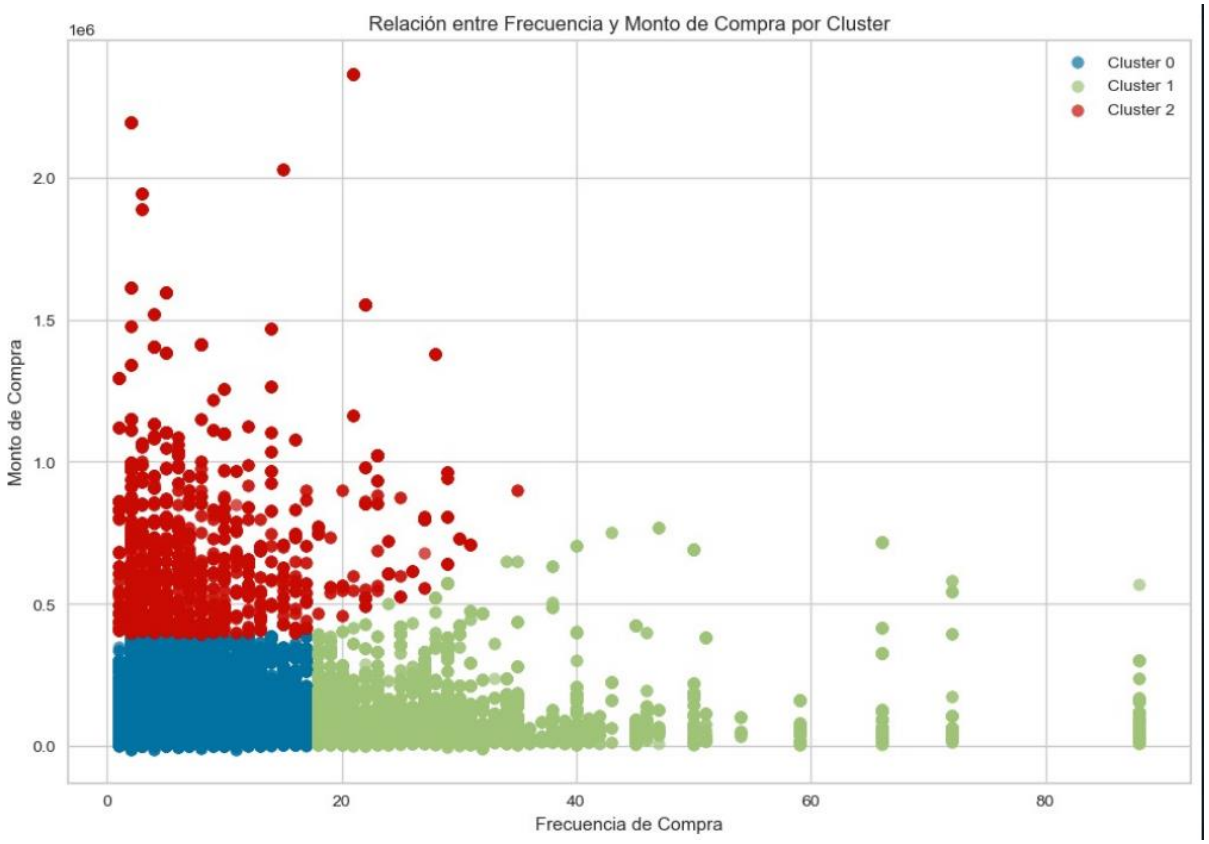


Imagen 15: Gráfico de frecuencia de compra vs monto de compra de los usuarios.

	frecuencia	monto_dep	frecuencia_dep
0	9.071995	1.914987e+07	1572.632281
1	9.259678	2.495401e+07	17417.412533
2	9.151515	9.753759e+07	4538.214578

Silhouette Score: 0.78327307122697847

Imagen 16: Ejemplo Resultados del modelo de agrupamiento

Cluster	edad_actual	Genero_numerico	departamento_num	tenure_m	\
0	34.008923	1.367666	155.357872	26.032451	
1	38.988889	1.376878	198.538224	21.988126	
2	36.000002	1.377023	197.733006	25.966767	

Cluster	suma_monto	ingreso_promedio	monto	Region_numerica
0	124860.044073	869888.328908	7487.183515	13.002889
1	157385.644559	902789.899023	14772.041852	12.989021
2	577692.061425	899643.892312	390959.615889	13.006540

Imagen 17: Ejemplo características de los clústeres

	cluster	departamento	frecuencia
0	0	D417.-DECORACION	15.000000
1	0	D130.-INSTRUMENTOS MUSICALES	12.000000
2	0	D384.-CHECK OUT TECNOLOGIA	9.166667
3	0	D331.-ROPA INTERIOR ESCOLAR	8.750000
4	0	D107.-CALZADO ESCOLAR	8.333333
5	0	D398.-MARCAS NACIONAL NINO	7.828571
6	0	D102.-ALFOMBRA	7.705882
7	0	D192.-BICICLETAS Y MAQUINAS	7.476190
8	0	D433.-HERRAMIENTAS Y MAQUINARIAS	7.472222
9	0	D161.-RODADOS	7.327869
10	1	D175.-JUGUETERIA	44.000000
11	1	D328.-PERFUMERIA	38.625000
12	1	D172.-VIDEOJUEGOS	38.090909
13	1	D171.-TV-VIDEO	37.750000
14	1	D314.-TEXTIL MAS	34.304348
15	1	D397.-LICENCIAS NINO	34.153846
16	1	D198.-PROMOCIONES INFANTIL	33.409091
17	1	D367.-COMPLEMENTOS DECO	33.333333
18	1	D338.-REGATTA HOMBRES	33.200000
19	1	D320.-CACHAREL MUJER	33.071429
20	2	D129.-INDEX MUJER	15.000000
21	2	D307.-INDEX HOMBRE	15.000000
22	2	D402.-MARCAS NACIONAL BEBES	15.000000
23	2	D314.-TEXTIL MAS	11.736842
24	2	D176.-PROPIA RECIEN NACIDO	11.684211
25	2	D103.-AUDIO	11.250000
26	2	D420.-NUTRICION	10.750000
27	2	D151.-NAVIDAD	9.809524
28	2	D320.-CACHAREL MUJER	9.580645
29	2	D200.-CUIDADO PERSONAL	9.529412

Imagen 18: Ejemplo “departamentos” más frecuentados por clúster

Para la creación de las campañas, se identificaron las preferencias y/o intereses de los 3 clústeres, según “departamento” más concurrido en las tiendas Ripley. Dando como resultados, “Index Mujer (vestuario mujer)” (clúster 2), “Index Hombre (vestuario hombre)” (clúster 2), “Juguetería” (clúster 1), “Perfumería” (clúster 1) y “Decoración” (clúster 0).

Dado esto, se generaron banners (se pidieron las gráficas a la diseñadora de Chek) dentro de la aplicación y comunicaciones personalizadas para cada clúster

a través de WhatsApp, Mail y notificaciones push, ofreciendo incentivos monetarios para que capten la tarjeta de crédito Chek.

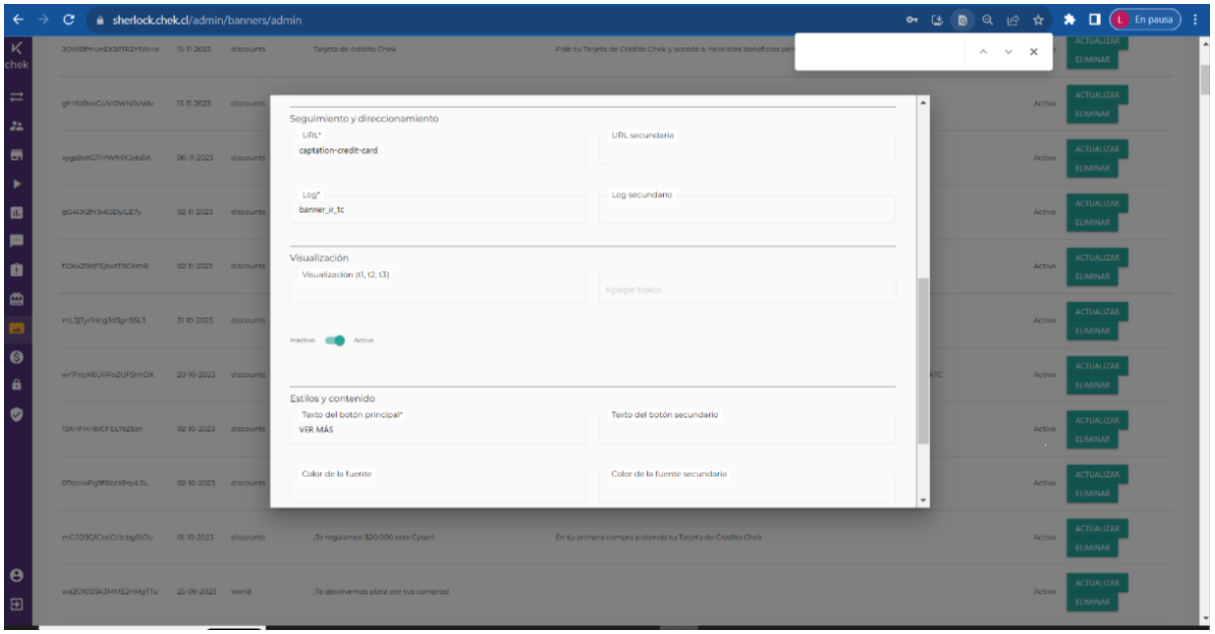


Imagen 19: Ejemplo creación banner en plataforma Sherlock

Imagen 20: Ejemplo banner "perfumería" dentro de la aplicación.

Para analizar los resultados, fue necesario hacer seguimiento de los distintos canales de comunicación, en particular, para las notificaciones push y los banners, las cuales tienen un ID determinado, se debió hacer uso de la plataforma

Big Query, donde a través de un código de SQL, y posteriormente cruzar (a través de la fórmula de Excel BUSCARX) con las bases de las campañas, se pudo verificar si las altas corresponden a alguna de las campañas comerciales generada en en proyecto. Para WhatsApp y Mail, envían una planilla desde el área de Business Intelligence con los reportes de las campañas.

Resultados

En primer lugar, decir que los resultados del proyecto contribuyeron positivamente en la gestión de la relación del banco neo digital Chek con sus usuarios, otorgando una nueva estrategia para la generación y análisis de campañas comerciales, provocando un mayor interés de parte de los clientes de la empresa.

En términos de resultados, de acuerdo con el desarrollo del proyecto explicado anteriormente, para el primer objetivo de aumentar la tasa de conversión de tarjeta de crédito Chek, se logró obtener una tasa de conversión promedio de 4% para las campañas comerciales generadas en este proyecto, superando considerablemente la meta de aumento de 1%.

Por otro lado, el segundo objetivo específico de lograr una tasa de apertura de al menos un 5% para las notificaciones push, debido a que estas son el canal de comunicación más débil (en la situación inicial tenían una tasa de apertura promedio de 1%), es posible decir que se logró cumplir con el objetivo propuesto para cada campaña, incluso superando por 3 puntos porcentuales en la campaña relacionada a juguetería. Respecto al resto de canales utilizados se mantuvieron estables en relación con la tasa de apertura promedio obtenidas respecto a la situación inicial. En promedio para las notificaciones push se obtuvo una tasa de apertura de 6% una mejora considerable respecto a la situación inicial.

Metrica/Campaña	JUGUETERIA	DECORACION	INDEX HOMBRE	INDEX MUJER	PERFUMERIA	PROMEDIO
Tasa Push	8%	5%	6%	6%	7%	6%
Tasa Whatsapp	66%	NA	NA	NA	66%	66%
Tasa Mail	NA	NA	36%	36%	NA	36%

Tabla 12: Tasa de apertura para las campañas comerciales

Además, para el tercer objetivo específico de aumentar el interés (clics en banner relacionados a la tarjeta de crédito) de los usuarios de Chek, se propuso un aumento de 5 puntos porcentuales, el cual en la situación inicial promediaba un 17%. A continuación, se mostrará en una tabla con datos obtenidos a través de la plataforma Big Query, el interés generado por cada banner perteneciente a las campañas comerciales efectuadas por el proyecto. También se presenta un funnel donde se muestra los usuarios que entraron a la aplicación durante el periodo de las campañas comerciales generadas por el proyecto, cuantos de estos usuarios cuentan con oferta de tarjeta de crédito, cuántos son clientes ready (cumplen con los requisitos para obtener la tarjeta de crédito) y cuantos usuarios mostraron interés (hicieron clic en algún banner relacionado a tarjeta de crédito). Es posible decir, que se obtuvo una cantidad de 3002 clics en banners relacionados a las campañas comerciales efectuadas por el proyecto, lo que representa un 25% respecto a los clientes ready, logrando superar las metas propuestas por la empresa. En la siguiente tabla y funnel es posible visualizar los resultados para los objetivos específicos de aumentar la tasa de conversión e interés.

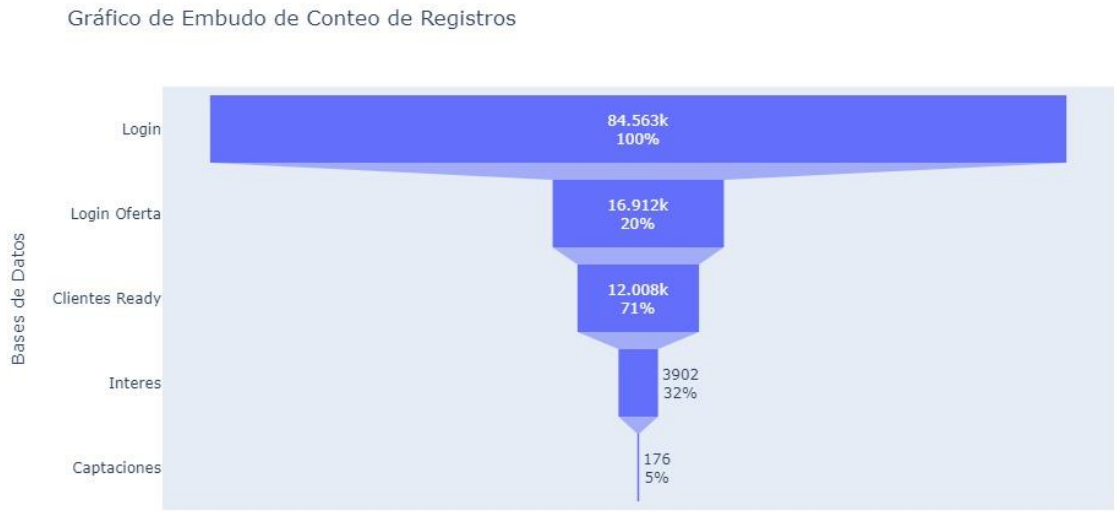


Imagen 32: Funnel de la campaña comercial generada por el proyecto.

Metrica/Campaña	JUGUETERIA	DECORACION	INDEX HOMBRE	INDEX MUJER	PERFUMERIA	TOTAL
Interés	675	525	598	579	625	3002
Captaciones	27	21	24	23	25	120
Tasa Conversión	4,00%	4,00%	4,01%	3,97%	4,00%	4,00%

Tabla 13: Interés, captaciones y tasa de conversión por campaña.

Finalmente, se concluye que, debido al cumplimiento de los 3 objetivos específicos nombrados anteriormente, por consecuencia, se logró cumplir con el objetivo general de mejorar el rendimiento y aumentar la cantidad de conversiones en 100 unidades de la tarjeta de crédito del banco neo-digital Chek mediante una estrategia de segmentación de usuarios efectiva. Obteniendo resultados mejor a lo esperado para los distintos objetivos específicos, y, por ende, en el objetivo general de contribuir con 100 captaciones de tarjetas de crédito, se lograron 120 captaciones atribuibles a las campañas comerciales generadas en el proyecto. Es necesario decir que para las campañas de juguetería y perfumería (enviadas a los usuarios pertenecientes al “clúster 1”) se utilizaron los canales de comunicación de notificaciones push, banners dentro de la aplicación y WhatsApp, para las campañas de vestuario hombre y vestuario mujer (enviadas a los usuarios pertenecientes al “clúster 2”) se utilizaron los canales de comunicación de notificaciones push, banners dentro de la aplicación y mail, y para la campaña de “Decoración” (enviada a los usuarios pertenecientes al “clúster 0” se utilizaron los canales de comunicación de notificaciones push y banners dentro de la aplicación.

Es destacable, que las métricas esperadas para cada 1 de las 5 campañas era como mínimo captar 20 usuarios de tarjeta de crédito, superando las expectativas en cada una de estas. Siendo las campañas de “juguetería” y “perfumería” las que tuvieron mejor performance, 135% y 125% respectivamente versus la meta propuesta. Se hace necesario decir, que para validar los resultados de las campañas del proyecto se utilizó un grupo de control del 10% de la base de usuarios con oferta para el cual no se les comunicarían campañas comerciales. Los resultados del grupo de control en comparación con el periodo de las campañas son de 7 captaciones de usuarios de tarjeta de crédito.

Metrica/Campaña	JUGUETERIA	DECORACION	INDEX HOMBRE	INDEX MUJER	PERFUMERIA	TOTAL
Captaciones	27	21	24	23	25	120
Meta	20	20	20	20	20	100
Cumplimiento	135%	105%	120%	115%	125%	120%

Tabla 14: Captaciones, metas y cumplimiento de las campañas.

Conclusiones

Para concluir, la implementación de este proyecto conllevó comprender el rubro bancario y cómo se desenvuelve la empresa dentro de esta, analizando distintas bases de datos, comportamientos de los clientes, con el fin de lograr un resultado óptimo del objetivo propuesto de generar una cantidad adecuada de captaciones de tarjetas de crédito Chek. Dado los resultados anteriormente expuestos, es posible decir que la realización de este proyecto dará espacio para que exista un análisis más profundo y detallado de los clientes del banco neo digital. De esta forma, la empresa, y en específico, el área comercial donde se desarrolló este proyecto tiene la posibilidad de crear campañas comerciales más personalizadas, generando mayor interés por parte de los usuarios de Chek, y por consecuencia, aumentar el stock de clientes de tarjeta de crédito de la empresa.

Cabe destacar, que, para cumplir con el objetivo planteado por el proyecto, se tuvieron que afrontar distintas adversidades, tales como, la recopilación de datos trabajando de la mano con el área de business intelligence (BI) de la empresa, donde muchas veces los datos traen errores, lagunas, entre otros problemas. Por otra parte, la colaboración y comunicación con el equipo del área comercial cumplió un rol fundamental dentro del proyecto, siendo necesario para crear los banners (diseñadora del área comercial), crear la campaña comercial con una comunicación efectiva, conseguir las bases de las altas de la tarjeta de crédito Chek para hacer el seguimiento y concluir si es que las altas eran atribuibles a alguna de las segmentaciones generadas en el proyecto, etc.

De esta forma, el proyecto impactó el área comercial de la empresa en relación con la metodología que se usa para generar las campañas comerciales, brindado así una nueva estrategia para crear estas, haciendo más eficiente la comunicación con los usuarios, logrando comprender los distintos intereses y por ende mejorando la gestión de la relación con los clientes.

Dentro de los aprendizajes rescatables de este proyecto, está la colaboración y comunicación con el equipo, lo cual fue fundamental para cumplir cada objetivo

exigido por la empresa. Por otra parte, la planificación del proyecto contribuyó para cumplir con los plazos exigidos y propuesto por parte de la empresa. Otro aprendizaje es el cómo se abordó el problema, en base a las necesidades de la empresa, se implementó una solución pertinente que cumpliera con las expectativas de la empresa, permitiendo así desarrollar el proyecto que resuelve la problemática planteada.

Es posible decir, que los objetivos específicos propuestos en este proyecto se pueden clasificar como resueltos, ya que se superó la meta de captación de usuario de tarjetas de crédito, superando las expectativas en 20 unidades, se aumentó a un promedio de 4% la tasa de conversión de tarjetas de crédito, se mejoró la tasa de apertura de las notificaciones push,promediando un 6% de tasa de apertura, superando el 5% propuesto en los objetivos, y aumentando el interés en la tarjeta de crédito (clics generados por los banners relacionados a alguna de las campañas comerciales), logrando 3002 clics en total, dando una tasa respecto a los “clientes ready” de 25% (la meta era llegar a una tasa de 22%).

Finalmente, hacer las siguientes recomendaciones a la empresa.

En primer lugar, la ampliación de datos, explorando la posibilidad de incluir información adicional que pueda influir en el comportamiento de los usuarios, ya sea, interacciones en plataformas digitales, comportamientos de compra o preferencias específicas.

Por otro lado, la inclusión de datos temporales, es decir, recopilar información sobre tendencias y/o cambios estacionales que afectan el patrón de comportamiento de los usuarios.

Por último, fomentar la participación de los usuarios, implementando estrategias para obtener información directamente de los clientes, ya sea, encuestas, focus group, comentarios, etc. Con el fin de enriquecer la comprensión de los usuarios.

Anexos

```python
#Revisamos los datos en primera instancia
df.head(10)
# Crear una nueva columna 'captaron_tarjeta' en df_ e inicializarla con 0
df['captaron_tarjeta'] = 0
# Marcar como 1 aquellos usuarios que están en df_captaron
df.loc[df['rut'].isin(altas['rut']), 'captaron_tarjeta'] = 1
#Descripción de la columna captaron_tarjeta
df['captaron_tarjeta'].describe()
df['captaron_tarjeta'].value_counts()
#Descripción de la base
df.info()
df.nunique()
#Conteo de nulos
df.isnull().sum()
#Eliminamos la variable nationalid ya que es la que contiene mas nulos y además esá repetida con la columna rut
df=df.dropna(subset='nationalid')
#Vemos las variables categoricas y las variables numericas
cat_cols=df.select_dtypes(include=['object']).columns
num_cols = df.select_dtypes(include=np.number).columns.tolist()
print("Variables Categoricas:")
print(cat_cols)
print("Variables Numericas:")
print(num_cols)
#Rellenar valores nulos para las variables numericas con un valor especifico 0
df=df.fillna(0)
print(df.isnull().sum())
# Seleccionar solo las variables numéricas
df_numerico = df.select_dtypes(include='number')

# Aplicar describe() a las variables numéricas
descripcion_numerica = df_numerico.describe()

# Imprimir la descripción
print(descripcion_numerica)
```

1: Ejemplo de código EDA

```python
# Iterar sobre cada columna numérica y generar un histograma
for columna in df_numerico.columns:
    plt.figure(figsize=(8, 6))
    sns.histplot(df_numerico[columna], bins=30, kde=True)
    plt.title(f'Histograma de {columna}')
    plt.xlabel(columna)
    plt.ylabel('Frecuencia')
    plt.show()
    # Iterar sobre cada columna numérica y generar un boxplot
for columna in df_numerico.columns:
    plt.figure(figsize=(8, 6))
    sns.boxplot(x=df_numerico[columna])
    plt.title(f'Boxplot de {columna}')
    plt.xlabel(columna)
    plt.show()
    # Filtra las columnas numéricas
numeric_data = df.select_dtypes(include='number')

# Calcula la matriz de correlación
correlation_matrix = numeric_data.corr()

# Obtén una lista de todas las columnas numéricas
columnas_numericas = numeric_data.columns

# Define la cantidad de columnas más correlacionadas que deseas obtener
top_n = 3

# Itera a través de cada columna y encuentra las 3 columnas más correlacionadas
for columna in columnas_numericas:
    top_correlations = correlation_matrix[columna].nlargest(top_n + 1)[1:]
    print("Las", top_n, "columnas más correlacionadas con", columna, "son:")
    print(top_correlations)
    print("\n")
```

2: Ejemplo de código EDA

```python
# Top 10 de la columna 'comuna'
top_comuna = df['comuna'].value_counts().head(10)

# Visualizar el top 10 de la columna 'comuna'
plt.figure(figsize=(10, 6))
sns.barplot(x=top_comuna.index, y=top_comuna.values)
plt.title('Top 10 Comunas')
plt.xlabel('Comuna')
plt.ylabel('Frecuencia')
plt.xticks(rotation=45)
plt.show()

# Top 10 de la columna 'provincia'
top_provincia = df['provincia'].value_counts().head(10)

# Visualizar el top 10 de la columna 'provincia'
plt.figure(figsize=(10, 6))
sns.barplot(x=top_provincia.index, y=top_provincia.values)
plt.title('Top 10 Provincias')
plt.xlabel('Provincia')
plt.ylabel('Frecuencia')
plt.xticks(rotation=45)
plt.show()
```

3: Ejemplo de código EDA

```python
# Diccionario de mapeo para género
mapeo_genero = {
    'M': 1,
    'F': 2
}
# Aplicar la transformación a la columna "Genero"
df['Genero_numerico'] = df['gender'].map(mapeo_genero)
#Eliminar algunos datos "inconclusos"
df = df.drop(df[df['gse_corp'] == 'S/I'].index)
df = df.drop(df[df['provincia'] == 'S/I'].index)
df = df.drop(df[df['comuna'] == 'S/I'].index)
df = df.drop(df[df['gender'] == 'S/I'].index)
df = df.drop(df[df['gender'] == 'N'].index)
# Diccionario de mapeo
mapeo_regiones = {
    'Metropolitana de Santiago': 13,
    'Arica y Parinacota': 15,
    'Tarapacá': 1,
    'Antofagasta': 2,
    'Atacama': 3,
    'Coquimbo': 4,
    'Valparaíso': 5,
    'Lib. Gral. Bernardo OHiggins': 6,
    'Maule': 7,
    'Ñuble': 16,
    'Biobío': 8,
    'La Araucanía': 9,
    'Los Ríos': 14,
    'Los Lagos': 10,
    'Aysén del General Carlos Ibáñez del Campo': 11,
    'Magallanes y Antártica Chilena ': 12
}
# Aplicar la transformación a la columna "Region"
df['Region_numerica'] = df['region'].map(mapeo_regiones)
```

4: Ejemplo de código transformación de variables

```python
# Preprocesamiento de datos
scaler = StandardScaler()
scaled_features = scaler.fit_transform(df_numerico)

#  Determinación del número óptimo de clústeres usando el método del codo
visualizer = KElbowVisualizer(KMeans(), k=(2, 10), metric='distortion', timings=False)
visualizer.fit(scaled_features)
visualizer.show()

#  Aplicación del algoritmo KMeans con el número óptimo de clústeres
num_clusters = visualizer.elbow_value_
kmeans = KMeans(n_clusters=num_clusters, random_state=42)
df['cluster'] = kmeans.fit_predict(scaled_features)

# Análisis de resultados
cluster_centers = scaler.inverse_transform(kmeans.cluster_centers_)  # Centroides en la escala original

#Interpretación de clústeres
df_cluster_analysis = pd.DataFrame(cluster_centers, columns=df_numerico.columns)
print("Características de cada clúster:")
print(df_cluster_analysis)


# 9. Validación y ajuste (opcional)
silhouette_avg = silhouette_score(scaled_features, df['cluster'])
print(f"Silhouette Score: {silhouette_avg}")
```

5: Ejemplo código modelo de agrupamiento

```python
# Obtener los índices del top 10
top_indices = np.argsort(mutual_info)[::-1][:10]

# Crear un gráfico de barras para visualizar la ganancia de información del top 10
fig, ax = plt.subplots(figsize=(10, 6))
plt.bar(range(10), mutual_info[top_indices], align="center")
plt.xticks(range(10), X.columns[top_indices], rotation=90)
plt.xlabel("Características")
plt.ylabel("Ganancia de Información")
plt.title("Top 10 de Ganancia de Información para Características")
plt.show()
```

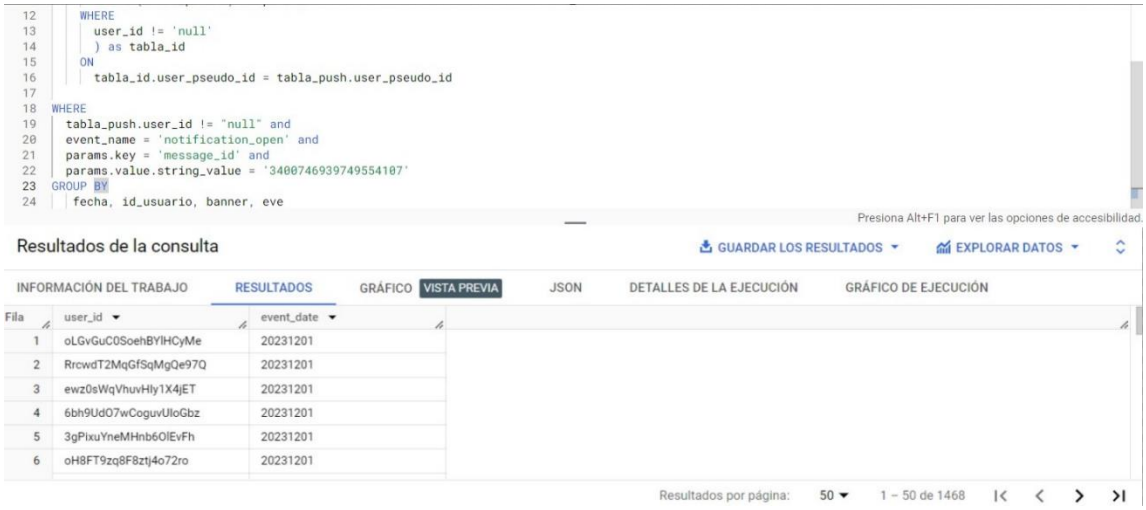
6: Ejemplo código selección de variables

```
#Visualizar captados y no captados
count_classes = pd.value_counts(df_numerico['captaron_tarjeta'], sort = True)
count_classes.plot(kind = 'bar', rot=0)
plt.xticks(range(2), LABELS)
plt.title("Frecuencia")
plt.xlabel("Clase")
plt.ylabel("Numero de observaciones");
#dividimos en sets de entrenamiento y test
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7)

#creamos una funcion que crea el modelo que usaremos cada vez
def run_model(X_train, X_test, y_train, y_test):
    clf_base = LogisticRegression(C=1.0,penalty='l2',random_state=1,solver="newton-cg")
    clf_base.fit(X_train, y_train)
    return clf_base

#ejecutamos el modelo "tal cual"
model = run_model(X_train, X_test, y_train, y_test)

#definimos funciona para mostrar los resultados
def mostrar_resultados(y_test, pred_y):
    print (classification_report(y_test, pred_y))
pred_y = model.predict(X_test)
mostrar_resultados(y_test, pred_y)
```

7: Ejemplo código modelo de clasificación

```
#Aplicamos el algoritmo BalancedBaggingClassifier
bbc = BalancedBaggingClassifier(base_estimator=LogisticRegression(),
                                sampling_strategy='auto',
                                replacement=False,
                                random_state=0)

#Train the classifier.
bbc.fit(X_train, y_train)
pred_y = bbc.predict(X_test)
mostrar_resultados(y_test, pred_y)
```

8: Ejemplo código modelo de balanceo



```
    WHERE
      user_id != 'null'
      ) as tabla_id
    ON
      tabla_id.user_pseudo_id = tabla_push.user_pseudo_id

WHERE
  tabla_push.user_id != "null" and
  event_name = 'notification_open' and
  params.key = 'message_id' and
  params.value.string_value = '3400746939749554107'
GROUP BY
  fecha, id_usuario, banner, eve
```

Presiona Alt+F1 para ver las opciones de accesibilidad.

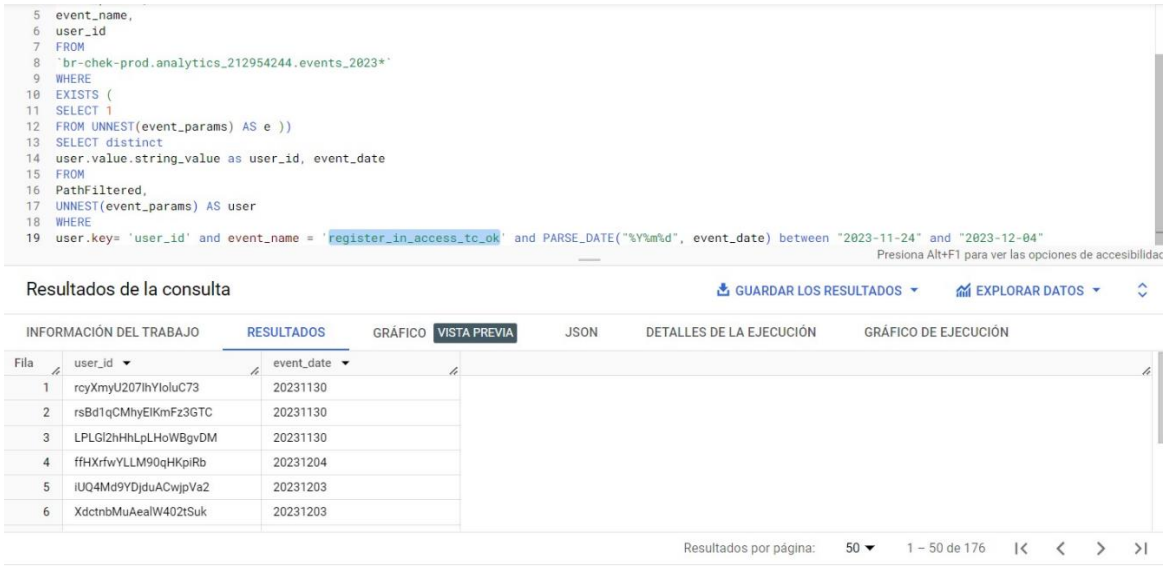
Resultados de la consulta — GUARDAR LOS RESULTADOS — EXPLORAR DATOS

INFORMACIÓN DEL TRABAJO | RESULTADOS | GRÁFICO | VISTA PREVIA | JSON | DETALLES DE LA EJECUCIÓN | GRÁFICO DE EJECUCIÓN

Fila	user_id	event_date
1	oLGvGuC0SoehBYlHCyMe	20231201
2	RrcwdT2MqGfSqMgQe97Q	20231201
3	ewz0sWqVhuvHly1X4jET	20231201
4	6bh9UdO7wCoguvUloGbz	20231201
5	3gPixuYneMHnb6OlEvFh	20231201
6	oH8FT9zq8F8ztj4o72ro	20231201

Resultados por página: 50 — 1 – 50 de 1468

9: Ejemplo código en Big Query para seguimiento de notificación push



```
event_name,
user_id
FROM
`br-chek-prod.analytics_212954244.events_2023*`
WHERE
EXISTS (
SELECT 1
FROM UNNEST(event_params) AS e ))
SELECT distinct
user.value.string_value as user_id, event_date
FROM
PathFiltered,
UNNEST(event_params) AS user
WHERE
user.key= 'user_id' and event_name = 'register_in_access_tc_ok' and PARSE_DATE("%Y%m%d", event_date) between "2023-11-24" and "2023-12-04"
```

Presiona Alt+F1 para ver las opciones de accesibilidad

Resultados de la consulta — GUARDAR LOS RESULTADOS — EXPLORAR DATOS

INFORMACIÓN DEL TRABAJO | RESULTADOS | GRÁFICO | VISTA PREVIA | JSON | DETALLES DE LA EJECUCIÓN | GRÁFICO DE EJECUCIÓN

Fila	user_id	event_date
1	rcyXmyU207lhYloluC73	20231130
2	rsBd1qCMhyElKmFz3GTC	20231130
3	LPLGl2hHhLpLHoWBgvDM	20231130
4	ffHXrfwYLLM90qHKpiRb	20231204
5	iUQ4Md9YDjduACwjpVa2	20231203
6	XdctnbMuAealW402tSuk	20231203

Resultados por página: 50 — 1 – 50 de 176

10: Ejemplo código en Big Query para seguimiento de captaciones

Campaña	Inicio	Fin	Estado	Segmentación	Última actualización	Envíos o impresiones	Clics o aperturas
$15.000 de regalo Obtén tu Tarjeta de Crédito Chek AQUÍ y ganas 15 luquitas ¡Pídela aquí!	20 nov 2023 16:00 p.m	-	Completadas	Piloto_predic	20 nov 2023	10,000-20,000	12%

Campaign ID: 3400746939749554107

Borrar | Duplicar

11: Ejemplo notificación push de piloto (donde aparece ID de campaña y tasa de apertura)

Referencias

1. Banco Chek, 2023. Reuniones internas área comercial

2. Banco Ripley, 2023. Reuniones

3. Clases de Fundamentos de Ciencias de Datos. Anriquez, Gonzalo. Universidad Adolfo Ibáñez.

4. Clases de Minería de Datos. Anriquez, Gonzalo. Universidad Adolfo Ibáñez.

5. BESMART Company. https://besmart.company

6. Metodología Fundamental para Ciencias de Datos. IBM Analytics.