UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
DEPARTAMENTO DE INGENIERIA INDUSTRIAL

“MODELO DE INACTIVIDAD DE CLIENTES TAXISTAS PARA UNA DISTRIBUIDORA DE COMBUSTIBLES NACIONAL”

MEMORIA PARA OPTAR AL TITULO DE INGENIERO CIVIL INDUSTRIAL

BENJAMÍN IGNACIO NORAMBUENA YUNG

PROFESOR GUÍA:
PABLO MARÍN VICUÑA

MIEMBROS DE LA COMISIÓN:
CAROLINA SEGOVIA RIQUELME
JOSÉ ANTONIO NALDA REYES

SANTIAGO DE CHILE
2020

RESUMEN DE LA MEMORIA PARA OPTAR AL TITULO DE: Ingeniero Civil Industrial
POR: Benjamín Ignacio Norambuena Yung
FECHA: 30/09/2020
PROFESOR GUÍA: Pablo Marín Vicuña

MODELO DE INACTIVIDAD DE CLIENTES TAXISTAS PARA UNA DISTRIBUIDORA DE COMBUSTIBLES NACIONAL

En los últimos años, cada vez han ido tomando mayor importancia los programas de fidelización en empresas, los que permiten manejar información personal de clientes, para luego planificar acciones comerciales acordes a distintas situaciones que se puedan presentar.

El trabajo desarrollado se enmarca en este contexto, donde se presenta una distribuidora de combustibles que dispone de un programa de fidelización para taxistas, el cual ha disminuido su actividad en el último tiempo, bajando el consumo de gasolina, así como también la cantidad de clientes activos.

Lo anterior representa un problema para la compañía, que se aborda definiendo como objetivo general, la identificación de taxistas propensos a la inactividad del programa de fidelización, mediante un modelo de predicción que permita optimizar campañas de retención de clientes vía marketing directo.

Para el modelamiento, se emplea la metodología KDD sobre una base de datos transaccional de la compañía, así como también una técnica conocida como RFMC, que permite clasificar quincenalmente a cada taxista en uno de 5 segmentos posibles. Basado en esos segmentos, se define la variable a predecir ("Inactividad") en cada macro grupo de clientes (3 posibles: Tarjetas, App y Tarjetas y App) y se ejecutan distintos modelos (cadenas de markov, árboles de decisión, *random forest* y *logit*), comparando luego sus resultados, y eligiendo los mejores en base a la maximización de la métrica ROC AUC.

Las conclusiones y principales hallazgos se centran en que sí es posible hacer el modelamiento de inactividad de taxistas, siendo *random forest* el modelo de mejor desempeño, alcanzando 85%, 83% y 87% en ROC AUC para los grupos Tarjetas, App y Tarjetas y App, respectivamente; y su utilización en campañas supondría un ahorro considerable en costos, los que se estiman en aproximadamente 37%, 53% y 76% para los 3 grupos anteriores. Cabe destacar que los modelos tienen la particularidad de poder predecir si un cliente se va a inactivar o no sólo analizando variables transaccionales de la última quincena de actividad, lo que entrega una alta capacidad de respuesta en el corto plazo a la compañía, sin la necesidad de tener que esperar observar variables acumulativas. Lo anterior además lleva a inferir que los taxistas analizados presentan un comportamiento de compra más bien definido en el tiempo inmediatamente anterior a su inactividad, donde generalmente, compran montos bajos y realizan pocas transacciones.

Dicho esto, distintas recomendaciones comerciales son propuestas, siendo la principal, la utilización de los modelos en futuras campañas de retención del programa. Así también, se entregan posibles lineamientos de trabajos posteriores, que sirvan como complemento al trabajo ya realizado.

Tabla de contenido

1. Introducción

1.1 Características de la organización

La organización donde se lleva a cabo el proyecto corresponde a una empresa chilena con fines de lucro, que tiene su principal actividad en la comercialización y distribución de combustibles líquidos y lubricantes. Esta compañía forma parte de un holding de empresas perteneciente a un grupo familiar chileno, y posee giro en distintos mercados.

La empresa actualmente presenta 653 estaciones de servicio de Arica a Puerto Williams, de las cuales 78 cuentan con la opción de autoservicio, 91 con restaurantes de comida rápida, y 65 de ellas tienen estaciones “Lub”, que ofrecen lubricantes, grasas, refrigerantes y espacialidades para automóviles. Aquí se da trabajo a un total de 9.500 atendedores en estaciones de servicio, 3.200 en tiendas de conveniencia, y 230 en estaciones “Lub”. Además, según declara la empresa, alrededor de 2.000 empleados son extranjeros, proviniendo la mayoría de Venezuela, Colombia y Haití, representando un 21% del total de atendedores de la red.

Durante el año 2018, la compañía atendió a más de 750.000 clientes en promedio diariamente en estaciones y tiendas de conveniencia, vendiendo aproximadamente 9.356.000 metros cúbicos de combustibles entre todos sus canales (54% de participación de mercado), además, en términos de venta de lubricantes, se logró comercializar alrededor de 91.000 metros cúbicos, adjudicándose más del 43% de participación de mercado (Compañía, 2018).

Con presencia en Chile, Colombia, Estados Unidos y otros países de América, tanto a nivel consolidado como individual, el buen desempeño de la compañía y sus filiales extranjeras ha ido generando crecimiento en los indicadores financieros durante los últimos años, tales como EBITDA, ventas e ingresos operacionales, entre otros. Es así como para el último año evaluado (2018), considerando la totalidad de sus ventas, la compañía logró adjudicarse una suma de $10.725.338 millones de pesos chilenos por ingresos operacionales, concentrándose la mayor cantidad de ellos en Chile ($5.405.638 millones), Colombia ($3.214.092 millones) y Estados Unidos ($930.895 millones). Así, a partir de todas estas ventas, la compañía obtuvo una cifra igual a $429.448 millones de pesos chilenos en su EBITDA[1], para finalmente quedarse con $170.239 millones de utilidad neta (Compañía, 2018).

[1] Utilidades antes de intereses, impuestos, depreciación y amortizaciones

1.2 Mercado institucional

El sector petrolífero (al cual pertenece la compañía) abarca un amplio mercado, dado que no sólo contempla la venta de combustibles a personas naturales, sino que también a aerolíneas, fuerzas armadas, empresas transportistas, pesqueras, entre otras; donde la mayor parte de este mercado es asistido a través de las estaciones de servicio distribuidas a lo largo y ancho del país.

La industria de venta de combustibles es extensa, tanto así que para el año 2018 se lograron transar 17.326.340 metros cúbicos de combustibles líquidos (que involucran petróleo, kerosene y gasolinas) (Comisión Nacional de Energía, 2020), logrando adjudicarse la compañía más del 53% de participación de estos combustibles (líderes en el mercado, la competencia que le sigue tiene un aproximado del 22%), y alrededor de un 44% en lubricantes (Compañía, 2018).

Por otra parte, es pertinente describir el marco regulatorio que concierne a la industria de combustibles de la compañía en Chile, declarado en el prospecto legal, citado a continuación:

"En virtud de la dictación del DFL1 del año 1979, la refinación, importación, distribución y comercialización de combustibles líquidos derivados del petróleo pueden ser efectuadas libremente por los particulares, haciendo presente que a partir del año 1978 se deroga el Decreto Número 20 del Ministerio de Minería del 1964, estableciéndose libertad de precios para los productos derivados del petróleo.

Debe tenerse presente que la Superintendencia de Electricidad y Combustibles, organismo que está regulado por la Ley 18.410, es el servicio dependiente del Ministerio de Economía encargado de fiscalizar y supervigilar el cumplimiento de las normas legales, reglamentarias y técnicas aplicables a la distribución de combustibles líquidos.

El Decreto Supremo N° 160 del Ministerio de Economía, publicado en el Diario Oficial del día 7 de Julio del año 2009, contiene el Reglamento de Seguridad para el Almacenamiento, Refinación, Transporte y Expendio al público de Combustibles Líquidos derivados del Petróleo, norma que es aplicable a todas las instalaciones destinadas a tal fin en cuanto a su operación, inspección, mantenimiento y término." (Compañía, 2016).

Finalmente, cabe destacar que dentro la regulación legal, son varios los organismos presentes en la industria de combustibles en Chile, lo que las hace formar parte del marco de la institución descrita. Dentro de estos organismos, se encuentran tres principales:

- **Empresa Nacional del Petróleo (ENAP):** Refina el petróleo para luego comercializarlo a distribuidoras.
- **Comisión Nacional de Energía (CNE):** Organismo técnico encargado de analizar precios, tarifas y normas técnicas a las que deben ceñirse las empresas de producción, generación, transporte y distribución de energía. Depende del ministerio de energía. (Comisión Nacional de Energía, 2020).
- **Superintendencia de Electricidad y Combustibles (SEC):** Fija los estándares técnicos de calidad de combustibles y fiscaliza su cumplimiento.

1.3 Desempeño institucional

A partir de los aproximadamente 600 puntos de venta a lo largo de Chile, la organización desarrolla su actividad comercial a través de distintas áreas de negocio (aparte de la venta de combustibles), pudiendo encontrar una amplia gama de servicios adicionales, entre los que destacan: baños, lavado de autos, venta de lubricantes, cafeterías, restaurantes, entre otros.

El origen de la empresa se remonta al año 1934, cuando un grupo de empresarios chilenos forman la compañía con el objetivo de distribuir y comercializar combustibles en el país. Con el correr de los años, la red de estaciones de servicio comenzó a marcar presencia en distintos lugares, logrando cubrir desde Coquimbo a Magallanes hacia finales de 1940. Con la marca ya conocida, para 1964, la compañía pasa a formar parte de la fundación de la Sociedad de Inversiones de Aviación Ltda. (SIAV), especializada en el suministro de combustibles a aviones en aeropuertos, y un año más tarde, la compañía crea una nueva unidad de negocio: minimarkets en estaciones de servicio. Ya en los años 90, la empresa estaba más que consolidada y estrena su primer restaurant en estación de servicio, contando con más de 60 de ellos en la actualidad. Luego, en el año 2010 la compañía inicia su expansión internacional con la adquisición de otra distribuidora de combustibles en Colombia. Finalmente, desde ese año en más, la compañía incrementa sus estrategias de innovación creando así estaciones de servicios sin atendedores, aplicaciones móviles para pagar desde smartphones, entre otras, ganando cada vez más clientes tanto nacional como internacionalmente (Compañía, 2020).

Todo lo anterior conlleva a que, dada la gran cartera de clientes (y sus variados comportamientos), se creen programas de fidelización para ellos, realizándose así, mes a mes, distintas campañas de marketing, promocionando tanto productos como beneficios que puedan haber por el uso de cada fidelidad.

2. Descripción del proyecto y justificación

2.1 Información del área de la organización

El trabajo se realiza bajo la gerencia de Marketing, específicamente, el área de Data Analytics, que tiene como objetivo encontrar *insights* en los clientes de la compañía, a partir del manejo y análisis de bases de datos pertenecientes a la empresa, lo que da paso al descubrimiento de conclusiones ocultas sobre el comportamiento humano, que permiten apoyar la toma de decisiones en el negocio, como también la creación de nuevas estrategias de marketing para el futuro.

La gerencia de Marketing está conformada por una totalidad de 46 profesionales, separados en 4 áreas, donde gran parte de ellos son ingenieros(as) civiles industriales, egresados de distintas universidades. Por otra parte, también hay ingenieros civiles en computación e ingenieros comerciales. A continuación, en la figura 1, se muestra cómo se organizan las subáreas de manera gráfica.

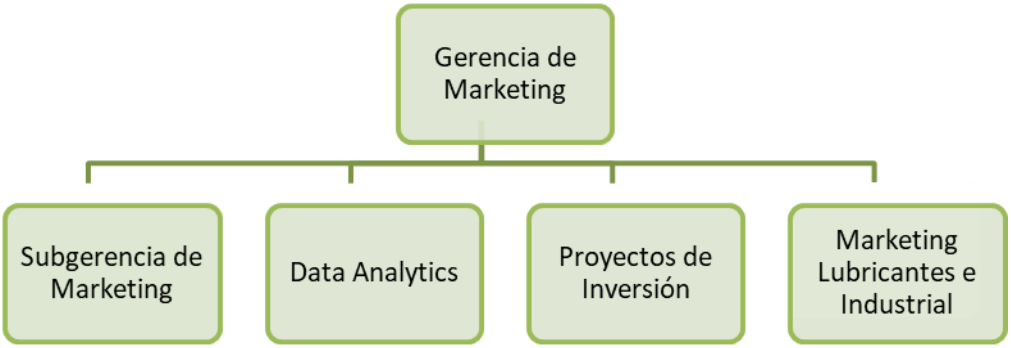


Figura 1: Organigrama gerencia de Marketing

Fuente: Elaboración propia

2.2 Identificación del problema y su relevancia

Actualmente en la institución, se realizan mes a mes una cantidad considerable de campañas publicitarias hacia sus clientes, promocionando distintas actividades dependiendo del tipo de negocio a las que pertenecen, utilizando tanto medios masivos (TV, prensa, redes sociales, etc.) como medios “uno a uno” (email, SMS, web, etc.) para su comunicación. Variados son los tipos de negocios existentes, sin embargo, para efectos del trabajo, sólo uno es el que interesa: el segmento de taxistas y colectivos, que representan el 1,5% del total de volumen de combustible cargado por personas naturales en la compañía.

La empresa, con el objetivo de captar una mayor cantidad de clientes, el año 2011, crea un programa de fidelización para taxis/colectivos, que entrega un beneficio de descuento

fijo por litro de combustible cargado (a cada taxista inscrito), dependiendo del volumen total cargado el mes anterior, teniendo así tres posibles descuentos:

- $10 pesos/lt: Si se cargó entre 0 y 199 lts.
- $12 pesos/lt: Si se cargó entre 200 y 400 lts.
- $15 pesos/lt: Si se cargó más de 400 lts.

Luego, a modo de ejemplo, si un taxista carga 350 litros de combustible el mes de enero, para el mes de febrero se le hace un descuento de $12 pesos por litro. Si un cliente es nuevo en el programa, se le asigna automáticamente el descuento máximo para el primer mes de uso del beneficio.

El programa inicia su funcionamiento utilizando tarjetas, es decir, a cada miembro del programa se le entregaba una tarjeta física al momento de fidelizarse, que contenía la patente asociada al vehículo del taxista, y que el cliente mostraba al momento de cargar combustible para acceder a los descuentos. Si la tarjeta se desgastaba o perdía, el taxista podía pedir su reposición y la compañía le fabricaba una nueva sin costo alguno. Sin embargo, con el correr de los años, lo anterior cambió. Esto porque en agosto 2018, se creó una App asociada al programa, la cual terminó con la creación de tarjetas, obligando a los taxistas a seguir usando sus tarjetas antiguas, o bien, cambiarse a la aplicación móvil para seguir accediendo a los descuentos.

Haciendo un seguimiento de clientes activos en el programa desde comienzos de 2019, es posible realizar distintos análisis, que se ven reflejados en variados gráficos[2]; primero el gráfico 1, donde se evidencia que a medida que transcurren los meses, hay una tendencia (línea roja punteada) a la disminución en la cantidad de clientes activos utilizadores del programa. Luego, analizando detalladamente los valores, se ve que, desde enero a diciembre 2019, hay una baja de aproximadamente 8% de taxistas activos en el programa, alertando de un posible problema clave para la compañía: la inactividad de clientes del programa de fidelización.

[2] Por temas de confidencialidad de la empresa, no es posible mostrar los números de la compañía, de ahí que se muestran porcentajes, donde, para el gráfico 1 se fijan los clientes activos de enero 2019 como el 100% de clientes del programa. Además, dada la baja actividad económica vivida entre octubre y noviembre producto de crisis social chilena, se optó por omitir esos meses en los análisis.

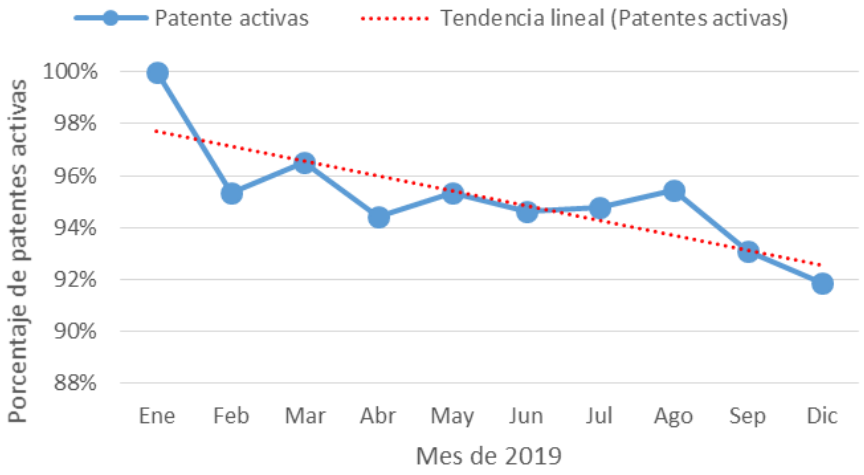


Gráfico 1: Evolución de porcentaje de clientes activos para el año 2019

Fuente: Elaboración propia

Lo anterior da paso a analizar si la tendencia a la disminución de clientes del programa de fidelización tiene algún efecto en la compañía, principalmente en el volumen de combustible total comprado por taxistas. Es así como se obtienen los resultados mostrados en el gráfico 2, donde, a partir de un porcentaje de crecimiento, se compara la cantidad de combustible cargada en 2019 v/s la cantidad cargada en 2018 para un mismo mes, es decir, se indica cuánto porciento más de volumen se cargó el 2019 con respecto a 2018 para el mes estudiado.

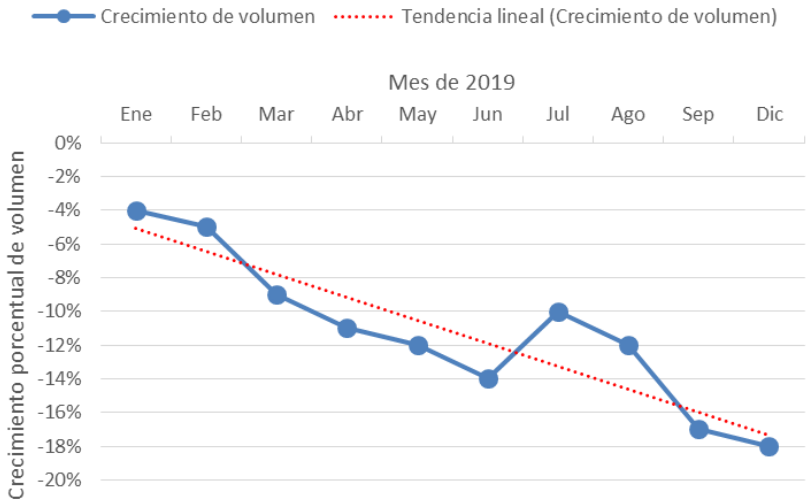


Gráfico 2: Crecimiento porcentual de volumen mensual (año 2019 con respecto a 2018)

Fuente: Elaboración propia

Observando el gráfico, se ve que para todos los meses de 2019 se tienen crecimientos negativos, lo que, dicho en otras palabras, significa que en cada mes de 2019 se cargó menos combustible que el mismo mes en 2018. Así, se destaca nuevamente la tendencia negativa mostrada, ahora para el crecimiento de volumen, que supone una continuación en la baja de combustible cargado en los meses del siguiente año, empeorando la situación del programa de fidelización si no se hacen cambios.

Indagando en la baja continua de crecimiento porcentual de volumen, surge como una posible explicación la irrupción de plataformas tecnológicas en el mercado (tales como *Uber*, *Didi* o *Beat*), sin embargo, la disminución es de un exceso tal que no alcanza a ser explicada sólo por ellas. Esto último queda demostrado en estudios de mercado realizados por una consultora externa a la compañía donde, para el año 2019, se estima que aproximadamente el 6,5% de la baja de litros cargados se debe a la llegada de Apps móviles, sin saber de dónde podría provenir una baja mayor. Así, por ejemplo, para el crecimiento negativo de mayo del gráfico 2, del 12% decrecido, hay un 5,5% que no se sabe a qué causa atribuir.

Todo lo anterior va de la mano con la inactividad de clientes diagnosticada, puesto que, a menor cantidad de clientes presentes, es de esperar que la compañía venda menos cada mes.

Por otra parte, según un estudio de mercado realizado por la compañía todos los años, desde su creación en 2011, el programa fue aumentando poco a poco su habitualidad[3] en el rubro de taxistas/colectivos - aunque con una leve caída entre 2013 y 2015, repuntando en 2016 - estando siempre por sobre la competencia. Sin embargo, como se puede ver en el gráfico 3, a partir de 2018, la compañía experimenta una disminución en la habitualidad de los clientes, mientras que la competencia directa inicia un aumento en su participación. Además, es posible percatarse de que lo anterior se agrava para el año siguiente (2019), cayendo aún más el indicador, aumentando a su vez el de la competencia, reforzando la idea de la posible inactividad de clientes, y profundizando el problema detectado: posiblemente los clientes inactivos son captados por la competencia.

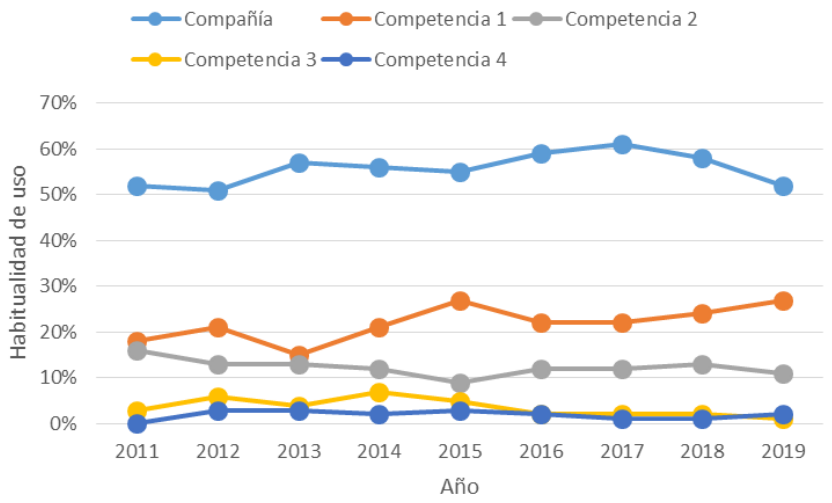


Gráfico 3: Evolución habitualidad de marcas en taxistas

Fuente: Estudio de mercado de la compañía

[3] Habitualidad indica qué tan habitual es para un taxista cargar en una compañía determinada

Finalmente, desde el área de marketing de la compañía señalan que para el 2019 no se hicieron campañas publicitarias relacionadas al programa de fidelización de taxis/colectivos, principalmente debido a descoordinaciones entre áreas involucradas en las promociones a realizar, así como también la llegada de la crisis social chilena en octubre, ocasionando un distanciamiento con el público objetivo (en términos de marketing). Además, expresan que, en toda campaña publicitaria realizada dentro de los últimos 2 años, sólo ha habido segmentaciones de clientes en base a la forma de uso del beneficio, es decir, si el cliente hizo uso del beneficio a través de la App, o bien, a través de la tarjeta física. Luego, la compañía no cuenta con la existencia de una segmentación que logre identificar la calidad de cada cliente que tiene fidelizado, imposibilitando la formulación de estrategias de retención hacia ellos, que podrían ser, por ejemplo, el ofrecimiento de regalías y/o beneficios extras en agradecimiento a su fidelidad y comportamiento con la empresa (para aquellos clientes sobresalientes).

Así, producto del problema de inactividad de clientes detectado y sus posibles causas, se pueden identificar dos efectos claves para la compañía, detallados en lo que sigue:

1) **Baja en el volumen general de ventas:** Un cliente taxista, al inactivarse, abandona consigo todo el volumen de combustible que cargaba utilizando su fidelidad, lo que, en promedio para 2019, se estima en 219 litros mensuales[4]. Luego, por cada cliente taxista inactivo que llenaba su estanque con bencina de 93 octanos (según la empresa, el común de los clientes), la compañía pierde en promedio 2.128.680 pesos[5] anuales por concepto de ventas.

2) **Aumento de costos asociados al programa de fidelización de taxis/colectivos:** Al inactivarse un taxista fidelizado, probablemente también se pierda la forma de contactarlo, ya que en el tiempo en que la compañía se percata de que el cliente dejó de cargar, él posiblemente cambie de número telefónico, vehículo, y/u otro; lo que aumenta la dificultad de seguir su rastro, dificultando así también su reactivación por medio de la gestión de marketing. Esto lleva a que la única opción para mantener o aumentar el número de clientes del programa sea adquiriendo nuevos, lo cual es entre 5 y 25 veces más caro que retener antiguos (dependiendo la industria en cuestión) (Harvard Business Review, 2014).

[4] Fuente: Bases de datos de la compañía

[5] Se fija el valor de un litro de combustible en $810 pesos chilenos (correspondiente a diciembre 2019)

2.3 Hipótesis y alternativas de solución para resolver el problema

Como hipótesis de las causas que dan origen al problema presentado, se tienen dos principales:

1) La no realización de campañas de marketing directo para clientes pertenecientes al programa de fidelización de taxis/colectivos genera un distanciamiento entre la empresa y aquellos taxistas utilizadores del programa, lo que es aprovechado por la competencia, logrando "quitar" mes a mes a todos aquellos clientes descontentos.

2) En la actualidad, la segmentación de clientes en las empresas suele estar a cargo de un área de Data Analytics u otra que apoye la toma de decisiones a partir del análisis e investigación de datos. En la compañía, esa área comenzó su funcionamiento en agosto de 2019, por lo que anterior a eso, las clasificaciones de clientes no eran óptimas de acuerdo con el objetivo que tenía cada campaña publicitaria a realizar (no existían segmentaciones basadas en datos). Luego, el marketing directo no necesariamente alcanzaba plena eficacia en el mensaje transmitido, dejando clientes dentro de campañas que en realidad no debían ser considerados, y otros fuera (que sí tenían que recibir el mensaje), los que finalmente optaron por distanciarse del programa.

Debido a que la decisión de realizar campañas publicitarias se resuelve a nivel gerencial, y de que es otra el área que gestiona su lanzamiento (no es posible interceder en ella), se descarta el trabajo en la primera hipótesis, luego, se decide abordar la segunda, principalmente porque el área de Data Analytics ya se encuentra activa en la compañía, y es allí donde quincenalmente se realiza una segmentación[6] de clientes del programa de fidelización, todo basado en los datos.

Los segmentos ya obtenidos, son tomados como información para la construcción de modelos de predicción de inactividad de clientes, que, para cada taxista, quincena a quincena, responden las preguntas:

- ¿Cuál es la probabilidad de que el cliente se inactive en el siguiente período?
- ¿Está el cliente en peligro de inactividad?
- ¿Se debe ejercer marketing directo sobre él? y si es así, ¿qué campaña se le debe aplicar?

Teniendo respuesta a lo anterior, los resultados son entregados al área de marketing, con el fin de que formulen estrategias para intentar evitar que clientes propensos a la inactivad, caigan en ella, todo a partir de la realización de marketing directo.

[6] La segmentación corresponde a una utilizando el método RFMC, detallado en la sección 4.1.1.

3. Objetivos

3.1 Objetivo general:

Identificación de taxistas propensos a la inactividad del programa de fidelización de una compañía[7], mediante un modelo de predicción que permita optimizar campañas de retención de clientes vía marketing directo.

3.2 Objetivos específicos:

Para lograr el cumplimiento del objetivo general se declaran los siguientes objetivos específicos:

1) Definir y construir base de datos analítica necesaria para la construcción de modelos
2) Caracterizar clientes taxistas según su comportamiento histórico
3) Definir criterio de inactividad
4) Construir y evaluar modelos que entreguen la probabilidad de que un cliente suspenda el uso del programa en el siguiente período (quincena), dado su comportamiento en el pasado
5) Definir rangos de probabilidades entre los cuales un cliente se debe catalogar como “en peligro de inactividad”
6) Identificar clientes que están en estado “en peligro de inactividad” para la quincena siguiente a la de estudio

[7] Por temas de confidencialidad no se puede decir en nombre

4. Marco conceptual

4.1 Marco metodológico

Para la clasificación de objetos en clases o categorías en particular, se utilizan distintos modelos que, dependiendo de su naturaleza, pueden entregar una decisión estricta sobre si un elemento pertenece a una de las clases definidas (a partir de una variable binaria), o bien, una probabilidad de que el elemento pertenezca a cada clase, luego, es posible identificar 4 modelos útiles para el trabajo realizado. Además, dado que los modelos se basan en una segmentación RFMC, y su comparación es bajo distintas métricas de desempeño, estas secciones también son detalladas.

4.1.1 Método RFMC (RFM + C)

RFM es una técnica utilizada en marketing para caracterizar y segmentar el comportamiento de clientes a partir del estudio de su historial de transacciones realizadas (Search Data Management, 2005). Luego, en base a esta caracterización, se pueden generar diversas aplicaciones, una de las cuales puede ser la identificación de qué consumidores son los más valiosos en una compañía (Investopedia, 2019). Para ello, sobre cada cliente, se inspeccionan tres variables claves:

- **Recencia** (R): Indica qué tan reciente fue la última compra del cliente analizado. Se mide en unidades de tiempo, pudiendo ser días, semanas, meses, entre otros. Esta variable responde a la pregunta: ¿Hace cuánto tiempo fue la última compra del cliente? Luego, a mayor cantidad de unidades de tiempo, peor cliente.

- **Frecuencia** (F): Se define como la cantidad de veces que un cliente compró dentro de un período de tiempo determinado (día, semana, quincena, mes, etc.). Responde a la pregunta: ¿Cuántas veces compró el cliente? Luego, a mayor cantidad de veces compradas, mejor cliente.

- **Monto** (M): Es la cantidad de dinero gastada por el cliente entre todas sus compras, dentro de un período de tiempo determinado. Esta variable responde a la pregunta: ¿Cuánto fue el gasto total del cliente dentro de la ventana de tiempo analizada? Así, a mayor monto gastado, mejor cliente.

Cada variable, dependiendo del valor que ocupa en cada cliente, se evalúa asignándole un score (suele ser un valor entero en un rango definido), donde a mayor score, mejor evaluación. Luego, juntando la evaluación de las tres variables (comúnmente sólo se concatenan los scores en el orden R, F, M), se obtiene una evaluación general (score RFM), que entrega una caracterización del comportamiento de cada cliente a partir de su recencia, frecuencia y monto total comprado, en ventanas de tiempo específicas (que, dependiendo de los requerimientos de cada compañía, pueden ser quincenas, meses, semestres, u otras).

El procedimiento anteriormente mencionado, en conjunto con reglas de negocio, permite la segmentación de cada cliente en distintos *clusters* (la cantidad de grupos se define con cada empresa), todo basado en su evaluación general RFM (Optimove, 2020).

Por otra parte, si bien el método RFM inspecciona y evalúa tres variables, hay ciertas modificaciones a la técnica, donde además de recencia, frecuencia y monto, también se incluye el análisis de otra(s) variable(s). Es así como surge el RFMC, que suma la "consistencia" al RFM, definida como:

- **Consistencia** (C): Variable que tiene por finalidad evaluar la distribución de compras de un cliente en días distintos. Responde a la pregunta: De la totalidad de días únicos comprendidos en el período de observación, ¿en cuántos de ellos compró el cliente? Luego, a mayor cantidad de días, mejor score de consistencia.

Finalmente, y análogo al procedimiento RFM, se evalúa cada variable para luego obtener un score RFMC que permite la asignación de cada cliente a uno (y sólo uno) de los *clusters* previamente definidos.

El éxito del RFM en la industria ha quedado de manifiesto en los últimos 10 años, esto a partir de su utilización en muchas de las grandes empresas existentes hoy en día, especialmente empresas retail (Doğan, Ayçin, & Bulut, 2018).

4.1.2 Árboles de decisión

Un árbol de decisión divide la totalidad de elementos (muestra o población) en dos o más particiones homogéneas de datos, basado en el separador o diferenciador más significativo de las variables independientes elegidas (atributos), que pueden ser tanto continuas como categóricas.

Cada árbol de decisión se compone de 5 elementos principales (Veloso, 2019):

1) **Nodo raíz:** Representa la población/muestra total, la cual se dividirá en 2 o más grupos de acuerdo a distintas reglas de clasificación

2) **Nodo de decisión:** Corresponden a sub-nodos que se separan en más sub-nodos

3) **Nodo hoja/terminal:** Son nodos que no se separan nuevamente

4) **Rama / Sub-árbol:** Corresponde a una sub-sección del árbol entero

5) **Nodos padre e hijo:** Un nodo que es dividido en sub-nodos es denominado un nodo padre de los nodos en que se dividió, que corresponden a sus nodos hijo.

Los caminos desde el primer nodo a cada hoja representan las reglas de clasificación de cada árbol de decisión. A modo de ejemplo, se presenta la ilustración 1, que modela una acción a partir de un árbol de decisión considerando variables meteorológicas.

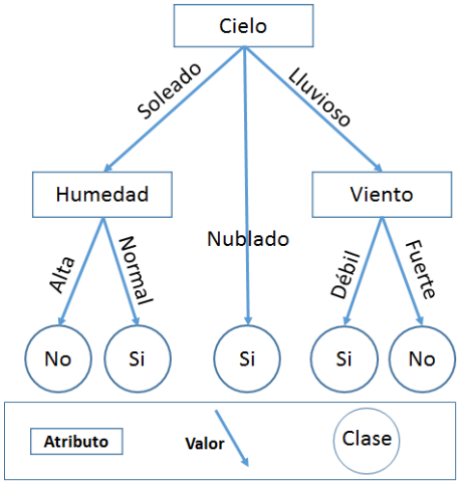


Ilustración 1: Ejemplo árbol de decisión (variable dependiente: "Salir a correr")

Fuente: (Numerentur, 2020)

Existen diferentes criterios de separación en un árbol, y entre los más utilizados se encuentran dos principales:

1) **Information Gain:** Este criterio se basa en la disminución de la entropía (medida de dispersión del conjunto de datos) después de que un conjunto de datos se divide según un atributo en particular. La entropía queda definida como muestra la fórmula 1.

$$Entropía(S) = -p_1 \log_2(p_1) - p_0 \log_2(p_0)$$

Fórmula 1: Cálculo de entropía generada por un atributo determinado

Fuente: (Quantinsti, 2019)

donde p_0 es la es la proporción de elementos del conjunto S denotados en la clase 0 y p_1 la proporción de elementos en la clase 1.

Luego, a partir del valor obtenido, para cada uno de los atributos se calcula la ganancia de información que genera el dividir la data en dos grupos distintos, lo que se traduce en la fórmula 2.

$$Information\ Gain(S, V) = Entropía(S) - \sum_{v \in V} \frac{|S_v|}{|S|} Entropía(S_v)$$

Fórmula 2: Cálculo de ganancia de información

Fuente: (Quantinsti, 2019)

donde S_v es el subconjunto de datos cuando se tiene el valor v de la variable V.

Así, el criterio de *Information Gain* hace elegir variables que generen una mayor ganancia de información al tomar determinados valores.

2) **Gini Index:** Consiste en un índice que mide la "impureza" de una partición de datos (clase), donde "impureza" se relaciona con la probabilidad de que una variable particular se clasifique erróneamente cuando se elige al azar. Este índice toma valores en el rango [0,1], siendo 0 una clasificación perfecta de la clase, luego, utilizando este criterio, primero se buscan las variables que tengan el menor *Gini index*. Su expresión está dada por la fórmula 3:

$$Gini\ index = 1 - \sum_{i=1}^{n} p_i^2$$

Fórmula 3: Cálculo de Gini index

Fuente: (Numerentur, 2020)

donde n es el número de clases y p_i es la probabilidad de que un objeto pertenezca a la clase i.

4.1.3 Random forest

En *machine learning*, *random forest* consiste en la realización de variados árboles de clasificación relativamente no correlacionados (técnica conocida como *bagging,* ver ilustración 2), donde cada árbol individual del "bosque" entrega una clase de predicción, luego, la clase con mayor cantidad de votos es donde finalmente queda el registro clasificado (Yiu, 2019).

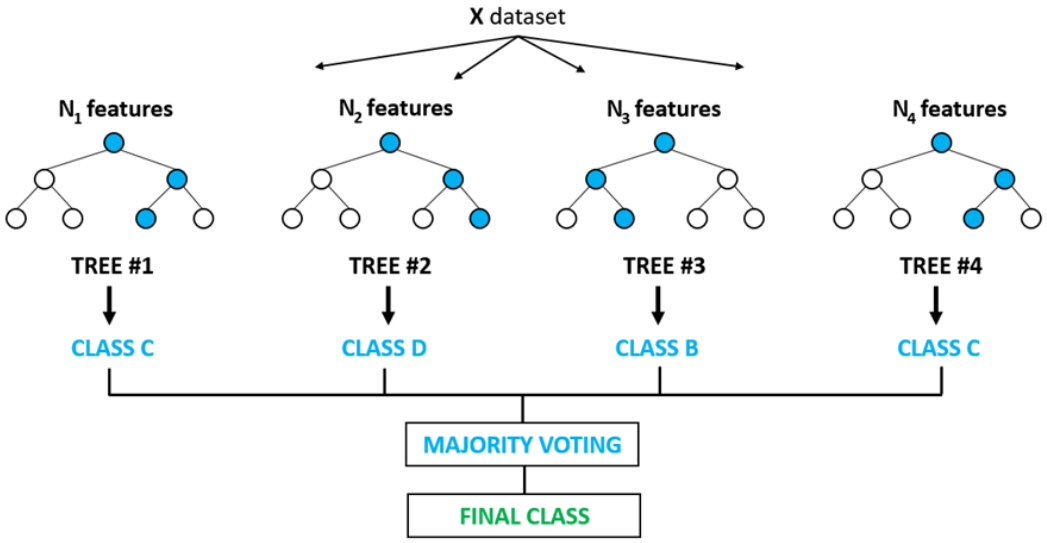


Ilustración 2: Visualización de técnica Bagging

Fuente: (Global software support, 2018)

Para hacer lo anterior, *random forest* utiliza un proceso de muestreo de datos "con reemplazo" denominado *bootstraping* (ilustración 3), donde, de la totalidad de N de datos, previo a la elaboración de cada árbol, se toma una muestra $M < N$ para el entrenamiento, dejando $(N - M)$ datos fuera (éstos se denominan *out of bag (OOB) samples*, y pueden utilizarse en testeo). Del grupo de entrenamiento de M datos, se replican sets de observaciones hasta completar la cantidad original de datos N.

Ilustración 3: Técnica Bootstrapping

Fuente: (Orellana Alvear, 2018)

Finalmente, utilizando lo anterior, se pasa al modelamiento y posterior testeo, siguiendo la metodología mostrada en la ilustración 4 (Orellana Alvear, 2018).

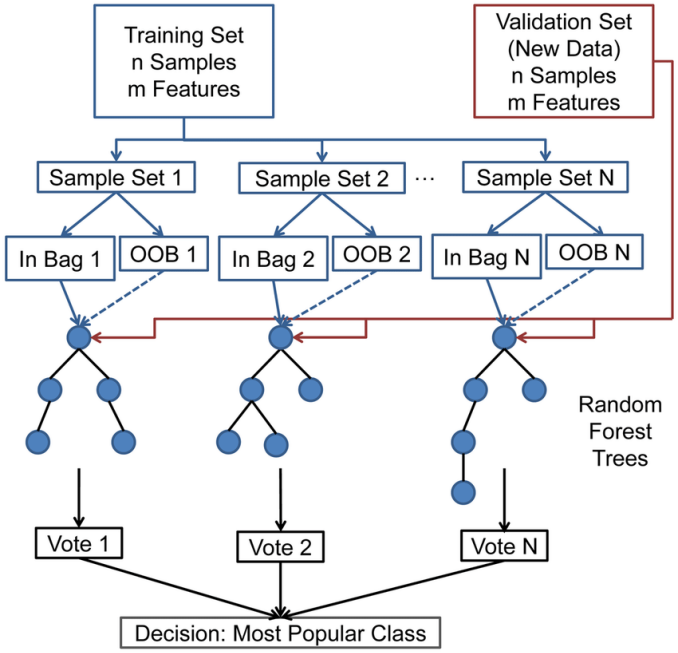


Ilustración 4: Metodología seguida por modelos random forest.

Fuente: (Orellana Alvear, 2018)

4.1.4 Regresión logística (*logit*)

Logit es un modelo de regresión que modela el logaritmo de la probabilidad de pertenecer a una clase en específico sobre dos posibles categorías, luego, la variable dependiente elegida en una regresión logística debe ser una del tipo binaria (Swaminathan, 2018). El modelo toma como input un conjunto de variables independientes determinado, y luego de entrenado, entrega la probabilidad de pertenecer a cada una de las dos clases, así como también los parámetros β asociados a las variables independientes (Weber, 2020).

Suponiendo que se dispone de una combinación lineal X de variables independientes, y una variable dependiente Y a modelar, el modelo *logit* responde a la ecuación 1 para modelar la probabilidad de pertenecer a la clase positiva.

$$Log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = X\beta$$

Ecuación 1: Modelamiento regresión logística

Fuente: (Weber, 2020)

Esta ecuación se puede reordenar de manera de obtener la probabilidad en sí (no el logaritmo), tal como se muestra en la ecuación 2.

$$P(Y=1|X) = \frac{1}{1+e^{-X\beta}}$$

Ecuación 2: Probabilidad de pertenecer a la clase positiva en logit

Fuente: (Weber, 2020)

Para medir el efecto de incidencia de cada una de las variables independientes en la variable dependiente, se utilizan los estimadores β, sin embargo, previo al modelamiento se debe certificar que las variables consideradas cumplan los supuestos de cualquier regresión lineal, de lo contrario se puede llegar a modelos erróneos y resultados sesgados.

4.1.5 Cadenas de markov

Una cadena de markov es un proceso estocástico que describe una secuencia de posibles eventos, donde la probabilidad de que ocurra cada evento sigue la propiedad de markov, es decir, sólo depende del estado alcanzado en el período anterior (cadena de orden 1). Luego, cada cadena de markov representa un sistema que varía su estado a lo largo del tiempo, siendo cada cambio una transición del sistema.

Las cadenas pueden ser procesos a tiempo continuo o discreto, así como también tener un número de estados finito o infinito, sin embargo, para efectos del proyecto, sólo interesan las cadenas de markov finitas a tiempo discreto.

Para trabajar con cadenas de markov finitas se debe determinar previamente una serie de elementos, detallados a continuación (Ching, Huang, Ng, & Siu, 2005):

1) **Un conjunto de estados del sistema** (E): Donde un estado se define como la variable descriptiva de la situación en que se encuentra el proceso en un momento determinado del tiempo. Luego, el conjunto de estados pasa a estar formado por todos los posibles valores que pueden tomar los estados.

2) **La definición de transición**: Momento en que una cadena de markov puede cambiar su estado. Es común que una transición se produzca a intervalos regulares de tiempo.

3) **Una ley de probabilidad condicional**: Que define la probabilidad del nuevo estado en función de los anteriores.

Definidos los elementos, se puede obtener la cadena de markov, que supone con ella la creación de una matriz de probabilidades de transición, también llamada matriz de la cadena, donde se detallan las probabilidades de pasar de un estado $\boldsymbol{i}$ a otro $\boldsymbol{j}$ de la cadena, desde el período $(\boldsymbol{t-1})$ al período $\boldsymbol{t}$. Luego, para todo período $\boldsymbol{t}$ las probabilidades de transición cumplen la siguiente propiedad:

$$\sum_{j=1}^{n} p_{ij} = 1\,, con\; p_{ij} \geq 0$$

Ecuación 3: Propiedad probabilidades de transición

Fuente: (Ching, Huang, Ng, & Siu, 2005)

Si bien lo dicho anteriormente es válido para cadenas de orden 1, también existen cadenas de orden superior ($\boldsymbol{k}$>1), que basan su teoría en las cadenas de markov, pero que siguen el teorema de Chapman-Kolmogorov[8] para el cálculo de las probabilidades de transición en $\boldsymbol{k}$ pasos (Yazlle, 2005). Así, una cadena de orden $\boldsymbol{k}$ con $\boldsymbol{n}$ estados, se puede adaptar a una cadena de orden 1 de manera que los estados de esta cadena sean definidos como los diferentes conjuntos en los que puede estar la cadena superior en las últimas $\boldsymbol{k}$ transiciones. Luego, los estados de la cadena de orden 1 están dados por:

$$X_t = \{E_t, E_{t-1}, \dots, E_{t-k}\}$$

Habiendo así $x = n^k$ posibles estados en la nueva cadena definida.

A partir de lo anterior, la intención es abordar el problema de inactividad definiendo cadenas de markov de orden superior, donde cada estado corresponde a cada posible segmento al que puede pertenecer un taxista según el RFMC ejecutado (detallados en la sección 4.2.1).

[8] Para más detalle, ver anexo I

4.1.6 Métricas de desempeño

Las medidas de evaluación técnica generalmente utilizadas en modelos de predicción, se basan en una tabla de contingencia que describe las instancias predichas correcta e incorrectamente, denominada matriz de confusión. Así, por ejemplo, en un caso de clasificación binaria, la matriz de confusión queda representada por la figura 2.

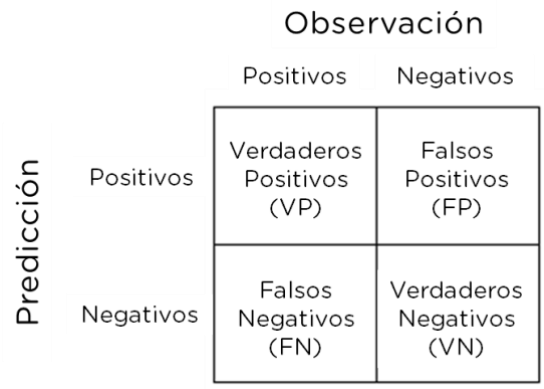


Figura 2: Matriz de confusión para un caso de clasificación binaria

Fuente: Elaboración propia

A partir de esta matriz, es posible definir distintas métricas de evaluación de modelos, tales como las que se detallan en lo que sigue (Riquelme, 2017):

1) ***Precision:*** Responde a la pregunta ¿Qué proporción de predicciones positivas fue correcta? Está dada por la expresión:

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

Fórmula 4: Precision

Fuente: (Riquelme, 2017)

2) ***Recall (o TPR):*** Responde a la pregunta ¿Qué proporción de positivos reales fue identificado correctamente? Está dada por la fórmula:

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

Fórmula 5: Recall

Fuente: (Riquelme, 2017)

3) ***F1-Score:*** También conocido como *F-Measure*, corresponde a la media harmónica entre las métricas *Precision* y *Recall*. Entrega una medida de la certeza del modelo, y se define como:

$$F1 - Score = 2 * \frac{Precision * Recall}{(Precision + Recall)}$$

Fórmula 6: F1-Score

Fuente: (Riquelme, 2017)

4) ***False Positive Rate (FPR):*** Entrega la proporción de negativos reales que se predijeron erróneamente. Sigue la fórmula:

$$FPR = \frac{FP}{FP + VN}$$

Fórmula 7: False positive rate

Fuente: (Riquelme, 2017)

5) ***Negative Predictive Value (NPV):*** Es el símil a *Precision*, pero para las predicciones negativas. Está dado por la fórmula:

$$NPV = \frac{VN}{VN + FN}$$

Fórmula 8: Negative Predictive Value

Fuente: (Riquelme, 2017)

6) ***True Negative Rate (TNR):*** Es el símil a *Recall*, pero para negativos reales. Se define como:

$$TNR = \frac{VN}{VN + FP}$$

Fórmula 9: True negative rate

Fuente: (Riquelme, 2017)

7) ***Accuracy:*** Corresponde a la fracción de predicciones que el modelo es capaz de predecir correctamente, dentro del universo total de observaciones. La métrica está dada por:

$$Accuracy = \frac{VP + VN}{VP + FP + VN + FN}$$

Fórmula 10: Accuracy

Fuente: (Riquelme, 2017)

8) ***ROC Curve:*** Corresponde a una representación gráfica de la tasa de verdaderos positivos (*TPR*) frente a la tasa de falsos positivos (*FPR*), en todos los umbrales de clasificación posibles (valor a partir del cual un caso es considerado positivo). Si bien la curva ROC no tiene una fórmula (puesto que sólo es el gráfico de dos indicadores), sí tiene una métrica de rendimiento asociada, denominada *AUC*, que corresponde al área bajo la curva, y que es interpretada como la probabilidad de que *un clasificador tenga más seguridad de que un ejemplo positivo elegido al azar sea realmente positivo con respecto a que un ejemplo negativo elegido al azar sea positivo* (Google Developers, 2020). Así, esta métrica proporciona una medición agregada del rendimiento en todos los posibles umbrales de clasificación, y permite la comparación entre modelos de distinto tipo. A modo de ejemplo, se muestra la ilustración 5, donde se simula una curva ROC (línea punteada de color azul) y su AUC (área gris bajo la curva).

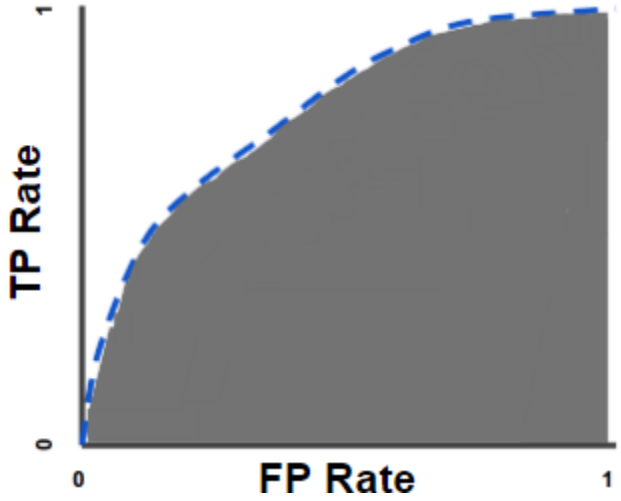


Ilustración 5: Curva ROC y AUC asociada

Fuente: (Google Developers, 2020)

Para la detección de inactividad de taxistas es necesario seleccionar un modelo entre varios posibles. Si bien esto se puede hacer basado en cualquier métrica de desempeño, se espera evitar la variación de resultados dependiendo del umbral de inactividad que se fije, luego, dado que ROC AUC entrega un rendimiento agregado de todos los umbrales, ésta es definida como la métrica de referencia, donde se tiene que a mayor valor, mejor modelo.

4.2 Conceptos propios de la empresa

Previo al modelamiento, es necesario manejar ciertos conocimientos inherentes a la compañía, que ayudan a comprender variables, distribución de clientes, comportamiento de ellos, entre otros, lo que es detallado en la actual sección.

4.2.1 Segmentos RFMC

Según el procedimiento de la sección 4.1.1, cada taxista, dependiendo del score RFMC quincenal que alcance, queda clasificado en un segmento en específico[9], pudiendo distinguirse 5 posibles:

1) **Oro**: Clientes más valorables del programa de fidelización. La compañía busca retenerlos durante toda su vida laboral, puesto que una pérdida o inactividad de ellos, es la que más afecta económicamente al programa.

2) **Plata**: Clientes de valoración media. El área de marketing busca incentivarles un mayor hábito de compra, para intentar que asciendan al segmento Oro.

3) **Bronce**: Clientes de valoración baja. La compañía intenta aumentar su consumo y fidelización con el programa, de tal manera de llevarlos al segmento Plata. Es considerado el segmento con mayores probabilidades de caer en la inactividad.

4) **Infrecuente**: Cliente que no realiza transacciones en la quincena en estudio.

5) **Atípico**: Cliente que su número de transacciones en la quincena, o bien, su monto total quincenal gastado, está fuera de rangos "comunes"[10].

4.2.2 Macro grupos de taxistas

Previo a la predicción de inactividad, por interés de la compañía, son definidos 3 grupos distintos de clientes de acuerdo a su comportamiento con el programa de fidelización (es decir, los grupos se diferencian en base al modo en que los clientes hacen uso del beneficio, ya sea por uso de la App o de tarjeta física).

Luego, dada la separación de clientes, cada grupo es tratado de manera independiente en el proyecto, obteniendo finalmente 3 modelos de inactividad (uno para cada tipo de taxista).

---

[9] Para ver qué scores conforman cada segmento, ver anexo II

[10] Los rangos "comunes" quedan definidos por el boxplot con $límite\ inferior = Q1 - 1{,}5 * IQR$ y $límite\ superior = Q3 + 1{,}5 * IQR$, donde IQR es el rango intercuartílico y $Q1$ y $Q3$ son el primer y tercer cuartil, respectivamente

Los macro grupos a los que se hace referencia son:

1) **Grupo Tarjetas:** Históricamente[11] sólo han usado tarjetas para acceder al beneficio de descuento.

2) **Grupo App:** Históricamente sólo han usado la app para acceder al beneficio de descuento.

3) **Grupo Tarjetas y App:** Históricamente han usado al menos 1 vez cada modalidad para acceder al beneficio de descuento.

[11] En la ventana de tiempo específica, en este caso, desde el 01 de enero de 2018 al 31 de diciembre de 2019

5. Metodología

El problema que se aborda con el proyecto requiere del estudio de las bases de datos transaccionales con las que cuenta la compañía, de manera tal de extraer conocimiento de ellas, así como también de los taxistas generadores de todos esos datos, para tener idea de sus hábitos y comportamiento en general.

En este contexto, la metodología a utilizada corresponde a KDD ("Knowledge Discovery in Databases"), que consiste en llevar a cabo diversos pasos lógicos para llegar a implementar un modelo de minería de datos óptimo y funcional. En estricto rigor, KDD es el proceso mediante el cual se busca descubrir e identificar patrones en los datos que permitan sostener o refutar hipótesis, que a la larga se pueden transformar en decisiones importantes para la empresa u organización. Los pasos de la metodología a aplicar se visualizan en la ilustración 6 y son detallados en lo que sigue:

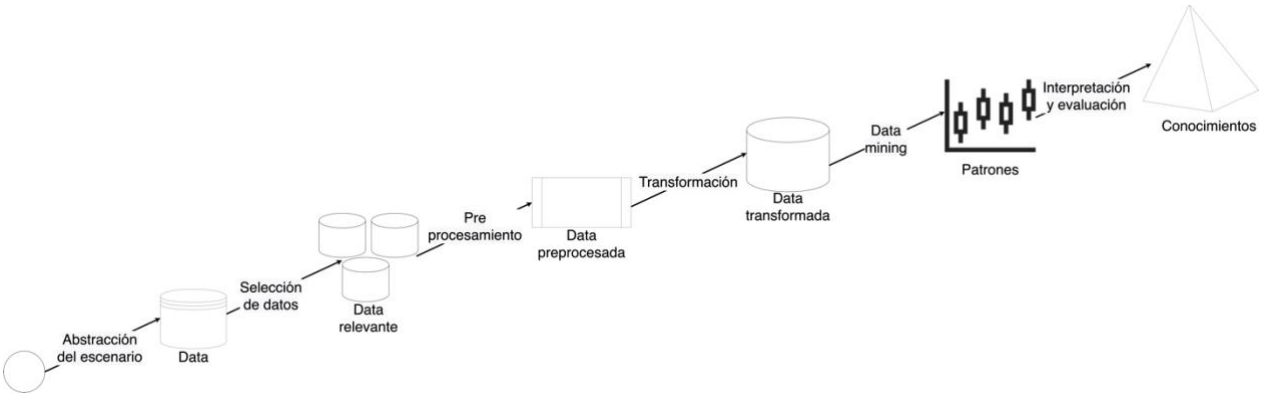


Ilustración 6: Metodología KDD

Fuente: Elaboración propia

5.1 Abstracción del escenario

En primer lugar, se debe entender el funcionamiento completo del negocio en general y todos los detalles del proceso involucrado en el proyecto, de tal manera de llegar a comprender claramente los objetivos de cada área de la organización. Además, en esta etapa se realiza una extensa investigación sobre el estado del arte de los modelos, metodologías y conceptos aplicables a la memoria. Un ejemplo de tarea a realizar es adentrarse en el funcionamiento del programa de fidelización de la compañía con taxis/colectivos, indagando en el paso a paso que ejercen los clientes antes de acceder al beneficio de descuento en combustibles.

5.2 Selección de los datos

Se seleccionan datos históricos de transacciones comprendidas entre enero 2018 y diciembre 2019, detallando fecha, volumen, descuento aplicado, monto total pagado, RUT del taxista que compra, y todas las variables disponibles, con el fin de analizar cuáles pueden ser incluidas en los modelos de inactividad a realizar.

5.3 Limpieza y pre-procesamiento

Se filtran las bases de datos extraídas de posibles datos faltantes (*missing values*), errores de medición, valores fuera de rangos comunes (*outliers*), RUT's de taxistas fraudulentos, entre otros. Así, con la eliminación de imperfecciones en las bases, se evita caer en problemas de sesgo en resultados.

5.4 Transformación de los datos

Se eliminan de las bases de datos todas aquellas variables que no aporten a la realización del objetivo general. Aquí también se considera la creación de nuevas variables a partir de las ya existentes (por ejemplo, monto total gastado por cada taxista en cada quincena de estudio), además de la transformación de otras que no están en formatos amigable con programas computacionales (por ejemplo, conversión de variables categóricas de texto a variables numéricas *dummies*).

5.5 Minería de datos

Hecho lo anterior, se procede a trabajar con herramientas computacionales para la minería de datos. En este paso, se selecciona el modelo que permita la posible detección de inactividad en los clientes analizados. Para este caso en específico, se utilizan cadenas de markov de distinto orden, comparándose luego sus resultados, así como también árboles de decisión, *random forest*, y regresiones logísticas.

Una vez definido el mejor modelo, se hace uso de él, tomando como input la base de datos final (luego del pre-procesamiento y transformación de datos), lo que entrega resultados evaluados en la etapa siguiente.

5.6 Análisis, interpretación y evaluación de resultados

A modo de finalización de la metodología, se realiza un análisis exhaustivo de los resultados obtenidos, de manera de poder interpretar los efectos que tienen en el negocio estudiado, para posteriormente evaluar situaciones hipotéticas sobre la actividad de clientes taxistas en el programa de fidelización, y su impacto en la empresa en general.

Por último, se procede a informar al área de marketing de la empresa sobre qué clientes taxistas son aquellos que se catalogan como “en peligro de inactividad”, para que luego se apliquen estrategias de marketing directo, con la intención de evitar su distanciamiento del programa.

6. Alcances

En la metodología, sólo se tienen en cuenta clientes pertenecientes al programa de fidelización de la compañía, es decir, conductores de taxis/colectivos que acceden al beneficio de descuento en combustibles utilizando su fidelidad. Además, para la realización de modelos, se utiliza RFMC para obtener segmentos quincenales de clientes, de manera tal de poder utilizar esos segmentos como estados en las cadenas de markov modeladas, así como también proporcionar información a los demás modelos a ejecutar.

Por otra parte, el entrenamiento y testeo de cada modelo sólo se efectúa con datos transaccionales comprendidos entre las fechas 01 de enero de 2018 a 31 de diciembre de 2019.

Por último, cabe mencionar que no se realiza experimento piloto.

7. Resultados esperados

Los resultados esperados del trabajo de título van de la mano con los objetivos planteados en la sección 3.2, luego, se esperan los siguientes entregables:

1) Base de datos estandarizada que permite la ejecución de modelos analíticos de manera quincenal

2) Segmento al que pertenecen los taxistas en cada una de las quincenas históricas en estudio

3) Un criterio que permita catalogar clientes como inactivos

4) Un modelo de predicción, capaz de asignar una probabilidad de inactividad a cada cliente (para cada período quincenal de observación), basado en su interacción histórica con la compañía

5) Un criterio que permita clasificar clientes en base a su probabilidad de inactividad

6) Una selección quincenal de clientes propensos a la inactividad, para su posterior inclusión en campañas de marketing directo

Lo anterior, en conjunto, forma un grupo de tareas que, cumplidas, logran alcanzar el objetivo general planteado en la sección 3.1.

8. Desarrollo metodológico

8.1 Limpieza y pre-procesamiento de los datos

Siguiendo la metodología planteada en la sección 5, luego de la interiorización en la empresa y del análisis y estudio de los tipos de modelos y conceptos aplicables al proyecto, se seleccionan como datos todas las transacciones realizadas por taxistas pertenecientes al programa de fidelización durante los años 2018 y 2019 (desde el 01 de enero al 31 de diciembre), donde, a cada transacción se le asocia una fecha, volumen, descuento aplicado, monto total pagado, RUT del comprador, estado de la transacción, entre otras variables disponibles.

Una vez obtenida la "*raw data*", se procede a la limpieza de ella, eliminando todas aquellas transacciones que presentaron errores en su ejecución, es decir, que por algún motivo la transacción finalmente no fue realizada (el campo estado de la transacción detalla esto último, indicando si la venta se completó de manera exitosa o no). Además, dado el objetivo que se tiene con el modelo (identificar a clientes propensos a la inactividad), se eliminan de la base de datos todos aquellos registros que en el campo RUT presentan valores dudosos (menores o iguales a 1.000.000 o mayores o iguales a 30.000.000), o bien, son RUT's de trabajadores pertenecientes a la compañía (los cuales corresponden a casos de fraude), quedando así con un total de 54.431 clientes identificables.

Con la data relevante pre-procesada, y basado en las definiciones entregadas en la sección 4.2.2, cada uno de los 54.431 taxistas identificados es clasificado en sólo uno de los tres macro grupos posibles, obteniendo resultados representados por el gráfico 4, donde se ve que la mayoría de los clientes queda asignado al grupo Tarjetas (69%), y la minoría en el grupo Tarjetas y App (14%).

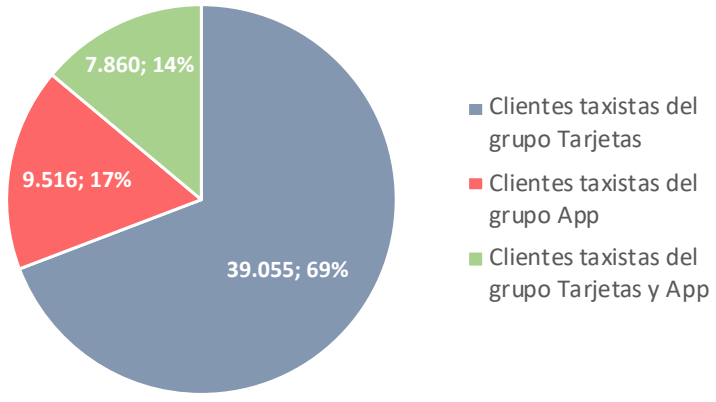


Gráfico 4: Distribución de grupos de muestra de taxistas considerados

Fuente: Elaboración propia

8.2 Definiciones previas

Una vez clarificada la muestra final de clientes, para proceder con el modelamiento, es necesario definir el criterio de inactividad en cada uno de los macro grupos de clientes, es decir, aquella ventana de tiempo de inactividad para la cual un taxista pasa a ser considerado como inactivo. Así, se realiza un "análisis de continuidad de infrecuencias", consistente en: para una ventana de tiempo comprendida entre los períodos $t = 0$ y $t = 11$[12], sobre todos aquellos clientes activos en $t = 0$ (es decir, que realizaron alguna transacción), se analiza cuántos clientes logran acumular desde 1 hasta 11 quincenas consecutivas sin comprar. Lo anterior luego, se repite para distintas ventanas temporales, y se analizan los resultados por medio de gráfico de barras, intentando encontrar el número de períodos a partir del cual la cantidad de clientes por barra se estabiliza[13], es decir, la diferencia de frecuencia entre barras no difiere significativamente.

A modo de ejemplo, en el gráfico 5 se muestran los resultados obtenidos para una de las ventanas temporales utilizadas en el grupo "Tarjetas"[14].

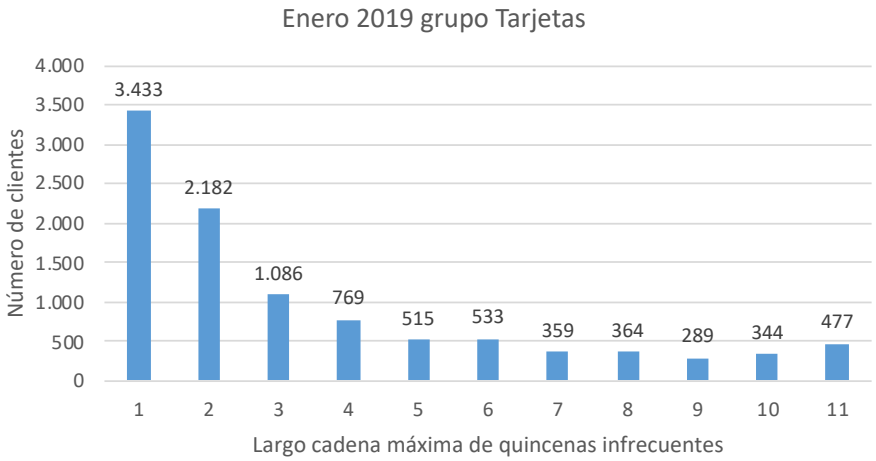


Gráfico 5: Análisis de continuidad de infrecuencias del grupo Tarjetas

(t=0 corresponde a la primera quincena de enero 2019)

Fuente: Elaboración propia

El eje X indica el largo de la cadena máxima de infrecuencias alcanzado, y el eje Y, el número de clientes asociado a esa cadena máxima. En este ejemplo específico, se ve que, después de las 5 infrecuencias, la cantidad de clientes no varía significativamente (lo que sí pasa entre la barra de 4 infrecuencias y la de 5), por lo que, para efectos del grupo tratado, la inactividad es definida luego de que un cliente alcance 5 períodos

[12] En este caso, cada período corresponde a una quincena

[13] Según conversaciones con la contraparte

[14] Los gráficos para todas las ventanas temporales de los grupos Tarjetas, App y Tarjetas y App se muestran en el anexo III

consecutivos como infrecuente (es decir, en 5 quincenas consecutivas o más no realiza transacción alguna).

Análogamente, para los clientes “App”, se obtienen los resultados exhibidos en el gráfico 6, que, en conjunto con decisiones de la contraparte, permite definir un nuevo criterio de inactividad en este grupo: 4 quincenas.

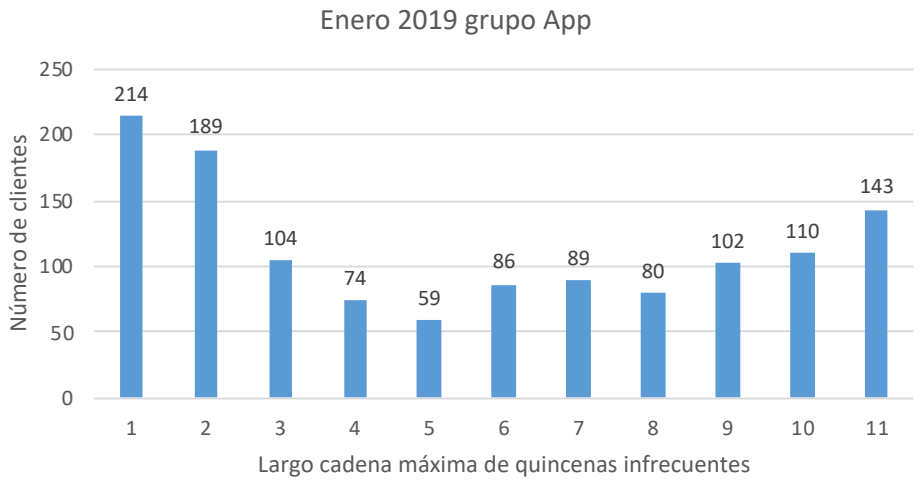


Gráfico 6: Análisis de continuidad de infrecuencias del grupo App

(t=0 corresponde a la primera quincena de enero 2019)

Fuente: Elaboración propia

Por último, se repite el análisis para el tercer grupo de clientes (grupo Tarjetas y App), obteniendo los resultados ilustrados en el gráfico 7, de donde se extrae que el criterio de inactividad para este grupo queda definido en 5 quincenas seguidas o más sin comprar (al igual que en el grupo Tarjetas).

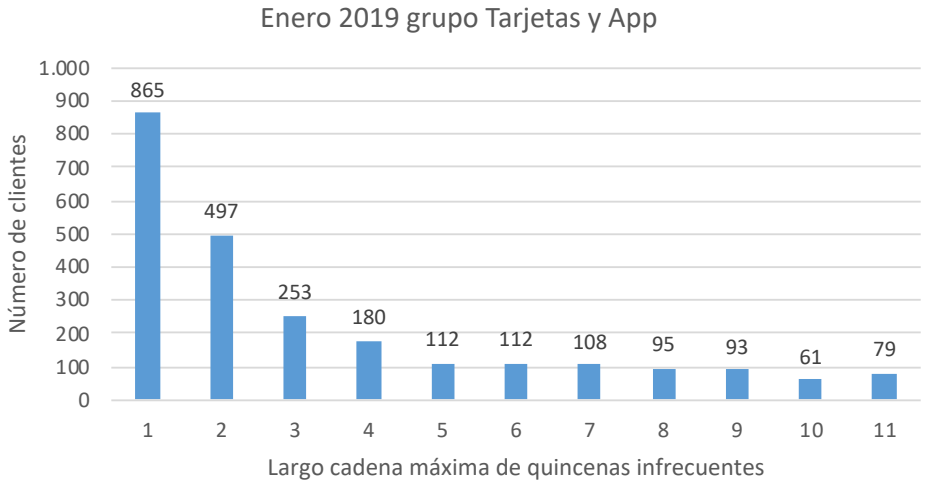


Gráfico 7: Análisis de continuidad de infrecuencias del grupo Tarjetas y App

(t=0 corresponde a la primera quincena de enero 2019)

Fuente: Elaboración propia

Así, a partir de lo desarrollado, se evidencia que el criterio de inactividad es definido según el grupo de clientes tratado, donde los resultados para los grupos Tarjetas y Tarjetas y App son similares, haciendo que sus clientes tengan el mismo criterio de inactividad (5 o más quincenas seguidas sin comprar), mientras que aquellos clientes del grupo App tengan un criterio más corto (4 o más quincenas seguidas sin comprar).

8.3 Análisis descriptivo de los datos

Con el objetivo de entender en profundidad la data con la que se trabaja, siguiendo una lógica de detalle de lo más general a lo más específico, distintos análisis son realizados.

Primero, se hace una exhaustiva inspección de transacciones efectuadas por taxistas, pasando luego a una muestra del comportamiento general de ellos con sus variables de compra quincenales (recencia, frecuencia, monto y consistencia). Detallado lo anterior, se prosigue con una caracterización de los clientes (por medio de la segmentación RFMC), para finalmente entregar descriptivos del problema tratado (inactividad de clientes).

8.3.1 Descripción de transacciones

Para ver el uso a nivel país del programa de fidelización de taxistas, se analizan las transacciones realizadas dentro de la ventana temporal enero 2018 a diciembre 2019 (ambos meses incluidos), obteniendo distinta información, descrita en lo que sigue.

Para comenzar, sobre el total de compras consideradas (N=10.666.240 transacciones), se separan las transacciones según lugar del país donde ocurrieron, dando con los resultados mostrados en el gráfico 8.

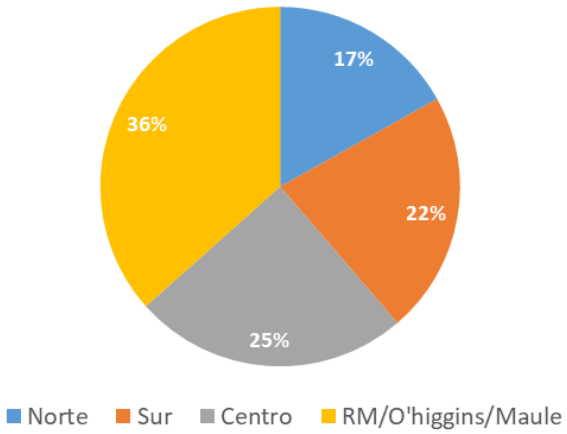


Gráfico 8: Distribución de transacciones totales (enero 2018 – diciembre 2019) según zona del país en que se realizaron.

(N=10.666.240 transacciones)

Fuente: Elaboración propia

En él, se ve que la mayoría de las transacciones ocurren en las regiones Metropolitana, O'Higgins, o del Maule (36%), siendo la zona norte del país (17%) la parte en que menos se ocupa el programa de fidelización. Además, en el gráfico 9 es posible ver que los porcentajes mostrados en el gráfico 8 permaneces más bien constantes en el tiempo, independiente del mes y año observado.

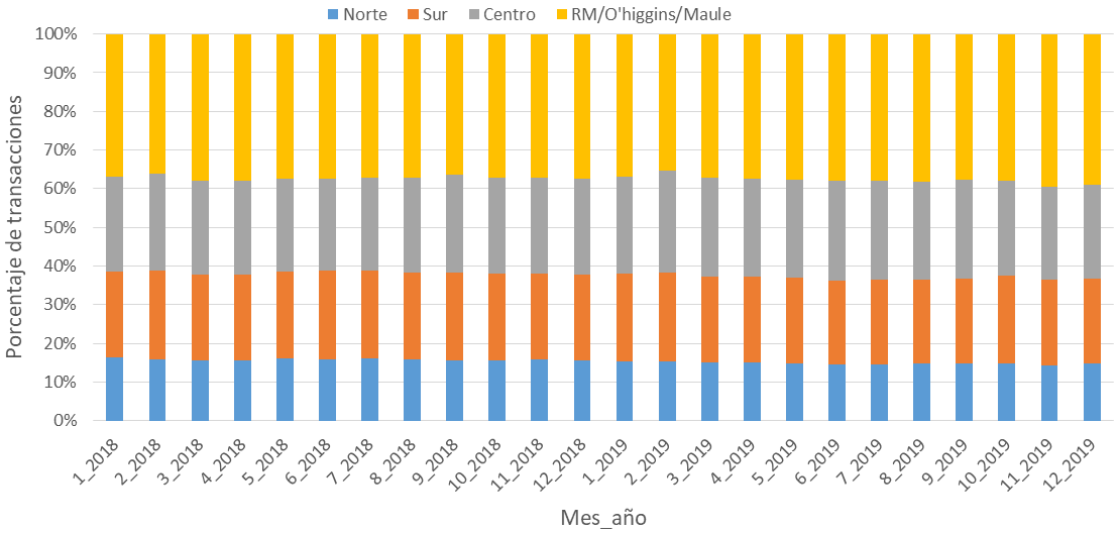


Gráfico 9: Distribución de transacciones mensuales según zona del país en que se realizaron

Fuente: Elaboración propia

Dentro de todas las transacciones, producto de las distintas formas de utilización del beneficio de descuento, también se examina la distribución de transacciones según la modalidad de uso del programa de fidelización, obteniendo que, del total de compras hechas por taxistas entre 2018 y 2019 (N=10.666.240), el 88% corresponde a accesos al beneficio de descuento usando tarjetas físicas, mientras que el 12% restante se atribuye a la app. La razón principal de esto es que el lanzamiento de la aplicación móvil ocurre en agosto de 2018, y desde ahí en adelante fue posible que los clientes la utilizasen.

Haciendo un seguimiento mensual según el gráfico 10, se puede ver también cómo la app va adquiriendo mayor importancia conforme avanza el tiempo desde su creación, logrando, para diciembre 2019, representar el 30% del total de las ventas del programa de fidelización, es decir, 3 de cada 10 transacciones son realizadas con app, mientras que el 70% restante corresponden a transacciones de tarjetas físicas.

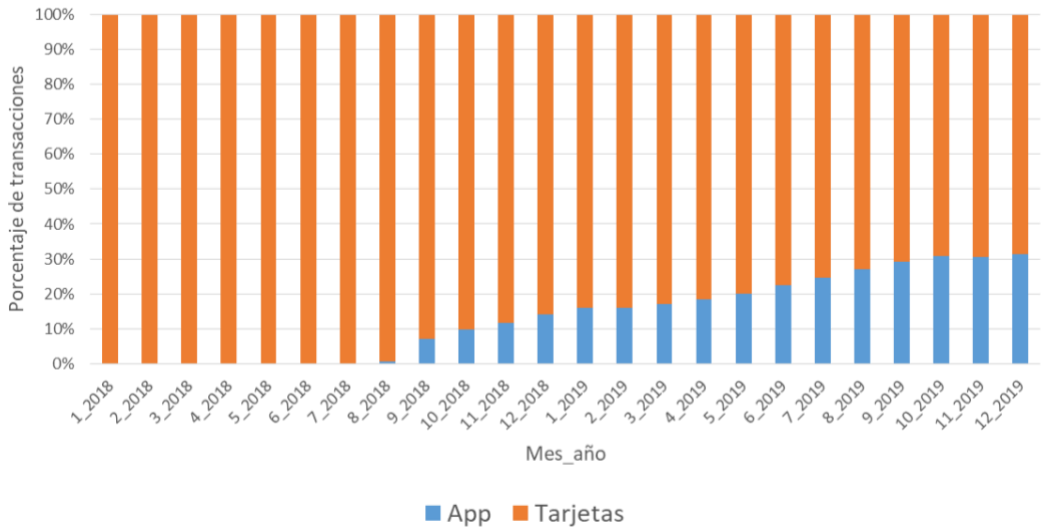


Gráfico 10: Distribución de transacciones mensuales según tipo de uso del beneficio

Fuente: Elaboración propia

Por último, según datos del programa, se tiene que para 2019, el número de transacciones bajó un 16,5% (comparado con 2018), entregando un atisbo de que la compañía disminuye sus ganancias para este último año, luego, con el objetivo de confirmar lo anterior, se realiza un seguimiento de volumen cargado mensualmente por taxistas, desde enero 2018 en adelante, ilustrando los resultados en el gráfico 11, que muestra la clara tendencia a la baja en el volumen cargado por taxistas mes a mes, concluyendo que efectivamente la compañía tiene una disminución de las ganancias percibidas por el programa de fidelización.

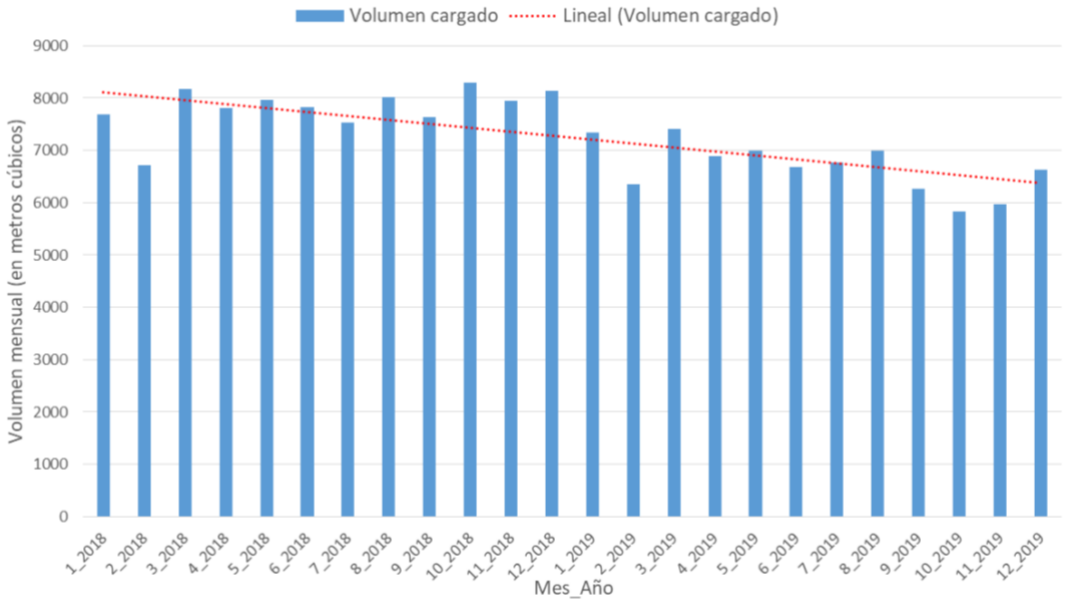


Gráfico 11: Volumen mensual cargado por taxistas utilizando el programa de fidelización

Fuente: Elaboración propia

8.3.2 Caracterización macro grupos de clientes y variables RFMC

En el afán por valorizar a los clientes del programa de fidelización, la compañía segmenta quincenalmente a los taxistas dependiendo del comportamiento que tuvo cada cliente en la quincena estudiada, midiéndolo como una combinación de scores de las variables recencia, frecuencia, monto total cargado y consistencia (tal como se detalla en le sección 4.1.1).

La segmentación se hace para cada uno de los 3 grupos definidos en la sección 4.2.2, es decir, en cada quincena estudiada, se tienen 3 segmentaciones distintas, una para el grupo Tarjetas, otra para el grupo App, y una última para el grupo Tarjetas y App.

La razón principal de la separación de grupos radica en la forma de utilización del programa de fidelización, lo que hace a los clientes comportarse de distinta manera en sus hábitos de compra dependiendo del grupo al que pertenecen. Así, por ejemplo, en el gráfico 12, se muestra quincenalmente el ticket promedio de cada grupo durante el año 2019 (se omite el año 2018, puesto que los grupos App, y Tarjetas y App no existían antes de agosto 2018).

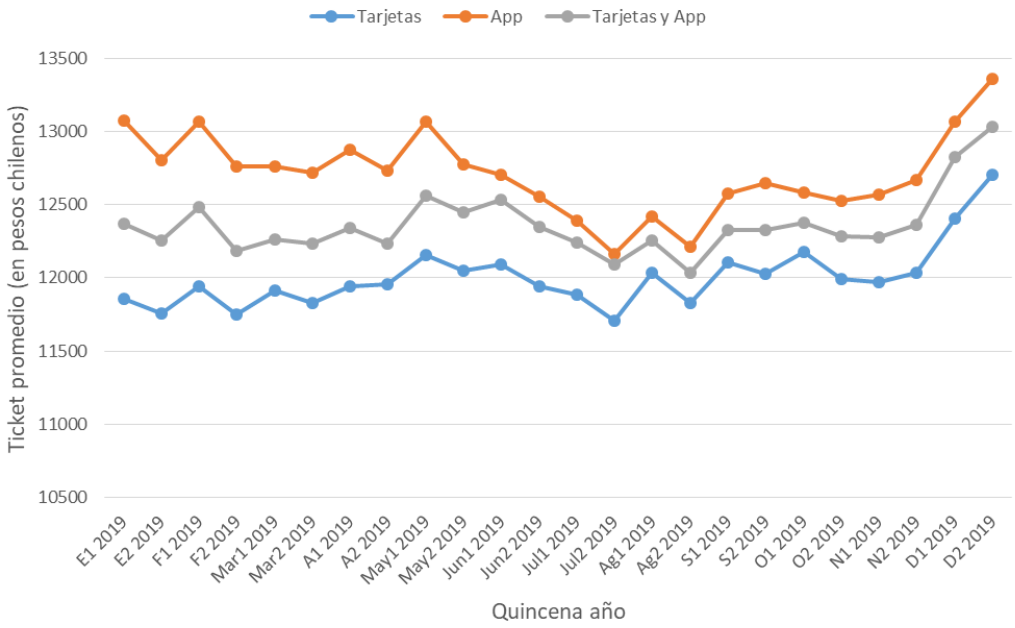


Gráfico 12: Ticket promedio de compra quincenal para los grupos Tarjetas, App, y Tarjetas y App durante el año 2019

Fuente: Elaboración propia

En el gráfico 12, es posible apreciar que los clientes que utilizan el programa sólo a partir de la aplicación móvil, tienen un ticket promedio considerablemente mayor a los que utilizan sólo tarjetas físicas. Por ejemplo, para la segunda quincena de mayo, se puede decir que quienes pertenecen al grupo App gastan, en promedio, aproximadamente $700 pesos más por compra que aquellos integrantes del grupo Tarjetas.

Una vez definidos los grupos, se ejecuta el RFMC, que basa la asignación de scores de las variables analizadas, en la distribución de valores de cada quincena. Así, para tener una concepción de los rangos en que se mueven las variables, se presentan en el cuadro

1, los indicadores mínimo, máximo, mediana, y promedio de recencia, frecuencia, monto y consistencia, para la segunda quincena de junio de 2019, y para todos los grupos a los que se aplica la segmentación[15].

Quincena	Variable	Grupo	Mínimo	Máximo	Mediana	Promedio
Junio 2 2019	Recencia	Tarjetas	0	14	1	2,39
		App	0	14	1	2,45
		Tarjetas y App	0	14	1	1,37
	Frecuencia	Tarjetas	1	23	6	7,26
		App	1	20	6	6,82
		Tarjetas y App	1	31	10	10,78
	Monto total	Tarjetas	$1.000	$267.506	$79.033	$87.085,11
		App	$1.000	$262.029	$82.126	$88.536,42
		Tarjetas y App	$5.009	$347.160	$117.000	$127.815,79
	Consistencia	Tarjetas	1	15	6	6,46
		App	1	15	6	6,19
		Tarjetas y App	1	15	8	8,33

Cuadro 1: Indicadores de variables RFMC para 2ª quincena de junio 2019

Fuente: Elaboración propia

Luego, utilizando algoritmos[16] sobre la distribución de datos de cada variable, se obtiene un nota individual de ellas, las cuales, al concatenarse (en el orden R, F, M, C), originan el score RFMC, definiéndose la categoría de cada cliente en cada grupo y quincena de observación, pudiendo ser Oro, Plata, Bronce, Infrecuente o Atípico, según se define en la sección 4.2.1.

8.3.3 Caracterización segmentos RFMC

Como se menciona en el punto anterior, 5 son las categorías a las cuales un cliente puede pertenecer en cada período de estudio (Oro, Plata, Bronce, Atípico o Infrecuente), y dependiendo de la quincena, varía la cantidad de clientes activos (es decir, aquellos que hacen transacciones), así como también la distribución de clasificación entre grupos. A modo de ejemplo, en el gráfico 13 se muestra la evolución de clientes activos del grupo

[15] Indicadores de variables de todas las quincenas se encuentran en el anexo IV

[16] *Jenks natural breaks*: K-means de 1 dimensión. Su output entrega valores determinados para separar la data en k grupos distintos (en este caso, k=3 grupos)

Tarjetas, donde se puede ver la cantidad de clientes perteneciente a cada segmento (Oro, Plata, Bronce o Atípico) según el RFMC de cada quincena[17].

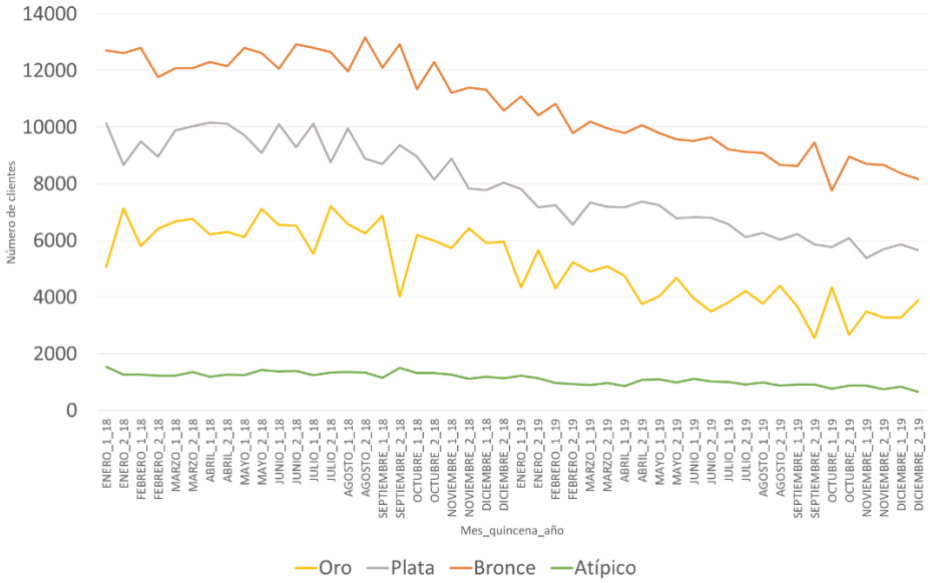


Gráfico 13: Evolución de clientes activos del grupo Tarjetas según segmento al que pertenecen

Fuente: Elaboración propia

En el gráfico 13, queda en evidencia que la cantidad de clientes totales del grupo Tarjetas ha ido en disminución desde enero 2018 a diciembre 2019, sin embargo, la distribución de clientes entre segmentos RFMC permanece relativamente constante, lo que se puede ver, en el gráfico 14, teniendo los segmentos así, porcentajes de clientes más bien establecidos[18].

[17] Dada la alta presencia de Infrecuentes en cada quincena, se opta por dejar de lado este grupo, para no alterar la escala utilizada en el gráfico

[18] Los gráficos 13 y 14 también se obtienen para los grupos App y Tarjetas y App, pero son mostrados en el anexo V

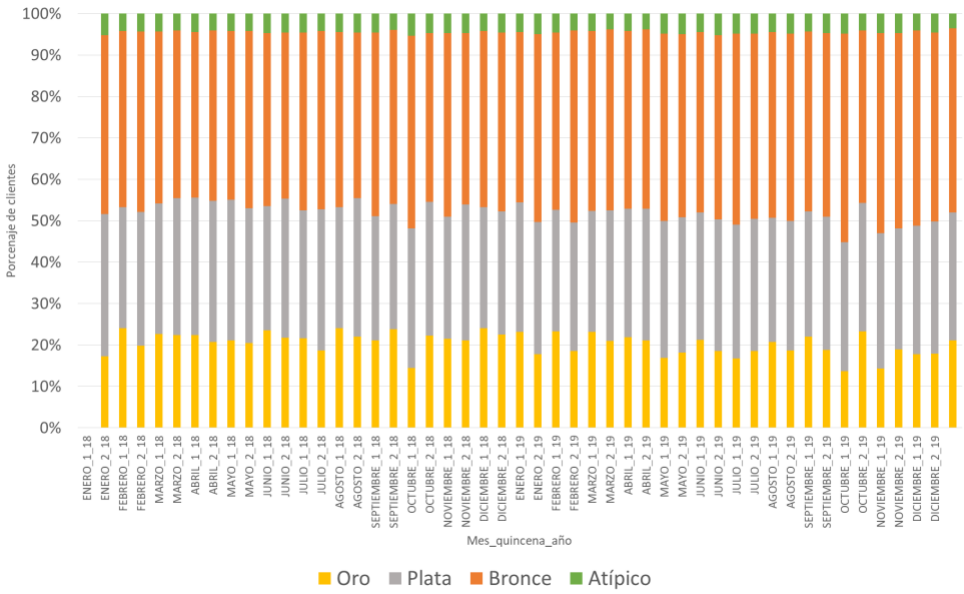


Gráfico 14: Distribución mensual de clientes activos del grupo Tarjetas según segmento al que pertenecen

Fuente: Elaboración propia

Lo anterior da paso a caracterizar los segmentos en términos de su importancia con la compañía, basándose en cuánto ingreso reporta cada uno según su composición de taxistas, obteniendo los resultados ilustrados en la figura 3, que muestran la relación quincenal promedio para los 3 macro grupos a los que se ejecuta el RFMC[19].

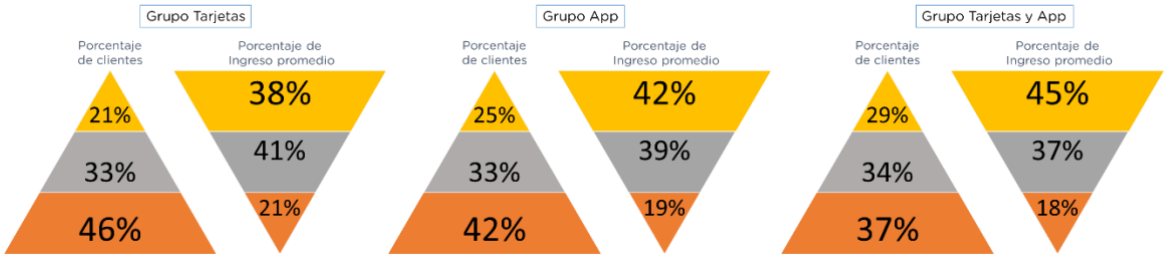


Figura 3: Relación porcentaje promedio de clientes segmentos Oro, Plata y Bronce v/s porcentaje de ingreso promedio que reportan a la compañía, para grupos Tarjetas, App, y Tarjetas y App, respectivamente

Fuente: Elaboración propia

Como se puede ver en la relación de pirámides, para los 3 grupos la composición quincenal del segmento Oro ronda, en promedio, entre el 21% y el 29%, reportando la mayoría de las utilidades del programa de fidelización, siendo el segmento Bronce siempre el de mayor concentración, quienes entregan la menor cantidad de dinero a la compañía, tal como se podría esperar.

[19] Debido a la característica de "outlier" del segmento "Atípico", éste no es considerado en la totalidad de clientes activos, al igual que sus montos totales.

8.3.4 Caracterización inactividad

La inactividad finalmente es definida tal como se detalla en la sección 8.2, es decir, clientes del grupo App pasan a ser catalogados como inactivos si luego de 4 períodos seguidos permanecen como infrecuentes (no han realizado transacciones), mientras que para los grupos Tarjetas y Tarjetas y App, su criterio es de 5 quincenas infrecuentes.

Dado que esta definición de criterio está basada en una combinación de análisis de datos y reglas de negocio, es probable que clientes catalogados como inactivos vuelvan a comprar en un futuro, es decir, vuelvan a formar parte de aquellos clientes activos. Es por esto que es adecuado estudiar qué porcentaje de clientes - de la totalidad de la cartera disponible en cada grupo - se inactiva (y qué porcentaje permanece siempre activo) dentro de la ventana temporal considerada.

Producto de la creación de la app en 2018, para no sesgar la data a utilizar en los grupos App y Tarjetas y App, se considera pertinente omitir los meses de agosto a diciembre 2018, tanto para el análisis de datos como para el modelamiento posterior que pueda haber, quedando así una ventana temporal desde enero 2019 a diciembre 2019 en estos grupos, y de enero 2018 a diciembre 2019 en el grupo Tarjetas.

Para comenzar, en cada grupo de estudio, se analiza la cantidad de clientes que alguna vez se inactivaron dentro de la ventana temporal considerada, es decir, que en algún momento cumplieron la condición de inactividad impuesta (dependiendo si pertenecen a App, Tarjetas o Tarjetas y App). Es así como se obtienen los resultados mostrados en la figura 4. En ellos se ve que el grupo Tarjetas y App es el que presenta menos casos de inactividad en sus clientes, donde el 28% se inactiva 1 o más veces, mientras que en el grupo Tarjetas, muy distinto es el caso, donde sólo el 38% de los taxistas permanece siempre activo, presentando todos los demás clientes al menos un caso de inactividad. Esto último puede estar explicado por la ventana temporal considerada en cada grupo, además de los distintos criterios de inactividad que tienen.

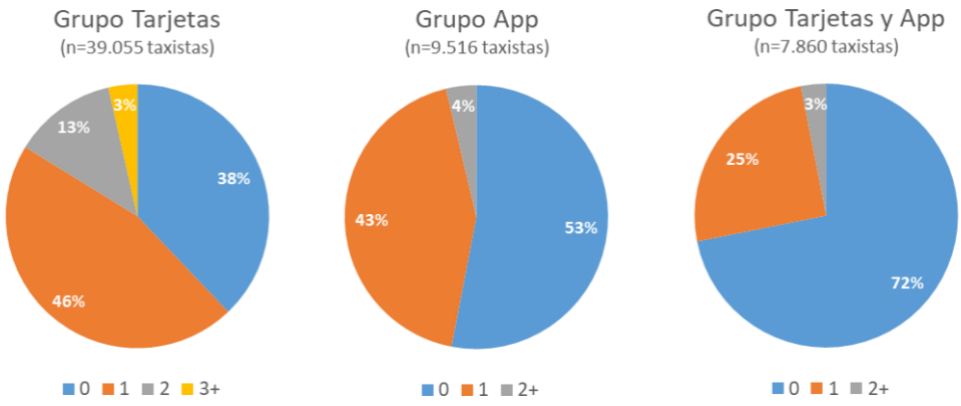


Figura 4: Proporción de número de inactividades detectadas en taxistas, para los grupos Tarjetas, App y Tarjetas y App, respectivamente

Fuente: Elaboración propia

Por otra parte, la importancia de la inactividad de un cliente varía dependiendo de lo valioso que él pueda ser para la compañía, es por esto que también se analiza, del total

de transiciones al estado “inactivo”, cuántas provienen desde segmentos Oro, Plata y Bronce (clasificados según el RFMC quincenal), obteniendo así los resultados mostrados en la figura 5.

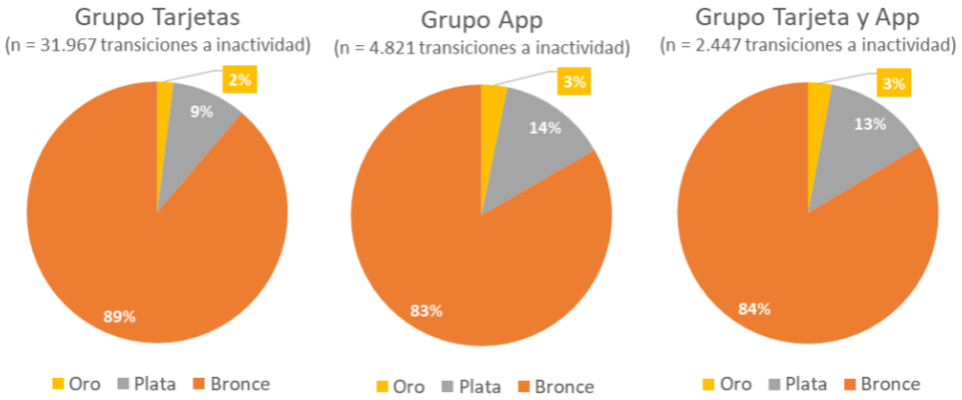


Figura 5: Proporción del total de transiciones a inactividad que provienen de estados Oro, Plata o Bronce, para los grupos Tarjetas, App y Tarjetas y App

Fuente: Elaboración propia

En la figura reciente se puede apreciar que, en la mayoría de los casos en que se declara inactividad de clientes, éstos provenían del segmento “Bronce” (estado previo a pasar a la inactividad), tal como podría esperarse, puesto que este segmento es el menos comprometido con la utilización del programa de fidelización. Por otra parte, también cabe destacar que hay un porcentaje importante de clientes de mayor valor (Oro y Plata) que, a pesar de tener una mayor preocupación con el programa de fidelización, igualmente reportan casos de inactividad. Es así como para el grupo Tarjeta, un 11% de las transiciones a estado “inactivo” provienen desde los dos segmentos más importantes de la compañía (Oro y Plata), aumentando esa cifra en los otros dos grupos, a 16% en grupo Tarjetas y App, y 17% en grupo App; siendo una cantidad considerable de casos de inactividad, y aquellos que la compañía más desea evitar.

A partir de esto, además, se puede hacer el análisis de la tasa de inactividad según segmento de cada grupo de clientes, es decir, qué porcentaje, de la totalidad de clientes pertenecientes a los segmentos Oro, Plata y Bronce, se inactivan quincenalmente (en promedio). Es así como se obtienen los resultados mostrados en el gráfico 15.

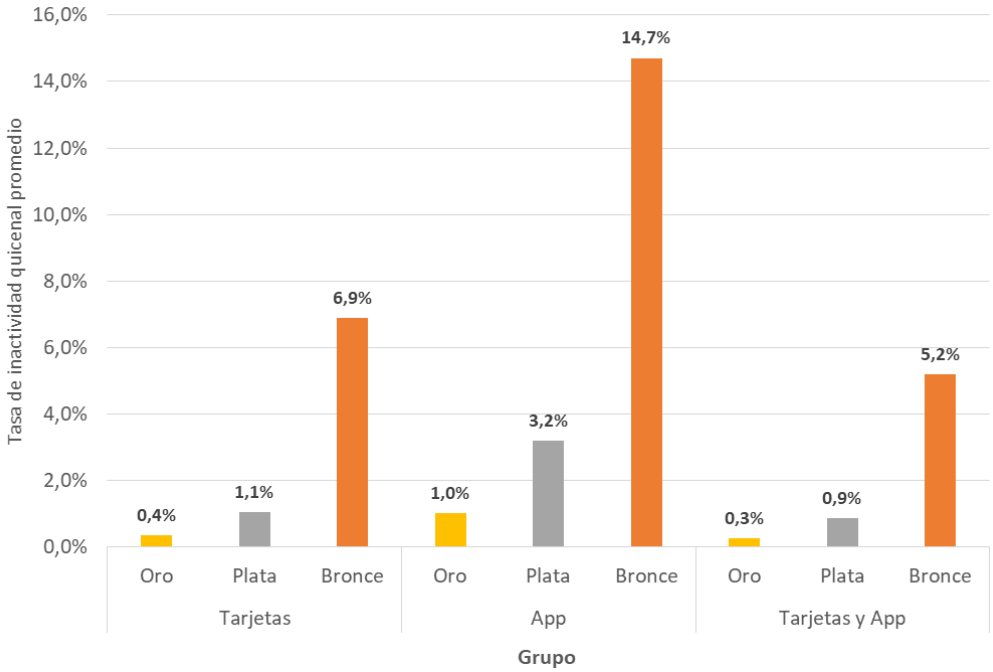


Gráfico 15: Tasa de inactividad quincenal promedio según segmento analizado en cada uno de los macro grupos de clientes

Fuente: Elaboración propia

Es preciso notar cómo el segmento Bronce tiene una alta tendencia a la caída en inactividad para los 3 grupos de estudio (reforzando lo analizado según la figura 5), especialmente en el grupo App, que alcanza un aproximado del 15%, lo que indica que, en promedio, por cada quincena, el 15% de los clientes bronces se inactivan la quincena posterior.

Finalmente, y con el objetivo de evidenciar aún más el problema de inactividad, se hace un seguimiento quincenal a la cartera total de clientes de cada grupo, comparando la cantidad de taxistas "inactivados" y "activados"[20] en cada período. La resta entre las cantidades comparadas supone la obtención de la adquisición neta de actividad, es decir, clientes que se activan o reactivan, menos aquellos que se inactivan. El seguimiento se hace para los tres grupos de estudio, dando con distintos resultados.

Cabe destacar que, dadas las definiciones de inactividad para cada grupo, se grafica desde las quincenas en que es posible observar clientes inactivados, es decir, para grupo Tarjetas, desde las segunda quincena de marzo de 2018 (puesto que los primeros clientes activos se identifican en la primera quincena de enero 2018), para el grupo App, desde la primera quincena de marzo de 2019, y para Tarjetas y App, desde el mismo mes, pero segunda quincena del año 2019.

[20] Un taxista puede ser considerado "activado" de dos formas: si realizó alguna compra en la quincena observada siendo que no había comprado en ninguna quincena anterior a la de observación, o bien, si es un cliente que estaba inactivo hasta la quincena anterior, pero se "reactiva" al nuevamente volver a cargar combustible en la quincena de estudio.

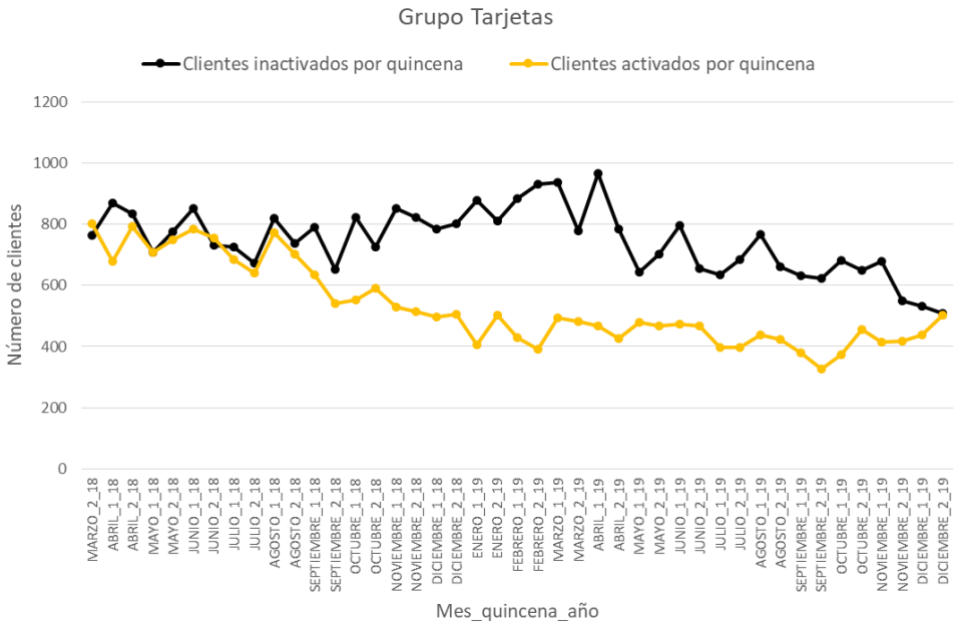


Gráfico 16: Cantidad de clientes inactivados y activados quincenalmente para el grupo Tarjetas

Fuente: Elaboración propia

Como es posible ver del gráfico 16, en el grupo Tarjetas, para casi la totalidad de quincenas, la cantidad de clientes inactivados supera con creces al número de clientes activados, dando una adquisición neta promedio negativa, lo que, ayudado por el gráfico 17, lleva a concluir que en este grupo, en promedio, se pierde actividad de 215 clientes por quincena.

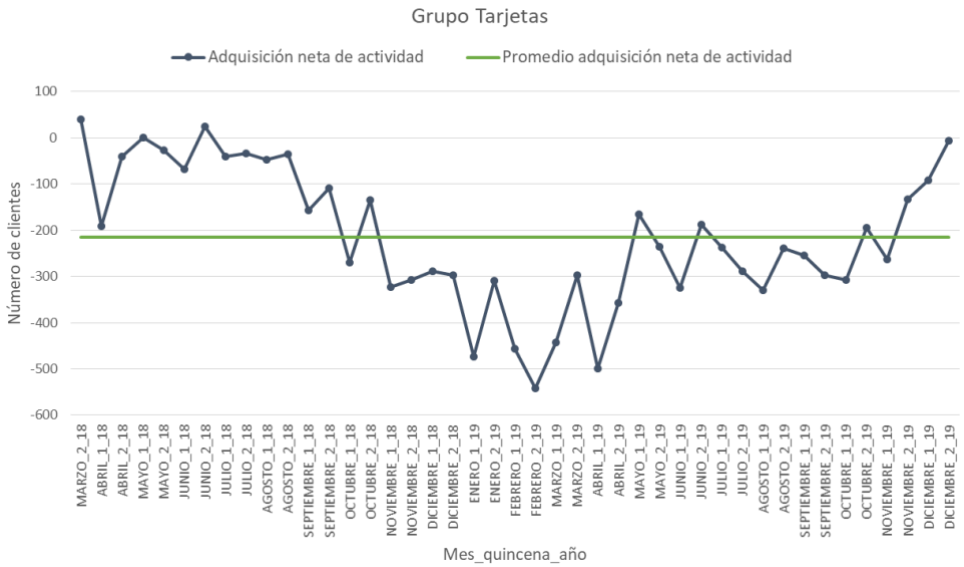


Gráfico 17: Adquisición quincenal neta de actividad de clientes para el grupo Tarjetas

Fuente: Elaboración propia

Para el grupo App se realiza el mismo análisis, obteniendo los resultados mostrados en el gráfico 18, donde es posible ver que este grupo experimenta el fenómeno contrario a lo percibido en el grupo Tarjetas, ya que para casi la totalidad de quincenas estudiadas, se tiene una mayor cantidad de clientes activados que inactivados. Esto se puede explicar, en parte, por la característica de nueva de la aplicación móvil.

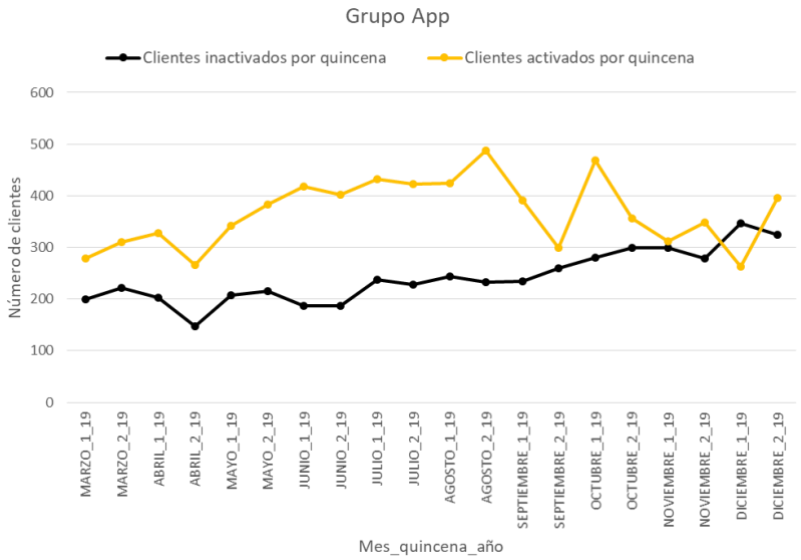


Gráfico 18: Cantidad de clientes inactivados y activados quincenalmente para el grupo App

Fuente: Elaboración propia

Lo anterior, traducido en adquisición neta de actividad, da con los resultados mostrados en el gráfico 19, donde se puede concluir que en el grupo App, en promedio, quincenalmente, se gana actividad de 125 clientes.

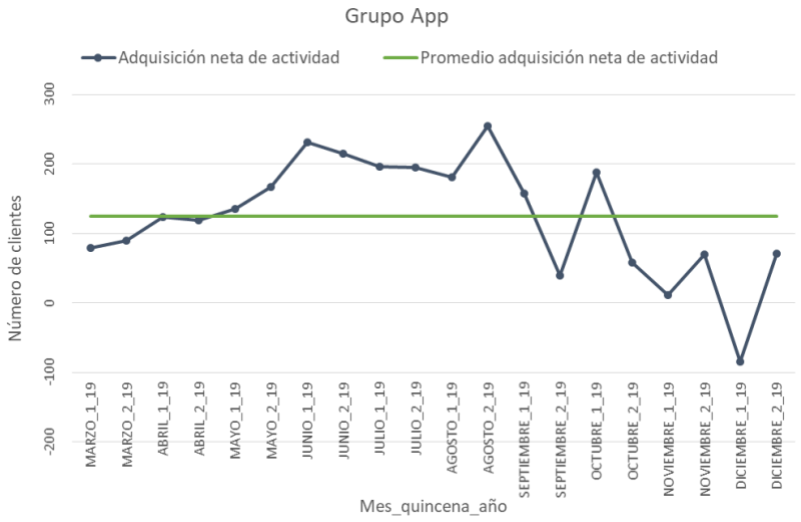


Gráfico 19: Adquisición quincenal neta de actividad de clientes para el grupo App

Fuente: Elaboración propia

Para finalizar, análogamente a lo hecho para los demás grupos de clientes, se realiza el análisis de actividad para el grupo Tarjetas y App, llegando a los resultados ilustrados en el gráfico 20.

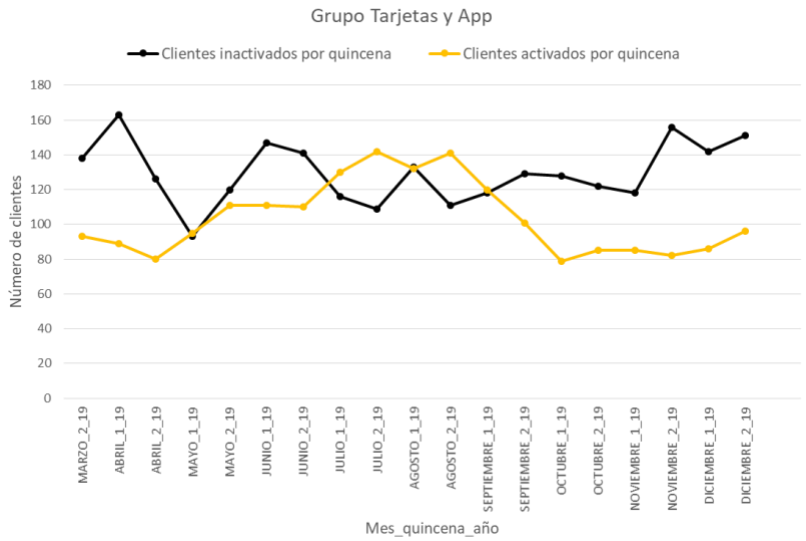


Gráfico 20: Cantidad de clientes inactivados y activados quincenalmente para el grupo Tarjetas y App

Fuente: Elaboración propia

En este grupo, nuevamente se ve una tendencia a haber mayor cantidad de clientes inactivados que activados por quincena, tal como sucede en el grupo Tarjetas, sin embargo, a pesar de ser igualmente negativa, la adquisición neta promedio para el grupo Tarjetas y App es mayor, perdiendo actividad de 26 clientes por quincena, tal como se puede ver en el gráfico 21.

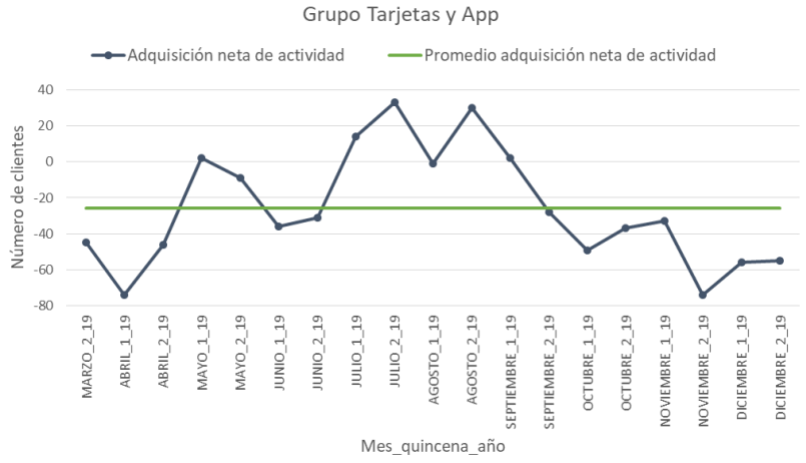


Gráfico 21: Adquisición quincenal neta de actividad de clientes para el grupo Tarjetas y App

Fuente: Elaboración propia

Ahora bien, englobando el programa de fidelización como un todo, es decir, agrupando los 3 grupos, el análisis de adquisición neta de actividad queda tal como se muestra en el gráfico 22.

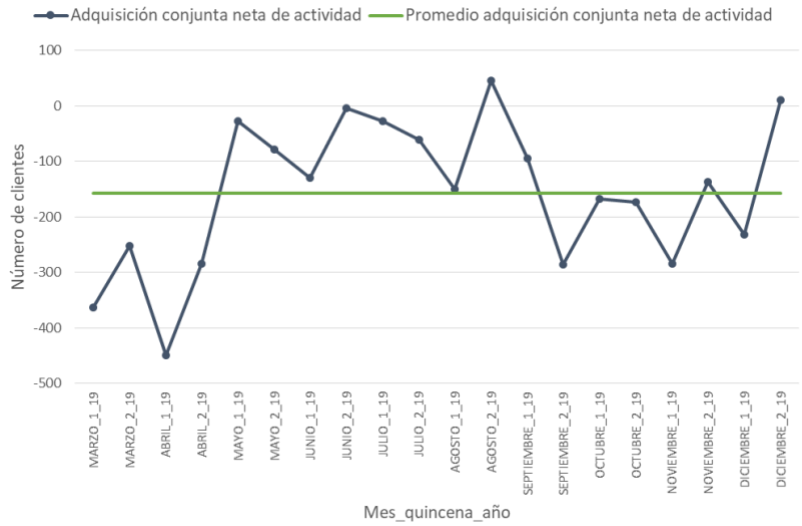


Gráfico 22: Adquisición conjunta neta quincenal

Fuente: Elaboración propia

Aquí se puede ver que el programa en sí, como composición de sus tres partes (grupos), tiene una activación neta negativa constante de clientes, que en promedio, se estima en aproximadamente -158 clientes por quincena. Luego, siguiendo la lógica de que atraer clientes nuevos es más costoso y difícil que retener antiguos, se hace necesario disminuir las tasas de inactividad mostradas para evitar ineficiencias en la compañía.

8.4 Obtención de la muestra

Previo al modelamiento, y descrita la data utilizada en el proyecto, buscando un correcto entendimiento de la ejecución de los modelos, es necesario describir el paso a paso seguido para la obtención de la muestra final de observaciones, sobre la cual se entrenan y testean los modelos de predicción de inactividad.

Para cada cliente taxista, basado en la segmentación RFMC, se tiene el historial de segmentos a los que perteneció en cada quincena pasada, lo que puede ser representado como una secuencia de transiciones. Estas secuencias, a fin de satisfacer las necesidades de cada modelo, son modificadas según pasos sucesivos, detallados en lo que sigue.

8.4.1 Identificación y "adelantamiento" de inactividad

Primeramente, para cada cliente de los distintos grupos, se identifica si su secuencia de transiciones presenta inactividad en algún momento o no (según los criterios de inactividad definidos en cada grupo, sección 8.2). Así, por ejemplo, para el cliente A mostrado en la figura 6, suponiendo que éste pertenece al grupo Tarjetas, se ve que entre los períodos 7 y 11, el taxista se mantuvo en el estado "Infrecuente" (es decir, no realizó transacción alguna), cumpliendo con el criterio de inactividad (5 períodos o más sin comprar). Con esto, se modifica su cadena de transiciones, cambiando el último estado "Infrecuente" por el estado "Inactivo".

Cabe destacar que, con los modelos, se busca predecir la inactividad desde un estado activo, es decir, hay un deseo de anteponerse a los hechos para tomar acciones comerciales antes de que los taxistas se distancien del programa. Luego, con el objetivo de evitar tener que esperar a que el cliente se inactive para hacer marketing, como último paso, el estado "Inactivo" identificado anteriormente, se adelanta hacia la primera posición de la cadena de infrecuencias, redefiniéndose la secuencia de transiciones del cliente A, quedando tal como se muestra en la última secuencia de la figura 6.

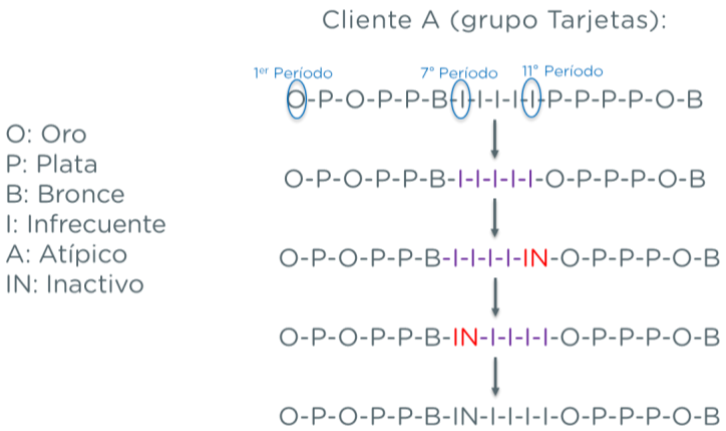


Figura 6: Ejemplo de redefinición de cadenas basado en el criterio de inactividad

Fuente: Elaboración propia

8.4.2 Selección de cadenas de actividad

Una vez redefinidas las secuencias de transiciones de todos los clientes (según el criterio de inactividad de cada grupo), se procede a particionar esas cadenas dependiendo de si el cliente cayó en inactividad o no (y en qué período lo hizo), lo que queda representado por la figura 7.

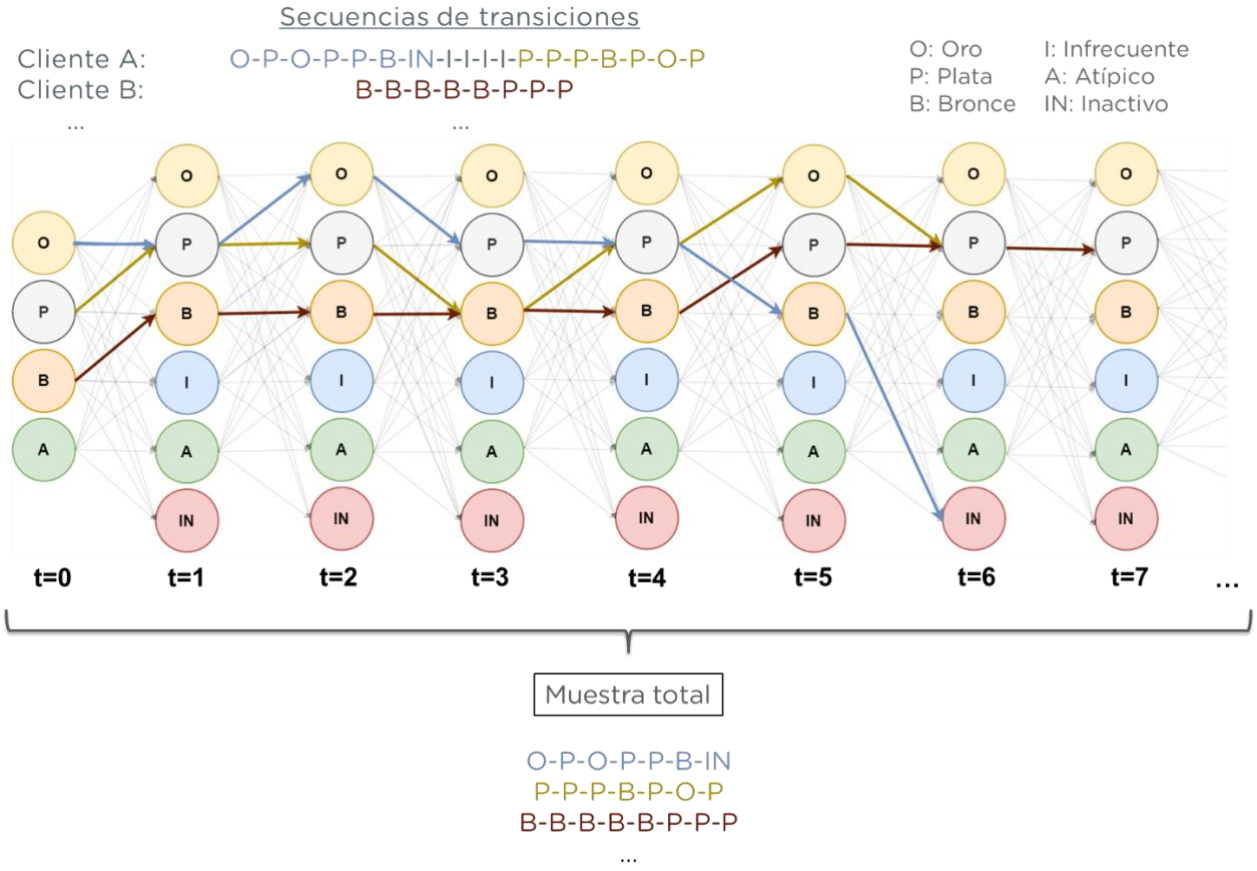


Figura 7: Ejemplos de selección de cadenas de actividad según cada cliente

Fuente: Elaboración propia

En la figura, utilizando el ejemplo del cliente A, se ve que él cae en inactividad una vez, pero luego vuelve a la actividad en el futuro, donde, de ahí en más, no registra inactividad nuevamente, pudiendo distinguirse así 2 cadenas de actividad independientes, aquella representada por las flechas de color azul, y la secuencia representada por flechas de color dorado.

Por otra parte, se muestra el cliente B, que nunca alcanza el estado “Inactivo” en su secuencia de transiciones, así, para él, se identifica sólo una cadena de actividad, correspondiente a la secuencia de transiciones original, representada con las flechas de color café.

Finalmente, se itera el algoritmo sobre la totalidad de clientes, extrayendo cadenas de actividad de manera independiente, las cuales se van almacenando en un conjunto de denominado “muestra total”.

8.5 Modelamiento

Una vez obtenida la muestra total de observaciones, a modo de poder comparar resultados, distintos modelos son ejecutados, para finalmente elegir uno en base a la mejor métrica ROC AUC.

Entre los modelos comparados se encuentran cadenas de markov de orden superior, árboles de decisión*, random forest* y *logit*.

8.5.1 Cadenas de markov

A partir de la muestra total de observaciones, se procede a encontrar el modelo supervisado de markov que mejor ajusta la predicción de inactividad, teniendo tres parámetros claves a definir: el orden del modelo a utilizar, el largo de las cadenas de entrenamiento, y el umbral de inactividad, es decir, la probabilidad desde la cual un cliente pasa a ser considerado en peligro de inactividad para la siguiente quincena en estudio.

Tal como se menciona en la sección 4.1.5, las cadenas de markov requieren definir el orden de la cadena, esto equivale a decir, cuántos períodos se analiza hacia atrás desde un tiempo $t = 0$, para predecir el estado en $t = 1$. Por ejemplo, orden 5, son 5 períodos (en este caso, 5 quincenas), por lo que, si se quisiera predecir el estado de un cliente para la 1ª quincena de diciembre 2019, el modelo utilizaría desde la 2ª quincena de septiembre 2019 hasta la 2ª quincena de noviembre 2019; entregando la probabilidad de terminar en cada posible estado en la quincena estudiada. Los estados son, en este proyecto, las clasificaciones RFMC de la sección 4.2.1, más la clase de inactividad descrita en 8.2.

Para el modelamiento, se cuenta con 2 años (2018 y 2019) de transiciones quincenales de clientes para el grupo Tarjetas, mientras que con 1 año (2019) para los grupos App y Tarjetas y App, lo que permite entrenar y testear distintos modelos de markov de orden superior, a partir de las fases sucesivas descritas a continuación.

8.5.1.1 Consideración de últimos movimientos

Sobre la “muestra total” de transiciones resultante de la obtención de la muestra (figura 7), dependiendo del orden n de la cadena de markov que se quiera entrenar, se seleccionan sólo los $(n + 1)$ últimos estados, para el posterior entrenamiento y testeo del modelo. Esto se fundamenta principalmente en que, cuando el largo temporal de la data (Z) es igual al orden del modelo (n), el modelo ajusta solo una vez; y si se tiene $Z > n$, la cadena ajusta todas las posibles combinaciones de n estados sucesivos, para finalmente entregar una distribución de probabilidad de pasar de una combinación de estados a otro estado determinado. Luego, al considerar la totalidad de transiciones, las probabilidades de inactividad entregadas por la matriz de transición son subestimadas, lo que no sucede

al considerar los $(n+1)$ últimos estados. La figura 8 ilustra un ejemplo para el pre-entrenamiento de cadenas de orden 3.

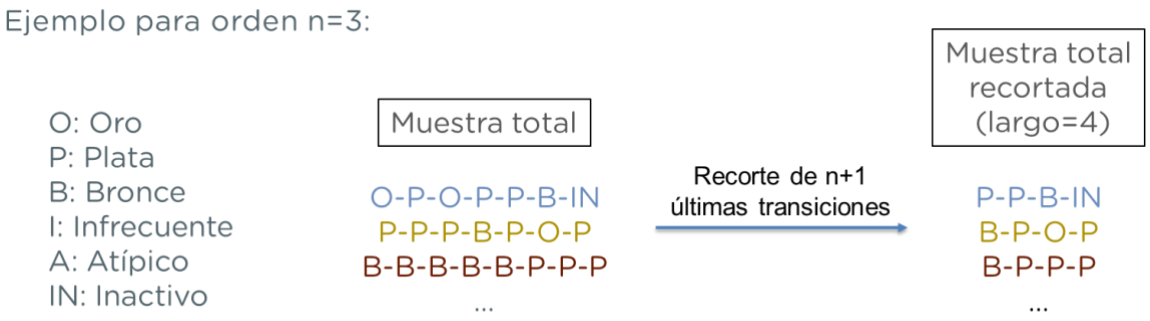


Figura 8: Ejemplo recorte de cadenas según el orden del modelo a entrenar

Fuente: Elaboración propia

8.5.1.2 Entrenamiento y testeo de modelos

Definido el modelo markoviano a entrenar y testear, y obtenida luego la "muestra total recortada" (figura 8), se separa ésta en sets de entrenamiento (80%) y testeo (20%), donde, a partir del entrenamiento, se genera una matriz de transición con las probabilidades de cambio entre estados, tal como simula la figura 9 para un modelo de markov de orden 3.

Ejemplo para orden n=3:

O: Oro
P: Plata
B: Bronce
I: Infrecuente
A: Atípico
IN: Inactivo

Estado en t

Estados en (t-3), (t-2) y (t-1)

Estado	O	P	B	I	A	IN
O-O-O	0.72	0.11	0.02	-	-	0.14
O-O-P	0.24	0.17	0.03	0.02	0.01	0.54
...	...	...	...	...	...	...
B-B-O	0.31	0.26	0.06	0.03	0.03	0.31
B-B-B	-	0.05	0.28	0.05	-	0.63
...	...	...	...	...	...	...

Figura 9: Matriz de transición para cadena de markov de orden 3

(La suma de cada fila puede no ser exactamente 1, puesto que las probabilidades son valores redondeados)

Fuente: Elaboración propia

A partir de la matriz de transición, se testea cada cadena perteneciente al set de testeo, según cada umbral de inactividad fijado. A modo de ejemplo, se presenta el cuadro 2, que simula el testeo del modelo de orden 3, para un umbral de inactividad del 25%.

Umbral de inactividad fijado	Set de testeo	Primeros n estados	Probabilidad de pasar a "Inactivo" (según matriz de transición)	¿Modelo predice inactividad? ($¿Prob. \geq umbral?$)	¿Ocurre Inactividad en la realidad? (¿Último estado es "IN"?)	Resultado testeo
25%	O-O-O-IN	O-O-O	14%	No	Sí	Falso negativo
	B-B-B-IN	B-B-B	63%	Sí	Sí	Verdadero positivo
	O-O-P-O	O-O-P	54%	Sí	No	Falso positivo
	...	...	...	...	...	...

Cuadro 2: Ejemplo mecanismo de testeo modelo de markov de orden 3, para un umbral de inactividad específico

Fuente: Elaboración propia

La primera observación del set de testeo, se ve que tuvo comportamiento "Oro" en los primeros tres períodos, lo que, según la matriz de transición extraída del entrenamiento (figura 9), entrega una probabilidad de 14% de pasar a la inactividad en el siguiente período. Este porcentaje es menor al umbral fijado (para este ejemplo, 25%), haciendo que el modelo no prediga inactividad, siendo que en la realidad sí ocurrió (el último estado es "Inactivo"), fallando así el modelo y presentándose un falso negativo como resultado.

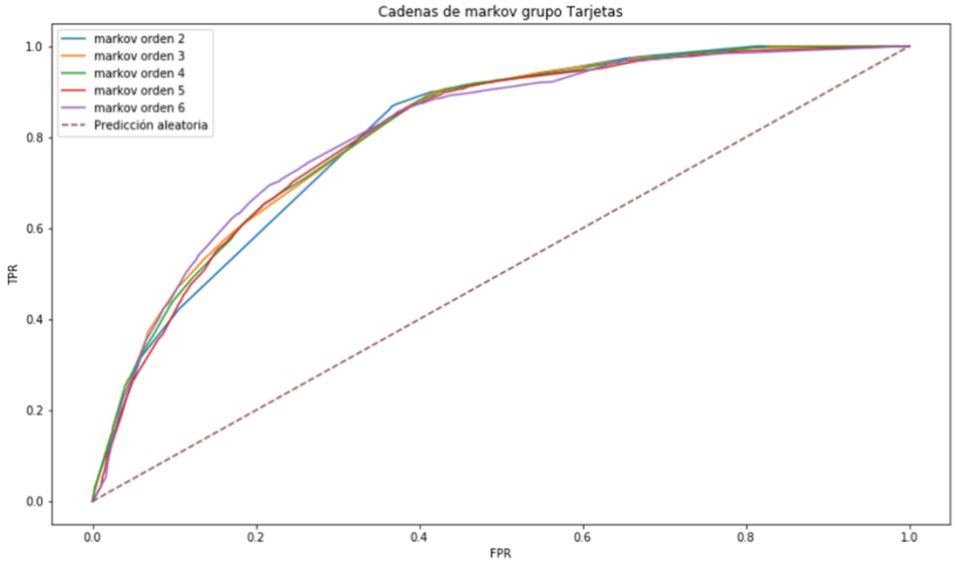
Análogamente para la segunda observación, se ve que el modelo sí predice inactividad, puesto que la probabilidad de terminar en estado "Inactivo" luego de transcurridos 3 períodos como "Bronce" es de 63%, valor superior a 25% (umbral), acertando en este caso, puesto que la transición a la inactividad sí ocurrió realmente, resultando en un verdadero positivo.

Iterando sobre la totalidad de cadenas pertenecientes al set de testeo, se obtiene una matriz de confusión por cada umbral fijado en el intervalo $[0\%, 100\%]$, que posibilita el cálculo de las métricas de rendimiento descritas en la sección 4.1.6.

8.5.1.3 Resultados

Ejecutados cinco modelos de markov distintos (de órdenes 2, 3, 4, 5 y 6), para cada umbral fijado, se calculan las métricas de desempeño, utilizadas luego para graficar las coordenadas (FPR, TPR) según los umbrales elegidos. Esto permite la construcción de la curva ROC, a partir de la cual se calcula también su área bajo la curva (AUC), obteniendo una métrica de rendimiento agregado de cada modelo markoviano, haciendo posible la comparación entre ellos.

Los resultados obtenidos para cada grupo analizado se ilustran en los gráficos 23, 24, y 25, donde, previo a su análisis, cabe señalar que hay ciertos órdenes para los cuales la métrica ROC AUC es la misma. En estos casos, se considera la cadena de menor orden como la de mejor desempeño, principalmente por la aplicación que el modelo tiene en la realidad (menor orden implica mayor capacidad de reacción en el corto plazo).



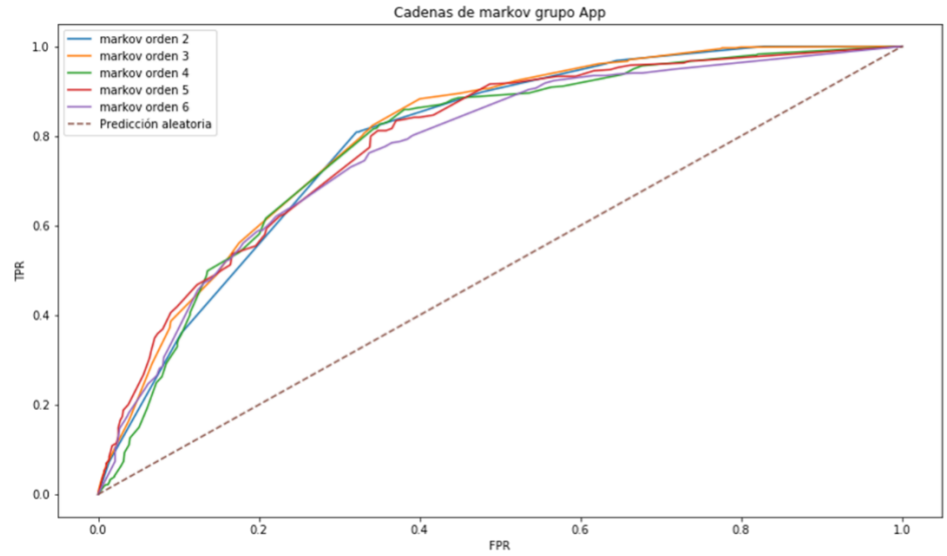
Grupo Tarjetas		
Variables	Markov de orden	ROC AUC
RFMC	2	81%
	3	81%
	4	81%
	5	81%
	6	81%

Gráfico 23: Resultados de cadenas de markov ejecutadas para el grupo Tarjetas

Fuente: Elaboración propia

Del gráfico 23, se puede ver que para el grupo Tarjetas, en todas las cadenas modeladas, se alcanza el mismo valor ROC AUC de 81%, sin embargo, por la razón expuesta anteriormente, se concluye que el modelo markoviano que reporta mejor rendimiento en este grupo corresponde la cadena de orden 2.

Por otra parte, analizando las cadenas de markov ejecutadas para el grupo App, es posible ver que los resultados cambian (gráfico 24), siendo para este grupo la cadena de orden 3 aquella con mejor rendimiento, logrando un valor de 80% en la métrica de interés.



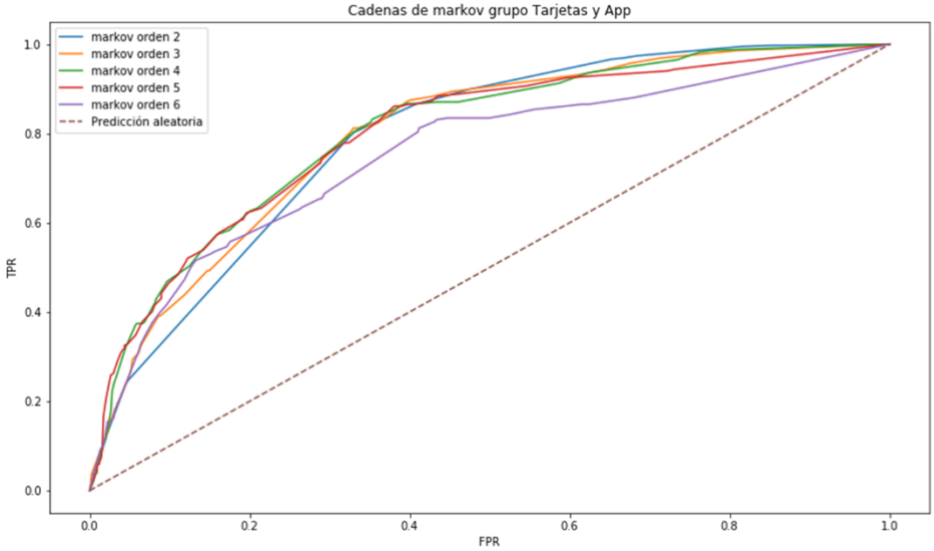
Grupo App		
Variables	Markov de orden	ROC AUC
RFMC	2	79%
	3	80%
	4	78%
	5	79%
	6	77%

Gráfico 24: Resultados de cadenas de markov ejecutadas para el grupo App

Fuente: Elaboración propia

Finalmente, y de manera análoga a lo hecho con los demás grupos, utilizando el gráfico 25, se que ve las cadenas de markov ejecutadas en el grupo Tarjetas y App presentan

un valor máximo de 80% en ROC AUC, que es común entre las cadenas de orden 3, 4 y 5. Luego, nuevamente basado en que a menor orden, mejor capacidad de reacción para la compañía, se concluye que para este grupo, la cadena de markov con mejor rendimiento es aquella de orden 3.



Grupo Tarjetas y App		
Variables	Markov de orden	ROC AUC
RFMC	2	79%
	3	80%
	4	80%
	5	80%
	6	75%

Gráfico 25: Resultados de cadenas de markov ejecutadas para el grupo Tarjetas y App

Fuente: Elaboración propia

8.5.2 Árboles de decisión, *random forest* y *logit*

Obtenidos los resultados para cadenas de markov, se procede al modelamiento de inactividad utilizando otros modelos, en específico, árboles de decisión, *random forest* y *logit*.

Producto de que este nuevo tipo de modelos admite el uso de data continua (a diferencia de las cadenas de markov utilizadas), y con el objetivo de hacer comparables los resultados de todos los modelos ejecutados, son hechos ciertos ajustes a la muestra total obtenida en la sección 8.4.2 (figura 7), según el procedimiento descrito en lo que sigue.

8.5.2.1 Obtención de variables

Para cada una de las observaciones de la muestra, primero se ve si termina en inactividad o no, lo que es denotado con una variable binaria. Ésta corresponderá a la variable dependiente de los modelos (ver figura 10).

O: Oro
P: Plata
B: Bronce
I: Infrecuente
A: Atípico
IN: Inactivo

Observación N°	Muestra total	¿Es Inactividad?
1	O-P-O-P-P-B-IN	1
2	P-P-P-B-P-O-P	0
3	B-B-B-B-B-P-P-P	0
…	…	…

Figura 10: Ejemplo de obtención de variable dependiente para distintas observaciones

Fuente: Elaboración propia

Por otra parte, como cada observación es finalmente una cadena de transiciones de un cliente en particular en un período determinado, es posible identificar su comportamiento transaccional precisamente en esa ventana temporal. A modo de ejemplo, se puede suponer que la observación N°2 de la figura 10, considera el período desde la 1ª quincena de marzo a la 1ª quincena de junio (7 quincenas sucesivas). Conociendo la persona que realizó tales transacciones, es posible calcular distintas variables que detallan el comportamiento de compra de esa observación, las cuales son utilizadas como covariables para los árboles de decisión, *random forest* y *logit* ejecutados.

El cuadro 3 entrega información sobre la totalidad de variables finalmente disponible, a partir de las cuales se construyen los distintos modelos de predicción de inactividad.

Variable	Definición
$es\_inactividad_i$	Variable binaria, indica si la observación i alcanza el estado "Inactivo" (1) o no (0)
Rut_i	Rut al que pertenece la observación i
$Q\_primera\_actividad_i$	Quincena de la primera transacción hecha por la observación i
$Q\_ultima\_actividad_i$	Quincena de la última transacción hecha por la observación i
fec_min_i	Fecha primera compra de la observación i
fec_max_i	Fecha última compra de la observación i
$diff\_fechas_i$	Diferencia en días de fec_min_i y fec_max_i
$zona\_moda_i$	Zona del país en que la observación i mayoritariamente hizo sus compras. 1: Norte, 2: Sur, 4: Centro, 6: RM/R. O'higgins/R. Maule
$monto\_total_{ik}$	Monto total gastado por la observación i entre las quincenas $(Q\_ultima\_actividad_i - k + 1)$ y $Q\_ultima\_actividad_i$. $k \in \{1, 2, \dots, 6\}$
$frecuencia\_total_{ik}$	Frecuencia total de compra de la observación i entre las quincenas $(Q\_ultima\_actividad_i - k + 1)$ y $Q\_ultima\_actividad_i$. $k \in \{1, 2, \dots, 6\}$
$consistencia\_total_{ik}$	Consistencia total de compra de la observación i entre las quincenas $(Q\_ultima\_actividad_i - k + 1)$ y $Q\_ultima\_actividad_i$. $k \in \{1, 2, \dots, 6\}$
$ultima\_recencia_i$	Recencia de i en $Q\_ultima\_actividad_i$
$M\_quincenal\_promedio\_i$	Monto total quincenal promedio de i entre $Q\_primera\_actividad_i$ y $Q\_ultima\_actividad_i$
$F\_quincenal\_promedio\_i$	Frecuencia quincenal promedio de i entre $Q\_primera\_actividad_i$ y $Q\_ultima\_actividad_i$
$R\_quincenal\_promedio\_i$	Recencia quincenal promedio de i entre $Q\_primera\_actividad_i$ y $Q\_ultima\_actividad_i$
$C\_quincenal\_promedio\_i$	Consistencia quincenal promedio de i entre $Q\_primera\_actividad_i$ y $Q\_ultima\_actividad_i$
$monto\_desvest\_i$	Desviación estándar del monto total quincenal promedio de i entre $Q\_primera\_actividad_i$ y $Q\_ultima\_actividad_i$. Si $fec_{min_i} = fec\_max_i$, la variable toma el valor 0
$frecuencia\_desvest\_i$	Desviación estándar de la frecuencia quincenal promedio de i entre $Q\_primera\_actividad_i$ y $Q\_ultima\_actividad_i$. Si $fec_{min_i} = fec\_max_i$, la variable toma el valor 0
$recencia\_desvest\_i$	Desviación estándar de la recencia quincenal promedio de i entre $Q\_primera\_actividad_i$ y $Q\_ultima\_actividad_i$. Si $fec_{min_i} = fec\_max_i$, la variable toma el valor 0
$consistencia\_desvest\_i$	Desviación estándar de la consistencia quincenal promedio de i entre $Q\_primera\_actividad_i$ y $Q\_ultima\_actividad_i$. Si $fec_{min_i} = fec\_max_i$, la variable toma el valor 0

Cuadro 3: Universo de variables calculadas para la ejecución de modelos predictivos

Fuente: Elaboración propia

Cabe destacar que la creación de las variables $monto\_total_{ik}$, $frecuencia\_total_{ik}$ y $consistencia\_total_{ik}$ (con $k \in \{1, 2, \dots, 6\}$), se debe a que las observaciones de la muestra total pueden estar comprendidas entre distintos períodos de la ventana temporal considerada, luego, buscando poder utilizar todas las observaciones para entrenar y testear, y que ellas tengan el mismo peso en cada modelo ejecutado, se consideran variables transaccionales totales (monto, frecuencia y consistencia) alcanzadas en sólo

las últimas k quincenas. Así, a modo de visualización, se presenta el cuadro 4, donde se simula el cálculo de algunas variables, para tres observaciones de la muestra total, fijando $k = 2$ quincenas.

k	Muestra total	Últimos k estados activos	$es\_inactividad_i$	$monto\_total_{ik}$	$frecuencia\_total_{ik}$	$consistencia\_total_{ik}$	...
2	O-P-O-P-P-B-IN	P-B	1	$90.000	10	10	...
	P-P-P-B-P-O-P	O-P	0	$130.000	23	19	...
	B-B-B-B-B-P-P-P	P-P	0	$110.000	16	13	...
	...	...	...	...	...	...	...

Cuadro 4: Simulación de cálculo de variables según parámetro k definido

Fuente: Elaboración propia

Es pertinente mencionar que la compañía dispone de una base de datos sociodemográficos de los taxistas miembros del programa de fidelización, sin embargo, la fuente no es confiable, puesto que posee una gran cantidad de campos nulos, y otros con errores en la data, luego, si bien está la intención de incorporar ese tipo de variables en los modelos, es descartada para el trabajo actual.

Con el set de variables calculado para cada una de las observaciones, se procede a modelar la inactividad, a partir de árboles de decisión, *random forest*, y *logit*, separando sets de entrenamiento (80%) y testeo (20%) de manera aleatoria sobre la muestra total, tal como se realiza con cadenas de markov; seleccionando distintas combinaciones de variables independientes a partir de la variación del parámetro k (entre todos sus posibles valores), analizando luego las métricas ROC AUC obtenidas, y definiendo finalmente el k específico a utilizar en cada tipo de modelo.

Por último, cabe destacar que, para modelar inactividad a partir de regresiones logísticas, éstas deben cumplir los supuestos de cualquier regresión lineal, luego, no es posible utilizar los mismos sets de variables que en árboles de decisión y random forest, puesto que violan el principio de multicolinealidad[21] (según matrices de correlación mostradas en el anexo VI). Así, para estos modelos en específico, se hacen nuevas combinaciones de covariables que satisfagan los supuestos requeridos.

[21] Criterio aceptable: Coeficiente de correlación de Pearson dentro del rango [-0,4, 0.4]

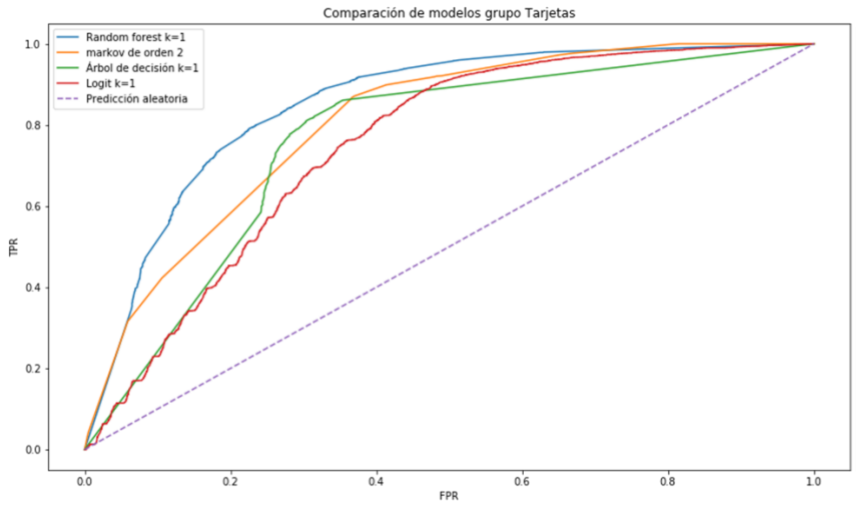
8.5.2.2 Resultados y selección de modelos

Las curvas ROC (y cada valor AUC asociado), para distintos k ($k \in \{1, 2, \dots, 6\}$) de los modelos ejecutados (árboles de decisión, *random forest* y *logit*), se muestran en anexos[22], de donde se extrae que $k = 1$ entrega las mejores métricas[23] en los 3 tipos de modelos propuestos, para todos los grupos estudiados.

Luego, a modo de comparación de los mejores modelos de cada tipo, se presentan los gráficos 26, 27 y 28, que ilustran las curvas ROC de cada modelo ejecutado, además del área bajo la curva (AUC) asociada. Aquí también, se visualizan los resultados de los mejores modelos markovianos obtenidos, todo con el objetivo de definir el modelo a utilizar finalmente en cada grupo, y así concluir con respecto a su aplicación a la realidad.

Grupo Tarjetas

A partir del análisis del gráfico 26, se puede ver que el modelo de mejor rendimiento para el grupo Tarjetas corresponde a *random forest*, logrando alcanzar 85% en la métrica medida, estando por sobre cadenas de markov, árboles de decisión y *logit*.



Grupo Tarjetas: Comparación de modelos		
Variables	**Modelo**	**ROC AUC**
RFMC	Markov de orden 2	81%
$monto\_total_{i1}$ $ultima\_recencia_i$ $frecuencia\_total_{i1}$ $consistencia\_total_{i1}$ $zona\_moda_i$	Árbol de decisión	76%
	Random forest	85%
$monto\_total_{i1}$ $ultima\_recencia_i$ $zona\_moda_i$	Logit	75%

Gráfico 26: Resultados mejores modelos de inactividad para grupo Tarjetas

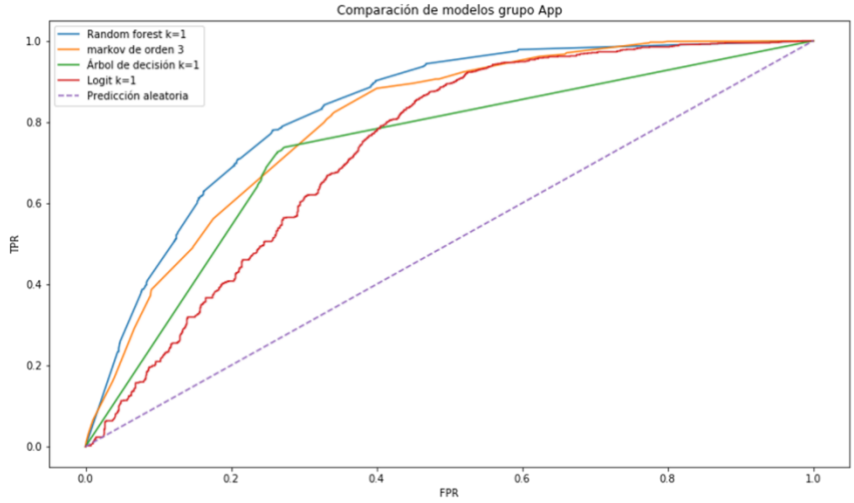
Fuente: Elaboración propia

[22] En específico, los anexos VII, VIII y IX

[23] Basado en la maximización de ROC AUC, igual como se hace en la comparación de cadenas de markov

Grupo App

Por otra parte, para el grupo App, mirando el gráfico 27, se ve ocurre algo similar, donde nuevamente, *random forest* se presenta como el modelo con la mayor métrica de interés (83% de ROC AUC), siendo árbol de decisión y *logit* los peores modelos, con una métrica de 73%.



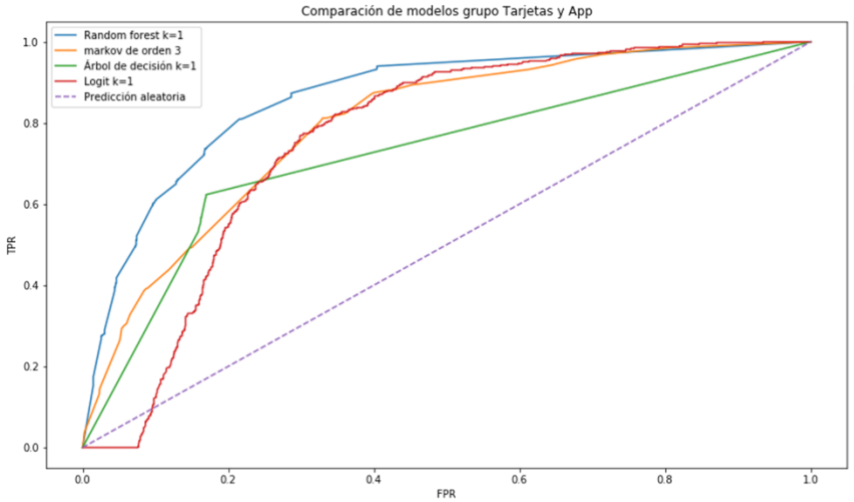
Grupo App: Comparación de modelos		
Variables	**Modelo**	**ROC AUC**
RFMC	Markov de orden 3	80%
$monto\_total_{i1}$ $ultima\_recencia_i$ $frecuencia\_total_{i1}$ $consistencia\_total_{i1}$ $zona\_moda_i$	Árbol de decisión	73%
	Random forest	83%
$monto\_total_{i1}$ $ultima\_recencia_i$ $zona\_moda_i$	Logit	73%

Gráfico 27: Resultados mejores modelos de inactividad para grupo App

Fuente: Elaboración propia

Grupo Tarjetas y App

Finalmente, y análogo a lo hecho para los demás grupos, se comparan los mejores modelos del grupo Tarjetas y App a partir del gráfico 28, donde se evidencia que la tónica continúa, ratificándose *random forest* como el modelo con ROC AUC más alto, alcanzando un valor de 87%.



Grupo Tarjetas y App: Comparación de modelos		
Variables	**Modelo**	**ROC AUC**
RFMC	Markov de orden 3	80%
$monto\_total_{i1}$ $ultima\_recencia_i$ $frecuencia\_total_{i1}$ $consistencia\_total_{i1}$ $zona\_moda_i$	Árbol de decisión	72%
	Random forest	87%
$monto\_total_{i1}$ $ultima\_recencia_i$ $zona\_moda_i$	Logit	76%

Gráfico 28: Resultados mejores modelos de inactividad para grupo Tarjetas y App

Fuente: Elaboración propia

Para concluir, y a modo de síntesis, se presenta el cuadro 5, donde es posible ver que los resultados coinciden: *random forest* corresponde al modelo que mejor desempeño tiene en todas las comparaciones, utilizando sólo variables extraídas de transacciones. Además, las cadenas de markov de orden superior se sitúan siempre como el segundo mejor modelo, tomando como input los segmentos a los que pertenecen los taxistas quincenalmente.

Comparación de modelos								
Grupo	Modelo	ROC AUC	Grupo	Modelo	ROC AUC	Grupo	Modelo	ROC AUC
Tarjetas	Markov de orden 2	81%	App	Markov de orden 3	80%	Tarjetas y App	Markov de orden 3	80%
	Árbol de decisión	76%		Árbol de decisión	73%		Árbol de decisión	72%
	Random forest	85%		Random forest	83%		Random forest	87%
	Logit	75%		Logit	73%		Logit	76%

Cuadro 5: Resumen de resultados de modelos ejecutados en cada grupo

Fuente: Elaboración propia

8.5.2.3 Importancia de variables

A partir de los resultados exhibidos, y elegido *random forest* como el mejor modelo en todos los casos, también es posible hacer un análisis de la importancia que toma cada variable considerada a la hora de modelar la inactividad, representado éste por el cuadro 6, que, basado en el criterio de impureza de *gini* (aquel con el que se obtienen los resultados de los modelos ejecutados), entrega la importancia relativa de cada variable en el modelo definido en cada grupo.

Importancia de Gini Random forest			
Variable	Grupo		
	Tarjetas	App	Tarjetas y App
$monto\_total_{i1}$	44%	45%	42%
$frecuencia\_total_{i1}$	22%	17%	21%
$ultima\_recencia_i$	20%	21%	18%
$consistencia\_total_{i1}$	10%	12%	13%
$zona\_moda_i$	4%	5%	6%

Cuadro 6: Importancia de variables consideradas en mejores modelos de cada grupo

Fuente: Elaboración propia

Del cuadro se puede ver que para los distintos grupos, el monto total gastado por los clientes en la última quincena transaccionada, es siempre la variable de mayor importancia a la hora de predecir inactividad, entregando un elevado aporte de información al modelo (sobre 41%), lo que luego es complementado por las demás variables consideradas, siendo la zona en que se realizan las transacciones aquella que menos incide.

Lo anterior va acorde a lo que se podría pensar (previo a la ejecución de cualquier modelo), puesto que, si un cliente va a disminuir su actividad en el programa, es de esperar que no gaste mucho dinero en sus últimas compras (debido a que ya no existe el interés por llegar a una meta de litros cargados, luego, no hay incentivo a cargar montos altos). Por otra parte, el rubro de cada taxista es igual a lo largo y ancho del país, por lo que, el conocer sólo la zona del cliente, no debería entregar gran cantidad de información relacionada a una posible baja de actividad.

Por último, del cuadro 6 también es destacable que en el grupo App, la importancia de la frecuencia es levemente menor a la de la recencia, mientras que en los otros dos grupos eso cambia, y se tiene que la recencia observada toma menor importancia que la frecuencia, concluyendo que en el grupo App, al modelar inactividad, es más valioso tener información sobre la última carga de cada cliente, que sobre la cantidad total de compras que él realizó en la última quincena de actividad.

8.5.2.4 Métricas de desempeño

Finalmente, a partir del testeo de cada modelo elegido, también es posible obtener la matriz de confusión asociada a los diferentes grupos estudiados, entregando así, detalles sobre la cantidad de aciertos y errores en la predicción. Estas matrices quedan representadas por la figura 11.

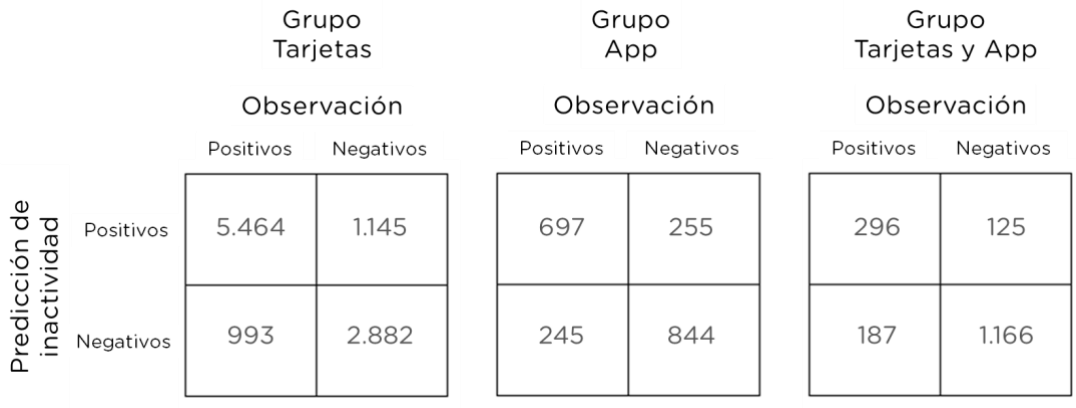


Figura 11: Matrices de confusión para modelos elegidos en los grupos Tarjetas, App y Tarjetas y App

Fuente: Elaboración propia

Tomando como referencia las matrices de confusión, es posible calcular métricas de desempeño adicionales a ROC AUC (aquellas definidas en la sección 4.1.6), que permiten la evaluación de los modelos en su aplicación a la realidad.

A partir de lo anterior, se elabora el cuadro 7, que ilustra un resumen global de resultados.

Modelo	Variables utilizadas	Grupo	Precision	Recall	F1 Score	NPV	TNR	ROC AUC	Accuracy
Random forest	$monto\_total_{i1}$ $ultima\_recencia_i$ $frecuencia\_total_{i1}$ $consistencia\_total_{i1}$ $zona\_moda_i$	Tarjetas	83%	85%	84%	74%	72%	85%	80%
		App	73%	74%	74%	78%	77%	83%	76%
		Tarjetas y App	70%	61%	65%	86%	90%	87%	82%

Cuadro 7: Métricas de desempeño según modelo seleccionado en cada grupo

Fuente: Elaboración propia

Componiendo la información de la figura 11 y el cuadro 7, se evidencia y destaca que, por ejemplo, para el grupo Tarjetas, hay una predicción de 6.609 observaciones como positivas a la caída en la inactividad, acertando en 5.464 de ellas, lo que se traduce en un 83% de *precision*. Por otra parte, de un total de 6.457 casos reales que pasaron al estado “Inactivo”, el modelo falla en 993 de ellos, lo que se traduce en un *recall* de 85%.

Análogamente para el grupo App, se puede ver que, de una totalidad de 952 casos predichos como positivos a la caída en inactividad, se logra identificar 697 de ellos correctamente, equivalente a un 73% en *precision* y 74% en *recall*, puesto que son 245 los falsos negativos.

Finalmente, centrándose en el grupo Tarjetas y App, es posible hacer el mismo análisis, teniendo esta vez 296 verdaderos positivos, 125 falsos positivos y 187 falsos negativos, logrando métricas de 70% en *precision*, y 61% en *recall*, tal como muestra el cuadro 7.

8.5.2.5 Análisis de sensibilidad

Continuando con los estudios, para cada uno de los 3 macro grupos de clientes, se realiza un análisis de sensibilidad, consistente en ver cómo cambian ciertos resultados al variar el criterio de inactividad definido, en específico, se analiza la cantidad de observaciones de actividad e inactividad identificadas, así como también la métrica ROC AUC asociada a los modelos *random forest* ejecutados sobre las nuevas muestras.

Cabe recordar que, bajo lo mencionado en la sección 8.2, la definición de inactividad es distinta según el grupo tratado, siendo catalogado un cliente como inactivo cuando pasa 4 quincenas seguidas o más sin comprar en el grupo App, y 5 quincenas seguidas o más en los grupos Tarjetas y Tarjetas y App.

Teniendo lo anterior en cuenta, se procede a variar esa definición de inactividad, ejecutándose modelos bajo 6 escenarios distintos, yendo desde considerar a clientes inactivos cuando pasan sólo 1 quincena sin comprar, hasta cambiando completamente el criterio, catalogando clientes como inactivos no antes de pasar 15 quincenas seguidas o más ausentes.

Los resultados para el grupo Tarjetas son mostrados a partir del cuadro 8, donde, tal como podría esperarse, se ve que al definir criterios de inactividad más flexibles (mayor cantidad de quincenas), el número de observaciones de inactividad disminuye, a la vez que la cantidad de casos de actividad aumenta. Así por ejemplo, para el criterio más estricto posible (1+ quincena), se aprecia que hay alrededor de 101.000 casos de inactividad detectados, muy por sobre los aproximadamente 32.000 del criterio original.

Grupo	Período	Criterio de inactividad (quincenas)	Observaciones de inactividad identificadas	Observaciones de actividad identificadas	ROC AUC	
Tarjetas	2018-2019	1+	101.529	16.341	86%	
		3+	45.003	18.931	83%	
		5+	32.136	20.284	85%	Criterio original
		7+	26.220	21.255	85%	
		10+	22.700	21.263	86%	
		15+	16.307	25.009	87%	

Cuadro 8: Resultados análisis de sensibilidad para el grupo Tarjetas

Fuente: Elaboración propia

Referente a la métrica de interés, se podría decir que existe una tendencia a aumentar conforme aumenta la flexibilidad del criterio, sin embargo, el criterio de 3+ quincenas presenta un menor valor ROC AUC que el criterio más estricto posible (1+ quincenas), contradiciendo la hipótesis.

Ahora bien, producto de que el modelamiento de inactividad es atemporal (no hay distinción entre quincenas), para descartar una posible aleatoriedad en los resultados, se hace el mismo análisis de sensibilidad, pero para períodos independientes, es decir, para 2018 y 2019 por sí solos, sin embargo, los resultados son similares (mostrados en el anexo X), nuevamente notando que el menor valor ROC AUC corresponde al criterio de 3+ quincenas.

Para el grupo App también se realiza el análisis, ilustrando los resultados en el cuadro 9. En él se ve que ocurren fenómenos parecidos a los presentados por el grupo Tarjetas, primero, porque se identifican cada vez menos casos de inactividad a medida que se incrementa la flexibilidad del criterio, pasando de alrededor de 10.000 observaciones (criterio más estricto) a aproximadamente 1.300 (criterio más flexible), y segundo, porque la métrica de interés se mantiene o incrementa (con respecto al criterio original) al flexibilizar el parámetro de inactividad, así como también al imponer el criterio más estricto posible, alcanzando un 87% en este último caso. Además, otra vez se ve que el criterio de 3+ quincenas tiene el peor rendimiento en ROC AUC.

Grupo	Período	Criterio de inactividad (quincenas)	Observaciones de inactividad identificadas	Observaciones de actividad identificadas	ROC AUC	
App	2019	1+	10.018	4.240	87%	
		3+	5.622	5.078	82%	
		4+	4.834	5.371	83%	Criterio original
		7+	3.374	6.270	85%	
		10+	2.388	7.005	88%	
		15+	1.303	7.950	87%	

Cuadro 9: Resultados análisis de sensibilidad para el grupo App

Fuente: Elaboración propia

Análogamente para el grupo Tarjetas y App, se presentan los resultados del análisis en el cuadro 10, donde es preciso notar que, una vez más, disminuyen de manera sucesiva los casos de inactividad (al dar más flexibilidad al criterio). Además, al imponer el criterio más rígido posible, se incrementa el ROC AUC, llegando a alcanzar un 89%, 2 p.p. por sobre lo alcanzado con que el criterio original. Por último, fijando el criterio en 3+ quincenas, el modelo alcanza peor rendimiento (85%) que con el criterio de 1+ quincenas (89%).

Grupo	Período	Criterio de inactividad (quincenas)	Observaciones de inactividad identificadas	Observaciones de actividad identificadas	ROC AUC
Tarjetas y App	2019	1+	9.025	5.533	89%
		3+	3.710	6.096	85%
		5+	2.461	6.405	87%
		7+	1.766	6.686	87%
		10+	1.145	6.991	87%
		15+	541	7.381	83%

Criterio original

Cuadro 10: Resultados análisis de sensibilidad para el grupo Tarjetas y App

Fuente: Elaboración propia

Para este grupo, cabe destacar que, con el criterio más flexible existente, la métrica de interés empeora (llega a 83%, 4 p.p. menor que con el criterio original), difiriendo de lo ocurrido en los demás grupos (donde siempre mejora). Esto puede ser explicado por la escasa cantidad de datos con los que se trabaja, específicamente, por la poca información de inactividad disponible, dado que el número de casos inactivos no supera los 550, siendo incluso menor al 10% de la data, afectando el rendimiento del modelo ejecutado.

Finalmente, a modo de síntesis de resultados, se destacan los siguientes hallazgos:

1) Tal como se podría esperar, a medida que aumenta el número de quincenas consideradas en el criterio de inactividad, disminuyen las observaciones de inactividad identificadas, a la vez que van aumentando los casos de actividad. Esto se explica porque a mayor cantidad de quincenas, hay mayor flexibilidad en el criterio.

2) Se presentan evidencias de que el rendimiento de modelos puede verse afectado por la cantidad de datos y distribución de clases que se tenga, esto a partir del grupo Tarjetas y App al imponer el criterio de inactividad de 15 quincenas, donde, de una escasa muestra de 7.922 observaciones, la clase de inactividad sólo representa el 7% de ella, alcanzando el mínimo ROC AUC entre todos los criterios probados (para todos los demás grupos, independiente del criterio elegido, la clase de inactividad está siempre por sobre el 10% de la totalidad de casos).

3) En los distintos grupos, con respecto al criterio original de inactividad definido, se tiene la hipótesis de que el ROC AUC se mantiene o mejora al llevar el parámetro a los extremos (ya sea a la máxima rigidez, o bien, a la máxima flexibilidad), lo que puede ser interpretado como que, bajo estas nuevas definiciones, es más fácil establecer criterios de corte para los modelos, diferenciando de mejor manera cada una de las clases posibles (propensos o no propensos a la inactividad), traducido luego en menos falsos positivos y más verdaderos positivos predichos.

   Ahora bien, a pesar de que hay tenues evidencias de lo anterior, ello no permite llegar a una conclusión concreta, puesto que en todos los grupos, se presenta un descenso en el ROC AUC al pasar del criterio original al criterio de 3+ quincenas, no pudiendo identificar la razón de esto con el análisis realizado.

Luego, como trabajo futuro, se propone realizar dos estudios. Primero, uno que explique el fenómeno percibido, donde se entreguen las razones que afectan a los modelos a la hora de definir el criterio de 3+ quincenas, pudiendo probarse distintas hipótesis, por ejemplo, que existe una cantidad considerable de taxistas que alternan su trabajo cada 3 quincenas, lo que hace “confundir” a los modelos a la hora de predecir inactividad. Y segundo, uno que cuantifique pérdidas y ganancias (tanto monetarias como estratégicas) asociadas a la elección de cada criterio de inactividad.

Lo anterior permitiría justificar si existiese una tendencia en el ROC AUC conforme se varía el parámetro de inactividad, para luego entregar recomendaciones sobre si se debiese o no cambiar la definición de inactividad tomada.

4) A pesar de que los criterios de inactividad originalmente considerados en cada grupo vienen de definiciones comerciales (acompañados de análisis cuantitativos), al cambiar el parámetro en cuestión dentro de una vecindad de 1 a 15 quincenas, la métrica de interés varía como máximo un 4% con respecto al criterio original, luego, se cree que independiente de la definición de inactividad que se tome, se llegaría a las mismas conclusiones y resultados en lo que a modelos concierne.

9. Conclusiones

Por medio del trabajo realizado, se declara el cumplimiento de cada uno de los objetivos planteados, tanto generales como específicos.

Primero, porque se da cuenta de la construcción de una base de datos analítica que permite la ejecución de modelos, esto a partir de definiciones comerciales y filtros aplicados (trata de datos faltantes, *outliers*, etc.) a una base de datos transaccional referente al programa de fidelización de taxistas estudiado.

Una vez definida la base analítica a utilizar, quincenalmente, distintas variables de compra son calculadas para los clientes (recencia, frecuencia, monto total comprado, entre otras). Con ellas, bajo el método RFMC, se clasifica a cada taxista en uno de cinco segmentos posibles (Oro, Plata, Bronce, Infrecuente o Atípico; según definiciones de la sección 4.2.1), entregando una caracterización histórica de su comportamiento con la compañía, dando respuesta al segundo objetivo específico planteado.

Como tercer paso, basado en análisis cuantitativos y decisiones comerciales, se define el criterio de inactividad a considerar en los modelos (sección 8.2). Este criterio tiene la particularidad de que depende de períodos consecutivos de no compra, siendo cada período, una quincena, y varía según el tipo de cliente analizado, definiéndose para el cliente App como 4 quincenas o más, y para los clientes Tarjetas y Tarjetas y App como 5 quincenas o más sin comprar. Así, se declara el cumplimiento de un nuevo objetivo específico.

Definido lo necesario para la ejecución de modelos, se modela la inactividad de clientes en cada uno de los macro grupos posibles, ejecutando 4 tipos de modelos: cadenas de markov, árboles de decisión, regresiones logísticas y *random forest*, siendo este último el con mejores resultados a la hora de la predicción, alcanzando 85%, 83% y 87% en la métrica de interés (ROC AUC) para los grupos Tarjetas, App y Tarjetas y App, respectivamente.

Cada modelo elegido, basado en la información tomada como input, entrega la probabilidad de que el cliente analizado caiga en la inactividad en el período siguiente. Si esa probabilidad está dentro del rango $[50\%, 100\%]$, el cliente se cataloga como “en peligro de inactividad”. Así luego, se obtiene una cartera de taxistas propensos a suspender su actividad del programa de fidelización, declarando con ello, el cumplimiento de los objetivos específicos 4, 5 y 6 planteados.

Por último, cumplido cada objetivo específico, se llega a la consecución del objetivo general del proyecto, concluyendo que es posible hacer el modelamiento de inactividad de taxistas del programa de fidelización, logrando métricas de desempeño aceptables por la contraparte en los tres grupos estudiados (Tarjetas, App, y Tarjetas y App), destacando un ROC AUC siempre superior a 80%, y métricas *Precision* y *Recall* relativamente altas, sobre todo en el modelo del grupo Tarjetas (el más importante, dada su alta concentración de clientes), alcanzando valores de 83% y 85% en ellas, respectivamente.

Los modelos seleccionados, logran predecir inactividad utilizando variables transaccionales (recencia, frecuencia, monto, consistencia y zona de compra) para ello, y tienen la particularidad de que sólo se analiza el último período (quincena) de actividad

que tuvo cada cliente, lo que entrega una alta capacidad de respuesta en el corto plazo a la compañía, sin la necesidad de tener que esperar observar otras variables acumulativas.

Lo anterior también lleva a inferir que los taxistas analizados presentan un comportamiento de compra más bien definido en el tiempo inmediatamente anterior a su inactividad, donde generalmente, compran montos bajos de combustible y realizan pocas transacciones.

Finalmente, con la utilización de los modelos *random forest* en cada grupo de clientes, basado en sus matrices de confusión (figura 11) y métricas de desempeño (cuadro 7), es posible atribuir 4 ganancias inmediatas para la compañía, detalladas en lo que sigue:

1) **Posibilidad de retención de clientes:** En el grupo Tarjetas (mayoritario) hay un 87% de casos de inactividad que el modelo logra predecir. A ellos se podría haber aplicado marketing directo para evitar el alejamiento del programa, lo que a su vez, evita costos de adquisición de clientes nuevos que reemplacen a los inactivados (es más costoso adquirir que retener).

2) **Ahorro en campañas:** Actualmente, al hacer campañas de marketing, la compañía envía SMS/mails a todos los miembros de un grupo específico de clientes (Tarjetas, App, o Tarjetas y App). Utilizando los modelos de predicción, se podría lograr un ahorro por campaña de 37,0%, 53,4% y 76,3%, en los grupos Tarjetas, App, y Tarjetas y App, respectivamente. Esto porque los modelos sólo harían enviar mensajes a aquellos propensos a la inactividad (los de predicción positiva), dejando fuera a los sin riesgo (predicción negativa). Ahora bien, todo lo anterior en el supuesto de que clientes propensos sean “estimulables” con campañas, lo que debe ser ratificado por medio de experimentación.

3) **Comunicación eficaz:** Siguiendo con el ejemplo del grupo Tarjetas, se tiene que el modelo entrega una baja tasa de falsos positivos (13%) sobre la totalidad de inactividades predichas, esto da signos de una alta calidad de clasificación, es decir, hay homogeneidad en el grupo en peligro de inactividad, facilitando la elaboración del mensaje publicitario que se enviará, así como también el alcance de la eficacia esperada.

4) **Reforzamiento de percepción de marca:** Al enviar comunicación directa a clientes que realmente no se iban a inactivar (falsos positivos), se refuerza la percepción de marca en ellos. De esto, la compañía puede salir beneficiada, ya sea por un aumento de compras de esos consumidores, o bien, por la recomendación “boca a boca” que ellos puedan comenzar en sus conocidos.

10. Recomendaciones comerciales

A partir de las conclusiones, cabe recordar que actualmente en la compañía, para promocionar el programa de fidelización, cada vez que se realizan campañas directas, éstas se efectúan sobre la totalidad de clientes pertenecientes a algún macro grupo específico (Tarjetas, App, o Tarjetas y App), enviando mensajes genéricos, que su única finalidad es funcionar como recordatorio del beneficio de descuento; luego, no existe distinción entre campañas de adquisición, retención o recuperación de taxistas. Además, las campañas no siguen una periodicidad establecida, pudiendo pasar varios meses entre distintas promociones.

Así, a modo de recomendación, se propone modificar la forma de hacer marketing relacionado al programa, comenzando por un reordenamiento de campañas de acuerdo a objetivos comerciales específicos (adquisición, retención o recuperación de clientes), utilizando en las estrategias de retención, el modelamiento de inactividad a partir de cada uno de los modelos *random forest* estudiados.

La utilización de los modelos supone además una serie de cambios en las campañas a realizar, pudiendo distinguirse 3 claves:

- **Recurrencia:** Se debe pasar de promociones no regulares en el tiempo, a campañas lanzadas quincenalmente, donde en cada quincena, se identifique a clientes propensos y no propensos a la inactividad.

- **Segmentación de clientes:** Cambiar el público objetivo de cada campaña, pasando de hacer marketing directo a todo cliente posible, a incluir en las campañas solamente a taxistas catalogados como “en peligro de inactividad”. Esto luego podría ser complementado con trabajos futuros, donde se identifique la “estimulabilidad” de cada taxista, a fin de evitar sobre invertir en clientes que realmente no requerían marketing para lograr el objetivo plateado con la campaña.

- **Mensaje transmitido:** Redefinir el mensaje elaborado en las campañas, pasando de uno genérico que funciona como recordatorio, a otro que busque generar una acción concreta en el segmento apuntado (evitar su alejamiento del programa).

Por otra parte, a partir del modelamiento, se obtienen pruebas de que el monto total cargado corresponde a la variable de mayor importancia a la hora de predecir inactividad, entregando un elevado aporte de información a los modelos (sobre 41% en todos los casos), luego, se plantea ratificar lo anterior por medio de experimentación, donde se realicen campañas piloto sobre muestras de clientes catalogados como “en peligro de inactividad”, para estimar el impacto de cada una de las variables independientes utilizadas en los modelos (recencia, frecuencia, monto total, consistencia, y zona de transacciones).

En caso de confirmarse que el monto es la variable más importante, se puede hacer uso de lo anterior ya sea en las mismas campañas de retención, o bien en campañas de adquisición o recuperación de clientes, donde se elaboren mensajes buscando incentivar fuertemente un alto consumo desde el primer periodo activo de cada cliente, reduciendo las probabilidades de una pronta inactivación.

11. Trabajo futuro

Finalmente, a modo de continuación del proyecto, se identifican variadas líneas de investigación posibles, que sirven como complemento al trabajo realizado.

1) Poner a prueba los modelos de inactividad en campañas, evaluando su performance y comparándola con los resultados que la estrategia actual (sin utilización de modelos) logra a la hora de evitar inactividades en clientes.

2) Realizar experimentación para identificar la "estimulabilidad" de cada taxista, a fin de dar con una cartera de clientes que tengan una alta probabilidad de responder positivamente al objetivo buscado con las campañas de retención. Esto puede ser complementado con los modelos de inactividad, para optimizar aún más las campañas, por ejemplo, mejorando la eficacia de cada mensaje publicitario enviado.

3) Dado que no se pudo incluir variables sociodemográficas en el estudio, se propone ejecutar nuevamente los modelos en el futuro, donde sí se incluyan estas variables, siempre y cuando se actualice la fuente con información personal fidedigna de todos los taxistas.

4) Experimentar con respecto a la importancia e impacto de variables a la hora de evitar inactividad, teniendo como primera hipótesis que el monto cargado corresponde a la más relevante. Esto serviría para todo tipo de campañas, tanto de retención como de adquisición y recuperación de clientes.

5) Estudiar consecuencias estratégicas y monetarias que ayuden a recomendar sobre si se debiese o no cambiar la definición de inactividad en taxistas (ya sea haciéndolo más corta o más larga).

6) Bajo la misma metodología utilizada en el modelamiento de inactividad, se plantea modelar la actividad sobresaliente de taxistas. esto quiere decir que, basado en el comportamiento histórico de compra de cada cliente, se intenta identificar a todos aquellos taxistas propensos a alcanzar la categoría "Oro" de la segmentación RFMC, para luego ejercer marketing directo sobre ellos, con la finalidad de ofrecer regalías e incentivar un mayor consumo, lo que se traduciría en menores posibilidades de pérdida de clientes destacados, y mayores ganancias para la compañía.

7) En cada macro grupo de taxistas, tomando como input la información quincenal entregada por los modelos de inactividad, se propone modelar el *Customer Lifetime Value* (CLV) de cada cliente en la compañía. Luego, se podría cuantificar el aporte futuro de cada taxista, así como también lo perdido en caso de que un cliente se distancia del programa.

8) Teniendo en cuenta que la compañía dispone de distintos programas de fidelización, se propone modelar la inactividad de clientes en cada uno de ellos, tomando como referencia los modelos del programa de taxistas.

Lo anterior ayudaría a la empresa descubrir nuevos hallazgos, que permitan apoyar la toma de decisiones en el negocio, así como también la creación de nuevas estrategias de marketing para el futuro.

12. Bibliografía

- Ching, Huang, Ng, & Siu. (2005). *Markov Chains Models, Algorithms and Applications.* Obtenido de [Ching & Ng 2005-12-05]
- Comisión Nacional de Energía. (Marzo de 2020). Obtenido de Sitio web Comisión Nacional de Energía: https://www.cne.cl/estadisticas/hidrocarburo/
- Compañía. (2016). *Inscripción en registro de valores Nº 0028, Prospecto legal emisión de bonos.*
- Compañía. (2018). *Memoria compañía.*
- Compañía. (2020). Obtenido de Sitio web compañía.
- Doğan, Ayçin, & Bulut. (2018). A customer segmentation by using RFM model and clustering methods: a case study in retail industry.
- Global software support. (Febrero de 2018). *Global software support: Random forest classifier*. Obtenido de https://www.globalsoftwaresupport.com/random-forest-classifier/
- Google Developers. (2020). *developers.google: Glosario sobre aprendizaje automático*. Obtenido de developers.google: https://developers.google.com/machine-learning/crash-course/glossary?hl=es
- Google Developers. (2020). *developers.google: ROC and AUC*. Obtenido de developers.google: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
- Harvard Business Review. (2014). The Value of Keeping the Right Customers.
- Investopedia. (5 de Julio de 2019). *Recency, Frequency, Monetary Value (RFM).* Obtenido de https://www.investopedia.com/terms/r/rfm-recency-frequency-monetary-value.asp#:~:text=Recency%2C%20frequency%2C%20monetary%20value%20is,customer%20has%20made%20a%20purchase
- Numerentur. (2020). *Numerentur*. Obtenido de Siti web Numerentur: http://numerentur.org/entropia-indice-de-ganancia/
- Optimove. (Mayo de 2020). *RFM segmentation.* Obtenido de https://www.optimove.com/resources/learning-center/rfm-segmentation
- Orellana Alvear, J. (Noviembre de 2018). *Bookdown: Ensambladores random forest*. Obtenido de https://bookdown.org/content/2031/ensambladores-random-forest-parte-i.html
- Quantinsti. (18 de Abril de 2019). Obtenido de https://blog.quantinsti.com/gini-index/
- Riquelme, F. (2017). *Memoria.*

- Search Data Management. (Noviembre de 2005). *RFM Analysis*. Obtenido de Search Data Management web site: https://searchdatamanagement.techtarget.com/definition/RFM-analysis
- Swaminathan, S. (15 de Marzo de 2018). *Towards data science: Logistic Regression — Detailed Overview*. Obtenido de https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
- Veloso, F. (5 de Septiembre de 2019). *feedingthemachine*. Obtenido de https://www.feedingthemachine.cl/arboles-de-decision-en-regresion-machine-learning/
- Weber, R. (Mayo de 2020). *U-cursos: Data Science for Social Network Analysis.* Obtenido de U-cursos: http://www.u-cursos.cl
- Yazlle, J. (2005). *Cadenas de markov.*
- Yiu, T. (12 de Junio de 2019). *Understanding random forest*. Obtenido de https://towardsdatascience.com/understanding-random-forest-58381e0602d2

13. Anexos

13.1 Anexo I: Ecuaciones de Chapman - Kolmogorov

Sea $p_{ij}^{(k)}$ la probabilidad de pasar de i a j en k pasos, se puede hacer el siguiente razonamiento:

Si al cabo de m<k pasos el sistema se encuentra en el estado e, la probabilidad de alcanzar el estado j después de $k - e$ pasos será:

$$p_{ie}^{(m)} . p_{ej}^{(k-m)}$$

Luego, como el estado intermedio e puede ser cualquiera, se puede determinar una expresión para la probabilidad de transición de k pasos dada por:

$$p_{ij}^{(k)} = \sum_{e=1}^{n} p_{ie}^{(m)} . p_{ej}^{(k-m)}$$

Haciendo m=1 y m= k −1 se obtienen las ecuaciones de Chapman – Kolmogorov, que permiten obtener las expresiones de las propiedades de transición en el estado k a partir de las de k-1.

$$p_{ij}^{(k)} = \sum_{e=1}^{n} p_{ie} . p_{ej}^{(k-1)} , con\ m = 1$$

$$p_{ij}^{(k)} = \sum_{e=1}^{n} p_{ie}^{(k-1)} . p_{ej} , con\ m = k - 1$$

Esto da paso a la obtención de matrices de transición de k pasos, a partir de las potencias de la matriz de transición P de la cadena de markov de orden 1. Luego se tiene:

$$\mathrm{P}^{(2)} = \mathrm{P} \cdot \mathrm{P} = \mathrm{P}^2$$

$$\mathrm{P}^{(3)} = \mathrm{P}^{(2)} \cdot \mathrm{P} = \mathrm{P}^2 \cdot \mathrm{P} = \mathrm{P}^3$$

$$\mathrm{P}^{(k)} = \mathrm{P}^{(k-1)} \cdot \mathrm{P} = \mathrm{P}^{k-1} \cdot \mathrm{P} = \mathrm{P}^{k-1}$$

Es decir, las sucesivas potencias de la matriz $\mathbf{P}$ indican las probabilidades de transición en tantas transiciones como se indica en el índice de la potencia.

13.2 Anexo II: Composición de segmentos según puntajes R, F, M, C

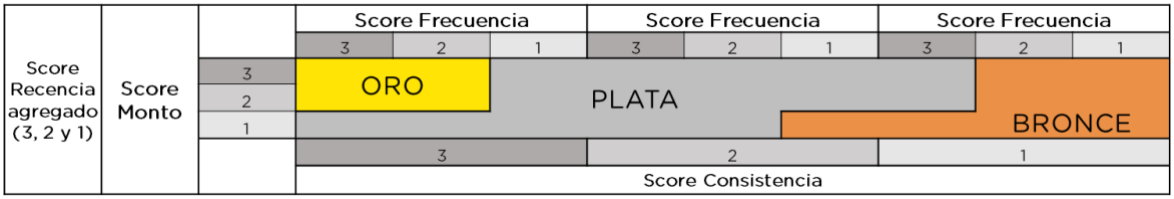


Figura 12: Composición de segmentos según puntajes R, F, M, C obtenidos

Fuente: Elaboración propia

13.3 Anexo III: Análisis continuidad de infrecuencias para los distintos grupos

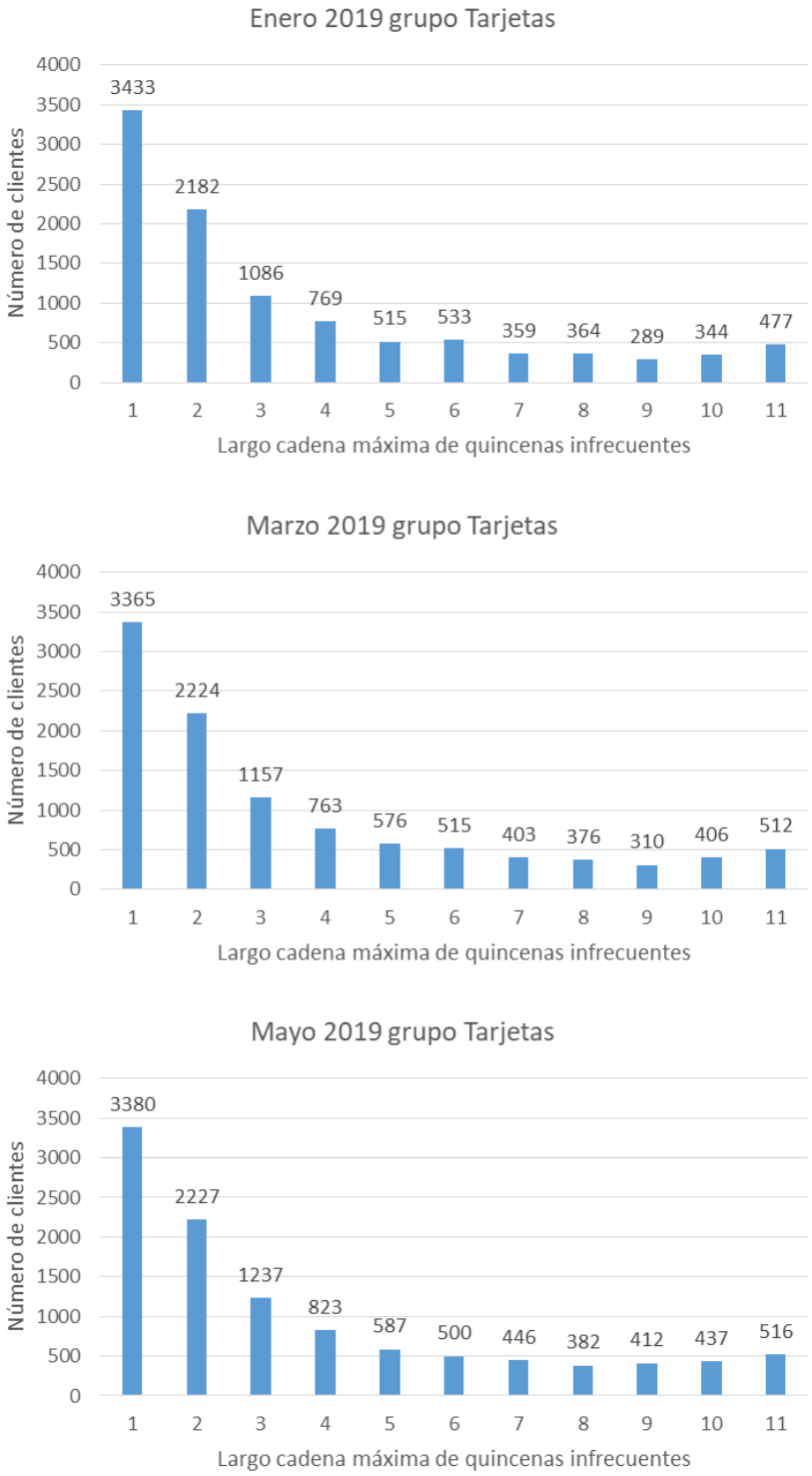


Gráfico 29: Resultados análisis continuidad de infrecuencias para grupo Tarjetas

Fuente: Elaboración propia

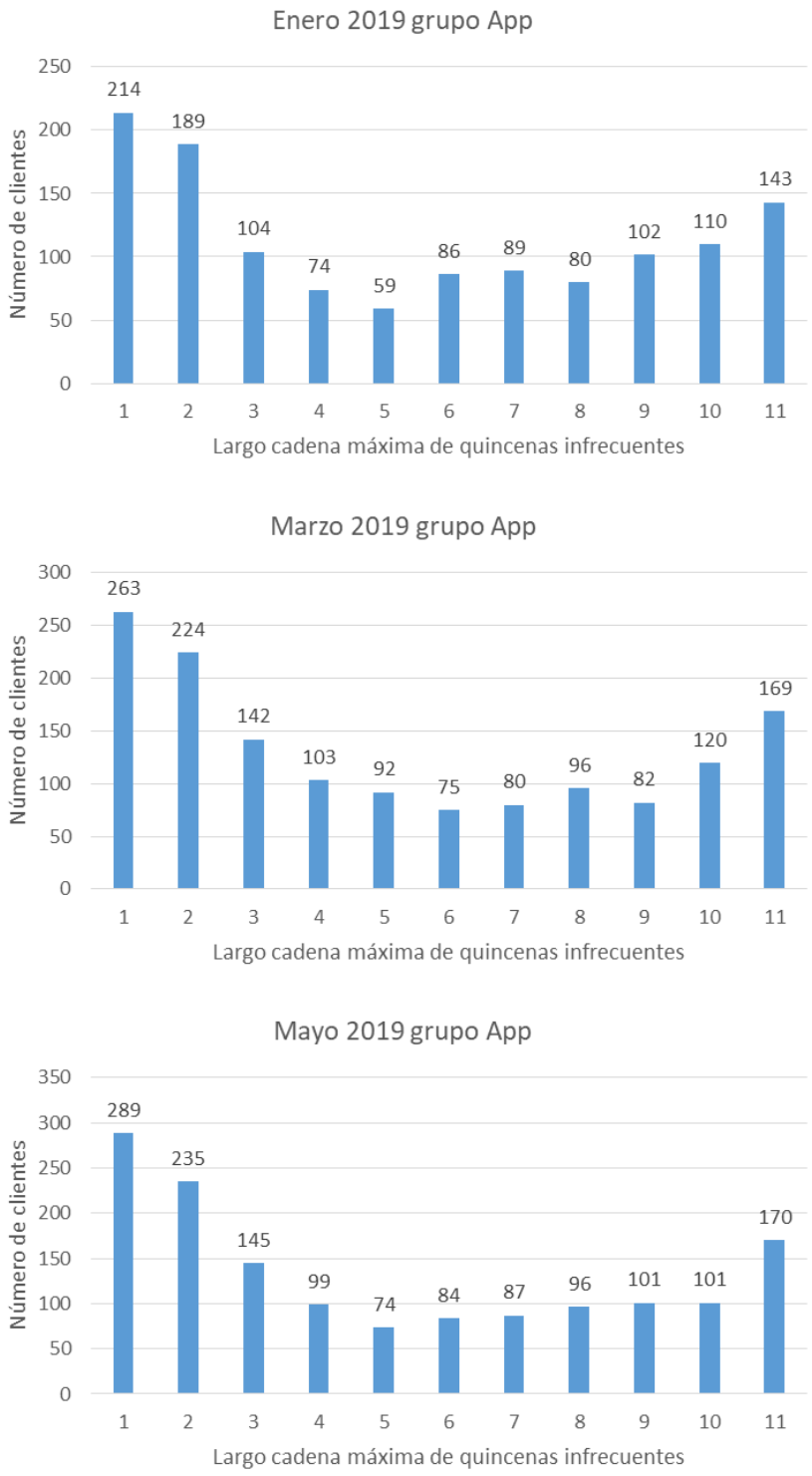


Gráfico 30: Resultados análisis continuidad de infrecuencias para grupo App

Fuente: Elaboración propia

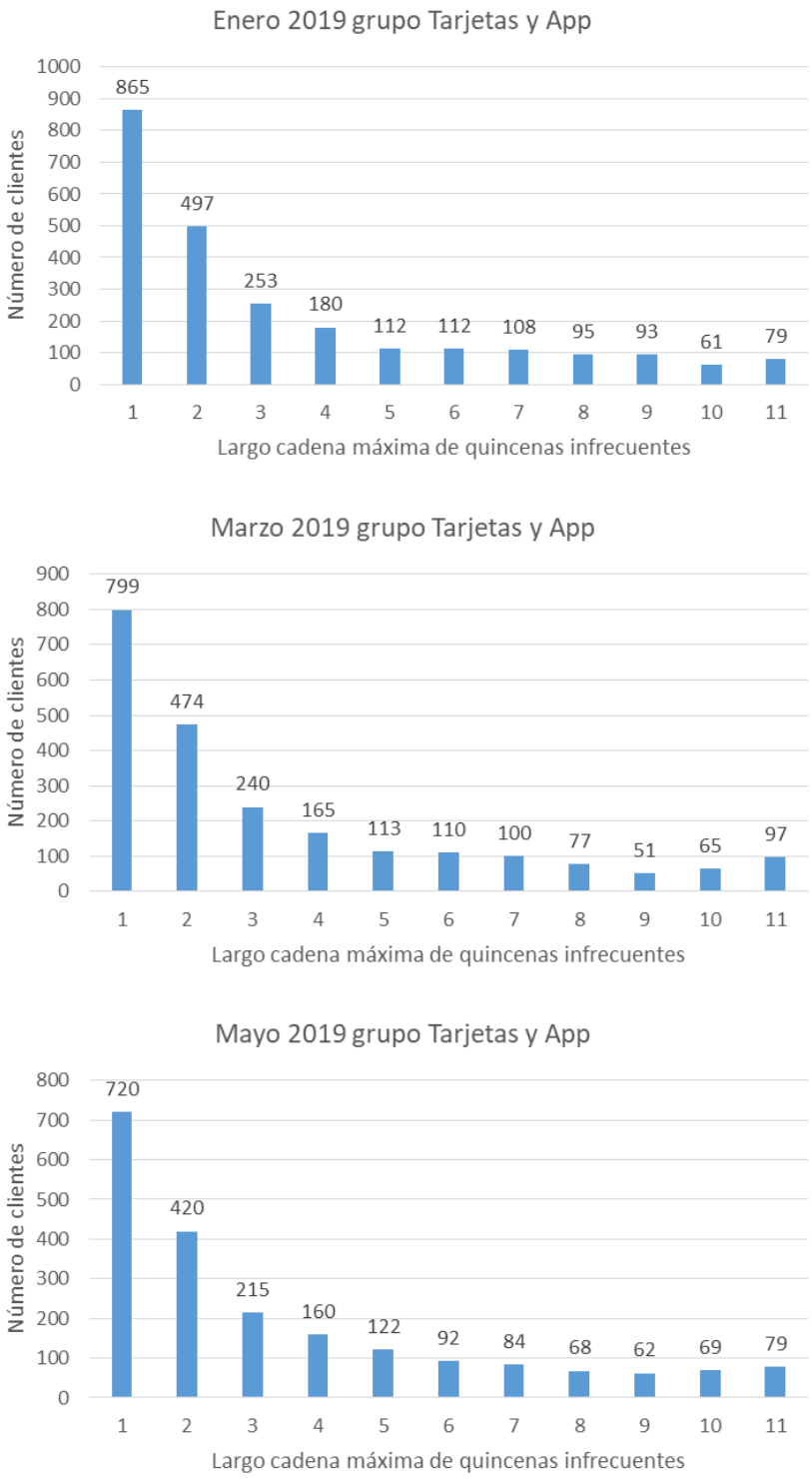


Gráfico 31: Resultados análisis continuidad de infrecuencias para grupo Tarjetas y App

Fuente: Elaboración propia

13.4 Anexo IV: Indicadores quincenales de variables R, F, M, C para los distintos grupos

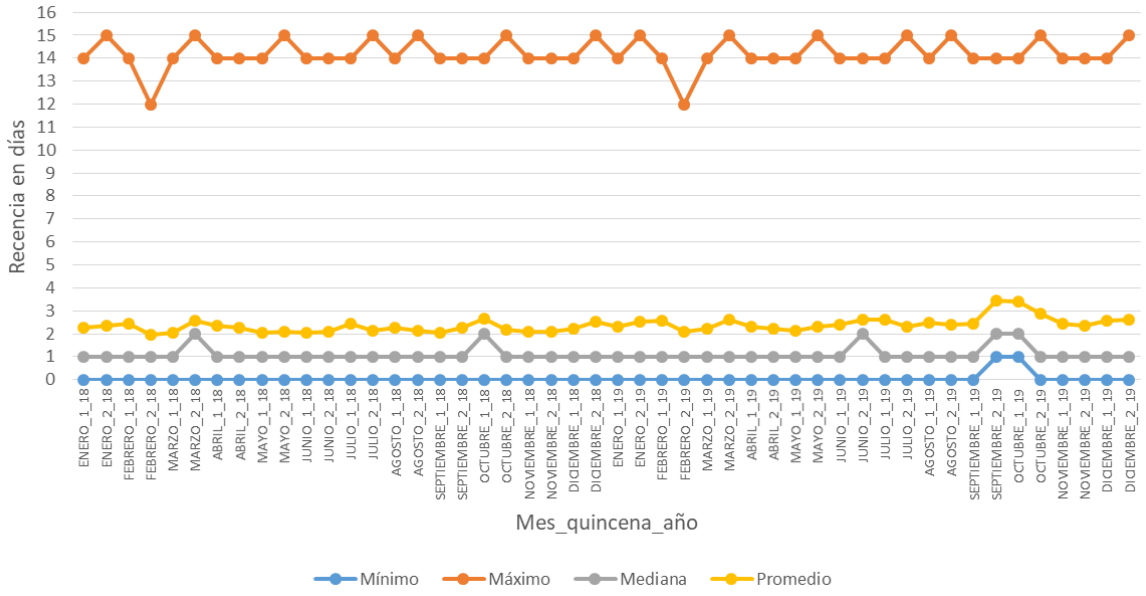


Gráfico 32: Indicadores quincenales de recencia para grupo Tarjetas

Fuente: Elaboración propia

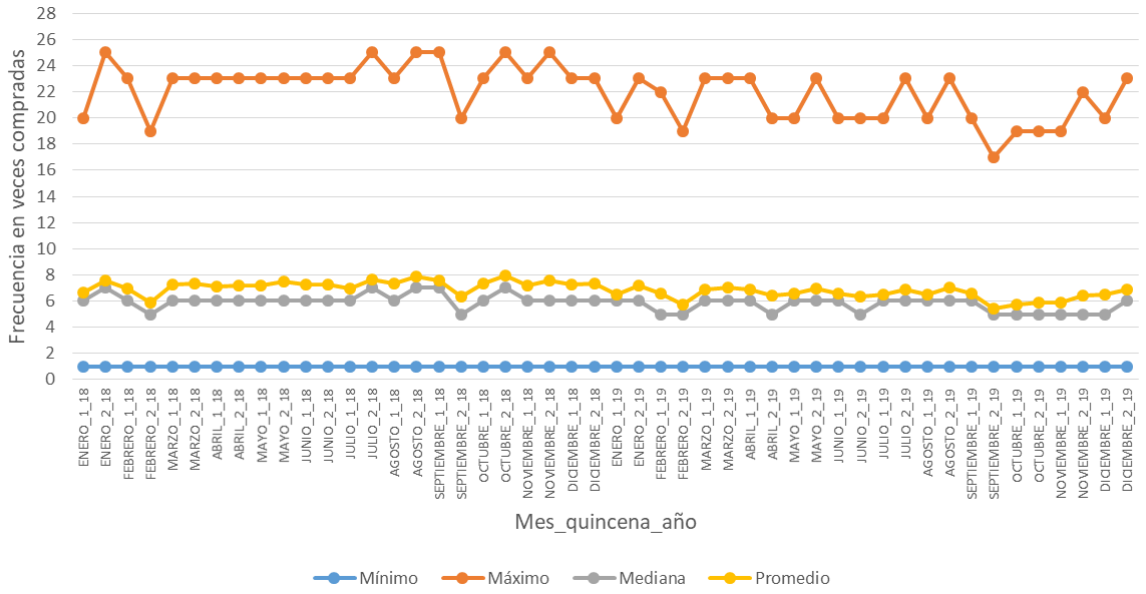


Gráfico 33: Indicadores quincenales de frecuencia para grupo Tarjetas

Fuente: Elaboración propia

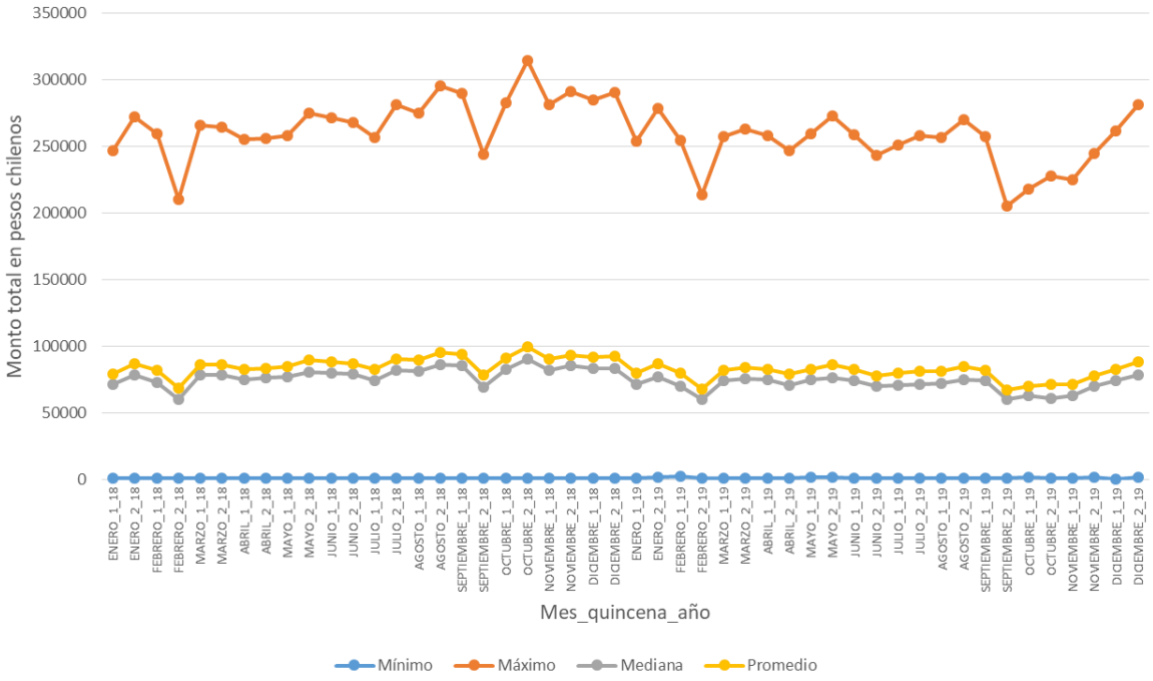


Gráfico 34: Indicadores quincenales de monto total para grupo Tarjetas

Fuente: Elaboración propia

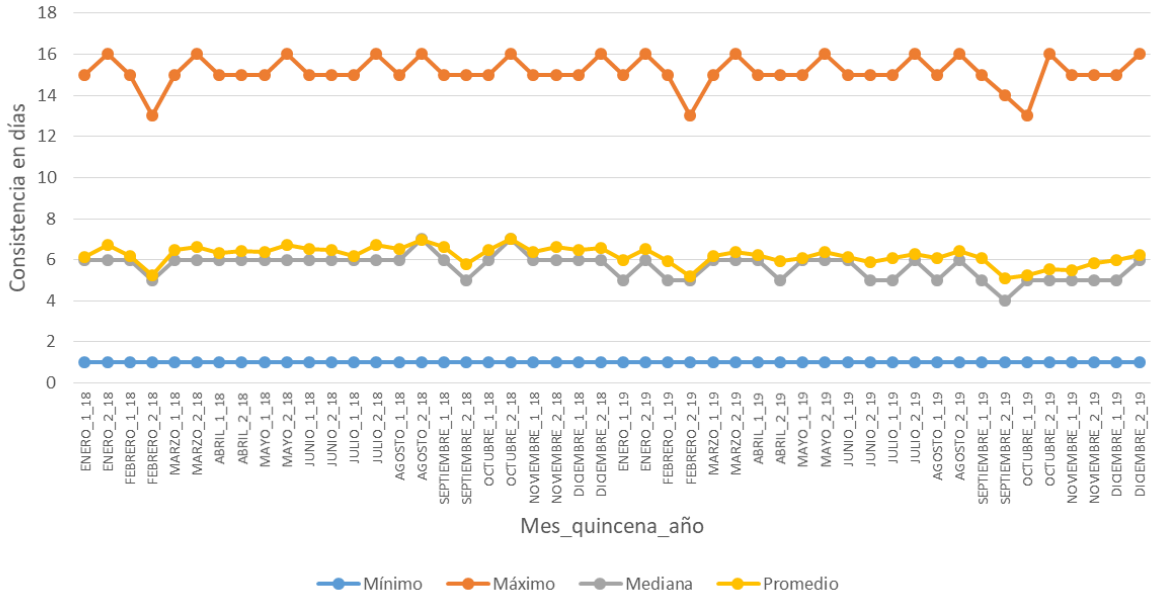


Gráfico 35: Indicadores quincenales de consistencia para grupo Tarjetas

Fuente: Elaboración propia

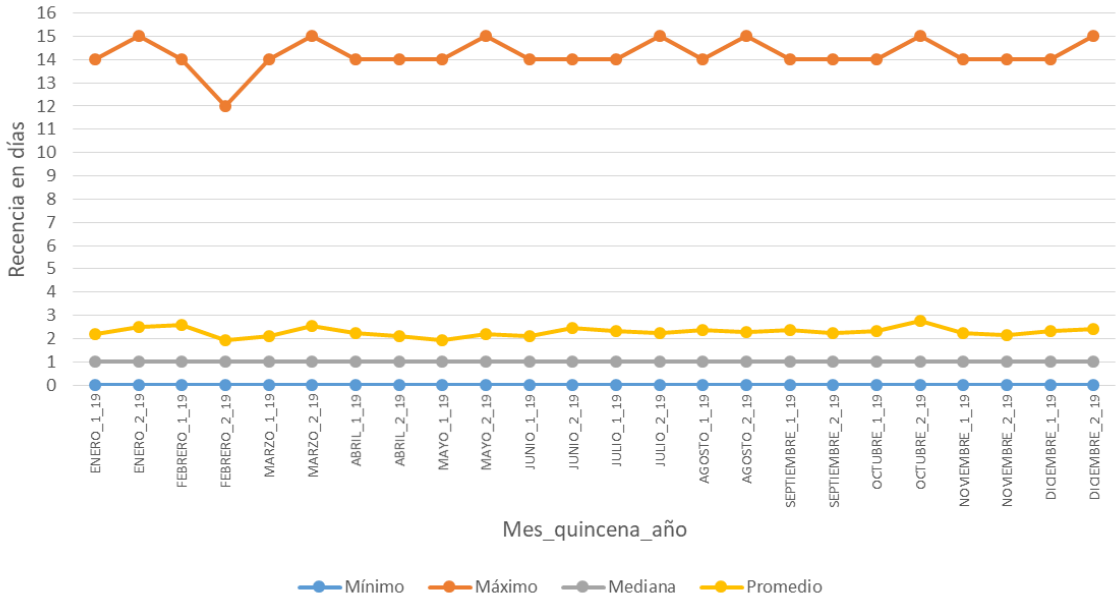


Gráfico 36: Indicadores quincenales de recencia para grupo App

Fuente: Elaboración propia

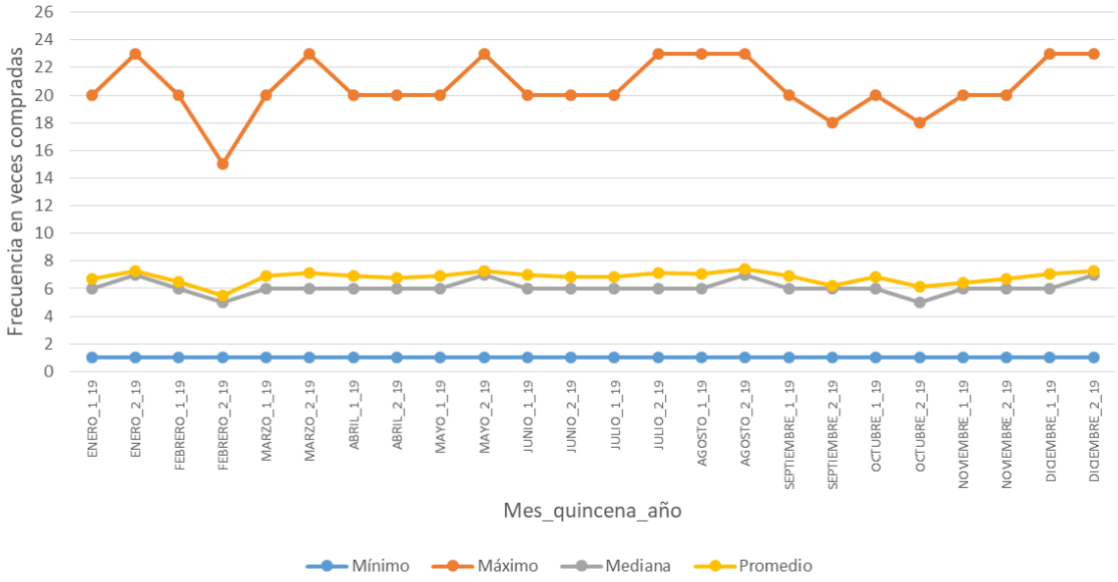


Gráfico 37: Indicadores quincenales de frecuencia para grupo App

Fuente: Elaboración propia

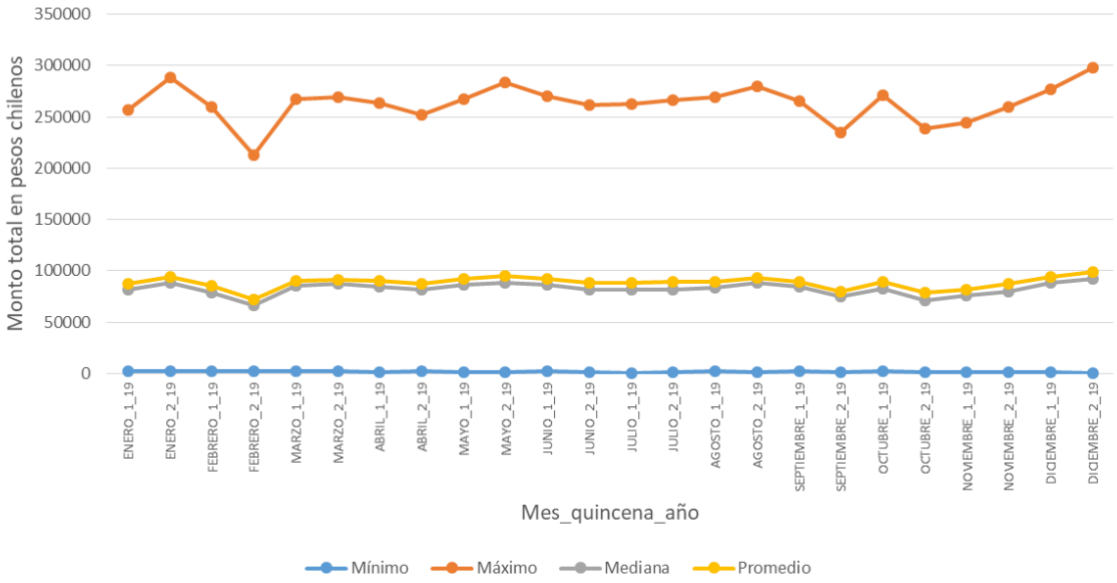


Gráfico 38: Indicadores quincenales de monto total para grupo App

Fuente: Elaboración propia

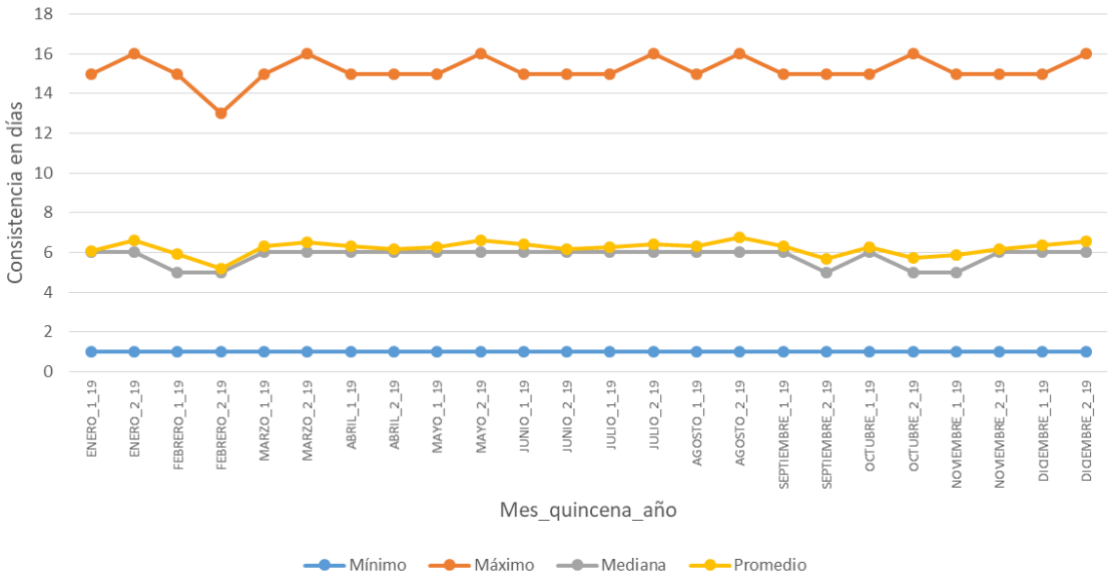


Gráfico 39: Indicadores quincenales de consistencia para grupo App

Fuente: Elaboración propia

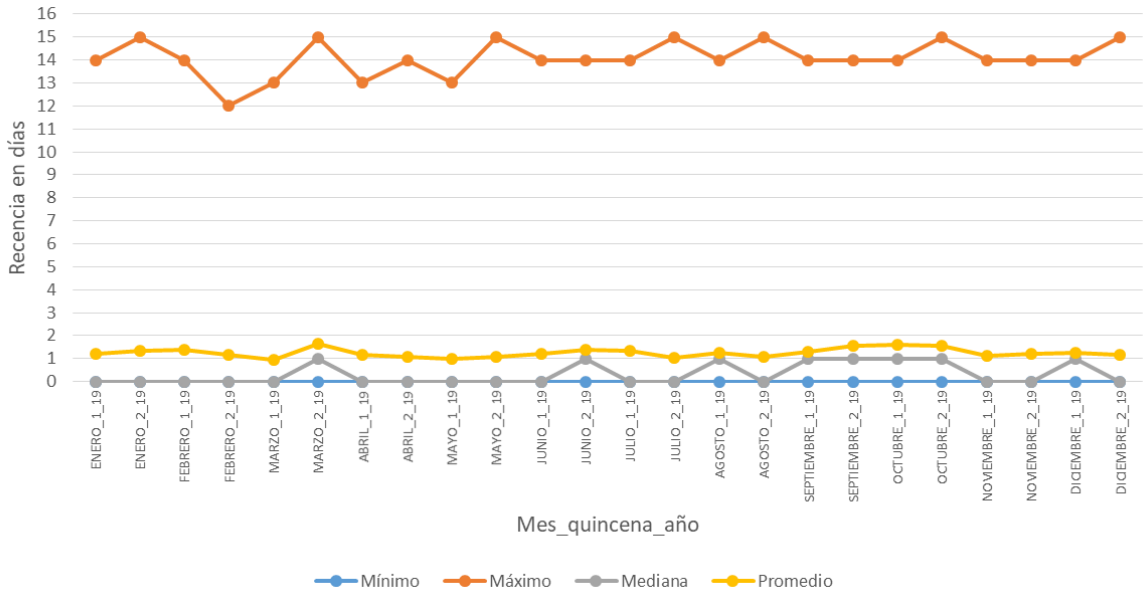


Gráfico 40: Indicadores quincenales de recencia para grupo Tarjetas y App

Fuente: Elaboración propia

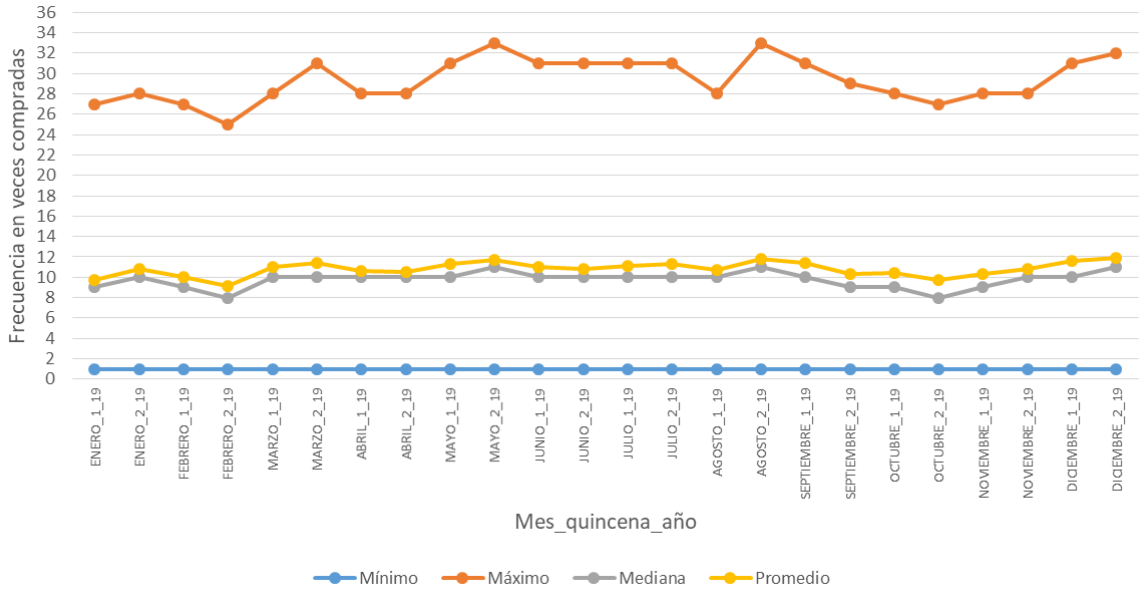


Gráfico 41: Indicadores quincenales de frecuencia para grupo Tarjetas y App

Fuente: Elaboración propia

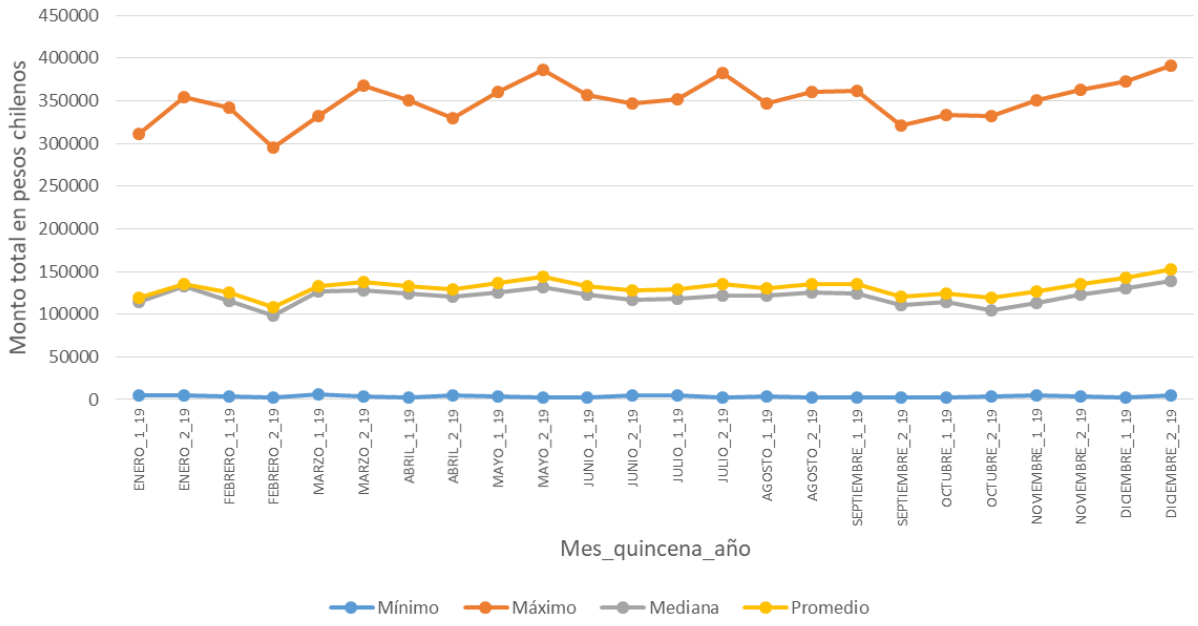


Gráfico 42: Indicadores quincenales de monto total para grupo Tarjetas y App

Fuente: Elaboración propia

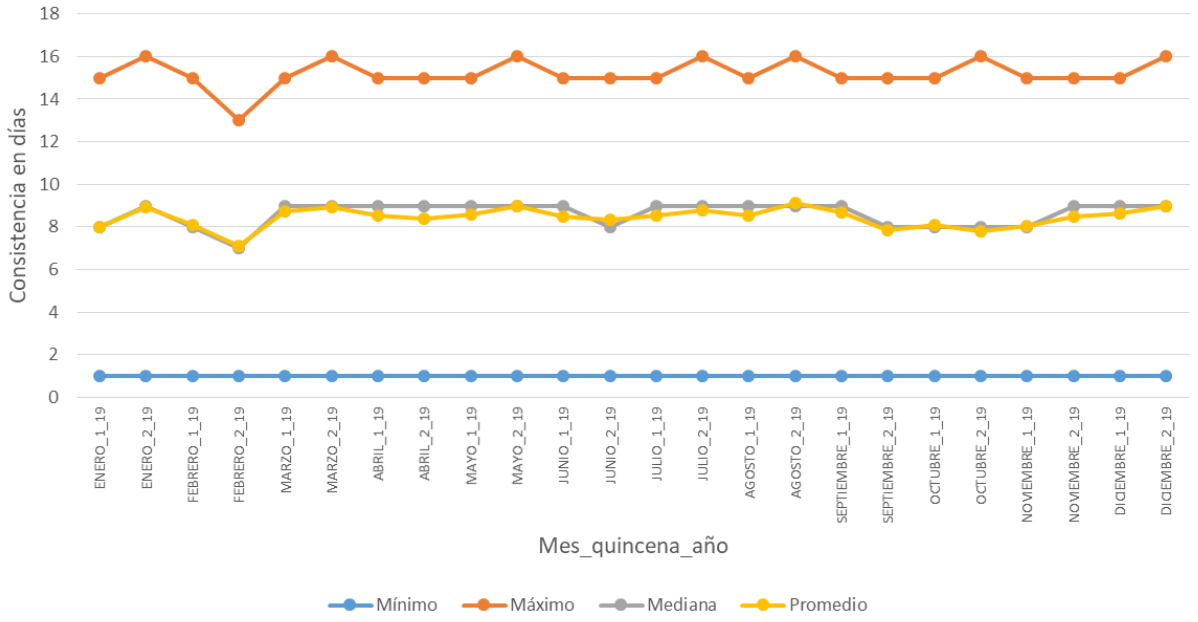


Gráfico 43: Indicadores quincenales de consistencia para grupo Tarjetas y App

Fuente: Elaboración propia

13.5 Anexo V: Cantidad y distribución quincenal de taxistas para grupo App y Tarjetas y App

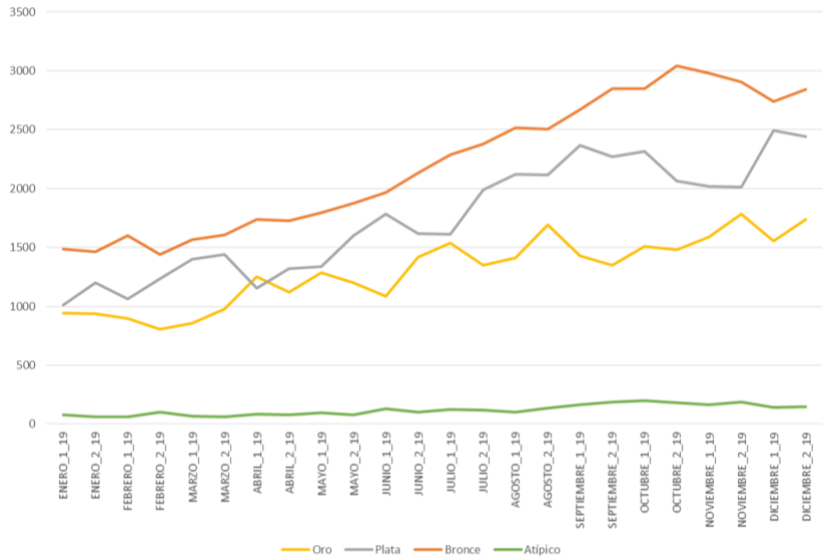


Gráfico 44: Evolución de clientes activos del grupo App según segmento al que pertenecen

Fuente: Elaboración propia

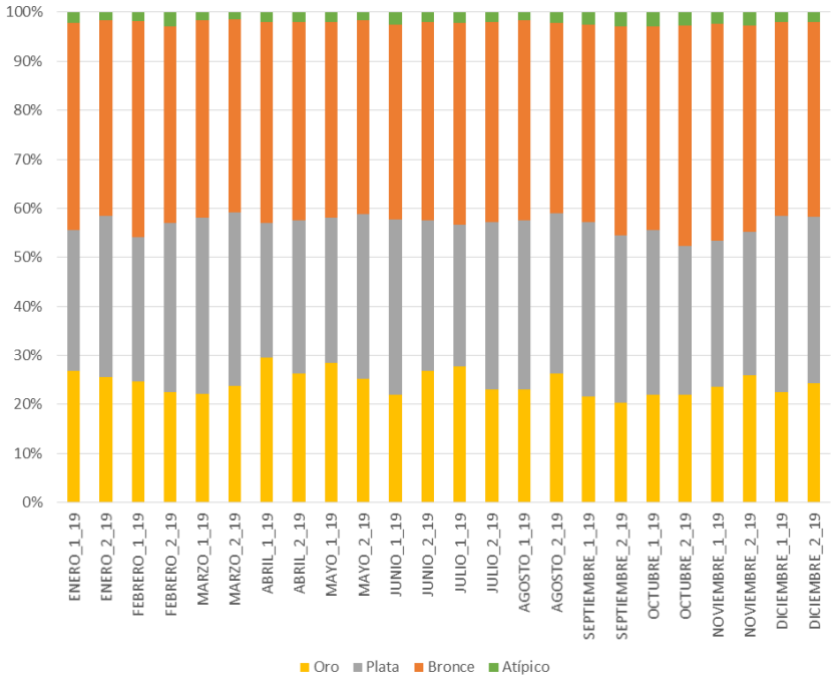


Gráfico 45: Distribución mensual de clientes activos del grupo App según segmento al que pertenece

Fuente: Elaboración propia

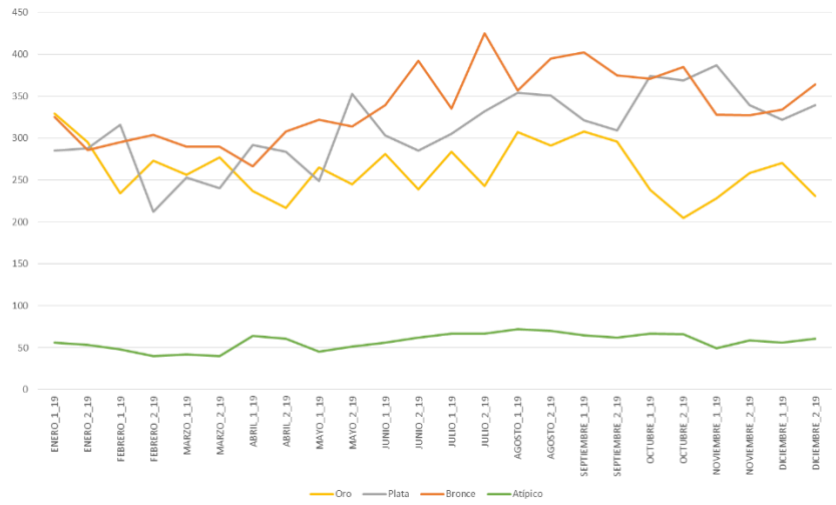


Gráfico 46: Evolución de clientes activos del grupo Tarjetas y App según segmento al que pertenecen

Fuente: Elaboración propia

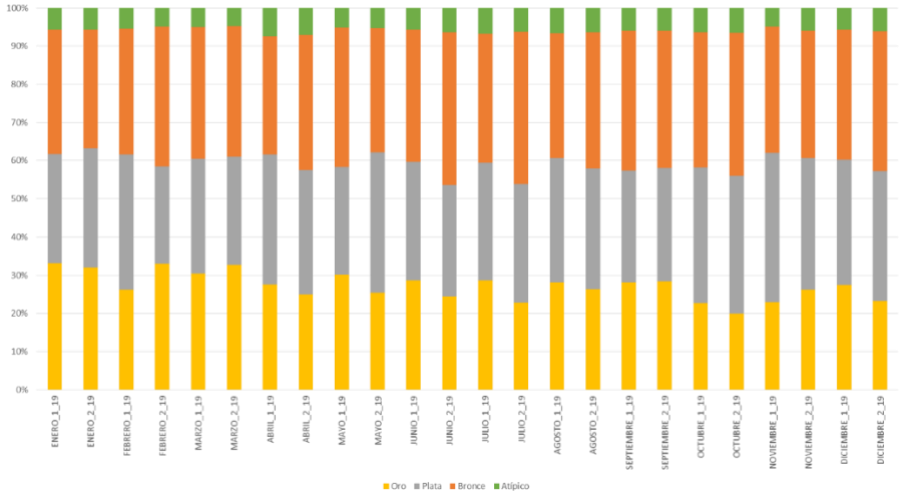


Gráfico 47: Distribución mensual de clientes activos del grupo Tarjetas y App según segmento al que pertenecen

Fuente: Elaboración propia

13.6 Anexo VI: Matrices de correlación de variables para los distintos grupos

	es_inactividad	zona_moda	monto_total	ultima_recencia	frecuencia_total	consistencia_total	recencia_promedio	frecuencia_promedio	monto_promedio	consistencia_promedio
es_inactividad	1,00	-0,03	-0,49	0,36	-0,46	0,36	0,25	-0,23	-0,27	0,24
zona_moda	-0,03	1,00	-0,02	0,00	-0,04	0,06	0,05	-0,09	-0,07	0,06
monto_total	-0,49	-0,02	1,00	-0,32	0,89	-0,56	-0,55	0,69	0,83	-0,50
ultima_recencia	0,36	0,00	-0,32	1,00	-0,32	0,53	0,48	-0,24	-0,25	0,19
frecuencia_total	-0,46	-0,04	0,89	-0,32	1,00	-0,62	-0,56	0,82	0,72	-0,59
consistencia_total	0,36	0,06	-0,56	0,53	-0,62	1,00	0,53	-0,65	-0,58	0,51
recencia_promedio	0,25	0,05	-0,55	0,48	-0,56	0,53	1,00	-0,66	-0,64	0,23
frecuencia_promedio	-0,23	-0,09	0,69	-0,24	0,82	-0,65	-0,66	1,00	0,82	-0,66
monto_promedio	-0,27	-0,07	0,83	-0,25	0,72	-0,58	-0,64	0,82	1,00	-0,54
consistencia_promedio	0,24	0,06	-0,50	0,19	-0,59	0,51	0,23	-0,66	-0,54	1,00

Figura 13: Matriz de correlación de variables para grupo Tarjetas

Fuente: Elaboración propia

	es_inactividad	zona_moda	monto_total	ultima_recencia	frecuencia_total	consistencia_total	recencia_promedio	frecuencia_promedio	monto_promedio	consistencia_promedio
es_inactividad	1,00	-0,02	-0,41	0,39	-0,40	0,34	0,25	-0,24	-0,29	0,33
zona_moda	-0,02	1,00	0,02	-0,01	0,00	0,04	0,00	-0,05	-0,02	0,05
monto_total	-0,41	0,02	1,00	-0,33	0,89	-0,50	-0,46	0,58	0,75	-0,53
ultima_recencia	0,39	-0,01	-0,33	1,00	-0,33	0,58	0,59	-0,34	-0,35	0,29
frecuencia_total	-0,40	0,00	0,89	-0,33	1,00	-0,57	-0,48	0,72	0,65	-0,62
consistencia_total	0,34	0,04	-0,50	0,58	-0,57	1,00	0,56	-0,73	-0,65	0,63
recencia_promedio	0,25	0,00	-0,46	0,59	-0,48	0,56	1,00	-0,65	-0,64	0,32
frecuencia_promedio	-0,24	-0,05	0,58	-0,34	0,72	-0,73	-0,65	1,00	0,80	-0,74
monto_promedio	-0,29	-0,02	0,75	-0,35	0,65	-0,65	-0,64	0,80	1,00	-0,62
consistencia_promedio	0,33	0,05	-0,53	0,29	-0,62	0,63	0,32	-0,74	-0,62	1,00

Figura 14: Matriz de correlación de variables para grupo App

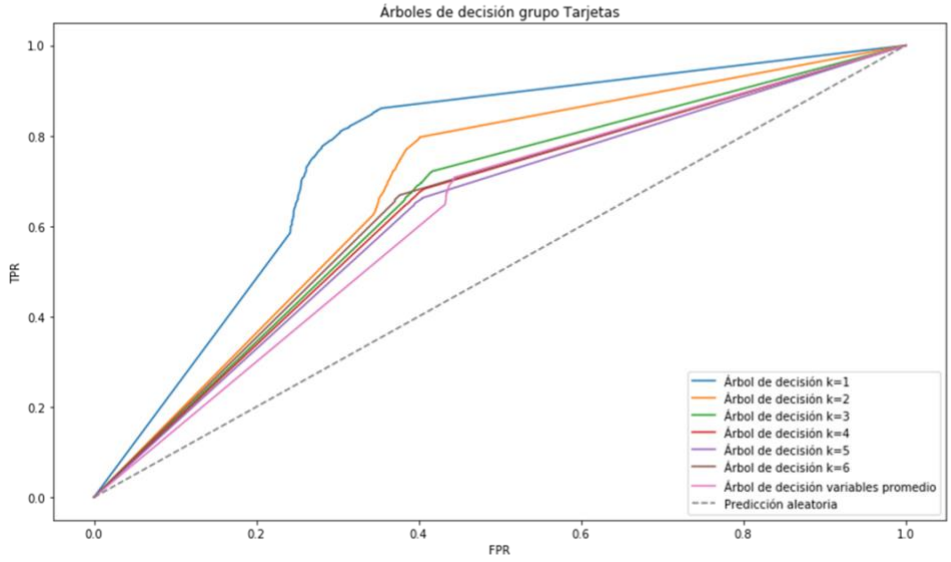
Fuente: Elaboración propia

	es_inactividad	zona_moda	monto_total	ultima_recencia	frecuencia_total	consistencia_total	recencia_promedio	frecuencia_promedio	monto_promedio	consistencia_promedio
es_inactividad	1,00	-0,04	-0,46	0,42	-0,44	0,35	0,32	-0,26	-0,29	0,30
zona_moda	-0,04	1,00	0,01	-0,02	-0,02	0,03	0,00	-0,05	-0,02	0,06
monto_total	-0,46	0,01	1,00	-0,39	0,86	-0,58	-0,59	0,73	0,91	-0,60
ultima_recencia	0,42	-0,02	-0,39	1,00	-0,40	0,61	0,56	-0,35	-0,35	0,30
frecuencia_total	-0,44	-0,02	0,86	-0,40	1,00	-0,67	-0,60	0,90	0,77	-0,72
consistencia_total	0,35	0,03	-0,58	0,61	-0,67	1,00	0,59	-0,70	-0,60	0,64
recencia_promedio	0,32	0,00	-0,59	0,56	-0,60	0,59	1,00	-0,65	-0,64	0,38
frecuencia_promedio	-0,26	-0,05	0,73	-0,35	0,90	-0,70	-0,65	1,00	0,80	-0,77
monto_promedio	-0,29	-0,02	0,91	-0,35	0,77	-0,60	-0,64	0,80	1,00	-0,62
consistencia_promedio	0,30	0,06	-0,60	0,30	-0,72	0,64	0,38	-0,77	-0,62	1,00

Figura 15: Matriz de correlación de variables para grupo Tarjetas y App

Fuente: Elaboración propia

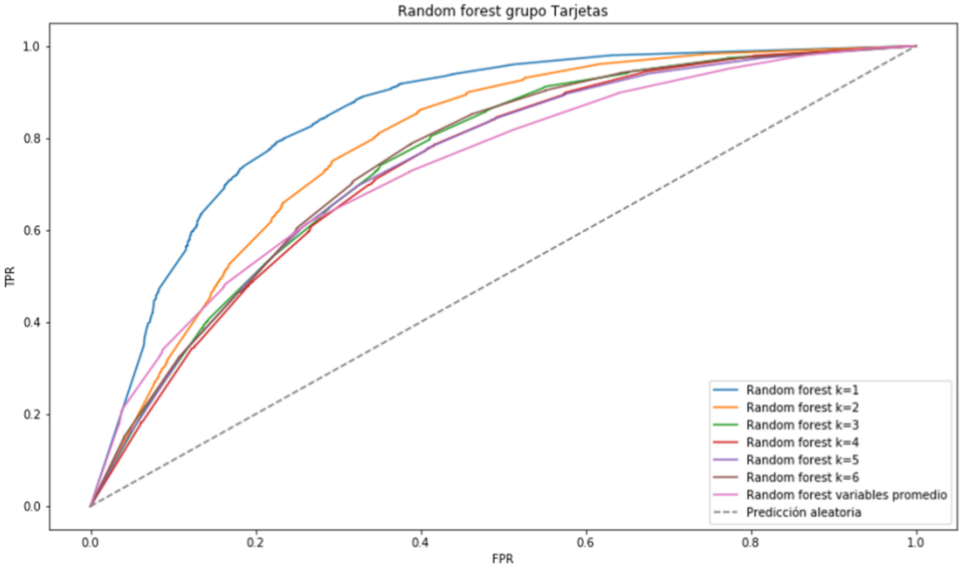
13.7 Anexo VII: Resultados para definición de parámetro k a utilizar en modelos del grupo Tarjetas



Grupo Tarjetas: Árboles de decisión		
Variables	**k**	**ROC AUC**
$monto\_total_{ik}$ $ultima\_recencia_i$ $frecuencia\_total_{ik}$ $consistencia\_total_{ik}$ $zona\_moda_i$	1	76%
	2	69%
	3	65%
	4	64%
	5	63%
	6	65%
$R\_quincenal\_promedio_i$ $recencia\_desvest_i$ $F\_quincenal\_promedio_i$ $frecuencia\_desvest_i$ $M\_quincenal\_promedio_i$ $monto\_desvest_i$ $C\_quincenal\_promedio_i$ $consistencia\_desvest_i$ $zona\_moda_i$		62%

Gráfico 48: Resultados modelos árboles de decisión para grupo Tarjetas

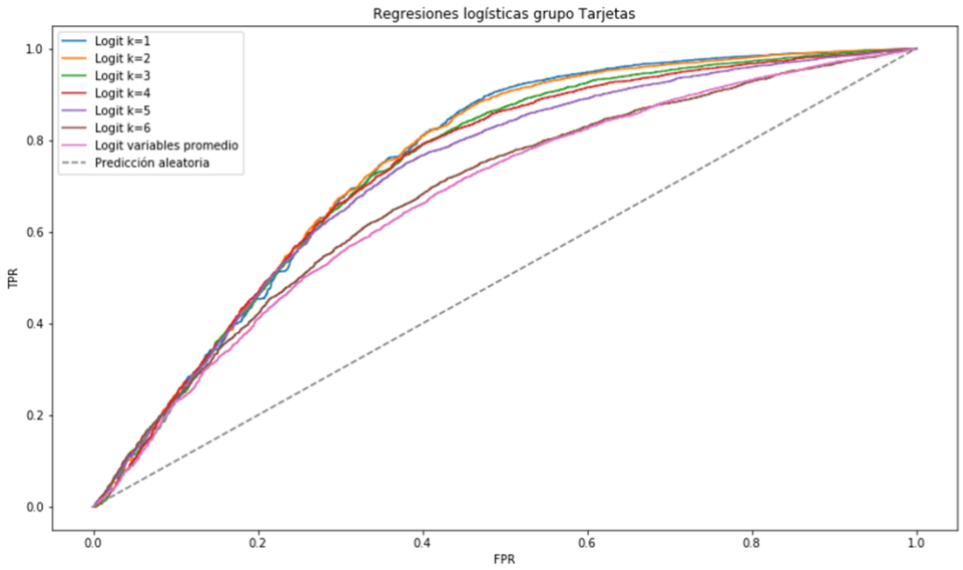
Fuente: Elaboración propia



Grupo Tarjetas: Random forest		
Variables	**k**	**ROC AUC**
$monto\_total_{ik}$ $ultima\_recencia_i$ $frecuencia\_total_{ik}$ $consistencia\_total_{ik}$ $zona\_moda_i$	1	85%
	2	79%
	3	75%
	4	74%
	5	74%
	6	75%
$R\_quincenal\_promedio_i$ $recencia\_desvest_i$ $F\_quincenal\_promedio_i$ $frecuencia\_desvest_i$ $M\_quincenal\_promedio_i$ $monto\_desvest_i$ $C\_quincenal\_promedio_i$ $consistencia\_desvest_i$ $zona\_moda_i$		74%

Gráfico 49: Resultados modelos random forest para grupo Tarjetas

Fuente: Elaboración propia

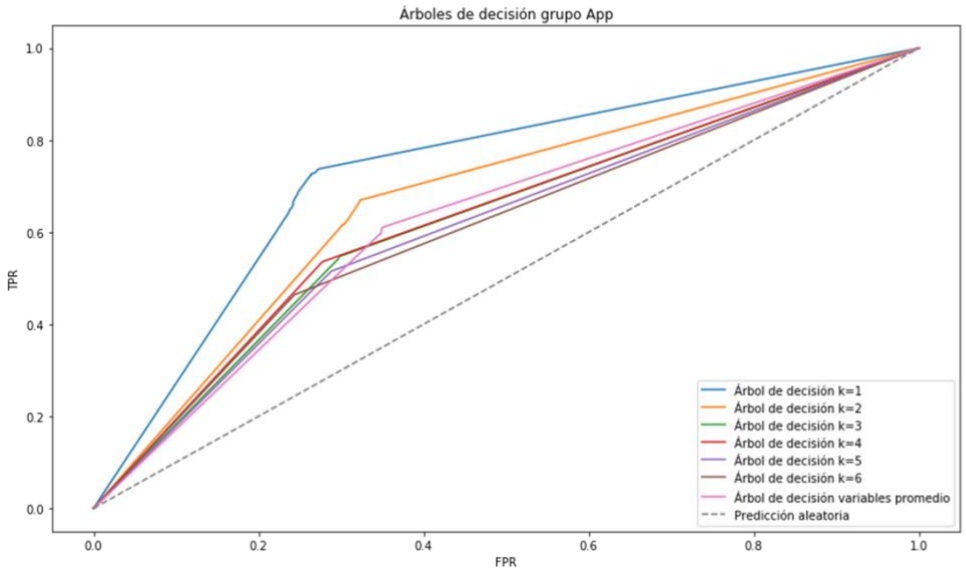


Grupo Tarjetas: Regresión logística (Logit)		
Variables	k	ROC AUC
$monto\_total_{ik}$ $ultima\_recencia_i$ $zona\_moda_i$	1	75%
	2	75%
	3	74%
	4	73%
	5	72%
	6	68%
$R\_quincenal\_promedio_i$ $M\_quincenal\_promedio_i$ $zona\_moda_i$		67%

Gráfico 50: Resultados modelos logit para grupo Tarjetas

Fuente: Elaboración propia

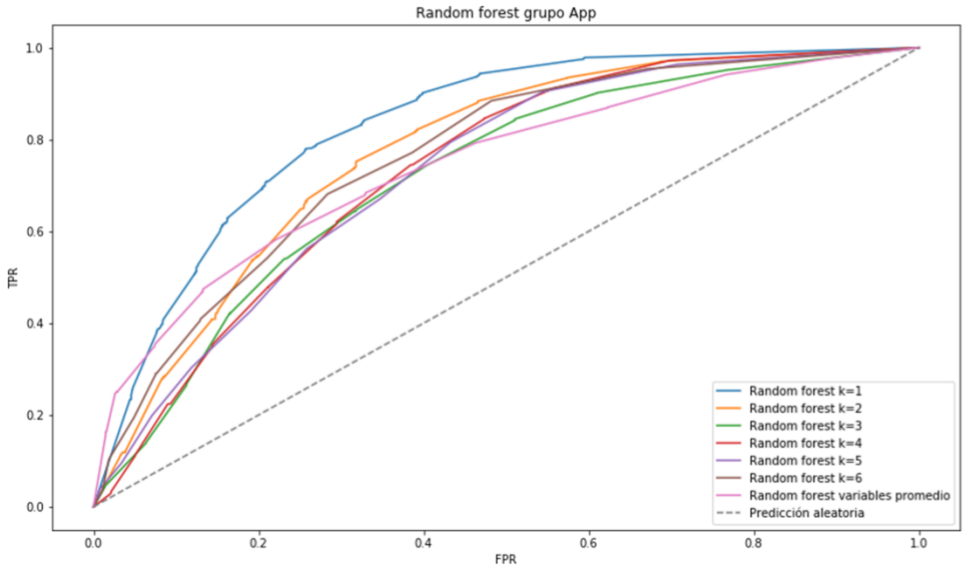
13.8 Anexo VIII: Resultados para definición de parámetro k a utilizar en modelos del grupo App



Grupo App: Árboles de decisión		
Variables	k	ROC AUC
$monto\_total_{ik}$ $ultima\_recencia_i$ $frecuencia\_total_{ik}$ $consistencia\_total_{ik}$ $zona\_moda_i$	1	73%
	2	67%
	3	62%
	4	63%
	5	61%
	6	61%
$R\_quincenal\_promedio_i$ $recencia\_desvest_i$ $F\_quincenal\_promedio_i$ $frecuencia\_desvest_i$ $M\_quincenal\_promedio_i$ $monto\_desvest_i$ $C\_quincenal\_promedio_i$ $consistencia\_desvest_i$ $zona\_moda_i$		63%

Gráfico 51: Resultados modelos árboles de decisión para grupo App

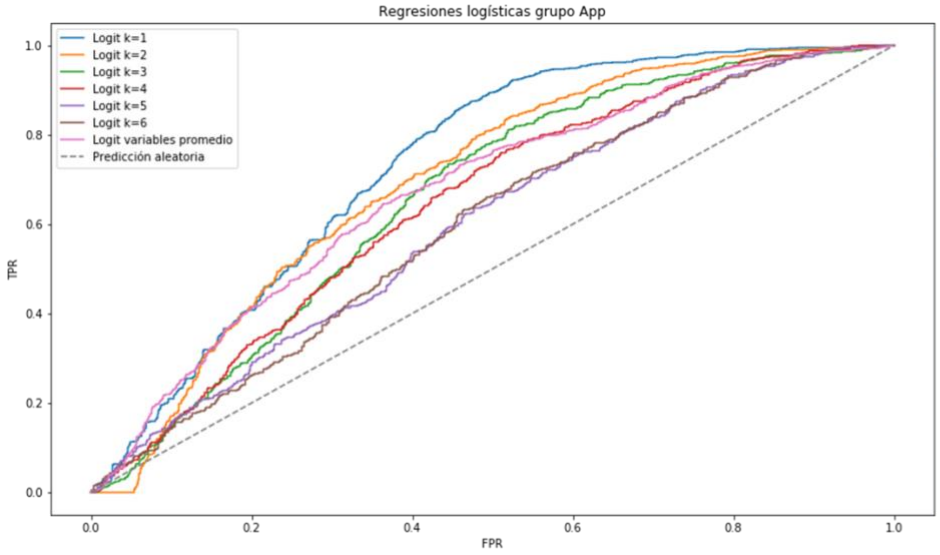
Fuente: Elaboración propia



Grupo App: Random forest		
Variables	**k**	**ROC AUC**
$monto\_total_{ik}$ $ultima\_recencia_i$ $frecuencia\_total_{ik}$ $consistencia\_total_{ik}$ $zona\_moda_i$	1	83%
	2	77%
	3	72%
	4	73%
	5	73%
	6	76%
$R\_quincenal\_promedio_i$ $recencia\_desvest_i$ $F\_quincenal\_promedio_i$ $frecuencia\_desvest_i$ $M\_quincenal\_promedio_i$ $monto\_desvest_i$ $C\_quincenal\_promedio_i$ $consistencia\_desvest_i$ $zona\_moda_i$		75%

Gráfico 52: Resultados modelos random forest para grupo App

Fuente: Elaboración propia

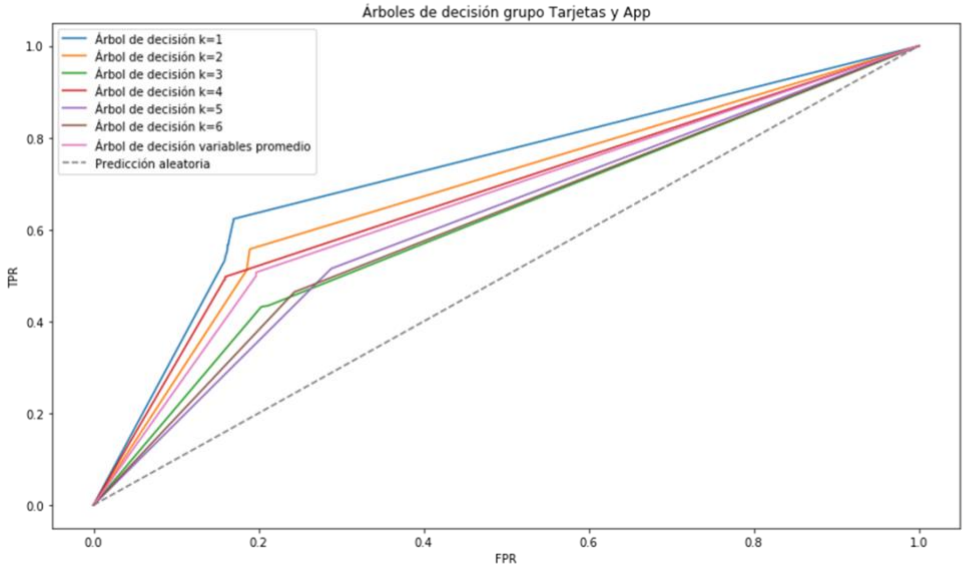


Grupo App: Regresión logística (Logit)		
Variables	**k**	**ROC AUC**
$monto\_total_{ik}$ $ultima\_recencia_i$ $zona\_moda_i$	1	73%
	2	70%
	3	66%
	4	65%
	5	60%
	6	60%
$R\_quincenal\_promedio_i$ $M\_quincenal\_promedio_i$ $zona\_moda_i$		67%

Gráfico 53: Resultados modelos logit para grupo App

Fuente: Elaboración propia

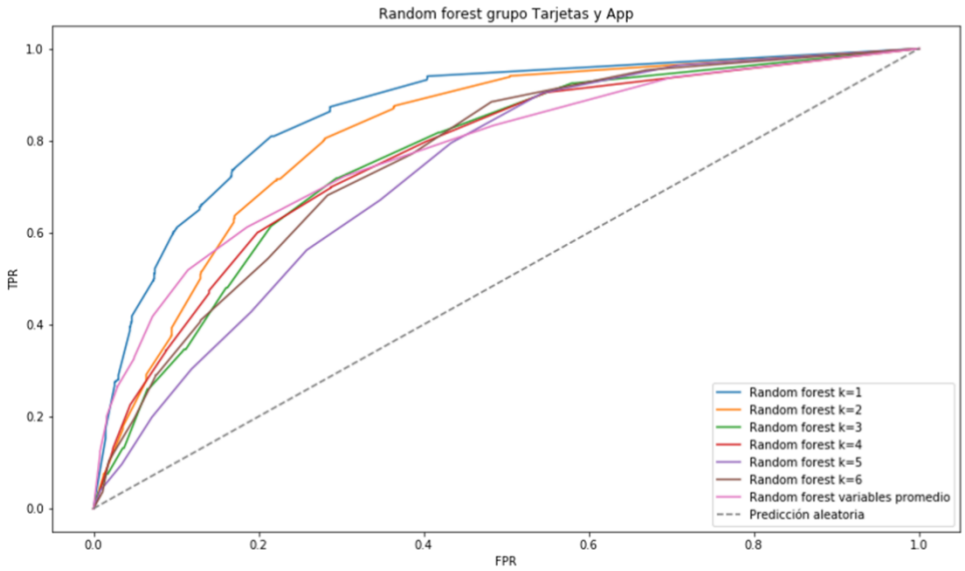
13.9 Anexo IX: Resultados para definición de parámetro k a utilizar en modelos del grupo Tarjetas y App



Grupo Tarjetas y App: Árboles de decisión		
Variables	k	ROC AUC
$monto\_total_{ik}$ $ultima\_recencia_i$ $frecuencia\_total_{ik}$ $consistencia\_total_{ik}$ $zona\_moda_i$	1	72%
	2	68%
	3	61%
	4	67%
	5	61%
	6	61%
$R\_quincenal\_promedio_i$ $recencia\_desvest_i$ $F\_quincenal\_promedio_i$ $frecuencia\_desvest_i$ $M\_quincenal\_promedio_i$ $monto\_desvest_i$ $C\_quincenal\_promedio_i$ $consistencia\_desvest_i$ $zona\_moda_i$		65%

Gráfico 54: Resultados modelos árboles de decisión para grupo Tarjetas y App

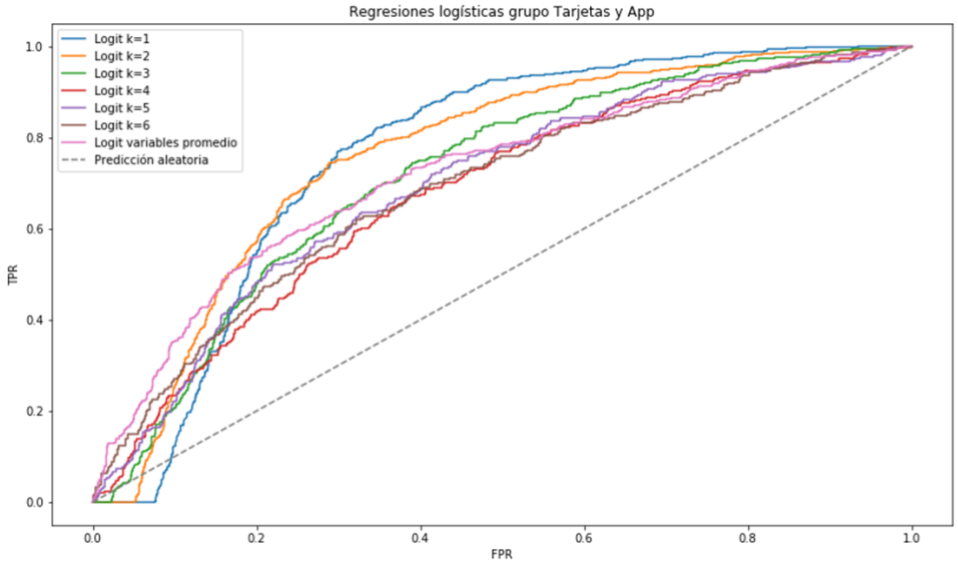
Fuente: Elaboración propia



Grupo Tarjetas y App: Random forest		
Variables	k	ROC AUC
$monto\_total_{ik}$ $ultima\_recencia_i$ $frecuencia\_total_{ik}$ $consistencia\_total_{ik}$ $zona\_moda_i$	1	87%
	2	82%
	3	77%
	4	77%
	5	73%
	6	76%
$R\_quincenal\_promedio_i$ $recencia\_desvest_i$ $F\_quincenal\_promedio_i$ $frecuencia\_desvest_i$ $M\_quincenal\_promedio_i$ $monto\_desvest_i$ $C\_quincenal\_promedio_i$ $consistencia\_desvest_i$ $zona\_moda_i$		78%

Gráfico 55: Resultados modelos random forest para grupo Tarjetas y App

Fuente: Elaboración propia



Grupo Tarjetas y App: Regresión logística (Logit)		
Variables	k	ROC AUC
$monto\_total_{ik}$ $ultima\_recencia_i$ $zona\_moda_i$	1	76%
	2	76%
	3	62%
	4	68%
	5	70%
	6	69%
$R\_quincenal\_promedio_i$ $M\_quincenal\_promedio_i$ $zona\_moda_i$		72%

Gráfico 56: Resultados modelos logit para grupo Tarjetas y App

Fuente: Elaboración propia

13.10 Anexo X: Análisis de sensibilidad para períodos independientes

Grupo	Período	Criterio de inactividad (quincenas)	Observaciones de inactividad identificadas	Observaciones de actividad identificadas	ROC AUC	
Tarjetas	2018	1+	53.733	21.439	89%	
		3+	25.123	22.261	83%	
		5+	14.754	27.147	84%	Criterio original
		7+	10.980	28.652	84%	
		10+	7.539	30.599	84%	
		15+	4.062	33.172	85%	
	2019	1+	45.042	16.281	89%	
		3+	19.122	18.871	84%	
		5+	12.936	20.224	85%	Criterio original
		7+	10.085	21.195	85%	
		10+	7.416	22.640	86%	
		15+	4.300	24.949	86%	

Cuadro 11: Resultados análisis de sensibilidad para períodos independientes (grupo Tarjetas)

Fuente: Elaboración propia