UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

DEFINICIÓN Y ANÁLISIS DE CLIENTES DIGITALES, CON TARJETA DE CRÉDITO ABIERTA, EN UNA COMPAÑÍA DE RETAIL FINANCIERO

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

FELIPE IGNACIO JORQUERA VALLADARES

PROFESORA GUÍA
CAROLINA SEGOVIA RIQUELME

MIEMBROS DE LA COMISIÓN
PABLO MARÍN VICUÑA
MÁXIMO BOSCH PASSALACQUA

SANTIAGO DE CHILE
2020

RESUMEN DE LA MEMORIA PARA OPTAR AL
TÍTULO DE: Ingeniero Civil Industrial
POR: Felipe Ignacio Jorquera Valladares
FECHA: 27/04/2020
PROFESORA GUÍA: Carolina Segovia Riquelme

DEFINICIÓN Y ANÁLISIS DE CLIENTES DIGITALES, CON TARJETA DE CRÉDITO ABIERTA, EN UNA COMPAÑÍA DE RETAIL FINANCIERO

El presente trabajo se desarrolla en una empresa perteneciente a la industria de Retail Financiero, la cual se referirá en este documento como empresa "RF". El objetivo de este trabajo es definir y analizar a los clientes digitales de la tarjeta RF, para sugerir mejoras en las campañas de marketing dirigido existentes.

Frente a esto, se utiliza una metodología mixta, considerando las metodologías CRISP-DM y KDD. El trabajo realizado se divide en tres etapas. La primera etapa corresponde a la ejecución de un modelo de segmentación, en base a hitos digitales, las cuales son diversas acciones que realizan los usuarios en los canales digitales. La segunda etapa es la realización de una regresión de Diferencias en Diferencias, que permite saber si existen cambios en transacciones y pagos, después de un proceso de digitalización. La tercera etapa se compone de tres modelos predictivos, dos de ellos entregan un pronóstico de digitalización de clientes a corto plazo, mientras que el tercer modelo predictivo entrega una predicción de uso de los canales digitales en los próximos 4 meses. Para ejecutar estos pasos, se utilizan datos sociodemográficos de los clientes, datos sobre el uso de canales digitales y datos de la tarjeta RF, como transacciones con tarjeta, pagos, deuda y monto de cupo total.

Como resultados de los modelos realizados, el modelo de segmentación se compone de 5 segmentos de clientes, con respecto a sus acciones digitales. Estos segmentos se denominan "No Digital", "App Incipiente", "Sólo Sitio Web", "Full App" y "Full Digital". Los segmentos "Full App" y "Full Digital" poseen promedios de edad menores a los segmentos restantes, por ende, se componen de clientes más jóvenes.

Desde la regresión de Diferencias en Diferencias, se observa que posterior a la digitalización de un cliente, este evento produce un aumento en transacciones de CLP $3.617 y una disminución de pagos de CLP $9.595. Con relación a los modelos predictivos generados, los dos primeros muestran resultados insatisfactorios, ya que la métrica de *Precision* posee cifras menores al 10%. El tercer modelo predictivo, de predicción de uso de canales digitales, entrega resultados superiores al 80% en las métricas *Accuracy*, *Precision* y *Recall*.

Dentro de las conclusiones del trabajo, se observan resultados concretos, pero limitados debido a que se consideran pocos meses en la obtención de datos. Por ello, se recomienda continuar con la realización de estos modelos y observar si existen cambios o se mantienen los resultados obtenidos. En base a los modelos predictivos, se recomienda establecer una predicción de un evento que no sea difícil de cumplir, para evitar datos desbalanceados entre las etiquetas positivas y negativas de la variable a predecir.

Agradecimientos

Primero que todo, quiero agradecer a mis padres, Pilar y Juan Carlos. Gracias a ellos aprendí sobre los principios y valores de la vida, sobre cómo enfrentar el mundo y los desafíos del día a día. Les agradezco a ustedes, por todo el esfuerzo que hicieron para que yo llegue hasta acá.

También les agradezco a mis hermanos, Nicolás y Claudia, al Nico lo recuerdo desde que nació, he vivido cerca de él todas las etapas de su vida y ha sido un apoyo muy importante siempre. A mi hermana, le agradezco por los consejos, el apoyo y la preocupación durante mi etapa universitaria y también por siempre ayudar a unirnos entre los 3 hermanos.

Les agradezco a los demás familiares, los que me acompañaron siempre y los que se preocuparon de cómo me estuvo yendo en la universidad, desde que entré hasta ahora, cuando ya estoy saliendo.

Dentro de mis amigos de la U, les agradezco a las personas geniales que conocí en estos años. Al grupo de Los Bultos, al Manuel, la Karen, Lema, Piccolis, Ja-Ja, Oliva, Carlete y Véliz. Los conocí a casi todos en primero y ahora nos seguimos juntando, ojalá los pueda seguir viendo y pasando bien en lo que viene. Le agradezco a grandes amigos que conocí en estos años, al Renato, Zelada, el Seba, Carlos y Pipe Alvarado. Muchos apañaron en el estudio, en los almuerzos y en uno o varios años de la u. Les agradezco también a las tremendas personas que conocí en el Proyecto Gota y especialmente a algunos con los que sigo hablando hoy en día, al Rafa, Guille, el Joaco y el Nico.

Les agradezco también a mis amigos del colegio, en especial al Franco y al Ignacio, los conozco de hace más de 10 años y siempre han estado conmigo cuando los necesito.

Por último, les agradezco a mis profesores guía y coguía, Carolina y Pablo, por la colaboración y disposición que tuvieron durante este año. En especial a Carolina, que me dio la oportunidad de realizar el trabajo de Memoria, depositando su confianza hacia mí y ayudándome siempre que lo necesité. Agradezco también a las personas que conocí durante el trabajo de Memoria, a la Fer, la Caro y Andrés. Me ayudaron mucho en los meses que hice este trabajo de Memoria.

Termino por agradecerle a la vida por lo que viví en estos 7 años. Tuve todo tipo de experiencias, buenas, malas y otros momentos de mucha exigencia para mí. También conocí la realidad de otras personas y sobre la realidad de este país. Sólo quiero que todos estos años vividos sirvan de aprendizaje para el futuro mío y de mi círculo cercano y ojalá lograr finalmente ser un aporte para la sociedad y para este país.

Tabla de Contenido

Índice de tablas

Índice de ilustraciones

Índice de fórmulas

1 Introducción

El presente informe se desarrolla en el marco del Trabajo de Título, de la carrera de Ingeniería Civil Industrial, en la Universidad de Chile. En específico, la temática a abordar es un proyecto de Data Science con una empresa como contraparte.

En este caso, el alumno desarrolla esta labor con una empresa de Retail financiero. El Retail financiero es una industria donde las cadenas de Retail añaden instrumentos financieros hacia sus clientes, generando un modelo de negocio adicional al Retail tradicional. Es relevante considerar que las compañías generan muchos datos sobre sus clientes, por lo que la compañía de contraparte es apta para desarrollar un Trabajo de Memoria en Data Sciente, en Ingeniería Civil Industrial.

En las primeras secciones del trabajo se abordan antecedentes principales, para introducir el contexto del Retail financiero en Chile y de la empresa de contraparte, en el momento de la realización de este trabajo. También se describe el proyecto a realizar, junto con sus objetivos generales y específicos.

En la siguiente sección se expone el marco conceptual del trabajo a realizar, donde se introducen conceptos importantes que previamente se deben conocer para comprender el proyecto descrito. Posteriormente, se expone la Metodología utilizada, con los pasos a realizar para llevar a cabo el trabajo y así cumplir con los objetivos deseados. Luego siguen las secciones de Alcances del trabajo de Título y Resultados esperados para este trabajo.

Finalmente, se muestran en el Desarrollo de la Metodología se muestran los resultados obtenidos en el presente trabajo. Se finaliza con las conclusiones asociadas al proyecto ejecutado, junto con recomendaciones futuras para considerar en un tentativo próximo proyecto, relacionado con el trabajo que se expone en este documento.

2 Antecedentes Generales

En esta sección se presentarán antecedentes principales, asociados a la industria y la empresa en la que se desenvuelve este trabajo. De esta forma, se introduce al lector en un contexto que permita comprender el trabajo realizado.

2.1 Industria Retail Financiero en Chile

La industria a la cual se enfoca el presente trabajo es el Retail Financiero en Chile, la cual es resultante de los negocios de las grandes cadenas de multitiendas, que durante los años 80 tuvieron gran éxito a nivel nacional, destacándose Ripley, Falabella y Almacenes París como los exponentes principales de esta industria en esta década. Dado esto, estas cadenas crearon sus respectivas empresas para administrar créditos a sus clientes [1].

Con el correr de los años, los avances tecnológicos y el desarrollo de los nuevos instrumentos de pago han generado una sustitución de instrumentos tradicionales, como efectivo y cheques, por aquellos electrónicos, tanto en Chile como en el resto del mundo [2]. Por ende, los negocios financieros en Chile se basan en el uso de tarjetas de crédito y en créditos de consumo para responder a las necesidades de sus clientes.

Dentro del área de retail financiero en Chile, se destaca el gran crecimiento de esta industria en los últimos años, dado el aumento de las tiendas físicas y los medios digitales, donde el cliente puede realizar transacciones y pagos con su tarjeta. Como ejemplo de ello, en 2017, un tercio de las ganancias de las compañías de retail se debe al negocio financiero [3].

Dado este aumento, se destaca como un actor relevante a la banca, la cual tiene una mayor cantidad de tarjetas de crédito disponibles, por ende, se convierte en un ente importante que recientemente ha participado en la adquisición de clientes de retail en el rubro financiero. Como ejemplos de este fenómeno se aprecian los casos de la fusión de Cencosud y Scotiabank, creando la tarjeta asociada a ambas empresas, también el traspaso de CMR al Banco Falabella y la venta de Presto al Banco BCI. Estos casos demuestran una tendencia incipiente a la fusión entre la banca y el retail financiero.

Dentro de las tendencias del mercado del retail, se destaca también la omnicanalidad y la relevancia del canal digital para lograr llegar al cliente. En los últimos años, las ventas en E-Commerce en Chile han aumentado de US$ 94 millones en 2004, a US$1.958 millones en 2014 y actualmente se proyecta una cifra de US$7.000 millones para el año 2019 [4]. Dentro de las prioridades de las compañías de retail,

en cuanto a su presupuesto, se destaca el caso de Falabella, la principal cadena de retail en Chile, que anuncia un 37% de sus inversiones destinado solo a la estrategia digital [5].

2.2 La Empresa

La empresa asociada a este trabajo de memoria es una empresa de retail financiero, la cual se referirá como empresa RF de ahora en adelante, que resulta de la fusión de las tiendas especializadas en productos electrónicos, línea blanca y muebles. Dicha fusión se produjo en el año 2008 y a esto se añade la adquisición de otra cadena de retail financiero en el año 2013.

Actualmente, la empresa RF vende variadas categorías de productos como muebles, electrodomésticos, tecnología y decoración del hogar. Sin embargo, el área de retail financiero tiene como servicio el ofrecimiento de diversos productos, donde el más importante es la tarjeta asociada a la empresa RF, la cual se mencionará como tarjeta RF de aquí en adelante. Esta tarjeta busca que sus clientes puedan adquirir productos en diversas tiendas asociadas al retail y luego pagar esa deuda a la compañía dentro de un plazo establecido en cuotas. Dentro de los usuarios de la compañía, se ubican todo tipo de consumidores de diverso rango etario y socioeconómico de Chile.

En cuanto al dimensionamiento de la empresa, ésta posee 151 tiendas alrededor del país, de las cuales 86 pertenecen a una de las cadenas asociadas y 65 pertenecen a la otra cadena de tiendas involucrada. Por otro lado, dentro de las 151 tiendas, 75 poseen un área de Servicios Financieros hacia sus clientes. Además, la tarjeta RF posee 5 millones de usuarios totales, que han abierto una cuenta con la tarjeta. De esta cantidad, 700 mil mantienen deuda con la entidad y se podría considerar por el momento como un grupo de clientes activos en la empresa de retail financiero.

Dentro de la ventaja competitiva que caracteriza a la empresa RF, se destaca la especialización por productos como tecnología, electro y decohogar, a diferencia de las cadenas de retail más importantes del país, las cuales se enfocan más en vestuario y no se caracterizan como especialistas en ninguna de las categorías de productos. En cuanto a los productos financieros, se observa un avance de la empresa en base la tasa de respuestas desfavorables (TRD) de reclamos de sus clientes, que disminuyó de un 46,3% a un 24,2% en el último año [6], dato favorable que se contrasta con la gran cantidad de reclamos que se reciben de parte de sus clientes, en comparación con otras tarjetas de retail.

Dentro de la estructura organizacional de la empresa RF, la División de Retail Financiero es la encargada de la gestión de clientes que poseen la tarjeta RF, la cual entrega diversas opciones de transacciones y deudas que pueden adquirir los consumidores con esta tarjeta de crédito. Estas transacciones en tiendas propias o en comercios asociados. Además, se ubican dentro de los servicios de Retail Financiero, la existencia de seguros hacia sus clientes, como seguros de salud, protección financiera y hogar y otros servicios de asistencia complementarios como servicios al hogar, salud y dentales.

2.3 Desempeño y objetivos en la empresa

En cuanto a la participación actual de la empresa RF, con respecto al mercado, se destacan las cifras presentadas por la empresa de investigación de mercado SCAN y Diario Financiero, que muestran la cantidad de clientes con tarjetas y actualmente operativas [7]. Excluyendo a las empresas líderes del mercado, la empresa RF concentra un 10% del total de tarjetas con operaciones y el líder en este rubro posee un 51,9% del total de tarjetas.

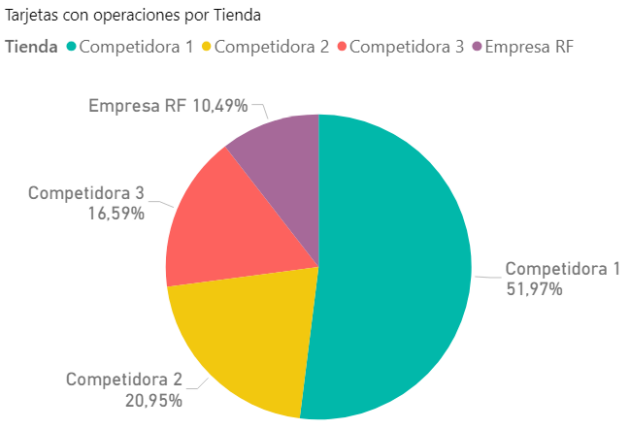


Ilustración 1: Participación de Mercado de empresas de retail financiero, excluyendo a las 3 compañías líderes del mercado

Dentro de la cantidad de clientes presentes en la empresa de Retail Financiero, se destaca la siguiente imagen con la evolución en los últimos meses:

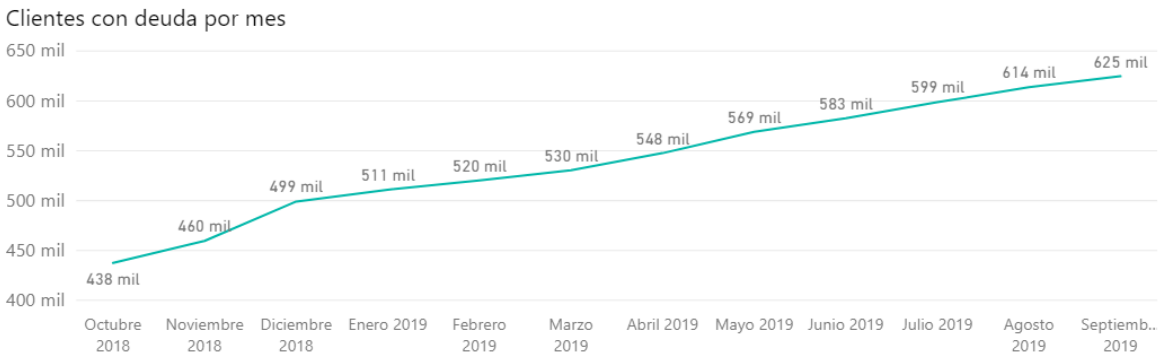


Ilustración 2: Clientes con deuda en la empresa RF, últimos 12 meses

En relación con las cifras presentadas, la compañía se encuentra en una etapa de crecimiento en base a la cantidad de clientes que posee la tarjeta RF. El enfoque que persigue en que el cliente realice la mayor cantidad posible de transacciones, utilizando los canales asociados a la tarjeta.

En adición a lo anterior, se destacan algunos aspectos relevantes sobre la relación entre cliente y la tarjeta RF. Actualmente, los clientes de la tarjeta pueden utilizar tiendas físicas o canales digitales para utilizar la tarjeta. Los canales digitales corresponden a la aplicación asociada a la tarjeta RF (denominada desde ahora como aplicación RF) y una plataforma de sitio web asociada. En base a esto, las medidas de desempeño relevantes dentro del negocio es la cantidad de usuarios que han ingresado a la aplicación y sitio web de la empresa RF, la cantidad de clientes que ha realizado pagos por los canales digitales, el porcentaje del total de clientes que actualmente son digitales y la cantidad total de clientes digitales.

Las últimas dos métricas mencionadas aún no se han definido y serán parte del trabajo de memoria a realizar, pero es relevante mencionar en esta sección los canales presentes hacia los clientes que poseen la tarjeta y también a la cantidad de clientes digitales como un indicador relevante para el negocio financiero de la empresa RF.

3 Descripción del proyecto y justificación

En esta sección se incluyen características principales del proyecto a realizar, junto con las diversas razones y beneficios que genera la ejecución de este trabajo, hacia la compañía asociada.

El proyecto incluye la definición y el análisis de los clientes digitales que poseen la tarjeta RF, en base al análisis de los datos que actualmente posee la compañía. Dentro de estos datos se consideran las acciones que realizan los clientes en los canales digitales, que corresponden a la aplicación y la plataforma de RF. Los datos utilizados involucran otras acciones de los clientes con su tarjeta, como por ejemplo acciones en E-Commerce externos, pero asociados a la tarjeta RF, montos de transacciones y pagos con la tarjeta RF, datos demográficos, entre otros.

En base a esto, primero se realiza una segmentación de clientes, en base a las acciones digitales con la tarjeta RF, que permite entregar una agrupación de los clientes que posee la compañía. Con una caracterización de cada segmento resultante, se pretenden construir criterios que definen a un cliente digital y a un cliente no digital. Finalmente, esta primera parte incluye una asignación de cada cliente al grupo de clientes digitales o no digitales, según los criterios que se definan en el paso anterior.

Por otra parte, dada la asignación de cada cliente a su estado de digitalización, el trabajo de Memoria incluye un análisis de los clientes digitales, enfocado principalmente a estimar el cambio en transacciones y pagos que conlleva el proceso de digitalización de un cliente. Es importante destacar que se requiere abordar transacciones y pagos de los clientes, debido a la relevancia para el negocio de que un cliente realice sus compras con la tarjeta RF y asimismo, que este cliente logre pagar sus cuotas de las compras realizadas anteriormente.

Finalmente, se desea estimar un modelo de propensión para conocer qué grupo de clientes son los más propensos a un proceso de digitalización, para considerarlo en las campañas de marketing dirigido que realiza la empresa hoy en día.

Dentro de las causas que motivan el trabajo por realizar, se destaca en primer lugar la disponibilidad de datos sobre las características que poseen los clientes digitales de la empresa RF y, por consiguiente, la oportunidad de extracción de conocimiento en base a estos datos que posee la empresa.

Por otro lado, no existe una cuantificación precisa del beneficio asociado al digitalizar un cliente, por lo que se agrega como otra causa del trabajo de Memoria a realizar. Esto se produce porque no se conoce un beneficio en los ingresos que otorga el proceso de digitalización, en cuanto a transacciones y pagos mensuales. Un dato que se conoce, desde la empresa, es que actualmente se estima una diferencia en

cuanto a costos de comisión, en proporción 1 a 9 entre costos de transacción digital y no digital, respectivamente, lo que quiere decir que las transacciones digitales son más convenientes que las transacciones no digitales en cuanto a costo.

La tercera causa que motiva el presente trabajo es la necesidad de focalizar las campañas de marketing y tener un indicador que permita saber a quién mandarle campañas asociadas con medios digitales o qué campañas enviar a los clientes de la tarjeta RF.

Como consecuencias del presente trabajo, la realización del proyecto descrito obtendrá un perfilamiento de clientes con una caracterización de cada segmento resultante, junto con una cuantificación de beneficio o costo en transacciones y pagos con la tarjeta RF y también con una predicción de digitalización de cada cliente. Así se pretende proporcionar valor sobre la información de los clientes que permitan mejorar las campañas de marketing digital.

De esta forma, la oportunidad para la empresa se basa en evaluar si los medios digitales están cumpliendo un beneficio en los indicadores de transacciones y pagos hacia la compañía, o también se puede saber si es una buena estrategia enfocar el uso de los canales digitales para todos los clientes o sólo para algunos. Los potenciales beneficiarios con este trabajo son las subgerencias de Business Intelligence y de Banca Digital, las cuales fueron descritas anteriormente.

Dentro de las métricas relevantes que se pueden mejorar en el área de retail financiero, se destacan un posible aumento de la cantidad de clientes digitales, un incremento en el gasto promedio, por cada cliente digital y también acrecentar el uso de los medios digitales para realizar las transacciones y pagos asociados al servicio de tarjeta de crédito.

Con relación a estas métricas y al trabajo por realizar, se pretende que el modelo de segmentación de clientes, junto con el modelo predictivo, puedan mejorar las campañas de marketing para aumentar en un 10% la cantidad de clientes digitales, usando la información de este trabajo de memoria. Esto se debe considerar cuando aún no existen cifras precisas sobre la cantidad de clientes digitales y tampoco se saben las condiciones que deben cumplir éstos para convertirse en clientes digitales. Por otro lado, en relación a las métricas, se recalca que no se tiene una estimación exacta del aumento monetario que genera una transformación de un cliente no digital a ser un cliente digital. Por ello, este trabajo de Memoria busca una primera cifra que refleje cambios en transacciones y pagos, dada la digitalización de los clientes.

4 Objetivos

A continuación, se enuncian los objetivos que busca el presente trabajo de Memoria:

4.1 Objetivo General

El objetivo General que persigue este trabajo de Memoria es el siguiente:

"Definir y analizar a los clientes digitales de la tarjeta RF, para sugerir mejoras en las campañas de marketing dirigido existentes"

Dado este objetivo por cumplir, se espera que éste se cumpla a través de distintos objetivos específicos, los cuales se describen en la siguiente sección.

4.2 Objetivos Específicos

En relación con el objetivo general propuesto, los objetivos específicos que busca este trabajo de Memoria son los siguientes:

- Establecer hitos digitales que pueden cumplir los clientes de la tarjeta RF.
- Establecer diferentes grupos de los clientes de la tarjeta RF, en base a su adopción de hitos digitales.
- Definir los criterios que requiere cumplir un cliente para ser considerado digital o no digital.
- Evaluar el efecto en transacciones y pagos generado por una transformación de un cliente no digital a un cliente digital.
- Estimar grado de propensión de la cartera de clientes no digitales de la tarjeta RF, para transformarse en clientes digitales.
- Generar lineamientos para un plan de digitalización de los clientes de la tarjeta RF.

5 Marco teórico

En este apartado se explican los conceptos previos que aborda el presente trabajo y que son relevantes, para comprender posteriormente la explicación y el desarrollo de la metodología de trabajo que compone esta Memoria.

5.1 Metodología CRISP-DM

Dentro del marco conceptual que abarca el presente trabajo de Memoria, se destacan primordialmente conceptos y temáticas asociadas al Data Science, dada la disponibilidad de datos de clientes en la compañía y la necesidad de extraer información útil en base a los registros que posee la empresa.

Dentro del primer tópico a abarcar en esta sección, se destaca la presencia de una temática relevante dentro de la carrera de Ingeniería Civil Industrial, que corresponde a Business Intelligence, área que se ha vuelto muy relevante en el último tiempo debido a la mayor facilidad para la obtención y procesamiento de los datos gracias al avance de la tecnología. Dado esto, es relevante para la industria que un Data Scientist logre una programación correcta de los algoritmos necesarios y que pueda comunicar correctamente una historia asociada con los datos, hacia cualquier interesado en esta actividad [8]. El trabajo de Memoria a realizar se incluye dentro de este marco, dada la existencia de fuentes de datos en la compañía, que son aptas para realizar técnicas de Data Science.

En cuanto a los conceptos específicos por abordar en el marco conceptual, se destaca una de las áreas más relevantes que componen Data Science, que es Machine Learning. Ésta corresponde al uso de algoritmos para realizar modelos de análisis en base a los datos y es uno de los tantos tópicos que abarcan a Data Science [9].

Además, se introducen algunas metodologías de trabajo asociadas a Data Science y particularmente, al proceso de Minería de Datos. El primero corresponde a la metodología KDD (Knowledge Discovery in Databases), la cual es la más nombrada dentro de esta rama. Es un proceso que establece diversos pasos para obtener conocimiento a partir de la Minería de Datos, tal como lo muestra la siguiente ilustración.

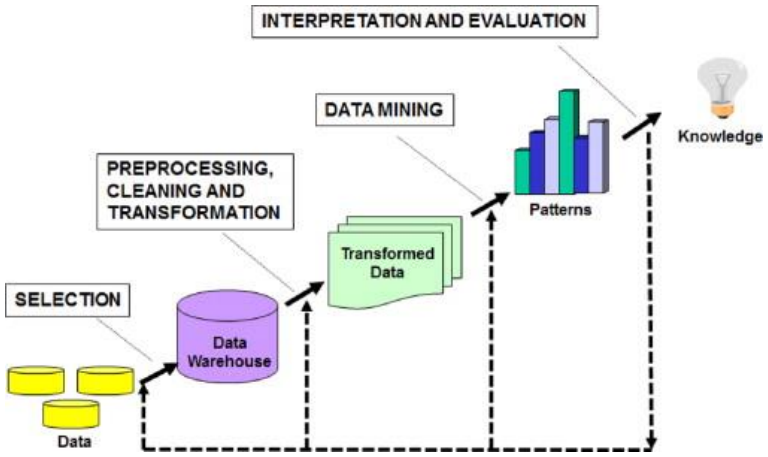


Ilustración 3: Pasos de metodología KDD

El segundo de ellos corresponde a la metodología CRISP (Cross Industry Standard Process for Data Mining), la cual se puede tomar como guía para estructurar el plan de trabajo por llevar a cabo durante la memoria.

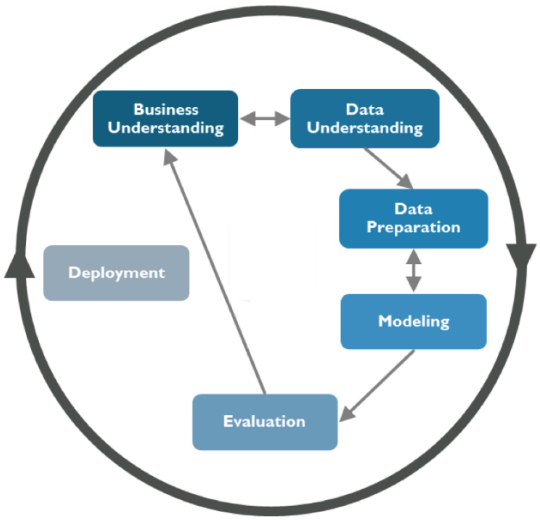


Ilustración 4: Pasos de metodología CRISP

Estas metodologías son las más utilizadas dentro de los procesos de Minería de Datos y de Machine Learning, bajo el cual se puede cumplir con un objetivo de extraer conocimiento y a diferencia del modelo KDD, sugiere iterar sobre el proceso realizado para mejorar la información final obtenida.

Debido a este mecanismo de iteración y mejoramiento del ciclo de minería de Datos, para el presente trabajo se utilizará la metodología CRISP, mediante la cual se pretende llevar a cabo el cumplimiento de los objetivos presentados anteriormente.

5.2 Análisis supervisado y no supervisado en Data Science

Dentro del concepto de Machine Learning se desglosan dos grandes ramas de algoritmos que la componen. En primer lugar, se encuentra el análisis supervisado, el cual requiere de una etiqueta, clase o resultado asociada a una fila de datos y el análisis no supervisado, donde la fila de datos no contiene una etiqueta necesaria para su análisis [10]. Considerando esto, los algoritmos de análisis supervisado tienen un objetivo de predicción del resultado de un registro, o de explicación del efecto de una variable, dadas las condiciones que especifica la fila de datos. Por otro lado, los algoritmos de análisis no supervisado solo evalúan las observaciones de los datos, es decir, no hay un conocimiento a priori de una variable dependiente y no se incluyen datos de salida o etiquetados en el modelo. Algunos de estos algoritmos poseen un fin de segmentación, como es el caso de Clustering, método que se abordará a continuación.

5.2.1 Análisis no supervisado y segmentación de clientes digitales

Dadas estas definiciones, el primer objetivo por lograr se puede abordar utilizando técnicas de análisis no supervisado, usando clustering. Dentro de los distintos algoritmos de Clustering que se pueden utilizar, se destacan los siguientes:

- K-Means: Segmentación que dentro de su algoritmo establece centroides, en base a las distancias entre los diversos puntos de los datos y luego asigna cada dato al segmento que posee el centroide más cercano, optimizando la suma de las distancias de cada punto al centroide asignado.

- K-Modes: Método con una lógica similar a k-means, es un algoritmo útil para incluir variables categóricas, que se pueden representar como variables binarias (con valor 0 o 1). La diferencia que posee con respecto a k-means es que no se basa en medir distancias entre los datos, sino que observa disimilitudes entre los datos, en su resultado, asigna modos a las variables en vez de medias. Un modo es un vector que minimiza las disimilitudes entre los vectores y cada punto de los datos. Este algoritmo fue desarrollado por Z. Huang[11].

- K- Prototypes: Modelo con un algoritmo mixto entre K-Means y K-Modes, ya que para las variables numéricas realiza un proceso de K-Means, efectuando una asignación de medias (promedios) para las variables y para las variables categóricas asigna modos. Este método también fue desarrollado por Z. Huang., como un complemento al modelo K-Modes [12].

En base a esto, se establece el modelo K-Prototypes como el indicado para realizar el procedimiento de segmentación, dada la naturaleza de los datos, con variables numéricas y categóricas, que se detallará en las siguientes secciones. Frente a esto el resultado deseado sería obtener un perfilamiento de clientes, utilizando los datos que posee la empresa sobre sus clientes asociados.

También se añade el concepto de Silhouette Score, o coeficiente de silueta, el cual es una métrica útil para evaluar un modelo de segmentación. Para cada punto, este score pondera la distancia promedio entre puntos intracluster, con la distancia media hacia elementos de otros segmentos. El coeficiente de silueta es la suma de este puntaje, para cada punto, su fórmula es:

$$\boldsymbol{Silhouette\ Score} = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \mathbf{s(i)}$$

$$\boldsymbol{s(i)} = \frac{\boldsymbol{b(i)} - \boldsymbol{a(i)}}{\boldsymbol{Max\{a(i), b(i)\}}}$$

b(i): Separación, mide distancia media del punto hacia otros clusters.
a(i): Cohesión, mide distancia media del punto hacia otros puntos del mismo cluster.

Fórmula 1: Silhouette Score

Mientras más cercano a 1, el score refleja un mejor desempeño del modelo de segmentación. Usualmente, se puede escoger un número óptimo de segmentos de un modelo [13], donde a cada valor de clusters de un algoritmo se le puede asociar un score, en donde el mayor valor de éste indica cuál es el número óptimo de clusters.

En cuanto a trabajos similares, asociados a la segmentación de clientes en banca, se destacan las labores realizadas por consultoras como Deloitte, McKinsey o Cognizant. Deloitte realizó una encuesta para observar el comportamiento de clientes online, generando 8 arquetipos de clientes en banca online [14], cuyo detalle se observa en el Anexo 1. Por su parte, McKinsey generó 5 perfiles de clientes de banca online, en un artículo que habla sobre retail financiero y su futuro [15], el detalle de los segmentos se ubica en la imagen del Anexo 2. Además, Cognizant, genera una segmentación de clientes de banca en base a su comportamiento por dispositivos móviles, diferenciándolos por edad e ingreso [16].

5.2.2 Análisis supervisado y modelos de propensión

Por otro lado, se identifica como un punto por abordar para el quinto objetivo propuesto, la realización de modelos predictivos para la digitalización de los clientes,

en base a su comportamiento con la tarjeta de crédito, en los canales digitales y con sus características demográficas. En base a esto, es importante tener dentro de los datos disponibles una etiqueta asociada a una base de datos, para lograr entrenar y ajustar un modelo que sea correcto. Considerando lo anterior, se pretende realizar un modelo de propensión que permita predecir qué clientes son más susceptibles a ser considerados clientes digitales. Dentro de esta tarea por abordar, se describen los algoritmos más utilizados para hacer análisis supervisado [17], [18].

- Regresión Logística: Este algoritmo permite ajustar una variable dependiente mediante el ajuste de coeficientes asociados a las variables independientes que se tienen en los datos. Dentro de sus ventajas se destaca la capacidad de interpretar los resultados de predicción del modelo a través del resultado de los coeficientes respectivos de cada variable.

- Árboles de Decisión: Permite predecir una variable, a través de distintas divisiones de los datos de variables predictoras, donde su resultado se puede representar a través de un árbol binario con nodos de decisión y nodos finales con el resultado de predicción.

- Random Forest: Algoritmo que es entrenado con la construcción de varios árboles de decisión, tiene su ventaja en evitar el sobreajuste que puede generar sólo un árbol de decisión, a través de la utilización de muchos árboles posibles.

- Naive Bayes: Método que estima los valores de la variable independiente utilizando el teorema de Bayes, con probabilidades condicionales de ocurrencia. Frente a esto el algoritmo asume la independencia de las variables dependientes para que sea válido esta utilización de probabilidades condicionales de Bayes. Tiene como ventaja la eficiencia computacional frente a bases de datos de gran tamaño.

- Support Vector Machine: Este algoritmo permite dividir la clasificación de los datos mediante vectores lineales o no lineales, según el parámetro que se fije en el modelo. Dado esto, es posible no hacer una separación perfecta de los datos, por lo que se calcula una función de costo que es minimizada para entrenar el modelo y entregar el mejor ajuste.

- KNN (K Nearest Neighbors): En un modelo que permite predecir una variable dependiente, considerando el resultado que generan datos similares, denominados vecinos cercanos. El algoritmo puede modificar la cantidad de vecinos cercanos a considerar para estimar una predicción de la variable dependiente.

Además, se introducen algunos conceptos asociados a los modelos de clasificación, los cuáles se utilizarán durante este Trabajo de Memoria.

1. Métodos Min-Max de normalización:

Esta técnica se utiliza para análisis supervisado y no supervisado de datos, es útil para tener una misma escala en todas las variables y evitar que existan variables con un rango muy grande y otro muy pequeño. Para modelos de segmentación, esta técnica es necesaria para poder comparar las variables y medir una distancia entre los puntos que sea coherente en la práctica. Para modelos predictivos, esta técnica es recomendable si se desean comparar coeficientes de regresiones, entre distintas variables.

Para cada dato de la variable se aplica la siguiente fórmula:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Fórmula 2: Normalización Min-Max

Con esta transformación, todas las variables tienen un mínimo de 0 y un máximo de 1.

2. Métodos de selección de variables:

Son técnicas previas a la generación de algoritmos para la base de datos y la realización del modelo de predicción. Estas técnicas permiten reducir dimensionalidad, la cual es útil para hacer un modelo con menor costo computacional. Las 3 técnicas de selección de variables a considerar para este trabajo son:

- Correlación con variable dependiente: Este método busca eliminar variables que no estén correlacionadas con la variable independiente a predecir [19].
- Correlación entre variables: El segundo método también es abordado en el mismo paper del método anterior, sostiene que la correlación entre variables no es aconsejable para los modelos de Machine Learning. Este fenómeno de variables correlacionadas se denomina como Multicolinealidad [20] y se soluciona eliminando variables hasta que no existan variables correlacionadas entre sí. Es importante destacar que este problema es más asociado con la econometría, más que con los modelos predictivos.
- Información Mutua: Mide la dependencia mutua entre dos variables, donde se pretende reducir la incertidumbre de una variable aleatoria Y,

considerando la entropía de una variable conocida X [21]. En base a esta definición, se instala un algoritmo para calcular ganancia de información en bases de datos, para guiar modelos de Machine Learning [22].

3. Oversampling:

Este método es indicado para bases de datos desbalanceadas, en un modelo de clasificación. Con desbalanceo se refiere a distintas proporciones de valores, en la etiqueta de la variable a predecir, donde una de ellas predomina con una mayor proporción. En particular, una técnica de oversampling a considerar para el presente trabajo es SMOTE [23], la cual genera datos sintéticos, con valores similares a los originales, para los datos presentes en la clase minoritaria. Así, se genera un balanceo que puede mejorar las métricas de desempeño de un modelo de clasificación.

4. Métricas de desempeño de modelos de clasificación:

Se consideran las siguientes métricas que miden la performance de un modelo predictivo:

- Accuracy: Mide qué proporción del total de datos fue predicho correctamente. Su fórmula es:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{FP} + \boldsymbol{TN} + \boldsymbol{FN}}$$

Fórmula 3: Métrica Accuracy

** TP son predicciones positivas y correctas, TN son predicciones negativas y correctas, FP son falsos positivos, es decir, predicciones positivas en datos que son negativos y FN son predicciones negativas e incorrectas.*

- Precision: Mide qué proporción de las predicciones positivas son realizadas correctamente. Su fórmula es:

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}}$$

Fórmula 4: Métrica Precision

- Recall: Mide qué porcentaje de los datos realmente positivos, son predichos correctamente. Su fórmula es:

$$\boldsymbol{Recall = \frac{TP}{TP + FN}}$$

Fórmula 5: Métrica Recall

- F1: Involucra los valores de Precision y Recall y pondera ambos indicadores. Su fórmula es:

$$\boldsymbol{F1 = 2 * \frac{Precision * Recall}{Precision + Recall}}$$

Fórmula 6: Métrica F1

- AUC: Mide el área bajo la curva ROC de un modelo predictivo. La curva ROC posee como ejes las tasas de verdaderos positivos y falsos positivos de un modelo y genera una recta al modificar cada una de estas tasas.

Para finalizar con los aspectos asociados a Data Science, se destaca como una herramienta muy útil para el retail y es un campo relevante, debido a la disponibilidad de grandes volúmenes de datos y la posibilidad de lograr insights hacia los objetivos de negocio, para las empresas de esta industria. Desde allí, se identifican variadas investigaciones asociadas que abordan el análisis de datos hacia el retail y con objetivos específicos como la segmentación de clientes [24] o la predicción asociada a clientes, por ejemplo, propensión de default, propensión de fuga o predicción de un monto de gasto [25]. Por otra parte, existe un punto a favor que es la existencia de canales online para abordar las técnicas de Data Mining hacia clientes de canales online [26].

Dados estos puntos, se pretende abordar un marco conceptual que logre planificar correctamente la metodología de trabajo y de esta forma, poder cumplir con los objetivos necesarios del trabajo de Memoria.

5.3 Estimación de Diferencias en Diferencias

La estimación de Diferencias en Diferencias (abreviada como DID o DD, como sigla en inglés), pertenece a un grupo de modelos que permite estimar el efecto de un tratamiento o evento que sucede en un grupo. Este modelo fue desarrollado por Card y Krueger (1994) [27], quienes desarrollaron una estimación del efecto de la tasa de empleo en Nueva Jersey, generado por un aumento del sueldo mínimo, utilizando los datos del caso de Pennsylvania, que efectuó un aumento del sueldo mínimo. Además de este caso, el modelo DID se utiliza frecuentemente para el estudio de experimentos realizados en las áreas de la educación, salud, trabajo, entre otros.

En cuanto a las características que cumple este modelo, las cuales permiten evaluar el efecto de un tratamiento o evento, el modelo de diferencias en diferencias corresponde a una regresión que permite estimar una variable dependiente considerando dos grupos, el grupo de control y de tratamiento.

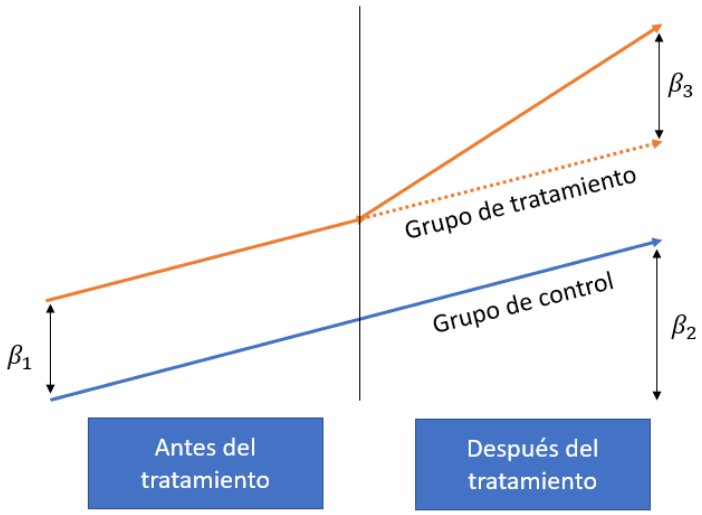


Ilustración 5: Diagrama Diff in Diff

Al grupo de tratamiento se realiza el experimento a estudiar, el cual puede ser una prueba de un medicamento o un test educativo; al grupo de control no se altera durante el experimento. Dado esto, se utiliza la regresión DID para estimar estas dos rectas, generando coeficientes asociados a las variables D y T:

$$Y_{it} = \alpha + \beta_1 D + \beta_2 T + \beta_3 DT + \varepsilon_{it}$$

Fórmula 7: Regresión DID

- D corresponde a la variable binaria que es 1 si el individuo el grupo de tratamiento, y cero si pertenece al grupo de control.
- T corresponde a la variable binaria que es 1 si el experimento ya fue realizado, o tiene el valor cero si se encuentra en un instante antes del experimento.
- DT corresponde a la variable binaria que es 1 si D y T equivalen a 1, por lo que permiten evaluar un cambio en la variable independiente, dada la realización del experimento hacia el grupo de tratamiento.

Con estas variables, se construye la ecuación presentada anteriormente, donde **α** es el coeficiente que permite estimar un aumento en la variable dependiente, dada la ocurrencia del experimento hacia el grupo de tratamiento. Así se logra estimar el efecto del experimento/evento y se espera que sea significativamente distinto de cero.

En adición a lo anterior, al modelo propuesto se le pueden incorporar otras variables "contaminantes" que no considera el primer modelo, como por ejemplo la edad, género, o salario, que permiten esclarecer de mejor forma si el efecto asociado a la variable DT es efectivamente positivo o no. Considerando esto, el modelo de regresión final que se considerará para el trabajo de Memoria es el siguiente.

$$Y_{it} = \alpha + \beta_1 D + \beta_2 T + \beta_3 DT + \beta_4 X' + \varepsilon_{it}$$

Fórmula 8: Regresión DID, agregando variables "contaminantes"

Además, a este modelo se le pueden incorporar variables que se vayan modificando durante la ventana de tiempo que se desea estudiar (por ejemplo, edad), lo que genera un modelo más complejo que el mostrado anteriormente. Sin embargo, este modelo no se considerará para el presente trabajo.

Finalmente, es relevante considerar el supuesto de tendencias paralelas [28] para realizar un modelo de diferencias en diferencias, es decir, se asume que el grupo de tratamiento tiene la misma tendencia del grupo de control, si es que no se realiza el experimento. Con esto es posible concluir que el coeficiente de la variable de interacción DT, corresponde a la estimación del efecto del experimento hacia el grupo de tratamiento.

6 Metodología

A continuación, se describen los diversos pasos que componen la metodología de trabajo. Dentro de ésta, se considera una metodología mixta compuesta por los pasos de KDD, agregando desde la metodología CRISP, un entendimiento básico del negocio financiero y los conceptos asociados a los canales digitales, los cuáles permiten tener un conocimiento previo que ayude a realizar un trabajo coherente, para cumplir con los objetivos deseados.

6.1 Selección y preparación de datos

Como paso previo al análisis de los datos, es primordial ejecutar los pasos de selección, limpieza y transformación de datos de forma correcta, para asegurar la calidad de un modelo y que éste sea representativo y confiable.

En primer lugar, se definen las bases de datos presentes en la compañía, donde se consideran las siguientes fuentes:

- Clientes: Datos de clientes por cada mes, se consideran los clientes con tarjeta activa, es decir que no tenga bloqueos y que pueda utilizarla en los comercios asociados, actualizados al mes de julio de 2019.

- Datos sociodemográficos: La compañía posee datos de sus clientes como edad, género, nacionalidad y ciudad en la que vive.

- Transacciones: Se tienen datos de transacciones de los clientes con la tarjeta RF, ya sea en tienda o por medio de E-Commerce, junto con su monto, fecha y lugar donde realizó la transacción.

- Pagos: Se tienen datos de los pagos de cuotas asociadas a las transacciones efectuadas con la tarjeta RF.

- Datos de navegación web: Esta base de datos engloba las acciones realizadas por los usuarios, en la aplicación y el sitio web de la compañía. Estas acciones abarcan el acceso a diversas secciones, como por ejemplo, la revisión del estado de cuenta, revisión de movimientos, ofertas de diversos productos, pago de cuotas (acción solo permitida en aplicación), entre otras.

- También se cuentan con otras variables, como el cupo total con la tarjeta, la deuda que posee un cliente en cada mes y el canal por el cual solicitaron la tarjeta RF.

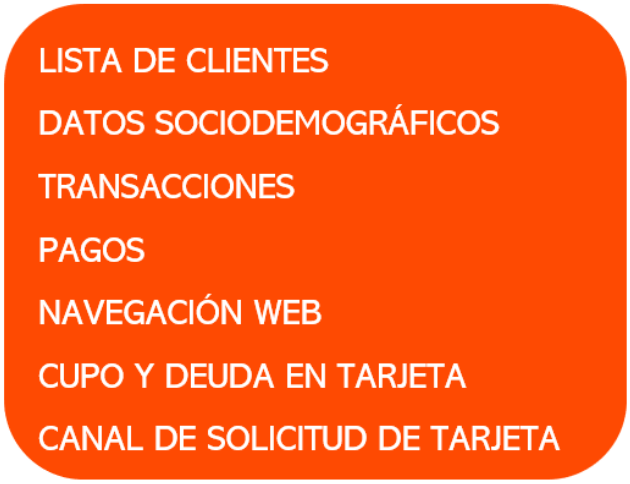


Ilustración 6: Bases de datos disponibles para el análisis de clientes digitales

Dada la extracción correspondiente, se construye una base de datos que consolida los campos presentes en las 5 bases de datos explicadas anteriormente. Además, en base a diversas columnas de esta nueva base de datos, se agregan variables binarias para marcar los hitos del cliente, los cuales se explicarán posteriormente, en la sección de Desarrollo de la Metodología.

Ya construidas las bases de datos, se realizan diversas técnicas de preprocesamiento para las etapas que componen el trabajo de Memoria. Para la primera parte de segmentación, se decide efectuar una transformación logarítmica a algunas variables, para que no tengan tantos valores cercanos al mínimo, que en este caso es cero. Por otro lado, se realiza la eliminación de filas con outliers o valores fuera de rango, para que eviten alterar los resultados del modelo. Como paso final de la primera etapa de segmentación, se realiza la transformación Min-Max, la cual permite que los valores de los datos del modelo estén entre 0 y 1, logrando así resultados no sesgados en las variables que poseen mayor varianza.

6.2 Primera Etapa: Segmentación de clientes y definición de cliente digital

En base a la cartera de clientes de la tarjeta RF y los datos mencionados anteriormente, se realiza la primera etapa de la metodología, la cual busca definir los criterios que cumple un cliente para ser considerado un cliente digital.

En relación a esto, esta primera etapa contiene diversas subetapas. En primer lugar, se busca identificar hitos digitales de los clientes en base a los datos disponibles, para evaluar el comportamiento de estos clientes en medios digitales y evaluar la cantidad de hitos que cumplen.

Dado esto, en conjunto con las Áreas de Business Intelligence y de Banca Digital, se definen los siguientes 10 hitos digitales, que corresponden en su mayoría a acciones que se pueden realizar en la aplicación o en el sitio web asociado a la compañía de retail financiero. Es importante destacar que estas acciones se identifican como variables binarias en la base de datos correspondiente a la primera etapa de segmentación:

Hito 1: Login en Aplicación RF
Hito 2: Login en sitio web RF
Hito 3: Revisión de estado de cuenta
Hito 4: Revisión de movimientos
Hito 5: Revisión de cupo disponible
Hito 6: Revisión de ofertas (sólo disponible en aplicación)
Hito 7: Pagos de cuenta (sólo disponible en aplicación)
Hito 8: Efectuar transacciones en E-Commerce
Hito 9: Efectuar transacciones en E-Commerce asociado a la empresa RF
Hito 10: Efectuar pagos en E-Commerce

En la siguiente ilustración se realiza un muestreo de 3.642.648 acciones de clientes, identificadas en app o sitio web, de las cuáles se identifica qué porcentaje de ellas son hitos digitales. Con esta ilustración se muestra que los hitos digitales cubren gran parte de las acciones totales en app y sitio web.

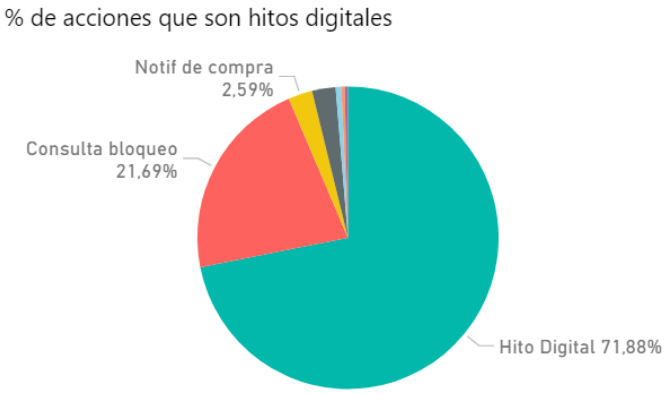


Ilustración 7: Proporción de hitos digitales con respecto al total de acciones

Es importante destacar que los hitos establecidos no corresponden al total de acciones que se puede realizar en la app, en particular, los hitos tienen características que las diferencian de una simple acción digital. En primer lugar, son acciones más frecuentes que las acciones que no son hitos digitales. En segundo lugar, corresponden a acciones que son directamente asociadas a transacciones, pagos o revisión del estado financiero, a diferencia de acciones que no son hitos digitales (por ej. una consulta de bloqueo de tarjeta), que son complementarias y no asociadas a transacciones y pagos.

Por otro lado, se considera una ventana de tiempo fija para realizar el análisis de hitos e identificar la cantidad de hitos que cumple el cliente durante ese lapso. Para el caso de la primera etapa de segmentación, se consideró una ventana de tiempo de 6 meses, desde febrero a julio de 2019. La ventana posee esa duración porque el sitio web RF fue creado a comienzos de este año y la aplicación RF fue creado en 2018, lo que impide realizar un análisis con mayor extensión de tiempo.

Luego, se realiza un proceso no supervisado de los datos, construyendo un modelo de segmentación que permita encontrar grupos de clientes, en base a los hitos digitales que cumplen, considerando la ventana de tiempo definida y también considerando el análisis exploratorio de los datos.

Finalmente, el análisis exploratorio y la segmentación se efectúan para definir los criterios que hacen digital o no a un cliente, para realizar una asignación que permita analizar a los clientes digitales y ver sus características relevantes para enfocar las campañas de marketing existentes. Es importante considerar que la definición de criterios que influyen en la digitalización de un cliente, se realizará en base a juicio experto, con ayuda de los resultados de análisis exploratorio y segmentación, por ende, esta decisión es un alcance del trabajo de Memoria, como se explicará posteriormente.

6.3 Segunda etapa: Análisis de transacciones y pagos de clientes digitales

La segunda etapa que conforma el presente trabajo corresponde a un análisis asociado a los clientes digitales y no digitales que poseen la tarjeta RF en la compañía. En base a esto, primero se pretende realizar un análisis del cambio que puede producir un proceso de digitalización de un cliente en los indicadores de transacciones y pagos.

Como primer paso de esta etapa, se realiza un análisis exploratorio asociado a los clientes digitales existentes, utilizando la asignación de cada cliente al grupo de clientes digitales o no digitales. En base a esto, se pretende observar diferencias del grupo de clientes digitales en base a variables demográficas, como edad, sexo, nacionalidad o estado civil, además de observar diferencias en los montos de transacción y pagos, con respecto al grupo de clientes no digitales. Frente a esto, se pretende tener ideas de un comportamiento distinto de un cliente al terminar un proceso de digitalización.

Luego, se realiza un análisis de transacciones y pagos de los clientes digitales, buscando diferencias con respecto a estos mismos clientes, cuando no eran digitales. En base a esto, se elabora un modelo de diferencias en diferencias, que evalúe si existe un efecto positivo en la nueva rentabilidad, dado el evento de digitalización de un cliente de la tarjeta RF. Para esto se deben considerar como datos el monto de transacciones del cliente digital antes de realizar el primer hito digital, el monto de transacciones de un cliente digital después de convertirse en un cliente digital. También se deben captar el nivel de pagos promedio de un cliente no digital al comienzo de la ventana de tiempo y al final de la ventana de tiempo. Se escogerá para esta caso una ventana de tiempo de 11 meses.

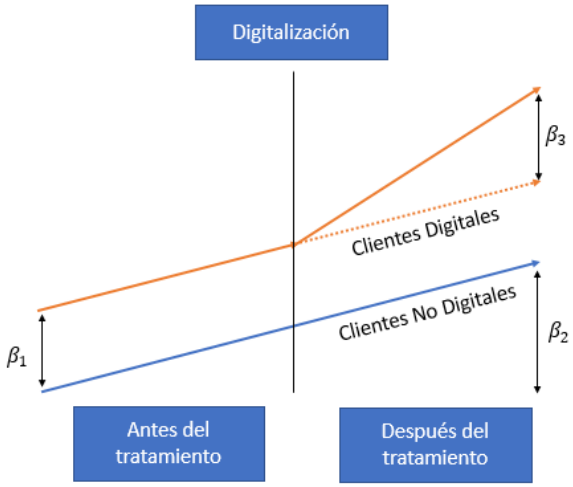


Ilustración 8: Diagrama Diff-in-Diff para el caso por aplicar

6.4 Tercera etapa: Estimación de propensión de digitalización de clientes

Finalmente, se realiza un modelo predictivo para calcular una probabilidad de propensión de un cliente no digital, para transformarse en cliente digital. Para esto, se utilizan variables demográficas y de comportamiento de uso con la tarjeta, en cuanto a transacciones y pagos con ésta.

Se construirán dos cálculos de propensión:

- El primer modelo considerará un cálculo de la probabilidad de digitalizarse en el siguiente mes, dado que el cliente no ha realizado ningún hito digital. Ésto se elaborará con el objetivo de incentivar el uso de canales digitales para nuevos clientes o clientes que aún no realizan sus acciones por medios online y así incentivar la adquisición de clientes digitales.

- El segundo modelo se realizará hacia los clientes que han cumplido hitos digitales, pero que aún no son considerados clientes digitales, ya que no usan frecuentemente los canales digitales. Este modelo tiene el objetivo de aumentar la conversión de clientes digitales y no sólo aumentar el uso de los canales online de la compañía, sino que guiar estas acciones para generar mayor cantidad de pagos y transacciones, para así aumentar la rentabilidad asociada a cada cliente.

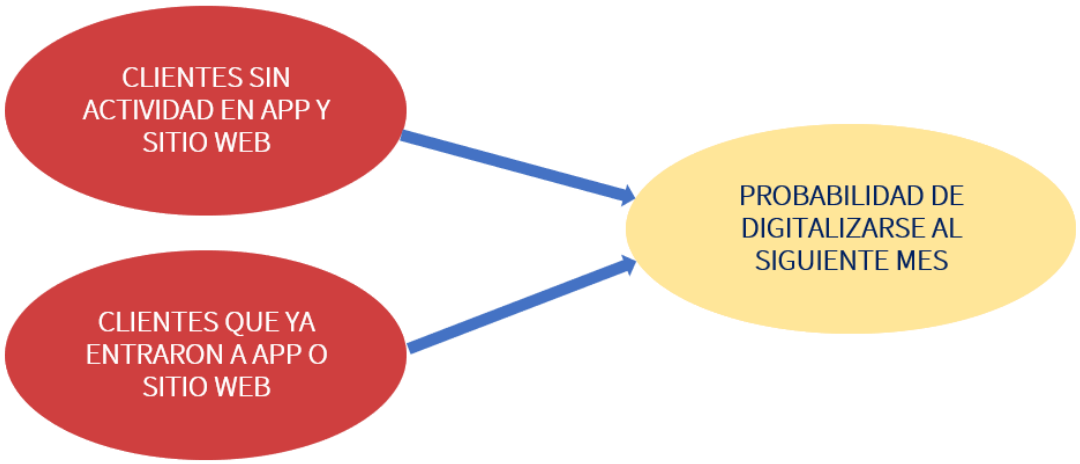


Ilustración 9: Grupos a considerar para modelo predictivo

Los modelos por utilizar son Logit, Árboles de Decisión, Random Forest, Gaussian Naive Bayes, KNN y Support Vector Machine. Frente a esto, se entrenan los datos y se calculan métricas de precisión para todos estos modelos, buscando un modelo que permita predecir lo mejor posible qué clientes se digitalizarán próximamente.

Dentro del proceso de modelos predictivos, se realiza una etapa de selección de variables, antes de ejecutar el algoritmo de predicción, Con esto se busca reducir dimensionalidad del modelo y para ello se utiliza un método filtro de 3 pasos, los cuáles son:

- Análisis de correlación de variables independientes, con la variable dependiente:
  Se escogen variables altamente correlacionadas con la variable dependiente. Este método tiene el objetivo de incluir variables que si tengan capacidad de predicción de la variable dependiente. Elimina variables con bajo peso predictivo y que pueden inducir a un modelo con alto costo computacional, considerando la gran cantidad de filas que pueden tener los datos escogidos.

- Análisis de correlación entre variables independientes:
  Se eligen variables que no tengan alta correlación entre sí. Con esto se eliminan las opciones de multicolinealidad del modelo. También esto permite reducir dimensionalidad y así disminuir los tiempos de ajuste del modelo.

- Análisis de ganancia de información entre variables independientes, frente a la variable dependiente:
  En este filtro se escogen variables con alta ganancia de información con respecto a la variable dependiente, eliminando atributos con bajo poder predictivo.

Frente a la aplicación de estos 3 métodos, se escoge para el modelo las variables que superan estos 3 filtros.

Por otra parte, si se da la situación de baja proporción de datos con etiqueta positiva, es decir, con su variable Y igual a 1, se aplicará el método de Oversampling “SMOTE”, sólo en la data de entrenamiento del modelo. Con esto se logra equiparar la base de datos y se evita que éste siempre haga predicciones a la clase mayoritaria, generando así mejores resultados en las métricas de los modelos.

Se utilizan diversas métricas de error para evaluar los modelos, como la curva AUC, los indicadores de Accuracy, Precision, Recall y F1. Sin embargo, para efectos de este trabajo, el Accuracy será la métrica clave que permitirá distinguir si un modelo tiene mejor desempeño que otro. Dado esto, se pretende tener como output final para el

modelo, la lista de clientes con su respectiva probabilidad de digitalizarse, dado el cálculo realizado con el mejor modelo estimado.

Frente a esto, el objetivo es recomendar hacia las campañas de marketing específicas, considerando los clientes más propensos a digitalizarse en el corto plazo.

7 Alcances y resultados esperados

En esta sección se mencionan los alcances asociados al presente trabajo, junto con los resultados que se pretenden obtener al desarrollar la metodología explicada anteriormente.

7.1 Alcances

Dados los objetivos solicitados por la empresa, sólo se considera en este trabajo la realización de los puntos destacados en la metodología. Por ende, se mencionan ciertos aspectos que quedan fuera del análisis propuesto dentro del trabajo de memoria, los cuáles se pueden abordar en las recomendaciones futuras del trabajo.

En primer lugar, se aborda como alcance la decisión específica de los criterios que digitalizan o no a un cliente. Si bien, el análisis exploratorio y la segmentación se realizará en el trabajo de Memoria, la definición de los criterios centrales de digitalización de clientes será ejecutada en conjunto con el área comercial de la empresa RF. Por lo tanto, esta decisión será especificada en la sección de desarrollo de la metodología, pero como un dato que servirá como input para los pasos siguientes del trabajo que corresponden a la asignación de digitalización para cada cliente y al análisis de transacciones y pagos de clientes digitales.

Por otro lado, no se abordarán predicciones complementarias a las especificadas en la tercera etapa de la metodología, dada la base de datos de clientes de la empresa. Este tipo de análisis pueden ser útiles con otros fines y algunos de ellos pueden ser ideas que surjan durante el trabajo de memoria, sin embargo, éstas no son viables de abordar en este trabajo para evitar sacar de foco el trabajo inicial y la modificación de objetivos y metodología planificada durante la Introducción al Trabajo de Título.

En adición, tampoco se abordarán recomendaciones hacia la construcción o modificación de las bases de datos existentes en la organización, estas labores si pueden ser ejecutados dentro de la relación alumno/empresa, pero no serán parte del trabajo de memoria ni del entregable final.

Finalmente, se decide no ejecutar aspectos específicos de las campañas de marketing digital. Se pretende entregar los hallazgos y recomendaciones del presente trabajo al área de Banca Digital de la compañía.

7.2 Resultados esperados

Los resultados esperados dentro del Trabajo de Memoria son los siguientes:

- Encontrar diversos perfiles de clientes en base a su comportamiento en los canales digitales, ya sean pagos en tiendas físicas/digitales, transacciones en tiendas físicas/digitales y frecuencia de realización de hitos digitales.
- Asociado a los diversos perfiles existentes, la lista de criterios de digitalización de un cliente se considera como otro entregable de este trabajo.
- Por otro lado, se establece como entregable un análisis de visualización que permita caracterizar a los clientes digitales, en base a sus características demográficas y en base a sus comportamientos de monto y frecuencia de compra en los comercios asociados a la tarjeta RF.
- Como cuarto entregable se tiene la cuantificación del cambio en transacciones y pagos con la tarjeta RF, que conlleva la transformación de un cliente no digital a digital.
- El quinto entregable corresponde a una predicción que logre estimar la probabilidad de un cliente del aumento del uso de la tarjeta RF, en los canales digitales. En particular, se pretende obtener las posibilidades de que un cliente no digital se transforme en digital.
- Finalmente, como sexto entregable se pretende entregar recomendaciones hacia las campañas de marketing que se realizan en la compañía, por medio de este mismo entregable, donde se pueda enlazar los resultados y conclusiones obtenidas desde este trabajo, en busca de un mejoramiento en las campañas que realiza actualmente la empresa de retail financiero.

8 Desarrollo de la metodología

En esta sección se muestra el desarrollo realizado en la metodología de trabajo, junto con sus resultados finales. Esta sección se divide de forma general en las tres secciones principales mencionadas en la metodología, la segmentación de clientes dadas sus acciones digitales; el análisis de transacciones y pagos, mediante el modelo de diferencias en diferencias y el desarrollo de modelos predictivos, que muestran la posibilidad de digitalización de los clientes que actualmente no son digitales.

8.1 Segmentación de clientes en base a acciones digitales

En esta etapa se efectúa una segmentación de los clientes que poseen la tarjeta RF, en base a las acciones que poseen en los canales digitales de la compañía. Frente a esto se pretenden obtener distintos perfiles de los clientes existentes y, además, un criterio que permita distinguir a los clientes digitales de los clientes no digitales, paso que será útil para las siguientes etapas del trabajo de Memoria.

8.1.1 Selección de datos

En cuanto a la selección de datos, la primera etapa es formar el universo de clientes para la etapa de segmentación, mediante diversos filtros de selección. Luego a cada cliente se asocia el estado de cumplimiento de hitos digitales, que considera una ventana de los últimos 6 meses, entre febrero y julio de 2019. Finalmente se escogen las variables asociadas a los hitos digitales, para cada cliente.

En los filtros de clientes se considera en primer lugar, excluir del análisis a clientes que posean su tarjeta con bloqueos. Además, sólo se consideran clientes que han realizado transacciones en los últimos 12 meses (entre agosto de 2018 y julio de 2019), para evitar considerar a clientes inactivos dentro del estudio. El siguiente diagrama muestra el universo total de clientes, considerado para la primera etapa de segmentación.

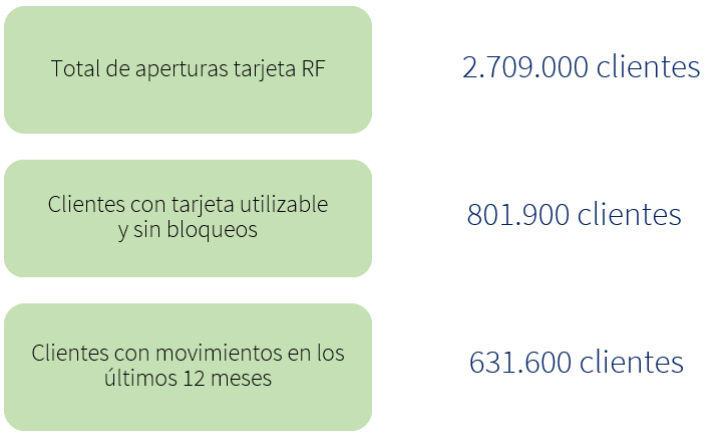


Ilustración 10: Filtro de universo de clientes, para etapa de segmentación

Con este total de clientes, el siguiente paso es mencionar las bases de datos existentes y que permitirán extraer las variables para el proceso de segmentación. En base a esto, la compañía posee bases de datos de transacciones y pagos de los clientes con la tarjeta RF, en tiendas propias y comercios asociados, solicitudes de tarjeta, datos demográficos de los clientes que poseen la tarjeta y registro de acciones en los canales digitales, es decir, en la aplicación y página web de RF.

Con los datos disponibles, se extraen las siguientes variables a considerar en la etapa de segmentación de clientes:

Variables binarias, asociadas a hitos:

1- Si el usuario realizó un login en aplicación de RF
2- Si el usuario realizó un login en sitio web de RF
3- Si el usuario realizó una revisión de cupo disponible, en app o sitio web
4- Si el usuario realizó una revisión del estado de cuenta, en app o sitio web
5- Si el usuario revisó los movimientos realizados, en app o en sitio web
6- Si el usuario revisó una oferta en la aplicación de RF
7- Si el usuario realizó un pago en la aplicación de RF
8- Si el usuario realizó una compra vía E-Commerce, en los últimos 6 meses
9- Si el usuario realizó una compra vía E-Commerce de la empresa RF, en los últimos 6 meses
10- Si el usuario realizó un pago vía E-Commerce, en los últimos 6 meses

Variables de frecuencia de acciones en app y sitio web, calculadas como promedio mensual:

1- Frecuencia de login en aplicación de RF
2- Frecuencia de login en sitio web de RF
3- Total de revisiones de cupo, en app o sitio web
4- Total de revisiones de estado de cuenta, en app o sitio web

5- Total de revisiones de movimientos, en app o sitio web
6- Total de revisiones de ofertas en aplicación
7- Total de pagos realizados en la aplicación

El valor de las variables binarias (1 o cero) dependen de si el cliente realizó la acción entre febrero y julio del presente año, considerando una ventana de 6 meses. Las frecuencias son el resultado del total de acciones entre febrero y julio, dividido en 6, lo que resulta en un promedio mensual de cada acción perteneciente a los canales digitales.

8.1.2 Análisis exploratorio de los datos

Además de las variables mencionadas anteriormente, se añaden las variables demográficas, de transacciones y pagos de cada cliente, para realizar un análisis exploratorio que permita entregar observaciones y deducciones de los datos disponibles.

En primer lugar, el análisis descriptivo más simple es visualizar los datos de clientes, en base a su edad, la nacionalidad que poseen y la antigüedad que tienen con respecto a la solicitud de la tarjeta RF. Estos gráficos asociados a estas 3 variables se muestran a continuación:

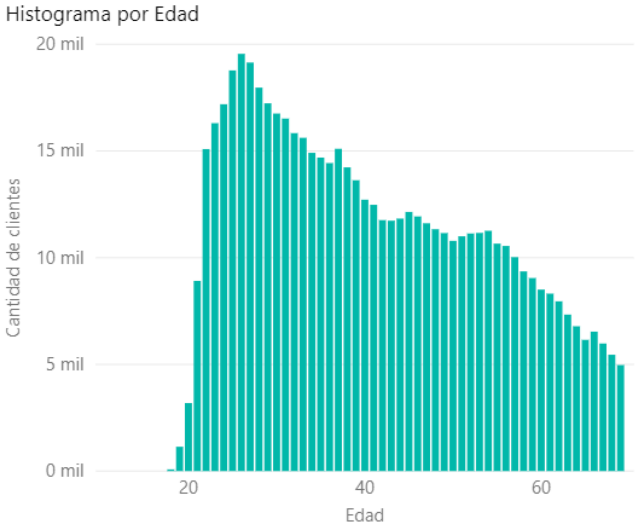


Ilustración 11: Histograma de cantidad de clientes por edad

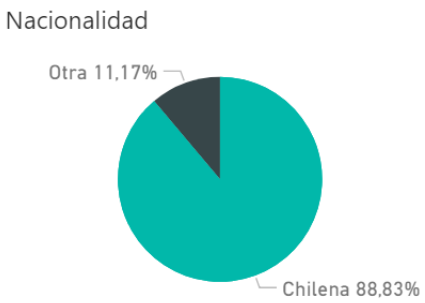


Ilustración 12: Nacionalidad de clientes de la tarjeta RF

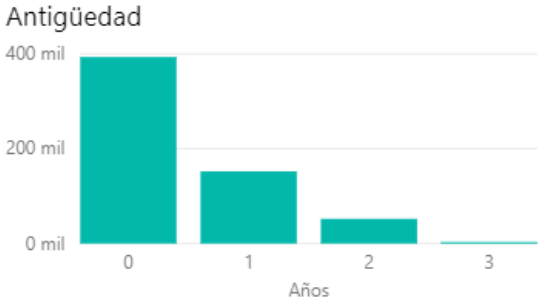


Ilustración 13: Antigüedad de clientes que poseen la tarjeta RF, en años

En este análisis se observa que muchos de los clientes son jóvenes, ya que la mayoría posee entre 22 y 40 años. Como otra observación, un 88% de los clientes poseen nacionalidad chilena, lo cual parece un resultado bajo con respecto a las cifras de inmigrantes que se encuentran habitando en Chile (en 2019, se estima que un 6,6% de la población es extranjera [29]). Por otro lado, una gran parte de los clientes fueron adquiridos en los últimos meses (el valor 0 del gráfico indica los clientes adquiridos entre mayo de 2018 y abril de 2019).

En adición a lo anterior, se analiza la variable de cantidad de hitos que cumplen los clientes, la cual tiene un rango de 0 a 10 hitos. Dado esto, el primer gráfico agrupa la cantidad de hitos promedio, según la edad que poseen los clientes y el segundo gráfico muestra el monto promedio de transacciones, agrupando a los clientes según los hitos digitales que poseen.

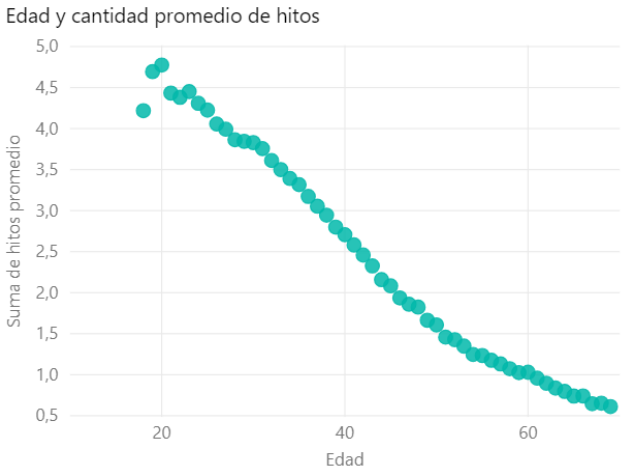


Ilustración 14: Cantidad promedio de hitos por edad

Ilustración 15: Monto promedio de transacciones (suma entre febrero y julio) por cantidad de hitos

Con estos gráficos, se deduce que el grupo con mayor promedio de hitos digitales se encuentra entre los 21 y 35 años, además, para personas mayores a 35 años, se observa una relación lineal inversa entre la edad y la cantidad de hitos que van cumpliendo. Este resultado podría explicarse si se piensa que, en general, la gente más joven es la más propensa a usar los canales digitales y por ende, más propensa a cumplir una mayor cantidad de hitos.

Por otro lado, al observar los montos de transacción promedio, se muestra al incrementar los hitos que posee el cliente, el promedio de monto de transacciones también va aumentando. Esto también es lógico si se considera que un mayor uso de canales debería generar mayor nivel de transacciones por parte de los clientes. No obstante, es útil verificar esto, considerando que posteriormente se debe analizar los ingresos que generan para la compañía, los clientes con mayor nivel de digitalización.

8.1.3 Preprocesamiento y tratamiento de datos

En este paso, se describen las técnicas utilizadas para el preprocesamiento realizado, hacia las variables del proceso de segmentación.

En primer lugar, se efectúa una transformación logarítmica a las variables de frecuencia de acciones, esto debido a que muchos valores son o se acercan a cero. Se pretende alisar el histograma de estas variables para que tengan un mejor

comportamiento en el proceso de segmentación. A la transformación logarítmica se le suma 1, para no indeterminar los valores originales iguales a 0. La siguiente imagen muestra un ejemplo de histograma de la variable de Login en la aplicación RF, junto con su transformación logarítmica asociada.

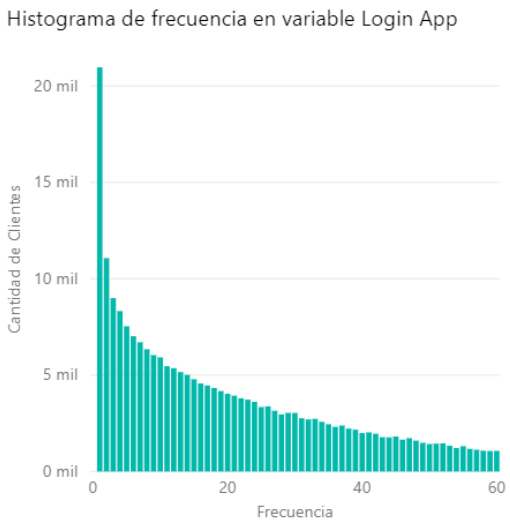


Ilustración 16: Cantidad de personas por frecuencia en variable "Login App"

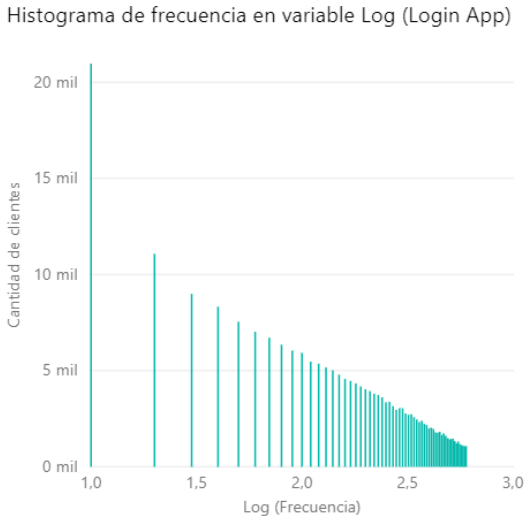


Ilustración 17: Cantidad de personas por frecuencia en variable "Logaritmo Login App"

Posteriormente, se realiza una identificación de outliers presentes en la base de datos, puntualmente en las variables de frecuencia que se pretenden llevar a la segmentación. Aquí se considera el criterio 2 Sigma, que elimina los valores superiores al valor de Media +2SD. Si bien, este criterio es el indicado para variables con distribución normal, en este caso es útil si se eliminan los valores muy altos y no se eliminan los valores cercanos a cero o que son cero.

Considerando lo anterior, se establecen los siguientes procesos de eliminación, asociados a las variables de frecuencia de acciones en canales digitales:

Variable	Criterio de eliminación
Logaritmo de frecuencia de login en aplicación RF	Valores superiores a 2,8
Logaritmo del total de revisiones de cupo, en app o sitio web	Valores superiores a 0,8
Logaritmo del total de revisiones de estado de cuenta, en app o sitio web	Valores superiores a 1,5
Logaritmo del total de revisiones de movimientos, en app o sitio web	Valores superiores a 1,2
Logaritmo del total de revisiones de ofertas en aplicación RF	Valores superiores a 3

Tabla 1: Valores límite para eliminación de outliers

Se efectúan estos criterios de eliminación para toda la base de datos. Cada fila debe cumplir con todas las restricciones mencionadas anteriormente. Después de realizar el filtro de outliers, de los 631.600 datos de clientes considerados en un principio, finalmente 532.936 de ellos (equivalente a un 84,4% del total) se consideran para el modelo de segmentación.

La tercera técnica utilizada para el preprocesamiento de datos es la de normalización de las variables, ya que se busca que todas las variables tengan un valor mínimo de 0 y un valor máximo de 1. Esto se efectúa porque en un modelo de segmentación, la normalización permite ponderar con el mismo peso a todas las variables y evita segmentos distinguidos en base a una o muy pocas variables (lo que sucede cuando alguna variable posee gran magnitud). Considerando que las variables binarias asociadas a los hitos digitales ya se encuentran entre 0 y 1, sólo se realiza la normalización de las variables de frecuencia de las acciones en app y sitio web.

Ya realizadas estas etapas de preprocesamiento, la base de datos de clientes se encuentra lista para realizar las técnicas de segmentación necesarias para entregar diversos perfiles de clientes.

8.1.4 Resultados del modelo de segmentación

En la etapa del proceso de segmentación, se utilizó el algoritmo de K-Prototypes para asignar los clientes al segmento más representativo, según los datos en las variables que posee. En primer lugar, es relevante destacar que para ejecutar este algoritmo se debe especificar el número de segmentos resultante. Dado esto, se debe realizar un análisis que deba indicar cuál es el número de clusters óptimo para el modelo.

Esta decisión no es exclusiva del trabajo de Memoria, ya que se realiza en conjunto con el área comercial de la compañía. Sin embargo, se añade a continuación, un análisis matemático y un análisis comercial que permite dilucidar cuál es el número de segmentos indicado, para llevar al modelo final.

A continuación, se analiza el indicador silhouette score para distintos números de segmentos. Este gráfico muestra que el modelo tiene un mejor desempeño, cuando el indicador es mayor, es decir, cuando el número de segmentos es mayor a 10. Además, en el caso de un número de segmentos menores a 10, los mejores resultados se muestran en los modelos con 5 y 6 segmentos, tal como se muestra en la siguiente ilustración:

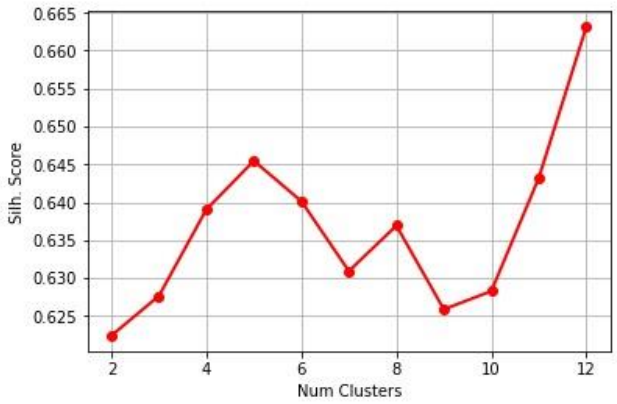


Ilustración 18: Silhouette Score por número de segmentos en Clustering

Analizando la situación desde un punto de vista comercial, se establecen como restricciones que el número de segmentos óptimo debe ser menor a 8 segmentos. Esta aseveración se debe a que un número de clusters muy grande impide resumir con claridad los resultados obtenidos hacia las distintas áreas de la compañía y además dificulta la planificación, realización de campañas y la medición de resultados futuros.

Por lo tanto, utilizando los resultados de silhouette score, se construyen modelos de 4, 5 y 6 segmentos, como propuestas de modelos para las áreas de Business Intelligence, Comercial y de Banca Digital. Estas áreas deciden en conjunto, cuántos

segmentos tendrá el modelo definitivo de clustering, en vista de la caracterización de cada modelo propuesto.

Se muestran resultados de las 3 segmentaciones de clientes mencionadas. En cada modelo, los segmentos respectivos se describen en base a los promedios de sus variables, por cada cluster, ya sean de cumplimiento de hitos o de frecuencia de acciones en app/sitio web.

En primer lugar, se construye el modelo de 4 segmentos, donde la caracterización se muestra a continuación y su detalle se ubica en el Anexo 3:

Segmento	Características	Tamaño (cantidad de clientes)
"No Digital"	Segmento con bajo cumplimiento de hitos	357.352
"App"	Segmento con alto cumplimiento de hito login en app, bajo cumplimiento de hito login en sitio web	78.309
"Sitio web"	Bajo cumplimiento de hito login en app y alto cumplimiento de hito login en sitio web	38.762
"Full Digital"	Segmento con alto cumplimiento de hito login en app, alta frecuencia de acciones en app y alta tasa de cumplimiento de hito de pago en aplicación	58.513

Tabla 2: Resumen de segmentos en modelo de 4 clusters

Por otro lado, en el resultado del modelo con 5 segmentos, detallado en el Anexo 4, sus clusters poseen las siguientes características:

Segmento	Características	Tamaño (cantidad de clientes)
"No Digital"	Segmento con bajo cumplimiento de hitos	350.913
"App Incipiente"	Segmento con alto cumplimiento de hito login en app, bajo cumplimiento de hito login en sitio web	58.916
"Sitio web"	Bajo cumplimiento de hito login en app y alto cumplimiento de hito login en sitio web	36.705
"App Frecuente "	Segmento con alto cumplimiento de hito login en app, bajo cumplimiento de hito login en sitio web y alta frecuencia de uso en aplicación en login y revisión de ofertas	40.764
"Full Digital"	Segmento con alto cumplimiento de hito login en app, alta frecuencia de acciones en app y alta tasa de cumplimiento de hito de pago en aplicación	45.638

Tabla 3: Resumen de segmentos en modelo de 5 clusters

Finalmente, el modelo con 6 segmentos, desglosado en el Anexo 5, tiene las siguientes características:

Segmento	Características	Tamaño (cantidad de clientes)
"No Digital"	Segmento con bajo cumplimiento de hitos	335.518
"App incipiente"	Segmento con alto cumplimiento de hito login en app y ofertas, bajo cumplimiento de los hitos restantes	34.333
"App intermedio"	Alto cumplimiento de hitos en general y con un mayor cumplimiento de hitos con respecto al segmento "App Incipiente"	53.643
"Sitio web"	Bajo cumplimiento de hitos en general y alto cumplimiento de hito login en sitio web	37.363
"App frecuente"	Alto cumplimiento de hitos, pero con una baja tasa de cumplimiento del hito de pagos en app	36.338
"Full Digital"	Alto cumplimiento de hitos y con alta tasa de cumplimiento del hito de pagos en app	35.741

Tabla 4: Resumen de segmentos en modelo de 6 clusters

En vista de los 3 modelos de segmentación propuestos, se decide escoger como modelo final de segmentación el que posee 5 segmentos, debido a que la caracterización de los segmentos resume de mejor forma el comportamiento de los clientes y sus niveles de digitalización.

Por otro lado, el modelo de 4 segmentos tenía las ventajas de dividir a los clientes por la plataforma de uso, sin embargo, la cantidad de personas en el segundo segmento de este modelo es de 8 mil en total. Se agrupa a gente que ha utilizado la aplicación, pero no distingue a la gente que usa mucho la aplicación de la gente que la ha utilizado poco. Dado esto, el modelo de 5 segmentos si establece esta distinción entre el segundo y cuarto segmento de este modelo.

En adición a lo anterior, en el modelo de 6 segmentos, 3 de ellos son segmentos para personas que utilizan mayoritariamente la aplicación, lo que no genera una mejor descripción que el modelo de 5 segmentos, que es más conciso y hace divisiones de los segmentos con información valiosa para el negocio de la compañía.

Después de escoger como modelo definitivo el ajuste de segmentación con 5 clusters, se elabora una caracterización más detallada de los segmentos. El análisis se concentra en la cantidad de hitos promedio, que posee cada segmento del modelo final de 5 segmentos. El siguiente gráfico muestra la cantidad de clientes para cada segmento y la cantidad de hitos promedio que cumple cada grupo:

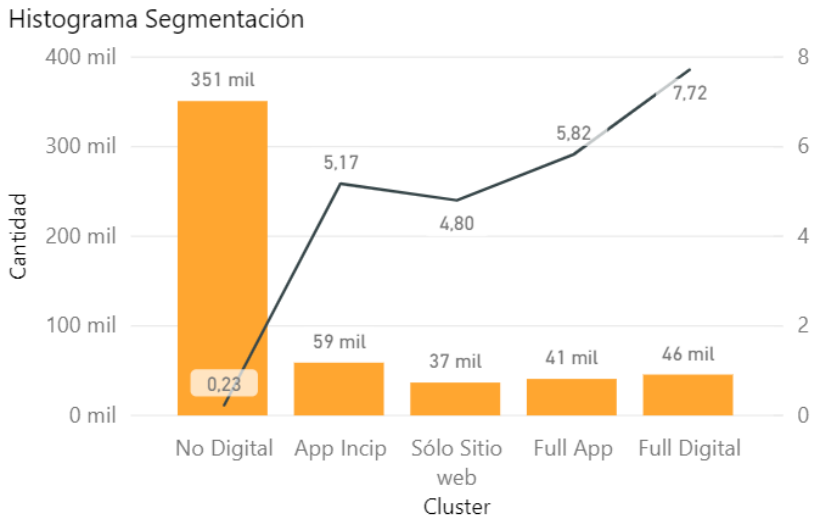


Ilustración 19: Cantidad de clientes y promedio de hitos por segmentos digitales

Por otro lado, se identifica a continuación la edad promedio para cada segmento, en conjunto con la cantidad de hitos promedio que posee cada cluster:

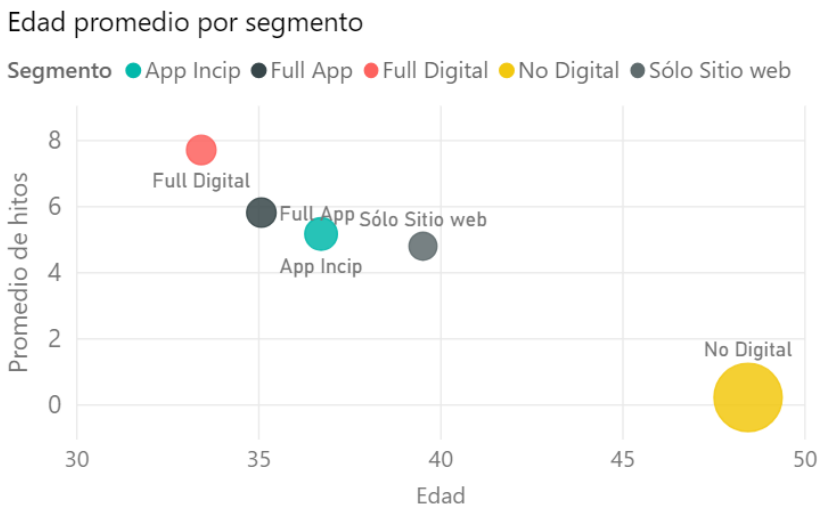


Ilustración 20: Edad y cantidad de hitos promedio por cada segmento digital

Luego se añade al análisis la variable de segmento etario, la cual agrupa a los clientes de acuerdo con la edad que poseen. Esta variable se encuentra previamente definida por la compañía y se utiliza con frecuencia en el análisis de clientes. En primer lugar, se detallan los segmentos etarios existentes:

- "Silencioso", nacidos antes de 1954.
- "Baby Boomers", nacidios entre 1954 y 1967.
- "X", nacidos entre 1968 y 1983.
- "Millenial", nacidos entre los años 1984 y 1997.
- "Z", nacidos después de 1998.
- "Extranjero", clientes que no poseen nacionalidad chilena (los segmentos anteriores son sólo compuestos por clientes de nacionalidad chilena).

Con esto, se grafica el % de personas de cada segmento digital, para cada segmento etario. Allí se muestra que el segmento "No Digital" va disminuyendo en proporción para los segmentos más jóvenes, Millenial y Z y además los segmentos "Full App" y "Full Digital" aumentan su proporción en los segmentos de menor edad.

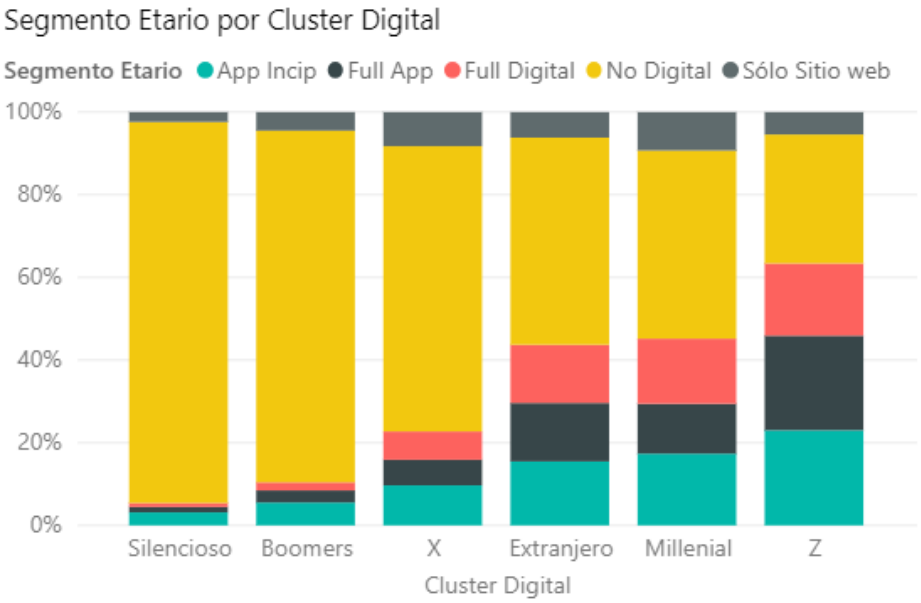


Ilustración 21: Proporción de segmentos etarios por cada segmento digital

Las ilustraciones anteriores muestran apreciaciones relevantes. Dentro de ellas, se destaca que existe un segmento con una cantidad de hitos mucho mayor a las demás, el grupo "Full Digital". Por otro lado, se observa que efectivamente los grupos de menor edad son más propensos a pertenecer a segmentos con mayor nivel de digitalización, desligándose en su proporción del grupo "No Digital" y donde los segmentos "Full App" y "Full Digital" son los que poseen menor edad promedio. Aunque una caracterización etaria es útil, en términos de descripción de cada segmento, es necesario establecer un análisis de mayor profundidad, que permitan dar indicios de los beneficios concretos que genera la digitalización de un cliente hacia la compañía.

En adición a los gráficos anteriores, se ubican otros gráficos adicionales, de descripción de los segmentos de clientes de la tarjeta RF, en base a acciones digitales. Se incluyen variables como Estado Civil, Género y % de clientes con nacionalidad chilena para cada segmento digital.

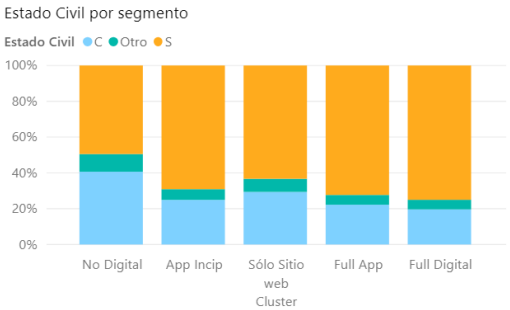


Ilustración 22: % de personas por estado civil, por cada segmento digital

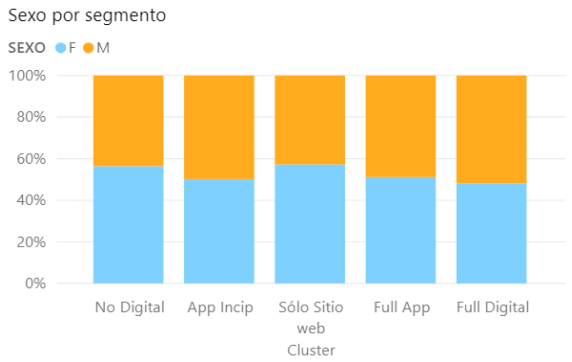


Ilustración 23: % de clientes por género, por cada segmento digital

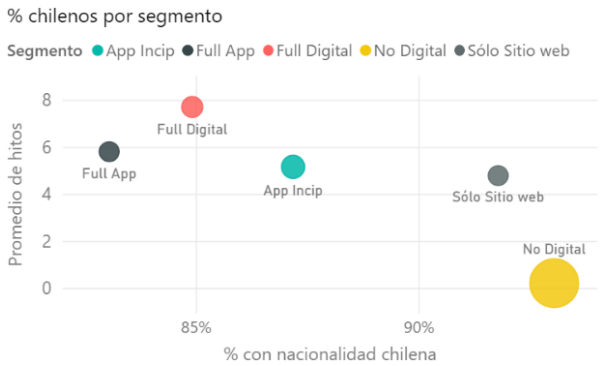


Ilustración 24: % de clientes de nacionalidad chilena, por cada segmento digital

Como último análisis asociado a la descripción de los segmentos resultantes del modelo final de clustering, se muestran gráficos que incluyen las variables de transacciones y pagos con la tarjeta RF, para instalar un primer acercamiento hacia el análisis de transacciones y pagos de los clientes digitales. A continuación, se grafica el promedio de transacciones con la tarjeta de crédito de RF, para los últimos 6 meses, entre febrero y julio de 2019, para cada uno de los segmentos resultantes en el proceso de clustering:

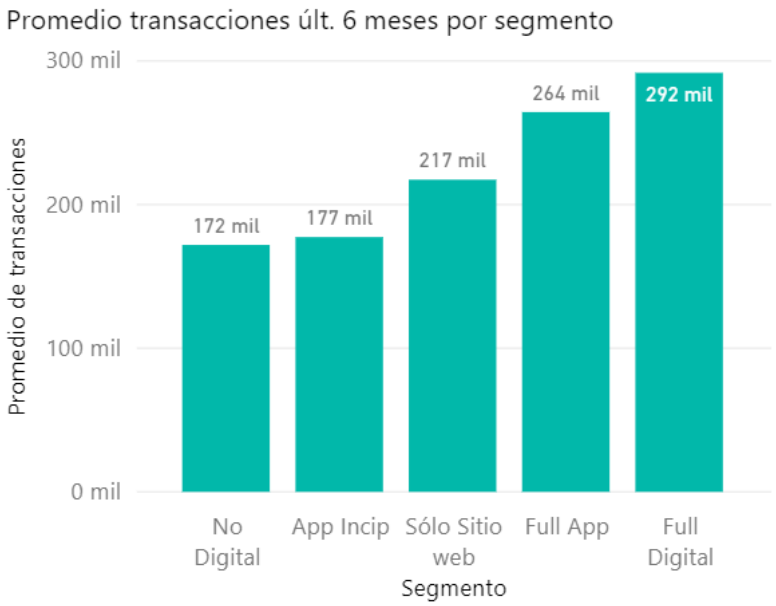


Ilustración 25: Promedio de monto de transacciones con tarjeta RF, entre febrero y julio, por segmento digital

Por otro lado, la siguiente ilustración muestra el promedio de pagos con la tarjeta RF, entre febrero y julio de 2019, para cada uno de los grupos correspondientes a la segmentación digital:

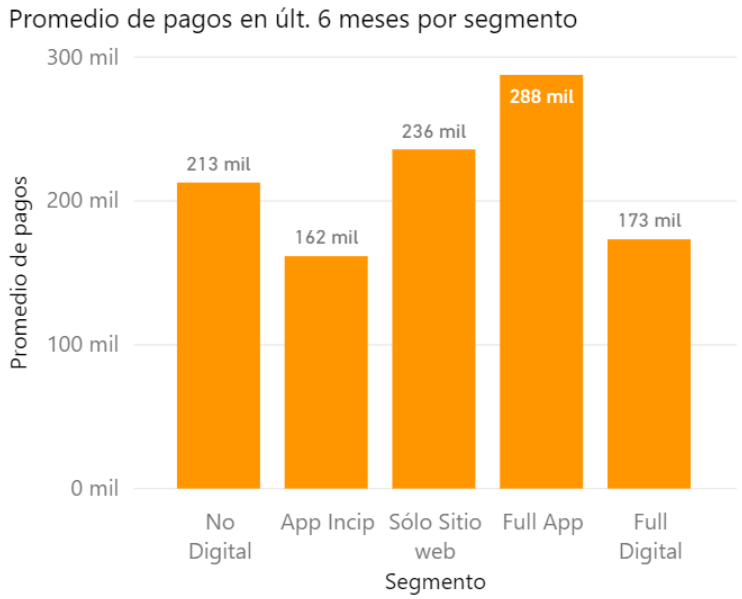


Ilustración 26: Promedio de monto de pagos con tarjeta RF, entre febrero y julio, por segmento digital

Con estas dos ilustraciones, se observa que los dos segmentos con mayor nivel de digitalización, el segmento “Full App” y “Full Digital”, son los que mayor promedio de transacciones poseen, sin embargo, el segmento “Full Digital” tiene un promedio

de pagos muy bajo y el segmento “Full App” es el que mejor desempeño posee en cuanto a los pagos con la tarjeta.

Este análisis es sobre promedios simples en montos de transacciones y pagos, aún es necesario profundizar mejor el análisis sobre las transacciones y pagos de los distintos grupos de clientes. Dado esto, en la sección posterior se pretende analizar los valores mensuales de transacciones y pagos de cada cliente y también calcular para cada mes del análisis, el segmento al cual pertenece el cliente con el modelo entrenado en esta primera etapa del trabajo.

8.2 Análisis de transacciones y pagos de clientes digitales

En esta fase del trabajo se realiza un análisis de las transacciones y pagos de los clientes digitales, generando una comparación con respecto a los clientes no digitales. Desde la etapa anterior, se tiene como resultado previo la caracterización de diversos segmentos, para los clientes que poseen la tarjeta RF. Con esto, se deben construir diversos criterios, que asignen a cada cliente el estado de "Digital" o "No digital". Se realizará esta agrupación para los clientes, en base a los hitos digitales que poseen entre febrero y junio de 2019, asignando de esta forma el estado digital o no digital del cliente en el mes de junio de 2019.

Por otro lado, el análisis de transacciones y pagos sugiere una ventana de tiempo anterior y posterior al evento de digitalización. En primer lugar, se escoge observar transacciones y pagos, como dos variables que influyen en la rentabilidad del cliente. Además, se considera observar el análisis anterior en noviembre de 2018 (3 meses antes del evento) y el análisis posterior en septiembre de 2019 (3 meses después). Por ende, el análisis de digitalización se realizará considerando el historial de hitos digitales entre febrero y junio de 2019.

Ilustración 27: Diagrama con fechas de análisis de transacciones y pagos

8.2.1 Selección de datos

Dentro de la selección de datos a considerar para esta etapa, se consideran algunos filtros utilizados en el proceso de segmentación, donde el cliente debe poseer su tarjeta sin bloqueos en agosto de 2019. Además, debe haber realizado algún movimiento en los últimos 12 meses, para observar que existe un mínimo nivel de actividad como cliente de la compañía.

En adición, para esta etapa deben considerar clientes aptos para generar un análisis previo de transacciones y pagos, en noviembre de 2018 y un análisis de digitalización entre febrero y junio de este año. Dado esto, no se consideran clientes que han

abierto la tarjeta durante el año 2019, sólo se toman en cuenta los clientes antiguos que han abierto la tarjeta antes de diciembre de 2018.

Cabe destacar que al considerar los clientes que abrieron la tarjeta en noviembre y diciembre de 2018, automáticamente no tendrán transacciones y pagos dentro de los datos en el período de "Análisis Previo", por lo que una opción a considerar era no incluirlos en el estudio. Sin embargo, estos clientes componen un porcentaje importante del total de clientes que abrieron la tarjeta en 2018, siendo un 24,88% del total. Por ello, se optó finalmente por considerarlos en el análisis de transacciones y pagos. Tomando en cuenta los filtros descritos, el universo de clientes a considerar en esta etapa es de 433.152 personas, para un análisis de 11 meses, desde noviembre de 2018 a septiembre de 2019.

8.2.2 Análisis de Transacciones y Pagos

En esta sección se muestra, como paso previo a la definición de cliente digital, un análisis de transacciones y pagos de los clientes de la tarjeta RF, con respecto a los distintos segmentos a los que puede pertenecer el cliente, utilizando el modelo de segmentación realizado en la primera etapa del trabajo. En primer lugar, se calcula el segmento digital al que pertenece el cliente, para el período de digitalización a analizar, entre febrero y junio de 2019. Luego se realiza un análisis exploratorio de cada segmento, considerando las transacciones y pagos que poseen, en la ventana anterior y posterior al proceso de digitalización. Finalmente, se realiza un análisis agregado (sin la separación por segmentos), del monto total de transacciones y pagos, para los meses de análisis del modelo Diff-in-Diff, entre noviembre de 2018 hasta septiembre de 2019, para observar estacionalidad de estas variables.

Dentro del cálculo de los segmentos de cada cliente al mes de junio, considerando las acciones que realiza en app y sitio web, se observa a continuación, el tamaño de cada segmento digital, donde se observa un segmento predominante que es el "No Digital":

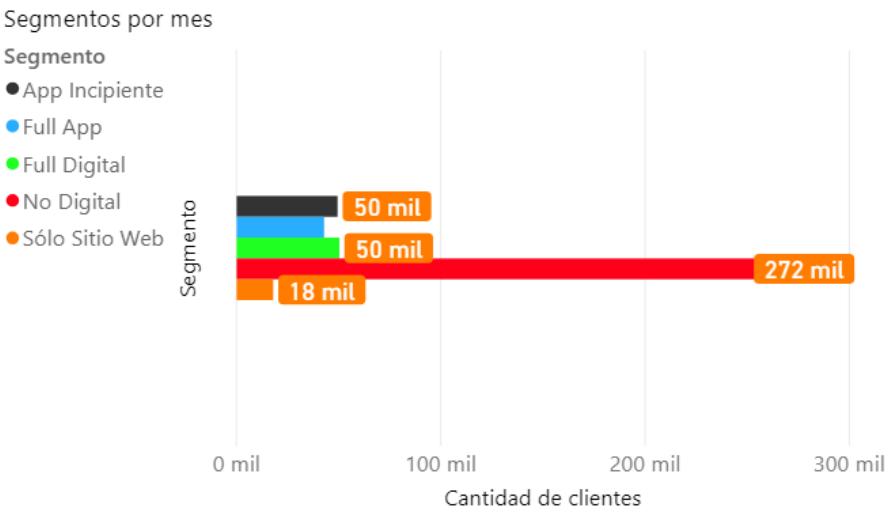


Ilustración 28: Cantidad de clientes de cada segmento digital, en el mes de junio

Con los segmentos digitales asignados para los clientes de la tarjeta, en la siguiente ilustración se observa el promedio de transacciones y de pagos que posee cada grupo, para los meses de análisis de digitalización:

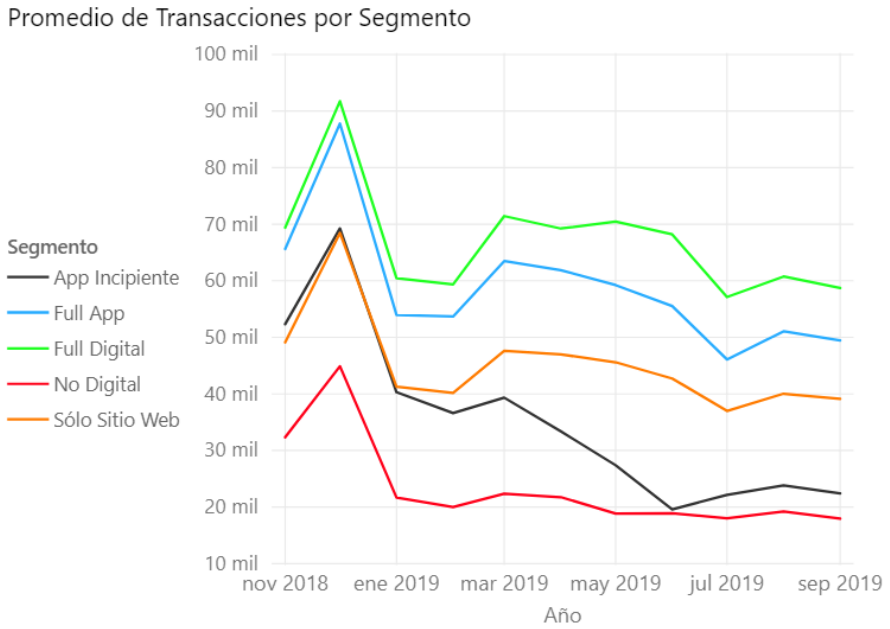


Ilustración 29: Promedio de monto de transacciones, por segmento entre febrero y junio 2019

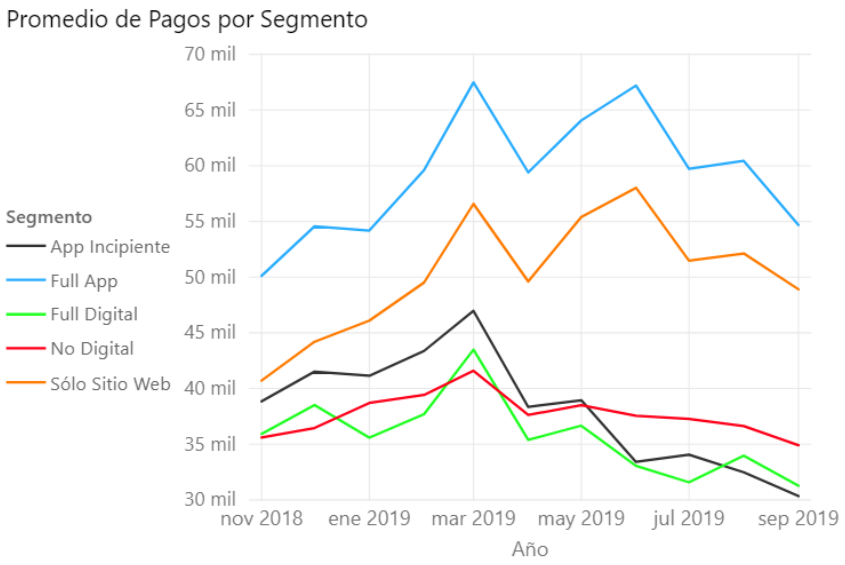


Ilustración 30: Promedio de monto de pagos, por segmento entre febrero y junio 2019

Dentro de las apreciaciones que se pueden destacar, se observan dos segmentos con buenos resultados en sus transacciones, el segmento "Full Digital" y "Full App". Sin embargo, al observar los pagos por cada mes, el segmento "Full Digital" tiene resultados muy bajos y los segmentos "Full App" y "Sólo Sitio Web" son los que tienen un mejor desempeño.

Por otro lado, se realiza un esquema que muestra el monto total de transacciones y pagos con la tarjeta RF por cada mes, para observar tendencias principales en estacionalidad que puedan ser relevantes para el modelo Diff-in-Diff.

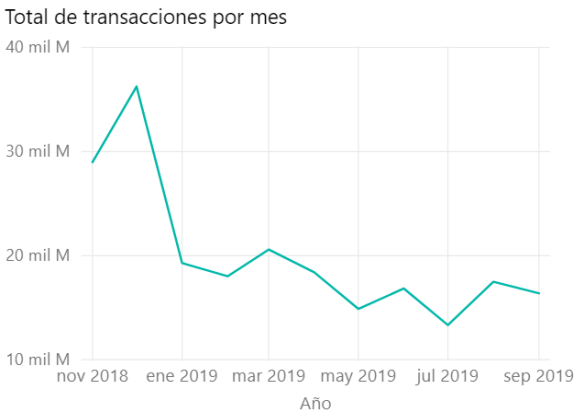


Ilustración 31: Monto total de transacciones con tarjeta, por mes

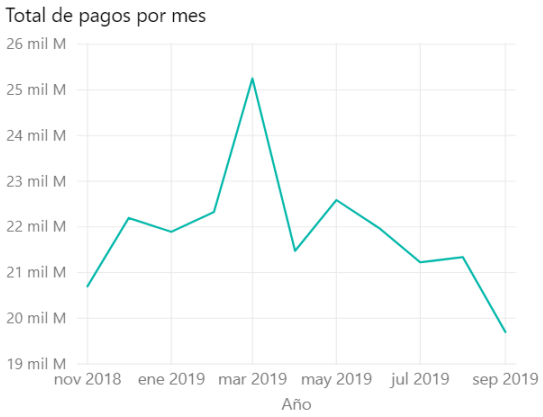


Ilustración 32: Monto total de pagos con tarjeta, por mes

Desde los gráficos anteriores, se muestra que las transacciones tienen un valor mucho mayor en noviembre y diciembre, a diferencia de los primeros meses del año. Por otro lado, el mes con mayor monto promedio de pagos es el de marzo y el comportamiento de los otros meses posee mayor estabilidad.

Es importante considerar, para el análisis de diferencias de transacciones, que los clientes podrían disminuir el monto de transacciones, al comparar los instantes antes y después de la digitalización. Debido a que las transacciones en noviembre de 2018 son mayores que en septiembre de 2019. Por ello, se espera que en el modelo de Diff-in-Diff, la variable T (Tiempo) resulte en un valor negativo, ya que incorpora esta estacionalidad. Por otro lado, en los pagos no se observan grandes diferencias al observar este análisis agregado de los clientes.

8.2.3 Definición Cliente Digital

Teniendo en cuenta las observaciones de transacciones y pagos y sus diferencias para cada segmento, se procede a definir criterios de digitalización de los clientes de la tarjeta RF. En líneas generales, hay dos principios que se espera que cumpla un cliente digital, el primero es que logre visitas recurrentes a los canales digitales, ya sean aplicación o sitio web y el segundo es que también cumpla con un grado de conversión final en la página, que sea directamente beneficioso hacia el negocio. Bajo estos lineamientos se establece una idea de lo que se pretende para la compañía, como definición de un cliente digital.

Para efectos prácticos del trabajo de Memoria, para cumplir el primer principio de cliente digital se considerará el segmento digital al cual pertenece el cliente, donde la condición clave es que pertenezca al cluster “Full App” o “Full Digital”, los cuáles se asocian de mejor forma con un cliente que visita frecuentemente los canales digitales. En adición a lo anterior, el segundo principio de cliente digital se cumple cuando un cliente logra hacer pagos con su tarjeta en los últimos 5 meses, por medio de la aplicación web o el sitio web de RF, este hito se establece como una condición clave que debe cumplir un cliente digital, ya que es la conversión final que se tiene en ambos canales digitales, de acuerdo con diversas áreas asociadas al negocio.

Resumiendo lo anterior, si los dos criterios descritos anteriormente se cumplen, un cliente es “Digital”, sino el cliente es considerado “No Digital”. Considerando esto, durante el mes de junio existen 43.860 clientes digitales, equivalentes al 10,12% del universo de clientes estudiado en esta sección.

8.2.4 Análisis Descriptivo de Clientes Digitales y Clientes No Digitales

Con la definición de criterios de digitalización de clientes y con la asignación realizada para los clientes que poseen la tarjeta, se realiza una tabla comparativa que permite observar características principales de los clientes digitales y no digitales.

Indicador	Clientes Digitales	Clientes No Digitales
Cantidad de clientes	43.860	389.292
Edad promedio	33,1	44,4
Antigüedad promedio (meses)	13,27	14,34
% mujeres/total	51,7%	55,26%
% extranjeros /total	16,28%	9,26%
Deuda total promedio	97.626	123.120
Cupo total promedio	412.526	408.200

Tabla 5: Promedio de indicadores en clientes digitales y no digitales

Dentro de las apreciaciones principales, se puede observar que los clientes digitales son mucho más jóvenes que los clientes no digitales y el porcentaje de extranjeros del grupo digital es mucho mayor, en comparación al grupo de clientes no digitales. Además, el monto de deuda es menor en el caso de los clientes digitales.

También se realiza una observación de la evolución de cada mes en las cifras promedio de transacciones y pagos, comparando los clientes digitales v/s clientes no digitales.

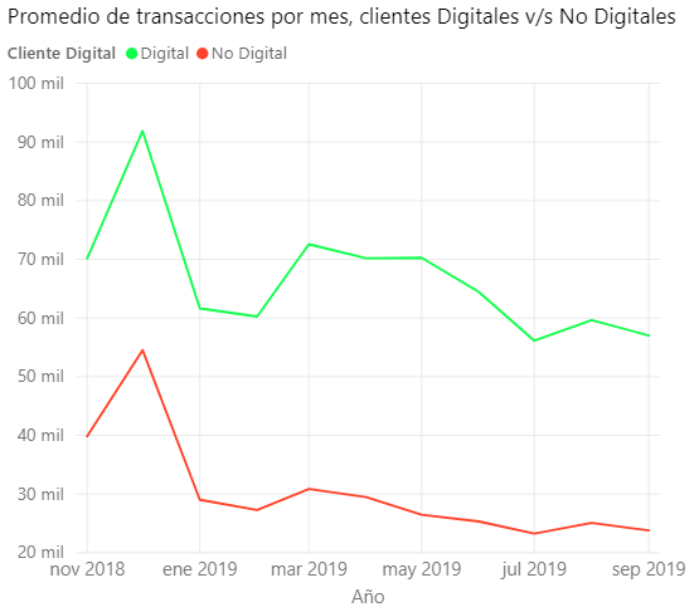


Ilustración 33: Promedio de transacciones por mes, en clientes digitales y no digitales

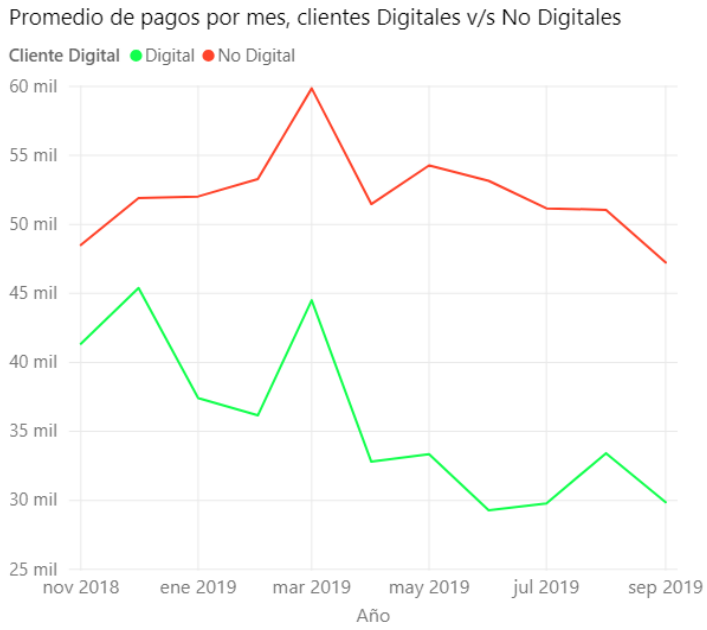


Ilustración 34: Promedio de pagos por mes, en clientes digitales y no digitales

En estas ilustraciones se muestra que las transacciones son mayores en promedio en el caso de clientes digitales, sin embargo, las variaciones existentes son parecidas para ambos grupos, donde la diferencia de promedios parece constante. Por otro lado, los pagos si muestran variaciones, donde además de observar estacionalidad, se aprecia una tendencia a la baja en los pagos promedio para el grupo de clientes

digitales y los clientes no digitales tienen leves variaciones en el monto de pagos promedio.

Este análisis es útil para obtener un resultado previo a la realización del modelo Diff-in-Diff, el cual permitirá confirmar si existe un cambio en las transacciones y pagos de los clientes, debido al evento de digitalización asociado.

8.2.5 Preprocesamiento de Datos para modelo Diff-in-Diff

Dentro de la etapa de preprocesamiento de datos, se tiene como resultado previo la cantidad de clientes a considerar, que es de 433.152 personas. En base a esto se ingresan 866.304 datos, ya que se evalúa el estado "Digital" o "No Digital" de los clientes en los meses de junio y julio de 2019, considerando su comportamiento en los canales digitales y los hitos digitales que cumplen en los 5 meses previos a los meses estudiados.

Con la primera cifra de datos a considerar para el modelo Diff-in-Diff, se llevan a cabo los procesos de eliminación de missing values y outliers antes de realizar el modelo. En primer lugar, se efectúa la eliminación de filas con valores nulos o fuera de rango, donde se generó un filtro para las siguientes variables:

- Nacionalidad, se eliminan valores no existentes. 38 datos eliminados.
- Edad, se eliminan clientes con edad menor a 18 y mayor que 110. 262 datos eliminados
- Antigüedad, se eliminan valores no existentes. 10 datos en total
- Estado Civil, se eliminan otros valores diferentes a soltero/casado o sin valor. 72.218 casos eliminados.
- Se eliminan clientes con valores sobre 500.000 en transacciones y sobre 250.000 en pagos. 29.469 casos en total.

Luego de realizar estos filtros, se consideran 764.433 datos para el modelo de diferencias en diferencias.

El último paso efectuado es la creación de las variables binarias D, T y DxT, donde D indica si el cliente es digital, T indica si la fila corresponde al período antes (noviembre 2018) o después (septiembre 2019) de la digitalización y DxT es la variable de interacción que permite capturar el efecto de la digitalización en transacciones y pagos.

8.2.6 Resultados modelo Diff-in-Diff

Para el modelo Diff-in-Diff se realizaron 4 regresiones. 2 de ellas calculan cambios en las transacciones de los clientes, comparando el período anterior a la digitalización (3 meses antes) y el período posterior a la digitalización (3 meses después) y las 2 regresiones restantes comparan los pagos de los clientes comparando los mismos períodos.

Modelo 1: Monto de transacciones, sin incluir variables anexas

En este modelo se consideran sólo las variables D, T y DxT, donde se muestran los efectos causados por estas variables que generan cambios en transacciones, los resultados entregados por esta regresión son los siguientes:

Variable	Coeficiente
Cliente Digital(D)	29.640
Post-Tratamiento (T)	-15.660
Efecto Digitalización (DxT)	3.581

Tabla 6: Coeficientes de influencia hacia el monto de transacciones, sin la inclusión de variables anexas

Modelo 2: Monto de pagos, sin incluir variables anexas

Para esta regresión se intenta ajustar el cambio en los pagos de los clientes, utilizando al igual que en el primer modelo, solamente las variables D, T y DxT. Los resultados asociados a este modelo son los siguientes:

Variable	Coeficiente
Cliente Digital(D)	-7.657
Post-Tratamiento (T)	-13
Efecto Digitalización (DxT)	-9.548

Tabla 7: Coeficientes de influencia hacia el monto de pagos, sin la inclusión de variables anexas

Modelo 3: Monto de transacciones, incluyendo variables anexas

Para este caso se consideran, al igual que en los casos anteriores, las variables D, T y DxT, pero se añaden las variables de Edad, Antigüedad, Sexo, Estado Civil y Nacionalidad. Estas variables pueden ejercer influencia en el modelo y modificar de esa forma el coeficiente de la variable DxT que se desea analizar. El ajuste resultante en los coeficientes de las variables D, T y DxT es el siguiente:

Variable	Coeficiente
Cliente Digital(D)	27.030
Post-Tratamiento (T)	-15.640
Efecto Digitalización (DxT)	3.617

Tabla 8: Coeficientes de influencia hacia el monto de transacciones, incluyendo variables anexas

Modelo 4: Monto de pagos, incluyendo variables anexas

Para este caso, también se añaden las variables de Edad, Antigüedad, Sexo, Estado Civil y Nacionalidad, pero ahora se realiza una regresión que ajusta la diferencia de pagos de los clientes. Los resultados que genera el modelo son los siguientes:

Variable	Coeficiente
Cliente Digital(D)	-5.039
Post-Tratamiento (T)	-96
Efecto Digitalización (DxT)	-9.595

Tabla 9: Coeficientes de influencia hacia el monto de pagos, incluyendo variables anexas

Para las tablas 8 y 9 sólo se muestran los coeficientes de las variables D, T y DxT. En el Anexo 6 se ubica el detalle de los valores de coeficientes asociados a todas las variables incluidas en el modelo.

Como apreciaciones principales de los modelos de regresión y sus resultados finales, es importante destacar que el efecto de digitalización muestra un cambio positivo para las transacciones de clientes, en el período posterior a la digitalización. Sin embargo, el coeficiente de DxT asociado a los pagos expone un efecto negativo de esta variable en el modelo. Además, no se producen grandes variaciones en los coeficientes de los modelos sin variables anexas y con variables anexas.

Por último, se destaca que los modelos y los coeficientes respectivos son significativamente distintos de cero, tal como se muestra en el Anexo 6. Los coeficientes de las variables DxT poseen un intervalo de confianza que se resume en la siguiente tabla:

Modelo	Valor mínimo	Valor máximo
Transacciones, con variables anexas	2.533,8	4.700,2
Pagos, con variables anexas	-10.400	-8.828,2

Tabla 10: Intervalo de confianza para coeficiente de variable DxT, en modelos de cambios de transacciones y pagos, con variables anexas

Es importante agregar que estos resultados no pueden ser generalizados para cualquier mes del año, ni para cualquier selección de clientes existente en la compañía, pero este primer resultado permite tener una idea inicial y un modelo que se debe ajustar con una mayor cantidad de datos para hacer una aseveración general.

8.3 Modelo de propensión de digitalización de clientes

En esta sección del trabajo de Memoria se muestran los resultados obtenidos al realizar diversos modelos de propensión, que buscan predecir si los clientes se digitalizarán o si no lo harán.

Se realizaron tres modelos predictivos de clasificación:

- El primer modelo busca predecir la digitalización de clientes, dado considerando el universo de clientes que si hicieron login en aplicación y sitio web de RF.
- El segundo modelo busca predecir si un cliente No Digital logra convertirse en un cliente Digital. En este caso el universo de clientes abarca a los que no efectuaron login en aplicación y sitio web de RF. Es importante destacar que son 4 los segmentos distintos al "No Digital" ("Full App", "Full Digital", "App Incipiente" y "Sólo Sitio Web") y su definición es distinta a la de cliente "Digital" o "No Digital", detallada en la segunda parte del presente trabajo.
- El tercer modelo predice si el cliente ocupa la aplicación o sitio web de RF, 4 meses después de la medición de los datos. Dentro del universo de clientes, se consideran personas que si han realizado login y también los que no han efectuado login en los canales digitales de RF.

Cada uno de los modelos posee distintos algoritmos de clasificación, donde se utilizaron los modelos Logit, Árbol de Decisión, Random Forest, Naive Bayes, Support Vector Machine y K-Nearest Neighbors. También se hicieron variaciones en los niveles de Oversampling que se aplican al modelo, debido a una gran proporción de etiquetas negativas, es decir, valores con la variable "Y" a predecir igual a cero.

8.3.1 Primer y segundo modelo predictivo

8.3.1.1 Selección y Preprocesamiento de Datos

Dentro de la selección y recopilación de datos, el primer y segundo modelo consideran registros de datos de los clientes por cada mes, desde junio hasta el mes de septiembre, con sus variables explicativas asociadas a cada uno de los meses. En el universo de clientes que abarca el modelo, se escoge el mismo filtro que en el modelo Diff-in-Diff, seleccionando clientes que han abierto tarjeta antes del mes de enero de 2019 y que hayan tenido movimientos con la tarjeta en los últimos 12 meses.

En total, 433.152 clientes son seleccionados en el primer filtro, lo que al considerar los registros de cada cliente por 4 meses, resulta en 1.732.608 datos para llevar inicialmente al modelo.

Por otro lado, las variables escogidas para el modelo predictivo consideran desde datos etarios, demográficos, mediciones de hitos digitales y otros montos asociados con la tarjeta. La lista de variables consideradas es la siguiente:

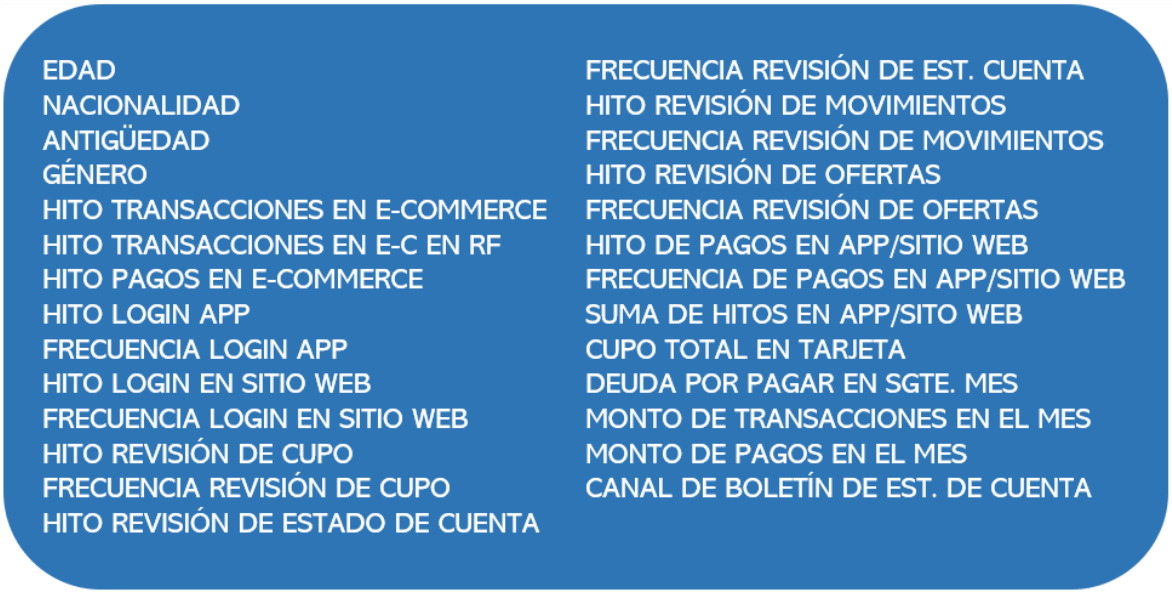


Ilustración 35: Lista de variables consideradas para los modelos predictivos

En la lista de variables, se consideran como variables binarias los hitos digitales, si el cliente efectuó el hito digital en los últimos 5 meses. El género y la nacionalidad también son variables binarias, con valor 1 si es mujer y chileno/a, respectivamente. La otra variable binaria presente es el canal de boletín del estado de cuenta, que es 1 si se envía por mail y 0 si se entrega directamente al hogar del cliente. Además, las variables de frecuencia de hitos se consideran para los últimos 5 meses, con respecto al mes de registro.

Dentro del preprocesamiento de datos, también se eliminan outliers y valores fuera de rango. Para ello se consideran los siguientes filtros, varios de ellos asociados a las variables de frecuencia de los hitos digitales:

- Variable de Frecuencia “Login App”: Se eliminan valores superiores a 77.
- Variable de Frecuencia “Cupos”: No se consideran valores superiores a 6.
- Variable de Frecuencia “Est. Cuenta”: Se eliminan valores superiores a 17.
- Variable de Frecuencia “Movimientos”: No se toman en cuenta datos superiores a 11.

- Variable de Frecuencia “Ofertas”: Se eliminan datos superiores a 95.
- Variable de Edad: Sólo se toman en cuenta personas entre 20 y 85 años. Se eliminan registros con datos faltantes, fuera de rango o outliers con más de 85 años.

* Los valores de outliers en las variables de frecuencia, utilizan el criterio de percentil 95% para establecer el valor máximo.

Después de ejecutar los diversos filtros de outliers y valores fuera de rango, los registros resultantes son 1.462.978.

Además, para cada par cliente-mes se asigna el segmento al cual pertenece y si es o no un cliente digital en el siguiente mes con respecto al mes de registro (por ej. etiqueta “Y” del mes de julio, para los registros “X” de junio; etiqueta “Y” de agosto, para los registros “X” de julio). Debido a la existencia de outliers, existen registros a los cuales no se les identifica el segmento al que pertenece y no poseen etiqueta. Estos registros se eliminan del modelo, en el filtro que se describirá a continuación.

El primer modelo requiere:

- Clientes no digitales (durante el mes de registro).
- Clientes que sí hayan tenido login en aplicación o sitio web de RF, en los últimos 5 meses.
- Clientes que si tengan etiqueta de cliente digital en el siguiente mes

El segundo modelo requiere

- Clientes no digitales
- Clientes que no hayan tenido login en aplicación o sitio web, considerando la misma ventana de 5 meses.
- Clientes que si tengan etiqueta de cliente digital en el siguiente mes

En base a esto, para el primer modelo, un total de 345.218 registros cumplen con los requisitos mencionados. Por otro lado, 999.097 datos superan los filtros requeridos en el segundo modelo y son seleccionados para calcular los algoritmos correspondientes.

Para la división en muestra de entrenamiento y de testeo, se escogieron 3 meses para la muestra de entrenamiento y el mes restante, para la muestra de testeo del modelo. Esto se efectuó para el primer y segundo modelo predictivo.

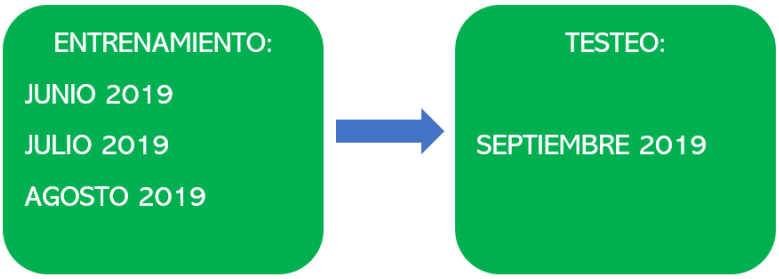


Ilustración 36: Meses de entrenamiento y testeo

8.3.1.2 Análisis exploratorio

Considerando los datos extraídos en la primera selección, se genera un análisis exploratorio de los datos, donde se realiza una división entre los registros con etiquetas positivas (clientes digitales al mes siguiente) y etiquetas negativas (no digitales al mes siguiente) y se analizan las variables con respecto a estos dos grupos.

A continuación, se muestra el análisis exploratorio del primer modelo, que considera a los clientes si tuvieron login en alguno de los canales digitales.

En cuanto a la cantidad de registros, aproximadamente 8.500 de ellos tienen etiqueta positiva, lo que equivale a un 2,48% del total de datos. Además, en la Ilustración 37, se muestra que los registros con etiqueta positiva son en promedio más jóvenes y levemente más antiguos que los clientes con etiqueta negativa.

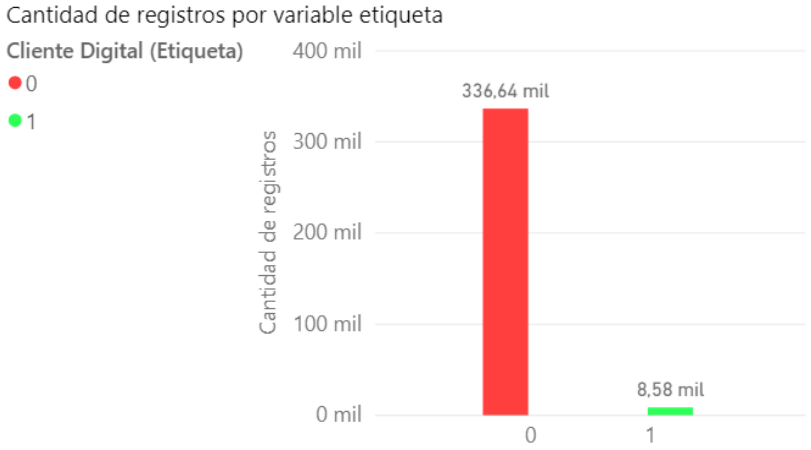


Ilustración 37: Cantidad de registros por variable etiqueta (1er modelo predictivo)

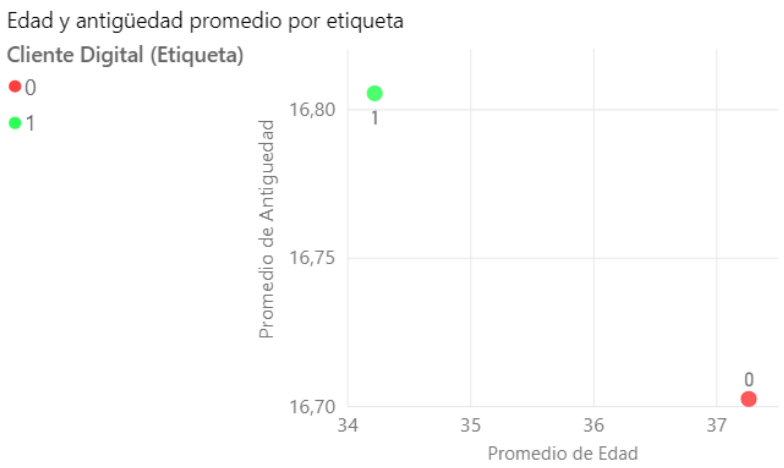


Ilustración 38: Promedio de edad y antigüedad, por variable etiqueta (1er modelo predictivo)

Observando otras variables demográficas, la proporción de etiquetas negativas varía en un leve porcentaje al hacer diferencias por género y nacionalidad.

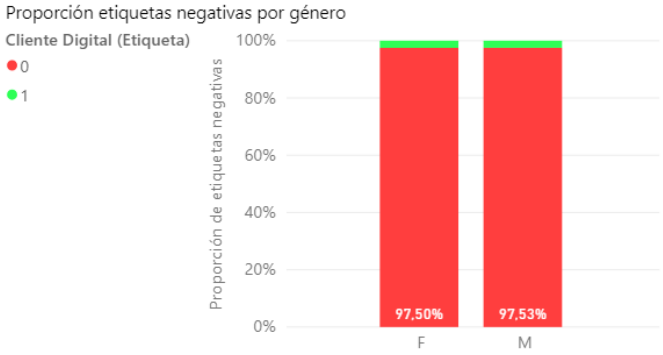


Ilustración 39: Proporción de etiquetas negativas, por género (1er modelo predictivo)

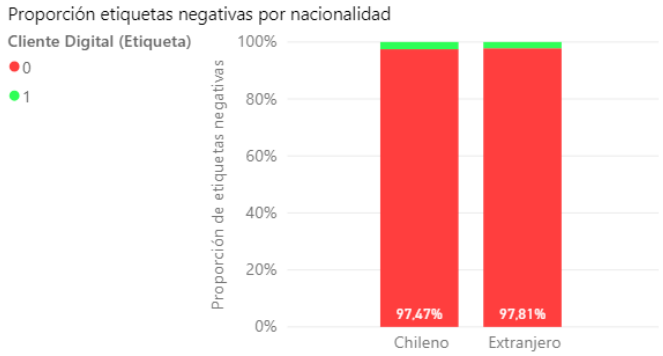


Ilustración 40: Proporción de etiquetas negativas, por nacionalidad chilena o extranjera (1er modelo predictivo)

Observando las variables de hitos digitales y de frecuencia de hitos que poseen los clientes, se muestra que para el grupo de clientes que se transformaron en digitales, ellos previamente poseen en los hitos de Cupos, Login App, Ofertas y Pagos, un porcentaje de cumplimiento mucho mayor que los clientes con etiqueta negativa, que

no se convirtieron en digitales. Con los resultados de la tabla 12, se observa que la cantidad promedio de hitos es mayor en el grupo de registros con etiqueta positiva. También se observa que las frecuencias promedio de los hitos Login App y Ofertas, son mayores en los registros con etiqueta positiva, en comparación con los datos que tienen etiqueta negativa.

% Cumplimiento / Etiqueta	Negativa (0)	Positiva (1)
% Hito Cupos	50,14%	69,02%
% Hito Estado de Cuenta	77,43%	88,82%
% Hito Login App	77,91%	98,86%
% Hito Login Plat	35,18%	27,21%
% Hito Movimientos	71,03%	73,8%
% Hito Ofertas	77,86%	98,86%
% Hito Pagos	4,55%	37,72%
% Hito Pagos E-Commerce	23,04%	37,24%
% Hito Transacc E-Commerce	10,9%	21,49%
% Hito ABCDIN	1,52%	2,73%

Tabla 11: % de cumplimiento de hitos, por variable etiqueta (1er modelo predictivo)

Promedio de hito / Etiqueta	Negativa (0)	Positiva (1)
Promedio de Cupos	1,05	1,33
Promedio de Estado de Cuenta	3,93	4,4
Promedio de Login App	8,72	13,03
Promedio de Login Plat	0,76	0,57
Promedio de Movim	2,17	2,34
Promedio de Ofertas	10,11	15,05
Promedio de Pagos	0,06	0,54
Promedio de Suma de Hitos	3,94	4,94

Tabla 12: Frecuencia promedio de hitos, por variable etiqueta (1er modelo predictivo)

Además, se observan las variables asociadas al uso de la tarjeta, en cuanto a transacciones, pagos, cupo total y deudas. En este análisis, se muestra que los clientes que se transformaron en digitales poseen en promedio mayores transacciones que el grupo de clientes que no se transformaron en digitales. Sin embargo, poseen un menor monto de pagos con la tarjeta. Si bien, las deudas promedio son menores, el bajo nivel de pagos concluye que los clientes con etiqueta positiva tienden a disminuir su monto de pagos, tal como se observó anteriormente en el análisis de Diferencias y Diferencias.

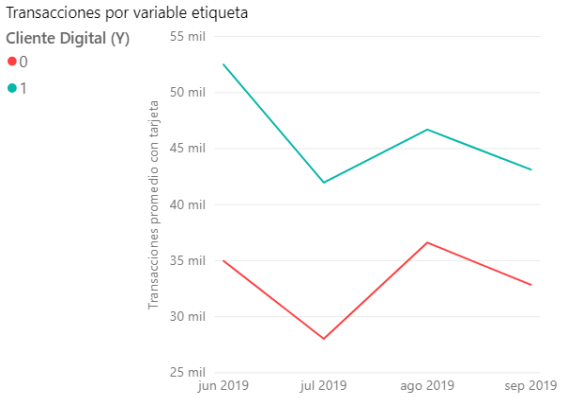


Ilustración 41: Transacciones promedio con tarjeta RF, por variable etiqueta (1er modelo predictivo)

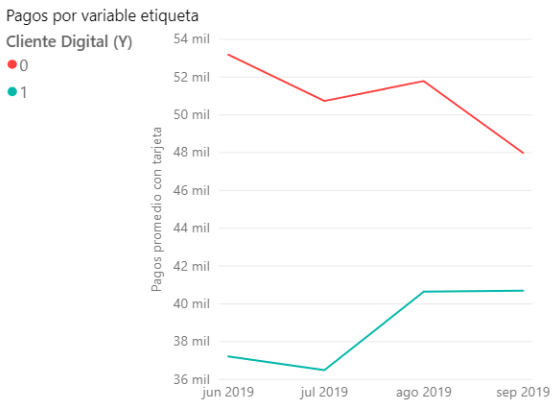


Ilustración 42: Pagos promedio con tarjeta, por variable etiqueta (1er modelo predictivo)

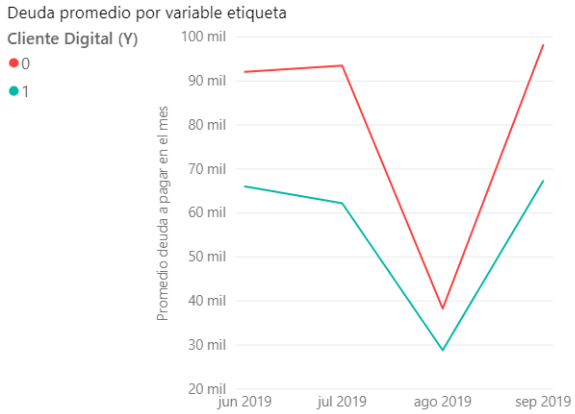


Ilustración 43: Deuda por pagar promedio en tarjeta, por variable etiqueta (1er modelo predictivo)

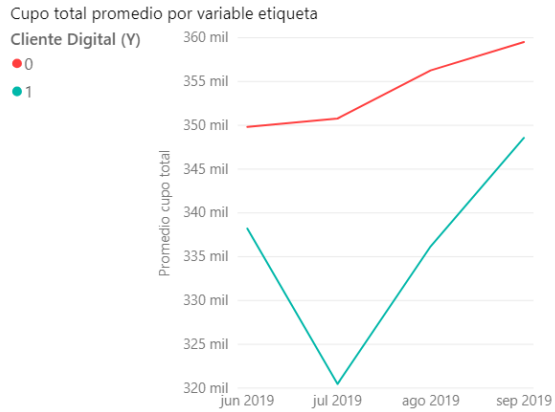


Ilustración 44: Cupo total promedio en tarjeta, por variable etiqueta (1er modelo predictivo)

En adición a los gráficos anteriores, también se realizó una observación más exhaustiva de algunas variables, para promover variables adicionales al modelo predictivo.

La primera variable por analizar es la de frecuencia del Hito Login App. Allí se observa para cada valor de la variable de frecuencia, la proporción de datos positivos con respecto al total de datos.

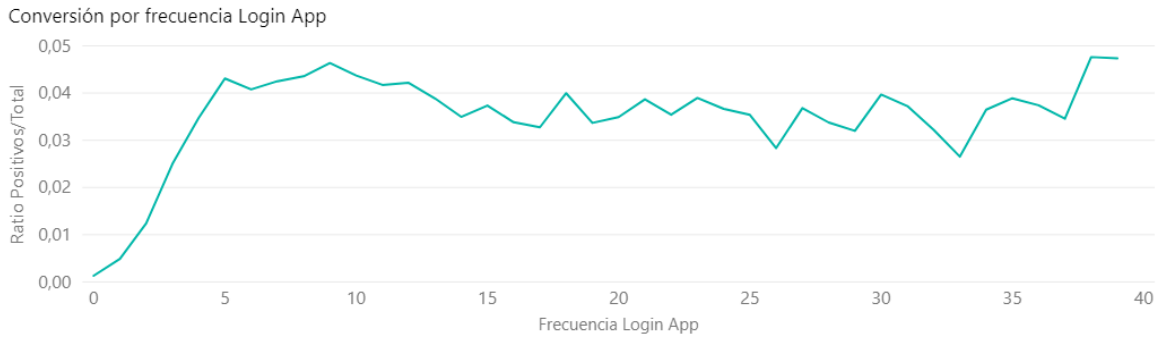


Ilustración 45: Ratio de etiquetas positivas v/s total de datos, por valor de frecuencia Login App

Observando este gráfico, se muestra una tendencia ascendente para los primeros valores de frecuencia. Cuando esta frecuencia es mayor a 5, el ratio de datos positivos comienza a estabilizarse y no aumenta notoriamente, sólo se observa ruido para valores más altos de la variable.

Considerando esto se construye la variable Login App 2 con los valores:

- Igual a Login App, si éste va de 1 a 5.
- Igual a 5, para valores de Login App mayores a 5.

Con esto se busca una codificación más precisa que relacione de mejor forma la frecuencia de Login en App con la variable Y de etiqueta de los datos.

Observando la variable de Edad, también se muestra una mayor conversión de etiquetas positivas, para las personas que poseen entre 23 y 35 años. Esta conversión empieza a descender de forma gradual para valores posteriores a 35.

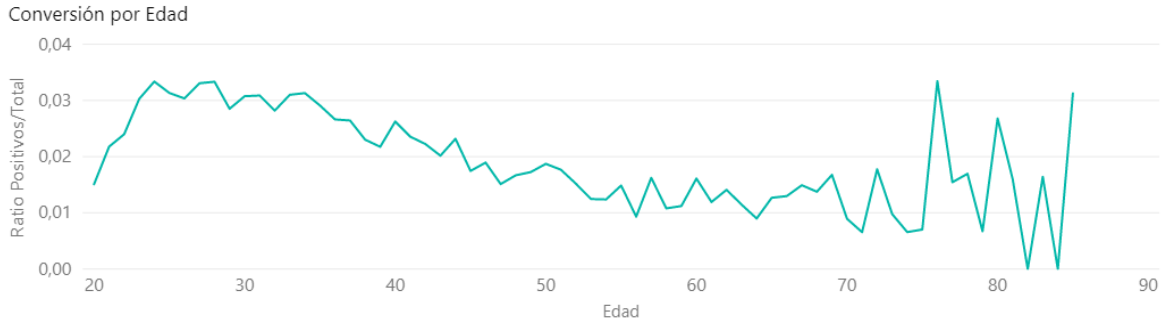


Ilustración 46: Ratio de etiquetas positivas v/s total de datos, por edad del cliente

En vista del gráfico anterior, se construye la variable binaria Bin Edad, la cuál es 1 cuando el valor de edad se encuentra entre 23 y 35 años. De esta forma, se intenta representar de mejor forma el efecto de la Edad, hacia la variable de etiqueta de clasificación.

A continuación, se construye la variable de ratio de Monto de Pagos en E-Commerce, con respecto a la Deuda que posee el cliente en el mes. Esta variable fue redondeada a la décima y su conversión de datos positivos se observa en el siguiente gráfico. Esta variable de ratio también se incluye en el modelo predictivo, como variable adicional.

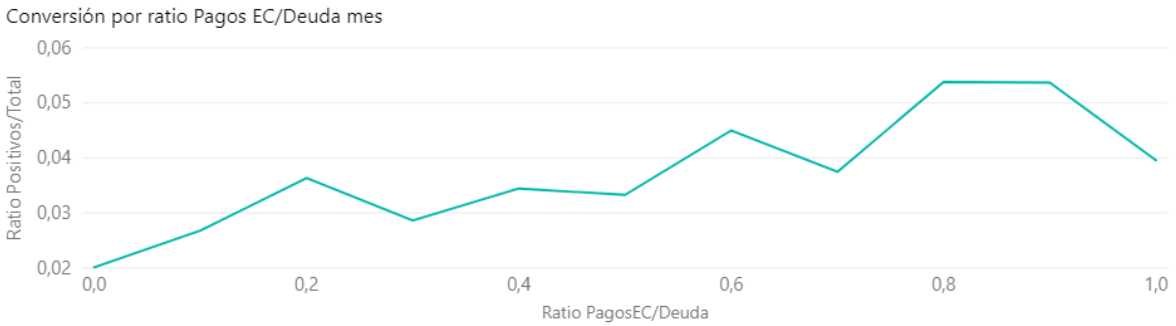


Ilustración 47: Ratio de etiquetas positivas v/s total de datos, por ratio de Pagos en E-Commerce/Deuda del mes

En adición a lo anterior, se agregan otras tres variables auxiliares, utilizando las variables de frecuencia de Hito de Cupos, de Estado de Cuenta y también observando la variable de Suma de Hitos en canales digitales. El detalle de la construcción de estas variables se ubica en el Anexo 7.

A continuación, se muestra el análisis exploratorio del segundo modelo, donde se consideran los clientes que no han tenido login en alguno de los canales digitales.

En primer lugar, se observa la cantidad de registros con etiqueta positiva y negativa. Se observa que aprox. 15.620 registros se convierten en digital al segundo mes, equivalente a un 1,56% del total.

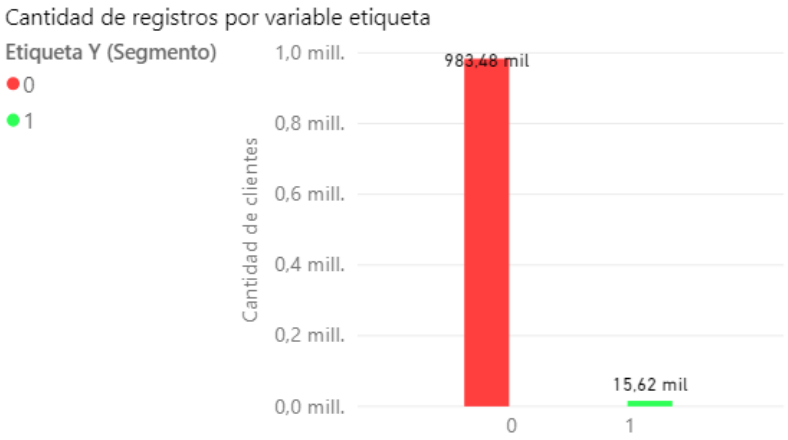


Ilustración 48: Cantidad de registros por variable etiqueta (2° modelo predictivo)

Por otra parte, los registros con etiqueta positiva poseen menor edad y son levemente más antiguos en promedio, con respecto a los datos con etiqueta negativa. El promedio de edad de los datos positivos es menor a 40, casi 10 años menor que el promedio de edad de los registros con etiqueta negativa.

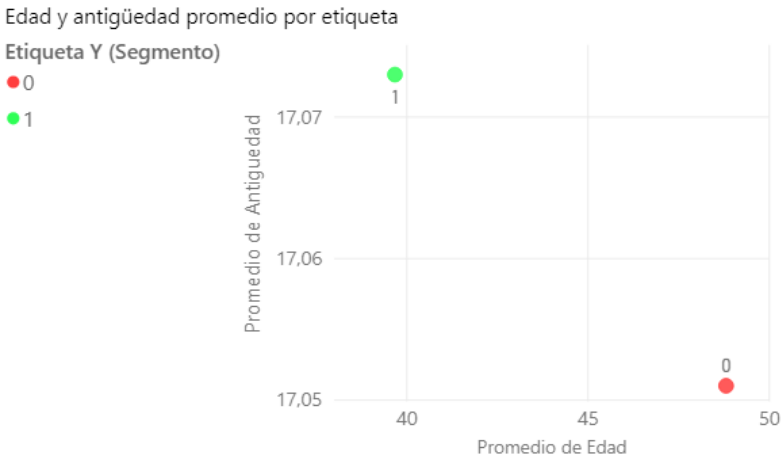


Ilustración 49: Promedio de edad y antigüedad, por variable etiqueta (2° modelo predictivo)

Además, al observar el cumplimiento de hitos, se muestran diferencias en los registros con etiqueta positiva, principalmente en la conversión del hito de Pagos en E-Commerce y en menor medida, en el hito de Transacciones en E-Commerce.

Etiqueta Y	% Hito PagosEC	% Hito TransacEC	% Hito ABCDIN
0	6,56 %	1,84 %	0,27 %
1	22,68 %	7,91 %	1,40 %

Ilustración 50: Porcentaje de cumplimiento de hitos, por variable etiqueta (2° modelo predictivo)

En adición a lo anterior, se realiza un análisis descriptivo de otras variables, como por ejemplo las transacciones y pagos promedio de los clientes, con etiqueta positiva o negativa en el segundo modelo. El detalle de estas observaciones se ubica en el Anexo 8.

8.3.1.3 Selección de Variables

Al ingresar los registros de los clientes, junto con la medición de sus atributos, el siguiente proceso es realizar la selección de variables para llevar al modelo predictivo. En esta etapa se busca reducir dimensionalidad, para hacer modelos más eficientes y sin un costo computacional tan alto. Por otro lado, se busca llevar una

cantidad de variables mínima, para que el modelo tenga la capacidad de predecir la variable dependiente con métricas aceptables.

Para realizar la selección de variables se utilizan los siguientes métodos de filtro hacia las variables:

- Análisis de correlación de variables independientes, con la variable dependiente: Se escogen variables con correlación sobre un 0,03 para el primer modelo, y sobre 0,02 para el segundo modelo. La lista de variables que superan este filtro se encuentra en el Anexo 9.
- Análisis de correlación entre variables independientes: Para los pares de variables que tengan un valor mayor a 0,8 en correlación, se elimina una de estas variables. Los gráficos de correlación, junto con las variables que superan este filtro se ubican en el Anexo 10.
- Análisis de ganancia de información entre variables independientes, frente a la variable dependiente: Para el primer modelo se eligen las variables que tengan un valor de ganancia de información mayor a 0,01. Para el segundo modelo se elimina sólo una variable, ya que ésta posee un indicador de ganancia de información mucho más bajo que las demás variables. El detalle de los gráficos que permiten escoger las variables se encuentra en el Anexo 11.

Considerando estos filtros de variables, las variables consideradas para realizar los algoritmos de predicción son las siguientes.

Para el primer modelo:

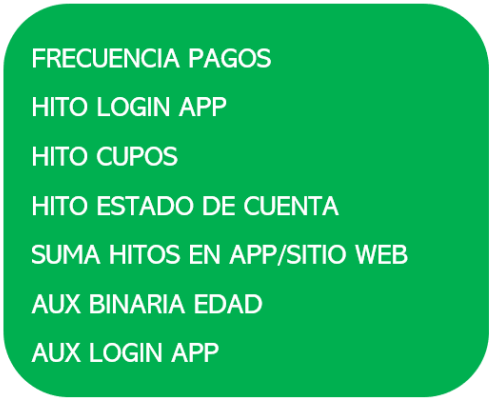


Ilustración 51: Variables seleccionadas para el primer modelo predictivo

Para el segundo modelo:

EDAD
MONTO TRANSACCIONES CON TARJETA EN E-COMMERCE
MONTO PAGOS CON TARJETA EN E-COMMERCE
HITO TRANSACCIONES EN E-COMMERCE
HITO PAGOS EN E-COMMERCE
CANAL SUSCRIPCIÓN EST CUENTA
BINARIA TRANSACCIONES

Ilustración 52: Variables seleccionadas para el segundo modelo predictivo

8.3.1.4 Resultados

Resultados de Primer Modelo Predictivo

Para el primer modelo propuesto, se realizan los 6 algoritmos de predicción mencionados anteriormente. Adicionalmente, se aplica el método SMOTE de oversampling y para cada uno de los 6 algoritmos, se realizan iteraciones con distintos valores de ratio de oversampling, es decir ratio entre valores positivos (que aumenta con el oversampling) y total de datos negativos (valor fijo). Con esto, se busca un parámetro o un rango de parámetros de oversampling que sean útiles para aplicar al modelo.

A continuación, se observa un gráfico donde se varía el ratio de oversampling. Utilizando el algoritmo logit de clasificación, se miden las métricas de predicción en testeo asociadas al modelo.

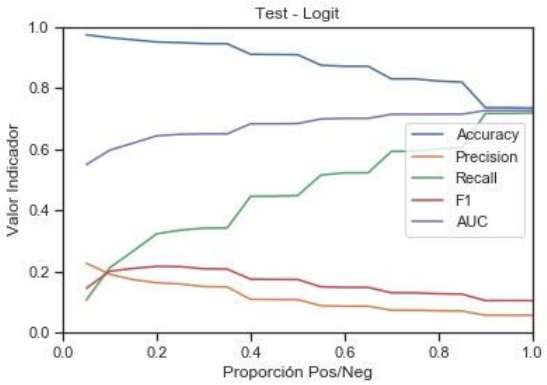


Ilustración 53: Métricas de ajuste en testeo, por ratio de oversampling, en 1er modelo predictivo

En el gráfico anterior, se muestra que para los valores de ratio de oversampling mayores a 0,9, se obtiene un mayor AUC y mayor valor de Recall, por lo que para un valor de ratio igual a 1, el modelo posee mejor desempeño. En el Anexo 12, se muestra el gráfico con los mismos ejes, para los algoritmos de Árbol de Decisión, Random Forest, Naive Bayes, SVM y KNN, los que respaldan que el valor de ratio igual a 1 es útil para el primer modelo de clasificación.

Utilizando un ratio de oversampling igual a 1, los resultados de las métricas de Accuracy, Precision y AUC son los siguientes:

Modelo	Accuracy Train	Accuracy Test	Precision Train	Precision Test	AUC Test
Logit	73,35%	73,49%	73,05%	5,6%	72,57%
Árbol de Decisión	74,22%	72,71%	73,14%	5,48%	72,39%
Random Forest	74,22%	72,76%	73,14%	5,49%	72,44%
Naive Bayes	70,77%	52,88%	64,88%	3,95%	70,97%
Support Vector Machine	72,63%	61,59%	68,26%	4,41%	71,47%
K-Nearest Neighbors	59,37%	95,99%	99,17%	16,48%	59,43%

Tabla 13: Métricas en testeo de 1er modelo predictivo, por algoritmo y con ratio de oversampling igual a 1

En el Anexo 13, se muestran los resultados de Accuracy, Recall y F1 para cada modelo utilizado. Los resultados muestran valores mayores a 70% en Accuracy en algunos algoritmos. También existen cifras de AUC mayores a 0,7.

Sin embargo, los valores de Precision en la base de testeo son extremadamente bajos. Debido a esto se realizan modificaciones en algunos parámetros, para intentar mejorar sus métricas de desempeño y en particular, del valor Precision.

Primero, se modifica el valor de threshold del modelo, es decir, la probabilidad límite donde un modelo diferencia una predicción positiva de una negativa. El valor de threshold por defecto es de 0,5 y con un valor mayor, se espera que existan menos predicciones positivas, pero más certeras, lo que tiende a aumentar la métrica Precision.

Con un valor de threshold de 0,8 se realiza la iteración sobre ratios de oversampling y con el modelo Logit se calculan las métricas de ajuste. El siguiente gráfico muestra que la métrica Precision no aumenta y por ende el modelo no mejora su desempeño.

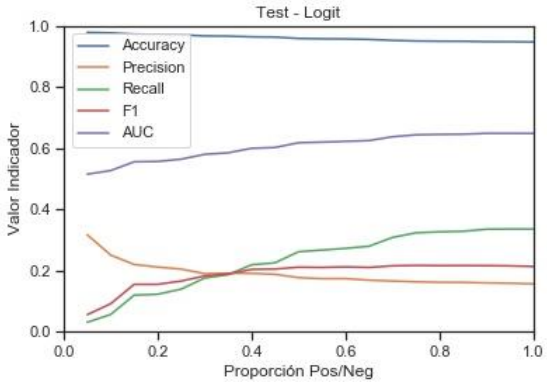


Ilustración 54: Métricas de ajuste en testeo en 1er modelo predictivo, por ratio de oversampling y con Threshold=0,8

Se proponen otras variaciones del modelo, para observar si existen mejoras en las métricas de desempeño. En primer lugar, se realiza un modelo con mayor cantidad de variables, que busque una mayor capacidad predictiva en los datos. Se agregan las variables "Bin Estado de Cuenta", "Bin Transacciones" y la de Hito de Pagos en E-Commerce, añadidas a las 7 variables del modelo anterior y sumando 10 variables en total. Se realiza el mismo gráfico de métricas de ajuste, variando el ratio de oversampling, donde se observa que no existen mejoras en la métrica de Precision.

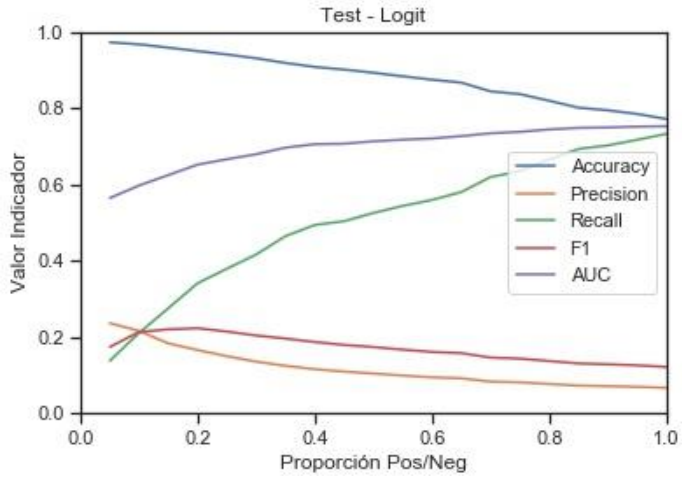


Ilustración 55: Métricas de ajuste en testeo en 1er modelo predictivo, por ratio de oversampling y con variación de cantidad de variables

Por último, se varía el método de sampling y en vez de SMOTE, se utiliza undersampling, técnica que disminuye la cantidad excesiva de etiquetas negativas

de los datos y que podría resultar en mejores métricas del modelo. Con undersampling, el gráfico de ratio de oversampling y métricas de desempeño es el siguiente:

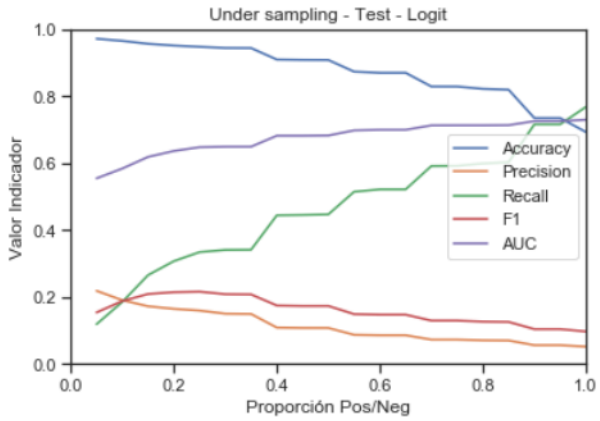


Ilustración 56: Métricas de ajuste en testeo, por ratio de undersampling, en 1er modelo predictivo

Con esta ilustración, tampoco se muestra un aumento en la métrica de Precision.

Se que deduce que las técnicas de oversampling y undersampling no son suficientes para generar un modelo confiable. El desbalance excesivo entre etiquetas positivas y negativas producen desajustes en las predicciones del modelo, donde se observan resultados de Accuracy aceptables debido al desbalance de datos y no por una capacidad predictiva del modelo utilizado.

Resultados de Segundo Modelo Predictivo

Para el segundo modelo de predicción se realiza un ejercicio similar al efectuado con el primer modelo. En primer lugar, se toman en cuenta los 6 algoritmos de clasificación propuestos y para cada uno de ellos se genera una iteración por diversos valores de ratio de oversampling. Con estos valores, se grafican las métricas de desempeño para cada modelo. En la siguiente ilustración, se muestra el gráfico descrito usando el modelo Logit, donde se observa que para valores de ratio de oversampling igual a 1, el modelo tiene mejores resultados de Recall y de AUC, sin embargo, su Accuracy y Precision disminuyen.

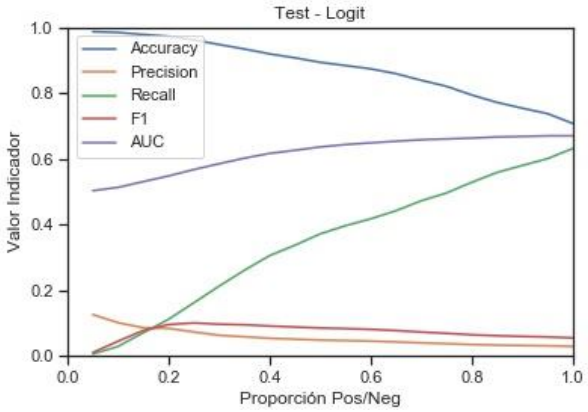


Ilustración 57: Métricas de ajuste en testeo, por ratio de oversampling, en 2° modelo predictivo

Se realizan gráficos de ratio de oversampling con los 5 algoritmos restantes, éstos se ubican en el Anexo 14. Estos gráficos reafirman que un valor de ratio de 1 es útil para aplicar al segundo modelo predictivo.

Utilizando un valor de ratio de oversampling igual a 1, se tienen los siguientes resultados de métricas de ajuste, para los 6 modelos de clasificación utilizados:

Modelo	Accuracy Train	Accuracy Test	Precision Train	Precision Test	AUC Test
Logit	69,5%	70,83%	69,87%	2,82%	66,93%
Árbol de Decisión	73,96%	75,45%	76,63%	2,52%	61,22%
Random Forest	73,72%	75,2%	76,37%	2,56%	61,76%
Naive Bayes	64,9%	86,14%	75,96%	4,04%	64,21%
Support Vector Machine	67,36%	54,88%	66,8%	2,22%	65,91%
K-Nearest Neighbors	69,78%	76,53%	76,58%	1,95%	55,53%

Tabla 14: Métricas en testeo de 2° modelo predictivo, por algoritmo y con ratio de oversampling igual a 1

Al igual que en el primer modelo predictivo, los resultados de Precision del modelo son extremadamente bajos, no mayores al 5% para los datos de testeo. Ésto indica la existencia de overfitting en el modelo, dado que existen resultados altos en los datos de entrenamiento y resultados muy bajos en la muestra de testeo. En el Anexo 15, se encuentran los resultados de Accuracy, Recall y F1 para cada uno de los modelos utilizados.

Dados los resultados de los dos primeros modelos, se genera un modelo alternativo que busque mejores métricas de desempeño y además, mantener el propósito de conversión de uso de los canales digitales por parte de los clientes de la tarjeta RF.

8.3.2 Tercer modelo predictivo

El tercer modelo del Trabajo de Memoria se presenta como un complemento a los modelos anteriores, debido a resultados insatisfactorios con respecto a la métrica de "Precision" que posee el modelo. En vista de las características de los modelos ya realizados, se buscan modificar 3 aspectos de los datos.

El primer aspecto por considerar es que en el primer y segundo modelo existen clientes que poseen datos muy similares entre un mes y otro, para un mismo cliente. Esto puede generar datos duplicados para el modelo y peor aún, con resultados distintos en su etiqueta. Por ello se construye sólo un registro por cliente y se evitan registros múltiples de clientes por cada mes.

La segunda característica para modificar es aumentar la ventana de tiempo de predicción, ya que realizar predicciones para un mes parece un evento muy inmediato y por ello se ha aumentado la ventana de tiempo a 4 meses, ya que se poseen los datos para ello y es una ventana lo suficientemente extensa para observar transformaciones de comportamiento digital, a diferencia de la ventana de 1 mes.

El tercer aspecto es establecer el hito al cual se le está haciendo el modelo de clasificación. Para los primeros modelos, el objetivo del cliente es transformarse de cliente no digital a digital. Sin embargo, cumplir con este objetivo de conversión es demasiado complicado y por ende, la proporción de datos con etiqueta positiva (variable Y igual a 1) es de un 2,48% y 0,15% para el primer y segundo modelo. Estas son cifras demasiado bajas para lograr un modelo de clasificación exitoso, pese a la ejecución de métodos de oversampling.

Considerando eso, en el tercer modelo la predicción por realizar es si el cliente no digital visita la aplicación o sitio web, 4 meses después de la medición de sus datos. Para este modelo, se toman en cuenta los datos del cliente en junio, considerando el cumplimiento de hitos y su frecuencia en el período de 5 meses (febrero-junio) y se observa el estado de actividad del cliente en el mes de octubre, donde se establece una etiqueta igual a 1 si el cliente visita alguno de los canales digitales (App/sitio web) o cero si es que no las visita en ese mes. Frente a esto, se busca aumentar la conversión de los registros y entregar al modelo las condiciones para generar mejores métricas.

8.3.2.1 Selección y Preprocesamiento de Datos

A diferencia del primer y segundo modelo construido (con datos de cada cliente por 3 meses distintos), el tercer modelo considera un registro por cada cliente, con los datos correspondientes solo al mes de junio. Por lo tanto, el primer filtro de datos se hace con el universo de clientes ocupado en los primeros modelos, de 433.152 datos.

Posteriormente, se efectúa la eliminación de outliers y datos erróneos. Aquí se hacen los mismos filtros que en el primer y segundo modelo predictivo. Estos son:

- Variable de Frecuencia "Login App": Se eliminan valores superiores a 77.
- Variable de Frecuencia "Cupos": No se consideran valores superiores a 6.
- Variable de Frecuencia "Est. Cuenta": Se eliminan valores superiores a 17.
- Variable de Frecuencia "Movimientos": No se toman en cuenta datos superiores a 11.
- Variable de Frecuencia "Ofertas": Se eliminan datos superiores a 95.
- Variable de Edad: Sólo se toman en cuenta personas entre 20 y 85 años. Se eliminan registros con datos faltantes, fuera de rango o outliers con más de 85 años.

Después de realizar estos filtros, los registros resultantes son 363.564 datos. Como filtro final para la selección de datos, el modelo requiere:

- Clientes no digitales.
- Clientes que si tengan etiqueta Y, si sigue visitando la app o sitio web 4 meses después.

Finalmente, 326.091 clientes cumplen con los requisitos mencionados y son llevados al modelo predictivo.

En relación con las variables ingresadas inicialmente al tercer modelo, se consideran las mismas variables del primer y segundo modelo predictivo, a las que se agregan también las 6 variables añadidas. después del análisis exploratorio del primer modelo.

Se considera un 80% de los datos para la muestra de entrenamiento y el 20% restante se lleva a los datos de testeo de la predicción.

8.3.2.2 Análisis exploratorio

A continuación, se observan las variables consideradas para el modelo, haciendo diferencias entre los registros con etiqueta positiva, donde los clientes si tienen login 4 meses después al registro y los registros negativos, donde los clientes ya no acceden a los canales digitales 4 meses después.

En cuanto a la cantidad de clientes, se muestran que aprox. 74.700 de ellos poseen etiqueta positiva, lo que equivale al 21,66% del total, mostrando un porcentaje de conversión mucho mayor que los dos modelos realizados anteriormente.

Ilustración 58: Cantidad de clientes, por variable etiqueta (3er modelo predictivo)

Por otro lado, se analiza la edad y antigüedad promedio de los grupos con etiqueta positiva y negativa. Esta variables son las más útiles para considerar en el análisis descriptivo y si existen diferencias importantes entre los grupos, pueden dar indicios de que influyen en el análisis predictivo posterior.

Al comparar ambos grupos, se muestra que el grupo con etiquetas positivas es levemente más joven y antiguo que el grupo con etiquetas negativas.

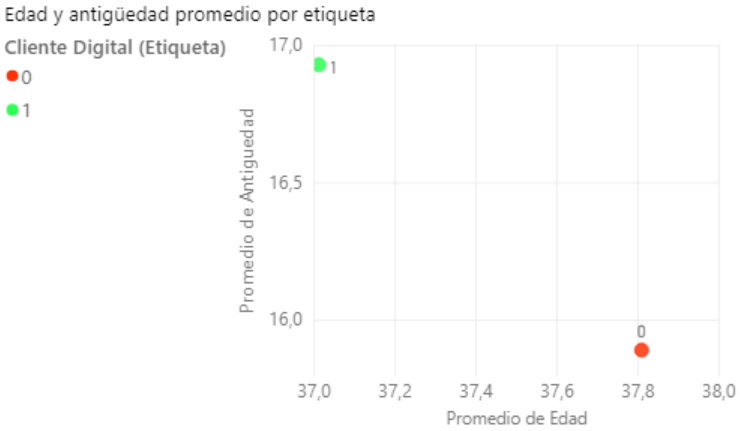


Ilustración 59: Promedio de edad y antigüedad, por variable etiqueta (3er modelo predictivo)

Finalmente, al observar el porcentaje de cumplimiento de hitos, junto con sus frecuencias asociadas, se muestra que los clientes con etiqueta positiva cumplen más los hitos de “Login App”, “Estado de Cuenta”, “Ofertas” y “Transacciones en E-Commerce”, en comparación con los registros con etiqueta negativa.

Etiqueta Y	% Hito Cupos	% Hito Est Cuenta	% Hito Login App	% Hito Login Plat	% Hito Movim	% Hito Ofertas	% Hito Pagos	% Hito PagosEC	% Hito TransacEC	% Hito ABCDIN
0	53,42 %	65,59 %	65,08 %	43,96 %	67,56 %	64,96 %	3,94 %	22,53 %	8,04 %	1,39 %
1	49,83 %	81,07 %	82,12 %	32,50 %	72,08 %	82,09 %	5,76 %	23,63 %	12,03 %	1,59 %

Tabla 15: Porcentaje de cumplimiento de hitos, por variable etiqueta (3er modelo predictivo)

Además, los clientes positivos cumplen con mayor frecuencia los hitos de “Estado de Cuenta”, “Login App” y “Ofertas”. Éstos mismos poseen una cantidad levemente mayor de promedio de suma de hitos, en los canales digitales.

Etiqueta Y	Promedio de Cupos	Promedio de Est cuenta	Promedio de Login_app	Promedio de Login_plat	Promedio de Movim	Promedio de Ofertas	Promedio de Pagos	Promedio de suma de hitos en App-plat
0	1,04	2,47	4,41	0,85	1,85	5,20	0,04	3,65
1	1,06	4,35	10,05	0,73	2,27	11,63	0,08	4,05

Tabla 16: Promedio de frecuencia de hitos, por variable etiqueta (3er modelo predictivo)

Frente al análisis exploratorio, se busca tener una primera noción de la existencia de diferencias entre los datos con etiqueta positiva y negativa, para observar si éstas se pueden diferenciar en el modelo predictivo. Además, se tiene un primer acercamiento hacia las variables que pueden ser consideradas para la selección de variables del modelo a realizar. En el Anexo 16, se muestran otros gráficos que forman parte del análisis exploratorio, como por ejemplo el análisis por género y nacionalidad, como también cifras promedio de transacciones y pagos de los clientes, según su valor etiqueta para el tercer modelo predictivo.

8.3.2.3 Selección de variables

Con relación al procedimiento de selección de variables de este modelo, se toman los mismos procesos de filtro de selección que el primer y segundo modelo.

- Análisis de correlación de variables independientes, con la variable dependiente: Se escogen las variables con correlación mayor a 0,3; con respecto a la variable dependiente. En el Anexo 17 se ubican los valores de indicadores de correlación, de las variables seleccionadas.
- Análisis de correlación entre variables independientes: Para los pares de variables con correlación mayor a 0,85; se elimina una de las variables involucradas. En el Anexo 18, se ubica el detalle de la matriz de correlación.

- Análisis de ganancia de información entre variables independientes, frente a la variable dependiente: Se escogen todas las variables, ya que todas poseen un valor mayor a 0,02; lo suficientemente alto para ingresar al modelo.

Las variables que superan los filtros mencionados son llevadas al modelo predictivo. Estas son:

EDAD	HITO LOGIN APP
FRECUENCIA LOGIN APP	HITO LOGIN SITIO WEB
FRECUENCIA CUPOS	HITO CUPOS
FRECUENCIA ESTADO DE CUENTA	HITO MOVIMIENTOS
FRECUENCIA MOVIMIENTOS	AUX BINARIA ESTADO DE CUENTA

Ilustración 60: Variables seleccionadas para 3er modelo predictivo

Estas 10 variables son ingresadas al tercer modelo predictivo, al cual se le aplican los mismos algoritmos de clasificación utilizados en el primer y segundo modelo predictivo.

8.3.2.4 Resultados

Se aplican los 6 algoritmos de clasificación propuestos, los cuales son Logit, Árbol de Decisión, Random Forest, Naive Bayes, SVM y KNN. Para cada uno de estos algoritmos se realiza una iteración, modificando los valores de ratio de Oversampling mediante la técnica de SMOTE, al igual que en los dos modelos anteriores.

La siguiente ilustración muestra los resultados de métricas de ajuste, para los distintos parámetros de ratio de oversampling aplicados en el modelo Logit.

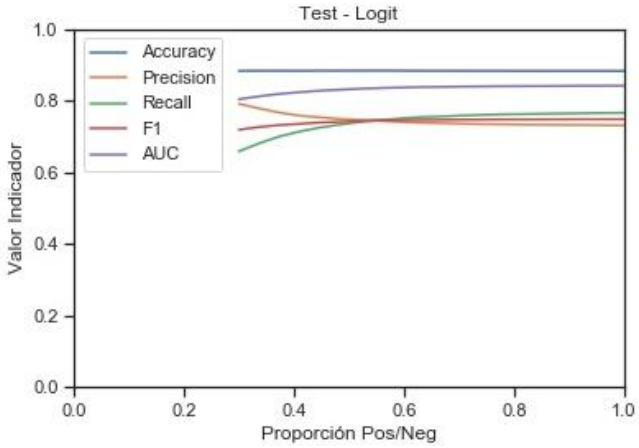


Ilustración 61: Métricas de ajuste en testeo, por ratio de oversampling, en 3er modelo predictivo

Además, se muestran en el Anexo 19 los resultados de ajuste en la misma iteración, utilizando los 5 modelos restantes.

Con el parámetro de oversampling fijo, los resultados de las métricas de ajuste, para cada algoritmo de clasificación, son los siguientes:

Modelo	Accuracy Train	Accuracy Test	Precision Train	Precision Test	AUC Test
Logit	83,86%	88,22%	90,02%	73,05%	84,09%
Árbol de Decisión	85,47%	86,8%	94,48%	73,21%	79,56%
Random Forest	85,36%	87,5%	93,79%	73,81%	81,29%
Naive Bayes	81,94%	88,27%	91,15%	75,88%	82,19%
Support Vector Machine	83,37%	88,2%	89,78%	72,86%	84,16%
K-Nearest Neighbors	83,87%	88,21%	93,07%	74,24%	83,14%

Tabla 17: Métricas en testeo de 3er modelo predictivo, por algoritmo y con ratio de oversampling igual a 1

Como se observa en la ilustración y tabla anterior, los resultados del tercer modelo predictivo logran mostrar métricas elevadas de ajuste, a diferencia del primer y segundo modelo predictivo. El algoritmo que posee mayor AUC es el modelo SVM, por su parte el modelo Naive Bayes tiene las mejores métricas de Accuracy y Precision en la base de testeo. En el Anexo 20, se ubican los resultados de las métricas de Accuracy, Recall, F1 asociadas a cada algoritmo utilizado.

9 Conclusiones

Dentro de las conclusiones asociadas al trabajo de Memoria, se destacan los siguientes puntos.

En cuanto al primer modelo de segmentación, se identifican claramente distintos segmentos de los clientes de la tarjeta RF, en base a sus hitos digitales. Estos segmentos también se pueden caracterizan y distinguir, en base a otras variables como edad, cantidad de hitos digitales y monto de transacciones y pagos con tarjeta. Por ende, se cumplen los dos primeros objetivos específicos planteados inicialmente.

No obstante, se recalca que se considera un historial de 6 meses para el modelo de segmentación, identificando sólo al mes de julio los diversos grupos de clientes. Por lo tanto, en una tentativa segmentación futura, al incorporar datos de los meses siguientes, podría variar el número óptimo de clusters escogido y también la caracterización de estos segmentos.

Con relación al segundo modelo de Diferencias en Diferencias, se destaca que existe significancia estadística de los coeficientes asociados a la variable DxT, que identifica el efecto de la digitalización de los clientes de la tarjeta RF. Para los modelos estimados, el cambio en transacciones es positivo para la ventana de tiempo de 11 meses. Por su parte, el cambio en pagos con la tarjeta RF es negativo, por lo que se deduce que la digitalización no produce una mejora en los pagos de los clientes de la compañía. Es necesario para el futuro estudiar las causas de este fenómeno y también pensar en posibles soluciones para mejorar estos indicadores.

A pesar de la existencia de resultados estadísticamente significativos, debido a la magnitud del número de clientes, el análisis sólo incluye una ventana de tiempo de los últimos 11 meses, por lo que no se puede concluir que estos efectos son perdurables en el tiempo y extrapolables hacia otros meses. Por ello, es necesario agregar otros meses al análisis para verificar los resultados obtenidos en este primer modelo. Sólo con eso se podrá estimar una cuantificación certera, del beneficio en transacciones y pagos de los clientes digitales.

Asociando el segundo modelo realizado con los objetivos presentados, se logran establecer criterios de digitalización de un cliente, considerando el uso de los canales digitales. Sin embargo, no es posible concluir con total seguridad el efecto de la digitalización en transacciones y pagos, debido a los puntos expuestos en el párrafo anterior.

Con relación al tercer procedimiento del trabajo de Memoria, los modelos predictivos generados poseen métricas aceptables de Accuracy, mayores al 70% en

su mayoría. Sin embargo, para el primer y segundo modelo generado, los resultados de Precision y Recall son insatisfactorios. Si bien se realizaron modificaciones en los parámetros del primer modelo, éstos no mejoraron su desempeño. Por esto, se deduce se generan métricas poco confiables debido al desbalance de las etiquetas positivas y negativas de la variable dependiente y en segundo lugar, por la dificultad para que las variables explicativas generen una predicción certera para todos los datos.

El tercer modelo predictivo si posee métricas aceptables, con resultados de Precision sobre el 70%. Considerando esto, se infiere que las modificaciones realizadas al modelo predictivo son recomendables para futuros modelos predictivos. En este caso, las recomendaciones apuntan a hacer una variable de clasificación más fácil de cumplir, también aumentar la ventana de predicción y considerar sólo un dato por cliente.

Asociando los modelos predictivos con los objetivos específicos propuestos, se logra estimar un grado de propensión de transformación de clientes no digitales a digitales. Pero este modelo no es 100% confiable, considerando los resultados obtenidos y las modificaciones explicadas en los dos párrafos anteriores. En relación con los lineamientos para un plan de digitalización, éstos se abordan dentro de las recomendaciones y trabajo futuro, donde se enlazan los resultados y conclusiones para entregar recomendaciones hacia la compañía, en términos comerciales.

En relación con el objetivo general expuesto, se cumple la definición de un cliente digital para la empresa. No obstante, en base al análisis realizado sobre los clientes de la compañía, en las tres etapas de este trabajo no se pueden extraer conclusiones permanentes para el negocio y el objetivo no se cumple en su totalidad.

Se concluye finalmente, que este trabajo es útil para obtener un primer resultado de los modelos ejecutados, pero es necesario aumentar la cantidad de meses de disponibilidad de datos, para verificar si estos resultados obtenidos se mantienen y agregan valor al negocio de esta compañía de retail financiero.

10 Recomendaciones y trabajo futuro

Desde los resultados y conclusiones que se generan en el presente trabajo, se extraen diversas recomendaciones para el trabajo futuro de la compañía de retail financiero.

En relación con el primer modelo de segmentación:

En primer lugar, las recomendaciones para la segmentación de clientes se basan en reiterar el modelo de segmentación, para actualizar los segmentos de clientes existentes. Frente a esto, la idea es observar si se mantiene la cantidad de segmentos y su caracterización, comparando indicadores como el tamaño del segmento, edad promedio o monto de transacciones y pagos.

Con relación a los hitos digitales, se aconseja agregar las nuevas funcionalidades que existan en la aplicación y sitio web asociado a RF. Sin embargo, deben existir varios meses de historia de datos, 6 meses en este caso, para incluir un hito en el modelo de segmentación y lograr que sea comparable con los otros hitos digitales, en el nuevo modelo. Como ejemplo claro de una futura funcionalidad que podría existir, se destaca la opción de hacer transacciones con la tarjeta RF, por medio de los canales digitales, acción inexistente hasta el momento y que sería muy útil para los clientes de RF.

Considerando las campañas de marketing hacia los clientes de la compañía, se recomienda considerar los diversos perfiles construidos y generar planes de marketing diferenciados para cada segmento, con el objetivo de generar un trato más personalizado con el cliente y también perseguir diferentes objetivos comerciales con los clientes, dado el segmento al que pertenecen. Por ejemplo, en un segmento “Full App” o “Full Digital” se tiene un objetivo de retención en el uso de canales digitales y en un segmento “No Digital” se tiene un objetivo de conversión o iniciación del cliente, hacia la utilización de app y plataforma respectivo.

Como última recomendación hacia el modelo de segmentación, se aconseja modificar la ventana de tiempo de activación de un hito digital. Para este caso se analizan los últimos 6 meses, pero también se podrían agregar estas mismas variables de hitos en un período más reducido de tiempo, por ejemplo, si ha realizado el hito en el último mes, o en los últimos 3 meses. Así, se busca tener una mayor cantidad de variables y poseer información más completa para una futura caracterización de segmentos. Por ejemplo, se podría generar un segmento de clientes que están dejando los canales digitales, si es que lo usaban hace 6 meses y si en el último mes ya no tuvieron acciones digitales.

En relación con el segundo modelo de Diferencias en Diferencias:

En cuanto al segundo modelo realizado, dado el resultado de diferencia negativa en los pagos de los clientes, después de la digitalización, se recomienda buscar causas que expliquen y soluciones que mejoren la situación de disminución de pagos de los clientes digitales, luego de cumplir su proceso de digitalización. La recomendación se dirige a abordar este tema en conjunto con el área comercial, área de riesgo y cobranzas y la idea es producir diversas acciones con el cliente, que impidan mantener una deuda impaga con la compañía RF. Ejemplos de estas acciones serían disminuir la cantidad de clientes morosos, impedir que los clientes en mora realicen acciones con la tarjeta RF, mejorar la comunicación con el cliente para que recuerde realizar los pagos de deuda con la tarjeta y generar una relación agradable hacia el cliente, que lo atraiga a pagar sus deudas en los canales digitales y que no lo inste a no realizar sus pagos por canales digitales y/o físicos.

Además, para el modelo de diferencias en diferencias, se recomienda realizar un nuevo análisis con mayor cantidad de meses, para verificar si existen las mismas diferencias de transacciones y pagos encontradas en el presente trabajo. Con esto se tendrá una cuantificación más certera del beneficio en transacciones y pagos que genera la digitalización de un cliente de la tarjeta RF. Se enfatiza que, para realizar un modelo con más de 1 mes de análisis, se debe extender la definición de cliente digital. Debido a que en cada mes, el cliente pertenece a un estado de cliente digital o no digital, lo que genera transiciones de estos clientes en los diversos meses de análisis del modelo.

En relación con los modelos predictivos realizados:

En relación con los resultados obtenidos en estos modelos, primero se recomienda que en la definición del modelo la situación a predecir no sea tan difícil de cumplir, para que no exista un desbalance excesivo entre la clase positiva y la clase negativa de la variable etiqueta a predecir. Con ello, se evitarán métricas bajas y poco confiables, como una baja precisión del primer y segundo modelo predictivo.

En segundo lugar, se aconseja aumentar la cantidad de meses de extracción de datos. Para el primer modelo predictivo se consideraron 4 meses de análisis, pero se recomienda ir agregando cada mes e ir actualizando la predicción del modelo.

Como tercera recomendación, se busca aumentar la cantidad de variables del modelo, en busca de un mejor ajuste futuro. Con esto, se recomienda agregar nuevas variables con una construcción más sofisticada, como por ejemplo variables conjuntas de interacción entre dos o más variables del modelo. También otra recomendación sería agregar variables que reflejen cambios entre un mes y otro,

como por ejemplo, el porcentaje de aumento de transacciones con respecto al mes anterior, o el aumento de pagos, o también cambios del número de hitos cumplidos, en comparación con el mes pasado.

Como cuarta recomendación hacia los modelos de clasificación, se aconseja modificar los parámetros de los algoritmos a utilizar, por ejemplo, la cantidad de árboles en Random Forest, el tipo de kernel en SVM, o la cantidad de vecinos en KNN. Con esto, se puede encontrar el mejor modelo para cada algoritmo utilizado y mejorar la capacidad predictiva del modelo generado.

Con relación a las campañas de marketing presentes y el enfoque final de un modelo predictivo, se recomienda elaborar planes de marketing diferenciados, considerando la propensión de cada cliente a digitalizarse. Específicamente, se recomienda dividir a los clientes en un número específico de grupos, preferentemente de 2 a 5 por temas de simplicidad, en base a los niveles de propensión a digitalizarse. Por ejemplo, generar una división de 3 grupos con clientes de propensión alta, media y baja. El objetivo principal de esta sugerencia es mejorar los índices de digitalización de los clientes y también generar datos para un futuro modelo econométrico, donde se pueda estudiar el efecto de las campañas de marketing, para cada grupo de propensión a digitalizarse.

Finalmente, en vista del trabajo realizado y el contexto del negocio financiero de la compañía, la recomendación es mantener el objetivo de conversión y fidelización de clientes hacia los canales digitales, ya que las conductas por estos medios son más rápidas y cómodas para el cliente y menos costosas para la compañía. Con esto se pretende que la aplicación y sitio web de RF sean los canales más utilizado por los clientes, para realizar las diversas acciones con la tarjeta asociada.

11 Bibliografía

[1] Retail Financiero. Historia del Retail en Chile. [En línea]. <http://retailfinanciero.org/quienes-somos/historia/>. [Consulta: 17 agosto 2019]

[2] Banco Central de Chile. 2018. Evolución de los Medios de Pago en Chile y su Incidencia en el Comportamiento de los Componentes de M1. Junio 2018. [En línea]. <http://si2.bcentral.cl/public/pdf/estudios-economicos-estadisticos/pdf/see125.pdf>. [Consulta:17 agosto 2019]

[3] El Mercurio. 3 de abril 2017. Economía y Negocios. [En línea]. <https://www.elmercurio.com/Inversiones/Noticias/Acciones/2017/04/03/Negocio-financiero-represento-un-tercio-de-las-ganancias-obtenidas-por-los-actores-del-retail-en-2016.aspx>. [Consulta: 16 diciembre 2019]

[4] América Retail. 2019. Chile: El incremento de las ventas online en nuestro país. [En línea]. <https://www.america-retail.com/chile/chile-el-incremento-de-las-ventas-online-en-nuestro-pais/>. [Consulta: 16 diciembre 2019]

[5] El Mercurio. 8 enero 2019. Inversiones. [En línea]. <https://www.elmercurio.com/Inversiones/Noticias/Analisis/2019/01/08/Falabella-vuelve-a-tener-crecimiento-en-sus-inversiones.aspx>. [Consulta: 16 diciembre 2019]

[6] SERNAC. Diciembre 2018. Tasa de respuesta desfavorable en tarjetas de crédito relacionadas al retail, pág 13. [En línea]. <https://www.sernac.cl/portal/619/articles-55160_archivo_01.pdf>. [Consulta: 23 agosto 2019]

[7] SCAN. 2018. Tarjetas con operaciones a marzo 2018. [En línea]. <https://www.scan.cl/2018/06/25/tras-salida-de-cmr-y-presto-la-banca-pasara-a-controlar-el-94-de-las-tarjetas-de-credito/>. [Consulta: 23 agosto 2019]

[8] HBR. 2012. Data scientist: The sexiest job of the 21st century. [En línea]. <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>. [Consulta: 25 agosto 2019]

[9] Data Science Central. 2017. Difference between Machine Learning, Data Science, AI, Deep Learning and Statistics. [En línea]. <https://www.datasciencecentral.com/profiles/blogs/difference-between-machine-learning-data-science-ai-deep-learning>. [Consulta: 25 agosto 2019]

[10] Github. 2018. Super cheatsheet: Machine Learning. [En línea]. <https://github.com/afshinea/stanford-cs-229-machine-learning/blob/master/en/super-cheatsheet-machine-learning.pdf>. [Consulta: 25 agosto 2019]

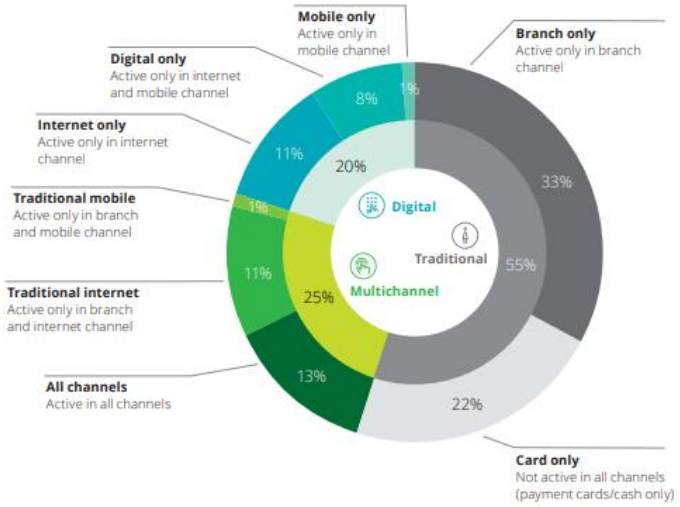
[11] HUANG Z. 1998. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. [En línea]. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.15.4028&rep=rep1&type=pdf>. [Consulta: 13 octubre 2019]

[12] HUANG Z. 1997. Clustering large data sets with mixed numeric and categorical values. [En línea]. <https://pdfs.semanticscholar.org/d42b/b5ad2d03be6d8fefa63d25d02c0711d19728.pdf>. [Consulta: 13 octubre 2019]
[13] SARKAR T. 2019. Clustering metrics better than the elbow-method. [En línea]. <https://towardsdatascience.com/clustering-metrics-better-than-the-elbow-method-6926e1f723a6>. [Consulta: 6 enero 2020]
[14] Deloitte. 2018. CEE PSD2 Survey. Voice of the customer. [En línea]. <https://www2.deloitte.com/content/dam/Deloitte/cz/Documents/financial-services/Deloitte_CEE_PSD2_Voice_of_the_Customer_Survey_012018_Short.pdf>. [Consulta: 26 agosto 2019]
[15] McKinsey. 2017. The future of customer-led retail banking distribution. [En línea]. <https://www.mckinsey.com/~/media/mckinsey/industries/financial%20services/our%20insights/the%20future%20of%20customer%20led%20retail%20banking%20distribution/the-future-of-customer-led-retail-banking-distribution-2017.ashx>. [Consulta: 26 agosto 2019]
[16] Cognizant. 2013. Segment-based strategies for mobile banking. [En línea]. <https://www.cognizant.com/InsightsWhitepapers/Segment-Based-Strategies-for-Mobile-Banking.pdf>. [Consulta: 26 agosto 2019]
[17] SIDANA M. 2017. Types of classification algorithms in Machine Learning. [En línea]. <https://medium.com/@Mandysidana/machine-learning-types-of-classification-9497bd4f2e14>. [Consulta: 13 octubre 2019]
[18] GAHUKAR G. 2018. Classification Algorithms in Machine Learning. [En línea]. <https://medium.com/datadriveninvestor/classification-algorithms-in-machine-learning-85c0ab65ff4>. [Consulta: 13 octubre 2019]
[19] HALL M. 1999. Correlation-based Feature Selection for Machine Learning [En línea]. <https://www.cs.waikato.ac.nz/~mhall/thesis.pdf>. [Consulta: 16 diciembre 2019]
[20] FARRAR D.E., GLAUBER R.R. 1964. Multicollinearity in Regression Analysis [En línea]. <https://dspace.mit.edu/bitstream/handle/1721.1/48530/multicollinearit00farr.pdf?sequence=1>. [Consulta: 16 diciembre 2019]
[21] Scholarpedia. 2009. Mutual Information. [En línea]. <http://www.scholarpedia.org/article/Mutual_information>. [Consulta: 16 diciembre 2019]
[22] KRASKOV A., STOGBAUER H., GRASSBERGER P. 2003. Estimating Mutual Information. [En línea]. <https://arxiv.org/abs/cond-mat/0305641>. [Consulta: 16 diciembre 2019]
[23] CHAWLA N., BOWYER K., HALL L., KEGELMEYER W.P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. [En línea]. <https://arxiv.org/pdf/1106.1813.pdf>. [Consulta: 16 diciembre 2019]

[24] ZIAFAT H., SHAKERI M. 2014. Using data mining techniques in customer segmentation. [En línea]. <https://www.ijera.com/papers/Vol4_issue9/Version%203/K49037079.pdf>. [Consulta: 26 agosto 2019]
[25] KUMAR S., KUMAR R. 2015. Analyzing Customer Behaviour through Data Mining.
[En línea].
<https://pdfs.semanticscholar.org/55e2/60f9a641e6df89b201aef5d87c9fef862892.pdf>. [Consulta: 26 agosto 2019]
[26] PAHWA B., TARUNA S., KASLIWAL N. 2017. Role of Data mining in analyzing consumer's online buying behavior. [En línea]. <https://www.ijbmi.org/papers/Vol(6)11/Version-3/G0611034551.pdf>. [Consulta: 26 agosto 2019]
[27] CARD D., KRUEGER A. 1994. Minimum Wages and Employment: A case study of the Fast-food industry in New Jersey and Pennsylvania. <http://davidcard.berkeley.edu/papers/njmin-aer.pdf>. [Consulta: 28 diciembre 2019]
[28] ATHEY S., IMBENS G. 2006. Identification and Inference in Nonlinear Difference-in-Differences Models. [En línea]. <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0262.2006.00668.x>. [Consulta: 14 noviembre 2019]
[29] Secretaría General de Gobierno de Chile. 2019. Estimación de personas extranjeras residentes en Chile. [En línea]. <https://msgg.gob.cl/wp/2019/02/14/ministro-s-de-interior-e-informe-estimacion-de-personas-extranjeras-residentes-en-chile-permite-visibilizar-a-la-poblacion-de-migrantes-y-construir-politicas-publicas-que-se-requi/>. [Consulta: 3 enero 2020]

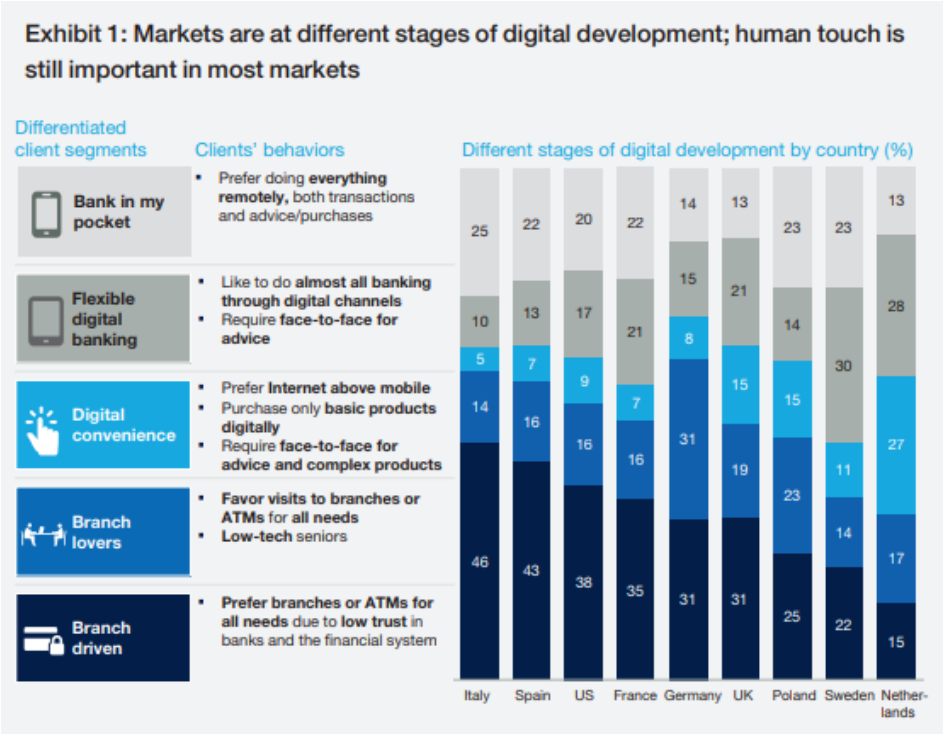
12 Anexos

Anexo 1: Segmentación de clientes de banca online en Europa Central y del Este, realizada por Deloitte



Fuente: Deloitte CEE PSD2 Survey 2018

Anexo 2: Segmentación de clientes de banca online, realizada por McKinsey en el marco de un artículo sobre retail financiero



Fuente: McKinsey Retail Banking Multichannel Survey 2016

Anexo 3: Resultados modelo de clustering con 4 segmentos, con cumplimiento de hitos por cada segmento

Segmentacion 1	% Cupos	% Est cuenta	% Login app	% Login plat	% Movim	% Ofertas	% Pagos app	% ABCDIN EC	% Pagos EC	% Transac EC
1	0,35	1,10	6,12	0,14	0,25	6,11	0,25	0,63	6,08	2,51
2	25,21	95,26	100,00	6,38	73,30	100,00	8,83	2,44	9,16	12,70
3	99,98	61,21	23,41	99,22	99,16	23,37	3,19	4,10	42,09	13,72
4	96,46	98,17	100,00	34,29	94,42	100,00	60,63	5,45	23,12	32,51

Anexo 4: Resultados modelo de clustering con 5 segmentos, con cumplimiento de hitos por cada segmento

Segmentación 2	% Cupos	% Est cuenta	% Login app	% Login plat	% Movim	% Ofertas	% Pagos app	% ABCDIN EC	% Pagos EC	% Transac EC
1	0,19	0,01	4,40	0,14	0,01	4,39	0,13	0,62	6,05	2,45
2	27,63	88,29	100,00	8,24	59,64	99,99	11,01	2,17	10,34	10,09
3	99,98	61,11	19,12	99,77	99,29	19,08	2,41	4,15	42,57	13,83
4	41,53	99,88	100,00	12,02	93,41	100,00	4,47	3,15	8,80	18,92
5	100,00	97,48	100,00	37,59	92,68	100,00	76,40	5,89	26,79	34,85

Anexo 5: Resultados modelo de clustering con 6 segmentos, con cumplimiento de hitos por cada segmento

Segmentacion 3	% Cupos	% Est cuenta	% Login app	% Login plat	% Movim	% Ofertas	% Pagos app	% ABCDIN EC	% Pagos EC	% Transac EC
1	0,02	0,01	0,01	0,14	0,01	0,00	0,00	0,60	6,09	2,41
2	11,92	37,39	100,00	2,71	10,67	99,98	7,20	1,38	7,63	5,44
3	25,59	98,34	100,00	6,79	79,75	100,00	9,81	2,47	9,55	12,52
4	99,98	62,00	20,54	99,60	99,38	20,50	2,44	4,13	42,27	13,70
5	69,33	99,94	100,00	23,03	98,76	100,00	13,27	3,84	12,73	23,44
6	100,00	96,75	100,00	37,53	90,91	100,00	86,89	6,14	28,39	36,19

Anexo 6: Output modelo Diff-In-Diff, donde se observan los coeficientes asociados a las diversas variables y también se muestra la significancia del coeficiente DxT. En los 4 modelos, la variable DxT no incluye al cero en su intervalo de confianza (cifras a la derecha de los valores entregados en el modelo), por lo que existe significancia estadística.

Modelo 1: Estimación de transacciones, sin incluir variables anexas

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.763e+04	126.857	296.623	0.000	3.74e+04	3.79e+04
Clientes Digitales	2.964e+04	392.757	75.460	0.000	2.89e+04	3.04e+04
Tiempo	-1.566e+04	178.846	-87.543	0.000	-1.6e+04	-1.53e+04
DxT	3581.8604	553.572	6.470	0.000	2496.877	4666.844

Modelo 2: Estimación de pagos, sin incluir variables anexas

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.772e+04	90.581	416.461	0.000	3.75e+04	3.79e+04
Clientes Digitales	-7657.2282	280.444	-27.304	0.000	-8206.889	-7107.567
Tiempo	-13.5083	127.703	-0.106	0.916	-263.803	236.786
DxT	-9548.8140	395.272	-24.158	0.000	-1.03e+04	-8774.093

Modelo 3: Estimación de transacciones, incluyendo variables anexas

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.63e+04	413.736	136.070	0.000	5.55e+04	5.71e+04
Nacionalidad_Chileno_Extranjero	-9811.6444	284.041	-34.543	0.000	-1.04e+04	-9254.933
Edad	-140.8611	6.794	-20.734	0.000	-154.177	-127.545
Antiguedad0	-213.9340	11.606	-18.434	0.000	-236.681	-191.187
Genero	-736.6527	170.088	-4.331	0.000	-1070.019	-403.286
Casado	-587.7850	202.346	-2.905	0.004	-984.376	-191.194
Clientes Digitales	2.703e+04	397.866	67.933	0.000	2.62e+04	2.78e+04
Tiempo	-1.564e+04	178.555	-87.596	0.000	-1.6e+04	-1.53e+04
DxT	3617.0340	552.661	6.545	0.000	2533.837	4700.231

Modelo 4: Estimación de pagos, incluyendo variables anexas

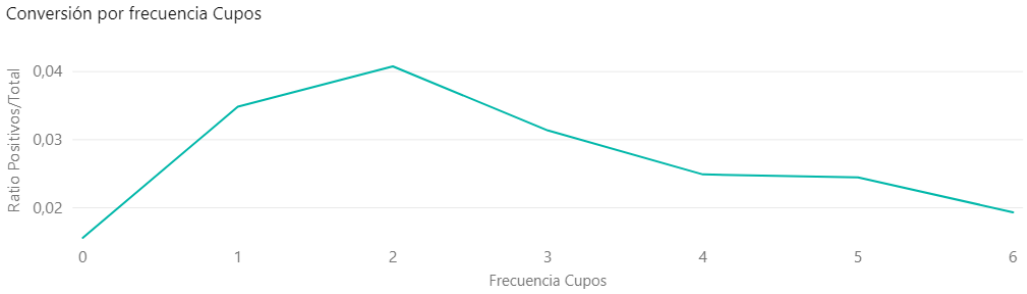
	coef	std err	t	P>|t|	[0.025	0.975]
const	2.54e+04	293.015	86.695	0.000	2.48e+04	2.6e+04
Nacionalidad_Chileno_Extranjero	-9634.6604	201.163	-47.895	0.000	-1e+04	-9240.388
Edad	215.2076	4.812	44.727	0.000	205.777	224.638
Antiguedad0	869.6221	8.219	105.802	0.000	853.512	885.732
Genero	-1114.6025	120.459	-9.253	0.000	-1350.699	-878.506
Casado	-150.3725	143.305	-1.049	0.294	-431.245	130.500
Clientes Digitales	-5039.7003	281.776	-17.885	0.000	-5591.972	-4487.429
Tiempo	-96.2207	126.456	-0.761	0.447	-344.070	151.628
DxT	-9595.3979	391.404	-24.515	0.000	-1.04e+04	-8828.258

Anexo 7: Creación de 3 variables auxiliares.

Anexo 7.1, 1ª variable auxiliar: Bin Cupos

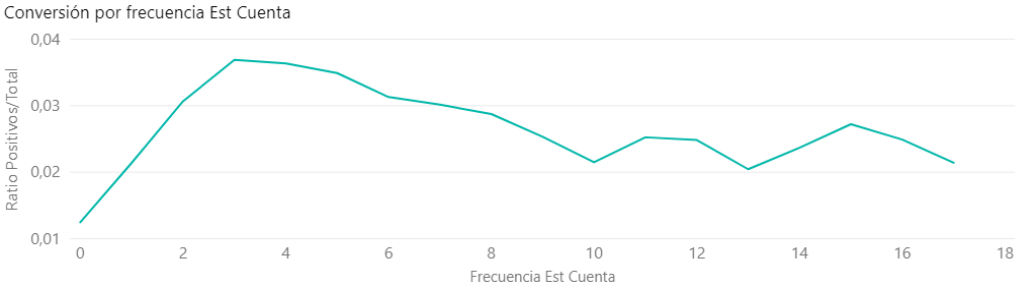
Bin Cupos es una variable binaria. Observando el gráfico adjunto, se muestra que para los valores de frecuencia de cupos de 1, 2 y 3, se tiene una mejor conversión de

la variable etiqueta, por lo que en esos valores se identifica como 1 la variable Bin Cupos. A continuación, se muestra el gráfico de frecuencia de hito de cupos, con su porcentaje de conversión respectivo.



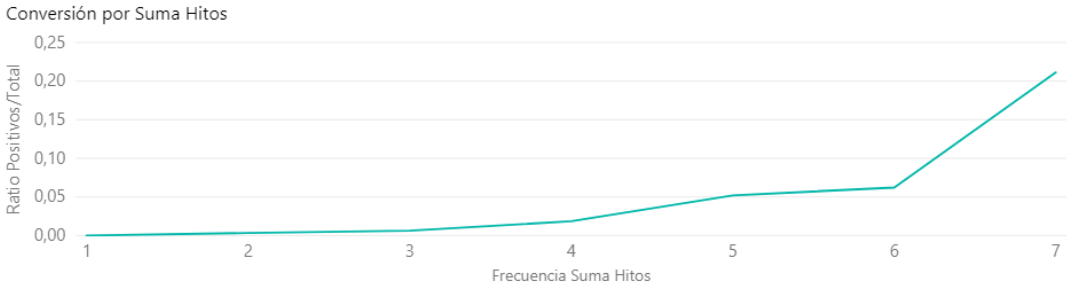
Anexo 7.2, 2ª variable auxiliar: Bin Estado de cuenta

Mediante el gráfico adjunto, se muestra que para los valores de frecuencia del hito "Estado de Cuenta", entre 3 y 7 se observa un mayor porcentaje de conversión. Por ello la variable binaria "Bin Estado de cuenta" es 1 si está entre esos valores en la variable original del hito.



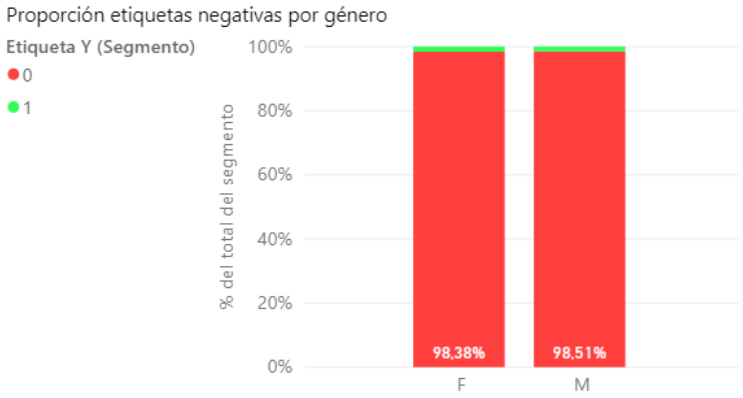
Anexo 7.3, 3ª variable auxiliar: Bin Suma Hitos

Esta variable es 1 cuando la suma de hitos es mayor a 4, esto se construye así ya que en el gráfico adjunto, se observa que los porcentajes de conversión de la etiqueta positiva son mayores cuando la cantidad de hitos es mayor a 4.

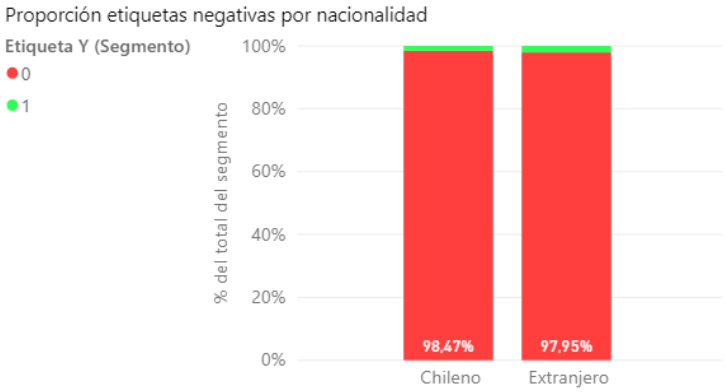


Anexo 8: Análisis exploratorio 2º modelo predictivo

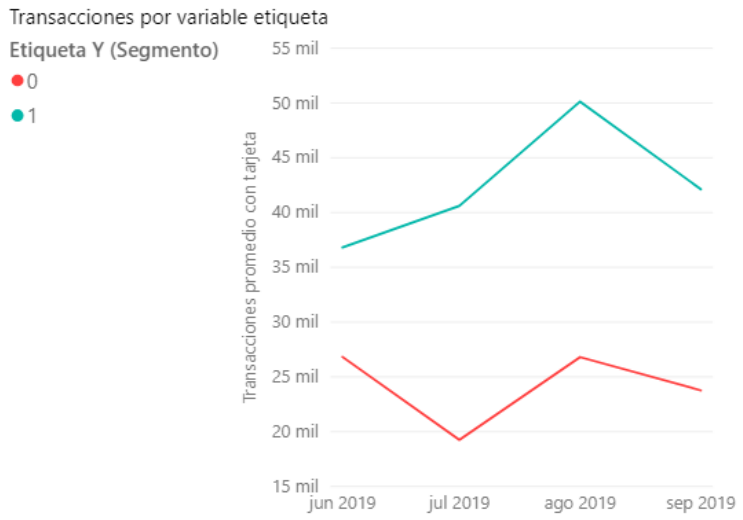
Anexo 8.1: Proporción etiquetas negativas con respecto al total de datos, por género



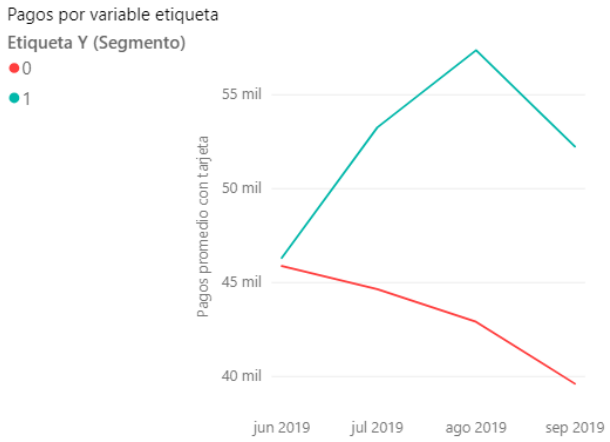
Anexo 8.2: Proporción etiquetas negativas con respecto al total de datos, por nacionalidad



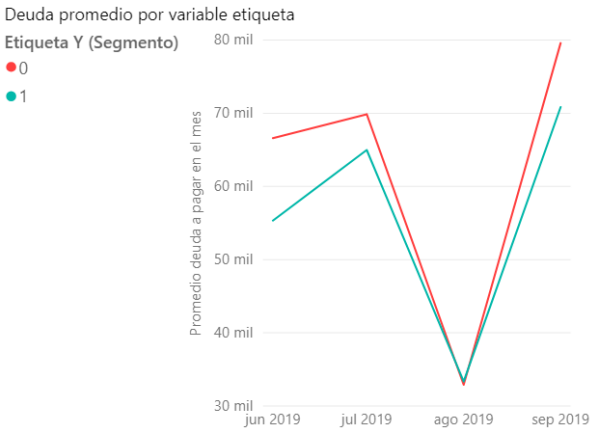
Anexo 8.3: Promedio de monto de transacciones, por división de datos entre positivos y negativos



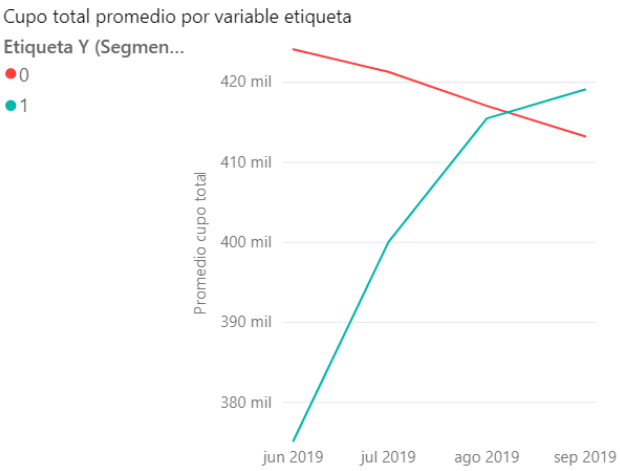
Anexo 8.4: Promedio de monto de pagos, por división de datos entre positivos y negativos



Anexo 8.5: Promedio de monto de deuda promedio por pagar en siguientes mes, por división de datos entre positivos y negativos



Anexo 8.6: Promedio de monto de cupo total, por división de datos entre positivos y negativos



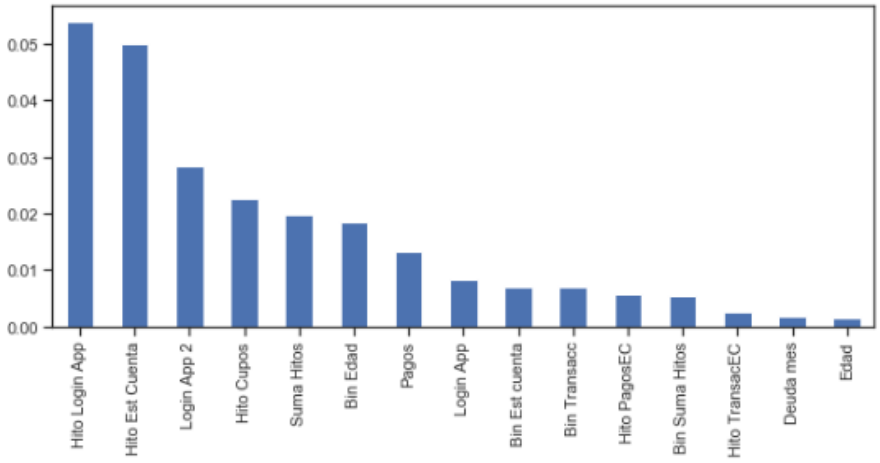
Anexo 9.1: 1er filtro de selección de variables en primer modelo predictivo. Correlación entre variable dependiente con las variables independientes.

Edad	0.038348
Login App	0.059218
Ofertas	0.060097
Pagos	0.236812
Hito Login App	0.079276
Hito Cupos	0.058800
Hito Est Cuenta	0.042591
Hito Ofertas	0.079402
Hito Pagos	0.229030
Hito TransacEC	0.052320
Hito PagosEC	0.052202
Suma Hitos	0.138336
Bin Transacc	0.052274
Deuda mes	0.033699
Bin Edad	0.037623
Login App 2	0.107746
Bin Cupos	0.060906
Bin Est cuenta	0.039218
Bin Suma Hitos	0.106527

Anexo 10.1: 2° filtro de selección de variables en primer modelo predictivo. Correlación entre variables dependientes. Con la selección de variables no correlacionadas entre si

	Edad	Login App	Pagos	Hito Login App	Hito Cupos	Hito Est Cuenta	Hito TransacEC	Hito PagosEC	Suma Hitos	Bin Transacc	Deuda mes	Bin Edad	Login App 2	Bin Est cuenta	Bin Suma Hitos
Edad	1	-0.13236	-0.0403195	-0.130551	0.024585	-0.0254813	-0.074718	-0.00650676	-0.0585684	0.0298825	0.010718	-0.670138	-0.151384	-0.0198685	-0.0427975
Login App	-0.13236	1	-0.0475917	0.408824	-0.0996364	0.293529	0.210527	-0.121893	0.279048	0.247421	0.0270718	0.0730673	0.626818	0.106175	0.0953557
Pagos	-0.0403195	-0.0475917	1	0.113935	0.173363	-0.0542594	-0.0153379	0.0159319	0.218144	-0.0620893	-0.0342766	0.0389005	0.101221	-0.0356093	0.120212
Hito Login App	-0.130551	0.408824	0.113935	1	-0.505008	0.313305	0.0205063	-0.298522	0.214852	0.024516	0.0673253	0.0818979	0.780976	0.202025	0.202118
Hito Cupos	0.024585	-0.0996364	0.173363	-0.505008	1	-0.104097	0.0696074	0.321263	0.554624	0.0140642	-0.0628655	-0.00595868	-0.318181	-0.0755671	0.38058
Hito Est Cuenta	-0.0254813	0.293529	-0.0542594	0.313305	-0.104097	1	0.0416959	-0.111279	0.506759	0.0610405	0.0280024	0.0111924	0.419356	0.333362	0.191136
Hito TransacEC	-0.074718	0.210527	-0.0153379	0.0205063	0.0696074	0.0416959	1	0.111269	0.123302	0.250028	0.00179272	0.051924	0.0936896	0.0164438	0.0902487
Hito PagosEC	-0.00650676	-0.121893	0.0159319	-0.298522	0.321263	-0.111279	0.111269	1	0.14066	0.00116976	-0.0843862	0.0231533	-0.237124	-0.0524073	0.193286
Suma Hitos	-0.0585684	0.279048	0.218144	0.214852	0.554624	0.506759	0.123302	0.14066	1	0.0756975	-0.0288494	0.045407	0.337519	0.202376	0.709029
Bin Transacc	0.0298825	0.247421	-0.0620893	0.024516	0.0140642	0.0610405	0.250028	0.00116976	0.0756975	1	-0.106868	-0.0184741	0.103474	0.0219949	0.0207651
Deuda mes	0.010718	0.0270718	-0.0342766	0.0673253	-0.0628655	0.0280024	0.00179272	-0.0843862	-0.0288494	-0.106868	1	-0.0382968	0.0415658	0.00903267	-0.0105978
Bin Edad	-0.670138	0.0730673	0.0389005	0.0818979	-0.00595868	0.0111924	0.051924	0.0231533	0.045407	-0.0184741	-0.0382968	1	0.0943765	0.0121716	0.0337179
Login App 2	-0.151384	0.626818	0.101221	0.780976	-0.318181	0.419356	0.0936896	-0.237124	0.337519	0.103474	0.0415658	0.0943765	1	0.291477	0.174859
Bin Est cuenta	-0.0198685	0.106175	-0.0356093	0.202025	-0.0755671	0.333362	0.0164438	-0.0524073	0.202376	0.0219949	0.00903267	0.0121716	0.291477	1	0.1079
Bin Suma Hitos	-0.0427975	0.0953557	0.120212	0.202118	0.38058	0.191136	0.0902487	0.193286	0.709029	0.0207651	-0.0105978	0.0337179	0.174859	0.1079	1

Anexo 11.1: Gráfico de ganancia de información, para el primer modelo predictivo. Donde sólo se escogen las variables que superan el índice 0,01.



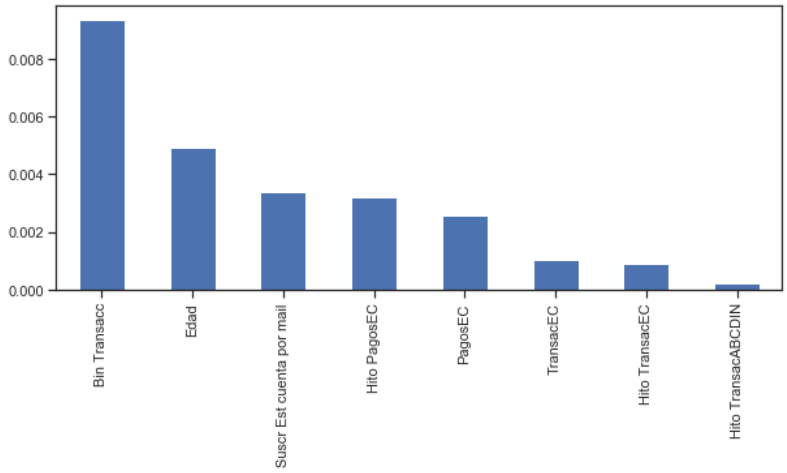
Anexo 9.2: 1er filtro de selección de variables en segundo modelo predictivo. Correlación entre variable dependiente con variables independientes.

Edad	0.077399
TransacEC	0.033873
PagosEC	0.056112
Hito TransacEC	0.054684
Hito PagosEC	0.079392
Hito TransacABCDIN	0.026195
Suscr Est cuenta por mail	0.063727
Bin Transacc	0.051548

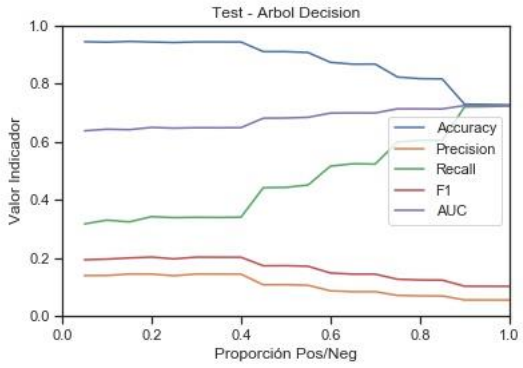
Anexo 10.2: 2° filtro de selección de variables en segundo modelo predictivo. Correlación entre variables dependientes. Con la selección de variables no correlacionadas entre si

	Edad	TransacEC	PagosEC	Hito TransacEC	Hito PagosEC	Hito TransacABCDIN	Suscr Est cuenta por mail	Bin Transacc
Edad	1	-0.0305377	-0.0779347	-0.0612195	-0.137431	-0.0150754	-0.075712	0.00914068
TransacEC	-0.0305377	1	0.0996411	0.603823	0.0630603	0.0238368	0.0271074	0.0849921
PagosEC	-0.0779347	0.0996411	1	0.101413	0.614545	0.0432778	0.0334472	0.0432644
Hito TransacEC	-0.0612195	0.603823	0.101413	1	0.097378	0.0383574	0.0466698	0.135323
Hito PagosEC	-0.137431	0.0630603	0.614545	0.097378	1	0.048694	0.0374506	0.016125
Hito TransacABCDIN	-0.0150754	0.0238368	0.0432778	0.0383574	0.048694	1	0.0251759	0.0348334
Suscr Est cuenta por mail	-0.075712	0.0271074	0.0334472	0.0466698	0.0374506	0.0251759	1	0.0565735
Bin Transacc	0.00914068	0.0849921	0.0432644	0.135323	0.016125	0.0348334	0.0565735	1

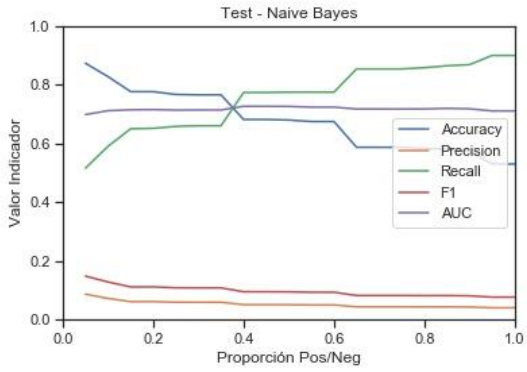
Anexo 11.2: Gráfico de ganancia de información, para el segundo modelo predictivo. Donde se elimina la variable con menor índice, dada su comparación de indicador de información, con respecto a las otras variables.



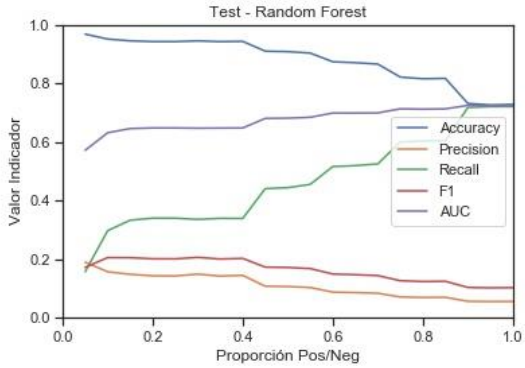
Anexo 12: Métricas de ajuste en testeo, por ratio de Oversampling. 1er modelo predictivo.



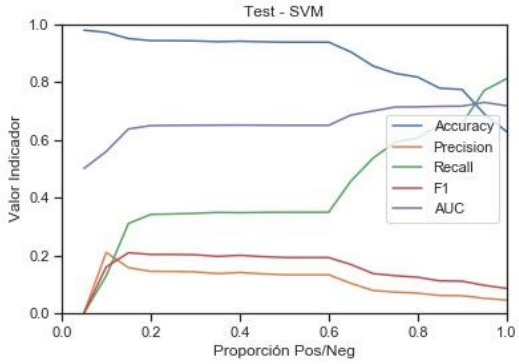
Árbol de Decisión



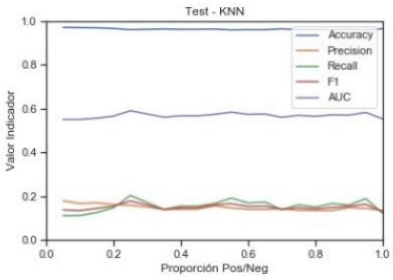
Naive Bayes



Random Forest



SVM

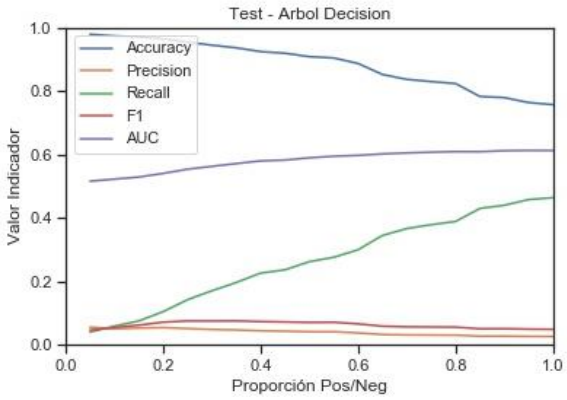


KNN

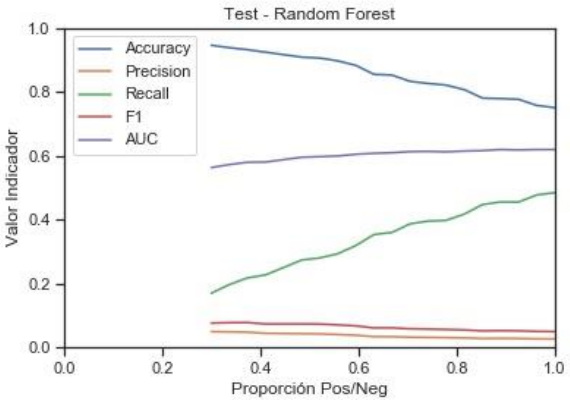
Anexo 13: Métricas Accuracy, Recall y F1 en 1er modelo predictivo, por algoritmo.

Modelo	Accuracy Train	Accuracy Test	Recall Train	Recall Test	F1 Train	F1 Test
Logit	73,35%	73,49%	74,01%	71,61%	73,53%	10,39%
Árbol de Decisión	74,22%	72,71%	76,55%	72,04%	74,81%	10,18%
Random Forest	74,22%	72,76%	76,55%	72,1%	74,81%	10,2%
Naive Bayes	70,77%	52,88%	90,57%	89,86%	75,6%	7,57%
Support Vector Machine	72,63%	61,59%	84,56%	81,79%	75,54%	8,37%
K-Nearest Neighbors	59,37%	95,99%	18,91&	21,23%	31,76%	18,56%

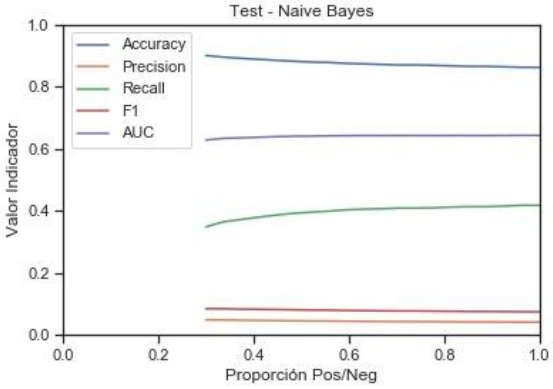
Anexo 14: Métricas de ajuste en testeo, por ratio de Oversampling. 2° modelo predictivo.



Árbol de Decision



Random Forest



Naive Bayes

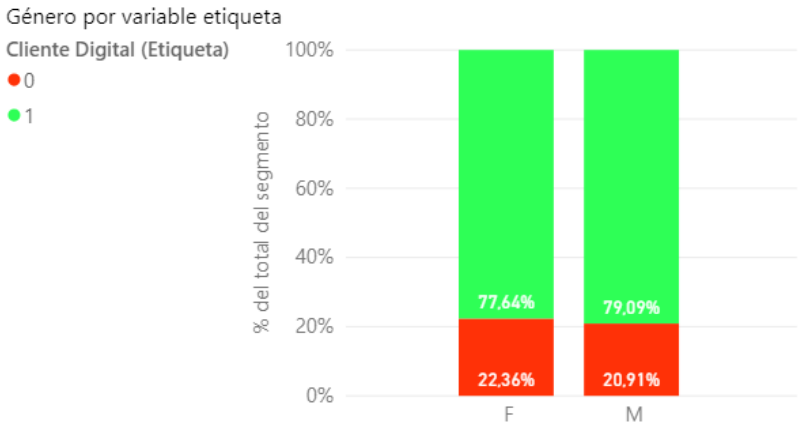
* No se realizó esta iteración en SVM y KNN, debido al gran volumen de datos y a las métricas de ajuste imprecisas que arrojó el modelo en los otros algoritmos.

Anexo 15: Métricas Accuracy, Recall y F1 en 2° modelo predictivo, por algoritmo.

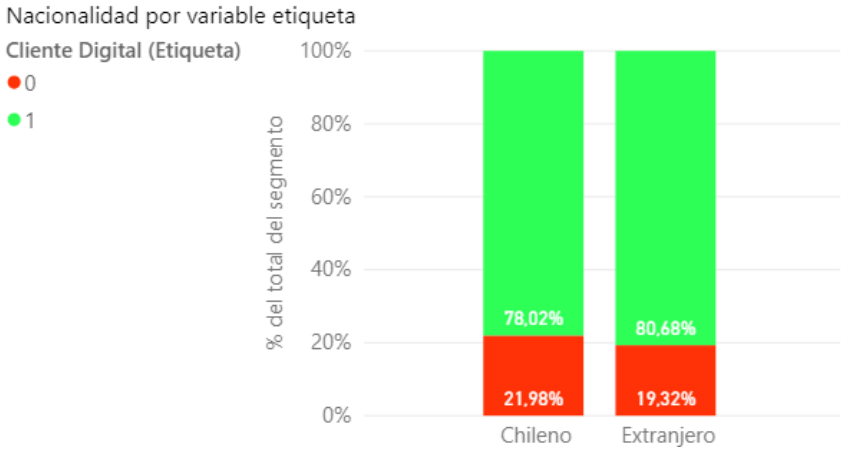
Modelo	Accuracy Train	Accuracy Test	Recall Train	Recall Test	F1 Train	F1 Test
Logit	69,5%	70,83%	68,56%	62,94%	69,21%	5,4%
Árbol de Decisión	73,96%	75,45%	68,95%	46,6%	72,59%	4,78%
Random Forest	73,72%	75,2%	68,68%	47,95%	72,32%	4,86%
Naive Bayes	64,9%	86,14%	43,6%	41,68%	55,4%	7,37%
Support Vector Machine	67,36%	54,88%	69,01%	77,23%	67,89%	4,33%
K-Nearest Neighbors	69,78%	76,53%	57%	33,96%	65,36%	3,68%

Anexo 16: Análisis exploratorio, 3er modelo predictivo.

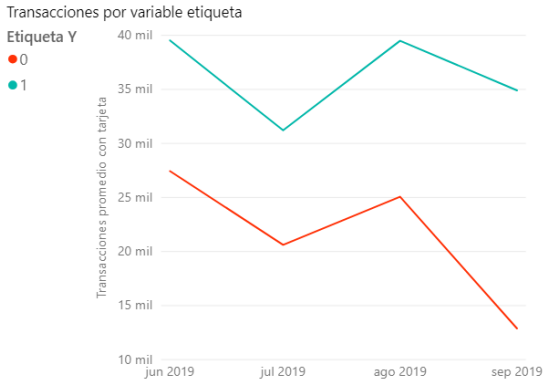
Anexo 16.1: Proporción etiquetas positivas y negativas c/r al total de datos, por género



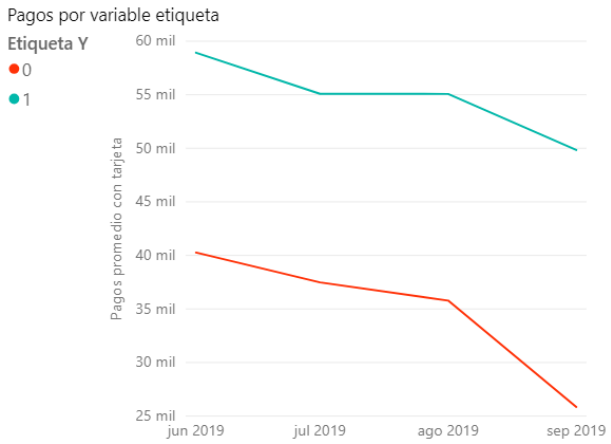
Anexo 16.2: Proporción etiquetas positivas y negativas con respecto al total de datos, por nacionalidad



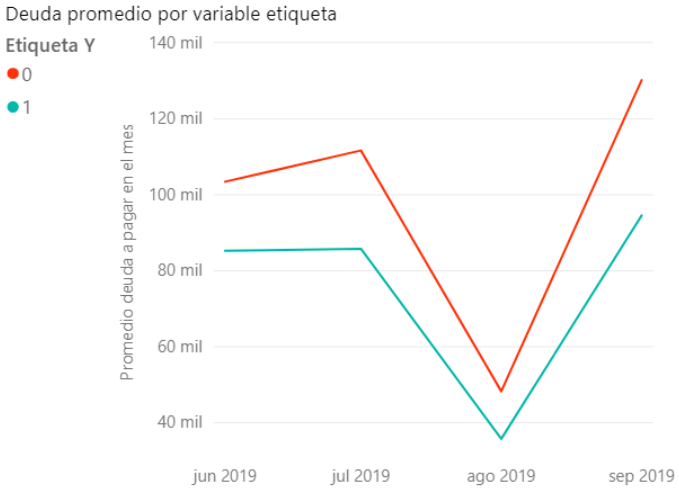
Anexo 16.3: Promedio de monto de transacciones, por división de datos entre positivos y negativos



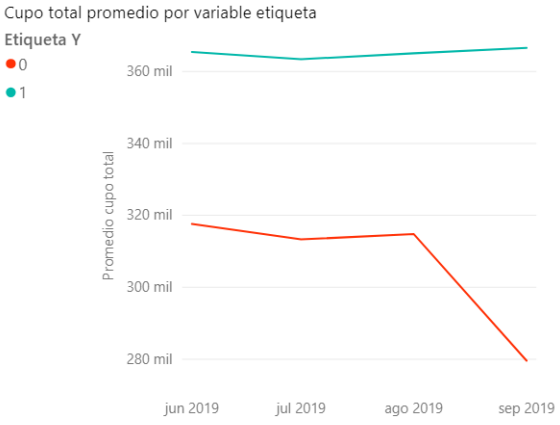
Anexo 16.4: Promedio de monto de pagos, por división de datos entre positivos y negativos



Anexo 16.5: Promedio de monto de deuda por pagar en siguiente mes, por división de datos entre positivos y negativos



Anexo 16.6: Promedio de monto de cupo total, por división de datos entre positivos y negativos



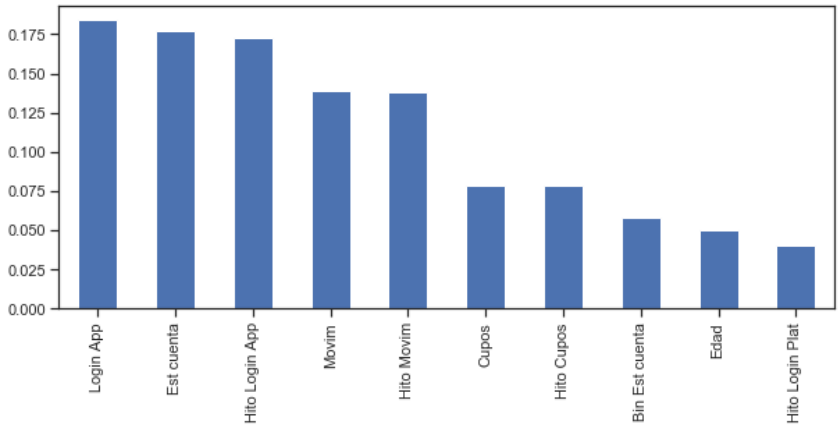
Anexo 17: 1er filtro de selección de variables para el tercer modelo predictivo. Correlación entre variable dependiente con variables independientes.

Edad	0.301571
Login App	0.504346
Cupos	0.373656
Est cuenta	0.531208
Movim	0.475402
Ofertas	0.512540
Hito Login App	0.629947
Hito Login Plat	0.309330
Hito Cupos	0.432574
Hito Est Cuenta	0.625635
Hito Movim	0.565478
Hito Ofertas	0.629908
Suma Hitos	0.667725
Login App 2	0.635982
Bin Cupos	0.378768
Bin Est cuenta	0.377456

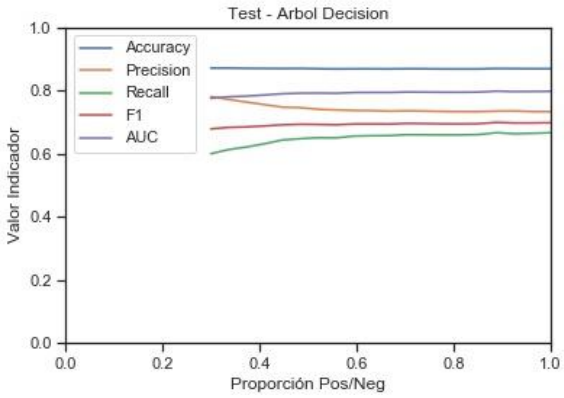
Anexo 18: 2° filtro de selección de variables en primer modelo predictivo. Correlación entre variables dependientes. Con la selección de variables no correlacionadas entre sí.

	Edad	Login App	Cupos	Est cuenta	Movim	Hito Login App	Hito Login Plat	Hito Cupos	Hito Movim	Bin Est cuenta
Edad	1	-0.236801	-0.174003	-0.22408	-0.212401	-0.312522	-0.15162	-0.215201	-0.266912	-0.171628
Login App	-0.236801	1	0.249182	0.720863	0.53485	0.661079	0.105043	0.314746	0.513345	0.359952
Cupos	-0.174003	0.249182	1	0.30969	0.703059	0.278891	0.731506	0.81346	0.621506	0.253099
Est cuenta	-0.22408	0.720863	0.30969	1	0.609436	0.68054	0.17084	0.352999	0.590136	0.374674
Movim	-0.212401	0.53485	0.703059	0.609436	1	0.480038	0.54437	0.623711	0.762649	0.352722
Hito Login App	-0.312522	0.661079	0.278891	0.68054	0.480038	1	0.153048	0.374932	0.600902	0.511565
Hito Login Plat	-0.15162	0.105043	0.731506	0.17084	0.54437	0.153048	1	0.803892	0.659764	0.164687
Hito Cupos	-0.215201	0.314746	0.81346	0.352999	0.623711	0.374932	0.803892	1	0.736023	0.26735
Hito Movim	-0.266912	0.513345	0.621506	0.590136	0.762649	0.600902	0.659764	0.736023	1	0.419874
Bin Est cuenta	-0.171628	0.359952	0.253099	0.374674	0.352722	0.511565	0.164687	0.26735	0.419874	1

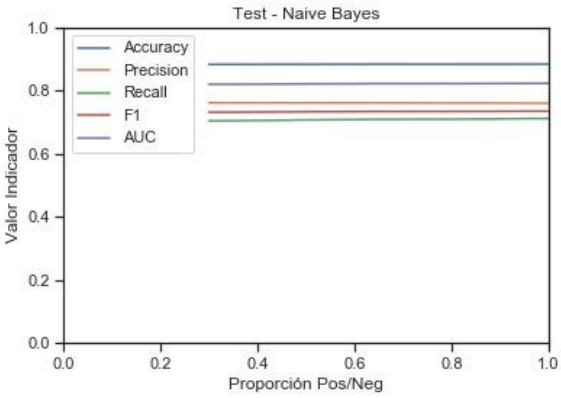
También se adjunta el gráfico de ganancia de información, para el tercer modelo predictivo, donde no se eliminan variables en esta filtro.



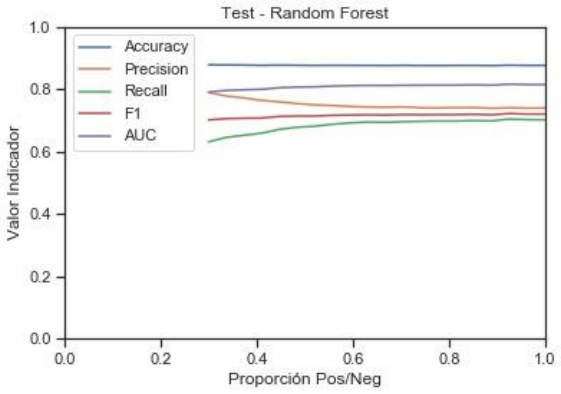
Anexo 19: Métricas de ajuste en testeo, por ratio de Oversampling. 3er modelo predictivo.



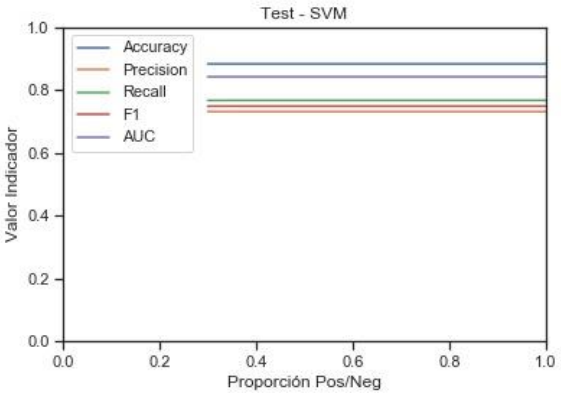
Árbol de Decisión



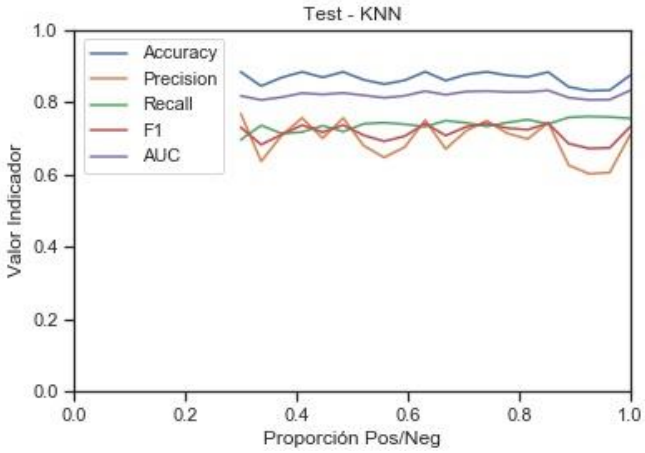
Naive Bayes



Random Forest



SVM



KNN

Anexo 20: Métricas Accuracy, Recall y F1 en 3er modelo predictivo, por algoritmo.

Modelo	Accuracy Train	Accuracy Test	Recall Train	Recall Test	F1 Train	F1 Test
Logit	83,86%	88,22%	76,17%	76,5%	82,52%	74,74%
Árbol de Decisión	85,47%	86,8%	75,34%	66,28%	83,83%	69,57%
Random Forest	85,36%	87,5%	75,74%	69,88%	83,8%	71,79%
Naive Bayes	81,94%	88,27%	70,75%	71,05%	79,67%	73,39%
Support Vector Machine	83,37%	88,2%	75,33%	76,75%	81,92%	74,76%
K-Nearest Neighbors	83,87%	88,21%	73,2%	73,83%	81,95%	74,03%