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

EVALUACIÓN DE UNA PROMOCIÓN PARA AUMENTAR OMNICANALIDAD EN UNA TIENDA POR DEPARTAMENTO

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

SEBASTIÁN IGNACIO ABARCA COCK

PROFESOR GUÍA:
ANDRÉS MUSALEM SAID

MIEMBROS DE LA COMISIÓN:
ALEJANDRA PUENTE CHANDÍA
JOSÉ NALDA REYES

SANTIAGO DE CHILE
2020

RESUMEN DE LA MEMORIA PARA OPTAR AL
TÍTULO DE: Ingeniero Civil Industrial
POR: Sebastián Ignacio Abarca Cock
FECHA: 13/01/2020
PROFESOR GUÍA: Andrés Musalem Said

EVALUACIÓN DE UNA PROMOCIÓN PARA AUMENTAR OMNICANALIDAD EN UNA TIENDA POR DEPARTAMENTO

Actualmente, y siguiendo la tendencia de algunos años atrás, la industria del retail está a la baja. Han caído las ventas e incluso han disminuido los empleos en el rubro. Lo anterior provoca el aumento de las promociones y liquidaciones en el sector, pero también es una oportunidad para probar soluciones innovadoras.

De manera de contraponerse a la situación actual de la industria, es que la tienda por departamento en la cual se trabaja ha puesto en desarrollo un framework de gestión de ciclo de vida del cliente. Esto para realizar estrategias de marketing a cada cliente, de acuerdo con la etapa en la que se encuentran con respecto a la tienda. Es decir, si están a punto de abandonar la empresa como clientes (fuga), se les envían ciertas promociones para propiciar su retención. Por otro lado, si son clientes habituales de la tienda, se les realiza un envío de promociones para que compren productos de categorías distintas a las usuales, para aumentar su rentabilidad dentro de la empresa.

De manera de optimizar la solución anterior, es que en este trabajo se realiza un modelo para predecir la propensión de clientes a concretar su primera compra en el canal web de la empresa, junto con un experimento que permite concluir a cuáles clientes enviar una promoción determinada, en función de su propensión ya calculada, para aumentar la cantidad de clientes omnicanal de la empresa en el corto plazo. Estos clientes a los cuales se pretende incentivar su compra en el sitio web de la empresa, no han comprado por ese canal en 1 año o más, pero han realizado transacciones en tiendas físicas de la empresa en los últimos 8 meses.

Los modelos evaluados usan variables transaccionales y sociodemográficas de los clientes, como cantidad de visitas con compra, suma total de compras, sexo, edad, etc. El modelo seleccionado posee un accuracy de 69.02% y un recall de 68.56% al usar un umbral óptimo de clasificación de casos, junto con un AUC de 0.7409. Mediante este modelo es que se segmentan a los clientes en propensión baja (3 deciles más bajos de propensión), propensión media (4 deciles siguientes) y propensión alta (3 deciles más altos de propensión).

Al analizar los resultados de un experimento en que se envía un cupón de descuento a clientes de todos los segmentos de propensión, se determina que sólo aquellos clientes del segmento de propensión media lograron una tasa de respuesta incremental significativamente positiva en el canal web, junto con una venta incremental positiva. Por lo que se recomienda enviar promociones similares sólo a este segmento de clientes. Pues son quienes pasan a ser omnicanal gracias al incentivo.

Como trabajo futuro se propone realizar experimentos con otros tipos de promociones, incorporar más datos de los clientes al modelo seleccionado, y evaluar la omnicanalidad a mediano y largo plazo de los clientes activados por la promoción.

AGRADECIMIENTOS

Ya en esta etapa de mi vida universitaria, me gustaría agradecer a todo aquel que me ayudó o colaboró con que haya llegado a este momento, a quienes fueron compañeros de grupo en ciertas tareas, a mis compañeros de sección, a mis compañeros de industrias, a mis colegas en mis prácticas profesionales, a mis colegas en mi trabajo de memorista y a mis profesores guía y co-guía.

Agradezco enormemente a la Fundación Moisés Mellado, sobre todo a Paulette Iribarne, su Gerente, por las becas y la ayuda otorgada durante mi carrera. Espero a futuro poder colaborar con la Fundación.

También quiero agradecer a mis amigos y amigas, que hicieron mi vida universitaria más llevadera, y me dieron ánimos para superar todas las dificultades. A Simón, Nicolás, Diland, Javier, Eduardo, Matías A., Fabián, Matías S., Alonso, Ricardo, Emilio, Lorena, Alejandro B., Alejandro P., Stefano y Kimberlyn, por todos esos almuerzos en la facultad, esas tardes de estudio, esos carretes y paseos. Momentos agradables que quedarán en mi memoria.

A Javiera, Franco y Guille por las salidas post trabajo, para despejar la mente y comentar nuestras vidas.

A Ispcia, Richard, Maxi, Caro y Anisa por hacer mucho más disfrutable mi intercambio en México y por cultivar la amistad una vez en Chile.

A Paulo por soportar mi difícil carácter en varias tareas grupales, por darme consejos y buenas vibras para salir adelante. A Vasty por siempre confiar en mis capacidades y habilidades. A Bárbara por escuchar mis anécdotas de vida y por ayudar a distraerme del trabajo y la universidad por unos momentos.

A mi hermano, Francisco, por invitarme a salir y a viajar junto a sus amigos, y por el regalo de hacerme tío de una sobrina hermosa. A esta última, Fran, por motivarme a volver a mi región a verla y consentirla.

Finalmente, agradezco a mis padres, sobre todo a mi mamá, por darme la oportunidad de estudiar en la Universidad de Chile y por costear todo lo que eso significa para un porteño sin contactos en la capital. Por preocuparse siempre de mi salud y mi desempeño académico. Por enseñarme a ser responsable, a cultivar el hábito del estudio y los valores que me hacen quien soy actualmente.

Infinitas gracias a todos aquellos con quienes compartí esta etapa de mi vida.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE GRÁFICOS E ILUSTRACIONES

1 Introducción

1.1 Antecedentes generales

La industria del retail en Chile lleva años con una tendencia negativa. De hecho, durante el año 2018, las grandes empresas realizaron con mucha más frecuencia liquidaciones y promociones de descuentos (El Mercurio, 2019). Más aún, las ventas registradas durante el primer trimestre del 2019 corresponden a las más bajas en los últimos cinco años, lo que derivó en un gran sobre stock de productos y liquidaciones. (La Tercera, 2019)

Es en este contexto que surge este trabajo de memoria. Cuyo fin es aumentar la cantidad de clientes activos en el canal web y el canal tiendas (omnicanalidad) de una tienda por departamento.

Entre las grandes organizaciones de retail en Chile, distingue la empresa en la cual se realiza este trabajo de memoria. Actualmente, ésta forma parte de un grupo de empresas, el cual está compuesto por las Tiendas por Departamento, Mejoramiento del Hogar, Supermercados, Financiero, Inmobiliaria y Tecnología.

La dotación de trabajadores del holding en Chile es de 54.472 empleados en total, donde 41.491 son colaboradores, 11.472 son profesionales y técnicos, y 1.509 son gerentes y ejecutivos. (Memoria anual empresa, 2019)

La organización, mediante sus tiendas físicas y su página web, comercializa productos para uso personal y del hogar, ordenados por múltiples categorías, entre las que se encuentran: vestuario y calzado, artículos de belleza, artículos electrónicos, electrodomésticos, muebles y accesorios de decoración; contando con marcas exclusivas internacionales y locales, además de marcas propias.

Si bien la empresa ofrece productos para ambos géneros, ésta se centra en la mujer como parte fundamental de su estrategia y el cliente a quien dirige sus esfuerzos publicitarios. Para posicionarse como la tienda preferida por este segmento, la compañía posee un modelo de negocio enfocado en la venta de productos a través de canales tanto presenciales, como virtuales.

En 2018, las tiendas de la empresa, realizaron ventas por $4.273 MM USD, lo que supuso un retroceso del 0,6% en comparación al año anterior (América-Retail, 2019). Actualmente, cuenta con 47 tiendas en todo el territorio nacional, alcanzando 338.698 metros cuadrados de superficie de venta. Junto con ello, es la empresa líder en la industria de comercio minorista, con un 23% de participación de mercado en el país. (Memoria anual empresa, 2019)

1.2 Motivación

Al momento de iniciado el trabajo de memoria, se terminaba de desarrollar un framework de gestión de clientes de la empresa, por el área de Inteligencia de Negocios. Este framework incluía el ciclo de vida de un cliente dentro de la empresa, focalizado en el canal web. El ciclo de vida puede ser visto en la Ilustración 1.

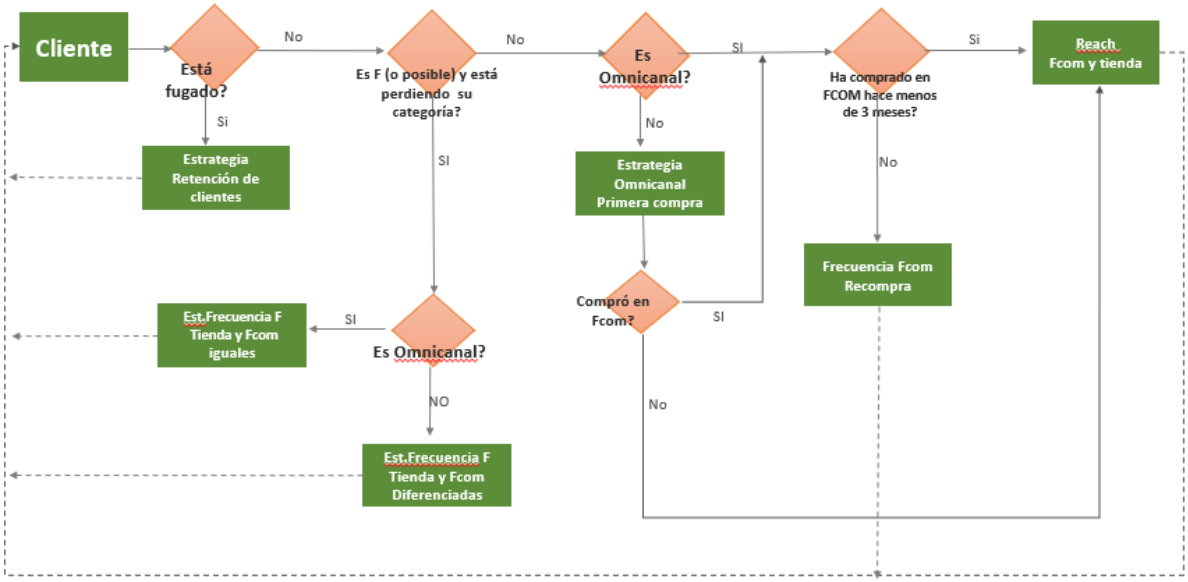

Ilustración 1: Framework de gestión de clientes. Fuente: Inteligencia de Negocios, Empresa.

Cabe mencionar que la empresa divide a sus clientes en dos categorías: Clientes F o Clientes Normales (o no F).

Clientes F: corresponden a esta categoría todos aquellos clientes que hayan visitado y comprado 7 o más veces en las tiendas de la empresa, y que hayan realizado transacciones acumuladas por un monto total igual o superior a $500.000, dentro de un año calendario. Incluye tiendas físicas o virtuales (página web de empresa).

Clientes Normales: corresponden a todos aquellos clientes que no cumplen las condiciones para ser F.

Por otro lado, la empresa etiqueta a un cliente como omnicanal si es que éste ha realizado alguna transacción en su página web en el corto plazo.

Las estrategias de marketing que incorpora el framework en función del ciclo de vida del cliente son las siguientes:

Primera compra web: estrategia que busca gatillar la primera compra en el canal web de la empresa. Incorpora información del historial de navegación. Enfocada en quienes no han comprado vía web en más de 1 año.

Recompra web: estrategia que busca mantener una frecuencia regular en las compras de los clientes por el canal web. Enfocada en quienes no han comprado vía web entre 3 meses a 1 año.

Frecuencia Categoría F: estrategia cuyo objetivo es que el cliente mantenga la condición de cliente F (frecuencia y gasto). Incorpora beneficios promocionales en las categorías blandas favoritas del cliente.

Reach: estrategia cuyo fin es que los clientes completen su closet con productos de la empresa. Diferencia entre clientes omnicanales y no omnicanales, ofreciendo a estos últimos promociones más atractivas en el canal web de la empresa. No posee diferenciación entre cliente F y normal.

Retención de clientes: estrategia que busca evitar la fuga de clientes de la organización. Campaña que se activa si la probabilidad de fuga supera el 50%. La acción se gatilla de manera prioritaria al resto de estrategias cuando aumenta el riesgo de fuga.

La empresa, sin embargo, pretende modelar cada estrategia de marketing para cada cliente, usando un método distinto al criterio comercial, que permita definir a qué clientes enviar promociones y a quiénes no, mediante un proceso de minado de datos. De manera de contrastar ambos resultados y elegir el método que sea óptimo para cada estrategia.

Así, como una forma de establecer a qué clientes enviar promociones para que hagan su primera compra en el canal digital es que se formuló la siguiente solución:

1. Calcular las probabilidades de **primera compra de cada cliente en el canal web de la empresa**.
2. Segmentar a los clientes según su propensity score.
3. Realizar un experimento de envíos de email promocionales según segmentos basados en propensity score.
4. Evaluar los resultados del experimento, para saber explícitamente, qué clientes reaccionan a cuáles incentivos.

Esto ayudaría a reducir los costos del área de Inteligencia de Negocios de la empresa, los cuales son directos e indirectos. Los costos directos incluyen los costos de comunicarse a un cliente objetivo (costos por enviar emails, cupones o llamadas), y los costos indirectos incluyen un costo latente que puede ser generado cuando una empresa elige como objetivo un cliente equivocado que no tiene intención de comprar un nuevo producto o servicio, lo cual conduce a deslealtad o insatisfacción del cliente, quien podría darse de baja de la lista de email marketing, o incluso dejar de comprar en la empresa.

1.3 Objetivos

1.3.1 Objetivo General

Desarrollar modelo predictivo para el área de Inteligencia de Negocios de una empresa de retail, que estime la propensión de cada cliente a realizar su primera compra en el canal web y que, mediante la determinación de la mejor acción promocional, permita aumentar la cantidad de clientes omnicanal.

1.3.2 Objetivos Específicos

1. Identificar las variables que afectan de manera relevante en la realización de la primera compra por canal online, y cuantificar su impacto.

2. Seleccionar el modelo predictivo de propensión a primera compra online, que posea el mejor rendimiento a partir de la comparación según métricas estadísticas.

3. Analizar el efecto de una promoción en la primera compra online de cada cliente, según su propensión.

4. Recomendar a cuáles clientes enviar un email promocional para lograr omnicanalidad en un corto plazo, según sus características (cantidad de visitas, ventas, edad, etc).

1.4 Alcances y Limitaciones

- Se trabajará usando solamente los datos disponibles dentro del data warehouse de la empresa. Específicamente con aquellos clientes que hayan realizado transacciones entre el 1 de enero de 2016 y el 31 de septiembre de 2019. Esto para tener datos suficientes para entrenar y validar los modelos.

- Los datos seleccionados sólo incluyen a clientes de todo el territorio chileno que sean identificables, esto es, aquellos para los cuales se tiene su rut correctamente ingresado y el registro de su comportamiento de compra (fecha, monto de compra, tienda, etc).

- Se pronosticará la realización de primera compra online para el próximo mes para cada cliente de la empresa.

- La promoción por evaluar será enviada únicamente por canal email, contemplando grupos de tratamiento y control. Cuidando la semejanza de características entre ambos grupos.

1.5 Resultados Esperados

Se espera obtener los siguientes resultados:

- Modelo predictivo capaz de estimar la propensión de cada cliente en realizar su primera compra por la página web de la empresa.

- Listado de variables relevantes para realizar primera compra online.

- Estrategia de marketing directo asociado a la acción promocional que incentiva la primera compra online de cada cliente, según su nivel de propensión.

2 Marco Conceptual

2.1 Metodología KDD

Con el fin de obtener un modelo probabilístico que estime la propensión de cada cliente a realizar su primera compra en el canal web de la empresa es que se hará uso de la metodología Knowledge Discovery in Databases (KDD), pues es una forma de obtener conocimientos en bases de datos altamente utilizada y documentada en problemáticas del ámbito del marketing. (Boselli et al, 2017) (Bate et al, 2008)

KDD es un proceso metodológico no-trivial para encontrar un modelo válido, útil y entendible que describa patrones de acuerdo con la información. No es un proceso

automático, sino uno reiterativo que explora exhaustivamente volúmenes muy grandes de datos para determinar relaciones. La Minería de Datos es sólo uno de los pasos en el camino hacia el descubrimiento de conocimiento dentro de los datos. (Fayyad et al, 1996)

El KDD requiere de un amplio y profundo conocimiento sobre el área de estudio en que se enmarcará, pues implica la evaluación e interpretación de patrones y modelos para tomar decisiones con respecto a lo que constituye conocimiento y lo que no lo es. Además, existen muchos métodos o algoritmos que podrían satisfacer los objetivos declarados. Por lo que también es necesario contar con conocimiento de conceptos de estadística.

Esta metodología está constituida por las siguientes etapas:

1. **Comprensión del dominio del estudio y establecimiento de objetivos:** se debe tener en claro cuáles son los límites y los objetivos que se pretenden lograr, junto con reconocer las fuentes de datos más importantes y quiénes tienen control sobre ellas. Además de dimensionar la cantidad de datos y formatos.

2. **Creación de un set de datos objetivo:** es la etapa donde los datos relevantes para el análisis son extraídos desde las fuentes de datos.

3. **Preprocesamiento:** se hace limpieza de los datos, esto es, tratamiento de datos perdidos o remoción de valores atípicos. Esto implica eliminar variables o atributos con datos faltantes o eliminar información no útil. Obteniéndose al final una estructura de datos adecuada para su posterior transformación.

4. **Transformación:** consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes. Operaciones de agregación y normalización, para consolidar los datos para la siguiente etapa.

5. **Data Mining:** proceso para descubrir los patrones o modelos presentes en los datos. Comprende tres pasos; la selección de la tarea, la selección del algoritmo y su uso.

    a. **Selección de la tarea:** se debe elegir el paradigma apropiado de minería de datos, ya sea clasificación, regresión, clusterización, sumarización, modelado de dependencia (estructural y cuantitativa) y modelado de cambios (cambios de valores previos o normativos), según los objetivos que se haya planteado. (Mackinnon et al, 1999)

    b. **Selección del algoritmo:** se debe elegir el o los algoritmos para la búsqueda del patrón y obtener conocimiento. Cada uno tiene su propia manera de trabajar y obtener los resultados, por lo que es recomendable conocer las propiedades de aquellos candidatos a utilizar y ver cuál se ajusta mejor a los datos.

c. **Aplicación del algoritmo:** se utiliza el algoritmo sobre el dataset, realizando varias iteraciones para ajustar sus parámetros, de ser necesario.

6. **Interpretación y evaluación de los resultados:** se identifican los patrones obtenidos y que son realmente interesantes. Se procede a evaluar los patrones que se generaron y el rendimiento que se obtuvo para verificar que cumpla los objetivos planteados en las primeras fases. A partir de aquí es posible regresar a cualquiera de los pasos anteriores.

7. **Integración al negocio:** se aplica el conocimiento encontrado al negocio, para comenzar a resolver sus problemáticas.

Estas etapas pueden apreciarse en la ilustración 2.

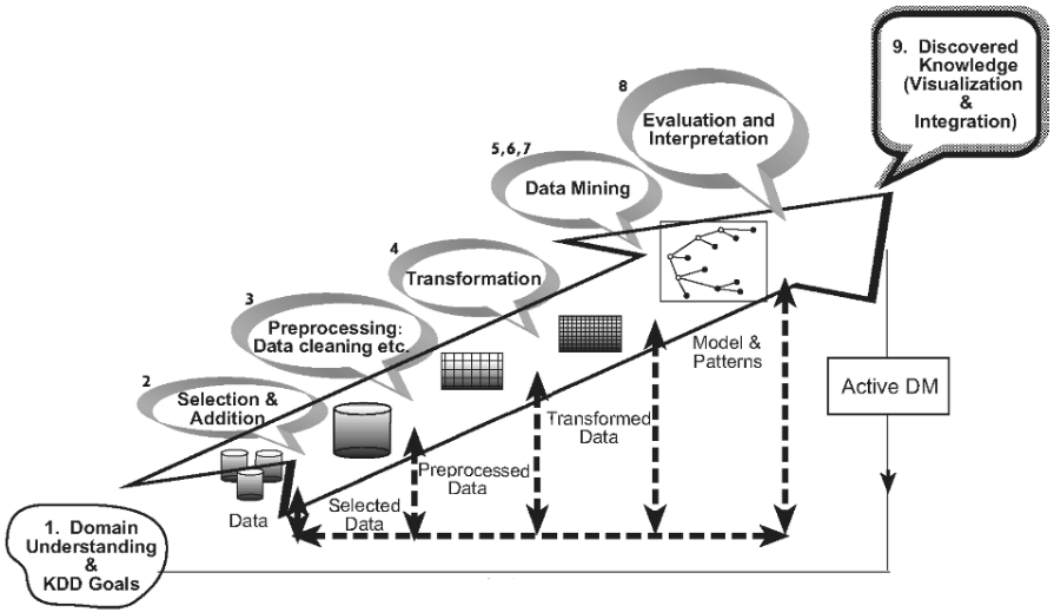

Ilustración 2: Etapas de la metodología KDD. Fuente: Data Mining and Knowledge Discovery Handbook. Oded Maimon and Lior Rokach.

2.2 Propensity Score

Los métodos de propensity score fueron propuestos por Rosenbaum y Rubin como herramientas centrales para ayudar a evaluar los efectos causales de las intervenciones (Rosenbaum y Rubin, 1983). Esto es de gran utilidad en el área de marketing, donde pueden ser evaluadas intervenciones como publicidad, promociones, etc. (Rubin y Waterman, 2006).
Se define el propensity score por cada cliente i (i = 1, …, N) como la probabilidad condicional de asignar un tratamiento particular ($Z_i = 1$) versus control ($Z_i = 0$) dado un vector de covariables observadas, x_i:

$$e(x_i) = pr(Z_i = 1 | X_i = x_i)$$

Donde se asume que, dados los X's, los Z_i son independientes:

$$pr(Z_1 = z_1, \dots, Z_N = z_N | X_1 = x_1, \dots, X_N = x_N) = \prod_{i=1}^{N} e(x_i)^{z_i}\{1 - e(x_i)\}^{1-z_i}$$

En este trabajo, el propensity score será usado para segmentar a los clientes, y así determinar cuáles segmentos de clientes reaccionan mejor a cuáles incentivos promocionales.

2.3 Adopción e intención de compra online

Tras una investigación del estado del arte relacionado a la adopción e intención de compra online, se recogieron las siguientes conclusiones:
Según un estudio estadístico realizado el 2001, en Estados Unidos, se concluyó que los factores que tienen un efecto en la intención de compra online son: tipo de producto, anterior compra vía online y género del cliente. Donde el efecto de la compra anterior en la intención de compra online era muy fuerte y también interactuaba con el tipo de producto. Además, se segmentó a los clientes en clusters con distintas orientaciones de compra. Para ello, se discriminó según sus niveles de disfrute, lealtad, precio, conveniencia y personalización a la hora de comprar. Los clusters encontrados fueron:

- Compradores personalizados.
- Compradores recreacionales.
- Compradores económicos.
- Compradores envueltos.
- Compradores recreacionales, orientados a la conveniencia
- Compradores orientados a la comunidad.
- Compradores apáticos, orientados a la conveniencia.

Segmentos para los cuales se recomendaron estrategias de marketing enfocadas en cumplir sus necesidades. (Brown M. et al, 2003).

Otro estudio, realizado el 2005, mediante encuestas a dueñas y dueños de casa en Reino Unido, concluyó que las variables que afectan significativamente en la adopción de compra online (primera compra), son: ingreso, edad, utilidad de internet y pensamientos de los clientes como: “comprar por internet se está haciendo más fácil”, “no compraría online porque el monto de dinero involucrado es alto”, “pagar deudas online es seguro”, “comprar por internet es muy conveniente”, “comprar por internet es conveniente sólo para ciertos productos” y “comprar por internet es arriesgado”. Donde la edad y los pensamientos de clientes “no compraría online porque el monto de dinero involucrado es alto” y “comprar por internet es arriesgado” afectan negativamente y las demás variables mencionadas afectan positivamente. Este estudio también segmenta a los clientes en “compradores online”, “navegadores” (quienes ven los precios e información de los productos online, pero compran en tienda), y “no compradores online” (sólo compran en tienda, sin navegar previamente). (Soopramanien, D. G., y Robertson, A, 2005).

Por otro lado, un estudio realizado el 2009, en Israel, modeló la raíz cuadrada del número de productos o servicios comprados online de un mercado en etapa temprana como variable dependiente, esto debido al efecto marginal decreciente de la publicidad online en ventas. Así, se concluyó que los hombres compran significativamente más que las mujeres, y las horas de navegación, los beneficios

percibidos de comprar online, la necesidad de búsqueda de información online, el tiempo de uso de internet, comprar por teléfono, la edad y ser comprador dominante afectan positiva y significativamente la variable dependiente. Los costos percibidos del proceso de compra, los costos percibidos de seguridad/privacidad y el disfrute de comprar afectaron negativa y significativamente la variable dependiente. No se encontraron efectos significativos en el estado marital y la actitud hacia publicidad online. (Liebermann, Y., y Stashevsky, S, 2009).

En Italia, en el año 2011, se realizó otro estudio estadístico para modelar la intención de compra online. Mediante el uso de árboles de decisión CART, en datos extraídos de encuestas realizadas a clientes de 24 sitios web, se concluyó que los ítems que más afectan la intención de compra online son: seguridad de pago, la conveniencia de precios, el costo de envío, la preferencia del sitio específico en relación a los demás y la frecuencia de compras hechas online. (Bonera, M,2011).

Así, se espera incluir variables como el sexo, edad, estado marital y navegación en la página web de la empresa, para estimar la propensión a primera compra online.

2.4 Problema de Clasificación

El objetivo del minado de datos es encontrar características inesperadas, atributos ocultos u otras relaciones no claras dentro de los datos, basado en combinaciones de técnicas (Mlambo, 2016). Las técnicas de Data Mining más usadas se categorizan en: agrupación, clasificación y regresión. Este trabajo de título, por su naturaleza, requiere de algoritmos de clasificación.

La clasificación es una técnica de minado de datos muy popular, la cual emplea un set de ejemplos pre-clasificados para desarrollar un modelo que pueda clasificar una población de registros. Puede ser usado para predecir etiquetas de clases categóricas y, tras clasificar datos en un set de entrenamiento, puede ser usado para clasificar nuevos datos disponibles.

En la industria del retail, la clasificación puede ayudar a los “direct marketers” otorgándoles tendencias útiles y precisas en el comportamiento de compra de sus clientes, y también ayudándolos a predecir cuáles productos sus clientes pueden estar interesados en comprar. De hecho, el minado de datos permite a los administradores de tiendas de retail a identificar a sus mejores clientes, a atraer clientes, a informarles vía email marketing, y maximizar sus utilidades mediante la identificación de los clientes más rentables.

Existen muchas técnicas de clasificación, entre las que destacan: redes neuronales, árboles de decisión, boosted logit y extreme gradient boosting.

2.5 Técnicas de muestreo para set de datos desbalanceados

En un set de datos que presenta observaciones clasificables dentro de dos clases distintas entre sí, se entiende como set de datos desbalanceado aquel en el cual existe una clara inequidad en la distribución de clases, tales como una proporción 1:100 o 1:1000 de observaciones de la clase minoritaria en relación a la mayoritaria.

El sesgo en el set de datos puede influenciar a muchos algoritmos de clasificación, llevando a algunos a ignorar a la clase minoritaria completamente. Esto es un problema cuando predecir la clase minoritaria es lo más importante.

Para resolver este problema, se han desarrollado distintas técnicas de muestreo, con el fin de lograr una distribución de clases más igualitaria dentro del set de datos. Algunas de las más usadas corresponden a las siguientes:

Downsampling: se eliminan aleatoriamente observaciones pertenecientes a la clase mayoritaria, hasta llegar a la distribución de clases deseada.

Upsampling: se duplican aleatoriamente observaciones pertenecientes a la clase minoritaria, hasta llegar a la distribución de clases deseada.

SMOTE: abreviación para Synthetic Minority Over-Sampling Technique. Se sobremuestrea la clase minoritaria tomando cada observación perteneciente a dicha clase e introduciendo observaciones sintéticas a lo largo de segmentos lineales que las unen con los k vecinos más cercanos pertenecientes a la misma clase. Dependiendo del sobremuestreo requerido, vecinos de los k-más cercanos son aleatoriamente escogidos. Este proceso se puede observar en la ilustración 3, donde x_i es el punto seleccionado, x_{i1} a x_{i4} son algunos de los vecinos más cercanos y r_1 a r_4 son los puntos de datos sintéticos creados por la interpolación aleatoria. Para generar las observaciones sintéticas se toma la diferencia entre el vector perteneciente a la clase minoritaria y su vecino más cercano. Esta diferencia es multiplicada por un número aleatorio entre 0 y 1 y es añadida al vector de la clase minoritaria. Esto provoca la selección de un punto aleatorio a lo largo del segmento lineal entre dos observaciones específicas.

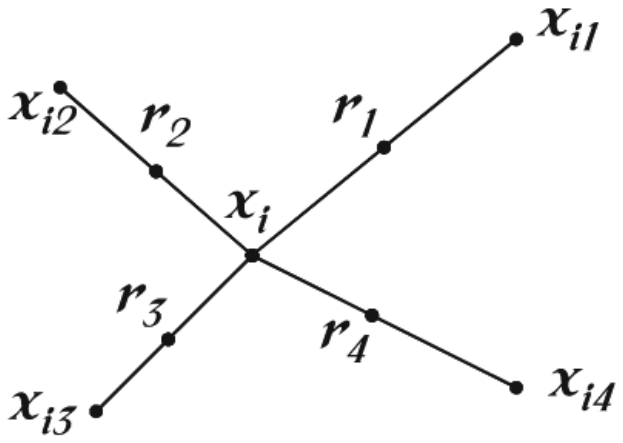

Ilustración 3: Creación de datos sintéticos mediante SMOTE. Fuente: SMOTE-RSB∗: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory.

2.6 Algoritmos de Clasificación

Los algoritmos de clasificación que serán testeados en este trabajo incluyen: árboles de decisión, boosted logit y extreme gradient boosting.

2.6.1 Árboles de Decisión

Los algoritmos de árboles de decisión surgieron de un enfoque no paramétrico, el cual no requiere de ningún supuesto sobre las clases ni distribución de los datos o escalas. Estos algoritmos conducen a un set de reglas “si-entonces” para mejorar la lectura humana. Usan la función valor discreto, la cual es robusta a datos ruidosos. Implementan datos ordinales y nominales como variables de ingreso. Existen diferentes métodos, como ID3, CART, CHAID, C5.0 y QUEST. Cada uno difiere en sus métodos de extracción de conocimiento, como también en su *accuracy*, velocidad de clasificación, valores perdidos, manejo de datos ruidosos, etc.

En este trabajo, se analizará el rendimiento de modelos basados en árboles de decisión C5.0. Este algoritmo fue desarrollado por Quinlan, creador de los algoritmos ID3 y C4.5. Incluye todas las funcionalidades del árbol de decisión C4.5, junto con nuevas tecnologías, como el “boosting”, capaz de mejorar la *accuracy* del modelo mediante iteraciones de árboles, o el uso de una matriz de costos para asignar distinto peso a los errores de la matriz de confusión. (Pang y Gong, 2009)

Otras características incluyen:

- La medida empleada para las divisiones del árbol es la entropía.
- Incorpora un método para la selección de predictores previo al ajuste del modelo, llamado “winnowing”.
- El algoritmo de boosting se detiene si la incorporación de nuevos árboles no aporta un mínimo de mejora al árbol original.

2.6.2 Boosted Logit

Los algoritmos de boosting combinan el rendimiento de muchos clasificadores “débiles” para producir un clasificador “poderoso”. Estos algoritmos fueron propuestos por Schapire (1990), Freund (1995), Freund y Schapire (1997), y desde entonces han recibido mucha atención. En particular, los modelos boosted logit constituyen la suma de varios modelos logit que van disminuyendo su error, gracias a los modelos anteriormente calculados.

2.6.3 Extreme Gradient Boosting

Es una implementación de Gradient Boosting Machines (GBM) creada por Tianqi Chen. Los modelos de extreme gradient boosting “aprenden” mediante los errores de modelos anteriores. En particular, utilizan el gradiente de una función de pérdida para minimizar el error. Se diferencia de los métodos tradicionales de gradient boosting al introducir un término de regularización para penalizar la complejidad de la función, logrando que el resultado corra menos riesgo de un sobreajuste. Soportan modelos de regresión y clasificación. Los hay del tipo “lineal” y del tipo “árbol”, diferenciándose en los modelos que utilizan para minimizar los errores de predicción.

El factor más importante detrás del éxito de este algoritmo es su escalabilidad en todos los escenarios. El proceso es diez veces más rápido que las soluciones populares ya existentes en una sola máquina. Esta escalabilidad es producto de varias optimizaciones de sistema y algorítmicas. (Chen y Guestrin, 2016)

Optimizaciones de sistema:

- **Paralelización** de la construcción de árboles al usar todos los núcleos de la CPU durante el entrenamiento.
- **Computación Distribuida** para entrenar modelos muy grandes al usar un clúster de máquinas.
- **Computación Fuera-de-Núcleo** para set de datos muy grandes que no caben dentro de la memoria.
- **Optimización de Caché** de las estructuras de datos y algoritmo para hacer un mejor uso del hardware.

Optimizaciones de algoritmo:

- **Consciente de Escasez** de datos, con un manejo automático de los valores perdidos.
- **Estructura de Bloques** para soportar la paralelización en la construcción de árboles.
- **Entrenamiento Continuado** para poder usar “boosting” en un modelo ya ajustado en datos nuevos.

2.7 Indicadores de Rendimiento

Luego de testear los algoritmos de clasificación, es necesario evaluarlos mediante determinadas métricas, con el fin de elegir el de mejor rendimiento entre ellos.

2.7.1 Matriz de confusión

Un modelo de clasificación binario (como el de este trabajo de título) clasifica cada observación en una clase entre dos alternativas: una clase verdadera y una clase falsa (IASRI, 2015). Esto da origen a cuatro posibles clasificaciones para cada observación: un verdadero positivo (True Positive), un verdadero negativo (True Negative), un falso positivo (False Positive), o un falso negativo (False Negative). Esta situación puede ser representada mediante una matriz de confusión (también llamada tabla de contingencias) como en la ilustración 4. La matriz de confusión yuxtapone las clasificaciones observadas o reales (columnas) con las clasificaciones predichas de un modelo (filas). En la ilustración 4, las clasificaciones que se ubican en la diagonal son las clasificaciones correctas, esto es, los verdaderos positivos y los verdaderos negativos. Los otros campos corresponden a errores del modelo. En un modelo perfecto, sólo los campos de la diagonal tendrían valores y los demás serían igual a cero.

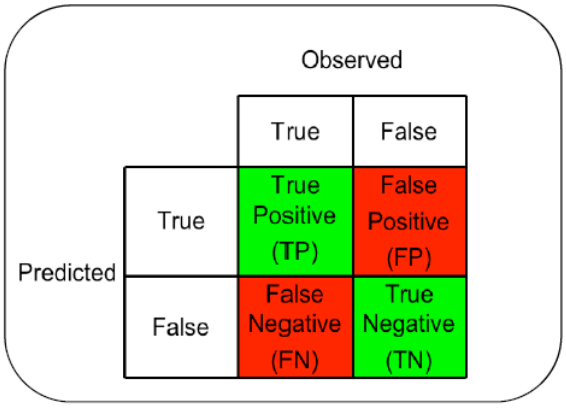

Ilustración 4: Matriz de confusión. Fuente: Evaluation Measures for Data Mining Tasks. Indian Agricultural Statistics Research Institute.

Varias métricas de rendimiento de modelos pueden ser derivadas de la matriz de confusión.

Accuracy: es la proporción del número total de predicciones correctas. Calculada como el ratio entre el número de casos clasificados correctamente y el número de casos totales. (Oprea, 2014)

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

Error: indica la proporción de casos clasificados incorrectamente.

$$\boldsymbol{Error} = \mathbf{1} - \boldsymbol{Accuracy}$$

Precision: indica la proporción de los casos clasificados como positivos que corresponden a la clase positiva.

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

Recall: es la proporción de casos pertenecientes a la clase positiva que son correctamente predichos como positivos.

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

F-Measure (F-score): es una combinación de las dos últimas métricas descritas (Precision y Recall), tomando la media armónica entre ellas.

$$\boldsymbol{F - Measure} = \mathbf{2} * \frac{\boldsymbol{Precision} * \boldsymbol{Recall}}{\boldsymbol{Precision} + \boldsymbol{Recall}}$$

$$\boldsymbol{F - Measure} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \frac{\boldsymbol{FN} + \boldsymbol{FP}}{\mathbf{2}}}$$

2.7.2 Curva ROC

Una curva ROC (Receiver Operating Characteristic) es un gráfico que muestra el rendimiento de un modelo de clasificación en todos los umbrales de clasificación. Esta

curva representa dos parámetros: la tasa de verdaderos positivos y la tasa de falsos positivos. (IASRI, 2015)

Tasa de verdaderos positivos (TPR), Sensibilidad o Recall: $TPR = \frac{TP}{TP+FN}$

Tasa de falsos positivos (FPR) o 1-Especificidad: $FPR = \frac{FP}{FP+TN}$

Dicho de otro modo, la curva ROC corresponde a un gráfico que muestra la Sensibilidad VS 1 – Especificidad. Así, el punto (0,1) de la curva ROC es la predicción perfecta e ideal, y clasifica correctamente todos los casos positivos y negativos. El punto (1,1) representa un clasificador que predice todos los casos como positivos. El punto (1,0) representa un clasificador que clasifica todas las observaciones incorrectamente. Ejemplos de puntos en la curva ROC están representados en la ilustración 5.

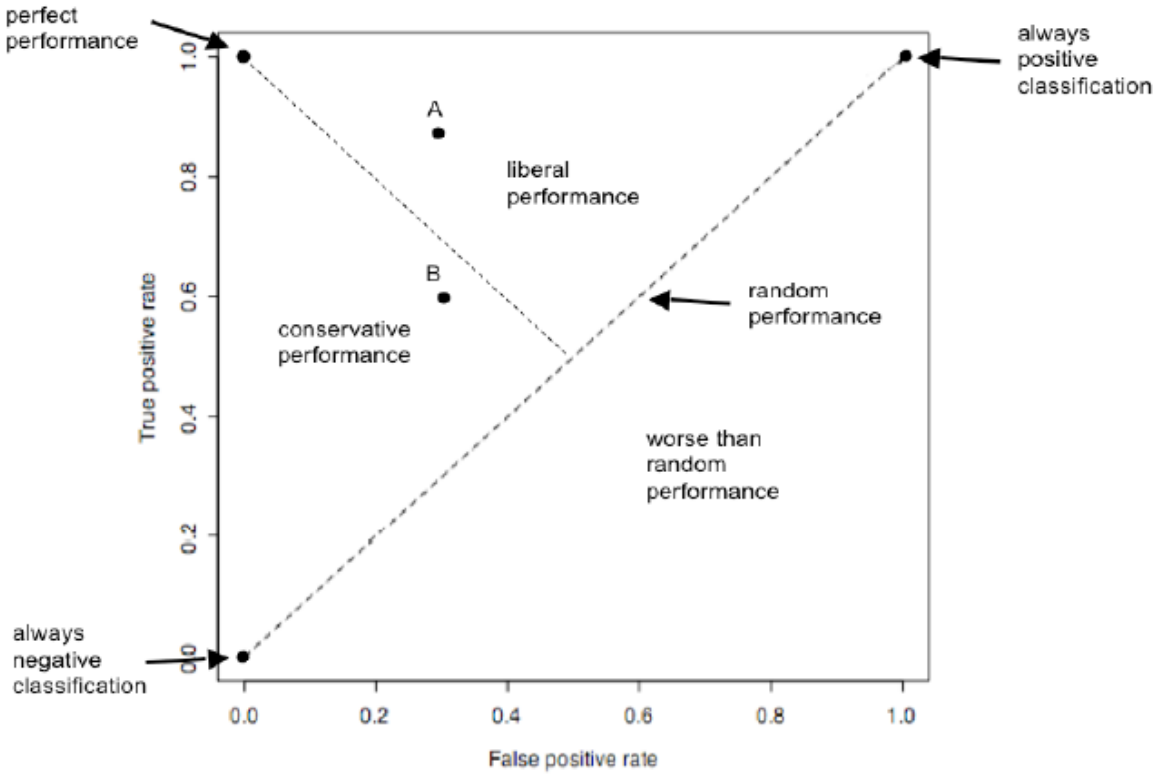

Ilustración 5: Regiones de una curva ROC. Fuente: Evaluation Measures for Data Mining Tasks. Indian Agricultural Statistics Research Institute.

El AUC (área bajo la curva ROC) mide toda el área bidimensional por debajo de la curva ROC completa, de (0,0) a (1,1). Proporciona una medición agregada del rendimiento en todos los umbrales de clasificación posibles. Una forma de interpretar el AUC es como la probabilidad de que el modelo clasifique un ejemplo positivo aleatorio más alto que un ejemplo negativo aleatorio. (Bradley, 1997)

2.8 Experimentos

Un **experimento**, en general, consiste en el establecimiento de ciertas condiciones para comprobar la influencia de una causa sobre unos resultados. (Ildefonso, G. & Elena, A., 2009)

Existen de dos tipos:

Experimentos directos: son parecidos a los que se desarrollan en el ámbito de las ciencias sociales. Suponen un previo conocimiento del fenómeno que se pretende estudiar, porque existe una teoría que relaciona las variables. Por ejemplo, existe una teoría económica que establece que existe una relación inversa entre precio y cantidad. Por lo que, aumentar el precio esperando una disminución en la cantidad comprada corresponde a un experimento directo.

Experimentos indirectos: estos consisten en crear unas condiciones para probar sus consecuencias, sin hipótesis formales previas. Por ejemplo, una empresa ingresa al mercado un nuevo producto para saber si tiene aceptación por sus clientes, interés de distribuidores, difusión, etc.

Todo experimento está formado por variables, las que pueden ser:

Variables independientes: son las variables que se controlan y que se suponen causales, por ejemplo, el descuento que se aplica sobre un producto, o la existencia de un aviso publicitario, son variables independientes.

Variable dependiente: es aquella que se supone tiene que responder ante los cambios en las variables independientes. Es la consecuencia, o efecto ante la causa. Por ejemplo, la cantidad vendida de un producto, que se supone alterable ante las modificaciones de precio si los demás factores causales permanecen constantes.

Variables de confusión: son aquellas distintas de las independientes que afectan a los resultados, o consecuencias. O no se controlan o no pueden ser controladas. Por ejemplo, una empresa puede disminuir el precio de un producto para estudiar su efecto sobre las ventas. Puede suceder que al mismo tiempo tenga lugar una campaña publicitaria. También es posible que los competidores imitaran a la empresa, y bajaran el precio de sus productos a la misma vez. La publicidad y la imitación de la competencia son variables de confusión, porque pudiendo ser controladas no lo fueron, como la publicidad, o no es posible controlarlas, como la bajada de precio de los competidores.

Unidades de prueba: son las que responden a las causas. Por ejemplo, personas, casas comerciales, mercados geográficos, etc.

Los experimentos deben poseer la propiedad de validez. Es decir, debe ser posible establecer relaciones causa-efecto. Esta validez puede ser interna y externa.

Validez interna: es una medida de la precisión de un experimento. Existe validez interna si en un experimento la causa origina un efecto. Por ejemplo, si la disminución de precios provoca un aumento en la cantidad vendida de un producto, se dice que existe validez interna.

Validez externa: permite la generalización de la relación de causalidad encontrada. Consiste en poder llegar a formular leyes de comportamiento. Si existiera validez externa quiere decir que se puede confiar en que las mismas causas provocan los mismos efectos.

Es deseable que se cumpla tanto la validez interna como externa en un experimento.

Los experimentos deben ser aleatorios. Las unidades de prueba no deben ser elegidas, sino que seleccionadas al azar. Esto para reducir el sesgo de la muestra. De esta manera, se controla por algunas variables ajenas al experimento.

Para poder estimar los efectos causales en un experimento es necesario la creación de dos grupos: el grupo de control y el grupo de tratamiento.

Grupo de control: es el grupo al cual no se le realiza ninguna intervención, y se controla por todas sus variables. Es igual al grupo de tratamiento, difiriendo sólo en la intervención no realizada.

Grupo de tratamiento: es el grupo al cual se le realiza la intervención, controlando las demás variables.

2.9 Cálculo de población de grupos de control y de tratamiento

Para calcular el tamaño de la población necesaria en los grupos de control y de tratamiento (clientes que no recibirán promoción y clientes que recibirán promoción) se hará uso del estadístico Z, para testear la diferencia entre dos proporciones. Esto, ya que, para evaluar los resultados de la campaña promocional, se compararán las tasas de respuestas de los clientes.

El estadístico Z permite contrastar la hipótesis nula $H_0: \pi_{GT} = \pi_{GC}$ frente a $H_1: \pi_{GT} \neq \pi_{GC}$ a partir de dos muestras independientes.

Este estadístico se calcula de la siguiente manera:

$$Z = \frac{p_{GT} - p_{GC}}{\sqrt{\frac{p(1-p)}{n_{GT}} + \frac{p(1-p)}{n_{GC}}}}$$

Donde:
p_{GT}: proporción grupo de tratamiento.
p_{GC}: proporción grupo de control.
p: estimación de proporción obtenida del total de observaciones.
n_{GT}: población grupo de tratamiento.
n_{GC}: población grupo de control.

Además, una aproximación de p puede calcularse como:

$$p = \frac{n_{GT} * p_{GT} + n_{GC} * p_{GC}}{n_{GT} + n_{GC}}$$

2.10 Indicadores de rendimiento campaña de primera compra web

Con el fin de poder medir y comparar el rendimiento de las campañas de primera compra web realizadas por la empresa, se tienen en consideración la tasa de respuesta, la tasa de respuesta incremental y la venta incremental, obtenidas luego de haber concluido el tiempo de canje de los cupones de descuento de cada campaña.

Tasa de respuesta: porcentaje de clientes que realizó compras web del total de clientes. Medido en el período de canje de los cupones de descuento de las campañas de email marketing.

Así, la tasa de respuesta del grupo control de cada segmento s de clientes (según su propensión) se calcula:

$$TR_{GCs} = \frac{Cantidad\ de\ clientes\ GC_s\ con\ compras\ web\ realizadas}{Cantidad\ de\ clientes\ GC_s\ en\ total}$$

Mientras que la tasa de respuesta del grupo de tratamiento de cada segmento de clientes se calcula:

$$TR_{GTs} = \frac{Cantidad\ de\ clientes\ GT_s\ con\ compras\ web\ realizadas}{Cantidad\ de\ clientes\ GT_s\ en\ total}$$

Tasa de respuesta incremental: diferencia entre tasas de respuesta del grupo de control y el grupo de tratamiento de una campaña.

De este modo, la tasa de respuesta incremental de cada segmento s de clientes se calcula como:

$$TRI_s = TR_{GTs} - TR_{GCs}$$

Por otro lado, la venta incremental en el canal web de cada segmento de clientes s se obtiene de la forma:

$$VI\ web_s = Venta\ GT_s\ web - \left(\frac{Venta\ GC_s\ web * Cantidad\ de\ clientes\ GT_s\ en\ total}{Cantidad\ de\ clientes\ GC_s\ en\ total}\right)$$

Para estos indicadores se toman en cuenta las compras web realizadas en cualquier categoría de productos de la empresa, no sólo las que poseen descuento gracias al email enviado. Esto para medir el efecto general de las campañas de email marketing.

2.11 Weight of Evidence e Information Value

El weight of evidence (o peso de la evidencia) muestra el poder predictivo de una variable independiente en relación con la variable dependiente. Al originarse del mundo de scoring de créditos, es generalmente descrito como una medida de la separación de clientes buenos y malos. Un “cliente malo” se refiere a un cliente que haya tenido un default en su préstamo, mientras que un “cliente bueno” se refiere a un cliente que haya devuelto su préstamo. Se calcula de la siguiente manera:

$$WOE = \ln\left(\frac{Distribución\ de\ Buenos}{Distribución\ de\ Malos}\right)$$

Donde:

Distribución de Buenos: porcentaje de clientes buenos en un grupo particular.
Distribución de Malos: porcentaje de clientes malos en un grupo particular.

Un WOE positivo indica Distribución de Buenos > Distribución de Malos.
Un WOE negativo indica Distribución de Buenos < Distribución de Malos.

En términos de eventos y no-eventos, el WOE puede ser calculado de la siguiente manera:

$$WOE = \ln\left(\frac{Porcentaje\ de\ no\ eventos}{Porcentaje\ de\ eventos}\right)$$

Weight of Evidence (WOE) ayuda a transformar una variable independiente continua o categórica en un set de grupos o categorías basadas en la similitud de distribución de la variable dependiente, esto es, el número de eventos y no eventos.

Así, los beneficios de su implementación son los siguientes:

- Puede tratar outliers, ya que los deja dentro de un grupo.
- Puede manejar datos faltantes, ya que son separados en un grupo aparte.
- Ya que maneja variables categóricas, no hay necesidad de crear variables dummy.
- La transformación mediante WOE ayuda a construir relaciones lineales estrictas con log odds (logaritmo de la razón de oportunidades).

Por otro lado, para medir el poder predictivo de una variable independiente puede usarse el Information Value (valor de información). Esta es una de las técnicas más útiles para seleccionar variables importantes en un modelo predictivo. Usa la siguiente fórmula:

$$IV = \sum(Porcentaje\ de\ no\ eventos - Porcentaje\ de\ eventos) * WOE$$

De acuerdo con Siddiqi (2006), por convención los valores de IV pueden ser interpretados de la siguiente forma:

- Menor a 0.02: el predictor no es útil para modelar (separar los casos buenos de los malos).
- Entre 0.02 y 0.1: el predictor tiene sólo una relación débil con el odds ratio entre casos buenos y malos.
- Entre 0.1 y 0.3: el predictor tiene una relación de fuerza media con el odds ratio entre casos buenos y malos.
- 0.3 o mayor: el predictor tiene una relación fuerte con el odds ratio entre casos buenos y malos.

3 Desarrollo Metodológico

3.1 Creación de un set de datos objetivo

Ya definidos los objetivos y alcances del trabajo, junto con haber estudiado el estado del arte asociado, se prosigue con la creación del set de datos para continuar con la metodología KDD.

Se dispone de varias bases de datos con información sociodemográfica, de navegabilidad y transaccional de los clientes de tiendas chilenas y el canal web de la empresa, desde el 1 de enero del 2016 al 31 de septiembre del 2019. Éstas fueron consolidadas, con los datos agrupados según el rut del cliente.

Para esta memoria, se trabajará con los datos de aquellos clientes identificables, esto es, de quienes se tiene su respectivo rut e información transaccional. Así, se poseen los datos de más de ocho millones de clientes.

Las variables previamente disponibles, que no fueron transformadas, y serán usadas en este trabajo corresponden a las de la tabla 1.

Variable	Tipo	Descripción
Rut	Numérica	Rut del cliente
Estado civil	Categórica	Estado civil del cliente (soltero, casado, sin información)
Zona de residencia	Categórica	RM Oriente, RM Poniente, Norte, Sur
Sexo	Categórica	Sexo del cliente (masculino, femenino, sin información)
Segmento Valor Tarjeta	Categórica	Tipo de tarjeta de crédito del cliente (Normal, Premium, Elite)

Tabla 1: Variables disponibles en bases de datos. Fuente: Elaboración propia.

Las demás variables a usar fueron preprocesadas para ser utilizadas en los modelos estadísticos.

3.2 Preprocesamiento de datos

Para una correcta predicción de los modelos estadísticos a usar, se debieron tratar los valores extremos (outliers). Para ello, se identificaron los outliers multivariados asociados a las visitas por año, la cantidad total de productos comprada en un año y las ventas totales en un año, ya que son unas de las variables más relevantes para el negocio. Como metodología para detectar estos valores extremos se usó la distancia de Mahalanobis.

Cabe mencionar que para usar esta metodología es necesario que los datos posean una distribución normal. Sin embargo, al graficar un histograma para las variables mencionadas anteriormente, se observa que siguen una distribución exponencial decreciente. Lo anterior se aprecia en los gráficos 1 y 2.

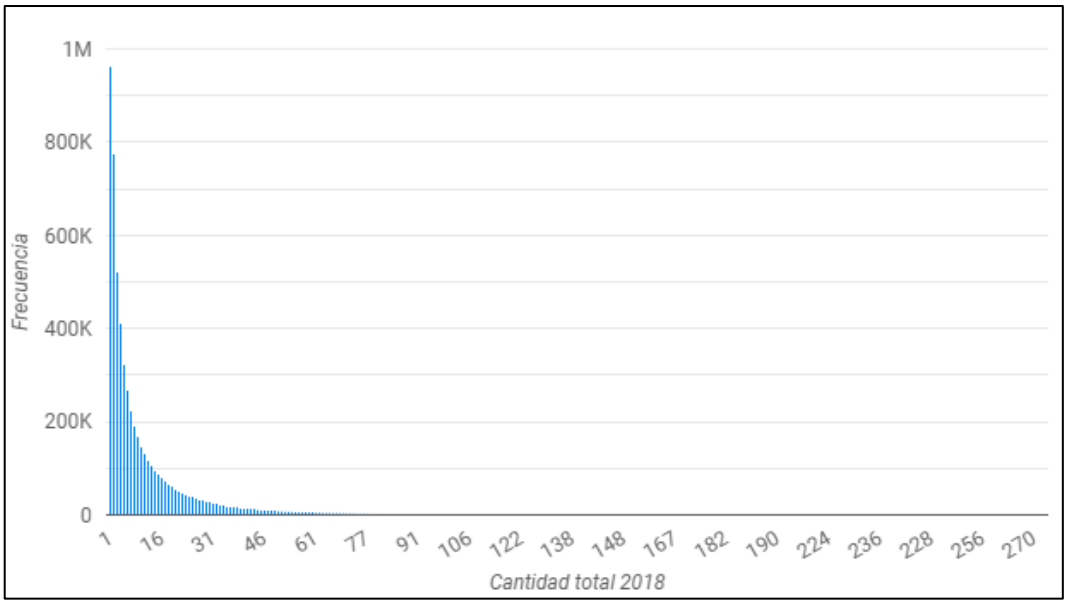

Gráfico 1: Histograma cantidad total productos por cliente año 2018. Fuente: Elaboración propia.

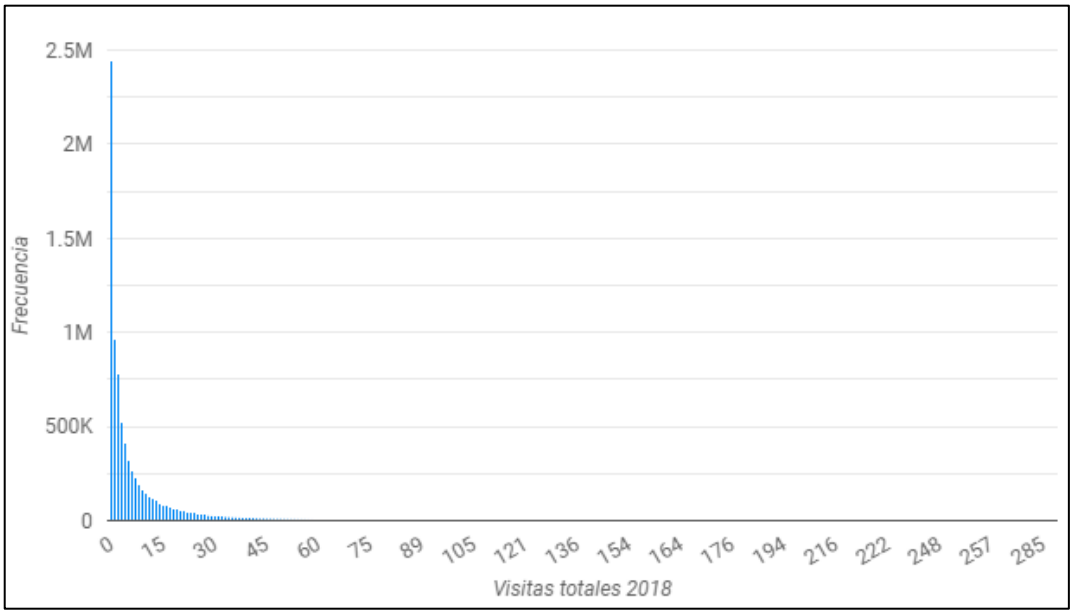

Gráfico 2: Histograma visitas totales por cliente año 2018. Fuente: Elaboración propia.

Es por esta razón que se realizó una transformación de dichas variables aplicando logaritmo sobre los valores. Obteniéndose distribuciones muy semejantes a una Normal.

Así, se calculó la distancia de Mahalanobis usando un umbral dado por la distribución chi-cuadrado con 3 grados de libertad (pues se usaron 3 variables), al 97.5% de confianza, para los años 2016, 2017 y 2018. Tras esto, se eliminaron cerca de 400,000 datos de los ~8,000,000 datos iniciales (aproximadamente 5% de los datos).

Además, para evitar una posible multicolinealidad de variables en los modelos predictivos a utilizar, se obtuvieron las correlaciones existentes entre las variables calculadas, a fin de eliminar ciertas variables que pudiesen entregar información redundante y para reducir el número de variables a utilizar. Las variables con valor absoluto mayor a 0.9 en sus coeficientes de correlación corresponden a las variables de la tabla 2. Se prefirió usar este valor en vez de 0.7 o 0.8, debido a que usando

estos valores se eliminarían demasiadas variables asociadas a las visitas y ventas acumuladas en tiendas físicas, las cuales son necesarias para manejar la estacionalidad del retail en ciertos meses específicos.

Variable
Cantidad de unidades compradas
Visitas con compra en tienda, últimos 21 meses acumulados
Visitas con compra en tienda, últimos 18 meses acumulados
Visitas con compra en tienda, últimos 15 meses acumulados
Visitas con compra en tienda, últimos 12 meses acumulados
Ventas en tienda, últimos 21 meses acumulados
Ventas en tienda, últimos 18 meses acumulados
Ventas en tienda, últimos 15 meses acumulados

Tabla 2: Variables altamente correlacionadas. Fuente: Elaboración propia.

3.2.1 Transformación de datos

En primer lugar, para modelar la probabilidad de primera compra web de cada cliente, se crearon las variables “Visita con compra web agosto 2019” y “Visita con compra web septiembre 2019”. Con valor 1 si el cliente compró al menos un producto en el canal digital de la empresa en agosto (o septiembre) del año 2019, y 0 en caso contrario.

Se agregaron las visitas con compra y la suma acumulada de ventas (compras del cliente, en pesos) para los 3, 6, 9, 12, 15, 18, 21 y 24 meses anteriores, para compras en tiendas y en el canal web, por separado. También se incorporaron estas variables para los días cyber, en primer y segundo semestre, para los años 2017, 2018 y 2019, si aplicaba.

Se agregaron las visitas con compra y la suma acumulada de ventas por categoría (vestuario, electro o decohogar), para los 24 meses anteriores, según canal tienda o canal web. La cantidad de SKUs distintos en total, comprados en 24 meses anteriores también se agregó al dataset.

Se calcularon las recencias de los clientes en sus compras en tienda y en web, tomando de fecha el 1 de agosto de 2019.

También, se calcularon la cantidad de días navegados en los 12 meses anteriores, y en los 24 meses anteriores. Junto con la cantidad de SKUs navegados por categoría (vestuario, electro o decohogar), para los 12 meses anteriores, y los 24 meses anteriores. Otra variable calculada fue la recencia de los clientes en su navegación en los 24 meses anteriores, para el 1 de agosto de 2019.

Además, las siguientes variables fueron transformadas y segmentadas según su “weight of evidence”, mediante la variable dependiente “Visitas con compra web agosto 2019”. El algoritmo usado para esta segmentación creaba distintos grupos con los valores de cada variable, entregando la combinación de grupos con la cual se maximizaba la Information Value de cada una.

Variable	Descripción
Cupo	Cupo ($) en tarjeta de crédito de empresa, del cliente
Número de hijos	Cantidad de hijos del cliente
Edad	Edad del cliente
Renta	Monto renta del cliente

Tras esto, la segmentación de cada variable quedó de la siguiente manera:

Cupo	WoE
[10000, 90000]	29.49
90000+	-71.26
NA	41.81

Hijos	WoE
[0; 1]	-12.12
[2; 3]	10.52
3+	34.50
NA	184.91

Edad	WoE
[18; 35]	-32.02
35+	11.10
NA	62.19

Renta	WoE
[1, 2]	4.40
3+	-58.93
NA	35.53

Donde un WoE negativo indica una mayor proporción de “compras web en agosto de 2019” sobre su total que “no compras web en agosto de 2019” sobre su propio total. La tabla final de variables transformadas se encuentra en Anexos.

3.3 Análisis Exploratorio

Se realizó un análisis exploratorio de los datos a modelar para obtener los primeros insights sobre los clientes de la empresa.

En primer lugar, se contabilizó a los clientes según su antigüedad de transacciones realizadas en la empresa. Los clientes antiguos corresponden a aquellos que realizaron al menos una compra en el último año y en el año predecesor, es decir, han comprado por dos años consecutivos, mientras que los clientes nuevos son aquellos que concretaron al menos una compra en el último año, pero ninguna en el año predecesor. La cantidad aproximada de clientes según su antigüedad se puede observar en las tablas 3 y 4.

Año 2017	Nuevos	Antiguos	Total
Cantidad de clientes	~1,600,000	~3,900,000	~5,500,000
Porcentaje de clientes	29.09%	70.91%	100.00%

Tabla 3: Cantidad de clientes nuevos y antiguos, año 2017. Fuente: Elaboración propia.

Año 2018	Nuevos	Antiguos	Total
Cantidad de clientes	~1,500,000	~4,030,000	~5,530,000
Porcentaje de clientes	27.12%	72.88%	100.00%

Tabla 4: Cantidad de clientes nuevos y antiguos, año 2018. Fuente: Elaboración propia.

Se tiene que la cantidad de clientes totales crece del año 2017 al 2018. Esto pues, aunque la cantidad de clientes nuevos el año 2018 sea menor a la del 2017, aún sigue siendo mayor a la cantidad de clientes fugados del 2018.

Se entiende a un cliente fugado como aquel que realizó al menos una compra en el año predecesor, pero ninguna en el último año. La cantidad de clientes fugados por año se aprecia en la tabla 5.

Año	2017	2018
Cantidad de clientes fugados	~1,400,000	~1,470,000

Tabla 5: Cantidad de clientes fugados de la empresa, años 2017 y 2018. Fuente: Elaboración propia.

Se calcularon, además, para los años 2017 y 2018, la cantidad de clientes que no compró en el canal digital de la empresa en más de un año. Lo anterior se puede observar en el gráfico 3.

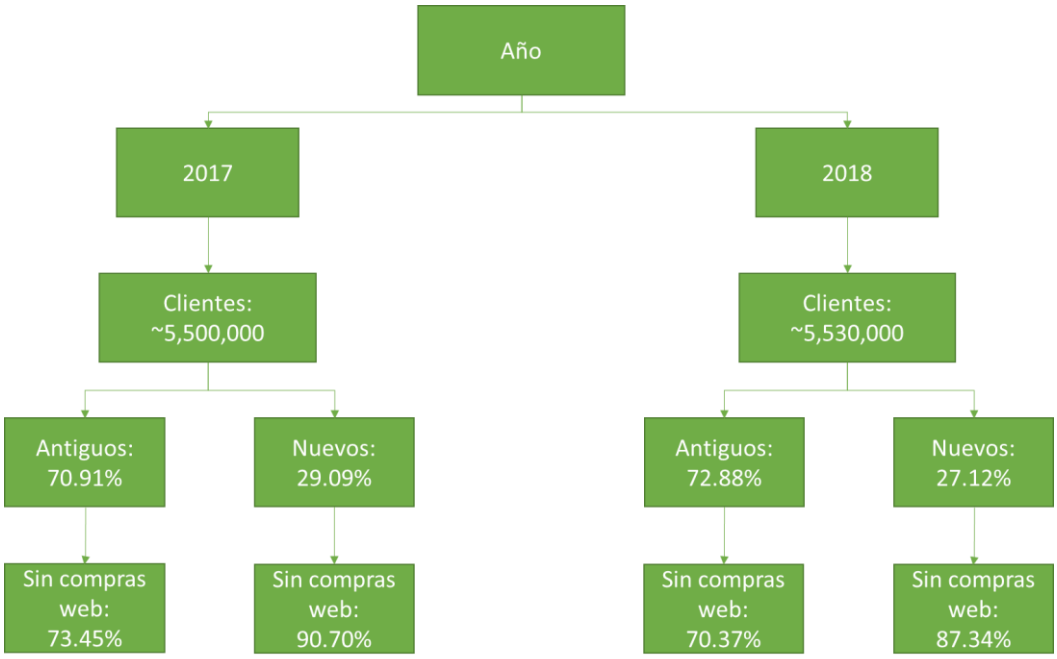

Gráfico 3: Árbol de clientes según antigüedad y compras web, año 2017 y 2018. Fuente: Elaboración propia.

Mediante el gráfico anterior, se puede observar que la cantidad de clientes que deja de comprar (o no ha comprado) en la página web de la empresa disminuye en su proporción en el año 2018 comparado al 2017. De todos modos, sigue siendo una cantidad alta, pues se concluye que sólo el 30% de clientes antiguos ha comprado en el canal web dentro del último año.

Se ha calculado, además, el porcentaje de clientes que realizaron compras en el canal web en el año 2018, según su cantidad de visitas (con compra) en el canal web de la empresa, para el año 2017, para clientes antiguos y nuevos. Estos datos se pueden apreciar en el gráfico 4 y el gráfico 5.

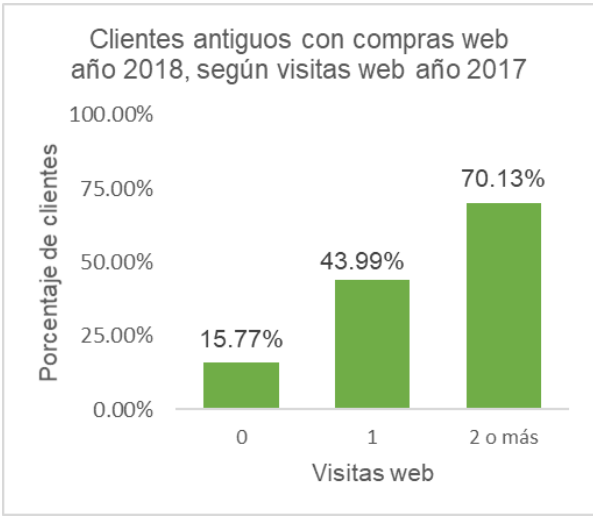

Gráfico 4: Porcentaje de clientes antiguos con compras en canal web año 2018, según visitas web año 2017. Fuente: Elaboración propia.

Gráfico 5: Porcentaje de clientes nuevos con compras en canal web año 2018. según visitas web año 2017. Fuente: Elaboración propia.

De ambos gráficos, es fácil notar que el porcentaje de clientes que nunca visitó (con compra) la web de la empresa el año 2017, es mucho mayor en los clientes que no compran en el canal web el año 2018 (con un 84.23%), comparado con los que sí compran (con un 15.77%). Lo cual es intuitivo, pues al año siguiente (2018), todos estos clientes dejan de comprar (o siguen sin comprar) vía web. Esta diferencia de porcentaje es menor en los clientes antiguos si se comparan con los clientes nuevos. Sin embargo, sigue siendo una alta cantidad. Por otro lado, se tiene que los clientes antiguos que realizaron 2 o más visitas web el año 2017, en su mayoría compran en el canal web el 2018 (70.13% contra un 29.87%). Mientras que para los clientes nuevos que realizaron 2 o más visitas web el año 2017, en su mayoría no compran en el canal web el 2018 (53.02% contra un 46.98%).

El gráfico 6 contiene el promedio de visitas web (con compra) en el año 2017, para los clientes antiguos y nuevos, según si compran en el canal web el año 2018. De este gráfico se observa que, en promedio, de los clientes que no compraron vía web el 2018, los clientes antiguos visitaron la web de la empresa el año 2017 aproximadamente 2 veces en promedio, mientras que los clientes nuevos la visitaron aproximadamente 1 vez en promedio. Por otro lado, los clientes que no compraron vía web el 2018 casi nunca visitaron el canal digital el año 2017, en promedio.

Gráfico 6: Promedio de visitas web año 2017, según compras en canal web año 2018. Fuente: Elaboración propia.

Los gráficos 7 y 8 contienen el porcentaje de clientes que realizan compras web el año 2018, según su sexo y antigüedad. De estos gráficos se observa que casi todos los clientes sin información de su sexo no compraron en el canal digital para el 2018. Además, no se observan grandes diferencias en el porcentaje entre quienes compran y no compran vía web entre sexos distintos.

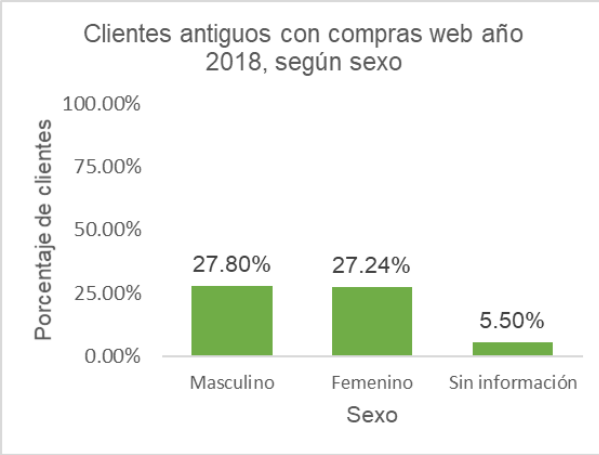

Gráfico 7: Porcentaje de clientes antiguos con compras en canal web año 2018, según sexo. Fuente: Elaboración propia.

Gráfico 8: Porcentaje de clientes nuevos con compras en canal web año 2018, según sexo. Fuente: Elaboración propia.

Según su edad y antigüedad, se graficaron los porcentajes de clientes con compras en el canal web en el año 2018, en los gráficos 9 y 10. Del gráfico 9, se observa que el mayor porcentaje de clientes antiguos con compras web está en los clientes con su edad en el rango de 26 a 45 años, con 32.08% de su total, seguido por quienes poseen una edad mayor a 46 años, con un 23.28% de su total, y por último, le siguen los clientes con su edad en el rango de 18 a 25 años, con un 21.12% de su total (sin contar a los clientes sin información). Por otro lado, del gráfico 10, se observa que el mayor porcentaje de clientes nuevos con compras web está en los clientes con su edad en el rango de 26 a 45 años, con un 13.54% de su total, seguido por los clientes cuya edad está entre los 18 y 25 años, con un 12.48% de su total, y por último, le siguen los clientes mayor a 46 años, con un 7.16% de su total.

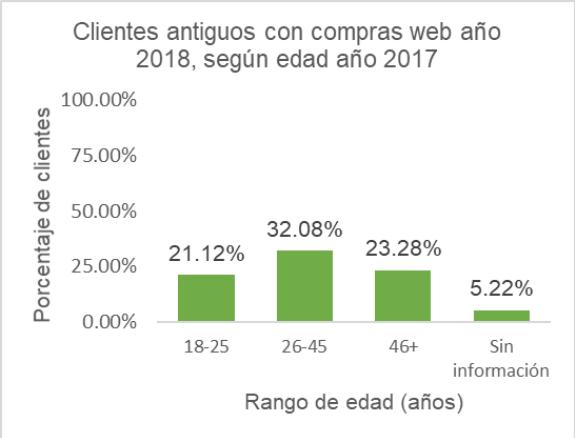

Gráfico 9: Porcentaje de clientes antiguos con compras en canal web año 2018, según edad año 2017. Fuente: Elaboración propia.

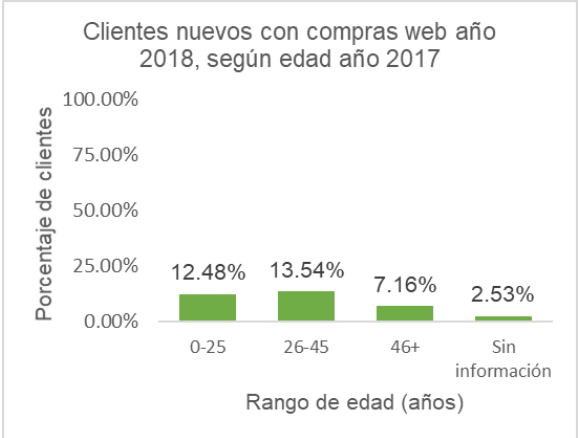

Gráfico 10: Porcentaje de clientes nuevos con compras en canal web año 2018, según edad año 2017. Fuente: Elaboración propia.

El porcentaje de clientes con compras web en el año 2018, según su zona de residencia y antigüedad, para el año 2017, se ha ilustrado en los gráficos 11 y 12. De estos, se aprecia que para los clientes antiguos y nuevos, la zona RM Oriente es la que mayor proporción de clientes con compras web el 2018 posee, con un 39.42% y un 20.28% de su total, respectivamente. Por otro lado, sin contar los clientes sin información de su zona de residencia, los clientes antiguos y nuevos de la zona Sur de Chile son quienes no compraron vía web el 2018 en mayor proporción, con un 76.61% y un 90.87% de su total, respectivamente.

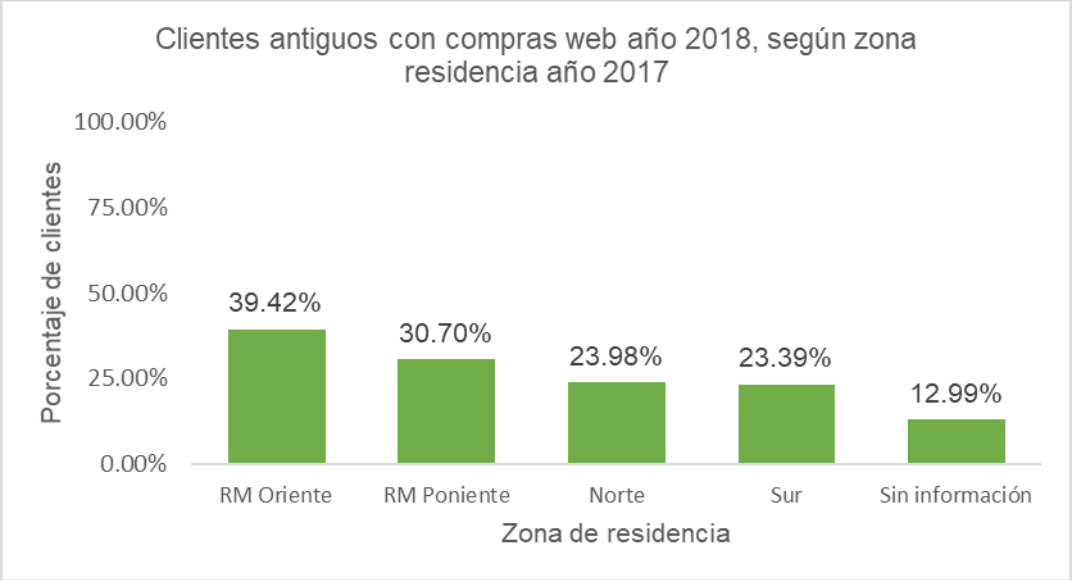

Gráfico 11: Porcentaje de clientes antiguos con compras en canal web año 2018, según zona de residencia año 2017. Fuente: Elaboración propia.

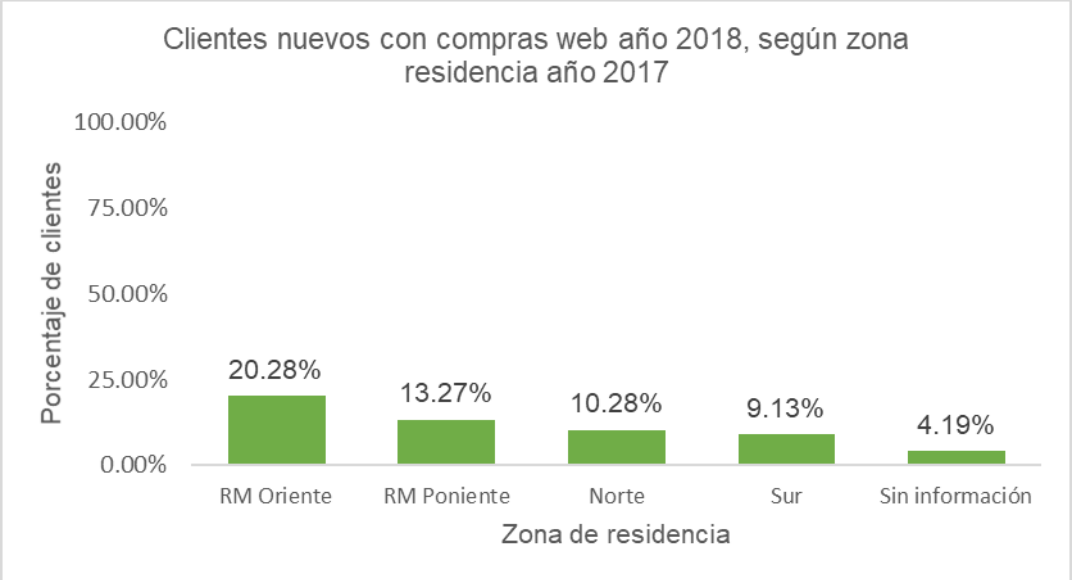

Gráfico 12: Porcentaje de clientes nuevos con compras en canal web año 2018, según zona de residencia año 2017. Fuente: Elaboración propia.

Los gráficos 13 y 14, contienen el porcentaje de clientes con compras web en el año 2018, según su estado civil y antigüedad, para el año 2017. Se observa que, en los clientes antiguos y nuevos, aquellos que eran solteros para el 2017, representaron un mayor porcentaje de compras web el 2018 que los casados, con un 27,81% contra un 25.89% en clientes antiguos, y un 11.12% contra un 8.45% en clientes nuevos.

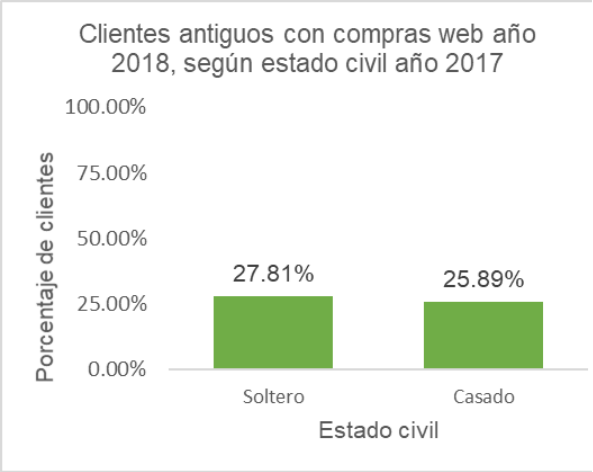

Gráfico 13: Porcentaje de clientes antiguos con compras web año 2018, según estado civil año 2017. Fuente: Elaboración propia.

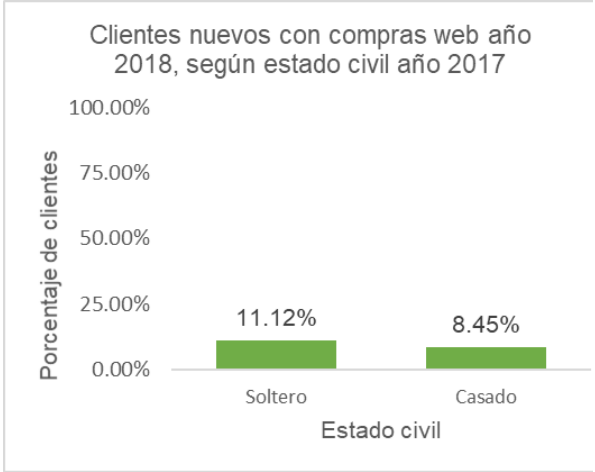

Gráfico 14: Porcentaje de clientes nuevos con compras web año 2018, según estado civil año 2017. Fuente: Elaboración propia.

En los gráficos 15 y 16 se ilustran los promedios de días navegados en el canal web de la tienda, para el año 2016 y 2017, según antigüedad de los clientes, y divididos en quienes compran y no compran en el canal digital para el año 2018. De ellos, se observa que los clientes antiguos son quienes más días navegan en promedio por año, con un 11.45 y 11.47 para los clientes sin compras web en el año 2016 y 2017, respectivamente, contra los 1.58 y 3.39 días en promedio que navegan por año los clientes con compras web, nuevos, en el año 2016 y 2017, respectivamente. Por otro lado, se tiene que los clientes sin compras web navegaron menos días en promedio para el año 2016 y 2017 que los clientes con compras web, con excepción en los clientes nuevos, donde los clientes sin compras web navegaron en promedio un poco más que aquellos con compras web (2.39 contra un 1.58).

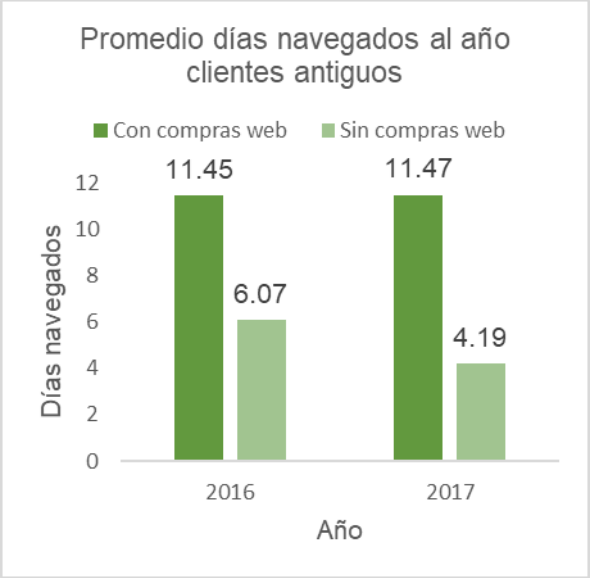

Gráfico 15: Promedio días navegados clientes antiguos año 2017, según compras canal web 2018. Fuente: Elaboración propia.

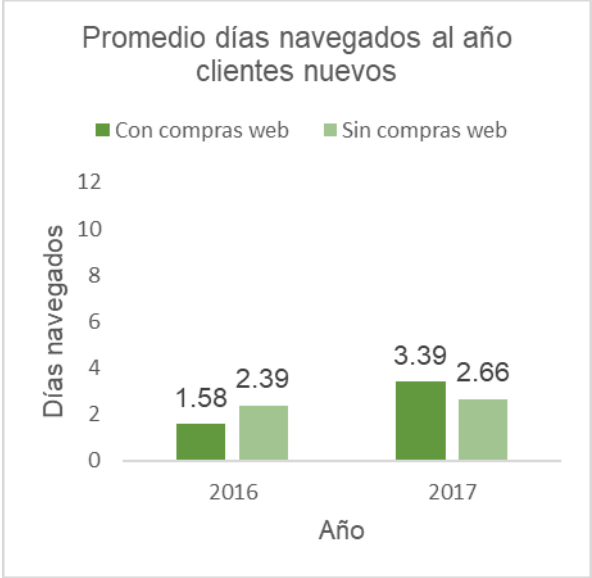

Gráfico 16: Promedio días navegados clientes nuevos año 2017, según compras canal web 2018. Fuente: Elaboración propia.

Los gráficos 17 y 18 ilustran el promedio de SKUs navegados por categoría y año, según si compraron vía web el año 2018, para clientes antiguos y nuevos del 2017, respectivamente. De ambos gráficos se observa que el mayor promedio de SKUs

navegados al año lo poseen los productos de categoría vestuario, le siguen los productos de categoría “electro” y luego “decohogar”. Por otro lado, se aprecia que el promedio de productos navegados de categoría vestuario aumentó en gran cantidad en los clientes que realizaron compras web el año 2018, pasando de un promedio de 43.41 SKUs navegados a 56.68, y de 7.6 a 27.58, para clientes antiguos y nuevos, respectivamente.

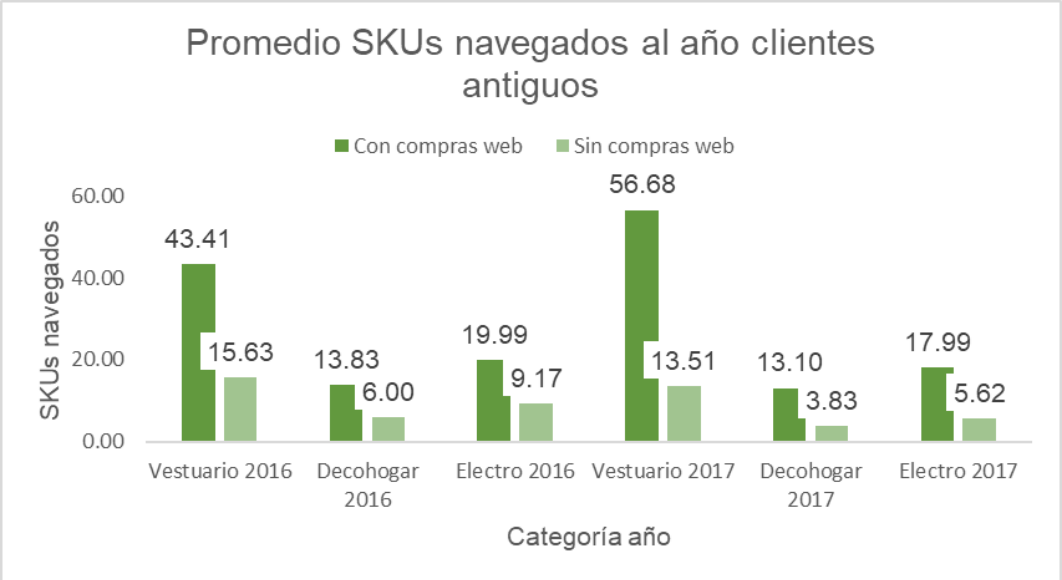

Gráfico 17: Promedio SKUs navegados por categoría clientes antiguos año 2017, según compras canal web 2018. Fuente: Elaboración propia.

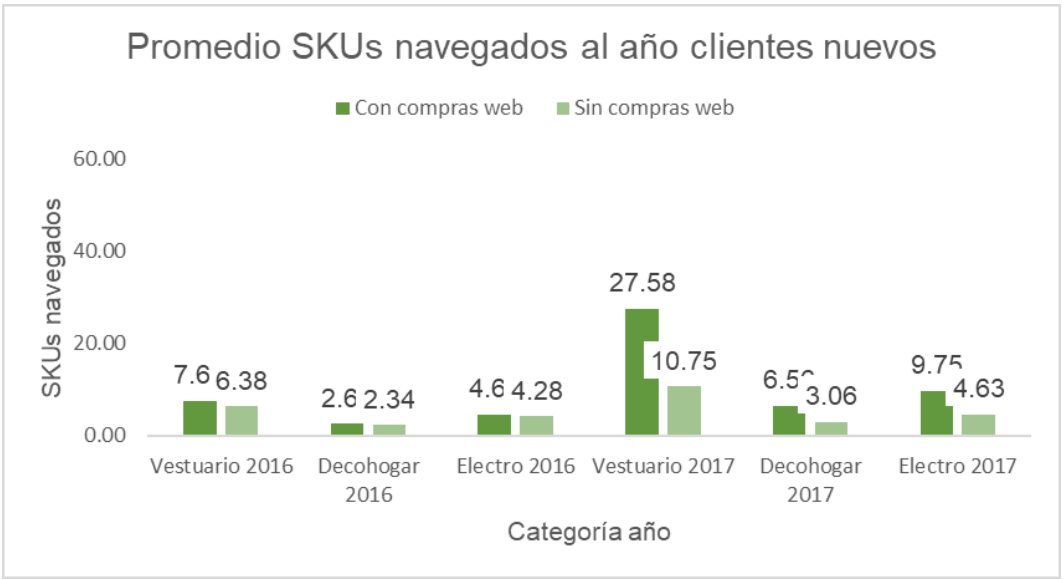

Gráfico 18: Promedio SKUs navegados por categoría clientes nuevos año 2017, según compras canal web 2018. Fuente: Elaboración propia.

Los gráficos 19 y 20 ilustran los promedios de visitas con compra por cliente según categoría y canal para el año 2017, por antigüedad, y tomando en cuenta si compran vía web el año 2018 o no. De ambos gráficos, se observa que los clientes antiguos visitan en mayor promedio la categoría vestuario en el canal tienda por sobre cualquier otra combinación, con 17.63 visitas con compra en promedio. Por otro lado, al ordenar el promedio de visitas con compra de mayor a menor, según categoría y canal para aquellos clientes antiguos con compra web el año 2018, se tiene el orden: tienda

vestuario, web vestuario, tienda electro, web electro, tienda decohogar y web decohogar; en cambio, para aquellos clientes antiguos sin compra web: tienda vestuario, tienda electro, tienda decohogar, web electro, web vestuario y web decohogar. Además, haciendo el mismo ordenamiento para clientes nuevos con compras web el año 2018, se tiene el orden: tienda vestuario, web vestuario, web electro, tienda electro, web decohogar y tienda decohogar; y para aquellos sin compras web: tienda vestuario, tienda electro, tienda decohogar, web vestuario y web electro, y web decohogar. Es decir, hay diferencias en las preferencias de categorías y canales según antigüedad de los clientes y según si compran vía web el año 2018.

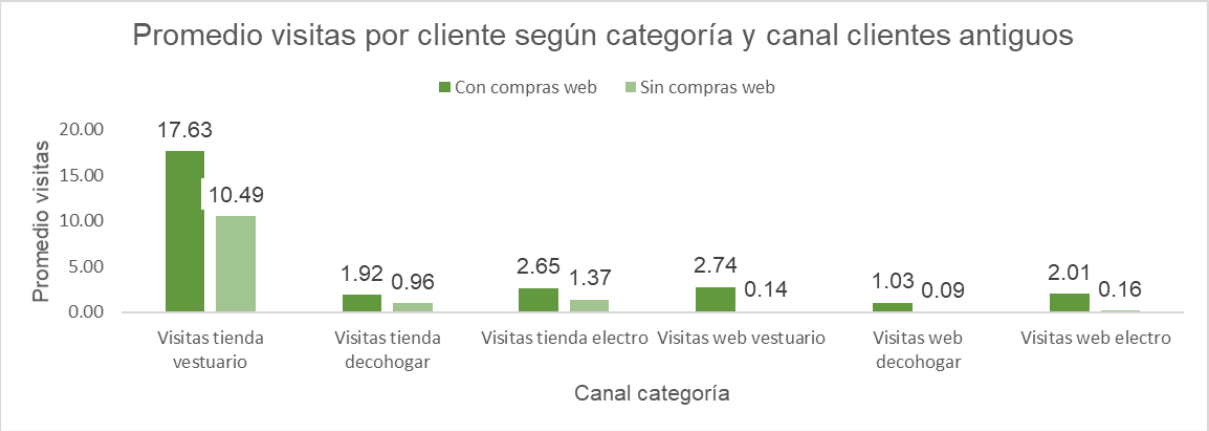

Gráfico 19: Promedio visitas por cliente según categoría y canal, clientes antiguos año 2017, según compras canal web 2018. Fuente: Elaboración propia.

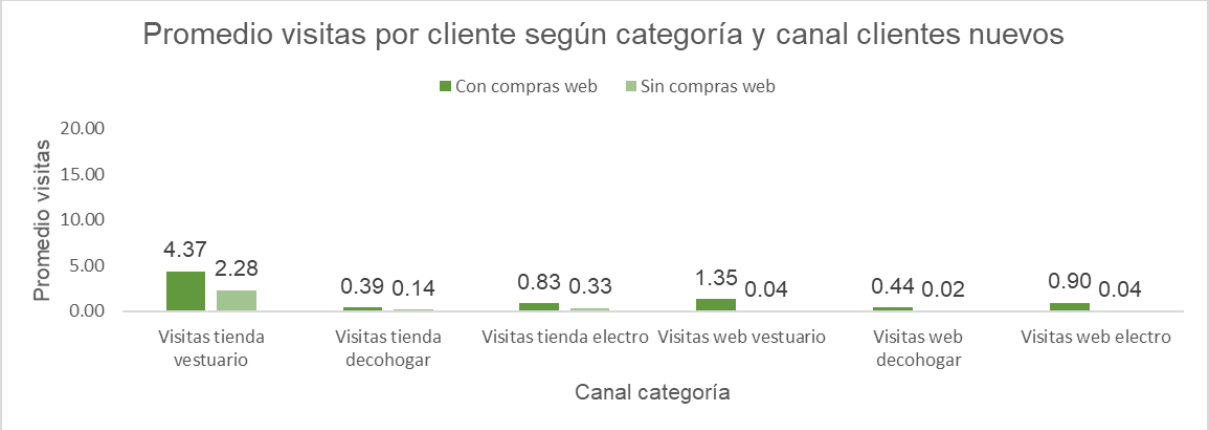

Gráfico 20: Promedio visitas por cliente según categoría y canal, clientes nuevos año 2017, según compras canal web 2018. Fuente: Elaboración propia.

3.4 Propensión de primera compra online

Para estimar la probabilidad de que un cliente compre en el canal web se utilizó como variable dependiente una variable binaria con valor 1 si el cliente compra en el canal web en agosto de 2019 y 0 en caso contrario.

Además, se trabajó con una muestra representativa usando el 2% de los datos, para no usar demasiados recursos computacionales, pero de manera de poder procesar datos suficientes para los modelos estadísticos. Se comprobó la representatividad de la muestra al comparar las distribuciones de ambos datasets (el original y la muestra), según la variable dependiente, el sexo, la edad, la zona de residencia, etc.

Posterior a ello, se dividió la muestra en un set de datos de entrenamiento y otro de testeo, usando una proporción 80:20, respectivamente, para evitar el sobreajuste de los parámetros de cada modelo.

Por otro lado, de los ~6,000,000 de clientes totales (que han realizado alguna transacción entre el 31 de julio de 2017 y el 31 de julio de 2019, pero no han comprado en el canal web en los últimos 12 meses), sólo ~70,000 realizaron al menos una compra en el canal web de la empresa en agosto de 2019, es decir, el 1.17% de los clientes totales posibles. Debido a esto, se realizó un balance de la base de datos, para que los modelos calculados no tiendan a predecir todas las observaciones como clases negativas (no compra web en agosto 2019), y así conseguir un 99% de Accuracy, pero 0% de Recall. Para ello, se probaron técnicas de downsampling, upsampling y SMOTE (Synthetic Minority Over-Sampling Technique). Siendo esta última la técnica seleccionada para este trabajo, debido a que el upsampling requería un alto nivel de procesamiento y el downsampling presentaba resultados de menor calidad a los de la técnica SMOTE. Con el fin de no sesgar los coeficientes obtenidos de los modelos estadísticos, estas técnicas fueron usadas sólo en el dataset de entrenamiento y no en el de testeo, pues, de esta manera, los resultados obtenidos por los modelos en el set de testeo son extrapolables a cualquier dataset de la empresa a futuro que posea los mismos atributos usados en este trabajo.

Luego, se realizó un procedimiento de Cross Validation usando 5 segmentos del dataset de entrenamiento, con una proporción de datos de 75:25, para cada set de training y validation, respectivamente. Así, los resultados de las métricas para cada modelo fueron calculados mediante el promedio obtenido entre los 5 segmentos de validación.

3.4.1 Entrenamiento de modelos

Se entrenaron los siguientes modelos, usando boosting (se reduce el error de predicción al aprender de iteraciones anteriores del modelo) para cada uno y probando distintas configuraciones.

Configuración de parámetros modelo árbol de decisión C5.0:

- Tipo: árbol o reglas.
- Con o sin “winnow” (usa sólo las variables más relevantes).
- Iteraciones: 1, 10 o 20.

N°	Modelo	Winnow	Iteraciones	ROC	Sens	Spec
1	Reglas	NO	1	0.8225	0.8845	0.7320
2	Reglas	NO	10	0.8877	0.7960	0.8105
3	Reglas	NO	20	0.8881	0.8290	0.7855
4	Reglas	SÍ	1	0.8264	0.8665	0.7505
5	Reglas	SÍ	10	0.8870	0.8065	0.8090
6	Reglas	SÍ	20	0.8882	0.8075	0.8045
7	Árbol	NO	1	0.8561	0.8645	0.7275
8	Árbol	NO	10	0.8870	0.8250	0.7865
9	Árbol	NO	20	0.8901	0.8385	0.7795
10	Árbol	SÍ	1	0.8616	0.8645	0.7390
11	Árbol	SÍ	10	0.8884	0.8300	0.7870
12	Árbol	SÍ	20	0.8903	0.8330	0.7720

Tabla 6: Configuración de parámetros árbol C5.0. Fuente: Elaboración propia.

Donde el modelo escogido basado en el ROC fue el tipo árbol, con winnow, usando 20 iteraciones para el boosting (modelo 12).

Configuración de parámetros modelo Extreme Gradient Boosting Lineal:

Cabe señalar que la función de pérdida que se busca minimizar es:

$$L = \sum_{i=0}^{n} loss(y_{res}, h(x)) + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 + \alpha \sum_{j=1}^{T} |w_j|$$

- Lambda: 0, 0.1 o 0.0001.
- Alpha: 0, 0.1 o 0.0001.
- Rounds para boosting: 50, 100 o 150.

N°	Lambda	Alpha	Rounds	ROC	Sens	Spec
1	0	0	50	0.8813	0.8330	0.7750
2	0	0	100	0.8734	0.8275	0.7760
3	0	0	150	0.8722	0.8275	0.7685
4	0	0.0001	50	0.8809	0.8300	0.7735
5	0	0.0001	100	0.8727	0.8230	0.7730
6	0	0.0001	150	0.8723	0.8275	0.7670
7	0	0.1	50	0.8805	0.8415	0.7640
8	0	0.1	100	0.8751	0.8195	0.7690
9	0	0.1	150	0.8749	0.8175	0.7745
10	0.0001	0	50	0.8804	0.8315	0.7730
11	0.0001	0	100	0.8732	0.8260	0.7745
12	0.0001	0	150	0.8727	0.8255	0.7740
13	0.0001	0.0001	50	0.8808	0.8350	0.7745
14	0.0001	0.0001	100	0.8745	0.8250	0.7730
15	0.0001	0.0001	150	0.8742	0.8260	0.7690
16	0.0001	0.1	50	0.8805	0.8415	0.7640
17	0.0001	0.1	100	0.8753	0.8200	0.7705
18	0.0001	0.1	150	0.8749	0.8155	0.7715
19	0.1	0	50	0.8768	0.8335	0.7665
20	0.1	0	100	0.8725	0.8330	0.7700
21	0.1	0	150	0.8724	0.8345	0.7670
22	0.1	0.0001	50	0.8768	0.8335	0.7665
23	0.1	0.0001	100	0.8725	0.8350	0.7690
24	0.1	0.0001	150	0.8709	0.8320	0.7665
25	0.1	0.1	50	0.8795	0.8360	0.7675
26	0.1	0.1	100	0.8746	0.8280	0.7700
27	0.1	0.1	150	0.8739	0.8250	0.7660

Tabla 7: Configuración de parámetros Extreme Gradient Boosting Lineal. Fuente: Elaboración propia.

Donde el modelo escogido según el mayor ROC fue aquel con Lambda = 0, Alpha = 0 y Rounds = 50 (modelo 1).

Configuración de parámetros modelo Boosted Logit:

- Iteraciones: 11, 21 o 31.

N°	Iteraciones	ROC	Sens	Spec
1	11	0.8165	0.8135	0.7330
2	21	0.8285	0.8595	0.7005
3	31	0.8301	0.8425	0.7055

Tabla 8: Configuración de parámetros Boosted Logit. Fuente: Elaboración propia.

Se observa que al pasar de 11 iteraciones a 21 el ROC aumenta en 0.012, la Sensitivity aumenta en 0.046 y la Specificity disminuye en 0.0325. Por otro lado, al pasar de 21 iteraciones a 31 el ROC aumenta en 0.0016, la Sensitivity disminuye en 0.017 y la Specificity aumenta en 0.050. Es por ello que, debido a la disminución en el crecimiento del ROC, al decaimiento en la Sensitivity, y al elevado tiempo de procesamiento, se decide mantener el máximo de iteraciones en 31.

Así, el modelo elegido por tener el mayor ROC y el segundo mayor valor de Sensitivity fue aquel con 31 iteraciones (modelo 3).

Configuración de parámetros modelo Extreme Gradient Boosting Árbol:

- Eta: 0.3 o 0.4. Tasa de aprendizaje del árbol.
- Máxima profundidad: 1, 2 o 3. Número de niveles del árbol.
- Ratio submuestral: 0.6 o 0.8. Ratio submuestral de columnas.
- Submuestra: 0.5, 0.75 o 1.0. Porcentaje submuestral.
- Rounds para boosting: 50, 100 o 150.

Los resultados de entrenamiento de este modelo se encuentran en Anexos. Donde se observa que el mejor modelo es aquel con Rounds = 50, Máxima profundidad = 3, eta = 0.3, Ratio submuestral = 0.8 y submuestra = 1 (modelo 52 del anexo).

3.4.2 Resultados de modelos

Los modelos estadísticos previamente entrenados fueron utilizados con el fin de obtener la probabilidad de que un cliente compre en el canal web de la empresa, prediciendo esto en el mes de septiembre de 2019. Para ello, se probaron los modelos con un umbral de clasificación de clases positivas (sí compra web) por defecto de 0.5, y con un umbral óptimo de clasificación, el cual varía para cada modelo, en función de optimizar el Recall.

Así, los resultados obtenidos fueron los siguientes:

Modelo	Accuracy	AUC	Recall (Sensitivity)	Precision	F-Score
C5.0	84.74%	0.7409	43.23%	2.93%	0.0548
C5.0 umbral óptimo	69.02%	0.7409	68.56%	2.24%	0.0433
Extreme Gradient Boosting Lineal	81.89%	0.7291	50.22%	2.28%	0.0537
Extreme Gradient Boosting Lineal umbral óptimo	77.72%	0.7291	56.33%	2.57%	0.0492
Boosted Logit	90.00%	0.6157	26.20%	2.82%	0.0509
Boosted Logit umbral óptimo	90.00%	0.6157	26.20%	2.82%	0.0509
Extreme Gradient Boosting Árbol	84.83%	0.7481	44.98%	3.06%	0.0572
Extreme Gradient Boosting Árbol umbral óptimo	64.05%	0.7481	75.98%	2.13%	0.0414

Tabla 9: Resultados modelos estadísticos probabilidad primera compra web en septiembre 2019. Fuente: Elaboración propia.

De la tabla 9 se puede observar que el modelo que obtiene una mayor Accuracy es el Boosted Logit, con un 90%, seguido por el modelo Extreme Gradient Boosting tipo árbol, con un 84.83%. Los modelos que poseen mayor AUC son el Extreme Gradient Boosting tipo árbol, con un 0.7481, seguidos por los modelos de árbol C5.0, con un 0.7409. El modelo con el mayor Recall es el Extreme Gradient Boosting tipo árbol con umbral óptimo, con un 75.98%, seguido por el modelo C5.0 con umbral óptimo, con un 68.56%. El modelo que obtuvo el mayor valor de Precision es el Extreme Gradient Boosting tipo árbol, con un 3.06%, seguido por el C5.0 con un 2.93%. Finalmente, el modelo con el mayor F-Score es el Extreme Gradient Boosting tipo árbol, con un 0.0572, seguido por el modelo C5.0, con un 0.0548.

De todos modos, en este caso, la Precision no es muy relevante, ya que los clientes que realmente compran en la web son una cantidad muy pequeña, y etiquetar (erróneamente) a un cliente que no compra realmente en el canal digital como uno que sí lo hace no es un error muy grave para sus consecuencias (se enviaría una promoción a un cliente que probablemente no la usaría, lo que tiene un costo casi despreciable).

Es así, que **se prefiere utilizar el modelo árbol de decisión C5.0**. Ya que usando el umbral óptimo de clasificación posee una Accuracy cercana al 70%, uno de los mayores valores de AUC, y un Recall cercano al 70%. Es decir, es un modelo que predice bien la mayoría de los datos, en todos los umbrales de clasificación, y predice bien la mayoría de los clientes que realmente realizan una compra web. Además, a diferencia de los modelos de Extreme Gradient Boosting, el árbol de decisión C5.0 es más intuitivo y puede ser replicado en la mayoría de las herramientas computacionales utilizadas por la empresa.

Por otro lado, en el gráfico 21, se presentan los valores de Lift Acumulado del modelo recién descrito, ordenado según los deciles con mayor propensión de los datos observados. Del gráfico, se aprecia que con un 10% o 20% de los datos totales con mayor propensión, se obtiene un lift acumulado de 3.49 o 2.58, respectivamente. Es decir, al seleccionar el 10% de los datos con mayor propensión, usando el modelo

escogido, uno puede esperar 3.49 veces el número total de “primeros compradores web” encontrados al seleccionar el 10% de los datos aleatoriamente sin un modelo.

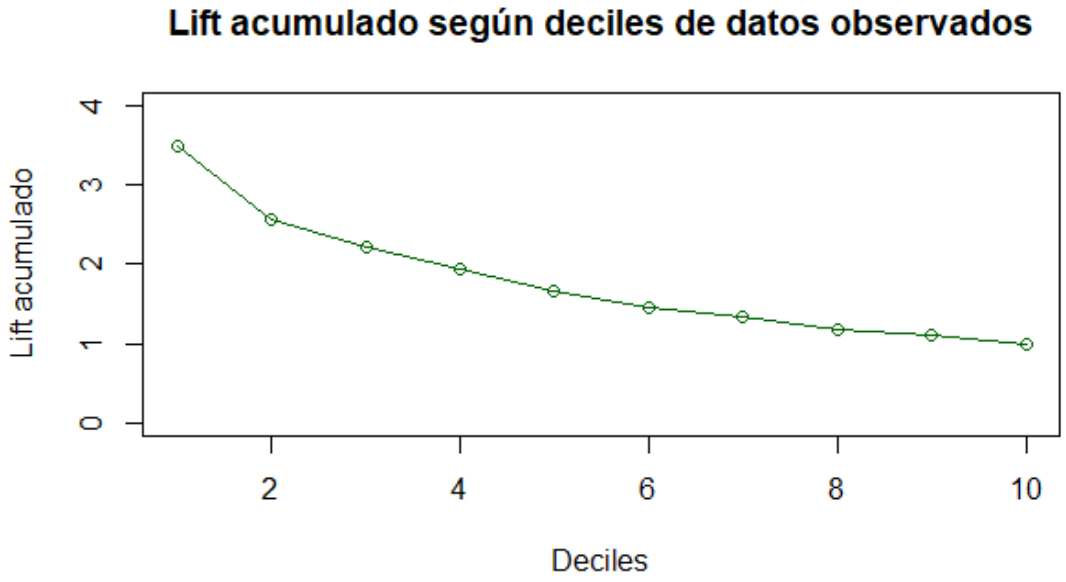

Gráfico 21: Lift acumulado según deciles de datos observados, modelo árbol C5.0. Fuente: Elaboración propia.

El gráfico 22, muestra la Ganancia del modelo ordenada según los deciles con mayor propensión de los datos observados. Donde se aprecia que al seleccionar el top 10% o 20% de los clientes más propensos a comprar vía web usando el modelo escogido, se cubre el 34.93% o 51.52% de los eventos totales (primeros compradores web). Además, con los tres deciles más propensos, se cubre el 66.38% de los eventos totales, y con los seis deciles más propensos, se cubre el 86.90% de los eventos totales.

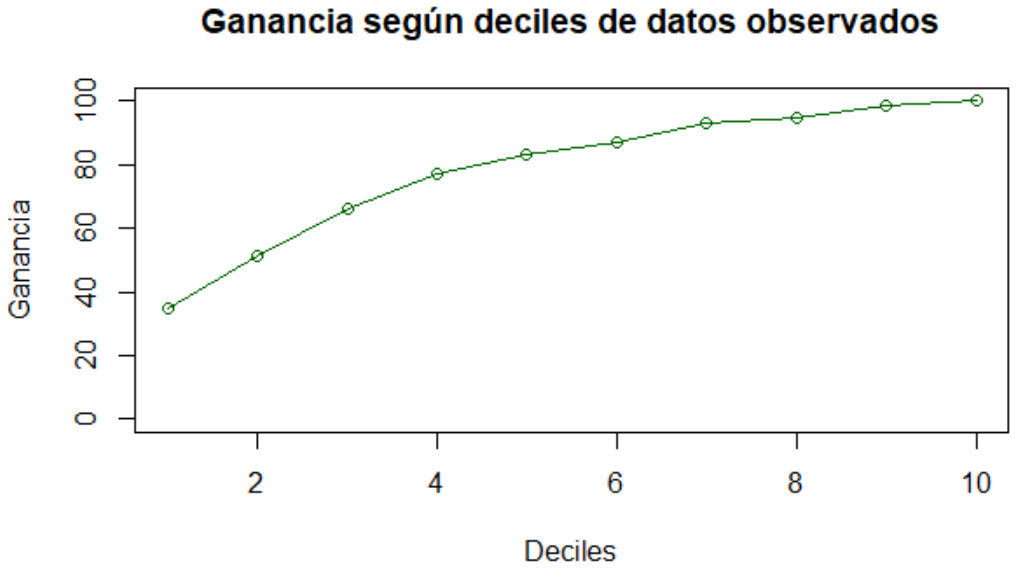

Gráfico 22: Ganancia según deciles de datos observados, modelo árbol C5.0. Fuente: Elaboración propia.

3.5 Variables más relevantes según modelo escogido

En la tabla 10, se pueden apreciar las 10 variables más relevantes para el modelo elegido según el porcentaje de observaciones de entrenamiento que caen en todos los nodos generados tras una división en el que ha participado el predictor. De ésta, se observa que el top 3 de variables más relevantes tiene relación con el comportamiento de compra de los clientes, sobre todo en el canal web, lo cual es intuitivo, pues a mayores visitas con compra anteriores, mayor probabilidad de volver a comprar el mes próximo. Luego, destacan datos de navegación (quien más navega es más proclive a comprar web), y cupo en la tarjeta de crédito de la empresa, es decir, a mayor capacidad financiera, mayor probabilidad de comprar en la web de la empresa.

Variable	Uso de atributo
Visitas con compra en tienda últimos 3 meses acumulados	100.00%
Visitas con compra en web últimos 18 meses acumulados	100.00%
Visitas con compra en vestuario, web, últimos 24 meses acumulados	100.00%
Cantidad de días navegados en los últimos 12 meses	100.00%
Cantidad de SKUs navegados vestuario, en los últimos 12 meses	100.00%
Cupo en tarjeta de crédito de la empresa	100.00%
Visitas con compra en web días Cyber mayo 2018	94.25%
Cantidad de SKUs navegados electro, en los últimos 24 meses	91.85%
Visitas con compra en electro, tienda, últimos 24 meses acumulados	91.38%
Cantidad de hijos	89.65%

Tabla 10: Variables más relevantes de árbol de decisión. Fuente: Elaboración propia.

3.6 Caracterización propensión primera compra web

Luego de aplicar el modelo elegido al set de datos de testeo, se procedió a graficar en boxplots los distintos cuartiles de clientes según su valor de propensión a realizar una primera compra web. Esto se hizo para cada variable categórica, sin mantener estáticas las demás variables dentro de cada categoría de clientes graficada.

El gráfico 23 ilustra la propensión a primera compra web según el tipo de tarjeta de crédito del cliente de la empresa. En éste, se puede observar que los cuartiles de clientes con menor probabilidad de primera compra web son de aquellos para los cuales no se posee información acerca del tipo de tarjeta de crédito que poseen, posiblemente porque no posean una tarjeta de crédito de la empresa. Además, se aprecia que, a mejor categoría de tarjeta de crédito, más altos son los cuartiles de propensión a primera compra web. Así, la mediana de la propensión descrita es aproximadamente 0.19 para clientes sin información de su tipo de tarjeta de crédito, cerca de 0.25 para clientes con tarjeta de crédito “Normal”, casi 0.5 para clientes con tarjeta de crédito “Premium”, y cerca de 0.6 para clientes con tarjeta de crédito “Elite”.

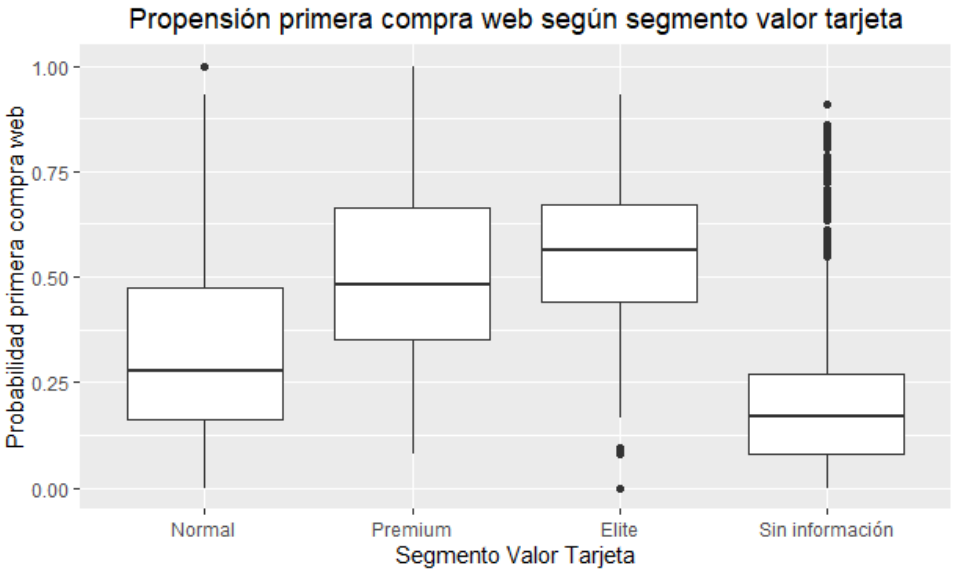

Gráfico 23: Propensión primera compra web según segmento valor tarjeta, modelo árbol C5.0. Fuente: Elaboración propia.

Por otro lado, en el gráfico 24 se observa la propensión a realizar la primera compra web, según el estado civil de los clientes de la empresa. Se puede apreciar que los tres primeros cuartiles de clientes con propensión a primera compra web más bajos (menores a 0.125) corresponden a clientes de los cuales no se tiene información de su estado civil. Los clientes solteros, por otra parte, poseen sus tres primeros cuartiles de propensión a primera compra web más altos que los demás clientes, con una mediana de aproximadamente 0.25, versus la mediana de 0.19 de los clientes casados.

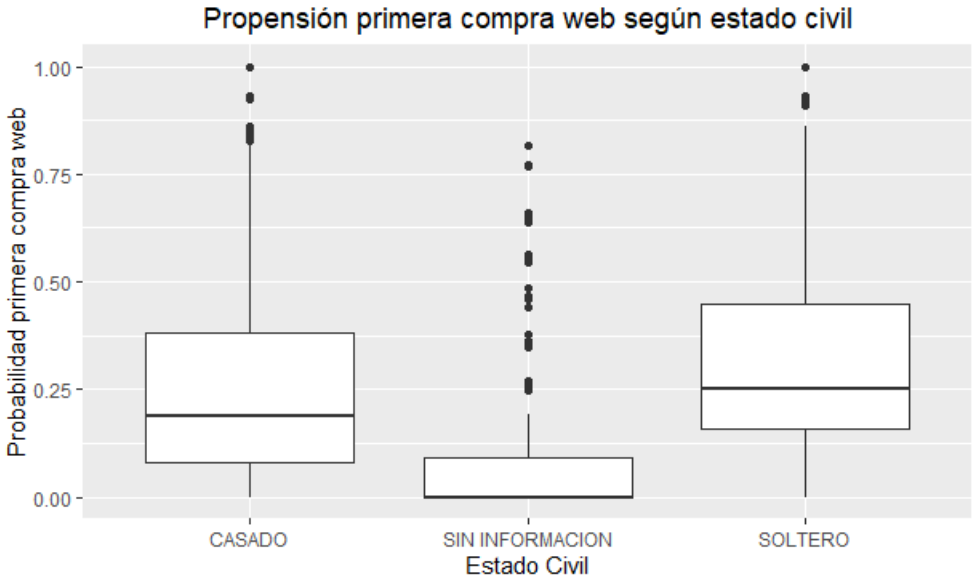

Gráfico 24: Propensión primera compra web según estado civil, modelo árbol C5.0. Fuente: Elaboración propia.

En el gráfico 25 se observa la propensión a primera compra online según la zona de residencia de los clientes. Mediante éste, se aprecia que no hay mayores diferencias entre los clientes para los cuales se tiene el registro de su zona de residencia. En cambio, los clientes de los cuales no se tiene información de su zona de residencia, poseen los menores valores de propensión a primera compra web. De hecho, sus tres

primeros cuartiles de propensión son menores a 0.26, en contraste a los demás clientes, cuyo tercer cuartil es cercano a 0.5.

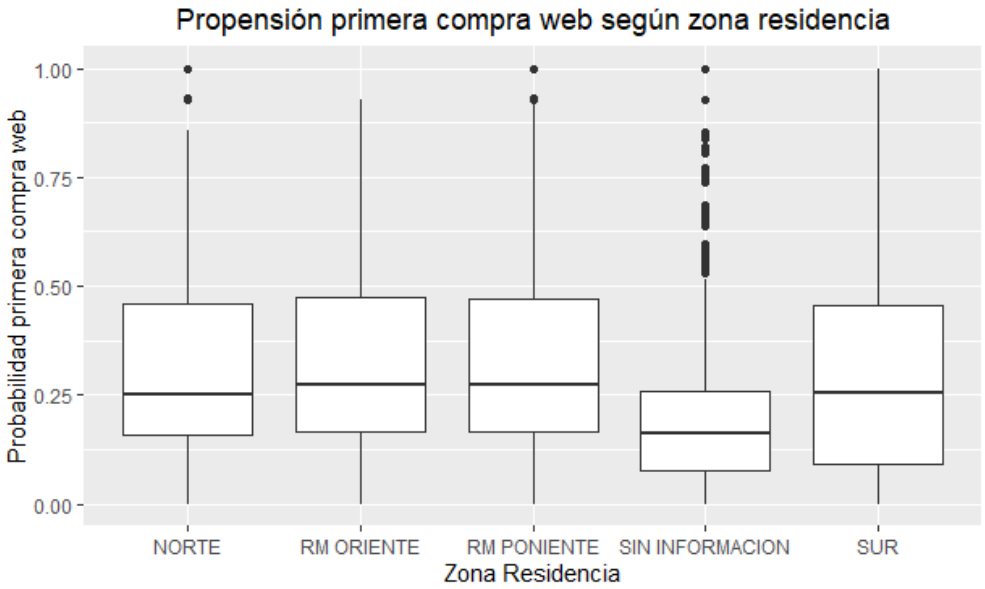

Gráfico 25: Propensión primera compra web según zona de residencia, modelo árbol C5.0. Fuente: Elaboración propia.

El gráfico 26 presenta la propensión a primera compra web según el sexo de los clientes. Se aprecia que los tres primeros cuartiles de propensión a primera compra web de los clientes de sexo femenino son ligeramente más altos que aquellos de los clientes de sexo masculino. Ambos sexos poseen una mediana de aproximadamente 0.25 en su propensión. Sin embargo, los clientes para los cuales no se tiene información acerca de su sexo poseen sus tres primeros cuartiles de propensión con valores menores a 0.14.

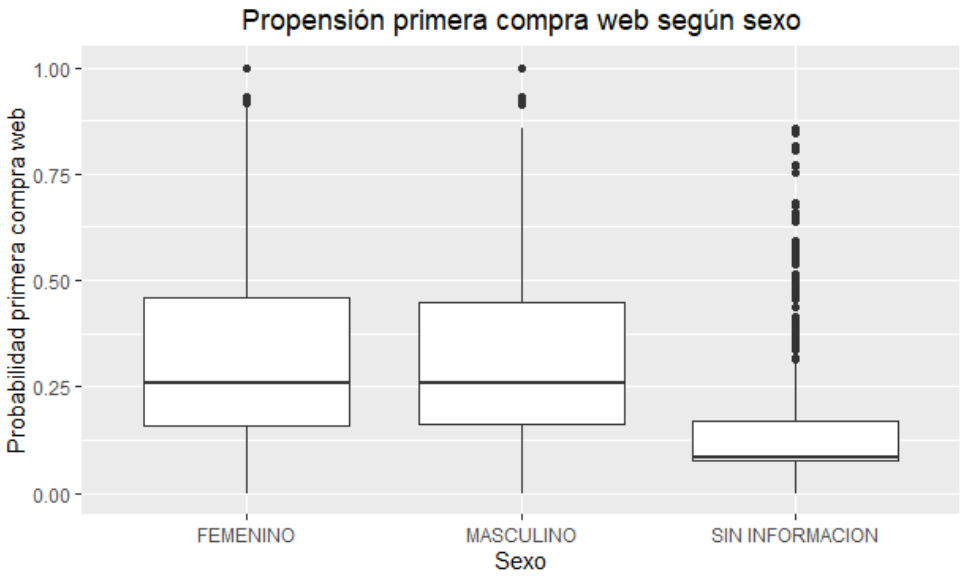

Gráfico 26: Propensión primera compra web según sexo, modelo árbol C5.0. Fuente: Elaboración propia.

El gráfico 27 ilustra la probabilidad de primera compra web según el cupo en la tarjeta de crédito del cliente de la empresa. Se observa que la mediana de la propensión de los clientes para los cuales no se tiene información acerca de su cupo crediticio

(posiblemente porque no tengan tarjeta de crédito en la empresa) es cercana a 0.19, mientras que en los clientes con cupo entre $10.000 y $90.000 la mediana es cercana a 0.26, y en los clientes con un cupo mayor a $90.000 este valor es cercano a 0.5.

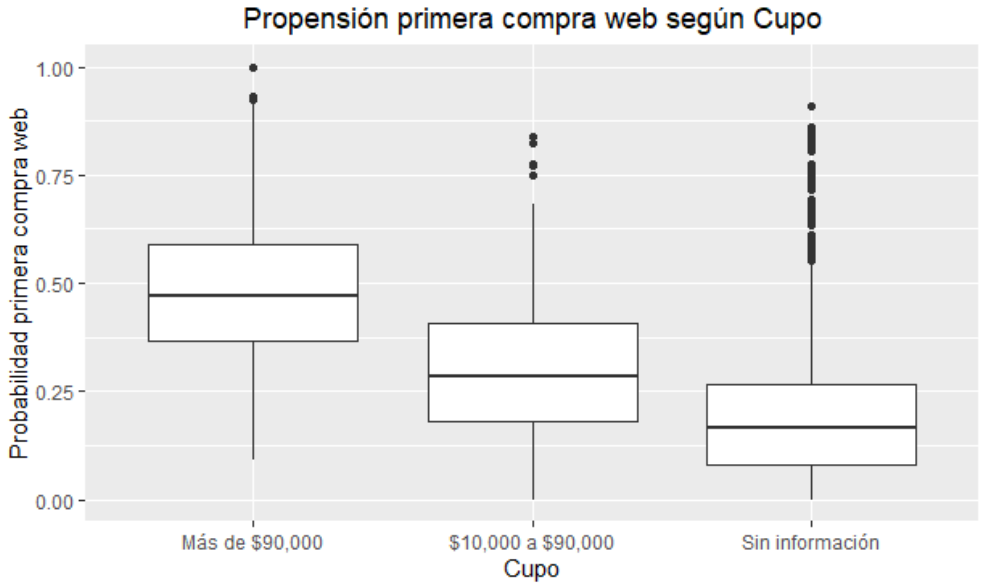

Gráfico 27: Propensión primera compra web según cupo en tarjeta de crédito, modelo árbol C5.0. Fuente: Elaboración propia.

En el gráfico 28, se observa la propensión a primera compra web según la renta de los clientes segmentada en deciles. En este gráfico se aprecia que la mediana de la propensión para los clientes de los cuales no se tiene información acerca de su renta es cercana a 0.125, mientras que los 3 primeros cuartiles poseen valores menores a 0.260. Por otro lado, la mediana de los clientes con un decil de renta entre 1 y 2 es cercana a 0.260, y sus 3 primeros cuartiles poseen valores menores a 0.490. Por último, los clientes con un decil de renta igual o mayor a 3, poseen la mediana de su propensión a primera compra web cercana a 0.375, y sus 3 primeros cuartiles son menores a 0.600.

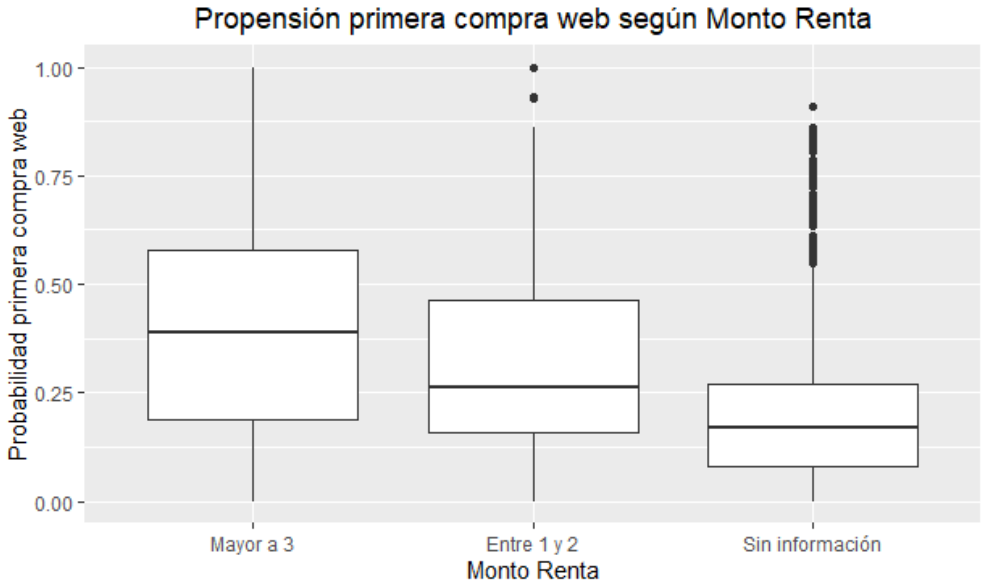

Gráfico 28: Propensión primera compra web según renta de clientes, modelo árbol C5.0. Fuente: Elaboración propia.

El gráfico 29 ilustra los valores de propensión a realizar la primera compra en el canal web de la empresa para los clientes de ésta, según su edad. Mediante este gráfico, se observa que para los clientes de los cuales no se tiene información de su edad, la mediana de su propensión tiene un valor cercano a 0.100, mientras que los 3 primeros cuartiles poseen valores menores a 0.170. Los demás clientes tienen una mediana de propensión cercana a 0.250 y sus 3 primeros cuartiles con valores menores a 0.480.

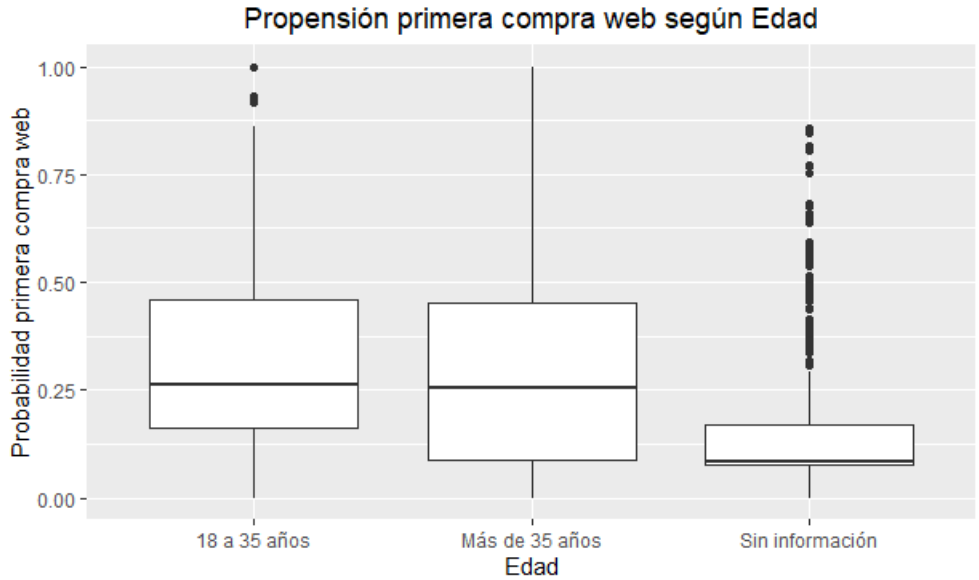

Gráfico 29: Propensión primera compra web según edad de clientes, modelo árbol C5.0. Fuente: Elaboración propia.

En el gráfico 30 se observan los valores de propensión a primera compra web de los clientes, clasificados según su cantidad de hijos. Mediante este gráfico se aprecia que los clientes para los cuales no se tiene información acerca de su cantidad de hijos poseen su mediana de propensión cercana a 0 y sus 3 primeros cuartiles tienen valores menores a 0.1. Aquellos clientes con más de 3 hijos, en cambio, tienen una mediana de propensión cercana a 0.190, y sus 3 primeros cuartiles tienen valores menores a 0.270. Los clientes con 2 a 3 hijos tienen una mediana de propensión aproximada de 0.190, mientras que sus 3 primeros cuartiles tienen valores menores a 0.375. Finalmente, los clientes con 0 o 1 hijo tienen una mediana de propensión cercana a 0.250, y sus 3 primeros cuartiles tienen valores menores a 0.490.

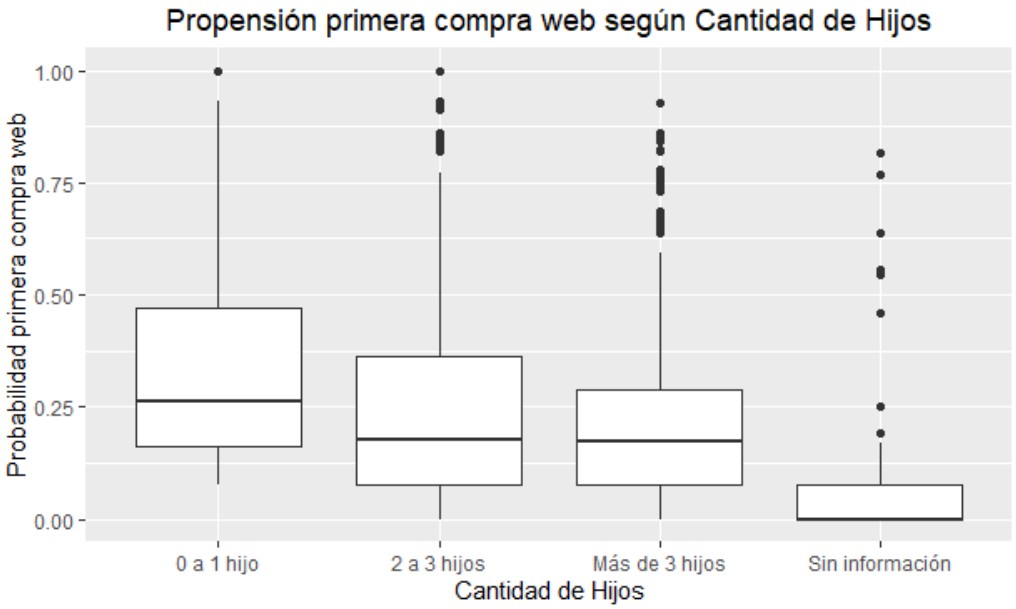

Gráfico 30: Propensión primera compra web según cantidad de hijos de clientes, modelo árbol C5.0. Fuente: Elaboración propia.

Cabe señalar que los gráficos y observaciones anteriores se hicieron tomando en cuenta los distintos niveles de variables categóricas obtenidas de los clientes, sin mantener las demás variables estáticas. De hecho, un cliente para el cual no se tiene información acerca de su segmento de valor de su tarjeta, generalmente tampoco se tiene información acerca de su renta, cupo, etc., mientras que un cliente para el cual se tiene información de su segmento de valor de su tarjeta, generalmente también se tiene información acerca de su renta, cupo, etc. Por lo tanto, el mayor aumento en la mediana de la propensión de los clientes a realizar su primera compra web tomando en cuenta sólo las variables categóricas, se logra al adquirir una tarjeta de crédito de la empresa, pues con ello se obtienen diversos datos sociodemográficos de los clientes que ayudan a pronosticar mejor su compra web.

3.7 Diseño experimental

Se ha seleccionado el email como canal único para el envío de las promociones que incentivan la primera compra online. Esto debido al costo de realizar envíos mediante SMS, y a la baja tasa de respuesta esperada al incluir promociones en las boletas de los clientes tras sus compras en tiendas (POS).

Las promociones sólo son válidas para comprar en la página web de la empresa, y corresponden a un descuento en las categorías de vestuario y calzado, para hombre, mujer y niños, de $10.000 al realizar compras totales por un monto superior a $20.000.

3.8 Hipótesis por testear

En el presente trabajo, con el fin de ganar conocimiento sobre el comportamiento de los clientes que no han comprado en el canal web de la empresa. se pretende testear las siguientes hipótesis:

1. **Promoción no afecta positivamente en la tasa de respuesta incremental en clientes con propensión alta de compra web.**

Se espera este comportamiento, ya que al ser alta la probabilidad de compra web, los clientes tratados y no tratados (grupo control), comprarán productos online independiente de la existencia de una promoción.

2. **Promoción afecta positivamente en la tasa de respuesta incremental en clientes con propensión media de compra web.**
   Como estos clientes podrían comprar o no comprar online (el modelo no los clasificó como propensión alta o baja), se espera que los clientes tratados reaccionen a la promoción enviada y opten por comprar online en mayor proporción que aquellos clientes no tratados (grupo control).

3. **Promoción no afecta positivamente en la tasa de respuesta incremental en clientes con propensión baja de compra web.**
   Es posible que parte importante de los clientes con propensión baja de compra web, no realice transacciones online debido a factores como: carecer de la tecnología o conocimiento necesario para realizar compra, poseer un bajo nivel socioeconómico, disgusto con modo de compra web, etc. Por lo que, aún cuando reciban una promoción, los clientes tratados no compren productos online en mayor proporción que los clientes no tratados (grupo control).

3.9 Variable experimental y tiempo de medición

La variable experimental que será usada en este trabajo es el tipo de promoción a enviar.

Tipo de promoción	Líneas
$10.000 al realizar compras mayores a $20,000	Vestuario y Calzado. Mujer, hombre y niños.

Tabla 11: Promoción experimento. Fuente: Elaboración propia.

El tiempo de medición para las promociones enviadas será el período comprendido entre el **día 9 de diciembre y el 16 de diciembre de 2019**.

3.10 Muestra total de clientes

Los clientes de la empresa que califican dentro del grupo objetivo de este trabajo (no han realizado compras vía web en más de 1 año, pero han realizado compras en tiendas físicas en los últimos 9 meses) son un total de 3.324.953 clientes.

De estos clientes, se descartaron a todos aquellos para los cuales no se poseía su email de contacto, junto con los que se dieron de baja de los correos de marketing de la empresa y los clientes pertenecientes a un grupo de control anual de la organización.

Así, no se consideró a los clientes presentes en las bases de datos:

- Grupo control anual.
- Lista negra SERNAC

Tras este descarte, se obtuvo una base de datos con 548,662 clientes totales.

3.11 Segmentación de clientes según propensión

En primer lugar, se usaron todas las variables transaccionales y sociodemográficas de los clientes que fueron descritas en este trabajo como predictores del modelo seleccionado (árbol de decisión C5.0). También se incluyeron 3 nuevas variables: cantidad de días navegados en el último mes, visitas con compra en tiendas acumuladas en el último mes y ventas acumuladas en el último mes. Esto ante la inquietud del área de Inteligencia de Clientes de saber el impacto de la información transaccional del último mes de cada cliente. Se realizó el preprocesamiento descrito en secciones anteriores y el balanceo y entrenamiento de la base de datos. Este entrenamiento fue realizado con los clientes que cumplían las condiciones de primera compra web, pero para el año anterior, pues el comportamiento debería ser similar a los clientes de este año, por estacionalidad. Por esta misma razón, y por la tendencia en el comportamiento de compra de los clientes es que se volvió a estimar el modelo.

Los resultados de las métricas en la partición de testing fueron las siguientes:

Modelo	AUC	Accuracy	Recall	Precision	F-Score
C5.0	0.7431	81.77%	49.53%	5.20%	0.0941
C5.0 umbral optimo	0.7431	65.88%	70.87%	3.88%	0.0735

Tabla 12: Resultados modelos estadísticos probabilidad primera compra web en diciembre 2018. Fuente: Elaboración propia.

Mediante la tabla 12, se observa que el modelo tiene sus métricas de rendimiento similares a las obtenidas anteriormente. Es decir, independiente de los clientes seleccionados y las ventanas temporales de los datos que se le otorga al modelo, éste predice de manera similar, al menos a simple vista.

Por otro lado, las variables más relevantes en función de su uso en las ramas del árbol de decisión son las siguientes:

Variable	Uso de atributo
Días navegados en los últimos 24 meses	100.00%
Días navegados en el último mes	100.00%
Visitas con compra web en días cyber de mayo, año anterior	100.00%
Cantidad de SKUs navegados de productos Electro, en los útimos 24 meses	100.00%
Cantidad de visitas con compra web de productos DecoHogar, últimos 24 meses	100.00%
Sexo: “Sin información”	100.00%
Visitas con compra en tiendas físicas, últimos 24 meses	98.06%
WoE de cupo crediticio	95.15%
WoE de monto de renta	92.66%
Segmento de valor tarjeta de crédito de empresa: Elite	87.41%

Tabla 13: Variables más relevantes de árbol de decisión. Fuente: Elaboración propia.

A partir de la tabla 13, se observa que la mayoría de las variables más relevantes, a la hora de clasificar a los clientes en “sí compra web” y “no compra web” son aquellas relacionadas con las compras web y la navegación del cliente.

Finalmente, mediante el modelo de árbol de decisión, se obtuvo la propensión a realizar la primera compra web de cada cliente.

Luego de ello, los clientes fueron clasificados en deciles según su propensión a realizar su primera compra en el canal web de la empresa. Dicha clasificación quedó de la siguiente manera:

Decil	Clientes totales	Propensión compra web
1	53493	[0 ; 0.642)
2	53990	[0.642 ; 0.725)
3	54427	[0.725 ; 0.770)
4	54870	[0.770 ; 0.801)
5	54587	[0.801 ; 0819)
6	55605	[0.819 ; 0.846)
7	54877	[0.846 ; 0.860)
8	55829	[0.860 ; 0.886)
9	55352	[0.886 ; 0.902)
10	55632	[0.902 ; 1]

Tabla 14: Cantidad de clientes por segmento y grupo experimental. Fuente: Elaboración propia.

Mediante la tabla 14, se observa que las propensiones de los clientes en general son altas (mayores a 0.5), esto puede ser debido a una tendencia positiva en el comportamiento de los clientes dentro del último año, que no fue capturada por el modelo estadístico, junto con el efecto del "estallido social" acontecido en Chile, donde varios locales físicos de la empresa fueron cerrados, provocando que muchas compras y cotizaciones de productos se realizaran en la web de la empresa, en vez de las tiendas físicas. De todos modos, se agrupó a los clientes usando varios deciles para la segmentación final, y de esta manera aislar la tendencia. Donde los tres deciles con menor propensión (deciles 1, 2 y 3) fueron segmentados como "propensión baja", los cuatro siguientes (deciles 4, 5, 6 y 7) como "propensión media", y los 3 últimos (deciles 8, 9 y 10) como "propensión alta.

3.12 Tamaño muestral

Para el cálculo del tamaño muestral se usó la ecuación presente en la sección **2.9 Cálculo de población de grupos de control y de tratamiento**. Los valores que se utilizaron como tasas de respuesta esperadas para estimar el tamaño muestral provienen del análisis de campañas realizadas anteriormente, específicamente de resultados de envíos efectuados en mayo y agosto del año 2019, que tenían el fin de lograr la omnicanalidad en el corto plazo de clientes que no realizaron compras web en más de un año. En concreto, el efecto observado y utilizado en la ecuación recién mencionada fue de un 0.04% (tasa de respuesta incremental).

Así, usando la ecuación anterior y, siguiendo criterios estadísticos y comerciales de la empresa, para lograr un nivel de confianza del 90% en tests de diferencia de proporciones de "una cola", se establecieron grupos de control (no tratados) con un tamaño igual al 33% de la población total de cada grupo de clientes segmentados en función de su propensión. Los tamaños de los grupos quedaron de la siguiente forma:

Segmento propensión	Cantidad clientes GC	Cantidad clientes GT	Cantidad clientes Total	Propensión compra web
Baja	52,972	108,839	161,811	[0 ; 0.77)
Media	72,564	147,375	219,939	[0.77 ; 0.86)
Alta	54,999	111,814	166,814	[0.86 ; 1]

Tabla 15: Cantidad de clientes por segmento y grupo experimental. Fuente: Elaboración propia.

3.13 Comprobación de balance de grupos experimentales

Para comprobar el balance en las características transaccionales y demográficas de los grupos de control y tratamiento, se realizaron tests de igualdad de medias y tests de igualdad de proporciones, para cada segmento de la población de clientes. Dichos resultados se observan en las tablas 16 y 17.

Segmento propensión	Variable	Media GC	Media GT	p-valor t-test
Alta	Ventas tienda, 12 meses	$280,034	$279,211	0.5814
	Visitas tienda, 12 meses	5.985	6.019	0.2248
	Boletas totales, 12 meses	7.092	7.144	0.1185
	Boletas web, entre 18 y 12 meses anteriores	0.360	0.360	1
Media	Ventas tienda, 12 meses	$190,312	$190,750	0.7093
	Visitas tienda, 12 meses	4.6	4.6	1
	Boletas totales, 12 meses	5.33	5.33	1
	Boletas web, entre 18 y 12 meses anteriores	0.04	0.04	1
Baja	Ventas tienda, 12 meses	$81,166	$80,942	0.7502
	Visitas tienda, 12 meses	2.47	2.47	1
	Boletas totales, 12 meses	2.73	2.73	1
	Boletas web, entre 18 y 12 meses anteriores	0	0	1

Tabla 16: Test de igualdad de medias, variables transaccionales. Fuente: Elaboración propia.

A partir de los resultados de la tabla 16, se observa que los grupos experimentales se encuentran balanceados para cada segmento de clientes, pues el p-valor de los t-tests de las variables transaccionales es mayor a 0.05, por lo que la diferencia entre las

medias de estas variables no es estadísticamente significativa con un nivel de confianza de 95%.

Segmento propensión	Variable	Proporción GC	Proporción GT	p-valor test z
Alta	Edad entre 18 y 45 años	0.6483	0.6511	0.1970
	Edad mayor a 45 años	0.3516	0.3488	0.1970
Media	Edad entre 18 y 45 años	0.5408	0.5377	0.1936
	Edad mayor a 45 años	0.4591	0.4622	0.1936
Baja	Edad entre 18 y 45 años	0.4468	0.4481	0.6744
	Edad mayor a 45 años	0.5531	0.5518	0.6744

Tabla 17: Test de igualdad de proporciones, variable edad. Fuente: Elaboración propia.

Mediante los resultados de la tabla 17, se aprecia que los grupos experimentales para cada segmento se encuentran balanceados entre sí, debido a que el p-valor de los tests z es mayor a 0.05, por lo que la diferencia en proporciones de la variable edad no es estadísticamente significativa con un nivel de confianza de 95%.

3.14 Resultados experimento

Ya realizada la campaña promocional, se procedió a analizar los resultados según las segmentaciones por propensión y por tratamiento.

Así, en cuestión de venta incremental en canal digital, se obtuvieron los siguientes resultados:

Grupo Tratamiento	Grupo Control	Venta Incremental	Venta Incremental por cliente
368,028	180,634	$-2,887,788	$-8

Tabla 18: Venta incremental en canal digital sin tomar en cuenta segmentación de propensión a primera compra web. Fuente: Elaboración propia.

Segmento	Grupo Tratamiento	Grupo Control	Venta Incremental	Venta Incremental por cliente
Baja	108,839	52,972	$-684,641	$-6
Media	147,375	72,663	$576,284	$4
Alta	111,814	54,999	$-2,739,142	$-24

Tabla 19: Ventas incrementales en canal digital por segmento de propensión a primera compra web. Fuente: Elaboración propia.

Mediante la tabla 18, se observa que, sin tomar en cuenta la segmentación por propensión, la venta incremental es negativa, es decir, los clientes tratados registraron compras con un valor total menor a los clientes que no fueron tratados. Esto puede ser debido al tipo de promoción. Los clientes tratados realizaron compras con valores mayores a $20,000 en total, luego, al aplicar la promoción, su boleta tenía un

descuento de $10,000. Mientras que los clientes no tratados realizaron compras sin el descuento, por lo que su boleta no disminuyó en valor. Además, los clientes tratados no realizaron las compras suficientes para superar la venta total de los clientes no tratados.

Sin embargo, al analizar los resultados por segmento (tabla 19), se tiene que, pese a que las ventas incrementales son negativas para los clientes del segmento bajo y alto, éstas son positivas para el segmento con propensión media. Lo anterior indica que las compras de los clientes tratados del segmento con propensión media fueron más que suficientes para superar las compras de los clientes no tratados del mismo segmento, pese al descuento aplicado a las boletas.

En cuestión de ticket promedio en el canal web de la empresa, los resultados obtenidos fueron los siguientes:

Ticket promedio GT	Ticket promedio GC
$50,423	$54,670

Tabla 20: Ticket promedio en canal digital sin tomar en cuenta segmentación de propensión a primera compra web. Fuente: Elaboración propia.

Segmento propensión	Ticket promedio GT	Ticket promedio GC
Baja	$50,717	$56,697
Media	$49,863	$53,719
Alta	$50,840	$54,559

Tabla 21: Ticket promedio en canal digital sin tomar en cuenta segmentación de propensión a primera compra web. Fuente: Elaboración propia.

A partir de la tabla 20, se observa que, en promedio, un cliente tratado realizó compras por un total de $50,423 en el período evaluado, en el canal web de la empresa. Cifra que es menor en comparación a los $54,670 gastados por los clientes no tratados. Esto es intuitivo, considerando los $10,000 de descuento que se entregó de promoción a los clientes tratados y que algunos de ellos usaron.

Al comparar los resultados usando la segmentación según la propensión de los clientes a realizar su primera compra web, se observa, mediante la tabla 21, un fenómeno similar. Para cada segmento, el ticket promedio de los clientes tratados fue menor al de los clientes no tratados.

Por otro lado, según la tasa de respuesta (porcentaje de clientes que realizó compras en el canal web), los resultados fueron los siguientes:

GT	GC	Tasa Respuesta GT	Tasa Respuesta GC	Tasa Respuesta Incremental	Nivel de confianza
368,028	180,634	0.75%	0.70%	0.05 pp.	**96.25%**

Tabla 22: Tasas de respuesta en canal digital, sin tomar en cuenta propensión a primera compra web. Fuente: Elaboración propia.

Mediante la tabla 22, se observa que, sin tomar en cuenta la propensión de los clientes a realizar su primera compra web, la tasa de respuesta en el canal digital fue de un 0.75% para los clientes tratados, en comparación al 0.70% de los clientes no tratados. Además, esta diferencia fue significativa, con un nivel de confianza de 96.25%. Es

decir, la promoción provoca que un mayor porcentaje de clientes realice su primera compra web en comparación al porcentaje de clientes que se "activa" bajo condiciones normales.

Los resultados en tasas de respuesta para los clientes segmentados según la propensión a realizar su primera compra web fueron usados para evaluar las hipótesis anteriormente presentadas.

3.15 Test de hipótesis

Las hipótesis planteadas en la sección **3.8 Hipótesis a testear**, serán evaluadas haciendo uso de tests de igualdad de proporciones con los resultados del envío promocional realizado. En particular, se analizará la diferencia en las tasas de respuesta de los clientes en el canal digital de la empresa, para clientes tratados y no tratados.

Se desea un nivel de confianza de 90% en tests de una cola para la evaluación de las hipótesis. Esto siguiendo los criterios que usa actualmente la empresa.

3.15.1 Hipótesis 1

Se evaluará lo siguiente: *"Promoción no afecta positivamente en la tasa de respuesta incremental en clientes con propensión alta de compra web".*

Para el segmento con propensión alta a realizar su primera compra web, los resultados fueron los de la tabla 23.

Segmento propensión	TR GT	TR GC	TR Incremental	Nivel de confianza
Alta	0.95%	0.93%	0.02 pp.	65.17%

Tabla 23: Tasas respuesta en canal digital, segmento propensión alta a realizar primera compra web. Fuente: Elaboración propia.

A partir de la tabla anterior, se observa que los clientes tratados del segmento con alta propensión tuvieron una tasa de respuesta de 0.95%, en comparación al 0.93% de los clientes no tratados. Sin embargo, esta diferencia no alcanza a ser significativa con un nivel de confianza de 90% en un test de diferencia de proporciones de "una cola". Razón por la cual no se puede rechazar la hipótesis planteada. Es decir, el impacto de la promoción en la tasa de respuesta es cercano a cero para el segmento de propensión alta.

3.15.2 Hipótesis 2

Se evaluará lo siguiente: *"Promoción afecta positivamente en la tasa de respuesta incremental en clientes con propensión media de compra web".*

Para el segmento con propensión media a realizar su primera compra web, los resultados fueron los de la tabla 24.

Segmento propensión	TR GT	TR GC	TR Incremental	Nivel de confianza
Media	0.75%	0.68%	0.06 pp.	**94.30%**

Tabla 24: Tasas de respuesta canal digital, segmento propensión media a realizar primera compra web. Fuente: Elaboración propia.

Mediante la tabla anterior, se observa que los clientes tratados del segmento medio de propensión tuvieron una tasa de respuesta en el canal web de la empresa de 0.75%, en comparación al 0.68% de los clientes no tratados del mismo segmento. Esta diferencia es significativa con un nivel de confianza de 94.30%, en un test de diferencia de proporciones de "una cola". Por lo que no se puede rechazar la hipótesis planteada. Es decir, la promoción provoca un aumento en la tasa de respuesta en el canal web de los clientes del segmento de propensión media.

3.15.3 Hipótesis 3

Se evaluará lo siguiente: *"Promoción no afecta positivamente en la tasa de respuesta incremental en clientes con propensión baja de compra web".*

Para el segmento con propensión baja a realizar su primera compra web, los resultados fueron los de la tabla 25.

Segmento propensión	TR GT	TR GC	TR Incremental	Nivel de confianza
Baja	0.54%	0.49%	0.05 pp.	88.49%

Tabla 25: Tasas de respuesta canal digital, segmento propensión baja a realizar primera compra web. Fuente: Elaboración propia.

A partir de la tabla 25, se aprecia que los clientes tratados del segmento bajo en propensión tuvieron una tasa de respuesta en el canal web de la empresa de 0.54%, en comparación al 0.49% de los clientes no tratados. Sin embargo, esta diferencia no alcanza a ser significativa con un nivel de confianza de 90% en un test de diferencia de proporciones de "una cola". Por lo cual no se puede rechazar la hipótesis planteada. Así, la promoción no provoca un aumento en la tasa de respuesta en el canal web de la empresa para los clientes del segmento de propensión baja.

3.16 Análisis de resultados experimento

Mediante los resultados obtenidos y recién descritos, se observa que existe un trade off producto de la promoción realizada, ya que, si bien provoca un aumento en la tasa de respuesta de los clientes en el canal web de la empresa a nivel general (sin tomar en cuenta la propensión de estos mismos), también se genera una venta incremental negativa, es decir, el incentivo no logra generar las compras necesarias de los clientes para que la empresa "recupere" el descuento aplicado.

Este trade off no ocurre al analizar los resultados de los clientes del segmento de propensión media a comprar en el canal web de la empresa. De hecho, es el único segmento con venta incremental positiva y con una tasa de respuesta incremental significativamente positiva.

Estos resultados sugieren seleccionar únicamente a los clientes clasificados como segmento de propensión media como los receptores de promociones similares, pues se esperarían ventas incrementales positivas y un aumento en la tasa de respuesta en el canal web.

Por otro lado, tomando en cuenta que el objetivo principal del área de Inteligencia de Clientes es que los clientes que no han comprado en el canal web de la empresa (o no lo han hecho en más de un año) vuelvan a hacerlo, podría decretarse la tasa de respuesta incremental como la única métrica importante, pues, a pesar de obtener ventas incrementales negativas o muy bajas, en el corto y mediano plazo esto se podría compensar mediante otras estrategias de marketing que mantengan a los clientes comprando en el canal web de la empresa y que generen mayores ventas.

4 Conclusiones

Los clientes de la empresa de retail en la cual se realizó este trabajo que no han comprado algún producto en el canal web de esta organización hace más de un año, pero que sí han realizado compras en sus tiendas físicas en los últimos 8 meses, llegan a ser más de 3 millones en total. Sin embargo, de éstos, aproximadamente el 1% realizaría su primera compra web el próximo mes.

Lo anterior provoca que sea muy difícil de pronosticar cuáles clientes serán los que realizarán su primera compra en el canal digital el mes siguiente, pues es una cantidad muy pequeña en relación con el total. En este trabajo, se testearon y evaluaron distintos modelos estadísticos con el fin de predecir quiénes serían estos clientes.

El modelo con mejor desempeño de los evaluados fue el árbol de decisión C5.0, pues usando un umbral óptimo de clasificación, su recall (68.56%), accuracy (69.02%) y AUC (0.7409) eran de los mejores, además de ser el más sencillo de replicar en otras plataformas computacionales.

Por otro lado, las variables más relevantes según el porcentaje de observaciones que fueron clasificadas en cada nodo del árbol corresponden a: las visitas con compra en los últimos meses en tiendas físicas y online, a la cantidad de días y productos navegados, al cupo en la tarjeta de crédito de la empresa, y a la cantidad de hijos de cada cliente. Estos predictores clasifican a más del 90% de las observaciones totales del dataset de entrenamiento, dentro del árbol de decisión. Por lo tanto, se desprende que el poder adquisitivo, la cantidad de compras y la navegación de cada cliente en la tienda, son muy importantes al momento de predecir si un cliente comprará en el canal digital de la empresa el próximo mes.

El uso del propensity score demostró ser una buena métrica bajo la cual segmentar a los clientes. Pues, mediante ésta se pudo clasificar a los clientes en propensión baja, media o alta. Y gracias a ello, observar cuáles clientes reaccionan mejor a una promoción determinada.

Al analizar la reacción de los clientes sin diferenciarlos por su propensión, la promoción enviada por email obtuvo una tasa de respuesta incremental positiva de 0.05 puntos porcentuales, siendo significativa con un nivel de confianza de 96.25%. Por lo tanto, el grupo de clientes que recibió la promoción tuvo un porcentaje significativamente mayor de clientes que compraron online que el grupo que no recibió la promoción.

Luego de analizar los resultados del envío tomando en cuenta los segmentos basados en el propensity score, se recomienda enviar promociones similares a clientes pertenecientes al segmento de propensión media, pues son los únicos que lograron ventas incrementales positivas ($576,284 en total) y una tasa de respuesta incremental significativamente positiva (0.06 puntos porcentuales, con un nivel confianza de 94.30%). Siendo esta última la métrica más importante de éxito para el área.

Hay que considerar que el envío promocional fue realizado en un período con muchas transacciones, como suele ser el mes de diciembre, pues se realizan muchas compras asociadas a los regalos de navidad y celebraciones de fin de año. Junto con ello, la

campaña se realizó apenas dos meses después del inicio del estallido social, el cual provocó una gran baja en las ventas de la industria del retail en general. Por lo anterior, las tasas de respuesta y las ventas obtenidas en la campaña promocional no corresponden a las de un período “normal” de la economía nacional (sin perturbaciones externas). Pese a esto, gracias a la comparación en las transacciones de un grupo de clientes tratados y otro grupo de clientes no tratados, se confirma la validez de los resultados obtenidos, pues ambos grupos de clientes experimentaron la misma inestabilidad en la economía.

5 Trabajo futuro

Posterior a la segmentación de clientes según su propensión a realizar la primera compra web el próximo mes, se sugiere caracterizar al “cliente promedio” de cada segmento. Así, se podrían idear promociones más acordes a las características de cada segmento de clientes, y de esta forma anticiparse a las necesidades de cada uno.

Para mejorar la predicción del modelo estadístico utilizado, se propone incorporar nuevas variables para cada cliente, tales como: su tasa de apertura de correos, su tasa de clicks dentro de correos recibidos, la frecuencia de compra del cliente, la cantidad de familiares que son clientes de la empresa y su tasa de canje de cupones promocionales.

Sobre las campañas promocionales, sería de gran utilidad el probar distintos tipos de descuentos, tanto en dinero como en porcentaje de la boleta, para cada tipo de cliente según su segmentación en propensión, en distintas líneas de productos y en distintos meses del año. De esta manera, se podría concluir cuáles promociones son más efectivas sobre cuáles tipos de clientes, disminuyendo los costos de la empresa en descuentos e incrementando la omnicanalidad de los clientes de distintos segmentos.

Junto con lo anterior, sería relevante para el área de la empresa, recoger la opinión de sus clientes mediante encuestas en línea o en tiendas físicas, para conocer de primera fuente las razones del por qué no han realizado sus compras por el canal web de la empresa. Estas razones podrían incorporarse como variables para el modelo predictivo o incluso se podrían idear soluciones para estos clientes que fueran más efectivas y/o a menor costo que el envío de promociones mediante marketing directo.

Por último, se propone realizar un seguimiento de transacciones a los clientes que realizaron su primera compra web gracias a la campaña promocional, para averiguar si los clientes mantuvieron la omnicanalidad en el mediano o largo plazo, o si sólo compraron en el canal digital debido al descuento ofrecido.

6 BIBLIOGRAFÍA

1. Alegría, C. 2019. Analistas adelantan una mala temporada de resultados para el retail. [en línea] El Mercurio. 18 de febrero, 2019. <https://www.elmercurio.com/Inversiones/Noticias/Analisis/2019/02/15/Continuaran-los-malos-resultados-del-sector-retail.aspx> [consulta: 23 julio 2019]

2. Ayala, M. 2019. Finanzas: Falabella tambalea en retail: reducen sus ventas en 2018. [en línea] América Retail. 4 de marzo, 2019. <https://www.america-retail.com/finanzas/finanzas-falabella-tambalea-en-retail-reducen-sus-ventas-en-2018/> [consulta: 23 julio 2019]

3. Bate, A., Lindquist, M., & Edwards, I. R. (2008). The application of knowledge discovery in databases to post-marketing drug safety: example of the WHO database. Fundamental & Clinical Pharmacology, 22(2), 127–140.

4. Bonera, M. (2011). The propensity of e-commerce usage: the influencing variables. Management Research Review, 34(7), 821-837.

5. Boselli R., Cesarini M., Mercorio F., Mezzanzanica M. (2017) Using Machine Learning for Labour Market Intelligence. In: Altun Y. et al. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2017. Lecture Notes in Computer Science, vol 10536. Springer, Cham.

6. Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7), 1145–1159.

7. Brown, M., Pope, N., & Voges, K. (2003). Buying or browsing? An exploration of shopping orientations and online purchase intention. European Journal of Marketing, 37(11/12), 1666-1684.

8. Cardoza, L. (2018). Evaluación de promociones en la retención de un segmento de clientes de una tienda por departamento. Memoria para optar al título de Ingeniero Civil Industrial, Universidad de Chile, Departamento de Ingeniería Industrial.

9. Cárdenas, L. 2019. Retail a la baja, suben las liquidaciones: Ventas caen a su menor nivel en cinco años y empleos disminuyen. [en línea] La Tercera. 4 de junio, 2019. <https://www.latercera.com/la-tercera-pm/noticia/retail-la-baja-suben-las-liquidaciones-ventas-caen-menor-nivel-cinco-anos-empleos-disminuyen/684221/#> [consulta: 23 julio 2019]

10. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD ’16.

11. Falabella Retail S.A. 2019. Nuestra empresa. [en línea] Falabella. <https://www.falabella.com/falabella-cl/category/cat40006/Nuestra-empresa> [consulta: 23 julio 2019]

12. Memoria Anual Empresa 2018. Santiago, Chile.

13. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-37.

14. Ildefonso, G., & Elena, A. (2009). Fundamentos y Técnicas de Investigación Comercial.

15. Indian Agricultural Statistics Research Institute. 2015. Evaluation Measures for Data Mining Tasks. Winter School on “Data Mining Techniques and Tools for Knowledge Discovery in Agricultural Datasets”, 145-152.

16. Liebermann, Y., & Stashevsky, S. (2009). Determinants of online shopping: Examination of an early-stage online market. Canadian Journal of Administrative Sciences/Revue Canadienne des Sciences de l'Administration, 26(4), 316-331.

17. Mackinnon, M. J., & Glick, N. (1999). Applications: Data Mining and Knowledge Discovery in Databases–An Overview. Australian & New Zealand Journal of Statistics, 41(3), 255-275.

18. Mlambo, N. (2016). Classification-Based Data Mining in Target Marketing. International Journal of Advanced Research in Computed Science and Software Engineering. 6(3):29-31.

19. Oprea, C. (2014). Performance evaluation of the data mining classification methods. Information society and sustainable development, 2344, 249-253.

20. Pang, S., & Gong, J. (2009). C5.0 Classification algorithm and application on individual credit evaluation of Banks. Systems Engineering – Theory & Practice, 29(12): 94-104.

21. Riquelme, F. (2017). Evaluación de la efectividad de promociones personalizadas en la retención y aumento de los clientes de alto valor de una tienda por departamento. Memoria para optar al título de Ingeniera Civil Industrial, Universidad de Chile, Departamento de Ingeniería Industrial.

22. Rosenbaum, P., & Rubin, D. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika, 70(1), 41-55.

23. Rubin, D., & Waterman, R. (2006). Estimating the Causal Effects of Marketing Interventions Using Propensity Score Methodology. Statistical Science, 21(2), 206-222.

24. Soopramanien, D. G., & Robertson, A. (2007). Adoption and usage of online shopping: An empirical analysis of the characteristics of “buyers”“browsers” and “non-internet shoppers”. Journal of Retailing and Consumer Services, 14(1), 73-82.

7 Anexos

Variable	Descripción
Visitas tienda 24 meses	Visitas con compra tienda física, acumulado 24 meses anteriores
Visitas tienda 9 meses	Visitas con compra tienda física, acumulado 9 meses anteriores
Visitas tienda 6 meses	Visitas con compra tienda física, acumulado 6 meses anteriores
Visitas tienda 3 meses	Visitas con compra tienda física, acumulado 3 meses anteriores
Visitas web 24 meses	Visitas con compra online, acumulado 24 meses anteriores
Visitas web 21 meses	Visitas con compra online, acumulado 21 meses anteriores
Visitas web 18 meses	Visitas con compra online, acumulado 18 meses anteriores
Visitas web 15 meses	Visitas con compra online, acumulado 15 meses anteriores
Visitas web cyber nov 17	Visitas con compra online, días cyber noviembre 2017
Visitas web cyber mayo 18	Visitas con compra online, días cyber mayo 2018
Visitas web cyber oct 18	Visitas con compra online, días cyber octubre 2018
Ventas tienda 24 meses	Ventas tienda física, acumulado 24 meses anteriores
Ventas tienda 12 meses	Ventas tienda física, acumulado 12 meses anteriores
Ventas tienda 9 meses	Ventas tienda física, acumulado 9 meses anteriores
Ventas tienda 6 meses	Ventas tienda física, acumulado 6 meses anteriores
Ventas tienda 3 meses	Ventas tienda física, acumulado 3 meses anteriores
Ventas web 24 meses	Ventas web, acumulado 24 meses anteriores
Ventas web 21 meses	Ventas web, acumulado 21 meses anteriores
Ventas web 18 meses	Ventas web, acumulado 18 meses anteriores
Ventas web 15 meses	Ventas web, acumulado 15 meses anteriores
Ventas web cyber nov 17	Ventas web, días cyber noviembre 2017
Ventas web cyber mayo 18	Ventas web, días cyber mayo 2018
Ventas web cyber oct 18	Ventas web, días cyber octubre 2018
Visitas tienda Vestuario	Visitas con compra tienda física, acumulado 24 meses anteriores, en Vestuario
Visitas tienda Decohogar	Visitas con compra tienda física, acumulado 24 meses anteriores, en Decohogar
Visitas tienda Electro	Visitas con compra tienda física, acumulado 24 meses anteriores, en Electro
Visitas web Vestuario	Visitas con compra web, acumulado 24 meses anteriores, en Vestuario
Visitas web Decohogar	Visitas con compra web, acumulado 24 meses anteriores, en Decohogar
Visitas web Electro	Visitas con compra web, acumulado 24 meses anteriores, en Electro
Ventas tienda Vestuario	Ventas tienda física, acumulado 24 meses anteriores, en Vestuario
Ventas tienda Decohogar	Ventas tienda física, acumulado 24 meses anteriores, en Decohogar
Ventas tienda Electro	Ventas tienda física, acumulado 24 meses anteriores, en Electro
Ventas web Vestuario	Ventas web, acumulado 24 meses anteriores, en Vestuario
Ventas web Decohogar	Ventas web, acumulado 24 meses anteriores, en Decohogar
Ventas web Electro	Ventas web, acumulado 24 meses anteriores, en Electro
Productos distintos	Cantidad de SKUs distintos comprados en 24 meses anteriores
Navegación año anterior	Días navegados, año anterior
Navegación año actual	Días navegados, año actual
Navegación Vestuario año anterior	Cantidad de SKUs distintos navegados año anterior, en Vestuario

Navegación Electro año anterior	Cantidad de SKUs distintos navegados año anterior, en Electro
Navegación Decohogar año anterior	Cantidad de SKUs distintos navegados año anterior, en Decohogar
Navegación Vestuario año actual	Cantidad de SKUs distintos navegados año actual, en Vestuario
Navegación Electro año actual	Cantidad de SKUs distintos navegados año actual, en Electro
Navegación Decohogar año actual	Cantidad de SKUs distintos navegados año actual, en Decohogar
Recency tienda	Recencia de cliente en tienda física
Recency web	Recencia de cliente en compras web
Recency navegación	Recencia de cliente en navegación web
Edad	Edad de cliente
Cupo	Cupo en crédito de cliente
Monto Renta	Monto de renta de cliente
Cantidad Hijos	Cantidad de hijos de cliente

Tabla 26: Variables numéricas calculadas. Fuente: Elaboración propia.

N°	eta	Máx profundidad	Ratio submuestral de columnas	Submuestra	Rounds	ROC	Sens	Spec
1	**0.3**	1	0.6	0.5	50	0.871	0.813	0.758
2	**0.3**	1	0.6	0.5	100	0.878	0.838	0.763
3	**0.3**	1	0.6	0.5	150	0.880	0.839	0.759
4	**0.3**	1	0.6	0.75	50	0.873	0.811	0.774
5	**0.3**	1	0.6	0.75	100	0.880	0.836	0.759
6	**0.3**	1	0.6	0.75	150	0.882	0.843	0.760
7	**0.3**	1	0.6	1	50	0.872	0.814	0.768
8	**0.3**	1	0.6	1	100	0.880	0.831	0.762
9	**0.3**	1	0.6	1	150	0.882	0.842	0.760
10	**0.3**	1	0.8	0.5	50	0.875	0.821	0.764
11	**0.3**	1	0.8	0.5	100	0.880	0.844	0.759
12	**0.3**	1	0.8	0.5	150	0.883	0.850	0.756
13	**0.3**	1	0.8	0.75	50	0.873	0.815	0.765
14	**0.3**	1	0.8	0.75	100	0.880	0.834	0.762
15	**0.3**	1	0.8	0.75	150	0.882	0.843	0.758
16	**0.3**	1	0.8	1	50	0.873	0.816	0.766
17	**0.3**	1	0.8	1	100	0.880	0.834	0.762
18	**0.3**	1	0.8	1	150	0.883	0.843	0.763
19	**0.3**	2	0.6	0.5	50	0.887	0.849	0.758
20	**0.3**	2	0.6	0.5	100	0.884	0.851	0.756
21	**0.3**	2	0.6	0.5	150	0.881	0.842	0.756
22	**0.3**	2	0.6	0.75	50	0.885	0.848	0.761
23	**0.3**	2	0.6	0.75	100	0.885	0.846	0.753
24	**0.3**	2	0.6	0.75	150	0.883	0.842	0.753
25	**0.3**	2	0.6	1	50	0.889	0.859	0.757
26	**0.3**	2	0.6	1	100	0.889	0.866	0.751
27	**0.3**	2	0.6	1	150	0.887	0.861	0.750
28	**0.3**	2	0.8	0.5	50	0.884	0.851	0.750
29	**0.3**	2	0.8	0.5	100	0.882	0.842	0.758
30	**0.3**	2	0.8	0.5	150	0.877	0.843	0.760
31	**0.3**	2	0.8	0.75	50	0.888	0.859	0.759
32	**0.3**	2	0.8	0.75	100	0.887	0.856	0.753
33	**0.3**	2	0.8	0.75	150	0.884	0.855	0.761
34	**0.3**	2	0.8	1	50	0.889	0.857	0.760
35	**0.3**	2	0.8	1	100	0.889	0.857	0.763
36	**0.3**	2	0.8	1	150	0.887	0.853	0.755
37	**0.3**	3	0.6	0.5	50	0.881	0.839	0.766
38	**0.3**	3	0.6	0.5	100	0.881	0.836	0.770
39	**0.3**	3	0.6	0.5	150	0.875	0.830	0.768
40	**0.3**	3	0.6	0.75	50	0.888	0.855	0.762
41	**0.3**	3	0.6	0.75	100	0.884	0.848	0.761
42	**0.3**	3	0.6	0.75	150	0.881	0.845	0.767
43	**0.3**	3	0.6	1	50	0.889	0.852	0.757

44	**0.3**	3	0.6	1	100	0.887	0.850	0.761
45	**0.3**	3	0.6	1	150	0.885	0.845	0.767
46	**0.3**	3	0.8	0.5	50	0.884	0.845	0.763
47	**0.3**	3	0.8	0.5	100	0.880	0.834	0.761
48	**0.3**	3	0.8	0.5	150	0.874	0.824	0.761
49	**0.3**	3	0.8	0.75	50	0.887	0.842	0.769
50	**0.3**	3	0.8	0.75	100	0.882	0.840	0.765
51	**0.3**	3	0.8	0.75	150	0.879	0.839	0.764
52	**0.3**	3	0.8	1	50	0.890	0.854	0.760
53	**0.3**	3	0.8	1	100	0.886	0.854	0.764
54	**0.3**	3	0.8	1	150	0.883	0.842	0.767
55	**0.4**	1	0.6	0.5	50	0.876	0.822	0.766
56	**0.4**	1	0.6	0.5	100	0.881	0.842	0.753
57	**0.4**	1	0.6	0.5	150	0.882	0.842	0.753
58	**0.4**	1	0.6	0.75	50	0.877	0.825	0.772
59	**0.4**	1	0.6	0.75	100	0.881	0.848	0.760
60	**0.4**	1	0.6	0.75	150	0.883	0.85	0.755
61	**0.4**	1	0.6	1	50	0.876	0.823	0.767
62	**0.4**	1	0.6	1	100	0.883	0.840	0.763
63	**0.4**	1	0.6	1	150	0.883	0.847	0.756
64	**0.4**	1	0.8	0.5	50	0.875	0.831	0.760
65	**0.4**	1	0.8	0.5	100	0.879	0.846	0.749
66	**0.4**	1	0.8	0.5	150	0.881	0.851	0.748
67	**0.4**	1	0.8	0.75	50	0.876	0.821	0.763
68	**0.4**	1	0.8	0.75	100	0.882	0.846	0.759
69	**0.4**	1	0.8	0.75	150	0.883	0.844	0.759
70	**0.4**	1	0.8	1	50	0.877	0.825	0.763
71	**0.4**	1	0.8	1	100	0.882	0.837	0.762
72	**0.4**	1	0.8	1	150	0.884	0.845	0.759
73	**0.4**	2	0.6	0.5	50	0.888	0.851	0.757
74	**0.4**	2	0.6	0.5	100	0.881	0.837	0.756
75	**0.4**	2	0.6	0.5	150	0.879	0.839	0.756
76	**0.4**	2	0.6	0.75	50	0.886	0.850	0.760
77	**0.4**	2	0.6	0.75	100	0.882	0.842	0.758
78	**0.4**	2	0.6	0.75	150	0.880	0.844	0.754
79	**0.4**	2	0.6	1	50	0.888	0.855	0.757
80	**0.4**	2	0.6	1	100	0.886	0.855	0.755
81	**0.4**	2	0.6	1	150	0.884	0.847	0.758
82	**0.4**	2	0.8	0.5	50	0.879	0.842	0.753
83	**0.4**	2	0.8	0.5	100	0.877	0.837	0.771
84	**0.4**	2	0.8	0.5	150	0.872	0.833	0.769
85	**0.4**	2	0.8	0.75	50	0.884	0.845	0.759
86	**0.4**	2	0.8	0.75	100	0.879	0.845	0.749
87	**0.4**	2	0.8	0.75	150	0.874	0.834	0.758
88	**0.4**	2	0.8	1	50	0.889	0.854	0.763

89	**0.4**	2	0.8	1	100	0.885	0.848	0.757
90	**0.4**	2	0.8	1	150	0.881	0.847	0.755
91	**0.4**	3	0.6	0.5	50	0.876	0.837	0.755
92	**0.4**	3	0.6	0.5	100	0.870	0.826	0.766
93	**0.4**	3	0.6	0.5	150	0.865	0.816	0.763
94	**0.4**	3	0.6	0.75	50	0.880	0.844	0.755
95	**0.4**	3	0.6	0.75	100	0.878	0.832	0.762
96	**0.4**	3	0.6	0.75	150	0.871	0.829	0.769
97	**0.4**	3	0.6	1	50	0.885	0.842	0.760
98	**0.4**	3	0.6	1	100	0.881	0.840	0.766
99	**0.4**	3	0.6	1	150	0.875	0.839	0.769
101	**0.4**	3	0.8	0.5	50	0.872	0.829	0.760
102	**0.4**	3	0.8	0.5	100	0.867	0.813	0.754
103	**0.4**	3	0.8	0.5	150	0.864	0.814	0.755
104	**0.4**	3	0.8	0.75	50	0.885	0.857	0.763
105	**0.4**	3	0.8	0.75	100	0.878	0.847	0.766
106	**0.4**	3	0.8	0.75	150	0.875	0.829	0.768
107	**0.4**	3	0.8	1	50	0.886	0.846	0.764
108	**0.4**	3	0.8	1	100	0.879	0.842	0.762
109	**0.4**	3	0.8	1	150	0.877	0.837	0.767

Tabla 27: Configuración de parámetros Extreme Gradient Boosting Árbol. Fuente: Elaboración propia.

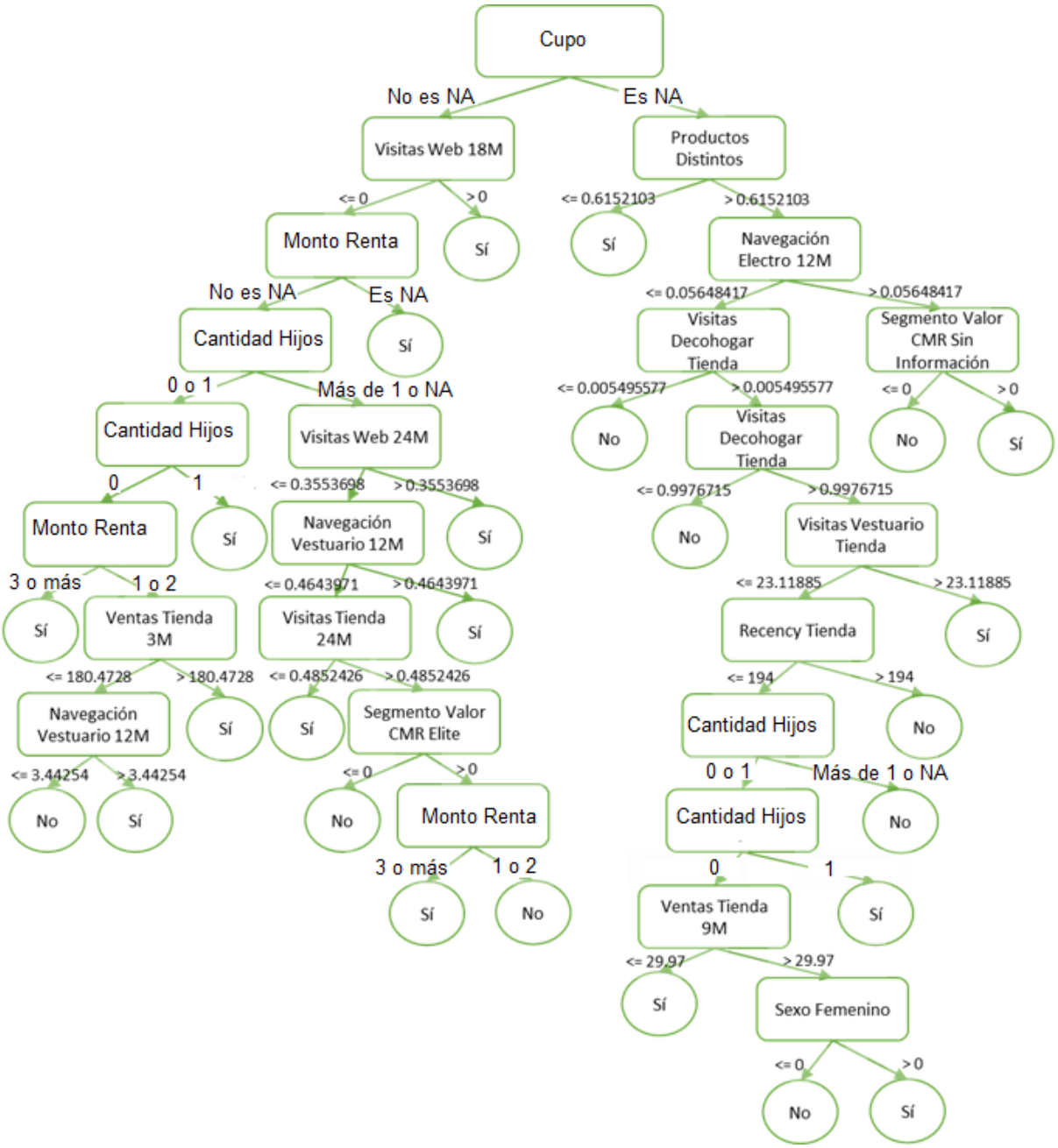

Gráfico 31: Árbol de decisión C5.0 primera compra web en agosto 2019. Fuente: Elaboración propia.