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

ANÁLISIS PREDICTIVO DE SATISFACCIÓN EN TELECOMUNICACIONES Y SUS IMPACTOS EN FUGA, CAMBIOS DE EQUIPO Y CONVERSION DE CAMPAÑAS.

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

TOMÁS PABLO LEYTON CARRASCO

PROFESORA GUÍA:
ALEJANDRA PUENTE CHANDÍA

MIEMBROS DE LA COMISIÓN:
PABLO MARÍN VICUÑA
DANIELA CARREÑO ROJAS

SANTIAGO DE CHILE
2020

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE: INGENIERO CIVIL INDUSTRIAL
POR: **TOMÁS PABLO LEYTON CARRASCO**
FECHA: 2020
PROF. GUÍA: ALEJANDRA PUENTE CHANDÍA

ANÁLISIS PREDICTIVO DE SATISFACCIÓN EN TELECOMUNICACIONES Y SUS IMPACTOS EN FUGA, CAMBIOS DE EQUIPO Y CONVERSION DE CAMPAÑAS.

Una de las características del mercado de las telecomunicaciones es ser altamente competitivo y con una clara tendencia a mantenerse así en el futuro. La firma para la que se realiza el trabajo busca conseguir ventajas competitivas a través de un mejor análisis de la satisfacción de sus clientes. El objetivo del presente informe es el de identificar la existencia de relaciones entre la satisfacción de un cliente y distintos ámbitos comerciales, a través de un modelo que prediga la satisfacción del cliente con la firma. Los ámbitos estudiados corresponden a la fuga de clientes, la tasa de conversión a campañas de la firma y el share de cambio de equipo por canales de la firma. Para la consecución de este objetivo se aplican distintas técnicas de minería de datos y machine learning, a través de la metodología CRISP-DM. Se utiliza información principalmente de interacciones entre el cliente y la firma, complementando con información socio-demográfica y de mercado.

A partir de regresiones logísticas, se observa que los clientes que interaccionan por los canales del call center, USSD e IVR tienen mayores probabilidades de insatisfacción, como también ocurre para los clientes que han tenido quejas con la firma o problemas de facturación. Luego se prueban distintos modelos de caja negra, donde los mejores resultados de predicción de satisfacción los obtuvo el algoritmo random forest, obteniendo un lift máximo cercano a 4 y un accuracy de 0,68 en el set de testeo. Dentro de los clientes clasificados como insatisfechos por el modelo, se obtiene una mejora de 17 puntos porcentuales por sobre un clasificador aleatorio.

Se estudian las relaciones entre la predicción de satisfacción y los efectos mencionados para 3 meses diferentes. Se confirma la existencia de relaciones entre la satisfacción y fuga de clientes, donde los clientes más insatisfechos tienen en promedio una tasa de fuga hasta 3 veces más alta que la tasa base. Se encuentra también una relación entre la satisfacción y el share de cambio de equipo a través de la firma, donde los clientes más satisfechos tienden a cambiar más sus equipos a través de los canales de la firma, superando en el mejor mes por 6 puntos porcentuales a la tasa de cambio base. No se encuentra relación evidente entre la satisfacción y la tasa de conversión de campañas donde se ofrece una línea adicional, donde parecen ser otros los factores que determinan la conversión de este tipo de campañas.

Los resultados muestran que los modelos presentados tienen potencial para obtener mejores resultados, a través de la inclusión de más datos que den cuenta de las diferencias en el servicio ofrecido según la zona geográfica, como cobertura, señal, tiendas disponibles, competencia presente, etc. Como trabajo futuro se propone incorporar esta información a los modelos desarrollados.

A mi familia, con la que comencé a caminar
y a la familia que he encontrado caminando.

Saludos

Agradecimientos

El mayor agradecimiento de este trabajo va para mi papá, mamá y mi hermana, quienes son parte indispensable en mi vida y siempre están ahí para apoyarme en todo, este trabajo es para y por ustedes.

Agradezco también a todos mis amiges por brindarme momentos de paz, alegría y risas para poder sobrellevar los momentos difíciles y recargarme de energías cada vez que lo necesito.

Muchas gracias a la gente con la que compartí en los meses de tesis, quienes me brindaron ayuda cuando la necesitaba y de quienes prendí mucho.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

El mercado de las telecomunicaciones se caracteriza por ser altamente competitivo y con una clara tendencia a mantenerse así en el futuro. Las diferentes compañías ponen gran parte de sus esfuerzos en aumentar la retención de sus clientes actuales, ya que resulta menos costoso para estas el conseguir que un cliente actual se quede con la firma una mayor cantidad de tiempo que el conseguir un cliente nuevo [20].

Dentro de este escenario, la firma busca conseguir una ventaja competitiva que le permita gestionar a sus clientes de una manera más efectiva en comparación a sus principales competidores. Una de las formas en que busca conseguir este objetivo es a través de un mejor entendimiento y manejo de la satisfacción de sus clientes.

La forma en que usualmente la firma ha trabajado la satisfacción de sus clientes se ha relegado a estudios centrados solo en clientes que son encuestados por la firma, a partir de lo que se obtienen distintas reglas de negocio relacionadas con satisfacción.

Este trabajo busca contribuir a cambiar el paradigma reinante en este ámbito de trabajo, en donde se realizan principalmente modelos centrados en su interpretabilidad, complementando lo anterior con el desarrollo de modelos centrados principalmente en la predicción realizada [4] (pero perdiendo en parte su interpretabilidad).

Para conseguir esto, el presente trabajo busca construir modelos que permitan predecir la satisfacción de un cliente a partir de encuestas de satisfacción realizadas a un conjunto de clientes, pero que luego puedan ser utilizados para obtener una predicción de satisfacción para todos los clientes de la firma. Con esto se busca mejorar la gestión de clientes y entregar nueva información que permita mejorar los criterios que utiliza la firma hoy en día para seleccionar a los clientes que forman parte de sus diversas estrategias comerciales.

Por ejemplo, se sabe de diversos estudios que la satisfacción tiene una fuerte relación con la fuga de un cliente [24, 13], es más, en estos estudios se llega a la conclusión de que la satisfacción es mejor predictor de la fuga que la información meramente económica (como por ejemplo gastos en distintos servicios como planes o bolsas).

A pesar de existir numerosos resultados de este estilo, la firma no ha intentado la elaboración de modelos que intenten predecir la satisfacción de sus clientes, ver los beneficios de

utilizar esta predicción en la población completa y de esta forma entender mejor que clientes son más propensos a fugarse debido a su satisfacción con la firma que por otros factores

El estudio busca no solo entender como una predicción de satisfacción ayudaría a la firma en cuanto a fuga, si no que también se buscara entender si existen relaciones entre la satisfacción y los cambios de equipo que realizan los clientes o con la tasa de conversión a campañas.

Para conseguir esta predicción, el trabajo hará uso de diversos métodos de minería de datos y machine learning, buscando complementar las respuestas de los encuestados con información que se tenga disponible para todos los clientes y que de esta forma permitan identificar clientes como satisfechos o insatisfechos con la firma.

Capítulo 2

Identificación del problema y Objetivos

2.1. Sector industrial

Una de las cualidades de la industria de las telecomunicaciones es su constante transformación a través del tiempo, partiendo en Chile por centrarse en lograr conectar las distintas zonas a lo largo del país a través del teléfono fijo, a competir por entregar el mejor servicio posible tanto en voz como internet móvil y hogar. Desde eso punto en adelante la industria no ha dejado de crecer y para diciembre de 2019 Chile alcanza ya una penetración aproximada de 130 abonados por cada 100 habitantes, con un total 25,1 millones de abonados a la fecha [9].

Según la SUBTEL[1] los ingresos brutos de la industria de las Telecomunicaciones y Servicios TI alcanzaron los $6.004 millones de pesos durante los últimos 12 meses (8,8 billones USD). Esto significó un leve crecimiento de 0,75 % en relación con el año 2018, impulsado por Banda Ancha, Servicios TI y Telefonía Móvil.

La demanda por servicios de datos y voz móvil es creciente, con incrementos en los abonados a telefonía móvil en un 5,9 % y un incremento de las conexiones móviles a internet de un 18,7 % entre septiembre de los años 2017 y 2018, crecimiento que se mantiene para el 2019 [6]. Lo opuesto ocurre en telefonía fija, donde las líneas en servicio exhibieron una disminución interanual de 6,4 %, con una tendencia decreciente de varios años.

[1] Subsecretaría de telecomunicaciones.

Participación de mercado	Dic 18	Dic 19
Movistar	28,1%	25,4%
ENTEL	31,1%	30,6%
Claro	24,0%	23,0%
Virgin	0,9%	0,8%
WOM	14,9%	19,0%
VTR	1,0%	1,2%
Otros	<0,1%	<0,1%

Figura 2.1: Participación de mercado principales empresas. Fuente: Informe anual SUBTEL 2019.

Existen varias compañías que ofrecen sus servicios dentro de la industria Chilena, siendo Movistar, Claro, Wom y Entel las principales compañías dentro del país. Los tres principales operadores (Entel, Movistar y Claro) poseen el 79,0 % del mercado a diciembre 2019 (ver figura 2.1). Los otros operadores en su conjunto alcanzan el 21 % del mercado, destacando la evolución de Wom con un 27,2 % de crecimiento en los últimos 12 meses.

Durante los últimos años, el mercado de clientes móviles ha presentado una migración constante de los clientes prepago (compra de bolsas) a clientes pospago (contratación de planes), tendencia que se espera continúe en el tiempo y que se puede ver a partir de los datos mostrados en las figuras 2.2 y 2.3.

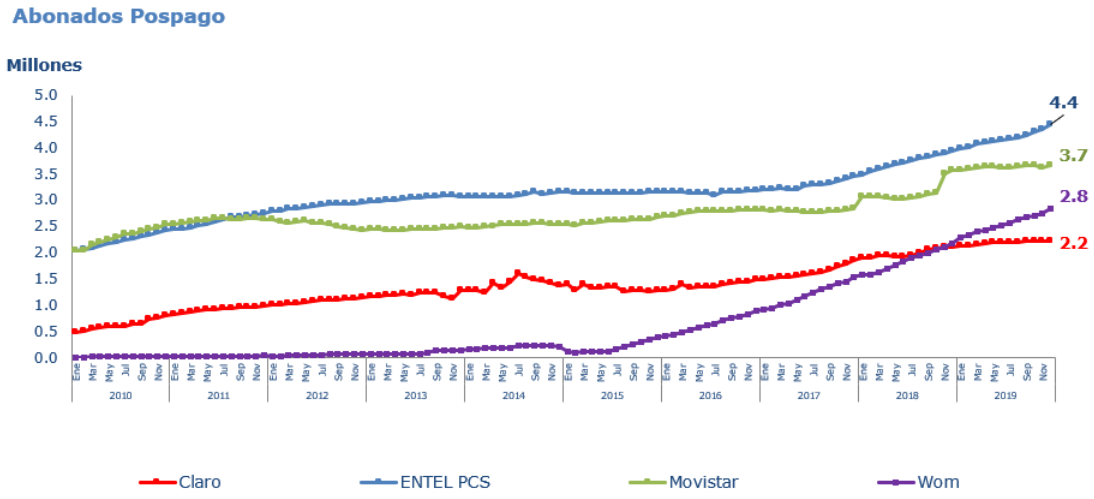

Figura 2.2: Abonados pospago por empresa. Fuente: Informe anual SUBTEL 2019.

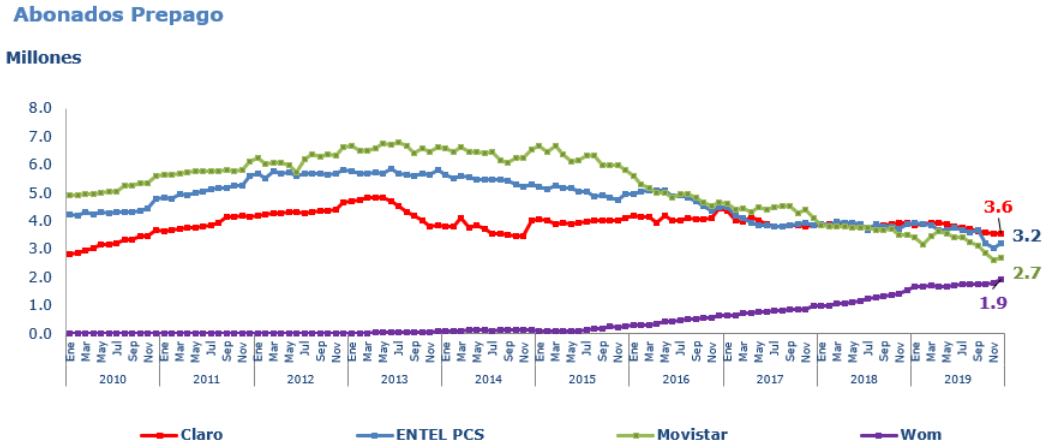

Figura 2.3: Abonados prepago por empresa. Fuente: Informe anual SUBTEL 2019.

Entel mantiene la mayor participación de mercado en los clientes pospago (33 %), donde luego le siguen Movistar (27,2 %), Wom (21,1 %) y Claro (16,5 %). Dentro de prepago las empresas que han presentado un mayor descenso en su universo de clientes corresponden a Entel y Movistar (figura 2.3). Si bien pareciera que Wom se escapa del comportamiento general de la industria, esto ocurre principalmente al ser una empresa nueva en comparación a las demás, por lo cual en estos momentos obtiene parte del mercado prepago que aún no considera el cambio a pospago como opción.

En cuanto a internet móvil, los accesos 4G alcanzaron los 16,51 millones de conexiones

a diciembre 2019, con un crecimiento de 13,4 % en los últimos 12 meses. La tecnología 4G está sustituyendo a los accesos 3G, los cuales decrecieron un 30,9 % en el mismo periodo. Con todo, el crecimiento neto del mercado de internet móvil fue de 852.177 en los últimos 12 meses.

A nivel mundial, desde finales 2018, comenzó la introducción de la tecnología del 5G. Esta nueva tecnología será introducida rápidamente, y se espera que la venta de los dispositivos móviles capaces de soportar el 5G sea muy veloz, producto de la caída de los precios de estos equipos. De esta manera, más usuarios podrán optar por esta nueva tecnología.

Los principales beneficios del 5G, en comparación con el 4G, son la mayor velocidad, mayor inmediatez por la casi nula latencia y la entrega de más valor en contenido y datos. Sumado a a esto, se podrá innovar, mejorar y masificar el streaming de datos y dar mejor soporte a diversas aplicaciones de tecnologías con realidad virtual.

Ante la inminente llegada de la red 5G al país, junto con la tendencia del mercado a migrar a la contratación de planes y la alta competitividad existente debido a la gran cantidad de firmas presentes en el mercado Chileno, es que se vuelve prioridad para estas firmas el poder identificar ventajas competitivas que les ayuden a servir este mercado de la mejor manera posible y estar preparados para servir a los clientes la red 5G de la mejor manera posible.

Uno de los métodos en que la firma busca obtener ventajas competitivas es a través de la identificación del grado de satisfacción que los clientes tienen con la empresa y a partir de esto entender como impacta la satisfacción del cliente en la fuga, conversiones a campañas y la tasa de cambio de equipos móviles a través de canales de la firma.

2.2. Estado de la satisfacción en la empresa

La satisfacción es un aspecto muy importante para varios mercados, incluido el de telecomunicaciones. Se considera que un cliente es realmente leal con una compañía cuando se encuentra satisfecha con sus servicios [13].

En un mercado competitivo un cliente que posee una alta satisfacción con el producto y/o servicio entregado por la firma, se mantendrá con esta una mayor cantidad de tiempo que alguien ubicado en el espectro contrario [12], aumentando así el valor del cliente para la firma. En cambio, un cliente insatisfecho genera un impacto negativo tanto en la rentabilidad esperada de esta persona, como en la reputación de la firma, pudiendo incluso afectar el comportamiento de otros clientes de la empresa.

Para entender y conocer la satisfacción de sus clientes, la firma realiza mensualmente encuestas de satisfacción y a la vez contrata a una consultora para que realice de forma independiente encuestas de satisfacción (las cuales serán descritas en mayor profundidad en la sección 5.1).

Dentro de estas encuestas se pregunta sobre la satisfacción por diferentes servicios, pero la atención de este trabajo estará centrada en la primera pregunta realizada por ambas

encuestas, la cual corresponde a la nota con la que el cliente evalúa su satisfacción general con la firma.

Ambas encuestas utilizan la misma escala de evaluación, la cual corresponde a una nota de 1 a 7, lo cual permite evaluarlas en conjunto. La figura 2.4 muestra el promedio de satisfacción obtenido en los meses estudiados por la firma.

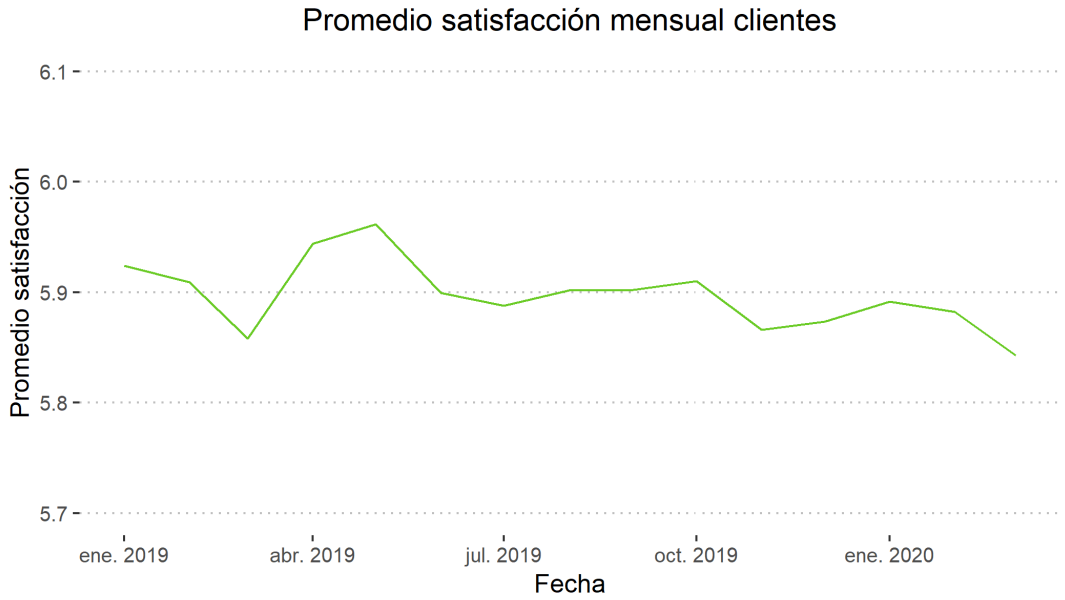

Figura 2.4: Promedio nota satisfacción general con la compañía.

Como se puede ver, la satisfacción promedio dentro del periodo estudiado se mantiene siempre cercana al valor 5,9, lo cual muestra que en promedio el desempeño global de la firma en este ámbito se ha mantenido relativamente constante.

Los resultados siguieren que dentro de este periodo cualquier esfuerzo por realizar cambios en la satisfacción de los clientes no se ha traducido en un aumento en la satisfacción promedio de clientes que logre llegar al 6 o sobrepasarlo. Como se puede ver, los últimos meses no se ha podido no se ha podido volver a superar una evaluación promedio de 5,9, presentando su mínimo valor en marzo.

Dentro del mercado la principal diferencia entre los clientes viene dada por su medio de pago del servicio, pudiendo ser prepago, donde el cliente paga de forma previa por los servicios de voz y datos, o clientes pospago, donde el usuario tiene un contrato o plan con la firma y así a cambio de un pago mensual, el cliente recibe un servicio constante todos los meses.

Para indagar si existen diferencias entre estos dos tipos de clientes, se grafica en la figura 2.5 la satisfacción promedio de los clientes separando por el tipo de pago que realizan.

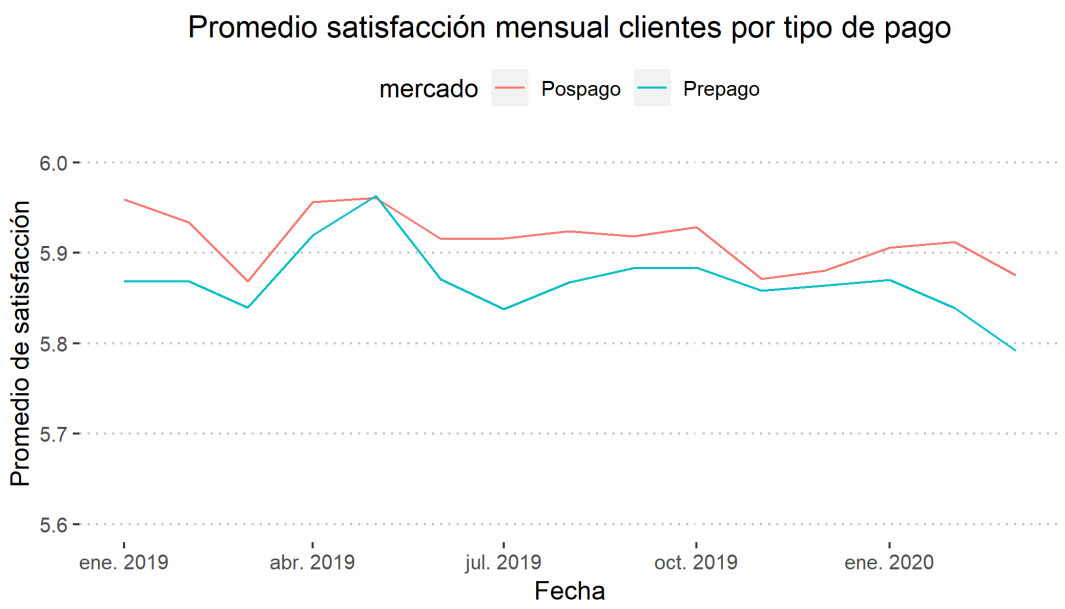

Figura 2.5: Promedio satisfacción mensual según medio de pago.

Al separar por este criterio, se puede ver que la satisfacción de los clientes pospago es consistentemente mayor o igual a la satisfacción dentro de los clientes prepago. Esto sugiere la existencia de servicios que influyen de maneras diferentes en la satisfacción de cada segmento, aunque la magnitud de estas diferencias no genera un desvío significativo entre los promedios obtenidos por los dos tipos de clientes.

Además de estas diferencias, existen otras relacionadas a la forma en que se gestiona a los clientes de cada segmento, diferencias en la disponibilidad de información y calidad de los datos que se tienen. Estas diferencias producen que sea necesario trabajar ambos mercados de forma separada.

Debido a estas diferencias, donde la disponibilidad de información en prepago es menor a la que se tiene en pospago, el presente trabajo pone su atención en predecir la satisfacción para los clientes pospago.

Como el universo de clientes prepago sigue siendo una parte importante para la firma, se decide que es necesario incluir una sección posterior al desarrollo del trabajo que determine la factibilidad de replicar el trabajo realizado dentro del mercado de prepago. En esta sección también se discutirá la necesidad o no de incluir información adicional que solo es relevante dentro de este prepago.

Volviendo a la satisfacción, si bien los estudios realizados sobre encuestas permiten encontrar insight interesantes sobre los motivos que provocan satisfacción o insatisfacción en los clientes, los resultados de los gráficos, donde no se ven cambios notorios en la satisfacción a través del tiempo, sugieren que es necesario poder identificar de antemano a los clientes satisfechos de los insatisfechos de una manera más específica para enfocar cualquier insight de satisfacción a los clientes que realmente están insatisfechos (o satisfechos) con la firma.

Como ya se mencionó, trabajos relacionados a la satisfacción ya concluyen que esta influye

de manera significativa en la fuga de clientes [13], y a la vez nos dicen que un cliente satisfecho tiene efectos positivos para la firma, como que el cliente se vuelva leal a esta.

Sin embargo, estos estudios no nos permiten entender el grado en que ocurren estos factores a nivel de toda la población. No se sabe si los clientes insatisfechos son 2 o 3 o más veces propensos a fugarse. Tampoco se conocen todas las áreas en las que un cliente más satisfecho que otro le trae beneficios a la firma al volverse leal a esta.

Para poder entender mejor estos factores, la empresa se ha propuesto el desarrollar una medida que le informe de la satisfacción que tiene cada cliente con la firma, clasificándolos como satisfecho o no con su servicio y saber de que formas esto afecta los esfuerzos comerciales de la firma.

Los estudios nos dicen que las decisiones de los clientes se basan en dos factores principales, los factores económicos, como los costos de cambio de servicio, y los factores de satisfacción, donde las decisiones del cliente se basan en la evaluación del desempeño del servicio ofrecido por la firma con el.

Un modelo de satisfacción se presenta como una oportunidad de diferenciar entre estos dos tipos de factores, entre clientes que basan sus decisiones en motivos relacionados solo a factores económicos o clientes que ponen mayor peso a la satisfacción que tienen con el servicio y su relación con la firma.

2.3. Objetivo general

El objetivo general corresponde a **identificar la existencia de relaciones entre la satisfacción de un cliente y distintos ámbitos comerciales, a través de un modelo que prediga la satisfacción del cliente con la firma**.

En específico, los ámbitos comerciales estudiados correspondes a: fuga de clientes, share de cambio de equipos a través de canales de la empresa y la tasa de conversión de clientes con campañas de la empresa.

2.4. Objetivos específicos

Para conseguir el objetivo general recién descrito, se presentan a continuación los objetivos específicos que permitirán acercarse a los resultados finales.

1. Identificar posibles sesgos y errores presentes en las encuestas, limpiar los datos con errores y mejorar la representatividad de las encuestas elaborando pesos en caso de estimarse necesario.

2. Construir y/o incorporar diferentes variables que permitan diferenciar de mejor manera a los clientes según criterios socio-demográficos, de mercado e interacciones realizadas.

3. Decidir criterio de corte a utilizar para seleccionar a un encuestado como satisfecho o insatisfecho con los servicios de la empresa, para utilizar como variable dependiente en modelos posteriores.

4. Entrenar diferentes modelos que predigan la satisfacción de un cliente y seleccionar el modelo con el mejor desempeño para esta tarea. Con esto, predecir la satisfacción del universo de clientes pospago para diferentes meses.

5. Identificar la existencia o no de relaciones entre la satisfacción de los clientes pospago con la fuga, share de cambio de equipo por canales de la firma y tasa de conversión a campañas de marketing.

6. Discutir sobre la factibilidad de replicar el trabajo realizado dentro del mercado de clientes prepago. Dar a conocer que cosas se puede implementar de forma directa y que cambios deben ser considerados para trabajar dentro de este mercado.

2.5. Alcances

Como ya se menciono, este trabajo estará centrado en los clientes **pospago** de la firma, por lo que al hablar de los clientes en las secciones posteriores, se dará por entendido que se habla de este segmento en particular en caso de no haber sido especificado.

Los principales motivos para esto tienen que ver con diferencias existentes entre ambos mercados que justifican el trabajarlos de forma separada, y diferencias en la cantidad y calidad de información disponible para cada segmento de clientes.

Hacia el final del trabajo, dado que el segmento de clientes prepago corresponde a un porcentaje importante de clientes de la firma, se incluye una sección donde se abarca como replicar el trabajo realizado en pospago dentro del mercado de prepago.

Esta sección incluye una discusión junto con un modelo con fines descriptivos para dar a entender mejor las diferencias existentes entre los dos tipos de clientes y como esto significa que para prepago se debe considerar información adicional relevante solo para este mercado.

Es importante destacar que como se quiere predecir la satisfacción para la totalidad de clientes, a la hora de seleccionar las variables para lograr este objetivo, no se podrá utilizar información que solo se tenga para los clientes encuestados (como utilizar las respuestas entregadas a otras preguntas dentro de la encuesta), ya que esta información no estaría disponible para el resto de clientes.

El trabajo debe ser reproducible en el tiempo, por lo que se deja afuera cualquier tipo de información o variables que puedan impedir la replicación del trabajo a futuro por parte de la empresa.

Debido a que la cantidad de variables disponible en la firma es muy alto, el énfasis de este trabajo estará en la elaboración de atributos a partir de información proveniente de diversas interacciones realizadas entre el cliente y la firma, en la forma de 3 bases, visitas a tiendas,

uso de la aplicación para celulares de la firma y un resumen mensual de diversas interacciones realizadas por los clientes en diferentes canales.

A estos datos se agregará también información socio demográfica y de mercado, cambios de equipos y señal recibida por los clientes. A diferencia de las 3 primeras tablas mencionadas, estos datos son utilizados de forma directa, sin mayores cambios previos más que limpieza y verificación de que sean fidedignos.

Las encuestas utilizadas para el presente trabajo fueron realizadas entre enero del 2019 hasta comienzos de marzo del 2020, por lo que los análisis realizados solo toman en cuenta la información contenida dentro de este periodo. En particular el trabajo se enmarca fuera de los efectos de pandemia que llegaron a Chile a mediados de marzo del 2020.

Capítulo 3

Marco conceptual

3.1. Customer relationship management

EL trabajo a realizar esta enfocada en el área de marketing, centrándose en la interacción que existe entre el cliente y la empresa. Customer relationship management (CRM) es un enfoque para gestionar esta interacción, buscado utilizar los datos históricos que los clientes tienen con la empresa para poder mejorar las relaciones comerciales futuras, buscando así impulsar el crecimiento de ventas.

Segun IBM, CRM busca mejorar el rendimiento de los productos, mejorar el servicio al consumidor, perfeccionar el valor entregado al cliente y la satisfacción de este, establecer relaciones de confianza mutua a largo plazo, atraer nuevos clientes y mantener a los clientes antiguos dentro de la empresa[25].

3.2. Machine learning

Corresponde a una disciplina científica del ámbito de la inteligencia artificial cuyo objetivo es programar computadores para optimizar un criterio de desempeño utilizando datos de ejemplo o información de experiencias pasadas [2]. Aprender en este contexto hace referencia a la ejecución de un algoritmo que optimiza los parámetros del modelo a partir de los datos que le son entregados.

Los modelos pueden ser predictivos, buscando identificar comportamientos futuros, o bien descriptivos, donde se espera adquirir nuevos conocimientos a partir de los datos, o buscar ambos objetivos a la vez.

Estos algoritmos se dividen principalmente en dos tipos: supervisados y no supervisados. El primero hace referencia a modelos en donde se busca predecir cierto atributo de los datos a partir de información para la cual se conoce este atributo, por ejemplo, elaborar un modelo que identifica si un correo es spam o no, utilizando para esto información histórica de mensajes para los cuales ya se conoce que mensajes corresponden a spam y cuales no.

Por otro lado, en el aprendizaje no supervisado solo se tiene información de entrada

sin una variable dependiente y el objetivo es encontrar regularidades en la información que permitan agrupar los datos [2]. Una de las áreas donde se utiliza este tipo de modelos es en la segmentación de clientes, donde se busca encontrar agrupaciones que diferencien a los clientes de una firma y de esta forma poder realizar diferentes estrategias, como por ejemplo, ofrecer distintos productos y servicios para cada agrupación de clientes que entrega el modelo.

El presente trabajo cae dentro de la categoría de modelo supervisado, ya que el objetivo es predecir la satisfacción de los clientes de la empresa a partir de clientes encuestados sobre este tema, por lo que se conoce su nivel de satisfacción con la firma.

Existen hoy en día diversos algoritmos que pueden ser utilizados para tareas de aprendizaje supervisado, de entre los cuales se utilizan los siguientes dentro del trabajo: regresiones logísticas y algoritmos basados en árboles de decisión, distributed random forest (DRF), gradient boosting machine (GBM) y extreme gradient boosting (XGBoost).

La elección de estos modelos viene dada a partir del potencial que poseen estos algoritmos de aprendizaje para el problema de fuga de clientes en telecomunicaciones y otros mercados [15, 21, 23], por lo que se espera que también sean adecuados para la tarea de predicción de satisfacción.

3.2.1. Regresión logística

Este tipo de regresión es de gran utilidad para problemas en los que la variable dependiente toma valores pertenecientes a un conjunto finito y/o corresponde a una variable cualitativa.

Supongamos que la variable dependiente Y corresponde a la ocurrencia o no de un suceso, tomando un valor de 1 si el suceso ocurre y 0 si no. Nuestro interés esta en estudiar la relación entre una o más variables independientes o explicativas $x_1, x_2, ..., x_i$ y la variable Y. Un modelo logístico establece la siguiente relación entre la probabilidad de que ocurra el suceso Y para un individuo dadas sus características $x_1, x_2, ..., x_i$ [14].

$$Pr(Y = 1|x_1, x_2, ..., x_i) = \frac{1}{1 + exp(-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - ... - \beta_i x_i)}$$

Esta relación normalmente se trabaja como una regresión lineal al aplicarle la transformación logit:

$$logit(Pr(Y = 1|x)) = ln\left(\frac{Pr(Y = 1|x)}{1 - Pr(Y = 1|x)}\right) = -\beta_0 - \beta_1 x_1 - \beta_2 x_2 - ... - \beta_i x_i$$

Esto permite trabajar la regresión logística como si fuera una regresión lineal y obtener muchas de las propiedades deseables de esta.

3.2.2. Árboles de decisión

Este es un algoritmo que divide los datos de forma recursiva utilizando declaraciones de tipo *if-else* según un criterio de partición estipulado, con el objetivo de predecir la clase a la que pertenecen los datos de los grupos resultantes.

A modo de ejemplo, supongamos que una nueva especie es descubierta por científicos [17]. ¿Cómo podemos decidir si es un mamífero o no? Una forma de resolver esto es realizar una serie de preguntas sobre las características de la especie, comenzando por si tiene sangre caliente o sangre fría. Si tiene sangre fría, entonces definitivamente no es un mamífero.

En caso de ser de sangre caliente, debemos seguir realizando preguntas como por ejemplo, ¿el género femenino de esta especie da a luz a sus nuevas generaciones o no?. Esta serie de preguntas y sus posibles respuestas pueden ser ordenadas como un árbol de decisión, donde el objetivo de cada pregunta es permitir una mejor clasificación de los grupos resultantes.

El criterio que suele utilizarse para decidir el mejor punto de partición de los datos es el de clasificación errónea o tasa de error de los nodos resultantes, existiendo también otros criterios como la entropía o el coeficiente de Gini.

$$Error\ de\ clasificación = 1 - max_i[p_i(t)]$$

Donde $p_i(t)$ corresponde a la frecuencia relativa de instancias que pertenecen a la clase i en el nodo t.

3.2.3. Distributed random forest

La idea base detrás del algoritmo random forest es la de utilizar varios árboles de decisión en conjunto para así armar un modelo con mayor robustez que el obtenido por cada uno de estos árboles por separado.

Para crear cada árbol, se selecciona un subconjunto al azar de observaciones del total de la muestra y luego se selecciona al azar también un subconjunto de los predictores disponibles para entrenar, esto para evitar que todos los árboles sean idénticos y así obtener beneficios de entrenar más de un árbol.

La decisión de la clase a la que pertenece una observación se realiza agregando los distintos resultados, ya sea mediante asignarle la clase mayoritaria u obteniendo el promedio de los resultados, en caso de ser predicción de algún valor numérico en el último caso.

Distributed hace referencia a una implementación específica del algoritmo random forest en el que todos los árboles son entrenados de forma paralela en vez de secuencial, por lo que se diferencia solo en términos de rapidez computacional a una versión del algoritmo que no es distribuida.

3.2.4. Gradient boosting machine y Extreme gradient boosting

Al igual que ranfom forest, GBM obtiene sus resultados a través del entrenamiento de varios modelos por separado, pudiendo ser estos árboles de decisión u otro modelo. GBM involucra 3 elementos principales:

- Una función de pérdida a ser optimizada, la cual dependiendo del tipo de problema puede ser mínimos cuadrados o la pérdida logarítmica, entre otros [11].
- Un modelo base para hacer predicciones que, como ya se mencionó, suele utilizarse árboles de regresión (en esta ocasión se usan estos como base).
- Combinar los modelos base en un modelo aditivo que busque optimizar la función de pérdida. GBM entrena los modelos de forma secuencial, donde cada nuevo modelo es entrenado a partir de los errores de clasificación del modelo anterior (el primer modelo se entrena sobre los errores de una primera predicción aleatoria basada en la distribución observada de las clases a predecir).

Existen diversas variaciones del algoritmo GBM, como lo es el caso del algoritmo Extreme Gradient Boosting (XGBoost), el cual se diferencia por implementar una versión regularizada de la función de pérdida, lo que consigue agregando costos adicionales por incorporar nuevas variables al modelo. De esta forma el algoritmo intenta que la mayor cantidad de coeficientes sean cero y así reducir el costo de la función quedándose solo con las variables esenciales.

El beneficio principal de esta modificación es la reducción esperada del sobre ajuste que pueda tener el modelo. XGBoost adicionalmente incorpora varias mejoras en cuanto a rapidez y eficiencia en el uso de recursos computacionales.

3.2.5. Criterio de información de Akaike

Una de las tareas que se deben realizar luego de elaborar varias regresiones que buscan explicar un suceso, es el de elegir cual de estos modelos consigue explicar de mejor manera la variable independiente. Para esto existen varios criterios dentro de los cuales se encuentra el criterio de información de Akaike o AIC[1].

AIC corresponde a una medida que busca realizar un trade off entre la bondad de ajuste de un modelo (lo bien que este se ajusta a la variabilidad de los datos) y la complejidad de este en término de cantidad de variables explicativas utilizadas. Este valor es calculado según la siguiente función de pérdida.

$$AIC = -2\ln(L) + 2k$$

Donde k corresponde a la cantidad de parámetros y L a la máxima verosimilitud del modelo. El problema se reduce a encontrar el modelo que minimiza el valor de esta función,

[1] Acrónimo de la palabra en inglés Akaike's Information Criterion.

el cual corresponderá al modelo preferido según este criterio. Como el objetivo es minimizar el valor del AIC, una cantidad mayor de variables significa una penalización mayor al valor de la ecuación.

Este criterio sirve como ayuda para discernir entre un modelo y otro, pero no es capaz de decir nada sobre la calidad de un modelo por si sólo, por lo que si los modelos estudiados presentan fallas este criterio no es capaz de dar aviso de ello de ninguna forma.

3.3. Métricas de evaluación de modelos

Además del uso del criterio AIC para regresiones logísticas, es necesario elegir las métricas de evaluación para poder decidir que modelo es mejor que otro entre los distintos algoritmos a utilizar. Dentro de las principales métricas utilizadas se encuentran el Lift, Precision, Recall y Accuracy.

Estas 4 métricas se construyen a partir de la matriz de confusión asociada a la predicción realizada por los modelos de clasificación y lo que realmente sucedió en el intervalo predicho (ver figura 3.1).

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos positivos (VP)	Falsos negativos (FN)
	Negativos	Falsos positivos (FP)	Verdaderos negativos (VN)

Figura 3.1: Matriz de confusión de un modelo de clasificación. Fuente: Elaboración propia.

- **accuracy:** Corresponde al porcentaje de los datos clasificados de manera correcta. Esta métrica es útil pues es la más simple de entender.

$$Accuracy = \frac{VP + VN}{total}$$

- **Precision:** Este valor corresponde al porcentaje de datos clasificados correctamente como positivos sobre el total de datos asignados como positivos por el modelo. A la hora de aplicar recomendaciones en base a los resultados de un modelo, este valor sirve para tener una estimación de la cantidad de gente que resultaría afectada por la recomendación sin necesitarla realmente.

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

- **Recall:** Compete a la proporción entre los datos clasificados correctamente como positivos y la totalidad de datos que eran realmente positivos. si este valor es muy bajo quiere decir que una gran parte de los casos que nos importan están siendo mal clasificados, por lo que este dato entrega información valiosa a la hora de elegir entre modelos.

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

- **Lift:** El lift corresponde a un ratio entre probabilidades aplicado a un conjunto de los datos donde se espera observar una probabilidad de pertenecer a la clase de interés más alta que el normal de la población total. Así, por ejemplo, si se quiere realizar una estrategia de marketing de envió de correo donde la tasa de apertura del mail de la población es de un 5 %, se trata de identificar grupos donde la tasa de respuesta esperada sea mayor que este valor. Continuando el ejemplo, si se identifica un grupo especial con una tasa de apertura del mail de un 20 %, este grupo tendrá un lift de $\frac{20}{5} = 4$.

$$Lift = \frac{P(A \cap B)}{P(A) * P(B)} = \frac{Tasa\ dentro\ de\ grupo}{Tasa\ gobal}$$

Existen casos en los cuales la métrica de accuracy resulta no ser una buena métrica para evaluar modelos. Esto ocurre cuando los datos por naturaleza están desbalanceados, por lo que existen categorías en las cuales están la mayor parte de los datos mientras otras categorías tienen un porcentaje menor.

Enfrentados a este tipo de problemas, algunos modelos tienden a identificar a la mayoría de los casos como perteneciente a la clase que se repite más en los datos. Esto normalmente se traduce en un accuracy bastante alto, pero que falla completamente en identificar la clase que nos interesa diferenciar, la cual corresponde normalmente a la que se presenta en menor medida.

Como se vera más adelante, esto es exactamente lo que ocurre dentro de las encuestas de satisfacción, donde los casos de gente que declara estar satisfecha con el servicio superan en gran medida a los casos que declaran insatisfacción, lo que dificulta el poder identificarlos.

Debido esto, en el presente trabajo el accuracy no será el principal valor para determinar que modelo es mejor que otro, si no que se basará principalmente en el lift, recall y precision.

3.4. Teoría de encuestas

A día de hoy las encuestas deben ser una de las técnicas más utilizadas para obtener conocimientos de la población [1], la utilizan desde empresas a universidades y gobiernos,

con objetivos que pueden ser académicos, comerciales, descriptivos, etc. Su alto grado de uso actual produce que sean vistas como algo sencillo de realizar, por lo que en muchos lugares son realizadas sin conocimientos previos sobre la teoría que hay detrás de una encuesta.

En su gran mayoría, la encuesta a consiste esencialmente en efectuar individualmente una serie de preguntas a un grupo de personas que han sido previamente seleccionadas de modo que constituyen una muestra representativa de la sociedad[2]. La idea básica es que mediante este procedimiento es posible cuantificar y descubrir determinadas características presentes en la población.

Las encuestas pueden variar en muchos aspectos, según el tipo de muestreo utilizado (por ejemplo probabilístico o estratificado), el medio o canal por el cual se realiza, el tipo de medición que implementa, la duración de la encuesta, la complejidad de las preguntas, etc.

Dentro de todos los factores mencionados, se debe tener el cuidado de no cometer errores o sesgos, los cuales pueden invalidar los resultados de una encuesta. A continuación se explican algunos de los principales sesgos que puede presentar una encuesta.

3.4.1. Sesgo en las encuestas

El sesgo en las encuestas corresponde a errores que ocurren en la preparación o realización de esta misma, provocando como consecuencia que las respuestas se inclinen en un determinado sentido, lo cual se puede traducir mas tarde en conclusiones equivocas a partir de sus resultados.

Existe una cantidad muy amplia de sesgos que pueden desviar los resultados, estos pueden ocurrir al momento de seleccionar a los encuestados, deberse a problemas con la encuesta, problemas con el entrevistador, ocurrir debido a una alta tasa de gente que no responde, etc.

Hay trabajos que han llegado lo bastante lejos como para identificar hasta 48 tipos diferentes de sesgos que pueden estar presentes dentro de una encuesta [7]. Se describen a continuación los principales tipos de sesgo que pueden estar presentes dentro de una encuesta [1].

- **Sesgo de selección de encuestados:** Ocurre cuando ciertos grupos de personas no son considerados para responder la encuesta, lo que invisibiliza las preferencias de ese grupo.
- **Sesgo de no respuesta:** Se genera a partir de las personas que reciben la oferta de ser encuestados pero deciden no realizarla. El mismo hecho de no contestar la encuesta puede estar diciendo que estas personas pertenecen a un grupo diferente que los que si quieren contestarla. Este sesgo puede no presentar efectos en los resultados de las encuestas si es que la representatividad de quienes si contestan la encuesta sigue siendo similar a la población.

[2] Estrictamente, también pueden haber encuestas donde no hay selección previa o donde se busca representar un segmento específico de la población, pero estas no corresponden al tipo de encuestas revisadas dentro de este trabajo, por lo que se prefiere mantener esta definición.

- **Sesgo en las respuestas del encuestado:** Producido por efectos conscientes y subconscientes de la persona entrevistada. Esto puede deberse a diversos motivos como no tener claro el objetivo de la encuesta, la existencia de sesgo de consentimiento, donde el encuestado quiere generar una buena impresión al entrevistador, falsedad de las respuesta entregadas, influencias culturales, etc.
- **Sesgo en las preguntas:** Este tipo de sesgo puede ocurrir a partir de preguntas muy complejas, una mala estructura de la encuesta o aparecer debido a la secuencia en que se presentan las preguntas (las primeras preguntas pueden influir en las respuestas de las siguientes preguntas), etc.
- **Sesgos por el entrevistador:** Los entrevistadores pueden ser una fuente que introduzca errores en la obtención de la información, por ejemplo, al modificar las preguntas, acortándolas para terminar luego la entrevista, o al cometer errores en la transcripción de las respuestas, entre muchos otros posibles casos.

Estos son solo algunos de los sesgos que se pueden presentar en una encuesta. Se analizará cuales de ellos pueden estar presentes en las encuestas utilizadas para el trabajo y para cuales de ellas se tiene la opción de corregir el sesgo y para cuales de estas no (por requerir acciones fuera del alcance del trabajo o por que solo son mejoras aplicables a futuras encuestas).

No es tarea del presente trabajo decidir si las encuestas que realiza la firma están hechas de una buena manera o no, lo que se busca entender es cuales de los sesgos mencionados están presentes dentro de las encuestas a utilizar y corregir estos casos cuando sea posible a través del manejo de los datos.

3.4.2. Diseño de pesos para encuestas

Los pesos o ponderaciones a encuestas corresponden a un valor que se asigna a cada observación (encuesta) y cumplen el objetivo de modificar cuanto más o cuanto menos de una observación se considera en los procedimientos estadísticos donde se utilicen. Lo que se busca con esto es aumentar la representatividad de la encuesta.

Por ejemplo, un peso de 2 significa que la observación es considerada como 2 casos idénticos y un peso de 1 significa que la observación es considerada como un solo caso. Si los encuestados superan a la población en cierta característica 5 veces a 1, entonces el peso que se le asigna a la observación es de $\frac{1}{5}$ [3].

Los 2 tipos más comunes de pesos son los pesos de diseño y los pesos de no respuesta o post estratificación. El primero asigna pesos para compensar por decisiones tomadas de manera previa, como sobre representar ciertos segmentos minoritarios de la población para poder obtener respuestas de ese segmento.

[3] Puede ser el caso también que se busque representar solo a partes específicas de la población, por lo que no sería necesario aplicar pesos en estos casos. No se profundiza en estos casos al no ser este el caso para las encuestas utilizadas en este trabajo.

$$peso\ de\ dise\tilde{n}o = \frac{1}{factor\ de\ muestreo}$$

El segundo tipo de pesos busca compensar por las diferencias obtenidas entre una encuesta y la población debido a las diferentes tasas de respuesta que el respectivo segmento pueda tener. Algunas de las características que tienden a generar diferencias son el sexo, edad, educación, lugar de residencia, etc. Para poder realizar los pesos para este tipo de casos es necesario conocer además las distribuciones reales que presenta la población de estudio dentro de estas características.

$$peso\ de\ no\ respuesta = \frac{proporción\ en\ la\ población}{proporción\ en\ la\ muestra}$$

Se preferirá siempre aplicar la segunda formula pues esta da cuenta también de las diferencias por no respuesta, pero en casos donde no se tenga las distribuciones originales de la característica en estudio la primera formula puede ser una opción.

Para los análisis tan solo se utiliza un solo peso, por lo cual si se quiere compensar más de un factor de la población estos deben ser combinados en un solo valor. Para conseguir esto existen varios métodos y el que se utiliza en el presente trabajo corresponde al siguiente proceso iterativo[16].

Considerando que se tienen 3 características A, B y C sobre los que se quiere elaborar pesos:

1. Se elabora el peso para A (W_A) y se pondera la información por este valor. Con esto se genera la tabla de frecuencias resultante para la característica B.

2. Se elabora el peso para B (W_B) y se pondera la información por $W_A * W_B$. Con esto se calcula la tabla de frecuencias resultante resultante para la característica C.

3. Se elabora el peso para C (W_C) y se pondera la información por $W_A * W_B * W_C$. Con esto se calcula la tabla de frecuencias resultante resultante para la característica A.

4. Se elabora un segundo peso para A (W_{A2}) y se pondera la información por $W_A * W_B * W_C * W_{A2}$. Con esto se calcula la tabla de frecuencias resultante resultante para la característica B.

5. El proceso se continúa hasta que los pesos de cada observación convergen a un valor estable.

Este proceso suele converger con 2 o 3 iteraciones. La formula utilizada para la elaboración de pesos corresponde a la de pesos de no respuesta pues se conoce las distribuciones de los clientes para las características en estudio.

Capítulo 4

Metodología

Al ser este un problema perteneciente al área de minería de datos se cuenta con varias metodologías a disposición, como KDD, SEMMA o CRISP-DM, pero que en la práctica no presentan diferencias significativas en el desarrollo de un trabajo. Se opta por la utilización de la metodología CRISP-DM con una modificación a uno de los pasos iniciales para dar mayor énfasis al hecho de trabajar con encuestas.

4.1. Metodología CRISP-DM

La metodología *Cross Industry Standard Process for Data Mining* o CRISP-DM contempla el proceso de análisis de datos como un proyecto profesional, estableciendo así un contexto mucho más rico que influye en la elaboración de los modelos [22]. Este contexto tiene en cuenta la existencia de un cliente que no es parte del equipo de desarrollo, así como el hecho de que el proyecto no sólo acaba una vez se halla el modelo idóneo (ya que después se requiere un despliegue y un mantenimiento), sino que está relacionado con otros proyectos, y es preciso documentarlo de forma exhaustiva para que otros equipos de desarrollo utilicen el conocimiento adquirido y trabajen a partir de él.

El ciclo de vida del proyecto de minería de datos según esta metodología consiste en seis fases mostradas en la siguiente figura.

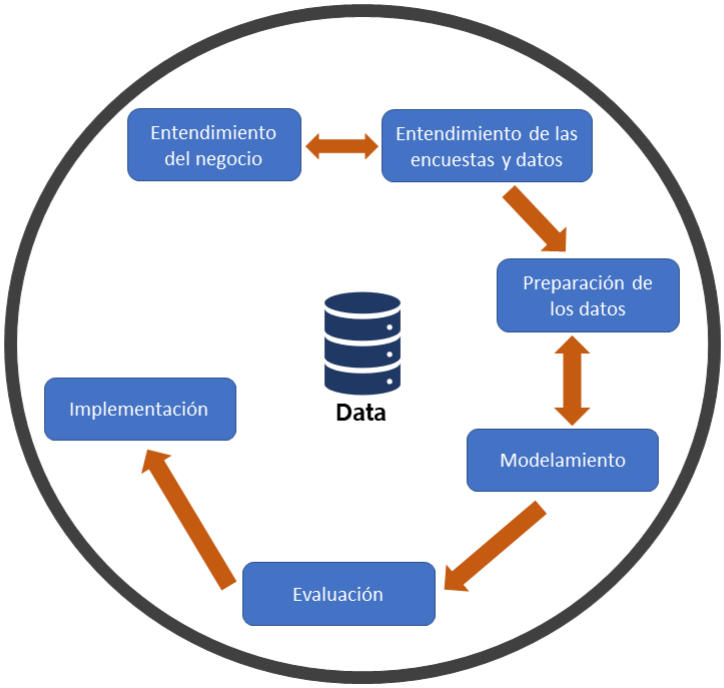

Figura 4.1: Representación metodología CRISP-DM modificada. Fuente: Elaboración propia.

- **Entendimiento del negocio:** Esta fase inicial se centra en el entendimiento de los objetivos del proyecto y los requerimientos desde una perspectiva del negocio, para convertir este conocimiento en un problema de definición de minería de datos y un plan preliminar diseñado para alcanzar los objetivos.
- **Entendimiento de las encuestas y los datos:** Esta fase abarca la colección de datos y las actividades exploratorias para la familiarización con ellos. La sección de entendimiento de las encuestas es añadida a esta etapa para dar énfasis a la importancia de familiarizarse también con las encuestas, identificar sesgos, detectar problemas en la calidad de la información o descubrir conjuntos interesantes que permitan elaborar hipótesis sobre los que trabajar más adelante.
- **Preparación de los datos:** Cubre todas las actividades para construir la base final de datos (datos que serán el alimento de las herramientas de modelado) desde una base en bruto. Es preferible que las tareas de preparación de datos se realicen varias veces y no en un orden preestablecido. Estas tareas incluyen tabulación, documentación y selección de atributos, como también transformación y limpieza de datos para las herramientas de modelado.
- **Modelado:** Se seleccionan y aplican varias técnicas, y sus parámetros son calibrados a los valores óptimos definidos. Por lo general hay varias técnicas para el mismo tipo de problema. Algunas técnicas tienen requerimientos específicos en la forma de los datos, por lo tanto será a menudo necesario devolverse a la fase de preparación de datos.
- **Evaluación:** Al llegar a esta fase se ha construido un modelo (o modelos) que aparentan tener una alta calidad desde la perspectiva del análisis de datos. Antes de proceder a

la entrega final del modelo es importante evaluarlo más a fondo y revisar los pasos ejecutados para construirlo, de tal forma que esté lo más cercano posible de alcanzar los objetivos del negocio. Un objetivo clave es determinar si hay algún evento importante del negocio que no haya sido considerado lo suficiente. Al final de esta fase, se debe tener una decisión sobre el uso de los resultados de minería de datos.

- **Despliegue:** La creación del modelo por lo general no es el final del proyecto. Incluso si el propósito del modelo es incrementar conocimiento sobre los datos, el conocimiento ganado necesitará ser organizado y presentado de una manera que el cliente lo pueda usar. A menudo implica aplicar modelos "en vivo" dentro del proceso de toma de decisiones de una organización, como por ejemplo, en la actualización de los resultados al ir obteniendo información mas reciente para este. Sin embargo, dependiendo de los requerimientos, la fase de despliegue puede ser tan simple como generar un reporte o tan compleja como implementar un proceso repetible de minería de datos a través de la empresa [5]. En muchos casos es el cliente, no el analista de datos, quien realiza los pasos de despliegue. Sin embargo, incluso si el analista no carga con el esfuerzo de despliegue, es importante que el cliente entienda que acciones deben ser llevadas a cabo para hacer uso de los modelos creados.

Dentro de la etapa de análisis de datos, parte de la información a utilizar no requerirá de transformaciones adicionales ya que se encuentran en un estado idóneo para su utilización (las tareas de limpieza y revisión de información si se realizan), mientras que se trabajarán 3 fuentes de información especialmente para este trabajo: datos de visitas de clientes a tiendas de la firma, datos provenientes del uso de la aplicación que tiene la empresa y por último, una base de datos que agrupa interacciones de diferentes fuentes como call center, reclamos, IVR[1], USSD[2], entre otros. Esto se realiza para acotar la etapa de preparación de datos ya que en la firma se cuenta con una cantidad muy elevada de información por lo que no es factible abarcar todos los datos disponibles.

Como paso inicial dentro de la etapa de modelado estará la construcción de la variable dependiente a ser utilizada dentro de esta sección del trabajo. En cuanto a la evaluación del modelo, este será centrado tanto en los resultados del modelo, como a las relaciones que se encuentren entre la satisfacción del cliente y los 3 factores que serán revisados, fuga, share de cambio de equipo y tasa de conversión de campañas.

El carácter de este trabajo es el de generar un aumento de conocimiento de los clientes para la firma, por lo que no se profundiza en una implementación específica del trabajo y todos los pasos que esto requiere, si no que se centra en ser un trabajo replicable por la firma a futuro.

[1] Interactive Voice Response, corresponde a los servicios de la compañía donde se realiza una llamada a una máquina con respuestas pre-grabadas.

[2] Acrónimo para Unstructured Supplementary Service Data, el cual corresponde a los servicios activados a través de la digitación de códigos, como por ejemplo el $*103\#$.

Capítulo 5

Desarrollo

5.1. Revisión de las encuestas

Se comienza el trabajo con la descripción de la estructura de las encuestas, los criterios que utilizan para la selección de encuestados, y una revisión de cuales son las preguntas realizadas tanto para las encuestas internas de la firma como las realizadas por la consultora contratada. Para el presente trabajo se cuenta con información de encuestas realizadas desde enero del 2019 hasta marzo del 2020.

Ambas encuestas son realizadas de manera mensual y son realizadas únicamente mediante llamadas telefónicas. La forma en que se eligen los encuestados por parte de la firma es la siguiente, los números son elegidos mediante un muestreo probabilístico y una vez seleccionados estos deben cumplir las siguientes restricciones.

- El cliente debe ser mayor de 14 años.
- Deben ser clientes que no trabajen en empresas de investigación o telecomunicaciones.
- El cliente debe presentar interacciones de datos dentro del mes anterior a la encuesta.

La última restricción se realiza a partir de los datos que maneja la compañía mientras que las 2 primeras son preguntadas al encuestado durante la entrevista y en caso de cumplir una de estas dos, se termina la realización de la encuesta. Se busca tener cada mes una cantidad similar de encuestas contestadas, por lo que la cantidad de móviles llamados varía cada mes. La tasa de respuesta promedio corresponde a un 53 %.

La escala utilizada para las respuestas es del 1 a 7 siendo el 1 la peor nota que se puede asignar y un 7 la mejor, sin admitir valores no enteros. Estas notas no están explícitamente relacionadas a una categoría como “muy satisfecho” o “levemente insatisfecho” como suele ocurrir en otras encuestas, como formularios online.

Las preguntas realizadas en cuanto a satisfacción se formulan de la siguiente manera, “en una escala del 1 al 7, que tan satisfecho se encuentra con...”, continuando con una de las siguientes 6 categorías, la satisfacción general con la compañía, con los servicios de voz, con los servicios de datos, en el caso de ser cliente pospago, la satisfacción con el plan contratado

mientras que para los clientes prepago se pregunta por la satisfacción con los precios y satisfacción con el proceso de recarga. Para el caso de la satisfacción general, luego de que el encuestado entrega su nota, se le pide entregar el motivo por el cual se le asignó esa nota a la compañía.

Además de lo anterior, se consulta la edad, la región donde vive el encuestado, su sexo, el mercado al que pertenece (empresa, prepago, pospago) y si el encuestado es quien paga por el servicio o no.

Las preguntas siguen la siguiente secuencia, se comienza por preguntar al encuestado si es que trabaja en investigación de mercado o no. luego se consulta por el mercado al que pertenece el cliente. Luego de esto vienen las preguntas de satisfacción, comenzando por la satisfacción general y la justificación de la nota. Se sigue con la satisfacción en los servicios de voz, datos, y luego si el cliente es pospago se pregunta por la satisfacción con el plan mientras que a los prepago se les pregunta por el precio de las recargas y el proceso de recarga. La encuesta finaliza con las preguntas sobre la edad región y sexo.

En cuanto a las encuestas realizadas por la consultora, estas comparten algunas similitudes con las encuestas de la empresa, se pregunta por la edad, sexo y región a manera de control. Se utiliza la misma escala de respuestas de 1 a 7 y se realizan las mismas preguntas de satisfacción anteriores, con la adición de más preguntas que extienden la duración de esta encuesta.

Se pregunta por la satisfacción con el servicio al cliente, si ha utilizado las ofertas del club de beneficios que posee la firma y su evaluación al respecto en el caso de que si los haya usado. Se pregunta si el cliente ha utilizado algún canal digital entre la aplicación o la web y si el acceso (en el caso de haber utilizado alguno de estos canales) fue realizado desde celular o computador y luego la evaluación respectiva de cada uno.

Se pregunta si en el último mes se ha usado o intentado usar algún canal de contacto como sucursal, call center, app, etc. y se registra cuales. También se pregunta desde hace cuanto tiempo el encuestado es cliente de de la compañía.

La consultora entrega a la firma solo los resultados de quienes responden la encuesta, por lo que no se tiene el porcentaje de no respuesta para este caso. La consultora también entrega una cantidad de respuestas fija cada mes, por lo que se desprende que llama a la cantidad necesaria de clientes para poder alcanzar esta cantidad cada mes.

Ambas encuestas tienen una estructura clara, fácil de entender y una misma escala de valores del 1 al 7. La razón de utilizar esta escala es el de buscar semejanza con la escala de notas que se utiliza dentro del país, la cual va de 1 a 7 y de esta forma facilitar al cliente la tarea de decidir su respuesta a partir de una escala que ya le es familiar.

La primera evaluación que se pide realizar en ambas encuestas es la nota a la satisfacción general con la empresa. Que esta sea la primera pregunta significa que su respuesta no pudo ser influenciada por preguntas anteriores, evitando así caer en sesgo en las preguntas. Otra conclusión obtenida es que desde este punto en adelante se pueden trabajar ambas encuestas de manera conjunta, ya que se trabajara sólo con esta pregunta, por lo que no es necesario

considerar factores que si pueden estar influyendo al resto de preguntas, como lo puede ser la duración de cada encuesta debido a la cantidad de preguntas que tiene cada una.

Por estructura, el sesgo de selección de encuestados no debería estar presente en ninguna de las 2 encuestas, ya que el medio por el cual se encuesta es a la vez el servicio que se entrega. Esto permite el llegar a la totalidad de clientes, a diferencia de una encuesta online o por correo, donde resulta inverosímil el poder llegar a todos los clientes.

En efecto, esto es lo que ocurre con las encuestas realizadas por la consultora, sin embargo, las encuestas realizadas de forma interna por la firma al tener restricciones adicionales, puede estar cayendo en este sesgo.

Es bastante probable que la restricción de haber tenido tráfico de datos el mes anterior a ser encuestados pueda invisibilizar la opinión de usuarios que solo utilicen los servicios de voz [1]. Lamentablemente a partir de las encuestas contestadas no hay forma de aseverar con seguridad si este es el caso o no. Una posibilidad es comparar si existen diferencias entre la distribución de edades de los clientes que responden le encuesta y el universo completo de clientes, ya que se sabe que las personas de mayor edad son los que hacen menor uso de internet.

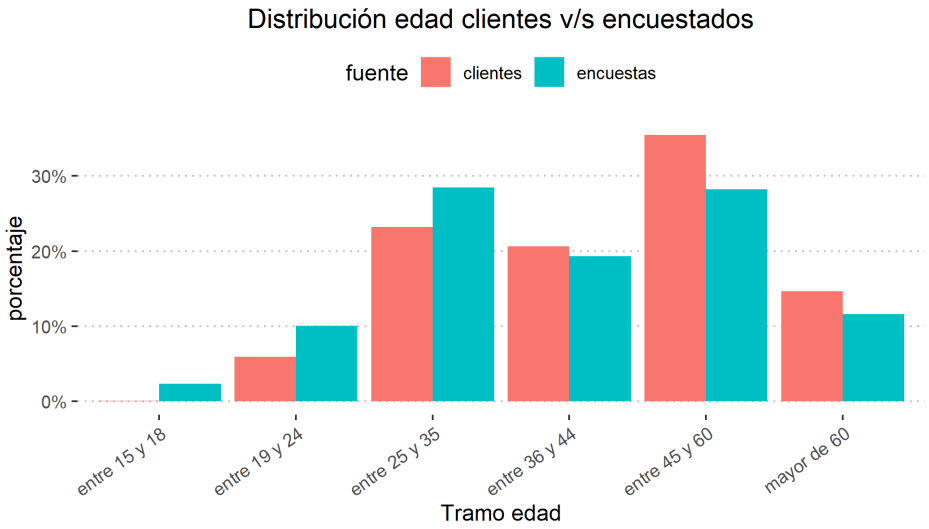

Figura 5.1: Distribución edades encuestas y total de clientes.

Una primera apreciación de la figura 5.1 muestra que existen diferencias entre ambas distribuciones, pero que todas estas son menores a los 10 puntos porcentuales para cada tramo de edad. se puede ver que los porcentajes de gente que contestan las encuestas tienen una concentración mayor en los tramos de clientes mas jóvenes. Este hecho puede deberse tanto a la restricción establecida como a la existencia de sesgo de no respuesta.

[1] Estrictamente, la restricción de edad es algo que también genera sesgo de selección, pero este caso corresponde a una decisión de negocio por parte de la firma que quiere omitir este segmento, no así con los clientes que solo utilizan servicios de voz, por lo que no se intenta compensar por la restricción de ser mayor de 14 años de edad.

Se aplica un test de diferencia de distribuciones chi cuadrado, este test tiene como hipótesis nula que las distribuciones de ambas muestras son iguales. El resultado de este test arroja un p-valor con el cual se rechaza la hipótesis nula al 95 % de confianza y se concluye que las distribuciones no son iguales. Continuando con otras características de los clientes, la figura 5.2 compara la distribución de sexos, región, GSE y segmento al que pertenecen los encuestados y el universo de clientes pospago.

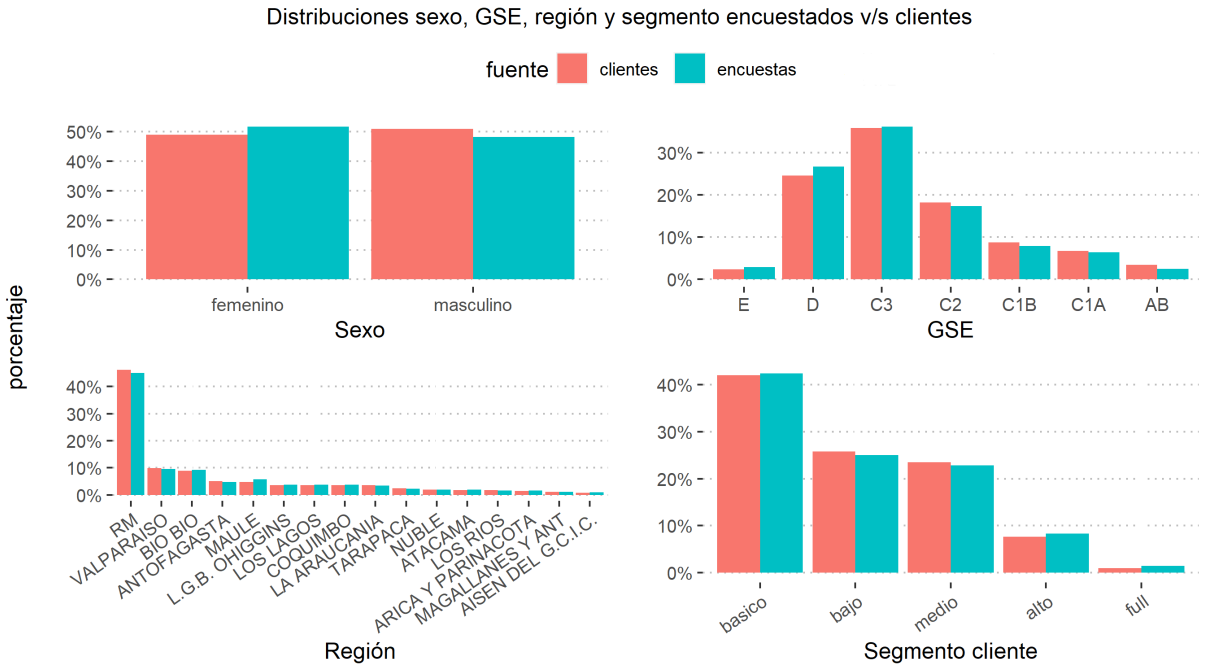

Figura 5.2: Comparación de distribuciones para el sexo, región, GSE y segmento cliente.

Las conclusiones son similares, las diferencias son tan solo de unos cuantos puntos porcentuales en los casos con mayores diferencias y el resultado de los test de distribución concluyen que estas son diferentes para todos los casos. Debido a estas diferencias se concluye que es necesario incorporar pesos en las respuestas de los clientes para conseguir resultados que sean mayormente representativos de la población de la firma.

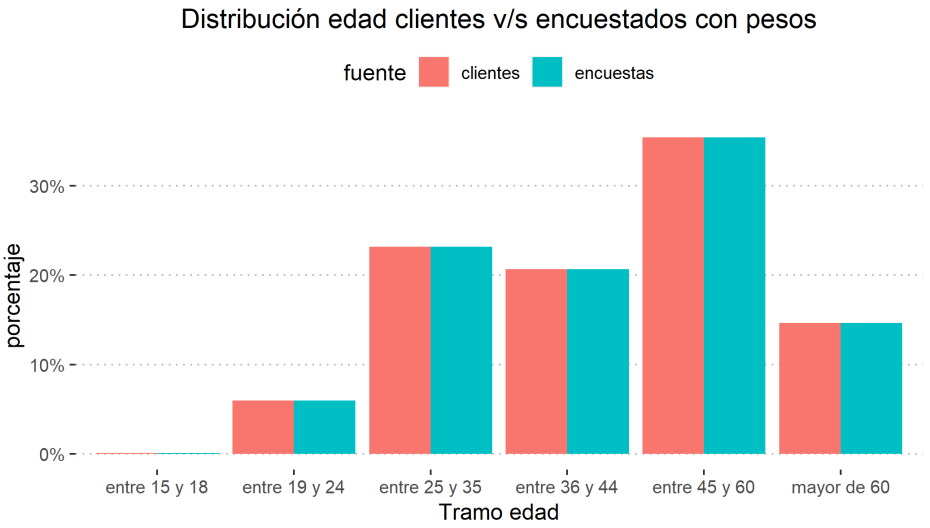

Figura 5.3: Distribución edades encuestados con pesos y total de clientes.

La figura 5.3 muestra el resultado de comparar la distribución de edades luego de incorporar los pesos mencionados a las respuestas de los clientes. El mayor beneficio obtenido dentro de las 5 dimensiones estudiadas esta en la mayor consideración que los pesos darán a las respuestas de gente de 45 años hacia arriba, quienes estaban siendo subrepresentados.

Pasando a la existencia de otros posibles sesgos, no se cuenta con grabaciones de las encuestas ni conocimiento de quienes fueron los encuestadores, por lo cual no es posible aseverar nada respecto a la existencia o no de sesgos producidos por parte del entrevistador durante la realización de las preguntas. Lo que si se posible es revisar el porcentaje de errores ocasionados por una mala transcripción de las respuestas a algún valor fuera del rango 1 a 7, lo cual para la pregunta de satisfacción general corresponde a un porcentaje cercano al 1 %. Estos casos son descartados del resto de respuestas al ser una cantidad despreciable de casos.

La estructura de ambas encuestas es simple y las preguntas son de una baja complejidad, por lo cual se descartan la existencia de posibles problemas por estos factores y la existencia de un sesgo en las preguntas. Sumando también el hecho de que la pregunta sobre satisfacción general es la primera que se realiza, se concluye que el estado de las respuestas es de buena calidad para el trabajo posterior luego de la limpieza de respuestas fuera de rango y el agregado de pesos.

Se hace mención a factores que no pudieron ser revisados dentro de esta revisión. Dentro de estos se encuentra el identificar errores donde el encuestador se equivocara transcribiendo otra nota pero que aún así esta se encuentre dentro del rango 1 a 7, tampoco se identificaron casos donde el encuestado respondiera entregando respuestas falsas, ni la presencia o no de otros sesgos provenientes por parte del cliente, como sesgo de consentimiento. Se trabajará bajo la hipótesis de que estos efectos tienen su menor impacto en la pregunta de satisfacción general debido a que esta es la primera pregunta realizada.

5.2. Selección y elaboración de atributos

Para la tarea de predicción de satisfacción, se utilizará información relacionada a diferentes aspectos de los clientes, los cuales se encuentran resumidos en la siguiente figura.

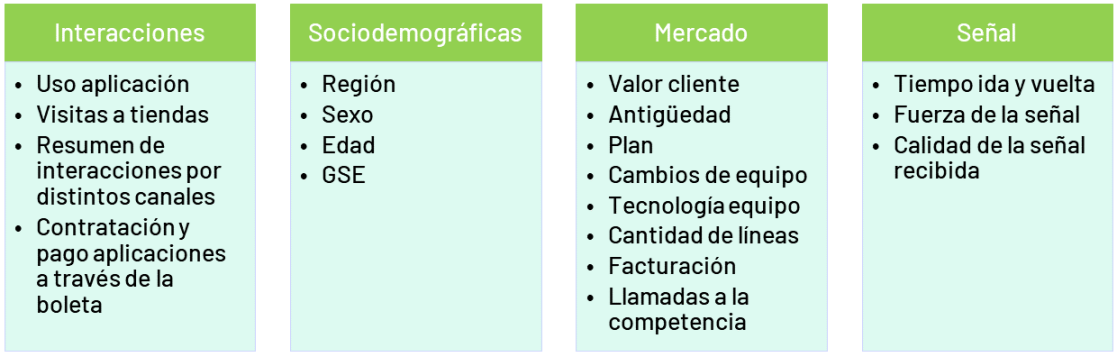

Figura 5.4: Resumen de variables a utilizar para modelos de satisfacción.

La información socio demográfica, de mercado y relacionada con la señal es obtenida de forma directa a partir de bases de datos que maneja la empresa y que esta agregada al nivel necesario para ser utilizada de forma directa, por lo que solo se realizan las tareas de limpieza y verificación de los datos en estos casos.

Si bien el trabajo se centra en el efecto de las variables relacionadas a interacciones, es de esperar que variables relacionadas a otros factores como la señal o uso de internet tengan un peso significativo en el nivel de satisfacción del cliente. Por esta razón se decide complementar la información de interacciones con datos sobre señal, para luego ver cuales de estas tienen mayor importancia en la predicción de satisfacción.

La información relacionada a interacciones no ha sido trabajada en ocasiones anteriores por la firma, por lo que esta desagregada y requiere un trabajo previo, el cual se aborda en la siguiente sección. La información a trabajar corresponde a la del uso de la aplicación, visitas a tiendas de la firma y un resumen mensual de interacciones realizadas por clientes en diferentes canales (en anexos se encuentra un resumen de todas las variables utilizadas dentro del trabajo con una breve descripción para cada una).

5.2.1. Visitas a tiendas

La primera base de datos revisada consiste a los datos obtenidos a partir de los turnomáticos ubicados en varias de las tiendas de la empresa. Estos corresponden a totems dentro de estas tiendas donde uno debe ingresar su rut de manera previa a ser atendidos.

La data se actualiza de forma mensual y cuenta con ingresos de ruts históricos desde enero del 2019 hasta la fecha actual, por lo que cuenta con información dentro de todo el periodo estudiado. Considerando la información hasta marzo del 2020, la tabla tiene un tamaño aproximado de 4,8 millones de ingresos de ruts en tiendas diferentes a lo largo del país. Los datos cuentan con información que identifica la tienda a la que fue el cliente y fecha en que se ingreso el rut, el mercado al que pertenece el cliente en caso de conocerse esta información y el tipo de tienda, pudiendo ser una tienda propia de la compañía o una franquicia.

Para comenzar se cruza la información con los ruts de los encuestados, se limpia información con errores y para cada cliente encuestado se elimina cualquier visita que ocurriera en una fecha posterior a la fecha en que se realizó la encuesta. Con esto se obtiene información de visitas para un 30 % de los encuestados aproximadamente, por lo que se cuenta información suficiente para estudiar esta tabla.

A continuación, se quiere estudiar la relación entre las visitas realizadas por un cliente y su satisfacción con la firma. Una de las opciones es obtener el total de visitas históricas por cliente, pero esto generaría que los clientes encuestados en los meses más recientes tengan una ventana mayor de información que aquellos encuestados en meses cercanos a enero del 2019.

Para evitar ambigüedades en la cantidad de información que representa la variable, se decide acotar la ventana de tiempo a considerar para cada cliente. Se utilizan 3 ventanas de tiempo, 1, 2 o 3 meses de información previa a la encuesta para construir la variable de cantidad de visitas realizadas por los clientes. La figura 5.5 muestra como distribuye la cantidad de visitas realizadas por los encuestados en el caso que donde se considera una ventana de 3 meses de historia.

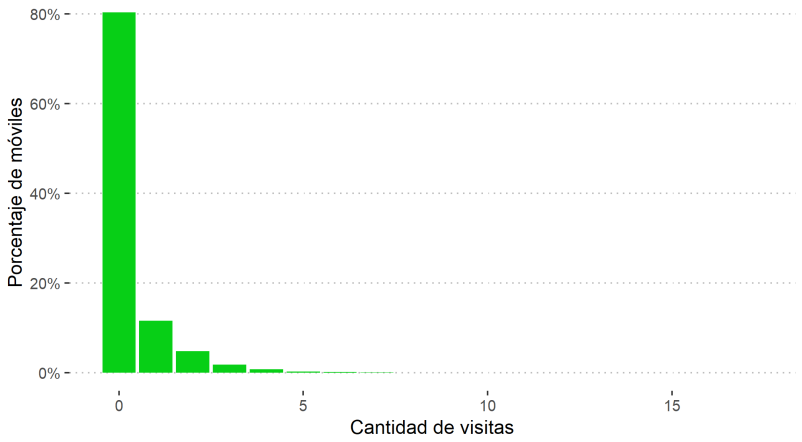

Figura 5.5: Conteo de clientes según cantidad de visitas realizadas en un periodo de 3 meses.

Al utilizar 2 meses de historia se descartan las encuestas realizadas en enero del 2019 y comienzos de febrero debido a que no se tiene información completa para estos casos. Cuando se utilizan 3 meses se descartan también las encuestas realizadas desde comienzos de marzo 2019 hacia atrás.

En el caso mostrado en la figura 5.5 utilizando una ventana de 3 meses, se tiene que aproximadamente un 20 % de los clientes encuestados realizó una o más visitas a tiendas de la firma. Este porcentaje se reduce con al ocupar solo 2 meses de historia y aún más al ocupar un solo mes, donde solo un 5 % de los clientes encuestados realiza una o más visitas a tiendas dentro de esa ventana.

A modo de conseguir una primera mirada del efecto de las visitas en la predicción de satisfacción, se presenta a continuación la satisfacción promedio mensual de los clientes que realizaron 1 sola visita y los clientes que realizaron 2 o más visitas contra la satisfacción general del mismo mes (utilizando 3 meses de historia de los clientes).

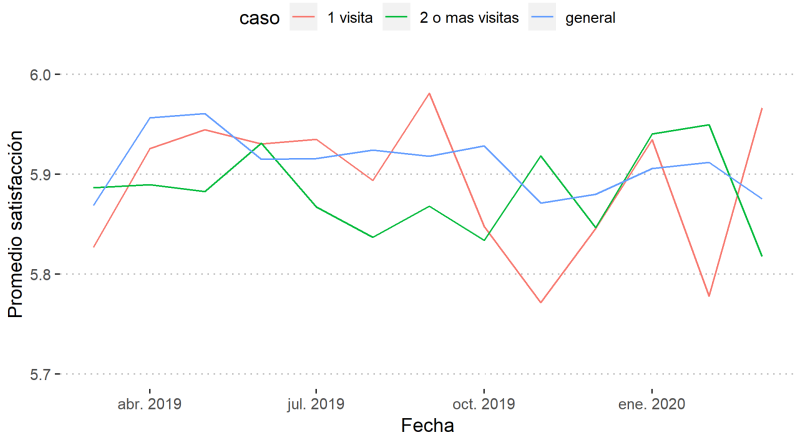

Figura 5.6: Satisfacción promedio mensual según cantidad de visitas a tiendas realizadas.

De la figura 5.6 se puede ver que la variabilidad de la satisfacción para clientes que visitan tiendas aumenta en comparación a los resultados generales, lo cual es de esperar al tener una muestra de menor tamaño según lo visto en la figura 5.5. No se puede concluir a partir de este gráfico que los clientes que han realizado una sola visita a tiendas o dos o más visitas tengan una satisfacción consistentemente diferente que la satisfacción general.

Dado esto, se propone diferenciar las visitas entre visitas que fueron hechas a tiendas propias de la empresa o si estas fueron visitas a franquicias, manteniendo las mismas ventanas de tiempo. Las tiendas propias de la firma tienen a ser más grandes y contar con mayor personal para atender los diversos tipos de requerimiento del cliente, como servicio al cliente o compra de productos, en comparación a una franquicia. La figura 5.7 muestra la satisfacción promedio de clientes que realizaron alguna visita a una tienda propia junto con la satisfacción de los clientes que visitaron alguna tienda franquiciada, comparando con la satisfacción general.

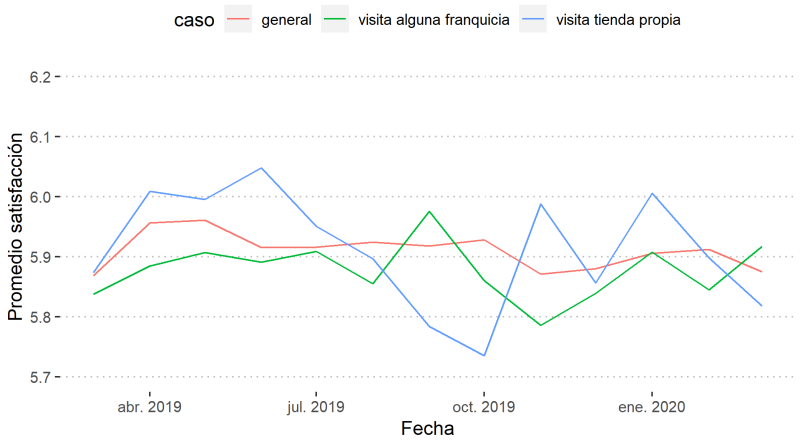

Figura 5.7: Satisfacción promedio mensual según tipo de tienda visitada.

La firma cuenta con una mayor cantidad de franquicias que de tiendas propias, por lo que una menor cantidad de visitas se realiza a estas últimas, lo que explica la variabilidad de los resultados obtenidos para este caso comparado a las visitas a franquicias. Se puede ver que, salvo un par de meses, disminuye levemente la satisfacción promedio si el cliente realizó alguna visita a una franquicia dentro de los últimos 3 meses, mientras que no se ve un efecto claro entre la visita a una tienda propia de la firma y la satisfacción del cliente.

Si dos variables tienen una correlación muy elevada significa que están describiendo la misma información. Las visitas totales y las visitas separando por tiendas explican el mismo evento, por lo que se estima que su correlación debe ser alta y esto produciría problemas en modelos futuros ya que podría concluirse que una variable es significativa cuando en verdad cualquiera de las dos podría haber sido elegida.

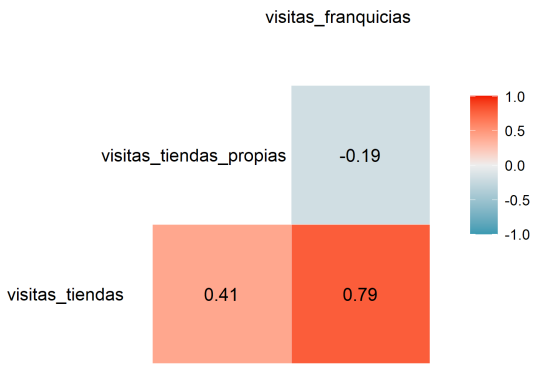

Figura 5.8: Correlación variables de visitas a tiendas con 3 meses de historia.

la figura 5.8 muestra que las visitas totales tienen una correlación elevada con las visitas a franquicias. Como ya se mencionó, la firma cuenta con una mayor cantidad de franquicias que de tiendas propias, lo que explica por que la correlación es mayor entre estas dos variables. Se decide mantener las visitas a tiendas con la diferenciación del tipo de tienda por sobre las visitas totales, dejando finalmente 2 variables en 3 ventanas de tiempo diferentes.

- Número de visitas realizadas a tiendas propias o tiendas franquiciadas *visitas_TP_1M, 2M, 3M y visitas_TF_1M, 2M y 3M.* respectivamente.

Estas 3 ventanas de tiempo, al representar la misma información también tienen una correlación alta entre ellas, por lo cual se deberá decidir cual de estas ventanas tiene mayor poder predictivo de satisfacción en los modelos y eliminar las otras dos. El trade off de aumentar la ventana de tiempo es que se puede tomar en consideración una mayor cantidad de visitas de clientes a tiendas, pero se tiene la hipótesis de que visitas realizadas en un periodo más corto de tiempo influyen más en la satisfacción que una visita realizada en un plazo mayor.

5.2.2. Uso de la aplicación

La compañía cuenta con una aplicación donde ofrecen diversos servicios y ofertas a sus clientes, como consulta de saldo y datos usados, servicio técnico, ofertas de cambios de equipo y plan, etc. Para comenzar a utilizar la app es necesario ingresar un número telefónico, nombre y rut. Con esta información la aplicación te deriva a una versión específica de la misma que entregará funcionas diferentes dependiendo del tipo de usuario, el cual puede ser prepago, pospago, cliente hogar o usuario no cliente.

Los datos que se tienen de interacciones realizadas dentro de la aplicación no han sido exploradas de manera previa por la firma debido a no contar con suficientes meses de historia de su uso. Para esta ocasión se tienen datos sobre el uso de la aplicación disponibles desde noviembre del 2019 en adelante.

El uso de la aplicación por parte de los clientes pospago de la firma va en aumento, siendo utilizada por un 28 % de los clientes al menos una vez durante marzo del 2020.

Debido a que se tiene una cantidad menor de historia para los clientes sobre el su uso de la aplicación, se fija la ventana de tiempo para las variables derivadas de esta tabla a 1 mes de historia, lo que significa tener datos para encuestados de diciembre hasta marzo. A pesar de contar con menos meses de información sobre estos datos, la empresa esta interesada en entender si ya este volumen de dato permite a la información proveniente de la aplicación aportar en la predicción de satisfacción.

En la aplicación pospago se puede revisar la boleta de gastos, consultar el consumo del plan y revisar el plan contratado. Se puede acceder a servicios de ayuda y soporte técnico (la cual consiste en enviar el celular a revisión con un técnico) y se tiene la posibilidad de acceder a un chat por Whatsapp con un bot desarrollado por la empresa. También se puede ingresar a una sección donde se pueden canjear diversos beneficios por ser cliente de la empresa.

Por último se cuenta con la posibilidad de adquirir los servicios de Spotify y Netflix a través de la aplicación (clasificado como sección de entretenimiento), donde la diferencia al adquirir los servicios a través de la firma es que estos serán cobrados luego a través de la boleta del plan, buscando ofrecer un beneficio tanto en ofertas de precio de estos servicios como también en la organización de gastos del cliente.

Tabla 5.1: Interacciones con la app pospago durante marzo

interacción	**% de uso del servicio de la app**	**% de uso del total de encuestados**
menu y dashboard	82 %	21 %
boleta	64 %	17 %
mi plan	55 %	14 %
consumo	46 %	12 %
equipo	39 %	10 %
ofertas	38 %	10 %
chat	29 %	7 %
beneficios	23 %	6 %
entretenimiento	16 %	4 %
soporte	11 %	3 %

La tabla 5.1 desglosa el nivel de uso que tuvo la aplicación por tipo de servicio ofrecido dentro de los meses mencionados. La primera columna detalla el servicio dentro de la app, la segunda columna describe el porcentaje de uso que tiene el respectivo servicio entre todos los clientes encuestados que tienen la app. La tercera columna muestra cual es el porcentaje de clientes que utilizó cada servicio entre el total de clientes encuestados.

Dejando de lado la interacción de menu y dashboard, ya que la función de esta es el de acceder al resto de servicios, de la tabla 5.1 se puede ver que el uso principal que se le esta da a la aplicación es de revisión de la boleta, seguido de la revisión del plan y el consumo realizado. Los servicios de chat, beneficios, entretenimiento y soporte son las que presentan

menor utilización. Si bien el uso de estos últimos es bastante bajo, se tiene la hipótesis de que estos servicios generan un impacto significativo en la satisfacción, por lo que se mantienen como variables a pesar de su bajo porcentaje de uso.

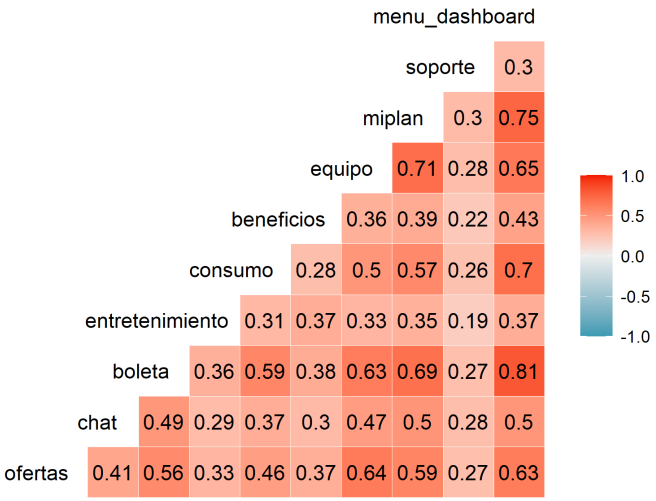

Figura 5.9: Correlación entre variables derivadas del uso de la app.

En primera instancia se decide estudiar la relación entre la cantidad de interacciones que realiza un cliente por tipo de servicio y la satisfacción. De manera previa a ver sus efectos se revisa la correlación entre estos servicios en la figura 5.9. Se puede ver que en general el uso de los distintos servicios de la app tiene correlaciones positivas de distinta magnitud.

Una correlación muy elevada significa que las variables están describiendo la misma información, por lo que se decide tomar como criterio de corte o modificación las variables con correlaciones alrededor de 0,6 hacia arriba. Se ve una correlación elevada entre los servicios relacionados a la consulta de estado de servicios que corresponden al estado del plan, equipo, revisión de boletas y de consumo, por lo que se toma la decisión de unir estas categorías en una sola. En cuanto a la categoría de menu y dashboard, se opta por eliminarla del estudio al ser una interacción para moverse entre los servicios de la aplicación y no representa un tipo de uso particular.

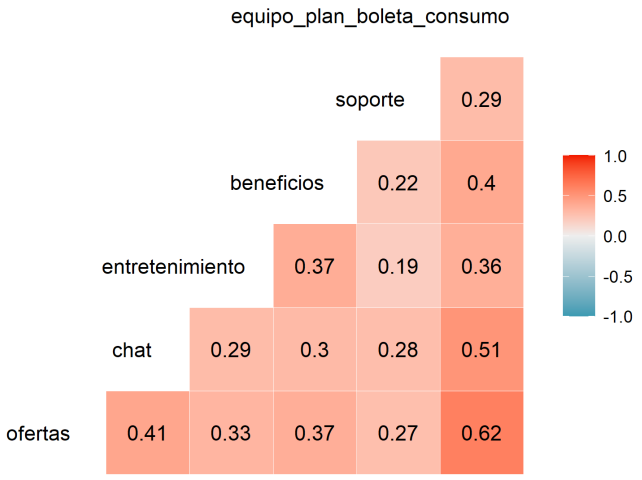

Figura 5.10: Correlación final entre variables de la app.

La figura 5.10 muestra las correlaciones luego de los cambios. Se puede ver que las interacciones sobre ofertas y la unión de interacciones realizada bordean el criterio mencionado anteriormente, pero últimamente se mantienen separadas al no ser una interacción relacionada al chequeo del estado del servicio. Con esto, la figura 5.11 muestra las diferencias entre la satisfacción promedio de los clientes que tuvieron una o mas interacciones por cada uno de los servicios mencionados.

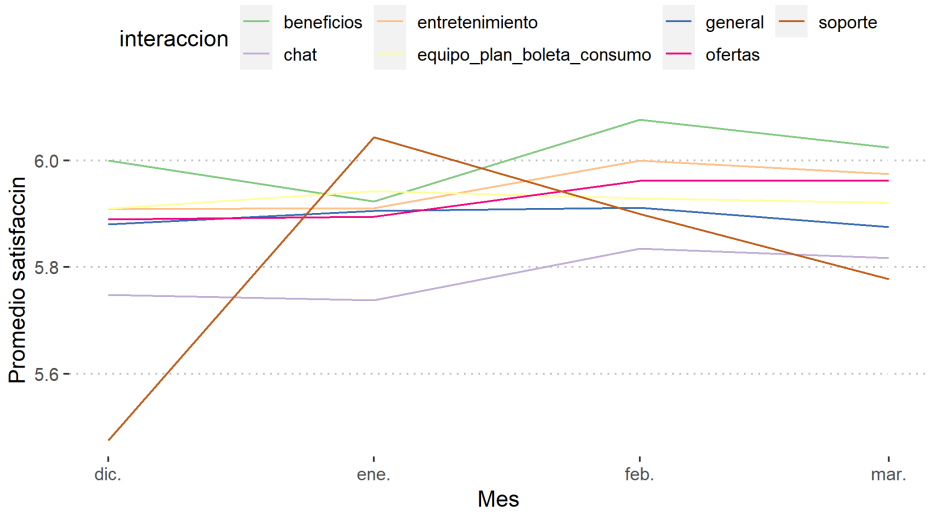

Figura 5.11: Satisfacción promedio mensual diferenciando por servicio utilizado dentro de la aplicación.

Como se puede ver del gráfico 5.11 la satisfacción de clientes que tienen interacciones de plan o boleta, ofertas y entretenimiento son similares a la satisfacción general. Los clientes que tienen interacciones con el servicio de beneficios presentan un promedio mayor de satisfacción mensual en comparación al resto de servicios, mientras que los clientes que acceden al chat tienen satisfacción promedio constantemente menor en el tiempo.

Estos resultados son esperados ya que el primer caso busca un aumento en la satisfacción de los clientes mientras que el segundo tiene por objetivo el ser un canal para clientes con problemas con su servicio o equipo. Lamentablemente la diferencia que estos dos servicios tienen con la satisfacción general es bastante leve por lo que no son significativas para afirmar si cumplen o no sus objetivos.

El otro servicio que destaca es el de soporte, el cual presenta la mayor variabilidad de todos debido a ser el servicio con menor utilización por parte de los clientes, impidiendo obtener una respuesta clara sobre su efecto en la satisfacción. Si bien el bajo uso de este servicio justifica el no incluirla en análisis posteriores, este servicio es el único que involucra la acción de mandar el celular a ser revisado por un técnico de la firma, acción que involucra varias interacciones con los servicios de la firma, por lo que se decide mantenerla para los modelos posteriores.

Se destaca el hecho de que la satisfacción promedio de los clientes que hacen uso de algún servicio de la aplicación parecen tener una variabilidad baja, pero lamentablemente la falta de una mayor cantidad de meses para estos datos imposibilita el determinar que tan consistente es este resultado en el tiempo. Con esto, las variables elaboradas para modelar corresponden a las siguientes:

- Cantidad de interacciones por tipo de interacción realizada en ventana de un mes, *ofertas, chat, beneficios, soporte, entretenimiento y revisión del consumo/equipo/plan/boleta*.

5.2.3. Base resumen de interacciones

Las siguiente información a revisar corresponde a una tabla que resume diversas interacciones realizadas por los clientes en una sola fuente de forma mensual. Esta tabla describe la fecha y canal donde se realizó la interacción, junto con las columnas *nivel_1 y nivel_2*, que describen la interacción realizada.

Luego del cruce con los clientes encuestados, se tiene que en promedio un 70 % de clientes tiene alguna interacción dentro de esta tabla cada mes. La figura 5.12 resume la cantidad de registros que se tiene por canal. Se puede ver que la mayor cantidad de interacciones de los clientes se dan a través de los canales USSD, web, call center e IVR. Del resto de canales, se descartan aquellas realizadas por tiendas al estar consideradas ya en la fuente revisada anteriormente, a la vez que se descarta el uso de la categoría “otro” al desconocer que canales representa y contener una cantidad baja de registros.

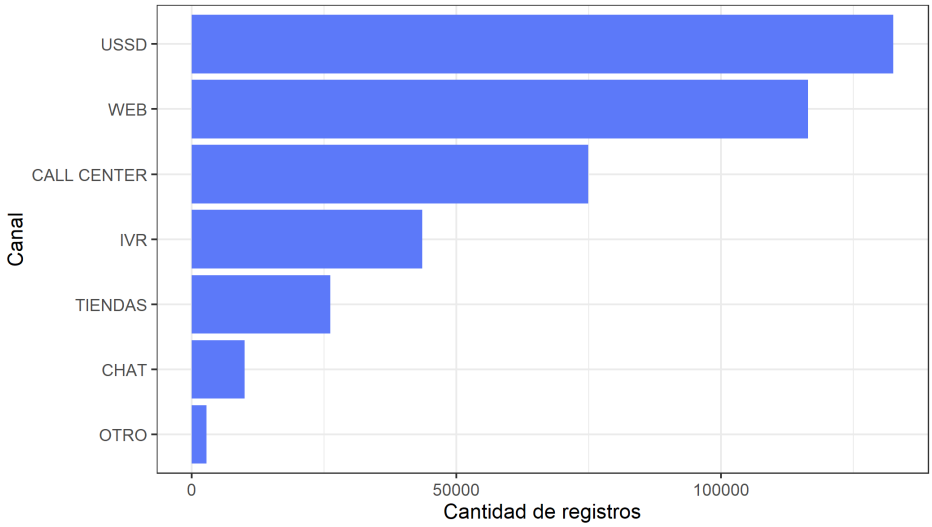

Figura 5.12: Conteo de registros disponibles agrupando por canales.

Sobre las dos columnas que describen la interacción. El *nivel_1* las clasifica dentro de 4 categorías, servicio, distribución, consulta o solicitud de información y problema o reclamo, las cuales se pueden ver en la figura 5.13.

De estas categorías, servicio corresponde a la asignación que tienen la interacción por defecto, donde está incluida toda la mayor parte de la navegación web, llamadas de call center y los servicios de consultas de saldo, boleta, consumo, de cualquiera de los canales donde esto es posible (USSD, WEB, IVR, CHAT). Dentro de distribución se encuentran las interacciones que terminan en la compra o cambio de servicios como la compra o cambio de un plan, termino de algún servicio, compra de equipos, etc.

En consulta de información caen los casos en que el cliente termina termina comunicándose con un ejecutivo por motivos como por ejemplo la consulta de disponibilidad de algún servicio en cierto sector, sin que esta consulta terminara en una compra. En caso de que el motivo de llamar a un ejecutivo sea por una queja con el servicio entonces se clasifica como problema o reclamo. Estos casos se pueden dar para los canales de chat, call center y web.

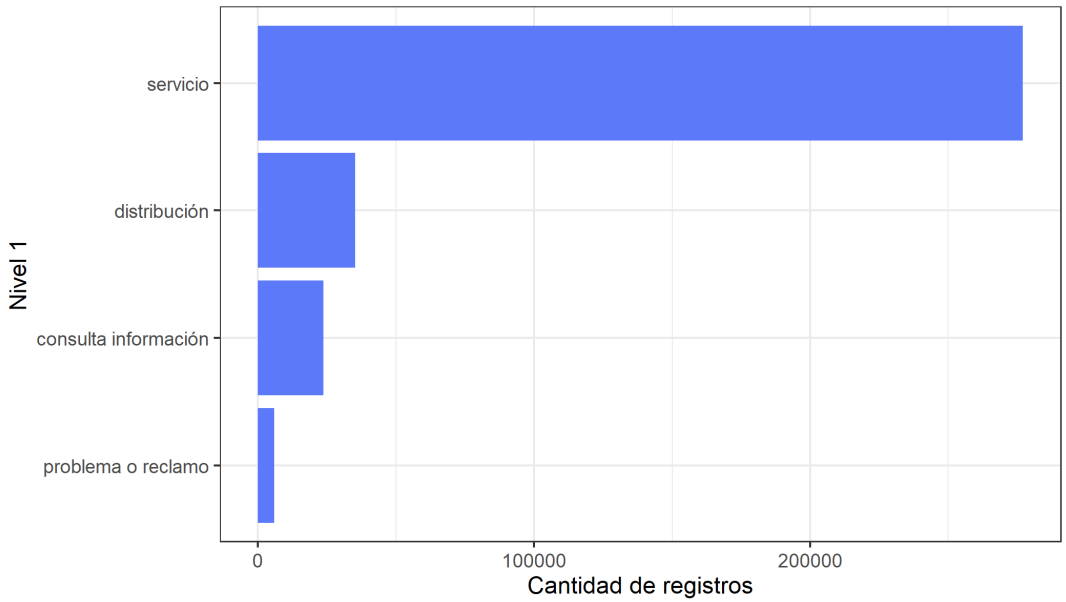

Figura 5.13: Cantidad de registros por clasificación en nivel 1.

El *nivel_2* corresponde a un campo de comentarios por parte de ejecutivos para los casos en los cuales se tuvo interacción con alguno, por lo que este campo cuenta con información principalmente para las interacciones con el call center, la pagina web de la firma y los chats. Estos comentarios son libres, por lo que existe una gran cantidad de casos en los cuales no hay información en este campo o comentarios que no se repiten en ninguna otra interacción.

Para los casos en los que si hay información, se realiza una limpieza a los comentarios de caracteres especiales y de stopwords, con lo que luego se realiza un conteo de palabras. Con esto se identifican 3 casos con mayor repetición dentro de este campo, cuando la interacción tiene que ver con problemas de facturación del cliente, cuando el cliente declara estar insatisfecho con el servicio o tener intenciones de renuncia y casos donde el cliente quiere realizar una anulación o bloqueo de algún servicio. La cantidad de registros dentro de estos 3 grupos se puede ver en la figura 5.14.

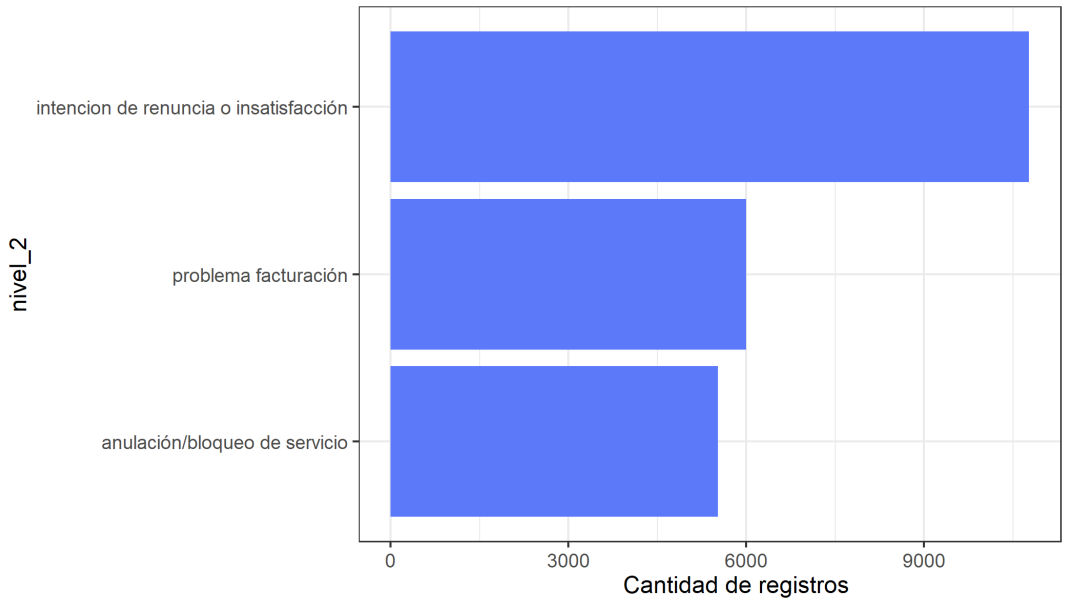

Figura 5.14: Cantidad de registros por clasificación en nivel 2.

Con la información de interacciones por canal y las clasificaciones del *nivel_1 y nivel_2*, se crean variables que cuentan la cantidad interacciones para cada canal y descripción, para cada cliente. Nuevamente se establecen como ventanas de tiempo utilizar 1, 2 y 3 meses de historia para la creación de cada variable.

Con esto se procede a replicar el análisis de correlaciones (para este caso las correlaciones se encuentran en anexos), a partir del cual se decide eliminar las clasificaciones de servicio y consulta de información, al tener correlaciones altas con con interacciones dentro del canal USSD y call center respectivamente. Esto nos deja con las siguientes variables par revisar dentro de las interacciones en pospago.

- Conteo de interacciones dentro de cada canal en las 3 ventanas de tiempo, *Call center, IVR, USSD, Web, Chat.*
- Conteo de interacciones que tuvo el cliente dentro de las clasificaciones *problema o reclamo, distribución, anulación/bloqueo de servicio, problema de facturación, intención de renuncia o insatisfacción* para las 3 ventanas de tiempo estudiadas.

La figura 5.15 muestra la satisfacción mensual promedio general en comparación a la de clientes que tienen al menos una interacción dentro de los casos considerados. Se puede ver como los clientes que presentan quejas tienen la menor satisfacción promedio vista hasta ahora y se ve una mejora durante el año en la satisfacción de estos clientes, pero que aún es más baja que la satisfacción general.

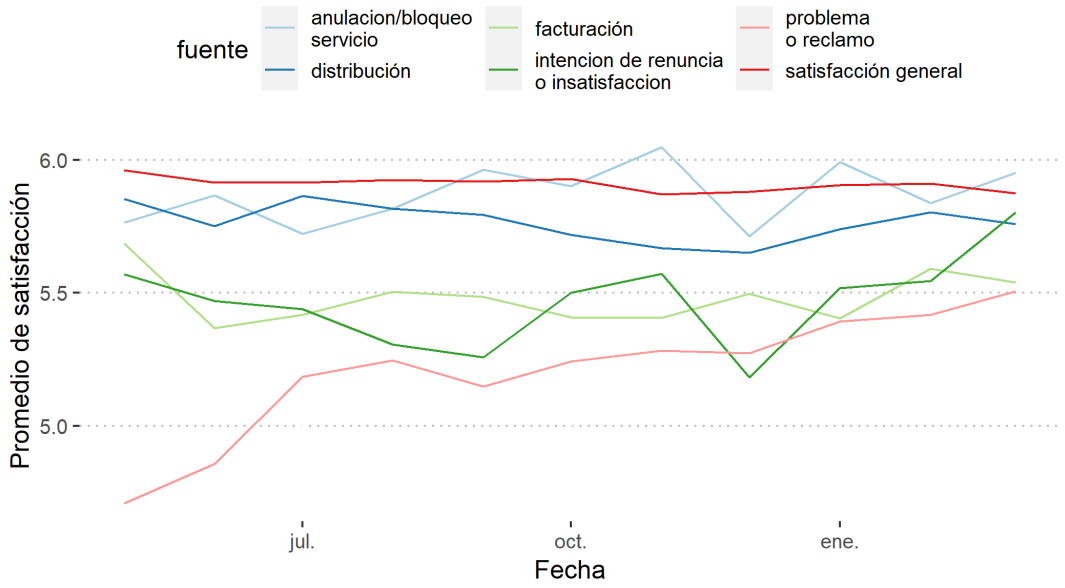

Figura 5.15: Satisfacción mensual pospago según tipo de interacción.

Los clientes que demuestran intenciones de renuncia o expresan su insatisfacción con la firma junto con los clientes que tienen problemas de facturación, tienen una satisfacción más baja, lo cual tiene sentido, pero que se ha mantenido de esta forma a lo largo del tiempo, por lo que se deduce que este tipo de situaciones aún requieren de un esfuerzo adicional por parte de la firma para poder mejorar en estos ámbitos, en especial en problemas relacionados a la facturación de clientes.

Al comparar los resultados de este gráfico con los elaborados en las figuras 5.7 y 5.11, se espera que las variables de esta última fuente de información tengan una importancia mayor en la tarea de predecir la satisfacción de clientes. Aún así, estos gráficos solo muestran los promedios de satisfacción con la información agregada, por lo que a partir de estos gráficos no se puede aseverar que información tendrá mayor importancia.

5.3. Elección de ventana de tiempo y criterio de satisfacción

Con la elaboración de los atributos descritos en las sección anterior, se procede a determinar cual de las 3 ventanas de tiempo estudiadas permite clasificar la satisfacción de los clientes de mejor forma. El efecto de utilizar una ventana de tiempo más larga es que aumenta la cantidad de clientes con información dentro de ese atributo, pero se tiene la hipótesis de que las interacciones más cercanas a la fecha de la encuesta tienen un efecto mayor en la satisfacción que aquellas realizadas en un plazo mayor.

Para tomar la decisión de que ventana de tiempo conservar para las secciones posteriores se utilizan regresiones logísticas. A través de los resultados de las regresiones y el criterio AIC se decide que ventana de tiempo conservar. Junto con esto, es necesario también determinar

de manera previa la variable dependiente que se utilizará para determinar la satisfacción de un cliente.

Partiendo por esto último, es necesario destacar las similitudes y diferencias de la escala de medición utilizada en la encuesta estudiada contra una escala likert. Una escala likert utiliza usualmente 5 a 7 categorías que buscan medir las preferencias o nivel de aceptación con algún enunciado, con opciones en un rango desde *muy satisfecho*, hasta *muy insatisfecho* y usualmente incorporando una opción intermedia *indiferente* [3]. Normalmente se suele asociar números a las respuestas y en una escala de 7 alternativas, donde el 4 correspondería a la opción de indiferencia.

Pero en el caso de estas encuestas se evalúa con una "nota del 1 al 7" buscando asemejarse a la escala de notas que se utiliza en las evaluaciones escolares del país. Debido a esta diferencia, no es inmediato identificar la opción "intermedia", pudiendo esta corresponder a la nota 4 o 5.

Estudios internos de análisis de sentimiento realizados por la firma concluyen que una nota 4 corresponde a una evaluación negativa mientras que el 5 a una alternativa mixta. Por otra parte, la consultora presenta los resultados de sus encuestas a la firma considerando las respuestas bajo esta misma clasificación. Dado que esta es la forma en que se trabaja los resultados de satisfacción dentro de la firma y por la consultora, se decide trabajar bajo este mismo esquema.

Una de las similitudes que comparten la escala likert con la evaluación con notas, es que ambos resultados son ordinales pues las respuestas tienen un orden intrínseco que permite diferenciar la escala en que un cliente esta satisfecho o insatisfecho. Si bien se puede concluir que una nota 1 es peor que un 2, desde un punto de vista estadístico no se puede aseverar cuanto más una nota es peor o mejor que la otra. Esto ocurre por que a pesar de que las respuestas son ordinales, estas no son de intervalos fijos, por lo que la distancia entre notas adyacentes no es clara.

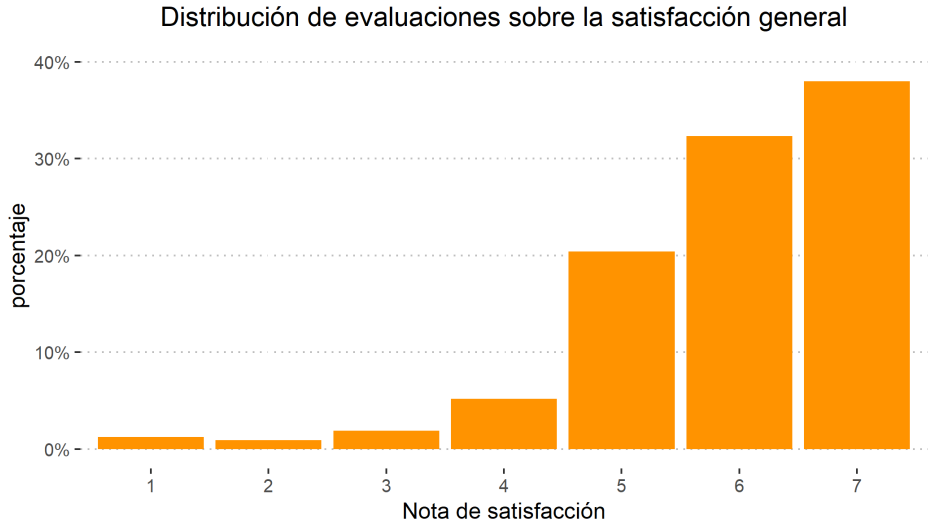

Figura 5.16: Distribución de notas puestas por los encuestados a la satisfacción general.

La distribución final de respuestas obtenida a la pregunta de satisfacción general se puede ver en la figura 5.16. Alrededor del 70 % de los encuestados evalúa a la firma con notas 6 o 7, lo que aumenta a alrededor de un 90 % al considerar también la nota 5. En consecuencia se tiene que un 10 % de los clientes declaró tener una evaluación negativa de la firma (1 a 4).

Dada la estructura de las respuestas, la tarea de predecir satisfacción puede ser realizada utilizando varios métodos. Una de las formas de abarcar este problema es realizando modelos del estilo *one v/s all* donde se predice la probabilidad de que un cliente tenga una evaluación igual a x con $x \in [1, ..., 7]$, lo que significaría realizar 7 modelos diferentes. Otras técnicas agrupan las notas puestas por los clientes para reducir la complejidad de la tarea, usualmente considerando solo 2 o 3 categorías diferenciando a los clientes como satisfechos, insatisfechos o neutros.

A partir de la discusión realizada, se descarta utilizar el primer método debido a la dificultad existente en determinar si notas adyacentes presentan diferencias significativas que puedan ser luego captadas por cada uno de estos modelos. El objetivo del trabajo esta en identificar la dirección de la satisfacción (cliente satisfecho o insatisfecho) mas que en la magnitud específica de este valor, en consecuencia, agrupar las evaluaciones es el mejor método para la tarea planteada.

Desde un punto de vista de negocios, si se quiere, por ejemplo, realizar una estrategia de marketing enfocada a clientes insatisfechos, un cliente tiene dos opciones, o estar considerado dentro de esta campaña o no. El tener una categoría de neutros implicaría una decisión extra que tomar sobre si incorporar a este grupo de clientes dentro o fuera de esta campaña. Casi cualquier otro tipo de gestión va a requerir del tomar este misma decisión de forma previa. En consecuencia se decide agrupar a los clientes solo como satisfechos e insatisfechos.

En consecuencia, se clasifican como insatisfechos las respuestas con notas de 1 a 4 y como satisfechos a quienes contestas 6 o 7. En cuanto a los clientes que responden con nota 5, se definen como pertenecientes al grupo de clientes satisfechos. Si bien las respuestas de estos clientes están consideradas como mixtas, estos son clientes que a pesar de tener motivos para no dar una calificación mayor a los servicios de la firma, no tienen una evaluación estrictamente negativa de la firma, a diferencia de quienes responden con notas de 1 a 4, y la firma tiene mayor interés en diferenciar a este grupo del resto.

Ya con la variable dependiente definida, queda elegir la mejor ventana de tiempo para la información elaborada. En la tabla 5.2 se pueden ver tres regresiones logísticas, utilizando cada una el mismo set de variables pero cambiando la ventana de tiempo utilizada. A modo de control se agregan el sexo, edad, antigüedad del cliente en meses y segmento comercial al que pertenece el encuestado, el cual se asigna en función de su cargo fijo, *básico, bajo, medio, alto y full*, en forma ascendente de valor.

La variable dependiente *Cliente satisfecho* corresponde en este caso a una binaria que toma el valor de 1 si el cliente evaluó la satisfacción general de la compañía con una nota mayor o igual a 5 y 0 en el resto de casos. Comenzando por las variables de control, se puede ver como la edad no muestra un efecto significativo en ninguno de los 3 casos estudiados, mientras que la antigüedad del cliente tiene un efecto significativo y positivo en el logaritmo de la probabilidad de ser un cliente satisfecho, el cual es bastante leve, pero dado que se

utilizó la antigüedad en meses, este toma un mayor valor para los clientes con varios años de antigüedad.

Destaca el efecto positivo y significativo en la satisfacción que tiene el ser mujer, lo que está diciendo que los hombres parecen tener un nivel de exigencia mayor en cuanto a la calidad del servicio esperado. En cuanto a los resultados asociados al segmento de valor del cliente, los resultados se comparan con la base de ser un cliente dentro de la categoría *alto*. Se puede ver como no hay una diferencia significativa en la satisfacción entre los clientes *full* con el caso base, mientras que si existe una diferencia entre los clientes de valor *alto* con los clientes *básico, bajo o medio*, categorías donde los clientes tienen mayor probabilidad de tener una satisfacción más alta.

Tabla 5.2: Regresiones logísticas variando la ventana de tiempo

	Dependent variable:		
	Cliente satisfecho		
	1 mes	2 meses	3 meses
	(1)	(2)	(3)
edad	−0.001 (0.001)	−0.001 (0.001)	−0.001 (0.001)
sexo.femenino	0.252*** (0.043)	0.249*** (0.043)	0.247*** (0.043)
Antiguedad cliente	0.001** (0.001)	0.001** (0.001)	0.001** (0.001)
Segmento.Bajo	0.269*** (0.091)	0.275*** (0.092)	0.267*** (0.092)
Segmento.Básico	0.193** (0.087)	0.195** (0.088)	0.186** (0.088)
Segmento.Full	−0.177 (0.164)	−0.153 (0.165)	−0.153 (0.165)
Segmento.Medio	0.243*** (0.091)	0.252*** (0.092)	0.247*** (0.092)
CALL CENTER	−0.058*** (0.016)	−0.055*** (0.012)	−0.047*** (0.010)
IVR	−0.055*** (0.019)	−0.029** (0.013)	−0.026*** (0.010)
USSD	−0.021*** (0.007)	−0.017*** (0.004)	−0.012*** (0.003)
WEB	0.022 (0.015)	0.011 (0.008)	0.009 (0.006)
CHAT	0.002 (0.018)	0.021 (0.017)	0.027* (0.016)
problema o reclamo	−0.334*** (0.066)	−0.320*** (0.047)	−0.251*** (0.039)
distribución	−0.015 (0.019)	−0.003 (0.012)	−0.004 (0.009)
intención renuncia/insatis.	−0.120*** (0.044)	−0.102*** (0.036)	−0.107*** (0.033)
anulación/bloqueo de serv.	0.043 (0.121)	0.023 (0.082)	0.005 (0.064)
facturación	−0.379*** (0.091)	−0.160** (0.066)	−0.099* (0.056)
visitas tienda propia	0.181* (0.097)	0.123** (0.062)	0.076 (0.047)
visitas franquicias	−0.080* (0.046)	−0.059* (0.034)	−0.025 (0.029)
Constant	2.002*** (0.101)	2.026*** (0.102)	2.039*** (0.102)
Observations	25,389	25,389	25,389
Log Likelihood	−7,947.300	−7,917.200	−7,916.900
Akaike Inf. Crit.	15,935.000	15,874.000	15,874.000
Note:			*p<0.1; **p<0.05; ***p<0.01

Estos resultados confirman conocimientos previos dentro del área de CRM sobre la relación precio-expectativa de los clientes donde los clientes que pagan una cantidad mayor por un servicio tienden a tener exigencias mayores sobre el desempeño del servicio. Tomando esta misma relación, se debería tener que mientras menos se paga las exigencias deberían ser menores y en consecuencia los clientes del segmento básico (menor cobro fijo) deberían ser los que tengan la mayor diferencia con los clientes del segmento alto, lo cual no ocurre en este caso y son los clientes del segmento bajo quienes tienen el mayor coeficiente.

Este hecho se repite a lo largo de las 3 regresiones, y muestra que si bien la relación precio-expectativa se tiene a grandes rasgos, esta no parece ser una relación estrictamente lineal, lo que también se aprecia al ver que en ninguno de los 3 casos el segmento full, donde el cargo fijo es mayor, tiene diferencias significativas con el segmento alto.

Pasando a las variables sobre interacciones y visitas a tiendas, se puede ver que todas las variables que resultaron significativas tienen un efecto negativo en la satisfacción excepto las visitas a tiendas propias de la firma, aunque esto se obtienen solo para las regresiones con ventanas de tiempo de 1 y dos meses de historia. Al usar 3 meses de información, el efecto de una visita a tiendas deja de ser significativo para las visitas a cualquiera de los dos casos.

Las interacciones realizadas en los canales tanto del call center, como IVR y USSD muestran un efecto negativo en la probabilidad de ser un cliente satisfecho, siendo las interacciones por call center las que tienen un mayor efecto negativo por cada interacción extra realizada por el cliente, mientras que las interacciones por USSD las que tienen un efecto menor. Las interacciones por la web y chat no tienen un efecto significativo en la satisfacción salvo por las interacciones a través de chat al utilizar 3 meses de historia.

Resulta interesante destacar que los clientes que utilizan o tienen interacciones por los canales más antiguos, call center, IVR y USSD, tengan una mayor probabilidad de ser clientes insatisfechos, mientras este no es el caso para los clientes que interaccionan por el web y chat, resultado que también se repite a lo largo de las 3 ventanas de tiempo.

Una posible interpretación es que la interacción que ocurre en estos 3 canales entre el cliente y la firma tengan algún factor en común que provoque estos resultados. Estos resultados muestran que puede ser favorable para la firma el estudiar estos 3 canales en busca de cual puede ser el factor que disminuye la satisfacción del usuario.

Continuando con el resto de variables, los mayores efectos en la satisfacción lo tienen los clientes que realizan problemas o reclamos, quienes dan a conocer su insatisfacción e intención de renuncia o tienen problemas asociados a la facturación del servicio. Estos resultados eran los esperados en torno a estas variables y dan cuenta de que la información utilizada es fidedigna. Estos resultados muestran que existe un área clara donde la firma puede mejorar sus servicios, el cual corresponde a los servicios de reclamos.

En general, se puede ver como los valores asociados a cada variable significativa de interacciones y visitas a tiendas, primero, mantienen su signo a través de los tres modelos y, segundo, son mayores al utilizar un mes de historia y descienden en magnitud con cada nuevo mes de información. Estos resultados confirman la hipótesis planteada anteriormente de que los efectos de interacciones entre el cliente y la firma son mayores conforme ocurren mas cerca

del momento en el que el cliente fue encuestado.

Al comparar el valor del criterio *akaike* obtenido por los 3 modelos, se puede ver como los modelos que utilizan 2 y 3 meses de historia obtienen el mismo puntaje y que este es menor al valor obtenido por el modelo con 1 mes de historia, por lo que aumentar la ventana de tiempo mejora la capacidad que tiene el modelo de explicar la variabilidad de los datos, pero que esta mejora se tiene principalmente al aumentar de 1 mes a dos meses, luego no se percibe diferencia al pasar de 2 a 3 meses.

Como última observación del modelo, el valor de la constante es positiva y significativa en los 3 modelos. Ya que esta variable de cuenta del efecto promedio de cualquier información por la que no se esta controlando, esto quiere decir que el set de variables utilizado no es suficiente para explicar completamente la variable dependiente, por lo cual se requiere controlar por otros factores, donde buenos candidatos parecen ser factores como la señal y ubicación geográfica, las cuales se consideran dentro de los modelos de caja negra en la siguiente sección.

Al mirar el criterio *akaike* los mejores resultados se tienen al ocupar 2 o 3 meses ventana de tiempo. Si nos fijamos en los valores de los coeficientes asociados a las variables significativas, los efectos son mayores al utilizar solo 1 mes de información y menores al utilizar 3 meses. Como consecuencia no hay una ventana de tiempo que sea estrictamente mejor a partir de estos 2 criterios, pero finalmente se decide utilizar la ventana de **3 meses de historia**.

Dado el objetivo trabajo, se le da un valor mayor a que las variables puedan diferenciar a una mayor cantidad de clientes por sobre la magnitud del efecto de cada variable. Por ejemplo, los clientes encuestados que presentaron algún reclamo o problema considerando 1 mes de historia, fueron solo un 2 % del total de clientes, porcentaje que aumenta a un 5 % al utilizar 3 meses de historia, lo cual significa que en el último caso se tienen a más del doble de clientes que se pueden diferenciar según los reclamos realizados. Este efecto de aumento de porcentajes al usar 3 meses de historia se repite para todas las variables elaboradas en sus respectivas magnitudes, motivo principal para decidir utilizar 3 meses de historia.

Los resultados de la tabla 5.2 permitieron tener una primera evaluación entre la relación existente entre las variables desarrolladas y la satisfacción, menos para las variables sobre el uso de la aplicación, que se dejaron fuera al no tener diferentes ventanas de tiempo.

Con motivo de entregar una primer insight también de la existencia o no de relaciones entre la satisfacción de los clientes de la empresa y el uso de la aplicación, se realiza una regresión logística con estos datos, la cual se encuentra en la tabla 5.3. Se utilizan las mismas variables de control que en el caso anterior.

Tabla 5.3: Resultados regresión logística variables APP.

	Dependent variable:
	mas_promotores
edad	−0.001 (0.005)
Sexo.femenino	0.279** (0.132)
Antiguedad cliente	0.004** (0.002)
Segmento.Bajo	0.011 (0.298)
Segmento.basico	0.216 (0.297)
Segmento.Full	−0.620 (0.554)
Segmento.Medio	0.099 (0.302)
ofertas_app	0.004 (0.023)
chat_app	−0.321*** (0.098)
entretenimiento_app	0.042 (0.050)
beneficios_app	0.029 (0.040)
soporte_app	−0.105 (0.123)
plan/equipo_boleta/consumo_app	0.008 (0.014)
Constant	2.216*** (0.336)
Observations	3,407
Log Likelihood	−905.350
Akaike Inf. Crit.	1,838.700
Note:	*p<0.1; **p<0.05; ***p<0.01

Como solo se tiene información entre diciembre 2019 y marzo 2020, se acotaron las encuestas para solo utilizar estos meses. De las variables de control nuevamente se mantiene la significancia del sexo femenino y la antigüedad del cliente, pero con este set de variables y solo 4 meses de información no se identifica diferencia en la satisfacción a partir del segmento al que pertenece el cliente.

De todas las variables elaboradas en la sección sobre la aplicación, tan solo la cantidad de veces que un cliente accede al servicio del chat tiene un efecto significativo en la satisfacción, donde el incremento en el logaritmo de la probabilidad de ser un cliente insatisfecho por cada acceso extra realizado por el cliente es de -0.321 y con una significancia del 99 %.

Los resultados muestran que el resto de servicios no guardan una relación significativa con la satisfacción del cliente, ya sea este al acceder a los beneficios que se ofrecen por la app, revisar el estado del plan, equipo, boleta o consumo realizado, ofertas o servicios de entretenimiento.

Esto quiere decir que el chat al cual se puede acceder a través de la aplicación de la firma esta funcionando como un canal por el cual los clientes intentan resolver los problemas que se les presentan, el cual es el objetivo de este servicio. Por otra parte, se esperaba que los servicios de entretenimiento o beneficios tuvieran relación significativa y positiva con la

probabilidad de ser un cliente satisfecho con la firma, pero los resultados muestran que no hay un efecto significativo.

Los resultados muestran que probablemente el aporte de la información del uso de la aplicación en la predicción de satisfacción en los modelos posteriores será bajo, exceptuando por el servicio de chat. La siguiente sección busca determinar la capacidad de los datos utilizado en la predicción de satisfacción al utilizar algoritmos de caja negra.

5.4. Modelos de satisfacción de caja negra

Con la ventana de tiempo escogida y las variables elaboradas, se procede al entrenamiento de modelos de caja negra. Para esta tarea se utilizan los algoritmos de *random forest, gradient boosting machine y extreme gradient boosting*. Se ha demostrado ya en otros trabajos que estos modelos pueden entregar buenos resultados para la tarea de predicción de fuga de clientes en el mercado de telecomunicaciones [23, 15, 21], pero no se ha visto un intento de aplicar este tipo de modelos específicamente para la predicción de satisfacción. Este trabajo permitirá evaluar el desempeño que estos modelos tienen en este tipo de análisis.

Un problema con ocurrencia frecuente dentro del ámbito de machine learning es el problema de "desbalance de clases". Este ocurre cuando la distribución de las categorías de la variable dependiente es desigual, usualmente con una sola clase prevaleciente sobre las demás [18]. Para el caso de una variable dependiente binaria, casos de desbalanceo extremo se pueden observar cuando la clase de interés ocurre en un 1 % de los casos o menos. Usualmente en estos casos el objetivo principal es poder identificar esta clase minoritaria.

Esta diferencia en la frecuencia de cada clase produce que gran parte de los algoritmos de aprendizaje, al ser aplicados a problemas de este tipo, identifiquen a la totalidad de observaciones como pertenecientes a la clase mayoritaria. La razón de este suceso radica en uno de los supuestos que normalmente realizan estos algoritmos, que es el de maximizar el accuracy del modelo, supuesto bajo el cual clasificar todos los datos como pertenecientes a la clase mayoritaria es normalmente la mejor solución.

El supuesto de maximizar accuracy esta basado en otro supuesto, que es el de que el costo de cometer alguno de los dos tipos de errores posibles (falso negativo o falso positivo) es idéntico. Cuando este supuesto no se cumple se cae en el problema de desbalanceo de clases, donde es usual también que el costo de clasificar erróneamente una observación de la clase minoritaria sea mayor al de clasificar erróneamente una observación de la clase mayoritaria.

Si recordamos la distribución del problema a entrenar, tan solo un 9 % de los encuestados evaluó a la firma con una nota de 1 a 4, por lo cual, si bien no se esta en un caso extremo de desbalance de clases, la diferencia es lo suficientemente grande para caer en este problema. En cuanto al costo de clasificación errónea, se tiene que para este problema el costo de ambos errores es diferente.

Ya es conocido que la satisfacción influye en aspectos como la fuga de clientes, por lo que identificar erróneamente a un cliente satisfecho como insatisfecho no tiene el mismo costo que

en el otro sentido. La firma realiza de forma seguida diversas estrategias comerciales para retener a sus clientes, donde se ofrecen mejores ofertas a clientes con mayor propensión a fugarse.

En este contexto, clasificar equivocadamente a un cliente insatisfecho como satisfecho, significaría dejarlo fuera de estos esfuerzos comerciales, lo cual provocaría un aumento en sus probabilidades de fugarse de la compañía (tomando como supuesto que el esfuerzo comercial produce una disminución de la probabilidad de ocurrencia de la fuga). Para el caso contrario, esto significaría que un cliente satisfecho estaría incorporado a una oferta comercial extra, pero dado que este cliente esta satisfecho con la compañía, la reducción de la probabilidad de fuga producto de la oferta es marginal ya que esta ya era más baja de manera previa a la oferta.

Para lidiar con este problema en el entrenamiento de modelos se aplican 2 estrategias diferentes, balanceo de clases y cambio del parámetro a optimizar por los modelos. La primera estrategia consiste en generar nuevas observaciones de la clase minoritaria a partir de la información disponible o eliminar observaciones de la clase mayoritaria, buscando conseguir una proporción similar entre las dos categorías.

Lo anterior es realizado a través de un random oversampling y random subsampling, técnicas que eligen observaciones al azar para luego duplicarlas (oversampling) o eliminarlas (undersampling).La técnica de random sampling permite entrenar manteniendo el accuracy como parámetro a optimizar.

La segunda estrategia consiste en optimizar una métrica diferente al accuracy pero que incorpore diferencias en el los tipos de error. La métrica utilizada para optimizar es el área bajo la curva precision-recall, AUCPR (area under curve precisión recall), la cual ha sido vista como un mejor parámetro a utilizar en la optimización de problemas con desbalance de clases [19, 8]. Esto es conseguido por esta curva ya que las métricas de precision y recall se olvidan de la tasa de verdaderos negativos, concentrándose en las métricas relacionadas a la predicción de la clase positiva[2]. Para poder ajustar los datos utilizados a esto, en los modelos presentados más adelante la clase positiva corresponde a los clientes insatisfechos, mientras que la clase negativa representa a los clientes satisfechos.

Para buscar mejorar los resultados de los modelos, se aplica la técnica de grid search, que consiste en entrenar diferentes combinaciones de parámetros dentro de los modelos, en busca de obtener mejores resultados y reducir la sobre representación. Este se aplica a los siguientes parámetros de los modelos presentados:

- Cantidad de árboles entrenados.
- Profundidad máxima de los arboles.
- Cantidad mínima de observaciones en el nodo terminal de un árbol.
- Balance de clases, diferentes ratios de oversampling y subsampling.
- Parámetro a optimizar, accuracy y AUCPR.

[2] Este método asume, sin perdida de generalidad, que la clase minoritaria corresponde a la clase positiva.

Para validar los resultados obtenidos por el entrenamiento de modelos, la data es dividida en datos de entrenamiento y datos de testeo. Para testear se utilizan las encuestas del último mes disponible (marzo) y se entrena con todos los meses anteriores disponibles. Este método es el preferido frente a una selección al azar de encuestados de todos los meses para realizar el testeo, pues esta es la forma en que se utilizan los resultados de modelos dentro de la firma, donde los resultados son utilizados para gestionar al mes siguiente.

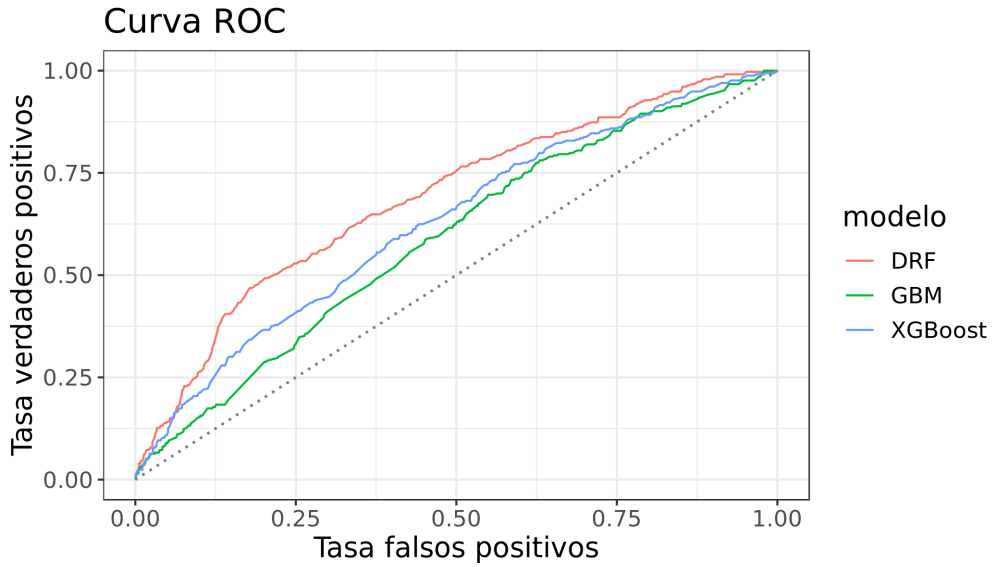

Figura 5.17: Curvas ROC obtenidas por los distintos modelos en el set de testeo.

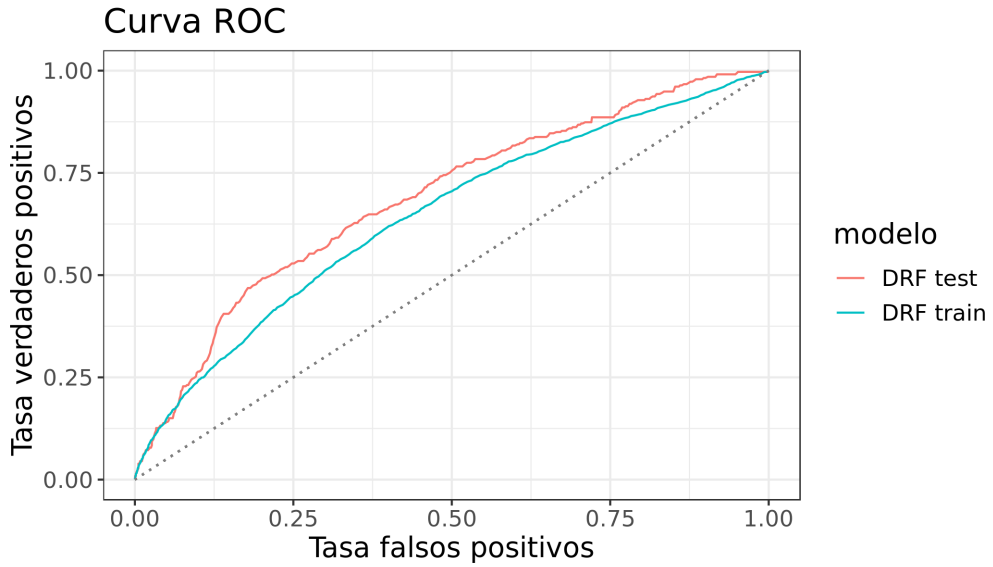

Figura 5.18: Curvas ROC testeo y entrenamiento modelo random forest.

En la figura 5.17 se puede ver el desempeño obtenido por los diferentes modelos luego de la aplicación del grid search, en la predicción de la base de testeo. Los mejores resultados se

obtuvieron con el algoritmo *random forest*. Este algoritmo obtuvo un área bajo la curva de 0,68 en el testeo y 0,64 en el entrenamiento, lo que se puede ver en la figura 5.18. El desempeño de los otros dos modelos en la base de entrenamiento fue similar, con scores alrededor del 0,65, pero los algoritmos *GBM* y *XGBOOST* cayeron en el problema de *overfitting*, donde hay una diferencia notoria entre el desempeño del modelo en la base de testeo contra la base de entrenamiento.

Una observación interesante es que los mejores resultados se obtuvieron al optimizar la métrica AUCPR por sobre el accuracy y sin aplicar balanceo de clases. Los siguientes gráficos muestran las curvas precisión recall obtenidas por los modelos.

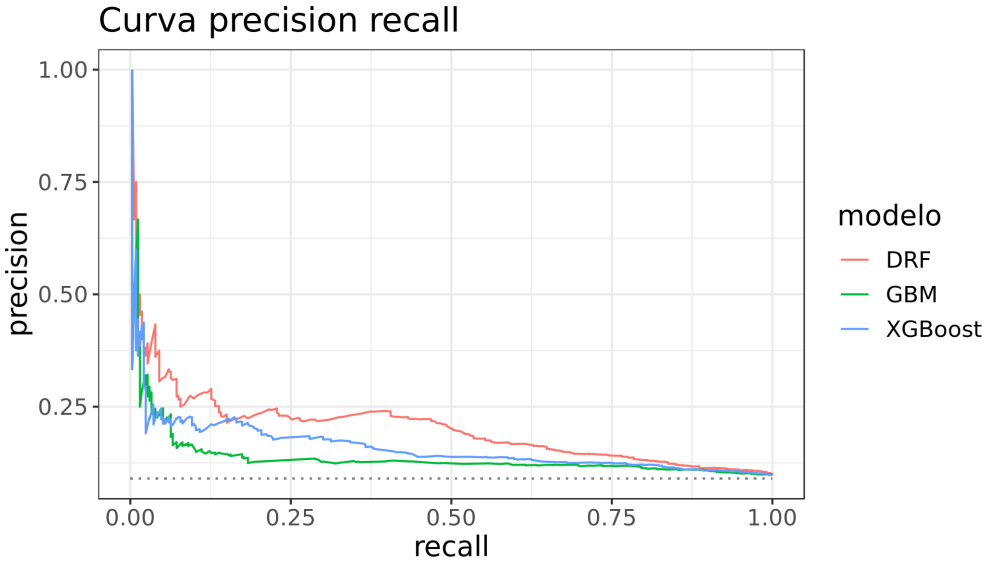

Figura 5.19: Curvas precision-recall obtenidas por los modelos en el set de testeo.

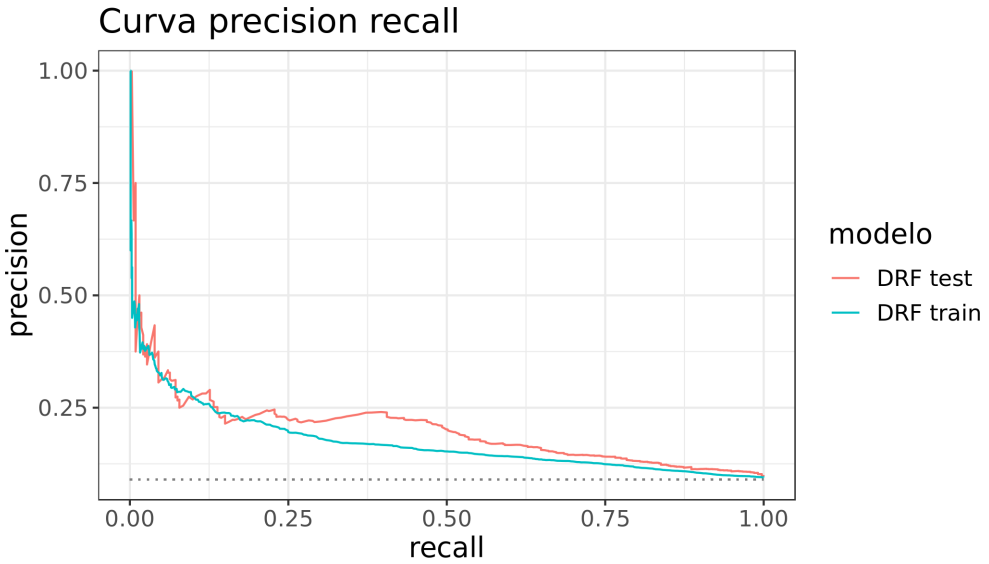

Figura 5.20: Curvas precision-recall testeo y entrenamiento modelo random forest.

A diferencia de la curva ROC, los gráficos AUCPR no tienen una línea base universal contra el que compararse, si no que esta depende de la distribución específica del problema que se está resolviendo. Un clasificador random tiene precisión igual a la distribución de positivos del problema, lo que en este caso corresponde a un 9 % aproximadamente, esto corresponde a la línea punteada gris en las figuras 5.19 y 5.20, donde la primera figura muestra el desempeño en esta curva para los distintos modelos y la comparación del testeo con el entrenamiento para el algoritmo random forest en la segunda figura.

Las conclusiones dentro de estos gráficos son similares a los de la curva ROC, nuevamente el modelo que diferencia de mejor manera a los clientes insatisfechos por sobre los satisfechos es el algoritmo random forest y sus resultados tanto en testeo como entrenamiento son similares. Lamentablemente las curvas AUCPR no tienen una interpretación significativa de su área más allá de poder identificar modelos con mejor desempeño en cuanto a su precisión al moverse en el espectro de la recall [10].

El mejor modelo obtenido consigue una precisión más de 2 veces mayor al de un clasificador random en la primera mitad de los valores de recall, que luego desciende linealmente hasta llegar al valor obtenido por un clasificador que siempre asigna el valor positivo a las observaciones (símil al punto 1,1 en la curva ROC). Estos resultados muestran que los modelos logran identificar clientes insatisfechos de mejor manera que un clasificador aleatorio, pero que en los 3 casos, aún existe posibilidad de mejorar los resultados para conseguir una diferenciación aún mejor.

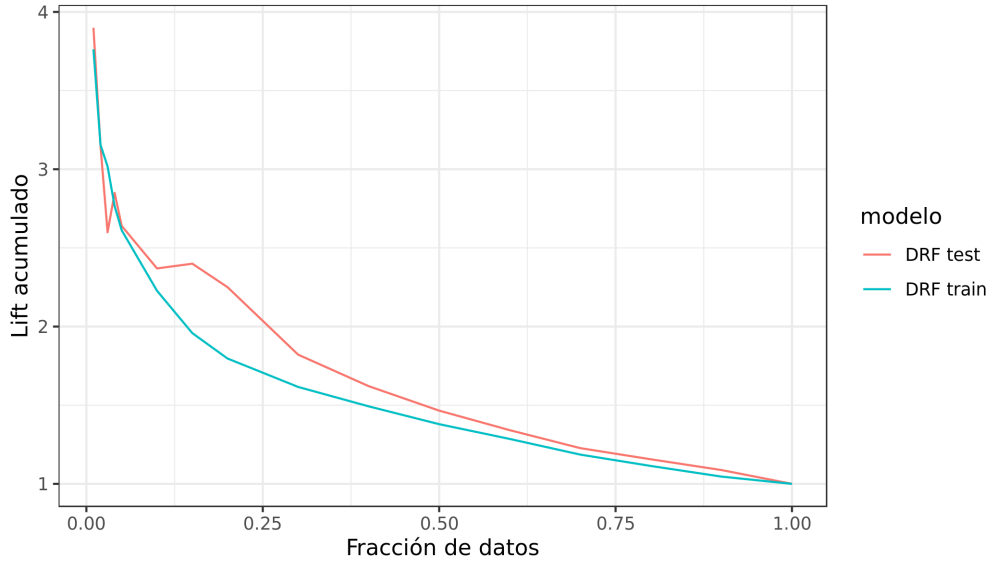

Figura 5.21: Lifts obtenidos por modelo random forest en sets de testeo y entrenamiento.

Los resultados de la curva figura 5.20 muestran que el modelo logra encontrar criterios donde se tiene mayor propensión de ser cliente insatisfecho. Una forma en que también se evalúa esto último es a través de las curva lift, la cual se muestra para el modelo random forest en la figura 5.21. Un lift obtenido aproximadamente de 4 que se mantiene tanto en el testeo como el entrenamiento, que se mantiene superior a 2 para el 25 % de la base más

propensa a ser insatisfecho en el set de testeo, o sea que para un cuarto de los datos se tiene una tasa 2 veces más alta de insatisfacción, lo que sería alrededor del 18 %, cercano a 1 de cada 5 clientes .

Revisando todos los resultados, el accuracy obtenido por los diferentes modelos no supera el valor de 0,7 en ninguno de los casos, lo que en primera instancia no es un resultado alentador. Sin embargo, los resultados en cuanto al lift y la curva AUCPR nos dicen que esta pérdida de accuracy en los modelos es a costa de obtener una mayor capacidad de predecir a los clientes insatisfechos. Un modelo que siempre predice a los clientes como satisfechos, obtendría un accuracy aproximado de 91 % en este problema, por lo que esta métrica no se ve como la mejor de las 3 para decidir sobre la utilidad de los modelos y se favorece el lift y AUCPR.

Dentro de estos, el lift nos muestra un aumento sustancial en la probabilidad de ser un cliente insatisfecho para ciertos grupos de clientes, mientras que el gráfico AUCPR muestra de manera gráfica que el modelo logra diferenciar mejor a clientes insatisfechos que una selección aleatoria. Para seguir entendiendo la calidad de resultados se revisa las variables que tuvieron mayor importancia en el modelo random forest para la predicción de satisfacción.

Variable	Importancia escalada
Región	1
Problemas o reclamos 3 meses	0.82
Call center 3 meses	0.65
Tiempo ida y vuelta	0.58
Edad	0.46
Resta facturación – cargo fijo	0.44
GSE	0.40
Antigüedad cliente	0.38
Meses uso último equipo	0.36
Fuerza de la señal	0.35

Figura 5.22: Top 10 variables más importantes del modelo random forest.

La tabla 5.22 muestra las 10 variables más importantes según el modelo random forest. Como se puede ver, dentro de las variables elaboradas particularmente para este trabajo, se encuentran dentro de este top dos variables, la cantidad de problemas o reclamos del cliente (segundo puesto) y las interacciones realizadas a través del call center (tercer puesto), luego

de esto le siguen fuera del top los problemas por facturación en el puesto 15 y la cantidad de interacciones en los canales USSD junto a IVR en los puestos 17 y 18 respectivamente. Del resto de variables desarrolladas, las relacionadas con la app de la firma son las que tuvieron la menor importancia entre el conjunto completo de variables, ocupando los últimos puestos en la importancia escalada.

Los resultados muestran como se utiliza distinto tipo de información de los clientes para la elección del modelo. Habiendo presentes en el top de variables información socio demográfica (región, GSE, edad), de mercado (antigüedad del cliente, meses de uso del último equipo), de señal (tiempo ida y vuelta, fuerza de la señal) e interacciones (problemas o reclamos y llamadas del call center). Estos resultados sugieren que la satisfacción es algo complejo de predecir y que no esta relacionado a un único factor dentro del servicio entregado por la firma, si no es que una suma de distintos factores tanto provenientes del servicio entregado como de las características del cliente en si.

De estas variables, destaca el hecho de que la región sea la variable con mayor importancia dentro de este modelo, lo que en primera instancia indica que la región del cliente tiene la mayor importancia a la hora de determinar la satisfacción de un cliente entre todas las variables utilizadas dentro del modelo.

Los servicios entregados por la firma difieren mucho de región a región, tanto en la cobertura de señal, como en la cantidad de tiendas para atender a clientes y la disponibilidad o no de otros servicios como internet hogar. La principal hipótesis que se tiene para que la región aparezca como primera variable es el que este dato este dando cuenta de estas diferencias que se presentan dependiendo del lugar donde reside el cliente (y probablemente de más diferencias que solo las mencionadas), información que no está incluida en el set de variables utilizado.

Junto con lo anterior, también se tiene que 2 de las 3 variables incorporadas con respecto a la señal están dentro de las más importantes, el tiempo de ida y vuelta promedio que tiene el enviar una solicitud por el teléfono a la aplicación de Facebook y la fuerza promedio de la señal recibida.

Esto confirma conocimientos que ya se poseen sobre la importancia de la señal en la satisfacción del cliente dentro de la industria. Los resultados obtenidos muestran también que el efecto de la señal sería mayor a los percibidos por el cliente mediante interacciones con la firma por los distintos canales, exceptuando el call center, siendo este el único canal que supera a la señal en importancia para la predicción de satisfacción del modelo.

Dentro del resto de variable se encuentra la información socio demográfica (edad y GSE) y la antigüedad del cliente. La aparición del GSE se relaciona con la satisfacción de la misma forma en que lo hace el valor comercial del cliente, un mayor nivel socio económico se relaciona con un mayor valor de plan esperado, lo que aumenta la exigencia del cliente por el servicio.

Destaca también la variable “resta facturacion - cargo fijo” que, como su nombre lo dice, da cuenta de la diferencia entre la última facturación realizada por el cliente y su cargo fijo dentro de ese mes. Es probable que esta variable este relacionada con la desarrollada anteriormente que da cuenta de los clientes con problemas de facturación. Esta variable muestra que es

necesario indagar más en esta información y el por que es importante en la satisfacción, si se relaciona con problemas de facturación o no, o si estar pagando mensualmente montos superiores o inferiores al cargo fijo esta impactando de manera significativa a la satisfacción.

Los resultados obtenidos muestran insights importantes sobre que tipo de información es relevante para la satisfacción, las que corresponden a variables relacionadas con la señal, interacciones, aspectos socio demográficos y información que de cuenta de problemas de cualquier índole y relacionados con la facturación de clientes, junto con información que de cuenta de los cambios de equipo realizados por los clientes y la calidad de estos.

De la información elaborada para este trabajo, tanto la información relacionada con las visitas a tiendas y la de interacciones dentro de la aplicación de la firma demostraron tener un peso bajo en la predicción de satisfacción del cliente, por este motivo se considera que su inclusión en los modelos puede ser despreciada, sin afectar los resultados obtenidos de manera significativa. Otra opción posible a futuro consiste en cambiar la ventana de tiempo para las visitas a tiendas que se consideran dentro del modelo.

Como última evaluación, se revisa en la figura 5.23 la matriz de confusión que se obtiene al asignar según los resultados del modelo a los clientes como satisfechos e insatisfechos.

		Valor Predicho	
		Satisfecho	Insatisfecho
Valor Real	Satisfecho	7598	392
	Insatisfecho	699	138

Figura 5.23: Matriz de confusión modelo random forest en set de testeo.

Dado el umbral que utiliza el modelo para clasificar a un cliente como satisfecho o insatisfecho, se puede ver que este clasifica correctamente a 138 clientes insatisfechos del total, que corresponde a 837. Esto significa que se clasifica correctamente a aproximadamente un 16 % de los clientes insatisfechos, por lo que queda un gran porcentaje de clientes insatisfechos incorrectamente clasificados.

Los clientes que fueron clasificados como insatisfechos corresponden a 530, lo cual representa un 6 % del total de clientes en la base de testeo. Dentro de este 6 % de clientes, un 26 % corresponden a clientes insatisfechos, lo que significa que dentro de este grupo se obtuvo una

mejora de 17 puntos porcentuales por sobre lo que un clasificador aleatorio hubiera obtenido en esperanza en un grupo de igual tamaño (este hubiera obtenido en esperanza solo un 9 %).

A partir de estos resultados se concluye lo siguiente:

- Dentro del 6 % de clientes clasificados como insatisfechos, se puede ver una mejora significativa en identificación de clientes con insatisfacción. El principal beneficio que se tiene es que al ser un porcentaje bajo de clientes, estos podrían ser gestionados con ofertas comerciales más fuertes que al resto de clientes, abriendo la posibilidad de una mejora en la gestión de clientes actual de la firma.
- Si bien de los resultados se identifica un grupo de clientes donde el porcentaje de insatisfacción es bastante más alto al porcentaje base, si solo se gestiona a este grupo, se estaría dejando afuera de una mejor gestión a un gran porcentaje de clientes insatisfechos. Si la firma quisiera gestionar a los clientes seleccionados por este modelo, deberá acompañar esto con estrategias comerciales de mayor cobertura, por lo que este modelo aparece como una oportunidad de mejorar la gestión en un porcentaje de los clientes insatisfechos, pero no pueden ser estos trabajados solo por los resultados de este modelo.

A partir de los resultados obtenidos se concluye, en primer lugar, que el mejor modelo obtenido logra identificar a los clientes insatisfechos mejor que un clasificador aleatorio, lo que se ve en los gráficos de lift y AUCPR. En segundo lugar, los resultados también muestran que si bien se identifica de mejor manera a los clientes insatisfechos, los modelos tienen aún espacio para mejorar a través dela inclusión de más variables que den cuenta de otros aspectos de una manera más específica, de los cuales destaca incorporar información que de cuenta de las diferencias del servicio entregado por la firma dependiendo de la ubicación geográfica donde reside el cliente.

5.5. Aplicación del modelo

Luego de la evaluación del modelo, se procede a la tarea de identificar como afecta la satisfacción del universo completo de clientes pospago en 3 ámbitos de importancia para la firma, la fuga de clientes de la firma, la conversión a campañas de marketing y el share de cambio de equipo a través de canales de la firma.

La metodología realizada es la siguiente, los 3 aspectos mencionados son estudiados para 3 meses diferentes, enero, febrero y marzo del 2020, esto para evitar caer en generalizaciones y conclusiones erróneas a las que se pueda llegar estudian0do los efectos de la satisfacción en un solo mes.

Para cada mes estudiado, se entrena el modelo random forest obtenido en la sección anterior para obtener la predicción de satisfacción del mes anterior al estudiado. Utilizando enero como ejemplo, se entrena el modelo utilizando datos de encuestas hasta noviembre y con esto se predice la satisfacción de los clientes para diciembre. Una vez con la predicción de satisfacción lista para todos los clientes, se estudia como se relaciona la satisfacción de

diciembre con la tasa de fuga, conversión a campañas y share de cambio de equipo de los clientes al mes siguiente (enero). Esto se resume en la figura 5.24. Este proceso es repetido luego para evaluar los efectos en febrero y marzo

Este proceso requiere de un constante re-entrenamiento del modelo de satisfacción mes a mes. El comportamiento de los clientes y la industria de las telecomunicaciones es dinámico y con constantes cambios, por lo que los modelos aplicados en la industria deben ser capaces de ir reconociendo estos cambios y la forma en que se logra esto es incorporando a los modelos la información más reciente que se tenga disponible. También es necesario realizar esto para asegurar que las predicciones son realizadas con información actual de los clientes, tanto de uso de servicios como socio-demográfica.

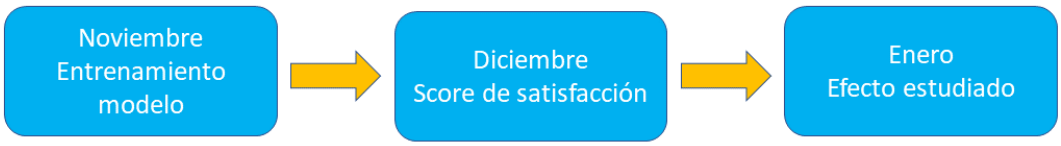

Figura 5.24: Esquema de evaluación de efectos de la satisfacción.

A diferencia de la sección anterior, donde se clasificó a los clientes como satisfechos o insatisfechos, para esta sección se opta por ordenar a los clientes según su probabilidad de ser un cliente insatisfecho y asignar un score de satisfacción.

Para esto, luego de ordenar a los clientes según su probabilidad, se divide a los clientes de un mes dado en percentiles, donde el percentil 1 corresponde a los clientes con mayor probabilidad de ser clientes insatisfechos y el percentil 100 a los clientes con las probabilidades más bajas de ser insatisfechos (o mayor probabilidad de ser cliente satisfecho).

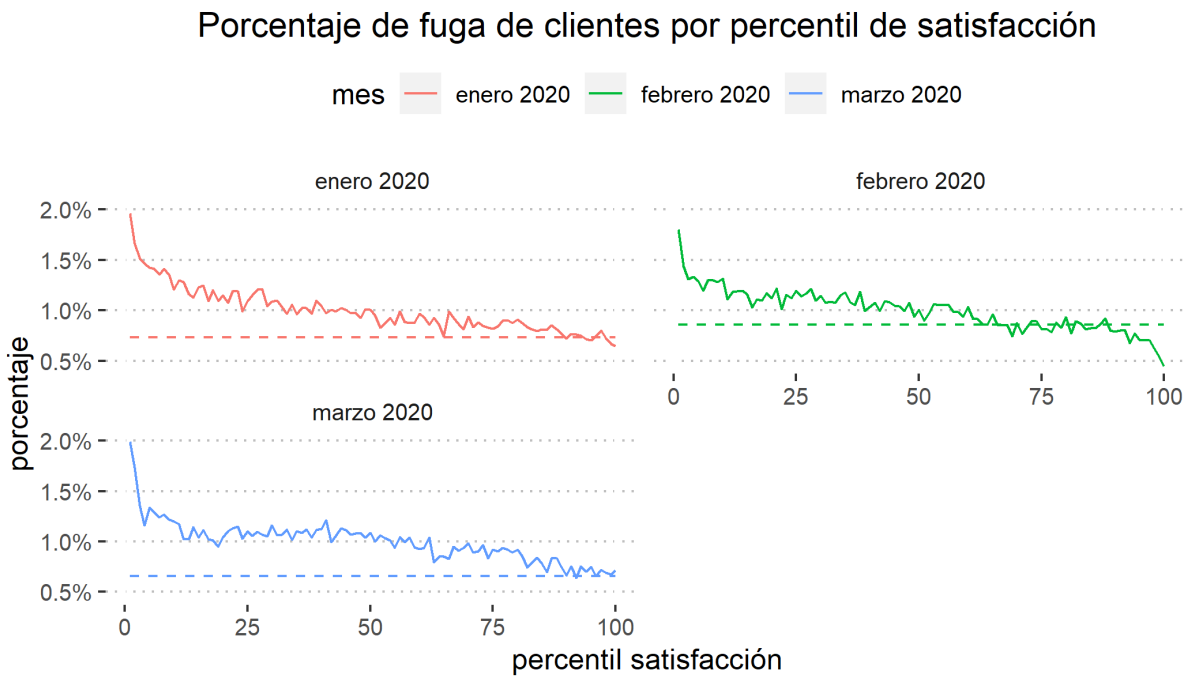

Figura 5.25: Comparación tasa de fuga por percentil de satisfacción contra tasa de fuga base.

El primer efecto que se estudia es la fuga de clientes, lo cual se puede ver en la figura 5.25. Los gráficos muestran para cada mes estudiado la tasa de fuga que hubo dentro de cada percentil de satisfacción construido.

Los resultados muestran que en los 3 meses estudiados hay una relación clara entre la satisfacción de un cliente y su probabilidad de fuga. Estos resultados confirman nuevamente trabajos previos que concluyen que la satisfacción es un predictor significativo de la fuga [13]. El efecto es más grande para el primer percentil de satisfacción, donde en promedio la tasa de fuga es alrededor de 3 veces la tasa base para los meses estudiados.

Si bien de estos estudios se conocía esta relación, esta evaluación de impacto de la satisfacción en la fuga permite entregar una muestra visual de esta relación y su magnitud a nivel global de clientes, algo que no se puede realizar en los otros estudios al solo poder basar los resultados en los clientes encuestados.

Al comparar el extremo izquierdo con el derecho en los 3 meses, se puede apreciar que el efecto de la satisfacción de un cliente produce en mayor magnitud un aumento en la probabilidad de fuga que una disminución en la probabilidad de esta. Febrero es el único mes en donde los clientes con mayor probabilidad a ser clientes satisfechos tienen una disminución notoria en su tasa de fuga, mientras que para enero y marzo estos solo se acercan a la tasa base.

Esto quiere decir que mientras más insatisfecho sea un cliente, el efecto de lograr reducir su insatisfacción (aumentando así su probabilidad de ser un cliente satisfecho) tendrá un impacto mayor en la reducción de su probabilidad de fuga, que aumentar la satisfacción de

un cliente que ya esta previamente dentro del grupo de clientes satisfechos, donde el efecto de tratar de "llevarlos" al percentil 100 no es igual que "sacarlos" del percentil 1.

La literatura nos dice que existen dos ejes principales para la retención de clientes, uno a través de la satisfacción de clientes, que contiene factores mayormente emocionales y otro a través de los costos de cambio y competencia existente, donde se encuentran las decisiones más racionales y basado en los costos de cambio y ofertas de la competencia.

Los gráficos presentados muestran el grado en que el primer eje mencionado, la satisfacción de clientes, está impactando a la fuga dentro de la firma. Esta información puede resultar esencial para diferenciar clientes para los cuales la satisfacción tiene un rol significativo en su decisión de fuga versus clientes que toman esta decisión basándose principalmente en los costos asociados al cambio con la competencia.

Diferenciar entre estos tipos de clientes puede ser fundamental para mejorar el targeting de estrategias comerciales y la elaboración de planes a largo plazo. En el corto plazo se puede evaluar a que clientes es más probable que estrategias relacionadas al costo del servicio serán más efectivas.

En el largo plazo, esta identificación de clientes insatisfechos puede ser fundamental para poder identificar sectores en los cuales utilizar ofertas como cambios de equipo o un mejor plan puedan no solucionar el problema de fondo, una señal con deficiencias, por lo que realizar efectos comerciales de este tipo no estarían realmente disminuyendo la probabilidad de fuga del cliente, dado que el motivo de su insatisfacción no ha sido solucionado. A partir de los resultados del modelo se podría priorizar sectores en los cuales se requiere mejoras de servicio para realmente mejorar la satisfacción y de esta forma disminuir las tasas de fuga.

Con esto resultados en mente, se investiga a continuación la existencia de relaciones entre la satisfacción de clientes y conversiones a campañas de la firma. Debido a que los modelos utilizados requieren de que los clientes tengan información dentro de la firma, se prueban dos campañas, abreviadas A y B, en las cuales se ofrece a clientes que ya son parte de la compañía el adquirir un servicio adicional con la firma, específicamente se les oferta el adquirir una línea extra.

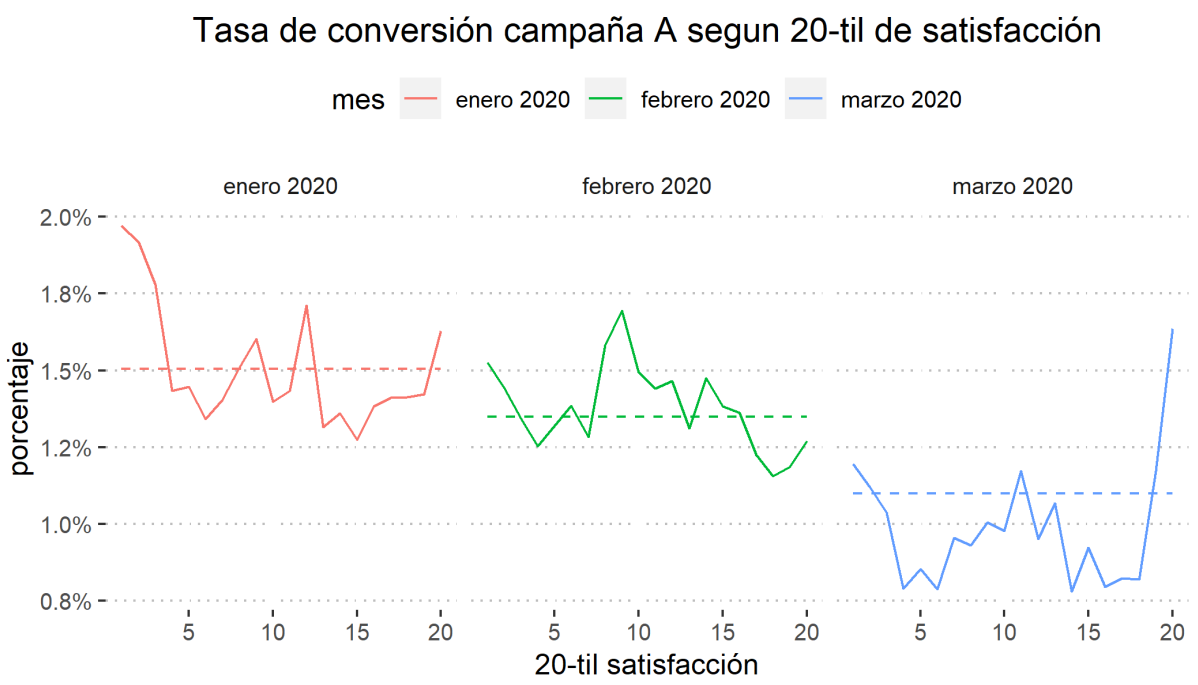

Figura 5.26: Comparación tasa de conversión campaña A por 20-til de satisfacción contra tasa de conversión base.

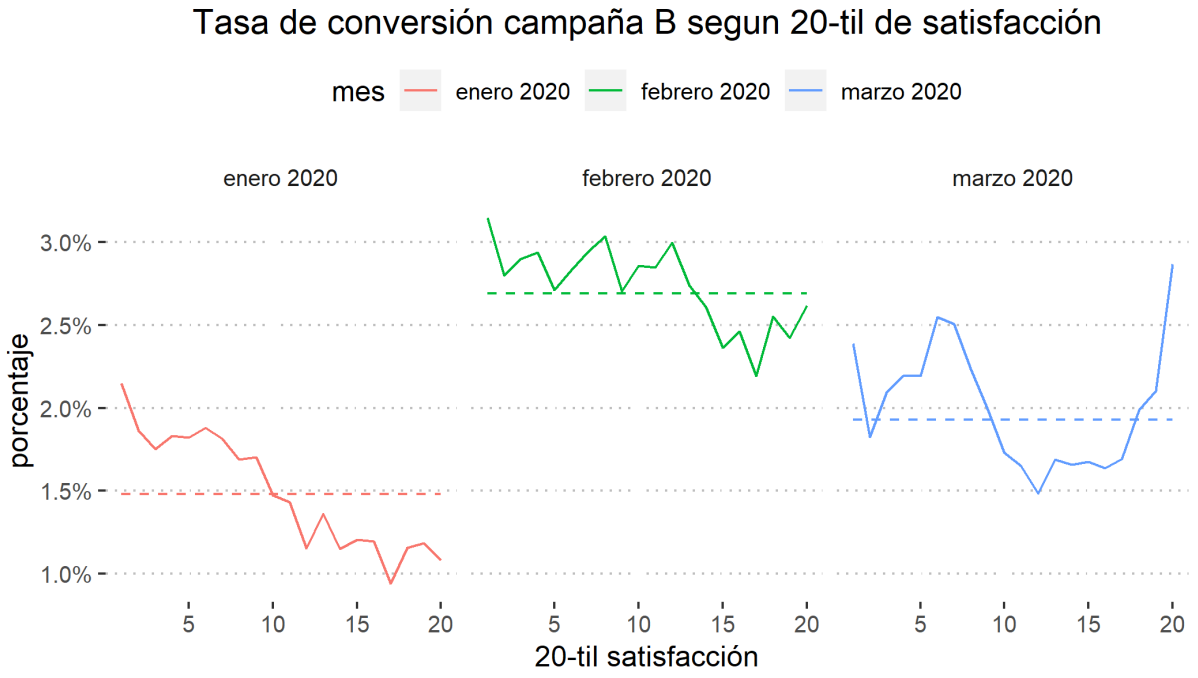

Figura 5.27: Comparación tasa de conversión campaña B por 20-til de satisfacción contra tasa base.

Las figuras 5.26 y 5.27 muestran los resultados obtenidos en los meses estudiados para las campañas A y B respectivamente. Para cada mes, la línea punteada corresponde a la tasa de conversión general que obtuvo la campaña, comparada contra la conversión agrupando según

satisfacción, donde los percentiles fueron esta vez agrupados en 20 grupos de 5 percentiles cada uno para suavizar las curvas y efectos (los cuales llamamos 20-tiles).

La diferencia entre ambas campañas radica principalmente en el precio de la línea extra ofrecida y el segmento de valor al que pertenece el cliente, la campaña A esta dirigida a clientes de valor básico y bajo, mientras que la campaña B esta dirigida a los segmentos medio, alto y full.

A diferencia de la relación que se obtuvo entre la satisfacción y fuga de clientes, para la conversión de campañas no se puede ver una relación clara entre estos dos factores para ninguna de las dos campañas revisadas. En cuanto a la campaña A, se ve como la conversión por 20-til oscila en torno a la tasa base, por lo que los resultados de un mes no son parecidos a otro. Se ve como los clientes con mayor insatisfacción convierten más en enero, luego en febrero la mayor conversión se obtuvo para los clientes del área intermedia, mientras que en marzo fueron los clientes más satisfechos quienes tuvieron la conversión más alta.

En cuanto a la campaña B, al mirar enero y febrero se puede ver como tienden a convertir más los clientes insatisfechos, conclusión que no se puede hacer en marzo, donde los resultados no siguen la misma tendencia.

Estos resultados son interesantes ya muestran que en las campañas estudiadas, donde se ofrece al cliente adquirir una línea extra, no existe una relación directa entre la satisfacción de los clientes y la tasa de conversión esperada, por lo que son otros los factores que definen mejor la probabilidad de convertir en este caso.

Esto nos indica que independiente de la satisfacción previa del cliente, si se identifican que factores son los que determinan la probabilidad de conversión a estas campañas, se podría incluir a estas tanto a clientes satisfechos como insatisfechos, sin esperar que su satisfacción previa influya en su probabilidad de adquirir el producto.

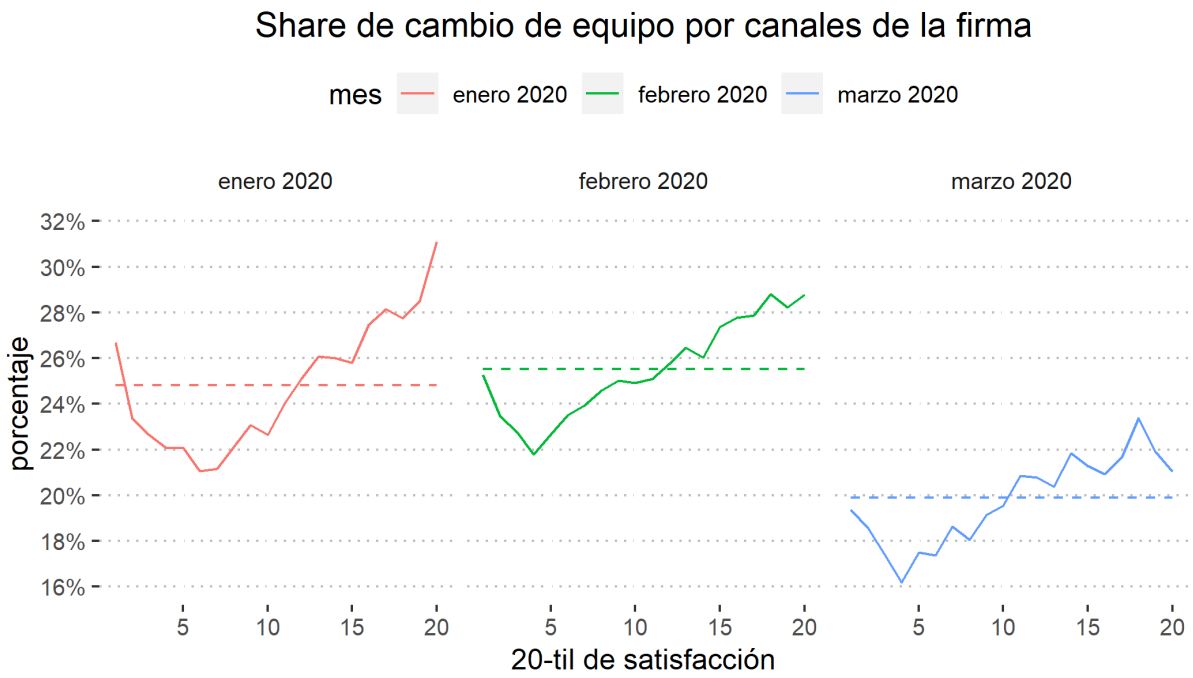

Figura 5.28: Comparación share de cambio de equipo por 20-til de satisfacción contra el share base.

El último aspecto que queda revisar es la relación entre la satisfacción de los clientes y el share de cambio de equipo a través de canales de la firma, lo que se puede ver en los gráficos de la figura 5.28. En este caso las líneas punteadas corresponden al share general dentro del respectivo mes, el cual tiende a ser entre un 20 % a un 25 % del total de cambios de equipos realizados por los clientes.

Para este caso se puede ver que a mayor satisfacción de los clientes, mayor es el share de cambio de equipo por canales de la firma, subiendo varios puntos porcentuales por sobre la tasa base, donde el mayor aumento se obtuvo en el mes de enero, en el cual los clientes en el 20-til veinte tuvieron un share 6 puntos porcentuales más alto.

Si se divide a los clientes desde el 20-til diez, se puede ver que los clientes dentro del 1 al 10 tienen en promedio un share de cambio de equipo por debajo de la tasa base, mientras que los del 11 al 20 superan la tasa base.

Se ve un efecto interesante al aumentar la insatisfacción de los clientes. A medida que disminuye la satisfacción, disminuye el share de cambio de equipo dentro de la compañía, pero este efecto se revierte para los clientes más insatisfechos, quienes vuelven a estar cerca del share base. Si bien no se puede explicar el por qué ocurre esto a partir de estos gráficos, se pueden generar hipótesis, como que los clientes más insatisfechos pueden tener problemas que son derivados tanto del servicio recibido como provocados por problemas en los equipos que utilizan, motivo por el cual dentro de los clientes más insatisfechos aumenta el share de cambio de equipo.

Dejando de lado esta hipótesis, lo que si se puede concluir a partir de los resultados de

estos gráficos, junto con los de la figura 5.25 es la existencia de un efecto sinérgico entre la satisfacción, la fuga y el share de cambio de equipo de la firma. Los clientes más satisfechos no solo tienen una tasa de fuga menor que los clientes insatisfechos, si no que además aumenta la probabilidad de que decidan cambiar su equipo a través de canales de la firma.

Ofertas de cambio de equipo aparecen como mejores alternativas para ser ofrecidas a los clientes más satisfechos con la firma. Los resultados sugieren que dentro de los clientes insatisfechos también existen grupos propensos a cambiar de equipo por la firma.

A partir de los resultados se estima conveniente agregar a iteraciones futuras de un modelo de satisfacción información relacionada al estado del equipo que utiliza el cliente. Probablemente, si se puede diferenciar mejor casos en los cuales el cliente esta insatisfecho por razones ligadas al teléfono en vez del servicio de la firma, se pueda identificar mejor a quienes de los clientes insatisfechos una oferta de equipo pueda ser la mejor opción a ofrecer.

5.6. Replicabilidad en prepago

El trabajo desarrollado fue implementado solo para el segmento pospago, pero en términos de cantidad de móviles, cada segmento corresponde a alrededor de un 50 % del total de la firma, por lo cual se deja una gran parte de los clientes fuera del análisis. La siguiente sección discute los motivos por los cuales no se incluyeron a los clientes prepago dentro de este trabajo y por que resulta conveniente trabajar los mercados de forma separada, agregando un análisis exploratorio de variables que sólo son relevantes dentro de este mercado mediante una regresión logística.

De la figura 2.5 vimos que la satisfacción de los clientes pospago no es igual a la que tienen los clientes prepago. Uno de los insights que derivan de esta diferencia es que los servicios entregados por la firma impactan en diferentes magnitudes a los clientes de cada segmento, magnitudes que no podrían ser captadas si nos se diferencia a los clientes según su segmento.

Las diferencias no solo existen en los mercados en cuanto a su percepción de la calidad de los servicios, si no que además la cantidad de información disponible en cada categoría de clientes es diferente. Este es uno de los motivos fundamentales para trabajar los mercados por separado, y también el motivo por el cual se trabaja en pospago de forma inicial, ya que dentro de pospago la disponibilidad de información sobre el cliente es mucho mayor.

Dentro de pospago, al adquirir un plan el cliente debe entregar su rut y otros datos para realizar el contrato, mientras que en prepago esta información no es entregada por el cliente. La falta de esta información genera una gran disparidad en cuanto a la disponibilidad de información, ya sea esta demográfica o relacionada a otros factores. Por ejemplo, las información de visitas a tiendas revisada en la sección 5.2.1 se tiene a nivel de rut y no por móvil, y como no se cuenta con el rut dentro de prepago, no se pueden identificar de manera aceptable las visitas realizadas por clientes a partir de estos datos.

Sumado a la falta de información, el uso de ciertos servicios también es menor en prepago. Dentro de los meses estudiados de la información relacionada a la aplicación de la compañía,

donde un 20 % de los clientes pospago hace uso de la aplicación en promedio, solo un 6 % de los clientes *activos* en prepago hicieron uso de la app entre diciembre y marzo, por lo cual el valor potencial de estudiar el uso de la app en prepago es reducido en comparación a pospago.

De lo anterior se desprende otra diferencia con pospago, la cual da cuenta del estado del móvil, pudiendo ser activo o inactivo. En prepago existe un gran porcentaje de móviles que tienen un uso esporádico y pueden pasar meses sin ser utilizados, produciendo móviles con muy poca información disponible para realizar estudios sobre este tipo de usuarios, agregando más diferencias junto a lo ya descrito.

Existen todavía más diferencias que justifican trabajar ambos mercados por separado. En pospago se vio que los problemas por facturación influyen significativamente en la satisfacción, lo cual no es relevante para explicar la satisfacción en prepago, mientras que en este último la compra de bolsas es un servicio principal entregado por la firma que es relevante solo dentro de este segmento.

Comenzar el análisis con pospago tiene un motivo estratégico. Ya que con este segmento se puede acceder a mayor información de los clientes, esto se utiliza luego para entender qué información dentro de la que se tiene disponible para prepago es relevante y si se debe o no priorizar la obtención de información con la que no se cuenta actualmente.

A partir de los resultados en pospago, se concluyó que la información de visitas a tiendas y los datos de interacciones dentro de la app tienen la menor influencia en la satisfacción en comparación al resto de variables. Si bien esto no asegura que en prepago los resultados deban ser idénticos, si se concluye que es más probable que la información socio demográfica sea también más relevante dentro de prepago, por lo cual obtener este tipo de información es prioritario por sobre los otros dos casos (a partir de los resultados en la figura 5.22).

Por otra parte, de la información que tuvo mayor relevancia, reclamos, interacciones con call center y el resto de canales, datos relacionados con la señal y los cambios de equipo, estos están disponibles para prepago de la misma forma que en pospago, por lo que se cuenta con gran parte de la información que tuvo mayor relevancia en pospago y que probablemente también sea relevante para prepago.

Para una correcta replicación del trabajo desarrollado dentro de prepago, faltaría analizar aspectos de prepago que tienen mayor relevancia dentro de este mercado. Como ya se mencionó, uno de estos aspectos corresponde a la compra de bolsas. se estudia a continuación una base de compras históricas de bolsas realizadas por clientes prepago a modo de entender la influencia de variables relevantes para prepago en la satisfacción de estos clientes.

La información cuenta con la fecha en que fue comprada la bolsa, el precio, el canal por donde fue comprada y el tipo de bolsa que corresponde, pudiendo ser de datos, voz, mixtas, de roaming o larga distancia internacional (LDI), o centradas en el uso de redes sociales o de video. Existe 5 canales principales para la compra de bolsas en la firma, a través de la aplicación, la web, IVR,USSD y una "página blanca", la cual corresponde a un portal web para el cual se requiere acceder una vez utilizando datos, se guardar la página y luego se puede volver a ingresar sin necesidad de utilizar datos móviles. Existen más medios de compra de

bolsas con un uso menor, que se agrupan en la categoría "otros".

Se propone estudiar la cantidad de bolsas compradas diferenciando por el tipo de bolsa, junto con una variable que da cuenta del canal "preferido" del cliente, que asigna como canal preferido aquel donde el cliente realizo la mayor cantidad de compra de bolsas. Lo anterior se realiza utilizando una ventana de 3 meses de historia y se considera también el promedio de recarga de esos meses. A modo de control, agregamos la edad, antigüedad del cliente y el valor del cliente, donde en este caso se diferencia a los clientes en bajo valor, alto valor y clientes que no han realizado recargas en más de 3 meses, "recarga 0").

Tabla 5.4: Resultados regresión logística sobre compra de bolsas en prepago.

	Dependent variable:
	Cliente satisfecho 1 mes
edad	0.001 (0.002)
Antigüedad cliente	0.0001 (0.001)
Promedio gasto bolsas	0.0001** (0.00003)
Valor_cliente.bajo valor	0.157*** (0.061)
Valor_cliente.recarga 0	0.168** (0.086)
Canal_preferido.APP	−0.028 (0.133)
Canal_preferido.IVR	−0.123 (0.075)
Canal_preferido.PaginaBlanca	−0.316* (0.167)
Canal_preferido.Otros	−0.927** (0.406)
Canal_preferido.USSD	−0.150 (0.112)
Canal_preferido.WEB	−0.219 (0.166)
Bolsas datos	−0.013** (0.006)
Bolsas mix	0.004 (0.005)
Bolsas datos_nocturno	−0.102*** (0.037)
Bolsas Voz	0.036 (0.061)
Bolsas rrss_video	0.212 (0.202)
Bolsas roam_ldi	0.216 (0.229)
Constant	1.978*** (0.110)
Observations	15,043
Log Likelihood	−5,259.400
Akaike Inf. Crit.	10,555.000
Note:	*p<0.1; **p<0.05; ***p<0.01

La tabla 5.4 muestra los resultados de una regresión logística de las variables descritas tomando como variable dependiente nuevamente la satisfacción del cliente (tomando valor 1 si el cliente declaro estar satisfecho y 0 si no), clasificando las respuestas de los clientes de la forma ya discutida en la sección 5.3. Los resultados muestran que la probabilidad de estar

satisfecho con la firma aumenta para los clientes de bajo valor y quienes no han realizado recargas en los últimos 3 meses, lo cual concuerda con la hipótesis de que los clientes de mayor valor son más exigentes con el servicio en comparación al resto de clientes.

Con respecto al canal preferido de compra de bolsa del cliente, los resultados se comparan con la categoría "ninguno" que corresponde a los clientes que no compraron bolsas en el periodo estudiado. Se concluye que los clientes que utilizan como medio principal de compra de bolsas la *página blanca* tienen mayor probabilidad de estar insatisfechos con el servicio en comparación a los clientes que no han comprado bolsas (pero con una significancia solo del 90 %). Junto con estos, los clientes que utilizan principalmente los medios no convencionales de la firma, agrupados en "otros", tienen un aumento aún mayor en la probabilidad de ser clientes insatisfechos (y con una sgnificancia del 95 %).

Estos resultados indican que dentro de todos los canales disponibles es necesario indagar en las razones que puedan generar que la página blanca y los canales dentro de "otros" provoquen una mayor insatisfacción en los clientes. Sobre las variables que cuentan la cantidad de bolsas compradas por tipo de bolsa, se puede ver que solo la compra de bolsas de datos y datos nocturnos (como su nombre lo dice, permite acceder a servicios de datos solo en horas de noche) tienen un impacto significativo en la satisfacción. Cada compra extra de bolsas de datos realizada por el cliente genera un aumento en la probabilidad de estar insatisfecho con la firma, lo cual no ocurre con las bolsas mixtas, de voz, redes sociales o video ni roaming con larga distancia internacional.

La hipótesis que se construye a partir de estos resultados tiene que ver con la señal. Los servicios de voz pueden funcionar de manera aceptable en gran parte del rango de calidad de la señal, mientras que la calidad de los servicios de datos tienen una variabilidad mucho mayor en función de la calidad de la señal. Se cree que los clientes que compran bolsas solo centrado en el servicio de datos son más propensos a ser afectados en mayor medida al notar una falla en el servicio a diferencia de clientes que compran bolsas con servicios de voz incluidos, como las mixtas.

Todos estos resultados muestran que es necesario incluir información relacionada con la compra de bolsas dentro de un modelo de predicción de satisfacción en prepago y que es necesario estudiar ambos mercados por separado. Agregando información específica para prepago junto con los insights del trabajo en pospago, se concluye a continuación las consideraciones para poder replicar el trabajo desarrollado dentro de prepago:

- La información relacionada a visitas a tiendas al no ser de fácil obtención para prepago, puede ser omitida en un comienzo. Los motivos para esto son que los modelos de la sección 5.4 mostraron que esta información tiene menor relevancia que otros datos, y como conseguir esta información dentro de prepago tiene un costo elevado, no se justifica su inclusión.
- El uso de la aplicación en prepago aún es bajo en comparación a pospago (6 % en comparación a 20 %), esto sumado al bajo aporte que entrega esta información a la predicción de satisfacción provocan que su inclusión dentro de este mercado no se justifique. Una vez el uso de la aplicación dentro de prepago sea mayor se puede decidir analizar estos datos dentro de prepago.

- Dentro de los datos analizados que es necesario incorporar dentro de un modelo de satisfacción en prepago, se encuentra la información de interacciones, cambios de equipo y reclamos. Es bastante seguro concluir que es esencial agregar dentro de este mercado una mayor cantidad de información relacionada a la señal que reciben los clientes de la firma, como se concluyó ya dentro de pospago.

- La información socio demográfica resultó significativa para la predicción de satisfacción, datos que no se tienen disponibles dentro de prepago. Se deben priorizar formas de obtener este tipo de información para los clientes de este segmento de una forma confiable.

- Se debe agregar información que es relevante dentro de este segmento. La información revisada en esta ocasión de compra de bolsas debe ser agregada a un modelo de satisfacción para poder determinar su importancia frente al resto de variables en este mercado.

Capítulo 6

Conclusiones

La primera conclusión del trabajo es que es posible realizar modelos de predicción de satisfacción aplicables a todos los clientes de pospago a partir de los resultados de encuestas de satisfacción. De las 3 fuentes de datos trabajadas específicamente para este trabajo, la información de visitas a tiendas y de uso de la aplicación resultó tener un aporte despreciable en la predicción de la satisfacción de los clientes de la firma.

Los resultados de las regresiones logísticas donde se probaron distintas ventanas de tiempo entregaron resultados interesantes. Exceptuando las visitas a tiendas, las variables que fueron significativas, lo fueron en las 3 ventanas utilizadas. De estas, la magnitud de los efectos son más grandes en todos los casos al utilizar 1 mes de información, disminuyendo con cada mes de información que se agrega.

Sin embargo esta disminución es bastante leve y la diferencia en el criterio akaike entre el modelo utilizando 1 mes con el que utiliza 3 meses de historia es tan solo de un 0.4 %. Dada esta diferencia, se elige como mejor ventana el utilizar 3 meses de historia, ya que de esta forma cada variable contiene información relevante para una mayor cantidad de clientes.

En cuanto a los modelos de caja negra, los mejores resultados se obtuvieron al optimizar el área bajo la curva precision-recall por sobre la métrica de accuracy, lo que confirma las conclusiones de otros trabajos revisados que sugieren que este parámetro obtiene mejores resultados cuando se trabaja con problemas con desbalance de clases. Para iteraciones futuras del modelo de satisfacción y problemas que cumplan con estas características se aconseja seguir entrenando optimizando este parámetro.

Para estos modelos, las variables provenientes de la información de interacciones mostraron ser fundamentales para los modelos, y se pudo identificar a partir de la importancia de variables obtenida que diversos aspectos tanto del servicio como del cliente influyen en la predicción de satisfacción, por lo que no hay un solo factor predominante para predecir satisfacción.

Destaca que la región corresponda a la variable con mayor importancia, lo cual es interpretado como un efecto de falta de datos que den cuenta sobre las diferencias que existe en los servicios entregados por la firma en las distintas regiones. Estas diferencias pueden ser de cobertura y calidad de la señal, cantidad de tiendas disponibles, servicios adicionales como

internet hogar o cuanta competencia de firmas hay en la región.

Este modelo entrega a la firma la oportunidad de gestionar de manera más efectiva a aproximadamente un 16 % de los clientes insatisfechos, a través de la gestión de solo a un 6 % de sus clientes en un mes dado. Para abarcar a una cantidad más amplia de clientes insatisfechos, todavía es necesario acompañar esta gestión con estrategias comerciales de mayor cobertura.

La aplicación de los modelos en los distintos factores estudiados entregaron resultados interesantes para la firma. En cuanto a la fuga, se vio que existe una relación clara entre la satisfacción del cliente y su probabilidad de fuga, donde en promedio para los 3 meses vistos, la tasa de fuga para los clientes más insatisfechos (percentil 1) era 3 veces mayor a la tasa de fuga base.

También se pudo ver que a medida que aumenta el percentil de satisfacción, lo que se gana al pasar al siguiente percentil es cada vez menor, por lo que los mayores beneficios están en concentrarse en los clientes más insatisfechos que en los de la zona intermedia y alta.

Los resultados sugieren que la predicción de satisfacción puede ser un input relevante para modelos de fuga, donde se podría diferenciar a los clientes propensos a fugarse por motivos relacionados solo a precios contra los clientes que también consideran su satisfacción y relación con la firma. Esto permitiría mejorar las estrategias comerciales que se realicen para los clientes dependiendo del caso al que pertenezcan.

Existe un aumento en el share de cambio de equipo a través de canales de la firma al aumentar la satisfacción del cliente de varios puntos porcentuales. Los resultados también sugieren que dentro de los clientes más insatisfechos también existen grupos que tienden a cambiar más sus equipos por la firma.

De estos dos resultados se ve un efecto sinérgico muy beneficioso para la firma, por una parte, la principal ventaja de dejar de ser un cliente insatisfecho es la reducción en la tasa de fuga, mientras que al aumentar la satisfacción del cliente, este se vuelve más propenso a cambiar su equipo a través de los canales de la firma.

Dentro del tercer aspecto estudiado, la tasa de conversión a campañas donde se ofrece una línea extra, no se puede concluir sobre la existencia de una relación entre la satisfacción y al probabilidad de adquirir una nueva línea. Esto significa que existen aspectos del negocio dentro de los cuales la satisfacción del cliente se ve superada por otros factores que influyen con mayor prevalencia en la decisión de compra del cliente.

De los datos trabajados, La información socio demográfica resulto ser de gran importancia, información que a la fecha no se tiene disponible para todos los clientes prepago de la firma y que dificulta aplicar lo trabajado en pospago con los clientes prepago.

La firma debería priorizar la elaboración de métodos que permitan identificar a sus clientes de manera confiable. Una propuesta para poder conseguir esto sería potenciando el ingreso a la aplicación por parte de los clientes, ya que en esta plataforma el cliente debe ingresar su número y rut. Esto podría lograrse de varias maneras, crear un proceso en el cual sea necesario

ingresar a la aplicación, incentivar un primer uso de la aplicación a través de beneficios en bolsas, saldo u otros servicios, entregar un beneficio mayor por comprar bolsa a través de la aplicación en lugar de otros canales, etc.

Existe una gran cantidad de trabajos en satisfacción que concluyen sobre los distintos beneficios que se ganan al tener un cliente satisfecho o se pierden con los clientes insatisfechos. Este trabajo resalta de estos trabajos en satisfacción principalmente al salirse de la evaluación de resultados solo dentro de los clientes encuestados y pasar a la aplicación de un score de satisfacción para el universo completo de clientes pospago y buscar encontrar estos beneficios para todos los clientes de la firma.

Los resultados permiten cuantificar los efectos en fuga, share de cambio de equipo y conversión a campañas, donde otros trabajos solo llegan a la evaluación de los coeficientes obtenidos por sus modelos, donde si bien casi siempre se concluye que la satisfacción es un aspecto fundamental en aspectos como la fuga, no se ven intentos de agregar la satisfacción como un input más para gestionar de mejor forma a los clientes.

6.1. Recomendaciones futuras

Los resultados mostraron al necesidad de incorporar nuevas variables que identifiquen mejor ciertos aspectos del servicio, lo cual se desprende del hecho que la región apareciera como la información más importante del modelo random forest.

Específicamente es necesario incorporar una mayor cantidad de información que de cuenta de las diferencias en el servicio ofrecido según la zona geográfica, como cobertura, señal, tiendas disponibles, competencia presente, etc. Como trabajo futuro se propone incorporar esta información y revisar como cambian las predicciones del modelo.

Se recomienda utilizar información que especifique más como son las interacciones entre el cliente y la firma. Por ejemplo, con la información utilizada no se pudieron estudiar efectos como el tiempo que pasan los clientes en espera al realizar una llamada al call center, o el efecto que tiene la duración promedio de las llamadas realizadas en la satisfacción. Estos son efectos que sería interesante revisar.

Como se vio del análisis de ventanas de tiempo, al utilizar solo un mes de historia los efectos revisados tienen un mayor impacto en la satisfacción al ser más recientes, pero involucran a una menor cantidad de clientes. En este trabajo se le dio prioridad al poder abarcar a una mayor cantidad de clientes sacrificando parte del efecto en satisfacción de la información utilizada. Futuras versiones del trabajo podrían centrar su foco en obtener efectos mas grandes, para lo cual cambiar la selección de una ventana de 3 meses a la ventana de 1 mes sería una opción válida.

Este trabajo se centro en estudiar a partir de la predicción de satisfacción 3 aspectos específicos, pero un trabajo futuro no tiene por que evaluar solo estos mismos 3 aspectos. Se recomienda evaluar otros tipos de campañas en los cuales se ofrezca un producto diferente a una línea extra para identificar si existen o no relaciones entre la satisfacción y otras ofertas

que pueda realizar la compañía.

Dentro de la sección de análisis de las encuestas, se mencionaron distintos tipos de sesgo sobre los que no se pudo controlar en este trabajo. Se recomienda para próximos trabajos intentar conseguir una mayor cantidad de información previa sobre la realización de estas encuestas y así poder confirmar o no la presencia de estos sesgos y corregirlos en caso de que sea factible.

Se recomienda también explorar técnicas para limpiar respuestas con errores o discrepancias que luego ensucien los resultados futuros, como la creación de reglas para eliminar respuestas donde la justificación a la nota general no concuerda con la nota entregada (por ejemplo, nota a satisfacción general aparece como un 1, pero el cliente justifica su nota con una explicación que concuerda con la de alguien satisfecho, lo que podría implicar que el 1 corresponde a un error de transcripción).

Bibliografía

[1] Antonio Alaminos. Teoría y práctica de la encuesta. aplicación en los países en vías de desarrollo, 1998.

[2] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[3] Dane Bertram. Likert scales. *Retrieved November*, 2:2013, 2007.

[4] Leo Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[5] Hernandol Canargo and Mario Silva. Dos caminos en la búsqueda de patrones por medio de minería de datos: Semma y crisp. *Journal of Technology*, 9(1):12–17.

[6] ICR Chile. El desafiante entorno de una industria altamente competitiva, análisis de la industria de telecomunicaciones en chile, 2019. URL https://www.icrchile.cl/index.php/estudios/3357-analisis-de-la-industria-de-telecomunicaciones-en-chile/file.

[7] Bernard Choi, Ricardo Granero, and Anita Pak. Catálogo de sesgos o errores en cuestionarios sobre salud. *Revista Costarricense de Salud Pública*, 19(2):106–118, 2010.

[8] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[9] Ministerio de Transportes y Telecomunicaciones. Subsecretaría de Telecomunicaciones. Informe anual de actividad del sector telecomunicaciones, 2019. URL https://www.subtel.gob.cl/estudios-y-estadisticas/informes-sectoriales-anuales/.

[10] Peter Flach and Meelis Kull. Precision-recall-gain curves: Pr analysis done right. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 838–846. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5867-precision-recall-gain-curves-pr-analysis-done-right.pdf.

[11] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[12] Teresa Garín-Muñoz, Teodosio Pérez-Amaral, Covadonga Gijón, and Rafael López. Consumer complaint behaviour in telecommunications: The case of mobile phone users in spain. *Telecommunications Policy*, 40(8):804–820, 2016.

[13] Anders Gustafsson, Michael D Johnson, and Inger Roos. The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention. *Journal of marketing*, 69(4):210–218, 2005.

[14] David W.; Stanley Lemeshow Hosmer. *Applied Logistic Regression, 2nd ed.* New York; Chichester, Wiley., 2000.

[15] Shin-Yuan Hung, David C Yen, and Hsiu-Yu Wang. Applying data mining to telecom churn management. *Expert Systems with Applications*, 31(3):515–524, 2006. URL https://www.sciencedirect.com/science/article/pii/S0957417405002654.

[16] David R. Jhonson. Using weights in the analysis of survey data, 2008. URL http://www.nyu.edu/classes/jackson/design.of.social.research/Readings/Johnson%20-%20Introduction%20to%20survey%20weights%20%28PRI%20version%29.pdf.

[17] Pang-Ning Tan; Michael Steinbach; Anuj Karpatne; Vipin Kumar. *Introduction to Data Mining.* Pearson, 2019.

[18] Charles X. Ling and Victor S. Sheng. *Class Imbalance Problem*, pages 171–171. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_110. URL https://doi.org/10.1007/978-0-387-30164-8_110.

[19] Zhongkai Liu and Howard D Bondell. Binormal precision–recall curves for optimal classification of imbalanced data. *Statistics in Biosciences*, 11(1):141–161, 2019.

[20] Sungwook Min, Xubing Zhang, Namwoon Kim, and Rajendra K Srivastava. Customer acquisition and retention spending: An analytical model and empirical investigation in wireless telecommunications markets. *Journal of marketing research*, 53(5):728–744, 2016.

[21] Saad Ahmed Qureshi, Ammar Saleem Rehman, Ali Mustafa Qamar, Aatif Kamal, and Ahsan Rehman. Telecommunication subscribers' churn prediction model using machine learning. In *Eighth International Conference on Digital Information Management (ICDIM 2013)*, pages 131–136. IEEE, 2013.

[22] Julio Villena Román. Crisp-dm: La metodología para poner orden en los proyectos, 2016. URL https://www.sngular.com/es/data-science-crisp-dm-metodologia/.

[23] Thanasis Vafeiadis, Konstantinos I Diamantaras, George Sarigiannidis, and K Ch Chatzisavvas. A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 55:1–9, 2015. URL https://www.sciencedirect.com/science/article/abs/pii/S1569190X15000386.

[24] Peter C Verhoef. Understanding the effect of customer relationship management efforts on customer retention and customer share development. *Journal of marketing*, 67(4): 30–45, 2003.

[25] Yingwei Zhang, Zhenji Zhang, and Ruize Gao. Study on customer relation management based on data mining sliq algorithm. In *LISS 2014*, pages 475–481. Springer, 2015. URL https://link.springer.com/chapter/10.1007/978-3-662-43871-8_69.

Anexo A

Variables utilizadas

A continuación se presenta una breve explicación del contenido de cada variable utilizada dentro de este trabajo en la sección 5.4.

- *Edad:* Edad del cliente en años.
- *Sexo:* Sexo del cliente
- *Región:* Región donde vive el cliente.
- *GSE:* Nivel socio económico al que pertenece el cliente.
- *Segmento:* Segmento de valor comercial al que pertenece el cliente, los cuales corresponden a básico, bajo, medio, alto y full, ordenados de menor a mayor valor.
- *Antigüedad cliente:* Cantidad de meses que el cliente lleva con la firma.
- *Negocio:* tipo de cuenta adquirida por el cliente, pudiendo ser cuenta controlado o suscripción.
- *Tecnología del equipo:* Tecnología de internet más alta que el equipo puede recibir dentro de las posibles que son ofrecidas por la firma, pudiendo ser 2G, 3G o 4G.
- *Cantidad de líneas:* Cantidad de móviles adquiridos por el cliente.
- *Cantidad de llamadas de la competencia:* cantidad de llamadas recibidas por el cliente que fueron realizadas por números que no pertenecen a clientes de la firma.
- *Cargo fijo:* cobro mensual que tiene el cliente.
- *Facturación:* Facturación del último mes del cliente.
- *Diferencia entre la factura y el cargo fijo:* Resta entre la última facturación del cliente y su cargo fijo.
- *Googleplay:* Variable categórica que indica si el cliente ha pagado servicios de la plataforma Googleplay a través de la boleta de la firma.

- *Spotify:* Variable categórica que indica si el cliente paga o ha pagado el servicio de Spotify a través de la boleta de la firma.
- *Netflix:* Variable categórica que indica si el cliente paga o ha pagado el servicio de Netflix a través de la boleta de la firma.
- *Antigüedad Spotify:* Cantidad de meses acumulados que el cliente ha pagado el servicio de Spotify a través de la boleta de la firma en caso de tener activado el servicio actualmente.
- *Antigüedad Netflix:* Cantidad de meses acumulados que el cliente ha pagado el servicio de Netflix a través de la boleta de la firma en caso de tener activado el servicio actualmente.
- *Promedio gasto Googleplay:* Promedio de gastos en los últimos 3 meses realizados por la plataforma Googleplay pagados a través de la boleta de la firma.
- *Cantidad de equipos:* Cantidad de equipos móviles que ha utilizado el cliente dentro de la firma.
- *Meses uso último equipo:* Cantidad de meses que cliente ha utilizado su equipo actual.
- *Tiempo ida y vuelta:* Tiempo promedio mensual de ida y vuelta en enviar una petición a un servidor de la aplicación de Facebook por comuna donde es realizada.
- *Fuerza de la señal: Potencia promedio mensual de las señales recibidas por un dispositivo en redes inalámbricas. Va desde 0 dBm a -100dBm y mientras más cercano sea el valor a 0, más fuerte será la señal. Valor obtenido a nivel de comuna.*
- *Calidad de la señal recibida: Valor que toma en consideración la fuerza de la señal, la demanda de la red, la interferencia y el ruido. Valor obtenido a nivel de comuna. Valor promedio mensual obtenido a nivel de comuna.*
- *Cambio de equipo en últimos 3 meses:* Variable categórica que indica si el cliente realizó o no un cambio de equipo en los últimos 3 meses.
- *Cambio de equipo en últimos 3 meses por canal de la firma:* Variable categórica que indica si el cliente realizó o no un cambio de equipo en los últimos 3 meses por algún canal de la firma.
- *Visitas tiendas propias 3m:* Cantidad de visitas realizadas por el cliente a tiendas de la firma en los últimos 3 meses.
- *Visitas franquicias 3m: Cantidad de visitas realizadas por el cliente a franquicias de la firma en los últimos 3 meses.*
- *App_ofertas:* Cantidad de interacciones realizadas en la aplicación de la firma por el cliente dentro de los servicios de ofertas.
- *App_chat:* Cantidad de interacciones realizadas en la aplicación de la firma por el cliente dentro de los servicios para acceder al chat con el bot elaborado por la firma.

- *App_beneficios:* Cantidad de interacciones realizadas en la aplicación de la firma por el cliente dentro de los servicios de beneficios para el cliente.
- *App_soporte:* Cantidad de interacciones realizadas en la aplicación de la firma por el cliente dentro de los servicios de soporte ofrecidos.
- *App_entretenimiento:* Cantidad de interacciones realizadas en la aplicación de la firma por el cliente dentro de los servicios de entretenimiento.
- *App_consumo/equipo/plan/boleta:* Cantidad de interacciones realizadas en la aplicación de la firma por el cliente dentro de los servicios de revisión de consumo realizado, revisión de equipo del cliente, consulta del plan contratado y estado de las boletas de la firma.
- *Call center:* Cantidad de interacciones realizadas entre el cliente y el call center en los últimos 3 meses.
- *IVR:* Cantidad de interacciones realizadas entre el cliente y los servicios de IVR en los últimos 3 meses.
- *USSD:* Cantidad de interacciones realizadas entre el cliente y los servicios de USSD en los últimos 3 meses.
- *Web:* Cantidad de interacciones realizadas por el cliente dentro de la web de la firma en los últimos 3 meses.
- *Chat:* Cantidad de interacciones realizadas por el cliente en chats con ejecutivos de la firma.
- *Problema o reclamo:* Cantidad de problemas o reclamos realizados por el cliente en los últimos 3 meses.
- *Distribución:* Cantidad de interacciones de distribución realizadas por el cliente en los últimos 3 meses.
- *Anulación/bloqueo de servicio:* Cantidad de anulaciones o bloqueos de servicio realizados por el cliente en los últimos 3 meses.
- *Problema facturación:* Cantidad de problemas o reclamos realizados por el cliente que están relacionados a la facturación de los servicios de la firma en los últimos 3 meses.
- *Intención de renuncia o insatisfacción:* Cantidad de veces que el cliente ha declarado tener intenciones de renunciar a la firma o que ha declarado estar insatisfecho con el servicio entregado, en los últimos 3 meses.

Anexo B

Interacciones

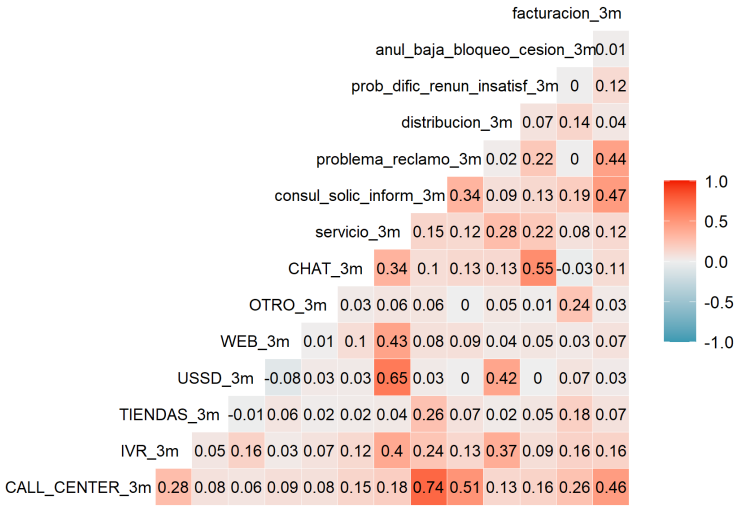

Figura B.1: Correlaciones interacciones pospago 3 meses de historia.