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

MODELOS DE PROPENSIÓN DE FUGA Y RELACIÓN DE LAS INTERACCIONES CON CLIENTES EN UNA EMPRESA DE TELECOMUNICACIONES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL

NATALIA VALERIA UBILLA SABABA

PROFESOR GUÍA:
DANIEL SCHWARTZ PERLROTH

MIEMBROS DE LA COMISIÓN:
PABLO MARÍN VICUÑA
MARÍA FERNANDA VARGAS COURBIS

SANTIAGO DE CHILE
2022

RESUMEN DE LA MEMORIA PARA OPTAR AL TITULO DE: Ingeniera Civil Industrial
POR: Natalia Valeria Ubilla Sababa
FECHA: 2022
PROFESOR GUIA: Daniel Schwartz Perlroth

MODELOS DE PROPENSIÓN DE FUGA Y RELACIÓN DE LAS INTERACCIONES CON CLIENTES EN UNA EMPRESA DE TELECOMUNICACIONES

La retención de clientes es un aspecto clave para la sobrevivencia de las empresas, especialmente considerando que la adquisición de un cliente es más costosa que retener uno ya existente. Por esta razón, es necesario identificar a los clientes que cancelarán el servicio con el fin de caracterizarlos y poder ejercer campañas de retención proactivas sobre los mismos.

En el presente trabajo de título se realizan modelos de predicción de fuga de clientes en una empresa de telecomunicaciones para el servicio de internet fibra óptica. Para esto se utiliza la metodología Cross Industry Standard Process for Data Mining y se entrenan 5 modelos de machine learning: Gradient Boosting Machine, Random Forest, Extremely Randomized Tress, Stacked Ensemble Best of Family y Stacked Ensemble All Models. Además, para solucionar el problema de clases desbalanceadas, cada modelo se entrena con 3 bases de datos distintas: una simple, una con algoritmo de random oversampling y una con algoritmo de random undersampling.

Los modelos que obtienen un mejor desempeño para la problemática planteada fueron los modelos de Random Forest y Stacked Ensemble. Por otro lado, los algoritmos de balanceo de clase no mostraron mejores resultados para las métricas de mayor importancia en este problema. Utilizando los modelos de mejor desempeño, se prevé que si se realizan acciones de retención proactiva a tan solo el 3% más propenso a fugarse se estarían contactando al 25% de las fugas totales que se reportan a nivel global en el servicio, lo que permitiría enfocar recursos en un grupo reducido de clientes, pero con un nivel alto de detección temprana de fugas.

También se identifica que las interacciones que realiza el cliente con la empresa a través del canal de Interactive Voice Response resultan un buen predictor de fuga de clientes, donde aquellas personas que más se contactan con la empresa poseen hasta un 57% mayor probabilidad de fugarse respecto a quienes no se contactan previamente. Para finalizar, se diseña un experimento que busca evaluar si las acciones de contactabilidad con el cliente o la entrega de descuentos podrían evitar la fuga en ciertos grupos establecidos como ‘propensos a fugarse’ por el modelo entrenado.

DEDICATORIA

Dedicada a los sueños cumplidos
y a los por cumplir.

TABLA DE CONTENIDO

INDICE DE TABLAS

INDICE DE GRÁFICOS

INDICE DE FIGURAS

CAPÍTULO 1: INTRODUCCIÓN

1.1 CONTEXTO DE LA EMPRESA Y RUBRO

El trabajo de título se desarrolla en una empresa chilena de telecomunicaciones que fue fundada en 1964. La empresa ofrece una variedad amplia de servicios, los que incluye telefonía móvil y operadores de red fija, tales como datos, integración TI, internet, telefonía local, larga distancia y otros servicios relacionados.

La empresa atiende las necesidades de sus clientes según segmentos de mercado: personas, empresas y corporaciones. A la vez, cada una de estas divisiones cuenta con un equipo propio que se encarga de la innovación y desarrollo de productos, precios, marketing, ventas y servicio al cliente.

Actualmente, se perfila como la empresa más grande de telecomunicaciones de Chile, donde sus ventas el año 2020 alcanzaron los $2.147 mil millones de pesos, lo que equivale a utilidades de $84.466 millones de pesos. Según lo estipulado en la memoria anual del año 2019, la empresa suma más de 17,4 millones de clientes, de los cuales 9,1 millones corresponden a clientes en Chile.

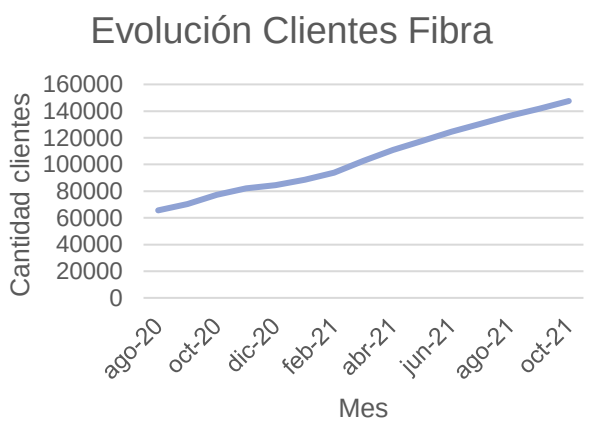
La empresa se desarrolla y ofrece sus servicios en el sector industrial de las telecomunicaciones. Este rubro se caracteriza por ser desafiante debido a su constante y ágil evolución, donde la exigencia de las personas y la capacidad de estar conectados es permanente e incrementa cada día más. Este sector incluye todos los servicios y productos relacionados a datos y voz móvil, TV paga, banda ancha, telefonía fija, servicio TI y datos fijos empresariales. Esta industria también se caracteriza por presentar un dinamismo constante, donde innovar, transformarse y buscar elementos diferenciadores se ha vuelto clave para mantenerse en el mercado. A nivel país, esta industria genera ingresos brutos de $6.004 millones de pesos anuales al año 2019, lo que equivale a un crecimiento del 0,75% respecto al año anterior, un incremento que fue impulsado por la banda ancha, los servicios TI y la telefonía móvil.

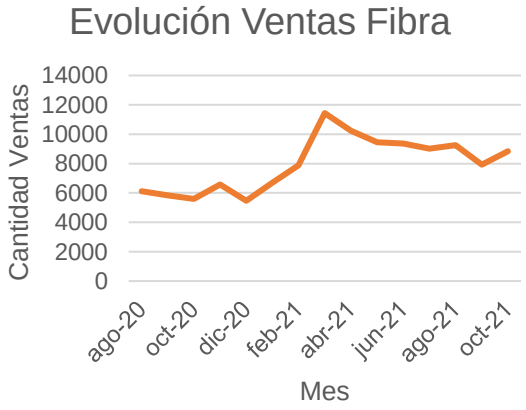
En los servicios de internet fijo, los principales actores del mercado corresponden a VTR y Telefónica, quienes poseen una participación de mercado del 33,9% y 26,9%, respectivamente. Otros actores relevantes son Claro, Mundo Pacífico y GTD.

La información descrita en este apartado se obtuvo de la Memoria Corporativa 2019 de la Empresa Nacional de Telecomunicaciones S.A. (Empresa Nacional de Telecomunicaciones, 2020) y la Subsecretaría de Telecomunicaciones de Chile (Subsecretaría de telecomunicaciones de Chile, 2020).

1.2 PROBLEMÁTICA DE LA EMPRESA

Para la realización de la memoria se trabajará en el mercado hogar con el servicio de internet fijo, específicamente con la tecnología de internet fibra óptica. Esta tecnología es bastante novedosa y ha presentado un crecimiento considerable en el último año, tanto en cantidad de clientes como en ventas del servicio (ver Gráficos 1 y 2), considerando especialmente el inicio de la pandemia COVID-19 y la cantidad de hogares que comenzaron a trabajar en modalidad de teletrabajo. El servicio de fibra óptica ofrece conexión fija a internet más rápida y con una mejor calidad de conexión que los servicios comunes de internet inalámbrico hogar. A modo de contraste, en febrero del año 2020 la empresa trabajaba con 43.958 clientes en fibra óptica y tan solo un año después, en febrero de 2021, poseía 100.728 clientes con este servicio, lo que equivale a un 230% de la base inicial de clientes de este mercado.





Gráficos 1 y 2: Evolución de base clientes y ventas de servicio fibra óptica entre agosto 2020 y octubre 2021.

La fibra óptica es una tecnología que para ser implementada requiere de una infraestructura específica, razón por la cual su factibilidad no se encuentra disponible en todas las zonas del país. Debido a esto, solo una cantidad determinada y reducida de hogares poseen factibilidad, la cual va expandiéndose a nuevas zonas cada mes. Esta tecnología comenzó a desplegarse en la compañía el año 2017 únicamente en la Región Metropolitana y, al año 2021, ya ha alcanzado una cobertura en más de 400 mil hogares del país.

El aumento de clientes del servicio de fibra óptica le entrega cada vez mayor peso a este servicio en la empresa, llevando atención y preocupación a lo que ocurre con la fuga de este (churn). El churn corresponde a un Key Performance Indicator (KPI) que se calcula tomando la cantidad de personas que cancelan el servicio a lo largo de un mes, dividido en el total de clientes que tienen contratado el servicio a inicios de este. Esta tasa

es calculada a inicios de cada mes, para así evaluar cómo varía en el negocio a lo largo del tiempo.

Existen dos tipos de fugas. La primera es denominada *fuga involuntaria,* la cual ocurre cuando la empresa decide cancelar el servicio con el cliente debido a razones que se encuentran por lo general vinculadas al no pago del servicio en los plazos estipulados. Por otro lado, la *fuga voluntaria* se refiere a aquella donde el cliente decide cancelar el servicio por su cuenta debido a distintas razones. En el trabajo de tesis se trabajará con la fuga conjunta de estas, principalmente debido a los deseos de la empresa de tener un modelo global que permita identificar cualquier tipo de cliente que dejará los servicios de la compañía, sumado a que la información de la fuga involuntaria previa a mayo no está correctamente etiquetada. En el Gráfico 3 se adjunta la evolución de la tasa de fuga total y voluntaria desde septiembre de 2020 hasta octubre de 2021, evidenciando que la mayor fuga corresponde a aquella de tipo voluntaria.

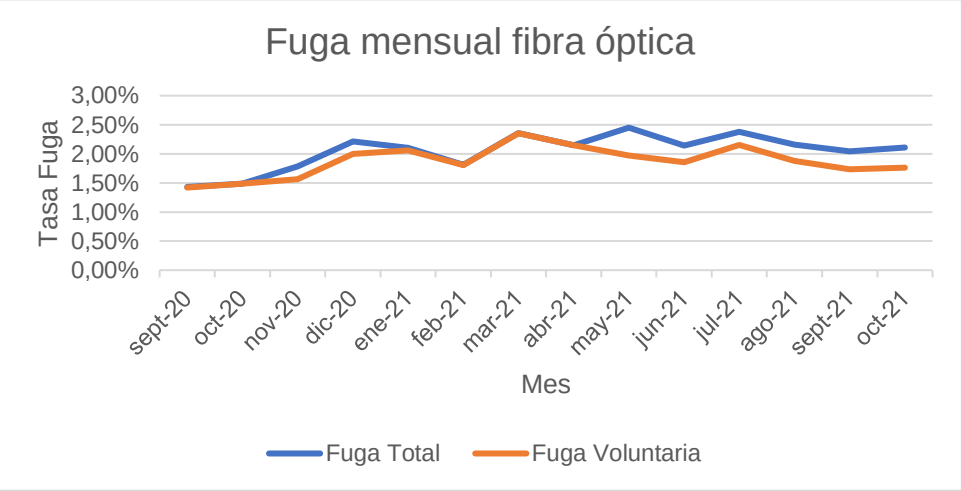


Gráfico 3: Evolución de fuga total y voluntaria para el servicio de fibra óptica entre septiembre 2020 y octubre 2021.

Al momento de evaluar económicamente el proyecto de fibra óptica se estableció que, dados los altos costos de instalación y mantención, la factibilidad y viabilidad de este quedaba sujeta a mantener una tasa de fuga total no mayor al 2%. Como se puede ver en el gráfico anterior, esto no se ha cumplido a cabalidad durante los últimos meses, lo que ha levantado alertas en el negocio. Tal así, que el 2% de fuga total que posee la compañía al mes de agosto del 2021 equivale a una pérdida de 2.700 clientes mensuales, lo que, sumado a la competitividad de la industria de las telecomunicaciones donde cada vez las ofertas se vuelven más agresivas, causa preocupación en distintas áreas de la empresa.

Se recalca de igual manera el alto costo de instalación del servicio de internet fibra en los hogares (el cual es cubierto por la empresa en su totalidad), un costo que en muchos clientes no se logra recuperar debido a la fuga precipitada del mismo. Uno de los grandes problemas con los proyectos relacionados al rubro de las telecomunicaciones es que el servicio de internet es considerado como un servicio básico en Chile, razón por la cual no es posible forzar la contratación del servicio por una cantidad mínima de meses, y, en consecuencia, el cliente puede elegir la cancelación del servicio en cualquier momento, lo que hace más difícil la viabilidad del proyecto al no poder definir un contrato con horizonte mínimo que el cliente deba permanecer con el servicio para retornar la inversión inicial realizada en ellos.

Hoy en día las políticas de retención se activan una vez que los clientes llaman solicitando cancelar el servicio, lo cual consiste en el ofrecimiento de distintos descuentos seleccionados en base al tipo de servicio que tiene contratado el cliente y la cantidad de competidores con factibilidad fibra en la dirección donde reside la persona. En caso de que el cliente insista en cancelar el servicio, se le pregunta para términos internos de la compañía la razón de su salida. Estas políticas de retención pueden llegar a ser ineficientes debido a que solo permiten reaccionar cuando la fuga ya está ocurriendo, momento en el cual el cliente ya ha tomado su decisión.

Entender los factores que intervienen en la decisión del cliente de cancelar el servicio se vuelve clave para saber cómo y con quiénes se debe comunicar la empresa para evitar que cancelen los servicios de la compañía y ejercer así campañas preventivas sobre ellos. Estas variables pueden ser de factores propios de los individuos (género, estatus socioeconómico, miembros del hogar, etc), factores de los servicios ofrecidos (precio, cantidad, calidad, etc) o también factores de la competencia (presencia de la competencia).

Buscando solucionar esta problemática, en los últimos años se ha propuesto utilizar modelos predictivos que permitan aprender de los datos existentes y el comportamiento de los clientes con la compañía para adquirir insights de negocio sobre aquellas variables que podrían relacionarse y afectar de cierta forma la fuga en el mercado, seleccionando también un score de fuga para cada cliente, pudiendo detectar y adelantarse a su comportamiento. Para esto se han utilizado una variedad de modelos de distinta índole, donde cada año se generan nuevos algoritmos que buscan mejorar el desempeño de estos, permitiendo adelantarse con mayor precisión a las acciones que tomará el cliente.

Por esta razón, para este trabajo de título se propone la realización de modelos estadísticos que sean capaces de predecir aquellos clientes más propensos a cancelar el servicio de internet fibra óptica, comparando el desempeño de cada modelo y pudiendo establecer aquellos que se comportaron de mejor forma ante esta problemática. Estos modelos le entregarán información importante a la empresa, permitiendo a futuro generar

campañas que le permitan a la compañía tener el panorama completo de sus clientes y los factores explicativos que podrían hacer que estos decidan cancelar los servicios.

Un desafío que se presentará al momento de generar los modelos predictivos se relaciona con el problema existente de clases desbalanceadas en los datos. Esto se debe a que la fuga mensual equivale a aproximadamente un 2% de la base total, generando así un desbalance entre la clase mayoritaria (clientes no fugados) y la clase minoritaria (clientes fugados) pudiendo generar un sesgo y una tendencia de parte del modelo hacía la clasificación de no fuga para la mayor parte de las observaciones buscando minimizar el error de predicción global del mismo.

Por último, los beneficios económicos que se obtiene reduciendo el churn son bastantes significativos para las empresas. En primer lugar, investigaciones han demostrado que el costo de adquirir un nuevo cliente es entre cinco y seis veces más alto que retener uno ya existente (Bhattacharya, 1998), razón por la cual el entendimiento del churn, la identificación de aquellos clientes que podrían cancelar su servicio y las políticas de retención se han vuelto parte crucial para todas las empresas y las áreas de Business Intelligence de las mismas.

Específicamente para la empresa, dada la cantidad de clientes que posee el mes de agosto de 2021, bajar la tasa de fuga en 1 punto porcentual equivaldría a que 1.400 clientes permanezcan contratando el servicio respecto a la situación actual que se está viviendo, lo que en el largo plazo acumularía una cantidad considerable de clientes y mantendría al proyecto dentro de los estándares de viabilidad establecidos. En términos de ganancias para la empresa, el área de Customer Value Management de la empresa estimó que bajar la tasa de fuga en tan solo un 0,3% permitiría mejorar el Valor Presente Neto (VPN) en casi $50.000 por cliente.

CAPÍTULO 2: LITERATURA PREVIA

A continuación, se mencionan algunas investigaciones y proyectos de títulos que se utilizaron como referencia para el presente trabajo.

En primer lugar, se identificó la realización de un proyecto similar pero enfocado en otro mercado de las telecomunicaciones y poniendo especial énfasis en las redes de contacto que poseen los clientes de telefonía y cómo estas impactan la decisión de fuga de estos, llegando a resultados que indicarían que el modelo que presenta mejor desempeño en esta problemática son los árboles de decisión (Pérez, 2014).

De igual manera, se han utilizado los árboles de decisión para evaluar la propensión de fuga en clientes pospago de televisión digital, donde se destacan los buenos resultados y simplicidad de interpretación que se obtiene con la utilización de estos tipos de modelos. (Contreras et al., 2017). Este estudio tiene la similaridad de evaluar un mismo mercado hogar, pero enfocándose en los servicios de televisión, el cual se diferencia bastante del mercado de internet de hogar, el cual en la práctica es más imprescindible para las personas lo que provocaría un comportamiento distinto y las razones de fuga podrían estar asociadas a otros factores.

También, en otra investigación se probaron cuatro algoritmos distintos para predecir el churn de una empresa de telecomunicaciones: Árboles de Decisión, Random Forest, Gradient Boosted Machine Tree “GBM” y Extreme Gradient Boosting “XGBOOST”. (Ahmad et al., 2019). De aquí se obtuvo que aquel que presenta mejor desempeño es el último mencionado, el cual es utilizado para predecir fugas a futuro.

Por otro lado, se han utilizado técnicas bastante llamativas para resolver el problema, por ejemplo, no solo incluir características de servicios y demográficas de las personas, sino que también agregar como nuevas características de las personas el score obtenido de otros modelos de predicción de fuga (regresiones logísticas, clasificaciones lineales, naive bayes, árboles de decisión, redes neuronales, support vector machines y algoritmos de data mining evolutivos) (Huang et al., 2012). Se concluye que utilizar estos resultados como nuevas variables predictoras mejoran los resultados de predicción, respecto a los modelos originales sin estas.

Otra técnica utilizada en este tipo de problemáticas es el Uplift, donde investigaciones han evaluado el desempeño de la predicción de churn a través de Uplift en contraste con la obtenida con modelos predictivos comunes, concluyendo que utilizando la métrica del aumento de beneficio máximo (MPU) el modelo de Uplift predice de manera más efectiva la fuga de clientes en la industria financiera. (Devriendt et al, 2021). De igual manera, se

han testeado tanto modelos de árboles de decisión como de redes neuronales, donde el primero muestra mejor desempeño para predecir la fuga de los clientes, además de destacar su facilidad de comprensión (Umayaparvathi & Iyakutti, 2012).

Igualmente se han probado distintos modelos de predicción, que incluyen J48, Random Forest, Regresión logística, OnetoR , Naive Bayes. De aquí se ha obtenido que el modelo de random forest fue aquel que presenta mejor accuracy y menor error. (Kumar, 2021). También se encontró que las redes neuronales convolucionales presentan mejor resultado y desempeño que otros algoritmos de machine learning utilizados previamente. (Chouiekh & Haj, 2020). Por otro lado, se ha utilizado un modelo de cadenas de Markov destacando su mayor adaptabilidad y facilidad de incluir nuevas variables. (Oludele et al, 2020).

En cuanto al problema de las clases desbalanceadas, en literatura previa se han utilizado dos metodologías para solucionar este problema. Por un lado, se tiene el undersampling que busca reducir la clase más grande y, por otro lado, el oversampling donde se duplica la clase más pequeña (Ahmad, 2019). El algoritmo de oversampling soluciona el problema de clases desbalanceadas generando una sobre muestra de la clase minoritaria que vendría siendo la fuga de clientes. Para esto se ha utilizado el algoritmo ADASYN, el cual muestra buenos resultados respecto al caso donde las muestras no son balanceadas (Faris, 2018). Al comparar el rendimiento de random sampling, advanced undersampling, gradient boosting model y weighted random forest, se obtuvo que la técnica de undersampling es aquella que ha mostrado mejor desempeño en ese trabajo (Burez & Van den Poel, 2008).

CAPÍTULO 3: OBJETIVOS

El objetivo general del trabajo de memoria se define como:

Generar modelos de propensión de fuga del servicio de internet de una empresa de telecomunicaciones, con el fin de analizar las características de dichos individuos y establecer el diseño de campañas de retención proactivas de los mismos.

Con el fin de organizar de mejor manera los resultados y entregables, se explicitan cinco objetivos específicos que se deben realizar para llegar a cumplir el objetivo general propuesto. Estos objetivos específicos para el trabajo de memoria son:

- Identificar variables que sean relevantes para predecir la fuga de los clientes.

- Evaluar los resultados de distintos modelos de propensión de fuga que establezcan una probabilidad de que el cliente deje la compañía.

- En base a los resultados del modelo, identificar aquellas variables relevantes para los clientes al momento de tomar la decisión de cancelar el servicio de la empresa, permitiendo caracterizar a los clientes propensos a fugarse y mejorar la toma de decisiones en campañas futuras.

- Proponer un experimento que permita medir la efectividad de generar acciones de interacción con el cliente por parte de la empresa en grupos propensos a fugarse.

CAPÍTULO 4: ALCANCES

- Se trabajará solo con datos de clientes que poseen contratación del servicio internet fibra óptica para esta empresa en Chile. Esto implica que las conclusiones y resultados encontrados no deben ser extendido a otros tipos de servicios ofrecidos por la misma empresa. No se trabajará con fugas de otros servicios de la compañía debido al comportamiento diverso que se espera que tenga cada mercado, especialmente considerando que este es un servicio que funciona para un ente familiar/hogar y no para una persona natural, como lo sería el servicio pospago donde la decisión de fuga viene dada a factores únicos de cada persona.

- Se trabajará con información que la empresa posee disponible desde agosto del año 2020 hasta octubre del año 2021, debido a que desde entonces se tiene información completa de los clientes y sus contratos.

- Debido al tiempo disponible para realizar la memoria, se evaluará las predicciones del modelo en una base de testeo que se utilizará para establecer aquel modelo más apto y, además, se evaluará con datos nueva para el mes previo a la finalización del informe, resultados que a futuro deben seguir midiéndose para evaluar posibles fallas predictivas de gran magnitud.

- No se puede asumir causalidad en las variables predictoras de fuga, debido a los posibles problemas de heterogeneidad presentes al no obtener los resultados directamente de un experimento.

- Debido a la complejidad económica que presenta la realización de un experimento, este no se llevará a cabo en el presente trabajo de título, sino que se dejará como una propuesta a futuro, estableciendo qué se debe ofrecer, cómo dividir la datos y lo que se podría obtener de los resultados. Quedará a decisión de la empresa la futura realización de este.

CAPÍTULO 5: MARCO TEÓRICO

5.1 MODELOS

Dado que se busca entender y trabajar con variables predictoras de fuga con datos disponible en la empresa, se utilizará la disciplina de Data Science para procesar los datos disponible y crear modelos que permitan establecer la probabilidad que presenta cada cliente a fugarse en base a características propias de este y su entorno.

Los modelos que se entrenarán presentan distintos pros y contras que se deberán considerar al momento de evaluar cada modelo y sus resultados, entre ellos el sobreajuste, la facilidad de interpretación y la complejidad para trabajar con los datos que presentan.

En base a la bibliografía leída, los modelos que mejor desempeño han presentado en trabajos previos para resolver este tipo de problemática y que se utilizarán en este trabajo de título son los que se describen a continuación:

1. Random Forest (RF): es una técnica de aprendizaje automático donde se combinan los resultados de diversos árboles predictores. Para esto se generan múltiples muestras de largo k provenientes de la base de datos completa, así como también de las variables de estas, tras lo cual se entrenan distintos modelos para cada muestra agregando los resultados finales. De esta forma se obtiene un estimador promedio con menos varianza, evitando también el sobreajuste respecto a los modelos de aprendizaje automático clásicos (Kroese et al, 2019).

2. Extreme Randomized Trees (XRT): es una técnica que entrena de manera similar a random forest, con la diferencia de que al escoger una característica y un valor de corte que divida al árbol no solo evalúa aquel que minimice el criterio de impureza, sino que además selecciona un valor aleatorio de corte para esa característica. Finalmente, se queda con la opción que generó una mejor división en el resultado final (Pedregosa et al, 2011).

3. Gradient Boosting Machine (GBM): técnica de aprendizaje automático que produce un conjunto de árboles de decisión y se construye de forma escalonada con el fin de optimizar los resultados de forma gradual, aditiva y secuencial. Utiliza la función de pérdida como una medida de qué tan buenos son los coeficientes del modelo en clasificar los datos (Kroese et al, 2019).

4. Stacked Ensemble: técnica de aprendizaje automático que utiliza un algoritmo de meta aprendizaje para elegir cómo combinar los resultados de distintos modelos de machine learning, logrando construir un nuevo modelo que proporcione mejor desempeño en el modelo general y las predicciones de este respecto a los modelos individuales. Este metamodelo hace uso de los datos para evaluar las predicciones que cada modelo realiza de estos y contrastarlos con los resultados esperados, logrando así la mejor combinación de los modelos originales. Este algoritmo es especialmente útil cuando se tienen distintos modelos que se poseen buenas habilidades predictoras, pero que son habilidades distintas entre sí, es decir, que las correlaciones entre las predicciones de los modelos son bajas (Brownlee, 2020).

5.2 MÉTRICAS

Los resultados de los modelos se evaluarán según distintas métricas, donde las más utilizadas en la disciplina del Data Science se mencionan a continuación.

En primer lugar, se contrastarán los modelos utilizando las distintas tasas de la matriz de confusión que se muestra en la Tabla 1.

Observación \ Predicción	Positivos	Negativos
Positivos	Verdaderos Positivos (TP)	Falsos Negativos (FN)
Negativos	Falsos Positivos (FP)	Verdaderos Negativos (TN)

Tabla 1: Matriz de confusión

En esta matriz se visualizan cuatro escenarios posibles:

- Falsos positivos (FP): casos donde el modelo predice fuga y no es fuga.
- Falsos negativos (FN): casos donde el modelo no predice fuga y sí es fuga.
- Verdadero positivo (TP): casos donde el modelo predice fuga y sí es fuga.

- Verdadero negativo (TN): casos donde el modelo no predice fuga y no es fuga.

Para obtener las tasas respectivas, se divide cada escenario en la cantidad de casos totales predichos en esa categoría (cliente se fuga o no se fuga). Aquí se pueden obtener una variedad de métricas como las que se mencionan a continuación.

El *recall* (sensibilidad) corresponde al porcentaje de la cantidad positiva que hemos sido capaces de identificar, es decir:

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

El *precision* nos permite evaluar la calidad de la predicción, estableciendo el porcentaje de los que el modelo predijo como fuga que realmente resultaron ser así, es decir:

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

El *accuracy* mide el porcentaje de casos que el modelo pudo clasificar de forma correcta, es decir:

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

Por otro lado, se considera el LIFT acumulado que presenten los distintos modelos, métrica que permite medir cuánto se aumenta la tasa natural de fuga en cada percentil de clientes. Esto permite por un lado evaluar si la predicción dejó a los clientes de grupos más propensos con un alto LIFT respecto a aquellos grupos menos propensos, o si esta diferencia fue casi nula, lo que sería un indicador de que el modelo no está dividiendo de forma correcta a los grupos.

También, se contrasta entre los distintos modelos los resultados del área bajo la curva ROC ($\boldsymbol{AUC_{roc}}$). La curva ROC establece la relación de las tasas de verdadero positivos y los falsos positivos, por lo que se busca que el área bajo la curva sea lo más grande

posible, donde un AUC de 0.5 representa un modelo que establece la fuga de clientes de forma aleatoria y un AUC de 1 representa un modelo que predice las fugas sin errores.

En la literatura previa se ha especificado que, dado los limitados recursos existentes de contactabilidad con clientes en las empresas del rubro de las telecomunicaciones, la capacidad de predecir la fuga se debe medir en base a la capacidad de identificar fugas reales entre el 0.1% y 5% más propenso a fugarse (Richter et al, 2010). Luego, se considera que los modelos deben medir su desempeño en base a su capacidad de predecir fuga entre estos clientes, argumento que se considerará en la parte de la evaluación de los modelos.

5.3 BALANCEO DE MUESTRA

Como se mencionó en el apartado anterior, la tasa de fuga es tan solo un porcentaje pequeño de la muestra general, siendo en promedio un 2% del total de la base de clientes, lo que desemboca en una muestra que se encuentra desbalanceada entre el grupo de clientes que se fugó y aquellos que no. Sin embargo, las muestras deben estar balanceadas con el fin de dar la misma importancia a ambas clases al momento de entrenar los modelos, en caso contrario el modelo tenderá a clasificar todos los individuos como la clase mayoritaria presentando de esta forma mayor cantidad de aciertos, pero a costa de perjudicar a la clase minoritaria.

Para solucionar el desbalanceo de clases existen distintas técnicas para trabajar los datos (Fernández et al, 2018). En este trabajo de título se probarán dos tipos de algoritmos para resolver esta problemática.

En primer lugar, se utilizará un método que duplica las observaciones de la clase minoritaria, el cual tiene por nombre *Random Oversampling* (*ROS*). Lo que este algoritmo hace es tomar una muestra aleatoria de la clase desbalanceada y las duplica una cierta cantidad de veces, logrando así emparejar la diferencia entre las distintas clases. Puede realizarse múltiples veces hasta que se obtenga la cantidad de muestra deseada para esa clase (Hoens & Chawla, 2013).

Por otro lado, existen técnicas de undersampling que se encargan de eliminar datos de la clase mayoritaria, donde se utilizará el método de *Random Undersampling (RUS*). Similar a ROS, este método selecciona muestras al azar de la clase mayoritaria y las elimina de la base original (Hoens & Chawla, 2013)..

CAPÍTULO 6: DESARROLLO METODOLÓGICO

Para el desarrollo del trabajo de título se utilizará la metodología Cross Industry Standard Process for Data Mining (CRISP DM), la cual consiste en 5 etapas: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Esta metodología fue seleccionada para el presente trabajo de título debido a que ha mostrado un buen desempeño resolviendo proyectos de Data Mining independiente del sector industrial y tecnología del que se hace uso, donde provee tanto la estructura como la flexibilidad necesaria para ejecutar distintos proyectos de esta índole (Wirth & Hipp, 2000).

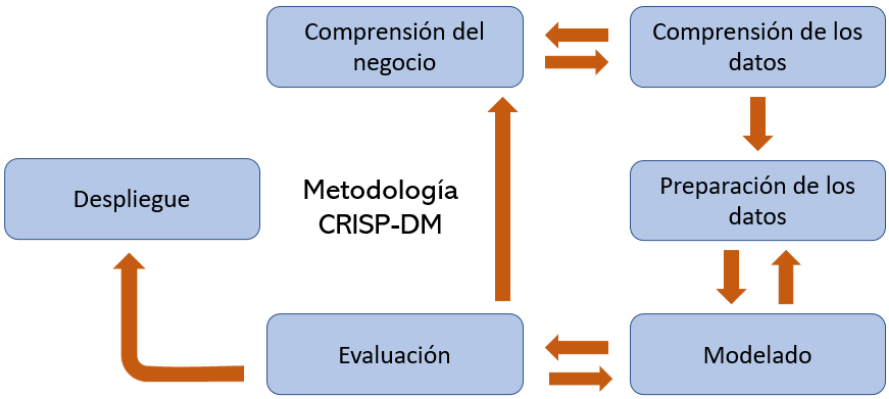


Figura 1: Metodología CRISP DM

Estas etapas se detallan y llevan a cabo en esta sección.

6.1 COMPRENSIÓN DEL NEGOCIO

Esta es la primera etapa, en la cual se busca comprender cómo funciona el negocio y la industria de las telecomunicaciones, con especial foco en lo que es el mercado hogar y el internet de fibra óptica en Chile. Con esto, se puede tener una idea de las cosas claves que podrían afectar la decisión de los clientes de cancelar o no su servicio contratado.

Es importante recalcar que el mercado hogar es bastante nuevo en la empresa, lo que lo hace un mercado desconocido y que se ha debido explorar de forma rápida y adaptándose a las situaciones cambiantes que se viven debido al aumento considerable de clientes que ha tenido la compañía para estos servicios. Por esta razón, la empresa ha dedicado muchos recursos a conocer cómo se comporta este nuevo mercado, qué es lo que les gusta y qué pueden hacer tanto para atraer más clientes, como también para retenerlos.

Según las investigaciones realizadas en conversaciones con ingenieros de la empresa, las variables que se ha observado en el último tiempo que podrían estar relacionadas con la fuga son las incidencias que han reportado los distintos tipos de packs que se ofrecen, considerando especialmente que el servicio de fibra óptica se vende en muchas ocasiones en packs dúos o triples, que incluyen televisión y/o telefonía móvil. La empresa ha reportado muchas caídas y problemas especialmente con el servicio de televisión, lo que se traduciría en consumidores más insatisfechos que aquellos que tienen contratado un servicio únicamente de internet hogar. Por otro lado, en la empresa se ha identificado que aquellos que se contactan con esta para reportar los problemas sí presentarían una mayor propensión hacia la fuga debido a una posible disconformidad con el servicio entregado.

Además, se encontraron comentarios de parte de trabajadores de esta área sobre de la importancia que tendría la comunicación y la forma en que se interactúa con el cliente al momento de que estos decidan cancelar su servicio. En el caso del mercado hogar, más de la mitad de las interacciones post venta con los clientes son a través del canal de IVR (Interactive Voice Response), donde todas las llamadas que los clientes realizan quedan registradas. Estas hipótesis se generan bajo la creencia de que aquellos clientes que más llaman podrían estar más molestos con el servicio. De la misma forma, se espera que los clientes a quienes se les contacta con el fin de saber cómo funciona el servicio presentarían una mejor evaluación del servicio ofrecido al sentir un acompañamiento y preocupación por parte de la empresa. Es por estas razones que se desea evaluar específicamente cómo la forma de interactuar con los clientes puede llegar a ser una variable importante para determinar la fuga de estos. Con este trabajo se pondrá especial énfasis y buscará contestar esta hipótesis planteada.

Una de las causas posibles del aumento en la tasa de fuga se puede asociar a lo imprescindible que se ha vuelto el internet del hogar debido a la pandemia COVID-19, la cual ha forzado a una cantidad considerable de familias a trabajar desde su casa, razón que ha vuelto necesario no solo poseer un buen servicio de internet, sino que también una conexión estable que permita a todos los integrantes del hogar trabajar de forma rápida y con un internet de mejor calidad. En base a conversaciones con trabajadores del mercado hogar, este hecho habría desembocado en clientes más intolerantes a fallas del servicio y más dispuestos a cambiarse de compañía por una que ofrezca mejores oportunidades, lo que se habría visto traducido en las alzas de las tasas de fuga.

Otro aspecto para comentar es que, al momento de contratar el servicio, los clientes obtienen un descuento inicial con una duración de seis meses, razón por la cual trabajadores de la empresa han notado que algunos clientes deciden cancelar su servicio cuando se les acaba este descuento y en muchos casos incluso contratarlo nuevamente al mes siguiente, obteniendo así el descuento por segunda vez.

En cuanto al comportamiento de fuga en la industria de las telecomunicaciones diversos estudios han concluido que existen tres tipos de variables que podrían incidir en la fuga:

- Variables relacionadas con el consumo del servicio, como uso del internet y tipo de pack contratado (Kisiogly & Topcu, 2011) (Hou et al, 2021).

- Variables estadísticas y propias de las personas, como la edad, los ingresos y la composición del grupo familiar (Verhoef, 2003) (Reinartz & Kumar, 2003) (Huang et al, 2012).

- Variables relacionadas con la empresa, como la calidad del servicio ofrecido y otros servicios que el cliente tiene contratado con la empresa (Rehman et al, 2016) (Su et al, 2019).

6.2 COMPRENSIÓN DE LOS DATOS

En esta sección se analiza que datos se poseen, que datos no se tiene y que datos son posibles de conseguir tanto dentro como fuera de la empresa. Es importante no solo analizar las tablas de información disponibles, sino que también comunicarse con otros empleados de la empresa que podrían tener conocimiento o acceso a otras tablas de utilidad para el modelo.

Entender los datos con los que se dispone y la calidad de estos es clave para comprender y comunicar los resultados que se esperan lograr y los posibles problemas que se pueden generar respecto a la idea inicial de proyecto propuesta.

En base a lo recopilado, la empresa posee múltiples bases de datos que se encuentran disponibles en distintos gestores de información y poseen datos desde agosto del año 2020 hasta octubre de 2021. Aquellas bases que se analizarán y evaluará la posibilidad de inclusión con su respectiva descripción se detallan a continuación:

- **Base de clientes fibra**: contiene todos los clientes que tienen contratado el servicio de fibra óptica al día 28 de cada mes. Contiene información de la dirección del cliente, el tipo de pack que posee y fecha de compra del producto.

- **Base de fugas**: contiene todas las cancelaciones del servicio fibra óptica que se realizaron cada mes, con la información del tipo de fuga que fue y el id de contrato al que correspondía.

- **Base competencia**: contiene la factibilidad del servicio de internet de seis empresas de la industria para cada dirección del país.

- **Base clientes móvil**: contiene la información de los servicios móviles que tiene contratado en la empresa cada persona, así como la cantidad de líneas telefónicas de la que es titular, el plan que posee contratado y la antigüedad.

- **Base malla parental**: posee información del hogar de gran parte de los chilenos. Aquí se tiene las relaciones familiares directas de las personas, así como el id de cada hogar e información estadística de cada individuo al año 2021.

- **Base bienes raíces:** contiene las propiedades que existen y sus respectivos dueños reportados al año 2020, así como información del avaluó fiscal de las mismas.

- **Bases automóviles inscritos**: posee la información de los automóviles que están inscritos en Chile y sus respectivos dueños al año 2020.

- **Base cartográfico**: posee información en base a las direcciones, donde se reporta el grupo socioeconómico y la comuna a la que pertenece.

- **Descuentos retenciones**: posee todos los descuentos que se han entregado para el servicio de fibra, así como su duración, el monto y la fecha en que el descuento termina.

- **Base interacciones**: contiene los tickets generados de las llamadas recibidas por los clientes al IVR, con la especificación del motivo de la llamada y la persona que la realizó.

- **Base profesiones**: contiene la información de las personas sobre sus profesiones reportadas, lugar de estudio y año de titulación.

- **Base gabinetes**: posee la información de la ocupación mensual para cada gabinete del que se conecta el servicio fibra, pudiendo establecer cuántas puertas tiene ocupadas y libres en cada periodo.

6.3 PREPARACIÓN DE LOS DATOS

En base a los análisis realizados en las secciones anteriores, se procesan los datos para dejarlos en un formato acorde a lo que se requiere en el modelo. Para esta parte se utiliza el lenguaje SQL, el cual permite establecer consultas de forma amigable e interactuando con varias tablas simultáneamente.

Haciendo uso de las bases detalladas en la sección anterior, se describen las variables que se posee y/o fueron creadas en la Tabla 24, disponible en la sección Anexos de este informe.

Se poseen datos desde agosto de 2020 hasta octubre de 2021, lo que en total genera 1.2 millones de observaciones, de las cuales 23.492 corresponden a fugas.

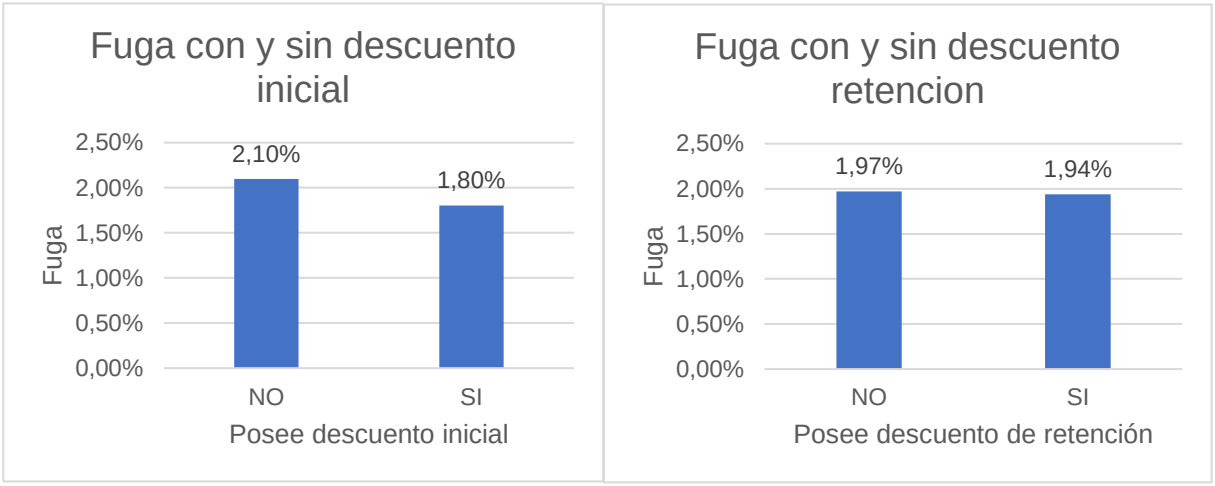
Para hacer el cálculo de las fugas, se comienza con la base de clientes fibra a fin de mes y se calcula cuánto porcentaje de estos clientes iniciales cancelan el servicio al mes siguiente. La cantidad de fugas que se tienen cada mes utilizando solo un mes de desfase se detallan en la Tabla 2.

Mes	Cantidad de clientes	Fugas mes siguiente	Fuga porcentual
Agosto 2020	65.604	939	1,43%
Septiembre 2020	70.439	1.049	1,49%
Octubre 2020	77.216	1.378	1,78%
Noviembre 2020	82.154	1.819	2,21%
Diciembre 2020	84.472	1.778	2,10%
Enero 2021	88.682	1.606	1,81%
Febrero 2021	93.958	2.216	2,36%
Marzo 2021	102.764	2.208	2,15%
Abril 2021	110.920	2.718	2,45%
Mayo 2021	117.505	2.516	2,14%
Junio 2021	124.438	2.958	2,38%
Julio 2021	130.433	2.813	2,16%
Agosto 2021	136.611	2.793	2,04%
Septiembre 2021	141.720	2.989	2,11%

Tabla 2: Fuga mensual utilizando 1 mes de desfase.

Se desea también observar cómo se comportan los clientes que poseen un descuento activo y cómo se relaciona con la probabilidad de fuga. En primer lugar, se observa que

un 46% de la base aún posee el descuento inicial de seis meses, y, a la vez, un 27% de la base posee algún otro tipo de descuento de retención otorgado. La relación que estas variables tienen con la tasa de fuga se observa en los Gráficos 4 y 5. Se observa que aquellos clientes que aún poseen el descuento inicial poseen menor tasa de fuga, sin embargo, se debe tener en consideración que esto podría ocurrir debido principalmente a que son clientes más bien nuevos, por ende, no puede asociarse su permanencia con el descuento mismo. Por otro lado, la tenencia de un descuento de retención parecería no variar en gran medida la probabilidad de fuga, ya que, pese a tener un descuento, hay que considerar que son clientes que ya manifestaron anteriormente su intención de irse, por ende, siendo un grupo más propenso a querer cancelar el servicio nuevamente.

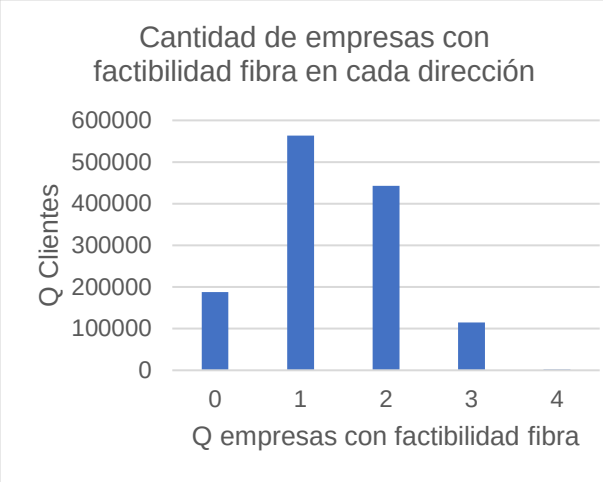


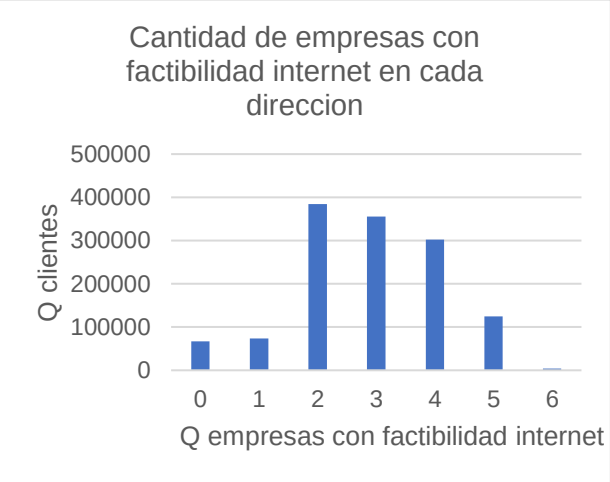
Gráficos 4 y 5: Fuga en función de la posesión o no de un descuento inicial sobre el precio del servicio o un descuento de retención sobre el mismo.

Es importante mencionar que existen ciertos datos que presentan fechas incorrectas, especialmente considerando que el servicio de fibra óptica comenzó a venderse desde el año 2017. Existen 869 datos que poseen una antigüedad mayor a 53 meses, los cuales representan errores de tipeo. El Gráfico 23 que se encuentra en la sección de anexos muestra la cantidad de clientes que presenta la base en función de la antigüedad desde que contrataron el servicio, dejando fuera las observaciones ya mencionadas que presentaban fechas erróneas. Se observa que la distribución es la esperada, donde la mayor parte de los clientes poseen pocos meses de antigüedad.

Por otro lado, al analizar la base de competencias, se observó que un 89% de clientes tiene factibilidad de algún servicio de internet con alguna empresa de la competencia. Los Gráficos 6 y 7 muestran cómo se distribuye la disponibilidad de opciones de la competencia que tienen los clientes tanto para el servicio fibra óptica, como también para algún tipo de internet hogar. Se puede notar que, dado que la fibra óptica es una tecnología bastante nueva, son pocas las empresas que han llegado a expandir su red y

factibilidad, lo que se traduce en que en la mayoría de los casos las personas solo pueden elegir entre 1 o 2 empresas de la competencia si desean cambiar su servicio. En cambio, si analizamos las opciones de internet tanto alámbrico como inalámbrico se puede ver que las opciones se incrementan, lo que podría aumentar también la probabilidad de fuga en aquellos clientes que estén dispuestos a contratar un servicio distinto a fibra.





Gráficos 6 y 7: Distribución de la cantidad de empresas con factibilidad fibra e internet en cada dirección respectivamente.

Otro aspecto importante para considerar es que un 46% de los clientes de fibra presentan también un servicio de telefonía móvil pospago en la compañía, lo que mostraría una cierta fidelización con la empresa. Más aún, un 21% de los clientes fibra es titular de más de 1 línea de teléfono en la compañía. Es importante entonces analizar si la fuga varía en función de si la persona posee también contratado un servicio externo al internet hogar en la empresa. El Gráfico 8 muestra cómo varía la tasa de fuga en cada mes, pero diferenciando por clientes móviles y no clientes móviles.

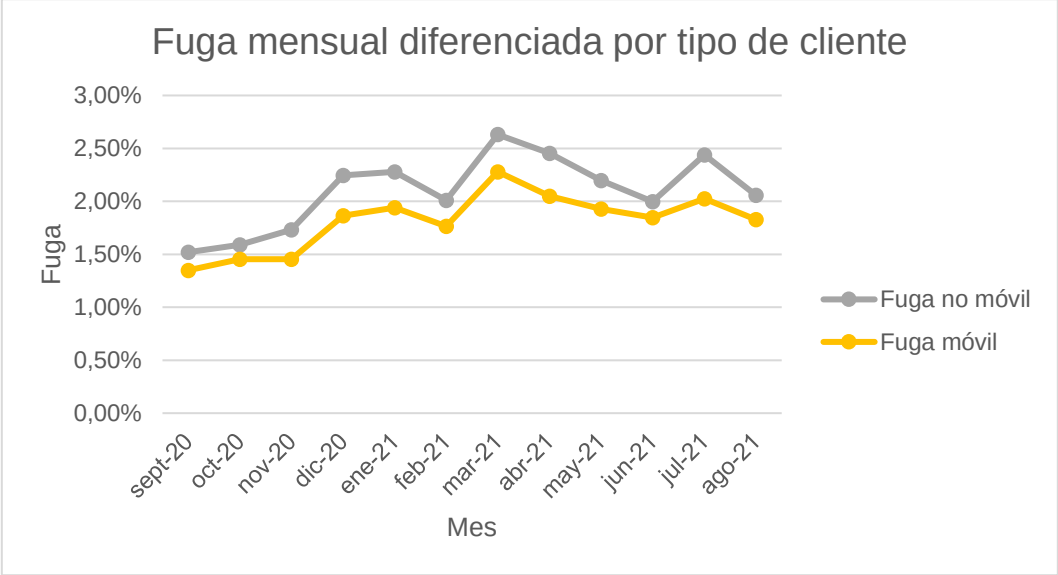


Gráfico 8: Fuga mensual del servicio de internet fibra óptica diferenciando por clientes móviles y no móviles.

Se puede visualizar que la fuga de los clientes móviles es siempre menor a la de clientes que no tienen servicio de pospago con la compañía, con un promedio de 0,28% menos de fuga en el primer grupo, donde los meses de abril y julio de 2021 incluso se genera una brecha superior al 0,4% entre estos. Este hecho da fuerza a la hipótesis planteada inicialmente de que existiría una fidelización con la compañía, traduciéndose en una mayor tolerancia a las posibles fallas o problemas y así una menor fuga para este grupo.

Otro tema que fue levantado en la etapa de comprensión del mercado es la importancia que podrían tener los llamados recibidos de los clientes por distintos motivos. Por esta razón se examina la base de interacciones que contiene todos los tickets generados a través del IVR, donde se obtiene que mensualmente en promedio 28.500 clientes llaman, generando un promedio de 2,3 tickets de atención mensuales cada uno. La cantidad de clientes que se ha contactado cada mes se observa en el Gráfico 9.

Gráfico 9: Cantidad de clientes que generaron tickets en el servicio de call center en cada periodo.

Estos números recalcan la importancia que pareciera tener la contactabilidad con los clientes en este rubro, ya que en promedio casi un 30% de la base total de clientes se comunica con la compañía por distintos motivos. En cuanto a los motivos de las llamadas, en el Gráfico 10 se visualiza cómo se distribuyen en relación a los tickets totales generados. La categoría que más recibe llamados se relaciona con las solicitudes, equivalente al 45% de las llamadas que se han recibido, seguido de los problemas, las consultas y en mucha menor medida los reclamos.

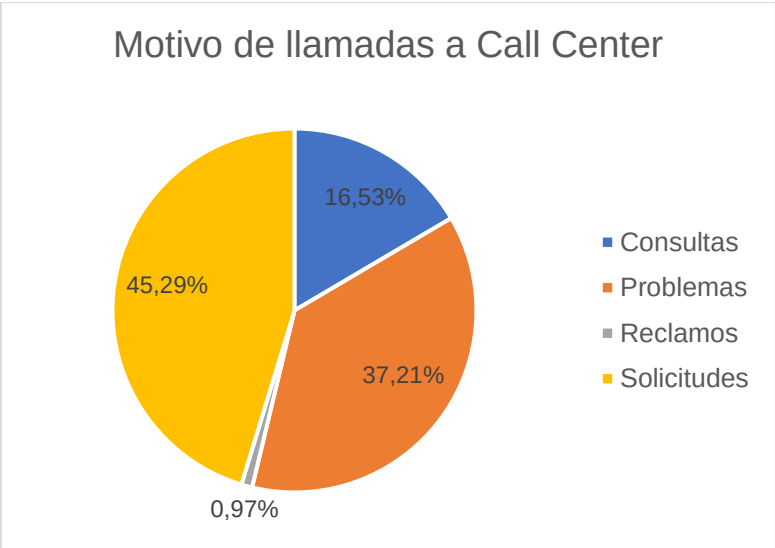


Gráfico 10: Motivos de tickets generados a través del servicio IVR.

Dado que el internet es un servicio que utilizan todos los integrantes del hogar es conveniente analizar el comportamiento de los hogares de los clientes. En base a los datos se obtiene que el promedio de edad de los hogares es de 38 años. La distribución de la cantidad de miembros de los hogares de los clientes se muestra en el Gráfico 24 que se encuentra en anexos, donde en promedio los hogares están compuestos de 2,8 personas. Se observa también que la tasa de fuga pareciera está correlacionada negativamente con la cantidad de miembros del grupo familiar, donde aquellos hogares en los que vive más personas tienden a tener tasas de fugas menores respecto a los hogares de una sola persona.

Por otro lado, en el Gráfico 11 se muestran las diez comunas con mayor cantidad de clientes fibra óptica en la empresa, donde se visualiza su distribución y la tasa de fuga presente en cada una. Pareciera ser que la fuga varía dependiendo de la comuna en que se encuentre el cliente, probablemente debido tanto a la situación socioeconómica como a las caídas o problemas con el servicio entregado en cada sector. Al momento de la investigación de mercado también se destacó que existían zonas más propensas a la caída del servicio, principalmente debido a ser zonas o comunas que presentan más disturbios, propiciando así el corte repentino de los cables de fibra.

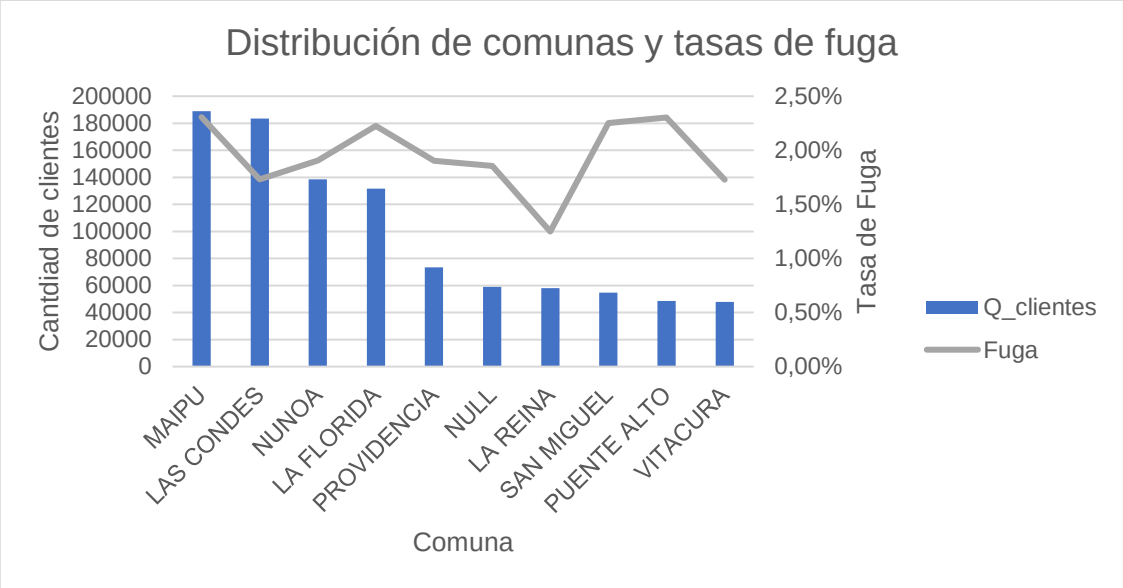


Gráfico 11: Distribución de frecuencia en las diez comunas con mayor cantidad de clientes y sus respectivas tasas de fugas.

Tomando en consideración que el precio de los servicios siempre ha sido una variable importante para los clientes, se procede a analizar cómo se comportan los distintos grupos socioeconómicos con la fuga del servicio. En el Gráfico 25 que se encuentra en anexos se observa que el grupo socioeconómico más alto resulta menos propenso a fugarse respecto a los otros, lo cual sigue la lógica ya establecida.

Como ya fue mencionado, el último año estuvo marcado por la pandemia COVID-19 y la explosión que esta produjo en la tasa de clientes del mercado fibra óptica, por esta razón es conveniente evaluar si este aumento de la tasa de fuga podría deberse más bien a la llegada de clientes con comportamientos distintos a los que ya existían en la empresa, los cuales podrían ser más propensos a fugarse que los que ya poseían el servicio. En base a esto, en el Gráfico 26 (ver anexos) se desglosa los comportamientos de fuga para clientes que tenían contratado el servicio previo a la llegada de la pandemia a Chile y aquellos clientes nuevos, considerados todos aquellos que contrataron el servicio posterior a febrero de 2020. Los resultados que se observan en el gráfico parecieran indicar que esta hipótesis no se cumple en el sentido esperado, donde incluso son aquellos clientes "antiguos" quienes presentan mayores tasas de fuga en los periodos analizados.

Es importante resaltar que existen distintos tipos de packs que se pueden contratar en fibra óptica, donde se puede incluir servicio de televisión o de telefonía pospago. Los distintos tipos de packs que se venden con sus respectivas tasas de fuga se muestran en el gráfico 12. Aquí se puede ver que los packs que presentan fuga más alta corresponden a DUO OTT (pack que incluye servicio de internet fibra óptica y televisión) y 3PLAY OTT (pack que incluye servicio de internet fibra óptica, televisión y telefonía), los cuales corresponden a packs que incluyen servicio de televisión, que como se comentó en la investigación de mercado es el servicio que presenta más incidencias y problemas. La descripción del contenido de cada pack ofrecido se detalla en la Tabla 25 (ver anexos).

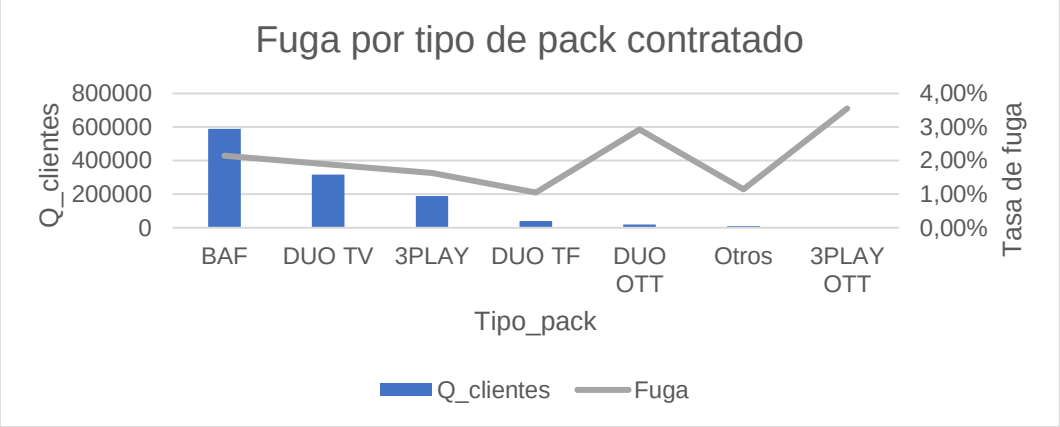


Gráfico 12: Fuga en el servicio de fibra óptica diferenciando por tipo de pack contratado.

Las propiedades que poseen los clientes también podrían ser un predictor de fuga. En el Gráfico 13 se evaluó si la dirección sobre la que se contrató el servicio es propiedad del cliente o un familiar directo, siendo alguno de estos los dueños de la propiedad o no.

Se observa que los clientes que son dueños de la propiedad sobre la que contratan el servicio son menos propensos a cancelar este respecto a aquellos que no. Además, en el Gráfico 14 se observa la fuga según si el cliente posee o no un automóvil inscrito a su nombre. Por último, en el Gráfico 27 (ver anexos) se muestra cómo varía la tasa de fuga según la cantidad de propiedades que posee el cliente y su cónyuge, donde pareciera existir una correlación negativa entre ambas variables.

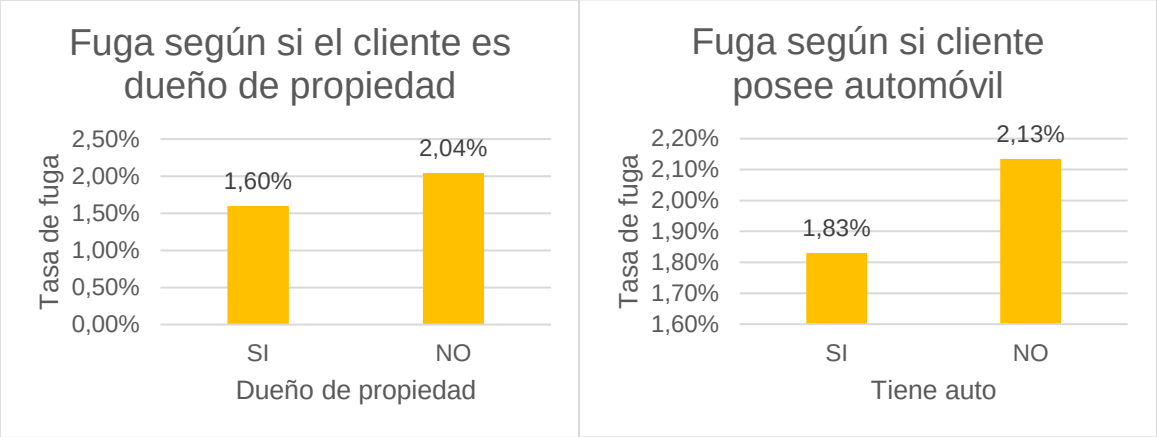


Gráfico 13 y 14: Tasa de fuga diferenciando si el cliente es dueño de la propiedad sobre la que tiene contratado el servicio de fibra óptica y si es propietario de un automóvil.

Por último, se establecen ciertas fallas o datos no encontrados en las bases originales que han conllevado a que existan datos nulos en estas. En primer lugar, un 2,6% de los datos no poseen un identificador de dirección factible, lo que lleva a que estos datos no posean información propia del lugar como la comuna o el grupo socioeconómico del hogar. Por otro lado, un 0,05% de los datos no poseen un identificador de contrato factible lo que imposibilita la identificación de detalles de la venta como la posible fuga, lo cual podría ser un punto ciego de varías fugas no detectadas. Además, un 16,3% de los datos no tienen información del identificador del hogar lo que no permitiría saber los miembros del grupo familiar y cómo se compone el mismo. De este grupo anterior, un 61,9% no son ni siquiera detectados en la malla parental, no pudiendo establecer información personal de la persona, como edad o sexo. Por último, un 48,9% de los datos no poseen información del canal a través del cual se realizó la venta.

6.4 EVALUACIÓN DE MODELOS DE PREDICCIÓN

6.4.1 MODELADO

En esta sección se procederá a trabajar con el lenguaje de programación R para ejecutar distintos modelos, los cuales fueron explicados en la sección de marco conceptual de este informe. Con el fin de poder evaluar cómo se comporta el modelo para predecir la fuga en datos con la que no se ha trabajado antes, se procederá a dividir la base de datos que se obtuvo de la sección *preparación de los datos* en dos conjuntos: entrenamiento y testeo. El primero incluirá todos los meses a excepción del último, el cual será usado para la base de testeo en la sección de *evaluación*.

Se utilizará la información y características de los individuos en un mes determinado para predecir si el cliente se fugó al mes siguiente o al subsiguiente, es decir, evaluando la fuga acumulada a dos meses (ver Figura 2). Se decidió utilizar la fuga acumulada a dos meses debido a que permitía mayor espacio de acción para realizar acciones futuras de retención de los clientes, pero considerando también información y características más actuales y, por ende, más precisa del cliente para predecir.

Figura 2: Explicación variable dependiente utilizada en el modelo de predicción de fuga para el servicio de internet fibra óptica.

Se debe tener en consideración este salto de temporalidad que existirá, razón por la cual la predicción podría no ser tan exacta debido a que no se utilizará los datos del cliente del mismo mes que se espera o no que ocurra la fuga. En la sección de despliegue se detallan las observaciones a tener en cuenta para la gestión de campañas preventivas de fuga.

Tal como se comentó, la base final que contiene datos de clientes desde agosto de 2020 a octubre de 2021 se divide en dos conjuntos de datos que se describen en la Tabla 3.

Datos	Periodo del cliente	Meses de fuga	Q_datos
Train	Agosto 2020 – Julio 2020	Agosto 2020 – septiembre 2021	1.148.172
Test	Agosto 2021	Septiembre 2021 – octubre 2021	136.715

Tabla 3: Descripción de conjuntos de datos de entrenamiento y testeo para modelo de fuga internet fibra óptica.

Se prueban entonces distintos algoritmos de machine learning con el fin de evaluar el desempeño de cada uno ante la problemática planteada. Estos algoritmos son:

- ✓ Gradient Boosting Machine.
- ✓ Random Forest.
- ✓ Extremely Randomized Trees.
- ✓ Stacked Ensembled All Models.
- ✓ Stacked Ensembled Best of Family.

Además, buscando solucionar el problema de desbalanceo de clases comentado en las secciones anteriores, se prueba cada algoritmo de tres formas distintas en la base de entrenamiento, buscando así que ambas clases posean mayor similaridad entre ellas. Las tres metodologías de balanceo utilizadas serán:

- ✓ Base simple, es decir, poseerá todos los datos disponibles sin ninguna alteración.
- ✓ Base con algoritmo de random undersampling, donde se buscará tener la misma cantidad de datos en ambas clases.
- ✓ Base con algoritmo de random oversampling, donde se buscará disminuir la clase mayoritaria para que no supere tres veces la cantidad de observaciones de la clase minoritaria, teniendo en consideración el trade-off entre generar bases similares, pero evitando reducir de manera drástica la cantidad de datos que se utilizarán.

Esto hace un total de 15 modelos distintos que se entrenan para solucionar la problemática planteada. Las bases utilizadas en el entrenamiento del modelo para cada opción se detallan en la Tabla 4.

Modelo	Cantidad de variables predictoras	Cantidad de no fugas en train	Cantidad de fugas en train
Modelos sin balanceo	88	1.099.534	48.638
Modelos con oversampling	88	1.099.534	1.099.534
Modelos con undersampling	88	145.914	48.638

Tabla 4: Descripción de conjuntos de datos de entrenamiento utilizados para cada tipo de modelo.

6.4.2 RESULTADOS

6.4.2.1 MODELOS SIN BALANCEO

Como se comentó anteriormente, se entrenaron los cinco modelos sin algoritmo de balanceo con los datos disponibles en la base de entrenamiento. Tras esto, se evalúan las predicciones que estos realizan en los datos disponibles de la base de testeo, cuyos resultados se muestran en la Tabla 5, donde se contrastan los modelos según 7 métricas distintas. Las matrices de confusión de cada modelo se detallan en anexos.

Métrica	GBM	Random forest	XRT	Stacked Ensemble BOF	Stacked Ensemble AM
AUC	0.70	0,72	0,70	0,73	0,73
LIFT acumulado 1%	6,69	14,35	11.54	14,15	14,15
LIFT acumulado 5%	4,12	5,82	5,16	6,00	6,00
LIFT acumulado 10%	3,14	3,82	3,5	3,97	3,97
Accuracy	0.93	0.94	0.94	0.95	0.95
Precision	0.21	0.28	0.25	0.37	0.37
Recall	0.26	0.26	0.22	0.23	0.23

Tabla 5: Resultados en métricas para modelos de predicción de fuga sin algoritmo de balanceo de las clases. Se evalúan los modelos según el área bajo la curva ROC (AUC), el lift acumulado en el 1%, el 5% y el 10% más propenso, el accuracy, el precision y el recall.

Por otro lado, en las Tablas 6 y 7 se muestran las tasas de fuga en los distintos deciles de propensión en que fueron clasificados los clientes en cada modelo.

	Tasa de fuga				
Modelo	**Decil 1**	**Decil 2**	**Decil 3**	**Decil 4**	**Decil 5**
GBM	12,9%	6,2%	4,4%	3,8%	3,5%
Random Forest	15,7%	5,2%	3,9%	3,1%	2,9%
XRT	14,4%	5,1%	4,0%	3,4%	3,1%
Stacked Ensemble AM	16,3%	5,3%	3,9%	3,1%	2,7%
Stacked Ensemble BoF	16,3%	5,3%	3,9%	3,1%	2,7%

Tabla 6: Tasa de fuga en deciles de propensión 1 al 5, en base a la clasificación realizada por los modelos de predicción de fuga sin algoritmo de balanceo.

	Tasa de fuga				
Modelo	**Decil 6**	**Decil 7**	**Decil 8**	**Decil 9**	**Decil 10**
GBM	2,7%	2,4%	2,0%	1,9%	1,4%
Random Forest	2,4%	2,5%	2,0%	2,0%	1,4%
XRT	2,7%	2,6%	2,3%	2,1%	1,5%
Stacked Ensemble AM	2,6%	2,2%	2,0%	1,6%	1,3%
Stacked Ensemble BoF	2,6%	2,2%	2,0%	1,6%	1,3%

Tabla 7: Tasa de fuga en deciles de propensión 6 al 10, en base a la clasificación realizada por los modelos de predicción de fuga sin algoritmo de balanceo.

Se observan resultados distintos para cada modelo en cada métrica, donde los siguientes modelos se destacan como aquellos de mejor desempeño al no emplear un algoritmo de balanceo:

- Mejor desempeño AUC: Stacked Ensemble
- Mejor desempeño Lift acumulado 1%: Random Forest
- Mejor desempeño Lift acumulado 5%: Stacked Ensemble
- Mejor desempeño Lift acumulado 10%: Stacked Ensemble
- Mejor desempeño accuracy: Stacked Ensemble
- Mejor desempeño precision: Stacked Ensemble
- Mejor desempeño recall: GBM y Random Forest

6.4.2.2 MODELOS CON UNDERSAMPLING

Nuevamente, se entrenan cinco modelos, pero utilizando las bases que fueron procesadas previamente con un algoritmo de undersampling, lo que redujo la cantidad de no fugados en la base de entrenamiento. En la Tabla 8 se muestran los resultados que se obtienen al probar los modelos en la base de testeo. Las matrices de confusión que se obtienen de los modelos se detallan en anexos.

Métrica	GBM	Random forest	XRT	Stacked Ensemble BOF	Stacked Ensemble AM
AUC	0,72	0,70	0,69	0,72	0,72
LIFT acumulado 1%	8,25	8,29	6,94	8,39	8,62
LIFT acumulado 5%	4,79	4,62	4,1	5,01	5,00
LIFT acumulado 10%	3,55	3,37	3,05	3,64	3,61
Accuracy	0.42	0.68	0.65	0.88	0.87
Precision	0.06	0.08	0.07	0.14	0.14
Recall	0.84	0.61	0.61	0.38	0.39

Tabla 8: Resultados en métricas para modelos de predicción de fuga con la aplicación de un algoritmo de undersampling en la base de entrenamiento. Se evalúan los modelos según el área bajo la curva ROC (AUC), el lift acumulado en el 1%, el 5% y el 10% más propenso, el accuracy, el precision y el recall.

Al igual que la sección anterior, en las Tablas 9 y 10 se muestran las tasas de fugas en cada uno de los grupos ya establecidos.

	Tasa de fuga				
Modelo	**Decil 1**	**Decil 2**	**Decil 3**	**Decil 4**	**Decil 5**
GBM	14,6%	6,1%	4,3%	3,8%	3,0%
Random Forest	13,9%	5,8%	4,2%	3,6%	3,0%
XRT	12,5%	6,0%	4,7%	3,6%	3,3%
Stacked Ensemble AM	14,8%	6,1%	4,2%	3,6%	3,0%
Stacked Ensemble BoF	15,0%	5,9%	4,2%	3,8%	2,9%

Tabla 9: Tasa de fuga en deciles de propensión 1 al 5, en base a la clasificación realizada por modelos de predicción de fuga con un algoritmo de undersampling aplicado en la base de entrenamiento.

	Tasa de fuga				
Modelo	**Decil 6**	**Decil 7**	**Decil 8**	**Decil 9**	**Decil 10**
GBM	2,5%	2,1%	2,0%	1,5%	1,2%
Random Forest	2,8%	2,4%	2,2%	1,8%	1,5%
XRT	2,9%	2,5%	2,3%	2,0%	1,3%
Stacked Ensemble AM	2,5%	2,1%	2,0%	1,6%	1,3%
Stacked Ensemble BoF	2,4%	2,2%	1,9%	1,6%	1,2%

Tabla 10: Tasa de fuga en deciles de propensión 6 al 10, en base a la clasificación realizada por modelos de predicción de fuga con un algoritmo de undersampling aplicado en la base de entrenamiento.

Nuevamente, los siguientes modelos se destacan como aquellos de mejor desempeño al emplear un algoritmo de balanceo de undersampling:

- Mejor desempeño AUC: Stacked Ensemble y GBM
- Mejor desempeño Lift acumulado 1%: Stacked Ensemble
- Mejor desempeño Lift acumulado 5%: Stacked Ensemble
- Mejor desempeño Lift acumulado 10%: Stacked Ensemble
- Mejor desempeño accuracy: Stacked Ensemble
- Mejor desempeño precision: Stacked Ensemble
- Mejor desempeño recall: GBM y Stacked Ensemble

6.4.2.3 MODELOS CON OVERSAMPLING

Finalmente, se entrenan los cinco modelos utilizando las bases que fueron procesadas con un algoritmo de oversampling, lo que incrementa la cantidad de fugados en la base de entrenamiento. En la Tabla 11 se muestran los resultados que se obtienen al probar los modelos en la base de testeo. Nuevamente, las matrices de confusión que se obtienen de cada modelo se detallan en anexos.

Métrica	GBM	Random forest	XRT	Stacked Ensemble BOF	Stacked Ensemble AM
AUC	0,66	0,73	0,72	0,73	0,73
LIFT acumulado 1%	4,61	10,31	7,79	10,31	10,31
LIFT acumulado 5%	2,95	5,65	4,58	5,65	5,65
LIFT acumulado 10%	2,36	3,92	3,48	3,92	3,92
Accuracy	0.35	0.95	0.89	0.95	0.95
Precision	0.05	0.34	0.15	0.32	0.32
Recall	0.84	0.18	0.34	0.19	0.19

Tabla 11: Resultados en métricas para modelos de predicción con algoritmo de oversampling aplicado en la base de entrenamiento. Se evalúan los modelos según el área bajo la curva ROC (AUC), el lift acumulado en el 1%, el 5% y el 10% más propenso, el accuracy, el precision y el recall.

Los resultados de la tasa de fuga para cada uno de los grupos ya establecidos se muestran en las Tablas 12 y 13.

	Tasa de fuga				
Modelo	**Decil 1**	**Decil 2**	**Decil 3**	**Decil 4**	**Decil 5**
GBM	9,7%	6,2%	4,8%	4,3%	3,8%
Random Forest	16,1%	5,6%	4,2%	3,2%	2,7%
XRT	14,3%	6,4%	4,7%	3,5%	2,9%
Stacked Ensemble AM	16,1%	5,6%	4,2%	3,2%	2,7%
Stacked Ensemble BoF	16,1%	5,6%	4,2%	3,2%	2,7%

Tabla 12: Tasa de fuga en deciles de propensión 1 al 5, en base a la clasificación realizada por modelos de predicción de fuga con un algoritmo de oversampling aplicado en la base de entrenamiento.

	Tasa de fuga				
Modelo	**Decil 6**	**Decil 7**	**Decil 8**	**Decil 9**	**Decil 10**
GBM	3,6%	2,9%	2,5%	1,9%	1,4%
Random Forest	2,7%	1,9%	1,8%	1,5%	1,3%
XRT	2,5%	2,2%	1,9%	1,6%	1,1%
Stacked Ensemble AM	2,7%	1,9%	1,8%	1,5%	1,3%
Stacked Ensemble BoF	2,7%	1,9%	1,8%	1,5%	1,3%

Tabla 13: Tasa de fuga en deciles de propensión 6 al 10, en base a la clasificación realizada por modelos de predicción de fuga con un algoritmo de oversampling aplicado en la base de entrenamiento.

Nuevamente, los siguientes modelos se destacan como aquellos de mejor desempeño al emplear un algoritmo de balanceo de oversampling:

- Mejor desempeño AUC: Stacked Ensemble y Random Forest
- Mejor desempeño Lift acumulado 1%: Stacked Ensemble y Random Forest
- Mejor desempeño Lift acumulado 5%: Stacked Ensemble y Random Forest
- Mejor desempeño Lift acumulado 10%: Stacked Ensemble y Random Forest
- Mejor desempeño accuracy: Stacked Ensemble y Random Forest
- Mejor desempeño precision: Random Forest
- Mejor desempeño recall: GBM

6.4.3 CONTRASTE DE MODELOS

Una vez que se obtienen las métricas de cada modelo, se procede a evaluar los resultados y se determina qué modelo presentó mejor desempeño. Esto se muestra en la Tabla 14, donde cada fila representa una métrica específica, así como aquel algoritmo de entrenamiento que presenta el mejor desempeño, cuál algoritmo de balanceo utilizó y el valor que alcanzó en esa métrica específica.

Métrica	Modelo	Balanceo	Valor
AUC	Stacked Ensemble	Simple, oversampling	0.73
Lift acumulado 1%	Random Forest	Simple	14.35
Lift acumulado 5%	Stacked Ensemble	Simple	6.00
Lift acumulado 10%	Stacked Ensemble	Simple	3.97
Accuracy	Random forest, Stacked Ensemble	Oversampling	0.95
Precision	Stacked Ensemble	Simple	0.37
Recall	Gradient Boosting Machine	Oversampling, undersampling	0.84

Tabla 14: Modelos de mejor desempeño para cada métrica establecida, se muestra también el algoritmo de balanceo que posee y su valor alcanzado.

Como se observa en la tabla, no existe un modelo que haya superado a todos a modo global, por lo que no hay una única respuesta ante la pregunta de cuál es el mejor modelo, debido a que esto dependerá de lo que se busca lograr y para qué se quiere utilizar este a futuro. Por esta razón, esta interrogante se retomará y se contestará en la sección de despliegue.

6.4.4 VARIABLES MÁS IMPORTANTES

Se procede a determinar cuáles fueron las variables de mayor importancia para el entrenamiento del modelo. Se tomará el resultado del modelo de random forest sin algoritmo de balanceo, el cual mostró mejor desempeño general entre los modelos de árboles, donde se detallan las 10 variables más importantes en el Gráfico 15. Esta importancia de variables se obtiene directamente del modelo, donde se considera cuánto aportó cada una a la configuración de este, tomando en consideración si la variable permitió dividir algún árbol y, en caso de haberlo hecho, cuánto disminuyó el error del modelo global.

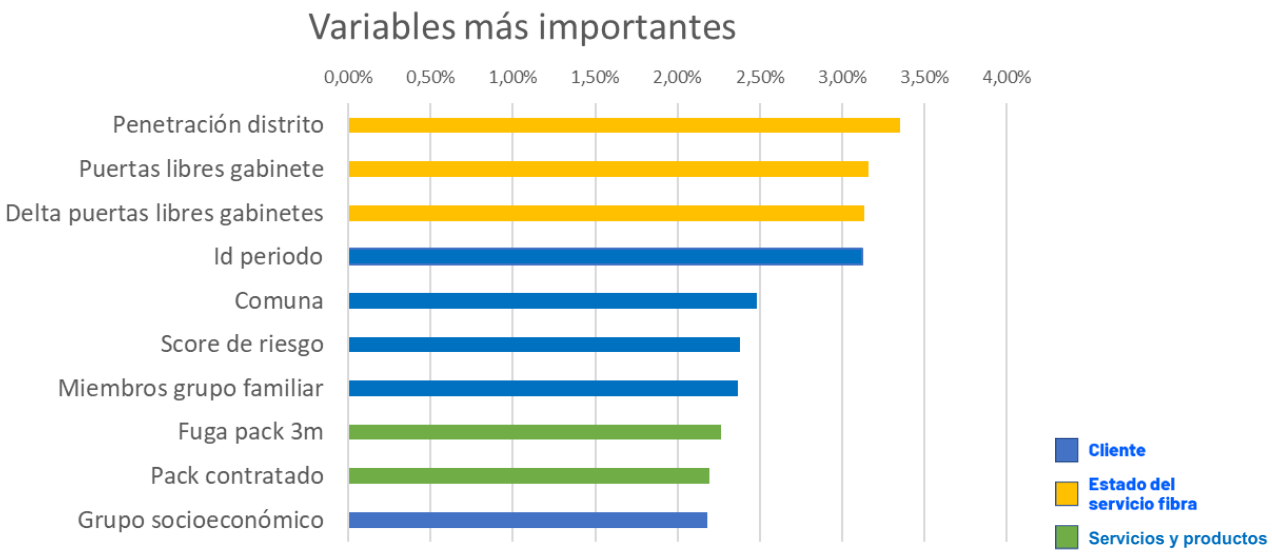


Gráfico 15: Variables más importantes seleccionadas por el modelo de random forest sin balanceo para la predicción de fuga.

En el gráfico se visualizan las variables clasificadas según tres colores. En azul se encuentran las variables descriptoras del cliente mismo, en amarillo se encuentran las variables que representan el estado del servicio fibra ofrecido y, por último, en verde se encuentran las variables que representan características del servicio que tiene contratado el cliente. Esta clasificación se realizó con el objetivo de evaluar las posibles ventanas de acción que podría tomar la empresa para reducir la fuga. Por ejemplo, aquellas variables azules no pueden modificarse ni cambiarse de parte de la empresa dejando fuera una oportunidad de acción con ellas. Por otro lado, variables de color amarillo o verde sí se relacionan con lo que la empresa hace o maneja, permitiendo que esta tome acciones y trabaje con ellas para poder disminuir la fuga a futuro.

Dado que una gran cantidad de variables importantes pertenecen a categorías con posibilidad de acción, los resultados son alentadores y muestran oportunidad de mejora al trabajar estas variables específicas.

Es importante mencionar también que ninguna variable presenta un porcentaje visiblemente mayor de importancia en comparación a las otras variables (a modo de ejemplo, la variable más importante no supera el 4%), lo que muestra resultados alentadores con un bajo sobreajuste del modelo hacia alguna variable específica.

Las 5 variables más importantes se mencionan y analizan a continuación:

1. **Penetración Distrito**: Aquella variable que obtuvo mayor importancia en el modelo fue la penetración de la empresa en el distrito. En el Gráfico 16 se muestra cómo se relaciona esta variable con la fuga, donde se notan dos tendencias claves. En primer lugar, tasas bajas de penetración de la empresa muestran altas tasas de fuga pudiéndose adjudicar a dos razones principales: por un lado, la calidad del servicio en esos distritos podría no ser buena llevando a que se tengan pocos clientes en la zona, y, por otro lado, podría existir una alta competencia, haciendo que existan mayores ofertas comerciales de los competidores e incrementando así la fuga del servicio. Por otro lado, se ve que al superar una penetración del 33% de clientes en el distrito la tasa de fuga se incrementa, que se podría asociar con el colapso del servicio, provocando que presente una peor calidad y llevando así a que clientes deseen cancelar el mismo. Otra hipótesis que se genera es que una alta tasa de penetración del servicio fibra muestra una alta disposición en el sector a contratar el servicio, provocando a la vez la llegada de nuevas competencias al sector y llevando a la fuga de clientes en la empresa.

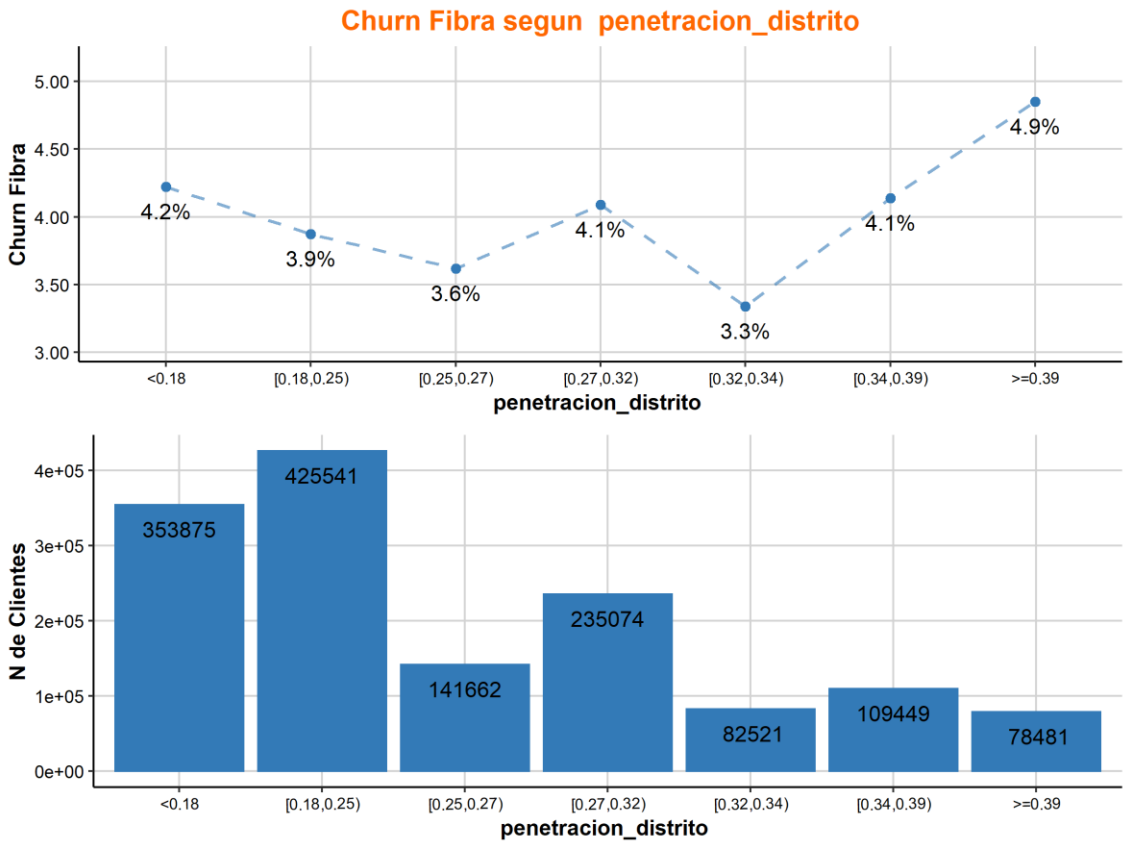


Gráfico 16: Distribución y tasa de fuga para la variable Penetración en Distrito en modelo de predicción de fuga.

2. **Total puertas libres en gabinete**: Por otro lado, la variable de puertas totales libres del gabinete muestra una tendencia a que mayor cantidad de puertas libres mayor es la tasa de fuga respectiva. Las razones que podrían ocultarse tras esto podrían relacionarse con la alta competitividad de la zona, donde esto se vería reflejado a través de la capacidad libre que tiene el gabinete de internet en esta. Por otro lado, la tasa más baja es aquella de los datos que no poseen información, los cuales resultan ser las filas de periodos más antiguos, relacionándose activamente con la antigüedad del cliente. Esta relación se puede ver en el Gráfico 17.

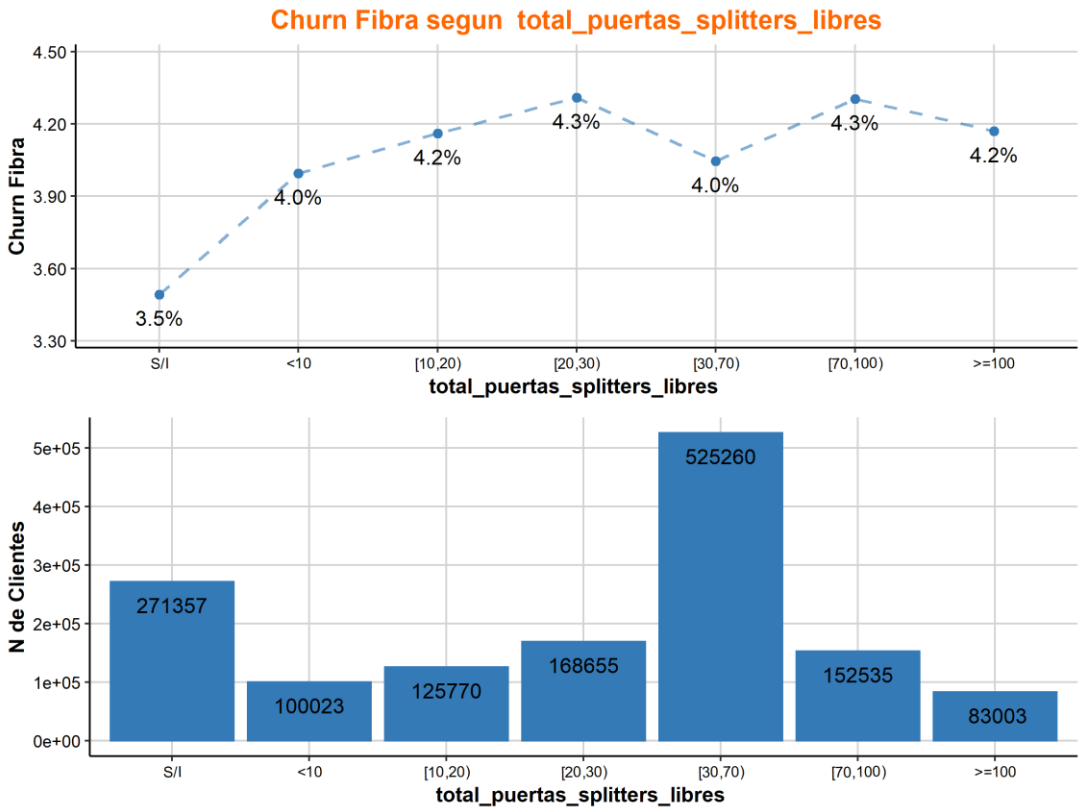


Gráfico 17: Distribución y tasa de fuga según variable de puertas libres de gabinete en modelo de predicción de fuga.

3. **Delta puertas libres gabinete**: En tercer lugar, se encuentra la variable que captura la variación de puertas libres en el gabinete entre cada mes. Por un lado, se ve el mismo fenómeno presente en la variable anterior, en donde los datos sin información presentarían una tasa de fuga baja, relacionado principalmente con la antigüedad del cliente. Además, se ven dos hitos claves en este gráfico. El primero muestra que cuando la cantidad de puertas libres baja en gran cantidad en contraste al mes anterior, la tasa de fuga es alta, lo que se puede relacionar con el aumento de clientes en la zona, llevando a un posible colapso del servicio y, por ende, una peor calidad del servicio entregado. Por otro lado, las tasas de fuga se incrementan al aumentar en gran cantidad las puertas libres entre un mes y otro, pudiéndose asociar a los esfuerzos de la empresa por mejorar la capacidad del internet en sectores que están teniendo problemas, los cuales podrían llegar tarde cuando el cliente ya ha tomado la decisión de cancelar el servicio. Este fenómeno también puede asociarse a que ya existen otros clientes en el sector que están cancelando el servicio de internet con anterioridad, liberando puertas en el gabinete, lo que podría deberse a problemas con el servicio entregado en la zona y llevar a que otros clientes también decidan hacerlo. La distribución de la variable se muestra en el Gráfico 18.

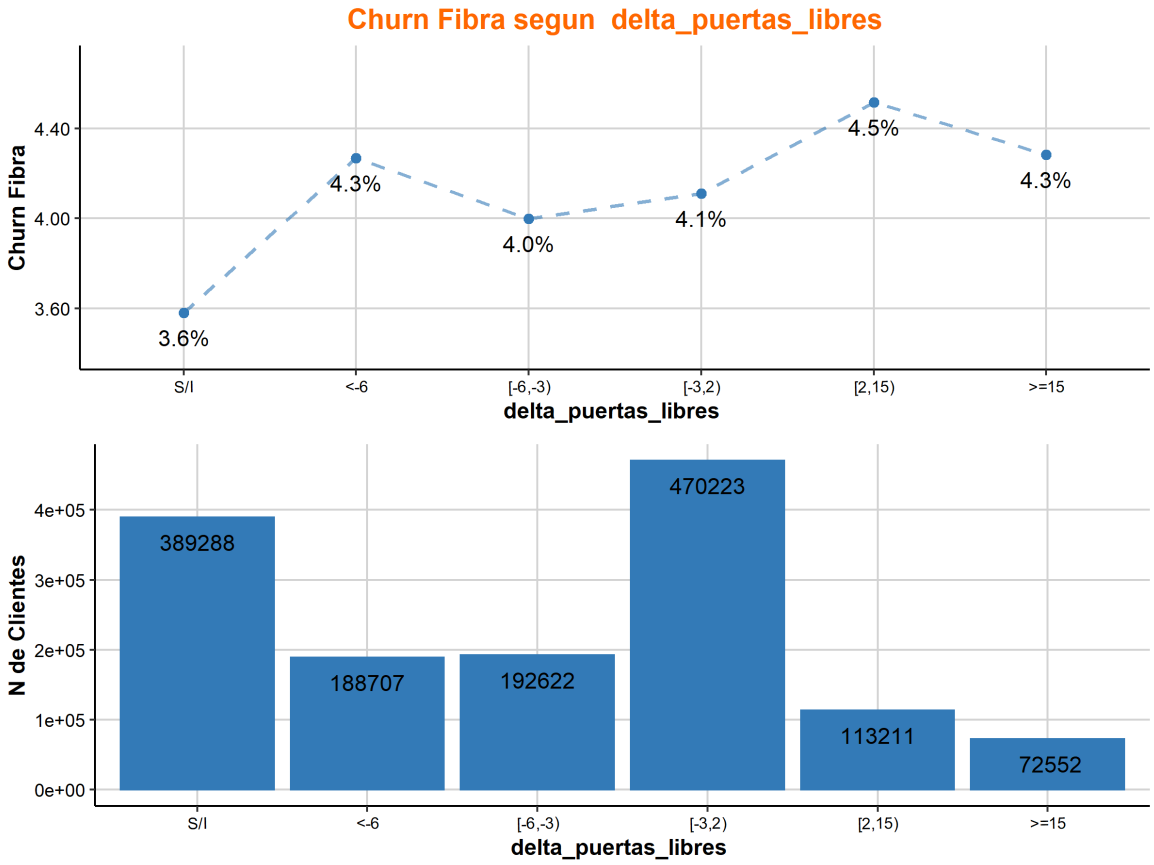


Gráfico 18: Distribución y tasa de fuga en variable de variación de puertas libres de gabinete en modelo de predicción de fuga.

4. **Id Periodo**: La cuarta variable de importancia en el modelo es una marca de qué tan antigua es la camada, la cual fue introducida como variable de entrenamiento con el objetivo de que el modelo la utilizara para discriminar y diferenciar las tasas de fuga que presenta el servicio a lo largo del tiempo. Por ejemplo, en las camadas más antiguas se tenía una menor tasa de fuga del servicio, la cual ha ido aumentando en el último año. Esto puede llevar a que características que en el pasado no provocaban la fuga de un cliente, hoy en día si pudieran llevar a una fuga de este debido a las distintas reacciones que tienen los clientes hacia distintas variables, las cuales van variando en el tiempo. La distribución de la variable se muestra en el Gráfico 19.

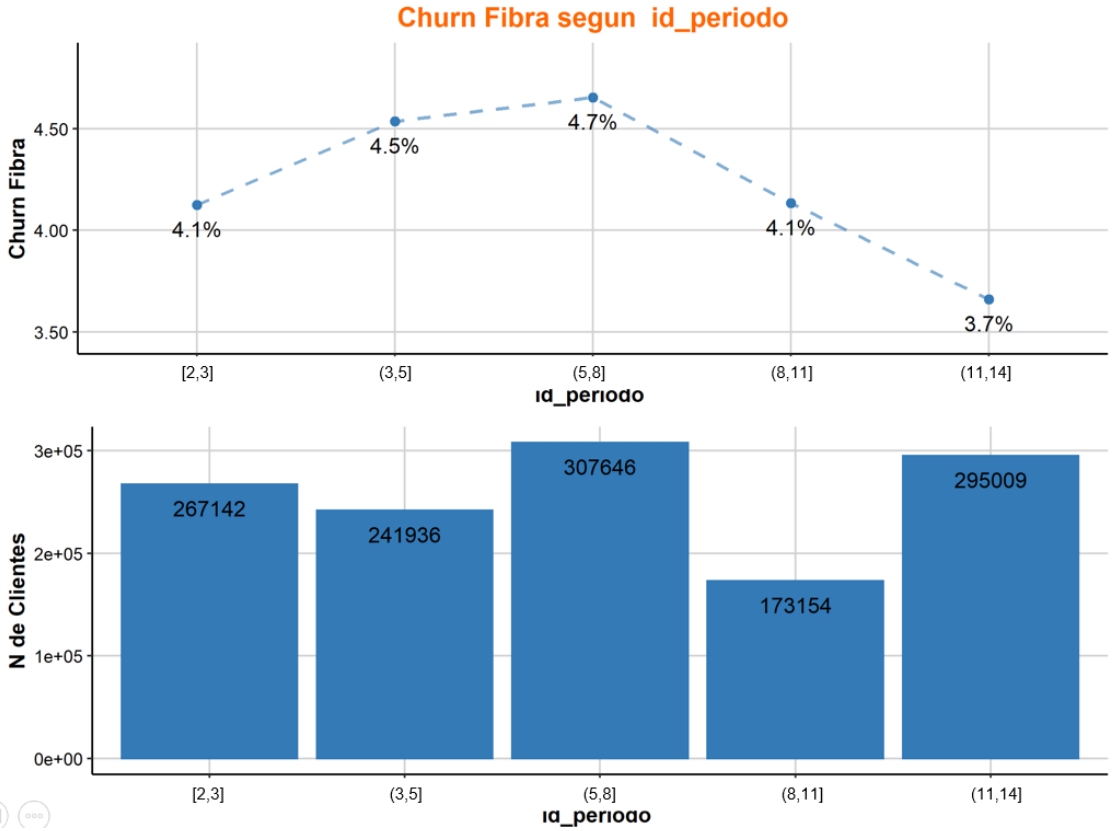


Gráfico 19: Distribución y tasa de fuga en variable id periodo de camada en modelo de predicción de fuga.

5. **Comuna**: Por último, se tiene la comuna del cliente. En el Gráfico 20 se observa la tasa de fuga respecto a la tasa promedio (4,3%) de las 8 comunas con mayor cantidad de clientes. En esta se puede notar la gran diferencia en la tasa de fuga que enfrentan algunas comunas. Es importante recalcar que esta variable posee también una alta correlación con la antigüedad del cliente, por ejemplo, la comuna de La Reina (quien posee la tasa más baja de fuga) es la comuna de mayor antigüedad en factibilidad de servicio con la empresa, mientras que la comuna de Maipú o Puente Alto tienen factibilidad del servicio fibra desde recién el año 2020, siendo así en su mayoría clientes más bien nuevos.

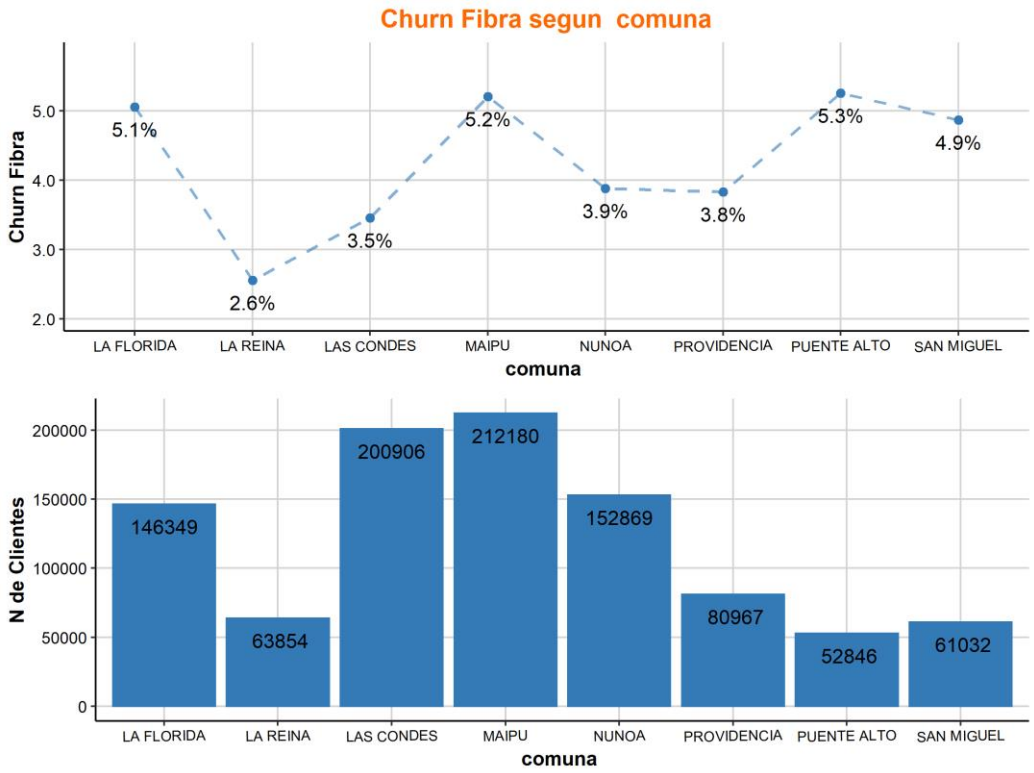


Gráfico 20: Distribución y tasa de fuga en variable comuna en modelo de predicción de fuga.

6.4.5 ANÁLISIS DE VARIABLES DE INTERACCIONES

Dado que se busca evaluar cómo las interacciones se relacionan y permiten predecir una posible fuga, se analizan aquellas que resultaron de mayor relevancia al entrenar el modelo de random forest sin algoritmo de balanceo. Las cinco variables de interacciones que mayor importancia tuvieron en el entrenamiento se muestran en la Tabla 15, así como también su porcentaje de peso para entrenar.

Variable	Importancia
Q_problemas_6m	1,66 %
Q_problemas_3m	1,39%
Q_días_interacciones_3m	1,23%
Q_días_interacciones_6m	1,10%
Q_consultas_6m	1,10%

Tabla 15: Variables más importantes de interacción para modelar en modelo de random forest sin algoritmo de balanceo.

En el Gráfico 21 y 22 se muestra cómo se relacionan la cantidad de interacciones que el cliente realizó los últimos seis meses debido a problemas y la cantidad de días que el cliente realizó interacciones los últimos 3 meses, respectivamente. Estas variables fueron consideradas importantes al momento de modelar, donde los problemas reportados los últimos 6 meses tuvo una importancia de un 1,66%, mientras que la cantidad de días que se realizaron interacciones a través del IVR los últimos 3 meses tuvo una importancia de un 1,23%.

Se puede ver que, a mayor cantidad de interacciones y problemas reportados, mayor es la probabilidad de fuga en la persona, lo cual es concordante con lo que se esperaba en base a la investigación realizada. Tal así que en aquellos clientes que reportaron problemas en más de tres ocasiones en los últimos 6 meses su probabilidad de fuga alcanza el 5,3%.

Por otro lado, cuando el cliente se contacta con la empresa más de dos veces en los últimos tres meses su probabilidad de fuga llega a alcanzar el 5,8%, siguiendo así la línea de tendencia esperada y mostrando una relación clara y marcada entre la cantidad de veces que el cliente se contacta con la empresa y la probabilidad que este se fugue a futuro.

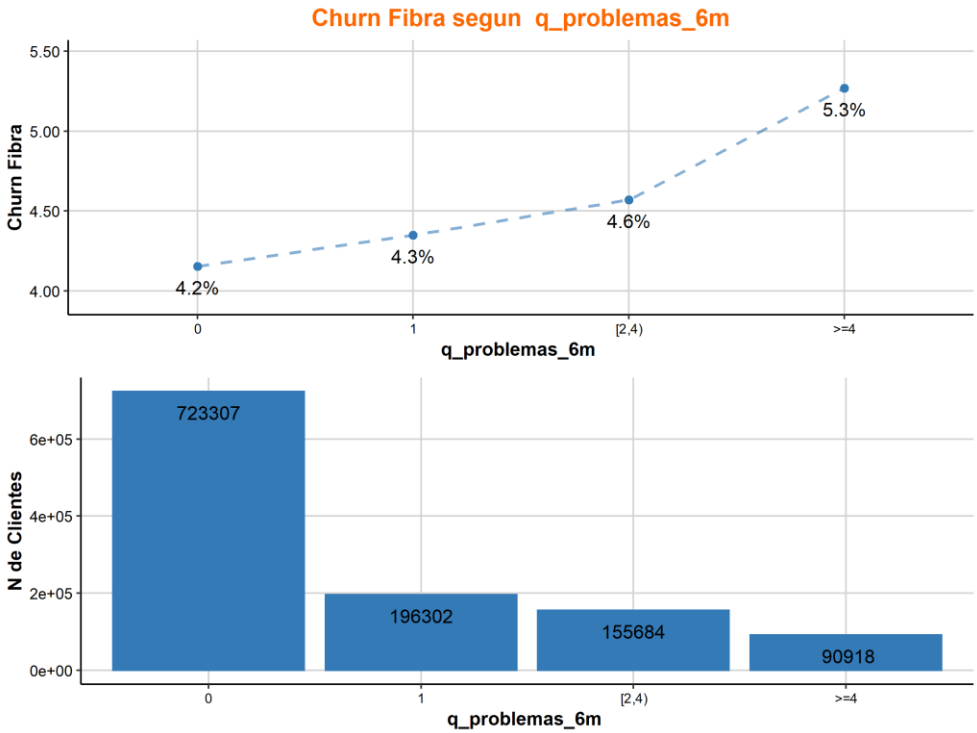


Gráfico 21: Distribución y tasa de fuga de variable cantidad de problemas en últimos 6 meses en modelo de predicción de fuga.

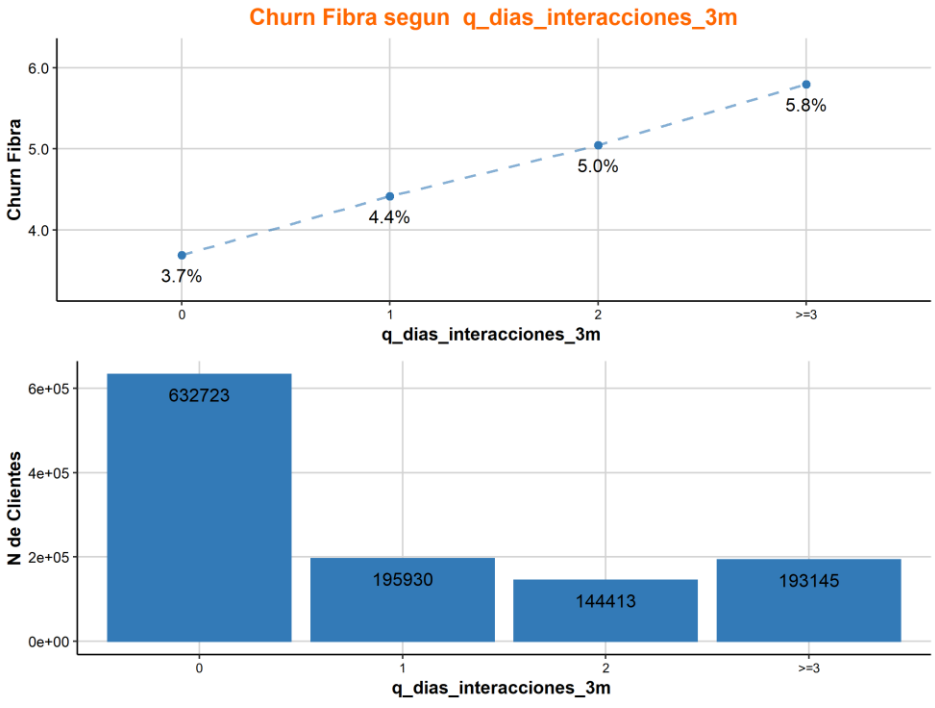


Gráfico 22: Distribución y tasa de fuga en variable de cantidad de días de interacciones en últimos 3 meses en modelo de predicción de fuga.

Entre las principales causas que son categorizadas como problemas se destacan:

- **Diferencia en navegación de datos**, que equivale al 42,6% del total de problemas reportados.

- **Diferencia en servicio de TV**, que equivale al 25,8% del total de problemas reportados.

- **Regulación de Front**, que equivale al 14,8% del total de problemas reportados.

- **Facturación y cobros**, que equivale al 8,2% del total de problemas reportados.

Es importante mencionar y barajar tres hipótesis que expliquen esta relación entre las variables de interacción y la fuga de los clientes, las cuales quedan propuestas como investigaciones futuras para la empresa.

1) **Relación interacciones – insatisfacción**: En primer lugar, se puede deber a que, dado que la mayoría de las interacciones ocurren cuando el cliente no está satisfecho con el precio/servicio o tiene algún problema técnico, la acumulación de las distintas interacciones que el cliente tiene con la empresa se ven reflejadas directamente en la satisfacción. En los últimos años se ha demostrado que las percepciones que los clientes van formando a través de distintas interacciones con la empresa determinan la relación existente entre el cliente y la misma (Magatef, 2015), lo cual se ve traducido en la satisfacción, que a la vez está altamente relacionado con la “lealtad” de los clientes, teniendo esto una influencia directa sobre la retención de los mismos. Luego, mientras más interacciones negativas presenten menor será su satisfacción, aumentando de esa manera su probabilidad de fuga.

2) **Relación interacciones – problemas no resueltos**: En segundo lugar, existe la posibilidad de que los clientes estén reportando problemas que no se están solucionando, lo que genera varias interacciones de parte de las mismas personas buscando solucionar sus problemáticas. Claramente si la empresa no les está solucionando los problemas, esto conllevará a la fuga oportuna de los clientes. En otras palabras, el cliente está dando aviso de problemas que a lo mejor no se están solucionando.

3) **Relación interacciones – Fin de descuento**: Por último, se baraja la posibilidad de que exista otra variable relacionada con la fuga y, a la vez, con las interacciones. Bajo esta línea se analizó cómo son las llamadas al IVR durante el mes específico en que se les acaba el descuento inicial a los clientes. Se observó que específicamente en el mes en que se cumple este hecho cerca de un 25% de los clientes realiza al menos una interacción con la empresa, en contraste con el 23% que se presenta en los otros periodos de antigüedad de contratación, lo que pareciera indicar a priori que efectivamente durante este periodo existe un leve aumento en la cantidad de interacciones. Sin embargo, al analizar la tasa de fuga en este periodo, se observó que esta no presenta una diferencia significativa con los otros meses por lo cual esta hipótesis queda descartada por el momento.

6.4.6 CARACTERIZACIÓN DE DECILES

Buscando evaluar cuánto es el peso real que les entrega el modelo a las distintas variables de predicción, se evaluará cómo se componen los deciles a los que se asocia cada cliente de la base de entrenamiento. Esta lógica se basa en utilizar el propio modelo entrenado para ver cómo predice la fuga en los mismos datos de entrenamiento, logrando entonces capturar el efecto real que el modelo le entrega a cada variable predictora. De esta manera se pueden sacar conclusiones o generar hipótesis observando la composición y características de las personas que se encuentran en cada decil de propensión: ¿Qué características tienen las personas más/menos propensas a fugarse?

Para hacer un contraste que muestre cómo varían los distintos deciles en torno a distintas variables descriptivas, se toman todos aquellos clientes de la base que se utiliza para entrenar el modelo de random forest sin algoritmo de balanceo y se verá la distribución de variables que tiene cada uno. Estos resultados se muestran en las Tablas 16 y 17, en donde se organizan los clientes más propensos a fugarse (decil 1), hasta llegar a los menos propensos a fugarse (decil 10).

Variable	**Tipo de variable**	**Decil 1**	**Decil 2**	**Decil 3**	**Decil 4**	**Decil 5**
Promedio antigüedad fibra	Cliente	11,1	11,3	10,9	10,4	10,2
Promedio cantidad interacciones 6 meses	Estado del servicio fibra	3,7	3,2	2,7	2,3	2,2
Promedio cantidad miembros del hogar	Cliente	2,1	2,0	2,0	2,0	2,2
Promedio propiedades familia	Cliente	0,8	0,7	0,8	0,8	1,0
Promedio edad	Cliente	36	34	36	37	39
% Clientes móviles	Servicios y productos	38%	33%	36%	36%	42%
% Clientes con descuento activo	Servicios y productos	18%	16%	16%	16%	17%
%Clientes que son dueños de propiedad	Cliente	18%	17%	18%	19%	21%

Tabla 16: Caracterización de deciles de propensión del 1 al 5 utilizando la base de entrenamiento en el modelo de random forest sin algoritmo de balanceo.

Variable	Tipo de variable	Decil 6	Decil 7	Decil 8	Decil 9	Decil 10
Promedio antigüedad fibra	Cliente	10,3	10,3	11,1	11,8	12,7
Promedio cantidad interacciones 6 meses	Estado del servicio fibra	2,1	1,8	1,9	2,0	2,3
Promedio cantidad miembros del hogar	Cliente	2,2	2,2	2,3	2,5	2,9
Promedio propiedades familia	Cliente	1,2	1,1	1,5	2,0	2,8
Promedio edad	Cliente	41	41	44	48	50
% Clientes móviles	Servicios y productos	45%	43%	49%	59%	79%
% Clientes con descuento activo	Servicios y productos	17%	16%	15%	13%	13%
%Clientes que son dueños de propiedad	Cliente	22%	21%	25%	31%	39%

Tabla 17: Caracterización de deciles de propensión del 6 al 10 utilizando la base de entrenamiento en el modelo de random forest sin algoritmo de balanceo.

De las tablas, se pueden sacar los siguientes insights de la agrupación realizada en deciles por el modelo:

- Los clientes de mayor antigüedad de contratación del servicio fibra se encuentran en los últimos deciles de propensión, mostrando la tendencia a la fidelización del cliente una vez que ya posee el servicio hace una cantidad considerable de tiempo. Existen diversas hipótesis que podrían explicar esta situación. Por un lado, una persona que ya lleva más de un año con el servicio es menos propensa a cancelarlo debido a que podría sentirse satisfecho con el mismo, lo cual se ve reflejado en la predicción que realiza el modelo mismo. También podría explicarse debido a la imposibilidad de sustituir por falta de competidores en aquel sector. Por último, la inacción de los clientes puede deberse al efecto estatus quo (también conocido como miedo al cambio), en donde los clientes que permanecen lo hacen por la estabilidad aceptable que les entrega en este servicio, lo cual prefieren frente a una opción nueva.

- Los grupos más propensos a fugarse de acuerdo con el modelo muestran en promedio una mayor cantidad de interacciones con la empresa en los últimos meses. Esta variable va de la mano con lo visto en la sección anterior, donde las interacciones con la empresa serían un claro predictor de la posible fuga del cliente en un futuro cercano.

- Hogares de familias más numerosas son clasificados con menor propensión a fugarse. Esto se puede asociar a que los hogares que solo están compuestos por una persona poseen alta tasa de fuga ya que están altamente correlacionado con personas de edad joven, quienes encuentran mejores ofertas y son quienes más podrían usar internet. Por otro lado, los hogares más grandes tendrían una menor probabilidad de fuga, probablemente asociado a que son un grupo socioeconómico más alto, donde se caracterizan familias más numerosas y, a la vez, menor capacidad de caer en default.

- Las propiedades familiares muestran una correlación positiva a medida que se incrementa el decil de propensión, demostrando que a aquellos clientes con mayor cantidad de propiedades el modelo les asignaría una menor probabilidad de fuga. Nuevamente podría existir una relación alta entre mayor cantidad de propiedades y un grupo socioeconómico más alto, donde el precio del servicio no sería un factor tan relevante a la hora de decidir si cancelarlo o no.

- Se observa que en los primeros deciles se encuentran los clientes más jóvenes, mientras que a medida que se avanza en el decil estos incrementan su edad promedio, llegando a existir una diferencia de edad promedio de 14 años entre los deciles más y menos propensos a fugarse. Una razón de esto puede darse por la facilidad y disposición que poseen la gente más joven a cambiar los servicios por una tecnología mejor, mientras que la gente de mayor edad tiende a conformarse con menos requisitos tecnológicos para realizar sus actividades, lo que estaría siendo capturado y utilizado por el modelo de predicción.

- El porcentaje de gente en el decil que además posee un servicio móvil pospago con la empresa también muestra una tendencia en los deciles. Este valor se incrementa considerablemente a medida que el modelo selecciona una menor propensión a fuga del cliente, tal así que en el primer decil tan solo un 38% de los clientes son clientes móviles, mientras que en el décimo decil un 79% es también un cliente móvil en la compañía. Esto muestra que el modelo le entrega una alta importancia a la fidelización que presentan los clientes con la compañía al momento de predecir una posible fuga.

- Los primeros deciles poseen mayor porcentaje de clientes con un descuento activo con la compañía, probablemente relacionado a la predisposición que poseen de cancelar el servicio, la cual ya ocurrió con anterioridad y podría darse nuevamente, situación que es capturada y utilizada por el modelo en su predicción.

- Los deciles más propensos a fugarse poseen menor porcentaje de clientes que son dueños del hogar sobre el que contratan el servicio, tal así que este valor se incrementa considerablemente a medida que se evalúa en los deciles de menor propensión. Este hecho se podría asociar a una alta correlación entre la capacidad económica de la persona y la menor probabilidad a cambiar su domicilio, factores que hacen que la persona tenga menor probabilidad de cancelar el servicio, donde se utiliza esta información para establecer un score de fuga a los clientes.

6.5 DESPLIEGUE

Para finalizar este trabajo, se busca dejar en claro cómo se utilizará este modelo, qué acciones tomar y los insights obtenidos para campañas futuras de la empresa. Esta es una de las secciones más importantes de la metodología, ya que permite que el trabajo realizado continue ejecutándose y sea manejable por otros empleados, entregando así un valor continúo y a largo plazo para la empresa.

La ejecución del modelo en campañas proactivas, la realización de un experimento y la evaluación del desempeño del modelo en fugas futuras se especifican en esta sección.

6.5.1 EJECUCIÓN DE MODELO

Dado que las gestiones de las distintas fuentes de datos se trabajan de manera separada en la empresa, la disponibilidad de la información completa y, por ende, la predicción del modelo se obtendrá la segunda semana de cada mes, por lo que las áreas dedicadas a la contactabilidad del cliente o la realización de las campañas de retención proactivas que podrían realizarse con los resultados del modelo deben eliminar de la base a aquellas fugas que ya ocurrieron al momento de la entrega de la información.

También es importante considerar que, si se desea realizar campañas proactivas de contactabilidad en grupos propensos a fugarse, esta debe hacerse dentro de un horizonte temporal de dos meses. En otras palabras, no se debe predecir la fuga todos los meses, sino que en camadas cada dos meses sobre los cuales puede ejercerse una acción de retención proactiva.

De igual manera, se deben excluir a clientes que el modelo predijo el periodo anterior y que fueron ya contactados, evitando así un hostigamiento de parte de la empresa al mismo y/o aplicar dos veces un mismo descuento a las personas.

El proceso entonces se realiza en el siguiente orden:

1) Se obtiene la base de clientes de cierre de mes y sus variables descriptivas, creando así un tablón que contenga toda la información necesaria para realizar la predicción. La disponibilidad de toda la información está prevista para el día 8 de cada mes.

2) Se calcula el score de propensión de fuga para cada cliente con el modelo ya entrenado.

3) Se eliminan de la base aquellas personas que ya se fugaron los primeros días del mes. Dado que el modelo predice fugas a dos meses, aún deberían presentarse una cantidad de fugas no menores en cada percentil en el tiempo restante de ese periodo de campaña.

4) Se elimina de la base a clientes que ya fueron contactados en la campaña previa.

En el caso de las telecomunicaciones ya se comentó anteriormente que se caracteriza por ser una industria con una gran cantidad de clientes y donde la mayor parte de las campañas se realizan a través de canales como Call Center, email o mensajes de texto. Luego, y considerando que estos canales tienen alcances limitados de contactabilidad debido tanto a la capacidad del negocio como también a los costos económicos de contactarse con cada cliente, solo se puede establecer contacto con una pequeña parte de la población total. Por esta razón y siguiendo la línea planteada en estudios anteriores (Richter et al, 2010), las métricas globales del modelo no aportan tanto como lo hacen las métricas específicas de comportamiento entre aquellos clientes que sí se espera contactar.

Bajo la lógica ya descrita, métricas globales como accuracy, recall y precision nos dan un indicio de la calidad del modelo, pero no serán aquellas que definan aquel modelo de mejor desempeño para este caso de estudio. Por el contrario, haremos un enfoque primordial de aquel modelo que presenta mejor desempeño para identificar fugas únicamente en el 1% más propenso de la base, que son aquellos que las áreas de gestión de campañas en la empresa contactan con regularidad en este tipo de acción proactiva.

Se vuelve entonces importante evaluar cuánta gente que realmente se fuga está quedando fuera de la base de acción de retención. Para esto, en las tablas 18, 19 y 20 se muestran el porcentaje de las fugas totales de la base de testeo que quedan seleccionadas en el 1% más propenso, el 2% más propenso, el 3% más propenso y el 97% restante para cada tipo de algoritmo de balanceo.

Modelo	Fugas totales 1%	Fugas totales 2%	Fugas totales 3%	Fugas totales 97%
GBM	6,7%	4,2%	3,7%	85,5%
Random Forest	14,4%	5,4%	3,8%	76,4%
XRT	11,6%	5,2%	3,4%	79,8%
Stacked Ensemble AM	14,2%	6,2%	4,0%	75,7%
Stacked Ensemble BoF	14,2%	6,2%	4,0%	75,7%

Tabla 18: Porcentaje de fugas totales en cada grupo de propensión para los modelos de predicción de fuga sin algoritmo de balanceo.

Modelo	Fugas totales 1%	Fugas totales 2%	Fugas totales 3%	Fugas totales 97%
GBM	8,3%	4,9%	4,3%	82,5%
Random Forest	8,3%	5,0%	3,8%	83,0%
XRT	6,9%	4,6%	3,3%	85,2%
Stacked Ensemble AM	8,6%	5,3%	4,4%	81,7%
Stacked Ensemble BoF	8,4%	5,4%	4,4%	81,8%

Tabla 19: Porcentaje de fugas totales en cada grupo de propensión para los modelos de predicción de fuga con algoritmo de undersampling.

Modelo	Fugas totales 1%	Fugas totales 2%	Fugas totales 3%	Fugas totales 97%
GBM	4,6%	3,1%	2,9%	89,4%
Random Forest	10,3%	6,5%	4,8%	78,4%
XRT	7,8%	4,8%	3,9%	83,5%
Stacked Ensemble AM	10,3%	6,5%	4,8%	78,4%
Stacked Ensemble BoF	10,3%	6,5%	4,8%	78,4%

Tabla 20: Porcentaje de fugas totales en cada grupo de propensión para los modelos de predicción de fuga con algoritmo de oversampling.

Se puede notar que, si se llegara a actuar únicamente en el 1% más propenso de la base, aquellos modelos que capturan la mayor cantidad de fugas totales respecto al total de fugas corresponden en su mayoría a aquellos sin algoritmo de balanceo. Por otro lado, en el caso que se desee hacer una campaña proactiva con mayor cantidad de clientes contactados, por ejemplo, sobre el 2% o 3% más propenso a fugarse, los modelos de mejor desempeño en estos percentiles son aquellos que poseen un algoritmo de oversampling para entrenar, pero considerando que esto se realiza a costa de detectar menos fugas en el primer percentil. Por ende, si se busca minimizar la cantidad de fugas que no fueron contactadas (97% menos propenso) dado que se contacta a la suma de los tres percentiles más propensos, conviene utilizar nuevamente los modelos sin algoritmo de balanceo, donde se dejarían sin contactar al 75% de las fugas totales en el mejor modelo entrenado.

En conclusión y en base a los resultados de los quince modelos generados, se obtiene que aquel de mejor resultado predictivo para este tipo de problemática es el modelo de random forest sin balanceo, el cual presenta mayor lift de fuga entre la población más propensa a cancelar el servicio. Al observar las métricas globales del modelo tampoco presenta resultados muy distintos a los otros modelos, razón por la cual no presentaría grandes problemas en caso de querer contactar más clientes. Por otro lado, si se deseara contactar al 3% más propenso de la base, los modelos de mejor desempeño para esta problemática serían aquellos de la familia Stacked Ensemble sin algoritmo de balanceo.

6.5.2 DISEÑO DE EXPERIMENTO

Con el fin de obtener resultados más completos, se deben plantear formas de evaluar las interacciones que ocurren desde la empresa hacia los clientes, estableciendo así si existe un posible beneficio en utilizar tal herramienta a modo de evitar fugas en grupos de clientes con alta predisposición a fugarse. Para esto se diseña un experimento que buscará evaluar los posibles beneficios que podrían obtenerse de la contactabilidad con los clientes previo a la fuga. Así, se establecerán distintos grupos con distintos tipos de contactabilidad, dejando siempre un grupo de control que permita capturar el verdadero efecto de la acción realizada.

Se propone entonces realizar un experimento a los percentiles más propensos a fugarse según el modelo. Cada percentil se dividirá en tres grupos, donde se ejercerá una acción distinta a cada uno:

- ✓ **Grupo 1: Contacto + Descuento**. La empresa se contactará con el cliente para consultar por el funcionamiento actual del servicio, donde además se les hará entrega de un descuento en el servicio.

- ✓ **Grupo 2: Contacto.** La empresa se contactará con el cliente para solucionar dudas y consultar por el funcionamiento actual del servicio.

- ✓ **Grupo 3: Control.** Corresponde a un grupo de control, donde no se recibe ninguna acción directa de parte de la empresa.

El objetivo de estos tres grupos es medir cómo reaccionan los clientes propensos a fugarse ante tanto incentivos económicos, como ante una preocupación de la empresa por ellos y el servicio que les está entregando.

Se propone realizar el experimento al decil más propenso a fugarse (que posee una tasa de fuga esperada de un 15,7%). Para obtener un resultado que asegure una significancia estadística se realiza una prueba de poder estadístico Chi-quadrado, donde se determina que con un alfa de 5%, un poder estadístico de 80% y un total de 14.760 clientes, se requiere de un tamaño de efecto mínimo a detectar de 0.026. Esta agrupación se detalla en la Tabla 21.

Grupo	N
Contacto + Descuento	4.920
Contacto	4.920
Control	4.920
Total Clientes Experimento	14.760

Tabla 21: Distribución grupos experimento.

Cada percentil se dividirá de forma aleatoria en tres subgrupos, buscando que la distribución de probabilidad de fuga sea igual entre cada uno de los grupos en el experimento. Para ayudar a entender la distribución del experimento por percentiles, se detalla la composición de cada uno de los grupos en la Tabla 22.

Percentil	Q Clientes Control	Q Clientes Contacto + Descuento	Q clientes Contacto
1	492	492	492
2	492	492	492
3	492	492	492
4	492	492	492
5	492	492	492
6	492	492	492
7	492	492	492
8	492	492	492
9	492	492	492
10	492	492	492
Total	3.444	3.444	3.444

Tabla 22: Distribución grupos experimento entre percentiles.

Estos tres grupos deben chequear su distribución de distintas variables, para así asegurar que la asignación fue aleatoria y no existe un sesgo que podría dañar la validez del experimento.

Finalmente, se debe realizar el experimento y analizar sus resultados al finalizar el segundo mes del que se prevé ocurra la fuga.

6.5.3 EVALUACIÓN DE MODELO EN NUEVAS FUGAS

Para finalizar este trabajo de título, se evalúa el desempeño del modelo en datos nuevos. Para esto se calcula el score de fuga con el modelo seleccionado en la sección anterior para los clientes de octubre 2021, es decir, considerando un modelo que prediga las fugas de estos en los meses de noviembre y diciembre del año 2021.

Cabe mencionar que el 23 de noviembre de 2021 la empresa utilizó los resultados del modelo para realizar campañas proactivas de retención entre los dos percentiles más propensos a la fuga, acción que podría haber generado un efecto sobre las fugas reales que el modelo preveía que ocurrirían. Por esta razón, para la presente sección del informe se analizarán los resultados únicamente para el periodo comprendido entre el día 1 y el 22 de noviembre del año 2021, es decir, previo a que se realizara la campaña ya mencionada.

Los resultados agrupados por decil de propensión se muestran en la Tabla 23. En esta se puede visualizar que el modelo efectivamente cumple la lógica prevista, donde la mayor cantidad de fugas se reportó en los deciles más altos, con una tasa que baja de forma monótona. Por ende, habiendo pasado los primeros 22 días de los 2 meses que se predice la fuga, se establece un buen desempeño del modelo en datos nueva, pero sí se nota un desempeño más bajo respecto a lo que se obtuvo en la base de testeo.

Decil	Total Clientes	Cantidad de fugas	Tasa de fuga
1	14.753	440	2,98%
2	14.753	305	2,07%
3	14.756	267	1,81%
4	14.754	248	1,68%
5	14.754	224	1,52%
6	14.755	202	1,37%
7	14.754	189	1,28%
8	14.752	176	1,19%
9	14.753	162	1,10%
10	14.752	134	0,91%

Tabla 23: Fugas ocurrida en cada decil de propensión seleccionado por el modelo de fuga entre el 1 y 22 de noviembre del año 2021.

Es posible observar que la tasa de fuga es menor a la que se esperaba, lo cual es normal ya que tan solo se está viendo una porción pequeña de los datos que se estaban prediciendo con el modelo (se predice quienes se fugarán a 2 meses, pero solo se están

evaluando los primeros 22 días del primer mes). Por ende, aún no es posible analizar cuál fue la tasa de error del modelo, debido a que aquellas personas que se predice que serán fuga aún podrían cancelar el servicio en el próximo mes y la predicción no estaría errónea. Esta sección no muestra resultados finales en datos nuevos, pero si una aproximación que permite observar si el modelo se encuentra bien encaminado, lo cual efectivamente ocurre.

La información y el detalle completo por percentil se encuentra en la Tabla 41 de la sección anexos de este trabajo, donde la tasa de fuga en los primeros 22 días del mes alcanza un 5,22% en el percentil más alto.

CAPÍTULO 7: CONCLUSIONES

Tras evaluar quince modelos de predicción de fuga para el servicio internet fibra óptica de una compañía de telecomunicaciones, se obtiene un buen resultado de predicción que permite identificar de forma apropiada a los clientes más propensos a cancelar el servicio contratado. En aquel modelo de mejor resultado se prevé que realizando acciones de retención proactiva a tan solo el 3% más propenso a fugarse se estarían contactando al 25% de las fugas totales que se reportan a nivel global en el servicio. Esto permitiría enfocar recursos en un grupo reducido de clientes, pero con un nivel alto de detección temprana de fugas.

El modelo de predicción que presentó mejor desempeño para este tipo de tareas es el modelo Random Forest, el cual entrena y genera predicciones en base a la realización de distintos árboles de decisión. De igual manera, los metamodelos de la familia de Stacked Ensemble presentan un buen desempeño si lo que se busca es contactar a bases más grandes de clientes, como lo serían los clientes 10% más propensos a fugarse.

Al evaluar dos algoritmos de balanceo de clases se determina que estos parecerían no mejorar el modelo en cuanto a las métricas que se consideran de mayor relevancia para resolver esta problemática, donde aquellos modelos sin algoritmo de balanceo predecirían de mejor forma las fugas en los percentiles de propensión más altos. Este hecho se debe principalmente a que los algoritmos de balanceo de clases lo que hacen es identificar una mayor cantidad de fugados respecto al umbral de fuga definido por el modelo, lo que no se ve reflejado en métricas como el Lift de propensión, la cual fue utilizada para analizar el desempeño de los modelos en este problema. Por otro lado, el método de oversampling supera de manera considerable al algoritmo de undersampling en esta problemática.

Al evaluar a los clientes con mayor propensión a fugarse en base a lo determinado por el modelo, se establece que estos grupos se caracterizan por estar compuestos por personas más jóvenes, de menor antigüedad de contratación del servicio, presentan mayor cantidad de interacciones con la empresa, no poseen contratación de servicios pospago con la compañía y tienen descuentos activos en el servicio. Por otro lado, en cuanto a la zona donde habitan son por lo general lugares con muy alta penetración de la empresa y una cantidad considerable de puertas libres en el gabinete que les entrega el servicio. En base a esto, se vuelve importante un análisis futuro que profundice en las razones de por qué este tipo de clientes presentaría una mayor propensión a fugarse, buscando analizar si las hipótesis planteadas en este trabajo son correctas o, en caso contrario, lo que se esconde realmente detrás de clientes con estas características.

Además, las interacciones que el cliente establece con la empresa a través del canal de IVR serían un buen predictor de fuga, donde aquellos clientes con mayor propensión a fugarse presentan también una cantidad considerable de contactos con la empresa los últimos meses, ya sea para reportar problemas, dejar constancia de reclamos, generar consultas o enviar solicitudes. Luego, es en estos clientes donde debe ponerse especial foco en campañas de retención proactivas, ya que muestran indicios tempranos de problemas y se demuestra que efectivamente las interacciones que el cliente realiza con la empresa son un buen predictor de una fuga pronta del mismo. El servicio que se les entrega post atención podría ser clave para retener al cliente, razón que vuelve importante que este mantenga un seguimiento sobre cómo se resuelven los problemas reportados y cómo varía la satisfacción de los clientes tras estos hechos.

La realización del experimento propuesto permite evaluar si ejercer acciones de contactabilidad o entrega de descuentos podría disminuir la probabilidad de fuga entre grupos propensos a cancelar el servicio según lo establecido en el modelo. De ser así, la contabilidad se volvería una pieza clave no solo para predecir la fuga (en base a la contactabilidad que el cliente realiza a través de los llamados al IVR), sino que también se volvería una herramienta de retención poderosa para evitar la fuga de los clientes a futuro. A la vez, la realización de este experimento debe ir de la mano con un análisis económico que evalúe el trade-off entre los costos de realización de este (considerando tanto los costos de contactabilidad, como los provenientes de los descuentos entregados) versus las ganancias que se obtienen por retención de clientes.

Es importante tener en consideración ciertos aspectos para futuros análisis. En primer lugar, este modelo fue entrenado con datos posteriores a la pandemia COVID-19, que como ya se mencionó fue un periodo que se destacó por un aumento en la cartera de clientes de la empresa. Dado que se espera que las restricciones sanitarias cambien en los meses venideros, se debe evaluar si estos cambios van también modificando tanto la performance del modelo, como también la importancia de las variables sobre el mismo. ¿Cómo los cambios y las menores restricciones sanitarias podrían llegar a afectar las predicciones del modelo y las conductas de los clientes? Es necesario no solo reentrenar el modelo cada cierto tiempo, sino que analizar constantemente los resultados de este y los cambios que se presentan.

Dado que el modelo fue entrenado con la fuga total, existe la posibilidad de que el mismo este prediciendo únicamente a aquellas personas con problemas de pago o mayor probabilidad de default (fuga involuntaria), lo cual podría ocurrir debido a correlaciones entre algunas variables y el nivel socioeconómico de las personas. Esto podría generar problemas al intentar retener personas que no poseen capacidad de pago o que, una vez que se les acabe el nuevo descuento, de igual manera se les cancele el servicio, lo cual sería contraproducente. Para evitar esto, se plantea a futuro generar dos modelos distintos: uno que prediga únicamente fuga voluntaria que se utilice para campañas proactivas de retención, y otro que prediga fuga involuntaria, el cual podría generar valor

especialmente para los proyectos de riesgo y facturación de clientes que tiene la empresa.

Bajo esta misma línea, en el presente trabajo se observó una relación entre comuna de residencia y fuga, donde algunas de estas presentaban tasas muy superiores a otras. Por esta razón surge la duda de si esto se debe a la correlación existente entre la comuna y el nivel socioeconómico (el cual también se observó que se relaciona con la fuga), o podría ocurrir que ciertos sectores sean más propensos a caídas del servicio y, por ende, mayor molestia de los clientes. En el primer caso, podría ser un indicador de que el modelo está prediciendo en mayor medida fugas involuntarias, mientras que en el segundo caso podría ser un problema de calidad en el servicio entregado. Para esto se necesita recopilar información de los problemas técnicos y caídas efectivas que se tienen en el servicio, no solo aquellas que reportan los clientes en los llamados.

Finalmente, se plantea evaluar la posibilidad de extrapolar los resultados de este modelo a otros servicios ofrecidos por la compañía. Ya se mencionó que el mercado hogar podría diferir bastante del mercado personas debido a que se utilizan por entes distintos y con objetivos distintos. Queda como un posible análisis futuro la creación de un modelo para el mercado de productos y servicios ofrecidos a personas, y la evaluación de qué tan distintos son los resultados respecto al modelo actual. ¿Es posible utilizar un único modelo global que simplifique las gestiones? ¿Disminuye en gran medida el desempeño del modelo al hacer esto? Analizar el trade-off entre simplicidad y desempeño es clave para resolver estas preguntas.

CAPÍTULO 8: BIBLIOGRAFÍA

[1] Empresa Nacional de Telecomunicaciones (2020). Memoria Corporativa 2019. <https://entel.modyocdn.com/uploads/f680cdbe-a10a-4e1a-83a5-80d62a0defe3/original/200427-Memoria-Entel-UV-FR_1_.pdf>

[2] Subsecretaría de telecomunicaciones de Chile (2020). Estadísticas – Internet. https://www.subtel.gob.cl/estudios-y-estadisticas/internet/

[3] C. B. Bhattacharya. (1998). “When customers are members: customer retention in paid membership contexts,” Journal of the Academy of Marketing Science, vol. 26, no. 1, pp. 31–44, 1998

[4] Pérez Villanueva, P. (2014). Modelo de predicción de fuga de clientes de telefonía móvil post pago. Disponible en http://repositorio.uchile.cl/handle/2250/115942

[5] Contreras, E., Ferreira, F., Valle, M. (2017). Diseño de un modelo predictivo de fuga de clientes utilizando árboles de decisión. Revista Ingeniería Industrial - Universidad del Bío Bío. http://revistas.ubiobio.cl/index.php/RI/article/view/3055

[6] Ahmad, A.K., Jafar, A. & Aljoumaa, K. Customer churn prediction in telecom using machine learning in big data platform. J Big Data 6, 28 (2019). https://doi.org/10.1186/s40537-019-0191-6

[7] Huang, B., Tahar, M., Buckley, K. Customer churn prediction in telecommunications,Expert Systems with Applications. Volume 39, Issue 1 (2012). https://doi.org/10.1016/j.eswa.2011.08.024.

[8] Brandusoiu, I., Toderean, G. Churn Prediction in the Telecommunications Sector Using Support Vector Machines. Annals of the Oradea University. Fascicle of Management and Technological Engineering. (2013). https://10.15660/AUOFMTE.2013-1.2772

[9] Devriendt, F., Berrevoets, I., Verbeke, W. Why you should stop predicting customer churn and start using uplift models, Information Sciences, Volume 548 (2021). https://doi.org/10.1016/j.ins.2019.12.075.

[10] Umayaparvathi, V., & Iyakutti, K. (2012). Applications of Data Mining Techniques in Telecom Churn Prediction. International Journal of Computer Applications, 42, 5-9.

[11] Mohamed F. H., Kasapbaşı M.,C. (2021). Churn Prediction with Ensemble Classifiers for Telecom Sectors. Journal of Technology and Applied Sciences 4(1), 57-71

[12] Kumar, S. (2021). Opportunities of Machine Learning on Telecom Sector: A Case Study at BSNL. International Journal of Research in Engineering and Science (IJRES), 9, 37-44.

[13] Chouiekh, Alae & Haj, El. (2020). Deep Convolutional Neural Networks for Customer Churn Prediction Analysis. International Journal of Cognitive Informatics and Natural Intelligence. 14. 1-16. 10.4018/IJCINI.2020010101.

[14] Oludele, Awodele & Ben, Adeniyi & A.C, Ogbonna & Kuyoro, Shade & Seun, Ebiesuwa. (2020). Enhanced Churn Prediction in theTelecommunication Industry. International Journal of Innovative Research in Computer Science & Technology. 8. 6-15. 10.21276/ijircst.2020.8.2.1.

[15] Ahmad, A.K., Jafar, A. & Aljoumaa, K. Customer churn prediction in telecom using machine learning in big data platform. J Big Data 6, 28 (2019). https://doi.org/10.1186/s40537-019-0191-6

[16] Faris, Hossam. (2018). A Hybrid Swarm Intelligent Neural Network Model for Customer Churn Prediction and Identifying the Influencing Factors. Information. 9. 288. 10.3390/info9110288.

[17] J. Burez & D. Van Den Poel, 2008. "Handling class imbalance in customer churn prediction," Working Papers of Faculty of Economics and Business Administration, Ghent University, Belgium 08/517, Ghent University, Faculty of Economics and Business Administration.

[18] Richter, Yossi & Yom-Tov, Elad & Slonim, Noam. (2010). Predicting Customer Churn in Mobile Networks through Analysis of Social Groups. Proceedings of the 10th SIAM

International Conference on Data Mining, SDM 2010. 732-741. 10.1137/1.9781611972801.64.

[19] Torres-Vásquez, M.; Hernández-Torruco, J.; Hernández-Ocaña, B. y Chávez-Bosquez, O. (2021). «Impacto de los algoritmos de sobremuestreo en la clasificación de subtipos principales del Síndrome de Guillain-Barré». Ingenius. N.° 25, (enero-junio). pp. 20-31. doi: https://doi.org/10.17163/ings.n25.2021.02

[20] P. Kisioglu and Y. I. Topcu, “Applying Bayesian belief network approach to customer churn analysis: a case study on the telecom industry of Turkey,” Expert Systems with Applications, vol. 38, no. 6, pp. 7151–7157, 2011.

[21] Y. Li, B. Hou, Y. Wu, D. Zhao, A. Xie, and P. Zou, “Giant fight: customer churn prediction in traditional broadcast industry,” Journal of Business Research, vol. 131, pp. 630–639, 2021

[22] P. C. Verhoef, “Understanding the effect of customer relationship management efforts on customer retention and customer share development,” Journal of Marketing, vol. 67, no. 4, pp. 30–45, 2003.

[23] W. J. Reinartz and V. Kumar, “*e impact of customer relationship characteristics on profitable lifetime duration,” Journal of Marketing, vol. 67, no. 1, pp. 77–99, 2003.

[24] B. Huang, M. T. Kechadi, and B. Buckley, “Customer churn prediction in telecommunications,” Expert Systems with Applications, vol. 39, no. 1, pp. 1414–1425, 2012.

[25] M. H. U. Rehman, V. Chang, A. Batool, and T. Y. Wah, “Big data reduction framework for value creation in sustainable enterprises,” International Journal of Information Management, vol. 36, no. 6, pp. 917–928, 2016.

[26] J. Su, Y. Yu, and X. Zhang, “Knowledge transfer efficiency measurement with application for open innovation networks,” International Journal of Technology Management, vol. 81, no. 1, pp. 118–142, 2019

[27] Wirth, R. & Hipp, Jochen. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.

[28] Brownlee, J. Stacking Ensemble Machine Learning With Python. Machine Learning Mastery, 2020.

[29] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[30] D.P. Kroese, Z.I. Botev, T. Taimre, R. Vaisman. Data Science and Machine Learning: Mathematical and Statistical Methods, Chapman and Hall/CRC, Boca Raton, 2019.

[31] Fernández, A. & García, S. & Galar, M. & Prati, R. & Krawczyk, B. & Herrera, F. (2018). Learning from Imbalanced Data Sets. 10.1007/978-3-319-98074-4.

[32] Hoens, T. & Chawla, Nitesh. (2013). Imbalanced Datasets: From Sampling to Classifiers. 10.1002/9781118646106.ch3.

[33] Magatef, Sima. (2015). The Impact of Customer Loyalty Programs on Customer Retention.

ANEXOS

ANEXO A: VARIABLES UTILIZADAS EN MODELO

Nombre variable	Tipo de variable	Descripción
Fuga	BINARIA	1 si cliente canceló el servicio al mes siguiente, 0 si no
Mes	INT	Fecha de la información en formato AAAAMM
Rut	STR	Número de documento único de cliente
Id_periodo	INT	Indicador que se refiere a cuántos meses hacia atrás se toma la muestra respecto al mes más actual.
Id_Bundle	INT	Identificador único de contrato
Bundle_name	STR	Detalle de tipo de servicio contratado
Fecha_activacion	FECHA	Fecha en que se activó el servicio de internet
Pack_hab	STR	Nombre de pack que contrató el cliente
Id_direccion	INT	Identificador único de dirección domiciliaria
Tipo_inmueble	STR	Tipo de domicilio
Grupo_canal	STR	A través de qué canal el cliente contrató el servicio
Antigüedad	INT	Meses que el cliente lleva contratando el servicio fibra
Cliente_pre_pandemia	BINARIA	1 si el cliente contrató el servicio antes de marzo 2020, 0 si no
Internet_competencia_x	BINARIA	1 si el cliente tiene factibilidad de internet en la competencia x, 0 si no
Q_competencia_fibra	INT	Cantidad de empresas de la competencia que presentan factibilidad fibra en esa dirección
Cliente_movil	BINARIA	1 si el cliente posee un número de teléfono en la compañía a su nombre, 0 si no
Antigüedad_movil	INT	Cantidad de meses que el cliente lleva contratando (si es

		que tiene) servicio de telefonía con la empresa
Lineas	INT	Cantidad de líneas móviles que el cliente es titular
Suma_cf	INT	Suma del costo fijo que posee el cliente en servicios móviles
Hogar_key	STR	Identificador único de hogares
Entre_X_Y_años	INT	Cantidad de personas que viven en el hogar entre X e Y años
Es_padre	BINARIA	1 si es padre, 0 si no
Q_hijos	INT	Cantidad de hijos
Promedio edad	INT	Promedio edad de los habitantes de cada hogar
Miembros	INT	Cantidad de personas que viven en el hogar
Gse	STR	Grupo Socioeconómico del hogar
Comuna	STR	Comuna del hogar
Q_propiedades	INT	Propiedades que tiene la persona a su nombre
Q_propiedades_familia	INT	Propiedades que tiene la persona o su cónyuge a su nombre
Avaluo_total	INT	Avaluó total de las propiedades
Es_dueño_propiedad	BINARIA	1 si la persona o su cónyuge son dueños de la propiedad donde tienen contratado el servicio, 0 si no
Tiene_auto	BINARIA	1 si tiene auto, 0 si no
Q_autos	INT	Cantidad de autos que la persona tiene a su nombre
Monto_descuento	INT	Monto de último descuento aplicado
Fin de descuento	INT	Fecha en que termina descuento en formato AAAAMM
Dscto_inicial	BINARIA	1 si el cliente se encuentra en los primeros 6 meses de contrato, 0 si no
Descuento_activo	BINARIA	1 si posee un descuento activo a la fecha, 0 si no
Meses_rest_desc	INT	Meses para que termine el descuento, si es negativo significa que ya finalizó
Fuga_pack_3m	FLOAT	% Fuga que ha tenido ese pack los últimos 3 meses

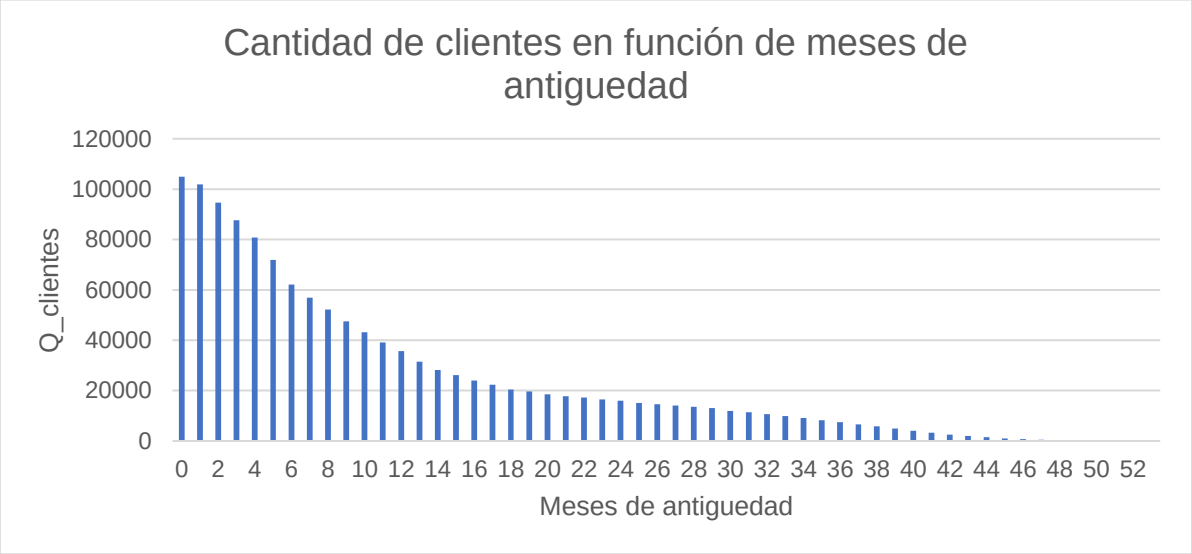
Q_interacciones_x_meses	INT	Cantidad de tickets generados por el cliente los últimos x meses
Q_días_interacciones_x_meses	INT	Cantidad de días distintos que se generaron tickets por el cliente los últimos x meses
Q_solicitudes_x_meses	INT	Cantidad de tickets generados por el cliente con motivo de “SOLICITUD”
Q_reclamos_x_meses	INT	Cantidad de tickets generados por el cliente con motivo de “RECLAMO”
Q_consultas_x_meses	INT	Cantidad de tickets generados por el cliente con motivo de “CONSULTA”
Q_problemas_x_meses	INT	Cantidad de tickets generados por el cliente con motivo de “PROBLEMA”
Profesión	STR	Ultima profesión reportada por la persona
Profesional	BINARIA	1 si posee profesión, 0 si no
Penetración Distrito	FLOAT	Muestra cuántos clientes tiene la empresa en el distrito respecto a la cantidad de clientes potenciales (con factibilidad) que se tiene.
Puertas libres gabinete	INT	Total de puertas libres que posee cada gabinete en un cierto mes. Una puerta equivale a una conexión que se puede realizar, donde un cliente se conecta a una sola puerta cada vez.
Puertas ocupadas gabinete	INT	Total de puertas ocupadas que posee cada gabinete en un cierto mes. Una puerta equivale a una conexión que se puede realizar, donde un cliente se conecta a una sola puerta cada vez.

Tabla 24: Variables disponibles para construcción de modelo de fuga.

Nombre pack fibra	Contenido
BAF	Internet fibra óptica
DUO TV	Internet fibra óptica + TV
3PLAY	Internet fibra óptica + TV + Telefonía
DUO TF	Internet fibra óptica + Telefonía
DUO OTT	Internet fibra óptica + TV con servicio OTT
3PLAY OTT	Internet fibra óptica + TV con servicio OTT + Telefonía

Tabla 25: Descripción pack de servicio fibra ofrecidos.

ANEXO B: GRÁFICOS ANÁLISIS DE VARIABLES



Gráficos 23: Distribución de los meses de antigüedad de los clientes de fibra óptica contratando el servicio.

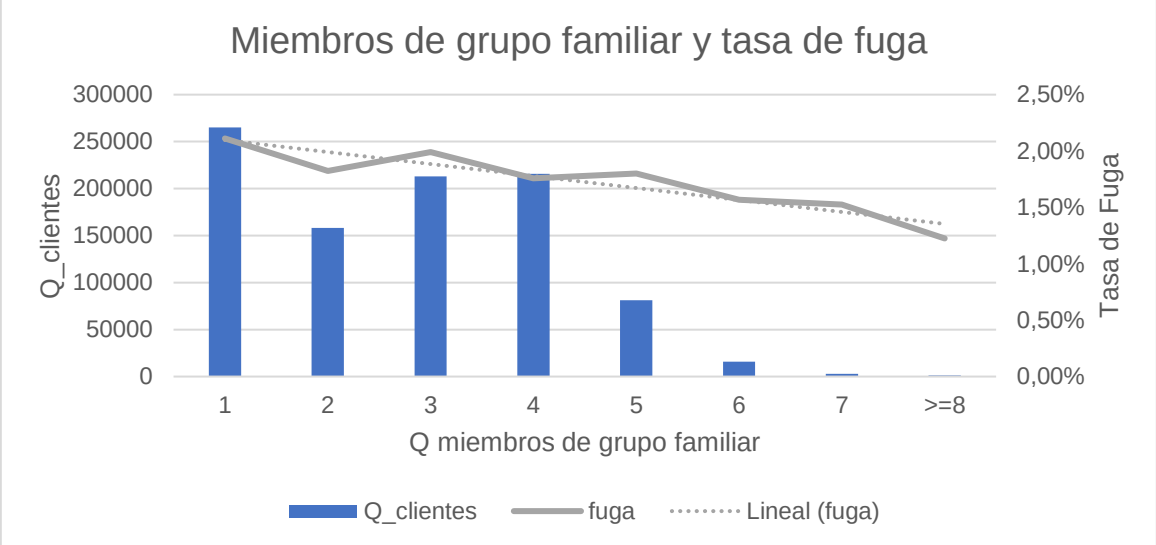


Gráfico 24: Miembros del grupo familiar y su tasa de fuga respectiva.

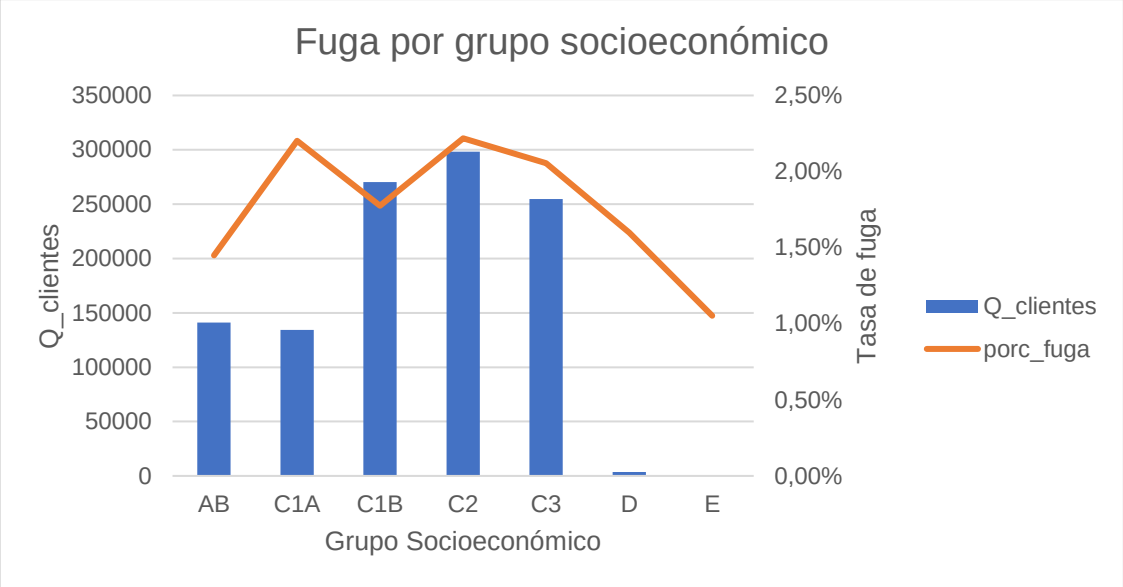


Gráfico 25: Fuga y distribución por grupo socioeconómico.

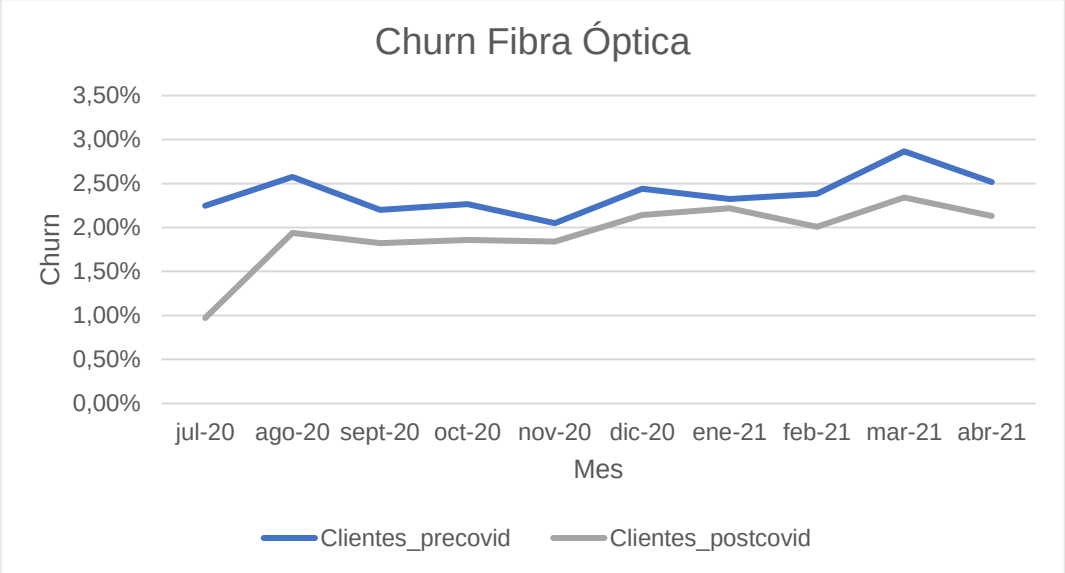


Gráfico 26: Evolución Churn para fibra óptica diferenciando por camadas pre y post pandémicas.

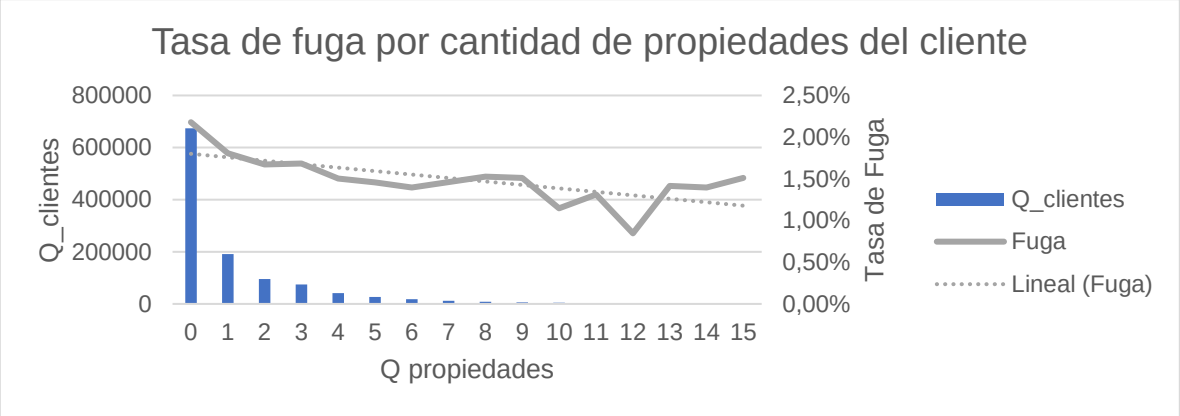


Gráfico 27: Tasa de fuga por cantidad de propiedades que poseen los clientes y sus cónyuges.

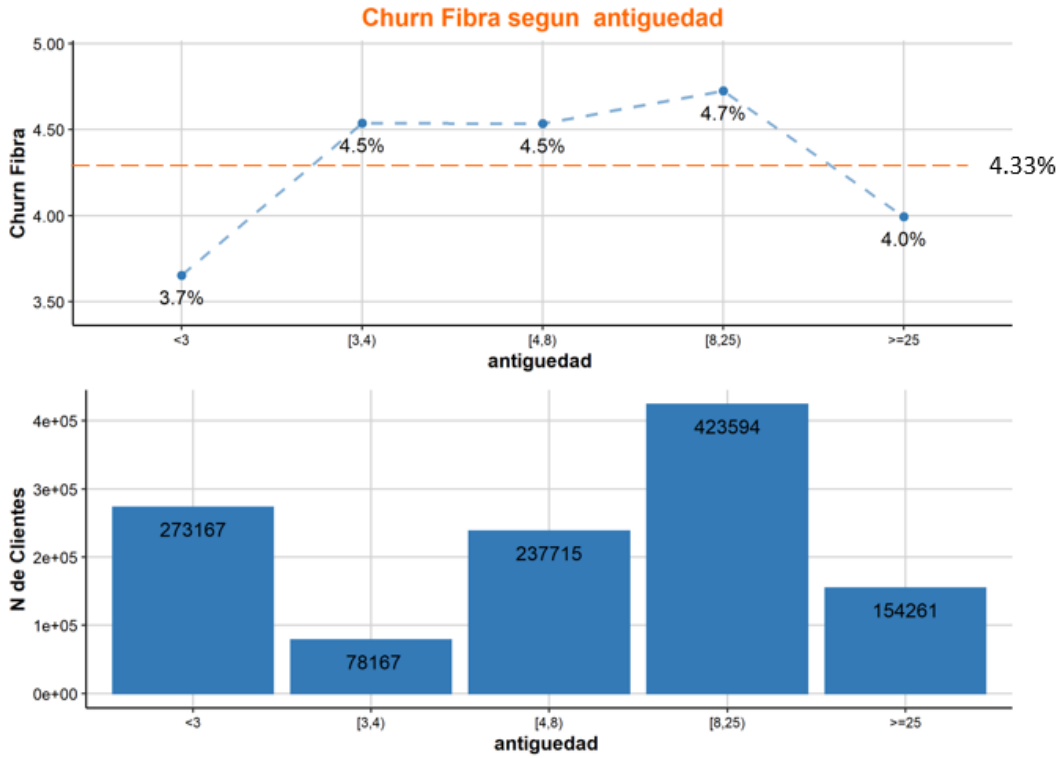


Gráfico 28: Distribución variable antigüedad en modelo de fuga.

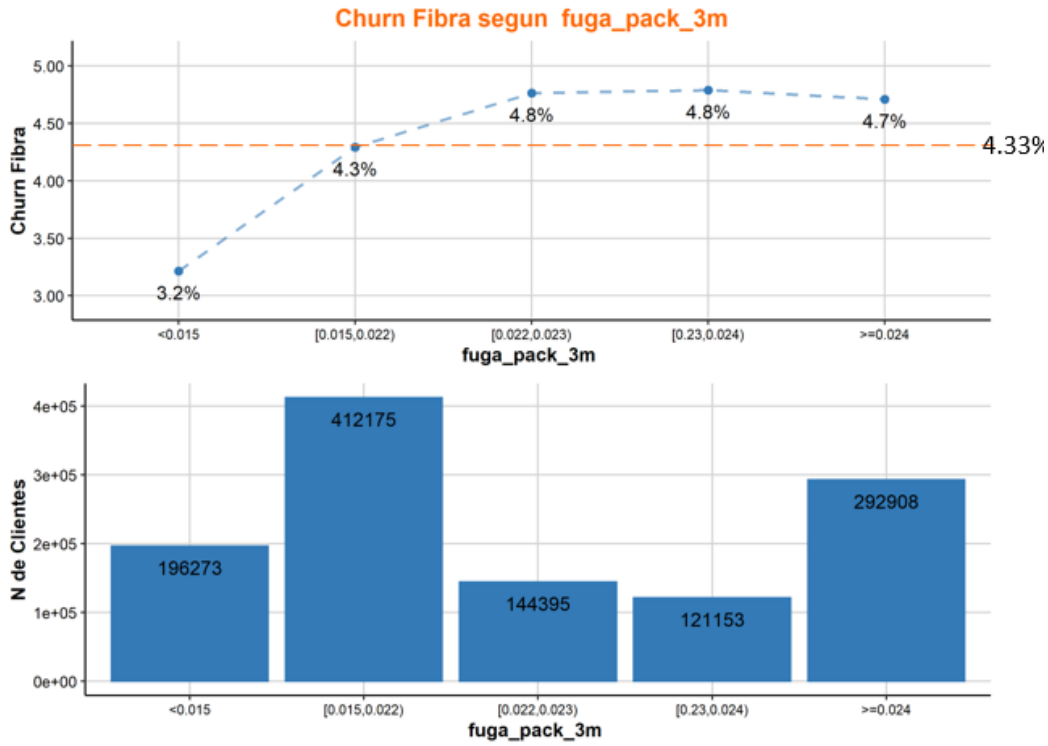


Gráfico 29: Distribución variable fuga pack de los últimos 3 meses en modelo de fuga.

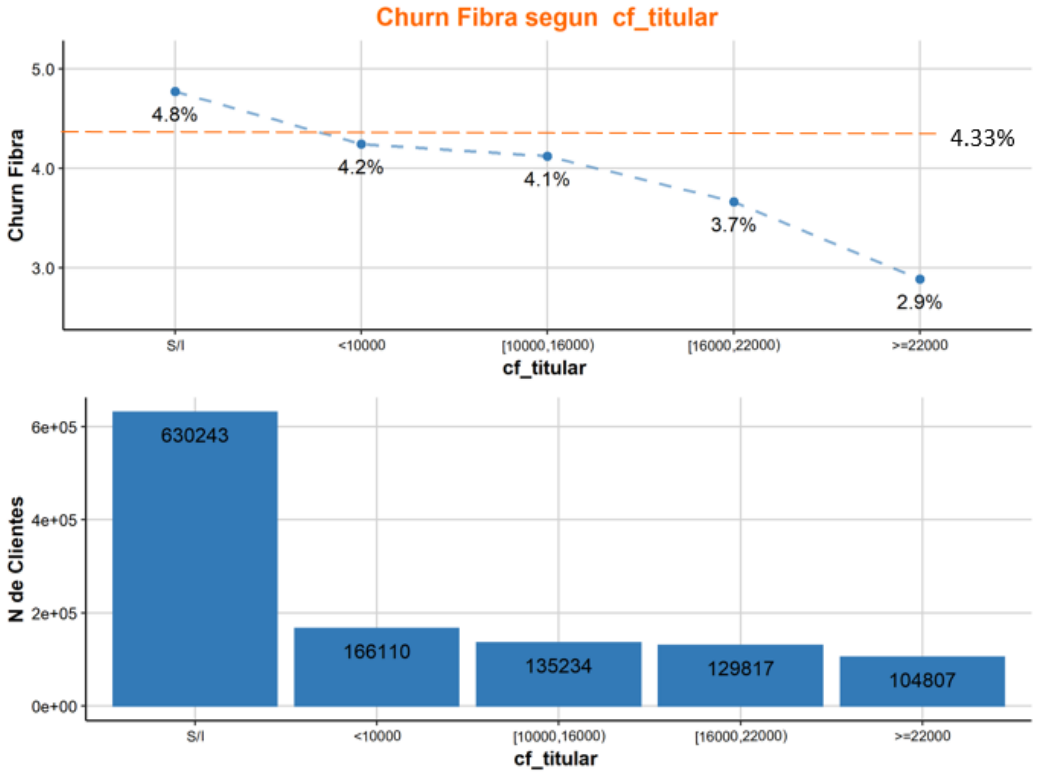


Gráfico 30: Distribución variable costo fijo móvil en modelo de fuga.

ANEXO C: MATRICES DE CONFUSIÓN DE MODELOS

		Valor real	
		Fuga	**No fuga**
Valor predicho	Fuga	1.439	5.260
	No fuga	4.181	125.835

		Valor real	
		Fuga	**No fuga**
Valor predicho	Fuga	1.443	3.631
	No fuga	4.177	127.464

Tablas 26 y 27: Matriz de confusión modelos GBM sin balanceo y random forest sin balanceo.

		Valor real	
		Fuga	**No fuga**
Valor predicho	Fuga	1.242	3.752
	No fuga	4.378	127.343

		Valor real	
		Fuga	**No fuga**
Valor predicho	Fuga	1.302	2.251
	No fuga	4.318	128.844

Tablas 28 y 29: Matriz de confusión modelos XRT sin balanceo y Stacked Ensemble AM sin balanceo.

		Valor real	
		Fuga	**No fuga**
Valor predicho	Fuga	1.302	2.251
	No fuga	4.318	128.844

		Valor real	
		Fuga	**No fuga**
Valor predicho	Fuga	4.721	78.765
	No fuga	899	52.330

Tablas 30 y 31: Matriz de confusión modelos Stacked Ensemble BOF sin balanceo y GBM con undersampling.

		Valor real	
		Fuga	**No fuga**
Valor predicho	Fuga	3.407	41.604
	No fuga	2.213	89.491

		Valor real	
		Fuga	**No fuga**
Valor predicho	Fuga	3.447	45.143
	No fuga	2.173	85.952

Tablas 32 y 33: Matriz de confusión modelos RF con undersampling y XRT con undersampling.

		Valor real	
		Fuga	**No fuga**
Valor predicho	Fuga	2.194	13.945
	No fuga	3.426	117.150

		Valor real	
		Fuga	**No fuga**
Valor predicho	Fuga	2.160	13.194
	No fuga	3.460	117.901

Tablas 34 y 35: Matriz de confusión modelos Stacked Ensamble AM con undersampling y Stacked Ensamble BOF con undersampling.

		Valor real	
		Fuga	**No fuga**
Valor predicho	Fuga	4.731	87.332
	No fuga	889	43.763

		Valor real	
		Fuga	**No fuga**
Valor predicho	Fuga	998	1.940
	No fuga	4.622	129.155

Tablas 36 y 37: Matriz de confusión modelos GBM con oversampling y RF con oversampling.

		Valor real	
		Fuga	**No fuga**
Valor predicho	Fuga	1.883	11.006
	No fuga	3.737	120.089

		Valor real	
		Fuga	**No fuga**
Valor predicho	Fuga	1.059	2.244
	No fuga	4.561	128.851

Tablas 38 y 39: Matriz de confusión modelos XRT con oversampling y Stacked Ensemble AM con oversampling.

		Valor real	
		Fuga	**No fuga**
Valor predicho	Fuga	1.072	2.305
	No fuga	4.548	128.790

Tabla 40: Matriz de confusión modelo Stacked Ensemble BOF con oversampling.

ANEXO D: RESULTADOS PREDICCIÓN EN NUEVOS DATOS

Percentil	Total Clientes	Cantidad de fugas	Tasa fuga	Mínimo score	Máximo score	LIFT
1	1476	77	5,22%	0,190694864	0,548509903	3,28
2	1476	48	3,25%	0,168514484	0,19068703	2,04
3	1476	44	2,98%	0,155443227	0,168452957	1,87
4	1476	49	3,32%	0,145815206	0,155438324	2,09
5	1476	35	2,37%	0,138460116	0,145814711	1,49
6	1476	33	2,24%	0,131923084	0,138453521	1,41
7	1476	36	2,44%	0,126743968	0,131920636	1,53
8	1476	38	2,57%	0,122089028	0,126741352	1,62
9	1476	35	2,37%	0,118032738	0,122081116	1,49
10	1476	45	3,05%	0,114297786	0,118030242	1,92
11	1476	28	1,90%	0,111021951	0,114297246	1,19
12	1476	43	2,91%	0,107999322	0,111015225	1,83
13	1476	28	1,90%	0,105264367	0,107998397	1,19
14	1476	34	2,30%	0,102649956	0,105263441	1,45
15	1476	27	1,83%	0,100160502	0,102649736	1,15
16	1476	21	1,42%	0,097608204	0,100159909	0,89
17	1476	37	2,51%	0,095267783	0,097607059	1,58
18	1476	29	1,96%	0,093173311	0,095266945	1,24
19	1476	30	2,03%	0,091288373	0,093171588	1,28
20	1476	30	2,03%	0,089399457	0,091287798	1,28
21	1476	28	1,90%	0,08757728	0,089398886	1,19
22	1476	32	2,17%	0,085901238	0,08757645	1,36
23	1476	25	1,69%	0,084293263	0,0858991	1,06
24	1476	25	1,69%	0,082632899	0,084292774	1,06
25	1476	24	1,63%	0,08100859	0,082631868	1,02
26	1477	20	1,35%	0,079356189	0,081007237	0,85
27	1476	30	2,03%	0,077717429	0,079356087	1,28
28	1475	28	1,90%	0,07621714	0,077717366	1,19
29	1476	21	1,42%	0,074841449	0,076216939	0,89
30	1476	32	2,17%	0,073504198	0,074838221	1,36
31	1475	27	1,83%	0,072148095	0,073501983	1,15
32	1475	25	1,69%	0,070918131	0,072148095	1,07
33	1476	35	2,37%	0,069716668	0,070915896	1,49
34	1475	23	1,56%	0,068556482	0,069716531	0,98
35	1475	25	1,69%	0,067353425	0,068556235	1,07
36	1475	28	1,90%	0,066226141	0,067353132	1,19
37	1475	32	2,17%	0,065110681	0,066224002	1,36
38	1475	19	1,29%	0,064060438	0,065110266	0,81

39	1475	17	1,15%	0,062979318	0,064059467	0,72
40	1475	17	1,15%	0,061943177	0,062978884	0,72
41	1475	25	1,69%	0,060872446	0,061941754	1,07
42	1475	17	1,15%	0,059796416	0,060872352	0,72
43	1476	27	1,83%	0,05865109	0,059796087	1,15
44	1475	22	1,49%	0,057530282	0,058650503	0,94
45	1475	23	1,56%	0,05644857	0,057530247	0,98
46	1475	24	1,63%	0,055421028	0,056448078	1,02
47	1475	22	1,49%	0,054441161	0,055420643	0,94
48	1475	20	1,36%	0,053475217	0,05444106	0,85
49	1475	23	1,56%	0,052545574	0,05347455	0,98
50	1475	21	1,42%	0,051674377	0,052544812	0,89
51	1476	18	1,22%	0,050792091	0,051674311	0,77
52	1475	20	1,36%	0,049907193	0,050791941	0,85
53	1476	24	1,63%	0,049024955	0,049907054	1,02
54	1475	24	1,63%	0,048102286	0,049024539	1,02
55	1475	20	1,36%	0,047263069	0,048101432	0,85
56	1475	19	1,29%	0,046428661	0,047263026	0,81
57	1475	17	1,15%	0,045577967	0,046427389	0,72
58	1475	20	1,36%	0,044752105	0,045577585	0,85
59	1475	21	1,42%	0,043961788	0,044752016	0,89
60	1475	19	1,29%	0,043131015	0,043961715	0,81
61	1475	20	1,36%	0,042337489	0,043130934	0,85
62	1475	20	1,36%	0,041537226	0,042337416	0,85
63	1475	16	1,08%	0,04065414	0,041536839	0,68
64	1475	17	1,15%	0,039770359	0,040653143	0,72
65	1475	21	1,42%	0,03884131	0,039770279	0,89
66	1476	23	1,56%	0,03790056	0,038841004	0,98
67	1475	14	0,95%	0,036992362	0,037900119	0,60
68	1475	18	1,22%	0,036104814	0,036992141	0,77
69	1475	22	1,49%	0,035230421	0,036104769	0,94
70	1475	18	1,22%	0,034352942	0,035229681	0,77
71	1475	16	1,08%	0,033568665	0,034352713	0,68
72	1475	15	1,02%	0,032735643	0,03356863	0,64
73	1475	24	1,63%	0,031915653	0,032735103	1,02
74	1475	21	1,42%	0,031144221	0,031915449	0,89
75	1475	16	1,08%	0,030316751	0,031143765	0,68
76	1475	21	1,42%	0,029499922	0,030316602	0,89
77	1475	20	1,36%	0,028722839	0,029499749	0,85
78	1475	18	1,22%	0,027910271	0,028722756	0,77
79	1475	17	1,15%	0,02711562	0,027910212	0,72
80	1475	8	0,54%	0,026276504	0,027115248	0,34
81	1475	18	1,22%	0,025420707	0,026275505	0,77

82	1475	27	1,83%	0,024609433	0,025419381	1,15
83	1476	26	1,76%	0,023776042	0,024607677	1,11
84	1475	12	0,81%	0,022901249	0,023775529	0,51
85	1475	13	0,88%	0,021985908	0,02290118	0,55
86	1475	15	1,02%	0,021046942	0,021985824	0,64
87	1475	11	0,75%	0,019934213	0,021046442	0,47
88	1475	18	1,22%	0,01862614	0,019934027	0,77
89	1475	13	0,88%	0,017282908	0,018624351	0,55
90	1475	9	0,61%	0,016059281	0,017282234	0,38
91	1475	17	1,15%	0,014733716	0,016059011	0,72
92	1475	21	1,42%	0,013469203	0,014733639	0,89
93	1475	18	1,22%	0,012194901	0,01346697	0,77
94	1475	18	1,22%	0,010852453	0,012194072	0,77
95	1475	14	0,95%	0,009468961	0,010851425	0,60
96	1475	8	0,54%	0,008059285	0,009468172	0,34
97	1475	14	0,95%	0,006616081	0,008059162	0,60
98	1475	9	0,61%	0,005018009	0,006615965	0,38
99	1475	10	0,68%	0,003037951	0,005016059	0,43
100	1475	5	0,34%	0	0,003037227	0,21

Tabla 41: Resultados por percentil de propensión para fugas entre el 1 y el 22 de noviembre de 2021.