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

MODELO PREDICTIVO PARA PORTABILIDAD DE LÍNEAS ADICIONALES MÓVILES POSTPAGO PARA UNA EMPRESA DE TELECOMUNICACIONES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL

MARÍA JOSÉ MUÑOZ HERRERA

PROFESOR GUÍA:
PABLO MARÍN VICUÑA

MIEMBROS DE LA COMISIÓN:
CAROLINA SEGOVIA RIQUELME

JOSÉ ANTONIO NALDA REYES

SANTIAGO DE CHILE
2022

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE: Ingeniero Civil Industrial
POR: María José Muñoz Herrera
FECHA: 2022
PROFESOR GUÍA: Pablo Marín Vicuña

MODELO PREDICTIVO PARA PORTABILIDAD DE LÍNEAS ADICIONALES MÓVILES POSTPAGO PARA UNA EMPRESA DE TELECOMUNICACIONES

Una de las estrategias comerciales que se utiliza dentro de la industria de las telecomunicaciones es la oferta de servicios postpago, adicionales a la línea titular, lo cual permite aumentar el volumen de clientes y la penetración de la compañía dentro de un grupo familiar. El impacto de esta estrategia se observa al comparar tasas de fuga, en donde aquellos hogares que poseen una penetración de un 100% registran una tasa de fuga menor que aquellos que no.

La motivación de este trabajo consiste en aumentar la penetración dentro de los hogares que poseen al menos un cliente postpago de una compañía telefónica. Por ello, se construye un modelo capaz de identificar clientes con mayor propensión a contratar líneas adicionales portadas, con el fin de mejorar la segmentación para la oferta de servicios existentes de la empresa. Dentro del estudio realizado se evidencia la importancia que representa las portabilidades numéricas para la empresa, debido a su baja tasa de fuga en comparación a otros tipos de contrataciones de líneas (migraciones o activaciones).

Parte de los análisis realizados en el estudio tuvo como objetivo analizar la influencia de clientes titulares dentro de su red social y el impacto en la adquisición de nuevos servicios. Sin embargo, no se obtuvieron resultados concluyentes para la predicción de líneas adicionales portadas. Como trabajos futuros se propone retomar el estudio de externalidades de red a través de la identificación de perfiles de clientes utilizando los registros de llamada o CDR, abordando el problema desde otra mirada.

Se evaluaron diferentes algoritmos de Machine Learning (GBM, RF y XGBoost) para la predicción del problema estudiado. El modelo que obtuvo una mayor capacidad predictiva e interpretativa corresponde a un XGBoost sin balanceo de datos, el cual posee un buen desempeño. Se obtiene un AUC cercano a 0.68 y se observa un correcto ordenamiento de los deciles de clientes a partir de la curva Lift acumulada (estrictamente decreciente). El perfil del cliente con mayor propensión es aquel que llegó a la compañía hace menos de 6 meses portando su línea titular, que no posee más de dos líneas asociadas a su cuenta y con un alto nivel de interacción con clientes de la competencia a través de llamadas telefónicas.

A partir del modelo seleccionado, se plantea un experimento de segmentación de oferta dentro del Call Center de la empresa, esperando confirmar los hallazgos encontrados.

A mis padres y hermanos,

por todo el amor y apoyo incondicional que me han dado.

A mis amigos,

por haberse cruzado en mi camino universitario.

A mi Pancha,

alma perruna que me alegra todos los días.

A mis compañeros de trabajo,

por haberme dado la oportunidad de aprender y crecer como profesional.

TABLA DE CONTENIDO

INDICE DE ILUSTRACIONES

INDICE DE TABLAS

INDICE DE ECUACIONES

1. INTRODUCCIÓN

1.1. CONTEXTO DE LA EMPRESA

El desarrollo del trabajo se encuentra inserto dentro de una empresa de telecomunicaciones chilena, líder dentro del sector industrial. Esta cuenta con operaciones en dentro de dos países latinoamericanos, acumulando hasta el año 2021 aproximadamente 10,2 y 9,9 millones de clientes en telefonía móvil en cada país respectivamente.

El sector industrial y la empresa se rigen bajo la Ley General de Telecomunicaciones N.º 18.168 en donde se define lo que es telecomunicación, se especifica el derecho libre de acceso a las telecomunicaciones para todos los habitantes del país, se designa a SUBTEL como el organismo de control y cumplimiento del marco regulatorio y se exponen tres puntos respecto al cumplimiento tratados de libre competencia y barreras de entrada al mercado: acceso y concesiones (la opción de poder optar a concesiones y permisos para el acceso libre e igualitario en telecomunicaciones para todos los actores del sector), interconexión (entre servicios públicos y los entregados por las distintas compañías de telefonía) y procesos tarifarios para la regulación de precios dentro del sector.

Esta empresa, cuenta con servicios de telefonía móvil y fija, redes de datos, internet y servicios segmentados según personas, empresas y mayoristas. Los ingresos consolidados de su operación en el país para el año 2021 fueron de aproximadamente $2.460.119 millones de pesos, con un EBITDA de $772.452 millones (un aumento de aproximadamente 10% con respecto al año 2020). Se observa un leve aumento en cuanto a la inversión y los ingresos desde el año 2019, lo que coincide con la expansión del mercado de fibra óptica en el país (inversión en infraestructura principalmente).

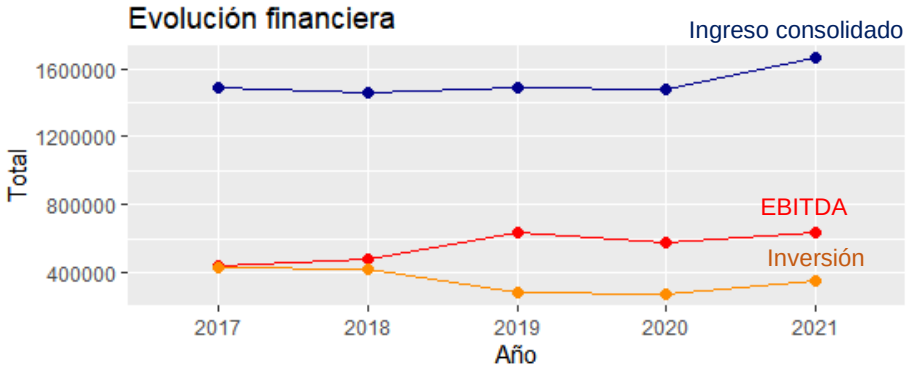


Ilustración 1: Evolución financiera de la empresa. Elaboración propia.

La empresa se divide en 4 segmentos de clientes: Personas, Empresas, Corporaciones y Mayorista. El presente trabajo se encuentra inserto dentro del primer segmento, por lo que se ahondará en este.

El segmento “Personas” provee a 4,49 millones de clientes con servicios de datos y voz móviles postpago (con contrato mensual) y 6,64 millones de clientes con telefonía y datos móviles prepago. Dentro de este mismo segmento se ofrecen servicios hogar como fibra óptica y servicios de valor agregado tales como Carrier Billing (servicios de Streaming como Netflix, Spotify, Google Play), Carrier a larga distancia internacional, venta de equipos y servicios financieros (seguros de viaje, SOAP, etc).

Según la distribución de ingresos por servicios en la industria de Telecomunicaciones en el gráfico 1, se observa que los servicios de datos y voz móviles dentro del segmento personas representan el mayor porcentaje de ingresos para las compañías telefónicas, alcanzando el 28,9% de las ganancias totales. Dentro del segmento de telefonía móvil, la empresa posee una participación de mercado de 32,2% (aumento de 0,3% respecto al año 2021), seguido por la competencia 1 C1 (24.9%), competencia 2 C2 (21.3%) y competencia 3 C3 (20).

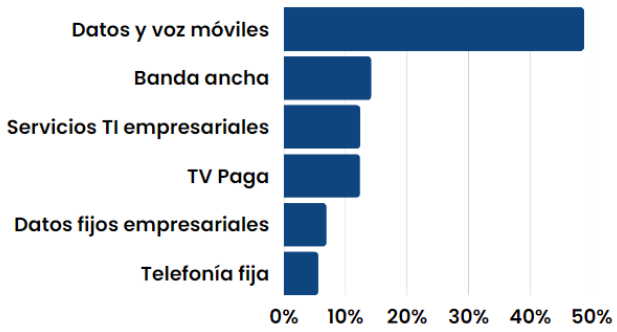


Ilustración 2: Distribución de ingresos por servicio en la industria. Elaboración propia.

Se poseen clasificaciones de clientes dentro del segmento de datos y voz móviles según el servicio que adquiere: prepago, refiriéndose a aquellos clientes que no poseen un contrato mensual pero que pueden acceder a bolsas de navegación (internet móvil) y minutos para llamadas, y clientes postpago, quienes tienen asociado una facturación mensual por planes de telefonía y datos móviles de manera mensual. El promedio de clientes postpago que mantuvo la empresa durante el año 2021 fue de 4,3 millones. Este número de cliente ha aumentado levemente a medida en el transcurso del tiempo.

Profundizando en la segunda clasificación, las líneas de clientes Postpago pueden provenir de 3 diferentes orígenes: líneas activadas o habilitadas (aquellos números nuevos que son obtenidos mediante planes de telefonía), líneas migradas desde prepago

a postpago (contratación de un plan de telefonía móvil desde un número existente prepago) y líneas portadas (números existentes en donde el cliente pertenecía a otra compañía previamente). Se ahondará en esta última clasificación en la siguiente sección.[1]

2.2. PORTABILIDAD NUMÉRICA

Desde sus inicios en 1990, hasta su llegada a Chile en 2012, la portabilidad numérica móvil se implementó con el objetivo de mejorar la calidad de servicio que perciben los clientes al intensificar la competitividad dentro de la industria de las telecomunicaciones. Es definida como la habilidad de aquellos “suscriptores” (clientes) de retener su número telefónico al momento de migrar desde un proveedor de servicios a otro, disminuyendo así el costo asociado a cambiarse a un nuevo número. [2]

Se aprecia una alta competitividad dentro del segmento de telefonía móvil postpago potenciado en parte por la llegada de la portabilidad, aumentando desde el año 2013 y superando las portabilidades de telefonía prepago. Esta categoría de línea es de especial interés dentro de la empresa debido a que no solo implica la adquisición de un cliente, si no que disminuye la ventaja de la competencia dentro de la industria al disminuir su participación dentro del mercado de voz y datos móviles.

En el siguiente grafico se observa la evolución de portabilidades entrantes (denominado “Port In”) correspondiente a los clientes que son adquiridos desde otras compañías y las portabilidades salientes (denominado “Port out”) correspondiente a la fuga voluntaria por parte de clientes de la empresa. En general la portabilidad neta ha sido positiva, excluyendo 4 meses que se pueden apreciar en el gráfico, en donde la portabilidad neta fue negativa.

[1] La información utilizada en la contextualización fue obtenida a partir del reporte integrado de la empresa correspondiente al año 2020. [1]

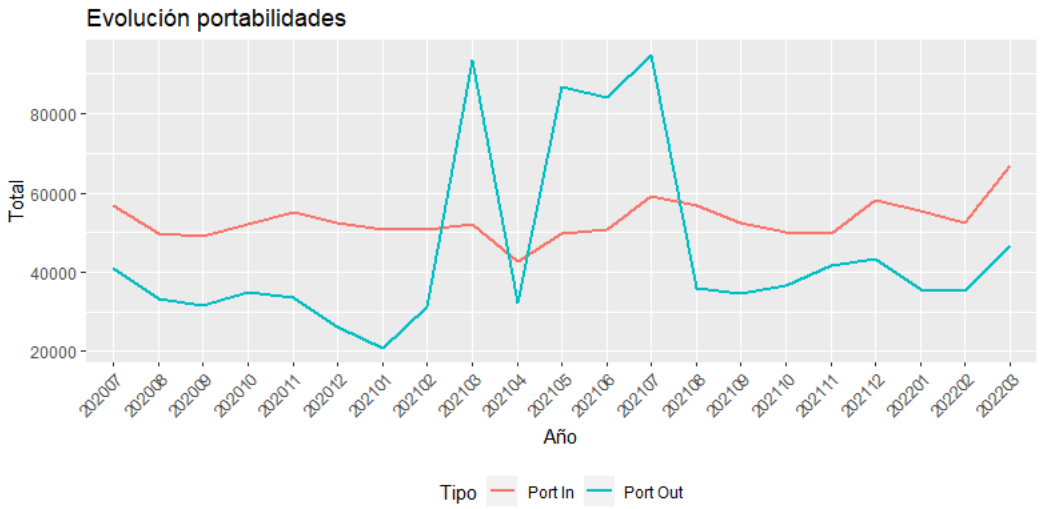


Ilustración 3: Evolución portabilidades. Elaboración propia.

2.3. PROBLEMÁTICA/MOTIVACIÓN

Respecto a los clientes postpago, se mantiene aproximadamente 4,3 millones de clientes de manera mensual, en donde según la actividad comercial reportada para cada mes, se captan 134 mil líneas en promedio. Tal como se menciona anteriormente, estas líneas pueden provenir desde una portabilidad, activación o migración.

Un 37% del total de ventas mensuales corresponden a líneas adicionales. Se entiende por línea adicional o segunda línea aquella que es asociada a un cliente titular que ya poseía un número móvil postpago. La variación de líneas adicionales según su clasificación de origen se aprecia en la ilustración 4. En esta imagen se puede observar que en su mayoría la actividad comercial se centra en activaciones de línea por sobre portabilidades, las cuales en promedio corresponden a un 20% del total de líneas adicionales.

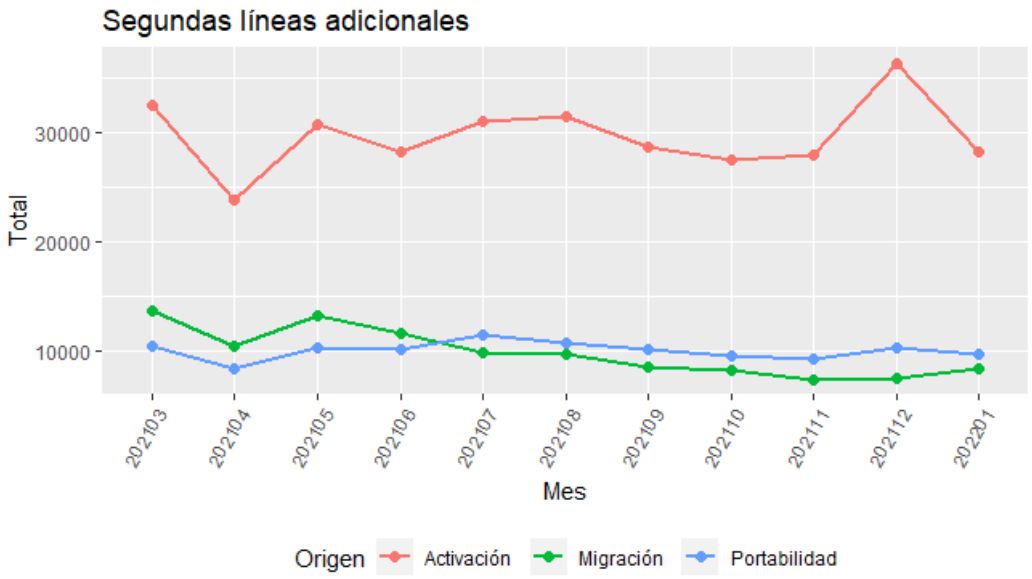


Ilustración 4: Evolución de segundas líneas a través del tiempo. Elaboración propia

La adquisición de líneas adicionales por parte de clientes titulares está relacionada con la estrategia comercial de la empresa denominada “Convergencia Hogar”. Este concepto se refiere al proceso de captar a todos los clientes identificados dentro de un grupo familiar, aumentando la penetración que posee la empresa dentro del grupo familiar. Este grupo familiar está definido a grandes rasgos según los siguientes criterios definidos por la empresa:

- Menores: todo individuo hasta los 26 años que no se encuentra casado ni posee un vínculo con otra persona debido al nacimiento de un hijo, forma un grupo familiar con sus padres.

- Adultos: personas que se encuentran casadas, poseen un vínculo por el nacimiento de un hijo o que tiene más de 26 años pasan a formar un grupo familiar.

- Divorciadas/os con hijos: forman un grupo familiar con todos los hijos menores cumpliendo el primer punto.

- Solteros/as con hijos: forman un grupo familiar en conjunto a los hijos que no estén casados, no tengan hijos o tengan menos de 26 años.

Según estos criterios de segmentación, se identifican 8.720.542 hogares (con al menos un integrante vivo), de los cuales el 27% tiene un integrante que es titular de un plan postpago o fibra (noviembre 2021). Existe una brecha entre el número de hogares

identificados por la empresa y los hogares identificados por el Censo realizado en 2017, sin embargo, se considera una estimación razonable a partir de validaciones realizadas dentro de la empresa.

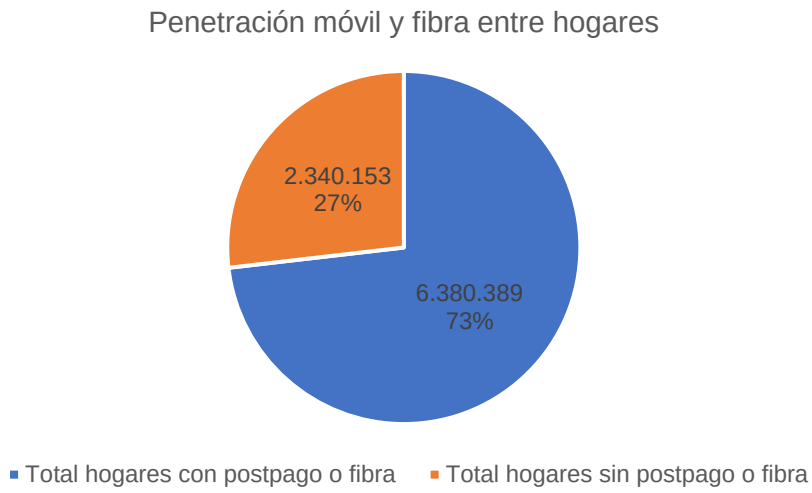


Ilustración 5: Identificación de hogares dentro de Malla Parental. Gráfico obtenido de presentación interna del área.

Esta asignación de grupos familiares se consolida dentro de una base denominada “Malla Parental”, de la cual surgen dos tipos de segmento hogar:

- Hogar abierto: grupo familiar en donde la suma de líneas que poseen los clientes titulares es menor al número de miembros. Un 46% de los hogares identificados corresponden a este segmento.

- Hogar cerrado: suma de líneas de los clientes titulares es mayor o igual a la cantidad de personas que componen el grupo familiar, es decir, la penetración de la empresa en este hogar es de un 100%. Un 54% de los hogares identificados corresponden a este segmento

La convergencia del hogar se encuentra dentro de los lineamientos estratégicos para la empresa, teniendo como objetivo aumentar el número de hogares cerrados debido a que disminuye la ventaja de los competidores (dentro de una familia no existiría puntos de comparación entre compañías lo que dificultaría la fuga), reduce el riesgo de contactabilidad de los competidores con miembros del grupo familiar y genera un “Lock-in” del cliente (accede a mejores beneficios, fugarse de la compañía con todas las líneas es un proceso más tedioso, etc).

El efecto de esta estrategia comercial se observa en la variación de tasas de fuga que se presenta en el siguiente gráfico, en donde el segmento cerrado se caracteriza por tener una menor tasa (1.7% para el segmento cerrado y 2.3% abierto respectivamente). Esto sustenta la ventaja mencionada anteriormente.

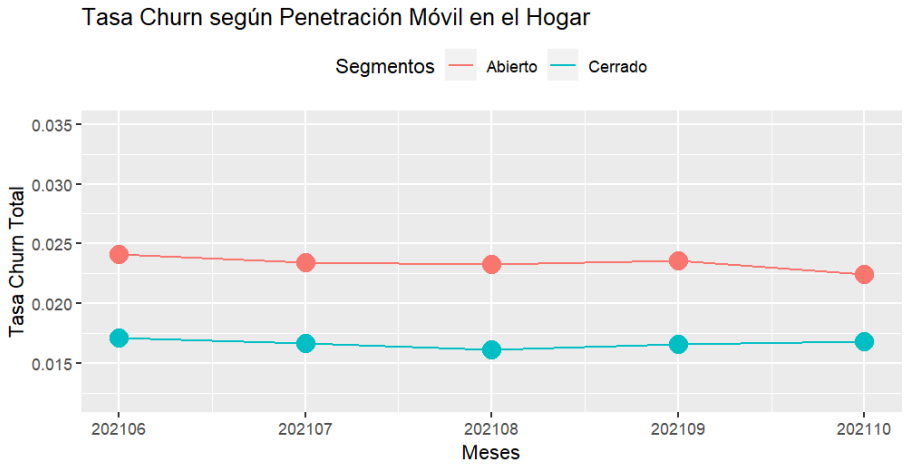


Ilustración 6: Variación de tasas de fuga según segmento hogar dentro del negocio móvil Postpago. Gráfico obtenido de presentación interna del área.

Se analiza además las características de los tipos de líneas que un cliente puede obtener, en base al cargo fijo inicial promedio de clientes que adquirieron una línea adicional en el último año y la tasa de fuga al sexto mes.

- Las líneas adicionales activadas representan un 60% de las ventas mensuales. Los nuevos clientes adquieren un plan con cargo fijo promedio de $14.197 y tienen una tendencia a disminuir este en un periodo de 6 meses de permanencia posterior a su llegada. Su tasa de fuga acumulada luego sexto mes es de 27,09%.

- Las líneas adicionales portadas representan un 20% de las ventas mensuales. Los clientes adquieren un plan con cargo fijo promedio cercano a las activaciones de línea con una tendencia a disminuir después de 6 meses. La tasa de fuga de esta categoría es de 12,78% luego del sexto mes.

- Las líneas adicionales migradas representan cerca de un 20% de las ventas mensuales, al igual que las portabilidades. En contraste a las otras categorías, el cargo fijo promedio que contratan estos clientes postpago nuevo es inferior a los $14.000 con una tasa de fuga acumulada al sexto mes de 26,69%.

Se observa un cargo fijo inicial similar entre las líneas portadas y activadas, sin embargo, las líneas que llegan mediante portabilidades poseen una tasa de fuga menor

al sexto mes. Estos clientes son relevantes para la empresa debido a que poseen una permanencia mayor, aportando ingresos similares a las activaciones.

Se analiza la distribución de líneas adicionales según el canal por el cual fueron adquiridas. Las tiendas propias, Inbound (Call Center con llamadas entrantes de clientes), plataformas online y Outbound (Call Center con llamadas salientes desde la empresa) son las que poseen una mayor cantidad de ventas de líneas adicionales.

Es importante mencionar que debido a la naturaleza de los primeros 3 canales, los cuales están orientados a la llegada orgánica de clientes, es difícil manejar una compaña de ventas enfocada en la contactabilidad de clientes. Por esto, se decide abordar un despliegue enfocado en el canal Outbound, el cual tiene un mayor porcentaje de portabilidades de líneas y efectividad de campañas de líneas adicionales (48,98% del total de ventas efectivas).

Actualmente solo se cuenta con un modelo que apoya la venta de segundas líneas, independiente del tipo de producto que se venta. Se espera que el desarrollo del proyecto permita una mejora en la segmentación de servicios móviles existentes de la compañía, con foco en clientes que sean propensos a adquirir segundas líneas postpago.

El cliente interno de la empresa corresponde al área de venta en canales, quienes en base a la propensión entregada por el modelo generarán el mix de canales por los cuales se gestionarán los diferentes grupos de clientes. Este mix tiene en consideración los costos del canal, la capacidad de gestión y los cargos fijos que se cobrarán para el cálculo de los ingresos.

2. OBJETIVOS

2.1. OBJETIVO GENERAL

Identificar clientes titulares postpago propensos a contratar una línea adicional portada utilizando modelos predictivos, para mejorar la segmentación de la oferta de servicios existentes pertenecientes a una empresa de telecomunicaciones.

2.2. OBJETIVOS ESPECÍFICOS

- Determinar variables explicativas para el problema, con el fin de identificar el comportamiento de clientes titulares en base a la decisión de portar una línea adicional.

- Evaluar el aporte de la información proveniente del registro de llamados a través del estudio de externalidades de red utilizando teoría de grafos.

- Evaluar modelos de predicción realizados, seleccionando aquel que posea una mayor capacidad predictiva e interpretativa para el caso de estudio.

- Proponer un experimento enfocado en mejorar las campañas de segundas líneas portadas, enviando mensualmente a los canales de contacto, clientes que sean propensos a adquirir un nuevo servicio de telefonía movil junto con una oferta diferenciada para portabilidades a partir de perfiles identificados.

3. ALCANCES

- Se analiza el fenómeno de líneas adicionales portadas enfocado en clientes titulares según la disponibilidad de información, de quienes se utiliza información 8 meses para el entrenamiento/testeo del modelo. Es decir, para el cálculo de propensión de junio 2022 se utilizarán datos desde septiembre 2021 hasta abril del 2022. Se utiliza esta limitación de meses para entrenar debido a que se mantiene 1 año de historia en la mayoría de las bases de datos productivas para agilizar el trabajo de los datos y liberar espacio.

- Debido al volumen y disponibilidad de la data correspondiente al registro de llamadas, se realiza un análisis de interacciones sociales utilizando data histórica de 4 meses (desde diciembre 2021 hasta marzo 2022). Esta selección es debido a la disponibilidad de la data procesada disponible y el elevado costo de procesamiento que generaría utilizar una mayor cantidad de data para el estudio preliminar.

- La evaluación del modelo que incluye las variables derivadas del análisis social es realizada mediante la comparación de métricas a partir de un modelo base entrenado y testeado con los 4 meses correspondientes.

- El experimento propuesto está enfocado en el canal de Call Center la factibilidad de contactarse directamente con los clientes perfilados con una mayor propensión a la portabilidad de líneas adicionales. El experimento no es implementado durante el desarrollo del trabajo de título y queda propuesto para la empresa.

4. MARCO TEÓRICO

4.1. LITERATURA PREVIA

Para el desarrollo del proyecto se han consultado ciertos estudios con el fin de entender el comportamiento del cliente y su intención de cambio o portabilidad dentro de los servicios de telefonía móvil al analizar variables que podrían estar explicando este fenómeno. La portabilidad numérica tiene como objetivo el disminuir los costos asociados a realizar un cambio de número telefónico, sin embargo, existen clientes que aún perciben estos costos muy altos. Los cuestionamientos que se han planteado en estos estudios pretenden responder preguntas tales como ¿Quién cambia y quién no? y ¿cómo las barreras de cambio y la satisfacción del cliente determinan la intención de migrar a un operador diferente? [3][4]

Variables consideradas en la bibliografía consultada que podrían estar influyendo en la intención o propensión de portabilidad en un cliente corresponden a las siguientes:

- Percepción de las barreras de cambio. Se sostiene la hipótesis que aquellos clientes que perciben altos niveles de barreras moderan la intención de cambio, inclusive si estos no están satisfechos con la calidad del servicio que reciben de su proveedor actual.

- Nivel de satisfacción, relacionada a la puntuación que otorga el cliente a la marca o proveedor basada en todos los encuentros e interacciones que históricamente ha tenido con la empresa. Se plantea la hipótesis de que aquellos clientes que poseen altos niveles de satisfacción son menos propensos a migrar en un futuro.

- Calidad del servicio, siendo esta característica algo intangible, percibida por el cliente como la impresión general relativa a la eficiencia de su proveedor de servicios. Es esperable que aquellos clientes que perciben una mejor calidad de servicio sean más fieles.

- Costo percibido por cambio, asociado al tiempo, dinero y costo psicológico incurrido en el momento de realizar la migración. Se plantea la hipótesis que relaciona altos niveles de costos por cambio con niveles altos de barreras de cambio.

- Lock-in del suscriptor o cliente, entendiéndose como los esfuerzos que el proveedor incurre con el fin de crear barreras de cambio, tales como contratos, “Bundling” o empaquetamiento de servicios, programas de lealtad, entre otros. Se relaciona niveles altos de Lock-In de clientes con altos niveles de barreras de cambio.

- Variables demográficas, tales como características socioeconómicas, sexo, educación, edad, estilo de vida, entre otras. Se sostiene que la edad podría ser una variable influyente en la intención de cambio, siendo los mayores más reacios a migrar de una compañía a otra.

A partir del quinto punto, se entiende que la estrategia de agrupar a los clientes por grupos familiares y aumentar de esta manera las líneas móviles postpago dentro de este grupo funciona como Lock-in. En secciones anteriores se observa la diferencia en la tasa de fuga dentro de los hogares que poseen una penetración del 100%, por lo que una hipótesis detrás de este efecto podría ser el Lock-in que se genera dentro de las familias que poseen mayor número de servicios móviles, las cuales serían menos propensas a migrar a otra compañía a futuro.

Otros estudios relacionados a la adquisición de productos con externalidades de red abordan la influencia social (definida como la habilidad que posee un cliente en influir en la decisión de otro) junto a la tendencia por gustos similares (también definida como *"Latent Homophily")*, como factores que ayudan a modelar el comportamiento del cliente frente a la adquisición de un producto (en este caso, servicio adicional de telefonía móvil). Esto permite abordar el problema en cuestión desde el enfoque de análisis de redes sociales. [5]

En específico, la influencia social es considerada como un factor observable, pudiendo identificar características de los clientes e identificación de comunicades cuyo comportamiento y/o perfil pueda ser similar.

Según un análisis de interacciones sociales, una buena comunicación e interacción entre los clientes de una industria de telecomunicaciones puede influir en la disminución de la tasa de fuga al final de un periodo, entregando además un espacio para difusión de productos y servicios. Dentro de estos análisis se han utilizado métodos de *teoría de grafos* y medidas de centralidad con el fin de analizar la influencia de cada nodo dentro de una red. Variables que fueron consideradas dentro de este estudio hacen referencia al tráfico de llamadas realizadas por un cliente (CDR o Call Detail Record), información entregada por las torres o antenas telefónicas (como la ubicación del cliente) y variables sociodemográficas. [6]

Trabajar con la data obtenida de CDR representa un desafío en cuanto al volumen de datos. En estos casos, la detección de comunidades puede ser una herramienta útil para el estudio de comportamiento de ciertos nodos dentro de una red social, utilizando métodos de clustering para agrupar ciertos perfiles de clientes que sean propensos a portar segundas líneas. Se han utilizado diferentes métodos para la detección de comunidades tales como particiones de grafos enfocados en vértices con una medida de centralidad Betweenness mayor, maximización de modularidad, clustering de vértices convencional, entre otros. [7]

4.2. MODELOS

La información contenida dentro de esta sección fue elaborada a partir de consultas realizadas al texto *"Data Science and Machine Learning, Mathematical and Statistical Methods"* del autor Dirk P. Kroese, complementando con otras fuentes citadas. [8]

4.2.1 Árboles de regresión (CART)

Los árboles de clasificación y regresión, o también denominados CART, son básicamente árboles de decisión con particiones binarias, dividiendo el espacio de regresores en subconjuntos, con el fin de poder utilizar un modelo simple dentro de cada división. Este modelo busca cada valor distinto de cada variable de entrada para encontrar el predictor y el valor de división que separa los datos en dos regiones R_1 y R_2 (con valores medios c_1 y c_2) para minimizar la suma de los errores al cuadrado. Como resultado, cada nodo del árbol de regresión representa valores numéricos.

$$minSSE = \sum_{i \in R_1} (y_i - c_1)^2 + \sum_{i \in R_2} (y_i - c_2)^2 \qquad (1)$$

Con el objetivo de evitar modelos que sean demasiados complejos (es decir, con demasiadas ramas), al finalizar el proceso los árboles son "podados" (se penaliza el número de nodos terminales). Por ende, es preferible aquellos que poseen una construcción más simple, con menor número de nodos.

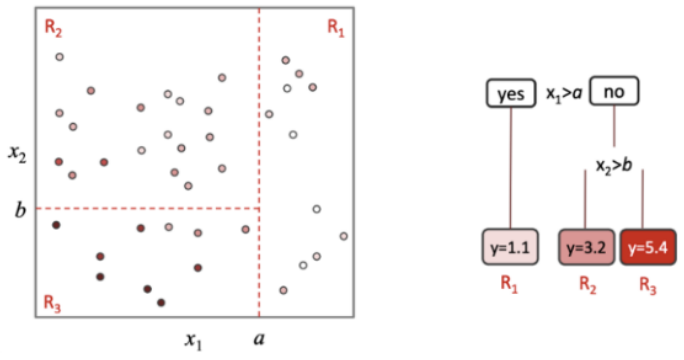


Ilustración 7: Ilustración gráfica del modelo CART

4.2.2 Random Forest

Random Forest corresponde a un algoritmo de clasificación el cual es usado para predecir una variable cualitativa por sobre una variable cuantitativa, como lo haría un árbol de regresión, aunque su construcción es similar a este último. Un árbol de clasificación predice si cada dato observado corresponde a la clase más común dentro de cierta región de los datos utilizados en el set de entrenamiento. Para interpretar los resultados de este modelo no solo es relevante la predicción de clasificación que realiza, sino también las proporciones de los datos que caen dentro de la región de entrenamiento.

Un problema de los árboles de regresión es que son sensibles a pequeñas variaciones en los datos. Es por esto que, para lidiar con este problema, se construyen varios árboles, en donde sus resultados son promediados para obtener la referencia de uno solo. Random Forest funciona de manera similar al proceso de Bagging o Boostrap Agregattion, pero no solo hace un muestreo de las observaciones, sino que también de las características.

Para cada muestreo de datos, se construye un árbol de decisión usando una elección aleatoria de características $n \leq p$ (siendo p el número total de variables independientes o características) como regla de separación. El algoritmo se observa a continuación.

- Sea el set de entrenamiento $\tau = (x_1, y_1)_{i=1}^{m}$, B el número de árboles, $n \leq p$ características a incluir aleatoriamente.

- Genera un muestreo de entrenamiento $\tau_1^*, \dots \tau_B^*$ para cada total de árboles a través del método de Boostrapping.

- Para cada set de entrenamiento, se selecciona aleatoriamente m características sin reemplazo.

- Usando estas características se entrena un árbol de decisión $g_{\tau_b^*}$

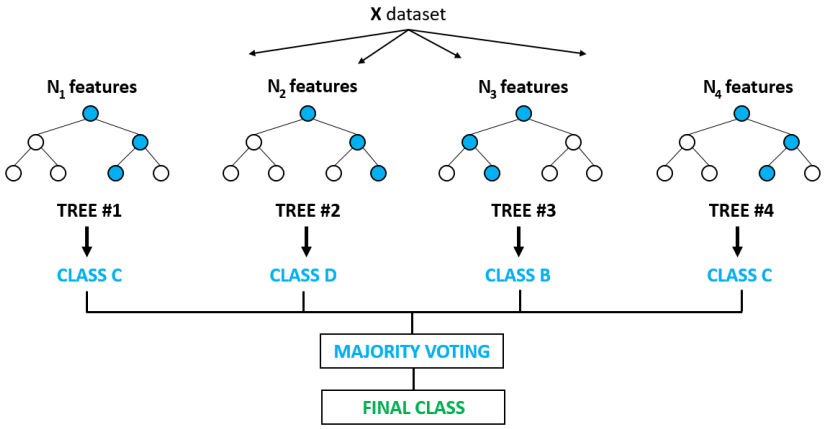


Ilustración 8: Ilustración gráfica del algoritmo Random Forest

4.2.3 Gradient Boosting Machine (GBM)

Boosting es un algoritmo de aprendizaje utilizado tanto para regresión como clasificación, el cual se entrena de manera secuencial a partir de clasificadores débiles, típicamente árboles de decisión como Random Forest. Gradient Boosting es una técnica la cual combina varios predictores débiles aditivamente, aprendiendo de los errores previos de los clasificadores para la construcción de un modelo con mejor predicción. [9]

GBM considera una función de pérdida a ser optimizada, un clasificador débil y un modelo que vaya agregando de manera progresiva estos clasificadores con el objetivo de minimizar la función de pérdida (la cual dependerá del problema a resolver, por ejemplo, pérdida logarítmica aplicable para este caso de estudio, en donde se desea predecir una probabilidad asociada a la decisión del cliente).

Este algoritmo suele obtener mejores resultados en comparación a otros modelos de predicción, sin embargo, se debe tener cierto cuidado con la preparación previa de los datos debido a que es sensible a outliers, tiende a tomar mayor tiempo de procesamiento y es propenso a realizar sobreajuste de los datos si es que el número de árboles es muy grande. Sin embargo, se considera como un buen modelo de clasificación para el problema a abordar, debido a su capacidad de aprender de los errores previos, pudiendo utilizar “Extreme Gradient Boosting” o XGBoost para agilizar el procesamiento con grandes cantidades de datos.

4.2.4 Teoría de grafos

Un grafo es definido como $G = (V, E)$, siendo V el conjunto de vértices o nodos (los cuales pueden representar individuos, organizaciones, objetos, etc) y E las interacciones

o edges (llamadas telefónicas, referencias dentro de un artículo de investigación, conexiones de vuelos, etc).

Existen diferentes características de un grafo basadas en el conjunto de edges o aristas, tales como un grafo no dirigido (la interacción o conexión no tiene una dirección) y los grafos dirigidos (interacción puede salir o entrar a un nodo en específico. Es posible además entregar un atributo de “peso” a cada arista de un grafo, basado en alguna variable relevante en el análisis (en este caso puede ser utilizada la duración de una llamada, frecuencia de llamados, etc). Los grafos también pueden estar compuestos por nodos pertenecientes a un solo tipo (homogéneos) o múltiples tipos (heterogéneos). En este caso se trabajará con un grafo heterogéneo dirigido debido a que es necesario representar la interacción entre clientes titulares y no clientes a los cuales puede llamar o del cual puede recibir llamadas. [10]

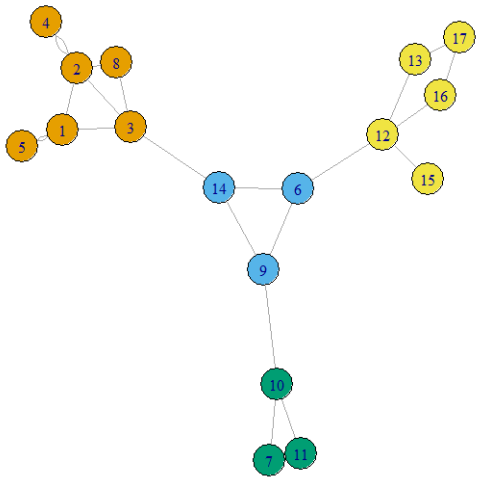


Ilustración 9: Representación de grafo no dirigido. Elaboración propia

Problemas que pueden ser resueltos mediante el análisis de un grafo pueden ser clasificaciones de nodos (inferencia de atributos para un nodo del cual no se posee información), predicciones de interacción (similar al caso anterior, pero determinando una posible conexión entre nodos) y detección de comunidades (clústers de nodos basado en sus similitudes e interacciones).

Existen métricas de centralidad que permiten analizar la importancia o influencia de un nodo dentro de una red social, los cuales son

- Degree centrality: cantidad de conexiones o vértices que posee un nodo, en donde d_v representa el grado del nodo v. Al dividir por el número total de nodos n menos 1 se normaliza esta medida de centralidad.

$$C_{deg}(v) = \frac{d_v}{n-1} \tag{2}$$

- Betweenness centrality: suma de los caminos que van desde s a t pasando por el nodo v ($\sigma_{s,t}(v)$) sobre el total de caminos más cortos entre s y t ($\sigma_{s,t}$)

$$C_{btw}(v) = \sum_{\{s,t \in N\}} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}} \tag{3}$$

- Closeness centrality: cercanía de un nodo con respecto a sus vecinos. Basado en la fórmula es el promedio de los caminos más cortos del nodo v a otros nodos de la red. En la fórmula, n representa el número total de nodos de la red y $d(s,t)$ el camino más corto desde el nodo s a t.

$$C_{closeness}(s) = \frac{n-1}{\sum_{t \in \frac{N}{s}} d(s,t)} \tag{4}$$

- Eigenvector centrality: conexiones de un nodo a otros basado en la importancia. Dentro de la fórmula se identifica v un nodo, $M(v)$ set de vecinos del nodo v, $A = a_{v,t}$ matriz de adyacencia y λ una constante correspondiente al eigenvalor.

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} \quad x_t = \frac{1}{\lambda} a_{v,t} x_t \tag{5}$$

$$\lambda x \; = \; Ax \tag{6}$$

4.3. MÉTRICAS DE EVALUACIÓN

Algunas métricas para evaluar la predicción de los modelos de clasificación se basan en la utilización de la "Matriz de confusión". Esta permite evaluar el costo debido a una incorrecta clasificación de datos (aciertos y errores), pudiendo observar la performance del algoritmo. En esta matriz se compara aquellos objetos predichos con la realidad, en donde se pueden obtener aciertos en la clasificación tales como *"True Positive"* (TP) o *"True Negative"* (TN) y errores tales como *"False Negative"* (FN) o *"False Positive"* (FN).

		True	
		Positive	**Negative**
Predicted	Positive	TP	FP
	Negative	FN	TN

Ilustración 10: Matriz de confusión.

A partir de estos resultados, se consideran las siguientes métricas de evaluación: sensitivity (porcentaje de respuestas positivas identificadas correctamente), specificity (porcentaje de respuestas negativas identificadas correctamente), precisión (proporción de casos positivos predichos que fueron correctos) y accuracy (proporción de predicciones correctas del modelo).

$$Sensitivity = \frac{TP}{TP + FN} \tag{7}$$

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Accuracy = \frac{TP + TN}{TN + TP + FN + FP} \tag{10}$$

Otra manera de evaluar los resultados de las clasificaciones de los modelos corresponde a la curva ROC (Receiving Operating Characteristic). Esta curva permite comparar modelos de manera de identificar cual tiene mejor rendimiento como clasificador, utilizando el área bajo la curva (AUC) como aproximación a la calidad del modelo, la cual aumenta a medida que la curva se desplaza a la esquina superior izquierda del gráfico (en donde se minimiza el ratio de falsos positivos y se maximiza el ratio de verdaderos positivos). [8]

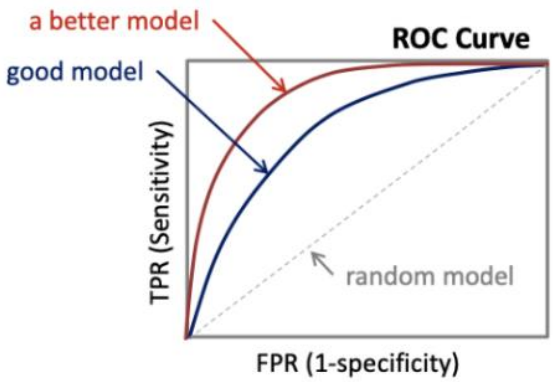


Ilustración 11: Curva ROC

Para la interpretación de modelos de Machine Learning es posible utilizar conceptos derivados de la teoría de juegos, tales como “Shapley Values”. Este método de interpretación determina cual es valor de contribución de una variable antes y después de agregarla al modelo de predicción, obteniendo así un ranking de las variables más importantes que explican el fenómeno. Es importante destacar que el orden en el cual se van “agregando” las variables al modelo afecta la predicción, por esto se mide el impacto agregando las variables en todos los órdenes posibles. De esta manera, el valor Shapley para cierta variable i contenida en un conjunto de variables N, dentro de una muestra de S < N variables seleccionadas, dada una predicción p se calcula de la siguiente manera. [11]

$$\phi_i(p) = \sum_{S \subseteq N/i} \frac{|S|!\,(n - |S| - 1)!}{n!} \big(p(S \cup i) - p(S)\big) \qquad (11)$$

A partir de esta definición matemática, es posible obtener gráficos de interpretación de variables los cuales entregan una interpretación similar a una regresión lineal. En la siguiente ilustración podemos observar un ejemplo. En el eje Y se observan las variables ordenadas de manera decreciente en cuanto a su importancia sobre el fenómeno estudiado, el eje X representa el impacto de una observación de la muestra en la predicción y la leyenda indica la magnitud de la variable. En este caso, al observar la variable “alcohol” podemos ver que a medida que aumenta en magnitud el impacto en el modelo o variable objetivo es mayor.

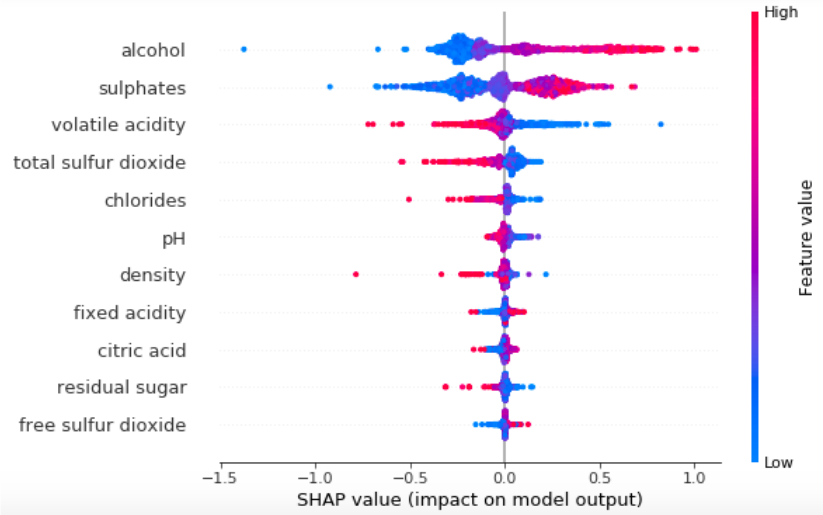


Ilustración 12: Ejemplo gráfico Shapley Values. [13]

También existen otros métodos de importancia de variables que se centran en la minimización del error del modelo al agregar una nueva variable además de analizar si fue utilizada para realizar una ramificación en un árbol de clasificación o decisión.

4.4. TÉCNICAS DE BALANCEO

La variable objetivo del problema a modelar (portabilidad de segunda línea o línea adicional) es minoría en comparación a las otras categorías (especialmente activaciones de líneas) representando en promedio un 20% mensualmente. Por eso, se observa un desbalanceo de datos que podría afectar la precisión del modelo para la clase minoritaria. Para solucionar este problema existen diferentes técnicas de balanceo de muestra. [12]

- Random Undersampling: balanceo de datos a partir de una reducción aleatoria de la clase mayoritaria con el fin de igualar el número de observaciones de la clase minoritaria. La desventaja de esta técnica es la pérdida de información importante para el modelo.

- Random Oversampling: balanceo de data al replicar datos de la clase minoritaria de manera aleatoria hasta igualar el número de observaciones de la clase mayoritaria. Esta técnica podría sesgar la predicción y afectar la precisión del modelo.

- SMOTE (Synthetic Minority Oversampling Technique): similar a la técnica de Oversampling ya que realiza un sobremuestreo de la clase minoritaria, sin embargo, toma un dato aleatorio dentro del espacio de muestra y genera nuevas

observaciones en base la información de características de los vecinos más cercanos en base a una métrica de distancia (distancia euclidiana), creando así una muestra sintética.

También existen modelos que no necesitan pasar por un proceso de balanceo, como es el caso de XGBoost, algoritmo que se hace cargo de este problema internamente. Se puede además una función de costos que penalice de diferente manera las clasificaciones erróneas de la clase mayoritaria y la clase minoritaria.

5. DESARROLLO METODOLÓGICO

La metodología CRISP-DM, abreviación de Cross-Industry Standard Process for Data Mining, modela el ciclo completo que sigue un proyecto dentro de la disciplina de ciencias de datos. Esta metodología consta de los siguientes pasos

1. Comprensión del negocio: determinar el contexto de la empresa, los objetivos del proyecto y la planificación de este.

2. Comprensión de los datos: etapa de exploración de los datos disponibles junto con un análisis preliminar para el entendimiento del problema. Se debe evaluar la calidad de los datos.

3. Procesamiento de datos: limpieza de los datos con el fin de armar el conjunto final de variables que permita el desarrollo de modelos predictivos. Dentro de esta etapa se pueden construir variables relevantes en base a la data existente.

4. Modelado: construir los modelos predictivos aplicando diversas técnicas con relación al problema, junto al diseño de evaluación que permitirá seleccionar aquel con mayor poder interpretativo y predictivo.

5. Evaluación: a partir de la construcción del modelo, se comparan los resultados y se definen los siguientes pasos a seguir.

6. Despliegue: se genera un plan de despliegue para la utilización del modelo, junto a un plan de monitoreo y mantención.

Se selecciona esta metodología de trabajo debido a que la secuencia entre cada una de las etapas no es estricta, otorgando amplitud y flexibilidad para volver a etapas anteriores si es necesario. Esto permite volver a evaluar variables importantes para el modelamiento que habían sido excluidas en un principio o incluso volver a la etapa de preparación de nuevos datos para el análisis.

A continuación, se detalla el desarrollo de las etapas planteadas en esta metodología.

5.1. COMPRENSIÓN DEL NEGOCIO

Dentro de esta etapa se espera entender el contexto y funcionamiento del negocio de las telecomunicaciones con el fin de entender los objetivos del desarrollo del proyecto y el impacto que este pueda generar dentro de la empresa. Por esto se realiza un estudio

respecto a la situación actual de la empresa, específicamente el servicio de telefonía móvil postpago.

El mercado de telefonía móvil en la empresa corresponde a uno de los servicios más importantes, grandes y por ende es el que genera mayores retornos a la compañía. Existen algunos modelos de predicción de los cuales nacen campañas que buscan atraer a clientes a través de ofertas de planes postpago, los cuales serán mencionados a continuación.

Primero encontramos el modelo de portabilidad numérica, enfocado en la identificación de clientes propensos a portar líneas móviles que previamente se fugaron de la empresa de manera voluntaria. A partir de este modelo se realiza una campaña comercial que apunta a la obtención de clientes titulares de líneas mas que a la portabilidad de segundas líneas asociadas a un cliente. Se perfilan aproximadamente 6,3 millones de clientes para las campañas de portabilidad numérica, de los cuales aquellos que se encuentran dentro de los primeros 4 deciles poseen una tasa de conversión cercana al 1%.

Por otro lado, existe un modelo predictivo que perfila aquellos clientes que poseen una mayor propensión a obtener una línea adicional, independiente del origen (migración, portabilidad o activación de línea). A partir de este modelo se realiza una campaña comercial que gestiona alrededor de 3 millones de clientes, de los cuales 400 mil son contactados a través del canal de Call Center. La efectividad de esta campaña dentro de los primeros 4 deciles es cercana al 1,5% (independiente si la llamada es contestada o no).

Pese a que existen campañas con foco en la adquisición de clientes portados desde otras compañías y adquisición de segundas líneas, estas están enfocadas a un grupo amplio de clientes y según requerimientos provenientes desde el área de CVM o Customer Value Management, es necesario identificar perfiles de clientes propensos a portar segundas líneas ya que no solo podría mejorar la efectividad en canales de contacto directo como lo es el Call Center, sino que además permitiría a la empresa aumentar su participación de mercado quitando líneas activas a la competencia.

Se debe mencionar que el modelo desarrollado en este trabajo es complementario a el modelo de líneas adicionales actual, por lo que el perfilamiento de los clientes propensos a una portabilidad será realizado sobre aquellos que poseen una propensión de adquirir una línea adicional.

5.2. COMPRENSIÓN DE LOS DATOS

Existen distintas fuentes de información dentro de la empresa, las cuales poseen variables que podrían explicar el fenómeno de portabilidad de líneas adicionales. Cabe destacar que toda la información que se posee dentro de estas bases es en referencia a los clientes titulares de una línea postpago, no es posible caracterizar al cliente cuya línea es propensa a ser portada como línea adicional. A continuación, se detalla cada fuente y las variables a utilizar de manera exploratoria.

5.2.1 Postpago consolidado

Esta tabla posee el registro histórico de clientes de telefonía postpago móvil al cierre de cada mes. Es actualizada de manera mensual considerando los movimientos provenientes de la actividad comercial considerando las ventas de líneas postpago y las líneas fugadas. Se mantienen aproximadamente 4,3 millones de líneas (tanto titulares como adicionales) y debido a limitaciones de almacenamiento de la empresa, solo se posee 1 año de historia, limitando el estudio de tendencias temporales que pueda existir. Para la construcción del modelo se considera solo aquellos móviles asociados a la línea titular.

Para el modelamiento del problema solo se usará una selección de variables de esta fuente de información, debido a que existen columnas que no han sido actualizadas durante el año 2021, lo cual no aportaría información relevante al caso estudiado. Las variables utilizadas son agrupadas a continuación.

- Caracterización de línea titular: origen o forma de ingreso de la línea utilizada por el cliente titular (portabilidad, migración o activación), cargo fijo por plan contratado, marca y antigüedad de equipo adquirido (en caso de haber comprado uno), cantidad de datos móviles consumidos durante el mes.

- Caracterización del cliente titular: cantidad de líneas asociadas a su cuenta (tanto titular como adicionales), antigüedad o permanencia en la compañía (desde la llegada con su primer línea), servicios hogar contratados (fibra óptica, televisión satelital, etc), número de productos contratados de valor agregado o VAS.

- Interacciones con la empresa: cantidad de reclamos realizados, canales de la empresa con los cuales interactuó durante el mes (reclamo, consulta, tramite), suma de interacciones realizadas durante el mes.

- Interacción con la competencia: duración y cantidad de llamadas recibidas por la línea titular provenientes de Call Centers de compañías externas.

5.2.2 Actividad comercial

La tabla de actividad comercial registra de manera mensual todas las nuevas líneas postpago que son adquiridas por la empresa, identificando aquellas que fueron portabilidades, activaciones o migraciones. En general se adquieren cerca de 134 mil líneas nuevas todos los meses, en donde las líneas portadas y activadas concentran la mayoría de las ventas totales.

Al filtrar esta data para aquellos clientes titulares (quienes se encuentran dentro de la base de clientes postpago con una antigüedad mínima de 2 meses) es posible obtener las líneas adicionales que fueron durante un mes. A diferencia de la venta total de líneas, se observa un bajo porcentaje de líneas adicionales portadas (cercano al 20%) en comparación a activaciones de líneas, por lo cual es probable tener cierto desbalanceo en la data que se utilizará para el modelamiento. No es posible observar temporalidad dentro de esta variación debido al corto periodo analizado producto de la disponibilidad de datos almacenados.

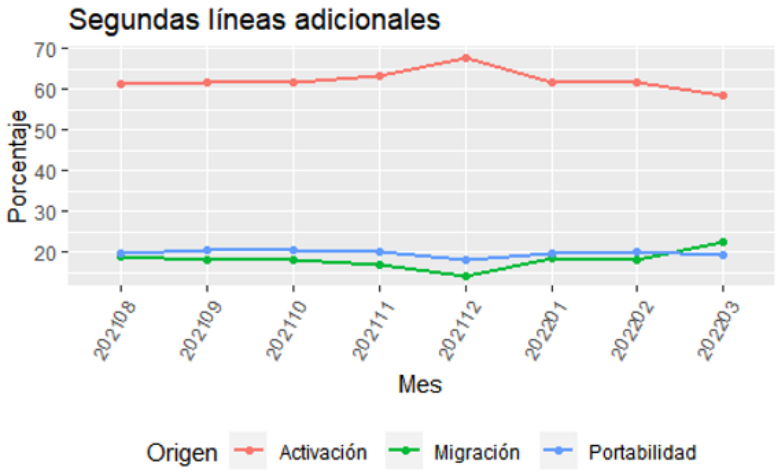


Ilustración 13: variación líneas adicionales móvil dentro de la actividad comercial desde abril del 2021 hasta enero del 2022. Elaboración propia

5.2.3 Malla parental

Esta tabla corresponde a una aproximación del grupo familiar a la cual correspondería un individuo, tanto clientes como no clientes. La información es construida a partir de nacimientos, matrimonios y defunciones obtenidos desde registro civil,

dirección de residencia obtenidos por SERVEL y/o Bienes Raíces y registros de extranjeros residentes en el país. Es posible asignar un grupo familiar a un individuo a partir de criterios definidos por la empresa, los cuales se detallan a continuación.

- Un individuo menor a 26 años conforma un grupo familiar con sus padres en el caso de que no esté casado ni tenga hijos. Si los padres se encuentran divorciados, este es asignado a un grupo familiar junto con la madre. En el caso del fallecimiento de uno de los padres, es asignado a aquel que aún siga vivo/identificado.

- Dos individuos que contraen matrimonio conforman un grupo familiar, junto con todos los hijos dentro de su matrimonio que cumplan el criterio anterior e hijos que provengan de parejas anteriores.

- Se crea un vínculo de pareja a partir del nacimiento de un hijo, independiente si contraen matrimonio, pasando a formar un grupo familiar.

- Una persona mayor a 26 años, soltero y sin hijo forma un grupo familiar de manera individual.

- Debido a que no es posible identificar la fecha de nacimiento de un extranjero residente, la asignación de un grupo familiar se realiza en la medida de que exista un vínculo de matrimonio o nacimiento de un hijo.

Con la definición de estos criterios se identifican 8.720.542 hogares o grupos familiares con al menos un integrante vivo. Un 27% de estos grupos familiares posee un titular postpago o fibra óptica[2]. Existe una brecha entre esta identificación de hogares con respecto a la información entregada por el censo 2017 (5.651.637 hogares identificados), sin embargo, se considera una buena aproximación debido a la falta de precisión que existe en los criterios definidos por la empresa.

Como resultado de este proceso de asignación de hogares, es posible identificar el Rut de un individuo, nombre, fecha de nacimiento (exceptuando a extranjeros), género, Rut del cónyuge (en el caso de estar casado), Ruts de sus padres y un identificador único del grupo familiar al cual corresponde.

Pudiendo asignar grupo familiar a cada individuo, es posible caracterizarlo no sólo demográficamente (edad, género, estado civil, si es padre), sino que también se pueden generar variables que hacen alusión a su composición familiar (el cual se puede

[2] Análisis interno realizado en noviembre del 2021

considerar como su círculo social más cercano). Las variables derivadas de esto corresponden al total de miembros en el hogar, número de miembros por tramos etarios, cantidad de clientes titulares de líneas postpago en el hogar, penetración global (cantidad de líneas sobre miembros del hogar) y penetración por miembros objetivo (cantidad de líneas sobre miembros del hogar mayores a 9 años).

5.2.4 Resumen de llamadas

Dentro de esta tabla podemos encontrar un resumen del tráfico de llamadas emitidas o recibidas por un cliente de la compañía. Este resumen se hace de manera mensual, considerando la data de registro de llamadas. De esta manera se hace más manejable la información, pero se pierde detalle en cuanto a la identificación de números con los cuales posee mayor interacción un cliente titular.

Las variables que serán utilizadas para el fenómeno de portabilidad de líneas adicionales se separan en duración de las llamadas emitidas o recibidas por un cliente titular postpago. Es posible identificar aquella compañía con la cual un cliente posee más del 50% de las interacciones durante el mes, lo cual es utilizado como variables adicionales para el modelo.

5.2.5 Call Detail Record (CDR)

A diferencia de la base anterior, el registro de llamadas o CDR (según sus siglas en inglés) posee mayor detalle con respecto a las interacciones de un cliente titular con terceros. Esta tabla contiene todo el tráfico de llamadas entre clientes de la empresa y no clientes (tanto llamadas recibidas como emitidas). El registro se encuentra diferenciado en dos tablas en base a la tecnología por la cual fue realizada la llamada o interacción, teniendo un CDR “Voz” (tecnología 2g y 3g) y “Volte” (tecnología 4g y 5g). Para el modelamiento del problema no se considera esta diferencia de tecnología.

Se posee la información de CDR desde diciembre del 2021 hasta marzo del 2022, por lo cual la evaluación de estas variables dentro del modelo es realizada en base a este periodo de tiempo. En la sección de procesamiento se profundizará en la limpieza de esta fuente de información.

5.2.6 Geolocalización y variables de competencia

Es posible identificar la comuna y región de residencia más probable de un cliente, a partir de la información entregada por las antenas telefónicas (proceso denominado “triangulación”). Por lo general esta estimación se realiza a partir de la información reportada por las antenas durante la noche, debido a que en este horario es más probable que el cliente se encuentre en su hogar de residencia.

Se observa una concentración de la data en regiones céntricas (específicamente RM y Valparaíso), teniendo aproximadamente un 10,24% de clientes sin información en región de residencia y aproximadamente un 5% de clientes cuya triangulación no fue identificada (S/T). En la ilustración 16 también se observa la tasa de portabilidad de líneas adicionales, representada por la línea azul. Esta tasa se mantiene entre el 15% y 20% para todas las regiones, excepto la región de Aysén.

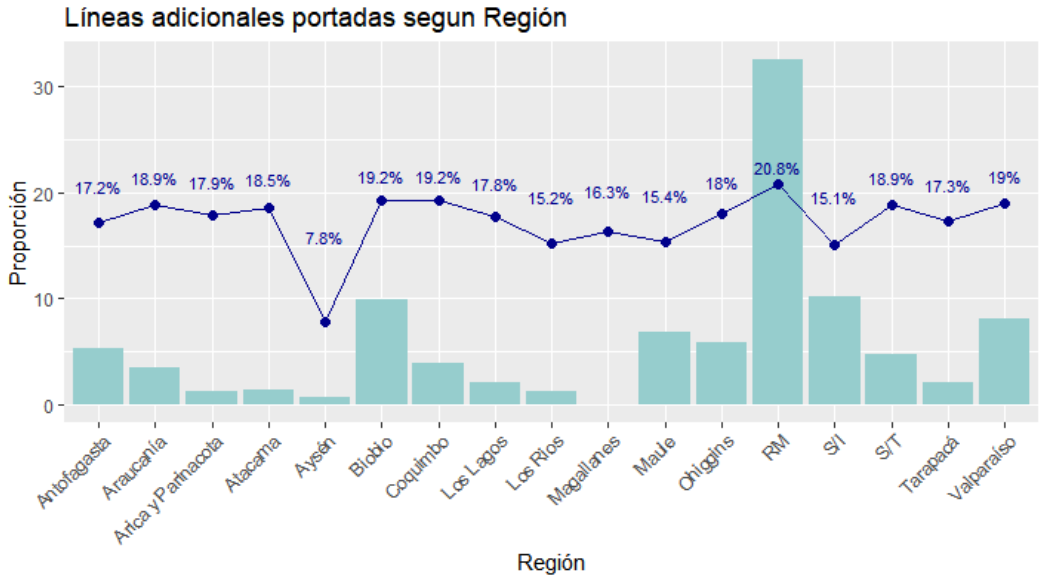


Ilustración 14: Tasa de portabilidad de líneas adicionales y distribución de datos según región.

En cuanto a la comuna, también se observa un 10,24% de datos sin información y 4,74% cuya ubicación no fue identificada. Las 5 comunas que registraron un mayor número de adquisición de líneas adicionales son Antofagasta, Puente Alto, Maipú, Santiago y Calama, con una tasa de portabilidad de 20% aproximadamente.

Teniendo la ubicación del cliente, se incorporan las siguientes variables de competencia con respecto a la calidad del servicio de una empresa (obtenidas a partir del uso de Facebook).

- TIV (tiempo de ida y vuelta): tiempo que toma enviar una solicitud desde el dispositivo al servidor y de vuelta al dispositivo. A menor tiempo de respuesta mayor calidad del servicio.

- DS (velocidad de descarga): tiempo que tarda la información en llegar al dispositivo como respuesta. A mayor velocidad de descarga, mejor calidad del servicio.

- NGP: porcentaje de tráfico de datos sobre una generación de red seleccionada (3g, 4g, 5g).

- SS (intensidad de la señal): medida de calidad de una señal en particular desde una distancia con respecto a su fuente. A mayor intensidad de señal, mejor calidad.

- DL (latencia de bajada): tiempo de respuesta desde que se realiza físicamente una acción hasta que el dispositivo la lleva a cabo. A menor latencia de bajada, mejor calidad.

- RSRQ: calidad de la señal de referencia de rendimiento recibida.

Estas variables se especifican tanto para una compañía en particular. Se considera la variación porcentual del estado de la empresa versus la competencia para el análisis y modelamiento del problema.

También se posee información de la calidad del servicio entregada por “Tutela”, empresa dedicada a recolectar información de uso de datos en más de 3000 aplicaciones utilizadas por sus usuarios a nivel global. Estas variables corresponden a la velocidad de descarga, velocidad de subida y latencia promedio para las 4 compañías más importantes de la industria de telecomunicaciones en Chile. Pese a que las métricas son similares a las recolectadas por Facebook, se consideran ambas fuentes de información para obtener dos puntos de comparación dentro de una misma localización.

5.2.7 Tráfico por aplicaciones

Detalle del tráfico de datos del cliente titular durante un mes a través de aplicaciones (bancos, nube, correo, delivery, free stream, información, juegos, laboral, mapas, música, diarios, ofertas, retail, redes sociales, deportes, series, telco, transporte, transporte corto, video y otros). También se registra el tráfico total de MB durante el mes del cual se deriva el porcentaje de uso por tipo de aplicación.

Se decide incorporar este set de variables con el fin de caracterizar el comportamiento del titular con respecto al uso de su línea titular, observando si existe alguna relación entre el perfil del cliente y aquel tipo de aplicación más utilizada con datos móviles.

5.3. PREPARACIÓN DE LOS DATOS

Para evitar sobreajuste en los modelos, dentro del procesamiento de los datos se utiliza la función “Woebin” de la librería “Scorecard”, para tramificar variables numéricas y categóricas. Esta función genera intervalos óptimos para las variables independientes utilizando métodos de segmentación de árboles. La tramificación se realiza sobre variables provenientes de las fuentes mencionadas anteriormente como variables creadas.

A partir de la cantidad de minutos emitidos o recibidos por un cliente titular se crean variables que indican el porcentaje de llamadas salientes y entrantes hacia clientes de la empresa y a la competencia. Estas variables son separadas en intervalos según la tramificación óptima obtenida. Además, se crean las variables que indica cual es la compañía con la cual se contacta más un cliente (compañía favorita), considerando aquella que posea un mayor porcentaje de minutos emitidos o recibidos.

Se observa que a medida que el porcentaje emitido a clientes de la empresa aumenta, la tasa de portabilidad de líneas adicionales disminuye. Esto puede describir el círculo cercano del cliente titular, el cual estaría compuesto en su mayoría por clientes de la compañía lo que coincide con una mayor proporción de llamadas emitidas dentro de la empresa. Es posible observar esta misma tendencia para llamadas recibidas por clientes de la compañía telefónica.

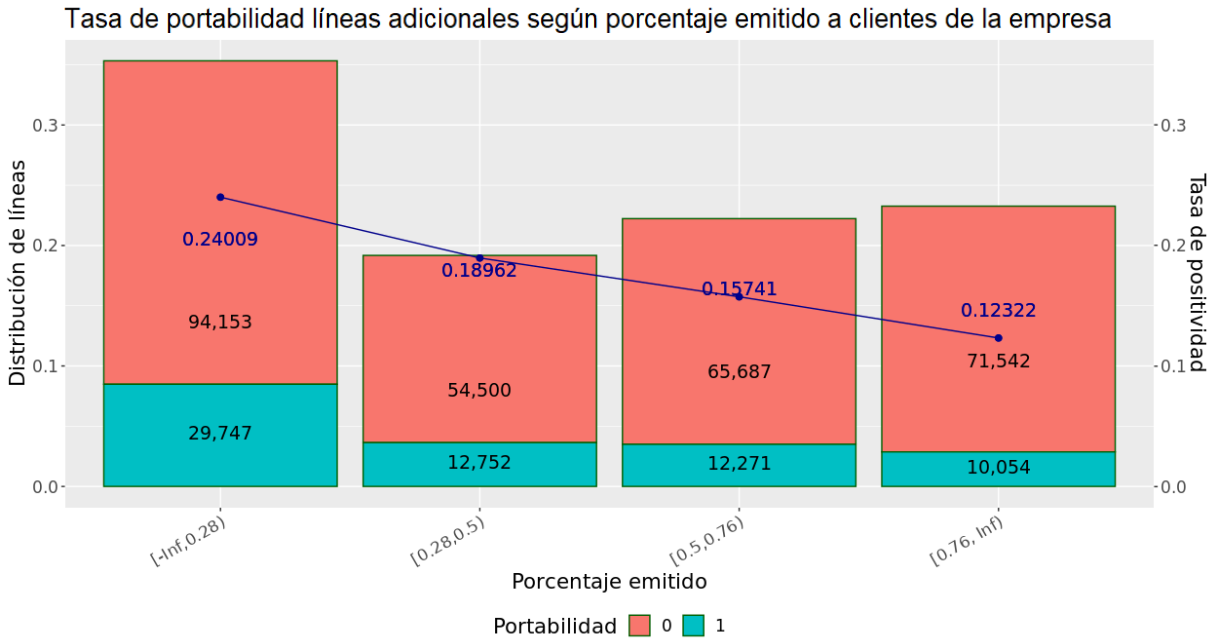


Ilustración 15: Distribución de líneas adicionales portadas con respecto al porcentaje de llamadas emitidas desde clientes titulares hacia otros clientes de la misma empresa. Elaboración propia.

Analizando el porcentaje de llamadas emitidas a clientes fuera de la compañía, se observa lo contrario al gráfico anterior. En medida que el porcentaje aumenta (en el siguiente grafico se observa el tráfico hacia una empresa de la competencia (denominada C1), pero en la tendencia es similar para las otras compañías), la tasa de líneas adicionales portadas también lo hace. Esto indicaría la existencia de números pertenecientes a otras compañías que son propensas a ser portadas bajo una cuenta adicional.

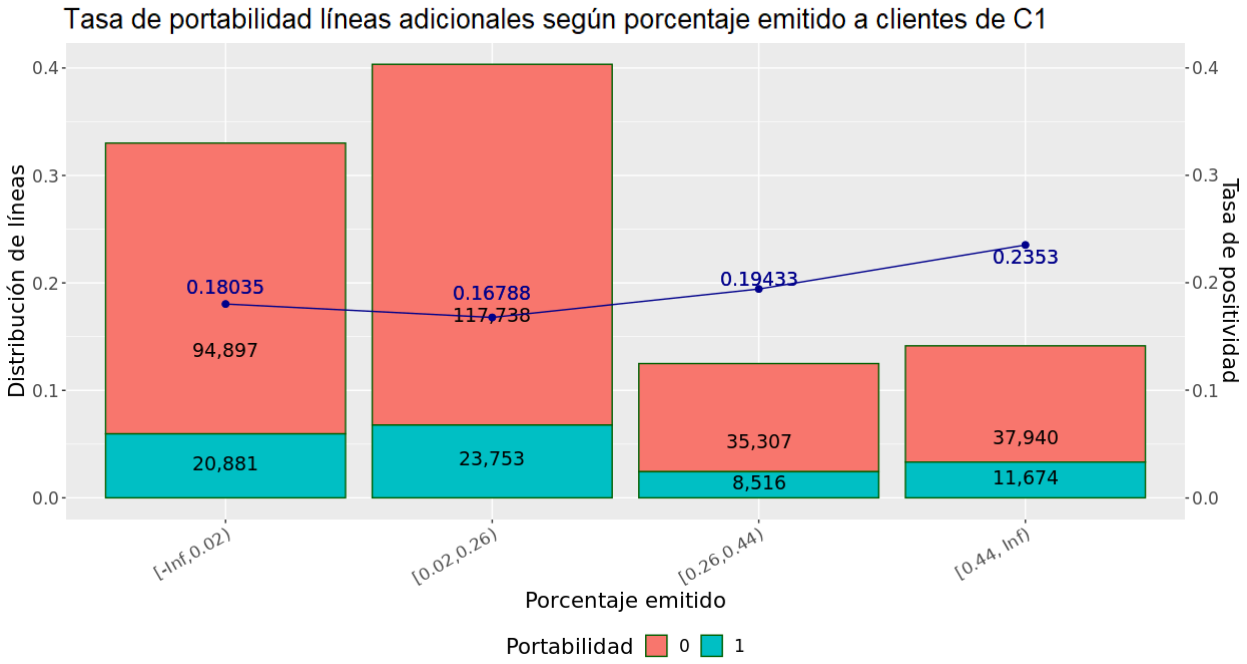


Ilustración 16: Distribución de líneas adicionales portadas con respecto al porcentaje de llamadas emitidas desde clientes titulares hacia clientes C1. Elaboración propia.

Con el fin de capturar tendencias históricas con respecto a la actividad comercial de líneas adicionales se crean variables que reflejan la tasa de portabilidad por sobre características del cliente (tipo de plan que posee, geolocalización, antigüedad de su móvil, marca del equipo y cargo fijo pagado de manera mensual).

Se procesan variables demográficas del cliente titular. Al tramificar y graficar la distribución de líneas adicionales según rangos etarios de clientes, se observa un aumento de la tasa de portabilidad a medida que aumenta la edad, específicamente concentrados en un rango medio-alto. La distribución de la data es esperable, debido aquella población menor a 25 años no tiene la posibilidad de contratar servicios de telefonía móvil y/o es adicional a la cuenta de otra persona. Otro factor que podría afectar en este aumento de portabilidades a medida que aumenta la edad corresponde al crecimiento del círculo familiar (hijos, sobrinos, nietos, etc) aumentando las posibilidades de identificar una línea propensa a portar a la compañía.

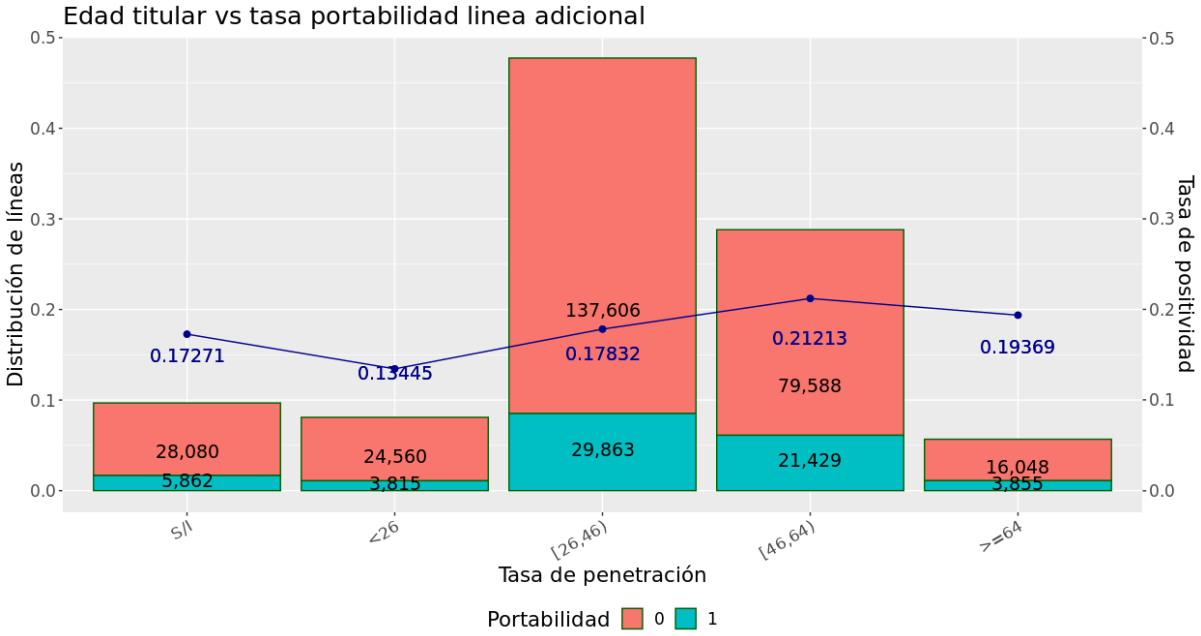


Ilustración 17: Distribución de líneas adicionales según el rango etario al cual pertenece el cliente titular. Elaboración propia.

Con respecto a la cantidad de líneas que posee un cliente titular, a medida que aumentan se observa una disminución de la tasa de portabilidad de líneas adicionales. Es esperable que a medida que el cliente posee más líneas bajo su cuenta no adquiera nuevas líneas pudiendo estar relacionado con el aumento de cargo fijo facturado mensualmente o al aumento en la penetración en el hogar, en donde no existirían líneas

propensas a ser portadas como líneas adicionales. La tendencia es la misma tanto para clientes dentro de un hogar cerrado o un hogar abierto, sin embargo, las tasas de portabilidad son mayores dentro del segmento abierto.

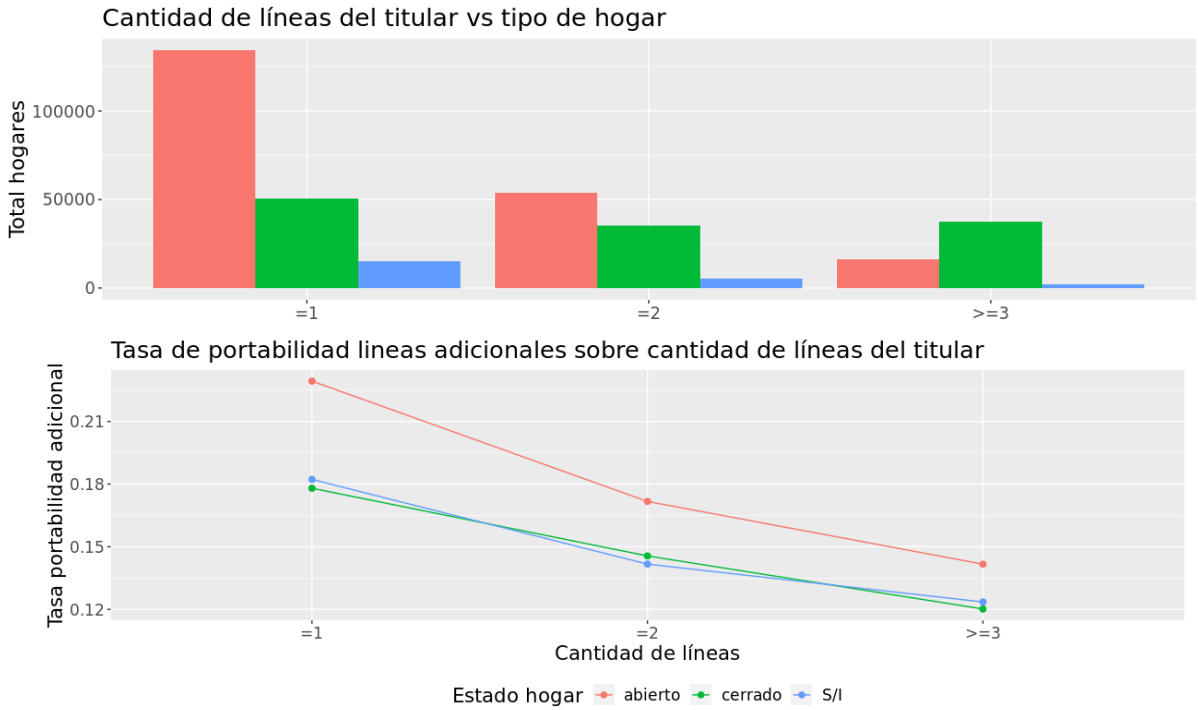


Ilustración 18: Distribución de líneas adicionales según la cantidad de líneas asociadas al cliente titular diferenciada por tipo de hogar. Elaboración propia

Observando el origen de los clientes con respecto a su línea titular, aquellos que llegaron a la compañía mediante portabilidad numérica tienen una tendencia a portar nuevas líneas adicionales en comparación a las otras categorías. Esto puede ocurrir debido a que el cliente ya está familiarizado con el proceso de portabilidad por lo que las líneas adicionales podrían seguir el mismo patrón. Se puede observar un gran número de clientes que no registran origen de la línea titular, superando un 50% del total de la data. Esta pérdida de información es producto a un cambio en el sistema de registro dentro de la compañía hace aproximadamente 2 años. Pese a la falta de información, esta variable se considera importante para el fenómeno estudiado, por lo que es incluida en el modelo, teniendo cuidado en la interpretación posterior a los resultados.

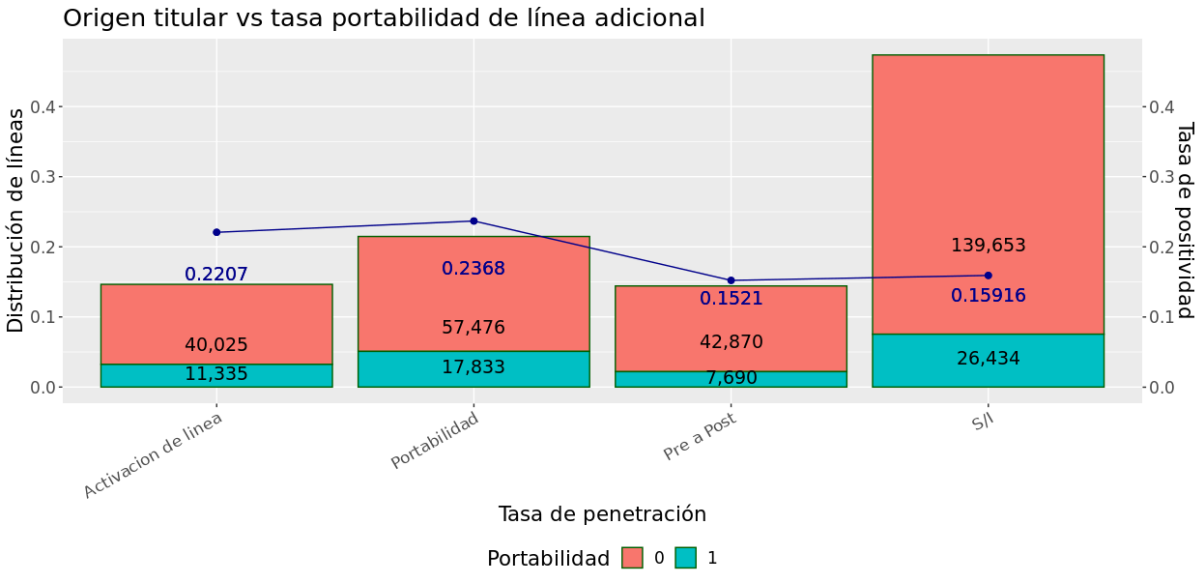


Ilustración 19: Distribución de líneas adicionales según el origen del cliente titular. Elaboración propia

Con respecto a la antigüedad del cliente en la compañía, la tasa de portabilidad de líneas adicionales decrece a medida que la permanencia aumenta. Esto podría tener múltiples interpretaciones: este cliente podría estar desinformado con respecto a las ofertas que entrega la compañía para líneas adicionales, podría no ser sensible a ofertas o ya contaría con líneas adicionales asociadas. La relación entre la antigüedad del cliente y el aumento de líneas se evidencia en la comparación de tasas de churn expuestas en el contexto inicial, en donde aquellos hogares cerrados (que poseen una cantidad de líneas igual o mayor a los miembros del hogar) se fugan en menor magnitud, teniendo una permanencia mayor en la empresa.

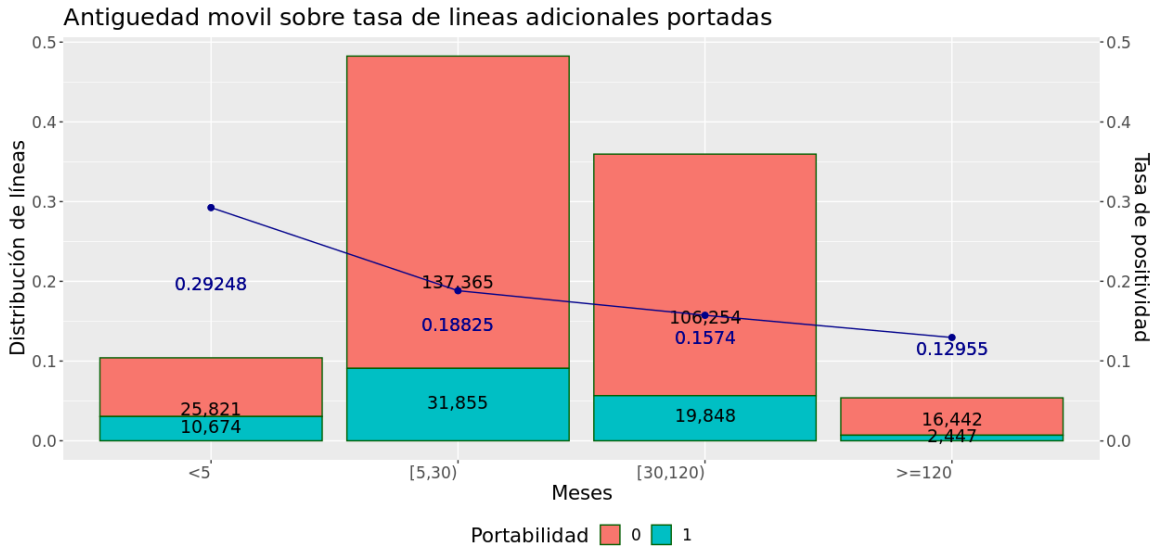


Ilustración 20: Distribución de líneas adicionales según la antigüedad del cliente titular. Elaboración propia

Al graficar el cargo fijo y su distribución de líneas adicionales no se observa una tendencia marcada en algún tramo específico. Sin embargo, hay una leve baja en la cantidad de líneas portadas para aquellos clientes que pagan más por su línea titular. Es probable que un comportamiento individual no representa una tendencia marcada, pero sí pueda haber cambios al hacer un análisis a nivel hogar.

Con esto en mente, se construyen variables que permiten caracterizar el hogar al cual pertenece. Con esto se crean las variables de penetración global que posee la empresa dentro del grupo familiar (creado a partir de la suma de líneas móviles postpago que posee la familia sobre el total de integrantes) y la penetración sobre los integrantes que poseen más de 9 años, debido a la baja probabilidad que posee un niño menor a esta edad a acceder a un teléfono con un contrato mensual de telefonía. También se caracteriza en base al promedio de edad familiar, integrantes por rango etario, suma de cargo fijo pagado de manera mensual, cargo fijo per cápita y variaciones de la facturación mensual dentro de 3 meses.

Analizando la suma del cargo fijo a nivel hogar se puede apreciar una tendencia a disminuir las portabilidades de líneas adicionales portadas a medida que el monto facturado aumenta.

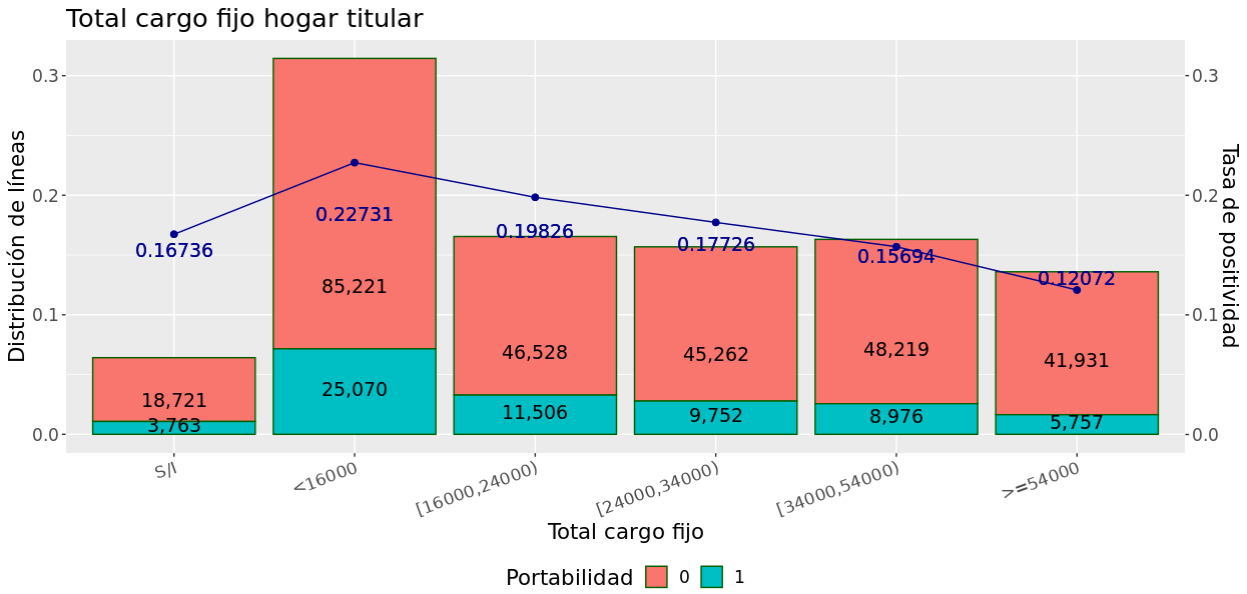


Ilustración 21: Distribución de líneas adicionales según suma de cargo fijo a nivel hogar. Elaboración propia

En el caso de las variables de competencia, tal como se mencionó antes, se poseen dos fuentes diferentes: Facebook y Tutela. En el primer caso solo se realiza una tramificación de las variaciones porcentuales. En el segundo caso, se genera la métrica de variación porcentual a partir de las variables desagregadas por compañía para luego ser tramificadas.

Analizando el tiempo de ida y vuelta o TIV, se puede observar que cuando el valor de la variación porcentual es positiva, el TIV de la competencia es mayor, por lo que la calidad de servicio es más baja comparado a lo que la empresa ofrece. A partir de un análisis de esta variable se observa lo siguiente.

- La variación porcentual promedio de la empresa respecto a la competencia C1 es de 22%, lo que indica que en general el tiempo de respuesta de la empresa con respecto a la competencia es menor. La tasa de portabilidad (o positividad) de líneas adicionales portadas se mantiene a medida que esta variación aumenta.

- La variación porcentual promedio entre la empresa y la competencia C2 es de -35%, indicando que el tiempo de respuesta de la competencia es menor. La tasa de portabilidad de líneas adicionales posee una leve tendencia a aumentar a medida que esta variación porcentual aumenta.

- La variación porcentual promedio entre la empresa y la competencia C3 es de -4%, indicando nuevamente que el tiempo de respuesta de la competencia es mejor. En este caso, la tasa de portabilidad de líneas adicionales disminuye a medida que el valor es positivo.

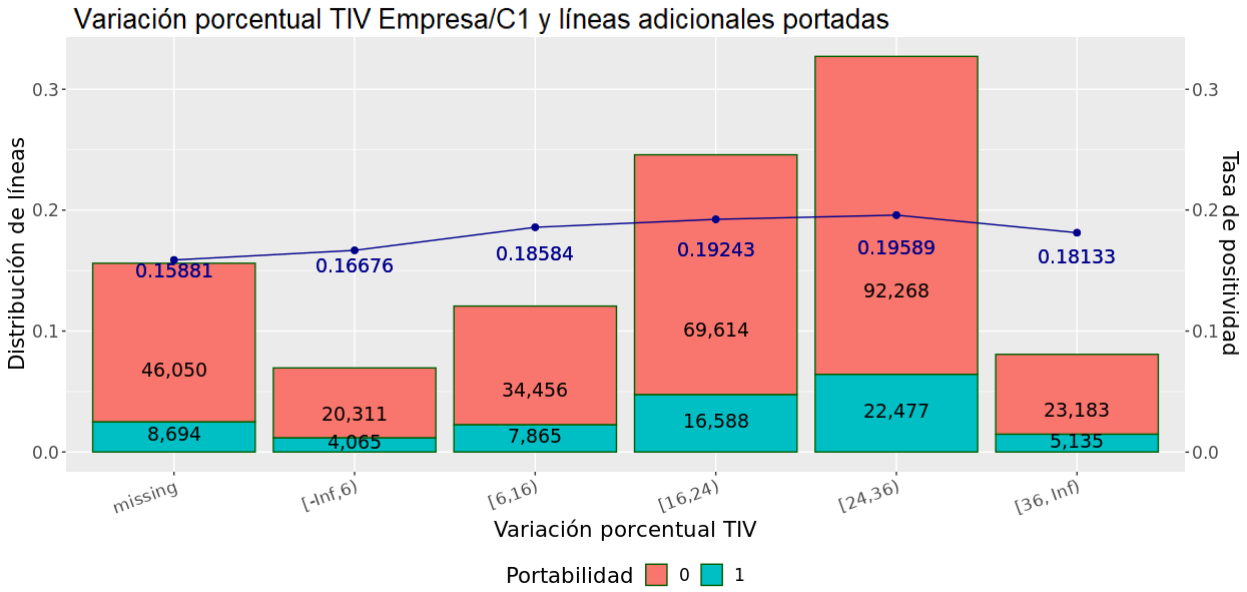


Ilustración 22: Tasa de portabilidad de líneas adicionales y distribución de líneas según variación porcentual TIV Empresa/C1.

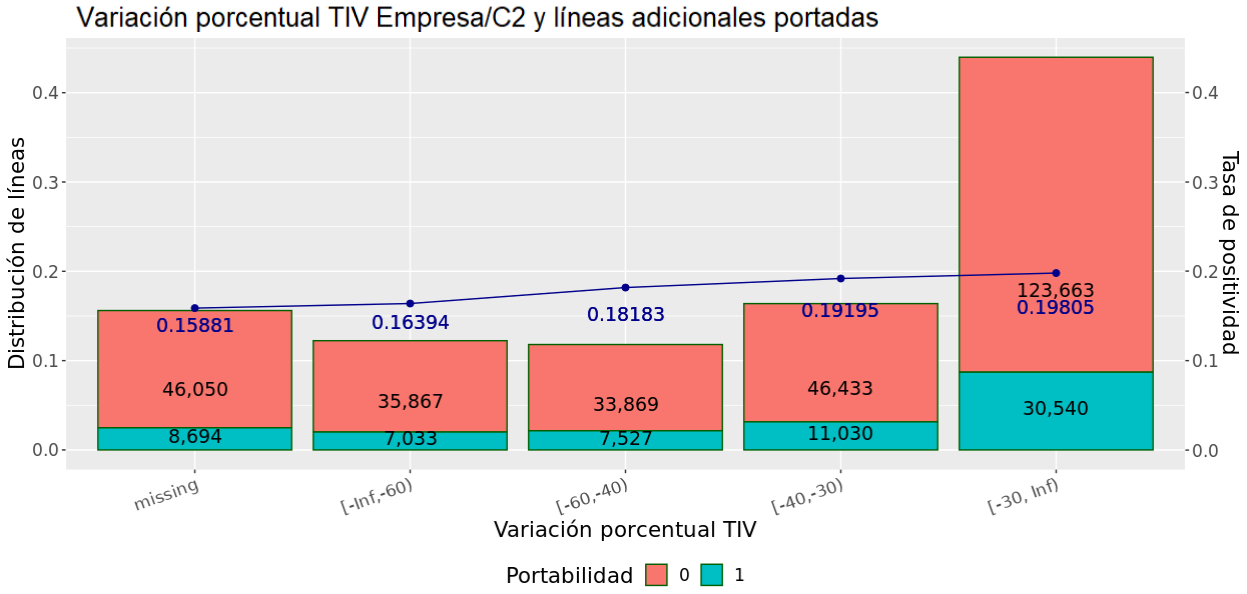


Ilustración 23: Tasa de portabilidad de líneas adicionales y distribución de líneas según variación porcentual TIV Empresa/C2.

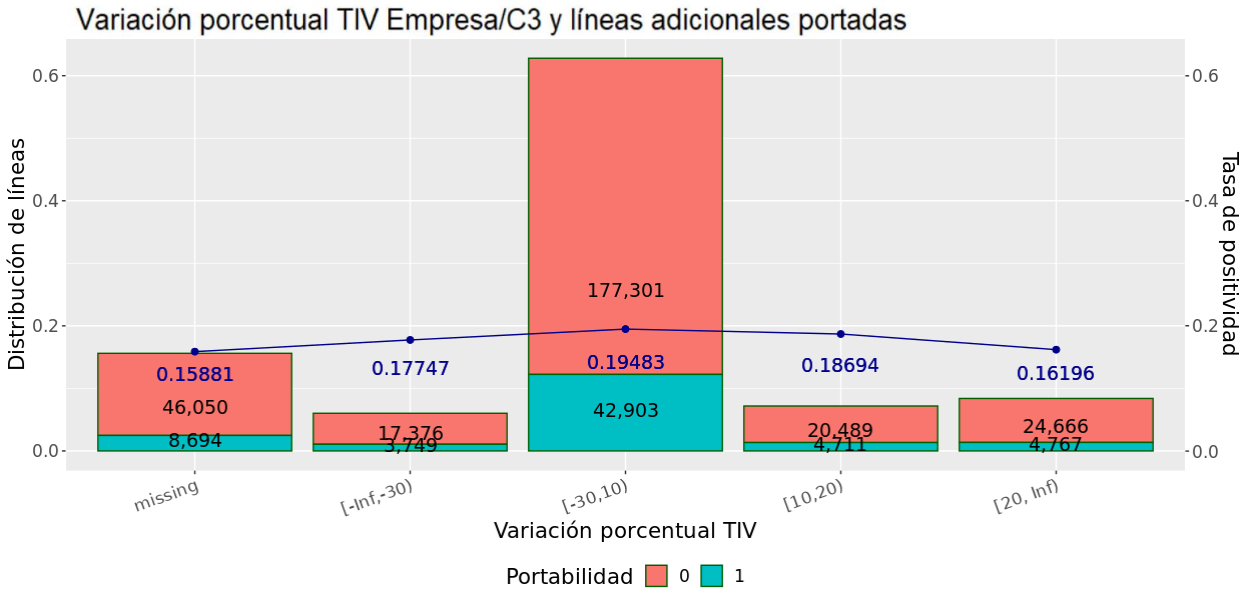


Ilustración 24: Tasa de portabilidad de líneas adicionales y distribución de líneas según variación porcentual TIV Empresa/C3.

Estas tendencias tienden a mantenerse en cuanto a las otras métricas analizadas, en donde la empresa estudiada supera en la mayoría a dos de sus competidores en cuanto a la calidad del servicio por sector.

Por último, se procesa la información del registro de llamadas o CDR. Se realiza una limpieza de los nodos o móviles que no son relevantes para el estudio y que podrían alterar el cálculo de importancia dentro de la red. Por eso, se eliminan aquellas llamadas que pudieron ser efecto de una equivocación, identificándolas como una única llamada durante el mes entre dos pares de móviles cuya duración es igual o inferior a 5 segundos. Por otro lado, se desea eliminar aquellos números telefónicos que puedan pertenecer a centros de emergencias, carabineros, bomberos, Call centers, etc. Por esto se definen reglas en cuanto a aquellos nodos o números que poseen una cantidad de llamadas salientes sobre 2000 pero que reciben menos de 100 llamadas (identificando así Call centers y similares) y aquellos que, por el contrario, la cantidad de llamadas entrantes superan las 2000 y emiten menos de 100 llamadas.

Luego de haber realizado la limpieza de nodos se caracteriza cada uno respecto al tipo de mercado al cual pertenecen: prepago o postpago (clientes de la empresa) o no clientes. Con estos atributos asignados a cada vértice o nodo se calcula, para cada movil dentro de la red el Degree o la cantidad de nodos totales con los que está conectado, diferenciando por mercado. En el grafo a continuación se ejemplifica un caso en donde se caracteriza cada nodo según el mercado al cual pertenece cada móvil, teniendo clientes postpago (pos), prepago (pre) y no clientes (nc). El cálculo de métricas para el

nodo 1 se muestra en la siguiente tabla. De la misma forma se calcula la proporción de minutos interactuados con clientes de distintos mercados.

Nodo	Degree	Degree pos	Degree pre	Degree nc	Proporción Degree pos	Proporción Degree pre	Proporción Degree nc
1	4	1	0	3	0.25	0	0.75

Tabla 1: Cálculo de métricas a partir de registro de llamadas para caracterización de nodos según mercado.

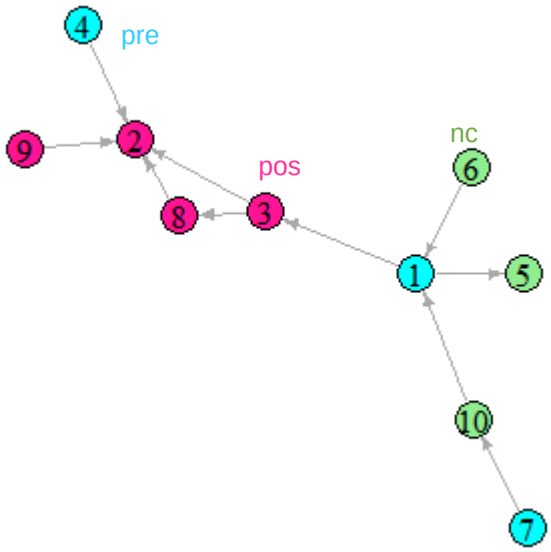
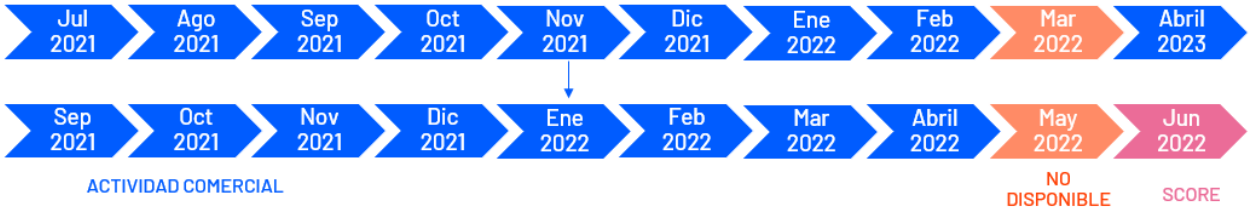


Ilustración 25: Ejemplificación de grafo. Elaboración propia.

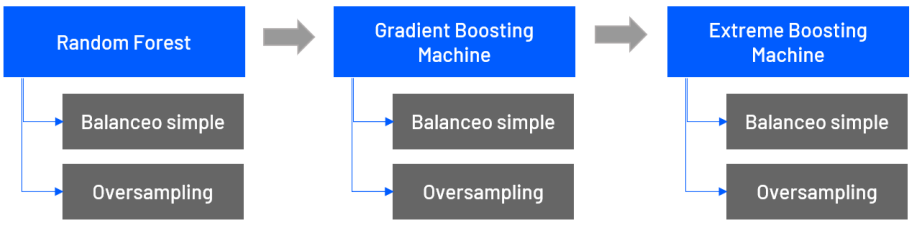
5.4. MODELACIÓN

Para la creación de los modelos predictivos a evaluar se utilizaron 8 meses de historia, separando esta muestra en una base de entrenamiento y una base de testeo para la validación. La variable objetivo o variable dependiente a predecir se construye a partir de la actividad comercial, identificando si un cliente porta una línea adicional durante el mes o no.

Debido a la disponibilidad de los datos existe un desfase de 2 meses, es decir, el cálculo de propensión de junio 2022 se realiza a finales de mayo 2022, en donde el cierre de postpago aún no está disponible, por lo que se utilizan los datos desde marzo 2022 hacia atrás. La construcción de variables independientes considera este desfase, por ejemplo, los clientes que obtuvieron líneas adicionales en enero del 2021 son caracterizados a partir de su comportamiento en la empresa en noviembre del 2021.



Se tiene un total de 350.706 observaciones en la muestra, de la cual un 80% es utilizada para entrenamiento y un 20% para testeo del modelo, procurando dividir la muestra de tal manera que no haya un cliente que se encuentre en ambas muestras a la vez. Los modelos fueron entrenados y testeados 98 variables. Estos se listan a continuación.



- Se entrenan modelos con Random Forest el cual corresponde a un algoritmo que entrega predicciones certeras. Se evalúa el modelo sin balanceo de data y aplicando Oversampling. No se utiliza Undersampling debido a que la muestra desbalanceada ya es reducida, por lo que aplicar esta técnica podría implicar pérdida importante de información. Lo mismo aplica para los siguientes modelos entrenados.

- Se evalúa modelos con Gradient Boosting Machine sin balanceo de data y aplicando Oversampling, con el objetivo de obtener una mejora en la predicción a través de la técnica de Boosting para entrenar un modelo que aprenda de predictores débiles.

- Similar al caso anterior, se aplica el algoritmo de XGBoost sin balanceo de data y aplicando Oversampling buscando una mejora en la predicción.

- Luego de evaluar el modelo con mejor desempeño según las métricas definidas, se procede a agregar variables derivadas del CDR pero entrenando y testeando con 4 meses (desde diciembre del 2021 hasta marzo del 2022) debido a la disponibilidad de la data de registro de llamadas. Se compara el modelo base (entrenado con los 4 meses) sin las métricas de llamadas y el modelo que considera estas variables adicionales con el objetivo de determinar si esto aporta información relevante para el caso estudiado.

Para el primer modelo con Random Forest no se balancean los datos y se prueba diferentes combinaciones con respecto al número de árboles y profundidad. Se obtienen los mejores resultados con un modelo que utiliza 45 árboles, cuya profundidad es de 8 y 246 hojas en promedio. Para la base de testeo se tiene un AUC de 0.6655 y un error cuadrado medio de 0.3787.

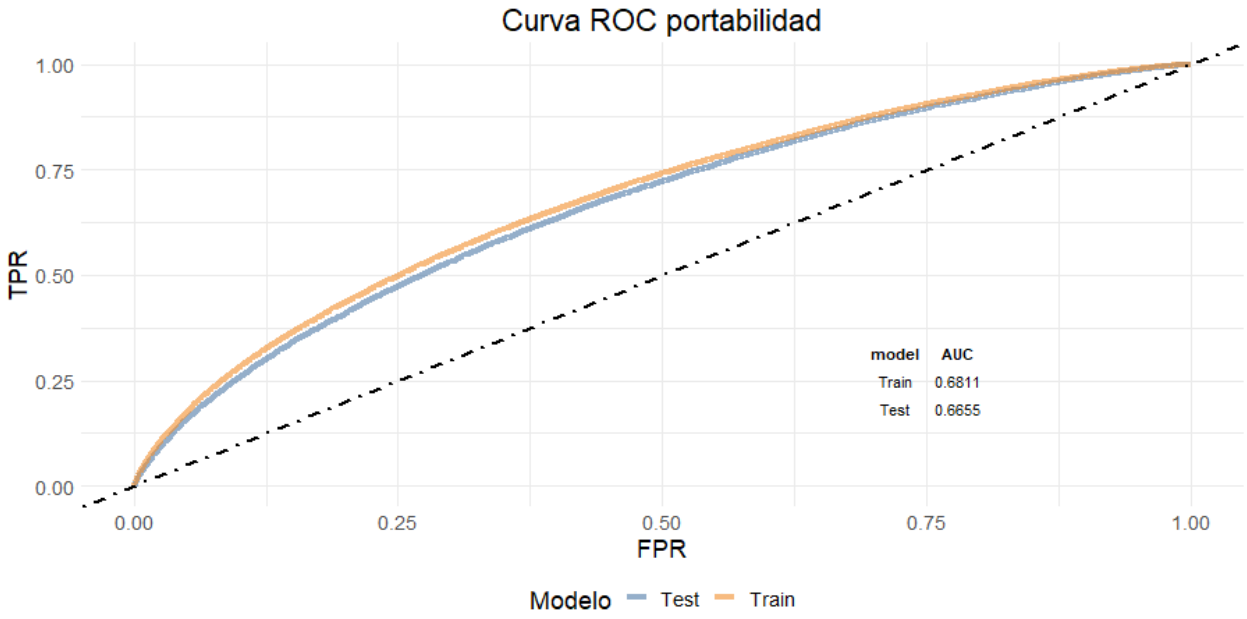


Ilustración 26: Curva ROC primer modelo RF sin balanceo de datos, 98 variables.

Se observan diferencias entre las curvas Lift tanto para la muestra de entrenamiento como testeo, sin embargo, no se presenta mayor sobreajuste en el

modelo. Para el percentil 1% obtienen 2.96 veces más casos positivos que respecto a una selección aleatoria. Ambas curvas son estrictamente decrecientes.

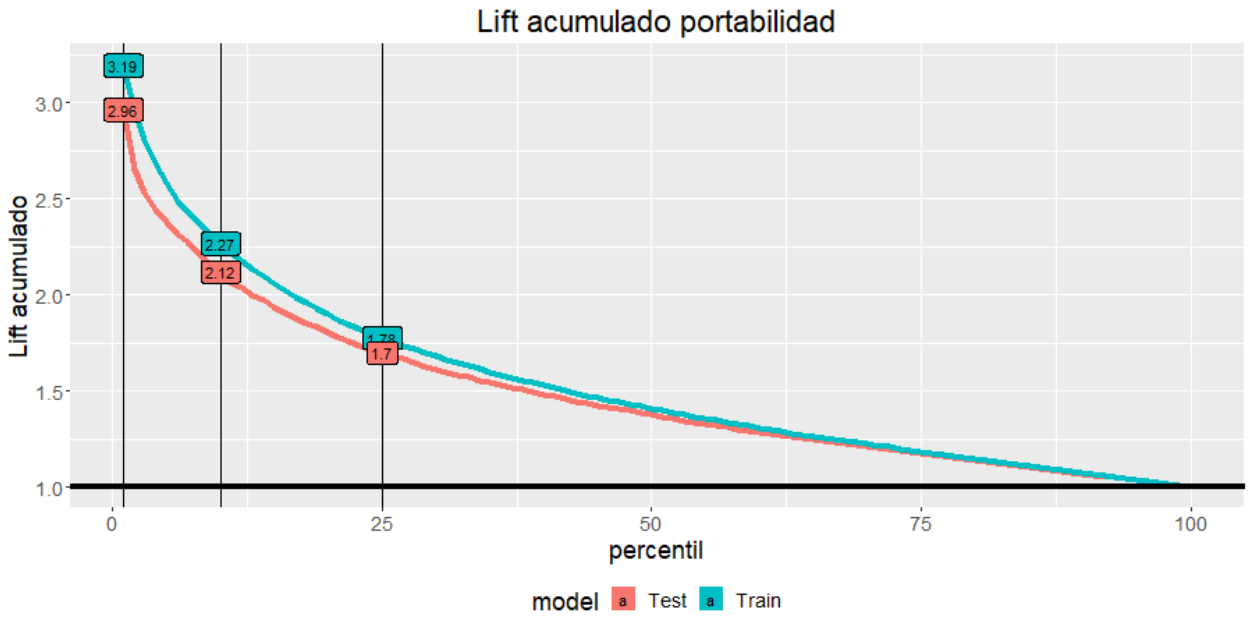


Ilustración 27: Curva Lift primer modelo RF sin balanceo de datos, 98 variables.

Se realiza las matrices de confusión a partir de los datos utilizados para el entrenamiento y validación del modelo. Se observan además las métricas de evaluación expuestas en las siguientes tablas, en donde la performance del modelo disminuye al utilizar la muestra de testeo.

		Predicted	
		0	**1**
Actual	**0**	38.151	19.035
	1	5.586	7.410

Tabla 2: Matriz de confusión RF (testeo)

Accuracy	**Precision**	**Sensitivity**	**Specificity**
0,649	0,280	0,570	0,667

Tabla 3: Métricas de evaluación desde matriz de confusión RF (testeo)

A partir de este modelo base, se procede a realizar un Oversampling para balancear la data, en donde se obtiene un modelo con un AUC de 0.6645 y una curva Lift en donde, para el set de testeo, se observa que dentro del percentil 1 es 2.88 veces más probable encontrar un caso positivo (portabilidad de línea adicional) sobre el modelo de selección aleatoria.

Se observa que ambos modelos no poseen gran sobreajuste pero con una baja capacidad predictiva. El balanceo de la muestra no presenta pequeñas mejorías en cuanto a las métricas utilizadas para la evaluación. Se procede a utilizar el algoritmo de Gradient Boosting Machine, el cual es entrenado a partir de predictores débiles como lo sería Random Forest.

Para el modelo GBM se consideran las mismas variables de los modelos anteriores y se comienza entrenando un modelo que no posee balanceo de datos. Se obtiene un AUC de 0.6838 para la base de testeo, presentando una mejora con respecto al modelo de Random Forest y tambien en comparación a un modelo base. Este modelo posee un error cuadrado medio de 0.3667. El mejor modelo se obtiene con 70 árboles, una profundidad de 6 y 63 hojas en promedio.

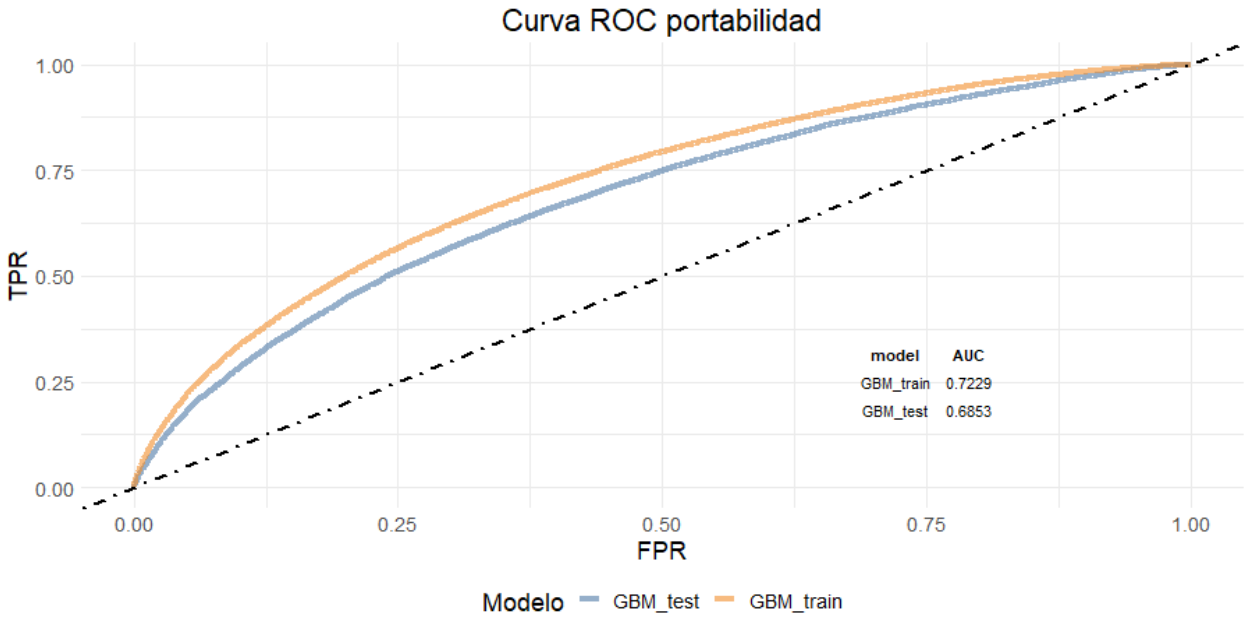


Ilustración 28: Curva ROC modelo GBM sin balanceo de datos, 98 variables.

Se observa una mejora en la predicción del modelo para la base de testeo. En el percentil 1 se obtienen 3.21 veces más casos positivos en comparación al modelo de selección aleatoria. Es importante destacar el comportamiento de la curva Lift, la cual es estrictamente decreciente, lo cual demuestra que es un modelo estable en cuanto a la predicción que se esta realizando a partir de los datos utilizados para entrenar y testear.

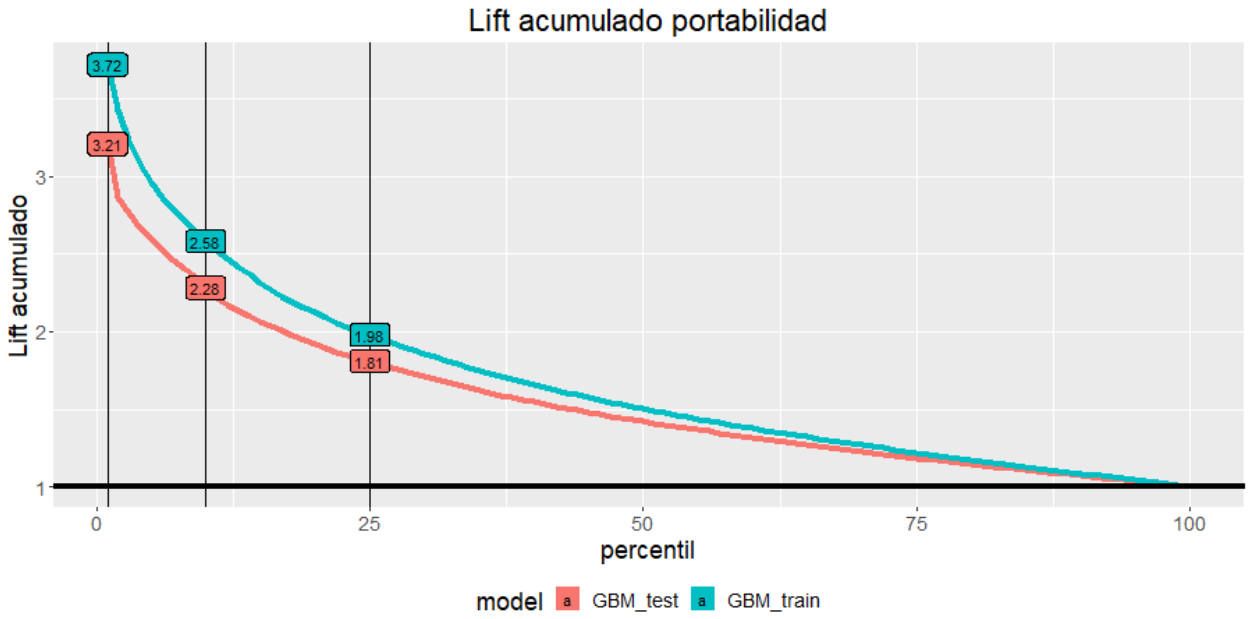


Ilustración 29: Curva Lift modelo GBM sin balanceo de datos, 98 variables.

Se realiza las matrices de confusión a partir de los datos utilizados para el entrenamiento y validación del modelo, en donde el umbral óptimo utilizado para la clasificación es de 0. 2122 y 0.2007 respectivamente. Se observan además las métricas de evaluación expuestas en las siguientes tablas

		Predicted	
		0	**1**
Actual	**0**	39.572	17.614
	1	5.505	7.491

Tabla 4: Matriz de confusión GBM (testeo)

Accuracy	Precision	Sensitivity	Specificity
0.671	0.298	0.576	0.692

Tabla 5: Métricas de evaluación desde matriz de confusión GBM (testeo)

Al realizar un balanceo utilizando Oversampling dentro de la muestra se obtiene un AUC de 0,6854 para la base de testeo, con una curva Lift en donde el percentil 1 presenta 3.13 más casos positivos que un modelo de clasificación aleatorio. Los gráficos y métricas obtenidas desde este modelo se pueden encontrar en el anexo. No se observa un aumento significativo en la performance del modelo tras balancear la muestra.

Buscando aumentar la capacidad predictiva de los modelos bases anteriores se utiliza el algoritmo XGBoost. En cuanto a la muestra de testeo se obtiene un AUC de

0.6836 y un RMSE de 0.3731. Se observa una leve disminución en las curvas ROC en comparación al modelo GBM, sin embargo, sigue siendo un modelo que discrimina de mejor manera en comparación al modelo aleatorio.

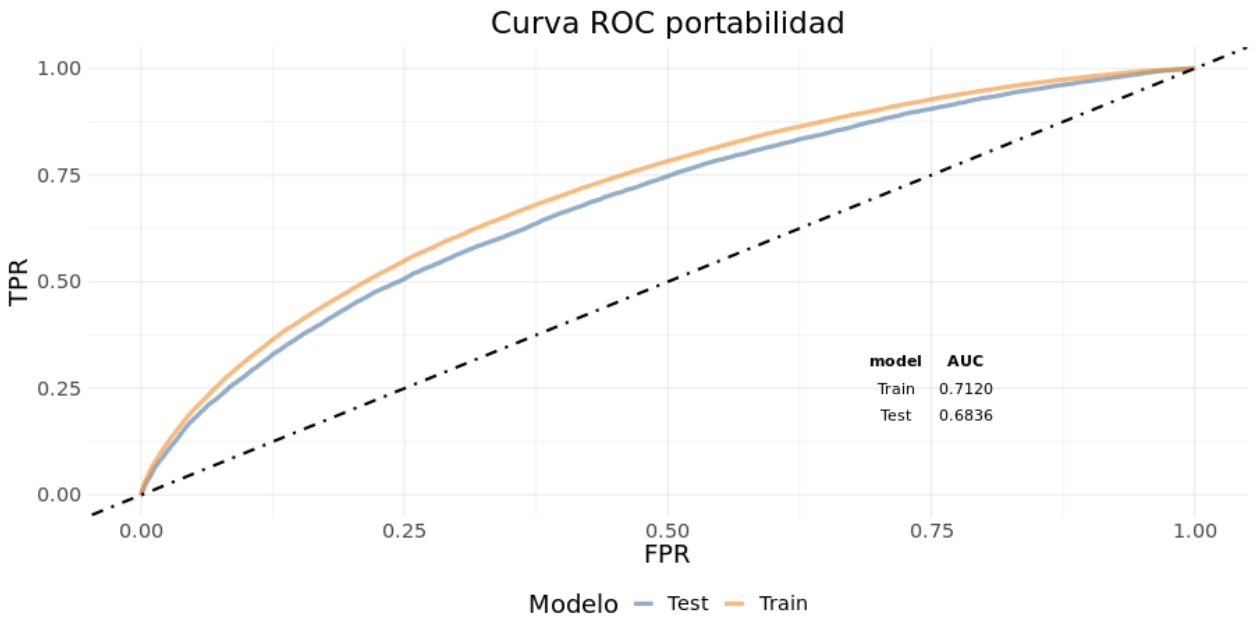


Ilustración 30: Curva ROC modelo XGBoost sin balanceo de datos, 98 variables.

Al observar la curva Lift no se observa sobreajuste de los datos, con un comportamiento decreciente y obteniendo 3.1 veces más casos positivos en el percentil 1 que un modelo de selección aleatoria.

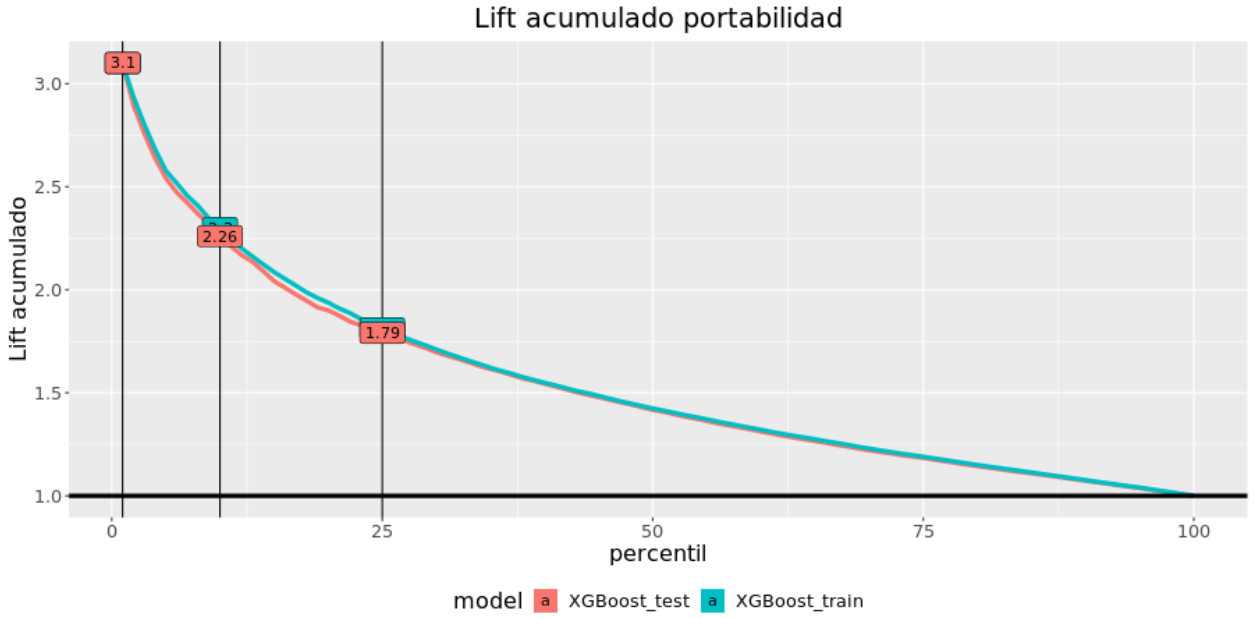


Ilustración 31: Curva Lift modelo XGBoost sin balanceo de datos, 98 variables.

Se realiza las matrices de confusión a partir de los datos utilizados para el entrenamiento y validación del modelo, en donde el umbral óptimo utilizado para la clasificación es de 0.1983 y 0.2026 respectivamente. Se observan además las métricas de evaluación expuestas en las siguientes tablas.

		Predicted	
		0	**1**
Actual	**0**	39.397	17.789
	1	5.496	7.500

Tabla 6: Matriz de confusión XGBoost sin balanceo de datos (testeo)

Accuracy	**Precision**	**Sensitivity**	**Specificity**
0.668	0.297	0.577	0.689

Tabla 7: Métricas de evaluación desde matriz de confusión XGBoost sin balanceo de datos (testeo)

Al igual que los modelos anteriores, se realiza un oversampling para correr el algoritmo de XGBoost, en donde se observa que a partir del AUC obtenido de 0.6823, no se obtienen mejoras apreciables con respecto al modelo sin balanceo. Al comparar el AUC de cada modelo entrenado se obtiene un mejor desempeño con un GBM con balanceo de datos, sin embargo, el modelo XGBoost sin balanceo corresponde a un modelo robusto. Es por esto que se utiliza el algoritmo XGBoost para el entrenamiento y testeo de un modelo que considera variables derivadas del registro de llamadas.

Debido a la disponibilidad de los datos del CDR (desde diciembre 2021 hasta marzo 2022), es que se entrena nuevamente un modelo base considerando estos 4 meses de historia. Esto se realiza con el fin de tener un punto de comparación previo a la adición de las variables de interacción de llamadas. De esta manera, se obtiene un AUC para el set de entrenamiento de 0.6770, detectando 3.27 veces más casos positivos en el primer percentil de la muestra.

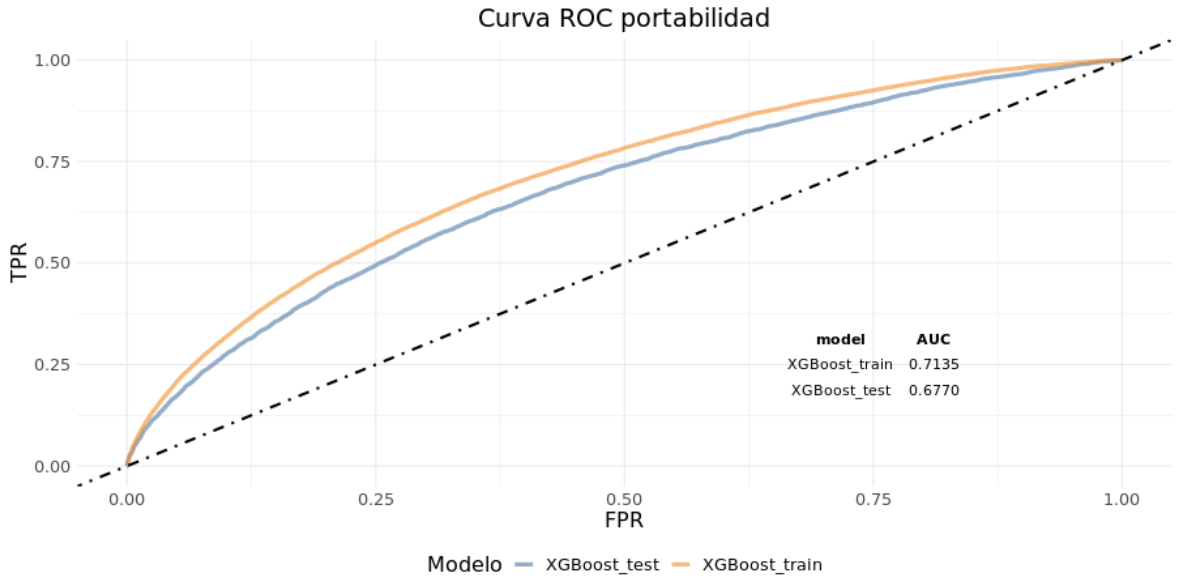


Ilustración 32: Curva ROC modelo XGBoost entrenado y testeado con 4 meses de historia (sin variables del CDR).

Posterior a esto, se entrena el modelo agregando variables derivadas del CDR que indica la proporción de llamadas y proporción de minutos hacia clientes postpago, prepago o no clientes por parte de un cliente titular de línea. También se incluye la métrica de centralidad Eigenvector, la cual indica la importancia del nodo dentro de la red. A partir de esto se obtiene un AUC de 0.6883. En el primer percentil se detectan 3.58 mas casos positivos en comparación al modelo de selección aleatoria.

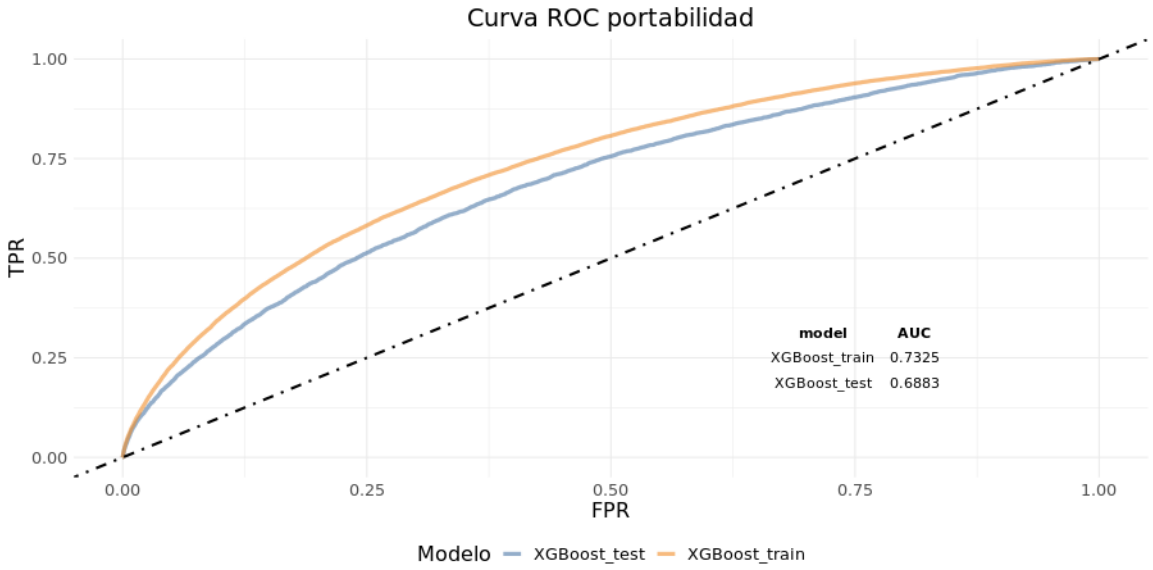


Ilustración 33: Curva ROC modelo XGBoost entrenado con 4 meses de historia (agregando métricas de centralidad derivadas del CDR).

Se observa una leve mejora en la performance, sin embargo, se esperaba un aumento mayor en la capacidad predictiva al agregar estas variables de interacción de llamadas. En este caso, se considera que la información entregada por el resumen de llamadas sería suficiente para explicar las interacciones del cliente con terceros, por lo que las variables del CDR no aportarían al modelo en cuestión.

Si bien, las variables de interacción del CDR no aportan valor para la predicción de líneas adicionales portadas, es una fuente de información que no ha sido explotada por la empresa y que entrega nuevas oportunidades de mejora de modelos existentes o el desarrollo de futuros modelos predictivos.

5.5. EVALUACIÓN

A partir de los modelos realizados en la sección anterior se procede a compararlos en base a las métricas de evaluación definidas, con el fin de seleccionar aquel que posea una mayor capacidad predictiva e interpretativa. Cabe destacar que dentro de los modelos comparados no fueron incluidas las métricas derivadas del análisis de grafos debido a la baja mejoría en cuanto a performance al incluir estas variables.

	RF sin balanceo	RF con balanceo	GBM sin balanceo	GBM con balanceo	XGBoost sin balanceo	XGBoost con balanceo
AUC	0.6655	0.6645	0.6853	0.6854	0.6825	0.6823
Lift 1%	2.96	2.88	3.21	3.13	3.1	3.05
Lift 10%	2.12	2.1	2.28	2.29	2.26	2.28
Lift 25%	1.7	1.69	1.81	1.81	1.79	1.79
Precision	0.280	0.270	0.298	0.306	0.297	0.312
Sensitivity	0.570	0.620	0.576	0.550	0.577	0.528
Specificity	0.667	0.619	0.692	0.717	0.689	0.735

Tabla 8: Tabla comparativa modelos ML.

A partir de la tabla comparativa no se obtiene un modelo que sobresalga en todas las métricas de evaluación, aunque se puede observar un comportamiento similar en cuanto a las métricas de evaluación para los modelos GBM y XGBoost, siendo un GBM con balanceo de datos el que mantiene una mejor performance. Debido a la similitud entre las métricas de los modelos con y sin balanceo es que se decide realizar las predicciones sin balanceo de datos. Por otra parte, los modelos XGBoost presentan un menor sobreajuste de los datos, por lo que serían menos sensibles a variaciones en la data entregando predicciones más precisas. A continuación, se detallan resultados derivados del modelo XGBoost sin balanceo.

Se presenta la curva Lift con respecto a los deciles de clasificación para el set de entrenamiento en la siguiente tabla. Se observa que aproximadamente un 51% de los efectivos detectados se concentran dentro de los 3 primeros deciles. Esta característica es importante para el negocio, debido a que no es necesario recorrer toda la base para llegar a la mitad de los clientes objetivo al momento de realizar una campaña de contactabilidad mediante un canal. La tasa de detección acumulada al tercer decil es cercana al 31%.

Decil	Total	Efectivos	Tasa	Tasa Acumulada
1	7.019	2.934	41,80%	41,80%
2	7.018	2.001	28,51%	35,16%
3	7.018	1.675	23,87%	31,39%
4	7.018	1.412	20,12%	28,58%
5	7.018	1.183	16,86%	26,23%
6	7.018	1.019	14,52%	24,28%
7	7.019	875	12,47%	22,59%
8	7.018	787	11,21%	21,17%
9	7.018	658	9,38%	19,86%
10	7.019	449	6,40%	18,51%
Total	70.182	12.993	18,51%	18,51%

Tabla 9: Tabla resumen curva Lift por decil, modelo XGBoost sin balanceo.

Es posible identificar aquellas variables cuyo aporte al modelo fue significativo. Esta importancia de variables es calculada a partir de la variación del error cuadrado medio al incluir la variable predictora como una divisora de un árbol de decisión. El siguiente gráfico muestra las 10 variables más importantes para el modelo, de las cuales se obtienen ciertas interpretaciones.

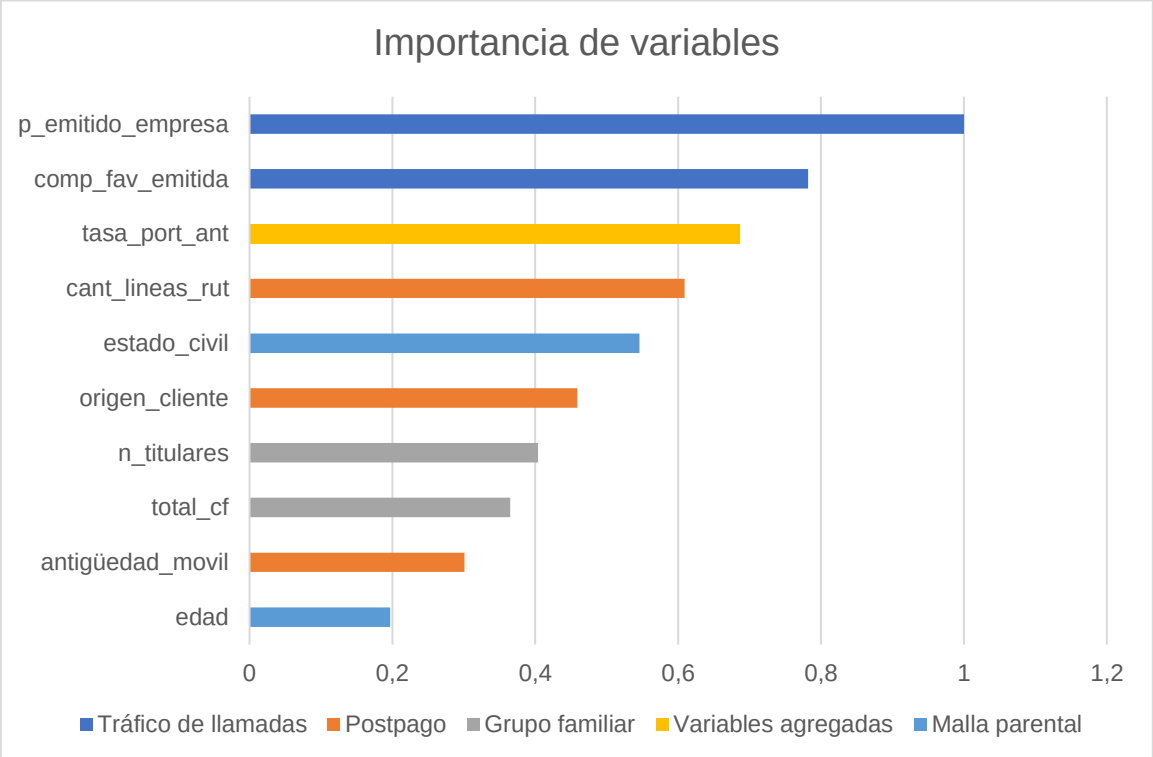


Ilustración 34: Importancia de variables modelo GBM sin balanceo de datos.

- **Porcentaje emitido empresa:** en la sección de preparación de datos se observó una tendencia marcada sobre la tasa de portabilidad de líneas adicionales. Con esto se obtiene que aquellos clientes que emiten un mayor porcentaje de llamadas hacia clientes dentro de la empresa tendrían un impacto negativo sobre la obtención de líneas adicionales. La interpretación de esta variable tiene relación con el círculo cercano del cliente, siendo clientes aquellas personas con las que más interactúa, por lo que habría menos líneas propensas a ser portadas bajo su cuenta titular.

- **Compañía favorita emitida:** en relación con el porcentaje de llamadas emitidas a clientes de la empresa, se observa una baja en la tasa de portabilidad de líneas en comparación a aquellos clientes cuya compañía favorita corresponde a la competencia. Este fenómeno puede estar relacionado con la cantidad de líneas propensas a adquirir como adicionales, ya que, si el cliente se comunica con terceros fuera de la empresa, estas líneas podrían ser objetivo de campañas de portabilidad, por lo que aumentaría su propensión a adquirir un nuevo servicio en el mes evaluado.

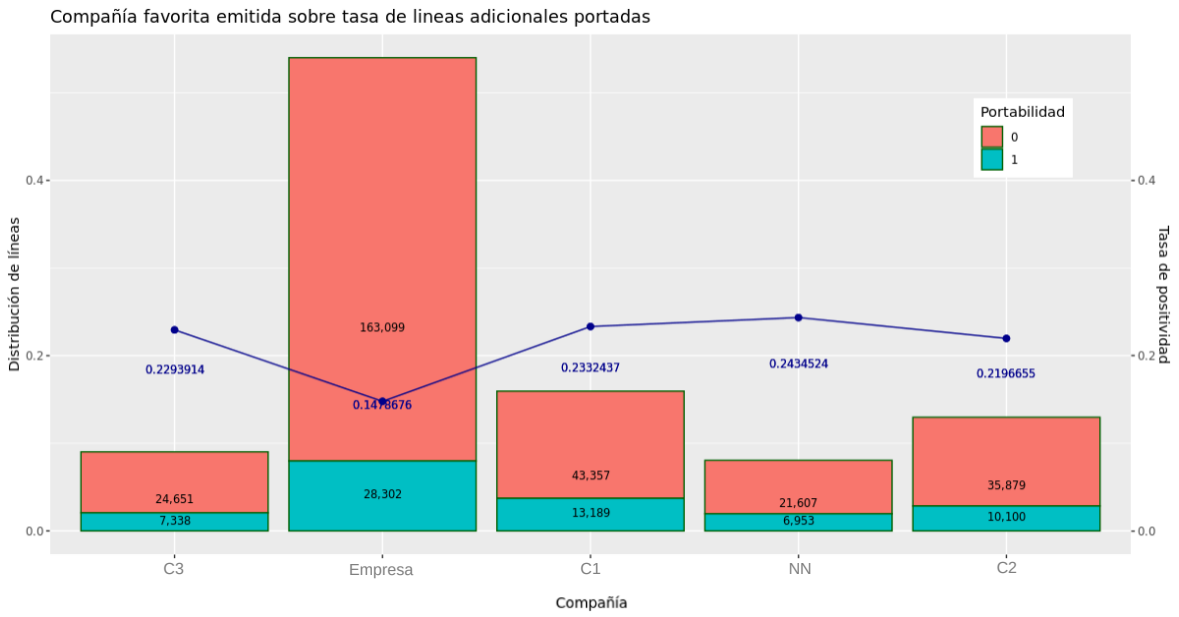


Ilustración 35: Tasa de portabilidad de líneas adicionales según compañía favorita emitida por cliente titular. Elaboración propia

- **Tasa portabilidad según antigüedad:** un aumento en la permanencia del cliente tiene un impacto negativo sobre el fenómeno estudiado. Se observa que a medida que aumenta la antigüedad de la línea titular dentro de la empresa la tasa de

portabilidad de líneas adicionales disminuye. Esto podría reflejar un perfil conservador del cliente al ser reacio a contratar nuevas líneas o, por el contrario, haber contratado líneas adicionales dentro de su núcleo familiar, aumentando la penetración de la empresa en este hogar. Esta última interpretación es validada gracias a la comparación de tasas de fuga, observando en aquellos hogares cerrados (aquellos que poseen una cantidad de líneas igual o mayor al número de integrantes) una mayor permanencia debido a la menor tasa de churn.

- **Cantidad de líneas por Rut:** aquellos clientes titulares que tienen más de una línea asociada a su cuenta (es decir, que ya posee al menos una línea adicional) tienen una menor propensión a portar líneas adicionales en un futuro. Este comportamiento es esperable no solo para líneas portadas, si no que para líneas migradas o activadas. La relación entre la tasa de portabilidad y la cantidad de líneas se analizó en secciones anteriores.

- **Estado civil:** dentro de la muestra de datos se observa que aproximadamente un 40% de los clientes están casados y poseen una tasa de portabilidad de líneas adicionales mayor en comparación a aquellos que no registran ninguna unión matrimonial. El hecho de estar casado implica un mínimo de dos integrantes dentro del grupo familiar, por lo que, si uno de ellos es cliente titular y el otro pertenece a otra compañía, entonces habría oportunidades de adquirir este cliente mediante ofertas de portabilidad.

- **Origen cliente:** esta variable indica el tipo de línea que obtuvo el cliente titular al llegar a la compañía. La tendencia a portar líneas adicionales es mayor en aquellos clientes que llegaron mediante portabilidades numéricas. Es probable que estos clientes repitan este patrón de adquisición de nuevos servicios al conocer previamente el proceso de portabilidad numérica.

- **Número de titulares:** aquellos grupos familiares en donde existen más de un titular tenderían a portar menos líneas adicionales. Es posible clasificar estos hogares como “monotitular” cuando existe solo un cliente titular o “multitular” cuando se identifican a más de uno. Por lo general, los clientes tienden a pertenecer a grupos familiares monotitulares debido a que adquirir una línea adicional es un proceso más rápido y con menos requisitos previos que adquirir una nueva línea. Sin embargo, si se ve que la familia tiene más de un titular es posible que consideren que la gestión de las líneas sea más sencilla de manera individual siendo menos propensos a adquirir líneas adicionales a su cuenta.

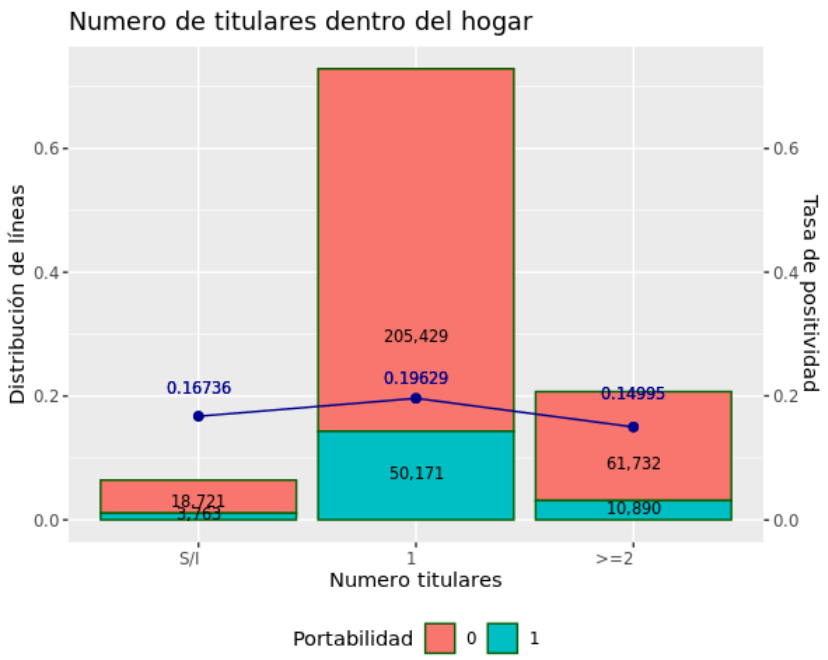


Ilustración 36: Tasa de portabilidad líneas adicionales sobre cantidad de titulares identificados dentro de un grupo familiar.

- **Total cargo fijo:** esta variable caracteriza a nivel hogar la facturación de los servicios contratados. Se en el gráfico presentado en la sección 5.3 se observa un impacto negativo en la tasa de portabilidad a medida que este monto aumenta. Este comportamiento es esperable, ya que un cliente que paga altos montos por los servicios se traduce en un mayor número de líneas contratadas (en general se adquieren planes con cargo fijo menor a 16mil pesos) por lo que tendría alto porcentaje de penetración, relacionando asi con la variable del punto anterior y la permanencia en la compañía. Otra interpretación derivada de esta variable corresponde a la baja propensión o reacción frente ofertas de nuevos servicios debido al aumento en la facturación mensual.

- **Antigüedad móvil:** interpretación de tasa de portabilidad según la antigüedad. A medida que la permanencia del cliente aumenta dentro de la compañía, la portabilidad de líneas adicionales decae.

- **Edad:** las portabilidades de líneas adicionales aumentan a medida que aumenta la edad del cliente titular, siendo baja para personas menores a 26 años debido a la dificultad de adquirir servicios de telefonía según su situación económica/laboral que pueda influir en el no pago del servicio y por ende un mayor riesgo para la empresa. Es esperable que la adquisición de líneas adicionales aumente a medida

que aumenta la edad del titular, ya que se espera que su grupo familiar o circulo social se componga de más personas. Es por esto que habrían más líneas propensas a ser portadas a la empresa.

En anexos se encuentra el gráfico de importancia de variables a partir de la contribución medida a través del método Shapley Values, teniendo un orden similar a la lista anterior. Desde este gráfico se puede interpretar el impacto de la variable sobre el fenómeno estudiado, en donde a modo de ejemplo, aquellos clientes que poseen una línea asociada a su cuenta registrarían una mayor propensión a portar una línea adicional (observaciones en rosado al lado derecho de la gráfica). Análogo al ejemplo, es posible obtener la interpretación de las 20 variables más importantes para el modelo.

Teniendo en consideración estas variables se puede caracterizar el perfil del cliente que se encontraría más propenso a portar una línea adicional. Este perfil sería un cliente que haya portado su línea recientemente (hace menos de 5 meses), quien posee solo su línea titular o pocas adicionales (teniendo en consideración la cantidad de miembros del grupo familiar al que pertenezca) cuya interacción de llamadas se concentra con clientes de la competencia. En general, el cargo fijo (a nivel familiar) pagado por este cliente se concentra en montos bajos.

5.6. DESPLIEGUE

5.6.1. DESARROLLO DE EXPERIMENTO

Para la implementación del modelo realizado se propone un experimento cuyo objetivo es direccionar la oferta de líneas adicionales a clientes titulares con foco en portabilidad. Este experimento queda propuesto para el canal Call Center, sin embargo, el proceso de perfilamiento de clientes y cálculo de propensión es aplicable para todos los canales de contacto de la empresa. Es importante recordar que este modelo es un complemento al modelo de líneas adicionales que se utiliza para la oferta en canales, en donde el canal Call Center gestiona hasta el sexto decil.

El perfilamiento de los clientes se detalla a continuación.

1. A finales de cada mes se entrena el modelo tomando los 8 meses de historia considerando el desfase ilustrado en la sección de modelamiento.

2. Por otro lado, se identifica la base potencial, la cual se obtiene a partir de los clientes perfilados por el modelo de líneas adicionales. Por lo general, esta base coincide con el cierre de clientes postpago mensual correspondiente al mes previo.

3. Sobre los clientes potenciales, se genera el tablón con las variables utilizadas para el entrenamiento del modelo.

4. Se calcula la propensión para portabilidad de línea adicional portada utilizando el modelo entrenado.

5. Se excluye todos los clientes que han sido contactado en la campaña del mes anterior y que adquirieron una línea adicional portada, debido a la baja probabilidad de que adquiera una segunda línea el nuevo mes.

6. Se genera una base de clientes la cual contiene el score de propensión de obtener una línea adicional y además la propensión de que esta sea provenga por una portabilidad numérica. A partir de esta base, es posible generar la intersección de los clientes por deciles de propensión de cada modelo. Esta matriz se utiliza para generar la división de clientes a gestionar y evaluación económica de la nueva segmentación.

De manera ilustrativa, se perfila los clientes potenciales a obtener líneas adicionales portadas en junio del 2022 de un total de 3.064.868, obteniendo la siguiente matriz de intersección.

Decil				Modelo segundas líneas portadas									
Tasa de efectividad				1	2	3	4	5	6	7	8	9	10
				41,8%	28,5%	23,9%	20,1%	16,9%	14,5%	12,5%	11,2%	9,4%	6,4%
Modelo segundas líneas	306487	1	2,31%	22.332	25.277	26.223	25.982	27.724	29.326	31.615	34.519	38.517	44.971
	306487	2	1,58%	27.477	26.885	26.405	25.924	27.863	29.630	31.912	32.556	35.532	42.301
	306487	3	1,14%	31.638	27.074	26.718	26.796	28.121	29.065	29.758	32.157	34.245	40.913
	306487	4	0,95%	35.336	28.982	27.351	27.148	28.078	28.365	28.762	30.528	33.527	38.411
	306487	5	0,90%	36.464	30.223	27.336	26.665	27.491	28.996	29.062	30.125	32.268	37.855
	306487	6	0,73%	37.614	30.349	29.249	28.387	27.934	28.239	28.937	29.494	31.158	35.126
	306487	7	0,55%	37.051	30.948	29.933	29.321	28.346	28.602	29.240	29.610	30.854	32.581
	306487	8	0,42%	34.001	30.478	29.388	29.544	29.809	29.797	31.646	30.544	31.441	29.839
	306487	9	0,31%	30.440	29.392	30.088	31.906	32.263	31.170	31.432	32.028	30.980	26.788
	306487	10	0,21%	25.271	35.095	37.062	37.147	35.589	34.225	31.525	29.075	24.497	17.001

Tabla 10:Matriz de priorización de campaña ventas líneas adicionales. Clientes perfilados para junio 2022 según combinación de deciles.

A partir de esta matriz de clientes perfilados para ambos modelos es posible generar muestras para medir la efectividad de campañas de venta por sobre la gestión realizada actualmente. La matriz anterior es dividida según colores, en amarillo aquellos clientes que recibirán una oferta de líneas adicionales con énfasis en portabilidad, en verde aquellos que recibirán oferta con énfasis en migración o activación y en rojo el área que no será gestionada en el experimento. En futuras campañas es posible extender la gestión según la efectividad obtenida después del experimento.

Como se puede observar, las tasas de efectividad varían en cada intersección, por lo que se simplifican estas áreas tomando una tasa de efectividad promedio. Esto ayuda en la selección de muestras aleatorias para la evaluación de la campaña. Con esto se seleccionan 3 grupos.

- Grupo 1 (área amarilla): oferta direccionada en líneas adicionales portadas. Tasa de efectividad para modelo de segundas líneas y segundas líneas portadas de 1,27% y 31,4% respectivamente.

- Grupo 2 (área verde): oferta direccionada en líneas adicionales migradas o portadas. Tasa de efectividad para modelo de segundas líneas y segundas líneas portadas de 1,27% y 13% respectivamente.

- Grupo 3 o control: clientes seleccionados de manera aleatoria de los grupos anteriores quienes recibirán la gestión de líneas adicionales que se utiliza actualmente, sin diferenciación entre tipo de línea.

Considerando la limitación de contactabilidad mensual que posee el Call Center con respecto a la campaña de líneas adicionales (la cual no superan los 50.000 clientes contactados de manera mensual) es que se selecciona una muestra aleatoria de clientes que recibirán ofertas direccionadas a cierta categoría de línea. Esta selección se realiza en base a las tasas de detección derivadas de los modelos con el fin de obtener un grupo de clientes efectivos comparables para concluir. A continuación, se muestra esta selección, en donde E_{SL} indica la probabilidad de ser una línea adicional o "segunda línea", E_{SLP} la probabilidad que esta línea adicional sea una portabilidad y E_{SLAM} la probabilidad que esta línea adicional sea una activación o migración (calculada como $1 - E_{SLP}$).

Grupo	E_{SL}	E_{SLP}	E_{SLAM}	Selección aleatoria	Efectivos
G1	1,27%	31,4%	68,6%	17.580	70
G2[3]	1,27%	13,0%	87,0%	6.343	70
Control G1	1,27%	31,4%	68,6%	17.580	70
Control G2	1,27%	13,0%	87,0%	6.343	70
Total				47.846	

Tabla 11: Distribución de base de clientes para desarrollo de experimento.

Con el objetivo de evitar sesgos de selección, es importante corroborar que las variables de la muestra seleccionada mantengan la distribución de la muestra total. Se espera que la segmentación realizada por clasificación de líneas potencie la venta de portabilidades por sobre la venta de líneas activadas o migradas debido al valor que reportan las portabilidades para la empresa.

En la siguiente sección se realiza una evaluación económica simplificada, en base al potencial de venta para la totalidad de clientes perfilados, analizando además la importancia de las portabilidades por sobre las otras clasificaciones de líneas.

[3] La detección y la selección aleatoria de este grupo se enfoca en la tasa para activación o migración ya que la oferta se realiza en base a clientes menos propensos a portar.

5.6.2. IMPACTO ECONÓMICO

En línea con la sección anterior, se realiza una evaluación económica de ventas derivadas de la segmentación obtenida por el modelo de líneas adicionales portadas. Esta evaluación es realizada para el perfilamiento de clientes para el mes de junio 2022. Se asume que se contacta a través de todos los canales que dispone la empresa a la totalidad de clientes potenciales para el mes, suponiendo además que todos los canales poseen la misma efectividad de venta.

Según la data histórica de actividad comercial, el cargo fijo de las líneas adicionales portadas evaluado al sexto mes después de la adquisición es de $12.552. Respecto a las líneas activadas o migradas, el cargo fijo promedio de ambas clasificaciones de línea es de $6.280. La evaluación económica es realizada con estos valores de cargo fijo debido a que además se estaría considerando la tasa de fuga de manera implícita (tasas de fuga al sexto mes para portabilidad, activación y migración corresponde a 12,78%, 27,09% y 26,69% respectivamente).

Considerando la distribución de líneas de la matriz de intersección, las tasas de efectividad entregadas por cada modelo y el cargo fijo por tipo de línea, es posible obtener la ganancia esperada calculada de la siguiente manera.

$$Ganancia_{esperada} = C_{ij} * P_i * P_j * CF_P + C_{ij} * P_i * \left(1 - P_j\right) * CF_{AM} \qquad (12)$$

En donde C_{ij} corresponde al grupo de clientes dentro de la intersección del decil i del modelo de segundas líneas con el decil j del modelo de segundas líneas portadas, P_i a la tasa de detección o probabilidad de ser una línea adicional, P_j a la probabilidad de ser una línea adicional portada, CF_P el cargo fijo al sexto mes de una línea adicional portada y CF_{AM} cargo fijo promedio al sexto mes de una línea activada o migrada.

Se observa una ganancia de $206M aproximadamente al gestionar toda la base (sobre el supuesto que todo cliente que es gestionado contrata una línea adicional). Si se recorre solo los primeros 3 deciles del modelo de líneas adicionales portadas, dentro de los cuales se espera obtener al 51% de los efectivos, se obtiene el 31% de las ganancias. Sin embargo, se evidencia una concentración de mayores ganancias en la parte superior de la matriz, por lo que se genera un trade off para la gestión de esta base con el objetivo de maximizar las ganancias de la campaña.

Decil			Modelo segundas líneas portadas				
			1	2	3	4	5
Tasa de efectividad			41,8%	28,5%	23,9%	20,1%	16,9%
Modelo segundas líneas	2,31%	1	$ 4.594.352	$ 4.713.335	$ 4.713.112	$ 4.528.735	$ 4.701.254
	1,58%	2	$ 3.868.048	$ 3.430.366	$ 3.247.413	$ 3.091.970	$ 3.233.024
	1,14%	3	$ 3.207.840	$ 2.488.047	$ 2.366.749	$ 2.301.915	$ 2.350.192
	0,95%	4	$ 2.992.577	$ 2.224.652	$ 2.023.663	$ 1.947.951	$ 1.959.971
	0,90%	5	$ 2.924.766	$ 2.197.213	$ 1.915.541	$ 1.812.068	$ 1.817.513
	0,73%	6	$ 2.431.314	$ 1.778.000	$ 1.651.706	$ 1.554.591	$ 1.488.261
	0,55%	7	$ 1.816.316	$ 1.375.068	$ 1.281.944	$ 1.217.794	$ 1.145.365
	0,42%	8	$ 1.272.773	$ 1.034.077	$ 961.089	$ 937.008	$ 919.757
	0,31%	9	$ 840.482	$ 735.559	$ 725.809	$ 746.389	$ 734.255
	0,21%	10	$ 462.136	$ 581.693	$ 592.119	$ 575.534	$ 536.434

Tabla 12: Ganancias esperadas según decil de propensión. Parte I.

Decil			Modelo segundas líneas portadas				
			6	7	8	9	10
Tasa de efectividad			41,8%	28,5%	23,9%	20,1%	16,9%
Modelo segundas líneas	2,31%	1	$ 4.873.416	$ 5.159.884	$ 5.571.025	$ 6.113.733	$ 6.943.986
	1,58%	2	$ 3.369.396	$ 3.563.933	$ 3.595.297	$ 3.859.255	$ 4.469.408
	1,14%	3	$ 2.380.564	$ 2.393.690	$ 2.557.856	$ 2.678.911	$ 3.113.530
	0,95%	4	$ 1.940.472	$ 1.932.425	$ 2.028.231	$ 2.190.744	$ 2.441.580
	0,90%	5	$ 1.878.720	$ 1.849.260	$ 1.895.575	$ 1.996.901	$ 2.278.896
	0,73%	6	$ 1.474.436	$ 1.483.858	$ 1.495.587	$ 1.553.889	$ 1.704.083
	0,55%	7	$ 1.132.600	$ 1.137.164	$ 1.138.718	$ 1.166.953	$ 1.198.769
	0,42%	8	$ 901.020	$ 939.798	$ 896.969	$ 908.076	$ 838.366
	0,31%	9	$ 695.216	$ 688.513	$ 693.767	$ 659.986	$ 555.149
	0,21%	10	$ 505.566	$ 457.353	$ 417.107	$ 345.629	$ 233.346

Tabla 13: Ganancias esperadas según decil de propensión. Parte II.

Tomando como supuesto que el costo por llamar a un cliente mediante el canal de Call Center y ofrecer productos móviles postpago es de $50, se puede estimar los ingresos de gestionar la base completa. Este cálculo excluye pago de comisiones, debido a que se desea entender el impacto monetario que tiene el modelo con la entrega de propensiones por clientes sin incentivos que puedan alterar los resultados.

De esta manera se observan ganancias positivas en la mayoría de los deciles que se gestionan para campañas de segundas líneas (a excepción de los últimos dos deciles correspondientes al ordenamiento del modelo de segundas líneas portadas). Por debajo de estos deciles se incurrirían en gastos que no serían compensados debido a que las propensiones de contrataciones son menores que el costo de gestión.

Decil			Modelo segundas líneas portadas				
			1	2	3	4	5
Tasa de efectividad			41,8%	28,5%	23,9%	20,1%	16,9%
Modelo segundas líneas	2,31%	1	$3.477.743	$ 3.449.476	$ 3.401.968	$ 3.229.619	$ 3.315.032
	1,58%	2	$2.494.189	$ 2.086.103	$ 1.927.170	$ 1.795.747	$ 1.839.863
	1,14%	3	$1.625.953	$ 1.134.371	$ 1.030.828	$ 962.096	$ 944.118
	0,95%	4	$1.225.798	$ 775.572	$ 656.120	$ 590.547	$ 556.095
	0,90%	5	$1.101.550	$ 686.046	$ 548.743	$ 478.806	$ 442.944
	0,73%	6	$ 550.596	$ 260.573	$ 189.255	$ 135.230	$ 91.559
	0,55%	7	$ -36.250	$ -172.322	$ -214.696	$ -248.261	$ -271.956
	0,42%	8	$ -427.255	$ -489.806	$ -508.294	$ -540.204	$ -570.703
	0,31%	9	$ -681.499	$ -734.016	$ -778.611	$ -848.903	$ -878.877
	0,21%	10	$ -801.428	$-1.173.053	$-1.260.996	$-1.281.810	$-1.243.015

Tabla 14: Ingresos luego de gestión de campaña. Parte I

Decil			Modelo segundas líneas portadas				
			6	7	8	9	10
Tasa de efectividad			14,5%	12,5%	11,2%	9,4%	6,4%
Modelo segundas líneas	2,31%	1	$ 3.407.141	$ 3.579.127	$ 3.845.092	$ 4.187.873	$ 4.695.420
	1,58%	2	$ 1.887.875	$ 1.968.318	$ 1.967.512	$ 2.082.632	$ 2.354.360
	1,14%	3	$ 927.289	$ 905.772	$ 949.987	$ 966.679	$ 1.067.860
	0,95%	4	$ 522.233	$ 494.329	$ 501.838	$ 514.373	$ 521.020
	0,90%	5	$ 428.899	$ 396.168	$ 389.315	$ 383.490	$ 386.154
	0,73%	6	$ 62.504	$ 37.009	$ 20.879	$ -4.027	$ -52.194
	0,55%	7	$ -297.492	$ -324.857	$ -341.789	$ -375.732	$ -430.293
	0,42%	8	$ -588.838	$ -642.499	$ -630.226	$ -663.983	$ -653.600
	0,31%	9	$ -863.281	$ -883.085	$ -907.658	$ -889.034	$ -784.250
	0,21%	10	$-1.205.670	$-1.118.903	$-1.036.631	$ -879.210	$ -616.708

Tabla 15: Ingresos luego de gestión de campaña. Parte II

6. CONCLUSIONES

6.1. CONCLUSIONES GENERALES

Mejorar la segmentación de clientes postpago con respecto a la adquisición de líneas adicionales fue la motivación del presente trabajo, sustentado en la importancia de aumentar la penetración de la empresa dentro de un grupo familiar de un cliente titular. Es por esto que se realizaron diferentes modelos predictivos con el objetivo de identificar clientes que posean una mayor propensión a obtener líneas adicionales contratadas a través de una portabilidad numérica, dada la importancia de esta clasificación por sobre las migraciones y activaciones de líneas.

Luego de evaluar diferentes modelos de predicción, se observan buenos resultados[4] tanto para el algoritmo Gradient Boosting Machine como XGBoost, entregando métricas de evaluación similares. Tras utilizar una técnica de balanceo de muestra no se obtienen mejoras significativas, por lo que se mantuvo un modelo simple para la predicción.

Uno de los objetivos planteados dentro del trabajo consistió en evaluar el aporte de variables de externalidad de redes a partir del análisis de datos obtenidos del registro de llamadas o CDR. Utilizando teoría de grafos se construyeron métricas de interacción por parte del cliente con terceros, sin embargo, no se obtuvieron los resultados esperados con respecto a mejorías en performance del modelo e interpretabilidad. Es probable que la información aportada por estas variables ya esté capturada en el resumen de llamadas, por lo que no aportaría valor para el modelo desarrollado. Pese a esto, este análisis entrega nuevas oportunidades que no han sido explotadas por la empresa. Específicamente, las variables construidas pueden ser insumo de otros modelos que la empresa ya posee o futuros modelos a desarrollar, tales como modelos de fuga, modelos de riesgo, identificación de perfiles de clientes, entre otros. Por esto, se considera un hallazgo relevante.

A partir de los resultados obtenidos con los modelos entrenados, se pudo obtener un ordenamiento efectivo de los clientes para la gestión de campañas con foco en portabilidad numérica. En esta línea, fue posible identificar al 51% de los clientes con mayor propensión dentro de los primeros 3 deciles, entregando la posibilidad de contactar a un grupo reducido de clientes con tasas de detección altas, característica relevante dada las limitaciones de contactabilidad de los canales de la empresa. A partir de esta

[4] Medido en base a criterio del área de Analytics de la empresa, donde valores dentro del rango 0.65 y 0.7 para el AUC es considerado como un buen modelo.

segmentación se espera una mejora en la oferta de servicios existentes de líneas adicionales.

Finalmente, se pudo caracterizar el perfil del cliente que se encuentra más propenso a portar una línea adicional. Este corresponde a aquel que portó su línea titular de manera reciente (hace menos de 5 meses), que posee pocas líneas asociadas a su cuenta, que interactúa en mayor proporción con clientes de la competencia y que pertenece a un grupo familiar con un cargo fijo facturado bajo. Se espera que el experimento planteado permita corroborar esta identificación de perfil basada en la interpretación del análisis exploratorio e importancia de variables obtenida por los modelos predictivos.

6.2. RECOMENDACIONES Y TRABAJO FUTURO

A continuación, se plantean espacios de mejora para el problema estudiado dentro del trabajo de título y que no fueron abordados. Estos puntos podrían ser relevantes para un trabajo futuro.

La modelación del problema fue realizada utilizando una clasificación binaria para el cálculo de propensión a portar la línea adicional. Sin embargo, debido a que se poseen tres categorías de líneas, existe la oportunidad de profundizar el estudio a través de un modelo de predicción multiclase. Esto permitiría realizar una segmentación más detallada en cuanto a la oferta a realizar dentro de las campañas de venta de líneas adicionales.

Por otra parte, la construcción de variables se realizó en base al comportamiento del cliente titular debido a la dificultad de predecir sobre una línea móvil que no pertenece a la empresa. El análisis de registro de llamadas también fue realizado en base al cliente titular, por lo que futuras mejoras podrían estar enfocadas en la identificación de perfiles de “no clientes”, entregando de esta manera información que puede ser relevante sobre la línea a portar y que no es posible obtener de manera directa. También es posible realizar un trabajo enfocado directamente en el análisis de grafos y el aporte que pueda entregar estas variables para otros modelos de predicción.

Finalmente, es importante realizar una evaluación luego de la implementación del modelo. El despliegue puede evidenciar espacios de mejora o problemas no considerados durante el desarrollo del trabajo, tales como los incentivos que poseen los ejecutivos frente a la oferta de un tipo de línea por sobre otras, lo que afectaría en la decisión de adquisición de un servicio por parte del cliente.

7. BIBLIOGRAFÍA

[1] *Memoria integrada de la empresa desde 2017 a 2021.*

[2] Odii, J. (s.f.). *A predictive model for evaluating Mobile Number Portability in Nigeria.* Obtenido de https://www.researchgate.net/publication/299443776_A_Predictive_Model_for_Evaluating_Mobile_Number_Portability_in_Nigeria

[3] Shin, D., & Kim, W. (2008). *Forecasting customer switching intention in mobile service: An exploratory study of predictive factors in mobile number portability.* Obtenido de https://www.researchgate.net/publication/222548252_Forecasting_customer_switching_intention_in_mobile_service_An_exploratory_study_of_predictive_factors_in_mobile_number_portability

[4] Kaur, G., & Sambyal, R. (2016). *Exploring Predictive Switching Factors for Mobile Number Portability.* Obtenido de https://journals.sagepub.com/doi/full/10.1177/0256090916631638

[5] Ma, L., Krishnan, R., & Montgomery, A. (2010). *Homophily or Influence? An Empirical Analysis of Purchase within a Social Network.* Obtenido de https://community.mis.temple.edu/seminars/files/2011/02/krishnan-homophily-influence.pdf

[6] Raeef Al-Molhem, N., Rahal, Y., & Dakkak, M. (2019). *Social Network analysis in Telcom data.* Obtenido de https://www.researchgate.net/publication/337289216_Social_network_analysis_in_Telecom_data

[7] Papadopoulos, S., Kompatsiaris, Y., Vakali, A., & Spyridonos, P. (2011). *Comunity detection in Social Media.*

[8] Kroese, D. P. (2020). *Data Science and Machine Learning, Mathematical and Statistical Methods.*

[9] Natekin, A., & Knoll, A. (2013). *Gradient boosting machines, a tutorial.* Obtenido de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3885826/

[10] Elinas, P. (s.f.). *KDnuggets*. Obtenido de https://www.bing.com/search?q=kdnuggets+knowing+your+neighbours+machine+learning+on+graphs&cvid=4ad1e84e82f34c86ae27ae7c99d56477&aqs=edge..69i57.17443j0j4&FORM=ANAB01&PC=U531

[11] Tseng, G. (20 de Junio de 2018). *medium*. Obtenido de https://medium.com/@gabrieltseng/interpreting-complex-models-with-shap-values-1c187db6ec83

[12] Wu, Y., & Radewagen, R. (s.f.). *KDnuggets*. Obtenido de https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html

[13] Kuo, C. (13 de Septiembre de 2019). *Medium.* Obtenido de Medium: https://medium.com/dataman-in-ai/explain-your-model-with-the-shap-values-bc36aac4de3d

ANEXOS

ANEXO A – CONTEXTO DE LA EMPRESA

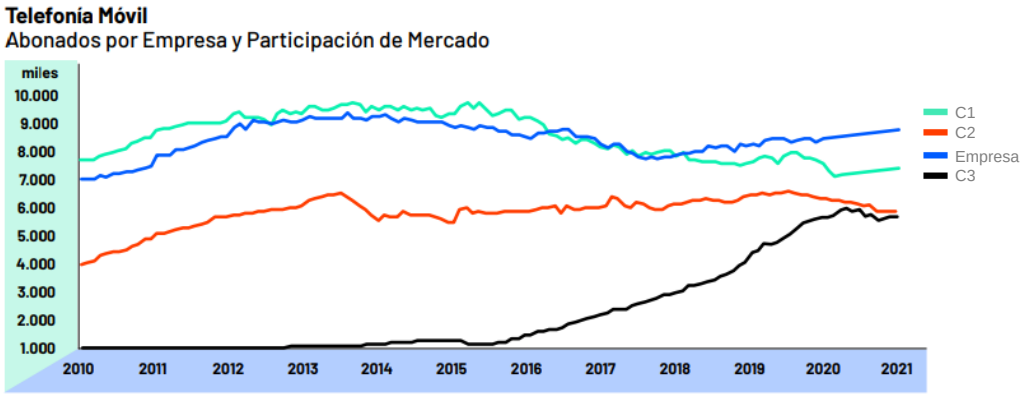


Ilustración 37: Abonados de telefonía móvil por empresa. Reporte integrado empresa.

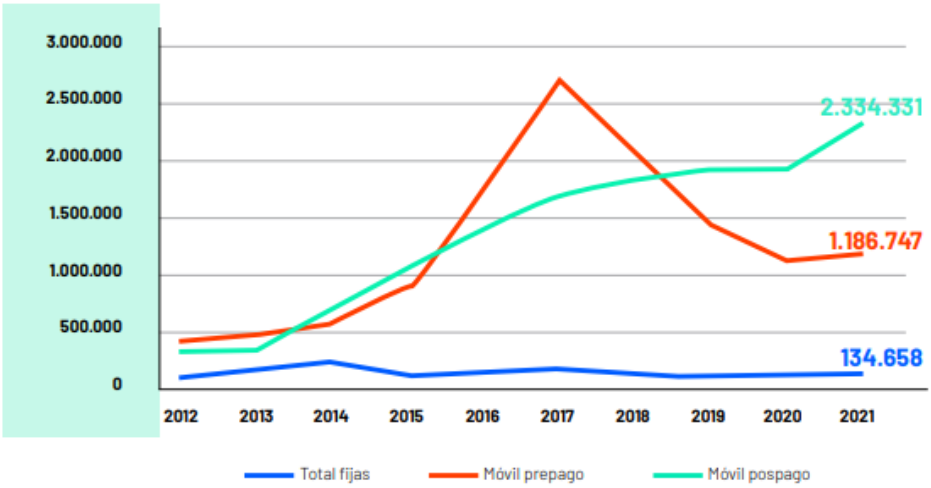


Ilustración 38: Evolución de portabilidades dentro de la empresa. Memoria integrada empresa.

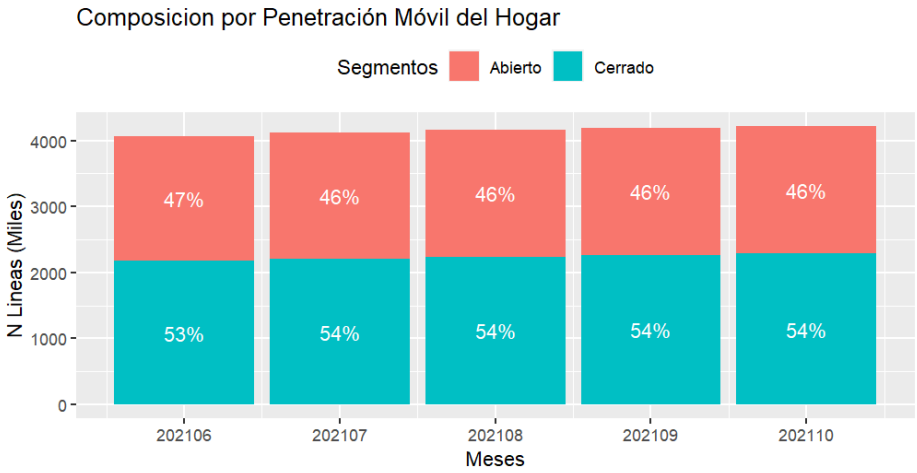


Ilustración 39: Composición de sector móvil según segmentación de hogar.

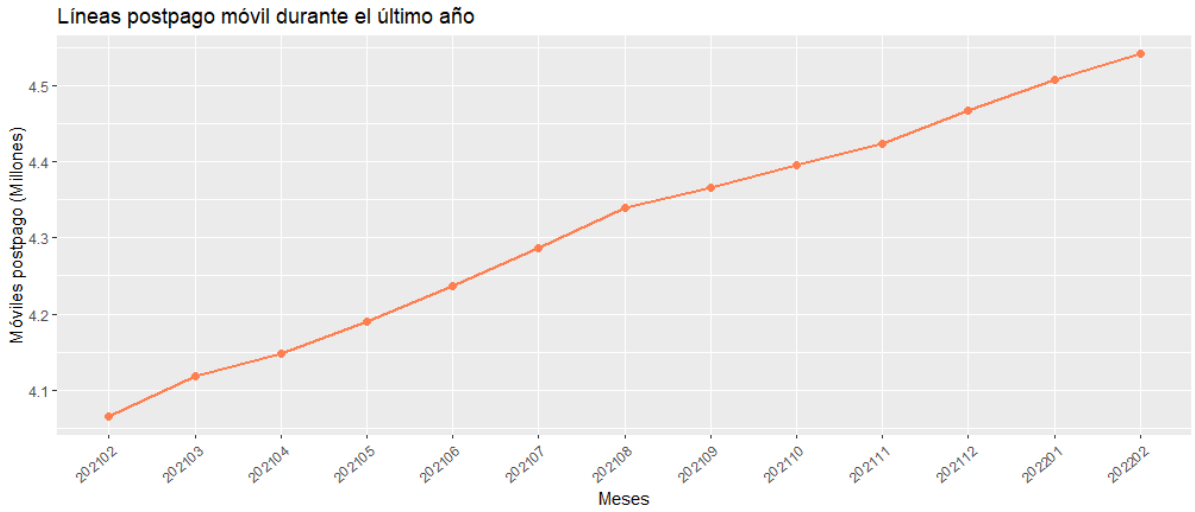


Ilustración 40: variación de clientes móviles postpago desde febrero del 2021 hasta febrero del 2022. Elaboración propia

ANEXO B – VARIABLES UTILIZADAS

VARIABLE	TIPO	DESCRIPCIÓN
id_mes	INT	Mes de desfase correspondiente a la información de postpago.
mes_campana	INT	Mes en donde se valida modelo. Se obtiene desde la habilitación comercial
id_rutcliente	STR	Rut del cliente titular
descr_movil	INT	Movil del cliente titular

portabilidad	BINARIA	Variable target. Indica si el cliente obtendrá una línea adicional portada durante el mes de campaña o no.
edad	INT	Edad del cliente titular. Construida en base a la fecha de nacimiento y una fecha intermedia entre los meses utilizados para entrenamiento y testeo.
genero	FACTOR	Género del titular de la línea postpago
estado_civil	BINARIA	Variable binaria indicando si el cliente tiene una unión matrimonial.
es_padre	BINARIA	Indica si el cliente titular tiene al menos un hijo.
cant_lineas_rut	INT	Cantidad de líneas que posee el cliente titular de un servicio móvil postpago asociado a su cuenta.
antigüedad_movil	INT	Meses transcurridos desde que el cliente titular adquirió su línea móvil (independiente de la última renovación del equipo).
origen_cliente	FACTOR	Clasificación de actividad comercial por la cual el cliente titular adquirió su línea principal (activación, migración o portabilidad)
flag_hogar	BINARIA	Variable que indica si el cliente titular posee contratación de servicios hogar (tales como televisión, internet, etc).
cargo_fijo	FACTOR	Monto cancelado mensualmente por el cliente titular respecto a su línea principal.
cant_interacciones	INT	Cantidad de interacciones del cliente con canales (reclamo, consulta, información, etc).
cant_trafico_datos	INT	Cantidad de MB traficados durante el mes por parte de la línea titular
meses_adq_equipo	INT	Meses desde la adquisición o renovación del equipo de la línea titular
monto_vas_total	INT	Monto cancelado mensualmente por el cliente respecto a servicios de valor agregado (VAS en sus siglas, tales como Spotify, Netflix, etc).
VAS	BINARIA	Indica si el cliente tiene al menos un servicio de valor agregado en su cuenta.
tasa_port_comuna	INT	Tasa de portabilidad histórica por comuna.
tasa_port_region	INT	Tasa de portabilidad histórica por región.
tasa_port_marca	INT	Tasa de portabilidad histórica por marca del movil titular.
tasa_port_plan	INT	Tasa de portabilidad histórica por id plan de postpago.
tasa_port_edad	INT	Tasa de portabilidad histórica por tramo etario.
tasa_port_cf	INT	Tasa de portabilidad histórica por tramo de cargo fijo pagado por linea titular.
tasa_port_ant	INT	Tasa de portabilidad histórica por tramo de antigüedad linea titular.
var_tiv_empresa_c1	INT	Variación porcentual del tiempo de respuesta o tiempo de ida y vuelta entre la empresa y la competencia c1

var_tiv_empresa_c2	INT	Variación porcentual del tiempo de respuesta o tiempo de ida y vuelta e ntre la empresa y la competencia c2
var_tiv_empresa_c3	INT	Variación porcentual del tiempo de respuesta o tiempo de ida y vuelta entre la empresa y la competencia c3
var_ds_empresa_c1	INT	Variación porcentual de velocidad de descarga entre la empresa y la competencia C1.
var_ds_empresa_c2	INT	Variación porcentual de velocidad de descarga entre la empresa y la competencia C2.
var_ds_empresa_c3	INT	Variación porcentual de velocidad de descarga entre la empresa y la competencia C3.
var_ngp_empresa_c1	INT	Variación porcentual con respecto al network generated percent entre la empresa y la competencia C1.
var_ngp_empresa_c2	INT	Variación porcentual con respecto al network generated percent entre la empresa y la competencia C2.
var_ngp_empresa_c3	INT	Variación porcentual con respecto al network generated percent entre la empresa y la competencia C3.
var_ss_empresa_c1	INT	Variación porcentual con respecto a la intensidad de la señal entre la empresa y la competencia C1.
var_ss_empresa_c2	INT	Variación porcentual con respecto a la intensidad de la señal entre la empresa y la competencia C2.
var_ss_empresa_c3	INT	Variación porcentual con respecto a la intensidad de la señal entre la empresa y la competencia C3.
var_dl_empresa_c1	INT	Variación porcentual con respecto a la latencia de bajada entre la empresa y la competencia C1.
var_dl_empresa_c2	INT	Variación porcentual con respecto a la latencia de bajada entre la empresa y la competencia C2.
var_dl_empresa_c3	INT	Variación porcentual con respecto a la latencia de bajada entre la empresa y la competencia C3.
var_rsrq_empresa_c1	INT	Variación porcentual con respecto a la calidad de señal de referencia entre la empresa y la competencia C1.
var_rsrq_empresa_c2	INT	Variación porcentual con respecto a la calidad de señal de referencia entre la empresa y la competencia C2
var_rsrq_empresa_c3	INT	Variación porcentual con respecto a la calidad de señal de referencia entre la empresa y la competencia C3
total_emitidos	INT	Total segundos emitidos durante el mes.
total_recibidos	INT	Total segundos recibidos durante el mes.
p_emitido_c2	INT	Porcentaje de llamadas emitidas a clientes de la competencia C2

p_emitido_c1	INT	Porcentaje de llamadas emitidas a clientes de la competencia C1
p_emitido_c3	INT	Porcentaje de llamadas emitidas a clientes de la competencia C3
p_emitido_empresa	INT	Porcentaje de llamadas emitidas a clientes de la empresa
comp_fav_emitida	FACTOR	Compañía que posee el mayor porcentaje de llamadas emitidas.
p_recibido_c2	INT	Porcentaje de llamadas recibidas desde clientes de la competencia C2
p_recibido_c1	INT	Porcentaje de llamadas recibidas desde clientes de la competencia C1
p_recibido_c3	INT	Porcentaje de llamadas recibidas desde clientes de la competencia C3
p_recibido_empresa	INT	Porcentaje de llamadas recibidas desde clientes dentro de la empresa
comp_fav_recibida	FACTOR	Compañía que posee el mayor porcentaje de llamadas recibidas.
apps	INT	Porcentaje del tráfico multimedia que realizó en Google Play y/o iTunes.
bancos	INT	Porcentaje del tráfico multimedia que realizó en aplicaciones bancarias.
nube	INT	Porcentaje del tráfico multimedia que realizó en aplicaciones de nube Google Drive, One Drive, etc.
correo	INT	Porcentaje del tráfico multimedia que realizó en aplicaciones de correos.
delivery	INT	Porcentaje de tráfico multimedia que realizo en aplicaciones de delivery.
free_stream	INT	Porcentaje del tráfico multimedia que realizó en aplicaciones de Streaming gratis.
info	INT	Porcentaje del tráfico multimedia que realizó en informaciones.
juegos	INT	Porcentaje del tráfico multimedia que realizó en juegos.
laboral	INT	Porcentaje del tráfico multimedia que realizó en aplicaciones laborales
mapas	INT	Porcentaje del tráfico multimedia que realizó en mapas-
música	INT	Porcentaje del tráfico multimedia que realizó en aplicaciones de música.
diarios	INT	Porcentaje del tráfico multimedia que realizó en diarios.
ofertas	INT	Porcentaje del tráfico multimedia que realizó en aplicaciones que entregan ofertas.
retail	INT	Porcentaje del tráfico multimedia que realizó en aplicaciones de retail.
rrss	INT	Porcentaje del tráfico multimedia que realizó en RRSS
deportes	INT	Porcentaje del tráfico multimedia que realizó en aplicaciones de deportes.

series	INT	Porcentaje del tráfico multimedia que realizó en aplicaciones de Streaming de series.
telco	INT	Porcentaje del tráfico multimedia que realizó en aplicaciones de empresas de telecomunicaciones.
transporte	INT	Porcentaje del tráfico multimedia que realizó en aplicaciones de transporte.
transporte_corto	INT	Porcentaje del tráfico multimedia que realizó en aplicaciones de transporte corto.
video	INT	Porcentaje del tráfico multimedia que realizó en aplicaciones de video.
otras_apps	INT	Porcentaje del tráfico multimedia que realizó en otras aplicaciones
trafico_total_mb	INT	Tráfico total en MB del cliente titular durante el mes.
var_download_empresa_c1	INT	Variación porcentual de velocidad de descarga, Empresa – competencia C1
var_download_empresa_c3	INT	Variación porcentual de velocidad de descarga, Empresa – competencia C3
var_download_empresa_c2	INT	Variación porcentual de velocidad de descarga, Empresa – competencia C2
var_upload_empresa_c1	INT	Variación porcentual de velocidad de subida, Empresa – competencia C1
var_upload_empresa_c3	INT	Variación porcentual de velocidad de subida, Empresa – competencia C3
var_upload_empresa_c2	INT	Variación porcentual de velocidad de subida, Empresa – competencia C2
var_latency_empresa_c1	INT	Variación porcentual de latencia, Empresa – competencia C1 (latencia: tiempo que tarda en transmitirse un paquete dentro de la red).
var_latency_empresa_c3	INT	Variación porcentual de latencia, Empresa – competencia C3
var_latency_empresa_c2	INT	Variación porcentual de latencia, Empresa – competencia C2
total_cf	INT	Cargo fijo total pagado por el grupo familiar.
cf_promedio	INT	Cargo fijo promedio pagado por el grupo familiar.
n_titulares	INT	Titulares dentro del grupo familiar.
n_miembros	INT	Integrantes dentro del grupo familiar.
Menor_9	INT	Miembros menores a 9 años del grupo familiar del cliente titular.
Entre_9_14	INT	Miembros entre 9 y 14 años pertenecientes al grupo familiar del cliente titular.
Entre_15_18	INT	Miembros entre 15 y 18 años pertenecientes al grupo familiar del cliente titular.
Entre_19_24	INT	Miembros entre 19 y 24 años pertenecientes al grupo familiar del cliente titular.
Entre_25_35	INT	Miembros entre 25 y 35 años pertenecientes al grupo familiar del cliente titular.
Entre_36_45	INT	Miembros entre 36 y 44 años pertenecientes al grupo familiar del cliente titular.

Entre_45_60	INT	Miembros entre 45 y 60 años pertenecientes al grupo familiar del cliente titular.
Mayor_60	INT	Miembros entre mayor a 60 años pertenecientes al grupo familiar del cliente titular.
edad_n_i	INT	Miembros del grupo familiar del cliente titular que no pudieron ser identificados según su edad debido a que no se tiene su fecha de nacimiento.
promedio_edad	INT	Promedio etario del grupo familiar.

ANEXO C – ANÁLISIS EXPLORATORIO

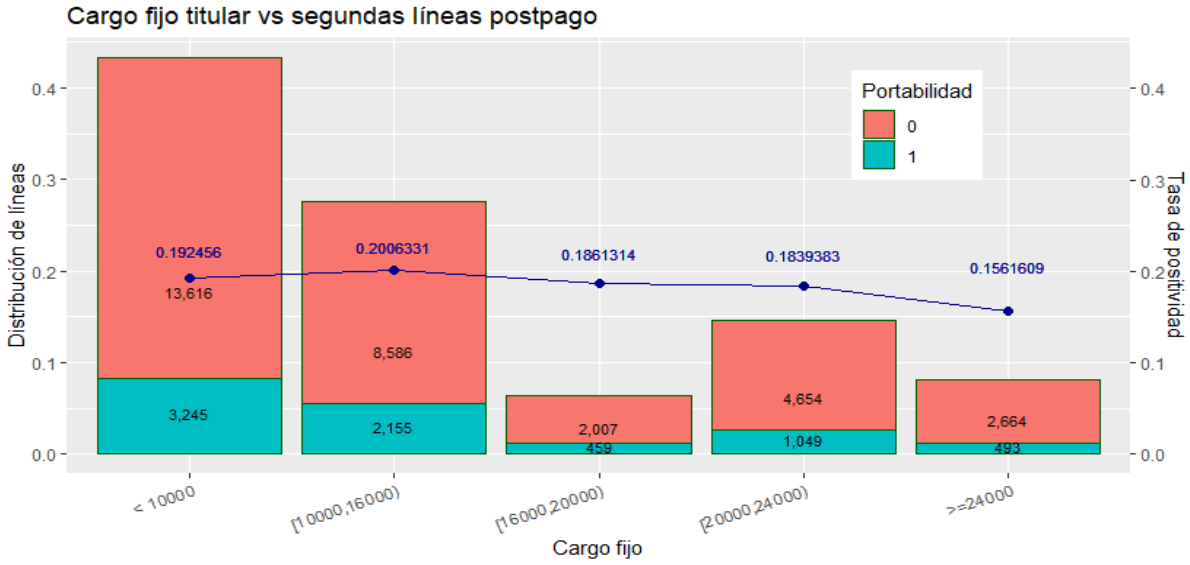


Ilustración 41: Distribución de líneas adicionales según el cargo fijo de la línea titular del cliente. Elaboración propia.

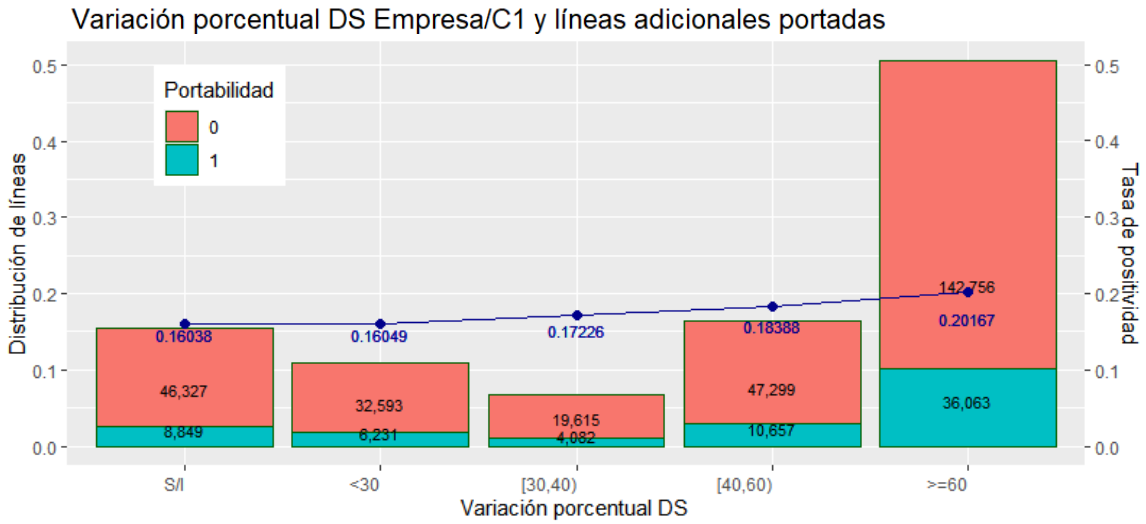


Ilustración 42: Tasa de portabilidad de líneas adicionales y distribución de líneas según variación porcentual DS Empresa/C1.

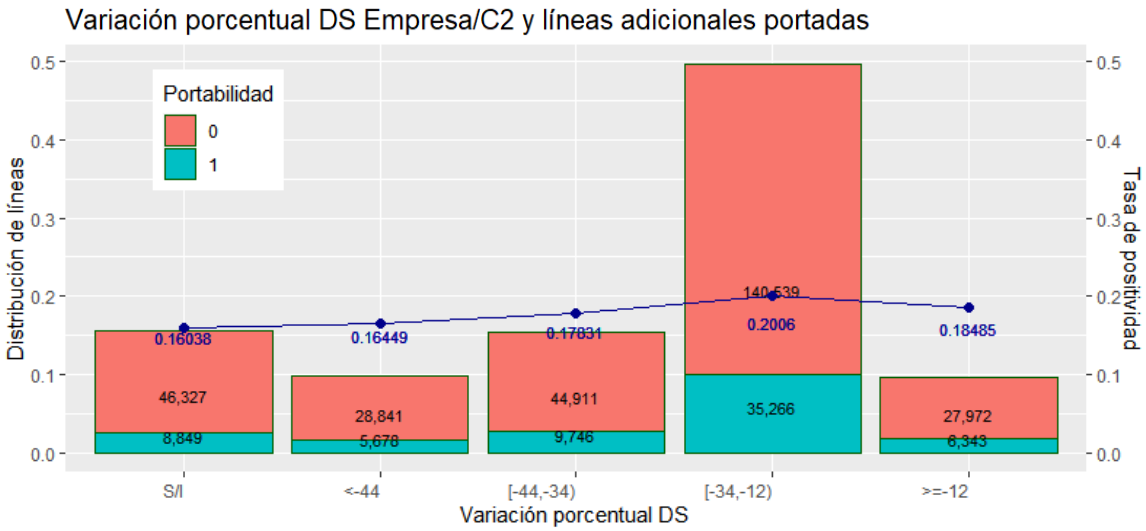


Ilustración 43: Tasa de portabilidad de líneas adicionales y distribución de líneas según variación porcentual DS Empresa/C2.

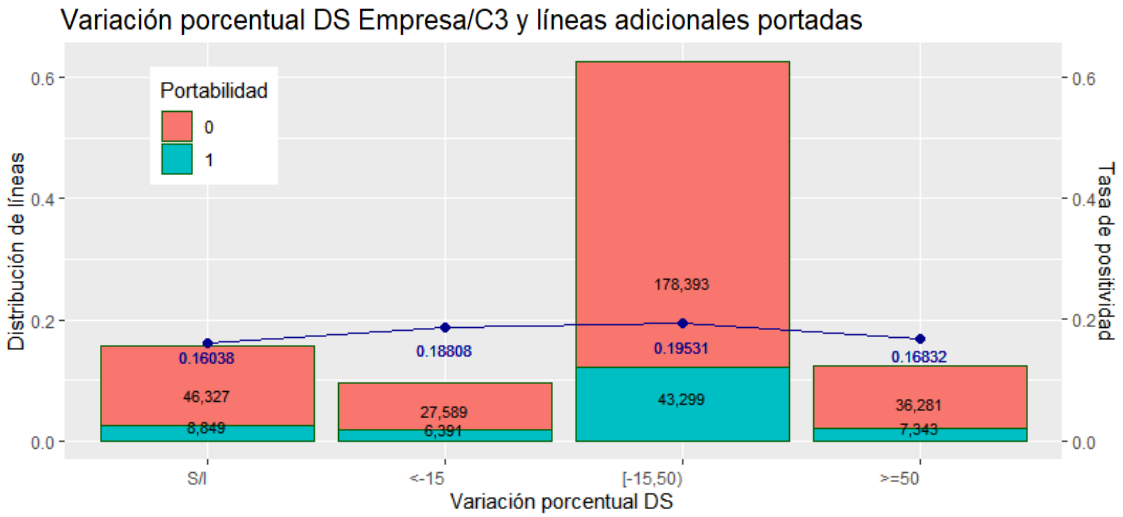


Ilustración 44: Tasa de portabilidad de líneas adicionales y distribución de líneas según variación porcentual DS Empresa/C3.

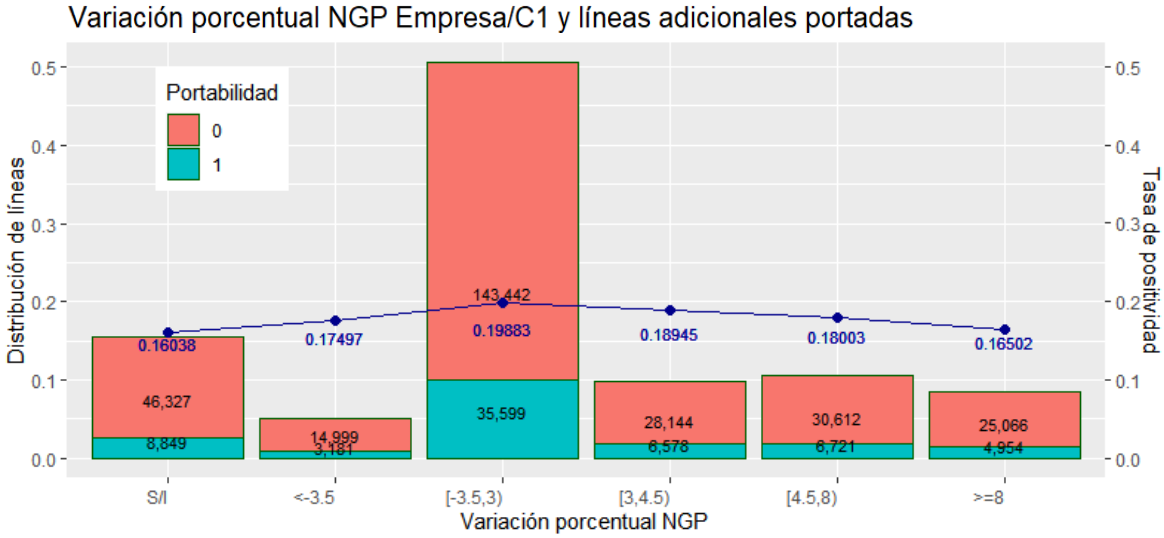


Ilustración 45: Tasa de portabilidad de líneas adicionales y distribución de líneas según variación porcentual NGP Empresa/C1.

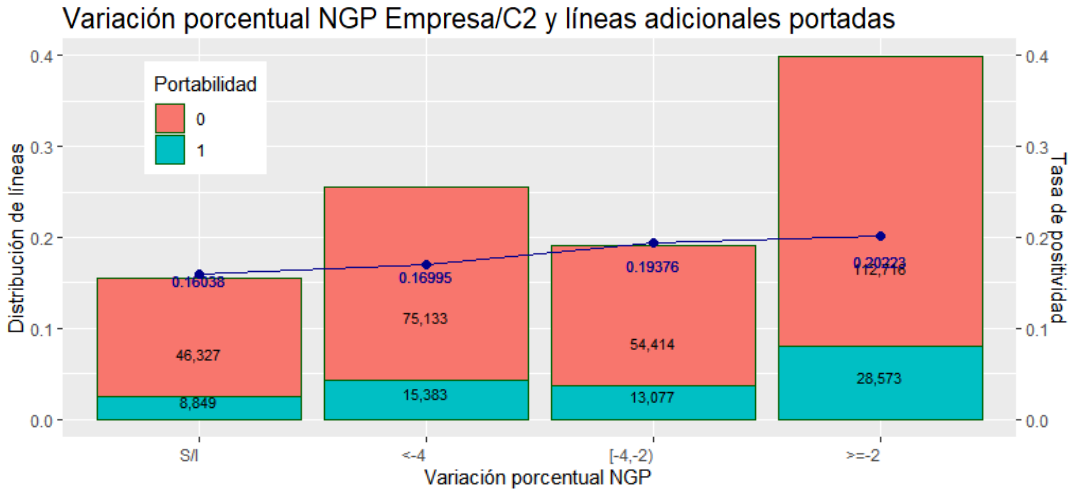


Ilustración 46: Tasa de portabilidad de líneas adicionales y distribución de líneas según variación porcentual NGP Empresa/C2.

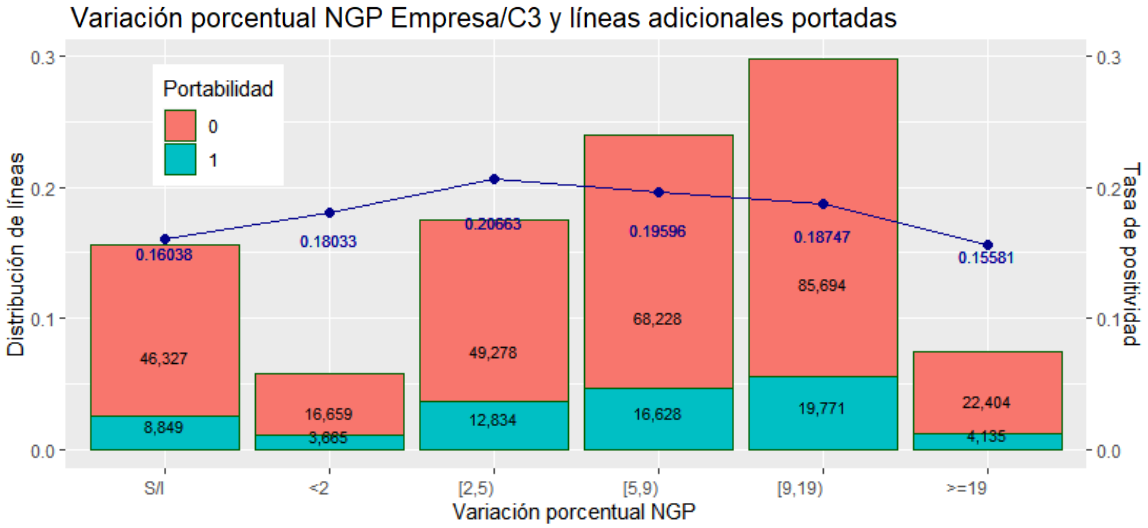


Ilustración 47: Tasa de portabilidad de líneas adicionales y distribución de líneas según variación porcentual NGP Empresa/C3.

ANEXO D – MATRICES DE CONFUSIÓN MODELOS

		Predicted	
		0	**1**
Actual	**0**	157.629	71.067
	1	22.351	29.477

Tabla 16: Matriz de confusión Random Forest (entrenamiento) sin balanceo de datos.

Accuracy	Precision	Sensitivity	Specificity
0,667	0,293	0,569	0,689

Tabla 17:Métricas de evaluación desde matriz de confusión Random Forest (entrenamiento) sin balanceo de datos

		Predicted	
		0	**1**
Actual	**0**	154.300	74.396
	1	21.653	30.175

Tabla 18: Matriz de confusión Random Forest con balanceo (entrenamiento)

Accuracy	Precision	Sensitivity	Specificity
0,658	0,289	0,582	0,675

Tabla 19: Métricas de evaluación desde matriz de confusión Random Forest (entrenamiento)

		Predicted	
		0	**1**
Actual	**0**	35.400	21.786
	1	4.939	8.057

Tabla 20: Matriz de confusión Random Forest (testeo)

Accuracy	Precision	Sensitivity	Specificity
0,619	0,270	0,620	0,619

Tabla 21: Métricas de evaluación desde matriz de confusión Random Forest (testeo)

		Predicted	
		0	**1**
Actual	**0**	172.448	56.248
	1	22.670	29.158

Tabla 22: Matriz de confusión GBM (entrenamiento) sin balanceo de datos.

Accuracy	Precision	Sensitivity	Specificity
0.719	0.341	0.563	0.754

Tabla 23: Métricas de evaluación desde matriz de confusión GBM (entrenamiento) sin balanceo de datos

		Predicted	
		0	**1**
Actual	**0**	165.013	63.683
	1	21.188	30.640

Tabla 24: Matriz de confusión GBM con balanceo (entrenamiento)

Accuracy	Precision	Sensitivity	Specificity
0.697	0.325	0.591	0.722

Tabla 25: Métricas de evaluación desde matriz de confusión GBM (entrenamiento)

		Predicted	
Actual		**0**	**1**
	0	40.989	16.197
	1	5.846	7.150

Tabla 26: Matriz de confusión GBM con balanceo (testeo)

Accuracy	Precision	Sensitivity	Specificity
0.686	0.306	0.550	0.717

Tabla 27: Métricas de evaluación desde matriz de confusión GBM con balanceo (testeo)

		Predicted	
Actual		**0**	**1**
	0	162.419	66.277
	1	22.997	28.831

Tabla 28: Matriz de confusión XGBoost sin balanceo de datos (entrenamiento)

Accuracy	Precision	Sensitivity	Specificity
0.682	0.303	0.556	0.710

Tabla 29: Métricas de evaluación desde matriz de confusión XGBoost sin balanceo de datos(entrenamiento)

Actual		**0**	**1**
	0	159.757	68.939
	1	22.031	29.797

Tabla 30:Matriz de confusión XGBoost con balanceo (entrenamiento)

Accuracy	Precision	Sensitivity	Specificity
0,676	0,302	0,575	0,699

Tabla 31:Métricas de evaluación desde matriz de confusión XGBoost con balanceo (entrenamiento)

		Predicted	
Actual		**0**	**1**
	0	42.034	15.152
	1	6.135	6.861

Tabla 32:Matriz de confusión XGBoost con balanceo (testeo)

Accuracy	Precision	Sensitivity	Specificity
0.697	0.312	0.528	0.735

Tabla 33: Métricas de evaluación desde matriz de confusión XGBoost con balanceo (testeo)

ANEXO E – EVALUACIÓN MODELOS

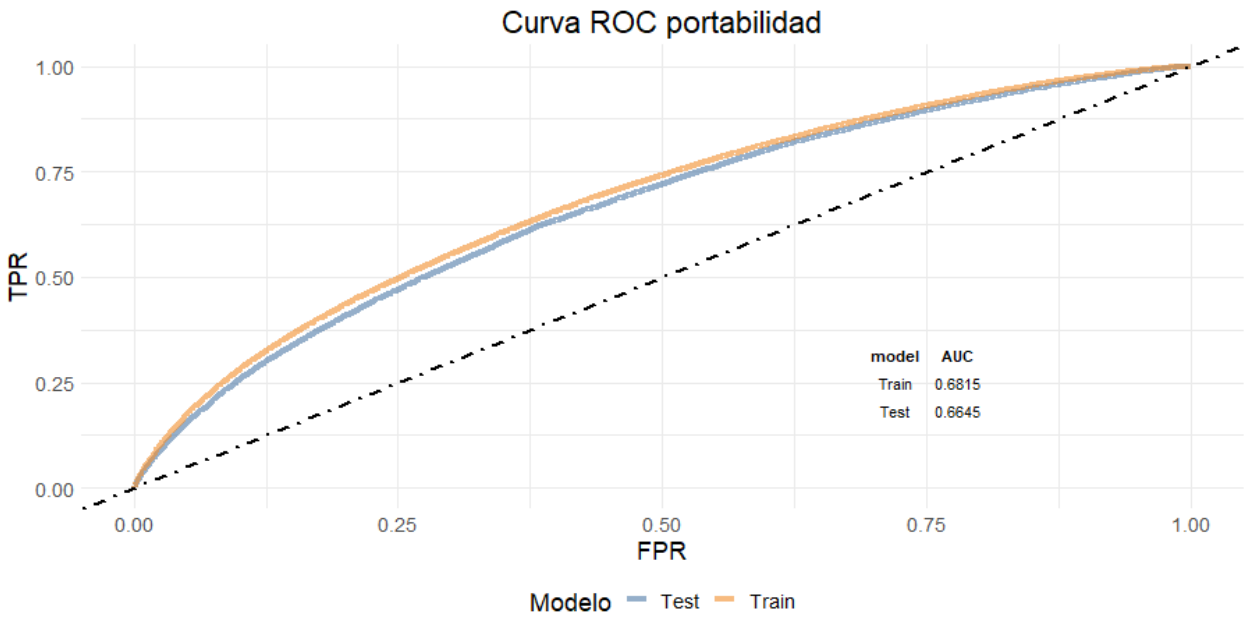


Ilustración 48: Curva ROC modelo RF con balanceo de datos

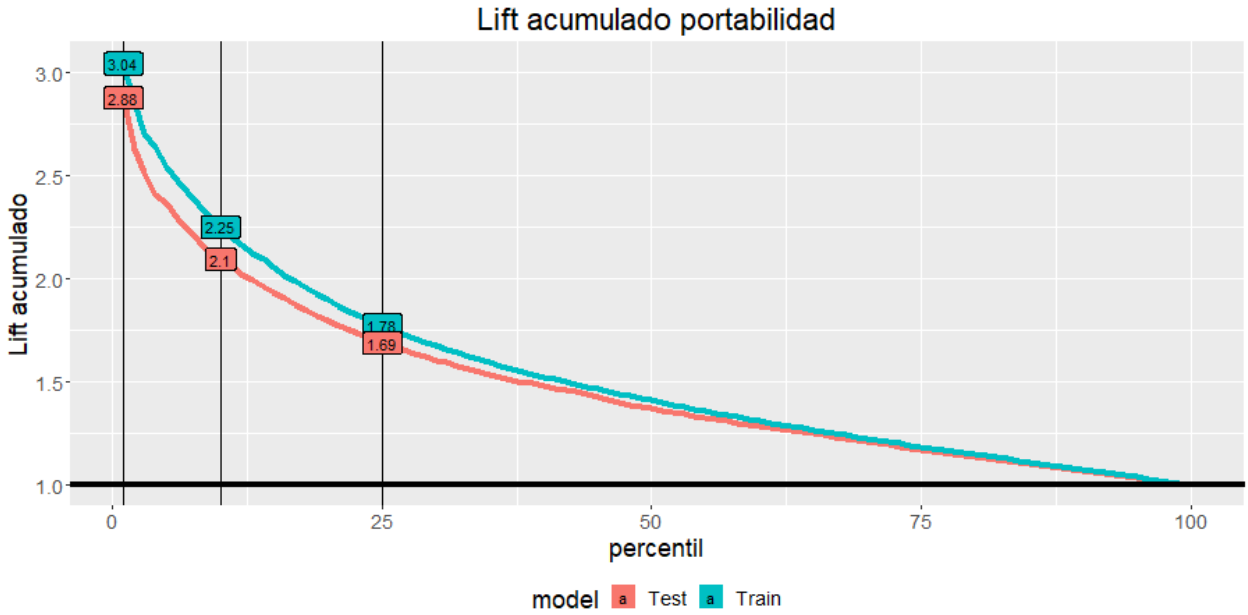


Ilustración 49: Curva Lift modelo RF con balanceo de datos

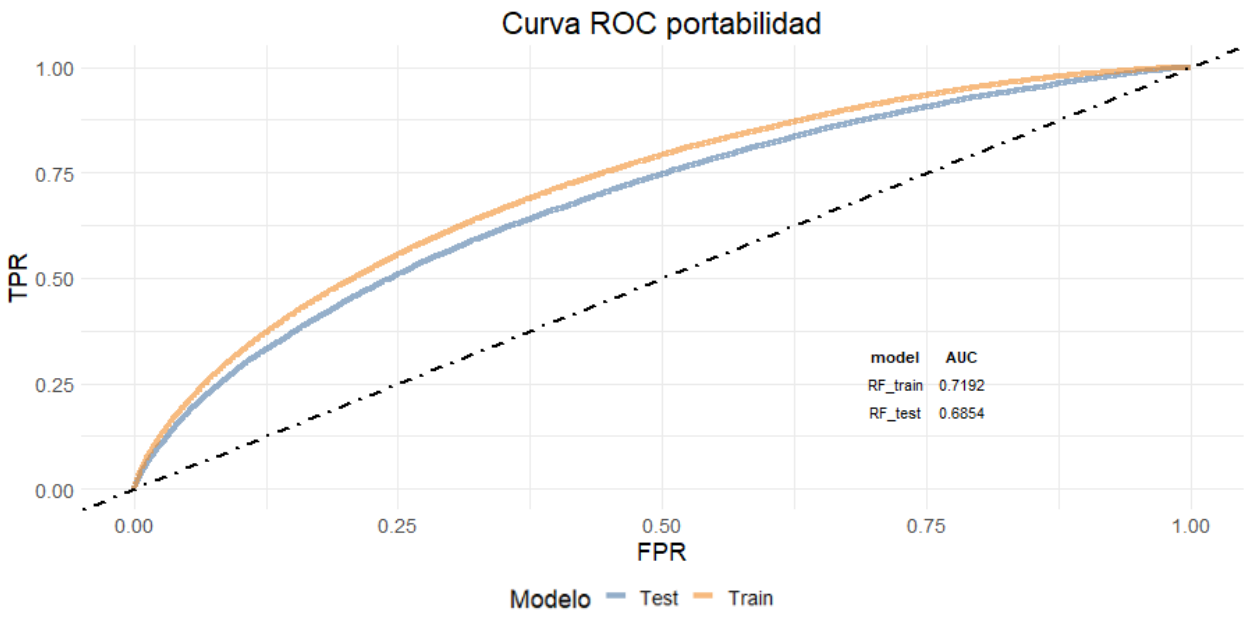


Ilustración 50: Curva ROC modelo GBM con balanceo de datos

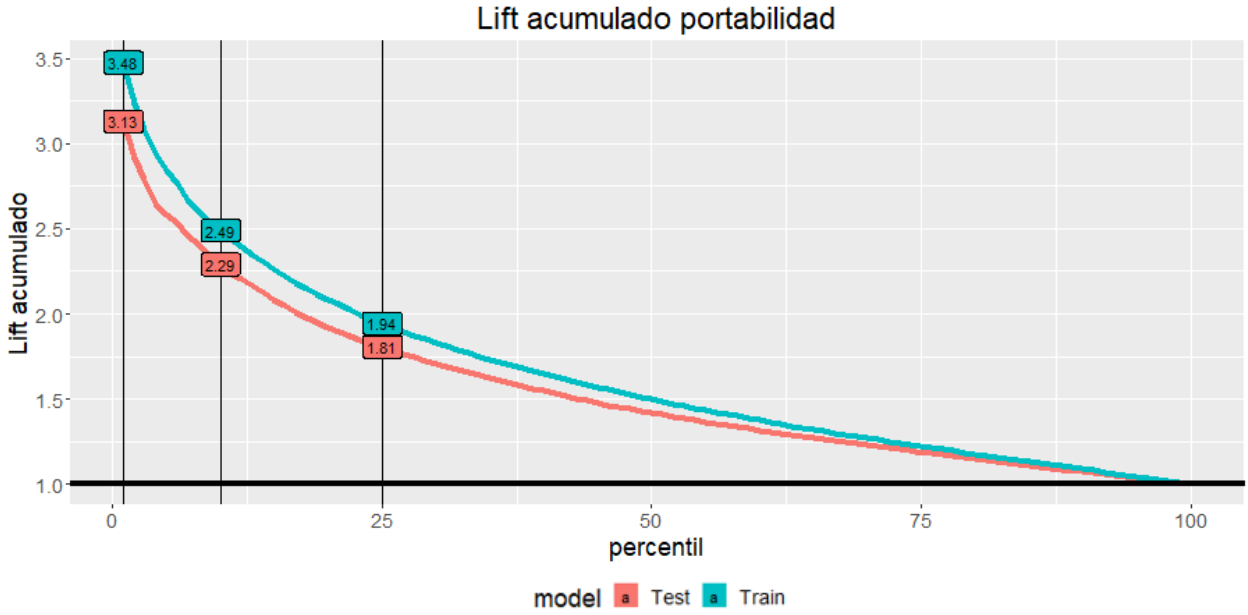


Ilustración 51: Curva Lift modelo GBM con balanceo de datos

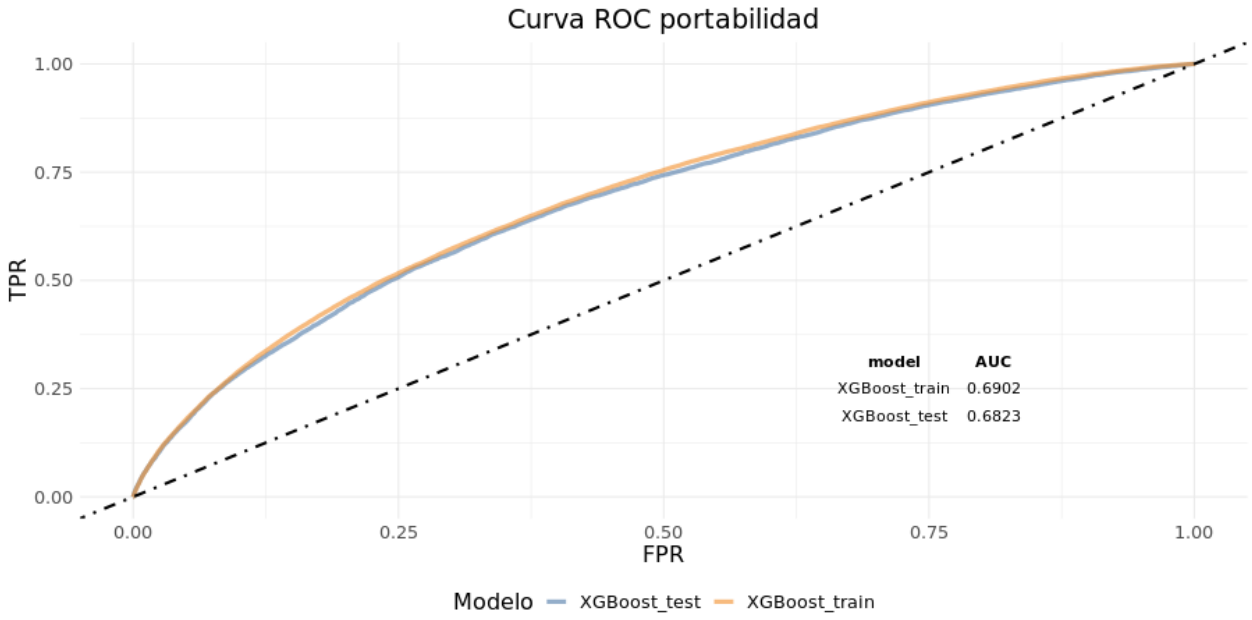


Ilustración 52: Curva ROC modelo XGBoost con balanceo de datos

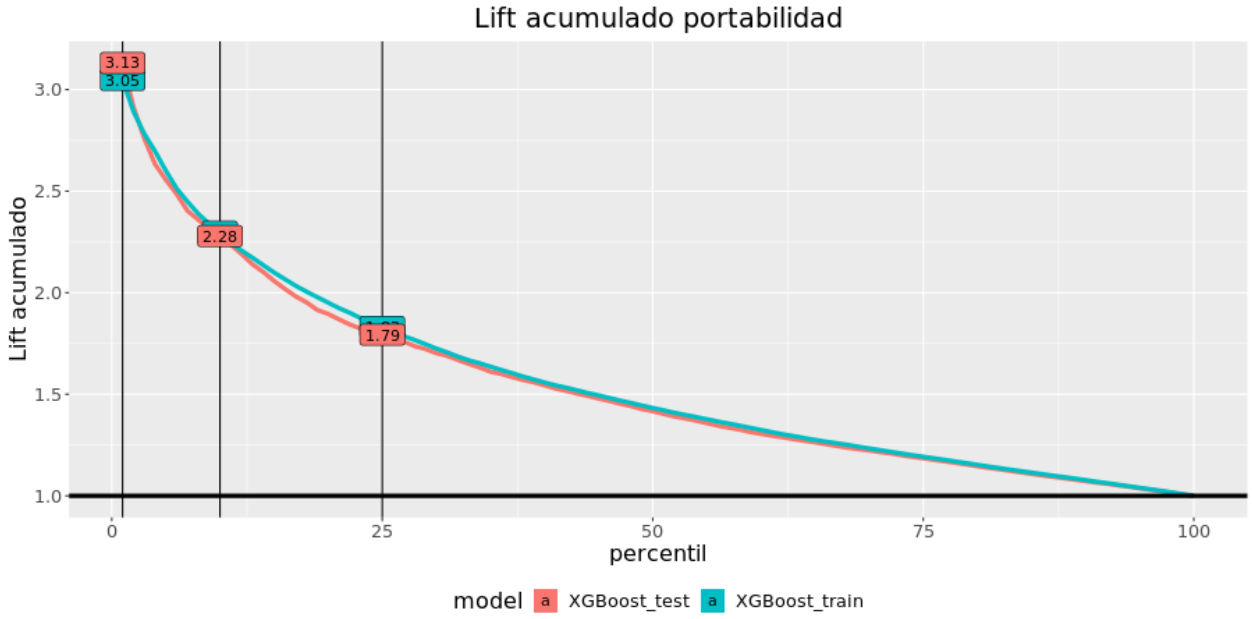


Ilustración 53: Curva Lift modelo XGBoost con balanceo de datos

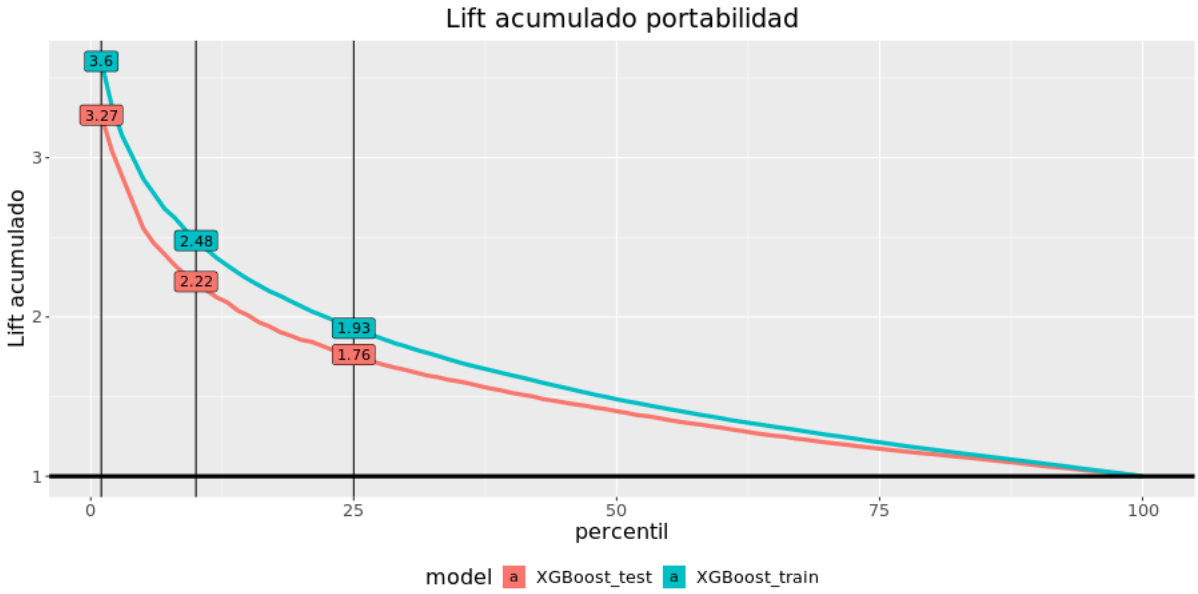


Ilustración 54: Curva Lift XGBoost entrenado y testeado con 4 meses de historia (sin variables del CDR).

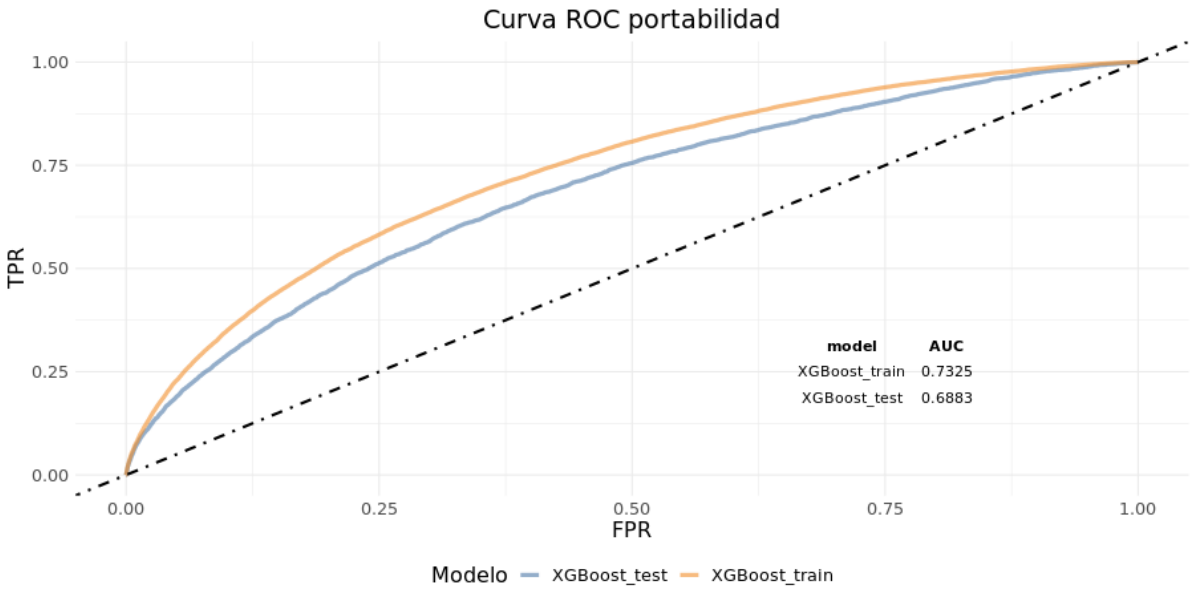


Ilustración 55: Curva Lift modelo XGBoost entrenado y testeado con 4 meses de historia (con variables de CDR).

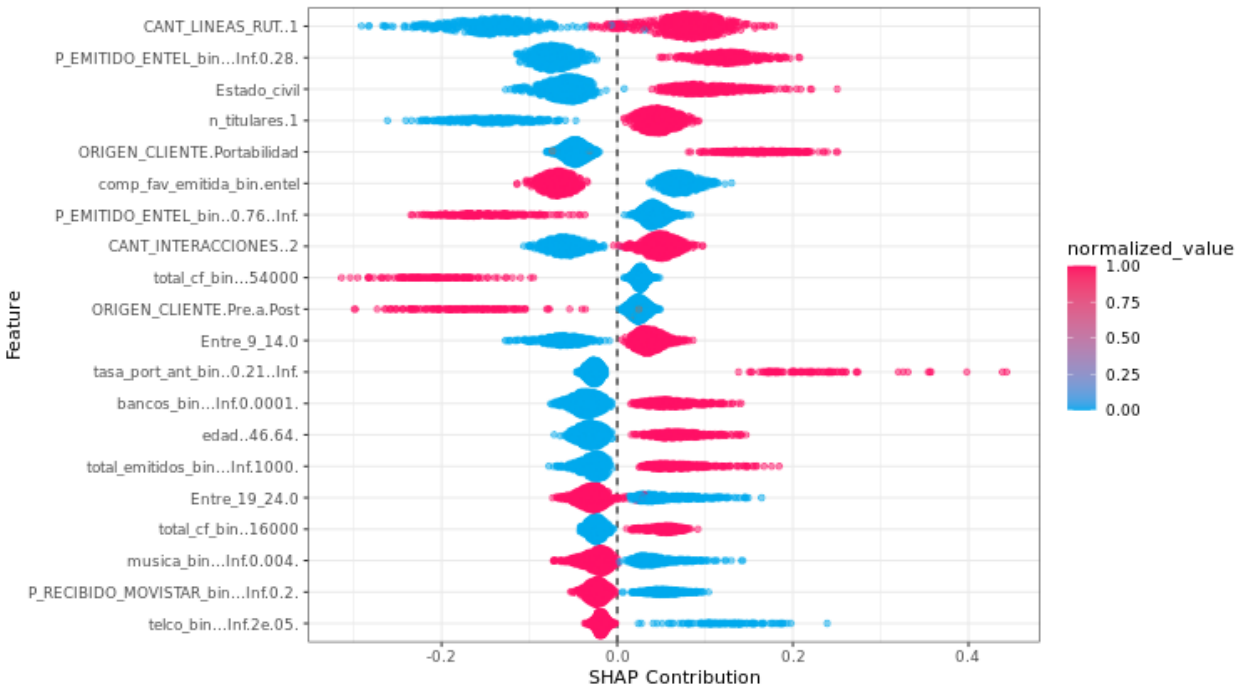


Ilustración 56: Importancia de variables, método Shapley Values.