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

MODELO DE PREDICCIÓN DE FUGA DE CLIENTES EN EMPRESA SAAS DE INTELIGENCIA LOGÍSTICA

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

CARLOS ALEN GALLARDO BURNS

PROFESORA GUÍA:
ALEJANDRA PUENTE CHANDÍA

PROFESOR CO-GUÍA:
PABLO MARÍN VICUÑA

COMISIÓN:
RODRIGO ASSAR CUEVAS

SANTIAGO DE CHILE
2021

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL
INDUSTRIAL
POR: CARLOS ALEN GALLARDO BURNS
FECHA: 2021
PROF. GUÍA: ALEJANDRA PUENTE CHANDÍA

MODELO DE PREDICCIÓN DE FUGA DE CLIENTES EN EMPRESA SAAS DE INTELIGENCIA LOGÍSTICA

En estos tiempos donde procesos y la forma en que nos relacionamos se está trasladando a lo digital, donde se espera que el mercado latinoamericano de tecnología de la nube crezca en un 250 % en 3 años a partir de 2021 [Vota et al., 2020], empresas SaaS enfrentarán cada día más competencia, aumentando con esto su riesgo de perder clientes. En este escenario, contar con herramientas que permitan anticiparse a una fuga se vuelve cada vez más relevante.

Esta memoria aborda la elaboración de un modelo de predicción de fuga de clientes y la interpretación de variables relevantes, en una empresa de inteligencia logística, con el objetivo de aportar con una herramienta a la organización. Esto se llevó a cabo utilizando técnicas de machine learning (XGBoost y Logit), interpretación de coeficientes de Logit y el análisis de shap values para la interpretación del efecto de variables en el modelo XGBoost. Vale destacar que para realizar este trabajo que se utilizó con una base de datos de 2372 instancias y 7 variables relevantes para la predicción.

Los resultados de este estudio indican que el modelo XGBoost tiene un mejor desempeño en la predicción de fuga para este contexto (AUC = 0.76, precisión = 0.22, sensibilidad = 0.56) en comparación con el modelo logit (AUC = 0.72, precisión = 0.04,sensibilidad = 0.76). Por otro lado, en cuanto a la interpretación de variables, el análisis de shap values permitió identificar que en los primeros 7 meses de actividad, los clientes son más propensos a la fuga, y que usuarios que registran niveles bajos en una de las variables que miden la frecuencia de uso ("n_planes_mes") tienden también a tener una probabilidad de fuga mayor. De esta última variable se destaca la capacidad de ser utilizada como clasificador simple a través de un árbol de decisión, donde clientes cuyas métricas estén bajo cierto umbral son clasificados como candidatos a la fuga, alcanzando un performance cercano al del modelo XGBoost (precisión = 0.14, sensibilidad = 0.55) con una implementación más sencilla.

En conclusión, los resultados muestran que el modelo XGBoost es el que logra la mejor predicción con los datos disponibles. Por su parte, el estudio de los shap values no solo permitió entender como cada variable influye en la predicción, sino que además contribuyó a generar preguntas como por ejemplo el por qué en los primeros 7 meses se tiene una mayor propensión a la fuga, cuya respuesta podría ayudar a generar estrategias efectivas de retención de cliente y que se propone investigar en trabajos futuros.

Para mi familia, Magda y amigos

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

1 | Antecedentes generales

1.1. Sobre el rubro de la inteligencia logística

En un mundo donde la forma de comunicarnos y de trabajar es cada vez más digital, la relevancia que tienen los datos crece día a día, conceptos como el "big data" y la inteligencia artificial están ganando protagonismo en la toma de decisiones y mejoramiento de procesos por su capacidad de mostrar lo que el ser humano no ve a simple vista, pudiendo realizar tanto predicciones acertadas como también automatizar diversos tipos de procesos.

Este cambio tecnológico, según un estudio de McKinsey & Company, tiene un especial impacto tanto en las actividades relacionadas a marketing y ventas como las de cadena de suministro y manufacturación [Chui et al., 2019]. Es en este segundo ámbito donde la inteligencia logística se posiciona, ya que utiliza la información disponible para mejorar todo tipo de procesos logísticos, como lo son la obtención de materias primas, la manufacturación y el despacho de los productos a su cliente final (la "última milla").

1.2. Sobre la empresa

Esta memoria se realiza en una empresa de inteligencia logística, dedicada específicamente a la mejorar la última milla de sus clientes a través de un servicio SaaS (Software as a Service) que resuelve la problemática de cómo organizar el despacho de productos de manera óptima.

A la problemática mencionada anteriormente se le conoce como el problema de ruteo de vehículos (VRP por sus siglas en inglés), el cual se ejemplifica en la figura 1.1, donde un cliente que posee un cierto número de vehículos repartidores (en este caso son 4) tiene que llegar a todos los puntos del mapa para entregar sus productos. Es claro ver en la imagen que existen múltiples formas de conectar a estos clientes (rutas), y si a esto se le agregan más puntos, más vehículos y la condición de que el vehículo llegue en un rango horario acordado con el cliente el problema se vuelve cada vez más complejo. Es aquí donde aplica la inteligencia logística, que en base a datos de GPS, registros históricos, modelos de optimización y de predicción es capaz de entregar una planificación óptima de rutas para el cliente.

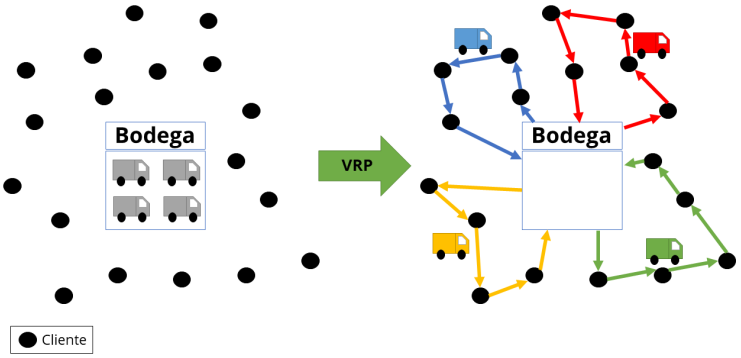

Figura 1.1: Diagrama VRP [Elaboración propia]

Entre los beneficios que un cliente percibe por esta solución y los servicios complementarios se destacan los siguientes:

- Ahorro de combustible por la disminución de kilómetros recorridos.
- Poder utilizar menos vehículos para realizar las mismas entregas.
- Ahorrar horas en planificación de rutas.
- Mayor satisfacción de clientes finales al poder ofrecer ventanas horarias de despacho más acotadas (Ej: el producto llegará entre las 10:00 y las 12:00 hrs.).
- Seguimiento en vivo del despacho para clientes finales.

Este servicio opera con una suscripción mensual por número de vehículos a utilizar, lo que ha permitido llamar la atención de empresas pequeñas como emprendimientos que buscan un aliado para poder escalar su negocio a un precio asequible. Además, el alto estándar de la solución permite disminuir los costos de todo tipo de empresas, donde se destacan sectores como el retail o supermercados, cuyo *core business* esta en otro lugar y encuentran un gran valor en externalizar esta tarea.

Vale destacar que los clientes no solo provienen del entorno local, sino que se distribuyen por gran parte de Latinoamérica, con un gran crecimiento extranjero en los últimos años, debido a un plan de expansión regional acompañado de la apertura de oficinas en 5 países.

2 | Descripción general del proyecto

2.1. La problemática y su justificación

Como se puede ver en la figura 2.1, la empresa ha tenido un buen desempeño en la captación de clientes con un crecimiento continuo a través del tiempo y con un nivel de fuga relativamente constante. Dentro de la historia, se destaca especialmente un hito marcado con la linea punteada roja, correspondiente al 18 de marzo de 2020, día en que la OMS caracterizó al COVID-19 como una pandemia [OMS, 2020], momento en que se implementaron cuarentenas y se forzó a potenciar el canal de delivery en el comercio, con lo que creció la demanda de este servicio a niveles nunca antes vistos.

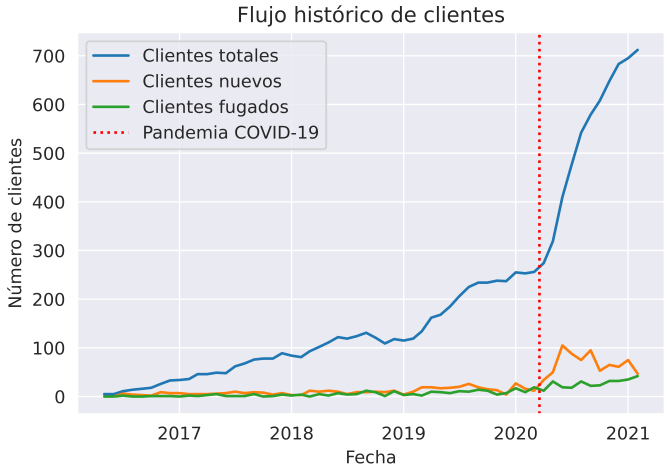

Figura 2.1: Flujo histórico de clientes [Elaboración propia]

En una revisión más profunda de los clientes fugados, en la figura 2.2 se observa que la tasa de fuga mensual tiene una tendencia al alza. De esto, si se complementa con lo exhibido por la figura 2.1, se puede concluir que cada mes se vuelve más relevante el desarrollar herramientas y políticas que tengan como objetivo la prevención de la fuga de clientes, puesto que, tanto el universo de clientes como la tasa de fuga están creciendo.

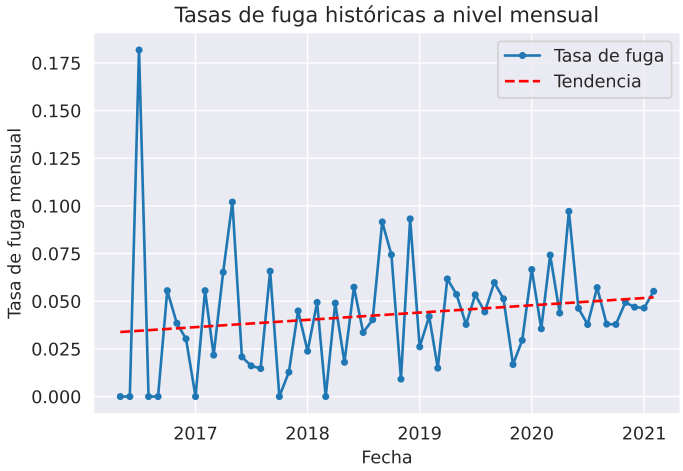

Figura 2.2: Tasas mensuales de fuga históricas [Elaboración propia]

Otro antecedente relevante es la manera en la que se trabajan las necesidades insatisfechas de los clientes. Actualmente es el cliente quien levanta sus necesidades, lo que deriva en que la empresa realice desarrollos para entregar un mejor servicio. Este enfoque es del tipo reactivo, y hace surgir preguntas como ¿Que está pasando con los que se fugan sin haber declarado que se encontraban disconformes? ¿Se pudo haber hecho algo para retenerlos?.

Como último antecedente, otra situación que aparece con el aumento de clientes es el aumento del ratio entre clientes y trabajadores que dan soporte, lo que se traduce en que un trabajador cada mes tiene que estar alerta de más empresas, incluso habiendo expandido al equipo. En este escenario, herramientas que permitan dar luces sobre dónde se debe focalizar el esfuerzo se vuelven aliadas importantes.

Por lo mencionado anteriormente, la elaboración de un modelo de predicción de fuga de clientes es una buena alternativa de solución, que de manera preventiva será capaz de identificar posibles fugas, entregando focos donde dirigir los esfuerzos del equipo en pos de mejorar la retención.

2.2. Objetivo del trabajo

2.2.1. Objetivo general

Elaborar un modelo predictivo para estimar la fuga de clientes que permita levantar alertas tempranas para poder atender necesidades insatisfechas de usuarios.

2.2.2. Objetivos específicos

- Construir, entrenar y validar modelos de predicción de fuga.
- Evaluar y escoger entre los modelos construidos, en base a métricas de desempeño.
- Medir el impacto de las variables utilizadas en la deserción de los clientes, en base a análisis estadístico.

- Recomendar a la empresa sobre que tipo de clientes deberían enfocar sus acciones de fidelización.

2.3. Alcances

En este trabajo se utilizarán solo los datos de clientes chilenos, puesto que son los únicos que tienen información complementaria completa sobre el tamaño, rubro, tipo de cliente entre otros campos. Además, estos clientes constituyen más del 50 % de la data (de granularidad cliente-mes) y son los que poseen más períodos de historia para modelar.

Debido a la completitud de variables importantes a estudiar, el horizonte temporal a estudiar es desde octubre de 2017 hasta diciembre de 2020.

3 | Marco teórico

3.1. Fuga de clientes

Para este trabajo, la fuga de un cliente corresponde a la acción del usuario de no renovar la suscripción al servicio ofrecido por la empresa para el mes posterior.

3.2. Modelos de predicción

3.2.1. Regresión logística

La regresión logística es un modelo aplicado en múltiples estudios para la predicción de probabilidades y de variables categóricas.

En esta, la probabilidad de que un evento suceda esta dada por las ecuaciones 3.1 y 3.2, donde x_i representa las variables explicativas y β_i sus coeficientes asociados respectivamente.

$$P(y = 1|x) = \sigma(z) = \frac{1}{(1 + \exp^{-z})} \tag{3.1}$$

Donde,

$$z = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \tag{3.2}$$

Para entrenar el modelo se procede a ajustar los coeficientes β de manera de maximizar el logaritmo de la función de verosimilitud (ecuación 3.3).

$$LL(\beta) = \sum_{j=1}^{m} y^{(j)} log(\sigma(z^{(j)})) + (1 - y^{(j)}) log(1 - \sigma(z^{(j)})) \tag{3.3}$$

La ventaja de utilizar este modelo en este tipo de problemas radica en la simplicidad en la interpretación de la influencia de las variables (a través de los coeficientes β) en la probabilidad de fuga de los clientes en general.

Como desventaja, hay que considerar que esta herramienta no puede de detectar efectos no lineales de las variables independientes sobre la dependiente.

3.2.2. Árbol de decisión

Modelo de predicción basado en la categorización por medio de reglas. La composición de un árbol de decisión se puede ver de manera gráfica en la figura 3.1, donde el primer cuadrado corresponde a un nodo donde se aplica una regla a un elemento a clasificar (en este caso si una variable es mayor o no a 10), luego, dependiendo de si se cumple o no la regla, las flechas dirigen al elemento a la clasificación correspondiente o a una nueva regla de clasificación.

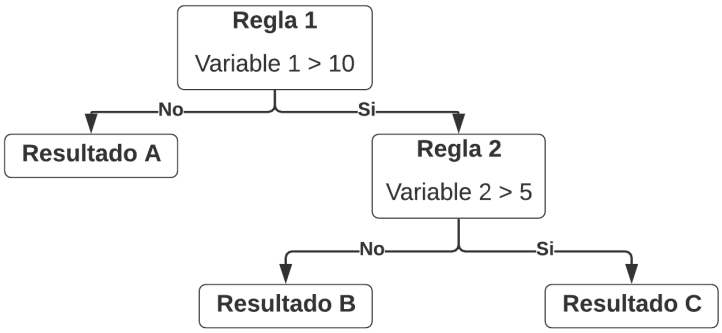

Figura 3.1: Diagrama de un árbol de decisión [Elaboración propia]

La fortaleza de este modelo radica en la facilidad para representar clasificaciones, las que se pueden complejizar añadiendo múltiples variables y nodos.

3.2.3. XGBoost

Extreme Gradient Boosting [Chen and Guestrin, 2016] es un algoritmo de machine learning basado en árboles, que utiliza la técnica gradient boosting para entrenar a través de una secuencia de modelos que aprenden de su predecesor. Para esto, se utiliza una función de pérdida $L(y, F(X))$ donde y es la variable dependiente y $F(X)$ el valor predicho por el modelo. Con esto, el valor de la predicción del modelo en cada iteración esta dado por la ecuación 3.4, donde $h_m(x)$ es el gradiente de la función de pérdida con respecto a $F(X)$ y γ es un multiplicador que se obtiene a partir de la ecuación 3.5

$$F_m(X) = F_{m-1}(X) + \gamma_m h_m(x)) \tag{3.4}$$

$$\gamma_m = argmin_\gamma \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) \tag{3.5}$$

Adicional a lo anterior, XGBoost añade un termino adicional a la función de pérdida, que mejora el desempeño del modelo en cuanto a sobreajuste y requerimiento computacional.

Junto con presentar un buen desempeño tanto en resultados como en uso de recursos, XGBoost cuenta con la ventaja de poder manejar los valores faltantes, lo que es un aporte para el trabajo de esta memoria.

3.3. Selección de hiperparámeteros

3.3.1. Grid search

Técnica de selección de hiper-parámetros basado en la confección de una grilla con múltiples configuraciones las cuales son testeadas con el conjunto de datos de entrenamiento y comparadas con el conjunto de datos de validación según una métrica de desempeño, para así poder encontrar la configuración óptima que entregue el mejor modelo.

3.4. Interpretación de variables

3.4.1. Shap analysis

Técnica de análisis de la influencia de variables independientes en la predicción de un modelo de machine learning [Lundberg and Lee, 2017]. Esta técnica está basada en teoría de juegos, donde a partir de la comparación de distintas configuraciones de variables, se calcula el aporte marginal de cada una por separado en la predicción. El resultado de la aplicación de esta técnica permite estudiar posibles tendencias en el efecto de una variable, las interacciones y la influencia general que tienen sobre el output de un modelo.

3.5. Métricas de desempeño

3.5.1. Matriz de confusión

La matriz de confusión (figura 3.2) es una herramienta de visualización del resultado de un modelo de clasificación. En esta, se presenta, como resumen, el conteo de casos donde: el modelo y el valor real son positivos (True Positive), el modelo y el valor real son negativos (True Negative), el modelo clasifica como negativo cuando realmente era positivo (False Negative) y cuando el modelo clasifica como positivo cuando realmente era negativo (False positive).

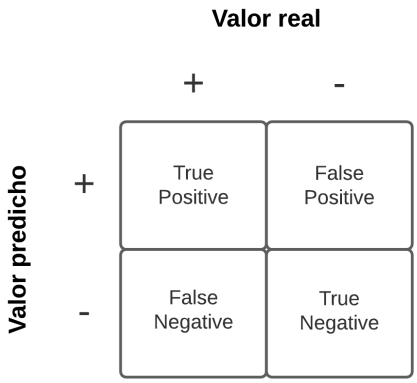

Figura 3.2: Matriz de confusión [Elaboración propia]

3.5.2. Sensibility

Métrica de desempeño que mide la fracción de casos positivos detectados (con respecto al total de positivos). Este indicador se calcula con la ecuación 3.6.

$$Sensibility = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{3.6}$$

3.5.3. Specificity

Métrica de desempeño que mide la fracción de positivos negativos (con respecto al total de negativos). Este indicador se calcula con la ecuación 3.6.

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \tag{3.7}$$

3.5.4. Precision

Métrica de desempeño que mide la fracción de clasificaciones verdaderamente positivas con respecto al total de clasificaciones positivas. Este indicador se calcula con la ecuación 3.8.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{3.8}$$

3.5.5. Curva ROC

Técnica de visualización que exhibe el trade off entre sensibilidad y especificidad de un modelo de clasificación al variar el umbral de discriminación. En la figura 3.3 se muestra el ejemplo de una curva ROC, cuyo eje y (True positive rate) corresponde a la sensibilidad, y su eje x (False positive rate) corresponde a la resta de $1 - especificidad$. Además, esta curva posee una linea azul punteada que representa el desempeño de un clasificador aleatorio que funciona como benchmark. Un modelo se considera un mejor clasificador cuando su curva se separa lo más posible (hacia arriba) de la linea azul.

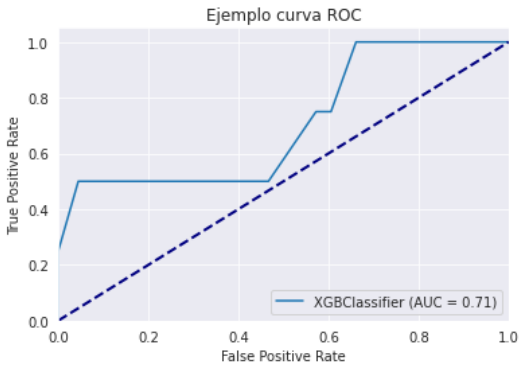

Figura 3.3: Ejemplo de curva ROC [Elaboración propia]

3.5.6. AUC (ROC_AUC)

La métrica AUC mide el área bajo la curva ROC y permite comparar un modelo con otro en todos los umbrales de clasificación. Esta métrica toma valores desde el 0 al 1, y modelos con un AUC mayor son considerados mejores clasificadores.

4 | Metodología

4.1. Entendimiento del negocio

En esta primera etapa se participó en reuniones con personas de áreas relacionadas a los clientes (Finanzas, Comercial, Customer Success) y con gente que del equipo de Data Science quienes conocen del funcionamiento interno del negocio.

En esta instancia fueron levantadas posibles causas que podrían influir en la fuga de clientes, como lo son el tener problemas de funcionamiento en la plataforma, utilizar mal el servicio, y no recibir soporte en un tiempo acotado. También se rescató que la disminución en el uso de la plataforma podría entregar información relevante para identificar una posible fuga.

4.2. Entendimiento de la data

4.2.1. Descripción general de la data

Para este estudio, la data principal a utilizar es la facturación de los clientes, donde se tiene información de cada mes referente al servicio que se está pagando y si el cliente sigue activo (o si se fugó), dentro del horizonte temporal entre agosto de 2015 a la actualidad. Esta data consta de 14.450 filas, donde cada una de estas representa el pago de un cliente en un mes en particular. En cuanto a los clientes, estos suman un total de 1.198 (con 581 fugas) localizados en más de 12 países, donde los clientes chilenos constituyen más del 50 % de la data (figura 4.1).

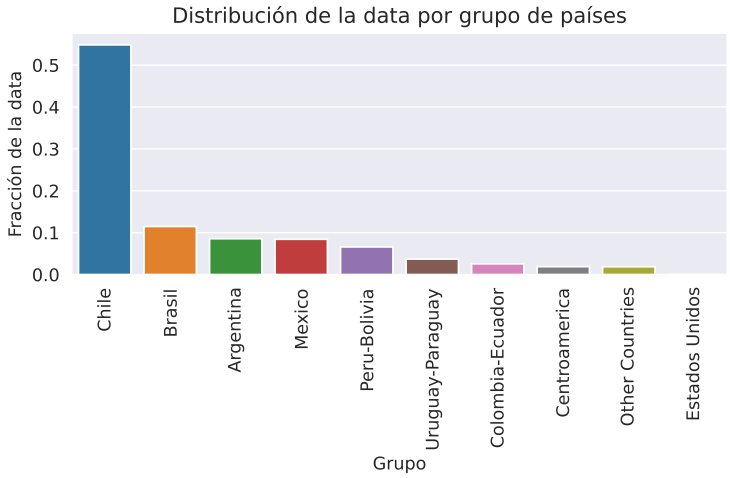

Figura 4.1: Histograma de fracción de la data, agrupada por países [Elaboración propia]

En esta memoria se trabajará solo con los clientes de Chile porque se cuenta con un mayor volumen de datos. Estos clientes conforman 8.359 filas de la data, 457 clientes y 159 fugas registradas a junio de 2021.

La otra fuente de información para este trabajo es la base de datos de la empresa, donde se almacena información sobre el uso del servicio por parte de los clientes, que contempla desde variables generales como las llamadas al servicio de optimización hasta un nivel granular de las visitas (entregas) que cada vehículo de los clientes ha realizado.

Uno de los problemas/desafíos para cruzar la información de facturación con la de uso del servicio es la falta de un mapeo completo de IDs por empresa (un ID corresponde a un usuario de empresa). En la situación actual solo el 55 % de los registros de las empresas tienen asociado un ID en la data de facturación (cuando no se conoce el ID de una empresa, todos sus registros tienen el campo vacío), además, como se puede ver en la figura 4.2, existen empresas con múltiples cuentas, lo que significó una dificultad puesto que empresas con un ID registrado podrían estar almacenando su información un un segundo o tercer ID desconocido. En la sección de preparación de la data se comenta como se trabajó esta situación para obtener un conjunto de datos con ambas fuentes, el cual resultó tener 2.327 filas correspondientes a 235 clientes (197 activos y 38 fugados).

Figura 4.2: Histograma de número de cuentas por cliente (con 2 o más cuentas) [Elaboración propia]

4.2.2. Variables para modelar

4.2.2.1. Variable dependiente

La variable dependiente de este modelo es una dicotómica llamada **"churn"**, que a nivel mensual que toma el valor 0 cuando el cliente sigue activo en el siguiente período y 1 cuando se fuga en dicho momento.

Para entender el comportamiento de los clientes fugados se generó la figura 4.3 donde se muestra que en los meses en torno al quinto período de actividad es donde se han registrado los peaks de fugas, por lo que puede ser relevante prestar especial atención a clientes con esta característica.

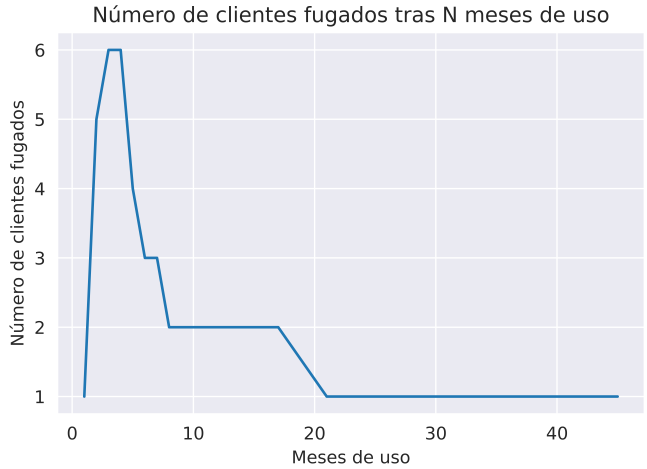

Figura 4.3: Número de clientes fugados tras N meses de uso [Elaboración propia]

También se visualizó los meses del año donde ocurren las fugas, como se puede ver en

la figura 4.4, donde los peaks de fuga ocurren en los trimestres (marzo, junio, septiembre, diciembre). Esto podría estar relacionado a que algunas empresas evalúen su funcionamiento a nivel a finales de trimestres y sea necesario un esfuerzo adicional en estos períodos para generar estrategias de fidelización.

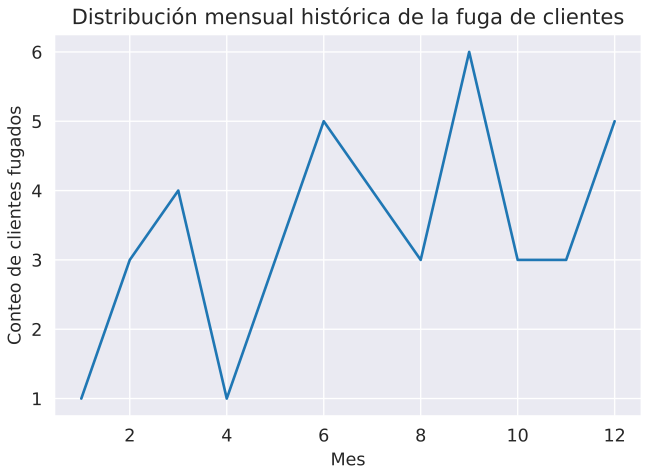

Figura 4.4: Distribución mensual de la fuga de clientes [Elaboración propia]

4.2.2.2. Variables a nivel cliente y mes

1. **Facturación:** Monto facturado (en $USD) para cada cliente y mes. El monto facturado se calcula a partir del número de vehículos a optimizar que el cliente contrata. La evolución del monto facturado promedio de los clientes se puede ver en la figura 4.5, donde se separó con la etiqueta "cliente_churn" (1 cuando el cliente se fuga en cualquier período y 0 si se mantiene activo a la fecha 2021-01). Con respecto a los clientes activos, en promedio han mantenido un nivel de facturación constante, con una leve disminución en 2020 (inicio de la pandemia) producto de la entrada de un número elevado de clientes pequeños. Para el caso de los clientes fugados, las bajas del final están relacionadas a la fuga de clientes de tamaño mediano.

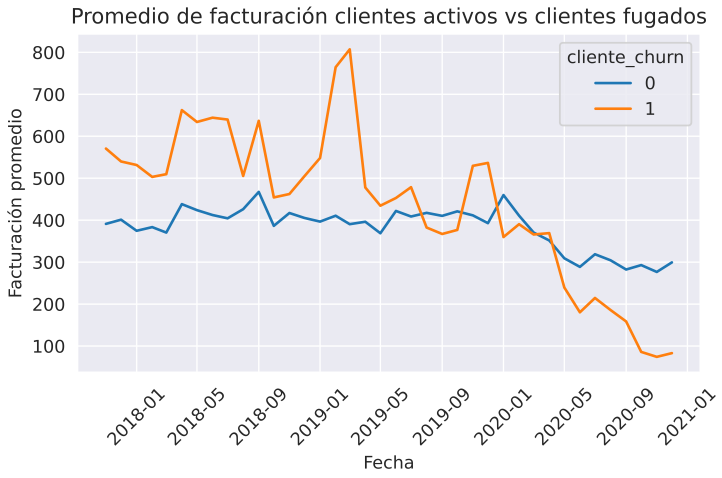

Figura 4.5: Promedio de facturación clientes activos v/s clientes fugados [Elaboración propia]

2. **Métricas de uso de la plataforma:**

   - **Optimizaciones:** Número de veces que un cliente utilizó el servicio de optimización en el mes. Una optimización consta de la planificación inteligente de varias rutas de reparto, la cual puede ser configurada mediante diversos parámetros que pueden influir en el desempeño del algoritmo de optimización. Si el resultado es guardado, entonces se almacena como plan. En la figura 4.6 se puede ver el número de optimizaciones promedio por cliente, donde no es clara una diferencia entre clientes fugados y clientes activos.

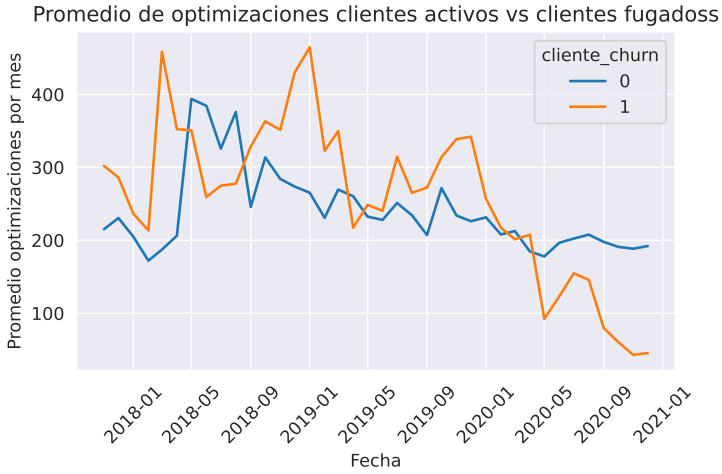

Figura 4.6: Promedio de optimizaciones clientes activos v/s clientes fugados [Elaboración propia]

   - **Planes:** Conjunto de rutas de reparto. Este número hace referencia al número de veces que el cliente planificó sus repartos, es decir, las instancias que el usuario requirió usar el servicio. En la figura 4.7 se puede ver el número de planes promedio por cliente, donde se aprecia que clientes activos guardan más planes que clientes

fugados. Dado que, en promedio, ambos grupos optimizan un número similar de veces, el que clientes fugados tengan menos planes indica que están siendo menos eficaces al usar la herramienta, requiriendo de más iteraciones para poder llegar a un resultado agradable.

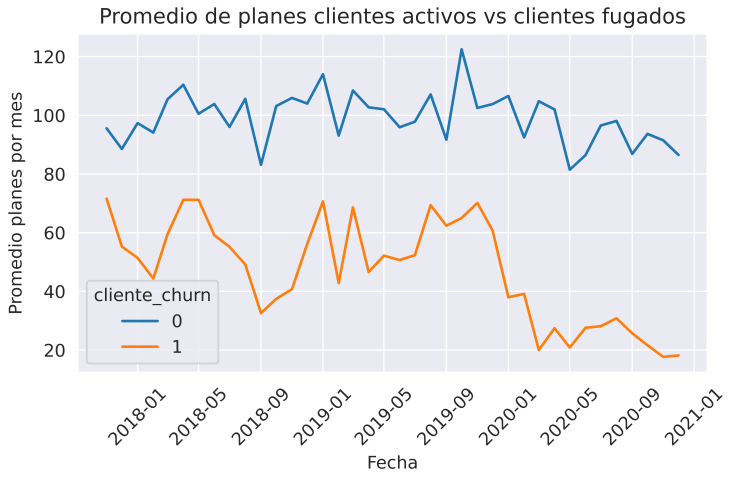

Figura 4.7: Promedio de planes clientes activos v/s clientes fugados [Elaboración propia]

- **Rutas:** Número de rutas que un cliente realizó en un mes. Una ruta contempla a un vehículo y una serie de localizaciones ordenadas de despacho. Este indicador da cuenta del desempeño del cliente en su propio negocio cuando es comparado con períodos anteriores. En la figura 4.8 se puede ver el número de rutas guardadas en promedio por cliente, donde para el caso de los clientes activos, se puede ver un alza en el período post pandemia (presumiblemente por el aumento del e-commerce); por su parte, no es clara la razón de las variaciones en los clientes fugados, estas podrían estar influenciadas por la aparición y fuga de clientes, por la estacionalidad de la demanda (hay peaks en enero), u otras razones.

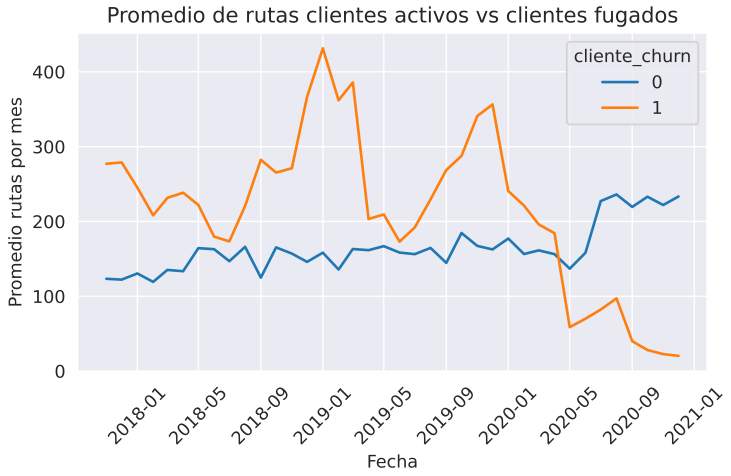

Figura 4.8: Promedio de rutas clientes activos v/s clientes fugados [Elaboración propia]

4.3. Preparación de la data

4.3.1. Tratamiento inicial de la data

1. **Data de facturación:**
   - Se le dió formato a las variables según el tipo de datos al que correspondían: entera, float, fecha o string.
   - Se colocaron en mayúscula todos los valores de las variables string para minimizar errores de tipeo entre los nombres de las empresas.
   - Se eliminaron las empresas que no habían contratado un servicio de manera recurrente (contrataciones por un mes en particular o clientes piloto que no prosperaron).
   - Se agruparon todos los clientes que poseían más de un centro de costos (que figuraban con más de una factura al mes).
   - Se corrigió la marca de fuga etiquetando aquellos casos en los que pese a que un cliente no había renovado su suscripción, aún figuraba como activo en su último mes de registro.
   - Se estandarizaron los nombres de los clientes que presentaban un cambio de nombre en los distintos períodos.
   - Se corrigió la agrupación de facturas por cliente y la marca de churn.
2. **Métricas de uso del servicio:**
   - Se generó un diccionario que asociaba los nombres de las cuentas (de la data de facturación) con los IDs de la base de datos de manera manual. Puesto que existían clientes con múltiples cuentas y que en la data de facturación habían IDs que no correspondían al cliente etiquetado, se determinó hacerlo desde cero, apoyado con los IDs candidatos de la data de facturación. Esta tarea consistió en buscar a través de consultas en SQL posibles matchs entre nombres de clientes 1 por 1 (probando con diversas combinaciones, apoyado con búsquedas en Google y consultas a personas de la empresa).
   - Se extrajo la data correspondiente a la información de optimizaciones/planes/rutas a nivel mensual por ID utilizando los IDs recolectados.
   - Se unificaron las multicuentas a través de una suma de período a período de cada una de sus métricas.

4.3.2. Agrupación y corte de la data

- Para la agrupación de la data, se unió mediante el nombre de la empresa, mes y año de la información.
- Debido a que las métricas de uso comenzaron a guardarse a partir de octubre de 2017 se modelará con datos de dicho mes en adelante.
- Por último, solo 235 clientes poseían información de las métricas en todos sus períodos, esto debido a que existen clientes con integraciones especiales que no almacenan estos datos y otros clientes que poseen pérdida de información. Para este estudio se decidió trabajar solo con los 235 clientes.

4.3.3. Generación de variables y preparación para modelar

- Para rescatar información de la historia de cada cliente para un período dado, se generaron las siguientes variables:

  1. **lag1_OPTI:** Número de optimizaciones en el período anterior.
  2. **lag1_PLAN:** Número de planes en el período anterior.
  3. **lag1_RUTA:** Número de rutas en el período anterior.
  4. **lag1_USD:** Monto facturado en el período anterior.
  5. **lag1_OPTI:** Número de optimizaciones en el período anterior.
  6. **dif1_OPTI:** Diferencia en el número de optimizaciones con el período anterior.
  7. **dif1_RUTA:** Diferencia en el número de rutas con el período anterior.
  8. **dif1_PLAN:** Diferencia en el número de planes con el período anterior.
  9. **dif1_USD:** Diferencia de facturación con el período anterior.
  10. **meses_uso:** Número de meses que el cliente lleva activo hasta la fecha.
  11. **mean_3m_USD:** Promedio del monto facturado en los últimos 3 meses.
  12. **mean_3m_RUTA:** Promedio del número de rutas en los últimos 3 meses.
  13. **mean_3m_PLAN:** Promedio del número de planes en los últimos 3 meses.
  14. **mean_3m_OPTI:** Promedio del número de optimizaciones en los últimos 3 meses.

- Luego, se generó una copia alternativa donde las variables fueron estandarizadas, para ser utilizada en el modelamiento de logit.

- Por último, para poder modelar se separó la data en entrenamiento/testeo aplicando k-folds cross-validation en ambos conjuntos de datos, considerando 5 grupos aleatorios y manteniendo las proporciones de clientes activos/fugados en cada grupo.

4.4. Modelamiento

Para este trabajo se utilizaron 10 combinaciones de variables, las que se muestran en la tabla 4.1, para el entrenamiento de modelos logit y XGBoost . Estos grupos fueron armados tomando las variables de fecha y antigüedad como base, y fueron combinados con las de métricas de uso y de facturación, con el objetivo de evaluar cuantos meses de historia combine utilizar para la predicción, y si es mejor trabajar con diferencias entre meses o no.

Tabla 4.1: Variables utilizadas en cada modelo

Variable	M 1	M 2	M 3	M 4	M 5	M 6	M 7	M 8	M 9	M 10
Año	x	x	x	x	x	x	x	x	x	x
Mes	x	x	x	x	x	x	x	x	x	x
meses_uso	x	x	x	x	x	x	x	x	x	x
USD	x	x		x		x		x	x	x
n_rutas_mes	x	x		x		x		x	x	x
n_planes_mes	x	x		x		x		x	x	x
n_optimizaciones_mes	x	x		x		x		x	x	x
lag1_USD		x	x	x			x	x	x	
lag1_RUTA		x	x	x			x	x	x	
lag1_PLAN		x	x	x			x	x	x	
lag1_OPTI		x	x	x			x	x	x	
dif_USD_1m				x	x	x	x			
dif_RUTA_1m				x	x	x	x			
dif_PLAN_1m				x	x	x	x			
dif_OPTI_1m				x	x	x	x			
dif_pr_USD_1m								x		
dif_pr_RUTA_1m								x		
dif_pr_PLAN_1m								x		
dif_pr_OPTI_1m								x		
mean_3m_USD									x	x
mean_3m_RUTA									x	x
mean_3m_PLAN									x	x
mean_3m_OPTI									x	x

4.4.1. XGBoost

Se entrenaron modelos de clasificación de XGBoost con la librería sklearn [Pedregosa et al., 2011] de Python, utilizando como función objetivo "binary:logistic" que se utiliza en clasificación dicotómica. Para el entrenamiento se utilizó validación cruzada de k-folds generada en la preparación de la data y se ajustaron hiperparámetros mediante grid search, maximizando las métricas de ROC_AUC y sensibility, cada una por separado. Vale destacar que, para tratar la problemática del desbalanceo de clases (hay 5 veces más clientes activos que fugados, y 61 filas de períodos activos por una fila de fuga), se utilizó el hiperparámetro scale_pos_weight que pondera el error de clasificación de la clase menos representada.

4.4.2. Regresión logística

Se entrenó un modelo de regresión logística multivariado con la librería sklearn [Pedregosa et al., 2011] de Python, utilizando como hiperparámetro la ponderación del error de la clase des-balanceada (sample_weight). Al igual que con el algoritmo anterior, el entrenamiento se realizó maximizando las métricas ROC_AUC y sensibility por separado.

4.4.3. Modelo alternativo: árbol de decisión

A partir del modelo con mejor desempeño se realizó un modelo de benchmark simple, correspondiente a un árbol de decisión con un nodo, donde se clasificó cortando las ramas con la variable más relevante del modelo seleccionado. Esto representa la capacidad predictiva actual, sin invertir esfuerzos en la implementación de un modelo de machine learning, y se espera que un modelo más complejo tenga un mejor desempeño que esta clasificación simple.

4.5. Evaluación

4.5.1. Métricas de desempeño

Los modelos se compararon utilizando las métricas de desempeño de ROC_AUC y precisión.

4.5.2. Interpretabilidad de los modelos

Se analizó la influencia de las variables en la probabilidad de fuga de cada modelo. En el caso de la regresión logística se hizo a través de la interpretación de sus coeficientes, y para el caso XGBoost se utilizó el análisis de Shap values.

5 | Resultados

5.1. Resultado de los modelos

5.1.1. XGBoost

Los resultados obtenidos en el entrenamiento de cada modelo de XGBoost son los presentados en la tabla 5.1, donde los modelos con nomenclatura M X.1 fueron entrenados maximizando la métrica de evaluación ROC_AUC mientras que los modelos M X.2 se entrenaron maximizando la sensibilidad.

Tabla 5.1: Resultados de entrenamiento de modelos de XGBoost

Modelo	ROC_AUC	Sensibility	Specifity	Precision
M 1.1	0.75	0.56	0.93	0.13
M 1.2	0.70	0.69	0.71	0.04
M 2.1	**0.76**	**0.56**	**0.96**	**0.22**
M 2.2	0.74	0.64	0.83	0.06
M 3.1	0.63	0.72	0.54	0.03
M 3.2	0.65	0.72	0.58	0.03
M 4.1	0.76	0.59	0.93	0.12
M 4.2	**0.77**	**0.59**	**0.94**	**0.14**
M 5.1	0.65	0.44	0.86	0.05
M 5.2	0.66	0.88	0.44	0.02
M 6.1	0.76	0.56	0.96	0.19
M 6.2	0.71	0.69	0.72	0.04
M 7.1	0.73	0.59	0.87	0.07
M 7.2	0.70	0.74	0.66	0.04
M 8.1	**0.78**	**0.67**	**0.90**	**0.10**
M 8.2	0.74	0.72	0.76	0.05
M 9.1	**0.76**	**0.56**	**0.97**	**0.23**
M 9.2	0.68	0.77	0.59	0.03
M 10.1	0.76	0.56	0.96	0.20
M 10.2	0.73	0.61	0.84	0.06

De la tabla 5.1 se destacan los modelos 4.2 y 8.1 que presentan un mejor desempeño en cuanto a ROC_AUC (0,77 y 0,78, respectivamente), también los modelos 2.1 y 9.1 que, además de tener niveles cercanos de ROC_AUC con respecto a los primeros (0,76), su precisiones alcanzan los valores más altos entre los modelos de XGBoost entrenados (0,22 y 0,23).

Debido a lo recién mencionado, el modelo de XGBoost escogido es el 2.1, puesto que posee niveles de ROC_AUC similares a los otros 3 modelos, y alcanza una precisión muy cercana a la del modelo 9.1 (-0.01) utilizando 4 variables menos.

5.1.2. Regresión logística

Los resultados obtenidos en el entrenamiento de cada modelo de regresión logística son los presentados en la tabla 5.2, donde, al igual que para XGBoost,los modelos con nomenclatura M X.1 fueron entrenados maximizando la métrica de evaluación ROC_AUC mientras que los modelos M X.2 se entrenaron maximizando la sensibilidad.

Tabla 5.2: Resultados entrenamiento modelos logit

Modelo	ROC_AUC	Sensibility	Specifity	Presicion
M 1.1	0.64	0.90	0.38	0.025
M 1.2	0.64	0.90	0.38	0.025
M 2.1	0.71	0.86	0.56	0.035
M 2.2	0.71	0.86	0.56	0.035
M 3.1	0.65	0.79	0.50	0.028
M 3.2	0.56	0.87	0.25	0.021
M 4.1	0.71	0.86	0.56	0.035
M 4.2	0.71	0.86	0.56	0.035
M 5.1	0.63	0.80	0.45	0.027
M 5.2	0.61	0.89	0.34	0.024
M 6.1	0.71	0.86	0.56	0.035
M 6.2	0.71	0.86	0.56	0.035
M 7.1	0.71	0.86	0.56	0.035
M 7.2	0.71	0.86	0.56	0.035
M 8.1	**0.72**	**0.76**	**0.67**	**0.041**
M 8.2	0.70	0.79	0.60	0.036
M 9.1	**0.70**	**0.68**	**0.72**	**0.041**
M 9.2	0.66	0.71	0.62	0.032
M 10.1	0.70	0.75	0.66	0.038
M 10.2	0.70	0.75	0.66	0.038

De la tabla 5.2 se destacan los modelos 8.1 y 9.1 por alcanzar la precisión más alta (0,041), particularmente el primero posee además el mejor ROC_AUC (0,72), por lo que se escoge como el mejor al modelo logit 8.1.

5.2. Interpretación de variables en modelos

5.2.1. XGBoost

A partir del modelo seleccionado de XGBoost (modelo 2.1) se confeccionó el gráfico de la figura 5.1, donde F-score representa el número de veces que una variable realiza un corte en el modelo. En el gráfico se puede notar que la variable "n_planes_mes" es la más utilizada por el modelo para discriminar si un cliente tiene más o menos probabilidad de fuga. Otras variables relevantes en la clasificación son "lag1_OPTI" y "n_optimizaciones_mes".

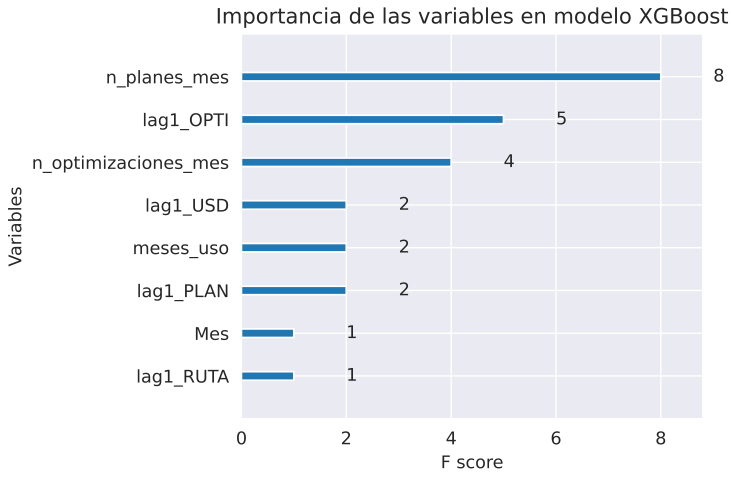

Figura 5.1: Importancia de variables en modelo XGBoost [Elaboración propia]

Para profundizar en el entendimiento de la influencia de las variables se estudiaron los shap values que explican la contribución marginal de cada variable en la asignación de probabilidad que el modelo entrega a cada cliente/mes. A partir de este estudio se descubrió que el modelo comienza con una probabilidad de fuga base de -1.246 (probabilidad negativa de fuga), a lo que se suma el efecto asociado al valor de cada variable. En la figura 5.2 se presenta cada uno de los efectos marginales de las variables para cada predicción, asociando colores de tonalidad roja a valores altos de la variable, y azul a los bajos. De este gráfico se destaca el efecto de "n_planes_mes", que cuando toma valores altos (rojos) se asocia a una reducción en la probabilidad de fuga, mientras que solo para valores bajos (azul) esta variable contribuye al aumento de dicha probabilidad. Otras variables de métricas de uso de la plataforma tienen un comportamiento similar a "n_planes_mes" pero con una influencia menor (en valor absoluto). Otra variable relevante es meses_uso, que a clientes que han usado la plataforma por más meses (rojos) tienen una influencia negativa en la probabilidad de fuga predicha, al contrario que los clientes más nuevos (azules).

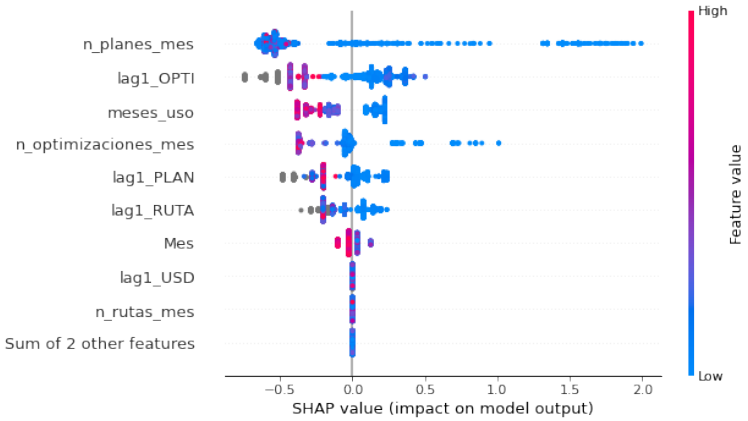

Figura 5.2: Influencia marginal de variables en modelo XGBoost [Elaboración propia]

A partir de las contribuciones marginales de cada variable, se confeccionó la figura 5.3

donde se muestra el promedio (en valor absoluto) de estas contribuciones en la predicción del modelo. En este gráfico se reafirma la gran influencia que tiene la variable "n_planes_mes" en la predicción del modelo, alcanzando un valor equivalente al doble que la segunda variable más influyente ("lag1_OPTI").

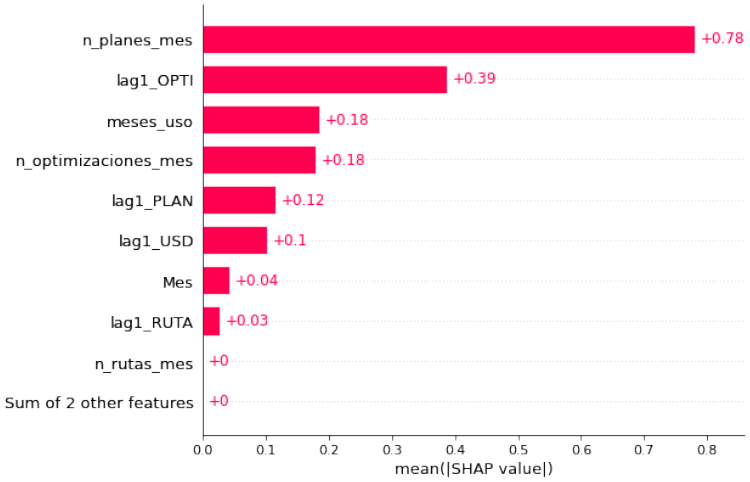

Figura 5.3: Influencia marginal promedio de variables (en valor absoluto) en modelo XGBoost [Elaboración propia]

Profundizando en el efecto de la variable "n_planes_mes", se elaboró la figura 5.4 donde se clarifica que la influencia positiva de esta variable en la predicción ocurre solamente en valores bajos de esta variable.

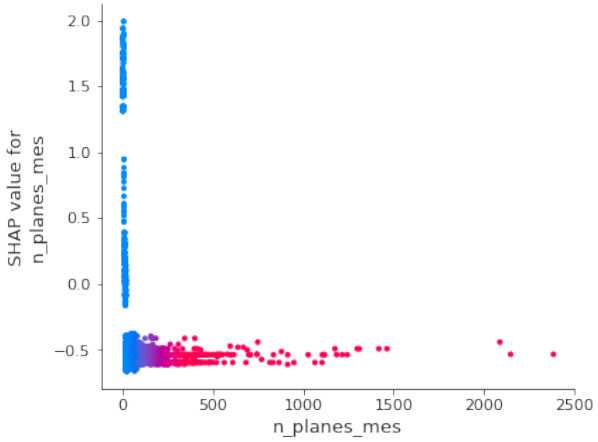

Figura 5.4: Influencia marginal de la variable "n_planes_mes" en modelo XGBoost [Elaboración propia]

Por último, se graficó la influencia de la variable "meses_uso" en la clasificación del modelo (figura 5.5), mostrando un tramo de influencia positiva en los primeros 7 meses de actividad de un cliente, luego, entre el mes 8 y el 17 la influencia es negativa, y a partir del mes 18 la influencia es aún más negativa.

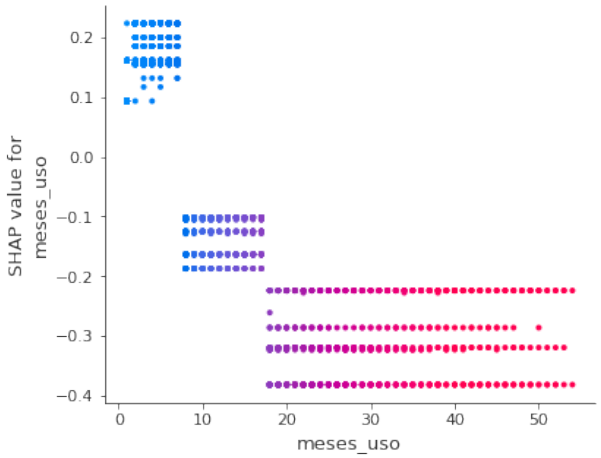

Figura 5.5: Influencia marginal de la variable “meses_uso” en modelo XGBoost [Elaboración propia]

5.2.2. Regresión logística

A continuación, en la tabla 5.3, se presentan los coeficientes de regresión logística asociados a las variables en el modelo 8.1. Entre las variables con coeficientes de mayor magnitud negativa se encuentra “n_rutas_mes”, “n_planes_mes”, “lag1_USD” y “lag1_OPTI”, mientras que las variables que mas influyen en una clasificación positiva por parte del modelo son “lag1_RUTA”, “USD” y “lag1_PLAN”. Vale destacar que en este modelo el intercepto toma un valor negativo (-2.39).

Tabla 5.3: Coeficientes modelo logit

Variables	Coeficientes	Promedio	Std
Intercepto	-2.39	-	-
Año	-0.02	2019.4	0.8
Mes	-0.33	7.2	3.2
meses_uso	-0.81	15.3	12.7
USD	4.04	372.7	511.7
n_rutas_mes	-10.35	193.2	466.8
n_planes_mes	-4.12	94.2	166.2
n_optimizaciones_mes	1.65	230.4	364.9
lag1_USD	-3.37	366.1	497.8
lag1_RUTA	5.19	181.6	431.1
lag1_PLAN	2.18	89.5	151.6
lag1_OPTI	-2.49	227.9	363.7
dif_pr_USD_1m	-1.47	0.1	0.5
dif_pr_RUTA_1m	-1.31	0.3	2.5
dif_pr_PLAN_1m	0.85	0.2	1.5
dif_pr_OPTI_1m	-0.17	0.2	1.4

5.3. Método alternativo: Árbol de decisión simple

A partir del modelo con mejor performance (XGBoost) se seleccionó su variable más relevante (“n_planes_mes”) para la elaboración de un árbol de decisión de un nodo para clasificar la posible fuga. El resultado de la clasificación para distintos cortes se presenta en

la figura 5.6, donde se destaca el corte en 8 planes (Sensibility = 0.55, Precision = 0.14) y el corte de 23 planes (Sensibility = 0.76, Precision = 0.032), ambos con una sensibilidad similar a la de los modelos seleccionados de XGBoost y Logit respectivamente.

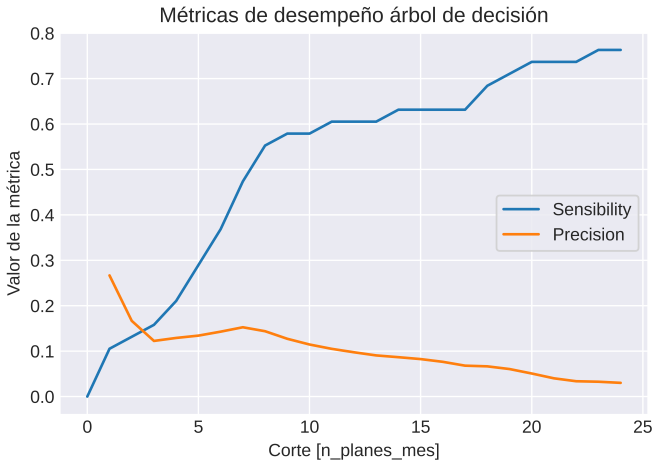

Figura 5.6: Métricas de desempeño árbol de decisión [Elaboración propia]

6 | Discusión

A partir de los resultados, en este capítulo se profundizará en torno al significado y la interpretación de estos, comenzando por lo referente al desempeño de los modelos, luego sobre la interpretación de las variables en los modelos para terminar con recomendaciones y limitaciones derivadas del estudio.

En primer lugar, para evaluar el desempeño de los modelos XGBoost y Logit es necesario profundizar en el modelo alternativo (árbol de decisión simple), que muestra el trade-off entre detectar más clientes fugados (Sensibility) versus que las alertas de posibles fugas sean más confiables (Precision). Para que sea conveniente destinar esfuerzos en la implementación de un modelo de machine learning, este debe, al menos, tener un mejor performance que el árbol de decisión de un nodo, que traducido a las métricas utilizadas, para un nivel dado de sensibilidad alcance una mayor precisión. Con esto en mente, el corte en 8 planes en el modelo alternativo es comparable al modelo XGBoost, donde el segundo alcanza una precisión más alta (+0.08 equivalente a una mejora del 57 %). De la misma forma, el corte en 23 planes es comparable con el modelo logit, donde alcanzan un nivel similar de precisión, con una pequeña ventaja para el modelo logit (+0.009 equivalente a una mejora del 28 %). De estas comparaciones se desprende que, en el escenario actual, utilizar el modelo de XGBoost podría añadir valor en la predicción de fuga, y en menor medida el modelo logit también.

El segundo punto a mencionar es la comparación en el desempeño de ambos modelos, donde, a través del ROC_AUC, se evidencia que XGBoost clasifica mejor que el modelo logit utilizando este conjunto de datos. Cabe destacar que la gran diferencia en precisión viene dada por el umbral de clasificación (0.5) que para el caso del modelo logit resulta en elevar el número de falsos positivos, los que, en caso de aumentar el umbral, deberían disminuir mejorando la precisión del modelo (a costa de reducir la sensibilidad). Este ajuste de umbral es un tema interesante a estudiar para un posible uso de cualquiera de los modelos, pues permitiría a la empresa evaluar que porcentaje de clientes fugados quiere detectar, tomando en cuenta la capacidad del equipo para realizar acciones de fidelización.

Con respecto a los resultados del estudio de las variables del modelo XGBoost, se puede ver que las métricas de uso capturan la idea intuitiva de que si un cliente utiliza más la plataforma este debería ser menos propenso a fugarse, efecto particularmente relevante con la variable "n_planes_mes" que por si sola permite generar una predicción de fuga interesante (con el modelo alternativo). Otro comportamiento intuitivo que se captura es que un cliente que se mantiene por más tiempo es menos propenso a la fuga. Asimismo, en este punto el modelo realiza 2 cortes con esta variable sugiriendo la existencia de 3 rangos de probabilidad de fuga relacionados con los "meses_uso", particularmente el rango de 0 a 7 meses (el de mayor

probabilidad de fuga) podría estar relacionado a clientes que no encontraron lo que buscaban en el servicio o no supieron sacarle provecho a la plataforma. Estudiar a este segmento de clientes fugados podría ayudar en el entendimiento sobre como poder llegar a otro segmento de clientes, o fidelizar de mejor manera al segmento que actualmente el el objetivo.

Pasando al análisis de los coeficientes del modelo logit, una interpretación de como opera es la siguiente: similar a lo ocurrido con el modelo XGBoost, las métricas de uso (en su mayoría) tienen un impacto negativo en la probabilidad de fuga entregada por el modelo, por otro lado, variables como "USD" y "lag1_RUTA" tienen coeficientes positivos lo que captura la idea de que clientes que pagan más dinero y utilizan poco la plataforma (métricas de uso bajas) recibirán una probabilidad de fuga mayor, lo mismo ocurre con un cliente que en el mes anterior realizó muchas rutas y en el mes actual el número de rutas disminuye.

Una limitante del trabajo corresponde a la disponibilidad de información para modelar. Si bien con la información existente se consiguió armar un modelo capaz de predecir la fuga, la existencia de nuevas variables podría mejorar el desempeño del modelo. En otros estudios de predicción de fuga de clientes en empresas SaaS se utilizaron métricas relacionadas a las sesiones de los usuarios [Ge et al., 2017], que para este contexto se podría emular guardando la duración de las sesiones de usuarios, número de inicios de sesión al mes. También existen variables interesantes que se encontraban incompletas en la base de datos, como la clasificación por rubro de clientes y tipos de empresas, que podrían ser completadas. Otra información con la que se cuenta es el registro de reclamos (tickets) por parte de los clientes, que no fue incluida por su horizonte temporal, pero que debería ser incluida en próximos estudios.

Otra limitante que tienen los modelos está en el tipo de uso recomendado. Puesto que el entrenamiento de los modelos se enfocó en maximizar la detección de la fuga de clientes, se recomienda que, en caso de generar acciones de retención, estas sean de bajo costo (como llamadas telefónicas). Para evaluar acciones de retención que significasen un costo mayor (por ejemplo ofrecer descuentos) se recomienda hacer un nuevo estudio donde se considere la función de utilidad, que en otros estudios ha logrado un mejor desempeño en el incremento de las ganancias [Verbeke et al., 2012].

7 | Conclusiones y trabajos futuros

7.1. Conclusiones

En este trabajo se buscó elaborar un modelo capaz de predecir la fuga de clientes de una empresa logística, además de buscar entender el cómo influyen las variables en el desempeño de la clasificación, utilizando un conjunto de datos con información de los clientes pequeño y altamente desbalanceado (7 variables base, 2327 filas, 235 clientes y 38 fugas).

Con respecto a los modelos, se entrenaron 2 modelos de clasificación (con 10 configuraciones de variables), por un lado XGBoost (ROC_AUC = 0.76, Sensibility = 0.56, Precision = 0.22) y por otro el modelo logit (ROC_AUC = 0.72, Sensibility = 0.76, Precision = 0.04) que obtuvieron un buen desempeño para abordar este desafío, lo que se observa en el ROC_AUC alcanzado. En este sentido, tomando en cuenta que el problema considera la clasificación desbalanceada, se puede concluir que para este tipo de datos las técnicas de ponderación del error en la clasificación son efectivas. Adicionalmente, entre ambos clasificadores, el XGBoost alcanzó un mejor performance en cuanto a la métrica ROC_AUC utilizando 4 variables menos, siendo, para este estudio, una alternativa más simple y efectiva.

El estudio de la influencia se realizó a través de los coeficientes de regresión para el modelo logit, y estudiando los shap values para XGBoost. Con respecto a la interpretación de coeficientes de logit, esto no fue una tarea sencilla dado que las variables presentan una desviación estándar alta (en la mayoría de los casos superior a su promedio), lo que dificultó discriminar si los coeficientes más altos realmente influían en la predicción. En cambio, el estudio de los shap values del modelo XGBoost entregó de manera gráfica y sencilla información sobre el efecto de las variables en la clasificación, resultando en hallazgos interesantes como que en los primeros 7 meses el modelo asigna una probabilidad de fuga mayor que en meses posteriores (bajo condiciones similares en otras variables), o que la variable "n_planes_mes" por si sola es capaz de predecir la fuga con un desempeño que se acerca al de los modelos entrenados (Sensibility = 0.55, Precision = 0.14). En este sentido, el estudio de shap values logró extraer información relevante de un modelo, la cual permite sugerir el monitoreo de clientes que registren menos de 9 planes en un mes, expecialmente en los primeros 7 meses de actividad, para poder aplicar estrategias de retención de clientes.

Para finalizar, es importante rescatar la limitación del estudio. Debido a que se trabajó con un segmento del total de clientes (clientes chilenos con información completa de sus métricas

de uso), los modelos entrenados deben ser utilizados para la predicción de probabilidad de fuga de dicho segmento. Con respecto al resto de los clientes, se recomienda enfocar esfuerzos para mejorar la base de datos, en particular estructurar la conexión entre ID's de usuarios de una misma empresa y el almacenamiento de métricas, de modo de poder replicar este estudio con todos los clientes.

7.2. Trabajos futuros

Para futuros estudios, con respecto a las técnicas de modelamiento, se propone probar otro enfoque en el entrenamiento de logit utilizando técnicas de oversampling como SMOTE, que podrían mejorar el performance en la clasificación. Además, sería muy valioso estudiar la fuga de clientes desde la perspectiva de la utilidad, esto permitiría levantar información útil para la toma de decisiones relacionadas con estrategias de fidelización.

Por otro lado, con respecto a la data de la empresa, se propone que, en caso de corregir los problemas detectados en el almacenamiento de la información, se repliquen los modelos de este estudio, con el objetivo de comprobar si los resultados son generalizables para la totalidad de clientes.

Por último, cuando se logre recopilar más meses de información, se propone analizar si los clientes que contrataron el servicio durante la pandemia se comportan distinto a los que eran clientes antes de la pandemia. Esta idea surge de la hipótesis de que, durante la pandemia, algunas empresas se vieron obligadas a modernizarse y no optaron por el servicio de forma natural. De esta forma, la empresa podría obtener información Este estudio sobre como atraer a un nuevo segmento de clientes.

Bibliografía

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

[Chui et al., 2019] Chui, M., Henke, N., and Miremadi, M. (2019). Most of AI's business uses will be in two areas. *McKinsey Analytics*. <https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Analytics/Our%20Insights/Most%20of%20AIs%20business%20uses%20will%20be%20in%20two%20areas/Most-of-AIs-business-uses-will-be-in-two-areas.pdf> [Consulta: 20 abril 2021].

[Ge et al., 2017] Ge, Y., He, S., Xiong, J., and Brown, D. E. (2017). Customer churn analysis for a software-as-a-service company. In *2017 Systems and Information Engineering Design Symposium (SIEDS)*, pages 106–111.

[Lundberg and Lee, 2017] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[OMS, 2020] OMS (2020). Covid-19: cronología de la actuación de la oms. <https://www.who.int/es/news/item/27-04-2020-who-timeline---covid-19> [Consulta: 20 abril 2021].

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Verbeke et al., 2012] Verbeke, W., Dejaeger, K., Martens, D., Hur, J., and Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218(1):211–229.

[Vota et al., 2020] Vota, D., Herman, J., Marín, D., and Robnett, S. (2020). The cloud market in latam: Jumping on a high-speed train. *BCG Platinion*. <https://bcgplatinion.com/insights/the-cloud-market-in-latam/> [Consulta: 13 septiembre 2021].