Universidad del Desarrollo
Facultad de Ingeniería

IMPLEMENTACIÓN Y ANÁLISIS DE MODELOS PREDICTIVOS DE DETECCIÓN DE FUGA PARA CLIENTES RESIDENCIALES DE LA EMPRESA LIPIGAS S.A

POR: JOSÉ PEDRO CORDERO BERENGUER

Capstone proyecto presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESOR GUÍA: DR. MAURICIO RENÉ HERRERA MARIN

ENERO, 2023

SANTIAGO, CHILE

AGRADECIMIENTO

Doy gracias a Mauricio René y Alejandro Álvarez quienes me guiaron a ver la luz al final del túnel

TABLA DE CONTENIDO

Resumen

La fuga de clientes es un problema relevante al que se enfrentan las empresas de servicios que operan en mercados saturados y altamente competitivos, y que les puede generar pérdidas económicas significativas. Para las empresas de servicio que ofrecen servicios no contractuales no hay una variable observable directamente asociada con la fuga, lo que hace a este problema un desafío. Sin embargo, es posible definir la fuga en base al comportamiento de los clientes y estimar una probabilidad de fuga asociada a cada uno de ellos.

En el presente trabajo se propuso una definición de fuga de clientes calculada a partir de los datos y se evaluaron los desempeños de 7 modelos para predecir la fuga clientes en el contexto de la venta de cilindros de gas para clientes residenciales de la empresa LIPIGAS S.A. Entre los modelos el que presentó mejor desempeño, en base a la métrica ROC (AUC), fue el modelo de Regresión Logística.

En cuanto a la interpretación de variables, el análisis de *shap values* permitió identificar que la variable "días desde la última compra" es la que mayor peso tuvo en la predicción de fuga; destacando que a mayor número de días desde la última compra la probabilidad de fuga crece.

Palabras Claves: Aprendizaje de Máquinas, Predicción de Fuga, Regresión Logística, Interpretabilidad de los Modelos de Aprendizaje de Máquinas

1. Introducción

El área comercial tiene un rol fundamental de poder establecer relaciones comerciales para atraer y retener clientes mediante campañas de marketing. Sin embargo, retener a un cliente resulta más rentable para una empresa que atraer a uno nuevo. Esta rentabilidad se debe a que un nuevo cliente implica altos gastos operacionales y gastos asociados a acuerdos comerciales, y se transforma en una fuente de beneficios para la empresa cuando comienza a consumir el servicio contratado y otros servicios derivados.

LIPIGAS S.A es una empresa que comercializa GLP en formato a granel[1], gas de medidor[2] y envasado en cilindros, siendo este último de vital interés comercial dado que es la principal forma de distribución del GLP. Los cilindros de gas varían en tamaño dependiendo de las necesidades del cliente y pueden ser entregados en forma directa a clientes residenciales que usan el gas para satisfacer necesidades de cocción de alimentos, calefacción y calentamiento de agua en hogares de todos los tamaños.

LIPIGAS S.A no cuenta con un modelo de aprendizaje de máquinas que permita predecir la fuga de clientes, por lo que las acciones de retención se realizan en base a la experiencia, por lo general se ejecutan de manera tardía y consideran solo la frecuencia de compra de los clientes. En este sentido, la empresa se beneficiaría al contar con un modelo que prediga la probabilidad de fuga de clientes de forma personalizada, en base a los patrones de compra históricos y características de cada cliente.

[1] Formato que permite abastecer hogares, comercios e industrias a través de un tanque individual, diseñado para almacenar GLP en forma segura y eficiente con distintas capacidades

[2] Formato está diseñado para suministrar GLP vía tuberías a múltiples puntos de consumo independientes desde una central de tanques común, las que también son abastecidas en forma regular por camiones inyectores.

Los algoritmos de aprendizaje automático han sido ampliamente utilizados para predecir la fuga de clientes; algunos de ellos son: Regresión Logística [1, 2, 3], Árboles de Clasificación [4, 5, 6, 7], Random Forests [8, 9], XGBoost [10], Máquinas de Soporte Vectorial [11,12,13] y Redes Neuronales Profundas [14]. En cuanto al problema de entrenar modelos de predicción de fuga con variables temporales, en la literatura previa se ha demostrado que particionar el conjunto de datos usando un método de validación fuera de tiempo[3] , el cual implica entrenar y testear el modelo con de datos extraídos de diferentes períodos de tiempo impacta positivamente en la predictibilidad del modelo [15].

La metodología aplicada para la resolución del problema de fuga parte por la selección de datos, donde se determinan las fuentes de datos y el tipo de información a utilizar. Continua con un etapa de preprocesamiento de la base de datos, con el fin de tener información de calidad, que aporte mayor valor a la predicción. Posteriormente, se construye una definición de fuga de clientes y se calcula dicha fuga. Se continua con la etapa de ingeniería de características, en donde se crean variables temporales a partir de variables transaccionales y de servicio. Luego se seleccionan las variables que tengan mayor grado de correlación lineal/no lineal con la fuga. Con estas variables se entrenan y evalúan un conjunto de algoritmos de clasificación binaria, y se selecciona el modelo que presente mejor desempeño en base a la métrica ROC (AUC). Para finalizar se realiza un estudio de interpretabilidad de las predicciones del modelo aplicando *shap values*.

[3] https://towardsdatascience.com/why-isnt-out-of-time-validation-more-ubiquitous-7397098c4ab6

2. Hipótesis y Objetivos

Hipótesis

A partir de datos transaccionales y demográficos del cliente, ambientales/estacionales (temperatura), entre otros tipos de datos de clientes del canal envasado residencial es posible determinar la probabilidad condicional de fuga de cada uno de ellos; y así poder dirigir las acciones comerciales de retención de manera más efectiva.

Objetivo General

Construir/implementar modelos predictivos de detección de fuga para clientes residenciales de la empresa LIPIGAS S.A, basados en datos del cliente: transaccionales y sociodemográficas; y de otras fuentes: temperaturas, precios de la competencia e indicadores macroeconómicos.

Objetivo Específicos

- Preprocesamiento y limpieza de datos, y análisis exploratorio de los datos (EDA).
- Definir criterios de fuga (variable *target*)
- Ingeniería de características para la construcción de variables que sean relevantes para la predicción de fuga de los clientes en el caso específico de la empresa LIPIGAS S.A
- Implementar modelos predictivos supervisados de clasificación binaria para la predicción de fuga.
- Evaluar los resultados de distintos modelos en base a las métricas de desempeño.
- Identificar, en base a los modelos, las variables relevantes que impactan en la fuga; permitiendo caracterizar a los clientes con riesgo de fuga.
- Aplicar el mejor modelo y obtener las predicciones de fuga.

3. Metodología

3.1. Datos

Los datos en este estudio comprende registros desde Enero 2019 hasta finales de Marzo de 2022, y provienen de las siguientes fuentes:

Información de Clientes (características/perfil del cliente): nombre, dirección, teléfono, correo electrónico, comuna, zona censal y región. Esta información es relativamente estática, ya que sólo cambia en horizontes de tiempo largos. Para efecto de este estudio será considerada como información estática

Información Transaccional (variables transaccionales de los clientes): id_cliente (identificador único del cliente), pedido (identificador único del pedido), teléfono, fechas de toma (fecha en la que el cliente realizó el pedido), fecha de entrega (fecha estimada en la que el pedido será entregado), fecha de atención (fecha en la cual el pedido es recepcionado por el cliente), camión, comuna, zona censal, cantidad de unidades vendidas por tipo de formato, kilos vendidos por tipo de formato y el descuento por tipo de cilindro de gas. Este tipo de información es dinámica y va cambiando con cada interacción que tiene el cliente con la empresa. Cabe destacar que esta información, como tal, no es valiosa de por sí, ya que debe ser procesada e interpretada.

Información Sociodemográfico: contiene variables sociodemográficas provenientes de dos fuentes de información, la primera es la encuesta CASEN año 2017[4] y la segunda EQUIFAX . Los atributos de la encuesta CASEN están construidos a nivel comunal, mientras que los provenientes de EQUIFAX están a nivel de persona natural.

Los atributos de la encuesta CASEN más relevantes para nuestro proyecto son: número de mujeres por zona censal, número personas por zona censal, número de casas por zona censal, número de hogares de estrato ACB1, D, S/C, C2, C3 y E, el sexo, ingreso promedio

[4] http://observatorio.ministeriodesarrollosocial.gob.cl/casen-multidimensional/casen/basedatos.php

percapita de la zona censal reportado y el rango etario entre 0 a 14 años, entre 15 a 64 años y mayor a 65 años.

Mientras que los atributos más relevantes de EFX son el índice socioeconómico y la renta promedio, ambas variables están a nivel de persona natural.

Información de Temperatura: contiene todas las mediciones de las temperaturas fue construida consultado una API pública que dispone el servicio climatológico Darksky [5]. Esta base cuenta con registros de temperaturas promedio diarias para las 320 comunas de CHILE donde LIPIGAS tiene presencia en la venta de cilindros de gas del canal envasado residencial.

Información Precios de la Competencia: contiene todos los precios de la competencia por comuna está construido a partir de un robot que extrae de la web de forma periódica los precios de la competencia. Los principales atributos de esta base son: tipo de distribuidor (GASCO, ABASTIBLE), región, comuna, producto (11 Kg o 15 Kg), tipo GLP (normal, catalítico) y el precio.

Información Macroeconómicas: contiene los principales indicadores macroeconómicos fue construido por medio de un proceso RPA de extracción diaria de datos desde dos fuentes de información. La primera tiene relación el los precios del Mont Belview Gas[6], y la segunda con el servicio web Mindicador[7] . Las principales variables de esta base son: dólar diario observador, precio del propano (valor spot), imacec mensual, desempleo mensual, ipc mensual y el pib trimestral.

[5] https://darksky.net
[6] Es el precio de spot del gas licuado de propano producido en Mont Belvieu, TX, en dólares estadounidenses por galón - https://ycharts.com/indicators/mont_belvieu_propane_spot_price
[7] Mindicador - https://mindicador.cl/ - es un servicio web de libre acceso que expone una API que publica los principales indicadores macroeconómicos diarios como históricos de Chile en formato JSON.

Construcción del dataset de entrenamiento

Para la construcción del pipeline de ingeniería de datos se usó en framework Kedro[8] de Python de modo de crear un proyecto de data science con código mantenible, modular y reproducible.

Para construir el conjunto de entrenamiento, los datos sufren una serie de transformaciones intermedias hasta obtener el tablón final, donde destacan las siguientes etapas de procesamiento de datos:

➢ **Limpieza y preparación de datos**

Básicamente esta etapa consiste en limpiar los datos crudos para cada base de datos descritas en la sección anterior (Cliente, Transaccional, Sociodemográfico, Temperatura, Precios de la Competencia, Macroeconómicas) y realizar operaciones tales como:

- Normalización de fechas en formato año-mes- día en la zona horaria UTC tanto en la base de clientes como en la transaccional.
- Cambio de tipos de tipos de variables, como por ejemplo castear variable de tipo string a datetime.
- Remediación de problemas de calidad de los datos: valores fuera de rango, identificadores duplicados, datos inconsistentes, errores de encoding, etc.

➢ **Creación de Features**

Esta etapa consistió en crear una features que no dependan del tiempo, y otras features temporales. Para el caso de las features temporales se crearon funciones para calcular ratios

[8] https://kedro.org/

de variable en relación al comportamiento de la misma aplicando funciones analíticas en ventanas móviles de 3, 6, y 12 meses.

- **Input del Modelo**

La última etapa fue la construcción del tablón final a partir de las features creadas en el punto anterior.

Definición de cliente

Definimos como cliente el hogar en donde se consumen cilindros de gas para diferentes casos de uso. Así, varias personas que compran directamente a LIPIGAS usando números telefónicos diferentes pueden ser considerados parte del mismo hogar si la dirección a la cual hacen el pedido es la misma. La finalidad es conocer y retener el nivel de consumo de un hogar más que de una persona en particular.

Validación de direcciones para la identificación de un cliente

Dada la definición de cliente asumida por el modelo es necesario la validación de las direcciones de los clientes disponibles en el sistema de call center de la compañía. La importancia radica en que si queremos identificar con la mayor precisión cada cliente luego la dirección asociada a cada pedido debe estar normalizada y ser invariante a través del tiempo. Esto permitiría al modelo hacer seguimiento de cada uno de los clientes a través del tiempo.

Construcción de Features

La creación de nuevas variables permite mejorar el poder de predictibilidad de los modelos de aprendizaje de máquinas, dado que estos son capaces de aprender patrones de estas variables permitiendo predecir la fuga de clientes con un mayor grado de asertividad.

Las variables fueron calculadas con periodicidad mensual, y se agrupan en las siguientes categorías:

Transaccionales:

- mes_1: variable dicotómica con valor 1 si la fecha es el mes 1, 0 de de otra forma.
- mes_2: variable dicotómica con valor 1 si la fecha es el mes 2, 0 de de otra forma.
- mes_3: variable dicotómica con valor 1 si la fecha es el mes 3, 0 de de otra forma.
- mes_4: variable dicotómica con valor 1 si la fecha es el mes 4, 0 de de otra forma.
- mes_5: variable dicotómica con valor 1 si la fecha es el mes 5, 0 de de otra forma.
- mes_6: variable dicotómica con valor 1 si la fecha es el mes 6, 0 de de otra forma.
- mes_7: variable dicotómica con valor 1 si la fecha es el mes 7, 0 de de otra forma.
- mes_8: variable dicotómica con valor 1 si la fecha es el mes 8, 0 de de otra forma.
- mes_9: variable dicotómica con valor 1 si la fecha es el mes 9, 0 de de otra forma.
- mes_10: variable dicotómica con valor 1 si la fecha es el mes 10, 0 de de otra forma.
- mes_11: variable dicotómica con valor 1 si la fecha es el mes 11, 0 de de otra forma .

- mes_12: variable dicotómica con valor 1 si la fecha es el mes 12, 0 de de otra forma.
- precio_kg_mes: valor comprado en el mes con respecto a los kilos comprados en el mes.
- valor_mes: valor comprado en el mes.
- valor_mes_pp: valor comprado en el mes con respecto al valor comprado en los últimos 12 meses.
- valor_ventana_3M_pp: valor comprado en los últimos 3 meses con respecto al valor comprado en los últimos 12 meses.
- valor_ventana_6M_pp: valor comprado en los últimos 6 meses con respecto al valor comprado en los últimos 12 meses.
- kilos_mes: kilos comprados en el mes
- kilos_mes_pp: kilos comprados en el mes con respecto a los kilos comprados en los últimos 12 meses.
- kilos_ventana_3M_pp: kilos comprados en los últimos 3 meses con respecto a los kilos comprados en los últimos 12 meses.
- kilos_ventana_6M_pp: kilos comprados en los últimos 6 meses con respecto a los kilos comprados en los últimos 12 meses.
- descuento_pesos_mes: descuento en el mes.
- descuento_pesos_mes_pp: descuento en el mes con respecto al total de compras en pesos en el mes.
- descuento_pesos_mes_pp_2: descuento en el mes con respecto al descuento en los últimos 12 meses.

- descuento_pesos_ventana_3M_pp: descuento en los últimos 3 meses con respecto al descuento en los últimos 12 meses.
- descuento_pesos_ventana_6M_pp: descuento en los últimos 6 meses con respecto al descuento en los últimos 12 meses.
- tenure: tenure del cliente en días.
- tam_ventana: tamaño de la ventana de compra (IPT_AVG+IPT_SD).
- ipt_avg: periodicidad de compra promedio.
- ipt_sd: periodicidad de compra desviación estándar.
- ipt_cv: periodicidad de compra coeficiente de variación.
- recencia_sobre_ipt_avg: recencia relativa a periodicidad de compra promedio.
- compras_12_meses: número de compras en los últimos 12 meses.

Para el caso de las variables precio_kg_mes y descuento_pp_mes se imputaron los nulos con valor cero.

Servicio:

- frustados_17_mes: número de pedidos frustrados en el mes.
- frustados_17_mes_pp: pedidos frustados en el mes con respecto a las compras en los últimos 12 meses.
- frustados_17_ventana_3M_pp: pedidos frustados en los últimos 3 meses con respecto a las compras en los últimos 12 meses.
- frustados_17_ventana_6M_pp: pedidos frustrados en los últimos 6 meses con respecto a las compras en los últimos 12 meses.

- frustados_17_ventana_12M_pp: pedidos frustados en los últimos 12 meses con respecto a las compras en los últimos 12 meses.
- anulados_7_mes: número de pedidos anulados en el mes.
- anulados_7_mes_pp: pedidos anulados en el mes con respecto a las compras en los últimos 12 meses.
- anulados_7_ventana_3M_pp: pedidos anulados en los últimos 3 meses con respecto a las compras en los últimos 12 meses.
- anulados_7_ventana_6M_pp: pedidos anulados en los últimos 6 meses con respecto a las compras en los últimos 12 meses.
- anulados_7_ventana_12M_pp: pedidos anulados en los últimos 12 meses con respecto a las compras en los últimos 12 meses.

Temperatura:

- temp_mean_1_month: Temperatura promedio en el último mes.

Macroeconómicas:

- dolar_diario: valor promedio de la tasa de cambio en relación al valor de dólar en el último mes.
- ipc_mensual: índice de precios al consumidor promedio en el último mes.
- desempleo_mensual_lag_3M: desempleo del anterior trimestre.
- imacec_mensual: índice IMACEC mensual.
- pib_trimestral: PIB trimestral.

Sociodemográficas:

- casas: porcentaje de casas en la zona censal.
- personas: número de personas en la zona censal.
- mujeres: porcentaje de mujeres en la zona censal.
- de_0_a14: porcentaje de población en edad de 0-14 años.
- de_15_a_64: porcentaje de población en edad de 15-64 años.
- de_65_mas: porcentaje de población en edad mayor a 65 años.
- ing_prom_percapita_casen: ingreso promedio per cápita de la zona censal reportado al CASEN.
- ing_prom_casen: ingreso promedio por zona censal.
- ing_prom_hogar_casen: ingreso promedio del hogar por zona censal.
- eq_renta_efx: Renta promedio por cliente reportada a Equifax.
- abc1: porcentaje de hogares con estrato ACB1.
- d: porcentaje de hogares con estrato D.
- s/c: porcentaje de hogares con estrato S/C.
- c2: porcentaje de hogares con estrato C2.
- c3: porcentaje de hogares con estrato C3.
- e: porcentaje de hogares con estrato E.

Precios competencia

- abastible_s_a_11kg_normal_res: diferencia entre el precio de lista de ABASTIBLE y el precio de lista de LIPIGAS para el formato 11 kg.

- gasco_glp_s_a_11kg_normal_res: diferencia entre el precio de lista de GASCO y el precio de lista de LIPIGAS para el formato 11 kg.
- abastible_s_a_15kg_normal_res: diferencia entre el precio de lista de ABASTIBLE y el precio de lista de LIPIGAS para el formato 15 kg.
- gasco_glp_s_a_15kg_normal_res: diferencia entre el precio de lista de GASCO y el precio de lista de LIPIGAS para el formato 15 kg.

Análisis de Correlaciones de Pearson

Existen dos caminos a usar la correlación de Pearson cómo método de selección de características, la primera es medir la correlación entre las variables y la segunda es medir la correlación entre la fuga y las variables independientes

Correlación entre pares de variables

Dado que contamos con 104 variables independientes se optó por crear un mapa de calor de correlaciones con agrupamiento jerárquico "clustermap" (figura 3.1) ,que es una forma mucho más eficiente de visualizar la gran cantidad de correlaciones existentes, y donde se puede distinguir zonas de pares de coeficientes de correlación entre variables con diferentes grados de correlación:

- No existe Correlación: los pares de variables son independientes una de la otra, es decir, una no se superpone a la otra, y el coeficiente de correlación varía entre −0.2 a 0.2. Se puede apreciar este tipo inexistencia de correlación de forma predominantemente en casi toda la matriz en zonas con tono (gris claro).

- Correlaciones positivas y negativas débiles: el coeficientes de correlación varía entre 0.2 a 0.4 para correlaciones positivas, y −0.4 a -0.2 para las correlaciones negativas. Las correlaciones positivas débiles destacan con un tono (celeste claro) en la diagonal de la matriz, mientras que las correlaciones negativas destacan tono (beige) en las esquinas de la matriz.

- Correlaciones positivas y negativas: el coeficientes de correlación varía entre 0.4 a 0.7 para correlaciones positivas, y −0.7 a -0.4 para las correlaciones negativas. Las correlaciones positivas destacan con un tono (azul claro) en la diagonal de la matriz, mientras que las correlaciones negativas destacan tono (salmón) en las esquinas de la matriz.

- Correlaciones positivas y negativas fuertes: el coeficientes de correlación varía entre 0.7 a 1.0 para correlaciones positivas fuertes, y −1.0 a −0.7 para las correlaciones negativas fuertes. Las correlaciones positivas fuerte destacan con un tono (azul oscuro) , en la diagonal de la matriz, mientras que las correlaciones negativas fuertes destacan tono (café) en las esquinas de la matriz.

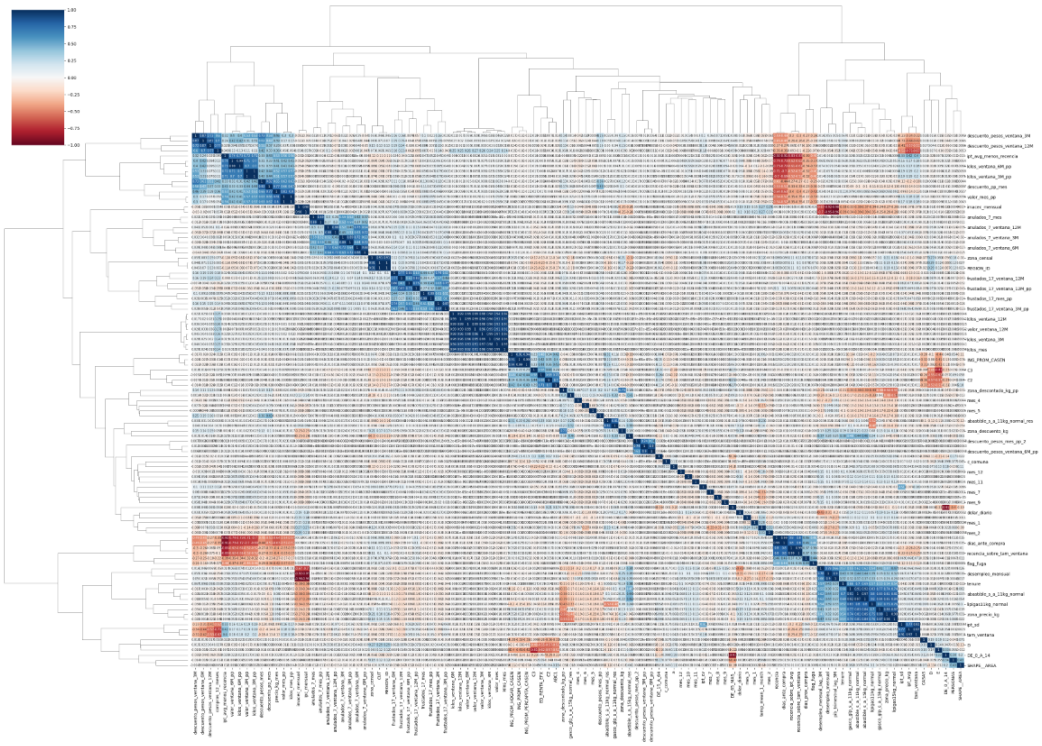

Figura 3.1 : Mapa de calor de Correlaciones con Agrupamiento Jerárquico

Como se puede ver en la figura 3.1, las correlaciones fuertes, tanto positivas (color azul oscuro) como negativas (color café) podrían ocasionar problemas de colinealidad en los modelos. Una situación no deseable en la que una de las variables independientes es una función lineal de otra variable independiente. Además, si estas variables altamente correlacionadas se usan para entrenar un modelo de aprendizaje de máquinas pueden hacer que las predicciones de los parámetros del modelo se tornen inestables.

Las correlaciones entre pares de variables son las siguientes:

- Correlaciones Fuertes Positivas entre pares de variables: el coeficiente de correlación varía entre 0.7 a 1.0 , Tabla 1 (Anexo A)

- Correlaciones Fuertes Negativas entre pares de variables: el coeficiente de correlación varía entre -1.0 a - 0.7, Tabla 2 (Anexo A)
- Correlaciones Positivas entre pares de variables: el coeficiente de correlación varía entre 0.4 to a 0.7, Tabla 3 (Anexo A)
- Correlaciones Negativas entre pares de variables: el coeficiente de correlación varía entre - 0.7 a -0.4 , Tabla 4 (Anexo A)
- Correlaciones Positivas Débiles entre pares de variables: el coeficiente de correlación varía entre 0.2 a 0.4, Tabla 5 (Anexo A)
- Correlaciones Negativas Débiles entre pares de variables: el coeficiente de correlación varía entre −0.4 a −0.2, Tabla 6 (Anexo A)
- No existe Correlación entre pares de variables: el coeficiente de correlación varía entre −0.2 a 0.2, Tabla 7 (Anexo A)

Correlación de la variable dependiente con respecto a las variables independientes

Se realizó la prueba de correlación de Pearson entre la variable fuga respecto a las variables independientes para medir el grado de correlación lineal entre ellas (figura 3.2).

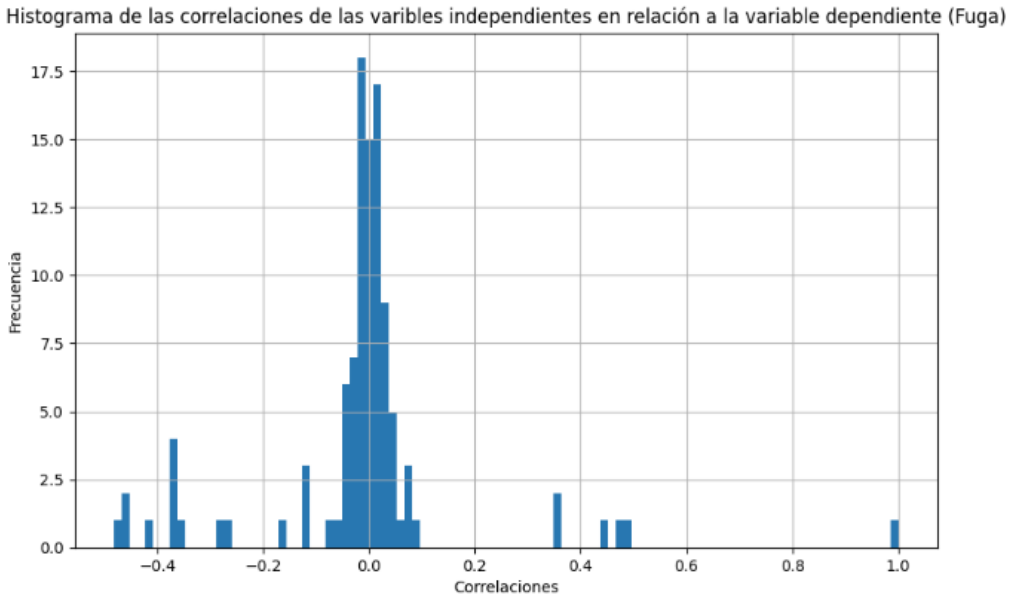

Figura 3.2: Histograma de las correlaciones de las variables independientes en relación a la variable dependiente (Fuga)

De las 104 variables se desprenden los siguientes tipos de correlaciones :

- El 83.8% de ellas no están correlacionadas [-0.2 a 0.2] con la variable fuga (figura 3.3).
- El 8.5% presenta correlaciones bajas [-0.2 a -0.4] y [0.2 a 0.4] respecto a la variable fuga (figura 3.4).
- El 2.8% presenta correlaciones fuertes [-0.4 a -0.7] y [0.4 a 0.7] respecto a la variable fuga (figura 3.5).

Los resultados demuestran que el 83.8% presentan correlaciones bajas con la fuga, el 8.5% de las variables presentan correlaciones moderadas o bajas, y el 2.8% (porcentaje bajo de variables) está fuertemente correlacionadas con el target (Tabla 3.1).

	Correlación
dias_ante_compra	0.899871
recencia	0.898806
recencia_sobre_ipt_avg	0.855166
recencia_sobre_tam_ventana	0.854571
dias_prox_compra	0.831039
temp_mean_1_month	0.276997
zona_censal	0.223148
CUT	0.216903
REGION_ID	0.214827
mes_2	0.212352

Tabla 3.1 : Las 10 variables más correlacionadas con la fuga

Selección de Características

La selección de características es una técnica en la que elegimos los mejores predictores para la variable objetivo (fuga) y cuyas principales ventajas son:

- Modelos más simples: que son fáciles de explicar; un modelo que es demasiado complejo e inexplicable no es valioso
- Reduce el tiempo de entrenamiento: al tener un conjunto reducido de variables el tiempo de entrenamiento se acorta..

- Reducción de la varianza: aumentar la precisión de las estimaciones al contar con variables con baja varianz se pueden obtener para una simulación dada
- Mejora la generalización del modelo: al reducir el sobreajuste

En el presente trabajo aplicaremos el método (Basado en Filtro) para selección de características en base a una métrica estadística, dentro de las cuales las más usadas para comprender las variables más relevantes son : la prueba de Información Mutua , la prueba de Chi-cuadrado, la prueba de Fisher, el coeficiente de correlación y el umbral de varianza.

Prueba de Información Mutua

En el contexto de selección de características el test de Información Mutua se utiliza para medir el grado de dependencia entre variables aleatorias que tienen correlaciones no lineales entre ellas.

La información mutua entre dos variables aleatorias X e Y se puede definir formalmente a través de la siguiente ecuación:

$$IM(X,Y) = H(X) - H(X|Y)$$

donde IM(X; Y) es la información mutua de X e Y, H(X) es la entropía de X y H(X | Y) es la entropía condicional de X dado Y.

El resultado se mide en unidades de bits. Mientras mayor sea este valor , mayor es el grado de dependencia entre las dos variables. Si el resultado calculado es cero, entonces quiere decir que las variables son independientes.

Aplicando el test de Información Mutua del módulo “SelectKBest” de la librería sklearn de Python a las variables dependientes y seteando el parámetro k igual a 60, se obtuvieron 60 variables con el mayor grado de correlación no lineal respecto a la fuga.

3.2. Metodología

Formulación matemática

El modelo de fuga fue creado como un modelo de probabilidad utilizando técnicas de aprendizaje automático, en particular aprendizaje supervisado y algoritmos de clasificación binaria. Matemáticamente, el modelo de fuga puede ser descrito como una función f que interpreta diferentes características de un cliente, que encapsulamos en X, para predecir una probabilidad P de fuga representada por variable dependiente Y. Esta probabilidad de fuga generalmente es un evento identificable a partir de los datos.
El modelo determina la probabilidad condicional:

$$P(Y/X) = f(X)$$

En particular, para nuestro caso tenemos que:

$$P(Y/información\ del\ cliente) = f\left(transaccional,\ servicio,\ macroeconómicas,\ precios\ relativos,\ contexto,\ temperatura\right)$$

Donde y, es la variable target del modelo que define las fuga mediante:

$$y = \{1\ \ si\ cliente\ se\ fuga\ en\ la\ próxima\ compra\ ,\ 0\ \ si\ cliente\ no\ se\ fuga\ en\ la\ próxima\ compra\}$$

Métricas

Las predicciones de los modelos se evaluarán según distintas métricas que aplican para los problemas de clasificación binaria como es el caso de la fuga de clientes.

En primer lugar, se evaluarán (contrastarán) los modelos por medio de la matriz de confusión que se muestra en la Tabla 3.1

Observación\Predicción	Negativo	Positivo
Negativo	Verdaderos Negativos (TN)	Falso Positivo (FP)
Positivo	Falso Negativo (FN)	Verdaderos Positivos (TP)

Tabla 3.2 : Matriz de Confusión

En esta matriz se visualizan cuatro escenarios posibles:

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

De la matriz de confusión se desprende una variedad de métricas como las que se mencionan a continuación.

El recall (sensibilidad) corresponde al porcentaje de fuga (etiquetas positivas) que el modelo es capaz de identificar, es decir:

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

La precisión nos permite evaluar la calidad de la predicción, estableciendo el porcentaje de clientes que el modelo predijo como fuga y que realmente resultaron ser así, es decir:

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

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

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

Otras métricas para comparar los distintos modelos es la Curva ROC (acrónimo de Receiver Operating Characteristic) y AUC (Area Under the Curve). La curva de ROC

mide la sensibilidad frente a la especificidad en una mtraiz de confusión para un clasificador binario según se varía el umbral de discriminación (valor a partir del cual decidimos que un caso es un positivo). Mientras que AUC que es área debajo de la curva de presenta el grado o medida de separabilidad entre la clase, e indica cuánto es capaz el modelo de distinguir entre clases. Cuanto mayor sea el AUC, mejor será el modelo para predecir la fuga.

Interpretación de variables

El método conocido como Shap Analysis [16] es framework para interpretación de la influencia de las variables independientes en la predicción de un modelo de aprendizaje de máquinas, además proporciona visualizaciones interactivas e intuitivas que muestran qué variables tienen mayor importancia por separado para una predicción determinada y para el modelo en general.

Definición de fuga

Uno de los problemas más difíciles de abordar en los modelos de comercialización no contractuales, es la falta de una buena definición de fuga. Por el contrario, en los modelos de comercialización contractuales, definir la fuga es una tarea simple, basta que un cliente deje de pagar la suscripción por el servicio para que dicha acción se considere como un evento de fuga. Sin embargo, para los modelos no contractuales no hay un evento observable para definirlo como fuga, también es probable que el evento de fuga no se observe en la venta de tiempo estudio, sino que la fuga ocurra en un evento futuro. En estadística a este fenómeno se le conoce como datos (censorados a la derecha), esto quiere decir que en un instante de tiempo de análisis no es factible saber si un evento ha ocurrido.

Definimos fuga si un cliente se separa más allá respecto al promedio de días entre compra más una desviación estándar del número de días entre compras.

Así matemáticamente un cliente se fuga si:

$$dias\_ante\_compra + dias\_prox\_compra > \alpha_1 \times IPT\_AVG + \alpha_2 \times IPT\_SD$$

Donde $dias\_ante\_compra$ es el número de días desde la anterior compra hasta la fecha de corrida del modelo, $dias\_prox\_compra$ el número de días desde la fecha de corrida del modelo hasta su próxima compra, IPT_AVG es el número de días promedio entre compras (periodicidad) en el último año y IPT_SD la desviación estándar del número de días entre compras.

De forma complementaria, un cliente no se fuga si:

$$dias\_ante\_compra + dias\_prox\_compra < \alpha_1 \times IPT\_AVG + \alpha_2 \times IPT\_SD$$

Intuitivamente, podemos interpretar la parte izquierda de la inecuación como el número de días entre compras, mientras que la parte derecha representa el tamaño esperado de una ventana de compra de un cliente, definida en función de la periodicidad promedio de compra y la desviación estándar de estas. Así, en otras palabras, *un cliente se fuga cuando el tiempo que tarda entre compras es mayor al tiempo promedio que esperábamos más una leve variabilidad evidenciada en su data histórica*. La importancia de esta definición de fuga radica en que estamos creando una definición personalizada de fuga para cada cliente de acuerdo con su consumo típico proveniente de la información transaccional.

En la definición de ventana de compra vale la pena aclarar la importancia de cada uno de sus componentes y los parámetros que se representan allí. Aquí α_1 y α_2 son de vital importancia ya que nos permiten calibrar en el modelo el tipo de fuga que queremos

identificar más adelante en producción. A la fecha de creación de este documento definimos:

$$tam\_ventana = IPT\_AVG + IPT\_SD$$

O de otra forma asignamos $\alpha_1 = 1$, $\alpha_2 = 1$. Intuitivamente al definir estos parámetros estamos construyendo como fuga todo el momento cuando un cliente compra después del tiempo esperado que teníamos para su próxima compra, o del mismo modo, el cliente pierde (no realiza) su próxima compra esperada. Sin embargo, al tener una definición de ventana parametrizable podemos identificar diferentes tipos de fuga con modificar los parámetros. Por ejemplo, si quisiéramos identificar la fuga como los momentos en que un cliente que pierde 3 compras esperadas de forma consecutiva, podríamos usar calibrar $\alpha_1 = 3$.

Validación de la Definición de Fuga

Para la determinación de los parámetros α_1 y α_2, se realizó un análisis del comportamiento de fuga en base a los datos transaccionales de los clientes considerando un período de observación desde Enero 2020 a Diciembre 2021, con un total de 3.964.079 de transacciones para 404.462 clientes, y donde se pudo observar que la tasa de fuga en relación al total de transacciones decrece (tabla 3.3), en la medida que α_1 crece.

alpha_1	alpha_2	total_fugas_vistas	fuga_sobre_trx
1	1	1916210	0.483393
2	1	1144444	0.288704
3	1	878839	0.221701

Tabla 3.3: Porcentaje de fuga en respecto total de transacciones para $\alpha_1\{1, 2, 3\}$ y α_2=1 constante

Otro enfoque es ver cómo se comporta la fuga total observada y la fuga promedio en el mismo período de observación desde Enero 2020 a Diciembre 2021, y donde se puede concluir que:

- Se captura una mayor volumen de *fuga promedio* con α_1=1, decrece con α_1=2, y llega al mínimo con α_1=3 (figura 3.6, 3.7 y 3.8) .
- La forma de curva de *fuga promedio* es similar para α_1=1 y α_1=2 , esto a pesar de que para α_1=3 la forma de la curva es distinta (figura 3.6, 3.7 y 3.8).
- Al observar la forma de la curva de *fuga promedio* se puede apreciar estacionalidades en el los meses de invierno y verano para α_1=1 y α_1=2 , no así para α_1=3 (figura 3.6, 3.7 y 3.8).

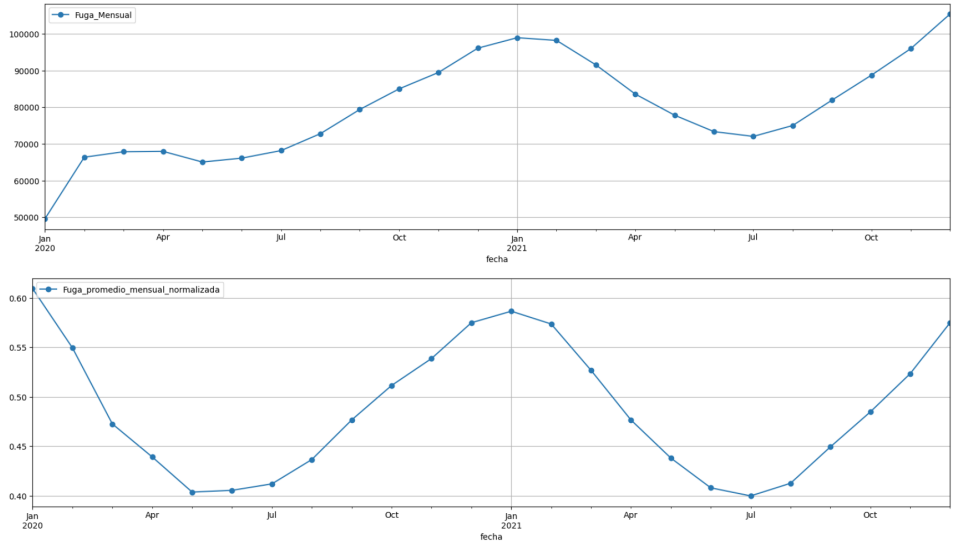

Figura 3.6: Gráfico superior: "Total Fuga" y Gráfico Inferior: "Fuga Promedio", entre los años 2020 y 2021, para $\alpha_1 = 1$ y $\alpha_2 = 1$,

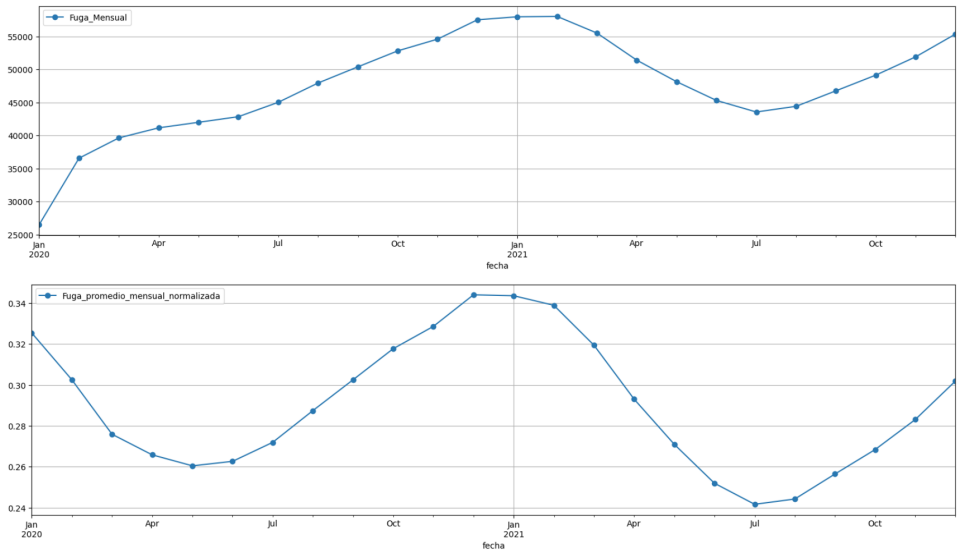

Figura 3.7: Gráfico superior: "Total Fuga" y Gráfico Inferior: "Fuga Promedio", entre los años 2020 y 2021, para α_1=2 y $\alpha_2 = 1$

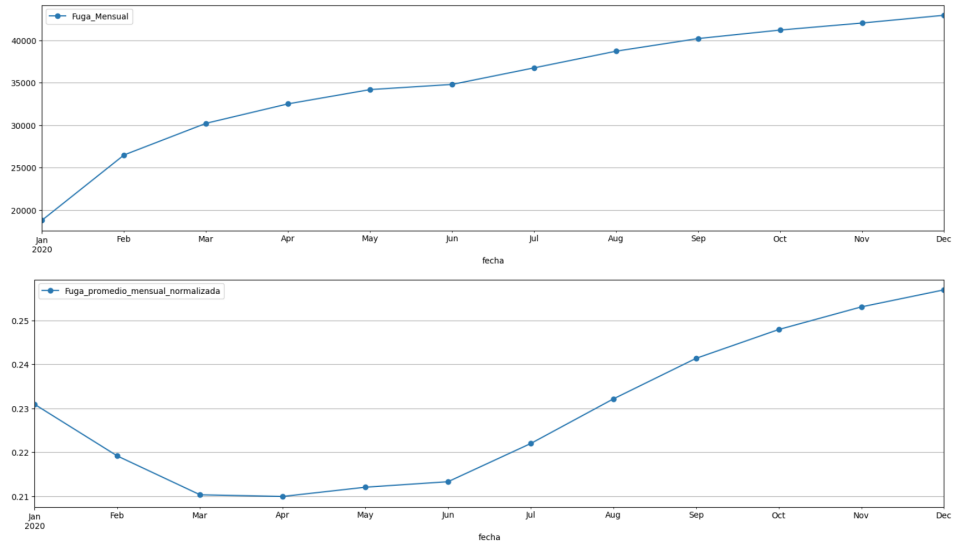

Figura 3.8: Gráfico superior: "Total Fuga" y Gráfico Inferior: "Fuga Promedio", entre los años 2020 y 2021, para α_1=3 y $\alpha_2 = 1$

Tomando en consideración el análisis previo, se optó por definir la fuga tomando $\alpha_1 = 1$ y $\alpha_2 = 1$:

$$dias\_ante\_compra + dias\_prox\_compra > \alpha_{1=1} \times IPT\_AVG + \alpha_{2=1} \times IPT\_SD$$

La elección de $\alpha_1 = 1$ implica además la aplicación de un criterio comercial de la empresa para capturar mayor volumen de fuga en el corto plazo; y así gatillar acciones comerciales agresivas de retención de clientes.

Cálculo del Tamaño de la Ventana Promedio de Compras

Se calculó la distribución de frecuencia promedio de compras entre los años 2020 y 2021, dando como resultado un promedio de días entre compras de 42 días y una desviación estándar del número de días entre compras de 47 días.

Tomando la definición del tamaño de la venta promedio de compra, tenemos que:

$$tam\_ventana = IPT\_AVG + IPT\_SD = 89\ días \sim 3\ meses$$

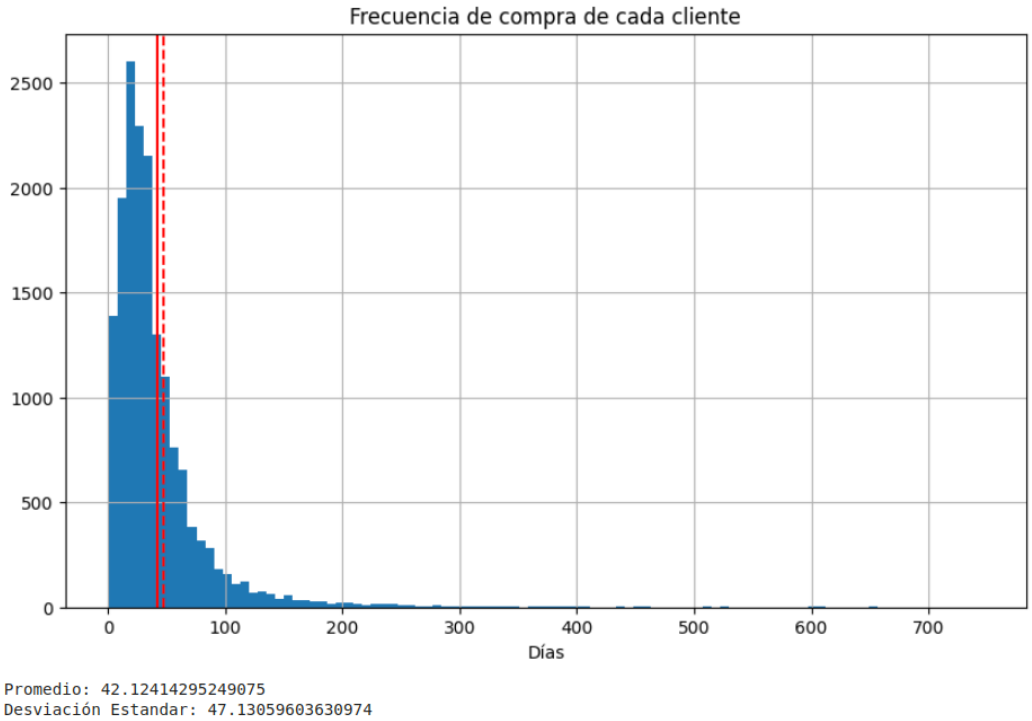

Figura 3.9: Frecuencia de compra de cada cliente

Definición de lealtad y validez

Definimos un cliente válido como todo aquel que tiene el número de compras suficientes sobre una ventana de tiempo (un año típicamente) para poder crear un patrón de compra. Por ejemplo, para el modelo del formato de 11 y 15 Kg definimos como válidos los clientes que tienen al menos 4 o más compras en una ventana de 12 meses.

Adicionalmente a validez, definimos clientes leales como todos los clientes que muestran un patrón de regularidad de compra estable en su historia. Esto con el fin de saber que, dada esa regularidad, podemos construir una definición de fuga como todos los momentos donde el cliente se aleja de su regularidad típica de compra. Para el caso de uso del modelo de 11 y 15 Kg definimos lealtad como todos los momentos donde un cliente cumple que:

$$IPT\_CV < percentile\_0.65(distribución\ de\ IPT\_CV)$$

Donde IPT_CV es el coeficiente de variación de la periodicidad de compra del cliente y $distribución\ de\ IPT\_CV$ es la distribución de los coeficientes de variación de todos los clientes en todos los meses del análisis (información de Enero 2020 hasta Marzo 2022). En particular el coeficiente de variación se define como:

$$IPT\_CV = \frac{IPT\_SD}{IPT\_AVG}$$

Descomposición de clientes para el análisis

La descomposición de los clientes (figura 3.10) consideró cuatro subgrupos:

- Clientes Totales: es todo aquel cliente que haya efectuado alguna transacción en el período de análisis (Enero 2020 a Marzo 2022).

- Clientes Válidos: del universo de clientes se consideran válidos aquellos con al menos 4 compras en la historia de análisis.
- Clientes Leales: son aquellos clientes que siendo válidos, tienen un comportamiento de compra regular según la siguientes definición matemática:

$$IPT\_SD \ / \ IPT\_AVG < quantil(65\%)$$

- Clientes Fugados: dentro del conjunto de clientes leales que tienen un comportamiento de compra regular, definimos a los clientes fugados como aquellos que teniendo compras regulares no ha comprado dentro de una ventana de tiempo esperada.

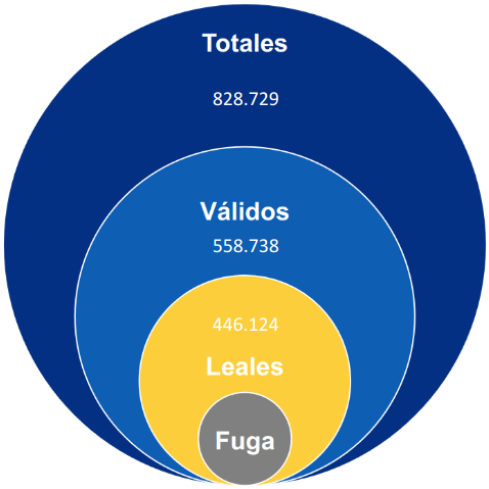

Figura 3.10 : Descomposición de clientes para análisis

4. Resultados

A fin de poder evaluar cómo se comportan los modelos para predecir la fuga, se procederá a dividir la base de datos que se obtuvo de la sección *datos* en los conjuntos de entrenamiento y testeo, como se puede ver en la Tabla 4.1. Para el particionamiento de los conjuntos de entrenamiento y testeo se tomó un sampleo aleatorio de 5.000 clientes para la correr los modelos, equivalente al 1.16% del total de clientes.

Datos	Periodo de Entrenamiento	Cantidad de datos
Train	Enero 2020 - Agosto 2021	3,492
Test	Enero 2021 - Marzo 2021	535

Tabla 4.1 : Descripción del conjunto de entrenamiento y testeo para el modelo de fuga

El particionamiento del conjunto de datos toma en consideración tanto la dimensión temporal de los datos como el muestreo aleatorio de clientes únicos que el modelo verá tanto en la etapa de entrenamiento como testeo, para ello se crearon dos grupos aleatorios de clientes (figura 4.1), el primero con un 80% de los clientes para entrenar el modelo, y el segundo el 20% restante para testear el modelo, donde el porcentaje de fuga presente en el conjunto de entrenamiento es del 48.19 %, mientras que conjunto de testeo es del 57.54 % (la fuga está balanceada).

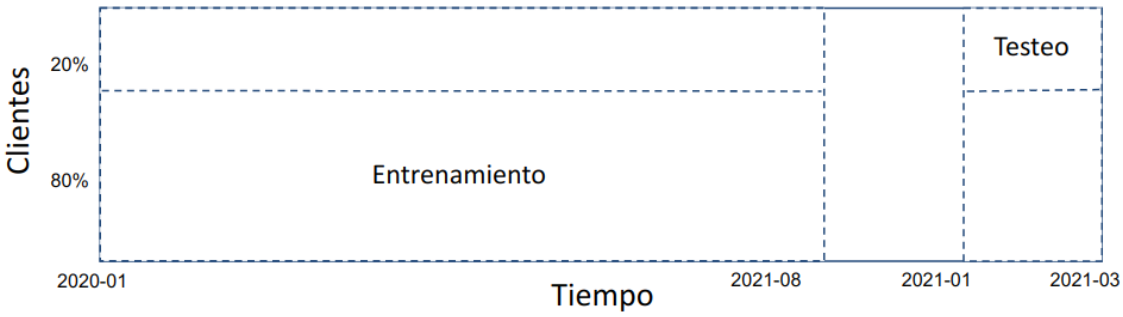

Figura 4.1: Particionamiento temporal del conjunto de entrenamiento y testeo

El primer conjunto de entrenamiento abarca un período de 20 meses, desde Enero 2020 a Agosto 2021. El fin de considerar más de 1 año observaciones tiene como objetivo incluir el efecto de la estacionalidad de la demanda, donde la demanda de gas residencial crece producto de las bajas temperatura, entre los meses de Abril a Agosto.

El segundo conjunto de testeo abarca un período de 3 meses, desde Enero 2021 a Marzo 2021. Esta ventana de tiempo de 3 meses corresponde al tamaño promedio de compras calculado a partir de los datos (Cálculo de la Frecuencia Promedio de Compras).

Entre el conjunto entrenamiento y testeo se consideró una (ventana ciega) de 3 meses, desde Septiembre 2021 a Diciembre 2021, que es el mismo tamaño de ventana que se usó en el testeo. Esta venta ciega donde el modelo tiene como propósito que el modelo pierda la memoria y no aprenda a predecir la fuga en dicha lapsus de tiempo consiguiendo así un mejor desempeño en la predicción de fuga.

Contraste de Modelos

En este apartado trabajaremos con el lenguaje de programación Python para ejecutar distintos modelos de aprendizaje de máquinas, los cuales fueron explicados en el apartado de *introducción* del presente informe.

Se prueban siete algoritmos de machine learning con el fin de evaluar el desempeño de cada uno en el problema planteado. Estos algoritmos son:

- Linear SVC (Máquina de Soporte Vectorial con Kernel Lineal)
- Logistic Regression
- RandomForest
- LightGBM
- Stochastic gradient descent
- XGBoost
- Decision Tree

Como parte de la etapa de modelado y experimentación se utilizaron distintas combinaciones de variables para entrenar los modelos, los cuales desempeñaban demasiado bien (ROC (AUC) ~ 98 %) en testeo, por lo que aplicando métodos de interpretabilidad de las predicciones estas mostraron que todo el peso de la predicción recae en la variable " dias_proxima_compra", básicamente porque la variable fuga por definición está construida a partir de la variable "dias_proxima_compra", por lo tanto contiene su información y es como si fuese la misma variable, es decir, una variable haciendo predicciones de sí misma.

Al descartar la variable “dias_proxima_compra” del conjunto de variables el peso de la predicción se distribuyó en más variables.

Luego evaluamos las predicciones de los 7 modelos en los datos disponibles de la base de testeo, cuyos resultados se muestran en la Tabla 4.2, donde se contrastan los modelos según 5 métricas distintas.

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
LGBMClassifier	0.75	0.77	0.77	0.75	0.43
RandomForestClassifier	0.73	0.75	0.75	0.73	16.97
LinearSVC	0.73	0.74	0.74	0.73	9.27
SGDClassifier	0.72	0.74	0.74	0.73	0.46
LogisticRegression	0.72	0.74	0.74	0.72	0.31
XGBClassifier	0.72	0.74	0.74	0.73	2.11
DecisionTreeClassifier	0.65	0.64	0.64	0.65	2.10

Tabla 4.2 : Resultados en métricas para modelos de predicción de fuga. Se evalúan los modelos según el área bajo la curva ROC (AUC)

La selección del mejor modelo se basó en aquel que presentó mejor desempeño en base a métrica ROC (AUC), por lo que se optó por seleccionar al modelo LightGBM ya que presentó un desempeño del 77% significativamente mejor en comparación con el resto de los modelos.

Ajuste de Hiperparámetros y Métricas de desempeño

Para el entrenamiento se utilizó validación cruzada con 5 particiones, únicamente sobre el conjunto de entrenamiento, y se ajustaron de los hiperparametros mediante grid search, maximizando la métrica ROC (AUC), y se definió un espacio de búsqueda para 5 combinaciones aleatorias de los siguientes hiperparámetros del modelo LightGBM:

- Regularización alpha: Regularización norma L1.
- Regularización lambda: Regularización norma L2.
- Tasa de aprendizaje: boosting learning rate .
- Máxima profundidad: Describe la profundidad máxima del árbol.
- Número de hojas: Este es el parámetro principal para controlar la complejidad del modelo de árbol. Idealmente, el valor del número de hojas debería ser menor o igual a 2^(Máxima profundidad). Un valor grande produce problemas sobreajuste del modelo.

Los hiperparámetros del mejor clasificador fueron los siguientes :

- Regularización alpha: 0.1
- Regularización lambda: 0.5
- Tasa de aprendizaje: 0.01
- Máxima profundidad: 5
- Número de hojas: 10

La para curva de la ROC (AUC) del modelo fue de 82 % (figura 4.2).

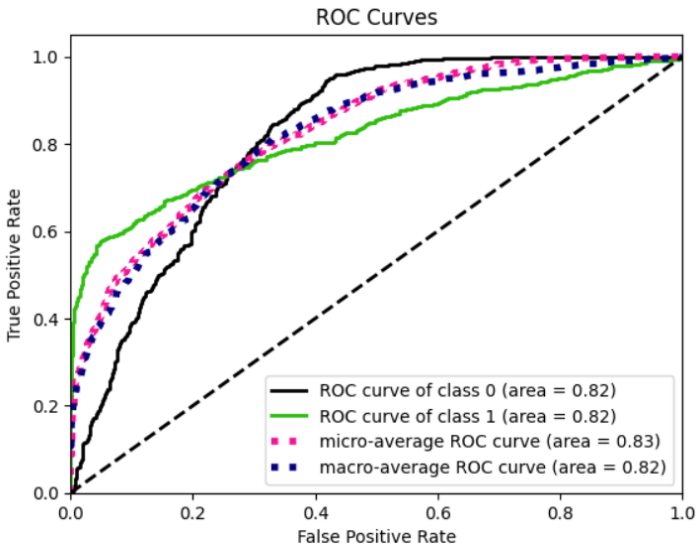

Figura 4.2: curva ROC sobre la clasificación de fuga (class 1) y no fuga (class 0)

Mientras que la ganancia del modelo (figura 4.3) en el primer decil (tabla 4.3) es más propenso a fugarse (que posee una tasa de fuga esperada de un 16%). Esto quiere decir, que al contactar al 10% de los clientes con mayor probabilidad de fuga en la base de clientes, el modelo es capaz de capturar el 16% de la fuga futura.

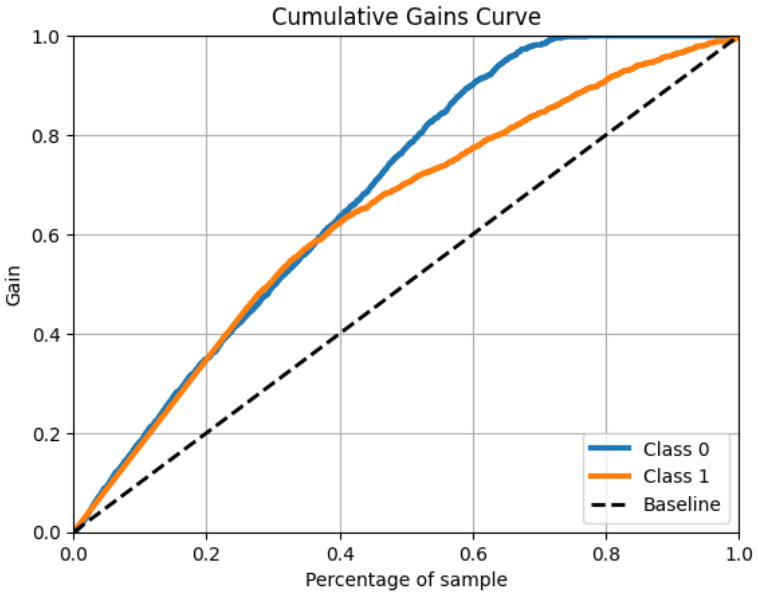

Figura 4.3 : Curva de Ganancia sobre la identificación de fuga (class 1) respecto al baseline aleatorio

Decil	% de Fuga
10%	0.16
20%	0.32
30%	0.47
40%	0.63
50%	0.68
60%	0.75
70%	0.81
80%	0.87
90%	0.93

Tabla 4.3 : Porcentaje de fuga observada por decil

Interpretabilidad del modelo

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 de fuga que el modelo entrega a cada cliente/mes.

A partir de las contribuciones marginales de cada variable, se confeccionó la figura 4.2 donde se muestra el promedio (en valor absoluto) de estas contribuciones en la predicción del modelo. En este gráfico resalta el gran peso que tiene la variable "recencia_sobre_tam_ventana" (recencia relativa al tamaño de una ventana de compra de un cliente) en la predicción del modelo, en segundo lugar le sigue la "recencia_sobre_ipt_avg" (recencia relativa a periodicidad de compra promedio) y tercer lugar la "ipt_avg_menos_recencia" (periodicidad de compra promedio menos la recencia).

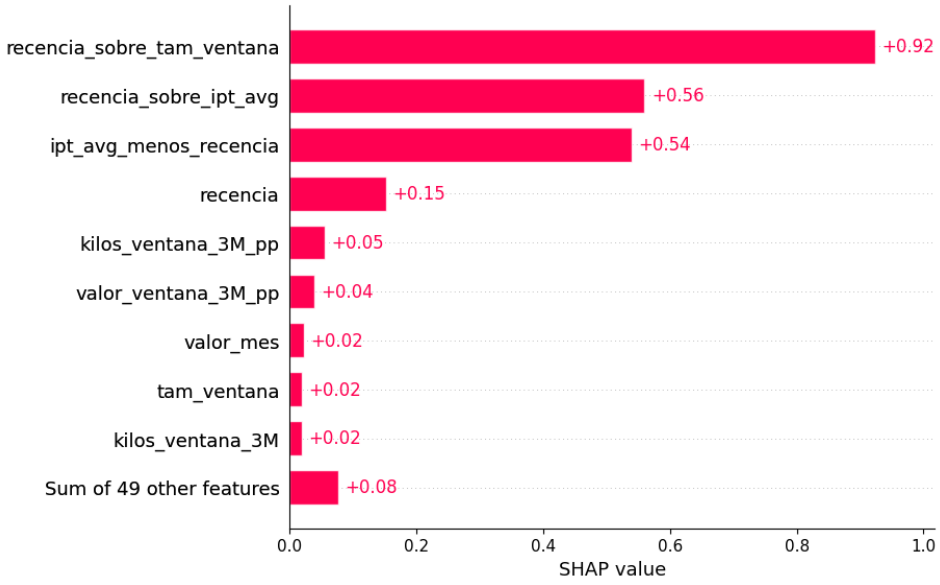

Figura 4.2: Influencia marginal promedio de variables (en valor absoluto) en modelo LightGBM

En la figura 4.3 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. Las variables cuando toman valores altos (rojos) se asocia a un aumento de la probabilidad de fuga, mientras que sólo para valores bajos (azul) se asocia a una reducción de la probabilidad de fuga. De este gráfico se destaca el efecto que tiene la recencia y otras variable derivadas de ella, como la "recencia_sobre_tam_ventana" y la "recencia_sobre_ipt_avg", en donde si han transcurrido pocos días desde la última compra la probabilidad de fuga disminuye, y esta última aumenta en caso contrario.

Otra variable relevante es la "temp_mean_1_month" (temperatura promedio en el último mes) , donde a menor temperatura decrece la probabilidad de fuga , y por el contrario cuando la temperatura aumenta la probabilidad de fuga también aumenta.

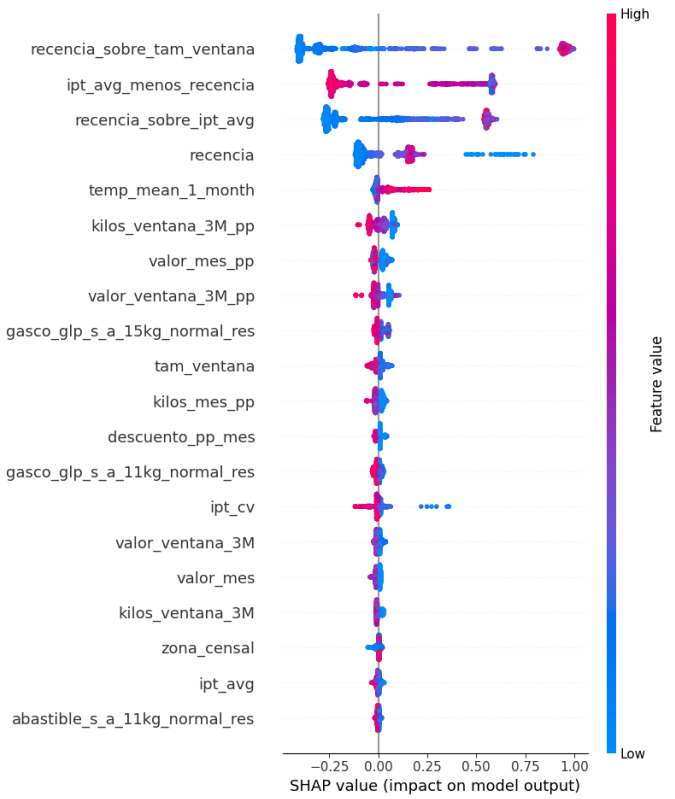

Figura 4.3: Influencia marginal promedio de variables en modelo LightGBM

5. Conclusiones

Uno de los objetivos del presente trabajo fue definir la fuga y predecir la probabilidad condicional de fuga para escenarios futuros no vistos por el modelo (predecir el futuro). De los 7 modelos de clasificación evaluados y entrenados para un subconjunto de clientes, el modelo de LightGBM fue el que tuvo mejor desempeño en base la métrica ROC (AUC), y luego de aplicar un ajuste de hiperparámetros el modelo presentó una ROC (AUC) de 82% sobre el conjunto de prueba, por lo que se consideró el modelo óptimo para el proyecto.

Respecto a la interpretabilidad del modelo, el estudio de los shap values del modelo LightGBM entregó, de manera gráfica y sencilla, información sobre el efecto de las variables en la clasificación. Así, se obtuvieron hallazgos interesantes como por ejemplo, el alto peso de la variable la recencia (número de días desde la última compra) y las variable derivadas de ella como la (recencia relativa al tamaño de una ventana de compra de un cliente), en la predicción del modelo, destacando que a mayor es el número de días desde la última compra la probabilidad de fuga aumenta, y esta disminuye en caso contrario. El alto protagonismo de la variable recencia se debe a información de esta variable esta contenida en la propia definición de fuga, por lo tanto la variable dependiente contiene su información y se produce el efecto de que las variables derivadas de la recencia están haciendo predicciones de sí misma.

Como trabajo futuro se propone productivizar el modelo, y ejecutarlo a fin de cada mes, obteniendo así las probabilidades de fuga para de cada cliente, con esta información el área comercial de la empresa puede priorizar aquellos clientes que tengan mayor probabilidad de fuga (primer decil), de manera que se puedan focalizar las acciones comerciales de retención más eficientemente en este subconjunto de clientes.

También se propone construcción de nuevas variables y profundizar en el estudio de selección de características para evaluar los efectos en el poder de predictibilidad del modelo.

Por otro último, puesto que la empresa en el último tiempo ha destinado esfuerzos para incorporar la captura de RUT's de los clientes en el procesos de compra, se propone cambiar la definición de cliente hogar (donde la la finalidad es retener el nivel de consumo de un hogar) a retener el consumo de un cliente en particular por medio del RUT.

Bibliografía

[1] Xiaohang Zhang, Ji Zhu, Shuhua Xu, and Yan Wan. Predicting customer churn through interpersonal influence. Knowledge-Based Systems, 28:97–104, 2012.

[2] Hans Risselada, Peter C Verhoef, and Tammo HA Bijmolt. Staying power of churn prediction models. Journal of Interactive Marketing, 24(3):198–208, 2010.

[3] Kristof Coussement, Dries F Benoit, and Dirk Van den Poel. Improved marketing decision making in a customer churn prediction context using generalized additive models. Expert Systems with Applications, 37(3):2132–2143, 2010.

[4] Xiaohang Zhang, Ji Zhu, Shuhua Xu, and Yan Wan. Predicting customer churn through interpersonal influence. Knowledge-Based Systems, 28:97–104, 2012.

[5] Chih-Fong Tsai and Mao-Yuan Chen. Variable selection by association rules for customer churn prediction of multimedia on demand. Expert Systems with Applications, 37(3):2006–2015, 2010.

[6] Hans Risselada, Peter C Verhoef, and Tammo HA Bijmolt. Staying power of churn prediction models. Journal of Interactive Marketing, 24(3):198–208, 2010.

[7] Zhao Xin, Wang Yi, and Cha Hong-wang. A mathematics model of customer churn based on pca analysis. In Computational Intelligence and Software Engineering,2009. CiSE 2009. International Conference on, pages 1–5. IEEE, 2009.

[8] Yaya Xie, Xiu Li, EWT Ngai, and Weiyun Ying. Customer churn prediction using improved balanced random forests. Expert Systems with Applications, 36(3):5445–5449, 2009.

[9] Koen W De Bock and Dirk Van den Poel. An empirical evaluation of rotation based ensemble classifiers for customer churn prediction. Expert Systems with Applications, 38(10):12293–12301, 2011.

[10] Gregory, B. Predicting Customer Churn: Extreme Gradient Boosting with Temporal Data. arXiv 2018, arXiv:1802.03396.

[11] Mohammed Abdul Haque Farquad, Vadlamani Ravi, and S Bapi Raju. Churn prediction using comprehensible support vector machine: An analytical crm application. Applied Soft Computing, 19:31–40, 2014.

[12] Jin Xiao, Yi Xiao, Anqiang Huang, Dunhu Liu, and Shouyang Wang. Feature selection based dynamic transfer ensemble model for customer churn prediction. Knowledge and Information Systems, 43(1):29–51, 2015.

[13] Troncoso et al., Predicción de fuga de clientes en una empresa de distribución de gas.

[14] Chence Shi, Zheye Deng, Yewen Xu, Weiping Song, Yichun Yin, Jile Zhu, Ming Zhang. Ensembling XGBoost and Neural Network for Churn Prediction with Relabeling and Data Augmentation.

[15] Gregory, B. Predicting Customer Churn: Extreme Gradient Boosting with Temporal Data. arXiv 2018, arXiv:1802.03396.

[16] 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.

Anexos

Anexo A: Tabla de Correlaciones entre pares de variables

par	coef_corr
(CUT, REGION_ID)	0.999079
(kilos_mes_pp, valor_mes_pp)	0.996539
(kilos_ventana_3M_pp, valor_ventana_3M_pp)	0.995130
(kilos_ventana_6M_pp, valor_ventana_6M_pp)	0.994725
(kilos_ventana_12M, valor_ventana_12M)	0.994462
(kilos_ventana_6M, valor_ventana_6M)	0.993261
(valor_ventana_12M, valor_ventana_6M)	0.992539
(kilos_ventana_12M, kilos_ventana_6M)	0.991062
(kilos_ventana_3M, valor_ventana_3M)	0.990983
(recencia_sobre_ipt_avg, recencia_sobre_tam_ve...	0.990508

Tabla 1: Correlaciones Fuertes Positivas entre pares de variables

par	coef corr
(imacec_mensual, pib_trimestral_lag_3M)	-0.980597
(ipc_mensual, pib_trimestral_lag_3M)	-0.963712
(desempleo_mensual, ipc_mensual)	-0.946538
(ipt_avg_menos_recencia, recencia)	-0.924120
(desempleo_mensual, imacec_mensual)	-0.920418
(dias_ante_compra, ipt_avg_menos_recencia)	-0.913150
(DE_0_A_14, DE_65_MAS_)	-0.908102
(ipt_avg_menos_recencia, recencia_sobre_ipt_avg)	-0.829408
(ipt_avg_menos_recencia, recencia_sobre_tam_ve...	-0.826474
(C3, E)	-0.810539

Tabla 2: Correlaciones Fuertes Negativas entre pares de variables

par	coef_corr
(gasco_glp_s_a_15kg_normal, lipigas11kg_normal)	0.697683
(abastible_s_a_15kg_normal, tenure)	0.691110
(C2, EQ_RENTA_EFX)	0.683799
(descuento_pp_mes, kilos_mes_pp)	0.667712
(desempleo_mensual_lag_3M, pib_trimestral_lag_3M)	0.664048
(descuento_pp_mes, valor_mes_pp)	0.662395
(gasco_glp_s_a_11kg_normal, tenure)	0.656672
(descuento_pesos_mes, precio_kg_mes)	0.650288
(anulados_7_ventana_12M, anulados_7_ventana_6M...	0.639141
(anulados_7_ventana_3M_pp, anulados_7_ventana_6M)	0.637810

Tabla 3 : Correlaciones Positivas entre pares de variables

par	coef_corr
(dias_ante_compra, valor_ventana_3M_pp)	-0.699853
(desempleo_mensual_lag_3M, ipc_mensual)	-0.696021
(dias_ante_compra, kilos_ventana_3M_pp)	-0.694606
(desempleo_mensual_lag_3M, imacec_mensual)	-0.674829
(compras_12_meses, tam_ventana)	-0.670486
(precio_kg_mes, recencia)	-0.640045
(dias_ante_compra, precio_kg_mes)	-0.626809
(recencia_sobre_ipt_avg, valor_ventana_6M_pp)	-0.619657
(recencia_sobre_tam_ventana, valor_ventana_6M_pp)	-0.618913
(E, EQ_RENTA_EFX)	-0.615682

Tabla 4: Correlaciones Negativas entre pares de variables

par	coef_corr
(desempleo_mensual, lipigas11kg_normal)	0.389884
(descuento_pesos_ventana_6M, precio_kg_mes)	0.387045
(desempleo_mensual, gasco_glp_s_a_11kg_normal)	0.383373
(abastible_s_a_15kg_normal_res, tenure)	0.382996
(frustados_17_mes, frustados_17_ventana_6M_pp)	0.377997
(anulados_7_mes_pp, anulados_7_ventana_6M)	0.373209
(abastible_s_a_11kg_normal, pib_trimestral_lag...	0.372240
(frustados_17_mes_pp, frustados_17_ventana_6M)	0.370460
(pib_trimestral_lag_3M, tenure)	0.369942
(dias_prox_compra, recencia_sobre_tam_ventana)	0.368473

Tabla 5: Correlaciones Positivas Débiles entre pares de variables

par	coef_corr
(precio_kg_mes, recencia_sobre_ipt_avg)	-0.398004
(precio_kg_mes, recencia_sobre_tam_ventana)	-0.395357
(dias_prox_compra, ipt_avg_menos_recencia)	-0.392906
(abastible_s_a_11kg_normal, ipc_mensual)	-0.392807
(abastible_s_a_15kg_normal, zona_descontado_kg...	-0.392132
(ipc_mensual, tenure)	-0.391107
(abastible_s_a_15kg_normal, ipc_mensual)	-0.387423
(abastible_s_a_11kg_normal, imacec_mensual)	-0.380066
(imacec_mensual, tenure)	-0.377527
(abastible_s_a_11kg_normal, zona_descontado_kg...	-0.376461

Tabla 6: Correlaciones Negativas Débiles entre pares de variables

par	coef_corr
(dolar_diario, ipc_mensual)	0.199917
(DE_0_A_14, SHAPE__AREA)	0.195603
(descuento_pesos_ventana_6M, frustados_17_vent...	0.191521
(tenure, zona_descuento_kg)	0.190990
(ipt_avg, kilos_ventana_3M_pp)	0.190742
(CASAS, E)	0.190713
(descuento_pesos_ventana_6M, frustados_17_vent...	0.190673
(descuento_pesos_ventana_12M, frustados_17_ven...	0.190663
(ipt_avg, valor_ventana_3M_pp)	0.190603
(ipt_sd, kilos_ventana_3M_pp)	0.189590

Tabla 7: No existe Correlación entre pares de variables

Anexo B: Histogramas de correlaciones

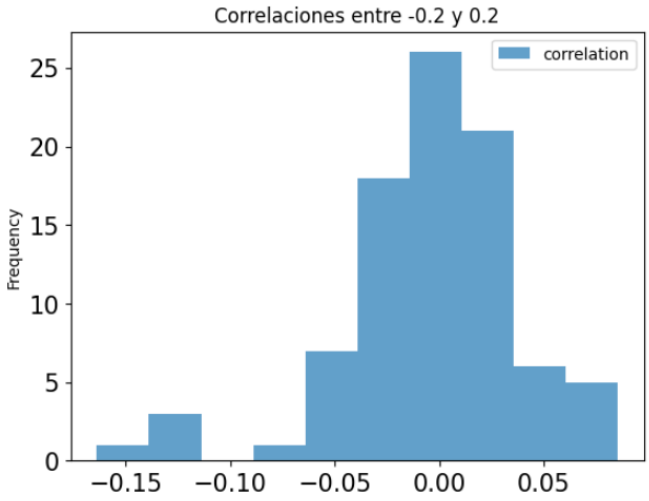

Figura 3.3: Correlaciones entre -0.2 y 0.2

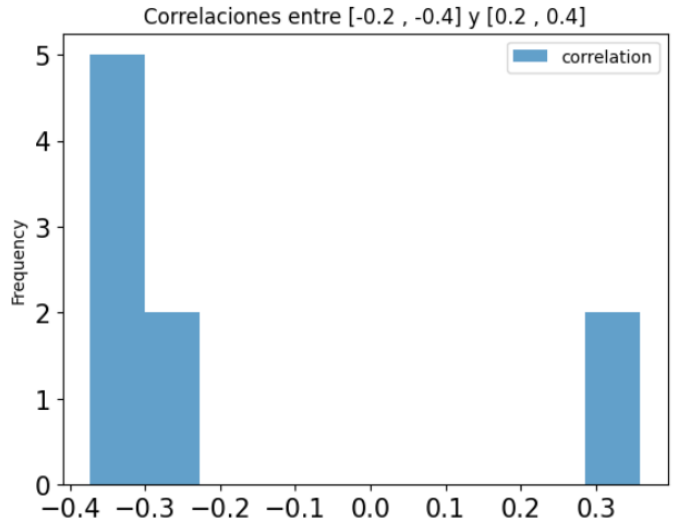

Figura 3.4: Correlaciones entre [-0.2 , -0.4] y [0.2 , 0.4]

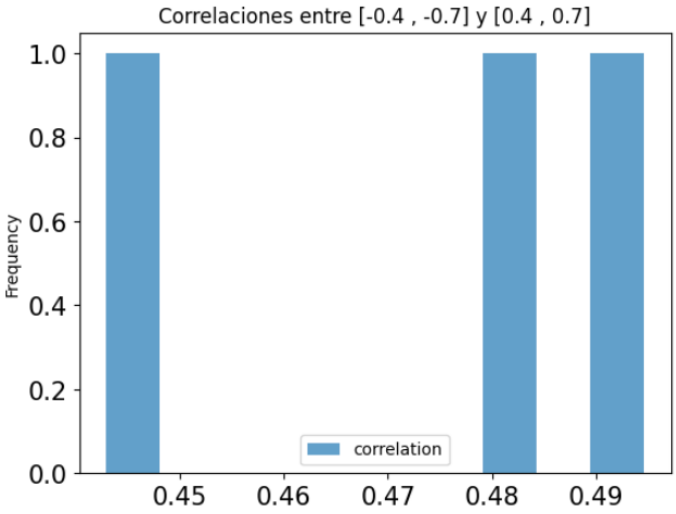

Figura 3.5: Correlaciones entre [-0.4 , -0.7] y [0.4 , 0.7]