Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Negocios y Administración Pública

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL DE ESPECIALIZACIÓN

Pérdida de clientes en una empresa mayorista de alimentos. Un modelo de predicción para un entorno B2B no contractual.

AUTOR: PABLO PAGANO
MENTOR: FACUNDO SANTIAGO

[SEPTIEMBRE 2022]

Resumen

Los modelos de predicción de pérdida de clientes tienen como objetivo detectar clientes con una alta propensión a la fuga. Es bien sabido, que a una organización le resulta mucho menos costoso retener un cliente que ganar uno nuevo. Construir un modelo de predicción de pérdida de clientes en entornos mayoristas B2B (*business-to-business*) con relaciones no contractuales, es desafiante puesto que el número de clientes es más reducido que en entornos B2C (*business-to-consumer*), y, adicionalmente, es complejo determinar cuándo abandonan la empresa, ya que no depende de una rescisión contractual. A ello se suma, el hecho de que es un tema poco abordado por la literatura, por lo que este trabajo, puede convertirse en un aporte relevante al conocimiento de la temática. El objetivo es desarrollar un modelo para predecir la pérdida de clientes en una empresa dedicada al comercio mayorista de alimentos. A tal fin, se compara la sensibilidad e interpretabilidad en la predicción de los modelos *random forest* y regresión logística. El estudio se realiza sobre los clientes que compraron durante el año 2021.

Palabras clave: Pérdida de clientes, Empresa mayorista, Modelos predictivos, Entorno B2B no contractual, Clasificación.

Índice

Introducción

Uno de los propósitos comerciales fundamentales de una empresa es lograr mantener o superar los niveles de ventas a lo largo del tiempo. Para ello es clave lograr una fidelización de los clientes actuales, y simultáneamente tratar de captar clientes nuevos, reconociendo el hecho de que a una organización le resulta mucho menos costoso retener un cliente que ganar uno desconocido. En el actual contexto de grandes volúmenes de datos, que se generan tanto dentro como fuera de la organización, existen potencialidades para quienes están dispuestos a analizar la información en búsqueda de soluciones para sus problemas de negocio. La problemática de la pérdida de clientes no es nueva, pero a raíz del fenómeno *Big Data* y la aparición del aprendizaje automático, existen modelos de predicción de pérdida de clientes que pueden detectar clientes con una alta propensión a la fuga.

El entorno en el que se aborda este estudio está caracterizado por ser mayorista, con relaciones B2B (*business-to-business*), sin vínculos contractuales con los clientes. Esta temática, escasamente abordada por la literatura, presenta un verdadero desafío, dado que el número de clientes es más reducido en comparación con entornos B2C (*business-to-consumer*), sumado a que es complejo determinar cuándo abandonan la empresa, ya que no depende de una rescisión contractual.

Dado el planteo del problema, el objetivo de este trabajo es encontrar el mejor modelo para predecir la pérdida de clientes en una empresa mayorista de alimentos, que se desempeña en un entorno B2B con relaciones no contractuales. Las opciones más populares en los estudios de predicción de fuga de clientes son la regresión logística, las máquinas de vectores de soporte (SVM) y los modelos de árboles, en particular *random forest* (De Caigny et al. 2018). Se espera que *random forest* sea el modelo con mejor perfomance en la predicción, a pesar de que probablemente no sea sencillo interpretar la importancia de las variables en la detección de la fuga de clientes. En relación con la medida de performance, no existe una métrica de oro que se pueda utilizar siempre, por lo cual es importante definir la métrica teniendo en cuenta el problema a resolver, es decir, considerando el contexto de negocio en el que finalmente será utilizado (Santiago, 2021).

Una vez seleccionado el mejor modelo para el problema de negocio en cuestión, en este caso la pérdida de clientes, se buscará identificar las variables relevantes en la predicción, se hará un análisis de los errores que comete cada modelo, y se visualizarán los resultados

de las predicciones de una manera clara y completa para los analistas de negocios de la empresa.

En consonancia al objetivo general planteado, el trabajo se estructurará de la siguiente forma. En primer lugar, se analizará el contexto *Big Data*, la predicción de pérdida de clientes en empresas mayoristas con relaciones B2B no contractuales y la relación de la problemática de pérdida de clientes con el modelo de negocios de la empresa bajo estudio. En segundo lugar, se describirá el proceso de generación del conjunto de datos y elección de las medidas de performance del modelo en el contexto de negocio. En tercer lugar, se hará la selección de modelos, la comparación de performance de cada uno, el análisis de los errores que cometen en las predicciones y la visualización de los resultados. Por último, se expondrán las conclusiones del trabajo.

El contexto *Big Data*, la predicción de pérdida de clientes en empresas mayoristas con relaciones B2B no contractuales y la relación de la problemática de pérdida de clientes con el modelo de negocios de la empresa bajo estudio

El objetivo de este apartado es analizar el contexto *Big Data*, desarrollar la problemática de predicción de pérdida de clientes en empresas mayoristas con relaciones B2B no contractuales y relacionar está problemática con el modelo de negocio de la empresa bajo estudio. Para ello, primero se analiza el contexto actual de disponibilidad de grandes volúmenes de datos de distintos tipos y velocidades, luego se desarrolla la problemática de predecir la pérdida de clientes en general y se examinan las particularidades del entorno mayorista, con relaciones "negocio a negocio" sin mediar ningún acuerdo escrito entre las partes, y por último, se relaciona la problemática de la pérdida de clientes con el modelo de negocio de la empresa mayorista de alimentos bajo estudio.

1.1. El contexto *Big Data*

Big Data (en español, grandes datos o grandes volúmenes de datos) es un término evolutivo que describe cualquier cantidad voluminosa de datos estructurados, semiestructurados y no estructurados que tienen el potencial de ser extraídos para obtener información.

En la actualidad, las empresas, conviven con distintas fuentes de información. Dado que tienen velocidades de generación y características disimiles, su captura se hará con la herramienta informática apropiada, según de qué tipo de dato se trate. Al respecto, existen del tipo no estructurado, como los comentarios en redes sociales, que se generan en tiempo real, mientras que simultáneamente hay del tipo estructurado, organizados en tablas, como el historial de transacciones de ventas, que son capturados por lotes. Adicionalmente, la tipología y el tamaño influye en la manera en que deben almacenarse. Mientras que, en un principio, para almacenar datos estructurados un *Data Warehouse* era lo más apropiado, actualmente, el auge de los del tipo no estructurado ha introducido la necesidad de componentes más flexibles como el *Lakehouse*.

Luego pueden suceder dos cosas: Que estos queden almacenados, como algo estático o rancio, cuya utilidad ya desapareció desde el momento en que se registró la transacción, o, por el contrario, ser usados como materia prima del negocio para crear una nueva forma de valor económico. Según Mayer-Schönberger y Cukier (2013) en la práctica, con la perspectiva

adecuada, los datos pueden reutilizarse inteligentemente para convertirse en un manantial de innovación y servicios nuevos. Según Verbeke et al. (2011) la minería de datos implica el proceso general de extraer conocimiento de estos. Las técnicas de minería de datos han tenido diversas aplicaciones como la detección del cáncer de mama en el sector biomédico, el análisis de la canasta de compra en el sector minorista (Berry y Linoff, 2004) y la calificación crediticia en el sector financiero (Baesens et al., 2003), como así también en la predicción de pérdida de clientes.

1.2. La problemática de la predicción de pérdida de clientes

Los modelos de predicción de abandono de clientes tienen como objetivo detectar aquellos con una alta propensión a la fuga y conocer las causas que motivan dicho comportamiento. Luego, esto permite diseñar y aplicar campañas de retención de clientes mucho más efectivas.

La retención de clientes es rentable para una empresa porque atraer nuevos clientes cuesta de cinco a seis veces más que retener clientes (Athanassopoulos, 2000; Bhattacharya, 1998; Colgate y Danaher, 2000); los clientes a largo plazo generan mayores ganancias y tienden a ser menos sensibles a las actividades de marketing de la competencia, luego, se vuelven menos costosos de atender y pueden proporcionar nuevas referencias a través del "boca en boca" positivo, mientras que los clientes insatisfechos pueden difundir el "boca en boca" negativo (Colgate et al., 1996; Ganesh et al, 2000); la pérdida de clientes genera costos de oportunidad debido a la reducción de las ventas (Rust y Zahorik, 1993). En consecuencia, una pequeña mejora en la retención de clientes puede conducir a un aumento significativo de las ganancias (Van den Poel y Larivière, 2004).

Existe un cuerpo sólido de literatura sobre opciones de modelado para la predicción de pérdida de clientes. Sin embargo, la literatura sobre aplicación de dichos modelos para la predicción y retención de clientes en la práctica es escasa y hasta ahora solo se han informado los resultados de unos pocos experimentos de campo (Gattermann-Itschert y Thonemann, 2021).

Burez y Van den Poel (2007) utilizan un modelo de predicción de abandono para identificar a los clientes con la mayor probabilidad de fuga de una empresa de televisión paga. Prueban tres medidas de retención en el campo que reducen la tasa de abandono. Preguntar a los clientes sobre su satisfacción logra el mayor efecto.

Ascarza et al. (2016, 2017) investigan la prevención proactiva de fuga en la industria de telecomunicaciones. No usan predicciones de fugas, sino que seleccionan clientes en función de otros criterios. Por ejemplo, se dirigen a clientes que se beneficiarían de cambiar su plan si

su patrón de uso continúa (Ascarza et al. 2016) o clientes cuyas cuentas se suspenden en breve (Ascarza et al. 2017).

Ringbeck et al. (2019) investigan las relaciones no contractuales en el comercio minorista. Llevan a cabo un experimento de campo a gran escala con 400.000 clientes, dirigidos al 10% superior de los clientes con las más altas probabilidades de fuga predichas en el grupo de tratamiento. Encuentran que la gestión proactiva de fuga de clientes con cupones disminuye la tasa de abandono y aumenta los ingresos.

Los estudios mencionados anteriormente consideran entornos B2C. El análisis de datos se ha utilizado ampliamente para respaldar el *customer relationship management* (CRM) en entornos B2C y ha recibido menos atención en entornos B2B (Wiersema 2013).

1.3. Particularidades del entorno mayorista con relaciones B2B no contractuales

Según el Diccionario panhispánico del español jurídico de la Real Academia Española (en línea), comercio mayorista es la actividad desarrollada profesionalmente con ánimo de lucro consistente en la adquisición de productos para su reventa a otros comerciantes o a empresarios para su incorporación en el proceso de producción o prestación de servicios. De esta definición se deriva que el comprador de los productos es otro comercio, de allí que las relaciones sean del tipo “negocio a negocio”, o B2B por su sigla en inglés.

En comparación con B2C, la cantidad de clientes en B2B suele ser pequeña (Lilien 2016), lo que hace que entrenar un modelo de predicción de pérdida de clientes y ejecutar un experimento de campo sea más desafiante.

Por otro lado, en el entorno de la empresa mayorista bajo estudio, no existen contratos que respalden las relaciones comerciales con los clientes. Esta situación, le otorga un tinte particular al análisis de la pérdida de clientes

Los ciclos de vida de los clientes son cada vez más transitorios debido al severo impacto de las acciones de los competidores en las relaciones existentes. Esto puede acentuarse en entornos no contractuales, debido a que los clientes tienen la oportunidad de cambiar continuamente su comportamiento de compra sin informar a la empresa al respecto. Además, en estos entornos los clientes no experimentan ningún costo de cambio al cambiar de proveedor (Reinartz y Kumar, 2000).

En muchos sectores, el cambio de comportamiento se define como deserción total. Los clientes cierran sus cuentas bancarias, cambian de operador de internet, TV o telefonía. En estas

industrias es fácil observar cuando se produce la deserción: las personas interrumpen totalmente su relación con la empresa. Como estas empresas se encuentran en un entorno contractual, pueden determinar el momento exacto en el que los clientes interrumpen su relación (Buckinx y Van den Poel, 2005).

En sectores donde no existen contratos, como el de la empresa bajo estudio, es más complejo determinar cuándo se van los clientes, ya que no necesariamente detienen abruptamente la compra a la empresa mayorista, sino que a menudo compran menos o reducen la frecuencia de sus pedidos gradualmente. La disminución sustancial en los volúmenes de compra se conoce como fuga parcial, que es el foco de predicción de los estudios en entornos no contractuales (Buckinx & Van den Poel 2005, Miguéis et al. 2012, 2013).

En este estudio se define como abandono parcial a la situación en la que un cliente deja de comprar por un período de cuatro meses o más. En dicho caso, el cliente es considerado como perdido.

1.4. Relación de la problemática de pérdida de clientes con el modelo de negocio de la empresa bajo estudio

En una empresa privada con ánimo de lucro, dedicada al comercio de alimentos al por mayor cuyo objetivo es crecer tanto en cantidad de clientes, abarcando un mercado cada vez más amplio, como en cantidad de proveedores, ofreciendo una gama cada vez más variada de productos, el buen uso de los datos es clave para logar este crecimiento. La empresa tiene tres canales de ventas bien marcados: Un canal clásico de venta personalizada gestionado por especialistas en cada segmento del cliente, y dos nuevos canales desarrollados recientemente durante la pandemia, como lo son la venta telefónica desde un *call center* interno a la empresa, y la venta web a través de un catálogo *online*. Estos tres canales están coordinados por un encargado general del departamento comercial.

La empresa también se encarga de la logística y reparto de las mercaderías, la cual en parte la realiza con una flota propia de vehículos y en parte de manera tercerizada con empresas de transporte. La propuesta de valor de la empresa consiste en que el cliente, en este caso un comercio minorista que desea comprar productos alimenticios para revender o usar como insumo en la prestación de su servicio, pueda comprar el producto que desee, de una manera ágil y cómoda por cualquiera de sus tres canales, y recibirlo en tiempo y forma de manera periódica en su negocio a un precio justo. Para mejorar aún más la experiencia del cliente, la empresa está interesada en implementar una aplicación para que cualquier cliente pueda rastrear

su pedido para saber cuándo llega. Otro punto clave en la creación de valor de esta empresa, consiste en contar con vendedores especializados en los distintos rubros del cliente. A modo de ejemplo, los hoteles, restaurants y cafés tienen asignados representantes comerciales distintos a los que atienden las panaderías y confiterías, de modo de asegurar que la experiencia de compra sea más placentera, en caso de que el cliente opte por el canal tradicional.

Si bien no se puede clasificar a esta empresa cómo un negocio de plataforma de intercambio de bienes, ya que es propietaria de los bienes que comercializa, si es un intermediario entre tres grupos de actores -clientes, proveedores, transportistas- ya que coordina las interacciones entre ellos. La relación que la compañía posee con sus clientes y proveedores se traduce en un círculo virtuoso, ya que, por una parte, es un aliado esencial para su red de proveedores que quieren llegar con sus productos a una masa de clientes distribuida en múltiples puntos geográficos, mientras que, por otra, eso ayuda a la compañía a atraer nuevos clientes quienes se benefician de la amplia oferta de productos disponible. Adicionalmente, una tercera parte en esta relación son las empresas de transporte que ofrecen prestar el servicio de flete a cambio del cobro de una comisión por tal servicio.

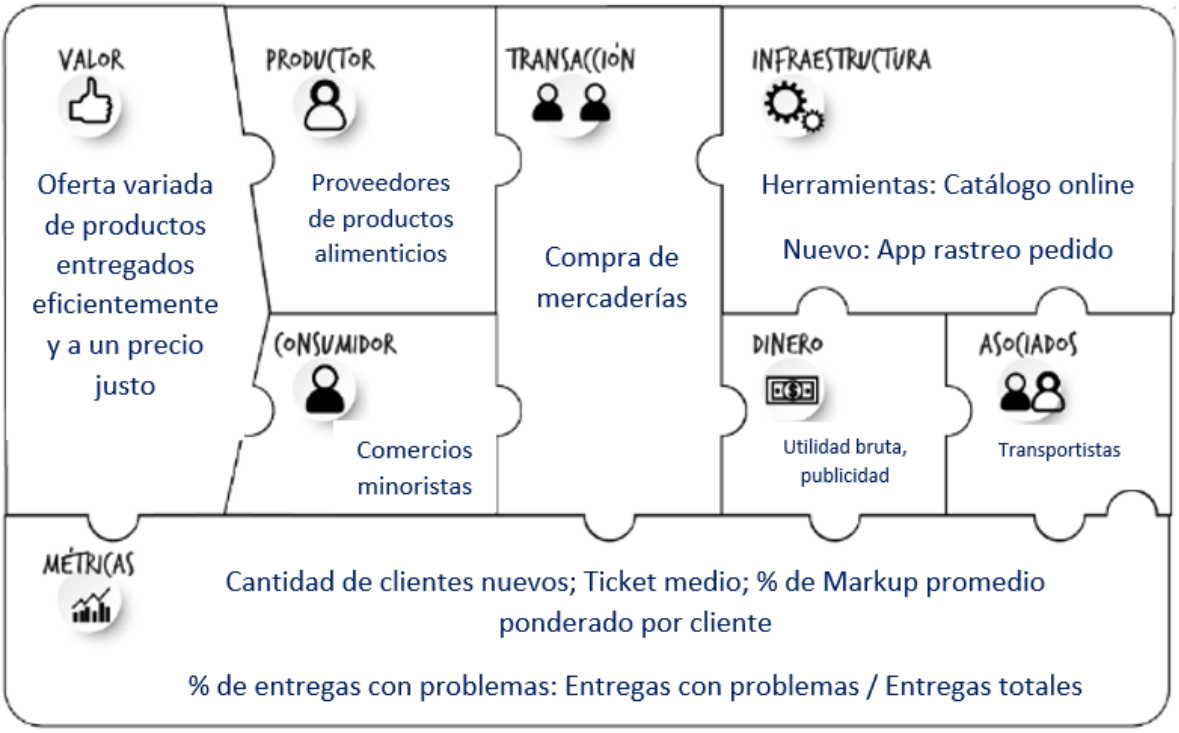
Siguiendo a Osterwalder y Pigneur (2010), se puede decir que la propuesta de valor para los clientes de la empresa consiste en “dejar el trabajo concluido” para el comercio minorista (se puede comprar por cualquiera de los tres canales, consultar por asesoramiento con un especialista, recibir las mercaderías en tiempo y forma), tener buen precio (al tener una amplia red de proveedores y un gran volumen de compra), la reducción o ahorro de costos para los clientes (al tener una amplia red de transportistas y muchos clientes en cada ruta el costo del flete baja), y la usabilidad del servicio de compra (al contar con tres canales de ventas uno de los cuales es totalmente online).

En relación a la forma en que se captura valor, al ser una empresa comercial, la ganancia se logra con la utilidad bruta, es decir, con la diferencia entre el precio de venta y el de compra que surge con la aplicación de un porcentaje de markup sobre el costo de la mercadería. Dicho porcentaje es estudiado detenidamente por el área comercial y varía en función del tamaño y segmento del cliente, como así también según la familia de producto de que se trate. Luego, en menor medida, existen ingresos por la publicidad brindada en su sitio web y catálogo online.

Según Osterwalder (2004), un modelo de negocio es una herramienta conceptual que contiene un conjunto de elementos y sus relaciones y permite expresar la lógica de una empresa para

ganar dinero. Los modelos de negocios se suelen representar con un diagrama de CANVAS. En Figura 1, se presenta el modelo de negocios de la empresa.

Figura 1. Modelo de negocio de la empresa mayorista de alimentos bajo estudio



Fuente: Elaboración propia

Para esta empresa es clave lograr que el grupo de clientes sea cada vez más grande, de modo de ir completando las distintas rutas de reparto y de esta forma economizar el costo del flete en cada entrega. También el hecho de tener un gran número de clientes favorece en aumentar el volumen de compra a los proveedores, logrando con ello la disminución de los precios de compra, lo que se traduce finalmente en disminución de precios de venta a los clientes.

A su vez, también es vital que el grupo de proveedores sea cada vez más grande, de modo que la oferta sea cada vez más variada, lo cual aumenta el ticket medio de compra. Adicionalmente, tener más de un proveedor por tipo de producto garantiza la disponibilidad de la mercadería para no perder ventas, y la competencia entre ellos asegura conseguir mejores precios de compra.

Por otro lado, también es crucial que la cartera de transportistas se vaya ampliando, ya que con ello se lograr abarcar más puntos geográficos, a un costo competitivo y con varias opciones, para elegir la que más se adapte a las necesidades del cliente.

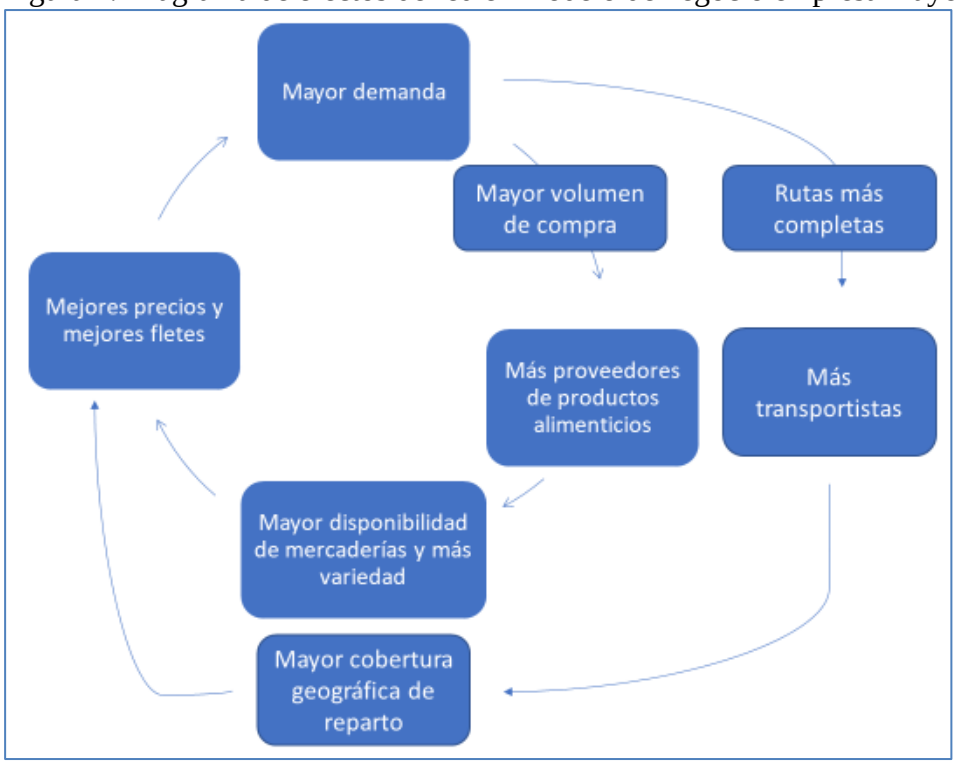
La empresa consiste en una red con tres lados o actores, donde es importante analizar los efectos de red que se producen. Los proveedores se benefician con un grupo de clientes más grande porque aumenta su poder de venta, mientras que los clientes se benefician con una cantidad de proveedores creciente porque aumenta la variedad, la disponibilidad y disminuyen los precios. Paralelamente, mientras mayor es el número de transportistas mejor es la calidad de entrega y a un costo más económico para el cliente. Todo este fenómeno, da origen a los efectos de red.

Los efectos de red refieren al valor incremental que obtiene una red por la incorporación de un nuevo usuario. En el párrafo anterior, se analizaron efectos de red que se dan entre los distintos lados de la red. Según Parker et al. (2016), estos se denominan efectos positivos de lado cruzado y se dan cuando los usuarios se benefician de un aumento en el número de participantes en el otro lado del mercado. Cabe destacar que no siempre los efectos de red son positivos.

Por ejemplo, si el número de clientes crece desproporcionadamente es probable que la mercadería no alcance para satisfacer de manera completa todos los pedidos, con la consiguiente pérdida de reputación para la empresa. A su vez, un crecimiento desmesurado de clientes en una determinada ruta de reparto puede traer aparejado que las entregas se hagan con demoras, lo cual también impacta negativamente en la imagen de la empresa. Por el lado de los proveedores, un crecimiento desordenado puede implicar que no se haga un adecuado control de calidad de las mercaderías ofrecidas, lo cual conlleva una mala experiencia del cliente. En cambio, un crecimiento no controlado de transportistas puede llevar a que se contraten empresas poco serias, lo que puede provocar problemas con las entregas. A su vez, un número muy elevado de transportistas para una misma ruta puede desincentivarlos, ya que probablemente reciban pocos fletes al mes. Estas situaciones desfavorables, se denominan efectos negativos de red y atentan contra la supervivencia de esta. Más específicamente, una de las consecuencias de estas situaciones desfavorables es la pérdida o fuga de clientes, problemática de negocio estudiada en este trabajo.

En Figura 2, se muestra un diagrama que expone el círculo virtuoso del modelo de negocio de esta empresa. Según se observa, si el crecimiento de la red está bien administrado, el ingreso de nuevos actores genera beneficios del otro lado.

Figura 2. Diagrama de efectos de red en modelo de negocio empresa mayorista de alimentos



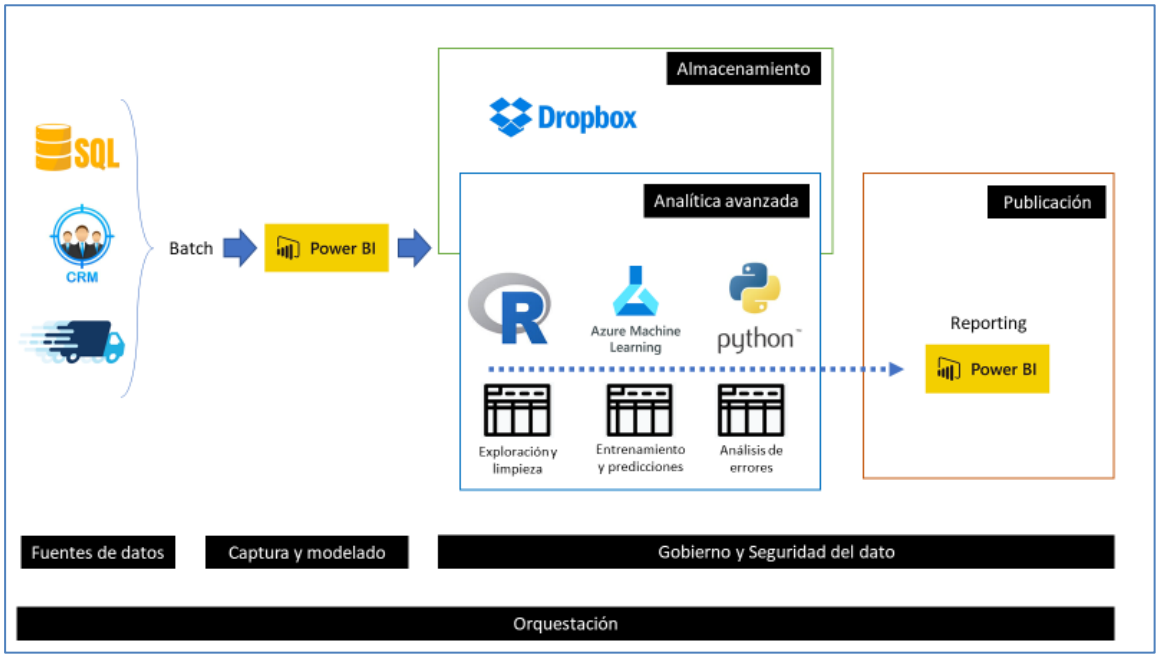
Fuente: Elaboración propia

Para cuidar la supervivencia de la red, es fundamental tratar de evitar que aquellos clientes mal atendidos se vayan de la empresa. Justamente, es el objeto de este trabajo lograr tal cometido haciendo uso del gran volumen de datos que dispone la compañía. Actualmente, además del sistema transaccional, la empresa cuenta con un CRM donde gestiona las relaciones con los clientes, y una base de datos de logística donde se almacena información histórica sobre las entregas de los pedidos. Toda esa información, de tipo estructurado, es condensada en el software de inteligencia empresarial Power BI, donde a través de tableros con visualizaciones interactivas se miden todas las variables críticas del negocio. Pero la tarea de análisis de datos, que hoy se limita a una analítica descriptiva tradicional, pretende ir más allá para lograr, a través de algoritmos de aprendizaje automático, predecir qué clientes dejaran de comprar y porqué, para poder diseñar estrategias para retenerlos.

Es por ello, que debe diseñarse una arquitectura de datos acorde a tal objetivo estratégico de negocio. En tal sentido, los datos que están en Power BI deben reorganizarse y exportarse para luego ser explorados y tratados con un software de análisis de datos como R, posteriormente entrenados con una herramienta como *Azure Machine Learning Studio*, los errores de clasificación del modelo deben ser analizados con lenguaje Python y por último los resultados deben ser expuestos a los directivos de la empresa a través del software de inteligencia

empresarial Power BI. En Figura 3 se expone la arquitectura de datos utilizada en este trabajo para hacer la predicción de pérdida de clientes.

Figura 3. Arquitectura de datos para predecir pérdida de clientes en empresa mayorista de alimentos



Fuente: Elaboración propia

Generación del conjunto de datos y elección de medidas de performance del modelo en el contexto de negocio

El objetivo de este apartado es generar el conjunto de datos y elegir las medidas de performance con las que se va a evaluar el modelo de predicción de pérdida de clientes en la empresa mayorista de alimentos bajo estudio. Para ello, primero se hace una selección de variables pertinentes para predecir la pérdida de clientes en empresas mayoristas, no sólo revisando la literatura sino también el criterio de especialistas en materia de negocios. En segundo lugar, se recolectan y tratan los datos con programas informáticos especializados en el análisis de datos como Power BI, R y *Azure Machine Learning Studio*. Por último, se eligen las medidas de performance con las que se va a evaluar el modelo de predicción de pérdida de clientes en la empresa mayorista de alimentos bajo estudio.

2.1. Selección de variables pertinentes para predecir la pérdida de clientes en empresas mayoristas

Para llevar a cabo esta tarea, se tienen en consideración tanto las variables mencionadas en la literatura, como las variables que, según el criterio de expertos en materia de negocios que asesoran a la empresa bajo estudio, deben ser tenidas en cuenta. En Tabla 1 se puede obtener una descripción general de las variables consideradas.

Numerosos estudios en el dominio B2C de predicción de abandono sugieren el uso de características relacionadas con la recencia, la frecuencia y el valor monetario (RFM) (Buckinx & Van den Poel 2005, Bose & Chen 2009, Migueis et al. 2013). Tamaddoni Jahromi et al. (2014) confirman esto en el contexto B2B al predecir la rotación clientes de un comercio minorista de bienes de consumo que vende en línea. Validan que la frecuencia de compras y el tiempo transcurrido desde la última compra son importantes.

Además de las variables RFM, existen otras características específicas del entorno B2B importantes. Entre ellas se destacan el porcentaje de markup promedio ponderado y la variedad de productos comercializados a cada cliente que se pueden extraer de la base de datos de transacciones de ventas, así como también, la antigüedad y el sector al que pertenece el cliente que se pueden obtener de la base de datos de clientes.

Tabla 1. Variables consideradas en este estudio

Tipo	Variable	Operacionalización	Fuente
RFM	dias_promedio_intercompra	Cantidad de dias promedio entre pedidos del cliente	Base ventas
	cantidad_de_pedidos	Cantidad total de pedidos en la historia del cliente	
	venta_promedio_mensual	Cociente entre el total de ventas al cliente y la cantidad de meses en los que compró	
Específicas de entornos B2B	valor_promedio_pedido	Cociente entre el total de ventas al cliente y la cantidad de pedidos	
	precio_prom_pond	Cociente entre el total de ventas al cliente y la cantidad de kilos que compró. Es un precio unitario promedio ponderado por kilo	
	porc_markup_prom_pond	Cociente entre la ulitidad bruta generada con el cliente y el costo de mercaderia vendida al cliente	
	variedad_de_productos	Cantidad de productos distintos comercializados al cliente	
	dias_antiguedad_cliente	Cantidad de dias desde la fecha de alta del cliente	Base clientes
	sector_cliente	Se generan 6 variables dummies: pequenos, medianos, restaurantes, dietéticas, panaderías, grandes	
	entregas_incompletas	Cantidad de pedidos incompletos entregados al cliente	Base logística
	cantidad_de_contactos	Cantidad de contactos telefónicos o personales del representante de ventas con el cliente	Base CRM
	dias_ultimo_contacto	Cantidad de dias transcurridos desde el último contacto con el cliente	
Propias de este estudio	cliente_cercano	1 si el cliente es de la misma ciudad que la empresa; 0 si es de otra ciudad	Combinación de variables
	vendedor_abandono	1 si el vendedor asignado esta en el top 5 de los con mayor fuga de clientes; 0 en caso contrario	
	provincia_abandono	1 si la provincia del cliente está en el top 15 de las con mayor fuga de clientes; 0 en caso contrario	
	disminuye_cant_pedidos	1 si el promedio diario de pedidos del ultimo mes es menor al del último trimestre; 0 en caso contrario	
	disminuye_ticket_medio	1 si el valor promedio por pedido del ultimo mes es menor al del último trimestre; 0 en caso contrario	
	cliente_perdido	1 si el cliente no vuelve a comprar por un período de cuatro meses o más; 0 en caso contrario. Es la variable a predecir	Base ventas

Fuente: Elaboración propia

Por otro lado, también se han considerado en este estudio, nuevas variables recomendadas por especialistas en materia de negocios que asesoran a la empresa, que no se mencionan en la literatura, y que se obtienen a través de combinaciones de otras variables. Capturan aspectos tales como si el cliente es de la misma ciudad que la empresa bajo estudio, o si el vendedor asignado al cliente es de los con mayores fugas, entre otros. A través de ellas, se pretende lograr un mayor poder predictivo del modelo.

Por último, está la variable a predecir, es decir, si el cliente va a dejar de comprar o no. Para ello, se considera que el cliente abandona la empresa, si deja de comprar por un período de cuatro meses o más.

2.2. Recolección y tratamiento de los datos

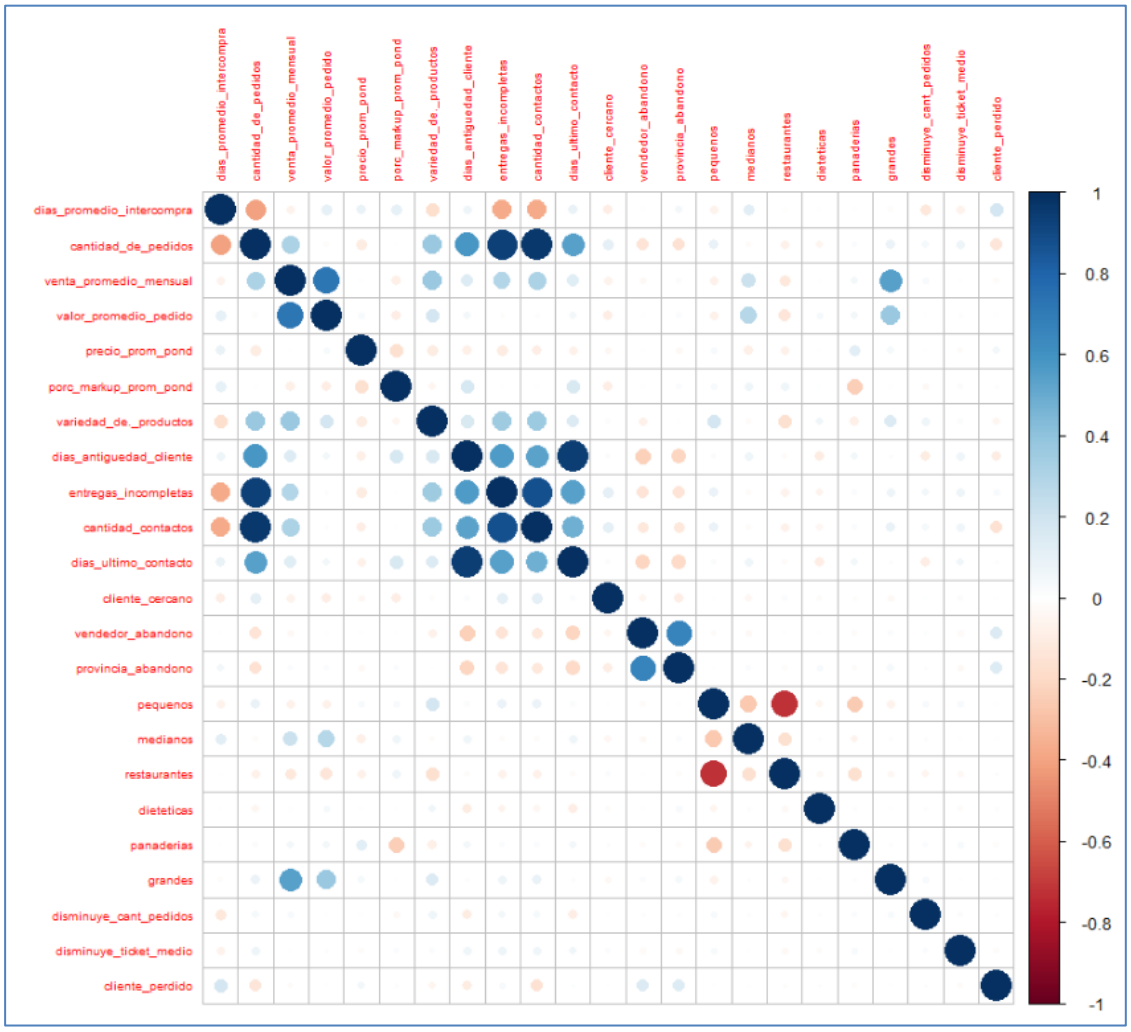
En el presente trabajo, se investigan las relaciones B2B de una empresa dedicada al comercio mayorista de alimentos. La empresa vende bebidas, alimentos y otros suministros esenciales a tiendas minoristas y las entrega usando logística propia.

En primer lugar, se recopilan datos maestros y datos transaccionales relevantes de diferentes fuentes primarias dentro de la empresa. Se cuenta con información sobre las transacciones de ventas realizadas durante el año 2021 y datos de identificación de cada cliente. Estos datos se combinan, agregan y relacionan en un *Data Warehouse* aplicando el *software* de inteligencia de negocios Microsoft Power BI.

A partir de la creación de métricas y atributos calculados, se pueden derivar características informativas a nivel de cada cliente, como la frecuencia y cantidad de pedidos, los ingresos por ventas, el ticket medio, la cantidad de días transcurridos desde la última compra, entre otras. Estás características, son las variables, que según se expuso en Tabla 1, conforman el conjunto de datos con el que se hará la predicción.

Previo a la aplicación de cualquier modelo de aprendizaje automático, es necesario hacer un análisis exploratorio y una posterior limpieza de los datos. A tal fin, y mediante el lenguaje de programación R, se obtiene la estadística descriptiva para una mejor comprensión de las variables del conjunto de datos y las relaciones entre ellas. En Figura 4 se puede apreciar el gráfico de la matriz de correlación. Allí se puede observar una alta correlación entre algunas variables. Tal es el caso de la cantidad de pedidos, altamente correlacionada con la cantidad de entregas incompletas y con la cantidad de contactos con representantes de ventas. Pero dado que se trata de variables que expresan informaciones distintas, se conservan en la muestra. Distinto es el caso de la variable provincia_abandono, que por estar altamente correlacionada con vendedor_abandono y expresar información similar (en general cada provincia tiene asignado su propio vendedor), se decide quitar del conjunto de datos.

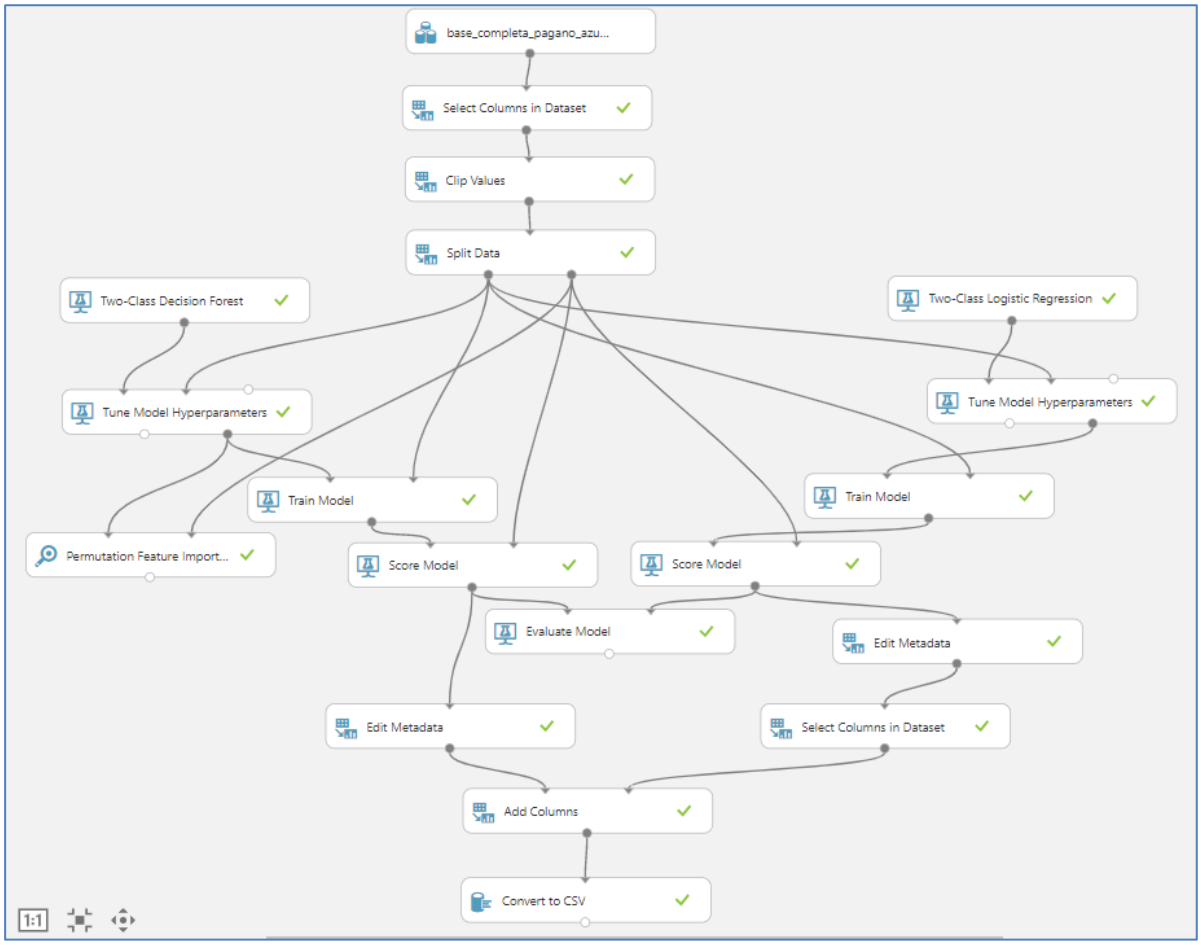
Figura 4. Gráfico de la matriz de correlación



Fuente: Elaboración propia

Luego en *Azure Machine Learning Studio*, se ejecutan algunas tareas de limpieza previas al entrenamiento del modelo. Entre ellas se destacan la eliminación de las variables id y provincia_abandono de la base, y el tratamiento de valores extremos en las variables dias_promedio_intercompra, cantidad_de_pedidos, venta_promedio_mensual, valor_promedio_pedido, precio_prom_pond, variedad_de_productos, dias_antiguedad_cliente, con la herramienta *clip values*. En Figura 5 se observa el diseño del pipeline datos en *Azure Machine Learning Studio* con los distintos pasos aplicados hasta llegar a la predicción. No solo se muestran las tareas de limpieza aplicadas, sino también la separación del conjunto de datos en entrenamiento y prueba, el tuneo de hiperparámetros, el entrenamiento de ambos modelos y la comparación de las predicciones y medidas de performance de cada uno.

Figura 5. Pipeline de datos en *Azure Machine Learning Studio*



Fuente: Elaboración propia

2.3. Medidas de performance del modelo en el contexto de negocio

En palabras de Gattermann-Itschert y Thonemann (2021), la performance de clasificación de un modelo de predicción de abandono de clientes se puede evaluar con medidas que se basan en la matriz de confusión. La matriz de confusión muestra cuántos clientes que se predijo que abandonarían realmente se fueron (verdaderos positivos: VP) o se quedaron (falsos positivos: FP) y cuántos de los que se predijo que no se irían se clasificaron correctamente (verdaderos negativos: VN) o fueron abandonos no detectados (falsos negativos: FN). La distribución de los clientes en los cuatro campos de la matriz de confusión depende del umbral de decisión. Los clientes se asignan a la clase positiva si su probabilidad de clase predicha está por encima de este umbral.

En este caso particular, el problema de negocio en cuestión es identificar qué clientes dejaran de comprar en la empresa y cuáles son las principales causas que motivan este comportamiento, para diseñar políticas efectivas para tratar de retenerlos. Dado que la tasa de abandono es baja, apenas un 3,05% (591 casos sobre un total de 19367 observaciones), y a que la empresa cuenta

con recursos para poder afrontar una campaña que trate evitar la pérdida de clientes de manera contundente, se decide optimizar una medida de performance que privilegie la detección de los casos verdaderamente positivos minimizando los falsos negativos. Al respecto, se opta por *Recall*, que mide la proporción de los casos realmente positivos detectados correctamente. En otras palabras, es el cociente entre los verdaderos positivos sobre todo lo que es positivo (verdaderos positivos + falsos negativos). También se puede denominar sensibilidad. Mejorar esta medida, perjudica otra medida llamada *Precision*, que mide cuanto de lo clasificado como positivo es realmente positivo. Es decir, es el cociente entre los verdaderos positivos sobre todo lo clasificado como positivo (verdaderos positivos + falsos positivos). En este contexto de negocio, para esta empresa particular, no preocupa tanto el valor que asuma está medida, dado que, en el peor de los casos, un falso positivo será un cliente innecesariamente contactado por una acción de retención. Incluso, antes de contactar al cliente, se puede designar un responsable que filtre aquellos clientes cuyos motivos de cese en la compra ya se conocen, limitando la acción de recuperación a los clientes que dejaron de comprar sin una razón aparente. En cambio, un falso negativo, representa un cliente que abandona la empresa, no detectado tempranamente. La empresa considera que el costo asociado a la pérdida de un cliente es considerablemente superior al costo de una acción de retención.

De todas formas, y con el objeto evaluar *Recall* y *Precision* conjuntamente, también se observa la métrica *F1-Score*, que se calcula haciendo la media harmónica entre ambas. Sin embargo, esta medida, supone que *Recall* y *Precision* tienen la misma importancia, lo cual, en el contexto actual, según lo comentado anteriormente no es cierto. En estos casos, se podría usar variaciones de esta medida, como *F beta-score* que permite ponderar con mayor importancia *Recall* con respecto a *Precision*.

Por otro lado, y dado que las clases están muy desbalanceadas, es decir, hay una muy baja proporción (cercana al 3%) de clientes que dejan de comprar, comparada con el total de la muestra, la medida *Accuracy*, no brinda mucha información. Esta métrica mide la frecuencia con la que el clasificador hace la predicción correcta. Es la relación entre el número de predicciones correctas (verdaderos positivos + verdaderos negativos) y el número total de predicciones. Tampoco es relevante para la elección del modelo en este problema de negocio, el área bajo la curva ROC, más conocida como AUC, dado que, al ser una medida independiente del umbral elegido, no considera que los costos de los falsos negativos puedan ser distintos a los costos de los falsos positivos. La curva ROC muestra la sensibilidad del clasificador graficando la tasa de verdaderos positivos contra la de falsos positivos. Muestra cuántas

clasificaciones positivas correctas se pueden obtener a medida que permite más y más falsos positivos.

Por último, en relación a la interpretabilidad del modelo, Miller (2017) la define como "el grado en que un ser humano puede comprender la causa de una decisión en un modelo". En este trabajo, se atenderá al peso relativo de las variables en la predicción para tratar de identificar las causas subyacentes en la fuga de clientes. Algoritmos simples como la regresión logística son fácilmente interpretables, mientras que algoritmos más complejos como *random forest*, boosted decision trees, o redes neuronales requieren de otras aproximaciones. Existen métodos flexibles que no dependen de ninguna particularidad del modelo que queremos interpretar. Uno de ellos es el cálculo de la importancia de variables, es decir, cuales de ellas tienen el mayor impacto en las predicciones, observando el error que se introduce en el modelo cuando falta la variable. Como no se puede evaluar un modelo sin algunas de las variables, se hace uso de *permutation feature importance*. Una variable es importante si la permutación de sus valores aumenta el error del modelo, lo que confirma que el modelo se basó en dicha variable para la predicción. De la misma manera, una variable no es importante si la permutación de sus valores mantiene el error del modelo sin cambios, porque el modelo ignoró la variable para la predicción (Santiago, 2020). Sin embargo, estos métodos son aproximaciones y tienen que ver más con técnicas de interpretación de modelos más que de identificación de relevancia para el modelo.

Selección de modelos, comparación de performance, análisis de errores y visualización de resultados

El objetivo de este apartado es seleccionar los modelos con los que se va a predecir la pérdida de clientes en la empresa mayorista de alimentos bajo estudio, para luego comparar su performance, en este caso sensibilidad e interpretabilidad, posteriormente analizar los errores que comete cada uno, y finalmente visualizar los resultados de una manera comprensible para los analistas de negocios. Para ello, primero se hace una selección y parametrización de modelos de aprendizaje automático para la predicción de pérdida de clientes. En segundo lugar, se compara la sensibilidad e interpretabilidad de cada modelo. En tercer lugar, se hace un análisis de errores sobre ambos modelos. Finalmente, se visualizan los resultados de las predicciones de una manera comprensible para los analistas de negocio de la empresa.

3.1. Selección y parametrización de modelos de aprendizaje automático para la predicción de pérdida de clientes

Se debe entrenar un modelo de predicción de abandono que clasifique a los clientes como futuros perdidos o no perdidos. Para una comparación de diferentes algoritmos de clasificación, se sigue a Verbeke et al. (2012). Las opciones más populares en los estudios de predicción de fuga de clientes son la regresión logística, las máquinas de vectores de soporte (SVM) y los modelos de árboles, en particular *random forest* (De Caigny et al. 2018). Este trabajo compara la sensibilidad y la interpretabilidad en la predicción de los modelos *random forest* y regresión logística.

Se espera que *random forest* sea el modelo con mejor perfomance en la predicción, a pesar de que probablemente no sea sencillo interpretar la importancia de las variables en la detección de la fuga de clientes. *Random forest* se aplica con frecuencia en la predicción de abandono debido a su robustez y bajo tiempo de ejecución en comparación con otras técnicas, al tiempo que ofrecen buenos resultados predictivos (Buckinx y Van den Poel 2005, Burez y Van den Poel 2009)

Con respecto a la optimización de los parámetros de cada modelo, se hace un barrido aleatorio de 10 corridas en cada modelo para optimizar la métrica *Recall*. En relación al modelo *random forest*, más precisamente *two-class decisión forest,* se opta por *resampling method: bagging* y se aleatorizan los parámetros *number of decision trees, maximum depth, number of random splits per node,* y *minimum number of samples per leaf node.* Con respecto al modelo regresión

logística, en este caso *two-class logistic regression*, se aleatorizan los parámetros *optimization tolerance, L1 y L2 regularization weight,* y *memory size for L-BFGS*. En Figura 6 se pueden observar los resultados de las 10 corridas en la búsqueda de hiperparámetros para *random forest*. La columna *Recall* muestra como decrece el valor a lo largo de las 10 corridas, obteniendo un máximo de 0,8179 para la combinación de 5 árboles, con *maximum depth* = 45, *number of random splits per node* = 490 y *minimum number of samples per leaf node* = 3.

En Figura 7 se pueden observar los resultados de las 10 corridas en la búsqueda de hiperparámetros para regresión logística. La mejor combinación de parámetros que maximizan el *Recall* a un nivel de 0,6879 (inferior al obtenido en *random forest*) es *optimization tolerance* = 0.000085, *L1 Weight* = 0,209605, *L2 Weight* = 0,018971 y *memory size* = 17.

Figura 6. Optimización de hiperparámetros en *random forest*

Minimum number of samples per leaf node	Number of random splits per node	Maximum depth of the decision trees	Number of decision trees	Accuracy	Precision	Recall	F-Score	AUC
3	490	45	5	0.991444	0.898701	0.817967	0.856436	0.976819
1	559	52	30	0.991886	0.926431	0.803783	0.860759	0.993193
8	601	16	20	0.991739	0.921409	0.803783	0.858586	0.99538
14	909	32	6	0.991444	0.920548	0.794326	0.852792	0.991432
14	1008	59	12	0.991444	0.922865	0.791962	0.852417	0.994045
1	59	47	4	0.991222	0.924581	0.782506	0.847631	0.961392
6	121	40	2	0.989599	0.873016	0.780142	0.82397	0.969239
12	117	12	12	0.991148	0.93913	0.765957	0.84375	0.99639
2	759	4	4	0.981707	0.903226	0.463357	0.6125	0.958561
13	215	1	9	0.968798	0	0	0	0.479178

Fuente: Elaboración propia

Figura 7. Optimización de hiperparámetros en regresión logística

OptimizationTolerance	L1Weight	L2Weight	MemorySize	Accuracy	Precision	Recall	F-Score	AUC
0.000085	0.209605	0.018971	17	0.989821	0.979798	0.687943	0.808333	0.994086
0.000013	0.741377	0.066427	9	0.985543	0.982979	0.546099	0.702128	0.991177
0.000076	0.113478	0.192541	22	0.983403	0.975962	0.479905	0.643423	0.984688
0.000053	0.586895	0.252595	33	0.981633	0.972826	0.423168	0.589786	0.984228
0.000093	0.888305	0.498072	13	0.978314	0.957447	0.319149	0.478723	0.977439
0.000036	0.117602	0.632062	7	0.977355	0.953125	0.288416	0.442831	0.975424
0.000004	0.057395	0.739404	10	0.976248	0.954955	0.250591	0.397004	0.97628
0.000016	0.478147	0.703134	11	0.976101	0.954128	0.245863	0.390977	0.976126
0.000001	0.546378	0.82493	48	0.97529	0.94898	0.219858	0.357006	0.974949
0.000087	0.985151	0.929787	21	0.974331	0.931034	0.191489	0.317647	0.972046

Fuente: Elaboración propia

3.2. Comparación de la sensibilidad e interpretabilidad de cada modelo

Una vez optimizados y entrenados ambos modelos se copian las predicciones sobre el conjunto de datos de prueba y se comparan las medidas de performance de cada modelo. La Figura 8 muestra las métricas de performance del modelo *random forest*. La medida elegida como más apropiada para el contexto de negocio (*Recall)*, indica que la sensibilidad del modelo, para un umbral de 0,5, es de 0,81. A su vez presenta *F1-Score* = 0,863, *Precision* = 0,925, *Accuracy* = 0,993 y AUC = 0,985. La Figura 9 expone las métricas de performance de la regresión logística. Para un umbral de 0,5, *Recall* es de 0,673 (notablemente inferior con respecto a *random forest*). Luego *F1-Score* es de 0,799, *Precision* es de 0,983, *Accuracy* = 0,990 y AUC 0,989.

Tanto *F1-Score* como *Accuracy*, al igual que *Recall*, son superiores en *random forest*, mientras que *Precision* y AUC son superiores en regresión logística.

Figura 8. Métricas de performance en *random forest*

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
136	32	0.993	0.925	0.5	0.985

False Positive	True Negative	Recall	F1 Score
11	5631	0.810	0.863

Positive Label	Negative Label
1	0

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	114	0	0.020	0.991	0.809	1.000	0.679	0.991	1.000	0.000
(0.800,0.900]	7	3	0.021	0.991	0.829	0.976	0.720	0.992	0.999	0.000
(0.700,0.800]	4	1	0.022	0.992	0.842	0.969	0.744	0.992	0.999	0.000
(0.600,0.700]	5	3	0.024	0.992	0.852	0.949	0.774	0.993	0.999	0.001
(0.500,0.600]	6	7	0.026	0.992	0.855	0.907	0.810	0.994	0.998	0.002
(0.400,0.500]	2	7	0.027	0.991	0.844	0.868	0.821	0.995	0.996	0.003
(0.300,0.400]	4	17	0.031	0.989	0.816	0.789	0.845	0.995	0.993	0.005
(0.200,0.300]	3	14	0.034	0.987	0.795	0.736	0.863	0.996	0.991	0.008
(0.100,0.200]	7	55	0.045	0.979	0.712	0.587	0.905	0.997	0.981	0.016
(0.000,0.100]	16	5535	1.000	0.029	0.056	0.029	1.000	1.000	0.000	0.985

Fuente: Elaboración propia

Figura 9. Métricas de performance en regresión logística

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
113	55	0.990	0.983	0.5	0.989

False Positive	True Negative	Recall	F1 Score
2	5640	0.673	0.799

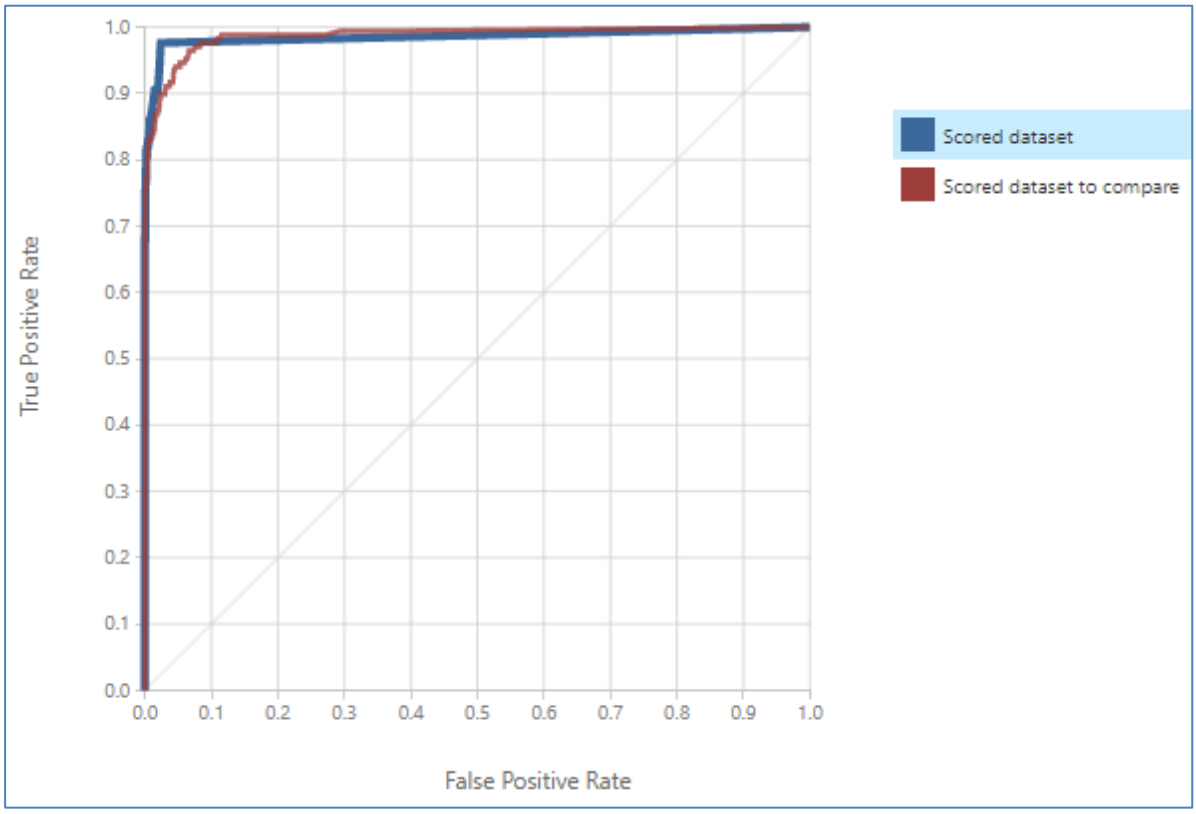
Positive Label	Negative Label
1	0

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	79	0	0.014	0.985	0.640	1.000	0.470	0.984	1.000	0.000
(0.800,0.900]	8	0	0.015	0.986	0.682	1.000	0.518	0.986	1.000	0.000
(0.700,0.800]	7	1	0.016	0.987	0.715	0.989	0.560	0.987	1.000	0.000
(0.600,0.700]	14	0	0.019	0.990	0.780	0.991	0.643	0.989	1.000	0.000
(0.500,0.600]	5	1	0.020	0.990	0.799	0.983	0.673	0.990	1.000	0.000
(0.400,0.500]	6	2	0.021	0.991	0.818	0.967	0.708	0.991	0.999	0.000
(0.300,0.400]	10	9	0.024	0.991	0.832	0.908	0.768	0.993	0.998	0.002
(0.200,0.300]	10	24	0.030	0.989	0.808	0.790	0.827	0.995	0.993	0.005
(0.100,0.200]	12	102	0.050	0.973	0.659	0.521	0.899	0.997	0.975	0.021
(0.000,0.100]	17	5503	1.000	0.029	0.056	0.029	1.000	1.000	0.000	0.989

Fuente: Elaboración propia

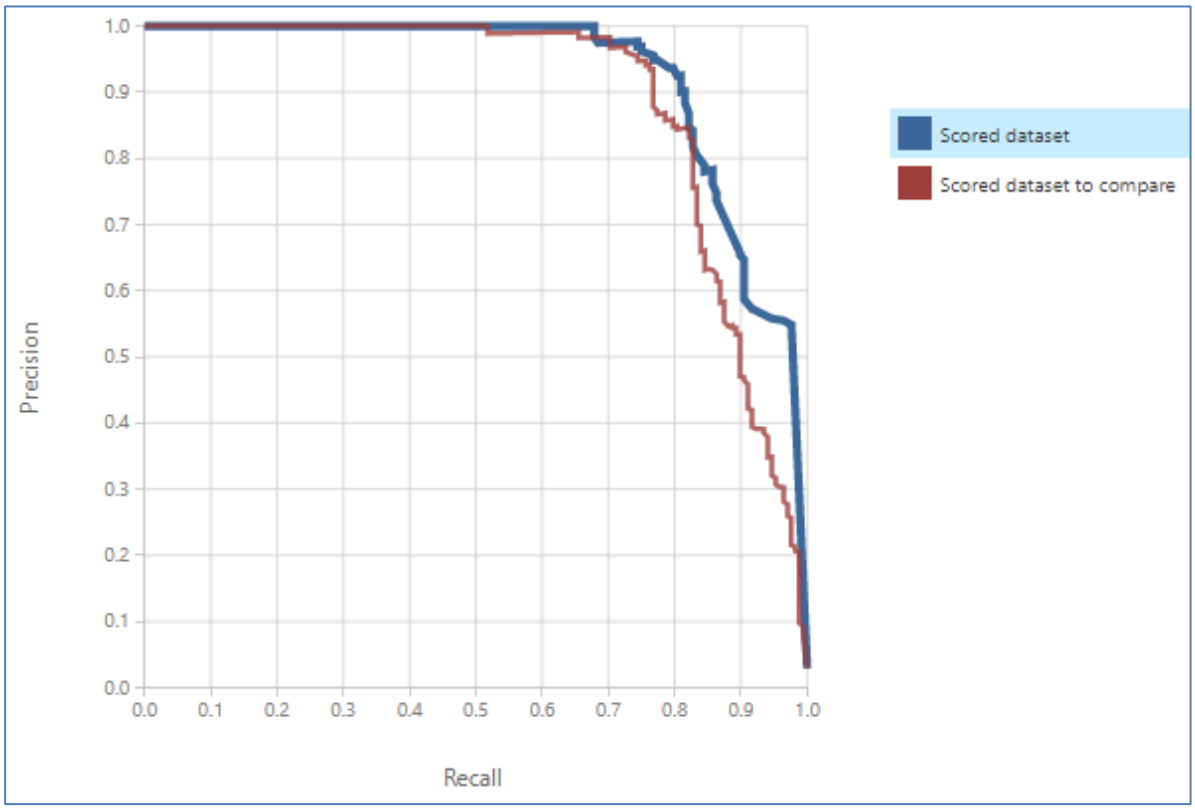
La Figura 10 superpone las curvas ROC de ambos modelos. La curva azul pertenece al modelo *random forest*, mientras que la curva roja pertenece a regresión logística. Como se puede apreciar, el modelo *random forest* presenta una mayor tasa de verdaderos positivos para valores de tasa de falsos positivos inferiores a 0,10. Mientras que la curva *Precision/Recall* expuesta en Figura 11, refleja que ante iguales valores de *Precision* el modelo *random forest* presenta valores superiores de *Recall*, en comparación con la regresión logística.

Figura 10. Compraración curvas ROC modelos *random forest* y regresión logística



Fuente: Elaboración propia

Figura 11. Compraración curvas Precision/Recall modelos *random forest* y regresión logística



Fuente: Elaboración propia

Como se comentó con anterioridad, la regresión logística es un modelo fácilmente interpretable. Una vez descartada la hipótesis de que todos los coeficientes son nulos, se necesita saber cuál es la contribución individual de los regresores, a la explicación de la variable dependiente. (Del Duca y Vietri, 2021). Esto se puede determinar observando el peso relativo de las variables en la predicción. En Figura 12, se puede identificar que la variable entregas_incompletas tiene un peso relativo de 81,9897, casi tres veces superior a la segunda variable en importancia que es días_último_contacto con 30,0964. En ambos casos, dado el signo positivo de los pesos, un aumento de estas variables implica un aumento en la probabilidad de clasificar a un cliente como perdido. Las siguientes variables en orden de importancia, tienen pesos negativos, lo que implica que un aumento en los valores de estas, disminuyen la probabilidad de clasificar a un cliente como perdido. Se pueden mencionar cantidad_contactos con -26,3158, cantidad_de_pedidos con -23,2561 y días_antigüedad_cliente con -21,4202, entre otras.

Figura 12. Peso relativo de variables en regresión logística

Feature Weights

Feature	Weight
entregas_incompletas	81.9897
dias_ultimo_contacto	30.0964
cantidad_contactos	-26.3158
cantidad_de_pedidos	-23.2561
dias_antiguedad_cliente	-21.4202
venta_promedio_mensual	-6.25206
Bias	-4.7121
valor_promedio_pedido	3.79584
dias_promedio_intercompra	3.13634
grandes	-2.59474
variedad_de _productos	-2.01022
porc_markup_prom_pond	1.95357
vendedor_abandono	1.54663
precio_prom_pond	1.06134
panaderias	-0.815094
medianos	-0.612774
restaurantes	-0.377948
pequenos	-0.203722
cliente_cercano	0.146221
disminuye_ticket_medio	-0.100922
disminuye_cant_pedidos	0.0740837
dieteticas	-0.0011684

Fuente: Elaboración propia

En el caso de *random forest*, según se explicó, para hallar la importancia de las variables se hizo uso de *permutation feature importance*. En Figura 13 se puede observar que la variable entregas_incompletas, al igual que en regresión logística, tiene la mayor importancia con un puntaje de 0,2065. Luego, la sigue cantidad_contactos con 0,1402, cantidad_de_pedidos con 0,03511 y días_ultimo_contacto con 0,0172. Si bien el orden no coincide, las variables más importantes en la predicción coinciden con las obtenidas a través de los pesos relativos en regresión logística.

Figura 13. Importancia de variables en *random forest*

Feature	Score
entregas_incompletas	0.20654
cantidad_contactos	0.140275
cantidad_de_pedidos	0.035112
dias_ultimo_contacto	0.017212
dias_promedio_intercompra	0.003442
dias_antiguedad_cliente	0.001549
precio_prom_pond	0.000688
venta_promedio_mensual	0.000516
valor_promedio_pedido	0.000516
pequenos	0.000516
porc_markup_prom_pond	0.000344
variedad_de _productos	0.000172
medianos	0
restaurantes	0
dieteticas	0
panaderias	0
grandes	0
disminuye_cant_pedidos	0
disminuye_ticket_medio	0
cliente_cercano	-0.000172
vendedor_abandono	-0.000344

Fuente: Elaboración propia

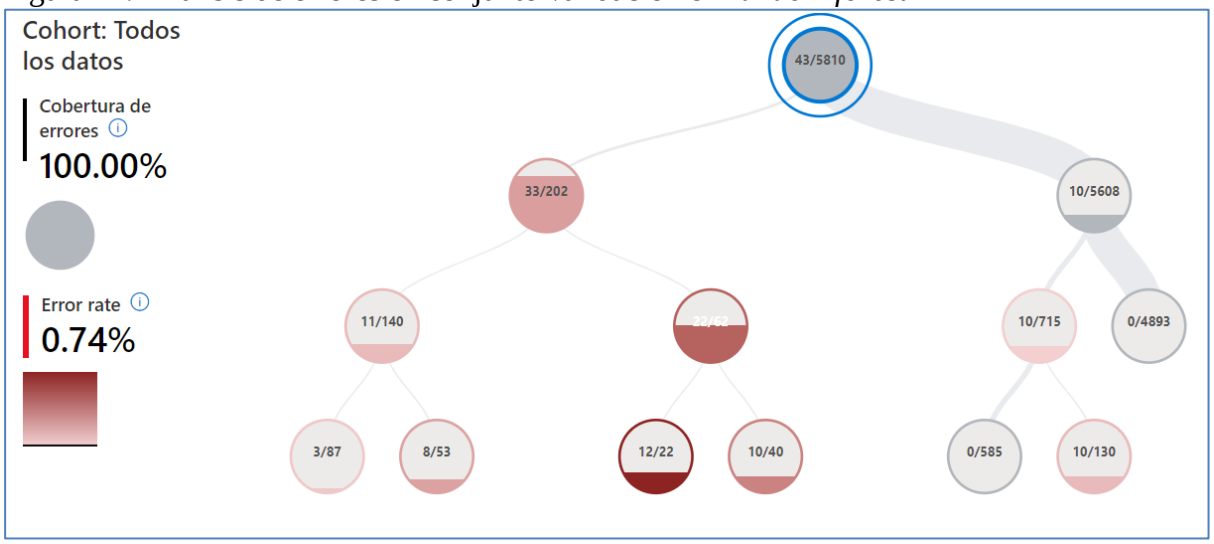
3.3. Análisis de errores sobre ambos modelos

También se puede indagar acerca de los errores que comete cada modelo y cómo se reparten en las distintas instancias. Según se observa en Figura 14, el árbol muestra cómo se distribuyen los errores que genera el modelo *random forest* y qué probabilidades hay de obtener errores en cada uno de los subconjuntos o áreas en las que se distribuye el error de una manera particular. Este modelo ha clasificado incorrectamente 43 casos de un total de 5810 observaciones que tiene el conjunto validación.

Para cada nodo seleccionado, la herramienta ofrece dos indicadores. *Error coverage* o cobertura de error, es el porcentaje de los errores totales del modelo cubiertos por el subconjunto seleccionado. En el nodo raíz este porcentaje es del 100%. El otro indicador es *error rate* o tasa

de error que mide cual es la probalidad de encontrar errores en el nodo seleccionado. En el nodo raíz, la tasa de error debe ser igual a 1 – *accuracy* del modelo. En este caso la tasa de error es 0,74%, que si se suma al *accuracy* obtenido del 99,26% se obtiene el 100%.

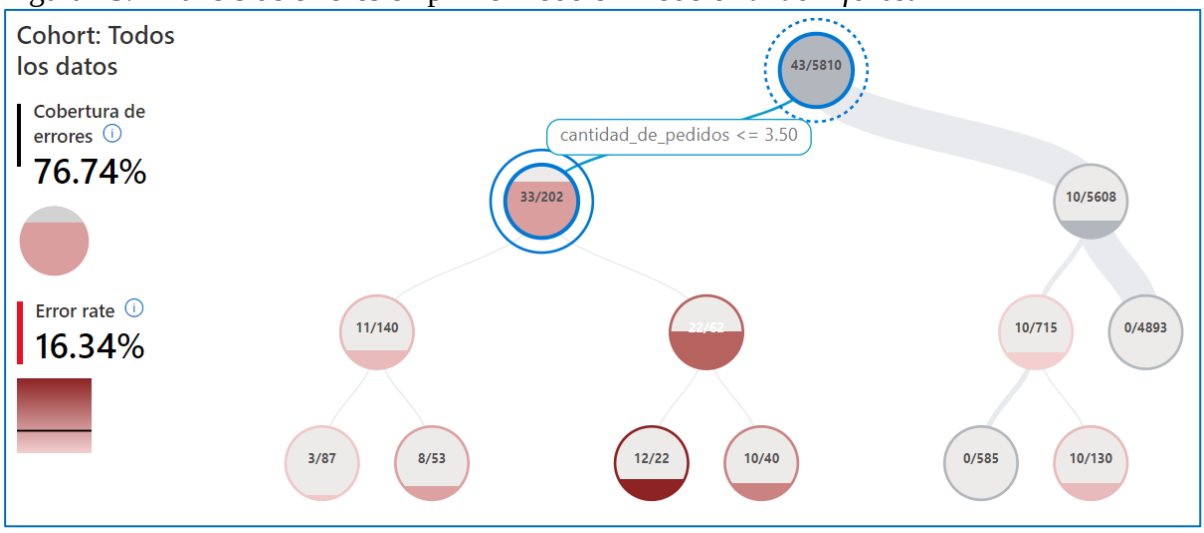
Figura 14. Análisis de errores en conjunto validación en *random forest*



Fuente: Elaboración propia

Según se expone en Figura 15, el primer nodo que se selecciona corresponde a todas las observaciones en las que la cantidad de pedidos es menor o igual a 3,5. Es decir se trata de clientes que no han registrado una alta frecuencia de compra. Este nodo representa un 76,74% del error que genera el modelo (33 errores de un total de 43 errores de clasificación del modelo) y para las instancias que están dentro de aquí hay un 16,34% de probabilidad de que el modelo prediga incorrectamente (33 errores de un total de 202 observaciones que están dentro de este subconjunto).

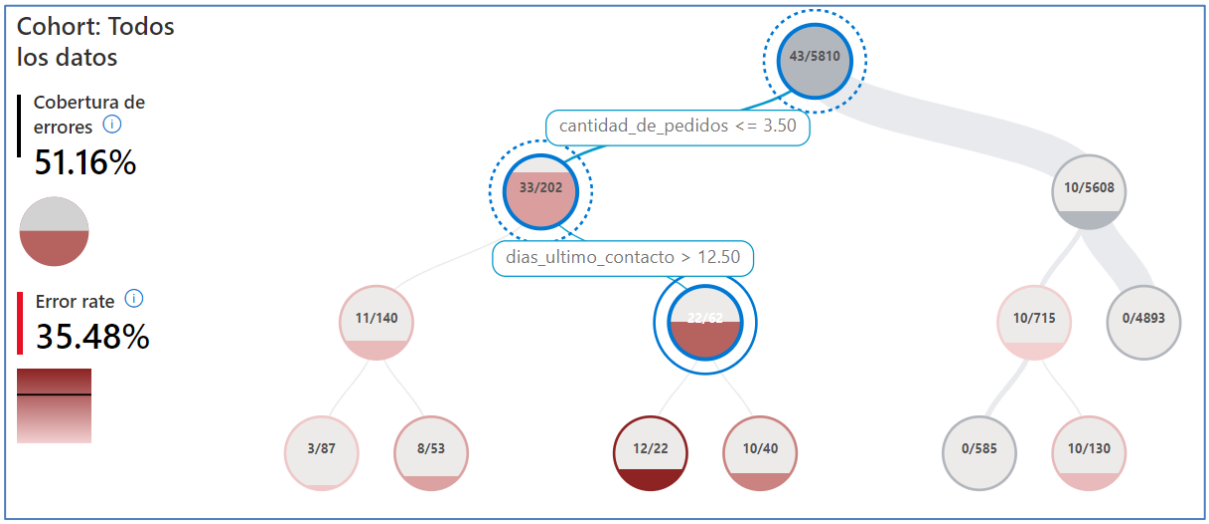
Figura 15. Análisis de errores en primer nodo en modelo *random forest*



Fuente: Elaboración propia

Si se sigue expandiendo el árbol hacia el área con más errores, la Figura 16 muestra el segundo nodo seleccionado que representa aquellas observaciones que hace más de 12,50 días desde el último contacto del representante de ventas. Es decir, son aquellos clientes que habiendo tenido una baja frecuencia de compra (<= 3,5 pedidos), no han sido recientemente contactados. Para este subconjunto el porcentaje de cobertura de errores disminuye a 51,16% pero la tasa de error aumenta considerablemente a 35,48%.

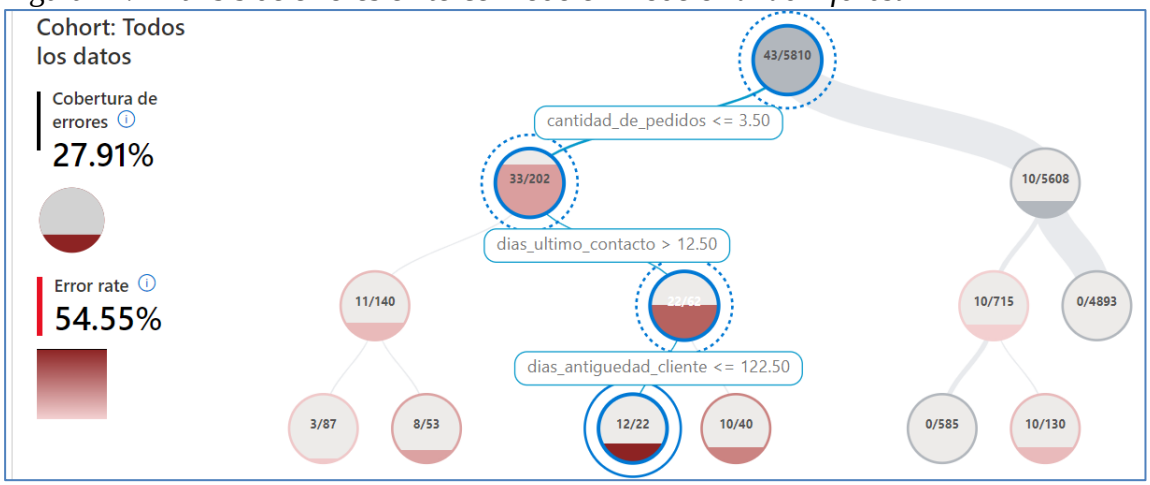
Figura 16. Análisis de errores en segundo nodo en modelo *random forest*



Fuente: Elaboración propia

Por último, se llega al siguiente nivel del árbol con más errores, que según se observa en Figura 17, es el tercer nodo seleccionado, que representa aquellos clientes que poseen una antigüedad menor o igual a 122,50 días. Es decir, son clientes que habiendo tenido una baja frecuencia de compra (<= 3,5 pedidos), no han sido recientemente contactados (>12,50 días) y adicionalmente son clientes relativamente nuevos (<=122,50 días de antigüedad). Este subconjunto tiene apenas 12 errores y 22 observaciones, pero el porcentaje de cobertura de errores es 27,91% y la tasa de error asciende a 54,55%. Esto implica que, revisando este pequeño grupo de casos, donde más de la mitad ha sido incorrectamente clasificado, se está tratando casi un tercio del total de los errores de clasificación del modelo. Esto podría indicar que faltan detalles de información alrededor de este tipo de clientes, o bien que no hay suficientes registros para que el modelo pueda aprender de manera más representativa como tratar este subconjunto de datos.

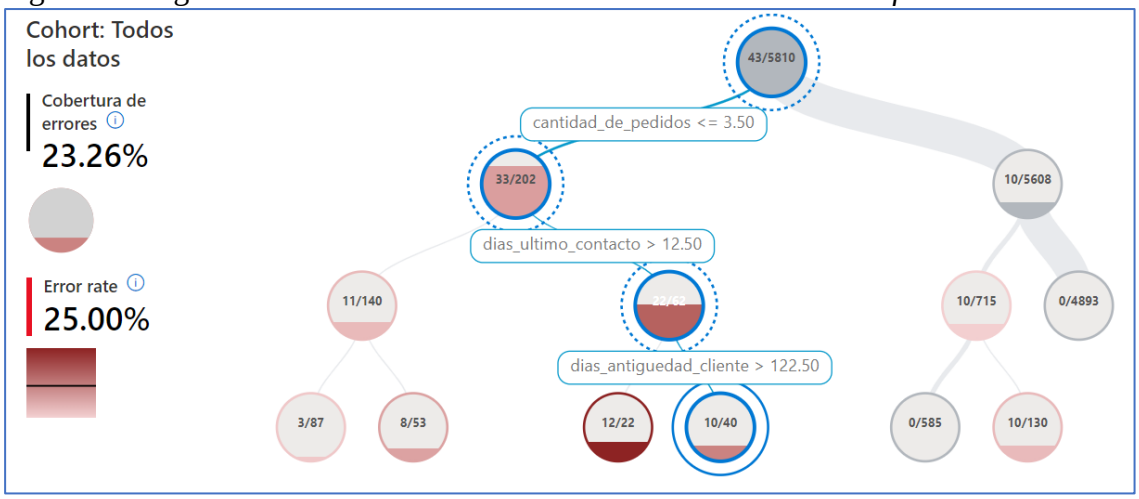
Figura 17. Análisis de errores en tercer nodo en modelo *random forest*



Fuente: Elaboración propia

Si bien se puede identificar otra ramificación con alto porcentaje de cobertura de errores la cual se expone en Figura 18, en este caso 23,26% comparado con el 27,91% de la primera ramificación analizada, esta posee una tasa de error considerablemente inferior, la cual asciende al 25%. Este subconjunto está compuesto por clientes que habiendo tenido una baja frecuencia de compra (<= 3,5 pedidos), no han sido recientemente contactados (>12,50 días) pero, a diferencia de la primera ramificación, son clientes más antiguos (>122,50 días de antigüedad).

Figura 18. Segunda ramificación con alto nivel de errores en *random forest*

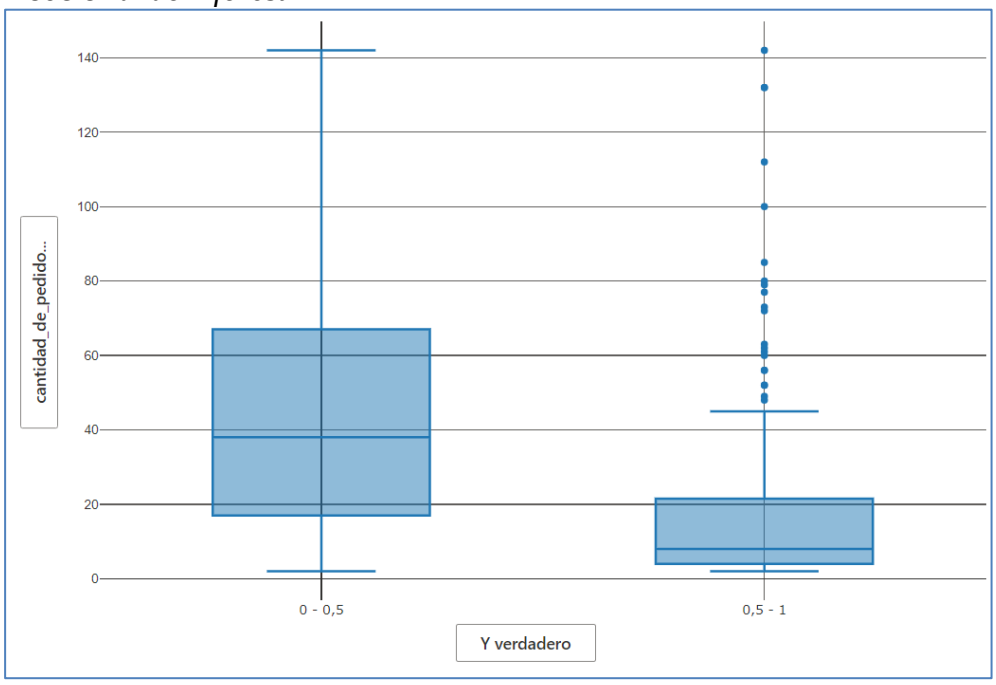


Fuente: Elaboración propia

Si se quiere ir más allá, y encontrar explicaciones de estos errores de clasificación, la herramienta permite graficar la distribución de una variable, ante distintos valores de otra variable. En Figura 19 se gráfica la distribución de valores de la variable cantidad_de_pedidos contra la variable Y verdadero. Como se puede observar, si bien en los casos de clientes que efectivamente abandonan la empresa (Y=1), la cantidad de pedidos se concentra en valores bajos (entre 1 y poco más de 20 pedidos), mientras que en los clientes que permanecen en la

empresa (Y=0), la cantidad de pedidos se concentra en valores notablemente superiores (entre 18 y 65 pedidos aproximadamente), existen valores atípicos que muestran clientes que abandonan la empresa a pesar de registrar elevada cantidad de pedidos. Estos *outliers*, favorecen la generación de errores de clasificación en el modelo en cuestión.

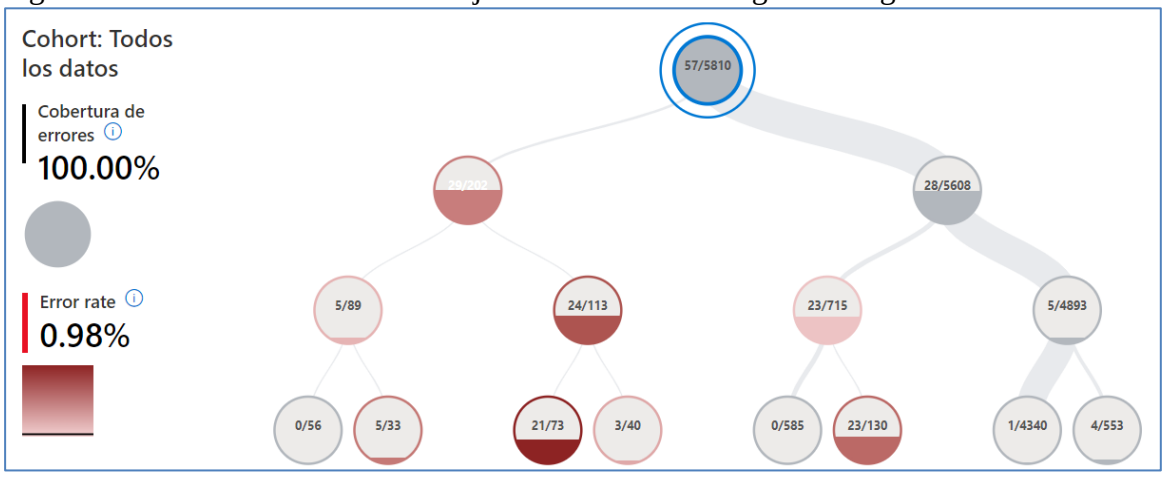
Figura 19. Distribución de la variable cantidad_de_pedidos contra la variable Y verdadero en modelo *random forest*



Fuente: Elaboración propia

Ahora si se analizan los errores de clasificación en regresión logística, según se observa en Figura 20, este modelo ha clasificado incorrectamente 57 casos de un total de 5810 observaciones que tiene el conjunto validación, lo cual arroja una tasa de error de 0,98%, que si se suma al *accuracy* obtenido del 99,02% se obtiene el 100%.

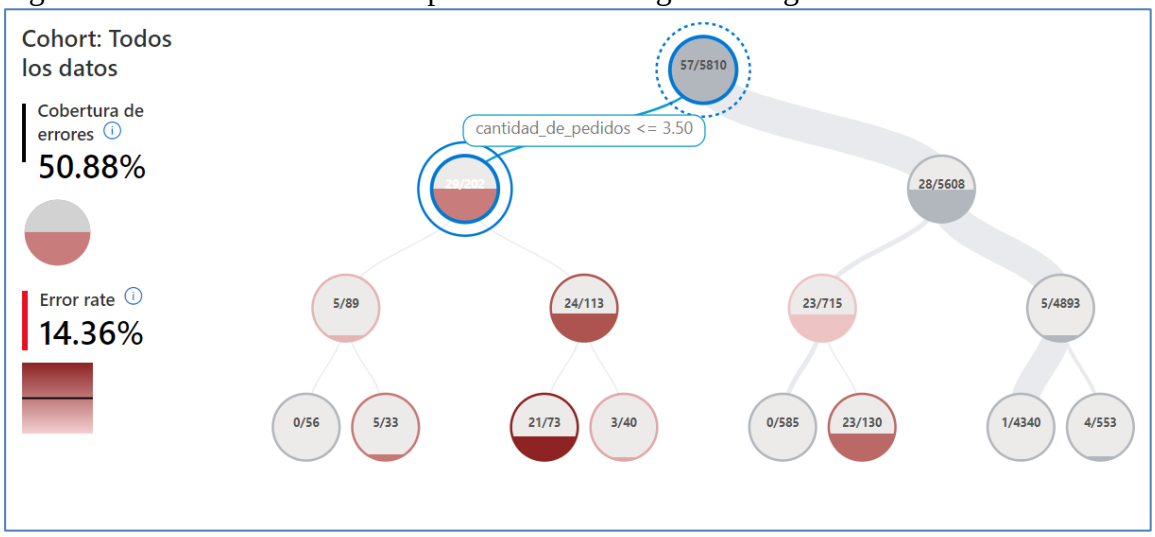
Figura 20. Análisis de errores en conjunto validación en regresión logística



Fuente: Elaboración propia

Según se expone en Figura 21, el primer nodo que se selecciona, al igual que en el modelo anterior, corresponde a todas las observaciones en las que la cantidad de pedidos es menor o igual a 3,5. Este nodo representa un 50,88% del error que genera el modelo (29 errores de un total de 57 errores de clasificación del modelo) y para las instancias que están dentro de aquí hay un 14,36% de chances de que el modelo prediga incorrectamente (29 errores de un total de 202 observaciones que están dentro de este subconjunto).

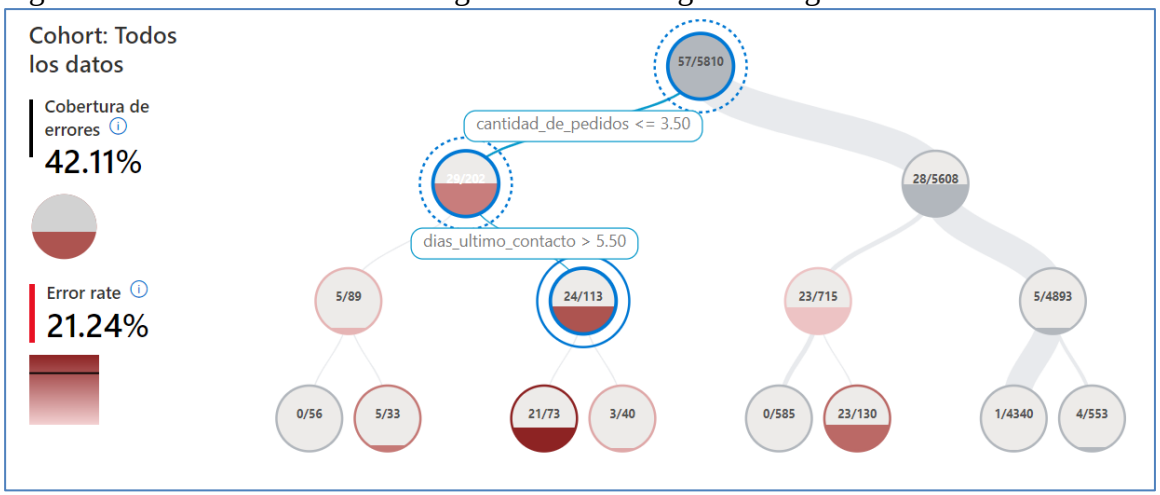
Figura 21. Análisis de errores en primer nodo en regresión logística



Fuente: Elaboración propia

Si se expande el árbol hacia el área con más errores, la Figura 22 muestra el segundo nodo seleccionado que representa aquellas observaciones que hace más de 5,50 días desde el último contacto del vendedor. Es decir, son clientes que habiendo tenido una baja frecuencia de compra (<= 3,5 pedidos), no han sido recientemente contactados. Para este subconjunto el porcentaje de cobertura de errores disminuye a 42,11% pero la tasa de error aumenta a 21,24%.

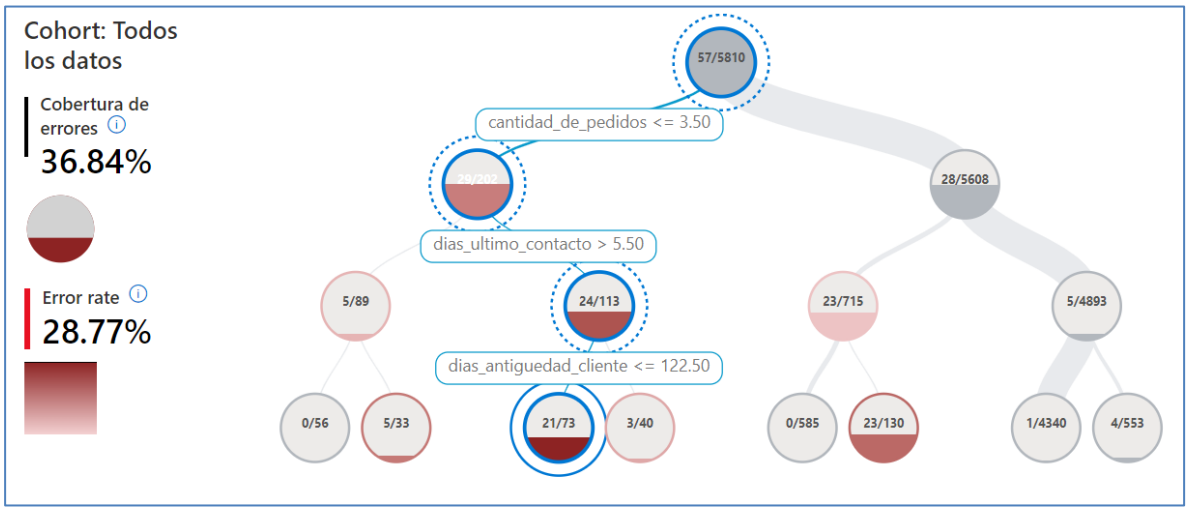
Figura 22. Análisis de errores en segundo nodo en regresión logística



Fuente: Elaboración propia

Por último, se llega al siguiente nivel del árbol con más errores, que según se observa en Figura 23, es el tercer nodo seleccionado, que al igual que el anterior modelo, representa aquellos clientes que poseen una antigüedad inferior o igual a 122,50 días. Es decir, son aquellos clientes que habiendo tenido una baja frecuencia de compra (<= 3,5 pedidos), no han sido recientemente contactados (>5,50 días) y adicionalmente son clientes relativamente nuevos (<=122,50 días de antigüedad). Este subconjunto con 21 errores y 73 observaciones, tiene un porcentaje de cobertura de errores del 36,84% y una tasa de error de 28,77%. A diferencia del modelo anterior, es mayor la cantidad de casos a revisar, y si bien representan una cobertura de error levemente superior, la cual representa poco más de un tercio del error total, la concentración de errores en estos casos, medido por la tasa de error, es considerablemente inferior.

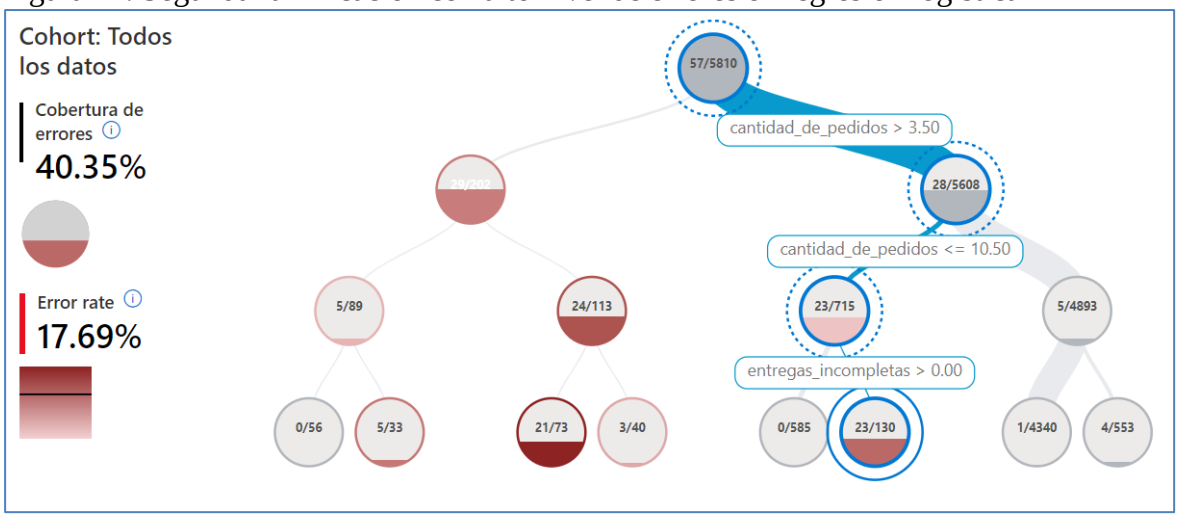
Figura 23. Análisis de errores en tercer nodo en regresión logística



Fuente: Elaboración propia

Adicionalmente, en este modelo, se observa otra ramificación del árbol con problemas de clasificación. En Figura 24 se observa un subconjunto de casos con cantidad de pedidos mayores a 3,5 y como máximo 10,5 y que presentan entregas incompletas, con un porcentaje de cobertura de errores superior a la primera ramificación analizada, que asciende a 40,35% (23 errores de un total de 57), pero con una tasa de error considerablemente inferior, apenas 17,69% (23 errores de un total de 130 observaciones).

Figura 24. Segunda ramificación con alto nivel de errores en regresión logística



Fuente: Elaboración propia

3.4. Visualización de los resultados

Con el fin de presentar los resultados obtenidos tras aplicar los modelos de aprendizaje automático, de una manera clara y completa para que cualquier persona que los visualice pueda interpretarlos, se crea un tablero interactivo en software de inteligencia empresarial Power BI.

La primera pregunta que se intenta responder es cuánto le cuesta a la empresa tener clientes perdidos sin tratar. A través de la variable venta mensual promedio del cliente, se calcula una nueva variable que agrega con la función suma, dichas ventas. Cabe aclarar que, dado que la base de datos de resultados no equivale a los clientes de un mes, sino que se trata una muestra de 5810 clientes que representa un 30% de los registros de ventas de un año -la cual surge del split data 70%/30% ejecutado previo al entrenamiento de los modelos- se estima que dicha muestra equivale a 3,6 meses -el 30% de un año-, por ende, se corrige la variable venta mensual promedio, dividiéndola por 3,6.

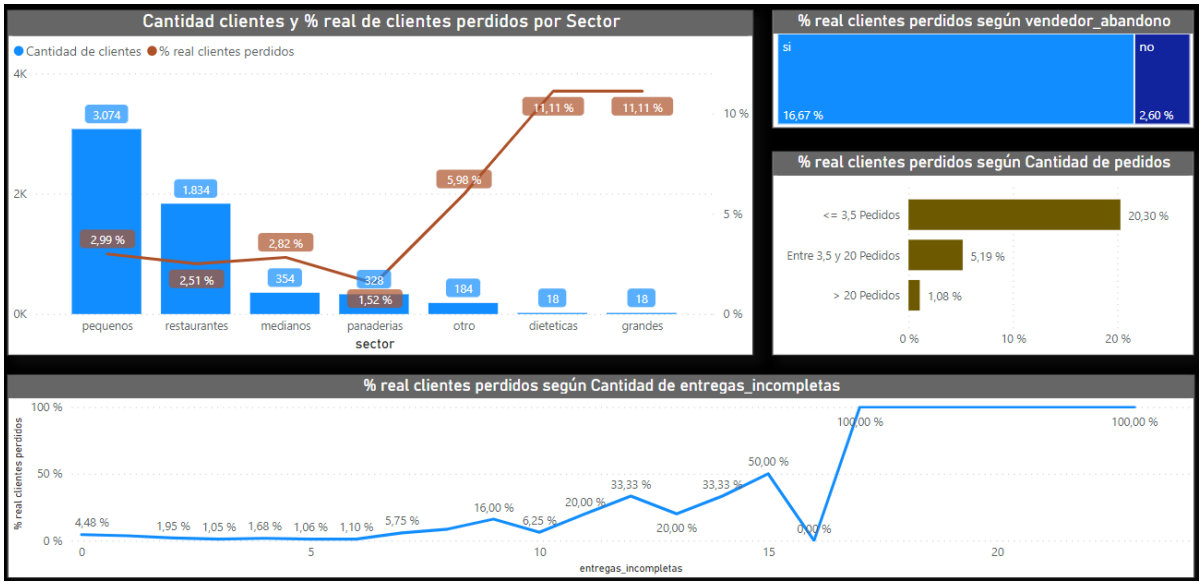
Luego, se calculan dos variables a partir de la venta mensual promedio. La primera se denomina “venta estimada perdida mensual”, y surge del producto entre la venta promedio mensual y el porcentaje real de clientes perdidos. La segunda se denomina “venta recuperable mensual”, y se calcula para cada modelo como el producto entre la venta estimada perdida mensual y el Recall del modelo respectivo. En Figura 25 se puede apreciar para la base de datos de resultados, la venta promedio mensual, el porcentaje real de clientes perdidos y la venta estimada perdida mensual. Dada una venta mensual promedio de $540.594 y un porcentaje de clientes perdidos del 2,89%, se estima una venta perdida mensual de $15.632.

Figura 25. Venta promedio mensual, porcentaje real de clientes perdidos y venta estimada perdida mensual en base de datos de resultados

Fuente: Elaboración propia

En Figura 26 se puede observar cómo se distribuyen los clientes perdidos en las distintas variables de la base de datos. Se puede advertir que si bien los sectores pequeños, restaurantes y medianos, son los con mayor cantidad de clientes, el porcentaje de clientes perdidos no llega al 3%, muy similar al promedio global. El segmento que sigue en cantidad de clientes es panaderías o reposterías, y es el que registra el menor porcentaje de clientes perdidos, con apenas un 1,52%. Luego, tanto en grandes como en dietéticas, ambos con pocos clientes, el porcentaje de clientes perdidos crece al 11,11%. También puede apreciarse que el porcentaje de clientes perdidos, en el grupo de vendedores con mayor deserción de clientes es del 16,67%. Por otro lado, el porcentaje de clientes perdidos es notablemente superior en clientes que han efectuado a lo sumo 3,5 pedidos en total. Por último, se puede afirmar que, en clientes con más de 16 entregas incompletas, el porcentaje de abandono es del 100%.

Figura 26. Distribución de clientes perdidos según distintas variables en base resultados



Fuente: Elaboración propia

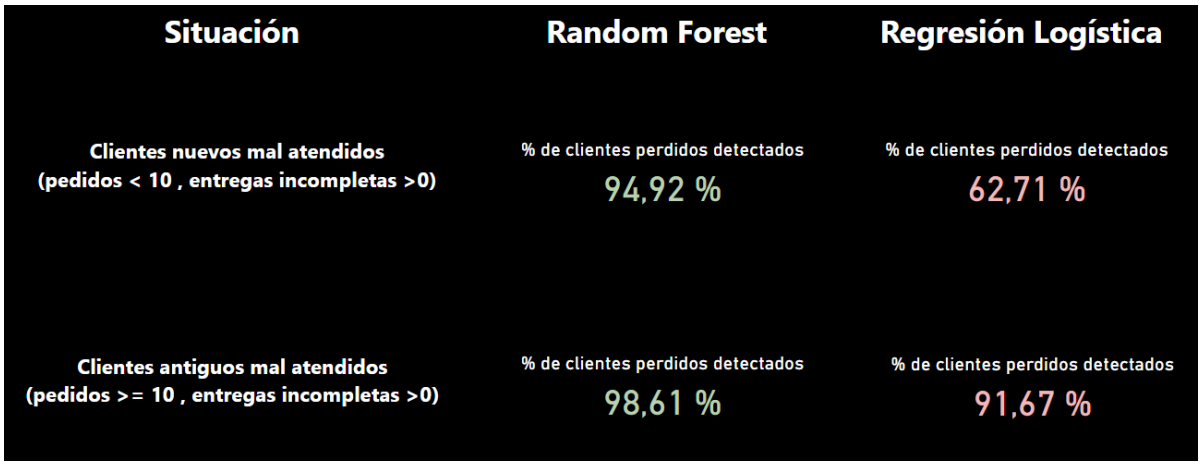
Luego en Figura 27 se compara el porcentaje de clientes perdidos detectados, más precisamente el Recall, de cada modelo, y la venta recuperable mensual estimada para saber cuál modelo conviene implementar. Dicho porcentaje, como ya se analizó, es mayor para *random forest*, y alcanza un 80,95% con una venta recuperable mensual estimada en $12.654.

Figura 27. Comparación de porcentaje de clientes perdidos detectado y venta recuperable mensual en *random forest* y regresión logística

Fuente: Elaboración propia

También puede compararse como se comporta cada modelo en situaciones concretas. En Figura 28 se analizan dos escenarios, el primero de clientes nuevos -con menos de 10 pedidos – mal atendidos – con entregas incompletas mayores que cero-, mientras que el segundo se ocupa de los clientes más antiguos – con 10 o más pedidos – mal atendidos – con entregas incompletas mayores que cero-. En ambas situaciones, *random forest* es el modelo con mayor porcentaje de clientes perdidos detectados.

Figura 28. Comparación comportamiento modelos en situaciones concretas

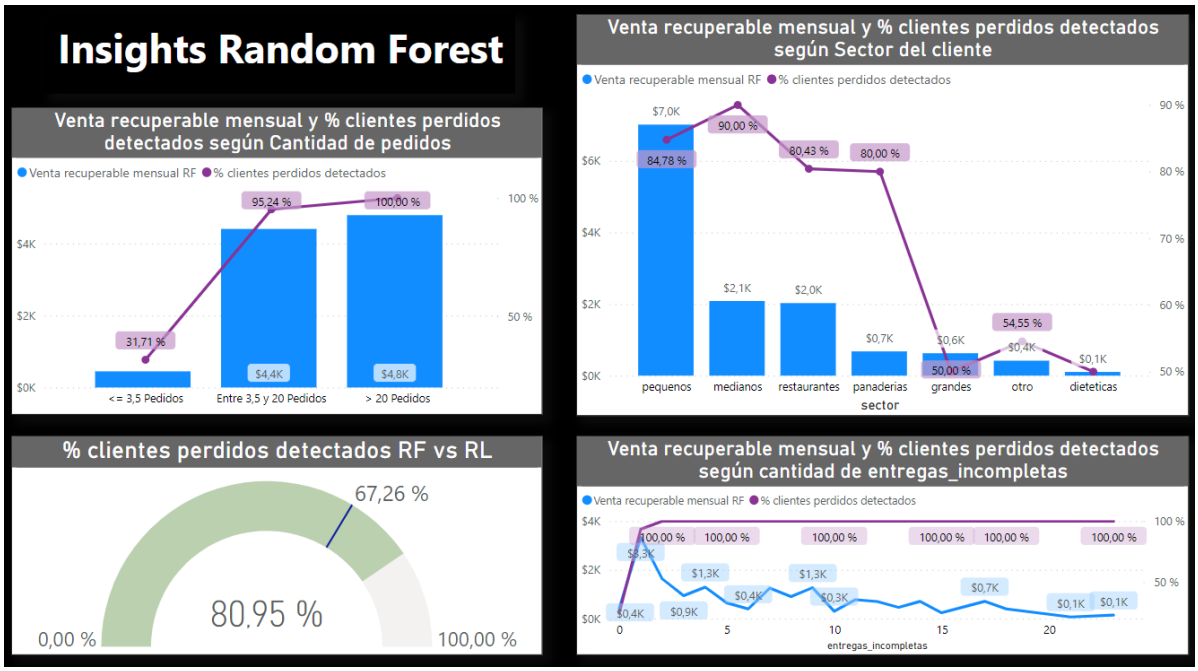


Fuente: Elaboración propia

Por último, en Figura 29 se hace un resumen de los principales *insights* en el caso de implementar *random forest*. En primer lugar, se puede apreciar que, si bien los clientes con 3,5 o menos pedidos eran los que registraban el mayor porcentaje de clientes perdidos, los resultados de las predicciones con *random forest* indican que es el grupo con menor porcentaje

de clientes perdidos detectados, y, por ende, la menor venta recuperable mensual estimada. La situación cambia para clientes con pedidos entre 3,5 y 20, con un porcentaje de clientes perdidos detectados del 95,24%, y luego para clientes con más de 20 pedidos dicho porcentaje alcanza el 100%. Posteriormente, si se analiza el comportamiento del modelo en los distintos sectores de clientes, se puede apreciar que, en los sectores pequeños, medianos, restaurantes y panaderías el porcentaje de clientes perdidos detectados es como mínimo 80%. Dichos sectores, como ya se había expuesto, son los sectores con mayor cantidad de clientes, por lo cual resulta muy conveniente que el modelo tenga buena performance dentro de ellos. Finalmente, se puede observar que para clientes con entregas incompletas mayores o iguales a 2, el modelo detecta el 100% de clientes perdidos.

Figura 29. Principales *insights random forest*



Fuente: Elaboración propia

Conclusión

Dentro de este trabajo, en primer lugar, se analizó el contexto *Big Data*, la predicción de pérdida de clientes en empresas mayoristas con relaciones B2B no contractuales y la relación de la problemática de pérdida de clientes con el modelo de negocios de la empresa bajo estudio. En segundo lugar, se describió el proceso de generación del conjunto de datos y elección de las medidas de performance del modelo en el contexto de negocio. En tercer lugar, se hizo la selección de modelos, la comparación de performance de cada uno y el análisis de los errores que cometieron en las predicciones.

El actual contexto de grandes volúmenes de datos, de diversos tipos y orígenes, ofrece un medio propicio para implementar modelos de aprendizaje automático que resuelvan problemas de negocio. El problema aquí es la pérdida de clientes, y dado que retener un cliente es menos oneroso que captar uno nuevo, el objetivo de este trabajo se centró en desarrollar un modelo que prediga qué clientes dejaran de comprar en la empresa y cuáles son las principales causas que motivan este comportamiento. Esta tarea no fue sencilla debido a que la empresa bajo estudio se desenvuelve en un entorno mayorista con relaciones B2B y sin contratos escritos con los clientes. Esto dificulta determinar cuándo un cliente se ha perdido, razón por la cual, se definió como tal a aquel que deja de comprar por un período de cuatro meses o más. Por otro lado, se analizó que, si bien el modelo de negocio busca que el número de clientes crezca, al igual que el número de proveedores y transportistas, si la red crece desmesuradamente, los efectos negativos de red atentan contra la misma, pudiendo alentar la fuga de clientes, entre otras consecuencias indeseadas.

También se hizo un estudio pormenorizado de las variables a tener en cuenta para el modelado del problema, tomando no sólo las recomendadas por la literatura sino también aquellas que se consideraron pertinentes. Al respecto, se consideraron diecisiete (17) atributos, provenientes de distintas fuentes como la base de datos de ventas, la base de clientes, el CRM y el sistema de logística, siendo que algunos se obtuvieron por combinación. Luego, en consonancia con el objetivo de retener clientes, se decidió optimizar una medida de performance que privilegie la detección de los casos verdaderamente positivos minimizando los falsos negativos. A tal fin, se optó por comparar la sensibilidad y la interpretabilidad en la predicción de los modelos.

Posteriormente, se seleccionaron los modelos *random forest* y regresión logística para hacer las predicciones. Con respecto a la sensibilidad o *Recall*, medida elegida como más apropiada para este contexto de negocio, para un umbral de 0,5, en *random forest* fue de 0,81, mientras que en

regresión logística fue de 0,673. A su vez, *random forest* presentó un F1-Score = 0,863, comparado con un *F1-Score* = 0,799 en regresión logística.

En relación a la interpretabilidad del modelo, es mayor en regresión logística ya que se puede observar el peso relativo de las variables en la predicción. Distinto es el caso en modelos como *random forest*, donde se requiere de otras aproximaciones, como lo es el cálculo de la importancia de variables con *permutation feature importance*. La variable entregas_incompletas, en ambos modelos, fue la que tuvo la mayor relevancia en la predicción. Luego, si bien el orden no coincide en los modelos comparados, las variables cantidad_contactos, cantidad_de_pedidos, y días_ultimo_contacto fueron las que siguieron en orden de importancia.

Si bien ambos modelos cometieron errores de clasificación, en *random forest* los errores estaban más concentrados, lo que simplifica la tarea de revisión.

Para concluir, se puede afirmar que, tanto por tener *Recall*, *F1-Score* y *Accuracy* superiores en *random forest*, como así también una mayor concentración de errores de clasificación en un pequeño subconjunto de observaciones, a pesar de no ser el modelo más fácilmente interpretable y tener *Precision* y AUC inferiores, se recomienda implementar el modelo *random forest* para cumplir con el objetivo de predecir la pérdida de clientes en una empresa dedicada al comercio mayorista de alimentos. Luego, si se quieren analizar los resultados en términos monetarios, también se concluye que conviene implementar *random forest* con la mayor venta recuperable mensual estimada en $12.654.

Este hallazgo es muy valioso para la organización debido a que, a partir de los resultados obtenidos, podrá adoptar políticas de retención de aquellos clientes que tempranamente se identifiquen como potenciales desertores. Esto puede conducir a un aumento significativo de las ganancias.

Sería interesante diseñar estrategias de retención de clientes basadas en la interpretación del peso relativo de las variables en la predicción. En trabajos futuros, se podrían diseñar experimentos de campo que evalúen la eficacia de adoptar dichas estrategias de retención sobre una fracción de la población de clientes, para confirmar si la implementación de este tipo de soluciones disminuye efectivamente la tasa de abandono de los clientes.

Referencias bibliográficas

Ascarza, E., Ebbes, P., Netzer, O., & Danielson, M. (2017). Beyond the target customer: Social effects of customer relationship management campaigns. *Journal of Marketing Research*, *54*(3), 347-363.

Ascarza, E., Iyengar, R., & Schleicher, M. (2016). The perils of proactive churn prevention using plan recommendations: Evidence from a field experiment. *Journal of Marketing Research*, *53*(1), 46-60.

Athanassopoulos, A. D. (2000). Customer satisfaction cues to support market segmentation and explain switching behavior. *Journal of business research*, *47*(3), 191-207.Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, *54*(6), 627-635.

Bhattacharya, C. B. (1998). When customers are members: Customer retention in paid membership contexts. *Journal of the academy of marketing science*, *26*(1), 31-44.

Berry, M. J., & Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.

Bose, I., & Chen, X. (2009). Quantitative models for direct marketing: A review from systems perspective. *European Journal of Operational Research*, *195*(1), 1-16.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European journal of operational research*, *164*(1), 252-268.

Burez, J., & Van den Poel, D. (2007). CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications*, *32*(2), 277-288.

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, *36*(3), 4626-4636.

Colgate, M. R., & Danaher, P. J. (2000). Implementing a customer relationship strategy: The asymmetric impact of poor versus excellent execution. *Journal of the Academy of marketing Science*, *28*(3), 375-387.

Colgate, M., Stewart, K., & Kinsella, R. (1996). Customer defection: a study of the student market in Ireland. *International journal of bank marketing*.

De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, *269*(2), 760-772.

Fader, P. S., Hardie, B. G., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of marketing research*, *42*(4), 415-430.

Ganesh, J., Arnold, M. J., & Reynolds, K. E. (2000). Understanding the customer base of service providers: an examination of the differences between switchers and stayers. *Journal of marketing*, *64*(3), 65-87.

Gattermann-Itschert, T., & Thonemann, U. W. (2021). Proactive customer retention management in a non-contractual B2B setting based on churn prediction with random forests.

Lilien, G. L. (2016). The B2B knowledge gap. *International Journal of Research in Marketing*, *33*(3), 543-556.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: la revolución de los datos masivos*. Turner.

Miguéis, V. L., Camanho, A., & e Cunha, J. F. (2013). Customer attrition in retailing: an application of multivariate adaptive regression splines. *Expert Systems with Applications*, *40*(16), 6225-6232.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1-38.

Osterwalder, A. (2004). *The business model ontology a proposition in a design science approach* (Doctoral dissertation, Université de Lausanne, Faculté des hautes études commerciales).

Osterwalder, A., & Pigneur, Y. (2010). *Business model generation: a handbook for visionaries, game changers, and challengers* (Vol. 1). John Wiley & Sons.

Parker, G., Van Alstyne, M. W., & Jiang, X. (2016). Platform ecosystems: How developers invert the firm. *Boston University Questrom School of Business Research Paper*, (2861574).

Ringbeck, D., Smirnov, D., & Huchzermeier, A. (2019). Proactive Retention Management in Retail: Field Experiment Evidence for Lasting Effects. *Available at SSRN 3378498*.

Real Academia Española: *Diccionario panhispánico del español jurídico (DPEJ)* [en línea]. < https://dpej.rae.es/ > [Fecha de la consulta: 16/08/2022]

Real Academia Española: Diccionario panhispánico del español jurídico (DPEJ) (2022).

Reinartz, W. J., & Kumar, V. (2000). On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *Journal of marketing*, *64*(4), 17-35.

Rust, R. T., & Zahorik, A. J. (1993). Customer satisfaction, customer retention, and market share. *Journal of retailing*, *69*(2), 193-215.

Santiago, F. (2021). Implementación de modelos de aprendizaje automático. Nota de cátedra extraída de: https://e72102.readthedocs.io/ el 16/08/2022.

Tamaddoni Jahromi, A., Stakhovych, S., & Ewing, M. (2014). Managing B2B customer churn, retention and profitability. *Industrial Marketing Management*, *43*(7), 1258-1268.

Van den Poel, D., & Lariviere, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European journal of operational research*, *157*(1), 196-217.

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert systems with applications*, *38*(3), 2354-2364.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & 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.

Wiersema, F. (2013). The B2B agenda: The current state of B2B marketing and a look ahead. *Industrial Marketing Management*, *4*(42), 470-488.

Apéndices

Sintaxis en el software R para el análisis exploratorio

```
rm(list = ls())
setwd(choose.dir())  ## Se setea el directorio de trabajo
## Se instalan paquetes
install.packages("pastecs")
install.packages("corrplot")
install.packages("psych")
install.packages("FactoMineR")
install.packages("factoextra")
base <- read.csv("base_completa_pagano_azure_1.csv",header = T) ## se levantan los datos del data frame
data.class(base) ## Se chequea que es un data frame
View(base) ## Se visualiza el data frame
library(pastecs)
dim(base) ## Se consulta la dimensionalidad (q filas y columnas)
BaseDescriptiva = base[2:24] ## Se crea una base sin la columna ID
View(BaseDescriptiva)
descriptiva <- stat.desc(BaseDescriptiva,basic=TRUE) ## Se crea un objeto con la estadística descriptiva
write.csv2(descriptiva, file = "Tabla descriptiva.csv") ## Se exporto a un csv
library(corrplot)
basetip <- scale(BaseDescriptiva, center = T, scale = T) ## Se tipifican todas las columnas de la base descriptiva
R <- cor(basetip, method = "pearson") ## Se genera la matriz R de correlación con la base tipificada
write.csv2(R, file = "Matriz correlación.csv") ## Se exporta a un csv
corrplot(R, sig.level=0.05,tl.cex=0.60) ## Se grafica el corrplot
```

Sintaxis en el software Python para el análisis de errores

Análisis de errores en conjunto de validación desde Azure ML Studio

Se realiza un análisis de errrores en Python utilizando un modelo entrenado en otra herramienta.

Se requiere instalar las librerias interpret-community, raiwidgets y error-analysis y lightgbm:

```
!wget
https://raw.githubusercontent.com/santiagxf/E72102/master/docs/develop/modeling/selection/code/error_analysis.txt \
    --quiet --no-clobber
!pip install -r error_analysis.txt --quiet
import pandas as pd
import numpy as np
validation = pd.read_csv('resultados_trabajo_pagano.csv')
validation
```

Se generan DataFrames con:

Los predictores del modelo. Los valores verdaderos. Las predicciones de cada modelo (two-class decision forest y two-class logistic regression).

```
X_val = validation.drop(['cliente_perdido', 'prediccion_df', 'probabilidad_df', 'prediccion_lr', 'probabilidad_lr'], axis=1)
y_val = validation['cliente_perdido'].to_numpy()
predictions_df = validation['prediccion_df'].to_numpy()
predictions_lr = validation['prediccion_lr'].to_numpy()
```

Las clases que nuestro modelo predice son:

* 0 cliente no perdido
* 1 cliente perdido

```
classes = validation['cliente_perdido'].unique().tolist()
classes
X_val
```

Los predictores disponibles son:

```
features = X_val.columns.values.tolist()
features
```

Los predictores con valores categóricos son:

```
X_val.dtypes
categorical_features = X_val.dtypes[X_val.dtypes == 'object'].index.tolist()
categorical_features
```

Análisis de errores

Antes de utilizar Error Analysis, se verifica que el dataset no contine valores nulos:

```
X_val.isna().any()
```

Para abrir la herramienta, se debe construir un ErrorAnalysisDashbord para cada modelo. Primero se construye para el modelo two-class decision forest:

```
from raiwidgets import ErrorAnalysisDashboard
from interpret_community.common.constants import ModelTask
ErrorAnalysisDashboard(dataset=X_val,
                       true_y=y_val,
                       categorical_features=categorical_features,
                       features=features,
                       pred_y=predictions_df,
                       model_task=ModelTask.Classification)
```

Luego se construye para el modelo two-class logistic regression:

```
from raiwidgets import ErrorAnalysisDashboard
from interpret_community.common.constants import ModelTask
ErrorAnalysisDashboard(dataset=X_val,
                       true_y=y_val,
                       categorical_features=categorical_features,
                       features=features,
                       pred_y=predictions_lr,
                       model_task=ModelTask.Classification)
```

Reporte del mentor

El presente TFE correctamente identifica una de las características más relevantes para organizaciones que operan en modalidad B2B que es la retención de clientes. Estas organizaciones tienen la característica de que el costo asociado de capturar un nuevo cliente es varias veces más alto que el costo de retener el mismo cliente. Sin embargo, saber que clientes realmente abandonaran la compañía con el suficiente tiempo de anticipación como para poder tomar una acción correctiva que evite esta decisión es una tarea desafiante. En este contexto, el trabajo plantea la posibilidad de utilizar las técnicas de analítica avanzada y aprendizaje automático para encontrar patrones que permitan predecir si un determinado cliente abandonará la empresa o no (hipótesis). El alumno efectivamente aborda la temática en el contexto y propone el desarrollo de un caso de negocio práctico donde resulte el problema planteado. El problema planteado requiere de un cuidadoso análisis de los resultados que se realiza de forma correcta y detallada en el trabajo. El alumno explora correctamente métricas relevantes en el contexto de algoritmos de detección (como un caso particular de los algoritmos de clasificación). Adicionalmente, destaco un correcto análisis de los errores que el modelo comete, explorando los diferentes resultados que obtiene al indagar las predicciones que realiza el modelo en determinadas cohortes de datos. Estas técnicas son de extremado valor ya que permite visualizar en el contexto del negocio los resultados. Si bien el trabajo no explora métodos avanzados o el estado del arte en lo que respecta a algoritmos de detección en estas características, se obtienen muy buenos resultados que pueden ser plasmados inmediatamente y se recomienda al alumno revisar este punto en un trabajo futuro.