Master in Management + Analytics

Modelo de Propensión de compra para plataforma de comercio electrónico

Autoría: Schettino, Macarena

Fecha de defensa de la tesis: 2022

UNIVERSIDAD TORCUATO DI TELLA

MASTER IN MANAGEMENT + ANALYTICS

MODELO DE PROPENSIÓN DE COMPRA PARA PLATAFORMA DE COMERCIO ELECTRÓNICO

Macarena Schettino

31 de Octubre de 2022

Tutor: Magdalena Cornejo

Resumen

La empresa sobre la que se desarrolla el presente trabajo está posicionada fuertemente dentro del mundo de la tecnología y digitalización del sector agropecuario puntualmente. Su negocio se basa en una plataforma de comercio electrónico dónde principalmente, productores y proveedores de dicho sector se reúnen para llevar a cabo transacciones.

Su crecimiento exponencial ha derivado en generar grandes volúmenes de datos día a día, llevando a la necesidad de encontrar una mejor manera de utilizarlos, a fin de obtener información precisa para la toma de decisiones acertadas.

En el presente trabajo se argumenta que dicha información no está pudiendo ser utilizada de una manera eficiente para la retención de los clientes, con la falta de una estrategia clara al momento de llevar a cabo las renegociaciones de los paquetes de membresías, los cuáles refieren al servicio contratado para publicitar y vender sus productos a través de la plataforma de manera *online*.

El foco del trabajo está puesto en la construcción de un modelo de aprendizaje automático que permita predecir la probabilidad de contratación de una membresía luego del proceso de renegociación por parte de los ejecutivos comerciales del área de ventas. Para lograrlo se aplicaron distintas técnicas de ingeniería de datos y aprendizaje supervisado, en el contexto de un problema de clasificación. Experimentando con distintos algoritmos de *Machine Learning* en búsqueda de aquel que presente una mejor *performance* en los resultados, otorgando un mayor poder predictivo.

Como resultado, se obtuvo un modelo de Random Forest con un rendimiento prometedor que alcanza un puntaje de 0.815 de área bajo la curva ROC en el conjunto de datos de validación, lo que les permitirá a los ejecutivos comerciales definir una estrategia superadora al momento de priorizar sus carteras de clientes.

Asimismo, tras otros análisis de probabilidades y puntos de corte óptimos se logra idear una potencial herramienta para el uso de los descuentos a ofrecer a los clientes. Dicha estrategia es simulada en distintos escenarios, optimista, intermedio y pesimista, en dónde se puede identificar el potencial ingreso extra o no generado a través de la implementación de la misma. Lo cuál, deja en evidencia la necesidad de ponerlo a prueba en producción.

Abstract

This work is focused on a company that is strongly positioned within the world of technology and digitalization of the agricultural sector. Its business is based on an electronic commerce platform where, mainly, producers and suppliers meet to carry out transactions.

Its exponential growth has led to the generation of large volumes of data every day, leading to the need to find a better way to use it, in order to obtain accurate information for making the right decisions.

In the present work, it is argued that this information is not being used for the retention of the clients, with the lack of a clear strategy at the moment of carrying out the renegotiations of the membership packages, which refer to the contracted service to advertise and sell their products through the online platform.

The project focuses on the construction of an automatic learning model that allows predicting the probability of contracting a membership after the renegotiation process by the commercial executives of the sales area. To achieve this, different data engineering and supervised learning techniques were applied in the context of a classification problem. Experimenting with different Machine Learning algorithms in search of the one that presents the best performance in the results to grant greater predictive power.

As a result, a Random Forest model was obtained with a promising performance that reaches a score of 0.815 for the area under the ROC curve in the validation data set, which will allow commercial executives to define a superior strategy when prioritizing their client portfolios.

Likewise, after other analyzes of probabilities and optimal thresholds, it is possible to devise a potential tool for the proper use of discounts to offer to customers. This strategy is simulated in different scenarios, optimistic, intermediate, and pessimistic, where the potential extra income generated or not through its implementation can be identified. That, reveals the need to test it in production.

ÍNDICE

1. Introducción

1.1. Marco teórico

1.1.1. Comercio electrónico

A raíz de la pandemia de COVID-19, la adopción del comercio electrónico se vio impulsada, marcando un punto de inflexión y un gran desafío. Frente a este contexto, el modelo de negocio de *e-commerce* se empezó a posicionar fuertemente, por lo que fue y seguirá siendo necesario que los avances tecnológicos acompañen este crecimiento exponencial, ofreciendo nuevas herramientas para optimizar los procesos de compra y venta a través de estas plataformas, alterándose la forma en que se hacen negocios.

Es así como, muchas empresas tuvieron que comenzar a repensar sus modelos de negocio afrontando nuevos desafíos, ya que la nueva realidad los obligó a realizar ventas habituales a través de transacciones electrónicas, trayendo consigo una necesidad de afrontar la digitalización. Consecuentemente, esto ha impulsado a pequeñas, medianas y grandes empresas a evolucionar en sus modelos de negocio, revaluando y replanteando su supervivencia en el mercado, siendo la transformación digital el camino para lograrlo (Rodríguez et al 2020).

El comercio electrónico, como se lo conoce mundialmente, es la actividad que implica la compra-venta de productos o servicios con la característica de que en este proceso intervienen medios o herramientas tecnológicas, como es el caso de internet (García, 2018), utilizando la tecnología de intercambio de datos, protocolos seguros y servicios de pago electrónico (Ramos, 2017).

El comercio electrónico ha permitido a muchas empresas alcanzar, con inversiones económicas mínimas, ventajas competitivas superadoras (Ramos, 2017), cambiando la facilidad de consumo de una manera única. Desde su inicio el fenómeno de comercio electrónico ha crecido constantemente influyendo en el

comportamiento de compras en todo el mundo (Meyer, 2011) convirtiéndose en una de las vías preferidas por los consumidores para realizar sus compras.

El foco de esta tesis se encuentra en una empresa de origen argentino dedicada al comercio electrónico, especializada específicamente en los agronegocios. Es una empresa de tecnología basada en una plataforma *online* dónde principalmente, productores y proveedores del sector agropecuario se reúnen para llevar a cabo transacciones. Entre otros, se comercializan principalmente productos pertenecientes a categorías como: maquinaria, campos, herramientas, seguros, equipamientos, vehículos e insumos agrícolas. Por un lado, los proveedores encuentran en dicha plataforma una herramienta que les facilita y complementa sus ya existentes procesos de cotizaciones, llegada a los compradores y canales de venta. Por el otro, los compradores logran tener un acceso rápido y sencillo a los productos que son de sus necesidades para continuar con sus procesos productivos.

Por cuestiones de confidencialidad de los datos prestados para llevar a cabo la presente tesis, no se podrá hacer mención del nombre de la empresa, por lo que se mantendrá en anonimato, evitando cualquier daño a la privacidad.

1.1.2. Aportes de las técnicas *Machine Learning* al problema en cuestión

El *Machine Learning*, es una ciencia dentro de la rama de la inteligencia artificial, que se refiere a un conjunto de algoritmos que usan estadísticas para encontrar patrones en cantidades masivas de datos con el fin de realizar predicciones. Según Tom Mitchell (2006) el aprendizaje automático, estudia algoritmos que buscan desarrollar sistemas informáticos que mejoren automáticamente su rendimiento o *performance* a través de la experiencia.

La mayoría de los problemas de aprendizaje estadístico refieren a dos categorías: aprendizaje supervisado o no supervisado.

Por un lado, en el dominio del aprendizaje supervisado, para cada una de las observaciones (i) de las variables predictoras (x), hay una respuesta Y asociada que

se puede medir. La idea es fitear un modelo que permita predecir la respuesta Y para futuras observaciones X (predicción), o entender la relación entre la variable respuesta y la predictora (inferencia) (James, et al., 2013).

Por el otro lado, en el dominio del aprendizaje no supervisado, para cada una de las observaciones (i) se cuenta con valores asociados a diferentes variables, formando un vector de medidas, pero sin tener asociada una respuesta (Y). Es decir, existen valores conocidos X, pero no variables respuestas Y, para predecir. Por lo cuál, la idea es encontrar patrones interesantes en los datos haciendo por ejemplo agrupaciones (detectar subgrupos similares de observaciones). Es más subjetivo, porque no se cuenta con una métrica para optimizar (James et al., 2013).

Para llevar a cabo la resolución del problema planteado a continuación, el cuál se desarrollará a lo largo del presente trabajo de tesis, se utilizaron distintos algoritmos de aprendiza automático, aplicando en definitiva diferentes técnicas de *Machine Learning* para encontrar patrones significativos en los datos, realizando predicciones precisas que ayuden a resolver el problema en cuestión.

Según Guillermo de Ockham y su principio de economía o parsimonia conocido como 'La Navaja de Ockham', las explicaciones nunca deben multiplicar las causas sin necesidad. Por lo tanto, cuándo se ofrecen dos o más explicaciones de un fenómeno, es preferible la explicación completa más simple, sin multiplicar las entidades sin necesidad. La explicación más sencilla suele ser la más probable (Rodriguez-Fernández, J., 1999).

De esta manera, en base a dicho teorema, se decidió comenzar con algoritmos simples e ir incrementando la complejidad hasta encontrar que algoritmos más complejos no brindaban una solución sustancialmente mejor. Sin embargo, es importante aclarar que como trabajo futuro sería interesante continuar probando algoritmos superiores en complejidad para ver el grado de mejora en la *performance* de los modelos predictivos.

1.2. Descripción del problema

La digitalización del mundo del agro llegó para quedarse, ofreciendo la posibilidad de hacer crecer el mercado *online* en este sector generando nuevas soluciones y servicios digitales cada vez más innovadores acorde a las necesidades que presentan las empresas y productores agropecuarios.

De esta manera, la empresa en cuestión desde sus inicios hace un par de años, comenzó a crecer de una manera exponencial cerrando financiamientos millonarios que la ayudaron a expandir y acelerar su crecimiento. Tras la necesidad de búsqueda de una mayor cantidad de clientes, es decir proveedores que publiciten y vendan sus productos en la plataforma *online*, el número de empleados comenzó a incrementarse para poder acompañar dicha expansión, debido a que no sólo una mayor cantidad de transacciones tenían que ser monitoreadas, sino que también se debía continuar expandiendo el negocio en búsqueda de nuevos clientes.

Gracias a ello, enormes volúmenes de datos se generan día a día, por lo que la posibilidad de acceder y analizar los mismos cobra una gran importancia. Con ellos, la empresa puede obtener información sobre una gran diversidad de aspectos que hacen al *core* del negocio, como puede ser, a modo de ejemplo el comportamiento de los clientes en la plataforma *online* y su *performance*. Todo esto, presenta un gran desafío a la hora de planificar, utilizar e interpretar dicha información disponible a la hora de tomar decisiones de una manera responsable y precisa teniendo un impacto positivo en el negocio.

Como la gran mayoría de las compañías nuevas que hoy en día se encuentran creciendo, esta empresa presenta una serie de contratiempos a la hora de utilizar dicha información para lograr dimensionar correctamente sus negocios y procesos de retención y adquisición de clientes. Puntualmente, resulta prioritario la retención de los vendedores, es decir aquellos que publicitan o venden sus productos en la plataforma *online*. Esto es así, ya que los mismos son la principal fuente de ingreso

para la empresa, la cual proviene de la contratación de membresías para publicitar o vender sus productos.

Cómo muchas otras empresas, se realizan distintos métodos de seguimiento sobre los clientes mediante estadística descriptiva, como ser tableros de control y KPIs, para que los ejecutivos cuenten con información detallada y previamente analizada antes de comenzar las renegociaciones. Esto además de brindarles una referencia de cómo está siendo el resultado de *performance* de sus clientes en la plataforma *online*, los ayuda a tener una mayor visibilidad de las fechas estipuladas de vencimiento de las membresías.

Sin embargo, a pesar de tener a disposición toda esta información descriptiva de los clientes, en este proceso de renegociación los ejecutivos comerciales no cuentan con una estrategia clara al momento de seleccionar los clientes cuyas membresías se encuentran próximas a vencer. La decisión de iniciar una renegociación la realizan de una manera aleatoria, sin tener un plan claro de cómo establecer una prioridad de clientes a contactar.

Sumado a ello, durante la negociación el ejecutivo cuenta con la posibilidad de realizar descuentos para cerrar el contrato de una nueva membresía. Sin embargo, nuevamente, estos no son utilizados con un plan concreto, lo que puede estar llevando a una ganancia menor. Según información brindada por la empresa, las negociaciones que realizan los ejecutivos comerciales ya tienen implícito un descuento del 15% por *default*.

Uno de los interrogantes de este proceso es si existe una mejor forma de priorizar los clientes a los cuales contactar y una mejor forma de utilizar los distintos tipos de descuentos para cerrar el contrato de manera tal de optimizar las ganancias aumentando la retención de los vendedores.

1.3. Objetivo de la tesis

El principal objetivo de este trabajo es encontrar un método basado en análisis de datos que permita predecir la propensión de los clientes a la hora de renovar una membresía. De esta manera, se podrá brindar a los ejecutivos comerciales información de mayor calidad al momento de renegociar la contratación de una membresía, permitiéndoles el armado de una estrategia de negociación robusta y una mejor priorización de la cartera de clientes activa a renovar.

El modelo intentará predecir la probabilidad de contratación de una membresía luego del proceso de renegociación. Para conseguirlo, se probarán distintos algoritmos de *Machien Learning* en búsqueda de obtener una mayor precisión en las predicciones.

A su vez, se harán unos análisis adicionales a fin de encontrar una estrategia de descuentos superior a la utilizada actualmente. Tras la implementación de distintos escenarios se probará y se buscará el potencial agregado de valor que podría generar a la empresa.

2. Materiales

En esta sección se procederá a realizar una descripción detallada de los datos utilizados para analizar el problema de negocio planteado a fin de encontrar una solución que permita ayudar a los ejecutivos comerciales a la hora de renegociar la contratación de las membresías.

2.1. Datos

2.1.1. Descripción y estructura de los datos

Los datos utilizados en la tesis fueron provistos por la empresa y por cuestiones de confidencialidad, debido a la sensibilidad de los datos utilizados, no se revelará la

identidad de la empresa de la cual provienen, evitando cualquier tipo de daño a la privacidad y preservando el anonimato en todo momento.

Los datos fueron extraídos de dos fuentes principales. Una de ellas es el sistema de gestión de relacionamiento con los clientes (también por cuestiones de confidencialidad no se expondrá el nombre del sistema utilizado por la empresa), de dónde se pueden destacar tres bases de datos que fueron utilizadas:

1) **Cuentas:** contiene información general sobre los clientes cargados en el sistema por parte de los ejecutivos comerciales. Estos clientes hacen referencia a los usuarios de tipo vendedores en la plataforma de comercio electrónico. De dicha base se utilizaron algunas variables de interés que serán detalladas en el anexo.

2) **Actividades:** contiene información sobre los distintos tipos de interacción que tienen los ejecutivos comerciales con los clientes. Se entiende por "tipo" de interacción a llamada, email ó tarea. De esta base se utilizaron las variables detalladas en el anexo.

3) **Oportunidades:** entendiéndose oportunidades por negociaciones, dicha base hace referencia a todas las negociaciones que tuvieron los ejecutivos comerciales con los clientes. Esta es la base principal sobre la cual se armó el *dataset* final que fue utilizado como input en los modelos de *Machine Learning* probados para alcanzar el objetivo propuesto. En el anexo se detalan cuáles son las variables utilizadas y a continuación se explica cómo se procedió al armado del *dataset* final.

La otra fuente de dónde se obtuvieron los datos restantes utilizados, refiere al *data warehouse* creado por la empresa para recopilar toda la información de la plataforma *online* dónde se comercializan los productos ofrecidos por los vendedores. De dicho origen se utilizaron las siguientes bases de datos. Las variables de cada una de ellas se detallan en el anexo.

1) **Merchant_performance:** contiene toda la información del comportamiento de los clientes en la plataforma *online*. Relacionada con las impresiones, leads, CR, CTR, tasas de lectura, respuesa, etcétera.

2) **Listado de productos:** presenta información sobre cada uno de los productos publicitados por el cliente en la plataforma. Contiene datos sobre si los productos se encuentran activos e inactivos, en que tipo de modelo de negocio se encuentran (más adelante se hará hincapié en esto y se explicará a que hace referencia), etcétera.

3) **Mensajes:** son 27 bases, una por cada mes, con información sobre interacciones que tuvieron los compradores con los vendedores. Se puede conocer si los vendedores leen y responden estos mensajes.

Considerando los datos que se disponían para abarcar el problema de negocio propuesto en el presente trabajo y encontrar una solución al mismo, se necesitó diferenciar cuál era la base de datos principal sobre la cuál se irían agregando distintas variables seleccionadas con diferentes criterios del resto de las bases de datos, para lograr nutrir al modelo y realizar así predicciones más precisas sobre la probabilidad de renovación de las membresías de los clientes.

Desde el punto de vista del negocio y la empresa en cuestión, fue necesario entender no sólo cuál era la base principal sino también la unidad de registro. La base de oportunidades, entendidas como negociaciones fue seleccionada para armar esta unidad de registro. Por lo tanto, hubo que realizar ciertas modificaciones a la base de datos original para poder llegar a obtenerla.

Para ello, se partió de la primera negociación ganada que haya tenido el cliente y luego se observó que sucedía en las posteriores negociaciones, en el caso de que las hubiera. Es decir, si estás se lograban ganar o perder por parte de los ejecutivos comerciales propietarios de dichas cuentas sobre las cuáles se hacen las negociaciones. De esta manera, considerando que para cada uno de los clientes se selecciona la primer negociación ganada que hayan tenido en su historial, fue

necesario descartar todas las anteriores a esta, es decir, las perdidas (en el caso de que las hubiera), como así tampoco se consideró a los clientes que únicamente tuvieron negociaciones perdidas.

Por lo tanto, se armó lo que se consideró una "tupla" como unidad de registro. En otras palabras, esta tupla está conformada en primer lugar por una negociación ganada, que luego está seguida de otra negociación inmediata que puede ser ganada o perdida. Este *label* de ganada o perdida, se encuentra establecido en la variable '*Stage*' de la base de negociaciones.

En línea con lo explicado en el párrafo anterior, para proceder al armado de estas tuplas, en un principio hay que considerar que hoy en día los ejecutivos comerciales cuentan con la posibilidad de comenzar a renegociar con sus clientes la renovación de las membresías actuales tres meses antes (90 días) del vencimiento estipulado de las mismas. La variable que hace referencia a esto es la '*Fecha de Fin de Servicio'* de la base de negociaciones. Sin embargo, esto no es una limitante para los ejecutivos al momento de decidir cuándo es conveniente llevar a cabo la renegociación, pudiendo de esta manera, comenzar con el proceso en un lapso de tiempo mayor a los tres meses. Debido a esto, se optó por elegir una ventana de tiempo de 120 días antes de la fecha de fin de servicio, para seleccionar la inmediata negociación a la primera ganada.

Durante el armado de las tuplas surgió el inconveniente de que había negociaciones que tenían únicamente un día de duración, es decir la fecha de fin de servicio era al día siguiente de la fecha de creación. Investigando internamente en la empresa con la gente apropiada, comentaron que esto se debe a un error de carga por parte de los ejecutivos comerciales, principalmente cuándo son nuevos y están aprendiendo el proceso de carga de las negociaciones en el sistema de gestión de clientes. Por lo tanto, para solventar este problema y tener un historial de información con un mayor impacto en la predicción del modelo, se decidió comenzar a buscar la siguiente negociación a la primera ganada, recién a los 90 días de la fecha de creación de la misma.

En resumen, el proceso de armado de los registros de la base principal sobre la cuál se agregaron las variables de interés del resto de las bases de datos, fue el siguiente:

- Primero, se seleccionaron las negociaciones cuyo estado sea igual a 'Ganada' con una fecha de creación posterior al 01/02/2020 (más adelante se detallará el motivo de elección de dicha fecha).
- Luego, se dejaron pasar 90 días desde la creación de esa negociación y teniendo en cuenta la fecha de fin de servicio de esa membresía contratada se establece una ventana de tiempo de 120 días previos. Por lo tanto, la negociación inmediata a la primera ganada se busca entre los 90 días posteriores a la fecha de creación de la primera ganada y 120 días antes de la fecha de fin de servicio, según la fecha de creación de esa segunda negociación. Es decir, la fecha de creación de esta segunda negociación que le sigue en la tupla tiene que estar comprendida entre ese lapso de tiempo.

A continuación se necesitó agregar información adicional a la base de negociaciones para que sea factible fusionar las bases seleccionando las variables necesarias desde el punto de vista del problema del negocio planteado. Es así como, desde la base de cuentas se incluyó a la base de negociaciones dos variables: *'ID Market Place'* y '*Billing Country*'. El país procedente de la cuenta fue necesario incluirlo ya que en la base de *performance* un mismo *'ID Market Place'* puede representar cuentas distintas en países distintos. Cabe aclarar que esto es un error del sistema de cómo está armado el *data warehouse*, y por lo tanto esta fue una de las maneras de solucionarlo. Esto permitió representar a un registro de tupla como único, a través del concatenado del *'ID Market Place'* y el país.

Otro punto a tener en cuenta antes de proceder al armado del *dataset* principal, es cuáles negociaciones incluir en base a las fechas a partir de las cuáles se cuenta con información en el resto de las bases de datos de dónde se seleccionaron las variables adicionales que son importantes incluir, para nutrir el modelo y dar una mayor robustez.

Para entender el motivo por el cual se procedió a realizar este filtro de fecha en la base de negociaciones, es necesario detallar para cada una de las bases de las cuales se extrajeron las variables a utilizar, desde cuándo se cuenta con información:

- **Oportunidades (negociaciones):** existen datos desde enero del 2019 a marzo del 2022.
- **Cuentas:** existen datos de todas las cuentas que hayan tenido cargadas negociaciones y actividades, por lo que no representa una limitante.
- **Actividades:** existen datos desde enero del 2020 a marzo 2022.
- ***Performance*:** existen datos desde febrero de 2020 a marzo 2022.
- **Productos:** existen datos desde febrero de 2020 a marzo 2022.
- **Mensajes:** existen datos desde enero 2020 a marzo de 2022.

Pudiendo observar lo anteriormente detallado, para poder utilizar las variables contenidas en las bases de las actividades, *performance*, productos y mensajes, no perdiendo dicha información, se necesitó seleccionar como fecha, febrero de 2020. Consecuentemente, las primeras negociaciones ganadas seleccionadas fueron aquellas cuya fecha de creación fue posterior al 01/02/2020.

Seguidamente, se mostrará una tabla que muestra las columnas y primeras filas del *dataset* principal que se armó con el detalle anteriormente mencionado:

	merchant_id	oportunidad_actual_id	oportunidad_label_id	fecha_opt_actual	fecha_opt_label	label_tag	label	mkt_place_id	country
6	0012S000024ujL8QAI	0062S00000vn0YdQAI	0062S00000yGbXYQA0	2020-03-05	2021-02-09	Ganada	1	11842	Argentina
7	0012S000024ujL8QAI	0062S00000yGbXYQA0	0062S000010xFuWQAU	2021-02-09	2021-12-27	Ganada	1	11842	Argentina
11	0012S000024uop7QAA	0062S00000wT40fQAC	0062S00000yTP0JQAW	2020-05-26	2021-03-17	Ganada	1	6210	Argentina
12	0012S000024uop7QAA	0062S00000yTP0JQAW	0068a00001FmlCNAAZ	2021-03-17	2022-03-28	Ganada	1	6210	Argentina
14	0012S000024uphRQAQ	0062S00000vo6wkQAA	0062S00000zMomHQAS	2020-04-14	2021-04-07	Ganada	1	11578	Argentina
...	...	...	...	...	...	...	...	...	...
4661	001E000001hXmrXIAS	0062S00000wTjM5QAK	0062S00000zPCSYQA4	2020-06-10	2021-06-03	Ganada	1	8041	Argentina
4665	001E000001hXmt9IAC	0062S00000wnsKgQAI	0062S00000zPTpdQAG	2020-08-07	2021-06-10	Perdida	0	8454	Argentina
4667	001E000001hXmtJIAS	0062S00000wqdgCQAQ	0062S000010Kbg0QAC	2020-10-21	2021-08-23	Perdida	0	1874	Argentina
4669	001E000001hXmu7IAC	0062S00000wV8LoQAK	0062S00000zO6KEQA0	2020-07-16	2021-05-07	Ganada	1	6144	Argentina
4671	001E000001hXmvAIAS	0062S00000wqkS3QAI	0062S000010MU2bQAG	2020-10-23	2021-10-06	Perdida	0	313	Argentina

2433 rows × 9 columns

Tabla 1: Muestra del *dataset* principal dónde se incluirán todas las variables de interés.

A esta altura, el *dataset* cuenta de 9 variables y 2433 registros (tuplas). Para lograr entender el problema, es importante entender los datos, y por lo tanto, que representan las variables creadas hasta ahora:

- ***'merchant_id'*:** identificador único de cliente (vendedor) en el sistema de gestión de clientes.
- ***'oportunidad_actual_id'*:** identificador único de la primera negociación ganada realizada al cliente.
- ***'oportunidad_label_id'*:** identificador único de la inmediata negociación realizada al cliente luego de la primera ganada, pudiendo esta ser ganada o perdida.
- ***'fecha_opt_actual'*:** fecha de creación de la primera negociación ganada.
- ***'fecha_opt_label'*:** fecha de creación de la inmediata negociación seleccionada luego de la primera ganada.
- ***'label_tag':*** estado de la inmediata negociación realizada al cliente luego de la primera ganada, es decir si fue ganada o perdida.
- ***'label'*:** '1' representa una negociación inmediata realizada al cliente luego de la primera ganada, como ganada, y '0' representa una negociación perdida.

- ***'mkt_place_id'*:** identificador único del cliente en la plataforma de comercio electrónico.
- ***'country'*:** país de dónde proviene la cuenta a la cual se realizó la negociación.

Por consiguiente, una vez que se contaba con el *dataset* principal, fue necesario comenzar a incluir las variables restantes. Éstas, en mayor medida, se relacionan no sólo con la *performance* y comportamiento que tuvieron los vendedores desde el momento en que comenzaron a ser clientes, sino que también con todas las interacciones que tuvieron los ejecutivos comerciales con estas cuentas en el lapso de tiempo estipulado anteriormente (90 días luego de la fecha de creación de la primera negociación ganada y a partir de ese momento, 120 días antes de la fecha de fin de servicio de la membresía contratada por el cliente).

Es importante hacer una aclaración respecto de la elección de las variables dentro de las diferentes bases de datos. Se seleccionaron aquellas que se sabía que la información se encontraba correctamente cargada en el sistema por parte de los ejecutivos comerciales.

Para comenzar se añadieron algunas variables de la base de negociaciones, como por ejemplo: *'rubro(UN)', 'división', 'total price', 'opportunity currency', 'categoria', 'subcategoria', 'product name', 'campaña activa', 'opp maturity (days)'*. Luego, se utilizó la base de cuentas para incluir la variable *'tipo de empresa'* y a continuación, se procedió con la base de actividades. En este caso, hubo que hacer un tratamiento de datos diferente.

Por un lado, se encontraba la variable *'task subtype'*, que desde el punto de vista del negocio cobra importancia en el tipo de relacionamiento a largo plazo que se genera con el cliente. Por lo tanto, se decidió hacer un tratamiento con la técnica *One Hot Encoding* (en la sección 2.3.1 se explica en detalle) a fin de tener agrupado para cada cliente la cantidad de los distintos tipos de interacción llevados a cabo con la cuenta en cuestión, en la ventana de tiempo explicada anteriormente, es decir entre *'fecha_opt_actual'* y *'fecha_opt_label'*, correspondiente a la tupla de negociaciones

establecida. Así, teniendo en cuenta el campo '*date'* de la base de actividades se pudo establecer la cantidad de los distintos tipos de interacción ('*task*', '*email'*, '*call'*) para cada uno de los registros. Con esta información se intentó entender de una mejor manera el comportamiento de los ejecutivos con las cuentas cuyas negociaciones fueron ganadas en una primera instancia y posteriormente proceden a una renegociación.

Por el otro lado, para cada uno de los registros se calculó la actividad media y la frecuencia de actividades en días entre ambas fechas que definen la tupla (*'fecha_opt_actual'* y *'fecha_opt_label'*). Con el fin de conocer cuál era el promedio y la mediana en días que pasaban entre la realización de una actividad y la siguiente, ayudando a la interpretación del comportamiento de los ejecutivos con sus clientes. A estas variables se las llamó: *'activity_frecuency_mean'* y *'activity_frequency_median'*.

Seguidamente, se continuó con las variables presentes en la base de *performance*. Para poder explicar la transformación de datos realizada, es necesario entender cómo se presentan los registros en dicha base. En la siguiente tabla, se muestra un ejemplo de las primeras columnas y filas que la componen:

Month Year	Country	Region	City	Id	Type Read Leads	Membership	Category N1 Ppal	Category N2 Ppal	Engmt
feb-2020	Argentina	Corrientes	Gdor Valentín Vi	12	Mi cuenta	Sucursal Primary	Maquinaria Agrícola	Tractores	Excelente
feb-2020	Argentina	La Pampa	General Pico	17	Mi cuenta	Sucursal Plus	Maquinaria Agrícola	Tractores	Excelente
feb-2020	Argentina	Santa Fe	Bombal	19	Mi cuenta	Sucursal Plus	Maquinaria Agrícola	Tractores	Malo
feb-2020	Argentina	Cordoba	Marcos Juárez	23	Envía Datos Usuario	Corporate Conquer	Maquinaria Agrícola	Pulverizadoras	-
feb-2020	Argentina	Cordoba	San Francisco	24	Envía Datos Usuario	Corporate Going Forward	Maquinaria Agrícola	Tolvas	-
feb-2020	Argentina	Buenos Aires	San Antonio De A	25	Mi cuenta	Sucursal Plus	Maquinaria Agrícola	Cosechadoras	Muy Malo
feb-2020	Argentina	-	-	26	Mi cuenta	Sucursal Plus	Vehículos	Acoplados	Malo
feb-2020	Argentina	Buenos Aires	Villa Luzuriaga	30	Mi cuenta	Sucursal Advanced	Maquinaria Agrícola	Tractores	Bueno
feb-2020	Argentina	Cordoba	Marcos Juárez	31	Mi cuenta	Corporate Expand	Maquinaria Agrícola	Fertilizadoras	Excelente
feb-2020	Argentina	Cordoba	Isla Verde	32	Integración CRM	Sucursal Plus	Maquinaria Agrícola	Cosechadoras	-
feb-2020	Argentina	-	-	33	Mi cuenta	Corporate Expand	Maquinaria Agrícola	Plataformas Y Cabezales	Bueno
feb-2020	Argentina	Chaco	Charata	34	Mi cuenta	Sucursal Advanced	Maquinaria Agrícola	Tractores	Excelente
feb-2020	Argentina	-	-	36	Mi cuenta	Sucursal Plus	Maquinaria Agrícola	Tractores	Muy Malo
feb-2020	Argentina	La Pampa	Realicó	38	Envía Datos Usuario	Corporate Starting	Maquinaria Agrícola	Tractores	-
feb-2020	Argentina	Buenos Aires	Bragado	39	Mi cuenta	Sucursal Plus	Maquinaria Agrícola	Tractores	Muy Malo
feb-2020	Argentina	Buenos Aires	9 De Julio	41	Mi cuenta	Sucursal Plus	Maquinaria Agrícola	Tolvas	Excelente
feb-2020	Argentina	Santa Fe	Cañada Rosquín	43	Mi cuenta	Sucursal Plus	Maquinaria Agrícola	Tractores	Regular
feb-2020	Argentina	Santa Fe	San Vicente	45	Mi cuenta	Sucursal Plus	Maquinaria Agrícola	Fertilizadoras	Malo

Tabla 2: Muestra de la base de datos de *performance*.

Como puede observarse cada registro representa una cuenta en un mes en particular, y no una fecha puntual por lo que se decidió crear variables tendencistas

que permitieran observar el comportamiento del rendimiento de los clientes en la plataforma *online* a lo largo de un lapso de tiempo determinado. En este caso, se utilizaron 30, 60 y 90 días antes desde la fecha de creación de la negociación inmediata a la primera ganada (*'fecha_opt_label'*).

Así también, la base de productos presenta la información de una manera similar a la base de *performance*. Un registro representa un producto de una cuenta en un mes en particular. A continuación en la siguiente tabla, un ejemplo de las primeras columnas y filas que componen dicha base:

MonthYear	Country	Merchant	Merchant ID	Model type	Product ID	Cantidad de pr	Product status	Categoría N1	Categoría N2
feb-20	Bolivia	Fertec	2	No transactional	2	1	Activo	Maquinaria Agrícola	Fertilizadoras
feb-20	Paraguay	Fertec	2	No transactional	2	1	Activo	Maquinaria Agrícola	Fertilizadoras
feb-20	Uruguay	Fertec	2	No transactional	2	1	Activo	Maquinaria Agrícola	Fertilizadoras
feb-20	Bolivia	Fertec	2	No transactional	3	1	Activo	Maquinaria Agrícola	Fertilizadoras
feb-20	Paraguay	Fertec	2	No transactional	3	1	Activo	Maquinaria Agrícola	Fertilizadoras
feb-20	Uruguay	Fertec	2	No transactional	3	1	Activo	Maquinaria Agrícola	Fertilizadoras
feb-20	Bolivia	Fertec	2	No transactional	5	1	Activo	Maquinaria Agrícola	Fertilizadoras
feb-20	Uruguay	Fertec	2	No transactional	5	1	Activo	Maquinaria Agrícola	Fertilizadoras
feb-20	Bolivia	Fertec	2	No transactional	6	1	Inactivo	Maquinaria Agrícola	Fertilizadoras
feb-20	Paraguay	Fertec	2	No transactional	6	1	Inactivo	Maquinaria Agrícola	Fertilizadoras
feb-20	Bolivia	Fertec	2	No transactional	7	1	Activo	Maquinaria Agrícola	Fertilizadoras
feb-20	Paraguay	Fertec	2	No transactional	7	1	Activo	Maquinaria Agrícola	Fertilizadoras
feb-20	Uruguay	Fertec	2	No transactional	7	1	Activo	Maquinaria Agrícola	Estercoleras
feb-20	Bolivia	Fertec	2	No transactional	8	1	Activo	Maquinaria Agrícola	Estercoleras
feb-20	Paraguay	Fertec	2	No transactional	8	1	Activo	Maquinaria Agrícola	Estercoleras
feb-20	Uruguay	Fertec	2	No transactional	8	1	Activo	Maquinaria Agrícola	Fertilizadoras
feb-20	Bolivia	Fertec	2	No transactional	9	1	Activo	Maquinaria Agrícola	Estercoleras

Tabla 3: Muestra de la base de datos de productos.

Antes de proceder al armado propiamente dicho de las variables de tendencia, se transformaron algunas otras de la base de productos para poder incluirlas de una manera correcta. En las siguientes líneas se explica cuáles fueron las transformaciones pertinentes.

Al igual que se utilizó la técnica de *One Hot Encoding* para la variable *'task subtype'* en la base de actividades, se la volverá a utilizar aplicándola en las *variables 'model type'* y *'status'*. Primero es necesario entender cómo están compuestas estas variables y que significado tienen sus distinas categorías para el negocio.

Existen diferentes tipos de membresías que pueden adquirir los clientes para publicar sus productos en la plataforma *online*. Estas difieren en varios aspectos que fueron explicados anteriormente, pero existe un ítem a destacar que es el tipo de

modelo de venta contratado (*'model type'*). El mismo abarca tres opciones posibles al momento de seleccionar un producto en particular por parte del comprador en la sección del vendedor:

- **No transaccional:** los compradores solamente pueden hacer una consulta sobre el producto. Se le solicita únicamente datos particulares (nombre, teléfono y *email*) del interesado.

- **Cotizable:** los compradores pueden seleccionar una cantidad del producto del cuál presentan interés para solicitar al vendedor una cotización. Pregunta no sólo por los datos particulares del interesado, sino también los de entrega y formas de pago. Por lo tanto, se genera un pedido de compra que le llega al vendedor.

- **Transaccional:** sigue los mismos pasos que la opción 'cotizable' con la diferencia de que al final del proceso se confirma la compra en lugar de realizarse un pedido de compra.

Por lo tanto, con la técnica de *One Hot Encoding* se agrupó para cada cliente la cantidad de productos ofrecidos en cada tipo de modelo de venta. Cabe aclarar que, un mismo vendedor puede tener distintos tipos de modelo de venta aplicados a distintos productos en función de su estrategia de negocios. De esta manera, se obtienen las variables que representan la cantidad de productos ofrecidos en forma no transaccional, cotizable y transaccional: *'cant_prod_no_transactional'*, *'cant_prod_quotable'* y *'cant_prod_transactional'*.

Además, se aplicó la misma técnica para la variable *'product status'*, la cuál hace referencia a la inactividad o no de los diferentes productos publicados por el vendedor. Al contarse con el historial de productos del cliente a lo largo del tiempo, puede existir la posibilidad de que haya tenido la necesidad de desactivar de su sección dentro de la plataforma algunos productos y mantener otros. Así, se agrupó para cada cliente la cantidad de productos activos e inactivos, obteniéndose las dos variables que lo representan: *'cant_prod_activos'* y *'cant_prod_inactivos'*.

En consecuencia, utilizando las variables de la base de *performance* y las creadas a través del *One Hot Encoding* de la base de productos y actividades, se procedió a crear las propiamente dichas variables de tendencia con 30, 60 y 90 días de anterioridad de la fecha de creación de la negociación inmediata a la primera ganada (*'fecha_opt_label'*). A modo de ejemplo se pueden mencionar algunas variables cómo: *'total_impressions_30d_label', 'total_impressions_60d_label', 'total_impressions_90d_label', 'leads_email_30d_label', 'leads_email_60d_label', 'leads_email_90d_label'*, *cant_prod_no_transactional_30d_label*, etcetera. En el anexo se brindará un detalle de todas las variables pertinentes.

Un inconveniente que se presentó al momento del armado de estas variables, fue que las fechas de creación de las negociaciones eran fechas exactas. Es decir día, mes y año, mientras que las fechas de la base de *performance* y productos estaban compuestas solamente por el mes y año, presentando para el día, el primer día del mes, es decir 1. Por lo tanto, cuándo por ejemplo una tupla tenía como *'fecha_opt_label'* el 03/03/2022, al restarle los 30 días para la generación de las variables de tendencia, la fecha que se originaría es 03/02/2022, y como el resto de las variables de *performance* y productos se encuentran con el formato 01/02/2022 no iba a estar seleccionando correctamente el mes de febrero, sino marzo. Entonces, para solucionarlo se hizo una transformación para que se considere el primer día del mes en la variable *'fecha_opt_label'*.

Por último al *dataset* principal se añadieron las variables necesarias de las bases de mensajes. Primero, es necesario explicar cómo se presentan estos datos. Se cuenta con una base por mes que contiene todos los registros de los distintos tipos de mensajes que recibió el cliente sobre sus productos a lo largo del período de tiempo en que los tenía publicado en la plataforma online. En la siguiente tabla se muestra a modo de ejemplo las primeras columnas y filas de estas bases:

Date	Hour	User Unique id	User Agro	Count User Unique	Transactions type	Merchant id	Buyer Message
06/02/2022	16:08:46	Argentina|154432	Yes	1	Messages	12573	Hola, estoy interesado en este
01/02/2022	18:13:19	Argentina|14367	Yes	1	Messages	12573	Buenas tardes. Necesito cotizac
23/02/2022	13:36:57	Argentina|37237	Yes	1	Messages	509404	Hola, estoy interesado en este
19/02/2022	13:30:50	Argentina|51285	Yes	1	Messages	509404	Hola, estoy interesado en este
10/02/2022	13:56:36	Argentina|534490	Yes	1	Messages	509404	Hola, estoy interesado en este
08/02/2022	22:44:19	Argentina|534490	Yes	1	Messages	509404	Hola, sigue disponible? Podrás
16/02/2022	18:47:00	Argentina|569079	No	1	Messages	267565	Hola, estoy interesado en este
24/02/2022	13:08:43	Argentina|142563	Yes	1	Messages	490019	Hola, estoy interesado en este
09/02/2022	13:00:19	Argentina|567018	Yes	1	Messages	339872	Hola, estoy interesado en este
24/02/2022	23:12:29	Argentina|438001	Yes	1	Messages	21194	Hola, estoy interesado en este
11/02/2022	22:32:24	Argentina|554660	Yes	1	Messages	21194	Hola, estoy interesado en este
11/02/2022	22:30:54	Argentina|554660	Yes	1	Messages	21194	Hola, estoy interesado en este
08/02/2022	10:24:32	Argentina|2620	Yes	1	Messages	21194	Hola, estoy interesado en este
08/02/2022	10:20:58	Argentina|554660	Yes	1	Messages	21194	Hola, estoy interesado en este
04/02/2022	01:38:21	Argentina|20170	Yes	1	Messages	21194	Hola,cuanto vale el tractor?.
23/02/2022	19:47:47	Argentina|552546	Yes	1	Messages	352418	Hola, estoy interesado en este
15/02/2022	12:18:09	Argentina|559753	Yes	1	Messages	352418	Hola, estoy interesado en este
15/02/2022	12:05:38	Argentina|559753	Yes	1	Messages	352418	Hola, estoy interesado en este
10/02/2022	14:18:50	Argentina|178915	Yes	1	Messages	352418	Hola, estoy interesado en este

Tabla 4: Muestra de la base de datos de mensajes correspondiente al mes de febrero 2022.

En primer lugar, se procedió a unificar todas las bases en un único *data frame* y a partir de ahí se seleccionaron distintas variables. Por un lado, las variables *'leído'* y *'respondido',* que a nivel negocio representan si el mensaje enviado por parte del potencial comprador fue leído o respondido por parte del vendedor. Por el otro, la variable *'transactions type'* la cuál hace referencia al tipo de interacción que comenzó el comprador ya sea un mensaje, un pedido de cotización o un pedido de compra. Resulta importante aclarar que únicamente los registros de tipo 'mensajes' presentan completo los campos 'leído y 'respondido'.

Una vez más se utilizó la técnica *One Hot Encoding* para la variable *'transactions type'* con el objetivo de conseguir una agrupación para cada cliente de la cantidad de distintos tipos de interacciones que presentó por parte del comprador, en la ventana de tiempo detallada en párrafos anteriores. Se obtuvieron las siguientes variables: *'msj_cant_messages', 'msj_cant_quoatations'* y *'msj_cant_sales_direct'*.

A su vez, para las variables 'leído' y 'respondido' , se definió un valor de '1' cuando definitivamente la categoría era 'si' para ambas, y un valor de '0' cuando era un 'no'. Pudiéndose obtener de esta forma la cantidad de mensajes leídos y respondidos por

parte del vendedor. Las variables se llamaron: *'msj_cant_leidos'* y *'msj_cant_respondidos'*.

En conclusión, luego del agregado de todas las variables anteriormente mencionadas, el *dataset* principal quedó conformado por la misma cantidad de registros, 2433, pero con 115 variables. El listado de las mismas se encuentra en el anexo. A continuación se muestra una tabla con las primeras columnas y registros de dicho *dataset* principal:

	merchant_id	oportunidad_actual_id	oportunidad_label_id	fecha_opt_actual	fecha_opt_label	label_tag	label	mkt_place_id	country
0	0012S000024ujL8QAI	0062S00000vn0YdQAI	0062S00000yGbXYQA0	2020-03-05	2021-02-09	Ganada	1	11842	Argentina
1	0012S000024ujL8QAI	0062S00000yGbXYQA0	0062S000010xFuWQAU	2021-02-09	2021-12-27	Ganada	1	11842	Argentina
2	0012S000024uop7QAA	0062S00000wT40fQAC	0062S00000yTP0JQAW	2020-05-26	2021-03-17	Ganada	1	6210	Argentina
3	0012S000024uop7QAA	0062S00000yTP0JQAW	0068a00001FmlCNAAZ	2021-03-17	2022-03-28	Ganada	1	6210	Argentina
4	0012S000024uphRQAQ	0062S00000vo6wkQAA	0062S00000zMomHQAS	2020-04-14	2021-04-07	Ganada	1	11578	Argentina
...	...	...	...	...	...	...	...	...	...
2428	001E000001hXmrXIAS	0062S00000wTjM5QAK	0062S00000zPCSYQA4	2020-06-10	2021-06-03	Ganada	1	8041	Argentina
2429	001E000001hXmt9IAC	0062S00000wnsKgQAI	0062S00000zPTpdQAG	2020-08-07	2021-06-10	Perdida	0	8454	Argentina
2430	001E000001hXmtJIAS	0062S00000wqdgCQAQ	0062S000010Kbg0QAC	2020-10-21	2021-08-23	Perdida	0	1874	Argentina
2431	001E000001hXmu7IAC	0062S00000wV8LoQAK	0062S00000zO6KEQA0	2020-07-16	2021-05-07	Ganada	1	6144	Argentina
2432	001E000001hXmvAIAS	0062S00000wqkS3QAI	0062S000010MU2bQAG	2020-10-23	2021-10-06	Perdida	0	313	Argentina

2433 rows × 115 columns

Tabla 5: Muestra del *dataset* principal.

2.2. Análisis inicial de exploración de datos

En la siguiente sección se presentará un análisis exploratorio de los datos analizados, buscando entender cómo se encuentran compuestos y cuáles son las principales características que deberán ser tenidas en cuenta a la hora de implementar los modelos de predicción.

Con dichos análisis, por un lado se podrá comprender cómo actualmente se están llevando a cabo los procesos de negociación por parte de los ejecutivos comerciales y por el otro, descubrir la variables relevantes para predecir y así incluirlas en los modelos.

Por lo tanto se hará un análisis descriptivo y exploratorio del *dataset* principal utilizando diferentes gráficos y estadísticas. Permitirá identificar las distribuciones de las variables y las relaciones existentes entre ellas. En esta etapa, en líneas generales se logran obtener patrones subyacentes en los datos. Esto conducirá a realizar una serie de transformaciones en las variables.

2.2.1. Conjuntos de entrenamiento, validación y testeo

Antes de proceder con los respectivos análisis que conlleva el EDA (análisis exploratorio de datos, por sus siglas en inglés), se armaron los conjuntos de entrenamiento, testeo y validación.

Esta práctica de partición de los datos es muy importante para luego corroborar con datos desconocidos (*dataset* de validación y testeo) la *performance* de los modelos implementados en el *dataset* de entrenamiento, lo que permite validar las decisiones de modelado cómo lo son la ingeniería de atributos, elección de hiperparámetros, modelos utilizados, etcétera.

Se buscó entender cuál era la distribución de los datos con los que se entrenarán, validarán y testearán los modelos respecto al *label*, la cuál podía tomar los valores de 'ganada' o 'perdida'.

La distribución que se observa en el *dataset* principal con todos los registros (2433) es de un 64.24% para el *label* 'ganada' y un 35.75% para el *label* 'perdida'. Como podrá observarse, no se encuentra perfectamente balanceado. Sin embargo, no se realizará un rebalanceo de las observaciones pero se utilizarán métricas de *performance* adecuadas a estas situaciones.

Luego de observar la distribución, se procede a dividir el *dataset* principal. El 80% de los registros se destinan al de entrenamiento y el 20% restante al de testeo. Luego, al *dataset* de testeo se lo vuelve a dividir en un 50%, quedando mitad de los registros para testeo y la otra mitad para el de validación. Se verifica que el balanceo

detallado anteriormente se mantenga en cada uno de los nuevos conjuntos de datos, obteniendo:

- **Conjunto de datos de entrenamiento:** 64.90% para el *label* 'ganada' y un 35.09% para *label* 'perdida'
- **Conjunto de datos de validación:** 60.24% para el *label* 'ganada' y un 39.75% para *label* 'perdida'
- **Conjunto de datos de testeo:** 62.96% para el *label* 'ganada' y un 37.03% para *label* 'perdida'

En consecuencia, se puede decir que ambos *set* de datos mantienen una distribución similar con respecto al *label*. El *dataset* de entrenamiento quedó con 1946 registros, validación con 244 y testeo con 243.

2.2.2. Calidad de los datos

Antes de comenzar con el análisis exploratorio de los datos propiamente dicho, se llevaron a cabo una serie de cambios en las variables del *dataset* de entrenamiento, debido a que se encontraron una serie de problemas que llevan a tener una calidad de datos no aceptable para incorporar a los modelos. También se describe qué solución se encontró para los mismos. Es importante tener en cuenta que este tratamiento de datos puede realizarse en esta instancia ya que existe un conocimiento previo del contexto del negocio.

Estas modificaciones se realizaron en el *dataset* de entrenamiento para no generar un sesgo sobre los datos de validación y testeo.

A continuación se detallarán los inconvenientes encontrados, cuáles fueron las variables involucradas y que solución se encontró.

Existencia de valores nulos y el guión '-' en lugar de un dato nulo

En el anexo (Variables del *dataset* principal) se muestra una tabla dónde se puede ver en detalle la cantidad de valores nulos existentes por cada una de las variables.

El tratamiento de los valores nulos se realizó tanto para variables categóricas como numéricas.

Se comienza con las categóricas, que se listan seguidamente con sus respectivos valores nulos. Tener en cuenta que el *dataset* de entrenamiento quedó conformado por 1946 registros luego de llevar adelante la división del *dataset* principal en los conjuntos de entrenamiento, testeo y validación.

- ***'subcategoria_t1'*:** 132

Esta variable presenta solamente un 6,8% aproximadamente de datos nulos y en este caso se decidió crear una categoría llamada 'Otros' ya que luego será tratada por variables dummies. Más adelante se explicará en detalle cómo se realizó esta transformación y se entenderá el motivo de elección de esta solución.

- ***'camp_activa_t1'*** y ***'camp_activa_t2'***: 1473

Ambas variables revelan un 75,7% aproximadamente de datos nulos cada una. Esto se debe a que durante los meses en que se llevó a cabo el análisis para esta tesis no se propusieron muchas campañas de incentivos por parte de la empresa. Por lo tanto, simplemente se creó una variable binaria, tomando el valor de '1' cuándo existió una campaña, y '0' cuando no, transformando esta variable en numérica.

- ***'tipo_empresa'***: 2

Dicha variable únicamente presenta 2 valores nulos. Esto se debe simplemente a un problema de carga en el sistema. Por lo tanto, se asignó el valor ya existente llamado 'Otros'.

- ***'region'*:** 115

Esta variable presenta un 5,9% aproximadamente de los datos nulos, y al igual que para la variable *'subcategoria_t1'* se procedió a crear una categoría llamada 'Otros'. Como así también a los registros que aparecen con un guión '-' se los asignó a 'Otros'.

- ***'membership_30d_label'***: 115, ***'membership_60d_label'***: 106 y ***'membership_90d_label'***: 131

Estas tres variables de tendencia creadas para '*membership'* presentan un 5,9%,

5,4% y 6,73% respectivamente de valores nulos. Aquí se realiza lo mismo que para la variable '*región'*, es decir se asignan los nulos a una nueva categoría 'Otros', como así también los registros que contienen un guión '-'.

- ***'cat_n1'*:** 115 y ***'cat_n2'*:** 115

A nivel de negocio estas dos variables representan lo mismo que la variable *'categoria_t1'* y *'subcategoria_t1'* respectivamente, por lo tanto se decidió eliminarlas del modelo para no agregar información redundante.

- ***'engagement_30d_label'***: 115, ***'engagement_60d_label'***: 106 y ***'engagement_90d_label'***: 131

Nuevamente, estas tres variables de tendencia creadas para '*engagement'* presentan un 5,9%, 5,4% y 6,73% respectivamente de valores nulos. En este caso, se generó una variable numérica ordinal, dónde los valores nulos, los '-' y la categoría 'Sin Leads' serán representados por un '0'. Luego, se hizo una asignación de '5,' '4', '3', '2', '1', para los valores 'Excelente', 'Bueno', 'Regular', 'Malo' y 'Muy Malo' respectivamente, transformando esta variable en numérica.

Se continúa haciendo el tratamiento de valores nulos con las variables numéricas:

- **Variables con los sufijos _30d, _60d y _90d listadas en el anexo con sus valores nulos**

Para todas estas variables que fueron creadas para poder analizar la tendencia, sus valores nulos y registros que contaban con un '-' fueron reemplazados por un '0'. Esto es así porque probablemente hubo casos en que algunos clientes se dieron de baja antes de finalizar el contrato de la membresía que tenían estipulado, y por lo tanto no queda registrada actividad en la plataforma *online*.

Debido a la existencia de registros con valores '-', las variables que los contienen aparecen listadas en el anexo como tipo 'object' en lugar de 'float64'. Luego de tratar estos valores con un '0' pasarán a quedar definidas como numéricas correctamente.

A su vez, dentro de este grupo de variables tendencistas hay algunas como: ***'cant_prod_no_transactional'***, ***'cant_prod_quotable'***, ***'cant_prod_transactional'***, ***'cant_prod_activos'*** y ***'cant_prod_inactivos'***, con sus respectivos sufijos _30d, _60d

y _90d, que se les imputó un valor de '0' a los valores nulos porque en este caso el cliente directamente no tiene productos publicitados en la plataforma, sino que simplemente adquirió una membresía que le permitía únicamente realizar publicidad en forma de banners.

Como puede observarse en el anexo las variables que surgieron de cada una de las distintas bases de datos (*performance* y productos) presentan la misma cantidad de valores nulos para cada uno de los grupos de tendencia armados (30, 60 y 90 días), lo que indica que fue confeccionado de una manera correcta.

- ***'activity_frequency_mean' y 'activity_frequency_median'*****:** 714

Para estas dos variables es necesario explicar en detalle cuál fue el problema encontrado y su solución.

Se presentaban cuentas que no tenían ninguna actividad cargada. Dejarlas con un valor de '0' no era correcto, ya que si existiera el caso de que una cuenta con una actividad realizada por día, su frecuencia daría un valor de '1', y por lo tanto ese valor de '0' se encontraría muy cercano a '1' y en la realidad no representarían situaciones similares. Una cuenta con una frecuencia de actividades de '1', puede considerarse un cliente que presenta una alta interacción con los ejecutivos comerciales, mientras que una cuenta con una frecuencia de '0', implica nada de relacionamiento con el cliente. De esta manera, se observó un máximo de '546' días entre dos actividades que se encontraban registradas en el sistema para una cuenta, por lo que se tomó la decisión de establecer por *default* un valor de '365' días a aquellos registros que no presentaban actividad, lo que indica que la frecuencia de comunicación entre el cliente y el ejecutivo es casi inexistente.

- ***'msj_cant_leidos'***, ***'msj_cant_respondidos'***, ***'msj_cant_messages'***, ***'msj_cant_quotations'***, ***'msj_cant_sales_direct'*** y ***'msj_cant_whatsap'*****:** 93

En este caso estas variables presentan un 4.77% aproximadamente de valores nulos y se les imputó un '0' ya que la ausencia de valor indica que el cliente no recibió ningún mensaje en el lapso de tiempo que duró su membresía.

- ***'sum_task', 'sum_email'* y *'sum_call'***

Para estás variables se decidió asignarles el valor de '0' a los valores nulos ya que la ausencia de datos representa la no interacción entre el ejecutivo comercial y el cliente. Este paso se realizó en el momento del armado del *dataset* principal, por lo que en la tabla del anexo se puede observar que no existen valores nulos.

Variables de tipo numéricas que aparecen como categóricas.

Se encontraron algunas variables categóricas que deberían ser numéricas. Esto se debe a que contenían registros con valores cómo un guión '-', por lo que se los trató como nulos y se hizo su imputación a '0', transformándose en numéricas correctamente (los casos se encuentran detallados en los párrafos anteriores).

Variables categóricas con una gran cantidad de valores distintos

En el *dataset* se encuentran algunas variables categóricas, que si bien no son la mayoría, es necesario hacerles un tratamiento para poder incluirlas en los modelos que se realizarán, considerando que los algoritmos de aprendizaje automático solamente aceptan atributos numéricos. Por lo tanto, primero se analizó cómo estaban compuestas y se las agrupó a fin de reducir la cantidad categorías que las componían. Luego, en la sección 2.3.1 (Transformación de datos e ingeniería de atributos) se procedió a transformarlas en variables *dummies*.

Una cuestión importante a tener en cuenta es que para aquellas categorías dentro de cada una de las variables que no representen más del 10% respecto al total de valores, se los asignó a una categoría llamada 'Otros', en excepción algunos casos que por contexto de negocio se pueden analizar de una manera diferente y se puede establecer un porcentaje menor.

A continuación se especificará para cada una de ellas cómo se realizó la modificación:

- ***'country'***

Estos son las distintas categorías y cantidad de registros presentes en esta variable:

- ➢ Argentina: 1475
- ➢ Brasil: 361
- ➢ Uruguay: 61
- ➢ Colombia: 48

Como podrá observarse la mayoría de las observaciones se concentran en los países de Argentina y Brasil, quedando los restantes representados como 'Otros'.

- ***'rubro_t1'***

Se detallan las diferentes categorías y cantidad de registros encontrados en estas variable:

- ➢ *Business*: 1479
- ➢ *Corporate*: 253
- ➢ *Advertising*: 91
- ➢ Demo: 89
- ➢ *Pack* Agro: 18 ➢ Sitios: 9
- ➢ Contenidos: 6
- ➢ *Mailing*: 1

Para esta situación, teniendo en cuenta que a nivel de negocio los rubros llamados '*Advertising'* y 'Sitios' hacen referencia al mismo paquete de membresía en cuánto a lo que ofrecen, se decidió llamar a la categoría 'Sitios' como '*Advertising'*.

Después, en lugar de tomar un 10% como *default* para buscar los registros que deberían ser incluidos en la categoría 'Otros', se utilizó un 5% ya que de esta manera se logra mantener el valor de '*Advertising'*. De caso contrario, quedaría incluido en el valor 'Otros', lo cuál no estaría bien ya que es un rubro importante a considerar por parte de la empresa para el análisis.

Por lo tanto se armaron los grupos: *Business*, *Corporate* y *Advertising.*

- ***'division_t1'***

Al analizar esta variable, se pueden destacar estas categorías y cantidad de registros:

- ➢ *Business* Argentina: 1110
- ➢ *Corporate* Argentina: 299
- ➢ *Business* Brasil: 253
- ➢ OLAC: 98
- ➢ *Planning*: 98
- ➢ *Corporate* Brasil: 88

Se decidió eliminarla ya que contiene información redundante que ya se encuentra analizada en las variables *'country'* y *'rubro_t1'.*

- ***'currency_t1'***

Sus categorías y cantidad de registros son los siguientes y no se realizaron cambios:

- ➢ ARS: 1478
- ➢ BRL: 358
- ➢ USD: 110

- ***'categoria_t1'***

Esta variable a diferencia de las otras, se encuentra conformado por una mayor cantidad de categorías, que se muestran a continuación:

- ➢ Maquinaria: 590
- ➢ Infraestructura: 326
- ➢ Campos: 217
- ➢ Rodados: 177
- ➢ Herramientas: 167
- ➢ Repuestos: 92
- ➢ Agencia de Publicidad: 87
- ➢ Insumos Agrícolas: 60
- ➢ Accesorios para vehículos: 60

- ➢ Maquinaria Vial: 48
- ➢ Tecnología para el Agro: 35
- ➢ Construcción: 25
- ➢ Servicios: 10
- ➢ Insumos Veterinarios: 8
- ➢ Seguros: 8

  Universidad: 8

  Bancos: 8
- ➢ Insumos de Máquinas y Vehículos: 7
- ➢ Equipamiento industrial: 4 ➢ Agricultura de Precisión:
- ➢ Merchandising: 2
- ➢ Viveros: 2
- ➢ Otros: 1
- ➢ Insumos Industriales: 1

Debido a que se quería mantener las siguientes categorías: 'Maquinaria', 'Infraestructura', 'Campos', 'Rodados' y 'Herramientas' se volvió a ajustar el valor por *default* de 10% por 8%, quedando el resto de los registros imputados a 'Otros'.

- ***'subcategoria_t1'***

Dicha variable presenta incluso muchas más categorías, y a diferencia de la anterior. Se puede ver en el siguiente listado un resumen de los primeros diez y su cantidad de registros. El resto se detalla en el anexo.

- ➢ Campos: 159
- ➢ Tractores: 154
- ➢ Otros: 133
- ➢ Sembradoras: 71
- ➢ Repuestos Agrícolas: 61
- ➢ Instalaciones para Ganadería: 56
- ➢ Acopio y Almacenaje: 54

- Camiones: 48
- Tanques: 48
- Acoplados: 42

Debido a que esta variable es casi unívoca, con un muy baja frecuencia por categoría, se procedió a eliminarla del modelo ya que no estaría aportando suficiente información.

- ***'product_name_t1'***

En el anexo se detallan todos las categorías presentes para esta variable, pero en la siguiente lista se puede ver un resumen de las primeros 10:

- ARG *Advanced* 12 meses: 373
- ARG *Primary* 12 Meses: 345
- ARG *Plus* 12 Meses: 299
- ARG *Select* 12 Meses: 116
- BRA *Primary* 12 Meses: 93
- BRA *Plus* 12 Meses: 53
- BRA Maquinaria *Business* Demo Gratis 3 Meses: 47
- BRA *Advanced* 12 Meses: 43
- ARG *Starting* Financiado 12 meses: 42
- Demo: 38

Frente a esta situación, al tener tantos valores distintos fue necesario hacer un análisis desde el punto de vista del negocio del significado e importancia de cada uno de estos productos. Se hicieron agrupaciones en base a las categorías que contengan las siguientes palabras: '*Advance*', '*Primary*', '*Plus*', '*Select*', '*Starting*', '*Going*', '*Conquer*', '*Move*'. Estos hacen referencia a los distintos tipos de membresías existentes, que se repiten en los diferentes países cambiando la sigla que hace referencia al mismo, pero a nivel de contenido, el servicio contratado por el cliente es el mismo.

Los ocho grupos mencionados anteriormente son los tipos de membresías típicos que se ofrecen a los clientes actualmente. El resto son simplemente productos puntuales que pueden haber sido creados para alguna campaña específica o por la necesidad de algún equipo, por lo tanto se los va a agrupar dentro de la categoría 'Otros'.

A continuación se muestra cómo quedó conformada la variable *´product_name_t1´*:

- *Primary*: 507
- *Advanced*: 442
- *Plus*: 381
- Otros: 254
- *Select*: 145
- *Starting*: 97
- *MoveIn*: 66
- *Going*: 35
- *Conquer*: 19

- ***'tipo_empresa'***

Esta variable no presenta tantas categorías distintas como las últimas mencionadas.

Se pueden observar a continuación:

- ➢ Fábrica: 711
- ➢ Concesionario Oficial: 337
- ➢ Otros: 328
- ➢ Concesionario Multimarca: 291
- ➢ Revendedor / Distribuidor: 180
- ➢ Inmobiliaria: 55
- ➢ Agencia: 40
- ➢ Ferreteria / Buloneria: 4

Se procedió a agruparlas en: 'Fábrica', 'Concesionario Oficial', 'Concesionario Multimarca' y 'Otros', utilizando el valor por defecto del 10% para armar el grupo 'Otros', los cuáles tienen poca representatividad dentro del total.

- ***'región'***

En el anexo se detallan todas las categorías presentes para esta variable, pero en la siguiente lista se puede ver un resumen de las primeras 10:

- ➢ Buenos Aires: 409
- ➢ Santa Fe: 361 ➢ Cordoba: 302
- ➢ Otros: 177
- ➢ São Paulo: 101
- ➢ Buenos Aires City: 93
- ➢ Entre Rios: 70
- ➢ Rio Grande Do Sul: 62
- ➢ Paraná: 50
- ➢ La Pampa: 44

En este caso también se mantuvo el 10% para seleccionar aquellas categorías que tenían una baja frecuencia dentro de la variable, y por lo tanto quedó conformada por los grupos: 'Buenos Aires', 'Santa Fe' y 'Cordoba'.

- ***'membership_30d_label', 'membership_60d_label'*** y ***'membership_90d_label'***

Similar a lo que sucedía con la variable *'product_name_t1'*, en este caso también existen una gran cantidad de cateogrías para estas variables, por lo que fue necesario entender el motivo pensándolo desde el punto de vista del negocio.

En primer lugar, en el anexo se detallan todas las categorías presentes en estas variables.

Luego, es importante aclarar que las membresías del tipo 'Sucursal' corresponden a las ofrecidas por los ejecutivos comerciales que se encuentran dentro del rubro *Business*, es decir, sus clientes son principalmente empresas pequeñas en comparación con las correspondientes al rubro *Corporate*.

Por lo tanto, se armaron tres grupos. Uno para aquellas categorías que contengan la palabra 'Sucursal', otro para los '*Corporate'* y por último el grupo 'Otros' que serían todas aquellas membresías que no están dentro de los paquetes *Business* y *Corporate*.

Nuevamente, estas membresías dentro del grupo 'Otros' pudieron haber sido creadas por alguna situación en particular, pero no refieren a la mayoría de los casos.

En conclusión, con todas estas modificaciones realizadas en las variables que conforman el *dataset* de entrenamiento, se consiguió mejorar la calidad de los datos, y se puede proceder a realizar el análisis exploratorio de los datos propiamente dicho.

2.2.3. Análisis univariante

Una manera de entender rápidamente lo que está sucediendo con el tipo de datos que se están trabajando, es hacer un histograma para las variables numéricas. Estos gráficos muestran el número de instancias, en el eje vertical, que presentan un rango de valores determinados,en el eje horizontal (Gerón Aurélien, 2019).

Tras el armado de los histogramas se pueden detectar algunos aspectos que llamaron la atención.

Para la variable ***'total_price_t1'*** se puede observar la siguiente distribución de los datos:

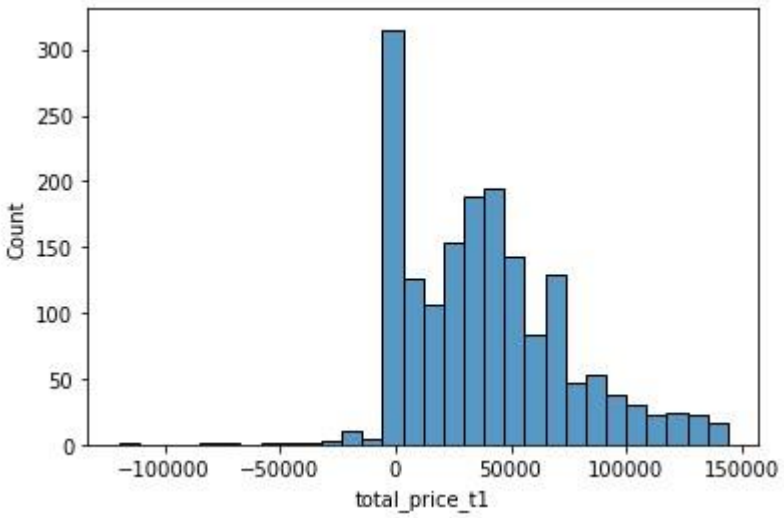

Figura 1: Histograma para la variable *'total_price_t1'*.

Lo primero que resulta interesante destacar los registros con precios negativos cercanos a 0. Esto tiene una explicación, y es que a nivel negocio existen las notas de crédito. Estas son solicitadas por los clientes cuando, por alguna razón, se necesitan dar de baja del contrato de membresía y por lo tanto se les hace una devolución del saldo restante. A nivel de sistema, esto se carga como una negociación con precio negativo por el valor correspondiente.

Luego, pueden observarse varios picos pero todos entre los precios de 80.000 y 30.000, lo que demuestra que las membresías que se adquieren en una mayor proporción son aquellas que no ofrecen los servicios más completos. En línea con esto, era de esperarse que la distribución sea de esta manera, ya que existen una mayor cantidad de clientes del rubro *Business*, quienes suelen adquirir membresías más económicas. En la sección 2.2.4 (Análisis multivariado) se dará más detalle acerca de esta variable y su relación con el target.

Las siguientes variables a analizar son las de tendencia referidas al '***engagement***' con sus respectivos sufijos _30d, _60d y _90d.

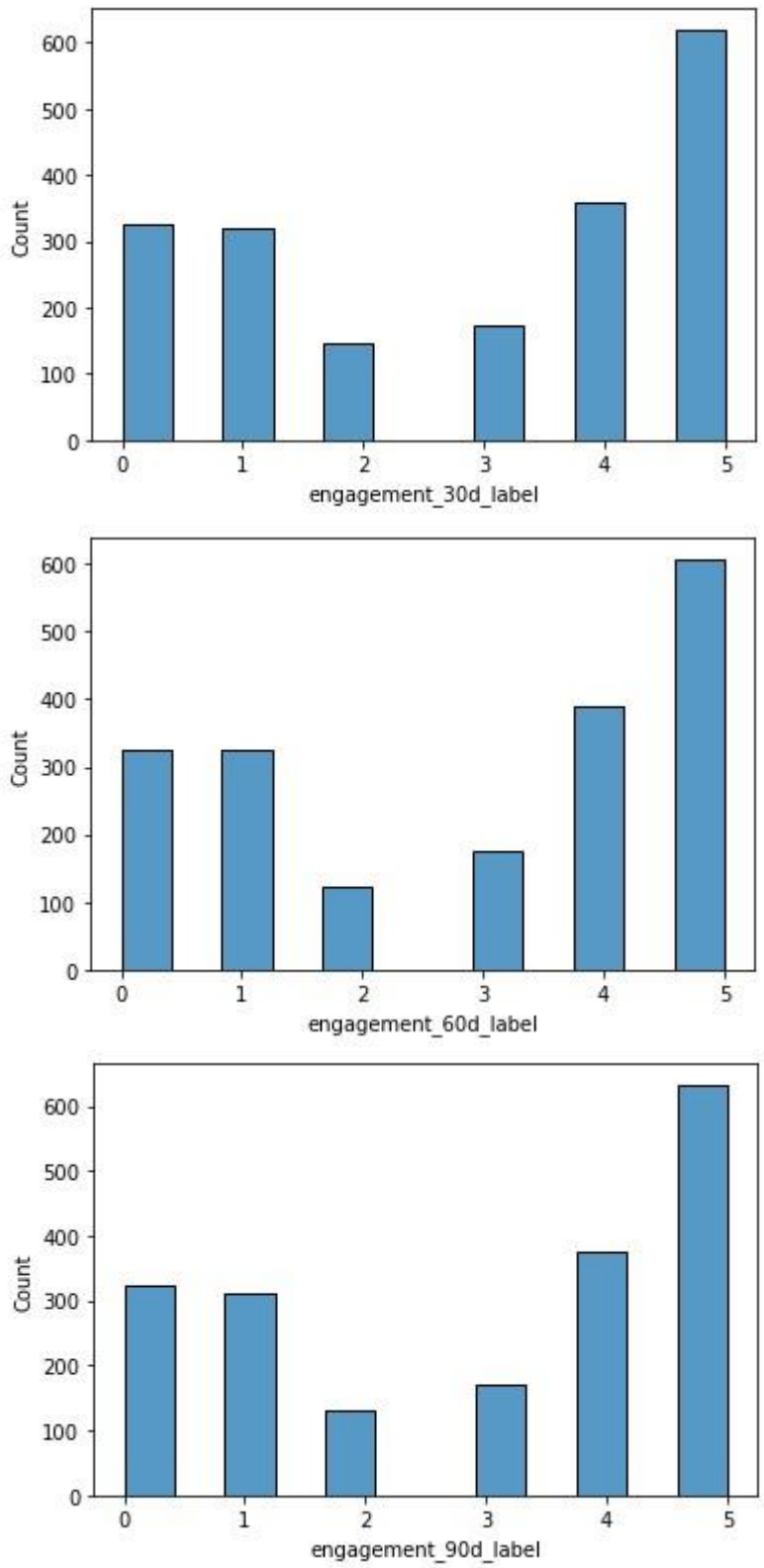

Figura 2: Histogramas para las variables *'engagement_30d_label', 'engagement_60d_label'* y *'engagement_90d_label'*.

Las tres variables presentan una distribución similar en los datos, lo cuál era de esperarse ya que la calificación que se le realiza al cliente en cuánto a su *engagement*

(Excelente, Muy Bueno, Bueno, Regular, Malo y Sin Leads) generalmente se indica una vez sóla dentro de su historial, pero sin embargo puede existir la posibilidad de que se realice alguna modificación. Esto indica que podría dejarse únicamente una variable, ya que sería información redundante para el modelo.

Lo que se puede destacar tras analizar dichos histogramas es que tienden a existir calificaciones principalmente positivas (Excelente y Muy Bueno) o negativas (Malo y Sin *Leads*), es decir el cliente se encuentra altamente motivado e involucrado o no.

En la sección 2.2.4 (Análisis multivariado) se dará más detalle acerca de esta variable.

Se continúa con el análisis univariante de la variable ***'opportunity_duration_t1'***, la cuál hace referencia a lo que en la empresa se llama 'madurez' de la negociación. Representa el tiempo (días) que pasa desde que una negociación se abre y se carga en el sistema, hasta que el ejecutivo logra finalmente cerrarla, es decir, la gana y el cliente contrata algún paquete de membresía o la pierde y no hay servicio adquirido.

Se puede observar su distribución en el siguiente gráfico:

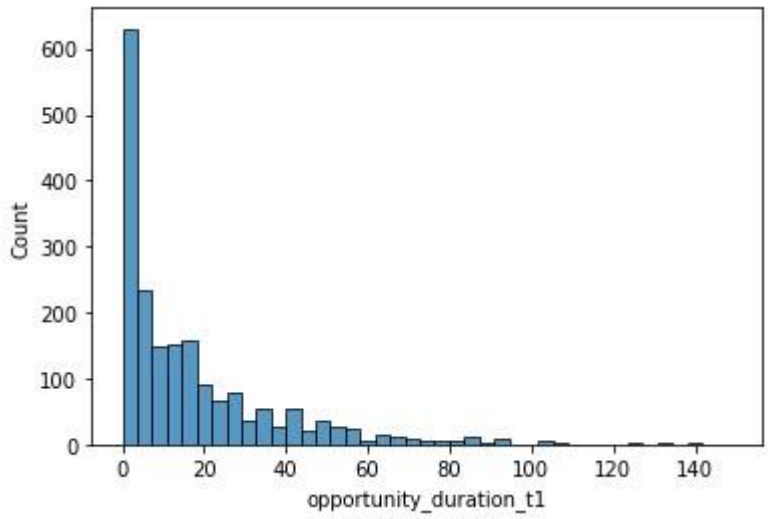

Figura 3: Histograma para la variable *'opportunity_duration_t1'*.

Esta figura es un claro ejemplo de los histogramas *tail-heavy*, es decir la distribución de los datos se extienden principalmente hacia un lado de la media, en este caso hacia la izquierda. A nivel de negocio implica que en su grán mayoría las negociaciones tienden a cerrarse rápidamente luego de ser abiertas, lo que resulta un comportamiento interesante de ser analizado en mayor profundidad. Por lo

tanto, en la sección 2.2.4 (Análisis multivariado) se analizará la relación de dicha variable con el target.

Continuamos avanzando el análisis con las variables ***'activity_frequency_median'*** y ***'activity_frequency_mean'***.

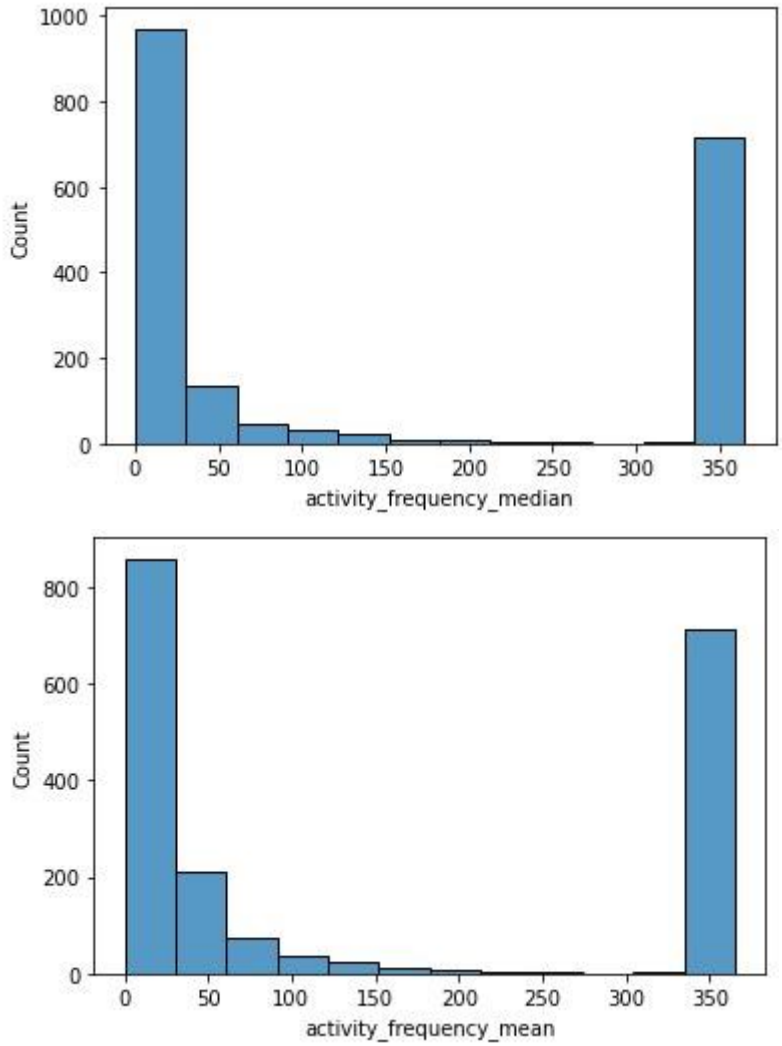

Figura 4: Histograma para las variables *'activity_frequency_median'* y *'activity_frequency_mean'*.

En estos histogramas quedan reflejadas las modificaciones que se realizaron anteriormente en estas dos variables en la sección 2.2.2 análisis de la calidad de los datos. Se había imputado un valor de '365' días para los valores nulos, lo cuál indicaba la muy baja frecuencia de comunicación entre el cliente y el ejecutivo. Existe una distribución con valores atípicos y una disposición de los datos hacía el lado izquierdo.

Hay que destacar que en el caso de existir algún tipo de actividad entre el cliente y el ejecutivo comercial, esta se suele realizar con una alta frecuencia. Es decir, existe un comportamiento por parte del ejecutivo de querer estar cerca del cliente y tener interacción con él. La mayor cantidad de valores se encuentran por debajo de los 60 días, lo que implica que el ejecutivo generalmente interactúa con sus clientes en un lapso menor a los dos meses.

Continuando con la variable ***'msj_cant_messages'*** se puede observar una marcada distribución de los datos hacia la izquierda del histograma:

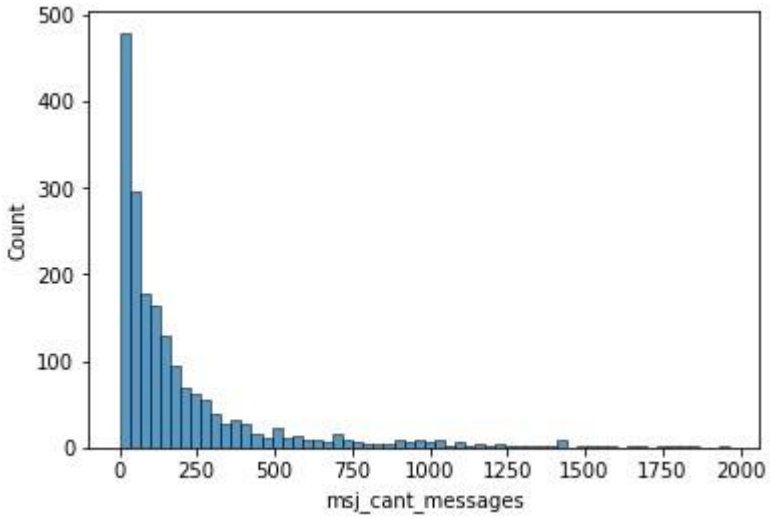

Figura 5: Histograma para la variable *'msj_cant_messages'*.

Esto permite sacar la conclusión de que durante el período de tiempo que dura la membresía, los clientes reciben, en una mayor proporción, una baja cantidad de mensajes de los potenciales clientes. Esta situación hace pensar si está resultando difícil para el cliente poder contactar al vendedor o si simplemente no existe un interés.

En línea con lo anterior, se prosigue analizando la distribución de los datos para las variables ***'msj_cant_leidos'*** y ***'msj_cant_respondidos'***.

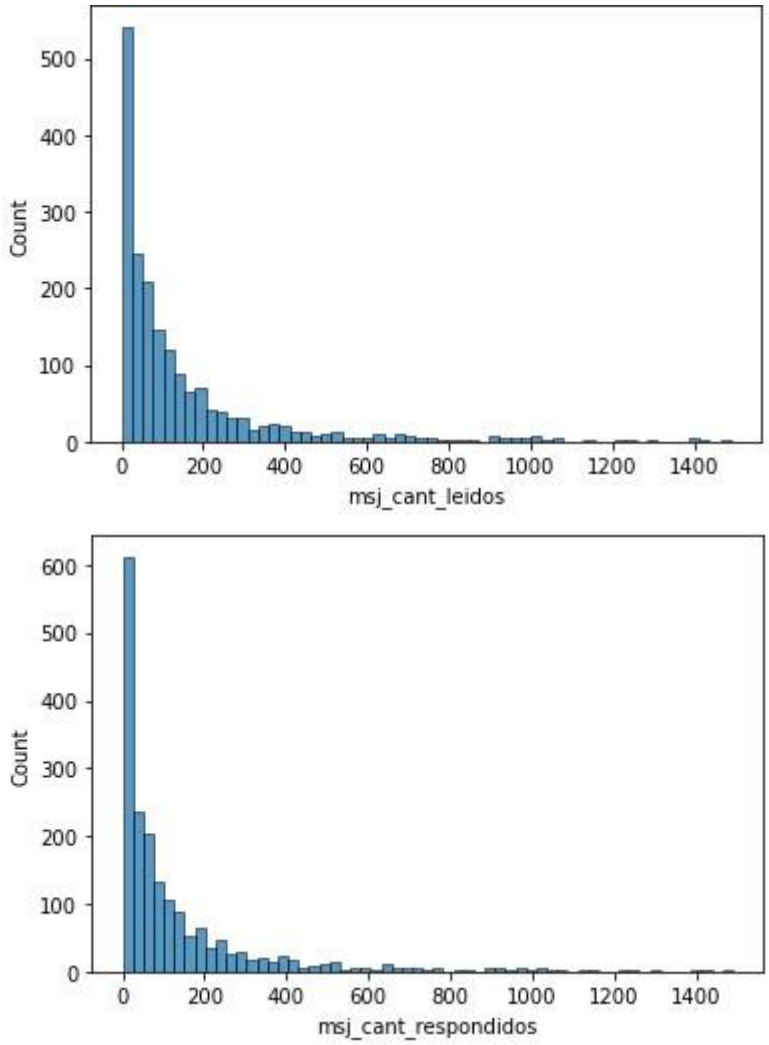

Figura 6: Histogramas para las variables *'msj_cant_leidos'* y *'msj_cant_respondidos'*.

Cómo era de esperarse, el comportamiento para ambas variables es muy similar, teniendo una una marcada distribución hacia la izquierda al igual que sucede con 'msj_cant_messages'**.** Al estar recibiendo pocos mensajes por parte del cliente, el vendedor no va a presentar una gran cantidad de mensajes leídos y respondidos.

2.2.4. Análisis multivariado

Hasta ahora, se realizó un análisis permitió obtener una comprensión general del tipo de datos que se están manipulando, su distribución y distintos aspectos que hacen a la calidad. Ahora, se avanzará hacia un nivel mayor de profundidad.

Se continuará con un análisis multivariado, observando el comportamiento de distintas variables en forma simultánea con el fin de encontrar información relevante en los datos. En un principio, se buscará entender la relación entre las variables categóricas y el *label*. Luego, la correlación existente entre las variables numéricas y por último el comportamiento de los atributos numéricos respecto al *label*.

Resulta interesante entender cuál es la relación entre las variables dependientes e independientes, lo que puede ayudar a identificar posibles enfoques para modelar el problema.

Comenzando con la variable ***'categoria_t1'*** y el *label*, se puede observar la siguiente figura:

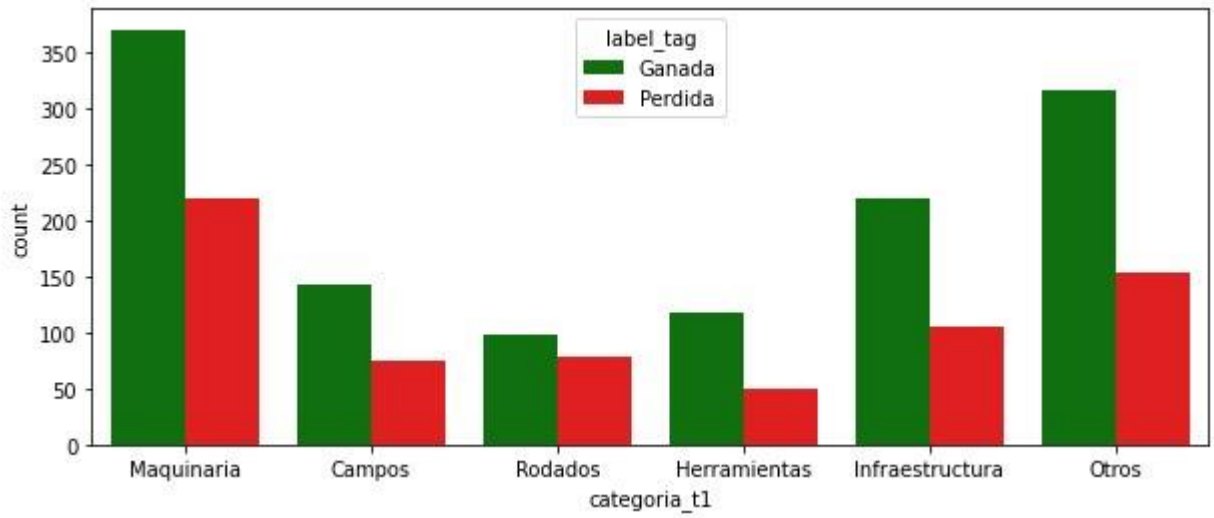

Figura 7: Gráfico de barras para la variable *'categoria_t1'* y el *label*.

Resulta interesante destacar la importancia que cobra la categoría de maquinaria frente al resto de categorías presentes. Es decir, los clientes que tienen una mayor participación en la plataforma son los pertenecientes a este grupo. Pensándolo desde el punto de vista del negocio, era de esperarse que este comportamiento fuera así ya que las bases con las que cuentan los ejecutivos para realizar las negociaciones están conformadas en su mayoría por esta categoría, ya que es la que resulta más sencillo conseguir información.

Luego, para cada una de las categorías siempre predomina la el *label* 'ganada' frente a 'pérdida'.

Seguidamente se analizan las variables ***'country'***, ***'currency'*** y ***'region'*** ya que se espera que sus comportamientos sean similares. Se puede observar la importancia que tiene el país Argentina a nivel de empresa, en comparación con Brasil y los demás clasificados como 'Otros'.

Además, Argentina en comparación con Brasil presenta una proporción de negociaciones ganadas superior respecto a las pérdidas, cuándo en Brasil sucede lo contrario, habiendo más negociaciones perdidas que ganadas.

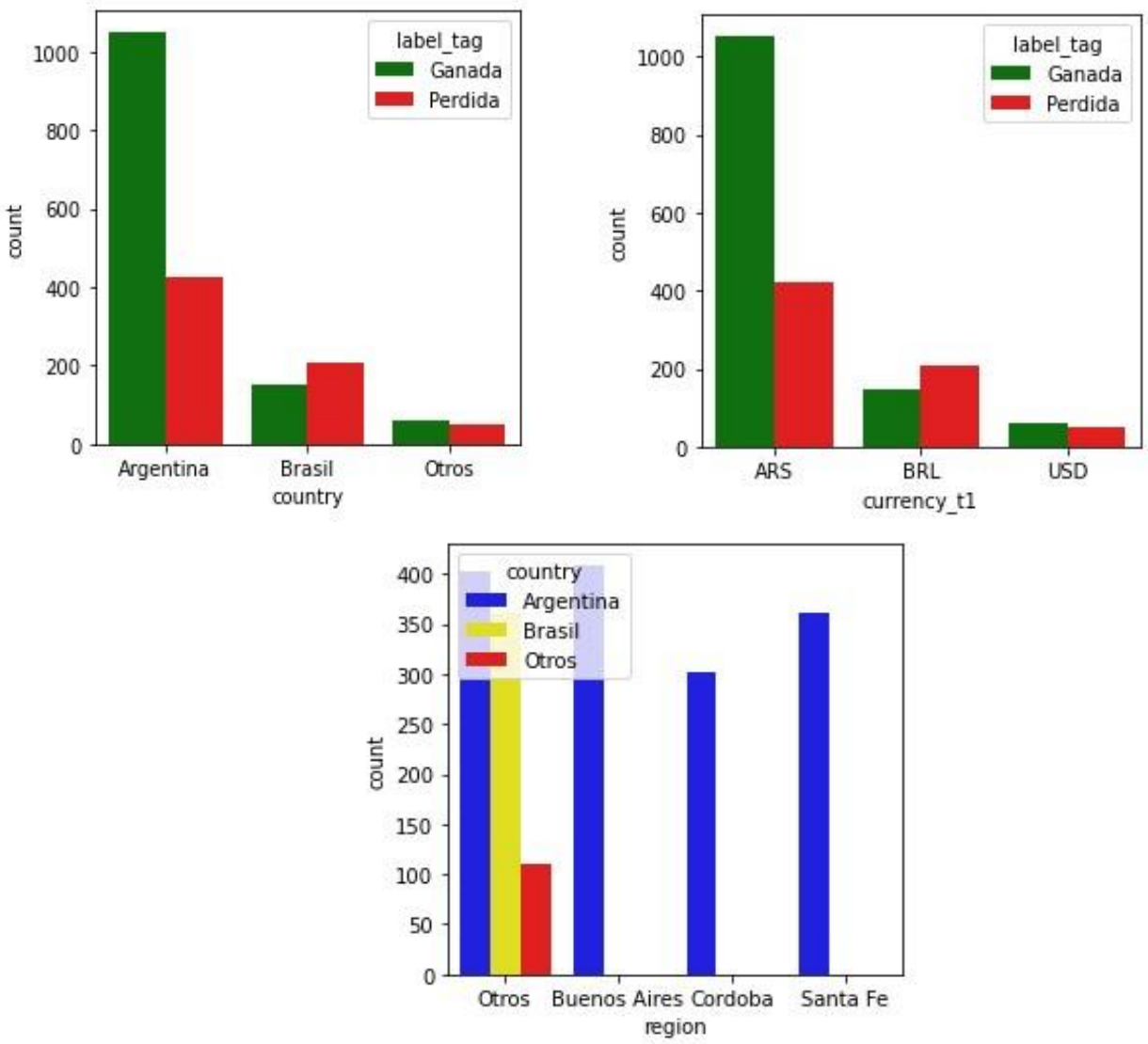

Figura 8: Gráfico de barras para las variables *'country'*, *'currency'*, *'region'* y el *label*.

En esta última figura dónde se muestra la variable 'región' se vuelve a remarcar la importancia de Argentina frente a los demás países, siendo la mayor cantidad de registros pertenecientes a regiones de dicho país.

A continuación, se muestra la preponderancia anteriormente mencionada en otras secciones del rubro *Business* sobre *Corporate*, *Advertising* y 'Otros'.

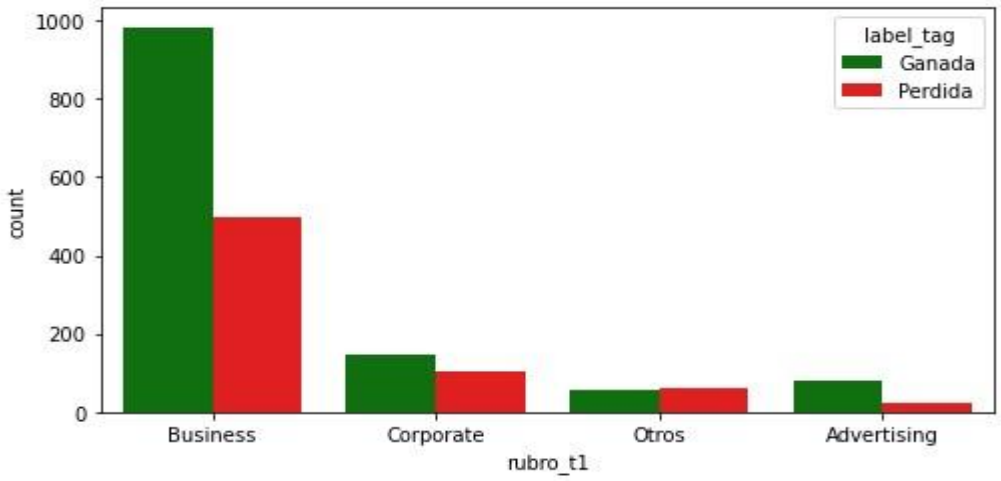

Figura 9: Gráfico de barras para la variable *'rubro_t1'* y el *label*.

Se puede notar para esta variable ***'rubro_t1'*** la gran cantidad negociaciones ganadas respecto a las pérdidas para el rubro *Business*, a diferencia de lo que sucede con el resto. Contextualizando con el negocio, es de esperarse que esto sea así, ya que las membresías ofrecidas a los clientes que conforman este grupo, resultan comparativamente económicas. En cambio, las membresías para los clientes *Corporate* tienen un precio superior y por lo tanto, a los ejecutivos comerciales les se les complica poder venderles el paquete, a pesar de que sean clientes con un mayor poder adquisitivo.

Luego, el rubro *Advertising* no suele tener un impacto grande en lo que son las ventas e ingresos de la empresa. Generalmente, los ejecutivos ofrecen estas membresías cuándo los clientes no tienen suficiente *budget* para destinar a la compra de un paquete *Business* o *Corporate*.

Como se había mencionado en líneas anteriores, el rubro 'Otros', incluye aquellas categorías que conforman la variable que no representan más de un 5% respecto al total de registros.

En línea con lo discutido en estos párrafos, en la siguiente figura se analiza la variable ***'membership_30d_label'*** y se encuentra un comportamiento similar a la variable ***'rubro_t1'***.

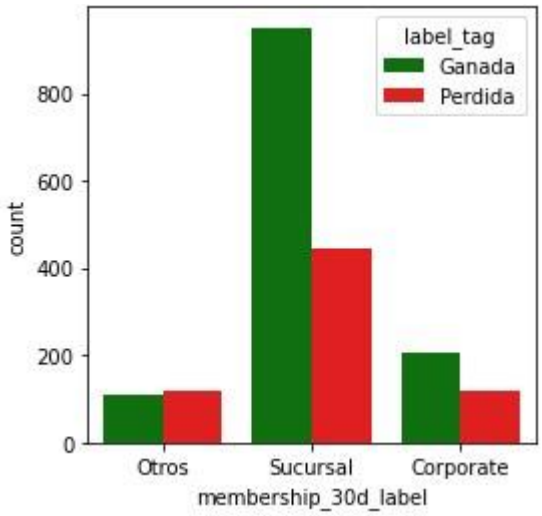

Figura 10: Gráfico de barras para la variable *'membership_30d_label'* y el *label*.

Ocurre esto ya que las membresías catalogadas como 'Sucursal' son aquellas ofrecidas por los ejecutivos comerciales a los clientes del rubro *Business*. Por lo tanto, se vuelve a comprobar que la mayoría de los clientes son de este rubro.

Avanzado con la variable ***'product_name_t1'*** se comprueba una vez más la importancia del rubro *Business* para la empresa. Los productos con nombres cómo '*Advance*', '*Plus*', '*Primary*' y '*Select*' corresponden a membresías *Business* ofrecidas a clientes que ingresan dentro de esta categoría. Mientras que, los que se llaman *'Starting', 'Move In', Conquer'* y *'Going'*, representan las membresías del rubro *Corporate*.

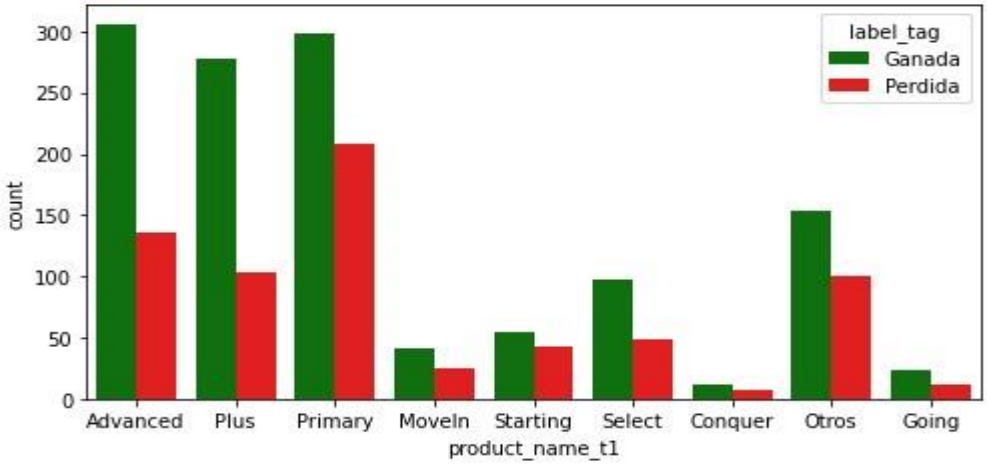

Figura 11: Gráfico de barras para la variable *'product_name_t1'* y el label.

Otro aspecto a destacar es que nuevamente se comprueba la mayor cantidad de negociaciones ganadas respecto a las perdidas para los productos correspondientes al rubro *Business*.

Respecto a las variables ***'camp_activa_t1'*** y ***'camp_activa_t2'*** se puede observar que no existe diferencia entre ambas, tanto para los casos de presencia de una campaña '1' o no '0'.

Hay situaciones en donde la primera negociación ganada realizada a la cuenta puede estar involucrada en una campaña propuesta por la empresa y entonces este campo se reporta en el sistema de gestión de clientes quedando asociado a la cuenta. Luego, cuando se lleva a cabo la renegociación para la contratación de una siguiente membresía, este campo no se modifica y la negociación queda asociada a la campaña anterior, y no necesariamente es así.

Por lo tanto, esto es un problema en la carga de datos por parte de la empresa, ya que debería modificarse dicho campo, en el caso de que la siguiente negociación que se realice al cliente, no esté involucrada en esa campaña a la cual hace referencia la primera negociación.

En el caso de los registros que no tienen una campaña asociada, no existe este problema, y es correcto que ambas variables presenten el mismo comportamiento ya que este dato se reporta en la cuenta asociada y no puntualmente a la negociación realizada.

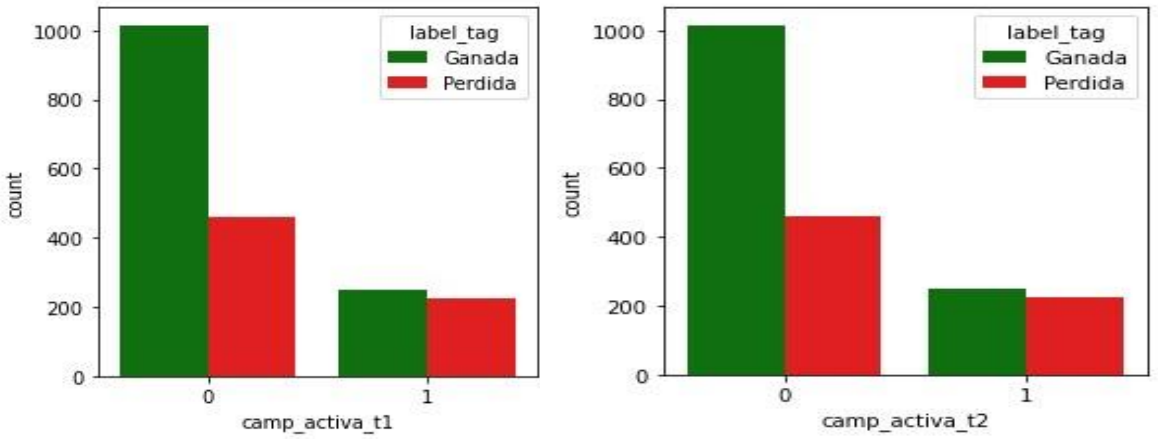

Figura 12: Gráficos de barras para las variables *'camp_act_t1', 'camp_act_t2'* y el

label.

Por último, resta analizar la variable ***'tipo_empresa'***. Para todas las categorías presentes, predominan las negociaciones ganadas sobre las perdidas, pudiendo notar una mayor diferencia para el valor 'Fábrica'.

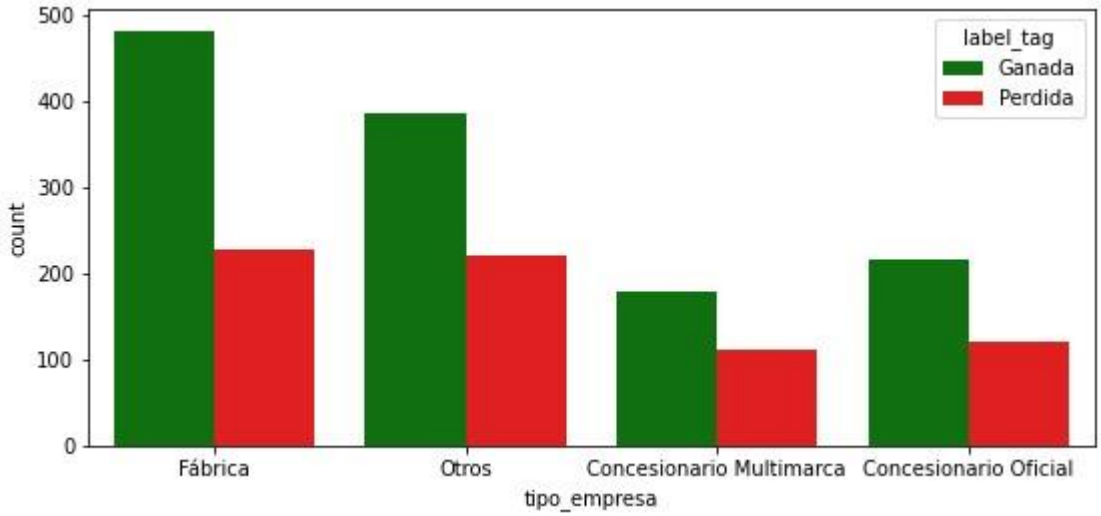

Figura 13: Gráfico de barras para la variable *'tipo_empresa'* y el *label*.

Para comprender mejor las variables y cómo estas se relacionan entre sí, se calculó la correlación existente entre las variables continuas.

La correlación indica el tipo de relación lineal entre cada atributo numérico, si es positiva o negativa, y la fuerza de la relación (si es fuerte, moderada, débil o si no tienen ningún tipo de relación). Según Wilson (n.d.), las variables que tienen una fuerte asociación tienen un coeficiente de correlación entre 0,5 a 1 o -1 a -0,5. En particular, si tienen una correlación lineal perfecta es igual a 1 o -1. Una relación moderada implica un valor de 0,3 a 0,5 o -0,5 a -0,3, y débil un valor de 0,1 a 0,3 o 0,3 a -0,1. Por último, las variables que tienen una relación nula o muy débil tendrán un valor entre -0.1 a 0.1.

Los siguientes *heatmap* detallan las 'Correlaciones de Pearson' para variables numéricas de tendencia creadas, comenzando con las de 30 días, continuando con las de 60 días y finalizando con las de 90.

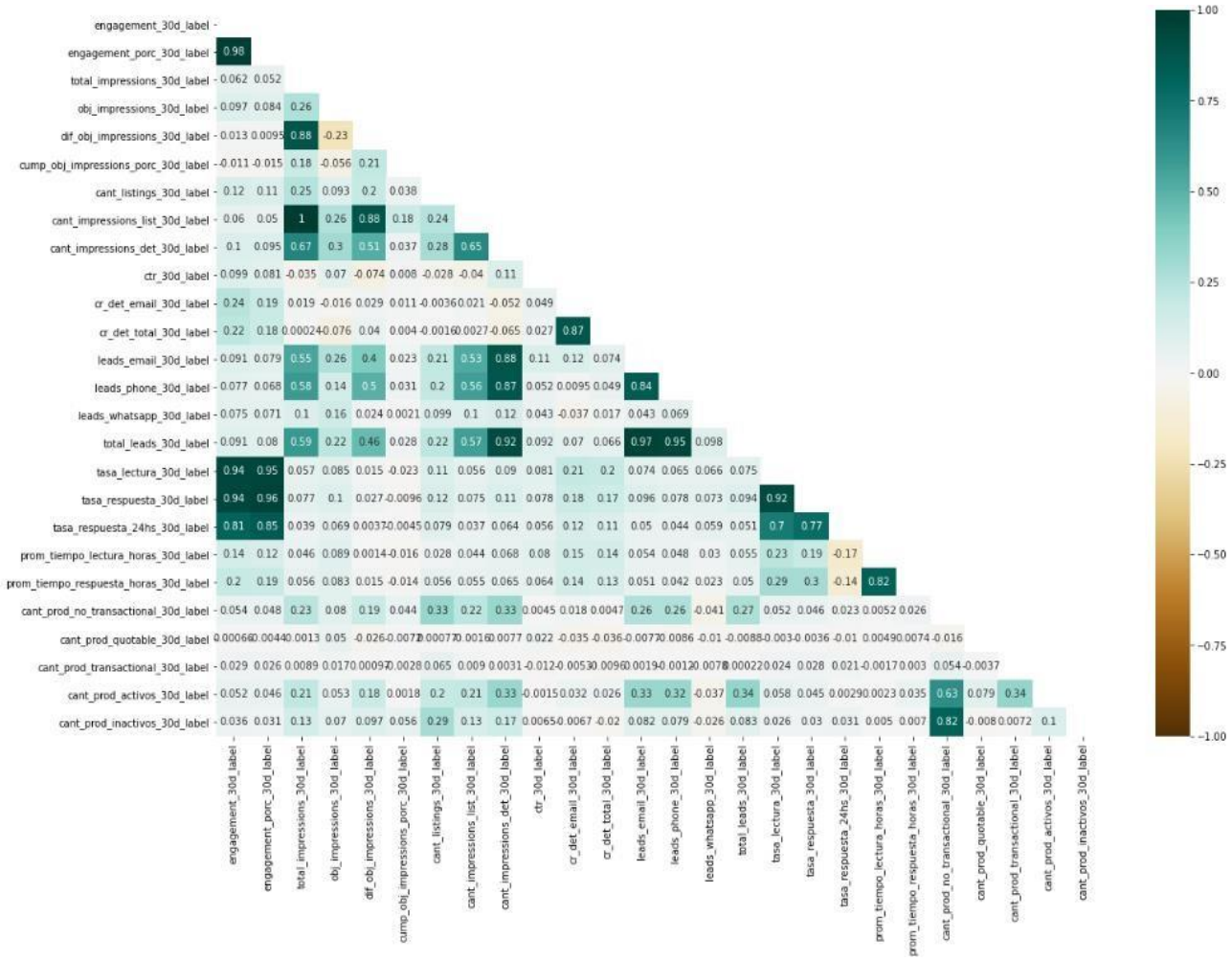

Figura 14: *Heatmap* que detalla la fuerza y el tipo de relación existente entre las variables continuas de tendencia para los 30 días.

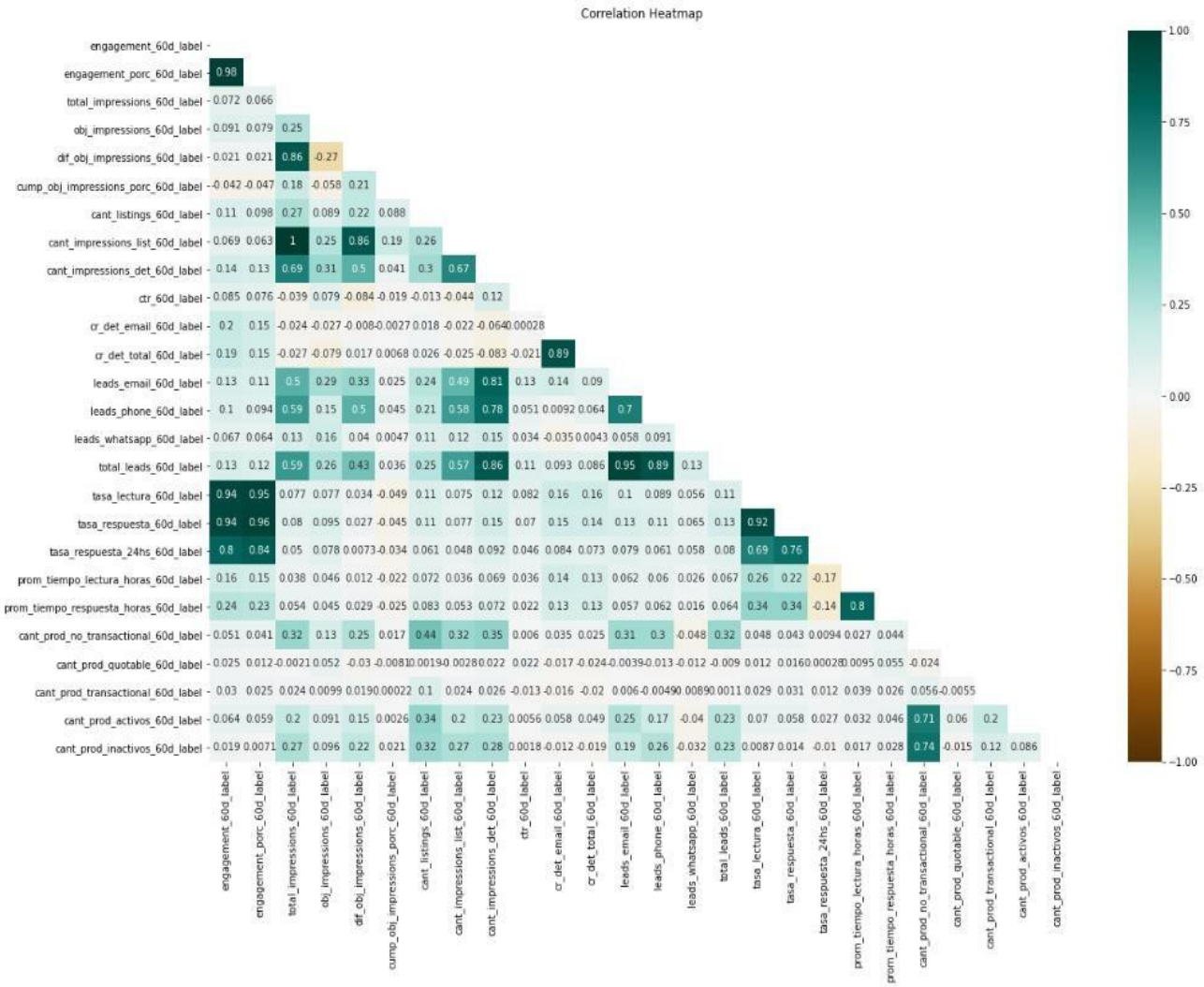

Figura 15: *Heatmap* que detalla la fuerza y el tipo de relación existente entre las variables continuas de tendencia para los 60 días.

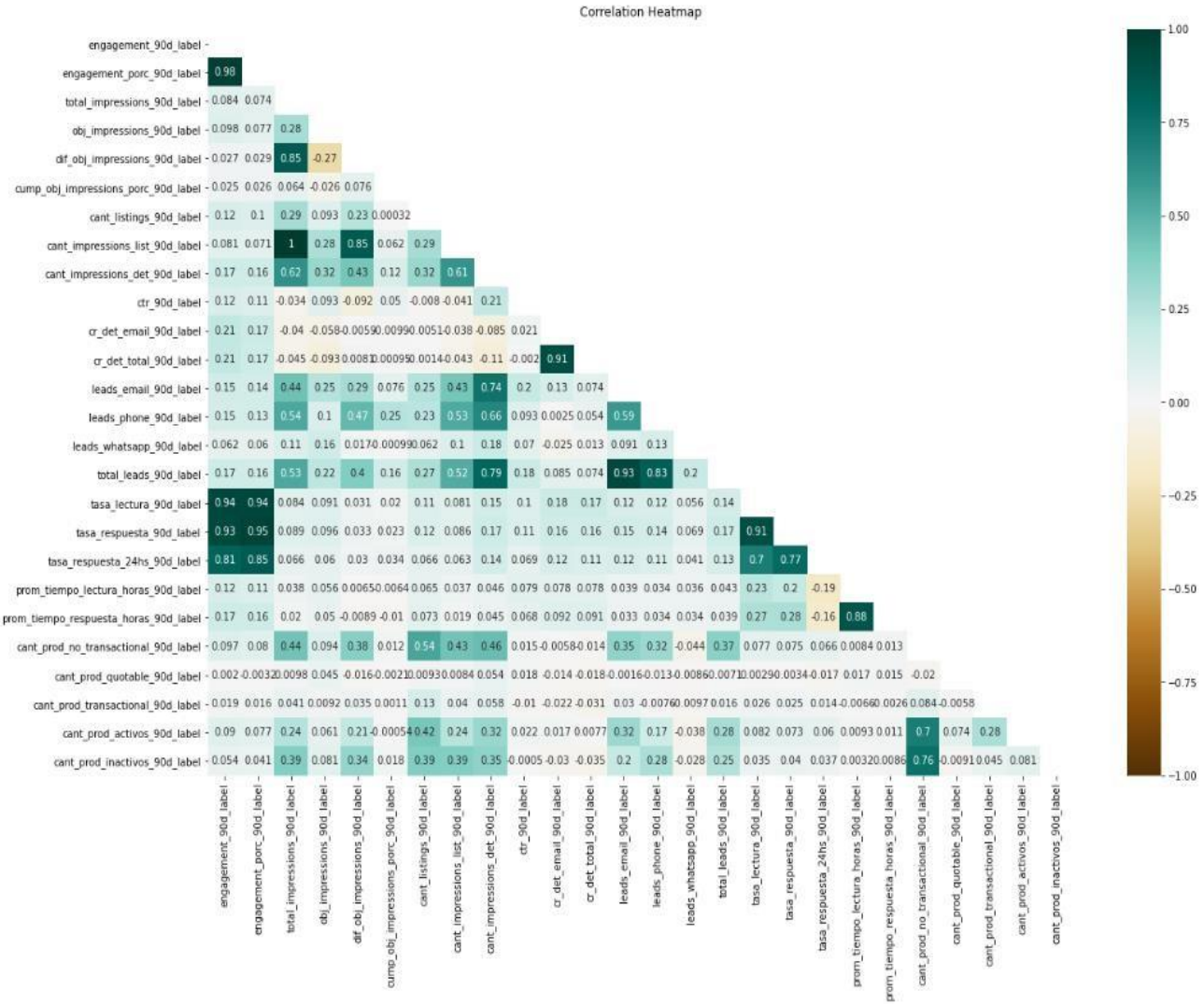

Figura 16: *Heatmap* que detalla la fuerza y el tipo de relación existente entre las variables continuas de tendencia para los 90 días.

Debido a que las correlaciones existentes entre las variables para los distintos *Heatmaps* son similares, el análisis en profundidad se realizará sobre el correspondiente a las variables de tendencia de los 30 días.

Es importante tener en consideración la correlación positiva y fuerte existente (0.82) entre la cantidad de productos inactivos (***'cant_prod_inactivos_30d_label'***) y la cantidad de productos operados en el formato no transaccional (***'cant_prod_no_transaciontal_30d_label'***), es decir aquellos que no aceptan pedidos de cotización ni de compra. Desde el punto de vista del negocio, es un punto a destacar y tener en cuenta, ya que implica que este método no transaccional que se lleva a cabo por algunos clientes está derivando en una mayor cantidad de

productos inactivos en la plataforma *online*. Esto llevaría a pensar que sería mejor enfocar los esfuerzos de los ejecutivos comerciales en intentar comercializar los otros dos métodos disponibles: '*quoatable'* y 'transaccional'.

Por lo tanto, para indagar sobre esta correlación, es interesante mostrar la relación entre estas variables y su comportamiento con el *label*. Se puede observar en la siguiente figura:

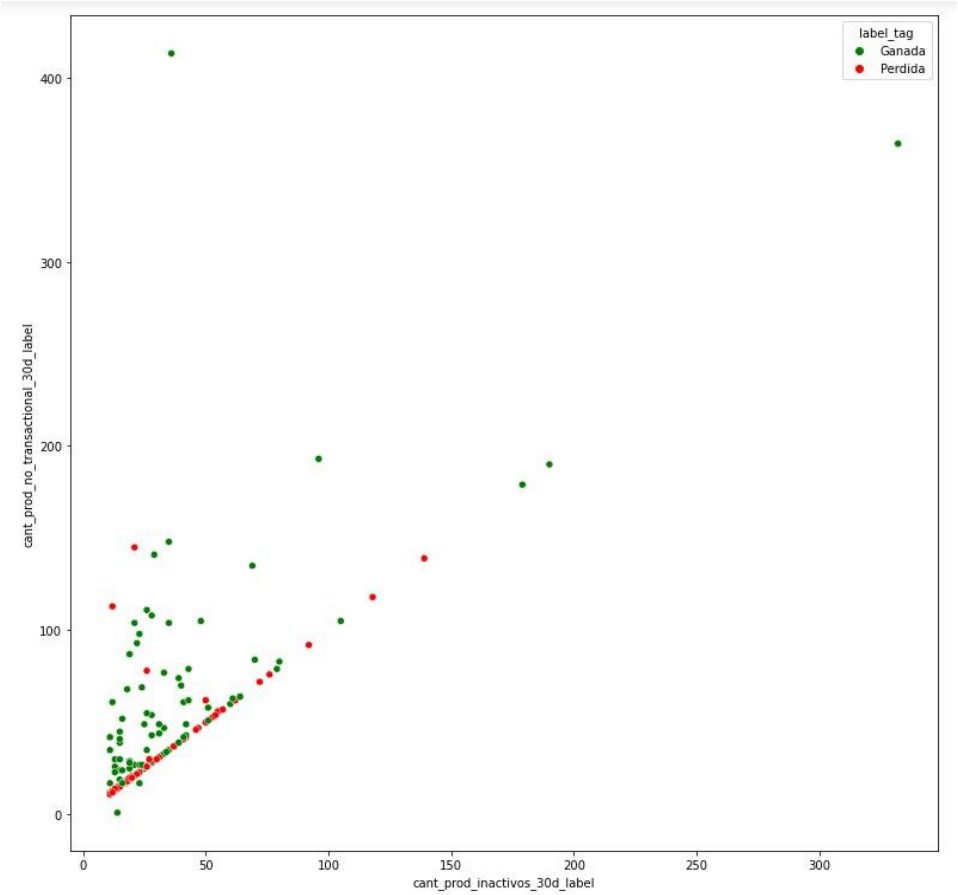

Figura 17: *Scatterplot* mostrando la relación entre las variables *'cant_prod_inactivos_30d_label'* y *'cant_prod_no_transaciontal_30d_label'* discriminado según el *label*.

Queda expuesta no solo la fuerte correlación positiva entre ambas variables, sino también se logra distinguir una correlación más marcada con las negociaciones perdidas que con las ganadas. Para ello, se decidió abrir el gráfico en dos, para verificar este detalle encontrado:

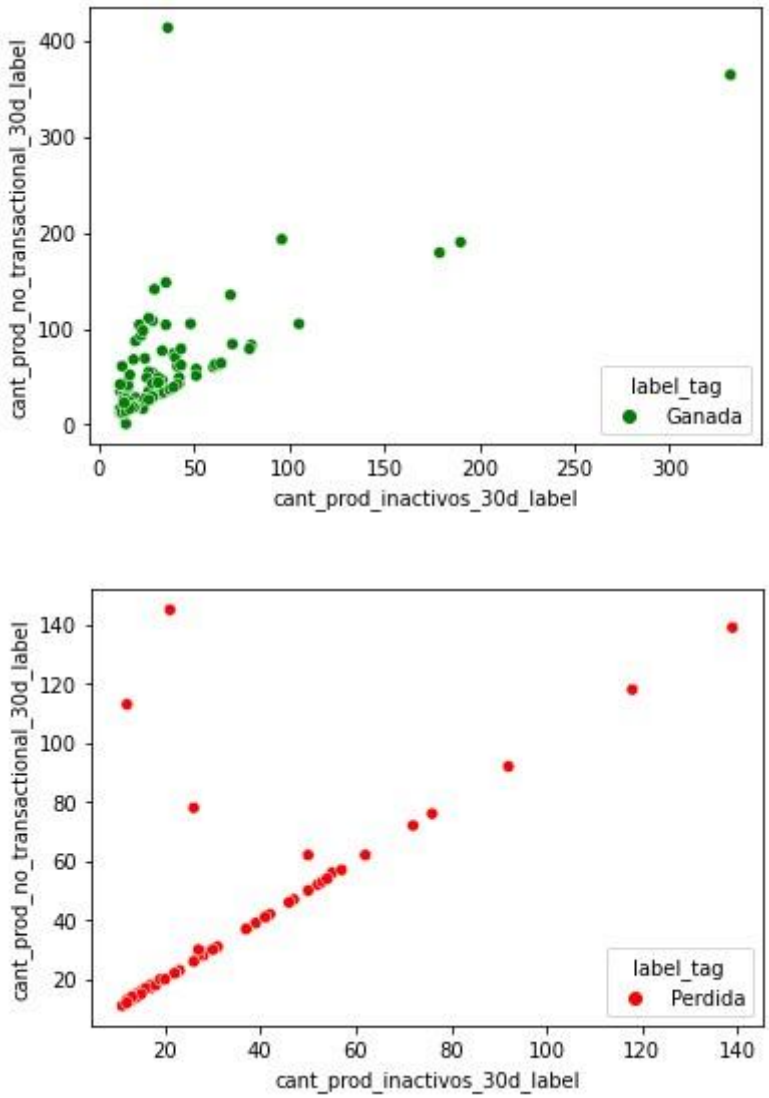

Figura 18: *Scatterplots* mostrando la relación entre las variables *'cant_prod_inactivos_30d_label'* y *'cant_prod_no_transaciontal_30d_label'* para negociaciones ganadas y perdidas respectivamente.

Así también existen otras fuertes correlaciones positivas entre distintas variables que es importante destacar. Por ejemplo, ***'engagement_30d_label'*** con la ***'tasa_respuesta_30d_label'*** (0.94), la ***'tasa_respuesta_24hs_30d_label'*** (0.81) y la ***'tasa_lectura_30d_label'*** (0.95),. Lo mismo sucede entre *'engagement_porc_30d_label'* con las variables anteriormente mencionadas, lo cuál tiene sentido ya que son dos maneras distintas de expresar el *engagement* de los clientes, tomando los valores de (0.96), (0.85) y (0.95) respectivamente. Esto demuestra que cuánto mayor es el nivel de compromiso, interacción, confianza que

tienen los clientes con la empresa, mayor va a ser la cantidad de mensajes de los potenciales compradores que se dediquen a leer y responder.

En relación a estas correlaciones, se encuentra la existente entre las variables ***'prom_tiempo_lectura_horas_30d_label'*** y ***'prom_tiempo_respuesta_horas_30d_label'*** (0.82), que simplemente hace referencia a el mismo indicador de rendimiento, nada más que es un promedio de horas en que tardan los vendedores en leer y responder los mensajes de los compradores.

Por otro lado, también se puede hacer mención a la correlación existente entre ***'cant_impressions_det_30d_label'*** con ***'total_leads_30d_label'*** (0.92), ***'leads_phone_30d_label'*** (0.87) y ***'leads_email_30d_label'*** (0.88). Esto también tiene lógica ya que a una mayor frecuencia de exposición de los anuncios de los productos, se espera que exista una mayor cantidad de clientes potenciales o prospectos.

A su vez, era de esperar que exista una correlación positiva entre ***'total_leads_30d_label'*** con ***'leads_email_30d_label'*** (0.97) y ***'leads_phone_30d_label'*** (0.95). La cantidad de leads totales están conformados por los leads generados a través de los llamados, mails y mensajes enviados al *Whatsapp*.

Por último, hay que tener en cuenta la correlación entre ***'cant_impressions_list_30d_label'*** y ***'total_impressions_30d_label'*** la cuál es igual a (1). Las impresiones de *listing*, a diferencia de las llamadas 'det' (detalle) son aquellas que se refieren al listado propiamente dicho de las categorías disponibles en la plataforma *online* en las que se encuentran categorizados los vendedores. Entonces, esta correlación de 1 demuestra que cuándo aumentan la cantidad de impresiones del tipo *listing*, aumentan en la misma proporción el total de impresiones. A diferencia de lo que ocurre con la correlación existente entre ***'cant_impressions_det_30d_label'*** y ***'total_impressions_30d_label'***, que toma un valor de (0.67), lo cuál implica que el total de impresiones no aumenta en la misma proporción cuándo aumentan las impresiones de tipo detalle. Estas impresiones 'det' hacen referencia a la visualización de los anuncios propiamente dichos dentro de la sección del cliente.

Para finalizar con el análisis multivariado, se buscará entender el comportamiento de las variables numéricas respecto al *label*, comenzando con la variable ***'opportunity_duration_t1'***.

Tal como se había explicado en la sección 2.2.3 (Análisis univariante), la distribución de los datos para dicha variable mostraba como las negociaciones, en su mayoría, tendían a cerrarse rápidamente luego de ser abiertas. Por lo cuál, resultaba interesante indagar un poco más en este comportamiento, y terminar de entender cómo es la distribución de las negociaciones ganadas y perdidas.

En la siguiente figura se puede observar cómo a medida que avanza el tiempo a los ejecutivos comerciales resulta en una mayor complejidad ganar las negociaciones, es decir cuándo la madurez se hace más larga, la distribución se acerca más a las de las negociaciones perdidas. A diferencia de lo que sucede en momentos más cercanos a cuándo se abre una negociación (madurez corta), donde se puede ver una separación más marcada entre la distribución de las perdidas y las ganadas, existiendo mayores chances de finalizar la negociación de una manera exitosa.

A su vez, hay muchas negociaciones que se cierran el mismo día en que se abren, y esto se puede dar cuándo los ejecutivos comienzan una negociación con un cliente que ya saben que van a cerrar positivamente, estableciéndose finalmente un contrato de membresía.

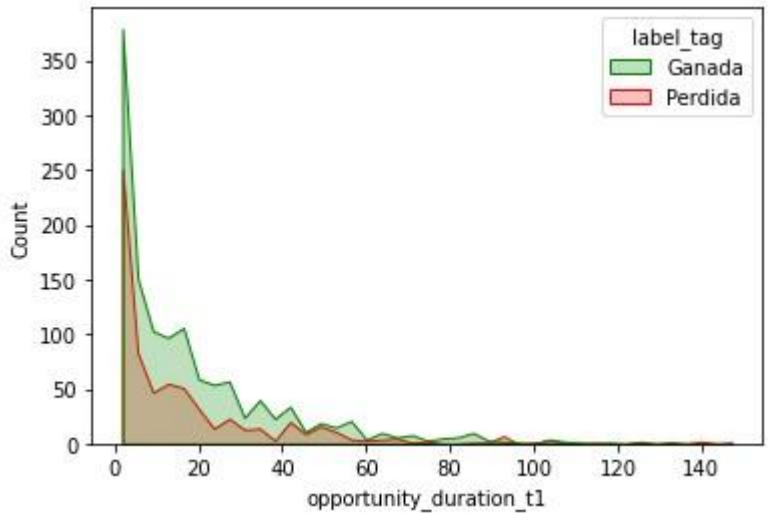

Figura 19: Distribución de negociaciones ganadas y perdidas (*label*) en función de la madurez de las oportunidades (*'opportunity_duration_t1'*).

A continuación, se prosigue con un análisis respecto a cómo se distribuyen las negociaciones ganadas y perdidas a lo largo de los meses del año en función de la fecha de creación de la primera oportunidad ganada (***'fecha_opt_actual'***).

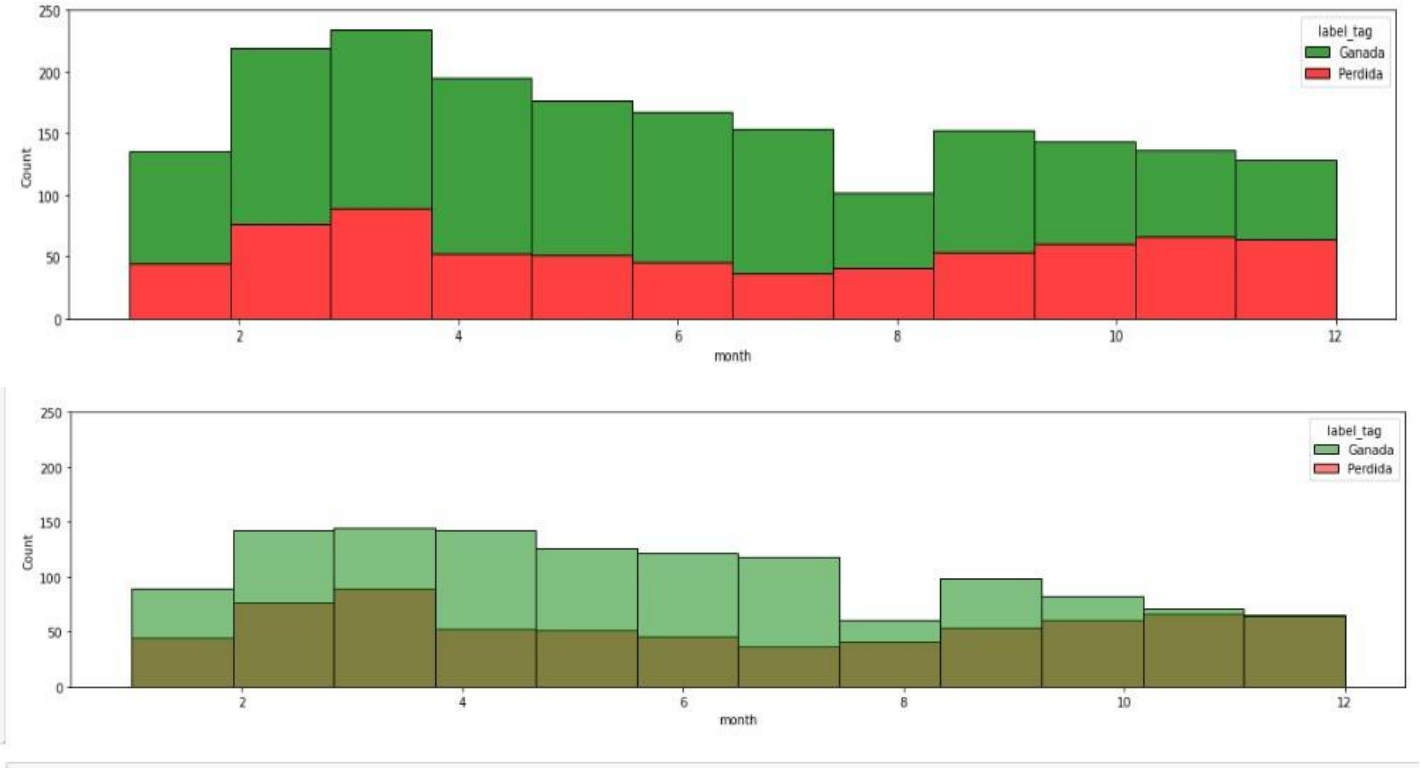

Figura 20: Distribución de negociaciones ganadas y perdidas (*label*) en función de la *'fecha_opt_actual'*.

Se puede notar como en el primer semestre del año ocurren una mayor cantidad de negociaciones independientemente si estás resultan ganadas o perdidas. A su vez, resulta ser el período dónde existe una mayor propensión a ganar las negociaciones. Se puede ver cómo en el mes de abril aproximadamente, de las 200 negociaciones que se abrieron, 50 se perdieron y 150 se ganaron, mientras que en diciembre hubo alrededor de un 50% y 50% respectivamente. A medida que van avanzando los meses la cantidad de negociaciones pérdidas van aumentando en proporción al total ocurrido en cada mes en particular.

Esto puede deberse a que en los primeros meses del año se realizan lo que se denomina internamente en la empresa, las 'ferias', dónde los ejecutivos comerciales se encuentran en diversos eventos con los clientes resultando facilitando el relacionamiento entre ellos. Otro de los motivos, es que a principio de año es donde

se tiende a planificar las estrategias de venta por parte de las distintas empresas (clientes).

En el análisis univariante realizado anteriormente se había considerado el comportamiento en la distribución de los datos para la variable ***'total_price_t1'***. Resumidamente se había comentado que existían negociaciones con precios negativos que hacían referencia a las notas de crédito generadas a los clientes cuando estos se dan de baja, como así también una concentración de las negociaciones principalmente en torno a los precios de 80.000 y 30.000 referentes a membresías del tipo *Business*.

Frente a estos *insights* encontrados, se decidió indagar más para entender en qué situaciones las probabilidades de ganar una siguiente negociación podrían ser mayores, y además analizar los valores negativos. Por lo tanto, se realizó un gráfico del tipo *boxplot* para mostrar el resultado obtenido en la segunda negociación, en base a cuál había sido el precio de la membresía contratada en la primera negociación ganada.

Puede observarse en la figura 21 que por alguna razón los productos con precios entre los 30.000 – 80.000 tienen más probabilidades de ser adquiridos. Uno de los motivos, como se explicó en líneas anteriores, es la mayor participación de clientes del tipo *Business*, quienes son consideradas empresas más pequeñas, que no tienen suficiente *budget* para destinar a la contratación de membresías de una jerarquía mayor.

A su vez, es interesante notar que cuándo se negocia la contratación de productos con valores inferiores a 30.000, estos tienden a rechazarse. Una de las razones podría ser que no existe mucha diferencia en el precio entre una membresía de 10.000 – 20.000 contra una de 30.000 – 40.000, pero si la hay en las características y particularidades ofrecidas en los paquetes, lo que termina siendo atractivo por la relación costo/beneficio.

Por último, en cuánto a los precios negativos, puede visualizarse que se dan en una mayor cantidad para las negociaciones pérdidas, lo cuál es correcto ya que implica que el cliente antes de finalizar su contrato, decidió darse de baja solicitando la devolución del dinero restante. Aún así, existen negociaciones con precios negativos pero ganadas. Esto resulta extraño, pero a nivel de negocio implica que el ejecutivo,

en el momento en que el cliente solicitó la baja, hizo un esfuerzo para evitarlo luego de emitirse la nota de crédito, y logró finalmente retenerlo. Entonces, si bien es una nota de crédito, está queda marcada como ganada ya que el cliente optó por continuar con su contrato.

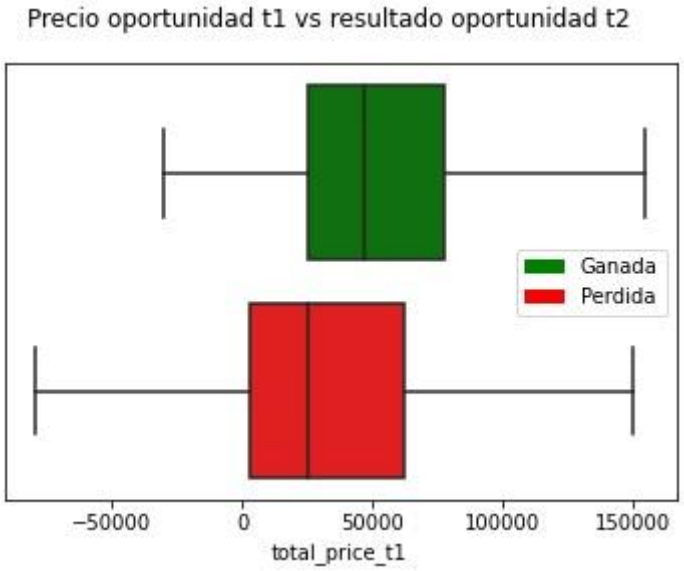

Figura 21: *Boxplot* mostrando la relación entre el precio de la primera oportunidad ganada (*'total_price_t1'*) y el resultado obtenido en la siguiente oportunidad negociada (*label*).

Al igual que para *'total_price_t1'*, se buscó explayarse en el análisis de la variable ***'engagement_30d_label'***. En la sección 2.2.3 (Análisis univarante) se mostró que las clasificaciones tienden a ser principalmente positivas, tomando valores de 4 ó 5, ó negativas, con valores de 1 ó 2.

Así en las siguientes figuras se trata de entender el resultado obtenido en la renegociación (oportunidad t2) en función de la clasificación del cliente durante la primera negociación ganada.

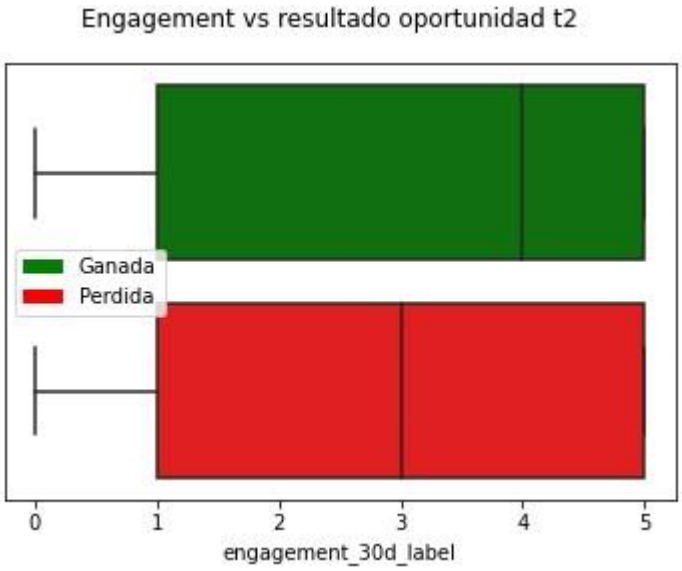

Figura 22: *Boxplot* mostrando la relación entre el *engagement* establecido en la primera oportunidad ganada (*'engagement_30d_label'*) y el resultado obtenido en la siguiente oportunidad negociada (*label*).

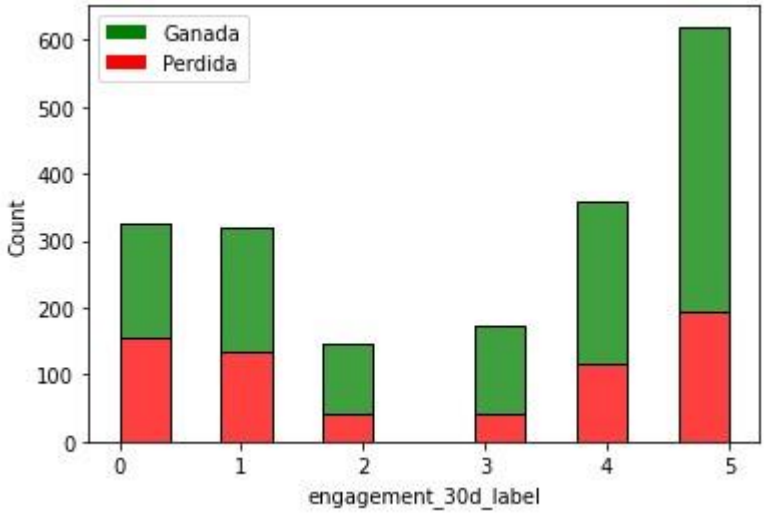

Figura 23: Histograma para la variable *'engagement_30d_label'* y su relación con el *label*.

De esta manera, puede verse que la mayor cantidad de negociaciones ganadas se obtienen cuándo el *engagement* del cliente ha sido positivo (Excelente y Muy Bueno), lo cuál tiene sentido. Si el cliente siente compromiso, confianza y una buena relación con la empresa, mayores van a ser las chances de que opte por re contratar el servicio de membresías. En la figura 22 se puede observar que la mediana se encuentra en el valor de 4 para las negociaciones ganadas, demostrando lo anteriormente mencionado. Mientras que, para las perdidas, esta

medida estadística se encuentra en el valor de 3, lo cuál implica un *engagement* 'Regular'.

2.3. Ingeniería de Atributos

2.3.1. Transformación de datos e ingeniería de atributos

La ingeniería de atributos cobra una gran importancia en cualquier modelo de aprendizaje automatizado, ya que ayuda a mejorar su rendimiento, proporcionando de esta manera mejores predicciones sobre la variable respuesta. Resumidamente, consiste en crear nuevas variables a partir de las ya existentes, permitiendo que los modelos extraigan información importante de los datos. Por lo tanto, una gran parte del esfuerzo en la implementación de los algoritmos de *Machine Learning* relacionados al aprendizaje automático, se dedica en el preprocesamiento y transformación de los datos (Bengio et al, 2013).

De acuerdo como se ha desarrollado a los largo de la sección 2.1.1 (Descripción y estructura de los datos), fue necesario llevar a cabo una parte del proceso de ingeniería de atributos con anterioridad al análisis exploratorio de los datos, para poder construir el *dataset* final que se utilizará en los modelos que realizarán las predicciones.

De esta manera, en la presente tesis se han aplicado diferentes técnicas con el objetivo de aprovechar al máximo los datos disponibles. A continuación se mencioan:

One Hot Encoding

Muchos algoritmos de aprendizaje automático no pueden operar directamente con variables categóricas, lo que significa que los datos categóricos deben convertirse a una forma numérica. Por lo tanto, la técnica de *One Hot Encoding* se utiliza para transformar todas las variables categóricas a un formato numérico que sea aceptado por los algoritmos de *Machine Learning*, mejorando la precisión de la predicción.

Cuándo las variables categóricas presentan muchas categorías distintas, el tratamiento con *One Hot Encoding* puede resultar en una gran cantidad de columnas nuevas, lo que llevará a un problema en el momento de entrenar los modelos, enlenteciendo y bajando la *performance* (Gerón Aurélien, 2019). Pero en este caso, debido a que las variables categóricas sobre las cuáles se llevó a cabo la transformación, presentaban una baja cantidad de categorías, fue factible utilizar dicha técnica.

A continuación se mencionan las variables que fueron tratadas con esta técnica durante el proceso de armado del *dataset* principal*: 'task subtype', 'model type', 'product status'* y *'transactions type'* . En la sección 2.2.1 (Descripción y estructura de los datos) se explicó en detalle cómo se realizó y el motivo. A su vez, en la misma sección se detalló la transformación que se realizó para las variables *'leído'* y *'respondido'* de la base de datos de mensajes.

Luego de realizar el análisis exploratorio de los datos (EDA) y antes de proceder a poner a prueba los modelos, se procedió a aplicar dicha técnica en las restantes variables categóricas que persistían en el modelo.

De esta manera, las siguientes variables categóricas fueron transformadas en numéricas: *'country', 'rubro_t1', 'currency_t1', 'product_name_t1', 'tipo_empresa', 'region', 'membership_30d_label', 'membership_60d_label', 'membership_90d_label'* y *'categoria_t1'*. Recordar que para estas variables en la sección 2.2.2 (Calidad de los datos), se explicó cómo se agruparon en función de las categorías que las conformaban, a fin de disminuir la cantidad categorías y poder así aplicar *One Hot Encoding* sin generar una gran cantidad de columnas.

Valores nulos

En la sección 2.2.2 (Calidad de los datos) se dejó explícito para cada una de las variables que presentaban valores nulos, cuál fue la decisión que se tomó para tratarlos. Los criterios tomados estuvieron basados según el contexto del negocio y sus implicancias.

Creación de variables de tendencia

Tal cómo se desarrolló en la sección 2.1.1 (Descripción y estructura de los datos), en dos bases de datos particulares, *performance* y productos, fue necesario realizar una arregló de las variables allí presentes ya que sus registros no representaban información específicamente para cada día, sino para un mes puntual. En dicha sección se explicó en detalle cómo se realizó el procedimiento de armado de estas variables.

3. Metodología

3.1. Métodos

El método a llevar a cabo está basado en un modelo de *Machine Learning* que intenta resolver un problema de clasificación binaria, infiriendo la probabilidad de que un cliente renueve o no su membresía en base a su experiencia de *performance* reciente, lo que permitirá entender cuando hay una mayor propensión a ganar las renegociaciones. Probando distintos algoritmos y técnicas de ingeniería de atributos se trató de encontrar el modelo que mejor se ajuste a los datos y tenga una mayor efectividad a la hora de realizar las predicciones.

Se espera que con esta información los ejecutivos puedan discernir entre vendedores que renovarán sin necesidad de negociar y potenciales vendedores que no renovarán sus membresías, para así poder priorizar las negociaciones y dirigir sus esfuerzos. Además, contar con la probabilidad de que los vendedores renueven, les permitirá hacer un mejor uso de los descuentos ofrecidos. De esta manera, cuando un vendedor con alta probabilidad de renovación entre en el proceso de negociación, el ejecutivo podrá monetizar esta información ofreciendo descuentos acordes y óptimos para el caso.

3.2. Modelos de Machine Learning

Se probarán distintos modelos de aprendizaje supervisado con el fin de encontrar cuál es el que presenta una mejor *performance* en los resultados, dando un mayor poder predictivo, ayudando a resolver el problema de negocio planteado.

Acorde a James, et al. (2013), un modelo de aprendizaje supervisado consiste en tener una variable respuesta yi para cada observación de la variables predictoras xi , con $i = 1, \dots, n$, buscando ajustar un modelo que relacione la variable respuesta con las predictoras, con el objetivo de predecir con precisión la respuesta para futuros observaciones (predicción).

Existen dos tipos de problemas supervisados, por un lado los llamados problemas de clasificación y por el otro los de regresión. En los primeros, la variable respuesta es categórica indicando a qué categoría pertenecen las observaciones del conjunto de datos. Mientras que en los segundos, la variable dependiente es continua y por lo tanto el modelo predecirá un valor numérico.

En la presente tesis, el problema de negocio se abordará como un problema de clasificación binaria, ya que la variable dependiente toma los valores de 0 ó 1.

Regresión Logística

El Modelo *Logit* es un tipo de regresión de clasificación binaria, que se usa comúnmente para estimar la probabilidad de que una instancia pertenezca a una clase en particular, es decir para predecir el resultado de una variable categórica a partir de las variables predictoras.

Si la probabilidad estimada es mayor al 50%, entonces el modelo predice que la instancia pertenece a esa clase (llamada clase positiva, etiquetado como "1"), y de lo contrario predice que pertenece a la clase negativa, etiquetada como "0". Esto lo convierte en un clasificador binario (Gerón Aurélien, 2019).

Según Gerón Aurélien, un modelo de regresión logística calcula una suma ponderada de los atributos de entrada (más un término de sesgo). En lugar de generar el resultado directamente como el modelo de regresión lineal, genera una función logística de este resultado.

$$\hat{p} = h_\theta(x) = \sigma(x^T\theta)$$

Ecuación 1: Modelo de regresión logística estimando la probabilidad (forma vectorizada)

La función logística, denominada σ(·), es una función sigmoidea que genera una número entre 0 y 1. Se define como se muestra en la Ecuación 2 y Figura 3.

$$\sigma(t) = \frac{1}{1 + \exp(-t)}$$

Ecuación 2: Función logística.

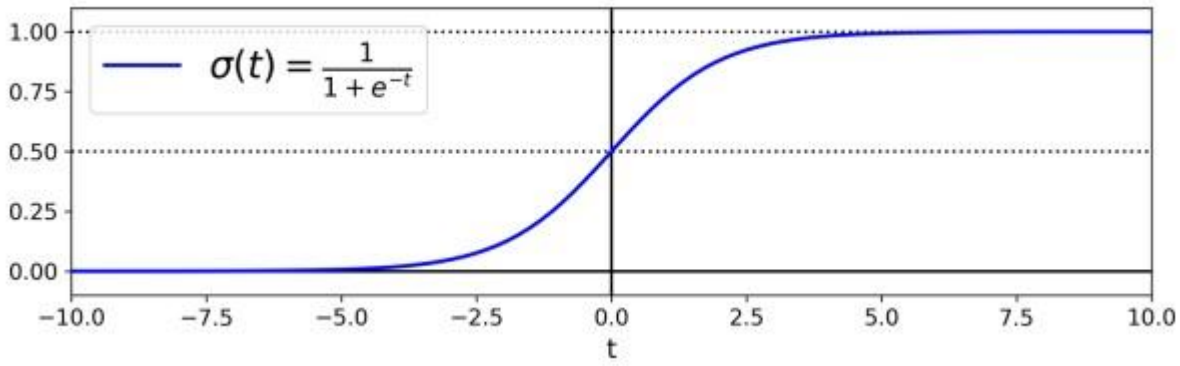

Figura 24: Función logística.

Una vez que el modelo de Regresión Logística ha estimado la probabilidad $\hat{p} = h_\theta(x)$ de que una instancia x pertenezca a la clase positiva, puede hacer su predicción $\hat{y}$ de la siguiente manera:

$$\hat{y} = \begin{cases} 0 & \text{if } \hat{p} < 0.5 \\ 1 & \text{if } \hat{p} \geq 0.5 \end{cases}$$

Ecuación 3: Regresión logística predicción.

Se puede observar que σ(t) < 0,5 cuando t < 0, y σ(t) ≥ 0,5 cuando t ≥ 0, por lo que un modelo de regresión logística predice 1 si x θ es positivo y 0 si es negativo.

Árboles de decisión

Los árboles de decisión son algoritmos versátiles de aprendizaje supervisado que pueden ayudar a resolver problemas tanto de clasificación cómo de regresión. Se los considera algoritmos poderosos, capaces de ajustar conjuntos de datos complejos

(Gerón Aurélien, 2019). Implementan una estrategia de *divide and conquer* en un formato no paramétrico, mediante la construcción de una estructura de datos jerárquica (Ke et al., 2017).

Acorde a James, et al. (2013), un árbol de clasificación es muy similar a un árbol de regresión, excepto que un árbol de clasificación se usa para predecir una respuesta cualitativa en lugar de una cuantitativa. Por un lado, la respuesta predicha para una observación en un árbol de regresión está conformada por la respuesta media de las observaciones de entrenamiento que pertenecen al mismo nodo terminal. Mientras que para un árbol de clasificación, se predice que cada observación pertenece a la clase más ocurrente de las observaciones de entrenamiento. Al interpretar los resultados de un árbol de clasificación nos interesa la predicción de clase correspondiente a una región de un nodo terminal en particular.

Entre las principales ventajas de trabajar con árboles de decisión, se encuentra la interpretabilidad y sencillez. Desafortunadamente, el algoritmo básico del árbol de decisiones, no es tan poderoso como otros enfoques de clasificación en términos de su precisión predictiva. Sin embargo, al agregar muchos árboles de decisión, usando métodos como *bagging* ó *boosting*, se puede lograr mejorar sustancialmente la *performance* (James, et al., 2013).

Random Forest y LightGBM

Random Forest y LightGBM son algoritmos de aprendizaje automático de *Machine Learning,* específicamente dentro de los métodos de ensamble conocidos como *bagging* (Random Forest) y *boosting* (LightGBM). Estos algoritmos utilizan árboles como bloques de construcción, para generar modelos predictivos más potentes, generalmente mejorando notablemente el nivel de precisión en las predicciones a expensas de alguna pérdida en la interpretación (James, et al., 2013).

Por un lado, ***bagging*** es una técnica de ensamblaje que busca reducir la varianza, y de esta manera aumentar la precisión de la predicción. Una forma natural de reducir la varianza es tomar muchos set de entrenamientos independientes, construir un

clasificador y luego promediar sus predicciones. Esto descrito anteriormente es exactamente lo que hace *bagging*.

El algoritmo de **Random Forest** utiliza el método *bagging* para entrenar, pero con una mejora sustancial. Introduce una aleatoriedad adicional al generar los árboles de decisión. Para realizar esto, en lugar de buscar el mejor atributo al dividir un nodo, busca la mejor característica entre un subconjunto aleatorio de atributos. Al construirse los árboles de decisión, cada vez que se considera una división en un árbol, se elige una selección aleatoria de *m* predictores del conjunto completo de *p* predictores, como candidatos para la división. La división puede usar solo uno de esos *m* predictores. De esta manera, el algoritmo da como resultado una mayor diversidad de árboles, descorrelacionados, intercambiando un sesgo más alto por una varianza más baja. Este enfoque ha demostrado ser particularmente eficaz en evitar el sobreajuste lo que generalmente produce un modelo con mayor poder de predicción (Gerón Aurélien, 2019).

Por último, ***boosting*** es otro método de ensamblado que ayuda a mejorar la *performance* de los árboles de decisión. Su manera de operar es distinta a *bagging*, ya que en lugar de entrenar árboles independientemente unos de otros, *boosting* propone un entrenamiento secuencial de los diferentes clasificadores: cada árbol se genera utilizando información de árboles generados previamente. La idea detrás de este procedimiento es evitar el ajuste de un solo gran árbol de decisión que probablemente genere *Overfitting* en los datos de entrenamiento (James, et al., 2013).

De esta manera, el algoritmo *boosting* aprende lentamente mediante la construcción de árboles relativamente pequeños cuyo objetivo es predecir los residuos del clasificador anterior en lugar del resultado *Y*. Es decir, dado el modelo actual, se ajusta un árbol de decisión a los residuos del modelo y se agrega este nuevo árbol a la función ajustada para actualizar los residuos. Cada uno de estos árboles puede ser bastante pequeño, con solo algunos nodos terminales, determinados por el parámetro *'d'* en el algoritmo. De esa manera, al ir ajustando pequeños árboles a los residuos, el algoritmo se enfoca en mejorar la función $\hat{f}$ en aquellas áreas donde no se está desempeñando bien (James, et al., 2013).

El parámetro *'shrinkage λ'* ralentiza aún más el proceso, permitiendo que más árboles de diferentes formas ataquen los residuos. En general, los enfoques de aprendizaje estadístico que aprenden lentamente tienden a desempeñarse bien (James, et al., 2013).

Existen varios métodos disponibles que utilizan la técnica *boosting*, pero los más comunes son *Adaptive Boosting* y *Gradient Boosting.*

LightGBM (*Light Gradient Boosting*) es un algoritmo de *Machine Learning* innovador que aplica el método *Gradient Boosting* a algoritmos basados en árboles, con una rápida velocidad de entrenamiento, alta *'performance'*, menor uso de memoria, mejor precisión y buena capacidad para manejar datos a gran escala.

Las técnicas utilizadas por el enfoque distintivo de LightGBM buscan reducir la cantidad de instancias de datos y la cantidad de atributos en búsqueda de acelerar el proceso de entrenamiento del tradicional *Gradient Boosting,* sin que el rendimiento se vea afectado. Con este objetivo, el algoritmo se basa en la aplicación de muestreo de *Gradient Based One-Side Sampling* (GOSS) y *Exclusive Feature Bundling* (EFB). La primera técnica (GOOS) excluye una proporción significativa de instancias de datos con gradientes pequeños y solo utiliza el resto para estimar la ganancia de información de las posibles divisiones, obteniendo así estimaciones bastante precisas con un tamaño de datos mucho menor. El segundo (EFB), agrupa características mutuamente excluyentes utilizando un enfoque codicioso pero efectivo y, por lo tanto, reduce la dimensión del espacio de datos. Esto resulta especialmente útil para conjuntos de datos con características categóricas de alta cardinalidad, típicamente abordadas con *One Hot Encoding*, generando enormes matrices dispersas.

Aplicando estas variantes, LightGBM, a diferencia del *Gradient Boosting* tradicional, ha demostrado acelerar el proceso de entrenamiento más de veinte veces y lograr casi la misma precisión (Ke et al, 2017, p. 3146).

3.3. Métrica de evaluación de modelos

Para evaluar los modelos y poder compararlos se utilizará el área bajo la curva ROC (AUC-ROC), métrica utilizada con clasificadores binarios, que mide el grado de separación de las clases, especialmente útil cuando se trata de problemas con clases muy desbalanceadas (Tan, Steinbac, Kumar, 2006).

La curva ROC se obtiene trazando las diferentes combinaciones de la tasa de verdaderos positivos (*TPR*) en el eje '*y*', y la tasa de falsos positivos (*FPR*) en el eje '*x*', para todos los posibles puntos de corte (*threshold*).

El *FPR* es la proporción de instancias negativas que son incorrectamente clasificadas como positivas. Es igual a 1 - la tasa de verdaderos negativos (*TNR*), que es la proporción de instancias negativas que se clasifican correctamente como negativas, a la cuál también se la conoce con el nombre de especificidad (*specificity)*. Por lo tanto, la curva ROC traza la sensibilidad (*recall*) versus 1 - especificidad (Gerón Aurélien, 2019).

De esta manera, la curva se obtiene iterando sobre el espacio del punto de corte [0;1] y trazando sus resultados. Cuanto más cerca esté la curva de la diagonal, menor será el grado de separabilidad de las clases que proporciona el clasificador. A diferencia de una curva ROC ideal, que sería aquella que esté lo más cercana posible a la esquina superior izquierda (lo que indicaría un alto valor de *TPR* y un bajo valor de *FPR*). Por lo tanto, cuanto mayor sea el área debajo de la curva (AUC-ROC), mejor será el clasificador. El área debajo de la curva determina la *performance* general del clasificador a lo largo de todos los posibles puntos de corte (James, et al., 2013).

Esta métrica toma valores en un rango de [0;1]. Un valor igual a 0,5 es sinónimo de predecir de forma azarosa, igual a 1 representa una separación perfecta, y menor a 0,5 sugiere una *performance* peor que el azar.

En la siguiente figura 4, se puede observar un ejemplo de la curva ROC:

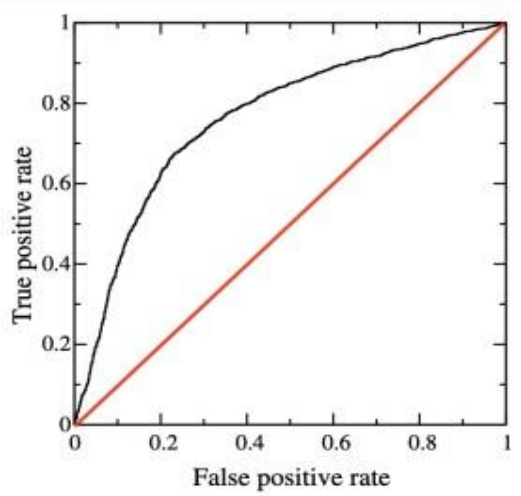

Figura 25: Ejemplo curva ROC.

3.4. Optimización de hiperparámetros

La optimización de los hiperparámetros de los algoritmos de aprendizaje automático es un paso muy importante a tener en consideración, ya que controlan directamente el comportamiento de los algoritmos de entrenamiento y tienen un efecto significativo en el rendimiento de los modelos (Wu et al, 2019, pág. 26).

El objetivo de la optimización de hiperparámetros consiste en encontrar un algoritmo de aprendizaje automático determinado que arroje la mejor *performance*, medida en un conjunto de validación (Koehrsen, 2018).

En el presente trabajo, la estrategia utilizada para encontrar los mejores hiperparámetros fue *Random search*. Alternativamente, se podrían buscar a través de un *Grid Search* que prueba todas las combinaciones posibles, pero esto es muy costoso computacionalmente y no trae mayores beneficios.

Random search es una forma mucho más eficiente y de igual (y a veces mayor) eficacia para encontrar los mejores valores. La metodología consiste en definir un rango de posibles valores para cada hiperparámetro y luego, aleatoriamente, seleccionar uno para cada uno. Es decir, en lugar de explorar cada una de las dimensiones de forma individual, las búsquedas aleatorias solo prueban una cierta cantidad de combinaciones que se seleccionan al azar, explorando el espacio con una mayor certeza**.** Es decir, se eligen aleatoriamente un set de coeficientes para cada

uno de los hiperparámetros, y luego se calcula la *performance*. Repitiendo esto varias veces, quedan armados distintos modelos a evaluar (Bengio, Bergstra, 2012).

A continuación, en la figura 26 se puede observar la disposición de ambos métodos de optimización de hiperparámetros mencionados anteriormente, *Grid Search* y *Random search.*

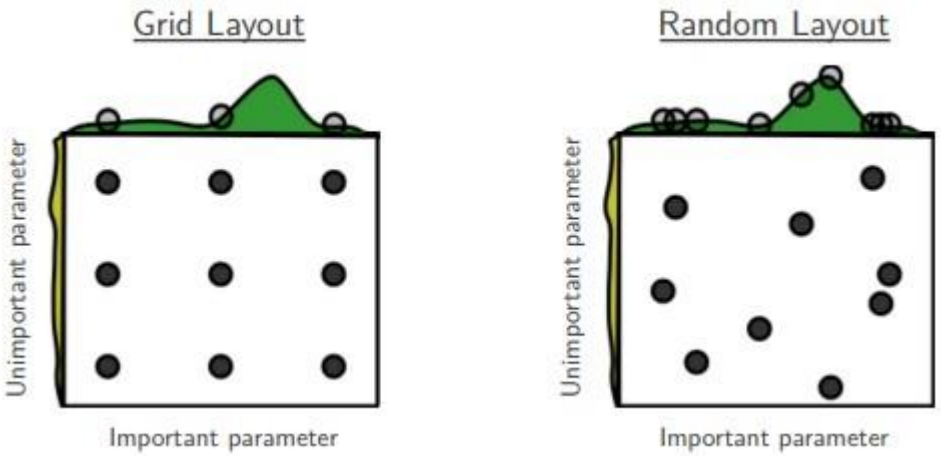

Figura 26: *Grid Search* y *Random search* de nueve pruebas para optimizar una función f(x, y) = g(x) + h(y) ≈ g(x) con baja dimensionalidad efectiva. Encima de cada cuadrado g(x) se muestra en verde, y a la izquierda de cada cuadrado h(y) se muestra en amarillo. Con *Grid Search* nueve intentos sólo prueban g(x) en tres lugares distintos. Con *Random search* los nueve ensayos exploran distintos valores de g. Este fracaso de la búsqueda en *Grid Search* es la regla y no la excepción en los casos de optimización de hiperparámetros de alta dimensión (Bengio, Bergstra, 2012).

En la presente tesis, *Random search* con *Cross-Validation* se aplicó en la búsqueda de optimización de los hiperparámetros para los algoritmos de Random Forest y LGBM.

El objetivo detrás de las técnicas de evaluación de modelos con validación cruzada (*Cross-Validation*) es contar con un conjunto de datos para la validación del modelo entrenado, en búsqueda de tener una estimación de cómo impactarán las decisiones de modelado en la *performance* del modelo con datos desconocidos. Entonces, la idea es simular la partición de los datos, en conocidos, que van a ser los que se van a utilizar para entrenar el modelo, y datos desconocidos que van a funcionar para validar las decisiones de modelado elegidas. Por lo tanto, la validación cruzada es el

proceso de medir la capacidad de generalización de diferentes modelos probándose con nuevos datos, no vistos durante el entrenamiento, y luego eligiendo el más preciso en este conjunto de datos. Hay diferentes técnicas disponibles para realizar una validación cruzada. En el proyecto actual se utilizó la validación cruzada *k-folds*.

Este enfoque implica dividir aleatoriamente el conjunto de observaciones en *k-folds* del mismo tamaño. La primera fold es tratada como el conjunto de validación, y el modelo es entrenado en el resto de las *k-1 folds*. Al final del proceso, hay *k* estimaciones diferentes de la *performance* del modelo, promediandose para estimar el rendimiento de la validación cruzada *k-fold*.

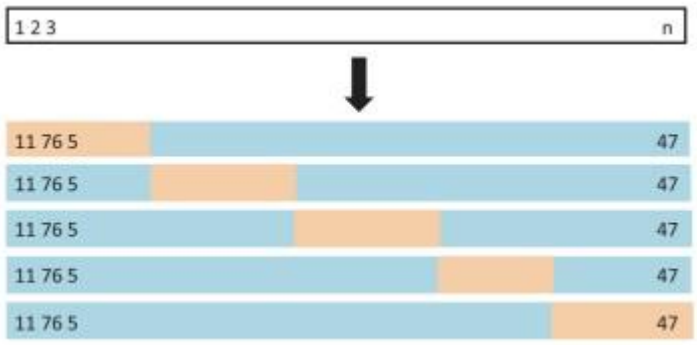

Figura 27: *K-fold Cross-Validation.*

Se utiliza este enfoque *k-folds* con el fin de comparar la *performance* entre diferentes combinaciones de hiperparámetros para el mismo algoritmo de aprendizaje automático evitando el sobreajuste en los datos de validación (*Overfitting*). Cuando se optimizan los hiperparámetros, se debe ajustar una cantidad significativa de modelos, y si todos ellos están entrenados y validados con la misma división estática de *'train-validation'* podría llevar a generar este sobre ajuste.

Se llevó a cabo la validación cruzada (*K-fold Cross-Validation)* para el algoritmo de Random Forest en primer lugar, y luego para el de LightGBM. El parámetro '*cv*', que determina la estrategia de división de validación cruzada se especificó igual a 3 para Random Forest y en 5 para LightGBM. El '*n_iter*', que representa el número de ajustes de parámetros que se muestran, se estableció en 200 para Random Forest y en 2000 para LightGBM.

Los rangos seleccionados para cada hiperparámetro se especifican a continuación:

Hiperparámetro	Rango
n_estimadores	[1000]
max_depth	[4, 5, 7]
criterion	['gini', 'entropy', 'log_loss']
min_samples_split	[2, 5, 10]
min_samples_leaf	[1,3,5]
max_features	['log2', 10, 20, 30]
max_leaf_nodes	[None]
boostrap	[True,False]

Tabla 5: Rango de cada hiperparámetro para la *K-fold Cross-Validation* de Random Forest.

Hiperparámetro	Rango
num_leaves	sp_randint(6, 50)
min_child_samples	sp_randint(100, 500)
min_child_weight	[1e-5, 1e-3, 1e-2, 1e-1, 1, 1e1, 1e2, 1e3, 1e4]
subsample	sp_uniform(loc=0.2, scale=0.8)
colsample_bytree	sp_uniform(loc=0.4, scale=0.6)
reg_alpha	[0, 1e-1, 1, 2, 5, 7, 10, 50, 100]
reg_lambda	[0, 1e-1, 1, 5, 10, 20, 50, 100]

Tabla 6: Rango de cada hiperparámetro para la *K-fold Cross-Validation* de LightGBM.

4. Resultados

En dicha sección se expondrán los resultados de *performance* obtenidos para cada uno de los modelos explicados anteriormente en la sección 3.2 (Modelos de *Machine Learning*):

1. Regresión Logística
2. Árboles de decisión
3. Random Forest
4. LightGBM

4.1. Performance

Antes de continuar, es importante aclarar que en una primera instancia se ejecutaron los modelos utilizando los parámetros por *default* y se utilizó *randomstate* = 42 para poder replicar los experimentos. Luego se llevó a cabo la optimización de hiperparámetros para aquellos que se consideró correcto y necesario.

Comenzando con el modelo de **Regresión Logística** se puede observar en la siguiente figura 28 la *performance*, obteniendo una AUC-ROC de 0.566 en el *dataset* de entrenamiento (*train*) y un 0.66 en el de validación. Por lo tanto, no solo este valor de 0.566 determina una AUC-ROC casi aleatoria, sino que también la AUC-ROC en validación es mayor a la de entrenamiento, lo que demuestra que la *performance* no es mejor que la 'suerte'.

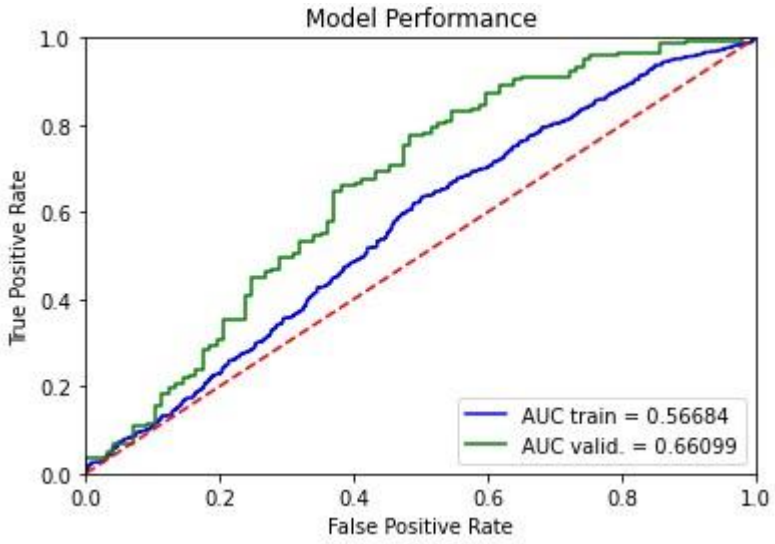

Figura 28: AUC-ROC, Regresión Logística en entrenamiento y validación.

En una segunda instancia, se continuó con el modelo de **Árboles de Decisión**. Si bien, se puede observar en la figura 29 que aumenta la AUC-ROC en el *dataset* de validación a 0.692 respecto a la Regresión Logística, la *performance* en entrenamiento da un valor muy alto de 0.999, lo que implica un sobre ajuste en los datos (*Overfitting*).

Esto quiere decir que se están ajustando patrones/particularidades en los datos de entrenamiento, que probablemente no se repiten en los datos de validación ("datos desconocidos"). De esta forma, este comportamiento deriva en una mala *performance* en validación (no generaliza muy bien con instancias nuevas) y una muy buena en entrenamiento, tal cómo sucede con este modelo.

Por lo tanto, es una situación indeseable porque el ajuste obtenido no producirá estimaciones precisas de la respuesta en nuevas observaciones que no formaban parte del conjunto de datos de entrenamiento original. A su vez, se puede decir que la varianza es alta y el sesgo (*bias*) es bajo. Un modelo con alta varianza presta mucha atención a los datos de entrenamiento y no generaliza sobre los datos que no ha visto antes. Cualquier alteración pequeña en los datos genera un gran cambio en las estimaciones (James, et al., 2013).

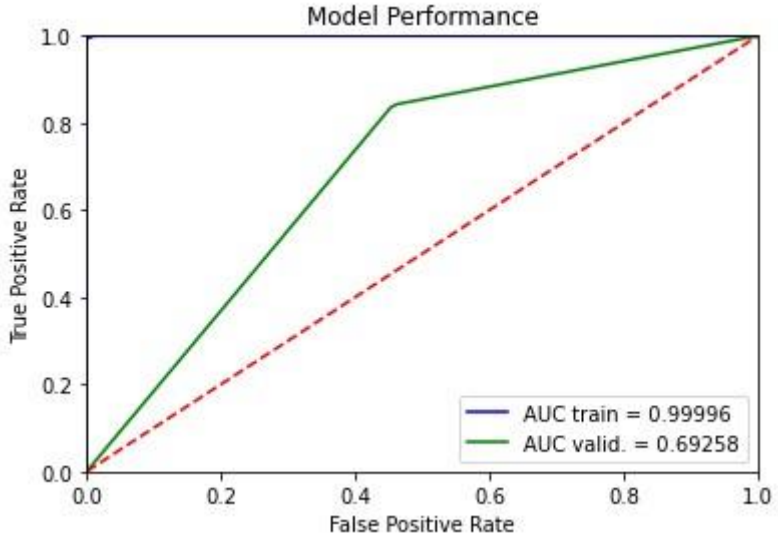

Figura 29: AUC-ROC, Árboles de Decisión en entrenamiento y validación.

En tercer lugar, se probó un modelo de **Random Forest**. Al igual que sucede con los Árboles de Decisión, se genera O*verfitting* en los datos, con un valor de AUC-ROC en entrenamiento de 0.999. Sin embargo, la gran diferencia es que incrementa considerablemente la AUC-ROC en validación a 0.857, por lo que se lo podría considerar un modelo bastante prometedor.

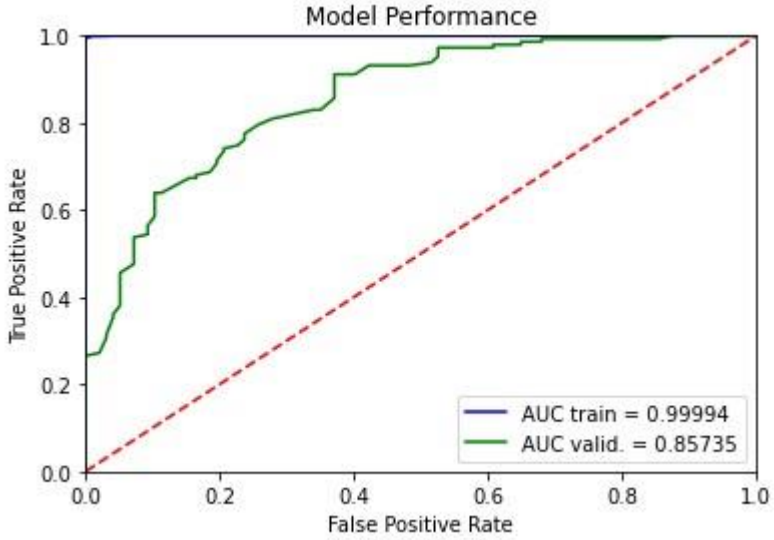

Figura 30: AUC-ROC, Random Forest en entrenamiento y validación.

Como cuarto modelo, se ejecutó LightGBM, obteniendo una *performance* en entrenamiento de 0.999 y en validación un valor de 0.831. Por lo tanto, se puede observar que los resultados de estos últimos dos modelos son muy similares, permitiendo pensar que con un tuneo correcto de sus hiperparámetros podrían mejorar su comportamiento y disminuir el sobre ajuste en los datos, manteniendo alta la *performance* en validación.

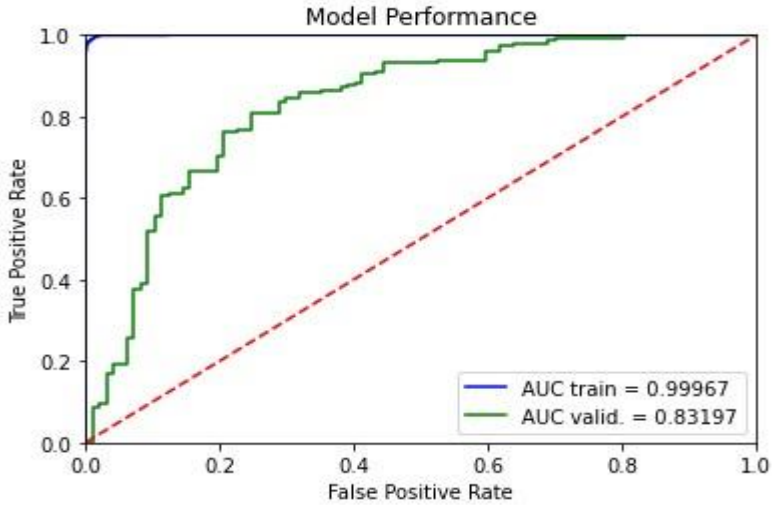

Figura 31: AUC-ROC, LightGBM en entrenamiento y validación.

Tal como se desarrolló en la Sección 3.4 (Optimización de hiperparámetros), en el presente trabajo se aplicó '*Random search Cross-Validation'* para la optimización de los hiperparámetros para los algoritmos de Random Forest y LightGBM.

Puntualmente para Random Forest y LightGBM, se obtuvieron los siguientes valores para los distintos hiperparámetros:

Hiperparámetros	Valores
n_estimadores	1000
max_depth	4
min_samples_leaf	3
min_samples_split	5
max_leaf_nodes	None
max_features	'log2'
criterion	'entropy'
bootstrap	false

Tabla 7: Hiperparámetros óptimos para Random Forest.

Hiperparámetros	Valor
colsample_bytree	0.5297944138219322
min_child_samples	108
min_child_weight	0.01
num_leaves	25
reg_alpha	0
reg_lambda	10
subsample	0.40818669436122185

Tabla 8: Hiperparámetros óptimos para LightGBM.

Con estos hiperparámetros óptimos se obtuvo una mejora en torno al sobreajuste en los datos para ambos algoritmos.

En cuánto a Random Forest, la AUC-ROC en entrenamiento disminuyó a 0.822, acercándose al valor de validación de 0.815. Si bien la *performance* en validación cae unos puntos, puede decirse que este modelo generaliza mejor con instancias desconocidas.

Luego, para LightGBM también disminuyó la AUC-ROC en entrenamiento a 0.946, pero en validación la performance aumentó a 0.844. Sin embargo con el modelo de LightGBM la diferencia entre entrenamiento y validación tiene una mayor amplitud que la obtenida con Random Forest, por lo que se estaría perdiendo un poco de poder de generalización. En las siguientes figuras se puede visualizar los resultados obtenidos para ambos modelos:

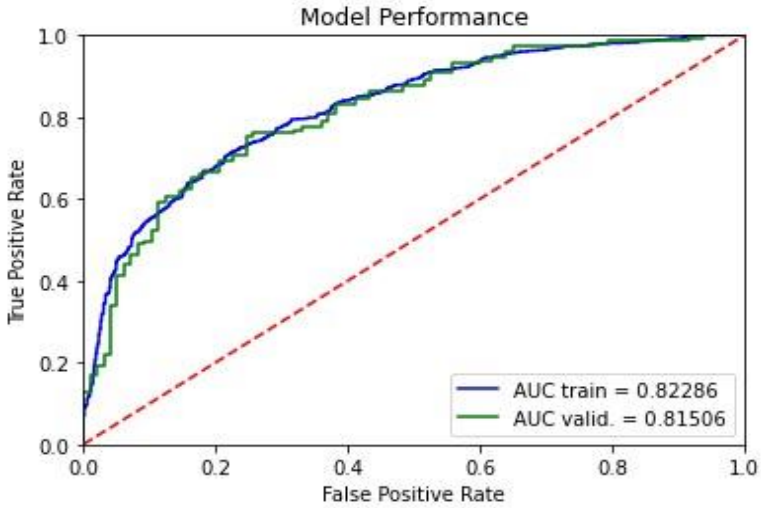

Figura 32: AUC-ROC, Random Forest en entrenamiento y validación luego de la optimización de los hiperparámetros.

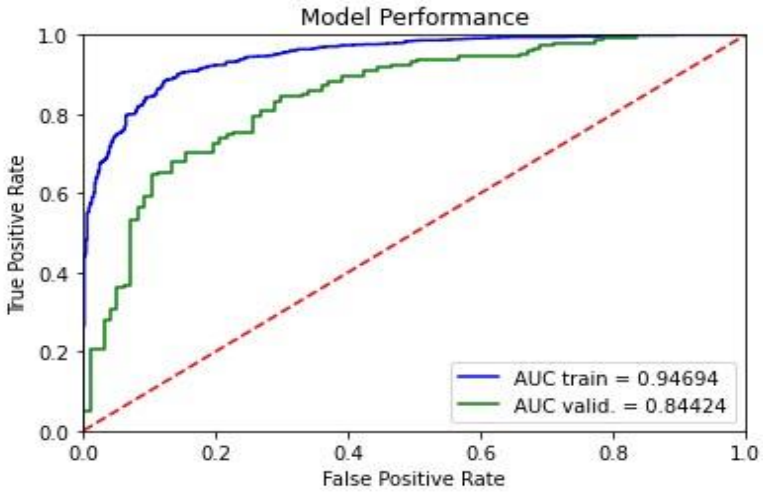

Figura 33: AUC-ROC, LightGBM en entrenamiento y validación luego de la optimización de los hiperparámetros.

4.2. Selección de modelo

Tras analizar los resultados obtenidos con los distintos modelos probados anteriormente, se llega a la decisión de seleccionar como modelo prioritario para la resolución del problema de negocio planteado, el de Random Forest, con la optimización de hiperparámetros encontrada.

El motivo de dicha decisión es que, si bien el modelo de LightGBM alcanzó una mayor *performance* en validación (0.844 *versus* 0.815), puede decirse que este no está generalizando mejor que Random Forest. Esto puede notarse en la diferencia existente entre la AUC-ROC en entrenamiento *versus* validación, la cuál tiene una mayor amplitud que la obtenida con Random Forest.

Por lo tanto, frente a la cantidad de datos con los que se cuenta, y las variables existentes, se opta por 'sacrificar' unos puntos de *performance* por un modelo que generalice mejor y genere un menor sobre ajuste en los datos.

4.3. Elección de punto de corte (*threshold*)

Luego de haber obtenido los resultados de *performance* de los distintos modelos, fue notorio el problema de sobreajuste en los datos. Por tal motivo, se decidió avanzar para lograr entender bien lo que estaba sucediendo.

Para realizar los modelos, se utilizó *scikit-learn*, una biblioteca de aprendizaje automático para el lenguaje de programación en Python. Los modelos que se generan a partir de la función *'predict'* de dicha librería se dice que son *Non-CostSensitive*, lo cuál significa que están prediciendo con un punto de corte (*threshold*) de 0.5. A diferencia de lo que sucede con la función *'predict_proba',* con la cuál se predicen las probabilidades de las clases buscando el mejor punto de corte. Razón por la cual, se decidió utilizar *'predict_proba'*, buscando cuál sería el mejor *threshold* para realizar las predicciones.

En la siguiente figura se intenta visualizar cómo es la distribución de probabilidades de las clases 0 (perdidas) y 1 (ganadas) en el *dataset* de entrenamiento, en función de cómo el modelo de Random Forest lo está prediciendo:

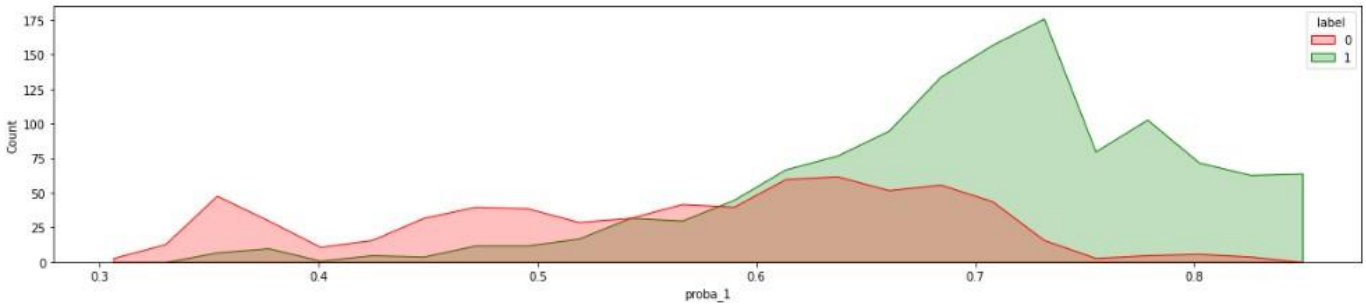

Figura 34: Distribución de probabilidades del *label* 0 (perdidas) y 1 (ganadas) del modelo Random Forest en el *dataset* de entrenamiento.

Aquí se puede observar que si se dejaría el punto de corte en 0.5 tal como lo haría la función *'predict'*, probablemente las negociaciones ganadas las prediga correctamente, pero van a existir muchas pérdidas que las va a estar clasificando como ganadas incorrectamente, por lo que queda claro visualmente que no es el mejor *threshold*.

A la vez, si se piensa en un punto de corte en torno al valor 0.75, aproximadamente dónde casi termina la distribución de probabilidades de las negociaciones perdidas, seguramente el modelo no clasifique de manera incorrecta las perdidas, pero sí se estarán prediciendo muchas negociaciones ganadas como perdidas.

De esta manera, es necesario encontrar el mejor punto de corte. Es decir, aquel que se encuentre lo más cercano del valor 1 de la AUC-ROC, lo que indicaría un alto valor de *TPR* y un bajo valor de *FPR.*

Una manera de calcular el punto de corte óptimo en una curva ROC es haciendo: '$TPR-FPR$'. En el punto en el que '$TPR-FPR$' está en su valor máximo, es dónde se considera el óptimo.

Tras realizar este análisis se obtiene que el mejor *threshold* es igual a 0.6518 en el modelo de Random Forest. Se puede apreciar a continuación:

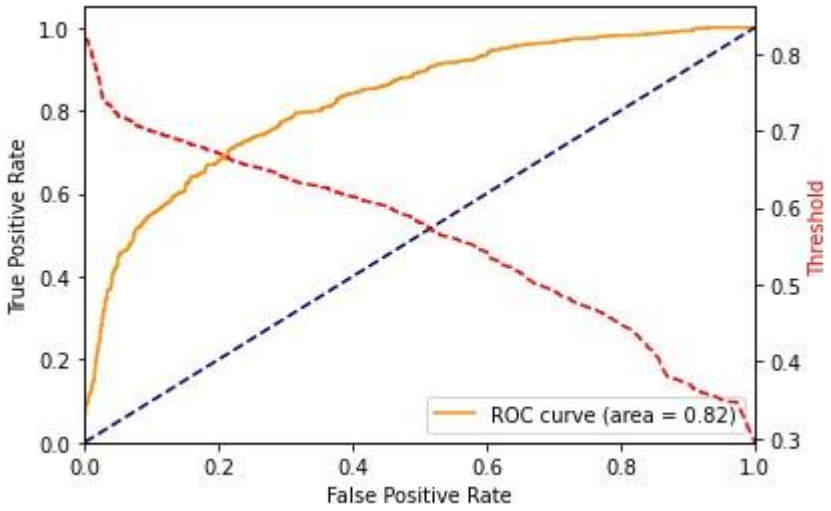

Figura 35: Obtención de punto de corte óptimo en la AUC-ROC.

Luego de realizar el análisis previamente explicado, se continuó evaluando la *performance* del modelo Random Forest en el *dataset* de testeo, obteniendo los siguientes resultados:

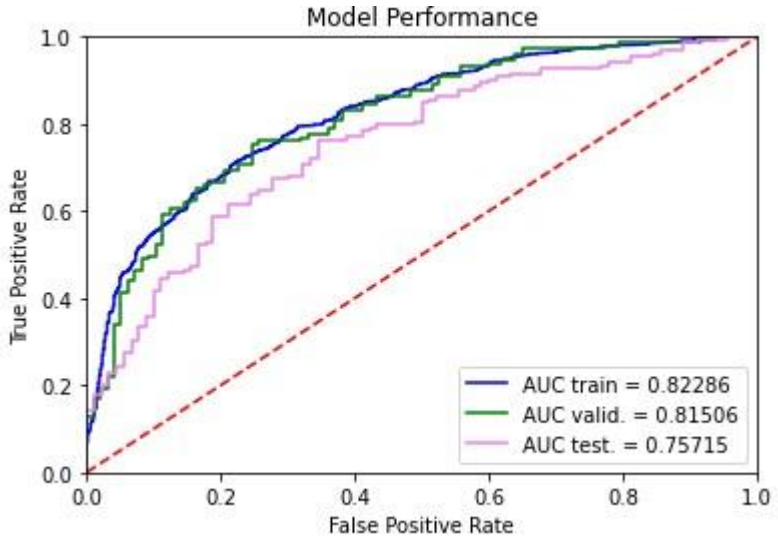

Figura 36: AUC-ROC, Random Forest en entrenamiento, validación y testeo luego de la optimización de los hiperparámetros.

Es notorio cómo el valor de AUC-ROC disminuye en testeo a 0.757. Esto denota un problema en la generalización lograda por el modelo. Por tal motivo, se decidió avanzar realizando un análisis con Matrices de Confusión.

A pesar de que las clases se encuentran desbalanceadas, se cuenta con el valor del *threshold* óptimo, lo que posibilita realizar una Matriz de Confusión, que permitirá evaluar con una mayor precisión el desempeño del clasificador.

La idea detrás de este análisis, es entender cuántas observaciones fueron correctamente o incorrectamente clasificadas y en donde falla el modelo cuando se lo prueba en un conjunto de datos nunca antes visto por el mismo. Es decir, contar el número de veces que las instancias de la clase positiva ('1' = Ganadas) se clasifican como negativas ('0' = Perdidas) y viceversa. Por lo tanto, no solo brinda información sobre los errores que comete el clasificador, sino que también que tipos de errores se generan.

En la siguiente figura se detallan los valores encontrados con la Matriz de Confusión realizada sobre del *dataset* de entrenamiento.

		Predict Values	
		0	1
Actual values	0	0,260	0,090
	1	0,169	0,479

Figura 37: Matriz de Confusión en el *dataset* de entrenamiento para modelo de Random Forest.

Se puede observar que, en términos relativos, los *TP* (*True Positive*) y *TN* (*True Negative*), es decir donde el modelo predice correctamente la clase positiva y negativa, se presentan valores altos de 47.9% y 26% respectivamente, por lo que es mejor prediciendo a la las negociaciones que son ganadas.

Luego, para los *FN* (*False Negative*) que representan la predicción incorrecta del modelo de la clase positiva como negativa, se obtiene un valor de 16.9%. Y para los *FP* (*False Positive*), los cuáles indican una predicción incorrecta de la clase negativa como positiva, se genera un valor más bajo de 9%.

Luego de evaluar la Matriz de Confusión en el *dataset* de entrenamiento, se procedió a realizarla utilizando los datos de testeo y el punto de corte encontrado.

		Predict Values	
		0	1
Actual values	0	0,242	0,127
	1	0,168	0,460

Figura 38: Matriz de Confusión en *dataset* de testeo para modelo de Random Forest.

En dicha matriz se puede notar que tanto los *TP* como los *TN* bajan unos puntos sus valores, a 46% y 24.2% respectivamente. Por otro lado, se incrementa notablemente el valor de los *FP* a 12.7%, manteniéndose similar el valor de los *FN*. Esto destaca un punto posible de mejora futura. Si bien el modelo generaliza bien, habría que intentar que aprenda mejor de los datos para separar las clases y evitar de esta forma que no queden tan superpuestas, lo que ayudaría a bajar el valor de los FP, dejando de predecir algunas de las negociaciones perdidas como ganadas.

Por lo tanto, frente a este análisis se pueden detallar algunos puntos considerables a tener en cuenta. Tal como se mencionó anteriormente, la tasa de falsos positivos (*FP*) se incrementó cuándo se probó en el *dataset* de testeo. Esto denota un problema. En este caso, si bien el modelo se está equivocado más en predecir las negociaciones perdidas cómo ganadas (*FN* = 16.8%), el valor de los *FP* no es menos importante, ya que a nivel de negocio esto tendría un mayor impacto. Implicaría que negociaciones perdidas sean predichas como ganadas, lo cuál se estarían perdiendo potenciales clientes, que con una estrategia de retención apropiada podría evitarse si el modelo los predijera correctamente como perdidas. A su vez, esto se ve acentuado por el menor valor de los verdaderos negativos (*TN* = 24.2%), respecto al de verdaderos positivos (*TP* = 46%).

4.4. Estrategia de descuentos

Viendo la distribución de probabilidades asignadas a las negociaciones perdidas y ganadas en la figura 34, se puede pensar en generar una política de descuentos distinta a la vigente actualmente. Tal como se mencionó en la introducción, según información brindada por la empresa, las negociaciones que realizan los ejecutivos comerciales ya tienen implícito un descuento del 15% por *default*.

Así, utilizando distintos rangos de probabilidades podría pensarse en una posible estrategia basada en información más acertada. Se puede observar en la figura 34 que a los clientes que el modelo le asigna una probabilidad entre 0 y 0.5 tienen

mayores chances de derivar en una negociación perdida, a diferencia de los que tienen una probabilidad entre 0.75 y 1, que tendrán una mayor probabilidad de terminar cerrando un nuevo contrato tras la renegociación. Pero, para aquellos que se encuentran entre 0.5 y 0.75 de probabilidad existe una mayor incertidumbre.

De esta manera, sería una buena estrategia pensar en aplicarles un descuento mayor (30%) a los que tengan una probabilidad asignada por el modelo entre 0 y 0.5, un descuento medio (15%) a los que están entre 0.5 y 0.75, y por último los que que tengan una probabilidad mayor a 0.75 no se les debería aplicar descuento.

Considerando esta posible estrategia basada en descuentos, la empresa estaría dejando de perder dinero cuando se le ofrece descuentos a clientes que potencialmente tengan más chances de renovar, compensando y utilizando dicho dinero para ofrecerles descuentos a los otros segmentos de clientes que tienen más probabilidades de rechazar la negociación y derivar en una baja. En definitiva, se estaría mejorando la retención de clientes haciendo un mejor uso de los cupones de descuentos ofrecidos.

A pesar de que sería particularmente interesante poner a prueba en producción dicha estrategia de descuentos mencionada anteriormente, en los siguientes párrafos se hará mención a distintos escenarios, para determinar en términos relativos el ingreso extra obtenido o no una vez aplicados los descuentos detallados.

Antes de continuar, es importante mencionar que se trabajará sobre la variable *'total_price_t1'* la cuál refiere al valor de la negociación. Este precio, ya tiene establecido un descuento del 15% por *default*. Es decir, los ejecutivos comerciales en el momento de negociar ya incluyen este 15% sin discriminar por cliente.

En primer lugar, pensando en un escenario totalmente optimista se tienen en cuenta los siguientes puntos para obtener el valor final de ingresos factibles a generar con la estrategia de descuentos planteada:

- Aquellas negociaciones ganadas con una probabilidad predicha por el modelo mayor a 0.75, se les quita ese 15% previamente realizado, llevando su valor al precio original. En este escenario optimista se está hipotetizando que todos los clientes dentro de estos parámetros no se darán de baja aunque se les quite ese 15% y continuarán con la adquisición de un paquete de

membresías. La razón por detrás, es que generalmente son clientes que se encuentran satisfechos y activos en la participación por lo que no los hará desertar esta nueva acción.

- Para las negociaciones perdidas con una probabilidad menor a 0.5, en lugar de dejarles ese 15%, se les aplica un 30% sobre el precio original, asumiendo que todos pasarán a contratar algún paquete y comenzarán a ser clientes de la empresa.
- Aquellas negociaciones ganadas con una probabilidad entre 0.5 y 0.75 se las considera tal cuál como están con ese 15% de descuento aplicado.
- Por último, las negociaciones ganadas con una probabilidad menor a 0.5 se les aplica un 30% en lugar de un 15%. Si bien esto no tiene mucho sentido ya que igualmente resultaron en negociaciones ganadas, la estrategia describe aplicar un 30% a aquellas predichas con una probabilidad menor a 0.5.

Así, teniendo en cuenta este detalle, se obtiene un ingreso aproximado del 21.8% superior al ya existente en base a las negociaciones ganadas antes de aplicar la estrategia de descuentos.

En segundo lugar, se puede pensar en un escenario intermedio. Aquí se considera:

- Igual que el escenario anterior, las negociaciones ganadas con una probabilidad predicha mayor a 0.75 continúan comprando a pesar de quitarles el 15% de descuento.
- A diferencia del escenario optimista, las pérdidas con una probabilidad menor a 0.5, siguen permaneciendo como tales aunque se les haya ofrecido un 30% de descuento.
- Para las ganadas con una probabilidad entre 0.5 y 0.75 aplica el mismo criterio que el escenario anterior.
- Aplica el mismo procedimiento para las negociaciones ganadas con una probabilidad menor a 0.5.

Con estas particularidades, el ingreso que se genera es un 15% superior al existente actualmente.

Tercer y último escenario que se genera es el pesimista en dónde:

- Al quitarles el 15% de descuento a las clientes cuyas negociaciones son ganadas con una probabilidad predicha mayor a 0.75, deciden darse de baja.
- Consideró las negociaciones ganadas con una probabilidad entre 0.5 y 0.75 con el descuento ya aplicado de 15%.
- También se tienen en cuenta las negociaciones ganadas con una probabilidad menor a 0.5 con un descuento del 30%.

Con este último escenario, la estrategia de descuentos termina afectando negativamente los ingresos, con una pérdida aproximada del 6.5%.

En resumen, con la supuesta implementación de está estrategia de descuentos, en base a los escenarios simulados, se espera obtener buenos resultados económicos, ya que únicamente en el caso pesimista se estaría perdiendo con un impacto negativo en los ingresos. Igualmente, hay que tener en cuenta la necesaria puesta en producción de lo planteado, lo cuál sería la única manera factible de comprobarlo y entender realmente, no sólo el comportamiento del modelo y su *performance*, sino también cómo los ejecutivos comerciales usarían dicha información para llevarla a la práctica, y entonces, agregar valor a la compañía.

4.5. Importancia de las variables

Luego de especificar los resultados obtenidos con los distintos modelos, se decidió buscar cuáles fueron las variables que influyeron a la hora de realizar las predicciones.

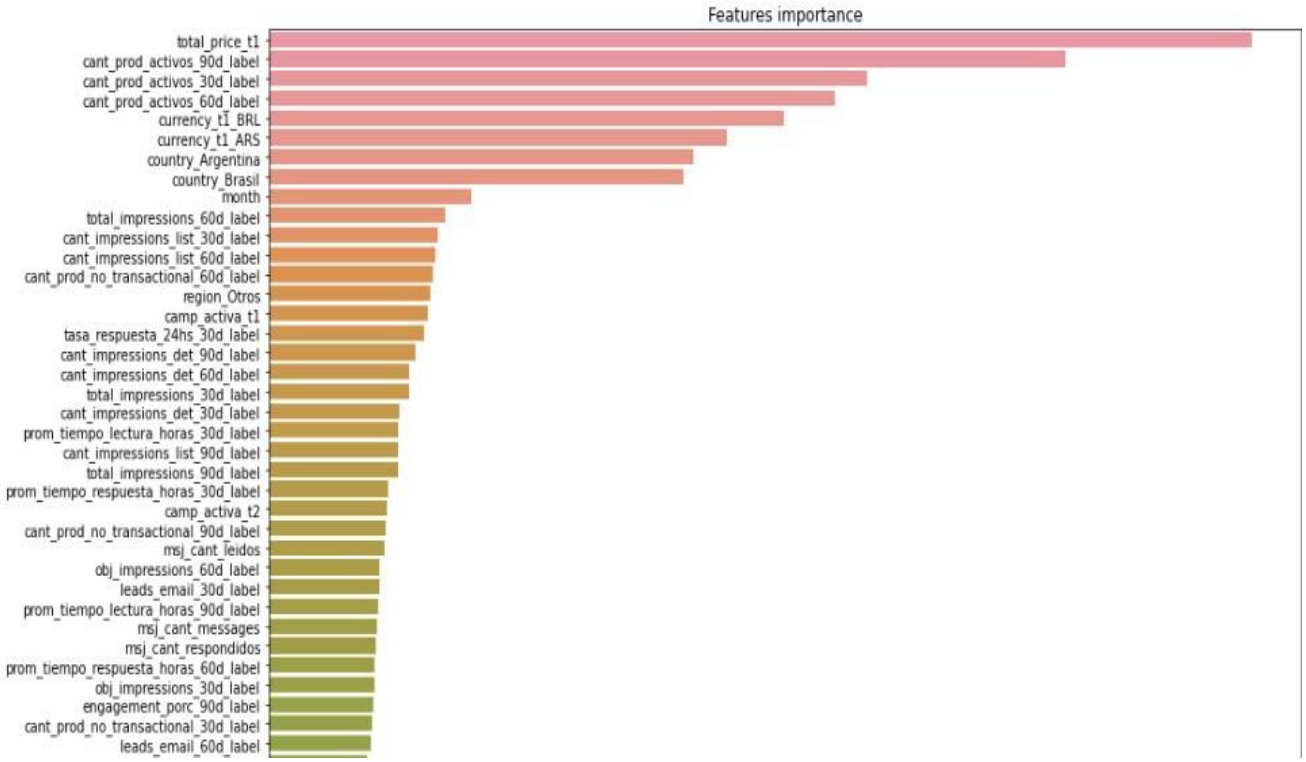

Figura 39: Top de las primeras 30 variables en términos de importancia para el modelo final.

Cómo puede observarse en el gráfico, la variable *'total_price_t1'* cobra una gran importancia. Es seguida por las variables *'cant_prod_activos_90d_label'*, *'cant_prod_activos_30d_label'* y *'cant_prod_activos_60d_label'* .

Un aspecto a destacar es que dentro de las primeras 30 variables más importantes tenidas en cuenta por el modelo, la gran mayoría refiere a las variables de tendencias creadas, salvo por algunas cómo: *'currency_t1_BRL'*, *'currency_t1_ARS'*, *'country_Argentina'*, *'country_Brasil'*, *'month'*, *'region_Otros'*, *'camp_activa_t1'*, *'camp_activat_t2'*, *'msj_cant_leidos'*, *'msj_cant_messages'* y *'msj_cant_respondidos'*.

5. Conclusión final y trabajo a futuro

5.1. Conclusión final

En el presente trabajo se ha demostrado cómo distintas técnicas de *Machine*

Learning pueden ser utilizadas para predecir la propensión de compra de un cliente. Particularmente, se buscaba encontrar un método que permita predecir la propensión a renovar las membresías por parte de los vendedores de la plataforma de comercio electrónico con la mayor *performance* posible y poder de generalización.

Tras los resultados obtenidos y la selección de modelo realizada, se puede decir que los ejecutivos comerciales podrían tener a disposición una nueva herramienta e información acertada al momento de llevar a cabo la renegociación de las membresías. Así, estarán mejor preparados para armar no solo una estrategia de negociación efectiva con un cliente puntual, sino también mejorar la priorización de la cartera de clientes activas que tienen a disposición.

En lo que respecta al análisis de distribución de las probabilidades asignadas por el modelo Random Forest, se puede destacar la estrategia de descuentos planteada con la cuál podría esperarse que se genere un impacto positivo y agregue valor a la compañía.

5.2. Propuestas de mejoras futuras

Existen algunos puntos a considerar para profundizar como trabajo futuro. A continuación se van a mencionar y explicar brevemente el motivo:

- Tras haber encontrado que el modelo de LightGBM lograba una mejor *performance* en validación, pero una peor generalización que la obtenida con Random Forest, sería interesante continuar indagando sobre dicho modelo, con no sólo un mejor tuneo de hiperparámetros, sino también con una mayor cantidad de datos, lo que ayudaría probablemente a conseguir una mejor generalización y menor sobreajuste, acercándose más la AUC-ROC de entrenamiento a la de validación. También sería interesante evaluar un *pooling* o combinación de modelos de pronóstico.
- Realizar *feature selection* en búsqueda de obtener una mayor *performance,* para eliminar del modelo ciertas variables que se consideren que tienen pocas probabilidades de tener poder predictivo.

- Sería interesante probar no sólo el modelo en producción, sino también la política y estrategia de descuentos.

6. Bibliografía

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of machine learning research, 13(2).

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). An introduction to statistical learning: with applications in R. Spinger

Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.".

Humphreys, E. (2015). Describir la cadena de valor del E-commerce en la República Argentina.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, pp. 3146-3154

Koehrsen, W., (2018). A Conceptual Explanation of Bayesian Hyperparameter Optimization for Machine Learning. Towards data science [online]. Viewed 14 June 2020. Available from: <https://towardsdatascience.com/a-conceptualexplanation-of-bayesian-model-based-hyperparameter-optimization-for-machinelearning-b8172278050f>

Mitchell, T. M., & Mitchell, T. M. (1997). Machine learning (Vol. 1, No. 9). New York: McGraw-hill.

Mitchell, T. M. (2006). The discipline of machine learning (Vol. 9). Pittsburgh: Carnegie Mellon University, School of Computer Science, Machine Learning Department.

Quinlan, J.R. (1986). Induction of Decision Trees. Machine Learning, vol. 1, pp. 81–106.

Ramos, J. (2017). E-Commerce 2.0. XinXii.

Rodríguez, K., Ortiz, O., Quiroz, A., & Parrales, M. (2020). El e-commerce y las Mipymes en tiempos de Covid-19. Revista espacios, 41(42), 100-118.

Rodriguez-Fernández, J. (1999). Ockham's razor. Endeavour, 23(3), 121-125.

Tom Fawcett, (2006). An introduction to ROC analysis. Pattern Recognition Learning 27, 8 (June 2006), 861–874. DOI: <https://doi.org/10.1016/j.patrec.2005.10.010>

7. Anexo

Variables base de datos - Cuentas: Id_Completo, Account Name, Billing Country, ID Market Place, Tipo de empresa

Variables base de datos - Actividades: Task Subtype, Account Name, Country, Date

Variables base de datos - Oportunidades: Opp_Id_Completo, Merchan ID, Billing Country, Account Name, Stage, Created Date, Fecha Fin Servicio, Rubro (UN), División, Total Price, Opportunity Currency, Categoría, Subcategoría, Product Name, Opp maturity, Campaña activa

Variables base de datos - *Merchant Performance*: Month Year, Country, Región, Id, Membership, Category N1 Ppal, Category N2 Ppal, Engmt, % Engmt, Total Impressions, Obj. Impressions, Dif. Obj. Impressions, % Cump. Obj. Impr, Listings, Impressions List, Impressions Det, CTR, CR Det (Email), CR Det (Total), Leads Email, , Leads Phone, Leads Whatsapp, Total Leads, Tasa Lectura, Tasa Rta., Tasa Rta. Menor 24 hs, Prom. Tiempo Lectura (horas), Prom. Tiempo Resp. (horas)

Variables base de datos - Listado productos: MonthYear, Country, Merchant ID, Model type, Product status

Variables base de datos - Mensajes: Date, Merchant id, Product Country, Transactions type, Leído, Respondido

Variables *dataset* principal

N°	Variable	Valores no nulos	Tipo de dato
0	merchant_id	1946 non-null	object
1	oportunidad_actual_id	1946 non-null	object
2	oportunidad_label_id	1946 non-null	object
3	fecha_opt_actual	1946 non-null	object
4	fecha_opt_label	1946 non-null	object
5	label_tag	1946 non-null	object
6	label	1946 non-null	int64
7	mkt_place_id	1946 non-null	int64
8	country	1946 non-null	object
9	rubro_t1	1946 non-null	object
10	division_t1	1946 non-null	object
11	total_price_t1	1946 non-null	float64
12	currency_t1	1946 non-null	object
13	categoria_t1	1946 non-null	object
14	subcategoria_t1	1814 non-null	object
15	product_name_t1	1946 non-null	object
16	camp_activa_t1	473 non-null	object
17	opportunity_duration_t1	1946 non-null	int64
18	camp_activa_t2	473 non-null	object
19	tipo_empresa	1944 non-null	object

20	sum_task	1946 non-null	float64
21	sum_call	1946 non-null	float64
22	sum_email	1946 non-null	float64
23	activity_frequency_mean	1232 non-null	float64
24	activity_frequency_median	1232 non-null	float64
25	region	1831 non-null	object
26	membership_30d_label	1831 non-null	object
27	cat_n1	1831 non-null	object
28	cat_n2	1831 non-null	object
29	engagement_30d_label	1831 non-null	object
30	engagement_porc_30d_label	1831 non-null	object
31	total_impressions_30d_label	1831 non-null	float64
32	obj_impressions_30d_label	1831 non-null	float64
33	dif_obj_impressions_30d_label	1831 non-null	float64
34	cump_obj_impressions_porc_30d_label	1831 non-null	object
35	cant_listings_30d_label	1831 non-null	float64
36	cant_impressions_list_30d_label	1831 non-null	float64
37	cant_impressions_det_30d_label	1831 non-null	float64
38	ctr_30d_label	1831 non-null	object
39	cr_det_email_30d_label	1831 non-null	object
40	cr_det_total_30d_label	1831 non-null	object
41	leads_email_30d_label	1831 non-null	float64

42	leads_phone_30d_label	1831 non-null	float64
43	leads_whatsapp_30d_label	1831 non-null	float64
44	total_leads_30d_label	1831 non-null	float64
45	tasa_lectura_30d_label	1831 non-null	object
46	tasa_respuesta_30d_label	1831 non-null	object
47	tasa_respuesta_24hs_30d_label	1831 non-null	object
48	prom_tiempo_lectura_horas_30d_label	1831 non-null	object
49	prom_tiempo_respuesta_horas_30d_label	1831 non-null	object
50	cant_prod_no_transactional_30d_label	510 non-null	float64
51	cant_prod_quotable_30d_label	510 non-null	float64
52	cant_prod_transactional_30d_label	510 non-null	float64
53	cant_prod_activos_30d_label	510 non-null	float64
54	cant_prod_inactivos_30d_label	510 non-null	float64
55	membership_60d_label	1840 non-null	object
56	engagement_60d_label	1840 non-null	object
57	engagement_porc_60d_label	1840 non-null	object
58	total_impressions_60d_label	1840 non-null	float64
59	obj_impressions_60d_label	1840 non-null	float64

60	dif_obj_impressions_60d_label	1840 non-null	float64
61	cump_obj_impressions_porc_60d_label	1840 non-null	object
62	cant_listings_60d_label	1840 non-null	float64
63	cant_impressions_list_60d_label	1840 non-null	float64

64	cant_impressions_det_60d_label	1840 non-null	float64
65	ctr_60d_label	1840 non-null	object
66	cr_det_email_60d_label	1840 non-null	object
67	cr_det_total_60d_label	1840 non-null	object
68	leads_email_60d_label	1840 non-null	float64
69	leads_phone_60d_label	1840 non-null	float64
70	leads_whatsapp_60d_label	1840 non-null	float64
71	total_leads_60d_label	1840 non-null	float64
72	tasa_lectura_60d_label	1840 non-null	object
73	tasa_respuesta_60d_label	1840 non-null	object
74	tasa_respuesta_24hs_60d_label	1840 non-null	object
75	prom_tiempo_lectura_horas_60d_label	1840 non-null	object
76	prom_tiempo_respuesta_horas_60d_label	1840 non-null	object
77	cant_prod_no_transactional_60d_label	699 non-null	float64
78	cant_prod_quotable_60d_label	699 non-null	float64
79	cant_prod_transactional_60d_label	699 non-null	float64
80	cant_prod_activos_60d_label	699 non-null	float64
81	cant_prod_inactivos_60d_label	699 non-null	float64
82	membership_90d_label	1815 non-null	object
83	engagement_90d_label	1815 non-null	object
84	engagement_porc_90d_label	1815 non-null	object
85	total_impressions_90d_label	1815 non-null	float64

86	obj_impressions_90d_label	1815 non-null	float64

87	dif_obj_impressions_90d_label	1815 non-null	float64
88	cump_obj_impressions_porc_90d_label	1815 non-null	object
89	cant_listings_90d_label	1815 non-null	float64
90	cant_impressions_list_90d_label	1815 non-null	float64
91	cant_impressions_det_90d_label	1815 non-null	float64
92	ctr_90d_label	1815 non-null	object
93	cr_det_email_90d_label	1815 non-null	object
94	cr_det_total_90d_label	1815 non-null	object
95	leads_email_90d_label	1815 non-null	float64
96	leads_phone_90d_label	1815 non-null	float64
97	leads_whatsapp_90d_label	1815 non-null	float64
98	total_leads_90d_label	1815 non-null	float64
99	tasa_lectura_90d_label	1815 non-null	object
100	tasa_respuesta_90d_label	1815 non-null	object
101	tasa_respuesta_24hs_90d_label	1815 non-null	object
102	prom_tiempo_lectura_horas_90d_label	1815 non-null	object
103	prom_tiempo_respuesta_horas_90d_label	1815 non-null	object
104	cant_prod_no_transactional_90d_label	886 non-null	float64
105	cant_prod_quotable_90d_label	886 non-null	float64
106	cant_prod_transactional_90d_label	886 non-null	float64
107	cant_prod_activos_90d_label	886 non-null	float64

108	cant_prod_inactivos_90d_label	886 non-null	float64
109	msj_cant_leidos	1853 non-null	float64
110	msj_cant_respondidos	1853 non-null	float64
111	msj_cant_messages	1853 non-null	float64
112	msj_cant_quotations	1853 non-null	float64
113	msj_cant_sales_direct	1853 non-null	float64
114	msj_cant_whatsapp	1853 non-null	float64

Variables numéricas tendencistas

Variables numéricas	**Valores nulos**
engagement_porc_30d_label	115
total_impressions_30d_label	115
obj_impressions_30d_label	115
dif_obj_impressions_30d_label	115
cump_obj_impressions_porc_30d_label	115
cant_listings_30d_label	115
cant_impressions_list_30d_label	115
cant_impressions_det_30d_label	115
ctr_30d_label	115
cr_det_email_30d_label	115
cr_det_total_30d_label	115
leads_email_30d_label	115

leads_phone_30d_label	115
leads_whatsapp_30d_label	115
total_leads_30d_label	115
tasa_lectura_30d_label	115
tasa_respuesta_30d_label	115
tasa_respuesta_24hs_30d_label	115
prom_tiempo_lectura_horas_30d_label	115
prom_tiempo_respuesta_horas_30d_label	115
cant_prod_no_transactional_30d_label	1436
cant_prod_quotable_30d_label	1436
cant_prod_transactional_30d_label	1436
cant_prod_activos_30d_label	1436
cant_prod_inactivos_30d_label	1436
engagement_porc_60d_label	106
total_impressions_60d_label	106

obj_impressions_60d_label	106
dif_obj_impressions_60d_label	106
cump_obj_impressions_porc_60d_label	106
cant_listings_60d_label	106
cant_impressions_list_60d_label	106
cant_impressions_det_60d_label	106
ctr_60d_label	106
cr_det_email_60d_label	106

cr_det_total_60d_label	106
leads_email_60d_label	106
leads_phone_60d_label	106
leads_whatsapp_60d_label	106
total_leads_60d_label	106
tasa_lectura_60d_label	106
tasa_respuesta_60d_label	106
tasa_respuesta_24hs_60d_label	106
prom_tiempo_lectura_horas_60d_label	106
prom_tiempo_respuesta_horas_60d_label	106
cant_prod_no_transactional_60d_label	1247
cant_prod_quotable_60d_label	1247
cant_prod_transactional_60d_label	1247
cant_prod_activos_60d_label	1247
cant_prod_inactivos_60d_label	1247
engagement_porc_90d_label	131
total_impressions_90d_label	131
obj_impressions_90d_label	131
dif_obj_impressions_90d_label	131
cump_obj_impressions_porc_90d_label	131
cant_listings_90d_label	131
cant_impressions_list_90d_label	131
cant_impressions_det_90d_label	131
ctr_90d_label	131
cr_det_email_90d_label	131
cr_det_total_90d_label	131

leads_email_90d_label	131
leads_phone_90d_label	131
leads_whatsapp_90d_label	131
total_leads_90d_label	131
tasa_lectura_90d_label	131
tasa_respuesta_90d_label	131
tasa_respuesta_24hs_90d_label	131
prom_tiempo_lectura_horas_90d_label	131
prom_tiempo_respuesta_horas_90d_label	131
cant_prod_no_transactional_90d_label	1060
cant_prod_quotable_90d_label	1060
cant_prod_transactional_90d_label	1060
cant_prod_activos_90d_label	1060
cant_prod_inactivos_90d_label	1060

Categorías y cantidades de registros de la variable categórica: 'subcategoria_t1'

Valores	**Cantidad registros**
Campos	159
Tractores	154
Otros	133
Sembradoras	71
Repuestos Agrícolas	61
Instalaciones para Ganadería	56
Acopio y Almacenaje	54
Camiones	48
Tanques	48

Acoplados	42
Tolvas	38
Cosechadoras	35
Generadores de energía	33
Neumáticos	30

Pulverizadoras	30
Casillas Rurales	27
Agricultura de precisión	26
Balanzas y básculas	25
Departamentos	25
Carrocerías	25
Desmalezadoras	23
Plataformas y Cabezales	23
Instalaciones para Riego	22
Fertilizantes	21
Mixers	21
Cortadoras de Césped	20
Rastras	20
Accesorios para maquinaria agrícola	20
Protección de cultivos	19
Autos	18
Instrumentos de Medición	18
Estructuras	18
Fertilizadoras	18
Otros elementos de Infraestructura	17
Alambrados y cercos	16

Palas cargadoras	15
Semirremolques	15
Autoelevadores	15
Materiales de Construcción	14
Semillas de Cultivos Extensivos	14
Casas	14
Otras Herramientas	14
Repuestos Automotor	13
Camionetas	13
Hoyadoras	13
Pulverizadoras Manuales	12
Bombas	11

Moledoras y Quebradoras	10
AgTech	10
Motosierras	10
Accesorios para Autos y Camionetas	9
Compresores de Aire	9
Rolos	9
Formación	9
Construcciones y Estructuras	9
Otras Maquinarias	8
Accesorios para Camiones	8
Llantas	8
Iluminación	8
Seguros para Patrimonios	7

Hidrolavadoras	7
Repuestos Viales	7
Instalaciones para Cerdos	6
Embolsadoras	6
Nutrición Animal	6
Terrenos	6
Cargadores y Transportadores de Rollos	6
Picadoras	6
Utilitarios	5
Retroexcavadoras	5
Plataformas Elevadoras	5
Otros Inmuebles	5
Otras	5
Créditos	5
Soldadoras	4
Fresadoras	4
Procesadoras de Alimentos balanceados	4
Otros Repuestos	4
Palas	4
Extractoras	4

Instalaciones para Tambos	4
Minicargadoras	4
Grúas	4
Instalaciones para Avicultura	4
Palas de Arrastre	3
Otros Servicios	3

Rastrillos	3
Palas y Niveladoras	3
Paratil	3
Taladros	3
Aviones	3
Seguridad vehicular	3
Amoladoras	3
Chimangos	3
Aceites para Motores	3
Enfardadoras	3
Asesoramiento	3
Instrumentos de Laboratorio	3
Casas de Campo	2
Cajas de herramientas	2
Tarjetas	2
Combustibles	2
Plantas	2
Barredoras	2
Otros Accesorios	2
Pinturas y Revestimientos	2
Carros Forrajeros y Compactadores	2
Engrasadoras	2
Segadoras	2
Inoculadores	2
Excavadoras	2
Aceites para Transmisión	2

Pinzas	2

Clasificadoras de Semillas	2
Cortacercos	2
Repuestos Rodados	2
Niveladoras	2
Bordeadoras	2
Ómnibus	2
Sierras Industriales	2
Sierras	2
Repuestos Herramientas	2
Subsoladores	2
Motorhomes	2
Estercoleras	2
Dosificadores Variables	2
Insumos Veterinarios	2
Inoculantes	2
Piezas para Construcción	1
Palas con retro	1
Contratistas	1
Seguros para Hacienda	1
Tijeras	1
Ropa	1
Aditivos	1
Llaves	1
Elevadores de Vehículos	1

Maquinaria Frutihortícola	1
Plegadoras de Chapas	1
Azadas	1
Cursos	1
Llaves de Impacto	1
ATV	1
Semillas Forrajeras	1
Silo Bolsas	1
Complejos de Departamentos	1
Motoguadañas	1
Pistolas para Pintar	1
Otros Insumos Agrícolas	1
Vajilla	1
GPS	1
Motos	1
Accesorios para Herramientas	1
Aparejos	1
Servicios Industriales	1
Martillos Hidráulicos	1
Chipeadoras	1
Aspiradoras	1
Consórcios	1
Motoniveladoras	1
Casas Quintas	1

Categorías y cantidades de registros de la variable categórica:

‘product_name_t1’

Valores	Cantidad registros
ARG Advanced 12 meses	373
ARG Primary 12 Meses	345
ARG Plus 12 Meses	299
ARG Select 12 Meses	116
BRA Primary 12 Meses	93
BRA Plus 12 Meses	53
BRA Maquinaria Business Demo Gratis 3 Meses	47
BRA Advanced 12 Meses	43
ARG Starting Financiado 12 meses	42
Demo	38
Membresía Silver	33
Banners Sueltos	31
OLAC Primary 12 meses	23
ARG Move In Financiado 12 meses	21
ARG Move In Contado 12 meses	21
BRA Starting 12 Meses	20
ARG Primary 6 meses	19
BRA Primary 6 Meses	17
BRA Move In 12 Meses	15
ARG Expand Financiado 12 meses	15
ARG Expand Contado 12 meses	13
OLAC Select 12 meses	12
ARG Going Forward Contado 12 meses	11
Membresía Gold 3 Meses	11

ARG Going Forward Financiado 12 meses	10
Desarrollo tecnológico	9
BRA Going Forward 12 Meses	8
OLAC Advanced 12 meses	8
ARG 3 Avisos	7
BRA Select 12 meses	7
ARG Conquer Contado 12 meses	7
OLAC Plus 12 meses	7
ARG Conquer Financiado 12 meses	6
OLAC Starting 6 meses	6
ARG Advanced Contado 12 Meses	6
OLAC Starting 12 Meses	6
Membresía Gold	6
Auspicios	6
BRA Primary 3 Meses	5
ARG Plus 6 meses	5
BRA Plus 6 Meses	5
ARG Starting Contado 12 meses	5
Membresía Platinum	5
BRA Starting 6 Meses	5
BRA Conquer 12 Meses	4

BRA Select 6 meses	4
Membresía Main Sponsor	4
Move-in 6 Meses	4
OLAC Select 6 meses	4
BRA Maquinaria Corporate Demo Gratis 3 Meses	4

ARG Starting Financiado 6 meses	4
ARG Advanced 6 meses	3
ARG 1 Aviso	3
OLAC Advanced 6 meses	3
ARG Move In Financiado 6 meses	3
ARG Primary Contado 12 Meses	3
OLAC Plus 3 meses	3
ARG Advanced Financiado 12 Meses	3
BRA Starting 9 Meses	3
ARG Starting Contado 6 meses	3
BRA Plus 3 Meses	3
Banner Listados Top 3 meses	3
ARG Going Forward Financiado 6 meses	2
BRA Plus 9 Meses	2
Banner Listados Low 3 meses	2
BRA 10 Avisos	2
OLAC Starting 12 meses	2
ARG Move In Contado 6 meses	2
ARG 5 Avisos	2
ARG Primary Financiado 12 Meses	2
BRA 5 Avisos	2
BRA Going Forward 6 Meses	2
ARG Plus Contado 12 Meses	2
BRA Advanced 3 Meses	2
ARG Maquinaria Premium B Financiado 12 Meses	1
OLAC Select 3 meses	1

ARG Rodados Corporate Plata B 12 Meses	1
ARG Expand Financiado 6 meses	1
OLAC Going Forward 12 meses	1
OLAC Plus 6 meses	1
BRA Conquer 9 Meses	1
ARG Expand Contado 6 meses	1
ARG Rodados Standard C&D Financiado 12 Meses	1
ARG Going Forward Contado 6 meses	1
OLAC Business Plus I	1
ARG Maquinaria Premium C&D Contado 12 Meses	1
OLAC Conquer 12 meses	1
BRA 3 Avisos	1
Banner Listados Top 6 meses	1
OLAC Advanced 3 meses	1
Email Marketing	1
BRA Starting 3 Meses	1
BRA Banner Subcategoria	1
ARG Select 6 meses	1
BRA 1 Aviso	1

Categorías y cantidades de registros de la variable categórica: 'region'

Valores	**Cantidad registros**
Buenos Aires	409
Santa Fe	361
Cordoba	302
Otros	177

São Paulo	101
Buenos Aires City	93
Entre Rios	70
Rio Grande Do Sul	62
Paraná	50
La Pampa	44
Minas Gerais	30

Montevideo	26
Chaco	21
Bogota	19
Mendoza	16
Santa Catarina	15
Canelones	13
Mato Grosso	11
Corrientes	9
Antioquia	9
Río Negro	8
Mato Grosso Do Sul	8
Goiás	8
San Luis	7
Tucuman	7

Neuquen	5
Santiago Del Estero	5
Jujuy	4
San Jose	4
Cundinamarca	4
Ceará	3
San Juan	3
Piauí	3
Colonia	3
Caldas	3
Soriano	2
Bahia	2
Tolima	2
Florida	2
Salta	2
Distrito Federal	2
Rio De Janeiro	2
Chubut	2
Magdalena	1
Paysandu	1
Casanare	1
Pará	1

Maldonado	1
Nariño	1
Espírito Santo	1
Catamarca	1
Risaralda	1
Valle Del Cauca	1
Formosa	1
Tocantins	1
Asunción	1
Lavalleja	1
Rio Negro	1
Treinta Y Tres	1
Rondônia	1

Categorías y cantidades de registros de las variables categóricas: 'membership_30d_label', 'membership_60d_label' y **'membership_90d_label'**

Valores	**Cantidad registros**
Sucursal Primary	481
Sucursal Advanced	413
Sucursal Plus	359
Sucursal Select	141
Otros	124
Corporate Starting	99
Corporate Move In	91

Corporate Going Forward	70
Corporate Conquer	51
Demo Liquidez	27
Demo Business	21
Business Churn	19
Empresa Registrada	8
Corporate Expand	8
Demo Business 0	8
Demo Bussiness 1	3
Business I	3
Select test (no usar - FDoming	2
Pack Agro 5	2
Pack Profesional 3	2
Business Select	2
Corporate Demo	2
Pack Agro 3	2
Plano Agro 5	2
Pack Agro 1	1
Business II	1
Profesionales Pack 10	1
Plano Agro 1	1
Demo Business 1	1
Corporate Churn	1