LICENCIATURA EN ECONOMÍA

Determinantes de Éxito en el *e-commerce*: Caso Mercado Libre

LORENZO, Laureano
Registro N° 33226
laureano.lorenzo@fce.uncu.edu.ar

BUSTOS, Benjamin
Registro N° 31903
benjamin.bustos@fce.uncu.edu.ar

Director: Pablo David Mahnic
Co-director: Gustavo Machín Urbay

Mendoza, 2024

Resumen

El éxito en el comercio electrónico depende de algoritmos que determinan qué productos logran mayor visibilidad, y por lo tanto, mayores chances de concretar ventas. Este estudio examina cómo variables como precio, reputación y logística influyen en el posicionamiento de publicaciones en Mercado Libre, tomando como referencia una muestra aleatoria y dos categorías con distinta capacidad de diferenciación: agua mineral y vino. Usando técnicas de machine learning, esta investigación proporciona una herramienta innovadora para que vendedores actuales y potenciales tomen decisiones informadas (*data-driven*) basadas en datos reales. Los hallazgos podrían extenderse a otros *marketplaces*, ofreciendo *insights* útiles para optimizar estrategias en un panorama digital competitivo.

Palabras clave: *e-commerce, machine learning, data mining*.

Índice

1.Introducción 4
2.Antecedentes del tema 7
2.1 Acerca del e-commerce y marketplaces virtuales 7
2.2 Machine learning aplicado a fines económicos 11
3. Metodología 13
3.1 Datos utilizados 13
3.2 Datos faltantes 20
3.3 Estandarización de los datos 21
3.4 Análisis y Validación de Resultados 22
4. Análisis exploratorio 24
4.1 Análisis exploratorio de la muestra total 24
4.2 Datos con títulos similares 28
4.3 Análisis exploratorio de la muestra de agua mineral 32
4.4 Análisis exploratorio de la muestra de vinos 38
5. Resultados 42
6. Conclusiones 50
7. Referencias 54

1.Introducción

Este trabajo de investigación se sumerge en el análisis de los determinantes para alcanzar el éxito en el *e-commerce*, tomando Mercado Libre como *marketplace* de referencia. Definiremos el éxito como la cantidad de visitas dentro de una categoría de productos específica, ya que entendemos al posicionamiento como factor determinante de ventas dentro de la plataforma. Sin dudas un tema que es de gran actualidad y relevancia tanto para vendedores actuales como potenciales de Mercado Libre y *marketplaces* similares.

La importancia del posicionamiento en plataformas como Mercado Libre radica en que, cuanto más arriba aparece un producto en el listado, mayor es la probabilidad de que los usuarios lo vean y, en consecuencia, lo compren. Este fenómeno se asemeja al comportamiento de los usuarios en motores de búsqueda como Google, donde la mayoría rara vez explora más allá de las primeras páginas de resultados. De hecho, el éxito de los vendedores depende de su capacidad para optimizar sus publicaciones en función de variables que el algoritmo valora. Aunque estudios previos han destacado la confianza y los precios como factores críticos de éxito en el *e-commerce,* poco se ha investigado sobre el rol crucial del posicionamiento dentro de los listados de productos, un aspecto decisivo para alcanzar el éxito en plataformas como Mercado Libre.

En el entorno actual de los *marketplaces* digitales, la competencia entre vendedores es cada vez más intensa, y el éxito de las operaciones comerciales depende en gran medida de factores no necesariamente relacionados con el producto en sí. En plataformas como Mercado Libre, el posicionamiento en los listados puede ser un factor decisivo en la rentabilidad, es por eso que la presente investigación se centra en una pregunta crucial: ¿qué variables determinan este posicionamiento y en qué medida influyen en la clasificación?

El objetivo principal de este trabajo es desentrañar los factores que el algoritmo de la plataforma prioriza, analizando cómo variables como el precio, la ubicación geográfica del vendedor, el rating, las métricas internas de la plataforma (como la salud de la publicación y los *tags* descriptivos), y la suscripción a servicios premium afectan el posicionamiento, la visibilidad y éxito de los artículos en la plataforma.

En esta investigación, el análisis se circunscribe territorialmente a Argentina, y se enfoca en datos recogidos durante el mes de septiembre de 2024. La población de estudio incluye dos grupos: por un lado, dos categorías de productos con un interés económico particular, y por otro, una muestra aleatoria de 100 categorías, lo que resulta en aproximadamente 60 mil registros. El objetivo es examinar los factores que influyen en el posicionamiento dentro de Mercado Libre, tomando como referencia los 20 artículos más vendidos en cada categoría seleccionada. Las dos categorías con interés específico se componen de: el vino y el agua embotellada (dos mercados de bienes con grados de diferenciación opuestos). La elección de estas categorías responde a la variación en el grado de homogeneidad de los productos, comparar dos mercados tan diferentes como el del vino y el agua mineral podría ofrecer una perspectiva interesante sobre cómo el algoritmo de Mercado Libre trata productos con distintos niveles de diferenciación.

Como resultado esperado de nuestra investigación, se postula que, además de las variables propuestas por Mercado Libre—tales como el título, la calidad y cantidad de fotografías, la ficha técnica, las categorías, el stock, el tipo de logística, el rating de la publicación, la reputación del vendedor, la categoría del vendedor y la publicación en catálogo—existen otras variables, como el precio y el financiamiento, que también podrían

influir en el posicionamiento de los productos en la plataforma. Esta influencia puede variar según el tipo de mercado, sugiriendo que el análisis del algoritmo de Mercado Libre debe considerar una gama más amplia de factores para comprender completamente su impacto en la clasificación de los productos.

Se prevé que los resultados de este estudio ofrezcan información respaldada en datos, útil para la toma de decisiones estratégicas para vendedores de las categorías mencionadas dentro de Mercado Libre Argentina, pudiendo dichos hallazgos ser aplicables a otros segmentos del *e-commerce*, en distintos *marketplaces* y en otros países de la región, aportando así un marco más amplio para la mejora de la competitividad en mercados digitales.

El trabajo se estructura de la siguiente manera: en el siguiente capítulo se ofrece una revisión de antecedentes. En el segundo, se describen los datos y la metodología empleada. El tercero introduce el análisis exploratorio y gráfico. En el cuarto, se presentan los resultados del machine learning. Finalmente, el último capítulo cierra con las conclusiones e implicancias económicas de la investigación.

2.Antecedentes del tema

2.1 *Acerca del e-commerce y marketplaces virtuales*

Munoz, Holsapple, y Sharath (2023) definen al *e-commerce* o comercio electrónico como "un enfoque para lograr objetivos de negocios en el que la tecnología se utiliza para manejar conocimientos y permitir la ejecución de actividades a lo largo de cadenas de valor, y mejorar la toma de decisiones que subyacen esas actividades". Hoy, el *e-commerce* efectivo requiere de sistemas mucho más complejos para complacer demandas cada vez más exigentes. Las órdenes son cada vez más pequeñas, la variedad de productos cada vez mayor, y los tiempos de envío cada vez más cortos.

Mercado Libre ha consolidado su posición como líder del comercio electrónico en América Latina, marcando un hito en la evolución del comercio digital en la región. En 2021, la empresa concentraba un 34,7% del tráfico electrónico en Latinoamérica y el Caribe, seguida por apenas un 7% de Olx (Díaz de Astarloa y Lotitto, 2023). Su plataforma ha creado un ecosistema integrado que no solo facilita la compra y venta de productos, sino que también promueve la innovación en servicios de pago y logística. Para vendedores y emprendedores, Mercado Libre ofrece un entorno que permite adaptarse a las demandas globales mediante enfoques estratégicos. Aprovechar las herramientas y recursos proporcionados por la plataforma facilita una respuesta ágil y efectiva a las tendencias del mercado, optimizando así las oportunidades y el desempeño en un panorama competitivo y en constante evolución.

El comercio electrónico global continúa su trayectoria ascendente, con proyecciones que indican que las ventas alcanzarán los 6,35 billones de dólares hacia 2027, según el *2023 Global Ecommerce Industry Report* (Benchmark International, 2023). En particular, Argentina se

destaca como uno de los mercados con mayor potencial de expansión, con un crecimiento proyectado del 14% hasta 2027. Este contexto plantea importantes oportunidades para los vendedores que operan en plataformas como Mercado Libre, que se encuentran en una posición ideal para capitalizar estas tendencias, tanto a nivel regional como global.

DeLone y McLean (2003) fueron los primeros en proponer un marco analítico para estudiar los factores de éxito en el comercio electrónico. Extiendendo su modelo de éxito a sistemas de información para aplicarlo al *e-commerce*, destacando varias dimensiones clave tales como:

- El uso: número de visitas, transacciones realizadas, y navegación dentro del sitio.
- La calidad del sistema: confiabilidad, accesibilidad, tiempo de respuesta, facilidad de uso.
- La satisfacción del usuario: desde la recolección de información hasta los servicios postventa.
- La calidad de la información: privacidad, seguridad, relevancia y completitud.

A través de una encuesta realizada en China dirigida a expertos en gestión de *e-commerce* y estudiantes de grado Pan y Chen (2010) realizan un análisis teórico y empírico con el objetivo de identificar los factores críticos que fomentan la confianza entre consumidores y sitios web de comercio electrónico. El estudio destaca varios elementos esenciales para construir relaciones de confianza tales como:

- *Funcionalidad*: Aspectos como la facilidad de uso, la navegación intuitiva y la posibilidad de interactuar con un representante humano son fundamentales para que los consumidores se sientan cómodos al realizar transacciones en línea.

- *Organizacionales*: La transparencia en la misión y los valores de la organización, junto con buenas relaciones públicas, contribuyen significativamente a la percepción de confianza por parte de los usuarios.
- *Seguridad*: La gestión segura de pagos y la protección de datos personales son vitales para que los consumidores confíen en un sitio web. La sensación de seguridad reduce la ansiedad asociada con las transacciones en línea.

Estos hallazgos son relevantes en el contexto de los *marketplaces*, ya que resaltan la importancia de lo que Kotler (2017) señala como la etapa de investigación que los consumidores atraviesan a la hora de comprar productos; pues es en esta etapa en la que clientes preguntan a su círculo, visitan foros para ver reviews y comparan precios; siendo de vital importancia trabajar en factores que pueden obstaculizar la construcción de confianza tales como: dificultades operacionales, problemas de seguridad y la falta de interacción; variables que en Mercado Libre hacen a la reputación del vendedor.

A través de un análisis exhaustivo de un amplio conjunto de literatura, Liu (2024), por su parte**,** identifica varios factores clave que influencian las decisiones de compra de consumidores en el *e-commerce*, incluyendo también la confianza, así como el precio, la calidad del producto y factores culturales, entre otros.Y si bien el autor propone que el precio se ha convertido en un factor clave debido a la intensa competencia entre vendedores, agrega que dado que los consumidores no tienen la posibilidad de probar los productos antes de realizar una compra, su confianza se basa en la reputación del vendedor y en la información disponible en la plataforma. Las descripciones detalladas, las calificaciones y las reseñas de otros compradores juegan un papel crucial en la percepción de calidad. Este contexto subraya la importancia de construir un entorno de confianza y credibilidad, elementos que son

esenciales para que los vendedores logren destacarse en un mercado cada vez más saturado y competitivo.

Gerpott y Berends (2022), a diferencia de enfoques anteriores, proponen que el precio es el factor más importante que determina el éxito en el *e-commerce*, esto debido a la feroz competencia que caracteriza estos mercados. Esta intensa competitividad es impulsada por los costos extremadamente bajos que tienen los consumidores para comparar precios a través de las plataformas en línea, lo que les permite encontrar rápidamente la oferta más ventajosa. Podemos interpretar, por lo tanto, que aquel vendedor más exitoso de una categoría dentro de un *marketplace* será aquel que ofrezca un producto de igual calidad a sus competidores pero al menor precio, siendo esta una oportunidad para que vendedores puedan beneficiarse de la implementación de mecanismos dinámicos de precios, ajustando sus estrategias en tiempo real para mantenerse competitivos.

Como se ha expuesto, no hay duda de la relevancia que tienen variables como el precio y la reputación del vendedor a la hora de tomar una decisión de compra dentro de un *marketplace*. Sin embargo, existe una notable escasez de literatura que aborde de manera exhaustiva el estudio del funcionamiento de grandes plataformas que actúan como intermediarios en el comercio electrónico, tales como Mercado Libre.

Además, se percibe una falta de análisis desde un enfoque económico sobre la importancia del posicionamiento en listados. Aunque el posicionamiento está determinado por algoritmos, en el marco de los *marketplaces* digitales este fenómeno refleja lo que la teoría de la organización industrial denomina diferenciación horizontal; pues en un entorno donde la calidad del producto puede ser homogénea, los consumidores no eligen según características intrínsecas del bien, sino según la accesibilidad y visibilidad. Así como en los mercados tradicionales un consumidor escogería al vendedor más cercano, en los *marketplaces* digitales,

el consumidor tiende a optar por el que aparece primero en los listados. En particular, el posicionamiento de productos dentro de los listados, determinado en gran parte por algoritmos, sigue siendo un área de investigación insuficientemente explorada.

2.2 Machine learning aplicado a fines económicos

Kharfan et al. (2021) hacen una predicción de demanda para la industria de la ropa y el calzado en Estados Unidos. Aplican una combinación de distintos modelos de aprendizaje automático (clustering, clasificación, entre otros) para segmentar los productos en estilos con características similares y encontrar demandas comparables a las de temporadas anteriores para el mismo grupo. De acuerdo con los autores, el estudio permite a las empresas del sector mejorar la precisión de sus estimaciones y reducir su inventario en un 40% .

Mathotaarachchi et al. (2024) analizan datos del registro público de propiedades de Inglaterra. Aplican modelos econométricos y de machine learning para predecir los precios de las propiedades en base a distintas características. Encuentran que la ubicación y el tipo de propiedad son las regresoras más importantes para explicar el precio en el conjunto de datos.

Pavlyshenko (2019) analiza el uso de modelos de aprendizaje automático para la predicción de ventas en series temporales. El estudio destaca la importancia de la generalización en machine learning, que permite realizar predicciones precisas incluso cuando se dispone de pocos datos históricos, como en el caso de nuevos productos o tiendas. Utiliza un enfoque de stacking, combinando múltiples modelos predictivos, lo que mejora la precisión de las predicciones al integrar las salidas de varios algoritmos, como Random Forest y ARIMA.

Hoang y Wiegratz (2023) discuten los beneficios de incorporar métodos de machine learning a la economía y las finanzas. Destacan la mejora en la precisión sobre la econometría tradicional, y la posibilidad de analizar variables no tradicionales como texto e imágenes. Finalmente, aplican diversos modelos para predecir el precio de las propiedades en un conjunto de más de 4 millones de registros en Alemania entre los años 2000 y 2020.

Zhang et al. (2023) hacen una revisión de literatura sobre las aplicaciones del machine learning en el *e-commerce* entre 2018 y 2023. Examinan 158 publicaciones y señalan que los objetivos más comunes son el análisis de sentimientos, los sistemas de recomendación y la detección de comentarios falsos. Además, destacan que los principales desafíos actuales en esta área incluyen los datos desbalanceados, el sobreajuste de modelos y la dificultad para generalizar resultados, así como la integración de datos procedentes de fuentes heterogéneas, entre otros.

3. Metodología

Para la investigación, se construye un conjunto de datos a partir de datos recolectados directamente de la API (*Application Programming Interface)* para desarrolladores de Mercado Libre (Mercado Libre, s.f). La empresa ofrece la posibilidad de acceder a diversos datos sobre publicaciones y vendedores a cambio de la producción de software y aplicaciones para vendedores. Para este trabajo se desarrolla una aplicación de prueba y se extraen datos para armar un conjunto de datos propio en un lapso de alrededor de 7 días durante el mes de septiembre.

El resultado luego del preprocesamiento es una base de datos de 71200 registros pertenecientes a más de 100 categorías de productos dentro de la plataforma. Además de las categorías de interés (vinos, agua mineral y libros), se seleccionan aleatoriamente categorías con menos de 4000 registros de una lista previamente recolectada, ya que Mercado Libre sólo comparte información de las primeras 4000 publicaciones por categoría. De otra forma, sería inviable analizar todos los ítems por motivos computacionales.

3.1 Datos utilizados

En base a los datos extraídos, se construye un conjunto de variables para el análisis y las predicciones. Las mismas se presentan a continuación:

- **total_visits:** variable objetivo. Visitas de la publicación desde su fecha de creación, con un período máximo de dos años.
- **total_visits_since_last_update:** Visitas de la publicación desde su última actualización.
- **api_seller_id:** identificador único del vendedor en la API.

- **price:** el precio final en pesos argentinos de cada producto.
- **original_price**: el precio antes de aplicar descuentos al producto.
- **number_of_pictures**: la cantidad de fotografías de la publicación.
- **local_pick_up:** sí permite retiro en persona.
- **free_shipping**: si ofrece envío gratis.
- **number_of_variations:** cantidad de variantes del producto.
- **number_of_tags:** cantidad de *tags* del producto.
- **positioning:** lugar que ocupa la publicación en el ranking de los 20 más vendidos (si aplica).
- **main_picture_aspect_ratio:** el ancho dividido la altura en píxeles de la imagen principal.
- **generic_specified**: si el vendedor aclara que la marca es genérica.
- **title_length:** cantidad de caracteres en el título de la publicación.
- **has_brand:** si el producto tiene marca o si, por el contrario, el vendedor no especifica o aclara que la marca es genérica.
- **health:** calidad de la publicación según Mercado Libre.
- **top_20:** si la publicación se posiciona en las 20 más vendidas de la categoría.
- **seller_total_transactions:** total de transacciones del vendedor.

- **rating_count:** cantidad de reviews de los compradores.
- **rating_one_star_count, rating_two_star_count, rating_three_star_count, rating_four_star_count, rating_five_star_count:** cantidad de *reviews* con calificaciones de 1,2,3,4 y 5 estrellas.
- **total_visits_since_last_update:** visitas totales desde la última actualización.
- **mandatory_free_shipping:** si por las características del producto, Mercado Libre exige envío gratis obligatorio.
- **lat y lng:** coordenadas geográficas de la ubicación del vendedor.
- **total_items:** items totales en la categoría.
- **warranty_days:** tiempo de garantía ofrecido en días.
- **catalog**: si la publicación es de catálogo (productos que cumplen ciertos estándares de calidad y se agrupan con otros similares).
- **is_official_store:** si el vendedor es la tienda oficial del producto.
- **quality_rating_average:** promedio de las valoraciones de los usuarios en la dimensión *calidad* del producto.
- **cost_benefit_rating_average:** promedio de las valoraciones de los usuarios en la dimensión *relación costo beneficio* del producto.
- **positive_tag_count,negative_tag_count, negative_tag_count**: cantidad de *tags* valorados como positivos, negativos y neutrales de cada producto. Esto se construye

mediante el uso de un modelo especializado para análisis de sentimiento, detallado en la sección de metodología.

- **seller_reputation_level_numeric:** reputación del vendedor según las reseñas pasadas de los usuarios. Se mide en una escala discreta que va desde 1 estrella (peor nivel de calificación) a 5 estrellas (mejor). Los vendedores reciben descuentos y privilegios según esta métrica.

En la tabla 1 se muestran algunas medidas descriptivas en referencia a cada variable.

Tabla 1: *medidas descriptivas sobre las variables numéricas.*

Nombre	Promedio	Desvío estándar	Mínimo	Mediana	Máximo
api_seller_id	325863625.25	382944037.79	10537	183897268	1987514572
price	123796.34	395504.65	100	31430	36605400
original_price	124956.07	396894.31	100	31825	36605400
number_of_pictures	5.03	6.41	1	4	279
local_pick_up	0.70	0.46	0	1	1
free_shipping	0.43	0.50	0	0	1
has_brand	0.85	0.35	0	1	1
number_of_variations	1.04	4.34	0	0	170
number_of_tags	3.98	1.26	1	4	10
health	0.77	0.12	0.33	0.77	1
positioning	9.51	5.62	1	9	20
top_20	0.01	0.08	0	0	1
seller_total_transactions	16633.43	47709.76	0	3161	919213
rating_count	3.44	34.44	0	0	4303
rating_average	1.26	2.09	0	0	5
rating_one_star_count	0.08	1.06	0	0	163
rating_two_star_count	0.05	0.59	0	0	58
rating_three_star_count	0.18	1.75	0	0	163
rating_four_star_count	0.56	5.87	0	0	755
rating_five_star_count	2.57	25.81	0	0	3180
cost_benefit_rating_average	4.53	0.40	1.80	4.63	5
total_visits	1561.47	11967.85	0	75	991473
total_visits_since_last_update	10.80	29.79	0	2	1307
mandatory_free_shipping	0.42	0.49	0	0	1
lat	-33.50	3.34	-54.81	-34.59	-22.05
lng	-59.76	2.43	-72.34	-58.54	-54.13
warranty_days	74.26	582.59	0	0	36135
catalog	0.22	0.41	0	0	1
title_length	50.02	13.09	4	53	200
is_official_store	0.11	0.32	0	0	1
generic_specified	0.03	0.16	0	0	1
main_picture_aspect_ratio	1.07	0.60	0.07	1.00	25
positive_tag_count	1.13	0.60	0	1	2
negative_tag_count	0.20	0.45	0	0	3
neutral_tag_count	2.65	0.97	1	2	8
seller_reputation_level_numeric	4.67	0.95	1.00	5.00	5

Fuente: elaboración propia en base a datos extraídos de Mercado Libre, septiembre de 2024.

Las siguientes columnas corresponden a variables categóricas:

- **title:** título de la publicación.
- **warranty_type:** si tiene garantía de fábrica, del vendedor, no aclara, o ninguna.
- **listing type**: tipo de publicación. Puede ser *free* (menor exposición), *gold_special* (exposición intermedia), o *gold_pro* (máxima exposición).
- **condition**: si el producto es nuevo o usado.
- **main_picture_size:** tamaño en pixeles de la foto principal.

- **in_house_shipping**: el tipo de envío. *Me2* significa que Mercado Libre se hace cargo. Para los envíos *me1* el vendedor se hace cargo.Finalmente, existen las categorías *custom* y *not_specified,* en casos donde no hay información.

- **logistic_type:** cómo se maneja la logística. En esta categoría se detalla si el vendedor deja el paquete en un punto designado del correo, si los productos están en un centro de distribución de Mercado Libre, o si se envía por correo directamente, entre otros.

- **seller_city:** ciudad de ubicación del producto autodeclarada por el vendedor.

- **seller_state:** provincia de ubicación del producto o Ciudad Autónoma de Buenos Aires si corresponde.

- **brand:** la marca del bien publicado.

- **status:** si la publicación está activa o si ha sido pausada.

- **date_created:** fecha de publicación inicial.

- **last_updated:** última fecha de actualización de la publicación.

- **total_visits:** visitas totales de la publicación.

- **total_visits_since_last_update:** visitas totales desde la última actualización.

- **tags:** características o valoraciones de la publicación de acuerdo con Mercado Libre.

- **power_seller_status:** si el vendedor es mercado líder. Este título tiene determinados requisitos de antigüedad, reputación, y volumen de ventas. Tiene 3 niveles distintos y otorga mayores beneficios a quienes los consiguen.

- **seller_user_type**: si el vendedor es usuario, marca o concesionaria. Esta última categoría se elimina del conjunto de datos por su baja representación (menos de 150 ejemplares).

- **category_name:** el nombre de la categoría del producto.

- **ubicacion:** si el vendedor proporcionó una ubicación exacta o aproximada. Esta variable se genera mediante la API de GeoNames (https://www.geonames.org/). En caso de no encontrar una ubicación precisa, se asigna a la fila la ubicación de la capital de su provincia.

- **region:** el grupo geográfico al que pertenece la localidad del vendedor. Se aparta la provincia de Buenos Aires por su importancia económica en la muestra. Inicialmente se intentó aplicar una clusterización mediante el algoritmo de *K-Means* con distancia de Haversine (James et al., 2023; Selvaraj y Sabarish, 2021). Sin embargo, los resultados fueron inconsistentes para distintas porciones del conjunto de datos.

Tabla 2: *medidas descriptivas sobre las variables categóricas.*

Nombre	Valores únicos	Más frecuente	Frecuencia
title	60168	Kit Cerco Eléctrico…	165
warranty_type	4	sin garantía	23777
listing_type	3	gold_special	60564
condition	2	new	63329
main_picture_size	30855	900x1200	1453
in_house_shipping	4	me2	64790
logistic_type	7	xd_drop_off	34095
seller_city	1618	Caseros	2431
seller_state	24	Buenos Aires	32728
brand	10586	No especifica	8327
status	1	active	70595
date_created	66962	2021-08-12 19:27:03	30
last_updated	62719	2024-09-12 09:54:42	47
tags	6391	good_quality_thumbnail…	20020
seller_reputation_level	5	5_green	57851
power_seller_status	4	platinum	37882
seller_user_type	2	normal	62463
category_name	69	Libros Físicos	7474
region	6	Buenos Aires	61921
ubicacion	2	exacta	68693

Fuente: elaboración propia en base a datos extraídos de Mercado Libre, septiembre de 2024.

3.2 Datos faltantes

A continuación, se exhiben las columnas que presentan datos nulos en la muestra.

Tabla 3: *valores faltantes por columna*

Nombre	Valores faltantes
main_picture_size	21
health	5064
seller_reputation_level	4632
quality_rating_average	70858
cost_benefit_rating_average	66425
seller_reputation_level_numeric	4632

Fuente: elaboración propia en base a datos extraídos de Mercado Libre, septiembre de 2024.

La columna **quality_rating_average** se elimina debido a su baja representatividad en el conjunto de datos (están asociadas con categorías muy específicas que no son de interés

particular para la investigación, como ropa y accesorios). **cost_benefit_rating_average** y **positioning** se mantienen como variables relevantes para el análisis gráfico, aunque no como predictoras para el aprendizaje automático.

Para tratar los datos faltantes en las demás columnas, se emplea un esquema de imputación mediante la mediana o la moda utilizando el algoritmo de *K-Nearest Neighbours* (Jadhav, Pramod, y Ramanathan, 2019). Este proceso, junto con la estandarización de los datos, se aplica exclusivamente a las muestras del conjunto de entrenamiento. Esta estrategia contribuye a reducir el riesgo de fugas de información y actúa como una medida adicional para mitigar el sobreajuste (Zheng y Casari, 2018).

3.3 Estandarización de los datos

Algunos modelos de machine learning, en particular aquellos que son funciones suaves de las variables independientes, pueden verse significativamente afectados por la escala de dichos atributos (Zheng y Casari, 2018). En estos casos, se torna oportuno aplicar estandarización o normalización a las variables predictoras para acotar los valores posibles de cada una, evitando un análisis menos informativo y que pondere las columnas incorrectas.

La estandarización es un componente vital de cualquier proceso de machine learning, y sin embargo rara vez es discutida como un tema en sí misma. Y aunque existen diversas variantes de este proceso, algunas son más robustas que otras (Lucas de Amorim, Cavalcanti, y Cruz, 2023).

Para el caso de las variables altamente dependientes de su categoría de producto (el precio, el tiempo de garantía ofrecido, el número de imágenes y el número de variantes), se calculan los nuevos valores a partir de la categoría y no de todo el conjunto de datos. En

particular, se utiliza el *Robust Scaler*, que aplica la siguiente transformación afín a cada observación:

$$x_i^{'} = \frac{x_i - Me(x)}{Q_3(x) - Q_1(x)},$$

siendo $Me(x)$ la mediana de las observaciones en x y $Q_3(x) - Q_1(x)$ el rango intercuartil de la columna. Al dividir por el rango intercuartil se mitigan los efectos de valores extremos y se consigue una estandarización con varianzas más homogéneas entre cada atributo.

3.4 Análisis y Validación de Resultados

Con el fin de estudiar las relaciones entre variables, se emplean dos enfoques diferentes. Por un lado, para probar la significancia estadística de las conclusiones derivadas del análisis exploratorio, se aplican pruebas de hipótesis sobre cada grupo de interés. En cuanto al análisis de los efectos de las variables sobre la regresada, se ajustan modelos de aprendizaje automático a los datos y se extraen conclusiones a partir de cómo estos hacen uso de cada variable predictora.

Para las pruebas estadísticas, se opta por un tratamiento de permutación. Aunque su aplicabilidad se puede ver limitada en presencia de numerosas covariables en problemas de predicción, diversos autores prefieren los tests de permutación sobre su contrapartida paramétrica. Esto es así porque requieren muy pocos supuestos sobre la distribución probabilística de los datos, y no sufren demasiada pérdida de eficiencia en el proceso de inferencia (Pesarin y Salmaso, 2010).

Específicamente, se aplican tests de diferencias de medias y análisis de varianza cuando las poblaciones están divididas de acuerdo con variables categóricas. Para el análisis de las variables numéricas se llevan a cabo permutaciones y se anota la frecuencia con la que se producen valores de correlaciones igual de extremos o más extremos que el observado.

Respecto a los métodos de machine learning, se entrenan tanto modelos lineales con y sin regularización (mínimos cuadrados ordinarios, Lasso, Ridge y Elastic Net) como modelos no paramétricos basados en árboles de decisión (CART y Random Forest) (Chan y Mátyás, 2022; Chan, Harris, Singh y Yeo, 2022). Al construir los modelos con regularización, se consideran las interacciones lineales entre columnas como candidatas a variables explicativas.

En todos los casos, se ajustan los estimadores mediante una búsqueda exhaustiva de parámetros y empleando validación cruzada para evitar el sobreajuste a los datos (James et al., 2023). Se decide omitir aquellas variables predictoras que el vendedor no controla directamente y que dependen de los propios consumidores, como el promedio de calificaciones de la publicación y la reputación del vendedor.

La evaluación de la capacidad predictiva de los modelos se desarrolla calculando el error absoluto medio y la raíz del error cuadrático medio (*MAE* y *RMSE* por sus siglas en inglés respectivamente), y el coeficiente de determinación R^2 para el estimador de mínimos cuadrados. Como punto de referencia, se toma un modelo base que genera la mediana de los datos de entrenamiento como respuesta para todas las observaciones. Esta toma el valor de 0.000047 en la muestra de entrenamiento.

Finalmente, se examinan los efectos de las regresoras sobre las predicciones de los modelos mediante técnicas de *feature importance* gráficos de dependencia parcial (Molnar, 2019;Chan, Harris, Singh y Yeo, 2022).

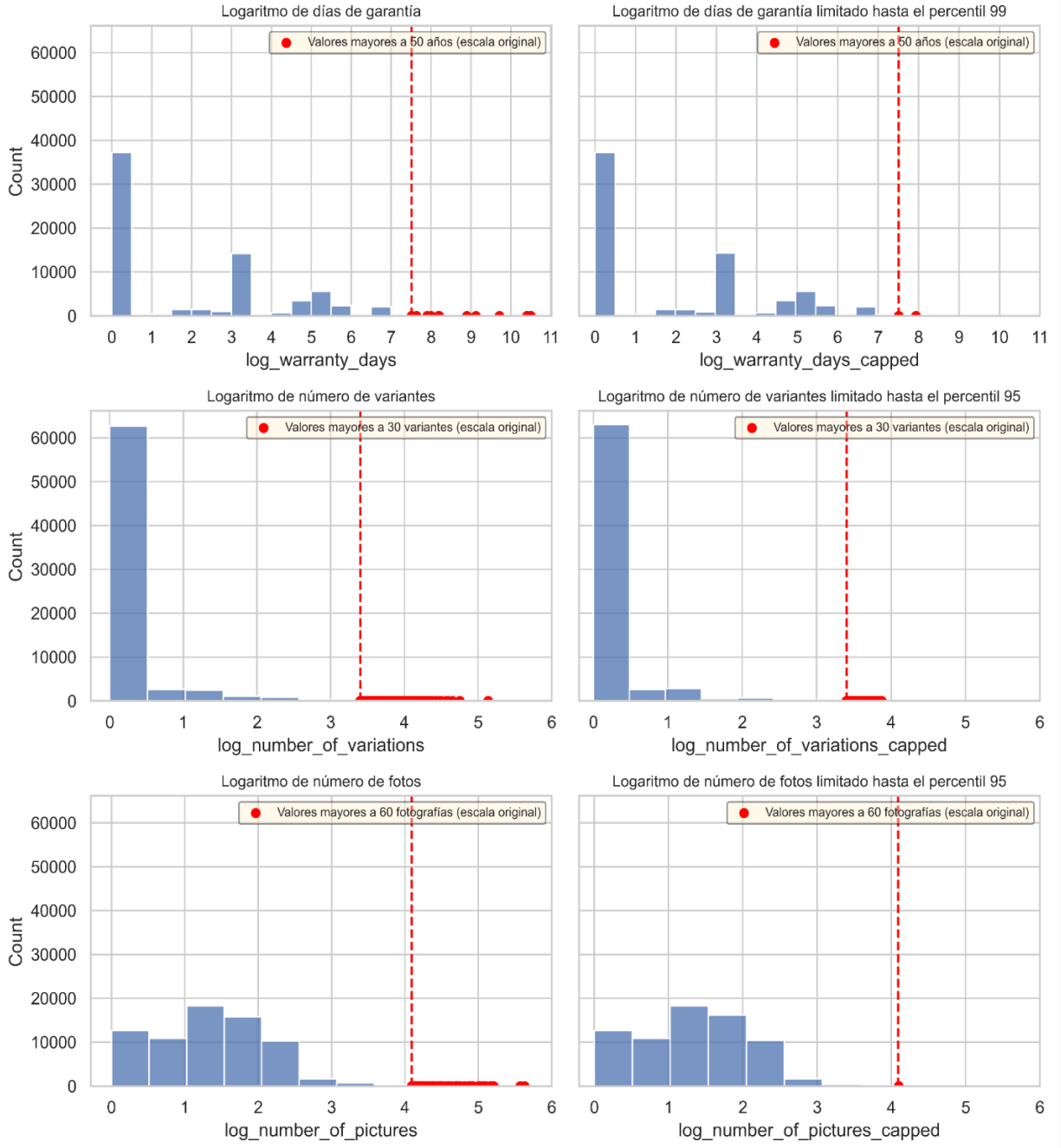
4. Análisis exploratorio

4.1 Análisis exploratorio de la muestra total

4.1.1 Valores extremos

En primer lugar, en cuanto a los días de garantía ofrecidos por los vendedores, se encuentra un pequeño grupo de publicaciones que ofrece 99 o 50 años de garantía, mientras que el promedio en sus respectivas categorías no supera el año. Para reducir el sesgo que estos valores extremos generan en la distribución de cada columna, se establece un límite en el percentil 99 de cada categoría y se crea una nueva variable dicotómica para capturar la presencia de garantías excepcionalmente largas. Del mismo modo, se aplican topes al número de fotografías por publicación y al número de variantes de cada producto para manejar estos valores atípicos.

Figura 1: *variables antes y después de limitar los valores. Se agrupan los negativos con el 0 para poder visualizar estos valores en escala logarítmica*

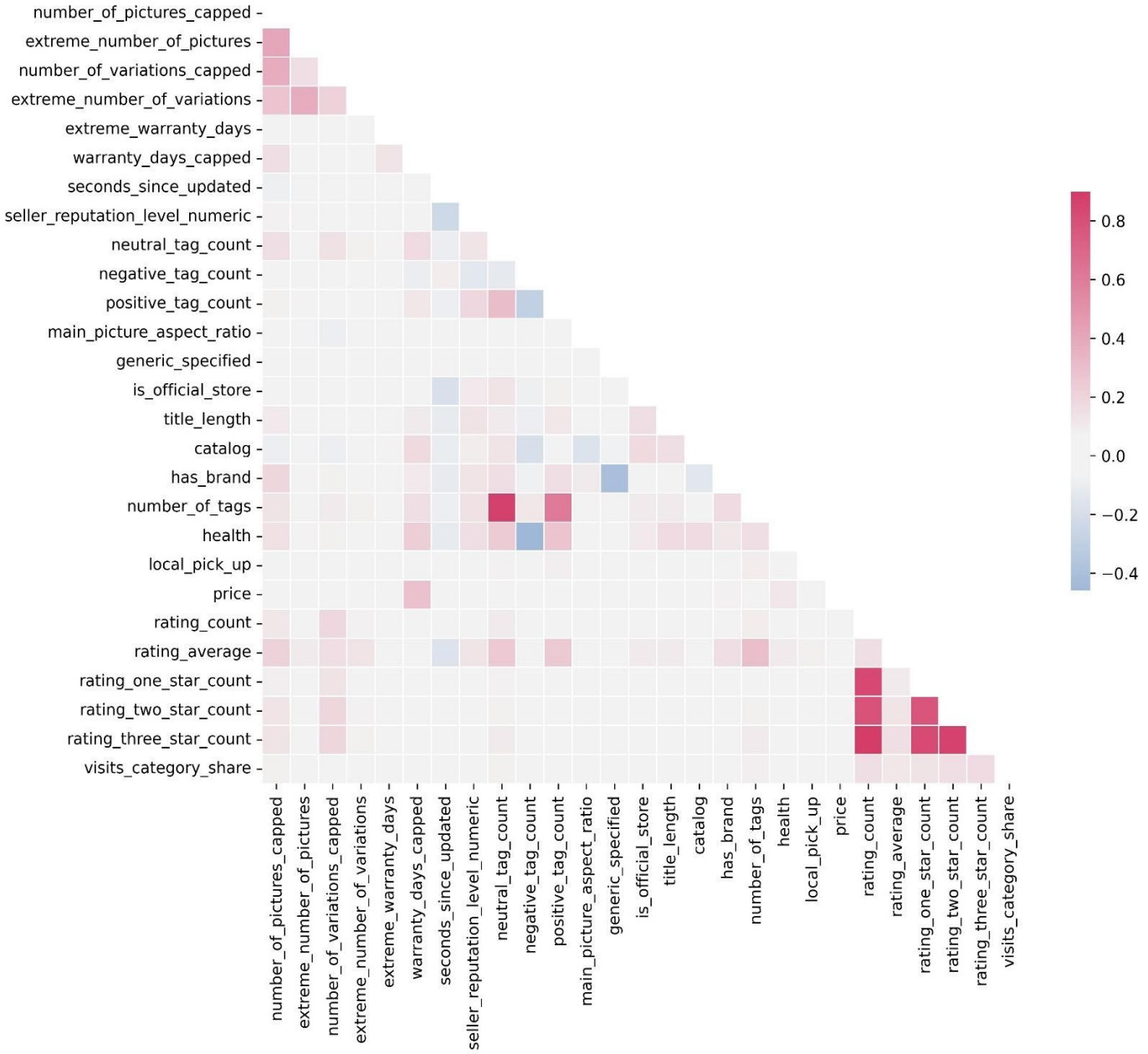


Fuente: datos extraídos de Mercado Libre, septiembre de 2024.

4.1.2 Análisis de correlaciones

El siguiente mapa de calor presenta las correlaciones entre las variables analizadas en el conjunto de datos:

Figura 2: *correlaciones entre variables numéricas*



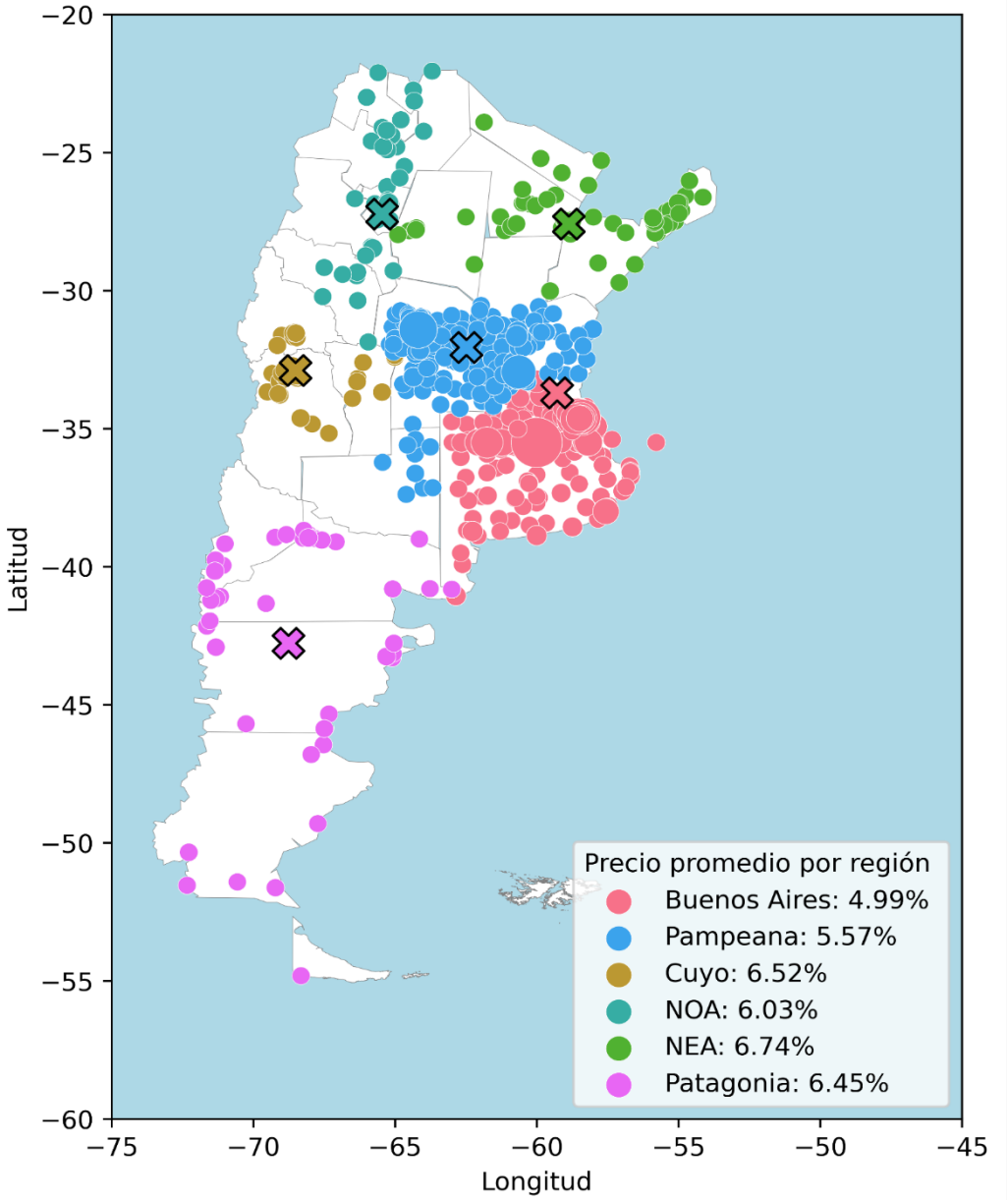
Fuente: elaboración propia en base a datos extraídos de Mercado Libre, septiembre de 2024.

Pudiendo extraerse las siguientes conclusiones principales:

- Existe una correlación positiva fuerte entre las variables relacionadas con las calificaciones (***rating_one_star_count***, ***rating_two_star_count***, ***rating_three_star_count***, etc.), lo cual es esperable, ya que estas variables se refieren a categorías similares que suelen comportarse de manera consistente.
- La variable ***visits_category_share*** muestra correlaciones positivas con algunas métricas relacionadas con las calificaciones promedio (***rating_average***) y el ***rating_count***, lo que podría sugerir que las publicaciones con mejor desempeño en calificaciones tienen una mayor proporción de visitas dentro de su categoría.
- No parece haber una correlación fuerte entre el precio y el ***visits_category_share***, lo que podría indicar que el precio no es un determinante directo de las visitas o calificaciones, al menos en este análisis.

En la Figura 3 se muestra la distribución geográfica de los vendedores y los promedios de los precios como porcentaje del máximo en cada categoría. Los precios en relación al máximo de la categoría son en promedio inferiores en la región de Buenos Aires. Sin embargo, ajustando una regresión lineal a los datos el coeficiente de determinación es casi nulo, indicando que los cambios en el precio no son explicados por la región o las distancias a los centros de cada una. Se decide no avanzar con este análisis.

Figura 3: *análisis geográfico y de precios relativos*



Fuente: elaboración propia en base a datos extraídos de Mercado Libre, septiembre de 2024.

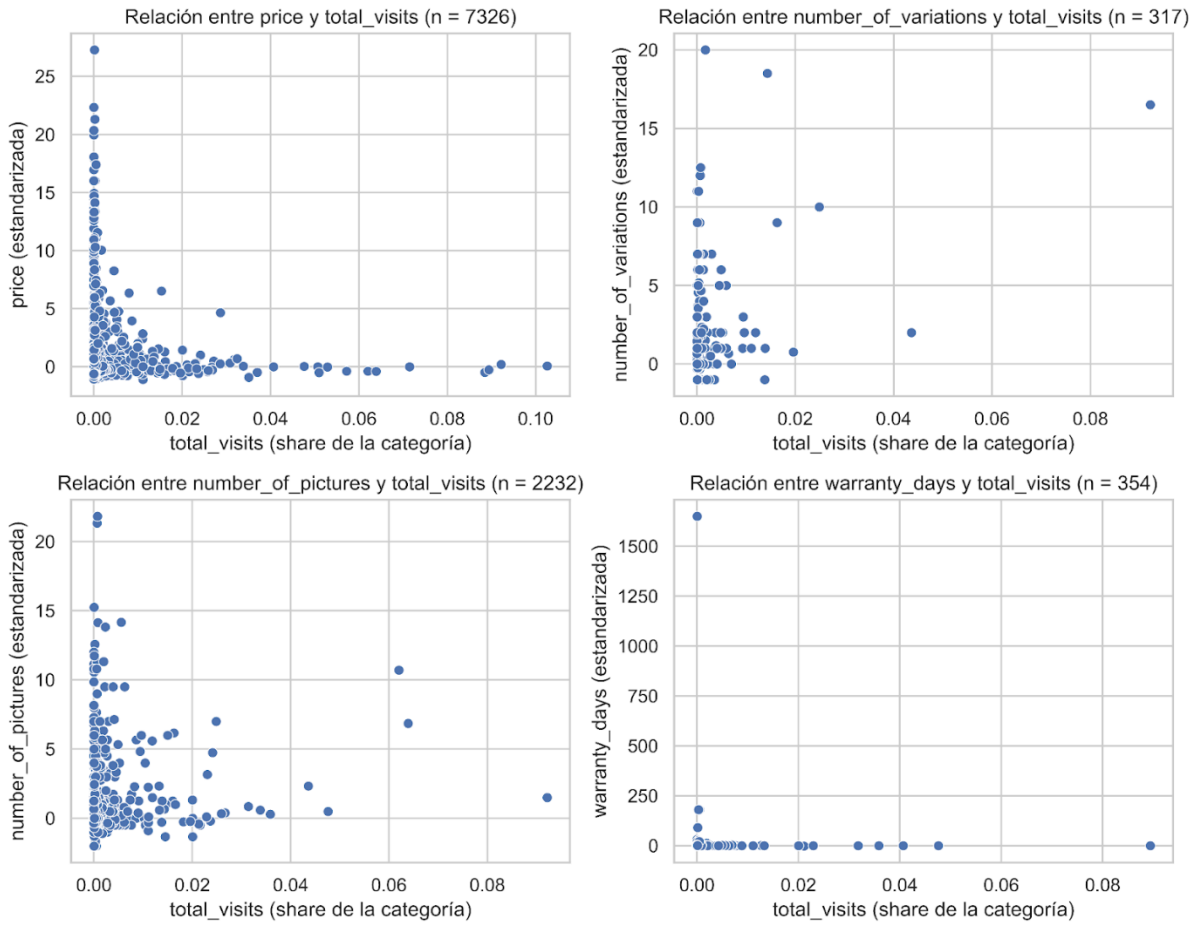
4.2 Datos con títulos similares

A través de una inspección exhaustiva de la muestra, se identifica un subconjunto de publicaciones duplicadas por los mismos vendedores, que comparten títulos idénticos pero con variaciones en otros atributos que el vendedor controla directamente. En las figuras a

continuación, se presentan los resultados de este grupo de publicaciones, focalizándose en aquellas variables susceptibles de ser modificadas por los vendedores. En otras palabras, se consideran publicaciones con títulos equivalentes, pero en las que el atributo específico bajo análisis varía (se excluyen las observaciones repetidas pero con un único valor de cada variable bajo estudio).

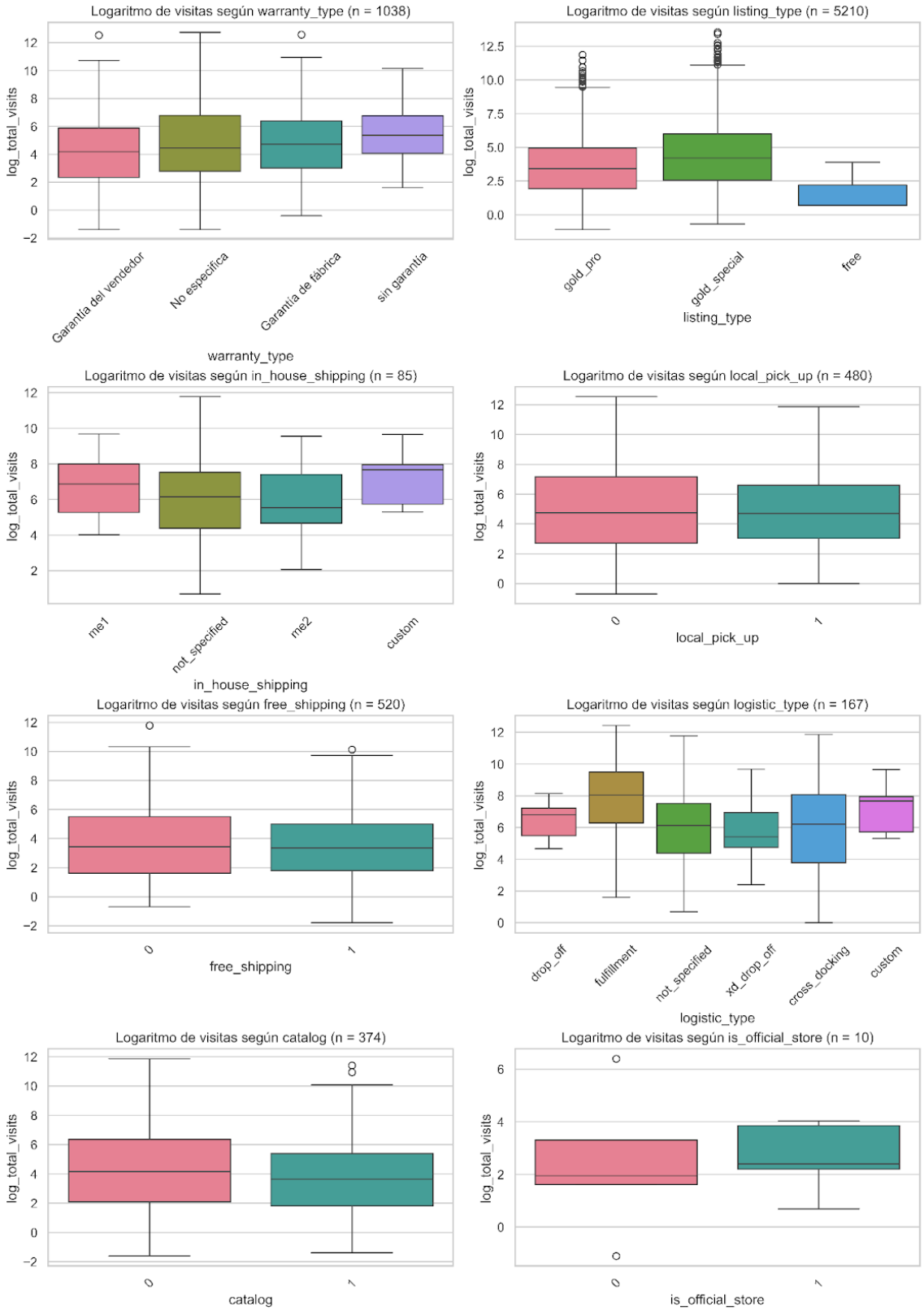
Mediante este proceso, se busca estimar el impacto directo que tienen los cambios en una variable independiente sobre la variable objetivo (total de visitas). No obstante, la disponibilidad limitada de datos impide realizar un análisis aislado de los efectos individuales de cada variable, manteniendo constantes las demás variables predictoras.

Figura 4: *variables numéricas y total de visitas en datos con títulos idénticos*



Fuente: elaboración propia en base a datos extraídos de Mercado Libre, septiembre de 2024.

Figura 5: *variables categóricas y total de visitas en datos con títulos idénticos*



Fuente: elaboración propia en base a datos extraídos de Mercado Libre, septiembre de 2024.

Gráficamente no se evidencian grandes diferencias entre las distribuciones para cada variable. Para poder hacer inferencia, se aplican tests de permutación sobre cada grupo de interés. Aunque su aplicabilidad se puede ver limitada en presencia de numerosas covariables en problemas de predicción, muchos autores los prefieren sobre su contrapartida paramétrica. Esto es así porque requieren muy pocos supuestos sobre la distribución probabilística de los datos, y no sufren demasiada pérdida de eficiencia en el proceso de inferencia (Pesarin y Salmaso, 2010).

Para evaluar los resultados se aplican tests de diferencias de media y análisis de varianza cuando las poblaciones están divididas de acuerdo con variables categóricas. Para el análisis de las variables numéricas se llevan a cabo permutaciones y se anota la frecuencia con la que se producen valores igual de extremos o más extremos que el observado. Los resultados se muestran en la tabla a continuación:

Tabla 4: *resultados de las pruebas en la muestra con datos repetidos..*

Tipo de test	Columna	Valor observado	p-valor	Corrección de Holm-Bonferroni
Correlación	price	-0.02	0.07	1
Correlación	number_of_variations	0.14	0.03	0.58
Correlación	number_of_pictures	0.20	0.00	0
Correlación	warranty_days	-0.03	0.12	1
ANOVA	warranty_type	2112194.90	0.35	1
ANOVA	listing_type	2096849.92	0.06	1
ANOVA	in_house_shipping	2617718.03	0.84	1
Diferencia de medias	local_pick_up	1088.31	0.27	1
Diferencia de medias	free_shipping	730.54	0.10	1
ANOVA	logistic_type	27633278.16	0.27	1
Diferencia de medias	catalog	546.54	0.32	1
Diferencia de medias	listing_type_agrupado	45.70	0.03	0.58
Diferencia de medias	is_official_store	102.47	0.43	1

Fuente: elaboración propia en base a datos extraídos de Mercado Libre, septiembre de 2024.

Los tests de permutación de análisis de varianza y diferencia de medias revelan que las diferencias entre cada grupo no son estadísticamente significativas para un nivel de tolerancia del error tipo I del 5%. Para controlar la *Family-wise Error Rate* (probabilidad de cometer al

menos un error tipo I) entre los p-valores, se aplica la corrección de Holm-Bonferroni (James et al., 2023). Según los resultados de este procedimiento, únicamente el número de fotografías está correlacionado con la cantidad de visitas de la publicación.

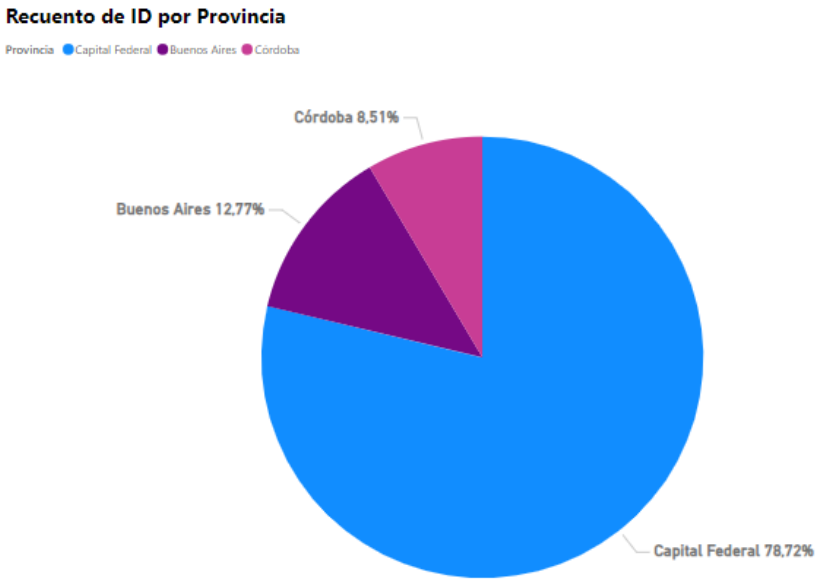
4.3 Análisis exploratorio de la muestra de agua mineral

A continuación se presentan los resultados obtenidos a partir del análisis exploratorio de la muestra de 'aguas', centrado en la categoría de agua mineral (500ml) en la plataforma Mercado Libre; para luego comparar estos resultados con los correspondientes al mercado del vino y así determinar si el algoritmo de Mercado Libre ajusta sus criterios de posicionamiento en función de las características y el grado de diferenciación propios de cada mercado.

El gráfico a continuación muestra que las ventas de agua mineral en Mercado Libre están concentradas en vendedores de solo tres provincias, predominando Buenos Aires y Capital Federal. Esta distribución sugiere que el producto, no siendo típicamente comercializado por *e-commerce*, tiene mayor demanda en zonas urbanas con alta densidad de población y servicios de entrega rápida. La presencia en estas áreas de grandes aglomeraciones permite despachos más

veloces, incluso en el día, lo que probablemente incentive la compra en estas localidades.

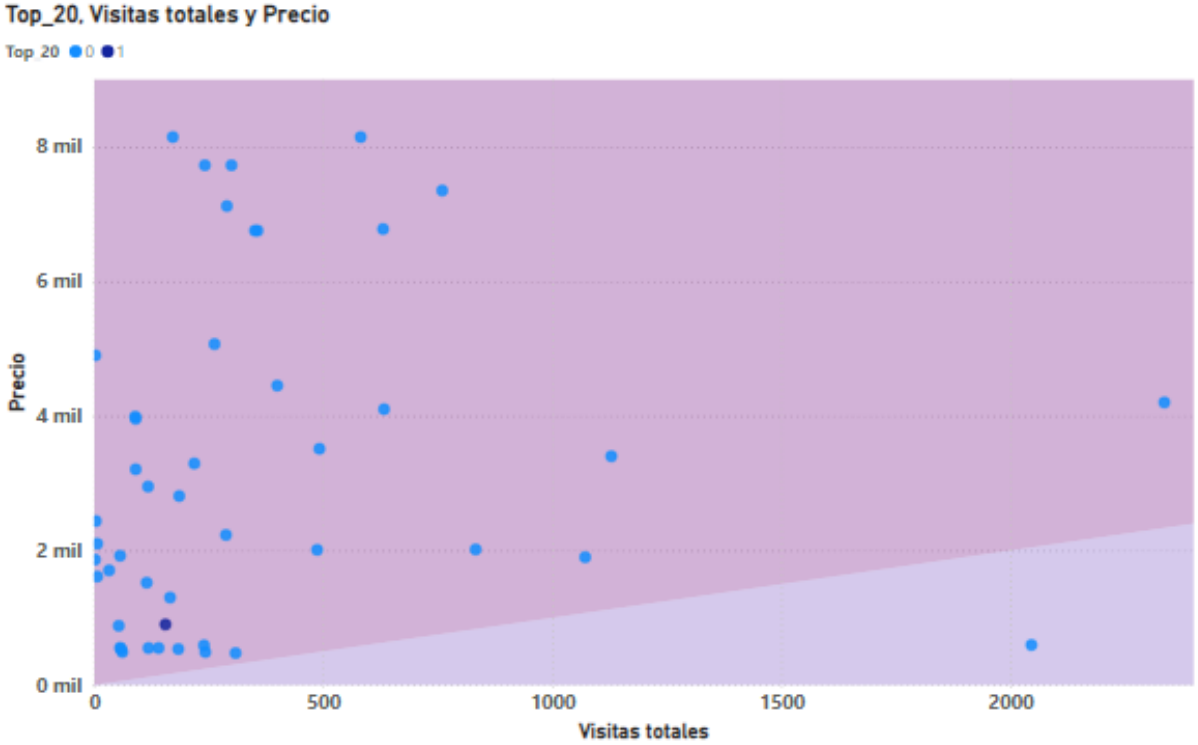
Figura 6: *vendedores por provincia en la categoría de agua mineral.*



Fuente: elaboración propia en base a datos extraídos de Mercado Libre, septiembre de 2024.

El siguiente gráfico explora uno de los ejes centrales de esta investigación: determinar si el precio tiene una influencia significativa en la cantidad de visitas dentro de la categoría de agua mineral, un mercado caracterizado por su baja capacidad de diferenciación. Aquí, se analiza el comportamiento de las visitas como una posible consecuencia del posicionamiento alcanzado, y este, a su vez, podría estar condicionado en parte por precios competitivos, es decir, más bajos en comparación con la competencia. Así, el análisis busca evidenciar si existe una correlación entre un buen precio y una mayor visibilidad en la plataforma.

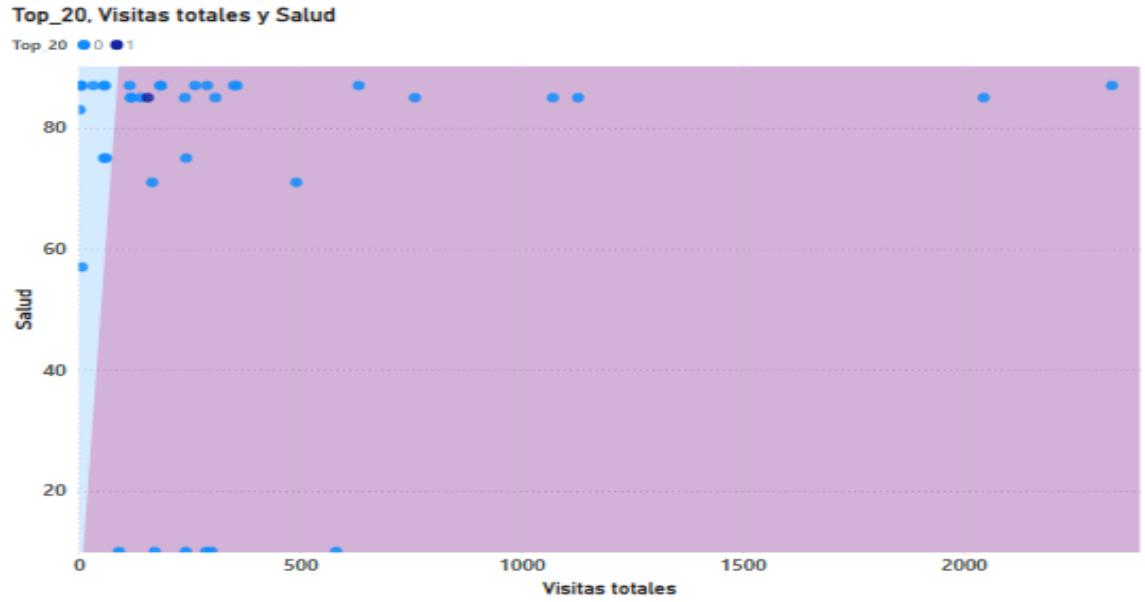
Figura 7: *relación entre las visitas y el precio en la categoría de aguas minerales.*



Fuente: elaboración propia en base a datos extraídos de Mercado Libre, septiembre de 2024.

El gráfico de dispersión muestra que, contrario a la expectativa de una relación inversa clara entre precio y visitas, no se observa una correlación tan definida; es decir, los precios más bajos no siempre resultan en un mayor volumen de visitas. Este patrón podría explicarse, en primer lugar, porque el agua mineral es un bien típicamente de bajo costo, lo cual limita el rango de precios, y, además, al no ser un producto comúnmente comercializado por *e-commerce*, cuenta con un nicho de visitas relativamente reducido. Estas características sugieren que factores adicionales, más allá del precio, podrían estar influyendo en el comportamiento de los usuarios en esta categoría.

Figura 8: *relación entre las visitas y la salud de la publicación en la categoría de aguas minerales.*



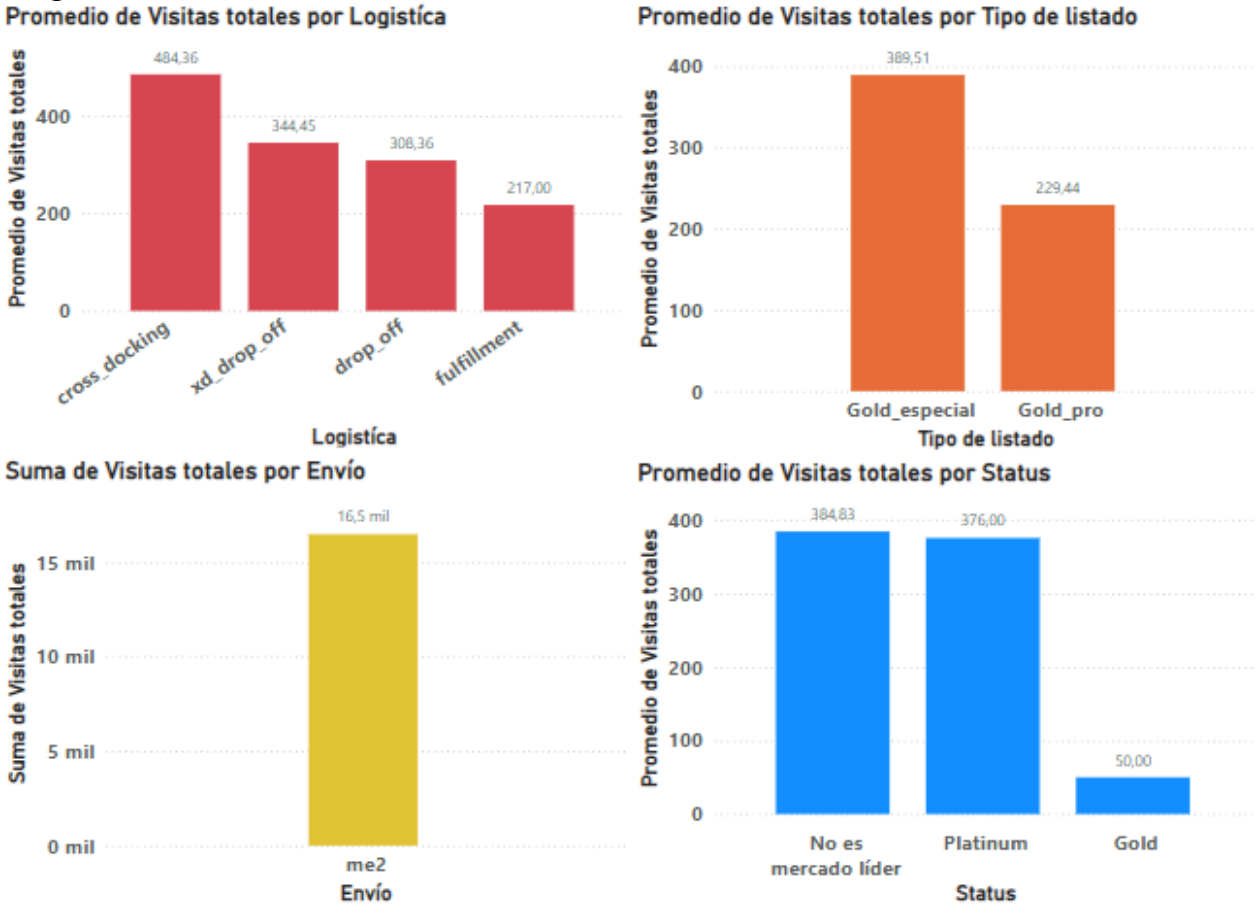
Fuente: elaboración propia en base a datos extraídos de Mercado Libre, septiembre de 2024.

El análisis del siguiente gráfico de dispersión sugiere que una alta "salud" de la publicación —una métrica subjetiva puntuada por Mercado Libre— no es una condición suficiente para asegurar un alto tráfico de visitas. Se observan casos en los que publicaciones con una salud baja logran alcanzar niveles de visitas comparables, e incluso superiores, a aquellas con puntuaciones de salud significativamente más altas. Esto indica que, si bien la salud de la publicación podría contribuir al posicionamiento, nuevamente, no es el único factor determinante en la atracción de visitas.

De los siguientes gráficos podemos extraer algunas conclusiones interesantes:

La primera es que la totalidad de las publicaciones de la muestra tienen envío me2, esto es, que el envío corre a cuenta de Mercado Libre.

Figura 9: *promedios de visitas según cada categoría para distintas variables en la muestra de aguas minerales.*



Fuente: elaboración propia en base a datos extraídos de Mercado Libre, septiembre de 2024.

La segunda es que, en promedio, las publicaciones **gold_special** tuvieron más visitas que aquellas **gold_pro** (una suscripción más costosa, que ofrece mayor visibilidad en listados).

La tercera es que, las publicaciones de vendedores que no forman parte de la categoría Mercado Lider tuvieron en promedio casi la misma cantidad de visitas (incluso más) que aquellas Platinum; superando ampliamente a las publicaciones de vendedores en la categoría gold (inferior a platinum).

La cuarta es que desglosando la manera en la que se envían los productos (a cargo de mercado Libre), vemos que la mayoría de las visitas en promedio se las llevan aquellas

publicaciones con logística cross-docking, es decir, que mercado libre recolecta el paquete y lo envía (el vendedor no tiene que dejarlos en puntos de entrega ni en ningún correo, por lo que la responsabilidad de mercado libre en el envío es mayor). Los tests de permutación revelan que efectivamente existen diferencias entre los grupos de publicaciones según el estatus del vendedor y el tipo de logística que se ofrece con cada producto:

Tabla 5*: resultados de las pruebas en la muestra de agua mineral.*

Tipo de test	Columna	Valor observado	p-valor	Corrección de Holm-Bonferroni
ANOVA	power_seller_status	36411.23	0	0
Diferencia de medias	listing_type	160.07	0.50	1
ANOVA	logistic_type	12327.35	0	0
Correlación	price	0.12	0.44	1
Correlación	health	0.06	0.69	1

Fuente: elaboración propia en base a datos extraídos de Mercado Libre, septiembre de 2024.

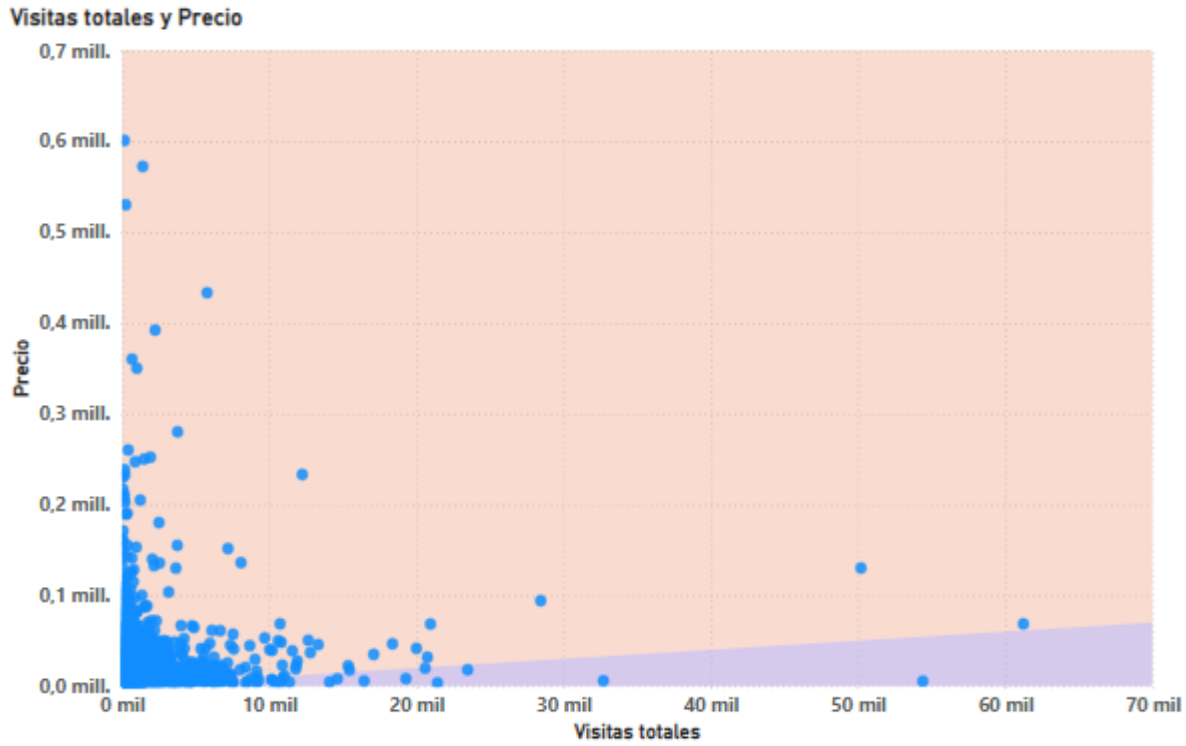
Con base en los resultados obtenidos en las pruebas estadísticas aplicadas a las variables, se puede determinar la significatividad estadística de los *insights* previamente discutidos. Según el cuadro de resultados, el análisis de varianza (ANOVA) aplicado a las variables ***power_seller_status*** y ***logistic_type*** muestra diferencias significativas entre los grupos (pv=0 en ambos casos, incluso después de la corrección de Holm-Bonferroni). Esto valida la hipótesis de que el estatus del vendedor y el tipo de logística tienen un impacto relevante en las visitas de las publicaciones.

Por otro lado, las pruebas de correlación realizadas sobre las variables ***price*** y ***health*** no revelaron una relación significativa con las visitas ($p-value = 0.44$ y $p-value = 0.69$ respectivamente). Finalmente, el test de diferencia de medias aplicado a *listing_type* tampoco arrojó diferencias estadísticamente significativas entre los tipos de publicaciones ($p-value = 0.50$). Estos resultados indican que, si bien algunas variables tienen un impacto claro, otras no presentan suficiente evidencia estadística para afirmar una relación significativa con las visitas.

4.4 Análisis exploratorio de la muestra de vinos

A partir del análisis exploratorio de la muestra de 'vinos' (botella de 750ml) dentro de la plataforma de Mercado Libre; podemos observar que tampoco parece haber una relación directa entre precio y visitas; aunque sí podríamos decir que un precio bajo podría ser considerado condición necesaria (pero no suficiente) para tener una buena performance en visitas (pues todos los puntos correspondientes a precios altos se corresponden con bajas visitas), una conclusión que resulta intuitiva y que podría extenderse a mercados distintos, pero que se evidencia claramente en un mercado con una elevada gama de calidades y precios como es el del vino.

Figura 10: *relación entre las visitas y el precio en la muestra de vinos.*

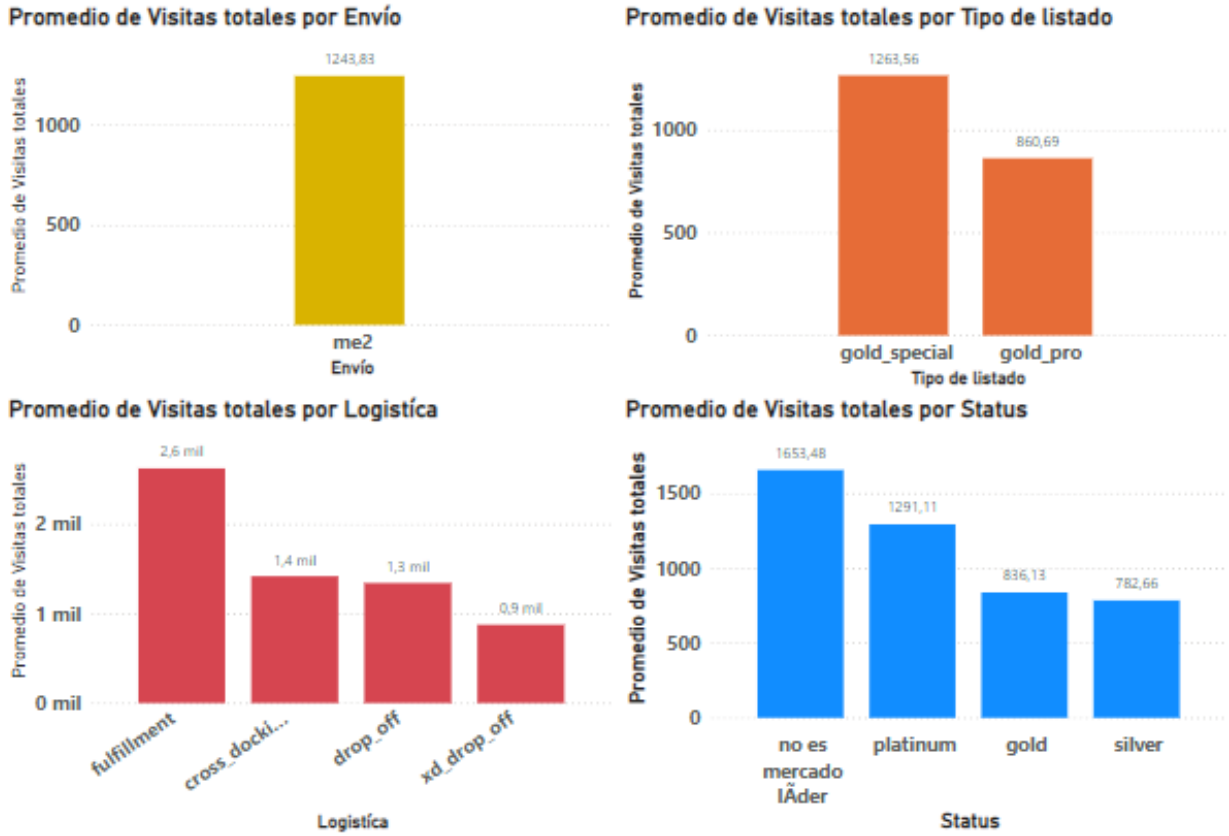


Fuente: elaboración propia en base a datos extraídos de Mercado Libre, septiembre de 2024.

Continuando con la misma estructura de análisis realizado en el mercado de aguas minerales, vemos que en este mercado tampoco parece haber una relación directa entre salud de la publicación y visitas; podríamos decir incluso que los gráficos son bastante similares en

ambos casos. Se podría inferir nuevamente que esto se debe a que si bien la salud de una publicación influye en sus visitas, hay otros numerosos factores que influyen también en esta variable.

Figura 11: *promedios de visitas según cada categoría para distintas variables en la muestra de vinos.*

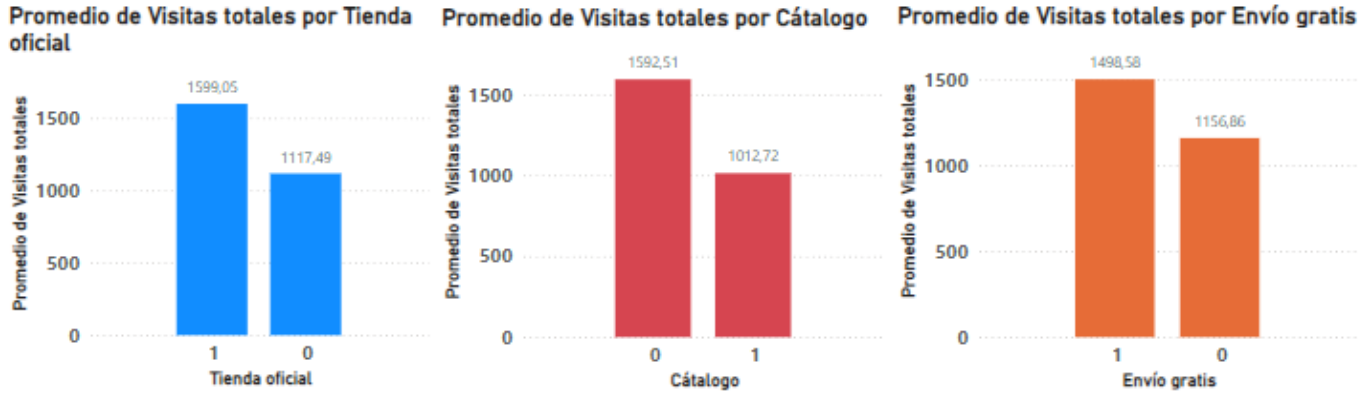


Fuente: elaboración propia en base a datos extraídos de Mercado Libre, septiembre de 2024.

Siguiendo la línea de análisis, vemos que las métricas promedio de visitas por tipo de listado, por status y por logística siguen teniendo comportamientos y relaciones similares a las del mercado del agua mineral; quizás siendo la mayor diferencia que en este mercado lidera el promedio de visitas la logística fullfilment (los productos se encuentran en un centro de distribución). También se puede apreciar que nuevamente todos los envíos para las publicaciones de este mercado son me2 (es decir, a cargo de Mercado Libre).

También vemos que no parece haber tampoco una relación directa y lineal entre niveles de reputación de vendedores y visitas en promedio; ni tampoco parece haberlo con las publicaciones en catálogo (las visitas promedio de las publicaciones en catálogo son considerablemente menores que aquellas no publicadas en catálogo).

Figura 12: *promedios de visitas según el valor de distintas variables dicotómicas en la muestra de vinos.*



Fuente: elaboración propia en base a datos extraídos de Mercado Libre, septiembre de 2024.

Por otro lado, sí parece haber influencia en las views promedio por parte de variables como envio gratís (superiores a aquellas que no ofrecen este servicio) y tiendas oficiales (sus publicaciones tienen más view en promedio, podría resultar intuitivo ya que inspiran mayor confianza a consumidores, incluso podrían hasta fidelizar clientes).

En cuanto a los tests de permutación, las variables significativas son nuevamente el estatus del vendedor y el tipo de logística con que se manejan los productos:

Tabla 6*: Resultados de las pruebas en la muestra de vinos.*

Tipo de test	Columna	Valor observado	p-valor	Corrección de Holm-Bonferroni
Correlación	price	0.01	1	1
Correlación	health	0.00	0.96	1
ANOVA	power_seller_status	168233.32	0	0
Diferencia de medias	listing_type	397.08	0.50	1
ANOVA	logistic_type	555845.80	0	0
Diferencia de medias	free_shipping	346.04	0.51	1
Diferencia de medias	catalog	578.20	1	1
Diferencia de medias	is_official_store	477.57	0.49	1

Fuente: elaboración propia en base a datos extraídos de Mercado Libre, septiembre de 2024.

En base a los resultados de las pruebas estadísticas aplicadas, se buscará validar la significatividad de los *insights* mencionados previamente. Los resultados obtenidos, reflejados en la tabla, indican que las variables *power_seller_status* y *logistic_type* presentan diferencias estadísticamente significativas (pv=0) incluso tras la corrección de Holm-Bonferroni, lo que refuerza la relevancia de estas variables en el comportamiento de las visitas promedio.

Por el contrario, variables como *price, health, free_shipping, catalog, is_official_store, y listing_type* no muestran una relación significativa con las visitas después de ajustar por la corrección, lo que sugiere que no existe evidencia suficiente para afirmar que estas variables tienen un impacto directo en el rendimiento de las publicaciones.

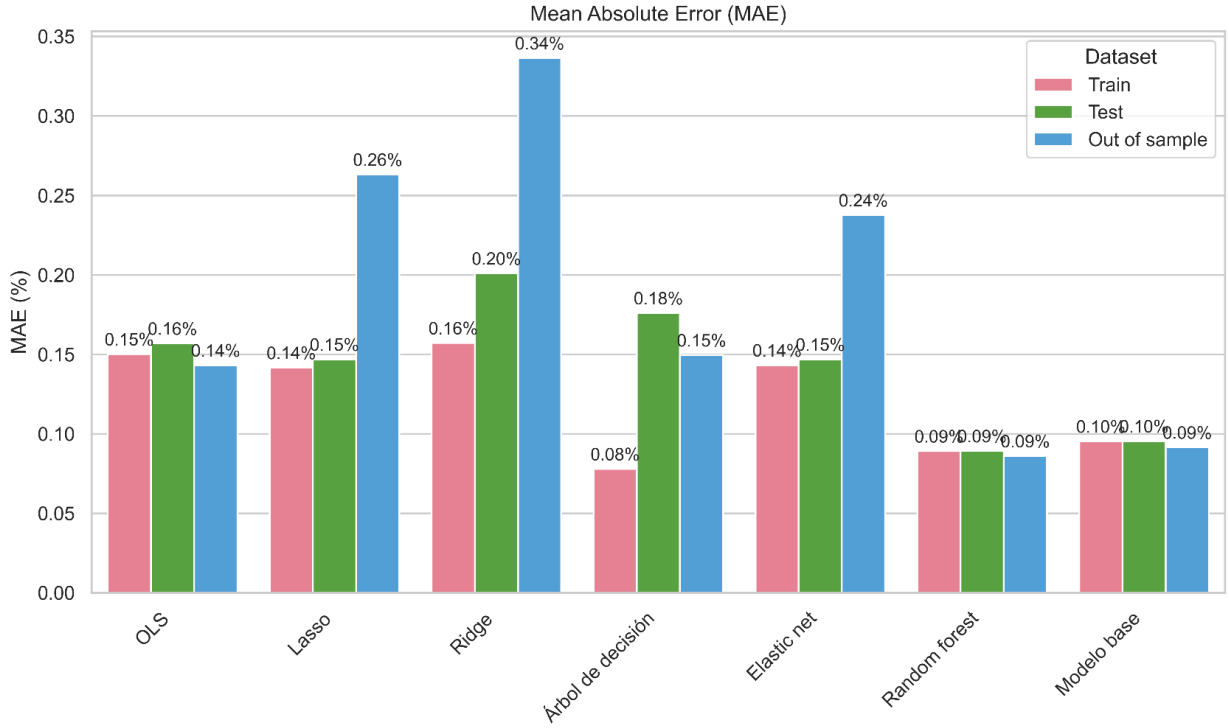
Estos hallazgos permiten confirmar que, si bien algunas características como el estatus del vendedor y el tipo de logística desempeñan un rol crucial, otras variables no parecen tener un efecto significativo en las visitas promedio, lo que coincide parcialmente con las conclusiones intuitivas del análisis exploratorio.

5. Resultados

A pesar de su rendimiento consistente en distintas muestras, la regresión lineal simple queda significativamente por debajo del modelo base en términos de precisión. Esta enfrenta importantes limitaciones, como heterocedasticidad, multicolinealidad y una distribución muy sesgada de los residuos (Gujarati, 2009), lo que compromete tanto su robustez como la interpretación confiable de los coeficientes. Además, el R^2 es extremadamente bajo (0.025), indicando una bondad de ajuste mínima. Debido a estos problemas, se descarta el uso del modelo lineal simple tanto para predicciones como para análisis descriptivo.

Los modelos con regularización tienen un desempeño aún peor, en especial considerando predicciones fuera de la muestra. El árbol de decisión, aunque apropiadamente podado, parece presentar un sobreajuste a los datos de entrenamiento de validación cruzada, y su desempeño no alcanza al del modelo base. Finalmente, el predictor random forest resulta superior a todos los demás, y es el único que logra superar la precisión del modelo base (aunque por un margen pequeño). Los errores de entrenamiento, validación y predicción fuera de la muestra de cada estimador se presentan a continuación:

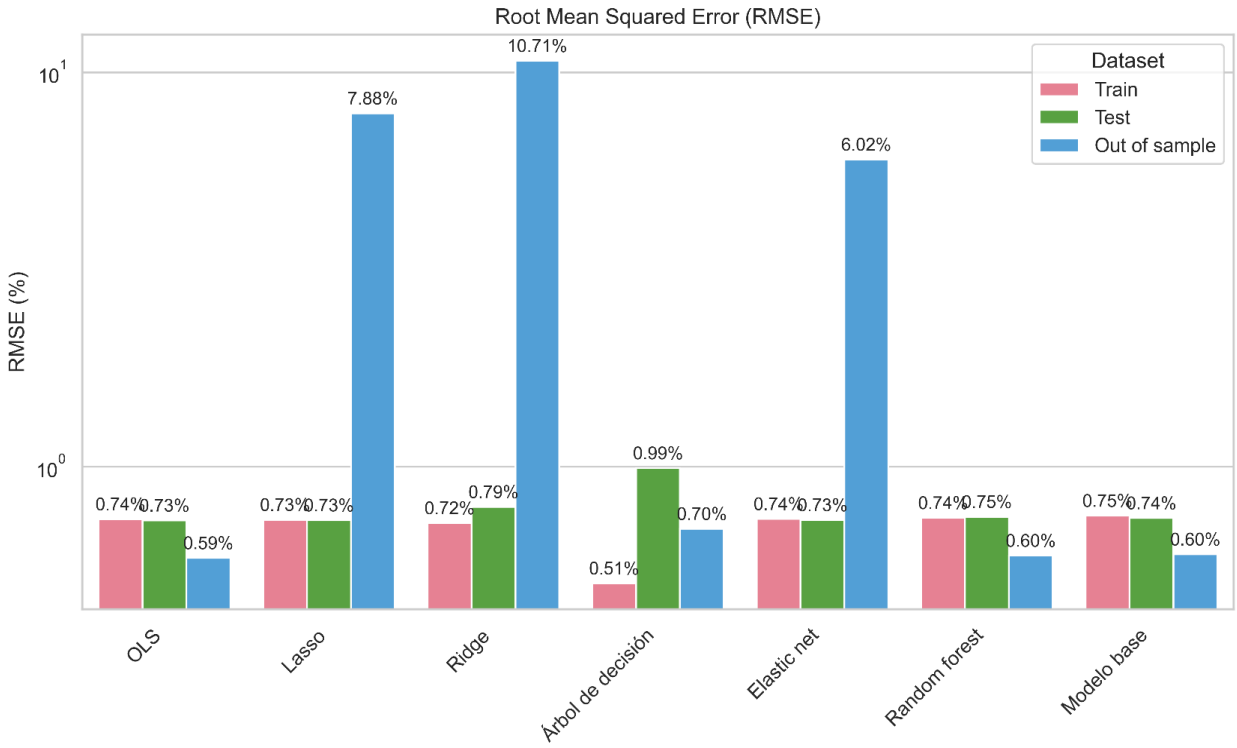
Figura 13: *error absoluto medio para distintos modelos según el conjunto de datos.*



Fuente: resultados propios en base a datos extraídos de Mercado Libre, septiembre de 2024.

Los modelos con regularización tienen un desempeño aún peor, en especial considerando predicciones fuera de la muestra. El árbol de decisión, aunque apropiadamente podado, parece presentar un sobreajuste a los datos de entrenamiento de validación cruzada, y su desempeño no alcanza al del modelo base. Finalmente, el predictor random forest resulta superior a todos los demás, y es el único que logra superar la precisión del modelo base (aunque por un margen pequeño). Los errores de entrenamiento, validación y predicción fuera de la muestra de cada estimador se presentan a continuación:

Figura 14: *raíz del error cuadrático medio para distintos modelos según el conjunto de datos.*

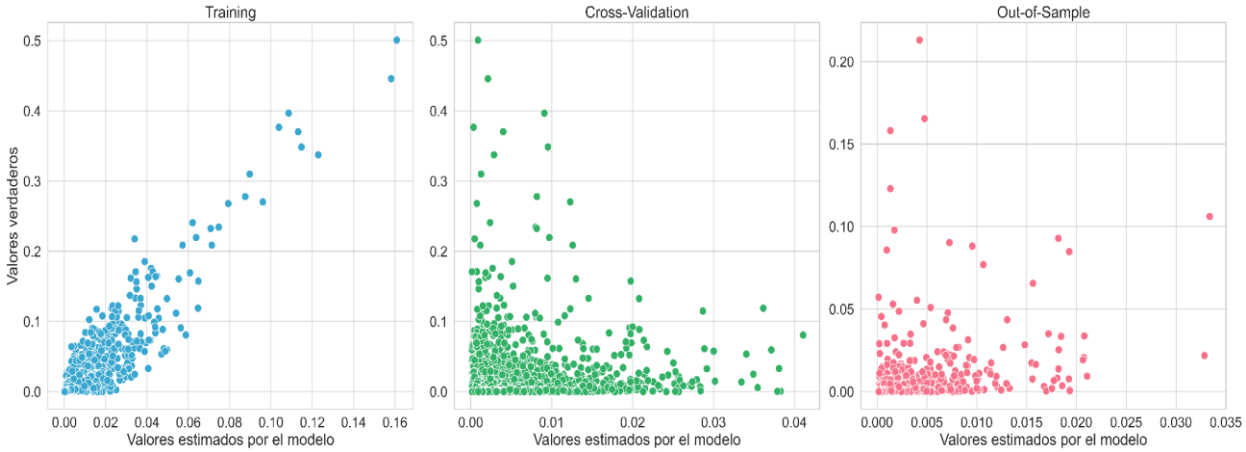


Fuente: resultados propios en base a datos extraídos de Mercado Libre, septiembre de 2024.

La distribución de la variable objetivo está muy sesgada hacia el 0 (el sesgo es tan extremo que se dificulta la visualización, incluso aplicando transformaciones). Por consiguiente, se vuelve muy difícil superar a las predicciones del modelo base, pues la mayoría de valores se encuentran muy cerca de la mediana. En este sentido, creemos que el modelo random forest puede aportar información valiosa sobre el efecto de las regresoras en el market *share* de visitas de cada artículo. Graficando las predicciones del modelo contra los valores verdaderos se observa su capacidad de distinguir las observaciones con alta participación en el número de visitas de la categoría (observaciones alejadas del cero y la mediana. Esta mayor calidad del ajuste es evidente especialmente en la muestra de entrenamiento:

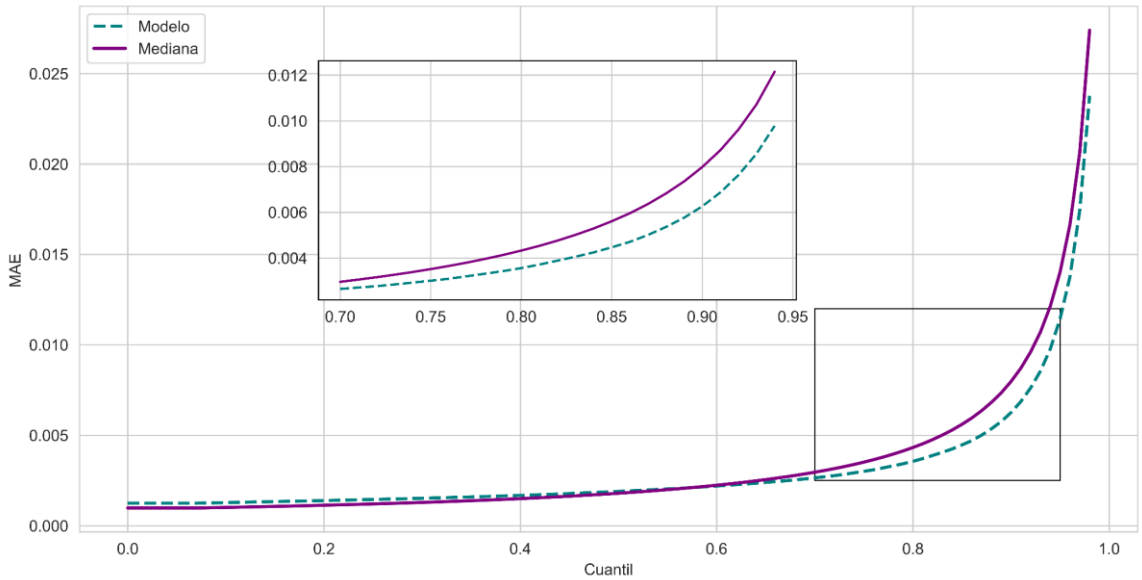
Examinando las predicciones a partir de cada cuantil de la variable objetivo, se detecta que la mejora se da aproximadamente a partir del cuantil 0.6 para la muestra de entrenamiento, y 0.7 fuera de la muestra. Para esta última la ventaja sobre el modelo base parece ser aún mayor.

Figura 15: *diagramas de dispersión entre las predicciones del modelo y los valores de la variable objetivo.*



Fuente: resultados propios en base a datos extraídos de Mercado Libre, septiembre de 2024.

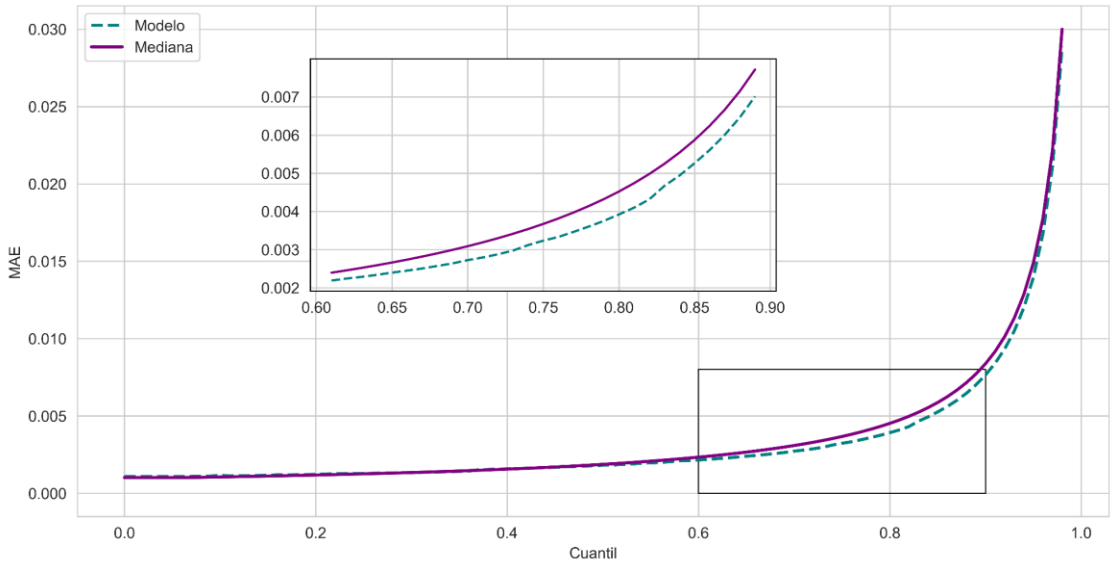
Figura 16: *error de predicción del modelo y la mediana a partir de distintos cuantiles de la variable objetivo.*



Fuente: resultados propios en base a datos extraídos de Mercado Libre, septiembre de 2024.

Luego de analizar los resultados del error de predicción dentro de la muestra, se presenta ahora el comportamiento del modelo fuera de la muestra; comparando resultados y evaluando así la capacidad de generalización del modelo:

Figura 17: *Error de predicción del modelo y la mediana a partir de distintos cuantiles de la variable objetivo fuera de la muestra.*



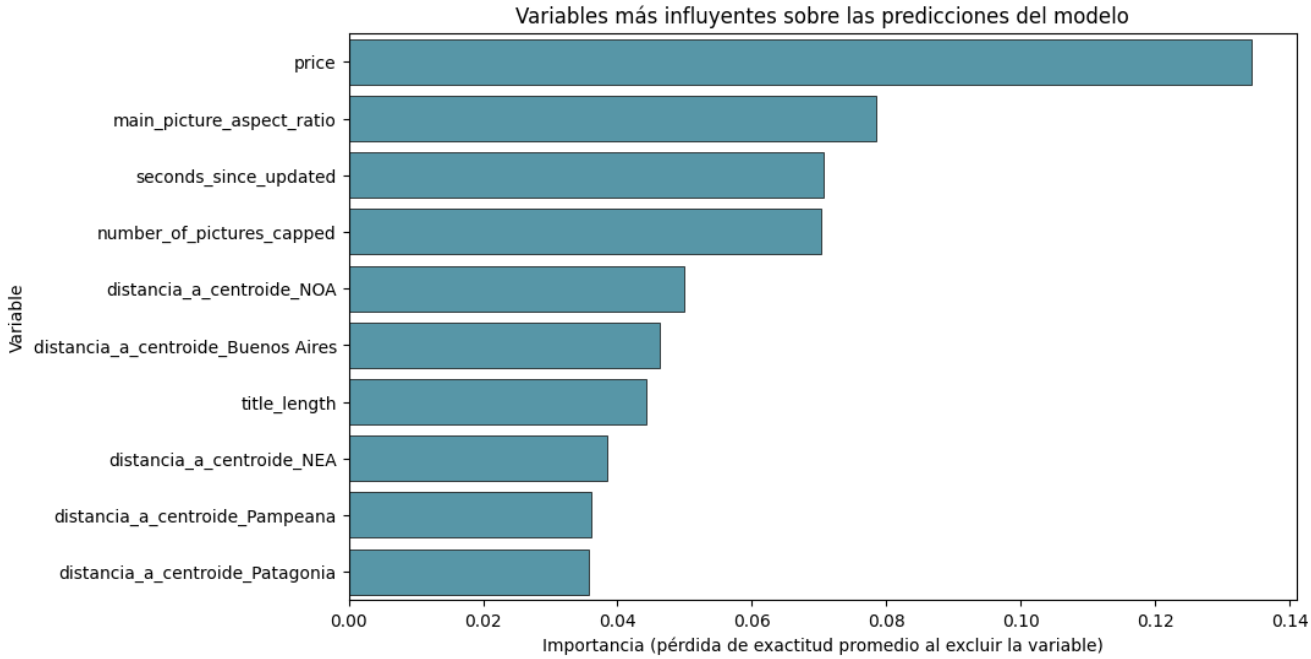
Fuente: resultados propios en base a datos extraídos de Mercado Libre, septiembre de 2024.

Con el objetivo de analizar las variables más relevantes para las predicciones del modelo, se obtienen las importancias de cada atributo en el estimador random forest. Estas se generan omitiendo cada variable una a la vez y calculando el impacto negativo que esto conlleva sobre la precisión a lo largo de todos los árboles del modelo (Chan, Harris, Singh y Yeo, 2022). Sin embargo, es mejor proceder con cautela al interpretar estos resultados, ya que el procedimiento tiende a preferir fuertemente las variables de alta cardinalidad (Scikit-learn, n.d). Por otro lado, este procedimiento simplemente muestra cómo las variables son utilizadas para predecir en la muestra de entrenamiento, por lo que no suelen ser aptos para explicar predicciones en otros conjuntos de datos, ni mucho menos hacer inferencia estadística.

Como se puede ver en el siguiente gráfico, el precio es holgadamente el atributo más influyente sobre el modelo. Resulta llamativo que este hace uso de varias de las columnas

geográficas. Como es de esperar, las variables categóricas pierden importancia relativa por tener pocas particiones en el espacio de parámetros de cada árbol.

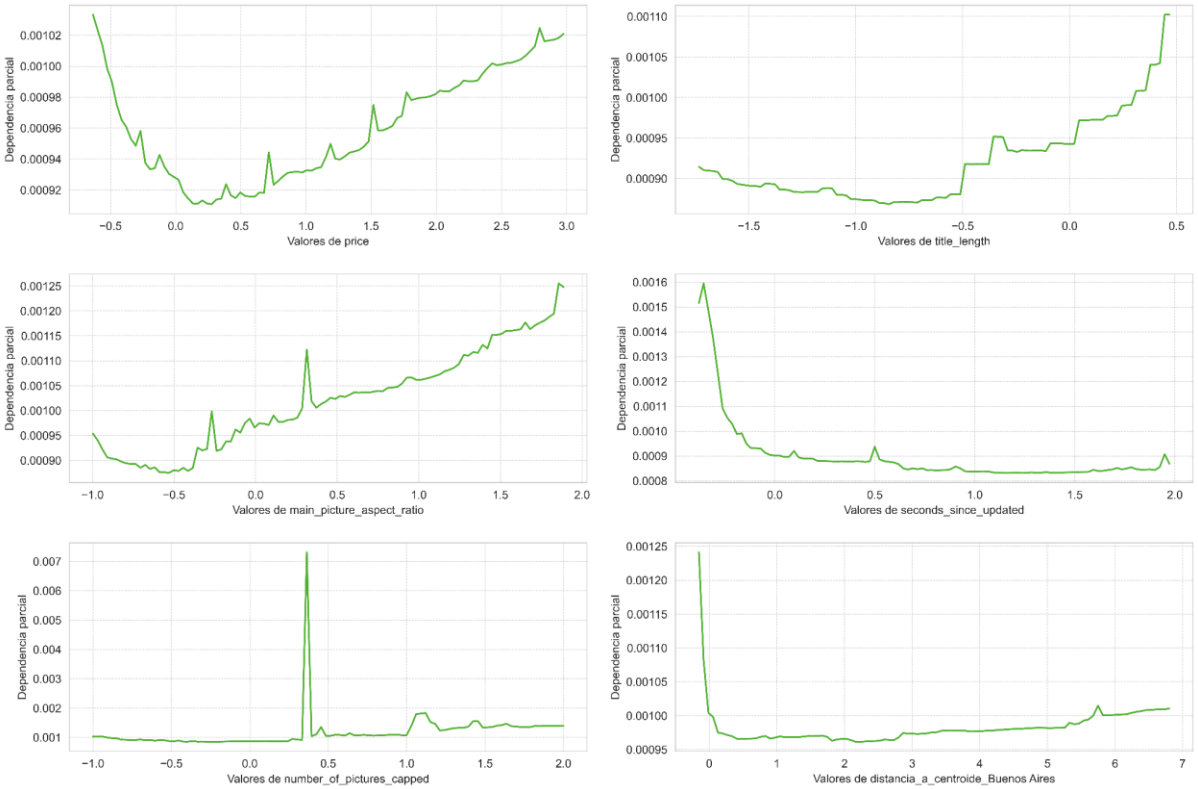
Figura 18: *variables más importantes para explicar las predicciones del modelo.*



Fuente: elaboración propia en base a los resultados mediante Scikit-Learn.

En cuanto a las relaciones específicas entre las variables y las estimaciones del modelo, se construyen gráficos de dependencia parcial para aquellas que resultan de interés económico. Esta técnica busca examinar el comportamiento de las predicciones para distintos valores de una determinada predictora, marginalizando sobre el resto de variables independientes.

Figura 19: *diagramas de dependencia parcial de las principales variables explicativas.*



Fuente: elaboración propia en base a los resultados mediante Scikit-Learn.

Al igual que la importancia de las variables, la dependencia parcial no está libre de complicaciones y supuestos limitantes. En el proceso de estimación se simulan valores que podrían ser muy difíciles o imposibles de observar si la variable de interés está correlacionada con el resto de predictoras (Molnar, 2020). Sin embargo, este tipo de recursos suelen ser la única opción a la hora de dar interpretabilidad a modelos complejos como el random forest.

El precio (escalado sobre la categoría) dibuja una curva no monótona, con las predicciones más altas para el *share* de visitas ubicadas en los extremos. Por su parte, el largo del título y el ratio del ancho sobre alto en píxeles de la imagen siguen una relación similar, aunque los valores más altos impactan más sobre las visitas estimadas. El tiempo desde la última actualización marca una tendencia clara: mientras más reciente es la actualización, más visitas predice el modelo. El número de imágenes de la publicación presenta un pico muy

elevado, cerca de la mitad, pero de lo contrario evidencia una trayectoria ascendente del número de visitas. Finalmente, las distancias a cada centroide presentan un patrón descendente y luego ascendente, indicando un mayor número de visitas cerca de los centros. Para Buenos Aires la cercanía al centroide marca un aumento todavía más fuerte en la predicción que en el resto de regiones.

6. Conclusiones

Los resultados indican que el número de fotografías está correlacionado positivamente con las visitas en muestras con títulos duplicados. Asimismo, la variable tiene un alto peso en explicar las predicciones del mejor modelo de aprendizaje automático, por lo que puede resultar valioso profundizar en su estudio.. Mercado libre provee información detallada sobre todas las imágenes de un producto (al mismo tiempo, brinda acceso a las propias imágenes), por lo que sería una opción interesante continuar explorando la relación entre estas y el éxito en las publicaciones. Este trabajo incluye únicamente datos sobre la cantidad de imágenes y la calidad de la imagen de portada, pero podría ampliarse el estudio y la captación de datos, desarrollando modelos más complejos y capaces de incorporar la información contenida en las imágenes en sí. Esto permitiría analizar cómo características novedosas (la iluminación, los colores, o la resolución, etc.) influyen sobre las visitas del producto.

Resulta interesante que los hallazgos en las muestras de vinos y aguas minerales son idénticos: el tipo de logística empleado y el estatus del vendedor parecen describir grupos diferentes en estos dos subconjuntos. En otras palabras, existen diferencias significativas en el *share* promedio de visitas de los productos según el estatus del vendedor y el tipo de logística.

Mediante una inspección visual de las diferencias, los grupos de *fulfillment* (envíos rápidos y desde los depósitos de la empresa) y *cross docking* (un agente recolecta los paquetes directamente en el domicilio del vendedor) se destacan por recibir un mayor número de visitas según la logística.Y por el lado de la posición del vendedor en la escala de *mercado líder*, curiosamente aquellos que no califican tienden a recibir un número de visitas similar o mayor en los grupos bajo estudio. Quizás alguna otra variable esté explicando este fenómeno, por lo

que podría resultar beneficioso continuar investigando en esa dirección. Desafortunadamente, estas columnas quedan rezagadas en el análisis de importancia. Al ser del tipo categóricas, se dificulta su comparación con el resto en relación al modelo predictivo y se pierde la posibilidad de elaborar conclusiones más complejas sobre su influencia en el algoritmo.

En cuanto al modelo predictivo en sí, el precio es la variable con mayor trascendencia. No resulta sorprendente encontrarla liderando. La influencia de los precios en el comportamiento de los agentes económicos ha sido ampliamente estudiada por la literatura económica, del marketing , y la psicología (Varian, 2014; Gourville y Soman, 2002; Kienzler y Kowalkowski, 2017). Sin embargo, lo destacable de este resultado es la relación entre esta variable y las predicciones del modelo. Aparentemente los productos que mayor proporción de visitas reciben son aquellos con precios extremos en relación a su categoría. Se podría asociar con dos comportamientos distintos por parte de los consumidores: un grupo que busca precio, y otro que busca calidad (o que al menos termina visitando los artículos más caros de la categoría de su interés).

También se destacan otras relaciones entre las regresoras y la regresada. Según los resultados, títulos más largos están asociados con mayores visitas. Igualmente, el ratio de píxeles (ancho por alto) de la imagen parece jugar un papel importante en la atención que recibe el artículo por parte de los compradores. Lamentablemente, para estas columnas no se aplicó estandarización respecto a la categoría, sino a toda la muestra. Por lo tanto, las conclusiones que surgen del análisis no son tan directas. En este sentido, podemos sospechar que publicaciones con títulos más largos o imágenes más anchas están asociadas a un nivel superior de visitas, pero no necesariamente comparado con los artículos de su categoría (un título más largo no implica en todos los casos un título más largo en relación a productos similares).

Para el número de imágenes de la publicación, se confirma el resultado encontrado en la muestra con títulos duplicados (una asociación positiva entre el número de imágenes y la porción de visitas del artículo). Aunque hay un pico en la dependencia que probablemente se deba a otras variables. En cuanto al tiempo de actualización de la publicación, hay una clara tendencia decreciente en las predicciones. Posiblemente los proveedores que consiguen mayor éxito en sus publicaciones hayan logrado mayores niveles de automatización en el inventario, las promociones y la estrategia de precios. No obstante, se interpretan los resultados como evidencia a favor de las actualizaciones frecuentes como buena práctica para aumentar la visibilidad en la plataforma.

Sin duda los clientes exhiben patrones de comportamiento complejos y dinámicos. Para poder adquirir mayores conocimientos sobre el efecto de estas variables en su decisión, creemos que sería necesario ajustar modelos paramétricos a los datos, aunque salvando los problemas de exactitud enfrentados en esta investigación. Como ya se mencionó, el mayor valor económico de las estimaciones reside en las observaciones atípicas, aquellas alejadas del cero y que resultan casi imposibles de predecir para los modelos de regresión tradicionales. Precisamente este es el compromiso que se asume al centrarse en algoritmos complejos como el random forest: se gana poder predictivo y ajuste a los datos, pero se pierde interpretabilidad y capacidad de inferencia.

A pesar de las limitaciones inherentes a este trabajo, los hallazgos aportan un conocimiento exploratorio valioso sobre los datos analizados. El estudio ha permitido identificar y describir patrones en la distribución de visitas según diversas variables independientes, generando indicios sobre temas clave que podrían guiar investigaciones futuras. Si bien los resultados no son plenamente generalizables, este trabajo sienta un

precedente en una línea de investigación que de otra manera sería inexplorada, abriendo nuevas oportunidades para estudios más profundos.

7. Referencias

DeLone, W. H., y McLean, E. R. (2003). The DeLone and McLean model of information systems success: A ten-year update. *Journal of Management Information Systems*, 19(4), 9-30. https://doi.org/10.1080/07421222.2003.11045748

Benchmark International. (2023). 2023 *Global eCommerce Industry Report*. Recuperado de https://www.benchmarkintl.com/insights/2023-global-ecommerce-industry-report/

Pan, J. y Chen, Y. (2010). Determinants of Success for Online Retailers: The Roles of the Consumer and the Technology. *Journal of Interactive Marketing*, 24(1), 42–56. https://doi.org/10.1016/j.intmar.2009.10.001

Kotler, P., Kartajaya, H., y Setiawan, I. (2017). *Marketing 4.0: Moving from traditional to digital* (A. Martín, Trad.). Wiley.

Liu, J. (2024). Research on the Influencing Factors of Cross-Border E-commerce Consumers. *Frontiers in Business, Economics and Management*, 13(1), 84-86.

Gerpott, T. J., y Berends, J. (2022). Competitive pricing on online markets: A literature review. *Journal of Revenue and Pricing Management*, 21(6), 596–622. https://doi.org/10.1057/s41272-022-00390-x

Munoz, F., Holsapple, C. W., y Sasidharan, S. (2023). E-commerce. *Springer Hand-book of Automation* (pp. 1411–1430). Springer.

Jadhav, A., Pramod, D., y Ramanathan, K. (2019). Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*, 33(10), 913–933. https://doi.org/10.1080/08839514.2019.1637138

Zheng, A. and Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and techniques for Data Scientists*. O'Reilly Media, Inc., Sebastopol.

Selvaraj, S., y Sabarish, B. (2021). Analysis of distance measures in spatial trajectory data clustering. *IOP Conference Series: Materials Science and Engineering*, 1085(1), 012021. https://doi.org/10.1088/1757-899X/1085/1/012021

Díaz de Astarloa, B., y Lotitto, E. (2023). The landscape of B2C e-commerce marketplaces in Latin America and the Caribbean. *Serie Desarrollo Productivo, CEPAL*. https://hdl.handle.net/11362/48583

Loureiro, D., Barbieri, F., Neves, L., Espinosa Anke, L., y Camacho-Collados, J. (2022). TimeLMs: Diachronic language models from Twitter. arXiv preprint arXiv:2202.03829. https://doi.org/10.48550/arXiv.2202.03829

Amorim, L. B. V., Cavalcanti, G. D. C., y Cruz, R. M. O. (2022). The choice of scaling technique matters for classification performance [Preprint]. arXiv.https://arxiv.org/abs/2212.12343

James, G., Witten, D., Hastie, T., Tibshirani, R., y Taylor, J. (2023). *An introduction to statistical learning with applications in Python*. Springer Texts in Statistics. Springer, Cham. https://doi.org/10.1007/978-3-031-38747-0

Tibshirani, R., Walther, G., y Hastie, T. (2001). Estimating the number of clusters in a data set via the Gap statistic. *Journal of the Royal Statistical Society: Series B* (Statistical Methodology), 63(2), 411-423. https://doi.org/10.1111/1467-9868.00293

Kharfan, M., Chan, V. W. K., y Efendigil, T. F. (2021). A data-driven forecasting approach for newly launched seasonal products by leveraging machine-learning approaches. *Annals of Operations Research*, 303(1), 1–22. https://doi.org/10.1007/s10479-020-03666-w

Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. Data, 4(1), 15. https://doi.org/10.3390/data4010015

Mathotaarachchi, K. V., Hasan, R., y Mahmood, S. (2024). Advanced machine learning techniques for predictive modeling of property prices. *Information*, 15(6), 295. https://doi.org/10.3390/info15060295

Zhang, X., Guo, F., Chen, T., Pan, L., Beliakov, G., y Wu, J. (2023). A brief survey of machine learning and deep learning techniques for e-commerce research. *Journal of Theoretical and Applied Electronic Commerce Research*, 18(4), 2188-2216. https://doi.org/10.3390/jtaer18040110

Hoang, D., y Wiegratz, K. (2023). Machine learning methods in finance: Recent applications and prospects. University of Hohenheim, Karlsruhe Institute of Technology. https://doi.org/10.2139/ssrn.4294396

Pesarin, F., y Salmaso, L. (2010). The permutation testing approach: A review. Statistica, 70(4). https://doi.org/10.6092/issn.1973-2201/3599

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. J*ournal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288. http://www.jstor.org/stable/2346178

Chan, F., y Mátyás, L. (2022). Linear econometric models with machine learning. En F. Chan y L. Mátyás (Eds.), *Econometrics with machine learning* (pp. 1–39). Springer. https://doi.org/

Chan, F., Harris, M. N., Singh, R. B., y Yeo, W. (B. E.). (2022). Nonlinear econometric models with machine learning. En F. Chan y L. Mátyás (Eds.), *Econometrics with machine learning* (Cap. 2). Springer. https://doi.org/

Molnar, C. (2020). Interpretable Machine Learning (A Guide for Making Black Box Models Explainable). La Biblia de la IA - *The Bible of AI™ Journal* (15 de November de 2024) https://editorialia.com/2020/06/23/r0identifier_b129021066d4fc15a561e0053c355588/.

Porter, D. C., Gujarati, D. N. (2009). *Basic Econometrics 5th ED (Fifth Edition)*. Boston: McGraw-Hill.

Pesarin, F., y Salmaso, L. (2010). The permutation testing approach: A review. STATISTICA, LXX(4), 481-501.

Scikit-learn. (n.d.). Permutation feature importance. Scikit-learn documentation. https://scikit-learn.org/1.5/auto_examples/inspection/plot_permutation_importance.html

Varian, Hal R., author. (2014). *Intermediate microeconomics : a modern approach*. New York :W.W. Norton y Company,

Gourville, J., y Soman, D. (2002). Pricing y the psychology of consumption. *Harvard Business Review*, 80(9), 90–96, 126. PMID: 12227149.

Kienzler, M., y Kowalkowski, C. (2017). Pricing strategy: A review of 22 years of marketing research. *Journal of Business Research*, 78, 101–110. https://doi.org/10.1016/j.jbusres.2017.05.005

UNCUYO UNIVERSIDAD NACIONAL DE CUYO | FACULTAD DE CIENCIAS ECONÓMICAS

DECLARACIÓN JURADA RESOLUCIÓN 212/99 CD

El autor de este trabajo declara que fue elaborado sin utilizar ningún otro material que no haya dado a conocer en las referencias que nunca fue presentado para su evaluación en carreras universitarias y que no transgrede o afecta los derechos de terceros.

Mendoza, 4/12/2024

Bayarski BENJAMIN BUSTOS

Firma y aclaración

31903

Número de registro

43270485

DNI

UNCUYO UNIVERSIDAD NACIONAL DE CUYO | FACULTAD DE CIENCIAS ECONÓMICAS

DECLARACIÓN JURADA RESOLUCIÓN 212/99 CD

El autor de este trabajo declara que fue elaborado sin utilizar ningún otro material que no haya dado a conocer en las referencias que nunca fue presentado para su evaluación en carreras universitarias y que no transgrede o afecta los derechos de terceros.

Mendoza, 5 de diciembre de 2024

Lorenzo Lorenzo

Firma y aclaración

33226

Número de registro

42863278

DNI