UNIVERSIDAD
TORCUATO DI TELLA

Master in Management + Analytics

Predicting used car prices from data collected with web scraping and machine learning techniques

TESIS

Lucía María Lopez Wallace

Mayo 2022

Tutor: Andrés Babino

Resumen

La compraventa en línea de vehículos viene creciendo a un ritmo exponencial tras los cambios en los hábitos de los consumidores que buscan maneras más simples para comprar un auto. En entornos cada vez más competitivos, aquellas empresas que logren una correcta estrategia de determinación de precios lograrán también traccionar ventas y tener una mejor posición en el mercado.

En la presente tesis se utilizaron técnicas de *web scraping* para recolectar información significativa del mercado que luego, mediante técnicas de aprendizaje automático, fue utilizada para predecir los precios de mercado para autos usados.

Se trabajaron diferentes modelos de aprendizaje automático con los datos obtenidos a través del *web scraping* del principal sitio en línea de compra y venta de autos usados. Estos datos fueron luego utilizados como *insights* para medir el precio del *stock* de una plataforma que comercializa autos usados.

Como resultado, se obtuvo un modelo con un rendimiento que alcanzó un puntaje de 371,4 miles de pesos de error absoluto versus valores reales. Dicho modelo tiene como objetivo ser la base para estimar los precios de mercado, sobre los cuales luego se trabajará la estrategia de precios que permitirá alcanzar la rentabilidad que la empresa necesita. Se midió el impacto económico de aplicar esta nueva metodología, y se obtuvo una ganancia de 1,3 millones de dólares para el stock actual de una compañía que compra y vende auto en línea.

Abstract

The online used car industry has been growing at an exponential rate following changes in consumer habits looking for simpler ways to buy a car. In increasingly competitive environments, those companies that achieve a correct pricing strategy will be able to transact sales and have a correct position in the market.

In this thesis, it is argued that, through web scraping techniques, powerful market information can be collected that will later, through machine learning techniques, be used to predict market prices for used cars.

Different machine learning models were used with the data obtained through the web scraper of the main marketplace for buying and selling used cars. That were later used as insights to measure the price of the stock of a platform that sells used cars.

As a result, a model was obtained with a performance that reached a score of 371,4 thousand pesos of absolute error versus real values. The objective of this model is to be the basis for estimating market prices, on which the pricing strategy that will allow the company to achieve the profitability it needs is then worked on. The economic impact of applying this new methodology was measured, obtaining a profit of 1.3 million dollars.

Índice

Índice de tablas

Índice de figuras

1. Introducción

1.1. Contexto

El proceso de compra de un automóvil sigue siendo relativamente anticuado y fragmentado en *retailers* que invierten poco en canales de venta digitales. A su vez, los usuarios todavía confían en estos métodos tradicionales, y esto está frenando a la industria. Las transacciones por medio de una plataforma digital prometen ayudar a superar varios puntos débiles existentes en el *customer journey*, como el engorroso papeleo y las visitas a múltiples puntos de venta. También prometen muchos beneficios adicionales y, no menos importante, el acceso instantáneo a una amplia gama de vehículos y la fácil evaluación comparativa de precios. En consecuencia, el mercado automovilístico tiene grandes desafíos con respecto a la digitalización y, más específicamente, las ventas en línea.

Panorama del mercado: un sector atractivo con un panorama competitivo fragmentado

Los autos usados son un mercado en crecimiento impulsado principalmente por la cantidad de vehículos en uso, que a su vez está determinado por el crecimiento de la población y las tasas de uso de vehículos per cápita [2]. La inercia de este crecimiento ayuda a que el mercado sea resistente. El comercio minorista de automóviles usados (B2C) está abrumadoramente dominado por los distribuidores físicos tradicionales. Esto deja solo una fracción del mercado ocupada por dos nuevas categorías: pequeños retailers en línea y nuevos retailers en línea [13].

Como sugiere el nombre, los retailers en línea ejecutan la mayor parte de su proceso de ventas en línea; los ejemplos en el mercado argentino incluyen Kavak, con sede en Mexico y operación en Chile, COlombia, Peru y Brasil; y Karvi con sede en Argentina y operación en Argentina y Brazil. Por el otro lado, los nuevos reatilers en línea son generalmente jugadores establecidos en la parte superior de la cadena de valor que también han comenzado a vender sus vehículos directamente, principalmente en línea. Incluyen arrendadores, compradores en efectivo C2B y plataformas de subastas, como Toyota con www.toyotausados.com.ar o Car One con www.carone.com.ar.

El auge *online*: cómo las ventas digitales están acelerando el mercado de autos usados

Comprar un automóvil usado de un concesionario es un proceso que presenta múltiples complicaciones y numerosos puntos débiles en el camino. La mayoría de los clientes potenciales comienzan investigando en Internet. Pero, de inmediato, se enfrentan a dificultades, ya que muchos sitios web de autos usados ofrecen una experiencia de usuario deficiente y poca información, como especificaciones y fotos, sobre los vehículos disponibles. Para

recopilar toda la información que necesitan, los clientes a menudo necesitan visitar varios puntos de venta diferentes. Esto es una pérdida de tiempo y también limita su elección a los vehículos disponibles en su localidad. Los clientes también se sienten frustrados cuando intentan obtener la mejor valoración para su vehículo de intercambio, ya que esto suele requerir varias visitas más a los concesionarios.

Los peligros de comprar un auto usado de la manera tradicional

El proceso de compra también es accidentado. A los clientes no les gusta la complejidad y la cantidad de papeleo involucrado, generando frustración entre los compradores de automóviles tal como muestra la Figura 1. También les frustra la falta de transparencia en los precios y son reacios a negociar. No tener un solo punto de contacto en el concesionario para facilitarles el proceso de venta es otra pesadilla. Por último, los clientes perciben las fechas de entrega estimadas como poco confiables y resienten los largos tiempos de espera que implica el proceso [3].

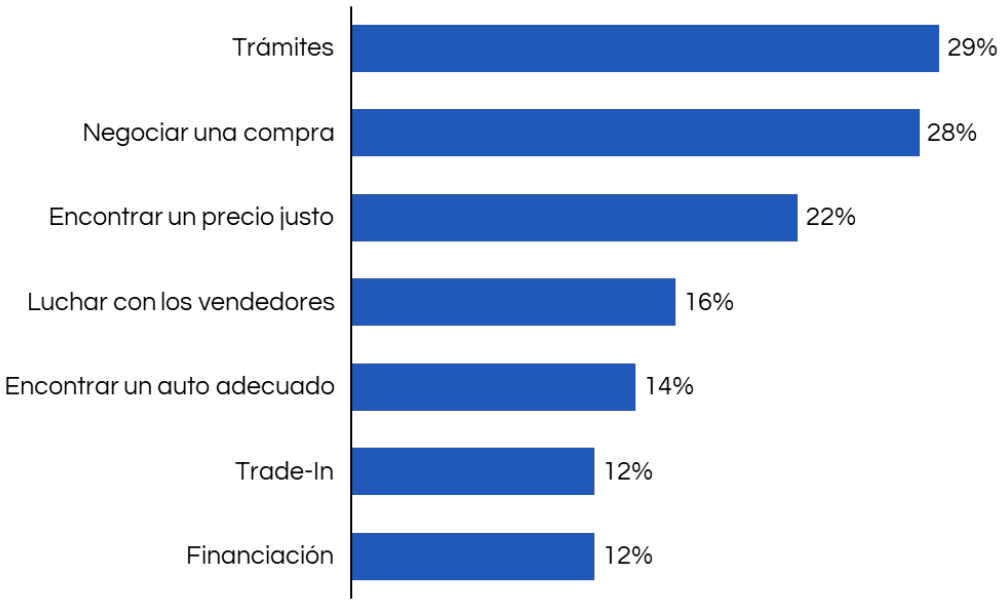

Figura 1. Estudio de CoxAutomotive 2019 para los compradores de auto en EE.UU.

El potencial de las ventas *online*

En el *retailer online*, las ventas *online* están ayudando a mejorar la experiencia del cliente. Cada vez más usuarios recurren a estos vendedores minoristas para comprar productos, y se agregan más y más categorías de productos todo el tiempo. Esto está generando confianza y creando hábitos de compra duraderos. La tendencia es impulsada tanto por los jóvenes como por las generaciones mayores que prueban las compras *online* por primera vez. Además, la pandemia de Covid-19 ha acelerado la tendencia hacia

las compras en línea, aumentando la penetración *online* y cambiando el comportamiento del consumidor. La industria minorista de autos usados también está aprovechando esta tendencia [1].

En respuesta, los vendedores están optimizando la experiencia del cliente para hacerla más simple y accesible, así como también, impulsando desarrollos tecnológicos en torno a la experiencia del usuario y a la simplificación del pago para facilitar el proceso.

Los clientes están empezando a esperar una experiencia completamente digital que brinde soluciones a los puntos débiles del proceso de compraventa de autos. Esto incluye acceso instantáneo a una amplia gama de vehículos, recomendaciones personalizadas, comparaciones de precios, menos o incluso ningún papeleo, y entregas a domicilio de vehículos. También está claro que los compradores valoran una plataforma y una experiencia fluida con el vendedor: Carvana, el canal minorista estadounidense de autos usados, ha logrado altas calificaciones de satisfacción del cliente, con un puntaje neto de promotor en 2020 de 82pp [4].

Los factores clave detrás de una transformación digital exitosa

La digitalización no solo significa tener un sitio web elegante; para ser una plataforma exitosa, las empresas necesitan un modelo de negocio orientado digitalmente, que difiera significativamente del modelo minorista tradicional. Podemos identificar seis factores clave de éxito para ayudar a las plataformas digitales a desarrollar dicho modelo [13]:

A. **Construir una marca fuerte**: Esto ayudará a la confianza de los potenciales compradores sobre la compra de un automóvil usado vía la plataforma.

B. **Adoptar un negocio basado en datos**: Se trata de generar estos datos a partir de los usuarios que navegan en su web y entender su comportamiento para luego fortalecerlos agregando datos externos del mercado y de la competencia. Por otro lado, para aprovechar estos valiosos datos, debe implementarse un procesamiento de datos avanzado para mejorar su evaluación.

C. **Combinar el mundo digital con el mundo físico**: Es esencial disponer de puntos físicos de retiro y entrega, tanto para el reconocimiento de la marca como para la optimización de costos.

D. **Establecer centros de reacondicionamiento avanzados**: El reacondicionamiento es un paso clave en la cadena de valor porque representa un elemento de gran costo y porque impulsa la calidad de los autos usados. Debe hacerse internamente ya que permite diferenciarse de la competencia en dos dimensiones principales. En

primer lugar, el proceso de renovación en sí mismo, que debe ser ágil para optimizar los costos y el tiempo de entrega. En segundo lugar, la maximización del valor de la renovación, que requiere una consideración detallada tanto del valor creado para los clientes como del costo de la renovación (es decir, qué renovación debe realizarse y mediante qué proceso).

E. **Controlar estrictamente los flujos logísticos**: el mercado de autos usados *online* genera importantes flujos logísticos, ya sea para la toma de autos, las entregas a domicilio o los traslados desde y hacia centros de rehabilitación. Una estrecha gestión de estos temas es clave para ofrecer entregas y recogidas rentables, rápidas y confiables a los clientes finales.

F. **Garantizar que el abastecimiento de vehículos sea eficiente y diversificado**: el mercado de autos usados es principalmente un negocio de suministro; por lo tanto, ofrecer la combinación adecuada de vehículos (en términos de marca, modelo de venta *online* de autos usados, la configuración, el kilometraje, la edad, etc.) para satisfacer la demanda es clave para superar a la competencia. Para hacerlo, los actores deben tener acceso rápido a los vehículos correctos a los mejores precios y, por lo tanto, deben aprovechar múltiples canales de abastecimiento, como intercambios de clientes, subastas y empresas de arrendamiento.

1.2. Problema

Un modelo comercial orientado digitalmente que integre con éxito estos factores claves de éxito garantizará que un canal de autos usados *online* sea competitivo en el precio del vehículo en comparación con un canal tradicional. De hecho, la competitividad seguirá aumentando junto con los volúmenes de venta *online* gracias a los efectos de escala (medios, TI, logística, etc.). Esta competitividad de precios, combinada con una mejor experiencia del cliente, permitirá que el canal *online* supere a los canales tradicionales y representen una mayor participación del mercado a largo plazo.

Además de la competitividad de los precios, los seis factores claves de éxito también influyen en la competencia en el mercado. Cada uno representa una importante barrera de entrada, especialmente la creación de marca y la adopción de factores comerciales basados en datos. Estas barreras son relevantes para todos los actores del mercado porque el comercio minorista *online* requiere activos que no son necesarios para las ventas tradicionales. La combinación de barreras de entrada y efectos de escala creará un mercado mucho más concentrado y rentable que el actual mercado tradicional de autos usados.

El principal desafío de estas plataformas es determinar cuál es el precio de mercado del auto usado que se quiere comprar y sobre ese precio definir la estrategia de precios

que les permita la rentabilidad. A su vez, ese desafío en el mercado argentino es mucho mayor dada la inflación que sufre Argentina (actualmente, se estima al menos un 50 % anual) y requiere una constante actualización de precios para garantizar la rentabilidad del negocio.

En este trabajo de tesis, se busca encontrar en mejor precio de mercado para una empresa privada que funciona como un retailers en línea. Esta empresa privada compra y vende autos usados y el éxitos de su negocio es conocer el precio de mercado de los autos usados a los cuáles luego se le aplica una estrategia de precios para alcanzar sus objetivos financieros.

1.3. Objetivo

Utilizando técnicas de *Web scraping*, se colectará información de un reconocido mercado de compra y venta de autos. Luego, con técnicas de aprendizaje automático (o machine learning), se utilizarán los datos recolectados para estimar el precio de mercado de ciertos autos usados.

Con los precios de mercado se buscará entender si el inventario de esta plataforma *online* requiere un cambio en la política de precios que tenga como resultado las siguientes sugerencias:

1. **Bajar el precio**: Para liquidar el stock inmovilizado
2. **Subir el precio**: En casos donde se esté ofreciendo un precio menor al mercado
3. **Mantener la estrategia de precios**

Y, por otro lado, la captura de perspectivas de mercado retroalimentará a los equipos de compra de autos para corregir la política de nuevas adquisiciones con el objetivo de tener un stock saludable en términos de rotación y valoración del mercado.

2. Datos

Para este estudio, se utilizó datos públicos y privados, que fueron relacionados para trabajar el problema de negocio.

Los datos públicos utilizados fueron extraídos de la web de Mercado Libre. Para la extracción de esta información se utilizaron técnicas de *Web scraping*. Por un lado, una técnica inicial sencilla de *web scraping* y, luego, se buscó enriquecer la base utilizando la API (Application Programming Interface) pública de Mercado Libre de donde se obtuvo un mayor grado de detalle de la información de las publicaciones.

Es importante destacar que se utilizó Mercado Libre como fuente de datos ya que es el sitio de *e-commerce* número uno del país para la compra y venta de autos usados. Y a su vez, se validó que las publicaciones de Mercado Libre sean marcas y modelos de autos que la empresa privada tiene disponible en su stock.

2.1. Datos privados

Estos datos fueron provistos por una empresa privada la cual compartió los datos del stock actual. En este momento contaba con un stock a la venta de 2.000 autos. Su stock cuenta con 23 marcas diferentes y 133 modelos distintos de autos. Se buscarán los precios asociados a estas marca-modelo en las publicaciones de Mercado Libre.

2.2. Datos públicos

2.2.1. Web scraping

Se utilizaron técnicas de *web scraping* para extraer información de Mercado Libre. *Scraping* es el proceso de extraer, copiar, filtrar o recopilar datos; la extracción de datos de la web (comúnmente conocidos como sitios web o páginas web, o recursos relacionados con Internet) normalmente se denomina *web scraping*.

Web scraping es un proceso de extracción de datos de la web que es adecuado para ciertos procesos. La recopilación y el análisis de datos, y su participación y toma de decisiones, además de las actividades relacionadas con la investigación, hacen que el proceso de *web scraping* sea sensible para todo tipo de industria.

La popularidad de Internet y sus recursos está provocando que los dominios de información evolucionen cada día, lo que también provoca una demanda creciente de datos sin procesar. Los datos son el requisito básico en los campos de la ciencia, la tecnología y la gestión. Los datos recopilados y organizados se procesan con diversos grados de lógica para obtener información y obtener más información[5].

Para realizar el *web scrapper* se utilizaron las bibliotecas Selenium y Beautiful soup para extraer los datos del código html de la web.

Selenium

Selenium es una poderosa herramienta de *web scraping* desarrollada originalmente para probar sitios web. En estos días, también se usa cuando se requiere una representación precisa de los sitios web, tal como aparecen en un navegador. *Selenium* funciona al automatizar los navegadores para cargar el sitio web, recuperar los datos requeridos e incluso tomar capturas de pantalla o afirmar que ciertas acciones ocurren en el sitio web. *Selenium* no contiene su propio navegador web; requiere integración con navegadores de terceros para ejecutarse. En este caso, se ejecutó *Selenium* con Firefox abriéndose una instancia de Firefox en la pantalla desde donde se navego al sitio web y realizó las acciones especificadas en el código[16].

Beautiful Soup

Beautiful Soup es una biblioteca de Python destinada principalmente a analizar y extraer información de una cadena HTML (HyperText Markup Language). Viene con una variedad de analizadores de HTML que permiten extraer información incluso de un HTML mal formateado, lo que desafortunadamente es más común de lo que se supone[18].

De la web de Mercado Libre se extrajo información de las publicaciones de autos usados; cada publicación cuenta con la siguiente información:

- Precio
- Moneda
- Localidad
- Provincia
- Kilometraje
- Año
- Marca
- Título

En la Tabla 1, se puede observar una muestra de estos datos.

precio	moneda	localidad	provincia	km	año	marca	título
1.950.000	$	La Matanza	Bs.As. G.B.A. Oeste	35900	2014	ford	Ford Fiesta Kinetic Design 1.6 Se Plus 120cv
3.750.000	$	Vicente López	Bs.As. G.B.A. Norte	35900	2018	citroen	Citroën C4 Cactus 1.2 Puretech 110 At6
3.900.000	$	La Plata	Bs.As. G.B.A. Sur	35900	2019	citroen	Citroën C4 Cactus 1.6 Vti 115 At6 Shine
3.500.000	$	Lobos	Buenos Aires Interior	36357	2020	peugeot	Peugeot Partner 1.6 Hdi Confort
2.560.000	$	Morón	Bs.As. G.B.A. Oeste	76695	2016	ford	Ford Focus Iii 1.6 S
2.000.000	$	Tigre	Bs.As. G.B.A. Norte	76835	2016	peugeot	Peugeot 408 1.6 Active 115cv
2.790.000	$	Capital Federal	Capital Federal	77958	2017	peugeot	Peugeot 408 1.6 Feline Hdi 115cv
1.790.000	$	San Fernando	Bs.As. G.B.A. Norte	78041	2017	fiat	Fiat Mobi 1.0 Easy
1.690.000	$	Avellaneda	Bs.As. G.B.A. Sur	78200	2016	fiat	Fiat Punto 1.4 Attractive Pack Top Uconnect
2.560.000	$	General San Martín	Bs.As. G.B.A. Norte	78360	2016	ford	Ford Focus Iii 1.6 S

Tabla 1. 10 primeras líneas de la base de publicaciones obtenidas mediante *Web scraping* de Mercado Libre

2.2.2. API de Mercado Libre

Por otro lado, se creó una aplicación de Mercado Libre para realizar consultas a la API (Application Programming Interface) con el objetivo de obtener mayor información y un mejor resultado de predicción. Una API define una sintaxis estandarizada que permite que una pieza de software se comunique con otra pieza de software, aunque puedan estar escritas en diferentes idiomas o estructuradas de manera diferente. La documentación de estas API generalmente describe rutas o puntos finales, como URL (Uniform Resource Locator) que puede solicitar, con parámetros variables, ya sea en la ruta de la URL o como parámetros GET. Mercado Libre proporciona en detalle la documentación de su API desde su sitio de desarrolladores. La respuesta de la API generalmente se devuelve en formato JSON (JavaScript Object Notation) o XML (Extensible Markup Language). JSON es mucho más popular en los tiempos modernos que XML, pero aún puede ver algunas respuestas XML, y el formato de respuesta de Mercado Libre [17].

A los campos mencionados arriba se sumaron los siguientes campos de interés:

- Cantidad de puertas
- Tipo de combustible

- Versión
- Tipo de Motor
- Transmisión

2.2.3. Preprocesamiento de datos: *Web scraping*

En primer lugar, se analizó la calidad de los datos del conjunto. Este conjunto contiene información de 43.736 publicaciones de Mercado Libre que se obtuvieron con el *web scraper* construido. A su vez, dado que el scrapper obtuvo publicaciones en pesos y dólares, se decidió retener únicamente aquellas en pesos y eliminar las publicaciones en dólares, achicando el conjunto de datos a 35.705 publicaciones.

Por otro lado, se observó si los datos tenían valores nulos y la única variable que presentó vacíos fue la variable 'Provincia' que contaba con nueve valores nulos. Por último, fue necesario reprocesar el título de manera tal de obtener el modelo del auto. Para ello, se buscó dentro del título partiendo del listado de modelo de la base privada. Tras crear esta variable, se observaron 3.836 modelos que no se encuentran en la base privada . Estos fueron eliminados resultando en un conjunto de datos final de 31.869 observaciones.

2.2.4. Preprocesamiento de datos: API

En primer lugar, se analizó la calidad de los datos de la base total, que cuenta con 72.890 publicaciones únicas con 27 características diferentes. Por otro lado, se observó si los datos tenían valores nulos; 23 de ellas cuentan con valores nulos, de las cuales 10 tenían al menos un 90 % de valores nulos por lo tanto fueron eliminadas del conjunto de datos.

A su vez, dado que se obtuvieron publicaciones en pesos y dólares, se conservaron únicamente aquellas en pesos, eliminando 10.454 publicaciones. Tras eliminar las publicaciones en dólares, se obtuvo una base de 49.058 publicaciones.

2.3. Análisis exploratorio

En esta sección, se hará un análisis descriptivo y exploratorio de las publicaciones extraídas de Mercado Libre para poder obtener los primeros *insights* sobre las características de las publicaciones.

Los datos de Mercado Libre fueron extraídos en el mismo período para ambas técnicas y se hubiera esperado recolectar un volumen de publicaciones similar; sin embargo, tal como se mencionó en la **Sección 2.2.3.** y **Sección 2.2.4.** las publicaciones obtenidas por

la API fueron 49.058 vs 35.705 publicaciones resultantes del webscraper. Al utilizar diferentes técnicas se nota un diferente alcance de las publicaciones pero con un distribución similar. Por lo tanto, usaremos la base de datos de la API para explorar el comportamiento de las diferentes variables.

2.3.1. Variable a predecir: precio

El precio, la variable dependiente a predecir con nuestros modelos, tiene un valor medio de 2.88 millones de pesos argentinos con un desvió estándar de 2.76 millones de pesos argentinos. En la Figura 2 podemos observar un histograma del precio con un 8,4 % de las publicaciones a más de un desvío estándar de la media.

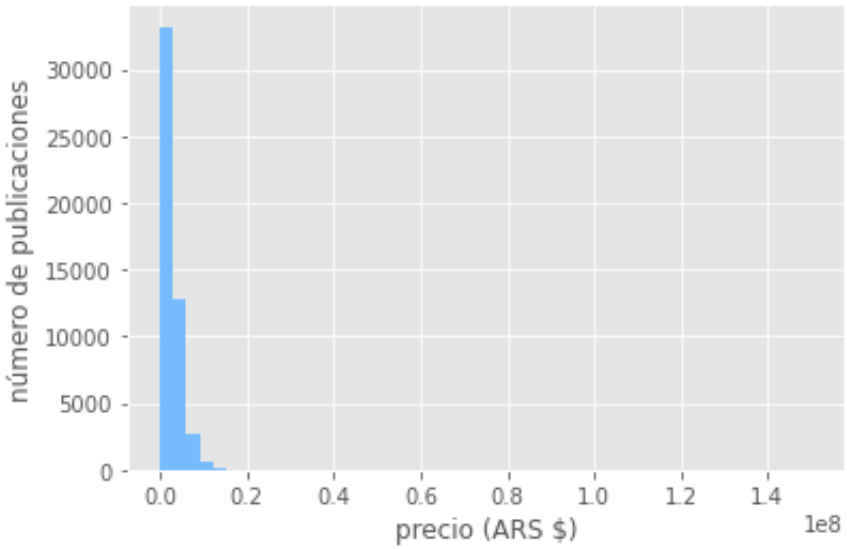

Figura 2. Histograma del Precio

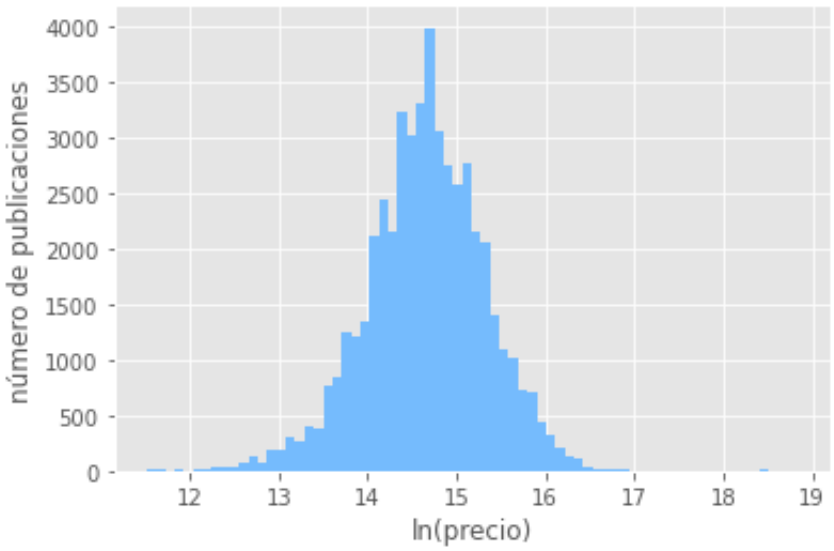

Figura 3. Histograma del Ln(Precio)

Por otro lado, se buscó entender cuál es la distribución del logaritmo del precio para comprender si el precio tiene una distribución log normal. Y en la Figura 3, se puede sospechar que lo tiene. A su vez, para confirmar la hipótesis de que el logaritmo del precio tiene una distribución normal, se realizó la prueba de Shapiro-Wilk. Esta prueba evalúa una muestra de datos y cuantifica la probabilidad de que los datos se hayan extraído de una distribución gaussiana, nombrada así por Samuel Shapiro y Martin Wilk. En la práctica, se cree que la prueba de Shapiro-Wilk es una prueba confiable de normalidad [22]. La función shapiro() de SciPy calculará el Shapiro-Wilk en un conjunto de datos determinado. La función devuelve tanto la estadística W calculada por la prueba como el p valor. Esta prueba tiene la hipótesis nula de que la variable analizada representa una distribución normal.

Tras realizar esta prueba en el logaritmo del precio y precio se puedo confirmar ambos test arrojaron un p valor menor a 5 % por lo tanto hay evidencia para rechazar la hipótesis nula de este test; y podemos confirmar que la variable logaritmo del precio no tiene un distribución normal.

Finalmente, se trazó un gráfico QQ, abreviatura de gráfico «cuantiles-cuantiles», es un tipo de gráfico que podemos usar para determinar si un conjunto de datos proviene o no de alguna distribución teórica. Muchas pruebas estadísticas suponen que un conjunto de datos sigue una distribución normal y, a menudo, se utiliza una gráfica QQ para evaluar si se cumple o no este supuesto. Aunque un gráfico QQ no es una prueba estadística formal, proporciona una manera fácil de verificar visualmente si un conjunto de datos sigue una distribución normal y, de no ser así, cómo se viola esta suposición y qué puntos de datos pueden causar esta violación. Podemos observar en la Figura 5 cómo luce un gráfico QQ para el precio y en la Figura 5 para el logaritmo del precio. Si bien el logaritmo del precio no es normal, la inspección del gráfico QQ nos muestra que esta variable es más parecida a una normal que el precio.

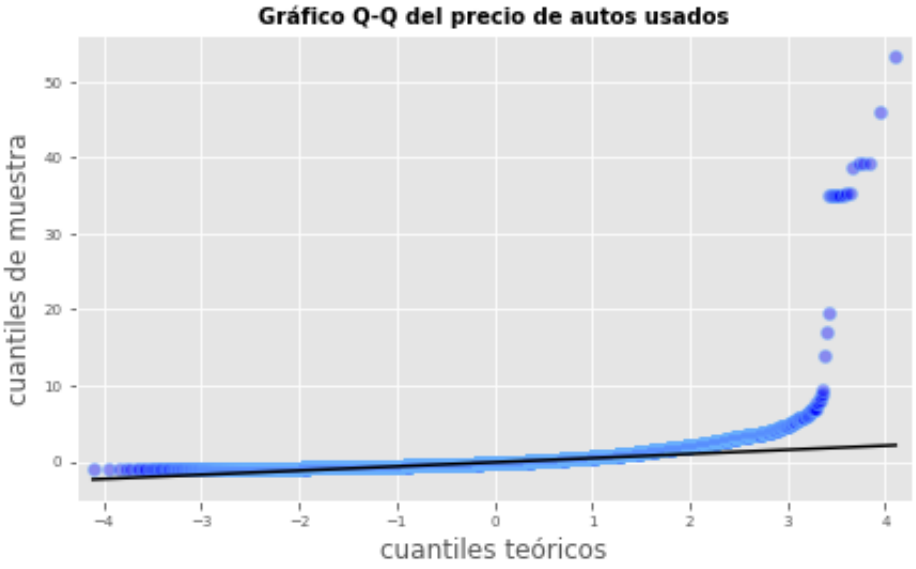

Figura 4. Gráfico Q-Q del precio de autos usados

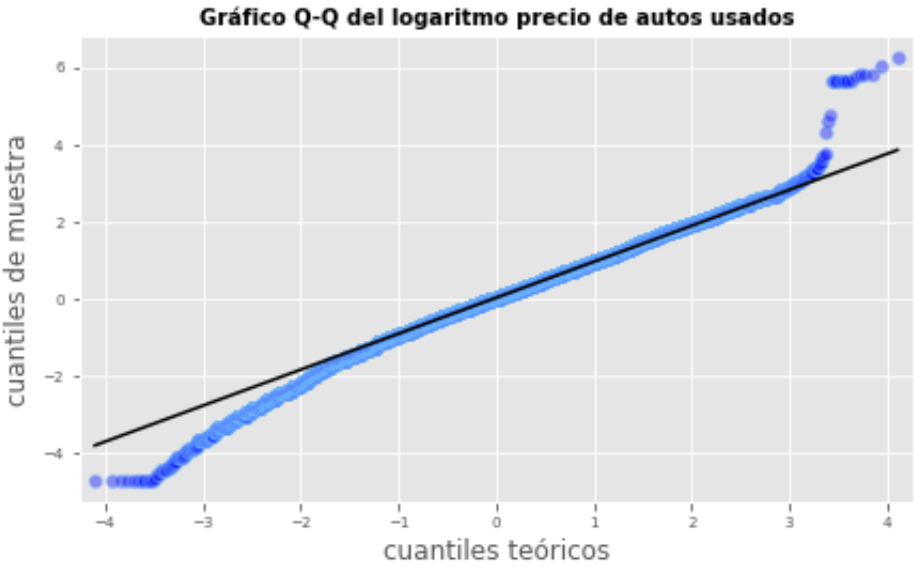

Figura 5. Gráfico Q-Q del Logaritmo del precio de autos usados

2.3.2. Kilometraje

Las publicaciones tienen un kilometraje promedio de 113.658 kilómetros, con un desvío estándar de 155.042 kilómetros. En la Figura 6 podemos observar un histograma de esta variable en donde se puede observar un 3,5 % de las publicaciones a más de un desvío estándar de la media. En la Figura 7, se puede observar que el conjunto de datos cuenta con publicaciones cuyo kilometraje supera los seis dígitos.

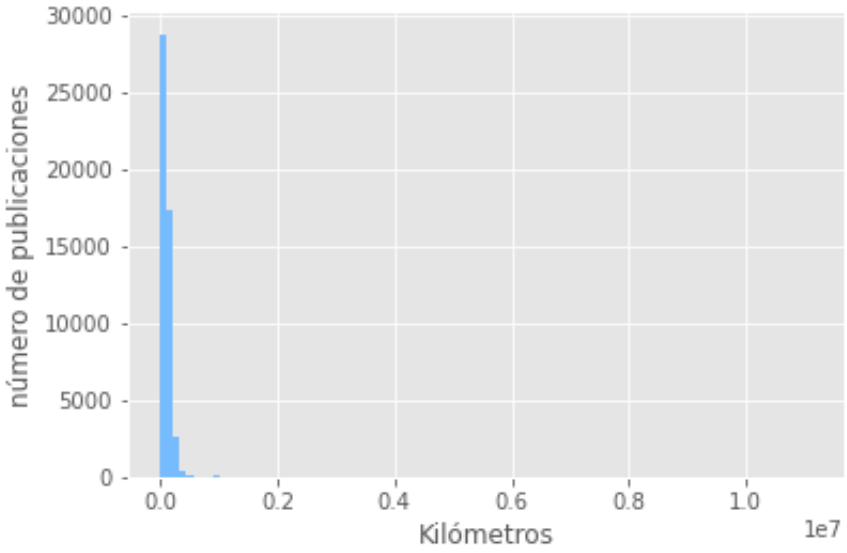

Figura 6. Histograma de kilómetros

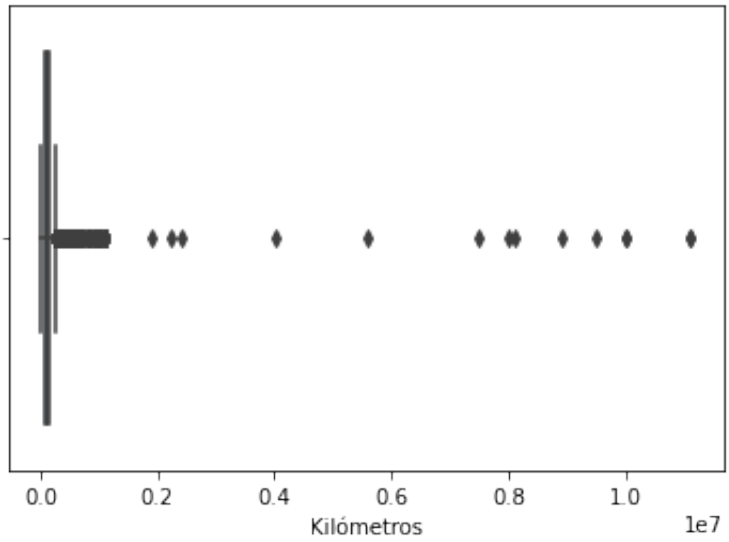

Figura 7. Boxplot de kilómetros

Con el objetivo de entender la calidad de estas publicaciones con la hipótesis de que son publicaciones con errores o poco representativas de la muestra que se busca analizar, se buscó mayor detalle en la publicación de Mercado Libre. En esta búsqueda se confirmo que aquellos autos de más de 999.999 kilómetros efectivamente son errores de publicación dado que en la imagen del tablero que muestra el kilometraje real del auto se observó un kilometraje menor o mismo en la descripción se indica que son autos 0 km. Con esta observación, se decidió eliminar aquellas publicaciones que superan los 999.999 kilómetros ya que no son representativos para estudiar el precio de mercado, se eliminaron 42 observaciones.

Con esta limpieza, el kilometraje promedio bajó 111.103 kilómetros, con un desvío estándar de 80.525 kilómetros. Se puede observar cómo se distribuyen los kilométros luego de esta corrección en la Figura 8.

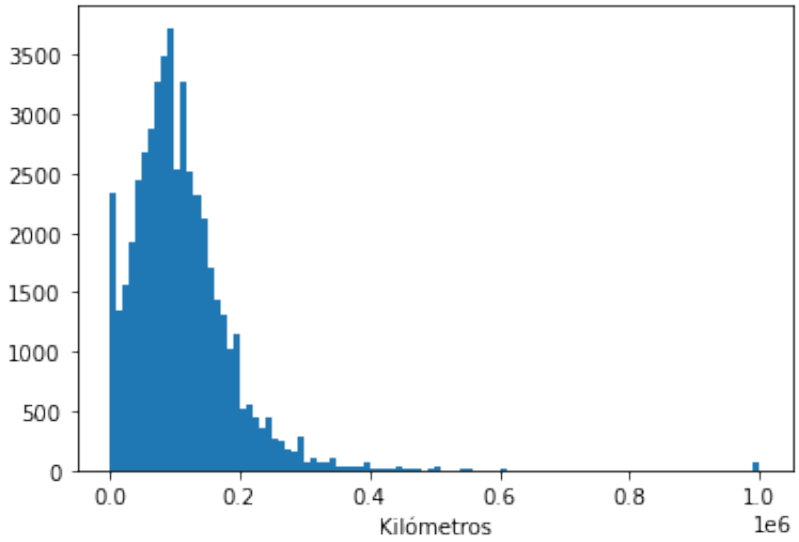

Figura 8. Histograma de kilómetros sin Outliers

2.3.3. Año

Las publicaciones tienen un año de fabricación promedio del 2.013, con un desvío estándar de 7 años. En la Figura 9 podemos observar un histograma de esta variable en donde observar un 7,2 % de las publicaciones a más de un desvío estándar de la media; con el auto más antiguo fabricado en el 1954 observable en la Figura 10año.

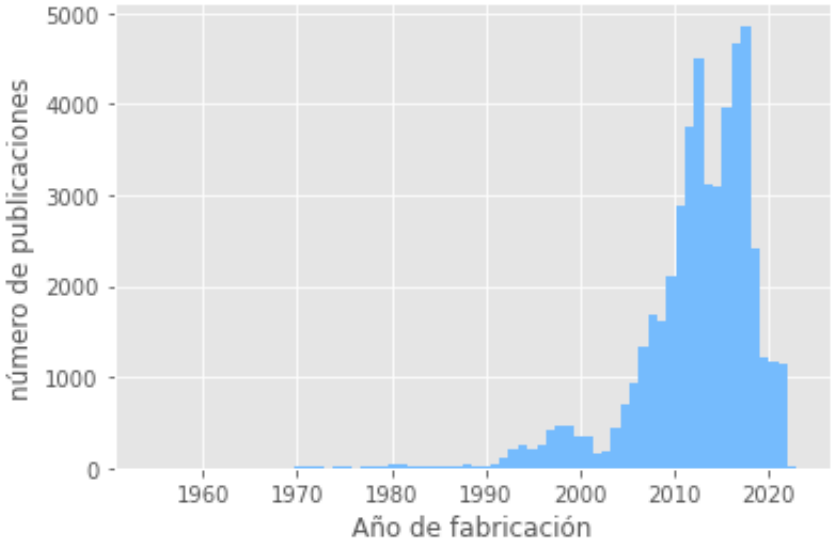

Figura 9. Histograma de año

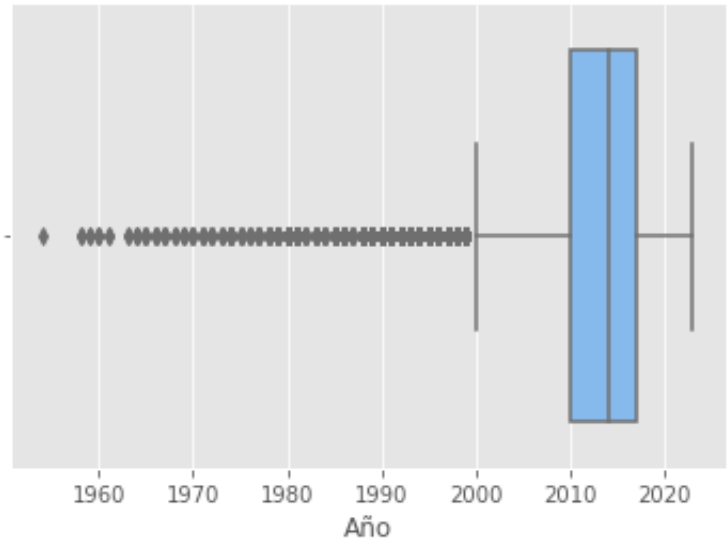

Figura 10. Boxplot de año

Al igual que para los kilómetros, se buscaron las publicaciones de año de fabricación más bajas. Y se observó que estas publicaciones eran autos en mal estado mecánico fuera de funcionamiento, por lo tanto se decidió solo trabajar con los autos fabricados de 1990 en adelante. A su vez, dado que la empresa privada trabaja con autos semi nuevos, las publicaciones de 1990 en adelante son más representativas de los autos que se quiere estudiar el precio de mercado. Se eliminaron del conjunto de datos 416 observaciones, reduciendo el desvío estándar a 5,8 años.

2.3.4. Variables numéricas y relación con el precio

Finalmente, si se observa cómo se comportan estas variables con el precio, es posible notar que el kilometraje tiene una correlación inversa al precio (Figura 11), mientras que el año tiene una correlación negativa al precio (Figura 12).

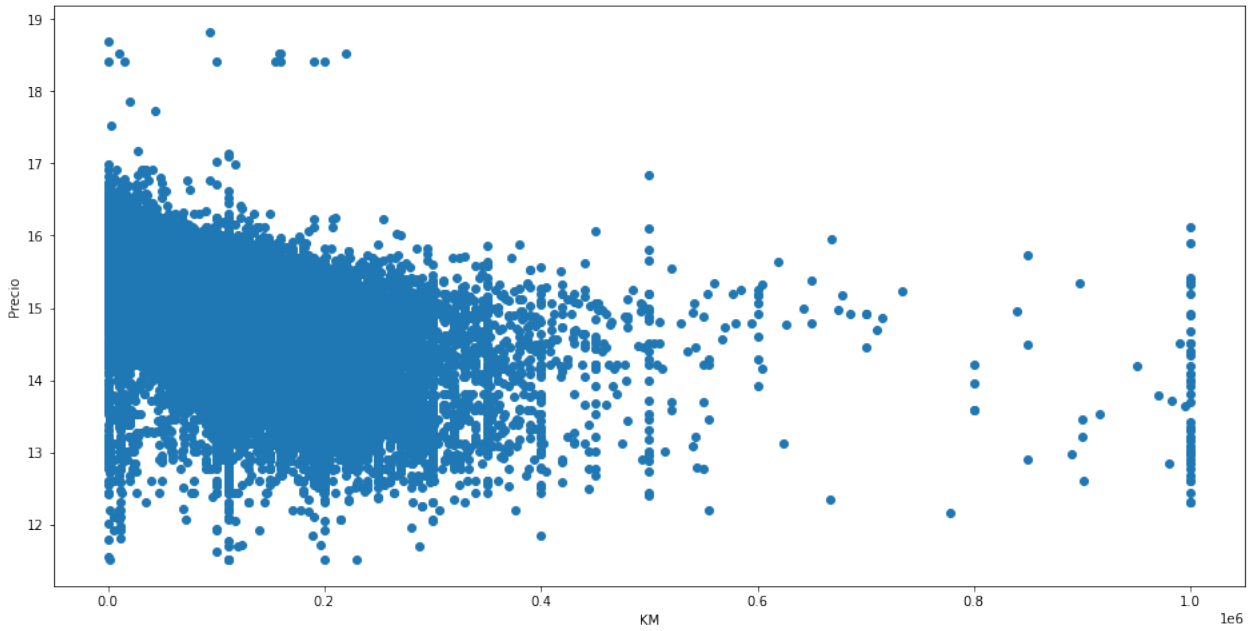

Figura 11. Relación entre kilometraje y logaritmo de precio

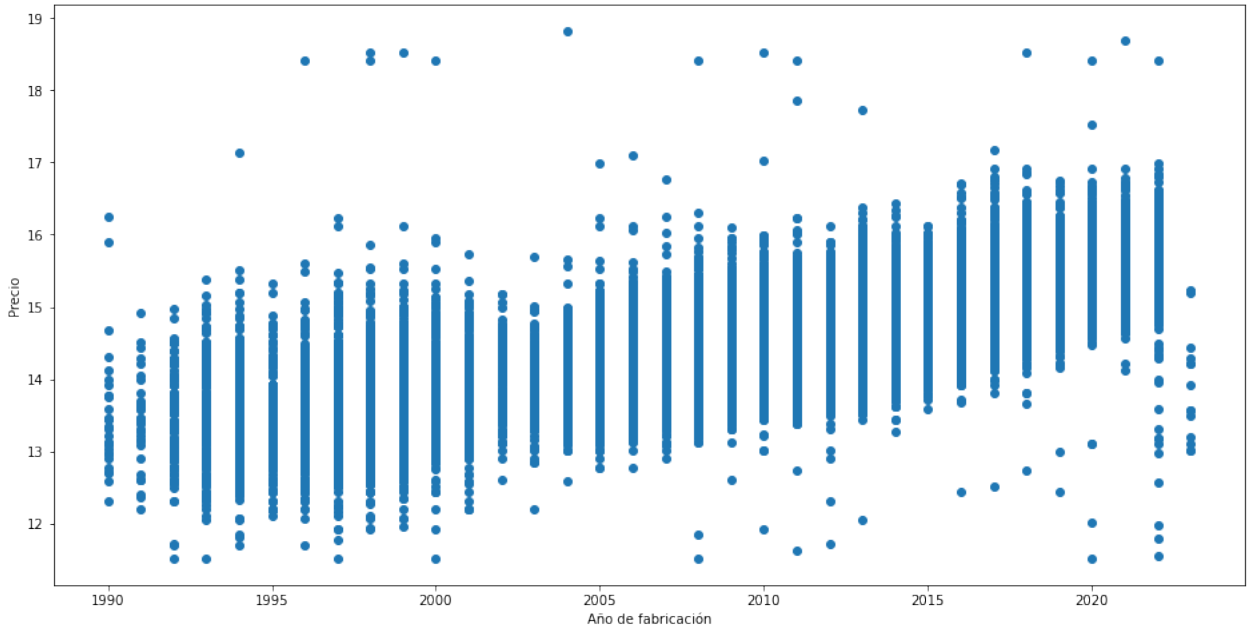

Figura 12. Relación entre el año y el logaritmo de precio

2.3.5. Variables categóricas

Por un lado, se dispone del detalle de la fabricación de los autos en las variables marca, modelo y versión.

La publicación de autos usados de Mercado Libre se encuentra representada por 23 marcas (representadas en la Figura 13). Sin embargo, solo 6 representan el 80 % de las

publicaciones. A su vez, las marcas tienen el promedio 5,3 modelos y los modelos tienen 11,5 versiones diferentes, cuyo resultado arroja 1.399 Marca-Modelo-Versión diferentes.

Figura 13. Ranking de marcas por publicaciones

Por ejemplo, si se toma el modelo Cruze II, su respectiva marca es Chevrolet, que cuenta con 8 versiones diferentes observables en la Tabla 2.

Marca	Modelo	Versión
Chevrolet	**Cruze II**	1.4 Lt 153cv
Chevrolet	**Cruze II**	1.4 Ltz 153cv
Chevrolet	**Cruze II**	1.4 Ltz At 153cv
Chevrolet	**Cruze II**	1.4 Ltz Plus 153cv
Chevrolet	**Cruze II**	1.4 Sedan At Ltz
Chevrolet	**Cruze II**	1.4 Sedan Lt
Chevrolet	**Cruze II**	1.4 Sedan Ltz
Chevrolet	**Cruze II**	1.4 Sedan Ltz Plus

Tabla 2. Versiones para un Chevrolet-Cruze II

Por otro lado, el detalle de la provincia y ciudad donde se encuentran estos autos, da como resultado el 26 % de las publicaciones en la Capital Federal.

2.4. One hot encoding

La base de datos contiene variables categóricas como, por ejemplo, la marca de los autos. Es por ello que se tuvo que recurrir a un método de codificación para convertir estas categorías no numéricas en un valor numérico para que puedan ser trabajadas en los modelos.

El método utilizado fue *One Hot Encoding* [25], donde cada bit representa una categoría posible. Si la variable no puede pertenecer a varias categorías a la vez, entonces solo un bit en el grupo puede estar "activado". Por lo tanto, una variable categórica con k categorías posibles se codifica como un vector de características de longitud k.

En la Figura 14, se puede observar distintos valores que toma la variable Marca y cómo quedaría esta variable una vez procesada con el método de One Hot Encoding.

Marca	Categorical #
Audi	1
Chevrolet	2
Citroen	3
Fiat	4

One Hot Encoding

marca_audi	marca_chevrolet	marca_citroen	marca_fiat
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Figura 14. One Hot Encoding de una variable categórica de 4 valores

2.5. Separación conjunto de entrenamiento, conjunto de testeo y evaluación de performance

Conocida la base de datos, se procedió a separarla en dos: conjunto de entrenamiento y conjunto de testeo. Esto se hizo con el objetivo de evaluar correctamente la performance de los modelos en datos desconocidos, el conjunto de testeo, y entrenar los modelos con el conjunto de entrenamiento (Figure 15). Con el conjunto de entrenamiento, se evaluarán los hiperparámetros mediante el aprendizaje supervisado por la variable a predecir, y posteriormente testear sus errores de predicción de la variable numérica en el conjunto de testeo.

La razón por la cual se decidió emplear esta técnica se debe a que es una de las maneras más simples de simular los comportamientos sobre los datos desconocidos, y su costo de cómputo es relativamente bajo. La desventaja a tener en cuenta es que tiene una dependencia del azar, es decir, de cómo se separan las submuestras dentro del conjunto de entrenamiento.

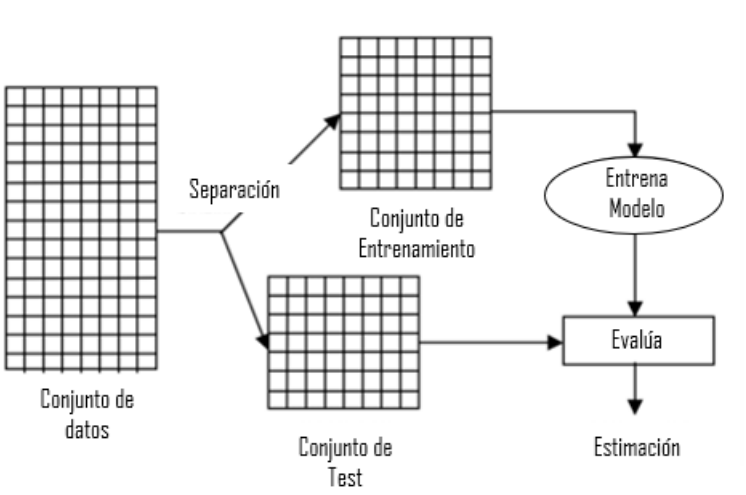

Figura 15. Separación de la base de datos en conjunto de entrenamiento y testeo

Con respecto a esta etapa, se tomaron dos decisiones para los modelos a entrenar:

- La separación de estos conjuntos se realizó de manera aleatoria, con el objetivo de disminuir los posibles sesgos dentro del conjunto de datos, y maximizar la performance de los modelos.
- Asignarle el 20 % del conjunto de datos al conjunto de testeo, dado que se cuentan con suficiente volumen de datos, y así poder mejorar el cómputo de la performance de los modelos.

2.6. Github: Repositorio de Tesis

A modo de contribución a la comunidad, se expusieron en un repositorio de GitHub (https://github.com/lucialopezwallace/tesis-mim) los códigos de esta tesis; compartiendo las técnicas de *web scrapping* utilizadas, las bases extraídas de Mercado Libre y los modelos utilizados con su respectivos hiperparámetros.

Los individuos interpretan y utilizan la información de Github sobre las acciones de otros. Se descubrió cuatro características clave del uso de plataformas colaborativas: generación de compromiso, mayor calidad del trabajo, generación de importancia para la comunidad y de relevancia personal. Estas características respaldan la colaboración, el aprendizaje y la gestión de la reputación en la comunidad. La transparencia en GitHub permite aprender de las acciones de otros desarrolladores; ser capaz de ver cómo alguien más codifica, a qué prestan atención los demás y cómo resolver problemas, respaldando el aprendizaje de la codificación y acceder a un conocimiento superior [7].

3. Metodología

3.1. Aprendizaje supervisado

Los problemas de aprendizaje automático pueden pertenecer a una de estas dos categorías: supervisado o no supervisado. Para un aprendizaje supervisado, cada observación de la medición del predictor xi, i=1, ..., n tendrá asociado una respuesta yi. El objetivo de dicho aprendizaje es el de ajustar un modelo que relacione la respuesta con la del predictor, con el objetivo de predecir con precisión la respuesta para futuras observaciones (predicción) o bien comprender mejor la relación entre la respuesta y los predictores (inferencia) [10].

Por el contrario, el aprendizaje no supervisado describe una situación de mayor complejidad donde para cada observación i= 1, ..., n, se observa un vector de medidas xi pero sin respuesta asociada yi. No es posible ajustar un modelo supervisado ya que no hay una variable de respuesta a predecir. La situación se denomina no supervisada ya que se carece de una variable de respuesta que pueda supervisar el análisis.

Dado que se dispone de precios de mercado según ciertas variables se abordó el problema con aprendizaje supervisado.

3.2. Problema de regresión

Las variables pueden ser caracterizadas cómo cualitativas o cuantitativas. Las variables cualitativas toman valores en una de K clases o categorías diferentes y hacen referencia a un problema de clasificación (ej. si una persona entra en default o no; la marca a la que pertenece un producto, entre otros); mientras que las variables cuantitativas toman un valor numérico y hacen referencia a los problemas de regresión (ej. precio de un inmueble, precio de una acción, edad, entre otros) [11].

Se abordará un problema de regresión dado que se busca predecir el valor de un auto con ciertas características.

3.3. Modelos

Con el objetivo de entrenar un modelo de regresión con buena capacidad de predicción, se utilizaron dos algoritmos de aprendizaje supervisado:

- Modelo de regresión lineal
- Modelo de XG-Boost

El modelo de regresión lineal actuó como modelo sencillo y de base para el análisis y luego se trabajó con el modelo XG-Boost buscando mejorar la predicción de dicho modelo.

Para estos dos algoritmos, se buscaron los hiperparámetros óptimos de cada uno, de forma tal de obtener la máxima capacidad predictiva, y se seleccionó aquel que logró el mayor grado de precisión.

3.3.1. Regresión lineal

La regresión lineal es uno de los algoritmos de aprendizaje automático más simples y comunes. Se usa comúnmente en métodos de investigación matemática, donde es posible medir los efectos predichos y modelarlos contra múltiples variables de entrada. Es un método de evaluación y modelado de datos que establece relaciones lineales entre variables que son dependientes e independientes. Este método modelaría así las relaciones entre las variables dependientes y las variables independientes desde el análisis y el aprendizaje hasta los resultados del entrenamiento actual [15].

A los efectos de este análisis, se trabajó con una regresión lineal múltiple ya que se dispone de más de una variable independiente. Para la implementación en Python, se utilizó scikit-learn. La librería Scikit-Learn proporciona una función llamada *LinearRegression* para hacer justamente este trabajo, y de esta forma ahorrarnos todos los cálculos matemáticos [19].

3.3.2. XGBoost

XGBoost es un algoritmo de la familia de los "boosting algorithms" muy eficiente en la reducción del sesgo y de las variantes de un modelo como así también es eficiente en memoria, rápido y de gran poder predictivo [9]. Particularmente implementa "gradiente boosting algorithm", donde se crean muchos árboles de decisión a partir de los datos de entrenamiento de manera secuencial y donde cada árbol utiliza información del árbol anterior. Puntualmente, cada nuevo árbol tiene el objetivo de explicar los residuos del modelo previo. Es por esto que se suele decir que son algoritmos que "aprenden despacio". Entrenar todos los árboles de forma simultánea es un problema muy complejo e imposible computacionalmente. Por lo tanto, se opta por entrenar sucesivamente uno por vez [6].

3.3.3. Optimización de hiperparámetros

La estrategia utilizada para encontrar los mejores hiperparámetros fue mediante una búsqueda aleatoria. Alternativamente, también se buscó optimizar los hiperparámetros; sin embargo, esta técnica no trajo mejores resultados de predicción. Se demostró que la búsqueda aleatoria es una forma mucho más eficiente en términos de error para nuestro

modelo. La metodología consiste en definir un rango de posibles valores para cada hiperparámetro y luego, aleatoriamente, seleccionar uno para cada uno. De esta forma quedan seleccionados los hiperparámetros correspondientes para un posible modelo. Repitiendo esto varias veces, quedan armados distintos modelos a evaluar.

En la Tabla 3, se puede observar una muestra de los hiperparámetros optimizados en nuestro modelo XGBoost.

Hiperparametro	Descripción	Rango
max_depth	Máxima profundidad de los arboles	(0;∞]
eta	Proporción que aprende de cada árbol	[0;1]
gamma	Mínima reducción del error necesaria en una hoja para generar una nueva	[0;∞]
colsample_bytree	Porcentaje de columnas elegidas (al azar) para construir un arbol	(0:1]
subsample	Porcentaje de observaciones elegidas (al azar) para construir el arbol	(0:1]
min_child_weight	Cantidad mínima exigida de observaciones por hoja	[0;∞]
nrounds	Cantidad de árboles a construir	(0;∞]

Tabla 3. Hiperparámetros de XGBoost a optimizar

3.4. Métrica de evaluación de modelos

Para poder medir la performance de cada uno de los modelos descriptos en la **Sección 3.1. Modelos**, y así poder elegir aquel de mayor performance, se utilizaron las métricas de raíz del error cuadrático medio (RMSE), error medio absoluto (MAE) y error cuadrático medio (MSE).

Como métrica principal de análisis y comparación de la performance de los modelos, se utilizó el MAE, considerando que esta métrica es más natural y menos ambigua que RMSE, tal como se debate en [24]. RMSE es menos apropiada por ser una función de tres características de un conjunto de errores, en lugar de uno (el error promedio). Finalmente, estas métricas serán calculadas utilizando el módulo sklearn.metrics.

3.4.1. Error medio absoluto

El error absoluto medio o "MAE" es la media de la diferencia absoluta entre los puntos de datos reales, y la salida predicha. Es una medida robusta a outliers; sin embargo, no ayuda a entender si se trata de una sobreestimación o subestimación de los datos.

3.4.2. Error cuadrático medio

El error cuadrático medio o “MSE” es la media de la diferencia entre los puntos reales de datos y la salida predicha, al cuadrado. Este método penaliza más las diferencias mayores y es el estándar en los problemas de regresión porque es la métrica que optimiza el algoritmo de cuadrados mínimos. A diferencia de MAE, no es robusta a valores *outliers*.

3.4.3. Raíz del error cuadrático medio

Dado que el MSE se encuentra elevado al cuadrado, el valor no puede ser comparado con las mismas unidades del valor real. El error cuadrático medio o ”RMSE” es la raíz de la media de los errores elevados al cuadrado, es decir, el MSE. Es una medida que ayuda a comparar qué tan dispersos están los errores. Suele ser mayor o igual a MAE.

4. Resultados

4.1. Descripción del análisis predictivo

Una vez realizado el análisis descriptivo e identificados los mejores insights y ajustes a realizar con el conjunto de datos, se procedió a realizar el análisis predictivo. Dicho análisis incluye la anticipación a futuros eventos/tendencias y minería de datos. Para esta sección, los objetivos del análisis fueron:

- Definir la variable del conjunto de datos que se va a predecir.
- Mediante enfoque de aprendizaje automático, entrenar diferentes modelos que predigan dicha variable, logrando los mayores niveles de certeza posibles.
- Medir resultados con mínimo de MAE.
- Asegurarse de emplear las técnicas correctamente, sin riesgo a cometer data leakage [26].

4.2. Modelo de regresión

Tal como se mencionó en la **Sección 3.1.1. Modelo de Regresión Lineal**, se trabajó con este modelo como base de nuestra predicción. Se analizarán los resultados tanto para el conjunto de datos obtenidos con el *web scapper* como el que se obtuvo con la API. Por otro lado, mediremos el modelo en la variable precio y luego en el logaritmo del precio para entender si esta transformación ayuda a reducir el error de predicción.

4.2.1. Resultados con datos del *web scraper*

El resultado de este modelo entrenado con los datos recolectados del *web scraper* fue de 990 mil pesos argentinos (Tabla 4). Mientras que entrenando el modelo de manera tal de predecir el logaritmo del precio, se redujo ese error a 880 mil pesos argentinos, alrededor de cien mil pesos de mejora (Tabla 5).

Por otro lado, el modelo que predijo el logaritmo del precio también arrojó un menor error al evaluarlo en el conjunto de datos de precios de la base privada.

Métrica	Modelo Regresion Lineal	Error evaluando Precio Conjunto de Datos Privado
Mean Absolute Error (miles ARS)	991	971
Mean Squared Error (miles)	9.201.689.968	1.423.226.784
Root Mean Squared Error (miles ARS)	3.033	1.193

Tabla 4. MAE, MSE y RMSE obtenidos en el modelo de regresión lineal

Métrica	Modelo Regresion Lineal ln(Precio)	Error evaluando ln(Precio) Conjunto de Datos Privado
Mean Absolute Error (miles ARS)	882	817
Mean Squared Error (miles)	9.166.898.967	1.153.609.424
Root Mean Squared Error (miles ARS)	3.028	1.074

Tabla 5. MAE, MSE y RMSE obtenidos en el modelo de regresión lineal del ln(Precio)

4.2.2. Resultados con datos de la API

El resultado de este modelo entrenado con los datos recolectados de la API de Mercado Libre; base más robusta y completa no solo en términos de características sino también en tamaño del conjunto de datos tal cómo se menciona en la **Sección 2.2.2. API de Mercado Libre** arrojó un error de 631 mil pesos argentinos (Tabla 6). Mientras que entrenando el modelo de manera tal de predecir el logaritmo del precio, se redujo ese error a 484 mil pesos argentinos, al rededor de 147 mil pesos de mejora (Tabla 7). A su vez, este modelo también supera a los resultados del modelo de regresión lineal entrenados con el *web scraper* en 398 mil pesos argentinos; mostrando ser un conjunto de datos mejor para la predicción del precio de autos.

Métrica	Modelo Regresion Lineal	Error evaluando Precio Conjunto de Datos Privado
Mean Absolute Error (miles ARS)	631	574
Mean Squared Error (miles)	4.664.058.074	1.258.248.431
Root Mean Squared Error (miles ARS)	2.159	1.121

Tabla 6. MAE, MSE y RMSE obtenidos en el modelo de regresión lineal a partir del conjunto de datos de la API

Métrica	Modelo Regresion Lineal ln(Precio)	Error evaluando ln(Precio) Conjunto de Datos Privado
Mean Absolute Error (miles ARS)	484	331
Mean Squared Error (miles)	8.661.092.723	274.555.797
Root Mean Squared Error (miles ARS)	2.942	523

Tabla 7. MAE, MSE y RMSE obtenidos en el modelo de regresión lineal del ln(Precio) a partir del conjunto de datos de la API

4.3. Modelo de XGBoost

Se trabajó con un modelo de XGBoost como segundo modelo de machine learning, descripto en la **Sección 3.1.3. Modelo de XGBoost**.

Acá se mostrarán dos resultados: en primer lugar, los que se obtuvieron utilizando el conjunto de datos más acotado que se obtuvo del web scraper y, en segundo lugar, se midió la mejora en la *performance* al utilizar un conjunto de datos compuesto por un mayor detalle de las publicaciones, el cual se obtuvo de la aplicación de Mercado Libre.

Dado que se trata de un modelo de mayor complejidad con la posibilidad de iterar con múltiples hiperparámetros, se procedió a entrenarlo con los hiperparámetros óptimos. Utilizando técnicas de optimización de hiperparámetros descriptas en la **Sección 3.1.3. Modelo de XGBoost** de forma tal de maximizar el MAE.

4.3.1. Resultados con datos del *web scraper*

Acá observaremos como respondió el modelo con los hiperparámetros optimizados al conjunto de datos que se contruyó con los datos del *web scrapper*. Este modelo busca reducir el error del precio de autos, nuestra variable target. En la tabla 8 vemos un error elevado superior a un millón de pesos argentinos.

Métrica	Modelo XGBoost
Mean Absolute Error (miles ARS)	739,2
Mean Squared Error (miles)	1.771.927.191
Root Mean Squared Error (miles ARS)	1.331,1

Tabla 8. MAE, MSE y RMSE obtenidos en el modelo XGBoost, con datos del *web scrapper*

4.3.2. Resultados con datos de la API

A su vez, se entrenó otro modelo de XGBoost con los datos obtenidos de la API con el objetivo de predecir el precio de mercado de autos usados, la variable target. Como se puede ver en los resultados de la Tabla 9, este segundo modelo entrenado con el conjunto de datos obtenido con la API de Mercado Libre obtuvo una precisión superior, con un MAE de 276,2 miles de pesos argentinos, superando a la performance del modelo de regresión lineal y a la del scraper.

Métrica	Modelo XGBoost	Resulado Y_privados
Mean Absolute Error (miles ARS)	271,4	371,4
Mean Squared Error (miles)	172.487.130,8	260.162.950,6
Root Mean Squared Error (miles ARS)	413,0	507,6

Tabla 9. MAE, MSE y RMSE obtenidos en el modelo XGBoost, con datos de la API

4.4. Importancia de variables

Los modelos de árboles tienen la particularidad de permitir computar la importancia de cada uno de las características de la base de datos respecto a la variable a predecir. Se utilizará esta característica del modelo XGBoost para analizar las variables de entrenamiento.

La importancia proporciona una puntuación que indica cuan útil o valiosa fue cada característica en la construcción de los árboles de decisión dentro del modelo. Cuanto más se utilice un atributo para tomar decisiones clave con árboles de decisión, mayor será su importancia relativa. Esta importancia se calcula explícitamente para cada atributo en el conjunto de datos, lo que permite clasificar los atributos y compararlos entre sí [14].

El concepto de importancia variable es una selección de características implícita realizada por *Random Forest* con una metodología de subespacio aleatorio, y se evalúa mediante el índice de criterio de impureza de Gini. El índice de Gini es una medida del poder de predicción de variables en regresión o clasificación, basado en el principio de reducción de impurezas [23]; es no paramétrico y por lo tanto no se basa en datos pertenecientes a un tipo particular de distribución.

Para dividir un nodo binario de la mejor manera, se debe maximizar la mejora en el índice de Gini. En otras palabras, un Gini bajo (es decir, una mayor disminución de Gini) significa que una función de predicción en particular juega un papel más importante en la partición de los datos en las dos clases. Por lo tanto, el índice de Gini se puede utilizar para clasificar la importancia de las características para un problema de regresión y clasificación [21] .

En la Figura 16 podemos observar las variables más relevantes del modelo. Es interesante destacar que dentro de las cinco variables más representativas se encuentran las variables Motor y Versión. Estas dos variables no se encontraban en el conjunto de datos del *web scraper* y fueron sumadas al utilizar la API de Mercado Libre. El sumar estas variables ayudó a reducir el error de predicción de la API versus el *web scraper*.

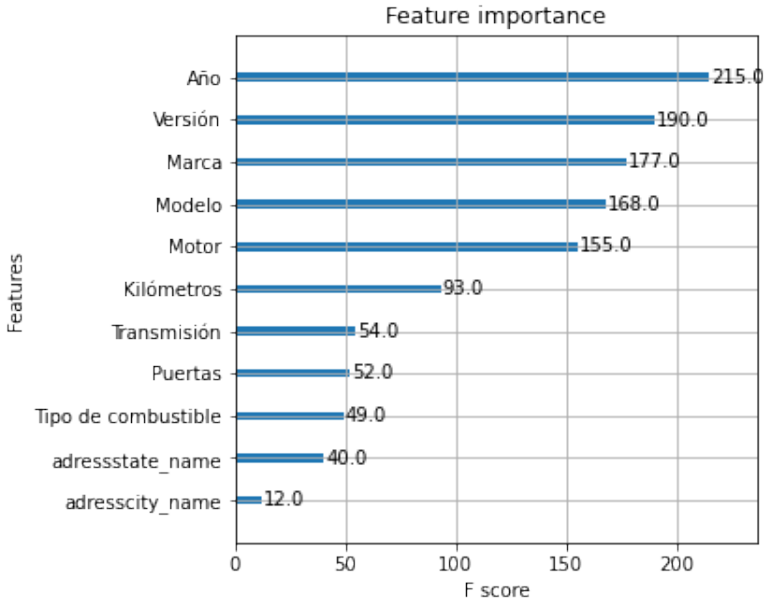

Figura 16. Importancia de variables del conjunto de datos respecto de la variable a predecir

4.5. Análisis de resultados

Tras haber analizado distintos modelos, se puede concluir que XGBoost es el modelo que más ayuda a aproximarse al precio de los autos y que a medida que se sume más variables se logrará reducir el error acercándose al precio correcto.

Con el objetivo de contrarestar los resultados con estudios similares, se contruyó la métrica error porcentual absoluto medio (MAPE) usado en el paper *Entity Embeddings of Categorical Variables* [12] para medir la performance de diversos modelos de aprendizaje automático para predecir las ventas de de una tienda europea. En la ecuación 1 se puede observar cómo se contruye esta métrica. En el estudio citado se observa un MAPE de 0.122 para el modelo de *gradient boosting* comparable con nuestro modelo de XGBoost que arrojó un MAPE de 0.078.

$$MAPE = \left| \frac{y - y_{predict}}{y} \right| \tag{1}$$

Por otro lado, cabe destacar que con la API se pudieron obtener aún más variables de las publicaciones, pero no fueron utilizadas ya que esas variables no se encontraban en la base privada. Sin embargo, para comprender si estas variables son relevantes para el análisis se ajustó el modelo XGBoost agregando estas variables y se obtuvieron resultados superadores a los vistos en la Tabla 9; dichos resultados puede observarse en la 10.

Métrica	Modelo XGBoost
Mean Absolute Error (miles ARS)	257,0
Mean Squared Error (miles)	147.893.721,7
Root Mean Squared Error (miles ARS)	381,8

Tabla 10. MAE, MSE y RMSE obtenidos en el modelo XGBoost, con datos de la API + extra features

4.6. Validación con especialista

Más allá de la performance arrojada por los modelos trabajados, se buscó validar el modelo con un especialista de pricing del negocio de autos usados. Para ello se construyó una muestra del conjunto de datos de la empresa privada con la siguiente metodología:

1. Se tomó 20 muestras aleatorias del conjunto de datos privados, con detalle de las características del auto.
2. Se sumó el dato del precio: predicho por el modelo entrenado y utilizado por la empresa privada.
3. Aleatoriamente, se alternaron esos 2 precios en 2 columnas, llamándolos **precio 1** y **precio 2**.
4. Se colocó una columna adicional para que el especialista seleccione cuál es el precio que más se asemeja al precio de mercado.

En la Tabla 11 podemos observar 10 observaciones del resultado de dicha tabla que se le entregó al especialista.

id	Marca	Modelo	Kilómetros	Año	Elección	Δ (Predicción - Modelo)	Δ % (Predicción - Modelo)
181241	Nissan	Kicks	16.599	2020	Real	959.054	23%
175599	Ford	Ecosport	20.350	2021	Real	- 234.956	-5%
179482	Peugeot	208	29.624	2020	Modelo	328.523	10%
179255	Volkswagen	Polo	37.591	2018	Real	- 349.928	-11%
178826	Chevrolet	S10	44.672	2017	Modelo	- 718.732	-12%
175566	Ford	Focus III	62.681	2015	Real	- 49.328	-2%
137006	Jeep	Renegade	70.661	2017	Real	536.980	14%
151885	Renault	Captur	74.292	2017	Real	176.589	5%
173926	Chevrolet	Cruze II	84.831	2017	Modelo	346.393	12%
159812	Chevrolet	Onix	98.077	2014	Modelo	- 150.976	-9%

Tabla 11. Muestra aleatoria de resultados

De la muestra, hubo 9 autos cuyo precio de mercado elegido fue el predicho por el modelo trabajado, es decir, un 45 % de precisión en los resultados. A su vez, para probar

la significación estadística, se realizó una prueba exacta de Fisher, la cual se utiliza en el análisis de tablas de contingencia y, en la práctica, se emplea cuando los tamaños de muestra son pequeños, como lo es en nuestro caso. La prueba de Fisher es el test exacto utilizado cuando se quiere estudiar si existe asociación entre dos variables cualitativas, es decir, si las proporciones de una variable son diferentes dependiendo del valor que adquiera la otra variable. En la gran mayoría de casos, el test de Fisher se aplica para comparar dos variables categóricas con dos niveles cada una (tabla 2x2) [8].

Particularmente, a través de la prueba de Fisher se analizó la tabla de contingencia (Tabla 12) de la elección del especialista con el fin de determinar si la elección del precio es independiente del origen del mismo (real vs modelo). El resultado de este test arrojo un p valor de 0.752; por lo tanto, si tomamos el nivel de significación en 0,05, no podemos rechazar la hipótesis nula ya que el valor p está por encima de 0,05. Por lo tanto, aunque puede parecer que hay una diferencia entre la elección y el origen de los datos, con estos datos, no hay evidencia suficiente para indicar que el origen de datos afecta la elección del precio. Con una muestra más grande, es probable que pueda demostrar una diferencia.

	Real	Modelo
Elegido	11	9
No elegido	9	11

Tabla 12. Tabla de contingencia

4.7. Estimación del impacto económico

En conjunto con la recomendación de negocio, se procedió a cuantificar el valor económico de implementar este algoritmo para predecir el precio de mercado.

Como primer paso, se observaron las predicciones de precio para los 2.230 autos en stock de autos a analizar.

Luego, se cuantificaron solo las diferencias mayores/menores a un 5 %, lo que representa un cambio de precios para el 80 % del stock, ya que solo el 20 % restante se encontraba con una dispersión de 5 %. En la Tabla 13 se puede observar en detalle esta distribución de estos gaps.

Buckets	# Autos	Share [%]
>30%	375	17%
[20% ; 30%]	361	16%
[10% ; 20%]	434	19%
[5% ; 10%]	231	10%
[0% ; 5%]	231	10%
[-5% ; 0%]	216	10%
[-10% ; -5%]	150	7%
[-20% ; -10%]	174	8%
[-30% ; -20%]	44	2%
<-30%	21	1%
Total general	**2.238**	**100%**

Tabla 13. Diferencias de las predicciones vs precio actual

Finalmente, aplicando estos nuevos precios el nuevo stock pasaría a tener una valoración incremental de 1,3 millones de USD, equivalente a un aumento de 588 USD por auto. Este aumento se debe a que, tal como se observa en la Tabla 14, el modelo desarrollado arroja una recomendación de aumento de precios para el 63 % del stock. Y al ser este efecto mayor al de una baja de precio, se obtiene una ganancia.

Buckets	# Autos	Share [%]
Aumentar el precio	1.402	63%
Reducir el precio	389	17%
Mantener	447	20%

Tabla 14. Propuesta de cambio de precios del stock

A esta ganancia bruta, se debería restar el costo del recurso que gestione y controle esta nueva herramienta; o incluso implementarla en alguna plataforma de gestión.

5. Conclusión

5.1. Recomendaciones para el negocio

Tras el análisis realizado en el Capítulo 4, se encontró que, mediante el uso de técnicas de machine learning, fue posible medir el precio de mercado de los autos usados a partir de variables públicas que describen las principales características de un auto. Esto sugiere que hay una oportunidad a la hora de definir el precio del stock, mejorar la estrategia de precio para la compra y venta de autos, y maximizar ingresos.

Tal como se investigó con los managers del equipo de estrategia de esta empresa, la predicción del precio es el corazón del negocio con impacto directo en el estado de resultados financieros de la empresa y el crecimiento futuro de la empresa.

La recomendación de negocio que se propone es revisar el precio del stock utilizando técnicas de aprendizaje automático que tengan como *input* publicaciones de autos usados del principal e-commerce del país y referente de los precios del mercado. Para complementar la visión del especialista, se propone al negocio realizar una prueba piloto con estos nuevos precios para entender qué tan atractivos son para el mercado. Para la construcción de este piloto se sugieren los siguientes pasos:

1. Buscar en el stock marcas-modelos-versión que tengan al menos 4 autos en stock.
2. Seleccionar aleatoriamente un 10 % de aquellas marcas-modelo-version que cumplan con el punto 1.
3. Asignar el nuevo precio al 50 % de las marcas-modelo-version elegidas en el punto 2.
4. Obervar el comportamiento de estas publicaciones.

Finalmente, será interesante medir el piloto en dos aspectos: popularidad y rentabilidad. Llamando popularidad al interes que se recibe por estos nuevos precios; en donde se puede medir según la cantidad de visitas que reciben. Mientras que la rentabilidad será medida en entender cuantos días se demora en encontrar un comprador a este precio, si es que se encuentra.

5.2. Limitaciones y futuras posibles mejoras

En línea con la recomendación de negocio, se debe hacer foco en las limitaciones que trae el análisis de la presente tesis. Existen oportunidades para mejorar el análisis predictivo y prescriptivo.

En primer lugar, los datos fueron extraídos con un solo visto. Este análisis se podría complementar con vistos diarios para entender qué publicaciones dejan de estar vigentes porque ya fueron vendidos los autos y de esta manera entender la rotación. El dato de la rotación es un dato relevante para entender la elasticidad de los diferentes precios.

En segundo lugar, se toma los precios publicados como precios de mercado y asumimos que los autos se transaccionan a dicho valor. Sin embargo, puede suceder que el precio publicado difiera al precio transaccionado ya que por fuera se dan negociaciones entre comprador y vendedor llegando a otro precio de transacción.

En tercer lugar, para complementar este análisis se podría explorar la evolución del precio de estas publicaciones para detectar cómo se mueven los precios en el mercado. Los precios de autos usados en Argentina se mueven con el aumento del dólar; pero, frente a una devaluación, la apreciación de los autos no siempre es inmediata sino que se ajusta de manera gradual. Es por ello que, frente a escenarios de devaluación, es importante entender cómo evoluciona este aumento del precio del mercado para evitar tener un stock devaluado.

En cuarto lugar, esta tesis abordo el error de manera absoluta; sin embargo, sería valioso entender cómo se comporta el MAE, MSE y RMSE a nivel marca, modelo y versión. De esta manera se podría comprender donde se encuentran los principales desvíos y buscar nuevas features que ayuden a reducir el error. Por otro lado, con respecto a la manera de entender el error se a modo de generar una métrica más intuitiva del error se propone sumar una métrica que mida el MAE sobre la mediana del precio de mercados para entender que desvio porcentual dejan los diferentes modelos analizados. Finalmente, sería intenresante buscar estudios similares de predicción de precios en autos usados dentro de América Latina (similares a los realizados en la Republica de Mauricio [20]) para comparar con los resultados arrojados en esta tesis.

Finalmente, dado que se dispone de la información de los vendedores, se podría hacer la estimación por vendedor con el objetivo de entender si alguno de ellos tiene un precio más alto o bajo como parte de su estrategia de posicionamiento de mercado.

5.3. Conclusiones finales

Como conclusión final de la tesis, y tras haber obtenido resultados aplicables por medio de técnicas de *Machine Learning*, se puede confirmar que utilizando datos del principal marketplace de Argentina, Mercado Libre, se puede obtener el precio de mercado de autos usados.

A su vez, la industria de autos usados es dinámica y los precios se ven afectados por situaciones macroeconómicas del país por lo cual es importante no perder de vista lo que sucede en el mercado; no solo para generar rentabilidad en el negocio sino también para ubicarse de manera estratégica frente a los principales competidores.

Referencias

[1] Karin von Abrams. *Global Ecommerce Forecast 2021*. eMarketer, 2021.

[2] Azoth Analytics. *Global Used Car Market (2021 Edition)*. Research y markets, 2021.

[3] Cox Automative. *2021 COX AUTOMOTIVE CAR BUYER JOURNEY STUDY OVERVIEW*. Cox Automative, 2021.

[4] Bazaarvoice. *Introduction to Carvana*. URL: https://investors.carvana.com/~/media/Files/C/Carvana-IR/documents/intro-to-carvana-march-2019.pdf.

[5] Anish Chapagain. *Hands-On Web Scraping with Python: Perform advanced scraping operations using various Python libraries and tools such as Selenium, Regex, and others*. Packt Publishing Ltd, 2019, pág. 8.

[6] Tianqi Chen y Carlos Guestrin. "Xgboost: A scalable tree boosting system". En: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, págs. 785-794.

[7] Laura Dabbish y col. "Social coding in GitHub: transparency and collaboration in an open software repository". En: *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 2012, págs. 1277-1286.

[8] Ronald A Fisher. "On the interpretation of χ 2 from contingency tables, and the calculation of P". En: *Journal of the Royal Statistical Society* 85.1 (1922), págs. 87-94.

[9] Jerome H Friedman. "Greedy function approximation: a gradient boosting machine". En: *Annals of statistics* (2001), págs. 1189-1232.

[10] James Gareth y col. *An introduction to statistical learning: with applications in R*. Spinger, 2013, págs. 26-27.

[11] James Gareth y col. *An introduction to statistical learning: with applications in R*. Spinger, 2013, págs. 28-29.

[12] Cheng Guo y Felix Berkhahn. "Entity embeddings of categorical variables". En: *arXiv preprint arXiv:1604.06737* (2016).

[13] Olivier Hanoulle. *The online boom in used-car sales*. Roland Berfer, 2021.

[14] Trevor Hastie y col. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009, págs. 367-369.

[15] Dastan Maulud y Adnan M Abdulazeez. "A review on linear regression comprehensive in machine learning". En: *Journal of Applied Science and Technology Trends* 1.4 (2020), págs. 140-147.

[16] Ryan Mitchell. *Web scraping with Python: Collecting more data from the modern web*. .º'Reilly Media, Inc.", 2018, pág. 166.

[17] Ryan Mitchell. *Web scraping with Python: Collecting more data from the modern web*. .º'Reilly Media, Inc.", 2018, pág. 176.

[18] Jay M Patel. *Getting Structured Data from the Internet: Running Web Crawlers/Scrapers on a Big Data Production Scale*. Springer, 2020, pág. 37.

[19] Fabian Pedregosa y col. "Scikit-learn: Machine learning in Python". En: *the Journal of machine Learning research* 12 (2011), págs. 2825-2830.

[20] Sameerchand Pudaruth. "Predicting the price of used cars using machine learning techniques". En: *Int. J. Inf. Comput. Technol* 4.7 (2014), págs. 753-764.

[21] Alessia Sarica, Antonio Cerasa y Aldo Quattrone. "Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review". En: *Frontiers in aging neuroscience* 9 (2017), pág. 329.

[22] Samuel Sanford Shapiro y Martin B Wilk. "An analysis of variance test for normality (complete samples)". En: *Biometrika* 52.3/4 (1965), págs. 591-611.

[23] Carolin Strobl, Anne-Laure Boulesteix y Thomas Augustin. "Unbiased split selection for classification trees based on the Gini index". En: *Computational Statistics & Data Analysis* 52.1 (2007), págs. 483-501.

[24] Cort J Willmott y Kenji Matsuura. "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance". En: *Climate research* 30.1 (2005), págs. 79-82.

[25] Alice Zheng y Amanda Casari. *Feature engineering for machine learning: principles and techniques for data scientists*. .º'Reilly Media, Inc.", 2018, pág. 78.

[26] Alice Zheng y Amanda Casari. *Feature engineering for machine learning: principles and techniques for data scientists*. .º'Reilly Media, Inc.", 2018, pág. 93.