Sistema de recomendación de precios para venta de vehículos usados.

Esteban Jaramillo.

Sebastián Giraldo G.

Introducción

El mercado de vehículos usados ha experimentado un crecimiento constante en los últimos años. Sin embargo, tanto compradores como vendedores enfrentan desafíos a la hora de materializar los negocios, relacionados con el alto componente de subjetividad y las asimetrías de información presentes en este mercado.

Por el lado de los compradores, puede ser difícil encontrar el vehículo que mejor se adapte a sus necesidades y presupuesto, debido a la gran cantidad de opciones disponibles y a la falta de información clara y objetiva, además del costo en tiempo y trámites en que incurren para disminuir el riesgo de comprar un auto en malas condiciones. Por el lado de los vendedores, pueden existir problemas para encontrar un precio que maximice sus ganancias esperadas, debido, por ejemplo, a una percepción sentimental que sobreestime el precio de mercado de su vehículo; por el contrario, puede existir el riesgo de asignar un precio muy inferior al promedio del mercado.

Además, en el caso de las empresas vendedoras, un precio muy alto puede dificultar la rotación de inventarios y afectar los costos. De igual manera, una empresa vendedora que presente un gran número de quejas y de usuarios inconformes, puede ver afectada su reputación. En este sentido, uno de los principales problemas en la compra de autos usados es la dificultad para comparar y evaluar las características de los vehículos disponibles en el mercado. Esto puede llevar a decisiones de compra poco informadas y, en última instancia, a la insatisfacción del cliente.

Dado lo anterior, el propósito de este proyecto es desarrollar un sistema de recomendación que, mediante el análisis de un dataset de autos usados, facilite la realización de las ventas en el mercado de carros usados, permitiendo a los compradores potenciales encontrar el vehículo que mejor se adapte a sus preferencias, necesidades y presupuesto. A su vez, que le facilite al vendedor comparar otras opciones similares y pueda maximizar su ganancia esperada, reduciendo, además, el tiempo esperado.

Objetivos

General

- Encontrar las principales variables que afectan el precio de los vehículos y con base en estas, desarrollar un sistema de recomendación simple, capaz de sugerir un precio de venta competitivo para vehículos particulares utilizando toda la información del vehículo e información del mercado actual.

Específicos

- Comprender las diferentes variables de los vehículos y cómo cada una de ellas podría afectar el precio.
- Identificar las variables más relevantes y con mayor impacto sobre el precio de los vehículos.
- Utilizar herramientas para el procesamiento de grandes volúmenes de información que permitan extraer información útil sobre las características de los vehículos
- Seleccionar el modelo que mejor se ajuste a los datos y recomiende el precio más acertado que beneficie tanto a vendedor como a comprador.
- Implementar el modelo preliminar con una muestra de los datos.

Revisión de literatura y estado del arte.

Cuando hablamos del uso de machine learning para predecir precios la literatura ha hondado en los métodos tradicionales como lo son la Regresión Lineal, Random Forest, XGBoost entre otros modelos de esta índole, sin embargo, la literatura no ha tratado de explorar el concepto de un consumidor-vendedor lo cual es extraño, pues hoy en día existen bastantes plataformas donde los usuarios pueden vender productos usados.

Los sistemas de recomendación se han tratado a lo largo de los años y se han enfocado en sugerir contenido o productos a usuarios, (Lu et al. 2015) hace un buen trabajo en resumir y explicar los múltiples tipos de sistemas de recomendación y lo que estos hacen. Nosotros nos enfocaremos en los dos principales filtros colaborativos y filtros basados en contenidos, Shafter et al. (2007) define los filtros colaborativos como el proceso de filtrar o evaluar un objeto usando las opiniones de otras personas. Este sistema se forma mediante una matriz que clasifica las preferencias de las personas y las calificaciones que ha recibido un elemento y con base en esto realizara recomendaciones por similitud y estos son filtros de tipo de vecinos más cercanos.

Los filtros basados en contenido los definin en Pazzani y Bilsus (2007) como sistemas que analizan las descripciones de los elementos para poder identificar objetos que son de interés específico para un usuario. Estos filtros se enfocan más en el objeto que en el usuario para la predicción, este método por su esencia se adapta más a nuestro problema pues nuestro problema no tiene usuarios como tal y la única información que tenemos del usuario es el vehículo que desea vender y las características de este.

Finalmente, en tema de predicción del precio de venta de vehículos usados, no se encontró un gran desarrollo del tema, en especial mediante sistemas de recomendación, en la literatura son más frecuentes la construcción de modelos de predicción de precios. Sin embargo, hay un consenso de cuáles son las variables más relevantes para este modelo (Gajera, et al (2021), Pal et al (2018) y Venkatasubuu (2019), encuentran que las variables más relevantes a la hora de vender un vehículo son como esperaríamos las siguientes: modelo del vehículo, año, kilometraje (millaje), cilindraje, marca y el tipo de carrocería. Además, los métodos de predicción que en general alcanzan un mejor desempeño son la regresión lineal, árboles de decisión y los métodos de ensamble como Random Forest o XGboost.

Metodología

Como se mencionó en el anteproyecto, para la realización del trabajo se seguirá la metodología CRISP-DM. A continuación, se explican cada una de las fases de desarrollo.

a) Comprensión del negocio:

El mercado de vehículos usados se está haciendo cada vez más grande, con más modelos y más personas en búsqueda cambiar su automóvil. En el caso puntual de Colombia, las cifras del Registro Único Nacional de Tránsito (RUNT) y la Asociación Nacional de Movilidad Sostenible (Andemos), mostraron que para 2022 se registraron alrededor de 1,1 millones de traspasos de vehículos usados que si bien tuvieron una leve reducción de 1% frente a 2021, fue el segundo año con un mayor número; a su vez, se espera que para 2023 la cifra esté entre 900.000 y 1 millón de traspasos. Es términos generales, se estima que por cada vehículo nuevo se venden cerca de 4,5 usados[1]. A manera de ejemplo, la figura 1 muestra las búsquedas del término "carros usados" vs "carros nuevos" en Google Trends, evidenciando que en los últimos 5 años el interés por los primeros ha sido mayor, situación que se aceleró en la pandemia debido en parte al fuerte incremento que tuvieron los vehículos nuevos.

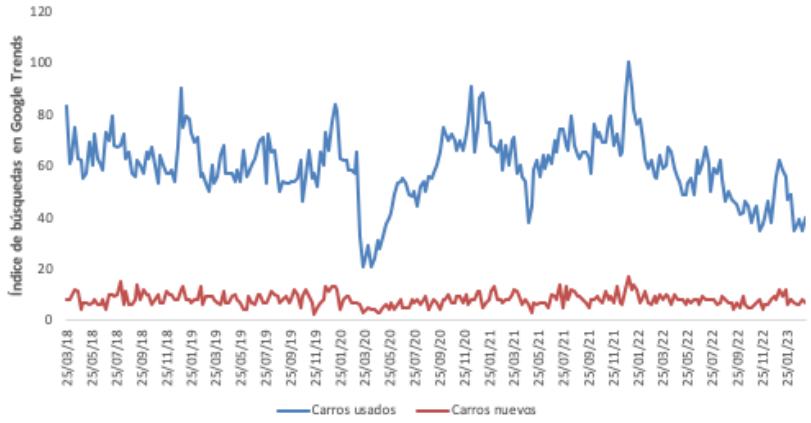

Figura 1. Búsqueda de los términos carros nuevos- carros usados en Google Trends.

En la figura 1 también se puede observar que el interés en las búsquedas de carros usados ha disminuido desde su punto máximo a comienzos de 2022, entre otros factores han incidido la desaceleración económica, el aumento sostenido en el precio de los vehículos y de las tasas de interés de los créditos para la compra.

Adicional a lo anterior, cabe destacar la existencia de diversas plataformas que ayudan a guiar las búsquedas y reducir la incertidumbre tanto de compradores como de vendedores. Entre algunas

[1] Revista Motor (2022). Cayó el mercado de carros usados durante 2022. Disponible en: https://www.motor.com.co/industria/Cayo-el-mercado-de-carros-usados-durante-2022-20230112-0005.html

de las más importantes se encuentran Carvana o Cargurus en Estados Unidos, Mercadolibre, OLX entre otros para Latinoamérica. El desarrollo de estas plataformas ha permitido ganancias de eficiencia en el mercado de vehículos que hace algunos años no hubieran podido suceder, en especial porque uno de los criterios por los cuales se guían los compradores es la calificación del vendedor del vehículo, que se construye con base en transacciones previas y comentarios de otros usuarios.

b) Comprensión de los datos:

El dataset que usaremos es de acceso público y fue obtenido de Kaggle, este contiene información de la plataforma Cargurus tomados con webscrapping en septiembre de 2020. El número de registros es de 3 millones de vehículos con 66 variables que describen las características más importantes de los carros, el tamaño total del archivo csv es de 10 GB[2]. Algunos de los atributos disponibles son: Precio, tiempo en el mercado, cilindraje, tipo de vehículo, transmisión, economía de combustible, ubicación, marca, modelo, kilometraje, accidentes, tipo de motor y tipo de combustible. Estas son algunas de las variables que según la literatura consultada podrían dar más información sobre el precio al que se venderá un vehículo. En el anexo 1 se muestran los principales atributos del dataset y una breve descripción de los mismos.

c) Preparación de los datos:

Tratamiento de datos repetidos:

Se examinó que el datase no tuviera registros duplicados mediante la columna "vin", que era el identificador único del carro. No se presentaron datos repetidos.

Tratamiento de datos faltantes:

Se hizo una exploración inicial del porcentaje total de datos faltantes por cada atributo, a continuación, se muestra el ranking de aquellos con mayor cantidad de datos omitidos:

[2] Kaggle (2023). Used Cars Dataset. Disponible en: https://www.kaggle.com/datasets/ananaymital/us-used-cars-dataset?datasetId=885427&sortBy=voteCount&select=used_cars_data.csv. Consultado el 15 de marzo de 2023.

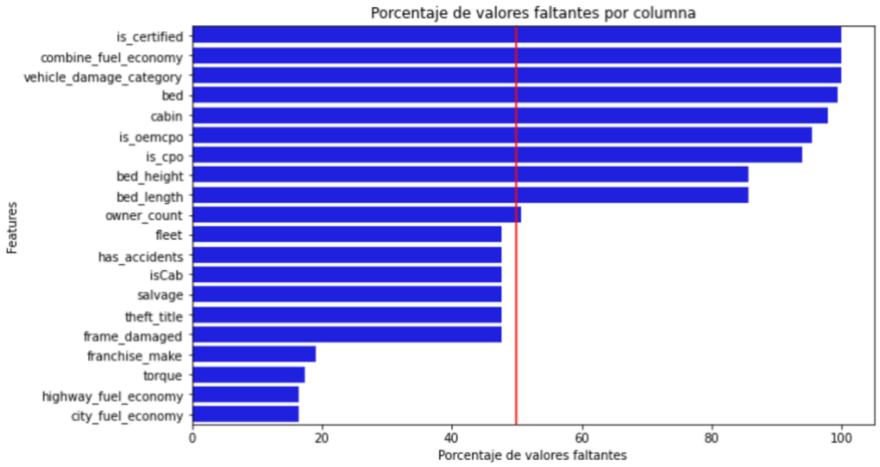

Figura 2. Porcentaje de datos faltantes en el dataset inicial.

De entrada, se eliminaron los tres primeros features los cuales no contenían información. En un paso siguiente, se examinaron las columnas con un porcentaje de missings superior a 50% y se consideró eliminar aquellas en las cuales no fuera posible su imputación con otra de las variables disponibles que sirviera de proxy. Este fue el caso de owner_count, bed, cabin, bed_height y bed_lenght, estas cuatro últimas, además, hacían referencia a un tipo de vehículo poco frecuente en la muestra. En otros casos de variables binarias como has_accidents, is_cab, salvage, theft_title y frame_damage, no era claro si el dato faltante pertenecía a la categoría 1 ó 0, por lo que también se excluyeron del análisis. Por otra parte, fue posible imputar los datos de franchise_make, highway_fuel_economy y city_fuel_economy buscando en el dataset el nombre del mismo modelo de vehículo (model_name) en otra de las filas. Además, se imputó el torque de los vehículos usando la columna de caballos de fuerza.

Eliminación de features con información redundante o no útil para el problema:

Además de las columnas que se eliminaron por contener una gran cantidad de faltantes y que no se pudieron imputar, se dejaron por fuera las columnas que parecían expresar características similares como engyne_cylinders y engine_type (cilindraje del vehículo), franchise_make y make_name (marca del vehículo), horse_power y power (caballos de fuerza), Wheel_system y Wheel_system_display (tracción en las ruedas). Además, se eliminó la columna que contenía la dirección URL con fotografías de los vehículos y las coordenadas geográficas de latitud y longitud y sp_name que hacía alusión al nombre del concesionario donde se ofrecía el vehículo.

Limpieza de datos en algunas columnas con presencia de letras y palabras:

En los features back_legroom, front_lengroom, fuel_tank_volume, height, length, wheel_base, width y maximum_seating, se limpiaron las celdas de letras para dejar solo el número que indicaba la magnitud respectiva. Por su parte, en la categría listing_color se examinó la categoría "unknown" que representaba la mayor parte de las observaciones para encontrar posibles agregaciones con colores ya existentes. En este caso, se encontró que la mayoría de colores clasificados como desconocidos en realidad pertenecían a distintos tipos de gris o plateado, principalmente, y en menor medida blanco y negro. De esta manera, se agruparon estos colores en su respectiva categoría. Otro de las columnas que se limpio fue torque, que contenía las letras de la unidad de medida (lb-ft).

Ingeniería de características:

Algunas de las variables podrían contener información de texto útil, por lo que se decidió procesarlas para volverlas una columna numérica. A continuación, se muestran cuales fueron:

Description: Esta columna contenía la descripción del vehículo en venta por parte del oferente. Mediante procesamiento de lenguaje natural y el uso de expresiones regulares, se transformó esta columna en un número igual al total de palabas usadas por el vendedor para describir el auto y se excluyeron los stopwords y los caracteres especiales. Se nombró como "word_count_description".

Major options: Contiene una lista con la descripción de los lujos o componentes extras que tenía el vehículo. Entre los más comunes se encontraron el GPS, bluetooth, calefacción en los asientos, rines de lujo, detección de colisiones, entre otros. Esta columna se transformó en un número igual a la cantidad de elementos en esas listas. A esta columna procesada se le puso el nombre de "extras".

Análisis exploratorio de datos:

El dataset completo contiene tanto variables numéricas como categóricas. Las primeras son 17, de las cuales 15 provienen del original y se crearon las dos anteriormente descritas "word_count" y "extras. Cabe mencionar que se eliminó el atributo de economía del vehículo en la autopista (highway_fuel_economy), por tener una correlación de 99,8% con el de economía del vehículo en la ciudad (city_fuel_economy). En la figura 3 se muestra el histograma de una muestra de variables numéricas y como líneas de referencia la media y la mediana:

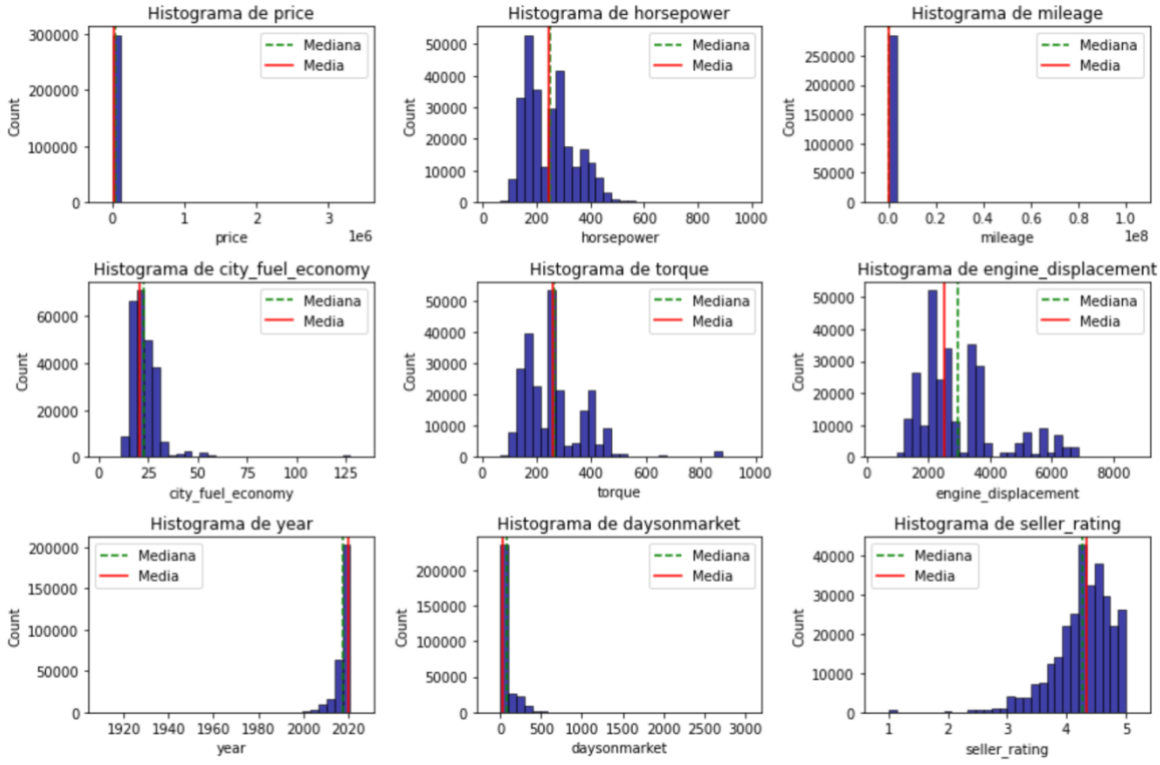

Figura 3. Histograma de un conjunto de variables numéricas del dataset.

De forma general, la mayoría de los histogramas fueron asimétricos, denotando la posible presencia de datos atípicos en las observaciones. Llama la atención el caso del precio, el millaje, el modelo (year) y los días en el mercado de los vehículos, con una pequeña proporción de sus valores muy alejados del centro, lo que hace ver los histogramas apiñados en valores más bajos. Algo más simétrica fue la distribución del calificativo del vendedor "seller rating", aunque con una cola moderada a la izquierda ya que gran parte de los valores se concentran en el límite superior en el rango de 4 a 5. Un poco más simétrico fue el histograma de economía del vehículo en la ciudad (city_fuel_economy), con más observaciones cercanas a la media y a la mediana de la muestra, mientras que los caballos de fuerza, el torque y el cilindraje (engine_displacement), mostraron histogramas con indicios de posible multimodalidad.

Ahora bien, como los histogramas dieron indicios de presencia de datos atípicos, se procedió a calcular el porcentaje de observaciones que se encontraba por fuera del rango intercuartil. En la tabla 1 puede verse que el atributo con mayor presencia de datos atípicos fue el de días el del número de días en el mercado, seguido de la distancia entre los ejes "wheelbase", aunque con valores relativamente moderados algo mayores al 10% del total. Por su parte, la mayoría de atributos, 12 en total, presentaron una presencia leve y tratable de outliers inferior al 5%.

Variable	**Porcentaje NAs (%)**
daysonmarket:	13,46
wheelbase:	12,48
maximum_seating:	9,30
year:	9,08
mileage:	7,75
engine_displacement:	4,62
price:	2,85
seller_rating:	2,80
city_fuel_economy:	2,66
torque:	1,54
back_legroom:	1,45
word_count_description	1,25
width:	0,81
front_legroom:	0,69
horsepower:	0,52
height:	0,40
extras	0,00

Tabla 1. Porcentaje de datos por fuera del rango intercuartil en las variables numéricas.

Ahora bien, se procedió a analizar la relación entre variables numéricas usando el coeficiente de correlación de Pearson. Los resultados para las variables con correlaciones mayores a 0,10 se muestran en la figura 4.

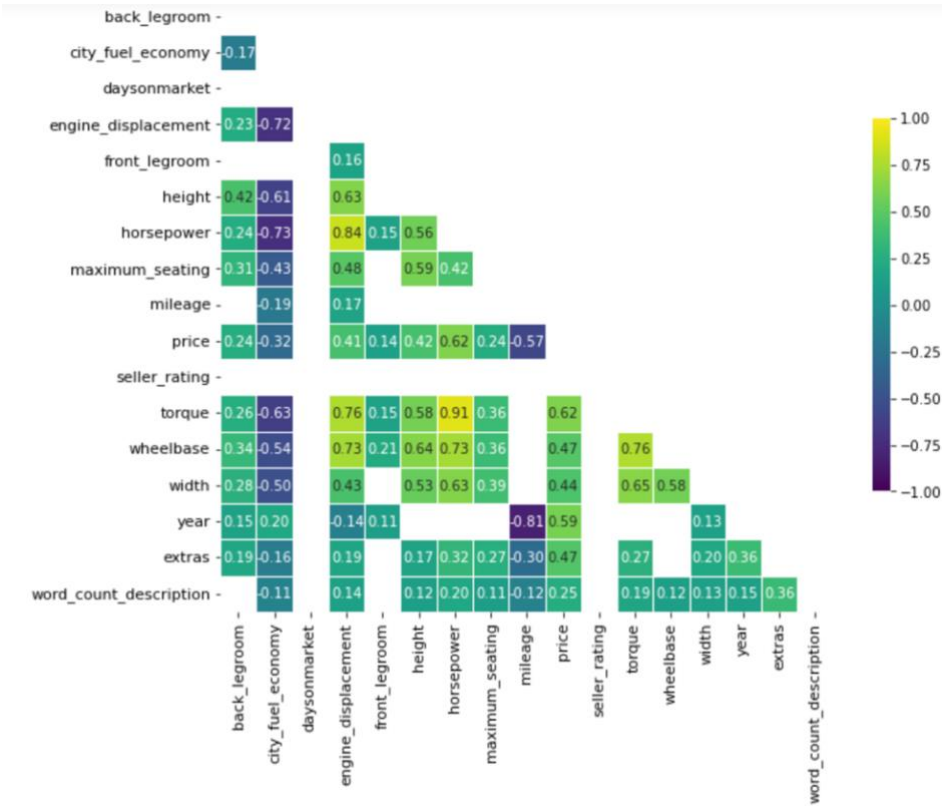

Figura 4. Coeficiente de correlación de Pearson entre las variables numéricas del dataset.

Con relación al precio, sobresale que las características más correlacionadas de forma lineal son los caballos de fuerza (0,62), el torque (0,62), el año de fabricación del vehículo (0,59) (mientras más reciente mayor es el precio); en menor medida, también hay asociación positiva con el cilindraje del vehículo y la distancia entre los ejes. Llama la atención también la leve relación que hay entre los dos features creados, cantidad de extras y conteo de palabas en la descripción del vehículo. De otra parte, el millaje del vehículo (-0,81) y el rendimiento del combustible en la ciudad (-0,32) son las variables que tiene una asociación negativa mayor con el precio.

Otros atributos que presentan alta correlación son los caballos de fuerza con el torque (0,91) y con el cilindraje (0,84) y también entre el cilindraje y el torque (0,76). Es importante tener en cuenta esto, en especial a la hora de construir un modelo de regresión lineal que tiene como supuesto importante del modelo la no correlación entre las variables explicativas.

Adicional a lo anterior, como parte de la limpieza y el tratamiento de outliers entre las variables numéricas, se utilizó la distancia de Mahalanobis para filtrar de forma multivariada el 2,5% de los datos más alejados del centro.

Descripción de las variables categóricas:

En lo que se refiere a los atributos categóricos, luego de realizar la limpieza de datos y descartar aquellos que tenían información redundante y que no era de interés para el proyecto, se dejaron las características que se muestran en la Tabla 2.

Variable	# de etiquetas	Etiqueta más repetida (%)	Etiqueta menos repetida (%)
city	4354	Houston (1,46)	Three Rivers (0,01)
model_name	1070	F-150 (4,32)	CJ-8 (0,01)
make_name	76	Ford (15,93)	Austin-Healey (0,01)
engine_type	35	I4 (48,83)	W12 Flex Fuel Vehicle (0,01)
listing_color	15	WHITE (22,78)	PINK (0,01)
body_type	9	SUV / Crossover (43,47)	Convertible (0,86)
fuel_type	8	Gasoline (86,55)	Propane (0,01)
wheel_system_display	5	Front-Wheel Drive (44,11)	4X2 (4,26)
transmission	4	A (82,33)	Dual Clutch (0,41)

Tabla 2. Variables categóricas, número de etiquetas y frecuencia mayor y menor.

Como se puede observar, varias categorías presentaban un número alto de etiquetas, especialmente ciudad, nombre del modelo del vehículo, y el fabricante. Más adelante se muestra cómo se trató de corregir esto para evitar problemas de dimensionalidad. Con base en lo anterior, también se muestra de manera exploratoria las etiquetas y su distribución de frecuencia en algunos de los features seleccionados en la figura 5. Como puede observarse los colores más frecuentes son blanco (22%), negro (20%), gris (17%), plateado (13%), rojo (9%) y azul (8%).

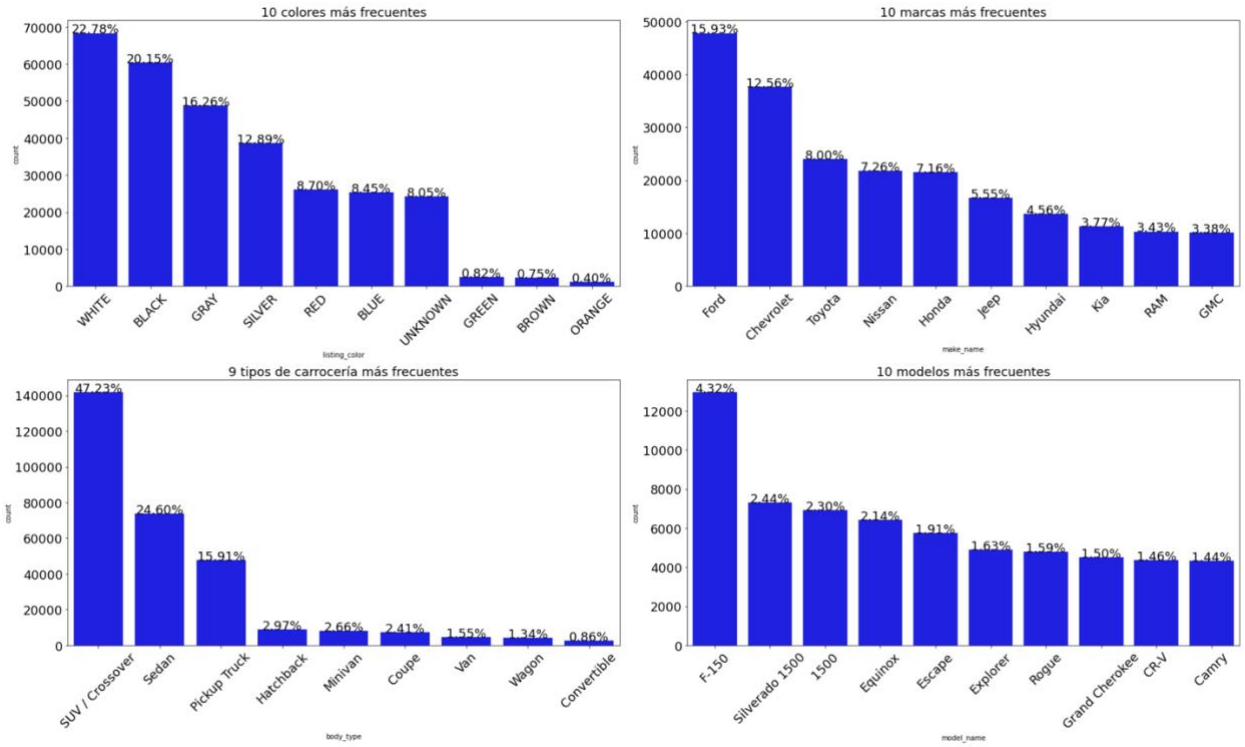

Figura 5. Distribución de las frecuencias de las etiquetas entre las principales variables categóricas:

En forma de complemento, a continuación se muestran los boxplots de las principales variables categóricas teniendo en cuenta el precio del vehículo. Parar una mejor observación del efecto por cada variable categórica, se excluyeron los valores por fuera del rango intercuartil:

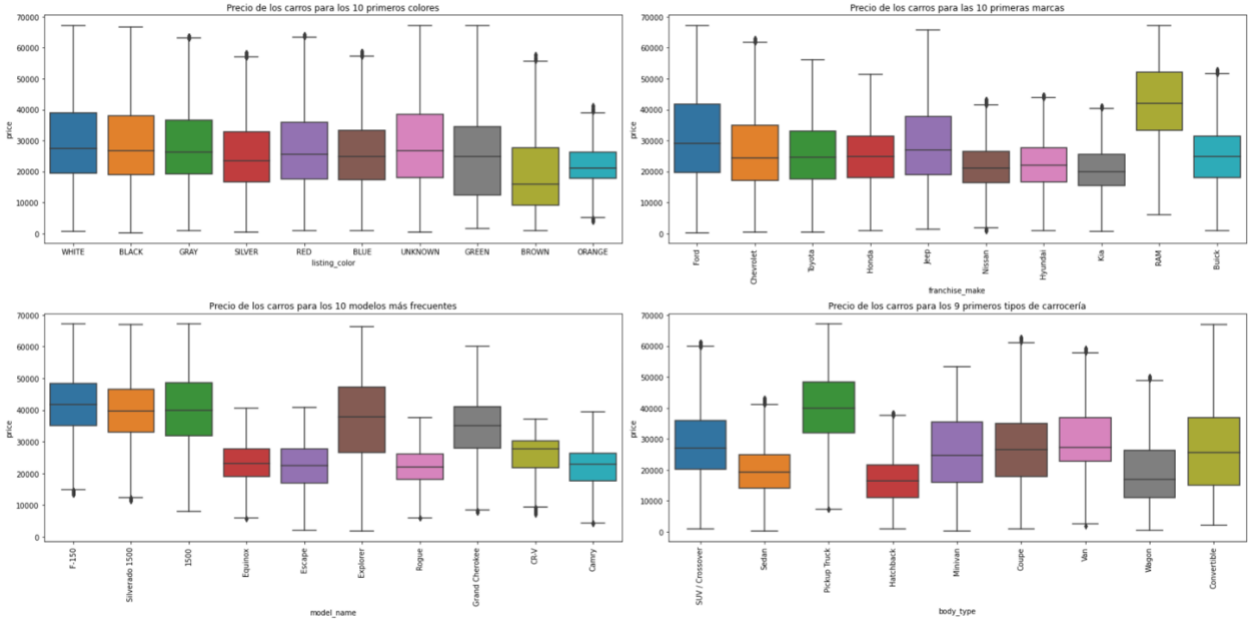

Figura 6. Distribución del precio según las etiquetas más frecuentes en algunas variables categóricas:

Resalta que para algunas de las características observadas pueden verse cambios significativos en el precio, principalmente en las marcas, en donde los vehículos de marca RAM parecen tener un mayor precio. Entre los modelos de los vehículos, sobresalen la Ford-F150, el Silverado 1500 y la RAM 1500 que junto con la Ford Explorer, parecieran tener rangos de precio más altos. En relación con lo anterior, en la categoría tipo de carrocería, la etiqueta que destaca por un precio más alta es la de Pickup Truck.

Finalmente, se realizó una prueba no paramétrica de Mann Whitney, también conocida como test de suma de rango de Wilcoxon, para definir si las distribuciones de los precios entre las variables categóricas eran iguales o no. La noción detrás de esta prueba es ver si una de las categorías tiende a tener valores más altos que los demás, comparando pares de categorías, esto daría un indicio de si efectivamente un atributo en particular incide en el precio. La hipótesis nula H_0 es que las distribuciones en las categorías son iguales y la alternativa que son distintas, se rechaza H_0 si el p-valor de la prueba es mayor al nivel crítico utilizado que para este caso fue de 5%.

En la figura 7 se construyen a manera de ejemplo dos matrices que contienen el resultado de las pruebas para las variables color del vehículo y fabricante del vehículo. Las casillas rojas indican pares de etiquetas en las que se rechaza la prueba, mientras la azul indica que las distribuciones del precio son estadísticamente similares. En general, se encontró que para casi todos los colores las distribuciones de los precios eran distintas solo en dos casos, gris y desconocido, y gris y verde se encontraron distribuciones similares. El resultado fue similar en el caso de los fabricantes de vehículos, con la mayoría de pares de marcas teniendo una distribución de precios distintas, exceptuando Toyota-Chevrolet y Honda-Buick.

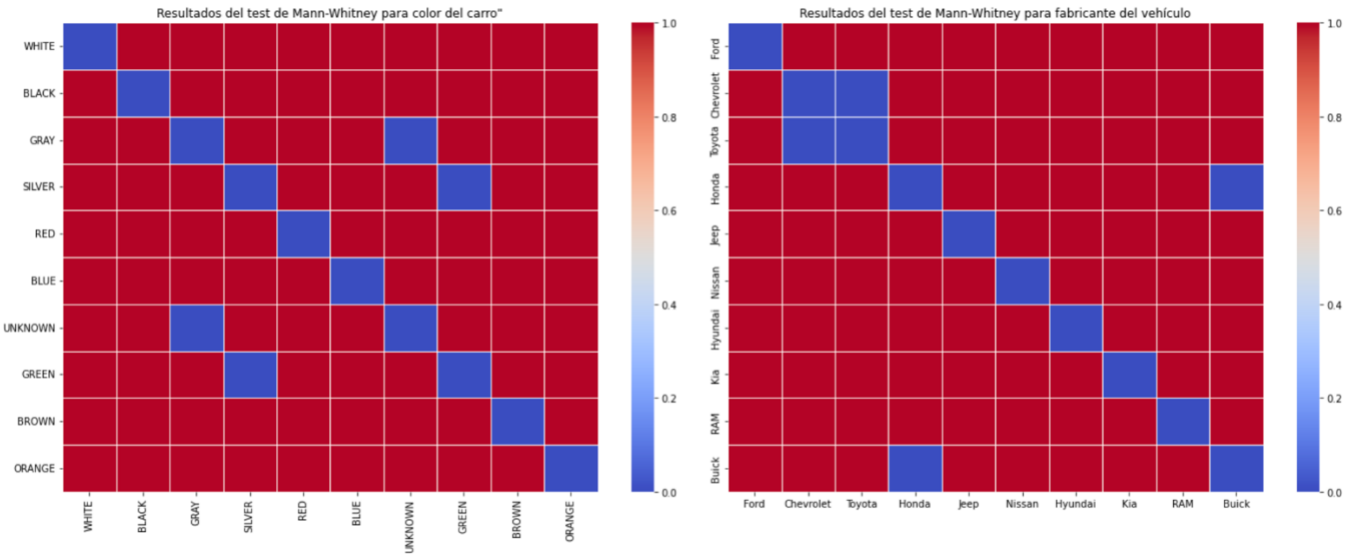

Figura 7. Distribución del precio según las etiquetas más frecuentes en algunas variables categóricas:

d) Modelado:

Para evaluar la capacidad predictiva de las variables del dataset y la eficacia de su uso en el sistema de recomendación se construyeron tres modelos de Machine Learning usando la libreía MLlib de Spark: Regresión lineal, Random Forest y Gradient Boosting. Para el caso particular del modelo de regresión, se excluyeron los atributos con un coeficiente de correlación mayor a 0,8% para evitar problemas de multicolinealidad (horsepower, cilindraje, torque) y se estandarizaron las variables numéricas continuas. A su vez, para los tres modelos se hizo one-hot encoding a las variables categóricas, en las cuales se agregaron las etiquetas que aparecían menos del 1% en una llamada "otros", para evitar problemas de alta dimensionalidad, especialmente en el caso de ciudad, nombre del modelo, y el fabricante del vehículo.

El propósito de elegir estos modelos era ver cuál podría tener métricas de desempeño más competitivas en los datos de prueba, acorde con lo visto en la literatura sobre predicción de precios de vehículos. Se encontró que en particular la raíz el error cuadrático medio (RMSE), el error porcentual absoluto medio (MAPE) y el R2 eran las métricas más usadas para evaluar la calidad de la predicciones. Otra ventaja que presentan los modelos empleados, es que permiten conoces cuáles son los features que más impactan la predicción.

e) Evaluación

Con base en lo anterior, en la tabla XX se muestran los resultados obtenidos:

Modelo	**Train**			**Test**			**Hiperparámetros**
	RMSE	MAPE (%)	R2 (%)	RMSE	MAPE (%)	R2 (%)	
Regresión Lineal	4136,5	19,8	85,78	4152,2	20,1	85,72	Ridge: C=1

Random Forest	2672,6	11,0	94,06	2829,2	11,7	93,37	n_estimators=300, max_depth=14, min_samples_split=6, min_samples_leaf=4; max_features=sqrt
Gradient Boosting	2903,8	11,3	92,99	2909,2	11,3	92,99	n_estimators=300, learning rate=0.1 min_samples_leaf=1, max_depth=3, max_features=sqrt

Tabla 3. Resultados de los modelos de regresión para la predicción del precio de los vehículos.

Sobresalió la capacidad predictiva de los modelos de ensamble, tanto de la regresión de Random Forest (RF) como la Gradient Boosting (GB). Inclusive, se encontró que en comparación con algunos modelos de la literatura consultada, los que se presentan en este trabajo tuvieron, en general, un mejor desempeño (Bukvić et al, 2022) (Gajera et al, 2021) (Pal et al, 2017). Esto se debe en gran parte al conjunto de variables utilizadas, las cuales sobrepasaban en cantidad y en calidad a la de estos trabajos; además, otro buen indicativo de los resultados acá vistos es que los modelos corrigen relativamente bien el sobreajuste, en especial el Gradient Boosting y la Regresión Lineal. En el caso del Random Forest, se modificaron los hiperparámetros de profundidad máxima, número mínimo de hojas en cada rama, y el número máximo de características para entrenar cada árbol, esto permitió un mejorar el sobreajuste.

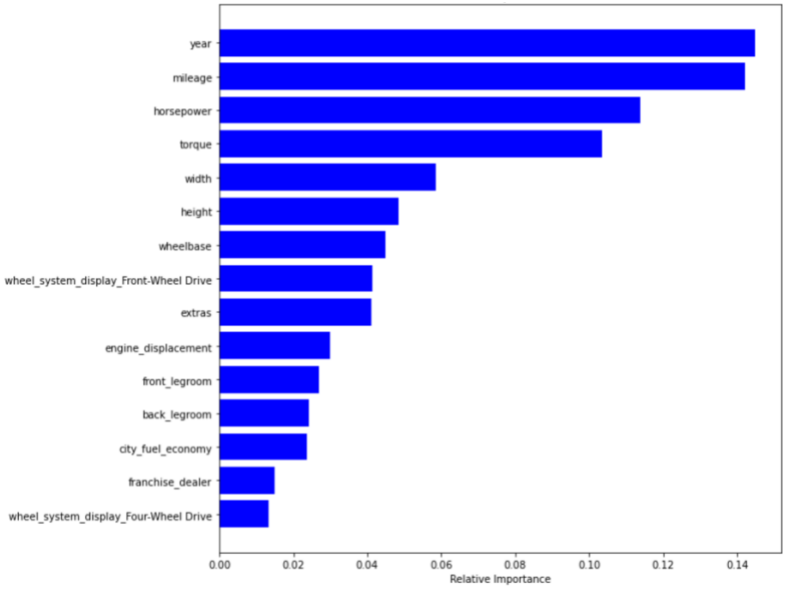

Figura 8. Importancia relativa de los features en el modelo de RF

En este sentido, como las métricas obtenidas con RF y GB fueron similares, se optó por extraer cuáles eran las características más importantes para cada modelo y se encontró que, como lo muestran las figuras 8 y 9, las más importantes para predecir el precio de los vehículos, en ambos casos, fueron el año de fabricación y el millaje del vehículo. Por su parte, estas características tuvieron una relevancia mayor en el GB (Figura 9), en donde además, se encontró que el torque, el alto del carro, los caballos

de fuerza, la tracción en las ruedas delanteras y la cantidad de extras eran los otros atributos que más ayudaban en la predicción del precio.
Por el lado del RF, los otros features que tenían importancia predicitiva en el modelo eran similares, pero en distinto orden. En tercer lugar se ubicaron los caballos de fuerza, luego el torque, el ancho, la altura, la distancia entre los ejes y la tracción en las ruedas delanteras.

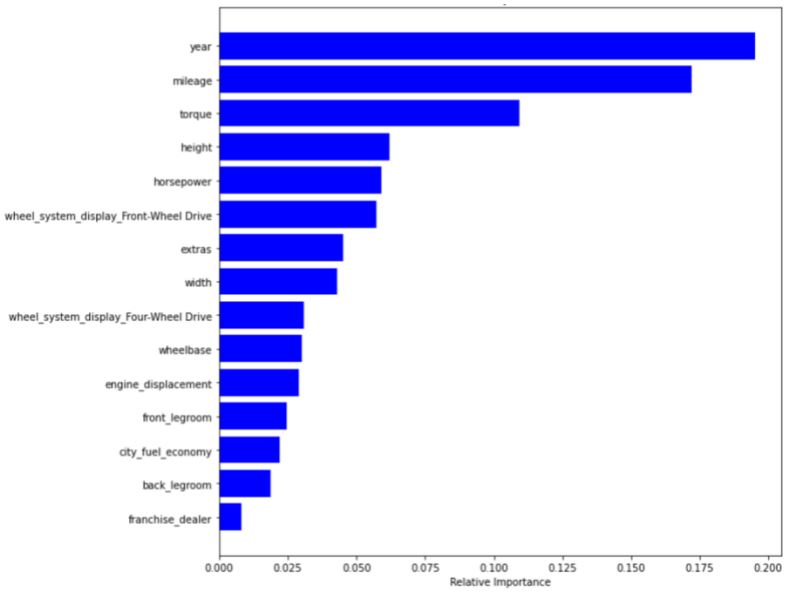

Figura 9. Importancia relativa de los features en el modelo de GB.

f) Despliegue

Esta sección esta por fuera del alcance del presente ejercicio que tiene propósitos fundamentalmente académicos. Sin embargo, en la sección siguiente se muestran algunos ejercicios de prueba del sistema de recomendación de precios.

Uso de herramientas de Big Data.

Para realizar el preprocesamiento y análisis exploratorio de los datos originales, se utilizó el servicio en nube de Microsoft Azure, para lo cual primero se cargaron los datos en una cuenta del servicio de almacenamiento de Azure Blob Storage dentro de un contenedor llamado "minería", como se muestra a continuación:

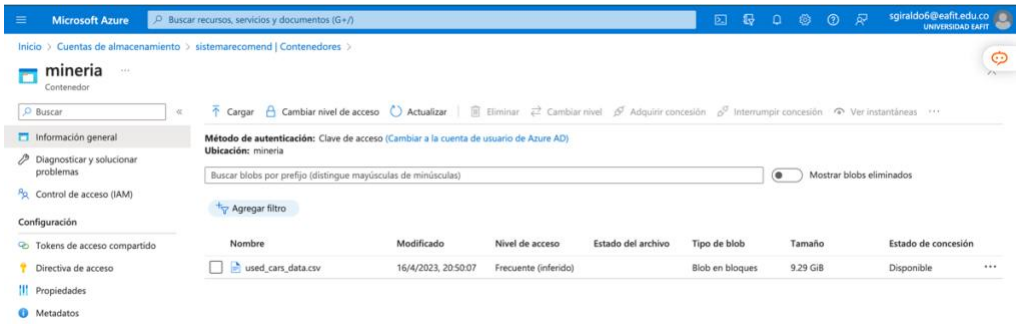

Figura 10. Almacenamiento de la base de datos original en la cuenta de almacenamiento de Azure Databricks.

Posteriormente, se utilizó una cuenta de estudiante en la herramienta Azure Databricks para crear un clúster en donde se realizó el preprocesamiento y el análisis exploratorio de los datos en Pyspark. La configuración se muestra a continuación:

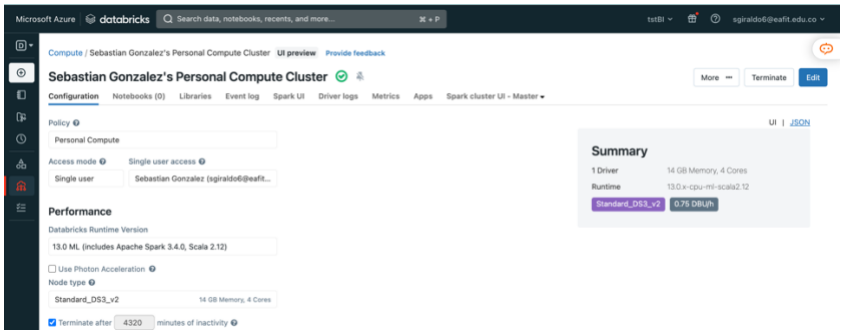

Figura 11. Configuración del clúster de trabajo en Azure Databricks.

Una vez procesados los datos, se sacó una muestra aleatoria del 10% que se guardó en una tabla llamada cleaned_data para construir algunos modelos de Machine Learning usando la herramienta de AutoML. En el experimento se usaron las librerías de lightgbm, xgboost y sklearn para construir los modelos de predicción de las regresiones.

Los resultados del experimento se muestran en la siguiente figura:

Run Name	Created	Duration	Source	Models	val_root_mean	model_type
righteous-doe-275	3 hours ago	2.7min	Noteboo...	sklearn	9340.9	xgboost_re...
vaunted-stag-408	3 hours ago	1.5min	Noteboo...	sklearn	9526.2	lightgbm_r...
abrasive-shoat-675	3 hours ago	54.5s	Noteboo...	sklearn	9874.8	xgboost_re...
suave-duck-100	3 hours ago	1.2min	Noteboo...	sklearn	10490.5	lightgbm_r...

32 matching runs

Figura 12. Resultados del experimento de AutoML en Databricks

No obstante, debido a limitaciones en la capacidad de cómputo del clúster y a que el proceso de AutoML necesitaba un tiempo de entrenamiento mayor para obtener mejores resultados, se optó por entrenar los modelos en una máquina local con la librería sklearn.

Desarrollo del sistema de recomendación.

Los sistemas de recomendación establecen un conjunto de criterios y valoraciones de los datos, con el objetivo de realizar predicciones que se consideren recomendaciones de elementos que pueden ser de utilidad para los usuarios, dentro de los sistemas de recomendación tenemos varios, primero tenemos los sistemas de popularidad, estos son implementados para recomendar productos que a un usuario le puedan interesar con el objetivo de aumentar ventas, luego están los sistemas de contenido que se basan en datos históricos almacenados y busca similaridades con la información almacenada u obtenida del usuario, nuestro sistema estará basado en este sistema pues buscamos que nos

recomiende el precio de venta o de compra adecuado para un vehículo tomando información histórica de vehículos actualmente en la plataforma. Finalmente tenemos los sistemas colaborativos, este es el más novedoso de todos pues genera recomendaciones analizando los datos que están almacenados y los datos del usuario comparando a este con otros usuarios similares y así hacer recomendaciones. algunos sistemas bastante conocidos son los utilizados por Netflix, Spotify y otros servicios de streaming para recomendar contenido nuevo a las personas.

Para el sistema de recomendación usaremos un sistema de recomendación por contenido item to item, este no es el más avanzado, pero es un ejercicio interesante para predecir el precio de venta de un vehículo, el filtro que nosotros haremos es un filtro de ranking el cual no es el más eficiente y depende de que la base de datos no sea muy grande pues debe ordenarla con base en los datos que tiene y. de esta manera buscar las similaridades entre los diferentes ítems y de esta manera generara el ranking con el cual hará la recomendación. Con el fin de evaluar este modelo y encontrar el modelo que mejor se adapta a los datos variaremos las métricas de distancia que usaremos para calcular el ranking, para esto usaremos la norma euclídea, coseno, Manhattan y chebyshev, una vez medidas todas usaremos la media del vehículo objetivo luego mediremos la diferencia entre la recomendación y dicha media y podríamos decir que la menor distancia entre la recomendación y la media se considera el mejor recomendador.

Muestra de recomendaciones de precio valores en USD				
model	**year**	**price**	**mean_price**	**confidence_interval**
F-150	2015	37602.2	36974.08	(35299.61 : 38648.55)
Elantra	2000	9006.8	7763.35	(6608.92 : 8917.78)
CR-V	2018	20715.6	20617.16	(19609.77 : 21624.55)
Fusion	2020	10895.8	7689.75	(6542.86 : 8836.64)
Grand Caravan	1999	5957	7266.98	(6233.9 : 8300.06)
Silverado 1500	2000	36633.6	35580.07	(33914.65 : 37245.49)
Altima	2010	12375.2	12301.74	(11515.57 : 13087.91)
RAV4	2011	10153.6	10286.66	(9625.5 : 10947.82)
Sentra	2019	8784.4	10900.74	(10118.46 : 11683.02)
Grand Cherokee	2015	33201	28584.24	(27145.37 : 30023.11)

Tabla 4. Resultados de una muestra de 10 vehículos de predicción de precio

Tratamiento de la muestra de datos.

Para el desarrollo de este modelo utilizamos un data set de una página de venta de vehículos donde contamos con datos como el modelo, la marca, cilindraje, distancia entre ejes, combustible, daños, millaje, entre otros. El data set contaba con 3,000,000 y un peso de 10 gb de datos por la limitación

de capacidad computacional limitaremos la muestra entre 30,000 y 50,000, el data set se limpió y se eliminaron datos nulos de algunas variables. Algunas de las técnicas para la eliminación de datos nulos fueron las siguientes. Millaje se tomó el promedio de millas de todos los vehículos de la muestra y esta se multiplico por la diferencia entre el año de fabricación del vehículo y el año de la data set 2022, así obtuvimos una aproximación de cuantas millas condujo el vehículo, claramente pueden existir eventos atípicos donde un vehículo más viejo tenga menos millaje, pero en general es una aproximación aceptable. Torque, para este utilizamos la fórmula para el cálculo del torque que es la siguiente Torque = (caballos de fuerza x 5252) / RPM. Para las RPM utilizamos 4000 RPM pues esta es la cantidad de RMP donde gran cantidad de los vehículos alcanzan el pico de torque ya que el torque no escala de manera lineal. Otras variables como eficiencia de combustible en ciudad y horse power tomamos un valor de un vehículo del mismo modelo y de esta manera obtuvimos una aproximación a los datos.

Test

Para probar el modelo tomamos 54 muestras creadas de manera aleatoria y ejecutamos el modelo con cada una, para esto tomamos la media de todos los vehículos en la muestra de datos con el mismo modelo y año que el que queremos comparar, luego calculamos la diferencia entre este y las recomendaciones con las diferentes métricas.

Distancia mediana en USD		**Distancia Media en USD**	
Euclidea	7273.89	Euclidea	9401.09
Coseno	8205.34	Coseno	9811.17
Manhattan	7278.14	Manhattan	9101.71
Chebyshev	6697.69	Chebyshev	9215.71

Tabla 5. Distancia media y mediana de las métricas en USD.

Como podemos observar, la media entre las tres métricas es muy similar lo que nos indica que cualquiera de las métricas podría utilizarse de manera efectiva para recomendar precios y las medianas de igual manera lo que nos da un indicio que no tenemos distancias outliers, sin embargo la mediana de chebysehv es la más baja por lo cual descartaremos esta por ser la más sensible, por el mismo motivo y por ser la más alta la distancia coseno también puede ser descartada, lo que nos deja con la distancia euclídea y la distancia Manhattan muy cerca la una de la otra y podremos utilizar cualquiera de las dos.

Con el objetivo de medir que tan distantes están las recomendaciones y los elementos, el modelo también calcula un intervalo de confianza con los 100 elementos más similares al elemento con esto también medimos las distancias para evaluar que tan viable es la implementación de este modelo, para observar esto notamos que hay una diferencia de 2000 dólares entre el límite inferior y el límite superior lo cual es un monto aceptable para este problema pues se podrían utilizar más variables a la hora de seleccionar las recomendaciones que podrían acercarnos más al precio real.

valor medio de las metricas en USD

Euclidea	2483.81
Coseno	1911.48
Manhattan	2486.29
Chebyshev	2468.23

Tabla 6. Valor medio de las métricas dólares.

Cabe resaltar que si los precios de los vehículos son más bajos es más probable que las diferencias con su precio medio sean más bajas a medida que la variabilidad del vehículo sea menor, además utilizamos menos variables para el modelo pues al aumentar el número de variables era necesario más poder de cómputo al cual no fue posible acceder.

Conclusiones y trabajo futuro.

Con base en los resultados vistos durante el desarrollo del proyecto, puede concluirse que la base de datos consultada tenía información que ayudaba a predecir con una buena precisión el precio de los vehículos usados. Esto se constató en el caso de los dos mejores modelos de predicción, Random Forest y Gradient Boosting, que obtuvieron métricas de desempeño similar y fueron explicados por atributos similares. Para la entrada del sistema a producción sería más conveniente usar el segundo, debido a un menor sobreajuste.

De igual manera, se puede concluir que este modelo nos permite hacer predicciones de manera efectiva y bastante precisa, sin embargo es un modelo que mejora con el volumen de datos, por lo tanto tiene dos elementos en su contra; el primero, es que es altamente intensivo tanto en memoria como en computo lo que no lo hace altamente escalable, por lo tanto es recomendable mejorar el rendimiento de este antes de una implementación a un cliente, sin embargo es un buen ejercicio para entender cómo funcionan los sistemas de recomendación y nos permitió idear y probar diferentes métodos que nos permitieran medir que tan preciso es el modelo.

En este sentido, el ejercicio aquí presentado tiene un margen amplio de mejora, que puede resumirse en dos frentes. El primero, es que es posible utilizar una arquitectura con más capacidad que permita realizar el despliegue y procesar un volumen de información más grande de forma más rápida. La segunda es que al sistema de recomendación se le pueden agregar más features para capturar condiciones globales de la economía que puedan afectar el precio de los carros, entre estas sobresale las tasas de interés, la inflación y una medida del crecimiento económico. Finalmente, el modelo debería incorporar las preferencias tanto de los vendedores como de los compradores, que permita hacer una búsqueda más precisa y eficiente del vehículo que están buscando los clientes.

Ejecución del plan

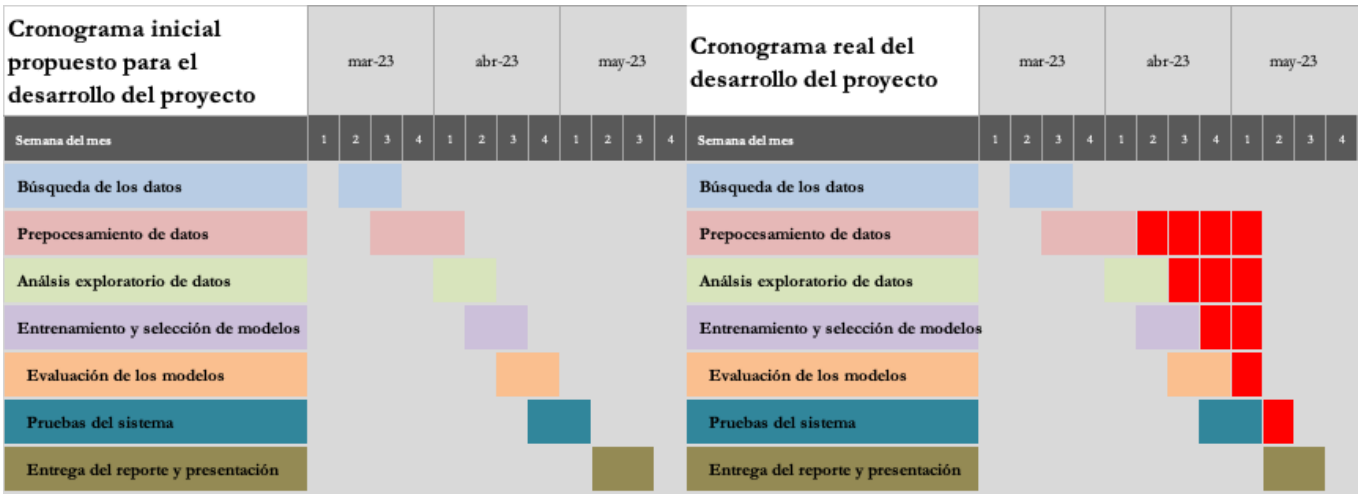

Figura 13. Cronograma planteado al inicio del proyecto vs Cronograma real de desarrollo

Con relación al cronograma planteado al principio del proyecto, las etapas en las que se presentó el cumplimiento esperado fueron la 1 de búsqueda de datos y la última de elaboración y entrega del reporte y la presentación. En las demás etapas hubo algunos atrasos o cambios de cronograma, especialmente en la limpieza y descripción de los datos que incluye en preprocesamiento y análisis exploratorio de datos. Esto se dio principalmente por el gran volumen de los datos que se usaron como insumo y a que la arquitectura para el desarrollo en general fue limitada. En esta parte, vale la pena decir que la cuenta educativa de Azure Databricks no permitía la creación de un clúster más potente para la realización de los cálculos. Se exploraron otras soluciones como Colab Pro de Google, pero en general el tiempo de ejecución de los procesos era lento. Dado lo anterior, para la construcción de los modelos de predicción se optó por tomar una una muestra de los datos originales y se ejecutaron en local, lo que a su vez facilitó la evaluación de estos y la posterior construcción del sistema de recomendación. En términos generales, la mayor parte del tiempo invertido, cerca del 70% del total, y en donde hubo mayores desfases con relación al plan inicial fue en la búsqueda de la arquitectura de desarrollo (incluida en la parte de procesamiento de datos), y la limpieza y procesamiento de los datos.

Implicaciones éticas

Este proyecto puede tener implicaciones éticas con temas de manipulación de mercados, el desarrollo de un modelo que prediga y recomiende precios debe ser tratado con seriedad pues quien sea el propietario, si este llegase a ser referente en el mercado, podría abusar de su poder y manipular los precios y las recomendaciones para beneficiar a algún agente o crear escenarios de corrupción, como hacer cobros irregulares para subir la reputación de algún participante. También los usuarios podrían abusar del algoritmo y manipularlo para obtener beneficios monetarios, un ejemplo sería una colusión entre vendedores corporativos, es decir, un concesionario para generar un alza de los precios de sus vehículos impactando de manera general los precios del mercado.

Aspectos legales comerciales

Hasta donde es de nuestro conocimiento, este proyecto en particular no tiene demasiadas limitaciones legales pues es con fines estrictamente académicos y los datos que se utilizan son de dominio público y están disponibles en Kaggle. Además, la información de cada vehículo es publicada varias

plataformas similares y puede ser consultada de manera libre y gratis por los usuarios. A su vez, su potencial comercial radica en que una plataforma que implemente un modelo de recomendación que logre equiparar de manera eficiente el precio solicitado por el vendedor y el ofrecido por el comprador, será una plataforma más deseada por el mercado, ya que en teoría se reduciría el tiempo que un vehículo está en el mercado y serán ofrecidos a un precio más cercano al óptimo un momento dado en el mercado, esto conllevaría a menores costos para los participantes y posiblemente mejores precios.

Bibliografía

Bukvić, L., Pašagić Škrinjar, J., Fratrović, T., & Abramović, B. (2022). Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning. Sustainability, 14(24), 17034. https://doi.org/10.3390/su142417034

Gajera, P., Gondaliya, A., & Kavathiya, J. (2021). Old Car Price Prediction with Machine Learning. International Research Journal of Modernization in Engineering, Technology and Science, 3(3). https://www.irjmets.com/uploadedfiles/paper/volume3/issue_3_march_2021/6681/1628083284.pdf

Lu, Jie, et al. "Recommender System Application Developments: A Survey." *Decision Support Systems*, Elsevier BV, June 2015, pp. 12–32. *Crossref*, doi:10.1016/j.dss.2015.03.008.

Pal, N., Arora, P., Sundararaman, D., Kohli, P., & Palakurthy, S.S. (2017). How much is my car worth? A methodology for predicting used cars prices using Random Forest. *ArXiv, abs/1711.06970*.

Pazzani, Michael J., and Daniel Billsus. "Content-Based Recommendation Systems." *The Adaptive Web*, Springer Berlin Heidelberg, pp. 325–41, http://dx.doi.org/10.1007/978-3-540-72079-9_10.

Schafer, J. Ben, et al. "Collaborative Filtering Recommender Systems." *The Adaptive Web*, Springer Berlin Heidelberg, pp. 291–324, http://dx.doi.org/10.1007/978-3-540-72079-9_9.

Venkatasubbu, Pattabiraman, and Mukkesh Ganesh. "Used Cars Price Prediction Using Supervised Learning Techniques." *International Journal of Engeniereing and Advanced Technology*, Blue Eyes Intelligence Engineering & Sciences Publication, 2019.