Universidad Francisco Marroquín
Christian Medina Armas
Machine Learning Model

Proyecto #2 -
Used vehicles price prediction model

Stephanie Grotewold
20210567
11 de abril del 2024

Introducción

En este informe se analizan los datos relacionados con las ventas de automóviles, centrándose en variables clave como modelo, año, precio, kilometraje, impuesto, millas por galón (mpg), tamaño del motor, tipo de transmisión, tipo de combustible y fabricante. El objetivo principal es proporcionar información detallada sobre las tendencias, patrones y relaciones dentro del conjunto de datos para comprender los factores que influyen en los precios y las ventas de automóviles.

Data

El conjunto de datos *vehicles_prices* es un conjunto de datos de características múltiples, con un total de 108,540 instancias y 10 características diferentes. Las variables son de diferentes tipos, incluyendo numéricas y categóricas. Las características incluyen información como el modelo del automóvil, el año de fabricación, el precio, la transmisión, el kilometraje, el tipo de combustible, el impuesto, el rendimiento en millas por galón, el tamaño del motor y la marca del automóvil.

Métodos

Análisis de datos numéricos

El conjunto de datos proporcionado ofrece una visión detallada de las características de varios vehículos. En promedio, los vehículos son del año 2017, aunque hay modelos que datan desde 1970 hasta 2060, lo cual es considerado un *outlier*, ya que ese año aún no ha pasado. El precio medio de los vehículos es de 16 mil, con un rango que va desde los 450 hasta los 159 mil. En cuanto al kilometraje, los vehículos han recorrido, en promedio, 23 mil millas, con un mínimo de 1 milla y un máximo de 323 mil millas. El impuesto medio que se paga por estos vehículos es de 120, aunque hay vehículos que no pagan impuestos y otros que pagan hasta 580. En términos de eficiencia de combustible, los vehículos consiguen en promedio 55 millas por galón, con un rango que va desde las 0.3 millas por galón hasta las 470.8 millas por galón. Finalmente, el tamaño medio del motor de estos vehículos es de 1.66, con motores tan pequeños como 0 y tan grandes como 6.6.

Análisis de datos categóricos

El conjunto de datos muestra que los cinco modelos de vehículos más comunes son el Focus, con 10,042 unidades; el C Class, con 7,646 unidades; el Fiesta, con 6,557 unidades; el Golf, con 4,863 unidades; y el Corsa, con 3,441 unidades. En cuanto a la transmisión, la mayoría de los vehículos (61,308) tienen transmisión manual, seguidos por los de transmisión semiautomática (24,903) y automática (22,319). Solo 10 vehículos tienen un tipo de transmisión categorizado como "Otro". Con relación al tipo de combustible, la mayoría de los vehículos (59,875) utilizan gasolina, seguidos por los que utilizan diésel (45,177). Los vehículos híbridos suman 3,229, mientras que 253 vehículos utilizan un tipo de combustible categorizado como "Otro". Sorprendentemente, solo hay 6 vehículos eléctricos. Por último, en cuanto a la marca, Ford es la más común con 23,419 vehículos, seguida por Mercedes (17,018), Volkswagen (15,157), Vauxhall (13,632), BMW (10,781), Audi (10,668), Toyota (6,738), Skoda (6,267) y Hyundai (4,860).

Análisis de normalidad

Los resultados de las pruebas de normalidad para las variables del conjunto de datos indican que ninguna de ellas sigue una distribución normal. Para el año, precio, kilometraje y tamaño del motor, tanto la prueba de D'Agostino's K^2 como la prueba de Anderson-Darling muestran que los datos no se ajustan a

una distribución normal, ya que el valor-p es 0.0 y el estadístico de Anderson-Darling es significativamente mayor que los valores críticos.

Matriz de correlación

La matriz de correlación revela varias relaciones entre las características de los vehículos. El año del vehículo tiene una correlación positiva con el precio y una correlación negativa con el kilometraje, sugiriendo que los vehículos más recientes suelen ser más caros y tener menos millas. El precio del vehículo está fuertemente correlacionado con el tamaño del motor. El impuesto sobre el vehículo tiene una correlación negativa con las millas por galón, indicando que los vehículos más eficientes suelen tener impuestos más bajos. Finalmente, el tamaño del motor está fuertemente correlacionado con el precio y negativamente con las millas por galón.

Variance Inflation Factor

El VIF es una medida que se utiliza para identificar la multicolinealidad en un modelo de regresión. En este caso, el año del vehículo, el precio, las millas por galón y el tamaño del motor muestran un VIF alto, lo que sugiere una alta correlación con otras variables. El impuesto y el kilometraje tienen un VIF más bajo, pero aún indican alguna correlación. Un VIF mayor a 5 puede indicar un problema de multicolinealidad, lo que puede hacer que los coeficientes de regresión sean inestables y difíciles de interpretar.

Modelos

Regresión Lineal: Este modelo asume una relación lineal entre las variables independientes (X) y la variable dependiente (Y). Ajusta una línea recta o superficie que minimiza las discrepancias entre los valores de salida predichos y reales. El cual es fácil de interpretar, eficiente, computacional, pero asume que la relación es lineal y es sensible a los valores atípicos.

Árbol de Decisión: Utiliza una estructura de árbol donde cada nodo representa una característica, cada rama representa una decisión y cada hoja representa un resultado. Es capaz de capturar relaciones no lineales y es fácil de interpretar, sin embargo, suele tener sobre ajuste.

Máquina de Vectores de Soporte (SVM): En este algoritmo, se traza cada elemento de datos como un punto en el espacio n-dimensional, donde n es el número de características, con el valor de cada característica siendo el valor de una coordenada particular. Tiende a tener un buen rendimiento en conjunto de datos medianos, pero es sensible a la escala de las características y requiere hiperparámetros.

Bosque Aleatorio: Los bosques aleatorios son una colección de árboles de decisión. Cada árbol en el bosque aleatorio escupe una predicción de clase y la clase con más votos se convierte en la predicción del modelo. Los bosques aleatorios evitan el problema de *overfitting* que enfrentan los árboles de decisión. Maneja automáticamente la importancia de las características y reduce la tendencia de sobreajuste, sin embargo, es más complejo computacionalmente.

Pasos a seguir de cada algoritmo:

1. **Separación de columnas numéricas y categóricas**: Las columnas del conjunto de datos se dividen en numéricas y categóricas para su posterior procesamiento.

2. **Definición de los pasos de preprocesamiento:** Para las columnas numéricas, se utiliza un *SimpleImputer* con la estrategia 'mean' para rellenar los valores perdidos con la media de la columna. Para las columnas categóricas, se utiliza un Pipeline que primero rellena los valores perdidos con la palabra *'missing'* y luego aplica un *OneHotEncoder* para convertir las categorías en variables binarias.
3. **Combinación de los pasos de preprocesamiento:** Se utiliza un *ColumnTransformer* para aplicar los pasos de preprocesamiento definidos anteriormente a las columnas numéricas y categóricas.
4. **Creación de un pipeline:** Se crea un Pipeline que primero aplica el preprocesador definido anteriormente y luego entrena un modelo específico (Regresión Lineal, Árbol de Decisión, Máquina de Vectores de Soporte o Bosque Aleatorio).
5. **División de los datos**: Los datos se dividen en conjuntos de entrenamiento y prueba, con un 70% de los datos para entrenamiento y un 30% para prueba.
6. **Entrenamiento del modelo**: El pipeline se ajusta a los datos de entrenamiento.
7. **Predicción**: Se hacen predicciones en los datos de prueba.
8. **Evaluación del modelo**: Se calcula el error cuadrático medio (MSE) y el error absoluto medio (MAE) para evaluar el rendimiento del modelo.

Las diferencias específicas entre los modelos se encuentran en el paso 4, donde se utiliza un modelo diferente para cada pipeline. Además, para el modelo de Máquina de Vectores de Soporte (SVM) y el modelo de Bosque Aleatorio, se añade un paso adicional de normalización de los datos numéricos utilizando *StandardScaler*, para normalizar los datos, en el paso de preprocesamiento.

Resultados

Las métricas utilizadas fueron el MAE y el MSE son métricas para evaluar modelos de regresión. El Error Absoluto Medio (MAE) mide el promedio de los errores absolutos entre las predicciones y los valores reales, siendo robusto ante valores atípicos. El Error Cuadrático Medio (MSE), por otro lado, calcula la media de las diferencias cuadráticas entre las predicciones y los valores reales, penalizando más los errores grandes.

Regresión de Vectores de Soporte (SVR):

- MSE: 62,188,055.67
- MAE: 4,594.35

Regresión Lineal:

- MSE: 21,846,034.94
- MAE: 2,898.03

Árbol de Decisión:

- MSE: 6,584,596.02
- MAE: 1,422.55

Regresión de Bosque Aleatorio:

- MSE: 4,198,148.76
- MAE: 1,157.62

Se intentó utilizar la búsqueda en cuadrícula para minimizar el error en dos modelos, el SVR y el Bosque Aleatorio. En el caso del SVR, el proceso tardó 602 minutos y 15.3 segundos, logrando reducir el MAE de 4,594.35 a 2,500.05.

Para el Bosque Aleatorio, se logró disminuir el MAE hasta 1140.99, pero el proceso tardó 277 minutos y 59.3 segundos. Además, se utilizó *RandomizedSearchCV*, donde se obtuvo un MAE de 1140.56. A pesar de las pruebas realizadas, se decidió utilizar el Bosque Aleatorio porque obtuvo el menor error. Durante la optimización del modelo de Bosque Aleatorio, se probaron diferentes configuraciones de hiperparámetros clave. El hiperparámetro *n_estimators* se evaluaron en valores de 100, 200 y 300, afectando la robustez del modelo y el tiempo de entrenamiento. La *max_depth* se ajusta entre *None*, 10 y 20 para evitar el sobreajuste limitando la complejidad de cada árbol. Por su parte, *min_samples_split* (valores de 2, 5 y 10) influye en la división de nodos internos, controlando el sobreajuste al requerir más muestras para dividir, lo que promueve divisiones más generalizables.

Se probó la eliminación de cada variable, pero se encontró que mantener todas las variables resultaba en el menor MAE. Aquí están los resultados:

- Excluyendo 'year': 1403.92
- Excluyendo 'model':1338.67
- Excluyendo 'mpg':1316.26
- Excluyendo 'mileage': 1302.63
- Excluyendo 'engineSize':1276.32
- Excluyendo 'transmission':1197.37
- Excluyendo 'make':1177.87
- Excluyendo 'tax':1170.17
- Excluyendo 'fuelType': 1165.89

Conclusión

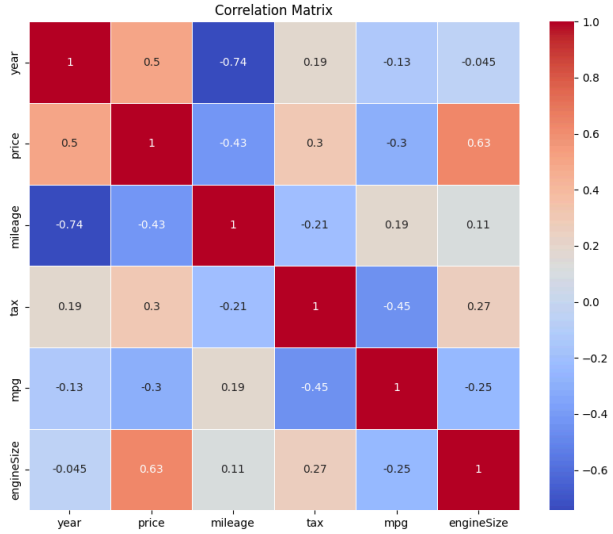
Al abordar el propósito de este informe, que es determinar la posibilidad de predecir los precios de los automóviles, se realizó un análisis numérico y categórico de un conjunto de datos. En este análisis, se identificaron relaciones significativas entre características como el año del vehículo y el precio, lo que sugiere que los modelos más recientes tienden a ser más caros. Además, se observó que el precio de los vehículos disminuye a medida que aumenta el kilometraje, y aumenta con el tamaño del motor y el impuesto. Esto indica que los coches más usados son generalmente más baratos, mientras que los coches con motores grandes y más valorados fiscalmente son más caros, destacando así su impacto en los precios.

Se aplicaron modelos de regresión, destacando que el Bosque Aleatorio es el más efectivo para predecir precios de automóviles. Destacando la importancia de variables como 'year', 'model' y 'mpg, y la exclusión de estas afecta el rendimiento. Se señaló la presencia de multicolinealidad entre variables, sugiriendo la necesidad de técnicas avanzadas de modelado o preprocesamiento de datos. Se enfatizó la importancia de una validación rigurosa de los modelos para evitar el sobreajuste y mantener la capacidad predictiva en situaciones reales. Finalmente, se recomienda mantener los datos actualizados y explorar técnicas avanzadas como el aprendizaje profundo para mejorar la precisión de los modelos predictivos.

Bibliografía

- *sklearn.model_selection.RandomizedSearchCV*. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html
- *sklearn.model_selection.GridSearchCV*. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html#sklearn.model_selection.GridSearchCV
- *¿Cuál es la diferencia entre el error cuadrático medio y el error absoluto medio en las métricas de aprendizaje automático?*(2024, April 6). www.linkedin.com. https://www.linkedin.com/advice/0/what-difference-between-mean-squared-error-tz1mc?lang=es&originalSubdomain=es

Anexos

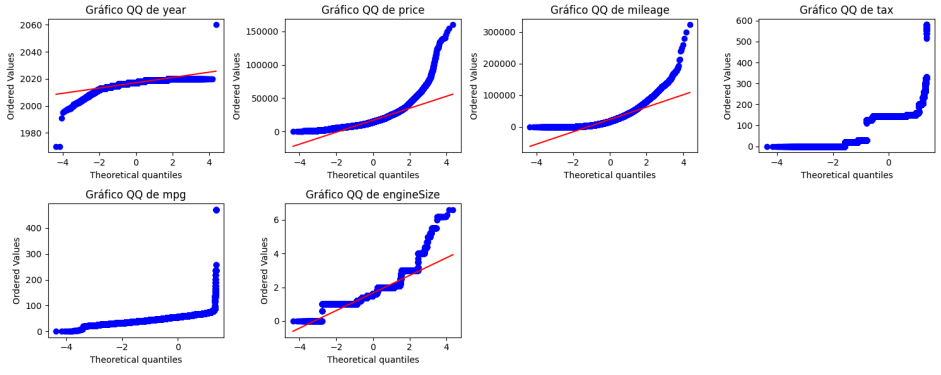


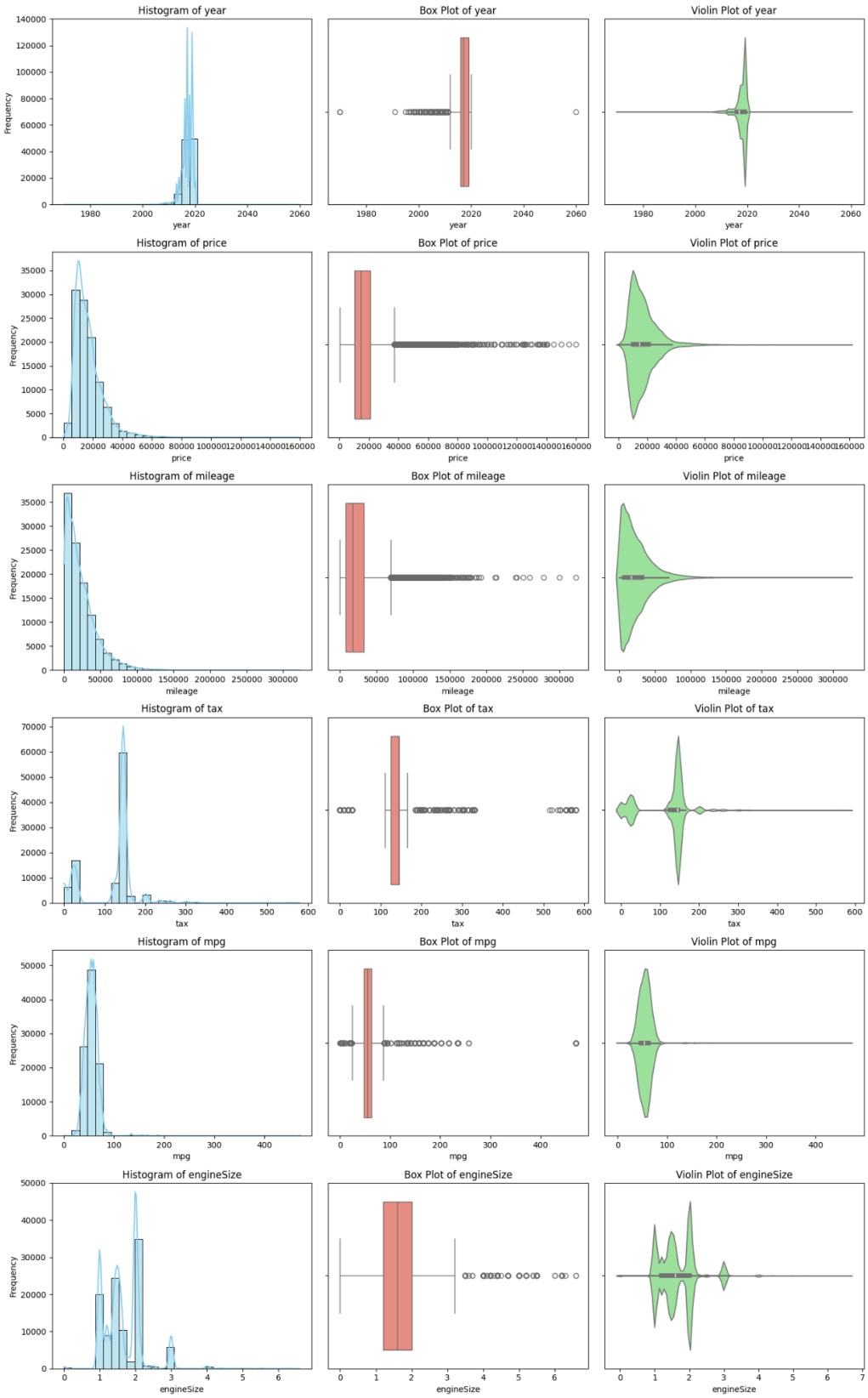

```
✓ 602m 15.3s

Best SVR Model - Mean Squared Error (MSE): 22017639.388901286
Best SVR Model - Mean Absolute Error (MAE): 2500.0569392267694
```

```
✓ 277m 59.3s

Best Random Forest Regression Model - Mean Squared Error (MSE): 4157175.006139692
Best Random Forest Regression Model - Mean Absolute Error (MAE): 1140.9944111919142
```




Histogram of year
Box Plot of year
Violin Plot of year
Frequency
year
Histogram of price
Box Plot of price
Violin Plot of price
price
Histogram of mileage
Box Plot of mileage
Violin Plot of mileage
mileage
Histogram of tax
Box Plot of tax
Violin Plot of tax
tax
Histogram of mpg
Box Plot of mpg
Violin Plot of mpg
mpg
Histogram of engineSize
Box Plot of engineSize
Violin Plot of engineSize
engineSize