Vehicle Prices

Steph Grotewold - 20210567

Agenda

01

EDA

02

Modelos

03

Resultados

04

Conclusiones

EDA

1. Se encontraron varios outliers en variables como año 2060 y valores faltantes en mpg y tax.
2. El automóvil más común es el Ford Focus con 10k.
3. La mayoría de vehículos son de transmission manual que utilizan gasolina.
4. Con relación al tipo de combustible, la mayoría de los vehículos (59,875) utilizan gasolina. Sorprendentemente, solo hay 6 vehículos eléctricos.
5. El año del vehículo y el precio tienen una relación positiva, 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.
6. El alto VIF de variables como el año del vehículo, el precio, las millas por galón y el tamaño del motor sugiere una fuerte correlación con otras variables, lo que podría afectar la estabilidad e interpretación de los coeficientes de regresión.

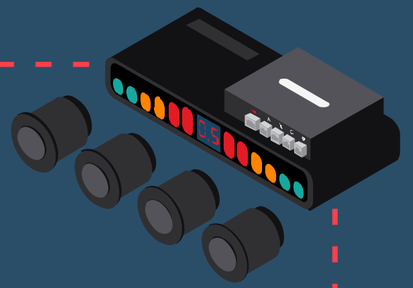
Pasos de los Modelos

1. **Separación de columnas numéricas y categóricas**
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.
5. **División de los datos 30/70**
6. **Entrenamiento del modelo**
7. **Predicción**
8. **Evaluación del modelo**: Se calcula el error cuadrático medio (MSE) y el error absoluto medio (MAE) para evaluar el rendimiento del modelo.

Error de los modelos

Modelos	MAE	MSE
SVR	4,594	62,188,055
Regresion Lineal	2,898	21,846,034
Árbol Decisión	1,422	6,584,596
Random Forest	1,157	4,198,148

Hiperparámetros



Sin hiperparámetros

MAE: 1,157

Con Grid Search CV

n_estimators: 100,200,300
max_depth: None, 10 y 20
min_sample_split 2,5,10

MAE:1,140.99

Con Randomized Search CV

regressor__n_estimators: 100, 200, 300
regressor__max_depth: None, 10, 20
regressor__min_samples_split:2, 5, 10

MAE: 1140.56

Error Absoluto Medio del modelo sin:

Con todas las variables el MAE disminuye a 1140

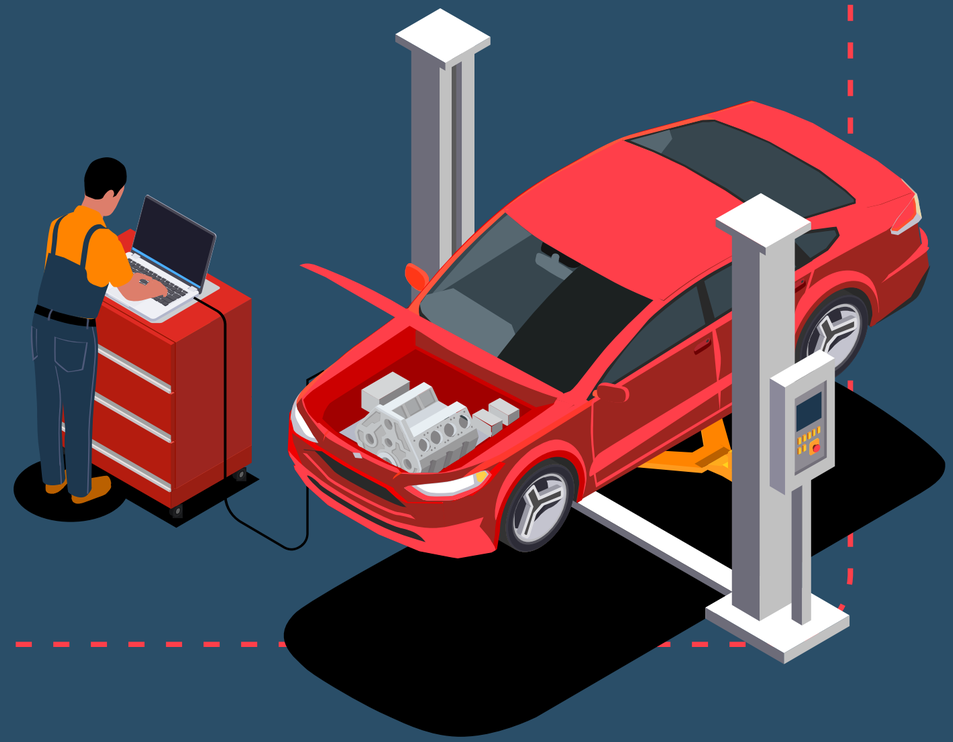
Conclusiones y recomendaciones

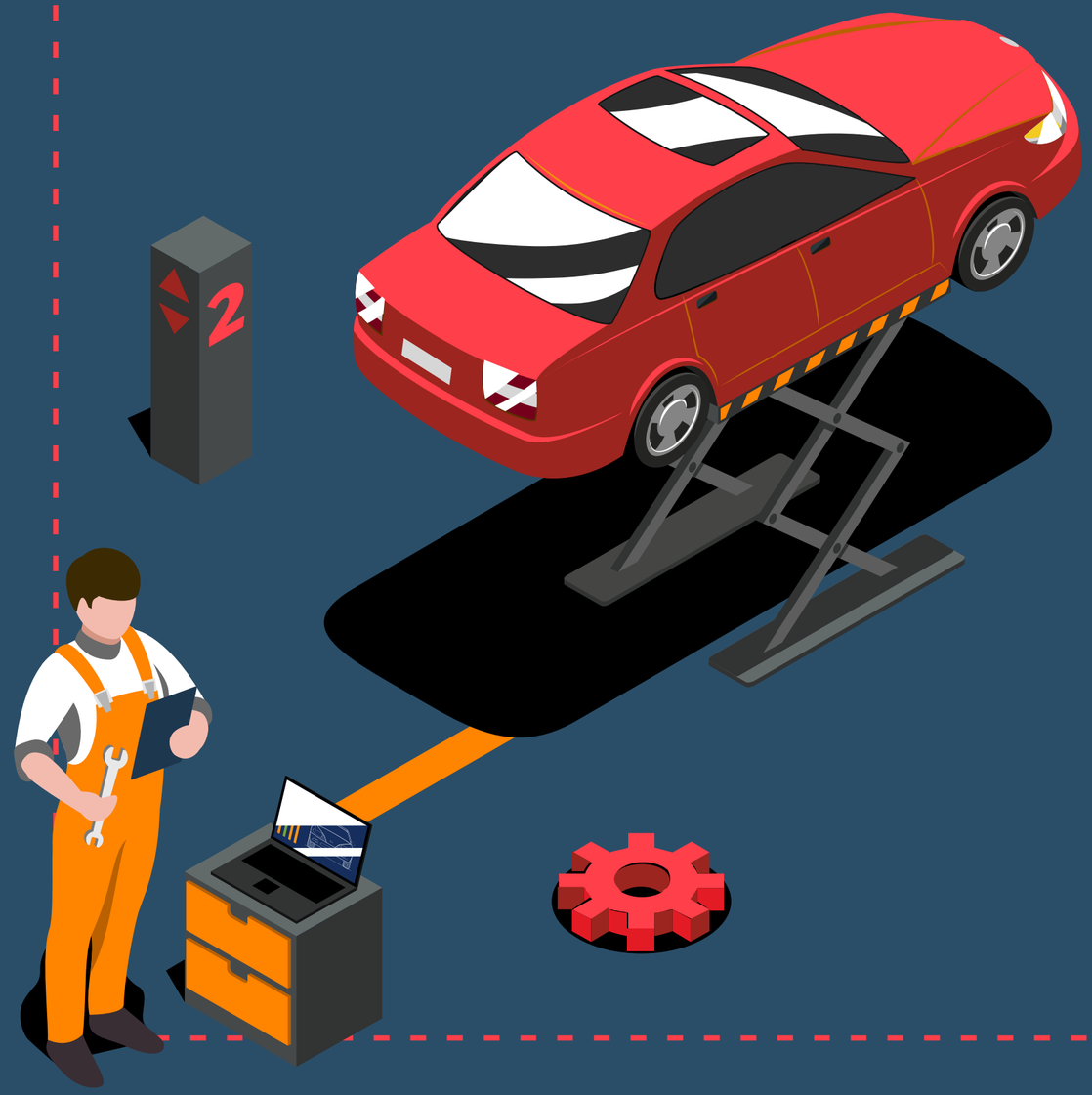
Conclusiones

- El Bosque Aleatorio es el más efectivo para predecir precios de automóviles.
- Los modelos más recientes tienden a ser más caros, mientras que los automóviles con más kilometraje suelen ser más baratos; además, los carros con motores grandes y más valorados fiscalmente también tienden a ser más costosos.
- Variables clave como 'year', 'model' y 'mpg' son importantes para la precisión del modelo.

Recomendaciones

- Mantener los datos actualizados para mejorar la precisión de los modelos.
- Balancear el data set porque hay una diferencia muy grande entre los tipos de combustibles.
- Considerar técnicas de deep learning para mejorar la precisión de los modelos predictivos.
- Considerar técnicas avanzadas de preprocesamiento de datos para abordar la multicolinealidad entre variables.





Thanks!