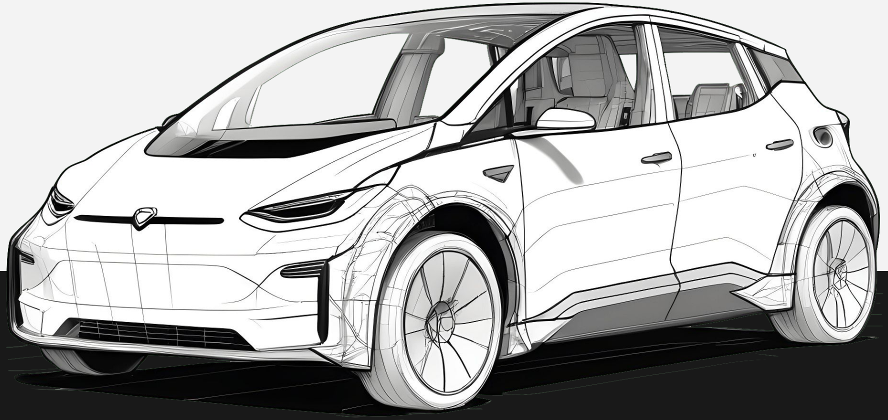
Predicting Prices of Second-Hand Cars

Aisyah Amatul Ghina
Flory
Paola Garay
Tiago Pedro



Project Overview

In the rapidly **growing used car market**, accurately predicting the price of second-hand vehicles is crucial for both buyers and sellers.

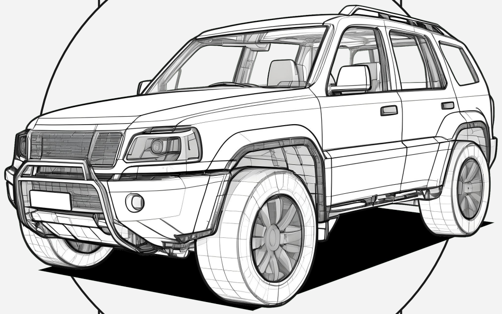
Estimating the right price based on its various attributes can help buyers make informed decisions, while sellers can set competitive prices.

Objective	to develop a machine learning model that can predict the price of used cars based on multiple attributes
Potential Impact	● Standardize price definition for second-hand cars ● Better readability of prices for consumers

Data Overview

Target

Price



Features

Brand
Model
Model year
Milage
Fuel type
Engine
Accident
Clean_title
Transmission
External color
Internal color

Shape

(187765, 13)

Source

Kaggle

Data Selection and Preparation

Understand every column

Researched technical details about cars

+

Fill in null and unclean data

Took some decisions based on cars characteristics' logic

→

Create functions for a clean dataset

Every cleaning functions have been sent to a separate file to scale the cleaning process for all of us

Feature Engineering

Normalization
One Hot Encoding

	Feature 1	Feature 2	Correlation
521	transmission_Automatic	transmission_Manual	0.900829
301	engine_size	cylinders	0.886290
262	horsepower	cylinders	0.697639
46	model_year	milage	0.675361
451	fuel_type_Gasoline	fuel_type_Hybrid	0.654406
261	horsepower	engine_size	0.605292
415	fuel_type_Electric	fuel_type_Gasoline	0.536552

Feature Selection

Low correlation between features
High correlation with target

```
# Checking features which has high correlation with Price
corr_matrix["price"].sort_values(ascending=False).head(10)
```

```
price            1.000000
milage           0.284189
horsepower       0.276135
model_year       0.236145
brand_ratio      0.214266
cylinders        0.132266
accident         0.125423
engine_size      0.096972
clean_title      0.089867
ext_col_Others   0.074695
Name: price, dtype: float64
```

Model Building and Evaluation

		TRAIN			TEST		
Model	**Parameters**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**
KNN Regression	K=10	16183.32	58457.25	0.25	17628.06	65764.58	0.08
Linear Regression		19194.52	63131.62	0.13	19068.93	63974.5	0.13
Decision Trees	Max depth: 5	17033.61	62159.81	0.15	16933.76	63331.70	0.14
Bagging and Pasting	Max depth: 5	17197.40	62313.73	0.15	17209.07	63574.70	0.14
Random Forest	Max depth: 5	16648.39	61360.77	0.17	16592.94	63239.19	0.15
Adaptive Boosting	Max depth: 5	333311.40	506211.80	-55.21	332764.11	507162.79	-53.98
Gradient Boosting	Max depth: 5	15782.15	56643.92	0.30	16320.64	63768.62	0.13

Hyperparameter Tuning Results

Random Forest (Best Params: Max Depth=5, Max Leaf Nodes=100):

- **Test MAE**: 16,570.11
- **Test RMSE**: 63,119.22
- **Test R^2**: 0.15

Gradient Boosting:

- **Test MAE**: 16,320.64
- **Test RMSE**: 63,768.62
- **Test R^2**: 0.13

The hyperparameter tuning, slightly improved the prediction accuracy (achieving a MAE of 16,570.11), **Random Forest is selected as the final model**

Key Findings and Insights

- **Top Features Influencing Price**:
  - **Mileage**, **Model Year**, and **Engine Size** were consistently the most important features across all models.
  - **Brand Ratio** provided additional value, helping models better distinguish between car brands and their impact on price.

- **Random Forest:** is the best-performing model, after hyperparameter tuning showed slightly higher $\mathbf{R^2}$ (0.15)

- **Gradient Boosting:** it has the lowest **MAE** (16,320.64) and competitive $\mathbf{R^2}$ (0.13), but shows more signs of overfitting.

Real-World Application and Impact

Practical Application

- Used Car Dealerships
- Insurance and Financial Services
- Online Car Marketplaces
- Consumers

Impact of the Model

- Market Efficiency
- Increased Accessibility
- Improved Decision-Making

Challenges

- Implement highly complex notions right after class
- Selecting the right model
- Merging our local work

Learnings

- Correlation between features and target are highly impacted by our data-cleaning choices
- Hard to create a good model right away

Future Work and Improvements

- **Improve Feature Engineering**
    - Transform categorical features, like external color, into ordinal variables based on their impact on price to improve model performance.

- **Incorporation of External Data Sources**
    - Consumer preferences and sentiment analysis
    - Macroeconomic Data: Car prices are influenced by macroeconomic factors such as interest rates, inflation, and fuel prices

Thanks!

Aisyah Amatul Ghina
Flory
Paola Garay
Tiago Pedro