Car Price Prediction using

Random Forest Regressor

Project Overview

The aim of this Data Science Project to focuses on building a machine learning model to predict car price for a company in Poland. Identifying car price according to their model, generation, year of production, mileage, and type and volume of engine.

Key Goals

- Perform extensive Exploratory Data Analysis (EDA) to understand feature behaviour.
- Pre-process the data (handling missing and duplicated values, encoding categorical features, scaling numerical features).
- Develop and tune a regression model to predict car price with low Mean Absolute error and achieve high variance (R2 score).

Data Source

The dataset used is the publicly available Kaggle Car Prices Poland dataset, which includes information about car features, year of formation, location and its prices**.**

Original Source: https://www.kaggle.com/datasets/aleksandrglotov/car-prices-poland

- **Target Variable:**

  Price: Price variable shows the cost of car in Poland currency PLN (approx. 1USD=1PLN)

- **Key Features:**

Unnamed: 0	It represent the serial number
mark	Car Manufacturer
model	Model of the car
generation_name	Formatted Generation Name of the car
year	Car Year of production
mileage	Car Mileage in Kilometres
vol_engine	Auto Engine Size

fuel	Engine Type according to their fuel type (electro, petrol, diesel)
city	Locality in Poland
province	Region of Poland

Methodology and Techniques

The project followed a standard Machine Learning workflow:

1. Exploratory Data Analysis (EDA)

- Histogram plot clearly shows that there is a lot of outlier in numerical features after removal of outlier we clearly notice the difference in histogram plot.
- Visualized the distribution of numerical features with histogram plot, noticing that car with latest model and low mileage have highest price.
- Analyzed the relationship between car feature and price rate with the help of bar graph.

2. Pre-processing

- **Handling Missing and duplicated Values:** Drop the missing and duplicated values from the data.
- **Remove outlier:** Remove the outlier from the data with the help of Quantile Method.
- **Encoding:** Used Label Encoding for the car feature variable i.e. mark, model, generation_name, city, province and fuel.
- **Drop features:** Drop the features i.e. {Unnamed: 0(because it is a serial number)}{city and Province (location not much effect the car prices)}.
- **Scaling:** Applied StandardScaler to the input features to standardize the input range.

3. Model Development

We compared three models: Decision tree, kn neighbour and Random Forest was selected due to its superior performance after initial benchmarking.

- **Hyper-parameter Tuning:** Performed GridSearchCV to optimize parameters such as n_estimators and random_state.

Results and Evaluation

The final tuned Random Forest Regressor achieved the following performance metrics on the test set:

Metric	Score
Price Mean value	$42062.14

Metric	Score
Mean Absolute Error	$5040.71
Mean Absolute Error %	14.8%
Root Mean Squared Error	$8040.00
R2 Score	0.93

Key Finding: The model demonstrated strong predictive power. Given the business goal of predicting car price according to their feature. The Mean Absolute Error percentage showing that the model predict prices with 14% of minute error and R2 Score of 0.93 indicates that the model explains 93% of the variance in car prices, demonstrating strong fitness for the business goal.