The UK Used Car Price Prediction

Yang Zheng
Data Science Initiative
https://github.com/yangzheng-brown/project_data1030/tree/main/Project

Introduction

In the car market, the price of a new car is mainly provided by manufacturers. On the other hand, the value of a used car is hard to determine since there are a variety of factors related to it. In my project, I am going to predict the price of used cars with the supervised machine learning models, and hope the best model will help people find the best deal on a used car. As the target variable of the project is car price which is a continuous variable, the problem is defined as a regression problem.

The dataset used in this project was from Kaggle created by Aditya (2020), it was collected from a UK online marketplace for car buyers and sellers called Exchange & Mart through web scraper. The dataset has six CVS files in total, and each file contains the same nine features of cars for a specific brand. I combined all six CVS files and added a new feature named "brand" as my input. Therefore, the processed dataset has 66,170 data points and 10 features.

The properties of the ten features are shown in Table 1. There are four categorical features and the rest are continuous. In categorical features, the brand feature has 6 categories, the model feature has 134 categories, the transmission has 4 categories, and the fuel type has 5 categories. The price feature contains the price of cars listed in Exchange & Mart by the time when data was collected. Feature of Road tax is the rate that the owner has to pay annually, which is determined by engine size or CO_2 emissions. Year, MPG, engine size, and mileage features describe the current conditions of the car.

Table 1. The Properties of Features

	Features	Feature Type	Data Type	Unit
1	Brand	Categorical	String	N/A
2	Model	Categorical	String	N/A
3	Transmission	Categorical	String	N/A
4	Fuel type	Categorical	String	N/A
5	Year	Continuous	Numeric	Year
6	Price	Continuous	Numeric	Pound (£)
7	Mileage	Continuous	Numeric	Mile

Table 1. The Properties of Features (Continued)

8	Road tax	Continuous	Numeric	Pound (£)
9	MPG	Continuous	Numeric	Mile per gallon
10	Engine size	Continuous	Numeric	Litre

The dataset for this project has been used in several other previous works already. One project is named "car_price_prediction" in Kaggle, and its author Sohail (2021) used this dataset to predict the prices of used cars and found that people in the UK like to drive cars with manual transmission because the manual transmission has the biggest portion compared to other types of transmissions. The prediction accuracy of this project through LGBMRegressor model is 0.947 in R^2 score, and 1,445.47 in MAE. Another project named "BMW price prediction" in Kaggle used only the BMW dataset and did the prediction for BMW cars. The author Leelakiatiwong (2020) found that more than 50% of BMW on sold in database are BMW 1 Series - 5 Series, and the accuracy of his model using XGBRegressor is 0.959 in R^2 score.

Exploratory Data Analysis (EDA)

In this project, the target variable is car price and the distribution of price (Figure 1) is right skewed. Through .describe function, I collected the information that the mean of the price was £16,798.16, the median is £14,690.00, and the standard deviation is £9,503.39. Because the maximum is £145,000.00 and the minimum is £495.00, the values vary over 3 orders of magnitudes. Through the calculation, the price above £36,182.5 is considered an outlier of the target variable, and there is a total of 2,510 outliers in the target variable.

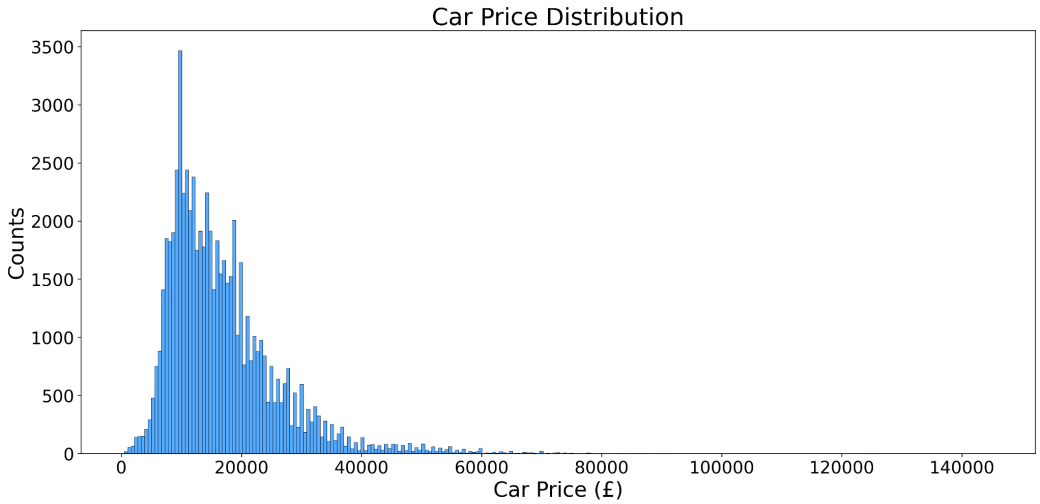

Figure 1. The target variable price is continuous variable. 68% of cars' selling price are between £7000 and £26000. The most expensive car is £145000, and the lowest price to buy a car is £495.

Figure 2 shows that most of the cars have MPG (Mile Per Gallon) of about 50-60, and most of the cars with MPG over 100 are hybrid cars. For the price of traditional types of cars, petrol and diesel powered vehicles, it has a negative correlation with the MPG feature, which means when MPG of a car increases, the price of that car decreases. However, for other types of cars, like hybrid and electric cars, there is no linear relationship between MPG and car price.

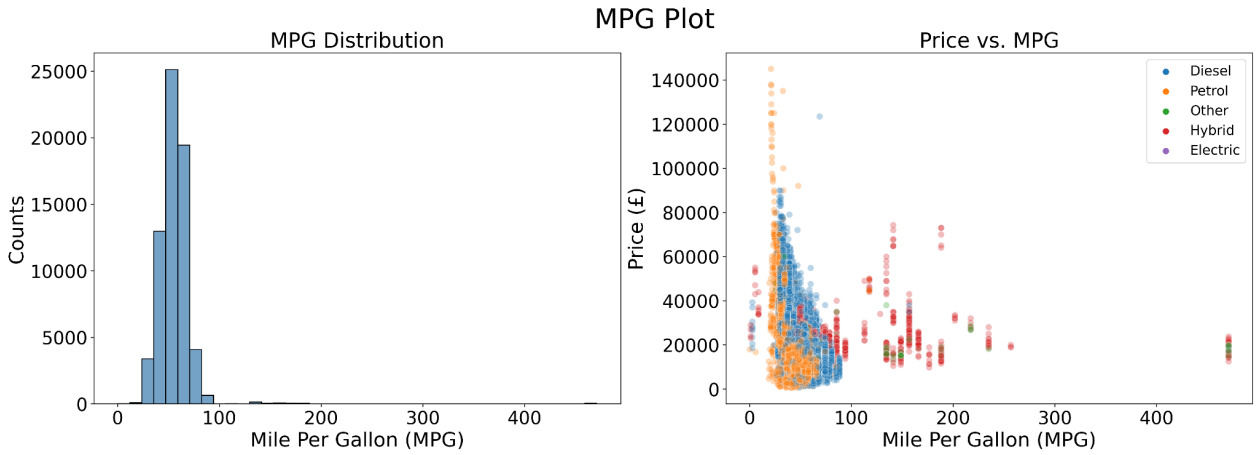

Figure 2. Majority of the cars have MPG around 50-60.
The most cars with MPG over 100 are hybrid cars.

From Figure 3 we can conclude that people in the UK prefer driving cars with manual transmissions rather than automatic and semi-auto transmissions. One of the reasons for this preference is that the average price of the manual car is lower than other types of cars. I did some additional research online and found that manual car is less fuel consumed than others (Gautam, 2020). The dataset also shows that most high-end cars in the UK are automatic or semi-auto.

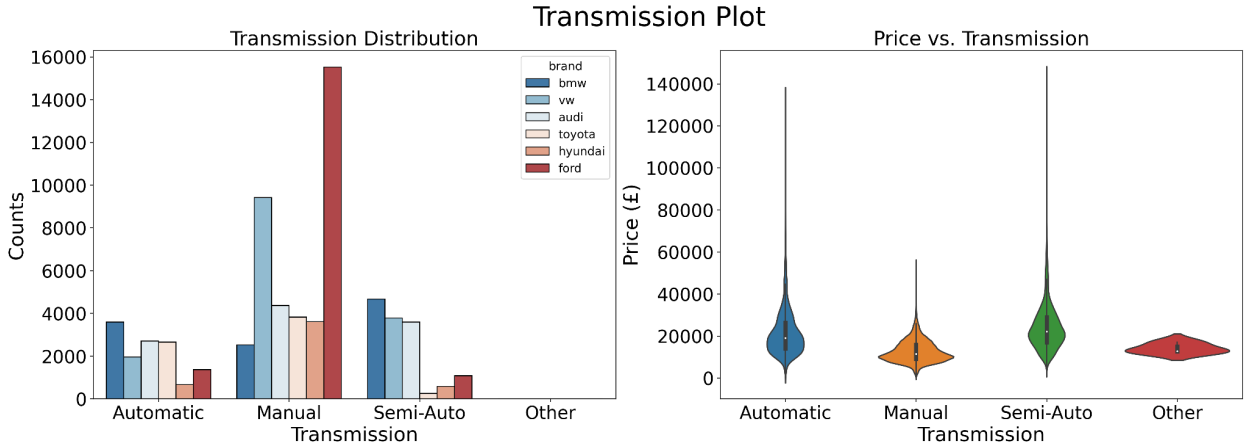

Figure 3. Most of cars in the UK are manual, and the
averge price of manual car is lower than others.

A correlation matrix (Figure 4) is a useful tool for understanding the relationship between the target variable and features. The plot below shows a moderate positive correlation between price and engine size, as well as price and year of manufacturing. A small correlation between the price and road tax can be observed. Furthermore, MPG and mileage both have low negative relations with the price.

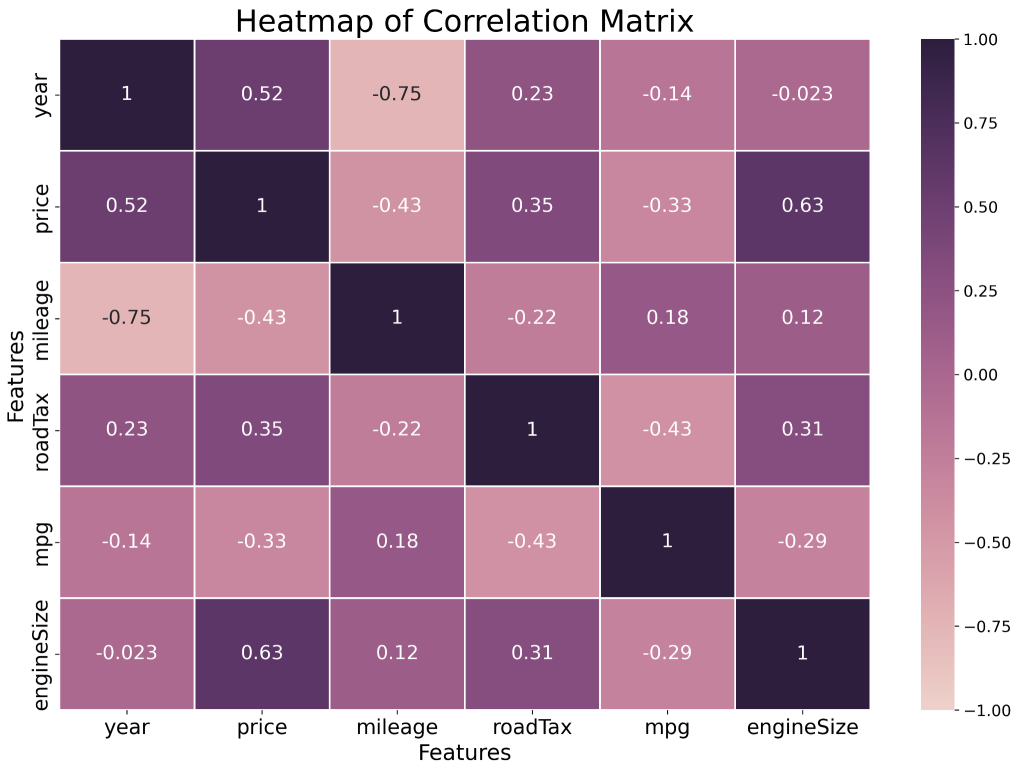

Figure 4. Year and engine size have a positive correlation with price. Mileage and MPG have a negative correlation with price.

Methods

The method of data splitting applied to this project is basic splitting because the size of the dataset is large and all data are independent and identically distributed (I.I.D.). Neither group structure nor time-series data are in the dataset. Thus, I split the dataset with the basic train test split function, and assign 80%-10%-10% as splitting ratio (80% of the total data goes into the train set and evaluation and test sets have 10% of the total data, respectively). As a result, the number of data points in the train set was 52,935, the number of data points in the validation set was 6,617, and the test set had the same number of instances as the validation set. In this way, I can ensure that there is enough data in the train set that could be used to train the model and guarantee the number of data points in validation and test sets is sufficient.

In the dataset, excluding the price variable, four features are categorical and the rest five features are continuous. Because all categorical data could not be ranked nor ordered, so I applied OneHotEncoder to categorical features. For the continuous features, I applied

MinMaxScaler to the engine size feature as it has upper and lower boundaries and StandardScaler to the rest of the continuous features as they have a tailed distribution. Before the preprocessing, the number of features in the train set was 9, while it became 153 after preprocessing. The reason train dataset became much larger after preprocessing is that OneHotEncoder was applied to the model feature which has 134 categories.

The ML algorithms used in this dataset are Lasso, Ridge, RandomForestRegressor, KNeighborsRegressor, and XGBRegressor. Lasso and Ridge are linear models, while others are non-linear models. The details about parameter tuning for ML models are summarized in Table 2. Additionally, RMSE (root mean squared error) was determined as the evaluation metric because the project is a regression problem.

Table 2. Parameter Tuning for ML Models

	ML model	Parameter	Values tuned
1	Lasso	alpha	10^{-3},10^{-2},10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}
2	Ridge	alpha	10^{-3},10^{-2},10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}
3	Random forest	max_depth	1, 3, 10, 30, 100
		max_features	0.25, 0.5, 0.75, 1.0
4	KNeighbors	n_neighbors	1, 125, 250, 375, 500
		weights	"uniform", "distance"
5	XGBoost	max_depth	1, 3, 10, 30, 100

In machine learning pipeline, two kinds of uncertainties could be generated. One is splitting uncertainty and the other is non-deterministic model uncertainty. Five random states were applied to the pipeline to measure the uncertainties due to splitting. As the result in Figure 5, linear models have lower uncertainties than non-linear ones and Ridge holds the lowest uncertainty among all models. However, the mean RMSE scores of linear models are much higher than non-linear models. The right-side plot in Figure 5 shows that Random Forest is the only non-deterministic model and all others are deterministic.

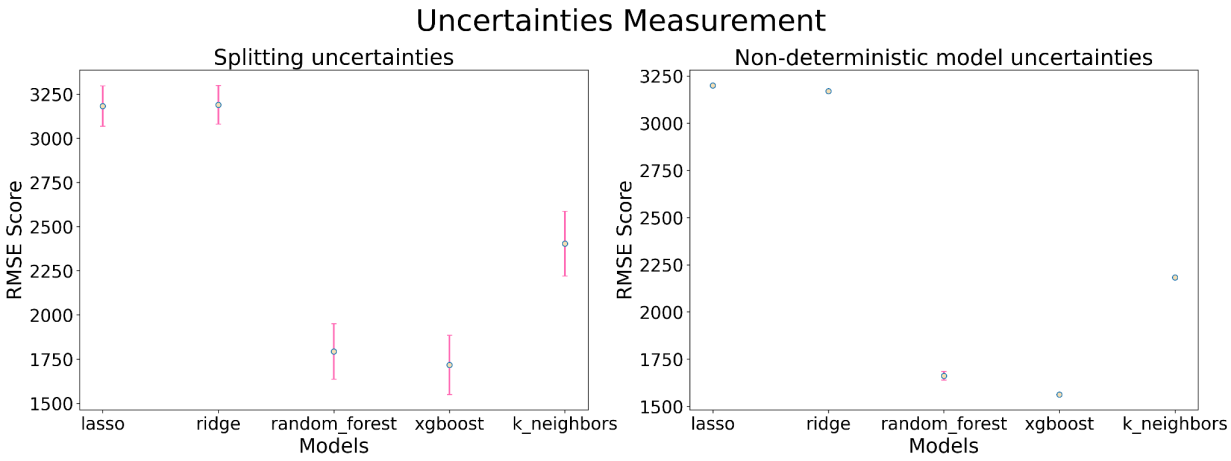

Figure 5. Ridge has the lowest uncertainty due to splitting among five models, and the random forest is the non-deterministic, and others are deterministic.

Results

To calculate the baseline RMSE score, the average value of test set was set to be the predicted value for all data points. With the formula of RMSE, the mean baseline score for five random states could be calculated, and the result was 9,340.22 (£). Among models, XGBoost is the most predictive model in this case, and it is 45.39 standard deviations above the baseline. The detailed performances of ML models are shown in Table 3.

Table 3. Performances of ML Models Based on the RMSE Score (£)

	ML model	Mean RMSE (£)	Standard deviation (std)	Baseline RMSE (£)
1	Lasso	3,182.45	114.12	9,340.22
2	Ridge	3,189.63	109.52	
3	Random forest	1,793.22	157.23	
4	KNeighbors	2,404.35	183.34	
5	XGBoost	1717.54	167.92	

To have a better understanding of the best model, an inspection of feature importance was necessary. Thus, three different methods—permutation (Figure 6), SHAP (Figure 7), and total gain—were applied to evaluate the feature importance. Although different methods used different metrics, the top 3 most important features in the three methods are identical: engine size, year, and MPG. Moreover, the result of feature importance mostly matches the correlation heatmap (Figure 4) that a greater absolute value of correlation with target variable means the larger importance of the feature. The least important feature is inconsistent in three methods, and they were onehot__model_Z4, onehot__model_Eos, and onehot__model_Escort, respectively.

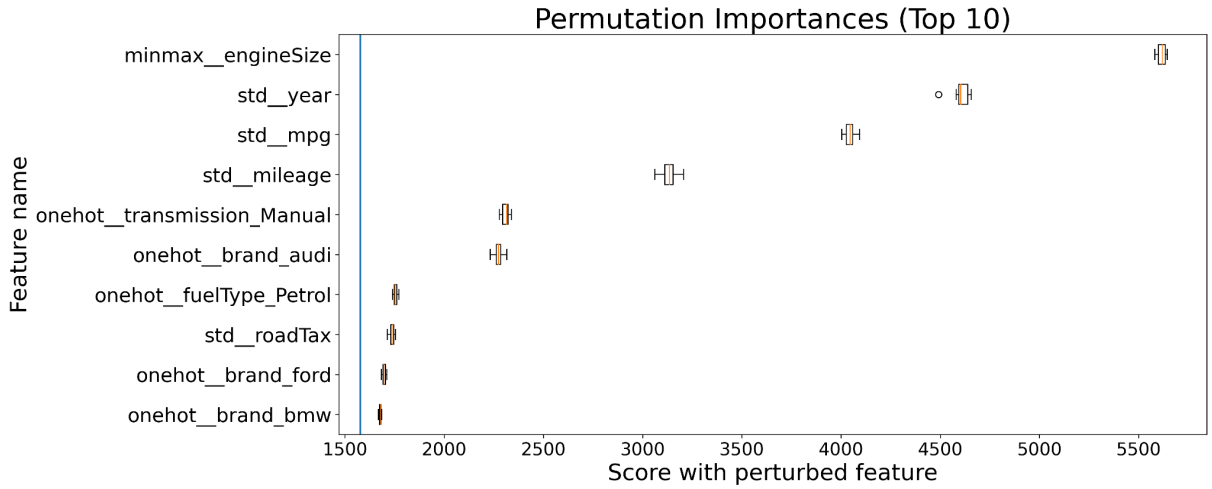

Figure 6. Engine size, year and MPG are the most important feature in permutation method.

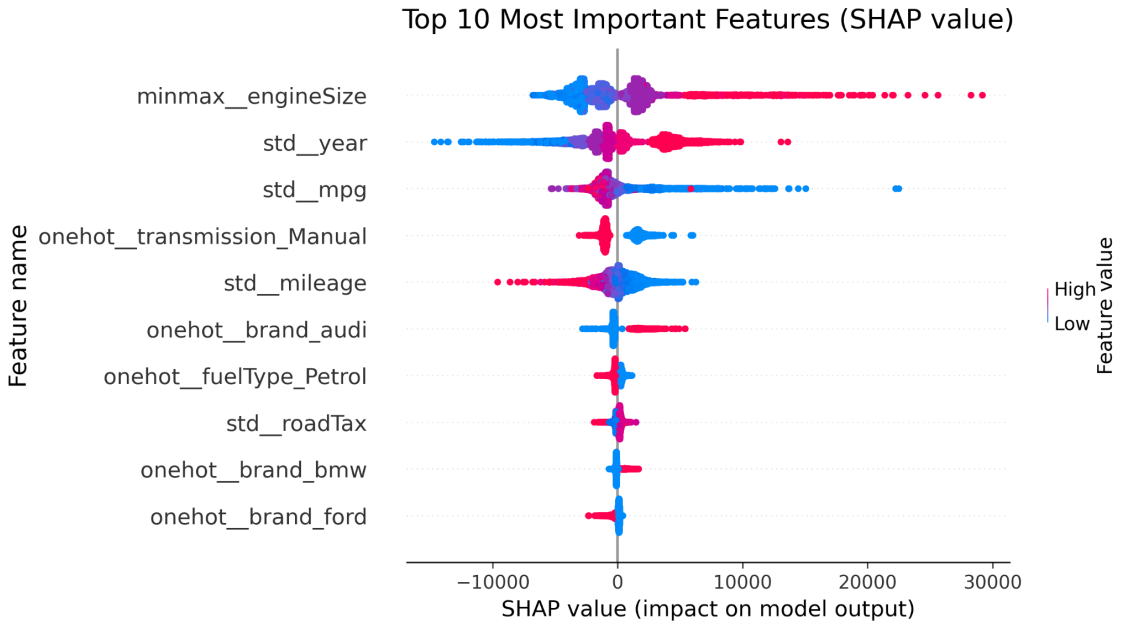

Figure 7. Engine size, year and MPG are the most important feature in SHAP method.

Regarding the local importance, data points with index 0 (Figure 8), 3000, and 6000 were analyzed. The major factors that push the car price up or down in disparate data points are alike, and they are engine size, year, mileage, and MPG. Those major attributes of specific points are also essential features in global feature importance. The conclusion of the feature importance analysis is that engine size, year, MPG and mileage could largely influence the price of the car, which was consistent with our intuition.

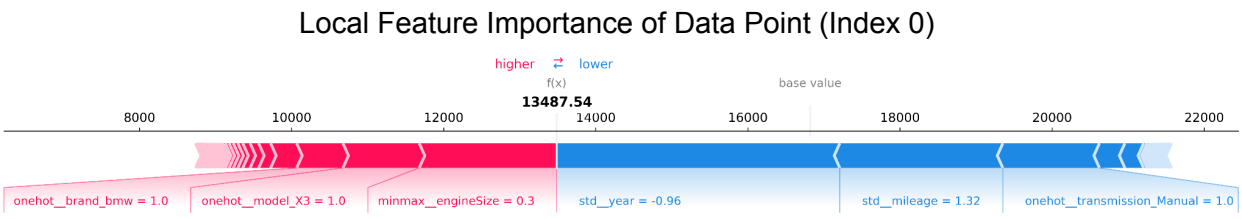

Figure 8. For data point with index 0, engine size is the major factor that increased the car price, and the manufacturing year is the factor that pulled the price down.

Outlook

Some hyperparameters are set with default values in ML models, however, those values may not be the best choice for this dataset, so tuning these parameters would potentially improve the model. Another effective way would be feature engineering which means we interpret some hidden relations between existing features and create new features for these relations. The third way of increasing accuracy is collecting additional features from Exchange & Mart, such as number of accidents, number of owners, interior quality, etc. Moreover, since only five ML models were applied in the project, the fourth way is to implement more models to see whether we could find a better model for this dataset.

Reference

Sohail, P., 2021, *car_price_prediction*, https://www.kaggle.com/code/parvezsohail/car-price-prediction

Leelakiatiwong, W., 2020, *BMW price prediction*, https://www.kaggle.com/code/wirachleelakiatiwong/bmw-price-prediction

Gautam, S., 2020, *Why Does Europe Prefer Manual Cars Over Automatic Ones*, https://blog.getmyparking.com/2020/01/20/why-does-europe-prefer-manual-cars-over-automatic-ones/

Aditya, 2020, *100,000 UK Used Car Data set*, https://www.kaggle.com/datasets/adityadesai13/used-car-dataset-ford-and-mercedes