The UK Used Car Price Prediction

Yang Zheng
Data Science Initiative

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 due to a variety of factors. 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 a reasonable deal on a used car. As the target variable car price is a continuous variable, the problem is defined as a regression problem.

The dataset was created by Aditya (2020) in Kaggle, but it was originally collected from a UK online marketplace for car buyers and sellers called Exchange & Mart. 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 CVS files and added a new feature named "brand" as input. After combining, the dataset has 66,170 data points and 10 features.

The properties of the ten features are shown in Table 1. Four of the features are categorical features and the remaining are continuous. In categorical features, there are 6 categories in the brand feature, 134 categories in the model feature, 4 categories in the transmission feature, and 5 categories in the fuel type feature. The price contains the car prices listed in Exchange & Mart by the time when data was collected. 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

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

Table 1. The Properties of Features (Continued)

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

The dataset for this project has been used in several other previous works already. One project is named "car_price_prediction" in Kaggle. 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 percentage compared to other types of transmissions. The prediction accuracy of his LGBMRegressor is 0.947 in R^2 score, and 1,445.47 in MAE (mean absolute error). Another Kaggle project named "BMW price prediction" used only the BMW dataset and did the prediction for BMW cars. The author Leelakiatiwong (2020) found that more than 50% of BMWs on sold in databases are BMW 1 Series - 5 Series, and the accuracy of his 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 discovered that the mean of the price was £16,798.16, the median was £14,690.00, and the standard deviation was £9,503.39. Because the maximum was £145,000.00 and the minimum was £495.00, the values varied over 3 orders of magnitudes. Through the calculation, the price above £36,182.5 was considered an outlier of the target variable, and there were a total of 2,510 outliers in the target variable.

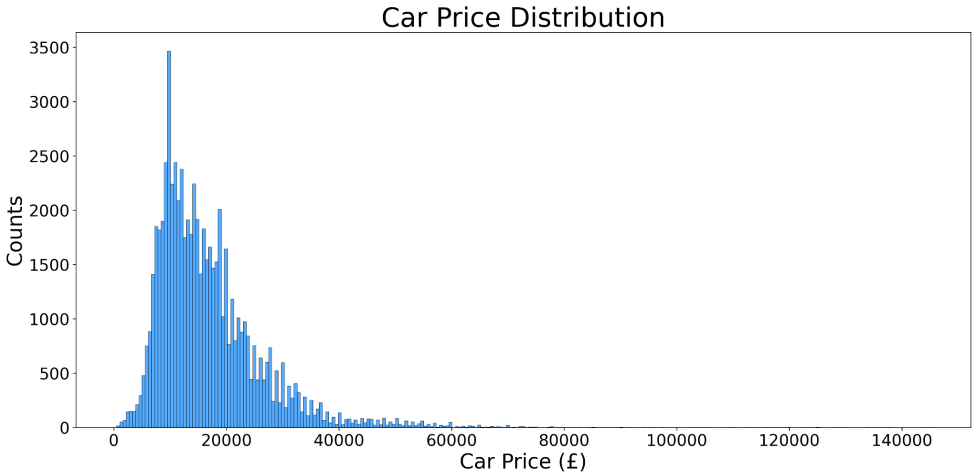

Figure 1. The target variable price is a continuous variable. 68% of selling price are between £7,000 and £26,000. The most expensive car is £145,000, and the cheapest is £495.

Figure 2 shows that most of the cars have 50-60 MPG (Mile Per Gallon), and most of the cars with MPG over 100 are hybrid cars. The price of traditional cars (petrol and diesel-powered vehicles) have a negative correlation with the MPG feature, which means when MPG 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.

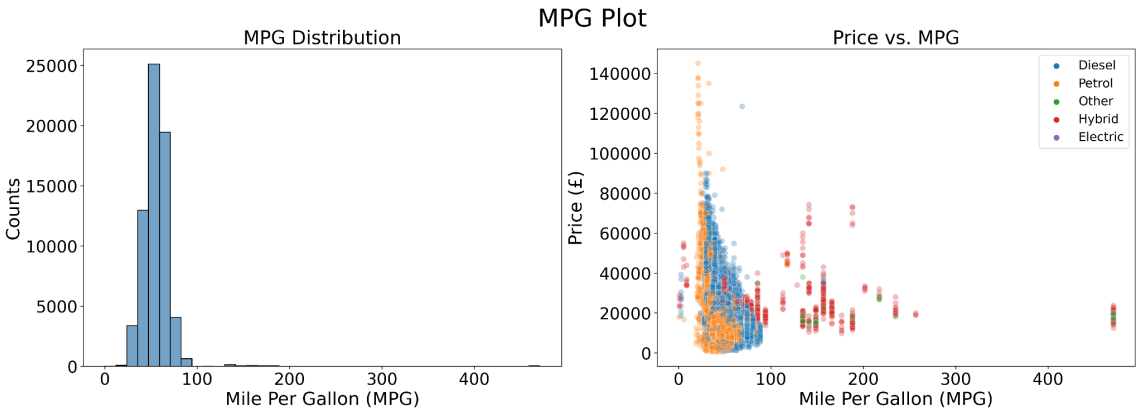

Figure 2. Majority of the cars have MPG around 50-60.
The most cars with MPG over 100 are hybrid cars.

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

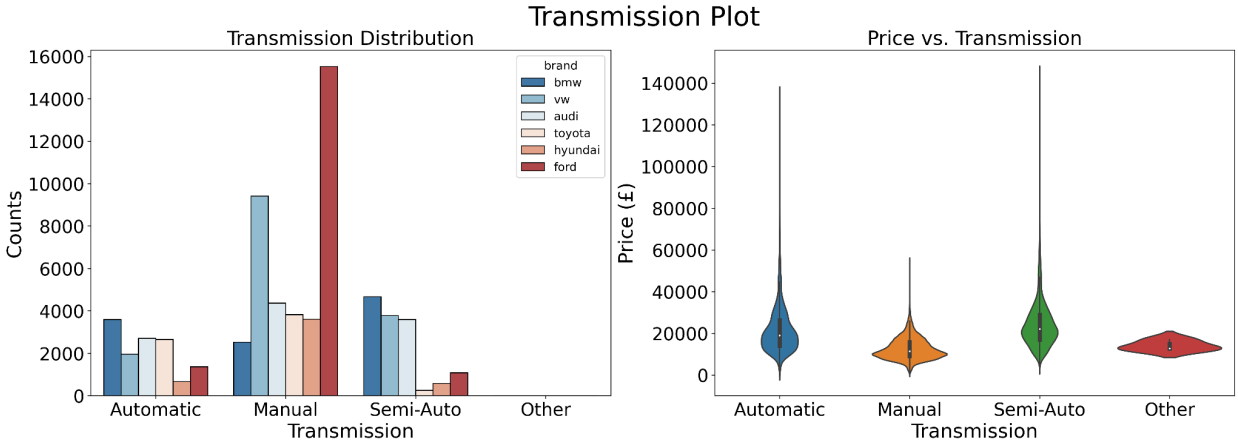

Figure 3. Most of cars in the UK are manual, and the
averge price of manual car is lower than others.

A correlation matrix is a useful tool for understanding the relationship between features. Figure 4 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 relation with the price.

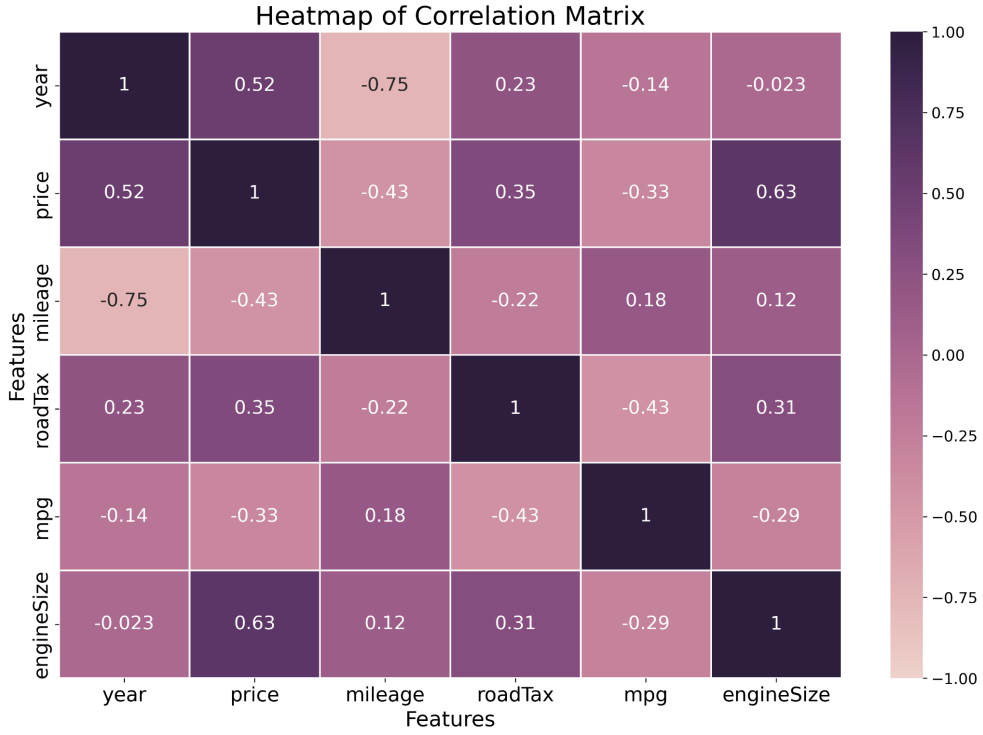

Figure 4. Year and engine size have a positive correlation with price. Mileage and MPG have a negative correlation with price.

Methods

Five random states were chosen to loop through the pipeline developed for the project. In each state, the pipeline would split and preprocess the data. Then perform the basic hyperparameter tuning for the ML algorithm to find the best parameters based on the evaluation metric. The project uses RMSE (root mean squared error) as evaluation metric since the project is a regression problem.

The method of data splitting 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 to train the model and guarantee the number of data points in validation and test sets is sufficient.

Because all categorical data could not be ranked nor ordered, during preprocessing I applied OneHotEncoder to categorical 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, there were 9 features in the train set and 153 afterward. The reason that the train dataset became much larger is that OneHotEncoder was applied to the model feature which has 134 categories.

The ML algorithms applied in this dataset are Lasso, Ridge, RandomForestRegressor, KNeighborsRegressor, and XGBRegressor. Lasso and Ridge are linear models, while the others are non-linear. The details about parameter tuning for ML models are summarized in Table 2.

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 a 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.

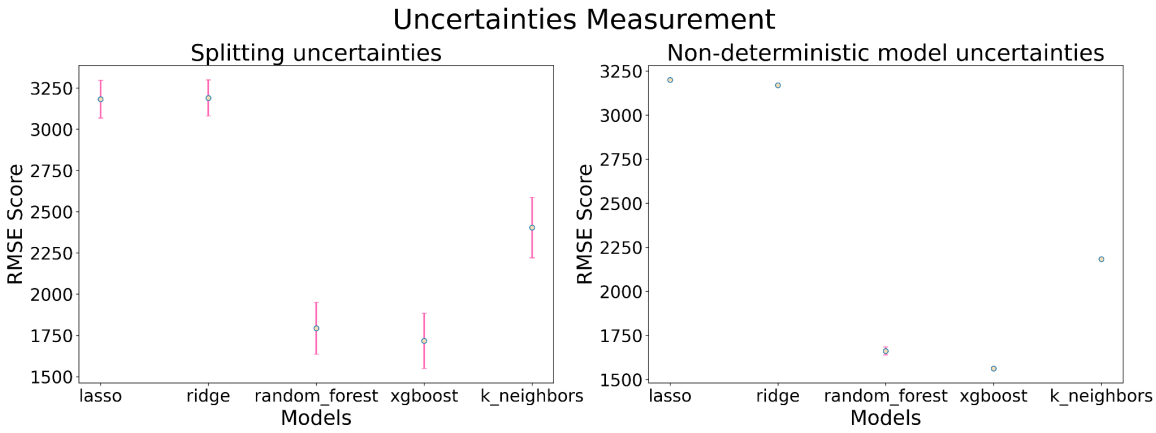

Figure 5. Ridge has the lowest uncertainty due to splitting among five models, and the random forest is the non-deterministic, and others are deterministic.

Results

To calculate the baseline RMSE score, the average price value of the 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 (£). After running through all models, XGBoost was the most predictive model, and it was 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 visualize the performance of XGBoost, a scatter plot of true vs. predicted values is provided in Figure 6.

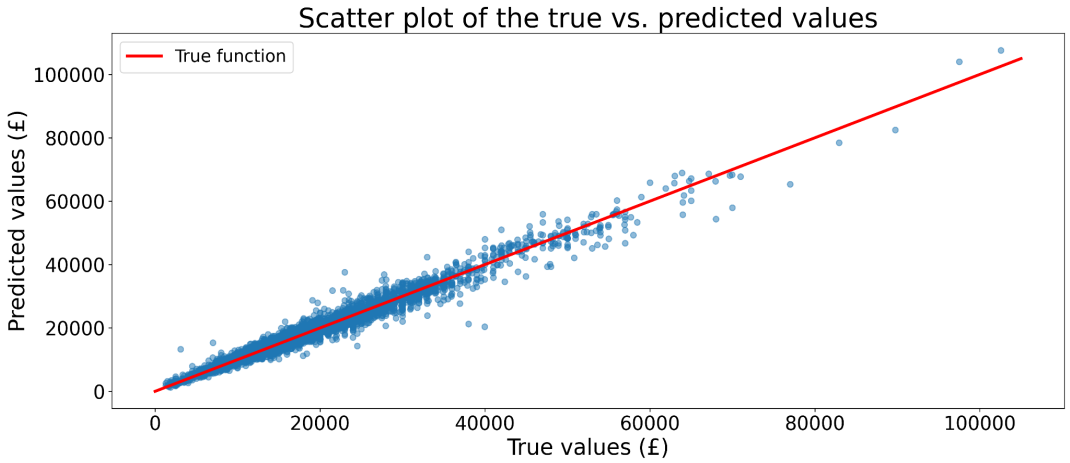

Figure 6. Most of predicted values are along the true function line.

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

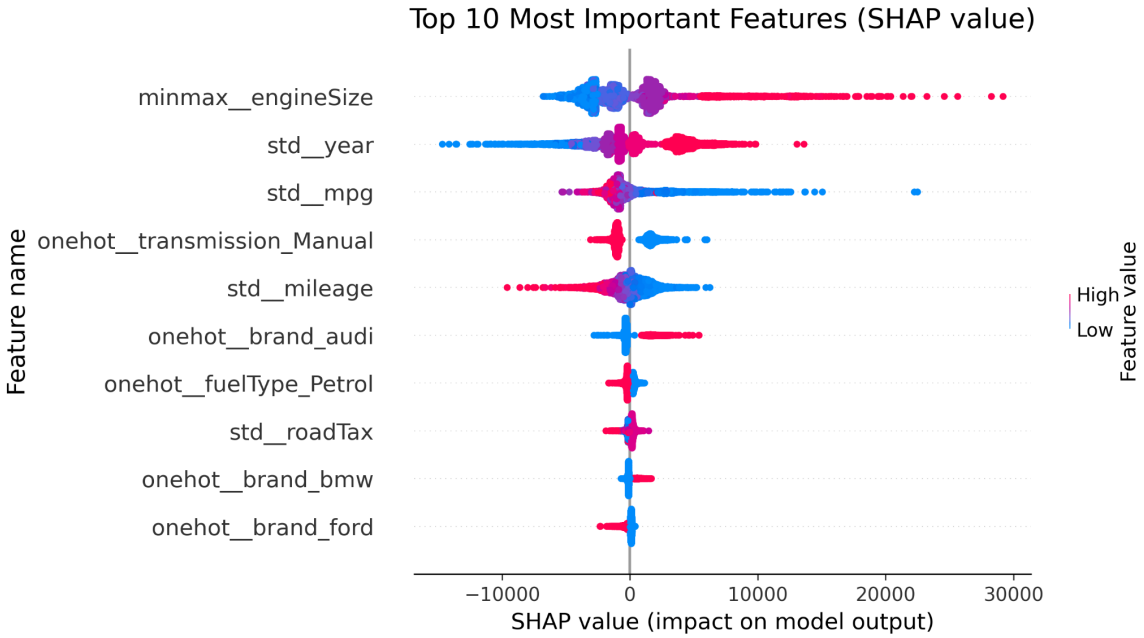

Figure 7. Engine size, year and MPG are the most important feature in SHAP method.

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: 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.

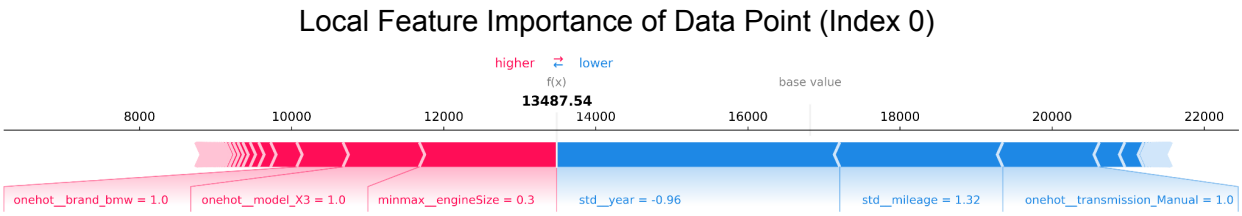

Figure 8. For data point with index 0, engine size is the major factor that increased the car price, and the manufacturing year is the factor that pulled the price down.

Outlook

Some hyperparameters are set with default values in ML models. However, those values may not be the best choice for this dataset. Thus, tuning these parameters would potentially improve the model. Another effective way would be feature engineering where 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 accidents, owners, interior quality, etc. Moreover, as 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

GitHub Repository

https://github.com/yangzheng-brown/project_data1030.git