The UK Used Car Price Prediction

Yang Zheng
Data Science Initiative

Introduction

In the car market, the price of a new car is typically set by manufacturers, but the value of a used car is hard to determine. This is because a used car's value can be affected by factors such as its age, mileage, and condition. In my project, I aim to predict the prices of used cars using 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 for this project was created by Aditya (2020) on Kaggle, and it was originally sourced from Exchange & Mart, a UK online marketplace for cars. The dataset includes six CVS files, each containing data on the same nine features (such as models, year, mileage, etc.) for a specific brand of car. To prepare the data for analysis, I combined all CVS files and added a new feature named “brand” to identify the brand of each car. The final dataset includes 66,170 data points, with 10 features per car (including the brand).

The properties of the ten features are shown in Table 1. Four of the features are categorical features and the remaining are continuous. The categorical variables have different numbers of categories, with the "brand" having six categories, the "model" having 134 categories, the "transmission" having four categories, and the "fuel type" having five categories. The continuous variables provide specific numeric values for each car, such as the listed price of cars, the annual road tax rate which is based on engine size or carbon dioxide emissions, and the manufacturing year, MPG (Mile Per Gallon), engine size, and mileage 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 a number of previous works that focus on predicting car price. For example, Sohail (2021) used the dataset in a Kaggle project called "car_price_prediction" to predict the prices of used cars in the UK. Sohail found that cars with manual transmission were the most popular, and he was able to achieve a prediction accuracy of 0.947 in R^2 score and 1,445.47 in MAE (mean absolute error) using a LGBMRegressor model. Another Kaggle project named ”BMW price prediction” used only the BMW dataset to predict prices of BMW cars. The author Leelakiatiwong (2020) found that more than 50% of BMWs in databases are BMW 1 Series - 5 Series. The accuracy of his XGBRegressor is 0.959 in R^2 score.

Exploratory Data Analysis (EDA)

In this project, the target variable is the price of used cars. As shown in Figure 1, the distribution of car prices is right-skewed, with a long tail of high-priced cars. 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. The range of prices spanned three orders of magnitude, from £495.00 to £145,000.00. 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.

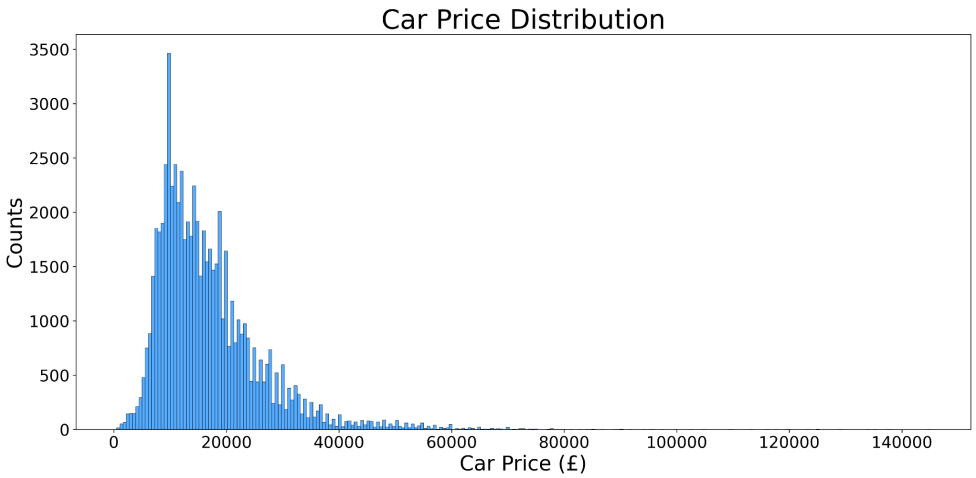

Figure 1. The target variable price is a continuous variable. 68% of selling price are between £7,000 and £26,000. The most expensive car is £145,000, and the cheapest is £495.

Figure 2 shows that majority of the cars have MPG values between 50 and 60. Most of the cars with MPG values 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 hybrid and electric cars, there is no linear relationship between MPG and car price.

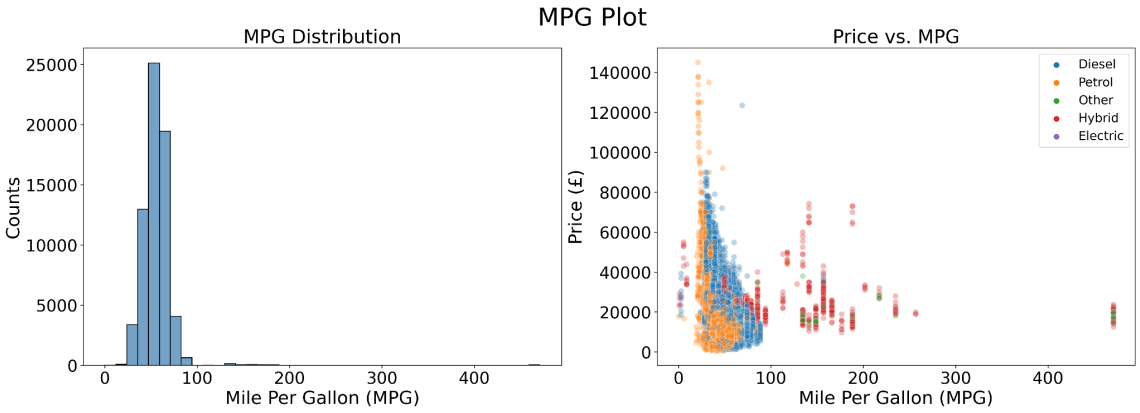

Figure 2. Majority of the cars have MPG around 50-60.
The most cars with MPG over 100 are hybrid cars.

Figure 3 shows that manual transmission cars are the most popular in the UK. 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 are more likely to be automatic or semi-automatic.

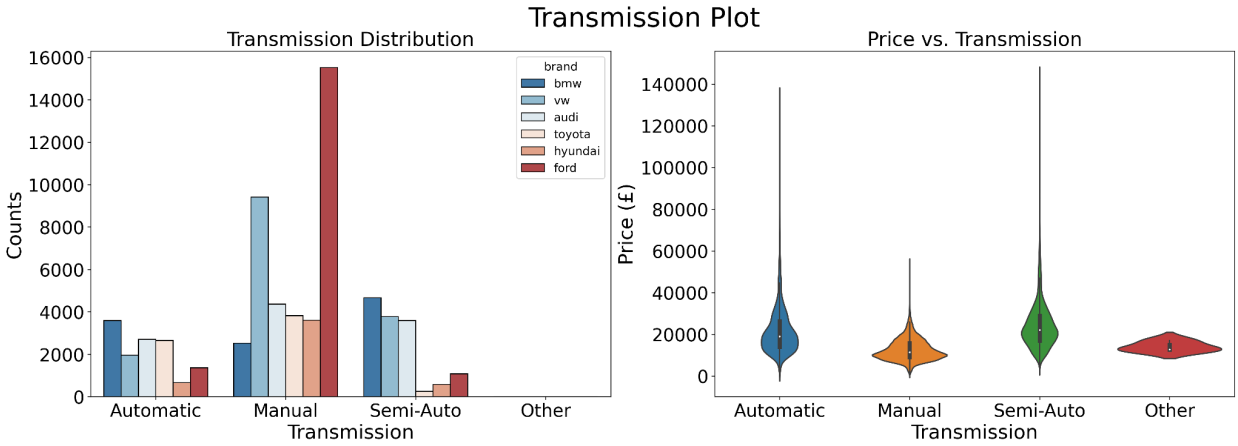

Figure 3. Most of cars in the UK are manual, and the
averge price of manual car is lower than others.

A correlation matrix is a useful tool for gain insight into the relationship between different 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.

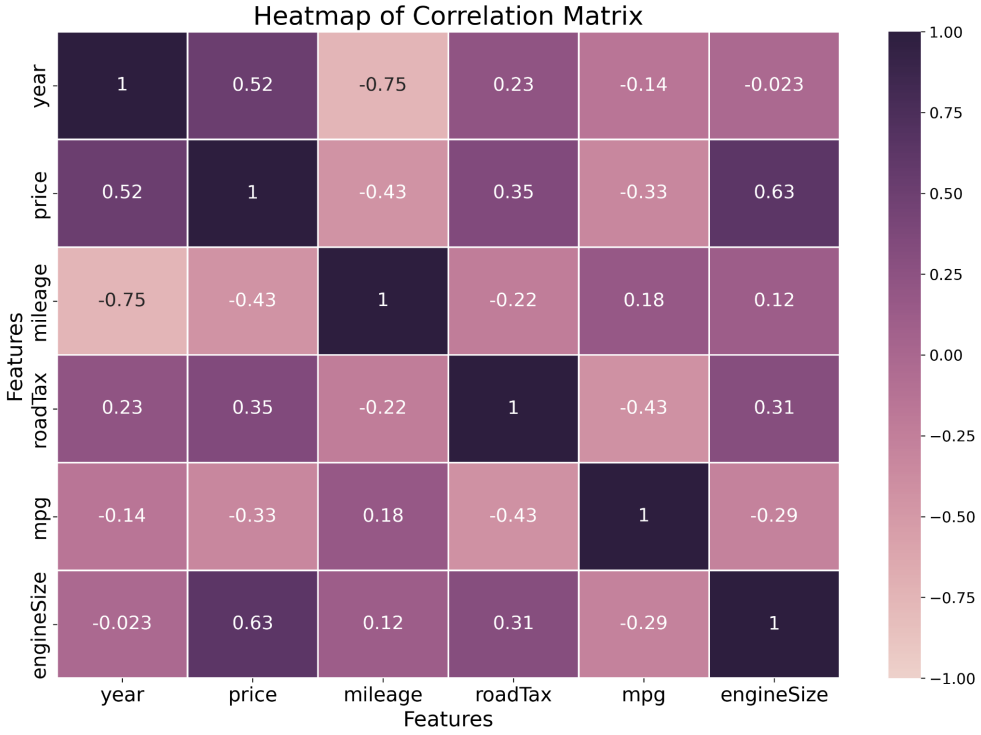

Figure 4. Year and engine size have a positive correlation with price. Mileage and MPG have a negative correlation with price.

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 to find the optimal parameters for the machine learning algorithm. based on the evaluation metric. The evaluation metric used for this project is RMSE (root mean squared error), as 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.). Since there is no group structure or time-series data in the dataset, I splitted the data into three sets: a training set, a validation set, and a test set. The splitting ratio used was 80%-10%-10%, with 80% of the data going into the training set, and the remaining 20% being split evenly between the validation and test sets. This ensured that there was enough data in the train set to train the model and guaranteed the number of data points in validation and test sets was sufficient. As a result, the training set contained 52,935 data points, the validation set contained 6,617 data points, and the test set had the same number of data points as the validation set.

During the preprocessing, OneHotEncoder was applied to all of the categorical features, as they could not be ranked or ordered. Additionally, I used MinMaxScaler on the engine size

feature as it has upper and lower boundaries and StandardScaler on the remaining 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 to this dataset are Lasso, Ridge, RandomForestRegressor, KNeighborsRegressor, and XGBRegressor. Lasso and Ridge are linear models, while the others are non-linear. Table 2 summarizes the details of the parameter tuning for each of these algorithms.

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, 3,10, 30, 100
		weights	"uniform", "distance"
5	XGBoost	max_depth	1, 3, 10, 30, 100

In a machine learning pipeline, there are two types of uncertainties that can arise. The first is splitting uncertainty, which is caused by the randomness of the train-test split. The second is non-deterministic model uncertainty, which is inherent to certain types of algorithms. 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 splitting uncertainties than non-linear ones, with Ridge having the lowest uncertainty of all. However, the mean RMSE scores of linear models were much higher than non-linear models.

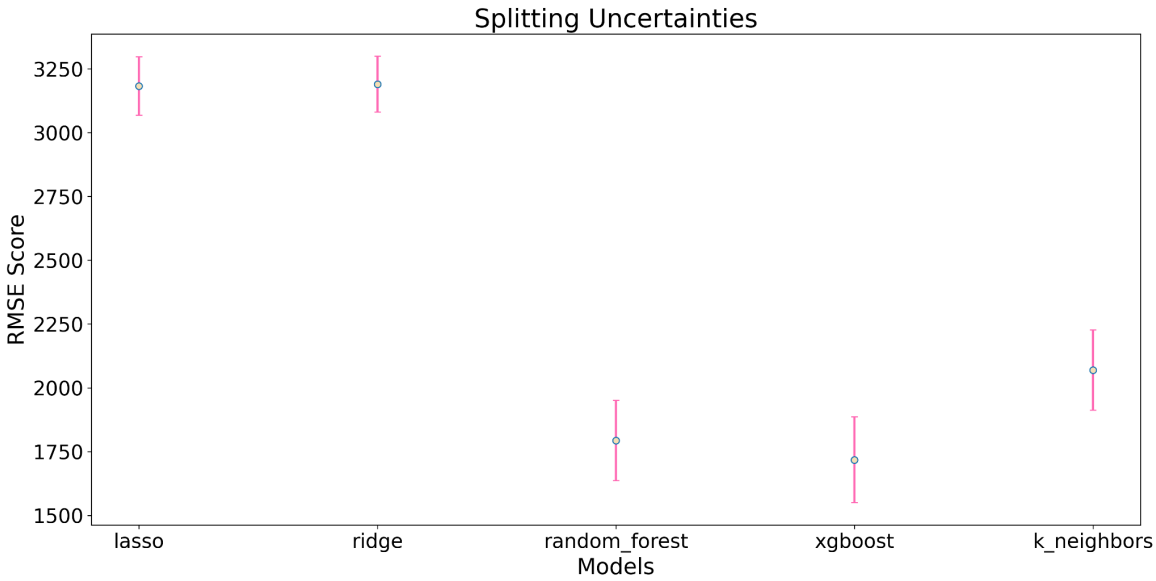

Figure 5. Ridge has the lowest uncertainty due to splitting among five models

Results

To calculate the baseline RMSE score, the average price value of the test set was set as 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, with a score that was 45.39 standard deviations above the baseline. The detailed performance of each of the models is 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	2069.67	157.19	
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.

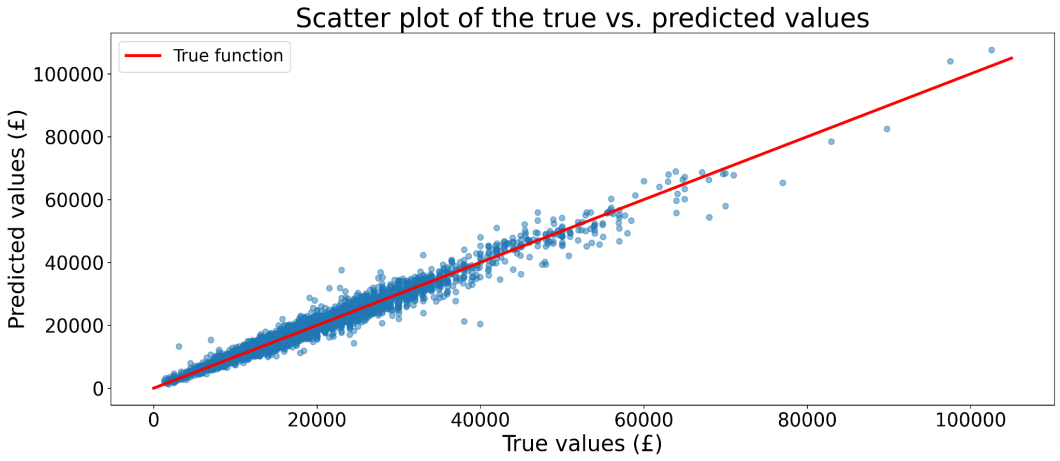

Figure 6. Most of predicted values are along the true function line.

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 the methods used different metrics, the top three most important features were the same in all cases: engine size, year, and MPG. These results also matched the findings of the correlation heatmap (Figure 4), which shows that features with a larger absolute value of correlation with the target variable were more important. The least important feature varied across the three methods, with onehot__model_Eos, onehot__model_Z4, and onehot__model_Escort being identified as the least important by each method, respectively. This analysis provides valuable insight into the factors that influence the performance of the models and allows us to better understand their behavior.

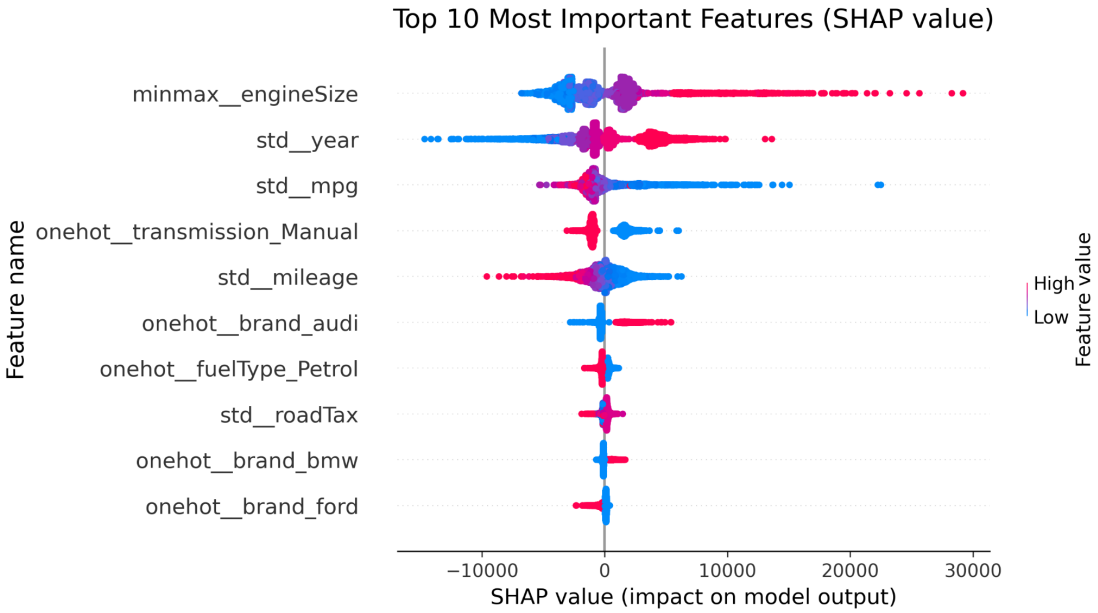

Figure 7. Engine size, year and MPG are the most important feature in SHAP method.

Regarding the local importance, data points with index 0 (Figure 8), 3000, and 6000 were analyzed. The major factors that affect the car price are consistent across these different data points, including engine size, year, mileage, and MPG. These features are also among the most important in the global feature importance analysis. This suggests that engine size, year, MPG, and mileage are the key factors that influence the price of a car, which aligns with our intuition.

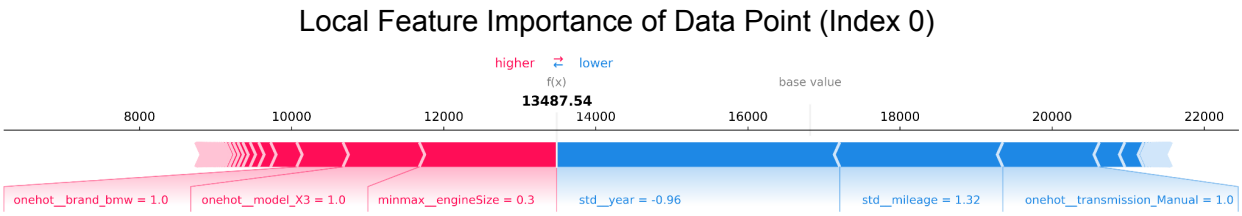

Figure 8. For data point with index 0, engine size is the major factor that increased the car price, and the manufacturing year is the factor that pulled the price down.

Outlook

There are several ways to improve the performance of the machine learning models used in this project. One way is to tune more hyperparameters of the models, as the default values may not be optimal for this dataset. Another approach is to perform feature engineering, which involves interpreting the relationships between existing features and creating new features based on those relationships. The third way of increasing accuracy is to collect additional features from Exchange & Mart, such as information on accidents, owners, and interior quality. Moreover, we could try using more machine learning models to see if there are any that perform better on 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