A red and blue text on a black background

Description automatically generated

**AIB 504**

**Machine Learning in Business**

**July 2023**

**Submission Date: 8 November 2023**

**Liu Kailiang (B2372684)**

# Executive Summary

The UK's used car market is an important segment of the automotive industry in the country. It is facing challenges in price regulation. This report introduces a machine learning approach to predict used car prices accurately and identify the determinants of their value utilising dataset from Auto Trader UK. Through adopting a combination of multivariable linear regression and random forest regression, the study reveals the significant impact of factors such as registration year and mileage on pricing, with variations observed across different car categories. Strategic recommendations for franchises and online platforms include focusing on younger, low-mileage vehicles to capitalize on higher resale values. While the models achieved strong prediction accuracy, future enhancements are suggested, particularly for the expanding data for electric vehicles.

# Introduction

In today's world, cars have become increasingly necessary, especially in developed countries such as the United Kingdom where car ownership rates are exceptionally high. According to the Department for Transport (2023), as of the end of March 2023, there were 40.8 million cars on the roads in the UK, indicating that about 60% of the population owns a car. The used car market has always played a significant role in the automobile industry. According to the Society of Motor Manufacturers and Traders (2023, February 10), in the UK, 6.8 million used cars were sold in the year 2022 alone. Despite the recent global semiconductor shortage causing a decline in new car production, thereby limiting the stock entering the used car market, this is still a significant figure. Another market research done by Mordor Intelligence revealed that the market size of used cars in the UK is projected to grow from USD 117.69 billion in 2021 to USD 226.16 billion by 2027. (Mordor Intelligence, n.d.)

Private sales had the highest volume in the UK used cars market in 2010, with total sales of 3.2 million units, equivalent to 47% of the market's overall volume. In comparison, sales of Franchised dealers had a volume of 1.9 million units in 2010, equating to 28% of the market total. (Datamonitor Plc, n.d.)

The used car market in the United Kingdom is fragmented, with similar market share among the players. This is a result of the rise in the number of pre-owned car retail outlets. Major players in this market include Arnold Clark, Cazoo, and the Aramis Group. (Mordor Intelligence, n.d.) Used cars are sold through various channels such as rental car companies, franchise and independent car dealers, auctions, and private sellers (Mordor Intelligence, n.d.). Unlike the new car market where customers purchase new cars from 4S stores and prices are determined by manufacturers, providing a level of transparency, the pricing of used cars is determined by the dealers. For instance, when an individual seller sells a car to a dealer, different dealers might offer different prices for the same car. Similarly, when a buyer is looking to purchase a used car, they might encounter varying prices for similar cars at different dealers. (Motorway Online Ltd., 2023)​ This situation leads to a lack of regulation standard in the pricing of used cars. As a result, addressing the problem of lack of regulation in used car prices is essential to avoid over-pricing and under-pricing and safeguarding the interests of individuals involved in the dealing processes. To address this issue, machine learning models can be developed to discover driven factors. This project aims to predict used car prices using supervised machine learning models, and through this process, identify the driving factors behind the prices, analyse how those factors influence used car prices. Another focus of the report is discovering how driven factors of prices shift when predicting different categories of cars.

# Methodology

Data utilised for this report was downloaded from Kaggle, where the original source of data was from Auto Trader UK, a popular automotive marketplace website in the UK. The dataset contains various information about the used cars sold which includes: "Title", "Price", "Mileage(miles)", "Registration(year)" "Previous Owners", "Fuel Type", "Body Type", "Engine", "Gearbox", "Seats", "Doors", "Emission Class", "Service history". These variables are informative and makes price prediction through machine learning feasible. Some of the variables contain missing values, data processing was performed prior to feature engineering to handle missing values.

A combination of machine learning models was used in the prediction of used car prices. This includes multivariable linear regression and random forest regression models. Models were created using functions from the scikit-learn packages in Python.

Linear regression is a popular model for predicting a target numeric variable. It estimates the relationship between the explanatory variable and the target variable. A single linear regression model consists of a single explanatory variable X, the function of a simple linear regression model can be expressed as: "Y=α + β\*X" where y is the target variable and x is the explanatory variable. α represents the intercept of the linear function and β is the coefficient of x which is the slope of x. It also can be interpreted as the effect of the explanatory variable on the target variable.

While simple linear regression estimates the effect of a single variable on the target variable, multivariable linear regression model is another type of linear regression model that incorporates multiple explanatory variables and estimates the association between each explanatory variable and the target variable, holding all other variables constant. The advantage of multivariable linear regression over simple linear regression is that through incorporating multiple independent variables, it accounts for confounding factors that may impact the effect of each explanatory variables on the target variable. (Boston University School of Public Health, n.d.)​ The function of a multivariable linear regression model can be written as "y=α + β1\*X1+β2\*X2+β3\*X3+…βi\*Xi where each β is the coefficient of each explanatory variable for i variables.

Multivariable linear regression was adopted because of its ability to quantify the impact of features on the target variable. With the coefficients, it is possible to observe the direction of the impact, for example, a negative coefficient of mileage indicates that prices are predicted to decrease as miles travelled by a car goes up.

Random forest is another popular supervised predictive machine learning model which is widely used because it is broadly applicable to many machine learning cases. It allows for features that have non-linear relationship with the target variable, and it can adapt to high-dimensional feature data (Borup, Christensen, Mühlbach, & Nielsen, 2023). In their presentation at the 2012 Strata Conference in New York, Howard and Bowle highlighted that Random Forest (RF) stands out as the most successful general-purpose algorithm of the modern era. (Howard & Bowles, 2012) Random Forest utilizes ensemble learning to enhance predictive performance by averaging the results of multiple decision trees. Each tree in the forest is applied to different subsets of the dataset, ensuring a more accurate prediction for complex problems. This method doesn't depend on one single decision tree; instead, it aggregates the predictions from all trees and bases the final output on the majority vote (Adetunji et al., 2022).

**Rationale**

This methodology used to predict used car prices incorporates the combination of multivariable linear regression and random forest regression. Rationale behind this strategy is that for random forest, it has an advantage in terms of prediction accuracy, and it allows for extraction of feature importance which measures the ability of features in purifying notes within each decision trees in the random forest. While feature importance indicates what variable are crucial for the prediction, however, it does not provide insights about how the features correlate with the target variable. Multivariable linear regression, although may not be as accurate as random forest in terms of its predictive power especially when non-linear relationship exists, is able to quantify the impact of the features on the target variable with the coefficients of the features. Utilising the advantages of both models, used car dealers are able to not only predict the prices of their cars, but also identify what factors are crucial in the prediction and how they impact prices. As the aim of the report is not only to accurately predict used car prices, but also identify and analyse how driven factors impact prices, this methodology aligns well with it.

Apart from applying random forest regression and multivariable linear regression to the complete dataset to build a general price prediction model, subsets of the data by various categories were created and models were also built to predict the prices. The purpose of this process is to observe how feature importance and coefficients vary across different types of cars and gain insights on driven factors of prices for specific types of cars. This is important because different types of cars may have different factors that influence their price. By creating separate models, we are able to capture these unique relationships more accurately.

**Regression Model Performance Evaluation Metrics**

Two model performance evaluation metrics for regression models were adopted to measure the accuracy of the models. The first metric is Rooted Mean Squared Error (RMSE) and the second metric is R Squared(R2).

RMSE is a metric that functions by square rooting the Mean Square Error (MSE), another metrics which is the squared difference between the true value and predicted value of the target variable. Advantage of using RMSE is that it is easier to interpret compared to MSE because it has the same scale as the target variable. A disadvantage of RMSE is that it is sensitive to outliers from the dataset.

R2 is another performance evaluation matric for regression models. It is expressed as "1-squared sum error of regression line / squared sum error of mean line. The value of R2 score is between 0 and 1, and it measures how good the regression fits the data, also known as "goodness of fit". However, it is only suitable for linear relationships. In the case of predicting used car prices, assumption made is that relationships between attributes of cars and their prices are mostly linear.

# Results and Discussion

**Data Description**

A screenshot of a computer

Description automatically generated

Fig 1. Data Info

The dataset contains 3685 records of used cars sold at Auto Trader UK with 12 variables of different data types. Variables of the dataset consist of:

* *title:* categorical variable indicating the model of the car
* *Price:* price of the car in pounds
* *Mileage(miles)*: numeric variable indicating number of miles travelled
* *Registration\_Year:* numeric variable indicating year of registration
* *Previous Owners:* numeric variable indicating number of previous owners
* *Fuel type:* categorical variable indicating type of fuel or power source if the car
* *Body type:*  categorical variable indicating the class of the car (e.g., SUV, Hatchback etc.)
* *Gearbox:* categorical variable indicating type of gearbox (manual/auto)
* *Doors:* numeric variable indicating number of doors of the car
* *Seats:* numeric variable indicating number of seats in the car
* *Emission Class:* categorical variable indicating emission class on the car according to the European Emission Standards
* *Service history:* categorical variable indicating service history of the car

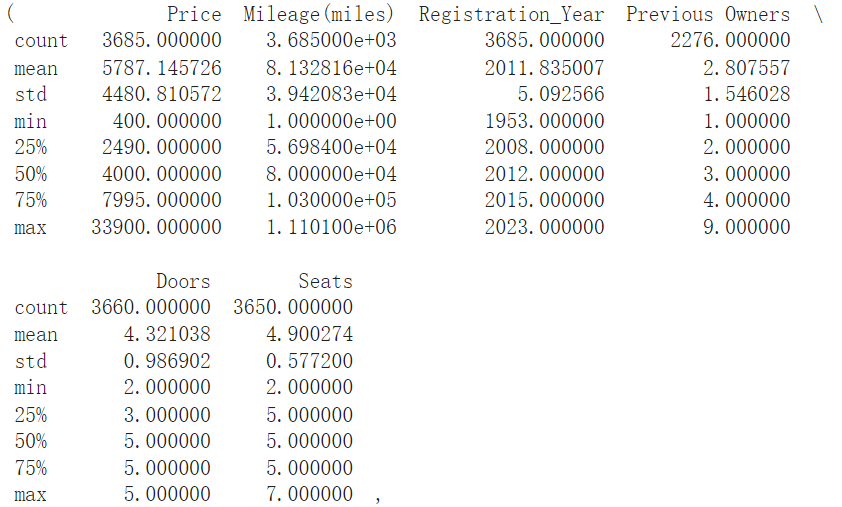


Fig 2. Summary Statistics

Summary statistics of numeric variables shows that there are outliers in the dataset. It can be inferred from the huge gap between the max value and the 75th percentile of price and mileage, and also between the minimum value and 25th percentile of registration year.

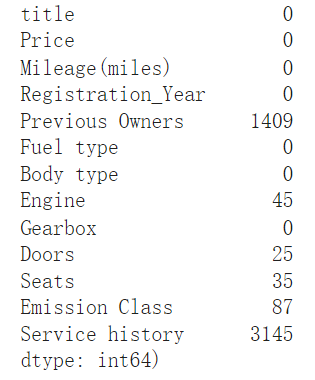
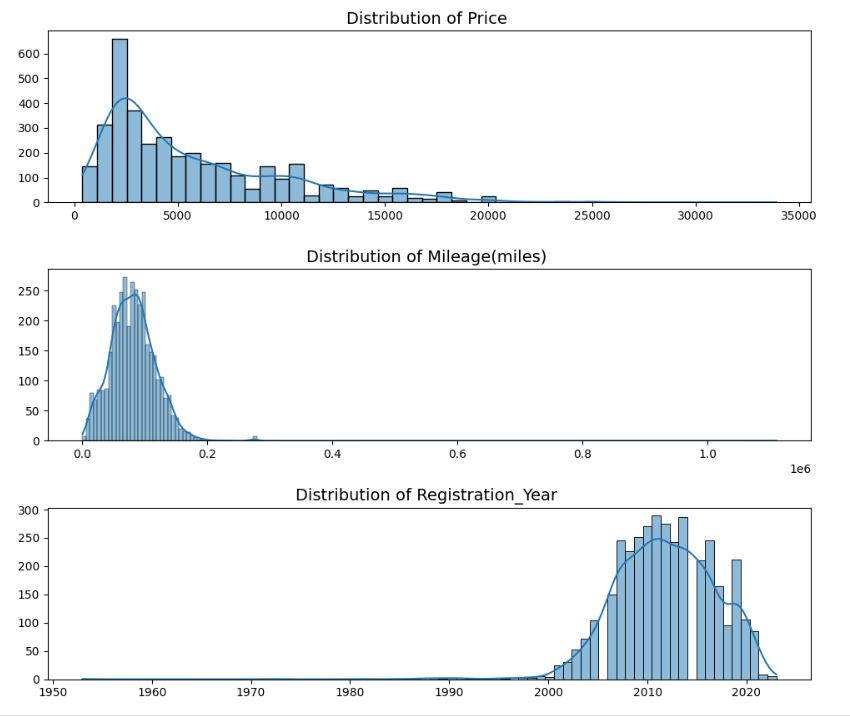


Fig 3. Missing Values

Missing values are found in Previous Owners, Engine, Doors, Seats, Emission Class and Service history. Service history containing too many missing values is dropped from the dataset. A possible method for dealing with the remaining of the missing values is imputation, however, for the purpose of visualising the actual data, this will only be done in the data preparation for machining learning stage.

**Distribution of Numeric Variables**



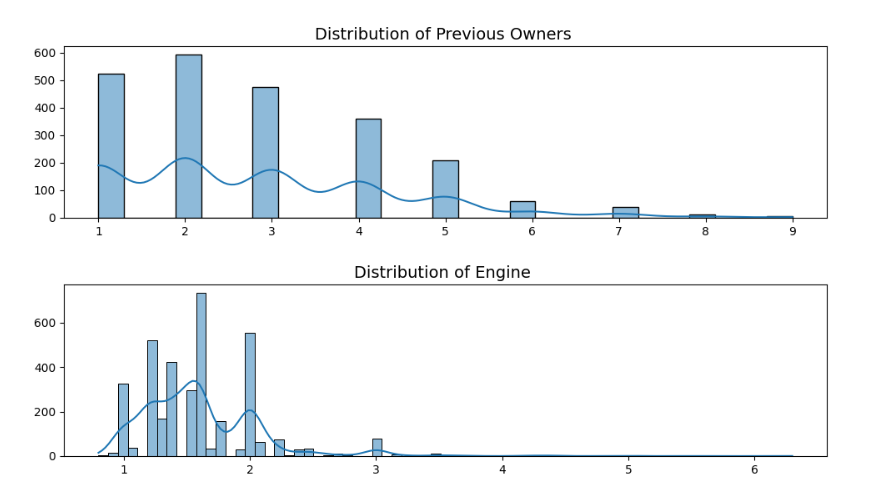
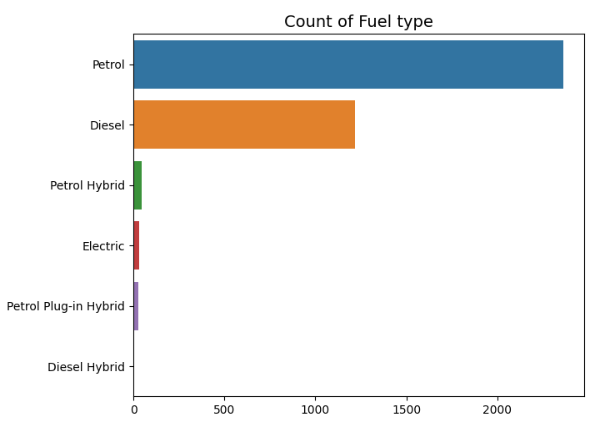
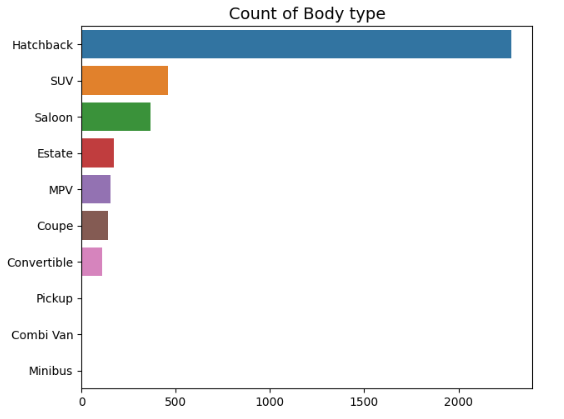
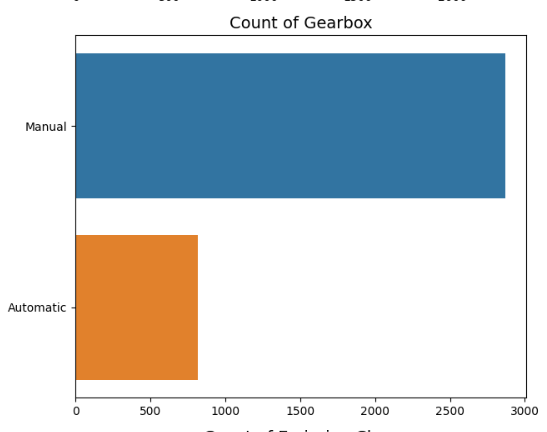
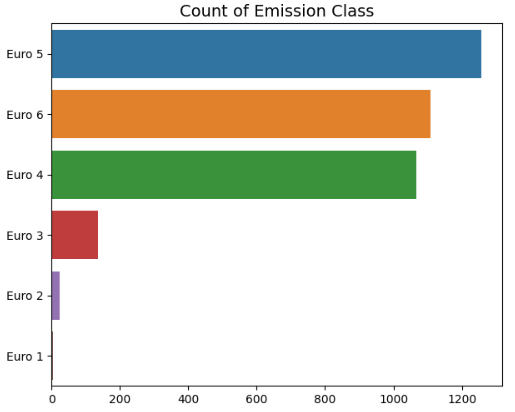


Fig 4. Noticeable Distribution Patterns Found in Numeric Variable

The scale of the x-axis of price, mileage and registration year proves that there are extreme outliers in the dataset. Similar to how missing values will be handled, for the purpose of visualising the actual full data, extreme outliers will be handled in a later stage.

Ignoring the distortion of the plots caused by extreme outliers, the distribution of price and previous owners is positively skewed with its tail on the right. Distributions of mileage and registration year fits a normal distribution pattern well, and the distribution of engine generally fits a normal distribution pattern as well. Notice that "Engine" has been transformed to a numeric variable by removing the letter 'L' from all of its records, this transformation will eliminate the need for creating dummy variables of engine when preparing the data for training the machine learning models in a later stage, reducing the levels of dimension.

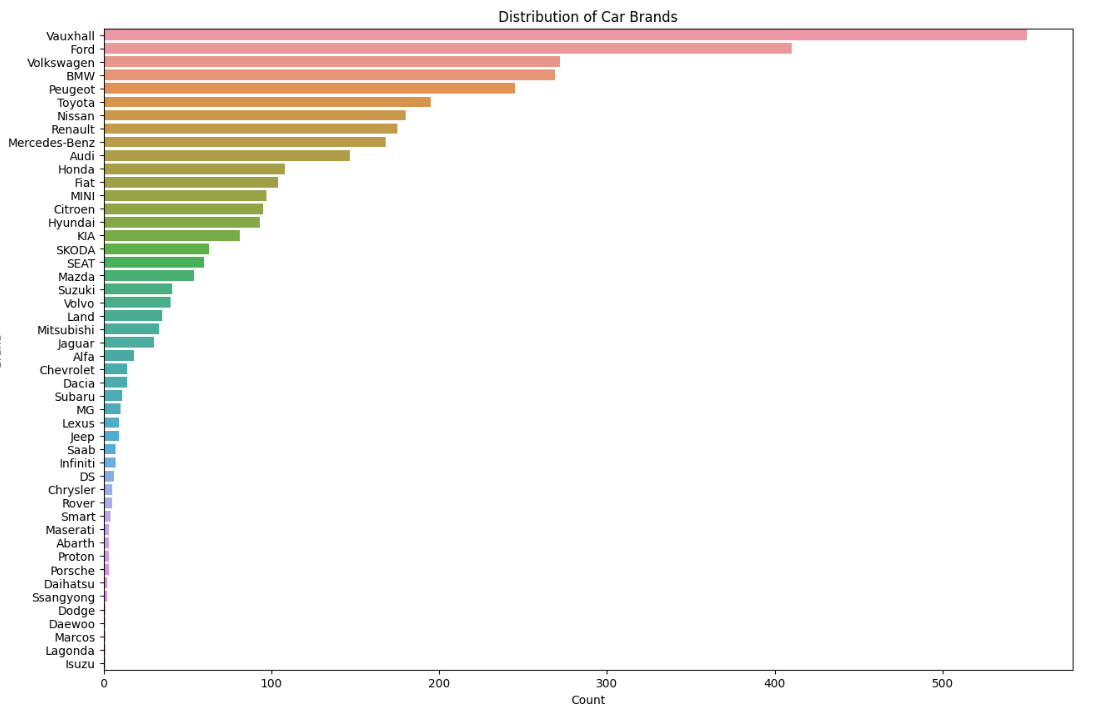


Fig 5. Distribution of Categorical Variables

The "Brand" variable is generated by extracting the first word in the 'title' variable. It is preferred over the 'title' variable because there are too many models of cars, replacing it with brands will reduce the levels of dimension when training machine learning models, and also make the distribution of brands readable.

The distributions of categorical variables show that:

* Majority of cars sold are non-EV cars
* Majority of cars are hatchbacks
* Majority of cars have manual gearbox
* Majority of cars belongs to emission classes Euro 4 to Euro 6
* Model of cars belongs to many brands and the distribution is imbalanced

Scatter Plots of Numeric Variables with Respect to Price

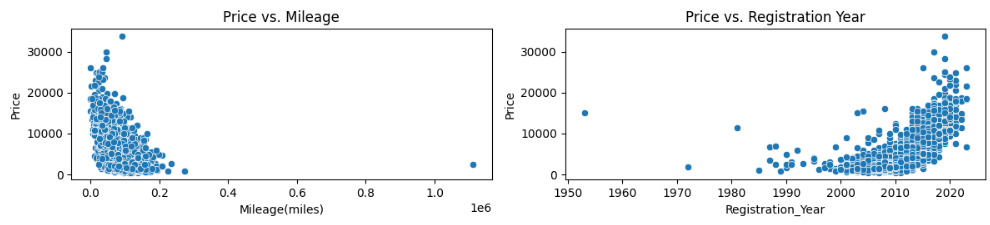
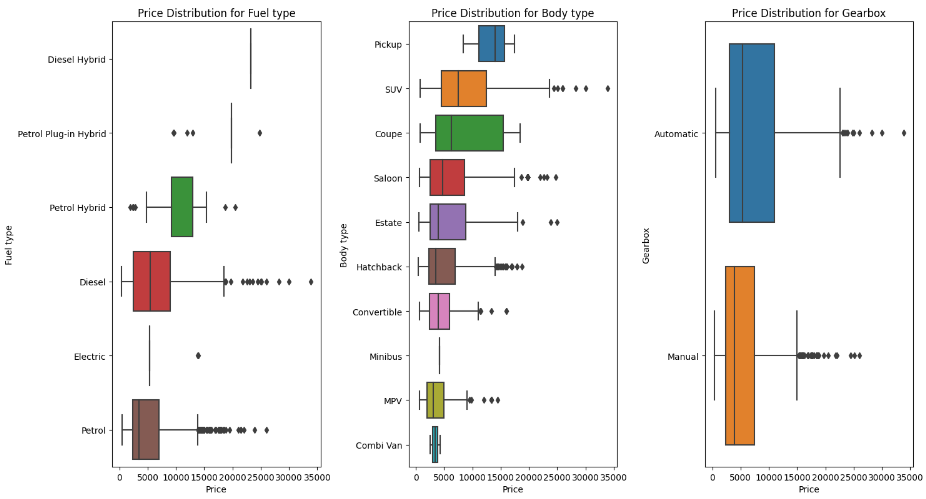




Fig 6. Noticeable Patterns Found in Scatter Plots

Despite the extreme outliers causing a distortion of the scale of price and milage, noticeable patterns are found in these scatter plots. It is observed from the plots that there is an obvious negative association between price and mileage and price. As mileage increases, price decrease. Moreover, there is positive association between registration year and price. As registration year gets close to present, price of cars goes up. It is also observed that price is negatively associated with number of previous owners, price of cars decreases as number of previous owners increases.

**Box Plots of Numeric Variables with Respect to Price**



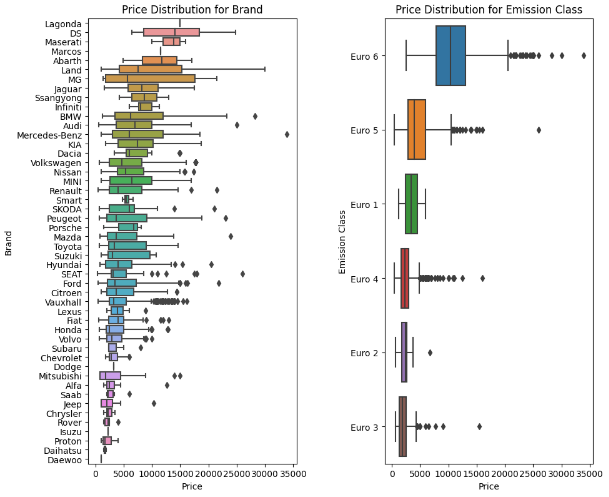


Fig 7. Boxplots of Categorical Variables with Respect to Price

The Price distributions of categorical variables show that:

* Hybrid cars are sold at higher prices compared to other fuel times
* Price varies more for coupe type of cars compared to other body types
* Cars with automatic gearboxes are mor expensive compared to manual gearbox cars
* Cars belonging to the most recently updated Euro 6 emission class are sold higher than other classes.

**Correlation heatmap**

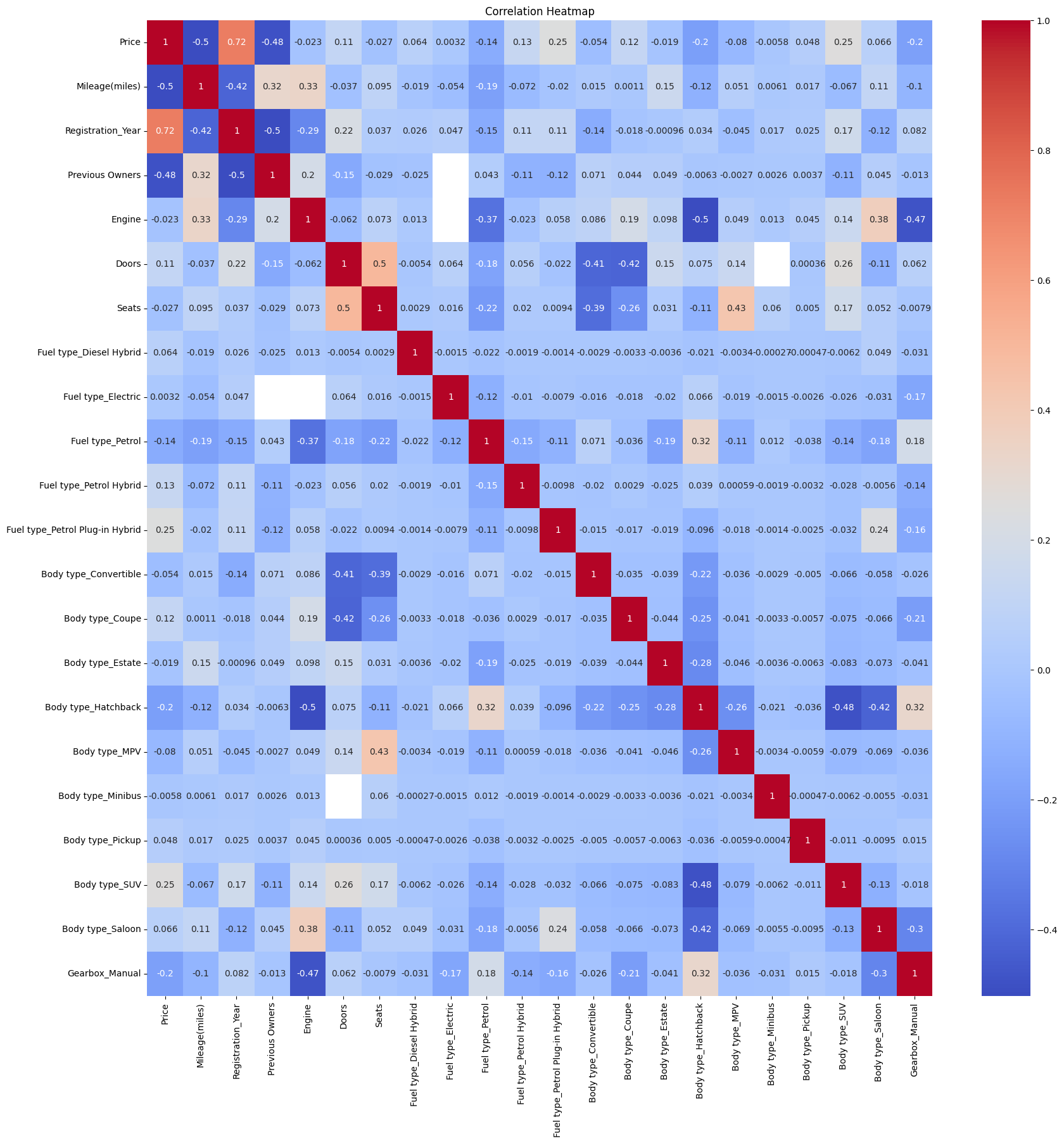


Fig 8. Correlation Heatmap

The correlation heatmap supports the previous findings from the visualisations and indicates that the correlation between price and mileage, and between price and previous owners are strongly negative, with the correlation between price and registration year is strongly positive. This heatmap also reveals new insights on the correlations among number of previous owners, mileage and registration year. It shows that mileage and previous owners positively correlates with each other, and both of them negatively correlate with registration year. This is reasonable because used cars that are relatively new usually have lesser number of owner and travel lesser distances.

## Predictive Modelling results

**Data pre-processing and Feature Engineering**

Before building the regression models to predict used car prices, data needs to be further processed as there are missing values and extreme outliers in the dataset. The dataset also requires feature engineering in order to train the models.

Missing values can be handled in different ways, the simplest way is removing them or dropping columns that contain a lot of missing values. However, these methods are not suitable for this dataset because the "Previous Owners" containing a substantial number of columns has been identified as an important variable. simply dropping it is not optimal and may lead to inaccurate prediction. Moreover, removing its null values will lead to a significant reduction of data, affecting the completeness of data and may lead to overfitting.

As a result, imputation which is a more refined approach for handle missing values is adopted by filling in the missing values with the median. Median is more robust for variable with skewed distributions. The same approach is applied to 'Engine', 'Doors' and 'Seats' which also contain a few missing values. For missing values in 'Emission Class', mode is used as it is a categorical variable. Other than missing values, there are also extreme outliers in the dataset, especially those have extremely high mileage. To handle this, outliers with a mileage of greater than the 99th percentile are removed.

To prepare the data for model training, feature engineering approaches are applied to the processed data. Categorical features are transformed to dummy variables with one-hot encoding, and the MinMax Scaler is applied to standardize the scales of the numeric features. After feature engineering the dataset is split into training and testing data and is ready for training the models.

**Model Evaluation for the general model**

A random forest regression model using the complete dataset is first built. This model functions as the general model for predict the prices and finding out the important features associating with price from the feature importance rank. In addition to random forest regression model, a multivariable linear regression model is also built, but the purpose of this model is only extracting the coefficients to measure the impact of the important features.

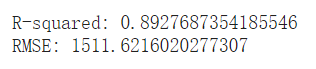


Fig 9. R2 and RMSE of RF (for Models Built on Complete Data)

The Random Forest Regression model has a R2 score of 0.89 and a RMSE of 1551,62，Indicating a acceptably accurate prediction of the target variable, price.

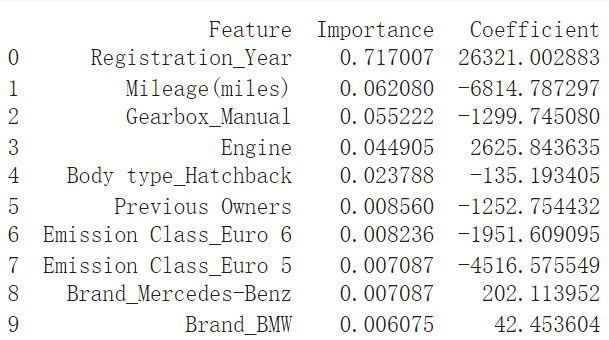


Fig 10. Feature Importance Rank and Coefficients (for Models Built on Complete Data)

Feature importance is dominated by registration year with an unexpectedly high importance value of 0.71. This indicates that registration year has a strong ability in purifying notes of the decision trees in the random forest model. It is indeed the most curtail feature for predicting the price. Other than registration year, mileage, manual gearbox type, and engine are also some important features. Although they are not as significant as registration year, they must not be neglected. Despite the fact that feature importance does not necessarily equal to the level of association between the features and the target variable, the coefficients extracted from the multivariable linear regression estimate the impact of the features on price and add a layer of understanding on the features. Registration year having a large positive coefficient implies that an increase in registration year may lead to a relatively significant rise in the price of the car.

**Model evaluation for EV and Non-EV cars**

Importance of features may vairy for different types of cars, and it is beneficial to discover how they shifts in order to better understand how each feature impact price for different categories of cars. As that reason, random forest and multivariable linear regression models are built for predicting price of EV cars and non-EV cars.

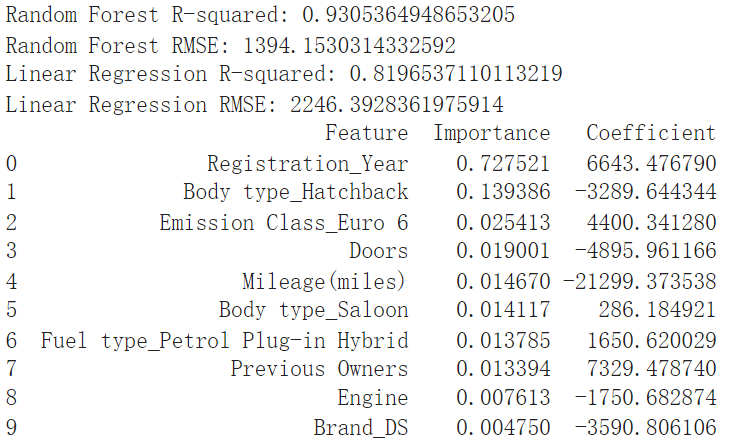


Fig 11. Feature Importance Rank and Coefficients (for EVs)

Similar to the previous model, registration year is still the dominant feature on the feature importance rank, however, there is a huge difference lies in its coefficient which is much smaller. This indicates that effect of registration year on price is much smaller for EV cars. The second important feature here being Hatchback body type is hard to explain in terms of its association with price, it only identifies itself as an important feature in the prediction process. However most importantly, the change in the confident of registration year is captured which highlight its specific impact on price of EV cars.

Data containing non-EV cars goes through the same process, the following figure shows the result.

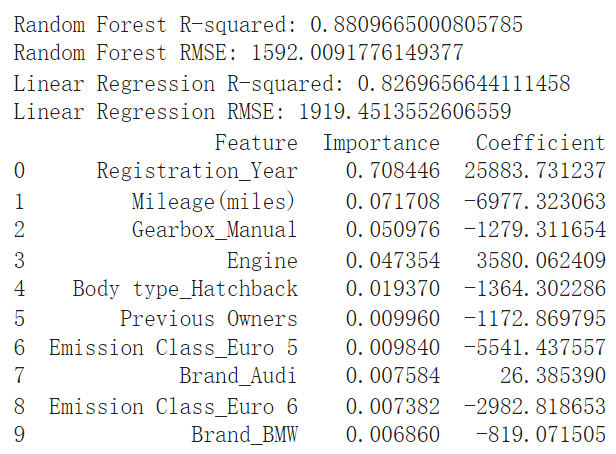


Fig 12. Feature Importance Rank and Coefficients (for non-EVs)

The feature importance rank and coefficients of the features are very similar to the result from the first random forest and linear regression model, possibly due to similar size of data record. However, this result can be interpreted as a reinforcement to the findings from the modelling results for EV cars.

**Model Evaluation for Luxury and Non-luxury Brands**

Many brands have been identified in the data, these brands can potentially be categorised into luxury brands and non-luxury brands. There are generally many differences between luxury brand and non-luxury brand in terms of varier aspects of the cars, differentiating them and applying the machine learning models separately can potentially discover new insight. As a result, dataset is subset for luxury brands and non-luxury brands and machine learning models are applied to both categories.

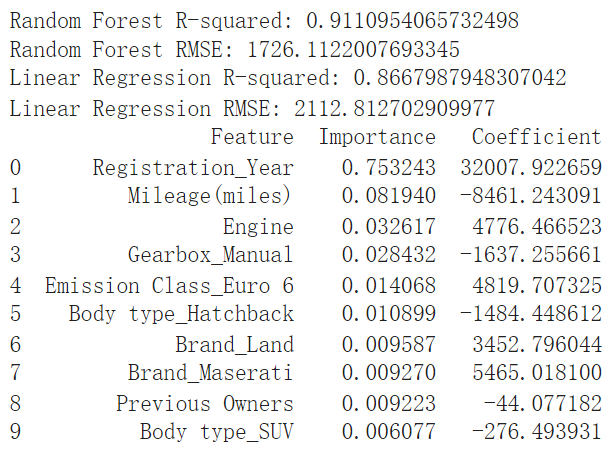


Fig 13. Feature Importance Rank and Coefficients (for luxury brands)

The feature importance rank of the random forest model for luxury brands is similar to the first general model in terms of order of the top few important features. The dominant feature, registration year is higher by less than 0.04. Major differences exist in the coefficients, where numbers are larger overall, compared to the first model. This is possibly due to the nature of luxury brands that they are relatively expensive.

The machine learning models are applied to non-luxury brands, result is shown in the following figure:

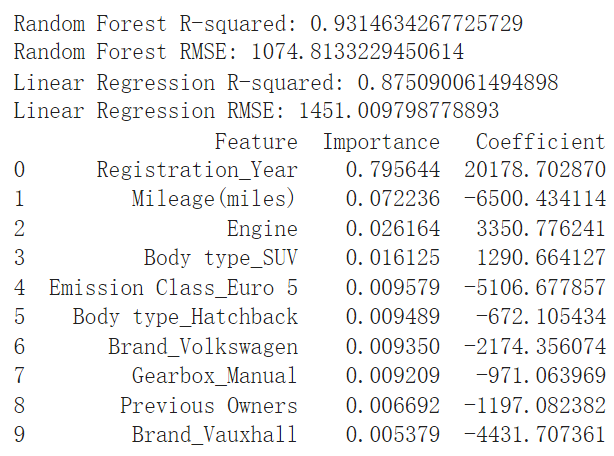


Fig 14. Feature Importance Rank and Coefficients (for non-luxury brands)

The importance of registration year for non-luxury brands is noticeably different from the first general model as it is higher by 0.07. This could mean that for non-luxury brands, registration year has even more predictive power when predicting price

# Conclusion and Recommendations

The aim of the project to build an accurate machine learning model for predicting used car prices in the UK has been successfully achieved through developing a random forest regression model that has an acceptably accurate R squared and RSME. The project has also identified crucial factors influencing the prediction of price by leveraging on the feature importance rank. By strategically utilising the strength of multivariable linear regression, the project has quantified the effects of the driven factors, allowing for a deeper understanding of how the driven factors impact used car prices. Additional models were also developed to observe the changes in the driven factors, and at the same time allowing for application capabilities in more specific scenarios. The models confirmed that the registration year and mileage are the most important variable significant predictors of a car's price. This finding is supported by Sung Jin Cho in his research on the determinant of used rental car prices where he concluded that distance travelled and age of cars are determinant of used car's resale value.

**Business Opportunities**

The models developed in the project can be adopted and applied by used car dealers to set prices that truly reflect a used car's market value. This could poetically increase turnover and customer satisfaction rate. The models developed could potentially be deployed as a online service for consumers planning to buy used cars. By providing accurate estimations of the cars, buying a over priced car from dealers can be avoided and consumers can negotiate better deals and make purchases with confidence.

**Recommendations**

Based on the findings from the model, franchises or online dealing platforms are recommended to strategically source for used cars of relatively young age, or cars that have not travelled long total distances. By focusing on acquiring and selling vehicles that are newer and with lower mileage, these businesses can benefit from selling those cars at higher prices Moreover, this approach could also satisfy consumer demand for reliable, newer used cars that are expected to have fewer maintenance issues and longer service life. when implementing this strategy, it is crucial for these dealers to balance their inventory to avoid over stocking which could potentially lead to value shrinking of unsold cars.

**Strengths and Limitations**

The random forest regression models developed in this project have the advantage of high prediction accuracy, while the multivariable linear regression models have the capability of quantifying the impact of the variables on price. limitation recognised from the models are the sensitivity to outliers and the reliance on linear relationships. Further exploration on EVs can be done by expanding the dataset with more EVs and include more features relevant to EVs. For example, types of battery, distance per charge, levels of autonomy. With more balanced dataset and more informative features, the models will be able to account for rapidly emerging trend of the shift towards electric vehicles.

# References

Department for Transport. (2023). Vehicle licensing statistics: January to March 2023. GOV.UK. Retrieved from <https://www.gov.uk/government/statistics/vehicle-licensing-statistics-january-to-march-2023/vehicle-licensing-statistics-january-to-march-2023>

Society of Motor Manufacturers and Traders. (2023, February 10). Used car market down but EVs buck trend. Retrieved from <https://www.smmt.co.uk/2023/02/used-car-market-down-but-evs-buck-trend/>

Mordor Intelligence. (n.d.). Used Car Market in UK - Trends, Share & Report. Retrieved [02.11.2023], from <https://www.mordorintelligence.com/industry-reports/unitedkingdom-used-car-market>

Used Cars Industry Profile: United Kingdom. (n.d.). Datamonitor Plc.

Motorway Online Ltd. (2023). Used car prices – the ultimate guide. Motorway. Retrieved from <https://motorway.co.uk/sell-my-car/guides/used-car-prices>

Boston University School of Public Health. (n.d.). The Multiple Linear Regression Equation. Retrieved from <https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704-ep713_multivariablemethods/BS704-EP713_MultivariableMethods2.html>

Borup, D., Christensen, B. J., Mühlbach, N. S., & Nielsen, M. S. (2023). Targeting predictors in random forest regression. International Journal of Forecasting, 39(2), 841-868. <https://doi.org/10.1016/j.ijforecast.2022.02.010>

The Two Most Important Algorithms in Predictive Modelling Today. Strata Conference Presentation, New York, February 28 2012.)

Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House Price Prediction using Random Forest Machine Learning Technique. Procedia Computer Science, 199, 806-813. <https://doi.org/10.1016/j.procs.2022.01.100>