Predicting Used Car Prices in the UK

Introduction

The used car market in the United Kingdom is a significant segment of the automotive industry, providing a wide range of options for consumers seeking affordability and value. In this project, I created a predictive model to estimate the prices of used cars in the United Kingdom. Using a dataset containing information such as model, year, mileage, fuel type, engine size, gearbox, etc I built models to predict the price of a car based on its characteristics.

A predictive model for used car prices can be extremely useful for both consumers and businesses, but pricing remains complex due to the many variables involved. Buyers can use such a tool to better understand what a fair price for a car should be, while sellers and dealerships can set competitive and realistic prices based on data rather than guesswork. Pricing vehicles appropriately can help cars sell faster and increase customer trust in the buying process.

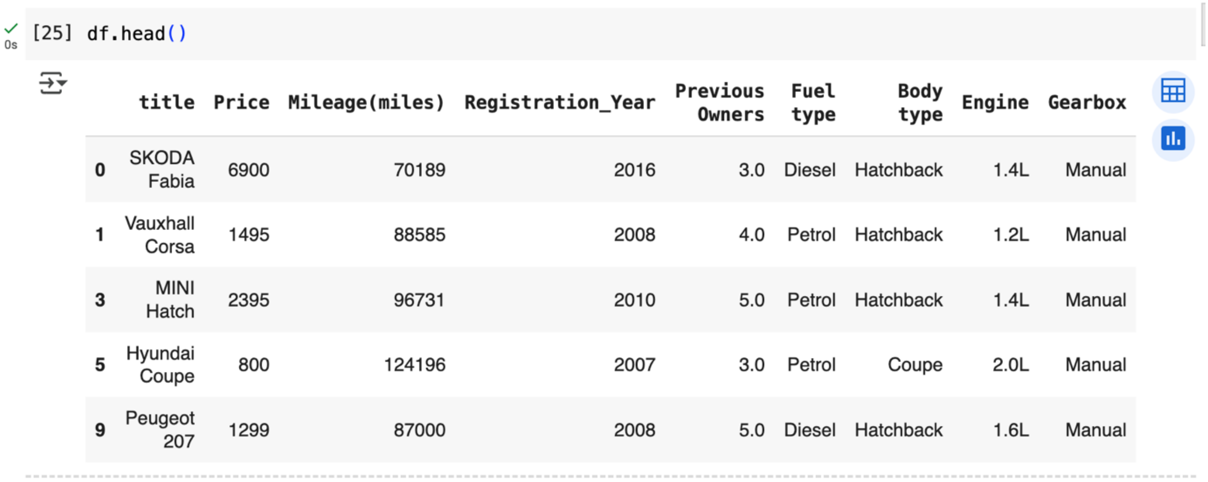
Using a dataset of over 2,200 used car listings from the UK, I performed exploratory data analysis to clean and understand the data before building predictive models to estimate the price. I tested Linear Regression, K-Nearest Neighbors, Decision Tree Regression, Random Forest Regression, and a Neural Network and evaluated the models using metrics like the R^2 score and Mean Squared Error (MSE). Ultimately, the Neural Network model performed the best, achieving the highest R^2 score and the lowest MSE among all models. Important factors influencing the price prediction included the year of registration, mileage, and engine size.

Data Description

The dataset used in this project comes from a Kaggle repository titled "Used Cars Prices in the UK ″. It contains information about thousands of used car listings posted online across the UK. This original dataset consists of 3,685 rows and 14 columns and Each row represents a used car listing from the United Kingdom.

Before modelling, significant data cleaning was required to improve the quality and usability of the dataset. First, several columns were removed. The 'Unnamed: 0' column, an index artifact from CSV export, was dropped as it contained no useful information. Additionally, the columns 'Doors', 'Seats', and 'Emission Class' were dropped because they were not directly relevant to predicting the selling price for this analysis. The 'Service history' column was also removed due to a high percentage of missing values, which made it unsuitable for reliable modelling.

After dropping unnecessary columns, the dataset still contained some missing values in other columns. To ensure that the models would not be negatively impacted by incomplete data, rows containing any null values were removed.



```
[25] df.head()
```

	title	Price	Mileage(miles)	Registration_Year	Previous Owners	Fuel type	Body type	Engine	Gearbox
0	SKODA Fabia	6900	70189	2016	3.0	Diesel	Hatchback	1.4L	Manual
1	Vauxhall Corsa	1495	88585	2008	4.0	Petrol	Hatchback	1.2L	Manual
3	MINI Hatch	2395	96731	2010	5.0	Petrol	Hatchback	1.4L	Manual
5	Hyundai Coupe	800	124196	2007	3.0	Petrol	Coupe	2.0L	Manual
9	Peugeot 207	1299	87000	2008	5.0	Diesel	Hatchback	1.6L	Manual

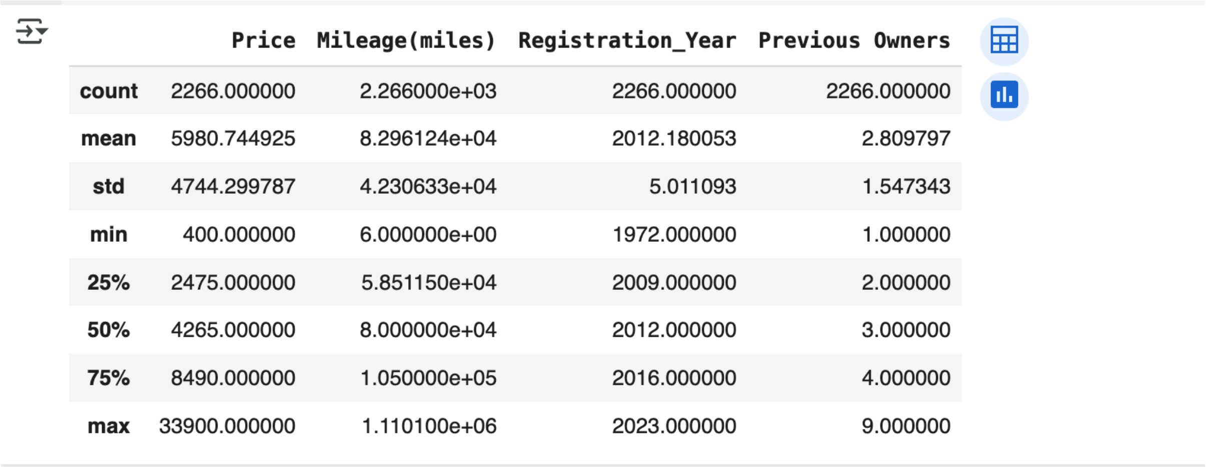
Figure 1: Cleaned Dataset

Following the cleaning process, the final dataset contained 2,266 rows and 9 columns with both numeric and categorical features. The cleaned dataset had no missing values, allowing the modeling process to proceed without the need for additional imputation or data balancing.

Exploratory Data Analysis

After cleaning the dataset, an exploratory data analysis (EDA) was conducted to better understand the distribution of the data, identify any underlying patterns, and detect potential relationships between features and the target variable, price.

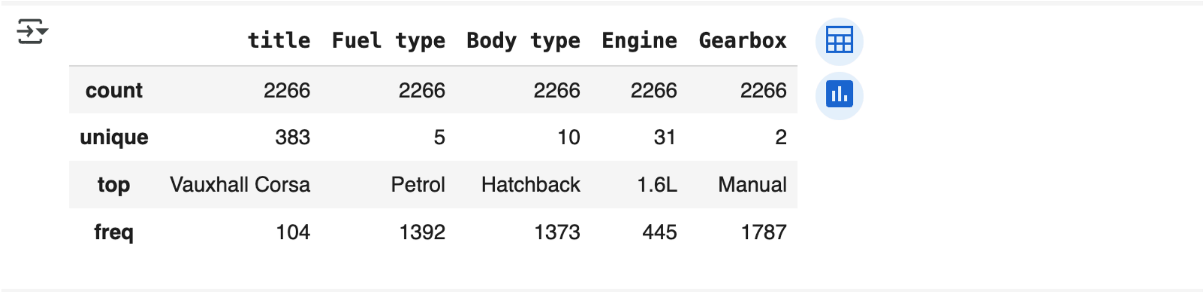
First, I reviewed the basic descriptive statistics for both the numerical and categorical features to better understand the overall structure and characteristics of the data.



	Price	Mileage(miles)	Registration_Year	Previous Owners
count	2266.000000	2.266000e+03	2266.000000	2266.000000
mean	5980.744925	8.296124e+04	2012.180053	2.809797
std	4744.299787	4.230633e+04	5.011093	1.547343
min	400.000000	6.000000e+00	1972.000000	1.000000
25%	2475.000000	5.851150e+04	2009.000000	2.000000
50%	4265.000000	8.000000e+04	2012.000000	3.000000
75%	8490.000000	1.050000e+05	2016.000000	4.000000
max	33900.000000	1.110100e+06	2023.000000	9.000000

Figure 2: Descriptive statistics for numerical columns

The average price of a used car in the dataset is approximately £5,981, but prices vary widely, with some vehicles priced as low as £400 and others reaching up to £33,900. Mileage ranges from nearly new cars (6 miles) to heavily used vehicles (over 1.1 million miles), although the majority fall between 58,000 and 105,000 miles. Most cars were registered around 2012, but the dataset includes cars as old as 1972 and as new as 2023.



	title	Fuel type	Body type	Engine	Gearbox
count	2266	2266	2266	2266	2266
unique	383	5	10	31	2
top	Vauxhall Corsa	Petrol	Hatchback	1.6L	Manual
freq	104	1392	1373	445	1787

Figure 3: Descriptive statistics for categorical columns

Petrol and Manual transmission cars dominate the dataset, accounting for the vast majority of listings. Additionally, The most frequent body type is Hatchback, aligning with the popularity of compact, fuel-efficient cars in the UK market.

I then continued the EDA by providing some visualisations to detect outliers and skewness and see how some features affect price.

Price Distribution

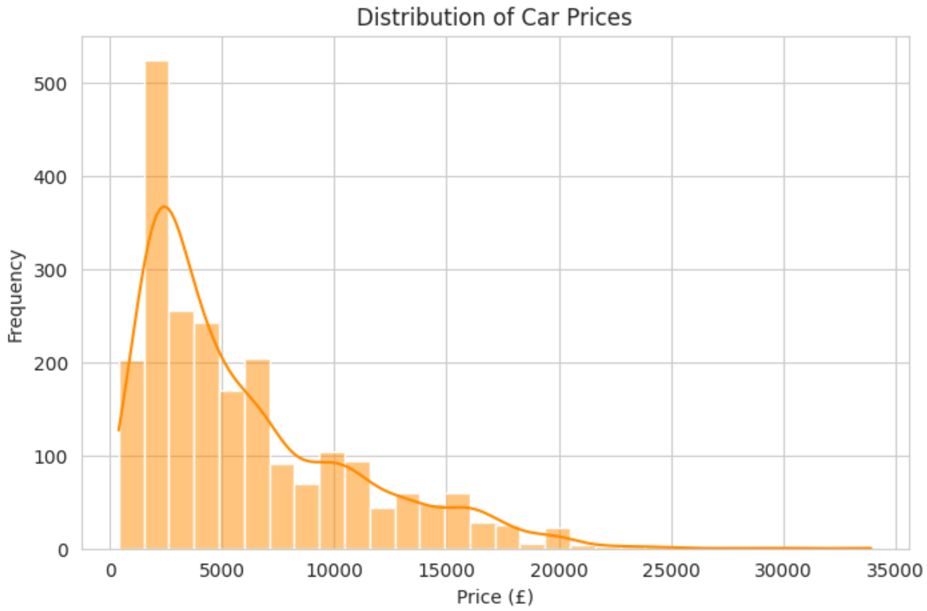


Figure 4: Distribution of Car Prices

The distribution of car prices is right-skewed, with the majority of vehicles priced below £20,000. A large cluster of cars are priced between £2,000 and £10,000, reflecting the typical price range for used cars in the UK. There are a few high-priced listings (above £25,000) which create a long tail on the right side of the distribution. These likely correspond to luxury or newer vehicles. The price data is not normally distributed, and the skewness suggests that models predicting price need to handle outliers carefully.

Relationship Between Price and Mileage

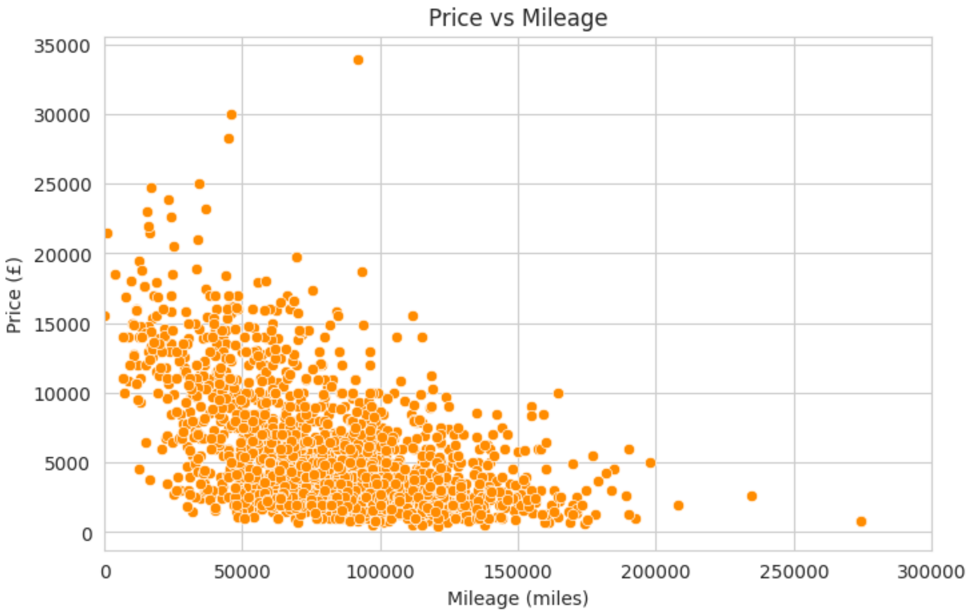


Figure 5: scatterplot of Price vs Mileage

The scatterplot of Price vs Mileage shows a clear negative relationship between the two variables. In general, as mileage increases, the price of the car decreases. Although the general trend is downward, there is noticeable scatter, especially among vehicles with low mileage, where prices vary widely depending on other factors. Mileage is expected to be an important predictor in the model and Nonlinear models may capture the scattered pattern better than simple linear models.

Relationship Between Price and Registration Year

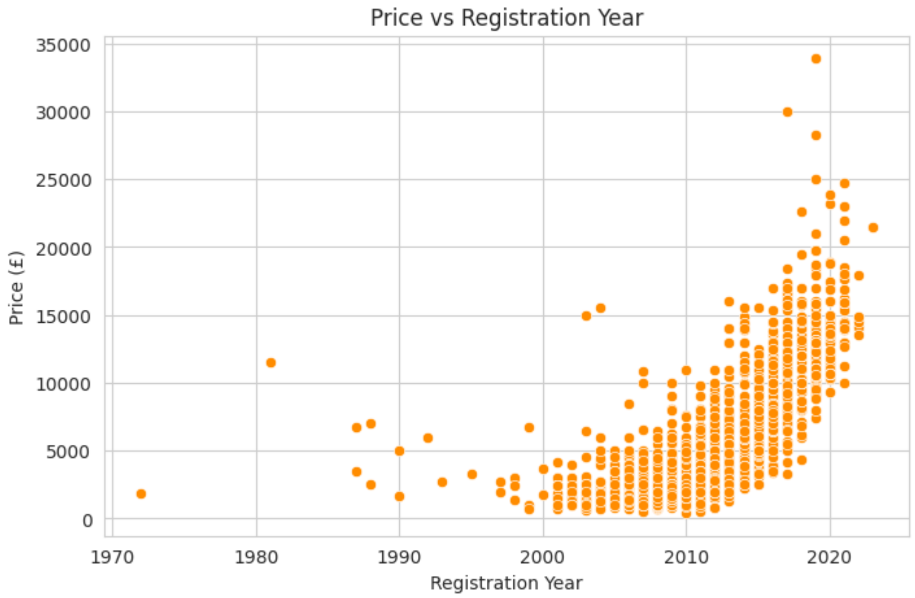


Figure 6: Scatterplot of Price vs Registration Year

The scatterplot of Price vs Registration Year shows a clear positive relationship between the two variables. As the registration year increases (i.e., the car is newer), the price tends to be higher. Vehicles registered after 2010 show a noticeable increase in price, with a steep rise for cars registered after 2015. Registration Year should be treated as a key feature during modeling and the non-linear relationship suggests that tree-based models or neural networks may model this relationship more accurately.

Relationship Between Price and Fuel Type

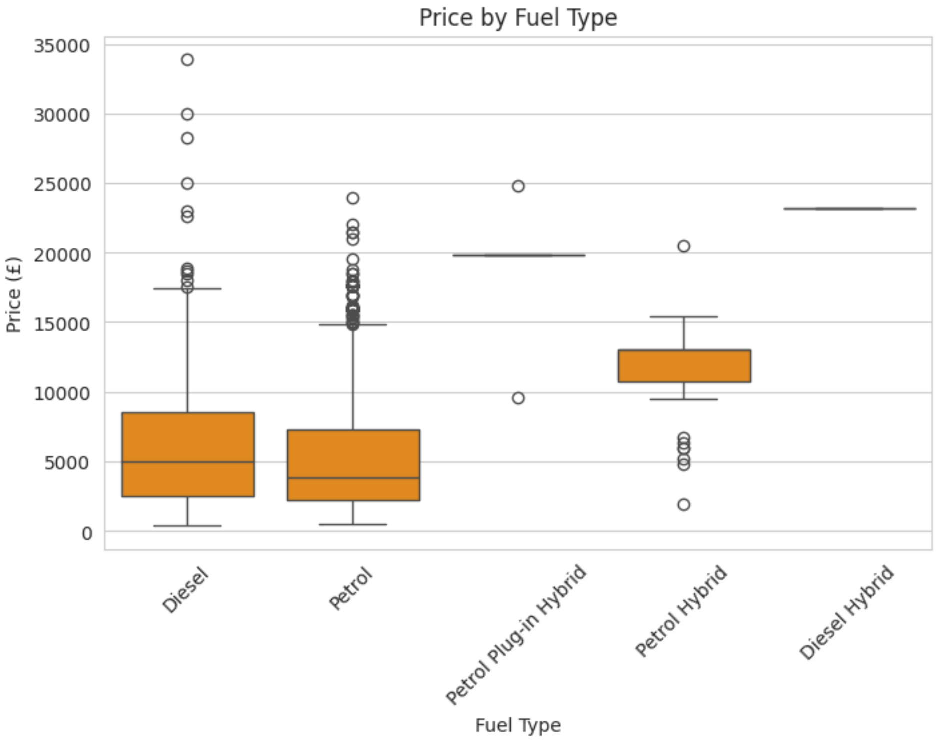


Figure 7: Boxplot of Price by Fuel Type

Fuel Type is a categorical feature that has a clear impact on price, suggesting it will be an important predictor. Because hybrid vehicles are consistently priced higher, models must be able to capture the effect of different fuel types accurately. Proper one-hot encoding of the Fuel Type variable will be critical for the models. Models that can naturally handle categorical variables (like Random Forests or Neural Networks) may model these differences better than purely linear models.

Relationship between Price and Body Type

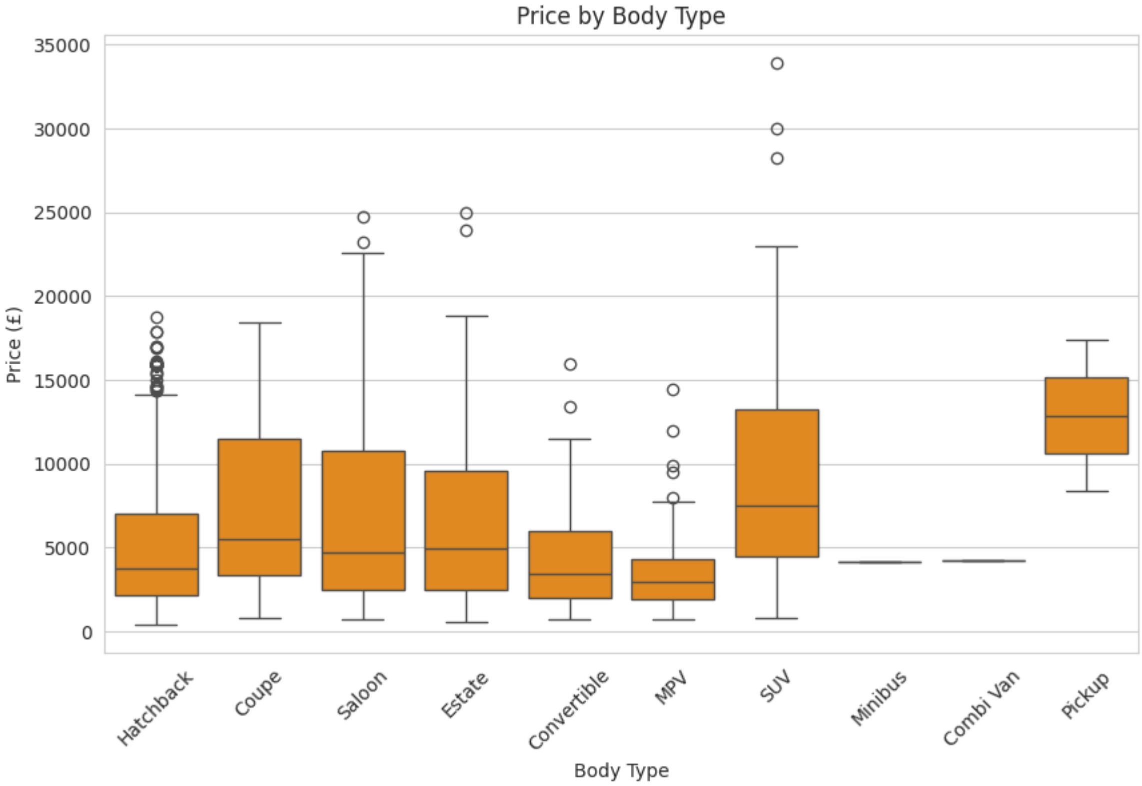


Figure 8: Boxplot of Price by Body Type

Body Type is an important categorical feature that impacts price and should be properly encoded before modelling. Some body types (like Minibus and Combi Van) have very few listings and low variability, which could introduce noise and models may either down-weight or struggle with these rare categories.

Relationship between Price and Number of Previous Owners

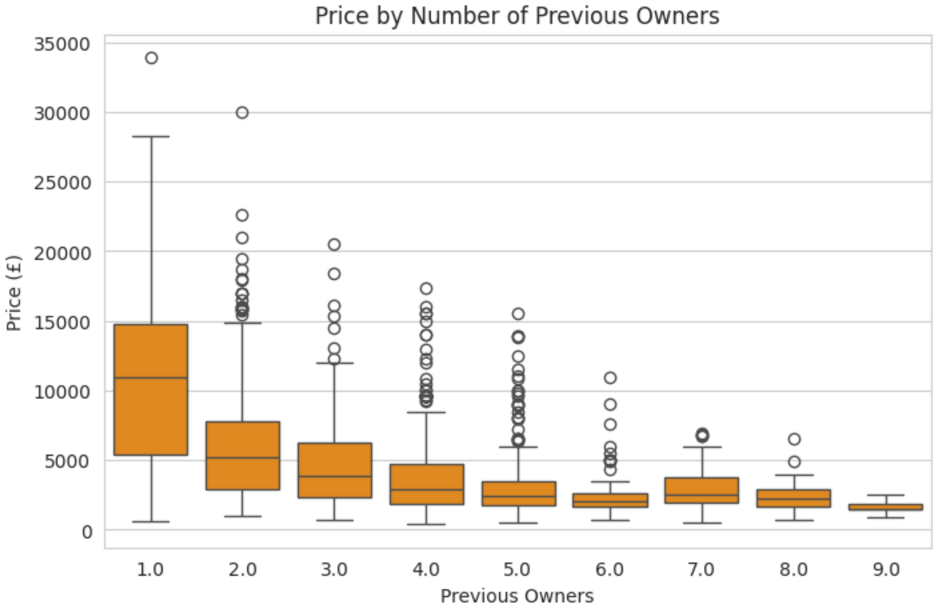


Figure 9: Boxplot of Previous Owners vs Price

The boxplot of Price by Number of Previous Owners shows a clear negative relationship between the number of previous owners and the price of the vehicle. The trend highlights that vehicles with fewer previous owners are considered more valuable, likely due to perceived better maintenance and lower likelihood of hidden issues. Extreme cases (e.g., 8 or 9 previous owners) could act as influential points in the model but are relatively rare.

Relationship between Price and Number of Engine Size

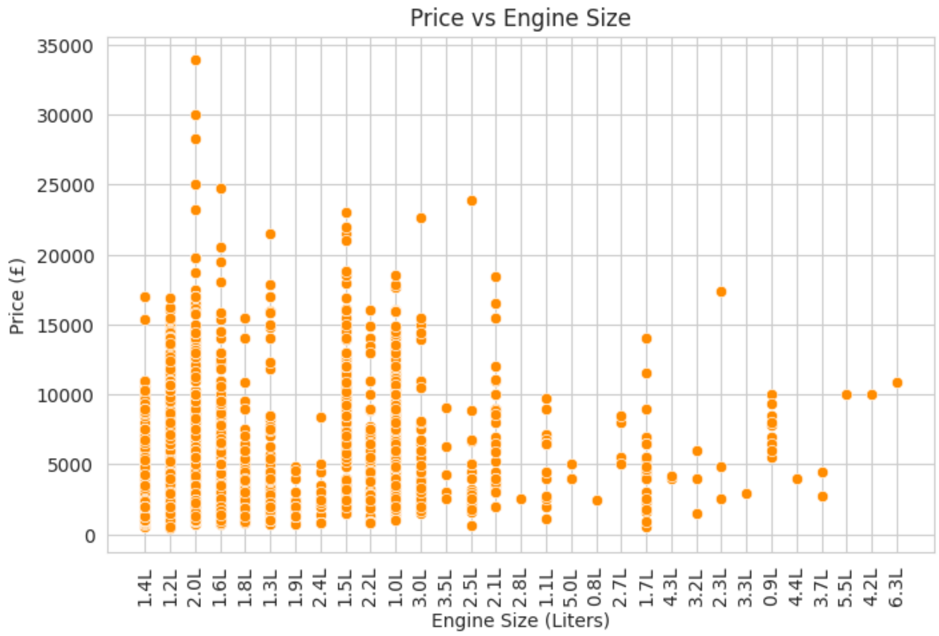


Figure 10: scatterplot of Price vs Engine Size

The scatterplot of Price vs Engine Size shows a complex relationship between engine size and car price. There is no simple linear trend and while some larger engine vehicles (above 2.5L) are priced higher, many have prices similar to smaller engine cars or even lower. Engine size may not be the best predictor but It could be included in the model along with other features.

Relationship between Price and Gearbox Type

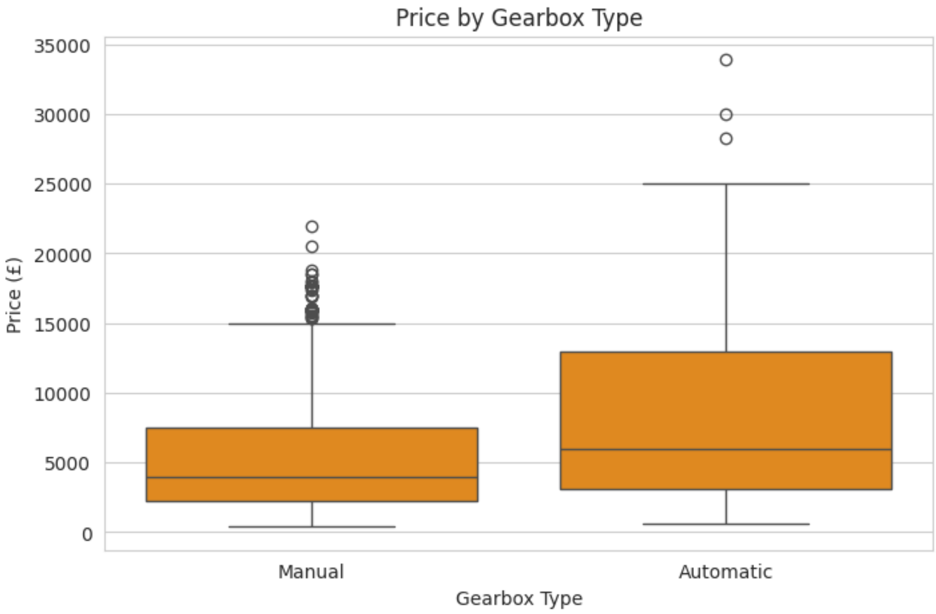


Figure 11: boxplot of Price by Gearbox Type

The median price for Automatic cars is noticeably higher, with a wider price range that extends beyond £25,000 for some listings. Gearbox Type is an important categorical variable that influences price and should be included in the model. Because the relationship between Gearbox Type and Price is strong but category-specific, one-hot encoding will be necessary to properly represent it for most models.

Correlation Analysis

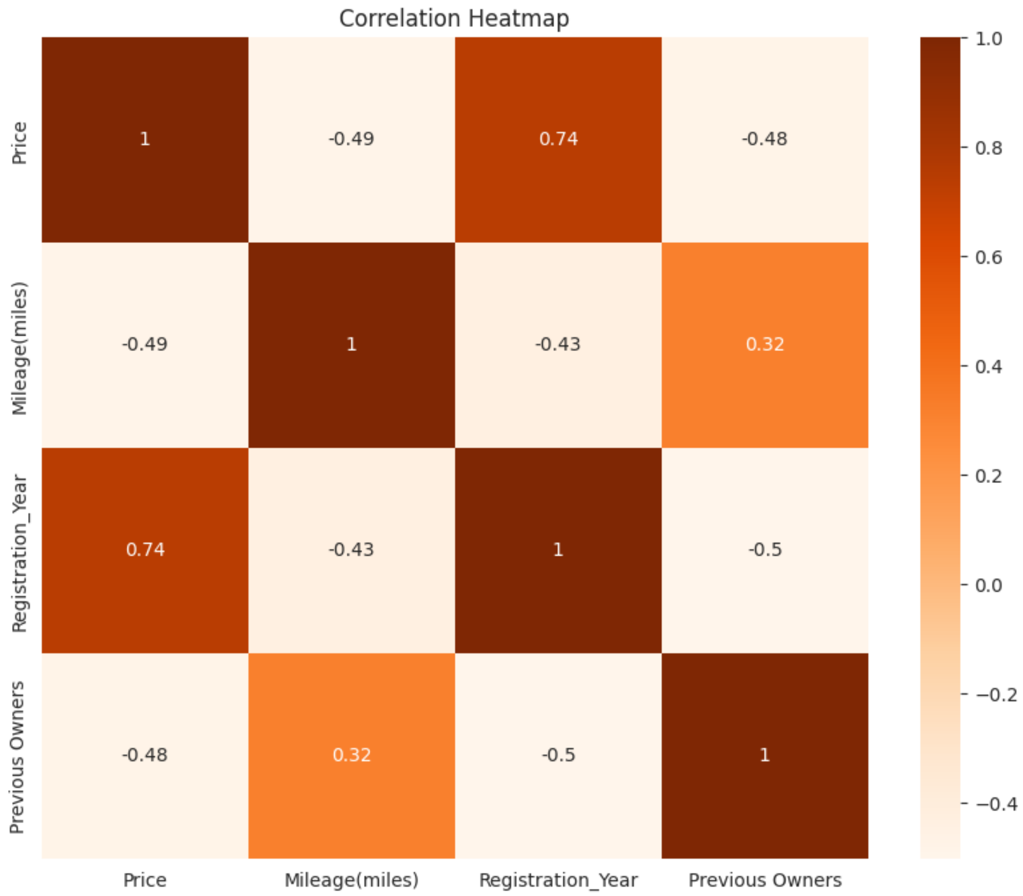


Figure 12: Correlation heatmap of numerical features

Features like Registration Year, Mileage, and Previous Owners are strongly associated with the target and are expected to be important predictors. While some features are moderately correlated, the dataset does not show perfect multicollinearity, meaning all variables can be retained without causing major issues for most models.

Models and Interpretation

To predict used car prices in the UK, I decided to test out several different regression models and see which one worked best. To prepare the data for modeling, the target variable (Price) was separated from the predictor variables (X). All columns except Price were used as predictors. After defining the features and target, I used an 80/20 train-test split for all my models, training on 80% of the data and testing on the remaining 20% to get a better sense of how each model would perform on new, unseen data. The random state was set to 42 to ensure reproducibility.

Baseline Model

Before training the more complex models, I established a simple baseline model to provide a reference point for model performance. The baseline model predicted the mean price from the training data for every test example, regardless of the car's features. This method represents a basic, average-based approach that does not take any vehicle-specific attributes into account.

The baseline model's predictions were evaluated using Mean Squared Error (MSE).
The resulting baseline MSE was approximately 23,290,338, reflecting the high variance in actual car prices. MSE measures the average squared difference between the predicted values and the actual values. A lower MSE indicates better model performance, while a higher MSE suggests that predictions are far from the true values.

The baseline Root Mean Squared Error (RMSE) was approximately 4,826. This means that, on average, the baseline model's price predictions were off by about £4,826. This large error highlights the need for more sophisticated models that can incorporate vehicle features to make more accurate predictions.

While the baseline model is not useful for practical predictions, it is important because it sets a minimum standard that more complex models must surpass to be considered effective. All subsequent models were compared against this baseline to ensure that they provided meaningful improvements in predictive accuracy.

Multiple Regression Model

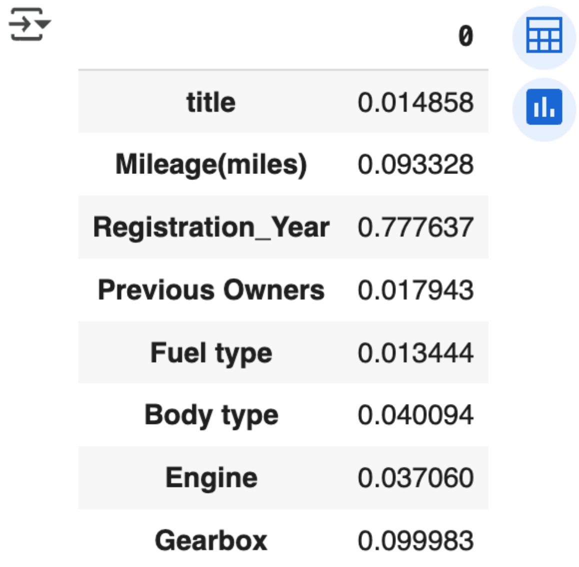
The first model trained was a Multiple Linear Regression model.

A preprocessing pipeline was created to one-hot encode the categorical features (Fuel Type, Body Type, Engine, Gearbox) while allowing numerical features (Mileage, Registration Year, Previous Owners) to pass through without scaling.

Compared to the baseline model, which had an RMSE of approximately £4,826, the multiple regression model achieved a much lower RMSE of approximately £2,448 (Test MSE of 5,994,153). This represents a significant improvement, demonstrating that using features such as mileage, registration year, and previous ownership history meaningfully enhances predictive accuracy over simply guessing the average price.

The R² score of 0.7426 indicates that the regression model was able to explain approximately 74% of the variance in used car prices, a strong result for an initial simple model. Additionally The Train MSE (5,646,676) and Test MSE values were close, indicating that the model did not suffer from overfitting.

I then calculated the permutation importance of each feature to see which are most relevant to this model.



	0
title	0.014858
Mileage(miles)	0.093328
Registration_Year	0.777637
Previous Owners	0.017943
Fuel type	0.013444
Body type	0.040094
Engine	0.037060
Gearbox	0.099983

Figure 13: Permutation importance in Multiple Regression model

Registration Year was by far the most important predictor, followed by Gearbox and Mileage. On the other hand, Fuel Type, Title, and Previous Owners had relatively small contributions to the model's predictions.

I then also found the coefficients for the numerical features. Registration Year was the strongest positive driver of price with each additional year increasing the predicted car price by about £590. Mileage had a negative impact with every extra mile reducing the predicted price slightly. Previous Owners also decreased the predicted price with each additional owner lowering the value by roughly £329.

While multiple regression significantly outperformed the baseline, it has some limitations. The model assumes a strictly linear relationship between features and price, which may oversimplify the more complex, nonlinear relationships observed during EDA (for example, sharp price increases for very new vehicles or variable mileage effects). It is also sensitive to outliers and may not perform well for cars priced at extreme values.

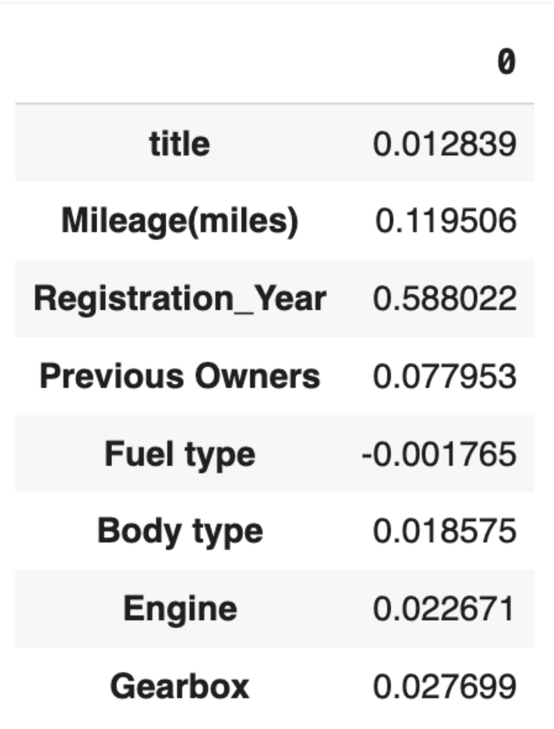
K-Nearest Neighbours (KNN) Regressor

After building the multiple linear regression model, a K-Nearest Neighbors (KNN) regression model was trained to explore a more flexible, non-parametric approach.
Unlike linear regression, KNN makes predictions based on the prices of nearby data points in feature space, allowing it to naturally capture nonlinear relationships.

Because KNN is sensitive to feature scaling, the preprocessing pipeline was modified to standardize numerical features using a StandardScaler. A GridSearchCV procedure was applied to find the optimal number of neighbors (k), trying values from 1 to 20 with 5-fold cross-validation. The best value found was k = 9.

The KNN model significantly improved upon the baseline and the multiple regression model with a test RMSE of approximately £1,771 (Test MSE of 3,136,259). The R-square value of the model also increased to 0.8653 and the Train MSE (2,881,639) and Test MSE values were close, indicating that the model did not suffer from overfitting.

I then calculated the permutation importance of each feature to see which are most relevant to this model.



	0
title	0.012839
Mileage(miles)	0.119506
Registration_Year	0.588022
Previous Owners	0.077953
Fuel type	-0.001765
Body type	0.018575
Engine	0.022671
Gearbox	0.027699

Figure 14: Permutation importance in KNN model

Registration Year remained the most influential feature. Mileage and Previous Owners also became more important compared to the linear model. Interestingly, Fuel Type had a slightly negative importance, suggesting it may introduce a small amount of noise.

The KNN model offered substantial improvements, especially by flexibly capturing nonlinearities between features and price. However, it is sensitive to noise and outliers or irrelevant features can heavily influence predictions.

Decision Tree Regression model

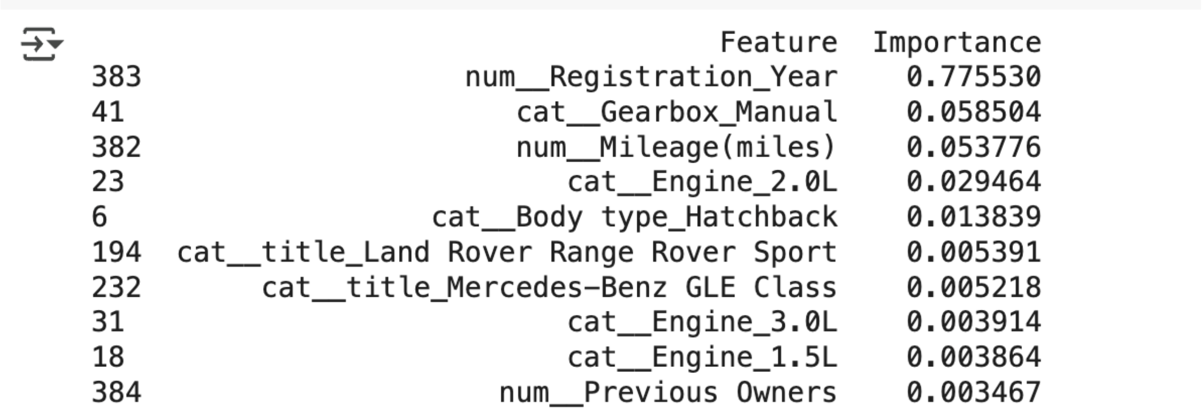
The next model trained was a Decision Tree Regressor, designed to capture nonlinear patterns in the data without requiring feature scaling. Decision Trees work by splitting the data at different feature thresholds to minimize prediction error.

A GridSearchCV procedure was used to tune hyperparameters, testing different values for max_depth and min_samples_split with 5-fold cross-validation to minimize mean squared error. The best hyperparameters found were: max_depth = 10, min_samples_split = 10

The Decision Tree model achieved strong predictive performance, with a test RMSE of approximately £1,755 (Test MSE of 3,082,804). The R² score on the test set was 0.8676 which shows that the model improved in comparison to the previous ones.

However, the Train MSE was 1,006,177, indicating that the model suffered from severe overfitting. The train MSE is significantly smaller than the Test MSE. Thus the model may be less stable for future unseen data.

I then calculated the feature importance from the trained Decision Tree model to identify the most relevant predictors:



```
                                     Feature  Importance
383                   num__Registration_Year    0.775530
41                       cat__Gearbox_Manual    0.058504
382                     num__Mileage(miles)    0.053776
23                          cat__Engine_2.0L    0.029464
6                    cat__Body type_Hatchback    0.013839
194  cat__title_Land Rover Range Rover Sport    0.005391
232       cat__title_Mercedes-Benz GLE Class    0.005218
31                          cat__Engine_3.0L    0.003914
18                          cat__Engine_1.5L    0.003864
384                     num__Previous Owners    0.003467
```

Figure 15: Feature importance for Decision tree model

Registration Year again emerged as the most influential feature by a large margin, similar to the previous models. Gearbox type, mileage, and specific engine sizes (such as 2.0L and 3.0L engines) were also significant contributors. Certain luxury car models (like Land Rover and Mercedes-Benz) showed small but non-negligible importance in predicting prices.

The Decision Tree model provided excellent improvements, especially by modeling complex interactions between features without requiring any transformations. However, while decision trees are powerful, they can still be sensitive to small data variations and risk overfitting like in this case.

Random Forrest Regressor

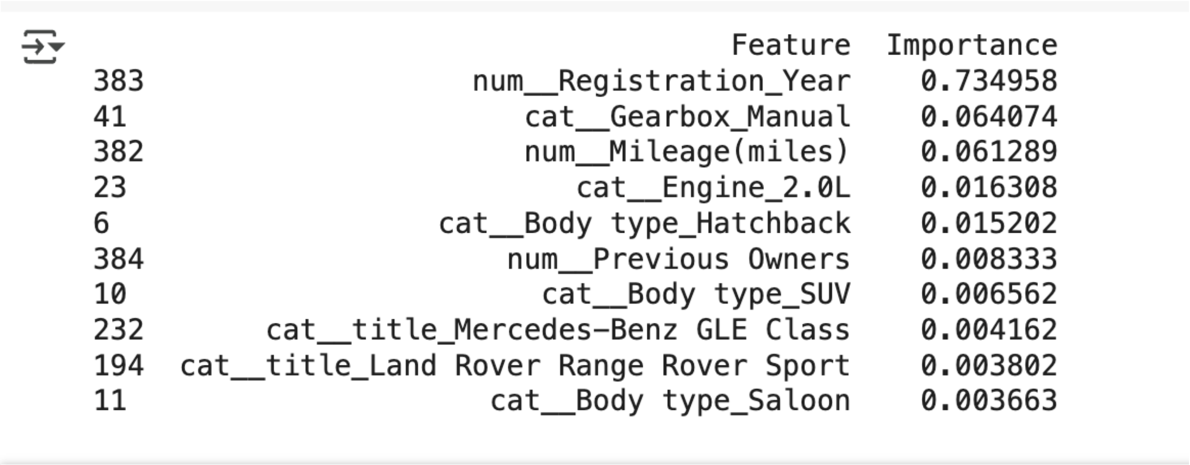
I then built a Random Forest Regressor to further improve prediction accuracy. Random Forest is an ensemble learning method that fits many decision trees and averages their predictions, making it robust to overfitting and able to capture complex relationships in the data.

A GridSearchCV was used to find the optimal combination of hyperparameters, testing different numbers of trees and maximum tree depths using 5-fold cross-validation. The best parameters were found to be 300 trees with no maximum depth constraint.

The Random Forest model achieved a strong performance, with a Test MSE of approximately 2,203,038, corresponding to a Root Mean Squared Error (RMSE) of about £1,484. The R^2 score on the test set was 0.9054, indicating that the model explained about 90.5% of the variance in the car prices. Much better than all previous models.

However, The training MSE (360,016) was much lower than the testing MSE, suggesting that the Random Forest overfit the training data slightly. This behavior is expected with ensemble methods when trees are allowed to grow without depth limitations, as they can perfectly fit the training data but generalize less well to unseen examples.

I also calculated feature importances from the Random Forest model to identify the most influential predictors.



```
                                   Feature  Importance
383                 num__Registration_Year    0.734958
41                     cat__Gearbox_Manual    0.064074
382                    num__Mileage(miles)    0.061289
23                        cat__Engine_2.0L    0.016308
6                   cat__Body type_Hatchback    0.015202
384                   num__Previous Owners    0.008333
10                      cat__Body type_SUV    0.006562
232        cat__title_Mercedes-Benz GLE Class    0.004162
194   cat__title_Land Rover Range Rover Sport    0.003802
11                   cat__Body type_Saloon    0.003663
```

Figure 16: Feature importance for Random Forrest Regressor

Registration Year continued to be by far the most important feature, similar to the findings in earlier models. Other important factors included whether the car had a manual gearbox, mileage, and engine size, with some specific vehicle models (Mercedes and Land Rover) also contributing.

The Random Forest Regressor was the best model so far in terms of test accuracy, but it showed mild overfitting due to unrestricted tree depth. Additionally, the model’s complexity made it less interpretable than the simpler models like linear regression.

Neural Network

I finally built a Neural Network model using a Multi-Layer Perceptron (MLP) regressor to explore more complex, nonlinear relationships in the data. Neural networks, unlike simpler models, are capable of approximating highly intricate patterns due to their multiple layers and neurons.

Since neural networks are sensitive to feature scaling, the preprocessing pipeline was modified to apply a StandardScaler to numerical features. I then used GridSearchCV to evaluate combinations of architectures with 50 or 100 neurons, and both 'relu' and 'tanh' activation functions. The best model used a hidden layer structure of (50, 50) and the 'relu' activation function.

The test MSE was the lowest out of all models at 1,898,002 corresponding to an RMSE of about £1,377. It means that the models predictions were only off by £1,377. The R² score of 0.9185 was the highest indicating that the MLP was able to explain over 91% of the variance in car prices on the test data.

The training MSE was 641,915, lower than the test MSE, suggesting some overfitting. However, the gap was not extreme enough to indicate severe overfitting, especially for a neural network model.

Additionally, the training loss curve (shown below) steadily decreased and plateaued, indicating that the model had converged properly during training without major instability.

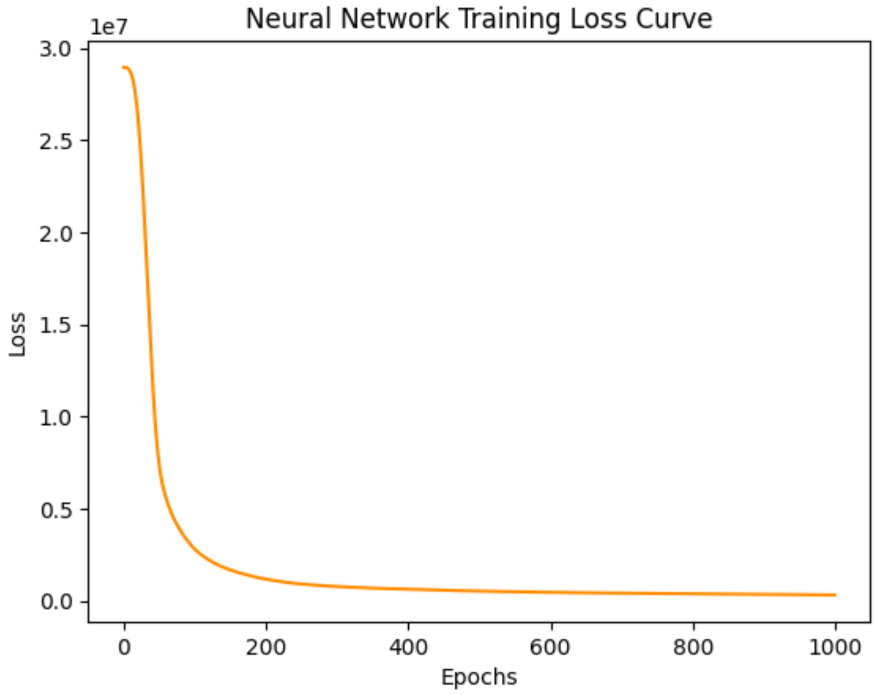
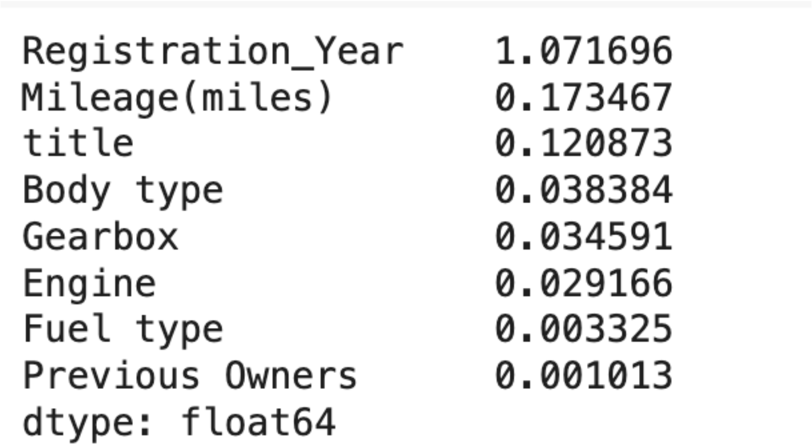


Figure 17: Neural network training loss curve

I then also calculated each feature’s importance.



```
Registration_Year    1.071696
Mileage(miles)       0.173467
title                0.120873
Body type            0.038384
Gearbox              0.034591
Engine               0.029166
Fuel type            0.003325
Previous Owners      0.001013
dtype: float64
```

Figure 18: Feature importance for Neural Network Model

Registration Year remained the most influential feature, with Mileage and Title (information about the car’s branding) following. Features like Previous Owners and Fuel Type had relatively low importance.

This model produced the best performance so far, capturing complex relationships in the data. Although it slightly overfit the training set, it still generalized well to the test data.

Conclusion

Summary of Findings

The baseline performed very poorly, highlighting the need for more complex models. Multiple linear regression improved accuracy but was limited by its inability to model nonlinear relationships.

Models like K-Nearest Neighbors, Decision Trees, and Random Forests captured nonlinearity better and significantly reduced errors. The best performance was achieved by the Neural Network (MLP Regressor), which had the lowest Test MSE and highest R² score (0.9185).

Throughout all models, Registration Year consistently emerged as the most important feature influencing price prediction.

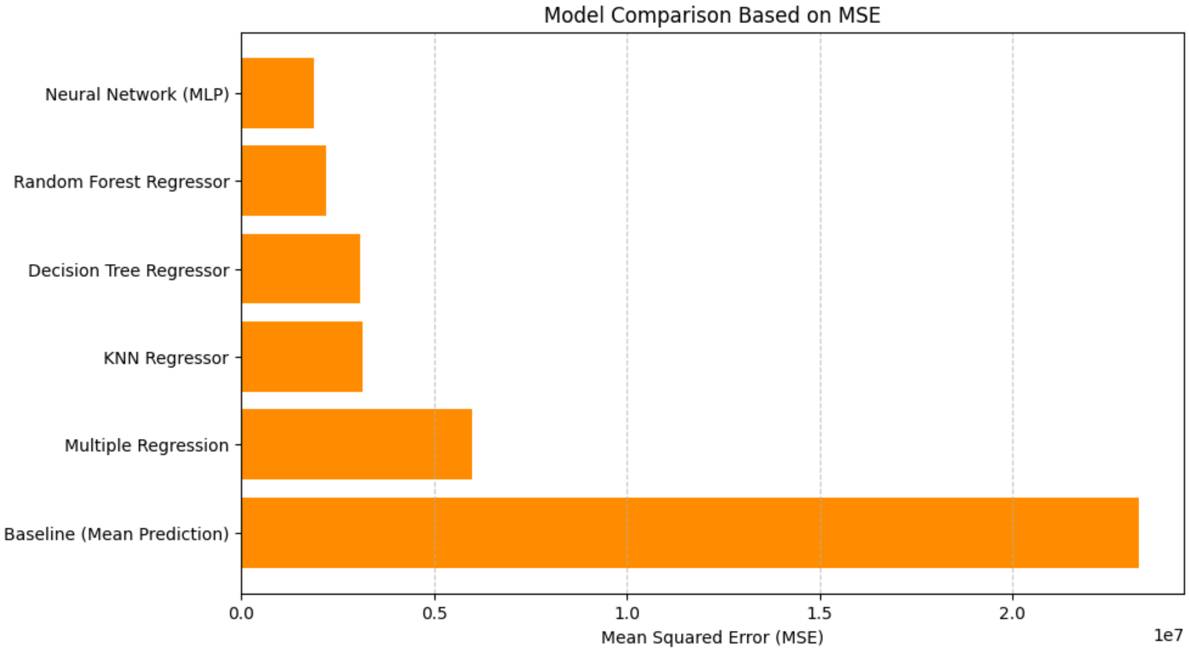


Figure 19: Model Comparison based on MSE

Key Insights

Feature Importance: Across all models, Registration Year was consistently the most important predictor of car price. Mileage, Gearbox, and Title features also played key roles.

Nonlinear Models Outperformed Linear Models: Models that could flexibly capture nonlinearity (KNN, Decision Trees, Random Forests, and Neural Networks) dramatically outperformed simple linear regression.

Model Complexity Tradeoffs: More complex models (especially Random Forests and Neural Networks) offered better accuracy but introduced greater risks of overfitting. Proper regularization, hyperparameter tuning, and careful evaluation were essential.

Next Steps/ Improvements

To enhance the predictive capabilities of the models and gain deeper insights into used car prices in the UK, I would want to incorporate these additional features into my models:

1) Hyperparameter Tuning: Conduct a more thorough grid search or use Bayesian optimization to explore a wider range of parameters, especially for complex models like Neural Networks and Random Forests.

2) Feature Engineering: Create new features like car age (Current Year - Registration Year), interaction terms (e.g., Mileage × Age), or advanced encodings for categorical variables (e.g., target encoding) to give models more predictive power.

3) Model Ensembling: Combine multiple models (e.g., averaging predictions from Random Forest and Neural Network) to potentially improve accuracy and robustness.

Bibliography

Dataset: https://www.kaggle.com/datasets/muhammadawaistayyab/used-cars-prices-in-uk

EDA Notebook: https://colab.research.google.com/drive/1EHo4ypn_cxKA_CrkgJ2S-Wo8gfaoqper?usp=sharing

Modelling Notebook: https://colab.research.google.com/drive/1KNy-GzQymHYRi5TxkV2zG3O03ihx3pxA?usp=sharing