Predicting prices of used cars

Harid Voranard

Supervisor: Dr. Marika Asgari

Abstract

Used car transactions, car types and variations in features have increased globally. This increased market complexity generates the need, amongst both buyers and sellers, for a model which evaluates a used car according to its features compared with other cars. Extensive data sets and machine learning algorithms now enable such predictions to be made. Previous studies have achieved predictive accuracy levels of between 75% and 95%, using linear regression, random forest and XGBoost techniques. In this study, these were applied to a data set of UK used car prices and another of car energy consumption in Canada, applicable worldwide. After merging and cleaning the data, a sample of 35,370 was created. The Random Forest proved to be the most efficient model in predicting used car prices, with r-squared 99 percent for the regression model. The most important factors are age and mileage - when age and mileage go up, the price, understandably, goes down. Cars with the lowest CO2 emission, the best fuel economy and the most eco-friendly were found to be provided by Toyota. The model has been deployed as a web application on Heroku Website[1]. In future studies electric cars will need to be added to the data set.

[1] https://used-car-prices-by-haridv.herokuapp.com/

Introduction & Background

The purchase of a car, either new or second-hand, represents a high-involvement activity in consumer behaviour. In the case of second-hand cars, there is a mechanism to establish a price which both buyer and seller can agree on. In the past, this was established mainly by the seller, according to guides produced by the trade, or on newspaper and other press formats. (Insurancefactory, 2020).

Today, a much more complex mechanism exists, based on extensive data sets and multiple variables. This process can now be enabled by applying machine learning techniques to publicly available data sets. These are explored in this report to forecast used car prices.

The main predictive techniques are linear regression method which establishes a linear relationship between the input (called 'X') and the prediction (called 'Y'), with a formula in the form of y = mx + b (Gupta, 2018). The Random Forest consists of a collection of decision trees. Random Forest is made up of many trees built in a "random" way. Each tree consists of a specific row of samples, and each node is divided into separate feature sets. The average of these predictions is then used to reach a single result (Mwiti, 2020).

Several studies have attempted to model used car prices, based on these techniques. For example, Pal et al., (2018) used a random forest that was trained on a Kaggle dataset to predict the cost of used cars and obtained high accuracy. The model was chosen after they had conducted a thorough exploratory data analysis to see how each variable affected pricing. A Random Forest was built with 500 decision trees to train the data. The model can predict with an empirical accuracy of 83.63% for testing and 95.82% for training. By choosing the most significant factors in the dataset, such as original price, kilometre, brand, and vehicle type, then removing outliers and unimportant features, the model can accurately anticipate vehicle pricing.

On the other hand, Puteri and Safitri (2020) used linear regression to predict used car prices in Indonesia from two producers, including Toyota and Honda and analyzed price against car mileage and price against the age of used cars. They obtained an accuracy prediction of no more than 75%.

This result fits with the findings of Mwiti (2020), who established that although linear regression is simpler than random forest, it does not work so well on large data sets.

In addition, XGBoost, a tool for building supervised regression models, has been used to predict used car prices in Bangladesh. The objective function and base learners can be used to determine the model's veracity. The objective function contains a loss function and a regularization term. It discusses the distinction between predicted and actual values (GeeksforGeeks, 2020). Amik et al., (2021) use of XGBoost for regression correctly predicted prices with more than 91%. Moreover, this person deployed on website on local machine.

As a result of the above assessment, two data sets were chosen for this project, along with Random Forest, Linear regression, and XGboost methods.

The purpose of the project is to gather pre-owned car datasets and identify the most important features that can be used to predict the price of used cars. The use of machine learning techniques to accurately anticipate the cost of used cars can assist a prospective buyer in making a more informed decision when buying a secondhand car. Recommended models, with the lowest CO2 emission and energy consumption, will be placed onto a website. Heroku was chosen because polyglot support, such as Python, HTML, and CSS, is needed. Heroku is a computer-based cloud application. It helps manage and deploy software. Heroku offers a Platform as a Service (PaaS) that enables scripts to run directly without extensive configuration. The Heroku experience provides services, tools, workflows, and polyglot support (Heroku, no date).

Data

Data Collection and combining the data

The first data set was obtained from the official open data website on Kaggle [2]. It comprises 85,555 used car prices in the UK from eight producers - Toyota, Audi, Hyundai, Volkswagen, Mercedes-Benz, BMW, Skoda, and Ford - covering year of used cars, transmission, fuel type, miles per gallon(mpg), engine size, model, mileage, tax, and price.

The second data set was obtained from the official open data website of the Canadian government[3]. This includes 7,385 vehicles' CO2 emissions and energy consumption. The data set contains make, vehicle class, cylinder, fuel type, CO2 emission, model, engine size, transmission, and fuel consumption.

The two data sets were obtained from different countries and needed to be merged. During this process, some issues arose. One was the linking of model with engine size to create a new column – model-engine size. In some cases, the models were not the same. This was the case with Skoda and Mercedes Benz, so these records were removed.

Data Preparation

The data contained text which was unnecessary and could not be analysed. An example of this is the use of "other". The transmission category contained Automatic, Semi-auto, Manual and other. The other category was removed so as to avoid inaccurate predictions. Under fuel type**,** Diesel, Petrol, Hybrid and other were listed. Again, the unspecific "other" category was removed.

After assessment, some of the data needed to be cleaned, as follows:

a) BMW and Hyundai cars represented 1.12 and 0.74 percent of the total database, respectively. They were removed.
b) The highest used car price was £145,000, but there was only 0.37% of observations between this figure and £70,000. These were removed to not skew the results.
c) Similarly, upper limit for mileage was adjusted from 176000 km to less than 150000 km because the range of more than 150000 km is only 0.02%.
d) In each year between 1998 and 2003, there were less than ten observations. Consequently, 2004 was adopted as the start year

[2] https://www.kaggle.com/datasets/adityadesai13/used-car-dataset-ford-and-mercedes
[3] https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64

Another factor was that the csv file of used car prices in the UK was released in 2020. To find differential such as 2020(data released) - 2020(model year) = 0 years because it is impossible to have 0-year-old cars. It should change from 0 years to less than 1 year.

The reduction of the data set to four companies – Audi, Volkswagen, Ford and Toyota – reduced the total sample size from 85,555 to 35,370. The loss of observations for Skoda was 6,267, Mercedes Benz, 13,119, BMW, 10,781 and Hyundai, 4,860.

In the next step of data encoding, a numerical representation of these categorical data was generated so as to be able to train our model. A one-hot encoder and dummy functions were applied to convert categorical features such as fuel type, transmission, name, and company into numeric data (binary values). It needs converting because the computer can understand when the values are numeric data. (Garg, 2022).

Variable Name	**Values**	**Type**
Company	Audi, Volkswagen, Ford, Toyota	categorical
Name	35 unique values such as Prius-1.8	categorical
Year	less than 1 year to 16 years.	categorical
Kms_driven	0 to 150000 Km	numeric
Fuel_type	Diesel, Petrol, Hybrid	categorical
Transmission	Manual, Automatic, Semi-auto	categorical

Table 1: Type variables

As part of the machine learning process, it is important to split the dataset into train and test sets because it is necessary to test the predictive accuracy of the model. The random values were split 80:20 from the dataset. The purpose is to be able to train the model in a single dataset, but testing in different new data sections. (Brownlee, 2020).

Training Data	**Testing Data**
28,296 cars (80 percent)	7,074 cars (20 percent)

Table 2: Train and test split.

Methodology

As discussed in the background section, three relevant previous studies, which have sought to predict used car prices, were identified. Based on the results of these studies, linear regression, XGBoost, and random forest models were chosen for this project.

Linear regression is a suitable place to start, as it is relatively straightforward. However, it only applies if the answer is linear. It might not be the case in many real-world circumstances (Varghese, 2018). On the other hand, Random Forest creates ensembles of decision trees to improve performance and robustness deficiencies (Dudoit et al., 2002). Whilst offering promise, random forests have drawbacks. For example, Random Forest cannot detect trends in extrapolated values outside the range of the training set. (Ballings et al., 2015). In comparison, XGBoost is the better choice for unbalanced datasets. When a model first fails to forecast an anomaly, XGBoost, gives it more weight and preferences in subsequent iterations, boosting its capacity to predict the class with low participation. It is the case that random forest cannot be guaranteed to handle the class imbalance in a good manner (Gupta, 2021).

The effectiveness of the three models will be assessed using statistical indicators including mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and R^2 Score.

The percentage of a dependent variable's variance that can be explained by an independent variable or set of independent variables in a regression model is expressed statistically as R^2 Score (Fernando, 2003). The original-to-predicted value difference is represented by MAE, which averages the absolute difference throughout the data set. The average difference over the data set squared represents the MSE, or mean square error, between the original and anticipated values. The square root of the MSE error rate is known as RMSE (DataTechNotes, 2019). Higher accuracy of a regression model is implied by lower values of MAE, MSE, RMSE, and MAPE. On the other hand, MAPE returns an error as a percent, simple for end users to understand and makes comparing model accuracy across use cases and datasets easier (Allwright, 2022).

The built machine learning model was saved in pickle (.pkl) format. This process is called serialization. It was decided to conduct price predictions for used car models on the web. The next process was to create a website with used car model price prediction as the backend using the Flask framework, Flask is the web framework written for python to the webserver (Flask, 2022). After that, the next step was to create an interface as the front end using Python, HTML, and CSS. Finally, the built system was deployed using the Heroku platform for cloud services.

Results

Following the assessment, adjustment and cleaning of the sample data, discussed in Section 2, a total sample of 35,370 resulted. Figure 2 sets out the distribution of observations by transmission type, fuel type and company.

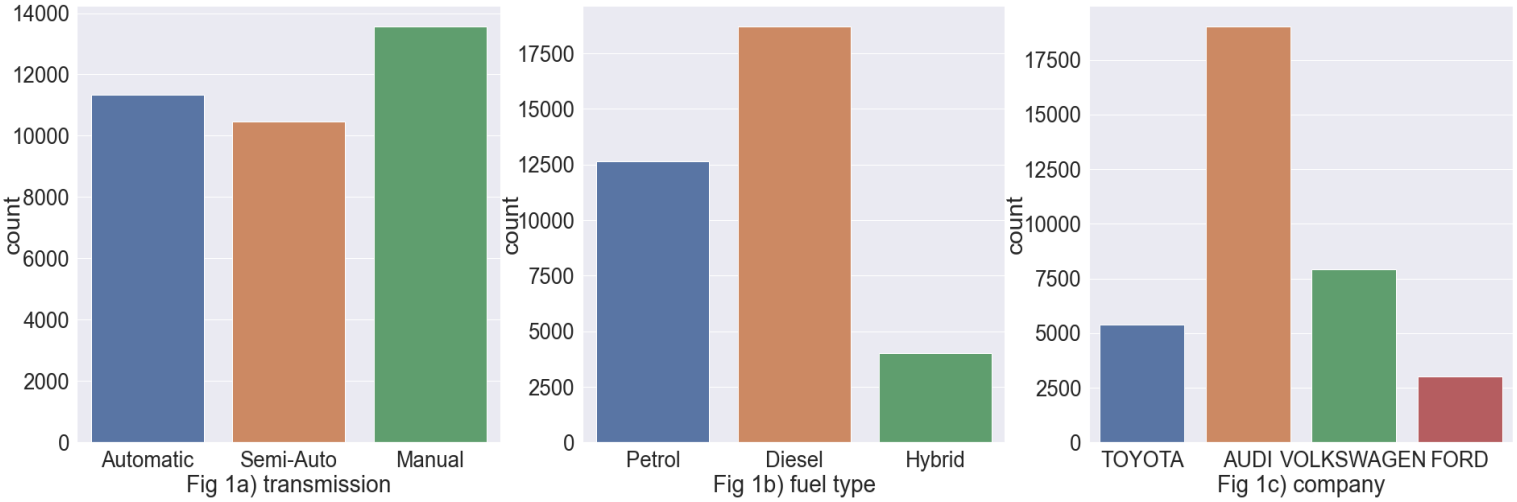

Figure 1: Number of cars with transmission, fuel type, and company. Figure1a illustrates that there is a slightly higher number of manual cars than automatic – 13,567 compared with 11,331, semi-automatic accounting for a further 10,472. Figure 1b reveals 18711 diesel-fueled cars, followed by 12,649 petrol-driven cars and 4,010 hybrid. In Figure 1c, Audi is shown to be the dominant brand 19,032 cars, followed by Volkswagen, 7,933 cars and Toyota with 5,405 cars.

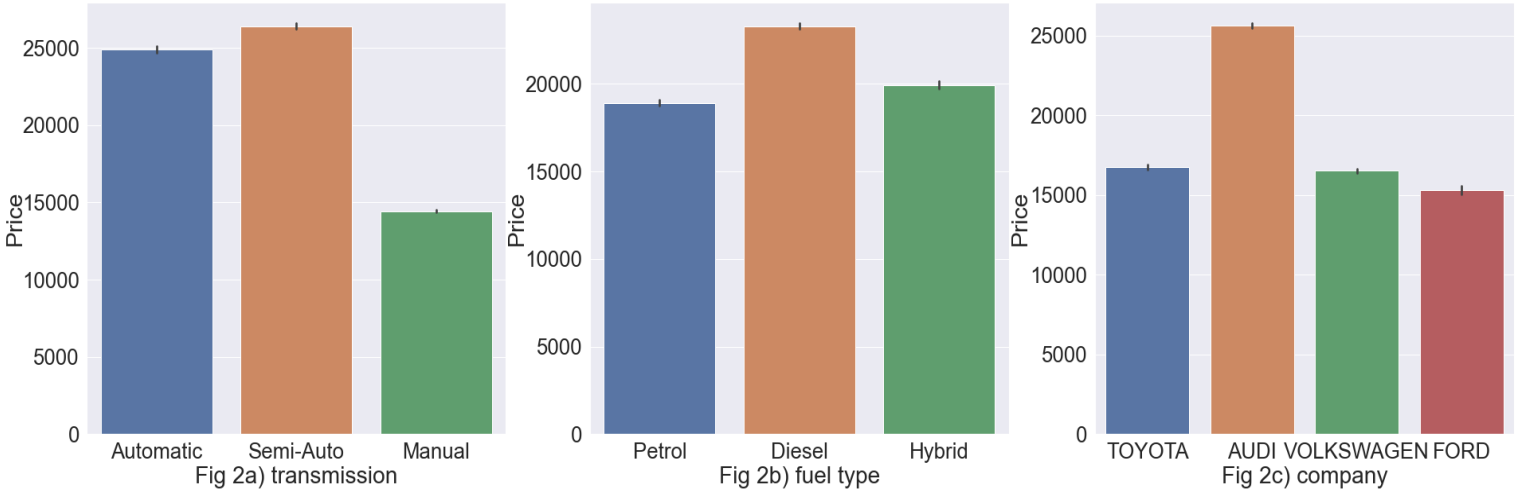

Figure 2: Average prices of transmission, fuel type and company. Figure 2a illustrates that semi-automatic cars have the highest price at £26411, followed by automatic at £24917 and manual at £14443. Figure 2b shows that both diesel and hybrid propelled cars, at £23294 and £19927, respectively, are higher priced than petrol, at £18902. Figure 2c indicates that the highest priced brand is Audi at £25613, followed by Toyota at £16748 and Volkswagen at £16507.

A linear regression was conducted on the sample, plotting price against distance driven in kms and taking account the age of the car.

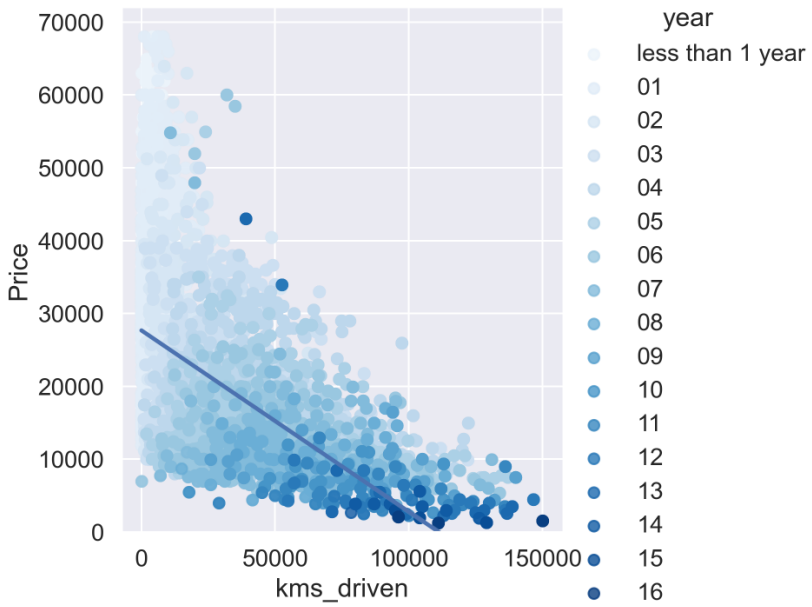

Figure 3: Linear regression of price against kms driven conditioned by year. The light blue of the scatter plot represents young cars and the dark blue signifies old cars. It can be seen that, the average initial price for young and less mileage (kms_driven) cars starts around £30000. After that, when the car is older and the mileage (kms_driven) increase, the prices reduce every year.

When estimating the price of used cars, not every feature has the same effect on the price. Therefore, the Pearson correlation can help people to understand factors that affect second-hand prices.

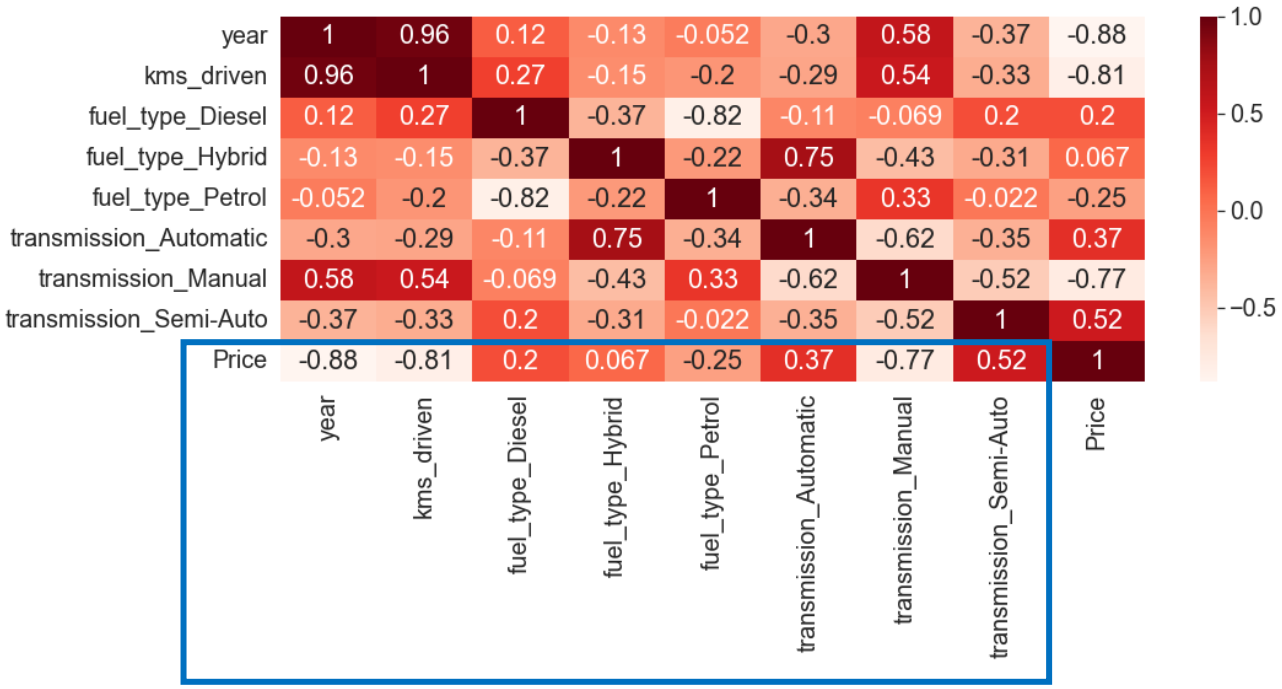

Figure 4 shows the most important feature by using the Pearson correlation technique. Nettleton (2014) explained that Pearson correlation assigns a value between − 1 and 1: − 1 is a negative correlation, 0 is no correlation, and 1 is a positive correlation.

The first focus is on the price in the last row, framed in the blue block, because this project seeks to predict the price of used cars. The second focus is on the cells on the left of the row, where the highest negative correlation is year with -0.88 followed by mileage (kms_driven) with -0.81. The reason why they are negative is because Figure 4 has shown that when the year and mileage (kms_driven) go up the price goes down.

In conclusion, the variable that most affects prices are year, followed by mileage (kms_driven).

In order to determine which algorithm serves as the best regression model statistical indicators were used. as set out in Table 3.

Algorithm /model	**R^2 Score**	**MSE**	**MAE**	**RMSE**	**MAPE**
Linear Regression	92%	8450468	2074	2907	11.32%
XGBoost	96%	3542213	1329	1882	6.59%
Random Forest	99%	1020627	446	1010	2.19%

Table 3 shows the R^2 Score, MSE (mean squared error), RMSE (root mean squared error), MAE (means absolute error), and MAPE (means absolute percentage error). The principle of viewing this table is comparing each model from the R^2 Score, the closest to 1 and the lowest error (MSE, RMSE, MAE, MAPE) are the best. (DataTechNotes, 2019).

The best algorithm is Random Forest because R^2 Score has the highest value of 99% and the lowest error MSE 1020627, MAE 446, RMSE 1010, and MAPE is 2.19%, followed by XGBoost 96% and linear regression 92%. Previous studies have achieved predictive accuracy levels of between 75% and 95%.

The removal of outliers can increase the efficiency of predicting used car prices because the first time prices were predicted in the maximum range of £145000, MAPE 3.46% was obtained, but when the outliers were removed, MAPE 2.19% was obtained. It is especially the case that, when the price of expensive cars (more than £70000) was predicted, the result was inaccurate because the distribution price of more than £70000 is only 0.37% of the total. This represents a low sample for training. The random Forest cannot detect trends in extrapolated values outside the range of the training set. (Ballings et al., 2015).

In order to investigate how well the Random Forest model predicts car prices, 7,074 actual prices from the test data were compared with forecast ones. Figure 5 illustrates the result.

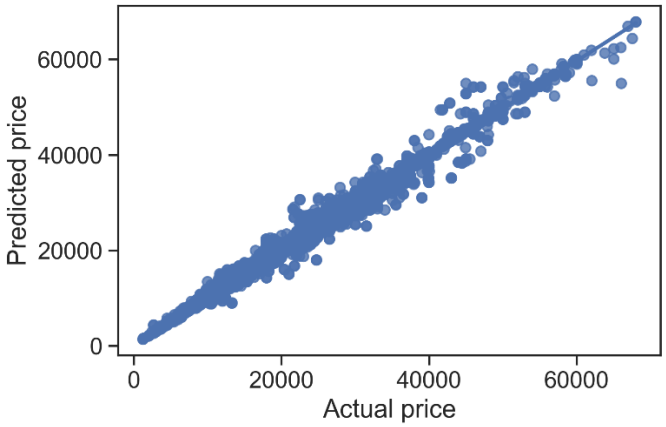

Figure 5: Predicted price vs actual price. If the scatter plot is close to linear regression, its accuracy is high.

However, it is difficult to know the standard deviation value. The error between a predicted value and the observed actual value can be seen by using residual plot, along with rotating linear model.

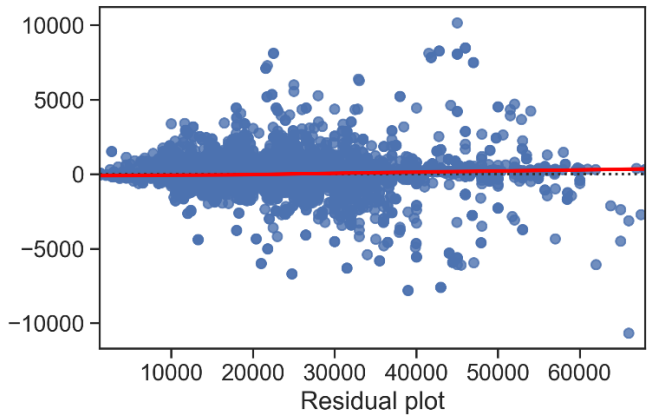

Figure 6: Residual plot. It can be seen that, the highest positive deviation around 10000. On the other hand, the highest negative deviation is around 10000. The red line shows the fit of the local regression method, locally weighted scatterplot smoothing (lowess), to the residual scatterplot. Local regression is another approach to fitting flexible nonlinear functions, where the fit at the target points is computed using only nearby training observations. (Kirenz, 2021).

Since the Random Forest model cannot detect trends in extrapolated values outside the range of the training set (Ballings et al., 2015), improved accuracy is required before the model is uploaded to the website. This is achieved as follows.

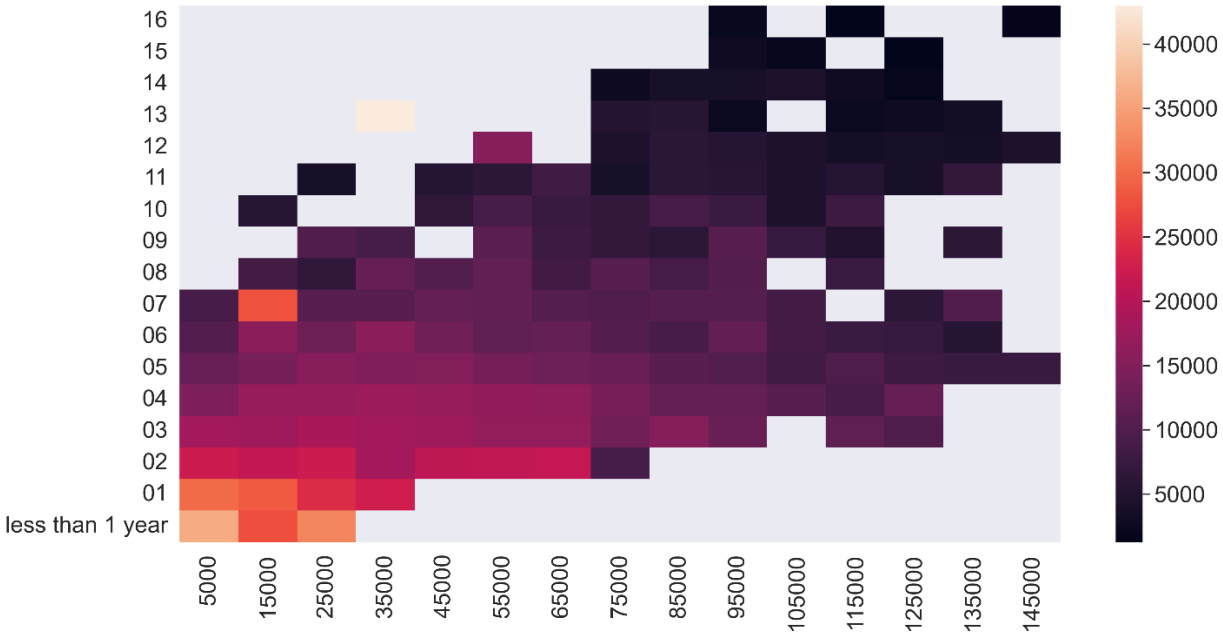

Figure 7: The record heat map of the year against mileage (kms_driven) conditioning by price. It can be seen that, the cream-colored block shows the expensive car for around £35000 at cars less than 1-year-old and the mileage around 0 to 5000. The colours show when the car is older and the mileage of car is increasing. Price trends go down every year.

	year	less than 1 year	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16
kms_driven	min	2.0	1.0	1000.0	917.0	1000.0	1800.0	80.0	7000.0	15000.0	25931.0	17733.0	29000.0	52500.0	39000.0	78356.0	95000.0	96000.0
	max	22601.0	33033.0	79032.0	122000.0	127000.0	140000.0	134000.0	131000.0	113500.0	139989.0	113000.0	136000.0	146604.0	138649.0	126323.0	129000.0	150000.0
	mean	2843.2	5967.3	16743.7	26903.4	37633.5	44418.9	48836.0	54506.4	67619.3	80453.5	77466.5	85310.6	95781.3	99395.1	98548.5	109714.3	116785.7

Table 4: The record table of the year against mileage (kms_driven).

Table 4 has been put on the website so that customers can use it to predict used-car prices and obtain better accuracy.

From the used car record, it can be seen that, understandably, young cars tend to have low mileage. On the other hand, it is not common that old cars have too low mileage.

A user of the website can input the desired values for parameters such as the company, model, year of car, transmission, fuel type, and mileage, as illustrated in Figure 8, below.

To see my website[4].

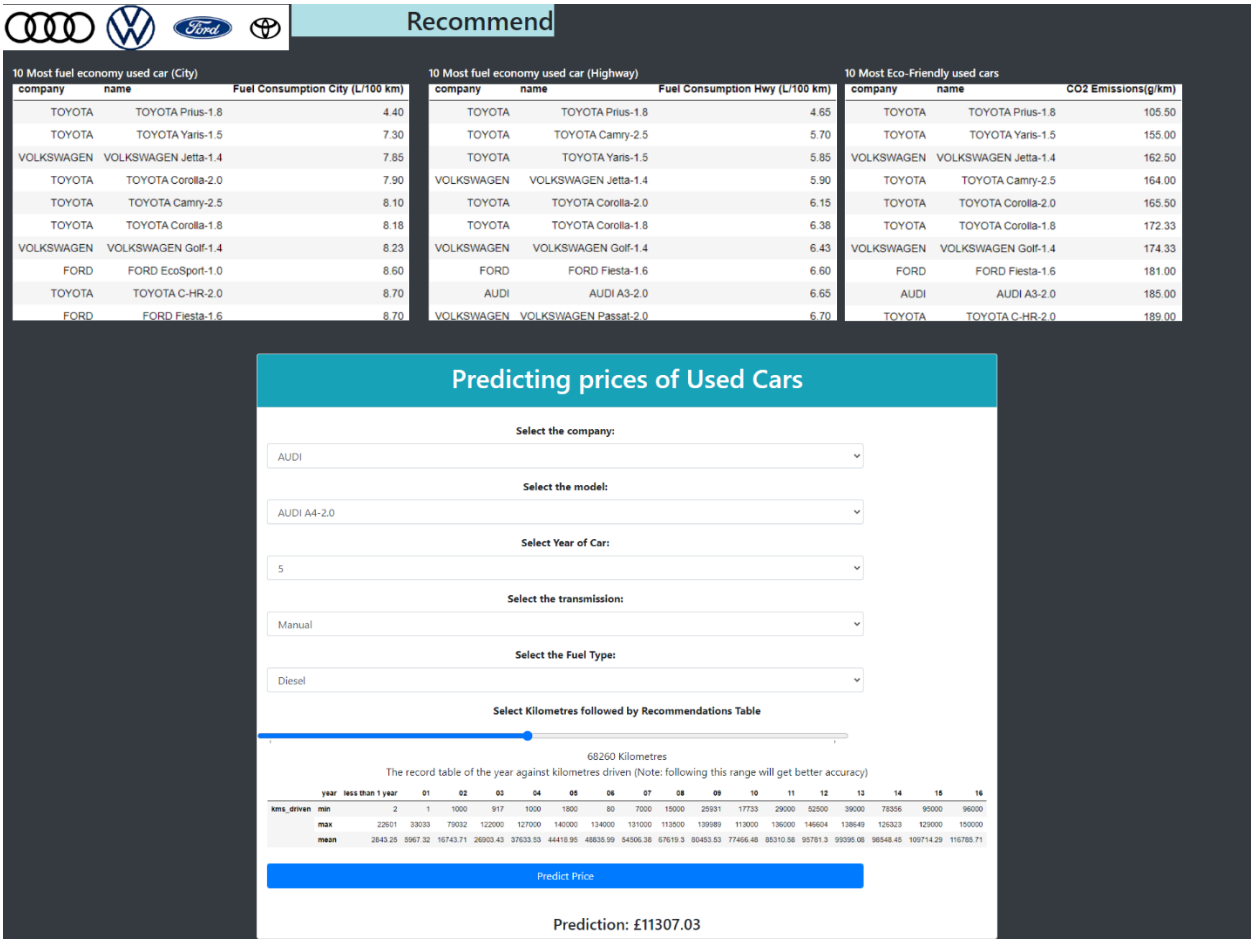

Figure 8: Prediction of used car website. The website includes a Table that recommends the best fuel economy used car (city). This is Toyota Prius-1.8 followed by Yaris-1.5 and Volkswagen Jetta-1.4, the best fuel economy used car(highway) is Toyota Prius-1.8 followed by Toyota Camry-2.5 and Toyota Yaris-1.5. The best eco-friendly used car is Toyota Prius-1.8 followed by Toyota Yaris-1.5 and Volkswagen Jetta-1.4.

1	name	company	year	kms_driven	fuel_type	transmission	Price
28997	AUDI A4-2.0	AUDI	5	68260	Diesel	Manual	11295

Table 5: Actual price from CSV file

In order to establish the accuracy of the prediction, a car was randomly chosen from the csv file, as detailed in Table 5. The predicted price, as shown in Figure 8 is 11307.03. It can be concluded that the prediction works very well.

[4] https://used-car-prices-by-haridv.herokuapp.com/

Conclusions

The purpose of this project was to establish how machine learning can assist buyers and sellers of used cars in the UK to predict prices. Previous, similar, studies were identified and assessed in terms of base data, methodology and results. Three proved to be relevant and informed the design and conduct of this project, for which two data sets were selected and three models applied – linear regression, random forest and XGboost.

The results show that machine learning can assist buyers or sellers to predict used car prices. This is based on a random forest R^2 Score of 99%, an XGboost of 96% and linear regression of 92%. These results compare favourably with those obtained in other studies.

Pal et al. (2018) used a random forest, trained on a Kaggle dataset and obtained high accuracy. The model predicted with an empirical accuracy of 83.63% for testing and 95.82% for training. The sample data included luxury cars (Porsche) and eco-cars and, although small in number, their prices skew the sample. In this project, outliers were removed and lead to a more accurate outcome. The sample data included luxury cars (Porsche) and eco-cars and, although small in number, their prices skew the sample. Outliers were removed from the data used in this project, leading to a more accurate outcome.

Puteri and Safitri (2020) used linear regression to predict used car prices in Indonesia from two producers Toyota and Honda, and analyzed price against car mileage and price against the age of used cars. They obtained a prediction accuracy of only 75%, lower than the results achieved by this project, largely because data cleaning and outlier removal was conducted.

Amik et al., (2021) predicted used car prices in Bangladesh, using XGBoost, which obtained accuracy of 83.63%. This compares to this study's result of r-squared with XGBoost at 96% which reaches 99%, using random forest instead of XGBoost. When comparing the predicted price vs actual price in this study with Amik et al, it was found that the distribution of scatter plots with median prices is too low. In addition, Amik et al, deployed a website on a local machine, compared with the use of Heroku on cloud by this study. It is better to be accessible by cloud publicly rather than a local machine cannot.

However, there is still a constraint. Even though the machine can predict car price very well, close examination and analysis of the data set reveals some anomalies. One example is that the higher the price of the car, the more inaccurate were the predictions, probably because less historic data is available.

Analysis of data used in this project generates the following recommendations for buyers and sellers.

Brand

From Figures 1c and 8, the brand of car that is recommended for the customer is Toyota Figure 8 shows that it has first rank for fuel economy and eco-friendliness Figure 2c, shows that Toyota is in a group with cheap cars. (Motors, 2021) mentions that Toyota has developed a solid reputation for dependability. Toyota stands out for dependability as well as the score is renowned for giving good performance with high mileage.

Age vs mileage decision for consumers

Figure 3 showed when the car is older and the distance driven is increasing, that the price tends to go down every year.

Cooper (2021) compared two types of cars - old car, lower mileage and newer car, higher mileage. The latter affects the price because of higher maintenance costs. However, old car, low mileage, low technology and low-security quality decreases the price and repairs may be less.

In conclusion, people pay more attention to newer cars than low mileage cars because Figure 4 showed that car age affects price more than mileage. The age of car correlation is -0.88 and mileage is -0.81 respectively.

Future work

Motors (2021) compared Petrol vs Diesel vs Hybrids vs Electric. This study is Petrol vs Diesel vs Hybrids. Electric cars could not be included as the number of observations was too low. In future studies, it will be necessary to include electric cars. In addition, it is recommended that advanced techniques, such as fuzzy logic, should be deployed to predict used car prices.

Acknowledgement

The authors wish to express their gratitude to my supervisor, Dr. Marika Asgari, for guidance. Additionally, the anonymous reviewers offered a number of incredibly helpful recommendations that significantly improved the paper. The recommendations are much appreciated by the authors.

References

1. Allwright, S. (2022) RMSE vs MAPE, which is the best regression metric?, Stephen Allwright. Available at: https://stephenallwright.com/rmse-vs-mape/(Accessed: August 7, 2022).

2. Amik, F. R. et al. (2021) "Application of machine learning techniques to predict the price of pre-owned cars in Bangladesh," Information (Basel), 12(12), p. 514.doi: 10.3390/info12120514.

3. Ballings, M., Van den Poel, D., Hespeels, N. & Gryp, R. (2015) Evaluating multiple classifiers for stock price direction prediction. Expert Systems with Applications, 42(20), 7046-7056.

4. Brownlee, J. (2020) How to train to the test set in machine learning, Machine Learning Mastery. Available at: https://machinelearningmastery.com/train-to-the-test-set-in-machine-learning/ (Accessed: August 18, 2022).

5. Cooper, J. (2021) *Should you buy an older car with lower mileage or a newer car with higher mileage?*, *Totallymotor.co.uk*. Available at: https://totallymotor.co.uk/should-you-buy-an-older-car-with-lower-mileage-or-a-newer-car-with-higher-mileage/(Accessed: July 29, 2022).

6. DataTechNotes (2019) DataTechNotes, Datatechnotes.com. Available at: https://www.datatechnotes.com/2019/02/regression-model-accuracy-mae-mse-rmse.html (Accessed: August 7, 2022).

7. Dudoit,S. et al. (2002) Comparison of discrimination methods for the classification oftumors using gene expression data. J. Am. Stat. Assoc.,97, 77–87.

8. Fernando, J. (2003) R-Squared, Investopedia. Available at: https://www.investopedia.com/terms/r/r-squared.asp(Accessed: August 7, 2022).

9. Flask (2022) Fullstackpython.com. Available at: https://www.fullstackpython.com/flask.html (Accessed: August 7, 2022).

10. Garg, S. (2022) How to deal with categorical data for machine learning, KDnuggets. Available at: https://www.kdnuggets.com/2021/05/deal-with-categorical-data-machine-learning.html (Accessed: August 18, 2022).

11. GeeksforGeeks (2020) XGBoost for regression. Available at:https://www.geeksforgeeks.org/xgboost-for-regression/(Accessed: August 6, 2022).

12. Gupta, A. (2021) XGBoost versus Random Forest - geek culture - medium, Geek Culture. Available at: https://medium.com/geekculture/xgboost-versus-random-forest-898e42870f30(Accessed: August 7, 2022).

13. Gupta, D. K. (2018) Mathematics for machine learning : Linear regression & least square regression Available at: https://towardsdatascience.com/mathematics-for-machine-learning-linear-regression-least-square-regression-de09cf53757c (Accessed: April 19, 2022).

14. Heroku (no date) About heroku, Heroku.com. Available at: https://www.heroku.com/about(Accessed: August 6, 2022).

15. Insurancefactory (2020) How the Car Buying Process Has changed in the last 30 years. Available at: https://www.insurancefactory.co.uk/news/January-2015/car-buying-changed-last-30-years (Accessed: August 18, 2022).

16. Kirenz, J. (2021) Linear regression diagnostics in Python, Jan Kirenz. Available at: https://www.kirenz.com/post/2021-11-14-linear-regression-diagnostics-in-python/linear-regression-diagnostics-in-python/ (Accessed: August 7, 2022).

17. Mwiti, D. (2020) Random Forest Regression: When does it fail and why?, neptune.ai. Available at: https://neptune.ai/blog/random-forest-regression-when-does-it-fail-and-why(Accessed: August 6, 2022).

18. Nettleton, D. (2014) "Selection of variables and factor derivation," in Commercial Data Mining. Elsevier, pp. 79–104.

19. Pal, N. *et al.* (2018) "How much is my car worth? A methodology for predicting used cars' prices using random forest," in *Advances in Intelligent Systems and Computing*. Cham: Springer International Publishing, pp. 413–422.

20. Puteri, C. K. and Safitri, L. N. (2020) "Analysis of linear regression on used car sales in Indonesia," Journal of physics. Conference series, 1469(1), p. 012143. doi: 10.1088/1742-6596/1469/1/012143. (Accessed: April 14, 2022).

21. Varghese, D. (2018) Comparative Study on Classic Machine learning Algorithms, Towards Data Science. Available at: https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222 (Accessed: August 7, 2022).