BMW Price Prediction Analysis

Areka Raza, 501071674

CIND820: Big Data Analytics Project

Instructor: Ceni Babaoglu

April 04, 2022

Github Repository: https://github.com/ArekaZR/CIND820.git

Abstract

In this paper, I investigated the application of supervised machine learning techniques to predict the price of pre-owned BMW vehicles. The dataset was retrieved from Kaggle, and consists of fictional, along with non-fictional values. Various techniques such as multiple linear regression, random forest regression, decision tree regression and cross fold validation were utilized. The models were then compared to retrieve the one with the highest accuracy. With each model, there was a slight issue with achieving a high accuracy, which is further outlined in the future work section.

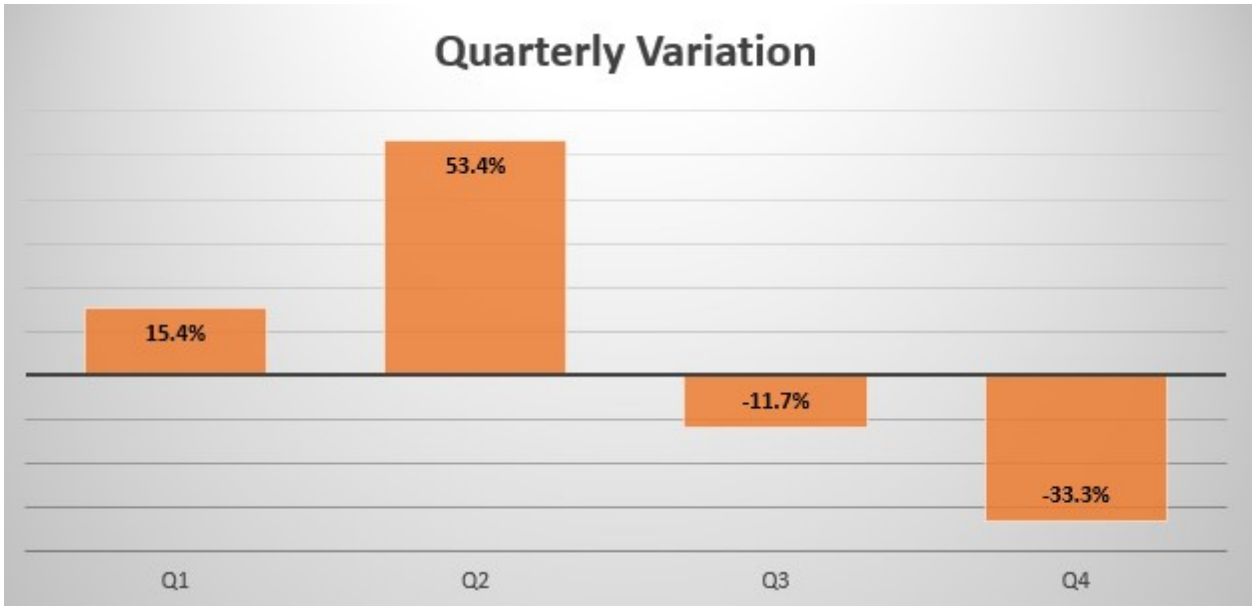
Introduction

The SARS-CoV-2 pandemic has led the entire world through changes one would not have deemed possible. The world is still recovering from its effects. A notable issue that was created due to SARS-CoV-2, was the microchip shortage. The ramifications of the microchip shortage has led to a decline in electronics and vehicles productions to name a few. Due to a shortage in supply, the demand has increased worldwide, causing an uptick in prices.

At the beginning of 2021, car sales were soaring, and crashed in the second half of the year as there were no available new cars to sell. In the first two quarters, sales had increased by 68.8% in comparison to 2020, however in quarters 3 and 4, sales had decreased by 45%.

Figure 1

Quarterly Variation



Note: Information retrieved from article published by Focus2Move

Due to the aforementioned decrease in supply, in July 2021, the average used car was being listed at a 12.8% markup, in comparison to July 2020 (Yun, 2021). It was reported by Global News that "pickup truck prices had increased at least 20 per cent" (*Canadian Used Car Sales Rose 5% in 2021 Amid Semiconductor Shortages - National | Globalnews.ca*, 2022). I noticed the increase in aforementioned prices while I was making the decision of whether or not I should purchase a used car. Observing prices of used cars soar intrigued me to learn what influences the pricing of used vehicles, BMW in specific as I work there.

When determining the price of a new vehicle, most individuals know what the dependent variables are, such as brand name, engine power and model type. However, unfamiliarity arises when establishing the price of a pre-owned vehicle. More factors come into light such as total mileage, condition of the vehicle's exterior, condition of a vehicle's interior, how many accidents the vehicle has been involved in and what the model year is, for example. "As a consequence,

customers who plan to purchase a pre-owned car often struggle to find an appropriate car within a budget. Even if a customer knows the type of car they want to purchase, it becomes challenging for them to estimate the price of the car" (Rahman et al., 2021). Not everyone has hours to spend researching which factor decreases or increases cost, and by how much. There are simply too many to consider, this presents a dire need to depend on a model that would predict the price and account for all the variables and factors involved.

Machine Learning for price prediction is beneficial for many reasons, notably 4, as stated by Rojewska in 2021:

1. Machine learning can cope with price volatility
2. Machine learning models can analyze multiple data sources at once
3. Machine learning improves the accuracy of price predictions
4. Machine learning can help you improve your profit margin

The aforementioned also make the automotive industry one that benefits majorly from price prediction (Rojewska, 2021).

The research questions that will be evaluated are:

1. Which variables will be the most strongly correlated with price?
2. Will a dataset with less than 5000 entries be sufficient?
3. What effect does engine power have on price?

The hypothesis is that mileage and engine power will have the strongest correlation with price. Some additional hypotheses are that there will be constraints with this dataset, and features that are unexpected like color may have a large impact on price.

Related Work

Similar studies have been conducted in the past whereby machine learning has been used to predict pricing of used vehicles. Pudaruth conducted a study in 2014 based on used vehicles in Mauritius. The author applied multiple linear regression, k nearest neighbors, naive bayes and decision trees for the predictive analysis. However, this study only consisted of 97 records and posed as a limitation, the sample size was small and only included car makes: Honda, Nissan and Toyota. Pudaruth found that the weakness discovered in decision trees and naive bayes was "their inability to handle output classes with numerical values" (Pudaruth, 2014). Another similar study was conducted by Babu et al., in 2021, they utilized linear regression, random forest algorithm and decision trees. The authors faced similar data issues during their analysis. They found the dataset was not large enough, they did not have many variables and also had missing values (Babu et al., 2021). In addition to this, Kiran S conducted a similar study in 2020 using only a linear regression model. The author predicted an accuracy of 90% with an error of 10% (S., 2020). A research paper by Ashok, K and Samrudhhi, K, focused on K-Nearest neighbor to build a model. The authors obtained an accuracy of 85%, and also validated the model with 5 along with 10 folds (Samruddhi & Kumar, 2020). And lastly the research paper titled "Vehicle Price Prediction System using Machine Learning Techniques" by Kanwal Noor and Sadaqat Jan in 2017 focused only on linear regression to create their model and selected only the highly

influential features, and removed the remainder of the features. They created a model with prediction precision of 98% (Noor & Jan, 2017).

Methodology

The purpose of this research is to utilize the aforementioned studies for reference while building a model that has higher accuracy and depicts a clearer picture when predicting the price of pre-owned BMW vehicles. I aim to apply the learnings into my research. The dataset was retrieved from Kaggle which consists of mostly verified vehicle sales in a business to business auction in 2018. Some initial data cleaning has already been conducted, such as removal of vehicles that had engine issues and would skew the results.

The dataset I will be using has 4843 observations, which highlights that records are at an average sample size. If it were too large, it would be difficult to conduct analysis on, and if it were minimal then results would not be depicting an accurate image. Furthermore, I will perform logistic regression to predict price using its attributes and will be evaluating the model performance. Furthermore, I will be using a train/test approach, along with a 10 fold cross validation approach to understand which one yields better results. Thus, my research will be different from studies done in the past as it will take the limitations of those to create an enhanced model, and will be using a mixture of all techniques utilized.

Figure 2

Descriptive Statistics

```
> sum(is.na(data))
[1] 0
```

```
> summary(data)
  maker_key          model_key           mileage
 Length:4843        Length:4843        Min.   :    -64
 Class :character   Class :character   1st Qu.: 102914
 Mode  :character   Mode  :character   Median : 141080
                                       Mean   : 140963
                                       3rd Qu.: 175196
                                       Max.   :1000376
  engine_power registration_date        fuel
 Min.   :  0   Min.   :1990-03-01   Length:4843
 1st Qu.:100   1st Qu.:2012-07-01   Class :character
 Median :120   Median :2013-07-01   Mode  :character
 Mean   :129   Mean   :2012-11-22
 3rd Qu.:135   3rd Qu.:2014-04-01
 Max.   :423   Max.   :2017-11-01
 paint_color          car_type         feature_1
 Length:4843        Length:4843        Mode :logical
 Class :character   Class :character   FALSE:2181
 Mode  :character   Mode  :character   TRUE :2662


 feature_2       feature_3       feature_4       feature_5
 Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:1004      FALSE:3865      FALSE:3881      FALSE:2613
 TRUE :3839      TRUE :978       TRUE :962       TRUE :2230


 feature_6       feature_7       feature_8           price
 Mode :logical   Mode :logical   Mode :logical   Min.   :   100
 FALSE:3674      FALSE:329       FALSE:2223      1st Qu.: 10800
 TRUE :1169      TRUE :4514      TRUE :2620      Median : 14200
                                                 Mean   : 15828
                                                 3rd Qu.: 18600
                                                 Max.   :178500

    sold_at
 Min.   :2018-01-01
 1st Qu.:2018-03-01
 Median :2018-05-01
 Mean   :2018-04-29
 3rd Qu.:2018-07-01
 Max.   :2018-09-01
```

There are 18 variables in the dataset, and none have missing values. Some data seems to be incorrect, for example the minimum mileage being listed as -64 and minimum engine power being listed as 0. Both are not possible, thus were amended. The mileage was changed to 64 and engine power was changed to the mean of the column which was approximately 128.98.

Furthermore, there were 8 unnamed features which were renamed to the following, this was done to better identify which features cause price to be affected:

1. Has Moon Roof
2. Has Leather Seats
3. Has Heated Seats
4. Has Navigation System
5. Has Bluetooth
6. Has Remote Start
7. Has Blindspot Monitoring
8. Has MSport Package

The aforementioned features were also changed from boolean to int to include them in my analysis. The maker key column was removed from the analysis as it only had one value for all rows, and would not have any contribution to my review. In addition to this, a new column titled "age" was created to better understand the current lifespan of each vehicle in days. To create this column, the columns "registration date" and "sold at" were converted to datetime. Furthermore, the columns fuel type, car type and paint color were converted to numbers to further include additional columns into the analysis, thus providing a holistic understanding of the dataset.

The numeric values will be normalized according to the dataset, thus standardizing the attributes selected. This will ensure that the attributes with larger scales will no longer affect other attributes with values of smaller scales. Price was then dropped from the dataframe in order to isolate it for prediction. The dataset was then divided into training and test sets, and random state

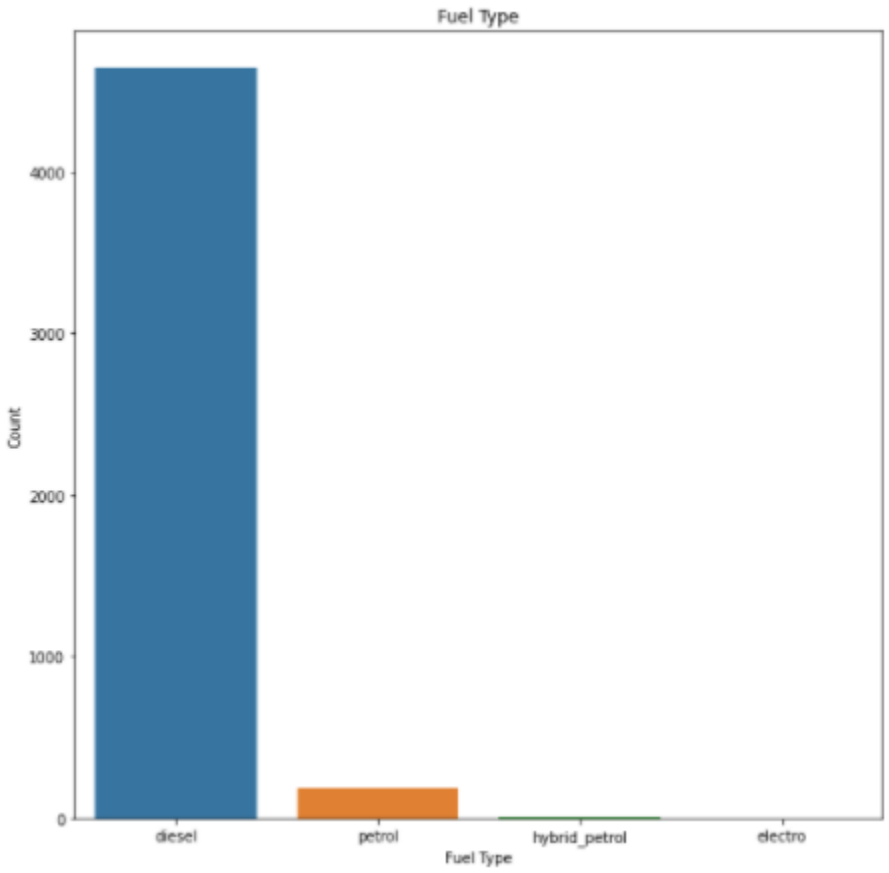
was included to ensure that results are reproducible. Multiple linear regression was then conducted and a specific instance for prediction was used, which will be discussed in the results and findings section. Random forest regression was then conducted, and a tree count along with the random data count was determined. After this, a decision tree regressor was also introduced. And lastly, 10 fold cross validation was conducted.

Results and Findings

Fuel Type

Figure 3

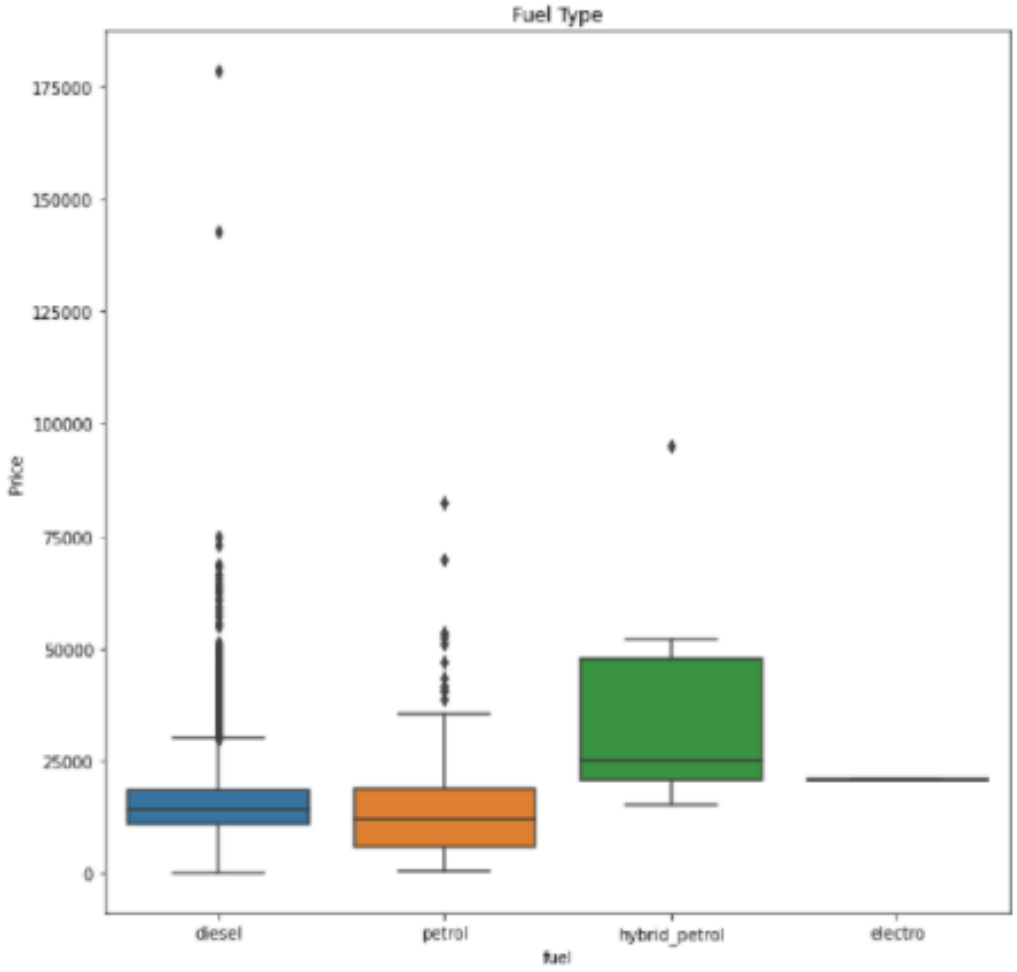
Bar chart signifying the volume of each fuel type of vehicles in the dataset.



In this bar chart, we see that the most common fuel type in this dataset was diesel, followed by petrol. Although, petrol was nowhere near diesel. In this case, diesel was over indexing.

Figure 4

Boxplot of each fuel type in comparison to price.

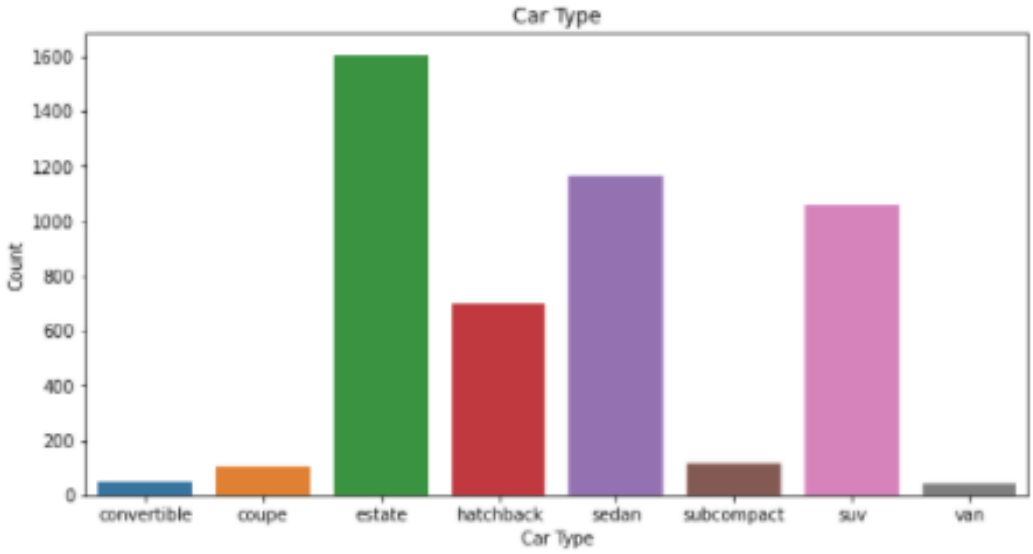


Hybrid Petrol, although was not identified as having a large volume in this dataset, it was the most expensive type of vehicle. Its minimum, maximum and median was a lot higher than diesel and petrol. It should also be noted that although an expected outlier, a vehicle with diesel as fuel type was the most expensive in this dataset, along with the second most expensive too.

Car Type

Figure 5

Bar charts illustrating the car type in this dataset.



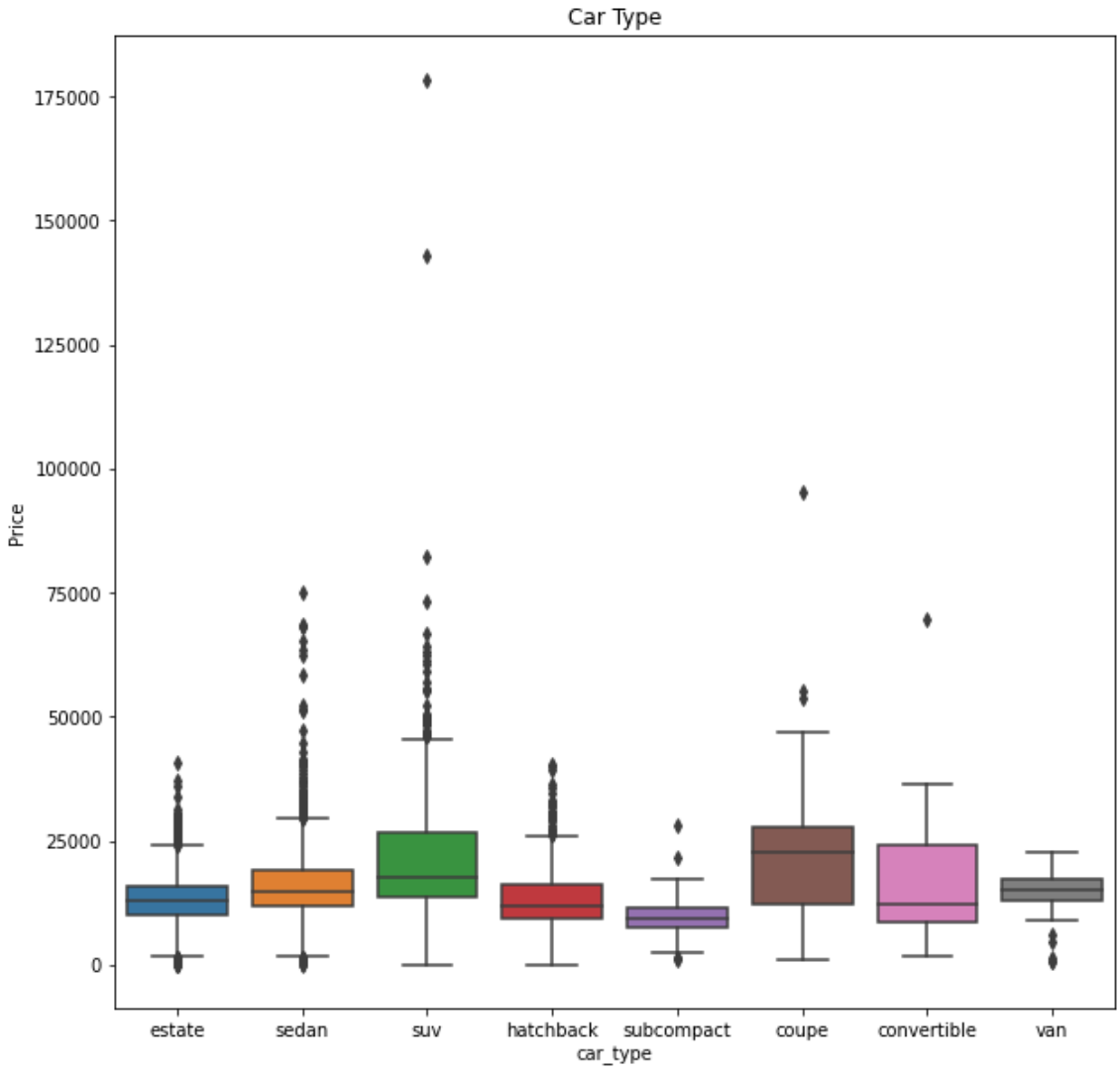
In this dataset, there were 8 different car types:

1. Convertible
2. Coupe
3. Estate
4. Hatchback
5. Sedan
6. Subcompact
7. SUV
8. Van

The largest volume for car type in the dataset was for estate vehicles, at approximately 1500 vehicles. This was followed by sedans, as expected. And closely followed by SUV's. In fourth place was hatchbacks, fifth place was subcompacts, sixth place was coupes, followed by vans, and lastly by convertible.

Figure 6

A boxplot of car type in comparison to price



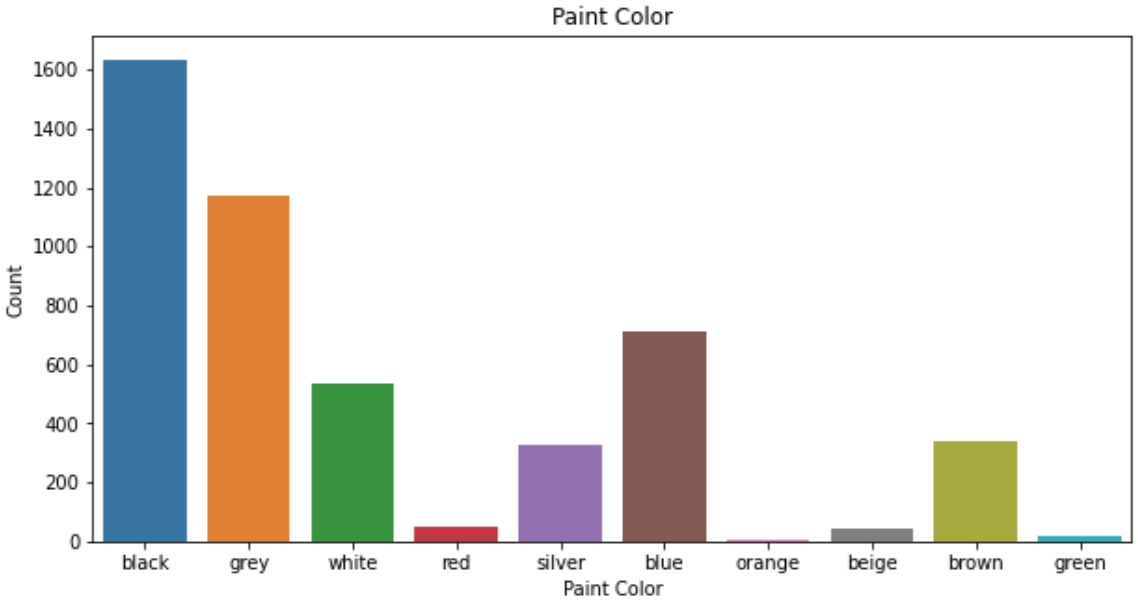
In this boxplot, we see that SUV has the highest maximum value, along with outliers too. Seeming as one of the most expensive vehicles in this dataset. When looking at the median

value, we see that coupe is the highest, and is a close tie with SUV in terms of maximum value, along with Q3 of the values for that attribute. The minimum for vans is higher than any other car type in this dataset. Furthermore, another finding was that although an outlier, SUV was responsible for the two most expensive vehicles in this dataset. This could be due to a number of reasons, such as the age of the vehicle, or the features it has.

Paint Color

Figure 7

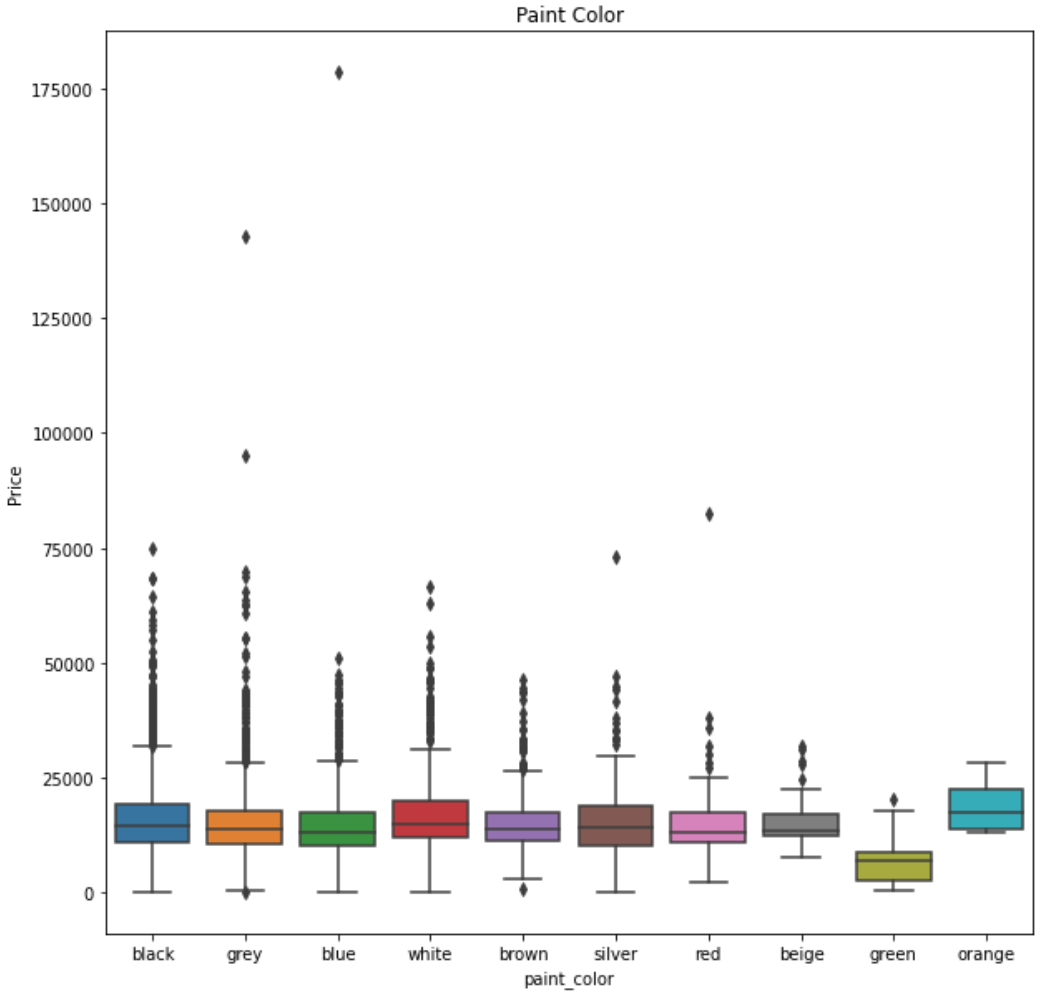
A bar chart of all paint colors available in this dataset, and their counts.



There were 10 colors listed in this dataset. The most frequently occurring color was black, followed by grey, then blue, at number four was white, at number five was brown, at number six was silver, followed by beige, then red, then green, and lastly orange.

Figure 8

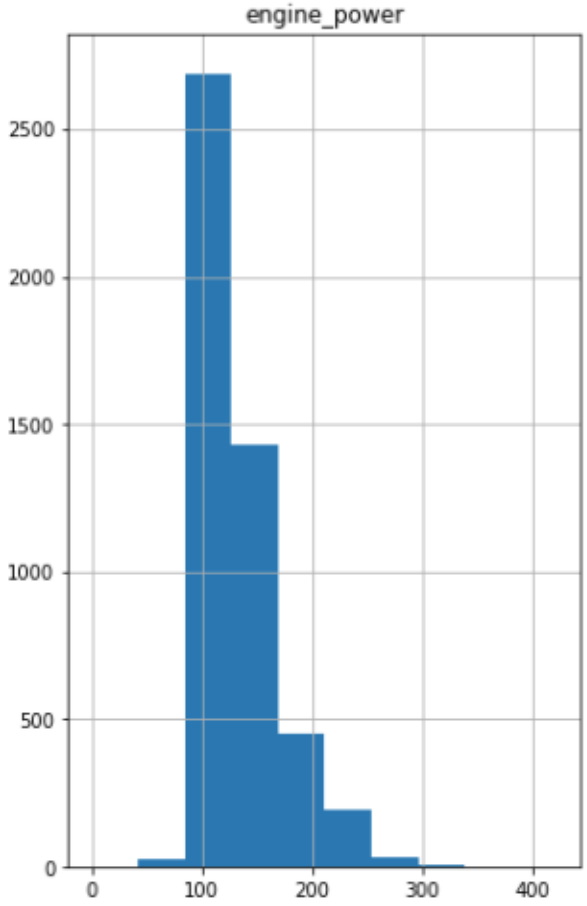
A boxplot of paint colors against price.



Grey, blue, brown and silver seem to be very similar in terms of price. Orange was an interesting color as its minimum is very high in comparison to all of the other colors, along with its median. Green seems to be the cheapest option in comparison to the rest of the colors as its maximum is below all others. The most expensive color in this dataset was for a vehicle that was blue.

Figure 9

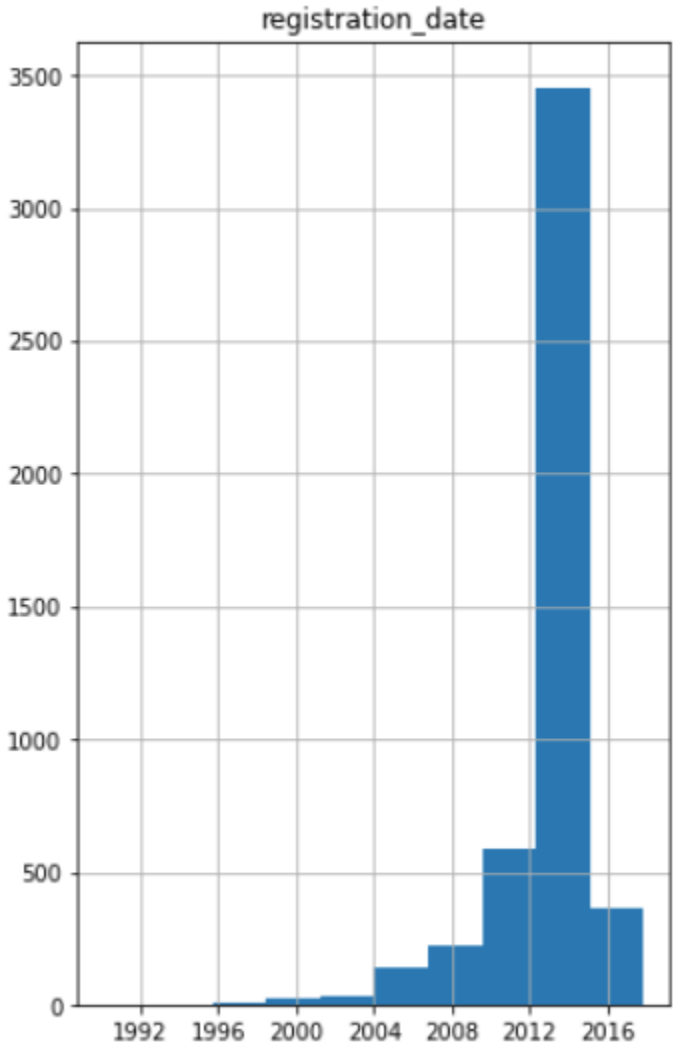
A histogram of engine power



The mean of this column is 128.98 HP, with a minimum of 25, a maximum of 423 and standard deviation of 38.9. The most common engine power in this dataset is within the range of 100-120HP.

Figure 10

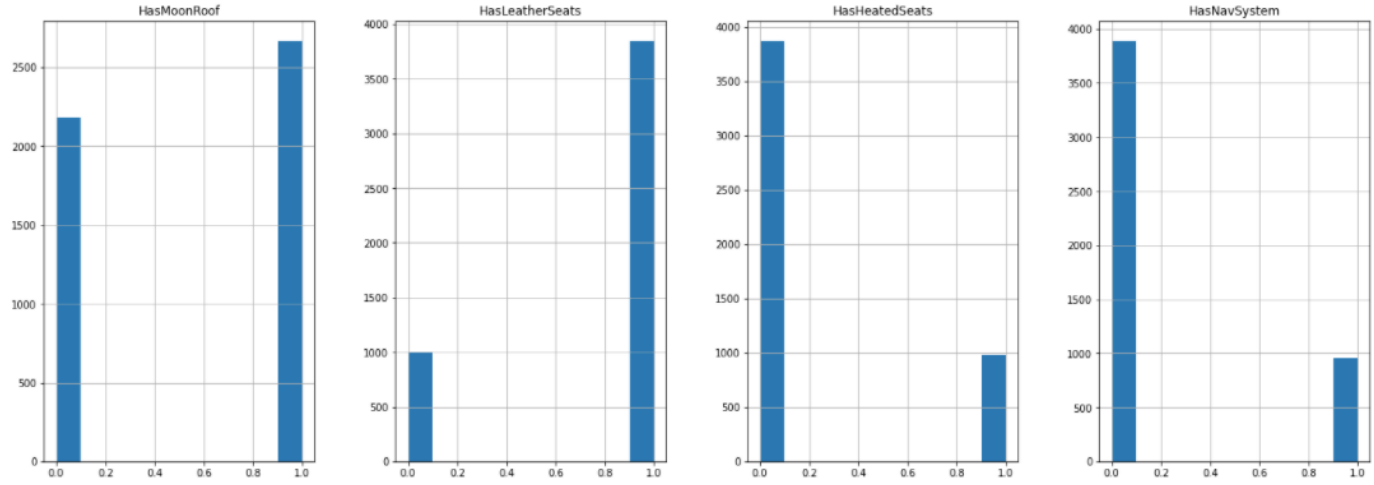
A histogram of registration date



This column refers to the date the car was registered, or initially sold. We see most of the occurrences are between 2012-2014. And very minimal data from 2000 and prior to that. It also seems that there was a steep decline after 2016.

Figure 11

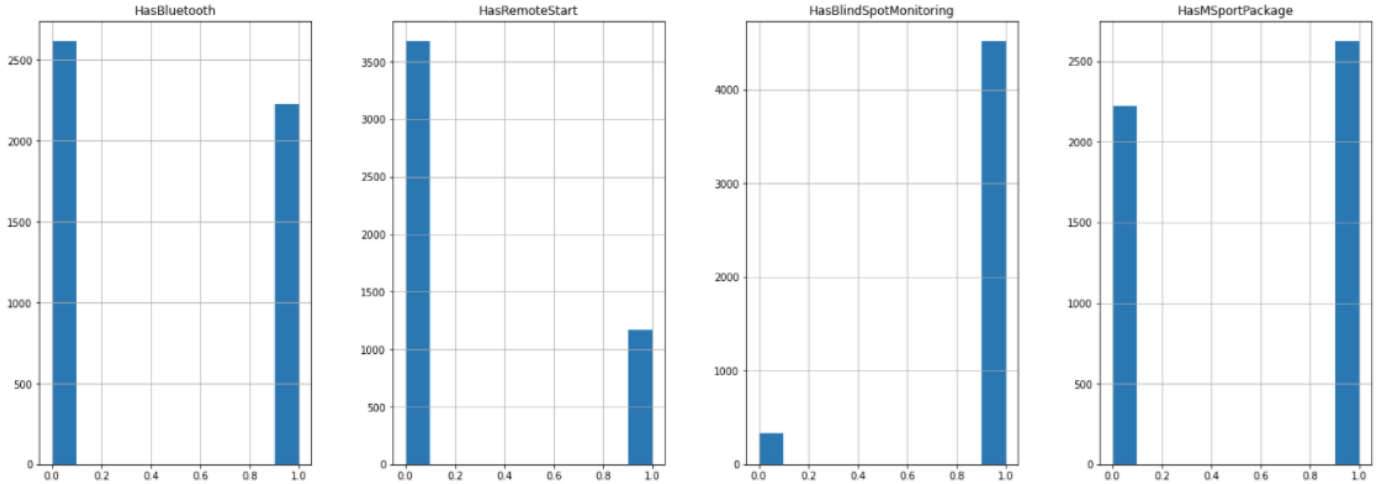
Evaluating histograms for renamed features



There were more vehicles that had the moon roof feature, than not. This was similar with vehicles consisting of leather seats. However, with heated seats and navigation systems, more vehicles did not have them in this data set.

Figure 12

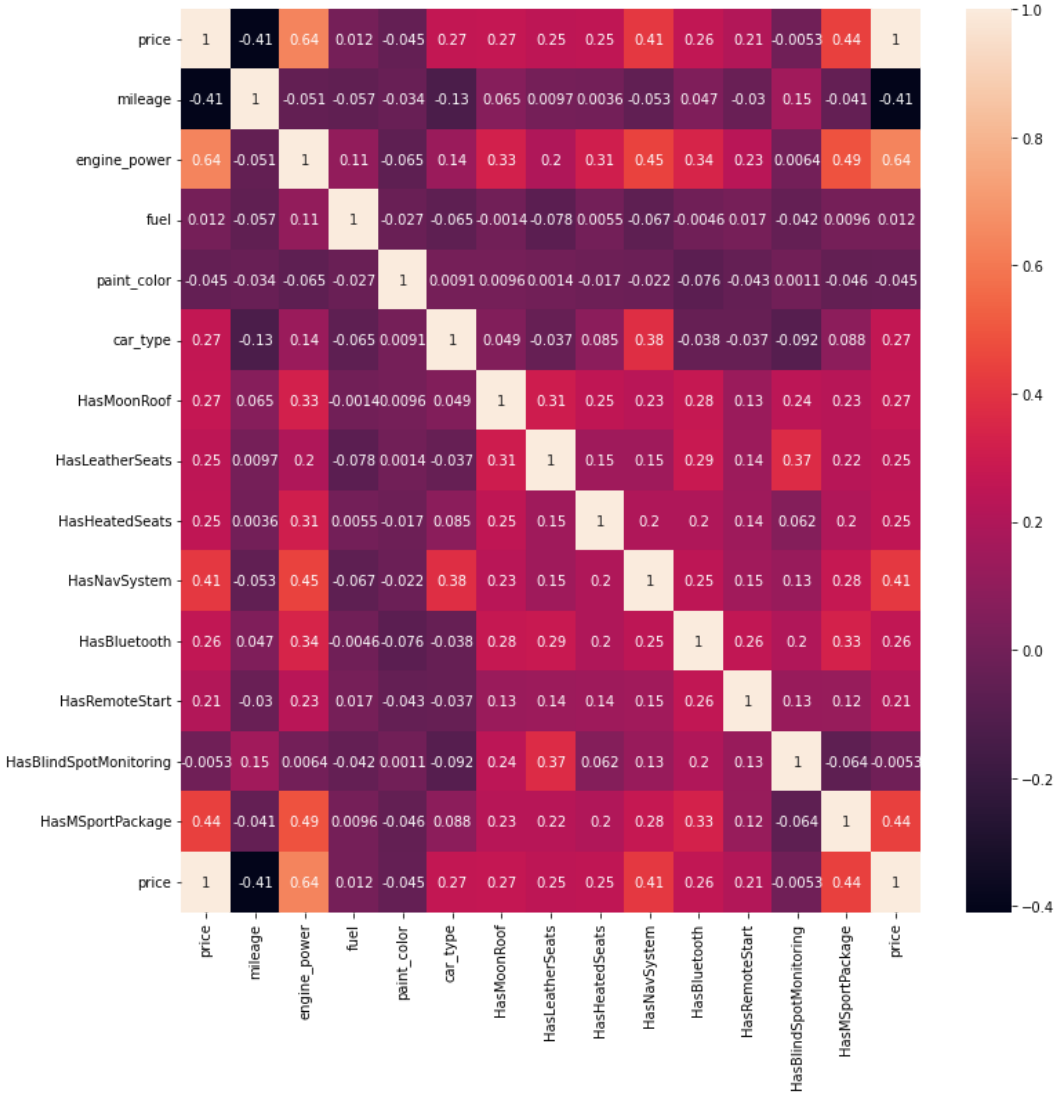
Evaluating histograms for renamed features continued



Looking at the remaining four features, we notice that there are more vehicles that do not have bluetooth, than do in this dataset. Although, it is quite close with vehicles that do have the bluetooth feature. As for remote start, there are more cars that do not have this option, which is understandable as this feature was introduced later than most of the registration dates. In addition to this, almost all vehicles have blind spot monitoring, than not in this dataset. The amount of vehicles that have this option is over indexing those that do not. And lastly, there are more vehicles that have the MSport package.

Figure 13

Evaluating correlation between each feature against price



A correlation analysis was conducted to analyze which features and attributes contribute to price the most. Each feature fairly has a strong correlation with price. As expected, mileage and engine power are highly correlated features. Some interesting ones to note are- has a navigation system and has an MSport package.

Figure 14

Histogram illustrating the variation in prices and their frequencies

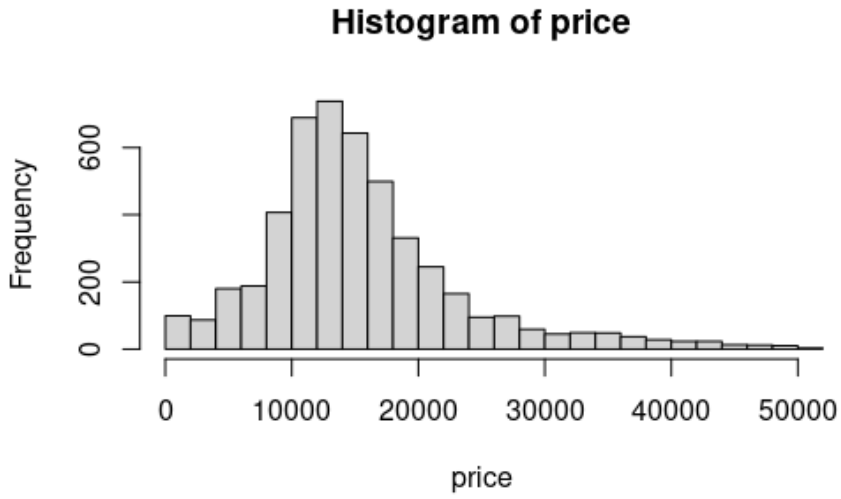
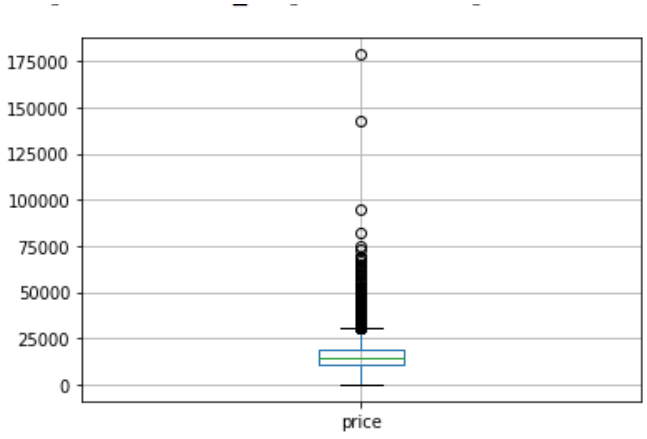


Figure 15

Boxplot of price



The independent variable in this research is price. And price is ranging from $100 to $17,8500 in the dataset, with mean being at $15,828. All of the listed variables clearly affect the price of the used vehicle, and a correlation analysis was conducted to understand which dependent variables are highly correlated with price. The hypothesis was that mileage and engine power will have the strongest correlation with price. This was proven to be true by the initial corr plot that was coded.

Model 1. Multiple Linear Regression

The dataset was normalized and divided into training as well as test sets. Training examples were 968, and testing examples were 3875. Multiple linear regression was used to visualize the dependent variables and display the linear relationship.

Figure 16

Initial Results of Multiple Linear Regression

```
print('Intercept: \n', LR.intercept_)
print('Coefficients: \n', LR.coef_)
```

```
Intercept:
 0.01777155673560478
Coefficients:
 [-0.38299976 -0.02469446  0.0914665  -0.0355812   0.50995183  0.0491526
  0.08185497  0.04989617  0.09284903 -0.04117361  0.03563168  0.02690171
  0.14397569]
```

```
print("R Square Value :",LR.score(x_test,y_test))
```

```
R Square Value : 0.5844520221195526
```

The intercept was 0.01777 with the above coefficients. In this case, the most correlated variables were mileage, car type and age. The coefficients stated names were in the following order:

1. Mileage
2. Engine power
3. Fuel
4. Paint color
5. Car type
6. Has moon roof
7. Has leather seats
8. Has heated seats
9. Has navigation system
10. Has bluetooth system
11. Has remote start
12. Has blind spot monitoring
13. Has MSport package
14. Age

After conducting the prediction of certain attributes, i found that the vehicle price was $827.73. The Prediction that was done was for the price of a car that has a mileage of 500, 1400 HP, uses petrol as fuel, is silver, is a sedan, has a moon roof, has leather seats, has bluetooth, has blind spot monitoring and is 7300 days old. This model had an accuracy of 58%, which was not as desirable.

Model 1b. Random Forest Regressor

This was done to fit a number of classifying decision trees on various sub-samples of the dataset and utilized the average to improve the prediction accuracy along with controlling over-fitting. The accuracy received on this was 64.5%, which was the highest percentage out of all models conducted. As we know, the closer the model accuracy is to 1, the better it is. In this case, it was a desirable outcome.

Model 1c. Decision Tree Regression

This model was conducting to break down the dataset into smaller subsets and decision trees were incrementally developed. This model returned an accuracy of 44.7%, which was the lowest for all models, and thus this was removed from the analysis.

Model 2. K Fold Cross Validation

The cross-validation method was defined and was used to building the multiple linear regression model. At the end, k-fold cross validation was used to evaluate model and viewing the mean absolute error. The mean absolute error (MAE) was ~6.87e-16. This is the average absolute error between my model prediction and the actual observed data. We know that the lower the MAE, the more closely a model is able to predict the actual observations. In this case, the MAE is very miniscule.

Conclusion

The most accurate model that was built was the random forest regressor, with an accuracy of 64.5%. We found that the most strongly correlated variables with price were mileage and engine

power, along with others. The age of the vehicle being a strongly correlated feature was expected, however car type was a surprising one. There were constraints with the size and information of the dataset.

Future Work

In the future, I would either find a dataset with more rows, or add to this dataset as there were a number of outliers, which in reality would not be counted as outliers since they depict accurate examples of vehicle features/models these days. Furthermore, I would aim to find a dataset that has the features listed for what they are, rather than renaming them as that is not accurate. In addition to this, I would ensure the dataset is entirely non-fictional, and each value is a real world example. And lastly, I would aim to create a model with higher accuracy- at least near the 75% percentile range, this may be achieved by not overfitting my models.

References

Babu, S. K., Sk., R., N, N., M, N., Kumari, L. K., & B, L. (2021). *VEHICLE RESALE PRICE PREDICTION USING MACHINE LEARNING*. Juni Khyat journal. Retrieved February 14, 2022, from http://junikhyatjournal.in/no_1_Online_21/68.pdf

Canadian used car sales rose 5% in 2021 amid semiconductor shortages - National | Globalnews.ca. (2022, February 9). Global News. Retrieved February 12, 2022, from https://globalnews.ca/news/8606989/canada-used-car-sales-rose-2021/

Focus2move| Canada Auto Sales - Facts & Data 2022. (2022, January 24). Focus2Move. Retrieved February 14, 2022, from https://www.focus2move.com/canada-auto-sales/

Noor, K., & Jan, S. (2017). *Vehicle Price Prediction System using Machine Learning Techniques*. International Journal of Computer Applications. Retrieved February 14, 2022, from https://www.ijcaonline.org/archives/volume167/number9/noor-2017-ijca-914373.pdf

Pudaruth, S. (2014). *Predicting the Price of Used Cars using Machine Learning Techniques*. Research India Publications. Retrieved February 14, 2022, from http://ripublication.com/irph/ijict_spl/ijictv4n7spl_17.pdf

Rahman, F., Lanard, A., & Ismat, A. (2021). *Application of Machine Learning Techniques to Predict the Price of Pre-Owned Cars in Bangladesh*. MDPI. Retrieved February 12, 2022, from https://www.mdpi.com/2078-2489/12/12/514/htm

Rojewska, K. (2021, September 15). *Price Prediction: How Machine Learning Can Help You Grow Your Sales*. DLabs.AI. Retrieved February 12, 2022, from https://dlabs.ai/blog/price-prediction-how-machine-learning-can-help-you-grow-your-sales/

S., K. (2020, July). *Prediction of Resale Value of the Car Using Linear Regression Algorithm*. https://ijisrt.com/assets/upload/files/IJISRT20JUL388.pdf

Samruddhi, K., & Kumar, A. (2020, September). *Used Car Price Prediction using K-Nearest Neighbor Based Model*. IJIRASE. Retrieved February 14, 2022, from https://www.ijirase.com/assets/paper/issue_1/volume_4/V4-Issue-3-686-689.pdf

Used car prices in Canada up 12.8 per cent from last year as microchip shortage continues. (2021, August 12). CTV News. Retrieved February 14, 2022, from https://www.ctvnews.ca/autos/used-car-prices-in-canada-up-12-8-per-cent-from-last-year-as-microchip-shortage-continues-1.5545158