UNIVERSITY OF BRITISH COLUMBIA

STATISTICS 447 COMPUTING IN STATISTICS

Used Car Prices in India

Kexin Chen, Chin Chen Lo, Yuxin Mao, Sisi Yang

Abstract

Background

The Indian used cars market is the fifth largest in the world. During the pandemic of coronavirus, there is an increasing number of people who prefer travelling by personal vehicle. Besides, used cars are usually at lower prices compared with completely new cars with the same configuration. All of these result in constant growth in demand in the used car market in India. However, it still contains great uncertainties in pricing for used cars without a preset standard of price, given the reality that all car sellers want the best resale prices for their cars. Currently, if any of them wants to estimate a car's price, they have to take their car to a respective company workshop or have to make an appointment for the company to get an estimate of the price, which will lead to a very time-consuming and resource-wasting process. Therefore, our purpose is to build a model to predict used cars price in India as accurately as possible based on given conditions. This model will not only save a lot of time for consumers but also help the company to reduce costs and streamline the process of selling used cars.

Methodology

This report presents several implementations of machine learning techniques to predict used car prices in India. Linear transformations, imputation methods, and feature selection are vastly used in the data pre-processing stages while simple OLS regression and tree algorithms are executed assisted by cross-validation. Highlights of this report include extensive manipulation of categorical variables and the creation of a new factor variable. Firstly, after the initial implementations of our models, further steps were needed to group categorical variables to ensure proper training and cross-validation fold splits. Secondly, a new variable *BodyType* was manually implemented to ensure algorithms with more sophisticated features and better predictive ability. As a result, our metrics such as the RMSE, average length of prediction intervals, interval scores or quantile random forests and coverage rates prove to be adequate in our summary metrics.

Results

A total of n=6002 observations were used for algorithm training with k=3 and k=5 cross-validation folds. Focusing on prediction intervals, interval scores, coverage rates, and root-mean-squared errors (RMSE) for evaluation, regression models, regression trees, random forests, and quantile regression forests were well-performed. The average lengths of prediction intervals (*avgleng*), interval scores *IS* were similar for the models. While coverage rates are approximate to the intended level. However, results prove that the quantile regression forest consistently performs better than the regression model and regression trees for a differed number of folds as it has smaller interval scores (i.e. less penalty).

Conclusions

Our summary statistics show that the price of used cars in India depends on a large variety of factors. With the diverse selection of used cars containing different features, our dataset suggests that the used car prices are particularly dependent on the car's horsepower, age, and brand name. However, in reason to the variety of configurations used cars can take, it is uncertain whether our algorithms have high predictive probability to observations outside of our dataset.

1 The Dataset

This dataset was downloaded from Kaggle, named "Used Cars Price Prediction". Initially, there were 6019 rows and 13 columns in our original dataset. These 13 columns included 1 response variable and 12 explanatory variables. We had deleted the variables "Location" and "New Price", since locations in which the cars were being sold or were available for purchase were all based in India, and the values of the majority of "new_price" were missing. In addition, we had deleted 3 rows in Excel directly due to missing values. We then have 6016 rows and 13 columns in our dataset.

1.1 Response Variable: Price

The response variable is the price of the used cars in INR Lakhs. We converted the currency unit from INR Lakhs to CAD using an exchange rate of 1 INR Lakh equals 1666.49 CAD (as of Mar 7, 2022). This conversion will make it easier for interpretation.

Summary Statistics	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD
Price (CAD)	733.3	5832.7	9390.7	15794.9	16581.6	266638.4	18648.8

Table 1: Summary Statistics for Price in CAD

2 Explanatory Variables

For detailed summaries of the explanatory variables, please refer to the appendix for tables and summaries.

Continuous Variables

- Kilometers Driven: Refers to the total kilometers driven in the car by the previous owner(s) in km. The kilometers driven reflect the extent the car was used. With capital depreciation in consideration, a car with higher kilometers driven would imply a lower selling price.

- Mileage: The standard mileage offered by the car company in kmpl or km/kg. With a higher mileage, vehicles are able to travel further with the same amount of fuel. Since the majority of the values of mileage in our dataset are in kmpl, we have converted the values in km/kg to kmpl directly in Excel using the scale 1 km/kg = 1.40 kmpl.

- Engine: The displacement volume of the engine in cc. A larger value of engine means that the car engine is able to produce higher power. Until fairly recently, car model designations often referred to the engine size–the bigger the number, the more expensive it usually costs the car.

- Power: The maximum power of the engine in bhp. Generally speaking, the more power a car produces, the better its acceleration, which is a strong factor in its overall performance. Initially, there was uncertainty behind the credibility of the variable as larger horsepower did not necessarily translate to better or preferred.

[1]

[1] Correlation between the quantitative explanatory variables was examined through the spearman matrix. Results show that LogEngine and SqrtPower may be highly correlated. Please refer to the appendix.

Categorical Variables

- Name: Refers to the brand of the car. The original dataset from Kaggle contained the "Full Name" of the used vehicle, containing the brand name, series, style, and perhaps even the trim (e.g.: Audi A8 L 3.0 TDI Quattro). However, due to a large number of unique classes for the original variable "Full Name", we extracted only the brand name at the initial cleaning stage. There are a total of 30 brand names of vehicles in our dataset. We extracted these 30 brand names since there will be a great number of models to account for without this extraction. Generally, cars that are known for high quality or prestige are more expensive. There is also a great difference between luxury and economy cars. The prices of luxury cars are much higher than economy ones even if these vehicles' configurations are exactly the same.

- Body Type: A factor variable we implemented to improve the predictability of our model. As mentioned in the previous bullet point, the original dataset provided a column containing a full description of the vehicle's series and style. However, due to the comprehensive naming, this led to the inability to group together the explanatory variable and relate it to the response variable. Thus, the unique approach to this problem was to extract the initial letter of every observation, giving us the brand "Name." To further sophisticate our explanatory variables, a new explanatory variable "BodyType" was manually implemented based. Although the process was semi-tedious, proper sources and previous knowledge allowed this variable to add to the predictive ability of the machine learning algorithms.

- Fuel Type: Refers to the type of fuel used by the car. Vehicles using diesel are usually more fuel-efficient, more powerful, and more environmentally friendly, which may lead to higher prices of diesel cars. There were 4 levels of fuel types. We combined levels 'CNG' and 'LPG' due to the small number of observations of them, naming this new class 'Others.'

- Transmission: Refers to the type of transmission used by the car. Manual cars have historically been cheaper than automatic cars due to their comparatively simple parts and lower construction costs to automatic cars.

- Owner Type: Means whether the ownership of the vehicle's previous owner is Firsthand, Secondhand or other. From firsthand cars to second-hand cars, there will be more wear and tear. Hence, firsthand cars are normally the most expensive. There were 4 levels of owner types in our dataset. "First" indicates the previous owner was the first owner of the current used car. We grouped "Second", "Third", and "Fourth Above" together to be a whole group named "Second Above".

- Seats: Refers to the number of seats in the car. This is a tricky variable since a higher number of seats does not necessarily mean a higher price. We combined the number of seats equal to or more than 6 to be a whole group, due to the unbalanced number of these classes.

Ordinal Variable

- Age: Year refers to the year or edition of the model. The dataset contains used cars manufactured between the years 1998 to 2019. In our project, we calculated the variable "Age" for the vehicles based on the our current year $(2022 - Year)$ to allow easier interpretation. Normally, the price of a vehicle will depreciate as its age gets older due to capital depreciation.

2.1 Transformations

In the original dataset, the variables *Price*, *Power*, and *Engine* were observed to be heavily skewed after initial exploratory data analysis. We attempted to transform these variables using cubic root, square

root, and log transformation then selecting the approach that most approximates the normal. *Price* was originally heavy right-tailed, indicating that a large proportion of the prices of used cars are on the lower end. After viewing the transformations, we concluded to use *log(Price)* as it most approximates the Normal. Furthermore, we applied the log transform to *Engine* and square root transform to *Power*. Refer to the appendix for verification on these transforms.

2.2 Data Cleaning - Removing Observations

- In the initial EDA, we came across a 2017 BMW with an unreliable value for 'Kilometers Driven'. Since an average car can normally last between 200,000 and 300,000 kilometers with proper care, the vehicle we found here is only 5 years old but has over 6 million kilometers driven, which is an extreme value even after all transformations. We will thus remove this observation.
- We ran into some problems with the variable "Name" where several types had very few observations and thus did not successfully split into the training and cross-validation fold simultaneously. There were 8 observations in this dataset and they would prevent us from training the algorithm to predict the car price. Hence, we will remove these 8 observations.
- We deleted body types "saloon", "roadster", and "hardtop" in Excel directly, and deleted "truck" and "pickup truck" using R in the data cleaning process. Additionally, we classified "muv" as "suv", re-grouping these 11 groups into 5 groups. Finally, there were 2 levels with very few observations. We will also remove these 5 observations.

In the process of data cleaning, we removed a total of 14 observations, which gives us a final dataset with 6002 rows and 12 columns to analyze.

3 Methodology

We applied two statistical modelling approaches, the OLS regression model and random forest, to predict the price of used cars in India second-handed market by lm() function and rpart library. To ensure the accuracy of the model, we used cross-validation to randomly split the data into k folds, and fitted the models using k-1 folds and validated the models using the remaining fold. This process was repeated until every fold was used as the testing set.

We used k=3 and k=5 to separately build the models and compared the performances of our models based on the RMSEs, interval scores, average lengths of prediction intervals and coverage rates.

3.1 Linear Regression

3.1.1 Linear Regression Assumptions

To ensure the unbiasedness of estimators, linear regression has four assumptions to be fulfilled.(The plots referred are in the appendix)

- Linearity: The residual plot shows a fitted pattern along the red line which is approximately horizontal at zero. The presence of this pattern indicates the linearity of the model.
- Homoscedasticity: The plot shows that residuals are approximately spread out equally throughout the range of fitted values. Hence, we assume the variance of residuals is unchanged, that is homoscedasticity.
- Normality of Residuals: Residuals are falling approximately along the straight line, so we assume residuals are normally distributed.

- Multicollinearity: For categorical predictors, we use squared scaled GVIF instead VIF to measure multicollinearity among predictors by vif() function under the R car library. If squared scaled GVIF is less than 4, then we assume there is no multicollinearity.

3.1.2 Feature Selection

Before the real analysis, feature selection should be applied to the linear regression to evaluate and reduce insignificant variables based on the scale of AIC (Akaike Information Criterion). The AIC is evaluated based on the following equation where K is the number of parameters in the model:

$$AIC = -2(loglikelihood) + 2K \tag{3.1}$$

AIC is an estimator of prediction error representing the missing information in the model. It concerns the risk of over and underfitting by looking into the tradeoff between goodness-of-fit and the simplicity of the model. The smaller the AIC value, the better the model. To ensure the accuracy of model variables selection, we implemented all of the following stepwise regression methods:

- Backward elimination: Starting with all possible variables, the insignificant and least affecting variables are dropped so that the model has the smallest AIC to optimize the performance of linear regression.
- Forward Selection: Starting with no variables, adding the variables which give the most statistical significance until none improves the model to a statistically significant extent and obtaining the smallest AIC.
- Bi-directional Elimination: A combination of the above, adding or deleting the predictor variables until obtaining the smallest AIC.

After three methods of feature selection, there was no change in the initial model where the AIC is the smallest. This suggested us to use the model: $logPrice\ Name + BodyType + KMsDriven + FuelType + Transmission + OwnerType + Mileage + LogEngine + SqrtPower + Seats + Age$

3.1.3 Linear Regression with Subset of Variables

- Subset1: The table below is a small portion of the linear regression summary output of all explanatory variables contained. Despite having the smallest AIC value, the p-values of each category under Seats (without baseline Seats2) significantly exceed 5%. Thus, we attempted the regression model after subsetting a data frame without the Seats variable.

p-value of Seats	Fold 1	Fold 2	Fold 3
Seats4	0.45	0.55	0.30
Seats5	0.93	0.89	0.19
Seats6 & Above	0.92	0.69	0.12

- Subset2: The regression tree works by finding a predictor which gives the best split. The regression tree suggested that SqrtPower, Age and Name are the most deterministic based on the plot shown below. Thus, we can apply these three to the linear regression model and check the performance.

In addition, we also used the rmse() and mae() function under the Metrics library to compare the RMSE and MAE of three regression models in each fold to compare the accuracy and performance on predicting the price.

3.1.4 Linear Regression Models Comparison

The tables below shows the average lengths of the prediction intervals, interval scores and cover rates of each fold at levels 50% and 80%. Average length measures the lengths of prediction intervals. The interval is narrower and the prediction is more precise if the average length is lower. Interval Score is an estimator to compare prediction intervals from different methods by rewarding for narrow prediction intervals and incurring a penalty if the interval does not contain the true test value. Thus, smaller interval scores are better. Besides, the cover rate is used to measure whether a model is calibrated. If the proportions are near α, then this model is well fitted.

Average (k=3)	MAE	RMSE
Linear Regression (all vars)	0.19	0.25
Linear Regression (subset 1)	0.19	0.25
Linear Regression (subset 2)	0.22	0.29

Average (k=5)	MAE	RMSE
Linear Regression (all vars)	0.19	0.25
Linear Regression (subset 1)	0.19	0.25
Linear Regression (subset 2)	0.22	0.29

Linear (All Vars, 50%)	avgleng	IS	cover
	5487	10417	0.57
	5443	10379	0.55
	4988	9751	0.54
	5111	10250	0.53
	5062	9014	0.56
avg	5218	9962	0.55

Linear (All Vars, 80%	avgleng	IS	cover
	10561	15460	0.85
	10475	15674	0.83
	9598	14848	0.83
	9831	14912	0.82
	9742	13407	0.85
avg	10041	14860	0.84

Linear (Subset 1), 50%	avgleng	IS	cover
	5493	10404	0.57
	5446	10381	0.55
	4984	9743	0.54
	5110	10284	0.53
	5056	9030	0.56
avg	5218	9968	0.55

Linear (Subset 1), 80%	avgleng	IS	cover
	10554	15495	0.85
	10480	15695	0.83
	9590	14823	0.83
	9828	14937	0.82
	9731	13439	0.85
avg	5218	14878	0.55

Linear (Subset 2), 50%	avgleng	IS	cover
	6445	12707	0.56
	6247	11529	0.53
	5729	11116	0.36
	5973	12155	0.54
	5786	10759	0.57
avg	6036	11651	0.55

Linear (Subset 2), 80%	avgleng	IS	cover
	12458	18843	0.85
	12072	16437	0.83
	11069	16243	0.84
	11535	18105	0.83
	11182	16076	0.84
avg	11663	17141	0.84

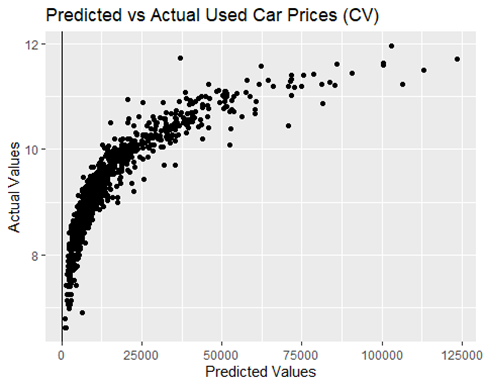


Figure 3.1: Predicted vs. Actual prices for regression model. (Y-axis in log-scale)

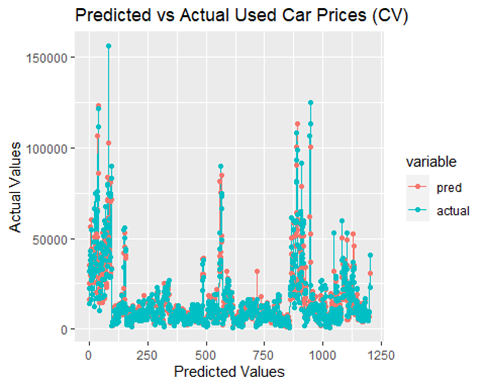


Figure 3.2: Predicted vs. Actual prices for regression model by observations.

In terms of MAE and RMSE, there is not a significant difference after removing *Seats* as an explanatory variable. However, the regression model only based on three predictors shows relatively lower accuracy. In addition, the interval scores of linear regression model with all variables are relatively the smallest among three models, which corresponds to the feature selection result based on AIC criterion. Hence, linear regression with all predictions is the best.

3.2 Tree Algorithms

3.2.1 Regression Trees

Starting with cross-validation k=3 folds, regression trees were produced for each fold but highly resembled one another. An example plot is shown below. One can observe that the regression trees has *SqrtPower* as the root note. Theoretically, *SqrtPower* being the root node variable indicates that it is the most deterministic variable for our response variable *logPrice*. The results also align with the regression model where the variable *SqrtPower* is statistically significant at the 1% significance level. The regression tree splits further based on the *Age* and *Name* variables then repeats for *SqrtPower* and *Age*. Although the splits based on the regression trees are surprisingly low in diversity, they suggest to us the variables relatively more important as to others.

By the summary tables below for k=5 folds at the 50% and 80% level, we can acknowledge that the coverage rates are very proximate to their intended levels, averaging to 0.5 and 0.79 respectively. Our average lengths and interval scores are also good considering the unit of our response variable, price. In calculating the interval scores, we once again exponentiated the *logPrice* variable as we want to convert it back to the original unit (price in Canadian dollars). Overall, we can assert that the regression tree algorithm performs well.

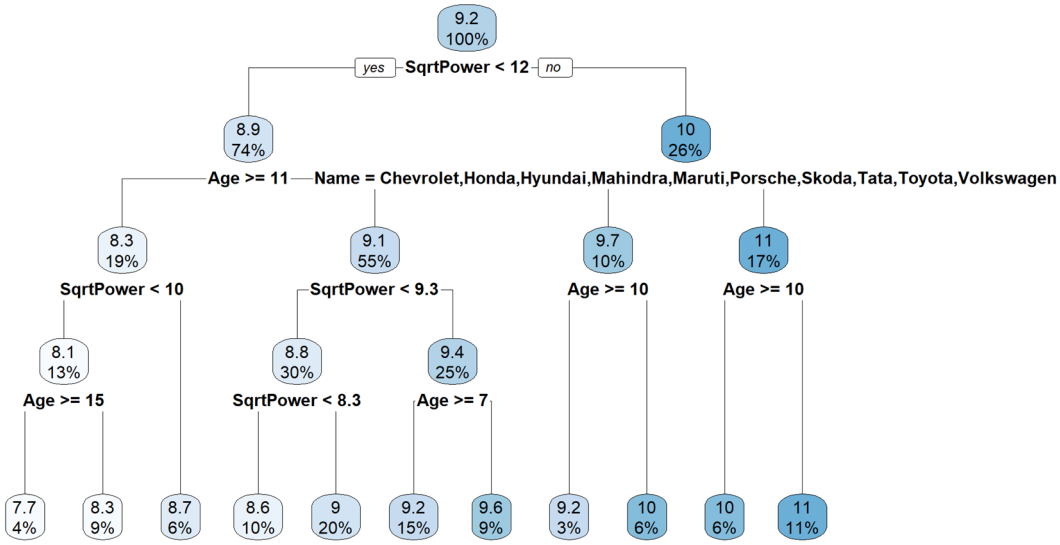
Regression Tree (50%)	avgleng	IS	cover
	7422	17769	0.49
	7877	16878	0.51
	7567	15310	0.51
	7475	16466	0.48
	7293	15527	0.51
Avg	7527	16390	0.5

Regression Tree (80%)	avgleng	IS	cover
	15336	26651	0.8
	16176	24930	0.79
	15444	22428	0.79
	14972	24379	0.79
	15011	22908	0.8
Avg	15389	24259	0.79

3.2.2 Random Forest

The random forest algorithm was evaluated on its RMSE's. We do not consider the random forest model very informative as we cannot make prediction intervals and thus calculate any lengths, interval scores, nor coverage rates. Regardless, on the log scale, the RMSE are small and the model seem to

Figure 3.3: One of the 5-fold cross-validation regression tree diagrams.



be well-performing. Plots of the predictions and actual used car prices are plotted and the points are unarguably close, showing the model's predictive power.

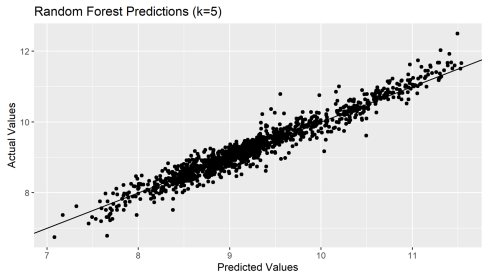


Figure 3.4: Plot for one of k=5 folds random forest predictions vs actual used car prices.

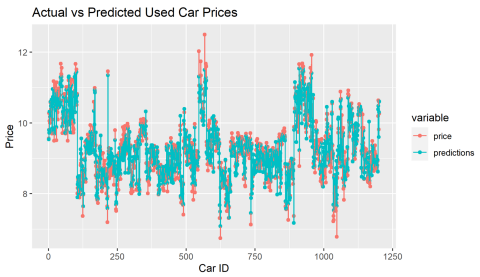


Figure 3.5: Visualizing the closeness of predictions.

Random Forest RMSE	0.22	0.22	0.23	0.24	0.22	**0.23**

3.2.3 Quantile Regression Forest (QR Tree)

The quantile regression forest is an approach predicts the intervals directly where the predictions correspond to specified quantiles rather than the arithmetic mean. Using the *quantregForest* library directly in R, we obtained metrics directly. In reference to the summary table, we observe that the coverage rates are very good as they are close to the intended levels although they tend to exceed by those of the regression model by a small degree. More importantly, the quantile regression forest provides evidence to having narrower prediction intervals and smaller interval scores in comparison to all other models. Thus, while retaining adequate coverage rates, the quantile regression forest is well-performing in terms of our metrics of interest.

QR Tree (50%)	avgleng	IS	cover	**QR Tree (80%)**	avgleng	IS	cover
	5095	9202	0.58		10526	13765	0.86
	5329	8764	0.57		10724	13007	0.86
	5059	8044	0.56		10318	12050	0.85
	5173	8501	0.58		10394	13178	0.87
	4850	7772	0.54		9864	11164	0.86
Avg	5101	8457	0.57	**Avg**	10365	12633	0.86

4 Discussion

4.1 Comparison of Models

In the methodology section, we saw the results to the regression models and tree algorithms using k=3 and k=5 cross-validation folds. While we can see slight difference in performance, the quantile regression forest consistently outperforms the other models for both respective quantity of folds–while retaining adequate coverage rates near the intended levels, the average lengths of prediction intervals and interval scores are significantly narrower and lower than those of regression trees and linear models for both 50% and 80% levels.

With stepwise regression of all approaches, the lowest AIC value was given to the regression model that used all of the variables from our dataset. Thus, giving us equation (3.2). At the 50% and 80% levels, the coverage rates are approximate to the intended levels with adequate average prediction interval lengths and interval scores. One must keep in mind the unit of our response variable that is large in magnitude. We then performed dataframe subsets with our regression model to attempt a simpler model by first removing *Seats* whose coefficients appeared to be insignificant in our model (despite having the lowest AIC value). The outcomes especially the average of all metrics (reported on page 10) did not vary by much. The second subset was taking the variables that the regression trees split on only–*SqrtPower,*

Age, and *Name*. This approach, however, downgraded the regression model's performance where the average lengths of prediction intervals and interval scores significantly increased for both levels. We can conclude that out of all the regression models, the original model with the full dataset or subset 1 are better performing.

As for the tree algorithms, we attempted a total of three separate algorithms. However, given that the regular random forest cannot provide interval scores, we can conclude that it is not informative despite showing signs of good performance in its RMSE and prediction plots. The regression trees and quantile regression forests both have great coverage rates that are near the intended levels although the quantile regression forests' coverage rates tend to exceed those of the regression trees. However, the regression trees have average prediction interval lengths and interval scores that are evidently wider than the quantile regression forests. Thus, out of the two, we can most definitely conclude that the quantile regression forests are more suitable for our analysis.

According to the result in the table, Linear Regression shows a slight advantage in average lengths while the quantile regression forests have significantly lower interval scores and larger cover rates than the intended levels. Besides, there is no obvious advantage shown in the regression model without Seats and a regression tree due to large interval score and average length. These two subsets do not work well. Overall, the metrics allow us to conclude that quantile regression forests have better performance in predicting the price of used cars in India–the algorithm consistently provides lower interval scores for all quantites of k-folds.

4.2 Variables

Based on the combination of our outputs from the regression models and tree algorithms, we can deduce down the most important features to be the horsepower of the car, age, as well as the brand name. The variable *SqrtPower* was the root note of our regression trees for both k=3 and k=5 folds and consistently statistically significant at $\alpha = 0.1\%$ for our regression models, having a positive association with the response variable. This indicates that the horsepower of used cars is suggested to be a great predictor for used car prices. While this conclusion came at a surprise to our original beliefs about the credibility of the variables, the results for the variable *Age* most definitely aligned with our hypothesis. This variable took part several times in our splits for the regression trees and is significant at $\alpha = 0.1\%$ for our regression models as well. The direction *Age* has with the used car price is negative which makes sense intuitively–the older the car, the lower the price due to capital depreciation. The *Name* of the used car is a little more complex in terms of its results. While the variable took part in splits of our regression trees, the regression summaries provided most levels of the variable with negative coefficients. While this can be further investigated, one can make an educational guess about the particularity of our dataset. Proven by our dataset, cars can be diversified in many ways and it is probable that the observations in our dataset have features for certain car brands influenced the direct price of them. In our dataset, most of the levels for *Name* have a negative relationship with the price and this can be further discussed in our following limitations section.

Overall, the stepwise regression approach and regression summaries suggested that the full set of explanatory variables were significant predictors for the variable *Price*. However, based on the regression trees, we observe horsepower, age, and brand name to be the most important.

5 Limitations

The main limitation of this analysis was the removal of the wide variety of car styles which were initially categorized down to brand names specifically. To ensure the model was trained with sophisticated features, we implemented our own variable "BodyType" along with the extraction of "Name". As mentioned in

section 2, this was done manually using sources like www.cars.com or previous knowledge. In this process, some information may have been lost and lowered the accuracy of each specific car model. Although our algorithm performs well in the testing set, there is uncertainty as to whether our model can be applied to observations out of the dataset; a high variance may exist. Besides, we can only predict the used car prices by their brand which is unrealistic if our model would be applied to true life. Future implementations can look into better categorizing the styles of the car or resampling techniques. Secondly, unlike the new car market, we can't level the brands of used cars to determine the brand value based on the average price due to the high variety caused by age and other related variables. These components worked together to affect the price of second-handed cars. As used car prices can be influenced by multiple factors, it is credible that our algorithms are specific to our dataset since a combination of inputs can greatly alter a vehicle's selling price; a high variance may exist in our algorithms. In addition to this, it is probable that our dataset does not include all the predictors which contribute to a used car's price. Many other features such as the car trim, whether the vehicle has been modified, and market valuation (cars that are widely recognized for performance and long-lasting value) can significantly influence its selling value. Thus, under the diversity of the automotive industry, there are many degrees in which data collection can be expanded and manipulated to improve an algorithm's predictive ability.

6 Contributions

- Kexin Chen: Feature Selection, Linear Regression code, Presentation 1, Final Report(Feature Selection, Regression models, Conclusion)
- Chin Chen Lo: EDA, Random Forest code, Presentation 1, Presentation 2, Final report (Abstract, Tree Algorithms, Discussion, Limitations)
- Yuxin Mao: EDA Plots, Presentation 1, Final Report (The Dataset, Appendix)
- Sisi Yang: EDA Plots, Presentation 1, Presentation 2

7 Appendix

7.1 Summary Statistics for Continuous Variables

Summary Statistics	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD
Kilometers Driven (km)	171	34000	53000	58748	73000	6500000	91290.4

Table 2: Summary Statistics for Kilometers Driven in km

Summary Statistics	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD
Mileage (kmpl)	0	15.5	18.0	17.7	21.0	33.0	4.6

Table 3: Summary Statistics for Mileage in kmpl

Summary Statistics	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD
Engine (cc)	624	1198	1493	1622	1984	5998	603.2

Table 4: Summary Statistics for Engine in cc

Summary Statistics	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD
Power (bhp)	8.0	74.0	93.0	112.4	138.0	560.0	54.0

Table 5: Summary Statistics for Power in bhp

7.2 Frequency Tables for Categorical Variables and Ordinal Variable

Name	Audi	BMW	Chevrolet	Datsun	Fiat	Ford	Honda	Hyundai
Frequency	236	267	121	13	28	300	607	1107
Name	Jaguar	Jeep	Land	Mahindra	Maruti	Benz	Mini	Mitsubishi
Frequency	40	15	60	271	1172	315	26	27
Name	Nissan	Porsche	Renault	Skoda	Tata	Toyota	Volkswagen	Volvo
Frequency	91	18	145	173	182	411	362	21
Name	Ambassador	Bentley	Force	ISUZU	Lamborghini	Smart		
Frequency	1	1	3	1	1	1		

Table 6: Frequency Table for Name

Body Type	hatchback	sedan	suv	minivan	convertible	saloon
Frequency	2213	2237	1497	40	11	4
Body Type	hardtop	muv	roadster	pickup truck	truck	
Frequency	3	3	3	3	2	

Table 7: Frequency Table for Body Type Before Grouping

Body Type	convertible	hatchback	minivan	suv	sedan
Frequency	11	2213	40	1500	2237

Table 8: Frequency Table for Body Type After Grouping

Fuel Type	CNG	Diesel	LPG	Petrol
Frequency	56	3205	10	2745

Table 9: Frequency Table for Fuel Type Before Grouping

Fuel Type	Diesel	Petrol	Others
Frequency	3205	2745	66

Table 10: Frequency Table for Fuel Type After Grouping

Transmission	Automatic	Manual
Frequency	1717	4299

Table 11: Frequency Table for Transmission

Owner Type	First	Second	Third	Fourth & Above
Frequency	4926	968	113	9

Table 12: Frequency Table for Owner Type Before Grouping

Owner Type	First	Second & Above
Frequency	4926	1090

Table 13: Frequency Table for Owner Type After Grouping

Seats	2	4	5	6	7	8	9	10
Frequency	16	98	5055	31	674	134	3	5

Table 14: Frequency Table for Seats Before Grouping

Seats	2	4	5	6 & Above
Frequency	16	98	5055	847

Table 15: Frequency Table for Seats After Grouping

Age(Years)	3	4	5	6	7	8	9	10
Frequency	102	298	587	740	744	797	648	580
Age(Years)	11	12	13	14	15	16	17	18
Frequency	465	342	198	174	125	78	57	31
Age(Years)	19	20	21	22	23	24		
Frequency	17	15	8	4	2	4		

Table 16: Frequency Table for Age(Years)

7.3 Plots Before Transformation and Data Cleaning

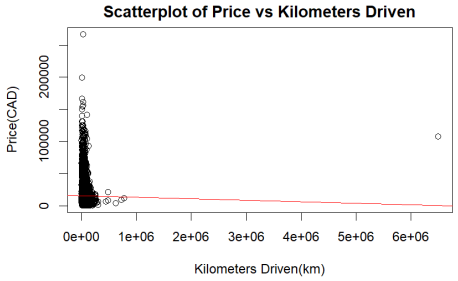


Figure 7.1: Scatterplot of Price vs Kilometers Driven Before Transformation

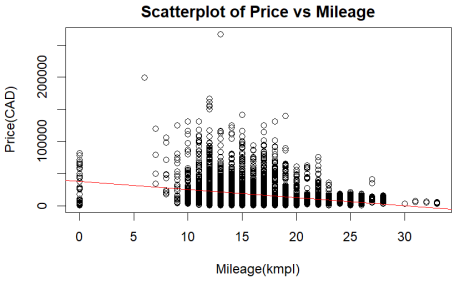


Figure 7.2: Scatterplot of Price vs Mileage Before Transformation

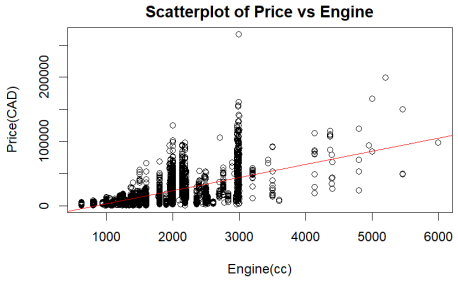


Figure 7.3: Scatterplot of Price vs Engine Before Transformation

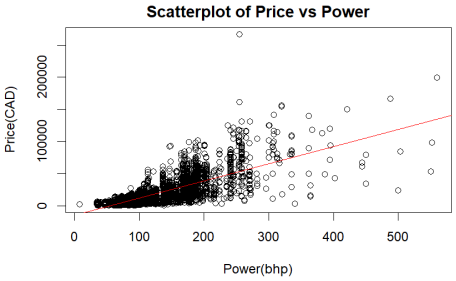


Figure 7.4: Scatterplot of Price vs Power Before Transformation

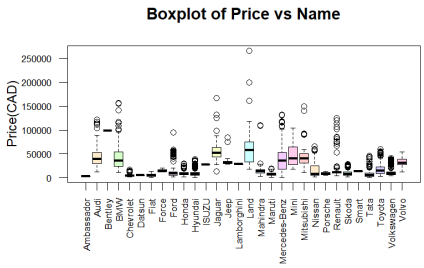


Figure 7.5: Boxplot of Price vs Name Before Transformation

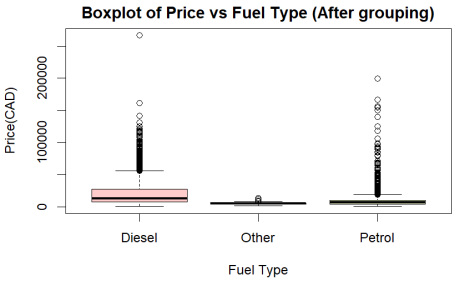


Figure 7.6: Boxplot of Price vs Fuel Type Before Transformation

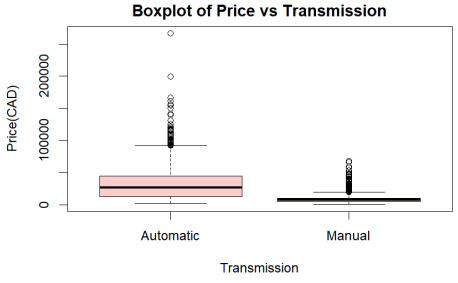


Figure 7.7: Boxplot of Price vs Transmission Before Transformation

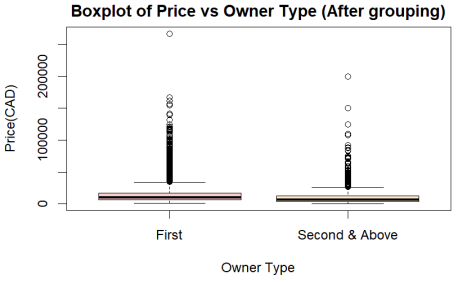


Figure 7.8: Boxplot of Price vs Owner Type Before Transformation

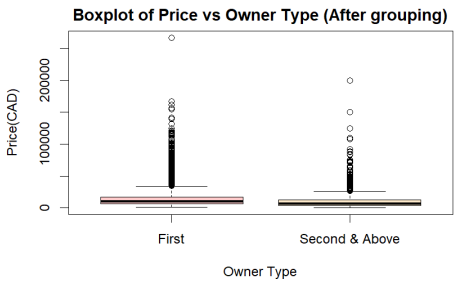


Figure 7.9: Boxplot of Price vs Seats Before Transformation

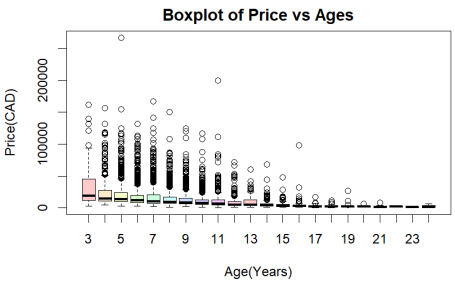


Figure 7.10: Boxplot of Price vs Ages Before Transformation

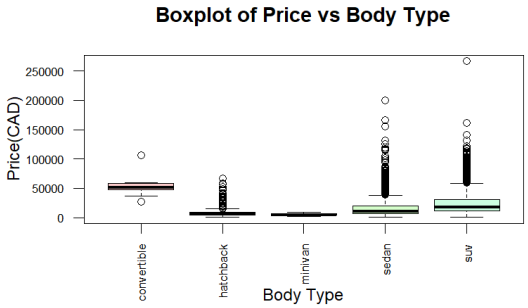


Figure 7.11: Boxplot of Price vs Body Type Before Transformation

7.4 Plots After Transformation and Data Cleaning

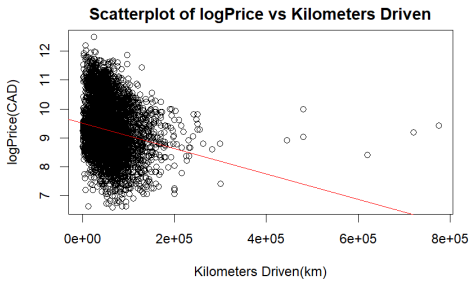


Figure 7.12: Scatterplot of logPrice vs Kilometers Driven Before Transformation

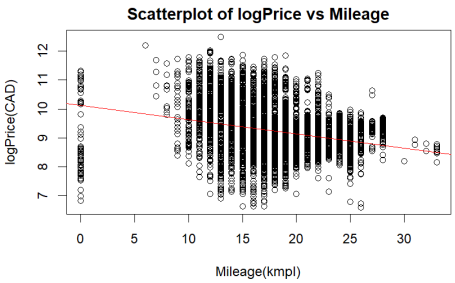


Figure 7.13: Scatterplot of logPrice vs Mileage Before Transformation

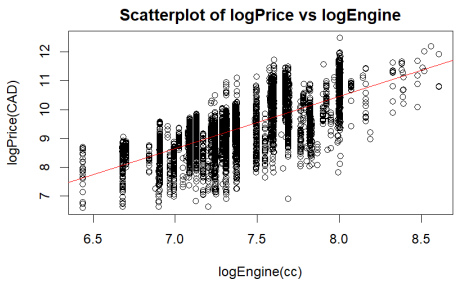


Figure 7.14: Scatterplot of logPrice vs logEngine Before Transformation

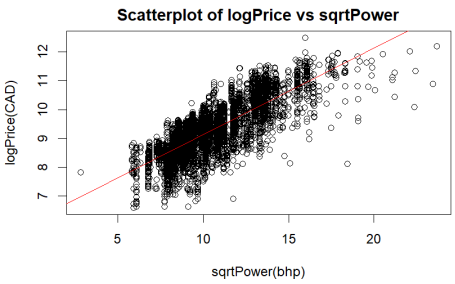


Figure 7.15: Scatterplot of logPrice vs sqrtPower Before Transformation

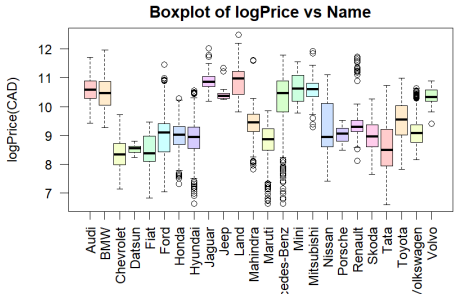


Figure 7.16: Boxplot of logPrice vs Name Before Transformation

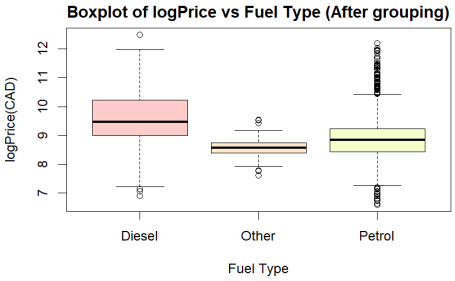


Figure 7.17: Boxplot of logPrice vs Fuel Type Before Transformation

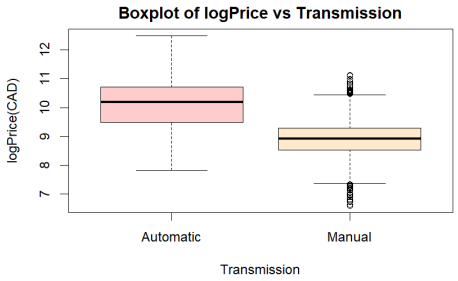


Figure 7.18: Boxplot of logPrice vs Transmission Before Transformation

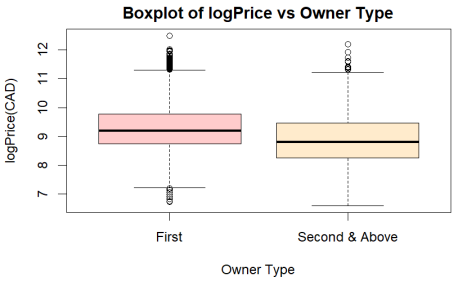


Figure 7.19: Boxplot of logPrice vs Owner Type Before Transformation

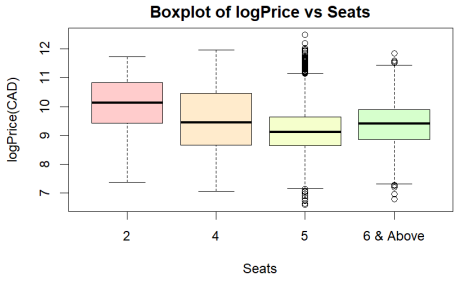


Figure 7.20: Boxplot of logPrice vs Seats Before Transformation

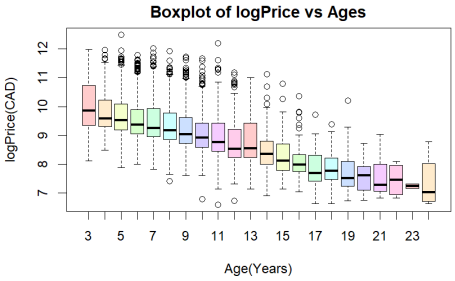


Figure 7.21: Boxplot of logPrice vs Ages Before Transformation

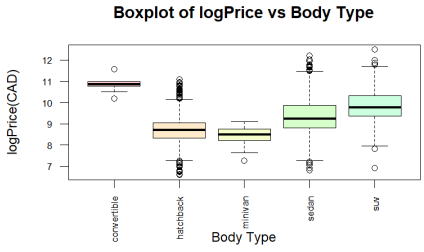


Figure 7.22: Boxplot of logPrice vs Body Type Before Transformation

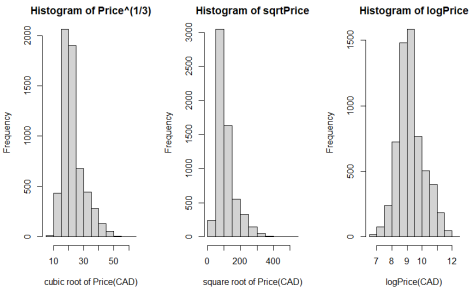


Figure 7.23: Transformation of Price

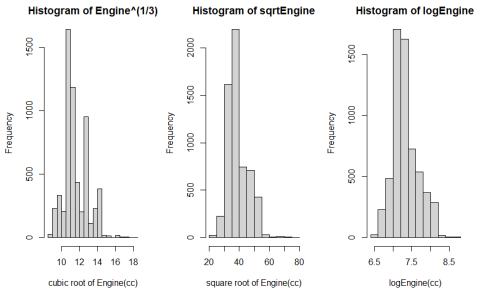


Figure 7.24: Transformation of Engine

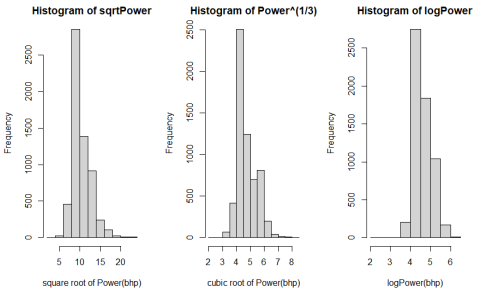


Figure 7.25: Transformation of Power

7.5 Regression Plots

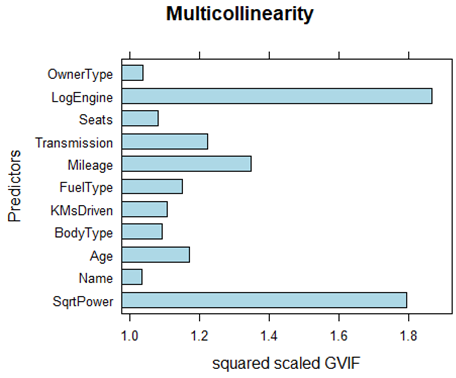


Figure 7.26: Multicollinearity plot

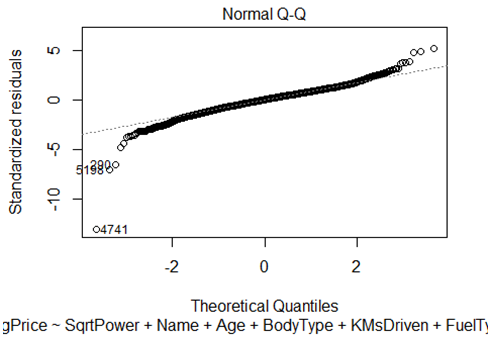


Figure 7.27: Normal QQ Plot

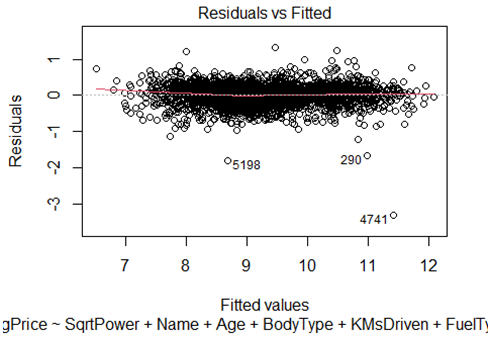


Figure 7.28: Residual vs fitted plot

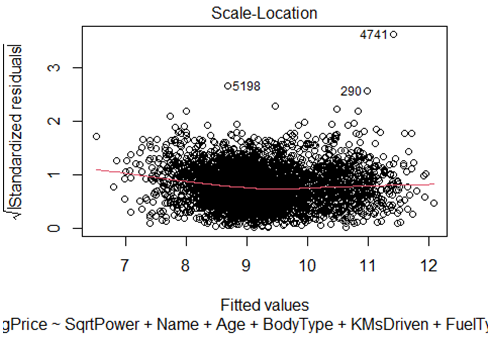


Figure 7.29: Scale Location

7.6 Correlation

Correlation	Kilometers Driven	Mileage	LogEngine	SqrtPower
Kilometers Driven	1.00			
Mileage	-0.15	1.00		
LogEngine	0.17	-0.58	1.00	
SqrtPower	0.01	-0.50	0.88	1.00
Age	0.45	-0.32	0.04	-0.05

Table 17: Correlation Between Quantitative Explanatory Variables