Final Project Report

5/12/2022

Car Selling Price Prediction

Rishika Reddy Aleti
G01349665

Introduction

Used Car Price forecast is a hot issue Because of the record volume of cars purchased and sold. People in poor countries prefer to buy secondhand cars since they are more affordable. The project's main goal is to estimate used car's selling price using factors that are substantially connected with different attributes of the dataset. Statistical Analysis, visualizations and regression modelling has been used to do this. During pre-processing, null, redundant, and missing values were omitted from the dataset. Two regressors (Random Forest Regressor, Linear Regression were trained, tested, and evaluated against a benchmark dataset in this supervised learning study. This project aims to deliver price prediction models to the public, to help guide the individuals looking to buy or sell cars and to give them a better insight into the automotive sector. Another goal of the project is to explore new methods to evaluate used cars prices and to compare their accuracies.

Data Description

Data was collected and Scrapped from an online website Kaggle which sourced this data from another website CarDekho. At first, the data is loaded from CSV file into R programming and displayed. The dataset contains 8128 observations with 13 variables. In the next step we check for the missing values to remove if there any missing values. Later, we remove the suffixes that are present in the current dataset and mold the dataset to make a regression model. Displaying the original dataset and the modified dataset after converting the characters into factors.

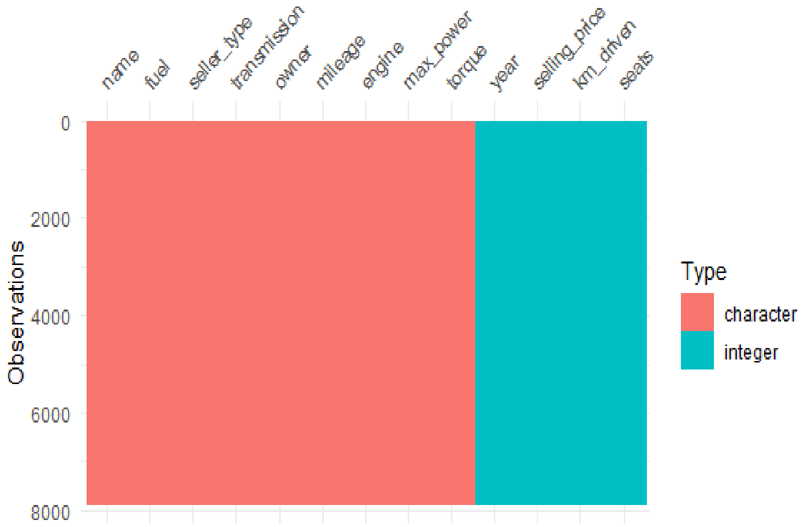


Fig 1- All the variables types originally.

From the visual interpretation of original dataset we can see that there are a lot of primary variables are structured in characters. Fitting a logistic regression model for this dataset using all other variable as predictors needs data preprocessing.

Data Preprocessing

1. Conversion of variables into factors.

```
car_df[c("fuel","transmission", "seller_type")] <-
sapply(car_df[c("fuel","transmission", "seller_type")],as.factor)
```

2. Assigning labels to owner_type.

```
car_df$owner <- factor(car_df$owner, labels = c(0,1,2,3,4), levels =
c("Test Drive Car", "First Owner", "Second Owner", "Third Owner", "Fourth
& Above Owner"))
```

3. Conversion of variables into numeric.

```
car_df[c("mileage","engine", "max_power", "torque", "owner")] <-
sapply(car_df[c("mileage","engine", "max_power", "torque",
"owner")],as.numeric)
```

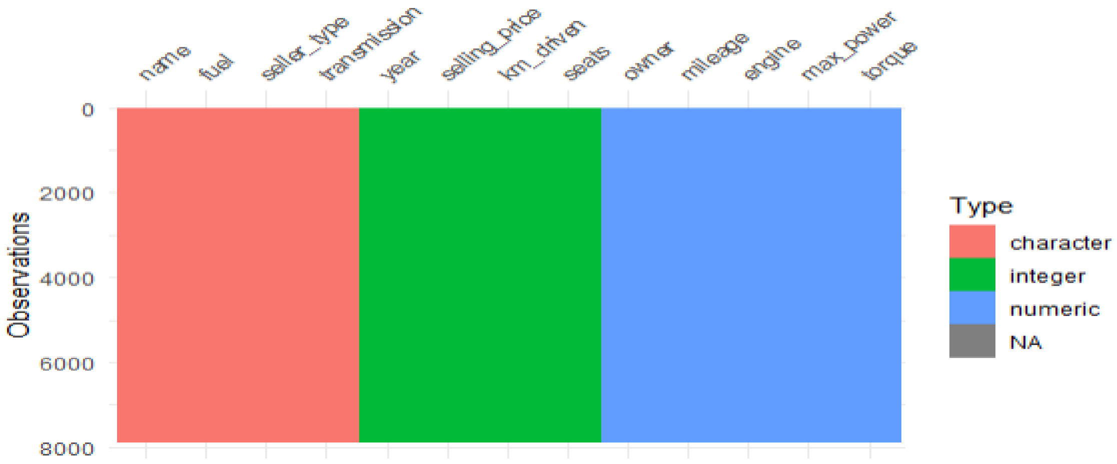
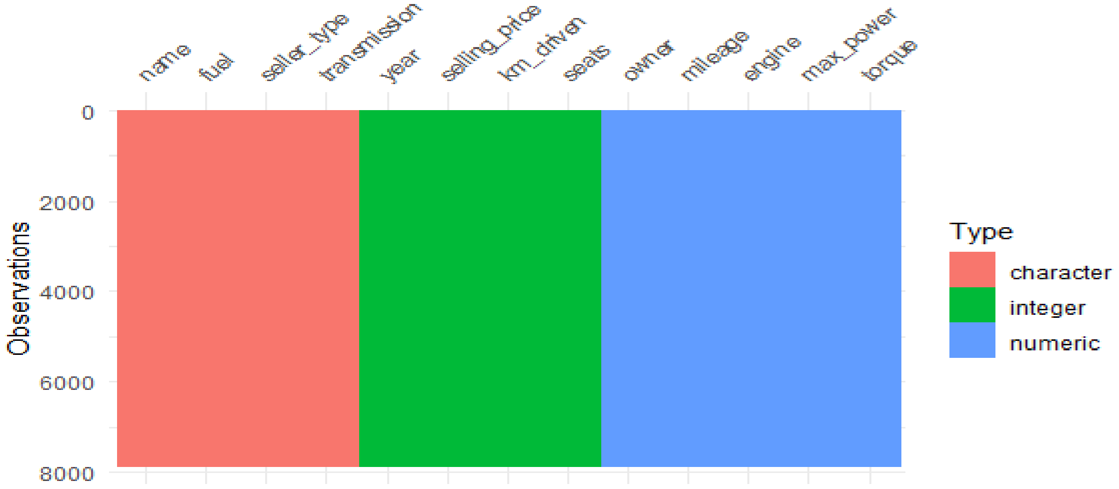


Fig 2- The structure of data after preprocessing.

4. Removing the NA values from the dataset.

```
car_df <- na.omit(car_df)
car_df
vis_dat(car_df)
```



5. Removing the missing values from the processed dataset.

```
cat("After removing the missing values there are:\n", nrow(car_df), "rows
and", ncol(car_df), "columns")
```

```
After removing the missing values there are:
 7881 rows and 13 columns
```

Data Analysis

Data analysis through visualization of selling price by fuel, type of seller and transmission.

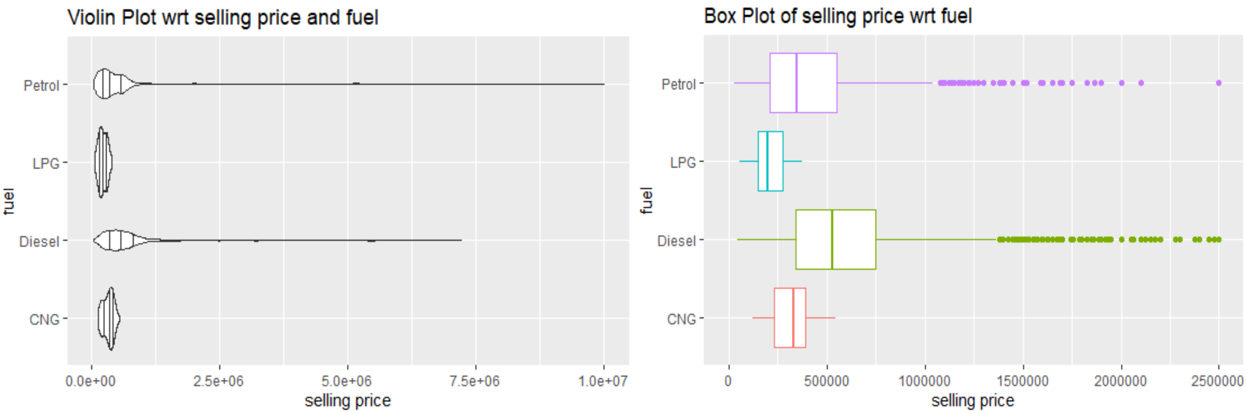
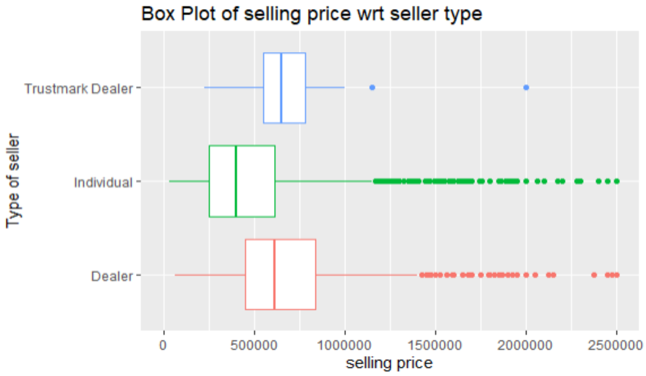


Fig 3- Plots to analyze selling price and fuel.

From the plots we can identify that diesel and petrol have the highest selling price and LPG has the least selling price. We can identify a pattern that is the CNG and LPG are the lowest.



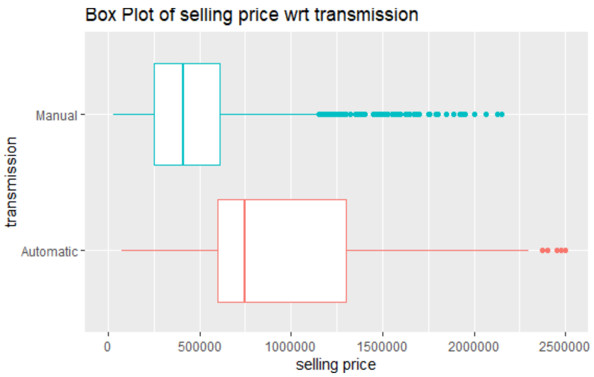


Fig 4- Box plot of selling price and seller type. *Fig 5- Box plot of selling price and transmission*.

From the plots we can identify that Manual tra the highest selling price and LPG has the least selling price. We can identify a pattern that is the CNG and LPG are the lowest.

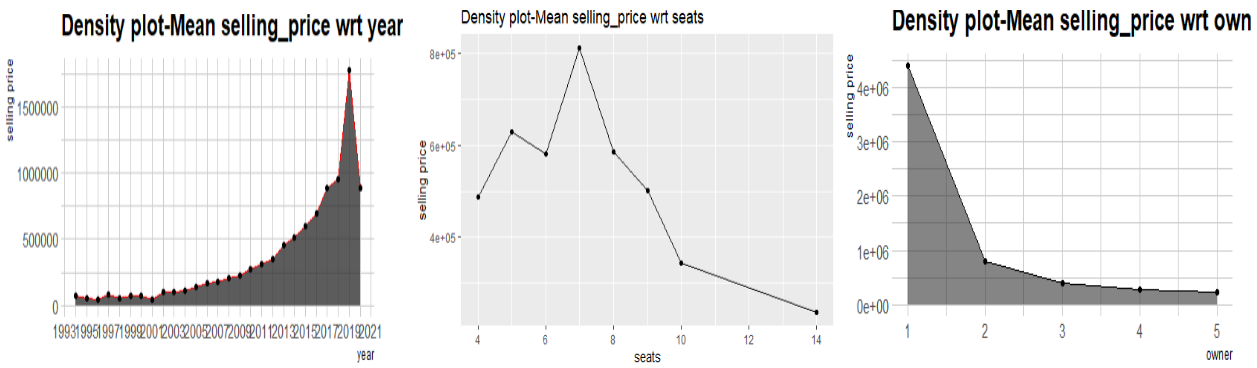


Fig 5- Density plots of mean selling price wrt to year, seats and owner.

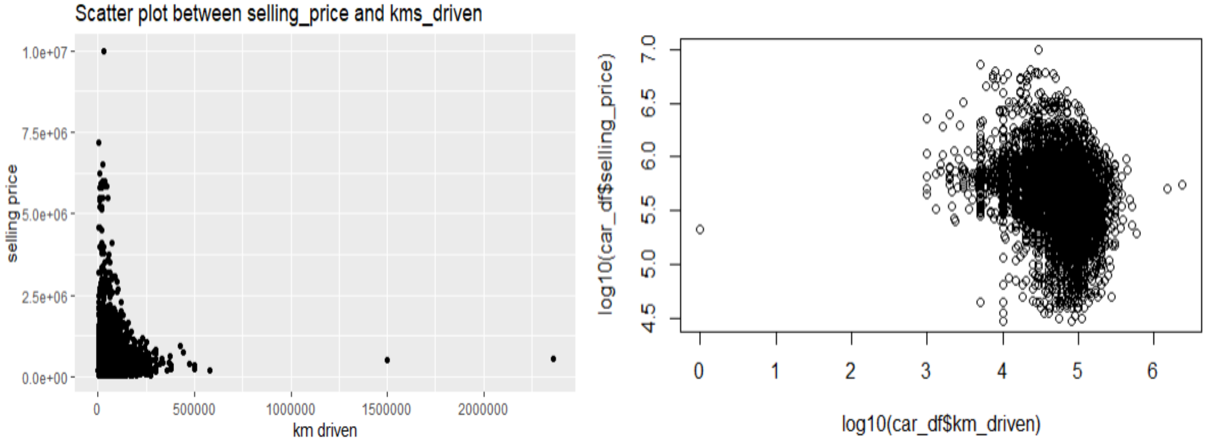


Fig 6- Determining the relationship between selling price and kms driven and improved plot for the same.

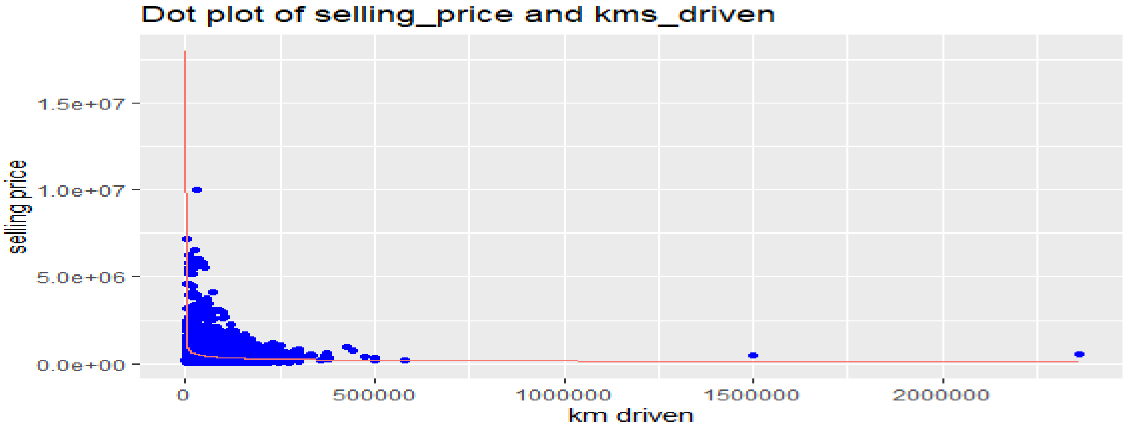
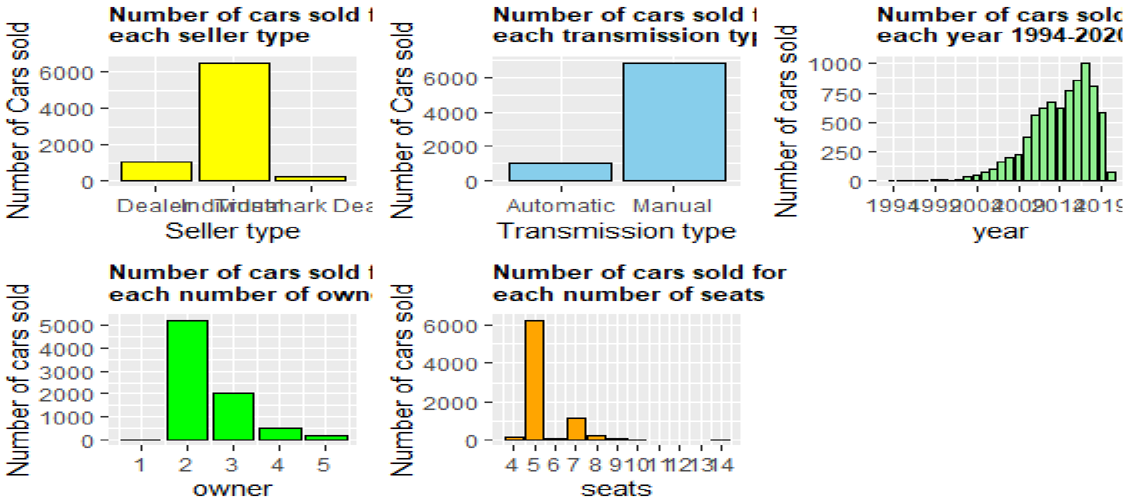


Fig 7- Dot plot of selling price and kms driven.

We can spot two outier's in the box that might cause difficulty while making a regression model.

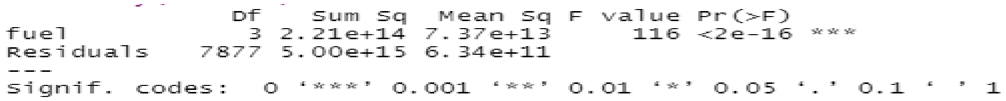
Exploratory Analysis



Data analysis through visualization of cars sold by seller types, transmission type, year, number of owners and seats.We can identify patterns in all of the above plots. We can see that in Fig 8 the individual seller have the highest rate of cars sold drastically differing from the other type of owner's, Dealer and Trustmark Dealer. In Fig 9 we can make out that the manual transmission type of cars are sold way more than that of the automatic transmission type. In fig 10 we can see that the number cars sold were comparatively more in the years 2014-2019 and the least cars were sold in the year 1994. In fig 11 we can see that type of owner also is an influential factor in the number of cars sold. In fig 12 the number of cars with 5 seats are sold more than the other number of cars seats.

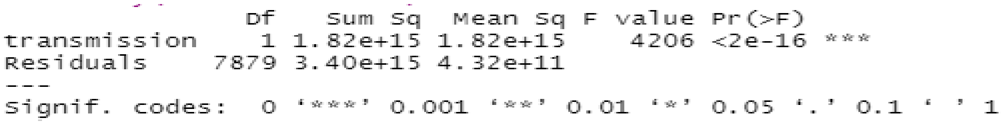
Correlation Analysis - Selling price and other variables.

1. The correlation of selling price and fuel.



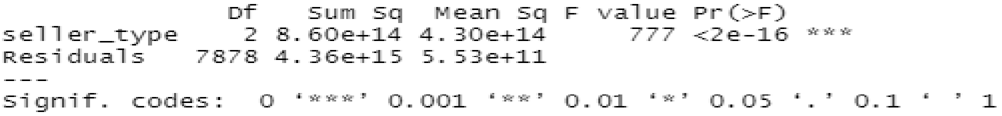
```
                Df   Sum Sq  Mean Sq F value Pr(>F)    
fuel             3 2.21e+14 7.37e+13     116 <2e-16 ***
Residuals     7877 5.00e+15 6.34e+11                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

2. The correlation of selling price and transmission.



```
                Df   Sum Sq  Mean Sq F value Pr(>F)    
transmission     1 1.82e+15 1.82e+15    4206 <2e-16 ***
Residuals     7879 3.40e+15 4.32e+11                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3. The correlation of selling price and transmission.



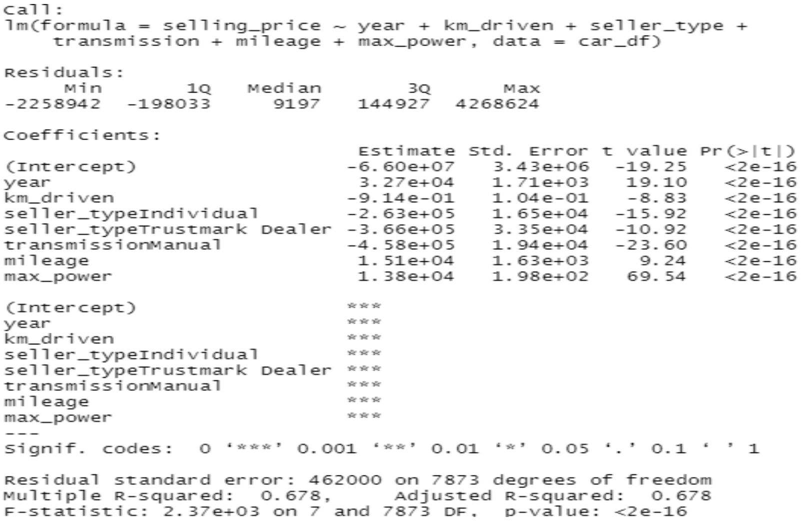
```
                Df   Sum Sq  Mean Sq F value Pr(>F)    
seller_type      2 8.60e+14 4.30e+14     777 <2e-16 ***
Residuals     7878 4.36e+15 5.53e+11                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

4. Visual depiction of the correlation

Predicting the selling price based on the variables.

Feature selection comparing methods is Forward Selection.

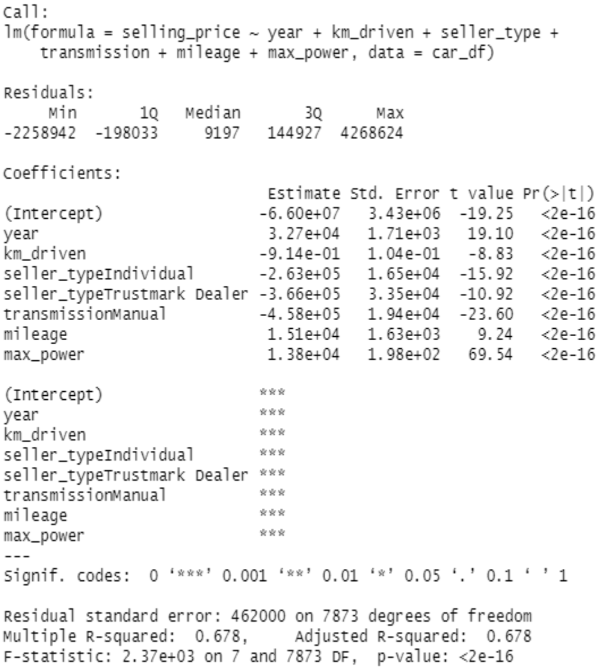
Forward Selection - Starts with no predictors in the model, iteratively adds the most contributive predictors, and stops when the improvement is no longer statistically significant.



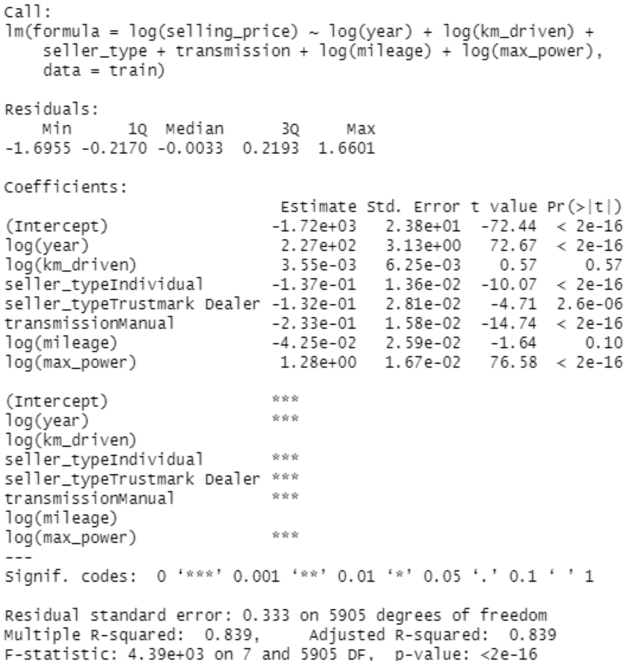
```
Call:
lm(formula = selling_price ~ year + km_driven + seller_type + 
    transmission + mileage + max_power, data = car_df)

Residuals:
     Min       1Q   Median       3Q      Max 
-2258942  -198033     9197   144927  4268624 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                -6.60e+07   3.43e+06  -19.25   <2e-16
year                        3.27e+04   1.71e+03   19.10   <2e-16
km_driven                  -9.14e-01   1.04e-01   -8.83   <2e-16
seller_typeIndividual      -2.63e+05   1.65e+04  -15.92   <2e-16
seller_typeTrustmark Dealer -3.66e+05   3.35e+04  -10.92   <2e-16
transmissionManual         -4.58e+05   1.94e+04  -23.60   <2e-16
mileage                     1.51e+04   1.63e+03    9.24   <2e-16
max_power                   1.38e+04   1.98e+02   69.54   <2e-16
                               
(Intercept)                ***
year                       ***
km_driven                  ***
seller_typeIndividual      ***
seller_typeTrustmark Dealer ***
transmissionManual         ***
mileage                    ***
max_power                  ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 462000 on 7873 degrees of freedom
Multiple R-squared:  0.678,	Adjusted R-squared:  0.678 
F-statistic: 2.37e+03 on 7 and 7873 DF,  p-value: <2e-16
```

Fig 13- The model of exhaustive linear regression.

Multiple Linear Regression Model.



```
Call:
lm(formula = selling_price ~ year + km_driven + seller_type +
    transmission + mileage + max_power, data = car_df)

Residuals:
     Min       1Q   Median       3Q      Max
-2258942  -198033     9197   144927  4268624

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               -6.60e+07   3.43e+06  -19.25   <2e-16
year                       3.27e+04   1.71e+03   19.10   <2e-16
km_driven                 -9.14e-01   1.04e-01   -8.83   <2e-16
seller_typeIndividual     -2.63e+05   1.65e+04  -15.92   <2e-16
seller_typeTrustmark Dealer -3.66e+05  3.35e+04  -10.92   <2e-16
transmissionManual        -4.58e+05   1.94e+04  -23.60   <2e-16
mileage                    1.51e+04   1.63e+03    9.24   <2e-16
max_power                  1.38e+04   1.98e+02   69.54   <2e-16

(Intercept)                ***
year                       ***
km_driven                  ***
seller_typeIndividual      ***
seller_typeTrustmark Dealer ***
transmissionManual         ***
mileage                    ***
max_power                  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 462000 on 7873 degrees of freedom
Multiple R-squared:  0.678,     Adjusted R-squared:  0.678
F-statistic: 2.37e+03 on 7 and 7873 DF,  p-value: <2e-16
```

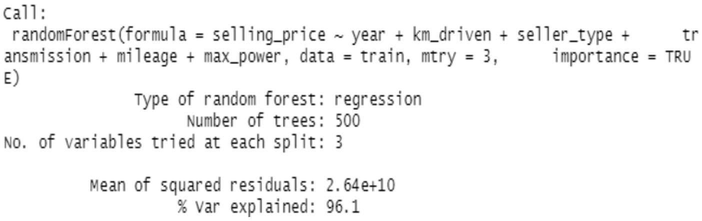


```
Call:
lm(formula = log(selling_price) ~ log(year) + log(km_driven) +
    seller_type + transmission + log(mileage) + log(max_power),
    data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6955 -0.2170 -0.0033  0.2193  1.6601

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -1.72e+03   2.38e+01  -72.44  < 2e-16
log(year)                    2.27e+02   3.13e+00   72.67  < 2e-16
log(km_driven)               3.55e-03   6.25e-03    0.57     0.57
seller_typeIndividual       -1.37e-01   1.36e-02  -10.07  < 2e-16
seller_typeTrustmark Dealer -1.32e-01   2.81e-02   -4.71  2.6e-06
transmissionManual          -2.33e-01   1.58e-02  -14.74  < 2e-16
log(mileage)                -4.25e-02   2.59e-02   -1.64     0.10
log(max_power)               1.28e+00   1.67e-02   76.58  < 2e-16

(Intercept)                 ***
log(year)                   ***
log(km_driven)
seller_typeIndividual       ***
seller_typeTrustmark Dealer ***
transmissionManual          ***
log(mileage)
log(max_power)              ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.333 on 5905 degrees of freedom
Multiple R-squared:  0.839,     Adjusted R-squared:  0.839
F-statistic: 4.39e+03 on 7 and 5905 DF,  p-value: <2e-16
```

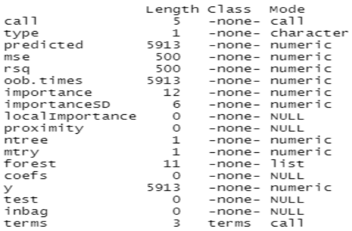
Fig 14- Linear Regression model of the dataset. *Fig 15- Improved linear regression model.*

In the first regression model we can see that the accuracy of the R-sqaure of training data is almost 67%. The accuracy of the original model is increased by adding log function to numeric variables, the R-sqaure of training data is almost 83% indicating that model is trained efficiently.

Random Forest Regression



```
Call:
 randomForest(formula = selling_price ~ year + km_driven + seller_type +      tr
ansmission + mileage + max_power, data = train, mtry = 3,      importance = TRU
E)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 2.64e+10
                    % Var explained: 96.1
```

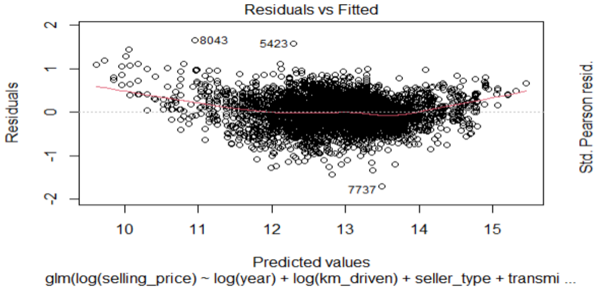


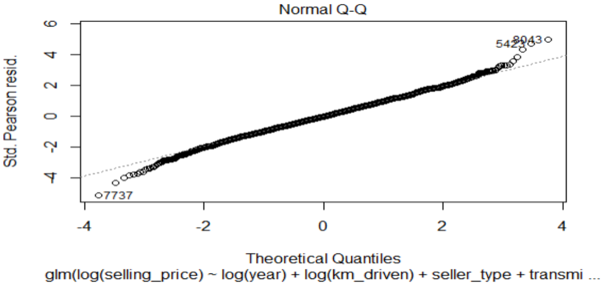
```
                Length Class  Mode
call               5   -none- call
type               1   -none- character
predicted       5913   -none- numeric
mse              500   -none- numeric
rsq              500   -none- numeric
oob.times       5913   -none- numeric
importance        12   -none- numeric
importanceSD       6   -none- numeric
localImportance    0   -none- NULL
proximity          0   -none- NULL
ntree              1   -none- numeric
mtry               1   -none- numeric
forest            11   -none- list
coefs              0   -none- NULL
y               5913   -none- numeric
test               0   -none- NULL
inbag              0   -none- NULL
terms              3   terms  call
```

Fig 16- Random Forest Regression model *Fig 17- Random Forest Regression model summary.*

The R-square of training data is almost ___% which indicates that model is trained efficiently.

Plotting Regression Model





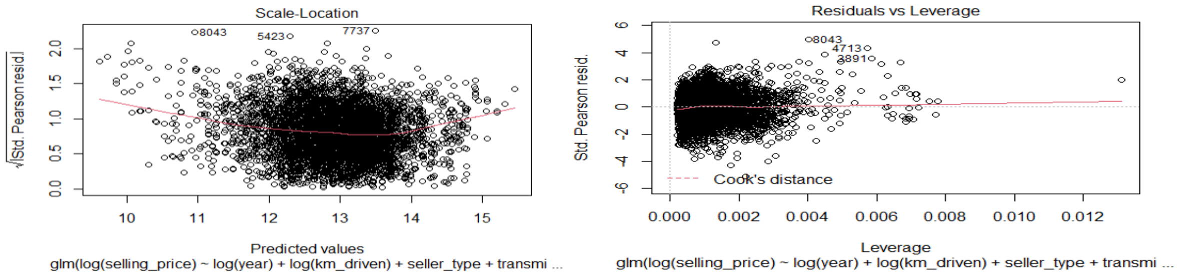


Fig 18- a) Plotting the linearity regression model (existing outliers).

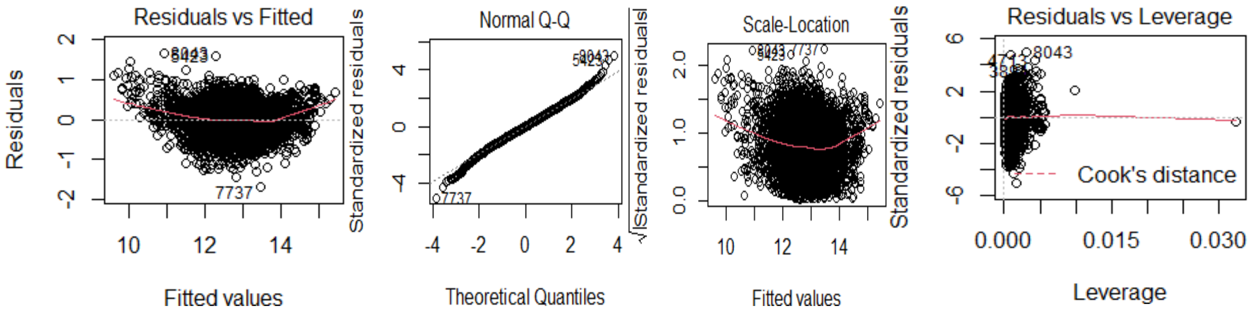


Fig 18- b)Plotting the linearity regression model without outliers.

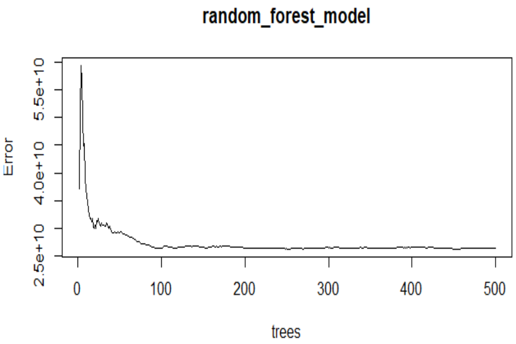


Fig 19- Plot of random forest model.

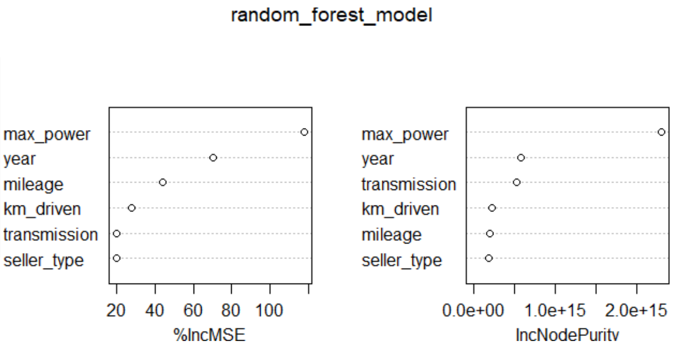


Fig 20- Plot of random forest model's variable importance.

As we can see the outlier's in the plot of the model we remove the outlier's in the same model and create another model with no outlier's and store it as a new model. After removing the outlier's the model's accuracy get increased and the model becomes more accurate to use as the outlier's are few values of the variables that change the entire model of the data and effect the accuracy percentile of the model.

Comparison

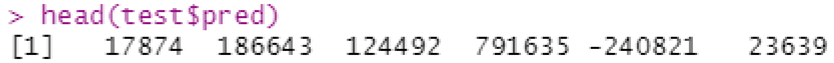
In machine learning, multiple linear regression is frequently used for prediction. A different type of regression is random forest regression. It does not use linear regression assumptions.

RFR appeared to be superior to MLR in terms of explanatory value and error for this data set. With this data set, this finding implies that RFR may have an advantage over MLR for selling price of car prediction, although MLR can still have strong predictive value in some circumstances. Given the evaluation parameters, the Random Forest Regressor beat the other algorithm, having the highest accuracy. Linear regression was the least accurate 67% and with an increase of accuracy by 6% percent after adding log function i.e., 83%, although having a smaller error value.

The best recommended model is Random Forest Regression model as the accuracy is 88.8%.

Research Question

The prediction of the selling price of car using all other variables as predictors predict in forward selection is as follows:-



```
> head(test$pred)
[1]   17874  186643  124492  791635 -240821   23639
```

The Random Forest approach outperforms Linear Regression by a little margin. Random Forests, on the other hand, tend to overfit the dataset due to their tendency to build longer trees. Because of the shorter training time, the regression model performed marginally better.

Limitations

In the past year the world of automobiles has seen a drastic change with the semiconductor shortages after the pandemic, which led to spike in used car prices. Hence, there was fast change in car prices during this study which will affect the actual car pricing prediction future. As the current dataset will undervalue the cars in the market. Therefore, a model that is built on real time data can be best integrated into a mobile app for public use would be the idea solution.

One of the most difficult aspects of this data set is that the predictor variables' distributions were non-normal. As a result, traditional statistical methods were ineffective in assessing this data. Furthermore, among the 14 variables in the dataset, there are only three numerical variables. This limits our ability to use Pearson's correlation coefficient.

As a result, we'll move on to the following portion. Machine learning models will be applied in the next part, and their performance will be evaluated using mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) (RMSE).

Future Analysis Techniques

The researchers of this project anticipate that in the near future, the most sophisticated algorithm should be used for making predictions, and then the model will be integrated into a mobile app or web page for the general public to use and understand the price prediction of the car easily. In the future, our machine learning model might be linked to a variety of websites that provide real-time data for price forecasting. We might also include a lot of previous car price data to help improve the machine learning model's accuracy. Train on clusters of data rather than the whole dataset is suggested. To correct for overfitting in Random Forest, different selections of features and number of trees will be tested to check for change in performance. Some future work for this project can be as follows:-

- Include more information in the price distribution plot to illustrate the pricing's appropriateness.
- Examine price methods among different dealers, geographies, seasons and elements that influence pricing in greater depth.
- Engine, transmission, and type of drive should be in numeric values.
- The cost of ownership (based on EPA information) should be mentioned exactly.
- Compile historical data and forecast price trends for each car.
- Find the rate of depreciation and the factors influencing depreciation.
- Create the output in such a way that it is appealing and engaging i.e., colors on the model.

Optimization of performance by designing deep learning network topologies, employing adaptive learning rates, and training on data clusters rather than the entire dataset.

References

Xtable – David B. Dahl [aut], David Scott [aut, cre], Charles Roosen [aut], Arni Magnusson [aut], Jonathan Swinton [aut], Ajay Shah [ctb], Arne Henningsen [ctb], Benno Puetz [ctb], Bernhard Pfaff [ctb], Claudio Agostinelli [ctb], Claudius Loehnert [ctb], David Mitchell [ctb], David Whiting [ctb], Fernando da Rosa [ctb], Guido Gay [ctb], Guido Schulz [ctb], Ian Fellows [ctb], Jeff Laake [ctb], John Walker [ctb], Jun Yan [ctb], Liviu Andronic [ctb], Markus Loecher [ctb], Martin Gubri [ctb], Matthieu Stigler [ctb], Robert Castelo [ctb], Seth Falcon [ctb], Stefan Edwards [ctb], Sven Garbade [ctb], Uwe Ligges [ctb]

kableExtra - Hao Zhu ORCID iD [aut, cre], Thomas Travison [ctb], Timothy Tsai [ctb], Will Beasley [ctb], Yihui Xie [ctb], GuangChuang Yu [ctb], Stéphane Laurent [ctb], Rob Shepherd [ctb], Yoni Sidi [ctb], Brian Salzer [ctb], George Gui [ctb], Yeliang Fan [ctb], Duncan Murdoch [ctb], Bill Evans [ctb]

stringi - Gagolewski M (2021). stringi: Fast and portable character string processing in R. R package version 1.7.6, https://stringi.gagolewski.com/

visdat -Tierney N (2017). “visdat: Visualising Whole Data Frames.” JOSS, 2(16), 355. doi: 10.21105/joss.00355, http://dx.doi.org/10.21105/joss.00355.

beanplot - Kampstra P (2008). “Beanplot: A Boxplot Alternative for Visual Comparison of Distributions.” Journal of Statistical Software, Code Snippets, 28(1), 1–9. https://doi.org/10.18637/jss.v028.c01

ggplot2 - Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org

gridExtra - Baptiste Auguie [aut, cre], Anton Antonov [ctb]

dplyr- Wickham H, François R, Henry L, Müller K (2022). dplyr: A Grammar of Data Manipulation. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr

viridis- Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, Pedro A, Sciaini, Marco, Scherer, Cédric (2021). viridis - Colorblind-Friendly Color Maps for R. doi: 10.5281/zenodo.4679424, R package version 0.6.2, https://sjmgarnier.github.io/viridis/.

ggcorrplot- Wei T, Simko V (2021). R package 'corrplot': Visualization of a Correlation Matrix. (Version 0.92), https://github.com/taiyun/corrplot.

tidyverse- Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). “Welcome to the tidyverse.” Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686.

caret- Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams,Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang and Can Candan. (2016). caret: Classification and Regression Training. R package version 6.0-71.https://CRAN.R-project.org/package=caret

randomForest- Liaw A, Wiener M (2002). “Classification and Regression by randomForest.” R News, 2(3), 18-22. https://CRAN.R-project.org/doc/Rnews/

tidymodels- Kuhn M, Wickham H (2020). Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.. https://www.tidymodels.org

car-Fox J, Weisberg S (2019). An R Companion to Applied Regression, Third edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Dataset -

Nishant Verma, N. K. (2020). Vehicle dataset. *Used Cars data form websites*. Retrieved May 2, 2022, from https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho/metadata