Predictive Analytics Project

Project Name: Predicting Pre-Owned Car Prices in India

Last updated: May 9, 2023
Deliverable Number: 4

1. Business Question and Case

1.1. Business Question: Can historical data of past car sales in India be used to predict the sale price of a used vehicle? By creating a model that can predict the sale price of a vehicle, businesses can use this information in considering purchases of potential inventory, in weighing repair expenses or in any other application of weighing the value of a vehicle. Additionally, this service could be sold to consumers to allow them to get a fair value for their trade-in, sale of used vehicle, or purchase of a used vehicle. The tool will be backed by past data and important variables which has affected the price of pre-owned cars in the past in India. In this project, our team attempts to come up with a statistical tool to predict the price of pre-owned car for both potential buyers and sellers and thereby bring transparency and confidence for buyers and sellers.

1.2. Business Case: The pre-owned care market was worth $23 billion[1] in FY 2021-22 and is projected to grow at a double rate at a CAGR[2] of 19.5% till FY 2026-27. The pre-owned car market has remained unorganized in India. The increasing trend in growth is mainly attributed to rapid urbanization, interruptions in the supply chain for manufacturers after COVID-19 pandemic, thereby resulting in delayed deliveries, budget constraints, and macroeconomic uncertainty. Further, COVID-19 fueled the growth of this market with increased demand of pre-owned cars and entry of new firms. However, there is not a standardized tool to predict pricing of pre-owned cars in India and the prices are generally governed by goodwill and industry knowledge of dealers and negotiating are often done without any standards. The business application for such an insight could take many shapes, such as a consumer facing service to give transparency to the car buying process, or, for a business, their ability to predict the sale price of a used vehicle most accurately among their competitors creates a competitive advantage.

2. Analytics Question

2.1. Analytical Goals: What is the effect of kilometers driven, age of vehicle, owner type, price of new car, transmission type etc. on the price of a pre-owned car in India? To what extent each of these variables affect the price of a pre-owned car in India and could affect the accuracy of the model developed. Ultimately the analytical goal is to have predictive accuracy.

2.2. Outcome Variable of Interest: The outcome variable will be the selling price of a pre-owned car which is going to be a continuous variable.

2.3. Main Predictors: The likely key predictors will be new price, model, kilometers driver, age of vehicle, owner type, fuel type, vehicle transmission, car condition, etc. This is a combination of both continuous and categorical variables.

[1] https://www.businesstoday.in/auto/story/heres-why-indias-used-car-market-is-touted-to-double-in-five-years-347737-2022-09-20.

[2] The compound annual growth rate (CAGR) is the rate of return (RoR) that would be required for an investment to grow from its beginning balance to its ending balance, assuming the profits were reinvested at the end of each period of the investment's life span.

3. Data Set Description

The model the team is trying to develop will be a predictive model that will predict the price of a pre-owned car based on pre-stated parameters. The team has obtained primary data on pre-owned cars from Kaggle (link provided in the footnote)[3] and supplemental statistical information from Statista and the team may refer to more data sources if needed later.

In part 1 of the deliverable, the team agreed upon a dataset from Kaggle website to do our subsequent analysis for prediction of used cars prices. Later, after analyzing the dataset the team realized that there's not enough datapoints for us to study. So, team did further research and finalized a new dataset with more predictors and datapoints. The predictors in this dataset are both continuous and categorical, whereas the "Y variable" i.e., the selling price of pre-owned vehicle is the continuous variable. The type of predictor variables includes but are not limited to – New Price, Car Model, Age of Vehicle, Kilometers Driven, Transmission Type etc. The total number of observations are 2,237.

4. Descriptive Analytics

The team performed descriptive analytics in R and the output is in Appendix of this paper. The team performed Descriptive Analytics by creating Correlation Matrix, Histogram, Box Plots for different continuous variables. The summary function was used to describe the dataset of 2,237 observations, including summary of the key variables in the model.

4.1. Descriptive Statistics of Key Variables: The analysis of the "Selling.Price" variable through the histogram and qq-plot indicates that the data is approximately normally distributed with mild skewness to the right. The histogram shows an approximate bell-shaped curve, while the qq-line shows most data points falling close to the line, indicating normality. However, some points at both ends deviate slightly from the line, suggesting some deviation from normal distribution.

4.2. Distribution or Key Variables: The dataset indicates a wide range of values for the selling price, kilometers driven, and age of vehicles, which can be useful in further analysis and modelling. To understand the data distribution and detect any outliers, team further created Box Plots for each variable. The analysis revealed outliers for all variables except "Age.of.Vehicle", indicating that further investigation and analysis is necessary to effectively interpret the data. This information could be useful for making decisions regarding the variables and their impact on the dependent variable.

Based on business rationale the team decides on the following initial predictors:

Variable	Variable Type	UOM (Units of measurement)	Brief Description	Business Rationale
Selling Price (Y Variable)	Continuous	₹	₹ is national currency of India called Indian Rupee	National Currency of India and acceptable and dominate transaction currency in India
New Price	Continuous	₹	₹ is national currency of India called Indian Rupee	The original selling price of a car is an important element for its future selling price

[3] https://www.kaggle.com/datasets/ankits29/used-car-price-data

Kilometers (Kms) driven	Categorical	Kilometers (km)	Km driven by car till date	As a car is driven more, the more its value depreciates.
Age	Continuous	Years	In numbers the age in years	As a car gets aged, its value depreciates with evolving technology and arrival of new models and competitors
Owner Type	Categorical	1st/2nd/3rd	This is the number of Owners through which car has been sold through	As a car is sold several times through different owners its value depreciates
Car Condition	Continuous	Continuous	The car is allocated a rating from 1.0 to 5.0 (in 0.1 increments) where 1.0 indicates that the vehicle is not in a bad condition and 5.0 is the best condition.	Based on condition of the car a seller allocates a rating on a scale of 1.0 to 5.0 and is a subjective rating allocated based on holistic condition of the car and on wisdom and judgement of the seller
Transmission	Categorical	Automatic/ Manual	Describes the transmission type of a car	In India both type of cars are prevalent, and the type of transmission is one of the factor deciding the car cost.

4.3. Correlation and Co-Variation Analysis: The team performed the correlation matrix to better understand the relationships between the variables in the dataset and to identify potential correlations that could be used in further analysis or modelling. The team found that car condition has a negative correlation with “Age.of.Vehicle” with a correlation of -0.65 and “Car.Condition” with a correlation of -0.45. “Current.Price” shows a positive correlation of 0.51 with “Selling.Price”, which basically aligns with the industry understanding that the new price should highly correlate with the selling price of a vehicle. Finally, “Age.of.Vehicle has a positive correlation with of 0.43 with “Kilometers.Driven” and negative correlation of -0.57 with “Selling.Price”.

4.4. Data Pre-Processing and Transformations: The data set contained several data elements which needed data pre-processing and transformations as follows:

4.4a Data Pre-Processing:

- ✓ *Current Price:* We standardized the data by converting values in "lacs" to absolute numbers for consistency.
- ✓ *New Price Availability:* We filled in missing data for new prices (28 entries) by conducting internet research on the corresponding vehicles and models.
- ✓ *Transmission Type:* 158 data sets had unknown values of Transmission Type. We pre-processed the data by filling the transmission type as manual or automatic based on the internet research for the corresponding vehicles and models.

4.4b Data Transformations: After doing the initial OLS regression and performing the CI and VIF tests on both fit and full model specification, we understood that there are issues of multi-collinearity in the model. To resolve this, the team standardized the OLS model.

5. Modeling Methods and Model Specifications

5.1. Initial Model Specification

To begin the model specification, the team started with two OLS models, the ‘fit’ model, which included only the continuous variables, and the ‘full’ model which included the categorical variables in addition to the continuous variables. This step's intention was to understand the predictive accuracy of a smaller model that only used the continuous variables and then to understand the variation in a model that then used the categorical variables as well.

5.2. Initial OLS

The original model generated was an OLS model using exclusively the continuous variables within the dataset (Age of Vehicle, Car Condition, Kilometers driven, and New Price). This resulted in a model with a R-squared value of 0.6282 with Age of Vehicle having a statistically significant inverse relationship to selling price, where, on average and all else equal, a 1 year increase in the age of the vehicle results in a ₹43,830 decrease in the selling price of a vehicle with a significance of $p < 0.001$. Additionally, Car Condition and Original Price were found have a positive relationship with a significance of $p < 0.001$, where, all else equal and on average, a 1 point increase on car condition, rated on a scale of 0 – 5, would result in an increase of ₹68,550, and an increase of ₹1 to current price, all else equal, increase the sale price of the vehicle by ₹0.2149 which is highly insignificant.

In reviewing the linear model that includes "Owner" and "Fuel.Type" variables, the team found that the Second and the Third Owner has no statistical significance in that model as compared with the First Owner, and that the categorical variable Fuel.type has statistical significance for the Petrol and Petrol+CNG relative to the reference value of Diesel, with a statistical significance of $p < 0.001$. CNG, or Compressed Natural Gas, is an emerging fuel alternative in the Indian car market, with Petrol+CNG vehicles having a hybrid engine that can run on either type of fuel. Relative to a diesel vehicle, and on average and all else being equal, a Petrol vehicle will sell for ₹91,900 less, and a Petrol+CNG will sell for ₹122,700 less. The addition of these variables results in a model with a R-squared value of 0.6613, which is roughly a 5.3% increase in the variance of sales price these models are able to predict.

With both linear models Low R-squared it was not likely for there to be a risk for multicollinearity. To confirm this, we performed a VIF test and each predicror resulted in a score below the threshold of 10, which is a general indicator that VIF values below to generally mean there is no risk of multicollinearity.

5.3. Assumption Tests

The team performed various assumption tests for the initial OLS model. The team created a histogram and a qq-plot to understand the normality in the Y variable(Continuous variables). Team understood that Y variable is not normally distributed and there is a certain degree of non-normality at the tail ends.

Team also created correlation plot to check the correlation among the predictors, and as highlighted in para 4.3, there is a high positive and negative correlations between few variables.

Team performed the Condition Index and Variance Inflation Factor tests to check for the multi-collinearity in the OLS model. From the results team understood that the CI was greater than 50 in OLS models, but the individual VIFs were less than 10. This basically meant that the collinearity was not introduced by the predictor variables, but the intercept.

5.4. Model Candidates and Rationale

The team considered multiple models to adjust for the OLS violation and multicollinearity. To address the non-normally distributed error, the team considered log-log and log-linear models. To control for the multicollinearity, a stepwise regression was run, with all predictors except for Owner having statistically significant results ($p < 0.05$). Because the multicollinearity was believed to be with the constant, in this case likely indicative of an issue with the scale of some of the predictors in the model, the team also considered a standardization transformation. We thought this was especially relevant not only because it would address the issue of predictors of dissimilar scale among in the model, the Car Condition variable is numerical, on a scale from 1.0 to 5.0 in 0.1 increments, so the true effect size of this variable had no meaning. For those reasons, a GLM model using a Standardization transformation was selected. Because the standardized translation resulted in only a marginal increase of the R-squared value of the model (more on this in the analysis), the team also selected a PCR (Principal Components Regression) model to test as a control of multicollinearity unresolved by the transformation.

For the non-parametric model to be tested, the selected random forest model as the model specification better control for the dominance of a single variable relative to a bagging model. Additionally, an advanced decision tree model is useful for predictive accuracy.

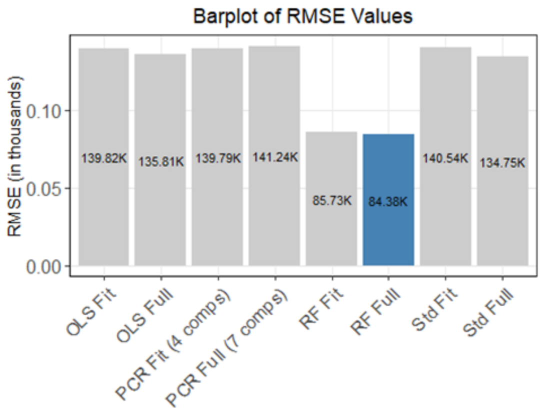
Sequence	Modelling Method	Specifications
1	OLS Model	Fit Model
2		Full Model
3	Standardization Model	Fit Model
4		Full Model
5	PCR Model	Fit Model
6		Full Model
7	Radom Forest (Non-parametric Model)	Fit Model
8		Full Model

5.5. Model Specification Candidates and Rationale

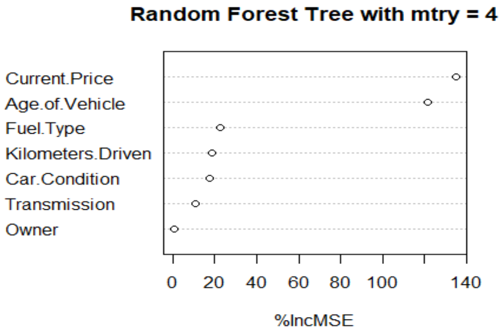
Two separate specifications were applied to each model. The same specifications from the original OLS models were use, the ‘fit’ model that only made use of the continuous variables, and the ‘full’ model that included categorical variables.

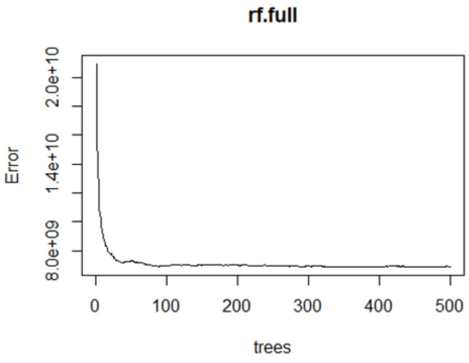
5.6. Cross-Validation Testing and Final Model Selection

K-Fold Cross validation ($K = 10$) was done to test each of the models and specifications for accuracy, and the RMSE (Root Mean Square Error) was calculated and used to measure the accuracy of the model. The RMSE in this case is the variance in the predicted variable unexplained by the model. The model and specification that yielded the strongest result was the random forest full fit. See below for full results.



The Random Forest of the full model specification resulted in the best result, with a RMSE of 84,380. This model used a M = 4 and had a predictive accuracy of 86.7%. The resulting variable importance is included below, and indicates that Age of Vehicle and Original Price of vehicle (Current.Price) being the two most important variables to include in the model.





```
##                      %IncMSE
## Age.of.Vehicle    121.2746005
## Car.Condition      17.6452243
## Kilometers.Driven  18.6990652
## Current.Price     134.7914906
## Owner               0.8726381
## Fuel.Type          22.6665210
## Transmission       10.9226760
```

The team also tested a random forest model with a M = 3 specification as a tuning parameter, with the results included in the final R output, but found that it did not improve the predictive accuracy.

6. Analysis of Results

In reviewing the results of the cross-validation testing, the team concluded that the Random Forest resulted in the lowest RMSE and was the model with the best predictive accuracy, with age of vehicle and original sale price being the two most important variables. To expand on these results, the team recommends that a larger data set, with both more rows of data and more predictors be given a similar treatment. Random Forests are effective when the number of predictors in the data set is large, and additional predictors could result in a model with stronger predictive accuracy. Turning from the non-parametric model to the model with the next best RMSE, the standardized transformation GLM model had the best predictive accuracy of the linear models evaluated, with the lowest RMSE and highest R-squared value with an R-squared of 0.6613. Additionally, this model found with statistical significance of $p < 0.05$ that all variables, with the exception of the LPG fuel type, to be of statistical significance.

7. Conclusions and Lessons Learned

7.1. Conclusions from the Analysis

The team concluded that of the models tested, the random forest had the best predictive accuracy, but that after testing OLS, a GLM model with standardized transformation, and the regression tree random forest, two measures should be taken for further analysis. These two are 1) conduct a similar analysis with a larger data set, both in samples and predictors, and 2) test a non-linear model. The models tested, while each working to address issues from the original OLS model, were linear models and did not sufficiently address the tail-wagging errors in the model.

7.2. Project Issues, Challenges and Lessons Learned

The team faced several issues and challenges while working on this project which are listed below:

- ✓ <u>Data Preprocessing and Transformations:</u> The team find challenging to fill missing value of 28 entries which required internet research and industry knowledge to correctly map the correct price for missing values. On further analysis and investigation found there were around 35 entries in the database which had Zero value of Current Price. The team ultimately captured Current Prices in the database as per internet research and accordingly the R-code could generate the output as desired.
- ✓ <u>Project Flow:</u> After the initial analysis on OLS model, creating a logical flow on how to proceed working with different models by remaining in the bounds of the model assumptions.
- ✓ <u>Uncertainty and Iterations:</u> Throughout the project team worked on a limited dataset and evolved its strategy to come up with an appropriate model meeting the needs of the client/project.
- ✓ <u>Team Collaboration:</u> The team learnt collaboration issues and learnt how to resolve them. Team strived to work on each team member’s strength and learnt ways to collaborate.

Appendices

A. Data Information

1. Summary of dataset

```
summary(cars.data)
```

```
##     Model                Year       Current.year  Selling.Price
##  Length:2237        Min.   :2006   Min.   :2021   Min.   :  75299
##  Class :character   1st Qu.:2012   1st Qu.:2021   1st Qu.: 272099
##  Mode  :character   Median :2014   Median :2021   Median : 355799
##                     Mean   :2014   Mean   :2021   Mean   : 418443
##                     3rd Qu.:2016   3rd Qu.:2021   3rd Qu.: 503299
##                     Max.   :2020   Max.   :2021   Max.   :1952397
##  Age.of.Vehicle   Car.Condition   Kilometers.Driven Current.Price
##  Min.   : 1.000   Min.   :3.000   Min.   :   913    Min.   : 190000
##  1st Qu.: 5.000   1st Qu.:4.200   1st Qu.: 32137    1st Qu.: 498000
##  Median : 7.000   Median :4.300   Median : 55430    Median : 594000
##  Mean   : 7.236   Mean   :4.371   Mean   : 61929    Mean   : 773230
##  3rd Qu.: 9.000   3rd Qu.:4.600   3rd Qu.: 83427    3rd Qu.: 892242
##  Max.   :15.000   Max.   :5.000   Max.   :855881    Max.   :8639399
##     Owner            Fuel.Type         Transmission        Insurance
##  Length:2237        Length:2237        Length:2237        Length:2237
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Current.Price..Unformatted. Price.Formatting
##  Length:2237                 Length:2237
##  Class :character            Class :character
##  Mode  :character            Mode  :character
```

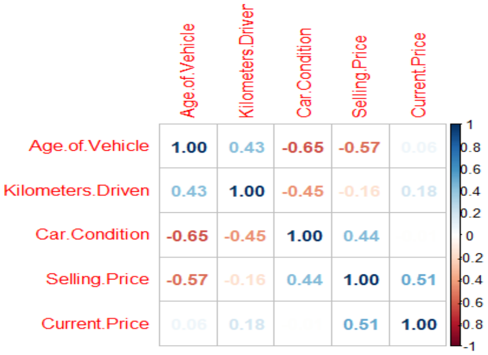
2. Class of variables

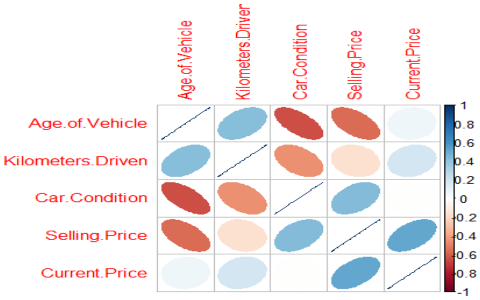
```
class(cars.data)

## [1] "data.frame"

class(cars.data$Selling.Price)

## [1] "integer"

class(cars.data$Kilometers.Driven)

## [1] "integer"

class(cars.data$Age.of.Vehicle)
```

```
## [1] "integer"
class(cars.data$Owner)
## [1] "character"
class(cars.data$Current.Price)
## [1] "integer"
```

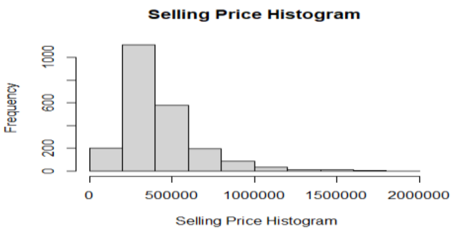
B. Visuals, Graphs and Plots

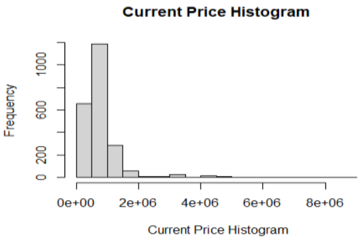
1. Correlation plots

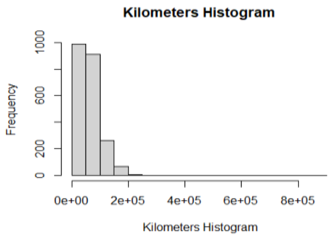




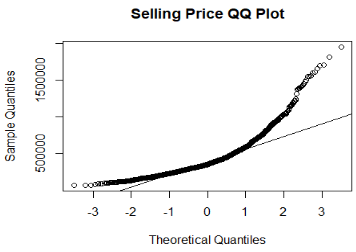
2. Variables histogram

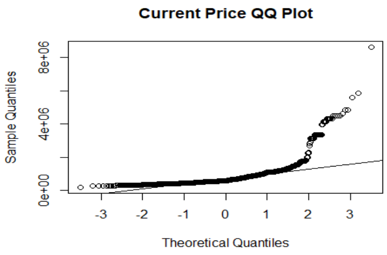


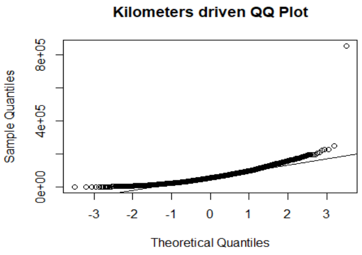




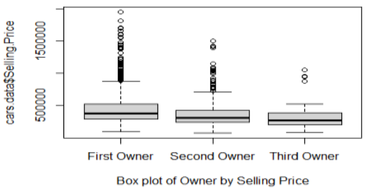
3. QQ-plots

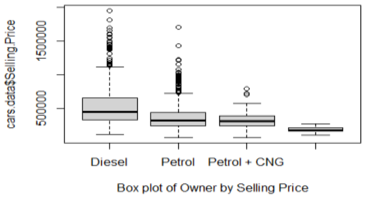


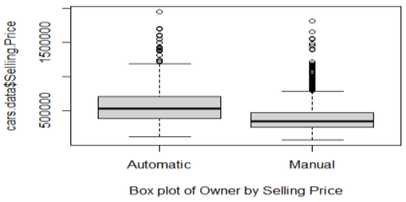




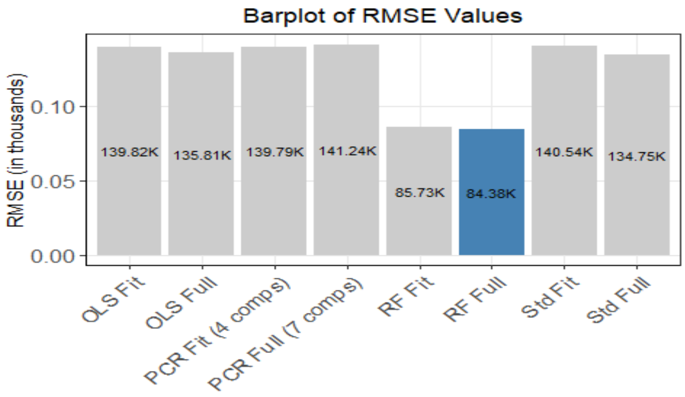
4. Box plots







5. RMSE bar plot



C. Quantitative R Output

```
importance(rf.full) # To view the importance of each variable

##                       %IncMSE IncNodePurity
## Age.of.Vehicle    121.2746005  3.482804e+13
## Car.Condition      17.6452243  8.867213e+12
## Kilometers.Driven  18.6990652  6.456076e+12
## Current.Price     134.7914906  5.643736e+13
## Owner               0.8726381  6.285300e+11
## Fuel.Type          22.6665210  4.960587e+12
## Transmission       10.9226760  1.836423e+12

importance(rf.full, type = 1)

##                       %IncMSE
## Age.of.Vehicle    121.2746005
## Car.Condition      17.6452243
## Kilometers.Driven  18.6990652
## Current.Price     134.7914906
## Owner               0.8726381
## Fuel.Type          22.6665210
## Transmission       10.9226760
```

```
mse.rfull <- mean(rf.full$mse)
rmse.rfull <- sqrt(mse.rfull)
cbind( "RF FULL RMSE" = rmse.rfull)

##      RF FULL RMSE
## [1,]     84380.08
```

RMSE Result Summary

	x
PCR Full (7 comps)	141245.00
Std Fit	140544.37
OLS Fit	139821.76
PCR Fit (4 comps)	139794.00
OLS Full	135809.76
Std Full	134750.57
RF Fit	85727.19

D. Other

E. References

https://www.businesstoday.in/auto/story/heres-why-indias-used-car-market-is-touted-to-double-in-five-years-347737-2022-09-20.

https://www.kaggle.com/datasets/ankits29/used-car-price-data

https://towardsdatascience.com/predicting-used-car-prices-with-machine-learning-techniques-8a9d8313952