Capstone 1

Automobile Price Prediction - Indian Market

Final Report

1. Introduction

In 2020, India was the fifth-largest auto market, with ~3.49 million units sold in the passenger and commercial vehicles categories. It was the seventh largest manufacturer of commercial vehicles in 2020. The 4-Wheeler Industry in India was ranked fourth (taking over Germany) and is expected to surpass Japan by the end of 2021.

Nearly 4 million units of four wheeler vehicles were sold in the year 2020 and are expected to grow at a rapid pace. Automobile exports grew at a rapid rate of 14.5% during FY 20 in India.

With such a lucrative environment for growth and opportunity, it is natural for international players to set up their foundations in India and profit from this huge economy.

My project is aimed to solve the challenging and complex problem of pricing in the Indian Automobile Industry.With around 1500 model variants floating in the Indian Market, it is essential to get the pricing correct when launching a new vehicle.

If a manufacturer misses the mark on price and over-prices its model, it is very likely that sales will suffer. On the other hand, if the prices are too low, the company can miss out on potential profit margins. Hence it is a must that the right note is hit with the pricing.

In this project I use an available data set of over 1200 car models in the Indian Industry in the year 2020 and perform an in-depth analysis from business analytics, using EDA to get an insight on how different factors weigh in on Price and some competition analysis, to a full blown linear regression model to predict the price of a new model with an accuracy of just under 90% on the test set.

There are of course a lot of different techniques which can improve our perspective on the Pricing and I suggest one such classification technique as a means of further improvement towards the end of the Project.

2. The Problem

2.1. The Problem Statement

Our target business is a French car manufacturer **Peugeot Automobiles**. Peugeot aims to enter the Indian Market by setting up a manufacturing unit in India and are planning to launch a new car model in the near future in competition with their European and American counterparts.

They have contracted an automobile consulting company to understand the factors on which the pricing of cars depends. Specifically, they want to understand the factors affecting the pricing of cars in the Indian market, since those may be very different from the French market.

The Company wants to figure out:

- Which variables significantly affect the price of a vehicle in the Indian Market

- How well do these variables explain the price

Based on a large scale market survey, the company has collected a database of over 1200 cars across the Indian market and now it is upto us to find the solution to their problem.

2.2. The Opportunity (Benefit to the customer)

Pricing is one of the most important aspects for the success of a product, especially in a price competitive market like India. If the company misses the mark on price and over-priced the model, it is very likely that its sales will suffer. On the other hand, if the prices are quoted to low, the company can miss out on potential profit. Hence it is a must that the right note is hit with the pricing. This is where we come in and help the company achieve its targets with the new car model.

2.3. The Solution

For the pricing model, we are going to use the database available for more than 1200 car models in the Indian market. We will attempt to determine the price by using a linear regression model. We will try different kinds of linear regression models to ascertain the best fit and test it on real-world data.

3. The Dataset

The dataset chosen is the Indian Cars Dataset from Kaggle (https://www.kaggle.com/medhekarabhinav5/indian-cars-dataset)

The dataset has reliable data though some amount of wrangling will be required due to missing data.

It contains a variety of features such as Model price, Engine related features, Body and comfort related features etc (Over 140). This makes it the optimal dataset for making a pricing model.

4. Data Wrangling

The following steps were performed to prepare the dataset for use:

4.1. Shortlist features based on usability and common sense:

Due to the sparsity of the dataset, 24 features are pre selected based on the kind of questions one could expect a prospective buyer to answer so that a reasonable prediction vector can be established. The choice of features were also based on the amount of information available for the features and a common sense understanding of correlation between the variables. Variables that had very little data (<50% of the dataframe size) or would not matter to a prospective buyer were removed.

Out of the 24 features selected, 11 were numerical, 10 were categorical and a combination of 3 would be used as the target variable for the clustering phase.

4.2. Addressing the target variables.

There were a few issues in the 'Make' and 'Model' entries, these were addressed by string search, splitting and replacing using the pandas module.

4.3. Cleaning and extracting the numerical features:

- To extract the numerical entries, the numerical columns were checked for string pattern consistency (since the units were a part of the numerical string entries). This was done using string comparisons with regex.
- Those with inconsistent entities were replaced by NaN entries.
- All the pattern consistent entries were then extracted using str accessors and regex.
- The numerical features were converted from object dtype to numerical dtype (float or int as required).

4.4. Filling the missing values in numeric features:

- The first attempt was to fill in the missing values by averaging by 'Model'. This would have ensured that accurate values are filled in. But this had the drawback of not filling up any if all the entries for a particular model and feature were all missing.
- Second attempt was to groupby 'Body_Type' and fill in the missing values by the median of 'Body_Type' for features that are functions of the Body style of a car. These are: Kerb_Weight, Ground_Clearance, Seating_Capacity, Wheelbase and Boot_Space
- The missing values for 'Mileage' were tackled by filling on an average by Body_Type and Fuel_Type. It was then converted into a more Robust feature: 'Cost_per_Km'.
- And finally 'Displacement' was tackled by sorting the dataframe using 'Power' and then interpolating.

4.5. Cleaning up the Categorical Variables:

- First, the unique values of the categorical variables were examined and the replaced for either duplication or relative usability to the customer.

  For example, for the category - 'Drivetrain', the categories of All_Wheel_Drive was replaced Four_Wheel_Drive since for practical purposes, both serve the same purpose.

- The missing values in the categorical variables were then filled out by taking the mode of the feature grouped by ‘Make’. Since these variables would roughly remain the same within the bounds of a manufacturer.

  These are: 'Drivetrain', 'Emission_Norm', 'Engine_Location', 'Fuel_System', 'Seats_Material', 'Transmission_Type'.

4.6. Removing Outliers:

The major outlier category that we are concerned with is ‘Price’. Primarily because we are predicting the price is our main motive.

When we remove the outliers, we see that the price still reaches Rs. 4 Cr. mark. We cut this down further and limit our database to cars that fall within the Rs. 2 Cr. range.

We remove outliers from the numerical predictor variables using scipy.stats package for observations with a z-score > 3.

We will later also remove some outliers during our linear regression stage to obtain linearity with the dataset. We will also transform the target variable ‘Price’.

With this the dataset is clean and ready for use.

5. EDA and Statistical Techniques

This report summarizes the Exploratory data analysis employed to get an initial view into the relationships between the predictor variables and the target variable (‘Price’).

We explore the data using bar charts and scatter plots and derive intuition from the visuals.

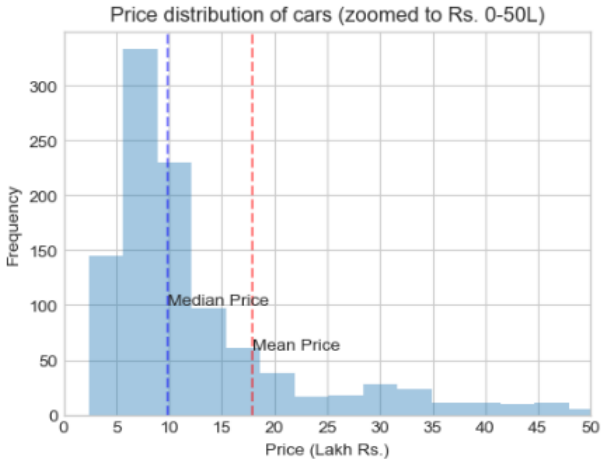
We will then use statistical methods to ascertain the trends from our visual observations, identify issues like multicollinearity, assess the importance of numerical and categorical variables and finally perform feature selection using the information gathered.

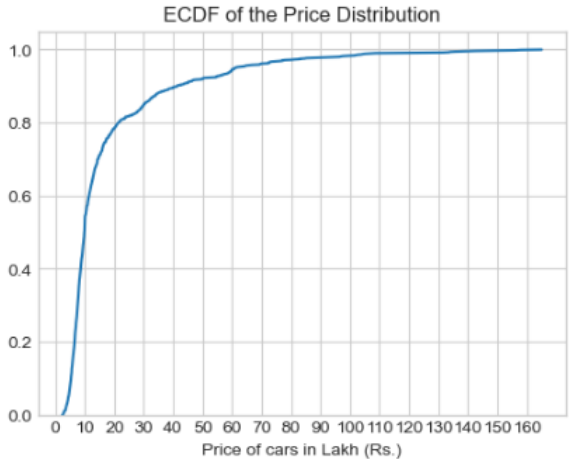
5.1. EDA Results

On completion of our EDA, we had the following observations about our data:

Price:

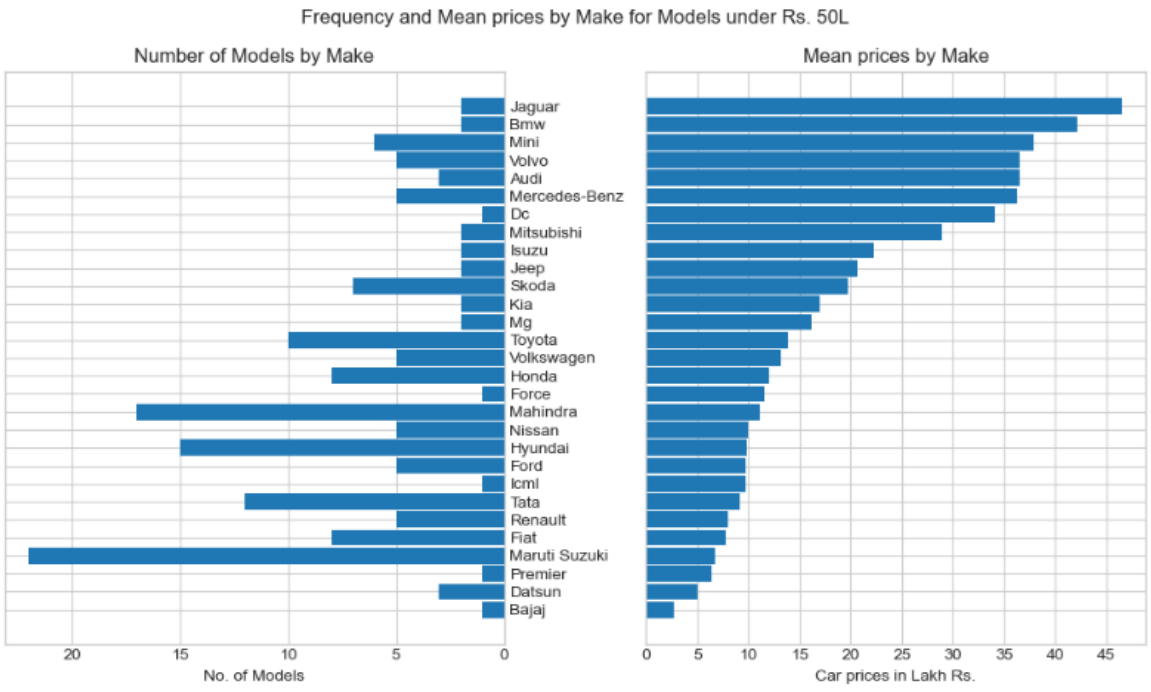
- The Price of the cars had a right-skewed distribution with 50% of the cars falling within the Rs. 10 Lakh mark.
- 80% of the cars fell within the Rs. 25 Lakh mark.
- The spread of the distribution is large with a large standard deviation.





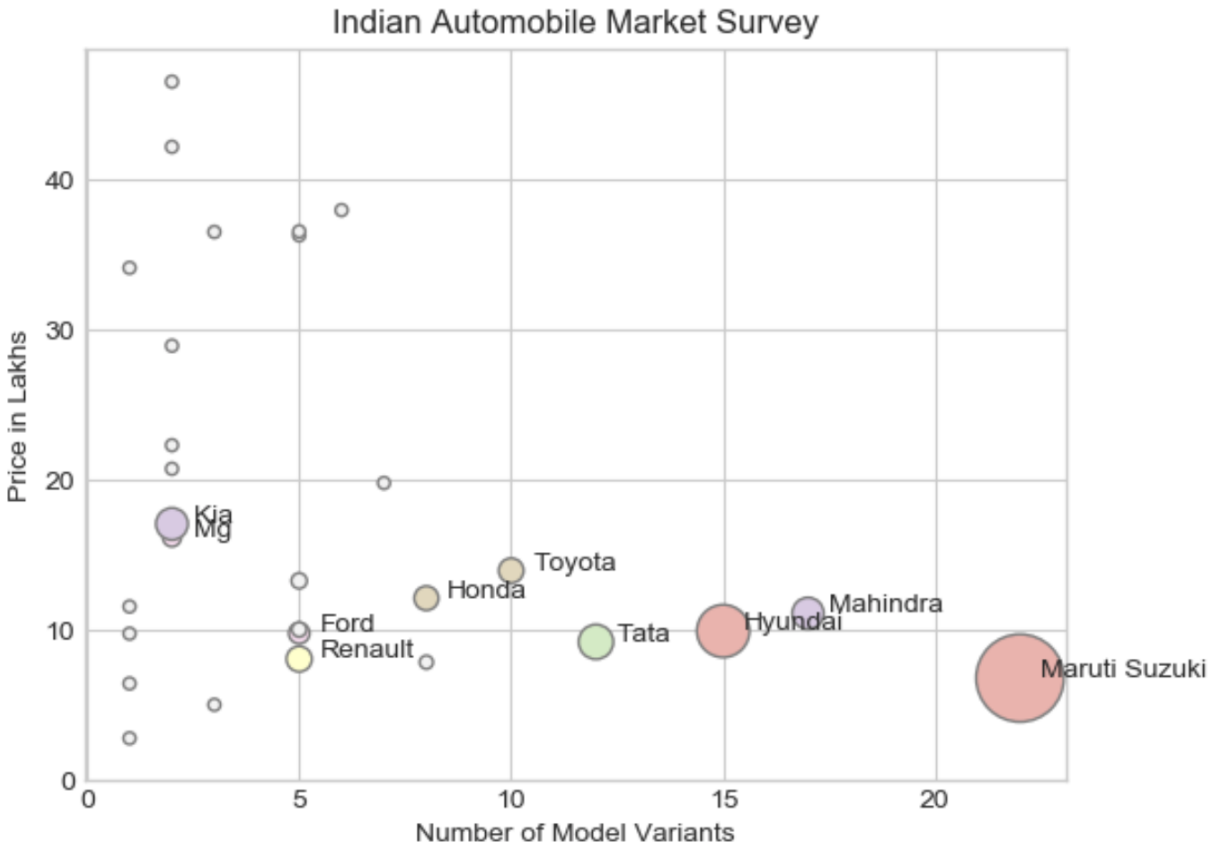
The Competition:

- We see a mix of product and pricing strategies in the Indian automobile market.
- Maruti Suzuki is the leader in the low price segment (< Rs. 7 Lakhs) and offers the largest range of model choices.
- Companies like Tata, Hyundai, Mahindra compete in the mid-segment with the mean model price ranging from Rs. 7-9 Lakhs. They also offer a wide variety of models.



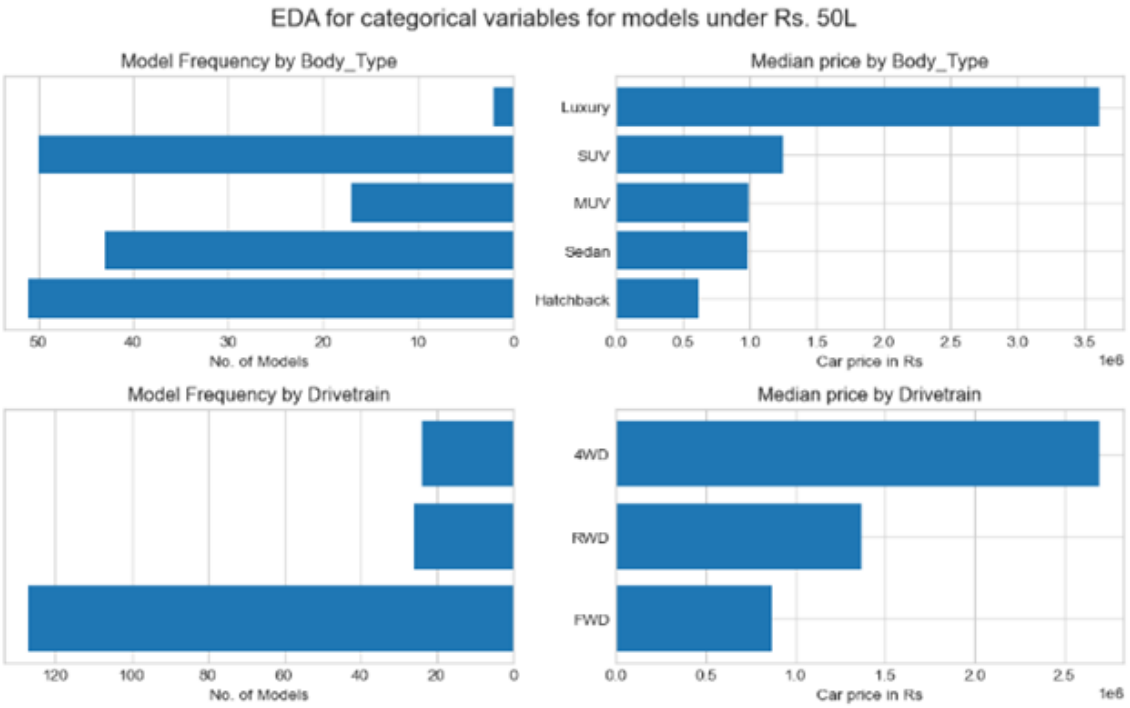
We see a pattern emerging when we include the market share of different companies:

Companies that offer a higher variety of models and low average pricing tend to dominate in terms of market share.



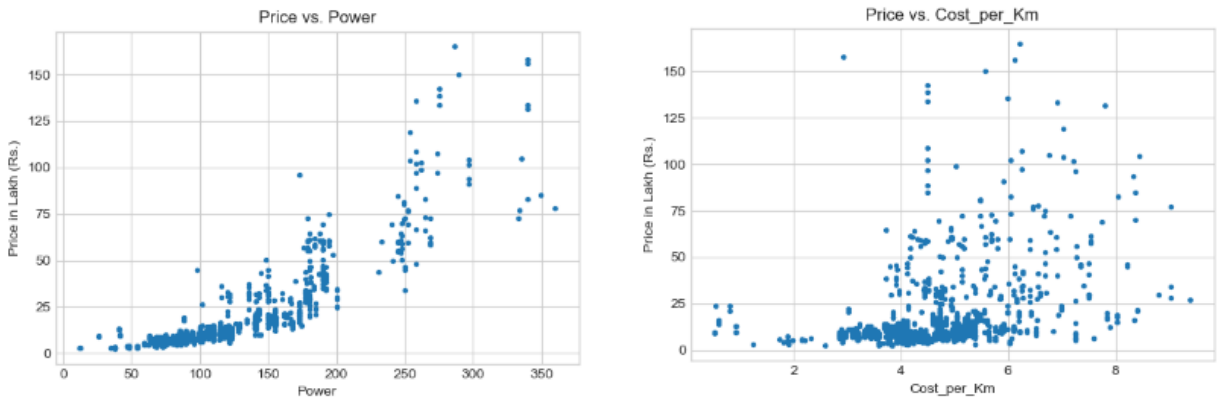
Significant relationship observed with Categorical Variables:

- The SUV segment has the same depth of variety as the Hatchback segment. Which is counter intuitive, as the median price of SUV's is twice that of hatchbacks.
- Models with Four wheel drive have a median price range almost double that Rear wheel drive models
- Electric vehicles are the most economical to run while also being the costliest on average and the maximum depth of models is found in petrol, followed by diesel



Significant relationship observed with Numerical Variables:

- The clearest trend wrt the target variable 'Price', amongst the numerical variables is seen with 'Power'. Similar trends can be observed between Price and Engine-Displacement and Torque.
- 'Kerb Weight' and 'Wheelbase' also show a positive trend with 'Price'
- 'Cost per Km' and 'Number of Airbags' shows a slight positive trend with price.



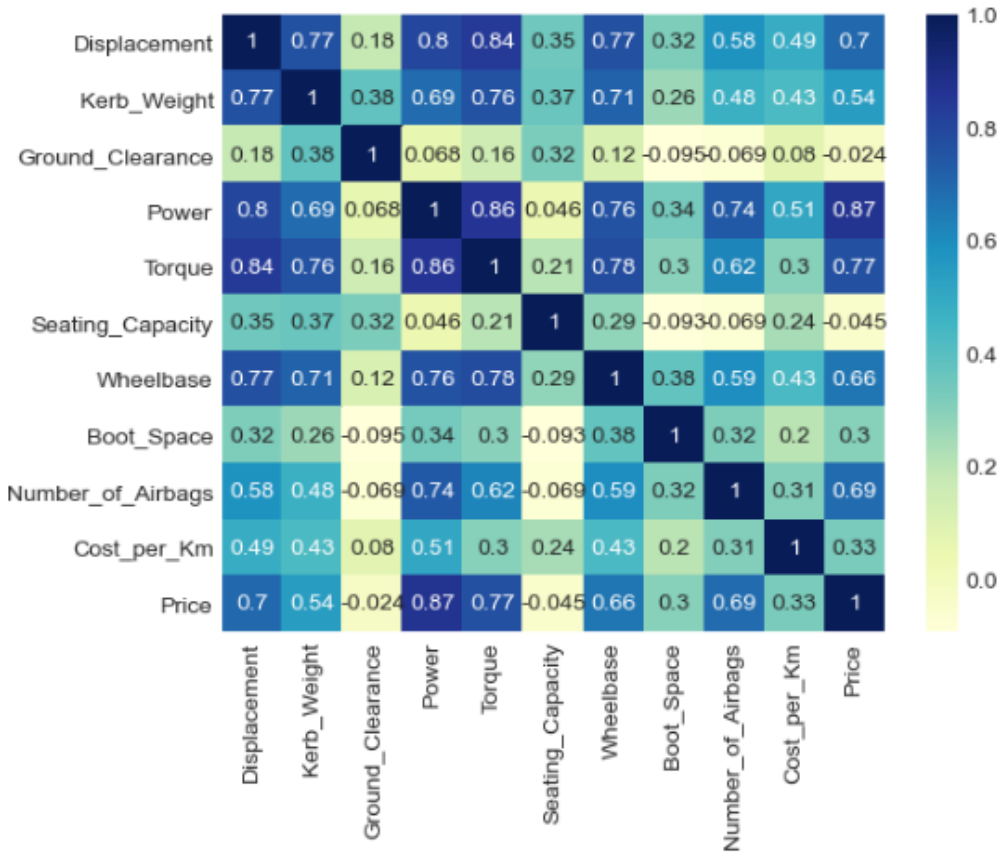
5.2. Statistical Techniques used:

Pearson's correlation coefficient - Correlation between numerical predictors and response variable.

In order to assess the correlation between the target variable 'Price' and the numerical variables, as well as to assess multicollinearity between various numerical variables, we use a correlation heat map using pearson's correlation coefficient.

We use the DataFrame.corr() function getting the correlation matrix.

We observe the following heat map and drop all variables except 'Power' and 'Cost per Km' due to the high degree of multicollinearity between the predictors.



ANOVA - Impact of categorical variables on Price.

ANOVA or Analysis of Variance is used to assess which categorical variables have a significant impact on the target variable 'Price'.

ANOVA is performed when one of the variables is a numerical variable while the other one is categorical. ANOVA uses the f-test to determine if the numerical means of the categories of the categorical variable come from the same distribution or not.

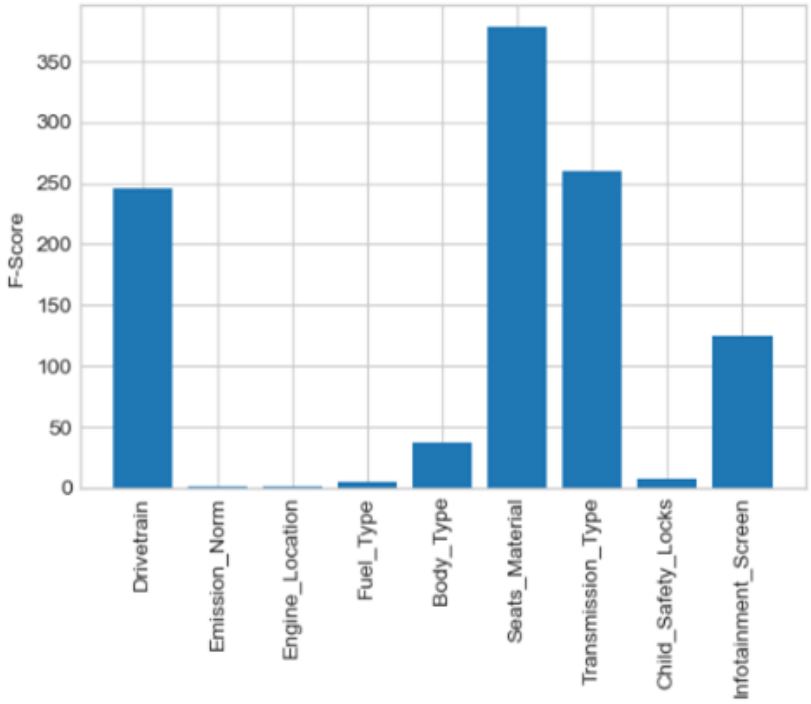
The null hypothesis : There is no difference between the means (obtained from the numerical variable) of the categories of the categorical variable.

The alternate hypothesis : The mean of at least one category is significantly different.

An F-statistic, or F-test, is a class of statistical tests that calculate the ratio between variances values, such as the variance from two different samples or the explained and unexplained variance by a statistical test, like ANOVA.

We use the **f_classif** function from the **sklearn.feature_selection** library for the same.

The results of performing ANOVA between the target variable 'Price' and the categorical variables can be seen from the graph below.



Tukey-HSD - Which categories are statistically significant?

ANOVA tells us if at least one of the categories is statistically significant in predicting a numerical variable (or vice-versa) but it does not tell us which of the categories are significant. In order to further short select our categorical variables, we perform a Tukey-HSD test on the variable 'Body_Type' to see which body types are actually significant.

We use **statsmodels Multicomparison** for the same. The following is the snapshot of the results:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Hatchback	Luxury	6876433.1846	0.001	4819804.7722	8933061.5971	True
Hatchback	MUV	542945.6981	0.3525	-264443.8902	1350335.2864	False
Hatchback	SUV	1611187.0738	0.001	1095950.0615	2126424.0861	True
Hatchback	Sedan	1357289.097	0.001	807518.3951	1907059.799	True
Luxury	MUV	-6333487.4865	0.001	-8474082.9973	-4192891.9758	True
Luxury	SUV	-5265246.1108	0.001	-7313596.7576	-3216895.4641	True
Luxury	Sedan	-5519144.0876	0.001	-7576452.7708	-3461835.4044	True
MUV	SUV	1068241.3757	0.002	282176.5613	1854306.1901	True
MUV	Sedan	814343.3989	0.0477	5222.5561	1623464.2418	True
SUV	Sedan	-253897.9768	0.6449	-771843.6977	264047.7441	False

We see that **pairs** Hatchback-MUV and SUV-Sedan do not reject the null hypothesis with p-values > 0.05 telling us that there isn't a significant difference in mean prices of Hatchbacks and MUVs as well as SUVs and Sedans.

We will still retain all the categories of Body_Type since overall, all categories have significant differences with at least one other category.

Using the above statistical techniques, we shortlist our feature set to the following predictors:

1. Power
2. Cost_per_Km
3. Drivetrain
4. Body_Type
5. Seats_Material
6. Transmission_Type

6. In-Depth Analysis and Machine Learning

Out of the various choices of machine learning models, the linear regression model was chosen for analysis. The justification of the choice of model lies in the fact that the response variable (the target) is a continuous variable whilst the estimators are numerical as well as categorical.

Hence we proceed with the linear model.

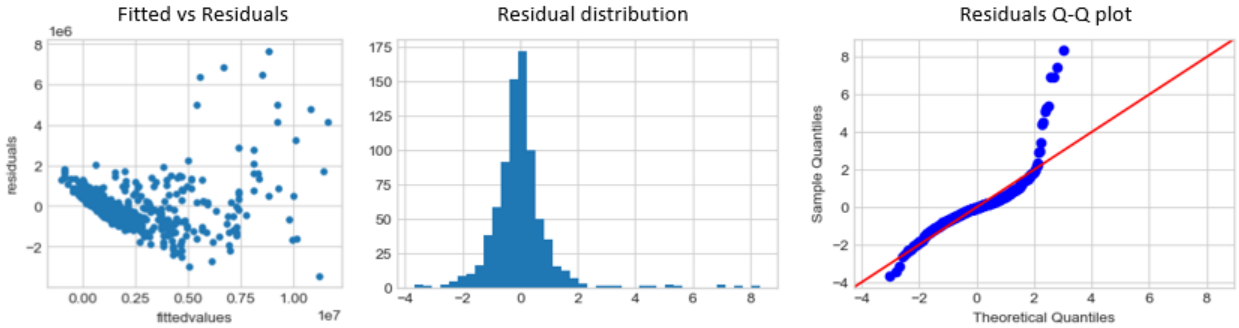
6.1. Verification of the Linear Model assumptions, transformation of data & removing outliers

The first thing we look into is verification of our linear model assumptions:

1. Linearity
2. Independence of observations
3. Homoscedasticity of Residuals
4. Normality of Residuals

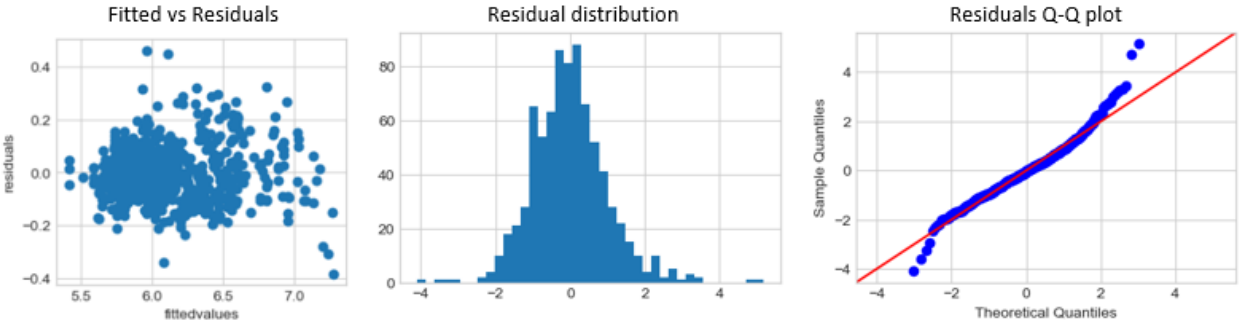
We check the above by building a makeshift model using statsmodels and plotting the residuals and predicted values. We quickly observe that the assumptions of linearity and homoscedasticity are violated.

First attempt model (Vanilla):

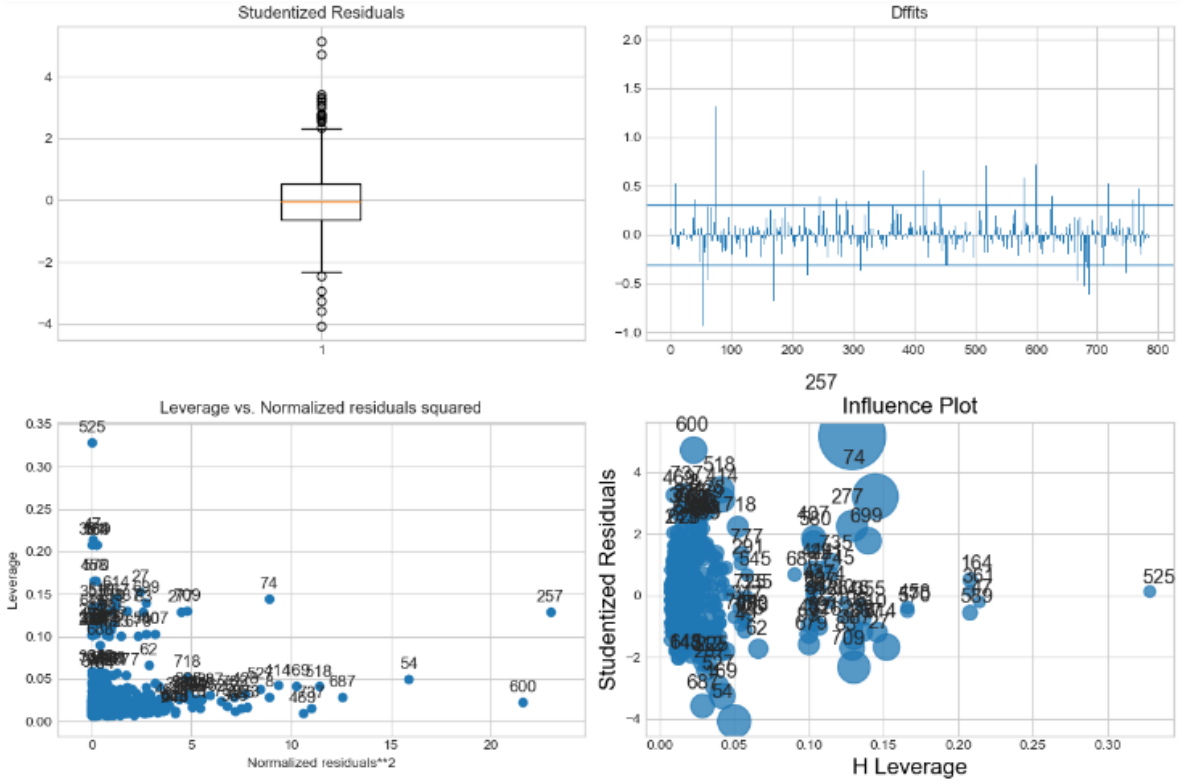


To correct the problems with linearity, we log-transform the target variable and check the plots again. This time we see that the problems of linearity and homoscedasticity are nearly resolved but the effect of outliers becomes prominent

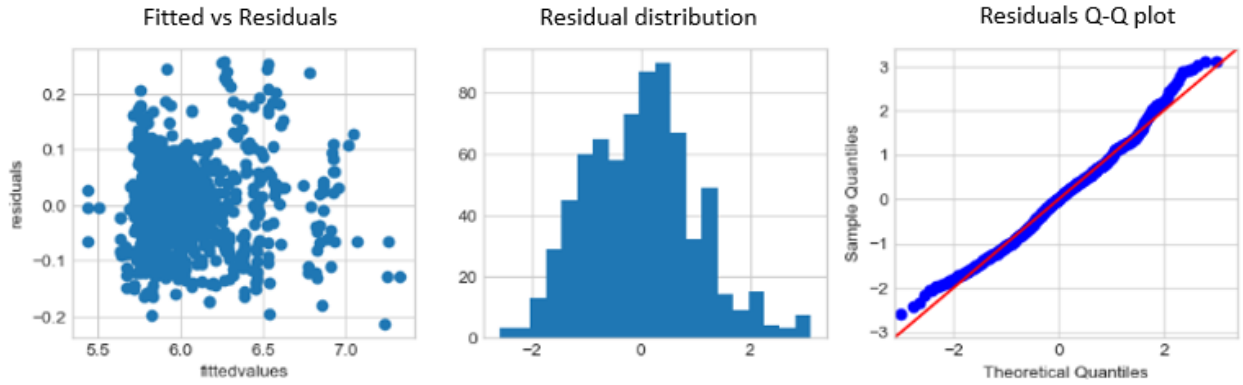
Target variable transformed:



Next we identify the outliers by looking at the studentized residuals, leverage and dffits and remove the same by limiting these values to the industry standards.



Log-transform of the target variable and elimination of outliers gives us a very well behaved linear model with good model fit statsitics.

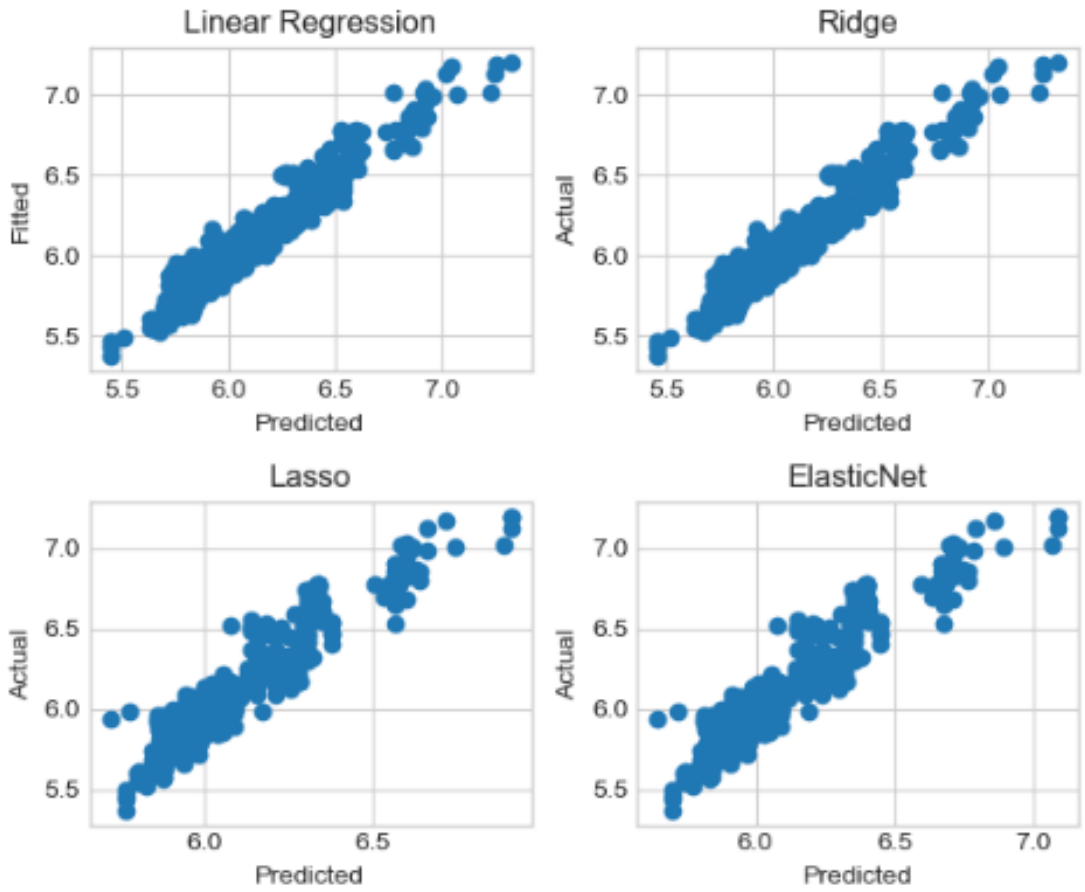


6.2. Final Model, Test scores and feature importance:

Next, we try out different linear regression models i.e. Vanilla linear regression, Ridge, lasso and Elasticnet using sklearn.

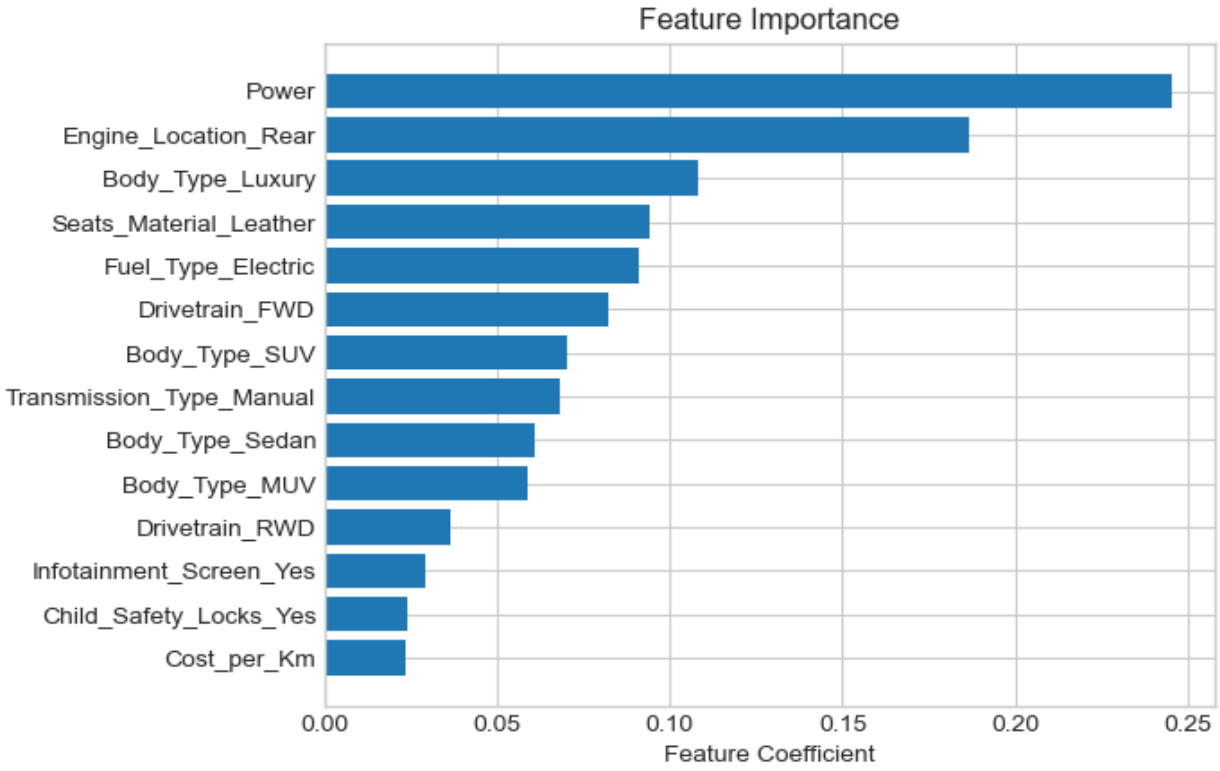
We use the linear_model library for the models and use GridSearchCV to tune the hyper-parameter alpha in order to obtain the best fit for these models and choose the one with the highest test set accuracy.

```
Comparison of accuracy
            Model  train_R2   test_R2         MSE
0  Linear Model  0.930558   0.89769  0.00698896
1         Ridge  0.930535  0.897905   0.0069912
2         Lasso  0.769006  0.724688   0.0232482
3    ElasticNet  0.838993  0.793745   0.0162044
```



Based on the results above, we choose Ridge regression as our preferred choice of linear regression algorithm with a test set accuracy of ~90%

Below we see the feature importance by comparing the coefficients of the linear regression model:



We test our model against unseen data from 3 new car models launched after the collection of the dataset:

1. Hyundai Creta S - Ex-showroom Price: 12.19 Lakhs
2. Kia Sonet GTX Plus Turbo DCT DT - Ex-Showroom Price: 13.09 Lakhs
3. Nissan Magnite Turbo CVT XV - Ex-showroom Price: 8.99 Lakhs

Our model performs fairly well predicting the price of these models when we feed it with the relevant feature information.

	Actual Price (Rs.)	Predicted Price (Rs.)	Offset %
Creta	1219000	1273909	4.504452
Sonet	1309000	1304528	-0.341583
Magnite	899000	934775	3.979445

7. Conclusions

As we can see, the model works well on the test data set (~90%) accuracy as well as on unseen data.

Power is the most significant factor in determining the Price of an automobile in the Indian Industry. Followed by engine location and body type.

The developed linear regression model has a test set prediction accuracy of 89% hence can be used by the customer to ballpark the entry price point of their new launch.

Further improvement Suggestions:

We can explore new avenues and new ways to approach this problem of price estimation by including techniques like classification in our work.

Due to the nature of linear regression, we have to remove quite a few outliers. Using classification, we would be better able to take into account the effect of these outlying car models and be able to identify what makes them unique.

One approach to go about this could be creating classes using body_type (Hatchback, MUV, Sedan, SUV and Luxury) and binning the cars that fall into these categories by quantiles. We can then try and classify our test data set into the identified bins.

This approach would not only give us a ballpark over our supposed Price Range, it will also point out directly which companies and car models we are competing against. This can be a huge leverage while launching a car in a new and competitive market like India.