Capstone Project: Used Cars Price Prediction

Executive Summary:

This project proposes a combined Regression model for the accurate prediction of used car prices. Linear Regression model performs better to fit the dataset and Ridge/Lasso Regression model to identify the most important features that impact the price prediction. The suggested model performs better with the observed data and comparatively low error. The Engine power and Year of manufacture plays a significant role in predicting the accurate price for used cars and hence the stakeholders consider these variables.

Problem Summary:

The used cars market contains uncertainties in pricing and supply. The sellers are in a difficult position to predict the actual price of the car as several factors such as mileage, brand, model, year can influence it. This issue is important to solve as it ensures the sellers are receiving accurate and fair benefits of selling their vehicle. It also protects the market growth. The key objective of this project is to come up with an effective pricing model that can accurately predict the price of used cars and can help the business in devising profitable strategies using differential pricing.

Solution design

Different regression models were explored as part of the design and the final proposed solution is Linear Regression Model which yields the better R-Squared and low RMSE values when compared to other models.

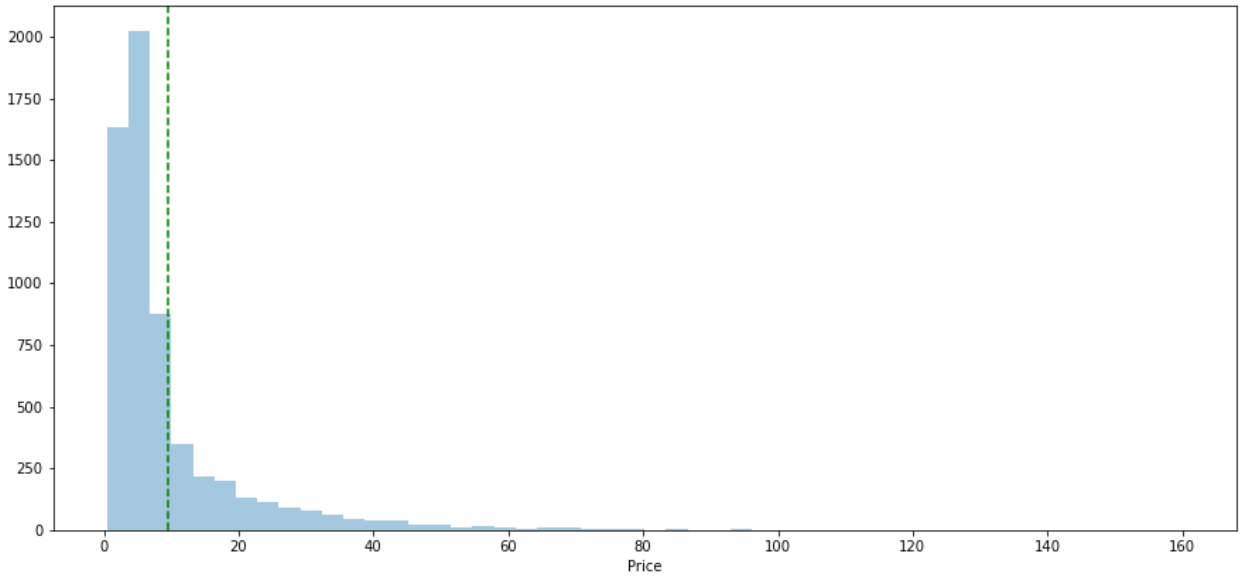
	Model	Train_r2	Test_r2	Train_RMSE	Test_RMSE
0	Linear Regression	0.704073	0.657405	5.852769	6.329361
1	Decision Tree	0.368083	0.379616	8.552623	8.517250
2	Ridge Regression	0.695206	0.671818	5.939807	6.194791
3	Random Forest Regressor	0.406375	0.411432	8.289439	8.295979

This appeared to be the robust prediction model with both the train and test data set.

Analysis & Key Insights

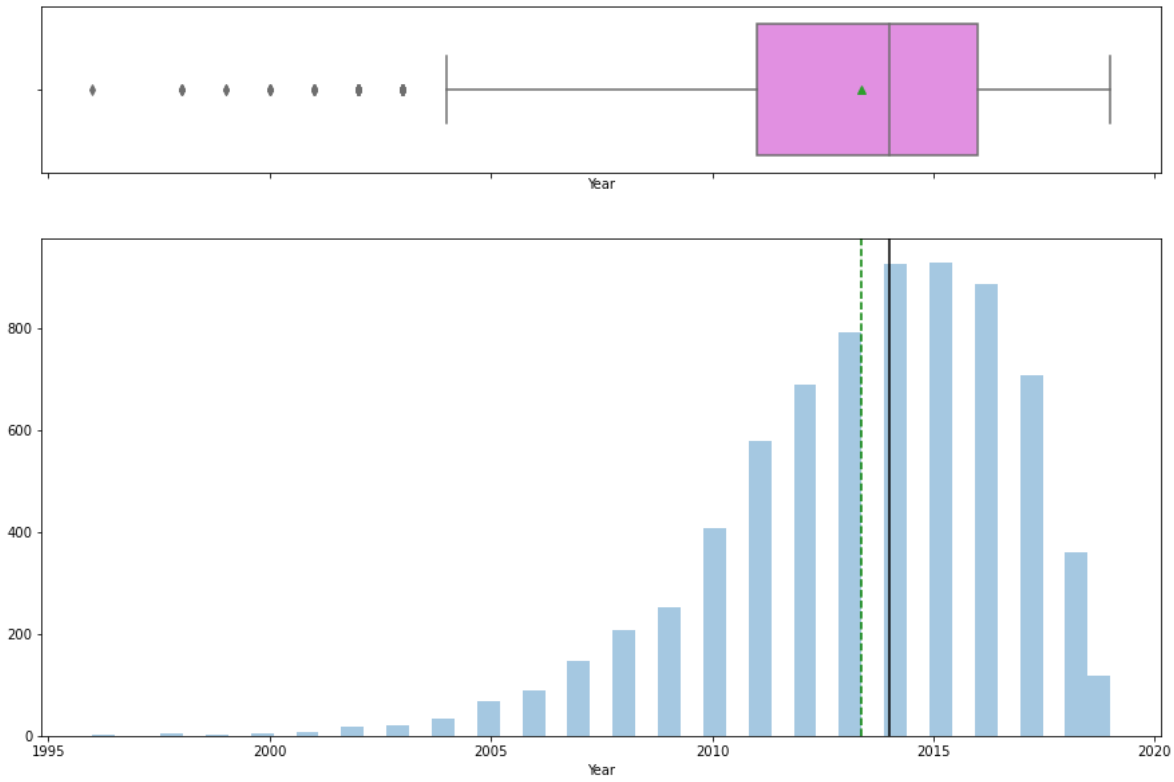
In order to arrive at a better prediction model, it is important to first examinate the given dataset, perform basic data preprocessing Feature engineering, treating the missing values, Univariate/Bivariate analysis.

- There are some discrepancies and missing values in the dataset for 'Kilometers_Driven' and 'Mileage'.
- Some of the attributes like Kilometers_Driven are highly skewed in the data

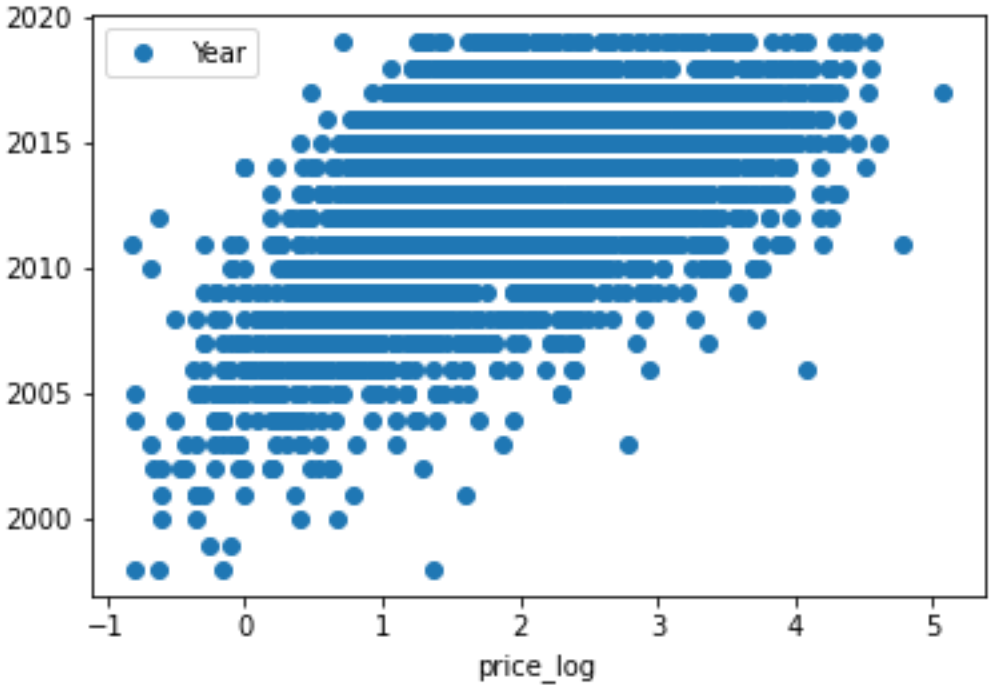


- Cars manufactured during 2011-2016 are more in the used cars market
- First hand owners are selling their cars more. This means second and third hand cars cannot be priced higher for sale and would not benefit the customer selling it
- Most people were selling the Manual transmission vehicles

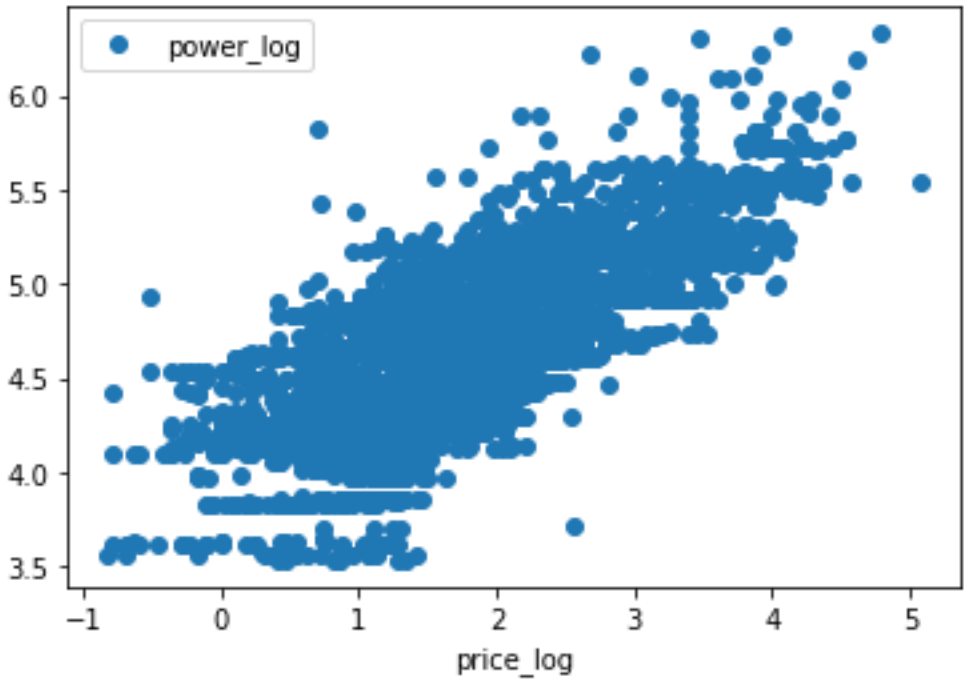
- Missing values in seats column were filled by using the Brand name and Model of the car
- Engine and Power have some manual entry errors and they are cleaned off.
- The standard deviation of `Kilometers_Driven` are high that indicates these values are more spread out. Hence a log transformation is applied on this parameter.
- Engine power is also highly rightly skewed so we have to apply a log transformation on power.
- More used cars are for sale in the metropolitan and cosmopolitan cities
- Mahindra XUVs are more sold in Mumbai in used car market
- Diesel vehicles tends to be driven for more duration and they are used by a single owner for a long time before they sell it. And they have the highest engine power
- There are few outliers in Year which doesn't have any significance, especially entries before year 2007 would not make any significance. So we have to ignore them.



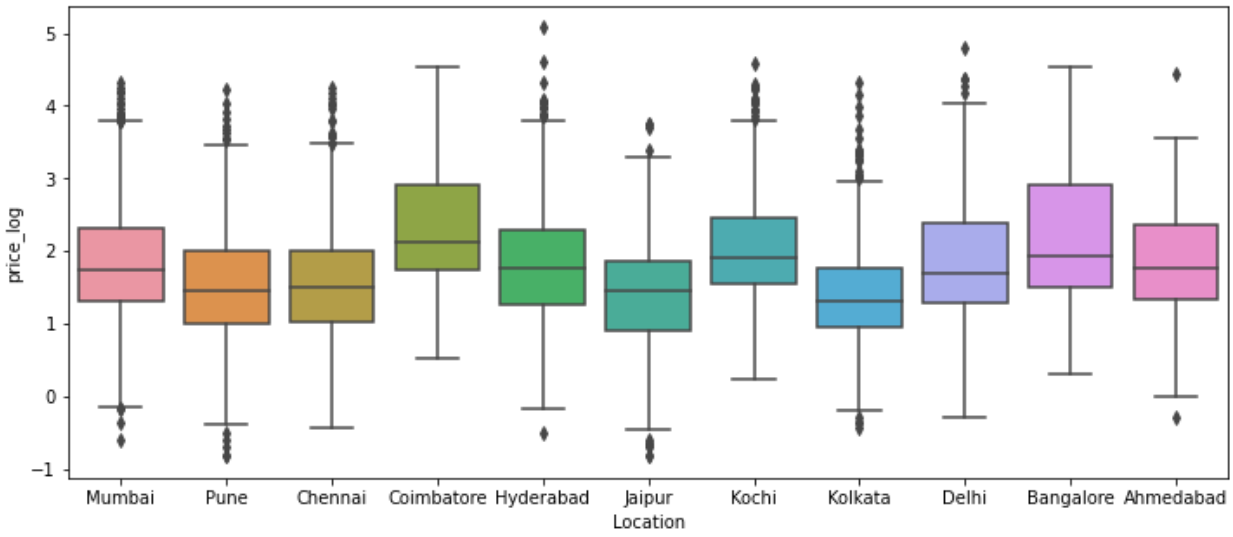
- Number of seats is not uniformly distributed with most of the data around 5 seats.
- Mumbai has the highest percentage of used cars for sale in market
- The price is higher for more recent cars. Meaning old cars are less priced comparatively with some outliers
- There doesn't seem to be a direct relationship between Kilometers driven and the price. Most of the cars are driven more than 80K kilometers but their price seems to be centered around 100,000 to 300,000
- Cars with more engine power tends to be expensive
- Engine volume doesn’t seem to impact the price of the car
- Some Luxurious cars with limited seats are priced higher which can be ignored
- Year and Price are positively correlated. Mileage and Year also have positive correlation.



- Mileage and Power, Mileage and Engine Volume have a significant negative correlation
- Price and Power, Price and engine volume, Price and New car price have the most significant positive correlation



- Vehicle price is higher in cities like Banglore and Coimbatore. Bangalore being a silicon valley of India with more population needing cars for their daily commute. Coimbatore is developing as an IT Hub and more educated people are migrating there.



- Diesel and Automatic transmission vehicles are expensive.
- First and second hand vehicles are priced more than the others.

Model Building

- Normalized version of price (price_log) is used for modelling the dataset.
- Categorical variable 'Name' can be removed as it does not have any significance to our model
- Kilometers_Driven, Price and price_log are the target dependent variables so they are dropped from train X dataset
- The entire dataset of 7253 used car entries are split into train and test dataset on a 70% - 30% ratio

Linear Regression Model

R-sqaure on training set : 0.7040732008196231
R-square on test set : 0.6574047216841202
RMSE on training set : 5.8527691291856705
RMSE on test set : 6.329361466562765

- The RMSE of test data is greater than the train data, that means the model over fitted the data.
- The R- squared values of Train and Test data > 0.6 and hence better in predicting the target value.

Drawbacks - There are no significant categorical variables identified by Linear Regression model.

Ridge/Lasso Regression Model

Ridge Model has a good prediction score of **0.8046030265770266**. Hence t his model is slightly better than Linear Regression. But Lasso was helpful in i mportant feature selection as there are only few parameters with more signif icances. Price and Year are identified as the significant features.

Decision Tree

The R-squared values of train and test data are closer. The RMSE values look underfitting the dataset.

The R-squared value for the model is less than 0.5 so its performing weaker than Linear Regression

Power is identified as the most important feature that can influence the price of the used cars. Year comes as the second most influencer. But the gap between their importance is higher.

Random Forest Regressor

The model has performed weaker with the R-squared and RMSE values.

RMSE value is also higher than linear regression so its weaker

But the RMSE values of Train and Test data are much closer than decision tree and there is no overfitting or underfitting of data.

Power and Year are identified as the most important features that can influence the price of the used cars. Both features have significant importance.

Refined Insights

- Overall, the Engine Power and Year has a significant impact on the price of the Used Cars. More the power of the engine, the car is priced higher. More the car is recently made, it's priced expensive and it's more from the first hand owner.
- Cities with more cosmopolitan population and cities with higher growth of IT and other technological job sectors have more demand for cars. More first-hand owners tend to switch to a new car with modern features.
- Overall, the Linear Regression and Ridge Regressor models performed better than a Decision Tree and Random Forest. So these models can be best fit. But feature importance were not identified by Linear Regression, so we have to use Lasso for retrieving the significant features.

Recommendations

- Identify used vehicles which are manufactured after 2017 and promote their sales.
- Introduce more campaigns in developing cities like Bangalore, Coimbatore.
- Shift the focus on vehicles with more engine power that attract the current generation of customers.
- Promote sales of Single hand (first owner) vehicles. Introduce more deals & discounts to boost their sales.