Capstone1

Automobile Price Prediction

In-Depth Analysis

1. Overview:

This report summarizes the In-depth Machine Learning applied to solve the problem of predicting the price of an automobile in the Indian market.

The database we are utilizing for the price prediction model is a database of over 1200 car models/variants from the Indian market updated till June 2020. These contain numerical as well as categorical features.

After cleaning & wrangling the database, EDA techniques were employed to get insights into the market and also to get a visual estimate of the important features that would drive our model.

Statistical methods were then deployed to make a scientific selection of the features. The techniques employed were:

1. Numerical predictor - Pearsons correlation coefficient
2. Categorical predictors - ANOVA f-tests and Tukey-HSD test

The above steps have been summarized and documented in preceding reports.

This document deals with the machine learning methods employed to arrive at the best fitted model for use for our project goal: Price prediction of an automotive based on the provided dataset.

2. Model Selection:

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.

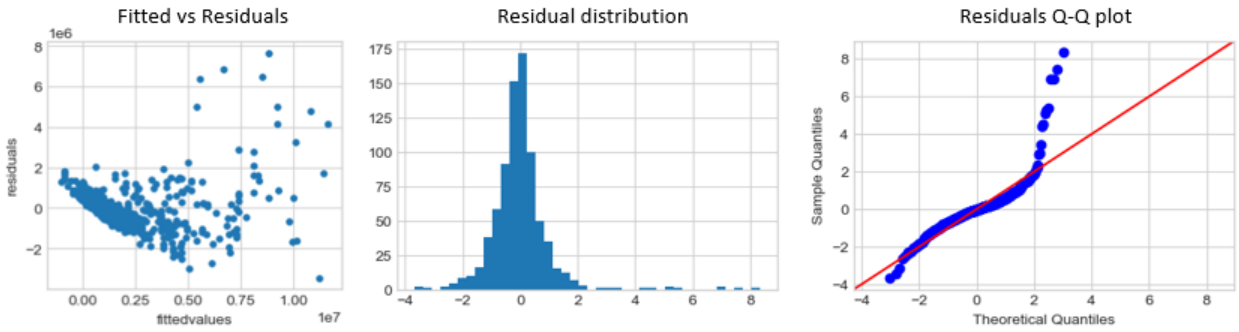
3. Verification of the Linear Model assumptions, transformation of data & removing outliers:

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

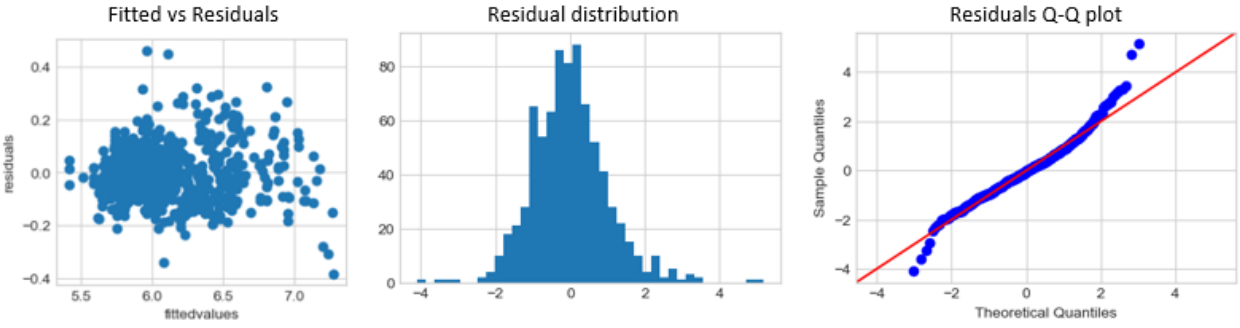
We check the above by building a makeshift model using statsmodels and plotting the residuals and predicted values. We quickly observe that the assumptions of linearity and homoscedasticity are violated.

First attempt model (Vanilla):

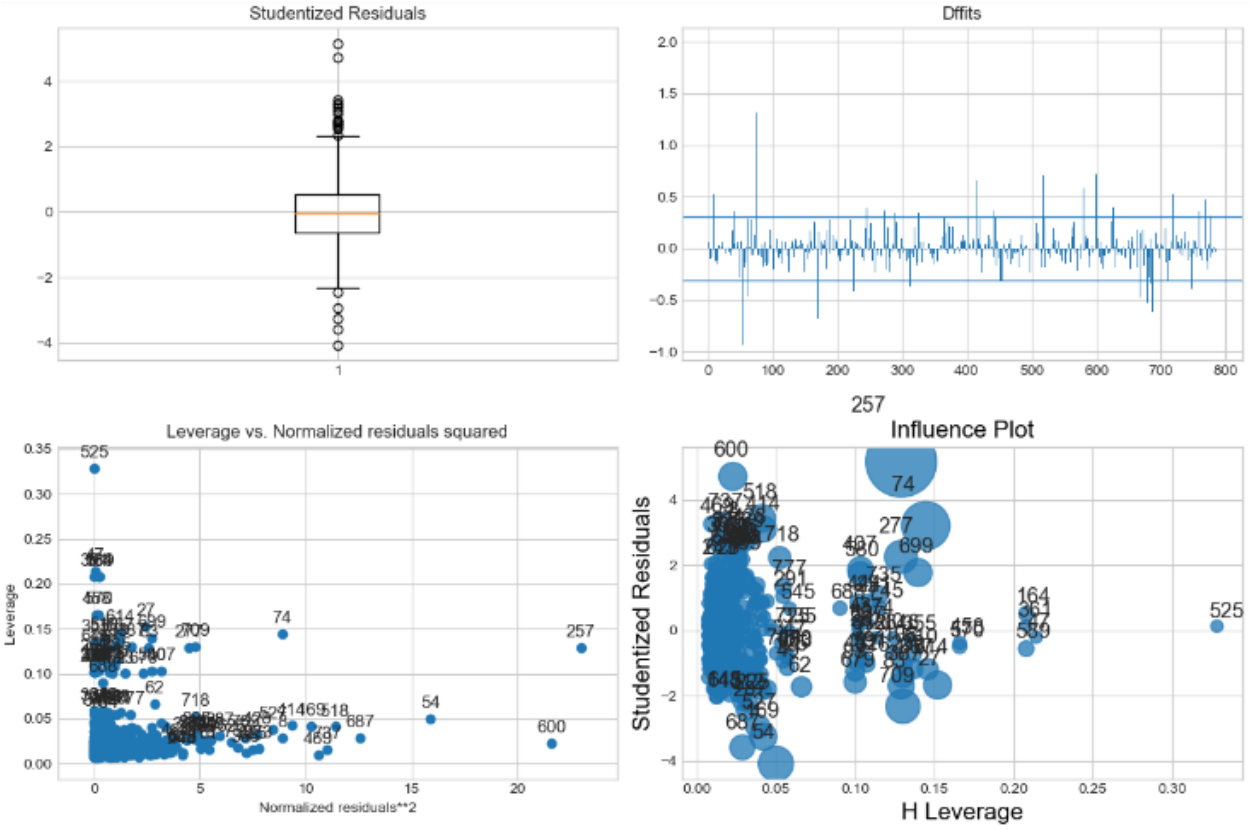


To correct the problems with linearity, we log-transform the target variable and check the plots again. This time we see that the problems of linearity and homoscedasticity are nearly resolved but the effect of outliers becomes prominent

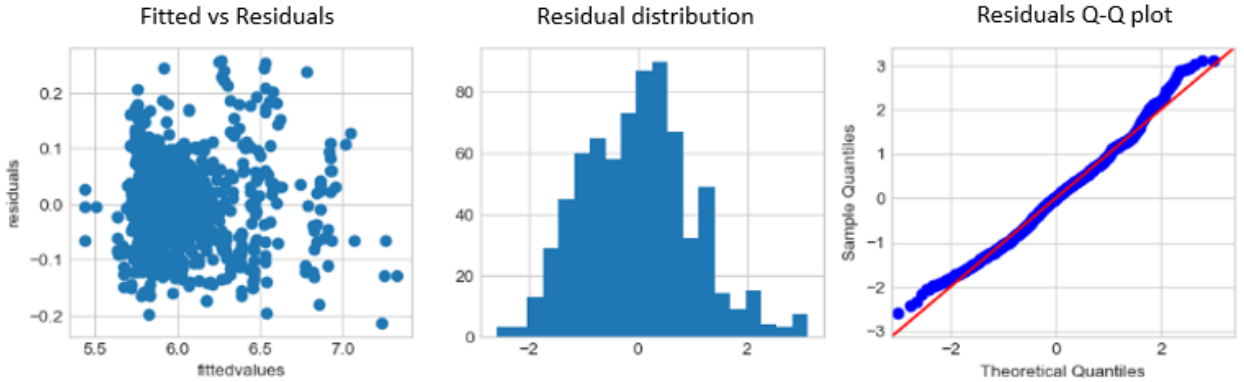
Target variable transformed:



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.



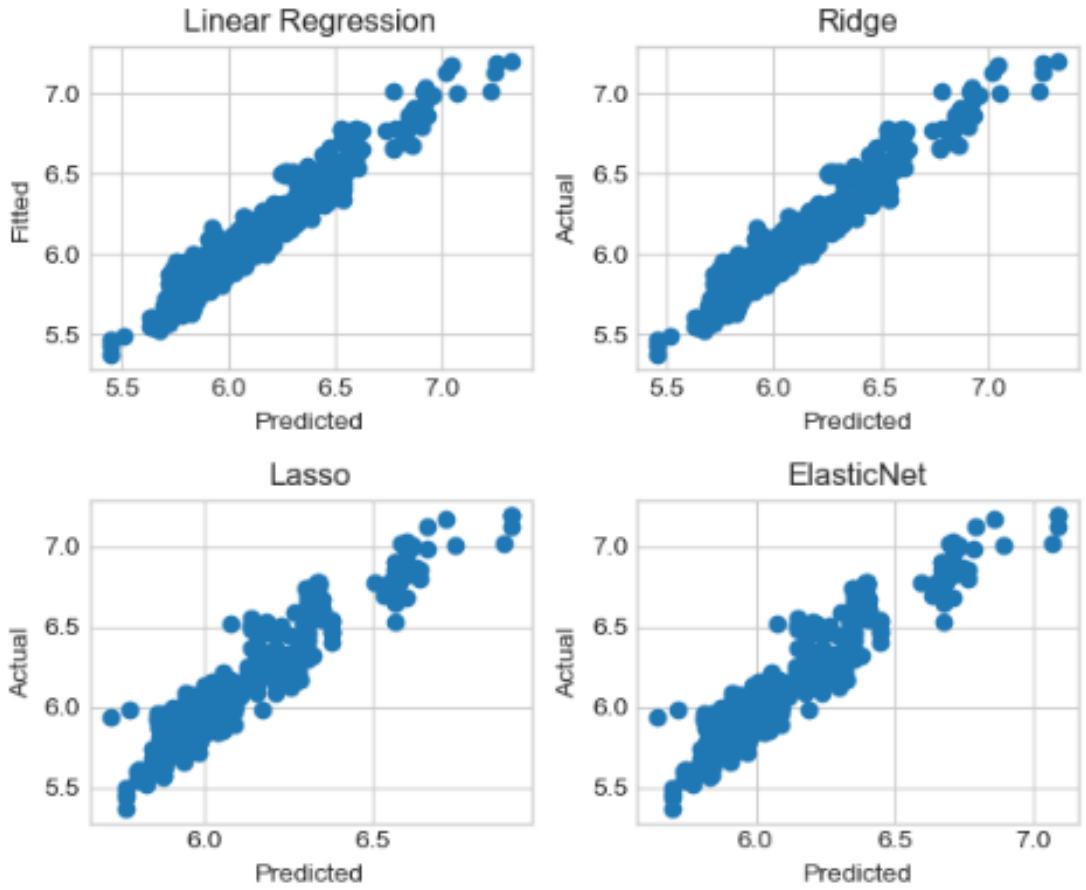
4. Final Model, Test scores and feature importance:

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.

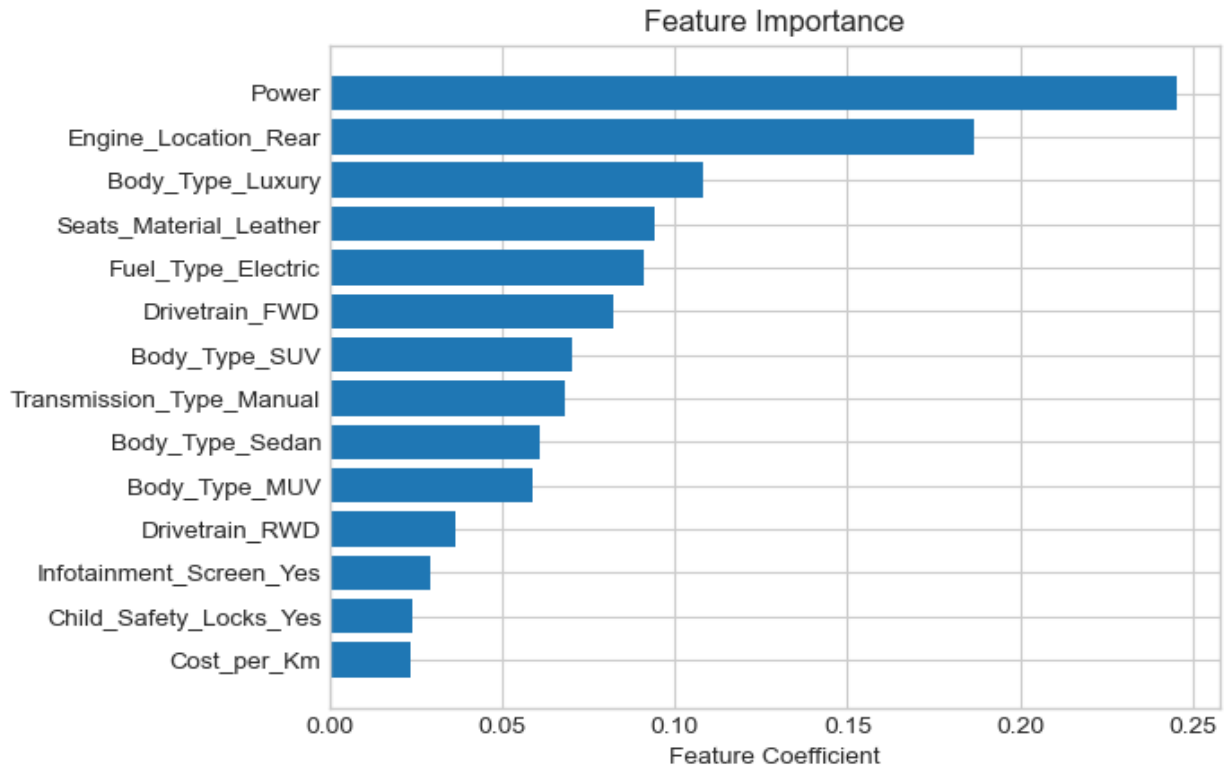
Comparison of accuracy

	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