Car Pricing Prediction

- Poojan Fegade

Overview (Context)

Car market is highly competitive in its pricing range. Companies select a certain market segment to launch a vehicle with attractive specifications along with its pricing.

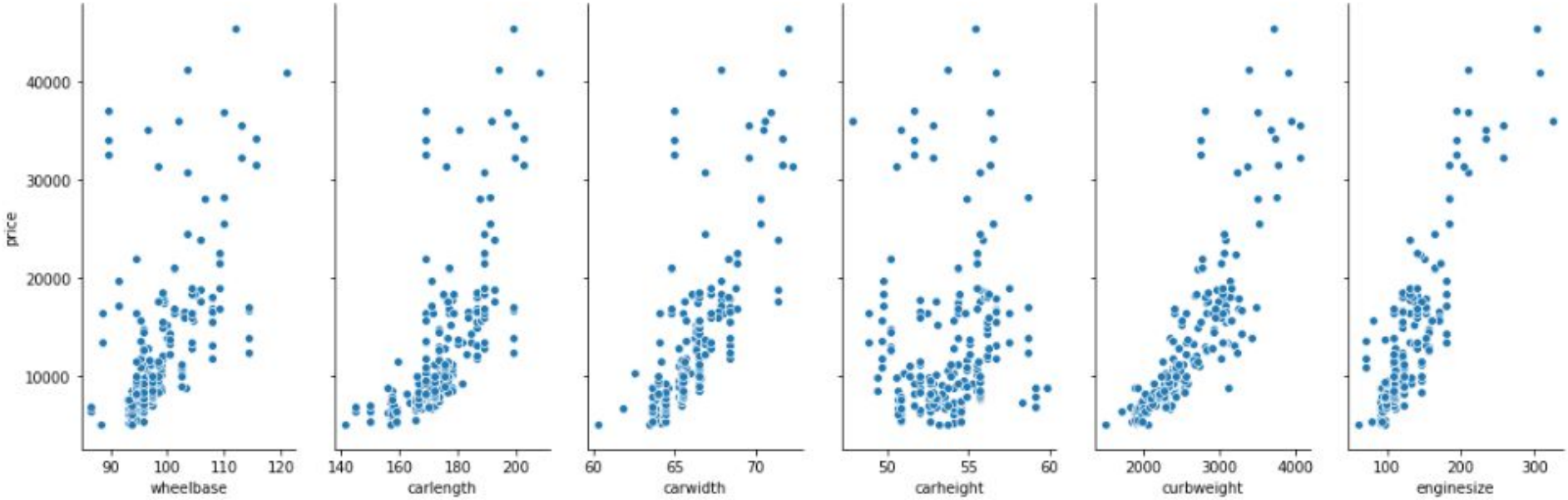
Problem Statement

There is need for a model to predict the car price given its specifications, as , there are certain details of the cars that are revealed in the car concept. These details can be used to predict the price of the competitor’s car in general and the company can take appropriate steps to counter the competitor.

Methodology

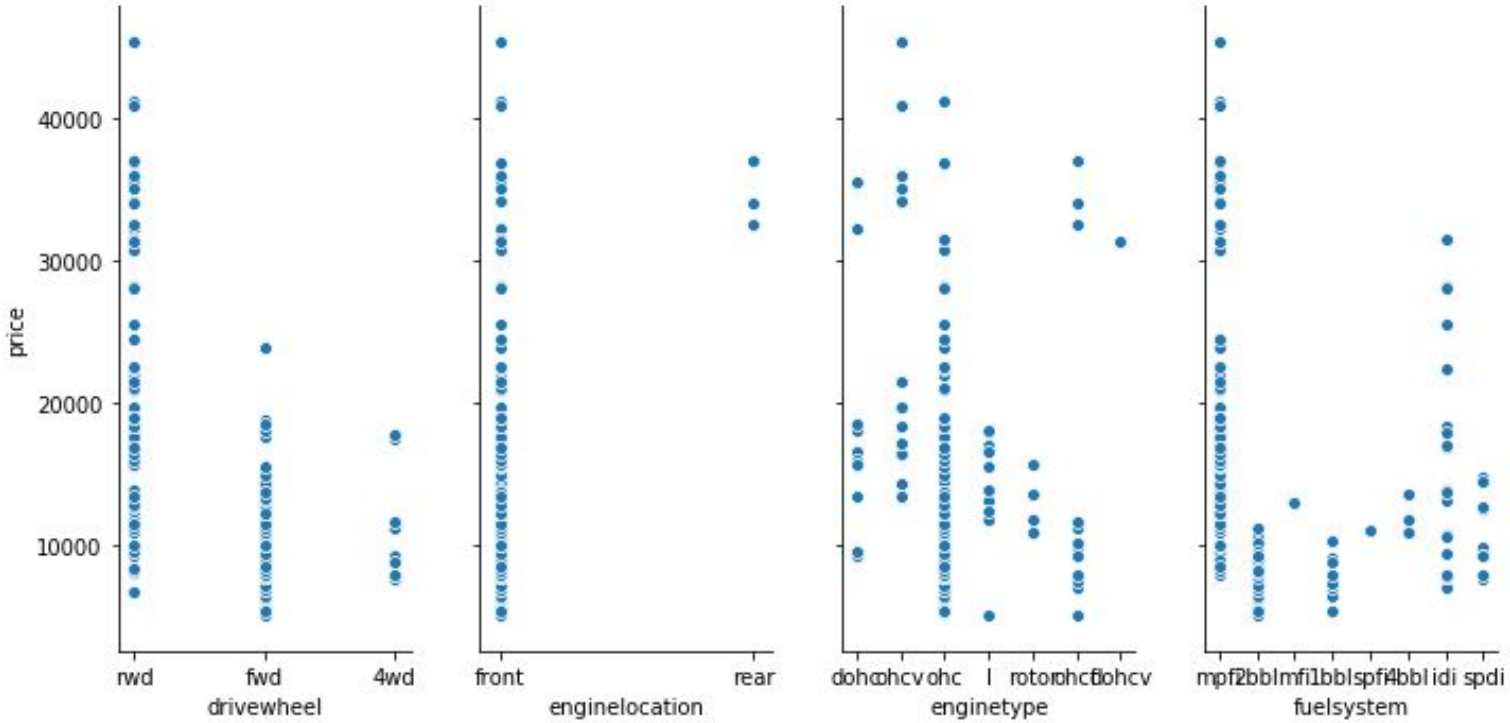
- Understanding the data fields as well as their basic statistical properties
- Identifying the categorical data fields and understanding it’s role
- Data cleaning by filling missing values or removing null record.
- Manipulate data by encoding categorical variables
- Split the data as training and testing data (7:3 ratio respectively)
- Fit model using features and validate using the test data.

Data Visualisation

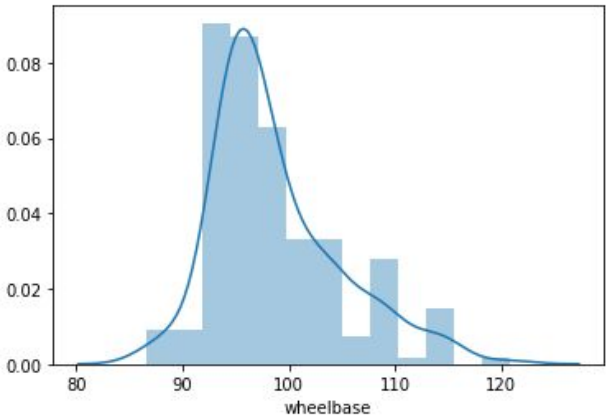


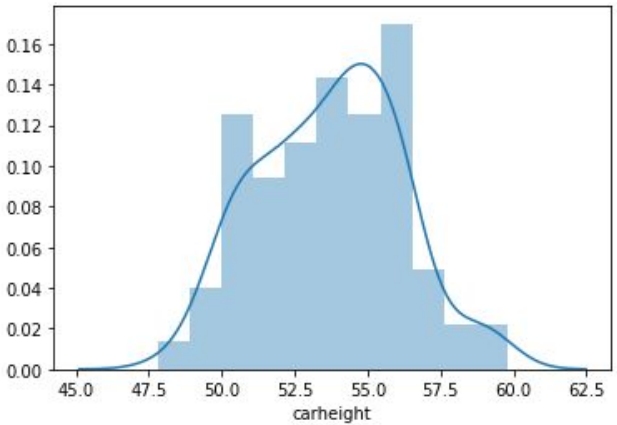
Here, from scatter plot we see that price is almost linearly dependent on wheelbase, weight, height, width and engine size of the vehicle.

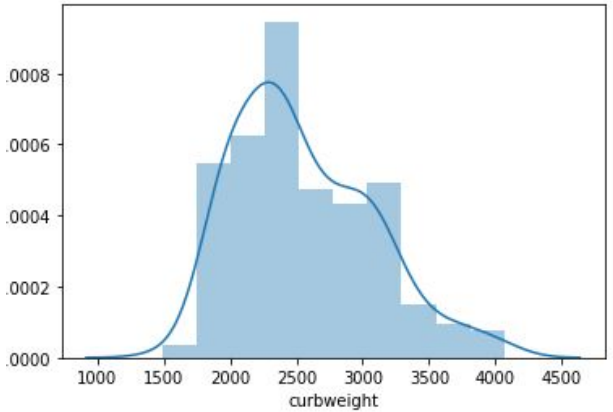
Price range covered for categorical variables such as drivewheel, enginelocation, enginetype and fuelsystem. These scatter plots tell us the approximate price range observed for a categorical variable (How costly a particular category can be)



Variation of wheelbase, carheight and cubweight is plotted to check the skewness in the distribution. This helps us in locating outliers and transform the data accordingly.

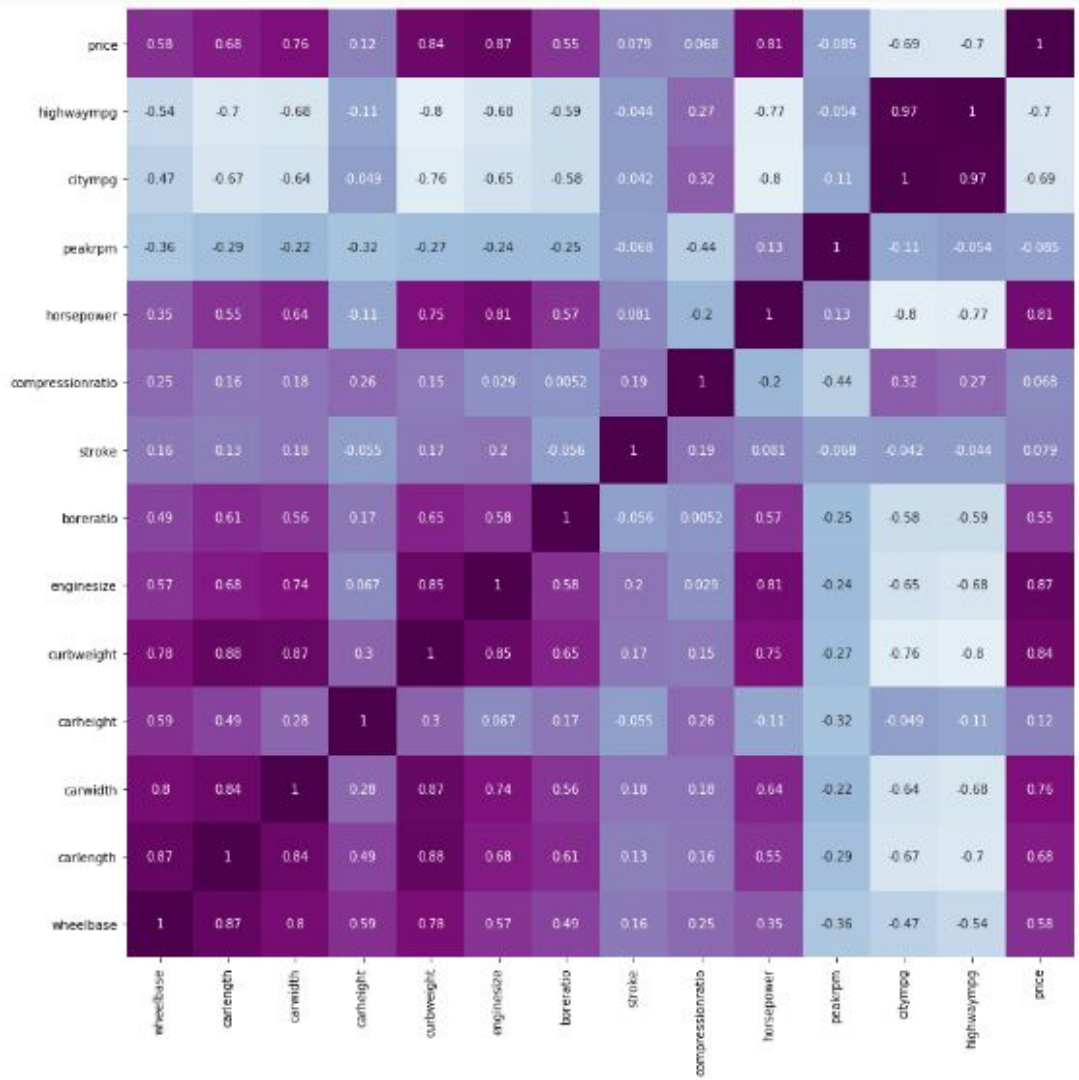






From the cross-correlation matrix, we find the following data fields important :-

- Enginesize
- Horsepower
- Curbweight
- Carwidth
- Highwaympg
- Carlength



Data Manipulation

- Created a new data fields called CarCompany and Carmodel for further analysis
- Corrected the misspelt CarCompany
- Preprocessed the categorical data by using OneHotEncoding.
- Used MinMaxScale in the numerical features as they didn’t have any outliers

Model

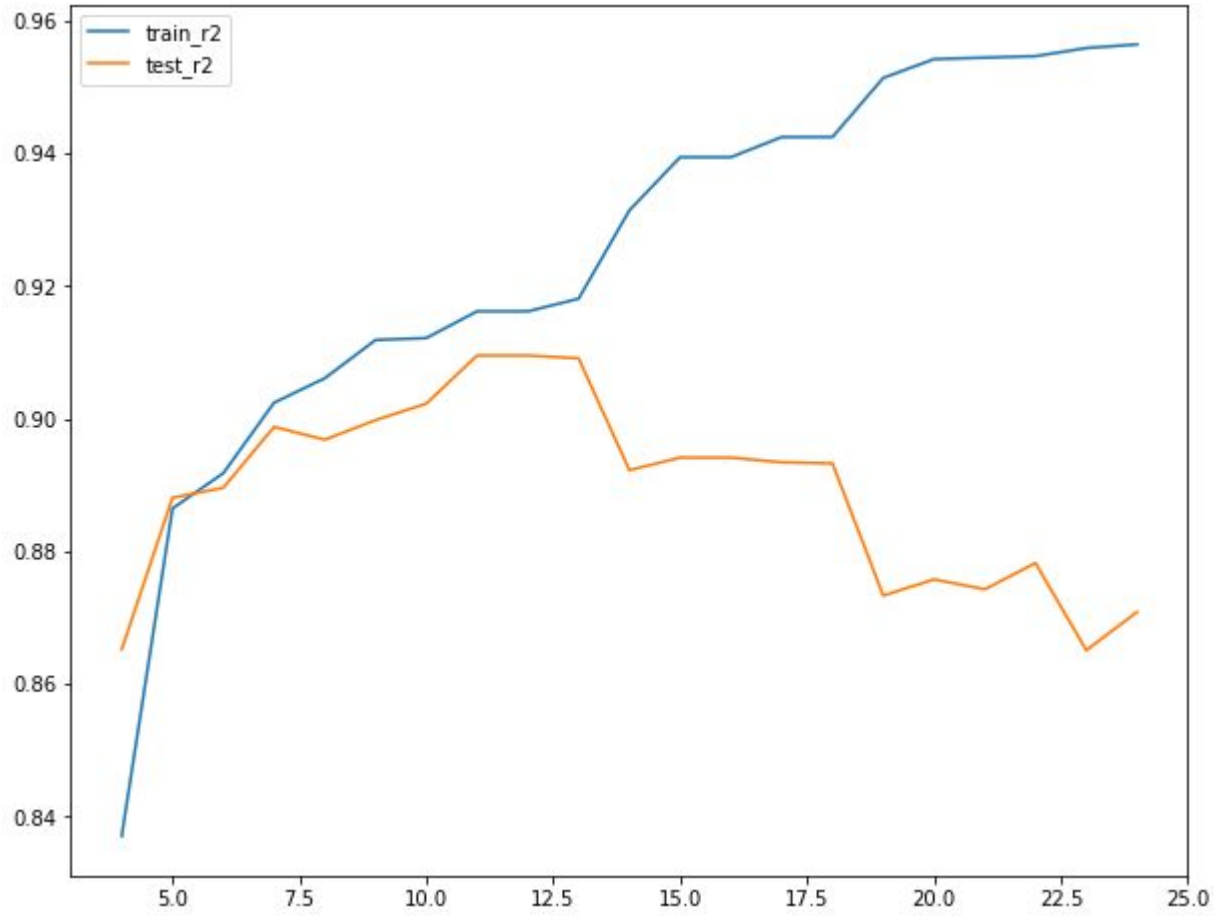
- Split train and test data in the ratio 7:3
- Fit Linear Regression model by using all the features and observed the coefficients and significance level of the features.
- However, not all the features contribute to the prediction variable. Hence removing features of low importance can increase accuracy, and reduce both model complexity and overfitting.
- Training time can be reduced for large datasets

Feature selection

- We use recursive feature elimination to select important features.
- We select a significance level
- Fit our model with all the independent variables
- Consider the variables with highest p-value. If the p-value is greater than the significance level we remove the feature
- We build the model again using the remaining independent variables
- Repeat the process until removal of variable changes the accuracy drastically
- In python we have RFE module to use

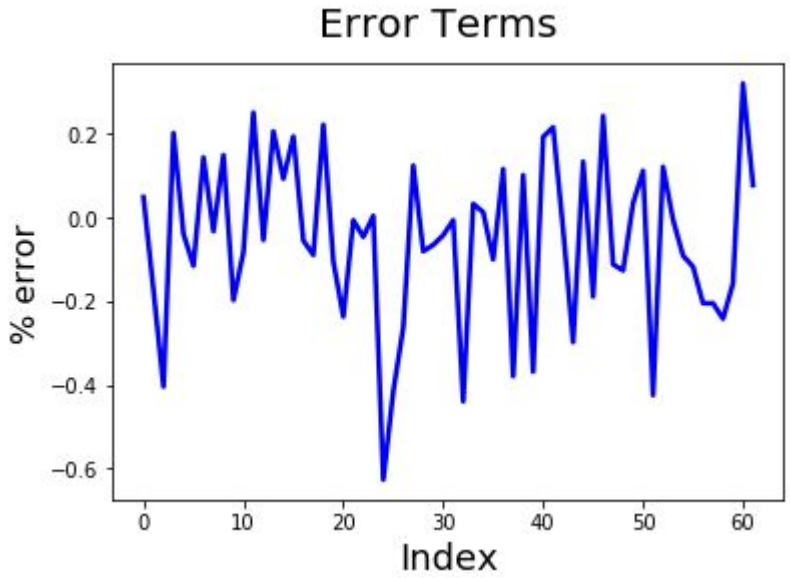
The adjacent graph shows the variation of r-squared with respect to the number of features selected.

The optimum area would be from 6-12 as the r-squared value doesn’t change much.



Final Model

- In our final model we used 7 features along with constant.
- Our level of significance is 0.05.



Dep. Variable:	price	**R-squared:**	0.902
Model:	OLS	**Adj. R-squared:**	0.897
Method:	Least Squares	**F-statistic:**	178.4
Date:	Sun, 15 Sep 2019	**Prob (F-statistic):**	5.64e-65
Time:	01:46:29	**Log-Likelihood:**	-1317.7
No. Observations:	143	**AIC:**	2651.
Df Residuals:	135	**BIC:**	2675.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1872.5321	598.553	-3.128	0.002	-3056.286	-688.778
carwidth	1.336e+04	2416.207	5.530	0.000	8583.500	1.81e+04
curbweight	7599.4427	2733.109	2.781	0.006	2194.195	1.3e+04
enginesize	2.063e+04	3055.575	6.750	0.000	1.46e+04	2.67e+04
enginelocation_rear	1.658e+04	2619.421	6.329	0.000	1.14e+04	2.18e+04
cylindernumber_three	7023.5321	2571.783	2.731	0.007	1937.336	1.21e+04
cylindernumber_two	5262.9474	1372.528	3.834	0.000	2548.510	7977.385
Car_Company_bmw	8820.2389	1089.708	8.094	0.000	6665.132	1.1e+04