IDENTIFIED SOLUTIONS

Data Analysis

This section performs the selling price prediction using a dataset consisting of 8128 used car details. This dataset is prepared by cardhekho.com and available on kaggle.

There are categorical as well as continuous features here

Correlation Matrix

Visualizing the correlation is an effective way of determining the dependencies. Sometimes selling price has high correlation with the manufacturing year engine Max power and transmission. The engine and the manufacturing year has the same approximate correlation so we can select any one of them in the final set of features.

Pair plot

Pair plots allows to see both distributions of single variable and relationship between 2 variable. They are great method to identify trends for follow up analysis and are easily implemented in python. Scatter plots in pair plots also help in visualization of outliers.

Random forest regression

Random forest is a supervised learning algorithm that uses an ensemble learning approach for regression and classification. The main principle behind the ensemble approach is that weak learners can learn from strong learners. Random forest operates by constructing multiple decision trees at training time. These decision trees are independently trained on bootstrap datasets. The final predicted value is calculated by taking the mean of the predictions by all the individual values.