Linear Regression - Report

Adarsh Singh

At first, we open the data set using the pandas library into the Jupiter notebook. We check for null values, and we find that we don't need to clean the data. The data frame comprised of 21613 observations and 21 explanatory variables (columns). The focus of the assignment is to predict the housing price. The target variable is set to be "price" in the dataset. The other variables excluding "id" are just features in the dataset and they are not used in price prediction. I have left out zip code and date as well. I have used data science related python libraries - pandas, numpy, matplotlib, sklearn, and seaborn. The distribution was left skewed with normal values. This means the mean is lower than the median value

The dataset comprised of 19 variables, so, I checked the correlation between the 'price' and other variables using corr() method and plotted a confusion matrix to check using the seaborn library. I decided to choose all variables other than 3 stated above with the highest correlation to perform linear regression. Further, the selected variables were plotted against 'price' in a scatter plot diagram. The presence of multicollinearity between variables is seen due to the similarity of data point distribution in the scatter plot and high co-linearity value between variables in the confusion matrix.

The dataset was divided into 80% training set and 20% testing set. Linear regression was performed on the training dataset. The evaluation of the regression model was conducted using R-square, and Root Mean Square Error (RMSE) is the measure of the predictive data analysis. R-square value from the linear regression on the training set was 0.614153344939728O6 which is derived from having a moderate effect on the dependent variable. It is not a bad indicator for the predictor of the housing prices. But, given the size of the dataset and the correlation between variables, the data points are about 54% (approx.) scattered around the regression line. This could also negatively indicate to show that the dataset is not strong enough to deliver accurate predictive analysis in order to determine the price.