Linear_Regression_Bike_Sharing_Assignment

Problem Statement:

A US bike-sharing provider BoomBikes has recently suffered considerable dips in their revenues due to the ongoing Corona pandemic. The company is finding it very difficult to sustain in the current market scenario. In such an attempt, BoomBikes aspires to understand the demand for shared bikes among the people after this ongoing quarantine situation ends across the nation due to Covid-19. They have planned this to prepare themselves to cater to the people's needs once the situation gets better all around and stand out from other service providers and make huge profits.

The company wants to know:

Which variables are significant in predicting the demand for shared bikes.

How well those variables describe the bike demands

Goal:

Develop a model to find the variables which are significant the demand for shared bikes with the available independent variables.

It will be used by the management to understand and manipulate the business strategy to meet the demand levels and meet the customer's expectations.

Conclusion:

Significant variables to predict the demand for shared bikes

holiday

temp

hum

windspeed

Season

months(January, July, September, November, December),

Year (2019)

Sunday

weathersit(Light Snow, Mist + Cloudy)

- The R-squared value of the train set is 82.71% whereas the test set has a value of 81.13% which suggests that our model broadly explains the variance quite accurately on the test set and thus we can conclude that it is a good model.

- Our developed model's mean squared error is almost 0 on both the training and testing datasets which suggests that the variance is accurately predicted on the test set. The p-values and VIF were used to select the significant variables. RFE was also conducted for automated selection of variables.

- The major steps included in the python notebook are data interpretation, data visualisation, data pre-processing, model training, feature selection, residual analysis, model evaluation on the test set.

- Concepts such as EDA, p-value, VIF, RFE were used and model building was done using statsmodels library