KJ Morgan

Dr. Lewis

Machine Learning

July 22, 2019

Machine Learning

From analyzing the data, one can see that the research question needs significant variables. To make this research question be changeable by machine learning, there will have to be two or more dependent variables being ran upon an independent variable. The different variables will be modeled to test for significance. The Origin variable is set to be the independent variable due to, this variable containing the two subjects that will be researched. Then the other 3 dependent variables will be measured under 3 different models. The problem here is supervised because the output variable is a real object being analyzed over a time period. Also, with respect to the variables, the best fit model is a regression model. To analyze the data in the cleaned dataset, a Linear Regression model must be used to find P-value, RMSE, and R-squared values. Detailed also, will be the standard error and p-value. These values give your insight to how much a dependent variable affects the independent variable.

There are 4 variables that will be analyzed in these models. Statistics show that there are 3 dependent variables and 1 independent variable. The independent variable is Origin, split into two subcomponents, ATL and LAX. These are the two major airports that were most significant in the analysis of two major airports. The three dependent variables are Security, Weather, and Late Aircraft delay. I will use the delay times listed under these variables to run statistics on and eventually put into the linear regression model.

The code used in machine learning found the standard error from the statistical analysis. Looking for a good value for a p-value, results in removal of a null model. A good p-value indicates a good model. In this technique, the standard error should be low. From this point, I used the “model2” to find my standard error and p-value. However, on the first model, my p-value was too low. Then I used the cross validation model, “model3”, and tried to find a lower p-value, using my three variables. An R-squared value determines the correlation between the points in the scatter plot. The range for R-squared values is 0-1. 1 stating a well precise model, 0, stating the independent variable may needs to be removed or furtherly analyzed. To get the R-squared to approach 1, I subtracted two variables and got a better value. The RMSE value was displayed and needed readjusting. Therefore, the cross validation method will be implemented. Next, cross validation was implemented to get a lower standard error on the validation dataset. The validation dataset is the dataset that holds the most efficient data. The last and most efficient model was ran on the validation dataset which proved and found an R-squared closer to 1.