EECS 731 – Report

10/12/2020

This was a little more difficult of a project than I initially believed it would be. I began the project by reading in the data and displaying how much information was available in different ways. It was very clear from early on that one particular product, and one particular warehouse, had much higher numbers than any of the others. The products in Category\_019 had a more than four to one ratio versus the next most demanded product. There was quite a bit of manipulation that needed to be done in order to make sure the data set was useable. The most time-consuming part of the project for me was trying to replace the negative values that were depicted with parenthesis with a numerical value that could be calculated.

When I was successfully able to do so, I then sorted the information based on the date and went about setting up the training samples and testing samples. These values would be used in the feature engineering portion of the program. I graphically represented the amount of orders there had been since the start date in 2012 until the final date in 2017. It would appear that based on this graph, that the company had been stable during this period and could be expected to be stable going into the future.

For the feature engineering portion of the program, I used ARIMA as was suggested by the professor of the course. I spent an abundant amount of time attempting to import “pmdarima” and use it as another feature engineering model, but even after attempting to use pip to install, I was unable to ever get that feature recognized by my system. I am not sure what the issue is, or why it would not install correctly, but it is something I will be looking into. After applying ARIMA to the dataset, I also applied R-squared regression in an attempt to get an accuracy score. I am not really sure how to interpret the results I received, if it is directly evident of the accuracy achieved then it was a very inaccurate modeling of the provided data.

I then applied three additional feature engineering models to the data set, Autoregression, Autoregressive Moving Average, and Seasonal Autoregressive Integrated Moving-Average. The scores I received based on these models was drastically different than my first results. Of the three, Season Autoregressive Integrated Moving-Average, or SARIMAX, had the highest score of all the models with a score of 117.85.