|  |
| --- |
| 3/11/21 |

|  |
| --- |
| Machine Learning Summary: ARIMA Forecasting |

# Description

Problem: Difficulty in forecasting volume during the declines in work volume as a result of Covid-19 in March of 2020.

Goal: Forecast the volume on a monthly basis. Compare results to actual to determine effectiveness of the model.

## Preprocessing

A SQL script is run on review data stored in a Google Cloud Platform account. This script provides a list of dates with the volume on each of those days going back to January 2019. Each month the list is extended by approximately 30 days. The list is downloaded into .csv file and stored in a folder with the Jupyter notebook.

The only noteworthy processing event within the notebook is filtering out all data prior to July 2019. This is due to an unexplained drop in May/June volume that most likely represent an issue with data integrity.

## Parameters

Completion of the ARIMA modeling requires inputting parameters p, d, q. This can be achieved by either manual means or autodetection (autoarima()). Notes are included throughout the file to assist with interpretation of the various graphs and visualization necessary for manual determination of parameters.

The main visualizations used for determining parameters are:

* Autocorrelation with differencing: d parameter, 1st order or 2nd order differencing
* Partial autocorrelation: p parameter, determine the number of lags
* Autocorrelation: q parameter, determine the

An auto\_arima() method was eventually added to test for the best combination of parameters.

## Model Selection

It is acknowledged that there are several nuances and subtleties with determining the p, d, q, parameters. The most effective method for selection of the best model was found to be, in part, based on trial error. Use of auto\_arima() is general effective, but understanding the manual selection process avoids pitfalls and traps associated with the auto\_arima() method.

After identifying parameters based on either manual or automatic methods, the user should test alternative parameters and select based on plotted residual errors. It should be noted the model with the lowest error rate does not always guarantee convergence of the model.

## Running the Model

This notebook is currently set to run with the data and parameters from January 2021. No changes are required. When new monthly data is added the following variables need to be manually changed:

* p, d, q parameters as needed
* train and test (due increased length of list size)
* model\_fit.forecast(step= )[0] (step = number of days in next month or predicted period)

## Results

Model accuracy has ranged between 89-96% over 9 months, with a mean of 93.17%. Efforts to improve the mean have been hampered by changes in volume due to operational changes I was unaware of, variation in volume related to Covid-19, and minor experiments on the part of the author in an effort to improve accuracy. Further knowledge of the business environment may be required to improve accuracy.