Capstone Report

ALY 6980 21311 Capstone SEC 02 Winter 2019 CPS

Instructor: Alice Mello

Liang Shi

03/28/2019

**Introduction**

Generally speaking, two models are used to predict the at once business in this capstone project, one is ARIMA model and the other is artificial intelligence neural network which based on Keras function and TensoreFlow backend. The main part of this essay will focus on the workflow of dealing with the data and the effects of the code.

**Analysis**

ARIMA

Since we are about to predict the sales order quantity in the future and the only variable we can depend on is its historical value, obviously we need to store all the historical order quantities in a time series. ARIMA can be an effective way to deal with the time series related problem while before constructing the model, we have to make some preparations and have an overview of the dataset.

Data Preparation & Data Overview

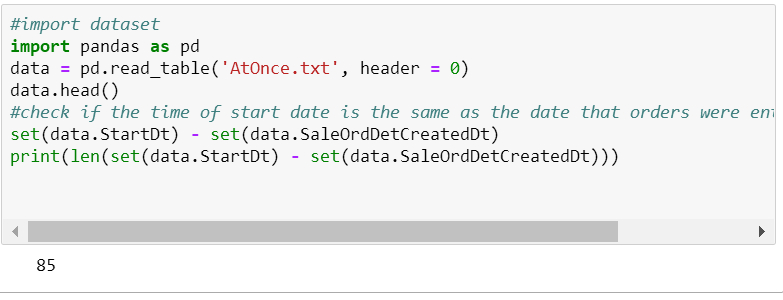


Figure 1

Since two columns are related to date in the dataset, and as the result shows, there are 85 different records between start date and the system recorded date. I finally choose the start date as the index since in the prediction, using start date can make sure the prediction start from January 2019, while using system recorded date will involve part of the results of 2018 which can be confusing for the decision maker.

After transforming the datatype of start date from string to datetime and set it as the index, I drew an overview by plotting the time series data.

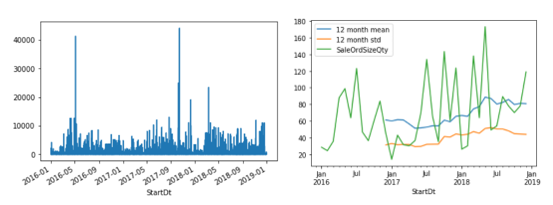


Figure 2

The graph in the left side is not smooth because the time series data is presented by days without any processing. To make it more smooth, I resampled the data by month and choose the average value of each month’s order quantity as the target. In addition, I also involved mean and standard deviation in the smoothed plot because ARIMA model requires the time series to be stationary which means the mean and standard deviation do not change across time. According to the right sides graph, mean and standard deviation are relatively flat.

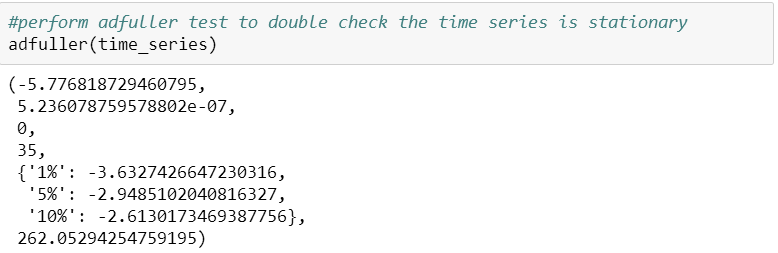


Figure 3

To double check its stationarity, adfuller test is also conducted which takes the null hypothesis as time series is not stationary. Then the P value is 5.236e-07 which is much smaller than 0.05, so we can reject the null hypothesis and significantly take the time series as stationary.

Build & Train ARIMA Model

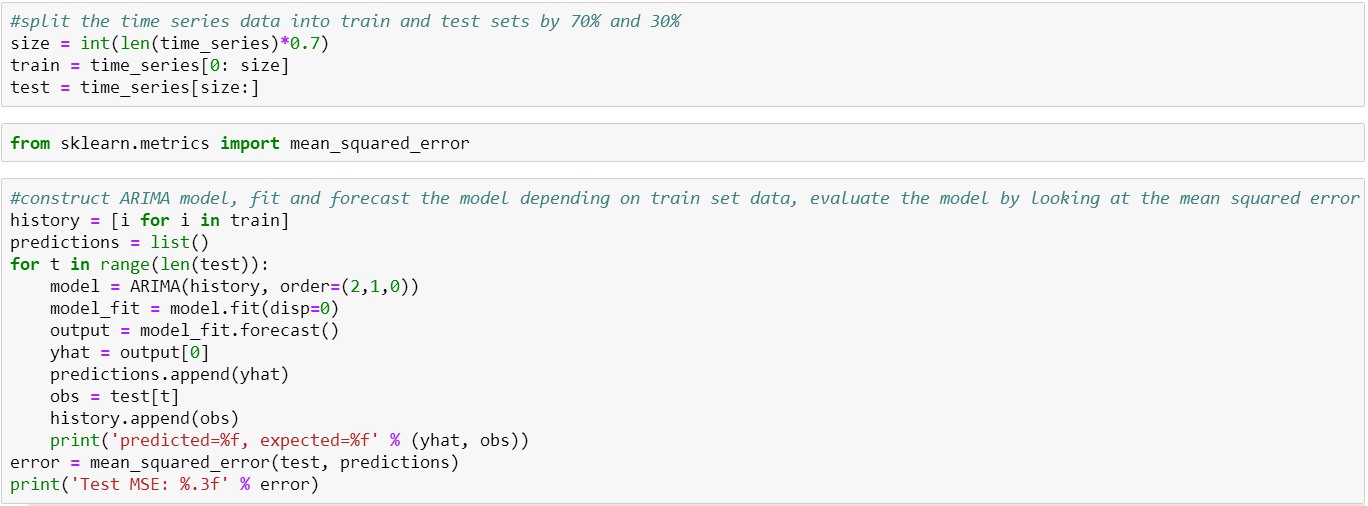


Figure 4

Before construct the model, I firstly split the dataset into train sets and test sets by 70% and 30%. The basic workflow of constructing the model is setting the parameters, fitting model by train datasets, making predictions based on test sets and evaluating the model by calculating mean squared error and visualizing the prediction.

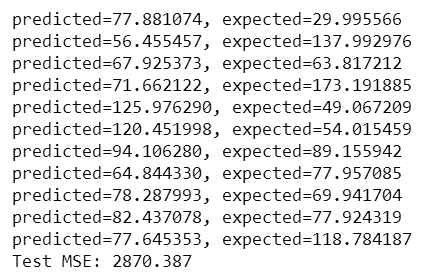


Figure 5

As you can see from the result, the test mean squared error is 2870.387. I don’t take its log value because it would be more direct if we look at its raw value. It is not a bad model because the fundamental value of average month order quantity is large, not to mention the squared one of the error.

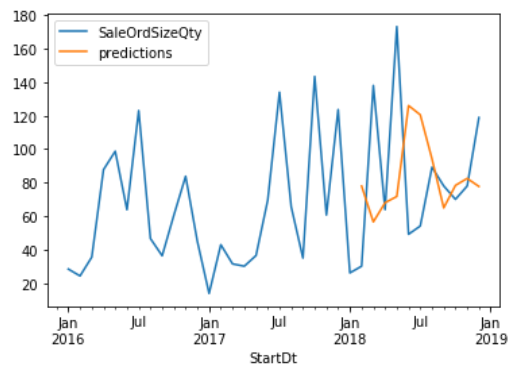


Figure 6

We can also tell that this model is not bad since there is not a huge bias between our prediction and true value.

Neural Network (NN)

In addition to the ARIMA model, I also constructed a neural network model as a great attempt of practicing this advanced technique and supplement of Cheng’s work. I didn’t present it in the class because I would like to focus more on ARIMA and let Cheng to focus on neural network while it is also part of my work.

My neural network is different from Cheng’s because I depend on the Keras function and take TensorFlow as the backend.

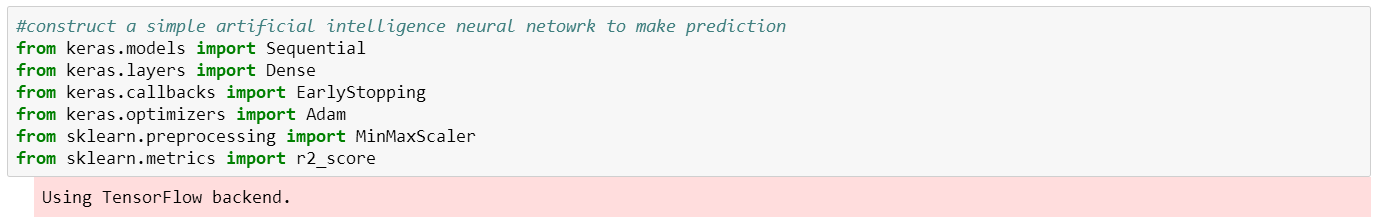


Figure 7

Before constructing the model, I normalized the data which can eliminate the scale difference of data and make the result more beautiful.

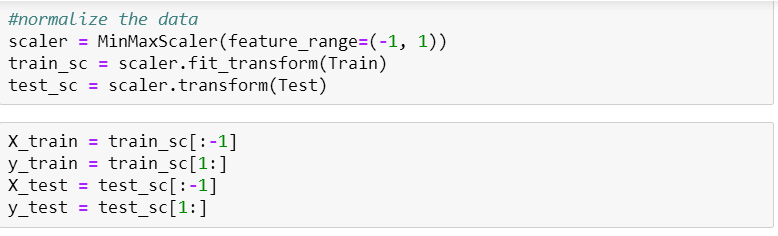


Figure 8

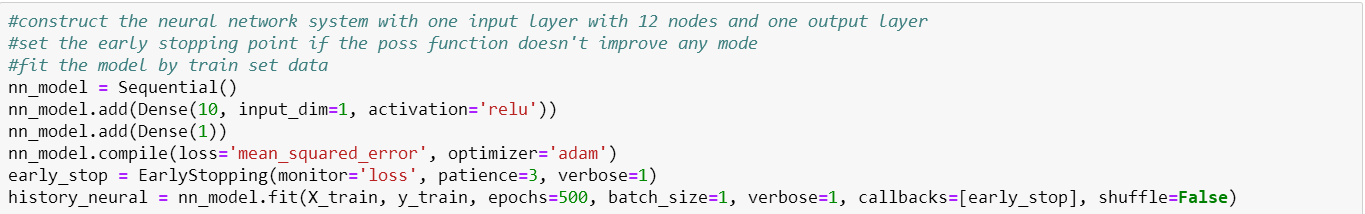


Figure 9

My neural network model is composed of only input layer and output layer. For input layer, there are 10 nodes in total and only one input dimension which is the historical order quantity value. Besides, I choose the relu as the activation function because the target is to get a sequence of numeric value and relu function can play the similar rules of regression. To compile the model, I choose the mean squared error as the loss function and adam as the optimizer. As what as tested by Cheng, adam works best among all the optimizers. And also, I set the early stopping point to improve the efficiency because once there is not obvious improvement with the loss function, the iteration of neural network will automatically stop just in case it takes too much time to run.

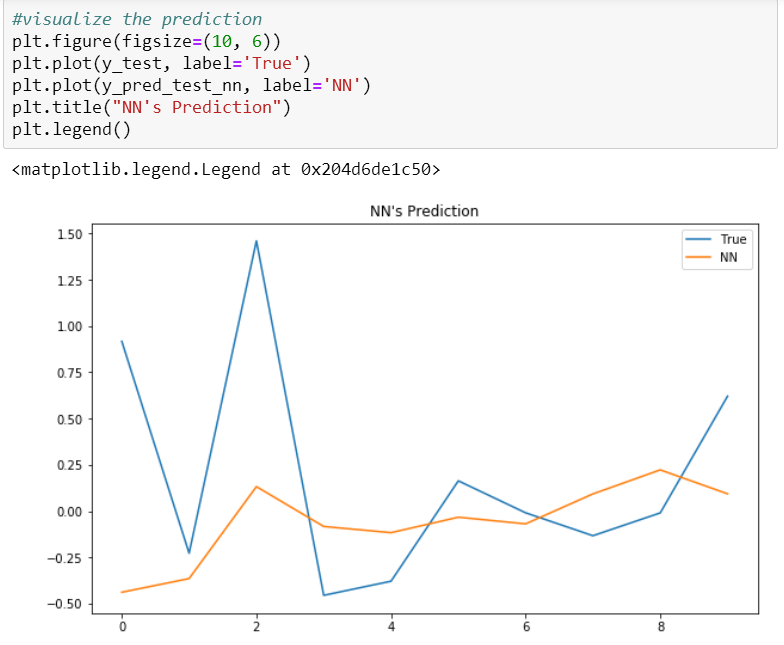


Figure 10

As we can see, the neural network prediction result is kind of flat and is not as well as the ARIMA model does.

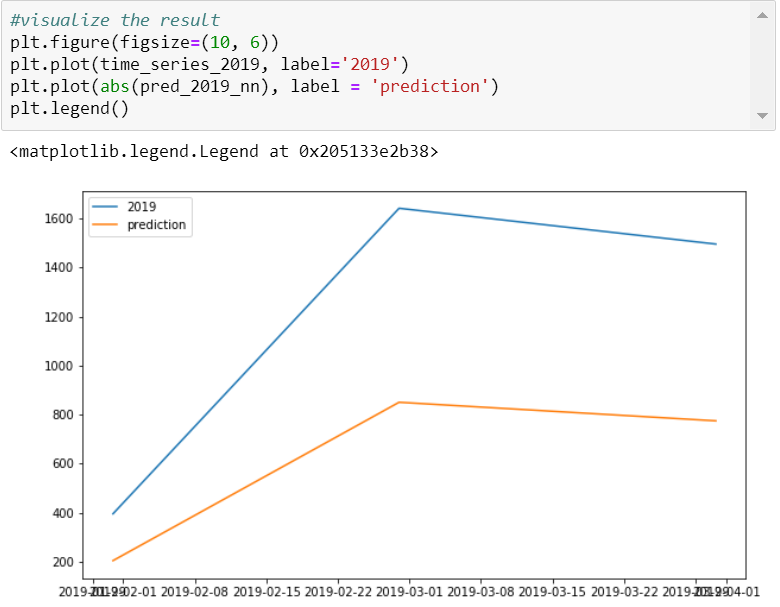


Figure 11

To test further, I put the 2019 at once data into the neural network model, and there is an obvious difference between the prediction and true value.

**Conclusion**

According to the existing dataset, ARIMA model works better than neural network because the dataset is not that large, and we only have current 3 years’ data to build and train the model. While it doesn’t necessarily mean that neural network does not apply for this project. At once business will always happen so the data volume will definitely boom in the future. Neural network will realize its power as the at once data accumulate over time.