Laura Valentina Gonzalez Bohorquez Final Project

MSDS 596: Regression and Time Series 12/08/2020

MSFT Stock- ARIMA Prediction

I. Introduction

Stocks are high-risk high-reward financial instruments and their potential for high-reward makes them quite lucrative for stock traders. Just in the NYSE, the average volume of stocks traded on a daily basis can range anywhere from 2 billion to 6 billion shares [2]. Therefore, stock value prediction models are of immense use to stock traders who prefer to use data to make purchase decisions instead of engaging in mere speculation. Since ARIMA models have a strong potential for short term predictions, such models can be attempted to be fit stock data.

In this analysis, multiple ARIMA models were fit for the Microsoft Corporation (MSFT) closing stock prices, pertaining to the last 20 years. These models were then used to predict the value of the stock for the following days. The predictions were compared with the actual data, in order to assess the accuracy of each model. The goal was to identify an ARIMA model that was simple, and had a good short term prediction accuracy.

II. Datasets

Stock datasets are publicly available through Yahoo Finance. The Historical data is quite extensive, starting from March 13, 1986 to present day. The data itself is classified by date and prices: open, high, low, close, average close, and volume. This analysis focuses on MSFT stock

closing prices with daily frequency over a 20 year period- from January 1st, 2000 to December 31st, 2020. All analyses used throughout this project are in R v.4.0.2.

III. Data Analytic Strategy

First, the MSFT stock data has been fetched from Yahoo finance and then plotted in a chart series to get a broad overview. The chart displayed open and close prices of the stock, as well as the traded volume, for each day in the selected 20 years. Since this analysis focused on closing prices, such data has been extracted and saved separately. The closing data's ACF and PACF plots were graphed, in order to analyze the lags. Later, the difference function was applied, which portrayed a constant mean after just one difference. Furthermore, in order to test for stationarity, the Augmented Dickey-Fuller (ADF) test was utilized. The test revealed clear non-stationarity in the data, and indicated that further visual inspection of the data was necessary. Initially, the 'auto.arima()' function was used to fit the given data into a model. The function suggested a (1, 2, 0) model which was labeled as 'Model 1' for the predictions. Subsequently, the ACF plot was analyzed to identify possible p values that can be considered for further models; and opted for the following:

#Model2: p = 7 because that's the position after which there is a significantly discernible cutoff beyond the blue line.

#Model3: p = 26 because that's the right most position immediately after which there is a cutoff beyond the blue line (but barely); This model would very likely be a classic case of over-fit.

#Model4: p = 1 because the cut off starts right at the beginning in the ACF plot

For each arima model generated, the residuals, ACF, and PACF plots have been generated. Even though there are still cut off points on the ACF plot for the ARIMA(1,1,7), they have been ignored, as the prediction accuracy of this model needs to be compared with the other chosen models, before further improvement. Similarly, such inference becomes applicable for the ARIMA(1, 1, 1) model.

The stock data for the following 50 days was extracted and used as a baseline measure, while evaluating the accuracy of each model's predictions. After obtaining the difference from both the actual stock price and the model predictions, the day-wise accuracy percentages were produced for each model and plotted accordingly. Additionally, least accurate prediction rate and the corresponding occurrence day, were determined for each model.

IV. Results

The Dickey-Fuller Test results (in Figure 3) showed a high p-value that indicated clear non-stationarity. Therefore, the data was fit into four different ARIMA models (based on analyzing the lags in the ACF plot) and captured the accuracy of these models to identify the best model for predicting future stock value of MSFT.

Model 1 – ARIMA(1, 2, 0): This model was suggested by the auto.arima() function.
It predicted the future values of stock with >95% accuracy for the first 7 days, then the accuracy of the predictions dropped sharply, and reached a minimum of 77.2% around the Day 33 mark.

Model 2 – ARIMA(1, 1, 7): This model predicted the future values of stock with >95% accuracy for the first 10 days, between 94% and 95% for the next 3 days, and then again >95% until Day 18 (except on Day 15). After Day 18, the accuracy of the predictions dropped sharply and reached a minimum of 83.6% around the Day 27 mark.

Model 3 – ARIMA(1, 1, 26): This model predicted the future values of stock with >95% accuracy for the first 11 days, between 94% and 95% for the next 2 days, and then again >95% until Day 18 (except on Day 15). After Day 18, the accuracy of the predictions dropped sharply and reached a minimum of 83.57% around the Day 27 mark.

Model 4 – ARIMA(1, 1, 1): This model predicted the future values of stock with >95% accuracy for the first 10 days, between 94% and 95% for the next 3 days, and then again >95% until Day 18 (except on Day 15). After Day 18, the accuracy of the predictions dropped sharply and reached a minimum of 83.59% around the Day 27 mark.

V. Conclusions

1. Model 4 (ARIMA(1, 1, 1)) has a prediction accuracy quite comparable to that of Models 2 and 3 which we generated after analyzing the ACF plots.
2. The increase in prediction accuracy with increase in model complexity (from Model 4 to Model 2 to Model 3) is quite small.
    a. Therefore, using ARIMA(1, 1, 1) can save time for creating the model, reduces the model complexity by quite a bit, and still provides predictions almost similar to the ARIMA (1, 1, 7) or the ARIMA (1, 1, 26) in this case.
    b. Additionally, the ARIMA(1, 1, 7) model does not need to be optimized any further.
3. The model suggested by the auto.arima() function performed the worst of all models.

4. As expected, Model 3 – ARIMA (1, 1, 26) – turned out to be an overfit and provided hardly any increment in prediction accuracy in comparison to Model 2 – the ARIMA(1, 1, 7).
5. However, as it turned out, Model 2 ARIMA(1, 1, 7) seems to be an overfit compared to ARIMA(1, 1, 1).
6. Therefore, ARIMA(1, 1, 1) model can be safely used to predict the value of MSFT stocks for the next 10 future days within a reasonable accuracy (>95%).
7. In the future, fitting ARIMA models for other stock's data should be considered, since the model potentially predicts future stock closing values within a reasonable accuracy.

Reference

[1]"Microsoft Corporation (MSFT) Stock Historical Prices & Data." *Yahoo! Finance*, Yahoo!, 6 Dec. 2020, finance.yahoo.com/quote/MSFT/history?p=MSFT.

[2] "This Day In Market History: Fewer Than 1 Million Shares Trade On The NYSE." *Yahoo! Finance*, Yahoo!, finance.yahoo.com/news/day-market-history-fewer-1-110000456.html.

Table 1. Day-wise accuracy percentages for each ARIMA Model

	Day-wise accuracy for each ARIMA model			
FutureDay/ARIMA Model	**(1,2,0)**	**(1,1,7)**	**(1,1,26)**	**(1,1,1)**
Day 1	97.77872	98.21237	98.22845	98.19107
Day 2	98.84781	99.39517	99.41201	99.43406
Day 3	98.30467	99.18678	99.22981	99.18035
Day 4	98.98285	99.8546	99.73326	99.90559
Day 5	97.17555	98.56607	98.60429	98.52582
Day 6	95.74075	97.32385	97.37471	97.31054
Day 7	95.94048	97.79038	97.9066	97.76311
Day 8	94.56243	96.6206	96.70703	96.60166
Day 9	94.99114	97.31043	97.37166	97.28692
Day 10	94.14056	96.68175	96.80842	96.66096
Day 11	92.21102	94.9435	95.02453	94.92169
Day 12	91.46381	94.41429	94.4574	94.39341
Day 13	91.55841	94.75498	94.76497	94.73357
Day 14	91.7644	95.2122	95.20731	95.19095
Day 15	90.96832	94.62983	94.57139	94.60857
Day 16	91.65731	95.59302	95.52055	95.57163
Day 17	92.97511	97.21888	97.13205	97.19707
Day 18	90.95178	95.35038	95.24991	95.32903
Day 19	89.32257	93.88645	93.80027	93.8654
Day 20	86.6457	91.31078	91.27718	91.29032
Day 21	87.71383	92.6786	92.62263	92.65783
Day 22	85.40203	90.47297	90.41934	90.45269
Day 23	82.46329	87.58981	87.565	87.57018
Day 24	82.34668	87.69693	87.66773	87.67727
Day 25	80.46097	85.91557	85.87143	85.89632
Day 26	80.13448	85.7941	85.75739	85.77487
Day 27	77.88453	83.60719	83.5669	83.58845
Day 28	79.47132	85.53826	85.49994	85.51909
Day 29	79.14337	85.41322	85.37314	85.39408
Day 30	79.36123	85.87816	85.839	85.85891
Day 31	78.44798	85.1183	85.07878	85.09922
Day 32	77.45134	84.26362	84.22494	84.24473
Day 33	77.22178	84.24112	84.20217	84.22224
Day 34	78.20722	85.54754	85.50816	85.52837
Day 35	80.54122	88.3402	88.29943	88.32041
Day 36	83.94135	92.32066	92.27812	92.29997
Day 37	85.11702	93.86967	93.82638	93.84864
Day 38	83.83675	92.71127	92.66854	92.69049
Day 39	89.94425	99.73876	99.69277	99.71641
Day 40	87.57645	97.38089	97.33599	97.35906
Day 41	81.88635	91.3055	91.2634	91.28504
Day 42	85.77	95.90102	95.85681	95.87953
Day 43	82.5031	92.5047	92.46205	92.48397
Day 44	84.39156	94.88589	94.84214	94.86462
Day 45	86.60435	97.64607	97.60105	97.62419
Day 46	92.64073	95.2551	95.30339	95.27857
Day 47	86.46798	98.0405	97.9953	98.01853
Day 48	90.31639	97.30732	97.35467	97.33033
Day 49	99.49798	86.5477	86.60001	86.57312
Day 50	86.8669	99.33058	99.28479	99.30832

Figure 1. Chart Series of MSFT stock

Figure 2. The Trend Plot for MSFT closing price

Figure 3. Results of the ADF Test

```
        Augmented Dickey-Fuller Test

data:  MSFT_ClosePrice
Dickey-Fuller = 1.6659, Lag order = 17, p-value = 0.99
alternative hypothesis: stationary
```

Figure 4. Model Residuals, ACF, and PACF plots for Model 1 ARIMA(1, 2, 0)

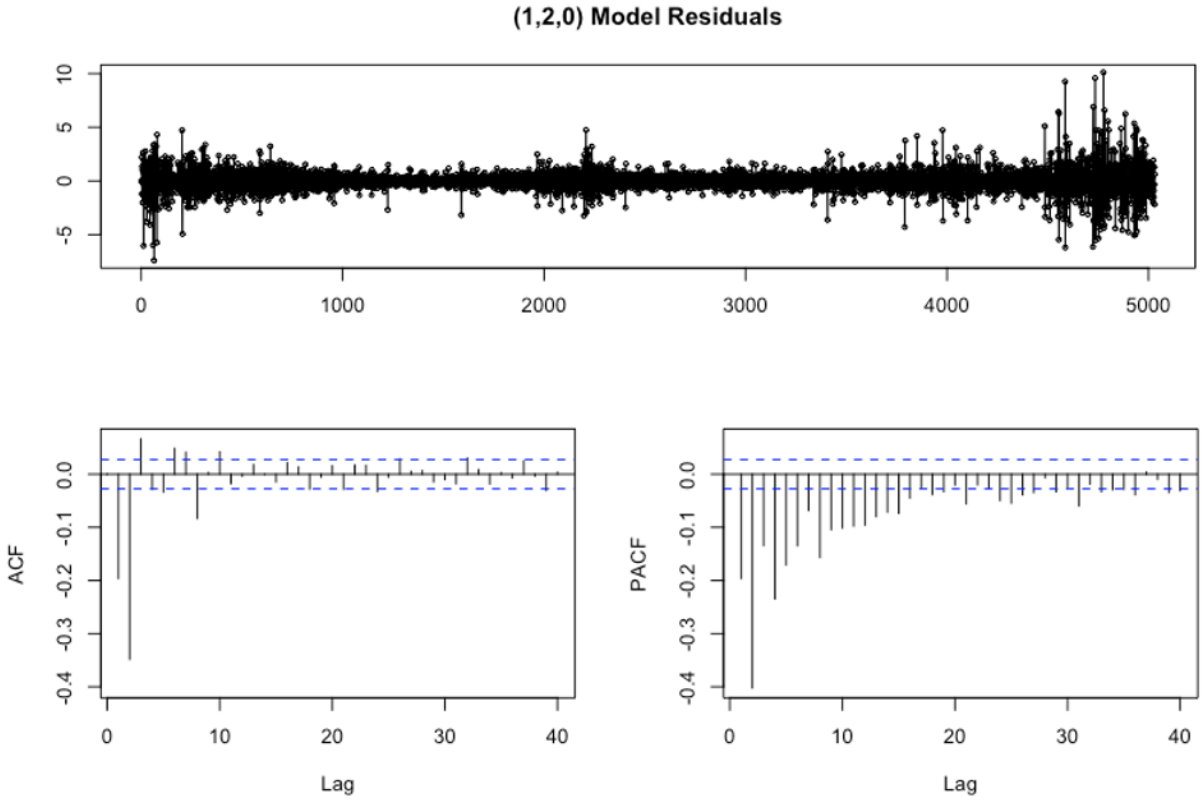

Figure 5. Model Residuals, ACF, and PACF plots for Model 2 ARIMA(1, 1, 7)

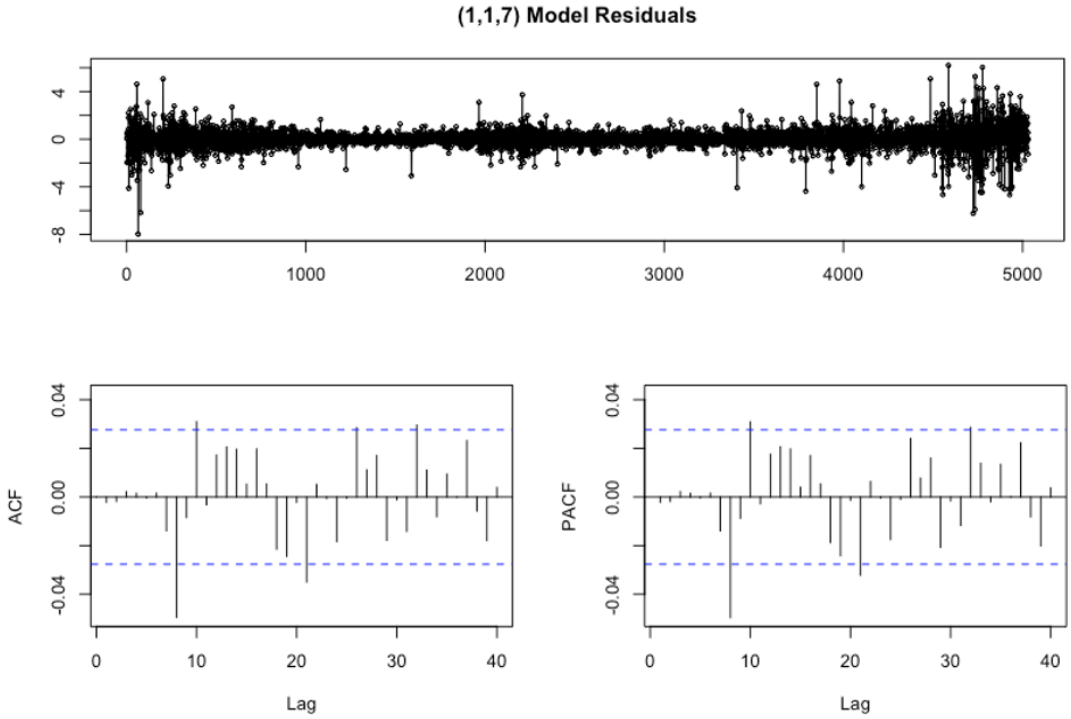

Figure 6. Model Residuals, ACF, and PACF plots for Model 3 ARIMA(1, 1, 26)

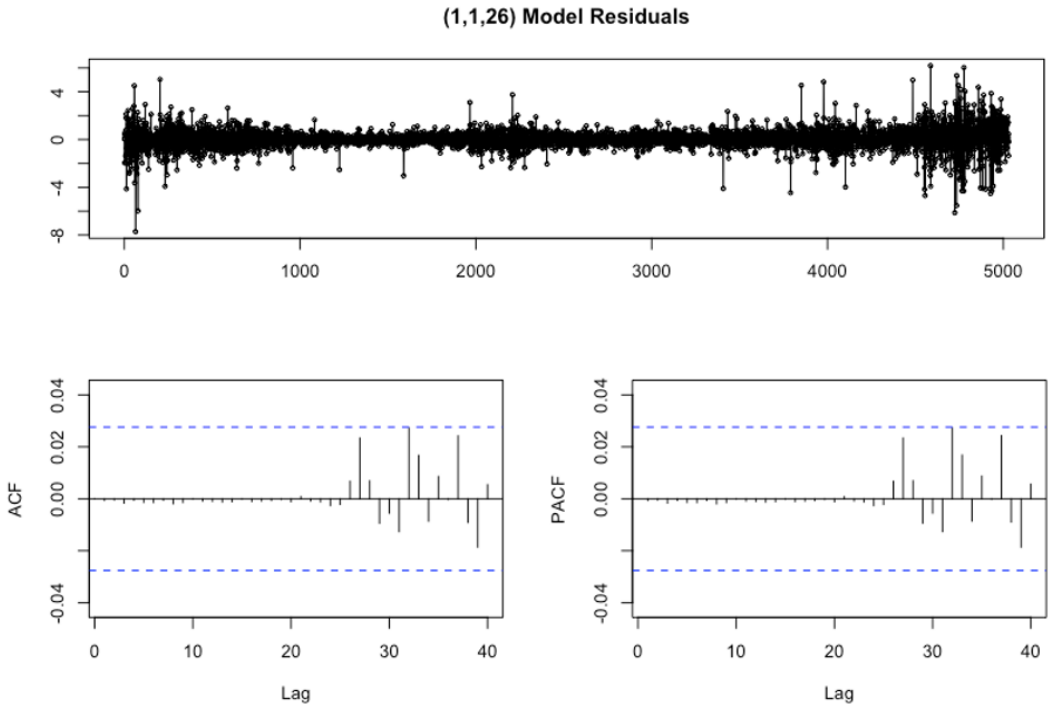

Figure 7. Model Residuals, ACF, and PACF plots for Model 4 ARIMA(1, 1, 1)

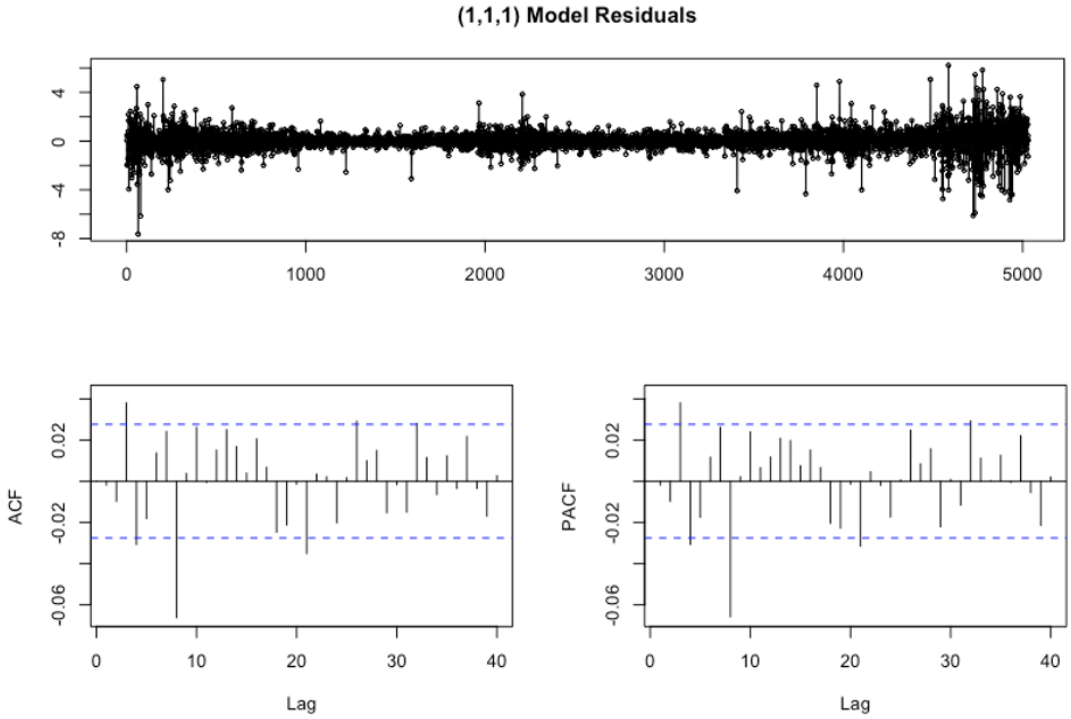

Figure 8. Day-wise prediction accuracy for the four ARIMA models