Lab Report

GROUP 6

February 26, 2022

Financial Analytics Lab [BM49002] Assignment 5

Pritam Mallick [18CS10042]
Rohit Jonwal [18CS10046]
Sigangsa Baglari [18EC3FP11]
G Vishnu Vamshi [19EE10023]

Introduction

ARIMA Forecasting: Auto-Regressive Integrated Moving-Average is a forecasting algorithm based on the assumption that previous values carry inherent information and can be used to predict future values. The ARIMA model takes in three parameters (p, d, q):

p is the order of the AR term
q is the order of the MA term
d is the number of differencing

Time Series Modeling: Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly.

Objective

In this assignment, our objective is to perform ARIMA forecasting on the **Microsoft Stock Price** time-series historical data obtained from Yahoo Finance. A sample of the dataset is shown in the following **Fig 5.1**.

Currency in USD

Download

Date	Open	High	Low	Close*	Adj Close**	Volume
Feb 28, 2022	294.31	298.54	293.81	294.59	294.59	14,267,312
Feb 25, 2022	295.14	297.63	291.65	297.31	297.31	32,528,000
Feb 24, 2022	272.51	295.16	271.52	294.59	294.59	56,989,700
Feb 23, 2022	290.18	291.70	280.10	280.27	280.27	37,811,200
Feb 22, 2022	285.00	291.54	284.50	287.72	287.72	41,736,100
Feb 18, 2022	293.05	293.86	286.31	287.93	287.93	34,223,200

Fig 5.1 MSFT Stock Price Time-series historical data

Theory and Definitions

Seasonal Decomposition: Time series decomposition involves thinking of a series as a combination of level, trend, seasonality, and noise components. Decomposition provides a useful abstract model for thinking about time series generally and for better understanding problems during time series analysis and forecasting.

Autocorrelation: Autocorrelation is a mathematical representation of the degree of similarity between a given time series and a *lagged* version of itself over successive time intervals. It's conceptually similar to the correlation between two different time series, but autocorrelation uses the same time series twice: once in its original form and once lagged one or more time periods.

Stationarity: In the most intuitive sense, stationarity means that the statistical properties of a process generating a time series do not change over time. It does not mean that the series does not change over time, just that the *way* it changes does not itself change over time.

Durbin Watson Statistic: The Durbin Watson (DW) statistic is a test for autocorrelation in the residuals from a statistical model or regression analysis. The Durbin-Watson statistic will always have a value ranging between 0 and 4. A value of 2.0 indicates there is *no autocorrelation* detected in the sample. Values from 0 to less than 2 point to *positive autocorrelation* and values from 2 to 4 mean *negative autocorrelation*.

Augmented Dickey-Fuller Test: ADF (Augmented Dickey-Fuller) test is a statistical significance test which means the test will give results in hypothesis tests with null and alternative hypotheses. As a result, we will have a p-value from which we will need to make inferences about the time series, whether it is stationary or not. If **P-value > 0.05**, then the given dataset does not reject the *H0 hypothesis*, hence is *non-stationary*. Otherwise, the H0 hypothesis is rejected and the dataset turns out to be stationary.

Data

The **MSFT Stock** dataset consists of *Open*, *Close* and *Adj Close* values over a span of many years. However, in this assignment, the time-series data analysis of **Adj Close** is our primary focus. Adjusted close is **the closing price after adjustments for all applicable splits and dividend distributions**. The time-series chart of the MSFT stock price is plotted in the following **Fig 5.2**.

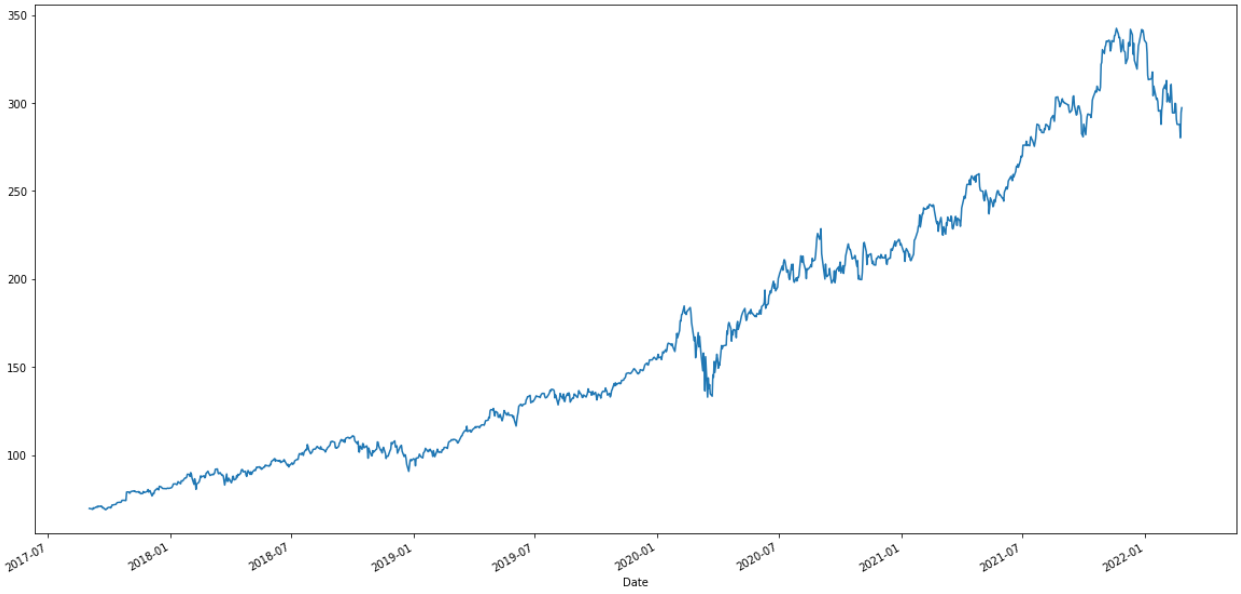


Fig 5.2 MSFT Stock Price Time-series chart

Note: It is apparent that the time-series chart shows a **growing trend** with **seasonality**. However, the case of stationarity is not observable in this trend, because the statistical properties seem to alter widely across windows of time, thus the data might be **non-stationary**. We proceeded by reserving the **last 10 data points** for **testing** purposes and the rest for **training** the model.

Time-series Analysis

We shall be performing certain tests to observe the properties and behaviour of the dataset to build a suitable model and make accurate predictions.

Durbin Watson Test (Autocorrelation)

The DW Statistic ranges from a value of 0 to 4 with the following statistics.

DW Statistic	Result
0 … 1.5	High Positive autocorrelations
1.5 … 2	Low Positive autocorrelations
2	No autocorrelation
2 … 2.5	Low Negative autocorrelations
2 .5 … 4	High Negative autocorrelations

After conducting the Durbin Watson test in Python Environment using "durbin_watson" imported from "statsmodels.stats.statstools", we obtained a value of DW Statistic equal to **0.00173758**. Hence, we concluded that the dataset has **very high positive autocorrelation** as it is close to 0.

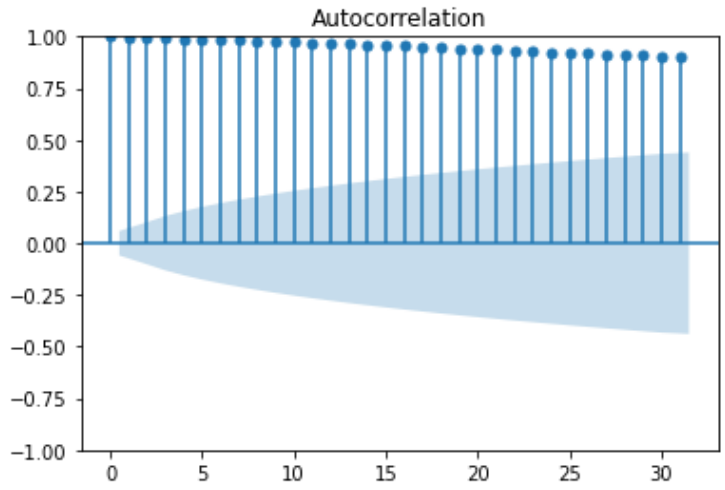


Fig 5.3 Autocorrelation on MSFT Stock Price data

Thus, we needed to remove autocorrelation issues from our dataset. To achieve this objective, we performed **first-order differencing** on the original dataset and then performed the Durbin Watson Test on what we obtained.

$$\mathbf{Y}_d(t) = \mathbf{Y}(t+1) - \mathbf{Y}(t)$$

This time we obtained a value of DW Statistic close to **2.4181**, indicating a **low negative autocorrelation**. The plots of the differenced values and autocorrelation statistics are shown in the following **Fig 5.4(a) and (b)**.

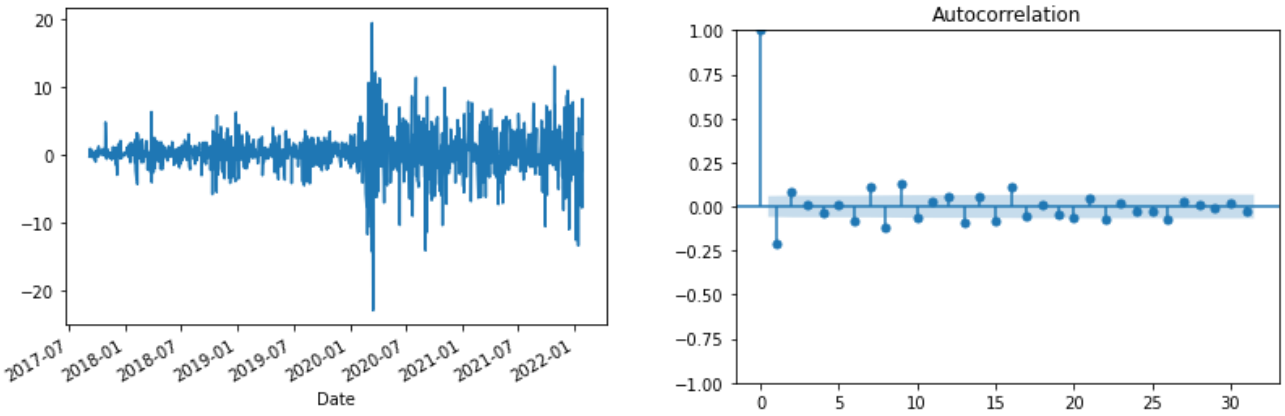


Fig 5.4(a) Plot of differenced data **(b)** Autocorrelation on differenced data

So, we concluded that the depth of the ARIMA model (value of the parameter **d**) is **1**, as we eliminated autocorrelation after one level of difference.

Augmented Dickey-Fuller Test (Stationarity)

The P-value of the ADF test is compared around the value 0.05

P-value	Result
> 0.05	H0 rejection fails => Non-stationary
< 0.05	H0 rejection passes => Stationary

After conducting the ADF test in Python Environment using "adfuller" imported from "statsmodels.tsa.statstools", we obtained a value of P-value equal to **0.9754595179946**. Hence, we concluded that the dataset is **non-stationary** as P-value is close to 1.

```
ADF:  0.2605330379552538
P-value:  0.9754595179946096
No. of Lags:  20
No. of observations used:  1088
Critical Values:
        1% :  -3.4363746281360426
        5% :  -2.864200133611212
        10% :  -2.568186343567528

The given dataset does not reject H0 hypothesis, hence is non-stationary.
```

However, we performed the ADF test once again on the differenced dataset that we previously obtained. This time we obtained a P-value almost equal to **zero**. Hence it was observed that the differenced dataset is **stationary**.

```
ADF:  -8.059630463360655
P-value:  1.656363324984809e-12
No. of Lags:  19
No. of observations used:  1088
Critical Values:
        1% :  -3.4363746281360426
        5% :  -2.864200133611212
        10% :  -2.568186343567528

The given dataset rejects H0 hypothesis, hence is stationary.
```

Seasonal Decomposition of Time Series

We conducted the **multiplicative** Seasonal decomposition of the time-series data using "seasonal_decompose" imported from "statsmodels.tsa.seasonal" in Python Environment. The following results were obtained shown in **Fig 5.5**.

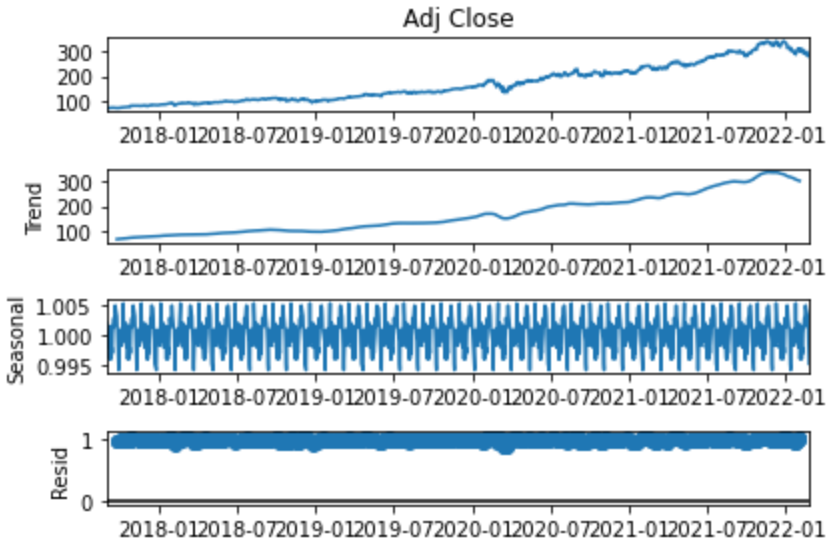


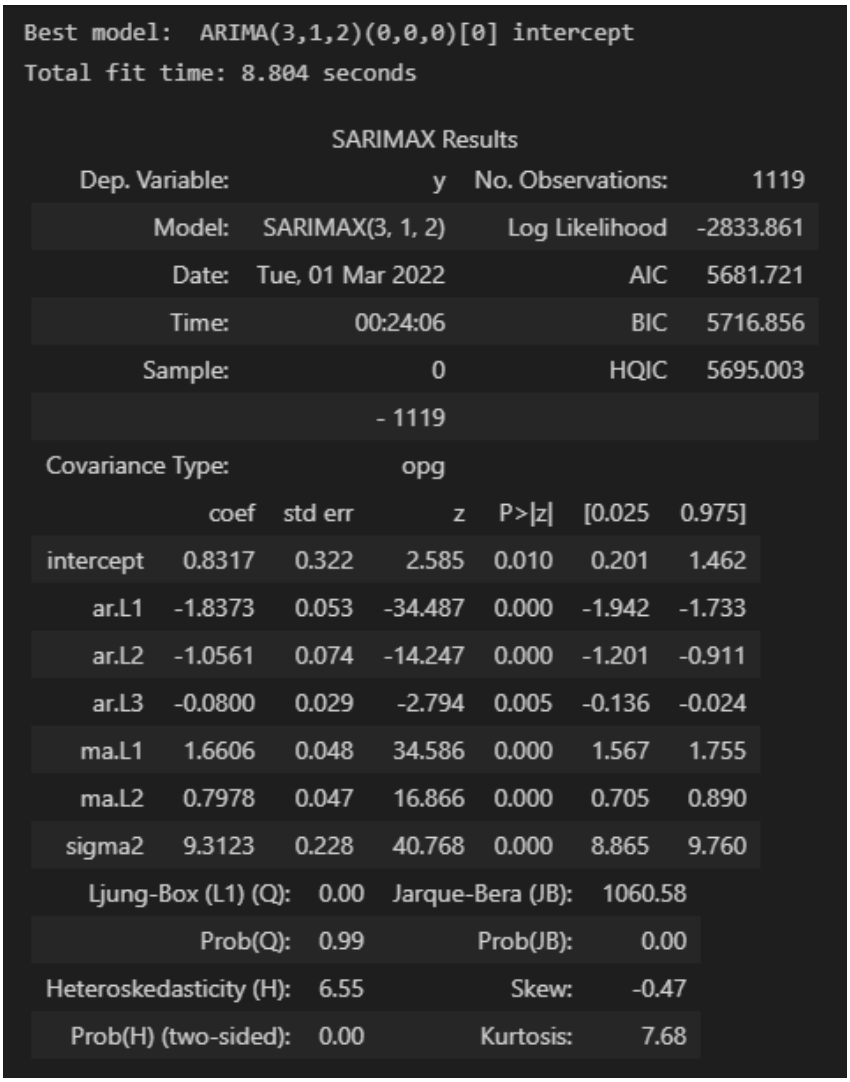
Fig 5.5 Seasonal Decomposition Plots

Y(t) = **Trend**(t) **X Seasonal**(t) **X Resid**(t)

As observable from the figure, we have **Trends**, **Seasonality** and **Residuals** in the time series. The trend plot is indicative of the increasing valuation of Microsoft Corporation within a span of 3-4 years. The seasonality explains the cyclic and repetitive pattern in stock price within every 30 data points. The residuals are the cause of the minor fluctuations in the daily stock price.

Time-series Model Specification

We ran the **auto_arima** test on our time-series data in order to obtain the parameters to be used in the building of our ARIMA model for prediction. The result of our test is shown in the following **Fig 5.6**.



```
Best model:  ARIMA(3,1,2)(0,0,0)[0] intercept
Total fit time: 8.804 seconds
```

SARIMAX Results

Dep. Variable:	y	No. Observations:	1119
Model:	SARIMAX(3, 1, 2)	Log Likelihood	-2833.861
Date:	Tue, 01 Mar 2022	AIC	5681.721
Time:	00:24:06	BIC	5716.856
Sample:	0	HQIC	5695.003
	- 1119		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	0.8317	0.322	2.585	0.010	0.201	1.462
ar.L1	-1.8373	0.053	-34.487	0.000	-1.942	-1.733
ar.L2	-1.0561	0.074	-14.247	0.000	-1.201	-0.911
ar.L3	-0.0800	0.029	-2.794	0.005	-0.136	-0.024
ma.L1	1.6606	0.048	34.586	0.000	1.567	1.755
ma.L2	0.7978	0.047	16.866	0.000	0.705	0.890
sigma2	9.3123	0.228	40.768	0.000	8.865	9.760

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	1060.58
Prob(Q):	0.99	Prob(JB):	0.00
Heteroskedasticity (H):	6.55	Skew:	-0.47
Prob(H) (two-sided):	0.00	Kurtosis:	7.68

Fig 5.6 Specifications of best fit ARIMA Model

As shown in the above figure, it was decided that the ARIMA**(3,1,2)** model was the best fit for our time-series predictive analysis. As expected, we obtained **d=1** from the auto_arima function which agrees with our Time-series analysis.

ARIMA(3,1,2) Prediction

We used the last 10 data points from the data to test the model, while the rest of the points were used for training. We obtained the following predictions that were compared with the testing data in the following **Fig 5.7**.

```
Pred:  305.247789633383  Actual:  294.431213
Pred:  302.15183108237693  Actual:  294.391296
Pred:  304.8075785827565  Actual:  299.850006
Pred:  302.9260364284111  Actual:  299.5
Pred:  303.8229490618605  Actual:  290.730011
Pred:  303.9508563012856  Actual:  287.929993
Pred:  302.91957641036424  Actual:  287.720001
Pred:  304.60572527410966  Actual:  280.269989
Pred:  302.5893966012313  Actual:  294.589996
Pred:  304.59259705225435  Actual:  297.309998
```

Fig 5.7 Predictions of trained ARIMA(3,1,2)

As shown in the above figure, the resulting model is healthy, as it predicts data with a very low standard error. We also evaluated the **Mean squared error** and compared it to the **Mean** of the data. **MSE = 12.5774 and Mean = 292.6722** with **MSE/Mean = 4.297%**

The plot of the 10 predicted values is compared with the 10 original test values in the following **Fig 5.8**.

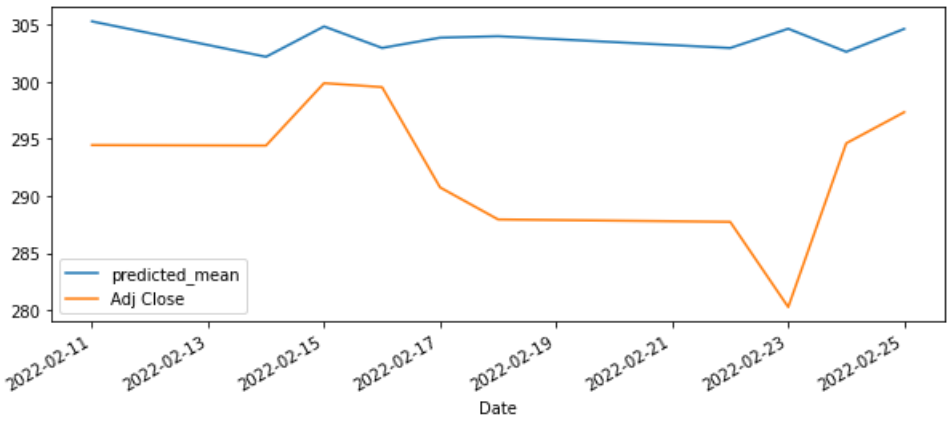


Fig 5.8 Predicted values vs.Original Test Values

We also plotted the residuals of the predicted data points, i.e. the differences between predicted and test data points. **Fig 5.9** displays the residual plot.

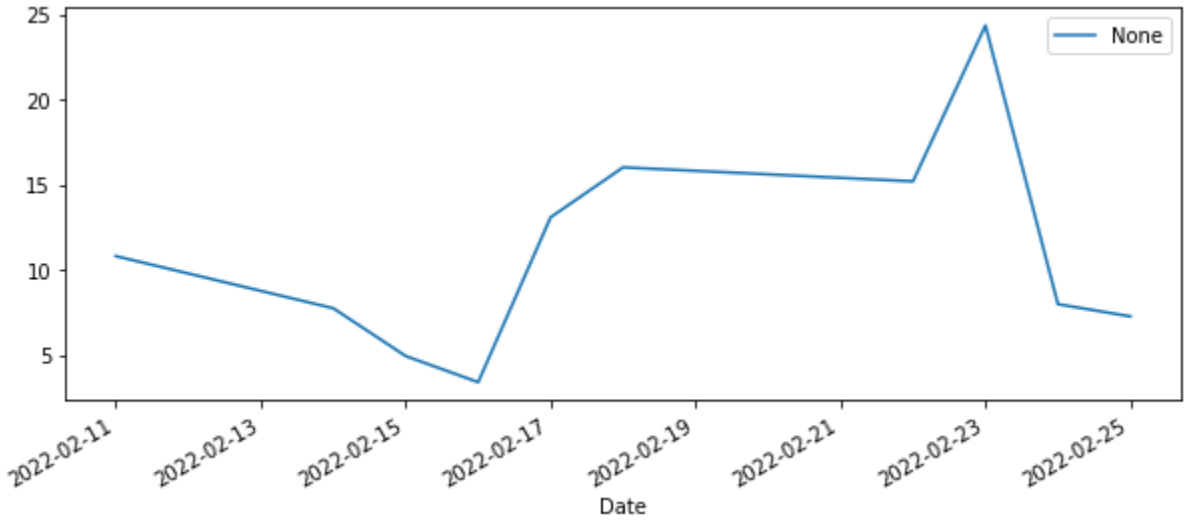


Fig 5.9 Residual plot

Conclusion

This assignment focused on predictive analysis of Microsoft Corporation Stock Price using ARIMA Forecasting models. The data was extracted from Yahoo Finance available here, starting from **09-01-2017** to **25-02-2022**. There are some missing days as these may account for weekends when the stock markets are closed or on some public holidays. In such cases, we eliminated the data points containing **'NaN'** values. We discovered **high-positive autocorrelation** and **non-stationarity** in our original dataset, which required one level of differencing in order to fit into a healthy ARIMA model. We also encountered unpredictable residuals in our predictions because of the recent events of the falling stock market.