Netflix Stock Price Forecasting using ARIMA Model in Python

Samriddhi Saxena

A thesis presented as a part of

B.Sc. (Hons.) Applied Statistics and Analytics, 2022-2023

Supervised by: Mrs. Snigdha Banerjee

School Of Statistics and School of Data Science and Forecasting Devi Ahilya University, Indore

DECLARATION

I, Samriddhi Saxena, declare that this is my original work and that it has never been presented to any institution or university for the award of Degree. In addition, I have referenced correctly all literature and sources used in this work and this work is fully compliant with the School of Statistics and School of Data Science and Forecasting, Devi Ahilya University's academic honesty policy.

ACKNOWLEDGMENTS

I would like to express my special thanks of gratitude to Mrs. Snigdha Banerjee, my teacher and project supervisor for her patient guidance from the beginning with constructive and valuable suggestions provided throughout the planning and improvement of my dissertation. She also helped me in doing a lot of Research, and her ability to give her time generously has been particularly valued.

Secondly, I would like to thank my family and friends for their support and encouragement throughout my course.

Samriddhi Saxena

B.Sc. (Hons.) Applied Statistics and Analytics, V Sem

ABSTRACT

This Thesis titled- “Stock price forecasting using time series model in python” focused on the comparison of the performance of time series models to predict the stock price for Netflix stocks. Forecasting and stock price analysis is important in finance and economics. Time series forecasting can be applied on any set of variables that change over time. For stocks or share prices, time series forecasting is common to track the price movement of the security over time. There is considerable past research work available on time series forecasting. In this thesis, a comparative study of time series forecasting using ARIMA (autoregressive integrated moving average) model has been explored. Historical stock price data was obtained from Yahoo Finance and used to build these models for comparative purposes.

CONTENT

List of Tables and Figures

List of Abbreviations

- ACF - Autocorrelation function
- PACF - Partial autocorrelation function
- adf – Augmented Dicky-Fuller test
- AR - Autoregressive
- ARIMA – Autoregressive Integrated Moving Average
- ARMA - Autoregressive moving average model
- TS - Time series
- EDA – Exploratory data Analysis ➢ CSV - Comma-separated values
- PDF - Portable Document Format

1. INTRODUCTION

“Forecasts create the mirage that the future is knowable.” — financial historian Peter L. Bernstein

1.1. Terms and Definitions

What is time series forecasting?

A time series is a collection of observations of well-defined data items obtained through repeated measurements over time. Time series analysis comprises methods for analysing time series data in order to extract meaningful statistics and other characteristics of the data. In the simplest terms, time-series forecasting is a technique that utilizes historical and current data to predict future values over a period of time or a specific point in the future.

Brockwell, P.J., Davis, R.A. and Calder, M.V., 2002, and Granger, C.W.J. and Newbold, P., 2014 stated that time series is a set of observations of N number of elements each of which recorded at a particular time (t). A time series is a sequence of values recorded in order by the time parameter. Such as medical, meteorological (rainfall, temperature), astronomy, economic forecasting, sales forecasting, stock market analysis, yield projections, process and quality control, inventory studies, workload projections, utility studies, census analysis, and many more. Time series studies do not typically make use of the majority of traditional

statistical techniques and procedures, hence new tools for Time Series analysis have been developed.

Time series forecasting is a method for making future predictions and inferences (Granger, C.W.J. and Newbold, 2014). With these methods, initial hypothetical probability models had to be built up, parameters had to be estimated, the data had to be checked for proper fit, and the fitted model might even be used to improve understanding of the mechanism generating the series. Depending on the specific application field, a model that has been produced to a suitable level may be used in a number of different ways. The modelling and forecasting of time series can be done in a variety of ways. The right technique is chosen based on the application type and user desire.

In essence, it's important to distinguish between a forecasted or predicted value of y_t that was made in a prior time period and a fitted value of y_t that was obtained by estimating the parameters in a time series model using historical data (Montgomery, D.C., Jennings, C.L., and Kulahci, 2015). The lead-time -τ forecast error is the outcome of a forecast of y_t that was made at time period t-τ.

$$e_t(\tau) = y_t - \hat{y}_t(t-\tau)$$

For example, the lead -1 forecast error is

$$e_1(1) = y_t - \hat{y}_t(t-1)$$

The difference between the observation y_t and the value obtained by fitting a time series model to the data, or a fitted value $\hat{y}$ defined earlier, is called a residual, and is denoted by

$$e_t = y_t - \hat{y}_t$$

What is Stock?

A stock (Akhilesh Ganti, Mar 2019) (also known as "shares" or "equity) is a kind of security that implies proportionate proprietorship in the issuing organization. This qualifies the investor for that extent of the company's advantages and profit. Stocks are purchased and sold dominatingly on stock trades, however there can be private deals also, and the establishment of about each portfolio.

These transactions have to conform to government regulations which are meant to protect investors from fraudulent practices. Historically, they have outperformed most other investments over the long run. These investments can be purchased from most online stock brokers

An analysis (Lo, A.W. and Wang, J., 2001) begins with I investors indexed by i=1,…,I and J stocks indexed by j=1,…,J. There is always an assumption that all the stocks are risky and non-redundant. For each stock j, let N_{jt} be its total number of shares outstanding, D_{jt} its dividend, and P_{jt} its ex-dividend price at date t.

For notational convenience and without loss of generality, assume throughout that the total number of shares outstanding for each stock is constant over time, i.e.
N_{jt}= N_j, j=1,…,J.

For each investor i, let S^i_{jt} denote the number of shares of stock j he holds at date t.
Let

$$P_j = [P_{1t} \dots P_{jt}] T$$

Which denote the vector of stock prices and shares held in a given portfolio, where A_t denotes the transpose of a vector or matrix A.

Time Series Analysis for Stock Prediction

Time series can be applied on any set of variables that change over time. Time series are frequently used to track a security's price over time for stocks or share prices. The price of a security from the start to the end of a business day, the closing price of a daily business day, or perhaps the last day of every month over the past 15-20 years can be tracked to show this over the short term. The high point of the stock market is the seasonal trend and flow.

This might be helpful to track the evolution of a company's capital, securities, or any other connected economic indicator. It can also be used to compare the shifts in the selected data point to changes in other variables over the same time frame. For example, suppose analysing a time series of daily closing stock prices for Netflix stock from Yahoo finance over a period of five years. Obtain a list of all the closing prices for the stock from each business day for the past five years and list them in chronological order. This would be five years all closing price time series for the stock.

The fig 1.1 shows the graphical presentation of stock price

Fig. 1.1: Netflix Stock Price from Yahoo Finance

1.2. Dissertation Roadmap

In order to improve performance and accuracy for time series datasets, this thesis explored forecasting models, the algorithms utilized inside the model, and other optimization strategies. The different performance evaluation parameters used for the automatic selection of appropriate input variables and model-dependent variables as well as optimizing the model parameters.

Applied ARIMA model for predicting future values.

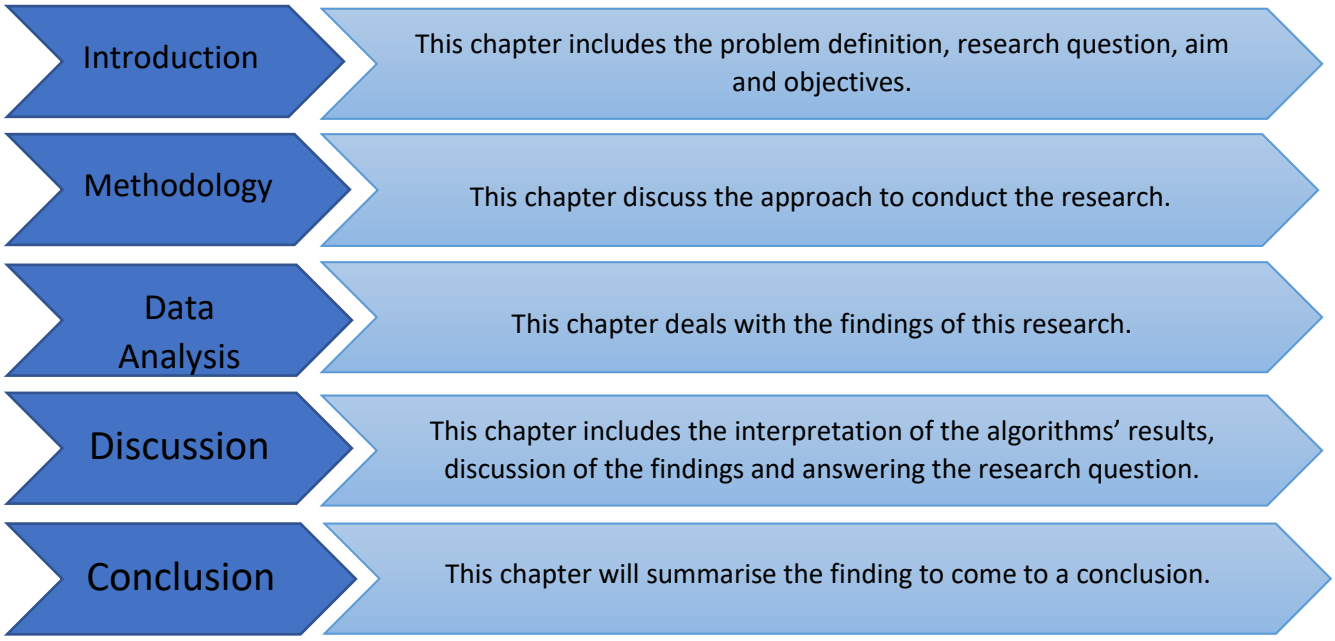

Fig. 1.2 Dissertation Roadmap

2. PROJECT AIM AND OBJECTIVE

Developing and using a predictive model by studying historical data that has a relationship between observations is a part of time series forecasting. Each and every step of the project, from the evaluation of forecast models to the fundamental challenges of the forecast research, is directly impacted by the decision that is made after applying the models. Forecasting stock prices is an example of attempting to determine a company's stock's future value.

Problem Statement: - *Time Series Analysis of Netflix Stock.*

Problem Type: Forecasting

Aim: Assist Netflix with decision relating to stock price forecasting using time series analysis.

Objective: To compare different statistical learning algorithms trained on Netflix stock data to forecast the estimated price of stocks.

The objectives are:

1) Model specification (or model identification);
2) Model fitting (or model estimation);
3) Model evaluation (or model accuracy assessment).

2.1. Forecasting Methods

1) Autoregressive integrated moving average (ARIMA)

2.2. Proposed approach

For this dissertation, proposed steps are:

Step 1: Data Understanding

1) Understanding the objective
2) Sample data collection
3) Describe and explore the data

Step 2: Data Preparation:

1) Feature selection
2) Transform the data to stabilize the attributes
3) Integrate and format data

Step 3: Data Modelling:

1) Examine the data to identify potential models
2) Testing and training
3) Forecasting

3. METHODOLOGY

It is a major shift to switch from machine learning to time-series forecasting. It was a difficult but rewarding exercise that advanced my knowledge of how machine learning may be used to forecast stock prices. An estimation of an unknown variable's value was the goal of a predictive model. Time (t) is an independent variable in a time series, and the target is a dependent variable. The anticipated value for y at time t(y') is the model's output.

To forecast stock prices, time series models, specifically ARIMA, were used. Sample data in CSV (Comma-separated values) format was collected from Yahoo Finance. To create the stock price data transformation and load in the Python script for model evaluation, the CRISP-DM methodology concept was used.

3.1. Time Series Components

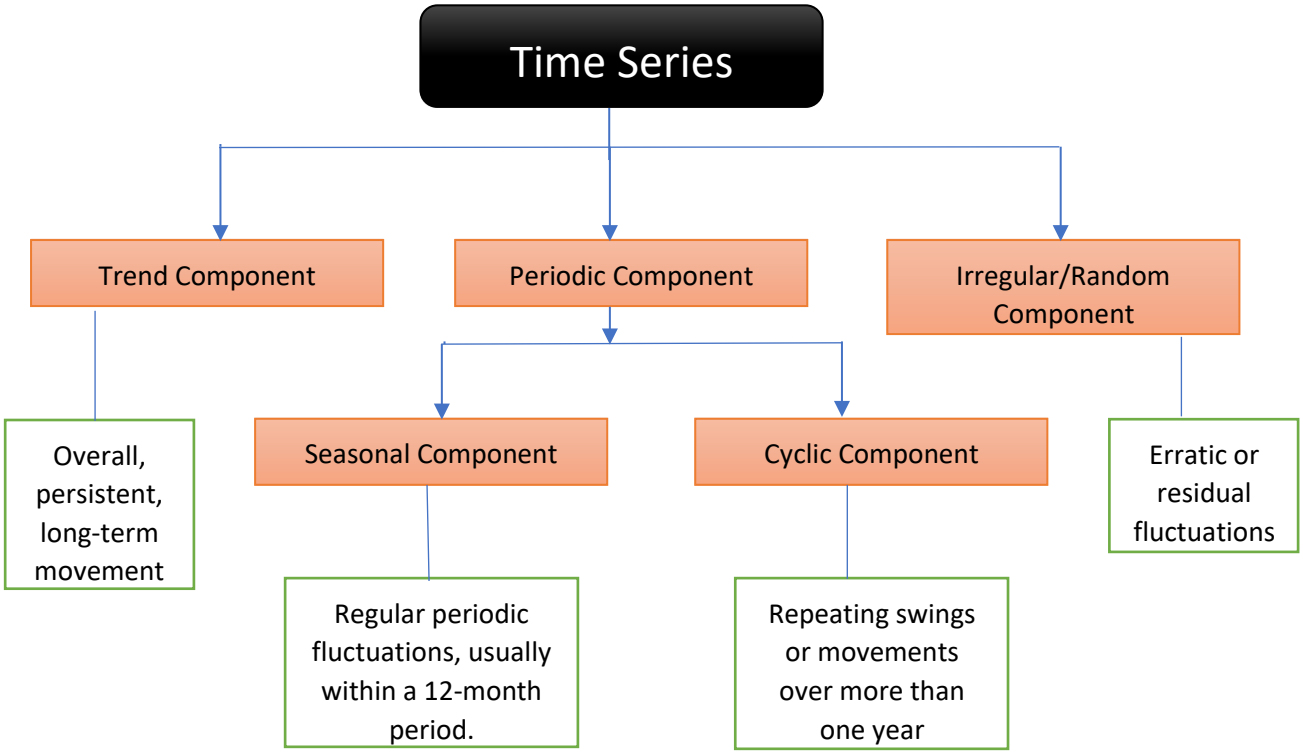

Fig. 3.1 Time Series Component

Trend: The trend is the component of a time series that represents variations of low frequency in a time series, the high and medium frequency fluctuations having been filtered out.

Seasonality: Seasonality is a characteristic of a time series in which the data experiences regular and predictable changes that recur every calendar year. Any predictable fluctuation or pattern that recurs or repeats over a one-year period is said to be seasonal.

Cycles: The cyclical component of a time series refers to (regular or periodic) fluctuations around the trend, excluding the irregular component, revealing a succession of phases of expansion and contraction.

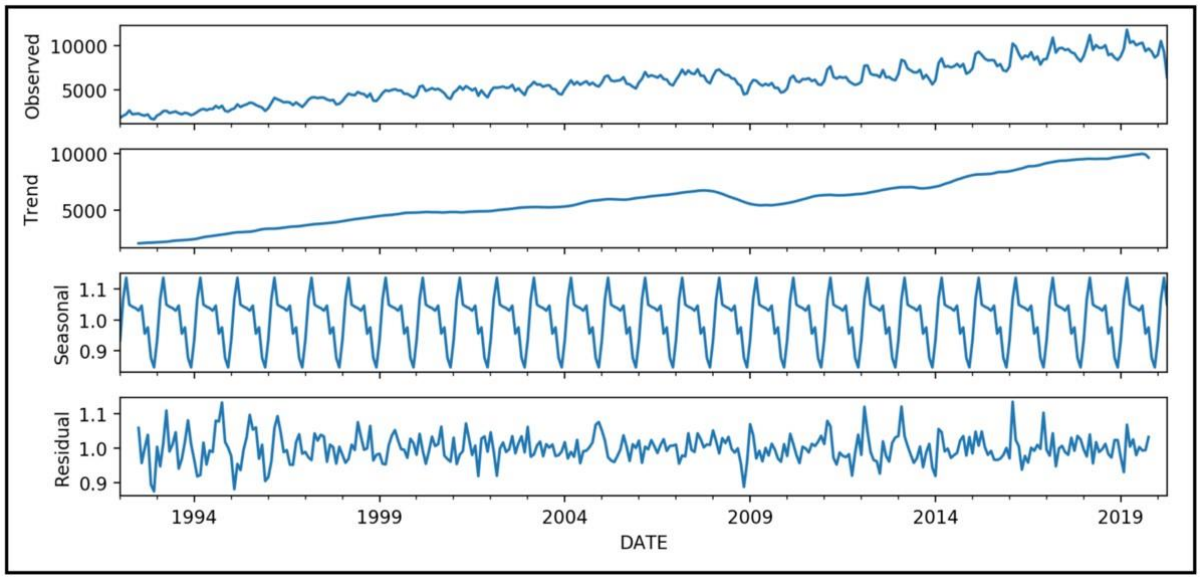

Fig. 3.2 Time Series Component Analysis

Time series components are highly important to analyse variable of interest in order to understand its behaviour, what patterns it has, and to be able to choose and fit an appropriate time-series model.

3.2. Procedures and Functions

For this thesis, applied models on training dataset to predict values for the test dataset. For seasonal time series, optimized model's parameters simultaneously to compare and discuss the result. Model chosen dependent based on the research done on time series forecasting is as follows:

3.2.1. ARIMA Model

ARIMA, abbreviated for 'Auto Regressive Integrated Moving Average', is a class of models that 'demonstrates' a given time series based on its previous values: its lags and the lagged errors in forecasting, so that equation can be utilized in order to forecast future values.

The ARIMA approach was popularized by Box and Jenkins (Devi, B.U., Sundar, D. and Alli, P., 2013), and ARIMA models are often referred to as Box-Jenkins models. The general transfer function model employed by the ARIMA procedure was discussed by Box and Tiao in 1975.

A class of statistical models known as the ARIMA model (Auto-Regressive Integrated Moving Average) is used to analyse and predict time series data. It specifically supports a number of

time series data standard structures. The evolving variable of interest is regressed on its prior values, as indicated by the AR component of ARIMA. The MA component of the equation shows that the regression error is essentially a linear combination of error components, the values of which happened simultaneously and at distinct points in the past.

Non-seasonal ARIMA models are generally denoted ARIMA (p, d, q). Where, ○ **p** =

the order of the AR term (number of time lags) ○ **q** = the order of the MA term

○ **d** = the number of differences required to make the time series stationary (degree of differencing/number of times the data have had past values subtracted)

Seasonal ARIMA models are usually denoted ARIMA(p, d, q) (P, D, Q)m, where m refers to the number of periods in each season, and the uppercase P, D, Q refers to the autoregressive, differencing, and moving average terms for the seasonal part of the ARIMA model.

ARIMA algorithm in three steps:

Step 1: Model identification

Step 2: Model estimation

Step 3: Forecasting

Stationary Series: A stationary series has no trends, its variations around its mean have a constant amplitude. A non-stationary series is made stationary by differencing.

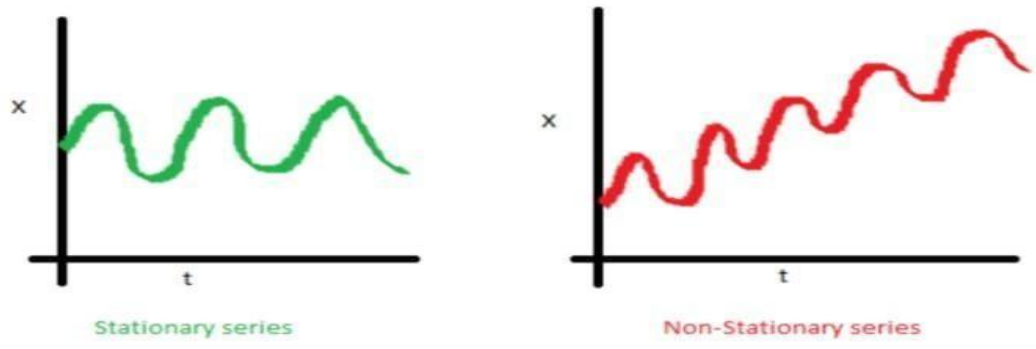

Fig. 3.3 Stationary and Non-stationary series

In an ARIMA model, the future value of a variable is supposed to be a linear combination of past values and past errors. In general, ARIMA model is denoted by
ARMA (p, q). The form of the ARMA (p, q) model is,

$$y_t = c + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \cdots + \Phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q y_{t-q}$$

Where $\in_t$ is an uncorrelated innovation process with mean zero and y_t is the actual value and $\in_t$ is the random error at time t, $\emptyset_1$ and $\emptyset_2$ are the coefficients, p and q are integers that are often referred to as autoregressive and moving average polynomials.

Here, applied for loop using (p, d, q) values between 0 to 2 to find least AIC value.

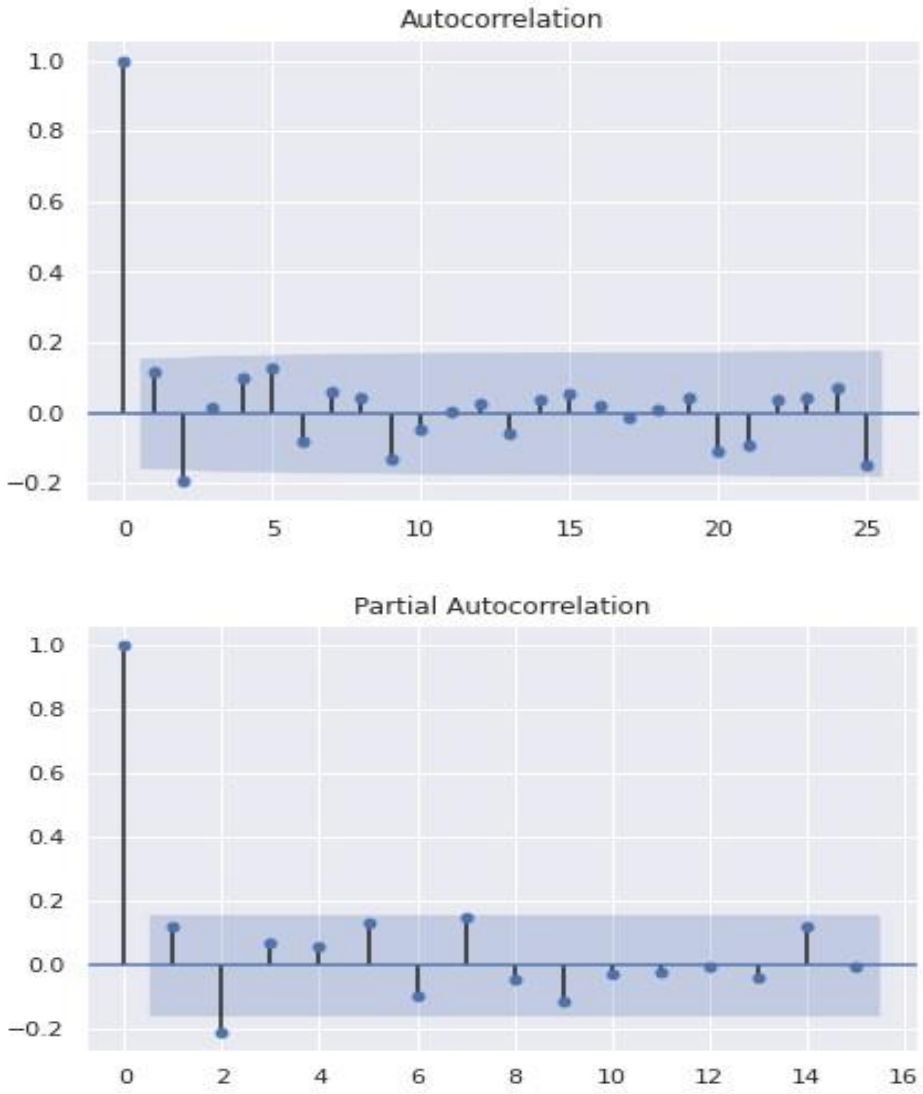

Fig. 3.4 Graphical representation of least AIC values (acf, pacf)

3.3. Software and Packages

Python is a popular computer programming language used to create software and websites, automate processes, and analyse data. Python is a general-purpose language, which means that it may be used to make a wide range of applications and is not tailored for solving any particular issues.

Its adaptability and beginner-friendliness have elevated it to the top of the list of programming languages in use today. It was the second-most popular programming

language among developers in 2021, according to a survey by the market research firm RedMonk.

Python is commonly used for developing websites and software, task automation, data analysis, and data visualization. Since it's relatively easy to learn, Python has been adopted by many non-programmers such as accountants and scientists, for a variety of everyday tasks, like organizing finances.

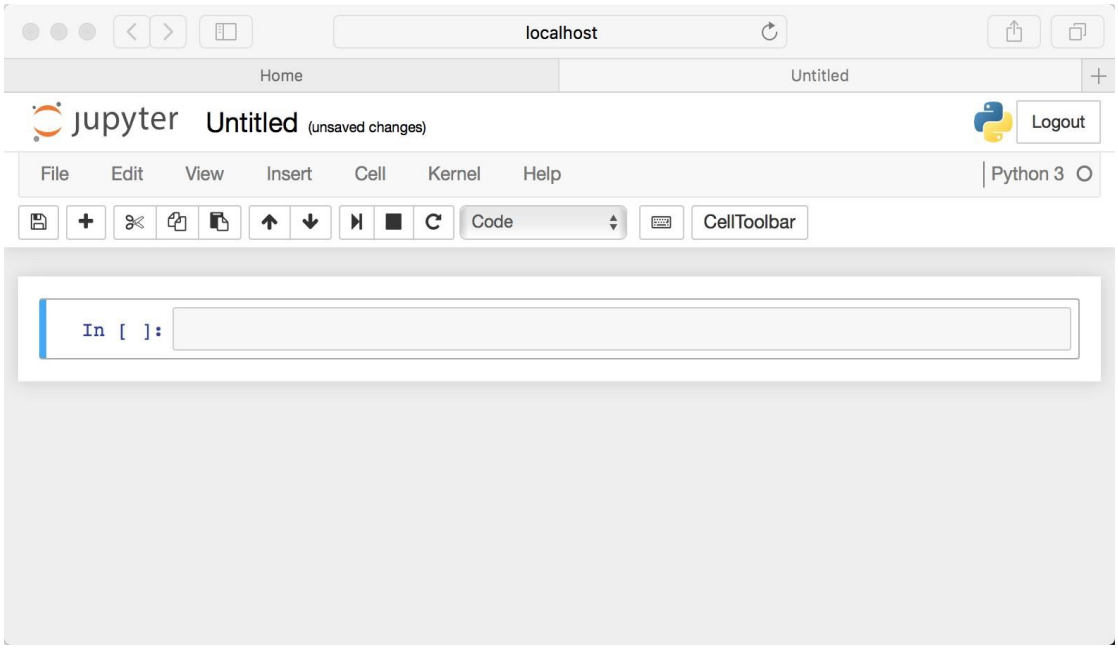

Fig. 3.5 Python console of Anaconda (Jupyter Notebook)

Packages and Libraries:

1. import numpy as np
2. import pandas as pd
3. import matplotlib.pyplot as plt
4. from statsmodels.tsa.stattools import adfuller
5. from statsmodels.tsa.seasonal import seasonal_decompose
6. from statsmodels.tsa.arima.model import ARIMA
7. from sklearn.metrics import mean_squared_error,
8. mean_absolute_error
9. import seaborn as sns

10. from statsmodels.graphics.gofplots import qqplot as qq
11. from statsmodels.tsa.seasonal import seasonal_decompose as sd
12. from statsmodels.tsa.stattools import adfuller
13. from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

Number of Files: 1 CSV file, Netflix.csv

Number of Records: The csv contain data of around 11-12 years, which has 3000+ records.

4. DATA ANALYSIS

4.1. EDA (Exploratory Data Analysis)

Exploratory data analysis (EDA) is a systematic way to explore the data using transformation and visualization. EDA is an iterative cycle without a set of rules, yet there are several fundamental procedures that should be followed to manage data in a methodical manner:

1. Generate questions about data
2. Search for answers by visualizing, transforming, and modelling data
3. Use what is to learn to refine questions and if required then generate new questions.

4.1.1. Sample Data Analysis

After sample data preparation the next task in modelling was sample data analysis and visualization. Fig. 4.1 representing the monthly average share price for Netflix.

Fig. 4.1 Monthly Average Netflix Stock

The stock closing price index was depicted in the graph above for about 11–12 years, from 2005 to 2017. Over the past 12 to 13 years, numerous up and down patterns have been observed. Since 2005, there has been constant growth in the stock price.

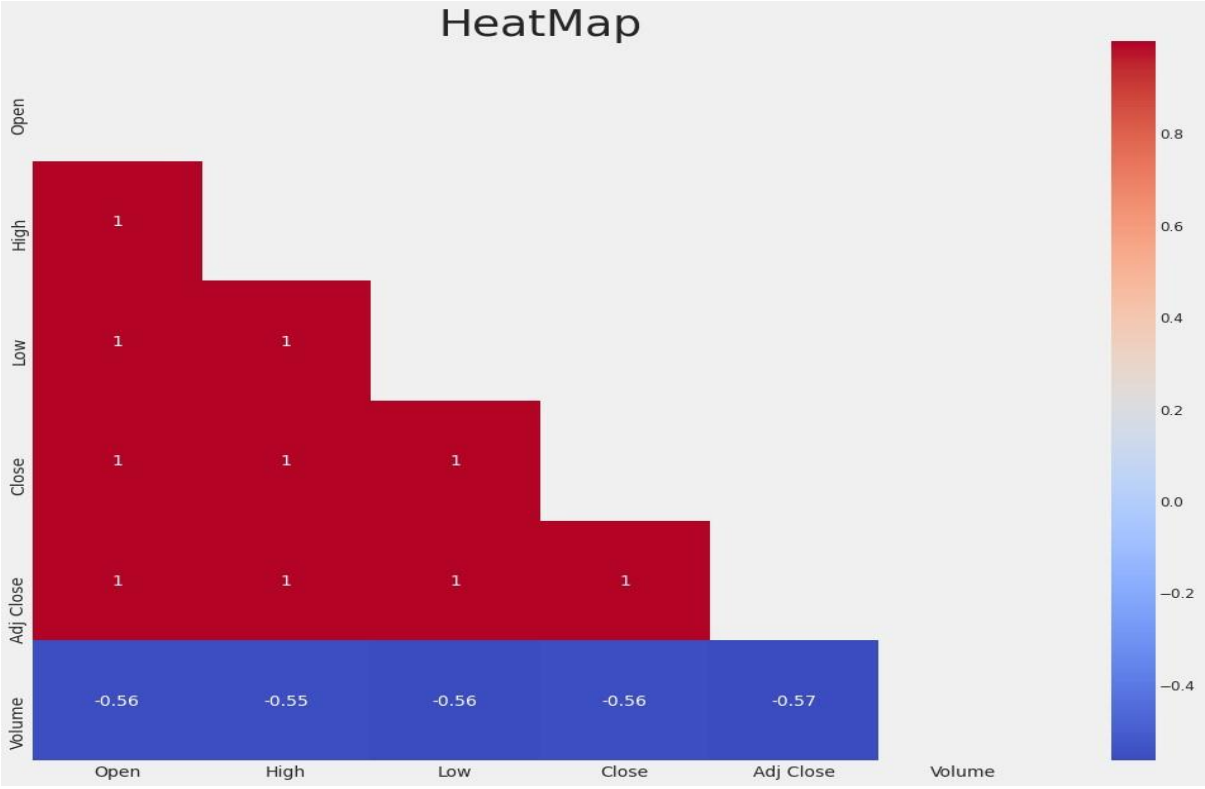

Fig. 4.2 Heatmap to Verify Multicollinearity between Features

Decomposition of Time Series

rcParams['figure.figsize'] = 18, 8 plt.figure(figsize=(20,16)) decomposed_series = sd(monthly_data['Close'],model='additive',freq=12) decomposed_series.plot() plt.show()

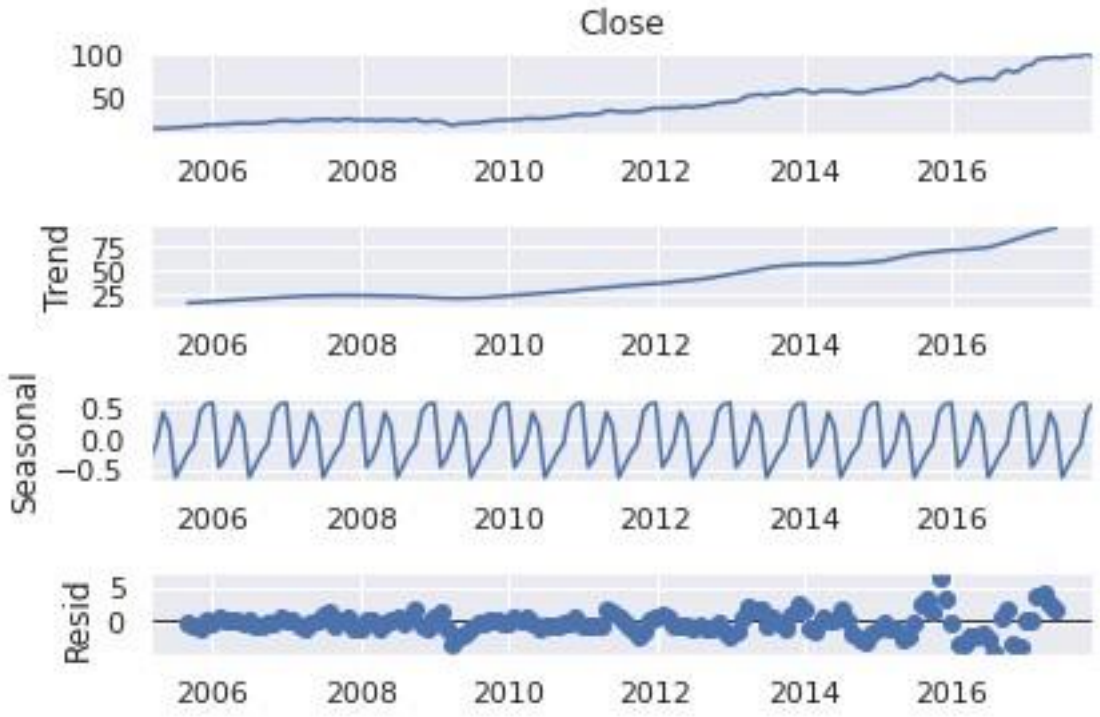

Fig. 4.3 graph showing decomposition of time series

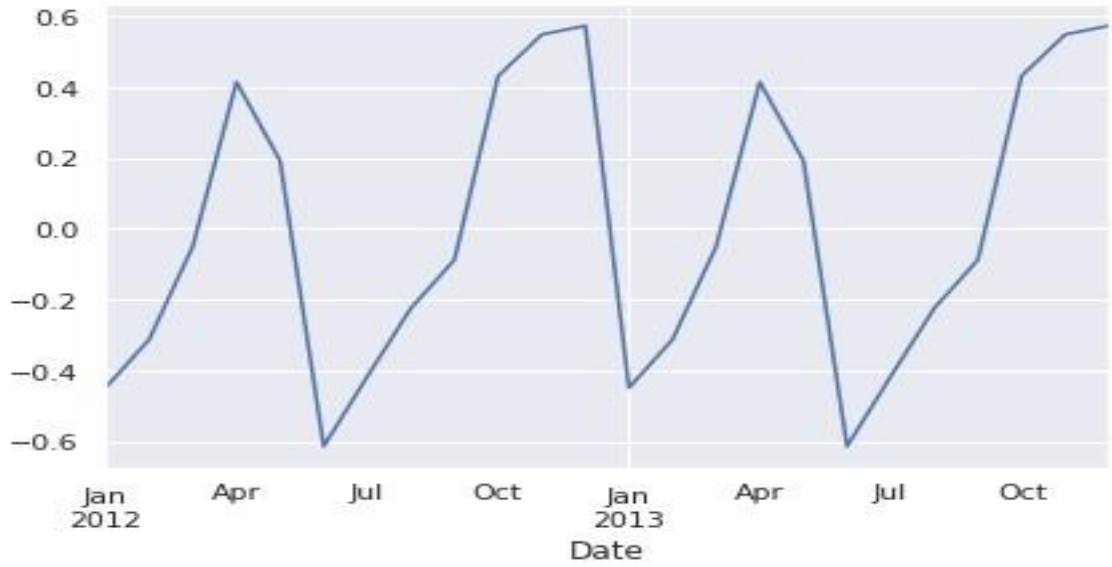

Fig. 4.4 Drilling down and observing seasonality

From Fig. 4.3 and Fig. 4.4 we can infer the results that overall, there is an upward trend in the data. But there appears to be seasonality, Netflix has rallied during holiday season which is around summers (April and May) and winters (December), which is obvious.

Data was non-stationary up until this point, but time series forecasting requires stationary data since it makes future prediction simpler. The mean and variance must remain constant over time for stationary time series.

Fig. 4.5 Non-Stationary Data

Utilizing the differencing method, data can be transformed from non-stationary to stationary time series. Graphical depiction has been used to validate the assumptions. The first differencing value is the difference between the current time and the preceding time period. Following the use of the differencing approach, the time series appeared as the stationary data below.

```
                    from statsmodels.tsa.stattools import adfuller
                    def ad_fuller_func(X):
                    rolmean = X.rolling(12).mean()
          rolstd = X.rolling(12).std()
#Plot rolling statistics:
                    plt.plot(X, color='blue',label='Original')
plt.plot(rolmean, color='red', label='Rolling Mean')
                    plt.plot(rolstd, color='black', label = 'Rolling Std')
          plt.legend(loc='best')                    plt.title('Rolling Mean
and Standard Deviation')
                    plt.show(block=False)
result_ad_fuller = adfuller(X)
                     print('ADF Statistic: %f' % result_ad_fuller[0])

                    print(f'p-value: {result_ad_fuller[1]}')
                    print('Critical Values:')                    for key,
value in result_ad_fuller[4].items(): print('\t%s: %.3f' % (key,
value))
                    output = pd.Series(result_ad_fuller[0:4],index=['Test Statistics','p-value','No. of lags used','No. of
          obvs used'])
                     if result_ad_fuller[0] < result_ad_fuller[4]['5%']:
                              print('Reject Null Hypothesis(Ho)-Time Series is Stationary')
          else:
```

print('Failed to Reject Ho-Time Series is Non-Stationary')
ad_fuller_func(monthly_data['Close'])

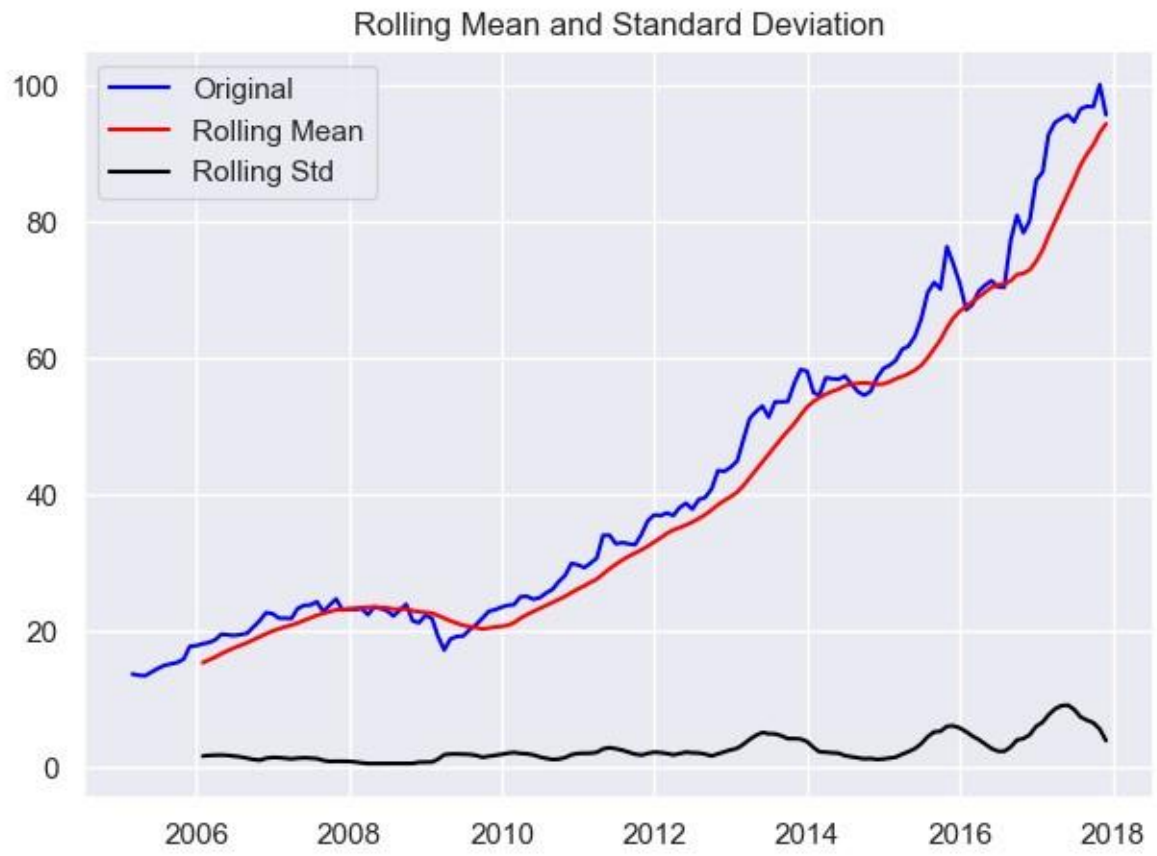

Fig. 4.6 Rolling mean and standard deviation of residuals

>>ADF Statistic: 1.874376 >>p-value:
0.9984856603902791 >>Critical
Values:
1%: -3.474
5%: -2.881
10%: -2.577
>>Failed to Reject Ho-Time Series is non-Stationary

Time Series is Not Stationary as observed earlier also by Decomposition (Trend and Seasonality Present). Statistically verified by ADF Test. Making series stationary

monthly_diff = monthly_data['Close'] -
monthly_data['Close'].shift(1) plt.figure(figsize=(12,8))
plt.plot(monthly_data.index, monthly_diff, label='stationary series')
plt.legend(loc='best') plt.title("Stationary Series") plt.show()

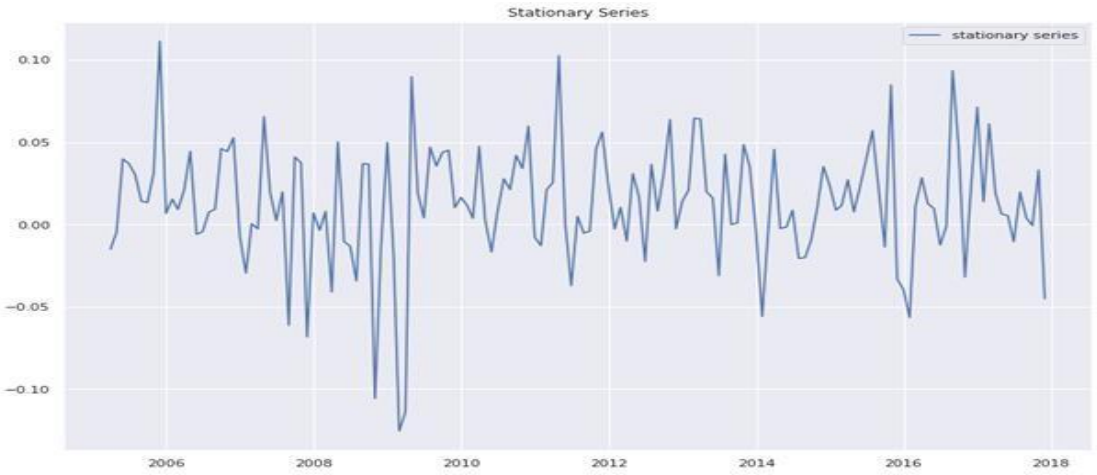

Fig. 4.7 Graphical representation after applying differencing method

ad_fuller_func(monthly_diff.dropna())

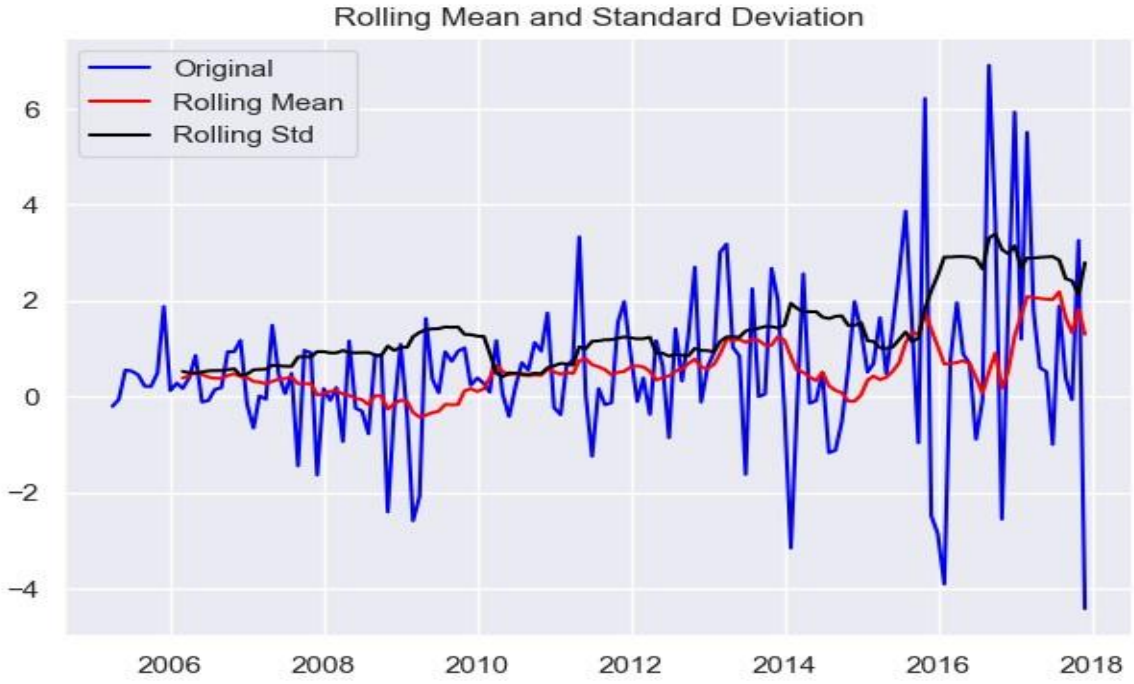

Fig. 4.8 rolling mean and standard deviation after differencing

>>ADF Statistic: -8.965249 >>p-value: 8.021789122750542e-15 >>Critical Values:

1%: -3.474

5%: -2.881

10%: -2.577

>>Reject Null Hypothesis(H_0)-Time Series is Stationary

5. RESULTS AND DISCUSSION

5.1. Result Review

The ARIMA model has been picked for a single time series ought not to change that much and pursue some pattern for order selection. There was 1 method (ARIMA) applied for Netflix stock data. For seasonal time series, optimized ARIMA model parameters to discuss the result. Let us review results.

5.1.1. ARIMA

The ARIMA model had been implemented for Netflix data using train dataset and least AIC (Akaike Information Criteria) value has been calculated to find the accuracy. Least AIC value is highlighted in pink in the below table. Train dataset has been divided in such a way so that the best result can be evaluated. Performing stepwise search to minimize AIC.

p	d	q	AIC
1	1	1	4481.583
0	1	0	4491.366
1	1	0	4486.634
0	1	1	4486.225
0	1	0	4498.110
2	1	1	4483.576
1	1	2	4483.979
0	1	2	4485.508
2	1	0	4485.848
2	1	2	4485.395
1	1	1	4491.385

Table 1: AIC Values

Best model: ARIMA (1,1,1) (0,0,0) [0] intercept

Here, applied for loop in python programming language using (p, d, q) values between 0 to 2 to find the least AIC value. As per for result, least AIC value has been measured which is (1,1,1) (0,0,0) [0] intercept.

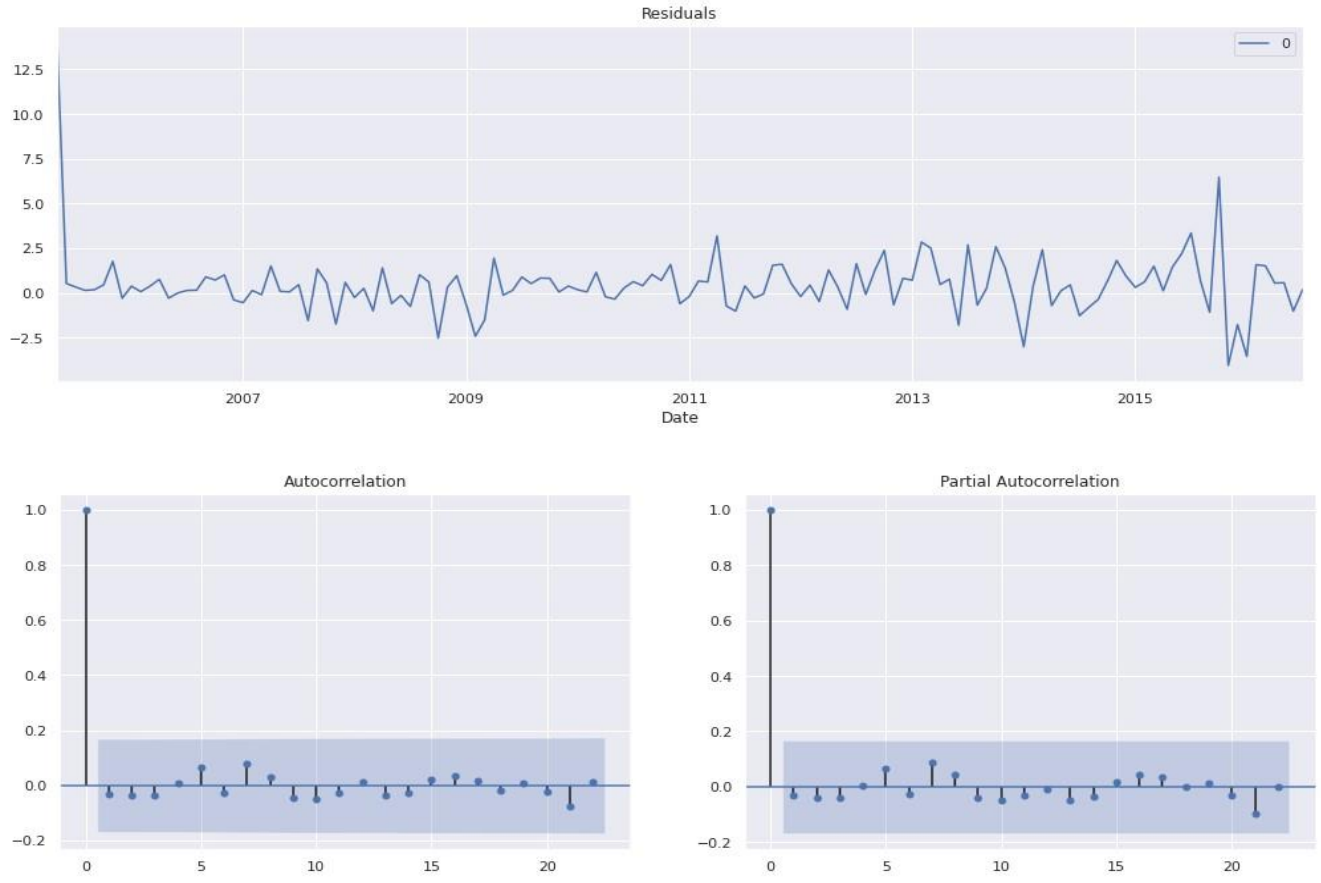

Fig.5.1 (1,1,1) ARIMA Model Residuals

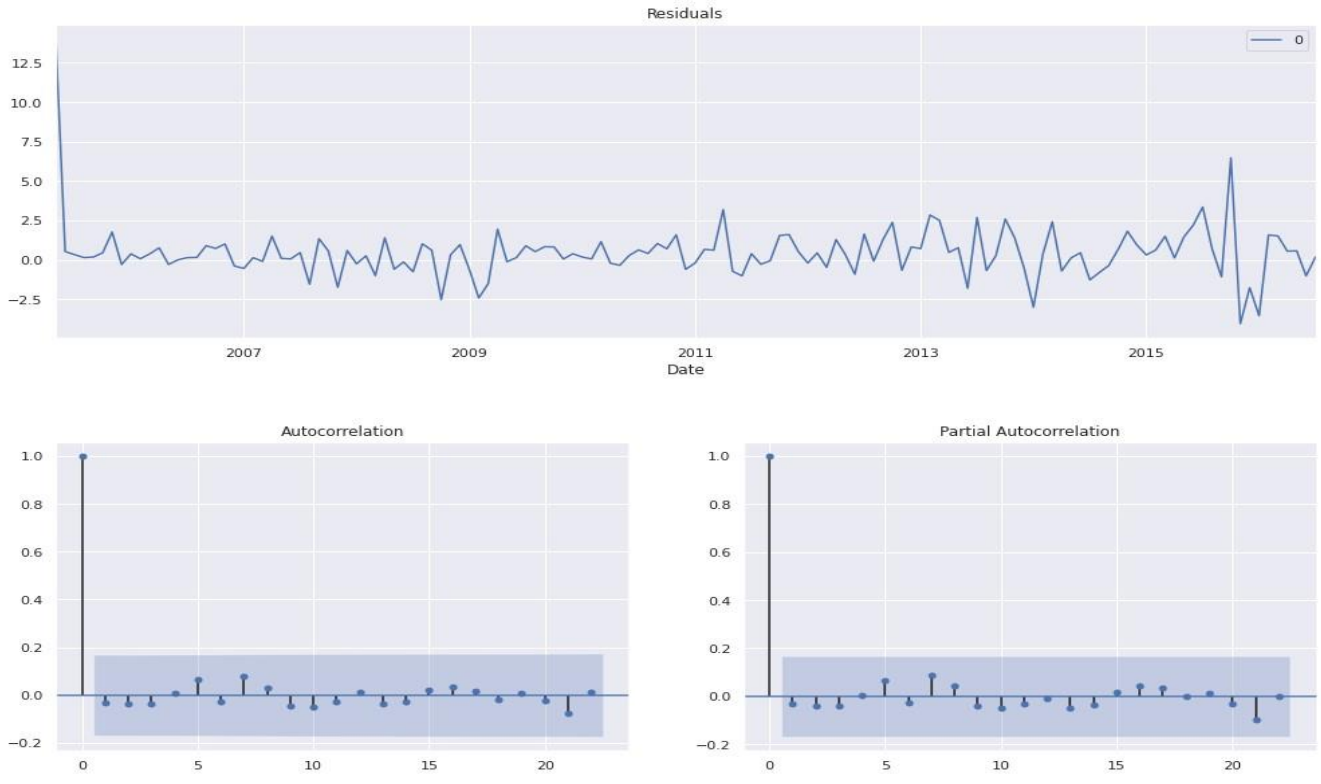

Fig. 5.2 (1,1,0) ARIMA Model Residuals

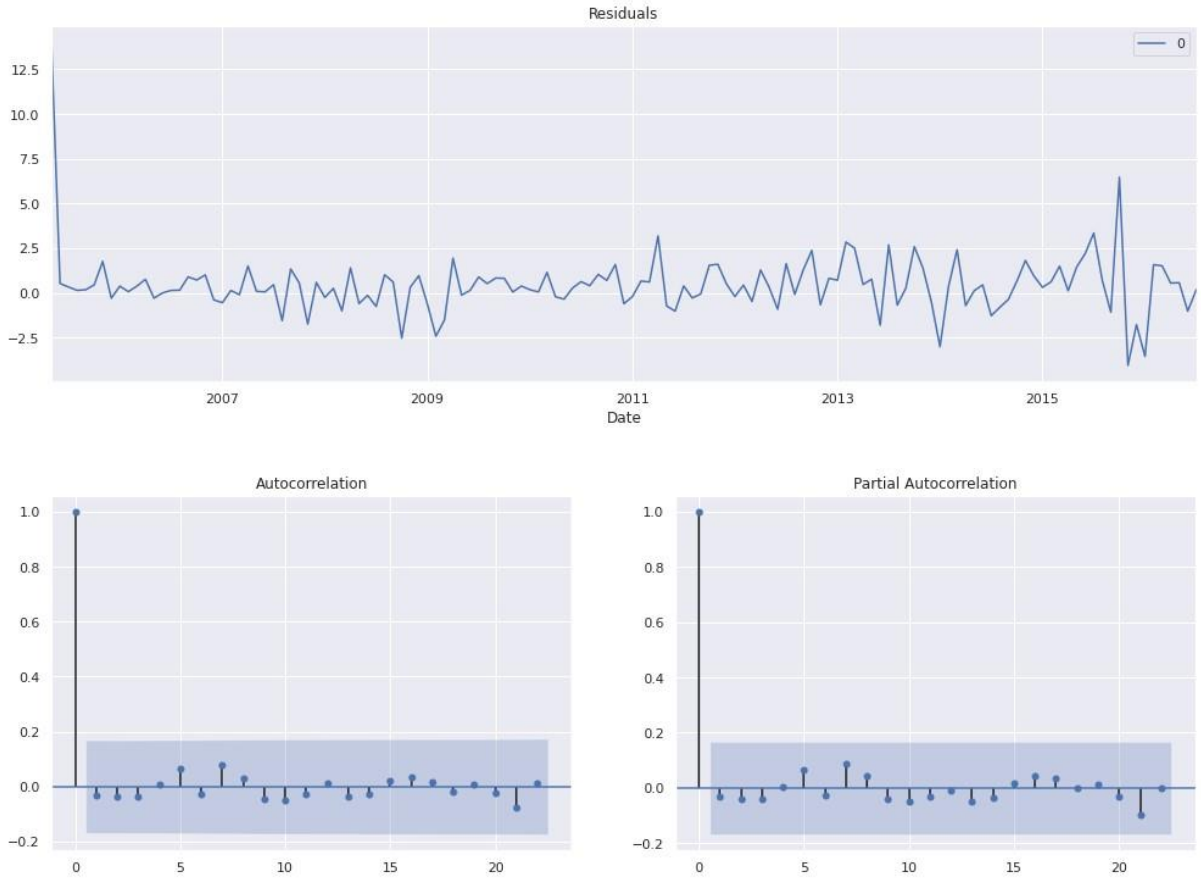

Fig. 5.3 (0,1,1) ARIMA Model Residuals

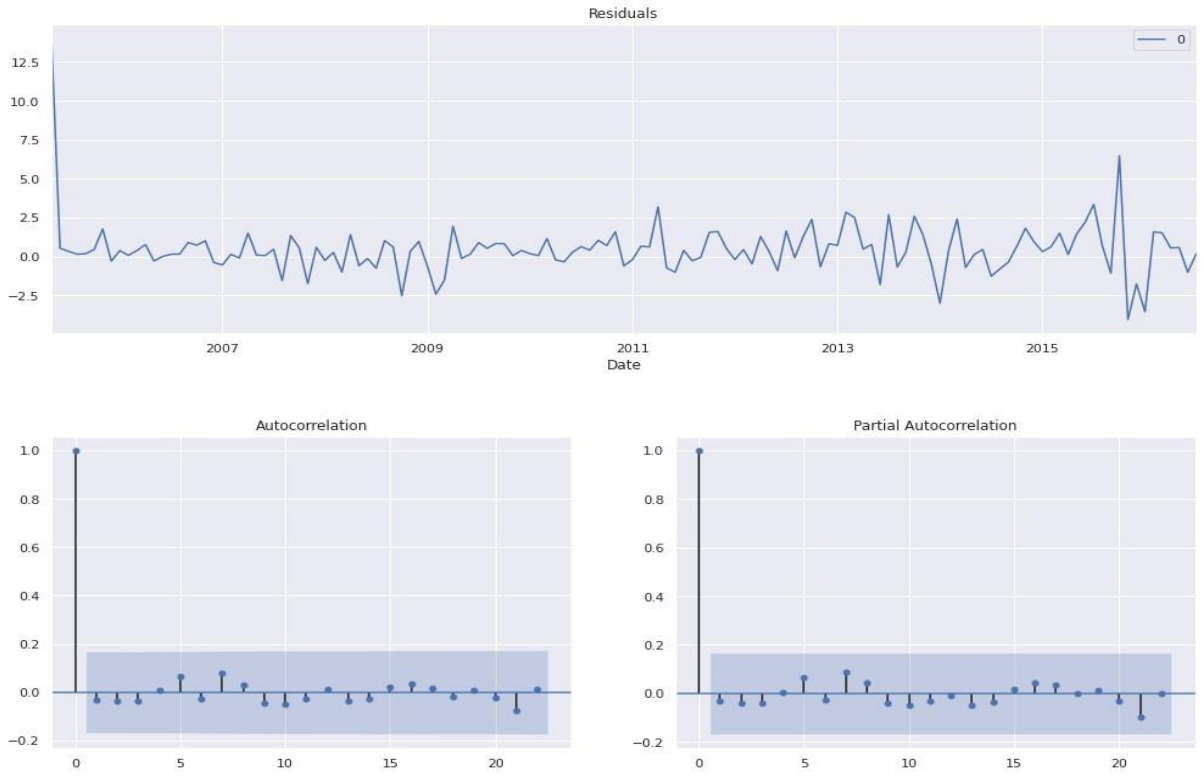

Fig. 5.4 (0,1,0) ARIMA Model Residuals

The graph below shows the plotted values of predicted stock price obtained using the training and the testing data.

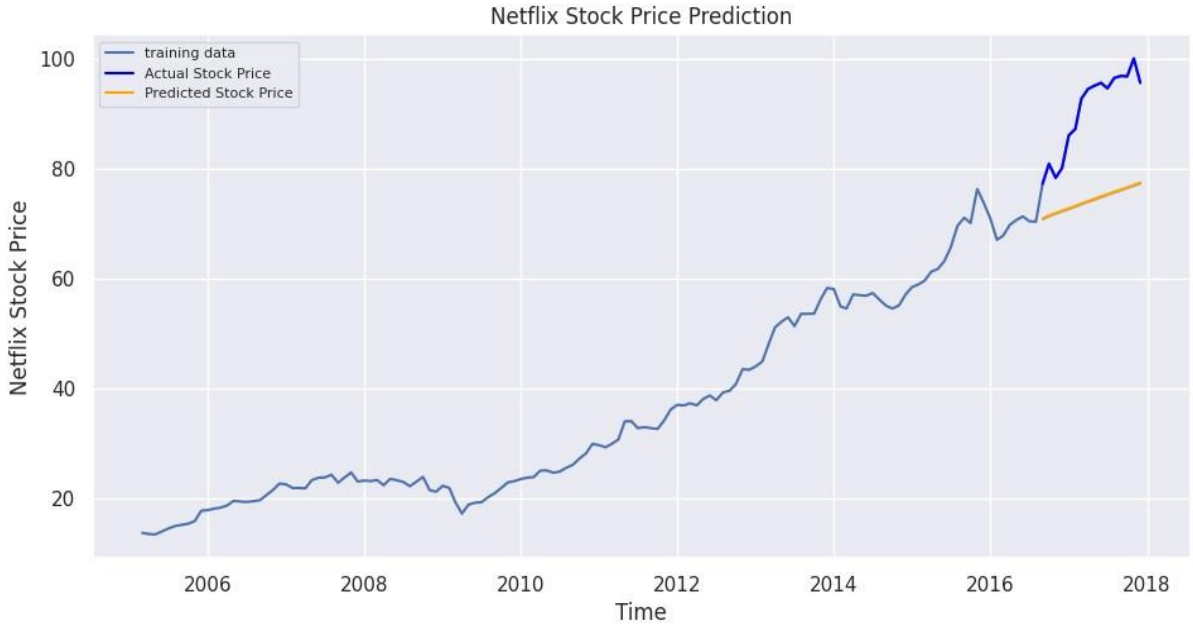

Fig. 5.5 Graph of predicted price

6. CONCLUSION AND FUTURE WORK

This study has presented the extensive procedure of ARIMA model for stock price prediction. The examinations of this model uncovered that stock data set of Netflix. The test results acquired showed the capability to predict stock prices satisfactory. This could direct speculators at stock price to settle on gainful investment decisions.

The ARIMA has a large amount of established time series techniques. The pattern segment of ARIMA was important to the point that it delivered very nearly a straight line. With regards to arranging future outstanding workloads, it has been observed that one could now do fine-tuning of the model. At last, keeping in mind that as of now there has been a decent pattern model which is prominent and it was not really dependably need complex machine learning calculations for determining models.

To anticipate the future value of a stock, we frequently train the model using data from NSE-listed businesses. This demonstrates that the suggested method can be used to distinguish between data relationships. The findings also show that an ARIMA model are able to detect changes in trends. The optimal model for the projected technique is known to be an ARIMA model. It makes predictions using the data available at a certain moment. This frequently happens as a result of the stock market's rapid movements. The stock market's fluctuations don't always follow a consistent cycle or a regular pattern. The duration of trends' existence and how long they last will vary depending on the firms and industries. The examination of these cycles and patterns can increase the returns on investment for investors. To increase the accuracy of forecasted stock prices, we will incorporate more stock market data in next work.

7. PLAGIARISM AND REFERNCE

- Brockwell, P.J., Davis, R.A. and Calder, M.V., 2002. Introduction to time series and forecasting (Vol. 2). New York: springer.
- Investopedia
- Yahoo Finance
- Java Point
- Wikipedia
- towardsdatascience
- datasciencewithmarco
- Montgomery, D.C., Jennings, C.L. and Kulahci, M., 2015. Introduction to time series analysis and forecasting. John Wiley & Sons.
- Devi, B.U., Sundar, D. and Alli, P., 2013. An effective time series analysis for stock trend prediction using ARIMA model for nifty midcap-50. International Journal of Data Mining & Knowledge Management Process, 3(1), p.65.
- Lo, A.W. and Wang, J., 2001. Stock market trading volume
- Fundamentals of Applied Statistics; SC Gupta and VK Kapoor
- Granger, C.W.J. and Newbold, P., 2014. Forecasting economic time series. Academic Press.