Automation of Mutual Funds Analysis and Price Prediction

DCB BANK

Submitted by

Ashwin Monpur

(M18-037 / MBA 2017-19)

In partial fulfillment for the award of the degree of

POST GRADUATE DIPLOMA IN MANAGEMENT

INSTITUTE FOR FINANCIAL MANAGEMENT AND RESEARCH

IFMR

5655, Central Express Way, Sector 24, Sri City, Andhra Pradesh

(2017-19)

Certificate by the Company Guide/Supervisor

[to be on the letterhead of the organization]

CERTIFICATE

This is to certify that this project report **"Automation of Mutual funds Analysis and Price Prediction"** is the bona fide work of **Ashwin Monpur** who carried out the project work under my supervision.

Project Guide comments:

Yuvraj Kamdar

Vice president and Head Analytics

DCB Bank

Date:

Place: Mumbai

Acknowledgment

I would like to express my gratitude to the following people who have assisted me in the successful completion of the internship tenure at DCB Bank Pvt Ltd.

Vice President and Head Analytics **Mr. Yuvraj Kamdar**, had given his constant support throughout my internship.

Project guide and mentor **Mr. Nilabja Ray** for his valuable time invested in helping me understand the processes of the company and guiding me through every step of the project work

All the faculty members who had helped me understand several concepts which came in handy during the internship.

Signature

Ashwin Monpur

EXECUTIVE SUMMARY

The primary objective of the project is to create a chatbot which can respond to all mutual funds-based queries. To this to be created large datasets are required, all FAQ's related to the mutual funds are collected, and this data is used for training the bot. Before training the bot, the data must be classified into intents so that the bot can understand the queries. To identify the question these intents are again classified into entities. Then rasa nlu and rasa core are used as the natural language processing packages (NLP). Spacy is used for training the data.

In the second phase of the project, the free source data from AMFI is stored in the database using API. This data is updated on a daily basis. In the same way, required benchmark index required for the mutual funds is stored in the database.

In this stage, the return and risk of the mutual funds are calculated and updated in the database. The critical aspect of the project is that bot should be able to suggest the better mutual fund for investment. The bot accesses the daily updated information, and on the user, request gives the required output.

The final stage of the project is to model the returns of the mutual funds. The time series data is modeled using machine learning algorithms like ARIMA and SARIMA, and also deep learning algorithms like RNN-LSTM is used to forecast the future NAV. This forecasted NAV's are updated in the database.

Keywords: Time series, Mutual Funds, ARIMA, SARIMA, LSTM, Natural language processing

List of Figures

List of Tables

TABLE OF CONTENTS

About the Organization

DCB Bank is a modern emerging new generation private sector bank with 318 branches across 19 states and three union territories. A scheduled commercial bank regulated by the RBI. DCB Bank has contemporary technology and infrastructure including state of the art internet banking for personal as well as business banking customers.

DCB Bank's business segments are Retail, micro-SME, SME, mid-Corporate, Agriculture, Commodities, Government, Public Sector, Indian Banks, Co-operative Banks and Non-Banking Finance Companies (NBFC). DCB Bank has approximately 450,000 customers.

DCB Bank has deep roots in India since its inception in the 1930s. Its promoter and promoter group the Aga Khan Fund for Economic Development (AKFED) & Platinum Jubilee Investments Ltd. holds over 15.01%.

INTRODUCTION

The project divided into two stages. Firstly, Mutual Funds Data (2006-2018 (present date)) updated daily from AMFI API and this data stored in the Database (HIVE, Local Database) then required data be plotted to observe the behavior of the Time series data. Then data rolling average and the standard deviation is calculated for 22 working days (step size=22) of a month. In this stage, Return and Risk of the mutual funds calculated monthly, Quarterly, Semi-Annual and Annually. Now to compare the mutual funds, past two years data is used to calculate Risk ratios and Return.

Risk Ratios: Depending on the benchmark index

- Active Risk
- Volatility
- Beta
- Jensen Alpha
- Sharpe Ratio
- Treynor Ratio
- Information Ratio
- Sortino Ratio

The second stage of the project is to predict the future prices of the Mutual Funds. Here the Machine learning and Deep learning models are used to predict the future prices of the funds. This modeling of data is used to forecast three months prices.

The whole process is automated using python, all the calculated numbers of return, risk, and predicted future prices are updated daily in the database.

Mutual Funds

A mutual fund is an investment firm that pools money from shareholders and invests in a variety of instruments, such as stocks, bonds, and money market instruments. Open-end Mutual funds stand ready to buy back its shares at their current net asset value, which depends on the total market value of the fund's investment portfolio at the time of redemption. Most open-end Mutual funds continuously offer new shares to investors. Mutual funds invest pooled cash of many investors to meet the fund's stated investment objective. Mutual funds stand ready to sell and redeem their shares at any time at the fund's current net.

Asset value: Total fund assets divided by shares outstanding.

In Simple Words, the Mutual fund is a process of pooling the investor's funds by issuing units to the investors and investing those funds in securities.

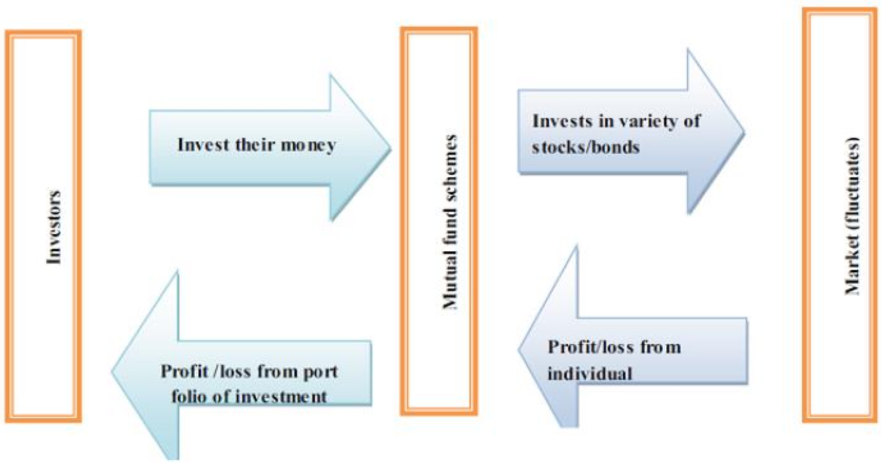


Figure 1: Mutual Funds Process

Investments in securities spread across broad sectors, and thus the risk minimizes. Diversification reduces the risk; all stocks may not move in the same direction in the same proportion at the same time. The investors share the profits or loss in proportion to their investments. The Mutual funds typically come out with some schemes with different investment objectives which are launched from time to time.

The investment manager would invest the money collected from the investor into assets that are permitted by the stated objective of the scheme. For example, an equity fund would invest equity, and equity related instruments and a debt fund would invest in bonds, debentures, gilts.

Concept Used

The funds pooled is invested in capital market instruments such as shares, debentures and other securities. Its unitholders share the income earned through these investments and the capital appreciations realized in proportion to the number of units owned by the investors.

Types of Mutual Funds

By structure

- Open – Ended Schemes.
- Close – Ended Schemes.
- Interval Schemes.

By investment objective

- Growth Schemes.
- Income Schemes.
- Balanced Schemes.

Other schemes

- Tax Saving Schemes.
- Special Schemes.
- Index Schemes.
- Sector Specific Schemes.

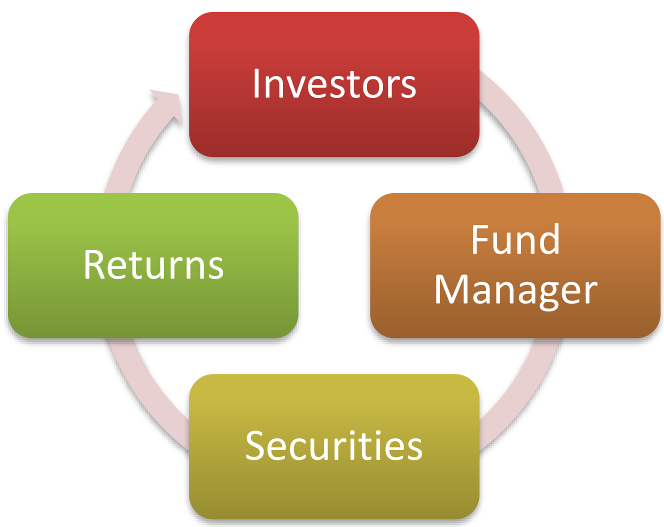


Figure 2: Mutual Funds Concept

Open-ended schemes

The units offered by these schemes are available for sale and repurchase on any business day at NAV based prices. Hence, the unit capital of the schemes keeps changing each day. Such schemes thus offer very high liquidity to investors and are becoming increasingly popular in India. Please note that an open-ended fund is NOT obliged to keep selling/issuing new units ways and may stop issuing further subscription to new investors. On the other hand, an open-ended fund rarely denies to its investor the facility to redeem existing units.

Closed-ended schemes

The unit capital of a close-ended product fixed as it makes a one-time sale of a fixed number of units. These schemes launched with an initial public offer (IPO) with a stated maturity period after which the units fully redeemed at NAV linked prices. Unlike open-ended schemes, the unit capital in closed-ended schemes usually remains unchanged. After an initial closed period, the scheme may offer direct repurchase facility to the investors. Closed-ended schemes are usually more illiquid as compared to open-ended schemes and hence trade at a discount to the NAV. This discount tends towards the NAV closer to the maturity date of the scheme.

Interval schemes

These schemes combine the features of open-ended and close-ended schemes. They may be traded on the stock exchange or may be open for sale or redemption during pre-determined intervals at NAV based prices.

Growth schemes

These schemes, also commonly called Equity Schemes, seek to invest a majority of their funds in equities and a small portion of money market instruments. Such schemes have the potential to deliver superior returns over the long term. However, because they invest in equities, these schemes are exposed to fluctuations in value especially in the short term.

Income Schemes

These schemes, also commonly called Debt Schemes, invest in debt securities such as corporate bonds, debentures, and government securities. The prices of these schemes tend to be more stable compared with equity schemes, and most of the returns to the investors generated through dividends or steady capital appreciation. These schemes are ideal for conservative investors or those not in a position to take higher equity risks, such as retired individuals. Compared to the money market schemes they did have a higher price fluctuation risk and compared to a Gilt fund they have higher credit risk.

RATIONALE OF THE PROJECT

The project helps to understand different types of Mutual Funds and their characteristics, and also to analyze their risk ratios compared with benchmark index and then rating the Mutual Funds for investment.

This project helps to understand handling the time series data in the platforms and its packages to do various calculations.

Understanding different types of Time series related machine learning and Deep learning algorithms and their implementation. Using those models to forecast future prices of the Mutual Funds.

In this project, automation is implemented to update the return, risk ratios, and forecasted prices in the database on a daily basis.

Objective

1. Understanding how mutual funds work
2. Understanding its types
3. Handling and understanding the Time series data
4. Calculating the risk ratios
5. Comparing Mutual funds ratios with benchmark ratios
6. Rating Mutual funds
7. Understanding different modeling techniques
8. Understanding how machine communicates and understands unstructured human language and communicates in the human language.

Need for Research

Mutual Funds is booming sector and is the second largest sector in the investment industry. It has much scope for generating income and providing returns to the investor. The need of research work is to evaluate the performance of different mutual funds available and automating the process to rank mutual funds according to their returns and other risk parameters.

Methodology

In the second phase of the project, machine learning and deep learning models used for the forecasting. To validate the model the data divided into two parts, one for training the model, i.e., the training data and other for validating the model. Training data is about 70% of the overall data, and test data is about 30% of the data. If the model fits best and validates the test data, then this model is used for forecasting the future NAV of the fund.

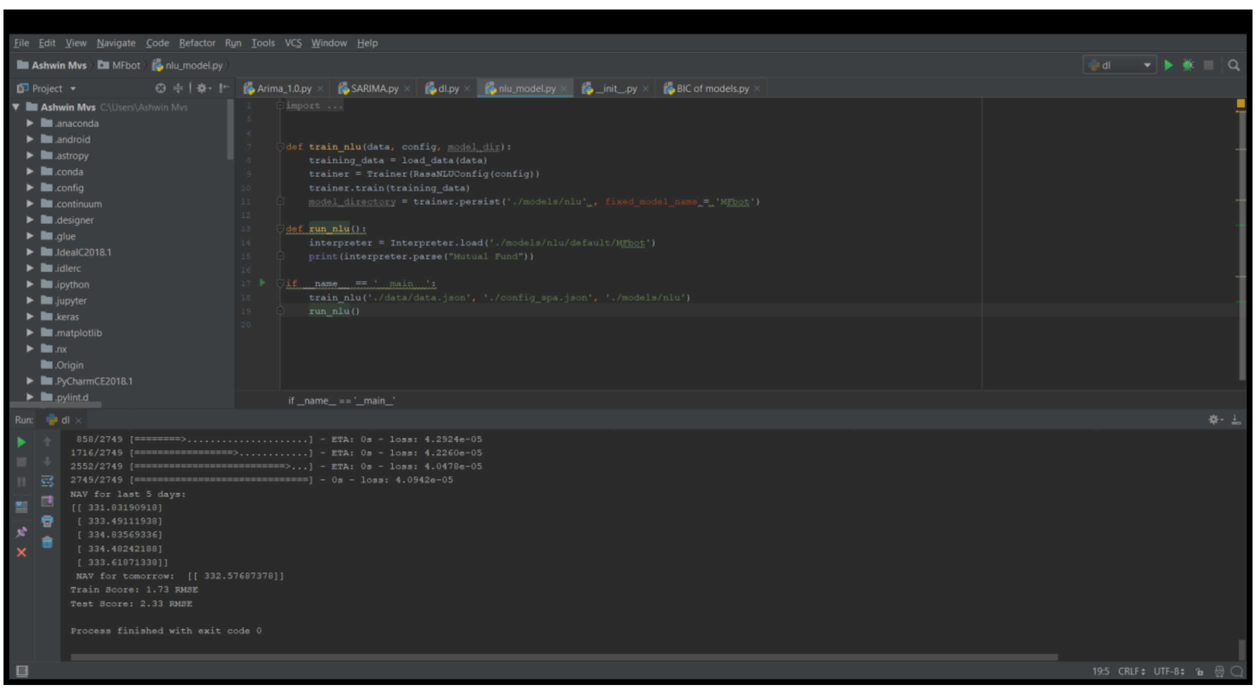


Figure 3: Pycharm IDE

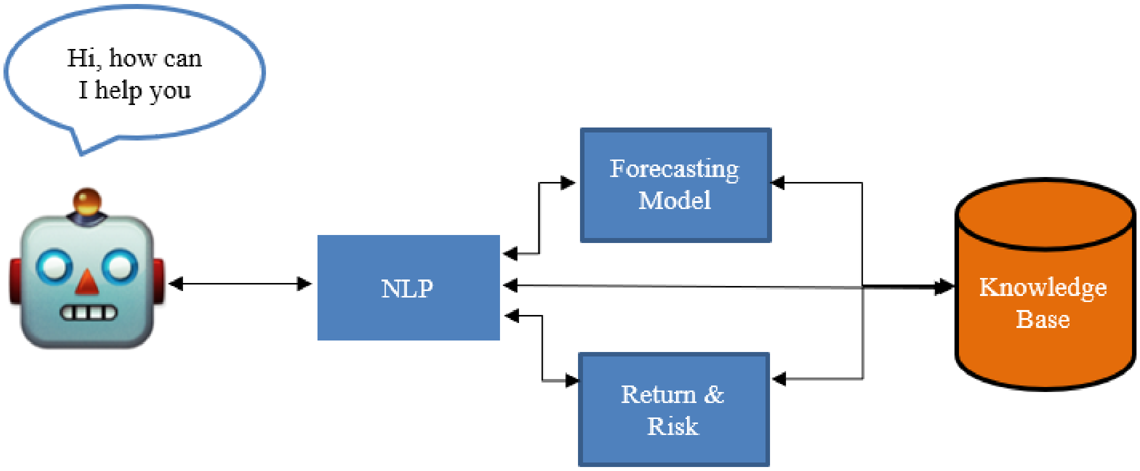


Figure 4: Chatbot

The whole operations and modeling are performed on the python platform. Python 3.5 and various packages like Pandas, NumPy, sklearn were used for calculating return, risk and for forecasting models. For designing chatbot, python natural language processing packages used. Pycharm is used for the development of the project.

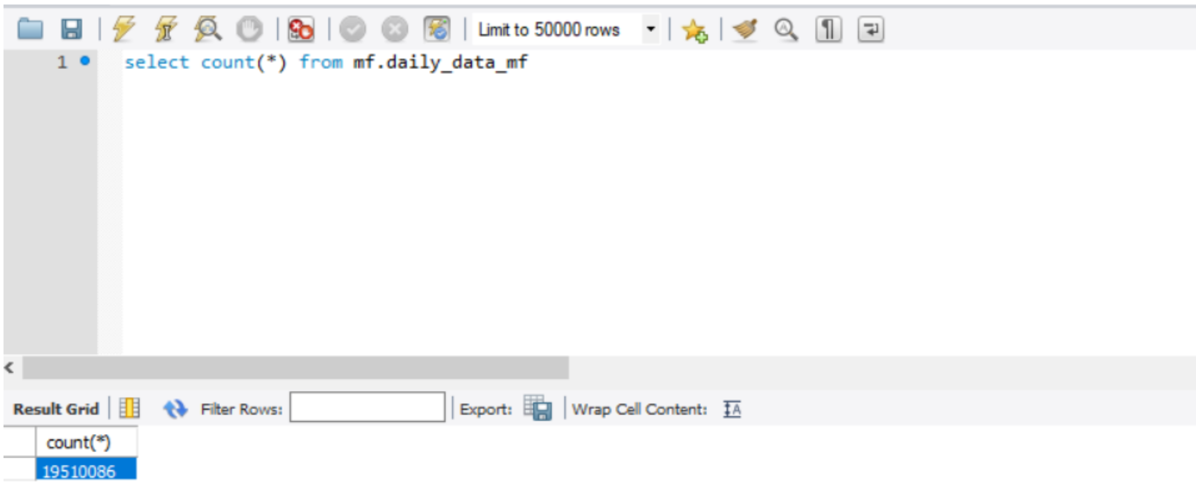
The Mutual fund's historical data is used to train the model like ARIMA, SARIMAX and LSTM models to predict the future NAV values. The models are selected on the best fit on the historical data, and this model is used to forecast the future NAV values.

The final stage of the project is to design the Mutual funds chatbot which helps customers to know all about the mutual funds. The whole module of the chatbot can help in answering the general question about the mutual funds and gives return, risk and forecasted prices of the mutual fund.

Data & Analysis

Firstly, the mutual fund's data and index data is collected from the AMFI website and stored in the database. For performance and comparison of mutual funds, the benchmark index of the data is required. Index data is collected from the NSE website.

As on 5/21/18 the total data points count in the database is about 19510086. The data is filtered according to the scheme code of the mutual fund. The ratios and the return of the fund is calculated and stored in the database. This process is repeated on a daily basis. The IDE shows the number of data points in the database



Mutual Fund: Franklin Blue-chip Fund

The data used for analysis is from 4-1-2006 to 15-5-2018. This data is retrieved from the database with a unique code called scheme code. Using the Franklin Blue-chip Net Asset Value, the below figure 5 is plotted using python to observe the behavior of the time series data. From the figure 5, observation is that the time series data is following a linear trend.

Note: This Fund is taken as the reference to calculate return and risk. Later the same process is automated for the rest of the funds in the database.

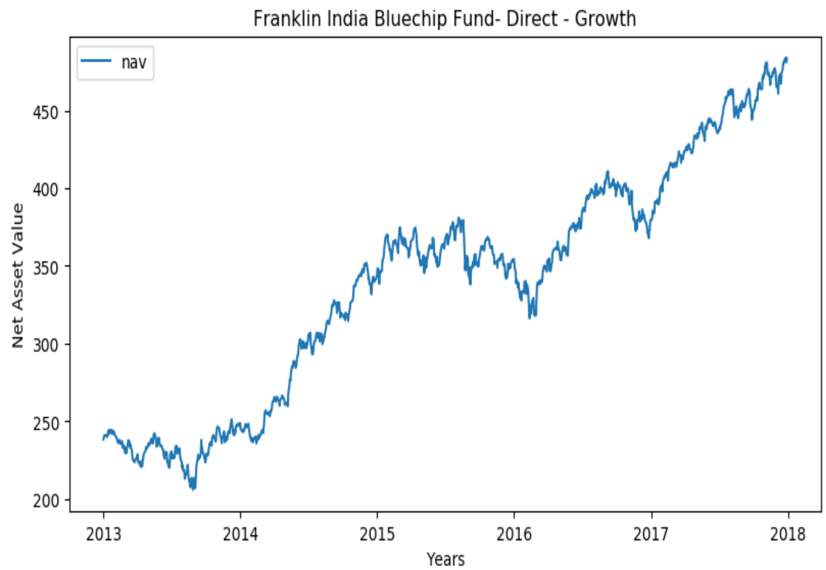


Figure 5: MF- Franklin Blue chip

Calculating Rolling Mean & Standard Deviation

Mean of the data is not constant it varies according to the time. So, the stochastic mean and standard deviation, i.e., the rolling mean and standard deviation calculated with a step size (window) of 22.

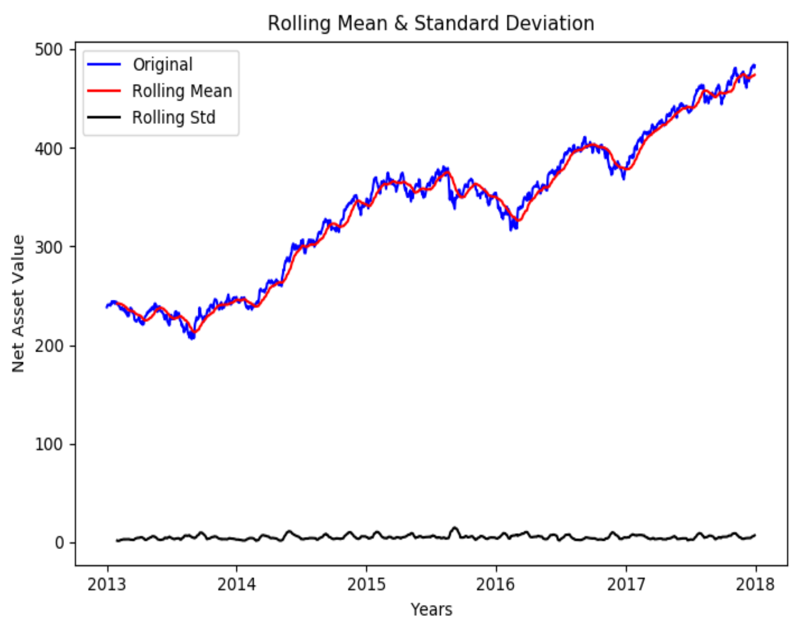


Figure 6: Rolling Mean and SD- Franklin Bluechip

Return & Risk Calculations

Return

The return of the fund calculated on day to day basis. Here percentage change formula is used to calculate the percentage change in the return.

$$\boldsymbol{Return} = \frac{Nav_{n+1} - Nav_n}{Nav_n}$$

Risk

The risk is calculated, using past two years data of mutual funds and benchmark.

- Past two years return is calculated taking present NAV of the fund and exactly two years back fund value.
- Return is calculates using percentage change formula

$$Return(\%) = \frac{Nav_{n+1} - Nav_n}{Nav_n} \times 100$$

$$Nav - Net\ Asset\ Value$$

- Two years return converted into the one-year effective return.

$$\boldsymbol{Effective\ Return}(\%) = (((1 + Return)^{\frac{1}{2}}) - 1) \times 100$$

Volatility

Standard deviation (SD) measures the volatility of the fund returns its average returns. It explains how a fund return can deviate from the historical mean return of scheme.

$$\boldsymbol{Variance} = \frac{Sum\ of\ squared\ Differences:\ daily\ return\ and\ mean\ of\ fund}{Number\ of\ daily\ return\ data\ points}$$

Note: If data is less than three years

$$Variance = \frac{Sum\ of\ squared\ Differences: daily\ return\ and\ mean\ of\ fund}{Number\ of\ daily\ return\ data\ points - 1}$$

$$\boldsymbol{Standard\ Deviation\ (SD)} = \sqrt[2]{Variance}$$

Note: If data is more than three years

***Study*: **Higher the Standard deviation more volatile is the funds return.

R Square

It measures the relationship between a portfolio and its benchmark. The value will be in percentages between 1% to 100%

$$Correlation\ Coefficient = \frac{Cov(Mutual\ Fund, Benchmark)}{SD_{MF} \times SD_B}$$

$$SD_{MF}: Standard\ Deviation\ of\ Mutual\ Funds\ Return$$
$$SD_B: Standard\ Deviation\ of\ Benchmark\ Return$$
$$\boldsymbol{R\ Square} = (Correlation\ Coefficient)^2$$

Study: Higher R Square value indicates the fund's performance pattern has been in line with index and low indicates that fund does not match with index

Beta

Beta measures the fund's volatility compared to that of the benchmark index. It explains, how much a fund’s performance would move upside down compared to its benchmark.

$$\boldsymbol{Beta} = \frac{SD_{MF}}{SD_B} \times RSquare$$
$$SD_{MF}: Standard\ Deviation\ of\ Mutual\ Funds\ Return$$
$$SD_B: Standard\ Deviation\ of\ Benchmark\ Return$$

Study: A fund with beta more than one would be more volatile than the market. If beta lower than one, then the fund is less volatile than the benchmark.

Jensen Alpha

Alpha is performance ratio to measure the risk-adjusted performance of a portfolio. It measures the difference between an actual fund return and its expected performance, given its level of risk.

$$\boldsymbol{Alpha} = \{(Fund\ Return - R_f) - Beta \times (Benchmark\ Return - R_f)\}$$
$$R_f: Risk\ free\ return$$

Note: Risk-free return is the return that would be obtained if invested in a government bond for the same duration as mutual funds.

Study: Alpha represented by percentage indicates underperformance or outperformance of a portfolio

- Positive Alpha represents outperformed its benchmark index vice versa.
- Fund return and risk both contribute to its alpha, two funds with same return could have different alphas

Active Risk

Active risk is the difference between a portfolio's returns and the benchmark.

Two ways to measure Active risk

- Subtract benchmark cumulative return from the portfolio return.

$$\boldsymbol{Active\ Risk} = Fund\ Return - Index\ Return$$

- Calculating the standard deviation of the difference in the fund and benchmark returns over time.

$$\boldsymbol{Active\ Risk\ (TE)} = \sqrt{\sum_{i=1}^{n} \frac{(R_{MFi} - R_{Bi})^2}{N-1}}$$

$TE: Tracking\ Error$

$R_{MFi}: Fund\ Returns$

$R_{Bi}: Benchmark\ Returns$

$N: Number\ of\ Return\ Periods$

Study: Lower Active risk means Fund closely follows its benchmark. High active risk means the opposite. Active Risk gives a sense of how a Mutual fund is around the benchmark or how volatile the portfolio is relative to its benchmark.

Sharpe Ratio

Sharpe ratio helps to know if a mutual fund delivers the return concerning the risk is taken by it comparing a fund with a risk-free rate of return

$$\boldsymbol{Sharpe\ Ratio} = \frac{(R_{MF} - R_f)}{SD_{MF}}$$

$$R_{MF}: Fund\ Returns$$

$$R_f: Risk\ free\ return$$

$$SD_{MF}: Standard\ Deviation\ of\ Mutual\ Funds\ Return$$

Study: Higher the value better the fund. If Sharpe ratio is negative, then Fund underperforms when compared to the risk-free rate of return.

Treynor ratio

Treynor ratio is also known as a reward to volatility ratio; it is the excess return generated by a fund over the risk-free rate. Like Sharpe ratio, but uses beta as a measure of volatility.

$$\boldsymbol{Treynor\ Ratio} = \frac{R_{MF} - R_f}{Beta}$$

$$R_{MF}: Fund\ Returns$$
$$R_f: Risk\ free\ return$$

Study: Higher the ratio implies that the fund has better risk-adjusted return than that of another fund with a lower Treynor ratio.

Information Ratio

A measure of risk-adjusted return of Mutual Funds. It is also known as Appraisal Ratio. It is the advanced version of Sharpe ratio. The information ratio is the active return to the relevant benchmark divided by standard deviation of the active return or tracking error.

$$\boldsymbol{Information\ Ratio} = \frac{Annualized\ R_{MF} - Annualized\ R_B}{(SD_{MF} - SD_B) \times \sqrt{252}}$$

Study: The information ratio is used to gauge the skill of managers of mutual funds. It measures the active return of managers portfolio divided by the amount of risk that the manager takes relative to the benchmark. Higher the Information ratio, better the performance of the fund manager.

Sortino Ratio

Sortino ratio measures the performance of the investment relative to the downward deviation.

- It uses the downward deviation for denominator instead of standard deviation.

- Sortino ratio gives a more realistic picture of downward risk ingrained in the fund

$$\boldsymbol{Sortino\ Ratio} = \frac{Expected\ R_{MF} - R_f}{SD_{MF-d}}$$

$$Expected\ R_{MF}: Expected\ Fund\ Returns$$

$$R_f: Risk\ free\ return$$

$$SD_{MF-d}: Standard\ Deviation\ of\ Negative\ Asset\ Returns$$

Study: Large Sortino ratio indicates a low probability of a large loss

Comparison for (ICICI Prudential Top 100 Fund – Growth)

Table 1: Comparison of Risk Ratios

S.no	Statistic	Calculated	Fundoo	Deviation
1	Return	19.281	18.52	0.761
2	Jensen Alpha	5.2	6.32	-1.12
3	Active Risk	5.51	5.2	0.31
4	Volatility	11.29	10.56	0.73
5	Beta	0.9911	0.9	0.0911
6	Sharpe Ratio	0.79	0.92	-0.13

7	Treynor Ratio	15.73	12.2	3.53
8	Information Ratio	0.3	0.55	-0.25
9	Sortino Ratio	1.63	1.46	0.17

Modeling

Mutual fund movement is unpredictable and random, so the fund which is equity and commodity-linked is considered as a random variable. So, the concept of stock modeling is used here to predict the future NAV values of mutual funds.

$$Random\ variable: X = w_1.x_1 + w_2.x_2 + w_3.x_3 \dots \dots \dots \dots \dots . w_n.x_n$$

Bernoulli distribution

A random experiment is called the Bernoulli experiment if the experiment has only two possible outcomes (1 or 0). So, one is indicated as success and zero as the failure. Here this distribution describes the movement of the fund NAV either it should go up or should go down.

$$P(\{1\}) = p$$

$$P(\{0\}) = 1 - p$$

$$where\ 0 \leq p \leq 1$$

So, the expected value and variance are

$$\boldsymbol{Expected\ value}: E(x) = p$$

$$\boldsymbol{Variance}: Var(x) = p(1 - p)$$

Binomial Distribution

Then the same experiment is conducted many times, so the random variable exhibits a binomial distribution.

$$P(\{X = k\}) = \left(C_k^n . p^k . (1 - p)^{(n-k)}\right), 0 \leq k \leq n$$

When Bernoulli experiment performed n times the variance and expected values are the statistics of the binomial distribution given below.

$$\boldsymbol{Expected\ value}: E(X) = np \text{ and } \boldsymbol{Variance}: Var(X) = np(1-p)$$

According to the Bernoulli, there are two possibilities. Either the fund NAV will go up or goes down, and this continues and follows a binomial distribution. So, the fund values(NAV) varies according to the time(T). Because the invested amount varies exponentially according to the time. Where T is the value of NAV at a definite time.

For n-period binomial model:

$$P_0(t) = e^{rt}$$

Here r is considered as the amount invested at the risk-free rate

Black-Scholes Model

It is a model of price variation over time of financial instruments.

The temporary development of the stock price modeled as $P_1(t)$

$$P_1(t) = P_1 . e^{\left(b-\frac{1}{2}\sigma^2\right)t+\sigma W(t)}, t \in [0, T]$$

Where W(t) is the random variable.

As the index included funds are time variable, that explained by random walk term W(t) in the above equation. As a result, the development of a random experiment described over a space of time is called a stochastic process. The random variable W(t) is Brownian movement.

Time Series

A time series is a sequential set of data points, generally measured over successive times. The measurements taken during an event in a time series are arranged in proper chronological order. Time series is affected by four primary elements, which can be separated from the observed data.

Components: Trend, Cyclical, Seasonal and Irregular Components

The general tendency of the TS is to increase, decrease over a long period is termed as simply trend. Thus, it can be said as the trend is a long-term movement in a time series.
Seasonal variations in a Time series are fluctuations within a year during the season. Factors causing the seasonal variations are climate, weather conditions, customs and traditional habits.
The cyclic variations in a time series describe the medium-term changes in the series, caused by circumstances, which repeat the cycles. The duration of cycles extends over the long period, usually two or more years. Most of the economic and financial time series show cyclical variation.

Irregular or random variations in a time series are caused by unpredictable influences, which are regular and do not repeat in a pattern. There is no defined statistical technique for measuring random fluctuations in a time series.

Multiplicative Model

$$Y(t) = T(t) \times S(t) \times C(t) \times I(t)$$

Additive Model

$$Y(t) = T(t) + S(t) + C(t) + I(t)$$

Here Y(t) is the observation and T(t), S(t), C(t) and I(t) are a trend, seasonality, cyclical and irregular variations at time t.

The multiplicative model assumes that the four components of a time series are not necessarily independent, and they can affect one another; whereas in the additive model it is assumed that the four components are independent of each other.

Time series Analysis

In practice, a suitable model is fitted to a given time series, and the corresponding parameters are estimated using the known data values. The procedure of fitting a time series to a proper model is termed as Time Series Analysis. It comprises methods that attempt to explain the pattern of the series data and often used for future forecasting and simulation.

In time series forecasting, past observations are collected and analyzed to develop a suitable mathematical model which captures the underlying data generating a process for the series. The future events are then predicted using the model. Time series forecasting has essential applications in different fields. Often valuable strategic decisions and precautionary measures are taken based on the forecast results. Thus, making a useful forecast, i.e., fitting an adequate model to a time series is very important.

Stochastic Process

A time series is non-deterministic, i.e., cannot predict with certainty what will occur in future. Generally, a time series is assumed to follow certain probability model which describes the joint distribution of the random variable. The mathematical expression describing the probability behavior of a time-series named as a stochastic process. Thus, the sequence of observations of the series is a sample realization of the stochastic process that produced it.

A usual assumption is that the time series variables are independent and identically distributed following the normal distribution. The interesting point is that they follow some regular pattern in the long term.

ARIMA Time Series Model

One of the conventional methods in time series forecasting is known as the ARIMA model, which is known as Autoregressive Integrated Moving Average. ARIMA model can be fitted to time series data to predict future points in the series.

AR: Autoregressive. Uses the dependent relationship between observation and some number of

lagged observations
I: Integrated. The use of differencing of raw observation to make the time series stationary.
MA: Moving Average. Uses the dependency between observation and residual error from a moving average model applied to lagged observations.

There are three distinct integers (p, d, q) that are used to parametrized ARIMA models. Due to this, the ARIMA model shown as ARIMA (p, d, q). Together these parameters account for seasonality. The trend, and noise in data.

- p is the autoregressive part of the model. It incorporates the effect of past values into our model.
- d is the integrated part of the ARIMA model. This incorporates the order of differencing to apply to the time series.
- q is the moving average part of the model allows us to set the error of our model as a linear combination of the error values observed at previous points in the past.

The linear regression model construction including the specified number and type of terms and the data prepared by a degree of differencing to make the series stationary, i.e., to remove trend and seasonal structure that negatively affect the regression model.

MF Data Modeling: ICICI Prudential Top 100 Fund – Growth

For Forecasting using ARIMA, ICICI Prudential Top 100 Fund – Growth took from the database, and it is divided into train and test data. Rolling means and rolling standard deviation of the mutual fund is calculated.

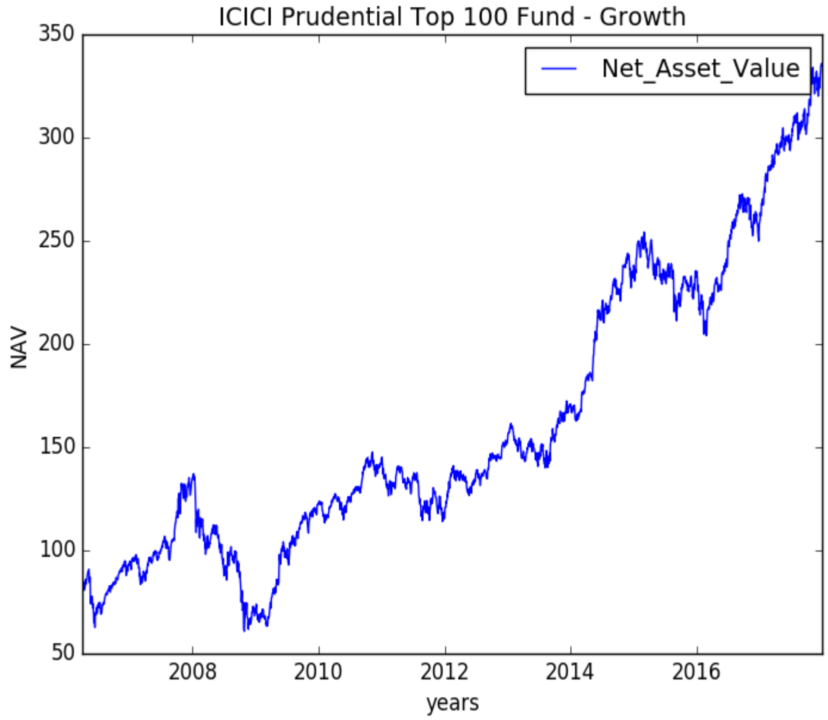


Figure 7: ICICI Prudential Top 100 - Growth (Original)

From the Figure 7, the plot of the graph shows that the Fund means is not stationary and mean, and standard deviation of the data is varying according to the time. As the mean and standard deviation changing with time is called as the stochastic mean and deviation. So, the data must be made stationary, only then ARIMA model can be applied to the data.

Firstly, the data must be made stationary, i.e., the trend, seasonality and cyclic from the data must be removed or minimized. Then we try to fit ARIMA model on the stationary data. If the model fits with the training data, then the forecasting of the model is made to predict future NAV of the fund.

Most of the time series model work on the assumption that the time series data is stationary. Stationarity is defined using strict criterion. We can consider the series to be stationary if it has constant statistical properties overtime

- Constant mean

- Constant variance
- Autocovariance that does not depend on time

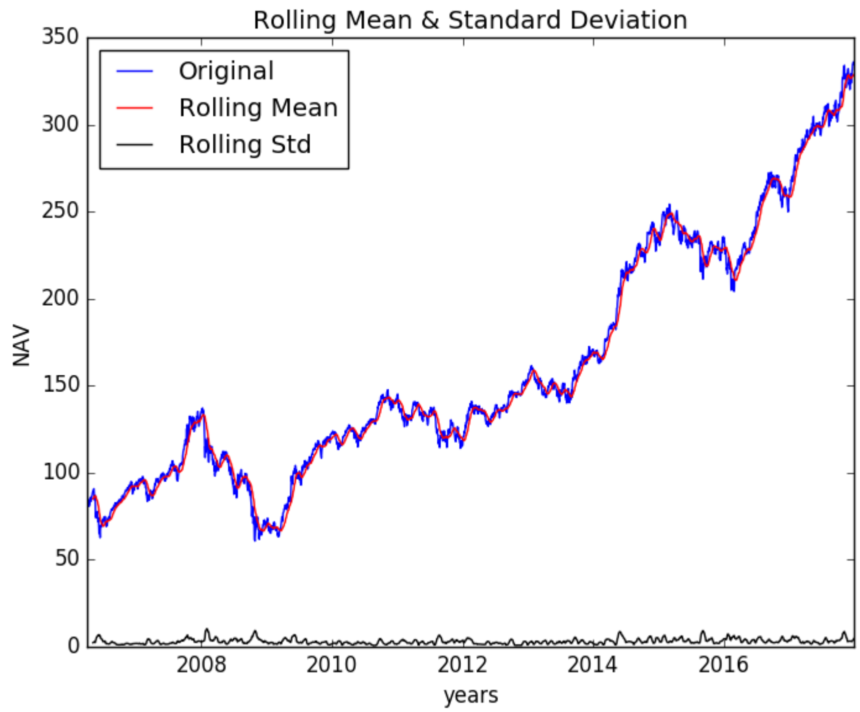


Figure 8: Rolling mean & SD

In the figure 7, the data is increasing trend with some seasonality; though seasonality is not visible in the graph, we need to check with a hypothesis test that if seasonality exists in the data. So, we can check stationarity using the following:

Plotting Rolling statistic: Plot the moving average or moving variance, and if the statistics vary with the time then stationarity does not exist.

Dickey-Fuller Test: This is the one of a statistical test to test stationarity. Here the null hypothesis is that time series is non-stationary. The test result comparison of a test statistic and critical value. If the test statistic is less than the critical value, we can reject the null hypothesis and can say that time series is stationary.

$$H_0: \textit{data is non stationary}$$
$$H_a: \textit{data is stationary}$$

From the rolling statistic, the data is not stationary. Using the Dickey-Fuller test we can conclude that the data is not stationary. However, according to the test, the test statistic is higher than the critical value. So, reject the null hypothesis, and the data is non-stationary. ARIMA model cannot apply to non-stationary data.

Observation: The test statistic is higher compared to the critical value. The rolling mean value varies with time and is not stationary. So, the series to be stationarized. We need to eliminate the trend and seasonality from the data to make it stationary.

```
Results of Dickey-Fuller Test:
Test Statistic                   1.116605
p-value                          0.995356
#Lags Used                       2.000000
Number of Observations Used   2892.000000
Critical Value (1%)             -3.432613
Critical Value (5%)             -2.862540
Critical Value (10%)            -2.567302
dtype: float64

Process finished with exit code 0
```

Transforming the data (eliminating trend)

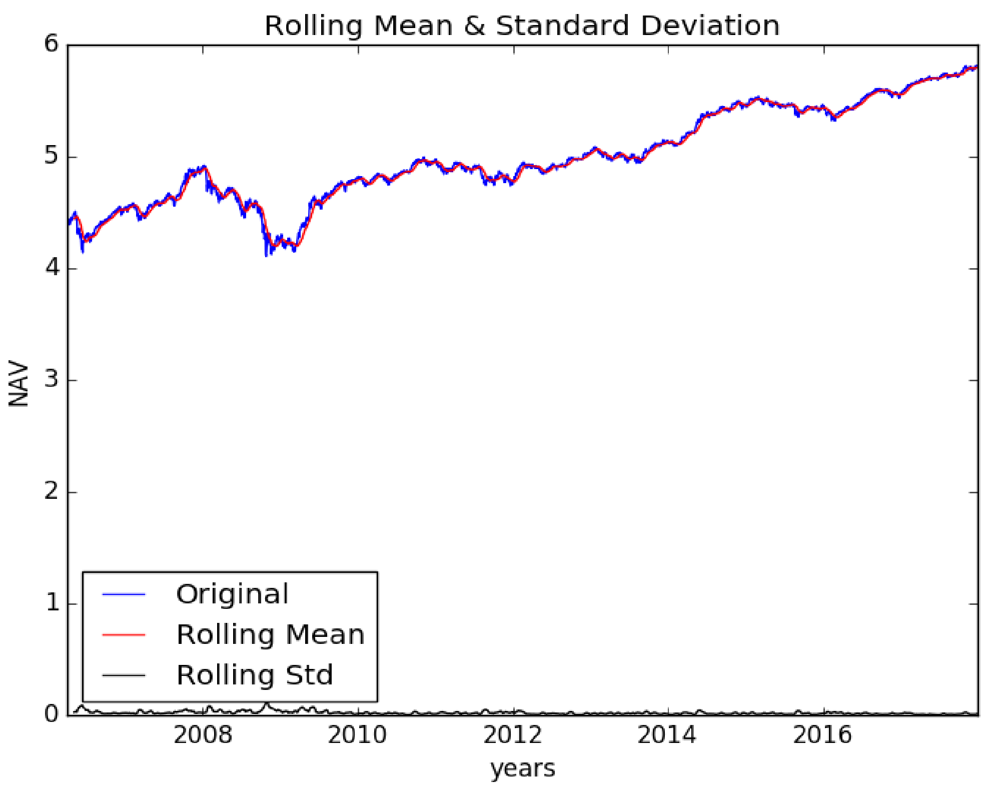


Figure 9: Log transformed plot

The series is log transformed to minimize the variance. In the figure 10 compared to the above figure 9, has more standard deviation then which was not log-transformed.

Checking for stationarity, the following result obtained.

```
Results of Dickey-Fuller Test:
Test Statistic                  -0.431590
p-value                          0.904689
#Lags Used                      28.000000
Number of Observations Used   2866.000000
Critical Value (1%)             -3.432634
Critical Value (5%)             -2.862549
Critical Value (10%)            -2.567307
dtype: float64

Process finished with exit code 0
```

From the Dickey-Fuller test, the obtained test statistic is still higher than the critical value, so the null hypothesis of the series is rejected. Still, the series not stationary, so the series has to transform to obtain stationarity.

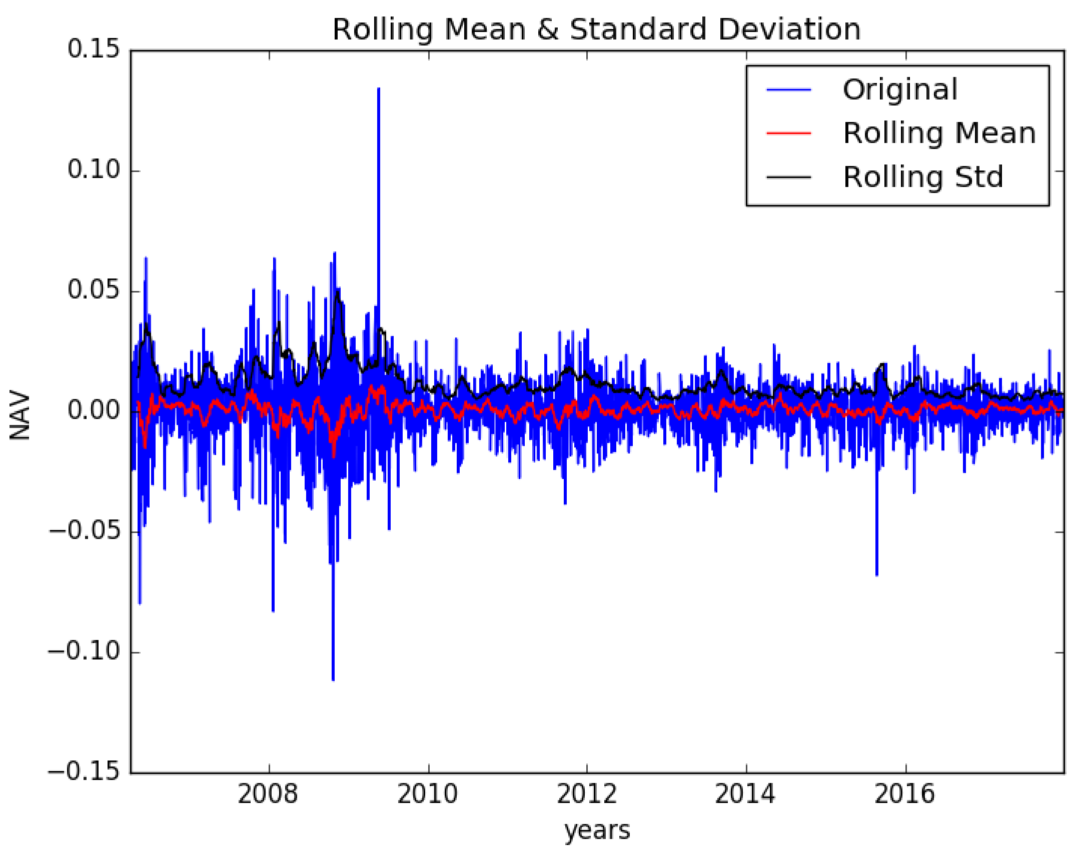


Figure 10: Log transformed differenced series

In the figure 10, the rolling mean and rolling standard deviation are not changing with time. The stationarity can only be confirmed using dickey fuller test.

From the dickey fuller test, the test statistic is far less compared to the critical value. So, the null hypothesis is rejected, and data is stationary. Further transformation of the data can be carried. Seasonality and trend existed in the data, that can be observed by decomposing the data.

```
Results of Dickey-Fuller Test:
Test Statistic                -9.381753e+00
p-value                        6.927213e-16
#Lags Used                     2.700000e+01
Number of Observations Used    2.866000e+03
Critical Value (1%)           -3.432634e+00
Critical Value (5%)           -2.862549e+00
Critical Value (10%)          -2.567307e+00
dtype: float64

Process finished with exit code 0
```

Decomposition of time series is a statistical process that deconstructs a time series into several components, each representing feature of the categories of data patterns.

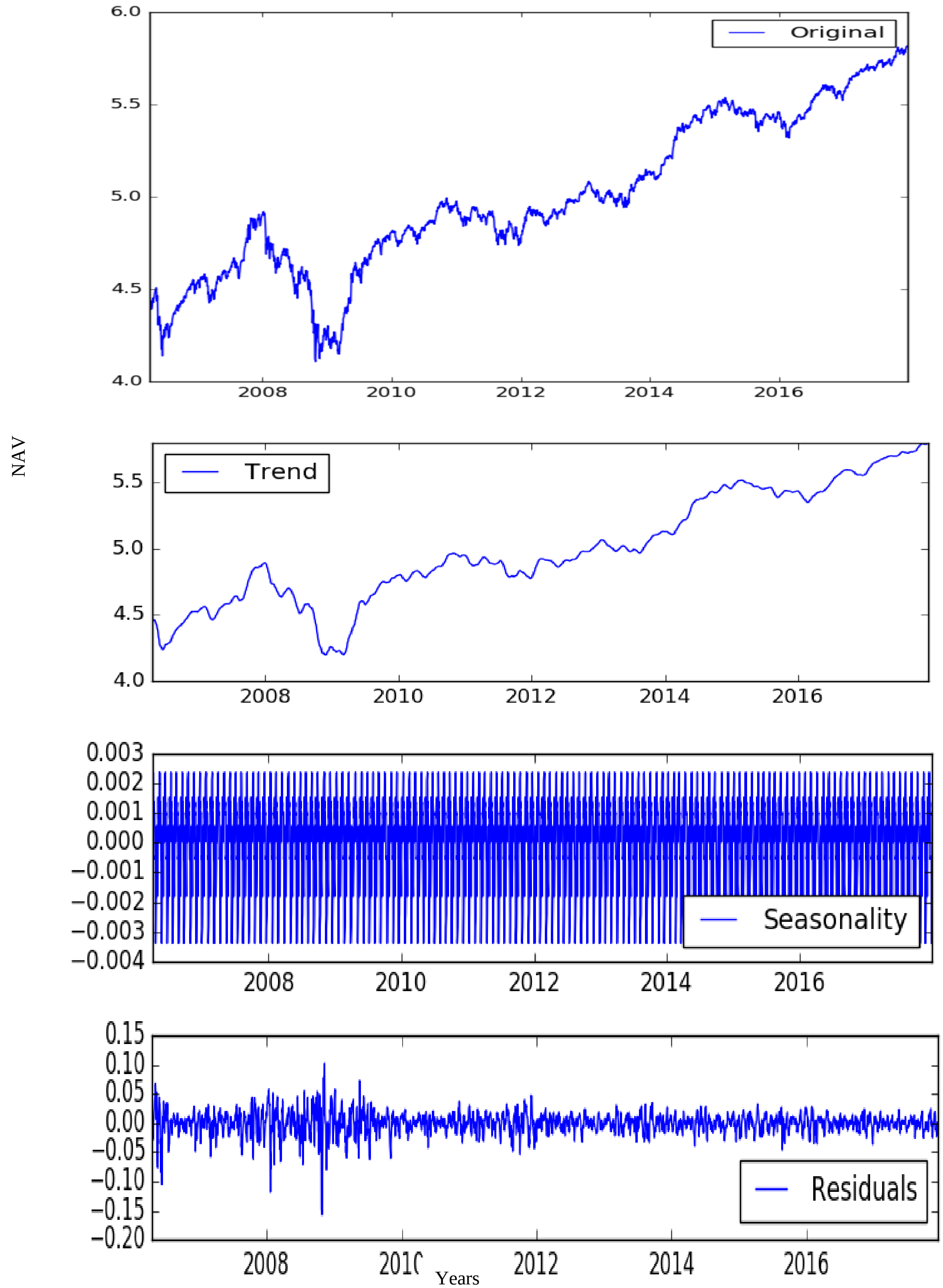
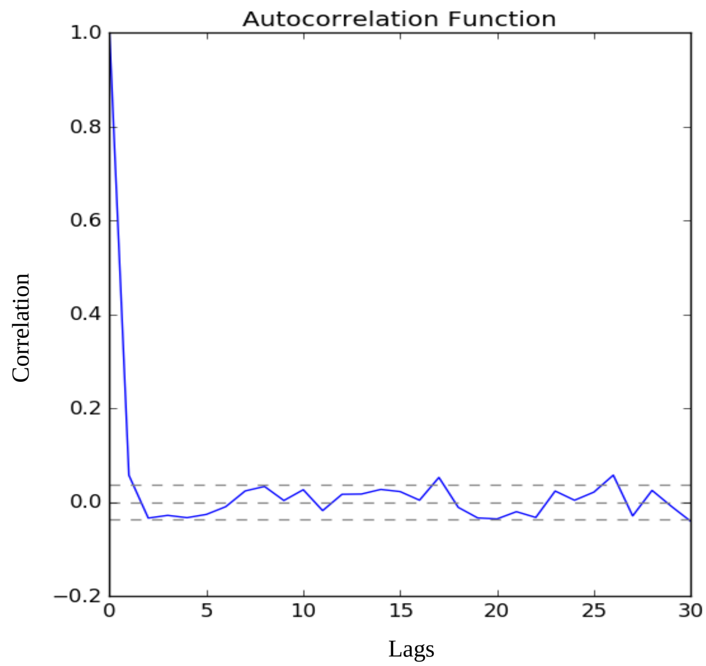


Figure 11: Decomposition Plots

Autocorrelation Function



It is also known as serial correlations, the correlation between two successive values x_t and x_{t+k}. Then correlation is called the autocorrelation of lag k of the series. Autocorrelations functions displays the autocorrelations on the vertical axis for successive values of k on the horizontal axis. The ACF represents the degree of persistence over respective lags of a variable. ACF will identify the order of the MA model.

Form the above figure; some lags cannot be identified, i.e., the number of times the ACF is crossing the upper confidence interval. The same figure transformed into the below (figure 12). The order of the MA is three according to the above number of lags.

The order of the MA, according to the ACF lags is four

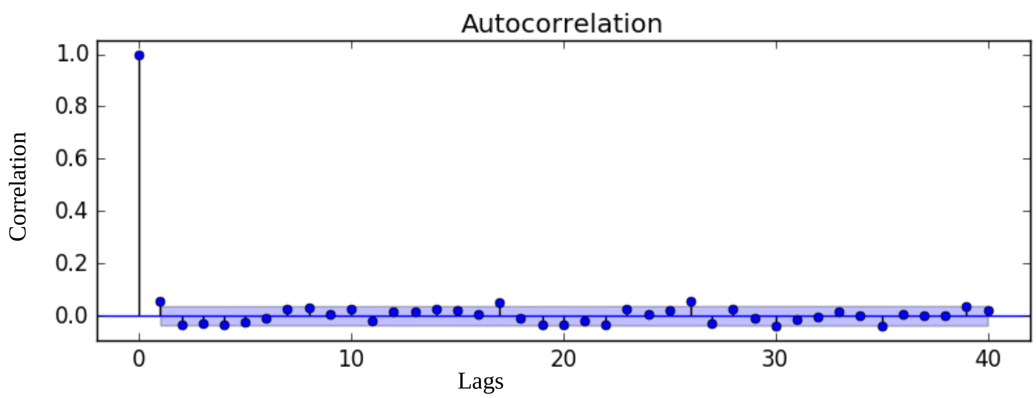
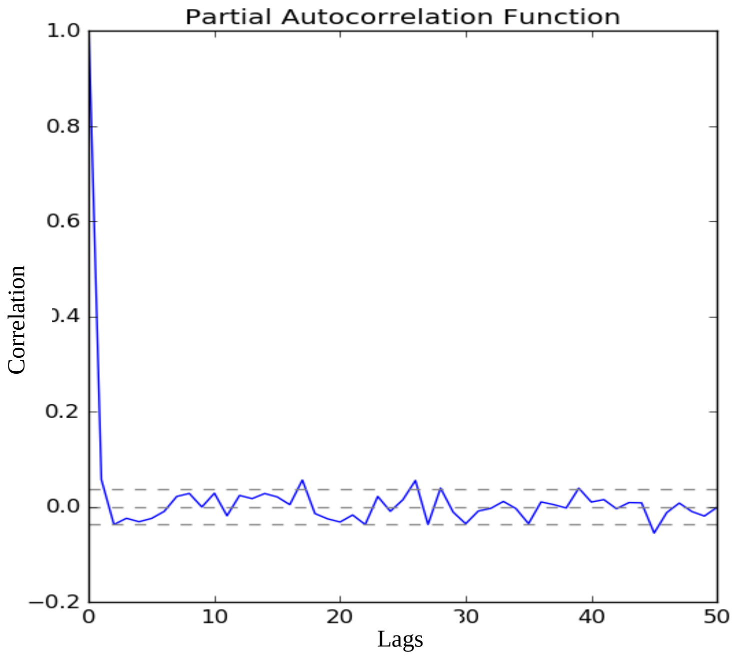


Figure 12: ACF Plot

Partial Autocorrelation function

PACF measures the correlation between the time series with a lagged version of itself but after eliminating the variations. In general, the partial correlation between variables is the amount of



correlation between them which is not explained by their mutual correlations with a specified set of other variables.

Partial correlations measure the degree of association between two random variables, with the effect of a set of controlling random variables removed.

Same way as of ACF, here the order of the PACF cannot be identified. So, the PACF figure transformed into the below form (figure 13).

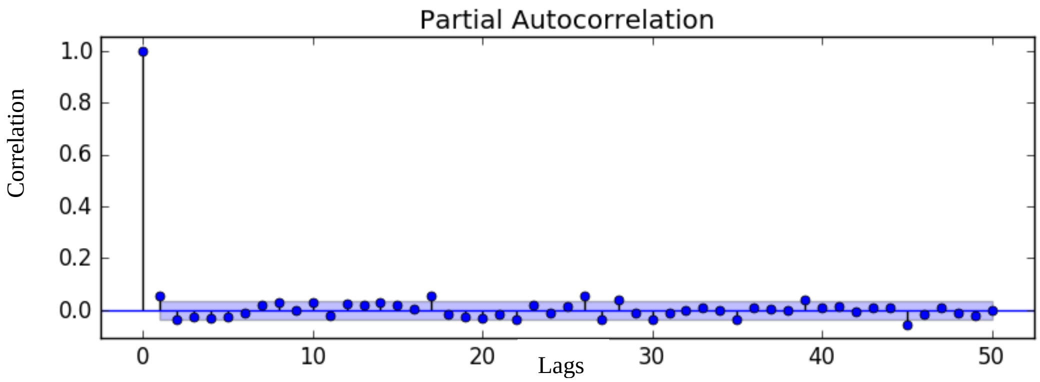


Figure 13: PACF Plot

The order of the AR, according to the PACF lags is four.

So, the parameter for the ARIMA:

p – 4

d -1

q - 4

The primary objective of the project is to automate the processes of selecting the time series model. This can be done by selecting the parameter which gives the lowest AIC and BIC values of the model. For this, we need to set the range of values for the parameters. The value of d can range from 0-2, p from 0 to 8 and q from 0 to 8. The automated code is run through the time series data until the best model is selected.

AIC and BIC

AIC and BIC are largely used in model selection. AIC is Akaike's information criteria, and BIC means Bayesian information criteria. AIC can be measured as the goodness of fit of any statistical model. The BIC is a type of model selection among a class of parametric models with a different number of parameters.

```
CONVERGENCE: REL_REDUCTION_OF_F_<=_FACTR*EPSMCH

 Cauchy                time 0.000E+00 seconds.
 Subspace minimization time 0.000E+00 seconds.
 Line search           time 0.000E+00 seconds.

 Total User time 0.000E+00 seconds.

The best suited paramters for ARIMA (4, 1, 4)

Process finished with exit code 0
```

AIC is useful for making asymptotic equivalent to cross-validation. Bayesian Information criteria are suitable for consistent estimation.
So here the model is parametric. So, BIC values are compared with all BIC values of different parameters of ARIMA model, and the min BIC parameters selected as the best model for forecasting. According to the PACF and ACF, the model section is validated, and forecasting of the model is done.

The parameters extracted from the result and these parameters assigned to the parameters of the $\mathbf{ARIMA}\ (\mathbf{p}, \mathbf{d}, \mathbf{q})$ and the model is executed. The executed model summary is generated, and model can be analyzed in more depth.

The coefficient column shows how weights of each feature and how each impacts the time series. Here each weight has p-value close to 0. So, it is reasonable to include the features in our model.

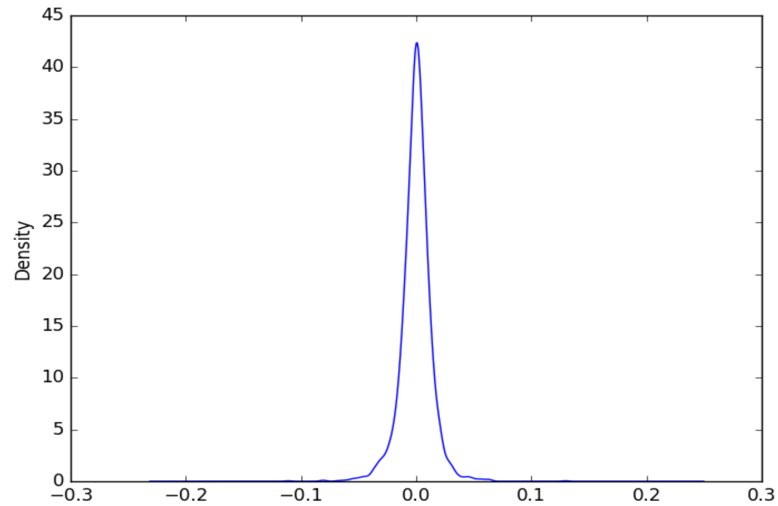


Figure 14: Density Plot

Before proceeding with the model, it is essential to run the model diagnosis to ensure that none of the assumptions made are violated. Firstly, check residual error plot, there may be some trend information not captured by the model. The figure 14 shows the distribution of residual errors. It shows a little bias in the prediction. Density plot of the residual error values, suggesting the errors are Gaussian, but may not be centered on zero.

```
                              ARIMA Model Results
==============================================================================
Dep. Variable:      D.Net_Asset_Value   No. Observations:                 2894
Model:                 ARIMA(4, 1, 4)   Log Likelihood                8373.109
Method:                       css-mle   S.D. of innovations              0.013
Date:                Fri, 25 May 2018   AIC                         -16726.217
Time:                        13:03:32   BIC                         -16666.513
Sample:                             1   HQIC                        -16704.703

===========================================================================================
                              coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                       0.0005      0.000      1.908      0.056   -1.32e-05       0.001
ar.L1.D.Net_Asset_Value    -0.3868      0.038    -10.295      0.000      -0.460      -0.313
ar.L2.D.Net_Asset_Value     0.5786      0.027     21.647      0.000       0.526       0.631
ar.L3.D.Net_Asset_Value    -0.5270      0.041    -12.745      0.000      -0.608      -0.446
ar.L4.D.Net_Asset_Value    -0.8925      0.049    -18.286      0.000      -0.988      -0.797
ma.L1.D.Net_Asset_Value     0.4276      0.032     13.397      0.000       0.365       0.490
ma.L2.D.Net_Asset_Value    -0.5893      0.026    -22.996      0.000      -0.639      -0.539
ma.L3.D.Net_Asset_Value     0.5105      0.035     14.784      0.000       0.443       0.578
ma.L4.D.Net_Asset_Value     0.9182      0.039     23.642      0.000       0.842       0.994
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            0.6973           -0.7189j            1.0015           -0.1274
AR.2            0.6973           +0.7189j            1.0015            0.1274
AR.3           -0.9926           -0.3631j            1.0569           -0.4442
AR.4           -0.9926           +0.3631j            1.0569            0.4442
MA.1            0.6988           -0.7222j            1.0049           -0.1276
MA.2            0.6988           +0.7222j            1.0049            0.1276
MA.3           -0.9767           -0.3527j            1.0385           -0.4448
MA.4           -0.9767           +0.3527j            1.0385            0.4448
-----------------------------------------------------------------------------

Process finished with exit code 0
```

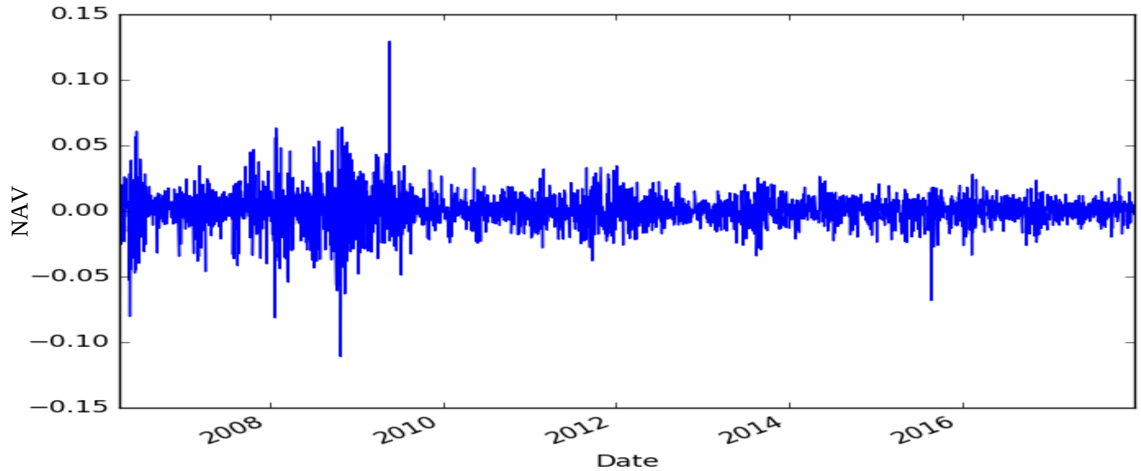


Figure 15: Residual Plot

Forecasting

Before forecasting the model must fit the historical data which was transformed by differencing the data to make it stationary. When the model fits the historical data, then the differenced data is transformed into exponential series to fit with the original data. In the figure 16 can be seen that the fitted model following the trend of the data.

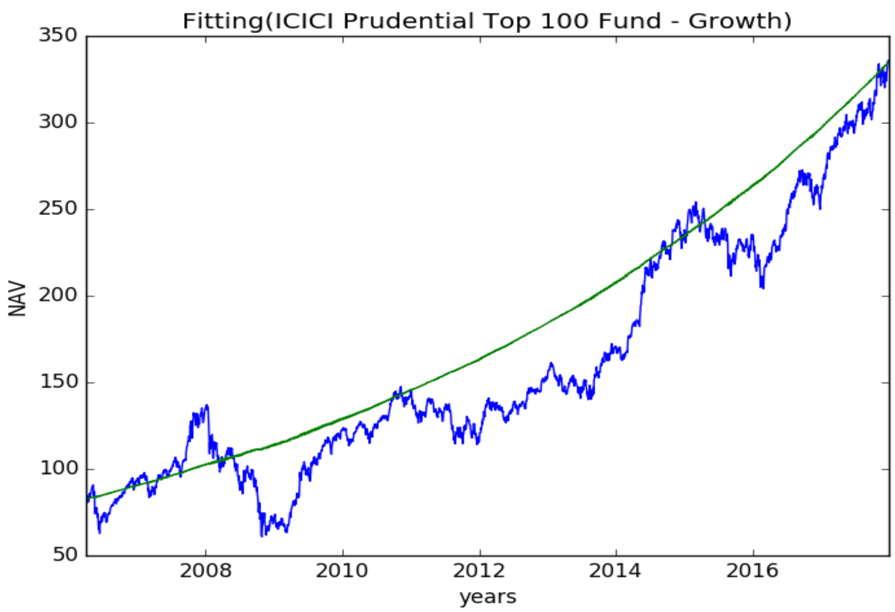


Figure 16: ARIMA-Original Fitting Plot

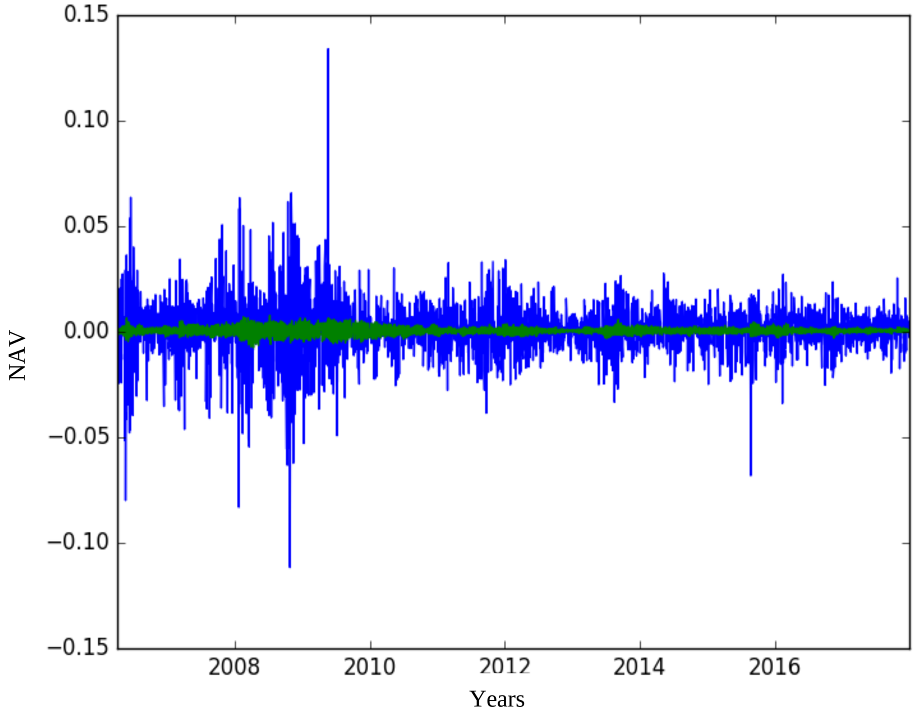


Figure 17: ARIMA-Residual Fitting Plot

In the figure 18, test data and forecasted values are plotted with a confidence interval. However, the model was not able to show the seasonality in the fitting data. So, the forecast is a linear trend with a positive slope.

The MSE of the forecasted data and test data.

```
... date_nav = pd.to_datetime(data['Date']) #dtype: datetime64[ns] Convert argument to datetime.
... ts_nav.index=date_nav
...
>>> np.sqrt(mean_squared_error(f_prediction_data,data_test['Net_Asset_Value']))
23.151045836253811
```

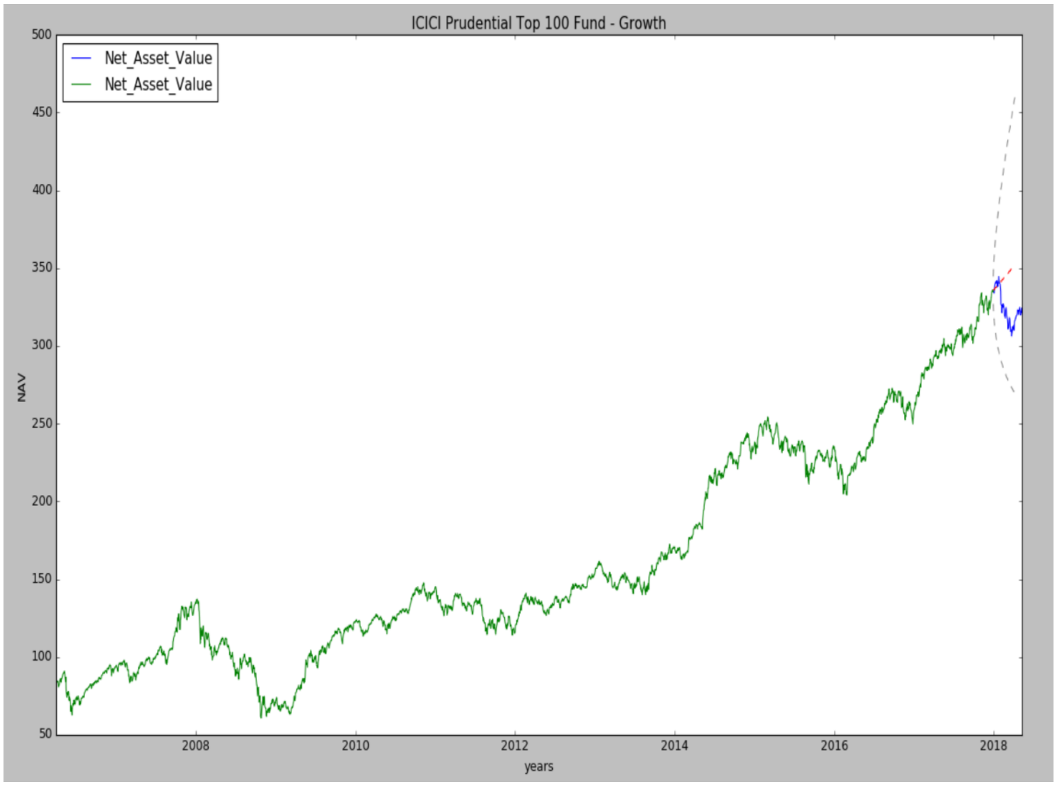


Figure 18: ARIMA-Forecast

SARIMA Model

SARIMA is also known as Seasonal Autoregressive Integrated Moving Average model. ARIMA is for non-seasonal and non-stationary data. In this model seasonal differencing of appropriate order is used to remove non-stationarity from series. A first-order differencing is differencing between an observation and the corresponding observation from the previous year.

The model generally termed as $\boldsymbol{SARIMA}\ (\boldsymbol{p}, \boldsymbol{d}, \boldsymbol{q}) \times (\boldsymbol{P}, \boldsymbol{D}, \boldsymbol{Q})$. The parameters of SARIMA

- p: Non-seasonal AR order
- d: Non-seasonal differencing
- q: Non-seasonal MA order

- P: Seasonal AR order
- D: Seasonal differencing
- Q: Seasonal MA order

- S: Time span of repeating seasonal pattern

$$\boldsymbol{SARIMA}\ (\boldsymbol{p}, \boldsymbol{d}, \boldsymbol{q}) \times (\boldsymbol{P}, \boldsymbol{D}, \boldsymbol{Q})._{\boldsymbol{S}}$$

The data used for the model is same as that used for ARIMA. So, the process is same as ARIMA. The automated process selects the required best model for forecasting.

```
-16700.0173499
The best suited paramters for ARIMA SARIMA(1, 0, 1)x(0, 0, 0, 3)
```

The above output is showing the best model selected. For the selection of the best model out of range of models, combinations of all parameters iterated so that the best model is with lowest BIC is selected, and the required parameters extracted. Only then the model is used for forecasting. Below showing the model summary.

```
>>> print(results.summary())
                           Statespace Model Results
==============================================================================
Dep. Variable:          Net_Asset_Value   No. Observations:                 2895
Model:                 SARIMAX(1, 0, 1)   Log Likelihood                8353.009
Date:                  Sun, 27 May 2018   AIC                         -16700.017
Time:                          22:50:59   BIC                         -16682.107
Sample:                               0   HQIC                        -16693.563
                               - 2895
Covariance Type:                    opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.0001   5.68e-05   1.76e+04      0.000       1.000       1.000
ma.L1          0.0593      0.011      5.341      0.000       0.038       0.081
sigma2         0.0002   2.12e-06     85.794      0.000       0.000       0.000
===================================================================================
Ljung-Box (Q):                       74.28   Jarque-Bera (JB):               8337.00
Prob(Q):                              0.00   Prob(JB):                          0.00
Heteroskedasticity (H):               0.19   Skew:                             -0.11
Prob(H) (two-sided):                  0.00   Kurtosis:                         11.31
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

The plot diagnostics object allows to generate model diagnostics and investigate for any unusual behavior quickly. Our primary concern is to ensure that the residuals of the model are uncorrelated and normally distributed with zero mean. If the seasonal ARIMA model does not satisfy these properties, it is a good indication that it can be further improved.

The model diagnostic suggests that the model residual usually is distributed based on the following:

- In the top right plot, the red KDE line follows closely with the N(0,1) line. Where N (0,1) is the standard notation for a normal distribution with mean 0 and standard deviation of 1. This is a good indication that the residuals are normally distributed. The forecast errors deviate somewhat from the straight line, indicating that the normal distribution is not a perfect model for the distribution of forecast errors, but it is not unreasonable.
- The qq-plot on the bottom left shows that the ordered distribution of residuals (blue dots) follows the linear trend of the samples taken from a standard normal distribution. Again, this is a strong indication that the residuals are normally distributed.

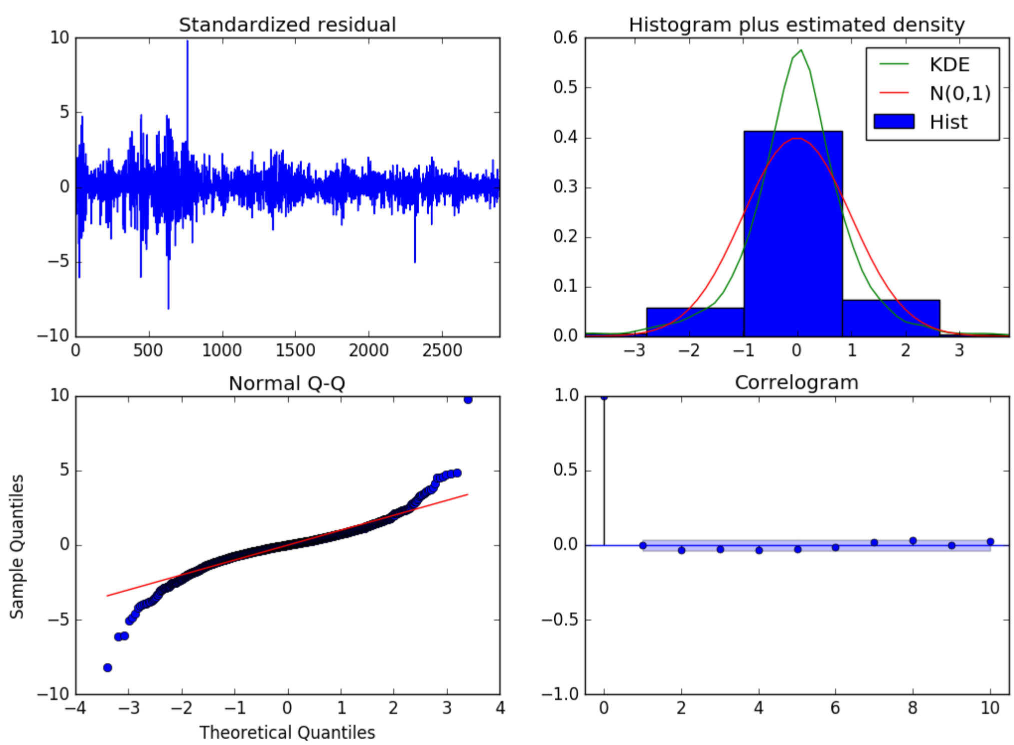


Figure 19: SARIMA - Model Diagnosis

- The residuals over time (top left plot) doesn't display any apparent seasonality and appear to be white noise. This is confirmed by the autocorrelation (i.e., correlogram) plot on the bottom right, which shows that the time series residuals have low correlation with lagged versions of itself.

```
>>> act_pred.head()
      lower Net_Asset_Value  upper Net_Asset_Value
2896             322.944765             348.798067
2897             320.192497             352.181586
2898             317.933914             355.072028
2899             315.985288             357.653124
2900             314.254468             360.017023
```

Those observations lead us to conclude that our model produces a satisfactory fit that could help us understand our time series data and forecast future values. So now we have a model that can forecast future NAV. By using the confidence interval generated by the model, compare the real values are in the generated confidence interval. If they are in the interval forecast the future NAV.

SARIMA Forecasting

The below figure 20 is model fitting. The model is forecasted and is plotted to validate the test data.

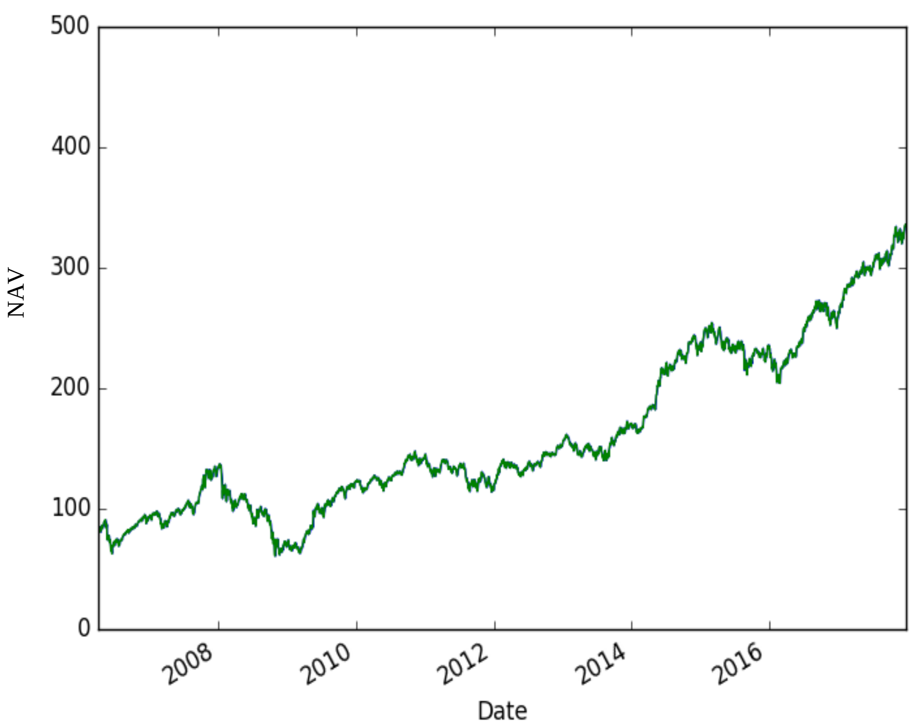
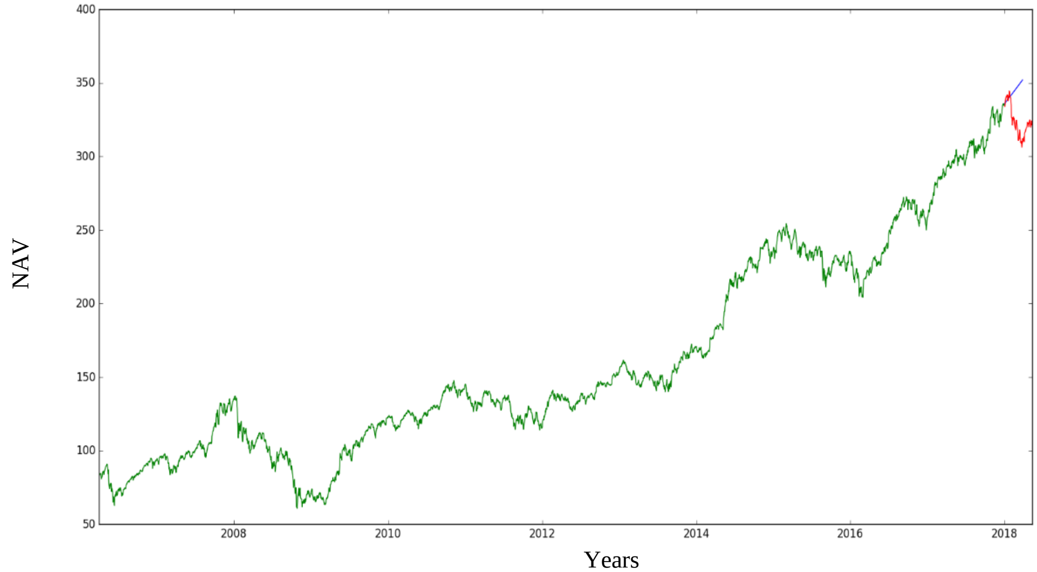


Figure 20: SARIMA- Fitting



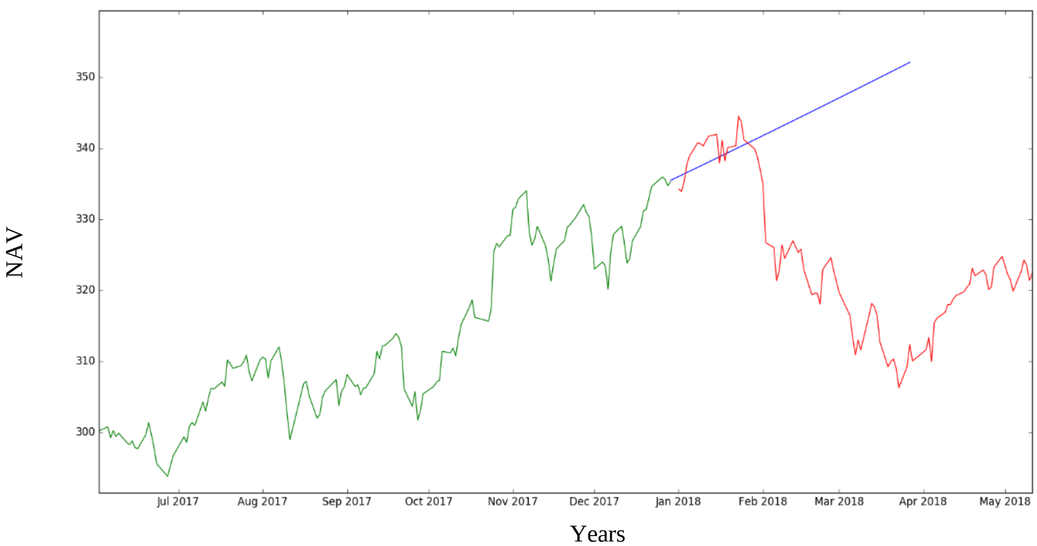


Figure 21: SARIMA-Forecasting

The above figure 21 shows the forecasted plot of the SARIMA. The model is unable to predict the test data.

The MSE of the model prediction with test data is higher than ARIMA model.

```
>>> np.sqrt(mean_squared_error(predicted_list,data_test['Net_Asset_Value']))
24.113583091948204
```

Neural Networks (LSTM)

An artificial neural network (ANN) is a computational model based on the structure and functions of biological neural networks. Information that flows through this neural network affects the structure of ANN because a neural network learns, in a sense based on input and output.

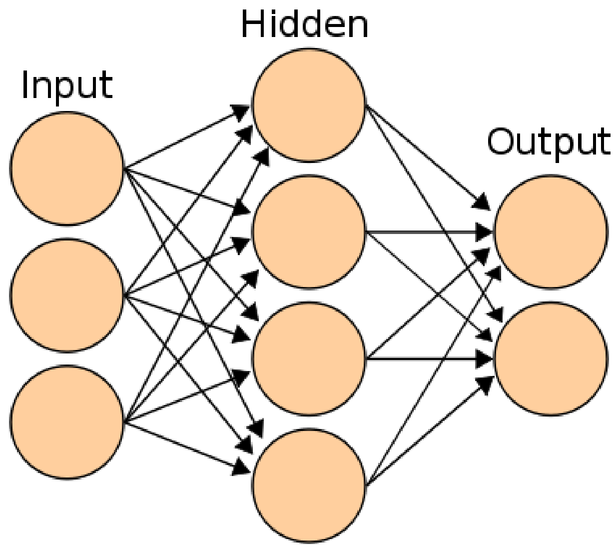


Figure 22: Neural Network

ANN is nonlinear statistical data modeling tool where the complex relationships between inputs and outputs are modeled, or patterns are found. ANN can learn from the observed data sets. ANN takes data samples rather than entire data sets to arrive at solutions. ANN is considered the relatively simple mathematical model to enhance existing data analysis technologies.

ANN has three layers; the first layer consists of input neurons. Those neurons send data on to the second layer, which in turn sends the output neurons to the third layer. Training an artificial neural network involves choosing from allowed models for which there are several associated algorithms.

Recurrent Neural Network (RNN)
The figure 23 is showing the Recurrent neural network. They are a network with loops in them, allowing information to persist.
An RNN is a class of Artificial neural network which process sequential data and takes in as

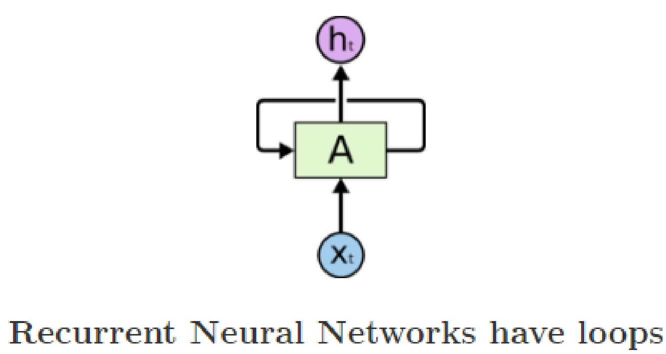
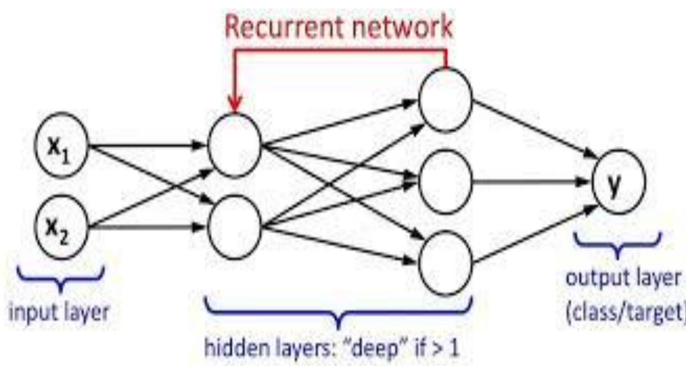


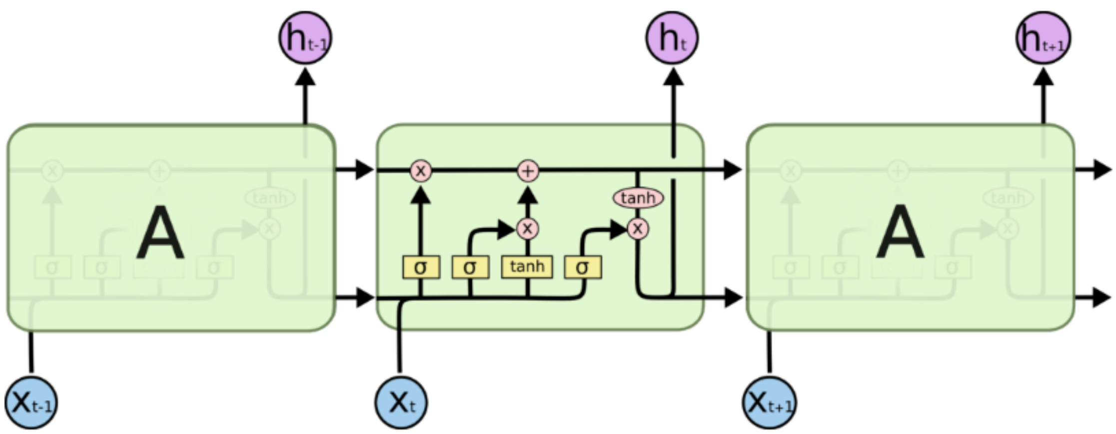
Figure 23: LSTM (ANN)

input both the new input at the current timestep and the output of the net in the previous timestep. The most popular type of RNN is probably the LSTM. Which has a cell state at each time step that changes with the new input.

Long Short-Term Memory Network (LSTM)

LSTM is the kind of RNN, capable of learning long-term dependencies. LSTM is explicitly designed to avoid long-term dependency problems. Remembering the information for a long period is practically their default behavior. LSTM networks have memory blocks that are connected through layers.





The repeating module in an LSTM contains four interacting layers.

A block has components that make the model smarter than a typical neuron and a memory for new sequences. A block contains gates that manage the block's state and output. A block operates upon an input sequence, and each gate within a block uses the sigmoid activation units to control whether they are triggered or not, making the change of state and addition of information flowing through the block conditional.

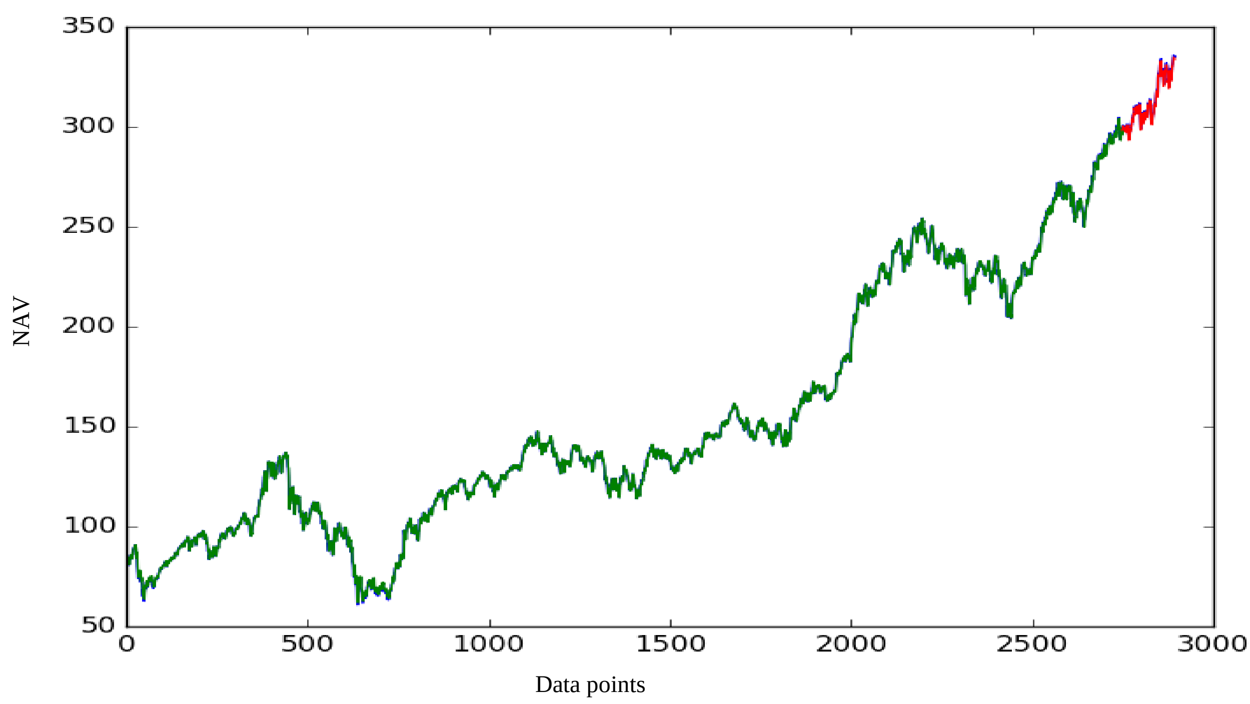
Within a gate, there are three types of gates:

- Forget Gate: conditionally decides what information to throw away from the block.
- Input Gate: conditionally decides which values from the input to update the memory state.
- Output Gate: conditionally decides what to output based on input and the memory of the block.

Each unit is like a mini-state machine, gates of the units have weights that learn during the training. LSTM networks in python using tensor flow deep learning library to address a time-series prediction problem.

Model Fitting

While fitting the model, i.e., predicting the NAV value, the process backed by the tensor flow. The below figure 24 shows the fitted model of the time-series data.



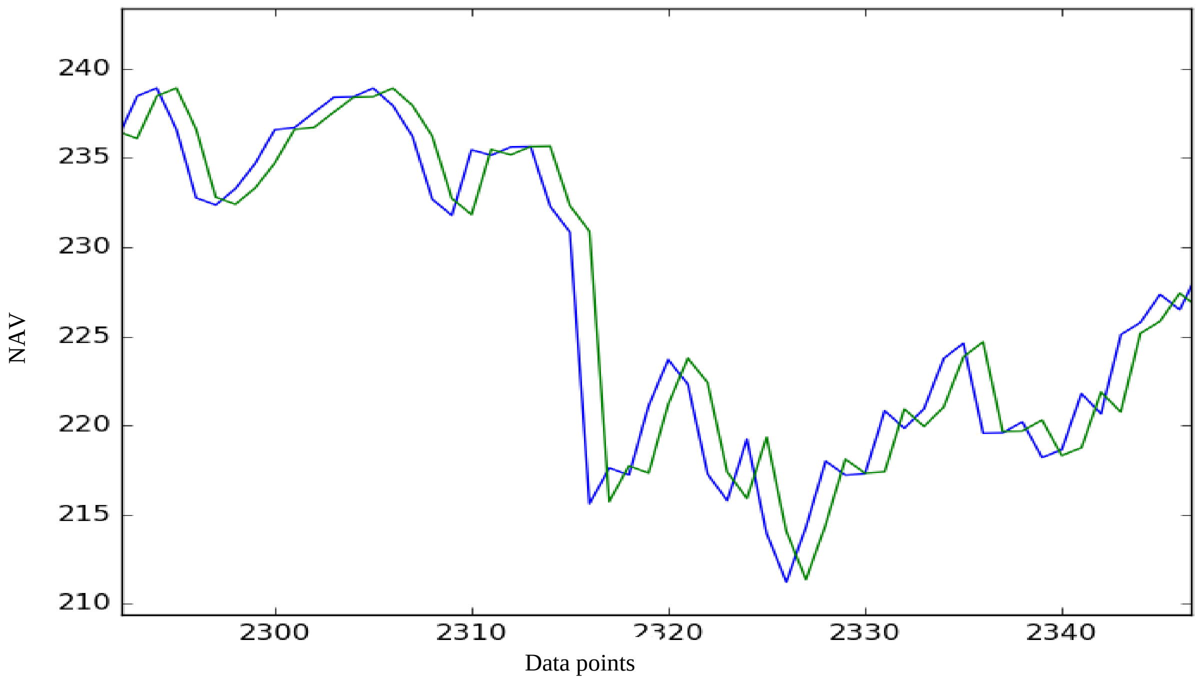


Figure 24: LSTM-Model Fitting

The above figure 24 is showing the close look of the LSTM prediction accuracy. This model has performed better than the ARIMA and SARIMA. However, the drawback of this model is that it can predict only one-step forecast. If the data is daily data, then the model can predict only one day NAV value

The LSTM model can validate the test data, but from the test data, the model can only forecast only one day NAV.

```
NAV for last 5 days:
[[ 331.83190918]
 [ 333.49111938]
 [ 334.83569336]
 [ 334.48242188]
 [ 333.61871338]]
 NAV for tomorrow:  [[ 332.57687378]]
```

The RMSE of the model shown after the model forecast shown in the figure 25.

```
Train Score: 1.73 RMSE
Test Score: 2.33 RMSE

Process finished with exit code 0
```

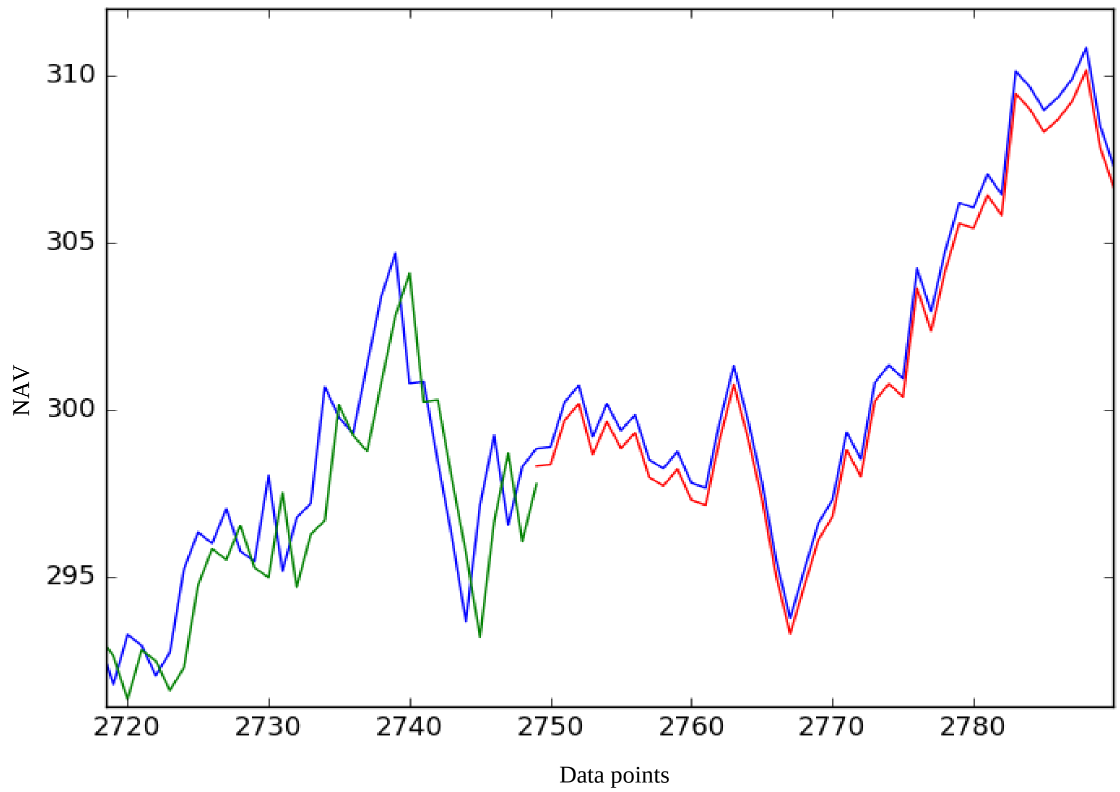


Figure 25: LSTM-Forecast

Table 2: RMSE of Models

S.no	Model	RMSE
1	ARIMA	24.11
2	SARIMA	23.15
3	LSTM	2.33

CHATBOT

A computer program designed to simulate communication between machine and humans. It is an assistant that communicates with us through text messages, a virtual companion that integrated into websites. Such a bot is an automated process of communication with users.

Chatbot works on Natural language processing. i.e., the user input message processed as the machine can understand and give the required input that a user can understand.

Objective

The objective of the developing the chatbot for DCB is to provide clients with required information about mutual funds.

Working

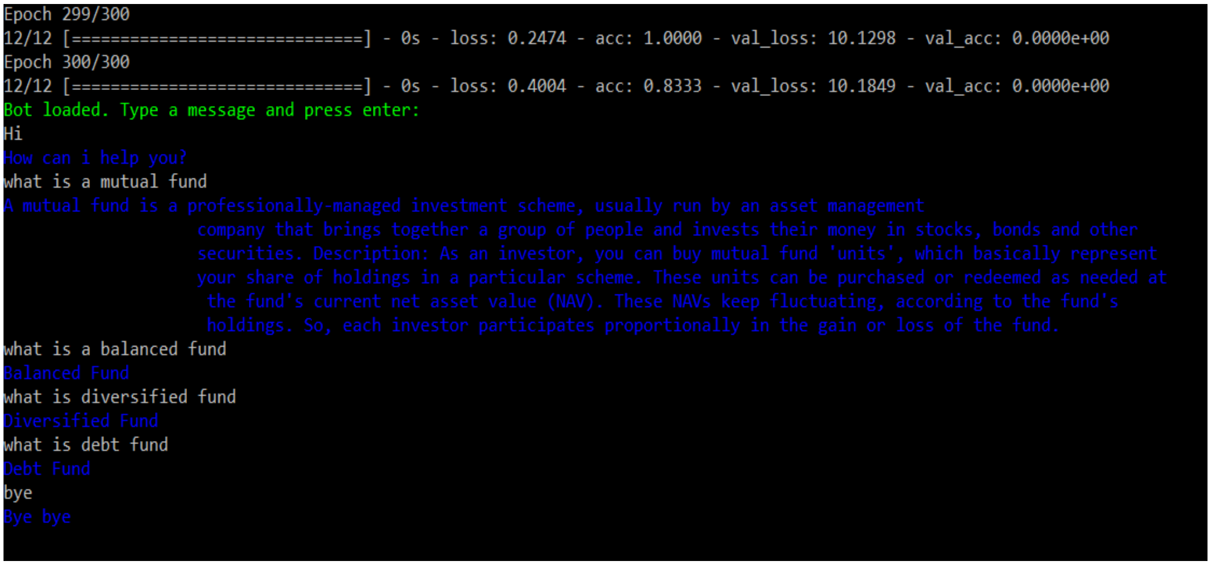
A chatbot integrated with Rasa nlu and Rasa core these are the natural language processing packages available on the python platform. Rasa only provides an interface for defining intents and entities of the training data. To process the given input a package named spacy is used to rank the input data according to their intents. In the spacy module, English 'en' is used to process the training data.

Natural Language processing

The field focuses on the interaction between human language and computers is called natural language processing. NLP is a way for the computer to analyze driven meaning from human language in a smart and useful way.

NLP algorithms typically based on machine-learning algorithms. Natural language processing can rely on machine learning to automatically learn these rules by analyzing a set of examples and make a statistical interface. The more the data analyzed, the more accurate the model will be.

The bot is trained with FAQ's related to the mutual funds, and these questions are mapped to the required intents and entities so that the bot can call the required action. This action gathers the required data from the knowledge base and posts the info related to the question.



```
Epoch 299/300
12/12 [==============================] - 0s - loss: 0.2474 - acc: 1.0000 - val_loss: 10.1298 - val_acc: 0.0000e+00
Epoch 300/300
12/12 [==============================] - 0s - loss: 0.4004 - acc: 0.8333 - val_loss: 10.1849 - val_acc: 0.0000e+00
Bot loaded. Type a message and press enter:
Hi
How can i help you?
what is a mutual fund
A mutual fund is a professionally-managed investment scheme, usually run by an asset management
                    company that brings together a group of people and invests their money in stocks, bonds and other
                    securities. Description: As an investor, you can buy mutual fund 'units', which basically represent
                    your share of holdings in a particular scheme. These units can be purchased or redeemed as needed at
                     the fund's current net asset value (NAV). These NAVs keep fluctuating, according to the fund's
                     holdings. So, each investor participates proportionally in the gain or loss of the fund.
what is a balanced fund
Balanced Fund
what is diversified fund
Diversified Fund
what is debt fund
Debt Fund
bye
Bye bye
```

Figure 26: Chatbot output

References

http://www.dcbbank.com/cms/showpage/page/about-us

https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

https://pypi.org/

http://fundpicker.thefundoo.com/

https://www.amfiindia.com/

https://www.slideshare.net/21_venkat/arima-26196965

https://newonlinecourses.science.psu.edu/stat510/node/67/

https://www.techopedia.com/definition/5967/artificial-neural-network-ann

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

https://www.nseindia.com/products/content/equities/equities/eq_security.htm

https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm/

https://nlu.rasa.com/tutorial.html

https://machinelearningmastery.com/natural-language-processing/

https://www.moneycontrol.com/mutualfundindia/

https://www.mutualfundssahihai.com/en/what-are-the-various-types-of-funds

https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/

https://towardsdatascience.com/time-series-analysis-in-python-an-introduction-70d5a5b1d52a

Abbreviations

MF	Mutual Fund
TS	Time Series
NAV	Net Asset Value
IDE	Integrated Development Environment
Rf	Risk Free Rate
ARIMA	Autoregressive Integrated moving averages
AR	Autoregressive
MA	Moving Averages
ARMA	Autoregressive Moving Averages
SARIMA	Seasonal Autoregressive Integrated Moving Averages
ANN	Artificial Neural Networks
RNN	Recursive Neural Network
LSTM	Long Short-Term memory
NLP	Natural Language Processing