A REPORT ON DATA ANALYSIS OF INVESTMENTS, PORTFOLIO MANAGEMENT AND TIME SERIES FORECASTING MODELS OF NATIONAL STOCK EXCHANGE(NSE) DATA

BY

ID Number	Name
2016B4A70166P	Aaryan Kapoor
2016B4A70632P	Satyansh Rai
2016A3PS0179P	Utkarsh Bajaj
2016B4A20603P	Mayank Prasad
2017B4A20750P	Rutuvi Narang
2016B4A10589P	Sumeet Kumar Singh
2016B4AB0552P	Archana
2017B4A80518P	Devesh Santosh Todarwal

Prepared for course MATH F432 (Applied Statistical Methods)

AT

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

November, 2019

Introduction

Stock market refers to a group of companies and individuals, where the companies come to sell their stocks/shares. A company can have many stocks (millions of stocks are sold and bought every day). Basically, if a company has ‘x’ stocks and a person buys 1 stock of the company this means person owes 1/x percent that company. The person also gains when the company is in profit and loses when it goes into loss. As more and more people tend to buy stocks, it means more people trust that company to profit. This makes the valuation of company to go up, eventually making the price of one stock to go up too. The person owning the stock can sell the stock whenever s/he feels it has made sufficient profit or when trying to cut their loss.

The National Stock Exchange of India Limited (NSE) is the largest financial market in India. The NSE, introduced in 1992, has become the world's fourth-largest electronics, electronics market in 2015. The trade began in 1994 with the introduction of the wholesale debt market and money market segment shortly thereafter.

Today, the exchange conducts transactions in the wholesale debt, equity, and derivative markets. One of the more popular offerings is the NIFTY 50 Index, which tracks the largest assets in the Indian equity market.

Market Feedback & Index Methodology Review NSE Indices Limited is committed to ensuring that all NIFTY indices are relevant for the market participants. **Nifty 50** is nothing but the calculated weighted average of the performance of 50 companies from various sectors which are listed on NSE. In order to ensure this, NSE Indices Limited on an on-going basis interacts with the stakeholders inviting feedback through various channels of communication. The feedback received from the market participants forms a key input for all index related aspects. Review of methodology of NIFTY indices is carried out on an annual basis. Additionally, NSE Indices Limited also considers any feedback that it may receive with regards to index methodology as part of on-going market interactions. Any changes to the index methodology is approved by the Committee and the same is announced through a press release.

Why should one invest?

Investing is a way of building wealth. Anyone can start investing. What differentiates investing from gambling is that it requires patience. Investing is the act of committing money or capital to an endeavor with the expectation of obtaining an additional income or profit. Benefits of investing:

1. Financial security
2. Financial independence
3. Build your wealth
4. Attain your goals

To understand the importance of investing, we present the following example.

Assume you earn Rs.50000 per month and you spend Rs.30000 towards your cost of living. Few simple assumptions.

- The cost of living is likely to go up by 8% year on year
- You are 30 years old and plan to retire at 50. This leaves you with 20 more years to earn
- The effect of personal income tax is ignored
- The employer is kind enough to give you a 10% salary hike every year
- Your expenses are fixed and don't foresee any other expenses

Your total savings after 20 years of hard work is Rs.1.7Crs. Now consider the following situations.

After you retire, assuming the expenses will continue to grow at 8%. So Rs.1.7Crs is enough to help you through roughly for about 8 years of post-retirement life. 8th year onwards you will be in a very tight spot with literally no savings left to back you up.

Consider another situation where you choose to invest the cash in an investment option that grows at 12% per annum.

YEAR	ANNUAL INCOME	YEARLY EXPENSE	MONEY LEFT	Retained Cash Invested @12%
1	600,000	360,000	2,40,000	20,67,063
2	6,60,000	3,88,800	2,71,200	20,85,519
3	7,26,000	4,19,904	3,06,096	21,01,668
...	...	...	...	...
18	30,32,682	13,32,006	17,00,676	21,33,328
19	33,35,950	14,38,567	18,97,383	21,25,069
20	36,69,545	15,53,652	21,15,893	21,15,893
		TOTAL	1,78,90,693	4,26,95,771

This is 2.4 times the regular amount. This means you will be in a much better situation to deal with your post retirement life.

Under fixed deposit an investment is more secure as compared to stocks. But this security comes at a cost of low returns. Investment in the stock market would yield a higher dividend. Another aspect diving these two would be liquidity. Stocks can be sold or bought at an easier rate. Whereas, fixed deposits come with a stipulated lock-in period. If one wishes to withdraw their FDs within this time period, they will attract a pre-closure charging fees and receive the amount at a lower rate of interest.

	Fixed Deposit	**Stock Market**
Rate of Interest	Low (4-7%)	Very High (14-15%)
Volatility	Low	Very High
Liquidity	Moderate	Very High
Taxation	Interest income is fully taxable	Short-term dividends are taxed. Long-term dividends are tax-free

Portfolio

Portfolio refers to any combination of monetary assets or investments. Portfolios are handled by individual investors and/or managed by financial professionals, hedge funds, banks and other financial institutions. These financial assets or investments often include

1. Stocks
2. Bonds
3. Mutual funds
4. Cash

Generally, a portfolio is designed according to the risk tolerance, time frame and investment objectives of the person investing. The monetary value of each asset may influence the risk/reward ratio of the portfolio.

Portfolio Management vs Financial Planning

“The art of managing an individual’s investment is called portfolio management.”

Portfolio Management	Financial Planning
Act of creating and maintaining an investment account	Process of developing financial goals and creating a plan of action to achieve them
Professional licensed portfolio managers are responsible for portfolio management	Individuals may choose to self-direct their own investments
The ultimate goal is to maximize the expected return at an appropriate level of risk	The ultimate goal is to ensure availability of funds whenever these are required

Objectives of Portfolio Management

1. To grow the capital
2. To form liquid and more stable investments.
3. To diversify the risks of asset classes by proper asset allocation.
4. Higher returns than market as well as consistent returns.
5. Also used for planning taxes.

Some of the Important Statistical Components you will find when you start to analyse a portfolio.

1. β : Beta (B) is a measure of the volatility, or systematic risk, of a security or a portfolio in comparison to the market as a whole
2. BMrk α: Benchmark Alpha (α), often considered the active return on an investment, gauges the performance of an investment against a market index used as a benchmark
3. CVaR: Conditional Value at Risk (CVaR), also known as the expected shortfall, is a risk assessment measure that quantifies the amount of tail risk an investment portfolio has.

Fig 1: Ratio tear sheet for a company

Monte Carlo Simulations

Monte Carlo simulation performs risk analysis by building models of possible results by substituting a range of values (based on a probability distribution) for any factor that has inherent uncertainty. It then calculates results over and over, each time using a different set of random values from the probability functions. It is used to understand the impact of risk and uncertainty in financial, project management, cost, and other forecasting models

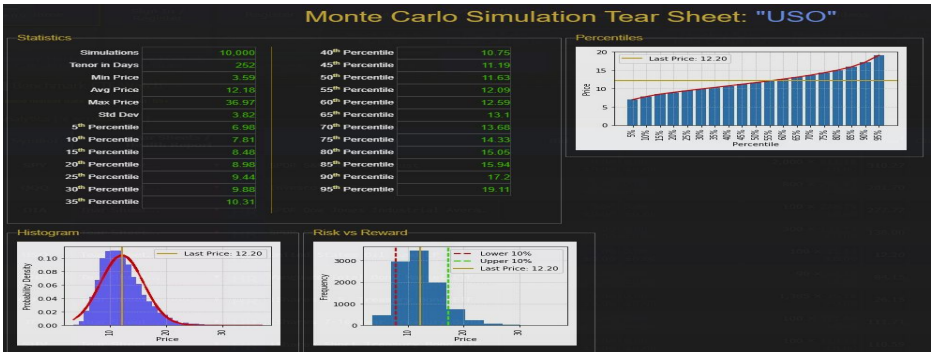

Fig 2: Monte carlo Simulation results

Portfolio performance evaluation essentially comprises of two functions, performance measurement and performance evaluation. Performance measurement is an accounting function which measures the return earned on a portfolio during the holding period or investment period. Performance evaluation, on the other hand, address such issues as whether the performance was superior or inferior, whether the performance was due to skill or luck etc.

Fig3: Snippets are taken from ZooNova.com. Such tools help us understand the performance of portfolio, which in turn helps us for smart investments.

Our Work

As a part of performing exploratory analysis, the aim of this report is to perform a comparative study on the time series forecasting of the NIFTY50 Index's price with time. The project deploys multiple statistical techniques taught as a part of the course MATH F432 Applied Statistical

Methods. The various algorithms, methodologies and outcomes of the project follow next upon which, the relevance of the project will be discussed.

The stock market has a large impact on the economy of a nation, this is why it is an interesting matter to see how stock market prediction can be used and whether or not the predicted results are valid.

1. Algorithms and Techniques

Stock market price production is a time series forecasting model, the solution for which can be approached using various classical statistical techniques and advanced models like Long Short Term Memory (LSTM) and ARIMA.

The aim of the project is to compare the performance of various statistical techniques in predicting the share market price. The following techniques have been deployed-

I. **Moving Average:** A moving average is calculated by taking the average of previous N stock prices, to give a single trend line. It is popular amongst traders because it can help to determine the direction of the current trend, while lessening the impact of random price spikes.

For each subsequent step, the predicted values are taken into consideration while removing the oldest observed value from the set. A moving average will enable us to examine the *levels of support and resistance*, by analyzing the previous movement of an asset's price. A moving average is primarily a lagging indicator, which makes it one of the most popular tools for technical time series analysis.

II. **Linear Trend Regression:** The linear regression model returns an equation that determines the relationship between independent variables and the dependent variable. A special case of a simple regression model is the linear trend regression, in which the independent variable is just a time index variable, i.e., 1, 2, 3 ... or some other equally spaced sequence of numbers. When it is estimated by regression, the trend line is the unique line that minimizes the sum of squared deviations from the data, measured in the vertical direction.

III. **K-Nearest Neighbors Algorithm (KNN):** KNN(similar to cluster analysis) is a non-parametric, lazy learning (does not build a model or function previously) statistical method. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point. It is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).

The prediction of stock market closing price is computed using kNN as follows:

a) Determine the number of nearest neighbors, k.
b) Compute the distance between the training samples and the query record.
c) Sort all training records according to the distance values.
d) Use a majority vote for the class labels of k nearest neighbors, and assign it as a prediction value of the query record.

IV. **Auto Regressive Integrated Moving Average (ARIMA):** It is a class of models that forecasts a given time series based on its own past values i.e. its own lags and the lagged forecast errors. A pure *Auto Regressive (AR only) model* is one where Yt depends only on its own lags. A pure *Moving Average (MA only) model* is one where Yt depends only on the lagged forecast errors.

A pure Auto Regressive model

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} + \epsilon_1$$

$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

A pure Moving Average model

An ARIMA model is one where the time series was differenced at least once to make it stationary and you combine the AR and MA terms. So the equation becomes:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

ARIMA Model

There are three important parameters in ARIMA:

a. p (order of auto-regression) :past values used for forecasting the next value

b. q (order of moving average): past forecast errors

c. d (order of differencing): to make time series stationary

Parameter tuning for ARIMA consumes a lot of time. So Auto ARIMA is used which automatically selects the best combination of (p,q,d) that provides the least error irrespective of existing seasonal, cyclical and trend components. ***SARIMA*** is a form of ARIMA considering seasonality of the data by calculating seasonal differences (similar to regular differencing, but, instead of subtracting consecutive terms, subtract the value from previous season).

V. **Long Short Term Memory (LSTM):** Long Short-Term Memory (LSTM) networks are a type of *recurrent neural network* capable of learning term dependence in sequence prediction problems. In the conventional feed-forward neural networks, all test cases are considered to be independent. Unlike traditional neural networks, at each timestep, LSTM takes the dataset as the input as well as the previous output, making it capable enough to learn better. LSTM uses *backpropagation* to minimize the *root-squared error* using optimization for each input. The weights of the network are adjusted accordingly for the 4 year training data. We have used 2 LSTM networks(each of layers=50) with 60 timesteps and a final output layer for our model. We have dropout techniques to prevent overfitting of data. The reason LSTM works so well is because LSTM is able to store past information that is important, and forget the information that is not. LSTM has following components:

~The input gate: The input gate adds information to the cell state

~The forget gate: It removes the information that is no longer required by the model

~The output gate: Output Gate at LSTM selects the information to be shown as output

~The memory cell

Prophet, developed by Facebook, also takes into consideration the trend as well as seasonality of the data.

2. Methodology

1. Data Pre-processing

The data to be dealt in the given problem set was that of the NIFTY50 Index for the past five years. There are multiple variables in the dataset – date, open, high, low, last, close, total trade quantity, and turnover. The data consisted of a total of 1234 records over the previous five years.

The columns Open and Close represent the starting and final price at which the stock is traded on a particular day. High, Low and Last represent the maximum, minimum, and last price of the share for the day. Total Trade Quantity is the number of shares bought or sold in the day and Turnover (Lacs) is the turnover of the particular company on a given date. Another important thing to note is that the market is closed on weekends and public holidays. The profit or loss calculation is usually determined by the closing price of a stock for the day, hence we will consider the *closing price as the target variable*. A line graph of the closing price against time is.

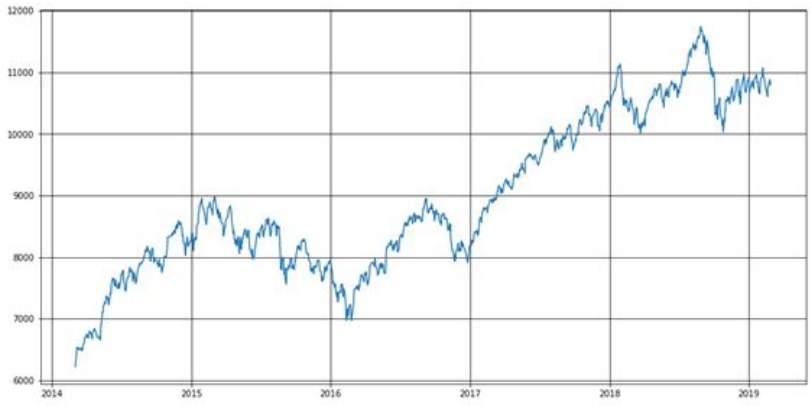

Figure 4: Share Price v/s Time

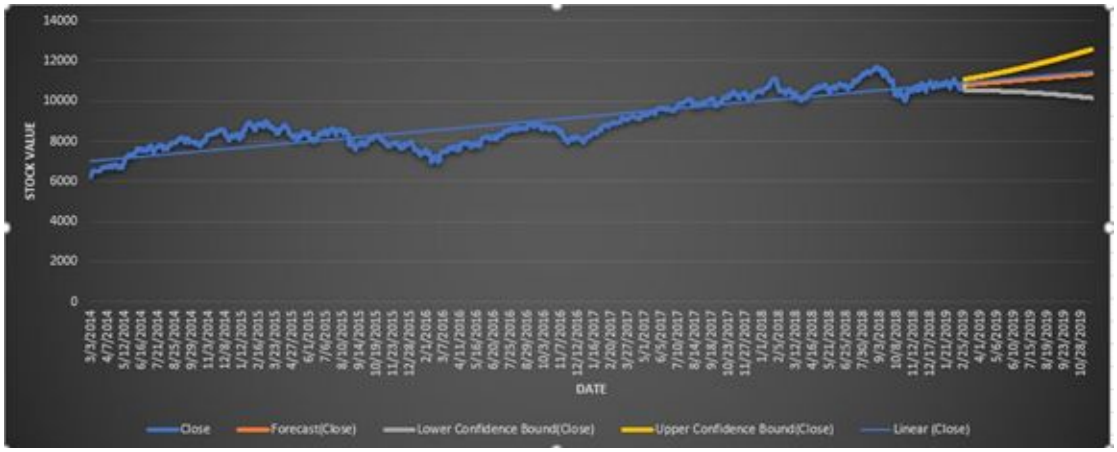

Figure 5: NIFTY50 Index Price Trend with upper and lower confidence bounds

2. Implementation: The data consisted of 1,234 records and to test the performance of the statistical models for the prediction of stock prices, the data set was split into training set (80% - 987 records) and the test set (20% - 247 records) for validating the prediction of the statistical models. The performance metrics for the statistical models included: RMSE, Comparison of prediction with the actual movement of NIFTY50 Index and finally, the robustness of the model's predictions. The results were as follows:

A. Moving Average- RMSE - 685.72259439

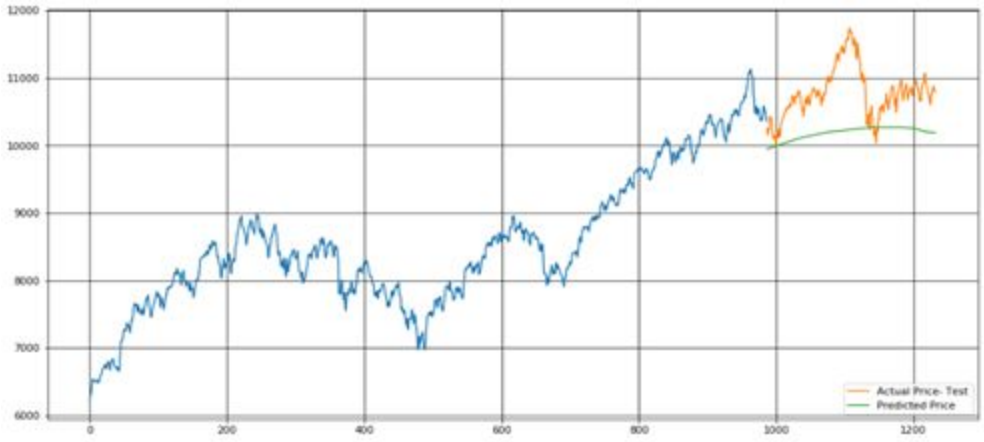

Figure 6: Moving Average Prediction (Price v/s Time)

The RMSE value is close to 685 but the results are not very promising (as you can gather from the plot). The predicted values are of the same range as the observed values in the train set (there is an increasing trend initially and then a slow decrease). A simple moving average model cannot work well because of two reasons. It takes time to account for sudden changes in data. Also, it gives a better idea about the trend but not the seasonality (up or down) from the trend.

B. **Linear Trend Regression –** RMSE - 564.3075943494691

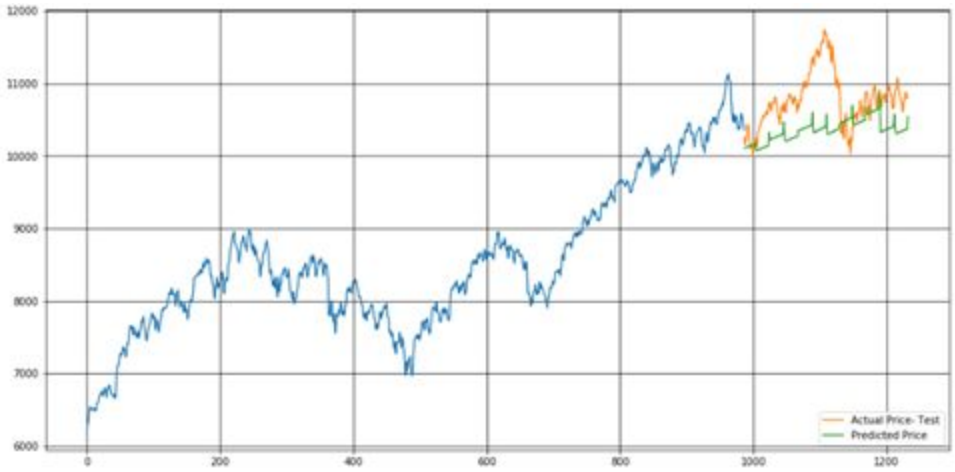

Figure 7(a): Linear Trend Regression Prediction considering Monday-Friday Factor

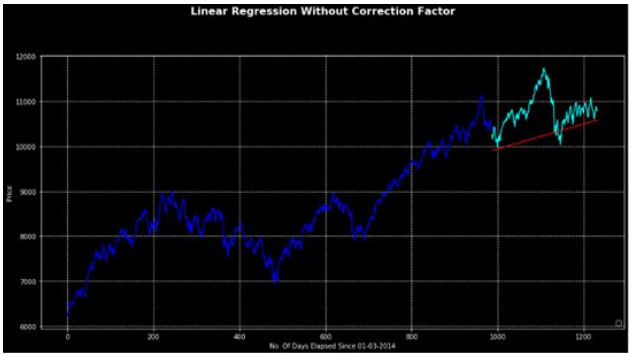

Figure 7(b): Linear Trend Regression Prediction not considering Monday-Friday Factor

Linear regression is a simple technique and quite easy to interpret, but there are a few obvious disadvantages. One problem in using regression algorithms is that the model overfits to the date and month column. Instead of taking into account the previous values from the point of prediction, the model will consider the value from the same date a month ago, or the same date/month a year ago. We have also considered a factor that market is most volatile on monday and friday, so our model charts a different value (peaks noticed in 7(a)). As seen from the plot, for January 2016 and January 2017, there was a drop in the stock price. The model has predicted the same for January 2018. After removing the effect, we notice a more “linear” looking pattern in fig 7(b)

C. **K Nearest Neighbor –** RMSE – 2859.252525533823

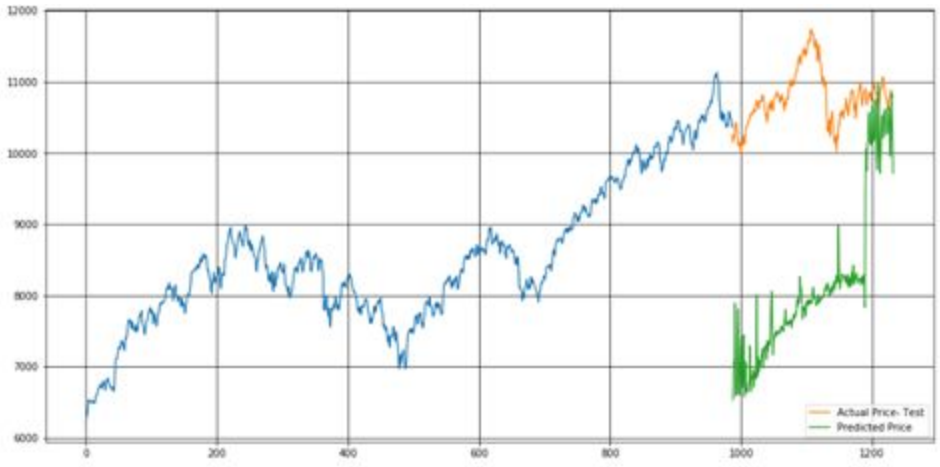

Figure 8: KNN Predictions (Index Price v/s Time)

The RMSE value is almost similar to the linear regression model and the plot shows the same pattern. For a very low value of k (suppose k=1), the model overfits on the training data, which leads to a high error rate on the validation set. On the other hand, for a high value of k, the model performs poorly on both train and validation set.Like linear regression, kNN also identified a drop in January 2018 since that has been the pattern for the past years. This algorithm works on forming clusters of data to find patterns. One likely reason for a very low predicted value when compared to actual value is that it is considering the data for first year ~(0-200) as the first cluster and predicted values according to it.

D. ARIMA – RMSE - 386.1390565770734

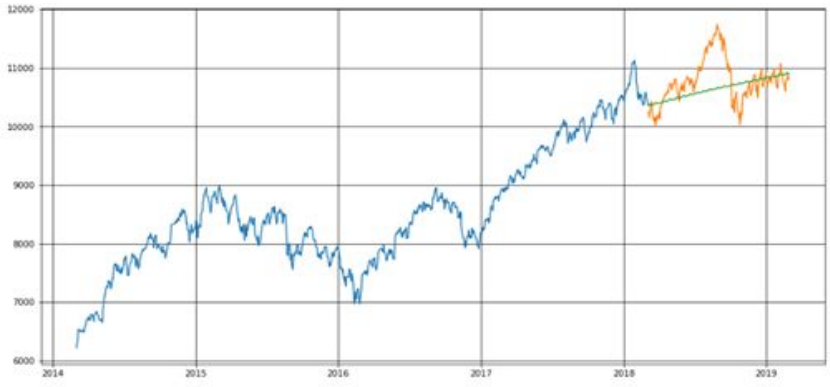

Figure 9 (a): Auto-ARIMA Predictions (Index Price v/s Time)

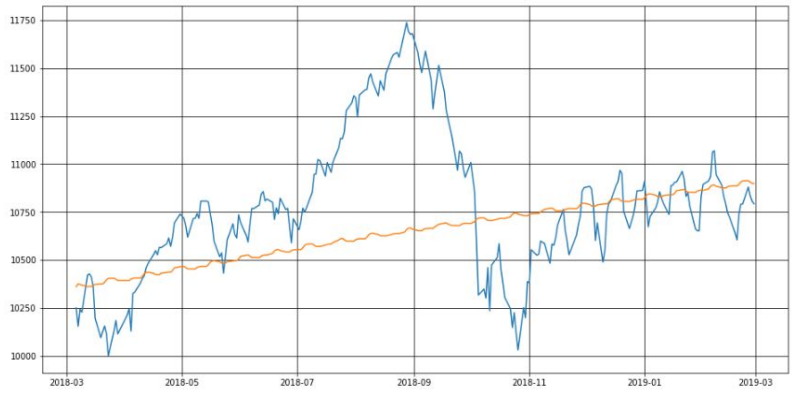

Fig 9(b): A closer look at predicted value using ARIMA to notice slight variations in prediction (Not a straight line)

An auto ARIMA model uses past data to understand the pattern in the time series. Using these values, the model captured an increasing trend in the series. Although the predictions using this technique are far better than that of the previously implemented machine learning models, these predictions are still not close to the real values. As is evident from the plot, the model has properly captured a trend in the series, but does not focus on the seasonal part. There is another variation of ARIMA available known as ***SARIMA***. *It can be used to account for seasonality of the data as well.*

E. LSTM – RMSE - 177.11003255955262

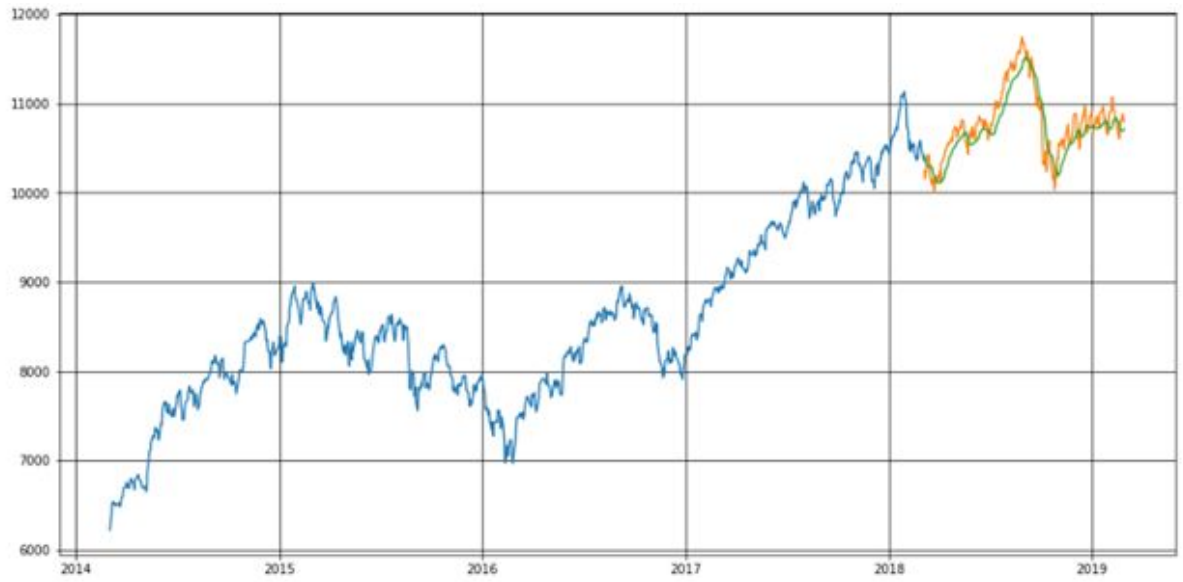

Figure 11: LSTM Prediction (Index Price v/s Time)

The LSTM Model performs exceptionally well in predicting the future prices of the stock market and from the comparative study conducted, emerges as the best prediction model for the share market prices. There is a small lag that can be observed in the output, but overall LSTM accounts for trend as well as seasonality from trend much quicker than the other models discussed.

State of Art: At its core, the stock market is a reflection of human emotions. Pure number crunching and analysis have their limitations; a possible extension of this stock prediction system would be to augment it with a *News feed analysis from social media platforms* such as Twitter, where emotions are gauged from the articles. This sentiment analysis can be linked with the LSTM to better train weights and further improve accuracy.

Note: For all the analysis done, we have written our code on Python using Jupyter Notebook. We used Pandas, Numpy, and Sklearn(for data normalization) for importing data and preprocessing. Keras was used for predictions using LSTM. (The different forecasting methods implemented can be found at: https://bit.ly/2DarwF1)

3. Results:

Models Comparison:

We have chosen *Root Mean Squared Error* as a basis for measuring the accuracy of different methods. The reason being, LSTM uses backpropagation algorithm to minimize the root mean square error and hence its accuracy measure uses RMSE.

We chose to use RMSE because they explicitly show the deviation of the prediction for continuous variables from the actual dataset. It measures the average magnitude of the error and ranges from 0 to infinity. The errors are squared and then they are averaged, RMSE gives a relatively high weight to large errors, and the errors in stock price prediction can be critical, so it is an appropriate metric to penalize the large errors. Therefore, to have a comparison between different stock market prediction techniques, we chose RMSE.

From the above results, we can see that LSTM prediction algorithm outperforms every other standard statistical methods based on the metric values. Amongst the standard models, Arima comes closest to giving a worthy prediction but it misses out on prediction of the seasonal movements of the market prices. Therefore, the performance of the models ranked from best to worst:

a. **LSTM –** RMSE - 152.11003255955262

b. **ARIMA –** RMSE - 386.1390565770734

c. **Linear Trend Regression** – RMSE - 564.3075943494691

d. **Moving Average Method** – RMSE - 685.72259439

e. **K Nearest Neighbor –** RMSE – 2859.252525533823

Estimating underlying Probability Distribution

1. **Anderson-Darling** tests the null hypothesis that a sample is drawn from a population that follows a particular distribution.
2. **Kolmogorov–Smirnov test (K–S test or KS test)** is a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test). Under the null hypothesis, the two distributions are identical, F(x)=G(x). The alternative hypothesis can be either 'two-sided' (default), 'less' or 'greater'. The KS test is only valid for continuous distributions.
3. **Shapiro–Wilk test** is a test of normality in statistics. It tests the null hypothesis that a sample came from a normally distributed population. Of all, this test has the maximum power to test normality.

Our results (for normality):

Anderson-Darling test: $P < 0.001$

Kolmogorov-Smirnov test: $P < 0.01$

Shapiro-Wilks test: $P < 0.001$

All of these results indicate that our data doesn't show normal distribution assuming confidence interval of 95% (at least not from the data for one year). (This testing was done on https://contchart.com/goodness-of-fit.aspx as well as other online available statistical tools)

D'Agostino's K-squared test, Jarque–Bera test, Lilliefors test are other goodness of fit tests to check the normality of the underlying population.

Assumptions used for various methods:

1. **Moving Average:** The time series is locally stationary with a slowly varying mean.
2. **Linear Trend Forecasting:** The model assumes that the data follows a linear model.
3. **K-nearest Neighbour:** It assumes data is in a metric space i.e we are working with scalars or multidimensional vectors
4. **ARIMA:** It considers that the data is stationary. If the data is not, it makes it stationary through differencing.
5. **LSTM:**No assumptions are taken here, making it the best model.