**NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY**

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM, APPROVED BY AICTE & GOVT.OF KARNATAKA)

****

**REPORT**

On

**LA 2 Project**

**Stock Market Prediction**

*Submitted in partial fulfilment of the requirement for the award of Degree of*

*Bachelor of Engineering*

*in*

*Computer Science and Engineering*

*Submitted by:*

|  |  |
| --- | --- |
| A ANKITH SAVIO  ABHISHEK CM  ADHRIT PAUL | 1NT19CS001  1NT19CS009  1NT19CS011 |
|  |  |

Under the Guidance of

Dr. Vani V

PROFESSOR, Dept. of CS&E, NMIT



Department of Computer Science and Engineering

**(Accredited by NBA Tier-1)**

2022-23

**NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY**

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM

, APPROVED BY AICTE & GOVT.OF KARNATAKA)

Department of Computer Science and Engineering

**(Accredited by NBA Tier-1)**

****

**CERTIFICATE**

This is to certify that the Report on **Stock Market Prediction** is an authentic work carried out by A ANKITH SAVIO (1NT19CS001),ABHISHEK CM (1NT19CS009) and ADHRITH PAUL (1NT19CS011) bonafide students of **Nitte Meenakshi Institute of Technology**, Bangalore in partial fulfilment for the award of the degree of **Bachelor of Engineering** in **COMPUTER SCIENCE AND ENGINEERING** of Visvesvaraya Technological University, Belagavi during the academic year **2021-2022*.*** It is certified that all corrections and suggestions indicated during the internal assessment have been incorporated in the report.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Internal Guide** | | **Signature of the HOD** | | **Signature of Principal** | |
|  | |  | |  | |
| Dr. Vani V  Professor, Dept. CSE,  NMIT Bangalore | | Dr Nalini N  Professor&Head,  Dept.CSE, NMIT Bangalore | | Dr. H. C. Nagaraj  Principal,  NMIT, Bangalore | |
|  | **Signature of Examiners** | | | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |

**ACKNOWLEDGEMENT**

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crowned our effort with success. I express my sincere gratitude to our Principal **Dr. H. C. Nagaraj**, Nitte Meenakshi Institute of Technology for providing facilities.

We wish to thank our HoD**, Dr.Nalini N** for the excellent environment created to further educational growth in our college. We also thank him for the invaluable guidance provided which has helped in the creation of a better project.

I hereby like to thank our Dr. Vani V**, Professor**, Department of Computer Science & Engineering on her periodic inspection, time to time evaluation of the project and help to bring the project to the present form.

Thanks to our Departmental Project coordinators. We also thank all our friends, teaching and non-teaching staff at NMIT, Bangalore, for all the direct and indirect help provided in the completion of the project.

|  |  |  |
| --- | --- | --- |
| **NAME** | **USN** | **Signature** |
| A ANKITH SAVIO | 1NT19CS001 |  |
| ABHISHEK CM | 1NT19CS009 |  |
| ADHRITH PAUL | 1NT19CS011 |  |

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| Sl No. | Title | PageNo. |
| 1 | Abstract | 3 |
| 2 | Introduction | 4 |
|  | 2.1 Motivation |  |
|  | 2.2 Problem Domain |  |
|  | 2.3 Aim and Objectives |  |
| 3 | Data Source and Data Quality | 5 |
| 4 | Data Pre-processing |  |
| 5 | Machine learning methods | 6 |
| 6 | Results and Discussions |  |
| 7 | Conclusion & Future Directions | 8 |
| 8 | Lessons Learnt | 9 |
| 9 | References |  |
| 10 | Appendix |  |
|  | 10.1 Python Codes Implemented |  |
|  | 10.2 Setup |  |

**ABSTRACT**

The stock market plays a pivotal role in the financial aspect of the nation's growth, but the stock market is highly volatile and complex in nature. It is affected by significant political issues, analyst calls, news articles, the company's future plans for expansions and growth, and many more. Hence, any investor would be interested in understanding the stock market over time and how the factors mentioned above affect the behavior of the stock market. Every business day, millions of traders invest in the stock market. Most of these investors lose money and others gain. However, considering any trading day, loss or gain is inconsistent. The demand to predict stock prices are extremely high hence is a need for stock market analysis.

In this project, we conducted a detailed analysis of the stock market, with a focus on predicting future stock prices using machine learning techniques. We analyzed the stock market data for Infosys, a publicly traded company listed on the National Stock Exchange of India, using four different machine learning algorithms: Random Forest, support vector regressor, artificial neural networks, and Long Short-Term Memory (LSTM). We compared the performance of these algorithms and found that the LSTM model was the most effective for stock market analysis.

The results of this study demonstrate the importance of stock market analysis for making informed investment decisions. By understanding the trends and patterns in the market and identifying key factors that influence stock performance, investors can make more informed decisions about their investments and potentially maximize returns while minimizing risks. The use of machine learning techniques, such as LSTMs, can also be a valuable tool for analyzing the stock market and predicting future performance.

**INTRODUCTION**

Stock market forecasting and analysis are among the most difficult tasks. There are several reasons for this, including market volatility and a variety of other dependent and independent variables that influence the market value of a specific stock. These variables make it extremely difficult for any stock market expert to predict the market's rise and fall with great precision. However, with the advent of Machine Learning and its powerful algorithms, recent market research and Stock Market Prediction advancements have begun to incorporate such approaches in analyzing stock market data. To summarize, many organizations use Machine Learning Algorithms to predict the stock market. This article will walk you through a basic analysis and forecasting implementation. Since forecast prediction will be a regression problem to predict the future stock trend based on past data, we will use models such as Random Forest Regressor, Support Vector Regressor, Artificial Neural Networks (ANN), and Long-short term memory ANN.

**2.1 MOTIVATION**

The stock market is a key component of any modern economy, and understanding its movements and trends is essential for both individual investors and businesses. By conducting a detailed analysis of the stock market, we can gain valuable insights into market behavior and make informed decisions about investments. Additionally, studying the stock market can help us to identify patterns and trends that may be indicative of future market performance. Overall, the motivation for this project is to increase our understanding of the stock market and to use this knowledge to make more informed investment decisions.

**2.2 PROBLEM DOMAIN**

1. Data availability and quality: The stock market generates a vast amount of data on a daily basis, and accessing and processing this data can be a challenge. Additionally, the quality of the data can vary, and ensuring that the data is accurate and reliable is essential for effective analysis.
2. Complexity of the market: The stock market is a complex system with many interconnected variables and factors that can influence its performance. Understanding these relationships and how they impact the market can be difficult.
3. Volatility and unpredictability: The stock market is often volatile and can be difficult to predict, making it challenging to make informed investment decisions.
4. Competition and information asymmetry: The stock market is highly competitive, and investors must constantly stay up-to-date with the latest market information in order to make informed decisions. However, information asymmetry, where some investors have access to more or better information than others, can create challenges for fair and efficient market functioning.
   1. **AIMS AND OBJECTIVES**
5. To understand the trends and patterns in the stock market and use this knowledge to make informed investment decisions.
6. To develop and test investment strategies that aim to maximize returns and minimize risks.
7. To forecast future market performance and identify potential opportunities and risks.
8. To evaluate the prediction among different learning algorithms.
9. To compare different algorithms for the problem.

**DATA SOURCE AND DATA QUALITY**

The dataset for this project is the stock market data for Infosys, a publicly traded company listed on the National Stock Exchange of India. The dataset has been sourced from Kaggle, a popular online platform for data science and machine learning resources.

The quality of the dataset is likely to be good, as Kaggle is known for hosting high-quality datasets and requiring contributors to follow certain guidelines to ensure the accuracy and reliability of the data they submit. However, it is always advisable to assess the quality of the dataset and ensure that it meets the requirements of the project before proceeding with analysis.

In this project, we are using the last two years of data for Infosys. This should provide a sufficient amount of data to analyze trends and patterns in the stock market and make informed decisions about investments. It is important to ensure that the dataset is complete and up-to-date, as missing or outdated data can impact the accuracy and reliability of the analysis.

The dataset is the INFY stocks dataset that contains seven records, including date, opening value of the stock, closing value of the stock, high and low values of stock that day, the adjacent closing value of the stock, and the volume of stocks that had been traded that day.

The stock market is all about numbers. Understanding these numbers and finding the pattern is an art, and we the data analysts are the artists. The main variable we are going to focus on in this dataset is the ‘Adj Close’, and it will be our target Variable that we want to determine using learning algorithms.

**DATA PRE-PROCESSING**

**1. Data Cleaning:**

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc. The methods we have taken in this project are as follows:

* Missing Data: This situation arises when some data is missing in the data
* Ignore the tuples: This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
* Drop rows or data entries with null values.

**2. Data Transformation:**

This step is taken in order to transform the data in appropriate forms suitable for feeding it into the learning algorithms. The methods we have taken are as follows:

* Normalization: It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0). We have used Min Max Scaling and optional Standard Scaling.
* Time Series Split: In time series analysis, we cannot use standard split since observations in our time series datasets are not independent.

**MACHINE LEARNING METHODS**

**Random Forest Regressor:**

Random Forest is a machine learning algorithm that is used for regression and classification tasks. It is a type of ensemble learning method, where a group of weak models are combined to create a powerful model.

In a Random Forest Regressor, a group of decision tree regressors are trained on different sub-samples of the dataset. Each decision tree makes a prediction, and the final prediction is made by averaging the predictions of all the decision trees. This can help to reduce overfitting, as the model is trained on different sub-samples of the data and the final prediction is made by averaging the predictions of all the decision trees, rather than relying on a single decision tree.

Random Forest Regressors are simple to use and can provide good results on many types of problems. They are a good choice for regression tasks when you have a large number of features and you want to reduce the risk of overfitting.

**SVR:**

Support Vector Regression (SVR) is a type of support vector machine (SVM) that is used for regression tasks. It is a supervised learning algorithm that can be used to predict a continuous-valued output (such as a real number) based on a set of input features.

SVR works by finding the hyperplane in a high-dimensional space that maximally separates the data points of different classes. In the case of SVR, the data points are not divided into different classes, but rather are fit to a continuous function. The distance between the data points and the hyperplane is called the margin. The goal of SVR is to find the hyperplane that has the maximum margin, while still allowing some of the data points to be on the wrong side of the margin, known as the "slack variables."

SVR is a powerful algorithm that can handle high-dimensional data and complex non-linear relationships between input features and output values. However, it can be sensitive to the choice of kernel and parameters, and it may not perform well on noisy or sparse data.

**Artificial Neural Networks:**

An Artificial Neural Network (ANN) is a computational model inspired by the structure and function of the brain. It is composed of a large number of interconnected processing units called artificial neurons, which are inspired by the structure and function of neurons in the brain.

ANNs are trained to recognize patterns in data, such as in images, audio, and text, by adjusting the strengths of the connections between the neurons based on input data and expected output. This process is known as training the network.

Once trained, an ANN can use the patterns it has learned to make predictions or decisions. For example, an ANN can be a good model for regression making it a good candidate for our project.

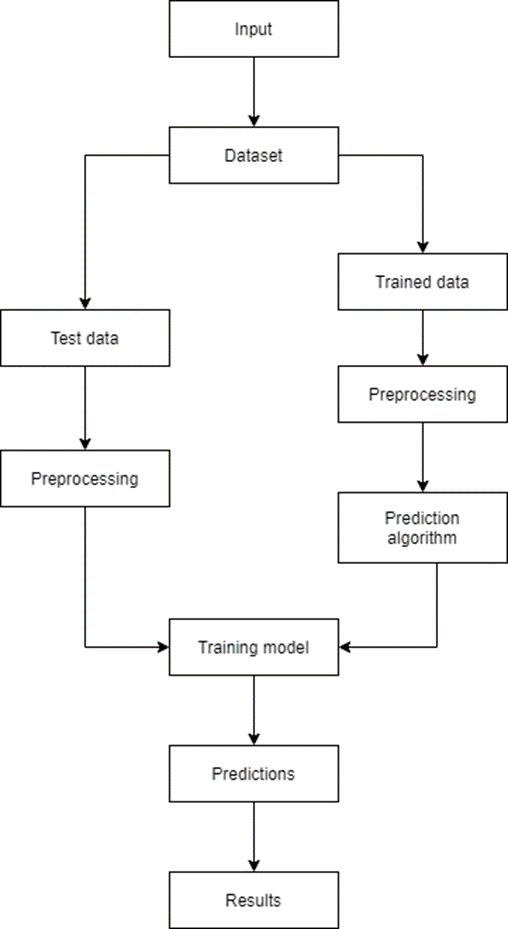
**LSTM:**

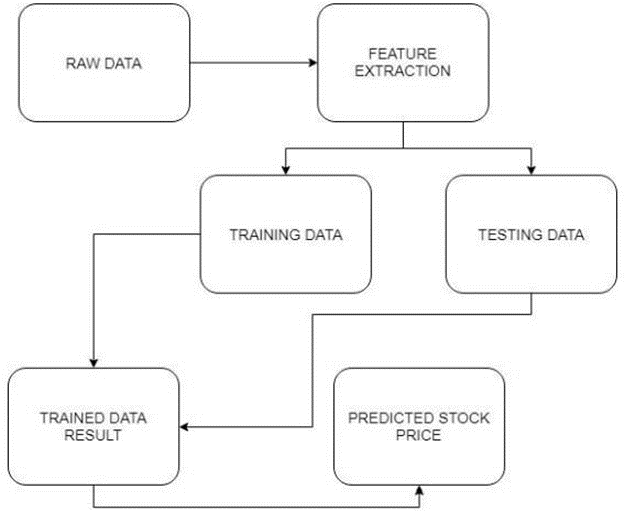
Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that is capable of learning long-term dependencies in data. It is a type of artificial neural network designed to process sequences of data, such as time series, natural language text, or audio.

LSTMs are a type of RNN that are particularly effective at learning long-term dependencies in data. They do this by using a special type of memory cell that can store information for long periods of time and can selectively retain or forget information as needed. This allows LSTMs to learn patterns in data that span long sequences of time steps.

LSTMs have been used in a wide range of applications, including natural language processing, speech recognition, and machine translation. They are a popular choice for tasks that involve sequential data and require the model to learn and retain long-term dependencies like stock market analysis and making this the best candidate for our project.

**RESULTS AND DISCUSSIONS**

**Architectural Design:**

**Data Flow Diagram:**

**Metrics for assessment:**

**Mean Absolute Error (MAE):**

MAE is a measure of errors between paired observations expressing the same phenomenon. Examples of Y versus X include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement versus an alternative technique of measurement. MAE is calculated as the sum of absolute errors divided by the sample size

**Mean Squared Error (MSE):**

MSE or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the error that is, the average squared difference between the estimated values and the actual value. MSE is a risk function, corresponding to the expected value of the squared error loss.

**Root Mean Squared Error (RMSE):**

Square root of Mean Squared Error

**Mean Absolute Percentage Error (MAPE):**

The mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD), is a measure of the prediction accuracy of a forecasting method in statistics. It usually expresses the accuracy as a ratio.

**Assessment**

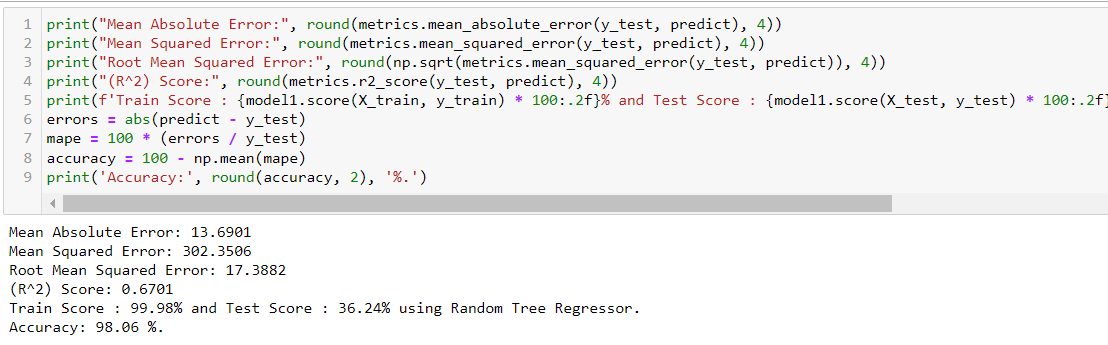
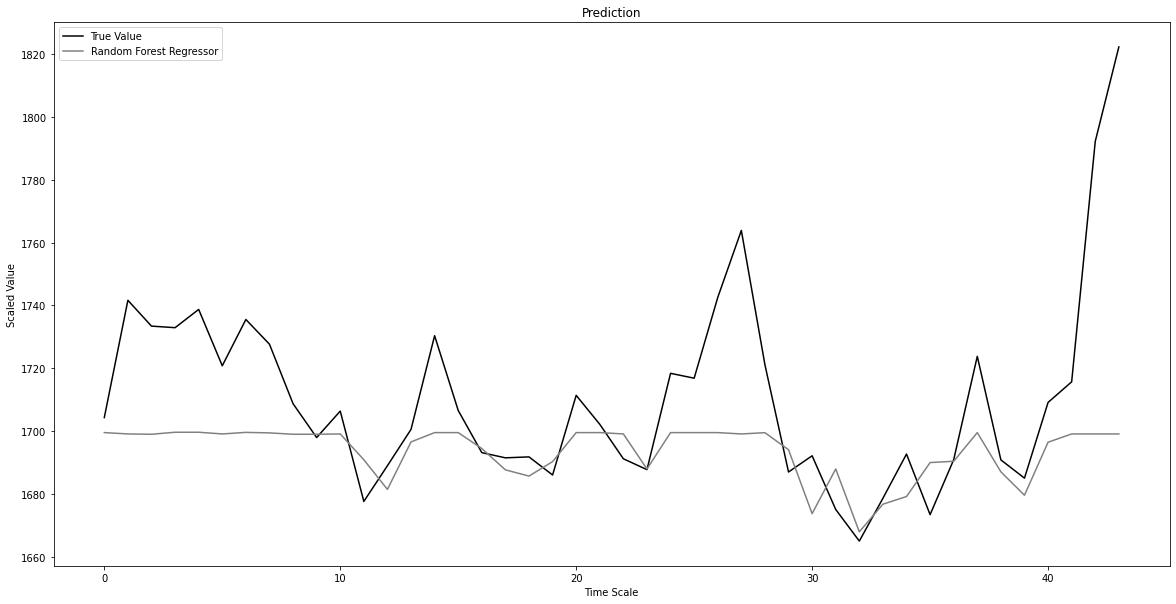
****

Figure Random Forest Regressor



Above we can see the line graph of True Test data and the predicted Test data of our dataset. The random forest regressor performs poorly for our problem getting a very low test score.

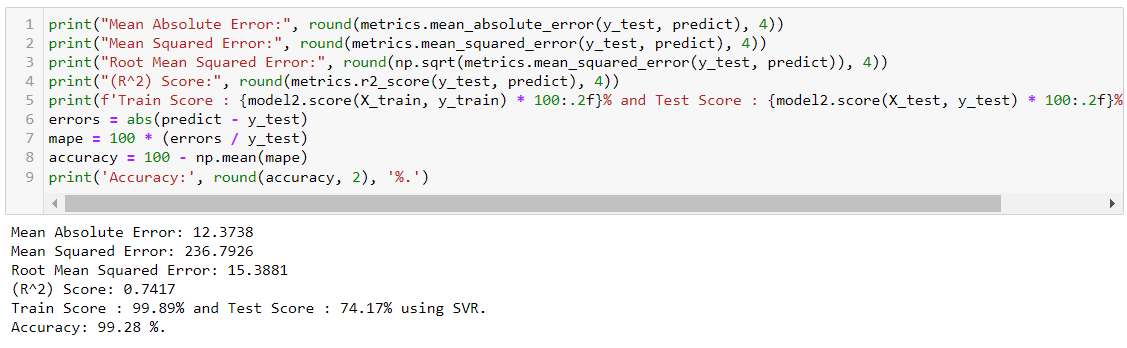
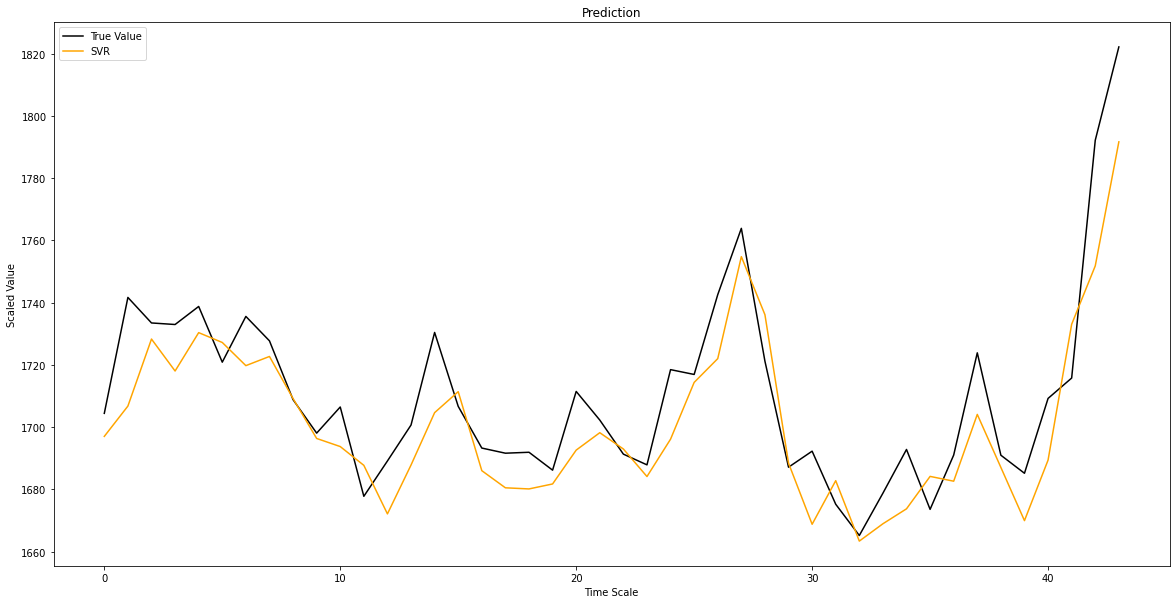
****

Figure SVR



Similarly, here we have prediction of SVR model plotted against True Test data, and performs pretty well coming second in our comparisons against other models.

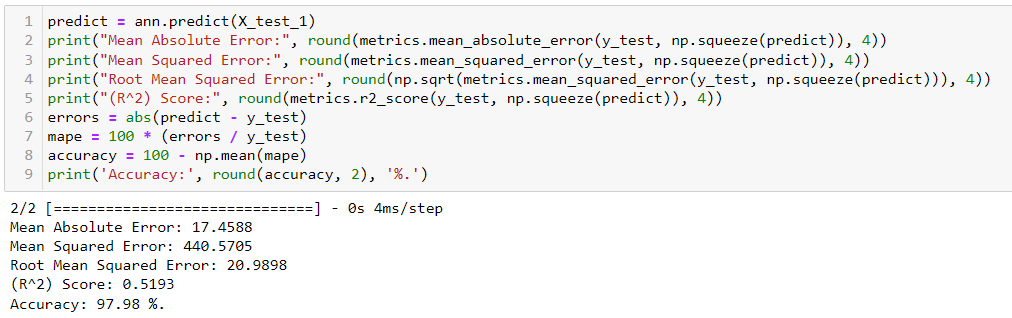
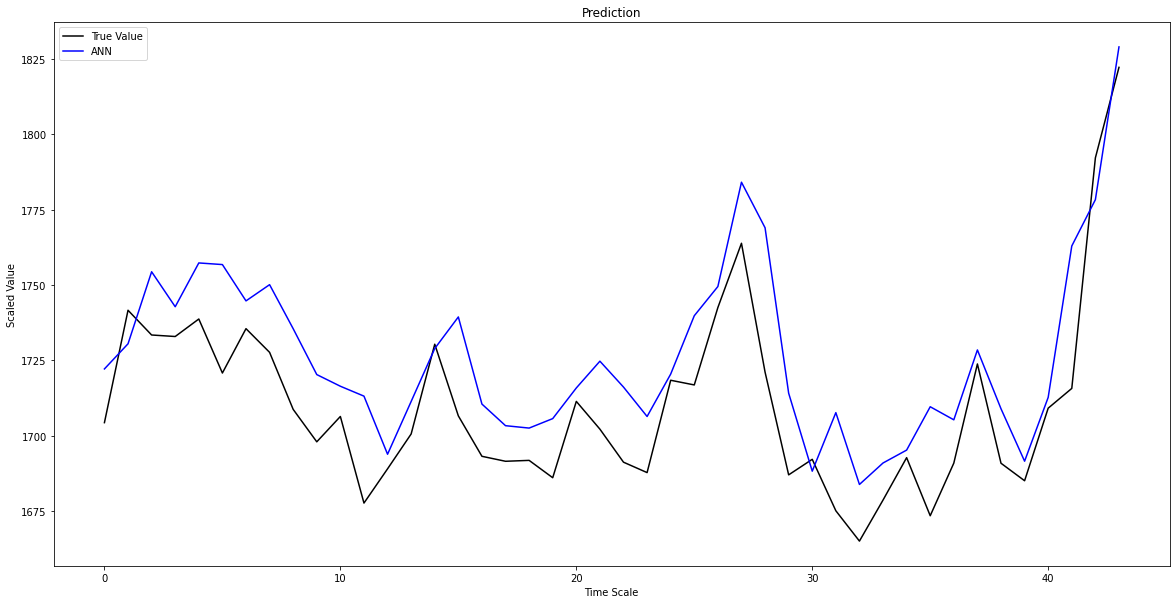
****

Figure Neural Networks



Here we have the line graph for ANN model, which performs quite well but still not as good as SVR model, and being way more expensive.

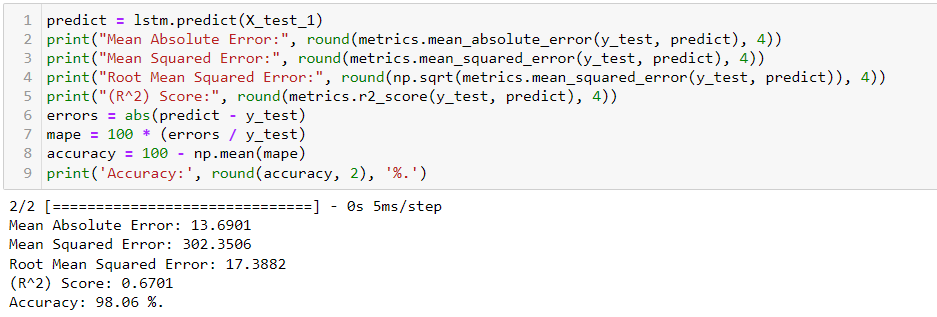
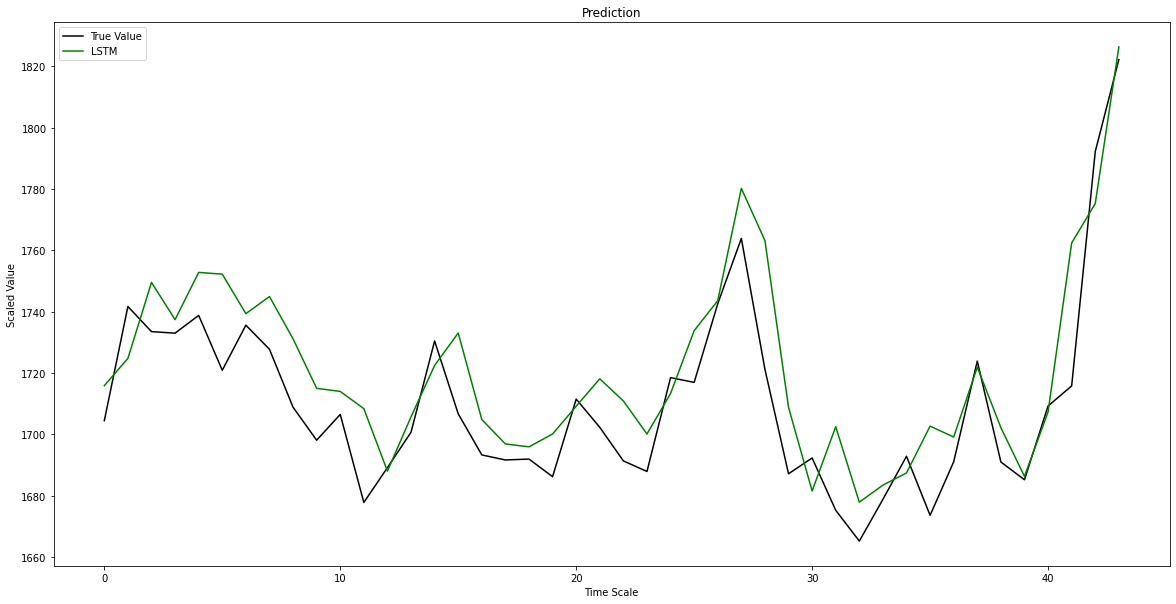
****

Figure Long Short Term Memory

****

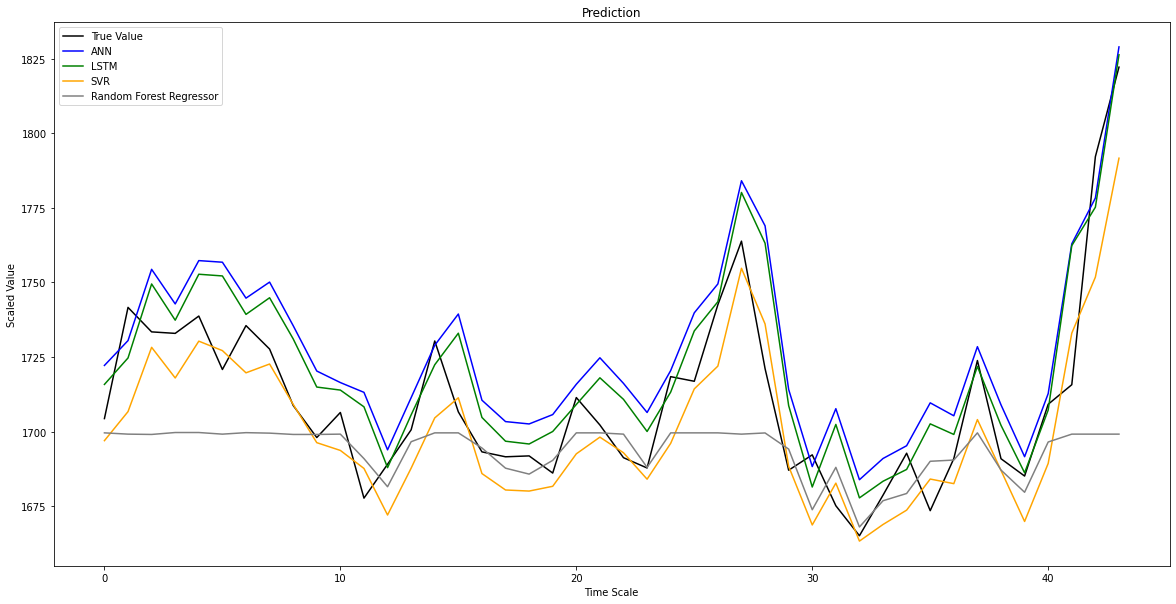
****LSTM model as we can see performs the best from the above prediction graph.

Figure 5Comparison among different models

**CONCLUSION & FUTURE DIRECTIONS**

Based on the results of the analysis, it appears that the Long Short-Term Memory (LSTM) model is the most effective for stock market analysis in this project. LSTMs are a type of recurrent neural network that are well-suited for analyzing time series data, such as stock market data, and have been shown to perform well in a variety of contexts.

Overall, the use of LSTMs in this project demonstrates the potential of machine learning techniques for improving our understanding of the stock market and making more informed investment decisions.

As a future direction for this project, it would be interesting to explore other machine learning techniques and compare their performance to that of LSTMs such as using Convolution 1D or the combination of LSTM + GRU. Additionally, it would be valuable to continue refining and improving the LSTM model through techniques such as hyperparameter tuning and feature engineering. Finally, it would be interesting to extend the analysis to other companies or sectors within the stock market to further validate the findings of this study.

**LESSON LEARNT**

* The importance of thoroughly cleaning and pre-processing the data before conducting analysis. Dirty or inconsistent data can lead to incorrect conclusions and poor decision-making.
* The effectiveness of machine learning techniques, such as LSTMs, for analyzing time series data and making predictions about the stock market.
* The importance of considering a range of factors that may influence the performance of the stock market, including economic indicators, company financial performance, and market sentiment.
* The value of regularly updating and re-evaluating the analysis to ensure that it remains relevant and accurate.
* The need to be aware of the limitations of any analysis and to carefully consider the risks and uncertainties associated with investment decisions.

**REFERENCES**

1. R. Ahuja, H. Rastogi, A. Choudhuri and B. Garg, "Stock market forecast using sentiment analysis," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 2015.
2. I. Parmar et al., "Stock Market Prediction Using Machine Learning," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), 20
3. S. Kalra and J. S. Prasad, "Efficacy of News Sentiment for Stock Market Prediction," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019.
4. M. K. Sophan, M. N. Fitriantama and S. Wahyuni, "Collaborative Filtering Item Recommendation Methods based on Matrix Factorization and Clustering Approaches," 2020 10th Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS), 2020.
5. Patel, Yagnesh G. and Vishal P. Patel. “A Survey on Various Techniques of Recommendation System in Web Mining.” (2015).
6. N. Sharma and A. Juneja, "Combining of random forest estimates using LSboost for stock market index prediction," 2017 2nd International Conference for Convergence in Technology (I2CT), 2017.
7. Sachin Gavhane, Kartik Sharma, Akilesh Rao, Rushabh Shah, Sujay Sail, Stock Market Analysis, International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887 Volume 6 Issue V, May 2018.
8. B. Panwar, G. Dhuriya, P. Johri, S. Singh Yadav and N. Gaur, "Stock Market Prediction Using Linear Regression and SVM," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2021.
9. Q. Yunneng, "A new stock price prediction model based on improved KNN," 2020 7th International Conference on Information Science and Control Engineering (ICISCE), 2020.

**APPENDIX**

Dataset: <https://www.kaggle.com/datasets/syedjaferk/infosys-stock-data-live-and-latest>

**10.1 PYTHON CODE IMPLEMENTED**

Code:

**10.2 SETUP**