# Stock Market Analysis and Stock Price Prediction using Machine Learning

Meet Popat (2020320)
meet20320@iiitd.ac.in
Pallav Singla (2020225)
pallav20225@iiitd.ac.in
Dux Pal Singh (2020297)
dux20297@iiitd.ac.in
Shourya Jindal (2020336)
shourya20336@iiitd.ac.in

## A. Abstract

Due to the volatility of the market caused by the COVID-19 pandemic and the increasing interest in financial markets among the general population, the stock market has become a popular topic of discussion. This has motivated us to conduct an analysis and leverage the power of data analysis and machine learning to gain insights into stock market trends. Our objective is to develop predictive models that can assist investors in making informed decisions. In this report, we present a comprehensive analysis of stock market data and stock price prediction using machine learning techniques.

### A.1. Motivation

The stock market is a crucial financial system that plays a significant role in the global economy. However, understanding the stock market and predicting stock prices can be challenging for the non-experts. To address this issue, we offer an optimal analysis method. Accurate stock price prediction is essential to investors, financial analysts, and policy-makers. In recent years, machine learning techniques have gained popularity for their ability to analyze vast amounts of historical stock market data and make predictions. This enables lenders to optimize lending decisions and leads to sound business economics experience. Our aim is to provide valuable insights to investors and financial professionals, enabling them to make more informed decisions.

### A.2. Introduction

Over the past few years, there has been a significant increase in the use of data-driven approaches and machine learning techniques to analyze financial markets, gain insights into their behavior, and develop predictive models for stock price movements. Our approach to this field is comprehensive, starting with the foundational steps of exploratory data analysis (EDA) and various model to understand and model stock price dynamics. Our study aims to:

- Utilize data-driven approaches to interpret stock market trends and behavior.
- Create predictive models that enable investors to make informed decisions about stock trading and investment.
- Research the capabilities of machine learning algorithms in accurately predicting stock prices.

This report will provide an overview of our modeling technique efforts, we discussed the application of various models and explored more advanced machine learning methods for stock price prediction. Our goal is to provide valuable insights into the complex world of stock market analysis and equip investors with the necessary tools to navigate the ever-changing landscape of financial markets.

## B. Literature Survey

”Stock Market Prediction Using Linear Regression and SVM”.

This paper discusses the application of machine learning algorithms, specifically Linear Regression and Support Vector Machine (SVM), for predicting stock market prices. The study involves data collection, feature engineering, model implementation, and performance evaluation. The paper emphasizes the dynamic nature of stock market data and its dependency on various factors. It also highlights the importance of having properly labeled training data for accurate predictions. Linear Regression Model outperformed SVM in predicting Amazon stock data, achieving an accuracy rate of 0.9876 compared to SVM’s 0.9432. The conclusion drawn from this analysis is that Linear Regression is a superior choice for stock market analysis when compared to SVM. Thus, the paper explores the use of machine learning algorithms in predicting stock prices, provides a comparison between Linear Regression and SVM, and concludes that Linear Regression is the better option for this specific application. [1]

”Stock Market Prediction Using Machine Learning”.
This research paper investigates the use of machine learning algorithms(Regression model and LSTM model) to forecast stock market trends. The dataset was taken from Yahoo Finance. They analyzed historical market data, applied diverse machine learning techniques, and evaluated their effectiveness. Regression Based Model has a confidence score of 0.86625. LSTM Based Model results in a Train Score of 0.00106 MSE (0.03 RMSE) and a Test Score of 0.00875 MSE (0.09 RMSE), indicating its accuracy. The conclusion highlights that both models show improvements in prediction accuracy, with the LSTM model outperforming the Regression model. The paper suggests that using a larger dataset and exploring other machine learning models could further enhance prediction accuracy. Additionally, sentiment analysis on news impact

and other deep learning models are potential areas for future research in stock market prediction. [2]

## C. DataSet

We have 2 datasets, indexInfo and indexProcessed both of them have been taken from kaggle.

Link to datasets:
indexInfo
indexProcessed

### C.1. IndexProcessed Dataset

#### C.1.1 DataSet Description

The data consists of Date, Open, High, Low, Close, Adj-Close, Volume and CloseUSD, this data set is our main data set combined with other dataset for increasing the features. It has all the values related to indexes in the stock market.

#### C.1.2 DataSet Extraction

The dataset is publicly available on the Kaggle website. We have used all the features of the data initially and all the available data, afterwards after computation we took data corresponding to 4 indexes.

#### C.1.3 Preprocessing

There are few null values in the dataset which we replaced with 0 and String components are labeled using label encoding and Date is changed to integer using python library.

#### C.1.4 Data Visualisation

This plot tells about the imbalanced data distribution in the indexProcessed dataset. The pie chart shows how the dataset gets distributed among different indexes so according to the given dataset we will evaluate four indexes that have more data according to the pie chart given above and in all others there are chances of underfit.

## D. Merged DataSet(indexInfo and indexProcessed

### D.1. DataSet Description

The dataset contains all the information present in the indexProcessed and the corresponding data information related to indexes present in the indexInfo this we used mainly for dataset features increment to get more idea how my dataset will work in more features. The Informative features include Index, Open, Close, Low, High, AdjClose, Volume, CloseUSD.

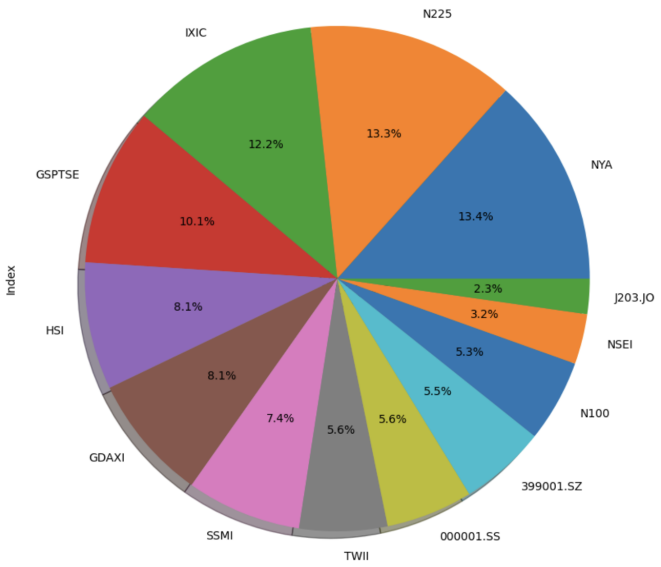


Figure 1. Pie chart showing data-set distribution between indexes

### D.2. DataSet Visualisation

We have made the Closing price vs Adj close graph for two indexes just to show the relation and the Difference between the Adj Close and Close value using matplotlib library, showing only two indexes here because of space constraints.
Difference between AdjClose and Close in Stock market:
- The adjusted closing price is the closing price adjusted to account for factors that can impact a stock's price but are unrelated to its fundamental performance.
-Adjustments include dividends, stock splits, reverse splits, and other corporate actions.
-The purpose is to provide a consistent view of a stock's value over time, considering these events.
This we did so that we can get the clear idea how much close differ with AdjClose so as to which to take while predicting.

We here showed the mean and standard deviation of each index for volume by this we are getting which stock we majorly getting traded interms of the volume this we did so as to work on four major stocks for prediction of their values.

We plotted the frequency of all the features using histogram plots of the matplot library and got various plots. All our plots are plotted using matplotlib and seaborn library.
Few of them are showed here:

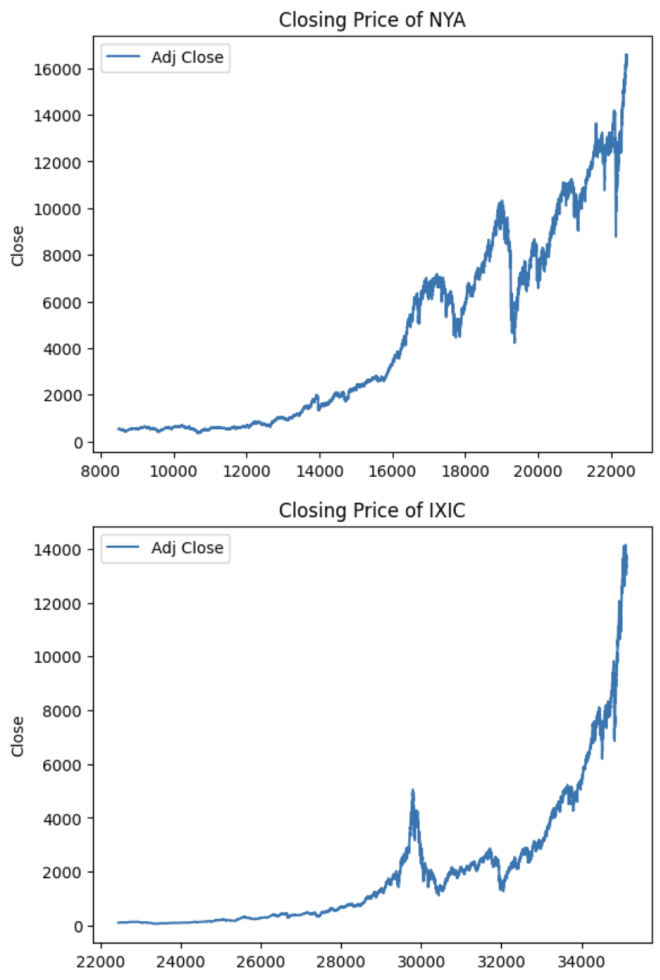


Figure 2. Showing the closing price of a stock

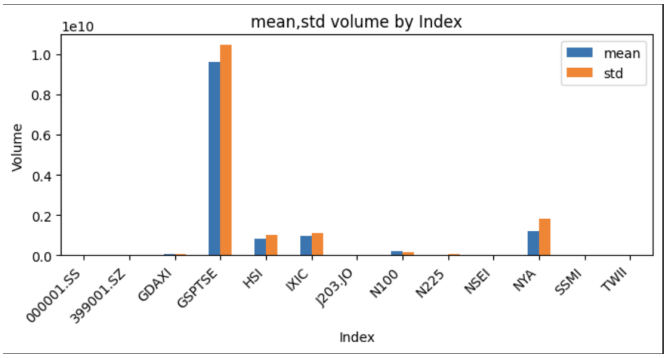


Figure 3. Mean & Standard Deviation volume by index

Then we did the Simple Moving average comparison of four major indexes with Close values, this technique is very much used in the actual stock market trading to predict the behaviour of the index for certain time. Here

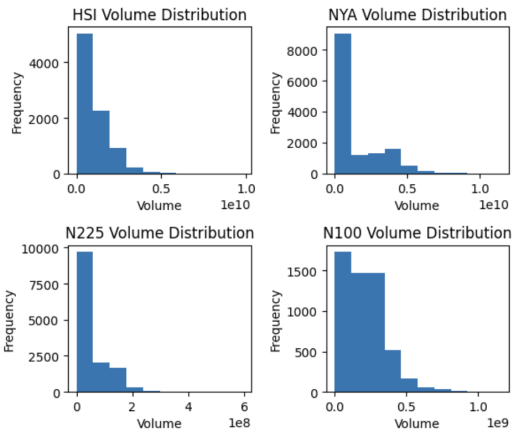


Figure 4. Frequency V/S Volume graph

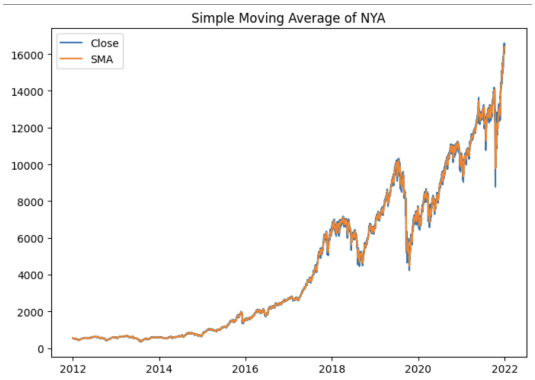


Figure 5. Comparing simple moving average and close value for easy visualisation

we took ten years and yearwise value can be seen in the plot.

Here we predicted the difference between Day high and low so as check whether their are outlier present in the high and low term or not.

## D.3. Feature Selection

We did some feature extraction as closing, opening, low, high are very much interrelated so looking at the initial heatmap we decided that in stock market role of difference will be crucial instead of individual so we made few columns with differences of the correlated features.

### D.3.1 Correlation Coefficient

It is a measure of the linear relationship between two or more variables. It helps in predicting a variable based on the value of another variable. It helps in deciding the features which are largely correlated with each other and

can be dropped after determining their correlation with the target variable and therefore help in feature selection.

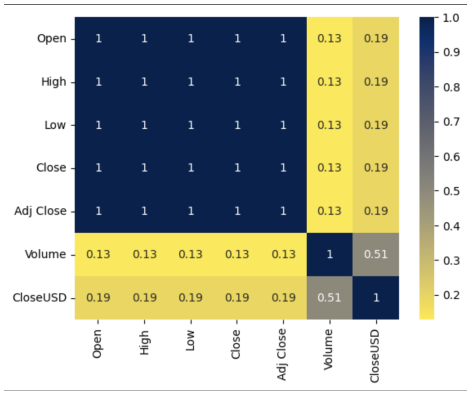


Figure 6. Heat map when we did not extract the data

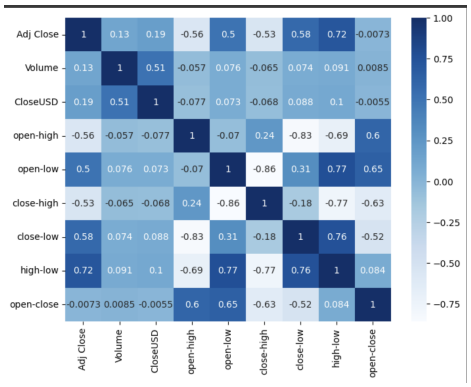


Figure 7. Heat map when we extracted the data and pre-processed it

## E. Methodology

The main objective of stock market analysis is to accurately predict the closing price of stock market indexes. This prediction task is framed as a regression problem, wherein the goal is to forecast a numerical value, in this case, the closing price. Various regression techniques are employed to achieve this objective.

We did Exploratory Data Analysis (EDA) to explore various options for gaining a better understanding of the dataset. Afterward, we preprocessed the data to further refine it.

To define the labels on the dataset, we started with date as an input feature and the closed USD on that date as a target variable. We trained different models to predict the closed USD for future as well. Multiple regression models like, Linear regression, Lasso regression, Ridge regression, Polynomial regression, Random Forest Regressor**(RFR)**, Long short-term memory **(LSTM)** model and Multi-Layer Perceptron **(MLP)**. We also used one unsupervised learning algorithm: K-Means to classify the data.

Further, we performed **GRID SEARCH** on the models: Lasso regression, Ridge regression, Random Forest Regressor(RFR), and Multi-Layer Perceptron (MLP), to do hyperparameter tuning.
The following hyperparameters were tuned:
**Lasso and Ridge Regression:** (Learning Rate, max iterations)
**Random Forest Regressor(RFR):** (max depth, n estimators)
**Multi-Layer Perceptron (MLP):** (hidden layer sizes, max iterations)

We used the following metrics to evaluate the performances of our models:

1. **R-squared:** It measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It provides a measure of how well the model's predictions fit the actual data.
2. **MAE (Mean Absolute Error)**: This is the average of the absolute differences between the predicted and actual values. The measure gives an equal weight to all big or small errors.
3. **RMSE (Root Mean Squared Error)**: This is the square root of the average of the squared differences between the predicted and actual values.

## F. Result and Analysis

### F.1. Linear Regression

We applied linear regression model on all the 13 stock indices, found its evaluation metrics and plotted the graph of the regressor achieved. Graph of index 'NYA' has been shown in Figure 8. The R2 score varied from 0.18 to 0.89 for different indices. This is because linear regression does not capture the sudden change in the stock market prices

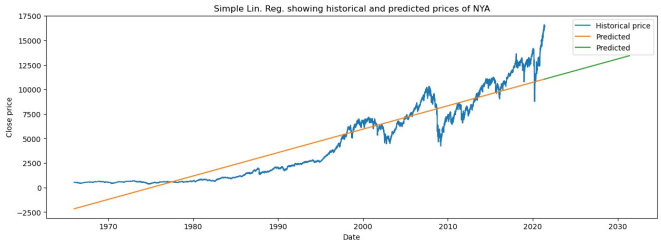


Figure 8. Accuracy score for NYA index 0.8747

## F.2. Polynomial Regression

Then we applied Polynomial Regression model on all the 13 stock indices, found its evaluation metrics and plotted the graph of the regressor achieved. Graph of index 'NYA' has been shown in Figure 9. The R2 score varied from 0.44 to 0.95 for different indices. Polynomial regression tries to solve the problem of underfitting and has a much improved accuracy than linear regression.

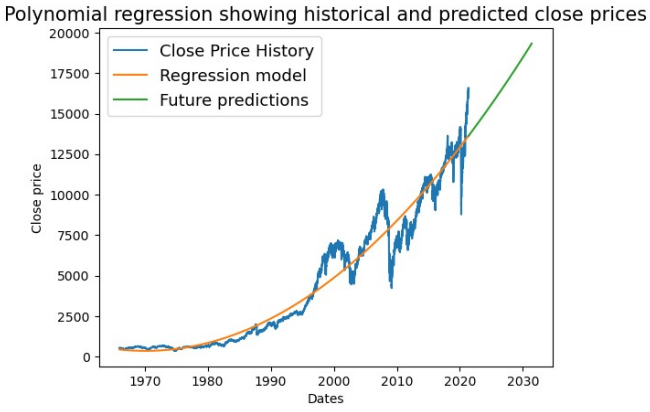


Figure 9. Accuracy score for NYA index 0.9551

## F.3. Ridge Regression

Then we applied, ridge regression and used grid search on the hyperparameters: (learning rate, max iterations). After grid search we found out that best learning rate is very close to zero (1e-05). Thus, ridge regression becomes exactly same as linear regression and thereby, producing almost similar results and accuracy. (Refer Figure 10)

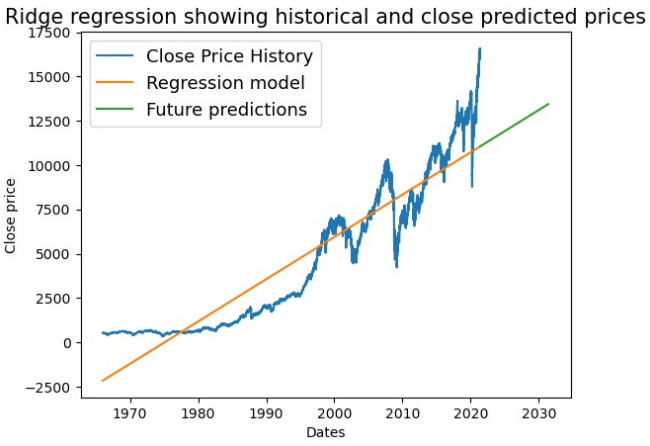


Figure 10. Accuracy score for NYA index 0.8747

## F.4. Random forest Regressor

Random forest is ensemble method that combines multiple decision trees to create a more robust model. We also performed grid search on this to fine tune the hyperparameters : (max Depth, n estimators). Accuracy achieved from this model were very high. R2 score was almost 1.00 for most of the indices (Refer Figure 11). This was due to the fact random forests is very robust outliers and non-linear data, which was a case in our data. So, to avoid this type overfitting we can applied the technique of **'Pruning'**. After pruning accuracy achieved was: 0.79.

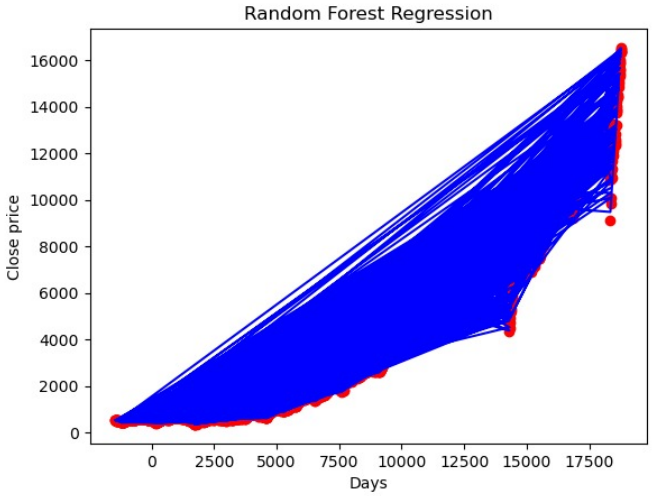


Figure 11. Accuracy score for NYA index 0.9998

## F.5. Multi Layer Perceptron

After this, we applied MLP model on all the 13 stock indices, found its evaluation metrics (refer Figure 12). Also applied grid search to tune hyperparameters: (hidden layer sizes, max iterations). The R2 score varied from 0.11 to 0.95 for different indices. This huge varaition in R2 score for different stock indices is beacause of: Random initialization of weights as the model may get stuck in different local minimas. Also, for some indices there is sudden change in stock market conditions and data available for some is quite less.

## F.6. K- Means Clustering

. Since there was huge difference in accuracy for different indices, we applied unsupervised learning- K-means Clustering to cluster the indices based on Compound Annual Growth Rate and Annual Volatilty of the stock indices prices. We found optimal number of clusters using **elbow curve method** and it came out to be **5 Clusters.** See Figure 13.

| | R-squared | MAE | RMSE | Best Params |
|---|---|---|---|---|
| HSI | 0.747 | 453.55079 | 21.297 | {'hidden_layer_sizes': (50, 50, 50), 'max_iter... |
| NYA | 0.952 | 554.63421 | 23.551 | {'hidden_layer_sizes': (50, 50, 50), 'max_iter... |
| IXIC | 0.537 | 1207.14987 | 34.744 | {'hidden_layer_sizes': (50, 100, 50), 'max_ite... |
| 000001.SS | 0.376 | 80.47724 | 8.971 | {'hidden_layer_sizes': (100, 100, 100), 'max_i... |
| N225 | 0.585 | 37.00860 | 6.083 | {'hidden_layer_sizes': (50, 50, 50), 'max_iter... |
| N100 | 0.118 | 169.57819 | 13.022 | {'hidden_layer_sizes': (50, 50, 50), 'max_iter... |
| 399001.SZ | 0.397 | 393.98141 | 19.849 | {'hidden_layer_sizes': (50, 100, 50), 'max_ite... |
| GSPTSE | 0.847 | 1421.32870 | 37.701 | {'hidden_layer_sizes': (50, 100, 50), 'max_ite... |
| NSEI | 0.372 | 19.26507 | 4.389 | {'hidden_layer_sizes': (100, 100, 100), 'max_i... |
| GDAXI | 0.643 | 2310.70475 | 48.070 | {'hidden_layer_sizes': (50, 50, 50), 'max_iter... |
| SSMI | 0.674 | 1342.87621 | 36.645 | {'hidden_layer_sizes': (50, 50, 50), 'max_iter... |
| TWII | 0.434 | 54.07342 | 7.353 | {'hidden_layer_sizes': (50, 100, 50), 'max_ite... |
| J203.JO | 0.490 | 279.76788 | 16.726 | {'hidden_layer_sizes': (50, 100, 50), 'max_ite... |

Figure 12. MLP Result metric for all index

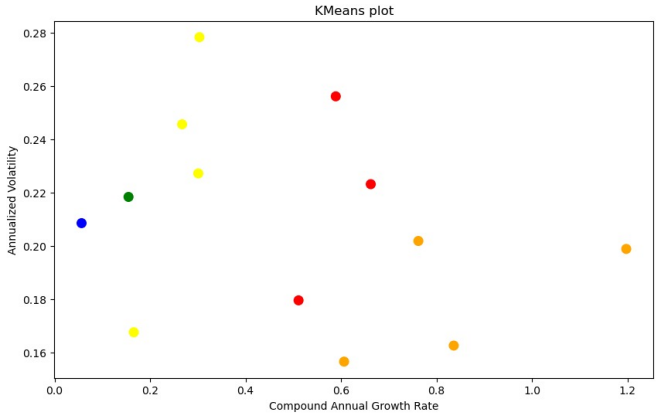


Figure 13. K-Means Clusters Of Stock Indices

### F.7. Novelty: LSTM (Long Short-Term Memory)

After reading multiple research papers and articles and also seeing so much variances in the accuracy of other models, we found out that LSTMs excel in stock market prediction due to their ability to capture complex temporal dependencies and handle long sequences of data. They effectively model non-linear patterns, inherent volatility, and intricate relationships in financial time series, crucial for accurate predictions. So, we applied this model and results came out to be very **consistent** across all indices. **Accuracy ranging from 0.82 to 0.99.** See Figure 14.

## G. Conclusion and Future Work

After applying multiple regression models, we conclude that in **LSTM model** is the best with most consistent accuracy across all the stock indices. While among the ML models best is **Random Forests** (with pruning). To date, we managed to follow the tentative timeline we had proposed. We have curated the datasets, trained and tested a model for stock market prediction data input, and received a satisfying result using all the regressor models.
In the future, we can work on more Deep Learning Models like CNN, RNN, and different algorithms for more improved scores and accuracy.

| | R-squared | MAE | RMSE |
|---|---|---|---|
| HSI | 0.829 | 148.98703 | 12.206 |
| NYA | 0.993 | 143.31517 | 11.971 |
| IXIC | 0.991 | 187.63250 | 13.698 |
| 000001.SS | 0.954 | 7.34691 | 2.711 |
| N225 | 0.996 | 2.61652 | 1.618 |
| N100 | 0.938 | 20.69862 | 4.550 |
| 399001.SZ | 0.985 | 27.13182 | 5.209 |
| GSPTSE | 0.951 | 249.17788 | 15.785 |
| NSEI | 0.861 | 5.23969 | 2.289 |
| GDAXI | 0.954 | 307.09372 | 17.524 |
| SSMI | 0.946 | 193.52196 | 13.911 |
| TWII | 0.992 | 5.23388 | 2.288 |
| J203.JO | 0.908 | 100.51614 | 10.026 |

Figure 14. LSTM Result metric for all index

## H. Member Contribution

**Pallav Singla** Literature review, Data Extraction, and Collection, EDA. Implementation of various models and Analysis and inference of the data and results.

**Dux Pal Singh** Literature review, Data Extraction, and Collection, EDA. Implementation of various models and Analysis and inference of the data and results.

**Meet Popat** Literature review, Data Extraction, and Collection, EDA. Implementation of various models and Analysis and inference of the data and results.

**Shourya Jindal** Literature review, Data Extraction, and Collection, EDA. Implementation of various models and Analysis and inference of the data and results.

## References

[1] Bhawna Panwar, Gaurav Dhuriya, Prashant Johri, Sudeept Singh Yadav, and Nitin Gaur. Stock market prediction using linear regression and svm. In *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 629–631. IEEE, 2021. 1

[2] Ishita Parmar, Navanshu Agarwal, Sheirsh Saxena, Ridam Arora, Shikhin Gupta, Himanshu Dhiman, and Lokesh Chouhan. Stock market prediction using machine learning. In *2018 first international conference on secure cyber computing and communication (ICSCCC)*, pages 574–576. IEEE, 2018. 2