Project Report

On

Stock Market Analysis and Prediction

Submitted

in the partial fulfillment for the award of

Post Graduate Diploma in Big Data Analytics (PG-DBDA)

from **Know-IT ATC, CDAC ACTS, Pune**

Guided by:

Mr. Anay Tamhankar & Mr. Prasad Deshmukh

Submitted By:

Amit Gavval (230343025017)

Ganesh Ingale (230343025020)

Satyam Rokade (230343025040)

Sushant Bhabutkar (230343025051)

CERTIFICATE

<u>TO WHOMSOEVER IT MAY CONCERN</u>

This is to certify that.

Amit Gavval (230343025017)

Ganesh Ingale (230343025020)

Satyam Rokade (230343025040)

Sushant Bhabutkar (230343025051)

Have successfully completed their project on

Stock Market Analysis and Prediction using Machine Learning

Under the guidance of

Mr. Anay Tamhankar

&

Mr. Prasad Deshmukh

ACKNOWLEDGEMENT

This project **“Stock Market Analysis and Prediction”** was a great learning experience for us, and we are submitting this work to Know-IT ATC, CDAC CTS, Pune.

We all are very glad to mention the name of **Mr. Anay Tamhankar & Mr. Prasad Deshmukh** for their valuable guidance at work on this project. Their continuous guidance and support helped us to overcome various obstacles during project work.

We are highly grateful to **Mr. Vaibhav Inamdar** (Center Coordinator, Know-it, Pune) for his guidance and support whenever necessary while doing this course Post Graduate Diploma in Big Data Analytics (PG-DBDA) through C-DAC ACTS, Pune.

Our most heartfelt thanks go to **Mr. Shrinivas Jadhav** (Vice-President, Know-it, Pune) who provided all the required support and his kind coordination to provide all the necessities like required hardware, internet facility and extra lab hours to complete the project, throughout the course and till date, here in Know-IT ATC, CDAC ACTS, Pune.

From:

Amit Gavval	(230343025017)
Ganesh Ingale	(230343025020)
Satyam Rokade	(230343025040)
Sushant Bhabutkar	(230343025051)

TABLE OF CONTENTS

ABSTRACT

This paper proposes a stock market analysis and prediction model using machine learning techniques. The model is designed to predict stock prices based on historical data and market trends. The approach involves data preprocessing, feature engineering, and model training using supervised learning algorithms such as Linear Regression, KNN, Random Forest, Elastic Net and Neural Networks. The performance of the models is evaluated using metrics such as mean squared error and accuracy. The results indicate that the proposed approach can accurately predict stock prices and outperform traditional statistical models. The model can be used by investors and financial analysts to make informed decisions about their investment portfolios.

1. INTRODUCTION

The stock market is a complex and dynamic system that involves various factors such as company financials, market trends, global events, and investor sentiment. Investors and financial analysts often rely on historical data and market trends to make informed decisions about their investment portfolios. However, traditional statistical models may not always accurately predict stock prices due to the complexity of the stock market.

Machine learning has emerged as a promising approach to predict stock prices by analyzing large amounts of historical data and identifying patterns and trends. This paper proposes a stock market analysis and prediction model using machine learning techniques. The model is designed to predict stock prices based on historical data and market trends.

The proposed approach involves data pre-processing, feature engineering, and model training using supervised learning algorithms such as Linear Regression, KNN, Random Forest, Elastic Net and Neural Networks. The performance of the models is evaluated using metrics such as mean squared error and accuracy.

The main objective of this paper is to develop a robust and accurate model for predicting stock prices that can be used by investors and financial analysts to make informed decisions about their investment portfolios. The rest of the paper is organized as follows: Section 2 provides a brief review of related work in the field of stock market prediction using machine learning. Section 3 describes the proposed approach in detail. Section 4 presents the experimental results and performance evaluation of the proposed approach. Finally, Section 5 concludes the paper and highlights future research directions.

2. OBJECTIVE OF PROJECT

The objective of this project is to develop a stock market analysis and prediction model using machine learning techniques. The model is designed to predict stock prices based on historical data and market trends. The main objectives of the project are as follows:

1. **Data Preprocessing:** The first objective is to preprocess the historical data by removing missing values, outliers, and noise to improve the quality of the data.
2. **Feature Engineering:** The second objective is to select the most relevant features that can best represent the data and improve the performance of the model.
3. **Model Training:** The third objective is to train the model using supervised learning algorithms such as Linear Regression, KNN, Random Forest, Elastic Net and Neural Networks to predict stock prices based on historical data and market trends.
4. **Performance Evaluation:** The fourth objective is to evaluate the performance of the model using metrics such as mean squared error, mean absolute error, and accuracy.
5. **Comparison with Traditional Models:** The fifth objective is to compare the performance of the proposed approach with traditional statistical models to demonstrate the effectiveness of the proposed approach.

The ultimate objective of this project is to develop a robust and accurate model for predicting stock prices that can be used by investors and financial analysts to make informed decisions about their investment portfolios.

3. FUNCTIONAL REQUIREMENTS

1) Python 3:

❿ Python is a high-level programming language that is easy to learn and use.

❿ Python is an interpreted language, which means that code can be executed on the fly, without theneed for compilation.

❿ Python is open source and free to use, with a large and active community of developers contributingto its development and maintenance.

❿ Python has a vast collection of third-party libraries and packages, such as NumPy, Pandas,Matplotlib, and Scikit-learn, among others, that make it easy to perform data analysis.

2) Tableau:

❿ Tableau is a data visualization and business intelligence software that allows users to connect,analyze, and share data in a visual and interactive way.

❿ It offers a user-friendly drag-and-drop interface that enables users to create interactive dashboards,reports, and charts without the need for complex coding or programming.

❿ Tableau supports various data sources, including spreadsheets, databases, cloud services, and big dataplatforms, such as Hadoop and Spark.

Data Cleaning: -

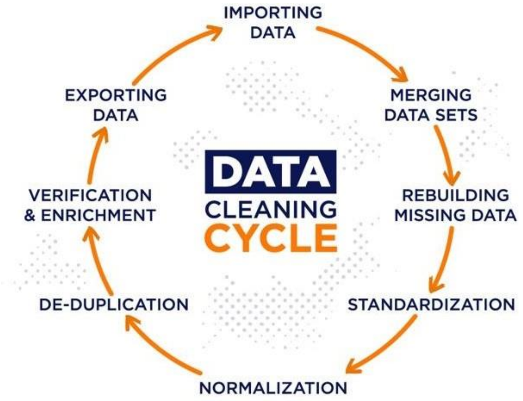


Fig: Data Cleaning Process

Data cleaning, also known as data cleansing or data scrubbing, is the process of identifying and correcting or removing inaccurate, incomplete, or irrelevant data in a dataset. Data cleansing is an essential step in the data preparation process, as it helps ensure the quality, consistency, and reliability of the data used for analysis.

Without proper data cleaning, data analysis and modelling can lead to erroneous or biased results, which can have serious consequences for businesses and organizations.

Hence, it is a critical step in the data preparation process, as it can significantly impact the accuracy and reliability of the insights and decisions that are derived from the data. By improving the quality of data, organizations can gain a better understanding of their operations, customers, and market trends, and make more informed and effective decisions.

4. SYSTEM ARCHITECTURE

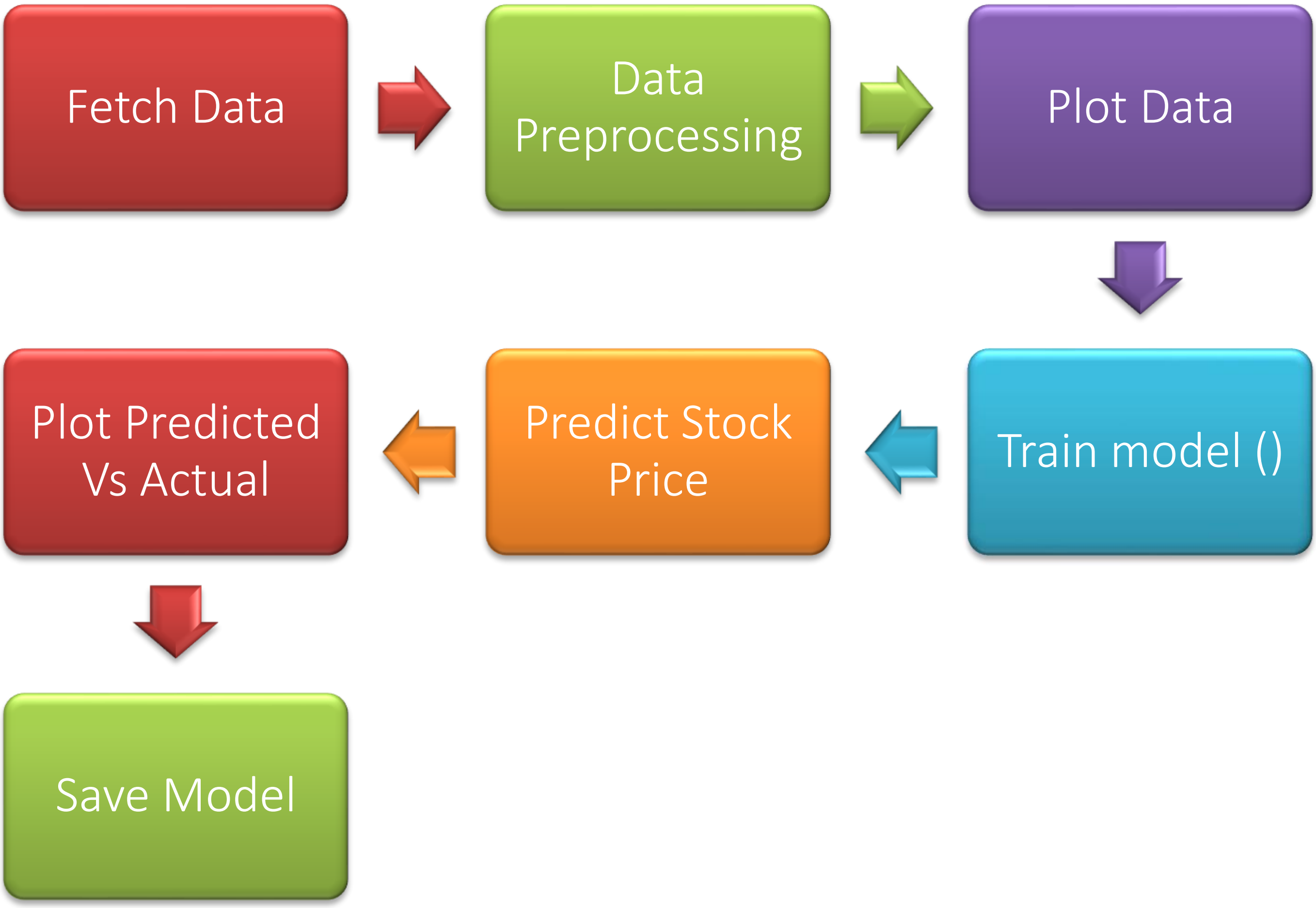
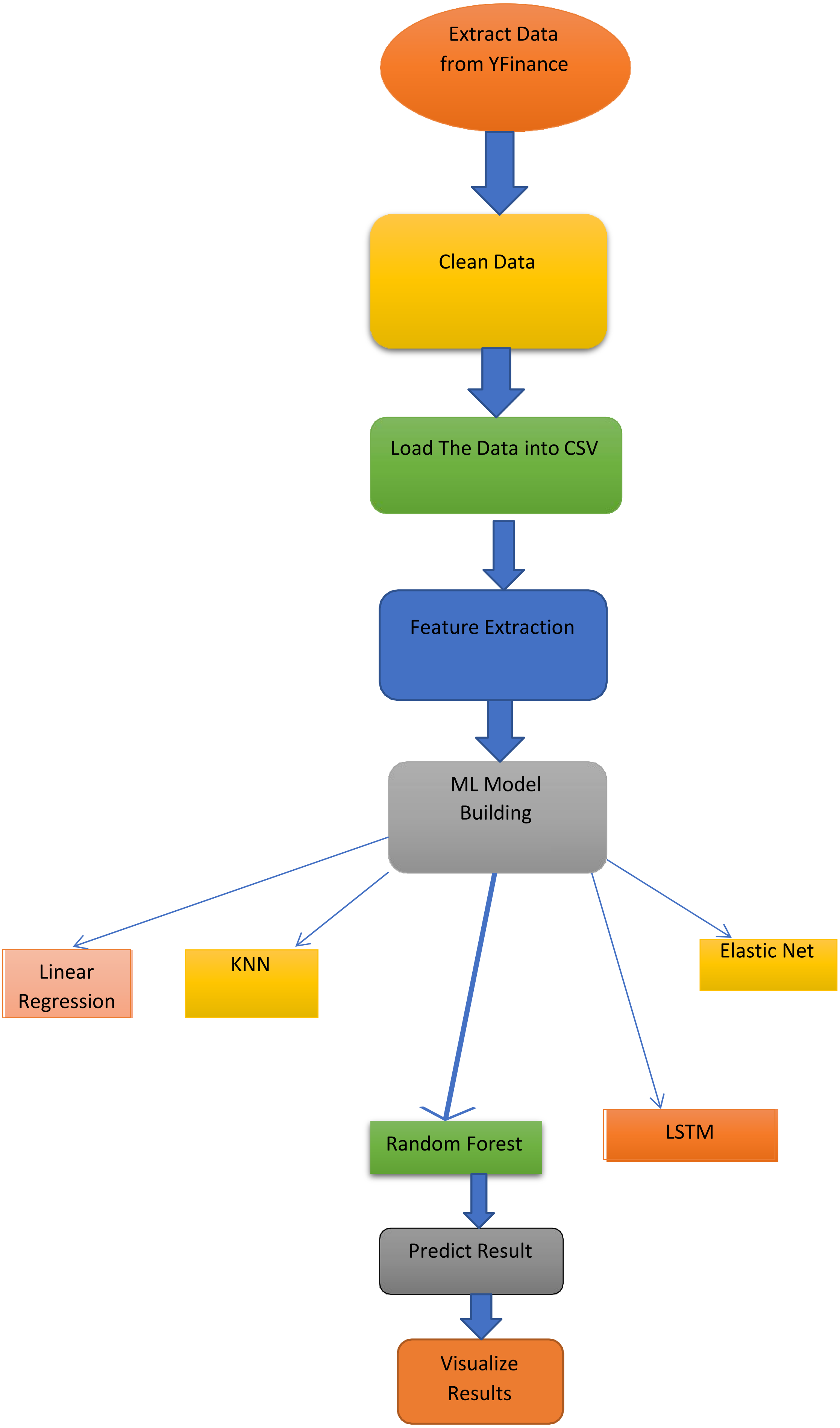


Fig: System Architecture of Stock Market Analysis and Prediction

5. METHODOLOGY



6. MACHINE LEARNING ALGORITHMS

- In our project, we applied various Regression Algorithms such as Linear Regression, KNN, Random Forest, Elastic Net and LSTM. After the implementation, were able to analyze the accuracy of the algorithms on our data.
- Machine learning is a subfield of artificial intelligence that involves developing algorithms and models that enable computers to learn from data and make predictions or decisions without being explicitly programmed. The goal of machine learning is to enable computers to improve their performance over time by learning from experience and feedback.
- The methods commonly used include Linear Regression, KNN, Random Forest, Elastic Net and LSTM with Hyperparameter Tuning. They were mainly used for classification and analysis. In our project, various machine learning algorithms which we used are as follows:

1. Linear Regression: -

- Linear regression is a statistical modeling technique that aims to establish a linear relationship between two or more variables. In the context of stock market data prediction, linear regression can be used to predict future stock prices based on past price movements.

- In a simple linear regression model, the relationship between the dependent variable (the stock price) and the independent variable (the time period) is modeled using a straight line. The equation of the line is determined by finding the slope and the intercept that minimize the sum of the squared residuals between the predicted values and the actual values.

- To use linear regression for stock market data prediction, historical stock prices are used to train the model. Once the model is trained, it can be used to predict future stock prices by plugging in values for the independent variable (i.e., time period) and solving for the dependent variable (i.e., stock price).

- It's important to note that while linear regression can be a useful tool for predicting stock prices, it's not always accurate. There are many factors that can influence stock prices, such as company performance, economic trends, and global events, that cannot be accounted for in a simple linear regression model. As such, it's important to use linear regression in conjunction with other analytical tools and to exercise caution when making investment decisions based on the results of a linear regression analysis.

2. K-Nearest Neighbors (KNN): -

- K-nearest neighbors (KNN) is a machine learning algorithm that can be used for stock market data prediction. KNN is a type of supervised learning algorithm that can be used for both classification and regression analysis.

- In KNN, the idea is to predict the value of a new data point by finding the "k" closest data points in the training data and taking the average of their values. The value of "k" is a hyperparameter that is chosen by the user and represents the number of neighbors that are considered.

- To use KNN for stock market data prediction, historical stock prices and other relevant variables are used to train the model. Once the model is trained, it can be used to predict future stock prices by finding the "k" closest data points and taking their average.

- One advantage of KNN is that it is easy to understand and implement, making it a good choice for beginners.It can also handle non-linear data and can be used for multivariate regression analysis.

- However, KNN also has some limitations. One limitation is that it can be sensitive to the choice of "k" and may require some trial and error to find the optimal value. It can also be computationally expensive, especially when dealing with large datasets.

- Overall, KNN is a useful machine learning algorithm that can be used for stock market data prediction. However, like any analytical tool, it should be used in conjunction with other methods and with caution whenmaking investment decisions based on its results.

3. Random Forest: -

- Random forest is a machine learning algorithm that can be used for stock market data prediction. It is anensemble learning method that combines multiple decision trees to create a more accurate and stable model.

- In random forest, many decision trees are created using random subsets of the input variables and the training data. Each decision tree is then used to predict the outcome of a new data point, and the final prediction is the average or majority vote of all the individual tree predictions.

- The advantage of using random forest for stock market data prediction is that it can handle high-dimensional data with complex interactions between variables. It is also less prone to overfitting than individual decision trees because the ensemble of trees can average out the individual errors.

- To use random forest for stock market data prediction, historical stock prices and other relevant variables areused to train the model. Once the model is trained, it can be used to predict future stock prices by plugging invalues for the independent variables and solving for the dependent variable.

- Overall, random forest is a powerful machine learning algorithm that can be used for stock market data prediction. However, like any analytical tool, it should be used in conjunction with other methods and with caution when making investment decisions based on its results.

4. Elastic Net: -

- Elastic Net is a regularization technique that combines both L1 (Lasso) and L2 (Ridge) regularization. It aims to address the limitations of individual Lasso and Ridge regularization methods by offering a balanced approach. In stock market analysis, Elastic Net can help mitigate issues like multicollinearity among features and feature selection, ultimately leading to improved model performance and more stable predictions.

- Elastic Net encourages sparsity in feature coefficients, automatically selecting relevant features for prediction. This is particularly useful when dealing with a large number of potential predictor variables in financial datasets.

- Stock market data often exhibits multicollinearity, where features are correlated. Elastic Net's combination of L1 and L2 regularization helps in handling this issue, preventing overfitting and providing more interpretable coefficients.

- Outliers can significantly impact stock market data. Elastic Net's robustness to outliers, especially when compared to Lasso, ensures more stable predictions in the presence of extreme data points.

- The Elastic Net parameter, which balances the L1 and L2 penalties, allows for adjusting the degree of regularization based on the specific dataset characteristics. This adaptability is crucial in capturing the inherent complexities of stock market dynamics.

5. LSTM: -

- Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) that can be used for stock market data prediction. LSTM is particularly useful for handling sequential data, such as stock prices, wherethe order of the data points is important.

- In LSTM, the network consists of one or more LSTM cells that can store information over a long period of time. Each cell is connected to other cells through a set of weights that are learned during the training process.

- To use LSTM for stock market data prediction, historical stock prices and other relevant variables are used totrain the model. The data is typically split into training, validation, and test sets. The LSTM network is trained on the training set using a backpropagation algorithm to adjust the weights in order to minimize the prediction error.

- Once the model is trained, it can be used to predict future stock prices by feeding in new data and using the LSTM cells to generate a prediction. The accuracy of the model can be evaluated on the validation set, and the final performance can be assessed on the test set.

- LSTM is particularly useful for stock market data prediction because it can capture long-term dependencies in the data and can handle noisy and irregularly spaced data. It can also be used for multivariate time series analysis by incorporating multiple input variables.

- However, LSTM can be computationally expensive and requires a large amount of training data to perform well. Additionally, like any analytical tool, it should be used in conjunction with other methods and with caution when making investment.

7. DATA VISUALIZATION AND REPRESENTATION

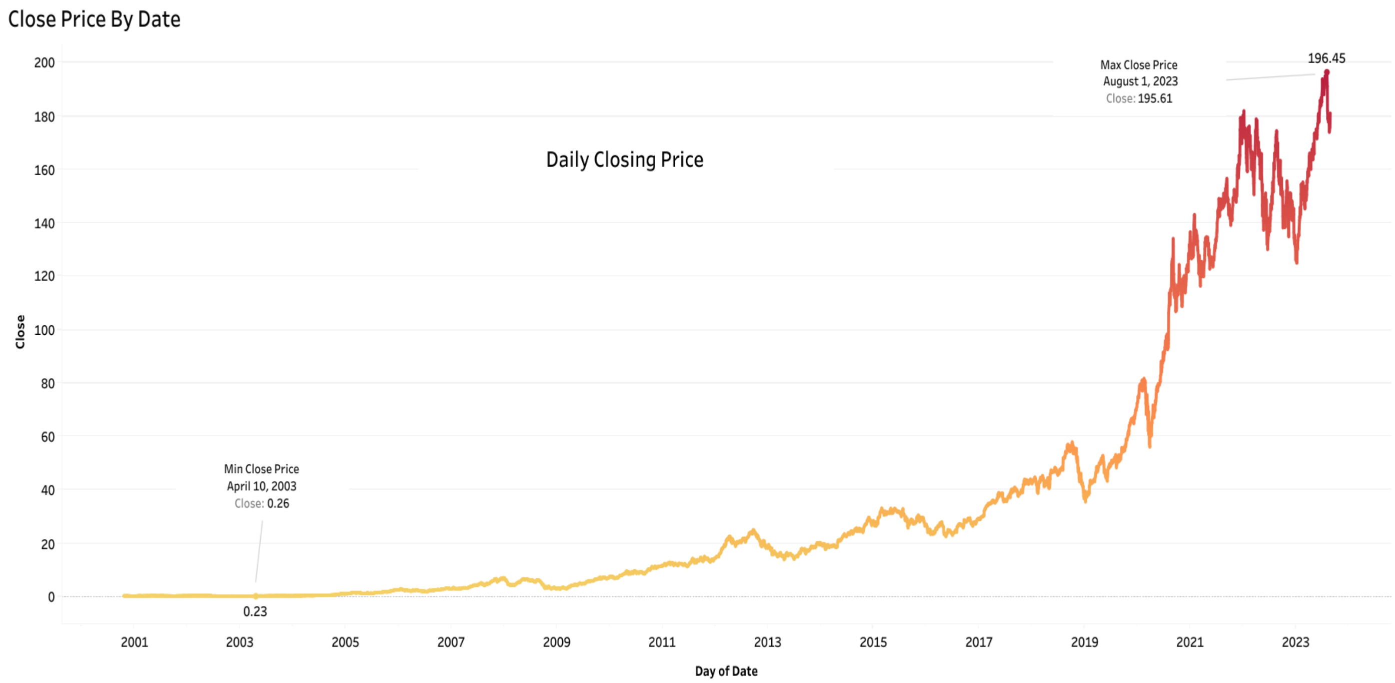


Fig.7.1 Graph for DATE(Day) vs CLOSE

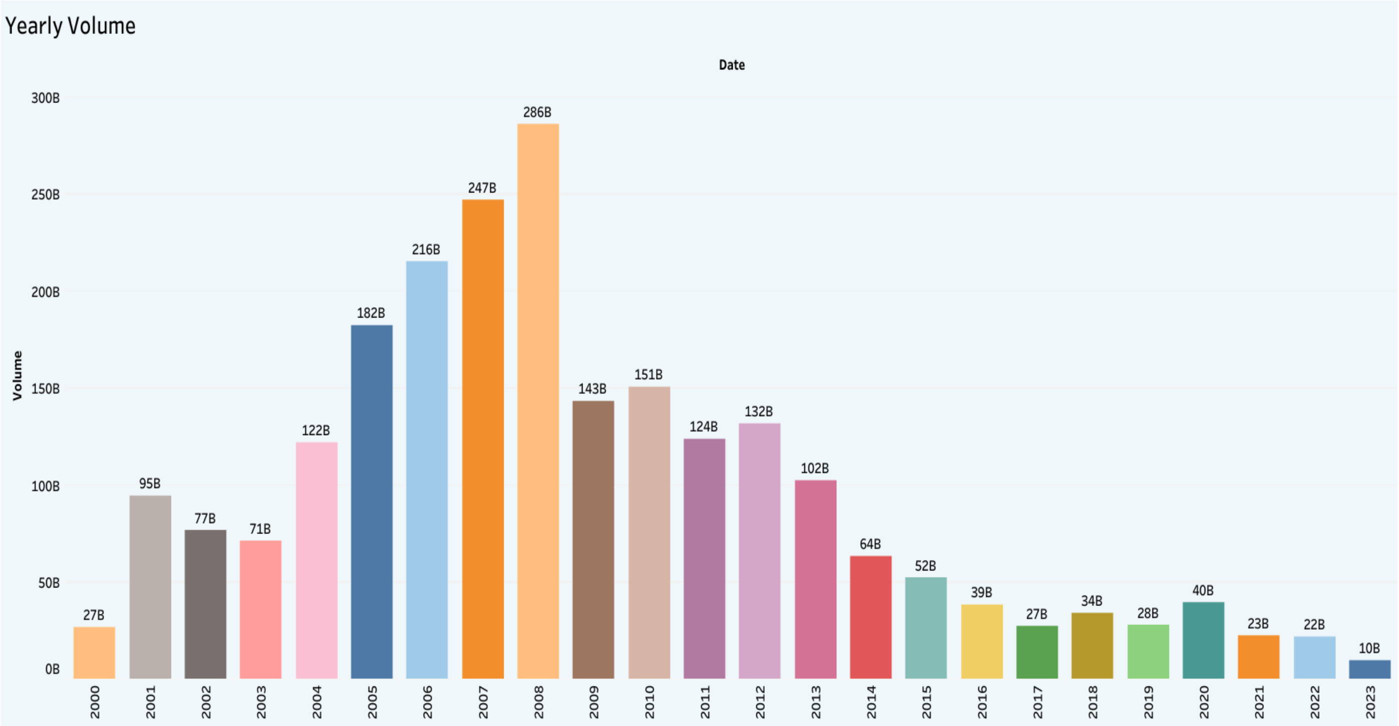


Fig.7.2 Graph for DATE vs VOLUME

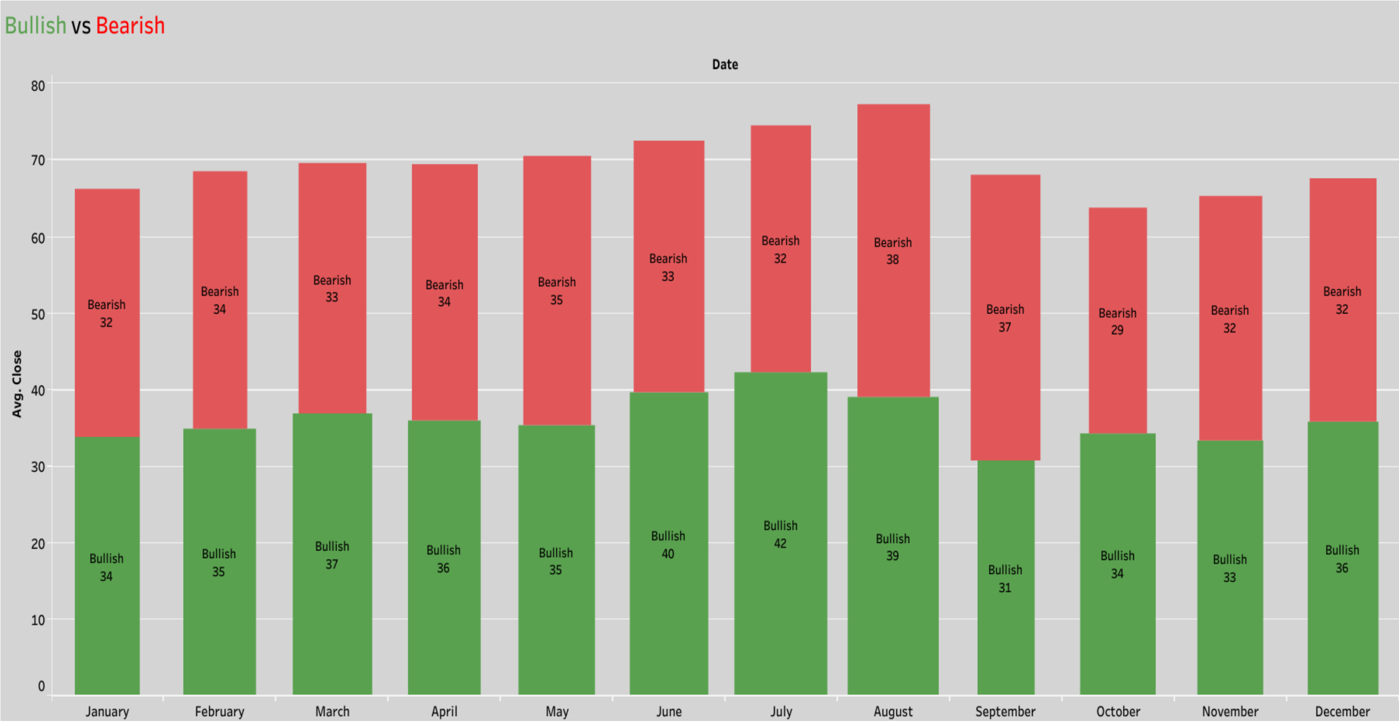


Fig.7.3 Graph for BULLISH vs BEARISH

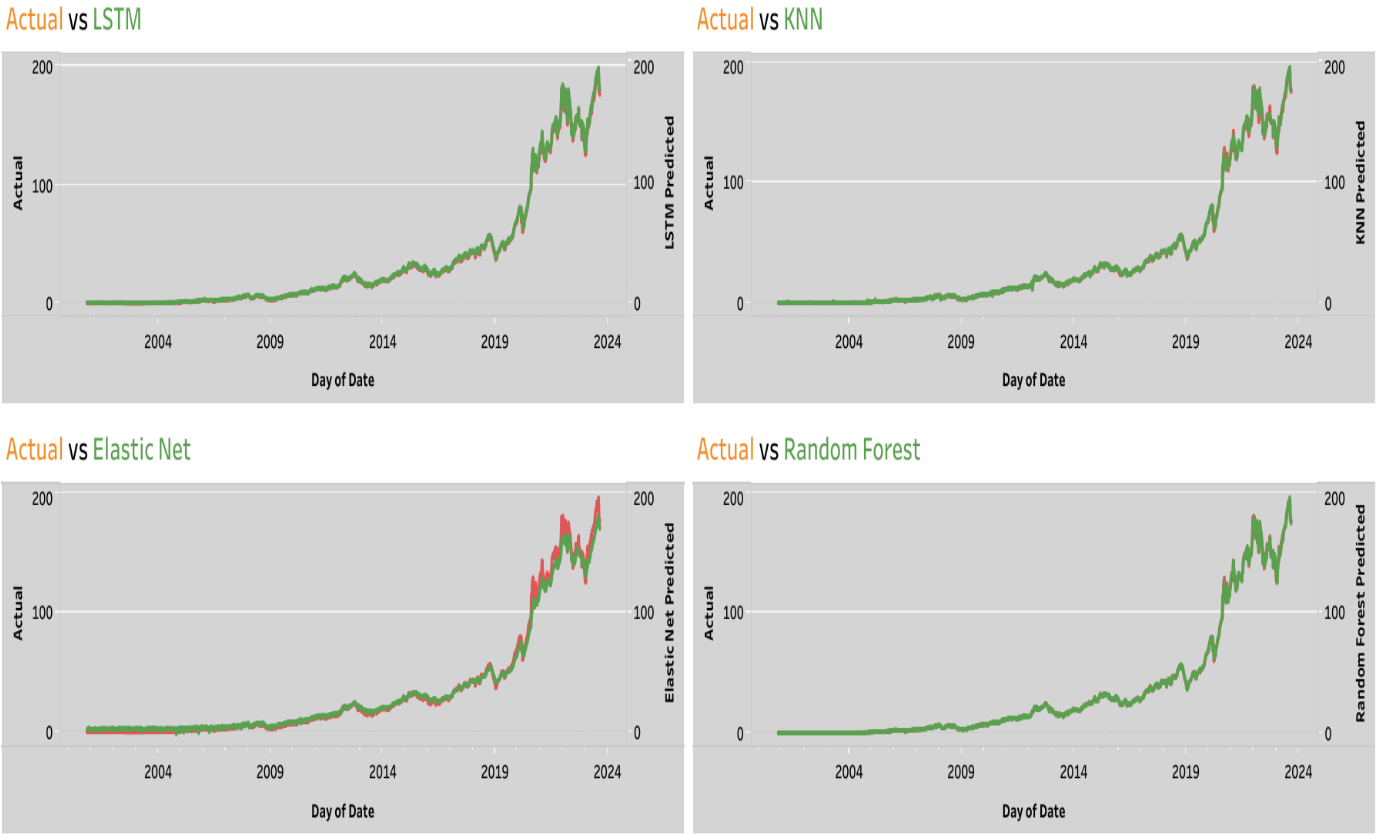


Fig.7.4 Graphs of ACTUAL vs PREDICTED

8. CONCLUSION AND FUTURE SCOPE

- In conclusion, our LSTM and KNN model had given the precise predicted values in terms of R2Score.

- With the introduction of Machine Learning and its strong algorithms, the most recent market research and Stock Price Prediction using machine learning advancements have begun to include such approaches in analyzing stock market data. The Opening Value of the stock, the Highest and Lowest values of that stock on the same day, as well as the Closing Value at the end of the day are all indicated for each date.

- Furthermore, the use of Machine Learning in Stock Market analysis and predictions can greatly improve the efficiency to predict the stock prices, as it can quickly analyze large volumes of data and provide accurate predictions in real-time.

9. FUTURE ENHANCEMENT

- The limitation of the proposed system is its computational speed, especially with respect to sliding-window validation as the computational cost increases with the number of forward day predictions.

- The proposed model does not predict well for sudden changes in the trend of stock data.

- This occurs due to external factors and real-world changes affecting the stock market.

- We can overcome this by implementing Sentiment Analysis and Neural Networks to enhance the proposed model.

References

1. https://finance.yahoo.com/
2. https://www.kaggle.com/code/faressayah/stock-market-analysis-prediction-using-lstm
3. https://www.simplilearn.com/tutorials/machine-learning-tutorial/stock-price-prediction-using-machine-learning
4. Brownlee, J. (2020). Introduction to Machine Learning with Python: A Guide for Data Scientists. Machine Learning Mastery.
5. Python documentation: https://docs.python.org/3/
6. Sklearn Documentation : https://scikit-learn.org/
7. "Machine Learning using Python" by Prof. U Dinesh Kumar, IIM Bangalore.