DAILY STOCK CLOSING PRICE PREDICTION USING NEURAL NETWORK AND MACHINE LEARNING TECHNIQUES

DENIS ROYSTAN RICHARD DALMEIDA

Thesis Report

AUGUST 2021

DEDICATION

This dissertation is dedicated wholeheartedly to my beloved parents, who have been nicely supportive in my whole research work.

I also dedicate this dissertation to my brother & sister, mentor, and friends who have been a source of inspiration and gave me strength when I thought of giving up, who continually provide their moral, spiritual, emotional, and financial support.

Thank you to my academic advisor(s) who guided me in this process and the upGrad Community who kept me on track.

And lastly, dedicated this dissertation to everyone who encouraged me to pursue my dreams and finish my dissertation.

ACKNOWLEDGEMENTS

I am honestly indebted to many admirable people without whom it would not be possible to complete the learning journey and research work. Until my accomplishment with my dissertation has one-hundred percent come to its end, my soul has been tied up to those inspiring individuals cheerfully round the clock. Reiteration, my endeavour without their advice is lacking effort. I am indeed grateful and overpowered too:

First of all, my supervisor, Ms. Drishti Singh, whose advice has been in my memory every day. Also, be the best and friendliest of all individuals and supervisors;

Miss Sneha Barathe, who has perfectly done everything to remind us of our research, and she has also softly encouraged us with her compassionate pleasing delicacy;

Dr Manoj Jayabalan, for his compelling lessons on research/thesis work, I could not linger for a while to attend most of them;

My tutors, who through the years have carefully taught us to capitalize on learning even more actively and independently;

My friends and classmates for their inspirational, motivational and passionate encouragement to complete my work with great success and determination.

ABSTRACT

In the 21st Century, the stock prediction has become one of the most popular topics for many trading / investing stakeholders for gaining better returns on investments. But the traditional methods were not enough to predict the stocks more precisely and quickly. Also, the haphazard and inconsistent historical series mark their prediction cumbersome. Availing the machine learning and deep learning approaches, we can overcome the lack in predicting & can also achieve enhanced results. In this paper we are using two models, Long short-term memory multivariate model and Extreme Gradient Boosting model with fine tuning of hyper-parameters to predict the stock's close price for the next business day. Model evaluation has been done by extensively used regression performance metrics: R-Squared, Root Mean Square Error, Mean Bias Error, Mean Absolute Percent Error. The financial dataset used here considers factors like; Open Price, Close Price, High Price, Low Price, Volume Weighted Average Price and newly created variables like Moving Averages. The best regression score achieved from R-Squared stands at 0.98 and the low values of Root Mean Square Error and Mean Absolute Percent Error indicators for Long short-term memory model shows that it is more efficient in predicting stock closing price than Extreme Gradient Boosting model. Thus it can be concluded through this research work that deep learning models are better than machine learning models in stock time series data forecasting.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

3MMA……. Three Month Moving Average
AI……. Artificial Intelligence
ANN……. Artificial Neural Network
ARIMA…… Auto Regressive Integrated Moving Average
AR…… Auto Regressive
CNN……. Convolutional Neural Network
CPU…….. Central Processing Unit
DL…… Deep Learning
DMA…… Day Moving Average
DNN…… Deep Neural Network
DRNN…….. Deep Recurrent Neural Network
EDA……. Exploratory Data Analysis
FE……. Feature Extraction
GA……. Genetic Algorithms
GPU…….. Graphics Processing Unit
LM……. Levenberg-Marquardt
LR……. Linear Regression
LSTM……..... Long Short-Term Memory
MAPE…… Mean Absolute Percentage Error
MBE…….. Mean Bias Error
ML……. Machine Learning
MLP……. Multilayer Perceptron
PCA……... Principal Component Analysis
PSR……... Phase Space Reconstruction
R2……..... R Squared
RAM……. Random Access Memory
RBF……... Radial Basis Function
RFE……... Recursive Feature Elimination
RF…….. Random Forest
RMSE…… Root Mean Square Error
RNN…….. Recurrent Neural Network
RSI……..... Relative Strength Index

SCG…….. Scaled Conjugate Gradient

SVR…….. Support Vector Regression

VWAP……... Volume Weighted Average Price

XGBoost…..... Extreme Gradient Boosting

CHAPTER 1

INTRODUCTION

Stock is in today's world very notable among stock investors and financing institutions. The stock market plays a crucial role in determining the economic strength of any country. Also, the country's economy is more dependent on its stock exchange indices. "The rise and fall of stock prices are influenced by many factors such as politics, economics, society, and the market. For stock investors, the forecast trend of the stock market is related directly to the acquisition of profits. The more accurate the forecast, the more effectively it can avoid risks" (Ding and Qin, 2020). It desires investors to predict the stock price in advance by time series analysis of the stock's continuous data, thus, proving to occur variations in the value of a stock.

Nevertheless, stock forecasting is a complicated job because of noise in continuous data, non-linearity, and non-stationary characteristics. In the economic writings, stock value foretelling has been compartmentalized as technical analysis, fundamental analysis, & machine learning (ML) and Artificial Intelligence (AI) approaches. In the research field also the stock forecasting plays a precise role as it decides a country's financial development. Stock prediction helps companies to undertake profit-making decisions for the future growth aspects of their products and services.

In the previous papers related to this research, researchers used different machine learning and deep learning techniques for solving prediction problems. The algorithm includes Random Forest, Support Vector Regression (SVR), Linear Regression, Time series analysis, Artificial Neural Network (ANN), Long Short Term Memory (LSTM), Convolutional Neural Network (CNN), and some even created hybrid models to tune the prediction results. Some of these papers achieved better results or around 95% of accuracy in prediction. These papers inspire the use of machine learning techniques in stock price prediction. Even we have found gaps in the past research papers on which we have worked in the current research.

The stock market data is highly volatile and depends on various external factors such as the economic strength, interest rates, inflation, unemployment, natural calamities, pandemic, politics, company-related factors, and sentiment of an investor. In numerous factors, the data is lacking that is available for predicting the stock prices. Even, many times the dataset size available for prediction purposes is very small, which can lead to appropriate predicted stock prices. Also, sometimes countless false tips from various sources lead to overvalued/undervalued stocks. Even one machine learning algorithm is not so good for better

prediction because of overfitting data and weak learners. Although, above reasons may lead the way to lower prediction accuracy, but, still, researchers are conducting various experiments and feeding all possible stock data to find out a better way or achieve better accuracy in terms of prediction results.

ML and Deep Learning (DL) models for stock prediction are created using Extreme Gradient Boosting (XGBoost) and LSTM variant models in the proposed research. It will take features, namely close price, open, high, low, volume weighted average price (VWAP), and moving averages as input and will provide the following day's predicted close price of a stock of a certain listed firm. Moreover, this paper will also show the inferences of volume traded of that particular stock over timeframes and draw the stock price trend direction upward or downward. Based on the above, it will recommend the stock to buy/sell on the next day. This paper will also discuss the future scope of extending this research with more features and a large dataset.

1.1 Background of the Study

In the past, lots of research papers have been submitted regarding the prediction of an individual stock or indices. Some of them used machine learning techniques like Random Forest, Linear Regression (LR), Support Vector Machine (SVM), & Feature Extraction and even many implemented the time series models like Autoregressive (AR), Auto-Regressive Integrated Moving Average (ARIMA), & Three Months Moving Average (3MMA). Some papers even introduced deep learning techniques like LSTM, ANN & Deep Recurrent Neural Network (DRNN) or hybrid models. Table 1.1 below is the list of similar papers that have been published in the past related to stock price prediction.

Table 1.1 Comparison of previous related research papers

Title & Author(s)	Dataset	Problems	Algorithm(s)	Result
"Stock Closing Price Prediction using Machine Learning Techniques (Mehar Vijh, Deeksha Chandola, Vinay	Yahoo Finance	Stock price predictor performance The limited set of data for prediction	ANN & Random Forest	ANN performs best with higher accuracy. Deep learning models are

Anand Tikkiwal, Arun Kumar)" (Vijh *et al.*, 2020)				faster than ML models
"Stock Market Prediction Using Machine Learning (Ishita Parmar, Navanshu Agarwal, Sheirsh Saxena, Ridam Arora, Shikhin Gupta, Himanshu Dhiman, Lokesh Chouhan)" (Parmar *et al.*, 2018)	Yahoo Finance	Effective future value of the stock price. Better prediction results than the traditional way. Vanishing Gradient problem in the regression model	Linear Regression & LSTM	The LSTM provides better results. Machine learning has proven to be effective in stock prediction
"Short-term stock market price trend prediction using a comprehensive deep learning system (Jingyi Shen, M. Omair Shafiq)" (Shen and Shafiq, 2020)	This dataset consists of 3558 stocks from the Chinese stock market. Chosen 2 years as the period of the dataset.	Important features to be selected for prediction A Hybrid tuned model of LSTM has been approached for a better outcome.	(Feature Extraction + Recursive Feature Extraction + Principal Component Analysis) and LSTM	The hybrid model performed well and achieved better results.
"STOCK PRICE PREDICTION USING ARTIFICIAL NEURAL NETWORKS	Infratel firm dataset is used. It has close, open, low, high &	Comparison of models better for stock prediction Limited dataset and fewer features	ANN, LSTM, AR, ARIMA	Comparing to various trained network models, ANN seems to be the best.

(Padmaja Dhenuvakonda, R. Anandan, N. Kumar)" (Dhenuvakonda, Anandan and Kumar, 2020)	volume as features.			
"Stock Market Prediction Using Machine Learning(ML) Algorithms (M Umer Ghani, M Awais, Muhammad Muzammul)" (Umer, Awais and Muzammul, 2019)	Diverse data is used such as research theories, data sets & resources related to financial presentation data.	To help investors invest with less risk Stock price prediction with better performance and accuracy.	Linear Regression; Exponential Smoothing; Time Series Forecasting	Exponential smoothing prediction turns out to be barely inaccurate. Subsequently, advised as elite stock predictor with general trend analysis
"Study on the prediction of stock price based on the associated network model of LSTM (Guangyu Ding, Liangxi Qin)" (Ding and Qin, 2020)	Shanghai composite index, other 2 are stocks of PetroChina on Shanghai stock exchange and ZTE Corporation on Shenzhen stock exchange	Limited dataset and its features. To predict with higher accuracy	LSTM network and LSTM-based deep-recurrent neural network (DRNN)	Experiments show that the average accuracy of Associated Net model is not only better than that of the other two models. Moreover, it can predict multiple values simultaneously, and the average accuracy of

				each predicted value is over 95%.
"Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms (Mahla Nikou, Gholamreza Mansourfar, Jamshid Bagherzadeh)" (Nikou, Mansourfar and Bagherzadeh, 2019)	iShares MSCI United Kingdom from January 2015 to June 2018. Collected from Yahoo Finance	Limited dataset with less features External factors data not considered	ANN, SVM, Random Forest (RF), Recurrent Neural Network (RNN), LSTM	The results of the study show that the recurrent network method with an LSTM block functions better in prediction. SVR method has higher precision than neural network and RF
"Stock Price Prediction Using Long Short Term Memory (Raghav Nandakumar, Uttamraj K R, Vishal R, Y V Lokeswar)" (Nandakumar *et al.*, 2018)	Yahoo Finance Google Finance	Less features for prediction To compare LSTM and ANN	LSTM	LSTM has a better prediction accuracy than ANN. An analysis of the results also indicates that both models give better accuracy when the size of the dataset increases.

"Stock Price Prediction (N P Samarth, Gowtham V Bhat, Hema N)" (N P Samarth, Gowtham V Bhat and Hema N, 2019)	Enron Corp dataset (2013 to 2017)	To eliminate human error as the decision process	Random Forest	This model is successful in predicting the futures stock prices with a good accuracy rate depending on the various attributes given by the user in the collected data
"Indian stock market prediction using artificial neural networks on tick data (Dharmaraja Selvamuthu, Vineet Kumar, Abhishek Mishra)" (Selvamuthu, Kumar and Mishra, 2019)	Tick data of Reliance Private Limited from period 30 NOV 2017 to 11 JAN 2018	To check if tick data helps bring seasonal and annual factors for prediction	ANN based on different leaning algorithms	All three algorithms provide an accuracy of 99.9% using tick data. The accuracy over 15-min dataset drops to 96.2%, 97.0% and 98.9% for Levenberg-Marquardt (LM), Scaled Conjugate Gradient (SCG) and Bayesian Regularization

"Stock Closing Price Prediction using Machine Learning SVM Model (Desai Mitesh Madhusudan)" (Madhusudan, 2020)	Yahoo Finance	To check SVM model for prediction Closest prediction to actual price	SVM	Radial Basis Function (RBF) SVR kernel is the best method for prediction. It predicted the stock closing prices closest comparing to the original value of same days
"Machine Learning Model For Stock Market Prediction (Ashwini Kanade, Sakshi Singh, Shweta Rajoria, Pooja Veer, Nayan Wandile)" (Kanade, 2020)	Yahoo Finance	To predict the stock market price fluctuation using ML	RNN, LSTM	Good results seen in the LSTM Model
"FORECASTING WITH DEEP LEARNING: S&P 500 INDEX (FIRUZ KAMALOV, LINDA SMAIL, IKHLAAS GURRIB)" (Kamalov, Smail and Gurrib, 2020)	Yahoo Finance	Time series forecasting CNN Model for prediction	Convolution-based neural network model	The model achieves the top accuracy rate of 56.21%.

1.2 Problem Statement

It is a common practice in the stock market, equity market, or share market for investors to strive for the best stock that will definitely return better perks. There are several financial advisory agencies that give a rating to the stock based on their past performance. Back in the days when there was no AI/ML in the finance domain, professionals started using fundamental and technical analysis to predict stock movements. It is termed the best tool available those days, but it needs thorough understanding of statistics and accounting. In the worst case scenario, investors unfailingly depend upon the tips/recommendation given by the brokers at the exchange or by the credit rating agencies. Henceforth, this leads to a biased outcome, and sometimes the large group of investors are left disappointed and hopeless. This study aims to predict the stock closing price proximately to actual prices. This will help the investors to deliberately choose the stocks and also walk in the path of a successful investment journey.

1.3 Aim and Objectives

The principal objective of this research is to forecast the next-day closing price of a stock using LSTM and XGBoost to provide better results in terms of precision and less erroneous results and also to prove that deep learning models are better than machine learning models.

The following research objectives are centered on the purpose of the study:

1. Effectiveness of LSTM & XGBoost models in predicting the stock closing price in terms of performance and less erroneous.
2. Explore how deep learning models are significantly more viable than machine learning models.
3. Conduct hypothesis testing to determine the effectiveness of past 10-year stock trend in the forecast.
4. Impact of prediction models on the recommendation for traders to buy/sell stock.
5. To study the volatility of market pre and post of Demonetization and Covid-19 time period.
6. Inferences drawn out of volume traded over 3 years, 5 years and 10 years and between VWAP and Closing Price.

1.4 Research Questions

1. To what extent are the LSTM (Deep Learning) and XGBoost (Machine Learning) models able to predict the closing price of stocks in terms of performance and are they less flawed?
2. To what extent do the DL models exceed the ML models for forecasting?
3. What hypotheses can we make based on the trend of stocks over a decade?
4. How better the prediction models can recommend traders to buy/sell stocks of a particular company?
5. To what extent the share markets are affected by pre and post of Demonetization and Covid-19 time period?
6. What inferences can we set through volume traded over 3 years, 5 years and 10 years timeline and between VWAP and Closing Price?

1.5 Scope of the Study

"Predicting stock market returns is a challenging task due to consistently changing stock values which are dependent on multiple parameters which form complex patterns" (Vijh *et al.*, 2020). The data obtained consists of only a few features that aren't much adequate in the real world. We are predicting stock based on Volume Weighted Average Price, Close Price and Moving Averages. In the near future, we can include various factors/features like investor's sentiment data, financial feeds, profit and loss statements. Also, we are determining the trend of the stock and the recommendation for traders to buy/sell the stock of a particular company. Hereafter, we can even show possible common indicators in order to estimate a stock's future direction and market psychology. In the scope of the study, we can extend this model by feeding external factors, such as data affecting the stock prices and even broadcast, bulletin & opinions of renowned personalities. We can even create multi-model using different ML and DL models to provide a better precision rate.

1.6 Significance of the Study

This study is important to explain how deep & machine learning approaches are nowadays much helpful in predicting the stock prices and how it impacts the traders, especially intraday traders. It also shows the comparison between the LSTM & XGBoost model related to performance. This study would help in verifying how many better results can be obtained via fewer features like close price, spread (High-Low, Open-Close), VWAP and moving averages. This study will reveal the inferences drawn from analysis on volume traded over 3 years, 5

years, 10 years timeframe. Also, it will look out for which features are better for prediction out of historical dataset used.

1.7 Structure of the Study

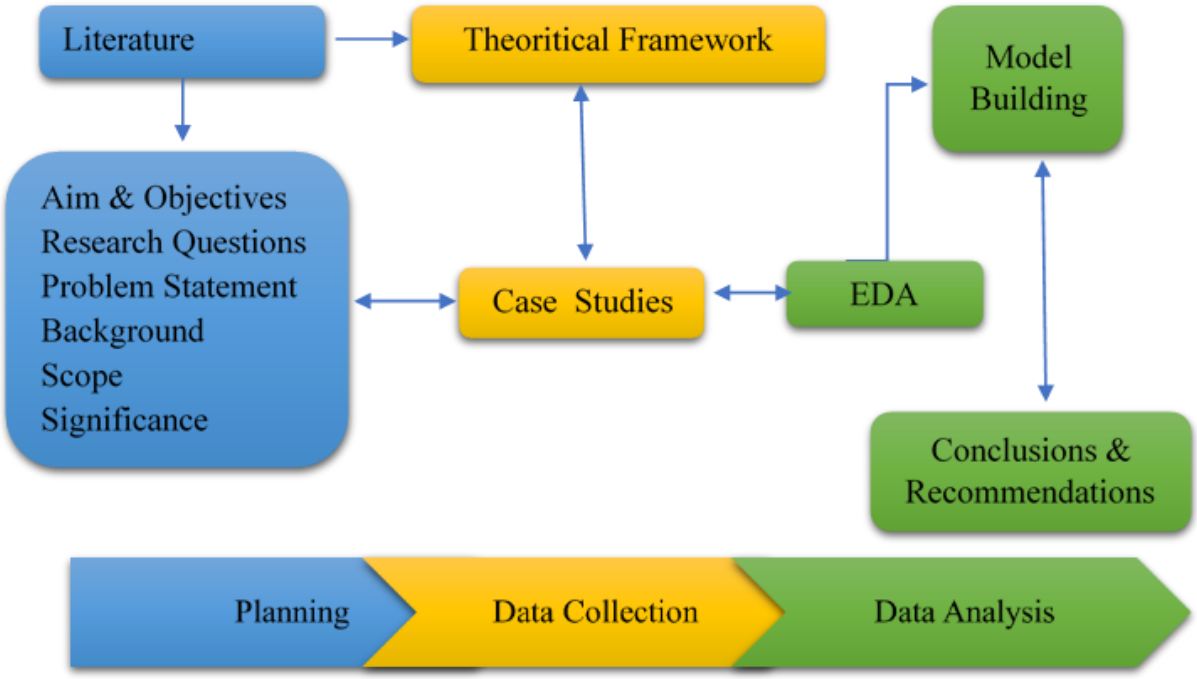


Figure 1.1 Structure of the study

Figure 1.1, depicts the thorough framework of the research work. Here, Chapter 1 and Chapter 2 belong to the research planning part of the study, Chapter 3 comprises collection of data and proposed model and evaluation technique, Chapter 4 and Chapter 5 collectively comprises of model building and evaluation results, finally, Chapter 6 comprised of conclusions and recommendations. To detailed additionally:

1. Chapter 1: Focus primarily on the context of the study, identify the problem statement, define the goal and objectives and research questions, categorize the scope of the research and the potential importance of the study.
2. Chapter 2: Primarily covers the literature survey of the research subject. Encompass Stock market theory, different machine learning models used in the past, challenges and related research publications.
3. Chapter 3: Data collection and description of features, proposed models and algorithm, evaluation, flow diagram and expected outcomes and resources or requirements.

4. Chapter 4: Data mining and analysis phase using exploratory data analysis package, final models implementation phase and fine tuning of architectures.
5. Chapter 5: Results and discussion phase using model outputs and evaluation metrics. Also, it includes the model validation phase along with the interpretation of visualizations.
6. Chapter 6: Conclusions and recommendations phase along with thesis contributions to new knowledge and further future works on the research paper.

CHAPTER 2

LITERATURE REVIEW

This chapter entails the work done on topics which are similar to the current research topic, bringing out where the gaps in the literature are, and how my research helps to fill in one or more of these gaps. It includes peer-reviewed articles, books, dissertations and conference papers.

2.1 Introduction

Here, we are discussing a topic related to the prediction of a company's stock future close price. Stock market traditionally is predicted using fundamental or technical analysis. But, that's not enough to accurately forecast the stock price. Recently, in the last decade, a lot of paper's have been published related to stock forecasting using machine learning or deep learning techniques. The papers mostly used the Yahoo finance dataset and algorithms like RF, SVM, LSTM, RNN, CNN, ANN and time series of models. Some of the papers also used hybrid models of deep learning algorithms. Even a few of the papers discussed sentimental analysis using news feed, opinions fetched from social media platform like Twitter, balance sheets and profit and loss statements.

Firstly, in this chapter we comprehend the business knowledge regarding the stock market and how the stock market/ exchange works and about stocks and historical data generated from those stock trades by traders on a day-to-day basis. Also, the following chapter covers the presence of data analytics in the stock market and the benefit of it.

In this chapter. It also discusses the predictive modelling on similar datasets and the different approaches for the model building and machine learning or deep learning algorithms used for the research purpose. It also mentions the comparison between the models used and their accuracy on validation or test data.

We have also set apart a section that mentions the challenges in the stock market and provide in detail discussion on the same. We do have the section on comparison with past literature reviews and also, lastly, we have a section comprising related research works or publications listing the latest three-year papers on the stock prediction.

2.2 Business Knowledge – Stock Market

All companies need money to run their business. Sometimes the profit acquired from selling goods and services is not sufficient to meet the working capital requirements. And so, companies invite normal people like you and me to put some money in their company so that they can run it efficiently and, in return, investors get a share of whatever profit they make.

The stock market is where people buy/sell shares of publicly listed companies. Trade in stock markets means the transfer (in exchange for money) of a stock or security from a seller to buyer. It offers an exchange trading platform to facilitate seamless trade of shares.

"A stock market is similar to a share market. The key difference is that a stock market helps you trade financial instruments like bonds, mutual funds, derivatives as well as shares of companies. A share market only allows trading of shares" (KOTAKSECURITIES.COM, 2021, para.2).

Shares are a way to own a part of the company's value. In proportion to the capital you invest, you can get ownership rights to a certain percentage in the company. Say you own 2% of the shares being traded in the market, you can say you have 2% ownership in the company. Hence, shares are units of ownership in the company and its financial assets. Shares are also known as stocks, equity, scrips etc. After purchasing them, you will be known as a stockholder or a shareholder of the company.

The main source is the stock exchange, the central platform which offers the facilities used for trading equities of enterprises and other securities. Shares may not be purchased or sold unless it is listed on a stock exchange. Thus, it is the assembly place of the stock buyers and sellers. India's primary stock exchanges are the Bombay Stock Exchange and the National Stock Exchange.

Here are two varieties of share markets, primary and second markets. The primary market is where a company gets a license to issue a certain number of shares and increase its income. This is also referred to as obtaining listed on a stock exchange. If the company sells shares for the initial, this is referred to as a public offering.

The shares sold in primary markets are the shares that are traded in the secondary market. It's about granting investors a possibility of coming out of a stake and selling the shares. Secondary market dealings are where an investor purchase shares from another investor at the current market price or at the price agreed by both parties. Normally, investors do this through a broker who ease the process.

2.3 Data Analytics in Stock Market

Data analysis has been defined as the approach chosen in the random data analysis and process to facilitate understanding. The companies accumulate the wealth of quantitative and qualitative information. These data can be very useful if they are properly reviewed and read to provide useful information and results.

Stock market trading has demanded accurate and timely inputs. The significance of data generated within the stock market on a daily basis is impracticable to be handled, evaluated and made sense of by human beings due to the large quantity of data generated and the rate at which this financial data is being produced from diverse sources.

Subsequently, data analytics can be used in the stock market towards recognition of stocks and shares with growth possibility, buying them at low prices and selling them when the share prices are sky-high. Data analysis makes it possible for people to find the best stocks to buy today and get short-term profits.

In this digital age, all stock trading is done in line with demat accounts. One of the benefits of using a demat account is it shows all the past securities or financial transaction history of the user and the past performance and returns of a stock or a share. Using data analysis, one can explore and uncover trends and the impact which determines traders' attitude towards buy/sell decisions which can help investors in the stock market to keep healthy investment and listed companies to take major judgment relating to their service or product, distribution, costing, advertising etc. to gain a stance of market advantage.

Today, financial analysis alone is not enough to look at stock prices and the behavior of stock prices. These financial analyses have to be incorporated with external factors such as the social and economic trends of the economy, the policy context, consumer attitude and choice, etc., that have the ability to have an impact on the price of equities of a specific stock or on the price of equities of a specific sector. Data analysis may use forecasting models to guess projected results and yield on stake. Rise in access to these outcomes and in data analysis precision level, investors can influence this information and forecast towards relieving their risks involved with stock market trading.

The robustness of algorithmic trading is within its illimitable proficiency of analyzing data, making real-time investing decisions and performing trades at a rapid pace and high rate using a broad array of organized and unorganized data acquired from varied sources such as stock market data, financial feeds on social platforms, latest information etc. in the direction of making spontaneous opinion. This analysis of circumstantial sentiments can be immensely worthwhile in stock market trading.

2.4 Predictive Modeling in Stock Market Prediction on Similar Datasets

Predictive modeling is the overall idea of constructing a predictive model that can make predictions. Typically, a similar model encompasses a machine learning algorithm that discovers a certain characteristic from a set of training data with a view to make those predictions.

Predictive modelling can be broken down into two sub-fields: regression and pattern classification. Regression models are determined on the examination of association between variables and trends with a view to make predictions about continuous variables, e.g., the prediction of the maximal heat for the forthcoming days in climate foretelling.

2.4.1 Machine Learning Approach

Many of the papers have used the machine learning techiniques to forecast the stock prices. Few papers even used machine learning for comparative analysis. The papers (Parmar *et al.*, 2018; Umer, Awais and Muzammul, 2019) have used Linear Regression algorithm for building basic model for predicting stock. Some of the papers (N P Samarth, Gowtham V Bhat and Hema N, 2019; Nikou, Mansourfar and Bagherzadeh, 2019) used Random Forest for building predictor using ensembles and decision trees to achieve better results with the help of fine-tuning the architecture. Even Random Forest is used for comparative analysis in some papers like (Vijh *et al.*, 2020; Ghosh, Neufeld and Sahoo, 2021) where in former paper we compare RF with hybrid ANN model, while in the latter one we have compared the prediction accuracy of Random Forest with LSTM model. In some other papers like (Nikou, Mansourfar and Bagherzadeh, 2019; Madhusudan, 2020) using machine learning techniques we can see Support Vector Regression (SVR) been used. In the former paper we can see SVR method has higher precision than neural networks and RF, while in the latter paper we can see that Radial Basis Function (RBF) SVR kernel method proves to be the best for prediction of stock. In one of the paper (Henrique, Sobreiro and Kimura, 2018) where Support Vector Regression (SVR) is used to predict stock prices for large and small capitalisations and in three different markets, employing prices with both daily and up-to-the-minute frequencies.

The machine learning algorithms used in past papers are mostly Random Forest, Support Vector Machine, Linear Regression. This machine learning techinques are even tuned for better results or in some papers the researcher have used this models for comparative analysis. In some papers, ensembled models are used that seems to have gone too deep to predict the probability. These stacked models combine every individual weak model and generate a best predictive model in the machine learning techniques.

2.4.2 Deep Learning Approach

Deep Learning is nowadays more well known modeling technique for prediction problems. Some literature provides a deep learning framework to forecast the direction of price movement based on historical data from financial time series. Deep learning is a subset of machine learning in artificial intelligence that has networks capable of learning unsupervised from data that is unstructured or unlabeled. Also known as deep neural learning or deep neural network.

In some of the papers like (Nikou, Mansourfar and Bagherzadeh, 2019; Dhenuvakonda, Anandan and Kumar, 2020; Vijh *et al.*, 2020) were ANN is used for building the prediction model and in the above mentioned second paper we can see that ANN model proves to be better than time series model like AR, ARIMA. In one of the paper (Selvamuthu, Kumar and Mishra, 2019) were ANN is used with three different algorithms and companies tick data and the 15 minute dataset. All three algorithms provide 99.9% of accuracy but when evaluated on 15 minute data the accuracy drops to 96.2%, 97.0% and 98.9% for Levenberg-Marquardt (LM), Scaled Conjugate Gradient (SCG) and Bayesian Regularization respectively which is significantly poor in comparison.

In other papers (Kamalov, Smail and Gurrib, 2020) were CNN model is used with 10 years historical data and 60 financial indicators with top accuracy rate of 56.21%. In one of the other paper (Hiransha *et al.*, 2018) were CNN is compared with other neural network models such as LSTM, RNN and Multilayer Perceptron (MLP) for stock prediction based on historical prices available and it has been observed that CNN is outperforming the other models.

In other papers we can see the LSTM models been used for building stock predictor. (Hiransha *et al.*, 2018; Parmar *et al.*, 2018; Kanade, 2020; Moghar and Hamiche, 2020) papers have used LSTM model as a implemented technique for future stock prediction. (Shen and Shafiq, 2020) paper have combined comprehensive feature engineering with LSTM to perform prediction whereas in (Das *et al.*, 2018) LSTM is used for real time sentiment analysis of Twitter streaming data for stock prediction. In some other papers like (Nandakumar *et al.*, 2018; Nikou, Mansourfar and Bagherzadeh, 2019; Dhenuvakonda, Anandan and Kumar, 2020) LSTM is used for comparative analysis with other neural networks. (Ghosh, Neufeld and Sahoo, 2021) paper used LSTM model with final accuracy of 69.67%.

In some of the papers we can see hybrid model of LSTM or LSTM deep recurrent neural network like (Ding and Qin, 2020) paper were deep neural network is combined with LSTM to provide accuracy of 95% whereas in other paper (Yu and Yan, 2020) a Deep Neural Network (DNN) based prediction model is designed based on the Phase Space Reconstruction (PSR) method and a LSTM for DL and used to predict stock prices. A comparison of the results shows

that the proposed prediction model has higher prediction accuracy. (Chung and Shin, 2018) paper proposed a hybrid approach integrating LSTM network and genetic algorithm (GA). The experimental result demonstrates that the hybrid model of LSTM network and GA outperforms the benchmark model.

Past related papers introduced ANN, RNN, and CNN on Multivariate timeseries. The deep learning models proved to be more efficient in terms of accuracy in predicting the stock prices as compared to machine learning models.

2.4.3 Other Approaches

In the past reseach papers, time series models were also introduced such as Auto Regressive model, Auto Regressive Integrated Moving Average model and Three Months Moving Average (3MMA). In one of the paper (Umer, Awais and Muzammul, 2019), Exponential smoothing method is used which obtained hypothesis that prediction results given are less error and with greater accuracy and we considered it as best stock market predictor with general trend analysis. In the same paper even 3MMA model is used for stock prediction but former one proves to be better in terms of accuracy. (Dhenuvakonda, Anandan and Kumar, 2020) paper used time series baseline models such as Auto Regressive (AR) and Auto Regressive Integrated Moving Average (ARIMA) for building stock predictor and also for comparative analysis with neural network models. (Xu and Cohen, 2018) paper even used ARIMA technique along with other models.

2.5 Challenges in Stock Market

Stock market prediction is a major challenge owing to non-stationary, blaring, and chaotic data, and thus, the prediction becomes challenging among the investors to invest the money for making profits. The core of equity investments is high risk and high profit, making them attractive to many companies, investors and economists. Traditionally, the assessment of the growth of the enterprise depends on the prediction of profits and cash flows by means of an appropriate discount rate of cash flows to arrive at the value of the enterprise. However, this traditional prediction of earnings is only possible when a company has positive profits, comparable companies or a long history of performance. Furthermore, stock market data are subject to external factors such as natural calamity and policy-making judgment; as a result, they are naturally noisy and unstable. The uncertainty of the stock data is also due to the insufficient data from the former behavior of the stock market to authorize capturing the dependency between forthcoming and preceding prices. Insufficient data on the stock market

is frequently considered a noisy feature, which complicates forecasting the later price of a stock. Speedy build-up in trading and investment is resulting in a double need for suitable tools and techniques to ease risks and enhance earnings.

2.6 Comparison of Literature Review

Many of the papers published used combination of deep learning and machine learning models. Some even tuned the basic available model and formed hybrid models like LSTM Hybrid model (Chung and Shin, 2018; Yu and Yan, 2020) or ANN Hybrid Model (Selvamuthu, Kumar and Mishra, 2019; Vijh *et al.*, 2020). Most of the researchers used deep learning models like CNN, LSTM, RNN together in papers like (Hiransha *et al.*, 2018; Nikou, Mansourfar and Bagherzadeh, 2019; Kanade, 2020). In most of the papers (Parmar *et al.*, 2018; Vijh *et al.*, 2020; Ghosh, Neufeld and Sahoo, 2021) it's been proved that deep learning models provides better accurate prediction results than machine learning models. The LSTM based Deep Recurrent Neural Network (DRNN) neural network model (Ding and Qin, 2020) provides average accuracy of each predicted value over 95%. Some of the papers (Umer, Awais and Muzammul, 2019; Dhenuvakonda, Anandan and Kumar, 2020) used time series models such as AR, ARIMA, 3MMA which does not prove to be better than deep learning models. Some of the other papers (Henrique, Sobreiro and Kimura, 2018; N P Samarth, Gowtham V Bhat and Hema N, 2019; Nikou, Mansourfar and Bagherzadeh, 2019; Madhusudan, 2020) used Random Forest and SVR models from which SVR proved to be best predictive machine learning model because it predicted the stock closing prices closest comparing to the original value of same days. Some of the papers (Parmar *et al.*, 2018; Umer, Awais and Muzammul, 2019) even used Linear Regression for stock prediction system.

2.7 Related Research Publications

The 2019 paper "Stock Closing Price Prediction using Machine Learning Techniques" (Vijh *et al.*, 2020) is one of the base paper related to my research work where ANN + RF model is implemented. Another 2018 paper "Stock Market Prediction Using Machine Learning" by (Parmar *et al.*, 2018) too uses Yahoo finance dataset where they used Linear Regression along with LSTM have shown an improvement in the accuracy of predictions, thereby yielding positive results. In another 2018 paper "Stock Price Prediction Using Long Short Term Memory" (Nandakumar *et al.*, 2018) wherein LSTM model is used for better accuracy even when size of the data increases. With more data, more patterns can be fleshed out by the model, and the weights of the layers can be better adjusted. In a recent 2020 paper "Stock Closing Price

Prediction using Machine Learning SVM Model" (Madhusudan, 2020) SVM model is used and the outcome of the paper states that RBF SVR kernel is the best method for prediction. In one more recent paper of 2020 "Machine Learning Model For Stock Market Prediction" (Kanade, 2020), proposed the utilization of the data collected from financial markets with ML algorithms in order to predict the stock price fluctuations. They used stock closing price to predict the stock and news heading. LSTM technique is used in this paper. Also, It uses Sentiment analysis of financial news and opinions fetched from social media platform like Twitter. Each of these paper's used different models for achieving better results and even stacked/combine models for comparing or creating highly precisive accurate model. In some other papers (Chung and Shin, 2018; Das *et al.*, 2018; Hiransha *et al.*, 2018; Moghar and Hamiche, 2020; Yu and Yan, 2020) LSTM models are used based on RNN or GA (Genetic Algorithm), or DNN to build the optimal model for better accuracy of predicted values.

2.8 Summary

This chapter wholly comprises past papers and the related research work which encompass different papers from different authors. These chapters shows the progressive work in the field of stock market and increasing demand of ML or DL methods for analysing the large stock data and also forecasting the trend and prices of the stock. It also justifies the usage of the data science in this stock market field to an extensive level by the financial professionals such as investors, brokers or advisory institutions.

In this chapter we have seen some published research or conference papers that are related to the subject and topic of the research work I am conducting here. These chapters have given more insight on the past paper's which will help us to find research gaps on which we can work. Also, this chapter put some lights on the machine learning or deep learning models used before in the related research work.

It has papers using similar datasets or custom datasets that contains the financial data, balance sheets, news feed or social apps data regarding finance and stock market. The chapter has even provided business knowledge on the stock market and challenges faced for stock prediction. It also described the data analytics in the stock market. In the last sections of the chapter, we have done a comparison of Literature Review, and there is also a discrete section that comprises of the related research publications from which some papers are used as base papers of this conducted research work.

Finally, we close this chapter on a good note that why it is so necessary to conduct research work in the stock market field and the importance of this research for the people surrounded by

the stock finance domain. Also, this chapter supports the work to be conducted as part of this research.

CHAPTER 3

RESEARCH METHODOLOGY

This chapter includes details on the dataset and models used for forecasting. It also unveils the data pre-processing and transformation techniques along with expected outcomes, logical flow design and tools used in this approach.

3.1 Introduction

This chapter gives an outline of the research methods that were followed in the study. The purpose of this chapter is to introduce the research methodology for this qualitative-ground theory study regarding the usefulness of machine learning in stock market prediction. The applicability of grounded theory and a constructivist approach for this study are discussed in-depth in this chapter.

This chapter also discusses the various tools and techniques used to conduct the research study. The approach, methodology, techniques, tools, research framework, data samples, data collection and evaluation techniques are some of the elements discussed in this section. It provides information on the data, that is, past stock closing prices of a listed company, data pre-processing, data transformation and how they were sampled. The researcher describes the research design that was chosen for the purpose of this study and the reasons for this choice. The instrument that was used for data collection is also described, and the procedures that were followed to carry out this study are included. The researcher also discusses the methods used to analyze the data. Lastly, the evaluation techniques that were followed in the process are also discussed.

The chapter begins with the data collection with extensive details on data such as the source of the data along with the time interval of the dataset. Dataset features are also described and their representation in the share market. It also mentions how the data would be interpreted, and the normalization technique used on the data. This chapter also highlights the main data attributes that will be used for predicting the model. It also talks about how the data is divided for testing and training purposes.

This chapter also mentions the new data variables created for training the model and prediction of the stock closing price. It also broadcast the proposed methods and what will be the input and output and how those models will help to solve your problem and why those models are

chosen. This chapter also portrays the logical design of the flow of the system which will outline the system along with inputs/outputs and flow direction along with the final result outputted. This chapter also brings up the evaluation techniques used in the process. It underlines the expected outcomes from the model built for the prediction of stock prices. This chapter also mentions the tools used in the data analysis and model building along with the minimum hardware resources required.

3.2 Research Methodology

This section describes the logical flow of the system and then the dataset and its features in detail. It also mentions the data pre-processing techniques used on the raw data for the research work. It even declares some more newly created variables out of raw data. This section also mentions the splitting of data for testing and training the model and important column/features that will be used for creating a stock price predicting model.

3.2.1 Logical flow of the system

Figure 3.1 shows the logical flow of the system from the starting phase of data collection to the creation of the final model. Following are the points that discuss the flow of the system in detail:

1. The starting phase is the collection of the data from the source, loading and reading the data in the acceptable format.
2. Now, the collected data is pre-processed by the following steps: data cleaning, data analysis, data transformation, creation of new variables, and normalization of data.
3. Now, the feature selection is carried out through dimensionality reduction of the dataset to remove unwanted features before passing it to the model.
4. After completing the above steps, the learning algorithm is configured with basic settings and the training data is passed to the model to learn and predict using historical data, and it's features and also to validate the model using validation data.
5. Next, we try to make predictions using test data and then evaluate results. If the results are not satisfying, then we again make some configuration changes or update hyper-parameters value to again build the model and evaluate it. Identical steps are carried on other learning models too.
6. Once we achieve good results, we considered that model the final model and made final predictions based on that model for outcomes.

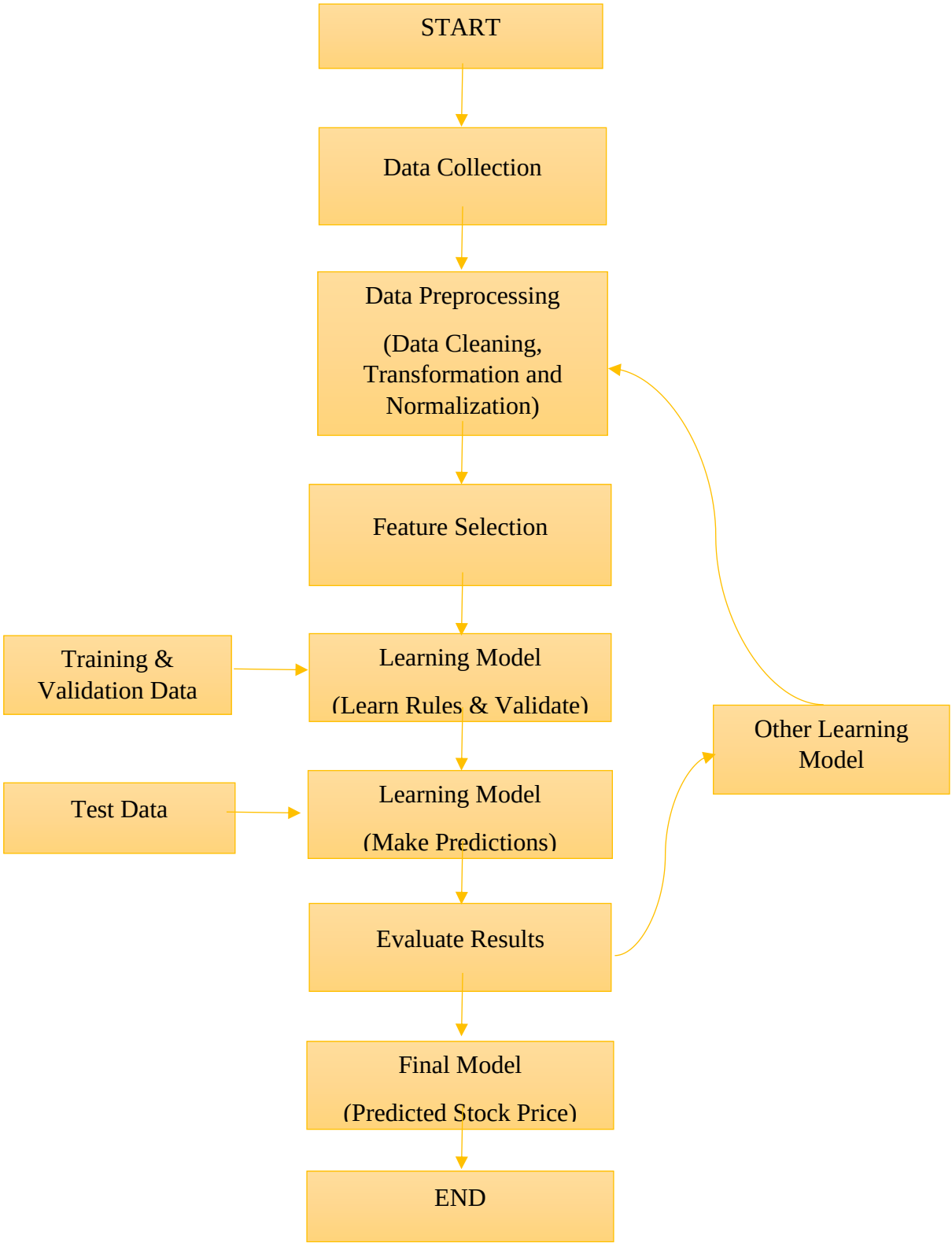


Figure 3.1 Flow chart of system

3.2.2 Data Description

The historical dataset used in this research is collected from BSE Historical Stock Prices Data (*BSE Stocks Price*, 2021). The time interval of the dataset is from 01/01/2011 to 01/03/2021. The dataset is collected for five companies from different sectors – Infosys Ltd, Glenmark Pharmaceuticals Ltd, Indiabulls Housing Finance Ltd, Maruti Suzuki India Ltd, Adani Ports and Special Economic Zone Ltd. A total of 12 attributes are present in the dataset obtained. Important columns/features that will be used for creating stock price predicting model – Close Price, WAP (Volume Weighted Average Price), No. Of Trades, Spread (Close-Open, High-Low). Table 3.1 shows the features available in the dataset along with their brief description.

Table 3.1 Description of dataset features

Dataset Features	**Description**
Open Price	It is the price at which the security first traded at the open of the day's trading on its stock exchange
Close Price	Final price for a day the stock exchanged.
High Price	The highest closing value of a stock over the past 52 weeks
Low Price	The lowest price at which a security trades on a given trading day.
WAP	It measures the mean price of the stock exchanged for that day.
No.of Shares	It measures the number of shares traded in a stock
No.of Trades	It measures the number of trades throughout the day
Total Turnover (Rs.)	It measures the overall quantity of shares exchanged considering their value
Deliverable Quantity	It is the number of shares that move from one set of people to another set of people
Deliverable Qty to Traded Qty (%)	It measures the deliverable quantity w.r.t. traded quantity
Spread (H-L)	It is a difference between high-low of stock for that day
Spread (C-O)	It is a difference between close-open of stock for that day

3.2.3 Data Preprocessing

The dataset at hand is in a CSV structure that will be interpreted and transformed into a data frame using the python pandas tool. From this, the unwanted columns/features are dropped from the data frame such as Open Price, High Price, Low Price, No. Of Shares, No. Of Trades, Total Turnover (Rs.), Deliverable Quantity, and % Deli. Qty to Traded Qty. After that, the dataset will be looked at to identify and correctly handle the missing values. We will drop those rows wherever missing values are present.

Once the dataset is cleaned, the data is normalized using "MinMaxScaler" from the "sklearn" library in Python. It converts every available feature in a specific range such as [0, 1] or [-1, 1]. To preserve the zero's in a sparse dataset, "MinMaxScaler" is a good option. Finally, the dataset is split into three different datasets – training, validation, testing. The training dataset was kept at 70% of the available dataset and the testing and validation dataset was kept at 15% each from the available dataset.

3.2.4 Data Transformation

For training the model and prediction of the stock closing price, we will create four more variables. Four of the new variables are moving averages which are used mostly by stock professionals for predicting stock price moving likely uptrend or downtrend. The four moving average variables are as below:

1. Stock price's 20 days' moving average (20 DMA)
2. Stock price's 50 days' moving average (50 DMA)
3. Stock price's 100 days' moving average (100 DMA)
4. Stock price's 200 days' moving average (200 DMA)

3.3 Proposed Methods

This section encompasses a detailed description on how you will carry out your research. It includes your research sketch, methodology and procedures that you envisage exploiting. Furthermore, the activities that you plan to carry out to complete your project and the blueprint of the work. The proposed techniques will also describe the inputs and outputs of the model and provide a rationale for why the following models are used in the research.

3.3.1 LSTM

Long Short Term Memory is competent enough to learn series dependence in forecasting problems. It's a special kind of RNN. This is accomplished on account of a recurring module of the model that has a blend of four layers interacting with everyone. The model learns unlearns and preserves details from all units using the cell state & 3 gates. The cell state in LSTM helps the details to flow from all units without being revised. The forget gate can modify the cell state and the input gate can adjust information inside the cell state. Every unit has an input, output and a forget gate which can override the details in the cell state. Using the sigmoid method, the forget gate can decide which facts from the preceding cell state should be wiped out. The current cell state information is controlled from the input gate using a point-wise product of 'sigmoid' & 'tanh'. Finally, the information proceeded on to the next hidden state, which is handled by the output state.

In the proposed research, the model functions using four fundamental layers. Figure 3.2 shows the architecture of the proposed research LSTM model. It comprises an input layer, LSTM layer that takes the sequence from the previous layer, followed by a dense layer with 5 neurons and then final dense layer that outputs the predicted value The input layer consists of new variables that include Close Price, Spread (H-L), Spread (C-O), and 20 DMA, 50 DMA, 100 DMA, 200 DMA, and Volume Weighted Average Price. The weight regularization technique used for imposing constraints on the input weights within the LSTM memory cell. The hidden layer or activation layer comprised of these LSTM memory cells. Each LSTM cell has its input and output streams. In each LSTM cell, the computation takes place between the functions "sigmoid" and "tanh" by multiplication or addition between two vectors/matrices. The output layer is composed of a single LSTM cell that will yield the expected value in terms of the closing price of the stock. The optimizer used here would be Adam for the LSTM network.

The proposed model would help us in intraday trading by giving close to accurate price of the stock on a daily basis, and it is also one of the best models with good performance and less computational costs. LSTMs are very powerful in sequence prediction problems because they're able to store past information and also solve the problem of vanishing and exploding gradients. This is important in our case because the previous price of a stock is crucial in predicting its future price.

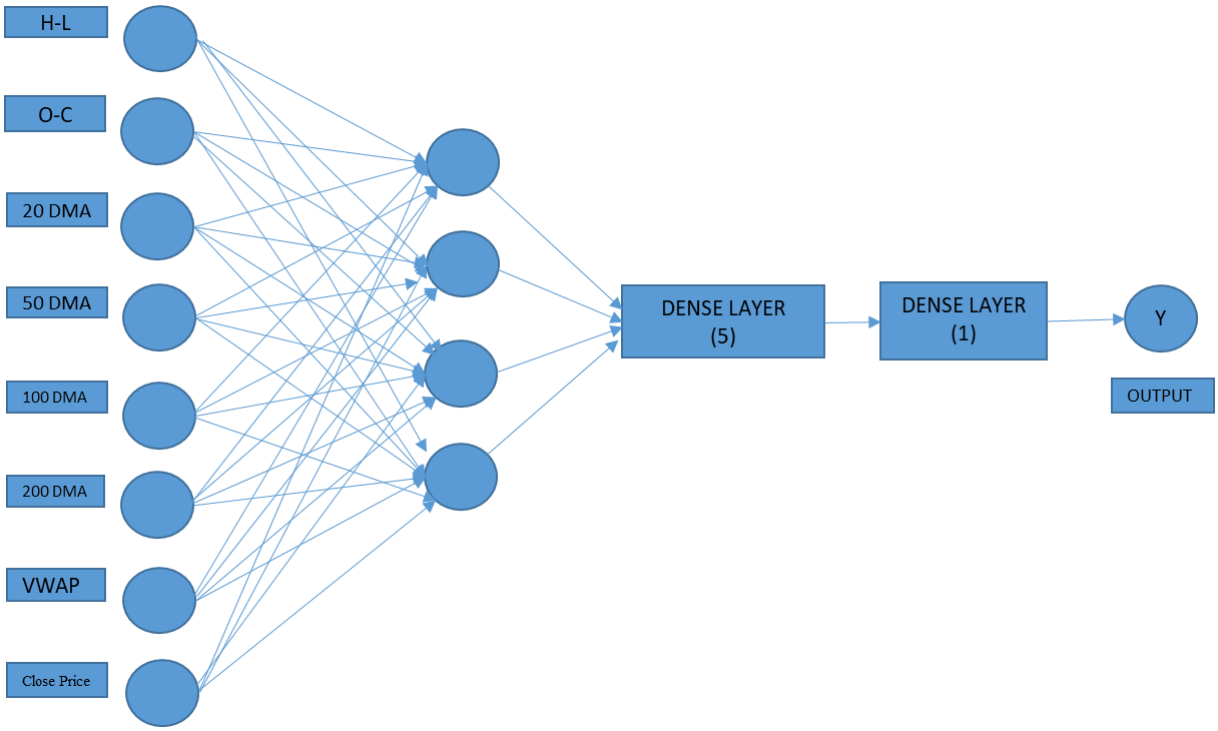


Figure 3.2 Architecture of long short term memory model

3.3.2 XGBoost

"XGBoost is short for Extreme Gradient Boosting and is an efficient implementation of the stochastic gradient boosting machine learning algorithm" (Jason Brownlee, 2020). For regression and classification, XGBoost is preferably better for gradient boosting implementation. It is speedy and efficient when compared to other ML predictive models. XGBoost can be utilized for time series prediction by altering the data set into supervised learning. It uses a peculiar method for model evaluation named walk-forward validation. It's a decision trees ensemble where novel-tree repairs faults already present in existing trees. Until we reach a point where no enhancement can be done, we add up trees. XGBoost bestow hyper-parameters to improvise the accuracy in model building. In terms of performance, XGBoost runs faster compared to other Decision Tree machine learning models. For predicting the stock closing price for a particular company, we would be tuning the hyper-parameters for better results.

In this proposed model, the following features – Spread (H-L), Spread (C-O), and 20 DMA, 50 DMA, 100 DMA, 200 DMA, and Volume Weighted Average Price for the training of each weak ensemble model which in turn combines them to produce improved results. Ensemble methods usually produce more accurate solutions than a single model would. These models, when used as inputs of ensemble methods, are called "base models". The noise in the stock data

would be much higher because of many factors, and that would cause the model to provide lower accuracy. It aims at minimizing forecasting error by hyper-tuning some parameters that would in turn optimize the final output, which would be the closing price of the stock. For XGBoost, there are several hyper-parameters that are tuned including n_estimators, max_depth, learning_rate, gamma and random state for our model.

The XGBoost model is both fast and efficient, performing well, if not the best, on a wide range of predictive modeling tasks. XGBoost is chosen because it is a gradient boosting model which uses a gradient descent algorithm to minimize the loss when adding new models. This helps to get the best model for prediction with minimized error.

3.4 Evaluation Techniques

The evaluation of the models is intended to assess the precision of the generalization of a model for future data (not observed/not disclosed). The techniques used to assess the performance of a model are divided into 2 types: holdout and cross-validation. Both approaches employ a set of tests (i.e., unknown data) to measure, model performance. Model evaluation metrics are required to quantify model performance. The choice of evaluation metrics depends on a given machine learning task. Different evaluation metrics are used for different kinds of problems. A significant characteristic of the assessed parameters is their ability to distinguish outcomes from the model. The mere fact that we create a predictive model is not our motive. This involves developing and choosing a model that is extremely precise from sample data. Hence, it is vital to watch the accuracy of your model precedence to computing predicted values. The following evaluation metrics are used in this research – R-Squared, Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE) and Mean Bias Error (MBE).

3.4.1 R-Squared or Coefficient of Determination

Coefficient of determination also called R-Squared (R2) score is used to evaluate the performance of a linear regression model. It is the amount of the variation in the output dependent attribute which is predictable from the input independent variable (s). It is used to check how well-observed results are reproduced by the model, depending on the ratio of total deviation of results described by the model.

$$R2 = 1 - SS_{res} \,/\, SS_{tot} \tag{3.1}$$

Where in Equation 3.1,

SS_{res} is the sum of squares of the residual errors.

SS_{tot} is the total sum of the errors.

We can import r2_score from sklearn.metrics in Python to compute R2 score. The best possible score is 1, which is obtained when the predicted values are the same as the actual values.

3.4.2 MAPE or Mean Absolute Percentage Error

The mean absolute percentage error (MAPE) is a measure of how accurate a forecast system is. MAPE can be considered as a loss function to define the error termed by the model evaluation. Using MAPE, we can estimate the accuracy in terms of the differences in the actual vs estimated values.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \quad (3.2)$$

As explained in Equation 3.2, we initially calculate the absolute difference between the Actual Value (A) and the Estimated/Forecast value (F). Furthermore, we apply the mean function on the result to get the MAPE value. MAPE can also be expressed in terms of percentage. The lower the MAPE, the better the model.

Figure 3.3 shows that the Python sklearn library offers us with a mean_absolute_percentage_error() function to calculate the MAPE value.

```
from sklearn.metrics import mean_absolute_percentage_error


# Evaluating predicted values
Y_actual = [1,2,3,4,5]
Y_Predicted = [1,2.5,3,4.1,4.9]
mape = mean_absolute_percentage_error (Y_actual, Y_Predicted)
```

Figure 3.3 Mean absolute percent error code snippet

MAPE output is non-negative floating point. The best value is 0.0. But note the fact that bad predictions can lead to arbitrarily large MAPE values, especially if some y_true values are very close to zero.

3.4.3 RMSE or Root Mean Squared Error

RMSE is the standard deviation of the errors which occur when a forecasting is performed on a data record. It is similar to mean squared error, but the mathematical root is calculated of the resulted term while deciding the model precision. Importantly, the square root of the error is calculated, which means that the units of the RMSE are the same as the authentic units of the target value that is being projected.

Use RMSE if you want to:

1. Penalize large errors.
2. Have the result be in the same units as the outcome variable.
3. Use a loss function for validation that can be quickly computed.

```
from sklearn.metrics import mean_squared_error
from math import sqrt

actual_values = [3, -0.5, 2, 7]
predicted_values = [2.5, 0.0, 2, 8]

mean_squared_error(actual_values, predicted_values)
# taking root of mean squared error
root_mean_squared_error = sqrt(mean_squared_error)
```

Figure 3.4 Root mean squared error code snippet

Figure 3.4 shows how to calculate RMSE using a mean squared error function in the sklearn python library. A fine-tune RMSE term is 0.0, in the sense that forecasted values matched the actual values accurately. A better RMSE is proportional to your particular data set.

3.4.4 MBE or Mean Bias Error

Mean bias error is primarily used to estimate the average bias in the model and to decide if any steps need to be taken to correct the model bias. Mean Bias Error (MBE) captures the average bias in the prediction. A positive bias or error in a variable represents the data from datasets is overestimated and vice versa. The lower values of errors and considerably higher value of correlation coefficient for the variable and direction are of greater importance.

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i) \quad (3.3)$$

Where in Equation 3.3, Oi is the observation value and Pi is the forecast value

3.5 Outcomes

1. Comparison of evaluation outcome in-between LSTM and XGBoost models w.r.t accuracy, R-Squared, MBE, MAPE & RMSE.
2. Positive or negative trend direction of the stock price of the last decade in stock price trend analysis.
3. VWAP dataset feature is a good indicator for prediction of stock close price.
4. Recommendation to buy/sell the stock based on the predicted price and trend.
5. Analysis on stock returns and increase/decrease in volume traded over the last decade.

3.6 Requirement/Resources

This section mentions the tools used for building the model for stock price prediction. It jot down the hardware and software minimum requirements. Also, it mentions application tools used to interpret the data collected and transformed to the format the model would be expecting as input. This section even mentions the python libraries used for exploratory data analysis and model training, testing and evaluation. In terms of hardware it mentions about the device used for model building and equipment requirements such as Random Access Memory (RAM), Central Processing Unit (CPU) and Graphics Processing Unit (GPU). This section provides all the requisite resources that will make it easier for some other researcher to recreate my research.

3.6.1 Hardware Requirements

1. Laptop or Desktop
2. 6+ Cores CPU with base speed more than 2.5 GHz
3. 32 GB RAM
4. K80 GPU

3.6.2 Software Requirements

1. Jupyter Notebook
2. Python 3.0
3. Pandas and Numpy
4. Matplotlib & Seaborn (Data Visualization Libraries)
5. Scikit-learn (Machine Learning Library)
6. Keras (Deep Learning Library)
7. Mplfinance (Matplotlib's finance API)

3.7 Summary

This chapter encompasses the research methodology used in the research work. It covers the dataset used for the research work and the techniques used to transform and interpret the data for finally using that data in the model building. This chapter also list down the data features and logical flow of the system that will help the researcher or any other to understand the process and dataset better in every possible way.

This chapter even throws light on the analysis of the data using exploratory data analysis and conduct different analysis on data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. It also shows the stock price trend across a decade for the selected companies. It also covers volume traded over the last decade for a stock. It even comprises the exploration on the difference between VWAP and Closing Price.

This chapter even discussed the proposed methods used – LSTM, XGBoost and the significance of both the machine and deep learning algorithms. This chapter also reveals the model building phase and the inputs and outputs through the final model for prediction purpose.

This chapter encloses the evaluation techniques used in the research methodology which are as follows – R-Squared, MBE, MAPE & RMSE. It gives a detailed explanation on the evaluation techniques used in research work and also list the basic code for using that evaluation method through the python library.

The research methodology chapter includes the expected outcomes and the requirements and resources used for the research work purpose. The outcomes mention the results expected or studied from the conducted research work.

CHAPTER 4

ANALYSIS

This chapter includes analysis of data and interactive visualizations conducted on five companies' stock dataset which is limited to the past 10-year stock data. Also, this chapter comprises model building steps and techniques along with hyper-parameter tuning for choosing a set of optimal hyper-parameters for a learning algorithm.

4.1 Introduction

This chapter outlines the interactive analysis on the data with the goal of discovering useful information, informing conclusions, and supporting decision-making, often using statistical graphics and other data visualization methods. It also discusses the different steps carried out on historical stock data before passing it to a training model as input which helps in extracting hidden patterns in data to provide valuable insights.

This chapter also contains discussion about the fine tuning of architecture using hyper-parameters in both the models while training for yielding optimal results and to minimize the predefined loss. A good choice of hyper-parameters can really make a model succeed in meeting the desired metric value or, on the contrary, it can lead to an unending cycle of continuous training and optimization.

This chapter further talk about the process of implementing the model which comprises of techniques used in model building for both the models along with the model building steps including the methods used, the number of layers, applied loss function, data split into train-test and validation data passed to regression model while training. It also discusses the learning curves which is a plot of model learning performance over experience or time. This helps in better understanding of the model.

This chapter overall discusses various things carried out on the historical data and model for achieving better and optimal results. It also gives in depth details on what all inferences we are able to extract from the data gathered from sources. Besides this, we are able to conclude some of our objectives and aims through the interactive visualizations. Also, it lists down various processes carried out and the model learning steps or methods along with the hyper-parameters used for building a favourable model.

4.2 Data Mining

Data mining is used in business to make better managerial decisions by: Automatic review of data, Extracting essence of information stored, Discovering patterns in raw data. Data Mining, also known as Knowledge Discovery in Databases, refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data stored in databases.

4.2.1 Data Inspection

Data Inspection is the act of viewing data for verification and debugging purposes, before, during or after the transformation. In this step we have inspected data to see the structure and dimensionality shape, the values in the stock dataset, and the data-types of every column in the dataset. We have seen that the data are of three types: date-time, float, and int. We have also seen that shape (rows, columns) of data is (2478, 13) for Adani Ports, Maruti Suzuki, Glenmark and Infosys or (1839, 13) for Indiabulls Housing Finance.

4.2.2 Data Cleaning

Data cleaning is defined as removal of noisy and irrelevant data from collection. Here, we have cleaned any rows with missing values. The missing values were almost less than 1% which is negligible and can be removed from the dataset before passing it to the next step of data mining.

4.2.3 Data Reduction

This technique is applied to obtain relevant data for analysis from the collection of data. The size of the representation is much smaller in volume while maintaining integrity. We have done dimensionality reduction which reduces the number of attributes in the dataset. The following attributes are removed (Open Price, High Price, Low Price, No. Of Shares, No. Of Trades, Total Turnover (Rs.), Deliverable Quantity, % Deli. Qty to Traded Qty.

4.2.4 Data Transformation

In this process, data is transformed into a form suitable for the data mining process. Data is consolidated so that the mining process is more efficient, and the patterns are easier to understand. In here we have used Normalization technique for scaling of data to fall within a smaller range. This is done to eliminate bias that occurs while computing distance between features whose values come in different ranges. We have used MinMaxScaler from sklearn

library to scale the data and for getting all the values in between the range 0 to 1. This helps in improving model performance and accuracy of prediction.

Also, here, new attributes are created from an existing set of attributes for visualization purpose or for passing it to a learning model as a feature. We have created Relative Strength Index (RSI) indicator feature to predict whether a stock is overbought/oversold, daily moving averages of 20 DMA, 50 DMA, 100 DMA, 200 DMA and daily returns feature column. In total, we have created 6 new features out of existing features from the historical stock dataset.

4.2.5 Data Partition

Data partitioning in data mining is the division of the whole data available into two or three non-overlapping sets: the training set, the validation set, and the test set. The basic idea of data partitioning is to keep a subset of available data out of analysis, and to use it later for verification of the model.

In here, we have split the whole dataset into following smaller datasets: train dataset which will be used for training the model which is set to 70%, validation dataset which will be used in the early stage of model building to validate and fit the model properly on unseen data which is set to 15% and lastly the test dataset which will be used for testing the model and evaluating the results which is set to 15%.

Figure 4.1 shows the graph with train, valid and test dataset of all five companies namely Adani Ports, Maruti Suzuki, Infosys, Indiabulls Housing Finance and Glenmark.

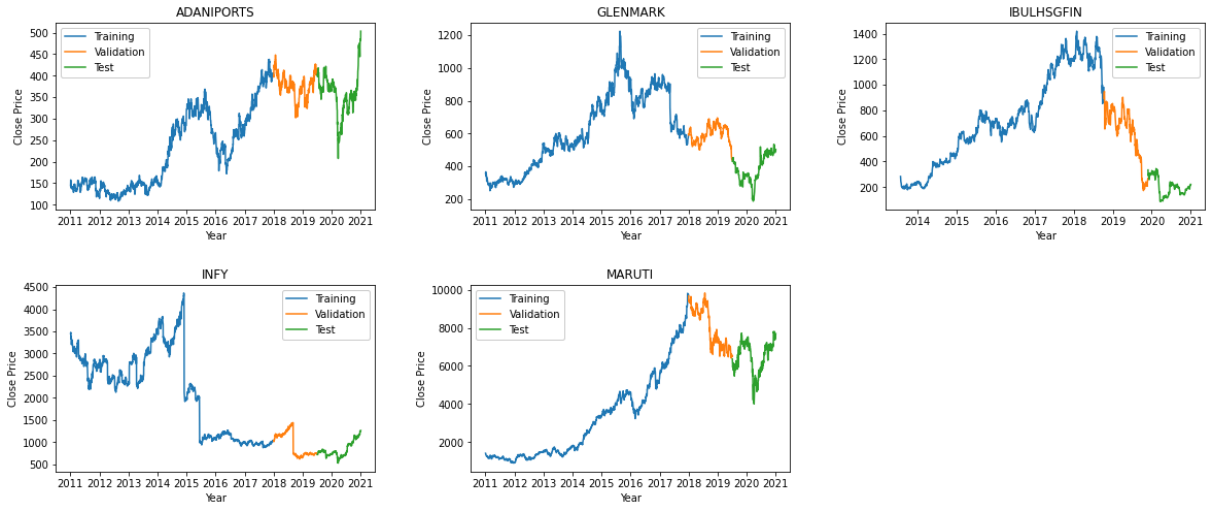


Figure 4.1 Dataset split into train-valid-test of stocks

4.3 Exploratory Data Analysis

Exploratory data analysis (EDA) is a term for certain kinds of initial analysis and findings done with data sets, usually early on in an analytical process. This section encloses the analysis of data from the dataset using exploratory data analysis. For analysis, here we used python libraries like "Matplotlib" and "Seaborn" for interactive visualizations. Examples of visualizations are scattered plot graph, line graph, histogram, etc. Those who are working with the data can expedite the process of figuring out what the data means, what it can be used for, and what conclusions can be drawn from it.

4.3.1 Boxplot

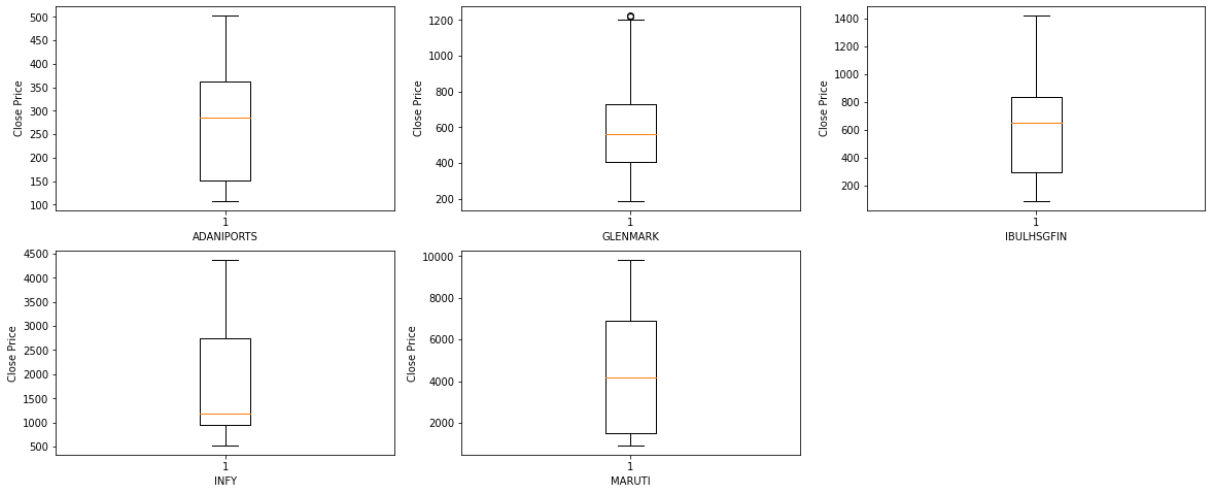


Figure 4.2 Boxplot graph showing close price feature

Figure 4.2, We can say that the Close Price of the selected five stocks has no outliers, and looking at the graph we can say that Adani Ports, Glenmark, Indiabulls Housing Finance have data negatively skewed and Infosys positively skewed. While Maruti shows the data as symmetric or normally distributed.

4.3.2 Scatter Plot (Close Price Vs Volume Weighted Average Price)

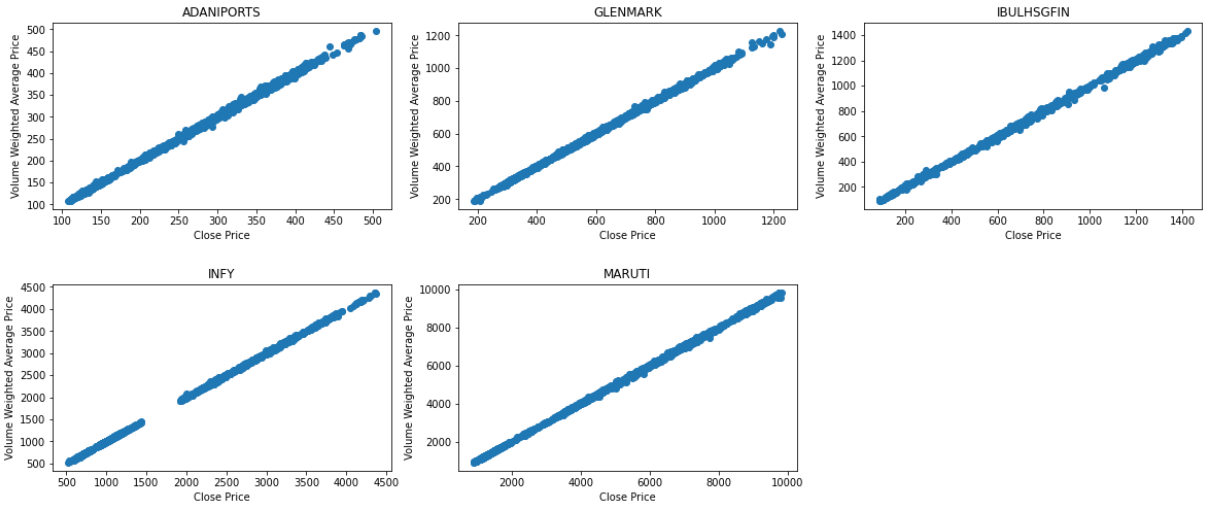


Figure 4.3 Close price vs volume weighted average price

Figure 4.3, Looking at the scatter plot, we can say that the closing price and the volume weighted average price are positively correlated as both the variables are increasing gradually with time. This means both the features have similar distribution and may present redundancy and even multicollinearity. We are going to use only one out of this for stock price prediction model.

4.3.3 Closing Price Trend Using Line Chart

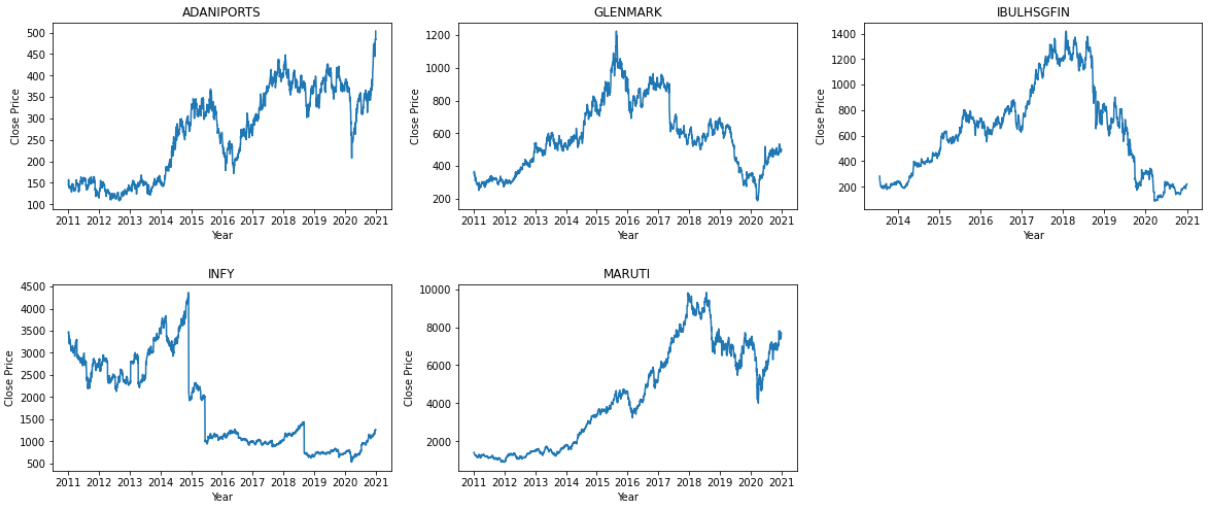


Figure 4.4 Closing price trend

Figure 4.4, In this figure we can see the trend of the closing price of all five stocks. Clearly, we can say that Adani Ports, Glenmark, Maruti stocks are on a positive trend and returning good returns to investors, but looking at the Indiabulls Housing Finance and Infosys stocks, it shows a negative trend and heavy losses for the investors in the past decade. Even through the graph we can depict that after 2017, Adani Ports, Glenmark, Maruti stocks skyrocketed.
From the time period 2017 to 2019, Indiabulls Housing Finance have grown drastically, just like Infosys which showed high growth in 2014-2015 but later both felled.

4.3.4 Volume Of Shares Traded Over Past Decade (2011-2021)

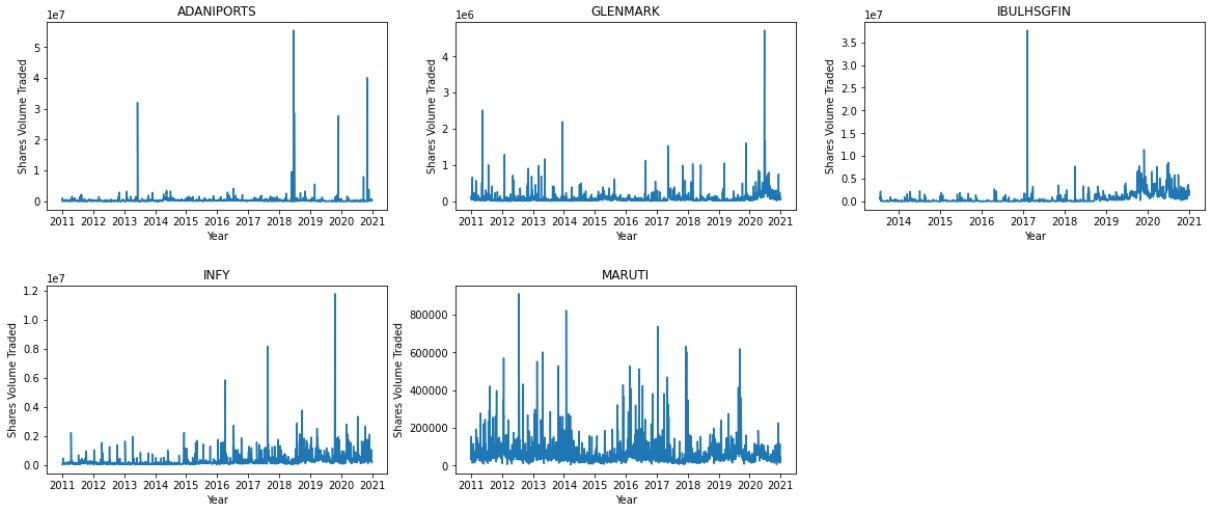


Figure 4.5 Volume of shares traded

Figure 4.5, We can say that shares traded for Maruti were always on peak over the past decade. This proves that investors show a lot of interests in Maruti shares. Apart from Maruti share, all others were traded at very standard volume in terms of the past decade. Some of these shares were highly volatile during some period of time like in years – 2017, 2019, 2020.

4.3.5 Moving Averages Over Period Of Time

Figure 4.6, We can say that moving averages were on a positive note for Adani Ports, Glenmark, Maruti over the last decade. But for the Indiabulls Housing Finance and Infosys, the moving averages were on a positive note at the start of the decade, but in the later half of the decade the moving averages started showing negative notes. This depicts that both Indiabulls Housing Finance and Infosys were bullish in the first half and then became very bearish in the later half of the decade.

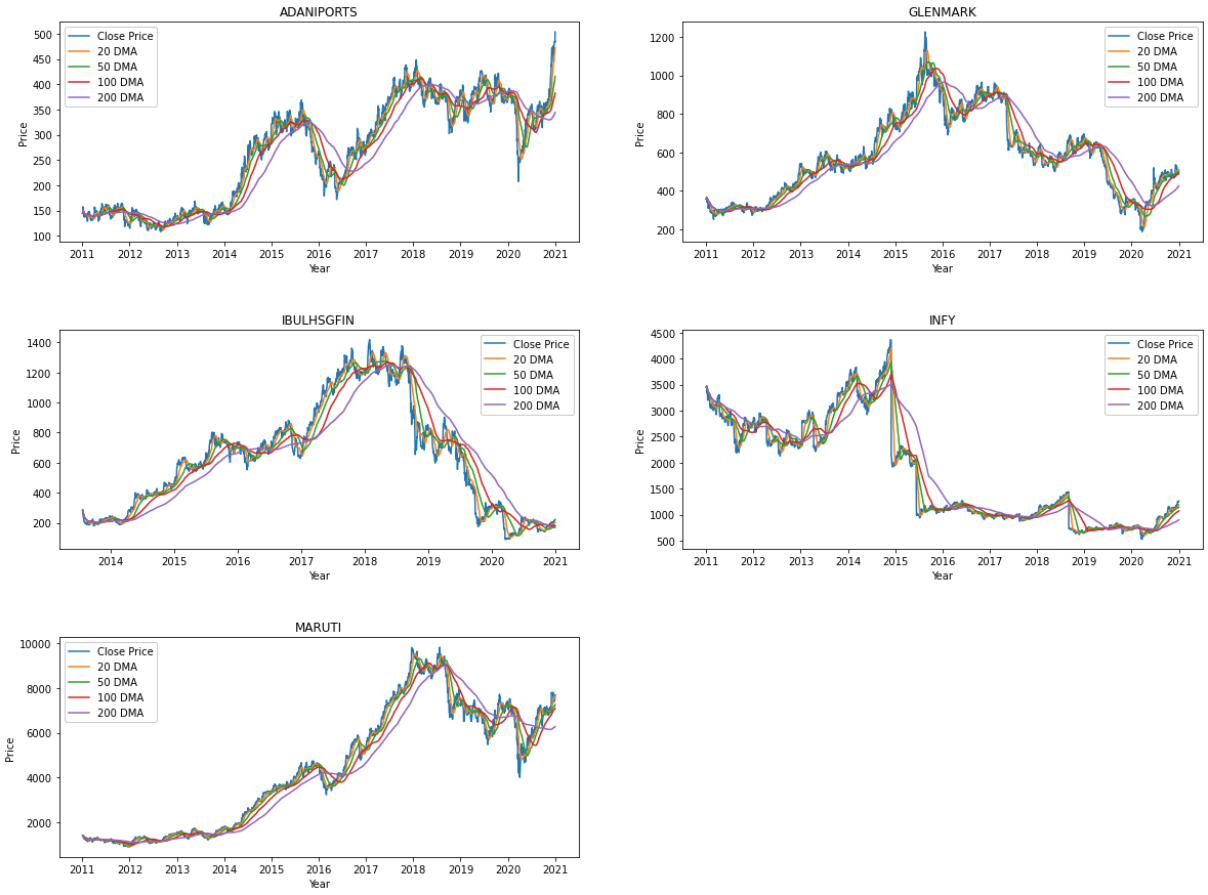


Figure 4.6 Moving averages over period of time

4.3.6 Daily Returns Graph

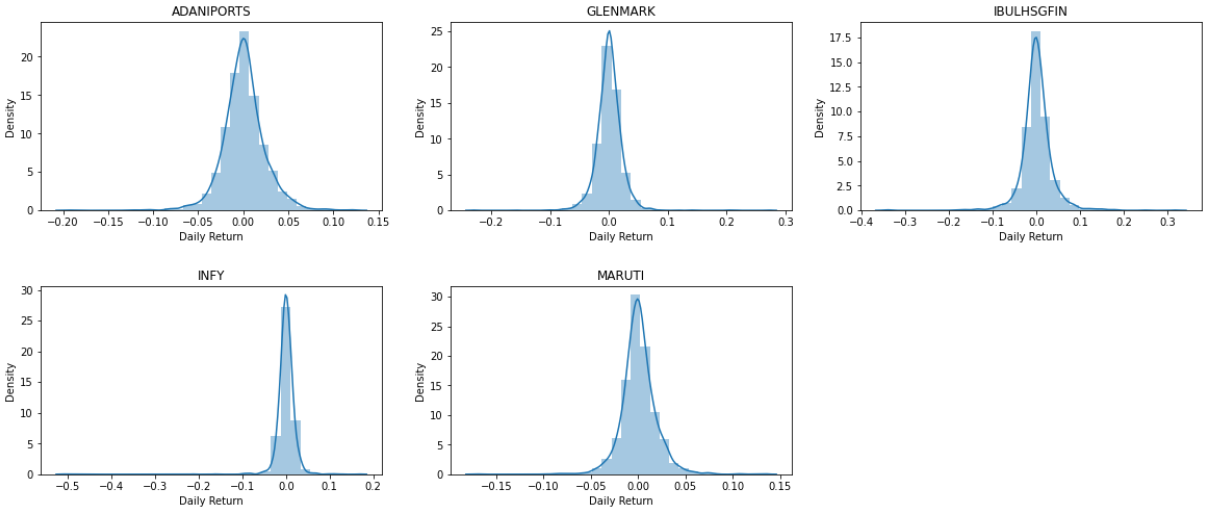


Figure 4.7 Daily returns of stocks

Figure 4.7, The above graph depicts that the daily returns of any of those five stocks were mostly negative. This means that there are fewer chances of getting profit on intraday trading but might have a possibility of getting high returns on a long-term investment.

4.3.7 Heatmap

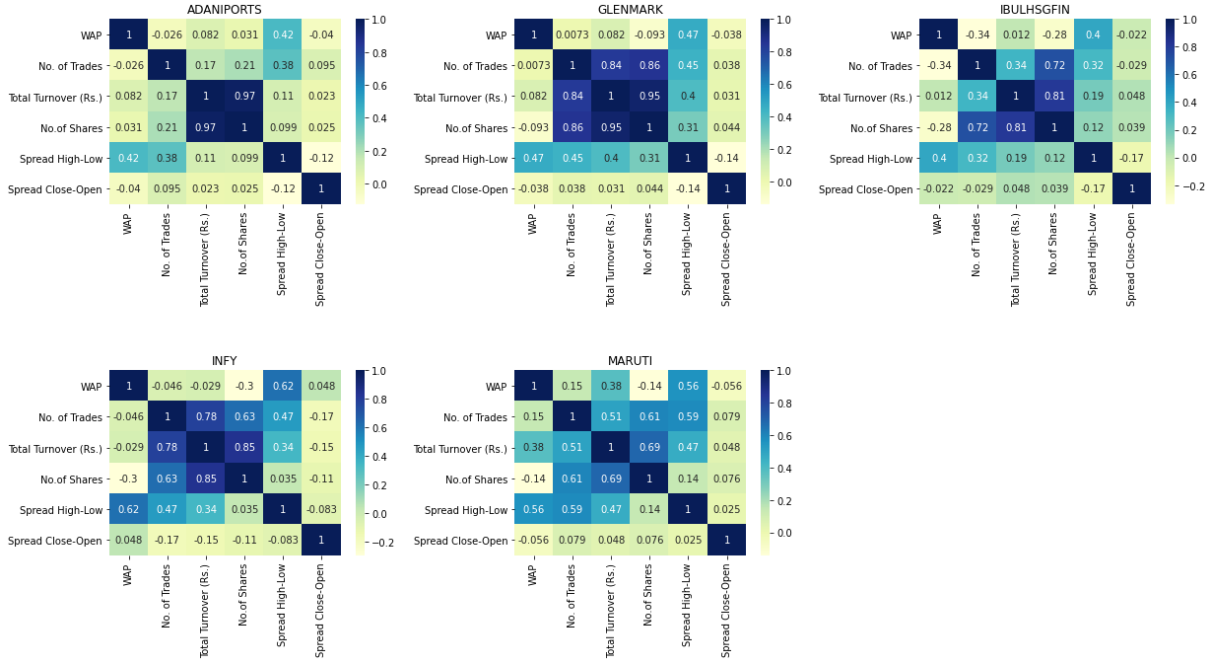


Figure 4.8 Correlation of data features

Figure 4.8, Looking at the heat-map, we can say that traded volume and Total Turnover, Traded volume and No.of Trades, VWAP and Spread (H-L), No.of Trades and Total Turnover are highly correlated. Volume is negatively correlated to VWAP and VWAP is also negatively correlated to Spread (C-O).

4.3.8 Relative Strength Index (RSI)

Figure 4.9, Looking at the plot, we can say that there was an uptrend or a bull market, the RSI tends to remain in the 40 to 80 ranges with the 40-50 zone acting as support. We can analyse that the stock historical strength and prominence was good, and there is no sign of overbought or oversold conditions for a longer period of time.

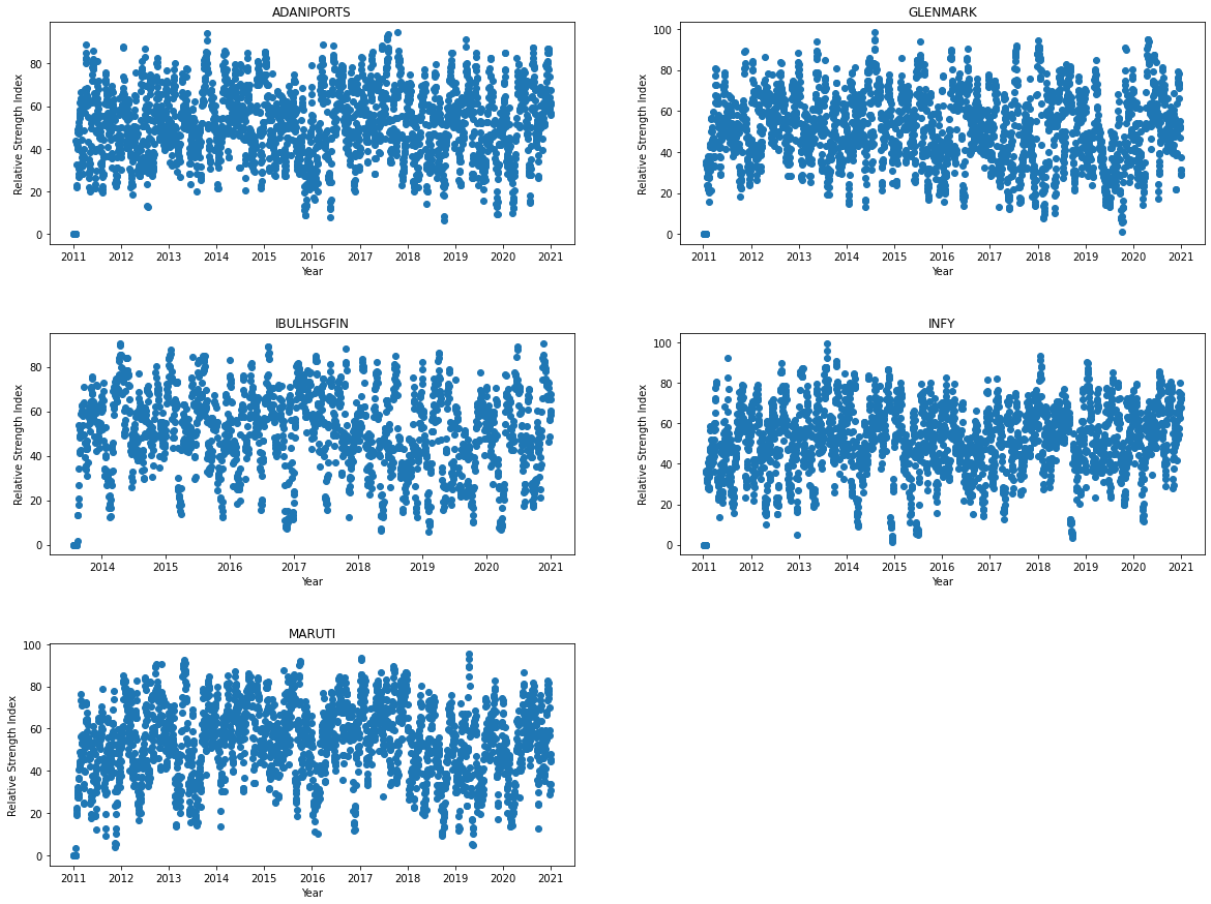


Figure 4.9 Relative strength index of stocks

The above analyses gives us a good idea about the dataset and is also useful while building the prediction model for stock price prediction. The above data analysis gives more clear picture of the data in terms of correlation between the different features of the dataset, outliers, multicollinearity, comparison, positive or negative trend over the past decade and returns in terms of daily trades. Apart from these, we also looked more into stocks as in which years the stocks were most actively traded and also more keen insights of the data.

4.4 Fine-Tuning the Architecture

Hyper-parameter tuning is choosing a set of optimal hyper-parameters for a learning algorithm. A hyper-parameter is a model argument whose value is set before the learning process begins. The key to machine learning algorithms is hyper-parameter tuning. The trade-off between variance-bias components is determined by the complexity of the model and the amount of training data. The optimal hyper-parameters help to avoid under-fitting (training and test error are both high) and over-fitting (Training error is low, but test error is high).

In the LSTM model we have used the following hyper-parameters for tuning the model. Starting with training epochs set to 50, batch size of 20 using batch_size attribute on the fit() function of multivariate LSTM sequential model, neurons are set to 400 with the help of n_neurons attribute on the fit() function which affects the learning capacity of the network. We have also added two layers for additional learning capacity of the network. We have even used the Adam optimization algorithm for an optimization technique for gradient descent and loss functions of mean_squared_error to lift performance. We have used 8 features (Close Price, VWAP, Spread Close-Open, Spread High-Low, 20 DMA, 50 DMA, 100 DMA, 200 DMA) and 50 as a sequence length time-steps to improve learning and predictive capability of the model. This together has made our final model give better and optimal results.

In the XGBoost Model too, we have used a number of hyper-parameters for tuning the model which are passed to param_grid attribute on the XGBRegressor() function. The following are the parameters, gamma of value (0.001) which specifies the minimum loss reduction required to make a split, max_depth of value (12) which is used to control over-fitting as higher depth will allow model to learn relations very specific to a particular sample, learning_rate of value (0.05) which is the step size shrinkage used in update to prevent over-fitting, n_estimators of value (300) which specifies the number of trees (or round) in a XGBRegressor model and objective hyper-parameter set to value 'reg: squarederror' which is regression with squared loss. The objective parameter specifies the function to be minimised and not to the model.

4.5 Model Implementation

This section contains and explains the various steps carried out towards model building and techniques or any specific libraries used for learning a model and even specifies the learning curve with the plot showing the performance of the model through the training loss values.

4.5.1 Sliding Window Technique

An essential step in time series prediction is to slice the data into multiple inputs data sequences with associated target values. For this process, we use a sliding windows algorithm. This algorithm moves a window step by step through the time series data, adding a sequence of multiple data points to the input data with each step. In addition, the algorithm stores the target value (e.g., Closing Price) following this sequence in a separate target data set. Then the algorithm pushes the window one step further and repeats these activities. In this way, the algorithm creates a data-set with many input sequences (mini-batches), each of which has a

corresponding target value in the target record. This process is applied both to the creation of the training and the test data.

We have applied the sliding window approach to our data. The result is a training set (x_train) that contains 2057 numbers for Adani Ports, Maruti Suzuki, Glenmark and Infosys or 1514 numbers for Indiabulls Housing Finance of input sequences, and each has 50 time-steps and 8 features. The corresponding target dataset (y_train) contains 371 (Adani Ports, Maruti Suzuki, Glenmark and Infosys) or 275 (Indiabulls Housing Finance) target values.

4.5.2 Model Building

In this section we have discussed in depth details of model building for both the models. It has enlisted the various libraries or methods used to train the model with a given training dataset. Below are given the details of models (LSTM and XGBoost) respectively in different sections.

4.5.2.1 LSTM Model

In the Multivariate LSTM model, we have passed the scaled training data of 70% along with the validation data of 15% for training the recurrent neural network for stock market prediction. This data is transformed using the sliding window technique for sequence time-steps and features creation. The architecture of our neural network consists of the following four layers.

1. LSTM layer, which takes our mini-batches as input and returns the whole sequence.
2. LSTM layer that takes the sequence from the previous layer, but only return 5 values.
3. Dense layer with 5 neurons.
4. Final dense layer that outputs the predicted value.

The number of neurons in the first layer must be equal to the size of a mini-batch of the input data. Each mini-batch in our dataset consists of a matrix with 50 steps and 8 features namely Close Price, VWAP, Spread Close-Open, Spread High-Low, 20 DMA, 50 DMA, 100 DMA and 200 DMA. Thus, the input layer of our recurrent neural network consists of 400 neurons.

In here, we have used sequential multivariate LSTM model and have used Adam optimizer and loss function as mean_squared_error. Later we start the training process by running the fit() function with x_train, y_train, batch_size (20), epochs (50), and validation_spilt (0.15) which is 15% of train dataset attributes. Lastly, we can see the output messages show the loss and val_loss over each epoch training.

4.5.2.2 XGBoost Model

In the XGBoost model, we have passed the same training data of 70% along with the validation data of 15% for training the gradient boosting model for stock market prediction. We have passed the following parameters for fine tuning the XGBRegressor method which are as follows: n_estimators [300], learning_rate [0.05], max_depth [12], gamma [0.001], and random_state [42]. For evaluating the model in the early stage itself, we have set the eval_set attribute passing train and validation data in it.

For building the model, we have used XGBRegressor with objective set to 'reg: squarederror' and the best parameters returned by the GridSearchCV method. Later we used the fit() function along with the following parameters or attributes: x_train, y_train, eval_set set to test and validation datasets and verbose set to False. The above configuration builds the optimal model for the stock prediction without any high computational time or cost.

4.5.3 Learning Curves

A learning curve is a plot of model learning performance over experience or time. Learning curves are a widely used diagnostic tool in machine learning for algorithms that learn from a training dataset incrementally. The model can be evaluated on the training dataset and on a hold out validation dataset after each update during training, and plots of the measured performance can be created to show learning curves. Reviewing learning curves of models during training can be used to diagnose problems with learning, such as an under-fit or over-fit model, as well as whether the training and validation datasets are suitably representative.

In our case, we have used Optimization Learning Curves, the learning curves calculated on the metric by which the parameters of the model are being optimized, e.g., loss.

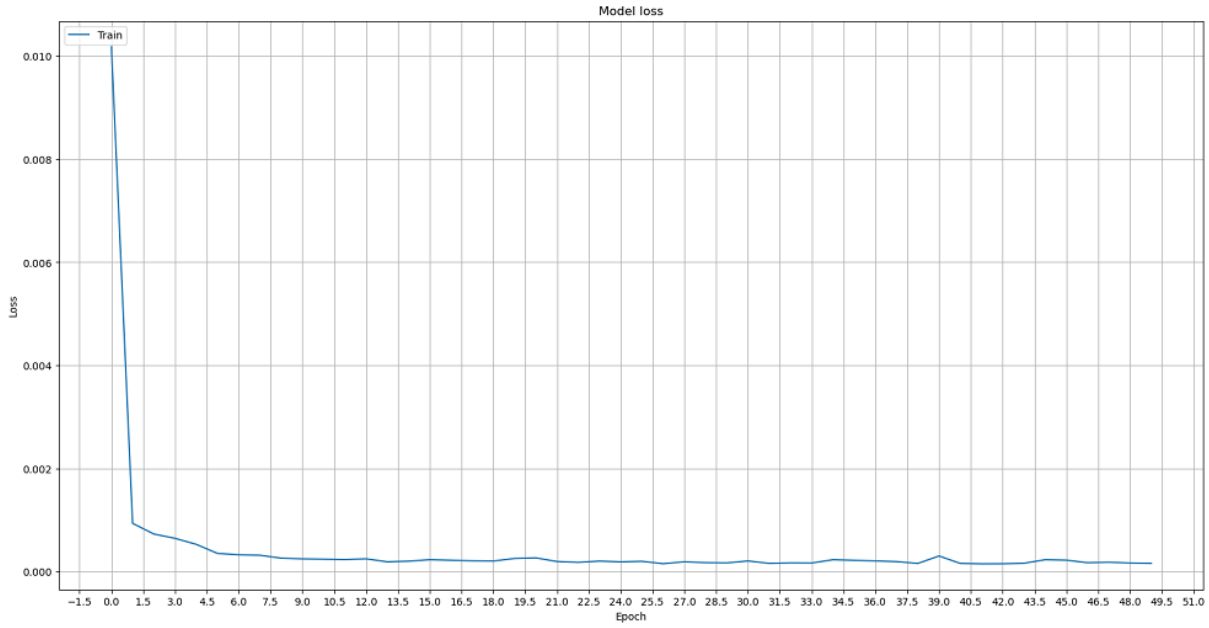


Figure 4.10 Learning curve chart of model loss during training

In the Figure 4.10, we can see clearly the plot showing the loss curve describing that the loss drops quickly to a lower plateau, which signals that the model has improved throughout the training process.

4.6 Summary

This chapter encompass the data mining steps that were carried out to extract the useful information and various insights to support the thesis aims and objectives and also to clean, dimensionality reduction and transformation of the data into the format acceptable by the model. Apart from these, it also discusses about the scaling of data and feature selection and about the sliding windows algorithm which helps in slicing the data into multiple inputs data sequences with associated target values for time series prediction.

This chapter further enlightened us on the data analysis carried out using interactive charts and graphs by exploratory data analysis. It shows how these charts and graphs helped us to come to a conclusion and make some helpful decisions out of the inferences drawn from it. It even shows the correlation drawn between the data features and list down various insights we have been able to make from the historical stock data.

Furthermore, it has shown the chosen hyper-parameters for availing the optimal loss and best results through the trained models (LSTM and XGBoost). It even summarizes how the change in parameters affect the model, and the results achieved from the prediction on the unseen data or test data given to the model.

Lastly, it discusses the model implementation steps that are carried out to build the model for stock prediction using the historical stock data and its features. Inside the model building it shows different methods and values passed to the model along with hyper-parameters. It also discussed the validation set, loss function, model algorithms, and the various libraries used for training and building models.

Overall, the chapter discussed the historical data along with its important features and the data pre-processing steps, splitting of data and, lastly, normalization of data and feature selection. Apart from these, it even shows useful insights drawn out of the past data which help in business decisions. It enlists the steps for model building, and the various techniques used for the training of the model, fine tuning architectures and along with that the plot of the learning curve to display the performance of the model.

CHAPTER 5

RESULTS AND DISCUSSIONS

This chapter includes the evaluation and results of the built final model. Also, here we discuss the interpretation of visualizations and the importance of features considered in a learned model. It even lists down the validation techniques used in model building and model outputs.

5.1 Introduction

This chapter includes the evaluation metrics along with the final obtained results and accuracy of the model. It also discusses and compares the results drawn out of both the models (LSTM and XGBoost) and even list down the table of derived evaluation metrics estimation of the five companies stock data based on the comparative study on predicted values and the actual values of the test data.

This chapter even discusses what new we have observed from the interactive visualizations and the interpretations from patterns or trends or what information the chart is meant to convey for achieving the goal to understand, interpret & reflect on the information represented & then infer new information based on the assessment.

This chapter further discusses the feature importance and which input features are highly useful at predicting a target variable while training the model and the role of feature importance in a predictive modeling problem. It even enlists the three most important features that are considered while predicting the closing price of stock using historical data in the learning model.

This chapter further discusses the validation methods used while training and learning the model for evaluating the performance of the models. It even lists down the different validation methods used on LSTM and XGBoost models like cross validation and automatic/manual validation sets.

This chapter begins with explaining the various inferences drawn out of the visual charts and graphs using historical stock data that supports the aims and objectives of a thesis or research work. Along with that, we can see new learnings out of these interactive charts which help in determining business decisions in the long run.

The results and evaluation section gives in depth details on the acquired accuracy and the different metrics used for error estimation evaluation out of those drawn predicted values. It

lists the comparison made between the four techniques (R2 Score, RMSE, MBE, MAPE) on five different sector companies.

At last, it discusses the important features used to predict values and even the validation sets, and it's types used here in this research work for fitting the model at its best on the stock time series data for better prediction.

5.2 Interpretation of Visualization

Data interpretation is the process of reviewing data through some predefined processes which will help assign some meaning to the data and arrive at a relevant conclusion. It involves taking the result of a data analysis. There are four steps to data interpretation: assemble the information you'll need, develop findings, develop conclusions, and develop recommendations. Here, data interpretation involves explaining patterns and trends shown through interactive charts and graphs as below.

5.2.1 Pre and Post of Demonetization and Covid-19

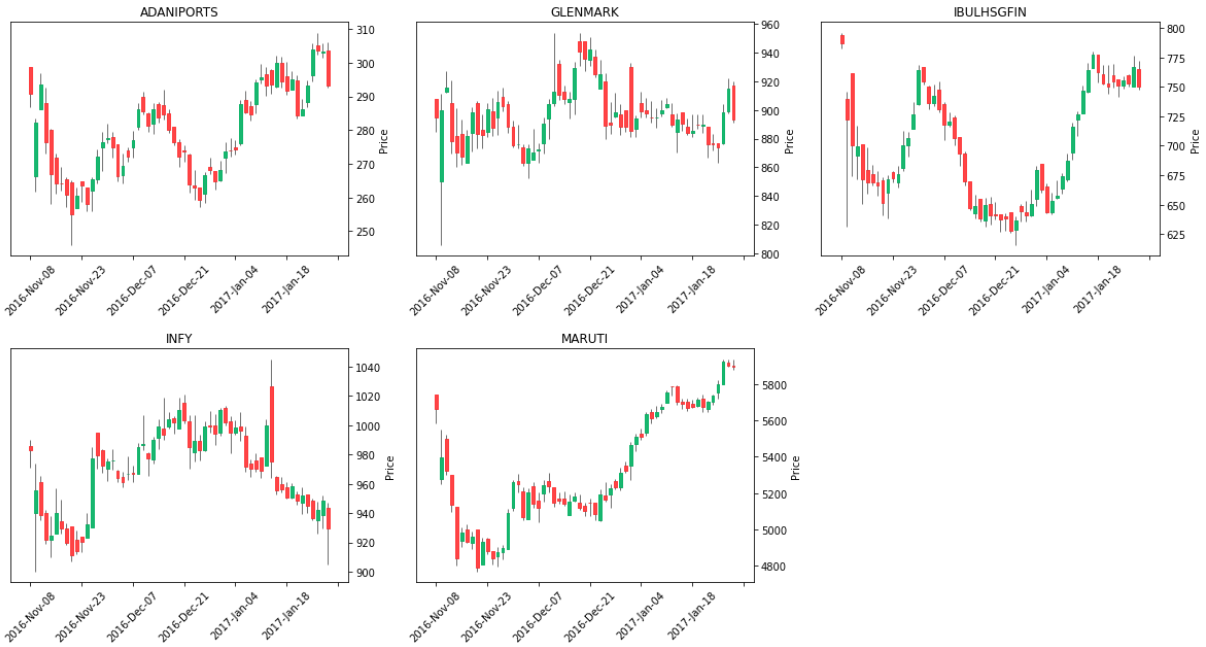


Figure 5.1 Candle chart for demonetization time period

Figure 5.1 shows that the economic policy of demonetization on 08/11/2016 has affected the stock market prices to a great extent, but the bear market lasted only for a month, and again the

prices started shooting up. Thus, we can say that the changes in economic policies do affect the share market.

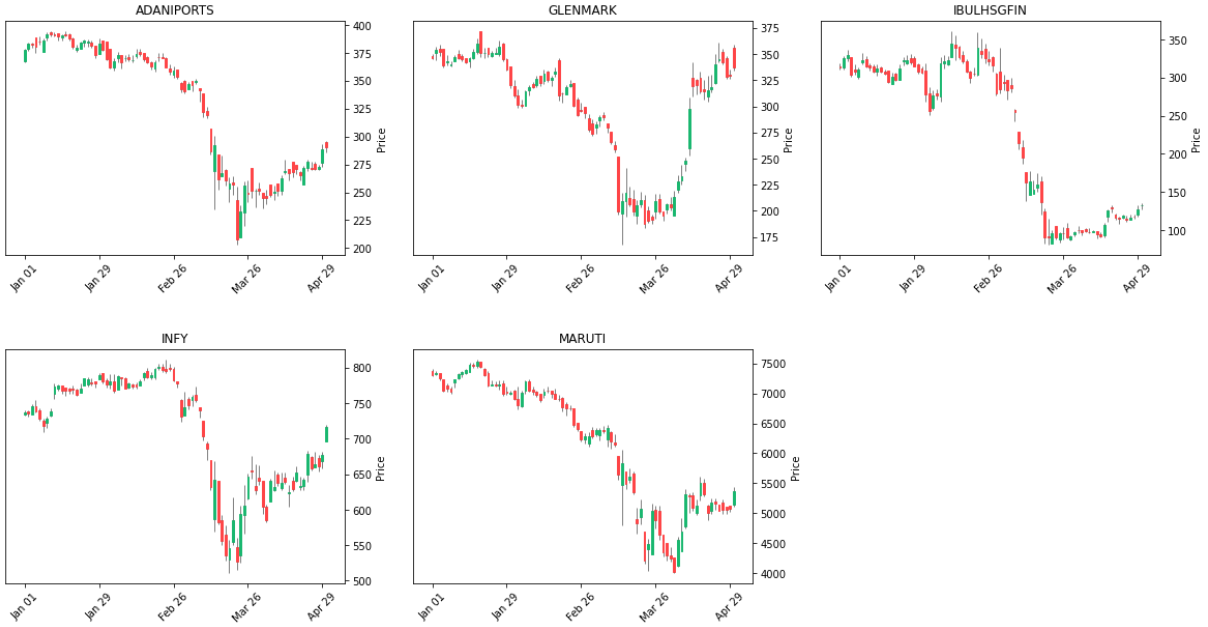


Figure 5.2 Candle chart for covid-19 time period

Figure 5.2 shows that during the pandemic start period of 01/02/2020 due to covid-19 has made the stock market fall very sharply. We can clearly see that prices have gone down upto 40% within a month. This clearly indicates that any unforeseen circumstances if occurred would highly influence the share market stock prices.

5.2.2 Price and Volume Correlation

Figure 5.3, From the above graph we can say that volume is positively correlated to the closing price of the stock. It might be the external factors that affect more the volume traded over a particular time period, but we can say that if the shares traded volume is increased, then there is a high chance of an increase in the stock price.

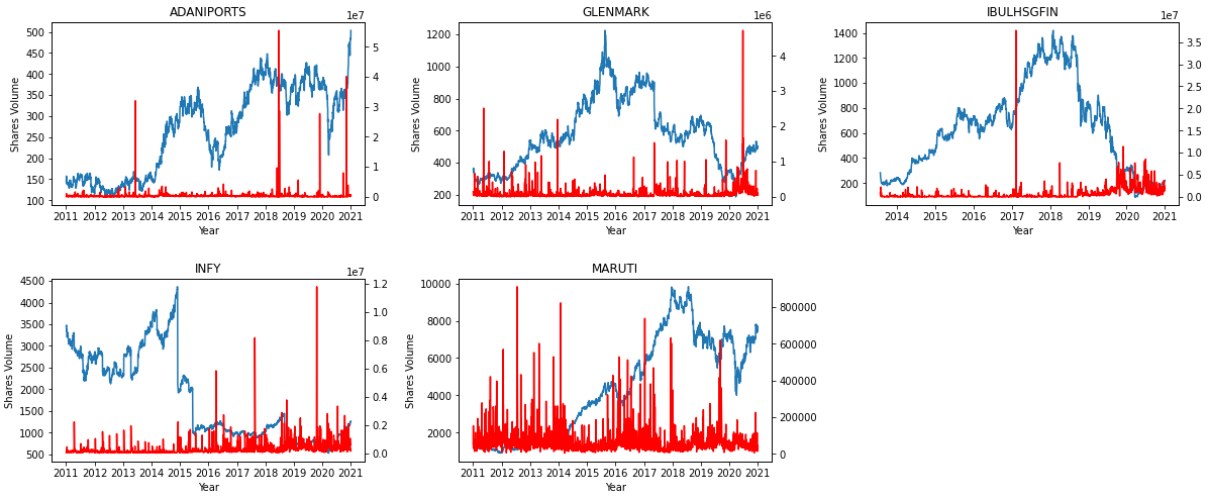


Figure 5.3 Price and volume correlation

5.3 Results and Evaluation

To evaluate the effectiveness of the models, a comparison is made between the four techniques on five different companies namely, Adani Ports, Glenmark, Indiabulls Housing Finance, Maruti Suzuki, and Infosys using both LSTM and XGBoost models. Predicted closing price are subjected to R-Squared (R2 Score), Mean Absolute Percentage Error (MAPE), Mean Bias Error (MBE) and Root Mean Square Error (RMSE) for finding the final minimalised errors in the predicted price.

Figure 5.4 represents the graphs showing the original closing price of the stock with respect to the predicted closing price of the stock of five different companies using LSTM. Figure 5.5 represents the graphs showing the original closing price of the stock with respect to predicted closing price across the test data along with train data using LSTM.

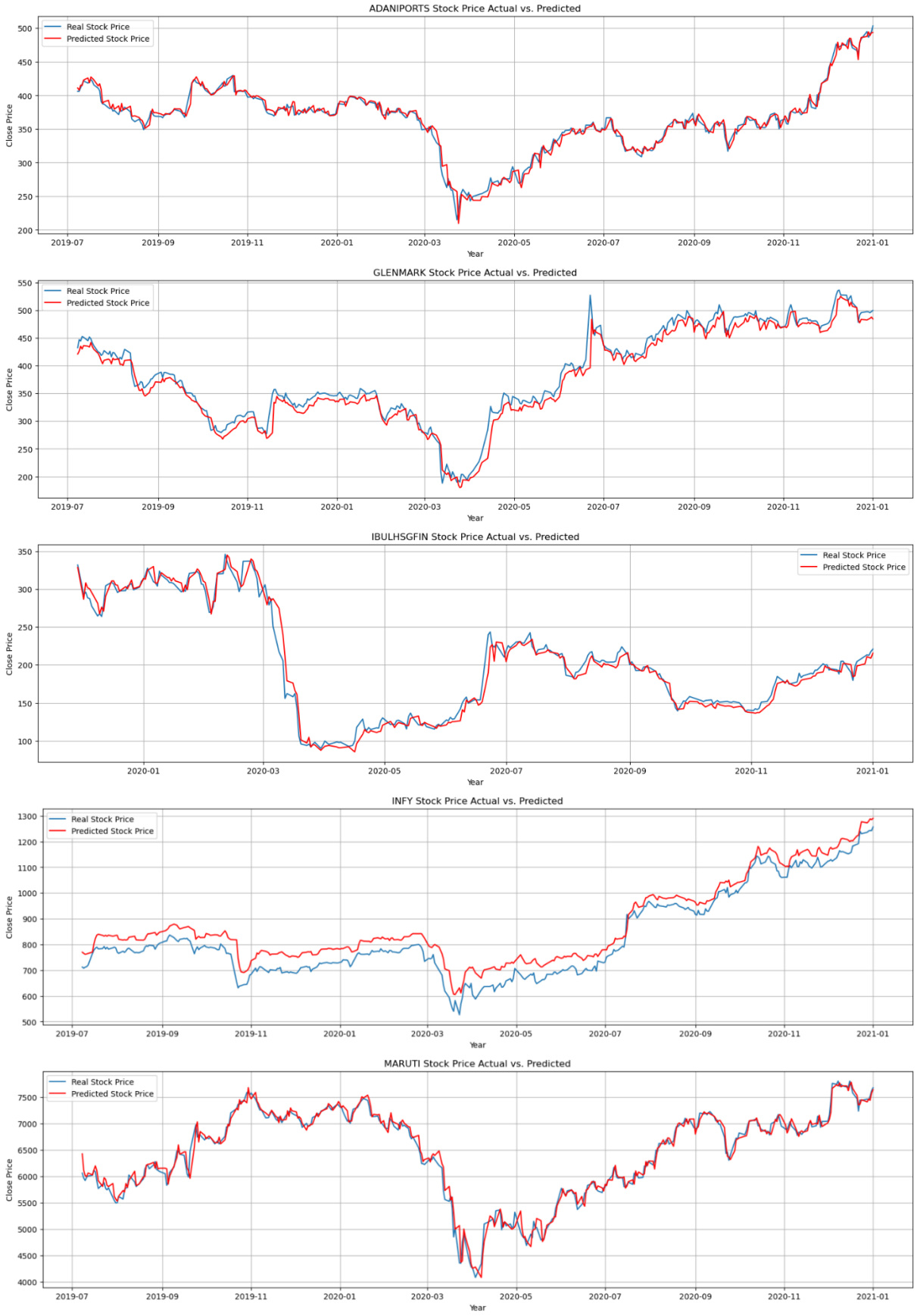


Figure 5.4 Predicted v/s actual close price using long short term memory

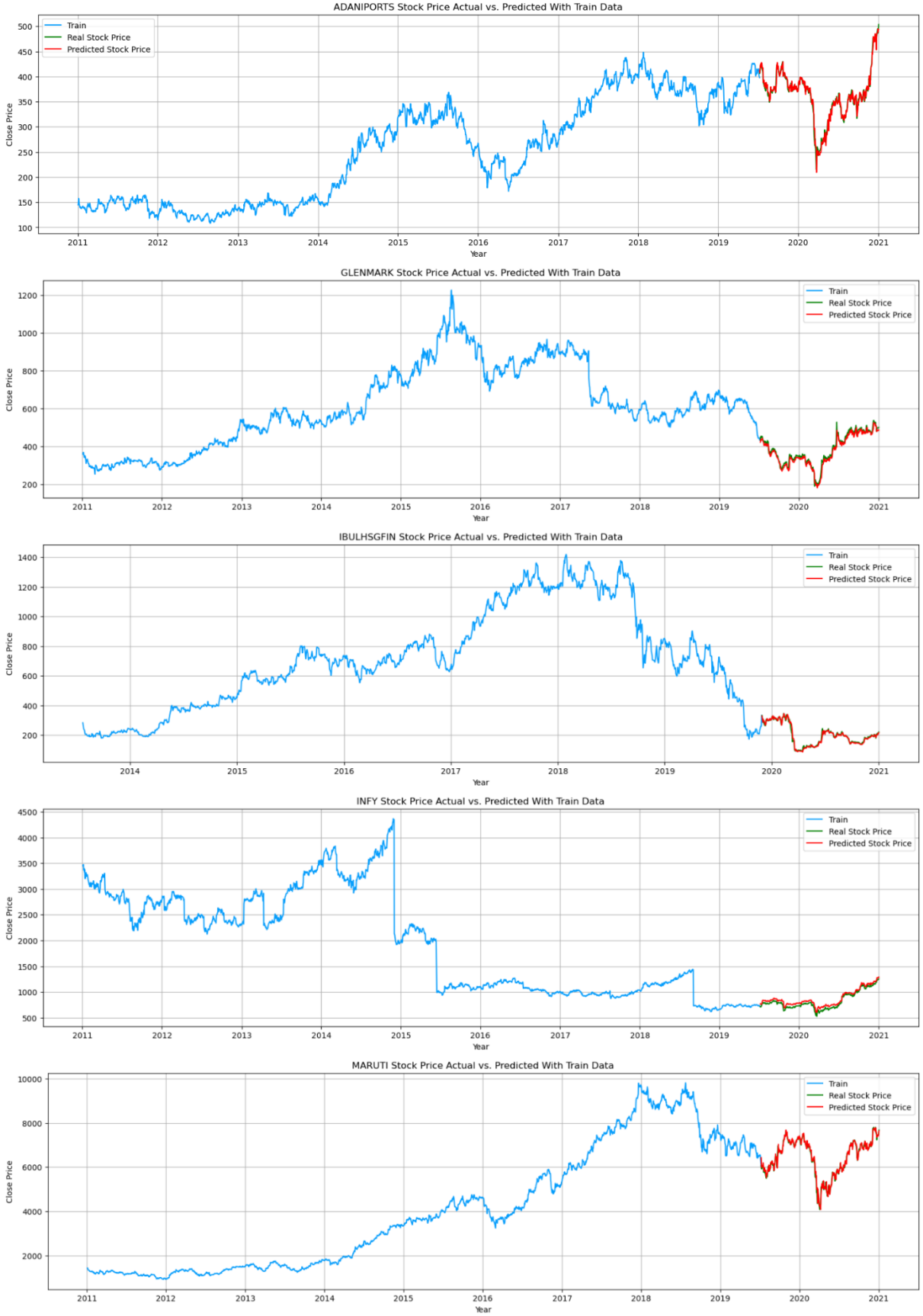


Figure 5.5 Long short term memory prediction chart with train data

Figure 5.6 represents the graphs showing the original closing price of the stock with respect to the predicted closing price of stock of five different companies using XGBoost. Figure 5.7 represents the graphs showing the original closing price of the stock with respect to predicted closing price across the test data along with train data using XGBoost.

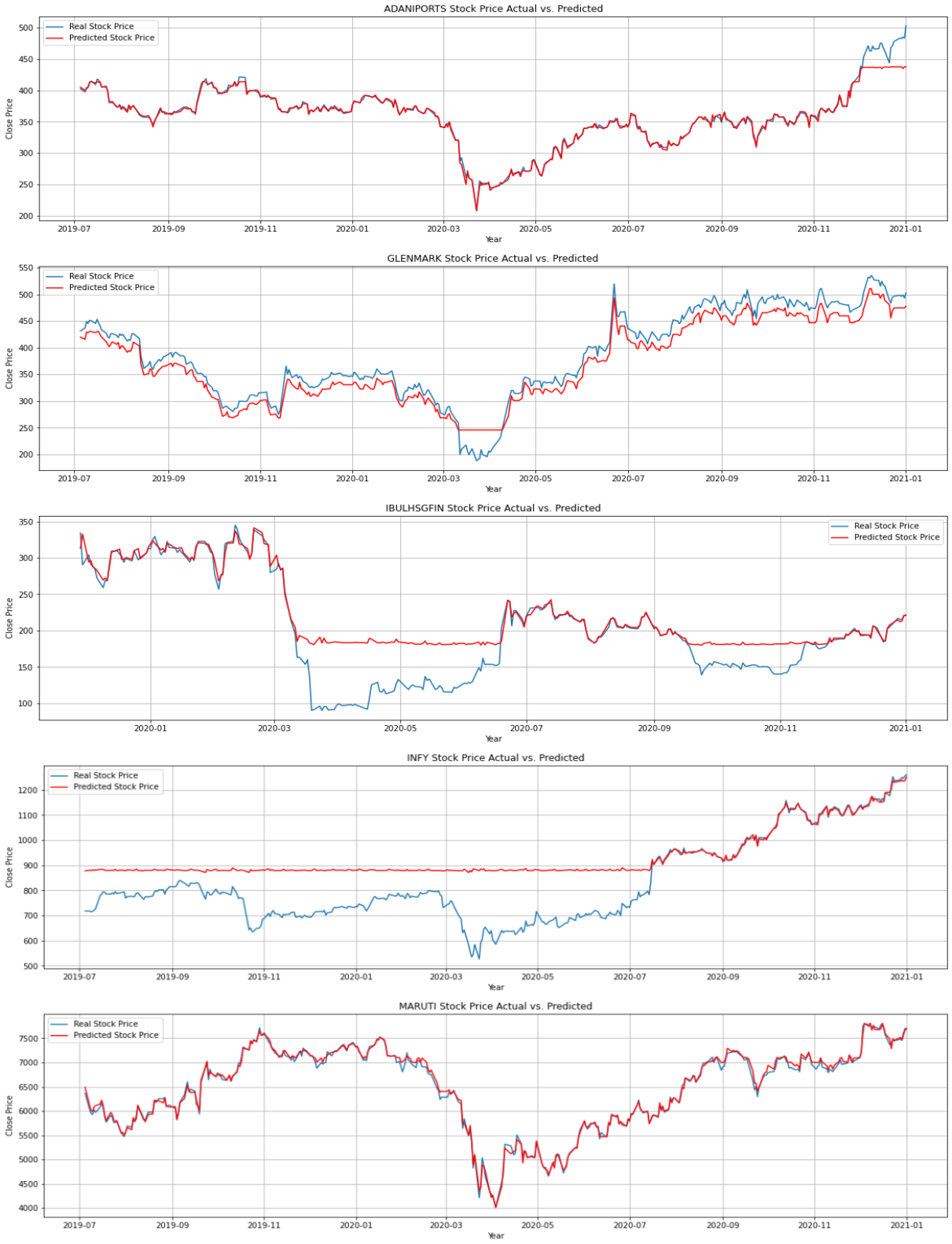


Figure 5.6 Predicted v/s actual close price using extreme gradient boosting

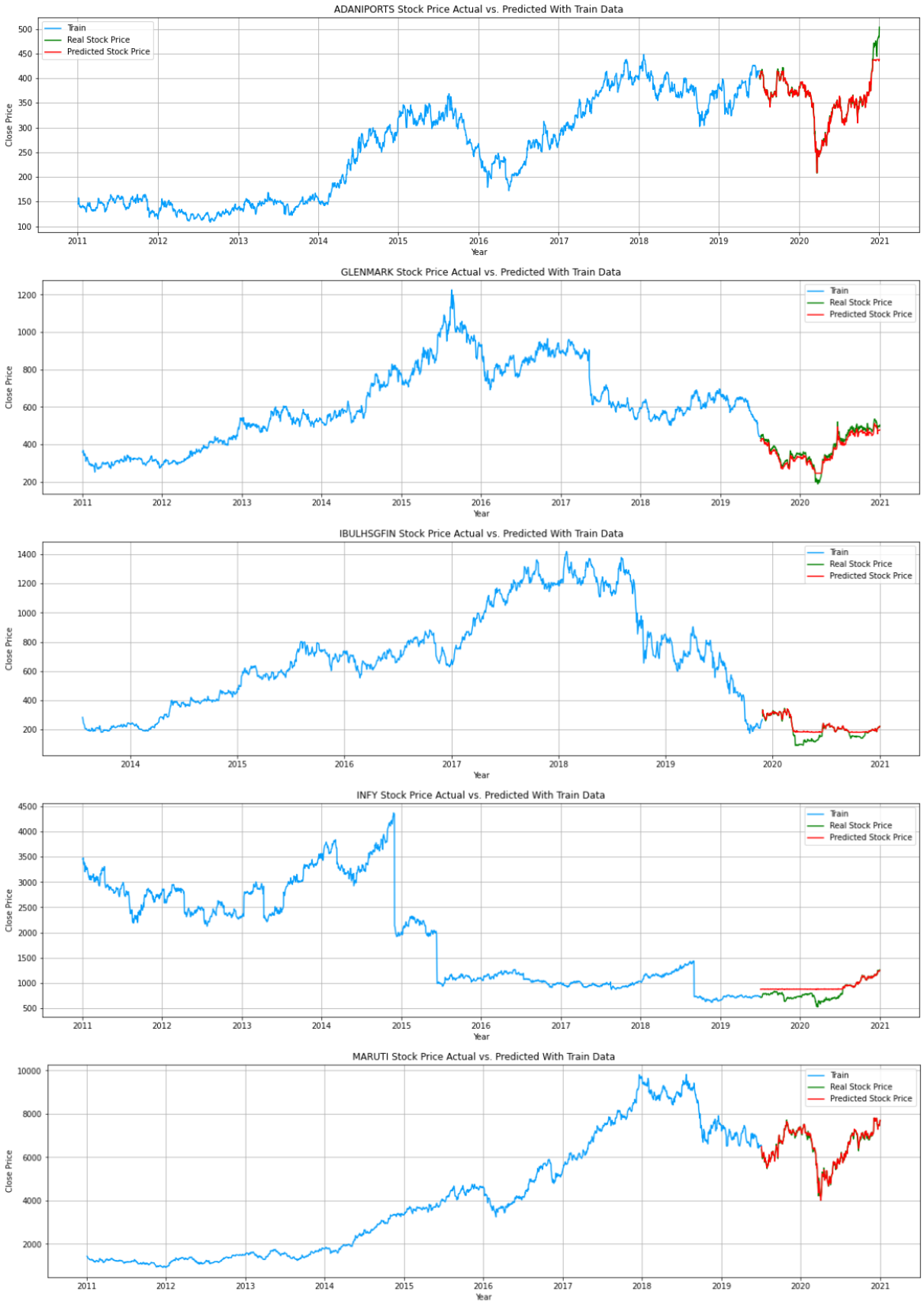


Figure 5.7 Extreme gradient boosting prediction chart with train data

Comparative analysis of the R-Squared (R2 Score), MAPE, RMSE, and MBE values obtained using LSTM and XGBoost model is shown in Table 5.1, it can be observed that LSTM shows better prediction results for stock prices.

Table 5.1 Comparative analysis of evaluation metrics

Company	**LSTM**				**XGBoost**			
	R2 Score	RMSE	MAPE	MBE	R2 Score	RMSE	MAPE	MBE
Adani Ports	0.98	7.31	0.02	-0.54	0.97	8.96	0.01	2.16
Glenmark	0.97	15.38	0.03	10.2	0.93	21.19	0.05	16
Indiabulls Housing Finance	0.98	10.54	0.04	1.26	0.76	33.72	0.16	-19.2
Infosys	0.91	51.7	0.06	-47.92	0.37	133.51	0.15	-100.87
Maruti	0.98	119.72	0.01	-18.57	0.99	60.06	0.01	-18.28

The comparative analysis indicates that for Adani Ports, Glenmark, Indiabulls Housing Finance, and Maruti Suzuki companies, LSTM proves to be a better technique, giving better metric values of R-Squared (R2 Score) and MAPE as shown in the Table 5.1.

5.4 Feature Importance

Feature importance refers to techniques that assign a score to input features based on how useful they are at predicting a target variable. The scores are useful and can be used in a range of situations in a predictive modeling problem, such as: Better understanding the data, Better understanding a model, Reducing the number of input features. In case of regression, it shows whether it has a negative or positive impact on the prediction, sorted by absolute impact descending.

Here we use XGBoost built-in function plot_importance() to plot features ordered by their importance. It takes the model as a parameter and outputs the plot of feature importance with the most important features from top to bottom in the horizontal bar chart. A benefit of using ensembles of decision tree methods like gradient boosting is that they can automatically provide estimates of feature importance from a trained predictive model.

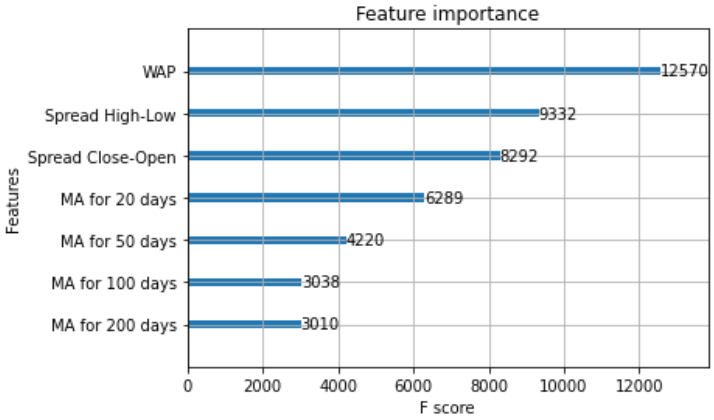


Figure 5.8 Feature importance chart

Looking at the plot above in Figure 5.8, we can clearly see that the most important features that were highly considered while predicting the target values are VWAP (volume weighted average price), Spread High-Low and lastly the Spread Open-Close. Thus, we can say that the previous day total traded value, and shares of that company contribute highly on the closing price value of the next day. The plot shows F-score that determines the importance of every feature for predicting target values.

5.5 Model Validations

Model validation is the process by which model outputs are (systematically) compared to independent real-world observations to judge the quantitative and qualitative correspondence with reality. Here we list down different types of validation sets used to validate the model while the model is at the learning stage. It helps the model to minimize the overfittting and to eliminate errors that can be caused for future predictions. It's a part of the training dataset which finds out the fine-tuned parameters for our algorithm. Below are the types of validation sets used in our models – Automatic/Manual Verification Dataset and Cross Validation.

5.5.1 Automatic/Manual Verification Dataset

In this type we either pass the separate validation dataset or mention the percentage of split to be done on training data while learning the model. This sets the separate validation dataset from training data to optimize the results.

In the LSTM model, we have used the automatic verification dataset which separates a portion of our training data into a validation dataset and evaluate the performance of our model on that validation dataset each epoch. We have done this by setting the validation_split argument on

the fit() function to a value of 0.15 (15%) of the size of our training dataset. We can see that the verbose output on each epoch shows the loss on both the training dataset and the validation dataset.

In the XGBoost model, we have passed the manual verification dataset which is a separate portion of data kept while splitting the datasets into train, valid and test datasets. This validation dataset is 15% of the total dataset and after scaling the data, it is passed by setting the eval_set parameter on the fit() function of the XGBoost's XGBRegressor function.

5.5.2 Cross Validation

The GridSearchCV class computes accuracy metrics for an algorithm on various combinations of parameters, over a cross-validation procedure. This is useful for finding the best set of parameters for a prediction algorithm. So based on all these possible combinations, we can get the best parameters by calling best_params_. This will become our final parameters to use in the final model. It also provides a mean cross-validated score of the estimator by calling best_score_.

The GridSearchCV will split the train dataset further into train and test to tune the hyper-parameters passed to it. And finally, fit the model on the whole train data with the best found parameters.

It takes the model as an estimator, parameters through param_grid attribute and verbose for printing info and debug messages. We used the 5-fold cross validation on the train dataset for fitting over each candidate and returning the best parameters and validation score. There were total 80 fits and 16 candidates for learning the model based on parameters passed to the model while training. It outputs following best parameters with values {'gamma': 0.001, 'learning_rate': 0.05, 'max_depth': 12, 'n_estimators': 300, 'random_state': 42} and best validation score of 0.96.

5.6 Summary

This chapter encompass the details of the results and evaluations and the discussions also made through the obtained results and the estimated error metrics. It initially discusses the decisions made through inferences drawn from the visual charts and graphs using past historical data that provides valuable insights from data features.

Here we can see how unseen circumstances affect the share market, or stock prices or even changes in economic policies or company policies that cause undesirable volatility in stock prices of the listed companies.

Later it displays the table of comparative results of the five top companies from different sectors which includes the evaluation metrics estimation values for both the models (LSTM and XGBoost). Even it concludes the best results obtained from the final models and also proves that deep learning models are best at time series stock data prediction by choosing LSTM model over XGBoost model from the comparative analysis done over the different evaluation metrics. It further showed the important features from the stock dataset that are very helpful in predicting the target variable and even lists down the three most important features which are Volume Weighted Average Price (VWAP), Spread Open-Close and Spread High-Low.

Last, it enlightened us on the validation sets used in the learning model and the types of validation sets – automatic/manual validation set and cross validation used in research work for both the models.

Overall, this chapter discussed the final outputs from the trained model and even the obtained accuracy and error estimations using regression evaluation metric techniques like R2 Score, MAPE, RMSE, MBE. It even discusses interpretations from visualizations of past historical data, the feature importance and the validation sets used in the learning model.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

This chapter includes the conclusion and recommendations for further work. It also discusses contributions to knowledge and various other discussions such as comparative study, achievements, areas where the solution could be improved etc. related to the thesis work conducted.

6.1 Introduction

This chapter outlines the conclusions drawn out of the thesis work and also recommends the ways in which this model can be further improved considering various other factors on which the stock price is depended upon. It outlines the good model from various model approaches and also justifies which algorithm is best for stock predictions. It even mentions how the limited set of historical data and features impediment the prediction results compared to actual values. This chapter also discusses the best results achieved and also how we can improve accuracy, regularize overfitting/underfitting of model via further tuning of models using more epochs/layers or optimal hyper-parameters or using other deep learning and machine learning algorithms or deep neural network. This chapter also discusses contribution to knowledge and lists the learnings from this research work along with what something new we have observed. Lastly, it also discussed the future work that can be done on the existing research work and also enlist recommendations for the further improvisation of conducted research work.

This chapter begins with discussions with extensive details on the limitation of historical data and the existing variables in the dataset, comparative analysis conducted and which technique is much preferable in the real world. It also mentions how we created new variables for obtaining better accuracy along with results that show best results in terms of accuracy achieved and evaluation metrics.

This chapter also mentions how viable the models are in the real world data and computation cost and infrastructure needed for those models to be implemented. They solve the problems of traders and investors in predicting the future stock prices, probably the next-day closing price of an individual company stock.

This chapter also brings up the knowledge gained by the researcher by conducting this research work and how it helped to achieve efficient results or comparative analysis by carrying out new

experimental work from other research work conducted previously. This helps in knowing how this research work would be helpful and informative for further research studies in the future.
This chapter ultimately discusses recommendations and future work that could be carried out on this paper by other researchers and also how we can improve the model by adding various extensive internal and external factors that very much influences the stock prices in the share market.

6.2 Discussion and Conclusion

Prediction of stock prices and returns are very difficult due to persistently changing company stock values because of dependable multiple factors which directly or indirectly influence the stock prices. Because of which one single model or limited data wouldn't be enough to accurately predict stock values in the real share market. The historical data at hand on BSE's website consists of only few features like open, high, close, low, volume-weighted average price, traded volume, etc. which are not adequate for better prediction. To achieve best results, we created new variables from existing ones which are frequently used by traders for recommending stocks to buy/sell and manually forecasting future stock prices.
The conducted study helps to summarize that we can use deep learning and machine learning models for stock prediction. It also states that the past data is highly effective in forecasting the future prices of stock. Using this model, we can even recommend stocks to investors for making their profits in the share market.
The features like VWAP, Spread Close-Open, and Spread High-Low are highly important features for predicting the prices of stock. If we compare actual stock prices to predicted stock values, we can clearly see that the values are almost similar. Thus, we can say that we can use prediction models for stock prices. Sometimes, predicted stock prices are not comparable to actual values for a time period due to the unforeseen circumstances like company policies, economic policies, demonetization, pandemic, etc.
Multivariate LSTM along with sliding window approach is used for predicting daily closing price of the stock and for a comparative examination, XGBoost model with cross-validation approach is also carried on historical dataset. In contrast to evaluation metrics based on R2 Score, RMSE, MAPE and MBE estimate evidently stipulate that LSTM provides better prediction of stock closing prices as compared to XGBoost. Results reveal that the finest values derived by LSTM model yield R2 score (0.98), MAPE (0.01), RMSE (0.42) and at last MBE (-0.54). This concludes that we can use LSTM model for predicting the stock values in the real world for better returns to investors.

6.3 Thesis Contribution

We have produced something worth knowing from our research, and it speaks with, and to, what is already known about our particular topic. The contribution is our offering to the scholarly conversation. It is the main idea of your paper and the main purpose of your research. This thesis makes various contributions which are discussed below:

1. Both theory and empirical findings contribute to our understanding of the stock prices. This study also contributes to our understanding of the question of how predictive models have swift and ease of stock prediction.
2. Prior research has shown that stock prediction using stock historical data is highly effective; however, we lack understanding of important features. I add to prior research showing the important features like technical indicators and volume-weighted average price.
3. We also found out that unforeseen circumstances do highly influence the individual listed company stock prices in the share market. This leads to the inability to predict the stock prices at that time period using past data.
4. We also found out that how suitable are the deep learning and machine learning models to forecast stock price using time series data. This study also depicts how efficient LSTM models are in time series prediction.
5. Conducting a comparative study, we also found that LSTM model is more efficient than XGBoost model. This concludes that deep learning models are better than machine learning models for time series prediction.
6. We also found out that the lack of internal and external factors affecting the stock prices gives poor prediction results.
7. We also found out that limited data and features suppress the prediction accuracy and better results.

6.4 Recommendations and Future Work

For future work, deep learning models could be developed which include financial news, articles along with financial parameters such as profit and loss statements, balance sheets, dividends, company announcements, stock-splits, bonus/rights issue, shares buyback, diverse types of information such as tweets, news, and other text-based data. Similarly, we can also pass technical indicators and fundamental details to achieve better results.

Even, the accuracy of the stock market prediction system can be further improved by utilizing a much bigger dataset to bring in seasonal and annual factors that affect the stock price

movement than the one being utilized currently. Furthermore, other emerging models of machine learning and deep learning could also be studied to check for the accuracy rate resulted by them.

We can further improve the model by deeper network (Deep Neural Networks), hyper-parameter tuning, adding more layers, regularizing overfitting/underfitting of model by Dropout, alternative loss functions, adaptive optimizers, features and time-steps, larger batch size and training for longer.

Since statements and opinions of renowned personalities are known to affect stock prices, a possible extension of this stock prediction system would be by combining latest sentiment analysis that can be linked with the LSTM to better train weights and further improve accuracy.

We can extend the stock market prediction system to be used in other stock exchanges like NYSE, LSE, SSE and even predict with newly provided features.

In the future, we can investigate different types of LSTM models, such as stacked LSTMs, encoder–decoder LSTMs, bidirectional LSTMs, CNN LSTMs, and generative LSTMs models in prediction of stock price.

REFERENCES

BSE Stocks Price (2021). Available at: https://www.bseindia.com/markets/equity/EQReports/StockPrcHistori.aspx (Accessed: 15 June 2021).

Chung, H. and Shin, K. S. (2018) 'Genetic algorithm-optimized long short-term memory network for stock market prediction', *Sustainability (Switzerland),* 10(10). doi: 10.3390/su10103765.

Das, S. *et al.* (2018) 'Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction', in *Procedia Computer Science*. doi: 10.1016/j.procs.2018.05.111.

Dhenuvakonda, P., Anandan, R. and Kumar, N. (2020) 'Stock price prediction using artificial neural networks', *Journal of Critical Reviews,* 7(11). doi: 10.31838/jcr.07.11.152.

Ding, G. and Qin, L. (2020) 'Study on the prediction of stock price based on the associated network model of LSTM', *International Journal of Machine Learning and Cybernetics,* 11(6). doi: 10.1007/s13042-019-01041-1.

Ghosh, P., Neufeld, A. and Sahoo, J. K. (2021) 'Forecasting directional movements of stock prices for intraday trading using LSTM and random forests', *Finance Research Letters*. doi: 10.1016/j.frl.2021.102280.

Henrique, B. M., Sobreiro, V. A. and Kimura, H. (2018) 'Stock price prediction using support vector regression on daily and up to the minute prices', *Journal of Finance and Data Science,* 4(3). doi: 10.1016/j.jfds.2018.04.003.

Hiransha, M. *et al.* (2018) 'NSE Stock Market Prediction Using Deep-Learning Models', in *Procedia Computer Science*. doi: 10.1016/j.procs.2018.05.050.

Jason Brownlee (2020) *Extreme Gradient Boosting (XGBoost) Ensemble in Python*. Available at: https://machinelearningmastery.com/extreme-gradient-boosting-ensemble-in-python/ (Accessed: 18 April 2021).

Kamalov, F., Smail, L. and Gurrib, I. (2020) 'Forecasting with Deep Learning: S&P 500 index', in *2020 13th International Symposium on Computational Intelligence and Design (ISCID)*. IEEE. doi: 10.1109/ISCID51228.2020.00102.

Kanade, P. A. (2020) 'Machine Learning Model for Stock Market Prediction', *International Journal for Research in Applied Science and Engineering Technology,* 8(6). doi: 10.22214/ijraset.2020.6030.

KOTAKSECURITIES.COM (2021) *Share Market Basics*. Available at: https://www.kotaksecurities.com/ksweb/share-market/share-market-basics (Accessed: 25 May

2021).
Madhusudan, D. M. (2020) ‘Stock Closing Price Prediction using Machine Learning SVM Model’, *International Journal for Research in Applied Science and Engineering Technology*, 8(11). doi: 10.22214/ijraset.2020.32154.
Moghar, A. and Hamiche, M. (2020) ‘Stock Market Prediction Using LSTM Recurrent Neural Network’, in *Procedia Computer Science*. doi: 10.1016/j.procs.2020.03.049.
N P Samarth, Gowtham V Bhat and Hema N (2019) ‘Stock Price Prediction’, *International Journal of Innovative Technology and Exploring Engineering*, 9(2S). doi: 10.35940/ijitee.B1042.1292S19.
Nandakumar, R. *et al.* (2018) ‘Stock Price Prediction Using Long Short Term Memory’, *International Research Journal of Engineering and Technology (IRJET)*, 05(03).
Nikou, M., Mansourfar, G. and Bagherzadeh, J. (2019) ‘Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms’, *Intelligent Systems in Accounting, Finance and Management*, 26(4). doi: 10.1002/isaf.1459.
Parmar, I. *et al.* (2018) ‘Stock Market Prediction Using Machine Learning’, in *ICSCCC 2018 - 1st International Conference on Secure Cyber Computing and Communications*. doi: 10.1109/ICSCCC.2018.8703332.
Selvamuthu, D., Kumar, V. and Mishra, A. (2019) ‘Indian stock market prediction using artificial neural networks on tick data’, *Financial Innovation*, 5(1). doi: 10.1186/s40854-019-0131-7.
Shen, J. and Shafiq, M. O. (2020) ‘Short-term stock market price trend prediction using a comprehensive deep learning system’, *Journal of Big Data*, 7(1). doi: 10.1186/s40537-020-00333-6.
Umer, M., Awais, M. and Muzammul, M. (2019) ‘Stock Market Prediction Using Machine Learning(ML)Algorithms’, *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, 8(4). doi: 10.14201/ADCAIJ20198497116.
Vijh, M. *et al.* (2020) ‘Stock Closing Price Prediction using Machine Learning Techniques’, in *Procedia Computer Science*, pp. 599–606. doi: 10.1016/j.procs.2020.03.326.
Xu, Y. and Cohen, S. B. (2018) ‘Stock movement prediction from tweets and historical prices’, in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*. doi: 10.18653/v1/p18-1183.
Yu, P. and Yan, X. (2020) ‘Stock price prediction based on deep neural networks’, *Neural Computing and Applications*, 32(6). doi: 10.1007/s00521-019-04212-x.

APPENDIX A: RESEARCH PROPOSAL

Abstract

In the 21st Century, the stock prediction has become one of the popular topics for many trading / investing stakeholders to gain better returns on investments. But the traditional methods aren't enough to predict the stocks more precisely and quick. Also, The haphazard and inconsistent historical series mark their prediction cumbersome. Availing the ML and DL approaches we can overcome the lack in prediction & can also achieve enhanced results. In this paper, we are using LSTM variant models and also XGBoost with hyperparameters tuned to predict the final price of a stock for a specific day. Model evaluation would be done using extensively used performance metrics: MAE, MAPE, RMSE. The financial dataset used here considers factors like open price, close price, high, low, and created new variables like volume-weighted average price (VWAP) and moving averages. The expected results from prediction should be more accurate with less error and much better compared to traditional methods output.

Table of Contents

1. Introduction

Stock is in today's world very notable among the stock investors and financing institutions. The stock market plays a major role in determining the economic strength of any country and also country's economy is more dependent on its stock exchanges indices. "The rise and fall of stock prices are influenced by many factors such as politics, economy, society and market. For stock investors, the trend forecast of the stock market is directly related to the acquisition of profits. The more accurate the forecast, the more effectively it can avoid risks"[1]. This desires investors to predict the stock price in advance by time series analysis of the stock's continuous data thus proving to occur variations in the value of a stock.

Stock forecasting is a nevertheless complicated job because of noise in continuous data and also non-linearity and non-stationarity characteristics. In the economic writings, stock value foretelling is compartmentalized as technical analysis, fundamental analysis, & ML and AI approach. In the research field too the stock forecasting plays a precise role as it decides a country's financial development. The stock predictions help the companies to take profit-making decisions and also the future growth aspect of their product and services.

In the proposed research, ML and DL models for stock prediction are created using XGBoost and LSTM variant models. It will take features namely close price, open, high, low, VWAP and moving averages as input and will provide the following day predicted close price of a stock of a certain listed firm. Moreover, this paper will also show the inferences of volume traded of that particular stock over timeframes and also will draw the stock price trend direction to be upward or downward. Based on the above, it will recommend the stock to buy/sell on the next day. This paper will also discuss the future scope of extending this research with more features and a large dataset.

2. Background and related research

In the past, lots of research papers have been submitted regarding the prediction of individual stock or indices. Some of them used machine learning techniques like Random Forest, LR (Linear Regression), SVM (Support Vector Machine), & Feature Extraction and even many implemented the time series models like Autoregressive (AR), ARIMA (Auto-Regressive Integrated Moving Average), & 3MMA (Three Months Moving Average). Some papers even introduced deep learning techniques like LSTM, ANN & DRNN (Deep Recurrent Neural

Network) or hybrid models. Given below is the list of similar papers that have been published in the past related to stock prediction.

Title & Author(s)	Dataset	Problems	Algorithm(s)	Result
"Stock Closing Price Prediction using Machine Learning Techniques (Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, Arun Kumar)"[2]	Yahoo Finance	Stock price predictor performance The limited set of data for prediction	ANN & Random Forest	ANN performs best with higher accuracy. Deep learning models are faster than ML models
"Stock Market Prediction Using Machine Learning (Ishita Parmar, Navanshu Agarwal, Sheirsh Saxena, Ridam Arora, Shikhin Gupta, Himanshu Dhiman, Lokesh Chouhan)"[3]	Yahoo Finance	Effective future value of the stock price. Better prediction results than the traditional way. Vanishing Gradient problem in the regression model	Linear Regression & LSTM	The LSTM provides better results. Machine learning has proven to be effective in stock prediction
"Short-term stock market price trend prediction using a comprehensive deep learning	This dataset consists of 3558 stocks from the Chinese stock market. Chosen 2	Important features to be selected for prediction	(FE + RFE + PCA) and LSTM	The hybrid model performed well and achieved better results.

system (Jingyi Shen, M. Omair Shafiq)"[4]	years as the period of the dataset.	A Hybrid tuned model of LSTM has been approached for a better outcome.		
"STOCK PRICE PREDICTION USING ARTIFICIAL NEURAL NETWORKS (Padmaja Dhenuvakonda, R. Anandan, N. Kumar)"[5]	Infratel firm dataset is used. It has close, open, low, high & volume as features.	Comparison of models better for stock prediction Limited dataset and fewer features	ANN, LSTM, AR, ARIMA	Comparing to various trained network models, ANN seems to be the best.
"Stock Market Prediction Using Machine Learning(ML) Algorithms (M Umer Ghani, M Awais, Muhammad Muzammul)"[6]	Diverse data is used such as research theories, data sets & resources related to financial presentation data.	To help investors invest with less risk Stock price prediction with better performance and accuracy.	Linear Regression; Exponential Smoothing; Time Series Forecasting	Exponential smoothing prediction turns out to be barely inaccurate. Subsequently, advised as elite stock predictor with general trend analysis

3. Research Questions

1. How effective are the LSTM (Deep Learning) and XGBoost (Machine Learning) models in predicting the stock closing price in terms of performance and less erroneous?
2. How are the DL models superior to the ML models for prediction?

3. What can we hypothesize from the trend of stock in a decade?
4. How better the prediction models can recommend traders to buy/sell stocks of a particular company?
5. What inferences can we set through volume traded over 3 years, 5 years and 10 years timeline and between VWAP and Closing Price?

4. Aim and Objectives

The main aim of this research is to predict the every day close price of a stock. Also to achieve a better outcome in terms of precision and less erroneous when put alongside other machine-learning models and deep-learning models.

The below are research objectives focused on the aim of this study which is as follows:

1. Effectiveness of LSTM & XGBoost models in predicting the stock closing price in terms of performance and less erroneous
2. Explanation on deep learning approach preferable instead of machine learning.
3. Hypothesis drew from the trend of stock in a decade.
4. Impact of prediction models on the recommendation for traders to buy/sell stock.
5. Inferences drawn out of volume traded over 3 years, 5 years and 10 years and between VWAP and Closing Price.

5. Significance of the study

This study is important to explain how deep & machine learning approach are nowadays much helpful in predicting the stock prices and how it impacts the traders especially intraday traders. It also shows the comparison between the ML & DL model related to performance. This study would help in verifying how much better results can be obtained via fewer features like close price, spread (High-Low, Open-Close), VWAP and Moving Averages. This study will reveal the inferences drawn from analysis on volume traded over 3 years, 5 years, 10 years timeframe. Also, the exploratory analysis on the dataset and T-test on the two continuous variables - VWAP and Close Price.

6. Scope of the study

“Predicting stock market returns is a challenging task due to consistently changing stock values which are dependent on multiple parameters which form complex patterns”[2]. Data obtained

consists of only a few features that aren't much adequate in the real world. We have created new variables to obtain better results and higher accuracy in the stock predicted closing price. Here we are using LSTM variant models for stock prediction and qualitative audit, we also implemented the XGBoost machine learning model. The audit is based on performance metrics such as RMSE, MAE, MAPE values. Also, we are determining the trend of the stock and the recommendation for traders' to buy/sell the stock of a particular company. In the possible future, we can extend this model by feeding the balance sheets, dividend payouts, and also twitter financial data, broadcast, bulletin & opinions of renowned personalities. We can even extend ML and DL models to provide a better precision rate.

7. Research Methodology

This section has details on the dataset and models used for forecasting. It also unveils the data pre-processing and transformation techniques.

Dataset

The historical dataset used in this research is collected from BSE Historical Stock Prices Data [7]. The time interval of the dataset is from 01/01/2011 to 01/03/2021. The dataset is collected for five companies from different sectors - Infosys Ltd, Glenmark Pharmaceuticals Ltd, Indiabulls Housing Finance Ltd, Maruti Suzuki India Ltd, Adani Ports and Special Economic Zone Ltd. A total of 12 attributes are present in the dataset obtained. Important columns/features that will be used for creating stock price predicting model – Close Price, WAP (Weighted Average Price), No.of Trades, Spread (Close-Open, High-Low).

Dataset Features	**Description**
Open Price	It is the price at which the security first traded at the open of the day's trading on its stock exchange
Close Price	Final price for a day the stock exchanged.
High Price	The highest closing value of a stock over the past 52 weeks
Low Price	The lowest price at which a security trades on a given trading day.

WAP	It measures the mean price of the stock exchanged for that day.
No.of Shares	It measures the number of shares traded in a stock
No.of Trades	It measures the number of trades throughout the day
Total Turnover (Rs.)	It measures the overall quantity of shares exchanged considering their value
Deliverable Quantity	It is the number of shares that move from one set of people to another set of people
Deliverable Qty to Traded Qty (%)	It measures the deliverable quantity w.r.t. traded quantity
Spread (H-L)	It is a difference between high-low of stock for that day
Spread (C-O)	It is a difference between close-open of stock for that day

Data Preprocessing

The dataset at hand is in a *CSV* structure that will be interpreted and transformed into a data frame using the python pandas tool. From this, the unwanted columns/features are dropped from the data frame such as No. of Trades, Total Turnover (Rs.), Deliverable Quantity, and % Deli. Qty to Traded Qty. After that, the dataset will be looked at to identify and correctly handle the missing values. We will use calculating the median to fill the missing values.

Once the dataset is cleaned, the data is normalized using MinMaxScaler from the 'sklearn' library in Python. It converts every available feature in a specific range such as [0, 1] or [-1, 1]. To preserve the zero's in a sparse dataset, MinMaxScaler is a good option. Finally, the dataset is split into three different datasets – training, validation, testing. The training dataset was kept at 60% of the available dataset and the testing and validation dataset was kept at 20% each from the available dataset.

Transformations

For training the model and prediction of stock closing price, we will create five more variables. Four of the new variables are moving averages which are used mostly by stock professionals for predicting stock price moving likely uptrend or downtrend. The 4 moving average variables are as below:

1. Stock price's 20 days' moving average (20 DMA)
2. Stock price's 50 days' moving average (50 DMA)
3. Stock price's 100 days' moving average (100 DMA)
4. Stock price's 200 days' moving average (200 DMA)

The last variable created would be Volume Weighted Average Price (VWAP). It is an average price of a stock weighted by volume.

Models

LSTM

Long Short Term Memory is competent enough to learn series dependence in forecasting problems. It's a special kind of RNN. This is accomplished on account of a recurring module of the model that has a blend of four layers interacting with everyone. The model learns, unlearn and preserve details from all units using the cell state & 3 gates. The cell state in LSTM helps the details to flow from all units without being revised. The forget gate can modify the cell state and the input gate can adjust information inside the cell state. Every unit has an input, output and a forget gate which can override the details in the cell state. Using the sigmoid method, the forget gate can decide which facts from the preceding cell state should be wiped out. The current cell state information is controlled from the input gate using a pointwise product of 'sigmoid' & 'tanh'. Finally, the information proceeded on to the next hidden state is handled by the output state.

XGBoost

"XGBoost is short for Extreme Gradient Boosting and is an efficient implementation of the stochastic gradient boosting machine learning algorithm"[8]. For regression and classification, XGBoost is preferably better for gradient boosting implementation. It is speedy and efficient when compared to other ML predictive models. XGBoost can be utilized for time series prediction by altering the data set into supervised learning. It uses a peculiar method for model evaluation named walk-forward validation. It's a decision trees ensemble where novel tree

repair faults already present in existed trees. Until we reach a point where additionally no enhancement can be done, we add up trees. XGBoost bestow hyperparameters to improvise the accuracy in model building. In terms of performance, XGBoost runs faster compare to other Decision Tree machine learning models. For predicting the stock closing price for a particular company, we would be tuning the hyperparameters for better results.

8. Expected Outcomes

- Comparison of evaluation outcome in-between machine learning and deep learning models w.r.t accuracy, MAE, MAPE & RMSE.
- Positive or Negative trend direction of the stock price of the last decade in stock price trend analysis.
- Recommendation to buy/sell the stock based on predicted price and trend.
- Analysis on stock returns and increase/decrease in volume traded over last decade.

9. Requirements / Resources

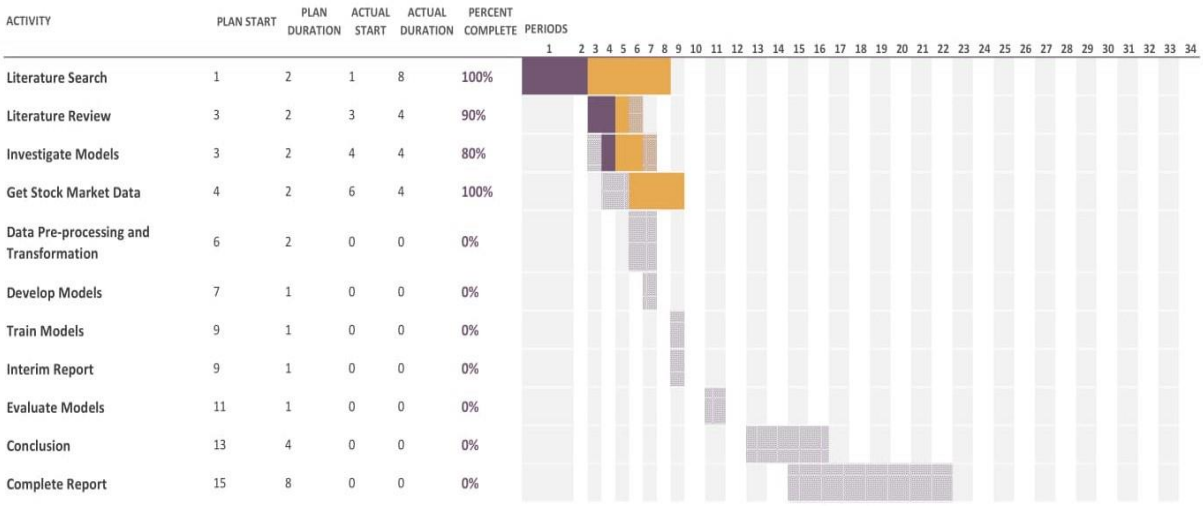
Hardware Requirements:

1. Laptop or Desktop
2. 6+ Cores CPU with base speed more than 2.5 GHz
3. 32 GB RAM
4. K80 GPU

Software Requirements:

1. Jupyter Notebook
2. Python 3.0
3. Pandas and Numpy
4. Plotly & Seaborn (Data Visualization Libraries)
5. Scikit-learn (Machine Learning Library)
6. Keras (Deep Learning Library)

10. Research Plan



ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Literature Search	1	2	1	8	100%
Literature Review	3	2	3	4	90%
Investigate Models	3	2	4	4	80%
Get Stock Market Data	4	2	6	4	100%
Data Pre-processing and Transformation	6	2	0	0	0%
Develop Models	7	1	0	0	0%
Train Models	9	1	0	0	0%
Interim Report	9	1	0	0	0%
Evaluate Models	11	1	0	0	0%
Conclusion	13	4	0	0	0%
Complete Report	15	8	0	0	0%

References

[1] G. Ding and L. Qin, “Study on the prediction of stock price based on the associated network model of LSTM,” *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 6, 2020, doi: 10.1007/s13042-019-01041-1.

[2] M. Vijh, D. Chandola, V. A. Tikkiwal, and A. Kumar, “Stock Closing Price Prediction using Machine Learning Techniques,” in *Procedia Computer Science*, 2020, vol. 167, pp. 599–606, doi: 10.1016/j.procs.2020.03.326.

[3] I. Parmar *et al.*, “Stock Market Prediction Using Machine Learning,” 2018, doi: 10.1109/ICSCCC.2018.8703332.

[4] J. Shen and M. O. Shafiq, “Short-term stock market price trend prediction using a comprehensive deep learning system,” *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00333-6.

[5] P. Dhenuvakonda, R. Anandan, and N. Kumar, “Stock price prediction using artificial neural networks,” *J. Crit. Rev.*, vol. 7, no. 11, 2020, doi: 10.31838/jcr.07.11.152.

[6] M. Umer, M. Awais, and M. Muzammul, “Stock Market Prediction Using Machine

Learning(ML)Algorithms," *ADCAIJ Adv. Distrib. Comput. Artif. Intell. J.*, vol. 8, no. 4, Sep. 2019, doi: 10.14201/ADCAIJ20198497116.

[7] "Stock Prices." https://www.bseindia.com/markets/equity/EQReports/StockPrcHistori.aspx (accessed Apr. 13, 2021).

[8] "Extreme Gradient Boosting (XGBoost) Ensemble in Python." https://machinelearningmastery.com/extreme-gradient-boosting-ensemble-in-python/ (accessed Apr. 18, 2021).