A Project Report on

Trading Analytics for Day Trading in Stock Market

Submitted in partial fulfilment for the award of the degree of

Master of Business Administration

In **Business Analytics**

Submitted by

Anand Mohan

R19MBA53

Under the Guidance of

Dr. JB Simha

Chief Mentor - RACE

REVA Academy for Corporate Excellence

REVA University

Rukmini Knowledge Park, Kattigenahalli,

Yelahanka, Bangalore – 560064

August 2022

REVA UNIVERSITY
Bengaluru, India

Candidate's Declaration

I, **Anand Mohan** hereby declare that I have completed the project work towards the first year of Master of Business Administration in Business Analytics at, REVA University on the topic entitled **Trading Analytics for Day Trading in Stock Market** under the supervision of Dr. **JB Simha, Chief Mentor-RACE**. This report embodies the original work done by me in partial fulfilment of the requirements for the award of the degree for the academic year **2022**.

Place: Bengaluru

Name of the Student: Anand Mohan

Date: 15 August. 22

Signature of Student

Certificate

This is to Certify that the Project work entitled **Trading Analytics for Day Trading in Stock Market** carried out by **Anand Mohan** with **SRN R19MBA53**, a bonafide student of REVA University, is submitting the first-year project report in fulfilment of the award of **Master of Business Administration in Business Analytics** during the academic year **2022**. The Project report has been tested for plagiarism and has passed the plagiarism test with a similarity score of less than 15%. The project report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the said Degree.

Signature of the Guide	Signature of the Director
Name of the Guide	Name of the Director
Guide	Director

External Viva

Names of the Examiners

1. <Name><Designation><Signature>
2. <Name><Designation><Signature>

Place: Bengaluru

Date:

Acknowledgment

I am highly indebted to Dr. Shinu Abhi, Director, Corporate Training for the guidance and support provided throughout the course and my project.

I would like to thank Chief Mentor, Dr. Jay Bharateesh Simha for the valuable guidance provided as my project guide to understand the concept and in executing this project. It is my gratitude towards Mithun Dolthody Jayaprakash and all other mentors including Ratnakar Pandey and Hrushikesha Shastry B S for the valuable guidance and suggestions in learning various data science aspects and for the support. I am grateful to them for their valuable guidance on several topics related to the project.

I am thankful to my classmates for their support, suggestions, and friendly advice during the project work. I would like to acknowledge the support provided by the founder and Hon'ble Chancellor, Dr. P Shayma Raju, Vice-Chancellor, Dr. M. Dhanamjaya, and Registrar, Dr. N Ramesh.

It is sincere thanks to all members of the program office of RACE who were always supportive of all requirements from the program office.

It is my sincere gratitude towards my parents and my family for their kind co-operation. Their encouragement also helped me in the completion of this project.

Place: Bengaluru
Date: 15 August. 22

Similarity Index Report

This is to certify that this project report titled **Trading Analytics for Day Trading in Stock Market** was scanned for similarity detection. Process and outcome are given below.

Software Used: **Turnitin**
Date of Report Generation:
Similarity Index in %:
Total word count:
Name of the Guide: Dr. Jay Bharateesh Simha

Place: Bengaluru
Date:

Name of the Student: **Anand Mohan**
Signature of Student

Verified by:

Signature
Dr. Shinu Abhi,
Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	OLS	Ordinary Least Squares
2	ARIMA	autoregressive integrated moving average
3	CV Lasso	cross-validation Least Absolute Shrinkage and Selection Operator
4	KNN	k-Nearest Neighbours
5	SMA	Simple Moving Averages
6	EMA	Exponential Moving Averages
7	CRISP-DM	Cross-Industry Standard Process for Data Mining
8	MAE	Mean Absolute error
9	MSE	Mean Square Error
10	R^2	R-squared (coefficient of determination)
11	RMSE	Root Mean square Error
12	MAPE	Mean Absolute Percentage Error
13	ADF	Augmented Dickey-Fuller
14	VWAP	volume-weighted average price
15	LSTM	Long Short-Term Memory

List of Figures

List of Tables

Abstract

The application of machine learning for stock prediction is attracting a great deal of attention in recent years. An enormous quantity of analysis has been conducted in this area and multiple existing results have shown that machine learning ways may well be with success used toward stock predicting using stocks' historical knowledge. Most of those existing approaches have targeted short-term prediction of stocks' historical value and technical indicators. during this thesis, twenty-one years' price of stock daily Returns is being utilized and investigated for accuracy of the predictions.

The objective of the project is to predict changes in stock price with higher accuracy while minimizing the degree of risks involved. The objective of the study is that all sorts of investors should still be able to utilize the paper's findings to assist them to guide their quality allocation and create buy-sell selections that best meet their needed returns expectations and the stock market predictions should work well in live testing environment as well.

A rule-based model is being developed to try and do hypothesis testing to see whether or not the chosen stock's value is crossing any of the subsequent moving averages: the 7-day, 13-day, 20-day, 100-day, and 200-day moving averages. It will be a purchase decision if the projection indicates that the value will be higher than various Moving Averages. Exponential statistic Models are then utilized to produce identical 5 hypothesis testing models. After that, any five ARIMA-based statistic models are created to support the buy or sell recommendation for the underlying stock.

Then various numerous Classification Models have been applied particularly K neighbors Classifier, Logistic Regression Modelling, and Auto Keras Classification Model using Structured knowledge classifier. The results show that AutoKeras Classification Model achieves the most effective prediction Accuracy followed by the Logistic Regression Classification Model and Then KNN Classification Model. SMA-7 samples and EMA-7 samples using T-test applied mathematics Hypothesis testing Models conjointly provided fairly smart accuracy.

Then various used regression Modelling Algorithms are used for predicting the close value and compared the Metrics, particularly MAE and MAPE.

The OLS-Linear Regression Model, Lasso Regression Model, Lasso regression Model using Cross Validation, The KNN rule, Decision Tree rule, GridSearchCV rule with Hyper-parameter standardization, Random Forest Regression Model, XGBoost Model, Using PCA with LSTM, Using PCA with LSTM with Moving Average variables (Feature Engineering), LSTM Neural Network Model, Regression Model using AutoKeras are the Regression Models used for predicting the close value.

The OLS-Linear Regression Model and Regression Model using AutoKeras offer the most effective results. Random Forest Regression Model and using PCA with LSTM conjointly provided smart results.

The project findings demonstrate that machine learning models may well be utilized to aid basic analysts with decisions relating to stock investment.

Keywords: Stock prediction, Hypothesis testing, ARIMA, Classification Models, Regression Model, LSTM, PCA, AutoKeras

Contents

Chapter 1: Introduction

The Stock market, as a result of its high volatility, maybe a new field for researchers, scholars, traders, investors, and companies. The number of Machine-Learning associated techniques that are developed have created the potential to predict the market to an extent (Sonkiya et al., 2021).

An oversized inventory of stock prediction techniques has been developed over the years, though the consistency of the particular prediction performance of most of those techniques remains debatable. In recent years, the recognition of applying numerous machine learning and data processing techniques to stock prediction has been growing. Results from several of the studies have shown that prediction models trained with historical worth and volume information may be with success used in predicting. For trading stocks through a broker, there is usually a commission paid to the broker for every purchase and sale. The rate of commission varies from broker to broker; however, it will nearly eat up the potential profit because the Trading frequency will increase, even with discount brokers (Huang et al., 2021).

The requirement is to beat the deficiencies of Fundamental and technical analysis, and also the evident advancement within the modeling techniques has driven numerous researchers to review new strategies for stock value prediction. A replacement type of collective intelligence has emerged, and new innovative strategies square measure being used for stock price predictions. The methodologies incorporate the work of machine learning algorithms for exchange shares analysis and prediction (Rouf et al., 2021).

Chapter 1 discusses the importance of Machine-Learning associated techniques that are developed for investments in the stock market. The chapter discusses that an oversized inventory of stock prediction techniques has been developed over the years and also informs that the evident advancement within the modeling techniques has driven numerous researchers to review new strategies for stock value prediction. In chapter 2, some of the available literature will be scanned which would throw light on various related aspects of Machine-Learning methods and other methodologies, and also study and research other related issues which would help assist better in Day trading in Stock Market.

Chapter 2: Literature Review

Financial markets are going through eventual transformations via the foremost fascinating inventions of the present time. They will have a significant impact on several areas like business, education, jobs, technology, and therefore on the economy. Analysing exchange movements and worth behaviours is extraordinarily difficult as a result of the market's dynamic, nonlinear, nonstationary, statistic, noisy, and chaotic nature and also because stock markets are being influenced by several extremely interrelated factors that embrace economic, political, psychological, and company-specific variables (Shah et al., 2019).

Some literature has used both supervised and unsupervised machine learning techniques for securities market predictive modelling and located that both kinds of models will create predictions with some accuracy. The assumption is being shared that even machine learning techniques haven't been ready to predict monthly securities market returns with high accuracy and this belief is being reiterated in this paper (Alhomadi, 2021).

Hypothesis testing could be a technique that helps to see whether or not a particular treatment has an impression on the people in a population. it's a proper procedure employed by statisticians to just accept or reject applied math hypotheses. the most effective process to verify whether or not an applied math hypothesis is true would be to look at the whole population. Since that's typically impractical, researchers generally examine a random sample from the population. If sample information doesn’t seem to be according to the applied math hypothesis, the hypothesis is rejected (Сороко, 2017).

ARIMA models have proven their economical capability to provide a short forecast and have unendingly outperformed refined structural models within the short prediction. This model in monetary time-series statistics is particularly economical and solid as the commonest Artificial Neural Network techniques. ARIMA model building phases involve model identification, diagnostic management, and also parameter analysis. One of the variants of the RNN flavour is the LSTM model. The self-loop style is employed as a vital input to construct a steep path that may be freely followed for a protracted time. A method exploring nonlinear parameters is employed to model a time series statistic (Biswas et al., 2021).

The central plan of PCA is to scale back the spatiality of a data set consisting of an outsized variety of interrelated variables, whereas holding the maximum amount as attainable of the variation within the data set. this is often achieved by remodelling a brand-new set of variables so that the first few derived variables explain most of the existing variations of that of the actual variables. The goals of PCA are to extract the foremost necessary data from the info table, compress the dimensions of the info set by keeping solely the necessary information, modify the outline of the data, reanalyse the structure of the observations and therefore the variables, and compress the info, by reducing the number of dimensions, while preventing abundant loss of information. eigenvectors and eigenvalues are the basic foundational principles used to implement PCA (López del Val & Alonso Pérez de Agreda, 1993).

Baek and Kim propose a framework referred to as ModAugNet, that is constructed on an associate LSTM deep learning model. Among the 10 models, four of them are designed on variants of convolutional neural network architectures, whereas the remaining six are made applying different LSTM architectures. The models are trained by applying the records of the first year, and they're tested on the remaining records. The cumulative RMSE and the RMSE for every day in a very week are computed to judge the prediction accuracy of the models. The results disclosed some fascinating observations. First, it's found that whereas the convolutional neural network models are quicker, in general, the accuracies of each convolutional neural network and LSTM model are comparable. Second, the univariate models are quicker and more correct than their multivariate counterparts (Series, 2021).

Based on the projected neural design search technique, an open-source AutoML system, particularly Auto-Keras was conceived. The goal is to help domain consultants WHO aren't aware of machine learning technologies to use machine learning techniques with ease. However, Auto-Keras is specializing in deep learning tasks, which is completely different from the systems specializing in shallow models. Although there are many AutoML services out there on giant cloud computing platforms, cloud services aren't cheaper. Also, the cloud-based AutoML sometimes needs difficult configurations of Docker containers and Kubernetes, which isn't straightforward Also, the AutoML service suppliers on cloud platforms cannot guarantee the safety and privacy of the information provided To bridge the gap, Auto-Keras was developed (Vreeken & Yamanishi, 2019).

The R-square is the proportion of the expected variable that's explained by a regression model. MSE measures the mean square error between the expected and actual variables. The addition of all the square values is calculated and divided by the no. of points. because of the squaring of errors, the negative values, and positive values don't diminish one another. RMSE measures the average magnitude of absolute error between the expected and actual variables. The MAE is commonly referred to as the mean absolute deviation. As compared with MAE, the RMSE includes a comparatively high weight for big errors, as a result of the errors being squared before averaging. The MAPE calculates the average percentage error. The MAPE is employed as the loss measurement for regression models in machine learning since it's more intuitive to elucidate the relative error. MAPE ought to be avoided for data existing at a low scale (Jierula et al., 2021).

Chapter 3: Problem Statement

Stock market analysis and prediction is still interesting and tough prospect. Today, big components of the population are excluded from the prospect of exploring monetary investments. Financial analysts investing in Stock markets generally do not appear to be tuned in to the exchange behaviors. They are facing issues in stock Trading as they are not able to understand which stock to buy and which to sell to achieve a lot of profits. The need for participation in these markets is also a high level of capital, so these markets are dominated by big investors, perpetuating the wealth divide. The ability to understand Exchange for making profits is required but balancing risks moderately at the same time is also very important.

Algorithmic Trading systems have changed the approach by which stock markets perform. Most of the commerce volumes in equity futures are generated by rule-based Methodology and not by humans. Whereas algorithmic Trading gives benefits like reduced expenses, reduced latency, and no dependence on sentiments, it brings up challenges for retail investors as they do not have the desired technology to create such systems. Today, it is common to look at events where panic selling is triggered due to these systems and thence the markets overreact. As a result, it becomes harder to gauge market behaviors. With new algorithms continuing to flood the markets every day, comparison of the effectiveness and accuracy of these algorithms pose nonetheless an added challenge.

Any one or two associated formulas or techniques may go fine on back testing in controlled environments, but the main challenge is live testing, as a result of many things like price variations, quiet news, and existing noise. Hence, a viable analysis direction would be to grasp a variety of the favored stock analysis techniques and implement those best practices in live or simulated environments.

Chapter3 mentions that stock market predictions with high accuracy while minimizing risks are required. As discussed, algorithmic Trading gives many benefits but retail investors do not have the desired technology to create such systems. As explained earlier, any derived and associated formula may go fine on back testing in controlled environments, but the main challenge is live testing. Chapter 4 will introspect more on objectives that would be more probably achieved through the effort put into the present project.

Chapter 4: Objectives of the Study

- firstly, the objective would be to use the knowledge given by the user and predict the results with the maximum amount of accuracy as they'll be. The project can use Hypothesis testing and numerous Classification Modelling Techniques to evaluate the accuracy of stock exchange predictions.

- Secondly, the project will use an ensemble of various Machine Learning and Deep Learning Algorithms and ARIMA modelling and will suggest to the user on investments for not simply solely the most profit but conjointly for minimizing the chance of loss. It is not solely necessary to know the quantum of profits created through trading in the Market however conjointly also evaluate the degree of risks concerned. The project can confirm what proportion errors in predictions are attainable by utilizing numerous estimation metrics specifically MAE, MSE, RMSE, and MAPE.

- Thirdly, the objective of the project is that the machine learning techniques explained here ought to be accessible to all sorts of small and big retail investors in contrast to Algorithmic Trading systems which can need advanced technology to form such systems. Even if it is believed that predicting stock exchange returns with high accuracy using daily or monthly returns is hard, investors can still use the paper’s findings to help them to guide their quality allocation, make buy-sell picks, and formulate optimum portfolios that best meet the clients’ required returns.

- fourthly, Machine learning algorithms can add real-time and manipulate the info in real-time, providing the most effective resolution. With the assistance of machine learning, the system can acknowledge the previous pattern and check out to suggest the output of what can be the long-run value of the stock. The objective of the project is to form a Leader board system using varied Statistical, ML, and Deep Learning Algorithms and supported by back testing on the dataset so that the most optimum and acceptable algorithm from the Leader board with the least MAE and MAPE ought to be chosen in the live testing setting.

Chapter 5: Project Methodology

Chapter 4 mentions that the main motivation for predicting returns in stock price is higher accuracy while minimizing the degree of risks involved. Chapter 4 discusses that even if it is believed that predicting securities market returns with high accuracy using daily or monthly returns is troublesome, the objective of the study is that all sorts of investors should still be able to utilize the paper's findings to assist them to guide their quality allocation and create buy-sell selections that best meet their needed returns expectations and the stock market predictions should work well in live testing environment as well. Chapter 5 will introspect more on the project Methodology that would be implemented and endeavours for continuous improvement that will be taken up while working on the project.

The CRISP-DM framework has been used for the project. The process of CRISP-DM is split into Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment

The CRISP-DM may execute in a very not-strict manner (could travel and forth between completely different phases). The arrows indicating the requirement between phases also are vital to one another phase; the outer circle represents the cyclic properties of the framework. CRISP-DM itself is not a one-time method, even as the outer circle diagram shows. Each method may be a new learning expertise, that new things are being learnt throughout the method, and it may trigger alternative business queries.

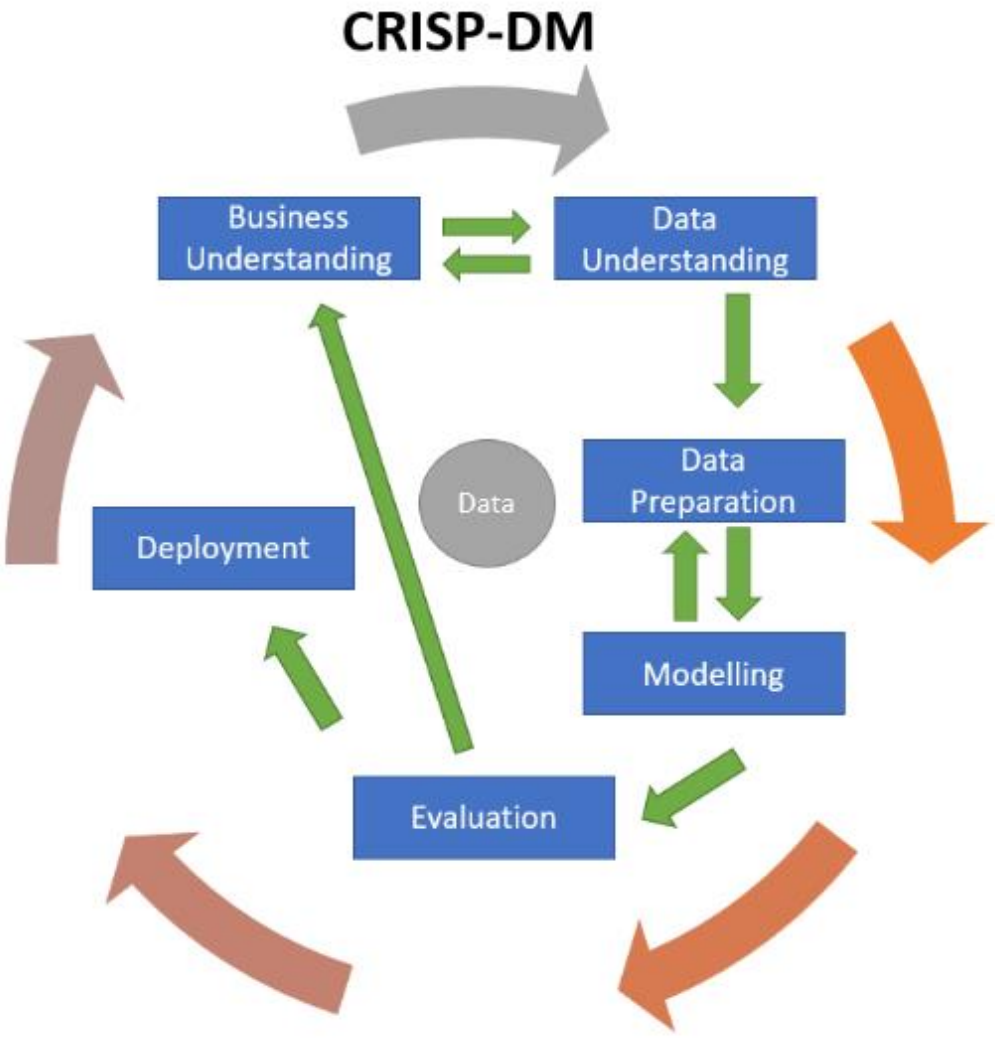

Figure 5.1 CRISP-DM Process Diagram

Chapter 5 explains the CRISP-DM framework. The framework comprises 6 different phases. The next chapter 6 will help to get a clear picture of the business problem and provides the foundation to remain more focused and engaged solely on the business objectives. Threads from Business understanding are gathered to more or less get a complete overview and blue wire print of the different consecutive phases of the data mining process.

Chapter 6: Business Understanding

There exist 2 main conventional approaches to the analysis of the stock markets: (1) Fundamental analysis and (2) Technical analysis.

Fundamental Analysis:

The fundamental analysis calculates a real worth of a sector or company and determines the number that one share of that company ought to price. The following are major methods that might be thought of in fundamental Analysis.

Valuations Strategies: Valuations square measure used as a very important strategy for selecting smart stocks at an occasional value or undervalued value with an honest margin of safety. The margin of safety is the distinction between the current value and intrinsic price, i.e., the Current value ought to be less than the intrinsic price of the stock. However, one has got to take care as valuation alone may be dishonorable. So, in conjunction with valuation, a corporation should even have quality and growth which can facilitate the “when” or the proper Entry rate or right time to take a position. following square measure, the parameters that require to require into thought whereas selecting Valuation’s strategies. These square measure DCF valuation, Graham valuation, Earning valuation, Yearly PE ratio, Quarter trailing PE, Latest PB ratio, Price/Sales, Enterprise Value/EBIT, and EBIT/Enterprise price.

Action or Momentum Strategies: Action or Momentum ways square measure based on value. therefore "Action" offers the proper Entry rate or “when” to take a position and offers one of the winners within the market. One should always watch out concerning investment in corporations that has quality and growth fundamentals in conjunction with momentum. Volume conjointly plays an integral part in momentum. following square measure, the parameters that require to require into thought whereas selecting Action or Momentum strategies. These square measures Last one Year performance,1M, 3M, 6M Performance,1 Year performance ignoring the last one month, Number of days positive value performance in a Year, return from fifty-two week high, return from fifty-two week low, Support & Resistance levels.

Long-term Quality Strategies: Long-term Quality is the most vital strategy to select Quality stocks. This can be the inspiration stone, that tells one "what" and “why” some stocks square measure higher than the opposite and helps to get rid of concern & Greed while investing. once one follows these elementary parameters, then the stocks perform even in falling markets. Most of the components of management integrity are covered by quality. Following square measure, the parameters that require to require into thought when selecting long-run Quality strategies. These square measure ROE & ROCE > fifteen, Free cash flow > zero, Debt to Equity magnitude relation < 0.30.

Using Growth Strategies: using Growth may be a strategy that focuses on parameters like sales and net growth in corporations. it's supported “what” the businesses have achieved Quarterly, 0.5 yearly & annually, and "why" must invest in them. Following square measure, the parameters that require to require into thought when selecting Growth strategies. These squares measure Sales, EBIT, Net Profit, and EPS.

Exit or Risk Parameters: Exit or Risk Parameters square measure supported those parameters and values, that build some stocks risky to take a position in. following square measure the parameters that require to require into thought whereas selecting Exit or Risk Parameters. These square measure High DE ratio, Promoter Pledge, terribly low Volume or turnover, Yearly & Quarterly net loss, Negative Book value, Mutual Funds Holding - zero or low, establishment Holding – zero, quarterly de growth in Sales & EPS.

Technical Analysis:

Technical analysis is the study of stock prices to create a profit or to create higher investment selections. The technical analysis predicts the direction of the longer-term value movements of stocks supported by their historical knowledge and helps to research financial time series knowledge using technical indicators to forecast stock prices. Meanwhile, it is assumed that the price moves according to a trend and has momentum. The price would be thought-about high, low or open, or the close value of the stock, wherever the time points would be daily, weekly, monthly, or yearly. Dow's theory puts forward the most important principles for technical analysis which says that the market value discounts everything, worth value moves in trends, and historic trends sometimes repeat identical patterns.

There are many technical indicators, like the Moving Average (MA), Moving Average Convergence/Divergence (MACD), the Aroon indicator, and also the cash flow index, etc. The evident flaws of technical analysis square measure that expert's opinions outline rules in technical analysis, that remains static and are reluctant to vary.

Modern Approaches for stock exchange Prediction:

Hypothesis Testing: Hypothesis testing may be considered a mathematical tool for confirming a monetary or business claim or plan. Hypothesis testing is beneficial for investors attempting to decide on what to take a position in and whether or not the instrument is probably going to produce a satisfactory return.

Despite the existence of various methodologies of hypothesis testing, constant four steps are used: outline the hypothesis, set the factors, calculate the data points, and reach a conclusion. This mathematical model, like most applied mathematics tools and models, has limitations and is liable to bound errors, necessitating investors conjointly consider different models in conjunction with this one.

Hypothesis testing starts by stating and assuming a null hypothesis(H0) then the method determines whether or not the belief is probably going to be true or false. The vital purpose to notice is that it is being square measured about testing the null hypothesis as a result of which there is a component of doubt concerning its validity. whatever data that is against the declared null hypothesis is captured within the alternative Hypothesis (H1). If the likelihood of obtaining a sample mean is smaller amount than five-percent, then the conclusion is to reject the null hypothesis. Otherwise, settling for and retaining the null hypothesis is being concluded.

ARIMA Model: ARIMA may be a technique for forecasting or projecting future outcomes supported by historical statistics. it's supported by the applied math construct of serial correlation, wherever past information points influence future information points.

Machine Learning Approach: In the supervised learning approach, the named input data and so the specified output are given to the learning algorithms. Meanwhile, in the unsupervised

learning approach, the unlabeled input data is provided to the learning formula, and so the formula identifies the patterns and generates the output consequently.

Prediction with Deep Learning Approach: Deep learning can be considered a kind of machine learning, that is a neural network with 3 or a lot of layers. These neural networks plan to simulate the behavior of the human brain—although aloof from matching its ability—allowing it to "train" from giant amounts of information. whereas a neural network with one layer will still build approximate predictions, extra hidden layers will facilitate optimizing and refining for accuracy.

Sentiment Analysis Approach: One of the phenomena of current times that are remodeling the world is the world's availability of information superhighway. The most-used platforms on the information superhighway are social media. it's calculated those social media users in every place around the globe will range from around 3.07 billion. There's a high association between stock prices and events related to stocks on the information superhighway. The event data is extracted from an information superhighway to predict stock prices; the such approach is known as the event-driven stock prediction. Through social networks, people generate tremendous amounts of data that are full of emotions. Much of this data is expounded on user perceptions and problems. Sentiment analysis may be a field of study that deals with people's problems, beliefs, emotions, perceptions, and sentiments towards some entity. it is the technique of analyzing text corpora, e.g., news feeds or stock exchange-specific tweets, for stock trend prediction. Stock Twits, Twitter, and Yahoo Finance are the standard medium for sentiment analysis and Topic Modelling.

Chapter 6 explained the different approaches helpful in predicting share market returns starting initially with the conventional approaches namely Fundamental Analysis and Technical analysis and then later getting a walkthrough on Modern Approaches for stock exchange Prediction namely Hypothesis Testing, ARIMA Modelling, Machine Learning Approach, Prediction with Deep Learning and then finally the Sentiment Analysis Approach.

The successive chapter 7 explains the Data Understanding section of the CRISP-DM framework. The data Understanding section will get a clear understanding of the dataset before data preparation, process, and analysis.

Chapter 7: Data Understanding

Daily Data of HDFC company from the year 2000 to 2021 which is traded on the stock exchange in India, is being used for this study. The National Stock Exchange of India Ltd. and the Bombay Stock Exchange Limited both list the shares of the Bank. NSE and BSE are the Indian electronic market-places for selling and buying stocks and securities. The stock tables are loaded with data that may facilitate anyone to become a savvy capitalist.

To properly scan stocks, it should first be perceived what every column within the stock chart means:

Name and symbol: This column tell us the corporate name (usually abbreviated) and also the symbol mentioned thereto. Share tables list stocks in alphabetical order symbol-wise, and anybody would like to use them all together in all stock communications.

There are completely different series columns utilized by NSE and BSE Stock exchanges. The dataset under consideration for the project is EQ. It stands for Equity. For this series, intraday commerce is feasible additionally to Delivery Trades.

The previous close nearly always refers to the previous day's final worth of security once the market formally closes for the day. It will apply to a stock, bond, commodity, futures or options contract, market index, or other security.

The opening price is the first trade worth that was recorded throughout the day’s commerce. The high is the highest worth at that a stock is listed during a period. The low is the lowest worth of the period. The previous closing is going to be a consecutive session's opening price. The last price is the one at which the foremost recent transaction happens. The close is the last commerce worth recording once the market is closed on the day

The volume-weighted average worth (VWAP) may be a technical analysis indicator used on intraday charts that resets at the beginning of each new commerce session. it is a commerce benchmark that represents the typical worth which the security listed throughout the day, based on both volume and worth.

Trading Volume shows the number of shares listed for the day, listed in lots of 100 quantities of shares. Share turnover may be an estimation of stock liquidity, calculated by dividing the whole number of shares traded throughout some period by the average number of shares outstanding for the same duration of time.

Chapter 7 explained the HDFC stock-related feature variables that may be used as the independent variables. The close price of the HDFC stock represents the Target or dependent variable utilized in the Modelling algorithms. Different Modelling algorithms are utilized one by one for the target variable which is the close price of the HDFC stock and the findings are being compared in Leader Boards for the Target variable. The successive chapter 8 explains the Data Preparation section of our CRISP-DM framework. Within the data preparation section, the data will be cleaned and remodeled before process and analysis.

Chapter 8: Data Preparation

Handling Missing values: Three of the features—Trades', 'Deliverable Volume','% Deliverable had quite one hundred periods missing values therefore those columns need to be dropped as they are having several missing values. Implementing the mean, median, and mode imputation methodology needs to have refrained commonly because those might render values that may introduce bias into the dataset. Second, the strategy solely looks at the variable itself and therefore might come up with values that don't seem to be representative of trends within the dataset.

Features Addition: Additionally, computed variables were added to the dataset that for sure would influence stock returns. These are moving averages for rolling periods of seven days,13 days,20 days,100 days, and two hundred days. conjointly enclosed were EMA for seven days,13 days,20 days,100 days, and two hundred days. That's going to be useful in evaluating the securities market returns. one day's previous lag values of volume are also added in the concert of the input feature. The prediction has its uncertainty; however, these indicators have helped monetary economists in the past perceive the longer-term movement of the stock costs. Analysis of the connection between extra added features and securities market returns are explored and therefore the analysis findings indicate that there are key options just like the ones that are embraced in the analysis, which demonstrated the existence of a correlation between those options and stock markets' returns.

Data Scaling using MinMax Scaler: Many machine learning algorithms work higher when features are on a relatively similar scale and close to normally distributed. MinMaxScaler, RobustScaler, StandardScaler, and normalizer are scikit-learn ways to preprocess info for machine learning. The methodology which is needed to be deployed depends on the model kind and feature values.

Data Scaling is a data preprocessing step for numerical variables. several machine learning algorithms like the Gradient descent process, KNN algorithmic rule, linear and logistical regression, etc. need data scaling to supply sensible results. varied scalers are defined for this purpose. The fit(data) methodology is employed to work out the mean and std dev for a given feature so that it will be used further for scaling. The transform(data) methodology is employed

to perform scaling using mean and std dev calculated using the fit () methodology. The fit transform () method does both fit and transform.

MinMax Scaler is one of the approaches to data scaling that is being used. Here, the minimum of features is created up to zero, and the most of features are up to one. MinMax Scaler shrinks the data inside the given range, sometimes from zero to one. It transforms data by scaling variables to a given range. It scales the worth to a selected value range while not varying the form of the initial distribution. Chapter 8 is intended on making ready the data to be future-ready for the Model Building processes. the successive chapter 9 explains the Data Modelling section of the CRISP-DM framework.

Chapter 9: Data Modeling

A rule-based model is being developed to do hypothesis testing to determine whether the chosen stock's price is crossing any of the following moving averages: the 7-day, 13-day, 20-day, 100-day, and 200-day moving averages. It will be a purchase decision if the projection indicates that the value will be higher than various Moving Averages. Exponential Time series Models will be used to create the same five hypothesis testing models. After that, five further ARIMA-based time series models will be created to support the buy or sell recommendation for every stock.

The idea is to determine how much profit, assuming $10,000 is invested in HDFC stock, will result from the forecasting outputs from these 15 various models.

HDFC excel data is put in Tabular form in step 1. The time series data is plotted for the HDFC stock that is provided as a dataset for the project for all ten years. The 7-day moving average time series data is added in step 3. The data for a 7-day moving average time series is being plotted. The data from a rolling 7-day moving average is included in the Data frame. It is determined whether the closing price value on a certain prior day was lower or higher than the current 7-day moving average.

If yesterday's closing price was below the 7-day moving average and the overall trend is upward, the stock price is likely to increase tomorrow. It will serve as the hypothesis testing rule. It is to be determined how frequently the price rise predicted by the hypothesis testing is the same as the actual price rise for the next day.

It is necessary to repeatedly verify the hypothesis testing rule's percentage accuracy. The T-test can be used to perform hypothesis testing if the sample size for testing is lesser than 30 samples. Z-Test can be used to validate null and alternate hypothesis testing for samples larger than 30.

The same step is performed for the moving averages of 13 days, 20 days, 100 days, and 200 days. Step 8: EMA is used to recreate the five different models created using SMA.

ARIMA Time series modelling is used to create an additional five different models. The construction of all 15 models, as seen above, will be used to forecast day trading in the stock market.

When the majority of the 15 various models or all of them move in the same direction, a choice on whether to purchase or sell the stock must be made. What works in the Indian stock market must be proven with evidence. Any stock on the stock market can utilize the same procedure to forecast buy or sell choices, which is helpful.

various Classification models namely AutoKeras Classification Model (Structured Data Classifier), K-neighbours Classifier Model, and Logistic Regression Classification Model deployed and their prediction accuracy is being compared with SMA Models, EMA Models, and ARIMA Models. further ahead various Regression Models including both Machine Learning and Deep learning techniques are deployed and Metrics namely MAE and MAPE are deployed to estimate the quality of the predictions on the close price of the HDFC share. These Regression Models are the OLS-Linear Regression Model, Lasso Regression Model, Lasso regression Model Using Cross Validation, The KNN Algorithm, Decision Tree Algorithm, GridSearchCV Algorithm with Hyperparameter Tuning, Random Forest Regression Model, XGBoost ML Model, Using PCA with LSTM, Using PCA with LSTM with Moving Average variables (Feature Engineering), LSTM Neural Network Model, Regression Model using AutoKeras.

The previous chapter 9 focuses on employing various Modelling algorithms to predict the Target variable value and determine the accuracy of the trend prediction as well. The next chapter 10 speaks about the Data Evaluation phase of the CRISP-DM framework. The Data Evaluation phase is the results of the Data Modelling phase and discusses the Metrics utilized to determine the extent of successes achieved from the different Modelling Algorithms employed on the Target Variable.

Chapter 10: Data Evaluation

Here the input variables or options are Open, High, Low, and the Last price for HDFC stocks collected on a day-to-day basis. The other Feature variables used are VWAP, Volume, and Turnover. The Target variable is taken as the close price for HDFC stocks. All Building models ought to be evaluated for all the anticipated close values of HDFC share vs. Actual values.

Feature Engineering comprised of explanation of further options from the close price particularly moving averages for rolling periods of seven days,13 days,20 days,100 days, and 2 hundred days. Different other feature Variables are also collectively derived, particularly EMA for seven days,13 days,20 days,100 days, and 2 hundred days.1day previous lag values of volume is being formulated and creates as well as part of the input feature variables.

Initially, A rule-based model is being developed to try to do hypothesis testing to determine whether or not the chosen stock's price is crossing any of the moving averages mentioned as on top. prediction based on the Hypothesis Testing Rule is compared with the actual trend to evaluate the accuracy for predicting the upward Trend or Downward trend of the HDFC shares.

Then a few Classifications Based Models will be conjointly built. Metrics being employed for classification Models would be accuracy score and confusion matrix which can facilitate in determining the accuracy of predicting the upward Trend or Downward trend of the HDFC shares.

The scikit learn accuracy score works with multilabel classification within which the accuracy score operates and calculates subset accuracy. Accuracy is solely the number of correct predictions divided by the overall number of examples.

One method for summarising a classification formula's performance is to use a confusion matrix. Classification accuracy alone is deceiving if there is an unequal range of observations in every category or if you have got more than 2 categories in your dataset. Calculating a confusion matrix will offer you a far better plan of what your classification model is obtaining the right and what varieties of errors it is creating.

A confusion matrix is an outline of prediction results on a classification Model. The number of correct and incorrect predictions are summarized with count values and split by every category. This is often the most important aspect related to the confusion matrix. The confusion matrix shows how the classification model is confused once it makes predictions. It offers insight not solely into the errors being created by the classifier however a lot more significantly it hints at the kinds of errors that are being created. It is this one aspect where it scores over classification accuracy.

Following that five ARIMA models are created using Moving Average as the Target variable because it would smoothen the curve for the close price of the HDFC stock worth.

When a model is created for prediction functions in statistic Time series analysis, a stationary time series is required for a higher prediction. Hence, the opening move to figure on modelling is to create a Time series stationary. Testing for stationarity may be an often-times used activity in autoregressive modelling. Numerous tests are performed just like the KPSS, Phillips–Perron, and ADF.

ADF test is a statistical significance test which means the test will end up in hypothesis tests with null and alternative hypotheses. As a result, it gives a p-value from those inferences needs to be formed regarding the Time series as to whether it is stationary or not.

To perform the ADF test in any statistic package, the stats model provides the implementation operation adfuller (). Function adfuller () provides the subsequent data particularly p-value, Value of the test statistic, Number of lags for testing consideration, and critical values.

if the results of the ADF test are bigger than 0.05 then it is required to fail to reject Null Hypothesis H0 and are available to reasoning that point Series is not Stationary. If the results of the ADF test would are lesser than 0.05 then it is required to reject Null Hypothesis H0 and is available to reasoning that point Series is Stationary.

In all results of the ADF test for ARIMA Modelling on the dataset for HDFC stock, the p-value obtained was bigger than 0.05 thus the null hypothesis is not rejected and it is concluded that the statistic for Dataset under consideration is non-stationary.

For most Time series patterns, one or a pair of differences is critical to creating a stationary Time series. ADF test would facilitate verifying the order of differencing needed to create the statistic stationary before the ARIMA Models are built for the Time series info.

Conjointly the Autocorrelation plot and also the partial autocorrelation plot of the statistic Time series information can be evaluated to work out Auto Regressive Moving Average (ARMA) models for Time series analysis. Understanding Autocorrelation operation (ACF), and Partial autocorrelation operation (PACF) plots of the series are necessary to work out the order of AR and/ or MA terms.

Auto ARIMA models will be built on the statistical Time series dataset. The auto Arima is an automatic Arima operation, which is formed to seek out the optimum order and also the optimum seasonal order, supported on determined criteria like AIC, BIC, and among the selected parameter restrictions, that matches the most effective model to one variable (univariable) time series.

Next, the different Regression Models are being built using each of the Machine Learning and Deep Learning algorithms to work out the Accuracy in predicting the expected close price of the HDFC stock that is that the Target or dependent variable for the Modelling Algorithms. The metrics that need to be verified for the accuracy of predictions in the case of regression Modelling are MAE, MSE, RMSE, Median Absolute Error (MAE), and MAPE. The Median absolute error is robust to outliers. The loss is calculated by taking the median of all absolute variations between the target and the prediction variable. Model performance is being evaluated supported on the above-discussed metrics for the various Models designed for the project.

Chapter 11: Deployment

Major Action Items Implemented:

SMA EMA T Test Metrics:

HDFC excel information is placed in Tabular type. The data from a rolling 7-day moving average is enclosed within the Data frame. It is determined whether or not the closing price worth on a particular previous day was lower or more than this 7-day moving average. If yesterday's terms were below the 7-day moving average and also the overall trend is upward, the stock worth is probably going to be bullish tomorrow. It will function as the hypothesis testing rule. it is to be determined how often the value rise expected by the hypothesis testing is the same as the actual price rise for the successive day.

The hypothesis testing rule's proportion accuracy is repeatedly verified. The T-test is employed to perform hypothesis testing because the sample size for testing is lesser than thirty samples. An equivalent step is performed for the moving averages of thirteen days, and 20 days. EMA with seven days,13 days, and twenty days spans are employed to recreate the various models that were created using SMA in earlier steps.

Date	Close	SMA_7	SMA_13	SMA_20	EMA_7	EMA_13	EMA_20
2000-01-03	170.00	170.0000	170.0000	170.0000	170.0000	170.0000	170.0000
2000-01-04	173.80	171.9000	171.9000	171.9000	170.950000	170.542857	170.361905
2000-01-05	166.95	170.2500	170.2500	170.2500	169.950000	170.029592	170.036961
2000-01-06	168.30	169.7625	169.7625	169.7625	169.537500	169.782507	169.871537
2000-01-07	168.35	169.4800	169.4800	169.4800	169.240625	169.577863	169.726628

Table 11.1– Top five rows for HDFC Dataset including SMA and EMA variables for the T-Test based on Hypothesis Testing

The Leader Board gives the following results:

SERIAL NUMBERS	DESCRIPTIONS	TOTAL	TRUE COUNT	FALSE COUNT	EFFICIENCY
SMA7	SMA-7 samples	5297	4114	1183	77 . 67
SMA13	SMA-13 samples	5291	3474	1817	65.66
SMA20	SMA-20 samples	5284	3217	2067	60.88
EMA7	EMA-7 samples	5297	4077	1220	76.97
EMA13	EMA-13 samples	5291	3486	1805	65.89
EMA20	EMA-20 samples	5284	3236	2048	61.24

Table 11.2– Leader Board-comparison of Metrics for SMA and EMA variables as per T Test based on Hypothesis Testing

From Table 11.2, It can be observed that T-test Hypothesis testing done for a rolling 7-day moving average data has given the highest efficiency in correctly predicting the upward or downward trend closely followed by EMA with a span of 7-days. however, prediction efficiency is least for 20 days of SMA data and 20-days of EMA data.

SMA EMA Z Test Metrics:

HDFC excel information is placed in Tabular type. The data from a rolling 100-day moving average is enclosed within the Data frame. It is determined whether or not the closing price worth on a particular previous day was lower or more than this 100-day moving average. If yesterday's value were below the 100-day moving average and also the overall trend is upward, the stock worth is probably going to be bullish tomorrow. It will function as the hypothesis testing rule. it is to be determined how often the value rise expected by the hypothesis testing is the same as the actual price rise for the successive day.

The hypothesis testing rule's proportion accuracy is repeatedly verified. Z-test is employed to perform hypothesis testing because the sample size for testing is more than 30 samples. An equivalent step is performed for the moving averages of 200 days. EMA with 100 days and 200 days spans are employed to recreate the various models that were created using SMA in earlier steps.

Date	Close	SMA_100	SMA_200	EMA_100	EMA_200
2000-01-03	170.00	170.0000	170.0000	170.0000	170.0000
2000-01-04	173.80	171.9000	171.9000	170.075248	170.037811
2000-01-05	166.95	170.2500	170.2500	170.013361	170.007086
2000-01-06	168.30	169.7625	169.7625	169.979433	169.990101
2000-01-07	168.35	169.4800	169.4800	169.947167	169.973781

Table 11.3– Top five rows for HDFC Dataset including SMA and EMA variables for the Z Test based on Hypothesis Testing

The Leader Board gives the following results:

SERIAL NUMBERS	DESCRIPTIONS	TOTAL	TRUE COUNT	FALSE COUNT	EFFICIENCY
SMA100	SMA-100 samples	5204	2798	2406	53.77
SMA200	SMA-200 samples	5104	2754	2350	53.96
EMA100	EMA-100 samples	5204	2829	2375	54.36
EMA200	EMA-200 samples	5104	2779	2325	54.45

Table 11.4– Leader Board-comparison of Metrics for SMA and EMA variables as per Z Test based on Hypothesis Testing

From Table 11.4, It can be observed that Z-test Hypothesis testing done for a rolling 100-day moving average and 200-day moving average has given lesser efficiency in correctly predicting the upward or downward trend compared to the prediction done with Hypothesis testing done on smaller samples using T-test Hypothesis testing. Similar inferences can be drawn for EMA with 100 days and 200 days span as well.

Classification Model Metrics:

Prev Close price, Open price, High price, Low price, Last, price, VWAP, Volume_lag_1d, and Close price are being used as Feature Variables.

Feature Engineering is being performed and direction is derived as Target Variable to predict the direction of the close price based on the Feature Variables as independent variables.

Auto Keras Classification Model (Structured Data Classifier), KNN Classification Model, and Logistic Regression Classification Modelling techniques are deployed to predict the direction of the close price.

Date	Prev Close	Open	High	Low	Last	VWAP	Volume_lag_1d	Close	direction
2000-01-04	170.00	182.00	183.45	171.00	174.00	174.99	33259.0	173.80	1
2000-01-05	173.80	170.00	173.90	165.00	168.00	169.20	168710.0	166.95	0
2000-01-06	166.95	168.00	170.00	165.30	168.95	168.44	159820.0	168.30	1
2000-01-07	168.30	162.15	171.00	162.15	170.75	166.79	85026.0	168.35	1
2000-01-10	168.35	172.90	179.50	165.00	166.30	167.79	85144.0	165.90	0

Table 11.5– Top five rows for HDFC Dataset including direction as Target Variable for Classification Modelling

The Leader Board gives the following results:

SERIAL NUMBERS	DESCRIPTIONS	TOTAL	TRUE COUNT	FALSE COUNT	EFFICIENCY
Structured Data Classifier	Auto Keras Classification Model	1061	901	160	84.92
K Neighbors Classifier	KNN Classification Model	1061	786	267	74.08
Logistic Regression	Logistic Regression Classification Model	1061	956	97	90.10

Table 11.6– Leader Board-comparison of Metrics for Accuracy Predictions on Close price of HDFC Share by different Classification Models

From Table 11.6, It can be observed that Logistic Regression Classification Model and Auto Keras classification Model have given the accuracy of near about 85 to 90% in able to correctly predict the direction of the close price. The highest Accuracy in predicting the direction by Hypothesis Testing using SMA and EMA was near about 77%. Hence, it can be safely concluded that Deep Learning models and Machine Learning Models were able to provide better outputs compared to Statistical methods of Hypothesis Testing.

ARIMA Models Metrics:

Time series data is being used as an input variable for the Auto Arima Time series Modelling Technique. Feature Engineering is done and rolling 7-day moving average,13-day moving average,20-day moving average,100-day moving average, and EMA with 200 days span as Target Variables for 5 different Auto Arima Models are being derived to predict the value of the close price based on the Time series data as an input variable.

The Leader Board gives the following results:

SERIAL NUMBERS	DESCRIPTIONS	MAE FOR TEST DATA	MSE FOR TEST DATA	RMSE FOR TEST DATA	Median Absolute Error FOR TEST DATA	MAPE FOR TEST DATA
EMA_200ARIMA	Auto Arima model using EMA-200 samples	84.21	9662.99	98.30	96.06	Nan
SMA_100ARIMA	Auto Arima model using SMA-100 samples	112.25	19404.28	139.30	95.51	9.42
SMA_20ARIMA	Auto Arima model using SMA-20 samples	183.76	45227.79	212.67	181.82	16.29
SMA_13ARIMA	Auto Arima model using SMA-13 samples	184.73	44482.52	210.91	172.64	16.171
SMA_7ARIMA	Auto Arima model using SMA-7 samples	185.64	47486.11	217.91	173.93	15.09

Table 11.7– Leader Board-comparison of Metrics for Accuracy Predictions on Close price of HDFC Share by ARIMA Models

from Table 11.7, In all results of the ADF test for ARIMA Modelling on the dataset for HDFC stock, it can be seen that the p-value obtained was bigger than 0.05 thus the null hypothesis is not rejected and concluded that the statistic for Dataset under consideration is non-stationary. It can be observed that MAE, MSE, RMSE, Median Absolute Error, and MAPE are far too high in the case of all Auto ARIMA Modelling. Hence, it can be concluded that the dataset under consideration was not suitable for Time series Modelling using the ARIMA Modelling algorithm.

Regression Models-Part1 Metrics:

Prev Close price, Open price, High price, Low price, Last, price, VWAP, and Volume_lag_1d are being used as Feature Variables. Close price is being used as Target Variable to predict the values of the close price based on the Feature Variables as independent variables. OLS-Linear Regression Model, Lasso Regression Model, and Lasso regression Model are being deployed Using Cross-Validation, and the KNN Algorithm as the Modelling techniques to predict the close price of the HDFC share.

Date	Prev Close	Open	High	Low	Last	VWAP	Volume_ lag_1d	Close
2000-01-04	170.00	182.00	183.45	171.00	174.00	174.99	33259.0	173.80
2000-01-05	173.80	170.00	173.90	165.00	168.00	169.20	168710.0	166.95
2000-01-06	166.95	168.00	170.00	165.30	168.95	168.44	159820.0	168.30
2000-01-07	168.30	162.15	171.00	162.15	170.75	166.79	85026.0	168.35
2000-01-10	168.35	172.90	179.50	165.00	166.30	167.79	85144.0	165.9

Table 11.8– Top five rows for HDFC Dataset including Close as Target Variable for Regression Modelling

The Leader Board gives the following results:

SERIAL NUMBERS	DESCRIPTIONS	MAE FOR TEST DATA	MSE FOR TEST DATA	RMSE FOR TEST DATA	Median Absolute Error FOR TEST DATA	MAPE FOR TEST DATA
OLS Model	OLS-Linear Regression Model	2.03	11.83	3.44	1.14	0.227
LASSO Model	Lasso Regression Model	7.56	132.63	11.52	4.67	0.85
CVLASSO Model	Lasso regression Model Using Cross-Validation	7.55	132.59	11.51	4.66	0.85
KNN Model	KNN Algorithm	5.42	132.08	11.49	3.16	0.59

Table 11.9– Leader Board-comparison of Metrics for Predicting Close price of HDFC Share by part1 Regression Models

From Table 11.9, It can be observed that MAE and MAPE were satisfactory for the OLS-Linear Regression Model. However, other Regression Models were not able to provide MAPE within the acceptable range.

Regression Models-Part2 Metrics:

Prev Close price, Open price, High price, Low price, Last, price, VWAP, and Volume_lag_1d are being used as Feature Variables. Close price is being used as Target Variable to predict the values of the close price based on the Feature Variables as independent variables. Decision Tree Algorithm, GridSearchCV Algorithm with Hyper-parameter Tuning, Random Forest Regression Model, and XGBoost ML Model is being deployed as the Modelling techniques to predict the close price of the HDFC share.

The Leader Board gives the following results:

SERIAL NUMBERS	DESCRIPTIONS	MAE FOR TEST DATA	MSE FOR TEST DATA	RMSE FOR TEST DATA	Median Absolute Error FOR TEST DATA	MAPE FOR TEST DATA
DT Model	Decision Tree Algorithm	3.26	23.95	4.89	2.10	0.383
GRIDSEARCHCV Model	GridSearchCV Algorithm with Hyper-parameter Tuning	3.22	23.16	4 . 81	2.10	0.38
RF Model	Random Forest Regression Model	2.45	15.25	3.90	1.49	0.29
XGBOOST Model	XGBoost ML Model	3.25	22.78	4.77	2.12	0.37

Table 11.10– Leader Board-comparison of Metrics for Predicting Close price of HDFC Share by part2 Regression Models

from Table 11.10, It can be observed that MAE and MAPE were satisfactory for Random Forest Regression Model. However, other Regression Models were able to provide fairly acceptable MAPE but still lower MAPE would have been better.

Regression Models-Part3 Metrics:

Prev Close price, Open price, High price, Low price, Last, price, VWAP, and Volume_lag_1d are being used as Feature variables. Close price is being used as Target Variable to predict the values of the close price based on the Feature Variables as independent variables. PCA with LSTM is being deployed Using PCA with LSTM with Moving Average variables (Feature Engineering), LSTM Neural Network Model, and Regression Model using AutoKeras as the Modelling techniques to predict the close price of the HDFC share.

The Leader Board gives the following results:

SERIAL NUMBERS	DESCRIPTIONS	MAE FOR TEST DATA	MSE FOR TEST DATA	RMSE FOR TEST DATA	Median Absolute Error FOR TEST DATA	MAPE FOR TEST DATA
PCA LSTM Model	Using PCA with LSTM	4.37	34.70	5.89	3.60	33.44
PCA LSTM Moving Averages Model	Using PCA with LSTM with Moving Average variables (Feature Engineering)	7.75	135.03	11 . 62	5.99	33.47
LSTM Model	LSTM Neural Network Model	9.71	159.01	12.61	8.20	33.40
Auto Keras Model	Regression Model using AutoKeras	2.59	242.51	15 . 57	1.10	0.27

Table 11.11– Leader Board-comparison of Metrics for Predicting Close price of HDFC Share by part2 Regression Models

from Table 11.11, It can be observed that MAE and MAPE were satisfactory for both Using PCA with LSTM and Regression Model using AutoKeras.However, other Regression Models were able to provide fairly acceptable MAPE but still, their MAE would have been better.

Chapter 12: Analysis and Results

Classification Metrics Comparison:

SERIAL NUMBERS	DESCRIPTIONS	EFFICIENCY>67%
SMA7	SMA-7 samples	YES-77.67
SMA13	SMA-13 samples	NO-65.66
SMA20	SMA-20 samples	NO-60.88
EMA7	EMA-7 samples	YES-76.97
EMA13	EMA-13 samples	NO-65.89
EMA20	EMA-20 samples	NO-61.24
SMA100	SMA-100 samples	NO-53.77
SMA200	SMA-200 samples	NO-53.96
EMA100	EMA-100 samples	NO-54.36
EMA200	EMA-200 samples	NO-54.45
Structured Data Classifier	Auto Keras Classification Model	yes-84.92
K Neighbours Classifier	KNN Classification Model	yes-74.08
Logistic Regression	Logistic Regression Classification Model	yes-90.10

Table 12.1– Leader Board-comparison of Metrics for Classification Models

From Table 12.1, It can be observed that Logistic Regression Classification Model and Auto Keras classification Model have given the accuracy of near about 85 to 90% in able to correctly predict the direction of the close price. The highest Accuracy in predicting the direction by Hypothesis Testing using SMA and EMA was near about 77%. other Hypothesis testing using T-test and Z-test statistical algorithms were not satisfactory in able to predict the direction of the close price of the HDFC share.

Regression Metrics Comparison:

SERIAL NUMBERS	DESCRIPTIONS	MAE<=5	MAPE<=0.33
OLS Model	OLS-Linear Regression Model	YES-2.034	YES-0.23
LASSO Model	Lasso Regression Model	NO-7.555	NO-0.85
CVLASSO Model	Lasso regression Model Using Cross-Validation	NO-7.55	NO-0.85
KNN Model	KNN Algorithm	NO-5.423	NO-0.59
DT Model	Decision Tree Algorithm	YES-3.26	NO-0.38
GRIDSEARCHCV Model	GridSearchCV Algorithm with Hyper-parameter Tuning	YES-3.218	NO-0.38
RF Model	Random Forest Regression Model	YES-2.45	YES-0.29
XG Boost Model	XGBoost ML Model	YES-3.25	NO-0.37
PCA LSTM Model	Using Principal Component Analysis (PCA) with LSTM	YES-4.366	YES-33.44
PCA LSTM Moving Averages Model	Using Principal Component Analysis (PCA) with LSTM with Moving Average variables (Feature Engineering)	NO-7.75	YES-33.47
LSTM Model	LSTM Neural Network Model	NO-9.71	YES-33.40
Auto Keras Model	Regression Model using AutoKeras	YES-2.59	YES-0.27

Table 12.2– Leader Board-comparison of Metrics for Regression Models

From Table 12.2, It can be observed that the OLS-Linear Regression Model, Random Forest Regression Model, Using PCA with LSTM, and Regression Model using AutoKeras provide MAE<=5 and MAPE<=0.33. Hence these Regression Models were most successful in predicting the close value of the stock price. XGBoost ML Model, Decision Tree Algorithm, GridSearchCV Algorithm with Hyper-parameter Tuning provided good MAE but were slightly higher with MAPE.

The implementation for the capstone project can be accessed at the link below:
https://github.com/Embedded-org/ACCOMPLISHMENTS/tree/master/RACE%20CAPSTONE%20PROJECT1

Chapter 13: Conclusions and Recommendations for future work

The hypothesis testing rule's percentage accuracy was repeatedly verified using five SMA Models. EMA was used to recreate the five other different models created using SMA. T-test was used to perform hypothesis testing if the sample size for testing was lesser than 30 samples. Z-Test was used to validate null and alternate hypothesis testing for samples larger than 30.

ARIMA Time series modelling was used to create an additional five different models. The construction of all 15 models, was used to forecast day trading in the stock market.

Prediction accuracy was then compared with Classification Model Algorithms. When the majority of the various models or all of them move in the same direction, a choice on whether to purchase or sell the stock must be made.

This project then solely focuses on predicting the close price of the HDFC stock using Regression algorithms deploying both Machine Learning and Deep Learning Techniques.
What works in the Indian stock market must be proven with evidence. Any stock on the stock market can utilize the same procedure to forecast buy or sell choices, which is helpful.

Recommendations for Future Work: it is assumed that returns are more or less constant over time. However, the assumption that the returns are constant over time is restrictive, and not true. Returns are highly dependent on time. This project has not discussed how to address one major drawback of stock prediction, namely that over different periods the stock returns can change drastically to either extremely low returns during stock market crashes or extremely high returns during stock market booming periods.

In future projects, it can be shown how to define Bullish and Bearish regimes using modern machine learning techniques. The techniques already discussed in this project will then be used to estimate the direction of close price for each of the Normal and Crash regimes. The Sentiment Analysis Approach may also need to be explored using Text Analytics for predicting stock market returns.

Bibliography

Alhomadi, A. (2021). Forecasting stock market prices : A machine learning approach. *Digital Commons*, *11*(2), 16–36.

Biswas, M., Nova, A. J., Mahbub, M. K., Chaki, S., Ahmed, S., & Islam, M. A. (2021). Stock Market Prediction: A Survey and Evaluation. *2021 International Conference on Science and Contemporary Technologies, ICSCT 2021*, *December*. https://doi.org/10.1109/ICSCT53883.2021.9642681

Huang, Y., Capretz, L. F., & Ho, D. (2021). Machine Learning for Stock Prediction Based on Fundamental Analysis. *2021 IEEE Symposium Series on Computational Intelligence, SSCI 2021 - Proceedings*, *5*. https://doi.org/10.1109/SSCI50451.2021.9660134

Jierula, A., Wang, S., & Oh, T. (2021). *applied sciences Study on Accuracy Metrics for Evaluating the Predictions of Damage Locations in Deep Piles Using Artificial Neural Networks with Acoustic Emission Data*.

López del Val, J. A., & Alonso Pérez de Agreda, J. P. (1993). Principal components analysis. *Atencion Primaria / Sociedad Española de Medicina de Familia y Comunitaria*, *12*(6), 333–338. https://doi.org/10.5455/ijlr.20170415115235

Rouf, N., Malik, M. B., Arif, T., Sharma, S., Singh, S., Aich, S., & Kim, H. C. (2021). Stock market prediction using machine learning techniques: A decade survey on methodologies, recent developments, and future directions. *Electronics (Switzerland)*, *10*(21). https://doi.org/10.3390/electronics10212717

Series, I. (2021). Machine Learning Algorithms and Applications. In *Machine Learning Algorithms and Applications* (Vol. 7). https://doi.org/10.1002/9781119769262

Shah, D., Isah, H., & Zulkernine, F. (2019). Stock market analysis: A review and taxonomy of prediction techniques. *International Journal of Financial Studies*, *7*(3). https://doi.org/10.3390/ijfs7020026

Sonkiya, P., Bajpai, V., & Bansal, A. (2021). *Stock price prediction using BERT and GAN*. *6*. http://arxiv.org/abs/2107.09055

Vreeken, J., & Yamanishi, K. (2019). *Proceedings of the 25th {ACM} {SIGKDD} International Conference on Knowledge Discovery & Data Mining, {KDD} 2019, Anchorage, AK, USA, August 4-8, 2019*. 1946–1956. https://doi.org/10.1145/3292500

Сороко, Н. В. (2017). *Масові Відкриті Європейські Он-Лайн Курси Для Вчителів (2017 Р.). Інформаційний Бюлетень№ 1*. *801*, 1–23.

Appendix

Plagiarism Report[1]

Publications in a Journal/Conference Presented/White Paper[2]

Any Additional Details

[1] Turnitin report to be attached from the University.

[2] URL of the white paper/Paper published in a Journal/Paper presented in a Conference/Certificates to be provided.