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

JB Simha

Chief Mentor - RACE

REVA Academy for Corporate Excellence

REVA University

Rukmini Knowledge Park, Kattigenahalli,

Yelahanka, Bangalore – 560064

July 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 **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: 31 July. 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**, is a bonafide student of REVA University, is submitting the first-year project report in fulfilment for 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 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 in 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: 31 July. 22

REVA UNIVERSITY
Bengaluru, India

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: Mr. Mithun Dolthody Jayaprakash

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 regression
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 Average
6	EMA	Exponential Moving Average
7	CRISP-DM	Cross-Industry Standard Process for Data Mining
8	MAE	Mean Absolute error
9	MSE	Mean Square Error
10	MAE	Median Absolute error
11	R^2	R-squared (coefficient of determination)
12	RMSE	Root Mean square Error
13	MAPE	Mean Absolute percentage Error
14	ADF	Augmented Dickey-Fuller
15	VWAP	volume-weighted average price

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 victimization 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 our predictions.

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 being utilized to produce identical 5 hypothesis testing models. After that, 5 any ARIMA-based statistic models are being created to support our buy or sell recommendation for underlying stock.

Then various numerous Classification Models is applied particularly K neighbours Classifier, Logistic Regression Modelling, Auto Keras Classification Model using Structured knowledge classifier. Our results show that AutoKeras Classification Model achieves the most effective prediction Accuracy followed by supply Regression Classification Model Then K Neighbors Classification Model. Simple moving average-7 samples and Exponential moving average-7 samples using ttest applied mathematics Hypothesis testing Models conjointly provided fairly smart accuracy.

we then used regression Modelling Algorithms for predicting the close value and compared the Metrics particularly Mean Absolute Error and Mean Absolute Percentage Error.

Ordinary method of least squares (OLS)-Linear Regression Model, Lasso Regression Model, Lasso regression Model using Cross Validation, The K-Nearest Neighbors (KNN) rule, Decision Tree rule, GridSearchCV rule with Hyper-parameter standardization, Random Forest Regression Model, XGBoost Model, Using Principal Component Analysis (PCA) with

LSTM, Using Principal Component Analysis (PCA) with LSTM with Moving Average variables(Feature Engineering),Long short Memory-LSTM Neural Network Model, Regression Model using AutoKeras were the Regression Models used for predicting the close value.

Ordinary method of least squares (OLS)-Linear Regression Model and Regression Model using AutoKeras offer the most effective results. Random Forest Regression Model and using Principal part Analysis (PCA) with LSTM conjointly provided smart results.

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

A large quantity of analysis has been conducted within the field of stock performance prediction since the birth of this investment instrument, as investors naturally would love to take a position in stocks that they need and expected that they can outmatch the others so as to come up with profit by selling them later. 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. (Huang et al., 2021)

Results from several of the studies have shown that prediction models trained with historical worth and volume information may be with success used towards predicting. However, there is one major downside for short prediction and high frequency Trading, that is resistance value or service commission. For trading stocks through a broker, there is usually a commission paid to the broker for every purchase and sell. 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. Since the short-term prediction models from several of the studies don't incorporate resistance value in analysis, the determinateness of the studies is also affected. (Huang et al., 2021)

In this thesis, we have a tendency to aim to gauge machine learning ways for stock prediction based on elementary analysis. we have a tendency to do therefore by comparison the prediction performance of 11 advanced machine learning ways based on elementary analysis using elementary Features. To develop and check the machine learning models, information extracted from the daily Returns reports of HDFC stocks is being utilized which appeared within the National securities market between 2000 and 2021. so as to gauge the performance of various machine learning ways. Here, the different Machine Learning Models is being ranked based on their expected relative outcome. (Huang et al., 2021)

The chapter has discussed the importance of Machine-Learning associated techniques that are developed for our investments in stock market. The chapter discussed that analysing exchange movements and worth behaviours is extraordinarily difficult. The chapter also talked about the downside for short prediction and high frequency Trading, and trading stocks through a broker by paying service commission and then finally the previous chapter talked about 11 advanced machine learning ways which has been developed in the project based on elementary analysis using elementary Features. In the following chapter, we will scan through some of the literature available which would throw light on various related aspects of Machine-Learning methods and other methodologies in lieu of studying and researching other related issues which would be helpful in assisting us better on Day trading in Stock Market.

Chapter 2: Literature Review

Financial markets are going through eventual transformations via the foremost fascinating inventions of our time. They will have a significant impact on several areas like business, education, jobs, technology and therefore on the economy. Over the years, investors and researchers are inquisitive about developing and testing models of stock worth behaviour. However, analysing exchange movements and worth behaviours is extraordinarily difficult as a result of the markets 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)

There is some literature that have used both supervised and unsupervised machine learning techniques for securities market predictive modelling and located that both kind of models will create predictions with some accuracy. we have a tendency to share the assumption that even machine learning techniques haven't been ready to predict monthly securities market returns with high accuracy and that we reiterate this belief in this paper. (Alhomadi, 2021)
The cross-industry standard process for data mining or CRISP-DM is open standard process framework model for data processing project coming up with. This is a framework which many have utilized in many industrial undertakings and proved productive within the application. (Wijaya, 2021)

The two commonest metrics accustomed predict long-run value movements yearly for elementary analysis are (a) the Price to Earnings ratio quantitative relation (P/E) and (b) the Price by Book quantitative relation (P/B). The P/E ratio is employed as a predictor. The businesses with a lower P/E ratio yield higher returns than companies with a high P/E ratio. money analysts additionally use this to prove their stock recommendations. (Rouf et al., 2021).

The requirement is to beat the deficiencies of Fundamental and technical analysis, and also the evident advancement within the modelling 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)

For time-series predictions, ARIMA models were studied. A linear combination of previous worth's and past mistakes within the ARIMA model represents the longer-term value of a variable. 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 statistic is particularly economical and solid as the commonest Artificial Neural Network techniques. This model was recognized for its long-range prediction. prognosticative ARIMA model building phases involve model identification, diagnostic management and also the parameter analysis. ARIMA models were utilized to construct a comprehensive short-term stock worth prediction methodology. (Biswas et al., 2021)

Because of world digitization, SMP has entered a technological era. Machine learning in share price prediction is employed to find patterns in data. Usually, an incredible quantity of structured and unstructured heterogeneous knowledge is generated from stock markets. By exploring machine learning algorithms, it is attainable to quickly analyze additional complicated heterogeneous knowledge and generate additionally precise results. numerous machine learning strategies are used for stock exchange Predictions. The machine learning approaches are chiefly categorized into supervised and unsupervised approaches. (Rouf et al., 2021)

Long Short-Term Memory (LSTM): One modification of the RNN is that 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. The LSTM model is effective at displaying the link between nonlinear time series statistic and therefore the stock prediction aims in delayed state space.(Biswas et al., 2021)

Deep Neural Network (DNN): a minimum of one hidden layer of neural network is gift in an exceedingly deep neural network. it's going to be ready to provide complicated non-linear functions further as a large abstraction capability, implying that the model's fitting power is significantly augmented.(Biswas et al., 2021)

Reinforcement Learning: Reinforcing learning could be a sort of profound learning that focuses on responding to profits in an exceedingly specific circumstance. The two essential elements of strengthening learning are state and action. Increasing learning, which supported buying, selling and holding possibilities as final output, outlined the neural net structure, the reward and therefore the behavior of the agents. (Biswas et al., 2021)

The information generated through social media permits us to explore large and various opinions. Exploring sentiments from social media additionally to numeric time-series stock knowledge would enhance the accuracy of the prediction. Exploiting time-series knowledge further as social media knowledge would intensify the prediction accuracy. completely different approaches and techniques are projected over time to anticipate stock costs through various methodologies, because of the dynamic and difficult panorama of stock markets. (Rouf et al., 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 a 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)

Machine learning and deep learning models have found widespread applications in planning predictive frameworks for stock Market. Baek and Kim propose a framework referred to as ModAugNet, that is constructed on associate LSTM deep learning model. Chou and Nguyen predetermined a window metaheuristic improvement methodology for stock value forecast. Gocken et al. propose a hybrid artificial neural network applying harmony search and genetic algorithms to investigate the connection between numerous technical indicators of stocks and therefore the index of the Turkish exchange. (Series, 2021)

Prediction of future stock Market and value movement patterns could be a difficult task if the stock value statistic encompasses a great deal of volatility. during this chapter, we tend to

apply ten deep learning-based regression models for sturdy and precise prediction of stock costs. Among the 10 models, four of them are designed on variants of CNN architectures, whereas the remaining six are made applying totally different LSTM architectures. The historical stock value records are collected applying the Metastock tool over a span of 2 years at 5 minutes intervals. The models are trained applying the records of the first year, and they're tested on the remaining records. The testing is dispensed applying associate approach called walk-forward validation, in which, supported on the last one- or two-weeks historical stock costs, the predictions of stock costs for the 5 days of succeeding week are created. The cumulative RMSE and the RMSE for every day in a very week are computed to judge the prediction accuracy of the models. The time required to finish one spherical of execution of every model is additionally noted so as to determine the speed of execution of the models. The results disclosed some fascinating observations. First, it's found that whereas the CNN models are quicker, in general, the accuracies of each CNN and LSTM models are comparable. Second, the univariate models are quicker and more correct than their multivariate counterparts. and at last, the number of variables in a model encompasses a vital result on its speed of execution apart from the univariate encoder decoder LSTM models. (Series, 2021)

Chapter 3: Problem Statement

Stock market analysis and prediction is still an interesting and tough prospect. Today, big components of the population are excluded from the prospect of exploring monetary investments. The exchange is one in all the few investment markets technically accessible to traditional voters, and even that phenomena is relatively new. The accessibility to various markets, like commodities or property, entirely exists in theory. 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. it is unimaginable to make substantial profits in these markets whereas not having access to important amounts of capital. The individual investors with small assets can technically trade stocks. However, their ability to understand important profits is restricted due to their restricted capability to make leverage and to balance risks moderately.

Financial analysts investing in Stock markets generally do not appear to be tuned in to the exchange behaviour. They are facing issues in stock Trading as they are not able to understand which stock to shop for and which to sell so as to achieve a lot of profits. (Falinouss, 2007)

As heaps of knowledge became available, we tend to face new challenges in obtaining and processing the data to extract information and analyse the result on stock prices. These challenges include issues with live testing, algorithmic Trading, unsuccessful, long-term predictions, and sentiment analysis on company filings. (Shah et al., 2019)

Most of the stock analysis and prediction literature make an assertion about live testing that proposed techniques are utilized in real time to make profits among the exchange. it is a huge claim to make as an associated formula may go fine on back testing in controlled environments, but the main challenge is live testing, as a results of many things like price variations, and quiet news and existing noise. Hence, a viable analysis direction would be to grasp but variety of the favoured stock analysis techniques and implement those best practices in a live or simulated environments. (Shah et al., 2019)

Algorithmic Trading systems have changed the approach by which stock markets perform. Most of the Trading volumes in equity futures are generated by algorithms 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 behaviour. With new algorithms continuing to flood the markets every day, comparison of the effectivity and accuracy of these algorithms pose nonetheless an added challenge. (Shah et al., 2019)

The chapter has mentioned that the exchange is technically accessible to traditional voters, however still these markets are dominated by big investors, perpetuating the wealth divide. As explained earlier, there are many challenges involved with live testing, algorithmic Trading, unsuccessful, long-term predictions, and sentiment analysis on company filings. Nevertheless, any derived and associated formula may go fine on back testing in controlled

environments, but the main challenge is live testing. Also, algorithmic Trading gives many benefits but retail investors do not have the desired technology to create such systems. Within the following chapters, we will introspect more on objectives that would be more probably achieved through the effort put on the present project.

Chapter 4: Objectives of the Study

The monetary market could be an advanced, organic process, and non-linear dynamical system. The sector of Trading forecasting is characterised by information intensity, noise, non-stationary, unstructured nature, high degree of uncertainty, and hidden relationships. several factors move in finance together with political events, general economic conditions, and traders' expectations. Therefore, predicting worth movement in monetary markets is kind of difficult. (Falinouss, 2007)

Majority of share market investment selections are extremely influenced by herd mentality. To avoid herd mentality, we must always not blindly invest in any stock based on what others do. once creating investments in Stocks, the target of our analysis paper is to review the connected factors a lot fastidiously.

In some industries, information is not simply available on the market and that we have to be compelled to place in further efforts to seek out and valuate information and create wiser selections consequently. The main motivation for predicting changes in stock price is the potential financial returns.

An oversized quantity of analysis has been conducted within the field of stock performance prediction since the birth of this investment instrument, as investors naturally would love to invest in stocks that they have predicted can outdo the others so as to generate profit through them later. (Huang et al., 2021)

We write this paper below with the assumption that markets are efficient and the type of efficiency markets replicate is the semi-strong form. we have a tendency to believe that the securities market reflects all publicly available information on the market data and consequently historical information can't be able to predict future securities market returns.

Nevertheless, machine learning techniques have created it fairly achievable with their process capabilities to use historical and high dimensional information to create multiple periods predictions and ensure to enable checking the market's potency with their advanced process capabilities. (Alhomadi, 2021)

This analysis topic can continuously be vital for investors who seeks to come up with excess returns, create buy-sell selections, and confirm portfolio allocation, significantly considering

development in machine learning techniques and algorithmic improvement. In an inefficient market, investors can have a troublesome time achieving those ends as a result of volatility within the market. Volatility creates opportunities for gains and losses; however, it makes investors investment decision-making tougher. (Alhomadi, 2021)

This paper is tributary to the continuous research initiatives towards analysis of assessing securities market returns predictability and market potency. Even if we have a tendency to believe predicting securities market returns with high accuracy using daily or monthly returns is troublesome, investors will still use the paper's findings to assist them guide their quality allocation, create buy-sell selections knowing that the National exchange market is economical, and formulate optimum portfolios that best meet the clients' needed returns. Also, machine learning techniques explained here will be useful in providing insights for profits, and can be considered vital in influencing the securities market returns. (Alhomadi, 2021)

The chapter mentions that the main motivation for predicting changes in stock price is the potential financial returns. The chapter discusses that even if we have a tendency to believe predicting securities market returns with high accuracy using daily or monthly returns is troublesome, objective of the study is that investors should still be able to utilize the paper's findings to assist them guide their quality allocation and create buy-sell selections that best meet their needed returns expectations. Within the following chapters, we will introspect more on project Methodology that would be implemented and endeavours for continuous improvement that will be taken up while working on the project.

Chapter 5: Project Methodology

The CRISP-DM framework has been used here 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 towards the requirement between phases also are vital with one another phase; the outer circle represents the cyclic properties of the framework. (Wijaya, 2021)
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 we will learn new things throughout the method, and it may trigger alternative business queries. (Wijaya, 2021)

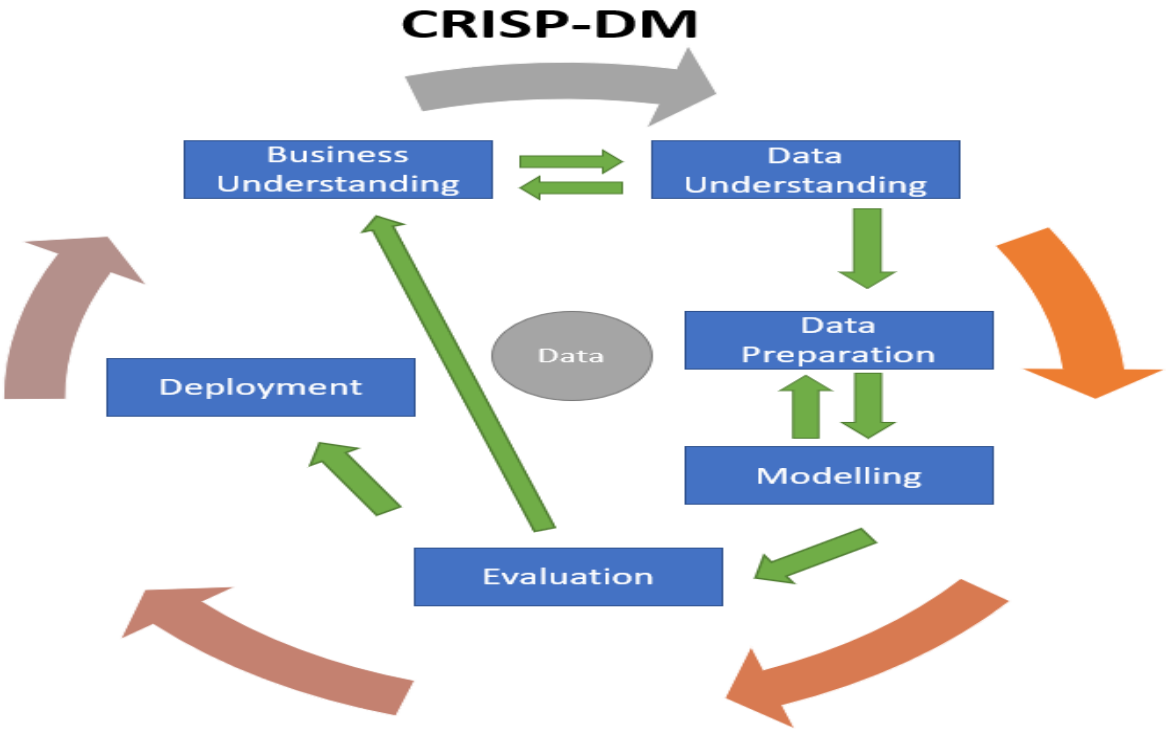

Figure 5.1 CRISP-DM Process Diagram (Wijaya, 2021)

The chapter explains the CRISP-DM framework. The framework comprises 6 different phases. The next chapter explained about Business Understanding phase. This helps us to get a clear picture of our business problem and provides the foundation to remain more focussed and engaged solely on our 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 our 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:

Fundamental analysis calculates a real worth of a sector or company and determines the number that one share of that company ought to price. A supposition is formed that, if given enough time, the corporate can move to a price believing the prediction. (Rouf et al., 2021) If a sector or company is undervalued, then the market price of that company ought to rise, and conversely, if an organization is overvalued, then the value ought to fall. The analysis is performed considering varied factors, like yearly business summaries and reports, balance sheets, a future prospectus, and also the company's work atmosphere. (Rouf et al., 2021)

following area unit 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. Margin of safety is the distinction between current value and intrinsic price, i.e., Current value ought to be less than 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 within 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, 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 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. This square measure Last one Year performance,1M, 3M, 6M Performance,1 Year performance ignoring 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 that 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. Quality covers most of the areas of management integrity. following square measure, the parameters that require to require into thought whereas 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" you must invest in them. Following square measure, the parameters that require to require into thought whereas selecting Growth strategies. These square measure Sales, EBIT, Net Profit, 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 terribly low, establishment Holding – zero, quarterly de growth in Sales & EPS.

Fundamental analysis may be used for the thought of monetary ratios to tell apart poor stocks from quality stocks. The P/B magnitude relation compares the corporate price specified by the market to the corporate price specified on paper. If the magnitude relation is high, the corporate is also overvalued, and also the company's price would possibly fall with time. Conversely, if the magnitude relation is low, the corporate is also underestimated, and also the value might rise with time. Of course, elementary analysis may be a powerful methodology. Still, it has some drawbacks. elementary analysis, firstly, lacks adequate data of the foundations governing the workings of the system, and second, there's non-linearity within the system. (Rouf et al., 2021)

Technical Analysis:

Technical analysis is the study of stock prices to create a profit, or to create higher investment selections. Technical analysis predicts the direction of the longer-term value movements of stocks supported 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. Technical analysis uses value charts and bound formulae, and studies patterns to predict future stock prices; it's primarily utilized by short-term investors. (Rouf et al., 2021)

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 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. (Rouf et al., 2021)

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. numerous parameters that have an effect on stock costs remain unnoticed. (Rouf et al., 2021)

Modern Approaches for stock exchange Prediction:

Hypothesis Testing:

Hypothesis testing may be considered as a mathematical tool for confirming a monetary or business claim or plan. Hypothesis testing is beneficial for investors making an attempt to

make a decision 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 considering different models in conjunction with this one.

Null Hypothesis (H0) is assumed to be true. Hypothesis testing starts by stating and assuming a null hypothesis then the method determines whether or not the belief is probably going to be true or false.

The vital purpose to notice is that we tend to square measure 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).

Calculate the statistic:
This step involves calculating test statistics (like mean, z-score, p-value for the chosen sample). With the computed value(s), the null hypothesis is determined. If the likelihood of obtaining a sample mean is a smaller amount than five-percent, then the conclusion is to reject the null hypothesis. Otherwise, settle for and retain the null hypothesis is being concluded.

ARIMA Model:
Since the prediction of the stock costs in finance and in economy is a vital issue, the eye of researchers rises to make a more robust forecast model capable of predicting correct stock costs. In 1970, Box and Jenkins planned the integrated moving average autoregressive model (ARIMA). (Biswas et al., 2021)

Machine Learning Approach:
In the supervised learning approach, named input file and therefore the desired output ar given to the educational algorithms. Meanwhile, within the unsupervised learning approach, unlabeled input file is provided to the educational formula, and therefore the formula identifies the patterns and generates the output consequently. moreover, completely different recursive approaches are utilized in SMP, like the Support Vector Machine (SVM), k Nearest Neighbors (kNN), Artificial Neural Networks (ANN), Decision Trees, Fuzzy Time-Series, and Evolutionary Algorithms. (Rouf et al., 2021)

The SVM could be a supervised machine learning technique that limits error and augments geometric margins, and could be a pattern classification formula. In terms of accuracy, the SVM is a very important machine learning formula compared to the contemporary classifiers. (Rouf et al., 2021)

In the kNN, stock prediction is mapped into a classification supported on closeness. By estimating Euclidian distance, the kNN classifies the “k” nearest neighbors within the training set. (Rouf et al., 2021)

The ANN could be a nonlinear machine structure for numerous machine learning algorithms to investigate and method complicated input data together. (Rouf et al., 2021)

The FIS (Fuzzy Inference Systems) apply rules to fuzzy sets so apply de-fuzzification to administer crisp outputs for deciding. The evolutionary algorithms embrace gene-inspired neuro-fuzzy and neuro-genetic algorithms, mimic the obvious selection theory of species, and may offer the best possible output. (Rouf et al., 2021)

Prediction with Deep Learning:

A. Convolutional Neural Network (CNN)
B. Recurrent Neural Network (RNN)
C. Long Short-Term Memory (LSTM
D. Deep Neural Network (DNN
E. Reinforcement Learning

Sentiment Analysis Approach:

One of the phenomena of current times that is transforming the globe is that the global availability of the net. The most-used platforms on the net are social media. it is calculable that social media users everywhere around the globe can number around 3.07 billion. there is a high association between stock costs and events associated with stocks on the net. The event info is extracted from the net to predict stock prices; such an approach is understood as event-driven stock prediction. Through social networks, folks generate tremendous amounts of knowledge that is stuffed with emotions. (Rouf et al., 2021)

Much of this knowledge is expounded to user perceptions and issues. Sentiment analysis could be a field of study that deals with the people’s issues, beliefs, emotions, perceptions, and sentiments towards some entity. it's the method of analyzing text corpora, e.g., news feeds or stock exchange specific tweets, for stock trend prediction. The Stock Twits, Twitter, Yahoo Finance, so on ar well-known platforms used for the extraction of sentiments. (Rouf et al., 2021)

There is a big importance of exploring sentimental knowledge for enhancing the prediction of volatility within the stock exchange. The ‘Wisdom of Crowds’ and sentiment analysis generate additional insights that may be wont to increase the performance in numerous fields, like box workplace sales, election outcomes, SMP, and so on. This means that an honest call is created by taking the opinions and insights of huge teams of individuals with varied forms of info. (Rouf et al., 2021)

The chapter explained the different approaches helpful in predicting on 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 explains the Data Understanding section of our CRISP-DM framework. Within the data Understanding section, we will get clear understanding on our dataset before data preparation, process and analysis.

Chapter 7: Data Understanding

Daily Data of HDFC company from year 2000 to 2021 which is traded on the stock exchange in India, is being used for this study. The Bank's shares are listed on the Bombay Stock Exchange Limited and The National Stock Exchange of India Ltd. NSE and BSE are the Indian electronic market-places for selling and buying of stocks and securities.

Learning a way to scan stocks by exploring stock tables meant to collect exchange data will facilitate to pick promising investment opportunities and monitor our stocks performance. The stock tables are loaded with data that may facilitate anyone to become a savvy capitalist.

A stock table could look discouraging initially as a result of tons of existent info. However, to be assured in a way to scan stocks, every information got to be understood well and Valuable insights got to be extracted from them. every column within the chart provides some clues regarding the present state of affairs for a specific company to assist oneself to build good investment selections.

To properly scan stocks, we should tend to firstly perceive 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 altogether in all stock communications.

There are completely different series columns utilized by NSE and BSE Stock exchanges.

EQ: It stands for Equity. For this series intraday commerce is feasible additionally to delivery Trades.

BE: It stands for Book Entry. Shares falling within the Trade-to-Trade or T-segment are listed during this series and no intraday is allowed. This implies trades will solely be settled by accepting or giving the delivery of shares.

BL: This series is for facilitating block deals. Block deal may be a trade, with a minimum amount of five lakh shares or minimum worth of Rs. 5 crores, implemented through one Block deal, on the special "Block Deal window". The window is opened for under thirty-five minutes within the morning from 9:15 to 9:50AM.

BT: This series provides associate exit route to penny investors having shares within the physical kind with a cap of most five hundred shares.

BZ: Stocks that are blacklisted for violation of exchange rules. This series stocks falls underneath Trade-to-Trade class and thus BTST (Buy nowadays Sell Tomorrow) and intraday isn't allowed in such stocks.

GC – This series permits Government Securities and Treasury Bills to be listed underneath this class.

IL – This series permits solely FIIs to trade among themselves. Permissible solely in those securities wherever most permissible limit for FIIs isn't broken.

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

The opening price is that the first trade worth that was recorded throughout the day's commerce. This figure is usually employed in relevance the present worth or the closing price from the previous commerce session in an endeavor to quantify the stock's movement.

Financial periodicals and websites usually embody a stock's "high" and "low" costs. The high is that the highest worth at that a stock is listed during a period. The low is that the lowest worth of the period. A stock's high and low points for the day are usually known as its intraday high and low.

High and low conjointly indicates the worth vary at that the stock has listed at throughout the day. In different words, these are the most and also the minimum costs that folks have procured the stock.

Typically, the previous closing is going to be consecutive session's opening price, however this can be not invariably the case. a pointy modification between the last listed worth and its open typically suggests that a stock is experiencing robust momentum, either positive or negative based on whether or not the present session's opening worth is higher or below the

previous session's close. It usually represents a stimulating commerce chance. The day’s high and low also are common knowledge points found at intervals between a stock quote. This knowledge is usually utilized by traders as a estimation of volatility.

The last price is that the one at that the foremost recent transaction happens, whereas the market value is actually the price the brokerage will estimate to satisfy our order as immediately as possible. If shopping for a stock, then the market value is the ask price at that moment.

The close is that the last commerce worth recorded once the market closed on the day. If the closing price is up or down over five-percent than the previous day's shut, the whole listing for that stock is listed in bold. it is to be noted that you are not almost certain to get this worth if you get the stock consecutive day the reason being that the share price is consistently dynamic (even when the exchange is closed for the day). The close is just associate indicator of past performance and except in extreme circumstances is a ballpark of what you ought to expect to pay.

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 a security has listed at throughout the day, based on both volume and worth.

The VWAP is employed as a benchmark to work out the standard of executions in massive orders. for instance, if a portfolio manager needs to accumulate thousands of shares, however conjointly needs to get the position below the typical worth for the day, the VWAP can typically be the price to beat.

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 of time by the average number of shares outstanding for same duration of time. The upper the share turnover, the more liquid the company shares will be.

The chapter 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 explains the Data Preparation section of our CRISP-DM framework. Within the data preparation section, we will clean and remodel our data before process and analysis.

Chapter 8: Data Preparation

Most machine learning algorithms need knowledge to be formatted with a very specific approach, therefore datasets typically need some quantity of preparation before they will yield helpful insights. Some datasets have value that area unit missing, invalid, or otherwise troublesome for an algorithmic rule to method. If info is missing, the algorithmic rule can't use it. If knowledge is invalid, the algorithmic rule produces less correct or perhaps deceptive outcomes. Some datasets area units are comparatively cleaner however they ought to be formed fittingly and lots of datasets area unit are simply lacking desirable business context thus necessitating the need for feature enrichment. sensible data preparation produces clean and well-curated info that results in more sensible and correct model outcomes.

Data preparation could also be one amongst the foremost troublesome steps in any machine learning project. The reason is that every dataset is totally different and extremely specific to the project. nonetheless, there are enough commonalities across predictive modeling that we can outline a loose sequence of steps and subtasks that are most likely to be performed.

This method provides a context during which we will take into account the data preparation needed for the project, necessitated both by the definition of the project performed before data preparation and Evaluation of machine learning algorithms performed later.

Handling Missing values:

Three of the features—Trades', 'Deliverable Volume','% Deliverable had quite one hundred periods missing values therefore we will drop those columns as they are having several missing values.

We are needed to refrain from implementing the mean, media, and mode imputation methodology, because those might render values that may introduce bias into our dataset. Second, the strategy solely looks at the variable itself and therefore might come up with values that don't seem to be actually representative of trends within the dataset.

Features Addition:

Additionally, in our dataset, we added computed variables that for sure would influence stock returns. These are moving averages for rolling periods of seven days,13 days,20 dyas,100 days and two hundred days. we tend to conjointly enclosed Exponential moving averages for a span of seven days,13 days,20 dyas,100 days and two hundred days. we think it's going to be useful in evaluating the securities market returns. We conjointly enclosed 1day previous lag values of volume in concert of our input feature.

We perceive that prediction has its uncertainty; however, we understand that these indicators have helped monetary economists within the past perceive the longer-term movement of the stock costs. previous analysis on 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 we tend to embrace in our analysis, that 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 normaliser are scikit-learn ways to preprocess info for machine learning. The methodology which is needed to be deployed depends on model kind and feature values.

Data Scaling is a data preprocessing step for numerical variables. several machine learning algorithms like 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 in order 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 among the approach of data scaling that is being used. Here, the minimum of feature is created up to zero and the most of feature up to one. MinMax Scaler shrinks the data inside the given range, sometimes of zero to one. It transforms data by scaling variables to a given range. It scales the worth's to a selected value range while not varying the form of the initial distribution.

The chapter intended on making ready the data to be future-ready for the Model Building processes. the successive chapter explains the Data Modelling section of our 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 our 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 our forecasting outputs from these 15 various models.

HDFC excel data is put in Tabular form in step 1.

Step 2: 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.

Step 4: The data for a 7-day moving average time series is being plotted.

Step 5: The data from a rolling 7-day moving average is included in the Data frame.

Step 6: 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. 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.

Step 7: The same step is performed for the moving averages of 13 days, 20 days, 100 days, and 200 days.

Step 8: Exponential Moving Average is used to recreate the five different models created using Simple Moving Average.

Step 9: 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 utilise 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 is deployed and their prediction accuracy is being compared with Simple Moving Average Models, Exponential Moving Average Models and ARIMA Models.

further ahead various Regression Models including both Machine Learning and Deep learning techniques are deployed and Metrics namely Mean Absolute error and Mean Absolute percentage errors are deployed to estimate the predictions quality on the close price of the HDFC share. These Regression Models are Ordinary Least Squares(OLS)-Linear Regression Model, Lasso Regression Model, Lasso regression Model Using Cross Validation, The k-Nearest Neighbours(KNN) Algorithm, Decision Tree Algorithm, GridSearchCV Algorithm with Hyperparameter Tuning, Random Forest Regression Model,XGBoost ML Model, Using Principal Component Analysis (PCA) with LSTM, Using Principal Component Analysis (PCA) with LSTM with Moving Average variables(Feature Engineering),Long Short-Term Memory(LSTM) Neural Network Model, Regression Model using AutoKeras.

The previous chapter focussed on employing various Modelling algorithms to predict the Target variable value and determine accuracy of the trend prediction as well. The next chapter speaks about the Data Evaluation phase of the CRISP-DM framework. The Data Evaluation phase is the results of our 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 Last price for HDFC stocks collected on day-to-day basis.

The other Feature variables used are VWAP, Volume, Turnover. The Target variable is taken as 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 dyas,100 days and 2 hundred days. we tend to collectively also derive different other feature Variables particularly Exponential moving averages for a span of seven days,13 days,20 dyas,100 days and 2 hundred days. we formulate and create 1day previous lag values of volume as well as part of our input feature variables.

Initially A rule-based model is being developed to try to do hypothesis testing to work out whether or not the chosen stock's worth is crossing any of the moving averages mentioned as on top. Then potency of the prediction based on Hypothesis Testing Rule is decided that is being employed as a Metric to work out the accuracy for predicting the upward Trend or Downward trend of the HDFC shares.
Then we will conjointly build few Classifications Based Models. Metrics being employed for classification Models would be accuracy score and confusion matrix which can facilitate in determining the accuracy for 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 operate calculates subset accuracy. Accuracy is solely the number of correct predictions divided by the overall number of examples.
A confusion matrix may be a technique for summarizing the performance of a classification formula. Classification accuracy alone is deceiving if you have got 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 right and what varieties of errors it is creating.

A confusion matrix may be 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 key to the confusion matrix. The confusion matrix shows the ways in which your 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 on 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 we create a model for prediction functions in statistic Time series analysis, we tend to need a stationary statistic for higher prediction. That the opening move to figure on modelling is to create a Time series stationary. Testing for stationarity may be an ofttimes used activity in autoregressive modelling. we will perform numerous tests just like the KPSS, Phillips–Perron, and Augmented Dickey-Fuller.

ADF (Augmented Dickey-Fuller) test is a statistical significance test which means the test will end up in hypothesis tests with null and alternative hypotheses. As a result, we will have a p-value from that we will have to be compelled to create inferences regarding the Time series, whether or not it is stationary or not.

To perform the ADF test in any statistic package, stats model provides the implementation operate 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 results of ADF test are bigger than 0.05 then we were required to fail to reject Null Hypothesis H0 and are available to a reasoning that point Series is not Stationary. If results of ADF test would are lesser than 0.05 then we were required to reject Null Hypothesis H0 and are available to a reasoning that point Series is Stationary.

In all results of ADF test for ARIMA Modelling on our dataset for HDFC stock, we will see that the p-value obtained were bigger than 0.05 thus we did not reject the null hypothesis and concluded that the statistic for Dataset under consideration is non-stationary.

For most Time series patterns, one or a pair of differencing is critical to create a stationary Time series. ADF test would facilitate to verify the order of differencing needed to create our statistic stationary before we tend to build ARIMA Models for our Time series info.

we would conjointly evaluate Autocorrelation plot and also the partial autocorrelation plot of our statistic information to work out Auto Regressive Moving Average (ARMA) models for Time series analysis. Understanding Autocorrelation operate (ACF), and Partial autocorrelation operate (PACF) plots of the series are necessary to work out the order of AR and/ or MA terms.

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

Next, we will build totally different Regression Models using each of 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 our Modelling Algorithms. The metrics that we want to verify for the accuracy of predictions in case of regression Modelling are Mean Absolute Error (MAE), Mean sq. Error (MSE), Root Mean sq. Error (RMSE), Median Absolute Error (MAE), Mean Absolute percentage Error (MAPE).

In the context of machine learning, absolute error refers to the magnitude of distinction between the prediction of observation and the true worth of that observation. MAE takes the mean of absolute errors for many predictions and observations as a measurement of the magnitude of errors for the complete unit.
The Mean square Error (MSE) is maybe the sole and commonest loss perform. To calculate the MSE, we are taking the distinction between model's predictions and so the ground truth, sq. it and average it out across the complete dataset.

Root Mean sq. Error (RMSE) is that the standard deviation of the residuals or the prediction errors. RMSE determines the spread of these residuals. In different words, it tells how focused the info is round the line of best fit

The Median absolute error is robust to outliers. The loss is calculated by taking the median of all absolute variations between the target and so the prediction.

The mean absolute proportion error (MAPE) is that the mean or average of absolutely the proportion errors of forecasts. Error is outlined as actual or discovered worth minus the forecasted worth. proportion errors are summed while not reference to sign to reckon MAPE.MSE is scale-dependent however MAPE isn't. therefore, MSE cannot be used for comparison accuracy across Time series with totally different scales. For business use, MAPE is commonly most popular.

R-squared is another applied statistic jargon that denotes the goodness of fit of a regression model. The perfect worth for r-square is one. The nearer the price of r-square to one, the better is that the model fitted. The worth of R-square may be negative once the models fitted are worse than the typical fitted model. R-square values are computed for each test and train data one by one.

Model performance is being evaluated supported on the above-discussed metrics for the various Models designed for our project.

Chapter 11: Deployment

Major Action Items Implemented:

The implementation for the capstone project can be accessed at the link below:

https://github.com/Embedded-org/ACCOMPLISHMENTS/tree/master/RACE%20CAPSTONE%20PROJECT1

SMA EMA T Test Metrics:

HDFC excel information is place 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 that the same as the actual price rise for successive day.

The hypothesis testing rule's proportion accuracy is repeatedly verified. 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, 20 days. Exponential Moving Average with seven days,13 days and twenty days span are employed to recreate the various models that was created using simple Moving Average 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 Simple Moving Average and Exponential Moving Averages variables for the T Test based on Hypothesis Testing

My Leader Board gives me the following results:

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

Figure 11.1– Leader Board-comparison of Metrics for Simple Moving Averages and Exponential Moving Averages variables as per T Test based on Hypothesis Testing

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 exponential moving averages with a span of 7-days. however, prediction efficiency is least for 20 days Simple moving average data and 20-days exponential moving average data.

SMA EMA Z Test Metrics:

HDFC excel information is place 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 that the same as the actual price rise for 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. Exponential Moving Average with 100 days and 200 days span are employed to recreate the various models that was created using simple Moving Average 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 11.2– Top five rows for HDFC Dataset including Simple Moving Average and Exponential Moving Averages variables for the Z Test based on Hypothesis Testing

My Leader Board gives me the following results:

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

Figure 11.2– Leader Board-comparison of Metrics for Simple Moving Averages and Exponential Moving Averages variables as per Z Test based on Hypothesis Testing

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 Exponential moving average with 100 days and 200 days span as well.

Classification Model Metrics:

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

I have done Feature Engineering and derived direction as Target Variable to predict the direction of the close price based on the Feature Variables as independent variables.

I have employed Auto Keras Classification Model (Structured Data Classifier), KNN Classification Model and Logistic Regression Classification Modelling techniques 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 11.3– Top five rows for HDFC Dataset including direction as Target Variable for Classification Modelling

My Leader Board gives me 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

Figure 11.3– Leader Board-comparison of Metrics for Accuracy Predictions on Close price of HDFC Share by different Classification Models

It can be observed that Logistic Regression Classification Model and Auto Keras classification Model have given 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 simple moving Average and Exponential Moving averages were near about 77%. Hence, we can safely conclude that Deep Learning Model and Machine Learning Models were able to provide better outputs compared to Statistical methods of Hypothesis Testing.

ARIMA Models Metrics:

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

My Leader Board gives me the following results:

SERIAL NUMBERS	DESCRIPTIONS	MEAN ABSOLUTE ERROR (MAE) FOR TEST DATA	MEAN SQUARE ERROR (MSE) FOR TEST DATA	ROOT MEAN SQUARE ERROR (MSE) FOR TEST DATA	Mean Absolute Percentage Error FOR TEST DATA	MAPE FOR TEST DATA
EMA_200ARIMA	Auto Arima model using Exponential moving average-200 samples	84.21	9662.99	98.30	96.06	Nan
SMA_100ARIMA	Auto Arima model using Simple moving average-100 samples	112.25	19404.28	139.30	95.51	9.42
SMA_20ARIMA	Auto Arima model using Simple moving average-20 samples	183.76	45227.79	212.67	181.82	16.29
SMA_13ARIMA	Auto Arima model using Simple moving average-13 samples	184.73	44482.52	210.91	172.64	16.171
SMA_7ARIMA	Auto Arima model using Simple moving average-7 samples	185.64	47486.11	217.91	173.93	15.09

Figure 11.4– Leader Board-comparison of Metrics for Accuracy Predictions on Close price of HDFC Share by ARIMA Models

In all results of ADF test for ARIMA Modelling on our dataset for HDFC stock, we had seen that the p-value obtained were bigger than 0.05 thus we did not reject the null hypothesis and concluded that the statistic for Dataset under consideration is non-stationary.

It can be observed that mean Absolute Error, Mean Square error, Root Mean Square Error, Median Absolute Error and Mean Absolute percentage Error are far too high in the case of all Auto ARIMA Modelling. Hence, we can conclude that the dataset under consideration was not suitable for Time series Modelling using ARIMA Modelling algorithm.

Regression Models-Part1 Metrics:

I have used Prev Close price, Open price, High price, Low price, Last, price, VWAP, and Volume_lag_1d as Feature Variables. I have used Close price as Target Variable to predict the values of the close price based on the Feature Variables as independent variables.

I have employed Ordinary Least Squares (OLS)-Linear Regression Model, Lasso Regression Model, Lasso regression Model Using Cross Validation and the k-Nearest Neighbours (KNN) Algorithm as the Modelling techniques to predict 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 11.4– Top five rows for HDFC Dataset including Close as Target Variable for Regression Modelling

My Leader Board gives me the following results:

SERIAL NUMBERS	DESCRIPTIONS	MEAN ABSOLUTE ERROR (MAE) FOR TEST DATA	MEAN SQUARE ERROR (MSE) FOR TEST DATA	ROOT MEAN SQUARE ERROR (MSE) FOR TEST DATA	Mean Absolute Percentage Error FOR TEST DATA	MAPE FOR TEST DATA
OLS Model	Ordinary Least Squares (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	The k-Nearest Neighbors (KNN) Algorithm	5.42	132.08	11.49	3.16	0.59

Figure 11.5– Leader Board-comparison of Metrics for Predicting Close price of HDFC Share by part1 Regression Models

It can be observed that mean Absolute Error and Mean Absolute percentage Error were satisfactory for Ordinary Least Squares (OLS)-Linear Regression Model. However, other Regression Models were not able to provide MAPE within acceptable range.

Regression Models-Part2 Metrics:

I have used Prev Close price, Open price, High price, Low price, Last, price, VWAP, and Volume_lag_1d as Feature Variables. I have used Close price as Target Variable to predict the values of the close price based on the Feature Variables as independent variables.

I have employed Decision Tree Algorithm, GridSearchCV Algorithm with Hyper-parameter Tuning, Random Forest Regression Model and XGBoost ML Model as the Modelling techniques to predict close price of the HDFC share.

My Leader Board gives me the following results:

SERIAL NUMBERS	DESCRIPTIONS	MEAN ABSOLUTE ERROR (MAE) FOR TEST DATA	MEAN SQUARE ERROR (MSE) FOR TEST DATA	ROOT MEAN SQUARE ERROR (MSE) FOR TEST DATA	Mean Absolute Percentage 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

Figure 11.6– Leader Board-comparison of Metrics for Predicting Close price of HDFC Share by part2 Regression Models

It can be observed that mean Absolute Error and Mean Absolute percentage Error 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:

I have used Prev Close price, Open price, High price, Low price, Last, price, VWAP, and Volume_lag_1d as Feature variables. I have used Close price as Target Variable to predict the values of the close price based on the Feature Variables as independent variables.

I have employed Using Principal Component Analysis (PCA) with LSTM, Using Principal Component Analysis (PCA) with LSTM with Moving Average variables (Feature Engineering), Long Short-Term Memory-LSTM Neural Network Model and Regression Model using AutoKeras as the Modelling techniques to predict close price of the HDFC share.

My Leader Board gives me the following results:

SERIAL NUMBERS	DESCRIPTIONS	MEAN ABSOLUTE ERROR (MAE) FOR TEST DATA	MEAN SQUARE ERROR (MSE) FOR TEST DATA	ROOT MEAN SQUARE ERROR (MSE) FOR TEST DATA	Mean Absolute Percentage Error FOR TEST DATA	MAPE FOR TEST DATA
PCA LSTM Model	Using Principal Component Analysis (PCA) with LSTM	4.37	34.70	5.89	3.60	33.44
PCA LSTM Moving Averages Model	Using Principal Component Analysis (PCA) with LSTM with Moving Average variables (Feature Engineering)	7.75	135.03	11.62	5.99	33.47
LSTM Model	Long Short-Term Memory-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

Figure 11.7– Leader Board-comparison of Metrics for Predicting Close price of HDFC Share by part2 Regression Models

It can be observed that mean Absolute Error and Mean Absolute percentage Error were satisfactory for both Using Principal Component Analysis (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	Simple moving average-7 samples	YES-77.67
SMA13	Simple moving average-13 samples	NO-65.66
SMA20	Simple moving average-20 samples	NO-60.88
EMA7	Exponential moving average-7 samples	YES-76.97
EMA13	Exponential moving average-13 samples	NO-65.89
EMA20	Exponential moving average-20 samples	NO-61.24
SMA100	Simple moving average-100 samples	NO-53.77
SMA200	Simple moving average-200 samples	NO-53.96
EMA100	Exponential moving average-100 samples	NO-54.36
EMA200	Exponential moving average-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

It can be observed that Logistic Regression Classification Model and Auto Keras classification Model have given 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 simple moving Average and Exponential Moving averages were 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	Ordinary Least Squares (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	The k-Nearest Neighbours (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
XGBOOST 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	Long Short-Term Memory-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

It can be observed that Ordinary Least Squares (OLS)-Linear Regression Model, Random Forest Regression Model, Using Principal Component Analysis (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 Mean absolute percentage error.

Chapter 13: Conclusions and Recommendations for future work

The hypothesis testing rule's percentage accuracy was repeatedly verified using five Simple Moving Averages Models. Exponential Moving Average were used to recreate the five other different models created using Simple Moving Average. 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 were used to create an additional five different models. The construction of all 15 models, were used to forecast day trading in the stock market.

Prediction accuracy were 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 utilise the same procedure to forecast buy or sell choices, which is helpful.

Recommendations for Future Work: we 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. We haven't 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, we can show how to define Bullish and Bearish regimes using modern machine learning techniques. We will then use the techniques already discussed in this project to estimate the direction of close price for each of the Normal and Crash regimes. We will also like to explore Sentiment Analysis Approach 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

Falinouss, P. (2007). *MASTER ' S THESIS Stock Trend Prediction Using News Articles*. *4*.

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

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

Wijaya, C. Y. (2021). *CRISP-DM Methodology For Your First Data Science Project*. Towards Data Science. https://towardsdatascience.com/crisp-dm-methodology-for-your-first-data-science-project-769f35e0346c

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.