Trading Analytics for Day Trading in Stock Market

by Anand Mohan

Submission date: 24-Aug-2022 05:34PM (UTC+0530)
Submission ID: 1886367726
File name: ing_Analytics_for_Day_Trading_in_Stock_Market_-_Anand_Mohan.docx (153.14K)
Word count: 7473
Character count: 40572

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 get the right stock and collect all relevant data to make correct forecasting. Build the right models by using multiple Modelling techniques and explore some of the state-of-the-art solutions to minimize the prediction errors.

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

Chapter 1: Introduction

Algorithmic Trading systems have changed the approach by which stock markets perform. 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. 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 (Shah et al., 2019).

The Stock market, as a result of its high volatility, is 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. 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).

The previous Chapter 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 the next chapter, 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. Analysing exchange movements and price momentum 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).

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: Valuation strategies are used as a very important strategy for selecting smart stocks at an occasional value or undervalued value with an honest margin of safety. The parameters that require consideration are 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: 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. The parameters that require consideration are 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. The parameters that need consideration are ROE & ROCE > fifteen, Free cash flow > zero, and 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. The parameters that need consideration are Sales, EBIT, Net Profit, and EPS.

Exit or Risk Parameters: Exit or Risk Parameters are determined by those parameters that build some stocks risky to take a position in. The parameters that need consideration are 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.

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. 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.

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. 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. 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. 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).

The previous chapter discusses all current techniques used to build better Forecasting or Trading Strategies. With all options discussed in the Literature review, still, the volatility of the market is a concern which is being discussed in the next chapter.

Chapter 3: Problem Statement

Investors are looking at algorithmic trading as an option to reduce volatility. Fundamental analysis is being used for evaluating a share's intrinsic value for long-term investment opportunities. Technical analysis on the other hand assists the traders to evaluate trends in the stock's price, momentum, and volume from a statistical perspective. However, the consistency of the prediction performance of most of these techniques remains debatable and the volatility of the market is still unpredictable. Therefore, it is the constant endeavor of investors to find better, easy, and simple Modelling techniques for forecasting any share’s price for day trading in the stock market. Such a process should also evaluate the degree of risks concerned and minimize the chances of loss with the highest possible accuracy.

Chapter 4: Objectives of the Study

Based on the problem statement mentioned in the previous chapter, the objectives of the project are as follows.

- Firstly, the objective of this project is to get the right stock and collect all relevant data to make correct forecasting. Understand the data pattern using Exploratory Data Analysis and Hypothesis testing and perform data preparation which enables the production of clean and well-curated info with extra Features addition that results in more sensible and correct model outcomes.

- Secondly, the objective of the project is to start with simple models whose iteration speed would be higher and can be understood easily namely linear regression and decision tree. Then move to something more complex by using multiple other Machine Learning and Deep Learning Techniques.

- Thirdly the objective of the project is to explore some state-of-the-art solutions to minimize prediction errors. For every forecasting Technique, there will be errors, and since the stock market has high volatility, hence the chances of errors are more. Therefore, some standard Error Metrics are being used in this project to measure the error of the forecasting models and quantitatively compare their performances.

Chapter 5: Project Methodology

The current Chapter 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.

Business understanding provides Fundamental and Technical analysis of HDFC stock to demonstrate why the HDFC stock dataset has been used for this project. Data understanding explains the different columns used in the HDFC dataset. Data preparation explains that Handling Missing values, Features Addition and Data Scaling using MinMax Scaler were the steps used for processing the dataset before being used for Modelling. Hypothesis testing, Classification Models, ARIMA Models, and different Regression Models were used in the Data Modelling phase. The data evaluation phase examines the results of different Modelling techniques which were used in the Data Modelling phase. Deployment speaks about developing a front-end API for the deployment Dashboard.

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. CRISP-DM itself is not a one-time method. Each method may be a new learning expertise, that new things are being learned throughout the method, and it may trigger alternative business queries (Cornellius Yudha Wijaya, 2021).

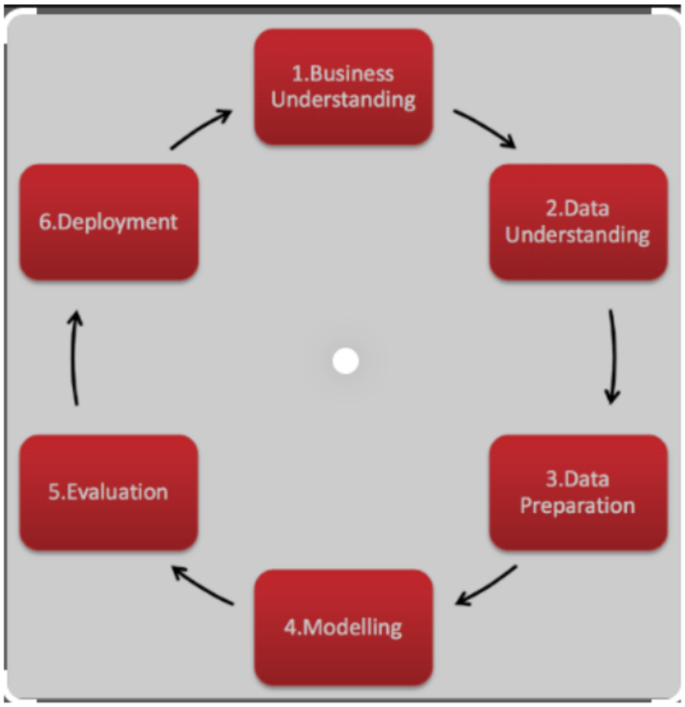


Figure 5.1 CRISP-DM Process Diagram

The previous Chapter explains the CRISP-DM framework. The framework comprises 6 different phases. 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

This chapter helps to determine whether HDFC Bank stock is the right stock which is the dataset under consideration for this capstone project. All relevant data is collected and inferences are made using Fundamental and Technical Analysis of HDFC stock.

Fundamental Analysis of HDFC stock:

The fifty-two weeks high for HDFC bank shares stands out at 1305 and its complete capitalization is INR 4.98 trillion making it a large cap company. HDFC Bank is India's largest personal sector investor in terms of assets. The corporate has been the fastest developing bank in India in the last decade.

It's conjointly the most important bank in India by capitalization as of March 2020 and operates across three verticals specifically Retail Banking, Wholesale Banking, and Treasury Services. The 3 vital factors resulting in this are often the market presence, CASA ratio (Current Account Savings Account), and low non-performing assets (NPAs). HDFC bank is spread across the country with 5000+ banking shops in 2748+ cities in India, it has a CASA ratio of roughly 45%, a low Gross of 1.36%, and net non-performing assets are evaluated as 0.39%.

Treasury services remain a profitable service in derivatives, foreign exchange, and debt securities markets. HDFC's Cards business is additionally one among the most important in India with 13.9 million credit cards. POS terminals are numbering more than 5lakhs.

Mr. Aditya Puri is that the director of HDFC Bank since its origination in 1994. he's known for utilizing technology to alter how banking happens in the country. The revenue has been increasing at a CAGR of about 20% in the last decade. Net profit margin has increased owing to the increasing scale and low NPAs. Overall, the corporate has managed to beat economic and interest rate cycles and win higher profits.

The company maintains a savings balance per account of INR 75,000+ together with the Floats from multiple transactional banking franchises. This helps the bank to keep up liquidity while increasing its reserves portfolio. The money deposit ratios have seen a small improvement and also the investment deposit ratio has seen a decline which implies lower deposits with RBI in government bonds and better loan advances.

Net profit as a share of total funds has conjointly seen improvement with stable loan and asset turnover ratios. this is often a healthy indicator of business potency. The increasing average price/ earnings ratio and the consistent average price to book value are also quite appreciable. Overall, the ROE has been decreasing thanks to the reduced leverage and not the profit which may be a smart indicator of the economic condition within the company (Aaron Patrick, 2020).

Technical Analysis of HDFC stock:

Relative Strength Index defines RSI. For 14 days, if RSI is in the range 25-45 it would mean that HDFC stock is trending downwards, RSI between 45-55 will mean that the HDFC stock indicates sideways movement. it will be trending upwards if RSI is in the range of 55-75. if RSI is below 25, HDFC stock is oversold and RSI more than 75 indicates HDFC stock is overbought. Presently RSI is 57.53 meaning that HDFC stock is moving in an upward trend.

MACD is defined as Moving Average Convergence Divergence. it is calculated by subtracting 26 days EMA from 12 days EMA. if the MACD is more than 0 and also greater than 9 days EMA, HDFC stock will be trending upwards. if the MACD is less than 0 and also lesser than 9 days EMA, HDFC stock will trend downwards. Currently, MACD is 34.09 indicating that HDFC stock is showing an upward trend.

For 20 days, the position of the close price for the High-low range will define the Stochastic indicator which determines the momentum in HDFC stock. Stochastic in the range 55-80 will indicate that the stock is trending upwards. Between 45 and 55, it will be in a sideways trend, and in the range 20-45, the HDFC stock will indicate trending downwards. Stochastic above 80 would mean that HDFC stock is overbought and less than 80 will tell that HDFC stock is oversold. Currently Stochastic is 86.62 which means that HDFC stock is overbought and hence the investor should wait for some time so that the Stochastic indicator gives a lesser value.

ADX is nothing but the Average Directional Index. We can decide how strongly HDFC stock is trending upwards or downwards using ADX. for 14 days, an increasing ADX will indicate

HDFC stock trending upwards or downwards very strongly. A decreasing ADX means that no strong trend will exist either upwards or downwards. Currently, HDFC stock ADX is 41.95 meaning it will show a very intense upward or downward trend.

Bollinger band is positive and negative standard deviations from SMA. For 20 days, if the close price of HDFC stock moves quite away from a positive standard deviation will mean that HDFC stock is overbought and if the close price of HDFC stock moves away from a negative standard deviation then the HDFC stock will be considered oversold. Currently, the upper band is 1514.69 and the lower band is 1,261.46. The close price of HDFC stock is 1493.05 which means HDFC stock is overbought (moneycontrol, n.d.)**.**

The previous Chapter performed the fundamental and technical analysis of HDFC stock. The next chapter 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 Trading Data of HDFC company from the year 2000 to 2021 is being used for this study. This study uses NSE Data. Following are the details for every column used in the HDFC dataset:

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 [9] 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 [10] 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.

The previous Chapter explains 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 next chapter 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

The HDFC data which is taken from NSE comes with a lot of limitations and that has to be processed which includes the following steps:

Handling Missing values: Three of the features’ trades, ‘Deliverable Volume’, and’% Deliverable had quite one hundred periods of 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: [8] 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 [8] to preprocess info for machine learning. The methodology which is needed to be deployed depends on the model kind and feature values.

[5] 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 [5] 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 [5] 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 [5] 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. The previous Chapter is intended on making ready the data to be future-ready for the Model Building processes. the next chapter 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 stock data is being used 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 SMA time series data is added in step 3. The data for a 7-day SMA time series is being plotted. The data from 7-day SMA 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 SMA.

If yesterday's closing price was below the 7-day SMA 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 SMA of 13 days, 20 days, 100 days, and 200 days. 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-[13]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 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 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

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 on top. prediction based on the Hypothesis Testing Rule is compared with the actual trend to evaluate the accuracy of predicting the upward Trend or Downward trend of the HDFC shares.

SMA EMA T Test Metrics:

The hypothesis testing rule’s accuracy is repeatedly verified. The T-test is employed to perform hypothesis testing because the sample size for testing is lesser than thirty samples. SMA of 7 days.13days, and 20 days and EMA with 7,13 days, and 20 days spans are employed to recreate the various models based on T-test Hypothesis Testing.

The Leader Board for T-test Hypothesis Testing gives the following results:

SERIAL NUMBERS	TOTAL	TRUE COUNT	FALSE COUNT	EFFICIENCY
SMA7	**5297**	**4114**	**1183**	**77.67**
SMA13	5291	3474	1817	65.66
SMA20	5284	3217	2067	60.88
EMA7	5297	4077	1220	76.97
EMA13	5291	3486	1805	65.89
EMA20	5284	3236	2048	61.24

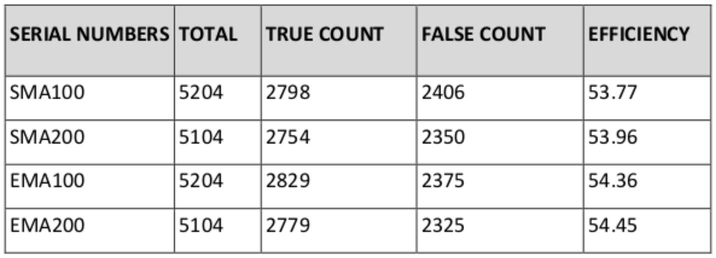
Table 10.1– Leader Board-comparison of Metrics for SMA and EMA variables as per T Test based on Hypothesis Testing

From Table 10.1, It can be observed that T-test Hypothesis testing done for 7-days SMA has given the highest efficiency in correctly predicting the upward or downward trend closely followed by 7-days EMA. However, prediction efficiency is the least for 20-day SMA and 20-days EMA.

SMA EMA Z Test Metrics:

The hypothesis testing rule's accuracy is repeatedly verified. Z-test is employed to perform hypothesis testing because the sample size for testing is more than 30 samples. SMA of 100,200 days and EMA with 100 days and 200 days spans are employed to recreate the various models.

The Leader Board for Z-test Hypothesis Testing gives the following results:



SERIAL NUMBERS	TOTAL	TRUE COUNT	FALSE COUNT	EFFICIENCY
SMA100	5204	2798	2406	53.77
SMA200	5104	2754	2350	53.96
EMA100	5204	2829	2375	54.36
EMA200	5104	2779	2325	54.45

Table 10.2– Leader Board-comparison of Metrics for SMA and EMA variables as per Z Test based on Hypothesis Testing

From Table 10.2, 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.

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 further in determining the accuracy of predicting the upward Trend or Downward trend of the HDFC shares.

Classification Model Metrics:

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.

The Leader Board for Classification Models gives the following results:

SERIAL NUMBERS	TOTAL	TRUE COUNT	FALSE COUNT	EFFICIENCY
Auto Keras Classification Model	**1061**	**901**	**160**	**84.92**
KNN Classification Model	1061	786	267	74.08
Logistic Regression Classification Model	**1061**	**956**	**97**	**90.10**

Table 10.3– Leader Board-comparison of Metrics for Accuracy Predictions on Close price of HDFC Share by different Classification Models

From Table 10.3, 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.

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. 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 and hence ARIMA Modelling is not suitable for HDFC stock.

ARIMA Models Metrics:

SMA of 7-days,13-days,20-days and100-days and EMA with 200 days span is used as Target Variables to recreate 5 different Auto Arima Models.

The Leader Board for ARIMA Models gives the following results:

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

Table 10.4– Leader Board-comparison of Metrics for Accuracy Predictions on Close price of HDFC Share by ARIMA Models

From Table 10.4, 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.

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. Model performance is being evaluated based on the above metrics for the various Regression Models designed for the project.

Regression Models Metrics:

The Leader Board Based on OLS, Lasso, CVLASSO, and KNN regression Models gives the following results:

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

Table 10.5– Leader Board-comparison of Metrics for Predicting Close price of HDFC Share by the First set of Regression Models

From Table 10.5, 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.

The Leader Board based on Decision Tree, GridSearchCV, Random Forest, and XGBoost Regression Models gives the following results:

DESCRIPTIONS	MAE FOR TEST DATA	MSE FOR TEST DATA	RMSE FOR TEST DATA	Median Absolute Error FOR TEST DATA	MAPE FOR TEST DATA
Decision Tree Algorithm	3.26	23.95	4.89	2.10	0.383
GridSearchCV Algorithm with Hyper-parameter Tuning	3.22	23.16	4.81	2.10	0.38
Random Forest Regression Model	**2.45**	**15.25**	**3.90**	**1.49**	**0.29**
XGBoost ML Model	3.25	22.78	4.77	2.12	0.37

Table 10.6– Leader Board-comparison of Metrics for Predicting Close price of HDFC Share by the Second set of Regression Models

From Table 10.6, 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.

The Leader Board based on PCA with LSTM, PCA with LSTM with Moving Average variables, LSTM Neural Network, and AutoKeras Regression Models gives the following results:

SERIAL NUMBERS	**MAE FOR TEST DATA**	**MSE FOR TEST DATA**	**RMSE FOR TEST DATA**	**Median Absolute Error FOR TEST DATA**	**MAPE FOR TEST DATA**
Using PCA with LSTM	4.37	34.70	5.89	3.60	33.44
Using PCA with LSTM with Moving Average variables	7.75	135.03	11.62	5.99	33.47
LSTM Neural Network Model	9.71	159.01	12.61	8.20	33.40
Regression Model using AutoKeras	**2.59**	**242.51**	**15.57**	**1.10**	**0.27**

Table 10.7– Leader Board-comparison of Metrics for Predicting Close price of HDFC Share by the Third set of Regression Models

From Table 10.7, 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.

The previous chapter discusses the accuracy of stock prediction using hypothesis testing and classification models. The Arima Models and various Regression Models predict the close value of HDFC stock and estimate using different error metrics. The Analysis and Results chapter will examine all the results derived from the various models and figure out the best model which has been most successful in minimizing the prediction errors.

Chapter 11: Deployment

In the Future, there is a deployment Dashboard proposed. The data pipeline shown below explains the deployment plan to be taken up where the business requirement would be to develop a front-end API as an executable application.

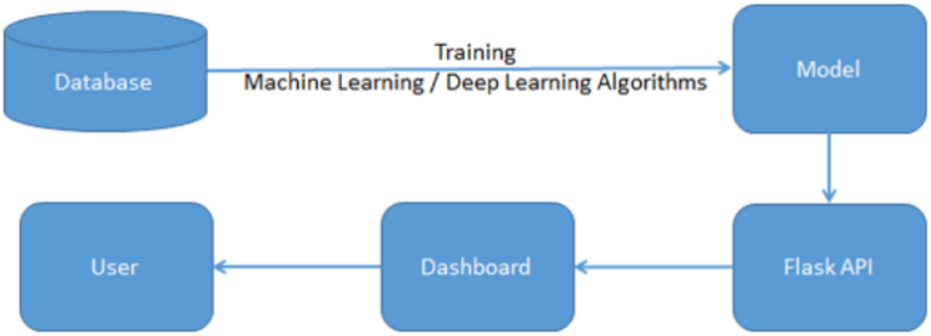


Figure 11.1 Deployment Proposal

As per the proposal for future assignments, the dashboard takes API as an input
Derived from themachine/deep learning algorithms for multi-label features with an end-to-end UI Interface.

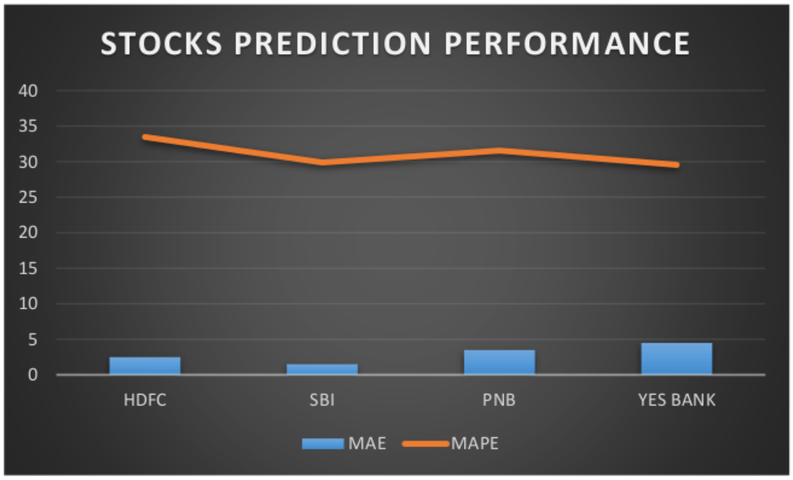


Figure 11.2 Illustration of

Dashboard

Chapter 12: Analysis and Results

All the models are now combined and below is the description for the final results.

Classification Metrics Comparison:

SERIAL NUMBERS	EFFICIENCY>67%
SMA-7 samples	**YES-77.67**
SMA-13 samples	NO-65.66
SMA-20 samples	NO-60.88
EMA-7 samples	YES-76.97
EMA-13 samples	NO-65.89
EMA-20 samples	NO-61.24
SMA-100 samples	NO-53.77
SMA-200 samples	NO-53.96
EMA-100 samples	NO-54.36
EMA-200 samples	NO-54.45
Auto Keras Classification Model	**yes-84.92**
KNN Classification Model	yes-74.08
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	**MAE<=5**	**MAPE<=0.33**
OLS-Linear Regression Model	**YES-2.034**	**YES-0.23**
Lasso Regression Model	NO-7.555	NO-0.85
Lasso regression Model Using Cross-Validation	NO-7.55	NO-0.85
KNN Algorithm	NO-5.423	NO-0.59
Decision Tree Algorithm	YES-3.26	NO-0.38
GridSearchCV Algorithm with Hyper-parameter Tuning	YES-3.218	NO-0.38
Random Forest Regression Model	**YES-2.45**	**YES-0.29**
XGBoost ML Model	YES-3.25	NO-0.37
Using Principal Component Analysis (PCA) with LSTM	**YES-4.366**	**YES-33.44**
Using Principal Component Analysis (PCA) with LSTM with Moving Average variables	NO-7.75	YES-33.47
LSTM Neural Network Model	NO-9.71	YES-33.40
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.

Chapter 13: Conclusions and Recommendations for future work

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

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.

Trading Analytics for Day Trading in Stock Market

ORIGINALITY REPORT

15%	12%	4%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	3%
2	billiondollarvaluation.com Internet Source	3%
3	ir.lib.uwo.ca Internet Source	2%
4	global.oup.com Internet Source	2%
5	Submitted to University of Leicester Student Paper	1%
6	Submitted to City University Student Paper	1%
7	www.researchgate.net Internet Source	1%
8	towardsdatascience.com Internet Source	<1%
9	www.investopedia.com Internet Source	<1%

10	mgt.sjp.ac.lk Internet Source	<1 %
11	Submitted to Korea National University of Transportation Student Paper	<1 %
12	www.newsbtc.com Internet Source	<1 %
13	Ning Dai, Yang Feng, Yuning Liu, Jian Li. "Analysis of the Influencing Factors of Users' Adoption Behavior in Social Q&A Community Based on Machine Learning Regression Algorithms", 2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), 2022 Publication	<1 %
14	Submitted to Georgia Institute of Technology Main Campus Student Paper	<1 %
15	qmro.qmul.ac.uk Internet Source	<1 %
16	investorplace.com Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		