

A Project Report on

**Directional Analytics for Day Trading in Stock Market**

Submitted in partial fulfilment for the award of the degree of

Master of Business Administration

In **Business Analytics**

Submitted by

**Anand Mohan**

**R19MBA53**

Under the Guidance of

Dr. JB Simha

Chief Mentor - RACE

REVA Academy for Corporate Excellence

**REVA University**

Rukmini Knowledge Park, Kattigenahalli,

Yelahanka, Bangalore – 560064

October 2022



# Candidate’s Declaration

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

Place: Bengaluru Name of the Student: Anand Mohan



Date: 10 October. 22 Signature of Student



# Certificate

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

Signature of the Guide Signature of the Director

Name of the Guide Name of the Director

Guide Director

External Viva

Names of the Examiners

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

Place: Bengaluru

Date:



# Acknowledgment

I am highly indebted to Dr. Shinu Abhi, Director, Corporate Training for the guidance and

support provided throughout the course and my project.

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

I am thankful to my classmates for their support, suggestions, and friendly advice during the

project work. I would like to acknowledge the support provided by the founder and Hon’ble Chancellor, Dr. P Shayma Raju, Vice-Chancellor, Dr. M. Dhanamjaya, and Registrar, Dr. N Ramesh.

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

It is my sincere gratitude towards my parents and my family for their kind co-operation.

Their encouragement also helped me in the completion of this project.

Place: Bengaluru

Date: 10 October. 22



# Similarity Index Report

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

Software Used: **Turnitin**

Date of Report Generation:

Similarity Index in %:

Total word count:

Name of the Guide: Dr. Jay Bharateesh Simha

Place: Bengaluru Name of the Student: **Anand Mohan**



Date: 10 October. 22 Signature of Student

Verified by:

Signature

Dr. Shinu Abhi,

Director, Corporate Training

# List of Abbreviations

|  |  |  |
| --- | --- | --- |
| **Sl. No** | **Abbreviation** | **Long Form** |
| 1 | **LR** | **Logistic Regression** |
| 2 | **DT** | **Decision Tree** |
| 3 | **RF** | **Random Forest** |
| 4 | KNN | k-Nearest Neighbours |
| 5 | **XG Boost** | Extreme Gradient Boosting |
| 6 | CRISP-DM | Cross-Industry Standard Process for Data Mining |
| 7 | VWAP | volume-weighted average price |
| 8 | NSE | National Stock Exchange |
| 9 | HDFC | Housing Development Finance Corporation Limited |
| 10 | SBI | State Bank of India |
| 11 | RSI | Relative Strength Index |
| 12 | MACD | Moving Average Convergence Divergence |
| 13 | ADX | Average Directional Index |

# List of Figures

|  |  |  |
| --- | --- | --- |
| **No.** | **Name** | **Page No.** |
| Figure 5.1 | CRISP-DM Process Diagram | 19 |
| Figure 11.1 | Deployment Proposal | 41 |
| Figure 11.2 | Illustration of Dashboard | 41 |

# List of Tables

|  |  |  |
| --- | --- | --- |
| **No.** | **Name** | **Page No.** |
| Table 6.1 | Fundamental Analysis of HDFC stock | 20 |
| Table 10.1 | Accuracy Predictions on Direction Detection by 6,10,14 days consecutive closing prices split week on week using RF Classifier Model | 30 |
| Table 10.2 | Accuracy Predictions on Direction Detection by 6,10,14 days consecutive closing prices split week on week using XG Boost Classifier Model | 31 |
| Table 10.3 | Go Long Direction Prediction with Technical Indicators as Feature Variables using LR Classifier Model | 32 |
| Table 10.4 | Go Long Direction Prediction with Technical Indicators as Feature Variables using DT Classifier Model | 33 |
| Table 10.5 | Go Long Direction Prediction with Technical Indicators as Feature Variables using RF Classifier Model | 34 |
| Table 10.6 | Go Long Direction Prediction with Technical Indicators as Feature Variables using KNN Classifier Model | 35 |
| Table 10.7 | Go Long Direction Prediction with Technical Indicators as Feature Variables using XG Boost Classifier Model | 36 |
| Table 10.8 | Go Short Direction Prediction with Technical Indicators as Feature Variables using LR Classifier Model | 37 |
| Table 10.9 | Go Short Direction Prediction with Technical Indicators as Feature Variables using DT Classifier Model | 38 |
| Table 10.10 | Go Short Direction Prediction with Technical Indicators as Feature Variables using RF Classifier Model | 39 |
| Table 10.11 | Go Short Direction Prediction with Technical Indicators as Feature Variables using KNN Classifier Model | 39 |
| Table 10.12 | Go Short Direction Prediction with Technical Indicators as Feature Variables using XG Boost Classifier Model | 40 |
| Table 12.1 | Leader Board-comparison of Metrics for Direction Detection by 6,10,14 days consecutive closing prices split week on week using RF Classifier Model | 42 |
| Table 12.2 | Leader Board-comparison of Metrics for Go Long Direction Prediction with Technical Indicators as features using Classification Models | 43 |
| Table 12.3 | Leader Board-comparison of Metrics for Go Short Direction Prediction with Technical Indicators as features using Classification Models | 44 |

# Abstract

Advance Machine learning techniques are getting remarkably popular in predicting stock market returns. Several research and development initiatives have been taken in able to predict stock market returns using historical data. During this capstone project, twenty-two years' price of the stock daily close price is being utilized and investigated for accuracy of the predictions of the direction of the close price for the stock under consideration.

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.

The 6-day consecutive closing price for the stock under consideration is being taken. These 6 days' consecutive closing prices will be tabulated week on week for the entire dataset and will be utilized as 6 different feature variables for building the classification Model. The difference between the 7th and 8th-day Closing price is determined. If the 8th-day closing price is seen an increase from the 7th day by 0.7% or more, the direction of the closing price can be made positive. If the 8th-day closing price is seen as a decrease from the 7th day by -0.7% or less, the direction of the closing price can be made negative. Between -0.7% and 0.7%, the direction of the closing price for the stock under consideration can be treated as sideways.

The rule is being set to determine what has to be seen as a direction change.0.7% difference,1% difference, and 1.5% difference -these are different classes of direction for which rule is being set which is to be followed for computing the direction change as either positive change, negative change, or no change. once it is determined to say for example 0.7% difference has the best prediction accuracy among all different classes of direction then a Similar process is again repeated for a range of consecutive days to be utilized as the feature variable increases to 10 days and 14 days. The prediction accuracy is determined to confirm that say 0.7% difference has the best prediction accuracy among all different classes of direction even when the range of consecutive days to be utilized as a feature variable is increased to 10 days and 14 days consecutively.

Similarly, all technical indicators can be utilized in Technical Analysis to build another set of classification Models. All different types of technical indicators namely momentum indicators, trend indicators, volatility indicators, and volume indicators can be utilized as feature variables based on the input dataset and different classification models can be built to determine their prediction accuracy. Various Classification models namely LR Classifier, DT Classifier, RF Classifier, KNN Classifier, and XG Boost Classifier is deployed and their prediction accuracy is being compared using Metrics namely Precision, recall,f1-score, accuracy score, and ROC AUC Score.

Direction Detection by 6,10,14 days consecutive closing prices split week on week is performed on the close price. RF modelling is done for percentage change in close price between upper-band +0.7% and lower band -.0.7% has given the highest efficiency in prediction among all Modelling techniques namely LR, DT, RF, KNN, and XG Boost Modelling. LR modelling is done for percentage change in close price between upper-band +0.5% and lower band -.0.5% has given the highest precision, recall, f1-score, and accuracy score for volume and momentum indicators whereas XG Boost Classifier provided the best prediction performance for trend and volatility indicators for predicting the upward trend of the close price.LR modelling is done for percentage change in close price between upper-band +0.5% and lower band -.0.5% has given the highest precision, recall, f1-score, and accuracy score for volume and trend indicators whereas XG Boost Classifier provided the best prediction performance for momentum indicators. Similarly, RF Classifier provided the best predictions for volatility indicators. These outcomes were useful in predicting the downward trend of the close price.

The invaluable takeaway from the capstone is that various classification modelling techniques had been remarkably useful in predicting the direction of the close price for the stock under consideration.

Keywords: Direction prediction, Stock Market, Technical Indicators, Classification Models, LR, DT, RF, KNN, XG Boost, PCA, HDFC

**Contents**

[Candidate’s Declaration 2](#_Toc47857456)

[Certificate 3](#_Toc47857457)

[List of Abbreviations 6](#_Toc47857458)

[List of Figures 6](#_Toc47857459)

[List of Tables 6](#_Toc47857460)

[Abstract 8](#_Toc47857461)

[Chapter 1: Introduction 11](#_Toc47857462)

[Chapter 2: Literature Review 13](#_Toc47857463)

[Chapter 3: Problem Statement 16](#_Toc47857464)

[Chapter 4: Objectives of Study 17](#_Toc47857465)

[Chapter 5: Project Methodology 18](#_Toc47857466)

[Chapter 6: Business Understanding 20](#_Toc47857467)

[Chapter 7: Data Understanding 23](#_Toc47857468)

[Chapter 8: Data Preparation 25](#_Toc47857469)

[Chapter 9: Data Modeling 27](#_Toc47857470)

[Chapter 10: Data Evaluation 30](#_Toc47857471)

[Chapter 11: Deployment 41](#_Toc47857472)

[Chapter 12: Analysis and Results 42](#_Toc47857473)

[Chapter 13: Conclusions and Recommendations for future work 45](#_Toc47857474)

[Bibliography 47](#_Toc47857475)

[Appendix 50](#_Toc47857476)

[Plagiarism Report 50](#_Toc47857477)

[Publications in a Journal/Conference Presented/White Paper 51](#_Toc47857478)

[Any Additional Details 51](#_Toc47857479)

# Chapter 1: Introduction

Whereas the use of Trading algorithms gives edges like decreased expenses, decreased latency, and bereft of emotions, it brings up challenging situations for retail traders because of the inaccessibility of required technologies to shape such systems. As more innovativeness has resulted in the generation of newer Trading algorithms, comparing the effectiveness and accuracy of these algorithms seems to be a daunting task. Trading algorithms might go fine occasionally on back testing in controlled environments; however, live validations are still becoming a grim prospect, because of several things like value variations, quiet news, and existing noise. Hence, a feasible solution could be to identify and implement more than a few popular stock evaluation strategies and enforce the best practices in 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 outsized inventory of stock prediction strategies has evolved over the years, though the consistency of the precise prediction and overall performance of a maximum of these strategies stays debatable. For the transaction of shares via a broker, there is mostly a fee paid to the broker for each buy and sale which will almost eat up the gains because the Trading frequency will increase, although brokers are discount brokers (Huang et al., 2021).

The requirement is to overcome the ambiguities of Fundamental and technical evaluation, and additionally, the glaring development in the modelling strategies has pushed several researchers to check new strategies for stock value forecasting. An alternative form of collective intelligence has emerged, and new innovative strategies square measure being used for stock price predictions. The mechanisms contain the work of machine learning algorithms for exchange shares analysis and forecast (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 outsized inventory of stock prediction strategies has evolved over the years and also informs additionally that the glaring development in the modelling strategies has pushed several researchers to check new strategies for stock value forecasting. 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

There are numerous parameters impacting value movements in varied sizes and layers in the stock market. consistent with economical market theory, the market corrects itself, which means that the present share value is averaging the past values that are neither too low nor too high (Rajkar et al., 2021).

Ultrafast algorithms improve traders’ ability to seize opportunities long before any human would be able to do the same. Nonetheless, machine learning expands the scope of information mining and processing and enables the capability to triumph over the markets with the help of existing patterns and correlations (Hansen, 2020).

Regulators have restrained algorithmic commerce, following accusations of market manipulation. Market manipulation has currently featured quite conspicuously in developed countries. one of the outcomes of this state of affairs is that algorithmic commerce has become relatively unsafe in hopes of constructing better profits (Mukerji et al., 2019).

Fundamental analysis helps to identify and implement short positions by selling the shares of companies showing downtrends and then covering these positions by buying back the shares of these companies when they start showing upward trends (Elbialy, 2019).

Corporations should use caution in increasing funding sources from debt, though this will be profitable, sometimes it may be enormously risky if the corporate fails to pay its obligations (Anjani & Syarif, 2019).

Investors try and realize additional info to assist in stock market commerce and contemplate that historical knowledge might offer indications of future value movements (Faijareon & Sornil, 2019).

Technical Analysis can demarcate and recognize commerce openings in the stock market by examining identifiable patterns similar to volume and price action movements (Thanekar & Shaikh, 2021).

Technical analysis involves the use of many technical indicators like MACD, OBV, Moving average, etc on the past costs (Kimbonguila et al., 2019).

The expectation of various cryptocurrencies like Bitcoin, Ethereum, Litecoin, and Ripple digital currency value in examination with the anticipated price by the volatility regression model and trend indicators gave pretty higher returns for the entire month (Dahham & Ibrahim, 2020).

Spikes in the implicit market volatility are an indicator for future increments in the stock market returns, which amplifies systemic risk and reduces the benefits of portfolio diversification as aftereffects (Magner et al., 2021).

Momentum-based Trading commerce is amongst proved investment strategies across major stock markets (Mohapatra & Misra, 2020).

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).

Machine learning and AI have been increasingly used for the analysis of image-based cellular screens. However, we suggest unsupervised exploratory data analysis should be performed as an initial step to gaining better data understanding (Omta et al., 2020).

The central plan of PCA is to spot correlations and patterns in a dataset with high dimensionality and scale back it to a considerably lower dimension without losing any important info (Dar, 2021).

LR is used instead of linear regression in situations where the target variable is not numeric, but a nominal or an ordinal variable (Al-Bairmani & Ismael, 2021).

In DT, the hidden rules along with the constraints are explored to map them with the nodes and branches of the tree. However, the model becomes more complex as the size of the datasets increases. This is being handled using more advanced algorithms in DT for classification and regression problems (Jena & Dehuri, 2020).

In DT, the tree originates from the root node whereas the decision nodes decide the rule for moving from one node to another. Leaf nodes are the output from DT (Hafeez et al., 2021).

RF is quite flexible to non-linearity in the dataset and is the most appropriate ensemble learning algorithm for medium-sized to very large-sized datasets (Schonlau & Zou, 2020).

KNN is the most popular statistical technique utilized in pattern identification over the last four decades (Wang, 2019).

XGBoost8 is highly flexible, scalable, and extremely robust. XGBoost is extensively recognized as an extremely useful ensemble learning algorithm. However, its performance needs more improvements ideally in scenarios where the dataset is imbalanced (Zhang et al., 2022).

The confusion matrix helps in exploring several classification metrics for comparing performances. Accuracy determines the precision of a model according to its correctly classified true positives and true negatives in the dataset. F1-score helps in determining false positives and false negatives. Additionally, the AUC compares the rates of false-positive and true-positive in the dataset (Silva & Naranjo, 2020).

The confusion matrix provides the fusion of predicted vs. actual values within a single matrix. It evaluates numerous performance metrics which include accuracy, precision, and recall (Markoulidakis 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 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, most of the fundamental, technical analyses and algorithmic trading failed to consistently predict with higher accuracy and the volatility of the market remained still unpredictable. Therefore, it is the constant endeavor of investors to find better, easy, and simple Modelling techniques for forecasting the direction of the 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 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 build the right models by using multiple Classification Modelling techniques namely LR Classifier, DT Classifier, RF Classifier, KNN Classifier, and XG Boost Classifier to determine the Modelling algorithm which would provide the best prediction accuracy.
* 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. LR Classifier, DT Classifier, RF Classifier, KNN Classifier, and XG Boost Classifier 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 one of the datasets under consideration for this capstone project. All relevant data is collected and inferences are made using Fundamental and Technical Analysis of HDFC stock. A similar analysis is made for SBI and KOTAK bank stock which are the other two datasets under consideration for this capstone project. The analysis made for SBI and KOTAK bank stock

Has been done in the implementation document which can be accessed in the GitHub link as provided in the Appendix section of the report.

## Fundamental Analysis of HDFC stock:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **PARTICULARS** | **JUN 2021** | **SEP 2021** | **DEC 2021** | **MAR 2022** | **JUN 2022** |
| **Quarterly Result (All Figures in Cr.)** | | | | | |
| Net Profit | 7,729.64 | 8,834.31 | 10,342.20 | 10,055.18 | 9,195.99 |
| **Promoters Details** | | | | | |
| Promoters | 25.89 | 25.83 | 25.80 | 25.78 | 25.73 |
| **Investors Details** | | | | | |
| Investors | 74.11 | 74.17 | 74.20 | 74.22 | 74.27 |
| **PARTICULARS** | **MAR 2018** | **MAR 2019** | **MAR 2020** | **MAR 2021** | **MAR 2022** |
| **Profit & Loss (All Figures in Cr. Adjusted EPS in Rs.)** | | | | | |
| Net Profit | 17,486.73 | 21,078.17 | 26,257.32 | 31,116.53 | 36,961.36 |
| Adjusted  EPS (Rs.) | 33.69 | 38.70 | 47.89 | 56.44 | 66.65 |
| **Balance Sheet (All Figures are in Crores.)** | | | | | |
| Total Liabilities | 10,63,934.32 | 12,44,540.69 | 15,30,511.26 | 17,46,870.52 | 20,68,535.05 |
| Total Assets | 10,63,934.32 | 12,44,540.69 | 15,30,511.26 | 17,46,870.52 | 20,68,535.05 |
| **Cashflow (All Figures are in Crores.)** | | | | | |
| Closing Cash | 123,062 | 81,818 | 87,940 | 121,273 | 155,386 |

Table 6.1– Fundamental Analysis of HDFC stock

HDFC Bank’s 52-week high is 1,725 and 52 weeks low is 1,271.60. It is located in India, Bahrain, Hong Kong, and Dubai. It has 6,378 branches,18,620 ATMs and 21,683 banking outlets. It was founded in 1994 and is headquartered in Mumbai, India.

## Technical Analysis of HDFC stock:

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 58.72 meaning that HDFC stock is moving in an upward trend.

MACD 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 18.97 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 89.62 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.

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 11.43 meaning it will show a weak upward or downward trend.

Bollinger's 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, KOTAK, and SBI 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, KOTAK, and SBI Bank from the year 2000 to 2022 is being used for this study. This study uses NSE Data. Following are the details for every column used in the HDFC, KOTAK, and SBI 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 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

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

The previous Chapter explains the HDFC stock-related feature variables that may be used as the independent variables. The direction of 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 direction of 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.

6,10,14 and 30 days consecutive closing prices are tabulated week on week for the entire dataset and utilized as different feature variables for building the classification Models.

For momentum indicators, Awesome Oscillator Indicator, KAMA Indicator, Percentage Price Oscillator, Percentage Volume Oscillator, ROC Indicator, RSI Indicator, Stochastic Oscillator, TSI Indicator, Ultimate Oscillator, WilliamsR Indicator are being utilized as the feature variables to predict the direction of the closing price and determine the prediction accuracy.

For trend indicators, ADX Indicator, Aroon Indicator, CCI Indicator, Ichimoku Indicator, KST Indicator, MACD, PSAR Indicator, EMA Indicator, WMA Indicator, Vortex Indicator are being utilized as the feature variables.

For volatility indicators, Average True Range, Bollinger Bands, Donchian Channel, Keltner Channel, and Ulcer Index are being used as feature variables. The lower and upper bands of these volatility indicators are also utilized as feature variables.

For volume indicators, AccDistIndex Indicator, ChaikinMoneyFlow Indicator, EaseOfMovement Indicator, ForceIndex Indicator, MFI Indicator, OnBalanceVolume Indicator, VolumePriceTrend Indicator, VolumeWeightedAveragePrice, NegativeVolumeIndex Indicator, DailyLogReturn Indicator are used as feature variables.

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

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

MinMax Scaler is one of the approaches to data scaling that is being used. Here, the minimum of features is created up to zero, and the most of features are up to one. MinMax Scaler shrinks the data inside the given range, sometimes from zero to one. It transforms data by scaling variables to a given range. It scales the worth to a selected value range while not varying the form of the initial distribution. 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

## Classification Modelling on close price:

The 6-day consecutive closing price for the stock under consideration is being taken. These 6 days' consecutive closing prices will be tabulated week on week for the entire dataset and will be utilized as 6 different feature variables for building the classification Model.

The difference between the 7th and 8th-day Closing price is determined. If the 8th-day closing price is seen as an increase from the 7th day by 0.7% or more, the direction of the closing price can be made positive.

If the 8th-day closing price is seen as a decrease from the 7th day by -0.7% or less, the direction of the closing price can be made negative. Between -0.7% and 0.7%, the direction of the closing price for the stock under consideration can be treated as sideways.

For data within the 0.7% and -0.7% band, usually, the advice to the investor will be to hold on to existing portfolios and wait for the direction of the closing price to show as either negative or positive change. If there is a negative change, usually the advice to the investor will be not to invest in such a circumstance. If there is a positive change the investor will be suggested to invest.

It is to be determined how many times the positive changes are identified by predicting and how many times positive changes are there in the actual data. This will be utilized to evaluate how many times true positives were detected and how many times the false positives were predicted in the prediction. A similar process is to be followed for detecting true negatives and false negatives. A similar process is to be followed for detecting true neutrals and false neutrals. Based on prediction accuracy, it can be suggested whether to invest or not to invest to the prospective investor.

Computation is being done to evaluate whether it is a positive change, negative change, or no change between the 7th and 8th-day closing price. The rule is being set to determine what has to be seen as a direction change.0.7% difference,1% difference, and 1.5% difference -these are different classes of direction for which rule is being set which is to be followed for computing the direction change as either positive change, negative change, or no change.

once it is determined say for example 0.7% difference has the best prediction accuracy among all different classes of direction namely 0.7% difference,1% difference, and 1.5% difference then the range of consecutive days to be utilized as feature variable is increased to 10 days. Therefore, a 10-day consecutive closing price for the stock under consideration is being taken. These 10 days' consecutive closing prices will be tabulated week on week for the entire dataset and will be utilized as different feature variables for building the classification Model.

A similar process is again repeated for a range of consecutive days to be utilized as the feature variable increases to 14 days. The prediction accuracy is determined to confirm that say 0.7% difference has the best prediction accuracy among all different classes of direction even when the range of consecutive days to be utilized as a feature variable is increased to 14 days.

## Classification Modelling on Technical Indicators:

Similarly, all technical indicators can be utilized in Technical Analysis to build another set of classification Models. All different types of technical indicators namely momentum indicators, trend indicators, volatility indicators, and volume indicators can be utilized as feature variables based on the input dataset and different classification models can be built to determine their prediction accuracy.

Generally, Open price, High price, low price, close price, and volume for the stock under consideration will be utilized to derive feature variables from technical indicators. These derived feature variables will then be used as the feature variables to predict the direction of the close price. The Actual direction of the close price is estimated as a percentage change of the close price between upper-band +0.5% and lower band -0.5% for all technical indicators-based classification Models. Eight different Classification models based on four different types of technical indicators are being built.

Various Classification models namely LR Classifier, DT Classifier, RF Classifier, KNN Classifier, and XG Boost Classifier are deployed and their prediction accuracy is being compared.

When the majority of the 20 various models or all of them move in the same direction, a choice on whether to invest or not to invest in the stock under consideration must be made. if for example say 10000 is invested in HDFC stock, and say it is predicted as a positive change for the next day. The same prediction process is repeated say 100 times and evaluated how much is the net gain and loss based on that.

The entire process is tried and tested for a different dataset altogether to ensure that Any stock on the stock market can utilize the same procedure to forecast whether to invest or not to invest, which is helpful. Daily Trading Data of SBI and Kotak Mahindra company from the year 2000 to 2022 is being used to repeat the entire process which had been implemented for the HDFC dataset.

The previous chapter focuses on employing various Modelling algorithms to determine the accuracy of the trend prediction. 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

The previous chapter discusses the accuracy of stock prediction using classification models. various Classification Models predict the direction of 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.

## Data Evaluation for HDFC Stock

Direction Detection by 6,10,14 days consecutive closing prices split week on the week:

**(0-Negative,1-Neutral,2-Positive)**

#### **RF Classifier**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Target Variable** | **precision** | **recall** | **f1-score** | **support** | **accuracy score** |
| percentage change between upper-band +0.7% and lower band -.07% (6 days consecutive closing prices split week on week) | | | | | |
| **0** | **0.91** | **0.81** | **0.86** | **544** | **0.87** |
| **1** | **0.85** | **0.90** | **0.88** | **580** |
| **2** | **0.85** | **0.89** | **0.87** | **547** |
| percentage change between upper-band +0.7% and lower band -.07%  (10 days consecutive closing prices split week on week) | | | | | |
| **0** | **0.87** | **0.86** | **0.87** | **559** | **0.87** |
| **1** | **0.87** | **0.87** | **0.87** | **550** |
| **2** | **0.87** | **0.88** | **0.87** | **561** |
| percentage change between upper-band +0.7% and lower band -.07%  (14 days consecutive closing prices split week on week) | | | | | |
| **0** | **0.80** | **0.77** | **0.79** | **536** | **0.80** |
| **1** | **0.79** | **0.81** | **0.80** | **543** |
| **2** | **0.80** | **0.81** | **0.80** | **590** |
| percentage change between upper-band +1.0% and lower band -.1.0% | | | | | |
| 0 | 0.90 | 0.09 | 0.16 | 425 | 0.53 |
| 1 | 0.50 | 0.97 | 0.66 | 759 |
| 2 | 0.63 | 0.22 | 0.32 | 487 |
| percentage change between upper-band +1.5% and lower band -.1.5% | | | | | |
| 0 | 1.00 | 0.02 | 0.03 | 234 | 0.70 |
| 1 | 0.70 | 1.00 | 0.82 | 1103 |
| 2 | 0.90 | 0.04 | 0.07 | 253 |

Table 10.1– Accuracy Predictions on Direction Detection by 6,10,14 days consecutive closing prices split week on week using RF Classifier Model

From Table 10.1, it can be observed that RF modelling done for percentage change in close price between upper-band +0.7% and lower band -.0.7% has given considerable efficiency in prediction.

#### **XG Boost Classifier**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Target Variable** | **precision** | **recall** | **f1-score** | **support** | **accuracy score** |
| percentage change between upper-band +0.7% and lower band -.07% | | | | | |
| 0 | 0.39 | 0.16 | 0.23 | 384 | 0.40 |
| 1 | 0.43 | 0.61 | 0.51 | 386 |
| 2 | 0.35 | 0.42 | 0.38 | 344 |
| percentage change between upper-band +1.0% and lower band -.1.0% | | | | | |
| 0 | 0.40 | 0.08 | 0.13 | 313 | 0.48 |
| 1 | 0.49 | 0.90 | 0.64 | 521 |
| 2 | 0.37 | 0.14 | 0.20 | 280 |
| percentage change between upper-band +1.5% and lower band -.1.5% | | | | | |
| 0 | 0.42 | 0.05 | 0.08 | 213 | 0.64 |
| 1 | 0.65 | 0.98 | 0.78 | 704 |
| 2 | 0.56 | 0.08 | 0.13 | 197 |

Table 10.2– Accuracy Predictions on Direction Detection by 6,10,14 days consecutive closing prices split week on week using XG Boost Classifier Model

From Table 10.2, it can be observed that LR modelling done for percentage change in close price between upper-band +1.5% and lower band -.1.5% has given the highest efficiency in prediction. However, it predicts only a neutral direction with 0.65 precision but its precision for predicting upward or downward trend should have been still better. Hence, **XG Boost** Modelling results can be considered but with caution.

### Go Long Direction Prediction using Technical Indicators

**(0-Non positive,1-Positive):** The direction of the close price is estimated as the percentage change of the close price between upper-band +0.5% and lower band -0.5%-if the percentage change of the closing price is more than 0.5%, the direction of the closing price is treated as positive and suitable for long Trading in the stock market. Otherwise, the direction of the close price is treated as non-positive and not suitable for long Trading in the stock market.

#### **LR Classifier**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Target Variable** | **precision** | **recall** | **f1-score** | **support** | **accuracy score** | **Roc AUC score** |
| Volume Indicators as Feature Variables | | | | | |  |
| **0** | **0.90** | **0.99** | **0.94** | **658** | **0.92** | **0.91** |
| **1** | **0.98** | **0.83** | **0.90** | **452** |
| Momentum Indicators as Feature Variables | | | | | |  |
| **0** | **0.79** | **0.84** | **0.81** | **685** | **0.76** | **0.74** |
| **1** | **0.71** | **0.63** | **0.67** | **423** |
| Trend Indicators as Feature Variables | | | | | |  |
| 0 | 0.78 | 0.92 | 0.85 | 679 | 0.80 | 0.76 |
| 1 | 0.83 | 0.59 | 0.69 | 431 |
| volatility Indicators as Feature Variables | | | | | |  |
| 0 | 0.73 | 0.98 | 0.84 | 658 | 0.77 | 0.73 |
| 1 | 0.93 | 0.47 | 0.63 | 452 |

Table 10.3– Go Long Direction Prediction with Technical Indicators as Feature Variables using LR Classifier Model

From Table 10.3, it can be observed that LR modelling done for percentage change in close price between upper-band +0.5% and lower band -.0.5% has given considerably good accuracy score for all technical categories of indicators namely Volume, momentum, trend, and volatility. Precision and f1-score are also satisfactory. Recall can be improved further for trend indicators. ROC AUC score has been considered satisfactory for all technical indicators.

#### **DT Classifier**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Target Variable** | **precision** | **recall** | **f1-score** | **support** | **accuracy score** | **Roc AUC score** |
| Volume Indicators as Feature Variables | | | | | |  |
| 0 | 0.75 | 0.87 | 0.81 | 658 | 0.75 | 0.73 |
| 1 | 0.75 | 0.59 | 0.66 | 452 |
| Momentum Indicators as Feature Variables | | | | | |  |
| 0 | 0.75 | 0.82 | 0.78 | 685 | 0.72 | 0.69 |
| 1 | 0.66 | 0.55 | 0.60 | 423 |
| Trend Indicators as Feature Variables | | | | | |  |
| 0 | 0.72 | 0.75 | 0.73 | 679 | 0.66 | 0.64 |
| 1 | 0.57 | 0.53 | 0.55 | 431 |
| volatility Indicators as Feature Variables | | | | | |  |
| 0 | 0.70 | 0.80 | 0.75 | 658 | 0.68 | 0.65 |
| 1 | 0.63 | 0.51 | 0.56 | 452 |

Table 10.4– Go Long Direction Prediction with Technical Indicators as Feature Variables using DT Classifier Model

From Table 10.4, it can be observed that DT modelling done for percentage change in close price between upper-band +0.5% and lower band -.0.5% has given considerably good accuracy score Volume indicators. Recall and accuracy can be improved further for trend and volatility indicators. ROC AUC score has been more than 50% for all technical indicators.

#### **RF Classifier**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Target Variable** | **precision** | **recall** | **f1-score** | **support** | **accuracy score** | **Roc AUC score** |
| Volume Indicators as Feature Variables | | | | | |  |
| 0 | 0.82 | 0.96 | 0.89 | 658 | 0.85 | 0.83 |
| 1 | 0.93 | 0.69 | 0.79 | 452 |
| Momentum Indicators as Feature Variables | | | | | |  |
| 0 | 0.75 | 0.90 | 0.82 | 685 | 0.75 | 0.70 |
| 1 | 0.76 | 0.51 | 0.61 | 423 |
| Trend Indicators as Feature Variables | | | | | |  |
| 0 | 0.77 | 0.95 | 0.85 | 679 | 0.80 | 0.75 |
| 1 | 0.87 | 0.56 | 0.68 | 431 |
| volatility Indicators as Feature Variables | | | | | |  |
| 0 | 0.75 | 0.97 | 0.84 | 658 | 0.79 | 0.75 |
| 1 | 0.92 | 0.53 | 0.67 | 452 |

Table 10.5– Go Long Direction Prediction with Technical Indicators as Feature Variables using RF Classifier Model

From Table 10.5, it can be observed that RF modelling done for percentage change in close price between upper-band +0.5% and lower band -.0.5% has given considerably good accuracy score for all technical indicators. Recall and accuracy can be improved further for all especially for predicting upward direction trends. ROC AUC score has been considered satisfactory for all technical indicators.

#### **KNN Classifier**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Target Variable** | **precision** | **recall** | **f1-score** | **support** | **accuracy score** | **Roc AUC score** |
| Volume Indicators as Feature Variables | | | | | |  |
| 0 | 0.61 | 0.89 | 0.72 | 658 | 0.60 | 0.83 |
| 1 | 0.51 | 0.17 | 0.26 | 452 |
| Momentum Indicators as Feature Variables | | | | | |  |
| 0 | 0.68 | 0.87 | 0.76 | 685 | 0.67 | 0.70 |
| 1 | 0.62 | 0.34 | 0.43 | 423 |
| Trend Indicators as Feature Variables | | | | | |  |
| 0 | 0.62 | 0.87 | 0.73 | 679 | 0.60 | 0.75 |
| 1 | 0.45 | 0.16 | 0.24 | 431 |
| volatility Indicators as Feature Variables | | | | | |  |
| 0 | 0.60 | 0.88 | 0.71 | 658 | 0.59 | 0.75 |
| 1 | 0.47 | 0.16 | 0.24 | 452 |

Table 10.6– Go Long Direction Prediction with Technical Indicators as Feature Variables using KNN Classifier Model

From Table 10.6, it can be observed that KNN Classifier modelling done for percentage change in close price between upper-band +0.5% and lower band -.0.5% can be improved further for accuracy score for all technical indicators. ROC AUC score has been considered satisfactory for all technical indicators.

#### **XG Boost Classifier**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Target Variable** | **precision** | **recall** | **f1-score** | **support** | **accuracy score** | **Roc AUC score** |
| Volume Indicators as Feature Variables | | | | | |  |
| 0 | 0.84 | 0.95 | 0.89 | 658 | 0.86 | 0.83 |
| 1 | 0.90 | 0.73 | 0.81 | 452 |
| Momentum Indicators as Feature Variables | | | | | |  |
| 0 | 0.78 | 0.83 | 0.80 | 685 | 0.75 | 0.70 |
| 1 | 0.70 | 0.61 | 0.65 | 423 |
| Trend Indicators as Feature Variables | | | | | |  |
| **0** | **0.81** | **0.92** | **0.86** | **679** | **0.82** | **0.75** |
| **1** | **0.85** | **0.65** | **0.74** | **431** |
| volatility Indicators as Feature Variables | | | | | |  |
| **0** | **0.81** | **0.91** | **0.86** | **658** | **0.82** | **0.75** |
| **1** | **0.84** | **0.69** | **0.76** | **452** |

Table 10.7– Go Long Direction Prediction with Technical Indicators as Feature Variables using XG Boost Classifier Model

From Table 10.7, it can be observed that XG Boost modelling done for percentage change in close price between upper-band +0.5% and lower band -.0.5% has given considerably good accuracy score for all technical categories of indicators namely Volume, momentum, trend, and volatility. Precision and f1-score are also satisfactory. Recall can be improved further for trend indicators. ROC AUC score has been considered satisfactory for all technical indicators.

### Go Short Direction Prediction using Technical Indicators

**(0-Negative,1-non-Negative):** The direction of the close price is estimated as the percentage change of the close price between upper-band +0.5% and lower band -0.5%-if the percentage change of the closing price is less than -0.5%, the direction of the closing price is treated as Negative and suitable for Short Trading in the stock market. Otherwise, the direction of the close price is treated as non-negative and not suitable for Short Trading in the stock market.

#### **LR Classifier**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Target Variable** | **precision** | **recall** | **f1-score** | **support** | **accuracy score** | **Roc AUC score** |
| Volume Indicators as Feature Variables | | | | | |  |
| **0** | **0.97** | **0.83** | **0.90** | **399** | **0.93** | **0.91** |
| **1** | **0.91** | **0.99** | **0.95** | **711** |
| Momentum Indicators as Feature Variables | | | | | |  |
| 0 | 0.70 | 0.59 | 0.64 | 394 | 0.76 | 0.73 |
| 1 | 0.79 | 0.86 | 0.82 | 714 |
| Trend Indicators as Feature Variables | | | | | |  |
| **0** | **0.91** | **0.56** | **0.69** | **414** | **0.81** | **0.76** |
| **1** | **0.79** | **0.97** | **0.87** | **696** |
| volatility Indicators as Feature Variables | | | | | |  |
| 0 | 0.89 | 0.44 | 0.59 | 399 | 0.78 | 0.70 |
| 1 | 0.75 | 0.97 | 0.85 | 711 |

Table 10.8– Go Short Direction Prediction with Technical Indicators as Feature Variables using LR Classifier Model

From Table 10.8, it can be observed that LR modelling done for percentage change in close price between upper-band +0.5% and lower band -.0.5% has given considerably good accuracy score for all technical categories of indicators namely Volume, momentum, trend, and volatility. Precision and f1-score are also satisfactory. Recall can be improved further for trend indicators. ROC AUC score has been considered satisfactory for all technical indicators.

#### **DT Classifier**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Target Variable** | **precision** | **recall** | **f1-score** | **support** | **accuracy score** | **Roc AUC score** |
| Volume Indicators as Feature Variables | | | | | |  |
| 0 | 0.66 | 0.67 | 0.67 | 399 | 0.76 | 0.74 |
| 1 | 0.81 | 0.81 | 0.81 | 711 |
| Momentum Indicators as Feature Variables | | | | | |  |
| 0 | 0.58 | 0.55 | 0.56 | 394 | 0.70 | 0.66 |
| 1 | 0.76 | 0.78 | 0.77 | 714 |
| Trend Indicators as Feature Variables | | | | | |  |
| 0 | 0.55 | 0.44 | 0.49 | 414 | 0.66 | 0.61 |
| 1 | 0.70 | 0.79 | 0.74 | 696 |
| volatility Indicators as Feature Variables | | | | | |  |
| 0 | 0.56 | 0.43 | 0.49 | 399 | 0.67 | 0.62 |
| 1 | 0.72 | 0.81 | 0.76 | 711 |

Table 10.9– Go Short Direction Prediction with Technical Indicators as Feature Variables using DT Classifier Model

From Table 10.9, it can be observed that DT modelling done for percentage change in close price between upper-band +0.5% and lower band -.0.5% has given considerably good accuracy score for volume and momentum indicators. Precision for predicting downward trends can be further improved. ROC AUC score has been more than 50% for all technical indicators.

#### **RF Classifier**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Target Variable** | **precision** | **recall** | **f1-score** | **support** | **accuracy score** | **Roc AUC score** |
| Volume Indicators as Feature Variables | | | | | |  |
| 0 | 0.87 | 0.71 | 0.78 | 399 | 0.85 | 0.82 |
| 1 | 0.85 | 0.94 | 0.89 | 711 |
| Momentum Indicators as Feature Variables | | | | | |  |
| 0 | 0.72 | 0.50 | 0.59 | 394 | 0.75 | 0.70 |
| 1 | 0.76 | 0.89 | 0.82 | 714 |
| Trend Indicators as Feature Variables | | | | | |  |
| 0 | 0.87 | 0.46 | 0.60 | 414 | 0.77 | 0.71 |
| 1 | 0.75 | 0.96 | 0.84 | 696 |
| volatility Indicators as Feature Variables | | | | | |  |
| **0** | **0.88** | **0.55** | **0.68** | **399** | **0.81** | **0.76** |
| **1** | **0.79** | **0.96** | **0.87** | **711** |

Table 10.10– Go Short Direction Prediction with Technical Indicators as Feature Variables using RF Classifier Model

From Table 10.10, it can be observed that RF modelling done for percentage change in close price between upper-band +0.5% and lower band -.0.5% has given considerably good accuracy score for all technical indicators. Recall and accuracy can be improved further for all especially for recalling the downward direction trend. ROC AUC score has been considered satisfactory for all technical indicators.

#### **KNN Classifier**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Target Variable** | **precision** | **recall** | **f1-score** | **support** | **accuracy score** | **Roc AUC score** |
| Volume Indicators as Feature Variables | | | | | |  |
| 0 | 0.45 | 0.42 | 0.44 | 399 | 0.61 | 0.82 |
| 1 | 0.69 | 0.71 | 0.70 | 711 |
| Momentum Indicators as Feature Variables | | | | | |  |
| 0 | 0.53 | 0.54 | 0.53 | 394 | 0.66 | 0.70 |
| 1 | 0.74 | 0.73 | 0.74 | 714 |
| Trend Indicators as Feature Variables | | | | | |  |
| 0 | 0.44 | 0.37 | 0.40 | 414 | 0.59 | 0.71 |
| 1 | 0.66 | 0.72 | 0.69 | 696 |
| volatility Indicators as Feature Variables | | | | | |  |
| 0 | 0.44 | 0.43 | 0.43 | 399 | 0.60 | 0.76 |
| 1 | 0.68 | 0.69 | 0.69 | 711 |

Table 10.11– Go Short Direction Prediction with Technical Indicators as Feature Variables using KNN Classifier Model

From Table 10.11, it can be observed that KNN modelling done for percentage change in close price between upper-band +0.5% and lower band -.0.5% can be improved further for accuracy score for all technical indicators. ROC AUC score has been considered satisfactory for all technical indicators.

#### **XG Boost Classifier**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Target Variable** | **precision** | **recall** | **f1-score** | **support** | **accuracy score** | **Roc AUC score** |
| Volume Indicators as Feature Variables | | | | | |  |
| 0 | 0.86 | 0.79 | 0.82 | 399 | 0.88 | 0.82 |
| 1 | 0.89 | 0.93 | 0.91 | 711 |
| Momentum Indicators as Feature Variables | | | | | |  |
| **0** | **0.72** | **0.59** | **0.64** | **394** | **0.77** | **0.70** |
| **1** | **0.79** | **0.87** | **0.83** | **714** |
| Trend Indicators as Feature Variables | | | | | |  |
| 0 | 0.84 | 0.60 | 0.70 | 414 | 0.81 | 0.71 |
| 1 | 0.80 | 0.93 | 0.86 | 696 |
| volatility Indicators as Feature Variables | | | | | |  |
| 0 | 0.79 | 0.64 | 0.71 | 399 | 0.81 | 0.76 |
| 1 | 0.82 | 0.91 | 0.86 | 711 |

Table 10.12– Go Short Direction Prediction with Technical Indicators as Feature Variables using XG Boost Classifier Model

From Table 10.12, it can be observed that XG Boost modelling done for percentage change in close price between upper-band +0.5% and lower band -.0.5% has given considerably good accuracy score for all technical categories of indicators namely Volume, momentum, trend, and volatility. Precision and f1-score are also satisfactory. Recall can be improved further for recalling downward trend direction. ROC AUC score has been considered satisfactory for all technical indicators.

# 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 machine 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.

Analysis for HDFC Stock is given below.

## **Direction Detection by 6,10,14 days consecutive closing prices split week on the week**:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Target Variable** | **precision** | **recall** | **f1-score** | **support** | **accuracy score** |
| percentage change between upper-band +0.7% and lower band -.07% (6 days consecutive closing prices split week on week) | | | | | |
| **0** | **0.91** | **0.81** | **0.86** | **544** | **0.87** |
| **1** | **0.85** | **0.90** | **0.88** | **580** |
| **2** | **0.85** | **0.89** | **0.87** | **547** |
| percentage change between upper-band +0.7% and lower band -.07%(10 days consecutive closing prices split week on week) | | | | | |
| **0** | **0.87** | **0.86** | **0.87** | **559** | **0.87** |
| **1** | **0.87** | **0.87** | **0.87** | **550** |
| **2** | **0.87** | **0.88** | **0.87** | **561** |
| percentage change between upper-band +0.7% and lower band -.07%(14 days consecutive closing prices split week on week) | | | | | |
| **0** | **0.80** | **0.77** | **0.79** | **536** | **0.80** |
| **1** | **0.79** | **0.81** | **0.80** | **543** |
| **2** | **0.80** | **0.81** | **0.80** | **590** |

Table 12.1– Leader Board-comparison of Metrics for Direction Detection by 6,10,14 days consecutive closing prices split week on week using RF Classifier Model

From Table 12.1, it can be observed that RF modelling done for percentage change in close price between upper-band +0.7% and lower band -.0.7% has given the highest efficiency in prediction among all Modelling techniques namely LR, DT, RF, KNN, and XG Boost Modelling. It predicts upward, neutral, and downward trend directions with reasonably good precision. F1-score combining the precision and recall of a classifier into a single metric is also reasonably good. This has been tested and proven with 6,10- and 14-day consecutive closing prices split week on week as 6,10 and 14 feature variables. Hence, RF Modelling provides a reasonably good modelling technique to be able to provide optimal prediction performance.

### Go Long Direction Prediction using Technical Indicators

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Target Variable** | **precision** | **recall** | **f1-score** | **support** | **accuracy score** | **Roc AUC score** |
| Volume Indicators as Feature Variables for LR Classifier | | | | | |  |
| **0** | **0.90** | **0.99** | **0.94** | **658** | **0.92** | **0.91** |
| **1** | **0.98** | **0.83** | **0.90** | **452** |
| Momentum Indicators as Feature Variables for  LR Classifier | | | | | |  |
| **0** | **0.79** | **0.84** | **0.81** | **685** | **0.76** | **0.74** |
| **1** | **0.71** | **0.63** | **0.67** | **423** |
| Trend Indicators as Feature Variables for  XG Boost Classifier | | | | | |  |
| **0** | **0.81** | **0.92** | **0.86** | **679** | **0.82** | **0.75** |
| **1** | **0.85** | **0.65** | **0.74** | **431** |
| volatility Indicators as Feature Variables for  XG Boost Classifier | | | | | |  |
| **0** | **0.81** | **0.91** | **0.86** | **658** | **0.82** | **0.75** |
| **1** | **0.84** | **0.69** | **0.76** | **452** |

Table 12.2– Leader Board-comparison of Metrics for Go Long Direction Prediction with Technical Indicators as features using Classification Models

From Table 12.2, it can be observed that LR modelling done for percentage change in close price between upper-band +0.5% and lower band -.0.5% has given the highest precision, recall, f1-score, and accuracy score for volume and momentum indicators whereas XG Boost Classifier provided best prediction performance for trend and volatility indicators.

### Go Short Direction Prediction using Technical Indicators

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Target Variable** | **precision** | **recall** | **f1-score** | **support** | **accuracy score** | **Roc AUC score** |
| Volume Indicators as Feature Variables for  LR Classifier | | | | | |  |
| **0** | **0.97** | **0.83** | **0.90** | **399** | **0.93** | **0.91** |
| **1** | **0.91** | **0.99** | **0.95** | **711** |
| Momentum Indicators as Feature Variables for  XG Boost Classifier | | | | | |  |
| **0** | **0.72** | **0.59** | **0.64** | **394** | **0.77** | **0.70** |
| **1** | **0.79** | **0.87** | **0.83** | **714** |
| Trend Indicators as Feature Variables for  LR Classifier | | | | | |  |
| **0** | **0.91** | **0.56** | **0.69** | **414** | **0.81** | **0.76** |
| **1** | **0.79** | **0.97** | **0.87** | **696** |
| volatility Indicators as Feature Variables for  RF Classifier | | | | | |  |
| **0** | **0.88** | **0.55** | **0.68** | **399** | **0.81** | **0.76** |
| **1** | **0.79** | **0.96** | **0.87** | **711** |

Table 12.3– Leader Board-comparison of Metrics for Go Short Direction Prediction with Technical Indicators as features using Classification Models

From Table 12.3, it can be observed that LR modelling done for percentage change in close price between upper-band +0.5% and lower band -.0.5% has given the highest precision, recall, f1-score, and accuracy score for volume and trend indicators whereas XG Boost Classifier provided best prediction performance for momentum indicators. Similarly, RF Classifier provided the best predictions for volatility indicators.

# Chapter 13: Conclusions and Recommendations for future work

The 6-day consecutive closing price for the stock under consideration is being taken. These 6 days' consecutive closing prices will be tabulated week on week for the entire dataset and will be utilized as 6 different feature variables for building the classification Model. The difference between the 7th and 8th-day Closing price is determined. The rule is being set to determine what has to be seen as a direction change.0.7% difference,1% difference, and 1.5% difference -these are different classes of direction for which rule is being set which is to be followed for computing the direction change as either positive change, negative change, or no change. Various Classification models namely LR Classifier, DT Classifier, RF Classifier, KNN Classifier, and XG Boost Classifier are deployed and their prediction accuracy is being compared using Metrics namely Precision, recall,f1-score, and accuracy score. Once it is determined say for example 0.7% difference has the best prediction accuracy among all different classes of direction then the Similar process is again repeated for a range of consecutive days to be utilized as the feature variable increased to 10 days and 14 days using the Classifier Modelling algorithm which provided the best directional prediction.

Similarly, all technical indicators are utilized in Technical Analysis to build another set of classification Models. All different types of technical indicators namely momentum indicators, trend indicators, volatility indicators, and volume indicators can be utilized as feature variables based on the input dataset and different classification models can be built to determine their prediction accuracy. Various Classification models namely LR Classifier, DT Classifier, RF Classifier, KNN Classifier, and XG Boost Classifier is deployed and their prediction accuracy is being compared using Metrics namely Precision, recall,f1-score, accuracy score, and ROC AUC Score.

The construction of all 20 models, was used to predict the direction of the close price for the stock under consideration. 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 solely focuses on predicting the direction of the close price of the HDFC stock using classification algorithms Techniques. A later similar process is applied for predicting the direction of the close price of other stocks in the banking sector namely SBI and KOTAK stocks. In the Future, there is a deployment Dashboard proposed. As per the proposal for future assignments, the dashboard takes API as an input Derived from the machine learning algorithms and can be utilized in predicting the direction of the close price for any stock in the Banking sector. Any stock on the stock market can utilize the same procedure to forecast buy or sell choices, which is helpful.

Recommendations for Future Work: it is assumed that returns are more or less constant over time. However, the assumption that the returns are constant over time is restrictive, and not true. Returns are highly dependent on time. This project has not discussed how to address one major drawback of stock prediction, namely that over different periods the stock returns can change drastically to either extremely low returns during stock market crashes or extremely high returns during stock market booming periods. In future projects, it can be shown how to define Bullish and Bearish regimes using modern machine learning techniques. The techniques already discussed in this project will then be used to estimate the direction of close price for each of the Normal and Crash regimes. The Sentiment Analysis Approach may also need to be explored using Text Analytics for predicting stock market returns.

# Bibliography

Al-Bairmani, Z. A. A., & Ismael, A. A. (2021). Using Logistic Regression Model to Study the Most Important Factors Which Affects Diabetes for the Elderly in the City of Hilla / 2019. *Journal of Physics: Conference Series*, *1818*(1). https://doi.org/10.1088/1742-6596/1818/1/012016

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

Anjani, T., & Syarif, A. D. (2019). The Effect of Fundamental Analysis on Stock Returns using Data Panels ; Evidence Pharmaceutical Companies listed on IDX. *International Journal of Innovate Science and Research Technology*, *4*(7), 500–505.

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

Dahham, A. Z. D., & Ibrahim, A. A. (2020). Effects of Volatility and Trend Indicator for Improving Price Prediction of Cryptocurrency. *IOP Conference Series: Materials Science and Engineering*, *928*(3). https://doi.org/10.1088/1757-899X/928/3/032043

Dar, A. N. (2021). PRINCIPAL COMPONENT ANALYSIS (PCA) (Using Eigen Decomposition). *Gsj*, *9*(7), 240–252. www.globalscientificjournal.com

Elbialy, B. A. (2019). The Effect of Using Technical and Fundamental Analysis on the Effectiveness of Investment Decisions of Traders on the Egyptian Stock Exchange. *International Journal of Applied Engineering Research*, *14*(24), 4492–4501. http://www.ripublication.com

Faijareon, C., & Sornil, O. (2019). Evolving and combining technical indicators to generate trading strategies. *Journal of Physics: Conference Series*, *1195*(1). https://doi.org/10.1088/1742-6596/1195/1/012010

Hafeez, M. A., Rashid, M., Tariq, H., Abideen, Z. U., Alotaibi, S. S., & Sinky, M. H. (2021). Performance improvement of decision tree: A robust classifier using tabu search algorithm. *Applied Sciences (Switzerland)*, *11*(15). https://doi.org/10.3390/app11156728

Hansen, K. B. (2020). The virtue of simplicity: On machine learning models in algorithmic trading. *Big Data and Society*, *7*(1). https://doi.org/10.1177/2053951720926558

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*. https://doi.org/10.1109/SSCI50451.2021.9660134

Jena, M., & Dehuri, S. (2020). Decision tree for classification and regression: A state-of-the art review. *Informatica (Slovenia)*, *44*(4), 405–420. https://doi.org/10.31449/INF.V44I4.3023

Kimbonguila, A., Matos, L., Petit, J., Scher, J., & Nzikou, J.-M. (2019). Effect of Physical Treatment on the Physicochemical, Rheological and Functional Properties of Yam Meal of the Cultivar “Ngumvu” From Dioscorea Alata L. of Congo. *International Journal of Recent Scientific Research*, *10*, 30693–30695. https://doi.org/10.24327/IJRSR

Magner, N., Lavin, J. F., Valle, M., & Hardy, N. (2021). The predictive power of stock market’s expectations volatility: A financial synchronization phenomenon. *PLoS ONE*, *16*(5 May), 1–21. https://doi.org/10.1371/journal.pone.0250846

Markoulidakis, I., Kopsiaftis, G., Rallis, I., & Georgoulas, I. (2021). Multi-Class Confusion Matrix Reduction method and its application on Net Promoter Score classification problem. *ACM International Conference Proceeding Series*, 412–419. https://doi.org/10.1145/3453892.3461323

Mohapatra, S., & Misra, A. K. (2020). Momentum returns: A portfolio-based empirical study to establish evidence, factors and profitability in Indian stock market. *IIMB Management Review*, *32*(1), 75–84. https://doi.org/10.1016/j.iimb.2019.07.007

moneycontrol. (n.d.). *HDFC Bank Ltd.TECHNICALS*. https://www.moneycontrol.com/technical-analysis/hdfcbank/HDF01/weekly

Mukerji, P., Chung, C., Walsh, T., & Xiong, B. (2019). The Impact of Algorithmic Trading in a Simulated Asset Market. *Journal of Risk and Financial Management*, *12*(2), 68. https://doi.org/10.3390/jrfm12020068

Omta, W. A., van Heesbeen, R. G., Shen, I., de Nobel, J., Robers, D., van der Velden, L. M., Medema, R. H., Siebes, A. P. J. M., Feelders, A. J., Brinkkemper, S., Klumperman, J. S., Spruit, M. R., Brinkhuis, M. J. S., & Egan, D. A. (2020). Combining Supervised and Unsupervised Machine Learning Methods for Phenotypic Functional Genomics Screening. *SLAS Discovery*, *25*(6), 655–664. https://doi.org/10.1177/2472555220919345

Rajkar, A., Kumaria, A., Raut, A., & Kulkarni, N. (2021). Stock Market Price Prediction and Analysis. *International Journal of Engineering Research & Technology*, *10*(06), 115–119.

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

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *Stata Journal*, *20*(1), 3–29. https://doi.org/10.1177/1536867X20909688

Shah, D., Isah, H., & Zulkernine, F. (2019). Stock market analysis: A review and taxonomy of prediction techniques. *International Journal of Financial Studies*, *7*(2). https://doi.org/10.3390/ijfs7020026

Silva, I., & Naranjo, J. E. (2020). A systematic methodology to evaluate prediction models for driving style classification. *Sensors (Switzerland)*, *20*(6), 1–21. https://doi.org/10.3390/s20061692

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

Thanekar, G. S., & Shaikh, Z. S. (2021). Analysis and Evaluation of Technical Indicators for Prediction of Stock Market. *International Journal of Engineering Research & Technology (IJERT)*, *10*(May), 341–344.

Wang, L. (2019). Research and Implementation of Machine Learning Classifier Based on KNN. *IOP Conference Series: Materials Science and Engineering*, *677*(5), 0–5. https://doi.org/10.1088/1757-899X/677/5/052038

Zhang, P., Jia, Y., & Shang, Y. (2022). Research and application of XGBoost in imbalanced data. *International Journal of Distributed Sensor Networks*, *18*(6). https://doi.org/10.1177/15501329221106935

**Appendix**

## Plagiarism Report[[1]](#footnote-1)

## Publications in a Journal/Conference Presented/White Paper[[2]](#footnote-2)

The publication of this work has been planned after the future work of deployment of state-of-the-art API with the dashboard.

## Any Additional Details

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

<https://github.com/Embedded-org/ACCOMPLISHMENTS/tree/master/RACE_CAPSTONE_PROJECT2>

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

<https://github.com/Embedded-org/ACCOMPLISHMENTS/blob/master/RACE_CAPSTONE_PROJECT2/Capstone2_implementation.docx>

1. Turnitin report to be attached from the University. [↑](#footnote-ref-1)
2. URL of the white paper/Paper published in a Journal/Paper presented in a Conference/Certificates to be provided. [↑](#footnote-ref-2)