

A Project Report on

**Modelling direction detection in selected stocks in**

**Indian BFSI sector**

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

February 2023



# 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 **Modelling direction detection in selected stocks in Indian BFSI sector** 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 **2023**.

Place: Bangalore Name of the Student: Anand Mohan



Date: 25 February 23 Signature of Student



# Certificate

This is to Certify that the Project work entitled **Modelling direction detection in selected stocks in Indian BFSI sector** 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 **2023**. 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: Bangalore

Date: 25 February 23



# Acknowledgment

I am highly indebted to Dr. Shinu Abhi, Director, Corporate Training for her 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 execute 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.

I sincerely thank 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: Bangalore

Date: 25 February 23



# Similarity Index Report

This is to certify that this project report titled **Modelling direction detection in selected stocks in the Indian BFSI sector** was scanned for similarity detection. The 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: Bangalore Name of the Student: **Anand Mohan**



Date: 25 February 23 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 | PCA | Principal Component Analysis |
| 6 | **XG Boost** | Extreme Gradient Boosting |
| 7 | CRISP-DM | Cross-Industry Standard Process for Data Mining |
| 8 | VWAP | volume-weighted average price |
| 9 | NSE | National Stock Exchange |
| 10 | HDFC | Housing Development Finance Corporation Limited |
| 11 | SBI | State Bank of India |
| 12 | RSI | Relative Strength Index |
| 13 | MACD | Moving Average Convergence Divergence |
| 14 | ADX | Average Directional Index |

# List of Figures

|  |  |  |
| --- | --- | --- |
| **No.** | **Name** | **Page No.** |
| Figure 5.1 | CRISP-DM Process Diagram | 18 |
| Figure7.1 | class distribution For HDFC, KOTAK, and SBI stock | 23 |
| Figure7.2 | Close values of HDFCBANK, KOTAK BANK, and SBIBANK stock from 2000 to 2022 | 23 |
| Figure7.3 | Distribution Plot for the HDFCBANK Stock | 23 |
| Figure7.4 | Distribution Plot for the KOTAKBANK Stock | 24 |
| Figure7.5 | Distribution Plot for the SBIBANK Stock | 24 |
| Figure7.6 | Customized Box plot for the HDFCBANK, KOTAK BANK, and SBIBANK stock from 2000 to 2022 | 24 |
| Figure7.7 | Customized Scatter Plot against close price for the HDFCBANK Stock from 2000 to 2022 | 25 |
| Figure7.8 | Customized Scatter Plot against close price for the KOTAKBANK Stock from 2000 to 2022 | 25 |
| Figure7.9 | Customized Scatter Plot against close price for the SBIBANK Stock from 2000 to 2022 | 26 |
| Figure11.1 | confusion matrix For KOTAKBANK Stock using  Volume Indicators as Feature variables | 36 |
| Figure11.2 | confusion matrix For KOTAKBANK Stock using  Momentum Indicators as Feature variables | 37 |
| Figure11.3 | confusion matrix For HDFCBANK Stock using  Trend Indicators as Feature variables | 38 |
| Figure11.4 | confusion matrix For HDFCBANK Stock using  Volatility Indicators as Feature variables | 38 |
| Figure12.1 | Deployment Proposal | 40 |
| Figure 12.2 | Illustration of Dashboard | 40 |

# List of Tables

|  |  |  |
| --- | --- | --- |
| **No.** | **Name** | **Page No.** |
| Table 6.1 | Fundamental Analysis of HDFC, KOTAK, and SBI stock | 19 |
| Table 9.1 | Modelling strategies and Model Evaluation Rule | 29 |
| Table 10.1 | Model Evaluation using LR Classifier for Go Long Direction Prediction | 32 |
| Table 10.2 | Model Evaluation using RF Classifier for Go Long Direction Prediction | 33 |
| Table 10.3 | Model Evaluation using XG Boost Classifier for Go Long Direction Prediction | 34 |
| Table 11.1 | Leader Board-comparison of Metrics for Direction Detection and Go Long Direction Prediction using the best classifier model | 35 |

# Abstract

Advanced 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's 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 understand the data pattern using Exploratory Data Analysis and perform data preparation before building the models. Build the right models by using multiple Modelling techniques and explore state-of-the-art solutions to minimize errors in direction prediction. Therefore, given the historical data, it should be correctly predicted whether the price will move up or move down.

The 6-day consecutive closing price for the stock under consideration is being taken. These 6 days' consecutive closing prices will be getting 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.0.5% difference,1% difference, and 1.5% difference are different classes of direction for which the rule is being set which is to be followed for computing the direction change as either positive change, negative change, or no change. 0.7% difference as a class of direction can also be used in place of 0.5% change if that gives a better directional indicator. 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 increased 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 are being 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 are being utilized as feature variables based on the input dataset and different classification models are built to determine their prediction accuracy. Various Classification models namely LR Classifier, DT Classifier, RF Classifier, KNN Classifier, and XG Boost Classifier are deployed and their prediction accuracy is 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 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 done for percentage change in close price as 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.

The invaluable take away 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 detection, Stock Market, Technical Indicators, Classification Models, LR, DT, RF, KNN, XGBoost, PCA, HDFC, KOTAK, SBI,

# Table of Contents

[Candidate’s Declaration 2](#_Toc47857456)

[Certificate 3](#_Toc47857457)

[List of Abbreviations 6](#_Toc47857458)

[List of Figures 6](#_Toc47857459)

[List of Tables 7](#_Toc47857460)

[Abstract 8](#_Toc47857461)

[Chapter 1: Introduction 11](#_Toc47857462)

[Chapter 2: Literature Review 12](#_Toc47857463)

[Chapter 3: Problem Statement 15](#_Toc47857464)

[Chapter 4: Objectives of Study 16](#_Toc47857465)

[Chapter 5: Project Methodology 17](#_Toc47857466)

[Chapter 6: Business Understanding 19](#_Toc47857467)

[Chapter 7: Data Understanding 22](#_Toc47857468)

[Chapter 8: Data Preparation 27](#_Toc47857469)

[Chapter 9: Modeling 29](#_Toc47857470)

[Chapter 10: Model Evaluation 32](#_Toc47857471)

[Chapter 11: Analysis and Results 35](#_Toc47857473)

[Chapter 12: Deployment 40](#_Toc47857472)

[Chapter 13: Conclusions and Recommendations for future work 41](#_Toc47857474)

[Bibliography 43](#_Toc47857475)

[Appendix 46](#_Toc47857476)

[Plagiarism Report 46](#_Toc47857477)

[Paper Publications in a Journal/Conference Presented/White Paper 47](#_Toc47857478)

[Certificate for the Conference Presentation 47](#_Toc47857479)

[GithubLink 47](#_Toc47857479)

# Chapter 1: Introduction

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 (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).

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. New innovative strategies are being used for stock price predictions (Rouf et al., 2021).

Daily Trading NSE Data of HDFC, KOTAK, and SBI Bank from the year 2000 to 2022 is being used for this capstone project which would broadly come under BFSI.BFSI comprises of Banking, Financial Services, and Insurance sector. Also, the BFSI industry includes financial service firms such as Broking and Asset Management. The BFSI industry is growing year on year at a 27% rate.

The previous chapter discusses the importance of Machine-Learning associated techniques that are developed for investments in the stock market. The chapter 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 direction detection in Stock Market.

# Chapter 2: Literature Review

There are numerous parameters impacting value movements in varied sizes and layers in the stock market (Rajkar et al., 2021).

Therefore, Different analysis namely technical and fundamental analysis is being done to invest in stock markets. 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).

Increasing funding sources from debt, though profitable, 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).

Fundamental analysis helps to identify stock quality and therefore, stock technical analysis done later performs better on the strong fundamental stock. Technical Analysis can identify 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 crypto currencies 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 of future increments in the stock market returns, which amplifies systemic risk and reduces the benefits of portfolio diversification as after-effects (Magner et al., 2021).

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

Taking the discussion further, Algorithmic trading is a systematic method of trading without subjective assessment through a manual trader using computer programs. Fast algorithms improve traders’ ability to seize opportunities long before any human would be able to do the same. Nonetheless, it 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).

However, Regulators have restrained algorithmic commerce, following accusations of market manipulation(Mukerji et al., 2019).

Machine learning and AI have been increasingly used in the field of business analytics. However, we suggest unsupervised exploratory data analysis should be performed as an initial step to gaining a better data understanding(Omta et al., 2020).

Machine learning can again be further categorized into supervised and unsupervised learning.

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(Alhomadi, 2021).

PCA as an unsupervised machine learning technique is being further researched thoroughly. 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).

Further, various supervised classification machine learning techniques used in this project namely LR, DT, RF, KNN, and XGBoost are being examined.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 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).

Various classification algorithms as discussed have to be built for the data. Subsequently, all these algorithms have to be tested. Confusion matrix for classification models is a step in that direction. 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

There are plenty of Regression algorithms that can be utilized to detect the closing price of any stock. However, risks are more predominant in predicting the exact closing price using both Linear and Non-linear Regression algorithms. When we have lesser data to build a regression model, under fitting scenarios may destroy the accuracy of our machine learning model especially when we are trying to build a linear model with a complex dataset. At certain times, while trying to cater to all kinds of both existent and nonexistent possibilities in data points, over fitting scenarios in regression models may again destroy the accuracy of test data while the accuracy of the trained data may work perfectly fine. The situation requires not completely relying only on regression algorithms to quantitatively predict the exact closing price of any stock. Investors can find plenty of algorithms that detect the exact closing price of any stock but will not tell the direction of the closing price. Therefore, we should try alternate approaches as well which enables us to decide objectively whether say the price of any stock will move up or move down or remain neutral.

# 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 accuracy in direction prediction.
* Thirdly the objective of the project is to explore state-of-the-art solutions to minimize errors in direction prediction. For every forecasting Technique, there will be errors, and since the stock market has high volatility, hence the chances of errors are more. Therefore, given the historical data, it should be correctly predicted whether the price will move up or move down utilizing precision, recall, and accuracy Metrics used in classification modelling techniques.

# 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 stocks to demonstrate why a particular stock dataset has been used for this project. In Data understanding the different feature variables used for the project are being studied and their Univariate analysis is performed. 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). CRISP-DM itself is not a one-time method. Each method may be a new learning expertise, new things are being learned throughout the method, and it may trigger alternative business queries(Cornellius Yudha Wijaya, 2021).



Fig 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.

Fundamental Analysis of HDFC, KOTAK, and SBI stock:

Table 6.1– Fundamental Analysis of HDFC, KOTAK, and SBI stock

|  |  |  |  |
| --- | --- | --- | --- |
| **PARTICULARS** | **HDFC** | **KOTAK** | **SBI** |
| Promoters | 25.73% | 25.97% | 57.57% |
| Investors | 74.27% | 74.03% | 42.43% |
| Adjusted EPS(Rs.) | 66.65 | 42.99 | 35.49 |
| All figures are in crores for the year 2022. | | | |
| Net Profit  (Profit and Loss) | 36,961.36 | 8572.69 | 31,675.98 |
| Total Liabilities  (Balance Sheet) | 20,68,535.05 | 4,29,428.40 | 49,87,597.41 |
| Total Assets  (Balance Sheet) | 20,68,535.05 | 4,29,428.40 | 49,87,597.41 |
| Closing Cash  (cash flow) | 155,386 | 52,665 | 398,905 |

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.

KOTAK Bank’s 52-week high is 2,253 and the 52-week low is 1,631. It is located in India, London, New York, California, Dubai, Abu Dhabi, Mauritius, and Singapore. It has 1,702 branches, and 2,761 ATMs. It was founded in 1985 and is headquartered in Mumbai, India.

State Bank of India’s 52-week high is 578.50 and its 52-week low is 425. It is located in India, Australia, Bangladesh, Belgium, Bhutan, Canada, China, Germany, and Hong Kong. It has 22,266 branches and 65,030 ATMs. It was founded in 1806 and is headquartered in Mumbai, India.

**Technical Analysis of** HDFC, KOTAK, and SBI stock:

For 14 days, if RSI is in the range of 25-45 it would mean that HDFC stock is trending downwards, RSI in the range of 45-55 will mean that the stock indicates sideways movement. It will be trending upwards if RSI is in the range of 55-75. If RSI is below 25, the stock is oversold and an RSI of more than 75 indicates the stock is overbought. Presently RSI for HDFC stock is 58.72meaning that HDFC stock is moving in an upward trend. Presently RSI for KOTAK stock is 60.33 meaning that KOTAK stock is moving in an upward trend. Presently RSI for SBI stock is 69.86meaning that SBI 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, the stock will be trending upwards. If the MACD is less than 0 and also lesser than 9 days EMA, the stock will trend downwards. Currently, MACD for HDFC stock is 18.97indicating that HDFC stock is showing an upward trend. Currently, MACD for KOTAK stock is 25.42 indicating that KOTAK stock is showing an upward trend. Currently, MACD for SBI stock is 14.07indicating that SBI 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 the stock. Stochastic in the range of 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 of 20-45, the stock will indicate trending downwards. Stochastic above 80 would mean that stock is overbought and less than 80 will tell that stock is oversold. Currently Stochastic or HDFC stock is89.62which means that HDFC stock is overbought and hence the investor should wait for some time so that the stochastic indicator gives a lesser value. Currently Stochastic for KOTAK stock is 76.32 which mean that KOTAK stock is showing an upward trend. Currently Stochastic for SBI stock is 95.02which means that SBI stock is overbought hence the investor should wait for some time so that the stochastic indicator gives a lesser value.

It can be decided how strongly the stock is trending upwards or downwards using ADX. For 14 days, an increasing ADX will indicate 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. Currently, KOTAK stock ADX is **37.66**meaning it will show a strong upward or downward trend. Currently, SBI stock ADX is 30.53 meaning it will show a strong upward or downward trend.

Bollinger's band is positive and negative standard deviations from SMA. For 20 days, if the close price of the stock moves quite away from a positive standard deviation will mean that the stock is overbought, and if the close price of the stock moves away from a negative standard deviation, then the stock will be considered oversold. Currently, the upper band and the lower band for HDFC stock are 1514.69 and 1,261.46 respectively. The close price of HDFC stock is 1493.05 which mean HDFC stock is overbought. Currently, the upper band and the lower band for KOTAK stock are 1,970.16 and 1,854.16 respectively. The close price of KOTAK stock is 1944.20 which mean KOTAK stock is showing a sideways trend but may soon show an upward trend. Currently, the upper band and the lower band for the SBI stock are 582.40 and 505.09 respectively. The close price of SBI stock is 575.05which means SBI stock is showing a sideways trend but may soon show an upward trend(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 are being used for this study. This study uses NSE Data. Following are the details for every column used in the HDFC, KOTAK, and SBI datasets.

Name and symbol: This column tells 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 in addition 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 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.



Fig7.1 class distribution For HDFC, KOTAK, and SBI stock

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. As per the actual data available, HDFC STOCK is moving 2140 times in an upward direction and is suitable for Long trading whereas 3435 times, it is not moving in an upward direction. KOTAK STOCK is moving 2055 times in an upward direction and is suitable for long trading whereas 3199 times, it is not moving in an upward direction. SBI STOCK is moving 2211 times in an upward direction and is suitable for long trading whereas 3364 times, it is not moving in an upward direction.



Fig7.2 Close values of HDFCBANK, KOTAK BANK, and SBIBANK stock from 2000 to 2022



Fig7.3 Distribution Plot for the HDFCBANK Stock



Fig7.4 Distribution Plot for the KOTAKBANK Stock



Fig7.5 Distribution Plot for the SBIBANK Stock

The mean value is greater than the median value meaning Data has a positively skewed distribution which is observed in all 3 stocks namely HDFC, KOTAK, and SBI bank stock. However, SBIBANK stock is looking as the least volatile stock followed by HDFC Bank stock. KOTAK Bank stocks exhibit maximum volatility compared to the other two stocks.



Fig7.6 Customized Box plot for the HDFCBANK, KOTAK BANK, and SBIBANK stock from 2000 to 2022

There is notably a large difference between the 75th %tile and max values of most of the feature variables for all 3 stocks. Therefore, it suggests that there are extreme values-Outliers in our data set.



Fig7.7 Customized Scatter Plot against close price for the HDFCBANK Stock from 2000 to 2022



Fig7.8 Customized Scatter Plot against close price for the KOTAKBANK Stock from 2000 to 2022



Fig7.9 Customized Scatter Plot against close price for the SBIBANK Stock from 2000 to 2022

A customized Scatter Plot is drawn for all feature variables against the close price of the HDFC, KOTAK, and SBI stock. It is observed that a linear relationship exists between Independent variables and the Target variable except for fewer outliers which is quite negligible.

The previous Chapter explains the HDFC, KOTAK, and SBI stock-related feature variables that may be used as the independent variables. The direction of the close price of the HDFC, KOTAK, and SBI 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, KOTAK, and SBI stock, and the findings are 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, KOTAK, and SBI data which is taken from NSE come with a lot of limitations that have to be processed which include 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:** Computed variables were added to the dataset that 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. 1 day's previous lag values of volume are also added in concert with 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, the 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, and 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: Modeling

**Modelling strategies and Model Evaluation Rule:**

Table 9.1– Modelling strategies and Model Evaluation Rule

|  |  |
| --- | --- |
| **Modelling Strategies** | **Model Evaluation Rule** |
| Direction Detection by 6, 10, and 14 days consecutive closing prices split week on the week. | percentage change on closing price>0.7% =>Positive Trend  percentage change on closing price<-0.7% =>Negative Trend  percentage change on closing price between 0.7 and 0.7% =>Neutral |
| Go Long Direction Prediction performed separately using Momentum, Trend, Volatility, and Volume Indicators | percentage change on closing price>0.5% =>Positive Trend  percentage change on closing price<=0.5% =>Not Positive Trend |

**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 getting 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 a 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. 0.5% difference,1% difference, and 1.5% difference -these are different classes of direction for which the rule is being set which is to be followed for computing the direction change as either positive change, negative change, or no change. 0.7% difference as a class of direction can also be used in place of 0.5% change if that gives a better directional indicator.

Once it is determined to 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 getting 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 increased 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 are being 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 are being utilized as feature variables based on the input dataset and different classification models are 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 as 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 redeployed 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 Bank from the year 2000 to 2022 are being used to repeat the entire process which had been implemented for the HDFC Bank 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 result of the Data Modelling phase and discusses the Metrics utilized to determine the extent of the success achieved from the different Modelling Algorithms employed on the Target Variable.

# Chapter 10: Model 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, KOTAK, and SBI 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.

**Model Evaluation using LR Classifier for Go Long Direction Prediction:**

Table 10.1– Model Evaluation using LR Classifier for Go Long Direction Prediction

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelling Strategies** | **HDFC** | **KOTAK** | **SBI** |
| Direction Detection by 6,10,14 days consecutive closing prices split week on the week | precision-0.35  recall-0.60  accuracy-0.35 | Precision-0.37  recall-0.74  accuracy-0.36 | Precision-0.36  recall-1.00  accuracy-0.36 |
| Go Long Direction Prediction using Volume Indicators | **precision-0.98**  **recall-0.83**  **accuracy-0.92** | **precision-0.99**  **recall-0.93**  **accuracy-0.97** | **precision-0.92**  **recall-0.80**  **accuracy-0.90** |
| Go Long Direction Prediction using Momentum Indicators | precision-0.71  recall-0.63  Accuracy-0.76 | precision-0.73  recall-0.61  accuracy-0.75 | precision-0.69  recall-0.62  accuracy-0.74 |
| Go Long Direction Prediction using Trend Indicators | precision-0.83  recall-0.59  Accuracy-0.80 | precision-0.76  recall-0.48  accuracy-0.72 | precision-0.78  recall-0.49  accuracy-0.74 |
| Go Long Direction Prediction using Volatility Indicators | precision-0.93  recall-0.47  Accuracy-0.77 | precision-0.90  recall-0.40  accuracy-0.74 | precision-0.81  recall-0.30  accuracy-0.70 |

From Table 10.1, it can be observed that Go Long Direction Prediction using Volume Indicators has given considerable precision, recall, and accuracy in direction prediction.

**Model Evaluation using RF Classifier for Go Long Direction Prediction:**

Table 10.2– Model Evaluation using RF Classifier for Go Long Direction Prediction

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelling Strategies** | **HDFC** | **KOTAK** | **SBI** |
| Direction Detection by 6,10,14 days consecutive closing prices split week on the week | **precision-0.85**  **recall-0.89**  **accuracy-0.87** | **Precision-0.71**  **recall-0.79**  **accuracy-0.74** | **Precision-0.83**  **recall-0.88**  **accuracy-0.85** |
| Go Long Direction Prediction using Volume Indicators | **precision-0.93**  **recall-0.69**  **accuracy-0.85** | **precision-0.92**  **recall-0.79**  **accuracy-0.89** | **precision-0.90**  **recall-0.73**  **accuracy-0.86** |
| Go Long Direction Prediction using Momentum Indicators | precision-0.76  recall-0.51  Accuracy-0.75 | precision-0.78  recall-0.50  accuracy-0.75 | precision-0.72  recall-0.55  accuracy-0.74 |
| Go Long Direction Prediction using Trend Indicators | precision-0.87  recall-0.56  Accuracy-0.80 | precision-0.85  recall-0.44  accuracy-0.74 | precision-0.83  recall-0.57  accuracy-0.78 |
| Go Long Direction Prediction using Volatility Indicators | precision-0.92  recall-0.53  Accuracy-0.79 | precision-0.89  recall-0.50  accuracy-0.78 | precision-0.83  recall-0.61  accuracy-0.80 |

From Table 10.2, it can be observed that Direction Detection has given the highest precision, accuracy, and recall in prediction. Also, Go Long Direction Prediction using Volume Indicators has given considerable precision and accuracy in direction prediction but recall can still be improved.

**Model Evaluation using XG Boost Classifier for Go Long Direction Prediction:**

Table 10.3– Model Evaluation using XG Boost Classifier for Go Long Direction Prediction

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelling Strategies** | **HDFC** | **KOTAK** | **SBI** |
| Direction Detection by 6,10,14 days consecutive closing prices split week on the week | precision-0.35  recall-0.42  accuracy-0.40 | Precision-0.38  recall-0.41  accuracy-0.40 | Precision-0.38  recall-0.47  accuracy-0.37 |
| Go Long Direction Prediction using Volume Indicators | **precision-0.90**  **recall-0.73**  **accuracy-0.86** | **precision-0.92**  **recall-0.87**  **accuracy-0.92** | **precision-0.88**  **recall-0.82**  **accuracy-0.89** |
| Go Long Direction Prediction using Momentum Indicators | precision-0.70  recall-0.61  Accuracy-0.75 | precision-0.74  recall-0.59  accuracy-0.75 | precision-0.70  recall-0.59  accuracy-0.74 |
| Go Long Direction Prediction using Trend Indicators | precision-0.85  recall-0.65  Accuracy-0.82 | precision-0.82  recall-0.61  accuracy-0.79 | precision-0.83  recall-0.67  accuracy-0.81 |
| Go Long Direction Prediction using Volatility Indicators | precision-0.84  recall-0.69  Accuracy-0.82 | precision-0.81  recall-0.63  accuracy-0.79 | precision-0.80  recall-0.67  accuracy-0.81 |

From Table 10.3, it can be observed that Go Long Direction Prediction using Volume Indicators has given considerable precision, recall, and accuracy in direction prediction.

# Chapter 11: Analysis and Results

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

Analysis for HDFC, KOTAK, and SBI Stock is given below.

Direction Detection and **Go Long Direction Prediction using the** best classifier model:

Table 11.1– Leader Board-comparison of Metrics for Direction Detection and Go Long Direction Prediction using the best classifier model

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelling Strategies** | **HDFC** | **KOTAK** | **SBI** |
| Direction Detection by 6,10,14 days consecutive closing prices split week on the week  (RF Classifier) | **precision-0.85**  **recall-0.89**  **accuracy-0.87** | **Precision-0.71**  **recall-0.79**  **accuracy-0.74** | **Precision-0.83**  **recall-0.88**  **accuracy-0.85** |
| Go Long Direction Prediction using  Volume Indicators  (LR Classifier) | **precision-0.98**  **recall-0.83**  **accuracy-0.92** | **precision-0.99**  **recall-0.93**  **accuracy-0.97** | **precision-0.92**  **recall-0.80**  **accuracy-0.90** |
| Go Long Direction Prediction using  Momentum Indicators  (LR Classifier) | precision-0.71  recall-0.63  Accuracy-0.76 | precision-0.73  recall-0.61  accuracy-0.75 | precision-0.69  recall-0.62  accuracy-0.74 |
| Go Long Direction Prediction using  Trend Indicators  (XG Boost Classifier) | precision-0.85  recall-0.65  Accuracy-0.82 | precision-0.82  recall-0.61  accuracy-0.79 | precision-0.83  recall-0.67  accuracy-0.81 |
| Go Long Direction Prediction using  Volatility Indicators  (XG Boost Classifier) | precision-0.84  recall-0.69  Accuracy-0.82 | precision-0.81  recall-0.63  accuracy-0.79 | precision-0.80  recall-0.67  accuracy-0.81 |

From Table 11.1, it can be observed that RFclassifier modelling has given the highest efficiency in Direction Detection among all Modelling techniques namely LR, DT, RF, KNN, and XG Boost Modelling. 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. Also, LR classifier modelling has provided the best precision, recall, and accuracy for Go Long Direction prediction using Volume Indicators.

Utility from the Business perspectives

For a stop loss of 2.0 reward-risk ratio for approximately 0.8 Precision would be 2\*.8/2\*.2=4:1 if a 0.5% difference in consecutive day close price for any stock is only 2.0.for higher percentage difference reward to risk ratio would be higher.

Here, Modelling Algorithms were provided for the close price of HDFCBANK, KOTAK BANK, and SBIBANK Stock over 20 years with the train test split of 70%:30%. If we invest Rs.10000 for 6 years and roughly calculate profit with 0.5% change on close price with the highest precision in detecting true positives then the following results are possible as per the formulae given below:



**Go Long Direction Prediction**: Using Volume Indicators with the highest precision of 0.99 for KOTAK BANK stock, the confusion matrix provides information as below:



Fig11.1 confusion matrix For KOTAKBANK Stock using

Volume Indicators as Feature variables

Therefore, Net Returns will be:

0.5\*10000\*386\*0.99/100-0.5\*10000\*4\*0.99/100

=Rs.18, 909 profit which would be 18909/ (10000\*6)\*100=31.52% returns.

Using Momentum Indicators with the highest precision of 0.73 for KOTAK BANK stock, the confusion matrix provides information as below:



Fig11.2 confusion matrix For KOTAKBANK Stock using

Momentum Indicators as Feature variables

Therefore, Net Returns will be:

0.5\*10000\*251\*0.73/100-0.5\*10000\*93\*0.73/100=Rs.5767 profit which would be 5767/ (10000\*6)\*100=9.61% returns.

Using Trend Indicators with the highest precision of 0.85 for HDFCBANK stock, the confusion matrix provides information as below:



Fig11.3 confusion matrix For HDFCBANK Stock using

Trend Indicators as Feature variables

Therefore, Net Returns will be:

0.5\*10000\*282\*0.85/100-0.5\*10000\*51\*0.85/100=Rs. 9817.5 profit which would be

9817.5 / (10000\*6)\*100=16.36% returns.

Using Volatility Indicators with the highest precision of 0.84 for HDFCBANK stock, the confusion matrix provides information as below:



Fig11.4 confusion matrix For HDFCBANK Stock using

Volatility Indicators as Feature variables

Therefore, Net Returns will be:

0.5\*10000\*313\*0.84/100-0.5\*10000\*58\*0.84/100=Rs.10710 profit which would be 10710// (10000\*6)\*100=17.85% returns.

Hence, average returns using Go Long Direction Prediction would be much higher compared to Average stock market returns using Bank interest returns which would range from 7.0-7.5%.

Risk-Adjusted Returns:

The real Data dump is imported for HDFC, KOTAK, and SBI stock between 2000 till 2022. Then the Return, Variance, and Volatility of these stocks are calculated following which the Annualized return to Risk ratio and finally, the Sharpe ratios are calculated.

The Sharpe ratio for HDFCBANK Stock is calculated as 0.173818.

The Sharpe ratio for KOTAK BANK Stock is calculated as 0.149589.

The Sharpe ratio for SBIBANK Stock is calculated as 0.005306.

Therefore, from the results obtained it becomes evident that HFDC shows a better Return vs. Risk performance over the specified period compared to KOTAK stock followed by the SBI stock which shows the least Return vs. Risk performance.

# Chapter 12: Deployment

In the Future, there is a deployment Dashboard proposed. The data pipelines shown below explain the deployment plan to be taken up where the business requirements would be to develop front-end API-based executable applications.



Fig12.1Deployment Proposal

As per the proposal for future assignments, the dashboard takes API as input.

This is derived from machine learning algorithms. The deployment will cater to multi-label features with an end-to-end UI Interface.



Fig12.2 Illustration of Dashboard

# 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 getting 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 the 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 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 are deployed and their prediction accuracy is 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. In the Future, there is a deployment Dashboard proposed. An intelligent automated system for Options Trading would be also the next step forward.

# 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-2)

Plagiarism Report with below 15% similarly index to be attached in the annexure. The title page and last pages with the similarity index report are attached.

## Paper Publications in a Journal/Conference Presented/White Paper[[2]](#footnote-3)

1. **Publication** in a Journal/Conference Presented/White Paper – **Full paper** extracted from the journal / full conference paper and **the certificate** must be attached.
   1. Those who have **published**: (Attach full paper)

[All authors Name in order], “[Article Title].” [Journal Name/Conference Name], [Volume Number], [Issue Number], [Year], [Pages], DOI.

* 1. For those who have **not published** yet may add submission information: (Attach full paper)

[All authors Name in order], “[Article Title].” [Journal Name/Conference Name], [Date of submission]

*Eg: Madhukeshwar R* K, Ratnakar Pandey, Shinu Abhi, “A Recommender System for Indian Credit Cards using Text Analytics.” NeuroQuantology, Volume 20, Issue8, 2022, Page 9021-9028, doi:10.14704/nq.2022.20.8.NQ44922

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Table

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

## Certificate for the Conference Presentation

Text

Description automatically generated

## GitHub Link

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

**Additional Instructions (To be removed while submitting)**

**Formatting for Printing:**

1. The printing of your capstone project must be done with a RACE-approved printer only. Details will be shared.
2. The cover page shall be in White hardbound (for both 1st and 2nd-year projects) as per the template shared.
3. The report shall be printed on an A4sizeExecutiveBondsheet.
4. The font used shall be Times New Roman, and the font size shall be 12. For the Heading, use **Times New Roman 14 in Bold,** and for the subheading, use **Times New Roman 12 in Bold**.
5. The top, bottom, left and right margins shall be 1” each.
6. The line spacing shall be fixed at1.5lines.
7. Table line spacing shall be single line spacing.
8. Page numbers shall be placed at the bottom right position.
9. Chapters shall be numbered 1, 2, 3, etc. The tables and charts shall be in the format of 1.1, 1.2, etc. i.e., 1.1 indicate that it is the first table in Chapter 1; 2.1 Indicates the first table in Chapter 2. Similarly, chart no. 1.1 indicates the first chart in Chapter 1.
10. The project report shall be a minimum of 40 pages and shall not exceed 75 pages for second-year projects and a minimum of 30 pages, and a maximum of 50 pages for first-year projects.
11. You must submit three hard copies duly signed by the mentor and guide (scanned signature will be sufficient) and the Director along with a soft copy in pdf format. (Two copies to submit to the university and one is your copy).
12. The title of the study, objectives, analysis, findings, and suggestions must tally.
13. Each chapter must start on a fresh page.

**Mandatory Inclusions**

1. **Plagiarism Report** with below 15% similarly index to be attached in the annexure. The title page and last pages with the similarity index report are attached.
2. **Publication** in a Journal/Conference Presented/White Paper – **Full paper** extracted from the journal / full conference paper and **the certificate** must be attached.
   1. Those who have **published**: (Attach full paper)

[Authors Name], “[Article Title].” [Journal Name/Conference Name], [Volume Number], [Issue Number], [Year], [Pages], DOI.

Those who have **not published** yet may add submission information: (Attach full paper)

[Authors Name], “[Article Title].” [Journal Name/Conference Name], [Date of submission]

1. GitHub Link
2. Any other annexure (optional).

1. Turnitin report to be attached from the University. [↑](#footnote-ref-2)
2. URL of the white paper/Paper published in a Journal/Paper presented in a Conference/Certificates to be provided. [↑](#footnote-ref-3)