Modelling Direction Detection in Selected Stocks in Indian BFSI Sector

by Anand Mohan

Submission date: 10-Mar-2023 03:08PM (UTC+0530)
Submission ID: 2033790533
File name: Report_Anand_Mohan_Direction_Detection_Selected_Stocks_2023.docx (542.95K)
Word count: 7942
Character count: 45347

Abstract

Advanced Machine learning techniques are getting remarkably popular in predicting stock market returns. 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. 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.

The six-day consecutive closing prices will be getting tabulated week on week for the entire dataset. The difference between the seventh and eighth-day Closing price is determined. 0.5%, 1%, and 1.5% differences are different classes of direction for which the rule is being set 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. A similar process is repeated for a range of consecutive days to be utilized as the feature variable increased to ten days and fourteen days. Similarly, momentum, trend, volatility, 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. Random Forest (RF) modelling has given the highest efficiency in direction detection. Logistic Regression (LR) modelling provided the best prediction performance for volume and momentum indicators, whereas Extreme Gradient (XG) Boost classifier provided the best prediction performance for trend and volatility indicators for predicting the upward trend of the close price. The invaluable takeaway from the capstone is that various classification modelling techniques have 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, Decision Tree (DT), Random Forest (RF), K-Nearest Neighbours (KNN), XGBoost, Principal Component Analysis (PCA), Housing Development Finance Corporation Limited (HDFC), KOTAK, State Bank of India (SBI)

Chapter 1: Introduction

The stock market encourages the free economy concept. It is one of the significant financial tools in the hands of the corporate and enterprises to raise their funds through investments done by the common man. In return for investors putting their stake in company stocks, it is expected that they earn profits through dividends and upward stock movements, which would also enhance their economic status apart from the growth of the participant company whose stocks are at stake in the public domain.

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).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).

There have been numerous technological advancements made ~~in the area of~~in fundamental and technical analysis, which aid conventional investors in their decision-making. Various technical indicators, namely momentum, trend, volatility, and volume indicators, are conceived based on open, close, high, and low volumes of the stock under consideration. Fund managers and investment managers help the common man improve their prospects also through algorithmic trading and other automated trading innovations existent in contemporary times. However, there has been numerous evidence of market manipulations and fraudulent practices by some dubious advisory firms, and therefore, the unguarded speculations have led to enormous losses for the common investors and resulted in a bad reputation for the stock market concerning some unfair practices that are existent in current financial markets.

The stock market, ~~as a result of~~because 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 has created the potential to predict the market to an extent (Sonkiya 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).

The need of the study is to showcase the merits of machine learning and artificial intelligence over manual and algorithmic trading. The scope of the study is to explore unsupervised and various supervised classification techniques to be able to predict the upward direction in any stock under consideration. Investors can find plenty of algorithms that detect the exact closing price of any stock. However, a unique attempt is being made for this thesis to predict the direction of the closing price for any stock under consideration and then evaluate the prediction performance via metrics for evaluation of the modelling techniques used for this project.

Daily trading National Stock Exchange (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.

This chapter ~~e current chapter~~ discussed the importance of machine-learning-associated techniques that are developed for investments in the stock market. The chapter additionally informs 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 is scanned, which throws light on various related aspects of machine-learning methods and other methodologies. It also studies and researches other related issues that help better in direction detection in the stock market.

Chapter 2: Literature Review

The stock market is a significant entity in the financial sector, allowing companies to improve their money prospects. In return for investors putting their stake in company stocks, they earn profits through dividends and upward stock movements. The literature review initially scans through technical and fundamental analysis of stocks. Further, it discusses how algorithmic trading based on fundamentals and technical indicators helps investors in their decision-making. Further, it emphasizes the merits of machine learning and artificial intelligence over algorithmic trading. It talks about unsupervised and various supervised classification techniques used in this thesis. Later it reviews the literature on confusion matrix, discussing various metrics for evaluation of the modelling techniques used for this project.

There are numerous parameters impacting value movements in varied sizes and layers in the stock market (Rajkar et al., 2021~~).Therefore~~). 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 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 Moving Average Convergence Divergence (MACD), moving average, ~~etc~~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 gives 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 (Magner et al., 2021).Momentum-based trading commerce is among proven 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 (Hansen, 2020~~).However~~). However, regulators have restrained algorithmic commerce, following accusations of market manipulation (Mukerji et al., 2019).

Machine learning and artificial intelligence have been increasingly used in the field of business analytics. However, it is suggested that exploratory data analysis should be performed to gain a better data understanding (Omta et al., 2020~~).Some~~). 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 researched further 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 has been the most popular statistical technique utilized in pattern identification over the last four decades (Wang, 2019).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~~must be tested. 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~~). Accuracy~~). Accuracy determines the precision of a model according to its correctly classified true positives and true negatives in the dataset. AUC compares the rates of false-positive and true-positive in the dataset (Silva & Naranjo, 2020).

The current 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 that is getting discussed in the next chapter.

Chapter 3: Problem Statement

There are plenty of regression algorithms that are 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 there is lesser data to build a regression model, underfitting scenarios may destroy the accuracy of our machine learning model, especially when a linear model is being built with a complex dataset. At certain times, while trying to cater to all kinds of both existent and nonexistent possibilities in data points, overfitting 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, alternate approaches need to be tried as well 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 added that result 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.

Commented [DSA1]: Start as Identify and collect the right stock...Avoid using Firstly, Secondly etc.

Chapter 5: Project Methodology

~~The current~~This chapter introspects more on the project methodology that is getting implemented and endeavours for continuous improvement that will be taken up while working on the project.

The Cross-Industry Standard Process for Data Mining (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, feature addition, and data scaling using MinMax Scaler are 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 are used in the data modelling phase. Deployment discusses developing a front-end Application Programming Interface (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).

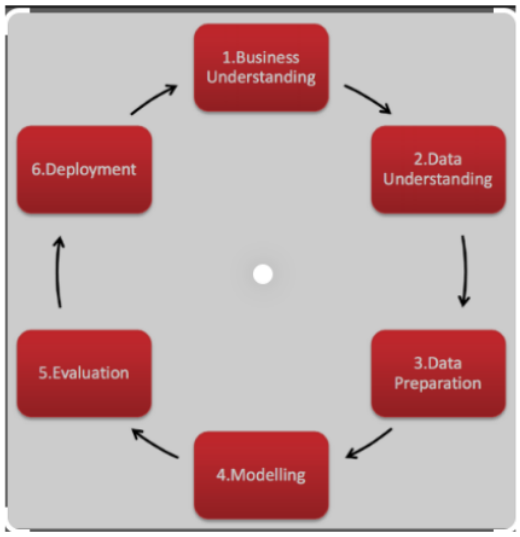

Fig. 5.1 CRISP-DM Process Diagram (Cornellius Yudha Wijaya, 2021)

Fig. 5.1describes the CRISP-DM framework. The framework comprises six different phases. Threads from business understanding are gathered to ~~more or less get~~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(moneycontrol, n.d.)

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

Table 6.1 performs a fundamental analysis of HDFC, KOTAK, and SBI stock. HDFC Bank's 52-week high is 1,725 and 52 weeks low is 1,271.60. It ~~is located in~~is in India, Bahrain, Hong Kong, and Dubai. It has 6,378 branches, 18,620 ATMs, and 21,683 banking outlets. It was established in 1994. It has its corporate office in Mumbai, India. KOTAK Bank's 52-week high is 2,253 and the 52-week low is 1,631. It ~~is located in~~is in India, London, New York, California, Dubai, Abu Dhabi, Mauritius, and Singapore. It has 1,702 branches and 2,761 ATMs. It was established in 1985. It has its corporate office 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~~is in India, Australia, Bangladesh, Belgium, Bhutan, Canada, China, Germany, and Hong Kong. It has 22,266 branches and 65,030 ATMs. It was established in 1806.It has its corporate office in Mumbai, India (moneycontrol, n.d.).

Commented [DSA2]: Add the date in Mendeley. Or else, Non dated message will appear.

Technical analysis of HDFC, KOTAK, and SBI stock:

For 14 days, if Relative Strength Index (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' exponential moving averages from 12 days' exponential moving averages. If the MACD is more than 0 and also greater than 9 days of exponential moving averages, the stock will be trending upwards. Any stock trends downwards If the MACD is less than zero and also lesser than nine days of exponential moving averages. 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 is 89.62which 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, the Stochastic for KOTAK stock is 76.32 which means that KOTAK stock is showing an upward trend. Currently, the 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 is decided how strongly the stock is trending upwards or downwards using Average Directional Index (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 simple moving averages. 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 means HDFC stock is overbought. Currently, the upper band and the lower band for KOTAK stock are 1,970.16 and1,854.16, respectively. The close price of KOTAK stock is 1944.20, which means 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 current~~This chapter discusses the fundamental and technical analysis of HDFC, KOTAK, and SBI stock. The next chapter discusses the data understanding of the CRISP-DM framework. The data Understanding section gets 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.

The name and symbol 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 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. *Volume-Weighted Average Price* (VWAP) represents the typical worth that the security listed throughout the day, based on both volume and worth. The 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.

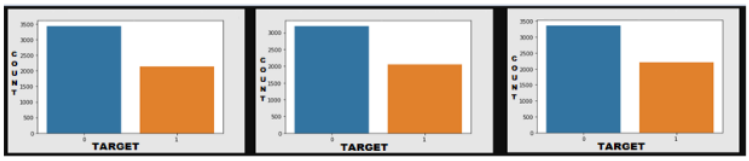

Fig. 7.1 Class distribution For HDFC, KOTAK, and SBI stock

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 shown in Fig 7.1, 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.

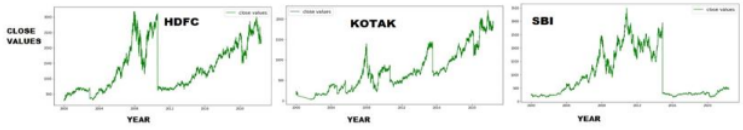

Fig. 7.2 Close values of HDFCBANK, KOTAK BANK, and SBIBANK stock from 2000 to 2022

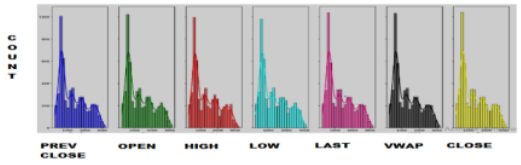

Fig. 7.3 Distribution Plot for the HDFCBANK Stock

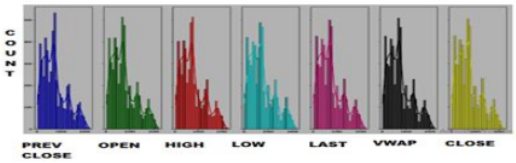

Fig. 7.4 Distribution Plot for the KOTAKBANK Stock

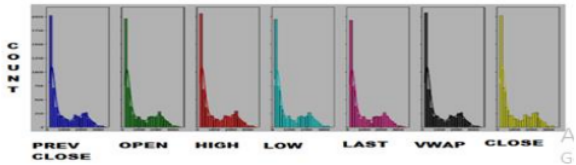

Fig. 7.5 Distribution Plot for the SBIBANK Stock

As shown in Fig. 7.2, 7.3, 7.4, and 7.5 the mean value is greater than the median value meaning data has a positively skewed distribution which is observed in all three 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.

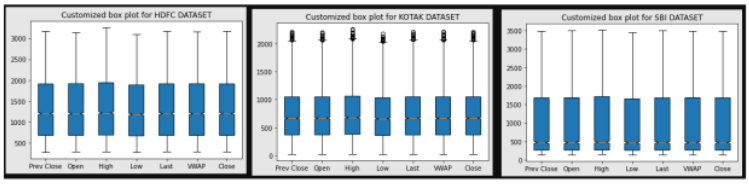

Fig. 7.6 Customized Box plot for the HDFCBANK, KOTAK BANK, and SBIBANK stock from 2000 to 2022

As shown in Fig.7.6, there is notably a large difference between the 75th percentile and max values of most of the feature variables for all three stocks. Therefore, it suggests that there are extreme values-Outliers in our data set.

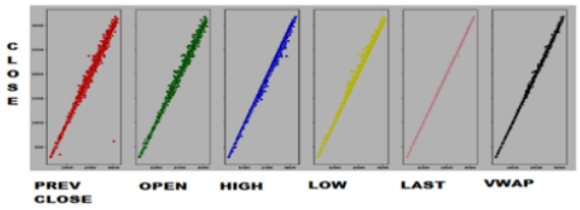

Fig. 7.7 Customized Scatter Plot against close price for the HDFCBANK Stock from 2000 to 2022

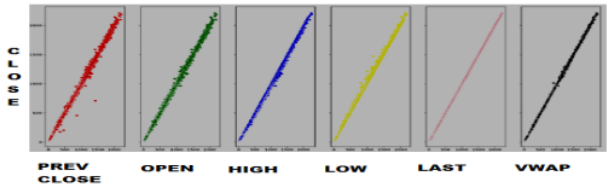

Fig. 7.8 Customized Scatter Plot against close price for the KOTAKBANK Stock from 2000 to 2022

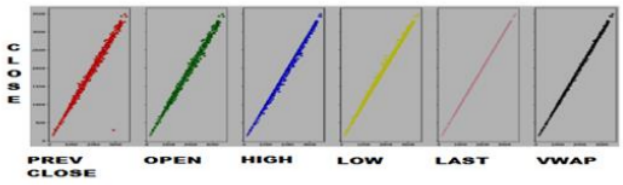

Fig. 7.9 Customized Scatter Plot against close price for the SBIBANK Stock from 2000 to 2022

As shown in Fig 7.7, 7.8, and 7.9 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 current chapter discusses 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 discusses the data preparation of our CRISP-DM framework. Within the data preparation, 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~~must 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 have 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, thirteen days, twenty days, hundred days, and two hundred days. Conjointly enclosed were exponential moving averages for seven days, thirteen days, twenty days, hundred days, and two hundred days. That's going to be useful in evaluating the securities market returns. One day of previous lag values of volume is also added in concert with the input feature. Six, ten, fourteen, and thirty 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, Kaufman's Adaptive Moving Average indicator, Percentage Price Oscillator, Percentage Volume Oscillator, Rate of Change indicator, RSI indicator, Stochastic Oscillator, True strength index indicator, Ultimate Oscillator, WilliamsR indicator are getting 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, Commodity Channel Index indicator, Ichimoku indicator, Know Sure Thing indicator, MACD, Parabolic Stop and Reverse indicator, Exponential Moving Average indicator, Weighted Moving Average 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, Chaikin Money Flow indicator, Ease of Movement indicator, Force Index indicator, Money Flow Index indicator, On Balance Volume indicator, Volume Price Trend indicator, VolumeWeightedAveragePrice, Negative Volume Index indicator, Daily Log Return indicator are used as feature variables.

MinMax Scaler for Scaling Data: Features should be approximately normally distributed and on relatively similar scales for getting higher performance with machine learning algorithms. 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 LDA, 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 standard deviation 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 standard deviation calculated using the fit () methodology. Both fit and transform accomplished using fit transform () method.

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 most of the 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 current chapter is intended to make the data ready for the model-building processes. The next chapter explains the data modelling of the CRISP-DM framework.

Chapter 9: Modeling

Based on direction detection accuracy, it can be suggested to the prospective investor whether to invest or not invest in stock. Direction prediction accuracy is further determined using momentum, trend, volatility, and volume indicators as feature variables and building different classification models on them. Table 9.1 explains the modelling strategies and model Evaluation Rule used for this project.

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

8 The difference between the seventh and ~~eighth day~~eighth day closing price is determined. If the ~~eighth day~~eighth day closing price is seen as an increase from the seventh day by 0.7% or more, the direction of the closing price can be made positive. If the ~~eighth day~~eighth day closing price is seen as a decrease from the seventh 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 seventh and eighth-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 the parameter for the best prediction accuracy is determined, say for example, 0.7% 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 ten days. Therefore, a ten-day consecutive closing price for the stock under consideration is being taken. These ten 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 increases to fourteen 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 ten and fourteen days, respectively.

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 0.5% percentage change of the close price 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 redeployed, and their prediction accuracy is being compared.

When the majority of the twenty 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 processes 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 current chapter focused on employing various modelling algorithms to determine the accuracy of the trend prediction. The next chapter discusses 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. In the current chapter, 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 examines all the results derived from the various models and figures 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 week	Pprecision-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** **uracy-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 acy-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 uracy-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 is observed that go-long direction prediction using volume indicators has given considerable precision, recall, and accuracy in direction prediction.

Commented [DSA3]: Keep consistency in caps.

Model evaluation using RF classifier for go long direction prediction:

From Table 10.2, it is observed that direction detection using RF classifier 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.

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

Model evaluation using XG Boost classifier for go long direction prediction:

From Table 10.3, it is observed that go-long direction prediction using volume indicators has given considerable precision, recall, and accuracy in 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

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 is 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 six, ten, and fourteen-day consecutive closing prices split week on week as six, ten, and fourteen 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 six 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 Equation (1):

$$n = numberofdaysitistrueposition$$
$$m = numberofdaysitisfalseposition$$
$$pr = precision$$
$$p = percentagechange$$
$$c = capitalinvested$$
$$netreturns = (p * c * n * pr - p * c * m * pr)/100 \quad (1)$$

Go long direction prediction:

Using volume indicators with the highest precision of 0.99for KOTAK BANK stock, the confusion matrix provides information as shown in Fig 11.1 below:

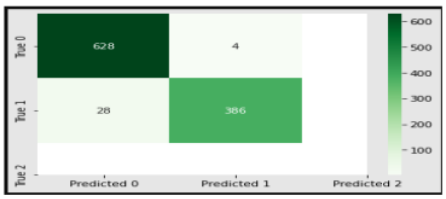

Fig. 11.1 Confusion matrix For KOTAKBANK Stock using Volume Indicators as Feature variables

Therefore, net returns are:
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 0.73for KOTAK BANK stock, the confusion matrix provides information as shown in Fig 11.2 below:

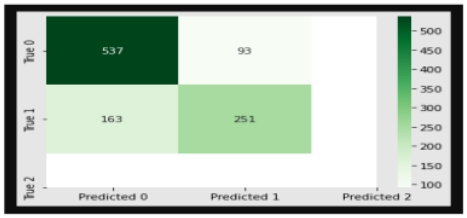

Fig. 11.2 Confusion matrix For KOTAKBANK Stock using Momentum Indicators as Feature variables

Therefore, net returns are:
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 shown in Fig 11.3 below:

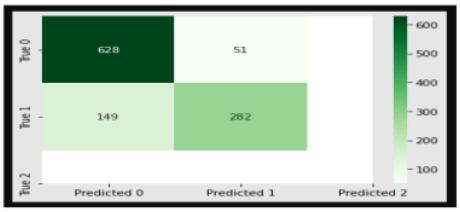

Fig. 11.3 Confusion matrix For HDFCBANK Stock using Trend Indicators as Feature variables

Therefore, net returns are:
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 0.84for HDFCBANK stock, the confusion matrix provides information as shown in Fig 11.4 below:

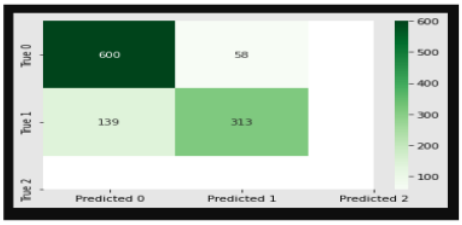

Fig. 11.4 Confusion matrix For HDFCBANK Stock using Volatility Indicators as Feature variables

Therefore, net returns are:
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 are 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 KOTAKBANK 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

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.

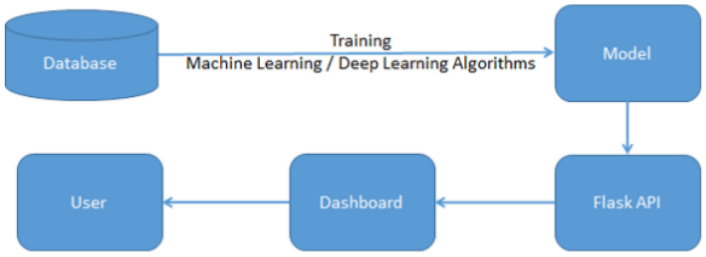

Fig. 12.1 Deployment Proposal

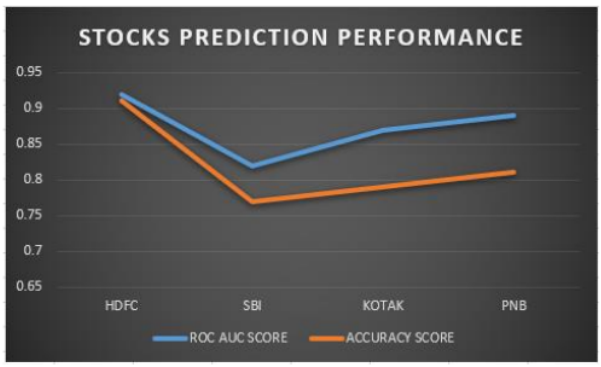

Fig. 12.2 Illustration of Dashboard

Fig12.1 showcases the deployment proposal and Fig12.2 performs further illustration of the dashboard. 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 User Interface.

Chapter 13: Conclusions and Future Scope

The six-day consecutive closing price for the stock under consideration is being taken. These six days' consecutive closing prices will be getting tabulated week on week for the entire dataset and will be utilized as six different feature variables for building the classification Model. The difference between the seventh and ~~eighth day~~eighth day Closing price is determined. The rule is being set to determine what ~~has to~~must 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~~deployed , and their prediction accuracy is compared using Metrics, namely precision, recall, f1-score, and accuracy score. Once the parameter for the best prediction accuracy is determined say for example 0.7% 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 ten days and fourteen 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 , accuracy score, and ROC AUC Score.

The construction of all twenty 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 will be 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 does not address a major drawback of stock prediction, which is, 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 will be a deployment dashboard proposed. An intelligent automated system for options trading would also be the next step forward.

Modelling Direction Detection in Selected Stocks in Indian BFSI Sector

ORIGINALITY REPORT

8%	5%	6%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Harrisburg University of Science and Technology Student Paper	2%
2	Pavan Kumar Nagula, Christos Alexakis. "A new hybrid machine learning model for predicting the bitcoin (BTC-USD) price", Journal of Behavioral and Experimental Finance, 2022 Publication	1%
3	Submitted to Leiden University Student Paper	1%
4	Paola Cerchiello. "Data Mining for Business and Industry", Statistics in Practice, 03/03/2008 Publication	<1%
5	Submitted to University of Salford Student Paper	<1%
6	www.investopedia.com Internet Source	<1%

7	Submitted to University of New England Student Paper	<1 %
8	irep.ntu.ac.uk Internet Source	<1 %
9	Submitted to Liverpool John Moores University Student Paper	<1 %
10	Marco Cristani, Vittorio Murino. "Chapter 10 Socially-Driven Computer Vision for Group Behavior Analysis", Springer Science and Business Media LLC, 2014 Publication	<1 %
11	hdl.handle.net Internet Source	<1 %
12	"Annotations.", The Lancet, 18970911 Publication	<1 %
13	Nagaraj Naik, Biju R Mohan. "Novel Stock Crisis Prediction Technique-A Study on Indian Stock Market", IEEE Access, 2021 Publication	<1 %
14	patternswizard.com Internet Source	<1 %
15	Submitted to Loughborough University Student Paper	<1 %
16	Submitted to University of South Australia Student Paper	

		<1%
17	Submitted to Queensland University of Technology Student Paper	<1%
18	Submitted to The Hong Kong Polytechnic University Student Paper	<1%
19	core.ac.uk Internet Source	<1%
20	media.neliti.com Internet Source	<1%
21	towardsdatascience.com Internet Source	<1%
22	insurance.kotak.com Internet Source	<1%
23	www.forexrealm.com Internet Source	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		