Stock Price Movement Prediction using Fundamental Analysis and Scope of Sentiment Analysis in the Domain.

Shruti Sindhi

3075407

Submitted in partial fulfillment for the degree of

Master of Science in Big Data Management and Analytics

Griffith College Dublin
September 2022

Under the supervision of Aqeel Kazmi

Disclaimer

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the Degree of Master of Science in Big Data Management and Analytics at Griffith College Dublin, is entirely my own work and has not been submitted for assessment for an academic purpose at this or any other academic institution other than in partial fulfilment of the requirements of that stated above.

Signed: *Shruti Sindhi* **Date: 10th Sept 2022**

Acknowledgement

Before I begin the thesis, I would like to take this opportunity to acknowledge the people who have helped in any way, shape, or form in the production of this thesis.

The completion of this study would not have been possible without the expertise of my supervisor Dr. Aqeel Kazmi. I express my sincere gratitude to him for providing invaluable advice, direction, knowledge, and encouragement since the commencement of this thesis. Without his mentorship, it would have been challenging for me to accomplish this body of work. I appreciate the wisdom of his supervision; it helped me to come up with ideas by myself and become relatively independent in completing the task. Moreover, there was never a delay from his side in answering my doubts and questions. It was my luck to study and work under his guidance.

A debt of gratitude is also owed to Dr. Faheem Bukhatwa for offering multiple seminar sessions, clearing doubts, and assisting set temporary milestones to complete all the tasks on time.

I am immensely thankful to my parents for their love, support, and caring. And for educating and preparing me for the future. Also, I thank both my sisters for their understanding and encouragement.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	**Definition**
BVPS	Book Value per Share
CAGR	Compounded Annual Growth rate
CapEx	Capital Expenditure
CF	Cash Flow
DCR	Dividend Coverage ratio
DFL	Degree of Financial Leverage
DPO	Days of Payable Outstanding
EBIT	Earnings Before Interest and Taxes
EBITDA	Earnings Before Interest, Taxes, Depreciation, and Amortization
EBT	Earnings Before Tax
EPS	Earnings Per Share
EV	Enterprise Value
EV/FCF	EV to Free Cash Flow

EV/R	Enterprise Value Multiple
FAT	Fixed Asset Turnover
FCF	Free Cash Flow
NI	Net Income
OCF	Operating cash flow
P/B	Price to book ratio
P/E	Price to Earnings ratio
P/OCF	Price to Operating Cash Flow ratio
P/S	Price to Sales ratio
POCF	High price to cash flow
PTB	Price-to-Book
ROA	Return on Assets
ROC	Rate of Change
ROCE	Return on Capital Employed
ROE	Return on Equity
SBC	Stock-based Compensation
SG&A	Selling, General, and Administration

Abstract

It is weened that there is always some helpful information hiding behind the massive and noisy data that may deliver insight into the financial market. The wobbling nature of the global stock market makes the task of prediction or classification challenging. There is tremendous market risk involved in the stock, bond, or cryptocurrency investment. Often, the risk is dependent on prediction error if not company performance.

Stock price movements are marked by various components like demand and supply, political stability, company earnings, foreign portfolio investment, and more.

The research addresses the problem of discovering a model that can presage the rise and fall of stocks, as it is an arduous job as there may be myriad aspects that can impact share prices. This thesis finds an effective prediction model that yields an accurate prognosis with the most subordinate error percentage for both fundamental and social media data (Twitter). This study reviews PCA (feature extraction model), LightGBM, RFE, Random Forest, Genetic Algorithm (GA), and five more models to get the best feature for the classification of the fundamental data. Later the models are compared to decide which approach is better for developing a robust model. to determine what approach is better for developing a robust model. The model results confirm that LightGBM is the best feature selection model compared to the Genetic Algorithm in the case of Fundamental Data. Likewise, algorithms like CatBoost, Multilayer Perceptron (MLP), Random Forest, and Support Vector Machines (SVM) were implemented and compared to uncover the best predictor for fundamental data. The correctness of the CatBoost prediction algorithm is more than other algorithms. On the other hand, the Random Forest model tends to overfit in the case of financial data associated with stocks.

The dissertation evaluates non-quantifiable data from Twitter to study and portray the relationship between stock-related tweets and stock movement by gauging their sentiment. The LSTM neural network approach evaluates a more accurate yet generalized model that reveals comments of investors or people on the general effects of stock direction.

Ultimately, the study proposes a method or technique to blend sentiment and fundamental analysis that verifies that the overall prediction value increases when the sentiment analysis is incorporated into the job.

Chapter 1. Introduction

1.1. Topic Overview

Data mining and machine learning approaches can be incorporated into business intelligence (BI) systems to help users to make decisions in many real-life applications[1]. Stock is an interesting BI application and has been broadly studied by researchers in different fields. Stocks are complex, dynamic, and unpredictable. Notable behavioural economists hypothesize that the value of the stock is not just correlated to the financial feature. But it is also affected by the public mood and sometimes the overall market performance. People in a generous mood can plausibly make more investments than regular investors. This may improve the overall stock market performance. Mood, the human factor, can be quantified with the aid of the internet. However, measuring the feelings of the investors is not a trivial task. To build a good prediction model or system with high accuracy and performance, one needs access to every possible person's mood pattern. These individuals may include regular stockholders/stock traders, prospective stock investors, venture capitalists, industrialists, philanthropists, and others. Observing all these individuals' mood patterns is impossible as it is a very tedious task, and the sample set may be minimal for good prediction. Also, it may not include swing traders and stock specialists expressing their feelings concerning a stock through a new article, comment, or any other statement on record.

A range of algorithms, from Logistic regression to Support Vectors Machine (SVM), has been used for better investment and management on the trading floor. Equity securities (including standard and preferred stock) are among the most traded securities[2] because these equity securities have attractive returns[3, 4]. However, despite all the attractive returns, equity investment has a high risk due to immense uncertainty and fluctuation in the stock market[5]. In other words, there can be (irrational) overreactions to the bull and bear markets[6]. But all the share investors need to make a persuasive investment decision at the correct juncture[7] and with the precise and proper amount of information[8]. The latency of a few milliseconds between commodities dealings in New York and London appears to be a priceless advantage to some people. Still, it can disadvantage others, depending on the situation. And so, a sound recommendation system with a single or combination of solid machine learning algorithms is required for sentiment analysis and fundamental analysis. The fundamental analysis done for engendering a better investment strategy differs from technical analysis. The main focus of technical analysis is on the market price dynamics and trading volume behaviour to speculate on the forthcoming behaviour of a share or trading floor[9]. The technical approach assumes that the price pattern that has struck in the past will also happen hereafter and so can be employed to augur future cost movements. [10]. It is also a pattern recognition problem[11, 12].

On the contrary, the fundamental analysis applies information like company-specific earnings and potentials to presage future cash flows and the corporation's value to foretell future share price movements. [9, 13]. Even though a technical model is extensively used in stock market prediction, it is seen that these methods are not always successful. There are also a few criticisms of technical analysis that it solely deems transactional data stocks and completely bypasses the fundamental aspects of corporations which might be valuable if the trading floor is in a weak form of efficiency[14].

The only problem is the use of a forecasting algorithm to pick the fads in the stock market prices that contradicts a primary rule in finance known as the Efficient Market Hypothesis (EMH)[15]. The hypothesis declares that asset prices mirror all the general and public information. But there, a direct implication is that it is beyond one's reach to "beat the market" consistently on a risk-adjusted ground as the market price should respond to only new data. This economic hypothesis further insinuates that the whole market would evolve aware of this benefit if somebody were to gain an advantage by analysing historical stock data. As a result, the price of the share would be repaired. Even though it is generally accepted, it is an approvingly controversial and usually debated theory. The researchers have deserted this theory by operating algorithms that can model intricate financial system dynamics.[16]. Ultimately, incorporating quantitative and qualitative information into the commodities exchange analysis can enhance projection ability[17].

The desired results of the stock market analysis can be achieved with the ensemble method, too, as studies reveal that a sole algorithm cannot always solve the problem effectively according to the funding by researchers[18]. And in such cases, ensemble methods lead to better prediction performance. An ensemble method is an approach in which multiple models are developed and merged to produce better results. As per the report by Nanni and Lumini, and Lessmann, the ensemble method performs better than the single AI and statistical methodology. But this research study doesn’t focus on ensemble learning to predict and classify the results[19].

1.2. Research Purpose and Objective

The prime purpose of the research is to investigate stock market (financial) data and learn the effect of sentiment and fundamental analysis on it. The study will also attempt to suggest the best machine learning model or a combination of algorithms for the task. Furthermore, the research aims to reveal the best feature selection method and why feature extraction and feature engineering are not conducted on financial stock-related data. Finally, the conclusive intent is to develop an intelligent model using a combination of algorithms that retains the financial market data and foretells the movement of the stock prices.

The other sub-objective compares the existing work and discovers which model best predicts which type and amount of data. The focus of the research is to give the fund manager, traders, investors, borrowers, and treasures better decision and planning abilities by foretelling the direction of the stock movement and the effect of the people's comments on the stock movement.

1.3. Future Goal

The future target is to create a recommendation system with numerous good features. Being able to categorize stock into different stock types. Other sub-objectives are to get a list of gainer, loser, and most active stocks, which is an outstretch version of whether a share will perform positively or negatively. Another intent is being disposed to inform about the commodities and bond rates of the company along with their stock report. A display of a complete recommendation system concerning a currency type would be a handy feature. The facility can be backed with a currency filter.

Presenting cryptocurrency investment guidance will be a great add-on to the system. Many designated market makers (DMM) believe there is a strong correlation between the price of cryptocurrency and the stock market. Therefore, even though cryptos are digital assets that can go around without the need for central monetary authority.

Efforts can be put into discovering how dependent stocks are on cryptocurrency or cryptocurrency information.

A culminating goal would manage to show a list of climate leader stock. To reiterate, it is the list of companies/ organization or their share indicating their strong ambition to decarbonize.

The future aim in case of sentimental analysis of stock-related textual data is to get the highest accuracy possible. The system estimation of consumers' true preference not only affects consumers' actions but also heavily affect the stock value, its overall performance, and stock return. The objective also includes translating not solely English textual comments into explicit ratings but also comments in all viable languages. The penultimate aim is to set up a dictionary for sentiment analysis, specifically including major financial lexicons in all languages.

The ultimate intent is to make the recommendation system more secure because of the increasing security issue. The goal is also to obtain the ticker symbol without delay and maintain the correct value if there is a company merger or a different exchange of the traded stock.

The thesis consists of 8 chapters, including an Introduction. The following chapter does a literature review. The third chapter presents feature engineering, extraction, and selection technique and which technique is best suited for the situation. The remaining dissertation is structured as follows: Section 4 koreroes different machine algorithms for both fundamental and sentimental modeling of the data and a comparison between them.

Section 5 confers text mining and sentiment analysis concepts and their influence on the stock market. Furthermore, section 6 briefly describes the datasets to predict the required results.

Section 7 evaluates the results fetched after both analyses. Finally, the last section summarizes the study and states the future goals.

Chapter 2. Background

2.1. Related Work

2.1.1. Fundamental Analysis

The machine learning techniques used in the financial and, specifically, the stock domain has improved efficiency by 60-86 percent compared to the past methods[20]. All classification, prediction, and clustering methods can be used for analysis related to stocks [21]. The early work includes the usage of conventional algorithms like Linear Regression[22] and Random Walk Theory (RWT)[23]. Furthermore, prominent researchers have used the Artificial Neural Network (ANN), Genetic Algorithm (GA), or Support Vector Machine (SVM) to herald shifts and modifications in the stock or the FX market[24]. The genetic Algorithm has also been handed-down for theoretical questions in the economic market by Andreoni and Miller (1990)[25], Arthur (1992), and Rust, Palmer, and Miller (1992), and to time series forecasting by Packard (1990) and Meyer and Packard (1992)[24]. Swarm Intelligence Algorithm presented the predicting procedure for the day-to-day exchange rates of the Japanese Yen to the US dollar and the US Dollar against the British Pound[26]. The recommended method included artificial neural network (ANN), evolutionary computing, and clustering technology. Whereas Particle Swarm Intelligence (PSO), a type of Swarm Optimization Algorithm, focuses on recommendations while making investments decision in the stock market[27]. The research by Anthony and Dozier worked on the same line of the idea, but it focused on adapting PSO to dynamic environments[28]. A stock buying-vending alert system was submitted employing a feed-forward neural network anointed NN5[29]. The Hong Kong and Shanghai Banking Corporation (HSBC) holding stock data was used, giving a more than 70% hit rate. Another team of researchers presented a hybrid technique using support vector regression (SVR), Self-Organizing Feature Map (SOFM), and filter-based feature selection to foretell the index value.[30]. Researcher CM Hsu implemented Self Organizing Map (SOM) to make a stock-related prediction[31]. SOM is an unsupervised learning network. An extended version of SOM (Self Organizing Map), Kohonen SOM made a smaller matrix than Backpropagation when data preprocessing was done on the available dataset[32]. Recurrent Neural networks (RNN) can be used to classify and forecast features related to stocks. [33]. RNN uses Backpropagation in learning but has a feedback method in its nodes. And so the models can foretell shares price based on contemporary history[34].

On comparison and analysis, it was learned that deep structured learning algorithms like ANN, bidirectional long short-term memory (BSTM), LSTM, RNN, and stacked long short-term memory (SLSTM) produce low percentage error[21]. Likewise, a radial basis function network, an ANN that encompasses the function as an activation function, can also be incorporated into stock market analysis. [35]. One of the first efforts to project the stock market changes was by expert Kimoto and his colleagues on the Tokyo stock market[36]. Researcher Mizuno and his team members utilized the same data and network. As a result, the neural network was predicted with 63% precision[37]. Nevertheless, when the Neural Network fused with Genetic Algorithms to foretell if Singapore trading would have an uprise or downfall, the precision value turned out to be 81%. A discussion on how data mining approaches can be involved in designing a market capital-specific prediction system for trading firms was released by Aditya Nawani in 2013[38].

Some researchers have used suitable kernels for features from multiple sources to haggle with the problem of foretelling the stock price. The method is called Multiple Kernel Learning [39]. MKL does sentiment analysis and fundamental analysis of the stocks. The study conducted in 2010 to predict the next day’s stock values of 1094 companies traded in the Tehran Stock Market (from 2000 to 2005) was done with the aid of Multilayer Perceptron (MLP) and Elman Neural Network[40]. The prediction error of the model was roughly 1.5%. The SVM algorithm with Cuckoo search optimization has high performance with an accuracy above 80%[41]. However, the Random Forest gave an exactitude value in the scope of 85-95% for long-term predictions of the outcomes on the same data set[42]. Despite the many advantages of SVM, it does have some drawbacks from a functional point of view. The first problem is selecting the kernel function parameters – for Gaussian kernels, the width parameter α and the value of ε in the ε loss insensitive function.[43]. Author Li and Yang used logistic regression for the supervised learning part because it conceded a accuracy of 55.66%[44].

The work of Darmadi Komo[45] implemented and compared two neural network models, Radial Bias Function (RBF) and Backpropagation (also known as multilayer perceptron (MLP)), for stock market prediction. According to Dase R K[46] literature review, speculating the stock index with traditional time series network analysis is intricate compared to the artificial neural network.

Hidden Markov Models (HMM) is broadly utilized for pattern recognition and classification problems for its established expediency for modeling active and dynamic systems. But predicting HMM employing upcoming occurrences is not straightforward. On implementing ANN and HMM models on presage, the stock values of four significant airlines, it was discovered HMM performed better. The Mean Absolute Percentage Error (MAPE) of both the models was less than ten and quite similar[47]. Th Hidden Markov Model is better than ANN because of ANN's inability to explain the model. Repley states, "the design and learning of feed-forward networks are hard". When the fusion of the HMM-ANN-CA algorithm with a weighted model is used, the MAPE value for apple stock is 1.925[47]. The value decreases further when HMM is implemented along with the fuzzy model. The value turns out to be 0.779.

Neural Network (NN) and extended versions of NN models were incorporated to classify if the stock had a downtrend, uptrend, or sideways trend. One vs. One (OAO-NN) and One vs. All (OAA-NN) neural networks were analogized with the traditional neural network on the Stock Exchange of Thailand (SET) for seven years and were used to conclude the outcomes. Further detailed analysis found that the OAA-NN outperformed OAO-NN and Neural Network with an accuracy of approximately 72.50% [48]. Back testing on indicators like RSI, MACD, OAA-NN, and STOCHASTIC also concluded that the One vs. All Network has the best performance of 57.67% return rate for every 35 trades[48]. RSI stood for Relative Strength Index and was proposed by Wilder[49]. Stochastic is nothing but a Stochastic Oscillator Technique. It is utilised to compare the current close price of the stock with its price range over a long time. Last but not least, Gerald Appel proposed Moving Average Convergence/ Divergence (MACD). [50].

The stock market prediction was made with a model employing initial centroid selection optimization for k-means with a Genetic Algorithm (GA). The accuracy of the algorithm was 89.31% on the data supported by Egypt Stock Exchange (EGX) and Egypt for Information Dissemination (EGID)[51].

Exploration by Nelson, Pereira. And de Oliveria showed exhibited that when the LSTM model was trained on 15-minute-interval observances for BOVESPA (Sao Paolo Stock Exchange) stock, it notified an exactness of 53-55% for the subsequent direction price[52].

Monte Carlo is established using the Central Limit Theorem and the Law of Large Numbers. Monte Carlo plays a paramount role in risk analysis. The statistical method can solve complicated mathematical and statistical problems. It is built on the logic of forming a sequence of random numbers.

A great benefit of the MC method is that one can use it to run scenario analysis. Scenario analysis is a process under which one computes risk outcomes for several different model assumptions. It has been a long that Monte Carlo has become an unavoidable ingredient in many quantitative investigations. The method has been employed in finance since the 1960s.

Research by Farid et al. [53] proposed a way to manage risk in stock exchanges using Monte Carlo Simulation (MCS) for Tehran Stock Exchange (TSE). However, the algorithm fails to include tail risk if the historical data doesn't have tail events. Another study examines the capacity of Monte Carlo simulation (MCs) to predict stock market returns in the Amman Stock Exchange (ASE)[54]. The other problem with the MC model is that an inappropriate input leads to wrong simulation results. The Monte Carlo simulation technique has also been applied to forecast future cash flows to improve long-term decisions in real estate[55-57]. From multiple studies, it could be concluded that the quality of the outputs from a Monte Carlo simulation largely depends on the quality of the inputs[58]. However, the current scope of Monte Carlo Methods in finance has expanded to include valuation and analysis of instruments, portfolios, and investments because it thrives in a situation where there is uncertainty. And so it is not used in the current study.

2.1.2. Sentiment Analysis

The primary motivation of sentiment analysis is to discover what other people think [59]. Sentiment Analysis is a study in natural language processing that analyses people’s opinions, sentiments, appraisals, attitudes, evaluations, and emotions concerning organizations, products, services, issues, topics, individuals, events, and their attributes. SA is a type of opinion mining. Expert O’Connor studied tweet sentiment and public opinion [60]. LeBaron foreran the relationship between new articles and stock price exists[61].A naïve Bayes and language model was implemented to predict the forthcoming trends in the stock price[62]. Kloptchenko [63], with other scholars, combined SOMs and prototype matching methods to analyse quarterly reports' quantitative and qualitative information. They also suggested that the textual information in the financial statements contain not only the description of the events but correspondingly explain why they have happened and how long the effect of such events will continue. Back et al., as well[64], used the SOM algorithm to cluster the companies based on the quantitative and qualitative information in the reports.

A sentiment dictionary was also created employing the Harvard IV-4 sentiment dictionary (HVD) and Loughran McDonald economic sentiment dictionary (LMD) to pull out data from the FINET financial news website of Hong Kong[65]. Contrastingly, the expert Wilson et al. [66] used OpinionFinder as a way and measured the tweet sentiment as the ratio of dark-light words in a tweet.

Anxiety, fear, worry, and fear from over 20 million posts on LiveJournal were found by Gilbert et al.[67], and it was discovered that an increase in negative expressions predicts downward pressure on the S&P 500 Index.

The Concept for the Imitation of the Mental Ability of Word Association (CIMAWA) was developed to find the association between words[68]. For the Sentiment Analysis component, researchers Deng and Shangkun[69] used SetiWordNet. Its English version consists of more than 100,000 words. This opinion lexicon is drawn from the WordNet database. The terms used are associated with numerical scores, where each score indicates the yin and yang of the sentiment through the information.

Gilbert et al. [67] used granger causality to detect the linking between the anxiety, fear, and worry expressed through the net and the stock market. The econometric technique checks whether the Anxiety index provides valuable information for estimating the future market price not already contained in the market. Researcher Nguyen and his team member worked on sentiment analysis of the Yahoo Finance Message Board stock market data. In 2015, they proposed a novel feature, ‘topic-sentiment’. It is also known as the Joint sentiment topic method (JST). JST-based and Aspect-based are the two methods to seize the topic sentiment association. They get the mood information of the stock.

2.2. Methodology and Research Strategy

The methodology employed for foretelling the stock price is categorized into fundamental, technical, and traditional time series forecasting. In technical analysis, the stock movement is foreseen utilizing statistics and other forecasting methods established on historical data or data volume. Fundamental analysis work on factors beyond the stock price, such as dividend payments, trading volumes, index trends, industry group trends, and stock volatility[70]. Alternatively, the rudimentary aspects included in the fundamental analysis are the business environment, overall financial performance, economic evidence, and social & political behaviour[71].

Since the present research lies in the information technology and finance domain, the research method used is quantitative. In general, the methodology applied depends upon the dataset's size and complexity. But Data Mining is an interdisciplinary field that moves beyond quantitative research; it also possesses qualitative research in some areas. One can also rephrase qualitative research to find meaningful patterns in large datasets. Software used for the study is Google Collab, Octoparse, and Tableau. The programming language used is Python. Content analysis is unpractised in this research as the aim is to comprehend the sentiment of the communications available on the web and not the quantitative and systematic description. Typically, content and sentiment analysis can be used together, but it is not preferred for the current research. Even though a part of the research does a qualitative study, the prediction and conclusion delineated are definitive. And there is no potential bias in the solution.

The research philosophy reflected in the research is objective, pragmatic, and functionalist. Implementing both fundamental and sentimental approaches give a picture of two different stances, indicating the usage of pragmatic philosophy.

The default research strategy operated for the thesis is experimental research because it plays a paramount role in progressing the field of data mining. Furthermore, this research strategy takes care of forming suitable and high-quality data testbeds, which are essential for the mining concept.

2.3. Financial and Economic Terms and their definitions

2.3.1. Important Financial Terms

1. **ROC**: Rate of Change analogizes the current price with the price n period ago. It is not used for trading; it simply alerts traders that a trend change may be underway. For instance, it indicates if the stock is overbought or oversold.
2. **Consolidation**: In consolidation, a stock or security neither continues nor reverses a more significant price trend. Consolidated stocks typically trade within limited price ranges and offer relatively few trading opportunities until another pattern emerges.
3. **Volume**: Volume indicates the number of shares bought or sold during a specific period or trading day.If you see a stock appreciating on high volume, it's more likely to be a sustainable move. Conversely, if you know a stock that's appreciating on low volume, it could be a dead cat bounce. Logically, when more money is drives a stock price, there is more demand for that stock.
4. **Cat bounce**: As per the economic definition, a dead cat bounce is a slight, brief recovery in the price of a declining stock. For a dead cat bounce to occur, a stock must gap lower

(i.e., dip at the open) by a significant percentage. As a rule of thumb, 5% might be a good number to look for in the process.

5. **Market Bottom**: A market bottom is the most subordinate price traded or published by financial security, commodity, or index within a particular referenced time frame. The time frame can be a year, month, or even an intraday period, but when referenced in financial media or studies, this term refers to a significant low point of interest.
6. **52-week value**: A 52-week high, as the name suggests, is the highest price that the security/ stock has traded over a 52-week, i.e., a year. It is a technical indicator used to analyse the security's current price. The 52-week high is also used to predict future movements as well.
7. **Freeriding investing**: The term freeriding refers to buying shares or other securities in a cash account and then selling them before the purchase has settled. When a trader freerides, they may pay for the stocks using money from the proceeds of the sale instead of cash. It can be avoided by using a margin account.
8. **Margin account:** A margin account is a brokerage account that permits investors to buy securities with borrowed funds, demanding a deposit of cash or assets as collateral to shield the risk of such transactions. A margin account increases purchasing power and allows investors to use someone else's money to increase financial leverage. Margin trading offers tremendous profit potential than traditional trading but also more substantial risks. In addition, purchasing stocks on margin amplifies the effects of losses.
9. **Premarket trading**: Premarket trading is the trading session before the regular trading session starts. The session permits institutional investors and personal traders to trade shares between 4:00 a.m. ET and 9:30 a.m. ET.
10. **Institutional investor**: A company or organization that finances money on behalf of clients or members is called an institutional investor.
11. **Extended trading**: Extended trading hours permit investors to respond to news and events even when closed markets. It is also believed to be a suitable way to trade for people who cannot buy and sell securities during the regular trading session.
12. **Accounting principles**: Commonly acknowledged Accounting Principles are the accounting benchmark adopted by the U.S. Securities and Exchange Commission. There are seven different principles of accounting. Accounting principles are the rules and guidelines companies, and other bodies must follow when reporting financial data.
13. **Hedge fund:** A hedge fund is nothing but a pool of money that levies many financial parameters to generate good returns at lower risk.

14. **Outstanding shares:** Outstanding shares are stocks currently held by all company shareholders. It includes both share blocks and restricted shares.
15. **Financial Solvency**: Financial solvency is the firm's ability to meet its long-term debts and financial obligation.
16. **Buyback:** It is the repurchase of the outstanding share of the company. A company does a buyback of the shares to lessen the number of shares in the market. And eventually boost the value of the remaining shares. It is a way in which a company re-invests in itself. A buy-back of more than 25% of the total paid-up is not approved. The maximum buy-back is 25% in a financial year. There is no increase in the value of earnings per share after buybacks.
17. **Defensive Investor**: A defensive investor is an individual who is unwilling to or unable to set in the time and effort required to be an enterprising investor[72].
18. **SG&A**: It is an abbreviation used to represent Selling, General, and Administration expenses in accounting that are all presented in the income statement. It is also known as the cost of doing business. SG&A does not include research and development (R&D) charges.
19. **Depreciation** is the fixed assets' estimated cutback value for a fiscal year. The financial feature helps conclude how much asset value has already been used or available.
20. **CapEx**: CapEx stands for capital expenditure. It is used to stage out the funds used by the company to acquire, upgrade, and maintain physical assets. The physical assets can also include all sorts of non-consumable assets. The risk escalates when the CapEx soars because it will take a company a relatively long time to recover from the amount. Abnormally higher value insinuates lower stock returns because the market discounts uncertainties in the share prices.
21. **Treasury Stock Method**: The Treasury Stock Method gauges the new shares that unexercised in-the-money warrants can potentially create. The technique expands the number of shares but decreases the earning amount of earning per share. The method presumes that the stock warrants are availed at the beginning of the year or the issued date.
22. **Stock Warrant:** A stock warrant is a contract between the company and investors. The way to access the warrant is through the broker. The warranties allow the investors the right to buy the shares. Companies also issue warrants to raise capital, even by selling the warrants. As a result, stock warrants offer slight protection during a bear market.
23. **Book value:** The financial term book value distinguishes between the establishment's assets and liabilities. It signifies a fair and accurate worth of the company.

24. **Share Dilution:** Share dilution is when a company releases additional stocks, lowering the ownership proportion of a current shareholder. Stock/share dilution is neither good nor bad; it depends more on the financial scenario. Share dilution decreases the value of the shares presently held by the active shareholders. Despite that, the dilution profits the shareholders in the long term. Share dilution occurs when a company wants to acquire another company or wants to raise additional capital.
25. **Dividends:** A stock dividend is financed to the shareholders as additional shares in exchange for cash. Dividends are not taxed until their proprietor sells them. To be eligible to gain access to stock dividends, one should buy the stock minimum of two days before the date of record and still hold the share at the closing date of trading. Dividends are paid monthly, quarterly, or annually depending upon the stock and dividend type. Dividends are usually considered profitable.

2.3.2. Major Features used for Fundamental Analysis Dataset

1. **Revenue Growth**: It is possible to have negative maturing in revenue and yet be profitable. It happens predominantly in the case of Start-ups because their risk/ reward profiles are completely different.
2. **Market cap**: The enterprise's market cap measures the enterprise's worth in the open market. Large-cap companies are companies with a market value of $10 billion and more. Companies increase the market cap by introducing new shares. As a result, the market cap is nothing but the company’s total outstanding shares.
3. **EPS**: EPS stands for Earning Per Share. It is the weight of earnings per outstanding share of the common stock. A stock with an 80 percent and higher rating has the best chance of success. EPS indicates how much the company makes for each share. It is rather valuable when compared against competitor metrics, companies of the same industry, or across a duration of time.
4. **EPS Diluted**: It would evaluate an establishment’s earnings per share if all the available securities (convertible) were converted. It is crucial because it is the value that act as a base value on which the analysts would publish their estimates. Extremely high dilution is the reason for lower EPS. If the company is too heavily diluted, then the spoils of war are reduced for shareholders. Therefore, diluted value furnishes a better assessment of a company's financial condition.

5. **PTB**: PTB is the Price-to-Book proportion. The value helps the investor understand if the company's market price seems reasonable compared to its balance sheet.
6. **Payout Ratio**: The payout ratio is dividends paid to the shareholders in relation to the company's total net income. Coverage of 0 to 35% is a good payout value. A negative payout ratio alludes to the company having to employ of existing cash or raise additional money to pay the dividend. A dividend yield of 2% to 4% is regarded as decisive. On the other hand, a result of more than 4% is a great buy but risky at the same time.
7. **POCF ratio**: A high price to cash flow (POCF) implies the high cost of trading a particular company. But conceptually, a low price/cash ratio signals that the stock value is better. This is because the proportion employs operating cash flow (OCF), which includes back non-cash expenses such as depreciation and amortization to the net income.
8. **10Y Revenue Growth:** It is the establishment's 10 Year Compound Annual Growth Rate. The parameter aids in verifying if the companies are consistent in increasing their earnings over the long run. A CAGR value of 8% to 12% is suitable for a firm with more than ten years of experience in the same business.
9. **Inventory Growth**: It is the investment of the company in its inventory. Slow demand after COVID-19 commotion can lead to inventory overhang. This can lead to recession. High inventory growth means a company has purchased more than it has sold. As per the efficient market view, lower stock prices mean lower inventory. However, this may not always be true. Inventory growth is directly related to stock market price.
10. **Free Cash Flow**: Free Cash Flow can also be represented as FCF. It showcases the company’s ability to generate cash after accounting for capital expenditure needed to maintain or maximize its asset base. In layman’s terms, it is the money left after a company has paid all its expenses, the interest on the loans, and taxes. If the value of the free cash flow falls, then the stock price also falls.
11. **Earning Yield:** Earning yield is the other way to inform investors about their earnings per share. The stock's profits are calculated over 12 months divided by its current (the latest) market price. The stockholder can also perceive if the stock is undervalued or overvalued with the chip-in of earnings yield value in the stock-related information.
12. **Asset Growth:** Assets, in general, signify everything a company or an individual owns; it involves money, security, equipment, and real estate. The assets of a company are listed on a balance sheet. Growing asset value signifies that a company is developing and efficiently generating revenue. But the value of the stock is affected by how the company's asset growth is financed.

13. **Operating Income Growth**: Operating income increases when there is an increase in sales and/or reduction in cost. Operating income is also known as operating profit or recurring profit. Operating income growth is the profit achieved from a business operation. Operating income growth should be evaluated and compared yearly with businesses in the same industry. A weak stock price can affect the operating performance. However, a decent operational income growth shows confidence in companies' business and creates credence among the investors that the stock price will grow.

14. **Effect of forex change on cash**: A company’s cash balance shifts occur due to see-sawing of currency exchange rates. This parameter exists due to the difference between the functional and reporting currency. Exchange rate fluctuation has a bit of bang on the stock returns of individual industries. The favourable or unfavourable exchange rate also conditions the stock price.
15. **Issuance (buybacks) of shares**: The buyback offer of the share remains active only after 15 days of its release. The offer remains open for not more than 30 days from the day of dispatch of the letter to the shareholders. The dispatched letter is the letter of recommendation. After an agreement with all the company members, the buyback duration can remain gumshoe for less than 15 days. A buyback alleviates the financial ratio, improving the stock price. The supply shock increases the stock's value and raises the shareholders' ownership stake.
16. **Graham Net-Net**: Net-net is a technique developed by the economist Benjamin Graham. The value generated using this technique solely depends on the net current assets. Net-net is nothing but the total amount (gross amount) with both standard and supplementary commission deducted. A score of 22.5 is retained as the rule of thumb in Graham’s assumption. Graham’s number is the upper edge value of the amount a defensive investor must pay for the stock.
17. **Cash per share**: Cash per Share is the ratio of available current cash with the company (in hands or company) to the total number of outstanding shares. A consistent or high cash per share indicates no interference with the number of stocks. A high cash per share ratio implies there is a positive movement in the stock price. A decent cash per share infers that there is no share dilution.
18. **Retained earnings (deficit)**: It is the total amount of a company's payments vamoosed behind after paying all the direct costs, indirect costs, income taxes, and dividends to the shareholders. It represents the money invested in new equipment, marketing, R&D, and

more. High retained earnings denote more investment opportunities, and a business is running well. All these parameters lead to high stock prices.

19. **Return on Tangible Assets**: Tangible assets are a company's physical assets or property. It may include equipment, building, inventory, etc. Tangible assets are the main assets. Return on tangible assets defines returns as an annualized percentage. Return on Net Tangible Assets is calculated by subtracting the liabilities, par value of preferred shares, goodwill, patents, and trademarks from the company's total assets. A stock price can negatively or positively affect an investment in physical assets, depending on the status quo.
20. **Company Equity Multiplier**: The company equity multiplier is one of the indicators of risk that calculates the portion of the company's assets financed by stockholder's equity, not the debt. The multiplier value is calculated by dividing total assets by shareholder equity. A higher equity multiplier value exhibits that the present shareholders clasp fewer assets than the current creditors. Conversely, a lower multiplier value means a company is less dependent on debt financing. There is no ideal equity multiplier; its value relies on the sector or industry a firm operates within.
21. **Current Ratio**: Current ratio is obtained by dividing Current Liabilities by the Current Total Assets. It is also known as the liquidity ratio, which quantifies a company's ability to pay short-term obligations or any obligations due within one year. Publicly listed companies in the United States of America delineated a current ratio (median) of 1.94 in 2020. The recent proportion has a strong positive influence on the stock price. If the current ratio increases, the stock price increases, and vice versa. The current ratio of 1.0 means that the company can cover the cost of its liabilities. Any value above 1.0 concludes that the company is financially solvent for a given year. Being financially solvent is good for the financial health of the firm.
22. **Profit Margin:** The profit margin ratio conveys the information on whether a company makes money. A profit margin of 20% and above is viewed as a good option. To obtain the value, one should divide the sales by the net profit. The parameter is a net income and revenue ratio. The higher the profit margin, the healthier is PS ratio of the stock. Profit margin value makes the investor get more share of the company.
23. **Net Profit Margin:** Net profit margin estimates the profit sired as a percentage of revenue. The value is depicted as a percentage or a decimal value. A higher net profit margin implies that the company is doing business efficiently. Net profit margin also has a positive

influence on the stock. It is an important indicator. One can compare two or more companies regardless of size using net profit margin.

24. **Price to Sales Ratio**: The price-to-sales ratio can also be expressed as P/S. A lower P/S value is more attractive for investment. The ratio is measured by dividing the company’s total revenue by the company’s market capitalization over the past 12 months. A considerably good price-to-sale ratio is above the price-to-sale value of the S&P 500. The value should be compared with another industry average to achieve a good benchmark value for price to sale. It also indicates how much an investor is ready to pay per dollar stock sale. P/S ratio is based on sales whereas the P/E ratio is based on earnings. Generally, P/E value is more casted-off then the P/S in financial analysis.
25. **Effective Tax Rate:** The effective tax rate is the portion of earnings a person or an organization pay in taxes. An effective tax rate is different from a tax rate. The advancing tax rate can sink the stocks but not always. Changing tax value makes the market tumble; there is not always clear evidence of the direction of the stock movements. Increasing taxes on individual income, corporate income, and capital gains affect the stocks.
26. **Fixed Asset Turnover**: The efficiency ratio, Fixed Asset Turnover (FAT), is acquired by separating net sales by net fixed assets. One calculates the ratio for a year. FAT specifies how finely or efficiently a business uses fixed assets to make sales—high or low stock prices in any industry result from a change in asset turnover values. For more efficient use of assets, the value of the turnover ratio should be high.
27. **Payout Ratio**: It reveals the portion of earnings a company pays its shareholders in the form of dividends. Dividing the total annual dividends by net income made in the process gives the ratio. A high payout ratio conveys that the company is endowing out a large share of its revenue to common shareholders, ultimately reducing the opportunity for reinvestment and any other future potential opportunities. A high payout ratio means a smaller percentage of reinvestment. A payout ratio above 50% is unstainable.
28. **Payable Turnover**: The payable turnover ratio designates how rapidly a business makes payments to all its creditors and suppliers that extend a line of credit. A high payable turnover value is a sign of high creditworthiness. The decision of turnover ratio of a company can be considered good or bad only in comparison with its competitors. If the liquidity is short-term, in that case, a high payable turnover is a good option. A low payable turnover does not always indicate that a business is struggling to pay its bills. A high payable turnover is a positive cue for potential investors.

29. **Quick Ratio**: Quick ratio proclaims a company’s ability to pay current liabilities without selling its inventory or obtaining additional capital. The quick ratio is regarded as more conventional than the current ratio. The quick ratio should be above 1:1. If the ratio is 1:1, it will head to technical solvency. Even though 1:1 indicates an ideal standard, such a ratio value means the company is not in the state to meet its instant liabilities. A high quick ratio value means better company liquidity and stock investment.
30. **Return on Equity**: Return on equity (ROE) determines the business’s productivity and profitability for owners and investors. It is attained by dividing the net income by the shareholder’s equity. A higher ROE insinuates the company can convert its equity into profits. The position of the company owner is better if the returns on the equity are higher. The amount available on return on equity is accessible to both ordinary shareholders and preferred shares. The higher value of ROE attracts more investors to invest in a company.
31. **Return on Capital Employed**: The financial ratio Return on capital employed (ROCE) is habituated to evaluate a company's profit and capital efficiency. The ratio aids in grasping a company's ability to generate profit from the available profit. A low value of ROCE hints that the capital is not utilized efficiently. High ROCE implies successful company growth, which in turn leads to more investment in the company. And finally, higher prices of the stock.
32. **Return on Assets**: Return on Assets (ROA) is also a financial ratio. The value portrays how much money a company makes in relation to the total assets. 5% of ROA is typically regarded as good, and any value above 20% is excellent. The ROA metric is a percentage that employs a company’s net income and average assets. Return on Assets partially does not influence the price of the stock. From this, it can also be inferred that a high or low value of ROA does not lead to high or low stock prices.
33. **Price to book ratio**: It is also symbolized as P/B. The ratio companies market value to its book value. The value indirectly releases the following information: the business is declining or undervalued. The higher the ratio value is higher the premium is ready to pay for its hard assets. When the price-to-book ratio is less than 1, it is the right time to buy the stock as it is undervalued. Here the book value is the net asset of the firm. Purchasing assets when the book value is low increase the margin of safety for the traders. A lower price-to-book ratio is a good value to make an investment. While trading value should not be less than one; otherwise, the traders must trade the stocks at a value less than the value of their assets.

34. **Net Debt**: Net Debt is a book value whose measure is obtained by summing up all short-term and long-term liabilities and subtracting the current assets from it. It unveils how much cash remains. Net debt values do not bear any interest. If the debt escalates, then the risk increases. Owing to this fact, one can conclude it could be more challenging for a company to pay back all its obligations to the shareholders and bondholders. Stock prices are less affected by bonds. The net debt helps decide if a company is overleveraged or has too much debt.
35. **Debt Growth**: The debt grows when the spending power of the company is more than its receiving power. The concept of debt financing comes into the picture when a company has a high debt growth or debt growth in general. The only advantage of the debt market is that it allows companies to raise funds.
36. **Deposit Liabilities**: Deposit Liabilities are the number of liabilities the banks hold from the depositors. The depositors are the people the bank pay back in the future. As per the accounting principle, deposits are always a current liability. The stock itself is a liability. Therefore, the stock liability is dependent on other liabilities.
37. **CapEx to Depreciation**: The CapEx to Depreciation ratio is nothing but the growing assets. An ever-increasing organization has a high ratio. Rising capital expenditure and falling depreciation value augment the risk. A regular company has a capital expenditure to depreciation ratio of 1. The ratio value is higher in the utility and energy fields. The value is between 1.8 to 2.1. As assets don't directly affect the share's worth, so do the growing assets.
38. **Dividends per share growth**: Dividend growth is computed by dividing the current dividend by the last dividend value and less one from the result. The dividend growth is presented in percentages. If the company is working well and cash flows are revamping, then there is a chance to pay shareholders higher dividends. After the declaration of the dividends value of the stock primarily increases. But the stock price is ultimately reduced to keep the company stable, so the book value per common share is diluted.
39. **Pre-Tax Profit Margin**: The margin value calculates the company's operating efficiency. The ratio states the number of cents made on each dollar before knocking off the tax from the amount. A good margin value varies from industry to industry. A reasonable margin for small businesses is between 7% and 10%. Lower overhead is noticed in the case of retail or food-related companies, as such firms tend to have high overhead costs. A consistently high Pre-tax profit margin hints that the industry is doing fall. A fall in the pre-tax profit

points out the inefficient business model and low pricing power. All these factors, in turn, affect the stock price.

40. **Enterprise value multiple**: Enterprise value multiple (EV/R) is a parameter that estimates the value of the stock. It compares a company's enterprise value to revenue. Market capitalization, the value of debt, minority interest, and preferred shares are summed, and then cash & cash equivalents are subtracted from it to get the enterprise value multiple. If the EV/R value is high, it is an indicator that the profit is declining. But the stock price may not always reflect the fall. The value indirectly gives the company's financial status. They play an important role in investment decisions because they are easy-to-use values for analysis.
41. **Stock-based compensation to Revenue:** Stock-based compensation (SBC) is also known as equity or share-based compensation. It is an exercise in which a company supplement employees' salary and bonuses with shares of ownership in the business. Compensation is commonly granted to employees in the form of stocks or restricted stocks. The SBC is recognized as a non-cash expense on the income statement. The compensation to revenue ratio is the amount of money a company spends on paying its workers to the portion it makes in net sales. Stock-based compensation should not be very high because disbursing too much money on employees can kill a business. And the corporation might not have enough money to reinvest. For all these reasons, the stock-based compensation to revenue ratio should be small to have a high stock price and enough shares to reinvest.
42. **Gross Profit Growth:** It is a salient measure to decide why a company's profits are increasing or decreasing by gazing at the sales, production cost, labour cost, and productivity. Gross profit is different from net income. The value specifies the company's success in generating revenue while keeping the expense as low as possible. Higher gross profit growth will make the stock earn a higher multiple. A riskier stock earns a lower multiple.
43. **Operating cash flow growth:** It represents the amount of money approaching the company from the firm. The operating cash flow growth can be calculated by subtracting a particular year's cash flow from its consecutive next year's cash flow and dividing the result by the former year's cash flow. Investors, creditors, and analysts prefer the operating cash flow ratio to be greater than 1. Firms with high or elevating operational cash flow are usually in good financial health. Increasing operating cash flow encompasses some material knowledge that influences stock market returns. Firms with consistent OCF growth will recover faster if economic crises occur[73].

44. **Free Cash Flow Yield:** The economic solvency Free Cash Flow (FCF) yield measure compares the free cash flow a company is expected to earn in opposition to its market value per share. The ratio gives the investors and stakeholders a better perspective of the company’s performance apart from the widely used P/E ratio. The higher the free cash flow is, the better and healthier the company is in paying the debt and dividends. An FCF ratio greater than 1% indicates that the company has more cash than it spends on capital expenditure. The ratio is measured by holding the free cash flow for each share divided by the current share price.
45. **Days of payable outstanding:** Days payable outstanding (DPO) is a value to figure out how long a company takes to pay its bill and invoices on average. The DPO value is calculated quarterly or on annual terms. A high DPO is always commendable because it implies that a company has some extra cash on hand that can be used as a short-term investment. Days of payable outstanding are the opposite of DSO. The DPO can also increase companies working capital and free cash flow. Direct Public Offering is also abbreviated as DPO. A firm having a low DPO value may also infer that the firm is taking complete advantage of early payment discounts offered by the suppliers. The money saved from the discounts can also be used for investment. So, whether a high or low value of DPO is good or not thoroughly depends upon the situation.
46. **EV to Free cash flow:** A low EV/FCF value indicates that the company is potentially undervalued. A low EV to Free cash flows also means the company can pay back quickly the cost of its acquisition. The ratio is the inverse of the Free Cash Flow Yield. The ratio exhibits and collates the company's total valuation with its ability to generate cash flow. In the case of this ratio, the value of the entire firm is taken into account, wherein in the P/E ratio, only the market price of equity is deemed.
47. **SG&A to Revenue:** The parameter SG&A to revenue is also known as SG&A to sales ratio. At times it is also referred to as the percent-of-sales method. The ratio measures what one gets when dividing the total SG&A costs by the total sales revenue. SG&A value should be 15 to 25 percent of the total sales revenue. Companies often focus on cutting SG&A costs when addressing the business as the ratio of SG&A to sales revenue soars over time. If the ratio is very high, it means that a swing is experienced in the company's financial health. Such a situation will ultimately show an alteration in stock price movement.
48. **3Y Revenue Growth (per share):** 3Y Revenue Growth (per share) is the same as the 3-Year CAGR. It is a three-year compounded annual growth rate (CAGR) of the Company Stock, which will be determined based on the appreciation of the Per Share Price during

the Performance Period, plus any dividends paid on the shares of Company Stock during the Performance Period[74]. In a stable scenario, 5% growth yearly is moderate revenue growth. The strategies to improve revenue growth are investing in the company's employees, reaching out to new customers, and using technology. An increase in revenue growth makes the price move up. It is not always the case; a company can have skyrocketing stocks even when they are not making enough money. The parameter is one of the most explicit measures to understand if the investment (stock price) will grow or decline over time. It may not be the perfect parameter but significantly preferred by financial experts.

49. **5Y Dividend per Share Growth (per share):** This parameter is similar to dividend per share growth. The only difference here is that the average value of 5 years is scanned to discover its effect on the stock price movement. Parameters like this play a prime role in determining the predicting quality of extrapolation. It checks if the extrapolation is linear, conic, or polynomial.
50. **5Y Net Income Growth (per Share):** Net Income Growth is estimated concerning the performance period, the sum of the actual, reported non-GAAP net income growth for a company for every year’s performance period divided by three[75]. The 5Y Net Income Growth calculates the net income growth for five years. Good net income growth is between 0% to 7% and 7% to 14% when the target values are linearly interlinked with the parameter.
51. **5Y Operating CF Growth (per Share):** The ratio computes how a company has grown its operating cash flow over the past years. It is also defined as the CFO's compounded annual growth rate (CAGR) for 5 years. The higher the 5-year operating cash flow growth rate, the better the stock performance.
52. **Book value per share growth:** The proportion is computed as a percentage change. The Book value per share growth depicts the speed at which a company has been growing its book value for each share. Book value per share (BVPS) should be high, or there should be growth because stock prices are perceived as more valuable when BVPS value is high. And ultimately, the stock price increases. Book value per share (BVPS) is the measure of the equity available to shareholders over the number of outstanding shares. Book value per share can be grown when a company operates a portion of its earnings to buy assets or recede its liabilities.
53. **SG&A Expenses Growth:** There are considerable ways to estimate SG&A Expenses. It can be measured as a percentage of sales revenue, a growth rate over a period, or a fixed

dollar value. SG&A represents a company's spending to promote, sell, and deliver products and services and manage daily operations. The SG&A expense growth stops after a company merger or acquisition. SG&A expenses include some fixed costs, which cannot be adjusted much. It is preferred that the SG&A expenses stay low. The expense does not directly affect the stock price movement.

54. **Enterprise Value over EBITDA:** The ratio is commonly employed to compare competitors in the same industry. The ratio gives a fair market value. A high EV/EBITDA value implies that the company is overvalued, which may not always be accurate. The ratio estimates the Enterprise Value (EV) over its Earnings Before Interest, Taxes, Depreciation & Amortization. The ratio is also independent of capital structure. The reciprocal value of the ratio is a measurement to state the company's return on investment. The ratio can determine at what multiple a company is trading. For example, 6x, 8x. It can sometimes be employed for computing a company's target price in the research report. EBITDA is the driver value of the ratio. Higher Growth rates impact the multiple, and Tesla is an excellent example as they trade on such high multiples. A low EV/EBITDA ratio implies that there is scope for potential investment.
55. **Net Cash/Market cap:** The net cash to market cap value notifies about the company's financial stability. The company's ratio is greater than 10%; it is reviewed as financially stable. The ratio is valuable and handy for comparison with competitors in the same industry. An immensely high ratio value hints that the company is not investing enough. The ratio does not directly affect the stock. But investing in companies with higher market caps is better than those with lower market caps because the associated risk is lower in the case of high cap companies.
56. **PRICE_VAR:** The metric represents the deviation of the average stock price from its mean value over a period. In this research, the time duration pondered is 52 weeks. The deal is earned by removing the standard unit cost from the actual unit cost and multiplying it with the exact quantity purchased. The price variance is the volatility indicator in the stock market dataset. In the commodities exchange, they are also known as Bollinger Bands. While opting for security for investment, traders, investors, and all the traders focus on earlier mentioned historical volatilities to find the proximate stake of a potential trade. A highly volatile stock is inherently riskier, but the risk cuts both ways. To maximize gain on a stock, traders exhibit high-risk tolerance.
57. **Other Assets:** In financial trading, other assets include terms like commodities, currency, and bonds. All these assets are financial assets and not tangible assets. Stock is also a

financial asset that can be seen or touched. Bond assets influence the stock market movement. If the bond value burn-down, then the stock price rockets. Commodities are goods used to begin the process and are interchangeable with other goods of the same type. The commodities sway the world economy by affecting prices. The commodity has a higher impact on related stocks. For instance, if the commodity value boosts by 20%, the stock price soars by more than 20%.

58. **Other Liabilities:** In investing, liabilities are the debts that must be paid and goods & services whose commitment must be settled. Recorded liabilities comprise loans account payable, mortgages, warranties, accrued expenses, and deferred revenues. A change in the liabilities concerns the stock price. If all the liabilities are not cleared on time and keep increasing, it makes investment risky.
59. **Weighted Average Shares Growth:** Investors and traders opt for average weight share if they are stuck at a position in a particular stock for a duration. Because the stock prices keep changing continuously, investors examine the weighted average value. The weighted average share growth is not just used to apprehend its influence on stock price movement. But it is also employed for getting insight into portfolio returns, inventory accounting, and valuation. Weighted average measures the varying degree of importance of the numbers in the dataset.
60. **Weighted Average Shares Out (Dil):** The feature displays the weighted average diluted share outstanding value. It is assessed using the weighted average number of shares of common stock and the effect of diluted potential common shares outstanding over a period using the treasury stock method.
61. **Capital Expenditure Coverage Ratios:** It is better to have a high coverage ratio because it would be easier to make interest payments on its debt or dividends. When the dividends and debts are cleared, then the risk on the stock decreases, and there is not much oscillation in the stock price movement.
62. **Dividend paid And Capex Coverage Ratios:** Dividend coverage ratio (DCR) is above 2. A deteriorating DCR or DCR below 1.5 perturbs the shareholders and investors. More fabulous the CapEx budget, the loftier the value of the stock. Therefore, the dividend paid divided by the CapEx coverage ratio should be small for the positive performance of the stock price. The ratio is different from DCR.
63. **Price to operating cash flows ratio:** The P/OCF ratio calculates the value of the cost in comparison to its operating cash flow per share. The operational cash flow includes both depreciation and amortization of the net income. A high P/OCF ratio implies that the

company is trading the stocks at a high price, but it is not generating enough operating cash flow to carry out the process. The value of the ratio depends on the firm, industry, or operation they perform. High P/OCF value is responsible for low or falling share price valuation.

64. **Weighted Average Shares Diluted Growth:** Diluted Weighted Average Share unveils the number of shares for diluted EPS calculation. Extraordinary items are excluded and included from the parameter while finding the Diluted EPS value depending upon the situation. The financial feature is used as a denominator while calculating the diluted EPS. The impact of EPS (Earnings per share) on the stock price movement is seldom inverse. So high EPS means high stock market price. Since diluted EPS is calculated with the aid of diluted average share, its growth is influencing the stock price in opposition.
65. **Weighted average Shares out:** The parameter in the dataset represents the weighted average of outstanding shares. Weighted average share outstanding brings up the calculated share of a company after rejigging the changes in the share capital over a reporting period. Investors get knowledge of the Earnings per share (EPS) with the aid of the weighted average share outstanding. A firm's weighted average share outstanding is not persistent and may change multiple times within a year. The shift in the value can be because of share buyback, new issues, conversion, or other things. The value of each share is conversely associated with the number of shares outstanding when the remaining economic/financial parameters are equal.
66. **EBIT per Revenue:** EBIT stands for Earnings before interest and taxes. EBIT is a measure of a company’s profitability. Operating profit and operating earnings are the same as EBIT. Earnings before interest are obtained by subtracting the tax and interest from the revenue. EBIT per revenue displays the operating margin of a company. The value is expressed as a percentage value and does not regard financial features like capital structure and tax burden. The profit margin values are compared to the company’s previous operating margin. If the operating profit is high, it indicates the profit generated on each dollar of revenue is high. And this conveys a high stock price.
67. **EBT per EBIT:** EBT ad EBIT are not the same. EBT value, in general, is more diminutive than EBIT. However, if a business has no interest expense or interest income, the ratio value equals 1. It was dividing Earning Before Tax (EBT) by Earning Before Interest and Tax (EBIT) expresses the Degree of Financial Leverage (DFL). A high degree of financial leverage steers more increased instability in the company’s earnings. This instability ultimately leads to a more labile stock price. Suppose there is a change in financial leverage

while adjusting the company’s debt capacity. In such a case, an increase in leverage will decrease the stock price.

68. **NI per EBT:** EBT is an abbreviation of Earnings Before Taxes. EBT is obtained by subtracting the expenses (that were incurred to earn the income) from gross income. Income before tax is the same as Earning before tax. While Net Income is the amount of the earnings left after all the taxes have been deducted. Ideally, NI per EBT value cannot be greater than 1. If the difference between the Net Income and Earnings before taxes is slight, it means the Net income is high—the stock price upsurges when there is an increase in the net income. Therefore, NI per EBT ratio close to 1 is preferred for a high stock price.
69. **Free Cash Flow Operating Cash Flow Ratio:** Operating Cash Flow (OCF) involves the cash generated from the operations. While on the contrary Free Cash Flow (FCF) covers all inflows and outflows of the procedure. A high FCF/OCF ratio means a company's financial health is good. The ratio is also operated to understand the quality of investment in the business. The free cash flow value should be greater than the operating cash flow; this implies that the ratio value will be higher than 1 in most cases. Investors and creditors prefer higher ratio values because they can cover the short-term liabilities and still have money left to continue the business and make stock investments. Higher FCF/OCF ratio is responsible for the positive movement of the stocks.

Chapter 3. Data Pre processing

3.1. Data Imputation or Discarding NAN values

Data imputation (imputing for short) is substituting the estimated value in place of consistent or missing fields in the data records. Imputation can be performed only if the baseline variable is absent. When dependent and baseline variables are missing, imputation can be a demanding goal. A simple imputation technique adds one value for a missing data element without defining an explicit model to fill the data. Imputation is employed because missing values cause distortion in the dataset and affect the final model results. The salient reason for using the imputation technique is because the missing data is incompatible with most of the python libraries used in the machine learning models. Imputation of several missing variables can be accomplished with routine multivariate imputation and iterative regression imputation.

Data imputation is straightforward to execute and does not need any data manipulation. The technique should be used when 5%-6% of the data or information is missing at random (MAR). Arbitrary imputation and Frequent Category imputation (mode imputation) are not preferred options in most real-world problems. For example, in the stock market fundamental data, all instances with more than 80-86% of information missing were withdrawn from the dataset. The features that had only had 1-5% of the missing value at random were preserved in the dataset. And the missing values that were present under those features were replaced with the KNN imputation technique. In this case, a single imputation strategy has cooperated. The standard error of estimates is low in such a situation. KNN imputer is implemented in this research because it is more accurate than the mean-median method.

In KNN imputation, the missing data is predicted concerning the mean of the neighbours. KNN is an effective model for imputation. Euclidean distance is the standard and commonly used distance measure in the KNN model. The Euclidean distance measure is aware of the NaN values and does not include them while calculating the distance between other member records. The method is also referred to as nearest neighbour imputation. KNNimputer is a sci-kit-learn class to predict the missing values. K-fold cross-validation is performed in addition to the imputation to overcome the missing values induced by data leakage or connection problems. The KNN imputer uses the feature similarity concept. The drawback of using the KNN imputer is that it is appealingly sensitive to outliers in the data records. Moreover, using this imputer is computationally expensive as KNN works by storing the whole training dataset in the memory.

Even though the Datawig approach is more accurate than KNN, it is not implemented as it imputes a single column at a time and is slow with large datasets.

KNN imputer does not operate well with categorical data. And so, the categorical variables were removed from the stock market dataset.

When listwise deletion of data records is not done, the methods that can be embodied to resolve the missing data problem are Substitution, Hot Deck Imputation, Cold Deck Imputation, Regression Imputation, Stochastic Regression Imputation, and Interpolation and Extrapolation. Interpolation and extrapolation typically work well in longitudinal data. Stock market fundamental data is also a type of longitudinal data. Other commonly used imputation techniques enclose multivariate implementation by chained equation (MICE) and imputation using Datawig.

Data imputation is always conducted; discarding records and parameters from the dataset is not a feasible option. It ushers in a reduction in the dataset size and a situation where lots of important information is lost. It may also lead to incorrect analysis. Therefore, the imputation method with a negligible overall error value is adopted to complete the task. But beyond that, one must experiment with different models and opt for the one that gives the overall best performance.

3.2. Dealing with Outliers

To discover the outliers in the dataset, one needs to view and examine the peaks and valleys in the data records. It is crucial to ensure that the target data make sense and that outliers oblige. Outliers occur because of mistyping, unreasonable values, measurement errors, sampling problems, or unusual conditions. There are two classes of outliers: Global Outliers and Contextual Outliers. Contextual Outliers are also known as Conditional Anomalies.

Stock market data and the present fundamental data employed for prediction comprise outliers. But these stock outliers are massive gains. Such stock outliers appear when companies have a good and consistent product, growth, and leadership. The stock outliers are data objects that deviate significantly from the entire data records. These stock outliers can be considered false negative outliers or false outliers. Therefore, they are not dropped from the data records during analysis.

The abnormally positive or negative returns in bonds, stocks, or any sort of investment do not come from a normal distribution of stock records or information. The returns are chiefly because of the outliers. Quite a similar but vague definition regarding outlier was given by Boudoukh [76] in inspecting the autocorrelation of stock exchange commodities returns. For all the stated reasons, when potential stock outliers are regarded, they improve the model's prediction. The prediction of outliers themselves can also be made to understand the behaviour of the stock data. Moreover, in an excellent paper, Frieman and Laibson (1989) present an alternative model for outliers in stock returns [77].

3.3. Feature Extraction, Feature Engineering, and Feature Selection

When an individual recognizes how the model works, it becomes easy to extract successful features since it is easy to reason out strong and weak elements.

All these feature assortment actions are not just to overcome the overfitting problem. It also tackles Occam's Razor principle and Garbage in Garbage out the situation. Razor's principle states that entities should not proliferate (multiplied) beyond necessity. Models must be simple and explainable. The dataset loses its interpretability and capability when there are too many features. As per the garbage in and garbage out scenario, non-informative features do not give optimal results. In other words, poor quality input will produce poor quality output. If the number of features is too elevated, the model also becomes bulky and takes longer to implement in production.

The feature assortment procedure generally has four key steps Subset Generation, Evaluation of Subset, Stopping Criteria, and Result Validation.

3.3.1. Feature Engineering

Feature Engineering transforms the data by pre-processing it into more meaningful features. Then, new features are formulated using the available data. Domain experts usually perform feature engineering. Feature Engineering is commonly used to enhance classification accuracy. Deep Learning models eliminate the need for feature engineering. It is mainly favoured in supervised learning models. The optimization loop in the feature engineering replaces the low-performing variables with high-performing ones.

Feature Engineering and feature selection process is not entirely exclusive. Complete automation of the feature engineering process is complex. The iterative steps of feature engineering include the following steps: brainstorming features, devising features, selecting features, and evaluating the features. A well-defined problem is required, so the feature engineering methods know when to stop the process and complete the model's training.

Feature Learning or Representation Engineering is an act of assembling the model for feature identification procedure automated. The feature Engineering process gives more control to the data scientists while opting for a feature. It also makes adjustments wherever necessary.

Feature engineering is mandated when high accuracy is not fetched with available features or when there is a crapshoot of overfitting employing the current features. Feature engineering is also abused to enhance the explainability of the model. Speeded-up training and improved data visualization are the other add-ups when feature engineering is implemented.

Normalization is another important sub-objective that should be achieved with feature engineering. It is essential to bring all the variables under a given interval. Feature engineering is believed to be challenging because it requires knowledge of not just machine learning algorithms. But a thorough and detailed domain knowledge as well. Moreover, conducting feature engineering manually can direct bias and errors.

3.3.1.1. Feature Engineering in Fundamental Analysis Data

In fundamental data analysis, feature engineering is not implemented because it has and will make the whole process resource intensive. There will be a high resource consumption because the dimensionality of the data in use is high. And it will make the entire process more complicated and problematic as there are already innumerable financial features. Also, the whole procedure will be time-consuming as feature engineering is an ongoing process that involves iterative testing, adjusting, and refining of resources.

3.3.1.2.Feature Engineering in Sentimental Analysis Data

Feature Engineering is a critical and binding task in text analysis. Raw text data from different sources constructs features. The primary purpose of feature engineering is to attain optimal variables for the job.

In sentimental analysis, feature engineering is the procedure in which domain knowledge of the information is acquired to concoct fields that the machine learning algorithm can operate to discover the answers to the question.

In feature engineering, varied concepts from NLP and linguistics are used to attain the features. A list of reliable and standard feature extraction methods during this feature engineering process includes parsing, PoS tagging, NER (Name Entity Recognition), and BoW (Bag of Words). The advanced technique for the study comprises the Word2Vec method. TF-IDF (Term Frequency-Inverse Document Frequency) is a probability and statistics-based way.

3.3.2. Feature Selection

3.3.3. Feature selection selects the most prominent variables without changing them. It can be contemplated as a search problem on the powerset of the set of unrestricted variables[78, 79]. Feature selection methods are employed to tackle over-fitting, lessen the training time, and simplify the data models to interpret the results better. The variables are picked in the selection process to eradicate the redundant variables and yank the required information. The selection mechanism should ensure that the model's extant overall accuracy and performance are not affected. In addition, the features selection technique makes some more space for storage.

The feature selection mechanism deals with feature generation and feature evaluation. Various selection algorithm addresses the issue differently. The features are selected and evaluated in a way that fits the requirement. The feature evaluation step makes services the evaluation function. There is a range of evaluation functions used for the selection. The processes can be either deterministic or non-deterministic. They can also be probabilistic estimates of a theoretical measure. There is a need to operate the evaluation function because it is a guide for the search process.

Evaluation can be further sorted based on independent and dependent measures. In the case of dependent measures, the features are opted based on the algorithm's performance. Evaluation without the assistance of a machine learning algorithm is accomplished for independent evaluation measures. Distance, Information, Dependency, and Consistency are the four different types of Measures. The principal approach when the greedy search of feature selection should sack is defining the size of the feature set to be appointed. [80]. Langley[81] states that the feature selection algorithm that searches through the space of feature subsets must address four main issues:

- The starting point of the search.
- The organization of the search.
- The evaluation of feature subsets.
- The criterion used to terminate the search.

3.3.2.1.Feature Weights

Feature weight is a concept in which the relative importance of the feature is evaluated. Weights are also assigned to each feature. The process that yields feature value is called feature weight. When correctly and accurately weighted, the most critical variable will have a significant weight value compared to the irrelevant features—feature weighting plays the first role in text categorization. Feature weighting is a part of the feature selection process. Weighted features can be obtained while implementing feature selection methods on the datasets. It is a pre-processing step, while feature weighting is a learning step. In it that the feature selection process is completed at this learning step. Correlated aided Neural Network (CANN) considers feature weight to be a correlation between input and output features[82]. The value of the feature weight is restricted between 0 and 1. A variable is used, or it is not for predicting the results. Algorithms that assign weight to the features (data variables) do not recede the dimensionality of the data. Instead, the magnitude of the consequences determines the degree of effectiveness of the feature. Research by Shankar et al. [83] shows that the feature weight accommodation increases the execution of centroid-based classifiers by 2-5%. Typically, the feature weight is acquired by allocating a continuous pertinence value to each variable by concentrating on the context or domain knowledge[84]. The risk of overfitting can be lessened by removing the noisy features.

3.3.2.2.Consistency Measures

The more the number of features in the consistency measure, the more consistent the hypothesis can be defined. The notion behind this evaluation measure is that the dataset with the selected feature must be compatible with predicting the instances' concept or class value. The consistency measure is dependent on class information[85]. It indicates that no two features can have the same predicting feature value if they have different concept values. In contrast, inconsistency is when two records have the same feature value but belong to another class.

3.3.2.3.Dependency Measures

In feature selection, distinct dependency functions exist, such as mutual, conditional, and joint-mutual information[86]. The dependency measure finds the reliance between two random features and is called the similarity or correlation measure. The dependency measure was defined by Fisher[87], but it was not formally proposed as a feature selection metric. The dependency measure can discern cohesive and distinctive features. The feature dependency score attained using the procedure is a sound source of information to determine how many variables are worth removing from the list of the dataset. The approach can be used in any symbolic clustering algorithm[88]. The prominent feature selection strategy sequential feature selection (stepwise sequential selection) may be combined with dependency measure to find if a more optimal subset feature can be acquired[88]. A version of this dependency metric is used for the feature selection process and can be evaluated in flexible prediction tasks[89].

3.3.2.4.Distance Measures

The concept of this measure is that different instances of the class are far-flung in the instance space. Many unsupervised feature selection techniques employ conventional distances like Euclidean distance to gauge the resemblance between occurrences in the dataset. The distance-based measure fails to reflect the dynamic structure information of the data. The distance measure can be used for subnormal and non-convex fuzzy sets. Sequential Feature Selection (SFS) can be evaluated with inter-class and probabilistic distance measures. There is no one soundest distance measure for a particular application; some distance measure method operates better than others in general, depending upon the characteristics of the data. Probabilistic distance measures are the divergence measure, the Matusita measure, the Mahalanobis distance measure, and the Patrick-Fisher measure[90].Conversely, the used inter-class distance measure is the Minkowski distance measure, city block distance measure, Euclidean distance measure, the Chebychev distance measure, and the nonlinear distance measure[90]. Huang and Yang utilize city block distance measures for feature selection in stock trend prediction[91].

3.3.2.5.Information Measures

Information is established on the information gained from the features. It is based on the concept of mutual information[92]. The information measure is also called an uncertainty measure. Information gain is the distinction between the prior uncertainty and expected subsequent uncertainty[93].

The three notably discussed styles of feature selection types are Wrapper Methods, Filter Methods, and Embedded Methods. Evaluation criteria can differentiate filter and wrapper.

3.3.4. Filtration Method for selecting features

Filter methods select features according to discriminant measures based on the quality of the data, independent of any classification algorithms[91, 94, 95]. In other words, the filter technique works without considering any learning mechanism to get the features. It is further classified into multi-variate and univariate. In the univariate filtering technique, all variables, called components, are reckoned separately. Filter methods rank the features of the feature subset. It uses a descriptive measure, not just the error rate, to determine the feature's benefit. All classes of data mining tasks cannot use the filtered features. The filtering process can be divided into classification, regression, and clustering depending upon the mining tasks. One downside of the method is that they are blind to any interaction or correlation between the variables. The known filtering techniques are ANOVA (Analysis of variances), Pearson correlation, and variance threshold feature selection. The filter method is more instantaneous than the wrapper method because filtering can be employed as a pre-processing step to lessen space dimensionality and overcome overfitting.[94].

3.3.5. Wrapper Method for selecting features

Wrapper methods utilize the predictive accuracy of predetermined classification algorithms (called base classifiers), such as the support vector machine (SVM), as the benchmark for determining and verifying the goodness of a subset of features[96, 97]. Wrapper Methods are better for picking the variables. But they are much slower than the filters in uncovering the subset of features as wrappers contingent on the resource demand of modeling the algorithm. Also, the wrappers are computationally pricey and do not scale reasonably to large datasets. Therefore, the modeling algorithm performance is reviewed as a black box evaluator. The two major drawbacks of this method are the extensive computation time for data with numerous features and the model's ability to overfit when there are not enough data points. A range of blends of the search strategy and modeling algorithms can be applied as a wrapper. But wrappers are only feasible for greedy search strategies and fast modeling algorithms. Wrappers consider the model hypothesis when training and testing the feature space. The known wrapper methods are backward selection, forward selection, and stepwise selection.

Forward selection starts with zero features, and the number of components increases after each iteration. The final model does not count all the features that don't have the needed p-values after each iteration. Backward selection commences with all the variables from the variable subset. In this scenario, the model runs multiple times and then calculates the p-value related to the t-test or the F-test. Here a feature is withdrawn from the set of features post each iteration if the conditions are not met. Stepwise selection is a hybrid of the forward and backward methods. There is no specific sequence for forward or backward movement. Instead, it depends upon the p-value of the current feature. Other techniques used under the wrapper method are Bi-directional, Exhaustive, and Recursive selection.

3.3.6. Embedded Method for selecting features

Embedded features are faster than the wrapper methods. The method performs feature selection during the model's algorithm execution and comprises multinominal logistic regression and its variants. The variety of decision tree algorithms for this method is CART, C4.5, and Random Forest[98]. Lasso and Ridge Regression are two common examples of multinominal logistic regression.

3.3.7. Feature Extraction

The theory of redundancy reduction drives the utilization of feature extraction[99]. Feature extraction is a type of dimensionality reduction process. It is the process of transforming the data into variables that preserve all the information as the original dataset. Feature extraction is widely used in pharmaceutical, oil, medical, and target marketing industries. The feature extraction process involves of selection and construction of the variables.

The feature extraction technique recasts the data into new fields, whereas in the case of the feature selection method, the actual variables are maintained. The work of Virbhadra and L. Rangarajan [100, 101] proposed bi-level dimensionality reduction methods that have combined feature selection and feature extraction methods with the aim of enhancing classification performance.

Feature extraction has four main aspects. They are feature construction, feature search strategy, evaluation assessment method, and last but not least, relevance index or predictive power. Predictive power is the ability to produce testable predictions. The search strategy is also known as feature subset generation. Feature extraction can be categorized into both Linear and Non-Linear forms.

The aggressive feature extraction technique is essential for high learning accuracy in the medical domain. It is central in the fields of pattern recognition. The feature extraction methodology is operated when the data dimensionality in features is very high. Feature extraction can be done depending on features on decision boundaries as well.

Heydorn [102] suggested a feature extraction method by expunging repetitious features where redundancy is defined as a marginal distribution function. The situation with feature extraction for classification is to find the lowest number of variables needed to accomplish the identical classification accuracy as in the original space. [103].

A shortcoming of feature extraction is that the linear combination of the original features is usually not explainable, and the information about how much an original variable contributes is often lost[104]. Feature extraction is also called dimensionality reduction explicitly or feature transformation. Selection from feature engineering selection and extraction depends upon the type of the data type or domain of applications.

3.4. Methods and Algorithms used for Feature Selection

A list of Methods and Algorithms which succeeded and failed during the feature selection process for the stock market are:

3.4.1. Principal Components Analysis (PCA)

Principal Component Analysis (PCA) comprises of mathematical approach that alters several possibly correlated features into a more diminutive number of uncorrelated features. These uncorrelated variables are anointed as principal components. [105]. The PCA extraction method is a technique that helps us overcome dimensionality reduction issues. It enhances the overall performance of the predictive model by reducing redundancy among the data. In PCA analysis, there is a notion that there is no unique variance; the total variance is equal to the common variance. The concept of variance in PCA aids in evaluating how vital each component is to the data records. Principal Component Analysis can facilitate the execution of the machine learning algorithms if the proximate correlation between input fields is explored, and the main component is picked carefully [58]. However, keeping the worth of the principal element very high or low can lead to some errors in classification and cause decrement in classification accuracy. Therefore, the PCA analysis method is preferred in scenarios where there is a diverse number of correlated dimensions. PCA can be operated as a feature extraction method as it can create valuable features. But it cannot be handled as a feature selection technique because all the variables are vital to calculating the new features.

PCA does not make any assumption on the form of the covariance matrix. It is a linear dimensionality reduction algorithm. PCA helps develop algorithms that can be utilized in different fields without prior knowledge of the system. Despite the effectiveness of PCA analysis, it does not have a widespread application in finance. It can also be operated as a backward-looking analytics tool.

Researchers Yu et al. second hand the SVM (Support Vector Machine) model to build a stock selection system and casted-off principal component analysis (PCA) to get low-dimensional and informative financial time series.[106].The data involved for the forecast here is time-stamped data. The prediction output revealed that the stock returns obtained by PCA-SVM were seemingly better than other benchmarks.

3.4.2. Variance Threshold Feature Selection

This feature selection method abolishes all the fields whose variance does not satisfy some threshold value. All the variables with zero variance are eradicated from the list of records. The eliminated features are attributes with the same value in all the subset samples. The presumption that is made in this method is that variables with higher variance consist of more valuable information. The technique is more practical in yanking variables from unsupervised models than the supervised model. The method does not scan the relationship between features or features and target variables; This property is a notable drawback of this filter method; therefore, it was not involved in the stock market analysis process.

3.4.3. Chi-squared Method

The chi-squared method does not take into consideration the interaction between the features. The technique gives the best comparison between categorical variables vs. categorical variables. Thus, it has a broad spectrum of applications in textual data. It is commonly utilized for univariate feature selection for classification.

In this scenario, all the features were retained.

3.4.4. Sequential Feature Selection (SFS)

Sequential feature selection allows backward as well as forward feature search. The methodology is different from RFE and SelectFromModel methods. It is a greedy algorithm that discovers the most exemplary set of features. It does not require coef_ or feature_importance_attribute parameter. In this setup, Sequential Feature Selection (SFS) is embedded K Nearest Neighbour. The section time would get more elevated with the higher number of features and data.

The two components of sequential feature selection are objective and search algorithm. The objective function minimizes the number of overall characteristics of the features. At the same time searching algorithm adds or removes the feature candidate while evaluating the objective function or criterion. The search technique follows only one direction when put in on the dataset. It either increases the number of features or reduces them. The filtering methods are faster than sequential search, except they are not as accurate as SFS. When the Search feature selection completes the modeling, it controls the procedure thoroughly. The desired result can't be achieved with the embedded method. The sequential search method has polynomial complexity, taking advantage of the hill-climbing search strategy. The Sequential Forward Selection method lies under the heuristic search category in data science. Sequential Forward Selection has the best performance when the dataset is small.

The limitation of this method is the paucity of theory to help choose the optimal value of L and R. Here, L represents added features. In contrast, R represents the features that are removed. Sequential Floating Feature Selection (SFFS) is an extended version of the SFS method.

3.4.5. Recursive Feature Elimination (RFE)

This wrapper method selects the attributes by repeatedly assessing a smaller and smaller set of variables. First, the less critical features are drawn out one at a time in each iteration. Then, the variables are dropped until an optimal number of features are procured.

The problem with the Recursive Feature Elimination method is that it is not always known how many optimal features to uncover in advance. This method is computationally expensive; therefore, one should pre-process the data to the maximum before applying the elimination technique. The wrapper-style feature selection algorithm also makes use of filter-based feature selection internally. Picking the algorithm for feature selection is an essential configuration option in the RFE method. The number of variables and hyperparameters of the algorithm can be traversed. However, the method's performance is not solely dependent on these parameters to work well. RFE generally uses an SVM model with a linear kernel to allocate a weighted score (feature relevance score) to every variable[107]. The key benefit of employing the RFE algorithm is that it directly operates with the accuracy measure of a shared classifier based on the feature importance factor to discover the optimal variable subset [108].

In the research proposed by Weng et al., the recursive feature elimination method (RFE) was used to select features for the next day's opening price movement[109]. Different research used Recursive Feature Selection (RFE) to uncover the pertinent technical indicators for Infosys and Reliance Stocks.[110]. The Recursive Feature Selection technique was the backbone for predicting the intra-day stock market movement using RNN (Recurrent Neural Network)[111]. To foretell short-term stock market price trends with a deep learning system, RFE was leveraged to make sure all the selected features were compelling[112]. Last but not least, research by Xu et al.[113] in a trend prediction algorithm, devoted two recursive feature elimination (RFE) methods, RF-RFE and SVM-RFE to stock price variable picking.

3.4.6. Lasso Method (Least Absolute Shrinkage and Selection Operator)

3.4.6.1 Lasso Method

The Least Absolute Shrinkage and Selection Operator is responsible for Feature Selection and Regularization tasks. This method allows interaction among features and encompasses its feature selection method.

The method makes lots of features' weight become zero. The regularizer shrunk the data values towards the central point of the mean. As a result, the Lasso curtails the total sum of the coefficients (L1 regularization).

In contrast, the Ridge regression cut down the squared sum of coefficients (L2 regularization[114]. The model involves the concept of shrinkage. The method improves the forecast accuracy as shrinking and eradicating the coefficients decrease the variance without a significant bias boost. It is beneficial when there are few observations and features [115]. The method also aids in improving the model's interpretability by eliminating irrelevant features not associated with the response variables. The technique eventually reduces overfitting. One of the limitations of the Lasso method is that if the variables are grouped, then LASSO tends to select one feature from each group and ignore the others. In the case of the small-n-large-p dataset, the Lasso selects at most n variables before it saturates. Finally, the lasso model shrinks the meagre contributed features to nil. Nevertheless, in the same situation, OLS (Ordinary Least Square regression) and Ridge Regression may sometimes fail to assess the exact zilch coefficients[114]. To overcome the limitation of the LASSO model, Elastic Net can be implemented[116].

3.4.6.2 LassoCV

For top financial feature selection, both Lasso and LassoCV methods are executed. However, the lasso is linear, while LassoCV uses an iterative model. The best model from the iteration of LassoCV is done by employing cross-validation. In the LassoCV model, iterative fitting is accomplished along a regularization path. The cross-validation splitting strategy should be done appropriately to get the finest and good quality results. Sometimes unnecessary duplication occurs while using the LassoCV method. The number and name of parameters can also be attained with the Lasso model.

The lasso optimization function is different for mono and multi-outputs.

The pipeline function is habituated to assemble many cross-validation steps while setting different parameters. There is also the incorporation of GridSearchCV functions while uncovering essential features. The function helps to loop through predefined parameters and fits the estimator on the training set. The process helps select the best parameters from the listed hyperparameters.

3.4.7. Logistic Regression

Logistic Regression can be applied for selecting fields as well. The algorithm requires a small number of restrictive assumptions. The Logistic Regression (LR) algorithm can even be employed when the data allotment is not usual, or the group sizes are disparate[117]. Logistic Regression shows good efficacy for soft classification. The logistic Regression Model is highly productive and efficient in foretelling commodities exchange trends. Many researchers have studied the stock market using Logistic Regression (LR), but not much work was found in the case of the feature selection process. Sulin Pang used Logistic Regression Model to predict the tendency of the stock price in 2004[59, 118, 119].

In this research. L2 regularization in place of L1 because the lbfgs function supports only the ‘L2’ or ‘none’ penalty. The regularization technique is treated to tune the model by counting the penalty and adding to the error function. Regularization can be employed to train the models and prevent the algorithm from overfitting the training dataset.

3.4.8. Tree based Model

In a tree-based selection model, each decision tree can ascertain feature importance using node impurities. It is a type of embedded method. In trees, binary decisions are the number of required splits. Binary decisions take value based on a particular feature and split it. Multiple binary decisions together form a decision tree. Tree-based selection models are suitable for managing features where a few splits can extract the significant signal. Tree-based models can naturally handle feature interaction. More the number of partitions needed to capture the feature; comparatively deeper trees are required to complete the task. Deeper trees with many leaves can lead to a higher risk of overfitting. Decision tree-like C4.5[120] are frequently used as embedded methods as they intrinsically perform feature selection on each node. The advantage of using tree-based model is that it accommodates both categorical and numerical parameters; there is no need for a separate method or model for categorical variables. Computational speed is moderate when using the tree-based model for feature selection. The model is well suited for large datasets and comparatively needs less data preparation time.

One can also train the models in a different order to recede the number of layers and fetch the optimal variables simultaneously. The tree-based model's most relevant variables are at the top of the tree hierarchy, as the features at the top have high information gain. The variable (node) importance in a tree-based model depends upon the impurities of the yes-no bucket of the dataset. Generally, the features selected for training have a loftier reputation than the mean importance of each variable by default. But this threshold value can be adjusted if necessary.

A deep tree doesn’t always indicate the tree will split on the desired feature. Sometimes unnecessary split can occur, dividing uncalled features while capturing the noise. The problem with a tree-based model is that they sometimes adopt a non-redundant feature set. Another issue is that the trees give equal importance to correlated features. Tree-based models, decision trees, and random forests give proclivity to features with high cardinality (features with more values). If the dataset holds several correlated features, then it is better to implement the tree model in a way that features are selected recursively rather than all together at once.

Tree based model is excessively used to discover essential pixels with a parallel forest of trees.

3.4.9. LightGBM Model

LightGBM is a type of histogram-based algorithm[121-123]. The gradient boosting framework LightGBM applies a tree-based algorithm. LightGBM obliges with classification problems. In the LightGBM model, trees grow leaf-wise. Optimization features of the model incorporate nifty dismemberment of data parallelism, eigenvalues, and feature parallelism. It also lessens the communication overhead and time complexity between the data[124].

LightGBM braces categorical features, but input values' data type should be integers, not strings. The model can handle continuous feature values. The continuous variables are bucketed into discrete bins.

The model has low memory consumption, high accuracy, and distributed support. It also can process an enormous amount of data rapidly. As a result, the communication cost in distributed learning is low. But the model is prone to overfitting if the dataset is small. LightGBM is almost seven times faster than the XGBoost model. Therefore, it is favoured over XGBoost Model in the feature selection process.

The accuracy of the model is better than other boosting algorithms. It supports GPU learning too.

3.4.10. XGBoost Model

XGBoost, as a decision tree promotion model, connects several trees to form a robust classifier[124]. The tree is grown by iteratively adding a new tree. The trees grow depth-wise. The XGBoost model consists of a regularization term in the target function that avoids overfitting. The model is good with tree pruning and supports parallelization.

Research by Yun et al. [125] submitted a hybrid GA-XGBoost prediction system with an improved feature engineering process, including feature set expansion, data preparation, and optimal field set selection while employing the GA-XGBoost (a hybrid algorithm). This algorithm assists in foretelling stock price direction. Another work by Han and Kim [126] suggested an N-Period Min-Max (NPMM) labelling that labels the data using XGBoost to overcome the minor price change sensitivity in the Nasdaq Stock Market. The XGBoost Model was also blended with Discrete Wavelet Transform (DWT), Principal Component Analysis (PCA), and a Multi-Objective Optimization Genetic Algorithm (MOO-GA) to attain high returns with minimum risk[126]. In addition, the cumulative abnormal return of stocks following earnings was predicted using XGBoost (optimized by a genetic algorithm)[127]. Finally, the research by Zolotareva [128] contends that the XGBoost algorithm aids with long-term investment decisions while handling imbalanced datasets and contradicting labels.

3.4.11. Select K Best Method

SelectKBest method works by adopting the features according to the k highest score. The technique selects those features from the dataset that contributes most to the target variable. Picking the most optimal feature is vital while preparing a large dataset for training. The Select K Best is a module in the sci-kit learn library that reduces training time, accuracy, and overfitting. The score of the best variables is calculated and established on univariate statistical analysis (analysis of variable one by one). The method uses a metric; the user needs to provide only a score function. SelectKBest includes these score functions needed to get the solution. If the SelectKBest method is not executed carefully, it may throw out many features for the wrong reasons. One can access the number and name of features seen in the fit with the Select K Best Method. The score and p-value of the variables can be attained as well.

This financial data study incorporates Mutual Information as a feature selector. Mutual Information (MI) is operated because it provides a fast solution. It is also model neutral, which means it can be used with different ML models. Mutual information regression is used instead of correlation analysis as one can understand how much knowledge one can gain of a particular variable by knowing the value of another variable.

3.4.12. Genetic Algorithm

A Genetic Algorithm (GA) is usually devoted to search and optimization problems. The Algorithm does not require ample information. A randomized search algorithm finds the solution by creating a replica of a natural selection process. The Algorithm is stimulated by the human genetic activity of passing genes from one generation to another. The nature-inspired selection process comprises selection, cross-over, and mutation. The Algorithm employs an iterative approach to devise the best result out of the subset of multiple solutions. The Algorithm offers optimal results even in noisy environments. Genetic Algorithms can also be employed to enhance the implementation of Deep Learning Algorithms. The Genetic Algorithm (GA) contains population, chromosome, phenotype, and fitness function parameters. The Algorithm forages from chromosomes or population point of view but not a single point.

The heuristic algorithm can be used for np problems, game theory, code-breaking, and more. The application of genetic Algorithms in real-time system controls is limited as a consequence of convergence and random solutions[129]. The Genetic Algorithm (GA) can be parallelized and efficiently work on continuous or discrete problems. GA is probabilistic, time-dependent nonlinear, and non-stationary. The Algorithm also braces multiple objective optimizations. Because of genetic algorithms' computational complexity and time-consuming nature, it is not favoured in many situations. The Algorithm also belongs to the family of evolutionary Algorithms[130, 131]. A study by Chung and Shin proposed a hybrid approach integrating long short-term memory (LSTM) network and a genetic Algorithm (GA) [132]. Genetic algorithm aid in optimizing the stock market prediction. The market prediction was made on Korea Stock Price Index (KOSPI) data. Another study by Schoreels, Logan, and Garibaldi displays that they developed a system that uses a simple genetic Algorithm to establish an optimized trading method for each agent, with their trading judgments set on a range of technical indicators (used for creating trading cues)[133]. The genetic Algorithm plays a meaningful role in handling agent reproduction.

Genetic algorithm is unable to solve dynamic problems.

The genetic algorithm is not just employed for the feature selection process. It is utilized otherwise as well. For example, researcher Lin and Cao used Genetic Algorithm (GA) to select a sub-domain and near-optimal value for stock market data[134]. Likewise, Allen and Karjalainen[24] designed a system that utilized the genetic algorithm to form compound trading rules on which ground trading decisions are made.

3.4.13. Why is Cuckoo Search Optimization algorithm not employed?

From the literature review of influence of different machine learning algorithms on stock market data, it was clocked that cuckoo search optimization was executed by numerous stock market and financial industry experts. Therefore, the cuckoo search optimization model can be deemed an expanded version of the genetic algorithm, as it lays out better answers[135]. Moreover, the cuckoo search optimization algorithm can be used for feature selection as the selection process can be broadly expounded as an optimization problem.

Cuckoo search optimization is a nature-based metaheuristic algorithm built by Xin-She Yang and Suash Deb[136]. It was observed that the bee algorithm, the firefly algorithm, the cockroach optimization, and the cuckoo search algorithm also belong to the same type of algorithms. [137, 138]. Layeb[139] introduced a model inspired by blending quantum and cuckoo search to solve the knapsack issue. Another study displayed a new algorithm which was a blend of Cuckoo Search (CS) And Particle Swarm Intelligence (PSO), to cure the defect of PSO [140].

Cuckoo search has three rules that need to be followed to complete the task:

- The cuckoo lays one egg at a time and places it in an aimlessly picked nest, as they don’t build a nest of their own.
- The most suitable nests with high-quality eggs (answers) are preserved for the next generations.
- The number of available and unrestricted host nests n is fixed, and a host can locate an alien egg with a probability pa ∈ [0, 1].

In such a circumstance, the host bird can either throw the egg away or abandon the nest to build an entirely new nest in a new spot when it recognizes the alien egg. In addition, cuckoos can make their egg look similar to the eggs of other species, which occasionally makes it challenging for the host to identify the foreign egg.

A modified Cuckoo Search (CS) method employs the memory-based approach. The mechanism facilitates the algorithm to remember the solution behaviour. The features triggered the improvements during the search operation and held a tally of each solution in the population[141].

Cuckoo search has a slower convergence (finding a stable point at the end of the sequence); because of this reason, it is not a feature selection algorithm of fundamental stock market data. The algorithm also has a slow search speed. This intelligent algorithm also has weak local search ability. In a stock market-related prediction, the results are significant, not how the solution was obtained or the path to the solution. Even the modified cuckoo search optimization algorithm is not a good option for the stock market feature selection process as the computational cost of memory storage will be very high.

Finally, the question of which feature should be deemed as applicable and inapplicable in such a multiple inference task is, in general, still a more complex problem. Even though the issue of feature assortment is addressed here, it remains open in the case of another domain.

Chapter 4. Algorithms used for Modeling stock data

4.1. Fundamental Analysis of the Stock Data

Research by D. Venugopal Setty[142] critiqued the application of data mining approaches to the trading floor's performance and concluded a gap between robust storage and retrieval methods. Therefore, some sort of technological leapfrog is essential to structure and prioritize information. This leapfrog can be achieved with data mining and machine learning tools.

To discover a model with limited complexity and a good generalization ability of previously unseen observations, noise (NaN) can be added during the training procedure[143].

4.1.1. Hyperparameter Tuning (Optimization)

Hyperparameters are unlike internal model parameters. They play an essential role in achieving good accuracy of machine learning algorithms[144]. They are the controllers of the learning approach. Conversely, the values of other parameters are learned in machine learning algorithms. Therefore, Hyperparameter tuning is paramount for controlling a machine learning model's behaviour and making it less prone to errors. However, Hyperparameter optimization is a strenuous process because one cannot mostly transcribe the actual mathematical formula for the functions they are optimizing. Therefore, Hyperparameter tuning should be done attentively and vigilantly to avoid overfitting the model. The hyperparameters are sometimes also known as nuisance parameters[145].Hyperparameters minimize the loss function.

Optimal hyperparameters settings hinge on the dataset. Hyperparameter tuning is a meta-optimization task[145]. Machine learning experts have endeavoured to auto-tune the process. The known optimization methods are grid search, random search, and smart hyperparameter tuning. Smart Hyperparameter tuning should not be employed when the wall clock time is the goal. Furthermore, smart tuning is not parallelizable.

4.1.2. Random Forest Model

Noises in the tree trigger the tree to grow wholly and unexpectedly differently. Random Forest confounds the problem by training numerous decision trees on a range of sub-space. It is achieved at the cost of a bit high bias. From this, it can be concluded that no tree in the forest has access to the entire training data. In a random forest, the data records are recursively divided into partitions. Researchers Saha and Khaidem[42] also estimated the Out of Bag (OOB) error while auguring the direction of movement of stock market values while implementing the random forest model. This groundwork does not focus on the Out of Bag (OOB) error. The model is used in numerous domains with scope for machine learning applications. It is a reliable model in the financial area of implementation despite neural networks' wide usage in economic prediction. It is an ensemble classifier with other advantages, including the models' ability to interact between features and its robustness to outlier values of the features[146]. The random forest model was developed by Breiman[147]. Trees with high depth tend to overfit. In addition, the possibility of overfitting in training sets increases more when highly irregular patterns exist.

Scientist Creamer and Freund [148] casted-off the Random Forest Regression technique for envisaging the performance and quantifying the corporate governance risk in the case of the Latin American Market. Concomitantly, Lariviere and Poel [149] used the Random Forest Regression algorithm to investigate both customer retention and profitability outcomes. But this research focuses on applying a random forest classifier.

4.1.3. CatBoost Model

CatBoost algorithm (Russian developed machine learning algorithm) is a type of gradient boosting decision tree (GBDT) algorithm that is good at handling categorical features well. The CatBoost algorithm is open source and allows both CPU and GPU implementation. GPUs trains the model 3X times faster than the CPUs. Graphics processing unit speeds up the process for deep learning. When the training is performed using GPU, the optimal tree depth is limited to 8. The CatBoost algorithm uses efficient strategy and help reduce the overfitting problem by using the entire dataset for training. The algorithm uses a novel schema for calculating leaf values which allows one to use several permutations without any problem. CatBoost is also a greedy algorithm.

CatBoost utilizes an oblivious tree as base learners. In oblivious trees, the same splitting criteria (splitting feature and splitting threshold) are operated across an entire level of the tree[150, 151]. The CatBoost algorithm also carries out feature discretization in a fixed number of bins to recede memory usage[152]. CatBoost algorithm uses ordered or plain boosting mode to complete the task. The plain mode is the default mode and is implemented in the research to complete the job.

The advantage of the CatBoost algorithm over other algorithms is that it can handle categorial features directly without any encoding. The model supports all numerical, categorical, text, and embedding features. Furthermore, the algorithm can convert categorical variables into numbers in the execution process. This can be done using statistics on a combination of categorical and numerical features. Although the CatBoost was modelled in 2017, it is much easier to use and user-friendly compared to other available models. The demand for comprehensive hyper-parameter tuning is less in CatBoost model because it comprises diverse parametrical conditions. The parameters include bagging temperature, fold size, learning rate, regularization, the number of trees, tree depth, and more.

The CatBoost model has a remarkable ability to fight gradient bias.

4.1.4. Support Vector Classifier

A supervised machine learning algorithm, a Support Vector Classifier, is one with a large amount of computation. In Support Vector Classifier (SVC), the data are supplied and computed in batches. It is slow to achieve classification using SVC if the problem involves large datasets. Therefore, the support vector classifier is not preferred in most real-time classification problems.

The Support Vector Classifier is not the same as the Support Vector Machine. In SVC, the hyperplane classifies the dataset linearly, whereas, with the SVM algorithm, the dataset is separated by a nonlinear approach. Support Vector Machine Algorithm has gained popularity because of its capacity to haggle with intricate nonlinear patterns. [153, 154]. Support Vector Machine can be operated for data analysis and pattern recognition. SVM algorithms are very resilient to overfitting problems. SVMs need a relatively high-dimensional model.

Instead of predicting the stock market indices using SVM, Chen and Hao (2017) submitted a hybrid framework to foretell the market indices with feature-weighted SVM and feature-weighted K-nearest[155].

4.1.5. Multilayer perceptron (MLP)

Feedforward NN is a typical architecture used in trading floor prediction[156]. ANN algorithm has the potential to learn and generalize from the non-linear data trend and is well suited for problems from the domain, such as stock market prediction. Many researchers believe that Bio-inspired algorithms like ANNs, evolutionary computation, swarm intelligence, artificial immune systems, and fuzzy systems[157] have shown tremendous success and improved research domain results.

MLP is a category of Artificial Neural network that incorporates backpropagation for training. All the classes in the model are separated via Hyperplanes. They are non-linear neural network models that approximate almost any function with higher accuracy[158]. Multi-layer perceptron utilizes distributed learning to complete the task.

For the research work of Mantri[158], it was established that SVM performs well compared to the MLP model when predicting the stock index trend over four years for the Bombay Stock Exchange (Sensex). Moreover, SVM was considered a more promising alternative for time series forecasting as it presented smaller MSE for all 4 years.

4.1.6. Why LSTM method is not chosen for analysis

Long-short-term memory (LSTM) is an artificial neural network that can be used for classification and regression. LSTM, the deep learning algorithm, has a feedback connection. Therefore, the LSTM model can be applied for classification problems that may not be time series data. Nonetheless, technology allows us to come up with a solution. Still, it does not make sense to build a time-dependent model with cross-sectional data to perceive an association. Another reason for not using LSTM for classification is its memory limitation. A lot of memory is utilized for storing a single observation over a very long number of input time steps.

4.2. Sentiment Analysis of the Stock Data

Mining opinion is a slick affair. Opinion analysis work and controls the perception of the product and understanding of the market through the lens of sentiment data. Opinions differ and can be categorized as direct, comparative, explicit, and implicit. Therefore, one should apprehend how to apply sentiment analysis in different business operations before utilizing a sentiment analysis algorithm. The type of sentiment analysis is fine-grained, emotion detection, aspect-based, and intent analysis.

A range of algorithms can be used for sentiment analysis. It includes Naive Bayes, Support Vector Machine (SVM), Decision Tree, XGBoost, K-Nearest Neighbour, and regression models. Commonly, deep learning models are favoured over traditional models because they give better results. A list of networks used in such cases is CNN (Convoluted Neural Network), DNN (Deep Neural Network), and RNN (Recurrent Neural Network). Hybrid Sentiment Analysis algorithms (which involve fusing rule-based and automated machine learning algorithms) propose the potency of machine learning with the flexibility of customization. This modern method is regarded to be a more efficient approach than others for sentiment analysis.

Keras, fasText, and DistilBERT (smaller but faster version of BERT) are the most known and used deep learning frameworks for sentiment analysis.

Businesses depend on sentiment analysis o gain a deeper understanding of the consumer mindset.

4.2.1. LSTM (Long short-term memory) Model

LSTM has feedback connections. Training the LSTM (long-short-term memory) neural network is done in TensorFlow. LSTM is a type of recurrent network that can solve several problems and a viral algorithm for time-series stock models. The Long short-term memory (LSTM) cell is a specially developed unit that helps an RNN memorize the long-term context better[159, 160]. Finally, the dropout regularization technique can be used in LSTM to prevent neural networks from overfitting**.** In this process, neurons are randomly disabled. As a result, their connections are also disabled. Gates controls the memorization process in LSTM to handle the increase or deletion of the information.

LSTM can be used to build a trend prediction model as well. LSTM's ability to harness long-term dependency makes it suitable for diverse applications like recommendation systems, stock market prediction, etc.

The research work by Xin et al.[161] proposed the Long Short-Term Memory (LSTM) model that employs the opinion analysis of social media to herald the real-time stock movement of cryptocurrency. When LSTM converges with a Twitter sentiment analysis, it surpasses other machine learning models such as Support Vector Machine in foretelling the stock price movement.[161]. LSTMs are picked for implementation in numerous stock market research because of the model's advantages in scrutinizing the relationship among stock time-series data via its memory function. Stock market closing price predictions could be made by adopting investors' sentiment empirical modal decomposition (EMD) and a revised long short-term memory (LSTM) model with the attention mechanism, according to work by Jin and Yang[162]. Whereas work by Guo coalesced quantitative sentiment score with stock recorded stock essential variables, using LSTM neural network to presage prospective stock close price and stock return[163]. In another research by Achkar et al.[164], they analogize two kinds of neural networks, Multilayer perceptron (MLP) and LSTM, and the outcome demonstrates that the percentage error in LSTM is small.

And from the research and study of sentiment analysis and related operations, it can be concluded that LSTM is the best fit for the task.

Chapter 5. Text Analysis and Sentimental Analysis

5.1. Text Mining

There is no distinction between text mining and text analysis. Both words direct to the same method of accumulating valuable insights from various unstructured and semi-structured text sources like email, social media feeds, and more.

Text mining sometimes leads to information overload, but its muscles lie in the optimization of machine learning algorithms.

5.2. Text Analysis

Text Analysis and Sentiment Analysis are not the same things. Text analysis must be accomplished before performing sentiment analysis. Text analysis is recognizing unstructured text, extracting relevant information, and transforming it to make better business decisions. It can process semi-structured data, too, and is employed in instances where there is a requirement to process extensive volume data. Text analysis, text mining, and text analytics are frequently stated interchangeably. But there is a difference between them. Text analysis offers qualitative outcomes, whereas text analytics provides quantitative results. In layman's terms, text analysis is about pulling models or schemas from recording/speech of human speech.

There are word spot and manual rules in text analysis. The word spotting is a straightforward approach, but it works well only with small datasets. The manual rule approach is relatively similar to the word spotting technique despite the complex matching pattern. All the functional word identification rules have good performance, though they are incompetent to identify the mood and behaviour of the situation/problem from the contextual data. And so, sentiment analysis comes into the picture.

Text Analysis (TA) has become paramount for business processes because of it, and the reliance of the business on quantitative survey data has sunken.

The benefits of text analysis are that it is scalable, consistent, and can analyse data in real time. However, the usability of the text model is not high. The challenges encountered in the case of text analysis are the ambiguity of human language and multi-lingual scenarios.

Types of text analysis techniques comprise text classification, text extraction, topic modeling (identifying theme or group-related keywords), and PII redaction (personally identifiable information), Concordance, Word Sense Disambiguation. Collocation can be recognized, too, with the aid of text analysis.

PII redaction support protects the privacy of individuals and concedes with local laws and regulations. Personally identifiable information (PII) redaction detects and removes personal information like names, addresses, ages, nationalities, account numbers, and more. Furthermore, the collocation technique identifies words that commonly co-occur. Collocation discovers the hidden semantic structures and enhances the acuities by counting bigrams and trigrams as one word.

Concordance offers the context surrounding the argument token. The step can decode the obscurity of human language to a specific capacity. It provides a quick understanding of how users are using a word.

Text analysis software proceeds on the guide of Deep Learning and Natural Language Processing (NLP). Text analysis needs to design and device customized text mining pipelines to achieve high accuracy results in a particular domain.

Text analysis has two broad types; linguistic tradition and sociological tradition[165]. The work by Tedlock[166] demonstrated the elucidative power linguistic methods bring to text analysis. The study by Jehn and Doucet exemplifies the rich assortment of qualitative and quantitative methods that are now available for text analysis[167].

5.2.1. NLP (Natural Language Processing)

Natural Language Processing (NLP) is a subset of Text Analysis. Text analysis aims to emanate acuities from the text without considering the semantics, whereas NLP understands the linguistics incorporated and the context behind the text.

Natural Language Processing aspires to construct appliances that can also react to voice data (in a way like humans). In addition, NLP is responsible for language translation in computer programs. It is also typically operated for text mining and automated question answering. Sentiment Analysis established on NLP can identify the textual information's feelings, opinions, or beliefs. NLP incorporates computational linguistics—rule-based modeling of human language—with statistical, machine learning, and deep learning models to process human language (text or voice) and comprehend the data's sentiment or intent.

As the name suggests, NLP is a computer reading language, whereas NLU is a computer understanding language. Natural language understanding (NLU) is a minor part of natural language processing (NLP).

Conversational applications like Alexa and Google Assistant are developed around the NLU concept. Other applications include simple profanity filters, sentiment detection, topic classification, entity detection, and more. NLP algorithms employed on textual information or data have been pre-processed in the pre-processing step. Both NLU and NLP are mandated to make more robust and autonomous machines. Ultimately the usage of either NLP or NLU depends upon the business condition.

The advantage of using natural language processing is that it enhances communication between computers and humans. Other benefits of enforcing NLP are that they are less expensive, easy to execute, and have quick customer service response time. In addition, NLP-related tools are scalable, and the models work reasonably well on large-scale data. They process a tremendous amount of data in just a few seconds or minutes. Furthermore, NLP streamlines the process and can be executed 24/7 in real-time. The only drawback with the NLP method is that it can never be 100% dependable as there is a possibility of prediction error. The error could be because of the ever evolving and partly ambiguous natural language. While text generation utilizes recordings and voiceover for the content, precision, tone of voice, and inflection can cause a problem and affect the overall outcome of the NLP algorithm.

Modification and revision in algorithms related to natural language processing have expanded the depth and breadth of textual data that can be analysed. Semantic and Syntax Analysis are two principal approaches employed with natural language processing.

5.2.1.1.Syntax Analysis

It is the process of organizing the words in the sentence so that they grammatical make sense. NLP levy meaning to textual content based on grammatical rules. Syntax analysis steps are parsing, word segmentation, sentence breaking, morphological segmentation, and stemming.

Parsing understands the grammar of the words and implicates splitting the sentence into different parts of speech. Therefore, it is advantageous for complex downstream processing tasks. The base structure of the word is reached in the stemming technique after reducing inflected words. While stemming algorithms learn and detect infected and root forms that are the same word with distinct letters or in different word forms. Morphological segmentation separates words into morphemes—all morphemes created after the split do not always form a word with a meaning. Bound morphemes don't make any sense when used by themselves. Morphological segmentation is helpful for machine learning and speech recognition.

5.2.1.1.Semantic Analysis

Semantic Analysis grasps the meaning of the text. First, the NLP devotes different algorithms to studying every individual word's meaning (lexical semantics). Next, it checks for the combination of terms and tries to understand the context.

The technique is further sub-categorized to word sense disambiguation and relationship extraction. Relationship extraction tries to comprehend how entities relate to each other in the available textual information. On the other hand, word sense disambiguation works on understanding in what sense the words are employed.

Innumerable researchers have endeavoured to develop improved technology and techniques to complete tasks necessary for NLP works. Research in 1999 submitted a model of lexicon that concerns the automatic accession of words and representation of the semantic scope of individual lexical entries[168]. The work of Kam-Fai et al.[169] conveys the fields of an NLP device named Chicon utilized for word segmentation in Chinese text.

Noun phrasing is supposed to be a critical Natural Language Processing (NLP) method employed in Information Retrieval and Web Search (IRWS). It is because it gives the probability of fusing traditional keywords and syntactic techniques with semantic techniques to improve the quality of information retrieval. Tolle and Chen [170]compared four noun phrase generation tools to evaluate their might to isolate noun phrases from medical journal abstracts databases.

Sentiment analysis is a subcategory of Natural Language processing. And RNN (Recurrent Neural Network) is the extremely used NLP algorithm for sentiment analysis

Chapter 6. Dataset used for Fundamental and Sentimental Analysis

This chapter reviews and gives a brief explanation of the datasets used for fundamental and sentimental analysis

6.1. Fundamental Analysis Dataset

A brief description of the significant fields used in all the models is explained in Chapter 2 of the thesis. The dataset initially comprises 225 financial indicators of US Stocks. The recorded stock data is from 2014 to 2018. The financial indicators used in the dissertation are commonly founded in the 10-K filings (public information available through several sources). Almost all the economic indicators are numerical variables except a few that fall under the string and have a categorical data type. This is because all the financial hands are discrete in nature. But, in time series and econometric models, these discrete variables are treated as continuous observations.

Five features were utterly discarded from the dataset before the feature selection process. As a result, more than 4k stock information is available in each year's dataset. Exploratory Data Analysis (EDA) found that the Financial Services sector had the maximum number of records tallied for all 5 years, whereas the Communication Services sector had the minimum.

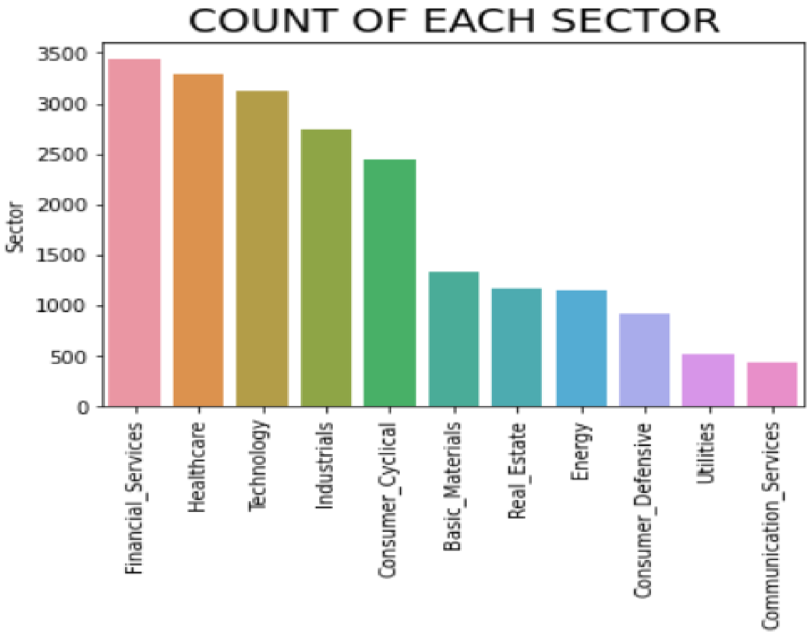

Figure 1: ***Total Count of records in each sector from 2014 to 2018***

6.2. Sentiment Analysis Dataset

For sentiment analysis, fetching stock-related information was done for the same period as fundamental data. The web scraping was performed on the Twitter website using the top tweets template on the Octoparse tool. The template was executed over 2 days to get sufficient data records for analysis. More than 137363 records were scrapped for analysis. In addition, filtration in the template was done so that no duplicated records were fetched. During the initial setup of the template, care was taken that the pulled content was semi-structured. The keywords conditioned while yanking the information are bear, bullish, closing price, 52-week range, market cap, stock, and American Stock Exchange. The keywords were selected after understanding and unveiling their frequency in general. The extraction was also conditioned for only top tweets (based on likes and retweets) as the relationship between sentiment score and number of likes and retweets are rarely inversely proportionate.

Other common keywords used for scrapping backup data are stockholder, wall street, alternative investment market, tanking stock, stock exchange, bearish, bullish, downtick, and listed companies' capitalization.

Chapter 7. Result Evaluation and Proposed Model

7.1. Evaluation of Results

Evaluation is a substantial part of any system development activity, and information science researchers put in an abundance of effort to deliver an appropriate valuation.

On the visualization of the data, it was discovered not all companies of any sector had their stock price variation in a given range. Instead, they were some outlier establishments. These companies outshined (in both directions) other companies that fall under the same subset.

As exhibited in Fig. 2, the highest oscillation was witnessed in the Industrial sector stock. Some companies had a maximum positive gain of close to 200. On the other hand, the most movement of the stock in the negative direction was 100. Glancing at the graph, one can presume that only a handful of businesses in the industrial sector are trading at relatively small amounts.

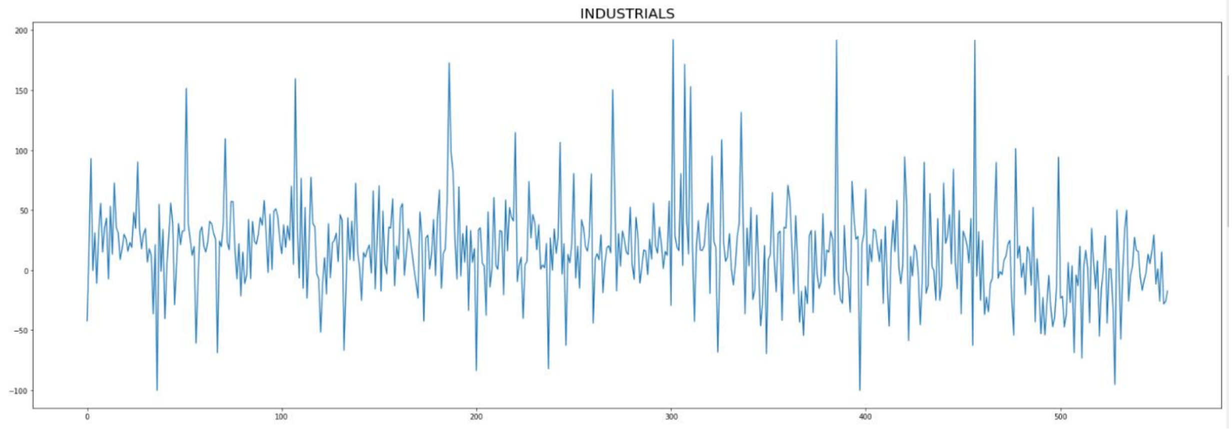

Figure 2: ***Expected price variation in Industrial stocks in 2017 based on 2016 financial data***

Fig. 3 also portrays flux in the price variation of the stock. The technology business sector comprises a couple of blue-chip stocks. The highest gain noted was approximately 1200. According to the 2016 technology graph, the technology industry does not have numerous bear markets.

Both Healthcare and Real Estate sectors have high maximum gain values. However, comparatively, healthcare gains are even higher than real estate. In addition, some healthcare stocks have such high healthy variance because they are an outstanding defensive play in an environment of rising interest rates and economic tension. Therefore, Healthcare shares are considered the most reliable defensive shares. Only a few enterprises in the real estate sector depict high swing in their price variation for the consecutive year when using 2016 financial data.

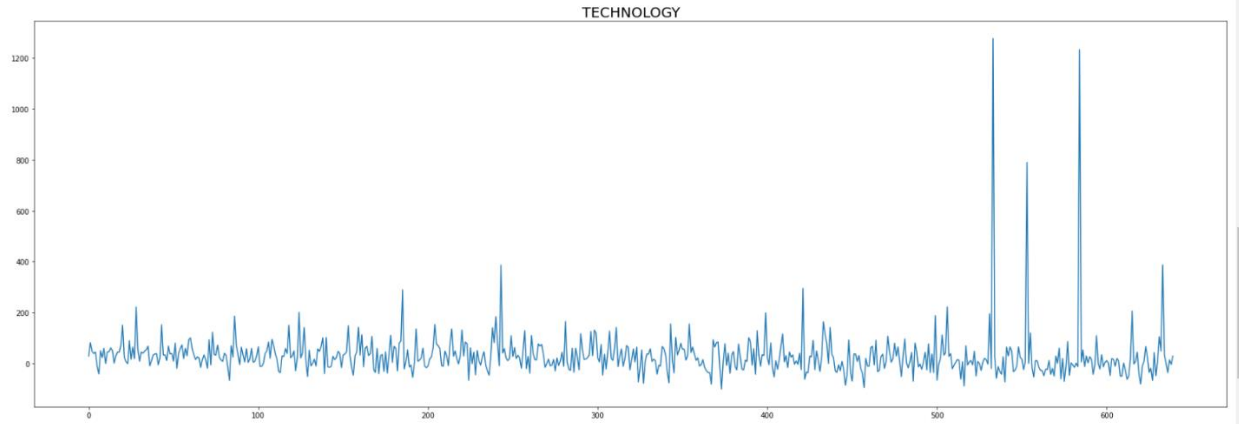

Figure 3: ***Expected price variation in Technology stocks in 2017 based on 2016 financial data***

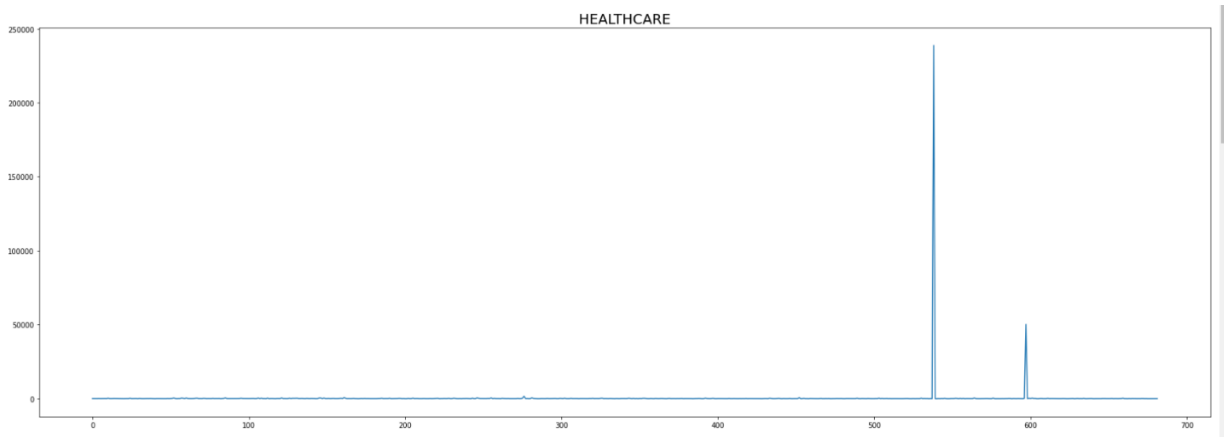

Figure 4: ***Expected price variation in Healthcare stocks in 2017 based on 2016 financial data***

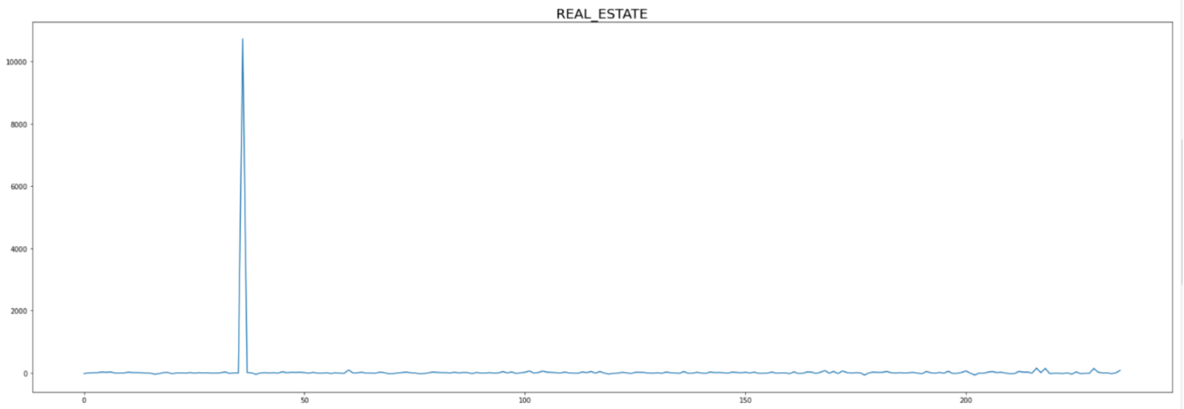

Figure 5: ***Expected price variation in Real Estate stocks in 2017 based on 2016 financial data***

In Fig 4. and Fig 5., the stock price of each enterprise is almost the same. A few small peaks and valley bounces were glimpsed in Real Estate stocks.

In feature selection comparison Table 1 the Variance Threshold feature selection and Cuckoo Optimization (CO) search method is not displayed. As variance threshold and CO approach were not implemented. Their conceptual knowledge was used to eradicate them and determine why it is not the best fit for implementation. But a comparative study of both is given in Chapter 3. The feature subset applied to all the approaches initially comprised 220 features. Sequential Feature Selection rows are empty because the computation time taken by the method was too long (more than 7 min) for the current stock market dataset. The script for Sequential Feature selection goes into an endless loop.

Lasso Method filters the important features by assigning variables values between 0 and 1. But all features value turns to 0, and therefore no important variables were extracted using this method. As a result, the number of fields remains the constant even after grid search attempts to maximize the performance. LassoCV may not have the same impact on complete execution. But it will also tend to move the value close to null. Therefore, it is also not a good option for feature selection.

When the tree-based model Random Forest Classifier was implemented for feature selection, only 1 feature was drawn out. Then, comparing the model's prediction with training-set accuracy, it was extrapolated that the model doesn't overfit. After that, the Random Forest Model was implemented again, but this time without the Price Variance feature to retrieve other essential features and to confirm if the model was performing correctly. All the parameters and hyperparameters were kept the same in the second run to develop a better comparative analysis. Random Forest Classifier extracted 116 features on the second run with an accuracy of 61.4%.

Even though the Genetic Algorithm (GA) model gives an accuracy of almost 1, it may be considered the best model for feature selection. But the model takes lots of computational time compared to LightGBM. It takes 2-3 minutes to train the algorithm. Also, the number of variables pulled by the Genetic Selection method was more than double the number of features obtained by the LightGBM Model.

Principal Component Analysis (PCA) downsized the features into a smaller number of components. In total, 135 pieces were obtained in this study. Data loss (Information loss) may occur if the correct number of features is not chosen. No method could be found to estimate the correct number of components. In PCA, each component is a linear combination of original features. Stock market data generally shows non-linearity. Interaction of noise, arbitrage traders, market volatility, and all these public information leads to non-linear results.

There was no information on whether the used dataset inherits linearity. Therefore, PCA analysis was not employed for making the final call for the feature selection process. Only 63.67% of the result was accurate when the Random Forest Classifier was in the Principal Component Analysis (PCA) process.

SelectKBest was implemented for approximating and assuring the features obtained by the LightGBM model. The mutual Information Regression model was used as an estimator, and the k value selected was 75, the same as the LightGBM model. However, all the features pulled from the SelectKBest method were not the same as LightGBM. From the solution, it could be inferred that when dependency between the components is considered, some of the features lose their importance, and a few uncommon ones are added to the list.

Method or Algorithm used	Number of Features retained	Accuracy (in %)
Chi-squared method	220	-
Sequential Feature Selection (With KNeighborClassifier)	-	-
Recursive Feature Elimination (With Logistic Regression)	220	61.24
Lasso	220	-
Recursive Feature Elimination (With Support Vector Estimator)	220	-
LassoCV	220	-
Logistic Regression	39	61.24
Random Forest Classifier (Using SelectFromModel method)	1	99.97
LightGBM	75	99.90
Genetic Algorithm	158	99.98

Table 1: ***Feature selection model comparison table***

LightGBM was the best model in terms of accuracy and the number of features for assigned research. Another reason the model is the best fit for feature selection is that the model did not overfit for such a large dataset. The model is also faster than XGBoost and supports GPU learning too. Another additional advantage of using the model was that it assigns weight to the features, and their importance could be evaluated using it. The most important feature has the highest weight value; in this matter, it was price variance.

Model	Accuracy (in %)	R-square score	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)
Random Forest	100	0.0	0.0	0.0
CatBoost Model	99.80	0.99215	0.00194	0.044
Multilayer Perceptron	56.67	-0.74851	0.43430	0.65901
Support Vector Machine (SVM)	54.57	-0.82894	0.45427	0.67400

Table 2: ***Comparison of different training model***

As revealed in Table 2, the accuracy of the Random Forest model on Fundamental Analysis data is 100%. However, there can be a possibility of overfitting as there is no such thing as a perfect model. The CatBoost Model has an accuracy of 99.80%. Even after k cross-validation, the correctness value remains the same. Hence, it can cease that there is no overfitting of the data.

On the other hand, Multilayer Perceptron and Support Vector Machine (SVM) have negative R-square scores. Therefore, it can be concluded that the model predicts worse than the mean of the target values. The accuracy of Multilayer Perceptron (MLP) and Supports Vector Machine (SVM) doesn't cross even 60% despite parameters and hyperparameters adjustments.

The research proved that a thriving prediction arrangement largely depends on a deliberate combination of feature selection methods with a baseline learning model. The whole process formed a satisfactory balance and harmony between the scourge of dimensionality and the blessing of dimensionality.

7.2. Proposed Model for the Future

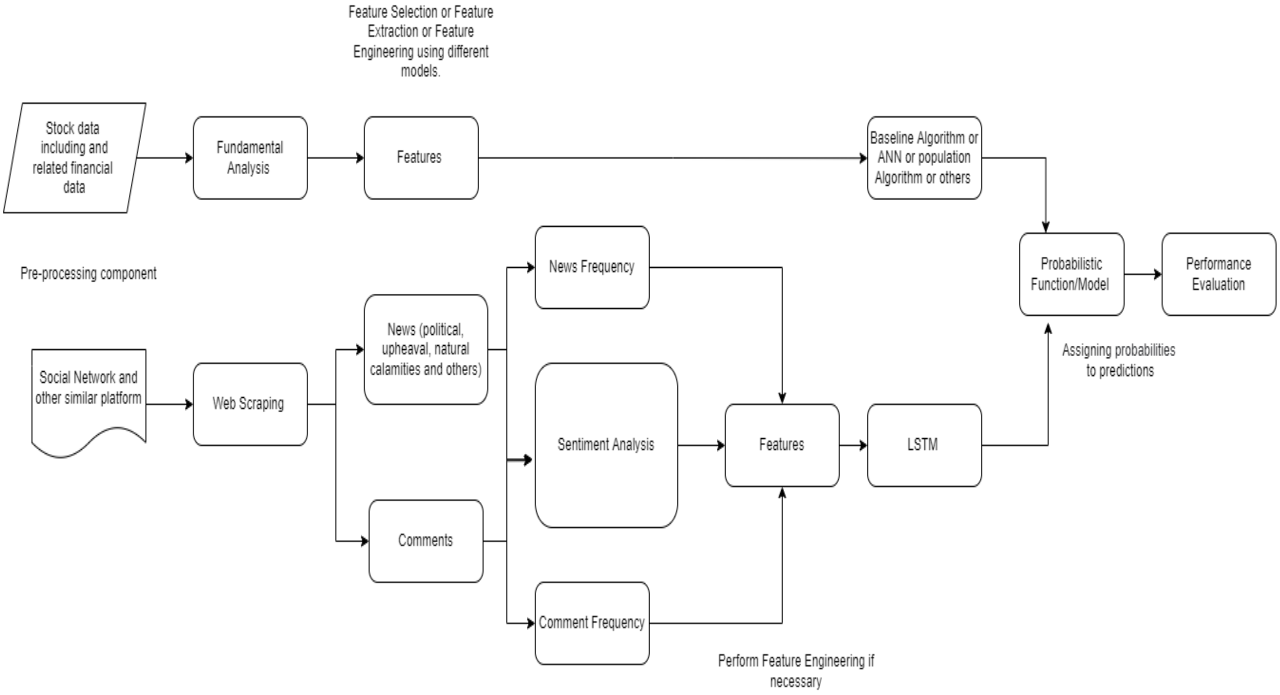
Feature Selection or Feature Extraction or Feature Engineering using different models.
Stock data including and related financial data
Fundamental Analysis
Features
Baseline Algorithm or ANN or population Algorithm or others
Pre-processing component
News Frequency
Probabilistic Function/Model
Performance Evaluation
News (political, upheaval, natural calamities and others)
Assigning probabilities to predictions
Social Network and other similar platform
Web Scraping
Sentiment Analysis
Features
LSTM
Comments
Comment Frequency
Perform Feature Engineering if necessary

Figure 6: ***Proposed Model***

As illustrated in Fig. 6, the suggested model will perform a fundamental and sentimental analysis of the stock market data. In addition, the model can be further enhanced by incorporating stock time series data (considering both historical data and historical price and volume). In the sentiment analysis part, the frequency of news and comments can be evaluated because a new feature can be generated. Simultaneously, the frequency can also aid in picking the most frequently used text and terms. Feature engineering can be performed on the variables acquired after sentiment analysis-related tasks if the features are semi-structured. In fundamental analysis, optimal features can be fetched using feature selection or feature extraction models. The use of feature selection, extraction, and engineering partially leans on the source of the data and its dimensionality.

Once the optimal features are accessed, a machine learning algorithm can be run on them to get the desired stock gain prediction. Finally, all the prediction results can be combined using a probability function or probability-based model. The weight assigned to each part of the analysis is not the same, as the effect of each analysis on a company's stock price may not be the same. The best weight value can be determined and fixed by using the data's k folds and adjusting the model's iteration parameter. The weight designation will improve the prediction accuracy and lessen the Mean Absolute Error (MAE).

Ultimately, the goal is to develop a model that is GPU supported.

Chapter 8. Conclusion

8.1. Summary and Conclusion

Since the stock market has a massive amount of data sets, it is viewed as a complex domain of research analysis. It has been the area of interest for numerous investors around the world.

The central claim of the thesis was to find the best feature selection and model for the analysis results. The study proved that a KNN imputation technique is the most reasonable when dealing with missing values in the dataset (comprising mostly numerical fields). It is best to apply imputation based on industry/sector. LightGBM is the most appropriate algorithm for feature selection. The price direction movement is accurately predicted using the CatBoost algorithm. The model tries to overfit when Random Forest Classifier is used to foretell the direction (1 means in upwards order and 0 means in downward order). No sample set was found where the price variation value remained unchanged. No price variation doesn't imply the stock value is 0. The value of the share cannot go below 0. If the value of the share for some reason becomes 0 or falls below the expected threshold value, then the share is delisted.

8.2. Future Work

Multifarious researchers have developed domain-dependent and concept-level sentiment dictionaries for better sentiment analysis[171, 172]. Separate sentiment dictionary for German political language[173], Chinese text [174, 175] , and adjective-based dictionary for social media[176] are available. Researcher Yu et al. [177] used a sentiment dictionary to foretell the direction of the stock for Korean text and market. But there is not much stock market-related research where a dictionary is incorporated for the task, which includes slang words. A dictionary (slang specific) and a model can be developed in the future to uncover the effect (estimated as probability) of slang words on the exchanges. The vernacular will include words and phrases like castles in the sky (meaning stock prices are highly overvalued), all the boat rises (it means the stock is rising quickly), painting the tape (investors illegally moving the stocks), and more. The future plan also includes updating sentiment analysis software or APIs to tackle issues like tone, polarity, sarcasm, emojis, idioms, and negations.

References

[1] C. K.-S. Leung, R. K. MacKinnon, and Y. Wang, "A machine learning approach for stock price prediction," 2014, pp. 274-277.

[2] F.-L. Lin, S.-Y. Yang, T. Marsh, and Y.-F. Chen, "Stock and bond return relations and stock market uncertainty: Evidence from wavelet analysis," *International Review of Economics & Finance,* vol. 55, pp. 285-294, 2018.

[3] J.-S. Chou and T.-K. Nguyen, "Forward forecast of stock price using sliding-window metaheuristic-optimized machine-learning regression," *IEEE Transactions on Industrial Informatics,* vol. 14, no. 7, pp. 3132-3142, 2018.

[4] J. He, L. Cai, P. Cheng, and J. Fan, "Optimal investment for retail company in electricity market," *IEEE Transactions on Industrial Informatics,* vol. 11, no. 5, pp. 1210-1219, 2015.

[5] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2018.

[6] G. Caginalp and H. Laurent, "The predictive power of price patterns," *Applied Mathematical Finance,* vol. 5, no. 3-4, pp. 181-205, 1998.

[7] A. D. Ijegwa, O. R. Vincent, O. Folorunso, and O. O. Isaac, "A Predictive Stock Market Technical Analysis Using Fuzzy Logic," *Comput. Inf. Sci.,* vol. 7, no. 3, pp. 1-17, 2014.

[8] T. H. Nguyen, K. Shirai, and J. Velcin, "Sentiment analysis on social media for stock movement prediction," *Expert Systems with Applications,* vol. 42, no. 24, pp. 9603-9611, 2015.

[9] W. Leigh, R. Purvis, and J. M. Ragusa, "Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support," *Decision support systems,* vol. 32, no. 4, pp. 361-377, 2002.

[10] L. L. Torbira, "Insurance risk management: a correlate of economic growth in Nigeria," *Research Journal of Finance and Accounting,* vol. 9, no. 7, pp. 1-10, 2018.

[11] J. Felsen, "Learning pattern recognition techniques applied to stock market forecasting," *IEEE Transactions on Systems, Man, and Cybernetics,* no. 6, pp. 583-594, 1975.

[12] Z. Guo, H. Wang, Q. Liu, and J. Yang, "A feature fusion based forecasting model for financial time series," *PloS one,* vol. 9, no. 6, p. e101113, 2014.

[13] Z. Bodie, A. Kane, and A. Marcus, *EBOOK: Essentials of Investments: Global Edition*. McGraw Hill, 2013.

[14] J. Zhang, S. Cui, Y. Xu, Q. Li, and T. Li, "A novel data-driven stock price trend prediction system," *Expert Systems with Applications,* vol. 97, pp. 60-69, 2018.

[15] E. F. Fama, "Efficient capital markets: A review of theory and empirical work," *The journal of Finance,* vol. 25, no. 2, pp. 383-417, 1970.

[16] B. G. Malkiel, "The efficient market hypothesis and its critics," *Journal of economic perspectives,* vol. 17, no. 1, pp. 59-82, 2003.

[17] K.-T. Chen, T.-J. Chen, and J.-C. Yen, "Predicting future earnings change using numeric and textual information in financial reports," 2009: Springer, pp. 54-63.

[18] L. Nanni and A. Lumini, "An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring," *Expert systems with applications,* vol. 36, no. 2, pp. 3028-3033, 2009.

[19] S. M. Mousavi Anzahaei and H. Nikoomaram, "A comparative study of the performance of Stock trading strategies based on LGBM and CatBoost algorithms," *International Journal of Finance & Managerial Accounting,* vol. 7, no. 26, pp. 63-75, 2022.

[20] L. Li, Y. Wu, Y. Ou, Q. Li, Y. Zhou, and D. Chen, "Research on machine learning algorithms and feature extraction for time series," 2017: IEEE, pp. 1-5.

[21] M. Obthong, N. Tantisantiwong, W. Jeamwatthanachai, and G. Wills, "A survey on machine learning for stock price prediction: algorithms and techniques," 2020.

[22] G. A. F. Seber and A. J. Lee, *Linear regression analysis*. John Wiley & Sons, 2012.

[23] N. Reichek and R. B. Devereux, "Reliable estimation of peak left ventricular systolic pressure by M-mode echographic-determined end-diastolic relative wall thickness: identification of severe valvular aortic stenosis in adult patients," *American heart journal,* vol. 103, no. 2, pp. 202-209, 1982.

[24] F. Allen and R. Karjalainen, "Using genetic algorithms to find technical trading rules," *Journal of financial Economics,* vol. 51, no. 2, pp. 245-271, 1999.

[25] J. Andreoni and J. H. Miller, "Auctions with artificial adaptive agents," *Games and economic behavior,* vol. 10, no. 1, pp. 39-64, 1995.

[26] N. G. Pavlidis, D. K. Tasoulis, and M. N. Vrahatis, "Financial forecasting through unsupervised clustering and evolutionary trained neural networks," 2003, vol. 4: IEEE, pp. 2314-2321.

[27] J. Nenortaite and R. Simutis, "Stocks’ trading system based on the particle swarm optimization algorithm," 2004: Springer, pp. 843-850.

[28] A. Carlisle and G. Dozier, "Adapting particle swarm optimization to dynamic environments," 2000, vol. 1: Citeseer, pp. 429-434.

[29] P. M. Tsang *et al.*, "Design and implementation of NN5 for Hong Kong stock price forecasting," *Engineering Applications of Artificial Intelligence,* vol. 20, no. 4, pp. 453-461, 2007.

[30] C.-L. Huang and C.-Y. Tsai, "A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting," *Expert Systems with applications,* vol. 36, no. 2, pp. 1529-1539, 2009.

[31] C.-M. Hsu, "A hybrid procedure for stock price prediction by integrating self-organizing map and genetic programming," *Expert Systems with Applications,* vol. 38, no. 11, pp. 14026-14036, 2011.

[32] M. O. Afolabi and O. Olude, "Predicting stock prices using a hybrid Kohonen self organizing map (SOM)," 2007: IEEE, pp. 48-48.

[33] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271,* 2018.

[34] J. Roman and A. Jameel, "Backpropagation and recurrent neural networks in financial analysis of multiple stock market returns," 1996, vol. 2: IEEE, pp. 454-460.

[35] J. B. Singh, "Current approaches in neural network modeling of financial time series," 2009.

[36] T. Kimoto, K. Asakawa, M. Yoda, and M. Takeoka, "Stock market prediction system with modular neural networks," 1990: IEEE, pp. 1-6.

[37] H. Mizuno, M. Kosaka, H. Yajima, and N. Komoda, "Application of neural network to technical analysis of stock market prediction," *Studies in Informatic and control,* vol. 7, no. 3, pp. 111-120, 1998.

[38] A. Nawani, H. Gupta, and N. Thakur, "Prediction of market capital for trading firms through data mining techniques," *International Journal of Computer Applications,* vol. 70, no. 18, 2013.

[39] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," 2004, p. 6.

[40] M. P. Naeini, H. Taremian, and H. B. Hashemi, "Stock market value prediction using neural networks," 2010: IEEE, pp. 132-136.

[41] K. N. Devi, V. M. Bhaskaran, and G. P. Kumar, "Cuckoo optimized SVM for stock market prediction," 2015: IEEE, pp. 1-5.

[42] L. Khaidem, S. Saha, and S. R. Dey, "Predicting the direction of stock market prices using random forest," *arXiv preprint arXiv:1605.00003,* 2016.

[43] J. A. K. Suykens, G. Horvath, and S. Basu, *Advances in learning theory: methods, models, and applications*. IOS Press, 2003.

[44] H. Li, Z. Yang, and T. Li, "Algorithmic trading strategy based on massive data mining," *Stanford University Stanford,* 2014.

[45] D. Komo, C.-I. Chang, and H. Ko, "Neural network technology for stock market index prediction," 1994: IEEE, pp. 543-546.

[46] R. K. Dase and D. D. Pawar, "Application of Artificial Neural Network for stock market predictions: A review of literature," *International Journal of Machine Intelligence,* vol. 2, no. 2, pp. 14-17, 2010.

[47] M. R. Hassan and B. Nath, "Stock market forecasting using hidden Markov model: a new approach," 2005: IEEE, pp. 192-196.

[48] S. Boonpeng and P. Jeatrakul, "Decision support system for investing in stock market by using OAA-neural network," 2016: IEEE, pp. 1-6.

[49] J. W. Wilder, *New concepts in technical trading systems*. Trend Research, 1978.

[50] G. Appel and M. Appel, "A quick tutorial in MACD: Basic concepts," Working Paper, 2008.

[51] E. N. Desokey, A. Badr, and A. F. Hegazy, "Enhancing stock prediction clustering using K-means with genetic algorithm," 2017: IEEE, pp. 256-261.

[52] D. M. Q. Nelson, A. C. M. Pereira, and R. A. De Oliveira, "Stock market's price movement prediction with LSTM neural networks," 2017: Ieee, pp. 1419-1426.

[53] D. Farid, A. R. Meybodi, and S. H. Mirfakhraddiny, "Investment risk management in Tehran Stock Exchange (TSE) using technique of Monte Carlo Simulation (MCS)," *Journal of Financial Crime,* 2010.

[54] D. W. H. Alrabadi and N. I. Abu Aljarayesh, "Forecasting Stock Market Returns Via Monte Carlo Simulation: The Case of Amman Stock Exchange," *Jordan Journal of Business Administration,* vol. 11, no. 3, 2015.

[55] W. F. Tucker, "A real estate portfolio optimizer utilizing spreadsheet modeling with Markowitz mean-variance optimization and Monte-Carlo simulation," Working paper, John Hopkins University, Maryland, 2001.

[56] C. F. Kelliher and L. S. Mahoney, "Using Monte Carlo simulation to improve long-term investment decisions," *The Appraisal Journal,* vol. 68, no. 1, p. 44, 2000.

[57] N. French and L. Gabrielli, "The uncertainty of valuation," *Journal of Property Investment & Finance,* 2004.

[58] L. H. Li, "Simple computer applications improve the versatility of discounted cash flow analysis," *The Appraisal Journal,* vol. 68, no. 1, p. 86, 2000.

[59] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends® in information retrieval,* vol. 2, no. 1–2, pp. 1-135, 2008.

[60] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," 2010.

[61] A. J. T. Lee, M.-C. Lin, R.-T. Kao, and K.-T. Chen, "An effective clustering approach to stock market prediction," 2010.

[62] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan, "Mining of concurrent text and time series," 2000, vol. 2000: Citeseer University Park, PA, USA, pp. 37-44.

[63] A. Kloptchenko, T. Eklund, J. Karlsson, B. Back, H. Vanharanta, and A. Visa, "Combining data and text mining techniques for analysing financial reports," *Intelligent Systems in Accounting, Finance & Management: International Journal,* vol. 12, no. 1, pp. 29-41, 2004.

[64] B. Back, J. Toivonen, H. Vanharanta, and A. Visa, "Comparing numerical data and text information from annual reports using self-organizing maps," *International journal of accounting information systems,* vol. 2, no. 4, pp. 249-269, 2001.

[65] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis," *Knowledge-Based Systems,* vol. 69, pp. 14-23, 2014.

[66] T. Wilson *et al.*, "OpinionFinder: A system for subjectivity analysis," 2005, pp. 34-35.

[67] E. Gilbert and K. Karahalios, "Widespread worry and the stock market," 2010.

[68] P. Uhr, A. Klahold, and M. Fathi, "Imitation of the human ability of word association," *International Journal of Soft Computing and Software Engineering (JSCSE),* vol. 3, no. 3, pp. 248-254, 2013.

[69] S. Deng, T. Mitsubuchi, K. Shioda, T. Shimada, and A. Sakurai, "Combining technical analysis with sentiment analysis for stock price prediction," 2011: IEEE, pp. 800-807.

[70] M. C. Thomsett, *Mastering technical analysis*. Dearborn Trade Publishing, 1999.

[71] E. Beyaz, F. Tekiner, X.-j. Zeng, and J. Keane, "Comparing technical and fundamental indicators in stock price forecasting," 2018: IEEE, pp. 1607-1613.

[72] B. Graham and J. Zweig, *The intelligent investor*. HarperBusiness Essentials New York, 2003.

[73] M. Meliana, H. Kesuma, D. Enjelina, A. Rijanto, and D. S. Saraswati, "IS CASH FLOW GROWTH HELPING STOCK PERFORMANCE DuRING THE COVID-19 OuTBREAK? EVIDENCE FROM INDONESIA," 2022.

[74] L. Insider. "3-Year CAGR   definition." https://www.lawinsider.com/dictionary/3-year-cagr (accessed.

[75] L. Insider. "5Y Dividend per Share Growth." https://www.lawinsider.com/dictionary/net-income-growth (accessed.

[76] J. Boudoukh, M. P. Richardson, and R. E. Whitelaw, "A tale of three schools: Insights on autocorrelations of short-horizon stock returns," *Review of financial studies,* vol. 7, no. 3, pp. 539-573, 1994.

[77] B. M. Friedman, D. I. Laibson, and H. P. Minsky, "Economic implications of extraordinary movements in stock prices," *Brookings Papers on Economic Activity,* vol. 1989, no. 2, pp. 137-189, 1989.

[78] R. Kohavi, "Feature subset selection as search with probabilistic estimates," 1994, vol. 224, pp. 109-113.

[79] P. Langley, "Selection of Relevant Features," 1994, pp. 171-182.

[80] D. Koller and M. Sahami, "Toward optimal feature selection," Stanford InfoLab, 1996.

[81] P. Langley and S. Sage, "Induction of selective Bayesian classifiers," in *Uncertainty Proceedings 1994*: Elsevier, 1994, pp. 399-406.

[82] R. A. Iqbal, "Using feature weights to improve performance of neural networks," *arXiv preprint arXiv:1101.4918,* 2011.

[83] S. Shankar and G. Karypis, "A feature weight adjustment algorithm for document categorization," 2000.

[84] H. Cai, P. Ruan, M. Ng, and T. Akutsu, "Feature weight estimation for gene selection: a local hyperlinear learning approach," *BMC bioinformatics,* vol. 15, no. 1, pp. 1-13, 2014.

[85] H. Almuallim and T. G. Dietterich, "Learning boolean concepts in the presence of many irrelevant features," *Artificial intelligence,* vol. 69, no. 1-2, pp. 279-305, 1994.

[86] Y. Lin, Q. Hu, J. Liu, and J. Duan, "Multi-label feature selection based on max-dependency and min-redundancy," *Neurocomputing,* vol. 168, pp. 92-103, 2015.

[87] D. H. Fisher, "Knowledge acquisition via incremental conceptual clustering," *Machine learning,* vol. 2, no. 2, pp. 139-172, 1987.

[88] L. Talavera, "Dependency-based feature selection for clustering symbolic data," *Intelligent Data Analysis,* vol. 4, no. 1, pp. 19-28, 2000.

[89] J. J. Furtado Vasco, "Determining property relevance in concept formation by computing correlation between properties," 1998: Springer, pp. 310-315.

[90] S. Piramuthu, "Evaluating feature selection methods for learning in data mining applications," *European journal of operational research,* vol. 156, no. 2, pp. 483-494, 2004.

[91] C.-J. Huang, D.-X. Yang, and Y.-T. Chuang, "Application of wrapper approach and composite classifier to the stock trend prediction," *Expert Systems with Applications,* vol. 34, no. 4, pp. 2870-2878, 2008.

[92] A. El Akadi, A. El Ouardighi, and D. Aboutajdine, "A powerful feature selection approach based on mutual information," *International Journal of Computer Science and Network Security,* vol. 8, no. 4, p. 116, 2008.

[93] V. Kumar and S. Minz, "Feature selection: a literature review," *SmartCR,* vol. 4, no. 3, pp. 211-229, 2014.

[94] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research,* vol. 3, no. Mar, pp. 1157-1182, 2003.

[95] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology,* vol. 3, no. 02, pp. 185-205, 2005.

[96] K.-B. Duan, J. C. Rajapakse, H. Wang, and F. Azuaje, "Multiple SVM-RFE for gene selection in cancer classification with expression data," *IEEE transactions on nanobioscience,* vol. 4, no. 3, pp. 228-234, 2005.

[97] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning,* vol. 46, no. 1, pp. 389-422, 2002.

[98] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern recognition letters,* vol. 31, no. 14, pp. 2225-2236, 2010.

[99] A. Hyvärinen, "Survey on independent component analysis," 1999.

[100] L. Rangarajan, "Bi-level dimensionality reduction methods using feature selection and feature extraction," *International Journal of Computer Applications,* vol. 4, no. 2, pp. 33-38, 2010.

[101] P. G. Lovaglio and G. Vittadini, "Multilevel dimensionality-reduction methods," *Statistical Methods & Applications,* vol. 22, no. 2, pp. 183-207, 2013.

[102] R. P. Heydorn, "Redundancy in feature extraction," *IEEE Transactions on Computers,* vol. 100, no. 9, pp. 1051-1054, 1971.

[103] C. Lee and D. A. Landgrebe, "Feature extraction based on decision boundaries," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 15, no. 4, pp. 388-400, 1993.

[104] A. Janecek, W. Gansterer, M. Demel, and G. Ecker, "On the relationship between feature selection and classification accuracy," 2008: PMLR, pp. 90-105.

[105] Z. Duszak and W. W. Loczkodaj, "Using Principal Component Transformation in Machine Learning," 1994, pp. 125-129.

[106] H. Yu, R. Chen, and G. Zhang, "A SVM stock selection model within PCA," *Procedia computer science,* vol. 31, pp. 406-412, 2014.

[107] H. Gunduz, "An efficient stock market prediction model using hybrid feature reduction method based on variational autoencoders and recursive feature elimination," *Financial Innovation,* vol. 7, no. 1, pp. 1-24, 2021.

[108] K. Kumar, M. Haider, and T. Uddin, "Enhanced prediction of intra-day stock market using metaheuristic optimization on RNN–LSTM network," *New Generation Computing,* vol. 39, no. 1, pp. 231-272, 2021.

[109] B. Weng, M. A. Ahmed, and F. M. Megahed, "Stock market one-day ahead movement prediction using disparate data sources," *Expert Systems with Applications,* vol. 79, pp. 153-163, 2017.

[110] N. Nagaraj, B. M. Vikranth, and N. Yogesh, "Recursive Feature Elimination Technique for Technical Indicators Selection," 2022: Springer, pp. 139-145.

[111] K. Kumar and M. T. U. Haider, "Blended computation of machine learning with the recurrent neural network for intra-day stock market movement prediction using a multi-level classifier," *International Journal of Computers and Applications,* vol. 43, no. 8, pp. 733-749, 2021.

[112] J. Shen and M. O. Shafiq, "Short-term stock market price trend prediction using a comprehensive deep learning system," *Journal of big Data,* vol. 7, no. 1, pp. 1-33, 2020.

[113] Y. Xu, Z. Li, and L. Luo, "A study on feature selection for trend prediction of stock trading price," 2013: IEEE, pp. 579-582.

[114] R. Muthukrishnan and R. Rohini, "LASSO: A feature selection technique in predictive modeling for machine learning," 2016: IEEE, pp. 18-20.

[115] M. Waqar, H. Dawood, P. Guo, M. B. Shahnawaz, and M. A. Ghazanfar, "Prediction of stock market by principal component analysis," 2017: IEEE, pp. 599-602.

[116] V. Fonti and E. Belitser, "Feature selection using lasso," *VU Amsterdam research paper in business analytics,* vol. 30, pp. 1-25, 2017.

[117] Q. Cheng, P. K. Varshney, and M. K. Arora, "Logistic regression for feature selection and soft classification of remote sensing data," *IEEE Geoscience and Remote Sensing Letters,* vol. 3, no. 4, pp. 491-494, 2006.

[118] S. Pang, "An application of logistic model in stock forecasting," 2004, vol. 2: IEEE, pp. 1491-1496.

[119] S. Pang, F. Deng, and Y. Wang, "A comparison of forecasting models of the volatility in Shenzhen stock market," *Acta Mathematica Scientia,* vol. 27, no. 1, pp. 125-136, 2007.

[120] S. L. Salzberg, "C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993," ed: Kluwer Academic Publishers, 1994.

[121] P. Li, Q. Wu, and C. Burges, "Mcrank: Learning to rank using multiple classification and gradient boosting," *Advances in neural information processing systems,* vol. 20, 2007.

[122] R. Jin and G. Agrawal, "Communication and memory efficient parallel decision tree construction," 2003: SIAM, pp. 119-129.

[123] S. Ranka and V. Singh, "CLOUDS: A decision tree classifier for large datasets," 1998, vol. 2, 8 ed.

[124] Y. Yang, Y. Wu, P. Wang, and X. Jiali, "Stock price prediction based on xgboost and lightgbm," 2021, vol. 275: EDP Sciences, p. 01040.

[125] K. K. Yun, S. W. Yoon, and D. Won, "Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process," *Expert Systems with Applications,* vol. 186, p. 115716, 2021.

[126] Y. Han, J. Kim, and D. Enke, "A machine learning trading system for the stock market based on N-period Min-Max labeling using XGBoost," *Expert Systems with Applications,* p. 118581, 2022.

[127] Z. J. Ye and B. W. Schuller, "Capturing dynamics of post-earnings-announcement drift using a genetic algorithm-optimized XGBoost," *Expert Systems with Applications,* vol. 177, p. 114892, 2021.

[128] E. Zolotareva, "Aiding long-term investment decisions with XGBoost machine learning model," 2021: Springer, pp. 414-427.

[129] K. T. Meetei, "A survey: swarm intelligence vs. genetic algorithm," *International Journal of Science and Research (IJSR),* vol. 3, pp. 231-235, 2014.

[130] T. V. Mathew, "Genetic algorithm," *Report submitted at IIT Bombay,* 2012.

[131] K. D. Joshi and A. A. Pandya, "Genetic algorithms and their applications-traveling sales person and antenna design," 2003.

[132] H. Chung and K.-s. Shin, "Genetic algorithm-optimized long short-term memory network for stock market prediction," *Sustainability,* vol. 10, no. 10, p. 3765, 2018.

[133] C. Schoreels, B. Logan, and J. M. Garibaldi, "Agent based genetic algorithm employing financial technical analysis for making trading decisions using historical equity market data," 2004: IEEE, pp. 421-424.

[134] L. Lin, L. Cao, J. Wang, and C. Zhang, "The applications of genetic algorithms in stock market data mining optimisation," *Management Information Systems,* 2004.

[135] A. Akbarzadeh and E. Shadkam, "The study of cuckoo optimization algorithm for production planning problem," *arXiv preprint arXiv:1508.01310,* 2015.

[136] X.-S. Yang and S. Deb, "Cuckoo search via Lévy flights," 2009: Ieee, pp. 210-214.

[137] J. Kwiecień and B. Filipowicz, "Comparison of firefly and cockroach algorithms in selected discrete and combinatorial problems," *Bulletin of the Polish Academy of Sciences. Technical Sciences,* vol. 62, no. 4, 2014.

[138] P. Civicioglu and E. Besdok, "A conceptual comparison of the Cuckoo-search, particle swarm optimization, differential evolution and artificial bee colony algorithms," *Artificial intelligence review,* vol. 39, no. 4, pp. 315-346, 2013.

[139] A. Layeb, "A novel quantum inspired cuckoo search for knapsack problems," *International Journal of Bio-Inspired Computation (IJBIC),* 2011.

[140] Y. Wang and Z. Cai, "A hybrid multi-swarm particle swarm optimization to solve constrained optimization problems," *Frontiers of Computer Science in China,* vol. 3, no. 1, pp. 38-52, 2009.

[141] M. Alzaqebah *et al.*, "Memory based cuckoo search algorithm for feature selection of gene expression dataset," *Informatics in Medicine Unlocked,* vol. 24, p. 100572, 2021.

[142] D. V. Setty, T. M. Rangaswamy, and K. N. Subramanya, "A review on data mining applications to the performance of stock marketing," *International Journal of Computer Applications,* vol. 1, no. 3, pp. 33-43, 2010.

[143] S. L. Özesmi, C. O. Tan, and U. Özesmi, "Methodological issues in building, training, and testing artificial neural networks in ecological applications," *Ecological Modelling,* vol. 195, no. 1-2, pp. 83-93, 2006.

[144] J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, and S.-H. Deng, "Hyperparameter optimization for machine learning models based on Bayesian optimization," *Journal of Electronic Science and Technology,* vol. 17, no. 1, pp. 26-40, 2019.

[145] A. Zheng, *Evaluating machine learning models: a beginner's guide to key concepts and pitfalls*. O'Reilly Media, 2015.

[146] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009.

[147] L. Breiman, "Random forests," *Machine learning,* vol. 45, no. 1, pp. 5-32, 2001.

[148] G. G. Creamer and Y. Freund, "Predicting performance and quantifying corporate governance risk for latin american adrs and banks," *Financial Engineering and Applications, MIT, Cambridge,* 2004.

[149] B. Larivière and D. Van den Poel, "Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services," *Expert Systems with Applications,* vol. 27, no. 2, pp. 277-285, 2004.

[150] R. Kohavi and C.-H. Li, "Oblivious decision trees, graphs, and top-down pruning," 1995: Citeseer, pp. 1071-1079.

[151] P. Langley and S. Sage, "Oblivious decision trees and abstract cases," 1994: Seattle, WA, pp. 113-117.

[152] A. Gulin, I. Kuralenok, and D. Pavlov, "Winning the transfer learning track of yahoo!'s learning to rank challenge with yetirank," 2011: PMLR, pp. 63-76.

[153] R. Ślepaczuk and M. Zenkova, "Robustness of support vector machines in algorithmic trading on cryptocurrency market," *Central European Economic Journal,* vol. 5, no. 52, pp. 186-205, 2018.

[154] C. Bousoño-Calzón, J. Bustarviejo-Muñoz, P. Aceituno-Aceituno, and J. J. Escudero-Garzás, "On the economic significance of stock market prediction and the no free lunch theorem," *IEEE Access,* vol. 7, pp. 75177-75188, 2019.

[155] Y. Chen and Y. Hao, "A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction," *Expert Systems with Applications,* vol. 80, pp. 340-355, 2017.

[156] C. S. Vui, G. K. Soon, C. K. On, R. Alfred, and P. Anthony, "A review of stock market prediction with Artificial neural network (ANN)," 2013: IEEE, pp. 477-482.

[157] A. Engelbrecht, "Computational Intelligence: An Introduction," *J. Artificial Societies and Social Simulation,* vol. 7, 01/01 2004, doi: 10.1002/9780470512517.

[158] J. K. Mantri, "Comparison between SVM and MLP in predicting stock index trends," *International Journal of Science and Modern Engineering,* vol. 1, no. 9, pp. 81-82, 2013.

[159] A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks,* pp. 37-45, 2012.

[160] C. Chen and L. Xia, "Recurrent neural network and long short-term memory," ed: Citeseer, 2015.

[161] X. Huang *et al.*, "Lstm based sentiment analysis for cryptocurrency prediction," 2021: Springer, pp. 617-621.

[162] Z. Jin, Y. Yang, and Y. Liu, "Stock closing price prediction based on sentiment analysis and LSTM," *Neural Computing and Applications,* vol. 32, no. 13, pp. 9713-9729, 2020.

[163] Y. Guo, "Stock price prediction based on LSTM neural network: the effectiveness of news sentiment analysis," 2020: IEEE, pp. 1018-1024.

[164] R. Achkar, F. Elias-Sleiman, H. Ezzidine, and N. Haidar, "Comparison of BPA-MLP and LSTM-RNN for stocks prediction," 2018: IEEE, pp. 48-51.

[165] H. R. Bernard and G. Ryan, "Text analysis," *Handbook of methods in cultural anthropology,* vol. 613, 1998.

[166] D. Tedlock, "Hearing a voice in an ancient text: Quiché Maya poetics in performance," *Native American discourse: Poetics and rhetoric,* pp. 140-175, 1987.

[167] K. A. Jehn and L. Doucet, "Developing categories for interview data: Consequences of different coding and analysis strategies in understanding text: Part 2," *CAM Journal,* vol. 9, no. 1, pp. 1-7, 1997.

[168] S. Dogru and J. R. Slagle, "Implementing a semantic lexicon," 1999: Springer, pp. 154-167.

[169] K. F. Wong, V. Y. Lum, and W. I. Lam, "Chicon—A Chinese text manipulation language," *Software: Practice and Experience,* vol. 28, no. 7, pp. 681-701, 1998.

[170] K. M. Tolle and H. Chen, "Comparing noun phrasing techniques for use with medical digital library tools," *Journal of the American society for information science,* vol. 51, no. 4, pp. 352-370, 2000.

[171] M. Ahmed, Q. Chen, and Z. Li, "Constructing domain-dependent sentiment dictionary for sentiment analysis," *Neural Computing and Applications,* vol. 32, no. 18, pp. 14719-14732, 2020.

[172] A. C.-R. Tsai, C.-E. Wu, R. T.-H. Tsai, and J. Y.-j. Hsu, "Building a concept-level sentiment dictionary based on commonsense knowledge," *IEEE Intelligent Systems,* vol. 28, no. 2, pp. 22-30, 2013.

[173] C. Rauh, "Validating a sentiment dictionary for German political language—a workbench note," *Journal of Information Technology & Politics,* vol. 15, no. 4, pp. 319-343, 2018.

[174] G. Xu, Z. Yu, H. Yao, F. Li, Y. Meng, and X. Wu, "Chinese text sentiment analysis based on extended sentiment dictionary," *IEEE Access,* vol. 7, pp. 43749-43762, 2019.

[175] S.-M. Wang and L.-W. Ku, "ANTUSD: A large Chinese sentiment dictionary," 2016, pp. 2697-2702.

[176] W. Peng and D. H. Park, "Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization," 2011.

[177] E. Yu, Y. Kim, N. Kim, and S. R. Jeong, "Predicting the direction of the stock index by using a domain-specific sentiment dictionary," *Journal of intelligence and information systems,* vol. 19, no. 1, pp. 95-110, 2013.