Effective Feature Selection for Real-time Stock Trading in Variable Time-frames and Multi Criteria Decision Theory based Efficient Stock Portfolio Management

by

A. K. M. Amanat Ullah
20166016

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
M.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
August 2021

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

A. K. M. Amanat Ullah
20166016

Approval

The thesis/project titled "Effective Feature Selection for Real-time Stock Trading in Variable Time-frames and Multi Criteria Decision Theory based Efficient Stock Portfolio Management" submitted by

1. A. K. M. Amanat Ullah (20166016)

of Summer, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science and Engineering on August 23, 2021.

Examining Committee:

Supervisor:
(Member)

Md. Golam Rabiul Alam, PhD
Associate Professor
Department of Computer Science and Engineering
Brac University

Examiner:
(External)

Chowdhury Farhan Ahmed, PhD
Professor
Department of Computer Science and Engineering
University of Dhaka

Examiner:
(Internal)

Md. Ashraful Alam, PhD
Assistant Professor
Department of Computer Science and Engineering
Brac University

Examiner:
(Internal)

Muhammad Iqbal Hossain, PhD
Assistant Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Amitabha Chakrabarty, PhD
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Ethics Statement

This thesis was carried out in complete compliance with research ethics norms, and the codes and practices set by BRAC University. I have ensured that all our sources have been cited. As the author of this thesis, I take full responsibility for any ethics code violations.

Abstract

The unpredictability and volatility of the stock market render it challenging to make a substantial profit using any generalized scheme. Many previous studies tried different techniques to build a machine learning model, which can make a significant profit in the US stock market by performing live trading. However, very few studies have focused on the importance of finding the best features for a particular period for trading. Our top approach used the performance to narrow down the features from a total of 148 to about 30. Furthermore, the top 25 features were dynamically selected before each time training our machine learning model. It uses ensemble learning with four classifiers: Gaussian Naive Bayes, Decision Tree, Logistic Regression with L1 regularization and Stochastic Gradient Descent, to decide whether to go long or short on a particular stock. Our best model performed daily trade between July 2011 and January 2019, generating 54.35% profit. We further propose a novel model which uses Ada-boost to find the weights of each of the features and then apply TOPSIS to select the best stocks. Lastly, we survey the machine learning techniques used for ethical decision-making in stock trading, which will benefit any further research work on Responsible AI in Finance.

Keywords: Feature Selection, Multi Criteria Decision Theory, Computational Finance, Automated Trading, Responsible AI

Dedication

I would like to dedicate this research to my parents who have brought me to this world and nurtured me to become an adult.

Acknowledgement

Firstly, thanks to our beloved family members to whom I am, and always will be indebted.

Secondly, to my supervisor Md. Golam Rabiul Alam for all his help and constant support of our endeavours.

Thirdly, I would like to thank all the faculty members, staff, our peers and other stakeholders related to BRAC University for providing us an environment in which we were able to develop ourselves, learn to our fullest extent and conduct our research properly.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Stocks are essentially small pieces of ownership of a company, and the stock market works like an auction where investors buy and sell stocks. Owning stock means the shareholder owns a proportion of the company equal to the number of shares the person bought, against the total outstanding shares of the company. For example, if a company has 1 million shares and an individual owns 50,000 shares of the company, the person has a 5% stake in it.

1.1.1 Investing in Stocks

According to numerous studies, stocks produce greater returns than other assets. Stock returns mainly come from capital gains and dividends. Capital gains are when you sell a particular stock at a higher price than at which you purchased it. Dividends are a share of the profit that the company whose stocks you purchased makes, and distributes it to its shareholders. According to S&P Dow Jones Indices, since 1926, dividends have contributed to a third of investment returns while the other two-thirds have been contributed by capital gains.
The prospect of buying shares from largely successful companies such as Apple, Amazon, Facebook, Google, and Netflix, together denoted by the famous acronym FAANG, during the early stages of stock trading can seem tempting. Investors with a high tolerance for risk would lean more towards capital gains for earning profit rather than dividends. Others who prefer a more conservative approach may choose to stick with stocks that have historically been known to provide consistent and significant dividends.

1.1.2 Stock classification

Several classification methods can be applied to categorize stocks. They are usually classified in two ways- according to their sector, or by market capitalization.
Market capitalization equals the total amount of outstanding shares of a company. This is found by multiplying the present market price of a share with the total number of shares outstanding. Companies with a market capitalization of $10 billion or more are classified as large-cap, companies that have between $2 billion and $10

billion are mid-cap, and small-cap companies are those with a market cap between $300 million and $2 billion.
The Global Industry Classification Standard (GICS) is the industry standard for the sector-wise classification of stocks. Developed by S&P Dow Jones Indices and MSCI (Morgan Stanley Capital International) in 1999, the GICS is a useful tool that reflects the dimensions and progress of the industry sectors. The four-tier industry classification system is made up of 24 industry groups across 11 sectors. The sectors areas are listed in table 1.1.

Table 1.1: Different Sectors of the Stock Market

Different Sectors
Health Care
Real Estate
Communication Services
Financial
Materials
Energy
Industrial
Consumer Staples
Information Technology
Utilities
Consumer Discretionary

Sector classification allows traders to invest with respect to their respective risk preferences. A conservative investor, for example, may opt to buy stocks from industries with more stable prices and which continually provide dividends. Others who opt for a high-risk high-return strategy may opt to buy stocks from sectors such as energy, financial and IT.

1.1.3 Long-short investment strategy

Traditionally, stock investing was focused on looking for stocks to buy long that is likely to appreciate [4]. There was little, if any, thought given to capitalizing on short-selling overvalued stocks. When investors began to employ both long and short strategies in their investment portfolio more benefits and opportunities presented themselves which was previously unavailable.
Buying long is simply buying a stock that you think will appreciate, and selling for profit when the stock price rises. For instance, imagine that you bought 500 shares of a particular stock, at $10 per share. This amounts to $5000. After a week, the price of a share of ABC rises to $55. You sell the stock, pocketing a profit of $500.
Shorting is when you borrow stocks that you expect will depreciate from a broker, at interest, and selling them while you wait for the price to drop. Once the price has lowered a significant amount, you pay back the lender by buying the same number of stocks that you borrowed in the first place, at the lower price. Your profit is the difference in price minus the interest and commissions.
For instance, you borrow 100 shares of XYZ, at $50 per share, and immediately sell them for $5000 while waiting for the share price to depreciate. Once the price per share of XYZ has dropped to $45, you buy 100 shares of XYZ and pay $4500 for it.

Return the 100 shares to the lender and whatever remains minus the interest and commissions is the profit. In this case, your profit is $500.

1.2 Motivation

"I will tell you how to become rich. Close the doors. Be fearful when others are greedy. Be greedy when others are fearful." — Warren Buffett. The quote suggests that trading decisions need to be made entirely based on logic and not based on human emotions. Often times people cannot control their emotions It is difficult to let out emotion while trading. Effective trading involves making decision without letting emotions getting in the way. The perfect way to solve this problem is to deploy a machine which solely relies on logic to make effective decisions.

On another note, current estimates show that automated trading accounts for 50-70 percent of equities trades in the United States, 40% in Canada, and 35% in London [36] [29]. Therefore there will come a time where all the trades will be managed by machines. To prepare the world for such a time more research into this field is vital.

We are all aware of the unpredictability of the stock market, and how difficult it is to predict. Some people believe that it is not possible to do so. We believe that with the advancements in Machine Learning algorithms and Artificial Intelligence, we can predict stock market trends sufficiently, given we provide sufficient, refined, data to our models. Many previous researches[26], [41], [84] worked with selecting features with different algorithms for maximizing profit. However most of them did not run the final test on an actual stock trading setting. Additionally, none of the research worked on which feature time-frame works best for how many days of trading. Our research aims to explore this research gap. Recent researches [67], [68], [72] concluded that using TOPSIS over different classifiers proved to be give a better result than conventionally using a machine learning classifier. We intend to introduce a new way of calculating weights for the TOPSIS method. Recently expert computer systems have been found to give better recommendations compared to domain experts. The conventional way of assigning weights was to survey domain experts.

1.3 Contributions

To our knowledge, Our model is the first to recommend which feature time-frame suitable for how many days of trading. To address this problem our novel approach calculated each of the features using different time-variants (default, 1day, 2 day, 5 day, 22 day) to find out which variant works best for daily trading, weekly trading and monthly trading. Using the recommended features our model further used dynamic feature selection techniques coupled with advanced machine learning algorithms to generate profit on real-time stock data from 1500 stock from the US stock market from 2011 to 2019 and generated significant profit which is on par and in some cases better than the state-of-the-art models. In our proposed model uses the TOPSIS method from Multi Criteria Decision Theory. The novelty in the work is that the weights were assigned automatically using the weights from the Adaboost classifier. Additionally, ethical decision making in automated trading systems has been an emerging topic. Therefore we go through the research done so far in this topic. The main contributions of this research are as follows:

- A Dynamic Feature Selection mechanism has been proposed to select discriminative features over multiple time-frames for holding long and short positions for effective stock trading.
- After the initial feature selection mechanism the proposed model uses ANOVA for finalizing the set of features and uses ensemble of various machine learning algorithms for stock trading that generated 54.35% profit on the initial investment.
- The research also proposed a Multi Criteria Decision Theory based Efficient stock Portfolio Management system which uses the weights generated by the Adaptive boosting algorithm to perform the TOPSIS method to select which stock to go into long position and which stocks to go into short position.
- A brief Review of Responsible AI in Financial Investment is done to answer why it is needed and what work has been done so far on the topic.

1.4 Thesis Organization

In the next section of our paper, we discuss how others used machine learning models in order to predict trends in the stock market. We discuss our model for Dynamic feature Selection over multiple time-frames for efficient stock trading in chapter 3. Chapter 4 is on Multi Criteria Decision Theory based Efficient stock Portfolio Management on TOPSIS and AdaBoost. Chapter 5 talks about the future directions of automated trading in finance considering Responsible use of AI. The final chapter concludes the paper and talks about the limitations in our thesis and future prospects.

Chapter 2

Literature Review

The prevalence of volatility in the stock market makes predicting stock prices anything but simple. Before investing, investors perform two kinds of analysis [59]. The first of these is fundamental analysis, where investors look into the value of stocks, the industry performance, economic factors, etc. and decide whether or not to invest. Technical analysis is the second, more advanced, analysis that involves evaluating those stocks through the use of statistics and activity in the current market, such as volume traded and previous price levels [59]. Technical analysts use charts to recognise patterns and try to predict how a stock price will change. Malkiel and Fama's Efficient market hypothesis states that predicting the values of stocks considering financial information is possible because the prices are informationally efficient [1]. As many unpredictable variables influence stocks and the stock market in general, it seems logical that factors such as the public image of the company and the political scenario of a country will be reflected in the prices. By sufficiently preprocessing the data obtained from stock prices and the algorithms and their factors are appropriate, it may be possible to predict stock or stock price index.

There were quite a few different implementations of machine learning algorithms for the purposes of making stock market price predictions. Different papers experimented with different machine learning algorithms that they implemented in order to figure out which models produced the best results. Dai and et al. attempted to narrow down the environment by selecting certain criteria [34]. Under these criteria, they were able to achieve a profit of 0.0123, recall 30.05%, with an accuracy of 38.39%, and 55.07% precision, using a logistic regression model, after training the model for an hour. Zheng and Jin observed that when compared with Logistic Regression, Bayesian Network, and a Simple Neural Network, a Support Vector Machine having radial kernel gave them the most satisfactory results [69]. Due to their limited processing power, they were only able to use a subset of their data for training their model and recommended that a more powerful processor be used to achieve better results. Similar recommendations were made by G. Chen and et al., stating that their preferred model, the Long Short-Term Memory (LSTM), would have performed better were they able to train the different layers and neurons using higher computing power [65]. Since the data was non-linear in nature, a Recurrent Neural Network (RNN) would be more suited to the task.

In [43] it was discussed that when performing stock price prediction, it came out to be that ANN the algorithm that was once popular for prediction suffers from overfitting due to large numbers of parameters that it needs to fix [14]. This is where support vector machine (SVM) came into play and, it was suggested, that this method could be used as an alternative to avoid such limitations, where according to the VC theory [42] SVM calculates globally obtained sol unlike the ones obtained through ANN which mostly tend to fall in the local minima. It was seen that using an SVM model the accuracy of the predicted output came out to be around 57% [12]. There is one other form of SVM and that is LS-SVM (Least squared support vector machine). In the paper [58] it was mentioned that if the input parameters of LS-SVM is tuned and refined then the output of this classification algorithm boosts even further and shows promise to be a very powerful method to keep an eye out for. SVM being this powerful and popular as is it, is now almost always taken into consideration when it comes to predicting price of a volatile market, and thus we think that incorporating this into our research will boost our chances of getting a positive result.

While classical regression was more commonly used back in the day, non-linear machine learning algorithms are also increasingly being used as trading data regarded as time-series data which is non-stationary in nature. However, Artificial Neural Networks and SVM remain among the most popular methods used today. Every algorithm has a unique learning process. ANN simulates the workings of a human brain by creating a network of neurons [59]. The Hidden Markov Model (HMM), Artificial Neural Networks (ANN) as well as Genetic Algorithms (GA) were combined into one fusion model in order to predict market behaviour [18]. The stock prices converted to distinct value sets using ANN, which then became the input for the HMM. Using a selection of features determined from ARIMA analyses, Wang and Leu [5] designed a prediction model which was helpful in predicting market trends in the Taiwanese stock market. This produced an acceptable level of accuracy in predicting market trends of up to 6 weeks, after the networks were trained using 4-year weekly data [59]. A hybridized soft computing algorithm was defined by Abraham and et al. for automatic market predictions and pattern analysis [8]. They made used the Nasdaq-100 index of the Nasdaq stock market for forecasting a day ahead with neural networks. A neuro-fuzzy system was used to analyze the predicted values. This system produced promising results. A PNN (probabilistic neural network) model was trained using historical data by Chen and et al. for investment purposes [11]. When set against other investment strategies, namely the buy and hold concept and those which made use of forecasts estimated by the random walk and parametric Gaussian Mixture Model, PNN-based investment strategies produced better results. By searching a higher dimension hyperplane, a well-known SVM algorithm which separates classes was developed by Vapnik [6]. To test the predictability of price trends in the NIKKEI 255 index, Wang and et al. used a SVM to make forecasts [16]. They also made comparisons with other methods of classification, such as Elman Backpropagation Neural Networks, Quadratic Discriminant Analysis, and Linear Discriminant Analysis, SVM produced better experimental results. Kim compared the use of SVM to predict the daily stock price direction against Case Based Reasoning and neural network in the Korean stock market [12]. The initial attributes were made up of twelve technical indicators. SVM was proven to have produced better results.

Ensemble methods such as random forests help to reduce the probability of the data overfitting. Random forests use decision trees and majority voting to obtain reliable results. In order to perform an analysis on stock returns, Lin and et al. tested a prediction model that used the classifier ensemble method [32] and took bagging and majority voting methods into consideration. It was found that models using single classifiers under-performed compared to the ones using multiple classifiers, in regards to ROI and accuracy when the performances of those using an ensemble of several classifiers and those using single baseline classifiers were compared [59]. An SVM ensemble based Financial Distress Prediction (FDP) was a new method proposed by Sun and Li [37]. Both individual performance and diversity analysis were used in selecting the base classifiers from potential candidates for the SVM ensemble. The SVM ensemble produced superior results when compared to the individual SVM classifier. A sum of ten data mining techniques, some of which included KNN, Naive Bayes using kernel estimation, Linear Discriminant Analysis (LDA), Least Squared SVM, were used by Ou and Wang to try and forecast price fluctuations in the stock market of Hong Kong [24]. The SVM and LS-SVM were shown to produce better predictions compared to the other models.

The approach taken by an algorithm when it comes to predicting changes in the stock market is unique to each algorithm, as discussed above. Likewise, each algorithm also has its own unique set of limitations to be considered. Moreover, it has to be noted that the output depends not only on your choice of algorithm, but also the representation of input. The prediction accuracy can thus be improved by identifying and using a set of important classifiers instead of all of them. By putting together support vector regression(SVR) and a self-organizing map(SOM), Hsu and his fellow researchers designed a two-layer architecture [22]. The input environment was split into spaces where data points clumped together in order to properly dissect the non-linear nature of financial data, using the SOM. The SVR was run once the heterogeneous data was transformed into several homogeneous regions, in order to make predictions. The two stage architecture model yielded potentially significant results for the purposes of predicting stock prices. Variants of Genetic Programming (GP) have also been tried for modelling financial markets. To ensure generalization of the model, the model was further added with Multi Expression Programming and Gene Expression Programming, boosting and deep learning methods [59]. While trying to model the stocks in NYSE (New York Stock Exchange) Garg and et al. analyzed to what degree model selection criteria had on the performance [40]. the FPE criteria was proven as the better fit for the GP model than other model selection criteria, as indicated by the results. In order to make predictions about the closing value of five international stock indices, Nair et al. made use of an adaptive neural network [31]. The genetic algorithm helped the system adapt to the dynamic market conditions by making fine adjustments to the neural network parameters after each trade [59]. Using different neural networks models, trained with 14 years of data from NSE Nifty and BSE Sensex, Mantri and et al. tried to calculate volatilities of the Indian stock market [46]. They came to the conclusion that, using the models mentioned previously, having no distinction in regards to volatility of Nifty and Sensex estimated [59]. Mishra and et al. tested the rate of returns series for the existence of nonlinear dependency and chaos, for 6 Indian market indices [30]. The research indicated that random walk process was not followed by the returns [59]. To analyze and predict variations in price, Liu and Wang implemented an improved NN model by making the assumption which was an investor's purchasing decision relies on historical stock market data [35]. Araújo and Ferreira proposed the Morphological Rank Linear Forecasting model, and compared their results with that of Time-delay

Added Evolutionary Forecasting and Multilayer Perceptron networks methods [39].

2.1 Background Study on Multi Criteria Decision Theory in stock trading

Applications of TOPSIS in can be seen in many domains by various authors. Many of them are used in the financial sector which include the Transport sector, Banking sector and the Trading sector[33]. In the paper [7] Feng and Wang developed a model to evaluate Taiwanese domestic airlines based on financial ratios. The TOPSIS method was proven effective in the evaluation for airlines. By using the financial features and technical indicators, Tien-Chin and Hsu in there proposed model [15] ranked 10 computer producing companies in the Stock Market of Taiwan. The TOPSIS method was used with the weights generated from using entropy method, to sort the companies based on the relative performance. Demirelli in [62] computed the performance of commercial banks by using TOPSIS from the financial features of 2001-2007 in Turkey. However in this paper, equal weights were given to the financial ratios in performance calculation.

The paper [57] intends to propose a multi-rules dynamic model to quantify and analyze the budgetary presentation of thirteen innovation firms exchanging Istanbul Stock Exchange. These organizations are inspected and evaluated as far as ten budgetary proportions which are consolidated to acquire a monetary exhibition score by utilizing Technique for Order Preference by Similarity to Ideal Solution Methods (TOPSIS). TOPSIS assists with positioning these organizations for three-year timespan somewhere in the range of 2009 and 2011. Study will see if the positioning consequences of TOPSIS and the positioning aftereffects of the organizations market an incentive being referred to cover or not. n this paper, the monetary information of thirteen innovation organizations, which are recorded in ISEM for three-year timeframe somewhere in the range of 2009 and 2011, are utilized. Most importantly, ten budgetary proportions as measures are determined from their equalization and income sheet for each organizations by utilizing a proportion examination strategy. At that point, choice networks (13 x 10) are shaped independently for the 2009, 2010 and 2011 years by utilizing determined ten monetary proportions, for example, Return on Equity, Acid Test Ratio, Return on Assets, Current Assets Turnover, Total Debt Ratio, Fixed Assets Turnover, Net Profit Margin, Current Ratio, Working Capital Turnover and Debt Equity Ratio and thirteen choice focuses (firms). Afterward, equivalent loads are given to all the rules as the proportions have same noteworthiness w.r.t this paper and semantic factors are not utilized. In conclusion, the positioning is finished utilizing TOPSIS.

The paper [45] had the methodologies, TOPSIS and DEA, have been acquainted with rate dynamic organizations in concrete industry acknowledged in Tehran securities exchange. The methodology embraced in this paper is appropriate and completed during 2006-2011 and the number of inhabitants in the exploration remembers acknowledged organizations for securities exchange in concrete industry (28 organizations) and toward the end an exact positioning of the organizations is introduced by integrative procedures. For this paper, TOPSIS and DEA were incorporated yet DEA examination doesn't think about the underlying estimations of information sources and yields factors subsequently the last estimation of target capacity will be faulty. Settling this predicament in this examination we utilized TOPSIS's to think about starting estimations of data sources and yields factors. There is consistently the chance of changes in weight (with respect to changes of the specialists test) at that point the got loads may contrast altogether. This mistake may bring about miss productivity calculation separately. Subsequently it is basic to use measurable certainty span procedures to eliminate the imperfection and control all loads. The essential load of factors dictated by 12 industry specialists and afterward the standard deviation of weight additionally registered to direct certainty level for info and yield factors say something.

In [21] it is assessed execution of Turkish Sort A common assets and benefits stock assets by utilizing TOPSIS technique which is a multicriteria dynamic methodology. Both of these assets make out of stocks in their portfolios, so it tends to be empowered to analyze one another. It is utilized customary execution estimation strategies of assets like Sharpe proportion, Sortino proportion, Treynor record and Jensen's alpha. TOPSIS strategy thinks about these reserve execution estimation strategies and gives more sensible execution estimation. Information utilized in this examination incorporates month to month returns of 11 Kind A stock shared assets and 11 Benefits stock common assets in January 2007-December 2008 investigation period. In this paper, ISE 100 Public File is utilized for benchmark to benefit stock assets and Type A stock assets. It is assumed fitting for annuity stock assets in "Singular Benefits Framework Progress Report 2008" which is set up by Annuity Observing Center. To affirm the propriety of this benchmark for Type A stock assets, connection investigation is utilized. It is discovered that connection coefficient of Type A stock subsidizes' profits between ISE 100 Public List returns is normal 0.90. Along these lines, ISE 100 public List is a proper benchmark for Type A stock assets. Month to month shutting costs of ISE 100 public Record are received from [76]. [44] TOPSIS method was used to analyze financial statements of Istanbul's largest conglomerates which are traded on Istanbul Stock Exchange (ISE). financial performance scores were given to these conglomerates by using TOPSIS method on the nineteen financial ratio calculated, and then sorting of the the conglomerates was done based on the results. in order to predict future behaviors Kazan and et al., compared the Financial performance scores of these conglomerates. Multi Objective Decision Making (MODM) is employed in research of many industries to conduct performance evaluation and identify the best choice. The initial use of MODM was in the in areas operations research and theory of decision making. The approach was later applied to solve the financing problems as well. TOPSIS Method has the ability to incorporate qualitative and quantitative data, there for it provides ease of use and is has been performing a crucial role in analyzing different criteria to generate performance index or score.

According to [57], there have been countless sectoral concentrates in homegrown and unfamiliar writing executed utilizing the TOPSIS technique, in spite of the fact that there have been basically no investigations led on the REIT Business utilizing the strategy. By far most of studies in the budgetary areas are identified with the banking industry. As we will mirror the monetary or budgetary part of REITs in this investigation, we have just evaluated articles in which the money related execution of the firm was surveyed utilizing TOPSIS. At the end of the day, non-money related execution measures were not considered in this examination. In this investigation, a budgetary examination of REITs between 2011Q1-2014Q3 inside the monetary market in Turkey was estimated utilizing Entropy based TOPSIS (Strategy for Request Inclination by Likeness to An Ideal Arrangement) which is a generally utilized Multi-Models Dynamic (MCDM) technique. As indicated by the experimental outcomes, for all the periods, Avrasya, Akmerkez, Sinpaş, Kiler and Iş were discovered to be the most effective REITs individually, while, Optimist, Atakule, Alarko, Nurol and Vakıf exhibited the most exceedingly awful money related exhibitions all through the entire time frame.

2.2 Background Study of Feature Selection in stock trading

The study[41] measured twelve technical indicators for further investigation using data from the Shanghai Stock Exchange Composite Index (SSECI) from March 24, 1997 to August 23, 2006. The stock market's input variables were chosen from a total of 12 indicators. SMA, EMA, ALF (Alexander's filter), Relative Strength, RSI, MFI, percent B Indicator, Volatility, Volatility Band, CHO (Chaikin Oscillator), MACD (Moving Average Convergence-Divergence), percent K Indicator, Accumulation and distribution (AD) oscillator, and Williams percent R indicator are some of the indicators used. Then, PCA (Principal Component Analysis), Genetic Algorithm, and Sequential Forward Feature Selection methods to select which features for optimal investment. However, The paper did not include any resulting analysis or graphical representations of the results.

Yuan and et al.[84] selected 60 features for their prediction. The data comes from the Chinese A-share market and dates from January 1, 2010 to January 1, 2018. The algorithms used for prediction were Support Vector Machine(SVM), Artificial Neural Networks(ANN) and Random Forest. For the Feature selection, the paper used Recursive Feature Elimination (RFE) and Random Forest Feature selection using the information gain values. The Random Forest(RF) for feature selection and RF model for prediction has the greatest annualized return when it picks the top 1% of companies, with a 29.51 percent annualized return. The RF-RF model's profitability is further investigated using the stratified back-testing technique, and the new long-short portfolio's annualized return from 2011 to 2018 is 21.92 percent, with a maximum drawdown of just 13.58 percent. This profit is not substantial for proving the success of their model because better result can be achieved.

To decrease the cost of training time and increase prediction accuracies, the work[23] of Hunag and et al. combined the Support Vector Regressor (SVR) with the self-organizing feature map (SOFM) method and a filter-based feature selection. Thirteen technical indicators were used as input variables to forecast the daily price in the Taiwan index futures (FITX) in order to forecast the price index for the next day. The SOFM-SVR with feature selection had a Mean Absolute Percentage Error (MAPE) of 1.7726 percent, which is higher than the single SVR with feature selection and the one without feature selection. However, They did not test their strategy in the real stock market which would further evaluate their model's actual performance.

Barak and et al proposed a hybrid feature selection method[53] using ANFIS (Adaptive Neural Fuzzy Inference System) and the ICA (Imperialist Competitive Algorithm) is used to choose the most suitable features. The trading signals generated by

the model achieved superior outcomes with 87 percent prediction accuracy, and the wrapper features selection achieves a 12 percent increase in predictive performance over the basic research. Furthermore, since wrapper-based feature selection models are much more time-consuming, the results of our wrapper ANFIS-ICA method are better in terms of reducing time and improving prediction accuracy when compared to other algorithms like the wrapper Genetic algorithm (GA). However, they worked on only 24 feature at max and did not test implement a long-short strategy.
The research[73] by Nti and et al. used Random Forest (RF) with an improved leave-one-out cross-validation strategy and a Long Short-Term Memory (LSTM) Network to evaluate the degree of importance between various sectors stock-price and MVs and forecasted a 30-day had stock-price. From January 2002 to December 2018, the research dataset was acquired from the GSE official website, and the 42 macroeconomic indicators dataset was collected from the Bank of Ghana (BoG) official website. The LSTM model performed better than the baseline ARIMA model. But real-time trading was not done in this study.
The paper[85] used 12 technical features. The features are selected using decision tree algorithm based on wrapper feature selection. The paper uses the chronological penguin Levenberg–Marquardt-based nonlinear autoregressive network (CPLM-based NARX) for prediction. The suggested paper showed that CPLM-based NARX outperformed the competition in terms of MAPE and RMSE, with values of 0.96 and 0.805, respectively in comparison with the Regression model, Deep Belief Network (DBN), and NeuroFuzzy-Neural Network. This study does not analyze between the different timeframes of each technical feature and only uses 12 technical features.
Principal Component Analysis (PCA), Genetic Algorithms (GA), and Decision Trees (CART) are all compared in the research article[26] by Tsai and et al. It examines their prediction accuracy and mistakes by combining them using union, intersection, and multi intersection methods. The findings of the experiments indicate that integrating several feature selection techniques may improve prediction performance over single feature selection methods. The intersection of PCA and GA, as well as the multi-intersection of PCA, GA, and CART, perform the best, with accuracy rates of 79 percent and 78.98 percent, respectively.
The causal feature selection (CFS) method is proposed in this research[51] by Zhang and et al., to choose more representative features for improved stock prediction modeling. Comparative tests were performed between CFS and three well-known feature selection methods, namely principle component analysis (PCA), decision trees (DT; CART), and the least absolute shrinkage and selection operator, using 13-year data from the Shanghai Stock Exchanges (LASSO). When coupled with each of the seven baseline models, CFS performs best in terms of accuracy and precision in most instances, and finds 18 key consistent characteristics out the the 50 initial input features given.

Chapter 3

Dynamic feature Selection over multiple time-frames for efficient stock trading

The goal of this research is to adequately describe a prediction model that is able to predict stock market trends with sufficient accuracy and profitability. While there have been many attempts at doing this in the past, we wanted to find out if we could do this using the resources available to us, free of cost. This thesis reflects our research on the stock market, various ML algorithms used to predict stock market trends in the past, and the specific features, classifiers, and datasets needed to do so accurately.
to select 40 best time-variant from these features. Among those 40 features we further used.

3.1 Real-time Stock Trading Strategy

The system is built using a Quantopian working environment. It provides a large variety of financial data of major US stocks, starting from 2002 to the current date. Quantopian has many factors at its disposal for us to use and is also flexible enough to let us create our CustomFactor. These factors are a necessity for predicting the future market price using any Machine learning algorithm.

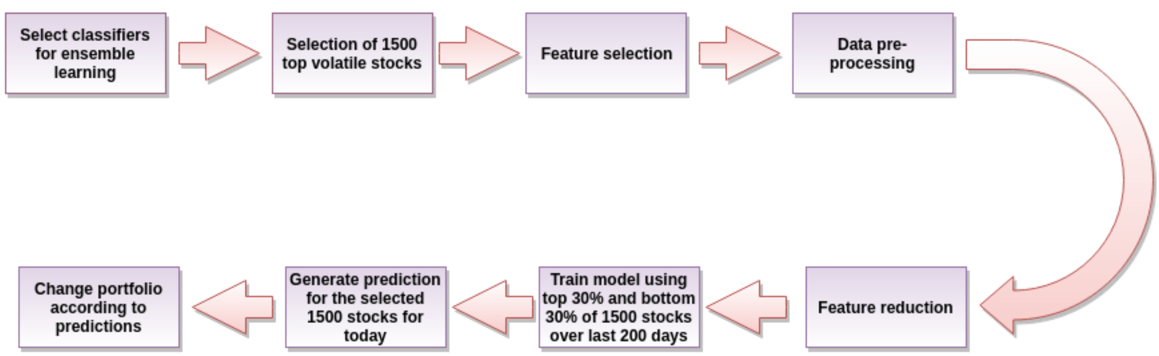


Figure 3.1: Ovreall System model for Real-time trading

The diagram in figure 3.1 shows the complete workflow of our model. The processes include selecting stocks, Feature selection, Data pre-processing, training and generating predictions using machine learning algorithms, and finally making changes to the portfolio according to the predictions.

Using the Quantopian algorithm feature, we implemented our strategy incorporating several machine learning algorithms into it. We started initially by collecting the data of 1500 of the top stocks in the market using the Q1500US function provided by Quantopian. Next, we imported all the factors provided whilst also including factors from TA-LIB (a significant financial factor provider). We also had to implement some custom factors ourselves. Some of the factors are Asset Growth 3M, Asset to Equity Ratio, Capex to Cash Flows, EBIT to Assets, EBITDA Yield, Earnings Quality, MACD Signal Line, Mean Reversion 1M, AD, ADX, APO, ATR, BETA, MFI etc. resulting in a total of 148 features.

But what we need to realize that due to the volatility of the market, just feeding all the factors into the ML algorithm won't give a very consistent result overall because a given factor can have both a positive or a negative effect on the prediction itself at the different given time concerning the market. To overcome this problem, we had to implement feature reduction dynamically, which is discussed in the 3.3 section. After the initial feature selection we further select the top 25 features based on the F-value of ANOVA.

After collecting the features, we set the number of stocks we wanted to trade, Machine learning window length, Nth forward day we wanted to predict, i.e. in this case of weekly trading, we had the variable set to 5, and also the trading frequency, i.e. the number of days after which we wanted to initiate the trade. From the 1500 stock data that we imported before, we sort and only trade on two different quantiles, upper 30% and lower 30%. We perform this slicing to make sure that we do not trade on stocks that have a very steady rate of change on their pricing, but only trade on stocks that are placed higher and lower down the ladder on which we could go long(upper 30%) and short(lower 30%) and have a significant success rate on it. We set the upper 30% to 1 i.e. long, lower 30% to -1 i.e. short and other 40% to 0 i.e. we do not perform any trade on them. The summation of these upper and lower quantiles results in 500 stocks, i.e. the number we set earlier. We had to strip the Label (Returns) from the zipline and perform a five-day computation on it. The T - 5 days data had to be discarded because there are no five days in forwarding time data for that given particular time resulting in NAN labels and thus was dropped from the zipline data frame. The Label column had to be kept separate to pass it onto the ML algorithm. Since Quantopian does not support machine learning and data preprocessing to be done inside the pipeline, we had to sort its entirety outside.

After preprocessing of the data, we make a new column in the pipeline called ML and call the Machine learning function to fill it up for each and every stock for that given day. The parameters of the ML function is the universe and all the columns of the pipeline, i.e. factors and label that we calculated. Here is the part where we perform the factor reduction that was talked about earlier. This process is performed dynamically throughout the training process, i.e. every single time we train the algorithm, we only train it with the top 25 features using the SelectKBest feature selection method.

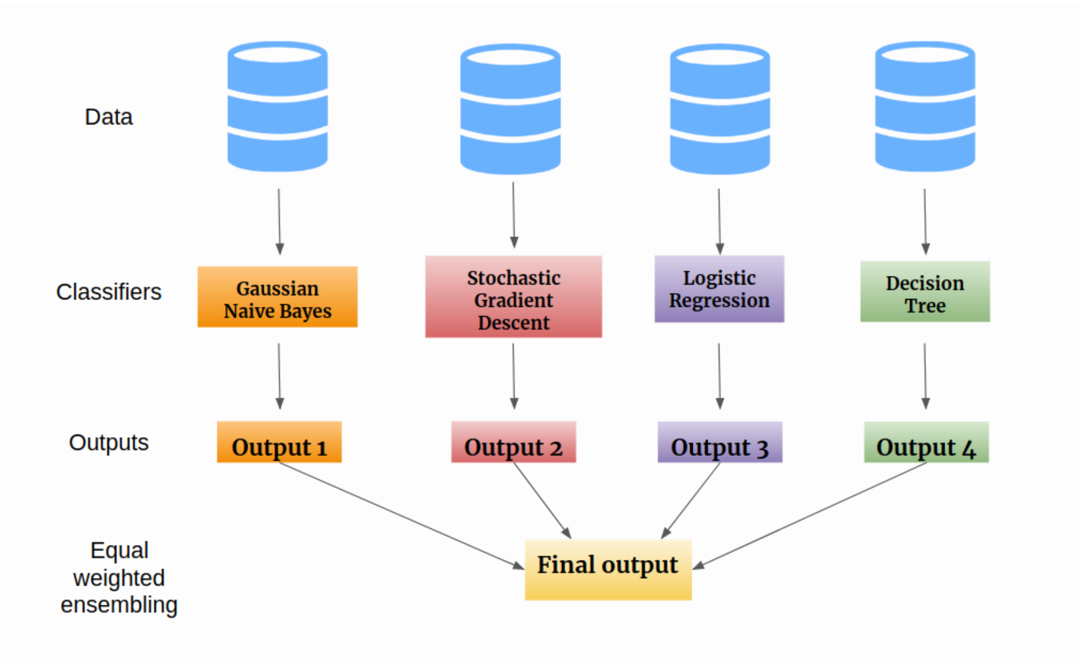


Figure 3.2: Structure of the Machine Learning Model Used

Figure 3.2 illustrates our Ensemble learning model. Here four machine learning algorithm each give us a hypothesis which each gives us an output. These four outputs are used for equal-weighted ensembling to generate the final output. An initial problem that we faced while implementing ML algo is that whenever we tried to have three or more high complexity classifiers(SVM, AdaBoost etc.) along with the dynamic feature selection, we use to run into TLE. But then we selected algorithms that have very small runtime, usually around 2-3 seconds to test and train, which is a very important thing to have in a live trading algorithm.

3.2 Data for Real-time Trading

3.2.1 Primary Features

Our primary data is the daily (1 day) open, low, high, close, volume of each stock in our tradable universe. Additionally we are retrieving daily balance sheet, cash flow statement, income statement, operation ratios, earning report, valuation, valuation ratios are taken as primary data. Using these primary data the 4* secondary factors were created. These factors are commonly used in financial prediction by traders.
bs = morningstar.balanceSheet
cfs = morningstar.cashFlowStatement
is = morningstar.incomeStatement
or = morningstar.operationRatios

er = morningstar.earningsReport
v = morningstar.valuation
vr = morningstar.valuationRatios

3.2.2 Secondary Features

Balance_sheet :- Balance sheet is a very important financial statement and is both financial modelling and accounting. It is used to portray a company's total assets. The balance sheet is usually calculated using the equation as follows.

$$\begin{aligned}
&A = L - SE\\
&where:\\
&A = Assets\\
&L = Liabilities\\
&SE = Shareholders\,Equity
\end{aligned} \tag{3.1}$$

Cash Flow Statement(CFS) It is a measure of how a company manages its financial strength and liquidity. It has a very high correlation ship with balance sheet and income statement. It can be used to analyze a company.
Income Statement: It is used for reporting a company's financial status over a specific accounting period. It summarizes a company's total returns, expenses over a period of time.

$$\begin{aligned}
&Net\,Income = \mathrm{R} + \mathrm{G} - \mathrm{E} + \mathrm{LE}\\
&where:\\
&\mathrm{R} = revenue\\
&\mathrm{G} = gains\\
&E = expenses\\
&\mathrm{LE} = losses\,equity
\end{aligned} \tag{3.2}$$

Operating Ratio: It shows the economy of a company by comparing total operating expenses to company net sales. The less the ratio the more efficient the company is at generating revenue.

$$\begin{aligned}
&OperatingRatio = \frac{OE+CG}{\text{Net sales}}\\
&where:\\
&\mathrm{OE} = Operating\,expenses\\
&\mathrm{CG} = cost\,of\,goods\,sold
\end{aligned} \tag{3.3}$$

Earnings Report: It is a quarterly earnings report made by companies to report their companies.
Valuation:- It is used to determine the current worth of assets of a company.
ADX and DX: It is a technical index used to indicate the strength of the trade. This strength can either be positive or negative and this is shown by two indicators +DI and -DI thus ADX commonly includes 3 separate lines. Additionally, It is a technical indicator that is used to predict the divergence side of the market. The

two components of DMI are +DI and -DI.

$$\begin{aligned}&\mathrm{DI}_{\text{Plus}} = \left(\tfrac{\text{Smoothed } +\text{DM}}{\text{ATR}}\right) \times 100\\&\mathrm{DI}_{\text{Minus}} = \left(\tfrac{\text{Smoothed } -\text{DM}}{\text{ATR}}\right) \times 100\\&\mathrm{DX} = \left(\tfrac{|\mathrm{DI}_{\text{Plus}}-\mathrm{DI}_{\text{Minus}}|}{|\mathrm{DI}_{\text{Plus}}+\mathrm{DI}_{\text{Minus}}|}\right) \times 100\\&\mathrm{ADX} = \tfrac{(\text{ Prior ADX}\times 13)+\text{ Current ADX}}{14}\end{aligned} \tag{3.4}$$

APO: It finds the absolute value and finds the difference between two different exponential moving averages . When the APO indicator goes above zero we go long i.e. Bullish and below zero we go short i.e. bearish.

$$\begin{aligned}&APO = FEMA - SEA\\&where:\\&FEMA = Fast\,Exponential\,Moving\,Average\\&SEA = Slow\,Exponential\,Average\end{aligned} \tag{3.5}$$

Mean Revision Theory: It is used to statistically analyze the market condition, which can overall affect the trading strategy. Mean revision also takes advantage of extreme price fluctuation of particular stocks. They can be applied for both buying and selling strategies.

$$\begin{aligned}&Mean\,revision = (\mathrm{MR} - \mathrm{Mean(MR)}) - \mathrm{Std(MR)}\\&where:\\&\mathrm{MR} = Monthly\,returns\\&Std = Standard\,deviation\end{aligned} \tag{3.6}$$

CMO: It is very similar to other similar momentum oscillators. It calculates momentum for both Market Up and Down days but it does not smooth out the results. The oscillator indicates between +100 and -100.

$$\begin{aligned}&\text{Chande Momentum Oscillator } = \tfrac{sH-sL}{sH+sL} \times 100\\&\text{where:}\\&sH = \text{ high close summation in N periods}\\&sL = \text{ low close summation in N periods}\end{aligned} \tag{3.7}$$

Returns

It is an indication of total money made or lost during transactions. Returns can be expressed as a ratio of profit to investment.

$$\text{Rate of Returns} = \frac{\text{current value-initial value}}{\text{initial value}} \times 100 \tag{3.8}$$

Williams %R It is known as Williams perfect Range, which is a type of momentum calculator that has an indicator range between 0 to -100 and measures the level of overbought and oversold. It is used to find the most optimal time for entry and exit the market.

$$
\begin{aligned}
&\text{Wiliams } \%R = \tfrac{\text{Highest High} - \text{Close}}{\text{Highest High} - \text{Lowest Low}}\\
&\text{where}\\
&\text{Highest High } = \text{ Peak price in the lookback}\\
&\text{time period, typically 2 weeks.}\\
&\text{Close} = \text{Latest closing price.}\\
&\text{Lowest Low} = \text{trough level price in the lookback}\\
&\text{time period, typically 2 weeks.}
\end{aligned}
\tag{3.9}
$$

ATR: It measures the market volatility, by dissolving the entire range of an asset price. Stocks with higher volatility has higher ATR and vice versa. This acts as an indicator for traders to exit and enter a trade.

$$
\begin{aligned}
&TY = \max\left[(H-L), \left|H - C_{prev}\right|, \left|L - C_{prev}\right|\right]\\
&ATR = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} TR_i\\
&\\
&\textbf{where:}\\
&H = \text{ High}\\
&L = \text{ Low}\\
&C = \text{ Close}\\
&TR_i = \text{ a particular true range; and}\\
&n = \text{ the time period employed (usually 14 days)}
\end{aligned}
\tag{3.10}
$$

AD: This is an indicator that makes use of volume and price to determine if a stock is accumulated or distributed. This factor looks for changes between stock price and volume flow, thus providing a hint of how strong a trend is.
The Formula for the Accumulation / Distribution Indicator.
where:

$$
\begin{aligned}
\text{CMFV} &= \frac{(P_C - P_L) - (P_H - P_C)}{P_H - P_L} \times \text{V}\\
CMFV &= \text{ Current Money Flow Volume}\\
P_L &= \text{ Losing price}\\
P_L &= \text{ Low price for the period}\\
P_H &= \text{ High price for the period}\\
\text{V} &= \text{ Volume for the period}
\end{aligned}
\tag{3.11}
$$

BETA: A coefficient measure of volatility for an individual stock in contrast to the entire market. Statistically beta is the gradient of the line. By default the market beta is 1.0.

$$
\begin{aligned}
&\text{Beta coefficient } (\beta)=\frac{\text{Covariance } (R_e,R_m)}{\text{Variance}(R_m)} \\
&\text{where:} \\
&R_e= \text{ revenue from a stock} \\
&R_m= \text{ revenue from overall market} \\
&\text{Covariance} = \text{Correlation of returns of a stock to} \\
&\text{returns of the market} \\
&\text{Variance} = \text{Divergence of the market's value} \\
&\text{from average}
\end{aligned} \tag{3.12}
$$

MP: It calculates the mean of the high and low of a stock candle.

$$
MedPrice = (high(t) + \text{low}(t))/2 \tag{3.13}
$$

MFI: It is a technical indicator that makes use of price and volume as a reference to identify if a stock is overvalued or undervalued. It can be used to spot the change in the daily price of the stock. The value of the oscillator ranges between 0-100.

$$
\begin{aligned}
&\text{Money Flow Index } = 100-\frac{100}{1+\text{ Money Flow Ratio}} \\
&\text{where:} \\
&\text{Money Flow Ratio } = \frac{\text{14 Period Positive Money Flow}}{\text{14 Period Negative Money Flow}} \\
&\text{Raw Money Flow } = \text{ Common Price } * \text{ Volume} \\
&\text{Common Price } = \frac{(\text{ High + Low + Close })}{3}
\end{aligned} \tag{3.14}
$$

PPO: This is used to show the correlation ship between two Moving averages between 0-1. It compares asset benchmarks, market volatility to come up with trend signals and help predict the trend of the market.

$$
\begin{aligned}
&\text{PPO} = \frac{\text{12periodEMA}-\text{26periodEMA}}{\text{26periodEMA}} \times 100 \\
&\text{Signal Line } = 9 \text{ -period EMA of PPO} \\
&\text{PPO Histogram } = \text{PPO} - \text{ Signal Line} \\
&\\
&\textbf{Where:} \\
&\text{EMA} = \text{Exponential Moving Average}
\end{aligned} \tag{3.15}
$$

Asset to Equity Ratio: It shows the correlation between the assets owned by a firm to the total percentage of the shareholders. The higher the ratio the greater the firm's debt.

Capex to Cash Flow: This is used to estimate a company's long term assets and also how much cash a company is able to generate.

$$
\text{Cash to capital Expenditures} = \frac{\text{Cash Flow from Operation}}{\text{Capital Expenditure}} \tag{3.16}
$$

Asset Growth: It is the growth of the overall asset of a company.

$$
\text{Asset Growth} = \frac{\text{Asset value prior} - \text{Asset value current}}{\text{Asset value prior}} \times 100 \tag{3.17}
$$

EBIT to Asset: It is a sign of a company's benefits which is generated from operations and trades and ignores tax burden and capital structure.

$$\begin{array}{l} \text{EBIT = R - COGS - OE} \\ \text{Or} \\ \text{EBIT} = \text{NI} + \text{I} + \text{T} \\ \\ \textbf{where}\text{:} \\ \text{R} = \text{Revenue} \\ \text{NI} = \text{Net Income} \\ \text{I} = \text{Interest} \\ \text{T} = \text{Taxes} \\ \text{OE} = \text{Operating Expenses} \\ \text{COGS} = \text{Cost of goods sold} \end{array} \tag{3.18}$$

EBITDA Yield: It is usually reported as a Quarterly earnings press release. It ignores taxes and non-operating expenses thus highlighting only important for the market analyst to focus on.

MACD Signal Line: This indicator shows the relationship between different stock's moving averages. After calculation, a nine day EMA-line is drawn over MACD line to use as a buy or sell indicator.

$$\text{MACD} = \text{12periodEMA} - \text{26periodEMA} \tag{3.19}$$

Money Flow Volume: Money flow is an indicator of when and at which price the stock was purchased. If more security was bought during the uptick time compared to downtick then the downtick time then the indicator is positive because almost all the investors participating in the trade were willing to give a high price for the stock and vice versa.

Operating Cash Flows to Assets: It is the flock of revenue generated by a company's normal business operation. It is an indicator of whether a company can generate a substantial amount of positive cash flow to maintain its growth. This indicator gives the market analysts a clear view of what a company is capable of.

Return on Invest Capital: This gives the market analyst an indicator of how well a company uses its resources to generate its revenue.

$$\begin{array}{l} \text{ROIC} = \frac{\text{NOPAT}}{\text{Invested Capital}} \\ \text{where:} \\ \text{NOPAT} = \text{Net operating profit after tax} \end{array} \tag{3.20}$$

39 Week Returns: Gives us the total returns over a period of 39 weeks

$$\begin{array}{l} 39WeekReturns = \frac{R(T)-R(T-215)}{R(T)} \times 100 \\ where: \\ \text{R} = Returns \end{array} \tag{3.21}$$

Trend line: It shows the momentum / Trend of the market from one given point to another.

Volume 22 Days: It gives us the total amount of volume generated over a period of 22 days.

3.3 Proposed Feature Selection Model

The paper analyzes a total of 148 features based on four criteria.

- Returns Analysis
- Information Coefficient Analysis
- Turnover Analysis
- Grouped Analysis

For the analysis we used Alphalens on the stock data from start_date='2011-03-06'and end_date='2012-03-06'. The feature selection method is showed in figure 3.3. The method went on long and short positions on the top and bottom quantile or reverse if the feature is negative. The trading is done for 1D (1 day hold period), 5D (5 day hold period) and 22D(22 day hold period). The features must have "mean return" greater than 0.05% or 0.5 basis point for both the long and short positions selected to trade for that certain time period. The Information coefficient must be greater than 0.005. In the turnover analysis, the mean turnover must be greater than 0.25. The stocks satisfying these criteria will be initially selected. In addition, Sklearn's SelectKbest method is used to select the best features out of the selected feature. For the hyper-parameter, "f_classif" is selected, which ranks the features using the T-scores from ANOVA.

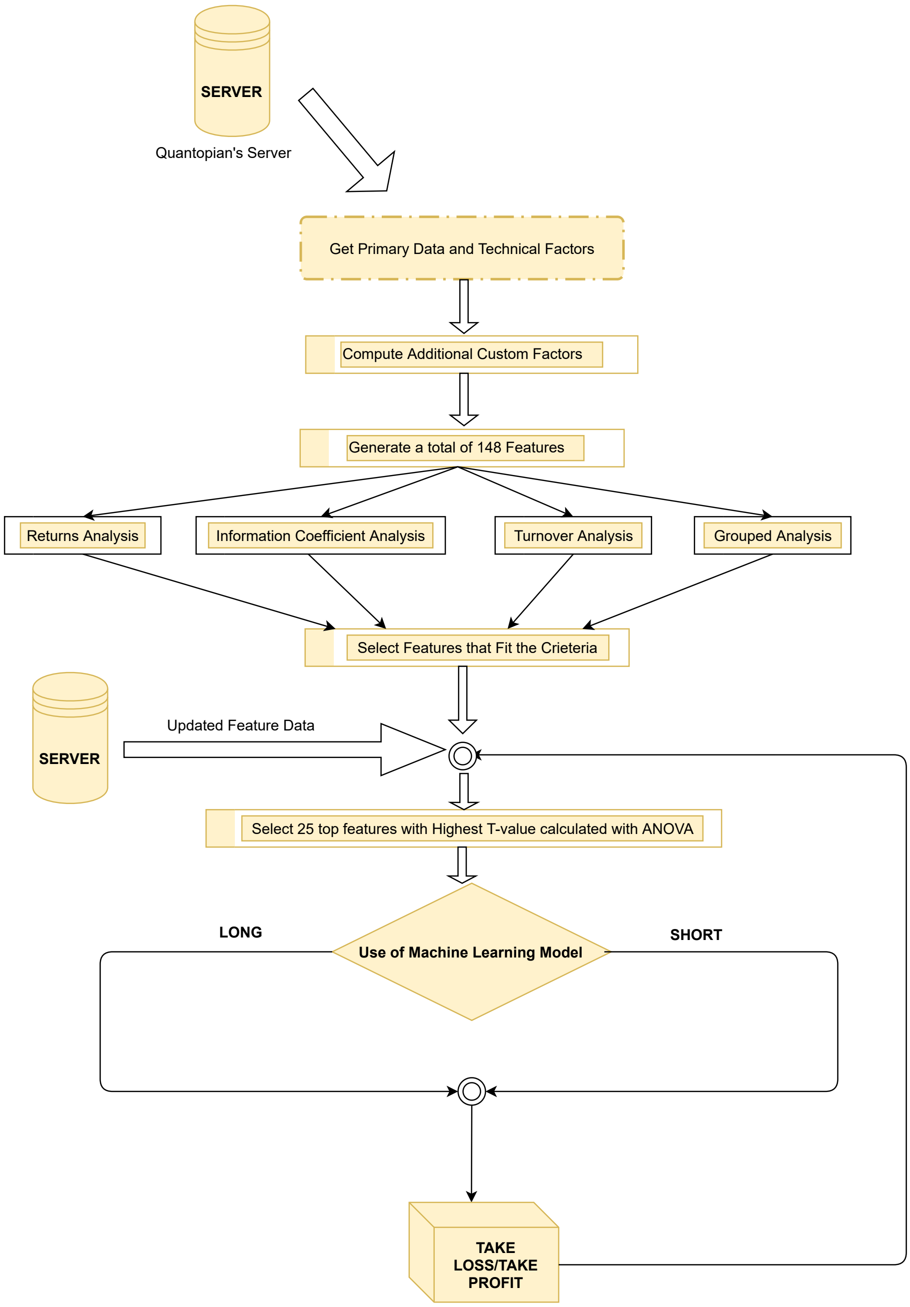


Figure 3.3: Proposed Feature Selection Model

The stock values were dvidied intro 3 equal quantile. The lower quantile, the middle quantile and the upper quantile.Each quantile had about 33.33% of the values. . Table 3.1 and 3.2 shows the selected features for daily, weekly and monthly trading.

Table 3.1: Selected Features for trading based on feature analysis for different time-frame (Part 1)

Feature Name	Trading Position			Feature Name	Trading Position		
	1D	1W	1M		1D	1W	1M
Asset_Growth_5D				BETA_6D			
Asset_Growth_2D	✓	✓	✓	BETA_1D			
Asset_Growth_5D				BETA_2D			
Asset_Growth_22D				BETA_5D			
Asset_To_Equity_Ratio	✓	✓	✓	BETA_22D	✓	✓	✓
Capex_To_Cashflows	✓	✓	✓	RSI_10D			✓
EBITDA_Yield		✓		BOP	✓		
EBIT_To_Assets	✓	✓	✓	CCI_14D			✓
Net_Income_Margin	✓	✓	✓	CCI_1D			
Return_On_Invest_Capital	✓	✓	✓	CCI_2D	✓		
Mean_Reversion_1M			✓	CCI_5D	✓	✓	
Mean_Reversion_2D	✓			CCI_22D			
Mean_Reversion_5D				CMO_15D			✓
Mean_Reversion_6D		✓		CMO_5D			
MACD_Signal_10d	✓	✓	✓	CMO_2D			
MACD_Signal_1d				CMO_22D		✓	
MACD_Signal_2d				DX_15D			
MACD_Signal_5d				DX_22D		✓	✓
MACD_Signal_22d				DX_2D	✓		
AD_14D				DX_5D			
AD_1D	✓			MAX			
AD_2D				MAXINDEX			
AD_5D		✓		MEDPRICE_1D	✓	✓	✓
AD_22D			✓	MEDPRICE_2D			
ADX_29D	✓		✓	MEDPRICE_5D			
ADX_1D				MEDPRICE_22D			
ADX_2D				MFI_15D			
ADX_5D				MFI_1D			
ADX_22D				MFI_2D			✓
APO_12D_26D				MFI_5D			
ATR_15D		✓	✓	MFI_22D	✓	✓	
ATR_2D	✓			MIDPOINT			
ATR_5D				MIN			
ATR_22D				MININDEX			

Table 3.2: Selected Features for trading based on feature analysis for different time-frame (Part 2)

Feature Name	Trading Position			Feature Name	Trading Position		
	1D	1W	1M		1D	1W	1M
MINUS_DI_15D			✓	WILLR_14D			✓
MINUS_DI_1D				WILLR_1D	✓		
MINUS_DI_2D				WILLR_2D			
MINUS_DI_5D				WILLR_5D			
MINUS_DI_22D				WILLR_22D		✓	
MINUS_DM_15D				Average_Dollar_Volume	✓	✓	✓
MINUS_DM_1D				Moneyflow_Volume_5D			
MINUS_DM_2D	✓			Moneyflow_Volume_1D			✓
MINUS_DM_5D				Moneyflow_Volume_2D		✓	
MINUS_DM_22D				Moneyflow_Volume_22D	✓		
PLUS_DI_15D				Annualized_Volatility	✓	✓	✓
PLUS_DI_1D				Operating_Cashflows_To_Assets	✓	✓	✓
PLUS_DI_2D	✓			Price_Momentum_3M	✓		✓
PLUS_DI_5D		✓	✓	Price_Oscillator_20D			
PLUS_DI_22D		✓		Price_Oscillator_1D	✓		
PLUS_DM_15D				Price_Oscillator_2D			
PLUS_DM_1D			✓	Price_Oscillator_5D			
PLUS_DM_2D				Price_Oscillator_22D			
PLUS_DM_5D				Returns_215D		✓	✓
PLUS_DM_22D				Returns_190D			
PPO_12D_26D	✓			Returns_160D			
PPO_8D_13D		✓		Returns_100D	✓		
PPO_1D_3D				Returns_50D			
PPO_24D_50D			✓	Returns_25D			
STDDEV				Trendline_252D	✓	✓	✓
TRANGE_2D		✓	✓	Trendline_25D			
TRANGE_1D				Trendline_50D			
TRANGE_5D	✓			Trendline_100D			
TRANGE_22D				Trendline_150D			
TYPPRICE_1D				Vol_3M			
TYPPRICE_2D				Vol_1D			
TYPPRICE_5D				Vol_2D			
TYPPRICE_22D				Vol_5D	✓		
Earnings_Quality				Vol_22D		✓	✓

3.3.1 Feature Evalutaion example for WILLR 14 Day

The same evaluation was done for all the features. For demonstrating purposes we only show the graphical results of the feature WILLR_14D.

1. **Mean Period Wise Return By Factor Quartile**

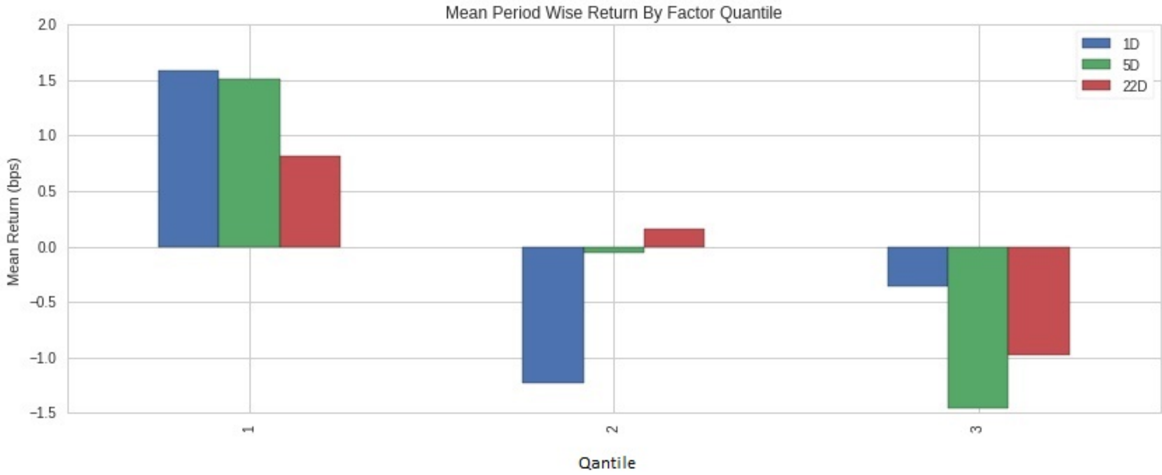


Figure 3.4: Mean Period Wise Return By Factor Quartile

The diagram in figure 3.4 represents the total return as per graph height.A positive graph represents long and negative short.In Our case we are taking 3 quartiles and breaking them into 3 separate days 1D 5D 22D, for which we can trade using the Quantopian environment. As per figure 3.4 what we can use this to observe for the factor taken works best with 5D trading as we have a good amount of return for both long and short, where as for 1d trading we can see that for the first quartile long gives a good result but is not the best for going short as displayed in the third quartile.

2. **Factor Weighted Long Short Portfolio Cumulative Return:**

   This graph represents the position of the portfolio of the trader given that person only traded taking that experimented factor into consideration alone. This represents the cumulative Returns on the portfolio of the trader.

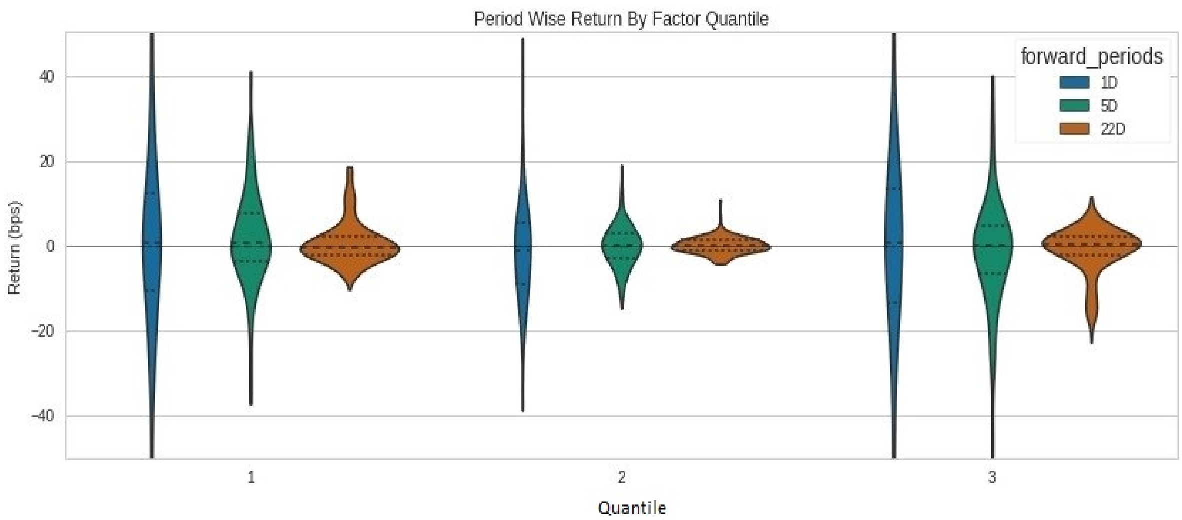
   The graphs in figure 3.5 display different positions of the portfolio given 3 different trading frequencies 1D, 5D, 22D as per quartile deceleration.

Figure 3.5: Factor Weighted Long Short Portfolio Cumulative Return (a)1D, (b)5D, (c)22D

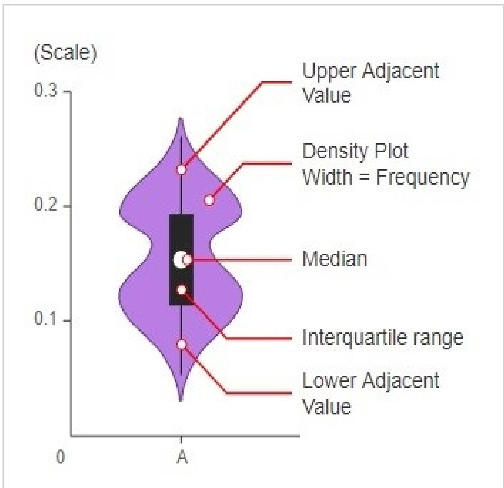
3. **Period Wise Return By Factor Quantile**

This graph is famously known as violin graph, and comes well in handy when only the median value is not a reliable option to use in order to judge the state of the data being experimented on.

This graph is very convenient when it comes to comparing the summary statistics of range of quartiles. The representation of this graph is very similar to that shown in the figure 3.6, but this time we get an idea of the density of where our returns are concentrated for each time period.



(a)



(b)

Figure 3.6: Period Wise Return By Factor Quantile

4. **Cumulative Returns by Quantile**

The cumulative quantiles of each time period are taken and are aggregated over the period of trading time. The main objective of this curve is to see of the quartiles spreads as far away from each other as possible. The far apart they are the better. The third quantile is very clearly above the first quartile and this gets more and more clearer as we move forward into time. The less overlapping between the graphs the better.

This is calculated for the 3 different quartiles over the period of time that we traded shown in figure 3.7.

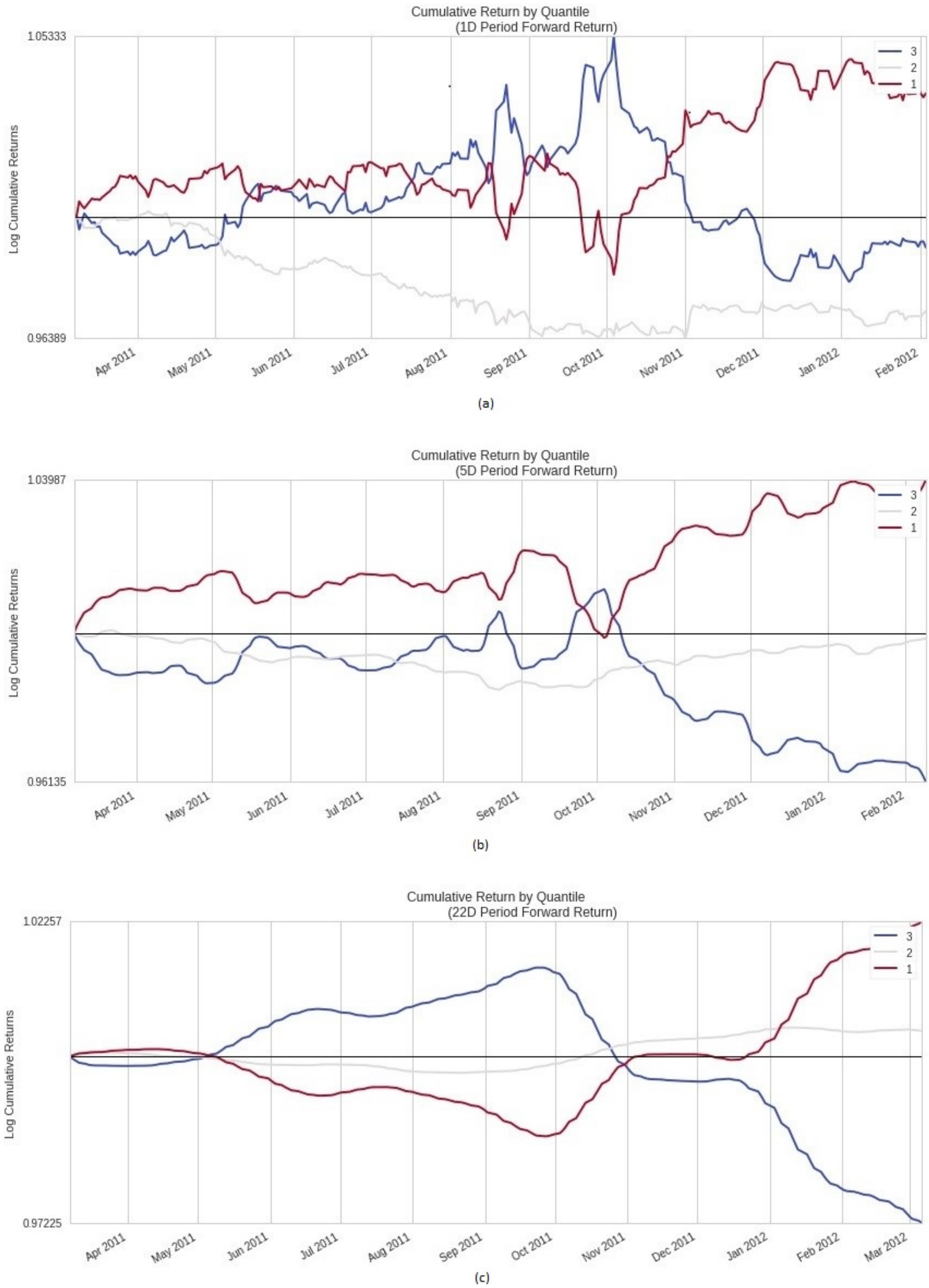


Figure 3.7: Cumulative Returns by Quantile (a)1D, (b)5D, (c)22D

5. **Top Minus Bottom Quantile Mean**

   This graph in figure 3.8 subtracts the top quantile from the bottom quantile and takes a mean of the answer to smoothen out the results for the given trading time period. The more positive the graph plot the more return we get over that period of trade time.

Figure 3.8: Top Minus Bottom Quantile Mean 1D, 5D, 22D

3.4 Machine Learning Algorithms

3.4.1 Naive Bayes

Naive Bayes classification uses Bayes' rule. Assuming Ck= a particular event x= feature vector

$$P\left(c_k|\mathbf{x}\right) = P\left(c_k\right) \times \frac{P\left(\mathbf{x}|c_k\right)}{P(\mathbf{x})} \tag{3.22}$$

Bayes' rule defines the probability of a particular event ck occurring for feature vector x, can be computed from the given formula.
For estimating $P(c_k|x)$ from a dataset we must first compute $P(c_k|x)$. The strategy used to find the distribution of x conditional on ck is specified by the following formula:

$$P\left(\mathbf{x}|c_k\right) = \prod_{j=1}^{d} P\left(x_j|c_k\right) \tag{3.23}$$

In this formula, we assume that x_j having a particular value is independent of the occurrence of any other x_j' from the x feature vector for a particular event c_k. By plugging in the estimates the equation 2 becomes:
Gaussian Naive Bayes is better for this case as the features are continuous. Whereas Bernoulli Naive Bayes works better when the features are binary.

3.4.2 Logistic Regression

Regression models are widely used for data-driven decision making. In many fields, logistic regression has become the standard method of data analysis in such situations. The key to any such kind of analysis is to find the model that fits best when explaining the relationship between a dependent and one or more independent variables. Unlike linear regression which most people are familiar with, where the outcome variable is usually continuous, a condition for logistic regression is that the outcome variable is binary. Logistic regression also allows us to determine to what degree a chosen independent variable affects the outcome.
Two reasons why logistic regression is so widely used is that it is 1) flexible and can be easily used in many situations, and 2) it allows for meaningful interpretations of the results. For simplicity, the quantity $\gamma(x) = E(Y given\, x)$ is used to represent Y's conditional mean, given a value x.
The logit transformation is integral for logistic regression. The transformation is as follows:

$$\begin{aligned} p(x) &= \ln\left[\frac{\gamma(x)}{1-\gamma(x)}\right] \\ &= \theta_0 + \theta_1 x \end{aligned} \tag{3.24}$$

The importance of the logit, p(x) lies in the fact that it contains many useful properties of Logistic regression. The logit takes linear vales which might be continuous, either positive or negative and depends on x range.
To summarize, when the outcome variable is dichotomous, in regression analysis:

- The logistic regression mean must be scaled to be between 1 and 0.
- The binomial, as opposed to the normal distribution, describes the distribution of errors.
- The principles of linear regression can also be applied to logistic regression.

For the model to produce reliable results, we need to have a large number of observations (at least 50).
In order to prevent overfitting of data, L1 (Lasso) and L2 (Ridge) regression is used. The difference between these two methods of regularization lie in the penalty term. L2 uses the "squared magnitude" as the penalty term to the loss function, while L2 uses the "absolute value of magnitude".

$$\underbrace{\textstyle\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\alpha_j\right)^2 + \gamma \sum_{j=1}^{p} \alpha_j^2}_{\text{cost function}} \tag{3.25}$$

$$\underbrace{\textstyle\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\alpha_j\right)^2 + \gamma \sum_{j=1}^{p} |\alpha_j|}_{\text{Cost function}} \tag{3.26}$$

The above equations show the cost function when using L2 and L1 regularization respectively where α is the weight put on a particular feature x and γ is the coefficient of the penalty term.

3.4.3 Stochastic Gradient Descent

It is a first-order optimizing supervised machine learning algorithm that specializes to fit a straight line over a series of data points with the least amount of error.

$$w_{t+1} = w_t - \gamma \frac{1}{n} \sum_{i=1}^{n} \nabla_w Q\left(z_i, w_t\right) \tag{3.27}$$

w is the weight we want to optimize with regards to cost $\nabla Q(z, w)$ where γ is the learning rate. It works by first assuming randomly a point of intercept to draw a straight line and for each individual weight of the graph we find the predicted Y value i.e. $\hat{y}$ using which we calculate the lingering or remaining value i.e. the difference between real y and $\hat{y}$ and then square the residual and square its value, and then find the sum of the squared residual for each and every Y value that exists for the X value. If we keep on increasing the intercept and for every intercept we get a sum of squared residual value, which if we plot a graph we will obtain a graph that looks similar to y = X^2. The maximum value of the graph will be the one that has the lowest squared residual. This is a very slow method but gradient descent uses this concept but works in a much faster way by taking big steps when it is far away from the optimal value and gradually decreases the step size when it gets closer. Gradient descent derives the sum of the squared residuals with respect to the intercept giving us the slope of the curve at that state of time. The closer we

get to intercept the closer the slope gets to 0. We then calculate the step size we multiply the slope that we got with αi.e. learning rate. Using the new intercept we got we then repeat the entire working process until we get a slope close to 0 or we reach our iterative limit, when we stop our algorithm.
Stochastic gradient descent on the other hand works very similarly but is very optimized and efficient when it comes to using it on large data sets.[25]
What stochastic gradient descent does is it picks random values from the weight and only uses that value to perform the entire working process and so on. Thus this reduces the calculation factor by F-1, here F is sum of the points. It also performs well when using it on data with a lot of redundancy, as it clusters them and only picks a random value from every single cluster to perform the working steps. Thus if there are 5 clusters stochastic gradient descent will pick 5 points to work with.
Thus stochastic gradient descent works well with stock prediction all whilst trading real time is it reduces the complexity of the algorithm by a whole lot thus not resulting in any TLE unlike gradient descent.
Hinge Loss : It is a loss function that is mainly used to train classifiers using the maximum margin classification.[13]
Elastic Net Penalty : Elastic net penalty that is mainly used to overcome the limitations Lasso regression. If there are highly correlated values lasso regression usually tends to pick one variable from the group that are highly correlated and ignore the rest but what elastic net does is it adds a squared penalty to it. Adding this term gives this loss function a unique minimum(strongly convex).

3.4.4 Support Vector Machine (SVM)

SVM is a classification and regression based algorithm. It is used to maximize predictive accuracy whilst avoiding the overfitting of data. It is used for applications such as handwriting, face, text and hypertext classification, Bioinformatics etc. SVM is used to achieve maximum separation between data points. Hyperplane is a part of SVM that maximize the separation of data points by increasing the line width with increments. It starts by drawing a line and two equidistant parallel lines. Next the algorithm picks a stopping point so that the algorithm does not run into an infinite loop and also picks an expanding factor close to 1 example 0.99. [17]

3.4.5 AdaBoost

A boosting algorithm increases the accuracy of weak learners. A weak learner as an algorithm that uses a simple assumption to output a hypothesis that comes from an easily learnable hypothesis class and performs at least slightly better than a random guess. If each weak learner is properly implemented then Boosting aggregates the weak hypotheses to provide a better predictor which will perform well on hard to learn problems.
Adaboost is the short form of Adaptive boosting. The AdaBoost algorithm outputs a “strong” function that is a weighted sum of weak classifiers. The algorithm follows an iterative process where in each iteration the algorithm focuses on the samples where the previous hypothesis gave incorrect answers. The weak learner is returns a weak function whose error is et such that

$$\epsilon_I \stackrel{\text{def}}{=} L_{\mathbf{D}^{(t)}}(h_t) \stackrel{\text{def}}{=} \sum_{i=1}^{m} D_i^{(t)} \mathbb{I}_{[h_t(\mathbf{x}_i) \neq \mathbf{y}_i]} \tag{3.28}$$

where L_D is the loss function and h is the hypothesis and then a specific classifier is assigned a weight for h_t as follows: $w_t = \frac{1}{2} \log \left(\frac{1}{\epsilon_t} - 1 \right)$. So, the weight given to that weak classifier is inversely proportional to the error of that weak classifier. [50]

3.4.6 Random Forest

A random forest uses a set of decision trees. As a class of decision trees of unspecified size has infinite VC dimension (Vapnik–Chervonenkis dimension), we restrict the size of the decision tree in order to prevent overfitting. Creating an ensemble of trees is another way to reduce the probability of overfitting [50].
An advantage of using Random Forest is that it both a classifier and regressor [9]. For our purposes, we applied Random Forest for classification. The algorithm works as follows:
Create a bootstrapped dataset from the original data (bagging). An important point about bootstrap samples is that the same sample can be chosen more than once, given they are chosen at random. Next, we create decision trees for each sample in our bootstrap dataset. At every tree node, choose a random subset of variables and the best split among those are chosen. We use the aggregate of the predictions of our trees in order to predict the classification of new data. For classification purposes, we use the majority vote. The average is used for regression.
We can then easily estimate the error in our results in the following manner: We take a sample from our original data, which was not used to create our decision trees. This sample is called an "Out of bag" (OOB) sample. We then try to predict the data of the out of bag sample using the tree we grew by applying bootstrapping. We then aggregate the predictions of the out of bag samples and calculate the rate of error. This is called the OOB estimate of the error rate.
Given enough trees have been grown, the OOB estimate of error rate is significantly accurate.

3.4.7 Decision Tree

Decision Tree Classifiers divides a problem into a collection of subproblems turning a complex problem easier to solve [8]. Using entropy as the criteria of splitting tress is useful when the problem contains numerous classes. The objective used for tree design in our model is to minimize uncertainty in each layer, or put differently increase entropy reduction. Shanon's entropy, defined as

$$H = \sum_i p_i \log p_i \tag{3.29}$$

Pi= a prior likelihood of class i

$$Gain(S, A) = Entropy(S) - \sum_{v \epsilon A} \frac{S_v}{S} Entropy(S_v) \tag{3.30}$$

Entropy is used to find the most gain of a particular factor. And the factor with the most gain is used to make a split. At the terminal nodes a decision is given about the classification.
One advantage of decision tree over other classifiers is that a training sample is tested for subsets of classes and not all classes. This reduces computation and improves performance of the classification task. The decision tree classifier also uses different subsets of the features of the given problem. This makes the classifier perform better than single layer classifiers. Decision tree classifier also overcomes high dimensionality problem as it takes limited factors for each decision tree.
Overlapping is one of the problems of using decision tree classifier. The classifier takes a large amount of time and space when the label class is large. But that is not the case in this system as the classification task is binary. There are a lot of difficulties involved in designing an ideal decision tree. Error may also add up in each level to reduce the accuracy of the model [3].

3.5 Live Trading Results

We used all the seven classifiers discussed to start to perform calculations on data from 2011-03-06 to 2011-09-7. We split by 80:20 ratio to form the train set and test set. Table 3.3 shows that ensemble methods work far better in this case. However, for ensemble methods, we only predicted the top and bottom values, as in real life we do not need to trade all the 1500 stocks. The ensemble 1 showing accuracy of 99.25% included LR, Gaussian_NB, Bernoulli_NB and SGDC whereas the ensemble 2 showing accuracy of 74.23% consisted of LR_L1Regress, LR_L2Regress, Gaussian_NB and Bernoulli_NB.

Table 3.3: Accuracy test on data from 1500 US stocks 2011-03-06 to 2011-09-7

Name of the Algorithm	Test accuracy
Naive Bayes(NB)	51.21 %
Logistic Regression(LR)	51.77 %
Stochastic Gradient Descent (SGDC)	50.56 %
Support Vector Machine (SVM)	54.06 %
Adaboost	53.29 %
Random Forest	52.43 %
Ensemble 1 (predict top and bottom)	99.25 %
Ensemble 2 (predict top and bottom)	74.23 %

3.5.1 Day Trading

- **RandomForest:** Using Random forest algorithm and daily trading we get a return of 18.08% with a sharpe ratio of 0.77.

- **AdaBoost:** Using AdaBoostClassifier in the mix we get a return of 11.69%

with a sharpe ratio of 0.49.

- **Ensemble 1 Classifiers:**

  1. GaussianNB
  2. LogisticRegression
  3. BernoulliNB
  4. Sgdc

  Using the mixed classiferes of all these algorithms together we get a return of 34.99% with a sharpe ratio of 0.67. Time complexity of all these algorithms combined is very less and thus is very feasible for our purpose.

- **Best Classifiers:**

  1. GaussianNB
  2. LogisticRegression
  3. DTC
  4. Sgdc

  Figure 3.9 shows, using Decision tree classifiers in the mix we get a return of 54.63% with a sharpe ratio of 1.16%.

  1. GaussianNB
  2. LogisticRegression
  3. AdaBoostClassifier
  4. Sgdc

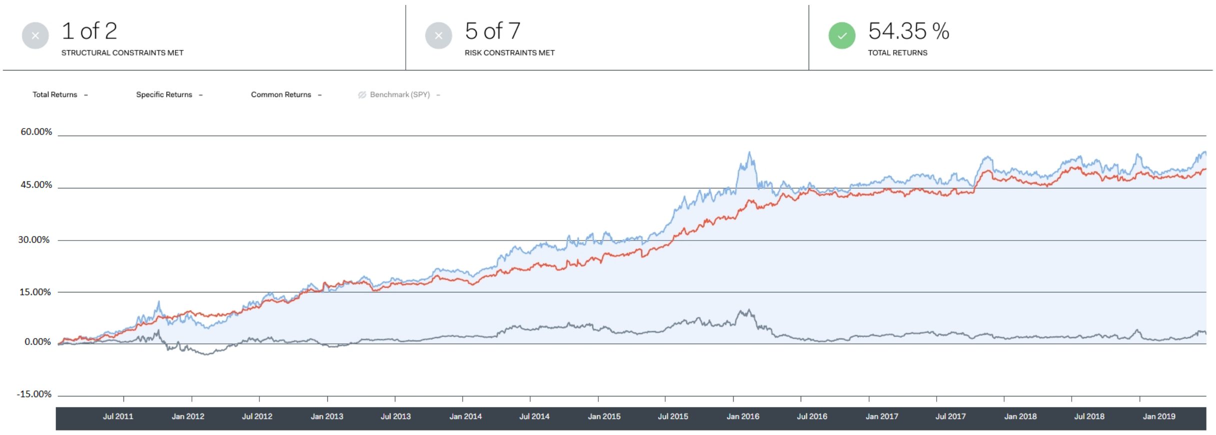


Figure 3.9: top result Daily Trade by ensembling GaussianNB, LogisticRegression, DTC and SGDC (best classifier)

3.5.2 Weekly Trading:

- **AdaBoostClassifier:** Using AdaBoost for weekly trading we get a return of 5.25%.
- **Decision Tree:**Decision tree for weekly trading we get a total return of 10.23%.
- **Random Forest:**Using random forest we get a total return of 7.86%.

3.5.3 Monthly Trading:

- **AdaBoostClassifier :** Using AdaBoost for weekly trading we get a return of 6.16%.
- **SVM :** Using AdaBoost for weekly trading we get a return of 13.05%.
- **Random Forest :** Using AdaBoost for weekly trading we get a return of 4.05%.

3.5.4 Performance of our best classifier

Total Returns: It is the total amount of returns of an investment over a given period of time.This accounts for two different categories of investment.

1. Fixed income investment
2. Distribution and capital appreciation

Common Returns: Common returns are how much of your total returns can be attributed to the common risk factors as modeled by Quantopian exposure to market beta, sectors, momentum, mean reversion, volatility, size, and value. If all your returns are common returns, it means your algorithm isn't doing anything unique and is therefore of little value. Table 3.4 shows 2.71% of common returns.
Specific Returns: It is an excess return that we get from an asset that is independent of specific returns of other assets. Table 3.4 shows 50.60% of common returns.
Sharpe Ratio: It is the measure of performance measure of an investment by risk adjustment. It measures the excess returns for every unit deviation of a trade. Our approach had a 1.16% sharpe ratio which is decent shown in table 3.4.

$$
\begin{aligned}
&\text{Sharpe Ratio } = \tfrac{E_p - E_f}{\sigma_p} \\
&\text{where:} \\
E_p &= \text{ return of portfolio} \\
E_f &= \text{ risk-free rate} \\
\sigma_p &= \text{ portfolio additional return's standard deviation}
\end{aligned}
\tag{3.31}
$$

Max Drawdown: It is the maximum observed loss from the maximum observed point of the graph to the minimum point. This is used to assess the relative risk of a stock strategy.

$$MDD = \frac{\text{Trough Value} - \text{Peak Value}}{\text{Peak Value}} \tag{3.32}$$

Volatility: It is the measure of risk

3.6 Performance Evaluation and Risk Evaluation:

Our best algorithm from all of the above, was the ensemble learning algorithm which incorporated Gaussian Naive Bayes classifier , Logistic Regression , decision tree classifier and Stochastic gradient descent classifier. The training day for each decision was set to be 200 days prior to that day and trading was done daily. Below are the few results that are got by running the algorithm from the date 01/04/2011 to 07/05/2019 with a capital of 10000000 USD.

Table 3.4: Performance of the System's Best Model

Total Returns	54.35%
Specific Returns	50.60%
Common Returns	2.71%
Sharp	1.16%
Max Draw Down	-8.31%
Volatility	0.05%

The table 3.4 depicts that returns calculated from the initial investment was 54.35% on the total capital. The average Sharpe ratio is 1.16 and the average volatility is 0.05 and the final max drawdown was -8.31. These values indicate that our model returns a portfolio that has a low level of risk.

Cumulative specific and total returns: Cumulative returns are independent of the time period and us the total amount of profit or loss from a particular investment. The common returns is very low which is a good sign for the model as it means that our algorithm has a low beta and performs well irrespective of whether the stock prices rise or fall. Which made the specific return very high (50.60%) as shown in figure 3.10.

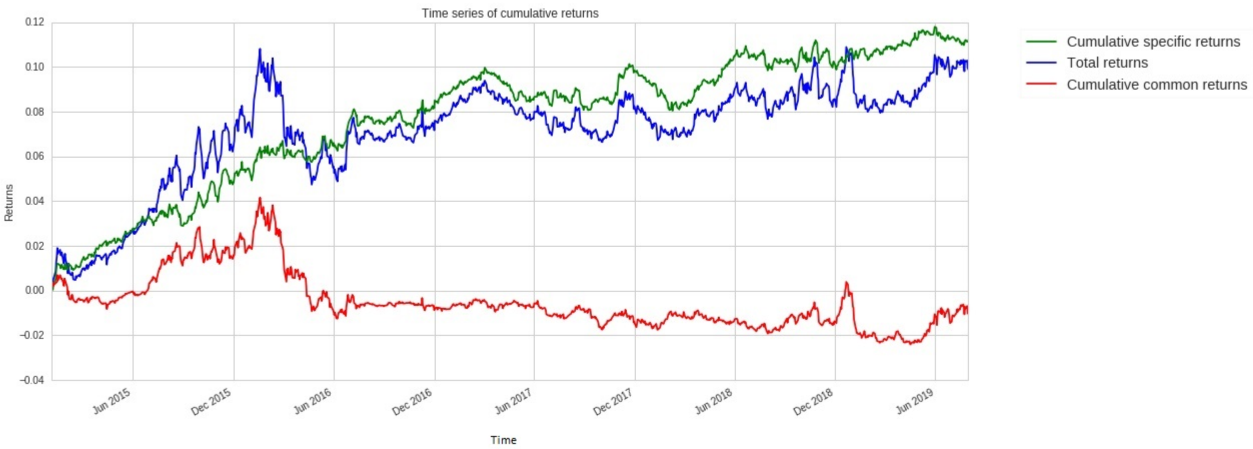


Figure 3.10: Cumulative specific and total returns

Returns over time : Returns are gains or losses made by a particular investment. Returns can be expressed as the percentage increase or decrease in a particular investment or it can be quantified in a particular currency. The figure 3.11 shows that the returns are mostly positive.

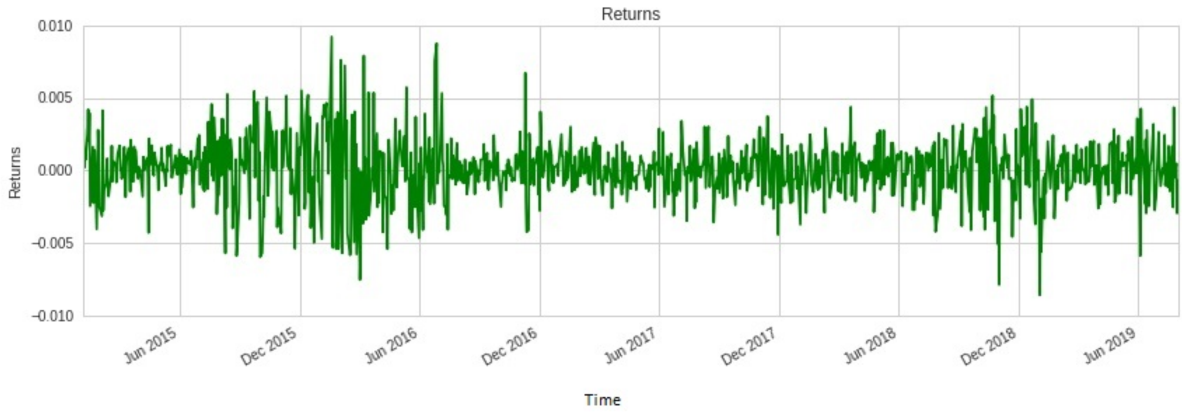


Figure 3.11: Returns over time

Rolling portfolio beta to Equity: This is shown in figure 3.12. The beta is the risk that can be attributed to the movement of the market. A beta having the value 1 signifies that a portfolio follows the trend of the market precisely. Whereas, a beta having a lower value than 1 means that a portfolio is less correlated with the overall market. Low beta value incorporated with high Alpha value will mean that the portfolio will make profit irrespective of the market movement.

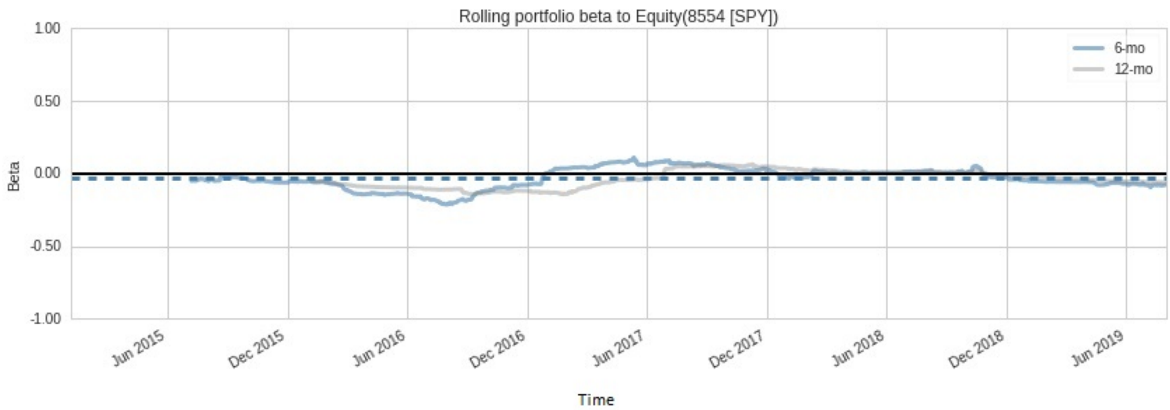


Figure 3.12: Rolling portfolio beta to Equity

Daily weekly and Monthly returns : Figure 3.13 illustrates, returns over the daily, weekly and monthly periods are indicated in the above figure. Each figure gives how much profit was made on a particular period. The daily, annual and monthly

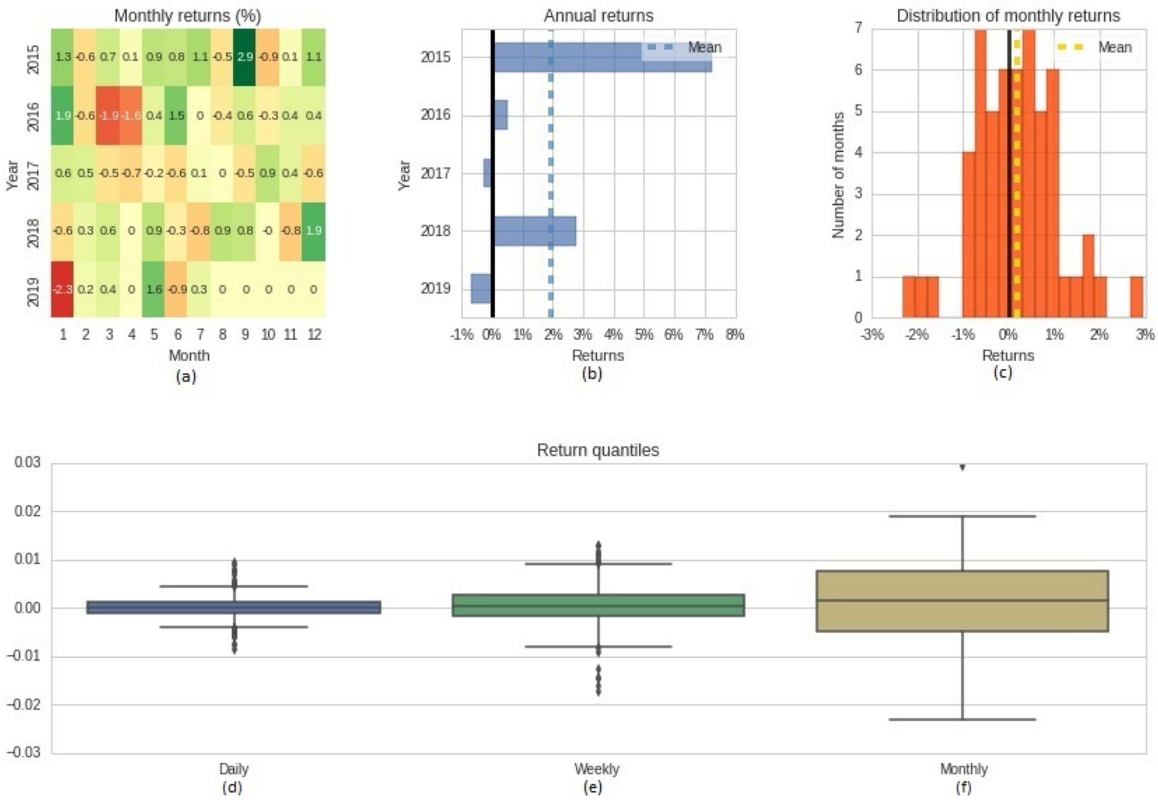


Figure 3.13: (a)Daily returns, (b) weekly returns, (c)Monthly returns, (d)daily, (e)weekly and (f) monthly quantiles

Style Exposure : Figure 3.14 shows, exposure to various investing styles. The values displayed are the rolling 63-day mean. The relevant styles are described below:

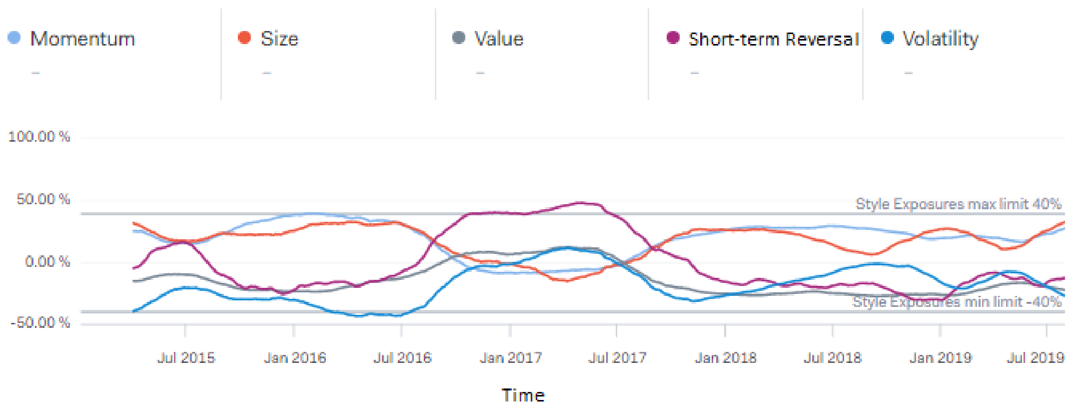


Figure 3.14: Expose to Momentum, Size, Value, Short-term Reversal and Volatility

Ratio of long and short position : We implemented an equal amount of long and short position strategy as shown in figure 3.15. So at a time we went long on 250 stocks and short on 250 stocks. This made our model to perform well both on bull market and bear market.

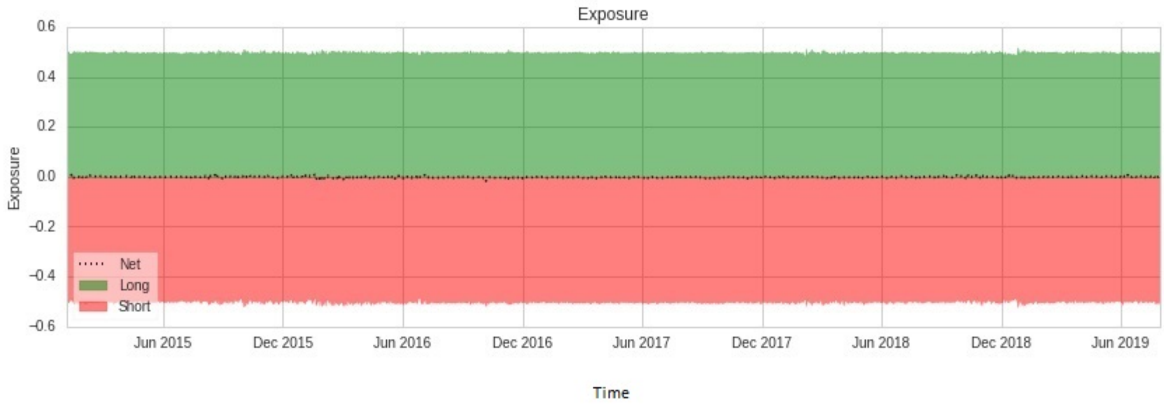


Figure 3.15: Ratio of long and short position

Daily Holdings: From the figure 3.16, we see the total daily holdings of our portfolio which never exceeds 500. As we set our maximum holding limit in our portfolio to be 500.

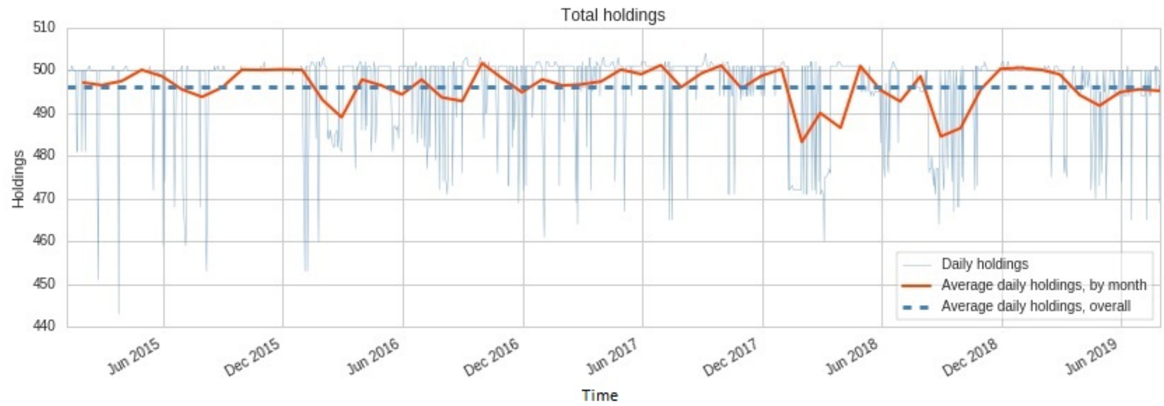


Figure 3.16: Daily Holdings

Gross Leverage : Figure 3.17 shows, we kept our leverage at max 1.05 and at least 0.96 so that our money would be utilized but avoided the risk of being liquidated.

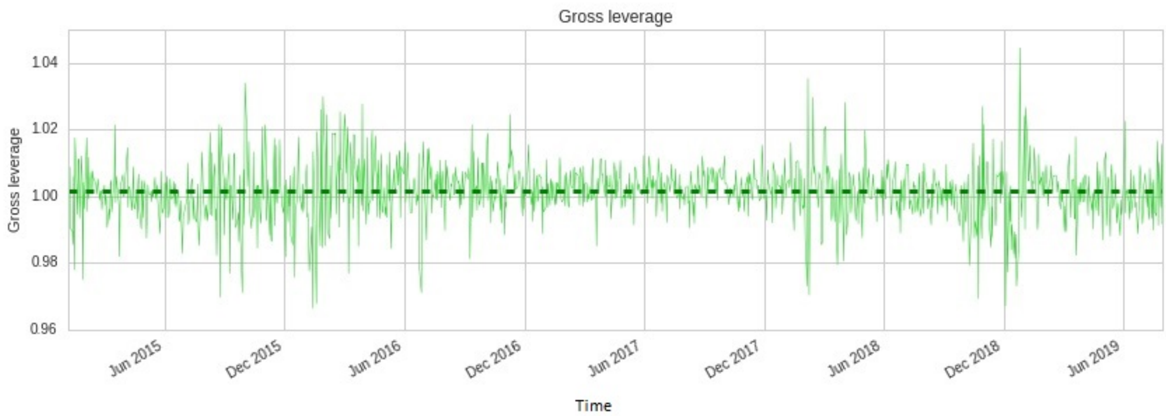


Figure 3.17: Gross Leverage

All the features that we calculated were later filtered out and grouped out into their specific dates for trading where they perform the best. The three categories are weekly trading, monthly trading and daily trading. We then used specific different algorithms to trade in order to compare their performance.

3.7 Evaluation on Synthetic Dataset

To further evaluate our model we created 2 synthetic datasets . In order to test if our model works both for normally distributed dataset and non-Gaussian dataset we use 2 different types of data generation techniques. All of the 148 data were used in both the datasets. After running our feature selection model 25 of the features were selected for final decision making.

	WILLR1	CCI_5D	TRANGE_5D	Price Oscillator1D	BOP	MFI2D	Earnings Quality	ATR2	PLUS_DI2	CMO2D	MEDPRICE
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000
mean	44.005103	3.516222	28.767159	0.495170	48.485151	36.918114	2.039838	5.980287	14.066248	35.192657	8.408863
std	21.577075	1.446376	16.026859	0.289945	13.483475	20.696564	1.147104	1.743082	5.751792	16.711553	4.339233
min	7.002618	1.000060	1.001068	0.000130	25.012245	1.005475	0.003682	3.000239	4.000696	6.004086	1.012838
25%	25.024638	2.262077	15.015692	0.243277	36.903169	19.160635	1.042545	4.453646	9.139007	20.585906	4.678866
50%	44.221253	3.529280	28.812670	0.492920	48.342528	36.547465	2.058611	5.992757	14.080048	35.295630	8.333083
75%	62.710065	4.772728	42.653096	0.750678	59.947694	55.220304	3.038629	7.513247	19.137138	49.952540	12.074140
max	80.986973	5.999387	55.995911	0.999993	71.994185	71.975806	3.999677	8.996947	23.958355	63.992037	15.995171

Figure 3.18: Descriptive Statistics of the Synthetic data

Figure 3.18 shows mean, standard deviation, minimum value, maximum value and specific percentile scores that give a basic idea about the synthetic datasets.

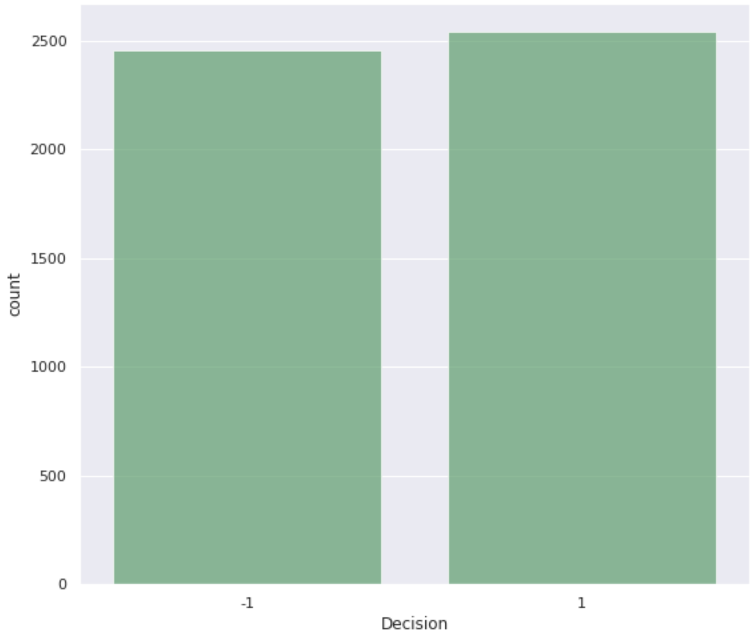


Figure 3.19: Labels of Dataset 1 (Uniform)

In figure 3.19, the labels are for long and short are represented by 1 and -1 respectively. The number of each label is close to 2500 adding up to a total of 5000 instances.

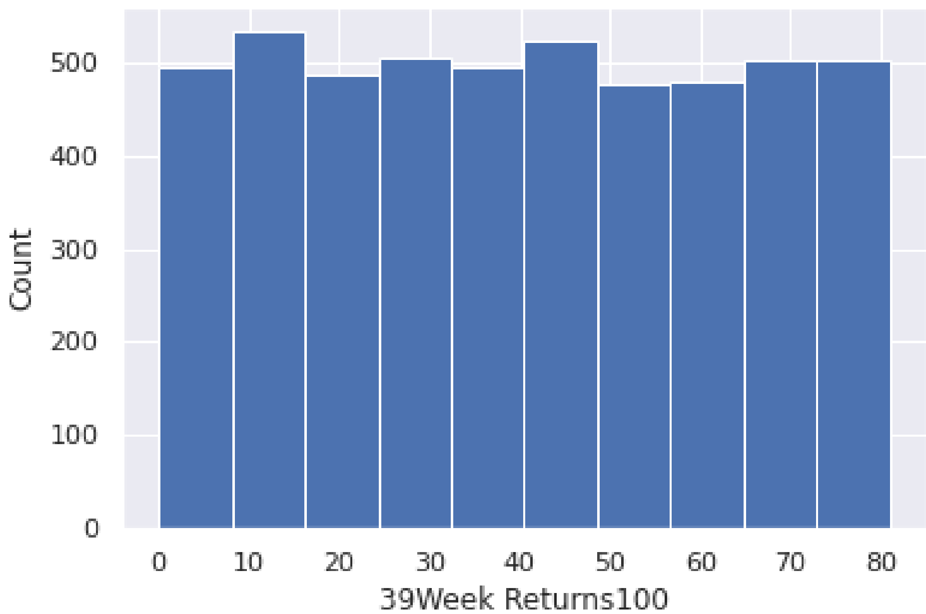


Figure 3.20: Distribution of features in Dataset 1 (Uniform)

In figure 3.20 the distribution of one of the features of the dataset 1 is shown.

The min and max value of the data was required to generate this non-Gaussian distribution.

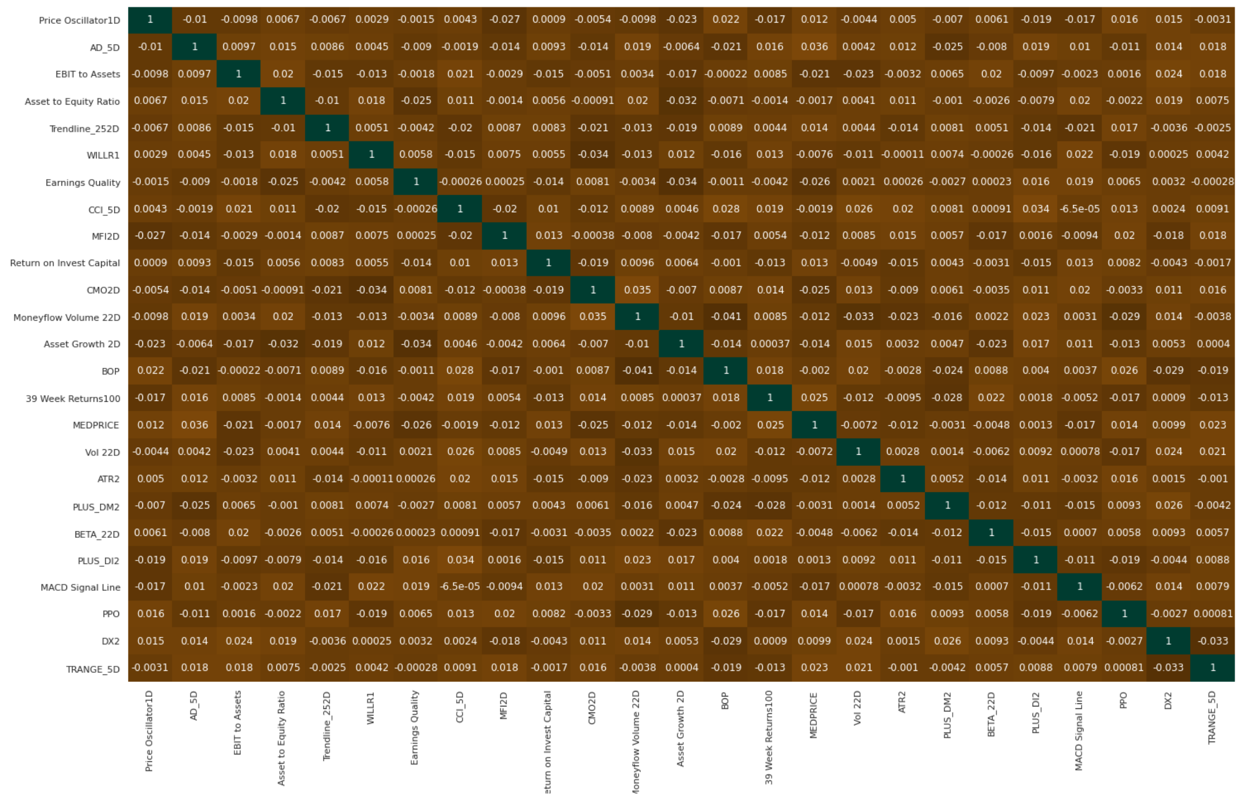


Figure 3.21: Correlation heatmap of Dataset 1 (Uniform)

The correlation of the final 25 features are shown in figure 3.21. The correlation of the features are important in determining the relationship between the features. Only one feature should out of two, if they are highly co-related.

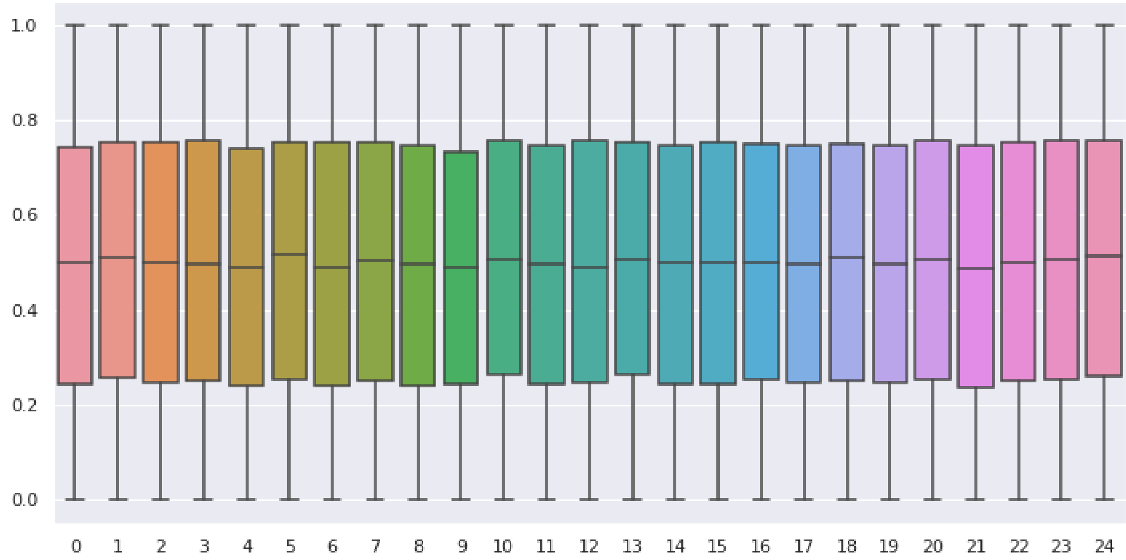


Figure 3.22: BoxPlot of the 25 selected features in Dataset 1 (Uniform)

Figure 3.22 shows the boxplot of the 25 selected features. The values are scaled from 0 to 1. The middle line shows the mean of the feature. The distribution of the feature values can be visualized from the figure 3.22.

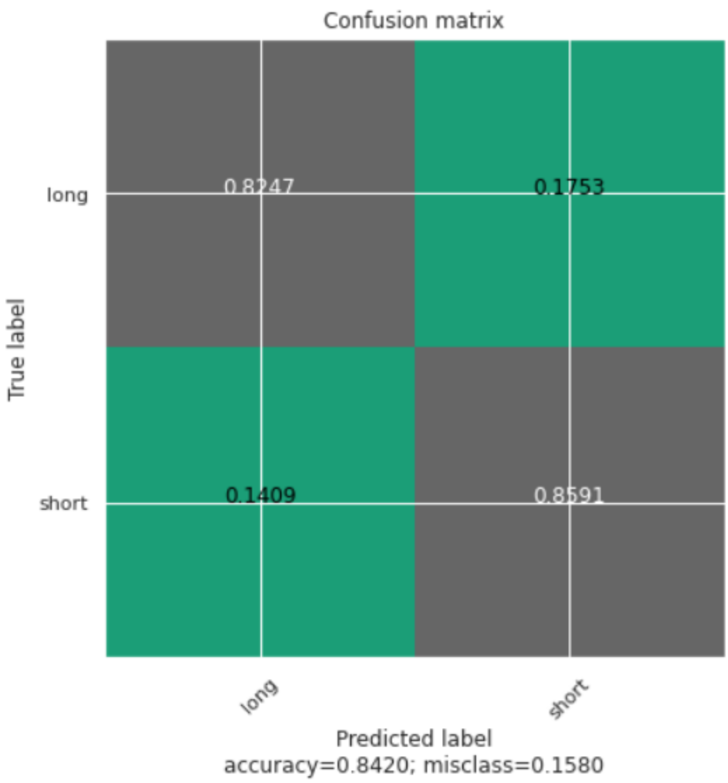


Figure 3.23: Confusion Matrix from the ensemble method on Dataset 1 (Uniform)

Figure 3.23 is the confusion matrix generated by the proposed best model. Our model achieves 84.20% accuracy in the non-Gaussian dataset. Using The values from the confusion matrix the Recall is 82.46%, Precision is 85.26% and the F1 score is 83.84%.

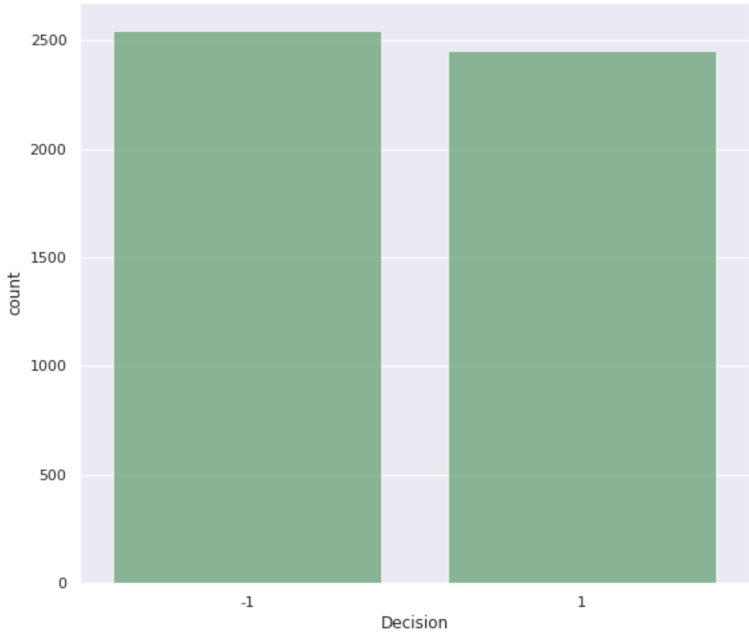


Figure 3.24: Labels of Dataset 2 (Gaussian)

In figure 3.24, the labels are for long and short are represented by 1 and -1 respectively. The number of each label is close to 2500 adding up to a total of 5000 instances.

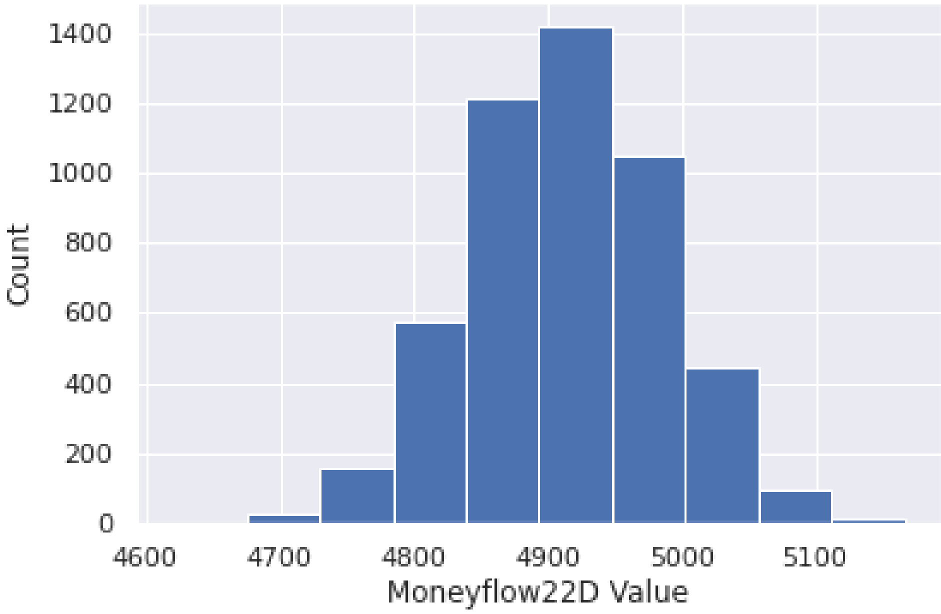


Figure 3.25: Distribution of features in Dataset 2 (Gaussian)

In figure 3.25 the distribution of one of the features of the dataset 2 is shown. The dataset is created using the mean value and standard deviation of the original data.

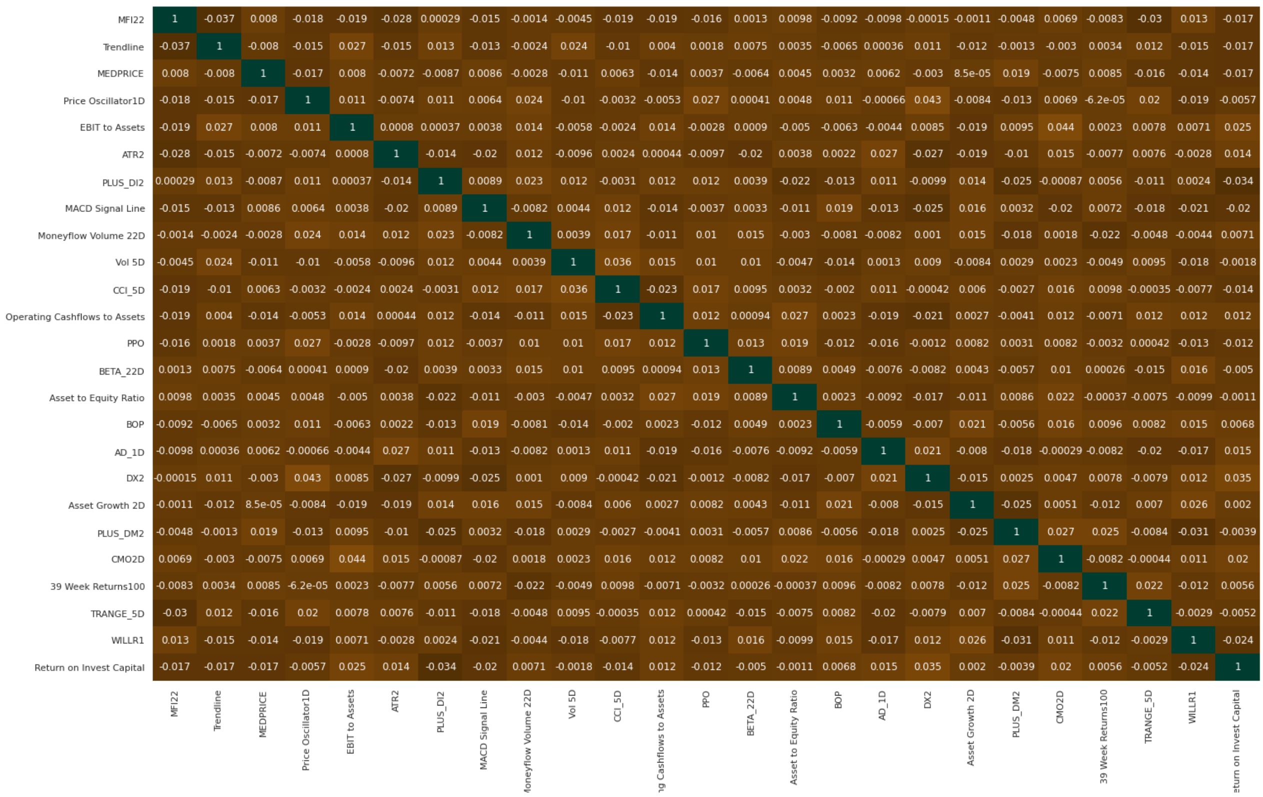


Figure 3.26: Correlation heatmap of Dataset 2 (Gaussian)

The correlation of the final 25 features are shown in figure 3.26. The correlation of the features are important in determining the relationship between the features. Only one feature should out of two, if they are highly co-related.

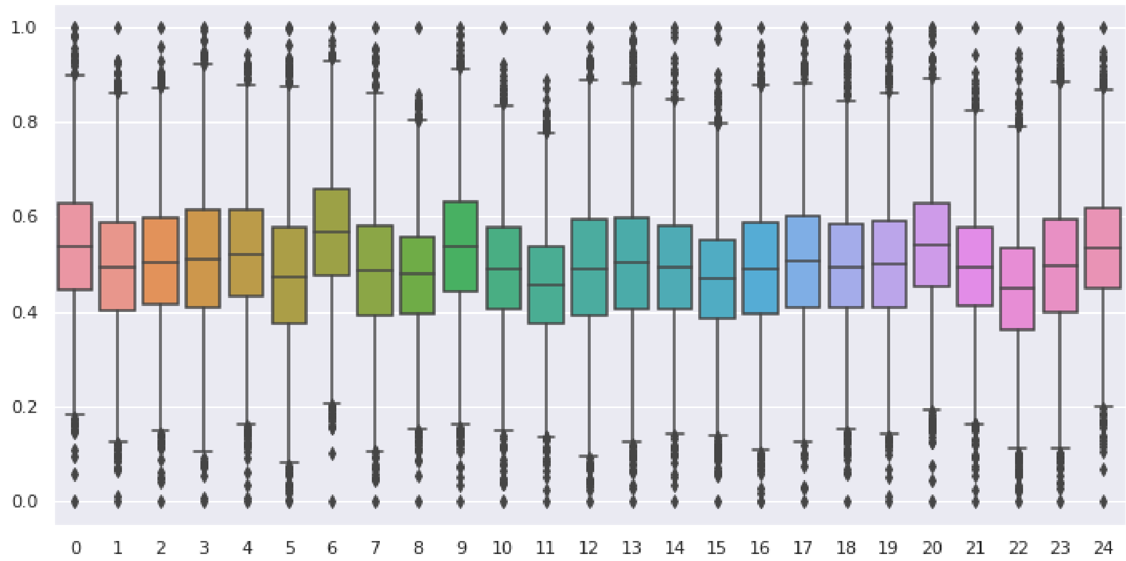


Figure 3.27: BoxPlot of the 25 selected features in Dataset 2 (Gaussian)

Figure 3.27 shows the boxplot of the 25 selected features. The values are scaled from 0 to 1. The middle line shows the mean of the feature. The distribution of the feature values can be visualized from the figure 3.22.

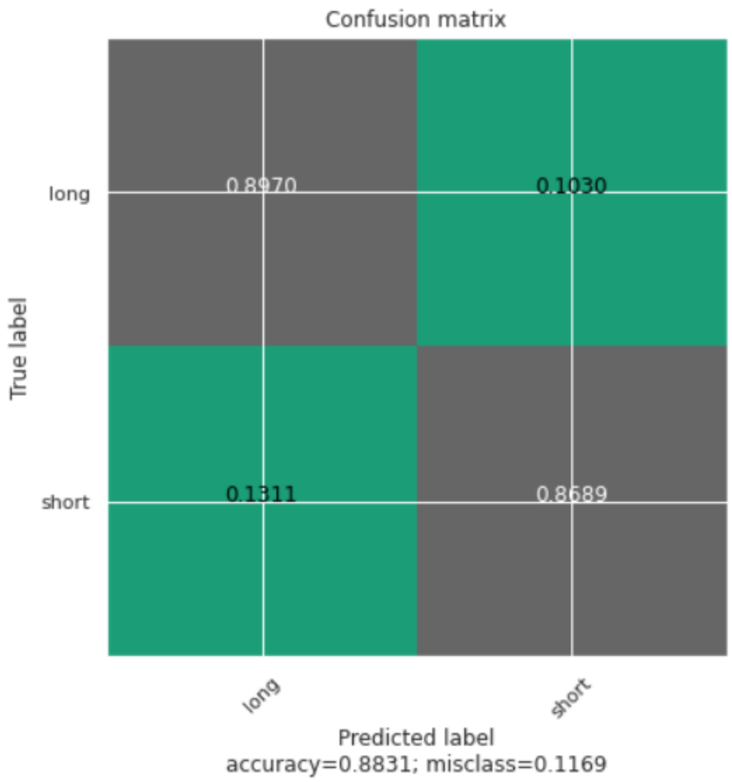


Figure 3.28: Confusion Matrix from the ensemble method on Dataset 2 (Gaussian)

Figure 3.28 is the confusion matrix generated by the proposed best model. Our model achieves 88.30% accuracy in the non-Gaussian dataset. Using The values from the confusion matrix the Recall is 89.70%, Precision is 87.36% and the F1 score is 88.51%.

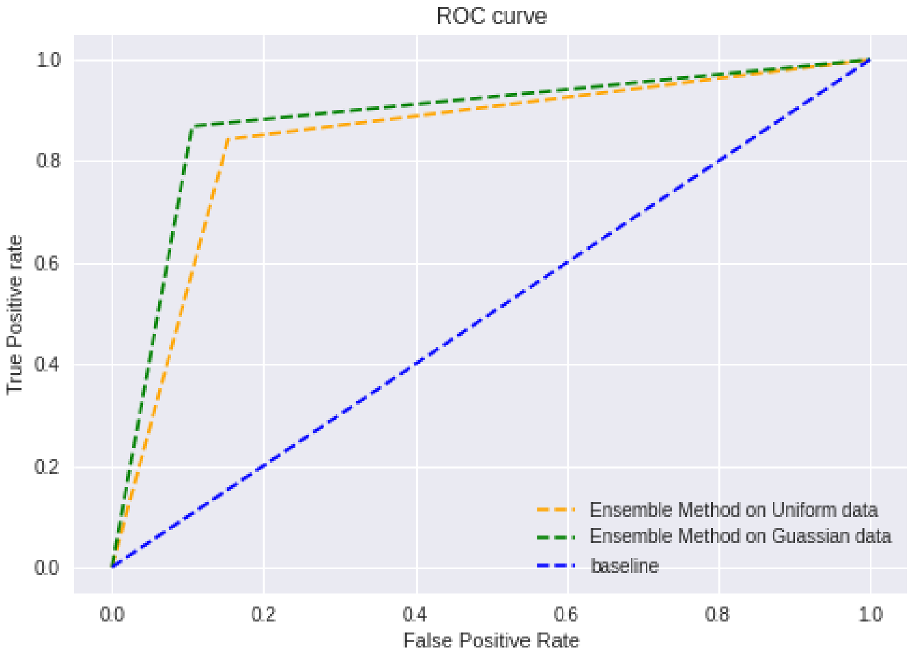


Figure 3.29: ROC curve of the Synthetic Datasets

A greater X-axis value in a ROC curve implies a larger number of False positives than True negatives. While a higher Y-axis value implies a greater number of True positives than False negatives, a lower Y-axis value suggests a lower number of True positives. As a result, the threshold is determined by the capacity to balance False positives and False negatives.
It is evident from the figure 3.29 that the AUC for the ensemble model on Gaussian data ROC curve is higher than that for the non-Gaussian data ROC curve. Therefore, we can say that our model does a better job of classifying the positive class in the normally distributed dataset. The AUC value of our model in the normally distributed dataset is 0.8814 and the AUC value of our model with uniformly distributed dataset is 0.8449.

Table 3.5: Proposed model performance on Synthetic data

Type of The Dataset	Accuracy	Precision	Recall	F1-score	AUC-score
Normally Distributed	88.30%	87.36%	89.70%	88.51%	0.8814
Uniformly Distributed	84.20%	85.26%	82.46%	83.84%	0.8449

Table 3.5 shows that the proposed model works better when dataset is normally distributed. As most stock data is normally distributed that is why the proposed model is finely tuned to work better with normally distributed data. However,the model also performs quite well achieving 84.20% accuracy and quite good in other performance matrices as shown in table 3.5.

3.8 Comparison with other Models and Reason for Proposed model being Winning

The most important part of our model is our novel feature calculation and selection method. For this reason even with the huge drawbacks of the Quantopian platform our model performs on par with the state-of-the-art models that perform quantitative Trading as can be seen from table 3.6. The biggest advantage of our proposed model is that the feature selection method can be added to any decision making model to make better predictions.

Table 3.6: Comparison of Proposed model with state-of-the-art models

Author	Time Period	Returns	Trading Frequency	Method Used
T. Dai, A. Shah and H. Zhong (2012)	5 years	30.66%	30 Days	SVM and Logistic Regression
G. Chen, Y. Chen and Fushimi (2017)	7 years	103%	30 Days	LSTM
Vo, N. N. Y., He, X., Liu, S., & Xu, G. (2019)	3 years	50.78%	1 year	Reinforcement Learning
Proposed Model	8 years	54.35%	1 Day	Ensemble Learning with Feature Selection

Table 3.6 shows that the proposed model performs better than the model of T. Dai and G. Chen. Moreover, the feature extraction and selection method can select the best feature for trading in any time-period. Therefore, this model can be incorporated with any model to significantly improve the quality of the features.

3.9 Discussion

Through experimentation, it is clear that ensemble learning produced a better result in case of stock market trading as compared to using a single algorithm. Furthermore, it also became clear that most important part of a stock trading algorithm is the feature extraction part. The 1 day trading algorithm made 54.35% over the course of 8 years profit due to the quality of the features that were used for 1 day trading. Whereas the weekly and the monthly algorithm did not perform that as well due to its features. Our most significant contribution is that we detected using statistical measures that which features should work well for which time-frame. The model can clearly capture the trend of the market over a one day period.

Chapter 4

Multi Criteria Decision Theory based Efficient stock Portfolio Management

4.1 TOPSIS Method

The figure 4.1 explains the goal, features, sub-criteria and alternatives of our TOPSIS(Technique for Order Preference by Similarity to Ideal Solution) model.
In this paper, we applied an MCDM strategy called the TOPSIS method. This model was proposed by Hwang and Yoon in 1981. This method is based on the concept that the best alternative should be closest to the ideal solution; in other words, it should have the shortest distance from the best solution. The fundamental thought of the TOPSIS technique is utilized to acquire an answer, which is nearest to the ideal arrangement and farthest from the negative ideal arrangement. The TOPSIS accepted that each quality takes the monotonically expanding (or, on the other hand diminishing) utility; at that point, it is easy to decide the "ideal" arrangement which comprises of the apparent multitude of best standards values reachable, and the "negative-ideal" arrangement comprising of all the most exceedingly terrible models esteems reachable. The TOPSIS technique comprises of the accompanying advances:

Step 1: The normalization of the decision matrix is performed using equation in 4.1 and by using that the normalization decision matrix $R = [r_{ij}]_{m \times n}$ is calculated.

$$n_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}} \quad i = 1, 2, \ldots, n; j = 1, 2, \ldots, m \tag{4.1}$$

Step 2: The columns of normalized matrix R multiplied by the related weight, w_j, and values of the weighted and normalized decision matrix are calculated by the following equation 4.2:

$$V_{ij} = W_j r_{ij}, i = 1, 2, \ldots\ldots, m; j = 1, 2, \ldots\ldots, n \tag{4.2}$$

Step 3: The ideal solution and negative-ideal solution are determined using Equations. 4.3 and 4.4 respectively. Here J and J' are sets of index of benefit and cost criteria in accordance.

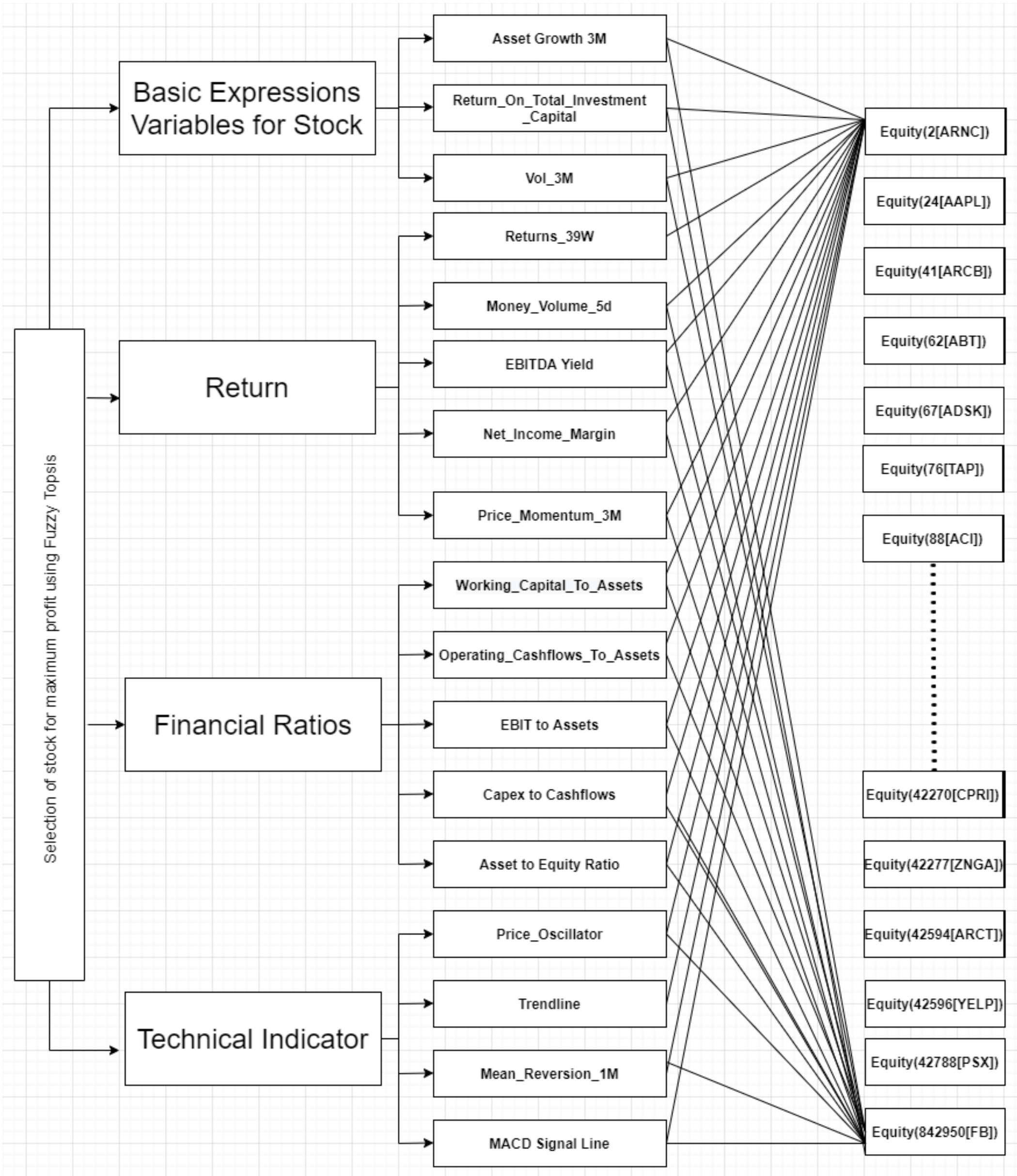


Figure 4.1: Hierarchical model of the features used for analysis

$$v_j^+ = \{(\max v_{ij} \mid j \in J), (\min v_{ij} \mid j \in J') \mid i = 1, 2, \ldots, m\} \quad (4.3)$$

$$v_j^- = \{(\min v_{ij} \mid j \in J), (\max v_{ij} \mid j \in J') \mid i = 1, 2, \ldots, m\} \quad (4.4)$$

Hence from the above equations, $A^+ = \{V_1^+, V_2^+, \ldots\ldots, V_n^+\}$ and $A^- = \{V_1^-, V_2^-, \ldots\ldots, V_n^-\}$ are obtained.

Step 4: Two Euclidean distances from the ideal(best) and anti-ideal(worst) solutions are calculated for each alternatives using the Equations. 4.5 and 4.6respectively.

$$(S_i^+) = \sqrt{\sum_{j=1}^n (V_{ij} - V_j^+)^2}, i = 1, 2, \ldots\ldots, m \quad (4.5)$$

$$(S_i^-) = \sqrt{\sum_{j=1}^n (V_{ij} - V_j^-)^2}, i = 1, 2, \ldots\ldots, m \quad (4.6)$$

Step 5: The relative closeness to the ideal solution C_i^+ is calculated as shown in the following equation 4.7 and a higher the value of the relative closeness to the ideal solution indicates a higher rank:

$$C_i^+ = \frac{s_i^-}{s_i^- + s_i^+}, i = 1, 2, \ldots\ldots, m; 0 \leq C_i^+ \leq 1 \quad (4.7)$$

4.2 Data of TOPSIS

4.2.1 Dataset Description

The paper used the data from total 6 months(2016-09-14 to 2016-03-06) of data to test perform testing. The data was obtained from Quantopian platform's US1500 stocks which includes the most volatile 1500 stocks of the US stock market. Quantopain allows that features could be extracted from the primary features that are provided by the environment. Our analysis used 17 features which can be seen from the table 4.2.

Features
39 Week Returns
Asset Growth 3M
Asset to Equity Ratio
Capex to Cashflows
EBIT to Assets
EBITDA Yield
MACD Signal Line
Mean Reversion 1M
Moneyflow Volume 5D
Net Income Margin
Operating Cashflows to Assets
Price Momentum 3M
Price Oscillator
Return on Invest Capital
Vol 3M
Working Capital to Assets
Trendline

Figure 4.2: List of 17 features used for analysis

security	39 Week Returns	Asset Growth 3M	Asset to Equity Ratio	Capex to Cashflows	EBIT to Assets	EBITDA Yield	MACD Signal Line	Mean Reversion 1M	Moneyflow Volume 5D	Net Income Margin	Operating Cashflows to Assets	Price Momentum 3M	Price Oscillator	Return on Invest Capital	Trendline	Vol 3M	Working Capital to Assets
ATVI	87.0	816.0	1244.0	75.0	234.0	911.0	474.0	1411.0	509.0	123.0	363.0	859.0	110.0	231.0	230.0	1181.0	538.0
CIM	1266.0	591.0	610.0	775.0	644.0	692.0	1385.0	452.0	260.0	959.0	802.0	1156.0	1281.0	693.0	1422.0	366.0	352.0
DOW	145.0	1073.0	160.0	483.0	1167.0	455.0	319.0	648.0	72.0	755.0	1346.0	421.0	142.0	1186.0	201.0	1031.0	1211.0
EFII	56.0	46.0	1045.0	820.0	994.0	1335.0	314.0	173.0	1003.0	494.0	1462.0	272.0	72.0	1030.0	207.0	1338.0	1132.0
ENS	671.0	663.0	194.0	710.0	689.0	978.0	609.0	254.0	618.0	1282.0	856.0	934.0	666.0	690.0	621.0	179.0	282.0
IMPV	97.0	889.0	1279.0	609.0	0.0	940.0	1045.0	1116.0	1167.0	655.0	1060.0	1354.0	94.0	959.0	140.0	1285.0	0.0
ISRG	109.0	7.0	288.0	1292.0	10.0	22.0	703.0	86.0	751.0	30.0	852.0	959.0	339.0	17.0	171.0	1347.0	281.0
NYCB	709.0	779.0	395.0	817.0	1311.0	273.0	1326.0	1136.0	376.0	896.0	1486.0	1462.0	1088.0	1357.0	421.0	1010.0	816.0
RRGB	2.0	1107.0	869.0	1262.0	229.0	25.0	511.0	60.0	921.0	11.0	443.0	2.0	3.0	219.0	352.0	1479.0	1266.0
XRAY	898.0	501.0	1303.0	869.0	1333.0	1385.0	1256.0	721.0	99.0	811.0	1454.0	442.0	743.0	1422.0	1287.0	87.0	580.0

Figure 4.3: Decision matrix

4.3 Application and Results of TOPSIS Paper

The entire dataset consists of 114478 instances/rows and 17 columns. The label is percent change in returns with a 5 day window length. We used AdaBoostClassifier with n_estimator = 150 as the hyperparameter. Post training the classifier with the dataset, we used the attribute of Sklearn's classifier called "feature_importances_" to determine the weight we should put on each feature.
The weight for each feature is shown graphically in the table 4.4.

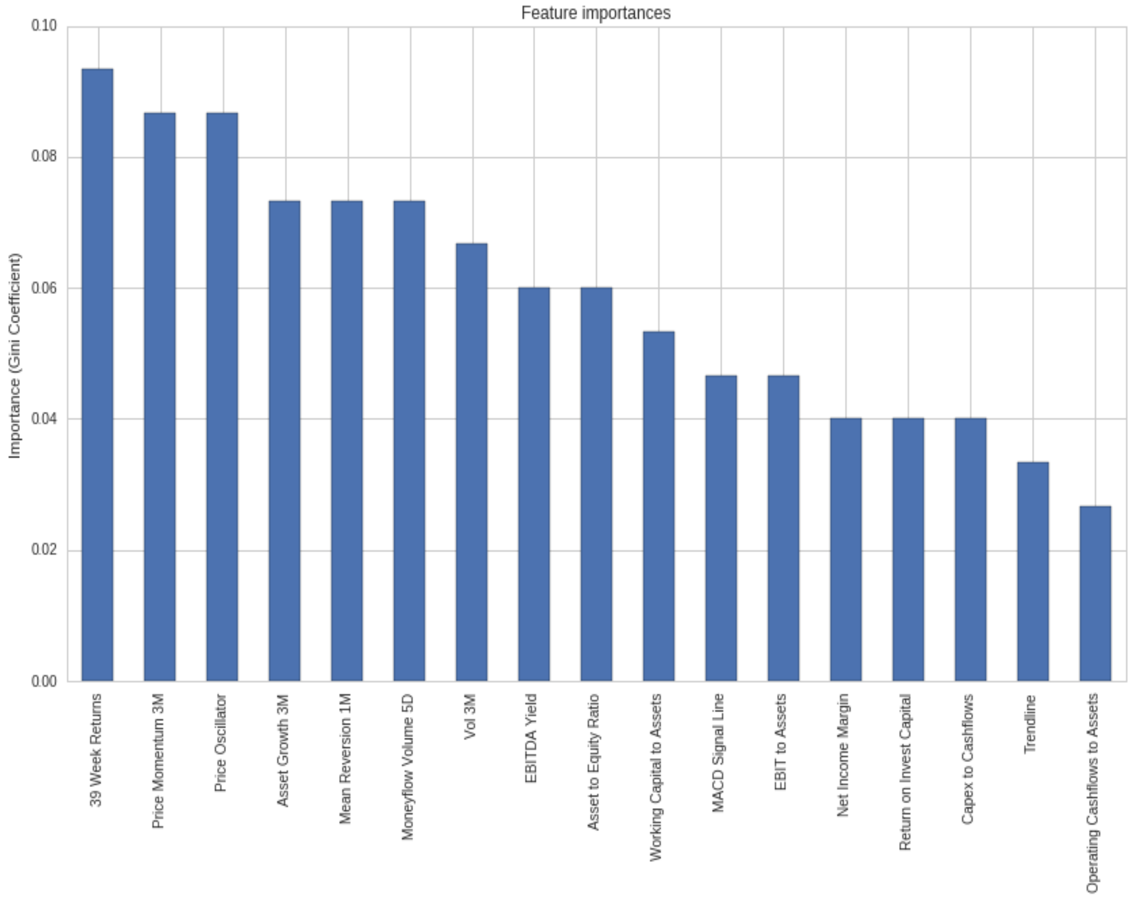


Figure 4.4: Weight distribution of features

security	39 Week Returns	Asset Growth 3M	Asset to Equity Ratio	Capex to Cashflows	EBIT to Assets	EBITDA Yield	MACD Signal Line	Mean Reversion 1M	Moneyflow Volume 5D	Net Income Margin	Operating Cashflows to Assets	Price Momentum 3M	Price Oscillator	Return on Invest Capital	Trendline	Vol 3M	Working Capital to Assets
ATVI	87.0	816.0	1244.0	75.0	234.0	911.0	474.0	1411.0	509.0	123.0	363.0	859.0	110.0	231.0	230.0	1181.0	538.0
CIM	1266.0	591.0	610.0	775.0	644.0	692.0	1385.0	452.0	260.0	959.0	802.0	1156.0	1281.0	693.0	1422.0	366.0	352.0
DOW	145.0	1073.0	160.0	483.0	1167.0	455.0	319.0	648.0	72.0	755.0	1346.0	421.0	142.0	1186.0	201.0	1031.0	1211.0
EFII	56.0	46.0	1045.0	820.0	994.0	1335.0	314.0	173.0	1003.0	494.0	1462.0	272.0	72.0	1030.0	207.0	1338.0	1132.0
ENS	671.0	663.0	194.0	710.0	689.0	978.0	609.0	254.0	618.0	1282.0	856.0	934.0	666.0	690.0	621.0	179.0	282.0
IMPV	97.0	889.0	1279.0	609.0	0.0	940.0	1045.0	1116.0	1167.0	655.0	1060.0	1354.0	94.0	959.0	140.0	1285.0	0.0
ISRG	109.0	7.0	288.0	1292.0	10.0	22.0	703.0	86.0	751.0	30.0	852.0	959.0	339.0	17.0	171.0	1347.0	281.0
NYCB	709.0	779.0	395.0	817.0	1311.0	273.0	1326.0	1136.0	376.0	896.0	1486.0	1462.0	1088.0	1357.0	421.0	1010.0	816.0
RRGB	2.0	1107.0	869.0	1262.0	229.0	25.0	511.0	60.0	921.0	11.0	443.0	2.0	3.0	219.0	352.0	1479.0	1266.0
XRAY	898.0	501.0	1303.0	869.0	1333.0	1385.0	1256.0	721.0	99.0	811.0	1454.0	442.0	743.0	1422.0	1287.0	87.0	580.0

Figure 4.5: Decision matrix

TOPSIS method: In figure 4.5 we see the initial decision matrix where the first column signifies the name of the stock. The following 17 columns are the features that are used for analysis. The normalization of the dataset is calculated using equation 4.1. The weight distribution of each of the features can be seen in table 4.1.

Table 4.1: Weight Distribution of the Features

39 Week Returns	Asset Growth 3M	Asset to Equity Ratio	Capex to Cashflows	EBIT to Assets	EBITDA Yield	MACD Signal Line	Mean Reversion 1M	Moneyflow Volume 5D
0.093	0.073	0.06	0.04	0.053	0.053	0.046	0.073	0.073
Net Income Margin	Operating Cashflows to Assets	Price Momentum 3M	Price Oscillator	Return on Invest Capital	Trendline	Vol 3M	Working Capital to Assets	
0.04	0.026	0.086	0.086	0.04	0.033	0.066	0.053	

The result after the normalization multiplied by the related weight, w_j, and values of the weighted and normalized decision matrix are calculated by the following equation 4.2. The transformation of the dataset can be seen from figure 4.6.

	39 Week Returns	Asset Growth 3M	Asset to Equity Ratio	Capex to Cashflows	EBIT to Assets	EBITDA Yield	MACD Signal Line	Mean Reversion 1M	Moneyflow Volume 5D	Net Income Margin	Operating Cashflows to Assets	Price Momentum 3M	Price Oscillator	Return on Invest Capital	Trendline	Vol 3M	Working Capital to Assets
ATVI	0.004394	0.025558	0.027433	0.001128	0.004771	0.018193	0.007871	0.043053	0.017319	0.002141	0.002815	0.025866	0.055127	0.080139	0.107823	0.354562	0.221027
CIM	0.063939	0.018511	0.013452	0.011658	0.013131	0.013820	0.022999	0.013792	0.008847	0.016689	0.006219	0.034810	0.641984	0.240418	0.666627	0.109881	0.144613
DOW	0.007323	0.033608	0.003528	0.007265	0.023794	0.009087	0.005297	0.019772	0.002450	0.013139	0.010438	0.012677	0.071165	0.411452	0.094228	0.309529	0.497517
EFII	0.002828	0.001441	0.023045	0.012334	0.020267	0.026661	0.005214	0.005279	0.034128	0.008597	0.011337	0.008191	0.036083	0.357332	0.097041	0.401697	0.465061
ENS	0.033889	0.020766	0.004278	0.010680	0.014048	0.019531	0.010113	0.007750	0.021028	0.022310	0.006638	0.028125	0.333772	0.239378	0.291122	0.053740	0.115854
IMPV	0.004899	0.027845	0.028205	0.009161	0.000000	0.018772	0.017353	0.034052	0.039708	0.011399	0.008220	0.040772	0.047109	0.332700	0.065631	0.385785	0.000000
ISRG	0.005505	0.000219	0.006351	0.019434	0.000204	0.000439	0.011674	0.002624	0.025553	0.000522	0.006607	0.028878	0.169893	0.005898	0.080164	0.404399	0.115444
NYCB	0.035808	0.024399	0.008711	0.012289	0.026730	0.005452	0.022020	0.034662	0.012794	0.015593	0.011524	0.044024	0.545261	0.470776	0.197363	0.303224	0.335238
RRGB	0.000101	0.034673	0.019164	0.018983	0.004669	0.000499	0.008486	0.001831	0.031337	0.000191	0.003435	0.000060	0.001503	0.075976	0.165016	0.444028	0.520112
XRAY	0.045354	0.015692	0.028734	0.013071	0.027179	0.027659	0.020857	0.021999	0.003369	0.014114	0.011275	0.013310	0.372361	0.493326	0.603340	0.026119	0.238282

Figure 4.6: Weighted Normalized Decision Matrix

The ideal solution and negative-ideal solution are determined using Equations.4.3 and 4.4, respectively. Two Euclidean distances from the ideal(best) and anti-ideal(worst) solutions are calculated for each alternatives using the Equations. 4.5 and 4.6, and can be seen rows of S+ and S- in figure 4.7. The relative closeness to the ideal solution C_i^+ is calculated as shown in equation 4.7 and a higher the value of the relative closeness resulted in a higher rank. Here 1 is the highest rank occupied by the stock IMPV in figure 4.7.

	ATVI	CIM	DOW	EFII	ENS	IMPV	ISRG	NYCB	RRGB	XRAY
S+	0.097963	0.062426	0.102771	0.106682	0.078504	0.092706	0.110325	0.056429	0.116269	0.071259
S-	0.069411	0.103070	0.062562	0.065485	0.070197	0.083325	0.051773	0.098543	0.064670	0.086799
C+	2.431038	-2.535923	1.555910	1.589555	8.450911	8.882330	0.884227	-2.339913	1.253322	-5.585531
Ranking	3.000000	9.000000	5.000000	4.000000	2.000000	1.000000	7.000000	8.000000	6.000000	10.000000

Figure 4.7: Euclidean distances from best and worst solutions and Final Ranking

4.4 Discussion

In our analysis, it showed IMPV, ENS and ATVI had a higher closeness score whereas XRAY, CIM and NYCB had the lowest score among the 10 stocks.The Multi Criteria Decision Theory based study aimed to find the best stock with respect to 17 financial features. The importance and weight selected for the features was done by using Ada-Boost. After finding the weights TOPSIS was used to see the results of 10 stocks from the dataset. In our analysis, it showed IMPV, ENS and ATVI had a higher closeness score whereas XRAY, CIM and NYCB had the lowest score among the 10 stocks.

Chapter 5

Future Directions: Responsible AI in Financial Investment

Automated trading is used in most of the major markets of our world. In order to ensure sustainable development, incorporating ethical and socially responsible ideas while designing these Artificial Intelligence (AI) systems has become a necessity. Both the industry and the academia are working towards Responsible AI, which can make Socially Responsible Investments (SRI). This paper reviews the research on SRI investment in the financial sector and evaluates these methods, which can help find future research directions in Computational Finance. This survey looks at the machine learning techniques used for ethical decision-making while stock or forex trading, which will benefit any further research work on Responsible AI in Finance.

In several exchanges across the globe, automated trading and high-frequency trading (HFT) has become the norm. HFT entails using automated algorithms to perform proprietary trading techniques. HFTs compete for consistent, albeit small returns on each trade by swiftly trading in and out of positions thousands of times a day without holding positions at the end of the day [52]. While figures vary due to the difficulties of determining whether each trade is an HFT, current estimates show that HFT accounts for 50-70 percent of equities trades in the United States, 40% in Canada, and 35% in London [36] [29]. As artificially intelligent systems have such a massive share in the market, only aiming for profit can lead to the fall of a particular industry, community, or country. If properly managed, these systems can help the economy bloom in all sectors on the flip side. Therefore, It is of utmost importance that investors who harness the power of Artificial Intelligence follow specific guidelines that make them socially responsible and considerate towards Environmental, Social, and Governance metrics.

The IEEE P7000 standards initiatives were established in 2019 to address ethical concerns in designing autonomous and intelligent systems. This decision was made in the wake of increasing public concern about the unintended effects of artificial intelligence (AI), exacerbated by the absence of a proactive procedure for addressing ethical issues in professional practice. The difficulty in transitioning from principles to practice, on the other hand, poses a major obstacle to the application of ethical standards[83].

According to Scopino, if a human builds an AI trader without intending to conduct market manipulation and the AI system performs market manipulation with

discretion, the person may not be held liable under the current US regulatory framework.[60]. Therefore, the need for evaluating AI trading systems have become essential. Investors and academic scholars are now trying to find ways to include ethcial decision making capabilities of algorithmic trading systems.
Investors have traditionally concentrated on investment returns by carefully reviewing financial data to identify the best-performing companies. Investors are growing more interested in other elements of businesses than simply profits because of a recent shift in attitudes about sensitive issues such as global warming and migrants[74].

5.1 Impacts of Automated Trading

Technical innovation has always resulted in more growth and a better average quality of life in the past; however, this does not imply that new technology adoption has always been without resistance [47]. Technological progress and the resulting displacement or change in the state of the economy has caused significant social discontent in the past, as shown by the Luddite movement in the 18th century and brilliantly portrayed in Charlie Chaplin's famous film "Modern Times" [55]. The current wave of AI research has already sparked widespread public debate about its implications for the economy and living conditions. A growing number of individuals and groups are warning about the potential detrimental impact of AI deployment on the financial sector. Some predict that AI will have even more drastic consequences than past technology revolutions [66].
HFT activity has increased dramatically in recent years, prompting both academics and authorities to wonder if it is helpful to financial markets. Although the findings are often inconclusive, recent academic study has looked at the effect of HFT on various metrics of market quality, such as liquidity, transaction costs, market integrity, and efficiency [54].

5.1.1 Stock Volatility

Automated trading is positively associated with stock price volatility[27]. The studies found that high-frequency trading has a detrimental impact on the market's capacity to integrate basic information about firms into asset pricing. When high-frequency trading is active, stock prices tend to overreact to fundamental news. Overall, high-frequency trading may have some negative consequences for the US capital market.

5.1.2 Ripple Effect

A key problem with HFT is that the combination of many distinct high-speed traders may put extra risk on the market and create excessive volatility. What if the different computer programs "misfire" in some way? Other investors may be seriously harmed as a result of this. The so-called "flash crash" of May 6, 2010 exemplifies our automated markets' potential to go haywire[38]. HFT may cause flash crashes, as occurred in 2010 on the BSE when volumes spiked unexpectedly owing to a glitch in a Delhi stock broker's trading algorithm, which resulted in buy and sell orders being executed repeatedly[71]. A flaw in a trading algorithm may have a cascading effect, triggering other trading algorithms and creating a hysterical run on stock prices causing a ripple effect in all related financial markets.

5.1.3 Manipulative Strategies

A variety of manipulative trading techniques try to shift prices away from their true value in order to benefit from the artificial difference[38]. Spoofing, Wash Sales, Quote Stuffing, Front running and other order Triggering Strategies can be easily used to make a profit in an illegal manner using the power of Artificial Intelligence in Trading.

5.2 Responsible AI

Because specific machine learning algorithms are opaque, there is a significant risk of unexpected outcomes, which may have real-world implications for humans and animals. Understanding how neural networks handle a particular input, even intuitively, may be difficult at the moment. This has many consequences in the Financial sector that are not yet completely understood. To begin with, determining an algorithm's implicit assumptions (for example, how much of the contextual background information it uses when making a decision) and, therefore, the possible dangers of utilizing it for this purpose may be challenging. Second, when an algorithm is expected to make predictions outside the boundaries of the training data, it may be not very clear. Making ensuring algorithms fail gracefully is, in fact, a significant research challenge[63]. Finally, an algorithm may be difficult to decipher as to why it made a certain choice. While these factors are well-acknowledged in the AI field, they have been largely ignored in recent debates about the technology's potential environmental advantages[75].

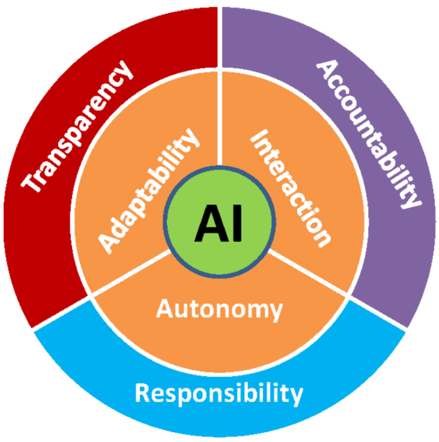


Figure 5.1: ART principles for Responsible AI [66]

A growing understanding is that a responsible approach to AI is required to guarantee that AI technologies are used safely, sound, and equitable. This involves the need to think about the ethical consequences of machine choices and define AI's legal position. Concrete methods to the responsible design of AI, on the other hand, are

almost non-existent. AI applications such as self-driving cars, companions, health-care robots, rating and profile algorithms, which are now impacting society or will be in a few years, need responsive design, development, and usage of AI systems. In all of these cases, AI thinking should evaluate societal values, moral and ethical concerns, assess the relative importance of values held by stakeholders in various multicultural settings, explain its reasoning, and ensure transparency[66].
Autonomy, interactivity, and adaptability are common characteristics of AI systems. We suggest that these characteristics be supplemented with the principles of accountability, responsibility, and transparency (ART) [66], as shown in figure 5.1, to reflect society concerns about AI ethics and guarantee that AI systems are created responsibly, integrating social and ethical norms.

5.3 Work done on Socially Responsible AI

Financial portfolio optimization has received a lot of attention. Many methods to developing decision assistance systems for stock trading have been explored. This includes Mean-Variance (MV)[10], TOPSIS[77], Logistic Regression[69], Autoregressive Integrated Moving Average (ARIMA) [5],Support Vector Machine (SVM)[12], [16], [58], Decision Tree[86], Long Short-Term Memory (LSTM)[65], Recurrent Neural Network (RNN) and Nerual Networks[18].

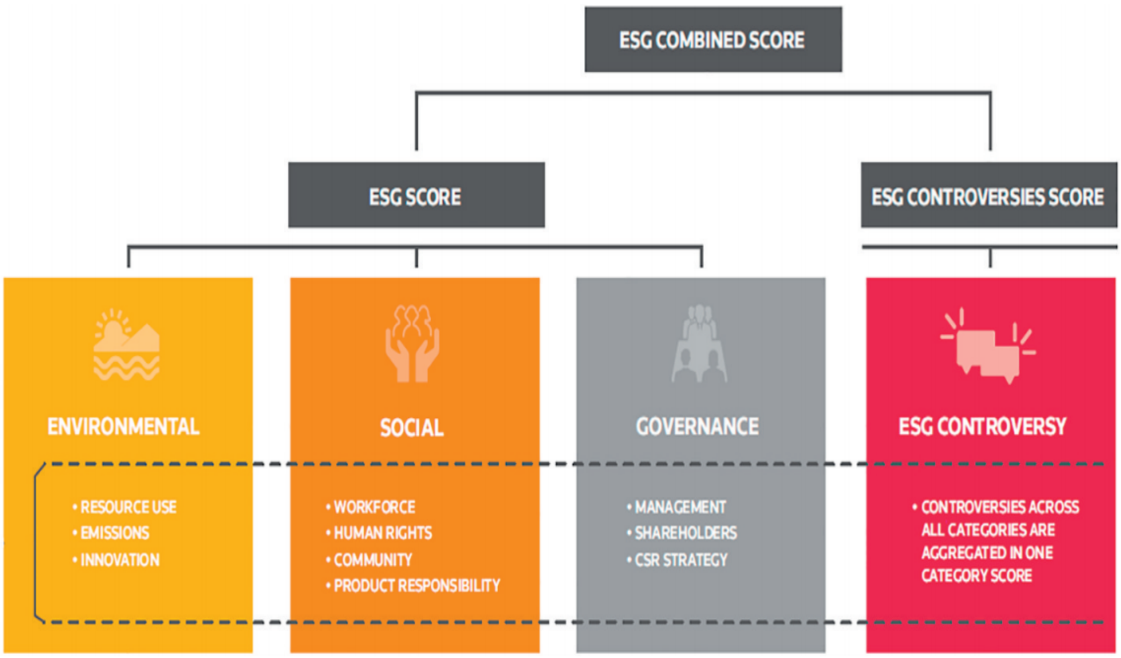


Figure 5.2: ESG rating criteria[88]

However, limited study has been done on socially responsible investing and Responsible AI. Although the concept of socially responsible investing was first suggested in the 1980s [2], it has only recently become a hot issue in academia and business [28]. During this period, studies have linked ESG ratings to company financial success [56], [70] or socially responsible fund performance [64] [48],[19]. More study and implementation in this field has been facilitated by the availability of environmental, social, and governance (ESG) ratings in industry and also the academic sector[61]. One of the earliest studies that used SRI for financial trading was done by Galema et al.[20]. The study used Fama-Macbeth, and Book-to-market regressions [20] to investigate the impact of SRI in returns of Financial assets. They aimed to clarify reasons behind the current gap between theoretical literature that suggests a link between SRI and stock returns and the majority of past research that finds no such link. The study[81], [82] uses machine learning to illustrate the impacts of automated trading in an artificially simulated agent-based stock market performed by Mizuta. There has been a recent trend in investing in companies that support CSR activities and have a high SRI score. Therefore12 advanced machine learning[78] techniques, namely Genetic Algorithm[82], Neural Networks[79], [80], and Decision Trees[87] used to evaluate Responsible AI and the impact of SRI metrics in the financial sector shown briefly in Table 5.1.

Table 5.1: Review of Responsible Investing in the Financial sector using AI & Machine Learning

Ref.	**Author(s)**	**Year**	**Dataset**	**Decision Making**	**Major Contributions**
[81],[82]	Mizuta, T.	2020	self-designed Financial Market by an AI agent	Genetic Algorithm	The AI trader found market manipulation to be the best investing technique. This means that, even if the AI trader's builder has no intention of manipulating market prices, the AI trader can discover market manipulation as an optimal investment strategy by learning with an artificial market simulation. The AI trader can learn the effects of its trades on market prices automatically.

[20]	Galema, R., et al.	2008	July 1992-June 2006 KLD Research & Analytics, Inc. data on social responsibility, and Datastream financial performance metrics	Fama-MacBeth and Book-to-market regressions	The study by Galema et al. showed that the employee relations score has a substantial beneficial impact on excess returns when it comes to SRI scores. The portfolio built based on community strength has a 3.4 percent excess return on investment (significant at the 10 percent level)
[74]	Vo, N. N., et al.	2019	The research looked at 100 companies from the Standard and Poor 500 index (S&P500) during the last 30 years, from December 31, 1988 to December 31, 2018 from Yahoo Finance	Multivariate BiLSTM Neural Networks, Reinforcement Learning	To create a reinforcement learning model, the model used deep learning methods and ESG ratings. In 2017, MAX-reinforcement ESG's learning portfolio produced a superior financial return of 50.78 percent at a reduced risk level of 19.19 percent, with a Sharpe ratio of 2.0634, proving the model's validity with real-world data.

[78]	De Franco, C., et al.	2020	Data Range: August 2009 to March 2018. The capitalization-weighted MSCI World Index USD defines the investment universe, which includes the biggest capitalization listed in the United States, Canada, Western Europe, Japan, Australia, New Zealand, Hong Kong, and Singapore.	Positive Machine Learning Screening	The Positive ML Screening beats all other portfolios on an annualized basis, outperforming the benchmark by 2.76 percent, the ESG best-in-class portfolio by 2.94 percent, and the Negative ML Screening by 4.77 percent. While actual yearly volatilities stay between 10.50 percent to 11.14 percent, the realized maximum drawdowns vary significantly: the Negative ML Screening shows a -22.47 percent loss from its peak, while the Positive ML Screening shows a -14.99 percent loss from its peak.
[87]	Guo, T., et al.	2020	2000 to 2014 Train set and January 2014 to June 2020 Test set. Unstructured data from internet news and climate data from Climate Change Report (2014) [49]	Neural Network, Natural Language Model and sentiment analysis	The paper proposes ClimateQuant deep learning framework to predict the relationship between stock behavior and climate. Compared to the benchmark portfolio the ClimateQuant portfolio had about 50% more returns while having 60% less carbon emissions

[80]	Lanza, A., et al.	2020	The study utilized the monthly total return of each stock from December 31, 2000 to April 30, 2019. The dataset was used from EURO STOXX 300.	Decision Trees	The model was used for trading for one year. Using the ESG indicators, the model had 1.2% return and had a sharpe ratio of 0.5 % in the test set. Using Environmental indicators, the model had 2.8 % return and 1.8 % in the out of sample data.
[79]	De Lucia, C., et al.	2020	The trial version of the Thomson Reuters ASSET4/EIKON database was utilized in the research. A maximum of 5000 observations may be downloaded in the trial edition.	Random forest, Artificial Neural Network (ANN), Support Vector Machine (SVM), k-Nearest Neighbour, and ridge regression	The preliminary results indicated that the majority of the algorithms could accurately estimate Return of Equity (ROE) and Return of Assets (ROA) and that the projections outperformed the baseline (the median value model). Between the fourth and tenth deciles, the authors discovered that sustainable development policy, diversity and opportunity policy, and the wage gap all had positive effects on ROE and ROA in the range of +10 percent to 16 percent.

The study by Mizuta focuses on whether an AI trader can learn to detect market manipulation even if the builder has no intention of manipulating the market[82]. This study is significant research in Responsible AI Domain. Though this was done with synthetic data, this research will pave the way towards further studies that can identify whether an AI trading bot is manipulating the market or not. [20] which was done in 2008.Vo et al. utilized Yahoo Finance to obtain both financial stock prices and public Environmental, Social Governance metrics ESG rating datasets. This is the first study that uses deep reinforcement[74] learning integrated with ESG ratings into a portfolio optimization model. The paper [78] there is a form of alpha in a company's ESG profile, but it can only be retrieved using sophisticated, non-linear methods like machine learning. The study[87] presents the ClimateQuant deep learning framework, which uses structured and unstructured climate data to predict the relationship between stock behaviour and climate. The study provides a quantitative analysis to show that ClimateQuant's investing approach reduces carbon intensity while maintaining a good return on investment. In [87] unstructured data is gathered by parsing internet news feeds for climate-related incidents or accidents affecting the businesses in our universe. The climate data for structural data were obtained from the Climate Change Report (2014) [49]. This paper[80] offers a new method for addressing existing discrepancies in ESG ratings by using Machine Learning (ML) techniques to discover variables that contribute more to creating efficient portfolios. The research[79] examines a debate that focuses on policy implications for the three major ESG indicators that were shown to have the greatest impact on ROE and ROA: 1.Environmental innovation; 2. Employment productivity; and 3. Diversity and opportunity.

5.4 Discussion

As SRI scores of the current year for a given company is not readily available or available in a vague way, some studies so far have used Artificial Intelligence to generate effective SRI scores. Some studies tried to show if the machine learning models are trying to manipulate the market and what sort of regulations be put in place so that Responsible use of AI is ensured. In contrast, other studies tried running various non-linear and deep learning methods on companies' limited available social and environmental scores to generate some form of alpha to gain an advantage over competitors while building a socially responsible portfolio.

To conclude, working with Responsible AI in the finance sector started to bloom in the year 2020. Therefore, there is much scope in this domain for research and improvement. Many studies could not generate a sizable return from investment as the SRI data available contains much noise and is unstructured. Steps to provide clean SRI data from data providers may mitigate this issue. Currently, the state-of-the-art time series models work like a black box where the internal decision-making process is unknown to the end-user. As for responsible AI, new time series models with superior explainability can help make automated trading machines more ethical and safer for society and the environment.

Chapter 6

Conclusion

6.1 Research Challenges

The Quantopian platform does not allow users to download their data; therefore, the model was restricted to the limitations of Quantopian. The highest data look back for daily trading was 200 days before the current trading day. In the case of weekly and monthly trading, the days were 150 days and 100 days, respectively. Due to these constraints, the weekly and the monthly trading algorithm could not perform as well as the daily trading algorithm.
Due to our limited resources, we were not able to use Pipeline to train our model and as such, were not able to train our models without exceeding the time limit for some algorithms. It is also difficult to implement high frequency trading such as hourly trading, as Quantopian does not provide features for hourly trading. We were also unable to implement any kind of neural networks as Quantopian would not allow us to import Keras for tensor flow. Furthermore, if we had better access to trading data, we would have been able to run our own neural network over the data for better results, but were unable to do so as Quantopian does not allow downloading of its datasets. With better resources, and better access to trading data, we would have been able to produce better and more accurate results.For future implementation purposes, we intend to design our own reinforcement learning algorithm that will be specifically tailored for this purpose. In order to get better results, we would like to try high-frequency trading, preferably minutely and hourly. We would have to calculate our own features in that case. When we start a trade, we will set an initial expected profit.

6.2 Conclusion and Future works

This thesis applied two novel methods on the stock market of the USA. In the 3rd chapter of the research, we demonstrated the feature selection method for trading with different time-frames. The results showed in the 3rd chapter reflects our success in deploying the model in a live trading environment. In the 4th chapter, the paper illustrates the weight distribution using Adaptive Boosting techniques coupled with the TOPSIS method to select an optimal portfolio. Finally, the 5th chapter of the paper discusses the work done in the ethical use of AI in Computational Finance. We hope to work on Socially Responsible Investment using AI and Machine learning in the Stock Market in the future.

Bibliography

[1] B. G. Malkiel and 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. DOI: 10.2307/2325486.

[2] H. Gray, *New directions in the investment and control of pension funds*. Investor Responsibility Research Center, 1983.

[3] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991. DOI: 10.1109/21.97458.

[4] B. I. Jacobs and K. N. Levy, "Long/short equity investing," *Journal of Portfolio Management*, vol. 20, no. 1, p. 52, 1993. [Online]. Available: https://jlem.com/documents/FG/jlem/articles/580182_LongShortEquityInvesting.pdf.

[5] J.-H. Wang and J.-Y. Leu, "Stock market trend prediction using arima-based neural networks," in *Proceedings of International Conference on Neural Networks (ICNN'96)*, IEEE, vol. 4, 1996, pp. 2160–2165. DOI: 10.1109/ICNN.1996.549236.

[6] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999. DOI: 10.1109/72.788640.

[7] C.-M. Feng and R.-T. Wang, "Performance evaluation for airlines including the consideration of financial ratios," *Journal of Air Transport Management*, vol. 6, no. 3, pp. 133–142, 2000.

[8] A. Abraham, B. Nath, and P. K. Mahanti, "Hybrid intelligent systems for stock market analysis," in *International Conference on Computational Science*, Springer, 2001, pp. 337–345. DOI: 10.1007/3-540-45718-6_38.

[9] A. Liaw, M. Wiener, *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: https://www.researchgate.net/profile/Andy_Liaw/publication/228451484_Classification_and_Regression_by_RandomForest/links/53fb24cc0cf20a45497047ab/Classification-and-Regression-by-RandomForest.pdf.

[10] M. Rubinstein, "Markowitz's" portfolio selection": A fifty-year retrospective," *The Journal of finance*, vol. 57, no. 3, pp. 1041–1045, 2002. DOI: 10.1111/j.1540-6261.1952.tb01525.x.

[11] A.-S. Chen, M. T. Leung, and H. Daouk, "Application of neural networks to an emerging financial market: Forecasting and trading the taiwan stock index," *Computers & Operations Research*, vol. 30, no. 6, pp. 901–923, 2003. DOI: 10.1016/S0305-0548(02)00037-0.

[12] K.-j. Kim, "Financial time series forecasting using support vector machines," *Neurocomputing*, vol. 55, no. 1-2, pp. 307–319, 2003. DOI: 10.1016/S0925-2312(03)00372-2.

[13] L. Rosasco, E. D. Vito, A. Caponnetto, M. Piana, and A. Verri, "Are loss functions all the same?" *Neural Computation*, vol. 16, no. 5, pp. 1063–1076, 2004. DOI: 10.1162/089976604773135104.

[14] X. Tao, H. Renmu, W. Peng, and X. Dongjie, "Input dimension reduction for load forecasting based on support vector machines," in *2004 IEEE International Conference on Electric Utility Deregulation, Restructuring and Power Technologies. Proceedings*, IEEE, vol. 2, 2004, pp. 510–514. DOI: 10.1109/DRPT.2004.1338036.

[15] T.-C. Wang and J.-C. Hsu, "Evaluation of the business operation performance of the listing companies by applying topsis method," in *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, IEEE, vol. 2, 2004, pp. 1286–1291.

[16] W. Huang, Y. Nakamori, and S.-Y. Wang, "Forecasting stock market movement direction with support vector machine," *Computers & operations research*, vol. 32, no. 10, pp. 2513–2522, 2005. DOI: 10.1016/j.cor.2004.03.016.

[17] V. Jakkula, "Tutorial on support vector machine (svm)," *School of EECS, Washington State University*, vol. 37, 2006. [Online]. Available: https://course.ccs.neu.edu/cs5100f11/resources/jakkula.pdf.

[18] M. R. Hassan, B. Nath, and M. Kirley, "A fusion model of hmm, ann and ga for stock market forecasting," *Expert systems with Applications*, vol. 33, no. 1, pp. 171–180, 2007. DOI: 10.1016/j.eswa.2006.04.007.

[19] A. Kempf and P. Osthoff, "The effect of socially responsible investing on portfolio performance," *European Financial Management*, vol. 13, no. 5, pp. 908–922, 2007. DOI: 10.1111/j.1468-036X.2007.00402.x.

[20] R. Galema, A. Plantinga, and B. Scholtens, "The stocks at stake: Return and risk in socially responsible investment," *Journal of Banking & Finance*, vol. 32, no. 12, pp. 2646–2654, 2008. DOI: 10.1016/j.jbankfin.2008.06.002.

[21] N. ALPTEKIN, "Performance evaluation of turkish type a mutual funds and pension stock funds by using topsis method," *International Journal of Economics and Finance Studies*, vol. 1, no. 2, pp. 11–22, 2009.

[22] S.-H. Hsu, J. P.-A. Hsieh, T.-C. Chih, and K.-C. Hsu, "A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression," *Expert Systems with Applications*, vol. 36, no. 4, pp. 7947–7951, 2009. DOI: 10.1016/j.eswa.2008.10.065.

[23] 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. DOI: 10.1016/j.eswa.2007.11.062.

[24] P. Ou and H. Wang, "Prediction of stock market index movement by ten data mining techniques," *Modern Applied Science*, vol. 3, no. 12, pp. 28–42, 2009. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.906.2882&rep=rep1&type=pdf.

[25] L. Bottou, “Large-scale machine learning with stochastic gradient descent,” in *Proceedings of COMPSTAT'2010*, Springer, 2010, pp. 177–186. DOI: 10.1007/978-3-7908-2604-3_16.

[26] C.-F. Tsai and Y.-C. Hsiao, “Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches,” *Decision Support Systems*, vol. 50, no. 1, pp. 258–269, 2010. DOI: 10.1016/j.dss.2010.08.028.

[27] F. Zhang, “High-frequency trading, stock volatility, and price discovery,” *Available at SSRN 1691679*, 2010. DOI: 10.2139/ssrn.1691679.

[28] N. Eccles and S. Viviers, “The origins and meanings of names describing investment practices that integrate a consideration of esg issues in the academic literature,” *Journal of business ethics*, vol. 104, no. 3, pp. 389–402, 2011. DOI: 10.1007/s10551-011-0917-7.

[29] J. Grant, “High-frequency boom time hits slowdown,” *Financial Times*, vol. 12, 2011.

[30] R. K. Mishra, S. Sehgal, and N. Bhanumurthy, “A search for long-range dependence and chaotic structure in indian stock market,” *Review of Financial Economics*, vol. 20, no. 2, pp. 96–104, 2011. DOI: 10.1016/j.rfe.2011.04.002.

[31] B. B. Nair, S. G. Sai, A. Naveen, A. Lakshmi, G. Venkatesh, and V. Mohandas, “A ga-artificial neural network hybrid system for financial time series forecasting,” in *International Conference on Advances in Information Technology and Mobile Communication*, Springer, 2011, pp. 499–506. DOI: 10.1007/978-3-642-20573-6_91.

[32] C.-F. Tsai, Y.-C. Lin, D. C. Yen, and Y.-M. Chen, “Predicting stock returns by classifier ensembles,” *Applied Soft Computing*, vol. 11, no. 2, pp. 2452–2459, 2011. DOI: 10.1016/j.asoc.2010.10.001.

[33] B. K. Bulgurcu, “Application of topsis technique for financial performance evaluation of technology firms in istanbul stock exchange market,” *Procedia-Social and Behavioral Sciences*, vol. 62, pp. 1033–1040, 2012.

[34] T. Dai, A. Shah, and H. Zhong, “Automated stock trading using machine learning algorithms,” Stanford University, Tech. Rep, Tech. Rep., 2012. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.278.5891&rep=rep1&type=pdf.

[35] F. Liu and J. Wang, “Fluctuation prediction of stock market index by legendre neural network with random time strength function,” *Neurocomputing*, vol. 83, pp. 12–21, 2012. DOI: 10.1016/j.neucom.2011.09.033.

[36] R. O'Reilly, “High frequency trading: Are our vital capital markets at risk from a rampant form of trading that ignored business fundamentals?” *The Analyst*, 2012.

[37] J. Sun and H. Li, “Financial distress prediction using support vector machines: Ensemble vs. individual,” *Applied Soft Computing*, vol. 12, no. 8, pp. 2254–2265, 2012. DOI: 10.1016/j.asoc.2012.03.028.

[38] J. J. Angel and D. McCabe, “Fairness in financial markets: The case of high frequency trading,” *Journal of Business Ethics*, vol. 112, no. 4, pp. 585–595, 2013. DOI: 10.1007/s10551-012-1559-0.

[39] R. D. A. AraúJo and T. A. Ferreira, "A morphological-rank-linear evolutionary method for stock market prediction," *Information Sciences*, vol. 237, pp. 3–17, 2013. DOI: 10.1016/j.ins.2009.07.007.

[40] A. Garg, S. Sriram, and K. Tai, "Empirical analysis of model selection criteria for genetic programming in modeling of time series system," in *2013 IEEE conference on computational intelligence for financial engineering & economics (CIFEr)*, IEEE, 2013, pp. 90–94. DOI: 10.1109/CIFEr.2013.6611702.

[41] Y. He, K. Fataliyev, and L. Wang, "Feature selection for stock market analysis," in *International conference on neural information processing*, Springer, 2013, pp. 737–744. DOI: 10.1007/978-3-642-42042-9_91.

[42] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013. [Online]. Available: http://bit.ly/statisticalLearningTheory.

[43] O. Hegazy, O. S. Soliman, and M. A. Salam, "A machine learning model for stock market prediction," *arXiv preprint arXiv:1402.7351*, 2014. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1402/1402.7351.pdf.

[44] H. Kazan and O. Ozdemir, "Financial performance assessment of large scale conglomerates via topsis and critic methods," *International Journal of Management and Sustainability*, vol. 3, no. 4, p. 203, 2014.

[45] A. Mansouri, E. Naser, and M. Ramazani, "Ranking of companies based on topsis-dea approach methods (case study of cement industry in tehran stock exchange)," *Pakistan Journal of Statistics and Operation Research*, vol. 10, no. 2, 2014.

[46] J. K. Mantri, P. Gahan, and B. B. Nayak, "Artificial neural networks–an application to stock market volatility," *Soft-Computing in Capital Market: Research and Methods of Computational Finance for Measuring Risk of Financial Instruments*, vol. 179, 2014. [Online]. Available: http://bit.ly/ANNStockMarket.

[47] M. Marien, "The second machine age: Work, progress, and prosperity in a time of brilliant technologies," *Cadmus*, vol. 2, no. 2, p. 174, 2014. [Online]. Available: https://www.proquest.com/openview/6409179cf1cc0bfc0da426b324acb95a/1.

[48] F. Munoz, M. Vargas, and I. Marco, "Environmental mutual funds: Financial performance and managerial abilities," *Journal of Business Ethics*, vol. 124, no. 4, pp. 551–569, 2014. DOI: 10.1007/s10551-013-1893-x.

[49] R. K. Pachauri, M. R. Allen, V. R. Barros, J. Broome, W. Cramer, R. Christ, J. A. Church, L. Clarke, Q. Dahe, P. Dasgupta, *et al.*, *Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change*. Ipcc, 2014. [Online]. Available: https://epic.awi.de/id/eprint/37530/1/IPCC_AR5_SYR_Final.pdf.

[50] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014. [Online]. Available: https://books.google.com.bd/books/about/Understanding_Machine_Learning.html.

[51] X. Zhang, Y. Hu, K. Xie, S. Wang, E. Ngai, and M. Liu, "A causal feature selection algorithm for stock prediction modeling," *Neurocomputing*, vol. 142, pp. 48–59, 2014. DOI: 10.1016/j.neucom.2014.01.057.

[52] M. Aitken, D. Cumming, and F. Zhan, "High frequency trading and end-of-day price dislocation," *Journal of Banking & Finance*, vol. 59, pp. 330–349, 2015. DOI: 10.1016/j.jbankfin.2015.06.011.

[53] S. Barak, J. H. Dahooie, and T. Tich, "Wrapper anfis-ica method to do stock market timing and feature selection on the basis of japanese candlestick," *Expert Systems with Applications*, vol. 42, no. 23, pp. 9221–9235, 2015. DOI: 10.1016/j.eswa.2015.08.010.

[54] V. Caivano, "The impact of high-frequency trading on volatility. evidence from the italian market," 2015. DOI: 10.2139/ssrn.2573677.

[55] R. van Est, I. van Keulen, L. Kool, A. Van Waes, F. Brom, F. van der Zee, G. Gijsbers, J. Korsten, H. Lintsen, and J. Schot, *Werken aan de robot-samenleving: Visies en inzichten uit de wetenschap over de relatie technologie en werkgelegenheid*. Rathenau Instituut, 2015. [Online]. Available: https://dspace.library.uu.nl/handle/1874/362889.

[56] G. Halbritter and G. Dorfleitner, "The wages of social responsibility — where are they? a critical review of esg investing," *Review of Financial Economics*, vol. 26, pp. 25–35, 2015, ISSN: 1058-3300. DOI: https://doi.org/10.1016/j.rfe.2015.03.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1058330015000233.

[57] M. Islamoglu, M. Apan, A. Oztel, *et al.*, "An evaluation of the financial performance of reits in borsa istanbul: A case study using the entropy-based topsis method," *International Journal of Financial Research*, vol. 6, no. 2, pp. 124–138, 2015.

[58] S. Madge and S. Bhatt, "Predicting stock price direction using support vector machines," *Independent work report spring*, 2015. [Online]. Available: https://www.cs.princeton.edu/sites/default/files/uploads/saahil_madge.pdf.

[59] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques," *Expert Systems with Applications*, vol. 42, no. 1, pp. 259–268, 2015. DOI: 10.1016/j.eswa.2014.07.040.

[60] G. Scopino, "Do automated trading systems dream of manipulating the price of futures contracts-policing markets for improper trading practices by algorithmic robots," *Fla. L. Rev.*, vol. 67, p. 221, 2015. [Online]. Available: https://scholarship.law.ufl.edu/flr/vol67/iss1/5.

[61] M. Von Wallis and C. Klein, "Ethical requirement and financial interest: A literature review on socially responsible investing," *Business Research*, vol. 8, no. 1, pp. 61–98, 2015. DOI: 10.1007/s40685-014-0015-7.

[62] S. Yuksel, H. Dincer, and U. Hacioglu, "Camels-based determinants for the credit rating of turkish deposit banks," *International Journal of Finance & Banking Studies (2147-4486)*, vol. 4, no. 4, pp. 1–17, 2015.

[63] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016. [Online]. Available: https://arxiv.org/abs/1606.06565.

[64] B. R. Auer and F. Schuhmacher, "Do socially (ir)responsible investments pay? new evidence from international esg data," *The Quarterly Review of Economics and Finance*, vol. 59, pp. 51–62, 2016, ISSN: 1062-9769. DOI: https://doi.org/10.1016/j.qref.2015.07.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1062976915000770.

[65] G. Chen, Y. Chen, and T. Fushimi, "Application of deep learning to algorithmic trading," Stanford University, Tech. Rep, Tech. Rep., 2017. [Online]. Available: http://cs229.stanford.edu/proj2017/final-reports/5241098.pdf.

[66] V. Dignum, "Responsible artificial intelligence: Designing ai for human values," 2017. [Online]. Available: https://www.itu.int/en/journal/001/Documents/itu2017-1.pdf.

[67] A. Hatami-Marbini and F. Kangi, "An extension of fuzzy topsis for a group decision making with an application to tehran stock exchange," *Applied Soft Computing*, vol. 52, pp. 1084–1097, 2017.

[68] E. Yılmaz and T. Aslan, "Evaluation of performance of tourism industry companies listed in istanbul stock exchange (bist) by topsis methodology," *EMAJ: Emerging Markets Journal*, vol. 7, no. 2, pp. 8–18, 2017.

[69] A. Zheng and J. Jin, "Using ai to make predictions on stock market," Stanford University, Tech. Rep, Tech. Rep., 2017. [Online]. Available: http://cs229.stanford.edu/proj2017/final-reports/5212256.pdf.

[70] A. Fatemi, M. Glaum, and S. Kaiser, "Esg performance and firm value: The moderating role of disclosure," *Global Finance Journal*, vol. 38, pp. 45–64, 2018, Special Issue on Corporate Social Responsibility and Ethics in Financial Markets, ISSN: 1044-0283. DOI: https://doi.org/10.1016/j.gfj.2017.03.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1044028316300680.

[71] V. Kutchu, "Growth and future of algorithmic trading in india," *Gavesana Journal of Management*, vol. 10, no. 2, pp. 83–93, 2018. [Online]. Available: https://vjim.edu.in/i/wp-content/uploads/2019/08/Gavesana-July_December_2018-issue2.pdf#page=88.

[72] R. Dash, S. Samal, R. Dash, and R. Rautray, "An integrated topsis crow search based classifier ensemble: In application to stock index price movement prediction," *Applied Soft Computing*, vol. 85, p. 105 784, 2019.

[73] K. O. Nti, A. Adekoya, and B. Weyori, "Random forest based feature selection of macroeconomic variables for stock market prediction," *American Journal of Applied Sciences*, vol. 16, no. 7, pp. 200–212, 2019. DOI: 10.3844/ajassp.2019.200.212.

[74] N. N. Vo, X. He, S. Liu, and G. Xu, "Deep learning for decision making and the optimization of socially responsible investments and portfolio," *Decision Support Systems*, vol. 124, p. 113 097, 2019. DOI: 10.1016/j.dss.2019.113097.

[75] O. R. Wearn, R. Freeman, and D. M. Jacoby, "Responsible ai for conservation," *Nature Machine Intelligence*, vol. 1, no. 2, pp. 72–73, 2019. DOI: 10.1038/s42256-019-0022-7.

[76] (2020), [Online]. Available: https://www.borsaistanbul.com/en/.%20Accessed%202020.

[77] A. K. M. Amanat Ullah, M. T. Mahtab, and M. G. R. Alam, "Efficient portfolio management using topsis and ada-boost," in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 2020, pp. 1–6. DOI: 10.1109/CSDE50874.2020.9411530.

[78] C. De Franco, C. Geissler, V. Margot, and B. Monnier, "Esg investments: Filtering versus machine learning approaches," *arXiv preprint arXiv:2002.07477*, 2020. [Online]. Available: https://arxiv.org/pdf/2002.07477.pdf.

[79] C. De Lucia, P. Pazienza, and M. Bartlett, "Does good esg lead to better financial performances by firms? machine learning and logistic regression models of public enterprises in europe," *Sustainability*, vol. 12, no. 13, p. 5317, 2020. DOI: 10.3390/su12135317.

[80] A. Lanza, E. Bernardini, and I. Faiella, "Mind the gap! machine learning, esg metrics and sustainable investment," *Machine Learning, ESG Metrics and Sustainable Investment (June 26, 2020). Bank of Italy Occasional Paper*, no. 561, 2020. DOI: 10.2139/ssrn.3659584.

[81] T. Mizuta, "An agent-based model for designing a financial market that works well," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, 2020, pp. 400–406. DOI: 10.1109/SSCI47803.2020.9308376.

[82] ——, "Can an ai perform market manipulation at its own discretion?–a genetic algorithm learns in an artificial market simulation–," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, 2020, pp. 407–412. DOI: 10.1109/SSCI47803.2020.9308349.

[83] D. Peters, K. Vold, D. Robinson, and R. A. Calvo, "Responsible ai—two frameworks for ethical design practice," *IEEE Transactions on Technology and Society*, vol. 1, no. 1, pp. 34–47, 2020. DOI: 10.1109/TTS.2020.2974991.

[84] X. Yuan, J. Yuan, T. Jiang, and Q. U. Ain, "Integrated long-term stock selection models based on feature selection and machine learning algorithms for china stock market," *IEEE Access*, vol. 8, pp. 22 672–22 685, 2020. DOI: 10.1109/ACCESS.2020.2969293.

[85] D. P. Gandhmal and K. Kumar, "Wrapper-enabled feature selection and cplm-based narx model for stock market prediction," *The Computer Journal*, vol. 64, no. 2, pp. 169–184, 2021. DOI: 10.1093/comjnl/bxaa099.

[86] A. K. M. A. Ullah, F. Imtiaz, M. U. M. Ihsan, M. G. R. Alam, and M. Majumdar, *Combining machine learning classifiers for stock trading with effective feature extraction*, 2021. arXiv: 2107.13148 `[q-fin.TR]`. [Online]. Available: https://arxiv.org/abs/2107.13148.

[87] T. Guo, N. Jamet, V. Betrix, L.-A. Piquet, E. Hauptmann, and R. A. Investments, “A deep learning framework for climate responsible investment,” [Online]. Available: https://ram-ai.com/upload/tiny/Paper/A%5C%20Deep%5C%20Learning%5C%20Framework%5C%20for%5C%20Climate%5C%20Responsible%5C%20Investment.pdf.

[88] *Thomson reuters esg scores - date of issue: February 2019*. [Online]. Available: https://www.readkong.com/page/thomson-reuters-esg-scores-date-of-issue-february-2019-4522592.