DL2

FYP Final Report

A System for Predicting Stock Price and Offering Financial Advice

by

Ho Chung Shun, Ling Hou Lam, SHI Jianhua, Tse Chun Lok

DL2

Advised by

Prof. Dik-Lun Lee

Submitted in partial fulfillment

of the requirements for COMP 4981 and CPEG 4901

in the

Department of Computer Science

The Hong Kong University of Science and Technology

2020-2021

Date of submission: April 14, 2020

Abstract

Stock price prediction is one of the most challenging problems in machine learning. In this project, we have developed a portfolio management system based on stock market prediction. Our stock prediction combines both technical analysis and sentiment analysis, which enables us to create a more accurate prediction than just using stock price data and predict the impact of financial events on stock prices. This allows us to select the best timing for buying, holding, and selling the stocks to maximize the gain and minimize losses from investment. After the prediction, our system further generates customized stock portfolios based on users' financial status, acceptable risk level and the current risks of stocks with the aim of achieving financial independence, retire early (FIRE). Our testing portfolios achieved an annualized return of 12%, indicating that using a combination of machine learning techniques and risk management strategies can generate a stock portfolio that assist investors to achieve FIRE.

Table of Contents

Abstract 3

1 Introduction 8

1.1 Overview 8

1.2 Objective 10

1.3 Literature Survey 12

1.3.1 Sentiment Analysis Using Financial News 12

1.3.2 Machine Learning for Stock Price Prediction 13

1.3.3 Risk Management 14

1.3.4 Financial Independence, Retire Early (FIRE) Movement 16

2 Methodology 17

2.1 Data Collection and Pre-processing 18

2.1.1 Design 18

2.1.2 Implementation 23

2.2 Event Analysis 27

2.2.1 Design 27

2.2.2 Implementation 27

2.3 Stock Price Prediction Models 29

2.3.1 Design 29

2.3.2 Implementation 33

2.3.3 Experiments 34

2.4 Risk Evaluation 37

2.4.1 Design 37

2.4.2 Implementation 41

2.5 Portfolio Recommendation 42

2.5.1 Design 42

2.5.2 Implementation 48

2.6 Web Application 49

2.6.1 Design 49

2.6.2 Implementation 58

2.7 Mobile Application 60

2.7.1 Design 60

2.7.2 Implementation 65

2.8 Testing 67

2.8.1 Data Collection and Pre-processing 67

2.8.2 Stock Price Prediction Models 67

2.8.3 Web Application 69

2.8.4 Mobile Application 70

2.8.5 Portfolio Management 71

2.9 Evaluation 73

2.9.1 Data Collection and Pre-processing 73

2.9.2 Stock Price Prediction Models 73

2.9.3 Visualization and Portfolio Management 73

3 Discussion 75

3.1 Data Collection 75

3.2 Stock Price Prediction Models 76

3.3 Visualization and Portfolio Management 76

4 Conclusion 78

4.1 Summary of Current Work 78

4.2 Future Research Directions 79

4.2.1 Incorporating More Factors 79

4.2.2 Applying Event Extraction 79

4.2.3 Towards Explainable Recommendations 79

4.2.4 Providing High Concurrency and Low Latency System 80

5 References 81

6 Appendix A: Technical Analysis 84

7 Appendix B: Format of Stock Price Data 87

8 Appendix C: Inputs to the Machine Learning model 88

9 Appendix D: Project Planning 89

9.1 Distribution of Work 89

9.2 GANTT Chart 91

10 Appendix E: Required Hardware & Software 92

10.1 Hardware 92

10.2 Software 92

11 Appendix F: Meeting Minutes 93

11.1 Minutes of the 1st Project Meeting 93

11.2 Minutes of the 2nd Project Meeting 94

11.3 Minutes of the 3rd Project Meeting 95

11.4 Minutes of the 4th Project Meeting 96

11.5 Minutes of the 5th Project Meeting 98

11.6 Minutes of the 6th Project Meeting 100

11.7 Minutes of the 7th Project Meeting 102

11.8 Minutes of the 8th Project Meeting 104

11.9 Minutes of the 9th Project Meeting 105

11.10 Minutes of the 10th Project Meeting 107

11.11 Minutes of the 11th Project Meeting 108

11.12 Minutes of the 12th Project Meeting 109

11.13 Minutes of the 13rd Project Meeting 110

11.14 Minutes of the 14th Project Meeting 112

11.15 Minutes of the 15th Project Meeting 113

11.16 Minutes of the 16th Project Meeting 114

11.17 Minutes of the 17th Project Meeting 116

11.18 Minutes of the 18th Project Meeting 117

11.19 Minutes of the 19th Project Meeting 118

11.20 Minutes of the 20th Project Meeting 119

11.21 Minutes of the 21st Project Meeting 120

11.22 Minutes of the 22nd Project Meeting 121

11.23 Minutes of the 23rd Project Meeting 122

11.24 Minutes of the 24th Project Meeting 123

11.25 Minutes of the 25th Project Meeting 124

11.26 Minutes of the 26th Project Meeting 126

11.27 Minutes of the 27th Project Meeting 127

11.28 Minutes of the 28th Project Meeting 129

11.29 Minutes of the 29th Project Meeting 130

11.30 Minutes of the 30th Project Meeting 131

11.31 Minutes of the 31st Project Meeting 132

1 Introduction

1.1 Overview

Stock price prediction is one of the most challenging problems because many factors affect stock prices and are usually unknown in advance. Nevertheless, many people have tried to predict stock price with different strategies and algorithms because correctly predicting future stock price or even just the trend of the stock is extremely useful for gaining profits and reducing risks during financial investment. Therefore, predicting stock prices has been the main target of investors since the birth of the stock market.

Traditionally, financial analysts utilize two separate methods of predicting stock prices. This includes technical analysis and fundamental analysis. Technical analysis examines and predicts the price movement in financial markets using historical prices and market statistics. Technical indicators such as RSI and MACD are used to identify market movement patterns to predict the future price trends of a stock. The indicators indicate whether the stock is in an overbought or oversold stage so as to perform selling or buying action accordingly. Whereas fundamental analysis focuses on the stock's true value, which takes external factors and intrinsic value into consideration. Fundamental analysis believes that financial reports, industry trends and the global economy play a significant role in affecting the price of stocks. Both methods are usually separately considered and are fairly accurate in predicting the price of a stock.

With the rise of machine learning, many researchers have started to devise machine-learning based strategies for predicting the stock market. The two types of analyses are used along with machine learning to produce a more accurate prediction of the stock prices. Technical analysis uses models such as random forest (RF), support vector machine (SVM), and recurrent neural network (RNN) to analyze historical stock prices and solve the original problem as a time series problem [1]. Recent fundamental analysis uses models like natural language processing (NLP) or sentiment analysis to extract useful information from textual data such as news related to the stock and financial reports to predict whether the stock will rise or fall [2].

In the recent years, more and more people are following the concept of "Financial Independence, Retire Early." (FIRE). The goal is to save up the majority part of the income and invest in the savings so that people can retire early at around 30s to 40s instead of the traditional retirement age of 65. The project follows the goal of FIRE to recommend a suitable portfolio.

1.2 Objective

The goal of this project is to create a financial prediction and advice system where both long-term investors and novice investors can take advantage of performing better action to increase profit while mitigating risks during stock investment. Our project mainly focuses on the following objectives:

1. Develop a system that automatically and regularly collects price and news data for different stocks.
2. Utilize machine learning techniques to predict the upcoming price of different stocks using the collected data.
3. Visualize the predictions, giving portfolio recommendations based on the financial status of users and providing trading alerts.

To achieve the first objective, we have found useful APIs for collecting stock data. Automation programs have been developed based on these APIs to crawl data automatically and save it into a database.

To achieve the second objective, we first pre-processed the data we obtained in objective 1 to extract useful features such as generating technical indicators and sentiment features. We then used this pre-processed data to train our model for each stock. Different model configurations have been tested and the settings with the highest accuracy were selected and deployed.

To achieve the third objective, we have provided a user-friendly website to display different stock recommendations and user profiles as well as stock advice based on individually calculated risk indexes. We have also provided a mobile application for investors to receive instant notifications about the latest stock news in order for them to react quickly to market changes.

The biggest challenge we faced is the machine learning model. To address this challenge, different machine learning architecture has been examined, tested, and evaluated to select the best model for final implementation in the system.

1.3 Literature Survey

We have found the following topics and systems related to our project.

1.3.1 Sentiment Analysis Using Financial News

1.3.1.1 Importance of financial news in Stock Price Prediction

There are various reasons leading to the change of stock price of a stock, including demand and supply, investors' expectation, dividends, global economy, and political climate. These aspects are intertwined to affect the stock market. One of the most useful data that includes all these events is the stock news. Stock news affects how people view and perceive one company's value. One example is that if the company released a better-than-expected financial report, the news will report such event to the public, and people will have higher expectation on the future of the company and believe that the stock is worth more than it is listed right now. Therefore, more people purchase that stock leading to a greater demand than supply, the price of the stock thus increases until reaching an equilibrium.

This research [3] found out that financial news is highly correlated to the change in the stock price of a stock. Therefore, we believe that the incorporation of financial news in stock price prediction is very important and can explain the changes in the stock price.

1.3.1.2 Sentiment Analysis for Text Feature Extraction

Sentiment analysis extracts sentimental features from a piece of text. Previously, economic and finance researchers used a finance-domain-specific dictionary, called the Loughran and McDonald dictionary, to determine the sentiment of a piece of financial text [4]. This method outperforms the generic bag-of-words approach in classifying sentiments. Recent research [5] showed that Bidirectional Encoder Representations from Transformers (BERT) [6] and BERT finetuned with financial news (FinanceBERT) has outperformed the Loughran and McDonald dictionary by 20% absolute accuracy in sentiment classification task. Furthermore, FinanceBERT achieved 3% higher accuracy than BERT. This indicates the effectiveness of applying finetuned BERT in solving downstream tasks such as sentiment classification.

1.3.2 Machine Learning for Stock Price Prediction

Recurrent Neural Network (RNN)

A RNN is an extension of feedforward neural network for handling sequential data typically involving variable-length input or output sequences. It allows previous outputs to be used as future input of the model, which makes RNN one of the most powerful models for processing sequential data. However, vanilla RNN has the problem of vanishing gradient which increases the difficulty of training the model. Long Short-Term memory (LSTM) is one of the most successful RNNs architectures introduced by Hochreiter & Schmidhuber in 1997 which prevents the vanishing gradient problem. It is a very advanced solution to sequence and time series related problems like stock price prediction. Much early research, like Saloni Mohan (2019) [7], have used LSTM as their models for technical and fundamental analyses of stocks and achieved fairly good results.

1.3.3 Risk Management

In order to recommend suitable investment portfolios for an investor, a specification of investment objectives and quantitative evaluations for the financial status of the investor and risk of different stocks could be referenced.

For financial status, it is essential to collect the information that helps to create a cash flow statement and a balance sheet for the investor. With the cash flow statement, we can tell the net income of a person and emergency savings needed in the portfolio. With the balance sheet, we can tell the assets, liabilities and the net worth of the investor [8].

We can say,

$$\text{financial health} \propto \frac{\text{cash flow}}{\text{debt equity ratio}}$$

Then, the risk investor expected, financial health and investment objectives are taken into concern for the expected risk level of the portfolio [9]:

$$\text{expected risk level of the portfolio}$$
$$\propto \frac{(\text{amount of investment goal}) \times (\text{risk investor expected}) \times (\text{financial health})}{\text{expected investment period}}$$

For the risk of different stocks, an analysis of the fundamental and technical sides of the stocks should be considered. The risk of a stock can be considered according to the factors in Table 1 [10]:

Factor	**Description**
Volatility	The volatility of the assets of the company including stocks
Momentum	The recent performance of the stock
Size	The market capital of the stock, indicating large stocks from small stocks
Liquidity	The recent volume traded of the stock
Growth	The past and anticipated earnings growth of the company
Value	Analysis of the fundamentals of the company, for example, ratios of earnings, dividends, cash flows, book value, sales, etc.
Earnings volatility	The volatility of the earnings of the company
Financial leverage	The debt-to-equity ratio of the company

Table 1: Factors Affecting Risk

1.3.4 Financial Independence, Retire Early (FIRE) Movement

The concept of "Financial Independence, Retire Early (FIRE)" is first originated from the 1992 book Your Money or Your Life by Vicki Robin and Joe Dominguez [11]. The main theory is to save and invest very aggressively in order to retire between the age of 30 and 40, which is significantly earlier than the traditional retirement age of 65. There has been an increased amount of discussion about the FIRE movement in recent years especially by the Millennials [12]. This movement suggests three key areas for achieving FIRE including keeping expenses low, boosting income, prioritizing saving, and investing [13]. In this project, we will focus on the investment side to help users achieve financial independence.

The investment goal is determined by the amount of annual expenditure. It is generally suggested to accumulate approximately 25-30 times the annual expenses to achieve FIRE [14]. This also suggests that one can withdraw up to 4% of the amount saved in the investment portfolio every year without running out of money. To account for inflation, some suggests that a portfolio that generates 8% return every year can be applied to achieve financial independence [15].

2 Methodology

Figure 1 and Figure 2 provide a comprehensive view of the system flow and the system architecture respectively.

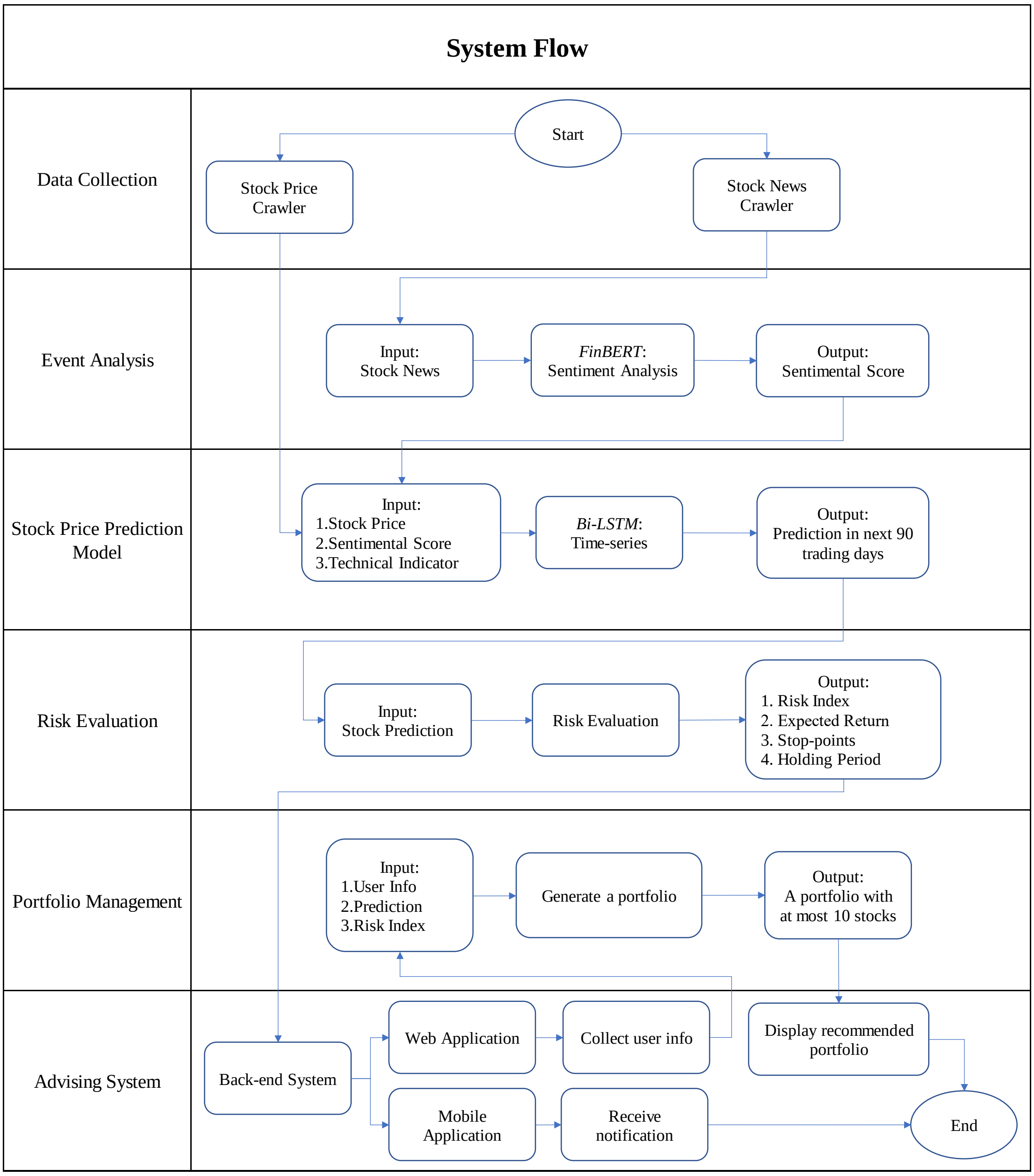

System Flow
Data Collection
Start
Stock Price Crawler
Stock News Crawler
Event Analysis
Input: Stock News
FinBERT: Sentiment Analysis
Output: Sentimental Score
Stock Price Prediction Model
Input: 1.Stock Price 2.Sentimental Score 3.Technical Indicator
Bi-LSTM: Time-series
Output: Prediction in next 90 trading days
Risk Evaluation
Input: Stock Prediction
Risk Evaluation
Output: 1. Risk Index 2. Expected Return 3. Stop-points 4. Holding Period
Portfolio Management
Input: 1.User Info 2.Prediction 3.Risk Index
Generate a portfolio
Output: A portfolio with at most 10 stocks
Advising System
Back-end System
Web Application
Collect user info
Display recommended portfolio
Mobile Application
Receive notification
End

Figure 1: System Flow

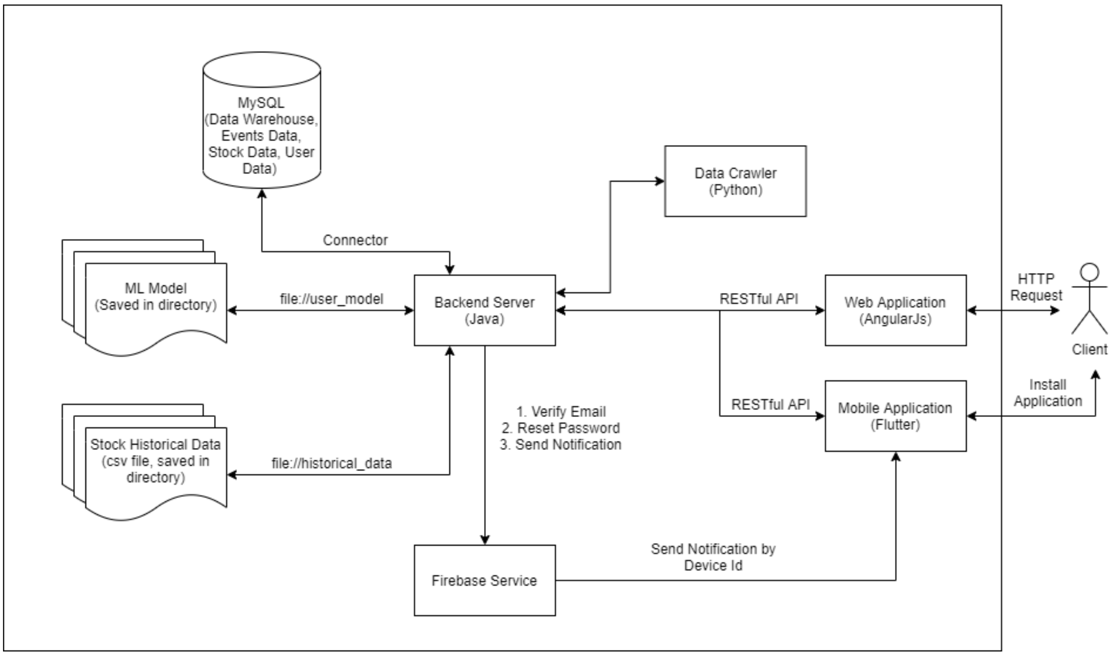

MySQL (Data Warehouse, Events Data, Stock Data, User Data)
Data Crawler (Python)
Connector
ML Model (Saved in directory)
file://user_model
Backend Server (Java)
RESTful API
Web Application (AngularJs)
HTTP Request
Client
Install Application
RESTful API
Mobile Application (Flutter)
1. Verify Email
2. Reset Password
3. Send Notification
Stock Historical Data (csv file, saved in directory)
file://historical_data
Firebase Service
Send Notification by Device Id

Figure 2: System Architecture Overview

2.1 Data Collection and Pre-processing

2.1.1 Design

2.1.1.1 Stock Market Selection

This project is expected to be applicable to every stock market in theory. With stock historical data in different markets, the model should also provide reliable financial advice. However, there are some other factors like the total trading volume, total market capitalization or even most of the industries in that country that vary in different markets, the result may be less accurate with the variance of these factors.

For the selection of the stock market, to find the most stable stock market for our current project goal, we have 3 main criteria which are the total market capital, number of stocks and diversity of stocks.

Therefore, our stock analysis focuses on US stocks. By the end of 30/06/2020, the total market capitalization of U.S. stocks is 35,503,373 [16]which is the highest among all the stock markets in the world. In 2020, there are 3,671 stocks in the US market, they can be separated into 11 sectors and further classified into 156 sub-industries [17]. Therefore, the US market should be more comprehensive compared to other stock markets, there are stocks with market capitalization from primary, secondary and tertiary industries.

For the selection of the stocks, as mentioned above, there are over 3,500 stocks in the US stock market, it might not be necessary to analyze all of them. It is not efficient for our system to keep track of every US stock to have real-time or daily analysis, it requires way more computing power. Also, there are some stocks that are not suitable for our prediction. Some less well-known companies may not have enough news for analysis. Besides, stocks with less market capital are more easily manipulated by some holding companies. The stocks with high volatility are not suitable neither, as our project targets to mostly investors who are not able to spend too much time monitoring the market. They might miss the chance of buying or selling out this kind of stock.

GICS Sector	**Count of GICS Sector**	**% of Market Cap in S&P500**	**Sum of Market Cap**	**% of Overall Market Cap**
Communication Services	26	14.75%	4,511 B	12.71%
Consumer Discretionary	61	11.31%	3,460 B	9.74%
Consumer Staples	33	7.40%	2,263 B	6.37%
Energy	26	2.36%	721 B	2.03%
Financials	66	9.55%	2,921 B	8.23%
Health Care	62	13.21%	4,041 B	11.38%
Industrials	73	7.50%	2,292 B	6.46%
Information Technology	71	26.31%	8,045 B	22.66%
Materials	28	2.37%	726 B	2.04%
Real Estate	31	2.49%	762 B	2.15%
Utilities	28	2.75%	840 B	2.37%
Grand Total	**505**	**100.00%**	**30,582 B**	**86.14%**

Table 2: Sector Details of US Stock Market

Stocks in S&P500 are chosen for analysis and portfolio recommendation. S&P500 is one of the important indexes for estimating the overall performance of the US stock market. As shown in Table 2, the stocks in S&P500 occupy around 80% of the total market capitalization and usually almost half of the trading volume in the market [18]. Therefore, they are more popular and stable for the investors, so it should be more efficient if our system focuses on S&P500 alone.

The list of stocks in S&P500 is constantly changing due to corporate activities such as company acquisition and therefore this project only considers stocks in the S&P500 list before October 12, 2020 and have at least 5 years of historical stock price data and is not delisted throughout the project period. A total of 460 stocks satisfied the requirements above.

2.1.1.2 Types of Data Required

In order to build a more robust stock price prediction system, technical indicators which are commonly used by financial analysts to determine the state of a stock are necessary. These technical indicators can be generated by using historical price data. As a result, 10 years of historical stock price data are collected on this project.

In addition, stock prices can be largely affected by news because investors often obtain stock information from the news. Positive news, such as a good earnings report and announcement of new products will generally cause investors to buy stocks. Stock prices will then increase with respect to demand. In contrast, negative news can reduce the likelihood of people buying their stocks, which can have a tremendous negative impact on their stock price. Therefore, it is very important for a clever investor to analyze real time and historical news to predict stock prices.

Moreover, the Efficient Market Hypothesis (EMH) states that the current market price fully reflects all the recently published news, which means that past and current information is being immediately incorporated into stock prices. Therefore, news data is also collected on our project.

2.1.1.3 Technical Indicators

Technical indicators in Table 3 are calculated from the historical stock prices. The description of these technical indicators is provided in the Appendix. These indicators are selected as they reflect the trend, momentum, volatility and the volume of a stock.

Technical Indicator	**Mathematical Formula**
Simple Moving Average SMA(t,N)	$\frac{\sum_{=1}^{N}(t-k)}{N}$
Exponential Moving Average EMA(t, Δ)	$\big(pc(t) - EMA(t-1)\big) * mult + EMA(t-1), mult = \frac{2}{\Delta + 1}, \Delta = \mathrm{timeperiodEMA}$
Parabolic Stop and Return	$\mathrm{Rising\ PSAR} = \mathrm{PSAR_{prev}} + \left[\mathrm{AF_{prev}}\left(\mathrm{EP_{prev}} - \mathrm{PSAR_{prev}}\right)\right]$ $Falling\ PSAR = PSAR_{prev} - \left[AF_{prev}\left(PSAR_{prev} - EP_{prev}\right)\right]$ $where$: $AF, starts\ at\ 0.02, increase\ by\ 0.02, up\ to\ maximum\ of\ 0.2,$ $each\ time\ EP\ is\ recorded.$ $EP, lowest\ point\ in\ current\ downtrend\ or\ highest\ point\ in\ the\ current\ uptrend.$
MACD(t)	$EMA(t, \Delta = 12) - \mathrm{EMA}(\mathrm{t}, \Delta = 16)$
Relative Strength Index RSI(t)	$\frac{100}{1 + RS(t)}, RS(t) = \frac{AvgGain(t)}{AvgLoss(t)}$
Bollinger Bands	$UpperBand(t) = 20 \times SMA(t, N) + 40 \times std(pc)$ $MiddleBand(t) = 20 \times SMA(t, N)$ $LowerBand(t) = 20 \times SMA(t, N)\big(40 \times std(pc)\big)$
Stochastic Oscillator KDJ(t)	$100\frac{pc(t) - MIN(pl)}{MAX(ph) - MIN(pl)}$
Average True Range ATR(t)	$TR(t) = MAX\big(ph(t) - pl(t); ph(t) - pc(t-1); pc(t-1) - pl(t)\big)$ $ATR(t) = \frac{13 \times ATR(t-1) + TR(t)}{14}$
Williams Indicator WR(t)	$\frac{MAX(ph) - pc(t)}{MAX(ph) - MIN(pl) * (-100)}$
On Balance Volume	$OBV = OBV_{prev} + \begin{cases} volume & if\ pc > pc_{prev} \\ 0 & if\ pc = pc_{prev} \\ -volume & if\ pc < pc_{prev} \end{cases}$
Chaikin Money Flow	$CLV = [(pc - pl) - (ph - pc)]/(ph - pl)$ $\mathrm{CMF} = \frac{\sum_{t=20}^{t} CLV_t \times volume_t}{\sum_{t=20}^{t} volume_t}$

With pc=close price, po=open price, pl=lowest price, ph=highest price and std=standard deviation

Table 3: Technical Indicators

2.1.2 Implementation

Figure 3 outlines the procedures of generating the input data required for model training and prediction.

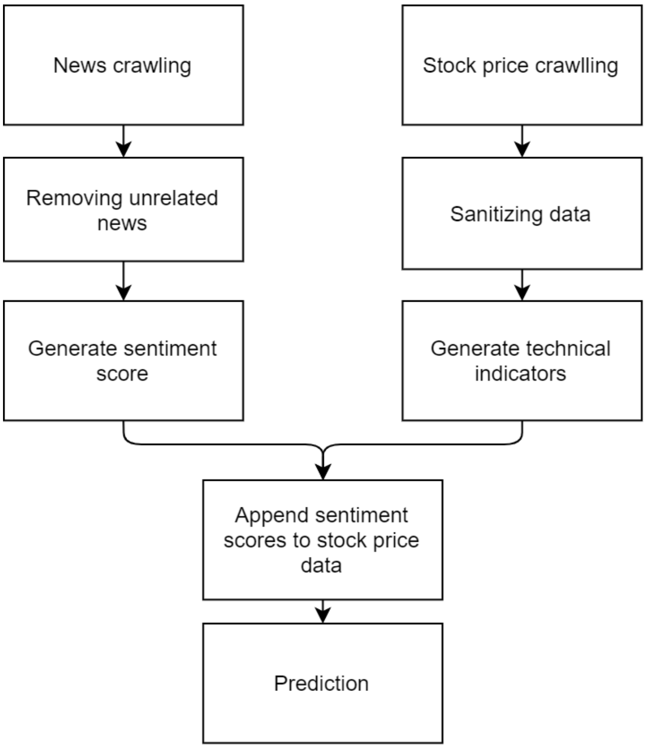


Figure 3: Data Preprocessing Flow Diagram

2.1.2.1 Stock Price Data

2.1.2.1.1 Collecting Stock Price Data

For market data, we used *yfinance* API to collect historical stock data from Yahoo Finance. This API is free, and it is possible to collect stock price data at different time intervals. As our goal focuses on long term investors which care less about short term stock price changes, we used daily data for this project instead of hourly or minute data. We have downloaded 10 years of daily historical stock price data and current data is collected daily after the market closes. The format of the data can be found in the Appendix.

2.1.2.1.2 Sanitizing Data

We created an automated check to see if there is any missing data inside the dataset. The missing data is filled with the average of the previous and next day using *Pandas* library in Python.

2.1.2.1.3 Generating Technical Indicators

The technical indicators are calculated using *Technical Analysis* library in Python [9]. The data generated is then concatenated with the stock price data.

2.1.2.2 Stock News Data

2.1.2.2.1 Collecting Stock News Data

Two different methods are used to collect historical and real-time stock news data. Only financial news on Yahoo Finance is crawled so that we can focus on news related to the financial market only and prevent collecting duplicates of news from different sources. Python is used as a programming language as it provides many convenient libraries for crawling data from the internet. *Requests* are used to send the HTTP requests to retrieve the webpage and *BeautifulSoup* is used to extract news passages. *Pandas* is used to store the data in CSV format and *Requests* is used to send the data to our backend server.

For historical stock news data, Google News is used to collect financial news because it provides convenient search with time range, while other free APIs only provide limited requests, e.g. 500 requests/day of news within one month for News API and 1 bulk requests/minute for Intrinio API, which are insufficient for our project.

However, Google blocked our crawlers with too many requests' errors and reCAPTCHAs when our crawlers have sent more than 70 requests. Therefore, we lowered the frequency of crawling and used multiple computers and IP addresses to crawl historical news data. We even used Virtual Private Network (VPN) during the crawling and automatically switched to another VPN server upon receiving too many request errors from Google using *NordVPN Switcher* in Python. With this technique, we crawled news data of all the stocks for more than 7 years. It is appended to real-time news data after completing the crawling of historical news.

For real-time stock news data, Finviz has listed about a hundred latest financial news of the stock in their stock page. They are sorted with time and stored statically on the page. More importantly, it does not block our bots while crawling. Therefore, we have used Finviz's stock page to crawl latest financial news about the stocks in S&P 500. News is crawled multiple times daily as some stocks may have a lot of news.

2.1.2.2.2 Filtering / Removing Unrelated News

Although Google search was very convenient to crawl historical news, it searched the entire page of Yahoo Finance to find relevant words. As the news page of Yahoo finance does not only contain news content we need, but also some thumbnails of other news and advertisement, Google search might have returned irrelevant news that was crawled with respect to our target stocks. Therefore, we filtered out this kind of news that was unrelated.

To filter out unrelated news, the number of occurrences of the relevant stock name, stock quote and company name inside the news content and title were counted. A piece of news article is regarded as irrelevant if it contains less than a fixed number throughout the entire document. Irrelevant news was removed and was not used in our model.

2.2 Event Analysis

2.2.1 Design

It is difficult for us to analyze an entire piece of news article directly in our prediction model. Therefore, it is necessary to extract meaningful information from each piece of news before using it. We have chosen sentiment analysis as the method of extracting information from the news articles because it is easier to identify the stock event as being either positive, neutral, or negative.

2.2.2 Implementation

FinBERT is the state-of-the-art sentiment analysis model that can be directly applied to our system. It is a pre-trained model that is specifically used to tackle financial natural language processing tasks. This model is generated by further training Bidirectional Encoder Representations from Transformers (BERT) [6] using financial domain specific texts. FinBERT has been validated to outperform BERT in financial texts sentiment analysis and other research had found out that finetuned BERT significantly outperforms finance-domain-specific dictionary-based analysis such as the Loughran and McDonald dictionary [5].

Thus, FinBERT is used in this project to perform financial sentiment analysis and generate sentiment scores for our news, which is treated as the current financial atmosphere of the stock event. A positive sentiment should correlate to an increase in stock price and a negative sentiment should correlate to a decrease in stock price. A pipeline is created to streamline the process of applying FinBERT for generating sentiment scores.

2.2.2.1 Applying Statistical Analysis

As there could be multiple news articles every day, the sentiment scores were grouped together so that there is a fixed number of features for the input of the model. Our group used different statistical analyses such as polarity scores from –1 to 1, and mean scores to combine those sentiment scores.

2.2.2.2 Appending to Stock Price Data

After performing all sentiment and statistical analysis, we incorporated the results into the stock price data according to the closest next trading day of that stock.

2.3 Stock Price Prediction Models

2.3.1 Design

We have designed multiple machine learning models to find the correlation between the historical stock price data and news data. The performance of different models has been evaluated. The best model and configurations generated by conducting experiments are used for the actual prediction.

The model is designed to predict P_{t+90} based on P_{t-w} to P_t where P_t is the stock price on the t-th trading day and w is the window size. Prediction is performed for 90 times to generate 90 days prediction from P_{t+1} to P_{t+90}.

2.3.1.1 Model Architecture

Due to the statement of Efficient Market Hypothesis and Adaptive Market Hypothesis, the stock market is mainly determined by two factors: stock price and news. Therefore, the input to the models includes the stock price, technical indicators generated using stock price and sentiment score generated from news articles. The exact list of inputs can be found in the Appendix. After the data has been inputted to the model, it generates our prediction result of P_{t+90}.

We have tried four different models for our project, Recurrent Neural Network (RNN), Unidirectional LSTM (Uni-LSTM), Bidirectional LSTM (Bi-LSTM) and Gate Recurrent Unit (GRU).

2.3.1.1.1 Recurrent Neural Network

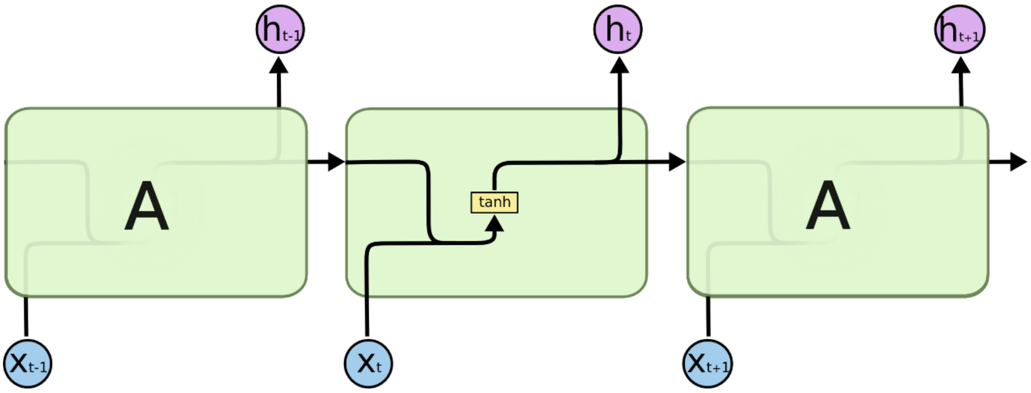


Figure 4: RNN Model

X_t denotes the input data for date t and h_t denotes the output P_{t+90}

RNN is designed to work with sequential data. As shown in Figure 4, it uses previous information in sequence combined with current input to generate the current output which is useful in time series prediction. However, this architecture suffers from a vanishing gradient problem. This is because when training an RNN, backpropagation through time is used so that gradients update from the final layer back to the first layer. The gradients can be exponentially decreasing or increasing due to their multiplicative nature. This results in the inability to capture long term dependencies.

2.3.1.1.2 Unidirectional LSTM

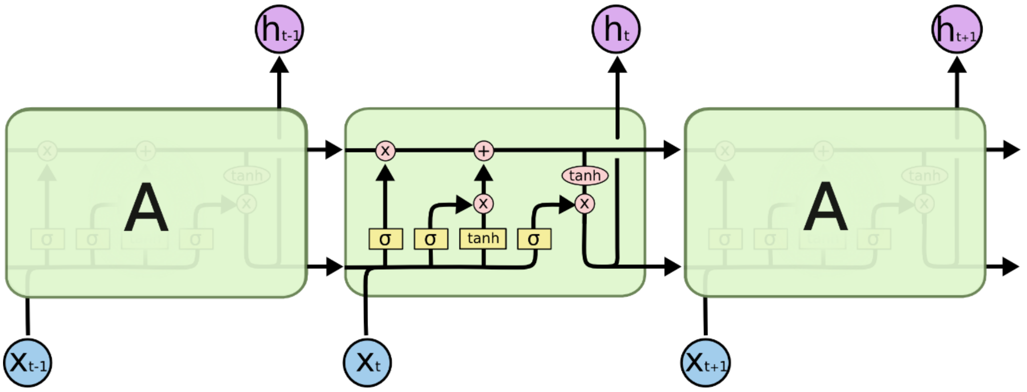


Figure 5: Uni-LSTM Model

X_t denotes the input data for date t and h_t denotes the output P_{t+90}

Compared to RNN, LSTM has three additional gates in its cell. This includes the input gate, output gate and forget gate as shown in Figure 5. LSTM has better performance than RNN in time series prediction because it is not only considering the short-term effects, but also the long-term effects. Thus, it can solve the long-term dependency problem in RNN. Hence, it should fit well with the situation of the stock market.

2.3.1.1.3 Bidirectional LSTM

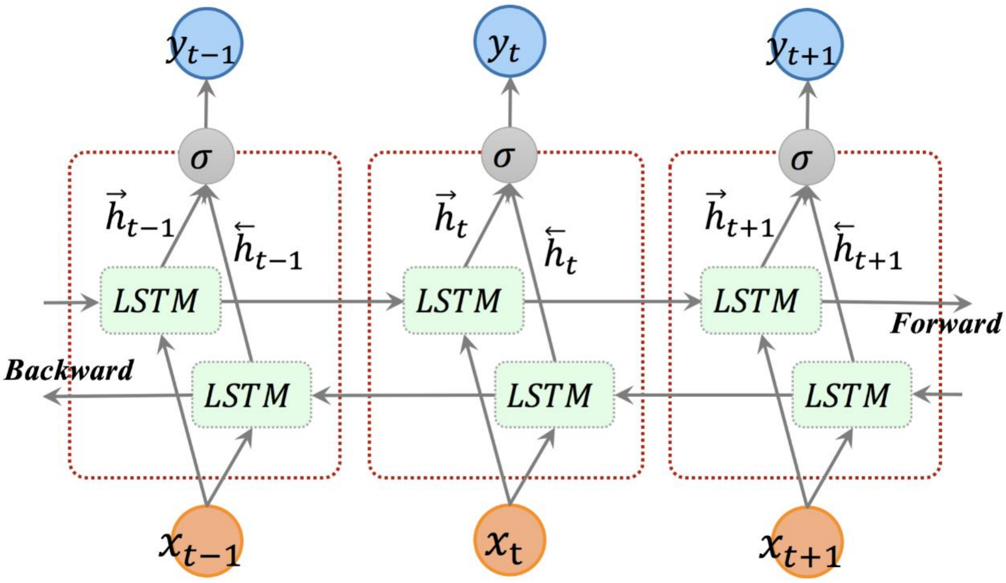


Figure 6: Bi-LSTM Model

X_t denotes the input data for date t, h_t denotes the output from the LSTM cells, the sigmoid function concatenates the two outputs to produce y_t which denotes P_{t+90}

Bi-LSTM is capable of capturing information not only from the backward direction but also the forward direction. This allows information to be better preserved compared to Uni-LSTM and thus Bi-LSTM usually outperforms Uni-LSTM.

2.3.1.1.4 Gate Recurrent Unit

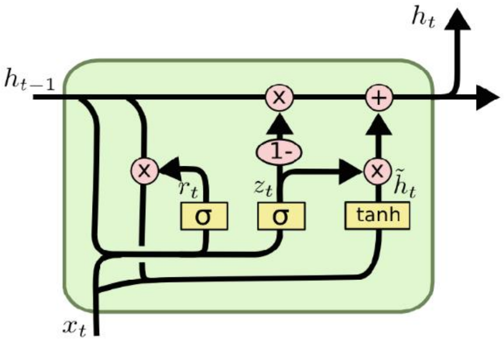


Figure 7: GRU Model

X_t denotes the input data for date t and h_t denotes the output P_{t+90}

GRU has a similar architecture to LSTM except it lacks an output gate. This leads to GRU having fewer parameters than LSTM and is able to reduce the training time as there are less units that have to be learned.

2.3.1.2 Evaluation Methodology

Mean Absolute Percentage Error (MAPE) is used as the loss function for training the models. MAPE is a popular measure of the prediction accuracy of a forecasting method in statistics, specifically in trend prediction. The mathematical interpretation is in terms of relative error as shown below:

$$MAPE = \frac{100}{n} \sum_{j=1}^{n} \left| \frac{y_j - y'_j}{y_j} \right| \%,$$

where n is the number of observations, y_j is the actual value and y_j' is the predicted value. MAPE provides errors in terms of percentages. It considers the ratio between error and the actual value, not only the value difference.

Using MAPE reduces the large deviation bias present in Root Mean Square Error (RMSE). MAPE only considers the difference between predicted values and observed values while the direction of the difference can be ignored.

The accuracy of the model is defined below:

$$Accuracy = (1 - MAPE) * 100\%$$

2.3.2 Implementation

Based on our experience with machine learning, we used *PyTorch* to build the models. We have randomly chosen a stock to evaluate and used grid search to finetune the model hyper-parameters such as hidden sizes, number of layers, dropout rate, window size and only store the model settings with lowest MAPE Loss. With this method, we are able to generate 90 days of predictions with reasonable accuracy.

2.3.2.1 Generating Datasets

We separated the entire dataset to generate training, validation, and testing data. We use the last 20 percent data as testing output and combine 720 previous data to generate a testing dataset. The last 20 percent of the previous data is used similarly to generate our validation set, and the remaining data is used as our training set. This ensures that there is sufficient data for validating and testing the model.

2.3.2.2 Feature Normalization

MinMaxScalers in *Scikit Learn* is applied to normalize the data to between 0 and 1. All the features are transformed using this scaler.

2.3.2.3 Activation Function

One important note is that we used linear activation for the output. We found out that this produces better results than using any non-linear activation functions since the prediction is a regression task.

2.3.3 Experiments

In order to search for the best model architecture and configurations that achieves the lowest validation loss, we deployed grid search to find out the optimal settings and hyperparameters systematically using *Optuna Framework*. Different architectures are compared including RNN, GRU, unidirectional LSTM and bidirectional LSTM.

The following table displays the hyperparameters that are tested out by grid search:

Hyperparameters	Configurations
Number of Hidden Units	[32, 64, 128]
Number of Layers	[1, 2, 3, 4]
Dropout rate	[0.1, 0.2, 0.4, 0.5]
Learning rate	[0.005, 0.01, 0.02, 0.05]
Window size	[90, 180, 360]

Table 4: Hyperparameters and Configurations

The following table displays the architecture and the lowest validation lost in grid search:

Model	Validation Loss
RNN	0.124
GRU	0.153
Unidirectional LSTM	0.122
Bidirectional LSTM	0.111

Table 5: Validation Loss of Different Models

Among the experiments, Bidirectional LSTM with the configurations in the following table achieved the lowest validation loss.

Hyperparameters	Configuration
Number of Hidden Units	32
Number of Layers	2
Dropout rate	0.5
Learning rate	0.01
Window size	90

Table 6: Configuration of Best Model

This model's architecture and settings are used for all of the stocks.

We further experimented with which combination of the datasets produces the least validation loss. We would like to compare whether stock price, technical indicators and sentiment scores combined would give the lowest validation loss when compared to any other input configurations. We used the best model architecture and settings from the grid search. Another stock is randomly chosen for this experiment, so the validation loss across the models is different from the previous experiment.

The experiment details are shown in the table below:

Model	Validation Loss
Bi-LSTM stock price only	0.251
Bi-LSTM stock price and sentiment score	0.258
Bi-LSTM stock price and technical indicators	0.217
Bi-LSTM stock price, technical indicators, and sentiment score	0.199

Table 7: Validation Loss of Models with different inputs

The experiment shows that the combined approach has higher accuracy over the models with technical analysis or sentiment analysis only. We input all of the features when training the final machine learning model.

2.4 Risk Evaluation

2.4.1 Design

2.4.1.1 Calculating Risk Index

No matter how sophisticated the algorithm is, there will always be errors in the prediction results. Therefore, to minimize the consequences of the error, stock risk indexes are calculated for the user as a reference. The formula uses the concepts of stop points and risk management.

A stop point is a selling price set based on the recent movement of a stock in the market. It includes upper stop and down stop. The upper stop is generated based on the results of the machine learning forecast.

$$\text{Upper stop } = Min(Lowest\ price\ in\ predicted\ 90\ days\ + 4 * ATR(90 days),\ Highest\ price\ in\ predicted\ 90\ days)$$

The down stop is a stopping point calculated based on historical price data to protect against potential losses due to forecast errors or large short-term fluctuations in the stock market. To calculate the down stop, Chandelier Stop and YoYo Stop are used. Chandelier Stop reflects the market fluctuations within a period, whereas YoYo Stop reflects the immediate market change by using last day's closing price. Down stop takes the larger one to get both advantages of the two stops.

$$\text{Chandelier Stop} = Highest\ price\ in\ historical\ 15\ days - 3 * ATR(15 days)$$

$$\text{YoYo Stop} = last\ day\ closing\ price - 1.5 * ATR(15 days)$$

$$\text{DownStop} = Max(Chandelier\ Stop, YoYo\ Stop)$$

The current price or investment price is taken as the cost price and the corresponding returns are calculated based on the two stops shown above. After that, the risk index is generated using the returns. The risk index indicates the risk that we should take with the expected return. The formula for calculating the risk index is shown below:

$$\text{Upper Stop Return } (USR) = \frac{Upper\ Stop}{Current\ price} * 100\%$$

$$\text{Down Stop Return}(DSR) = \frac{Down\ Stop}{Current\ price} * 100\%$$

$$\text{Expected Return } (ER) = \frac{Accuracy * Upper\ Stop + (1 - Accuracy) * Down\ Stop}{Current\ price} * 100\%$$

$$\text{Risk Index } (RI) = \sqrt{Accuracy * (USR - ER)^2 + (1 - Accuracy) * (DSR - ER)^2}$$

Accuracy is defined as the accuracy of the machine learning model in predicting the test set.

2.4.1.2 Selecting Appropriate Stocks to Recommend

After the calculation of risk index, the system applies screening to remove the stocks that are unlikely to generate profits.

There are 3 cases where the stock will not be recommended to users:

1. When the current down stop price is higher than the current price as shown in figure 8:

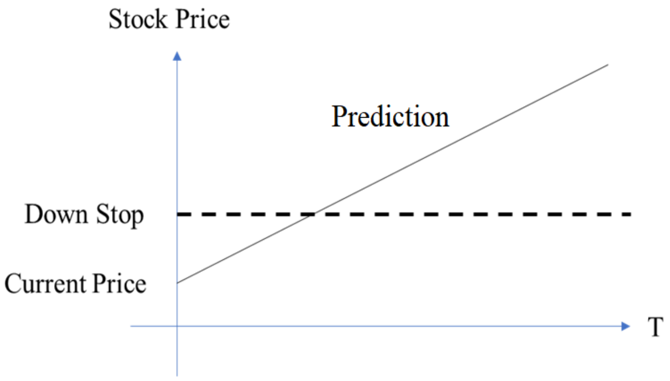


Figure 8: Prediction price through down stop

This case indicates that the current price is out of the fluctuation range of historical prices. Hence, it is assumed to drop further in the upcoming days, so it is not recommended to buy.

2. When the prediction trend is a bearish trend as shown in figure 9:

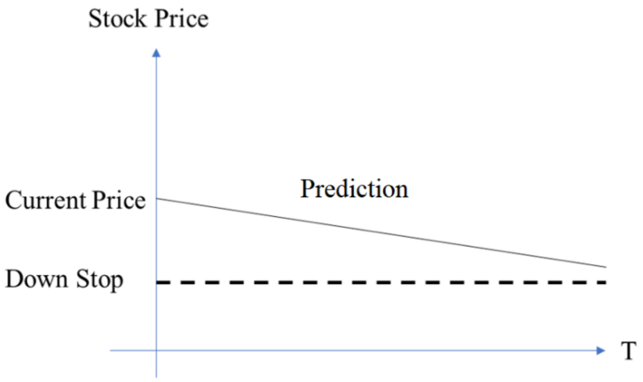


Figure 9: Bearish trend

3. When the prediction is V shape which the lowest prediction price is lower than the down stop price as shown in figure 10:

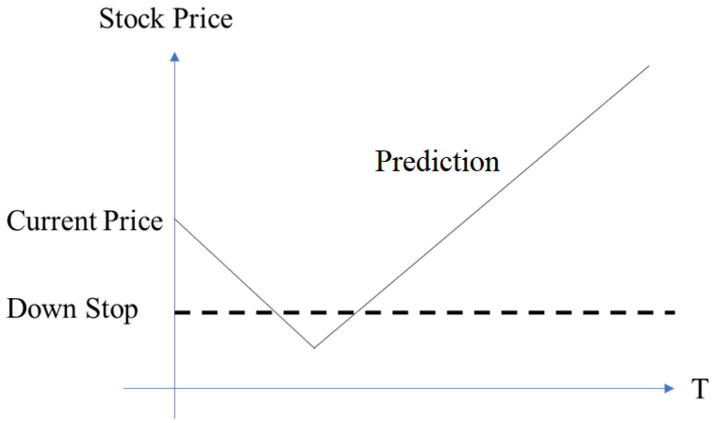


Figure 10: V shape through down stop

There are 2 cases where the stock will be recommended to users:

1. When the prediction is V shape which the lowest prediction price is higher than the down stop price as shown in figure 11:

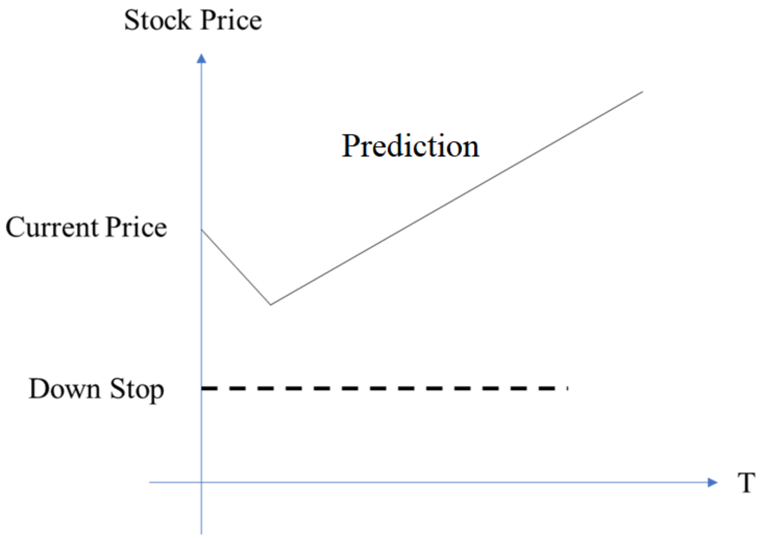


Figure 11: V shape above down stop

2. When the prediction trend is a bullish trend as shown in figure 12:

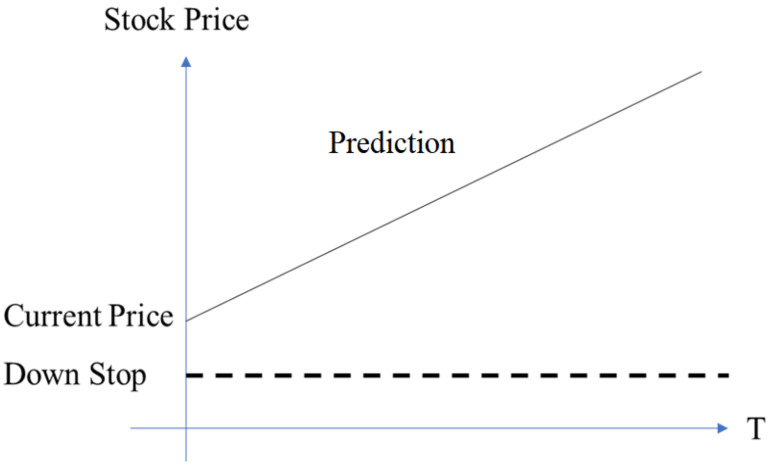


Figure 12: Bullish trend

2.4.2 Implementation

Python is used to implement the upper stop and down stop. The down stop is calculated based on the historical stock price. Therefore, it is implemented along with the generation of technical indicators after the stock price in the previous trading day are crawled. The upper stop is calculated based on the newly predicted upcoming stock price. Therefore, it is implemented after the generation of today's prediction prices. The risk index is calculated after both of the stop points and price predictions are updated to the database.

2.5 Portfolio Recommendation

2.5.1 Design

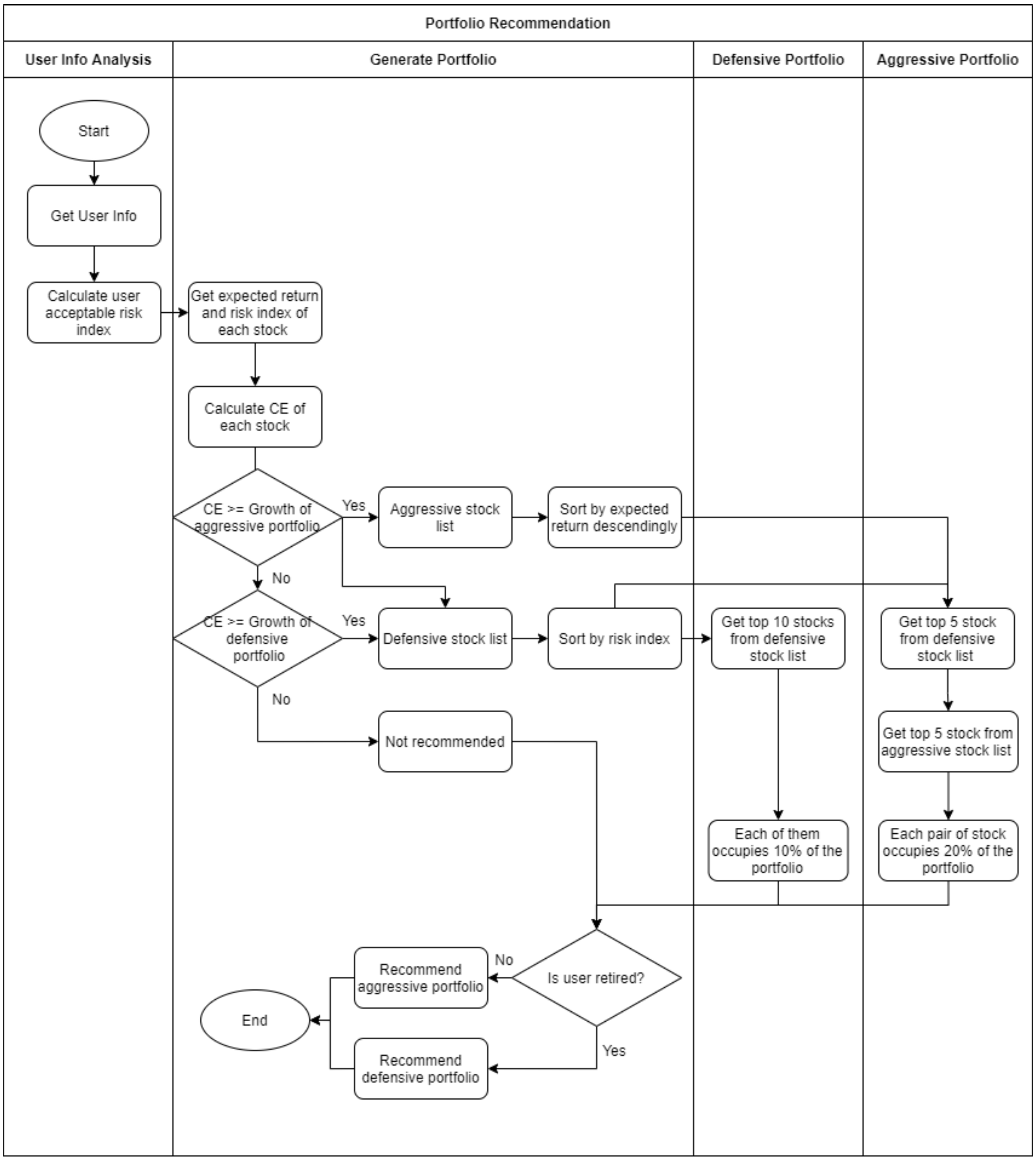

Portfolio Recommendation
User Info Analysis
Generate Portfolio
Defensive Portfolio
Aggressive Portfolio
Start
Get User Info
Calculate user acceptable risk index
Get expected return and risk index of each stock
Calculate CE of each stock
CE >= Growth of aggressive portfolio
Yes
Aggressive stock list
Sort by expected return descendingly
No
CE >= Growth of defensive portfolio
Yes
Defensive stock list
Sort by risk index
Get top 10 stocks from defensive stock list
Get top 5 stock from defensive stock list
Get top 5 stock from aggressive stock list
No
Not recommended
Each of them occupies 10% of the portfolio
Each pair of stock occupies 20% of the portfolio
Recommend aggressive portfolio
No
Is user retired?
End
Recommend defensive portfolio
Yes

Figure 13: Portfolio Recommendation Overview

Figure 13 gives a comprehensive overview of the portfolio recommendation system.

2.5.1.1 User Acceptable Risk Index

The user acceptable risk index is calculated using the financial information provided by the user. Our system aims to help user achieving "Financial Independence, Retire Early" (FIRE). Therefore, assets and expenses are collected from the advising system when users input their personal information as shown in the following table:

Raw Data (Dollar unit: USD)
Monthly Income (Monthly)
Living Expense (Monthly)
Housing Expense (Monthly)
Miscel Expense (Monthly)
Tax Expense (Yearly)
Equity
Cash Flow

Table 8: Data Collected in the Advising System

Current Asset amd Monthly Expense can be calculated by the following formulas:

$$Current\ Asset\ =\ Equity\ +\ Cash\ Flow$$

$$\begin{aligned} &Monthly\ Expense \\ &\quad = Living\ Expense\ +\ Housing\ Expense\ +\ Miscel\ Expense \\ &\quad + Tax\ Expense/12 \end{aligned}$$

The user acceptable risk level is calculated by the following formula:

$$Acceptable\ Risk\ Index\ (A)\ =\ 1-\left(\frac{Current\ Asset}{(Monthly\ Expense\ \times\ 12)/4\%}\right)^2$$

When the user has a high acceptable risk index, it indicates that the user can tolerate more risk. The 4% in the formula comes from the popular FIRE analogy that one can withdraw up to 4% of the amount saved in the investment portfolio every year without running out of money by saving up at least 25 times the annual expenditure. The function is exponential as the user has to be way more aggressive when the asset is low so that the asset can grow quickly in order to achieve FIRE.

2.5.1.2 Stock Selection

The risk index is based entirely on historical data and forecast models. Considering the differences in acceptable risks for different users, certainty equivalent is an economic way to quantify people's attitudes or preferences over risk and return.

$$Certainty\ Equivalent\ (CE)\ = ER\ -\ 1/2\ \times\ (1-A)\ \times\ 4\ \times\ RI^2$$

The A in the formula is the acceptable risk which is calculated in user info analysis. With a higher CE, the stock is relatively profitable at the same level of risk.

$$\begin{aligned}&Expected\ Holding\ Period\ (T)\\&= Time\ interval\ between\ today\ and\ the\ day\ of\ the\ stock\ reaching\ the\ highest\ price\\&\quad in\ prediction, where\ 1 \leq T \leq 90\end{aligned}$$

$$Expected\ Daily\ Growth\ = (1 + Expected\ annual\ growth)^{\frac{1}{365}} - 1,$$

$$Expected\ Growth\ in\ T\ trading\ days\ = (1 +\ Expected\ Daily\ Growth)^{T} - 1$$

Average annual inflation rate in U.S. in 2010-2019	1.40% [19]
Expected annual growth for aggressive portfolio	13.60% [20] (Average growth of S&P500 in 2010-2019)
Yearly return needed for retired user	4.00%
Expected annual growth for defensive portfolio	4.00% + 1.40% = 5.40%
Expected daily growth for aggressive portfolio	0.03494%
Expected daily growth for defensive portfolio	0.01440%

Table 9: Information for Calculating CE

After the system calculates the CE of each stock, it classifies them either as an aggressive stock, a defensive stock or a not-recommended stock.

Aggressive Stock:

If $Expected\ Growth\ in\ T\ trading\ days \geq (1 + Expected\ daily\ growth\ for\ aggressive\ portfolio)^T - 1$, it is classified as an aggressive stock.

The system takes up to 10 stocks with the highest expected return into the aggressive stock pool.

Defensive Stock:

If $Expected\ Growth\ in\ T\ trading\ days \geq (1 + Expected\ daily\ growth\ for\ defensive\ portfolio)^T - 1$, it is classified as a defensive stock.

The system takes up to 10 stocks with the lowest risk index into the defensive stock pool.

To avoid higher-risk stocks being recommended in the defensive stock pool, the stocks with risk index $\geq$ 0.043 are excluded, where the value is the intersection of the lines in the following figure.

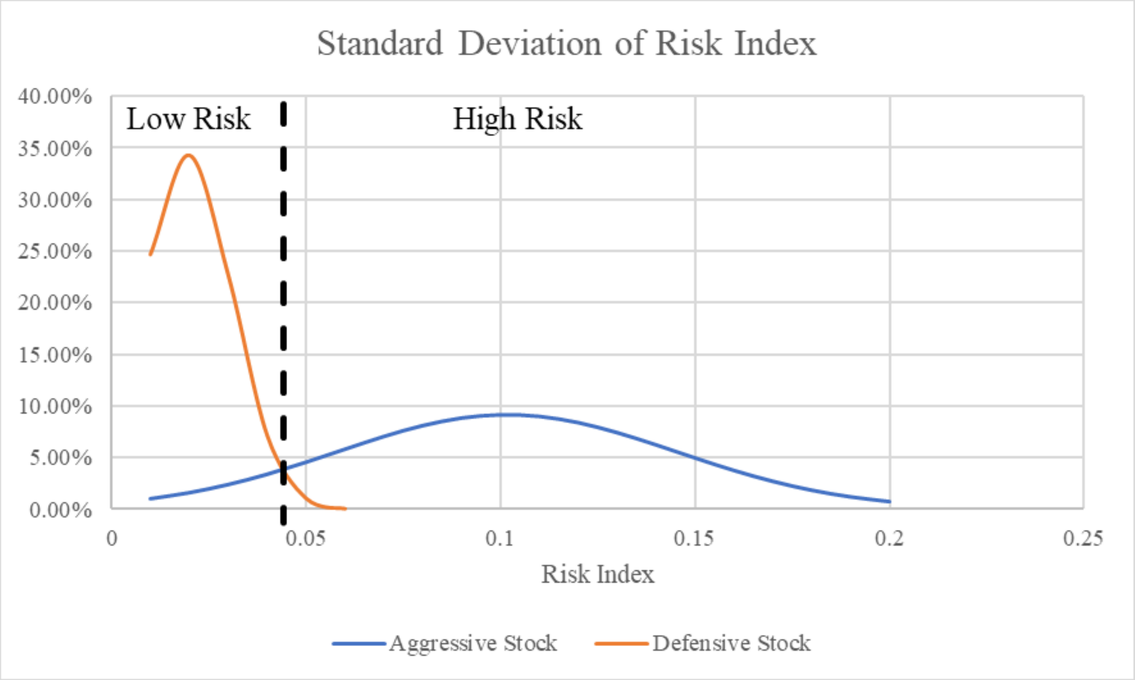
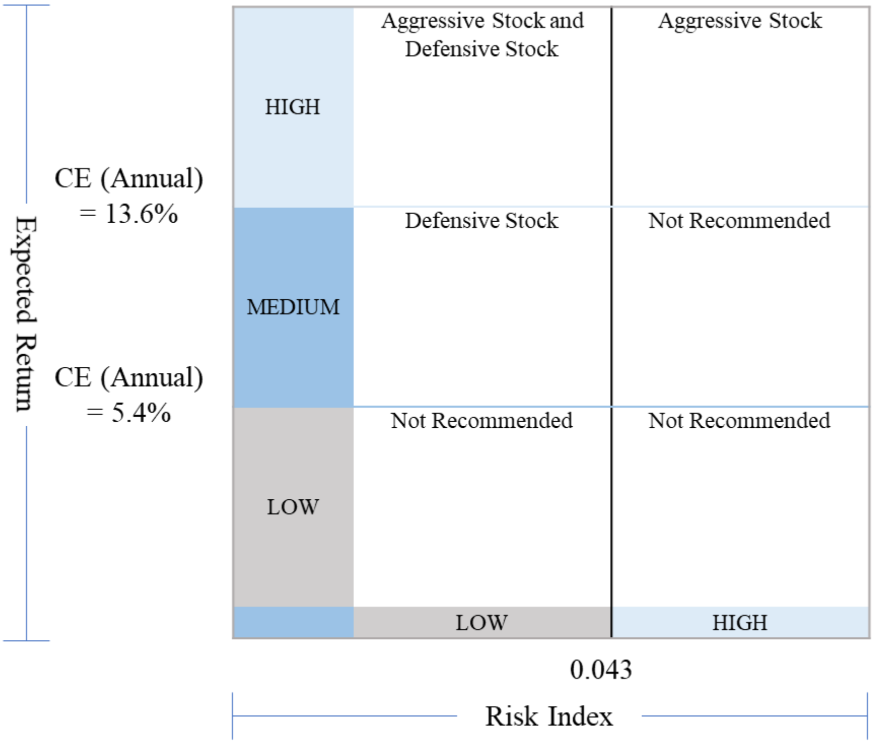


Figure 14: Standard Deviation of Risk Index

Stocks that are not classified as Aggressive stock or Defensive stock due to negative CE or a CE lower than our expected growth in the expected holding period (T, in trading days) are identified as **Not Recommended stock**. The following figure summarize the selection of stocks:



	LOW	HIGH
HIGH	Aggressive Stock and Defensive Stock	Aggressive Stock
MEDIUM	Defensive Stock	Not Recommended
LOW	Not Recommended	Not Recommended

Figure 15: Selection of Stocks

2.5.1.3 Portfolio Strategy

There are 2 types of portfolio the system would recommend to the users:

Aggressive Portfolio:

This portfolio is recommended to users that have not achieved FIRE yet. It consists of 5 aggressive stocks for maximizing profit and 5 defensive stocks for balancing the risk.

Each pair of aggressive stock and defensive stock occupies 20% of the portfolio.

The ratio of the aggressive stocks to defensive stocks depends on the $Acceptable\ Risk\ Index\ (A)$. If the user is retired, then A is equal to 0.

The ratio between Aggressive stock and defensive stock is $A : (1 - A)$.

Defensive Portfolio:

This portfolio is recommended to users that have achieved FIRE. It consists of 10 defensive stocks for maintaining the FIRE requirement at minimal amount of risks. Each stock occupies 10% of the portfolio.

2.5.2 Implementation

The implementation of the portfolio recommendation is done in the backend server using Java. The function is triggered when users access the Portfolio Management Page in the web application.

2.6 Web Application

2.6.1 Design

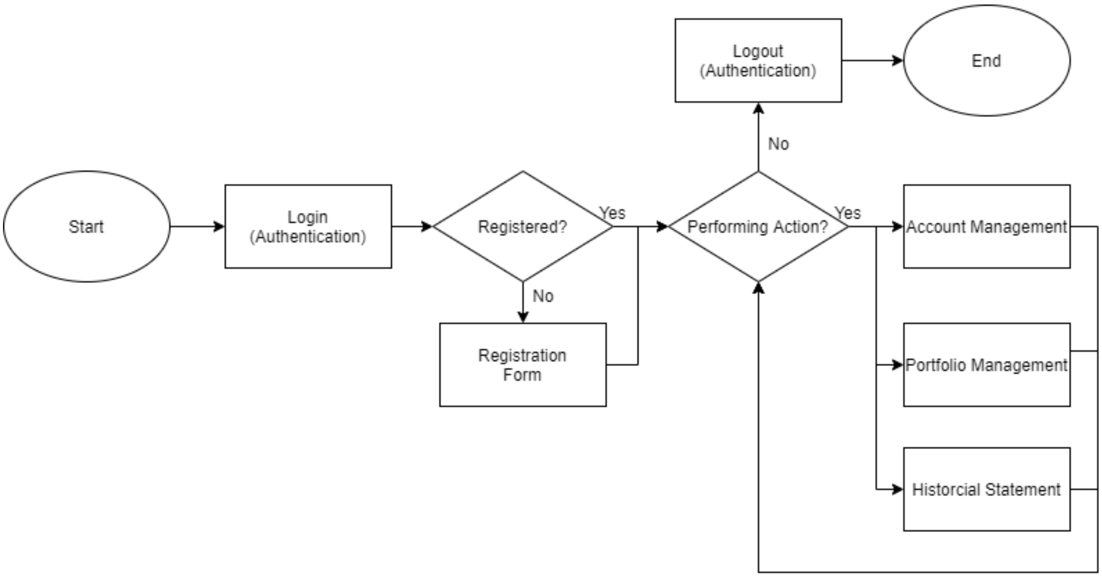


Figure 16: Web Application Design

Figure 16 provides an overview to the web application design.

There are 5 main functions in this system.

- Authentication
- Registration Form
- Account Management
- Portfolio Management
- Historical Statement

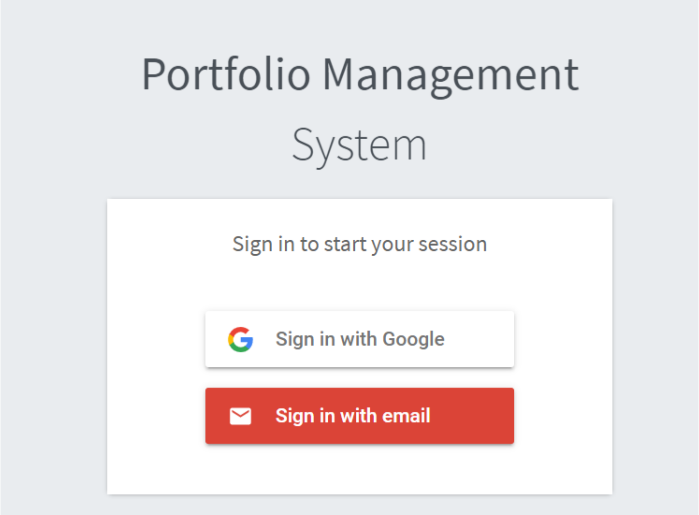
2.6.1.1 Authentication

A financial system collects lots of confidential information from users. To enforce security, we must make sure no one except the user can know the plain password, not even for the system administrator. Therefore, we have to salt and hash the password in the database. If a user forgets the password, a password reset action is needed.

2.6.1.1.1 Login

We provide 2 ways of authentication as shown in Figure 17:

1. Google account
2. Email account

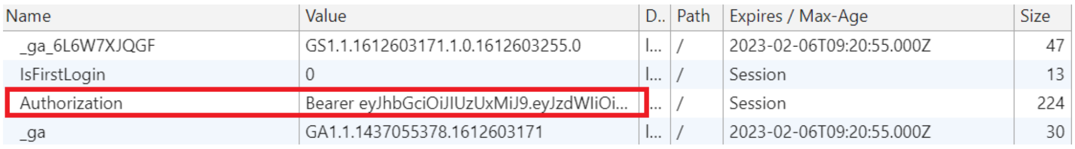


Portfolio Management System

Sign in to start your session

Figure 17: Login Page

Once a user logs in, a token will be returned to the browser and stored in a cookie as shown in Figure 18. Every action performed in the system requires a valid token. After a user has logged in for 30 minutes, the user has to renew the session. An account can only contain 1 token for web portal and 1 token for mobile application, if the user logs in from another computer or browser, the previous session is logged out and the token is expired.



Name	Value	D..	Path	Expires / Max-Age	Size
_ga_6L6W7XJQGF	GS1.1.1612603171.1.0.1612603255.0	l...	/	2023-02-06T09:20:55.000Z	47
IsFirstLogin	0	l...	/	Session	13
Authorization	Bearer eyJhbGciOiJIUzUxMiJ9.eyJzdWIiOi...	...	/	Session	224
_ga	GA1.1.1437055378.1612603171	l...	/	2023-02-06T09:20:55.000Z	30

Figure 18: Database View

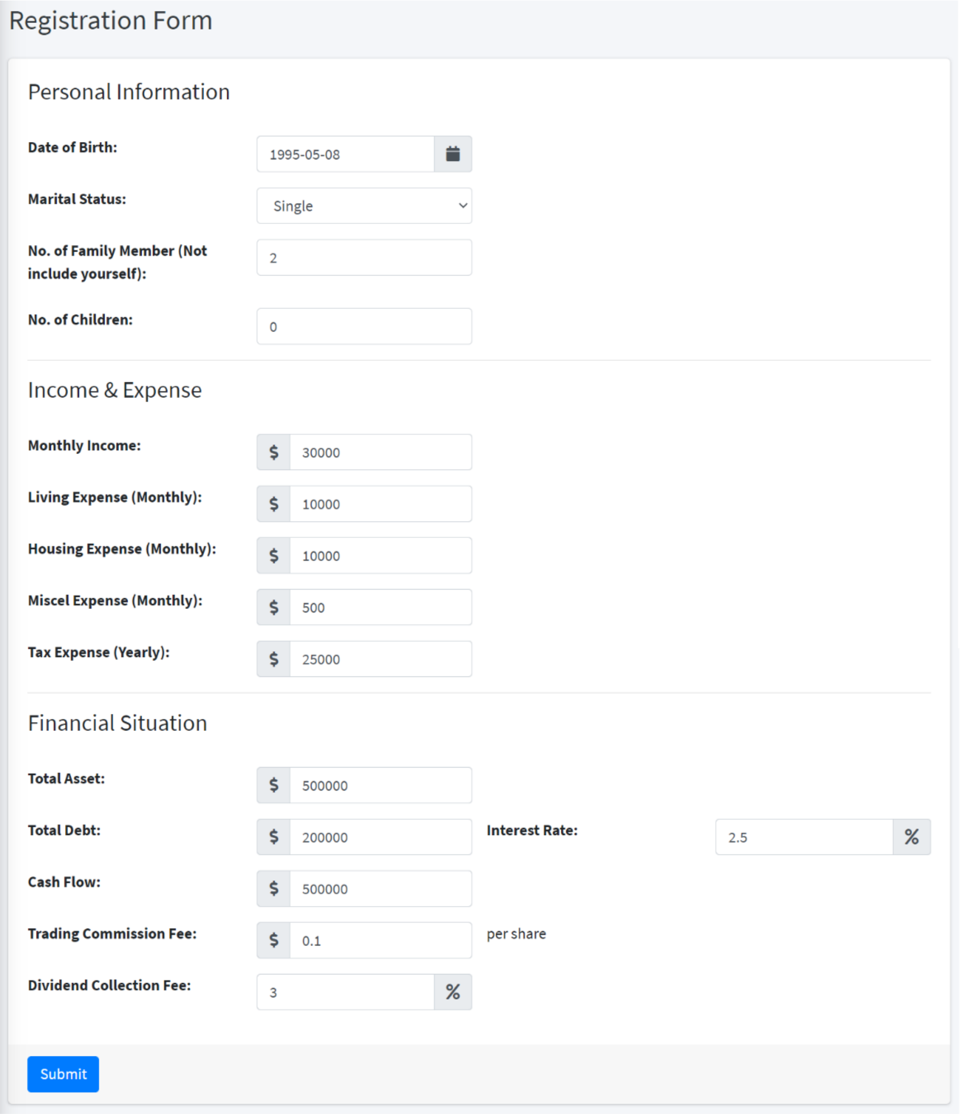
If the user logs in for the first time, the system will redirect the user to the registration page. Otherwise, the user will be redirected to the portfolio dashboard.

2.6.1.1.2 Logout

When the user clicks "logout" button, the token for accessing API will be cleared and the user will be redirected to the login page immediately.

2.6.1.2 Registration Form

The financial information about the user has to be collected before he or she can access the system. Otherwise, no recommendations could be made. The analysis of user info is handled in Portfolio Recommendation. Figure 19 shows the information collected.



Registration Form

Personal Information

Date of Birth: 1995-05-08

Marital Status: Single

No. of Family Member (Not include yourself): 2

No. of Children: 0

Income & Expense

Monthly Income: $ 30000

Living Expense (Monthly): $ 10000

Housing Expense (Monthly): $ 10000

Miscel Expense (Monthly): $ 500

Tax Expense (Yearly): $ 25000

Financial Situation

Total Asset: $ 500000

Total Debt: $ 200000 Interest Rate: 2.5 %

Cash Flow: $ 500000

Trading Commission Fee: $ 0.1 per share

Dividend Collection Fee: 3 %

Figure 19: Registration Form

2.6.1.3 Account Management

This function allows users to update their financial information. The system analyses their information to calculate the latest acceptable risk index. We highly recommend users update their information monthly or even weekly. This information will be integrated into the financial information analysis in order to provide a more accurate, acceptable risk index for the user.

2.6.1.4 Portfolio Management

This function aimed to provide a good user experience for users to manage their portfolio.

It consists of 3 features:

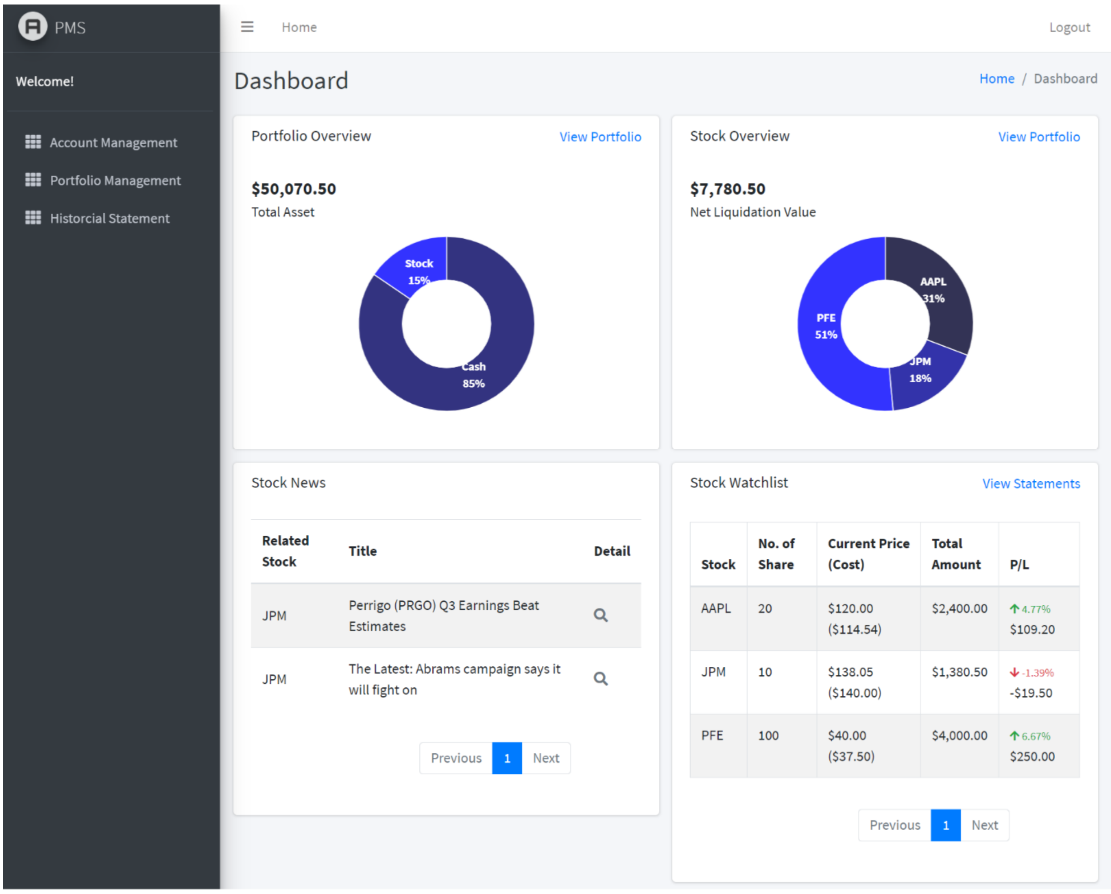
- Portfolio Dashboard
- Trade Reporting System
- Stock Recommendation

2.6.1.4.1 Portfolio Dashboard

Figure 20 shows the main page of the system. The system provides an at-a-glance view of the portfolio of the user.

It consists of 4 views:

- The total asset amount of different accounts and their distributions
- The net liquidation value of the stock account and its distribution
- The 7-day latest news of the stocks in the portfolio
- A watchlist showing the stocks in the portfolio



Portfolio Overview

$50,070.50
Total Asset

Stock Overview

$7,780.50
Net Liquidation Value

Stock News

Related Stock	Title	Detail
JPM	Perrigo (PRGO) Q3 Earnings Beat Estimates	🔍
JPM	The Latest: Abrams campaign says it will fight on	🔍

Stock Watchlist

Stock	No. of Share	Current Price (Cost)	Total Amount	P/L
AAPL	20	$120.00 ($114.54)	$2,400.00	↑ 4.77% $109.20
JPM	10	$138.05 ($140.00)	$1,380.50	↓ -1.39% -$19.50
PFE	100	$40.00 ($37.50)	$4,000.00	↑ 6.67% $250.00

Figure 20: Web Dashboard

2.6.1.4.2 Trade Reporting System

The following figure displays the workflow for the trade reporting system.

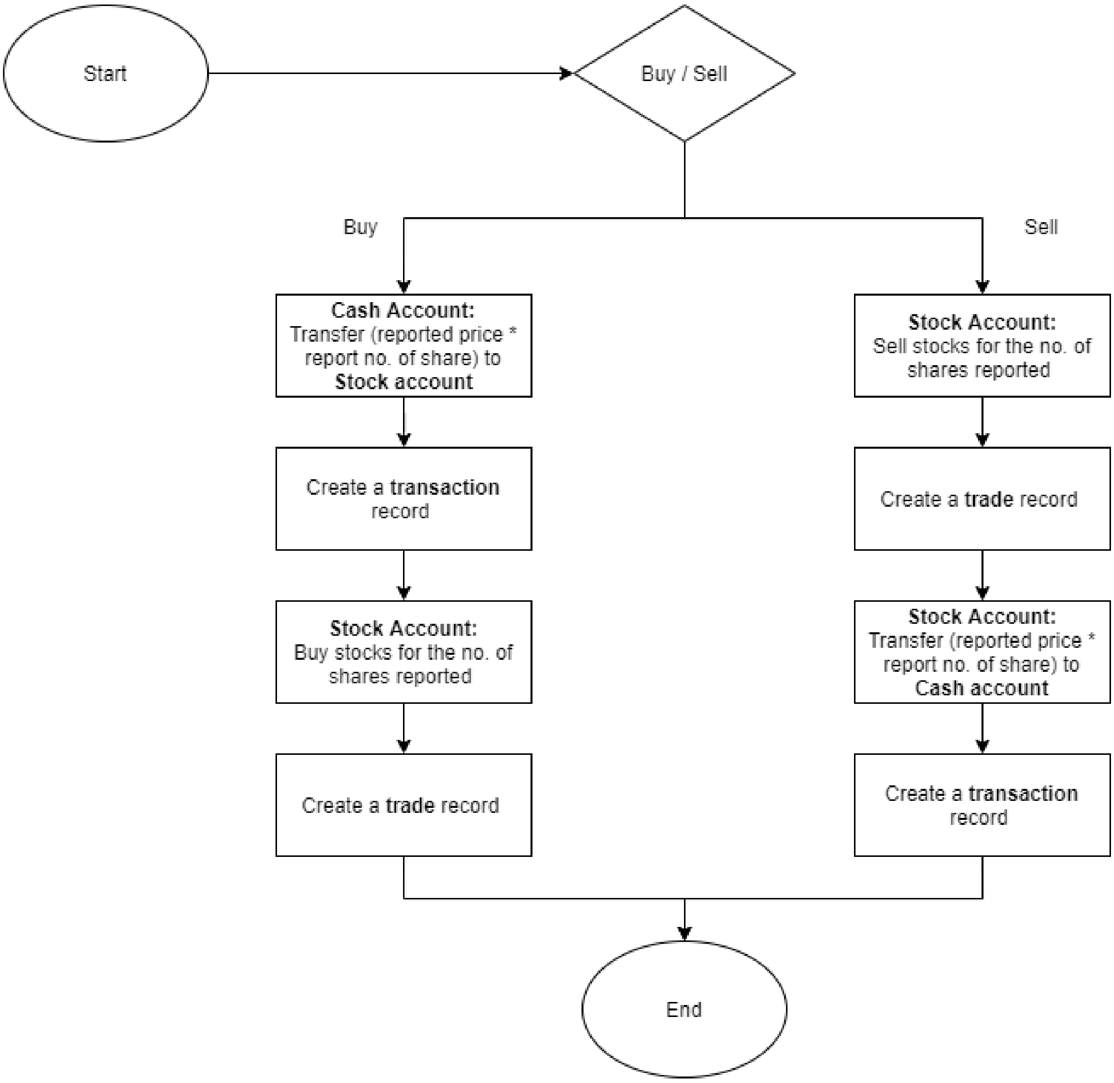
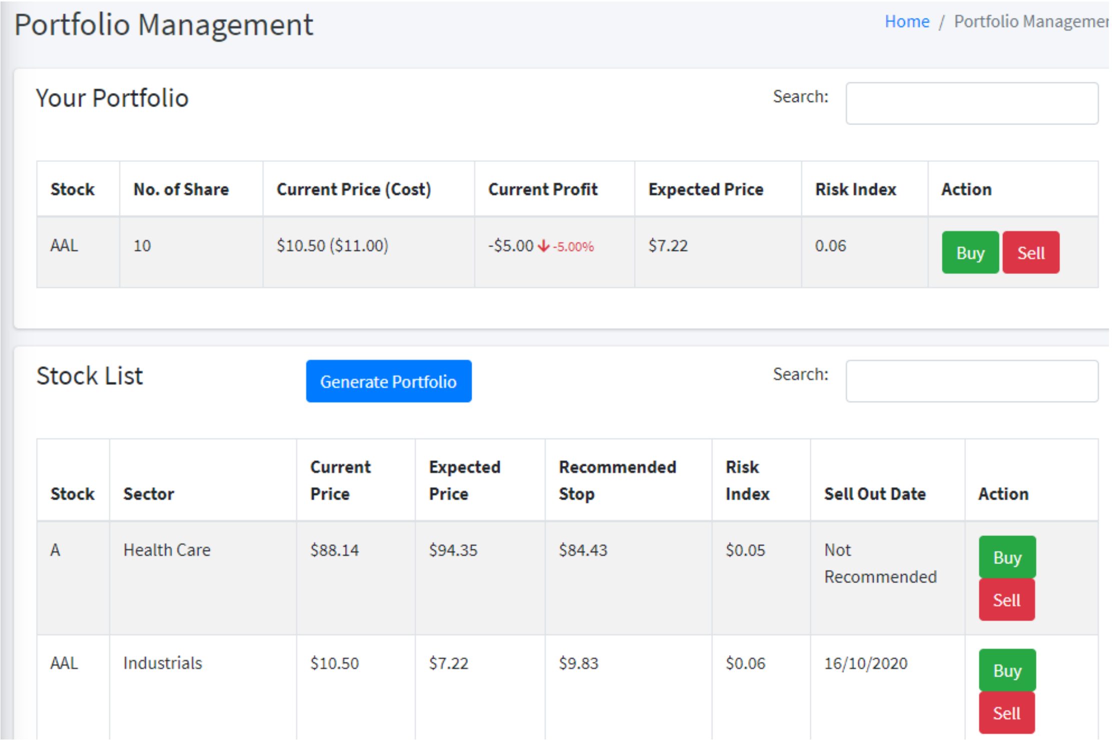

Start
Buy / Sell
Buy
Sell
Cash Account: Transfer (reported price * report no. of share) to Stock account
Create a transaction record
Stock Account: Buy stocks for the no. of shares reported
Create a trade record
Stock Account: Sell stocks for the no. of shares reported
Create a trade record
Stock Account: Transfer (reported price * report no. of share) to Cash account
Create a transaction record
End

Figure 21: Trade Reporting System Workflow

Our platform could not connect to the brokers directly due to security concerns. Therefore, we require users to input their trades on this page. User can view their stocks in trade in the upper section of the page and search for other stock in the "Recommended Portfolio" section as shown in Figure 22.



Portfolio Management

Your Portfolio

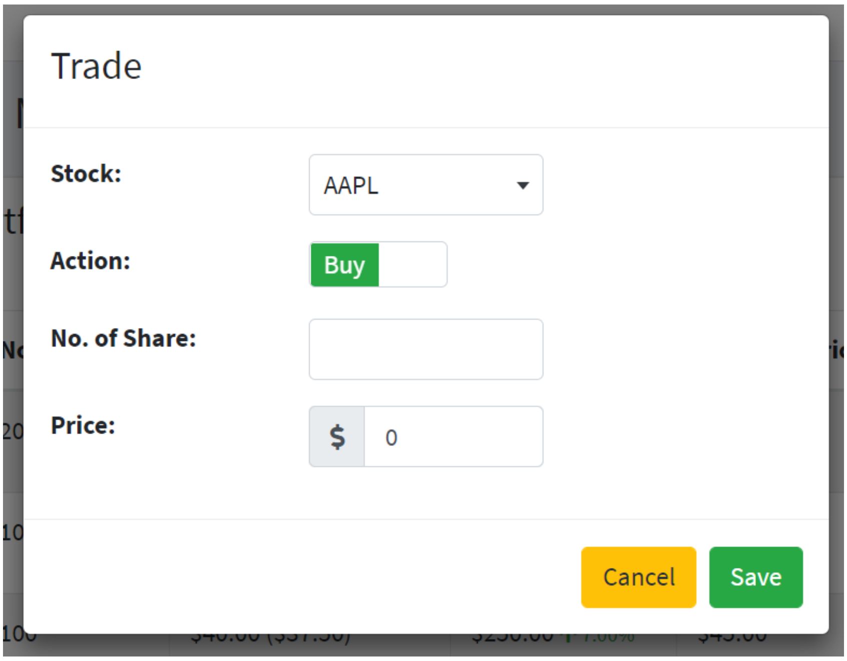
Stock	No. of Share	Current Price (Cost)	Current Profit	Expected Price	Risk Index	Action
AAL	10	$10.50 ($11.00)	-$5.00 ↓ -5.00%	$7.22	0.06	Buy Sell

Stock List

Stock	Sector	Current Price	Expected Price	Recommended Stop	Risk Index	Sell Out Date	Action
A	Health Care	$88.14	$94.35	$84.43	$0.05	Not Recommended	Buy Sell
AAL	Industrials	$10.50	$7.22	$9.83	$0.06	16/10/2020	Buy Sell

Figure 22: Portfolio Management Page

Once they click any of the action buttons, they can fill in the form as shown in the figure below to report a trade with the number of shares they brought or sold and the price at that moment.



Trade

Stock:	AAPL
Action:	Buy
No. of Share:	
Price:	$ 0

Figure 23: Trading Form

2.6.1.4.3 Stock Recommendation

The following figure displays the workflow for stock recommendation.

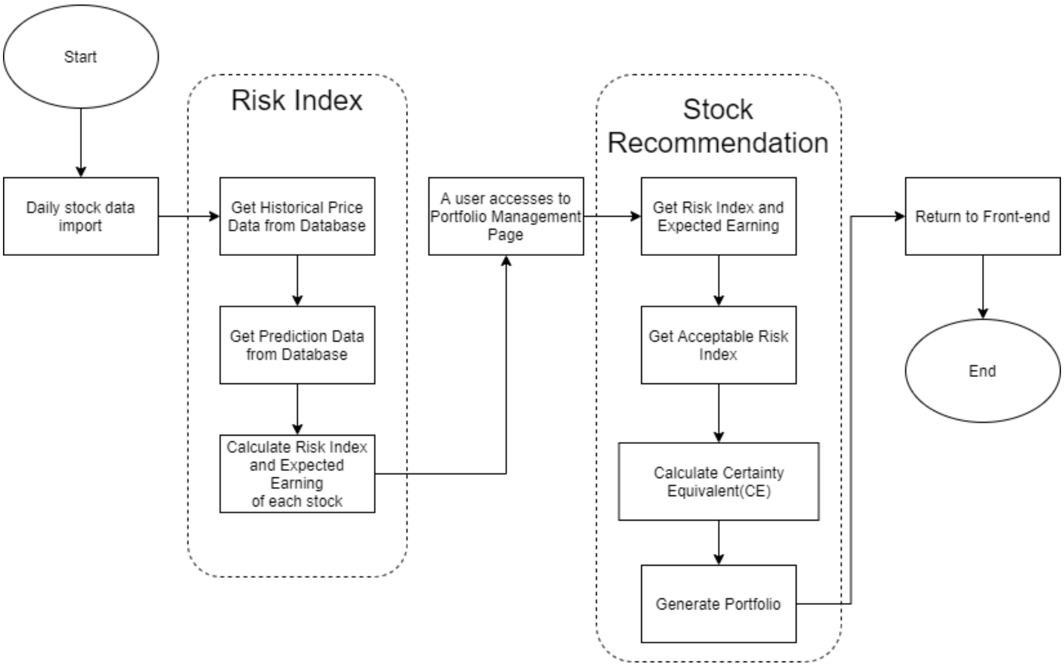
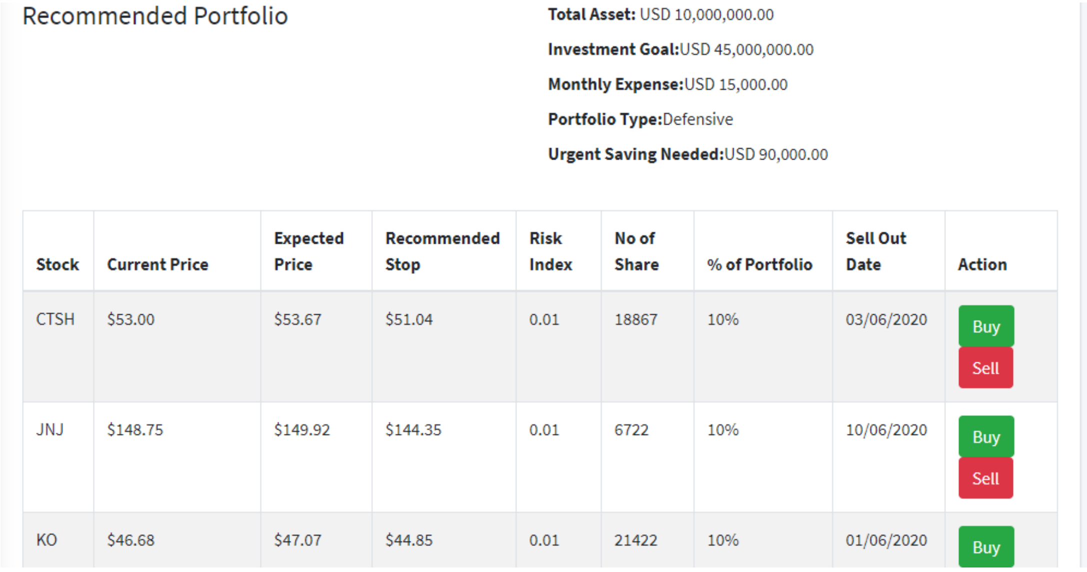

Start
Daily stock data import
Risk Index
Get Historical Price Data from Database
Get Prediction Data from Database
Calculate Risk Index and Expected Earning of each stock
A user accesses to Portfolio Management Page
Stock Recommendation
Get Risk Index and Expected Earning
Get Acceptable Risk Index
Calculate Certainty Equivalent(CE)
Generate Portfolio
Return to Front-end
End

Figure 24: Stock Recommendation Workflow

Users can view the recommendations in the Portfolio Management Page as shown in Figure 25. The recommendations are based on 2 parts of the system: Risk Index and Recommendation.



Recommended Portfolio

Total Asset: USD 10,000,000.00

Investment Goal:USD 45,000,000.00

Monthly Expense:USD 15,000.00

Portfolio Type: Defensive

Urgent Saving Needed:USD 90,000.00

Stock	Current Price	Expected Price	Recommended Stop	Risk Index	No of Share	% of Portfolio	Sell Out Date	Action
CTSH	$53.00	$53.67	$51.04	0.01	18867	10%	03/06/2020	Buy Sell
JNJ	$148.75	$149.92	$144.35	0.01	6722	10%	10/06/2020	Buy Sell
KO	$46.68	$47.07	$44.85	0.01	21422	10%	01/06/2020	Buy

Figure 25: Portfolio Recommendation

2.6.1.5 Historical Statement

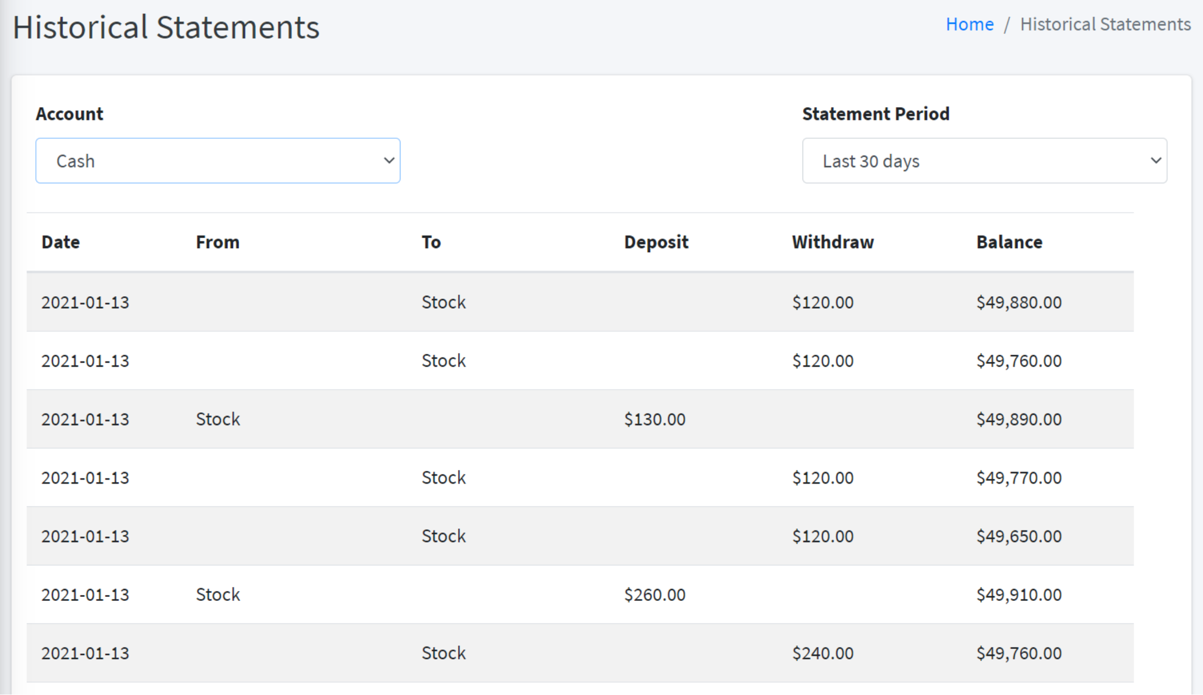
This function allows users to review their capital flows between different accounts and the profit or loss of the trading.

There are 2 kinds of statement:

- Account Statement
- Stock Statement

2.6.1.5.1 Account Statement

This statement as shown in Figure 26 allows users to review their capital flows between different accounts from 30 days to 180 days. Stock and Cash accounts are supported.



Account: Cash

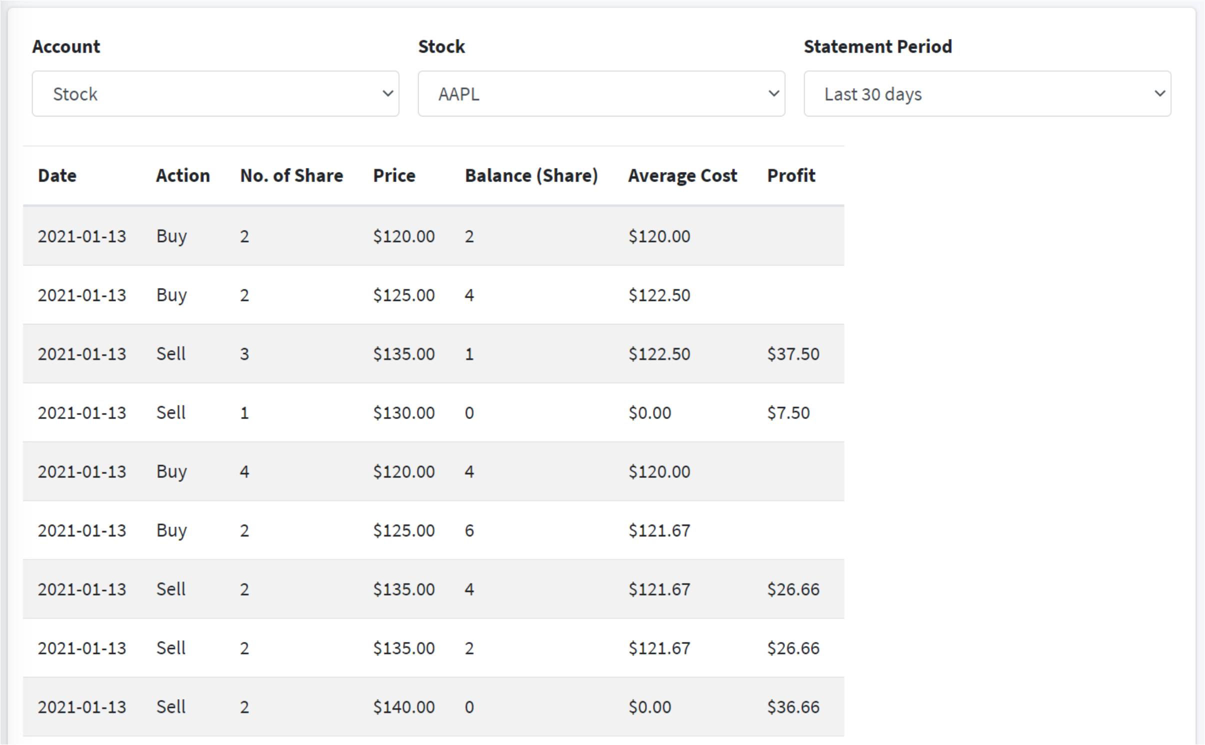
Statement Period: Last 30 days

Date	From	To	Deposit	Withdraw	Balance
2021-01-13		Stock		$120.00	$49,880.00
2021-01-13		Stock		$120.00	$49,760.00
2021-01-13	Stock		$130.00		$49,890.00
2021-01-13		Stock		$120.00	$49,770.00
2021-01-13		Stock		$120.00	$49,650.00
2021-01-13	Stock		$260.00		$49,910.00
2021-01-13		Stock		$240.00	$49,760.00

Figure 26: Account Statement

2.6.1.5.2 Stock Statement

This statement as shown in Figure 27 allows users to review their profit or loss of the trading for each stock from 30 days to 180 days.



Account: Stock

Stock: AAPL

Statement Period: Last 30 days

Date	Action	No. of Share	Price	Balance (Share)	Average Cost	Profit
2021-01-13	Buy	2	$120.00	2	$120.00	
2021-01-13	Buy	2	$125.00	4	$122.50	
2021-01-13	Sell	3	$135.00	1	$122.50	$37.50
2021-01-13	Sell	1	$130.00	0	$0.00	$7.50
2021-01-13	Buy	4	$120.00	4	$120.00	
2021-01-13	Buy	2	$125.00	6	$121.67	
2021-01-13	Sell	2	$135.00	4	$121.67	$26.66
2021-01-13	Sell	2	$135.00	2	$121.67	$26.66
2021-01-13	Sell	2	$140.00	0	$0.00	$36.66

Figure 27: Stock Statement

2.6.2 Implementation

Database Design:

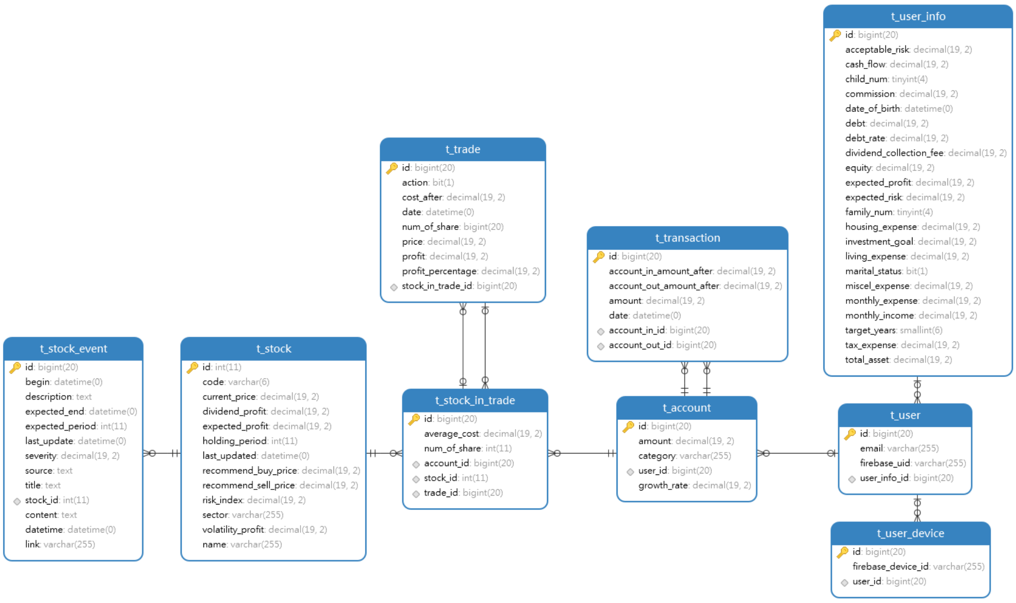

t_user_info
id: bigint(20)
acceptable_risk: decimal(19, 2)
cash_flow: decimal(19, 2)
child_num: tinyint(4)
commission: decimal(19, 2)
date_of_birth: datetime(0)
debt: decimal(19, 2)
debt_rate: decimal(19, 2)
dividend_collection_fee: decimal(19, 2)
equity: decimal(19, 2)
expected_profit: decimal(19, 2)
expected_risk: decimal(19, 2)
family_num: tinyint(4)
housing_expense: decimal(19, 2)
investment_goal: decimal(19, 2)
living_expense: decimal(19, 2)
marital_status: bit(1)
miscel_expense: decimal(19, 2)
monthly_expense: decimal(19, 2)
monthly_income: decimal(19, 2)
target_years: smallint(6)
tax_expense: decimal(19, 2)
total_asset: decimal(19, 2)
t_trade
id: bigint(20)
action: bit(1)
cost_after: decimal(19, 2)
date: datetime(0)
num_of_share: bigint(20)
price: decimal(19, 2)
profit: decimal(19, 2)
profit_percentage: decimal(19, 2)
stock_in_trade_id: bigint(20)
t_transaction
id: bigint(20)
account_in_amount_after: decimal(19, 2)
account_out_amount_after: decimal(19, 2)
amount: decimal(19, 2)
date: datetime(0)
account_in_id: bigint(20)
account_out_id: bigint(20)
t_stock_event
id: bigint(20)
begin: datetime(0)
description: text
expected_end: datetime(0)
expected_period: int(11)
last_update: datetime(0)
severity: decimal(19, 2)
source: text
title: text
stock_id: int(11)
content: text
datetime: datetime(0)
link: varchar(255)
t_stock
id: int(11)
code: varchar(6)
current_price: decimal(19, 2)
dividend_profit: decimal(19, 2)
expected_profit: decimal(19, 2)
holding_period: int(11)
last_updated: datetime(0)
recommend_buy_price: decimal(19, 2)
recommend_sell_price: decimal(19, 2)
risk_index: decimal(19, 2)
sector: varchar(255)
volatility_profit: decimal(19, 2)
name: varchar(255)
t_stock_in_trade
id: bigint(20)
average_cost: decimal(19, 2)
num_of_share: int(11)
account_id: bigint(20)
stock_id: int(11)
trade_id: bigint(20)
t_account
id: bigint(20)
amount: decimal(19, 2)
category: varchar(255)
user_id: bigint(20)
growth_rate: decimal(19, 2)
t_user
id: bigint(20)
email: varchar(255)
firebase_uid: varchar(255)
user_info_id: bigint(20)
t_user_device
id: bigint(20)
firebase_device_id: varchar(255)
user_id: bigint(20)

Figure 28: Database Design

The database design as shown in Figure 28 can support our functionalities in the web application.

2.6.2.1 Authentication

To enforce security, we used Firebase service for our authentication. Users register an account, reset password or login to an account through Firebase. We store the user id into our database for the first time and generate a JWT token further API calling every time when a user login.

2.6.2.2 Registration Form

Every time a user calls the APIs for inserting or updating the user information, we calculate the user acceptable index and store it into database.

2.6.2.3 Account Management

We have developed 3 APIs to support inserting, updating and getting user information. This enables users to update their information so that the system can generate a more accurate user acceptable index and recommend a more suitable portfolio.

2.6.2.4 Portfolio Management

We have developed an API to support the retrieving of the stock recommendation portfolio.

2.6.2.5 Historical Statement

We have developed 2 APIs to support getting account transactions and getting stock trade. Depending on the period, account and stock the user specified, we get the related transactions or trades and sort them by date in descending order.

2.7 Mobile Application

2.7.1 Design

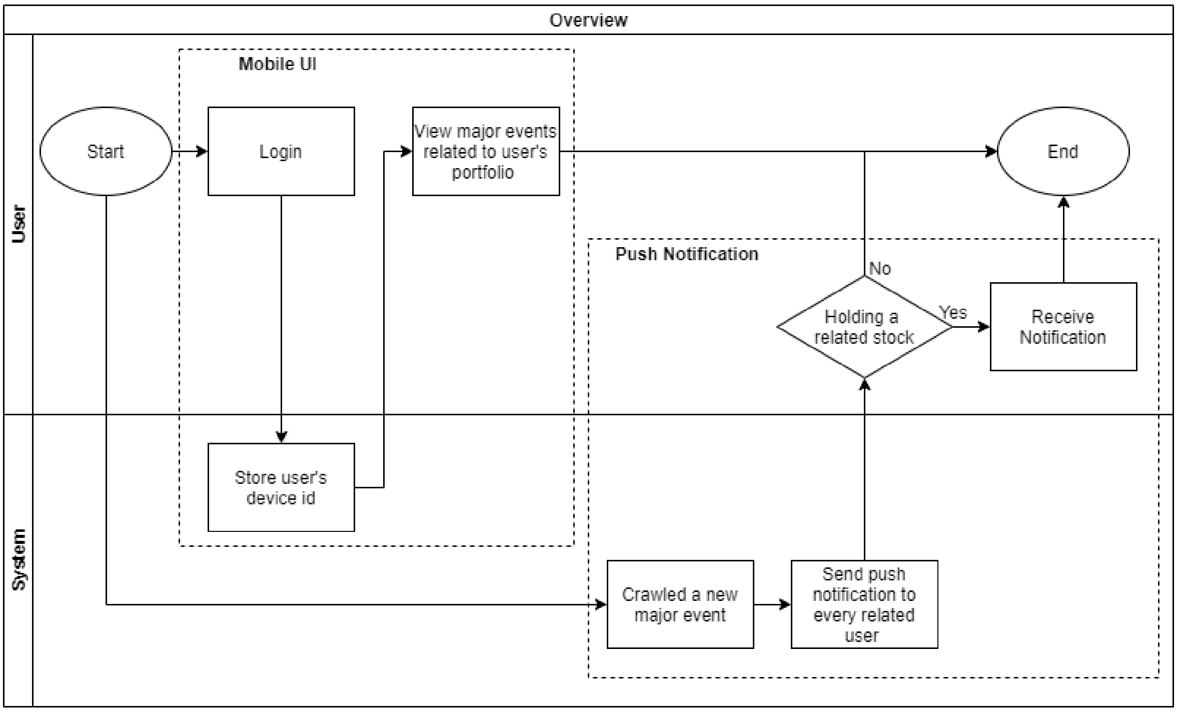


Figure 29: Overview of Mobile Application

Figure 29 displays the interaction between the user and system with the mobile application.

2.7.1.1 Authentication

Similar to the web application, the mobile application requires users to be logged in before accessing any of the functionalities.

2.7.1.1.1 Login

Mobile applications support two methods of signing in as shown in Figure 30:

1. Sign in with an Email
2. Sign in with Google

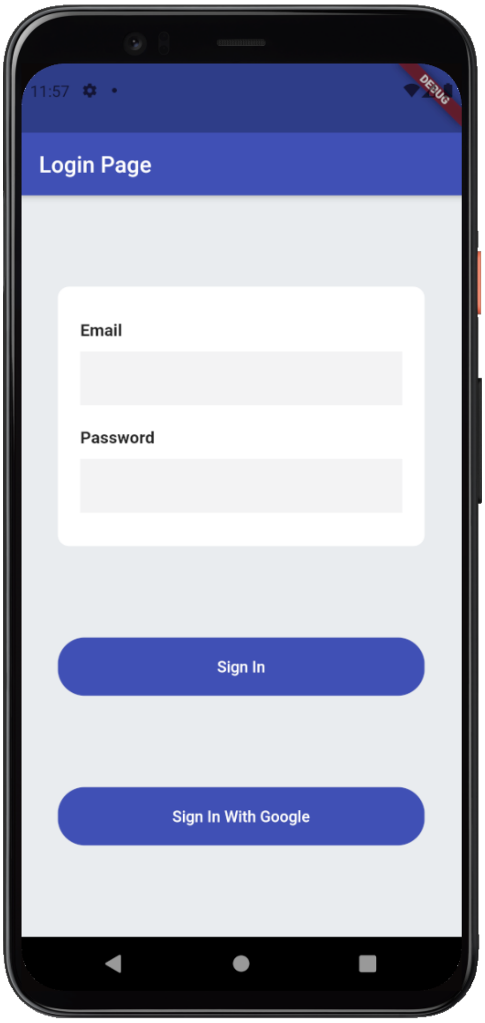


Figure 30: Mobile Login Page

The duration of the session is not limited since the main function of the mobile application is to receive notifications about the most-up-to-date news articles of the stocks that the users have purchased.

2.7.1.1.2 Logout

When the user clicks the "logout" button as shown in Figure 31, the token for accessing latest financial news will be cleared and the user will be redirected to the login page immediately. Users will no longer be able to receive news updates.

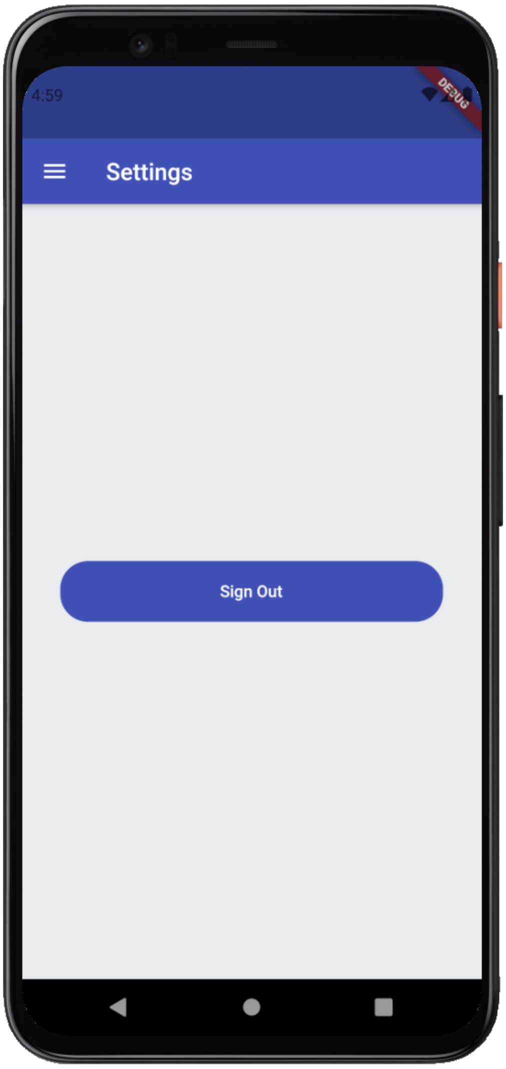
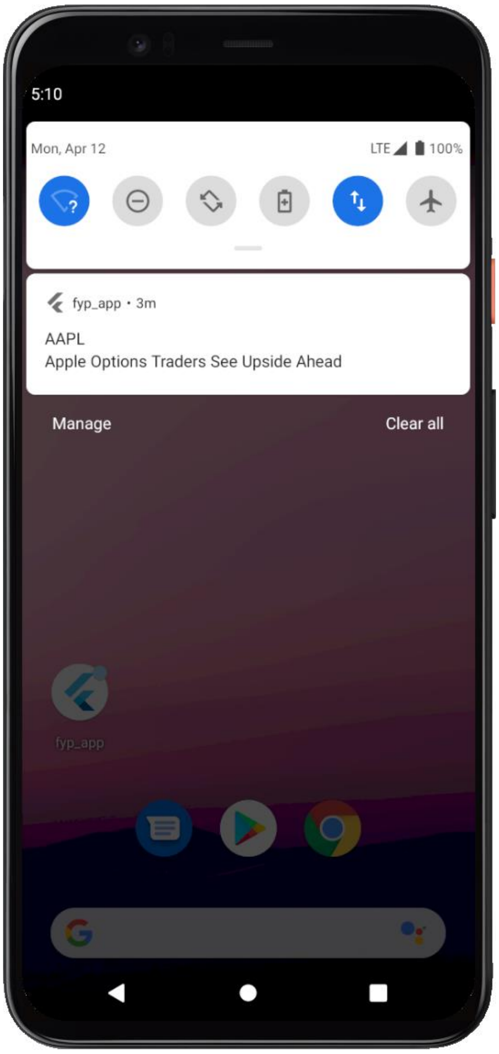


Figure 31: Mobile Logout Page

2.7.1.2 Push Notification

Once the server crawls a new article, it gets the device IDs of the related users from database and sends users a notification to the mobile application. Users can see a push notification with the related stock and the title of the news article as shown in Figure 32. Users can click on the notification to open the News page of the mobile application to view the details of the news.



5:10

Mon, Apr 12

LTE 100%

fyp_app • 3m

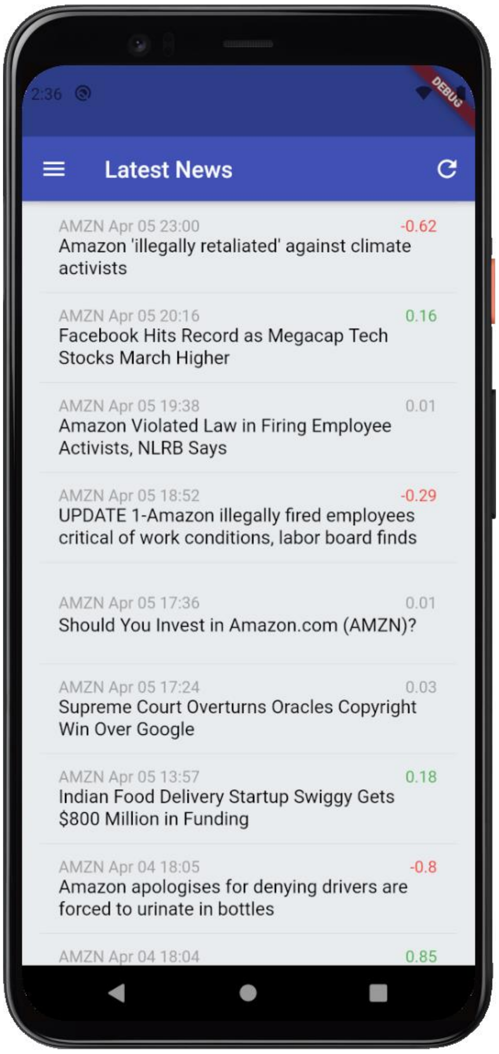
AAPL

Apple Options Traders See Upside Ahead

Figure 32: Mobile Notification

2.7.1.3 Displaying Latest News

The News page as shown in Figure 33 provides users to view the news articles of the purchased stocks published in the past 7 days. The severity is based on the sentiment score of the article's content. Positive news, neutral news, and negative news are displayed in green, grey and red respectively. The correlated stock and time of the article being published is displayed. Users can click on the tile to check out the news in the original website.



Latest News

AMZN Apr 05 23:00 -0.62
Amazon 'illegally retaliated' against climate activists

AMZN Apr 05 20:16 0.16
Facebook Hits Record as Megacap Tech Stocks March Higher

AMZN Apr 05 19:38 0.01
Amazon Violated Law in Firing Employee Activists, NLRB Says

AMZN Apr 05 18:52 -0.29
UPDATE 1-Amazon illegally fired employees critical of work conditions, labor board finds

AMZN Apr 05 17:36 0.01
Should You Invest in Amazon.com (AMZN)?

AMZN Apr 05 17:24 0.03
Supreme Court Overturns Oracles Copyright Win Over Google

AMZN Apr 05 13:57 0.18
Indian Food Delivery Startup Swiggy Gets $800 Million in Funding

AMZN Apr 04 18:05 -0.8
Amazon apologises for denying drivers are forced to urinate in bottles

AMZN Apr 04 18:04 0.85

Figure 33: Mobile News Page

2.7.2 Implementation

Flutter is used as the development toolkit for this application. It provides a framework for crafting the UI of the mobile application easily.

2.7.2.1 Authentication

We used Firebase for authentication. A Firebase ID Token is returned by Firebase after the user successfully sign in. Then, a POST request is sent to the backend API to pass the Firebase ID Token. The backend stores the device id to the database. A JWT token is returned to the user for authorization for using other APIs. *Flutter Secure Storage* is used to store the JWT token locally and securely. The Firebase Device ID for receiving notifications is then sent to the backend server via another API with the JWT token as the authorization using a POST request. The session ends if the account is logged in with a different device or the user pressed the logout button.

2.7.2.2 Push Notification

When the backend server crawls a new article, it retrieves the related user list from database and then passes the article title and device id to the Firebase Message Alert Queue. The queue proceeds with the entry by calling the Firebase service to send a push notification to the mobile device with this device id. *Firebase Messaging* and *Flutter Local Notification* are used to handle background notifications pop up and foreground notifications pop up respectively.

2.7.2.3 Displaying Latest News

Once the user directs to the News page, the mobile application sends a GET request with the JWT Token stored in *Flutter Secure Storage* to the backend to retrieve the last 7 days of news articles of the stocks that the user is currently holding. The sentiment polarity of each piece of news is colored accordingly to visualize sentiment. *URL Launcher* is used to redirect the website link to the default browser.

2.8 Testing

Unit testing was carried out to ensure all modules were built correctly in different phases. System integration testing was done after all the components were built and combined into the system. We tested the data collected, machine learning models, the user interface and the portfolio return.

2.8.1 Data Collection and Pre-processing

The following test cases are created to check the validity of the data.

Test case	Status
Goal: Check whether there are missing values in the stock price dataset Method: Loop through all dataset to check if there are missing price	Pass
Goal: Check whether there is empty news in the news dataset Method: Loop through all dataset to check if there are news article with empty link, title or content	Pass
Goal: Check whether there is duplicate news in the news dataset Method: Loop through all dataset to check if there are duplicate news	Pass
Goal: Check whether the news is related to the stock Method: Loop through all dataset to check if each piece of news has at least 5 mentions of the related stock	Pass

Table 10: Test cases for Data

2.8.2 Stock Price Prediction Models

To test the machine learning models, we have conducted black-box testing. We split our collected data into training set, validation set and test set. The test set represents the future of the training and validation set. The training set and validation set was used during the model training. The test set is used to benchmark the performance of the trained model.

The 90-days prediction of our current model achieved 80.95% accuracy on the test set of AAPL. The following graph visualized our predictions.

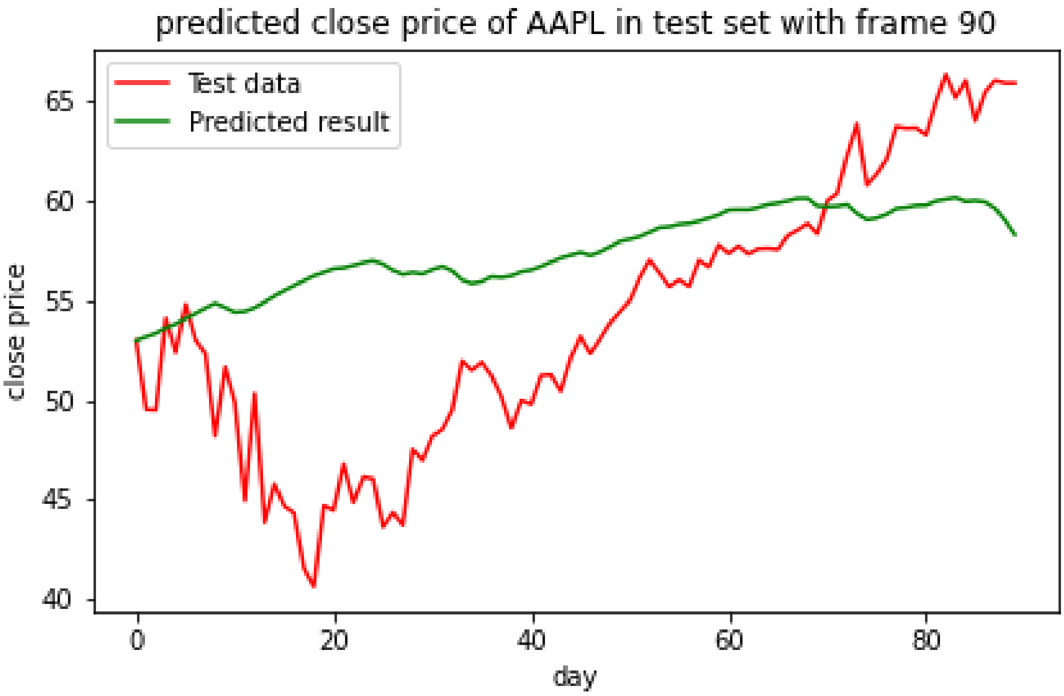


Figure 34: 90 Days Prediction of AAPL on test set

Overall, our machine learning models resulted in an average accuracy of 86.39% over 470 stocks in S&P500. This accuracy should be good enough for actual uses since we have also included the model accuracy in risk calculation.

2.8.3 Web Application

The following test cases are created to check the functionality of the web application.

Test case	Status
Goal: Check whether users can login successfully Method 1: Login with email and password Method 2: Login with Google account	Pass
Goal: Check whether the internal pages will be redirected to login page if the user is not logged in Method: Directly accessing to internal websites without logging in	Pass
Goal: Check whether the web UI displays correctly Method: Access the corresponding web pages	Pass
Goal: Check whether users can buy and sell stocks Method: Perform buy and sell actions in the portfolio management page	Pass

Table 11: Test Cases for Web Application

2.8.4 Mobile Application

The following test cases are created to check the functionality of the mobile application.

Test cases	Status
Goal: Check whether user can login successfully Method 1: Login with email and password Method 2: Login with Google account	Pass
Goal: Check whether the mobile application can retrieve latest news for last 7 days Method: Launch the news page of the mobile application	Pass
Goal: Check whether mobile notifications can be received Method: Backend sends a mobile notification	Pass

Table 12: Test Case for Mobile Application

2.8.5 Portfolio Management

Two types of portfolios are experimented to estimates their respective performance, namely the Defensive Portfolio and the Aggressive Portfolio. The Defensive Portfolio is generated based on a low user acceptable risk index whereas the Aggressive Portfolio is generated based on a high user acceptable risk index. Portfolio recommendation is carried out each month in this experiment. The stocks are sold when they reach the upper stop, down stop, or at the end of each month. Special Note is that since the model predicts up to 90 days ahead, the stocks purchased in November 2020 are sold in February 2021 if the stock price did not reach the upper stop or the down stop. The following figures display the monthly growth and cumulative growth of the respective portfolios:

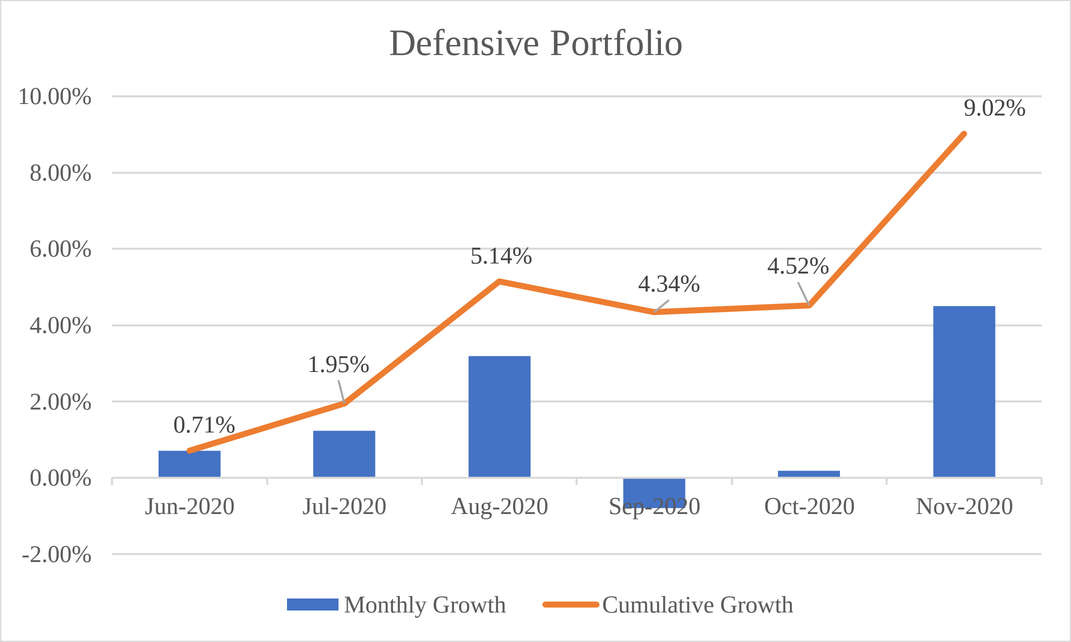


Figure 35: Performance of Defensive Portfolio

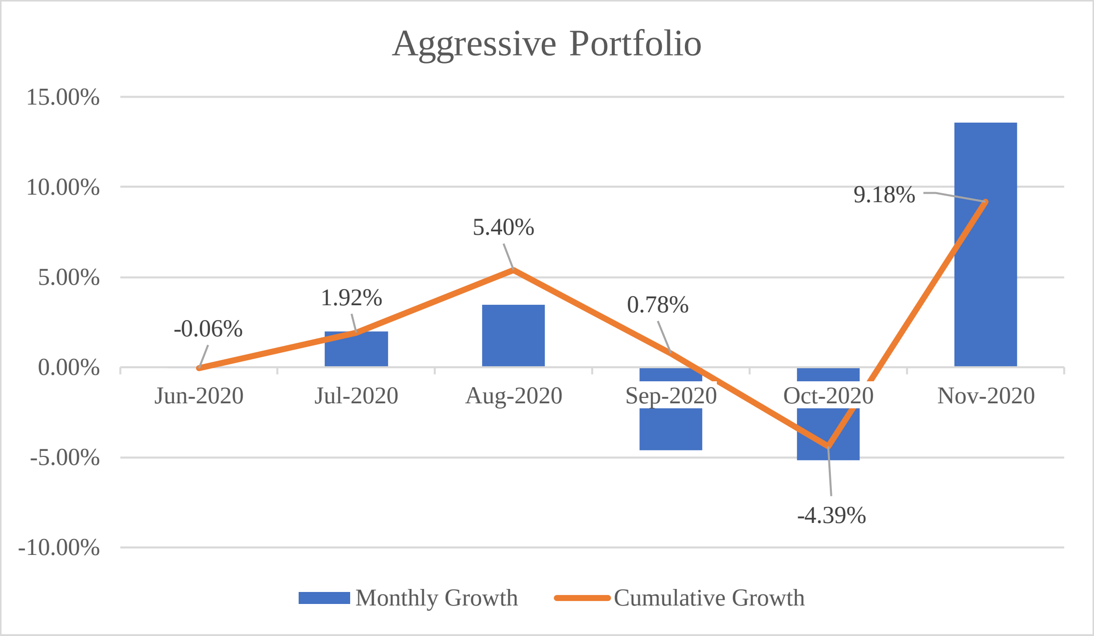


Figure 36: Performance of Aggressive Portfolio

Both portfolios returned profit in our test set during June 2020 to February 2021. The maximum monthly profit in the defensive portfolio and aggressive portfolio is 4.6% and 13.6% respectively. The maximum loss is 0.8% and 5.2% respectively. The cumulative profit attained is 9.02% and 9.18% respectively. Since our maximum prediction period is 90 days ahead, this indicates that our experiment has elapsed from June 2020 to February 2021, a total of 9 months. The annualized return for the portfolios is therefore 12.03% and 12.24% respectively.

2.9 Evaluation

We have completed all the objectives that we set for this project. The objectives include automatically and regularly downloading stock price and news data, creating a machine learning model that gives acceptable prediction accuracy, visualize the predictions for users to understand, give portfolio recommendations based on financial status of users and notify users about stock news and trading alerts. We evaluated how well we have completed for each of the objectives.

2.9.1 Data Collection and Pre-processing

Our data scrapers automatically and regularly collect price data after every trading day and news data every several hours. This ensures our system receives the latest data for model prediction as well as notifying users of important news updates. We are satisfied with the results as the stock price data collected is ordinary and news articles are successfully retrieved from Yahoo Finance. In general, this task is a success.

2.9.2 Stock Price Prediction Models

Our machine learning models achieved an average accuracy of 86.39% over 470 stocks in S&P500, which should be acceptable for calculating the risk index of the stocks. In general, this task is a success.

2.9.3 Visualization and Portfolio Management

AdminLTE is used as our website template and suitable tables, charts and graphs are used to visualize our predictions and portfolio recommendation. The web application functionalities are separated into different sections for easy understanding and operation for our users. The mobile application provides a simple UI for the users to interact with. In general, this task is a success.

A lower stop point is introduced so that the users can only lose up to a certain amount of money in each portfolio, preventing users to suffer from excessive loses due to huge price drop. Our portfolio testing results showed that the recommended portfolios are capable of generating good amount of profits. Risks and returns are balanced according to the users' acceptable risk index. Low-risk portfolios are suggested to users with low acceptable risk index and high-risk portfolios are suggested to users with high acceptable risk index. This can be verified by investigating the maximum loss and profits of the portfolios. The Defensive Portfolio has lower loss (-0.80%) and profits (4.50%) whereas the Aggressive Portfolio has higher loss (-5.17%) and profits (13.6%) in the testing, thus validating the style of the portfolios. In general, this task is a success.

3 Discussion

In this section, we will outline some of the processes that we have tried before finalizing our design and implementation for the system. We will also illustrate some of the challenges that we encountered and solutions to resolve the difficulties.

3.1 Data Collection

At the very beginning of our news scraper design, we tried to use yahoo finance directly to crawl our news data. However, during the implementation of our scraper, we found that the Yahoo finance website only generates news dynamically when we scroll the page to the bottom. This significantly increased the time needed for our news crawling. Luckily, we found out that Finviz has the latest stock news sorted with time and stored statically on their page. Hence, we decided to modify our data scraper to crawl news data using this website to increase the performance of our news scraper.

Although we have investigated some paid APIs that can be used and easily retrieve the data, we decided to build our crawlers from ground up since they did not really fit our needs entirely. We believed that some improvements can be made to increase both the quality and quantity of stock data if we can utilize these APIs in our project as they provide aggregated and sanitized results. We can also try to retrieve news not only from Yahoo Finance but also from more websites such as Bloomberg Finance, CNBC and Market Watch and aggregate the results to provide a comprehensive view on an event. This should provide a more accurate sentiment on the event.

3.2 Stock Price Prediction Models

Our initial design of the machine learning model was a multi-step model that predicts the upcoming 90 trading days (P_{t+1} to P_{t+90}) in a single shot. However, upon implementing the model, we found out that the trained model failed to generalize well even on the validation set. It was displaying a repeating pattern in different prediction periods and we were unable to fix this problem. Therefore, we instead tested and deployed a simpler single-step model that only predicts the upcoming 90th trading day. This model resulted in acceptable accuracy and was further finetuned for selecting the best hyperparameters. Although our model prediction results are quite promising, more sophisticated and advanced models can be used to further increase the accuracy of predictions.

3.3 Visualization and Portfolio Management

We have provided a user-friendly website for users to manage their portfolios and a simple mobile application for receiving news notification. This allows users to keep track of their profits and stock market activities easily. Both the web application and mobile application provides easy access of news articles for users to catch up with the latest events happened. This allows users to quickly react to huge stock market changes. More functionalities can be added such as allowing users to simulate their portfolio through the web application and manage the stock portfolio in the mobile application.

To consider the limitations of our machine learning model, we have involved risk management before making portfolio recommendations. We have used a risk index to indicate the risk the user requires to obey by quantifying the user's risk aversion which personalizes the portfolio recommendations for different users.

Risk management is an essential task to provide reliable portfolio recommendations. We have tried to utilize some basic financial concepts to formulate a risk management strategy. However, the strategy we adopted is far from perfect. We believe a more advanced solution can further decrease the loses on unexpected price drop of a stock.

We believe the performance of our testing portfolio can be better evaluated if the portfolio management is carried out daily. The upper stop, down stop and price prediction is generated every day, these values should be adjusted daily to allow profits to be further attained since the upper stop will not be executed so quickly when the stock is in a rising trend.

4 Conclusion

4.1 Summary of Current Work

In this project, we have developed a system that utilize stock price, technical indicators and news sentiments to predict the upcoming stock price and calculate the risk index of different stocks. The system then gives portfolio recommendations to investors based on their acceptable risk level and the corresponding risk index of different stocks. We then visualized the results through a user-friendly web application as well as allowing users to manage their portfolios. A mobile application is also created for notifying users of latest stock events.

Our testing shows that the portfolio generated is capable of generating 12% annualized returns which proved the feasibility of the recommendation system. We believe that long term investors and novice investors can use our system to better create a stock portfolio that offers suitable amount of risk and return based on their personal financial background to achieve FIRE.

4.2 Future Research Directions

4.2.1 Incorporating More Factors

Due to the complicated environment in stock markets, more factors can be considered to calculate the risk index, such as dividends, political issues, economic atmosphere, market sectors and the correlation between different stocks. More financial data can be retrieved and explored to identify deeper relationships. A better-evaluated risk index allows the system to generate more sophisticated portfolio recommendations to the user.

4.2.2 Applying Event Extraction

For retrieving useful information from news articles, we applied FinBERT which produces sentiment score indicating the polarity of a piece of news. Although the sentiment of financial text is correlated to the stock price, we believe that more sophisticated methods can be used to extract more information from news articles. If there is more time for this project, event extraction could be experimented to get a more accurate description of the event. The severity and implications of the event can be learnt by the machine learning model for more accurate prediction of stock price.

4.2.3 Towards Explainable Recommendations

We have yet to explore the possibilities for explainable decisions from the machine learning models and the system. If the system can carry out due diligence to identify and list out crucial news, important fundamental and technical analysis that supports the recommendations, users can have more faith towards the system as well as deciding whether that stock recommendation is suitable. In our project, users can only believe that our system will succeed based on our historical portfolios’ risk and return, which may not be ideal for actual use.

4.2.4 Providing High Concurrency and Low Latency System

We have already developed a system that manage the clients’ request, save users’ data and organize the crawled data. However, we have not considered the performance of our system when deployed for large scale use so it may not be optimized for such uses. To provide stable and high-quality services, distributed system can be used to solve this problem. For example, Read/Write Splitting for high-quality access and write on database, NoSQL techniques for quicker data access, Reverse Proxy for reducing the access pressure of server. Many current advanced techniques can be applied to develop a high concurrency and low latency system.

5 References

[1] A. Picasso, S. Merello, Y. Ma, L. Oneto and E. Cambria, "Technical Analysis and Sentiment Embeddings for Market Trend Prediction," *Expert Systems with Applications,* vol. 135, pp. 60-70, 2019.

[2] K. Joshi, B. H. N and J. Rao, "Stock Trend Prediction Using News Sentiment Analysis," *International Journal of Computer Science and Information Technology,* vol. 8, pp. 67-76, 2016.

[3] M. Alanyali, H. S. Moat and T. Preis, "Quantifying the Relationship Between Financial News and the Stock Market," *Scientific Reports,* vol. 3, p. 3578, 2013.

[4] T. Loughran and B. Mcdonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *The Journal of Finance,* vol. 66, pp. 35-65, 2011.

[5] A. Huang, H. Wang and Y. Yang, "The Informativeness of Text, the Deep Learning Approach," University of Hawai'i at Manoa, 2020.

[6] D. Jacob , C. Ming-Wei, L. Kenton and T. Kristina , "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *NAACL-HLT*, 2019.

[7] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia and D. Anastasiu, "Stock Price Prediction Using News Sentiment Analysis," *IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService),* pp. 205-208, 2019.

[8] J. Kagan, "Financial Health," Dotdash, 14 January 2020. [Online]. Available: https://www.investopedia.com/terms/f/financial-health.asp#:~:text=Financial%20health%20is%20a%20term,fixed%20or%20non%2Ddiscretionary%20expenses.. [Accessed 11 September 2020].

[9] F. K. Reilly and K. C. Brown, Investment Analysis and Portfolio Management, 10th Edition, Mason: Cengage Learning, 2012.

[10] R. C. Grinold and N. K. Ronald, Active portfolio management: a quantitative approach for providing superior returns and controlling risk, New York: McGraw-Hill, 2000.

[11] S. Kurutz, "How to Retire in Your 30s With $1 Million in the Bank," The New York Times, 1 September 2018. [Online]. Available: https://www.nytimes.com/2018/09/01/style/fire-financial-independence-retire-early.html. [Accessed 4 February 2021].

[12] A. Kerr, "Financial Independence, Retire Early (FIRE)," Investopedia, 14 January 2021. [Online]. Available: https://www.investopedia.com/terms/f/financial-independence-retire-early-fire.asp. [Accessed 4 February 2021].

[13] C. Hogan, "What Is the F.I.R.E. Movement?," Ramsey Solutions, 1 December 2020. [Online]. Available: https://www.daveramsey.com/blog/what-is-the-fire-movement. [Accessed 4 February 2021].

[14] R. Philps, "The Ultimate Guide to the FIRE Movement," NerdWallet UK, 20 January 2021. [Online]. Available: https://www.nerdwallet.com/uk/current-accounts/guide-to-the-fire-movement/. [Accessed 4 February 2021].

[15] K. Smith, "The Forbes Guide To FIRE," Forbes Advisor, 29 June 2020. [Online]. Available: https://www.forbes.com/advisor/retirement/the-forbes-guide-to-fire/. [Accessed 4 February 2021].

[16] "Total Market Value of U.S. Stock Market," Siblis Research Ltd, [Online]. Available: https://siblisresearch.com/data/us-stock-market-value/#:~:text=The%20total%20market%20capitalization%20of,about%20OTC%20markets%20from%20here.). [Accessed 19 August 2020].

[17] "MARKET INDICES," TradeView, 19 August 2020. [Online]. Available: https://www.tradingview.com/markets/indices/quotes-snp/. [Accessed 19 August 2020].

[18] "S&P 500 Index Chart, Components, Prices," Barchart, 19 August 2020. [Online]. Available: https://www.barchart.com/stocks/indices/sp/sp500?viewName=fundamental. [Accessed 19 August 2020].

[19] "Inflation, consumer prices (annual %) - United States," The World Bank, [Online]. Available: https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG?locations=US. [Accessed 14 April 2021].

[20] " S&P 500 (^GSPC) Historical Data," Yahoo Finance, [Online]. Available: https://finance.yahoo.com/quote/%5EGSPC/history/. [Accessed 14 April 2021].

[21] "RSI indicators," Gear of Trade, 23 October 2016. [Online]. Available: http://www.gearoftrade.com/index.php/technical-analysis/indicators/22712-rsi-indicators. [Accessed 13 September 2020].

[22] "Channel strategy based on ATR volatility indicator," Medium, 18 December 2018. [Online]. Available: https://medium.com/@FMZ_Quant/channel-

strategy-based-on-atr-volatility-indicator-b03aff84693f. [Accessed 13 September 2020].

6 Appendix A: Technical Analysis

Financial Indicators Used in This Project

Simple Moving Average (SMA):

It is to directly calculate the mean value of equal weight of time series, so it is easy to use. But it is most criticized for its hysteresis. As the computing window grows, the moving average becomes smoother, but also more and more backward.

Exponential Moving Average (EMA):

The weights of the values of the exponential moving average decrease exponentially. Due to its higher weight attached to recent data, the exponential moving average is more sensitive to recent changes than the simple moving average, and the exponential moving average also has hysteresis.

Moving Average Convergence / Divergence (MACD):

MACD is a trend index based on the construction principle of moving average after smoothing the closing price. MACD index is more of a reference for stock investors to buy and sell in the medium term but is not very effective for the short term.

Parabolic Stop And Reverse (PSAR):

PSAR is used to determine the trend direction and potential reversal in prices. It creates a series of dots to identify whether the stock price is rising or falling. Investors commonly use this technical indicator to identify suitable entry points, exit points and stop loss.

Relative Strength Index (RSI):

RSI index is a measure of the internal relative strength of securities. The general principles of the investment that investors' trading behavior reflects comprehensive results of various factors, the change of the market ultimately depends on supply and demand, and RSI indicator is based on the principle of the balance of supply and demand, by measuring one rose during the total accounted for the total average amplitude of the stock price changes, to assess the strength of the long-short power degree, which prompts specific operation [21].

Bollinger Bands (BB):

Bollinger bands use moving averages and standard deviations to estimate the value band. Since prices move up and down around value, a breakout above that band is overbought and a breakout below that band is oversold to gauge where prices are relative to value.

Stochastic Oscillator KDJ:

The KDJ measure measures the relative position of current stock prices over the most recent ups and downs. The higher the relative position of the current stock price, the higher the value of the three KDJ indicators will be, which is a signal that the stock price is in a strong rising trend. The lower the relative position of the current stock price, the lower the value of the three KDJ indicators, a sign that the stock price is in a strong downward trend.

Average True Range (ATR):

ATR index is mainly used to measure the intensity of market volatility and show the rate of market change [22].

Williams Indicator (WR):

WR uses oscillators to reflect the phenomenon of overbought and oversold in the market. It can predict the highs and lows of the cycle, thus showing effective buying and selling signals. It is a technical index used to analyze the short-term market trend.

On Balance Volume (OBV):

The theory behind OBV is that volume is the key force to move the stock price. This index calculates the running cumulative total trading volume of a stock to determine whether the volume is flowing in or out. An increasing OBV reflects positive volume pressure that can potentially lead to higher prices of a stock.

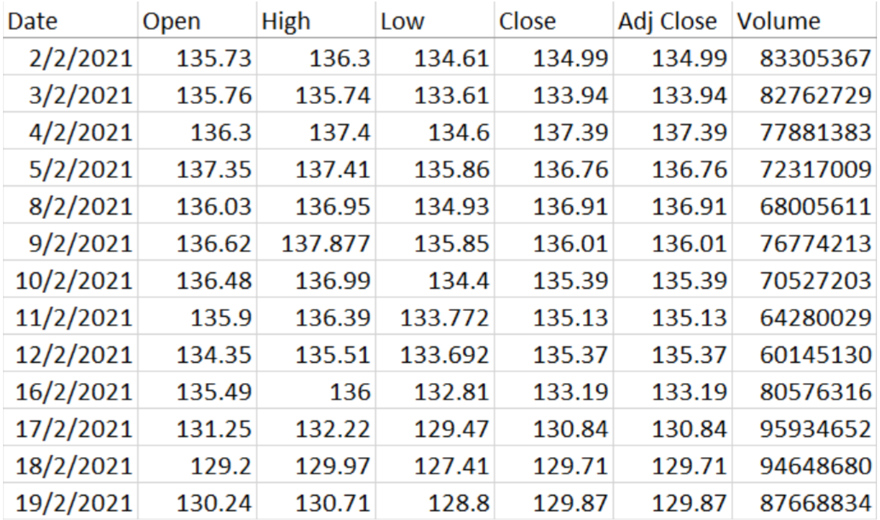
Chaikin Money Flow (CMF):

CMF is used to measure Money Flow Volume over a defined period. Money Flow Volume measures the buying and selling pressure of a stock in a single period and CMF sums up the Money Flow Volume in a defined period.

7 Appendix B: Format of Stock Price Data

The format of the downloaded data is as follows: (stock code).csv. Example such as AAPL.csv. This shows that the stock of this csv file is Apple, Inc.

The downloaded csv includes the stock's Date, Open, High, Low, Close, Adj. Close and Volume.



Date	Open	High	Low	Close	Adj Close	Volume
2/2/2021	135.73	136.3	134.61	134.99	134.99	83305367
3/2/2021	135.76	135.74	133.61	133.94	133.94	82762729
4/2/2021	136.3	137.4	134.6	137.39	137.39	77881383
5/2/2021	137.35	137.41	135.86	136.76	136.76	72317009
8/2/2021	136.03	136.95	134.93	136.91	136.91	68005611
9/2/2021	136.62	137.877	135.85	136.01	136.01	76774213
10/2/2021	136.48	136.99	134.4	135.39	135.39	70527203
11/2/2021	135.9	136.39	133.772	135.13	135.13	64280029
12/2/2021	134.35	135.51	133.692	135.37	135.37	60145130
16/2/2021	135.49	136	132.81	133.19	133.19	80576316
17/2/2021	131.25	132.22	129.47	130.84	130.84	95934652
18/2/2021	129.2	129.97	127.41	129.71	129.71	94648680
19/2/2021	130.24	130.71	128.8	129.87	129.87	87668834

Figure 37: Sample Data

Open indicates the price at which a stock started trading at the start of that day.

High indicates the highest price at which a stock is traded during that day.

Low indicates the lowest price at which a stock is traded during that day.

Close indicates the price at which a stock's last trade at the end of that day.

Adj. Close price factors in any corporate actions to reflect a stock's value.

Examples such as stock splits, dividends and rights offerings.

Volume indicates the number of shares that changed hands during a given day.

8 Appendix C: Inputs to the Machine Learning model

Three types of data are inputted to the machine learning model, including the stock price data, technical indicators and news sentiments. The following lists out each of the data inputted:

- Stock Price Data
    - Open Price
    - High Price
    - Low Price
    - Close Price
    - Volume
- Technical Indicators
    - 12-Day SMA
    - 26-Day SMA
    - 50-Day SMA
    - 12-Day EMA
    - 26-Day EMA
    - MACD (12, 26, 9)
    - MACD Signal
    - MACD Difference
    - PSAR
    - 6-Day RSI
    - 12-Day RSI
    - Stochastic Oscillator KDJ
    - WR
    - BB Middle Band
    - BB High Band
    - BB Low Band
    - BB Band Width
    - BB Percentage Band
    - ATR
    - OBV
    - CMF
- News Sentiments
    - Mean Sentiment Score

9 Appendix D: Project Planning

9.1 Distribution of Work

Task	Alan	Howard	Simon	Gordon
Do the Literature Survey	○	○	●	●
Analyze Stock Market	○	●	○	○
Types of Data Required	○	○	○	●
Stock market Selection	○	○	○	●
Design Data Scraping Techniques	○	○	○	●
Design Machine Learning algorithms	●	○	○	○
Design the Database	○	○	●	○
Design the Backend Server	○	●	●	○
Design the Web Application	○	●	○	○
Design the Mobile Application	●	○	○	○
Develop the Stock Price Data Scraper	○	○	○	●
Stock Price Data Preprocessing	○	○	○	●
Develop the Stock News Data Scraper	●	○	○	○
Stock News Data Preprocessing	●	○	○	○
Develop the Machine Learning Algorithms	●	○	○	○
Build the Database	○	○	●	○
Build the Backend Server	○	●	●	○
Build the Web Application	○	●	○	○
Build the Mobile Application	●	○	○	○
Test the Data Scraper	○	○	○	●
Test the Machine Learning Algorithms	●	○	○	○
Test the Database	○	○	●	○
Test the Backend Server	○	●	●	○
Test the Web Application	○	●	○	○
Test the Mobile Application	●	○	○	○
Write the Proposal	○	○	○	●
Write the Monthly Reports	○	○	○	●
Write the Progress Report	○	○	○	●
Write the Final Report	○	○	○	●
Prepare for the Presentation	○	○	○	●
Produce the FYP Short Video	○	●	○	○

● Leader ○ Assistant

9.2 GANTT Chart

Task	July	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr
Do the Literature Survey	■	■								
Analyze Stock Market		■	■							
Types of Data Required		■	■	■						
Stock market Selection		■	■	■						
Design Data Scraping Techniques		■								
Design Machine Learning algorithms		■	■	■	■	■	■	■		
Design the Database		■	■	■	■	■	■			
Design the Backend Server		■	■	■	■	■	■			
Design the Web Application			■	■	■	■	■			
Design the Mobile Application			■	■	■	■	■			
Develop the Stock Price Data Scraper		■	■							
Stock Price Data Preprocessing			■	■	■					
Develop the Stock News Data Scraper			■	■	■	■				
Stock News Data Preprocessing			■	■	■	■				
Develop the Machine Learning Algorithms				■	■	■	■			
Build the Database				■	■	■	■	■	■	
Build the Backend Server				■	■	■	■	■	■	
Build the Web Application				■	■	■	■	■	■	
Build the Mobile Application							■	■	■	
Test the Data Scraper			■	■	■	■				
Test the Machine Learning Algorithms				■	■	■	■	■		
Test the Database					■	■	■	■	■	■
Test the Backend Server					■	■	■	■	■	■
Test the Web Application					■	■	■	■	■	■
Test the Mobile Application							■	■	■	■
Write the Proposal		■	■							
Write the Monthly Reports				■	■	■				
Write the Progress Report						■	■	■		
Write the Final Report								■	■	■
Prepare for the Presentation									■	■
Produce the FYP Short Video									■	■

10 Appendix E: Required Hardware & Software

10.1 Hardware

Development PC:	PC with MS Windows 10
Server PC:	Linux environment with 10GB hard drive
GPU:	GTX1080Ti

10.2 Software

Programming Languages:	Python, Java, Dart
Deep Learning Library:	PyTorch
Data Crawling Libraries:	yfinance, Requests, NordVPN Switcher
Data Processing Libraries:	Pandas, BeautifulSoup, Technical Indicators
Database:	MySQL
Application Framework:	Spring
Web UI Framework:	Angular
Mobile App UI Framework:	Flutter

11 Appendix F: Meeting Minutes

11.1 Minutes of the 1st Project Meeting

Date:	24 April 2020
Time:	1745 - 1834
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**
   1.1. This is the first group meeting, so there are no minutes to approve.

2. **Report on Progress**
   2.1. All team members read through all Final Year Project Topics.
   2.2. All team members came up on interested topics.

3. **Discussion Items**
   3.1. All members discussed about interested topics.

4. **Goals of the Following Meeting**
   4.1. Finalize a topic for the FYP.

5. **Next Meeting**
   5.1. The next online team meeting will be held on 25 April 2020.

11.2 Minutes of the 2nd Project Meeting

Date:	25 April 2020
Time:	2100 - 2305
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**

   1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**

   2.1. All team members decided on the most interested topic.

3. **Discussion Items**

   3.1. All members agreed on working on a FYP about financial prediction.

   3.2. All members decided on working on "A system for predicting stock price and offering financial advice (COMP/DSCT/CPEG)" supervised by Professor Dik-Lun LEE.

   3.3. All members agreed on drafting an email to the professor showing why our group will be successful in this project.

4. **Goals of the Following Meeting**

   4.1. Draft an email of interest to the FYP advisor for the FYP project.

   4.2. Draft a project outline using PowerPoint to the FYP advisor for the FYP project.

5. **Next Meeting**

   5.1. The next online team meeting will be held on 26 April 2020.

11.3 Minutes of the 3rd Project Meeting

Date:	26 April 2020
Time:	2100 - 2138
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**

   1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**

   2.1. Gordon has drafted an email of interest to the FYP advisor.

   2.2. Howard has drafted a PowerPoint outline to the FYP advisor.

3. **Discussion Items**

   3.1. All members reviewed the email and PowerPoint.

   3.2. Small changes were made to the email to mention the added PowerPoint.

4. **Goals of the Following Meeting**

   4.1. Start looking into the details of the project.

   4.2. Research on the basics of financial investment.

   4.3. Research on stock market prediction.

   4.4. Research on algorithmic trading.

5. **Next Meeting**

   5.1. The next online team meeting will be held on 8 June 2020.

11.4 Minutes of the 4th Project Meeting

Date:	8 June 2020
Time:	2100 - 2325
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. Approval of Minutes

1.1. The minutes of the previous meeting were approved without amendment.

2. Report on Progress

2.1. All members have done some research on stock market details and news processing.

3. Discussion Items

3.1. Simon said one possible method is to use LDA for event extraction.

3.2. Another possibility is to use NER for event extraction.

3.3. We can also use knowledge graph to build a keyword relationship. So that we know what stocks are related to one piece of news.

3.4. We need to check how to map news events to stock market.

3.5. We believe US market is a more predictable market than HK / China stock market. We are thinking we can only do US market in order to achieve a higher accuracy and avoid having to process Chinese news article in order to get Chinese stock predictions.

3.6. Whether we should only do a few stocks prediction and master it or do many stocks and not getting as good prediction.

3.7. Is it possible to get intraday data because it provides a much better details of the stock market movements? Sometimes news events happened in the morning is already reflected in hours. This change can be better interpreted by the AI.

3.8. Are daily numbers good enough for daily prediction or we must use minute data in order to get a more accurate prediction.

3.9. We need to check what stock data we can use.

3.10. We also need to check on different kinds of investment strategies.

4. Goals of the Following Meeting

4.1. Research on available historical stock price API.

4.2. Research on US market.

4.3. Research on Machine Learning for text processing.

4.4. Research on financial investment strategies.

5. Next Meeting

5.1. The next online team meeting will be held on 12 June 2020.

11.5 Minutes of the 5th Project Meeting

Date:	12 June 2020
Time:	2100 - 2245
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**
   1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**
   2.1. Researched on available historical stock price API.
   2.2. Researched on US market.
   2.3. Researched on Machine Learning for text processing.
   2.4. Researched on financial investment strategies.

3. **Discussion Items**
   3.1. Alan has made a paper trading account for trying out investment for the first time.
   3.2. We discuss about the scope of the stock market that we will be doing in this project.
   3.3. We agreed to focus on U.S. stock market due to more variety and expected to be simpler in analyzing the news because of abundant English sentiment extraction tools.
   3.4. We discussed how to analyze the type of news and what should be the output of the news analysis.
   3.5. We need a news severity detector to analyze how much effect is a news towards a change in stock.
   3.6. We should try to search for big news, see how big its effect, determine how much it affects.
   3.7. We discussed about time series analysis and news summarization (PacSum) based on BERT.
   3.8. We discussed about portfolio management and risk management.

3.9. Portfolio management and risk management should be easy to do and no need machine learning involved. Probably use a survey to calculate the users' acceptable risks and produce an optimal portfolio.

4. Goals of the Following Meeting

4.1. Ask professor about whether doing the US market is a good idea.

4.2. What can we improve on the outline?

5. Next Meeting

5.1. The next online team meeting will be held on 13 June 2020.

11.6 Minutes of the 6th Project Meeting

Date:	13 June 2020
Time:	1015 - 1100
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon, Prof Lee
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**
   1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**
   2.1. All members have gathered some questions to ask Prof Lee.

3. **Discussion Items**
   3.1. Report on the scope. Howard suggests doing one or two stocks in the US stock market. Reasons being more diversity in sectors and afraid of stock manipulating.
   3.2. Prof Lee say that there should not much difference in USA and HK stock market so we should focus more on strategies, techniques, and algorithms like sentiment analysis. US has an advantage of bigger market and information are more readily available.
   3.3. Previous FYP uses Yahoo Finance which only has by day stock data. Howard suggests using a more detailed stock data but there is a cost of buying stock data, ~100 US for one stock for 5 years by hour. Professor suggests seeking help from Business school which have some Bloomberg data, but a permission may be needed to get the data. Professor also question that is 5 years data enough? Therefore, data collection became a serious problem of our project.
   3.4. Professor raises up a question about doing short-term or long-term stock prediction. Howard suggest it is not really good to use by day data. It is not common to have big events that influent the market over days. If this solution requires hourly price, then we need to look for hourly or even minute data. If it is a long-term event, then daily data is good enough. It is okay to do both. For example, international event, chain event. Trade war,

release of a report by a financial company. One of the possibilities is to look at how specific event will affect stock price.

3.5. How to determine whether an event will have short- or long-term impact. Simon suggests event analysis by stock which use the related keywords to search the news and then consolidate news from different source into one news. Some technique can be used like event analysis by sector, use authoritative source, do summarization and similarity check, and use LDA to classify the topics.

3.6. Professor Lee suggests we may apply BERT in time series analysis because building the knowledge graph for explanation is a huge topic. Also, Prof emphasize the topic is to include financial technique and with a little bit financial analysis. Therefore, reading research paper about financial technique for stock prediction is necessary. Moreover, the system should be data independent because it is supposed to work in different stock market.

4. Goals of the Following Meeting

4.1. All team members will research on event analysis, trend prediction, BERT and Time series analysis.

4.2. Howard will research on portfolio management.

4.3. Simon will research on event analysis.

4.4. Gordon, Simon, and Alan will research on stock price prediction.

4.5. Gordon will check out some libraries for collecting stock prediction data.

5. Next Meeting

5.1. The next online team meeting will be held on 20 June 2020.

11.7 Minutes of the 7th Project Meeting

Date:	20 June 2020
Time:	1400 - 1505
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**
   1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**
   2.1. Howard did some research on portfolio management.
   2.2. Simon did some research on event analysis.
   2.3. Gordon, Simon and Alan did some research on stock prediction.
   2.4. Gordon checked out some libraries for collecting stock prediction data.

3. **Discussion Items**
   3.1. Simon introduced some lecture notes about risk management. Simon found some useful event analysis tool.
   3.2. Gordon found a tool called Open IE which can extract the subject action and object as a tuple.
   3.3. We agreed on the news should be rather neutral and quick to produce so as to quickly process and feed it to the system to get output. We can do similarity search after that so as to avoid calculating multiple news source referring to the same news. Potential time series analysis including LSTM, ARIMA, SVM. These methods are good but far from using the state-of-the-art strategies. We are discussing on whether to include financial statement.
   3.4. Howard is reading a book about risk management.
   3.5. Howard suggested that we can also apply some business ideas. We can look at some business / investment books for more investment technical detail.

4. **Goals of the Following Meeting**
   4.1. Research on Simon’s mentioned event analysis paper.

5. Next Meeting

5.1. The next online team meeting will be held on 27 June 2020.

11.8 Minutes of the 8th Project Meeting

Date:	27 June 2020
Time:	2000 - 2120
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**
    1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**
    2.1. All members have researched on Simon's Event analysis paper and some other event extraction papers.

3. **Discussion Items**
    3.1. Gordon mentioned no libraries are present with the state-of-the-art event extraction strategies.
    3.2. We believe that we should not spend too much time on coding of event extraction. We should better find a library for news extraction.
    3.3. We discussed how news should be gathered, by crawling. We should only use the largest news companies. First news explaining the event will be considered.
    3.4. Gordon suggested the simplest method would be to use Google to search news with a keyword of a company.

4. **Goals of the Following Meeting**
    4.1. Find a method to put the extracted data into the time series model.
    4.2. Find an evaluation strategy of the model.

5. **Next Meeting**
    5.1. The next online team meeting will be held on 4 July 2020.

11.9 Minutes of the 9th Project Meeting

Date:	4 July 2020
Time:	2000 - 2130
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**

    1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**

    2.1. All members have done some research on time series model.

3. **Discussion Items**

    3.1. Gordon and Alan have knowledge on NLTK. NLTK has inherent issues about too much abstraction. It is going to be inaccurate and only slightly better than random guess.

    3.2. Simon suggests we should separate the price fluctuation and period consequences to compare sentiment analysis and the effect on the stock.

    3.3. Simon guesses we make an easy model first before getting into state-of-the-art strategies. This can be served as a baseline. We can crawl stock price and news. We can select large volume stocks for stock price and news crawling.

    3.4. With this model, it is not possible to generate an explainable AI to reference the prediction. We need to find a way to see how much error the AI is generating.

    3.5. We can use Google finance to look for Bloomberg historical news. We use Yahoo Finance to get historical daily news. Preprocessing we can directly put the news into sentiment analysis. Removing any html and irrelevant stuff.

4. **Goals of the Following Meeting**

    4.1. All team members will further research on appropriate machine learning methods to correlate the stock price and the financial news.

4.2. Howard will study about investment strategies and creating a survey to calculate the risk index of the users.

5. Next Meeting

5.1. The next online team meeting will be held on 11 July 2020.

11.10 Minutes of the 10th Project Meeting

Date:	11 July 2020
Time:	1400 - 1600
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. Approval of Minutes

1.1. The minutes of the previous meeting were approved without amendment.

2. Report on Progress

2.1. All team members had researched on appropriate machine learning methods to correlate the stock price and the financial news.

2.2. Howard studied investment strategies and created a survey to calculate the risk index of the users.

3. Discussion Items

3.1. We have completed the FYP status report as requested by Prof Lee.

3.2. We suggested our target users to be long term stock holding people who want to invest but have no finance background or knowledge. The instant news notification feature is also useful to day traders.

3.3. We plan to incorporate stock prices and dividends together so as to balance risks and return.

3.4. Currently the news crawler we find is not specific to financial events. We need to look for a better news crawling method.

3.5. Alan mentioned google news crawler, but this is not specific to any reputable news company and type of news.

4. Goals of the Following Meeting

4.1. All team members will try to find alternate methods to gather financial news.

5. Next Meeting

5.1. The next online team meeting will be held on 18 July 2020.

11.11 Minutes of the 11th Project Meeting

Date:	18 July 2020
Time:	2000 - 2120
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**
   1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**
   2.1. All members have researched on alternative methods to collect financial news.

3. **Discussion Items**
   3.1. Review professor's email. We think that it is better to focus more on strategies and portfolio management. The goal of the FYP is not to predict the exact price of the stocks but to minimize the risks and maximize profits of investment. We can include interests and different sectors.
   3.2. For crawling historical news, we will use GoogleNews API because there are no better and free methods in collecting a lot of historical financial news.

4. **Goals of the Following Meeting**
   4.1. All members start doing the FYP proposal.

5. **Next Meeting**
   5.1. The next online team meeting will be held on 25 July 2020.

11.12 Minutes of the 12th Project Meeting

Date:	25 July 2020
Time:	2015 - 2220
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**

    1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**

    2.1. All members started doing the FYP proposal.

3. **Discussion Items**

    3.1. We continue to work on project proposal as mentioned by professor to be submitted by the end of July.

    3.2. Mobile app notice will be sent to users if the change in predictions is largely affected by an event. We hope to immediately notify the user to earn some quick money or prevent some huge losses due to the financial events.

4. **Goals of the Following Meeting**

    4.1. Continue to work on FYP proposal.

5. **Next Meeting**

    5.1. The next online team meeting will be held on 31 July 2020.

11.13 Minutes of the 13rd Project Meeting

Date:	31 July 2020
Time:	2130 - 2210
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**
    1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**
    2.1. All members were working on the FYP proposal.

3. **Discussion Items**
    3.1. Professor Lee has reviewed our FYP proposal and has given some comments. It is better if we can find some historical heat or trending topic news to include global or macro financial, economic, and political events. We can try to identify heated news by searching news topic and see how many times it occurs in a few days. Real-time news can also be done in this way.
    3.2. We use the 500 largest stocks in S&P 500. We take S&P 500 as the average of all stocks because it contains a lot of big companies. We can use S&P 500 along with global affecting financial, economic, political news to see the macro scale effect of the news.
    3.3. We will use hourly data and daily data to predict and compare to see which one is better at predicting.
    3.4. Due to the difficulty of event extraction, it is simpler to use NER or other easier models to simplify it into a binary classification problem.

4. **Goals of the Following Meeting**
    4.1. All team members will be adjusting the proposal based on Prof Lee's advice. More details will also be added into the proposal to prepare for the FYP proposal to be submitted in September.
    4.2. All members will research on the details of the S&P500.

5. Next Meeting

5.1. The next online team meeting will be held on 8 August 2020.

11.14 Minutes of the 14th Project Meeting

Date:	8 August 2020
Time:	2000 - 2200
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**
   1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**
   2.1. All members have done some research on S&P500.

3. **Discussion Items**
   3.1. We discussed about S&P500 to see whether it is really a good range of stocks to select.
   3.2. Howard mentioned that the market focuses mainly on IT, at around 26% while some sectors' effect is tiny such as materials, real estate, utilities.
   3.3. We discussed about adding multiplier on different sectors because the market cap varies greatly.
   3.4. Gordon has crawled the stocks in S&P500.

4. **Goals of the Following Meeting**
   4.1. We will work on the data collection part of FYP proposal.

5. **Next Meeting**
   5.1. The next online team meeting will be held on 15 August 2020.

11.15 Minutes of the 15th Project Meeting

Date:	15 August 2020
Time:	2000 - 2110
Place:	Online Voice Call
Present:	Alan, Gordon, Howard
Absentees:	Simon
Recorder:	Gordon

1. **Approval of Minutes**
   1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**
   2.1. All members have worked on the data collection part of FYP proposal.
   2.2. Simon was quite busy last week so he did not complete the scheduled work.

3. **Discussion Items**
   3.1. All members reviewed past week's work.
   3.2. We need to prepare for predicting 500 models in backend.
   3.3. We separate the target users to two types of users, day traders and long-term traders.
   3.4. We have discussed about our project design, including web, user registration, survey, risk index, users' preference and capital, portfolio management.
   3.5. We will give alerts on mobile application when our system detects a rise or fall in stock price.

4. **Goals of the Following Meeting**
   4.1. Howard will draw a flow chart about the design and system architecture.
   4.2. Gordon and Alan will write the overview and objectives.
   4.3. Simon will write the design of the machine learning algorithm.

5. **Next Meeting**
   5.1. The next online team meeting will be held on 19 August 2020.

11.16 Minutes of the 16th Project Meeting

Date:	19 August 2020
Time:	2100 - 2200
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**

   1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**

   2.1. Howard drew a flow chart about the design and system architecture.

   2.2. Gordon and Alan wrote the overview and objectives.

   2.3. Simon wrote the design of machine learning algorithm.

3. **Discussion Items**

   3.1. All members reviewed past week’s work. Simon raised some questions about each part of the proposal that we have written so far. Some improvements and clarifications are made to Simons questions. More charts will be added to the proposal. Some grammars can be better expressed in the proposal. More detail is needed to explain the stock data and news data.

   3.2. The design is divided into 4 parts. Model, web, database, and alert. Alert will be sent by mobile application. For database, we are planning to use MySQL to store the data and the trained models. For front end design, we will use Angular because it supports multiple platforms. For model, we hope to deploy two machine learning models. The first one is used to calculate how many intervals one event should span. The second one is used to predict the upcoming stock price based on the stock data and the news data. More details of the models are needed to be discussed to get a clear view on how to actually write the model setting that can achieve the requirements.

4. **Goals of the Following Meeting**

   4.1. Howard will complete the flow chart and system architecture design.

4.2. Gordon and Alan will continue to write the overview and objectives.
4.3. Simon will complete the literature survey about technical analysis.

5. Next Meeting

5.1. The next online team meeting will be held on 22 August 2020.

11.17 Minutes of the 17th Project Meeting

Date:	22 August 2020
Time:	2000 - 2135
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**
   1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**
   2.1. Howard finished some major part of system design.
   2.2. Simon finished the fundamental analysis part of literature survey
   2.3. Gordon and Alan wrote most part of the overview and objectives.

3. **Discussion Items**
   3.1. Simon mentioned to follow a paper about news sentiment analysis of stock price prediction. We can use one of its approach and embedding method. Instead of word2vec, we will use BERT for title and dictionary for the article.
   3.2. Although we have collected 1-minute, 2-minute data, we may not be using in this project. We will mainly use 1-hour and 1-day data in this project.
   3.3. We will use Java and node.js on backend.
   3.4. We will use Angular on frontend.
   3.5. We will use Flutter on mobile application.

4. **Goals of the Following Meeting**
   4.1. Simon will finish all the description of financial indicators in the technical literature survey.
   4.2. Howard will continue to work about the design part of database, web, and mobile application.
   4.3. Gordon and Alan will continue to write the overview and objectives.

5. **Next Meeting**
   5.1. The next online team meeting will be held on 29 August 2020.

11.18 Minutes of the 18th Project Meeting

Date:	29 August 2020
Time:	2030 - 2130
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**
   1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**
   2.1. Simon finished all the description of financial indicators in the technical literature survey.
   2.2. Howard completed the design and implementation part of database, web, and mobile application.
   2.3. Gordon and Alan finished writing the overview and objectives.

3. **Discussion Items**
   3.1. Simon raised some advice on how to improve the FYP proposal. We are deciding on the topics of the literature survey to be included in the proposal. We agreed to use Technical Analysis and Fundamental Analysis as the terms for the topics.

4. **Goals of the Following Meeting**
   4.1. Howard will complete financial strategy part in literature survey.
   4.2. Alan, Gordon, Simon will complete the remaining part of the proposal in order to get an initial version of the proposal.

5. **Next Meeting**
   5.1. The next online team meeting will be held on 6 September 2020.

11.19 Minutes of the 19th Project Meeting

Date:	6 September 2020
Time:	1600 - 1700
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**

   1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**

   2.1. Howard completed financial strategy part in literature survey.

   2.2. Alan, Gordon, Simon completed the remaining part of the proposal.

3. **Discussion Items**

   3.1. FYP proposal is basically completed. We discussed details about the placement of the data collection section. Gordon thinks it is better to put this into the design and implementation sections. Details are being discussed whether this is a good idea or not. We decided to follow this idea.

   3.2. Simon raised questions about using MySQL as database. Gordon mentioned that Howard thinks MongoDB is better in this case because MongoDB is more lightweight, recent and supports caching. Simon raised that NoSQL is better in sorting because it uses key value pair. After some discussion we will use MySQL and MongoDB for now.

   3.3. Howard talked through the database, frontend, backend, mobile application design.

4. **Goals of the Following Meeting**

   4.1. Complete and submit the FYP proposal.

5. **Next Meeting**

   5.1. The next online team meeting will be held on 13 September 2020.

11.20 Minutes of the 20th Project Meeting

Date:	13 September 2020
Time:	1500 - 1800
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**
   1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**
   2.1. The FYP proposal is completed and submitted.

3. **Discussion Items**
   3.1. We discussed about the upcoming works that we should do.
   3.2. Gordon raised that there are limited periods, e.g., 2 years, of historical stock hourly data that we can collect, and we need to create a program to merges historical stock price of different periods into one file for preprocessing.
   3.3. Gordon raised some concerns about privacy and user data management since this project will require users to enter a lot of information about their financial status.
   3.4. We need to build front end and backend with authentication.

4. **Goals of the Following Meeting**
   4.1. Gordon will complete the software that merge multiple periods of stocks to one.
   4.2. Alan will complete the data scraper for crawling news on yahoo finance.
   4.3. Howard will build an Authentication system.

5. **Next Meeting**
   5.1. The next online team meeting will be held on 11 October 2020.

11.21 Minutes of the 21st Project Meeting

Date:	11 October 2020
Time:	1600 - 1700
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**
   1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**
   2.1. Gordon has completed the code that merges stock price.
   2.2. Data crawling and authentication are still processing.

3. **Discussion Items**
   3.1. We are working on the monthly report.
   3.2. Gordon found out that the minute data and hourly data are inconsistent from different downloads.
   3.3. Alan found out that crawling data from yahoo finance only provides limited amount of news articles. Some large stocks can only get 7 days of historical news using this method.

4. **Goals of the Following Meeting**
   4.1. Gordon will find some methods to deal with the data inconsistency problem.
   4.2. Alan will look for alternative methods to crawl news.
   4.3. Howard will continue to work on the authentication.

5. **Next Meeting**
   5.1. The next online team meeting will be held on 25 October 2020.

11.22 Minutes of the 22nd Project Meeting

Date:	25 October 2020
Time:	1330 - 1530
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**
   1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**
   2.1. Gordon has completed the software to combine data of different dates into one file.
   2.2. Alan is working on the news data crawler.
   2.3. Howard is working on the authentication system.

3. **Discussion Items**
   3.1. Howard proposed that using a template for our front-end website. AdminLTE can be used since it provides admin dashboard and control panel theme which fits our project.
   3.2. We agreed to this and decided to use AdminLTE for the front-end template.
   3.3. Howard emphasized that the mobile application is only used to get latest news. The web application should be used for checking the details of the portfolios.
   3.4. Alan mentioned that he is working on a crawler than uses Google search.

4. **Goals of the Following Meeting**
   4.1. Howard will complete the authentication system.
   4.2. Alan will continue to work on the news data crawler.
   4.3. Howard will investigate the use of AdminLTE for front-end UI.

5. **Next Meeting**
   5.1. The next online team meeting will be held on 8 November 2020.

11.23 Minutes of the 23rd Project Meeting

Date:	8 November 2020
Time:	1400 - 1515
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**
   1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**
   2.1. Howard completed Authentication in Web Application.
   2.2. Alan encountered some problems about the data crawlers
   2.3. Howard is developing Front-end using AdminLTE

3. **Discussion Items**
   3.1. We reviewed the functionality of Authentication. Howard demonstrated logging in using Google and email, token storage for login session and password recovering.
   3.2. We discussed the information required for analysing users' financial situations. Users have to fill in a form after registration.
   3.3. Simon would be responsible for designing entities for Back-end.
   3.4. Alan reported the issues on data crawlers. We were getting blocked by Google as we have too many requests from google search. We were unable to crawl news reliably and quickly for all companies.

4. **Goals of the Following Meeting**
   4.1. Alan will work on solving the problems related to with Google search.
   4.2. Simon will complete the design of the entities.

5. **Next Meeting**
   5.1. The next online team meeting will be held on 14 November 2020.

11.24 Minutes of the 24th Project Meeting

Date:	14 November 2020
Time:	1700 - 1840
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**

   1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**

   2.1. Alan has completed the news crawler.

   2.2. Simon has completed the design of the entities.

3. **Discussion Items**

   3.1. Simon added JWT and spring security for authentication.

   3.2. Simon updated database entity.

   3.3. Alan updated news crawler to avoid anti-scraping.

   3.4. Howard proposed that we change the login method to use firebase token so that users can use google account and simplify registration.

   3.5. Gordon used VPN to crawl news and expected to complete crawling of 500 stocks before February.

   3.6. Gordon proposed to use FinBERT for news sentiment analysis.

4. **Goals of the Following Meeting**

   4.1. Howard and Simon will complete front-end UI and back-end API.

   4.2. Alan and Gordon will implement stock data pre-processing and model implementation.

5. **Next Meeting**

   5.1. The next online team meeting will be held on 29 December 2020.

11.25 Minutes of the 25th Project Meeting

Date:	29 December 2020
Time:	1300 - 1930
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**

    1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**

    2.1. Howard and Simon are working on the front-end UI and back-end API.

    2.2. Alan and Gordon were working on the implementation of the stock price prediction model.

    2.3. Alan filtered news that are not actually related to the stock crawled.

    2.4. Gordon applied FinBERT to predict the article’s sentiment.

    2.5. Alan used Amazon EC2 to get latest news automatically.

3. **Discussion Items**

    3.1. We decided to append the article sentiment score to stock data according to the publish date of the news.

    3.2. We selected technical indicators for embedding stock price. Gordon will generate the technical indicators.

    3.3. Hourly data created too much variance and is harder to predict for upcoming 90 days. Therefore, we decided to move to use daily data.

    3.4. The initial model of directly predicting the upcoming prices in a single shot did not work well.

    3.5. We decided to train a model to predict 90 days later.

    3.6. Simon suggested using MAPE Loss as it minimizes the bias from RMSE.

4. **Goals of the Following Meeting**

    4.1. Complete prediction using daily data.

    4.2. Gordon will generate technical indicators.

    4.3. Alan will append the article sentiment score to the price data.

5. Next Meeting

5.1. The next online team meeting will be held on 2 January 2021.

11.26 Minutes of the 26th Project Meeting

Date:	2 January 2021
Time:	2000 - 2200
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**
    1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**
    2.1. Alan and Gordon tested out the machine learning model with daily data.
    2.2. Gordon generated the required technical indicators.
    2.3. Alan appended the sentiment score to the stock price data.

3. **Discussion Items**
    3.1. We have completed most of the system parts separately. We are preparing for system integration between crawlers, models, UI and backend web server
    3.2. Alan will prepare to format and import all news to a database
    3.3. Howard mentioned that he will add more functionalities to the UI
    3.4. We have to contact communication tutors for feedback on proposal.

4. **Goals of the Following Meeting**
    4.1. Gordon and Alan will finetune the machine learning model.
    4.2. Meet with the communication tutor to get feedback on our proposal.
    4.3. Howard will add some functionalities to the UI.

5. **Next Meeting**
    5.1. The next online team meeting will be held on 5 January 2021.

11.27 Minutes of the 27th Project Meeting

Date:	5 January 2021
Time:	1700 - 1800
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon, NoorLisa
Absentees:	None
Recorder:	Gordon

1. Approval of Minutes

1.1. The minutes of the previous meeting were approved without amendment.

2. Discussion Items

2.1. We reviewed our AI model part in proposal report. NoorLisa pointed out some problems:

2.1.1. Data need to be merged with pre-processing

2.1.2. We should find out the best algorithm for modeling based on the pre-processing

2.1.3. We should review the quality of dataset and set the expected output we needed

2.1.4. We should include the less important questions in our appendix. More important questions should be prioritized and answered in the main report.

2.1.5. The technical indicators listed in the literature survey should be placed in the appendix.

2.1.6. We should list out more possible algorithms and the reasons why not using them

2.1.7. We should talk more about pipeline input, hyperparameters and tuning.

2.1.8. We should try to find the places where it becomes very actively bullish or bearish and we want to know what has happened there.

2.1.9. We should try to proof we are able to show with accuracy where are the bottlenecks that affect the stock pricing

2.1.10. We should try to quantify news sentiment and see its effect on technical analysis.

2.1.11. We should explain why we use daily data used instead of hourly data.

2.2. We reviewed our Web Application part in proposal report. NoorLisa pointed out some problems:

2.2.1. We should present the reason why the information needed to display on the Front-end and what needs of the users were fulfilled.

2.2.2. We should explain more on risk management and include the mathematical formula and show the correlation between risk, price and indicators.

2.2.3. We should have a clearer structure for the database, schema and module

2.2.4. We should show more about use-cases of the system

2.3. We reviewed our Testing and Evaluation part in proposal report. NoorLisa pointed out some problems:

2.3.1. We should compare different methods and check the accuracy

2.3.2. We should give our conclusion based on formula and visualize the portfolio.

3. Next Meeting

3.1. The next online team meeting will be held on 9 January 2021.

11.28 Minutes of the 28th Project Meeting

Date:	9 January 2020
Time:	2030 - 2200
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**
    1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**
    2.1. Gordon and Alan finetuned the current machine learning model.
    2.2. Howard have completed the Import Data API for backend server.
    2.3. Howard added cash and stocks account for users.

3. **Discussion Items**
    3.1. Howard mentioned that we need to collect dividend data for each stock.
    3.2. Discuss the structure of progress report and our future plan.
    3.3. We discussed the features of the web application, including Financial Health Analysis, Trading Record and Statement, Stock Recommendation and Stock News.
    3.4. We have discussed the functionalities of the mobile application.

4. **Goals of the Following Meeting**
    4.1. Write progress report.
    4.2. Alan and Gordon will start implementing the mobile app.
    4.3. Gordon will collect the dividend data of each stock.
    4.4. Gordon and Alan will further update the machine learning model.

5. **Next Meeting**
    5.1. The next online team meeting will be held on 24 January 2021.

11.29 Minutes of the 29th Project Meeting

Date:	24 January 2021
Time:	1000 - 1200
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. **Approval of Minutes**

    1.1. The minutes of the previous meeting were approved without amendment.

2. **Report on Progress**

    2.1. Updated the prediction algorithm and the web application system.

    2.2. Gordon and Alan have completed most of the functionalities of the mobile app.

    2.3. Gordon has collected the dividend data.

    2.4. Part of the progress report is updated.

3. **Discussion Items**

    3.1. Simon suggested to add stock id column to historical price for data manipulation and add delisted to state out the delisted company.

    3.2. Simon introduced stop concept to calculate target buy and sell price.

    3.3. We have discussed the current functionalities of the mobile app.

    3.4. We have discussed the process about sending requests to the firebase database and backend server.

4. **Goals of the Following Meeting**

    4.1. Complete a stable version of our system.

    4.2. Gordon will implement the buy and sell price formula.

    4.3. Complete the progress report.

5. **Next Meeting**

    5.1. The next online team meeting will be held on 24 February 2021.

11.30 Minutes of the 30th Project Meeting

Date:	21 February 2021
Time:	1300 - 1400
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon
Absentees:	None
Recorder:	Gordon

1. Approval of Minutes

1.1. The minutes of the previous meeting were approved without amendment.

2. Report on Progress

2.1. Gordon has implemented the buy and sell price formula.

2.2. Professor Lee has provided feedback on our progress report.

3. Discussion Items

3.1. We discussed what areas we need to improve based on Professor Lee's feedback.

3.2. We need to better format the report to make the report clear and coherent.

3.3. We have to compare and evaluate results to show that using stock price, technical indicators and news sentiments outperforms the accuracy of any of the input data alone.

3.4. We have to add the exact lists of technical indicators used and inputs to the machine learning model.

3.5. We have to justify why bidirectional LSTM is chosen by comparing it with other models such as RNN, GRU and unidirectional LSTM.

3.6. We need to get feedback from communication tutor about the progress report.

4. Goals of the Following Meeting

4.1. Start working on the final report.

4.2. Meet with the communication tutor to get feedback on our progress report.

5. Next Meeting

5.1. The next online team meeting will be held on 24 March 2021.

11.31 Minutes of the 31st Project Meeting

Date:	24 March 2021
Time:	1500 - 1700
Place:	Online Voice Call
Present:	Alan, Gordon, Howard, Simon, NoorLisa
Absentees:	None
Recorder:	Gordon

1. Approval of Minutes

1.1. The minutes of the previous meeting were approved without amendment.

2. Discussion Items

2.1. NoorLisa reviewed our progress report and gave feedback on different parts of the report.

2.2. NoorLisa clarified what our FYP should be. She mentioned that our project should mention more details of the datasets, algorithms, machine learning models and what is unique instead of the web and mobile application. The project should also provide more details on risk calculation and portfolio advisory. We need to provide equations and explanations.

2.3. NoorLisa mentioned that the system overview diagram is insufficient to describe the entire system clearly. We need to add more details to the diagram so that readers can understand our entire system without reading the texts.

2.4. NoorLisa suggested us to think about what to include in the presentation slides first so as to better outline what we want to convey in our project.

3. Goals of the Following Meeting

3.1. Work on presentation slides and final report.

4. Next Meeting

4.1. The next online team meeting will be held on 14 April 2021.