ABSTRACT

Financial data forecasting has historically been a significant domain with a wide array of consequences and applications. In today’s tech driven era, it has facilitated the prevalence of data-driven decision making, reduced uncertainties in the financial arena and propelled monetary profits unprecedentedly. Numerous techno-social algorithms and technologies have been employed to facilitate financial data forecasting. Modeling these forecasters is not easy since financial markets are unpredictable, noisy, non-linear and volatile. However, the implementation of machine learning can prove to be vital. Many researchers across the globe have attempted to use sophisticated machine learning models to predict stock prices and stock price movements accurately. However, most studies do not pay attention to small-cap stocks. This is mainly because small-cap stocks have low share prices, are highly volatile and sometimes, do not have much data available. As a result, in order to adopt a more holistic approach to stock price prediction, this research project aims to predict prices of selected small-cap and blue-chip stocks in developing and developed markets. This project aims to explore various machine learning models and evaluate these models to discover the most accurate and predictive model.

Up to this stage, the project is progressing as scheduled. A thorough study of previous research work has been done to develop a better understanding of financial data projections. A project proposal has been written and submitted to the supervisor. This proposal includes a detailed description of the project objective, methodology and expected results. Additionally, a user friendly project website has been developed to showcase all documentation. Next, python scripts have been written to collect financial data from Yahoo Finance (numeric data: historic stock data for a period of 20 years) and Seeking Alpha (non-numeric data: textual data in form of news headlines). The data has been preprocessed and Augmented Dickey Fuller test has been performed to check if the time series is stationary or not. Some technical indicators have also been calculated. Additionally, exploratory data analysis has been performed to gain insights and sentiment analysis has been performed to generate sentiment scores. Finally, a classification based approach has been adopted to predict stock price movement and algorithms such as Logistic Regression and Support Vector Machines have been implemented. These algorithms give accuracies in the range of 49% to 52%.

Currently, the machine learning model is being optimized by dropping some less important input variables and applying penalties to wrong predictions. The immediate next step would be to continue the process of optimizing existing model and perhaps, exploring powerful algorithms if needed.
ACKNOWLEDGMENT

I would like to thank my supervisor, Dr. C.L. Yip for giving me the opportunity to undertake my Final Year Project under his supervision. I appreciate him for his unconditional support and treasured inputs throughout this project. His constant motivation, valuable feedback and guidance at each step has been of immense help.

Furthermore, I would also like to sincerely thank my second examiner, Professor George Mitcheson who has been very encouraging and supportive. His thought provoking questions have enabled me to streamline my thoughts and progress in the correct direction.

Lastly, I would like to extend a special token of gratitude to Miss. Grace Chang and Dr. Ken Ho from the Centre of Applied English Studies for helping me enhance my skills to express the work done during the project.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>ii.</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENT</td>
<td>iii.</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>v.</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi.</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>vii.</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1. PERSONAL MOTIVATION</td>
<td>1</td>
</tr>
<tr>
<td>1.2. BACKGROUND</td>
<td>1</td>
</tr>
<tr>
<td>1.3. SIGNIFICANCE</td>
<td>2</td>
</tr>
<tr>
<td>1.4. OBJECTIVES AND SCOPE</td>
<td>2</td>
</tr>
<tr>
<td>1.5. REPORT OUTLINE</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>2. LITERATURE REVIEW</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>3. METHODOLOGY</td>
<td>6</td>
</tr>
<tr>
<td>3.1. PROGRAMMING LANGUAGE</td>
<td>7</td>
</tr>
<tr>
<td>3.2. DATA COLLECTION</td>
<td>7</td>
</tr>
<tr>
<td>3.3. DATA PRE-PROCESSING</td>
<td>8</td>
</tr>
<tr>
<td>3.4. DATA ANALYSIS</td>
<td>8</td>
</tr>
<tr>
<td>3.5. FEATURE ENGINEERING</td>
<td>9</td>
</tr>
<tr>
<td>3.6. MACHINE LEARNING</td>
<td>9</td>
</tr>
<tr>
<td>3.6.1. TRAIN-TEST SPLIT</td>
<td>11</td>
</tr>
<tr>
<td>3.6.2. TRAIN-TEST MODEL</td>
<td>11</td>
</tr>
<tr>
<td>3.6.3. EVALUATION METRICS</td>
<td>12</td>
</tr>
<tr>
<td>3.6.4. OPTIMIZATION</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>4. PROGRESS</td>
<td>14</td>
</tr>
<tr>
<td>4.1. CURRENT WORK</td>
<td>14</td>
</tr>
<tr>
<td>4.1.1. PROJECT PROPOSAL AND PROJECT WEBSITE</td>
<td>15</td>
</tr>
<tr>
<td>4.1.2. DATA COLLECTION</td>
<td>16</td>
</tr>
<tr>
<td>4.1.2.1. NUMERIC DATA</td>
<td>16</td>
</tr>
<tr>
<td>4.1.2.2. NON-NUMERIC DATA</td>
<td>17</td>
</tr>
<tr>
<td>4.1.3. DATA PREPROCESSING</td>
<td>18</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

FIGURE 1: Methodology Overview ................................................................. 6
FIGURE 2: Machine Learning Overview ...................................................... 10
FIGURE 3: Train Test Split ........................................................................... 11
FIGURE 4: Numeric Data .............................................................................. 16
FIGURE 5: Non-Numeric Data ..................................................................... 18
FIGURE 6: Null-Values ............................................................................... 18
FIGURE 7: Augmented Dickey Fuller Test .................................................. 19
FIGURE 8: New Features ........................................................................... 19
FIGURE 9: Correlation ................................................................................ 22
FIGURE 10: Closing Prices Over 20 Years .................................................. 22
FIGURE 11: Mean Adjusted Closing Prices ................................................ 23
FIGURE 12: Competitive Firms .................................................................. 24
FIGURE 13: Sentiment Scores .................................................................... 25
FIGURE 14: Output Variable ....................................................................... 25
FIGURE 15: Feature Importance ................................................................. 27
LIST OF TABLES

TABLE 1: Current Work and Progress Table .................................................................14
TABLE 2: Technical Indicators .....................................................................................20
TABLE 3: Project Schedule Table ................................................................................27
LIST OF ABBREVIATIONS

ADF: Augmented Dickey Fuller
API: Application Program Interface
ARIMA: Auto Regressive Integrated Moving Average
ATR: Average True Returns
EMA: Exponential Moving Average
LSTM: Long Short Term Memory
MACD: Moving Average Convergence Divergence
MAPE: Mean Absolute Percentage Error
RMSE: Root Mean Square Error
RSI: Relative Strength Index
SMA: Simple Moving Average
SVM: Support Vector Machine
VADER: Valence Aware Dictionary for Sentiment Reasoning
WMA: Weighted Moving Average
1. INTRODUCTION

The following chapter gives an overview of this research project. Firstly, the personal motivation behind undertaking this research project is briefly explained. Then, the general background and the significance of the project are presented. Following this, the objectives and scope of this project are briefly discussed. Lastly, the detailed outline of this interim report is given.

1.1. PERSONAL MOTIVATION

I was working as a technology summer analyst at J.P. Morgan Chase & Co. in Hong Kong in 2020. Here, I got the opportunity to work on the company's interactive stock visualization tool called 'Perspective'. I used this tool to stream continuous stock data and compare different stocks on the basis of their ask price and bid price. This hands-on experience ignited my interest in the stock market and motivated me to delve more into this field. As a result, I was encouraged to undertake my final year project on the topic 'Financial Data Forecasting'.

1.2. BACKGROUND

In recent times, rapid globalization and exponential increase in trade have resulted in an increase in financial transactions across the globe [1]. As a result of this surge, massive amounts of financial data is being generated every minute. This financial data can be in the form of stock prices, market indices, foreign exchange, cryptocurrency prices etc.

If this data is leveraged and converted into insights to accurately predict future stock prices, then it will allow individuals and business organizations like hedge funds and investment banks to:

i) Make well-informed, data-driven strategic and operational decisions.

ii) Mitigate risks and losses while trading and investing.

iii) Significantly increase profits.

In order to achieve the aforementioned goals and facilitate the successful prediction of
financial data, machine learning has been playing an increasingly important role. In fact, the
growth rate for intelligent systems, and robo-advisors that implement machine learning is
around 70%. Moreover, there are $2.2 trillion in assets under management for such
automated predictive systems [2].

Thus, financial data forecasting by the means of sophisticated machine learning models
form the basis of this research project.

1.3. SIGNIFICANCE

According to the Efficient Market Hypothesis and the Random Walk Theory, it is extremely
difficult to predict market movements [3]. Moreover, stock markets are often described as
“unpredictable, non-linear and dynamic” in nature [4]. This is mainly because these markets
are affected by a variety of factors such as political agendas, public sentiment, terrorist
activities, leaders of different organizations and environmental factors.

However, since there is access to massive amounts of financial information, some
researchers believe that it is possible to anticipate stock prices to a great extent by the help
of machine learning. Thus, continuous efforts have been made to predict stock prices.

The existing studies mostly focus on blue-chip stocks which are stocks with a national
reputation for quality, reliability, and the ability to operate profitably in good and bad times
[5]. This is because more financial data is easily available and less risk is involved in making
investments.

On the other hand, not much emphasis is paid on small-cap stocks which are stocks
characterized by higher volatility and higher potential returns [6]. This is because they have
low share prices and are quite risky [7].

Thus, in order to attain a more holistic approach to stock price prediction, the aim of this
project is to consider both small-cap stocks and blue-chip stocks.

1.4. OBJECTIVES AND SCOPE

The scope of this research project includes selected stocks of different sizes: small-
cap stocks and blue-chip stocks.

The aim is to predict and analyze the aforementioned stocks in markets of different sizes
through machine learning. This involves the following goals:
i) Collecting time series of financial data (refer to section 3.2.) and pre-processing it (refer to section 3.3.).

ii) Analyzing the pre-processed data to identify trends, patterns and correlations (refer to section 3.4.).

iii) Applying machine learning algorithms to predict the stock value/stock movement (refer to section 3.6.).

iv) Continuously optimizing results to achieve high accuracy (refer to section 3.6.3. and section 3.6.4.).

The ultimate goal of this research project is to experiment and evaluate the machine learning algorithm that provides the highest accuracy in terms of stock prediction projection.

1.5. REPORT OUTLINE

This interim report is structured into seven chapters. Chapter 1 has given a brief introduction to the research project. It has thrown light on the personal motivation, background, significance and the objectives of the project. Chapter 2 gives a brief summary of past studies that have been conducted. Chapter 3 explains the methodology that will be followed to achieve the objectives. It covers the way in which data will be collected, prepared, analyzed, trained and evaluated. Chapter 4 entails the current progress of the project and discusses the tasks that have been completed and that are in progress. Further, it explains the immediate next steps to be taken and the future work to be conducted. Chapter 5 throws light on the challenges faced and potential solutions to mitigate these challenges. Chapter 6 concludes this report. It consists of a summary as well as some feasible recommendations for further developments. The last chapter, chapter 7 gives a bibliography of sources that were referenced and studied to prepare this report.
2. LITERATURE REVIEW

Machine learning and deep learning have often been used to anticipate market movements and predict prices of stocks. This chapter gives a brief summary of some research papers.

i) Financial time series of forecasting using Support Vector Machines (SVM)

Kyoung-jae Kim in his research paper Financial Time Series Forecasting using Support Vector Machines proposes SVM as a promising algorithm to predict financial time series. The rationale is that the solution of SVM maybe the global optimum and the probability of overfitting is less. This is not true in the case of other neural networks according to Kim. However, Kim also mentioned that SVM is sensitive to the value of parameters [8]. To increase accuracy, analyses of optimum parameters can be further delved into.

ii) Predicting the direction of stock market prices using Random Forests

The research paper predicting the direction of stock market prices using Random Forests is written by Luckuson Khaidem, Snehanush Saha and Sudeepa Roy Dey of Cornell University. It aims to reduce the error in forecasting of stock prices by considering prediction as a classification problem and makes use of ensemble learning. It considers the use of linear machine learning algorithms as “futile” and believes that the use of random forests gives much higher accuracy [9].

iii) Stock Market Prediction using Artificial Neural Network based on High Low Points (HLP)

The research conducted by Lei Wang and Qiang Wang called Stock Market Prediction using Artificial Neural Networks based on High Low Point makes use of neural networks in order to predict prices of stocks under the obvious assumption there is stock price data is noisy at high frequency. In this study, only the highest and lowest price of each stock in a given period of time were taken into consideration [10]. This paper was presented at International Conference on Intelligent Human-Machine Systems and Cybernetics.
iv) Predicting stock and stock prices index movement using trend deterministic data prediction and machine learning techniques.

In the aforementioned research paper, authors J. Patel, S. Shah, P Thakkar and K. Kotecha, use trend deterministic data to predict movement to produce final output which gives up or down movement signals. It makes use of machine learning algorithms such as ANN, Naïve Bayes, SVM and Random Forests and involves technical indicators such as momentum, stochastic SK, MACD etc. [11].
3. METHODOLOGY

It is essential to devise a cohesive methodology in order to ensure iterative development throughout the research. Thus, this chapter elucidates the detailed steps that are required to complete the project. Firstly, this chapter presents the choice of programming language used in this research project and justifies the choice. Further, it explains the approach to collect, prepare and analyze the collected data. It then describes the machine learning approach. Lastly, the way in which various machine learning algorithms will be evaluated is discussed with the help of mathematical equations. The following figure gives a brief overview of the methodology that will be followed. A comprehensive explanation of this figure is given in the subsections of this chapter.

![Figure 1: Methodology Overview](image)
3.1. PROGRAMMING LANGUAGE

Python programming language is used for this research project for a variety of reasons. Firstly, unlike any other programming language, python offers a wide range of powerful libraries such as NumPy, Pandas and Scikit-learn. These libraries allow easy handling and manipulation of enormous amounts of data. This is crucial for the research project as a lot of financial data is being dealt with. Secondly, python is easy to learn, is open-source and is being continuously improved to match current standards. Thus, in addition to simplicity, it will also be easy to modify our models in the future. Thirdly, python offers extensive Enterprise Application Integration. This essentially means that python allows for easy integration with other programming languages and frameworks. For example, it can implement COBRA/COM components and directly integrate with languages like Java, C and C++. Lastly, python is not implemented in any particular paradigm and thus, can be implemented with a multi-paradigm approach to support functional, procedural, and object oriented programming styles [12].

3.2. DATA COLLECTION

The first step is to collect legitimate and adequate amounts of financial data. This is an integral step because data acts as a fuel to machine learning models. Additionally, insufficient data can possibly lead to unsatisfactory and unreliable predictions. In this project, two types of data are collected, namely Numeric Data and Non-Numeric data. Numeric data is essentially historic stock data and includes the high, low, open, close and volume of stock traded over a period of time. On the other hand, non-numeric data is textual data that displays public and/ or political sentiments.

Both numeric and non-numeric data is collected because stock market is not only affected by historic prices. However, it is also affected by a number of other factors which include and are not limited to political sentiments, public opinions, human psychology and environmental factors [13]. An appropriate example for this could be how a simple tweet by Elon Musk about using Signal lead to an 1100% surge in unrelated stock with a similar name [14]. Another great example is the surge of Intel stocks on replacement of current CEO by with VMware's chief [15]. Thus, in order to make predictions more robust, this project aims to encompass both aspects of numeric and
non-numeric data.

As mentioned in section 1.3, only a few selected small-cap stocks and blue-chip stocks in developed and developing countries are considered. In order to obtain a list of these stocks, their symbols and metadata, screening tools such as the Zacks' Stock Screener Tool are used. Once the list is available, python scripts are written to retrieve data of the stocks from reliable Application Program Interfaces (API's) such as AlphaVantage, InvestPy, Yahoo Finance, IEX Cloud etc. Furthermore, non-numeric data is collected by scraping either investing platforms such as Seeking Alpha, news portals such as Google News or social media sites such as Twitter and Reddit.

3.3. DATA PRE-PROCESSING

The collected data can be disorganized, inconsistent and noisy. This is not desirable as the quality of data determines how accurate the predictive model will be. As a result, it becomes imperative to pre-process, clean and manipulate the collected data in order to prepare it for exploratory data analysis.

Firstly, inconsistent or duplicate values (if any) as well as missing values are handled. The missing values are either removed or replaced with the mean/median. The latter approach is better because simply removing values might lead to a risk of losing important information whereas replacing them will give more comprehensive results. Secondly, statistical tests such as the Augmented Dickey Fuller Test (ADF) are performed to check if the time series data collected is stationary or not. Such tests are important when handling time series data as non-stationary data might lead to fabricated regression results [16]. Lastly, standardization, normalization and vectorization will be performed if needed.

3.4. DATA ANALYSIS

Once the data is pre-processed, it is prepared for analysis. Data analysis is performed in order to understand the data better and draw insights from it. These insights are in the form of patterns, trends and correlations. Common techniques of exploratory data analysis are generating bar charts, pair plots, scatter plots, candlesticks line charts etc.

Since the scope of this research project involves different stocks in diverse markets,
thorough analysis can be done to devise interesting results. For instance, the deviation of
different stocks can be calculated and pair-plots can be created in order to visualize
these deviations. This will help in determination of risks and the returns on various
stocks. Further, analysis of the best and worst single day returns can be carried out and
the reason behind peculiar results can be explored.

3.5. FEATURE ENGINEERING

The next step is to perform feature engineering because it will prepare the dataset to
be compatible with the requirements of machine learning algorithms. It will also
enhance the performance of machine learning algorithms resulting in higher accuracy.

Feature engineering involves scaling severely skewed data using logarithmic
transformations and aggregation of functions on the basis of time-series
components.

Note that Feature engineering will be performed if and only required.

3.6. MACHINE LEARNING

Once the data has been prepared and feature engineering has been performed,
machine learning algorithms can be implemented.

The following figure gives an overview of the steps involved in machine learning:
As shown in the above figure, train-test split is performed in order to divide the data into two sets: training and testing datasets. After this split, the machine learning model is fit on the training data and it is tested on the testing data. The trained model is used to predict the prices and is evaluated to check the accuracy. It is then continuously optimized to achieve high accuracy.
3.6.1. TRAIN-TEST SPLIT

Data is split into training data and testing data in the ratio 70:30 or 80:20. The following figure depicts this split of data:

The training data will be used to build machine learning models while the testing data will be used to validate the results of the models.

3.6.2. TRAIN-TEST MODEL

Various input features as well as output features of the machine learning model are decided. The input variables are used to predict an output variable. The input features can either be the raw features that were gathered during the process of data collection or else, technical indicators can be calculated as well. On the other hand, the output variable can be the closing price, daily returns or stock movement.

The problem of financial data forecasting can be considered as a regression problem or a classification problem. Regression models are run on time series using the lagged features in order to predict the numeric value of the output variable i.e. stock price while classification models are used to predict binary value of output variable i.e. stock price movement.

An example of a classic regression algorithm could be simple linear regression which gives the relation between independent variables and dependent variables. The stock price on a particular day t can be modeled as a linear function of the stock price of a future
Apart from this, Auto Regressive Integrated Moving Average (ARIMA) which considers past values in order to make predictions about future values can be used as well. It can be particularly beneficial while predicting stock prices as historic data is being used [18]. In terms of classification algorithms, Random Forest can be implemented. It is an ensemble ML algorithm which combines numerous decision trees to make predictions. Usually, in stock markets, the noise is high. This can cause trees to become enormously large in a direction which is different from what is expected. Random forests aim to minimize this error by considering the prediction of stock prices as a classification problem.

### 3.6.3. EVALUATION METRICS

In order to ensure that the machine learning models are reliable, it is essential to evaluate their accuracy and find errors.

To evaluate the adequacy of the model’s price prediction, indicators such as Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and Confusion Matrix can be made use of.

i) Root Mean Squared Error (RMSE): RMSE is the square of root of the mean of squares of all the given errors. Although it is considered as a good measure of accuracy, it is scale-dependent i.e. it is not scaled to original error [19]. Below is the mathematical equation for RMSE:

\[
RMSE = \sqrt{MSE} 
\]

\[
MSE = \frac{1}{n} \sum (y - \hat{y})^2 
\]

where 'y' is actual value, 'ŷ' is predicted value and 'n' is number of samples.

ii) Mean Absolute Percentage Error (MAPE): MAPE is a very common measure of accuracy and is essentially the average of percentage errors. It is not suitable when there are extreme values and zeroes [19].
Below is the mathematical equation for MAPE:

\[
\text{MAPE} = \frac{100\%}{n} \sum \frac{|y - \hat{y}|}{|y|}
\]

(3)

where 'y' is actual value, '\(\hat{y}\)' is predicted value and 'n' is number of samples

Low values of RMSE and MAPE show that the models are effective and efficient in predicting the closing price of stocks.

iii) Confusion Matrix: Confusion Matrix is used to measure performance for machine learning classification problems. It gives values of True Positive, False Positive, True Negative and False Negative. Using these values the precision, recall and F1 score can be generated. These scores can help in interpreting the performance of the model. For instance, recall tells how many values have been predicted correctly out of all positive classes whereas Precision tells how many values are actually positive out of all positive classes that are predicted correctly [19].

3.6.4. OPTIMIZATION

In order to achieve high accuracy, it is important to continuously optimize the machine learning models. This is done by finding hyper parameters that give the best possible performance. For instance, grid search can be implemented to iterate through different combinations of parameters. After this iteration, optimal parameters that give the highest accuracy are selected. Alternatively, features with more importance can also be visualized in order to decide which feature should be considered as an input variable and which feature can be dropped. Additionally, in case of neural networks, the number of nodes per layer and the activation function can be varied to produce better results.
4. PROGRESS

This chapter gives a detailed synopsis of the work that has been completed until now. Additionally, it presents the work planned for the near future in order to ensure timely completion of the project. A tabular schedule highlighting the major milestones of the project has also been given.

4.1. CURRENT WORK

Until now, satisfactory progress has been made as planned. The following table summarizes the tasks that have been completed and the tasks that are in progress. Details of each task that has been completed have been given subsequently.

Table 1: Current work and progress schedule

<table>
<thead>
<tr>
<th>TASK</th>
<th>STATUS</th>
</tr>
</thead>
</table>
| • Conduct literature review, read research papers.  
• Discuss expectations with the supervisor.  
• Finalize objectives and scope. | Completed |
| • Formulate project proposal.  
• Setup project website. | Completed |
| • Collect financial data:  
  • Numeric Data.  
  • Non-Numeric Data. | Completed |
4.1.1. PROJECT PROPOSAL AND PROJECT WEBSITE

As indicated in Table 1, several research papers have been read in order to acquire a good understanding of prior research conducted in similar fields. After preliminary research and an in-depth discussion with the supervisor, the objectives and scope of the project were finalized. Next, a project proposal was formulated which includes a detailed description of the objective of the research project, project schedule, methodology and the outcomes expected. A user-friendly web application was also set up to showcase the overview of this project and present all documentation. The web application can be accessed through the link: [https://wp.cs.hku.hk/fyp20006/financial-data-forecaster/](https://wp.cs.hku.hk/fyp20006/financial-data-forecaster/).
### 4.1.2. DATA COLLECTION

As mentioned in section 3.2, two types of data have been collected: numeric and non-numeric data. The process of data collection is briefly described below:

#### 4.1.2.1. NUMERIC DATA

Python scripts were written in order to collect data from Yahoo Finance. The reason for selecting Yahoo Finance is that it provides full access to valuable stock information in real time for no cost.

Data collection was facilitated by installing relevant packages in Python and making use of the Yahoo Finance APIs to collect data between desired periods of time. Stock data over a period of 2 decades has been collected. The following figure is an excerpt of Bank of America stock data from 1999 to 2019:

![Figure 4: Numeric data](image)

Figure 4 depicts the stock data that contains 'Date', 'High', 'Low', 'Open', 'Close', 'Volume' and 'Adj. Close'.

The structure of the collected data is explained below:
• Date: Date is the index of the DataFrame.
• High: High denotes the highest value of the stock on a particular trading day.
• Low: Low denotes the lowest value of the stock on a particular trading day.
• Open: Open denotes the opening price of the stock on a particular trading day.
• Close: Close denotes the closing price of the stock on a particular trading day.
  o It simply gives the cash value of a share of the stock at the end of a particular trading day.
• Volume: Volume denotes the volume of stock that a company traded on a particular trading day.
• Adj. Close: Adj. Close is the adjusted closing price of the stock on a particular trading day.
  o Sometimes, the closing values of the stock are regulated by the companies. It analyses the stock's dividends, stock splits and new stock offerings.
  o Adjusted closing price is considered to give a better idea of the overall value of the stock [20].

Not that if the stock closing price is not regulated then close and adj. close are essentially the same.

4.1.2.2. NON-NUMERIC DATA
As mentioned in section 3.2, the scope of this project has been expanded by including non-numeric data in addition to numeric data. This was done in order to achieve a more robust and holistic approach towards financial data forecasting. Data was collected from a popular, New York based crowd sourced investment platform called Seeking Alpha. The following figure shows an excerpt of Bank of America stock headlines:
As shown in Figure 5, the data is structured in the form of 'News Headlines', 'Year' in which the headline was published and the 'Date' on which the headline was published.

4.1.3. DATA PREPROCESSING

As mentioned in section 3.3, the data might contain some inconsistencies and might not be stationary. As a result, preprocessing of the data was performed and is briefly described below:

4.1.3.1. NUMERIC DATA

In order to ensure that there are no null values, all empty values were dealt with and a heat map was generated to validate this, as can be seen in Figure 6.

In addition to this, the Augmented Dickey Fuller (ADF) Test was performed to check if the time series is stationary or not. According to the ADF Test which is a unit root test, the presence of a unit root indicates that the time series is not stationary. The p-value obtained should be less than 5% in order to reject the null hypothesis [16]. On performing this test, it was found that the time series is not stationary. This is because the p-value calculated is greater than 5% as can be seen in the figure below and thus, null-hypothesis cannot be rejected.
As a result of this, a number of additional features and technical indicators were calculated. These are explained in section 4.1.4.

Furthermore, currencies that are not in USD and converted to USD by using the google rate.

### 4.1.3.2. NON NUMERIC DATA

In the case of non-numeric data, initially, the date on which the headlines were posted was in the string format. It was then converted into date-time format in order to facilitate analysis with stock prices in later stages of the project. Next, a new column called 'Year' was created from the 'Date' column in order to analyze the number of headlines posted each year. Lastly, punctuation marks such as '!', ':' etc. and stop words such as 'a', 'am', 'are' etc. were removed from the headlines as well. This was done because punctuation marks and stop words are unnecessary and do not add to the information given by the headlines.

### 4.1.4. NEW FEATURES

A number of new features such as daily returns, moving averages, increase in volume etc. were calculated. These new features can potentially be used as input variables for the machine learning models in the later stage of this project. The new features created are shown in the following figure:

![Figure 8: New Features](image)

The new features calculated included some important technical indicators. Stock technical indicators are essentially statistical calculations based on the prices, volume, significance of share etc. [21].
The following table explains various technical indicators along with their mathematical formula:

**Table 2: Technical Indicators [21]**

<table>
<thead>
<tr>
<th>SNo.</th>
<th>Technical Indicator</th>
<th>Brief Explanation</th>
<th>Mathematical Formulae</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Simple Moving Average (SMA)</td>
<td>SMA is a simple average calculation of the closing price of any security for a given number of days.</td>
<td>( \frac{\text{Sum ( Price, } n \text{ )}}{n} ) where ( n ) = Time Period</td>
</tr>
<tr>
<td>2.</td>
<td>Weighted Moving Average (WMA)</td>
<td>WMA is same as SMA, except it applies more weight to recent forecasts and gradually less as we look back in time</td>
<td>((\text{Price } \times \text{weighting factor}) + (\text{Price previous period } \times \text{weighting factor-1}))</td>
</tr>
<tr>
<td>3.</td>
<td>Exponential Moving Average (EMA)</td>
<td>EMA assigns lesser weight to past data and it is based on a recursive formula that includes in its calculation all the past data in our price series.</td>
<td>( (\text{Price } - \text{previous EMA}) \times \frac{2}{n+1} + \text{previous EMA} ) where ( n ) = Time Period</td>
</tr>
<tr>
<td>4.</td>
<td>Relative Strength Index (RSI)</td>
<td>RSI calculates a ratio of the recent upward price movements to the absolute price movement.</td>
<td>(100 - \left\lfloor\frac{100}{1+\text{Average loss/Average gain}}\right\rfloor)</td>
</tr>
<tr>
<td>5.</td>
<td>Moving Average Convergence Divergence (MACD)</td>
<td>MACD reveals changes in the strength, direction, momentum, and duration of a trend in a stock's price.</td>
<td>(\text{EMA(26)} - \text{EMA(12)})</td>
</tr>
</tbody>
</table>
The ADF Test was run once again and this time, the p-value was less than 5%. Thus, the null hypothesis can be rejected and the time series is stationary.

4.1.5. DATA ANALYSIS

After pre-processing the data, conducting some fundamental statistical tests and generating new features, the data was prepared for analysis. Exploratory data analysis is essential because in-depth analysis of data allows identification of various patterns, correlations and trends and thus, facilitates better understanding of the data. Moreover, data analysis helps to gain insights and draw inferences. The data analysis performed is explained below by taking Bank of America as a reference stock:

4.1.5.1. NUMERIC DATA

- **CORRELATION**

Correlation provides a better understanding about the dependence of variables on each other and if any of the variables are highly correlated or not correlated at all. Thus, the correlation between original features as well as the new features was calculated. The following figure shows this correlations diagrammatically:
Correlation ranges from -1 to 1 where -1 indicates perfectly negative correlation, 0 indicates no correlation while 1 indicates perfectly positive correlation. It shows the strength between various features and it can be said that the cooler colors depict lower correlation between different features while the hotter colors depict stronger correlations. As we can see, some features are more related to other features and the more correlated features can be made use of as input variables of the machine learning model.

- **CLOSING PRICE**

In order to understand how the closing price fluctuates and what can be the potential reasons behind these fluctuations, a line graph was plotted.
Figure 10 is the line graph for Bank of America stocks and a couple of inferences can be drawn from this graph. Firstly, there is a sharp dip in 2001. On doing some research, it was found that this dip can be explained by the terrorist attacks in America in the same year. Next, there is a peak in prices from 2004-2007 which can be explained by the boom in housing market. Finally, there is a sharp dip in 2008-2009 and this dip can be explained by the bursting of the housing market bubble. These explanations behind rise and fall of stock prices make it quite evident that stock market is not only dependent on historic stock data and there are a wide range of underlying factors on which market predictions rely.

- **MEAN ADJUSTED CLOSING PRICE**

In order to further understand and spot market trends, time resampling was performed to aggregate the stock data into a defined period of time. The average adjusted closing prices on a yearly basis over a period of 20 years were calculated. In order to visualize this, the average adjusted closing prices for Bank of America stocks were displayed in the form of bar charts as shown below:

![Mean Adjusted Closing Price Chart](image)

*Figure 11: Mean Adjusted Closing Price*

From Figure 11, it can be seen that there is a fall in the average adjusted closing prices from 2009-2012. This was also the time period of the financial crisis caused due to housing bubble burst in America.
4.1.5.2. NON-NUMERIC DATA

Some elementary analysis was conducted to check the total number of headlines collected as well as the number of headlines posted each year. Furthermore, the presence of competitor firms was also detected in the headlines. As we can see in Figure 12, Citibank and J.P. Morgan Chase have been mentioned a significant number of times.

![Textual data](image)

*Figure 12: Competitor firms*

Lastly, visual displays of textual data are generated with the help of word clouds. Doing this highlights the most prominent words.

4.1.6. SENTIMENT ANALYSIS

Sentiment analysis helps to analyze textual data and gives a numeric value on a scale for the sentiment behind this data. It determines whether the piece of text is objective or subject and then determines if the subjective type of text contains positive or negative sentiments [22] Sentiment analysis is performed in order to gain an overview of public sentiments, emotions, attitude, evaluations and opinions behind the headlines. It thereby, helps to gauge the positive and negative nature of the headlines.

In this project, VADER (Valence Aware Dictionary for Sentiment Reasoning) Sentiment Analyzer which adopts a lexicon based approach in order to detect the polarity within texts has been used. Lexicon approach essentially means the individual words, phrases, or entire headlines in this data set will be labelled with a sentiment score. For instance, 1 could be extremely positive, 0 is neutral and -1 is extremely negative.

VADER Sentiment Analyzer is selected over other sentiment analyzers for a number of reasons. Firstly, it works exceedingly well on social media data and can handle emoticons, slangs and emoji’s very well. Additionally, it is extremely fast as compared to other complex models such as Support Vector Machines (SVM) which might take hours and is also computationally economic. It can be used with online streaming data and is not affected by the tradeoff between speed and performance. Apart from this, it is worth
noting that VADER Sentiment Analyzer does not require any training data as the scores are based on pre-trained human curated gold standard sentiment lexicon [23]. In order to perform sentiment analysis, NLTK library is installed and VADER Lexicon is downloaded. Then, the sentiment analyzer is run to calculate the sentiment scores for each news headline. This is done by using a loop to pass each headline into the analyzer. Finally, a sentiment score is assigned to each headline. The compound score of positive, negative and neutral polarities calculated is shown in the figure below:

![Figure 13: Sentiment Scores](image)

Now, some dates might have more than one headline being posted and so, the scores for all the headlines is combined to get one score.

### 4.1.7. MACHINE LEARNING MODEL AND RESULTS

Preliminary implementation of machine learning algorithms has been conducted by approaching the problem of financial data forecasting as a 'classification problem'. The main goal of the model that has been implemented is to predict the value of t+1 day based on previous days data. The output variable is a binary variable as can be seen in figure 14.

```python
df['Predict'].unique()
array([0, 1])
```

![Figure 14: Output variable](image)
It stores 1 if the closing price of the next day is greater than today and 0 otherwise. This is done to forecast if the price tomorrow will be lower or higher than the price today. The input variables are a combination of raw features collected and technical indicators calculated.

Firstly, the dataset was split into training and testing sets in the ratio 70:30. After this, the data was scaled by making use of MinMaxScaler which is available in the sklearn.preprocessing library in python. The MinMaxScaler basically subtracts the minimum value and then divides by the difference between the maximums and minimum value. In this way, it helps in transforming the features by scaling every feature to a given range.

Next, after the values are scaled, two classification machine learning algorithms, namely Logistic Regression and Support Vector Machine (SVM) are implemented on the training set and then on the test set. The training and testing accuracies for these models ranged from 49% to 52%.

Unfortunately, these accuracies are not very high. This is because a number of times, stock price movement is not due to technical indicators and historic stock data but a combination of other factors such as terrorist activities, human psychology, and political agendas as discussed earlier. However, an attempt to maximize the reliability of the model and to understand the relative importance of features, the feature importance was calculated and visualized as can be seen in figure 15:
This holds immense importance as the features with lower importance can be dropped and the algorithms can be run again in order to achieve more predictive, accurate and reliable results.

### 4.2. FUTURE WORK

In order to ensure that the project is completed in an accurate and timely manner, a detailed project schedule has been developed. This schedule indicates the internal deadlines and major milestones through the course of this project.

The following table outlines the dates and the tasks to be completed in the future:

**Table 3: Project Schedule Table**

<table>
<thead>
<tr>
<th>TIME PERIOD</th>
<th>TASK</th>
</tr>
</thead>
</table>
| January     | • Prepare interim presentation and project report.  
             | • Optimize machine learning algorithms that have been implemented.  
             | • Update project website |
As mentioned in the above table, the immediate next step is to optimize the machine learning model by tuning the hyper-parameters in order to achieve a higher accuracy and lower error. Next, in the month of February, some other machine learning regression models such as LSTM might be explored if needed. After this, in March, sentiment scores that have been calculated as mentioned in section 4.1.6. will also be made use of. This will be done by correlating the lagged score index against the stock prices. Furthermore, in end of March and beginning of April, machine learning will be implemented on other stocks as well in order to make some comparisons. The rest of April will be dedicated towards testing, completing documentation, updating web application, designing the poster and preparing the final report and final presentations.
5. CHALLENGES AND SOLUTIONS

As mentioned in section 3.2, data is a valuable resource for forecasting. It should ideally be collected at high frequencies than daily to achieve precise predictions. However, while writing python scripts to retrieve data, it was found that the APIs available do not provide information at high frequencies such as up-to-the-minute prices. This poses to be a major drawback for the research project as insufficient data can possibly reduce the accuracy of our prediction model.

A potential solution to this problem could be collecting stock prices over a long period of time at a lower frequency to ensure that there is sufficient data on which machine learning algorithms can be implemented.
6. CONCLUSION

This chapter concludes the report by providing a summary of each chapter and highlights the key results. Additionally, some feasible recommendations for future developments have also been proposed.

6.1. SUMMARY

Formulating predictive price models for the stock market is challenging due to the fact that stock markets are non-linear and noisy in nature. Moreover, stock market movements not only depend on historic stock data but also depend on a wide number of factors such as political activities, remarks by certain eminent personalities, leaders of different organizations, human psychology and opinions. However, financial data forecasting is important to make profitable strategies and decisions. Thus, this research project aims to implement machine learning on selected small-cap and blue-chip stocks in developed and developing countries. It also targets to potentially draw comparisons between them.

After studying various research works in related fields by renowned researchers, the objectives and scope of this research project were finalized. Following this, a project proposal was documented and a project website was also developed. This proposal includes a comprehensive explanation of objectives, methodology, schedule and anticipated results. Moreover, a user friendly and responsive website was developed to showcase project progress and documentation. Furthermore, python scripts were written to collect financial data. Historic financial data of 20 years along with news headlines has been collected. The data has been pre-processed and analyzed thoroughly in order to gain some insights. Sentiment scores have been generated by implementing VADER Sentiment Analyzer, Finally, machine learning using classification based approach was implemented to achieve accuracy scores in the range of 49% to 52%. Feature importance of various input variables has also been calculated which can aid optimization of the model. The roadblock encountered while doing this was the lack of
high frequency data available. In order to solve this challenge and compensate for the reduced frequency, data can be collected for a longer period of time. At this stage, attempts are being made to optimize the preliminary machine learning model to generate better results. In the future, some more sophisticated machine learning models such as LSTM might be considered.

6.2. RECOMMENDATIONS

As mentioned in section 1.3, machine learning is being implemented to forecast stock prices. However, the accuracies that have been achieved are not very high. In order to develop this project further, some deep learning algorithms such as Long-Short Term Memory (LSTM) Neural Network can be implemented. An LSTM block is merely a recurrent neural network and is used when forecasts deal with either past data or data with a sequence of events. In LSTM, the errors are back-propagated from the output and stay in the memory which helps to retain information for longer periods of time. This can be helpful when past data is made use of for stock price prediction [24].
7. REFERENCES

Retrieved from:

Retrieved from

Retrieved from

Retrieved from

Retrieved from
https://www.investopedia.com/terms/b/bluechipstock.asp

Retrieved from
https://www.investopedia.com/investing/introduction-to-small-cap-stocks/


Retrieved from
https://iknowfirst.com/stock-forecast-algorithm


Retrieved from


Retrieved from


Retrieved from
https://ieeexplore.ieee.org/document/6038160

Retrieved from

Novet, J. (2021). Elon Musk told his followers to ‘use Signal,’ leading to 1,100% surge in unrelated stock with similar name. CNBC. Retrieved from https://www.cnbc.com/2021/01/08/elon-musk-boosts-signal-app-signal-advance-stock-jumps-1100percent.html#:~:text=MenuElon%20Musk%20told%20his%20followers%20to%20use%20Signal,%20leading.unrelated%20stock%20with%20similar%20name&text=Elon%20Musk%20suggested%20that%20his,to%20Zoom%20Video%20and%20Twitter.)


Retrieved from:
https://www.researchgate.net/publication/333602124_A_Comprehensive_Study_on_Lexicon_Based_Approaches_for_Sentiment_Analysis

Retrieved from