TOPIC: FINANCIAL DATA FORECASTER

PROJECT PLAN

Supervisor: Dr. Yip Beta
Student: Anshika Rajiv
(3035489322)

OCTOBER 4, 2020
TABLE OF CONTENTS

LIST OF TABLES ................................................................................................................. 3
LIST OF FIGURES .................................................................................................................. 3

1. INTRODUCTION .................................................................................................................. 4
   1.1. BACKGROUND .............................................................................................................. 4
   1.2. OBJECTIVES AND SCOPE ......................................................................................... 4
   1.3. MOTIVATION AND SIGNIFICANCE ........................................................................... 5
   1.4. OUTLINE ...................................................................................................................... 5

2. LITERATURE REVIEW ....................................................................................................... 6

3. METHODOLOGY ................................................................................................................ 7
   2.1. DATA COLLECTION .................................................................................................... 8
   2.2. DATA PRE-PROCESSING ........................................................................................... 9
   2.3. DATA ANALYSIS ....................................................................................................... 9
   2.4. FEATURE ENGINEERING ......................................................................................... 9
   2.5. MACHINE LEARNING ............................................................................................. 9
   2.6. EVALUATION ............................................................................................................ 11

4. SCHEDULE ........................................................................................................................ 12

5. CONCLUSION .................................................................................................................... 14

6. BIBLIOGRAPHY ............................................................................................................... 15
LIST OF TABLES

TABLE 1: PROJECT SCHEDULE TABLE ................................................................. 12

LIST OF FIGURES

FIGURE 1: METHODOLOGY OVERVIEW ......................................................... 7
FIGURE 2: DATA ................................................................................................. 8
FIGURE 3: MACHINE LEARNING OVERVIEW ................................................. 10
1. INTRODUCTION

The following chapter gives an introduction of the research project. Firstly, the background and objectives of the research project are presented. Following this, the problem and motivation behind this project are discussed. Lastly, the detailed outline of this project plan is given.

1.1. BACKGROUND

In recent years, forecasting of financial data has been a topic of great interest for many researchers across the world [1]. Rapid globalization and exponential increase in trade between different countries has further resulted in an increase in financial transactions across the globe.

Every minute, massive amounts of financial data is being generated. This data includes stock prices, market indices, foreign exchange, cryptocurrency prices etc. If such financial data is predicted accurately, then it will allow individuals and business organisations like hedge funds, investment banks etc:

i) To make more well-informed data-driven strategic and operational decisions.

ii) To mitigate risks and losses while trading and investing.

iii) To significantly increase profits.

In order to achieve the aforementioned goals and facilitate successful prediction of financial data, data analytics, machine learning, deep learning and artificial intelligence are playing an increasingly important role. In fact, the growth rate for intelligent systems, and robo-advisors is around 70% with $ 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 forms the basis of this research project.

1.2. OBJECTIVES AND SCOPE

The scope of this research project includes 2 kinds of stocks: small-cap stocks and blue-chip stocks. This project aims to analyse and predict prices of stocks of different sizes (small-cap stocks and blue-chip stocks) in different markets (developed and developing countries) through machine learning.

This will be done by:

i) Collect time series of financial data (refer to section 2.1.) and pre-process it (refer to section 2.2.).

ii) Analyse the pre-processed data to identify trends, patterns, relationships and correlations (refer to section 2.3.).

iii) Apply machine learning algorithms to predict the value in the future and continuously optimise it (refer to section 2.5.).

iv) Achieve favourable accuracy of prediction (refer to section 2.6.).

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.
1.3. MOTIVATION AND SIGNIFICANCE

It is extremely difficult to predict market movements according to the Efficient Market Hypothesis and the Random Walk Theory [3]. Stock markets are often described as “unpredictable, non-linear and dynamic” in nature [4]. This is because these markets are affected by a variety of factors which include political agendas, public sentiment and environmental factors.

However, some believe that it is possible to anticipate stock prices to a certain extent and thus, efforts have been made to predict the stock prices. The existing studies mostly focus on mid-sized or large stocks as more financial data is easily available and less risk is involved in making investments. On the other hand, not much focus is paid on small-cap stocks essentially because they have low share prices and are highly volatile and risky. They are traded by individual traders and require more research [5].

As a result, through this research project, to attain a more holistic approach to stock price prediction, we aim to unfold the huge potential that lies in small cap stocks. Further, we wish to study market capitalisation of stocks (blue-chip stocks and small-cap stocks) in markets of different nature i.e. developing and developed markets.

1.4. OUTLINE

The project plan has been divided into 6 chapters. Chapter 1 has given a brief introduction of the research project. It has thrown light on the background, motivation and the objectives of the research project. Chapter 2 summarises a few researches that have been conducted in similar fields in the past. Chapter 3 explains the methodology that will be used to complete this project. It covers the way in which data will be collected, prepared, analysed, trained and evaluated. Chapter 4 describes a detailed project schedule that will be followed during the course of this research project. Chapter 5 concludes this report, mentions the progress made so far and gives a glimpse of the immediate next steps to be taken. The last chapter, chapter 6 gives a bibliography of sources that were referenced and studied to prepare this project plan.
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 \[ \alpha \]. 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, Snehanshu Saha an 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 \[ \text{[7].} \]

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 \[ \text{[8].} \] 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 \[ \text{[9].} \]
3. METHODOLOGY

It is important to devise a cohesive methodology in order to ensure iterative development and incorporate adaptability. This chapter covers the approach to collect, prepare and analyse the data. It then discusses the machine learning approach that will be adopted by explaining the rationale behind few selected algorithms.

The following figure gives a brief overview of the methodology that will be followed in this research project. The detailed explanation is given in the subsections of this chapter.

![Figure 1: Methodology Overview](image-url)
3.1. DATA COLLECTION

Collecting legitimate and adequate amount of financial data is a very integral part of this research project.

The stocks considered in this study can be broadly categorized into 2 types:

   i) **Small-cap stocks**: Small capitalization stocks, are those stocks which are characterised by higher volatility and higher potential returns, are mainly traded by individual investors [10].

   ii) **Blue-chip stocks**: Blue chip stocks are stocks with a national reputation for quality, reliability, and the ability to operate profitably in good and bad times [11].

These stocks will be chosen from countries of both developed (such as United States of America) and developing nature (such as India, China).

The following figure summarises the data that will be collected:

*Figure 2: Data*
Screening tools such as the Zacks’ Stock Screener Tool will be used to obtain a list of the aforementioned stocks, their symbols and metadata.

Stock prices will be collected from scrapping reliable sources such as AlphaVantage, InvestPy, Yahoo Finance, IEX Cloud etc.

Python scripts will be written that use the APIs to retrieve the stock price quotes for 10 years for small-cap and blue-chip stocks in markets of developing and developed nature.

3.2. DATA PRE-PROCESSING

The data needs to be pre-processed before it is ready for analysis. For this, data will be cleaned, data types will be explored, missing values will either be removed or be replaced with averages and inconsistent values will be dealt with. Normalisation and vectorisation will be performed if needed.

3.3. 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 in the form of patterns, trends and correlations. Since the scope of this research project involves different stocks in diverse markets, thorough analyses of these stocks can be done to devise some interesting results. This can potentially explain market capitalization in developing and developed countries.

In order to analyse the stock prices of various small-cap and blue-chip stocks, we can calculate the return deviation of different stocks and create pair-plots in order to visualise these rates. This will enable us to determine the risk and the returns of various stocks. Further, we can analyse the best and worst single day returns and try to understand the reason for peculiar results. Apart from this, rolling mean/moving averages can be analysed.

3.4. FEATURE ENGINEERING

The next step is to perform feature engineering because it will prepare our 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, logarithmic transforms, on-hot encoding etc.

3.5. MACHINE LEARNING

Now that the data is collected and prepared, we can implement machine learning. Data is split into training data and testing data in the ratio 70:30. 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 that we have built.
The following figure shows that the pre-proposed data on which feature engineering has been performed is split into two sets: training and testing data sets.

*Figure 3: Machine Learning Overview*
On the training data, we will implement various algorithms and continuously optimise them to increase accuracy. Some examples include but are not limited to the following:

i) **Simple linear regression**: Simple linear regression gives the relation between independent variables and dependent variables. We can model the stock price on a particular day as a linear function of the stock price of a future day \( t-30 \). This concept can be extended to **multiple linear regression** (to include multiple independent variables).

ii) **Auto Regressive Integrated Moving Average (ARIMA)**: ARIMA considers past values in order to make predictions about future values. It can be particularly beneficial while predicting stock prices.

iii) **Random Forest**: Random Forest 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 minimise this error by considering the prediction of stock prices as a classification problem.

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.

Other models such as Prophet, Support Vector Machine, Multiple Linear Regression etc can also be explored to predict prices.

### 3.6. EVALUATION

In order to ensure that the machine learning models are successful, it is essential to evaluate their accuracy. To evaluate the adequacy of model’s price prediction, we can make use of indicators such as **Root Mean Squared Error (RMSE)** and **Mean Absolute Percentage Error (MAPE)**:

i) **Root Mean Squared Error (RMSE)**: RMSE is the square of root of the mean of squares of all the given errors. It is considered as a good measure of accuracy. However, it is scale-dependent i.e. it is not scaled to original error.

ii) **Mean Absolute Percentage Error (MAPE)**: MAPE is a very common measure of accuracy and is essentially the average of percentage errors. However, it is not suitable when there are extreme values and zeroes.

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

In order to ensure that the research 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:

<table>
<thead>
<tr>
<th>DATE</th>
<th>TASKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>25.09.2020</td>
<td>• Literature Review: read research papers</td>
</tr>
<tr>
<td></td>
<td>• Finalize objectives and scope</td>
</tr>
<tr>
<td>04.10.2020</td>
<td>• Project Proposal</td>
</tr>
<tr>
<td></td>
<td>• Project Website</td>
</tr>
<tr>
<td>04.11.2020</td>
<td>• Write python scripts to collect financial data at higher frequency than daily</td>
</tr>
<tr>
<td></td>
<td>• Update project website</td>
</tr>
<tr>
<td>15.11.2020</td>
<td>• Pre-process collected data</td>
</tr>
<tr>
<td>30.11.2020</td>
<td>• Perform data analysis</td>
</tr>
<tr>
<td></td>
<td>• Update project website</td>
</tr>
<tr>
<td>10.12.2020</td>
<td>• Apply feature engineering</td>
</tr>
<tr>
<td>30.12.2020</td>
<td>• Implement machine learning algorithms</td>
</tr>
<tr>
<td></td>
<td>• Update project website</td>
</tr>
<tr>
<td>11.01.2021</td>
<td>• First presentation</td>
</tr>
<tr>
<td>24.01.2021</td>
<td>• Interim progress report</td>
</tr>
<tr>
<td></td>
<td>• Preliminary implementation</td>
</tr>
<tr>
<td>14.02.2021</td>
<td>• Implement machine learning algorithms</td>
</tr>
<tr>
<td>Date</td>
<td>Tasks</td>
</tr>
<tr>
<td>------------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>28.02.2021</td>
<td>• Update project website</td>
</tr>
<tr>
<td>15.03.2021</td>
<td>• Evaluate models</td>
</tr>
<tr>
<td></td>
<td>• Optimize models</td>
</tr>
<tr>
<td>30.03.2021</td>
<td>• Testing</td>
</tr>
<tr>
<td></td>
<td>• Review work done</td>
</tr>
<tr>
<td></td>
<td>• Final amendments</td>
</tr>
<tr>
<td></td>
<td>• Update project website</td>
</tr>
<tr>
<td>18.04.2021</td>
<td>• Final Tested Implementation</td>
</tr>
<tr>
<td></td>
<td>• Final Report</td>
</tr>
<tr>
<td>23.04.2021</td>
<td>• Final Presentation</td>
</tr>
<tr>
<td>04.05.2021</td>
<td>• Project Exhibition</td>
</tr>
</tbody>
</table>

*Table 1: Project Schedule Table*
5. CONCLUSION, PROGRESS and NEXT STEPS

In this project plan, the background, literature review, methodology and schedule have been elaborated on, in order to give insights on this research project.

So far, an in-depth study of previous research papers has been done to get a good understanding of the topic. Additionally, key objectives and scope of this research project have been finalised. A user friendly and responsive project website has also been set up.

The immediate next step would be to collect data by writing scripts to get data at higher frequency than daily. The project schedule as given in chapter 4 will be followed judiciously.
6. BIBLIOGRAPHY


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

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


