Project Plan

Project Name: Financial Data Forecaster

Supervisor: Dr. Beta Yip

Issue Date: 3/10/2015

Prepared by: Jordan Yew Tien Foo, 2012518358

Major: Computing & Data Analytics

Table of Contents

1 Summary

The purpose of this document is to illustrate project plan of financial data forecaster. In section 2, we will briefly discuss the objective and the motivation behind this project. This research is significant as it may help traders increase profit made in times of making investment decisions. In section 3, we will cover some background study of the domain of financial prediction models. In general, there are two analyses in stock prediction methodologies namely fundamental and technical. Basically a wise trader would determine his/her decision based on both analyses. For this project, we are going to cover only technical analysis based on machine learning knowledge in the field of computer science. Section 4 presents the scope of this project and some pre-requisites needed in the research. Section 5 discusses the detailed project approach about how the research is going to be conducted. In section 6, risks and its possible mitigations are listed out so that project supervisor may better understand the challenges of doing this project and guidance may be needed from supervisor from time to time. Section 7 will list out the forecast schedule of this project for the whole year. More modifications may be done in the future. Section 8 and 9 will wrap out this document with mini-conclusion and references.

2 Introduction

2.1 Objectives

Given the historical data of a number of stock prices, design an algorithm that would predict their values in the future.

2.2 Problem Statement

Stock market prediction has always been a popular topic among researchers and investors around the globe. Many people have been trying to make investment return by buying and selling financial instruments like stocks and bonds. For example, a trader would predict a future price of a stock and buy the stock before the price rises or sell it before its value declines. However, stock prices are considered to be very dynamic and susceptible to quick changes because of a mix of known parameters (closing prices, volumes of transactions) and unknown factors (rumors, election results, policy change in federal interest rates).

Despite of the volatile stock market, in this project we try to analyze the underlying price patterns and predict their stock prices based on various situations. The prediction model can then be taken in reference to making instantaneous investment decisions. If we can somewhat obtain a moderately accurate algorithm, this would result into **higher profits** made for investors.

3 Background

In the past, analysts used to make their investment decisions based on past performance, earning forecast and action plans of the companies. This fundamental analysis emphasizes more on a company rather than the actual stock of the company itself. With the evolvement of computing technologies nowadays, analysts would exploit machine learning algorithms in doing technical analysis based on the time-series data as an aid of making decisions in stock market.

Machine learning is a field of computer science that automates prediction model by learning patterns from past data using algorithms. It allows computer to find hidden insights without being explicitly programmed where to look for. In our context, we will use machine learning to train historical prices of stock data using various algorithms. After experimentations have been carried out, we will come out with the most efficient algorithm in predicting stock prices. Below are some quick literature reviews about similar research:

(a) *Shah, V. H. (2007). Machine learning techniques for stock prediction*
Vatsal (2007) has tried different algorithms like support vector machine (SVM), linear regression, decision stamps, online learning and boosting in building the prediction model. Of all the algorithms he applied, he concluded only SVM combined with boosting technique gave him satisfactory results (accuracy = 64.32%). He mentioned that there was another technique, expert weighting which looked promising but he did not manage to cover the evaluation of the results. This could be something we could look into in the latter part of this project. In his paper, he also introduced the idea of linguistic analysis of financial news results to predict stocks.

(b) Shen, S., Jiang, H., & Zhang, T. (2012). Stock market forecasting using machine learning algorithms

This report proposed the use of global stock data in association with financial data as the input features to machine learning algorithms such as SVM, particularly the correlation between the closing prices of the markets and the US markets. Shunrong, Haomiao and Tongda (2012) explained that since the connections between worldwide economies are tightened by globalization, external perturbations to local financial markets are no longer domestic. This was the reason why they believed data of oversea stock and other financial markets could be useful in predicting stock index movements. Numerical results suggested that there was a high accuracy of their model (76% on S&P 500) in trend prediction. We can verify this hypothesis in our future experimentation.

(c) *Iqbal, R. Ilyas, W. Shahzad, Z. Mahmood and J. Anjum (2013). Efficient machine learning techniques for stock market prediction*

The authors (2013) used many algorithms in stock market prediction and they had a focus on neural network algorithms whereas the previous two papers earlier did not do. We can conduct analysis on significance of neural network algorithms in the research later. After being tested with different algorithms including hybrid ones, it was observed that recurrent neural network (RNN) performed better than artificial neural network (ANN). At the same time, it was found that the success of the neural network algorithms relied heavily on the pre-processing and post-processing of data.

4 Scope

4.1 What will be included

This is a prediction problem so we will shift our focus mainly on supervised learning. Common machine learning algorithms like SVM, linear regression, ANN, boosting and other suitable supervised learning techniques will be tested on dataset.

4.2 What will not be included

We will omit unsupervised learning (clustering) for the time being. If time allows, we can explore further insights from clustering. For example, we may use K nearest neighbors clustering to check if technology stocks and oil stocks have similar attributes.

4.3 Pre-requisites

The whole project will be based on building software only. There are basically three popular tools for machine learning, namely SAS, R and Python. Since SAS is not open-source, so it will not be considered. In terms of ease of learning, Python will be better choice than R although R may have better graphical capabilities. Hence Python will be the language used and Ipython notebook will be the computing environment since it goes well with matplotlib. Machine learning tools like scikit-learn and scipy will be the main libraries used in the analysis.

4.4 Deliverables

The main deliverable of this project will be simulation model that predicts stock movements and the project website will be updated from time to time to show the progress of this project.

5 Project Approach

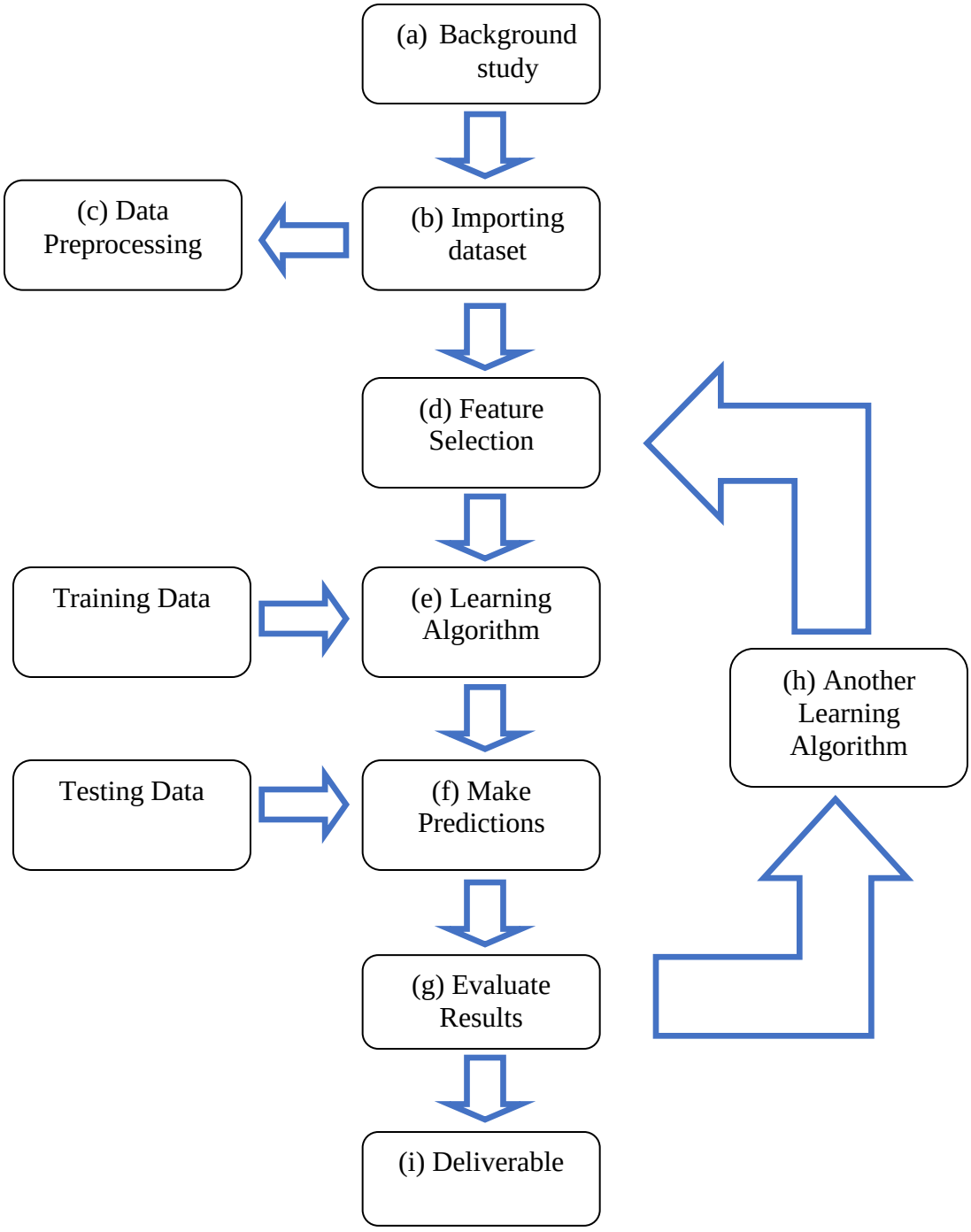

Figure 1 Flow chart of methodology

We will begin the project by doing some (a) background study about previous research at the same topic and try to adopt the useful findings and drop the insignificant methods. If we could somehow identify what has not been tried out in the past, we may add the lacking piece to our algorithms.

For the project, (b) historical data will be downloaded from http://finance.yahoo.com/ and at least fifteen stocks (or all) will be selected from Hang Seng Index to make up a portfolio. However, data collected may be dirty and contain a lot of noises so data-preprocessing (c) must be carried out. Outliers of the data should first be identified and removed. In addition, data normalization could be used to reduce data redundancy and dependency. The processed data will be cross-validated by splitting the dataset into training and testing dataset in the ratio of 7:3 as shown in the figure below.

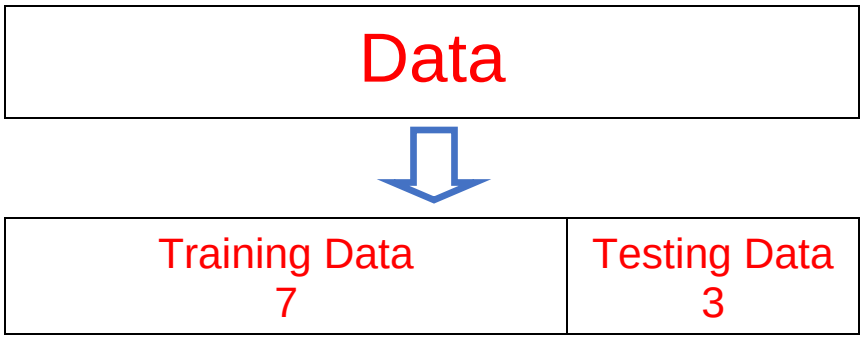

Figure 2 Cross-validation of data

Next, appropriate features that are normally used in the technical analysis of trend of stock prices (target) will be created in feature engineering. Some of them can be derived from initial raw features whereas some features like linguistic analysis from financial news need to be extracted from external sources using specific algorithms. Variable

transformation may be needed using logarithm, square or binning on features. For example, a skewed feature can be "logarithmed" to make its distribution more symmetric. Figure 3 shows a log transformation on features.

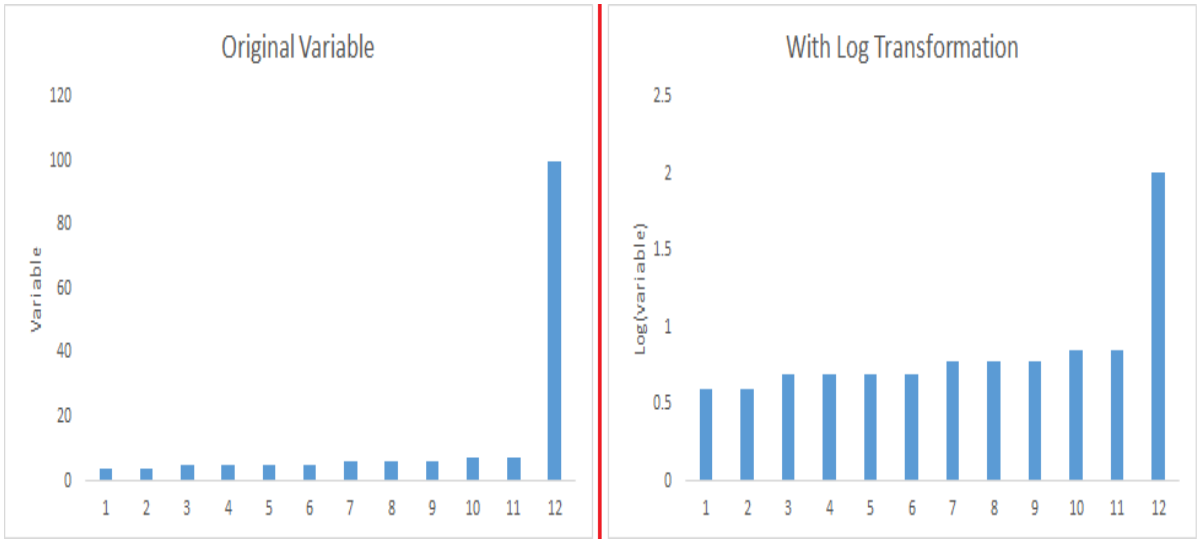

Figure 3 Log-transformation of features (variables)

Examples of basic features can be any of the following:

- Moving average: The average of the past n values till today.
- Time of day: Time of day at which the stock market is moving.
- Linguistic appearance: Frequency of stock being mentioned on financial news.

When number of features is large, computation cost (running time) could be expensive. Since not all the features are significant in this research, dimensionality reduction (d) will be necessary to eliminate unimportant features. This can be done using some algorithms like linear SVR feature ranking, permutation feature importance and different trees' feature ranking.

After feature selection, a series of machine learning algorithms (e) like regression, decision stamp, support vector machines and boosting tress will be experimented on training dataset first. Next, trained prediction model (f) will be tested on testing dataset.

At this moment, target variable of the project has not been fixed yet. It may be future prices of stock index (continuous) or stock movement (classification). It will be decided during the analysis part later since there are some difficulties in predicting stock prices directly and in such cases, stock movements will be a better target variable. Therefore, there are two possible evaluations of performance (g) based on metrics:

(i) Mean square error between true and predicted stock prices.

(ii) Accuracy percentage of stock movements.

The whole flow from (d) to (g) will be repeated to experiment various machine learning algorithms in a loop (h) until a feasible algorithm is found or the time of doing the project is running out. The best algorithm (or hybrid algorithms) will be the deliverable (i) of this project. Success criterion of the entire project relies on whether a prediction model that has accuracy higher than fifty percent on average all the time can be built. This would be more than enough to make money. Figure 4 is an example of deliverable taken from Shunrong. (2012) report.

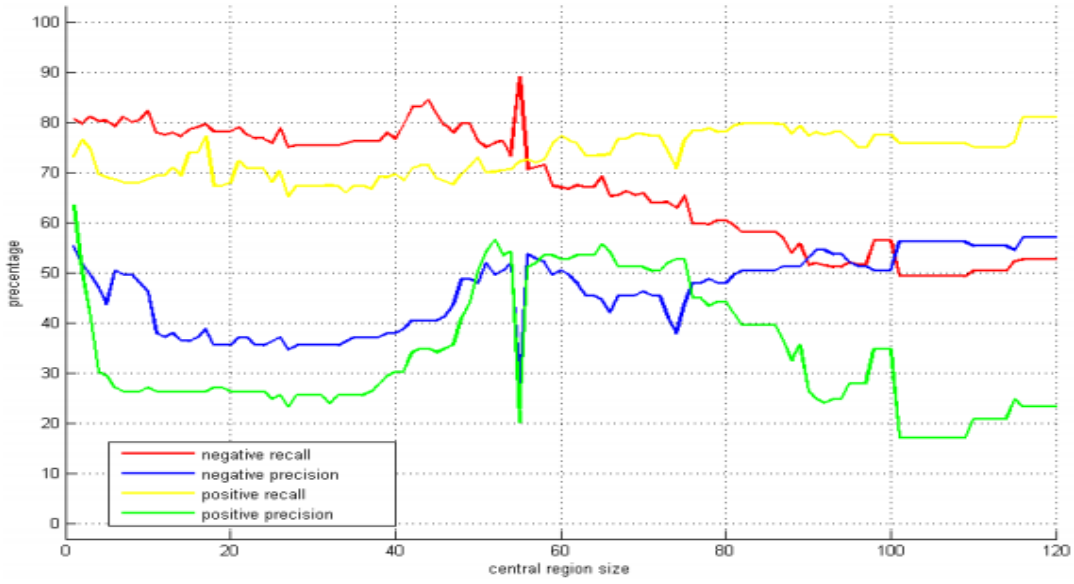

Figure 4 Sample deliverable of financial forecaster

6 Risks and Mitigations

Area of risks	Prob.	Impact	Mitigation
The distribution of data may change along the time series and we may fail to detect when the change point is and learn irrelevant data.	High	Learning irrelevant historical data will result in a lower accuracy in prediction model.	Use window size (of a year perhaps) to train model so that the model will be built based on more recent observations and hence risk is lowered.
Window of data would reduce number of data available for analysis. When big data becomes less 'bigger', accuracy may fall.	Medium	When sample size of data is not sufficient, results may not be convincing since over-fitting may happen.	To test the efficiency of using window, two models can be built based on window and all data points available respectively to make comparison in metrics.
It is difficult to capture all features in machine learning and we may overlook important features that should be considered.	High	Key features affect the success of machine learning algorithms.	More experimentations and studies need to be carried out to make sure important features are captured by the prediction model.
The Random Walk Hypothesis claims that stock prices do not depend on past stock prices, so patterns cannot be exploited.	High	This project could fail to deliver a feasible algorithm to predict stock prices since trend does not exist.	Keep tuning parameters or trying more algorithms including deep learning neural network until more efficient algorithmic model is built.

7 Schedule

	Title	Forecast Date
1	Research	1 Sep 15 – 4 Oct 15
2	Collection of data and experimentation	5 Oct 15 – 10 Jan 16
3	First presentation	11 Jan 16
4	Submission of interim report	24 Jan 16
5	More research and experimentation	25 Jan 16 – 16 Apr 16
6	Submission of final report	17 Apr 16
7	Final presentation	18 Apr 16
8	Project exhibition	3 May 16

8 Conclusion

There are many different machine learning techniques available in analyzing the problem, but machine learning itself is a challenging task especially in the context of analyzing non-linear time series data of financial stocks. It is expected that a lot of time will be used up in studying the algorithms and applying them on raw data practically. Nevertheless, I will try my best to learn the analytical tool, Python quickly and try out various techniques in tackling the problem with greatest effort.

To ensure I am working on this project in the correct direction, I will keep supervisor updated at least once every two weeks after this via email. The success of this project lies on my personal efforts and what I can learn from supervisor. Hereby, I would like to take this chance to thank my supervisor for his guidance in advance and I am looking forward to our cooperation in the coming days.

9 References

*Shah, V. H. (2007). Machine learning techniques for stock prediction.*Foundations of Machine Learning| Spring.

Shen, S., Jiang, H., & Zhang, T. (2012). Stock market forecasting using machine learning algorithms. url: http://cs229. stanford. edu/proj2012/ShenJiangZhang-StockMarketForecastingusingMachineLearningAlgorithms. pdf (visited on 05/08/2015).

Ahmad, W. (2014). Analyzing Different Machine Learning Techniques for Stock Market Prediction. International Journal of Computer Science and Information Security, 12*(12), 17.*