The University of Hong Kong

Department of Computer Science

Final Year Project 2017-2018:

Financial Data Forecaster

Detailed Project Plan

Date: 01/10/2017

Supervisor: Dr. C.L. Yip

Member: Ng Kwun Ting BEng(CompSC)

UID: 3035100897

Table of Contents

i. **Background and Objectives**

Background:

Investing on the financial markets is what a knowledge-demanded task that many professionals in the related fields and people are doing. With the computer technology nowadays, the term " Algorithmic Trading " is prevalent to the industry. In fact, there are lots of companies and individuals out there, who possess very complexing algorithms that can actually benefit from the financial market steadily. Although, there are many papers published to tell their trading algorithms, the really powerful ones are not published because of commercial secret and that is what we call the " Black Box ".

Objectives:

The goal of our project is to use scientific methods for building one of the powerful algorithms to forecast the financial data in the future. Hopefully, it can actually trade and benefit from the financial market, that is, " open the Black Box ".

ii. **Theoretical Background**

Pair trading:

Pair Trading is a method and trading strategy/concept that to find out a pair of stocks that are positive correlated. When there is a deviation between the two prices of the two stocks, there will be a force that the one (stock's price) gone above the average will go down and the other one(stock's price) gone below the average will go up until they reach an average/balance.

Artificial intelligence:

Artificial intelligence is a huge aspect that to design intelligent agent to complete a specific task optimally. In this project, the genetic algorithm, Simulated Annealing, artificial neural network(ANN) will be used as for optimization of the parameters.

Machine Learning:

Machine learning is a particular aspect of artificial intelligence, it mainly acts as a classifier to sort different items by types. In this project, classifier such as support vector machine(SVM) will be used to sort data to different types.

Time-series Analysis:

Time-series analysis is a hot aspect in statistics that considers the time-series data, based on regression methods. In this project, mainly two models will be used: Cointegration model - which proves the time-series data are positive correlated and "white noise" test - which has null hypothesis that the time-series data is a random process.

Multivariate data analysis:

Multivariate data analysis is a very important aspect in statistics that analyses various variables once. Methods such as principle component analysis(PCA) and factor analysis(FA) are used for dimensionality reduction, which is to select vital independent variables to dependent variables.

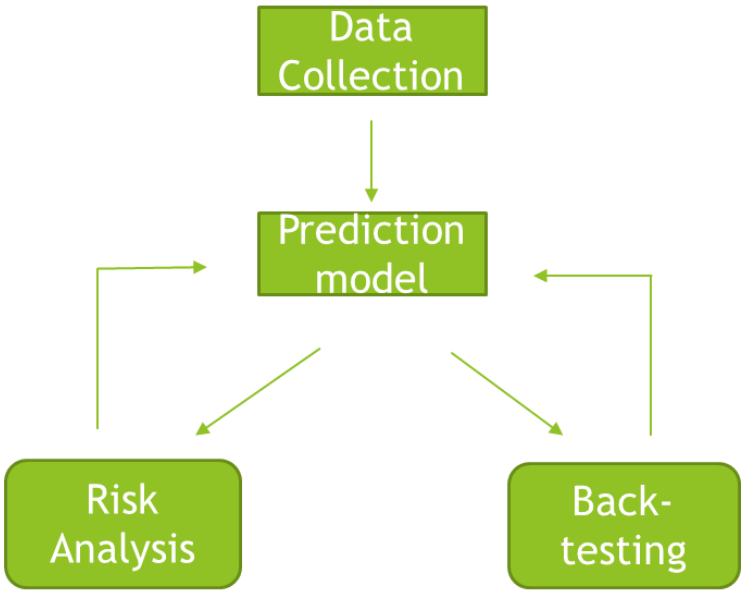
Technical indicators:

Technical indicators are developed by the finance industry to help understanding and predicting the trend of stock markets. It can be divided into two categories: trend indicators such as MACD and momentum indicators such as RSI and KD.

iii. **Methodology**

The general model of a financial forecaster is usually as follows:

1. Data Collection
2. Prediction model
3. Back-testing
4. Risk Analysis



As by the efficient market hypothesis of some economists, the financial market has reflected all the information of factors on the price , and therefore , it can not be predicted. However, some researches state that the efficient market is a concept of the whole financial markets in different countries. Within the efficient market, there are a mixture of smaller efficient and inefficient markets and this is why many predictions come.

As we believe in the second claims above that there are small inefficient financial markets contained in the big general efficient financial market, we plan to develop below method to find out the predictable stocks, which integrates pair trading, technical indicators, machine learning, statistics and artificial intelligence methods.

Description of the approach and logic of the designed method:

1. We will use the concept of pair trading to match some stocks in pairs by a time-series analysis model : Cointegration model , which can prove the two stocks are positive correlated statistically.

2. As every stock can be tested with the Cointegration model with each other stocks, there are numerous combinations of the pairs, therefore, we will first use some machine learning methods (classifiers such as SVM) to sort and reduce the combinations (it may base on the type of industries, annual financial report or derived formulas from the figures such as P/E ratio) needed to be tested with the Cointegration model.

3. Till this step, we know that, the pairs we find will have one stock's price(in a pair) go down and the other one's price(in a pair) go up in the future until the two prices attain an average(concept of pair trading). Based on this assumption, we investigate and estimate further the length of duration(by statistical or AI method) of the two stocks until the two prices attain the average.

4. Then, within the estimated duration, we will test the stocks in a pair individually with the "white noise" test (a "white noise" means a time-series is a random process that can not be predicted) to judge whether the duration(time-series) is a white noise or not, if it is not a white noise, it means an autocorrelation existed for the stock tested in that duration.

5. Till this step, we know that/assume that the selected stocks have been proved that there is a high chance that they are predictable and the future direction of their prices are also known(assumed from the pair trading concept) in that specific duration.

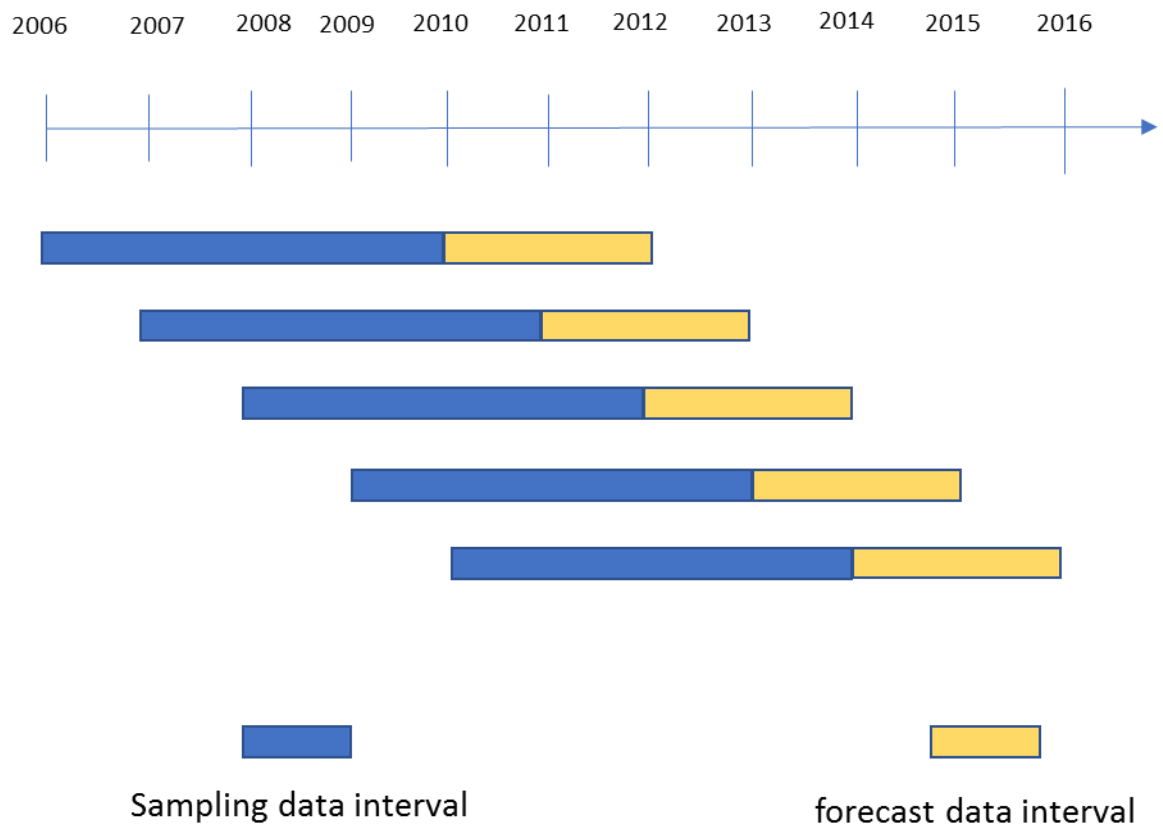
6. In this stage, we forecast the price in the future by two main approaches, the first one, we will apply the technical indicators(such as RSI, KD, MACD) and artificial intelligence methods(such as Genetic Algorithm, Simulated Annealing, ANN) to adjust and optimize the parameters of the indicators to forecast the price in that duration.

7. The other approach is to make an (logistic) regression equation by factors(derived formulas from financial industries such as P/E, Sharpe, alpha/beta risk, GVI). Then, by using Principal Component Analysis(PCA) or Factor Analysis(FA) for dimensionality reduction(to select the variables that are significant to the stock's price prediction) to get a simplified (logistic) regression equation to forecast the price in that duration.

8. Comparing the two approaches, try to mix the two approaches (technical indicators and factors) to get a more precise prediction.

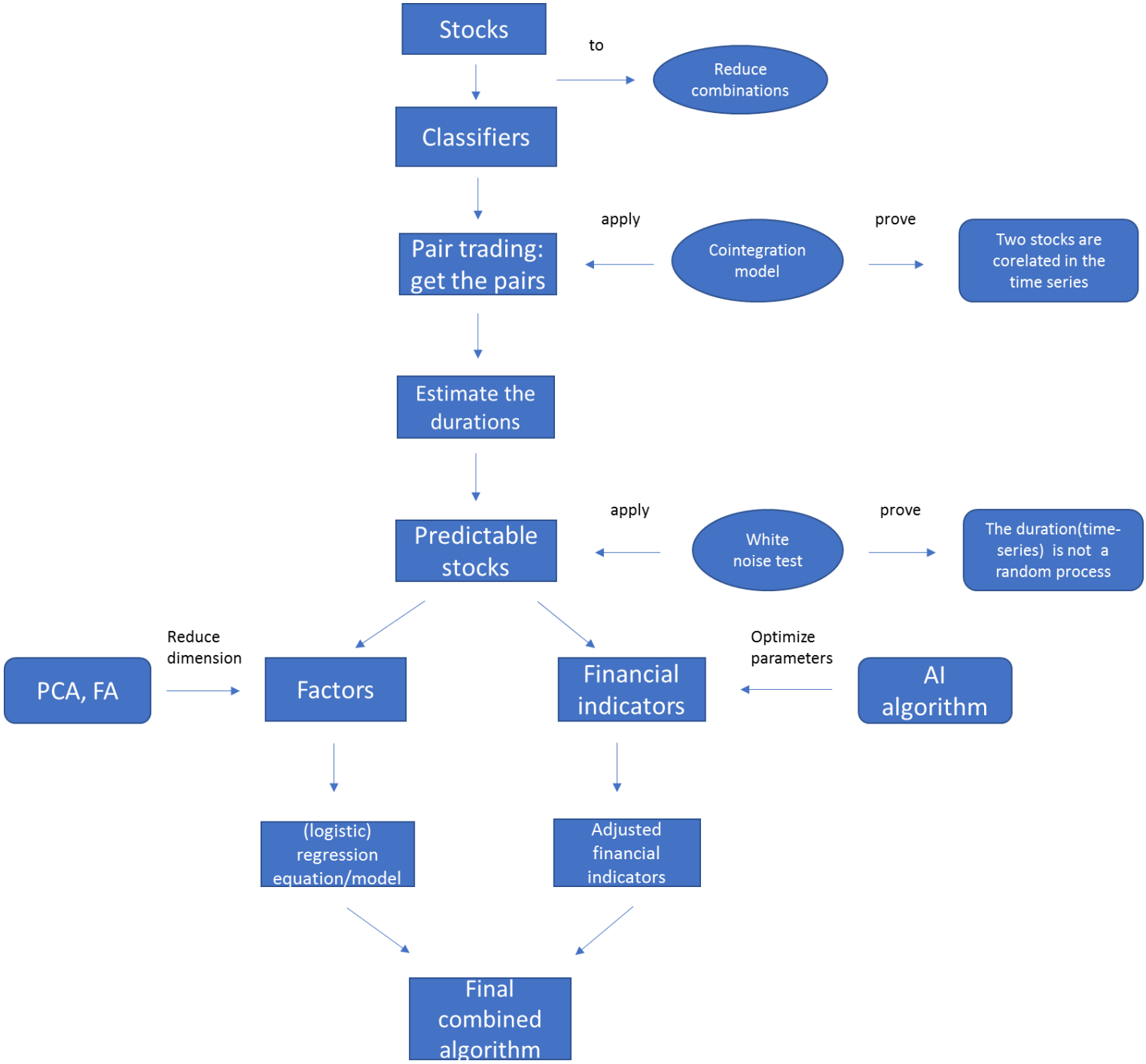
The above is the main idea and flow of our forecaster, within the process, historical data available from google/yahoo finance API (such as opening price and closing price) will be used. And more variables may be added to the model by web scraping technique from different websites. Python will be used for the project, and Python-Excel integration tool (such as xlwings) may be used too.

The historical data will be divided into two portions, one is for back-testing and fitting parameters, and the other is for the "Walk forward back-testing method" as the diagram below to reduce overfitting problem:



iv. **Scope**

Our idea can be modelled more concretely and clearly as the following diagram:



v. **Deliverables**

In total, there will be 6 deliverables:

1. Web scraper
2. pairs of stocks found, which fulfil the criterions
3. estimated durations of each pair of stocks will last for
4. adjusted/optimized/mixed financial indicators
5. simplified (logistic) regression model
6. final combined algorithm

Vi. **Challenges**

Data collection	-What variables to be used -How to get the data/variables -How difficult to get the data needed by web scraping(involve analysing the HTML structure)
Fulfilling all the criterions	-To fulfil all the criterions we set, the number of final buying/selling signals may not enough to be proved as powerful for statistical significance
Assumptions may not be true	-Although the above model has some statistical tests in some steps such as those related to time-series to make conclusion, most of statistical models have some assumptions behind and therefore the conclusion in the corresponding step may not be true

Vii. **Project Schedule**

Steps:	Deliverables:	Date:
More researches on the technical details/information required in each step		Oct
Data collection / build web scraper for the classifier to reduce combinations tested	data needed, and (Web scraper)	Oct-Nov
Test the combinations and find out pairs of stocks, which fulfil the criterions, estimate the duration of the pairs	pairs of stocks found, which fulfil the criterions, and estimated durations of each pair of stocks will last for	Nov-Dec
Applying financial indicators to the specific duration, adjust/optimize the parameters	adjusted/optimized/mixed financial indicators	Dec-Jan
Find out more variables by web scraper/ factors calculated from financial industries, then applying PCA/FA analysis	(web scraper), and simplified (logistic) regression model	Jan-Feb
Compare and Combine the findings to get a more precise model	final combined algorithm	Mar-Apr

Viii. **Summary**

Financial data prediction is a very attractive and challenging task that many people want to solve for years but is still a mystery(there is no sure winning strategy, at least, published for people). We know that to predict financial data in the future is an extremely complicated task and has a long way to go deeper and deeper. As ones of the people who want to solve this problem, we hope that through this project, we can build a possibly workable algorithm and find potential ways to go in the future researches.