David CHAN Chi Fung

8th August, 2016

P5 Capstone Project

Investment and Trading

Definition

1. Project Overview

Algorithmic trading is one of the most popular trading strategies nowadays. Algorithmic trading is a method of executing an order using automated pre-programmed trading instructions accounting They were developed so that traders do not need to constantly watch the stock and repeatedly send slices out manually.

Building up a reliable predictor is the main goal in this project, moreover a reliable predictor is a cornerstone of a successful trading-program.

In this project, a predictor is created to forecast Hang Seng Index[1] with machine learning skill. Hang Seng Index is the most widely quoted indicator of the performance of the Hong Kong stock market and is freefloat-adjusted MV weighted index with a 10% cap on individual securities.

The index level of Hang Seng Index will be predicted by related asset, which will be the input of this predictor, User can therefore choose to invest in those corresponding finance product, such as index future, index option and index ETF

All the code will be show in the ipython notebook **only** and the written project here **will not involve any code.**

[1] Hang Seng Index: https://www.hsi.com.hk/HSI-Net/HSI-Net

2. Problem Statement

The goal in this project is to predict the daily index level of Hang Seng Index which is representing the level of Hong Kong stock market. It means that there will probably be some high correlated asset with Heng Seng Index in Hong Kong

Therefore, we will firstly find out the relationship between Hang Seng Index with few popular and high traded turnover asset in Hong Kong before predicting the index level in order to find out suitable assets to be the input of the predictor. The following assets will be studied in this process:

I. Hang Seng China enterprises Index (^HSCE)

II. HSBC HOLDINGS (0005.hk)

III. AIA(1299.hk)

IV. China Mobile (0941.hk)

V. Tencent(0700.hk)

In order to run this program successfully, there are few python libraries which are necessary.

A. yahoo_finance[2]

B. pandas

C. datetime

D. matplotlib

E. sklearn

All the data will be gathered from yahoo finance[3]. In order to explore all the selected data easily, Pandas, datetime and matplotlib will help us to develop a fancy data frame and plotting all useful graph, like showing the daily price with some common used financial indicators and displaying the predicted index level in graph.

Finally, a back-testing from actual number will be displayed to estimated how reliable of this predictor. We can therefore determine whether to use this predictor.

---

[2] Yahoo Finance API: https://developer.yahoo.com/python/ , https://pypi.python.org/pypi/yahoo-finance

[3] Yahoo Finance: http://finance.yahoo.com/

Analysis

1.1 Dataset

Stock Market is affected by the whole world economic environment very much, such as exchange rate, interest rate, commodity-market and government policy. All these external factor is changing day by day, so **1 year** data is being suggested to be used in this project in order to have a similar external economic environment when the program is learning from the data.

Adjusted closing price[4], a stock's closing price on any given day of trading that has been amended to include any distributions and corporate actions that occurred at any time prior to the next day's open, is going to be utilized for analyzing and all the data will be gathered into a single dataframe to make it cleaner as shown:

	^HSI	^HSCE	0005.hk	0700.hk	1299.hk	0941.hk
2015-07-02	26282.32	12784.65	65.92	156.03	50.74	97.32
2015-07-03	26064.11	12608.98	65.92	154.64	50.00	96.98
2015-07-06	25236.28	12231.43	64.36	146.06	48.53	93.54
2015-07-07	24975.31	11827.30	64.88	144.27	49.07	92.91
2015-07-08	23516.56	11107.30	62.32	134.40	46.18	87.28

(Data.Head())[5]

1.2 Data Exploration

The correlation table below shows that the chosen assets have a high level of correlation to Hang Seng Index.

	^HSI	^HSCE	0005.hk	0700.hk	1299.hk	0941.hk
^HSI	1.000000	0.977592	0.876626	-0.030809	0.796288	0.845834
^HSCE	0.977592	1.000000	0.916670	-0.199795	0.679673	0.785587
0005.hk	0.876626	0.916670	1.000000	-0.422231	0.551301	0.682397
0700.hk	-0.030809	-0.199795	-0.422231	1.000000	0.476871	0.077976
1299.hk	0.796288	0.679673	0.551301	0.476871	1.000000	0.687294
0941.hk	0.845834	0.785587	0.682397	0.077976	0.687294	1.000000

(Correlation table)

For correlation,

$$\mathrm{corr}(X, Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y} = \frac{E\left[(X-\mu_x)(Y-\mu_y)\right]}{\sigma_x \sigma_y}$$

, where E is the expected value operator, cov means covariance, X and Y with expected values μX and μY and standard deviations σX and σY

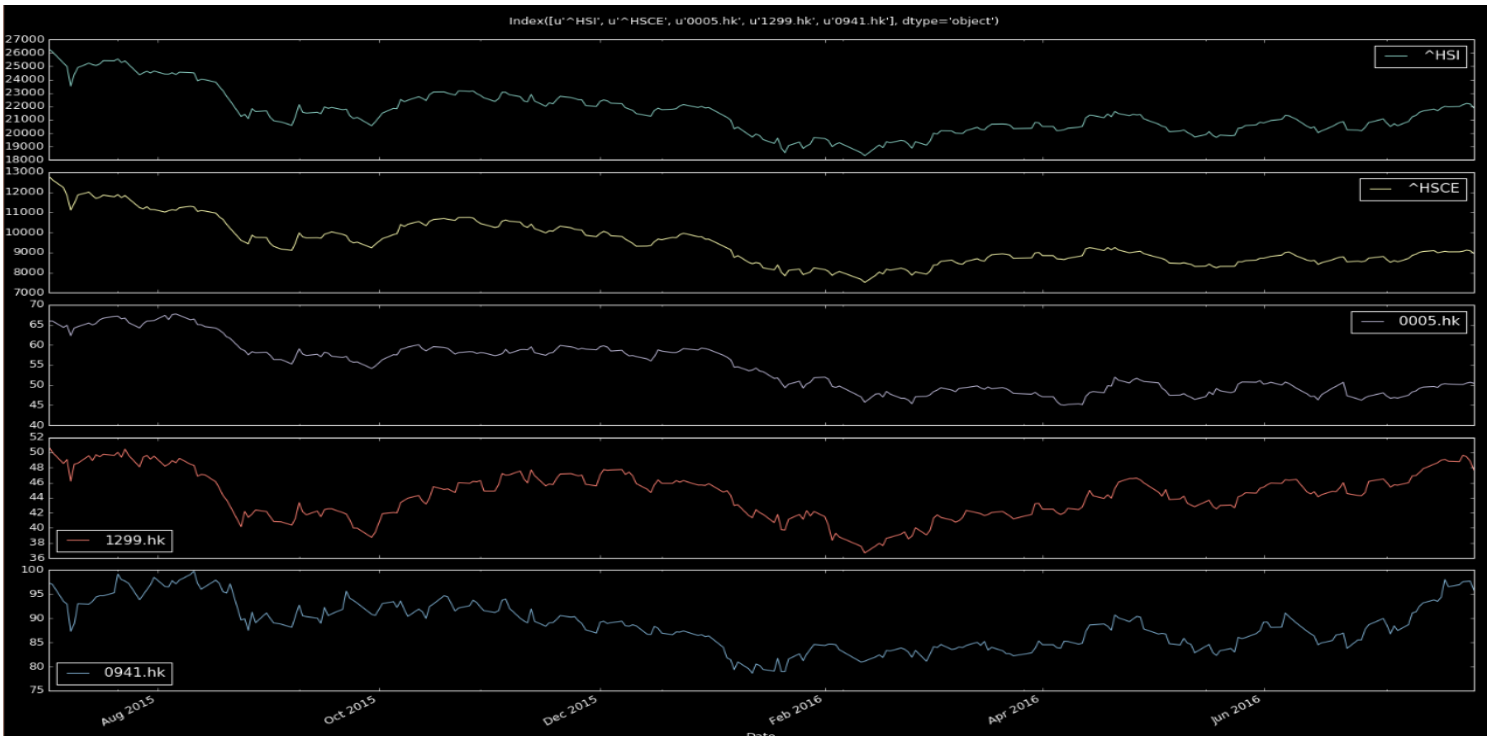
[4] Adjusted closing price: http://www.investopedia.com/terms/a/adjusted_closing_price.asp

[5] .head() is a python function to show first few row of the selected dataframe.

The correlation table shows us the relationship of each asset, the near to 1/-1, the higher relationship between asset. We can clearly find that Hang Seng index has a a significant relationship with Hang Seng China enterprises Index (^HSCE), HSBC HOLDINGS (0005.hk), AIA (1299.hk) and China Mobile (0941.hk).

Tencent (0700.hk) will be dropped from the dataset as it has a very low correlation with Hang Seng Index. Although it is one of the most popular and biggest listed company in Hong Kong Exchange, it seems that Tecent cannot provide an obvious information to predict the level of Hang Seng Index.

The follow figure is plotting the movement of each asset,



showing that the chosen assets have a very similar movement to Hang Seng index.

Finally, in order to make this project to be more financial and friendlier to investors, here is a defined function graph_with_indicator to exploring a single asset. This function might not be related to the machine learning part very much, however, it will be useful for investors to go via the asset as it can produce two common used financial indicators, Exponential Moving Average (EMA)[6] and Relative Strength Index (RSI)[7].

[6] Exponential Moving Average(EMA): http://www.investopedia.com/terms/e/ema.asp

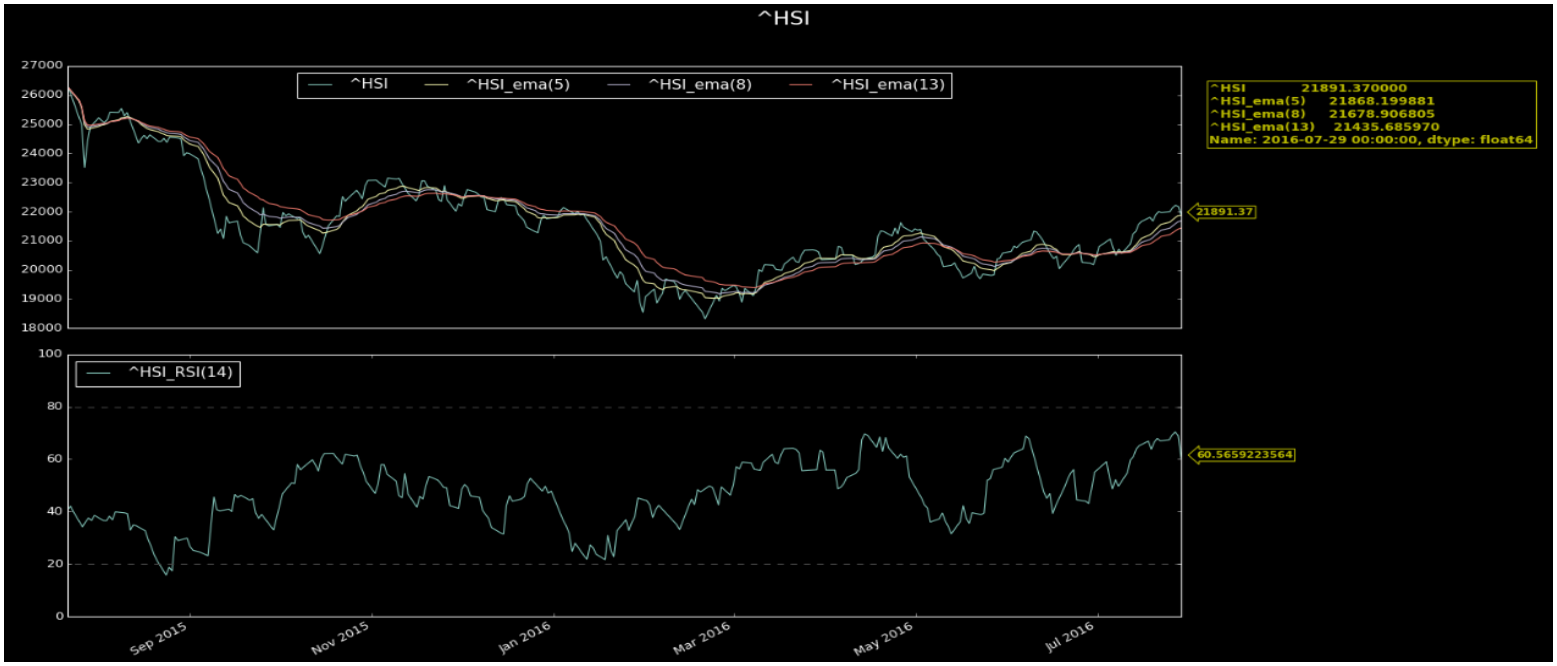
[7] Relative Strength Index (RSI): http://www.investopedia.com/terms/r/rsi.asp

Both Exponential Moving Average (EMA) and Relative Strength Index (RSI) are very common used technical analysis indicators in financial trading that will provide some kind of buying and selling signal.

As it is not the main part of the machine learning part, rather than explaining too much here, there will be two reference website providing related information of these. But it is important to note that this function is an extended function for investors to use this program rather than just focusing on machine learning and predicting.

Here is an example to display Hang Seng Index in graph, and more explanation will be in ipython notebook.

Example:



As shown above, this function will provide the EMA in 3 different windows, 5, 8 and 13, also the RSI in window =14. Moreover, there will be an annotation for the last price of the asset.

1.3 Benchmark

When we are determining how reliable or useful of this predictor, a benchmark will be necessary. As this project is trying to build a predictor to provide a market signal to make their trading decision, an absolute error will be here.

Within +5/-5 % error will be a preferable benchmark here in order to guarantee the user has at least 95% chance to earn money on average if they trade rely on this predictor.

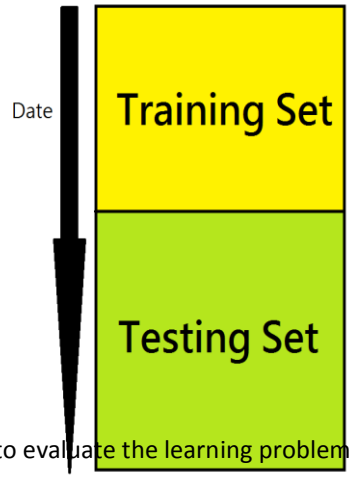
Predictor

In this part, some algorithms and Techniques will be involved here to explain how the predictor works. After defining all the metrics that will be utilized in this machine learning problem, an example will be shown to estimate how reliable this predictor will be.

1.1 Data splitting

Before inputting data into machine learning function, we should first splitting data into two different set, training set and testing set. Training set will be used to train our predictor to be the best estimator and testing set is being used to testing how reliable the predictor is.

As it is a time series data, if the time period time of testing set is before training set, the predictor will probably suffer from look-ahead bias. Therefore, the training set will be before the testing set.



1.2 Performance metric

In this project, R^2 will be used as performance to evaluate the learning problem as it is a continuous numerical data. Here is the formal of it:

$$R^2 = 1 - \frac{SSR}{SST}$$

$$\text{SSR} = \sum_{i=1}^{N} e_i^2$$

Where e is the error term.

$$\text{SST} = \sum_{i=1}^{N} (y_i - \bar{y})$$

Normally, it is as high as preferable for R^2 representing that indicates the proportion of the variance in the dependent variable that is predictable from the independent variable. In this project, the less error the more preferable, R^2 will be great here to explain how reliable our predictor is as it is focusing on error term mostly, the higher error from prediction, the lower R^2 it will be.

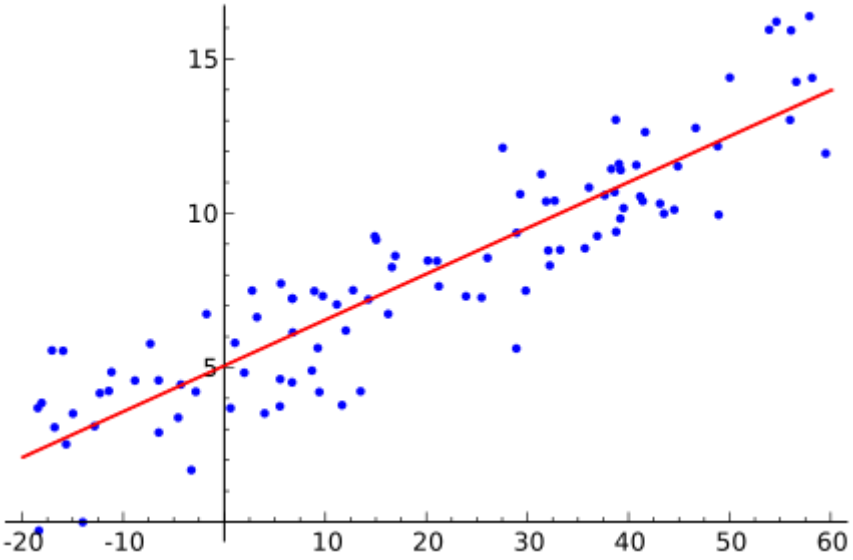
Therefore, a high R^2 will be preferable in this project.

1.3 Estimator

To solve the continuous number data, linear regression will be suitable in this case. The calculation of regression is like this:

$$y_i = \beta_0 + \beta_i x_i \quad , i = 1 \cdots n$$

It means that we can just simply input(x), then we can calculate the output(y) where the output will be the best fit of the correspond data. Here is the graph simple shows this logic.



In this project, the input will be the price of HSCE, AIA, HSBC holding and China Mobile and the output will be the Hang Seng Index.

It is important to note that the output value is the necessary equal to the true value, but it will be the best fit one. For a good regression estimate, the predicted value will very near to the true value.

1.6 Implementation

In this program, an updated data is very important as the finance market is changing day by day. It means that the more updated data the more desirable. Moreover, the finance market behavior is different in different period of time and the market information is too different in different period of time. Therefore, the length of the period in the whole dataset should not be too long.

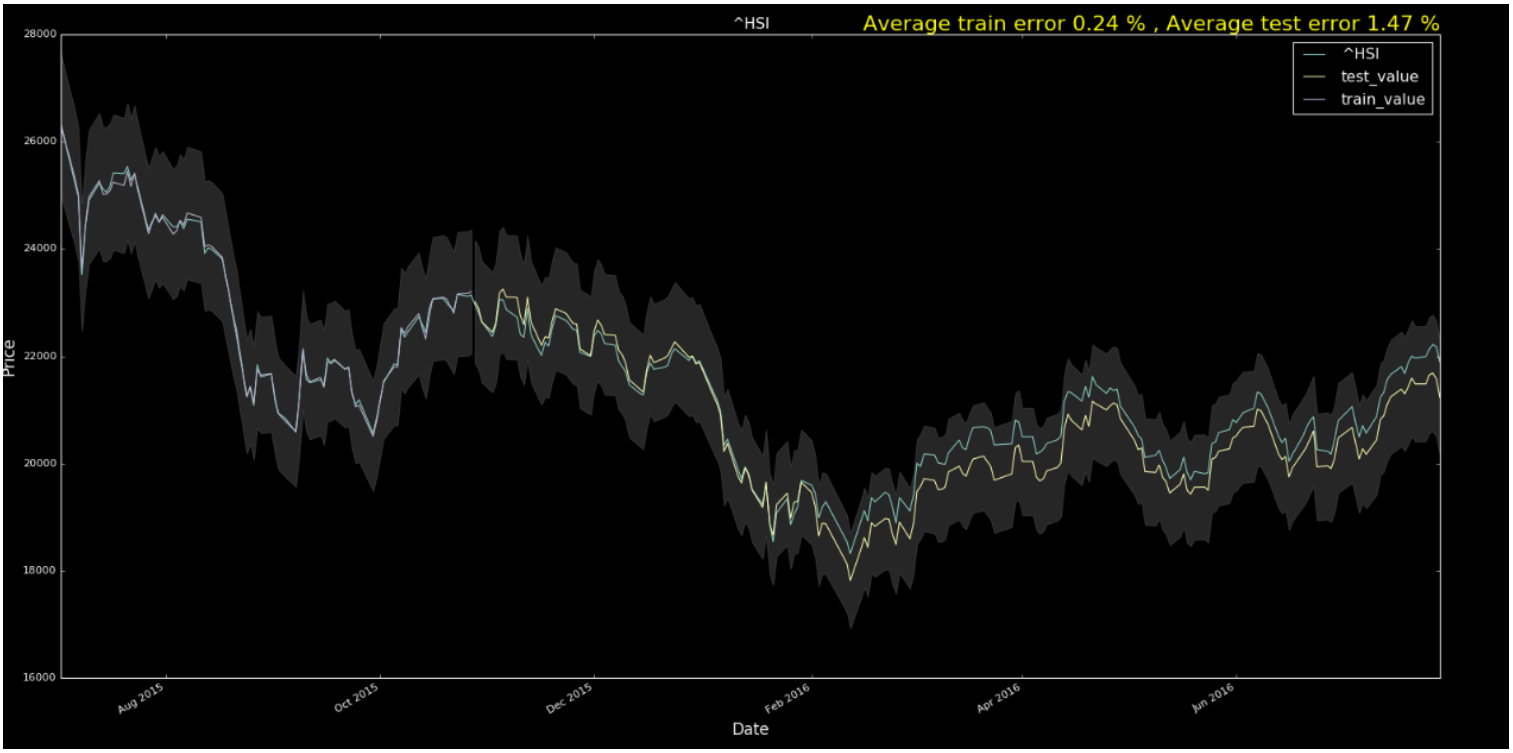
In our example, the data is starting from 2015-07-01 and the ending date of the data will be 2016-07-31. It is a year daily closing price.

In the training part, we use 30% of the data set as the training set to train the predictor and the rest to be testing set to estimate the predictor.

It is a very high R square, 0.997933120788 in the training set. It means that the program can produce a very good fit estimator here. Finally, we use the estimator to run the testing see to see how it performance in the data that it did not face to. It gives R= 0.903454754554. It seems to be a very good result here.

Conclusion

Rather than looking to R^2 other, it will be more practice to look at the actual percentage error between the prediction and actual value as investors or traders will look at the return much more than anything. To make it simple and general, we accept a +/-5% error. And the graph below will show us how reliable of this program.



The green line in the graph is showing the actual value of Hang Seng index, the blue line is the training value from the predictor, the yellow is the testing value from the predictor and the grey shadow is the +/- 5% Interval from the true value.

The graph here shows us that the predictor is reliable all the time as the error is very low, being closed to the actually value of the Hang Seng index very much.

In finance market, the market price will be overvalued or undervalued to the true value of an asset as we know that there are some over-bought or over-sold that lead to a mispricing problem. By using this program, it is possible to find out the true value of the Hang Seng Index, then, investors is possible to earn profit when the market price Is far away to the predicted value.

Improvement

It will be arguable that the trading chance might not be so many in this program and the index cannot be trade directly.

There are few ways to extend this program in a more powerful way. 1. The program is better to be implemented intraday, using 1 minutes or 5 minutes. In a intraday trading it is more chance to have a mispricing problem, for example there will be a time lagging that the price of AIA, HSCE or China Moble is changed, but the index do not response to it immediately yet, then it will be a high chance to earn profit.

Secondly, we can try to use this program to predict the future rather than index. It means that investor can directly trade to the predicted result. Or investor can try to use it to trade index ETF, both these asset will have a time lagging problem as mentioned in a intraday fast market.

However, these improvement ways need to have a huge intraday dataset that source is limited. Bloomberg professional will be a great platform to provide this data set, however, the cost is not cheap to register it