COMP 4971F – Independent Study (Winter 2016)

PREDICTION OF THE MOVEMENT OF NIKKEI 225 INDEX BASED ON JAPANESE YEN-USD CURRENCY PRICE

By
MUTHUKUMAR, Sivaraam
Year 5, Dual Degree in Technology and Management (MEGBA)
shiv_smk@yahoo.com

23rd January 2017

Supervised by:
Dr David Rossiter
Department of Computer Science and Engineering

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Table of Contents

ABSTRACT

A predictive algorithm was developed to predict the movement of the NIKKEI 225 Index (N225) based on the previous movement of the Japanese Yen-USD currency price (JPY). A machine learning algorithm was setup to identify the necessary conditions to maximize the predictive accuracy of the algorithm in the training datasets. With the maximizing conditions, an accuracy of 57% was achieved in predicting the movement of the N225 using JPY in the testing datasets.

INTRODUCTION

Trading has evolved in the recent years owing to the development of new technologies in trading and information sharing (data, news, etc.). A widely-accepted idea is that the market is inefficient and there exists a first-mover's advantage. This means that there exists a time gap between the identification of a profit-making strategy and prices reflecting the strategy. Hence, the goal of traders is to find a method to capitalize on the inefficiency period of the market and to keep the strategy private. Although a confidential strategy can generate consistent income, there is always a need to develop additional strategies to improve on the returns and to better cope with the complexity of the market's future movement. To develop a trading strategy, there is a need to develop a prediction model with which trading can be performed. In this project, a predictive algorithm is presented, while a trading algorithm will not be included; more on the remark section of the report.

For this project, the target is the Japanese Market. To trade in this market, two features are considered: NIKKEI 225 Index (N225) and the Japanese Yen-USD currency price (JPY). The N225 trades for five hours each day from 0800 to 1030 hour and from 1130 to 1400 hour. On the contrary, the JPY trades for 24 hours. The reason for choosing the N225 index is that it reflects the state of the Japanese market. JPY is used to predict the N225 because of the high turn-over in the currency foreign exchange market (FX), and the prices in the FX market converges faster due to the large number of frequent transactions. In short, people trade on the FX market, hence it is more reflective of the current trend and is suitable to predict the N225 index. Another reason in using a N225 and JYP pair for trading is because of the significant correlation between the two features. A quick calculation of the correlation yields a 68% correlation between the two and this is sufficient to develop a predictive and trading strategy based on them.

ASSUMPTION

The following are the assumptions made in this project.

1. A correlation of more than 50% is a sufficient for justifying the pairing of the N225 with JPY.
2. The trading hours per day is from 0800 – 1030 hours and from 1130 to 1400 hours only (shadowing the N225 index while JPY is 24 hours)
3. To predict the second half of the day (1130 to 1400 hour), the model is built based on the first half of the same day (0800 to 1030 hour). To predict the first half of the day, the model is built on the second half of the previous day.
4. Mean-reversion is present; values fluctuate over a constant mean.

PROCESS FLOW

The first step in this process is to acquire the 1-minute interval data for the N225 Index and the JPY from data source providers such as Thompson Reuters or Google Finance. Once the data had been acquired, a data pre-processing had to be performed on it to ready the data for the next step. The second step is to identify if there exists a correlation between the said entities. In regards to the N225 index and the JPY, a 68% correlation is present and is sufficient to proceed with the pairing. The third step is to split the data into multiple datasets to test the algorithm for repeatability and performance analysis. The datasets are called training sets and testing sets. The algorithm uses the training set to build the model, and the testing set is used to test the accuracy of the model based on the training set. The fourth and final step is to build a predictive model using machine learning to identify the conditions that generate the maximum predictive accuracy in the training set. The model is then fitted to the testing set and the predictive accuracy is measured.

Figure 1: Process Flow of the project

DATA PRE-PROCESSING

The data were retrieved from Thomson Reuters and Google Finance as 1-minute interval data from the 26th of December 2016 to 6th of January 2017. The trading dates in this period were 26th, 27th, 28th, 29th, and 30th December 2016, and 4th, 5th, and 6th January 2017. The N225 index trades for 5 hours a day while JPY trades for 24 hours. From this it was evident that the bottleneck in the quantity of data was the N225 index. To get a working dataset, the data of JPY that matches the timestamp of the N225 index were taken into a new file. Hence for each day, there exists data for N225 and JPY for the trading hours 0800 to 1030 hour and 1130 to 1400 hour. For each whole day, there are 300 data points. For the entire dataset, there are 2,400 data points.

An assumption made in this project is that to predict the first half of the day, a predictive model had to be built using the second half of the previous day, and to predict the second half of the day, the model is built using the first half of the same day. To create a set of training and testing datasets, each half-a-day data was saved in a separate file. This was done to make the retrieval of data more systematic and easy. Once the data had been prepared, the next step in the process flow could begin, finding the correlation between the data. The correlation of the data was found by taking all the available data and running the correlation function in Python. The output of the function is a 2D array with the cross-correlation of the N225 and JPY. From this output, the average correlation was determined to be 68%. This value is acceptable for the pairing of the trading entity. The final stage of the data pre-processing is by computing the logarithmic first difference of the data set.

Plotting the graph of the N225 and JPY will show a sporadic movement and the values of the two are of different order of magnitude. To normalize the values and remove the sporadic nature, the logarithmic first difference $(\ln(JPY_{i+1}) - \ln(JPY_i))$ and $(\ln(N225_{j+1}) - \ln(N225_j))$ was taken. This converts the dataset to a series that fluctuates along a constant mean and with a relatively constant deviation. The constant mean is very close to zero in both the cases and the standard deviation are comparable. These normalized values were then put into another array Array_JPY and Array_N225 respectively. A negative logarithmic difference value indicates a decrease in the price whereas a positive logarithmic difference value indicates an increase. With this step, the development of the predictive algorithm could begin.

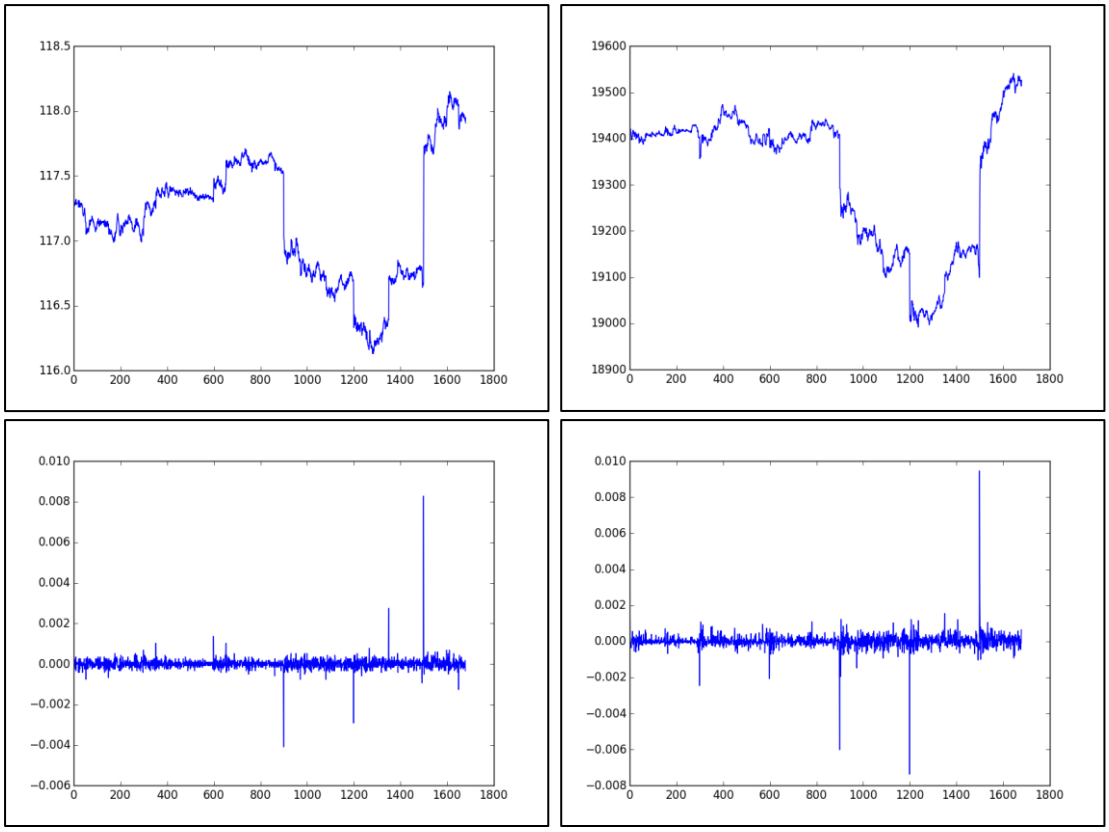

Figure 2: Plots of the actual JPYUSD currency price (top left), actual NIKKEI 225 Index (top right), and logarithmic difference of JPYUSD currency price (bottom left) and NIKKEI 225 Index (bottom right)

PREDICTIVE ALGORITHM

The predictive algorithm is a machine learning program that tries to maximize the desired output of the algorithm by looping through all the possible values for the variables. The algorithm uses the training set to find the maximizing conditions and then applies these conditions on the testing set to calculate the predictive accuracy of the algorithm. The variables used to identify a unique set of values for the given training set are as follows.

1. Lag – The lag variable is used to find after how many steps the movement of the JPY is reflected by the N225 index. For example, a lag of 1 indicates that the current timestamp's movement can help predict the next timestamp's movement. A lag of 3 indicates that the current timestamp's movement can help predict the movement 3 timestamps later.
2. Directional multipliers – The directional multipliers are to either maintain the current direction or reverse the direction. The possible values of this variable are 1, 0, -1, corresponding to maintain, not effect, reverse respectively.
3. Directional output variables– The directional output variables indicate the direction of the predicted movement. The possible values for this output are 1, 0, -1 which corresponds to increase, constant, and decreases respectively.

The predictive testing condition is based on the concept of mean-reversion. This mean that the values tend to fluctuate around the mean value. A value below the mean tend to increase and likewise a value above tends to decrease in the next step. To calculate the mean value for the mean-reversion, an average of the means of Array_JPY and Array_N225 were taken. The reasoning behind this is that the mean values of both these arrays are comparable. Since the logarithmic difference is also comparable to the mean values, a mean that is the average of both the arrays was deemed fit.

$$If\ [(X_i * s_1) + (Y_i * s_2)] \leq AVERAGE(Mean_X, Mean_Y)\ then \qquad (Eq.\ Set\ 1)$$
$$PredictionArray_{i+lag} = s_3$$
$$else$$
$$PredictionArray_{i+lag} = s_4$$

Where,

1. X_i is the i[th] value of the Array_JPY and i ranges from 0 to size(X)
2. Y_i is the i[th] value of the Array_N225 and i ranges from 0 to size(Y)
3. $Mean_X$ & $Mean_Y$ are the mean of the Array_JPY and Array_N225 respectively
4. lag is the lag variable mentioned above
5. s_1 & s_2 are the directional variables mentioned above. For example, $s_1 = 1$, $s_2 = -1$ means that check the mean-reversion based condition using the sum of the X_i in the same direction and Y_i in the opposite direction.
6. s_3 & s_4 are the directional output variables. For example, $s_3 = 1,\ s_4 = -1$ means that if the condition is true, the predictive array at the $(i + lag)$ position will be $+1$ (predicted going up) else, the predictive array at the position will be -1 (predicted going down).

The reasoning behind the use of directional multipliers was to allow the algorithm to decide which of the variables (JPY and N225) are contributing to the predictions. The use of the directional output variables was to allow the algorithm to decide in which direction will the predictions of the N225 be if the above Eq. Set 1 is true. Since there are two directional multipliers (s_1, s_2), two directional output variables (s_3, s_4), and three possible values (1, 0, -1) each, there are in total 3^4 (= 81) possible combinations for the directional multipliers and variables combined. For a reasonable consideration of the possible lag values, lags from 1 to 10 are considered in this project. Lags greater than 10 are considered very delayed and is therefore insignificant in the predictions of the N225 future movement. In the predictive algorithm, an outer loop is set to loop through the 81 possible combinations of the directional variables and an inner loop is set to loop through the 10 lag values. Inside the inner loop, the predictive testing condition is performed.

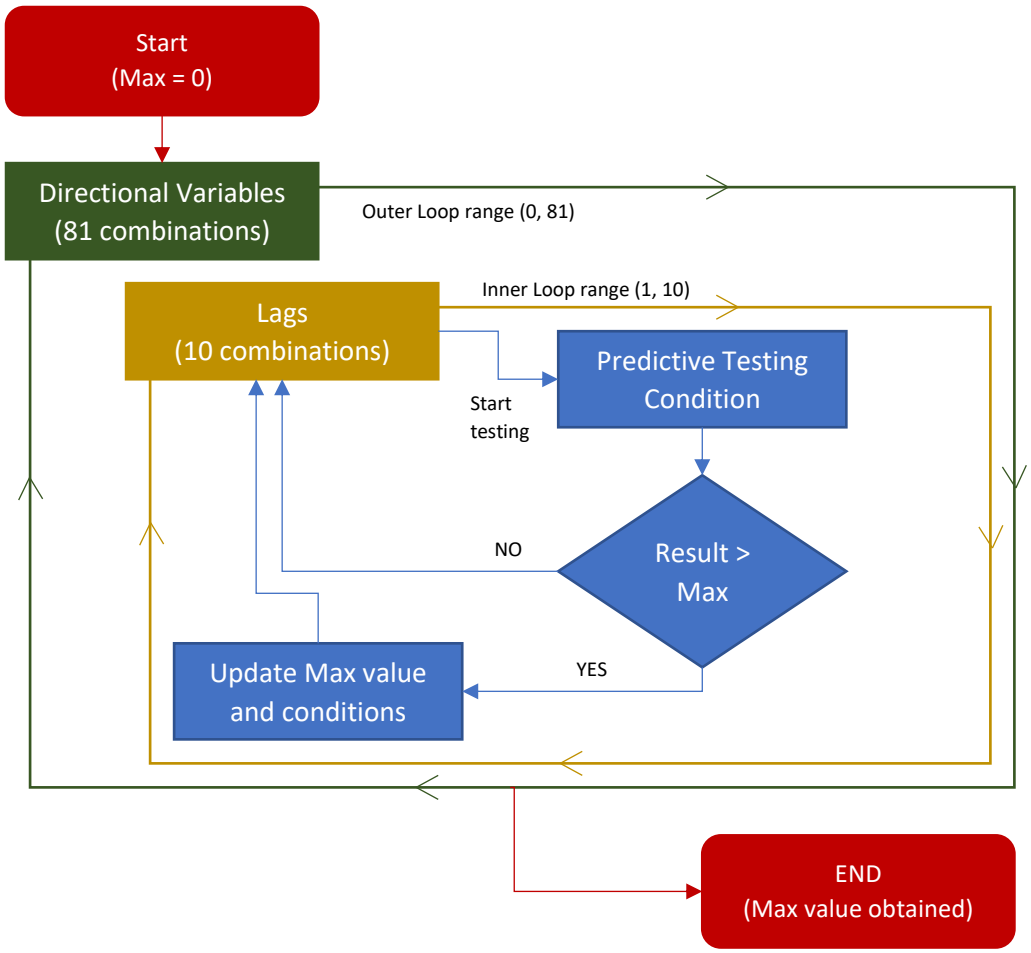

Figure 3: Flow chart of the steps involved in the predictive algorithm

Once the predictive array is generated, it is tested against the Array_N225 of the training set to measure the predictive level. The predictive level is measure by a simple custom function. It checks if the direction predicted by the predictive algorithm matches with the direction of the Array_N225 of the training set.

$$If\ predictiveArray_{i+lag} = Y_{i+lag}\ then \qquad (Eq.\ Set\ 2)$$
$$Var_{correct} = Var_{correct} + 1$$
$$else$$
$$Var_{incorrect} = Var_{incorrect} + 1$$

Where,

1. $predictiveArray_{i+lag}$ is the outcome of the Predictive Algorithm.
2. Y_{i+lag} is the Array_N225 of the training or testing set starting at the same index as the predictive array.
3. $Var_{correct}$ is the variable to count the number of correct predictions
4. $Var_{incorrect}$ is the variable to count the number of incorrect predictions

The looping of the array starts from the lag value. If the predictive array's direction (1, 0, -1) matches with the direction of the Array_N225 then the $Var_{correct}$ is

incremented by one, if not, then the $Var_{incorrect}$ is incremented by one. The performance of the algorithm, predictive level, is calculated by finding the percentage of $Var_{correct}$ in the total number of iterations.

This predictive level is maximized under the two loops, and its lag and directional variables are recorded at every maximum. The maximum value that remains at the end of the two loops are the maximizing conditions of the predictive algorithm. The maximizing conditions of the predictive algorithm are then put back into the model with the testing dataset, and the performance of the algorithm in the testing set is recorded. The result is then saved as a text file and is imported in Microsoft Excel for result analysis.

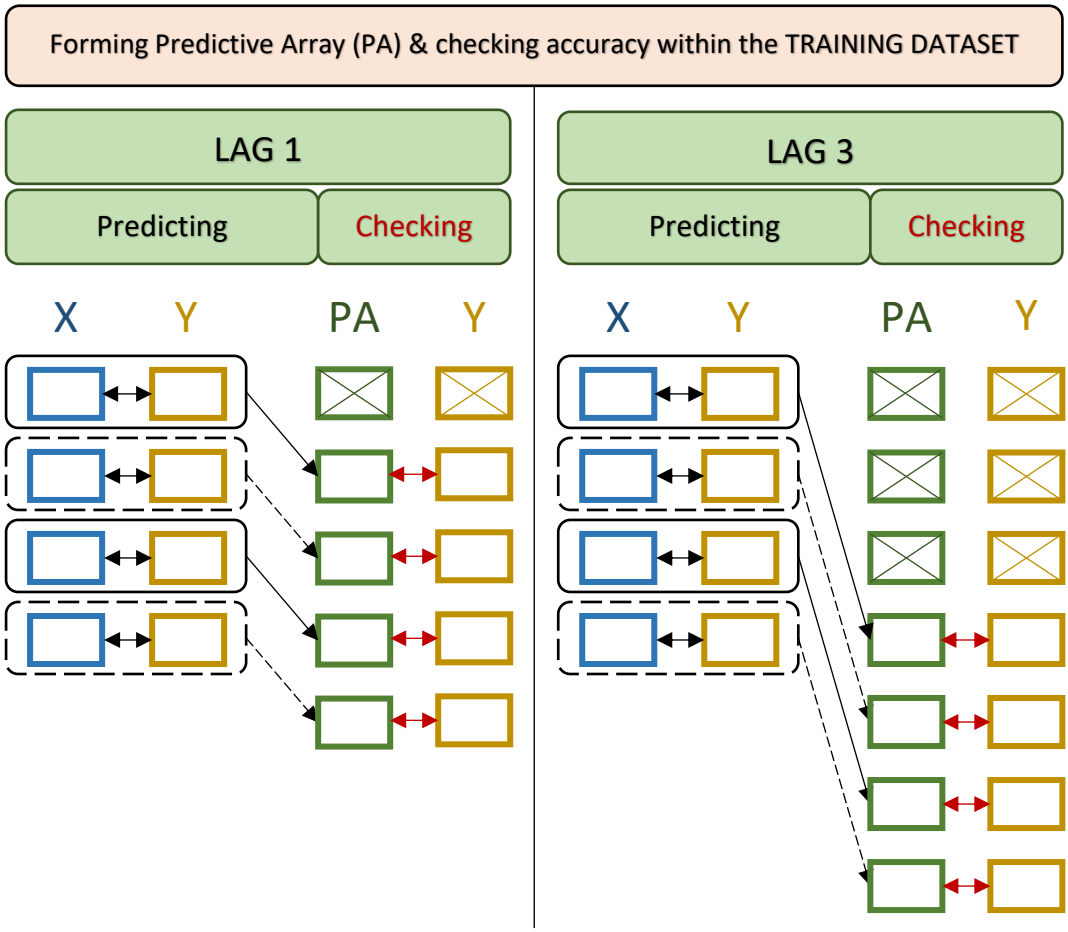

Figure 4: Mechanism behind the predictive algorithm in creating the Prediction Array and the testing method. Shown are the examples for Lag 1 and Lag 3.

RESULTS

The results imported in to Microsoft Excel are summarized below.

Table 1: Complete results table generated from the Python program and exported to excel. The result table shows the Dataset #, Maximum Prediction in training set, Prediction accuracy in testing set, lag, directional multipliers and directional output variables

Dataset #	Train Set Maximum Prediction	Test Set Prediction Accuracy	Lag	s_1	s_2	s_3	s_4
1	60.14%	44.76%	6	-1	-1	-1	1
2	60.54%	50.34%	2	-1	1	1	-1
3	63.57%	42.86%	9	1	1	1	-1
4	62.16%	54.73%	1	-1	0	1	-1
5	62.33%	52.74%	3	1	1	1	-1
6	60.14%	56.08%	1	1	1	-1	1
7	62.84%	62.84%	1	-1	0	1	-1
8	62.84%	63.51%	1	-1	0	1	-1
9	64.19%	56.76%	1	1	0	-1	1
10	59.46%	66.89%	1	-1	0	1	-1
11	66.89%	55.41%	1	-1	0	1	-1
12	60.42%	59.72%	5	0	1	1	-1
13	64.86%	62.84%	1	-1	0	1	-1
14	63.51%	66.22%	1	-1	0	1	-1
15	67.57%	57.43%	1	1	0	-1	1

From the above table the following summaries can be calculated.

Table 2: Summarizing the results table for Maximum Prediction and Prediction Accuracy

	Train Set Maximum Prediction	Test Set Prediction Accuracy
Simple Average	62.76%	56.88%
Win % (>50% coin-toss)	100.00%	86.67%

From the above table, we can see that the Predictive Algorithm can predict the future movement of the N225 ~57% of the time using the previous movements of JPY and N225. We can also see that the average in the training set is ~63% which is significantly higher than the testing set. This is because the algorithm is set to maximize this value. Since the future movement of the N225 should ideally be around 50% because of unavailable data, a prediction accuracy of ~57% is significant. This higher predictive ability could be due to the lag present in the movement of the JPY and the N225. To compare how much better is the predictive algorithm compared to the 50% predictive coin-toss, a win percentage is computed. The training set achieved a 100% in maximizing the conditions so that the prediction is better than the coin-toss. The testing set achieved 87% in predicting better than the coin-toss.

Table 3: Summarizing the Lag, Directional multipliers and directional output variables

	Lag	s_1	s_2	s_3	s_4
Mode	1	-1	0	1	-1
Percentage	66.70%	60.00%	60.00%	73.33%	73.33%

From the above table, we can see that the most frequently occurring value (mode) and the percentage of occurrence of the lag, directional multipliers (s_1, s_2), and directional output variable (s_3, s_4). Results indicate that the lag 1 is better at predicting the movement of the N225 using the JPY. This means that the current timestamp's movement of JPY can be used to predict the next timestamp's movement of N225. The mode of s_1 and s_2 are -1 and 0 respectively. This indicates that 60% of the time, the opposite direction of the JPY, alone, is sufficient to predict the next movement of N225 (0 indicates that 60% of the time, the effects of N225 does not influence the its next movement). The mode of s_3 and s_4 are 1 and -1 respectively. This indicates that 73% of the time, the value of the N225 undergoes mean-reversion. When the values is less than the mean-reversion (if condition is true), the next predicted movement is up (increases). When the value is more than the mean-reversion (if condition is false), the next predicted movement is down (decreases).

CONCLUSION

The Japanese Yen-USD currency can be used to predict the movement of the NIKKEI 225 Index with a 57% accuracy using a machine learning algorithm. The output indicates that the conditions opted from the predictive algorithm resembles the mean-reversion principle where the values fluctuate over the mean value. The predictive algorithm produces an accuracy that is better than the 50% probability of tossing a coin in making a future prediction. Such a 57% accuracy indicate that the future movement of N225 is partly influenced by the historic movement of the JPY and N225.

REMARKS

The following are the remarks of this project.

1. This project compares the performance with the 50% probability of a coin toss. In future study of such a nature, other predictive models can be used as a benchmark to compare the performance of the predictive algorithm.
2. A possible method to conclude the predictive algorithm is to develop a trading strategy. A trading strategy was developed to trade on the predictions made using the predictive algorithm. Due to the time constraint of the project window and the inaccessibility of the data, the trading section of the original project has been omitted in this report.