Forecasting Share Prices for Companies of a Country Using Companies of another Country

Sanjay Verma1 and Ujjawal Panchal2

1Professor, Indian Institute of Management, Ahmedabad

2B. Tech. Student, SRM Institute of Science and Technology, Kattankulathur

Email :1. sverma@iima.ac.in 2. ujjawalpanchal32@gmail.com

**Abstract**

*Stock price performance of companies is of major interest in both financial and academic studies. Performance of companies of a country in share market might indicate to some degree, economic growth of that country. Due to international relations between local companies and multinational corporations and many other international factors, the share market performance of a county’s companies may be dependent on the international factor to an extent. To anlayse the forecasting of share prices of a country, we have collected technical data for 66 companies of India, 67 companies of Japan and 67 companies of United States of America. In this paper, we have implanted random forest algorithm and it is trained on the training-set treating 1 company of India as the criterion variable at a time and all selected companies of Japan and United States of America as the predictor variables. We observed that if the dependence exists, and if it does, how strong it is, by computing the errors such as average of mean squared error, root mean squared error and mean of absolute error.*

***Keywords: Share market forecasting, Machine Learning, Random Forest, Stock Prices, Data Analysis***

1. **Introduction**

There are two main schools of thought in the financial markets, technical analysis and fundamental analysis. Fundamental analysis attempts to determine a stock’s value by focusing on underlying factors that affect a company’s actual business and its prospects. Fundamental analysis can be performed on industries or the economy. Technical analysis, on the other hand, looks at the price movement of a stock and uses this data to predict its future price movements. [1]

Data collected from source (listed in the data section), contains indicators such as Date, Open, High, Low, Close and Volume. For consistency, we perform analysis on High, Close and OHLC average (average of Open High Low and Close) of the technical data collected. The formula used for calculating OHLC average is shown below.

*OHLC Average = (Open + High + Low + Close) / 4* (1.1)

The default setting for many indicators is to use the close of the time frame as the input data. Changing this to the open, the high or low can dramatically affect how the indicator moves and the analytical insight it provides. The open, high, low and close average (OHLC average) is the average of all these settings combined. [2]

For analysis, we have implemented Random Forest Regression. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.[3][4]

The dataset is randomly shuffled and divided into two separate sub-sets, training-set and test-set. Training set consists of 80% of the dataset, Test set consists of remaining 20% of the dataset. Both training and test sets consists of α[assumption 1] attribute of each of the company listed in sequential fashion. All the India’s 66 companies are taken as criterion variables one by one, for each of these companies; a Random Forest Regression model is trained. All other variables of Japan and US are simultaneously provided to the model as predictor variables. For training, the training-set is used. For all these 67 companies, after training of the model, predictions are done on the test-set.

The predicted values for the criterion value are compared with the actual values. Now, using these 2, Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) is calculated. β[assumption 2] is calculated for all the 66 different models. γ[assumption 3] is calculated. If γ is very small as compared to 50, there is a strong relation between the share market performance of the companies of the countries in question as our model was able to predict criterion variable values that were very close to the actual criterion variable values of the test set on the basis of the values of the predictor variables.

1. **Data Collection and Processing**

In this paper, technical data is collected for 66 companies of India, 67 companies of Japan and 67 companies of United States of America is collected from the internet. We chose these companies mainly because of their popularity and due to the reason that they have existed for a long time in the share market (to get more data for better analysis). We have used the website performance.morningstar.com as the source for collecting this data. Daily data is collected from source from date 30th May 2018 till date 24th September 2004 (Sequentially).

The required fields for attribute α[assumption 1] used for analysis for each company is concatenated column wise into their own files alphabetically depending on what indicator, the analysis is being done. This encompasses the final dataset on which analysis will be done.The final dataset’s columns look like the table given in Table 1.1.

**Table 1.1 Columns of the final dataset**

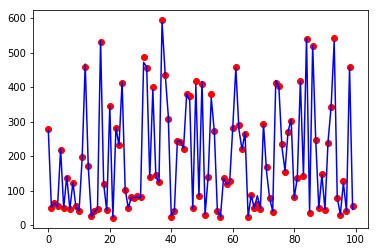
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| India 1 | … | India 66 | Japan 1 | … | Japan 67 | US 1 | … | US 67 |
| # | # | # | # | # | # | # | # | # |
| # | # | # | # | # | # | # | # | # |

In Table 1.1, the symbol ‘…’ indicate continuation of series. For example, India 1… India 3 would denote India 1, India 2, India 3. Symbol ‘#’ denotes values in the rows. The number of rows, columns are not precise and are just meant to give reader a view about the final dataset.

1. **Analysis**

A random forest algorithm is trained on the training-set treating 1 company of India as the criterion variable at a time and all companies of Japan and United States of America as the predictor variables. Now, the trained model is tested against a randomly arranged 20% of the test set. Following plots are given for the companies India 1 and India 66 as Figure 3.1 and Figure 3.2 respectively, where red points are values of criterion variables and blue lines are the models predictions. The x axis represents the index in the test set of the criterion variable (note that indexes are randomized from the dataset and are not a function of time as the dataset is shuffled before splitting.) The y axis represents the value of criterion variable’s α.

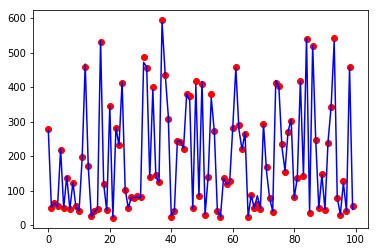
|  |
| --- |
| α of Value of criterion variable |



Index in the test set of the criterion variable

**Figure 3.1 For company India 1**

|  |
| --- |
| α of Value of criterion variable |



Index in the test set of the criterion variable

**Figure 3.2 For company India 66**

1. **Results**

Upon doing the above given experiment, the following results are obtained.

γ mean absolute error = 2.156315 (4.1)

γroot mean squared error= 5.816969 (4.2)

Average of Mean Squared Error = 16710.597253 (4.3)

Assumptions

1. The attribute for which analysis is being done (i.e. OHLC average), is assumed to be α.
2. Mean value of the criterion variable of the test set, is assumed to be β (Individual for all 66 models).
3. For mean absolute errors and root mean squared errors of the 66 models we trained, each value is divided by β and is multiplied by 100. This gives us the percentage part that is mean absolute error of β. Now all these values are averaged. This gives us the average percentage that the error is over all the 66 companies. Let this value be γ.
4. **Conclusion**

Upon conducting the analysis, we get small values for γ mean absolute error along with

γroot mean squared error and average of Mean Squared error. The model predicts α of the criterion variable for all 66 companies of India with satisfactory accuracy using only data from selected 67 companies of Japan and selected 67 companies of United States of America. This implies a strong dependence of the companies of India on the companies of Japan and United States of America.

**References**

1. Selene Yue Xu (UC Berkeley) “Stock Price Forecasting Using Information from Yahoo Finance and Google Trend”
2. Adam Milton “Which price data point is used will affect indicator performance” at the article “Whether to Use the Open, High, Low or Close on Trading Indicators” on thebalance.com updated on 5th February, 2017.
3. Ho, T.K. (1995) Random Decision Forest. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, 14-16 August 1995, 278-282.
4. Ho TK (1998). "The Random Subspace Method for Constructing Decision Forests" (PDF). IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 (8): 832–844.