Stock Forecasting Based on Interrelated Stock indexes

Anonymous

ABSTRACT

The financial market is a valuable resource for people and companies try to augment their returns. The financial market is a market where it is plausible to trade monetary products with low transaction costs. Between the different types of markets, there is one of the most attractive markets, the capital market (48%, of American adults have money in stocks, according to Bankrate's Money Pulse survey). This market is acknowledged to be a high-risk market (derived from it volatility and sensibility) that makes people fortunate or bankrupt. The stock market's (one of the types of a capital market) openness increased its reputation, and consequently the stock market price prediction become an imperative factor, which resulted in many diverse algorithms and routines.The present project proposes a data analysis approach using a particular methodology and an algorithm to reduce the risk on a stock market. The closing index returns of the stock exchange is predicted by using the knowledge acquired from other global stock exchanges that close before the target stock market and other financial constituents that influence the market. In this article, we predict the outcome of S&P stock index by using Non-US stock indexes, inflation, Exchange Trade Funds such as oil indexes, Gold, etc.This paper also presents the comparison of different machine learning algorithms for attaining the best results. Among all the models tried, Random Forest yielded the most precise results.

1. INTRODUCTION

The problem of stock prediction is a classic problem on which there are various models proposed. Predicting the stock empowers an individual to make the right decision whether to sell, buy or to stay put. Using the stock market prediction software has many advantages. Traditionally, people hire financial consultants who analyze the stock trends and predicts the outcome of the stock exchange. However, humans always tend to make mistakes and are prone to overlook various factors that play a crucial role in predicting the stock outcomes. Moreover, decisions made by a person may be prejudiced or may be fallacious because of not having access to all the data required to make adequate predictions. Due to developments in the fields of big data and artificial intelligence, there are many models available which perform the task of predicting the stocks eliminating the need for financial advisers. A stock market is quite volatile and is arduous to predict. Stock prices depend on numerous factors such as share prices of other stocks, oil prices, pandemonium in any country, investor sentiment, inflation, deflation, political events, changes in economic policy and many other. Since last decade, many models were proposed to predict the stocks prices. However, most of the models do not consider most of the above said factors. In the next three segments, we will discuss background work done on stock prediction techniques, approach to the proposed model and results of the model proposed. The last two sections address the future work and conclude the paper respectively.

2. BACKGROUND

As stock prediction is a classical problem. Many models introduced were able to predict the stock indexes precisely. However, the intricacy of predicting the stock market remains as one of the challenging tasks given the volatility of the stock exchange and various determinants to be considered. Many proposed models forecast the stock index outcome just by trend analysis of the different Stock indexes. These models demonstrated the effectiveness if other significant factors incorporated while training the model. In many of the prior works related to stock predictions, interrelations between the stocks markets were not taken into attention. In the paper by Shen et al. [7], the model, predicts the direction of the stock exchange, by leveraging the external stock exchange indexes of Japan, China, Germany, etc. Data used in the model comprises of Exchange Trade Fund indexes such as oil price, gold, silver, platinum ratio and also the exchange rates of Europe, Australia, Japan. This model did a decent job in predicting the outcome of the stock index. Nevertheless, this model do not consider the stock exchanges of the top financially sound countries and other economic factors like inflation, deflation, etc. As a result, this model does not acknowledge the accurate features for the model to be effective. In real world scenario, stock prices are significantly impacted by these external factors stocks. To address the shortcomings identified, along with the features recommended by Shen et al. [7], other factors are supplemented. The supplementary features include CAC index (a French stock market index), ESTX 50, Eu-

Table 1: Sample Stock data

NIK	INF	CAC	ENXT	Hang	spasx	STI	SP
0.67	1.43	0.95	0.79	1.06	0.43	0.11	Bull
-0.44	0.07	0.02	0.09	-0.45	0.43	0.08	Bear
-0.67	-1.53	-1.28	-1.27	-0.42	-0.59	-0.43	Bear
-1.57	-1.49	-2.69	-1.81	-1.29	0.34	-0.95	Bull

ronext (European stock indexes), S/P/ASX (an Australian stock index), STI (Straits Ties Index a Singapore stock index) and inflation.

3. APPROACH

3.1 Data Collection and Preprocessing

The global stock indexes are accumulated from various sources such as YAHOO Finance [1] , investing.com [2] and research.stlouisfed.org [3]. As the stock market indexes are more influenced by the recent data than the old historical data, the data collected is for a span of two years with daily frequency. The Data consolidated for this analysis includes features such as Nikkei, Hang Seng, SP, EUR/USD, GBP/USD, AUD/USD, YUAN/USD, Crude oil, Brent, Gold, DAX, CAC, STI, SSE, S&P/ASX, ESTX, EURONEXT and inflation. The data with the class attribute i.e. SPchange categorized with a class level of two (Bullish / Bearish) serves an input to the classification model.

The data consolidated undergo preprocessing to achieve compatibility with the current model. Several observations are discarded to limit the effect of missing values. Various datasets are assembled using inner joins to eliminate the missing values. The data is then normalized to avoid the ambiguous results induced by the presence of outliers. A sample data can seen in Table 1.

Not all the features in the data collected are effective in the predictive process of the class variable. So, a correlation of all the gathered data is plotted. Input for the training process includes only a subset of the strongly correlated features to the class attribute i.e. SPChange. As shown in Figure 1, only Nikkei, inflation, CAC, DAX, ESTX, EURONEXT, HANG SENG, S&P/ASX, STI are chosen for the model training.

4. MODEL TRAINING

4.1 Classification

Data prepared is fed into different classification models such as SVM, Decision Trees, Ada Boost, Random Forests all with default parameters. Date attribute is ignored during the training phase since it is unique for each observation that doesn't play any role in the prediction step.

SVM with default parameters results in an accuracy of 68.9%. SVM with polydot or vanilladot increases the efficiency to 73.8%.

4.2 Auto-Regression analysis

As an alternative model, the auto-regression analysis was made. To do that, the ARIMA model was used in three different ways, an ARMA model, an ARIMA model and an ARIMA model using exogenous variables. As a methodology to identify the best parameters of our models, it was used the follow steps:

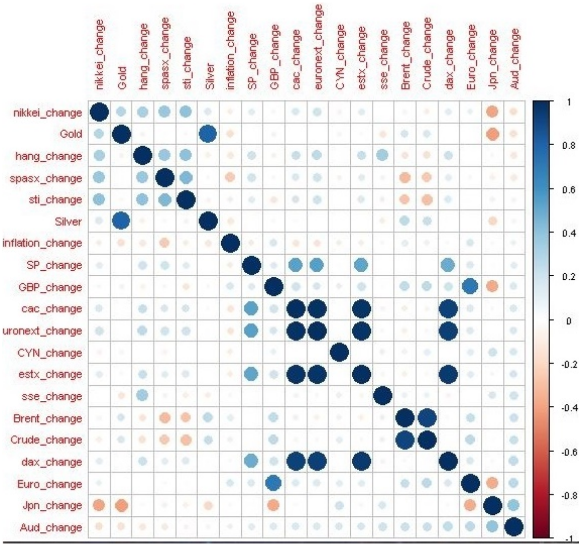
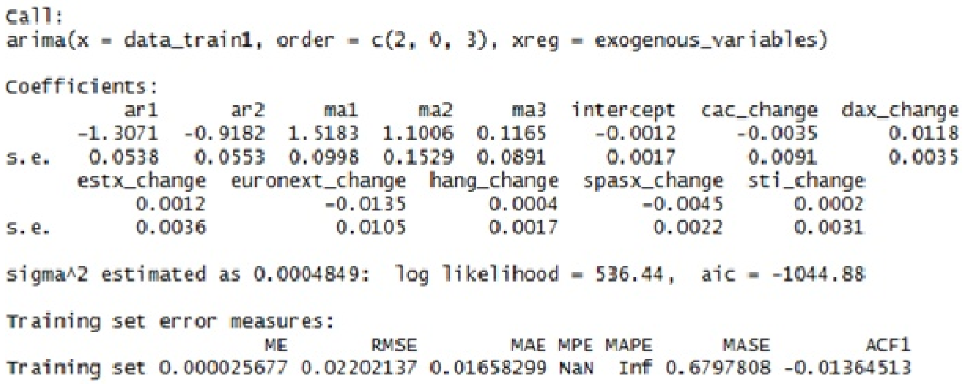


Figure 1: Correlation of various features

- Visualization analysis of the target attribute to determine if the data is stationary, non-stationary or trend stationary;
- If nonstationary or trend stationary takes the actions to stationary it;
- Dick-Fuller Test to confirm the stationarity of the target attribute;
- Identify the best auto-regression (AR), integration(I) and moving average (MA) orders;
- Divide the dataset on test and training set (90/10);
- Test the model without and with exogenous variables.

The variables used on the models with exogenous variables were `CAC_change` and a combination of `CAC_change`, **`dax_change`**, `estx_change`, `euronext_change`,`hang_change`, **`spa_change`** and `sti_change`. Also, the experimentations were made with different datasets. The first analysis were made with a small dataset of only 172 observations (daily observations). This analysis was our first test to check if the auto-regression would be a good alternative for the problem. After the confirmation of the auto-regression as a right approach, the second analysis were made with a dataset correspondent to a role stock price year. This analysis was chosen to amplify the accuracy of the model. The third study was made using two stock price years. This study was made to try to remove the seasonality of the series of the model. The approach that showed the best RMSE results were the ARIMA(2,0,3) with the combination of `CAC_change`, **`dax_change`**, `estx_change`, `euronext_change`, `hang_change`,**`spasx_change`** and `STI_change` as exogenous variables using the one-year dataset. The model showed an RMSE of 0.0.22 as presented in Figure 2. considering the target attribute variation of -4.13.



```
Call:
arima(x = data_train1, order = c(2, 0, 3), xreg = exogenous_variables)

Coefficients:
          ar1      ar2     ma1     ma2     ma3  intercept  cac_change  dax_change
      -1.3071  -0.9182  1.5183  1.1006  0.1165    -0.0012     -0.0035      0.0118
s.e.   0.0538   0.0553  0.0998  0.1529  0.0891     0.0017      0.0091      0.0035
      estx_change  euronext_change  hang_change  spasx_change  sti_change
           0.0012          -0.0135       0.0004       -0.0045      0.0002
s.e.       0.0036           0.0105       0.0017        0.0022      0.0031

sigma^2 estimated as 0.0004849:  log likelihood = 536.44,  aic = -1044.88

Training set error measures:
                      ME       RMSE        MAE MPE MAPE      MASE        ACF1
Training set 0.000025677 0.02202137 0.01658299 NaN  Inf 0.6797808 -0.01364513
```

Figure 2: Precision of ARIMA

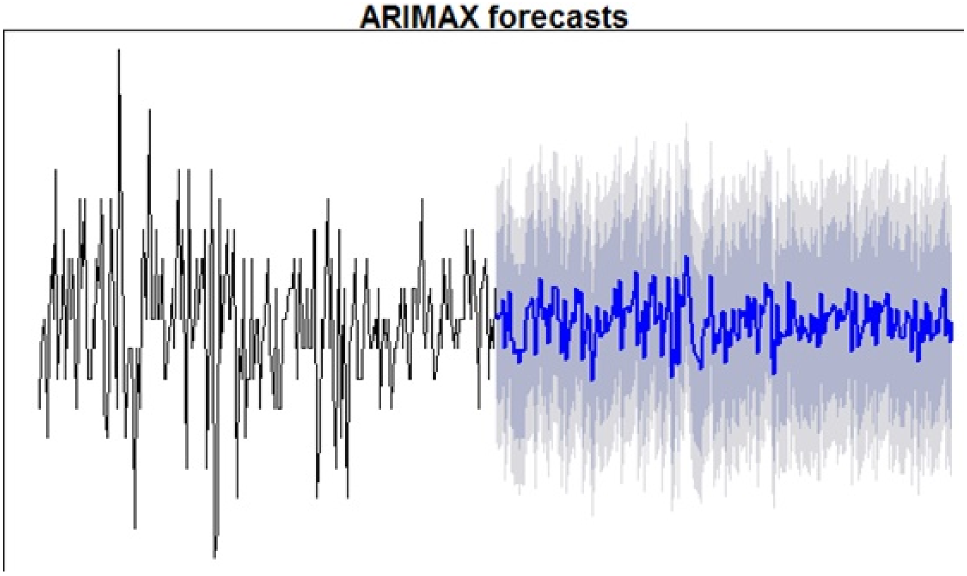


Figure 3: Forecasting results of ARIMA

As shown in Figure 4. For the Auto Regression analysis the ARIMA(2,0,3) with the combination of ”`CAC_change`”, ”`dax_change`”, ”`estx_change`”, ”`euronext_change`”, ”`hang_change`”, ”`spasx_change`” and ”`sti_change`” as exogenous variables using the 1 year dataset presented the best performance followed by the ARIMA(2,0,3) without exogenous variables.

5. RESULTS

As shown in Figure 2. Random Forests performed better than the other classification models followed by Boost Ada and Linear models. Figure 3. Displays the precision chart for Random Forest. The Area Under Curve (AUC) for Random Forest is around 82%.The Figure 3. presents the h= 25 prediction of the RMSE model described in the regression analysis portion. The dark and light blue zones represent 99% and 95% of confidence respectively.

6. OTHER RELATED WORKS

We have already discussed the past work on stock market prediction using the stock market indices. In this section other related works on stock market prediction. Some of these works predict the prices of a particular stock using the interrelated data, news article data, and Twitter data. The problem of stock market prediction has been studied for a long time, and there are a lot of works on the stock market prediction. A work by Gupta et. al.[5] uses the Hidden Markov model to predict the future price of a stock based on the historical data. This work uses the fractional changes of the assets as the attributes to predict stock prices [5].Kato et. al[6] used the interrelated time series data for stock market prediction. They propose a method to find the interrelations between different stocks and stock market indices to predict the future index price for a particular stock [6]. The interrelationships are identified by using evolution strategy on the variation patterns of the attributes[6].

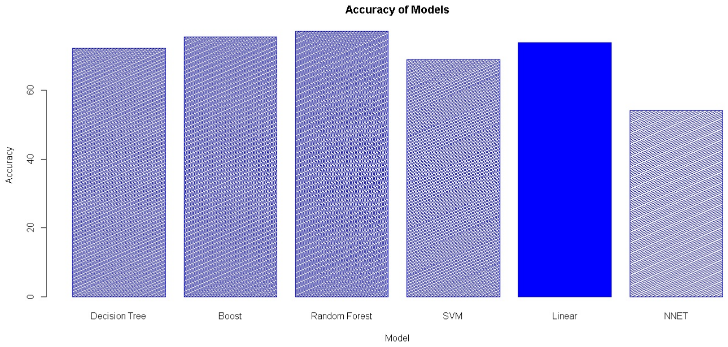


Figure 4: Accuracy of various classification models.

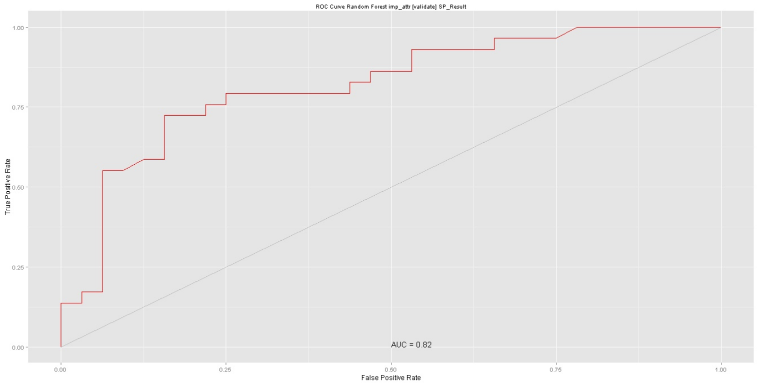


Figure 5: ROC Curve for Random Forest

Another work by Kato et. al. uses the sentiment analysis from the news articles in addition to the interrelated data to prediction[4].

7. CONCLUSION

In this paper, we introduced a model to predict the nature of stock market indices. Firstly, we have examined the correlation between various stock indices and we used the indices that affect the class variable(SP 500 in this case). Then, we have taken the indices that have the maximum affect our class variable and performed data analysis on these variables using Autoregression Analysis, SVM, Random forests, decision trees, Boost ADA. The experimental results show that random forests give the best results.

In the future, the presented work can be improved by analyzing the forecasts made and carry out the proper adjustments to the models. Also, tests with bigger datasets could be done to try to understand a little more about the seasonality of the stock prices. Another interesting work would be to test the models on different markets to identify the power of it and it singularities. Also, the next step of this work would be to try to forecast exactly the amounts of the changes on the prices. This work would help the financial analysts to identify the better profits among the stocks.

8. REFERENCES

[1] `http:finance.yahoo.com//`.

[2] `http://www.investing.com///`.

[3] `https://research.stlouisfed.org///`.

[4] K. Daigo and N. Tomoharu. Stock prediction using multiple time series of stock prices and news articles. In *Computers Informatics (ISCI), 2012 IEEE Symposium on*, pages 11–16, March 2012.

[5] A. Gupta and B. Dhingra. Stock market prediction using hidden markov models. In *Engineering and Systems (SCES), 2012 Students Conference on*, pages 1–4, March 2012.

[6] K. Ryota and N. Tomoharu. Stock market prediction based on interrelated time series data. In *Computers Informatics (ISCI), 2012 IEEE Symposium on*, pages 17–21, March 2012.

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