Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 194 (2021) 173–179

Procedia Computer Science

www.elsevier.com/locate/procedia

18th International Learning & Technology Conference 2021

Stock Market Prediction Using Machine Learning

Abdulhamit Subasi, Faria Amir, Kholoud Bagedo, Asmaa Shams, Akila Sarirete

Department of Computer Science, College of Engineering, Effat University, Jeddah, 21478, Saudi Arabia.

E-mail: absubasi@effatuniversity.edu.sa

Abstract

Due to the complex nature of stock market prediction, it has been a trending area of interest. This paper presents a comparison of the prediction by inputting different classifiers. Furthermore, the results of the comparison are done on an accuracy basis. Each machine learning algorithm is tested against the National Association of Securities Dealers Automated Quotations System (NASDAQ), New York Stock Exchange (NYSE), Nikkei, and Financial Times Stock Exchange (FTSE). Furthermore, several machine learning algorithms are compared with a normal and a leaked data set.

Keywords: Stock market, Machine learning, NYSE, FTSE, NASDAQ, Nikkei.

1. Introduction

Predicting the trends in stock market expenses is a completely difficult task due to many uncertainties involved and lots of variables that affect the marketplace value on a particular day including economic conditions, investors' sentiments towards a specific company, political activities, etc. Because of this, inventory markets are at risk of brief changes, inflicting random fluctuations inside the stock rate. Stock market collections are commonly dynamic, non-parametric, chaotic, and noisy. Hence, the stock market rate motion is considered to be a random method. Among the principal methodologies used to predict stock market prices are: 1) Technical Analysis, 2) Time-Series Forecasting, 3) Machine Learning and Data Mining and 4) Modelling and Predicting Volatility of stocks (Khaidem et al., 2016). The methodology that is discussed in this paper is Machine Learning and Data Mining applications in stock market. This paper aims at proposing a comparison of the prediction by utilizing different classifiers. Furthermore, the results of the comparison are done on different performance measures. Moreover, this paper makes a comparison between the normal and the leaked data set.

2. Literature Review

Chong et al. (Chong et al., 2017) examined the predictability of three different machine learning methods: principal component analysis, autoencoder, and the restricted Boltzmann machine. They used high frequency lagged stock returns as input data. They applied it to the Korean stock market and found that the DNNS performs better than the linear autoregressive model in the training set, while the regressive model does better in the test set. This discrepancy was explained by pointing out that the predicted part of a stock return is more influenced by the first three lagged returns of itself rather than the lagged returns of other stocks. Fischer and Krauss (Fischer & Krauss, 2018) focused on the performance of the short-term memory (LSTM) networks. They found that LSTM performance exceeds the

1877-0509
10.1016/j.procs.2021.10.071

methods of memory-free classification like the random forest, deep neural network, and the logistic regression classifier. However, the RAF was better performing in one case, during the global financial crisis. On the other hand, Ballings et al. (Ballings et al., 2015) found that the best algorithms as ranked from best to worst as follows: Random Forest, Support Vector Machines, AdaBoost, Neural Networks, K-Nearest Neighbors, and Logistic Regression. Similarly, Patel et al., (Patel et al., 2015) compared four prediction models, namely the ANN, SVM, the Random Forest, and the Naive-Bayes, and found that Random Forest showed the best performance among all. Zhong and Enke (Zhong & Enke, 2017b) found that the combining ANN with PCA (principal component analysis) gives more accuracy. In another study (Zhong & Enke, 2017a), they found that the ANNs gave higher classification accuracy compared to logistic regression. Moghaddam et al. (Moghaddam et al., 2016) investigated the ability of artificial neural networks (ANN) to predict the daily NASDAQ stock exchange rate. Their methodology used both short-term past stock prices and the day of the week as input. They found that there are no distinct differences between using thepast four days or the past nine days as input. Likewise, Pehlivanli et al. (Pehlivanli et al., 2016) found that predictionsusing a reduced dataset yield better results than predictions using a full dataset.

Malagrino et al. (Malagrino et al., 2018) used Bayesian Networks to see to what extent global indices influence the main index iBOVESPA (Sao Paulo, Brazil). Their model can serve as a basic block to more complex applications. Boyacioglu and Avci (Boyacioglu & Avci, 2010) used adaptive network-based fuzzy inference system ANFIS on the Istanbul Stock Exchange and found the model to have an accuracy success rate of 98.3%. Zhang et al. (Zhang et al., 2016) created a new system altogether. It used a heuristic algorithm that cuts stock data into multiple clips. These clips are classified. While, X. Zhang, Li, and Pan created a new approach named status box method, which classifies stock point into three categories of boxes—each box indicates different stock status. Their results show that the status box method has better classification accuracy and can solve the stock turning points classification problem.

3. Stock Market Data

This study employs different stock market data to be tested by different classifiers, in order to find the best machine learning algorithm for stock market forecasting. Further, our study focuses on comparing four datasets from different stock market indices, which are the National Association of Securities Dealers Automated Quotations System (NASDAQ), New York Stock Exchange (NYSE), Nikkei, and Financial Times Stock Exchange (FTSE). All stock datasets were uploaded from Yahoo Finance daily, for ten years starting from the 24th of March, 2010.

Yahoo Finance is a media that provides stock prices of different companies. Further, the stock price, in general, is determined by the basic rule of economics, supply, and demand. Besides, the market indices are calculated differently for each index type. For example, the FTSE index is calculated is by the total market capitalization of the constituent companies (*What Is the FTSE 100?*, n.d.), while NASDAQ is calculated by taking the total value of the share weights of all the stocks on the exchange, multiplied by each security's closing price (*NASDAQ Composite - Components, Methodology & Criteria for Inclusion*, n.d.). Further, NYSE calculations are based on their free-float market capitalization (Kenton, n.d.), and Nikkei is calculated by dividing the summation of the adjusted prices by the divisor (*Index Information - Nikkei Indexes*, n.d.). The datasets from Yahoo were downloaded in an excel file. The file includes seven columns, date of the stock price, opening and closing stock price, highest and lowest stock price, adjusted closing and the volume. Further, the opening and closing price denote the price at the beginning and end of the day, respectively. Also, the adjusted closing and volume denotes a more accurate reflection of a stock's value, and the number of shares traded, respectively.

Besides, NASDAQ and NYSE are American indices, while FTSE and NIKKEI are British and Japanese indices, respectively. Each index constitutes different companies. For example, the NASDAQ index contains over 3,300 companies, while the FTSE is an index of 100 companies. The oldest index is the NYSE, followed by the NASDAQ, and NIKKEI. Further, the newest index of all four is the FTSE which was founded in 1984 (What is the FTSE 100, n.d.).

4. Classification Methods

Machine learning is programming computers to use examples of data or past data to optimize an output criterion. We have a model specified to some parameters, and learning is done by running a computer program to optimize the

model's parameters using the training data or experience. The model can be predictive in making future forecasts or descriptive to gain new information (Alpaydin, 2014).

4.1 Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) are computational systems that conduct biological neural networks-like tasks. These systems can learn by using examples. ANN uses a collection of nodes called artificial neurons. Like the biological brain, these artificial neurons can send signals to each other. A signal is represented by a real number in the ANN implementation and the output of each neuron is calculated by a non-linear input sum function. The connection between each node is called an edge, and each edge and node have a weight associated with it. If the weight decreases, this means that the signal strength has improved. Usually, these neurons are placed into layers. On the inputs, each layer performs various functions. Signals move from the input layer (first layer) to the output layer (the last layer). Signals can move back and forth through the layers multiple times until reaching an output (Han et al., 2011).

4.2 K-Nearest Neighbor (k-NN)

K-Nearest Neighbor (k-NN) performs classification and regression without using parameters. The input is the k nearest neighboring examples within the feature space. The object is assigned to its class in k-NN classification, which is the most common among its closest k neighbors; the output is the class chosen. The output is the sum of all the values in k-NN regressions of its nearest k neighbors. K-NN is a type of lazy learning because the function is approximated locally. Each neighbor is given weight so that higher weights are given to the closest neighbors and lower weights given to the further (Han et al., 2011).

4.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is used to create classifications and regression analysis by using learning algorithms. SVM is given a training set of examples marked as either of two categories. Based on these examples, SVM can further categorize new examples. These examples are represented by points and are mapped into a space to show the divide between categorizations (Han et al., 2011).

4.4 C4.5 Decision Tree

C4.5 is an algorithm created by Ross Quinlan to construct a Decision Tree. C4.5 is built on Quinlan's earlier ID3 algorithm. C4.5 is a statistical classifier, meaning it is used for classification. Using the training data set, at each node of the tree, C4.5 decides what attribute best splits its set of examples into subsets of different classes. The attribute with the most standardized information gain is the decision-making attribute (Witten, Frank, Hall, & Pal, 2016).

4.5 Random Forests (RF)

Random Forests are used for classification, regression, and the creation of decision trees. RFs correct the habit of overfitting decision trees to the training set. The first algorithm for RF, the random subspace method, was created by Tin Kam Ho (Breiman, 2001).

4.6 Bagging

Bagging, or Bootstrap aggregating, is used to improve the performance of machine learning classification and regression. Bagging generates a new training set based on a given training set. It reduces variation and avoids overfitting. It is usually applied to decision trees but can be used on any other method. Bagging helps with outputting more accurate results for unstable methods (Witten et al., 2016).

4.7 AdaBoost

Boosting is an algorithm that moves weak learners to become a strong learner. Thus, improving machine learning methods by reducing bias and variance. Boosting works sequentially, each tries to train its predecessor. Usually, boosting algorithms work iteratively; the data weights are continuously readjusted. So future weak learners focus on what the previous weak learner made mistakes in (Hall et al., 2011).

AdaBoost, or Adaptive Boosting, was created by Yoav Freund and Robert Schapire. It is used in many methods to improve accuracy. The output of all the weak learners is combined into a weighted sum that becomes the final output. AdaBoost uses the boosting algorithm but it builds on it by combining all the outputs. Separate learners can be weak on their own but together they can prove to be strong learners (Hall et al., 2011).

5. Result and Discussion

In this study, four datasets (NASDAQ, NYSE, NIKKEI, and FTSE) and 7 classifiers (Random Forest, Bagging, AdaBoost, Decision Trees, SVM, K-NN, and ANN) have been used for stock market prediction. The comparison of the classifier constructed in this study with similar systems in the literature is a challenging task due to diversities in the classification techniques, and data extraction methods. All the classifiers used in this study were trained twice, once with normal data and then with leaked data.

5.1. Performance Evaluation Measures

A classification model evaluation aims to obtain a reliable evaluation of the reliability of the approximation of the target definition defined by the model, which is called the predictive performance of the model. Depending on the model's intended implementation, different performance metrics can be used. Given the fact that the model is created based on a training set, which is a usually small subset of the domain, it is its generalization properties that are essential for the approximation quality. For any performance measurement, it is important to distinguish between its value for a given dataset (dataset performance), particularly the training set (training performance), and its expected performance over the entire domain (true performance) (Cichosz, 2014). Cross-validation can be used to estimate or determine the appropriate amount of flexibility for a given statistical method test error. Model evaluation is the method of assessing the output of a model. Model selection is the process of selecting the appropriate level of flexibility for a model. (Alpaydin, 2014). In this study 10-fold cross validation is used.

5.2. Experimental Results

The classifiers ranked from highest to lowest accuracies are as following: Random Forest with leaked data (93%), Bagging with leaked data (93%), AdaBoost with the leaked dataset (82%), Decision Trees with the leaked dataset (79%), ANN with the leaked dataset (75%), K-NN with the leaked dataset (71%), SVM with the leaked dataset (61%), SVM (62%), K-NN (56%), ANN (56%), AdaBoost (54%), Random Forest (53%), Bagging (53%), and Decision Trees (49%). It is obvious from these results that retraining with leaked data results in higher accuracy results.

For the NASDAQ dataset, among all the classifiers with normal data set, SVM has the highest accuracy of 67%. The second highest belongs to Random Forest and KNN of 54%. The lowest accuracy belongs to Adaboost. On the other hand, among all the classifiers with the leaked data set, Bagging has the highest accuracy of 82%. Accordingly, the second-highest accuracy is 79% which belongs to the Random Forest classifier and the lowest accuracy of 61% belongs to SVM and ANN.

For the NYSE dataset, the highest accuracy of 62% belongs to SVM with normal data set and Bagging has the highest accuracy of 83% with the leaked dataset. The second highest for the normal dataset belongs to Adaboost (54%) and the lowest belongs to Decision Tree (46%). For the leaked dataset, the second-highest belongs to Random Forest (82%) and the lowest belongs to SVM (57%).

For the NIKKEI dataset, the highest accuracy of 56% belongs to SVM and KNN with normal data set and Random Forest has the highest accuracy of 85% with the leaked dataset. The second highest (53%) for the normal dataset belongs to Random Forest and Bagging and the lowest belongs to Adaboost (47%). For the leaked dataset, the second-highest belongs to Bagging (84%) and the lowest belongs to SVM (60%).

Again, for FTSE with normal data set, KNN has the highest accuracy of 46% and the lowest is 21% for SVM. For the leaked dataset, Random Forest and Bagging have the highest accuracy (93%) and Adaboost has the second highest of 82%. The least accuracy for the leaked dataset belongs to SVM (46%).

These results show that, compared to reported results in the literature, Random Forest with leaked dataset and Bagging with leaked dataset provides above satisfactory performance. In choosing an appropriate algorithm, performance is the most important concern, along with ease of use and interpretation. It should be noted due to the varieties in the related works in the literature, providing a completely fair and objective comparison is very difficult.

In several studies, other classifiers also have been used for stock market prediction. However, this study indicates that Bagging with the leaked dataset and Random Forest with the leaked dataset have the best performance and higher accuracy for Stock Market Prediction.

Table 1. Classifier Results for Nasdaq.

	Normal Dataset				Leaked Dataset			
Classifier	**Precision**	**Recall**	**F Score**	**Accuracy**	**Precision**	**Recall**	**F Score**	**Accuracy**
Random Fores	0.72	0.53	0.61	0.54	0.77	0.94	0.85	0.79
Bagging	0.71	0.43	0.54	0.5	0.79	0.94	0.86	0.82
AdaBoost	0.76	0.33	0.46	0.48	0.69	0.87	0.77	0.68
Decision Tree	0.66	0.49	0.56	0.49	0.78	0.82	0.8	0.75
SVM	0.67	1	0.8	0.67	0.61	1	0.76	0.61
K-NN	0.68	0.61	0.64	0.54	0.67	0.77	0.71	0.62
ANN	0.66	0.52	0.58	0.5	0.63	0.88	0.73	0.61

Table 2. Classifier Results for NYSE.

	Normal Dataset				Leaked Dataset			
Classifier	**Precision**	**Recall**	**F Score**	**Accuracy**	**Precision**	**Recall**	**F Score**	**Accuracy**
Random Forest	0.61	0.58	0.6	0.52	0.78	0.95	0.86	0.82
Bagging	0.61	0.63	0.62	0.52	0.81	0.92	0.86	0.83
AdaBoost	0.62	0.64	0.63	0.54	0.68	0.77	0.72	0.66
Decision Trees	0.59	0.45	0.51	0.46	0.78	0.78	0.78	0.74
SVM	0.62	1	0.76	0.62	0.57	1	0.73	0.57
K-NN	0.64	0.49	0.55	0.51	0.68	0.75	0.71	0.65
ANN	0.63	0.51	0.57	0.51	0.6	0.8	0.69	0.59

Table 3. Classifier Results for NIKKEI

	Normal Dataset				Leaked Dataset			
Classifier	**Precision**	**Recall**	**F Score**	**Accuracy**	**Precision**	**Recall**	**F Score**	**Accuracy**
Random Forest	0.56	0.71	0.63	0.53	0.85	0.9	0.88	0.85
Bagging	0.57	0.71	0.63	0.53	0.85	0.9	0.87	0.84
AdaBoost	0.53	0.56	0.55	0.47	0.73	0.74	0.74	0.68
Decision Trees	0.54	0.59	0.59	0.49	0.8	0.75	0.77	0.74
SVM	0.56	1	0.72	0.56	0.6	1	0.75	0.6
K-NN	0.59	0.73	0.65	0.56	0.68	0.68	0.68	0.61
ANN	0.56	0.65	0.6	0.51	0.64	0.8	0.71	0.61

Table 4. Classifier Results for FTSE

	Normal Dataset				Leaked Dataset			
Classifier	**Precision**	**Recall**	**F Score**	**Accuracy**	**Precision**	**Recall**	**F Score**	**Accuracy**
Random Forest	0.24	1	0.39	0.32	1	0.87	0.93	0.93
Bagging	0.26	1	0.41	0.39	1	0.87	0.93	0.93
AdaBoost	0.22	0.83	0.34	0.32	0.86	0.8	0.83	0.82
Decision Trees	0.12	0.33	0.18	0.36	0.91	0.67	0.77	0.79
SVM	0.21	1	0.35	0.21	0	0	0	0.46
K-NN	0.29	1	0.44	0.46	0.89	0.53	0.67	0.71
ANN	0.21	0.67	0.32	0.39	0.9	0.6	0.72	0.75

6. Conclusion

Stock market forecasting is a trending topic in the market nowadays. Therefore, our study focuses on comparing seven machine learning algorithms on four different stock indices datasets, NASDAQ, NYSE, NIKKEI, and FTSE, to facilitate the reduction of risk investment. Further, results concluded that Random Forest with leaked dataset and Bagging with leaked dataset provides above satisfactory performance.

Acknowledgments

The authors gratefully acknowledge the support from the College of Engineering at Effat University, Jeddah, Saudi Arabia as well as the technical support by Dr. Tayeb Brahimi.

References

Alpaydin, E. (2014). *Introduction to machine learning*. MIT press.

Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications*, *42*(20), 7046–7056. https://doi.org/10.1016/j.eswa.2015.05.013

Boyacioglu, M. A., & Avci, D. (2010). An Adaptive Network-Based Fuzzy Inference System (ANFIS) for the prediction of stock market return: The case of the Istanbul Stock Exchange. *Expert Systems with Applications*, *37*(12), 7908–7912. https://doi.org/10.1016/j.eswa.2010.04.045

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Chong, E., Han, C., & Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications*, *83*, 187–205. https://doi.org/10.1016/j.eswa.2017.04.030

Cichosz, P. (2014). *Data Mining Algorithms: Explained Using R*. John Wiley & Sons.

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, *270*(2), 654–669. https://doi.org/10.1016/j.ejor.2017.11.054

Hall, M., Witten, I., & Frank, E. (2011). Data mining: Practical machine learning tools and techniques. *Kaufmann, Burlington*.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Elsevier.

Index Information—Nikkei Indexes. (n.d.). Retrieved April 8, 2020, from https://indexes.nikkei.co.jp/en/nkave/index/profile?idx=nk225

Kenton, W. (n.d.). *NYSE Composite Index*. Investopedia. Retrieved April 8, 2020, from https://www.investopedia.com/terms/n/nysecompositeindex.asp

Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stock market prices using random forest. *ArXiv Preprint ArXiv:1605.00003*.

Malagrino, L. S., Roman, N. T., & Monteiro, A. M. (2018). Forecasting stock market index daily direction: A Bayesian Network approach. *Expert Systems with Applications*, *105*, 11–22. https://doi.org/10.1016/j.eswa.2018.03.039

Moghaddam, A. H., Moghaddam, M. H., & Esfandyari, M. (2016). Stock market index prediction using artificial neural network. *Journal of Economics, Finance and Administrative Science*, *21*(41), 89–93. https://doi.org/10.1016/j.jefas.2016.07.002

NASDAQ Composite—Components, Methodology & Criteria for Inclusion. (n.d.). Retrieved April 8, 2020, from https://corporatefinanceinstitute.com/resources/knowledge/trading-investing/nasdaq-composite/

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. *Expert Systems with Applications*, *42*(1), 259–268. https://doi.org/10.1016/j.eswa.2014.07.040

Pehlivanlı, A. Ç., Aşıkgil, B., & Gülay, G. (2016). Indicator selection with committee decision of filter methods for stock market price trend in ISE. *Applied Soft Computing*, *49*, 792–800. https://doi.org/10.1016/j.asoc.2016.09.004

What is the FTSE 100? | The Share Centre. (n.d.). Retrieved April 8, 2020, from https://www.share.com/a-guide-to-investing/before-you-start/what-is-the-ftse-100

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Zhang, X., Li, A., & Pan, R. (2016). Stock trend prediction based on a new status box method and AdaBoost probabilistic support vector machine. *Applied Soft Computing*, *49*, 385–398. https://doi.org/10.1016/j.asoc.2016.08.026

Zhong, X., & Enke, D. (2017a). Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications*, *67*, 126–139. https://doi.org/10.1016/j.eswa.2016.09.027

Zhong, X., & Enke, D. (2017b). A comprehensive cluster and classification mining procedure for daily stock market return forecasting. *Neurocomputing*, *267*, 152–168. https://doi.org/10.1016/j.neucom.2017.06.010