Stock Market Data Prediction Using Machine Learning Techniques: Proceedings of ICITS 2019

Chapter · February 2019

DOI: 10.1007/978-3-030-11890-7_52

4 authors:

Edgar P. Torres
Escuela Politécnica Nacional
12 PUBLICATIONS 83 CITATIONS
SEE PROFILE

Myriam Hernandez Alvarez
Escuela Politécnica Nacional
29 PUBLICATIONS 206 CITATIONS
SEE PROFILE

Edgar Torres
INVEX S.A.
7 PUBLICATIONS 36 CITATIONS
SEE PROFILE

Sang Guun Yoo
Escuela Politécnica Nacional
70 PUBLICATIONS 393 CITATIONS
SEE PROFILE

Stock Market Data Prediction Using Machine Learning Techniques

Edgar P. Torres P.[1](✉), Myriam Hernández-Álvarez[1],
Edgar A. Torres Hernández[2], and Sang Guun Yoo[1]

[1] Facultad de Ingeniería de Sistemas,
Escuela Politécnica Nacional, Quito, Ecuador
{edgar.torres,myriam.hernandez,sang.yoo}@epn.edu.ec
[2] Facultad de Ciencias Administrativas,
Pontificia Universidad Católica del Ecuador, Quito, Ecuador
etorresh777@gmail.com

Abstract. This paper studies the possibilities of making prediction of stock market prices using historical data and machine learning algorithms. We have experimented with stock market data of the Apple Inc. using random trees and multilayer perceptron algorithms to perform the predictions of closing prices. An accuracy analysis was also conducted to determine how useful can these types of supervised machine learning algorithms could be in the financial field. These types of studies could also be researched with data from the Ecuadorian stock market exchanges i.e. Bolsa de Valores de Quito (BVQ) and Bolsa de Valores de Guayaquil (BVG) to evaluate the effectiveness of the algorithms in less liquid markets and possibly help reduce inefficiency costs for market participants and stakeholders.

Keywords: Machine learning · Stock market · Artificial intelligence · Prediction

1 Introduction

This paper explores the possibility of using computers as tools to automatically process data gathered from financial activities and extract relevant information to achieve the goals of the participants involved in this field. The objective is to perform an automatic process of data using artificial intelligence techniques, in particular the application of machine learning algorithms. The final goal is to manage the extraction of quantitative data with relevant information from the stock market. For the purpose of this paper, we have used the historical data of opening price, closing price, highest price, lowest price, and volume traded of Apple Inc. stocks which has been gathered using Google Finance.

We think that the price of a stock incorporates all the relevant data related to the whole process of supply and demand of the stock market. This process of price discovery in the markets leaves traces in its historical prices.

Á. Rocha et al. (Eds.): ICITS 2019, AISC 918, pp. 539–547, 2019.
https://doi.org/10.1007/978-3-030-11890-7_52

This research is meant to be a first approximation to this subject by using machine learning techniques to process stock market data that are readily quantifiable. Furthermore, it is possible to perform predictions about the possible variations of Apple Inc.'s (NASDAQ: AAPL) stock through a historical price data set that includes different types of prices and their fluctuations while noting the impact of the traded volume for each market session. Finally, our goal is to evaluate how accurate these predictions are and measure its error.

The rest of the present paper is organized as follows. Section 2 explains some basic concepts and previous works related to the present paper. Then, Sect. 3 describes the details of the stock market price prediction activities. Finally, Sect. 4 delivers the conclusions of this work.

2 Background

2.1 Basic Concepts

Before proceeding with the stock market prices analysis, it's important to define specific concepts used in the present research:

- Data mining represents the use of automated techniques to analyze data to discover relationships between different variables in the data.
- Machine learning or automated learning is an Artificial Intelligence branch that develops techniques that allow computers to learn.
- Supervised Learning is a machine learning technique that generates a model from a training data set that is capable of predicting variables given input.
- Unsupervised Learning is a machine learning technique where models are formed by fitting it to the dataset used.
- Feed Forward Neural Network is an artificial neuronal network where connections between units don't form a cycle.
- WEKA is a collection of machine learning algorithms that can perform data mining operations.

2.2 Previous Works

Great effort has been placed to analyze financial data since the economic incentive is considerable for participants. Because of this reason, there are several researches about predictions and future tendencies of stock market prices [1]; some of them uses machine learning techniques such as Support Vector Machine (SVM) [3, 4] and others use text analysis to understand emotions and other quantifiable data from relevant financial information sources [1, 2, 4]. These types of techniques extract information regarding the participant's emotions and feelings about a particular subject, in this case, the stock market data. Generally speaking, it's possible to extract the data from multiple sources, like highs or lows in prices with a notable traded volume, but also financial statements, press releases, and other information from the company [2]. Another source for this type of information can be microblogs in social networks [3], and in particular stock tweets [3, 4], although this kind of analysis falls outside the scope of this article. In conclusion, there are many possibilities to trace the emotions and feelings of stock market participants. Finally, other approaches have tested this type of information from local data [2], expert data and other participants [3], and systematic processing of lexicographic, syntactic and natural language for sentimental analysis. [5].

2.3 Weka

Weka which means Waikato Environment for Knowledge Analysis is a free machine learning software developed at University of Waikato in New Zealand. Weka includes a collection of tools and algorithms for data analysis and prediction, and it is widely used for research [6, 7].

There is plenty of available information about WEKA’s algorithms capabilities and usage for financial data through machine learning techniques. These sources are available in WEKA’s official website and also other sites that offer tutorials on how to use its software, as well as papers researching the mathematical and technical aspects of WEKA’s algorithms. [8–12].

WEKA’s classifiers provides different models to predict nominal and numerical quantities. Among others, WEKA offers the following algorithms: Decision trees and lists, Classifiers based on instances, Support Vector Machines (SVM), Multilayer Perceptron, Logistical Regression, Bayesian Networks, and “Meta” Classifiers.

3 Stock Market Data Prediction Using Machine Learning Algorithms

3.1 Generalities

The objective of this research is to forecast or predict future closing prices of Apple Inc.’s (NASDAQ: AAPL) stock. To this end, we have used historical price data of the Open, Close, High, Low and Volume of the last 250 trading sessions. We obtained this information from Google Finance.

An important supposition for this research is that the different historical price types and their traded volume leave traces of the price discovery process. By extension, this method relies heavily on the emotions, feelings, and expectations of participants, some of which are often irrational and purely speculative. Hence, price jumps or crashes with heavy or light traded volume should have significance, and these subtle relations can be modeled by the application of machine learning techniques as previously mentioned. In this way, a prediction of the following closing prices can be made with a certain degree of acceptable accuracy. We’ve chosen to predict the variable Close because it’s the last price at which all participants agreed for the financial security in the traded session.

Much has been written about daily price fluctuation in stock market prices, often suggesting that these movements resemble a Random Walk [13], or a Brownian motion [14], among others. Nevertheless, it is accepted in finance that stocks have had a historically positive Baseline Drift which can be attributed to prices incorporating the growth of earnings and equity publicly traded companies. In this aspect, since in the last 250 stock market sessions, Apple Inc. has released multiple quarterly reports, it stands to reason that these results should be reflected in its stock price. On the other hand, the day to day trading of the stocks will provide with different types of data that machine learning algorithms should be able to model. We expect that by using this information, we can reduce the algorithms exposure to random noise and improve its accuracy for forecasting future closing session prices.

3.2 Algorithm Selection and Universe to Be Analyzed

The present work has used the WEKA software to execute the machine learning algorithms in our stock market data sets. As mentioned previously, we have obtained the data from Google Finance which contains the Open, High, Low and Close prices of Apple Inc. stock and the volume of trades of the last 250 sessions. Regarding to the machine learning algorithms, we have used the following WEKA packages:

- weka.classifiers.trees.RandomTree
- weka.classifiers.functions.Multilayer Perceptron

We tested both of these algorithms on the same dataset, i.e., the historical price values of the last 250 trading sessions of Apple Inc.'s (NASDAQ: AAPL) stock.

As previously mentioned, we considered six attributes for the algorithms: (1) Date, (2) Open, (3) High, (4) Low, (5) Close, and (6) Volume.

- Date: Market session date
- Open: The opening price for the market session
- High: The highest price reached during the market session
- Low: The lowest price reached during the market session
- Close: The closing price for the market session
- Volume: The total number of trades performed on the stock during the market session

The attribute Close is the one to be forecasted and compared to its real data, so that the accuracy of the algorithms can be tested and measured for errors and over fitting. Figures 1 and 2 present the results of the executions performed using Random Tree and Multilayer Perceptron algorithms.

```
Algorithm: weka.classifiers.trees.RandomTree
=== Run information ===

Scheme:     weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1
Relation:   prediccion 2-sep
Instances:  251
Attributes: 6
            Date
            Open
            High
            Low
            Close
            Volume
Test mode:  evaluate on training data

=== Classifier model (full training set) ===

RandomTree
==========
High < 104.4
|   High < 98.94
|   |   Open < 94.72
|   |   |   High < 94.07
…
|   |   |   |   High >= 123.09
|   |   |   |   |   Date < 41.5 : 122.57 (1/0)
|   |   |   |   |   Date >= 41.5 : 122 (1/0)

Size of the tree : 225

Time taken to build model: 0 seconds

=== Predictions on training set ===
   inst#     actual  predicted      error
       1   112.31    112.453      0.143
       2   110.15    110         -0.15
       3   112.57    112.453     -0.117
…
     250   106.73    106.73      -0
     251        ?    105.024       ?

=== Evaluation on training set ===

Time taken to test model on training data: 0.08 seconds

=== Summary ===
Correlation coefficient                  0.9998
Mean absolute error                      0.1151
Root mean squared error                  0.16
Relative absolute error                  1.6105 %
Root relative squared error              1.9553 %
Total Number of Instances              250
Ignored Class Unknown Instances                  1
```

Fig. 1. Execution of Random Tree

Algorithm: weka.classifiers.functions.MultilayerPerceptron

```
=== Run information ===
Scheme:      weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -
S 0 -E 20 -H a
Relation:    prediccion 2-sep
Instances:   251
Attributes:  6
             Date
             Open
             High
             Low
             Close
             Volume
Test mode:   evaluate on training data

=== Classifier model (full training set) ===

Linear Node 0
    Inputs    Weights
    Threshold    0.16407389183567725
    Node 1    -0.7409666964660168
    Node 2    1.7171258338000799
    Node 3    -1.371651766795869
Sigmoid Node 1
    Inputs    Weights
    Threshold    -1.225106292966126
    Attrib Date    0.01979747908695488
    Attrib Open    0.4749210034034923
    Attrib High    -0.5455483243023302
    Attrib Low    -0.7448374160532633
    Attrib Volume    0.5265070532721522
Sigmoid Node 2
    Inputs    Weights
    Threshold    -1.4617941741850469
    Attrib Date    0.006059752397296158
    Attrib Open    -1.3715046053801243
    Attrib High    1.6670130380975303
    Attrib Low    1.4209018925134493
    Attrib Volume    0.11134270375207257
Sigmoid Node 3
    Inputs    Weights
    Threshold    -1.3965598781124515
    Attrib Date    0.05273749507771454
    Attrib Open    0.7651119521019094
    Attrib High    -1.1018676853352654
    Attrib Low    -1.5648480563559939
    Attrib Volume    -0.23675701241891794
Class
    Input
    Node 0

Time taken to build model: 0.1 seconds

=== Predictions on training set ===

    inst#     actual  predicted      error
        1    112.31     111.523     -0.787
        2    110.15     111.099      0.949
…

      250    106.73     106.161     -0.569
      251         ?     107.005          ?

=== Evaluation on training set ===

Time taken to test model on training data: 0.15 seconds

=== Summary ===

Correlation coefficient                  0.9976
Mean absolute error                      0.4529
Root mean squared error                    0.6058
Relative absolute error                  6.3376 %
Root relative squared error                7.4024 %
Total Number of Instances                  250
Ignored Class Unknown Instances                  1
```

Fig. 2. Execution of Multilayer Perceptron algorithm

3.3 Results Evaluation and Analysis

Both executions of WEKA's algorithms fit the actual historical Price data (Correlation factor of 0.9998 for the first one and 0.9976 for the second with a maximum adjustment possible of 1.0) very closely and errors are tolerable (mean absolute error of 0.1151 for the first one and 0.4529 for the second). Hence, it is reasonable to conclude that the predictions using these two algorithms are suitable and acceptable for the application. Table 1 shows the details of the comparison of the results and Table 2 shows a portion of the historical data used for analysis.

Table 1. Comparison of algorithms

No.	Attributes	Algorithm 1	Algorithm 2
1	Correlation coefficient	0.9998	0.9976
2	Mean absolute error	0.1151	0.4529
3	Root mean squared error	0.16	0.6058
4	Relative absolute error	1.6105%	6.3376%
5	Root relative squared error	1.9553%	7.4024%
6	Total number of instances	250	250

Table 2. Examples of historical data

Date	Open	High	Low	Close	Volume
1	108.59	109.32	108.53	108.85	21257669
2	108.86	109.1	107.85	108.51	25820230
3	108.77	109.69	108.36	109.36	25368072
4	109.23	109.6	109.02	109.08	21984703
5	109.1	109.37	108.34	109.22	25355976
6	109.63	110.23	109.21	109.38	33794448
…	…	…	…	…	…
245	112.49	112.78	110.04	110.37	52906410
246	110.23	112.34	109.13	112.34	61520170
247	110.15	111.88	107.36	107.72	75988194
248	112.03	114.53	112	112.76	55962842
249	112.17	113.31	111.54	113.29	52896384
250	112.23	113.24	110.02	112.92	83265146

Table 3. Example of forecasted data

Date	Actual close	Predicted close	Error
251	105.17	105.024	0.146

Based on the previous analysis, it is possible to conclude that the first algorithm has a higher correlation coefficient and a lower error than the second one. Although, both algorithms present a good performance. Finally, Table 3 presents the forecast for the Close attribute for a test set using the Random Tree Algorithm (algorithm with better result).

4 Conclusions

The present work has shown how it is possible to perform forecasts and predictions of future stock market data using artificial intelligence techniques, specifically machine learning algorithms. The application of WEKA proved to be very valuable for this purpose and its use could be further researched in the financial field. This tool counts with many different algorithms which could be used for various types of economic data that could provide interesting insights to market participants and society. Additionally, in future works, it could be possible to quantify and analyze emotions and feelings expressed in a text through blogs, stock tweets, or other mediums, to increase the effectiveness of the predictions. Finally, we recommend that the Ecuadorian stock exchanges (BVQ and BVG) make this type of information (open, high, low, close, volume and otherwise data concerning the sentiment of market participants) public and readily available, through free online channels if possible. This transparency of information would facilitate the application machine learning algorithms and artificial intelligence to Ecuadorian financial securities to further research their application in other markets, which may help reduce costs of market inefficiencies. Ecuadorian government should pass laws to make transparent stock market information. This would constitute a valuable innovation for Ecuadorian stock market.

References

1. Schumacher, R.P., Chen, H.: Textual analysis of stock market prediction using breaking financial news: the AZFin text system. ACM Trans. Inf. Syst. **27**, 1–19 (2009)
2. Giannini, R.C., Irvine, P.J., Shu, T.: Do local investors know more? A direct examination of individual investors' information set. Working paper (2014)
3. Bar-Haim, R., Dinur, E., Feldman, R., Fresko, M., Goldstein, G.: Identifying and following expert investors in stock microblogs. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1310–1319 (2011)
4. Plakandaras, V., Papadimitriou, T., Gogas, P., Diamantaras, K.: Market sentiment and exchange rate directional forecasting. Algorithmic Financ. **4**(1–2), 69–79 (2015)
5. Ruiz-Martínez, J.M., Valencia-García, R., García-Sánchez, F.: Semantic-based sentiment analysis in financial news. In Proceedings of the 1st International Workshop on Finance and Economics on the Semantic Web, pp. 38–51 (2012)
6. Kumar, A., Malik, H., Chandel, S.S.: Selection of most relevant input parameters using WEKA for artificial neural network based solar radiation prediction models. Renew. Sustain. Energy Rev. **31**, 509–519 (2014)

7. Kalmegh, S.: Analysis of WEKA data mining algorithm REPTree, Simple Cart and RandomTree for classification of indian new. Int. J. Innov. Sci. Eng. Technol. **2**(2), 438–446 (2015)
8. Teuvo, K.: Self-Organizing Maps. Springer Science & Business Media, Heidelberg (2001)
9. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice-Hall, Upper Saddle River (1995)
10. Aha, D.W.: Tolerating noisy, irrelevant, and novel attributes in instance based learning algorithms. Int. J. Man Mach. Stud. **36**(2), 267–287 (1992)
11. Wettschereck, D., Aha, D.W., Mohri, T.: A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artif. Intell. Rev. **11**(1–5), 273–314 (1997)
12. Hornik, K., Buchta, C., Zeileis, A.: Open-source machine learning: R meets weka. Comput. Stat. **24**(2), 225–232 (2009)
13. Fama, E.F.: Random walks in sotck market prices. Financ. Anal. J. **51**(1), 75–80 (1995)
14. Osborne, M.F.M.: Brownian motion in the stock market. Oper. Res. **7**(2), 145–173 (1959)