|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured pagePhoto displaying partial image of two pie charts on a canvas-textured page |
| Stock Action Prediction using Data Mining |
| |  |  |  | | --- | --- | --- | | Harshan Nagulapally | 11/26/18 | MSIS 672 Data Mining | |

Contents

[I. Abstract 1](#_Toc532768225)

[II. Introduction 2](#_Toc532768226)

[III. Literature Review 3](#_Toc532768227)

[IV. Analysis 4](#_Toc532768228)

[1) Method of Study 4](#_Toc532768229)

[2) Data Collection 4](#_Toc532768230)

[3) Understanding the Data 4](#_Toc532768231)

[4) Preparing Data 4](#_Toc532768232)

[Table 1: 5](#_Toc532768233)

[Table 2: 6](#_Toc532768234)

[5) Building a Model 6](#_Toc532768235)

[Picture 1: 7](#_Toc532768236)

[Picture 2: 7](#_Toc532768237)

[V. RESULTS AND DISCUSSION 8](#_Toc532768238)

[VI. Conclusions 8](#_Toc532768239)

[References 8](#_Toc532768240)

## Abstract

*Researchers have spent several resources for a very long period to establish accurate stock price prediction techniques. They found patterns in the existing historical stock price data across different industries and analyzing them to understand and formulate it to forecast the stock price. This study helps the investors to safely spend their money for the greater interest of the market and the investors as well. The decision and the prediction patterns and the factors determining the stock price can be evaluated and shortlisted by implementing different techniques such as decision tree classification, neural network analysis and association rules.*

## Introduction

The stock market is essentially a non-linear, nonparametric system that is extremely hard to model with any reasonable accuracy [1]. Researchers have been trying to understand the pattern and conclude a way to predict the stock price. To achieve this we have to factor for several variables that contribute to varying price of stocks such as industrial sector of the company we are looking into, the macroeconomics and the current market situation in that industry due to the existing economy and it constrains and finally fundamental analysis of the price of the stock which depends on the returns on investment and intrinsic value of the price. But I assume considering these factors alone will not be enough to predict stock price. We should also account into the changes in the industry over time, public holidays, types of investors(full time, dedicated investors or part time investors) and type of investments(short term of long term). Another factor that affects the stock price is the company’s operations. These mentioned above are just few factors that contribute to changes in stock price. Usually analysts believe that fundamental analysis is enough to predict a stock price, but this will hold good only for long term investments. And for short- and medium- term investments, we need much stronger and precise microscopic analysis of stock price as this it is more volatile.

This all analysis relays on an assumption that these patterns repeat. We all understand history repeats itself based on the previous experiences, be it climatic conditions across a region, natural changes around the solar system, changes in the economies of different countries and thus affecting the global market and so on. There are several such examples where we can analyses and settle down to a pattern that can repeat in the future. Thus, some other researchers use the techniques of technical analysis [2]. Technical analysis basically uses the above assumed pattern theory to use the historical data and the existing volume to first identify a pattern and then use it to predict the future stock price of the same stock. But this method has not proved to be successful for several reasons such as other factors affecting the market or company has changed and those patterns do not tend to repeat themselves as that of the stock price pattern. For example, due to the advancement in technology, companies are now able to use solar energy for a part of their operations, this was not the case 10 years ago and thus any change in the price of fuel or other industries would show significant affect in the operational price for the companies and therefore stock price would be affected indirectly, but now the prediction based on the historical data will be less effective. Thus, we must come up with more effective way to predict and factor variables that account to changes in stock prices.

Data mining refers to business analytics methods that go beyond counts, descriptive techniques, reporting, and methods based on business rules. [3] It covers areas such as decision trees, neural networks and association rules to better analyze the data. Decision tress are graphical representation of possible outcomes that will provide important information about the data with easy classification rules. But there some constrains such as if the data is very large, it is difficult for the system to plot the tree and for us to analyze the output. Classification tree basically generates node on 0 or 1 corresponding to a non-acceptor (0) or acceptor (1), also known as terminal nodes. And then we have more advanced method of classification or prediction called neural networks. This method tries to replicate human level learning and analyzing skills. They are also called as black box, where complex algorithmic classification of the data takes place. Several successful applications have been reported in several industries, also in financial applications (see Trippi and Turban, 1996) such as bankruptcy predictions, currency market trading, picking stocks and commodity trading, detecting fraud in credit card and monetary transactions, and customer relationship management (CRM).[3] The idea behind neural networks is to combine the predictor information in a very flexible way that captures complicated relationships among these variables and between them and the outcome variable.

The challenge in using a database is to generate useful rules from raw data in a database for users to make decisions, and these rules may be hidden deeply in the raw data of the database. Association rules helps us tackle this situation by contributing to analyze and forecast what goes with what to deeply understand the data. The idea behind association rules is to examine all possible rules between items in an if–then format and select only those that are most likely to be indicators of true dependence.[3] This has opened new era of understanding patterns in customer purchases and disease analysis. Different measuring ratios such as support, confidence and lift ratio provide correctness of the rules and its significance to the available data.

In this paper I am trying to analyze the historical stock price to predict if the stock price will increase or reduce. This will help investors decide if they must buy or sell a stock. Analyzing stock price data involve thousands of records collected over the past. To make it compatible for the study, few hundreds of records will be selected as sample to develop techniques and test. The data will be used to implement different data mining techniques such as decision tress, neural networks and association rules. The data is collected from Yahoo Finance. And the stock I choose to perform this analysis is one that is of a common, stable and well-established company, this approach is to eliminate any other factors that account for price fluctuations of stock price for a company.

## Literature Review

Financial market has been one of the oldest and main driving factors for any economy and has been playing a major role ever since the beginning. And there has been several advancements in the finance market in the recent past and in the technology that plays a major role in deciding the advancements in the finance market, such as the development of powerful communication and trading facilities has enlarged the scope of selection for investors providing access to those who do not have a trading account. With such advancements has made it necessary for the finance sectors to closely observe and predict the stock market to make the most of the available opportunities. Thus, several researchers focused on forecasting the stock returns based on the publicly available historic data expecting to have some predictive relationships for the future returns.

In a paper published in 2006, researchers used advanced techniques to predict stocks. “Stock market trading rule discovery using two-layer bias decision tree” by Wang, J.L., Chan, S.H. [4], they applied the concept of serial topology to design a new decision system, namely the two-layer bias decision tree, for stock price prediction. The two aspects the study focused are to eliminate classification error after using decision models and achieve more accuracy by implementing two-layer bias decision tree. The results indicated with this decision model produced excellent accuracy, and it significantly outperformed other models that produced random purchase decision trees.

Similarly, there are several other studies and papers that focus on different aspects such as applying neural network techniques to forecast stocks, where the authors use the effectiveness of the neural networks to estimate different levels and classify them accordingly before interpreting the outcome. Others implement accuracy of univariate and multivariate neural network models, which involve deeper analysis of neural networks. And while some studies work on price prediction, others concentrate on stock selection and few others study the affect of corporate governance and performance indicators such as economic efficiencies and growth rates.

Though with such vast approach to cover all the aspects of finance, researchers are still looking into methods to precisely predict stock prices. There are several papers that work with association rules, factor analysis, machine learning techniques and artificial intelligence and other advances approaches where new high-speed time delay neural networks are used to predict stock price or operations decisions or growth rates predictions.

With all such advancements, future of finance shapes into a very interesting industry. With all the available data we can establish some amazing patterns and develop algorithms to precisely predict stock prices. Researches are using more advanced technologies for this field of study to establish accuracy. Mr. Hazem[5] proposed a new approach for fast forecasting of stock market prices. In the proposed approach, he uses new high-speed time delay neural networks(HSTDNNs) in MATLAB tool to simulate results to confirm the theoretical computations of the approach.

## Analysis

## Method of Study

Methodology of the study is designed to ensure that the results of the analysis leads to a stable model that successfully addresses the problem it is designed to solve. Various steps of the analysis involve gathering data, analyzing data, disseminating results and implementing results to forecast the increase or decrease in price. The objective of this study is to help the investors in the stock market to decide on buying or selling stocks based on the knowledge extracted from the historical prices of such stocks. The decision is based on the data mining techniques; the decision tree classifiers.

## Data Collection

The data is collected from Yahoo Finance. This is one of the commonly used and publicly available source where several researchers refer to for historical financial data of all the public companies. Yahoo is a very huge group of company that covers several other industries such as media, news and entertainment. The financial news focuses on data and commentary including stock quotes, press releases, financial reports, and original content. It also offers some online tools for personal finance management. Initially yahoo was known for its email service.

I opened the yahoo finance home page and navigated to Amazon Stock price page through the search bar. There we can choose, Historical data tab to extract the daily stock price changes. There I selected the date range and choose the max period date to load the stock price changes from the day Amazon went public, back in May of 1997. And we can choose Download option after we apply the date range. This saves the historical data of Amazon stock in a CSV file. The ticker for Amazon is “AMZN”.

## Understanding the Data

The data contains the following attributes and the data details as displayed below.

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | Description | Values/Type | Possible Value |
| Date | Date of the stock for AMZN | Date format | Date |
| Open | The opening price of on the given date | Number- price | Positive |
| High | The highest price stock reached on the given date | Number – price | Positive |
| Low | The lowest price stock reached on the given date | Number – price | Positive/negative |
| Close | The closing pricing of the stock on the given date | Number – price | Positive/negative |
| Adj Close | This is similar to close price of the stock. Usually refers to any adjustments made before the close of the stock on the given date. | Number – price | Positive/negative |
| Volume | The number of stock available in the market on the given date. | Number – quantity | Positive |

The amount of data is very huge, it contains 5467 records of historical data. The price values of stocks are recorded up to 6 decimals. The data will be cleaned and organized in the further steps.

## Preparing Data

The data collected has most of the attribute values as continuous numeric values. Data transformation techniques were applied by generalizing data to a higher-level concept so as all the values became discrete. If the opening price of a stock is greater than the close price of the previous day, then it is identified as a positive gain and is considered as positive increase in stock price, thus positive is marked in a new column generated. Similarly, when the open price of the stock is lower than the close price of the previous day, it is considered as negative and will be marked negative. And if the open price of the price is same as the previous day close price then it is marked as equal. The table below is an example from the actual data used for analysis.

### Table 1:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Date** | **Open** | **Previous Day Close** | **Discrete value assigned** | **Data Code** |
| 05/16/1997 | 1.968 | 1.958 | Positive | 0 |
| 05/19/1997 | 1.760 | 1.729 | Negative | 1 |
| 05/21/1997 | 1.635 | 1.635 | Equal | 2 |

The below picture is the screen shot of the sample data organized in the required format for further analysis after all the data process.

### Table 2:

![A close up of a piece of paper

Description automatically generated]()

## Building a Model

The model we build to understand and explore the existing data is using classification trees. Using classification trees for this model gives us several advantages. As the classification tree dose not require domain knowledge and it is preferred for descriptive models, it is the best method for this scenario, as we are trying to find the hidden pattern in the historical data and explain the same. Using metric gain and entropy, the steps of a decision tree are simple to understand, and the predictions can have high accuracy. The attributes are classified based on the gain ratio and are ranked accordingly. After implementing the decision tree, a set of rules were generated in which maximum number of attributes were considered.

### Picture 1:

![A close up of a map

Description automatically generated]()

The above image displays the classification tree for the training data. This tree is presenting options for positive action buy-sell decision. We need to further implement advanced methods such as ID3 and C4.5 algorithms to further build a decision tree and included all the data and present much deeper understanding of the action to be taken on daily basis.

### Picture 2:

The below classification tree is an example of how the training data looks after implementing punning.

![A picture containing map, text

Description automatically generated]()

## RESULTS AND DISCUSSION

The main purpose of this study is to understand how much role the numbers play and if there are other factors that influence the financial sector. Based on the accuracy attained from the above classification tree, it is conclusive that their financial sector is subject to several factors that are not presented in the historical data of a company. Factors include such as news about the company and the trends of the industry it belongs to, the financial performance of the company for that quarter or yearly financial reports subject to period of the stock analyzed and finally the overall performance of the market, which further is defined by several other political and economical driving factors. Thus, we cannot rely completely on classification tree to accurately predict the performance of a stock. But we can add come basic political and economic factors that follow the trend to the existing model along with the neural networks and increase the accuracy.

## Conclusions

The purpose of the study was to present the use of decision tree to classify and predict historical stock data of a company and check the accuracy of the model. Following the method proposed in the study would assist investors to trade the stock in the right time to gain maximum profit based on the available historical data. Based on the available data, the results of the proposed model were accurate, as we did not consider several other factors that play an important role in determining the stock price.

For the future of such model, we have a huge window to implement and research for more accurate methods to attain close prediction algorithms of the stock market. But this can only be achieved by collaborating all the other factors that affect the financial sector. This includes collecting more data and applying more advanced techniques such as neural networks and association rules.

## References

1. Wang, Y.F., (2003) “Mining stock price using fuzzy rough set system”, Expert Systems with Applications, 24, pp. 13-23.
2. Wu, M.C., Lin, S.Y., and Lin, C.H., (2006) “An effective application of decision tree to stock trading”, Expert Systems with Applications, 31, pp. 270-274.
3. Shmueli, Galit, et al. Data Mining for Business Analytics Concepts, Techniques, and Applications in R. Wiley, 2018. Page 5, Page 271, Page331
4. Wang, J.L., Chan, S.H. (2006) “Stock market trading rule discovery using two-layer bias decision tree”, Expert Systems with Applications, 30(4), pp. 605-611.
5. Hazem M. El-Bakry, and Wael A. Awad, Fast Forecasting of Stock Market Prices by using New High-Speed Time Delay Neural Networks, International Journal of Computer and Information Engineering 4:2 2010. Pp 138-144.