Predicting Stock Prices Using Data Mining Techniques

S.Thilagavathi¹, K.K. Suresh kumar²
Research Scholar¹, Assistant Professor²
Department of Computer Science (PG)
Kongu Arts and Science College (Autonomous), Erode, India

Abstract:
Forecasting stock return is an important financial that has attracted researchers attention for many years. It involves an assumption that fundamental information available in the past has some predictive relationships to the future stock prices. This study tries to help the investors in the stock market to decide for buying or selling stocks based on the historical prices of such stocks. The decision taken will be based on decision tree classifier which is one of the data mining techniques.

Keywords: Data Mining, Data Mining, Data Classification, Decision Tree, Future stock return, data mining techniques, decision tree classifiers.

I. INTRODUCTION
The stock market is a non linear, non parametric system that is extremely hard to model with any reasonable accuracy [1]. Investors have been tried to find a way to predict stock prices and find the right stocks and right timing to buy or sell. According to [2], some research used the techniques of fundamental analysis, where trading rules are developed based on information associated with macroeconomics, industry, and company. The authors [5] and [6] that fundamental analysis that the price of a stock depends on its intrinsic value and expected return on investment. The stock price can be predicted reasonably well. The fundamental analysis is a good method only on a long term basis. Some other research used the technical analysis [2], in which trading rules were developed based on the historical data of stock trading price and volume. The technical analysis as illustrated in [5] and [7] refers to the various methods to predict future price movements using stock prices and volume information. It is based on the assumption that history repeats and the future market can be determined by examining historical price data. It is assumed that price trends and the patterns that can be identified and utilized for profit. Most of the techniques used in technical analysis are highly subjective in nature and has been shown not to be statistically valid. The data mining techniques and artificial intelligence techniques like decision trees, rough set approach, and artificial neural networks have been applied to this area [8]. Data mining [9] refers to extracting or mining knowledge from large data stores or sets. The concept or class descriptions, associations and correlations, classification, prediction, clustering, trend analysis, deviation analysis, similarity analysis. Data classification can be done by many different methods; one of this method is the classification by using Decision Tree. It is a graphical representation of all possible outcomes and the paths by which may be reached. The grand challenge of using the database is to generate useful rules from raw data in a database for users to make decisions, and the rules may be hidden deeply in the raw data of the database. Traditionally, the method of turning data into knowledge relies on manual analysis; it is becoming impractical in many domain as data volumes grow exponentially. The problem with predicting stock prices is the volume of data is too large and huge. This paper is used one of the data mining methods which is the classification approach to the historical data available to help the investors to build their decision on whether buy or sell that stock in order to achieve profit. The main objective of this paper is to analyze the historical data available stocks using decision tree technique as one of the classification methods of data mining to help the investors to know when to buy new stocks or to sell their stocks. Analyzing the stock price data over several years may be involve a few hundreds or thousands of records, but there must be selected from millions. The data will be used this paper to build the decision tree will be the historical prices.

II. LITERATURE REVIEW
The research by [2] using decision tree technique to build on the work [12] where Lin tried to modify the filter rule that is to buy when the stock price rises k% above its past local low and sell when it falls k% from its past local high. The proposed modification to the filter rule in [12] was by combining three decision variables associated with fundamental analysis. An empirical test, using the stocks of electronics companies showed Lin’s method outperformed the filter rule. According to [2], in work, the criteria for clustering trading points involved the past information, the future information was not considered at all. The research by [2] aim to improve the filter rule by considering both the past and the future information in clustering the trading points. The researchers used to data on stock market and that of NASDAQ to carry out empirical tests. Test results showed that the proposed method outperformed both Lin’s method and the filter rule in the two stock markets. The model of [11] applied to the concept and designed a new decision system, namely the two layer bias decision tree for stock price prediction. The methodology developed by the authors differs from other studies in two respects, first to reduce the classification error, the decision model was modified into a bias decision model. Second a two layer bias decision tree is used to improve purchasing accuracy. The empirical results that indicates the presented decision model produced excellent purchasing accuracy. The authors of [10] presented an approach that used data mining methods and neural networks for forecasting stock market returns. An attempt has been made in this study to investigate the predictive power of financial and economic variables by adopting the variable relevance analysis technique in machine learning for data mining. The authors refers the effectiveness of the neural network models used for
level estimation and classification. The results showed that the trading strategies guide d by the neural network classification models generate higher profits under the same risk exposure than those suggested by other strategies. The research by [13] was basically a comparison between the work [14 - 15] and the artificial neural networks in order to try to predict the stock prices in the Chinese market. The purpose of the study is to demonstrate the accuracy of ANN in predicting stock price. In order to demonstrate the accuracy of ANN, the authors made a comparative analysis between Fama and French’s modeling and the predictive power of the univariate and multivariate neural network models. The results from this study indicated that artificial neural networks offer an opportunity for investors to improve their predictive power in selecting stocks, and more importantly, a simple univariate model appears to be more successful at predicting returns than a multivariate model. During the work, [16] presented a study that aimed to provide evidence of whether or not the corporate governance & performance indicators of the Jordanian industrial companies are affected by variables that were proposed and to provide the important indicators of the relationship of corporate governance & firms’ performance that can be used by the Jordanian industrial firms to solve the agency problem. The study random sample consists of (44) Jordanian industrial firms. The study founds a positive direct relationship between corporate governance and corporate performance. During the work [17] provided an overview of application of data mining techniques such as decision tree, neural network, association rules, and factor analysis and in stock markets. Prediction stock price has one of the biggest challenges to the AI community. Various technical, fundamental, and statistical indicators has been proposed and using varying results. During the work [18] surveyed some recent literature in the domain of machine learning techniques and artificial intelligence used to predict stock market movements. Artificial Neural Networks are identified to be the dominant machine learning technique in stock market prediction area. During the work, [19], proposed a new approach for fast forecasting of stock market prices. The proposed approach uses new high speed time delay neural networks. The authors used the MATLAB tool to simulate results to confirm the theoretical computations of the approach.

III. THE METHODOLOGY OF THE STUDY
Data mining methodology is designed to ensure that the data mining efforts leads to a stable model that successfully addresses the problem it is designed to solve. Various data mining methodology have been proposed to serve as blueprints for how to organize the process of gathering data, analyzing data, disseminating results, implementing results, and monitoring results [9]. To build the model that analyses the stock trends using the decision tree technique, the CRISP-DM (Cross-Industry Standard Process for data mining) [20] is used. This methodology was proposed in the mid-1990s by the European companies to serve as a non proprietary standard process model for data mining. The main reason and objective of this model is to try to help the investors in the stock market to decide the best timing for buying or selling stocks based on the knowledge extracted from the historical prices of such stocks. The decision taken will be based on one of the data mining techniques; the decision tree classifiers.

Understanding the collected data
The Oracle database contains the historical prices of the 230 companies listed in the exchange from the year 2000. As the amount of such data is very large and complicated, the decision was taken to choose three companies listed in the exchange. The selection of these companies was based on the following five criteria which represent the companies’ size and liquidity: Market capitalization, days traded, turnover ratio, value traded and the number of shares traded, also the sector representation was considered during the selection of these companies. These companies are “Arab Bank”, its code is in the stock market “ARBK” and its belongs to the banking sector, “United Arab Investors Company”, its code is “UAIC” and its belongs to the services sector, and “Middle East Complex for Engineering, Electronics and Heavy Industries”, its code is “MECE” and it belongs to the industrial sector. The period that was selected from April 2005 to May 2007, this presented the current and actual status of the market at that period of time.

Preparing the data
At the beginning, when the data was collected, all the values of the attributes selected were continuous numeric values. Data transformation was applied by generalizing data to a higher - level concept so as all the values became discrete. The criterion that was made to transform the numeric values of each attribute to discrete values depended on the previous day closing price of the stock. If the values of the attributes open, min, max, last were greater than the value of attribute previous for the same trading day, the numeric values of the attributes were replaced by the value Positive. If the values of those attributes were equal to the value of the attribute previous, the values were replaced by the value Equal. Table 2 shows a sample of the continuous numeric values of the data before selecting the 6 attributes manually and before generalizing them to discrete values, while table 3 shows the same sample after selecting the 6 attributes and after transforming them to discrete values.

Building the model
After the data has been prepared and transformed, the next step was to build the classification model using the decision tree technique. The decision tree technique was selected because [9] the construction of decision tree classifiers does not require any domain knowledge, thus its appropriate for exploratory knowledge discovery. Also, it can handle high dimensional data. Another benefit is that the steps of decision tree induction are simple and fast. Generally, decision tree accuracy is considered good. The decision tree method depends on using the information gain metric that determines the most useful attribute. The information gain depends on the entropy measure. The gain ratio is used to rank attributes and to build the decision tree where each attribute is located according to its gain ratio. When the decision tree model was applied on the data of the three companies using the WEKA software version 3.5 [21], the root attribute for both ARBK and UAIC company was the Open, while the attribute Last was the root for the decision tree of the MECE company. As the process of building the tree goes on, all the remaining attributes were used to continue with this process. After building the complete decision tree, the set of classification rules were generated by following all the paths of the tree. The maximum number of attributes that were used in some of the classification rules generated were 4 attributes, while some classification rules used only 1 attribute. Both the ID3 and C4.5 algorithms were used in building the decision trees and the pruning technique was used in the C4.5 algorithm in order to reduce the size of the produced decision trees. Table 4 gives...
a summary about the numbers of the classification rules that resulted after building the decision trees for each company using the C4.5 algorithm.

Deploying the model
The classification rules that were generated from the decision tree model can be used and integrated in a system that predict the best action and timing for the investors, either to buy or sell the stocks on that day.

IV. RESULTS
This section presents step 5 of the CRISP methodology which was used to build the model. It is simply about evaluating the model by using one or more of the well-known evaluation methods. In order to evaluate the model, the WEKA software was used to calculate the accuracy of the classification model. Two evaluation methods were used, the K – Fold Cross Validation (K-CV) where K= 10 folds and the percentage split method where 66% of the data was used for training and the remainder for testing. Both evaluation methods were used on the ID3 and C4.5 decision tree classification methods. It shows the accuracy of all the classifier s generated using both classification methods and both evaluation methods. As we can see from the table, the resultant classification accuracy from the decision tree model is not very high for the training data used and it varies from one company to another. The reason for such a low accuracy is that the company’s performance in the stock market is affected by internal financial factors such as: news about the company, financial reports, and the overall performance of the market. Also, external factors can affect the performance of the company in the market such as: political events and political decisions. Thus, it can be difficult to have a model that gives a high accuracy classification for all the companies at the same time as the performance of these companies differs.

V. CONCLUSION
This study presents a proposal to use the decision tree classifier on the historical prices of the stocks to create decision rules that give buy or sell recommendations in the stock market. Such proposed model can be a helpful tool for the investors to take the right decision regarding their investments, general economic conditions, and investor expectations influence stock market. As for the future work, there is still big room for testing and improving the proposed model by evaluating the model over the whole companies listed in the stock market. Also, the evaluation of a larger collection of learning techniques such as neural networks, genetic algorithms, and association rules can represent a rich area for future investigation. Finally, reconsidering the factors affecting the behavior of the stock markets, such as trading volume, news and financial reports which might impact stock price can be another rich field for future studying.

VI REFERENCES

