Communications of the IIMA

Volume 15 | Issue 2 Article 5

2017

Competitive Models to Detect Stock Manipulation

Jose J. Thoppan
Saintgits Institute of Management, India

Punniyamoorthy M.
National Institute of Technology – Tiruchirappalli, India

Ganesh K.
McKinsey & Company, Inc.

Recommended Citation

Thoppan, Jose J.; M., Punniyamoorthy; and K., Ganesh (2017) "Competitive Models to Detect Stock Manipulation," *Communications of the IIMA*: Vol. 15: Iss. 2, Article 5.
DOI: https://doi.org/10.58729/1941-6687.1385
Available at: https://scholarworks.lib.csusb.edu/ciima/vol15/iss2/5

Competitive Models to Detect Stock Manipulation

Jose Joy Thoppan
Saintgits Institute of Management, India
jose.thoppan@gmail.com

Dr. M. Punniyamoorthy
National Institute of Technology – Tiruchirappalli, India
punniya@nitt.edu

Dr. K. Ganesh,
McKinsey & Company, Inc.
k_ganesh@mckinsey.com

ABSTRACT

In this paper, data from the Indian stock market is used to study the prediction accuracy of various classification techniques that can be used to identify market manipulation. The data contains information regarding price, volume and volatility of various stocks. Techniques like discriminant analysis, a composite model based on artificial neural network – genetic algorithm (ann-ga) and support vector machine (svm) have been used for classifying stocks into manipulated and non manipulated categories. It is observed that the support vector machine based technique gives the best classification accuracy among the three techniques.

Key Words: Stock Price Manipulation, Support Vector Machine, Artificial Neural Network, Genetic Algorithm, Discriminant Analysis

INTRODUCTION

The price of a stock at any moment in time can be directly attributed to the number of buyers and sellers who want to acquire or part with the stock and the consideration that they are willing to pay. This can simply be equated to the demand and supply of the stock at any one moment. The consideration or the price of the stock adjusts itself to try and attain equilibrium. The stock price then attempts to sustain this equilibrium till such time that there is a change in the balance between the number of buyers or sellers for the security. This could be caused by some external influence (information) affecting the sentiments or an artificial imbalance caused by the activity of a participant. The rate at which the prices move in either direction favouring the buyer or the seller indicates its volatility. The efficient-market hypothesis (EMH) states that markets are inherently efficient with respect to its ability to disseminate information. It considers that, given the information available at the time of investing, no investor can achieve returns that are higher than the average market returns, on a sustained basis.

The volatility witnessed in recent years however is contrary to the postulates of the EMH and there are enough instances to show that markets are not perfectly efficient. This is especially true in the case of emerging markets, where there are few well-informed professional investors. A second theory based on Behavioural Finance states that humans are more often than not irrational in their decisions on buying and selling of securities. The theory attributes this behaviour to fears and misconceptions on the end result that often cause security prices to fluctuate from their rational, fundamental price. However not all fluctuations resulting in high volatility of stock prices can be attributed to irrational decision making by investors.

Recent fluctuations in the stock prices have raised serious concerns about the determination of stock prices, speculative tendencies and most of all illegal market manipulation. A manipulator having significant money power can choose a security, most often illiquid, and engage in sustained heavy buying thus driving up the prices. He might also plant some fraudulent stories in the business press to further his agenda. Once the prices have significantly ramped up, he hopes to sell his entire holdings at these high prices and walk away with a windfall. This is one of the most common types of market manipulation existing in the financial markets, typically the emerging markets.

Most manipulation is detrimental to the trading venue and its participants. Market manipulation impairs price discovery and misrepresents the fair value of a security. The distorted prices force investors to migrate to more efficient markets for deploying their capital. This reduces order flow and increases the cost of trading at a particular trading venue. This further motivates companies to come up with new issue to list their securities on other markets which are better regulated and more efficiently monitored. Hence, ways and means of understanding and eliminating manipulative practices attract great interest from researchers, regulators and exchanges.

Some of the most widely reported incidents on market manipulation include the “The Guinness Four Business Scandal” (BBC News, 2001) and the “The Livedoor Scandal” (TIME, 2006). The well-known scams in the Indian market (Basu and Dalal, 2009) are the twin scams of 1992 and 2001. In 1992, it was Harshad Mehta who, in collusion with Indian banks, businessmen, brokers, foreign banks and mutual funds, orchestrated a false bull market that ended in a meltdown. In an incredible recurrence of history, a different set of banks, brokers, foreign investors and companies connived with Ketan Parekh to produce a sequel which was equal, if not greater, in magnitude to the earlier scam causing a systemic collapse of the Indian Capital Market.

The underlying scenario is not much different today. Though the regulatory environment today is more stringent, the ways markets are controlled by some large players have not really changed. It is observed that even as recent as 2010, about 10 large participants controlled just under 25% of the turnover in the National Stock Exchange of India’s (NSE) cash equity segment. The scenario is graver in the derivatives segment. Here the top 10 participants accounted for about 38% of the turnover and the daily intra-day square off turnover accounted for about 67% of the total turnover during the same period. So it can be inferred that, unlike what is portrayed in the popular media, the Indian capital market remains narrow, shallow and illiquid with the pricing power concentrated in the hands of a few individuals located in a few centres. The Indian capital market is not as vibrant and broad-based as it is made out to be. This observation is not to cast aspersions on the integrity of the market, but to show that it is easy for a rogue trader, who could be among the leading volume contributors, to adversely impact the market efficiency. This also highlights the

need for effective and continuous monitoring of the market activity for early identification of any attempt to undertake a manipulative and illegal trade.

This is not a phenomenon restricted to India alone. Most major global exchanges including NYSE, Bursa Malaysia and Johannesburg Stock Exchange have such skewed figures. This shows that, worldwide, the larger participants have the ability to move the markets to suit their need. This again is not to be interpreted as a feeling of distrust on the credibility of the Global Capital Markets but as a pointer to the possibility of some rouge elements having the potential to destabilise the fair pricing mechanism of the market. Aggarwal and Wu (2006) provide evidence based on SEC actions that potentially informed participants, including corporate insiders, brokers, underwriters, large shareholders and market makers, are likely to be manipulators especially in illiquid securities.

The investors are to be protected from situations that are conducive to manipulations and from rapid fluctuations in stock prices, which can be detrimental to their interest. To achieve these objectives, there is a need for better administrative controls from the regulators. This is to be complemented appropriately by arming the regulators and SRO's with strong electronic safeguards. These safeguards can be in the form of real-time pre-trade risk and real-time market surveillance which will act as logical controls. Effective market surveillance helps the stock exchanges alternate trading destinations and the regulatory organizations spot objectionable situations and aberrant trading behaviour in the capital markets. The surveillance systems help these agencies pursue appropriate preventive/corrective actions against abusive, manipulative, or illegal trading practices. Effective monitoring of markets is achieved by scrutinizing the trading activity using mathematical models which analyze market data to identify potential manipulation.

Market surveillance is necessary to provide a free and fair trading environment to the investing public and assist the regulators and self regulating organizations (SROs) in their activities. Data from stock exchanges and other external systems could also be analyzed on a near real-time basis using these mathematical models to successfully counter the rogue elements that destroy investor confidence in a particular market and enable the determination of fair prices for different securities.

Various literature in the area of market surveillance and trading strategies identifies and applies different market parameters that can be monitored to explain the occurrence of a particular type of market behavior. In this research paper, key market parameters that are influenced by the actions of the manipulators are used as input to the three models to help predict stock market manipulation. The prediction accuracies of these models are compared to identify the best model that can be deployed in a market surveillance system at a regulator and thereby contribute in improving the market's efficiency.

LITERATURE REVIEW

The literature review is organized in multiple parts. The first part of the literature review helps understand the governing principles and definition of a transparent, liquid and efficient market. The next section of the literature review provides the theoretical foundation for the study of market manipulation. These theoretical studies help understand market manipulation better and identify the parameters that get significantly affected when a stock in a market is manipulated. The penultimate section is devoted to identifying the empirical studies done in this area and

understanding how the various theoretical models are applied to carry out the investigation of manipulative situation. The final section helps understand the techniques that are adopted by various researchers for detecting manipulation.

Madhavan (2000) defines market transparency as a market participant's ability to observe trade related data and information including, but not limited to, prices, quotes, volumes, sources and destinations of order flows. Black (1971) describes a liquid market as a continuous market where large amount of stocks can be bought or sold without delay. He further adds that an efficient market is one where small amounts of stocks can be bought and sold at prices very close to the current market price, and large amounts of stocks can be bought or sold over extended periods at prices that, on average, are very near the current market prices, so long as there is no additional material information on the security. He also states that in an efficient market an investor can buy or sell large blocks at a premium or discount depending on the block size.

Efficient-Market Hypothesis (EMH) asserts that markets are inherently efficient or all information (market, public or private) will be reflected in stock prices. It goes on to say that in an efficient market an investor cannot derive excess profit using any kind of information. Arefin and Rahman (2011) tested the EMH for Dhaka Stock Exchange (DSE) for the 2003–2005 period. They have used the excess return market model to confirm that the DSE is not semi-strong form efficient. They have further used Autoregressive Integrated Moving Average (ARIMA) and neural network to test the weak form efficiency, and have concluded that the DSE is not efficient. Tissaoui and Aloui (2011) investigated the dynamics of information flow between stock return and trading volume in the Tunisian Stock Market (TSE). Their results reveal that there is strong evidence of 'lead-lag' linkages in the mean of the return and the variance in volumes in major Tunisian stocks. This shows that the information flow in the TSE follows a sequential rather than a simultaneous process indicating that the market does not have information efficiency as assumed. These two papers together show that there is a possibility that markets are not always efficient and that there is a chance that an insider with superior information can manipulate the market.

The liquid, efficient market that Black describes should follow a firm guideline of market characteristics that meet his assumptions. If there are any aberrations from these normal occurrences, then they could be pointers to potential market manipulation.

Various literature has indicated the possibility of the occurrence of action-based manipulation, information-based manipulation and trade-based manipulation. Insider trading regulations have significantly curtailed the occurrences of action based manipulation. It has also been made difficult to carry out information-based manipulation through various legislation that mandates corporate disclosures. We can conclude that the regulator can be agnostic to the type of manipulation when identifying cases where there is a potential manipulation.

The following articles help us understand how the stocks can be manipulated. This also helps in identifying the parameters that get affected as a consequence of such manipulation.

Kyle and Viswanathan (2008) propose that a trading strategy should not be classified as "illegal price manipulation" unless the violator's intent is to simultaneously undermine both pricing

accuracy and market liquidity. Hence, we consider that price and volume are of the key indicators to identifying potential market manipulation.

Palshikar and Bahulkar (2000) propose that temporal patterns in trade data (which includes trade prices and trade volumes) that repeat themselves when stocks get manipulated can be detected using fuzzy temporal pattern recognition algorithms. This is based on the premise that each type of malpractice leaves a tell-tale trace in the trading databases.

Allen and Gorton (1992) have shown that it is possible for uninformed traders to carry out trade-based manipulation by buying stocks to drive prices up and then selling them at inflated prices to make a quick profit. They consider a trade-based uninformed manipulation model, in which asymmetry created by the buy and sell trades of these noise traders create the possibility of manipulation.

Allen and Gale (1992) also bring out the fact that it is difficult to eradicate trade-based manipulation, as the interested parties could be anybody trading the particular security. They show that great swings in prices or volatility are another key parameter in detecting market manipulation.

A market corner can involve any of the three types of manipulation. Allen, Litov and Mei (2005) studied the market corners during the robber-baron era. One of their key observations is that the price of a stock tends to be discontinuous and, more often than not, accompanied by large price jumps around the corner date, suggesting major disruptions to an orderly market. They also interpret that a market corner is accompanied by manipulation in market price of the stock, significant erosion of liquidity, increase in market volatility and adverse price impact on other assets. All these situations tend to hinder market efficiency.

Comerton-Forde and Putniņš (2011) used a sample of actual closing price manipulation to empirically demonstrate the impact of manipulation on equities. They show that the returns, spreads, trade size, trading activity at the end of the day, and price reversions the following morning, increase significantly for manipulated stocks. They constructed an index to measure the probability and intensity of closing price manipulation and obtained estimates of its classification accuracy.

Aggarwal and Wu (2006) have demonstrated, based on data from the US markets, that manipulation increases stock volatility. They demonstrate that a stock's price goes up during the manipulation period and then reverses direction in the post-manipulation period. They also point out that, most often, prices and liquidity are elevated when the manipulator sells rather than when he buys. This shows that changes in prices, volume and volatility are the critical parameters that are to be tracked to detect manipulation.

Jarrow (1992) investigates and asserts the existence of manipulative strategies in markets where there are large traders. He also goes on to say that the derivatives market could influence the possibility of manipulation in the underlying equity market if the markets are controlled by a group of large traders. As mentioned in the introduction sections and further elaborated in the sections below, both these conditions are true for the Indian stock exchanges.

Various statistical techniques are adopted by researchers to identify manipulations in stock markets. Dr. Longbing Cao of the Data Sciences & Knowledge Discovery Research Lab states that analytical techniques can play significant roles in market surveillance, covering various activities involved, including the identification of benchmarks, detecting abnormal behaviour, extraction of evidence, linkage analysis, case analysis, risk analysis and scenario analysis (Market Surveillance, 2007).

Gaganis, Sochos and Zopounidis (2010) initially identified nine financial ratios to be used in Discriminant Analysis and Neural Networks to classify firms into two categories, namely whether they are in a situation leading to impending bankruptcy or not. The firms under study were all operating in Greece. Palshikar and Apte (2008) further use a graph clustering algorithm to identify clusters that have significant trading among them, to detect circular trading patterns. Both these papers argue that financial and market data has significant information to help detect manipulation.

The SVMs, a classification and prediction technique, Neural Network, Logit and Discriminant Analysis had been applied in applications like bankruptcy prediction by Haardle, et al. (2005) and by Min and Lee (2005) . In these studies, SVM has given better results compared to the other techniques.

In response to the questions pertaining to the skewed trading statistics in the National Stock Exchange (NSE) put forth by the members of the Indian Parliament, Mr. Sukhdev Singh Dhindsa[i] and Mr. Mohammed Adeeb[ii], the minister of state for finance, Mr. Namo Narain Meena, responded that the National Stock Exchange (NSE) had more than 3.09 million clients participating in the cash equity segment during April to June 2010 period. In this period about 52% of the turnover was contributed by retail, high net-worth individuals (HNI), corporate clients etc., while the institutional and proprietary trading contributed 24% each. More than 557,000 clients traded on the Futures & Options (F&O) segments of NSE and here again about 52% of the Exchange turnover was contributed by retail, HNI, corporate clients etc. Institutional clients contributed about 12% and the proprietary traders contributed about 36% of the turnover. It is also observed that about 6% of the participants contribute to 90% of the total traded volume in the Cash Equity segment. The F&O segment, which is seven times the Cash Equity segment, had about 3% of the participants contributing 90% of the volumes. The top 25 trading members of NSE accounted for about 42% and 43% of the Cash Equity and equity stock F&O turnovers respectively during the period.

The data presented above portrays that the largest Indian Stock Exchange, which is a monopoly (96% market share, cash and derivatives put together), has neither depth nor diversity. It also goes on to show that only a small segment of the total population participate and that their participation is more speculative in nature, concentrated on a few stocks and indices like the S&P, CNX, and Nifty. The above data on the trading volumes and patterns in the Indian Exchanges are in line with Jarrow's (1992) assertion that markets that have a combination of equity and equity derivatives, and where trading volumes are concentrated in the hands of large traders, have a greater potential of being manipulated. With this understanding, data is collected from the Indian exchanges that include a set of manipulated and non-manipulated stocks.

The works of various researchers are covered to help define a liquid market and the parameters that help detect manipulation in a liquid market. The various techniques adopted by these researchers to help identify manipulation in stock market are also reviewed. Finally, a market that has the potential to be manipulated, in line with the findings of earlier studies discussed in the literature review, is also identified.

ISSUES

In all the surveyed literature that involves empirical analysis, it is observed that the researchers have used market data from different markets and different periods to investigate the performance of their models. This makes it difficult to carry out an objective comparison of the classification accuracies of each model. To make an objective comparison of the prediction accuracies of these models, they have to be evaluated using the same underlying data. Also, the choice of model for carrying out the studies was not tested for its suitability.

Towards achieving the stated purpose of testing suitability of the models and comparing the performance accuracies, the current research is divided into three parts, each with its own objective. The first objective is to identify the various techniques that are commonly used by researchers in detecting stock price manipulation. The extensive literature survey carried out helps narrow down the three most current and relevant techniques used in detection of financial crime. The second objective is to arrive independently at a comparable value of the predication accuracies of the various techniques. Finally, the last objective is to compare and contrast the results from the three models, to verify and quantify the results from the three models, and identify the model that gives the best prediction accuracy.

METHODOLOGY

In a market like India where there are more than 5000 securities listed on its major exchanges, it becomes difficult to monitor all of them for potential market abuse. Researchers have increasingly adopted a variety of statistical techniques to develop newer and effective models for detecting stock price manipulation. Based on the literature survey, Discriminant Analysis and Support Vector Machines were identified as two of the most popular techniques adopted by researchers. Additionally, Punniyamoorthy and Thoppan (2012b) used a hybrid model using Artificial Neural Network and Genetic Algorithm.

In this paper, the three models are analyzed on the same underlying dataset to find if a particular stock is witnessing abnormal activity indicative of manipulation. The three models help categorize stocks into two categories, namely manipulated and non-manipulated, to help investigators arrive at a shortlist of potentially manipulated stocks which could be taken up for further detailed investigation. Each of the model and the method for arriving at the results can then be compared one against the other using a confusion matrix as elaborated in the subsequent section. All the models use the same data set collected from the Indian Stock exchanges for the analysis so that the comparison of the results is possible.

Data

The trade data comprising manipulated and non-manipulated securities traded on the Indian Equity Exchanges were collected for the study. The manipulated securities were identified based on the adjudication orders passed by Securities and Exchange Board of India (SEBI, 2011). SEBI had identified manipulations in the Indian stock markets on various securities in multiple periods during 2003 to 2009. The adjudication orders were passed after investigating these incidents. These adjudication orders had pointed out instances of stock price manipulation in the Indian stock markets on different securities, along with the period in which they were manipulated.
The data collected, comprised 30 securities traded on the two leading Indian Exchanges, namely the National Stock Exchange (NSE) and the Bombay Stock Exchange (BSE). This data formed the foundation on which the research was carried out. Of these, 15 securities belonged to the manipulated group, i.e., the prices of these securities were known to have been manipulated during the period 2003 to 2009. For the same manipulation periods in which the prices of the 15 securities were rigged, data was collected for 15 securities of companies with comparable size and from the same industrial group. This constituted the second group, namely the non-manipulated group.

Based on previous studies by eminent researches, as indicated in the literature review, mainly from the works of Allen and Gale (1992), Allen et al. (2006), Aggarwal and Wu (2006) & Kyle and Vishwanathan (2008), a shortlist of attributes for equity stocks that were affected by the actions of a manipulator was identified. The attributes chosen were price, volume and volatility of individual securities. The average closing price, average trading volume and the variance in the stock price (volatility) for each company's stock during the study period were taken as the input data.

The list of the securities and the number of days for which they were manipulated during the 2003 to 2009 period is as follows: Thermax (38 days), Rajesh Exports (30 days), Geojit BNP Paribas (147 days), ABB (118 days), Bosch (30 days), Nahar Spinning (30 days), JM Financials (114 days), Arvind (47 days), United Spirits (37 days), Aarvee Denim (80 days), Aban Offshore (17 days), Sriram Transport Finance Corporation (23 days), DIC India (60 days), Peninsula Land (89 days) and Kwality Diary (134 days). To make an equitable comparison of different securities, each having a dissimilar duration of the manipulation period, the shortest manipulation period (17 days) was taken. This formed the initial data set. From this raw data, the average closing price, average trading volume and the variance in the stock price (volatility) for each company's stock during the study period was computed. This data was then taken as the input data for each of the three models developed.

Discriminant Analysis

One of the popular statistical techniques to classify data into two or more groups is through the discriminant analysis (Gaganis, Sochos and Zopounidis, 2010). The most common discriminant analysis method adopted by researchers to detect financial frauds using discriminant analysis is the Linear Discriminant Function.

Punniyamoorthy and Thoppan (2012a) have tested for the assumptions governing the use of the Linear Discriminant Function. The assumptions are that the data should be normally distributed (verified using the Q-Q Plot) and that the two groups should have equal variance-covariance matrices (tested using the Box's M Test). For the data collected, the normality was established, however the variance–covariance matrices for the two sets of data, namely manipulated and non-

manipulated was not the same. Thus, it was identified that the most commonly used classification technique was not appropriate for the data collected from the Indian exchanges.

It is identified that the Generalized Squared Distance Function, otherwise called the Quadratic Discriminant Function (Rencher, 2002), helps to preserve optimality even in cases where the variance – covariance matrices of the different groups are not equal. Hence, it is considered as the most appropriate technique to evaluate stock market data for potential manipulation.

The Quadratic Discriminant Function stated as below can be used for classifying the stocks into two categories by assigning a stock "y" to the group for which the value of $L_i(y)$ is the maximum.

$$L_i(y) = Ln\ P_i - \ 0.5 \ln |S_i| - \ 0.5\ (y - \overline{y_i})' S_i^{-1} (y - \overline{y_i})$$

Where,

P_i - Prior probabilities

S_i - Variance Covariance Matrix of the variables

The input to the above function will be the data for which the normality is established, but not necessarily having equal covariance matrices. The prior probabilities are assumed to be P_1, P_2, P_3 …P_i. In order to use a Quadratic Discriminant Function based on the covariance matrix, the number of observations in each group 'n_i' must be greater than 'p', the number of variables, so that the inverse of the covariance matrix will be present.

ANN-GA-based Hybrid Model

In this section we present the genetic algorithm-based neural network model to classify a known sample of stocks from the Indian capital market into either the manipulated or non-manipulated category. Further, the model's prediction accuracy is also analyzed.

Genetic algorithms (Holland, 1992) are a group of robust seeking, adjusting and optimizing techniques developed by Holland. In a genetic algorithm problem a potential solution set is arrived at through a natural selection over multiple generations through recombination. The recombination can be achieved using crossover and mutation operators. The fitness of each solution is evaluated and a better solution set is obtained. The crossover operation causes a controlled, yet unsystematic, exchange of inherited characteristics between solutions, under the premise that a 'good' parent will generate 'better' offspring. The final population provides a collection of solution candidates, one or more of which can be applied to the original problem. The optimal solution arising out of the genetic algorithm is used as the weights in a neural network

Artificial neural networks are a class of machine learning algorithms inspired by the way the nervous system of a human body functions. Similar to the nervous system, the machine learning algorithms use a network of computing units called neurons having input layer, multiple hidden layers and output layer. The weight in a neural network indicates the strength of the association between two neurons. Unlike the conventional neural network that uses steepest decent or back propagation, these coefficients or weights in the model are estimated using the genetic algorithm as described in the composite model proposed by Punniyamoorthy and Thoppan (2012b).

The model uses a single hidden layer. The calculations are carried out at three layers, namely input, hidden and output layers of the neural network. The network first computes the output of the input layer. The output of the input layer neurons is equal to the input of the input layer neurons. Next, the inputs of the hidden layer neurons are computed. This will be computed by multiplying the weights of synapses connecting input neurons and hidden neurons with the output of the input layer. In the hybrid model that is employed, the weights are obtained by using the genetic algorithm. The output of the hidden layer neurons will be calculated by sigmoidal activation function.

$$f(x) = \frac{1}{1 + e^{-I_H}}$$

Where,
I_H - The input to the hidden layer

The input of the output layer neurons is computed by multiplying the weights of synapses connecting hidden neurons and output neurons with the output of the hidden layer neurons. Then the output of the output layer neurons is calculated by sigmoidal activation function. If the output gives a value greater than zero, it is categorized as manipulated and if less than zero, it is categorized as non-manipulated.

Support Vector Machines

SVM can be defined as a method for creation of an optimal hyperplane in a multi dimensional space such that the hyperplane separates the two categories and has the lowest possible misclassification error (Burges, 1998). The hyperplane has the lowest misclassification error when it has the largest possible margin between the hyperplane and the nearest plot in the training set on either side of the hyperplane. Such a hyperplane can be called the maximum-margin hyperplane. SVMs are used to classify a security as 'manipulated' or 'non manipulated', based on the learning algorithm's ability to be trained on complex patterns and characteristics of interest that define the securities in the training set and recognise similar patterns in the observed variables of the security under investigation. SVM's can be broadly classified into three types, namely linearly separable classifier, linear soft margin classifier and nonlinear classifier. The nonlinear classifier is used in the model because the stock market data is so random that a linear classifier will not be able to classify the data into two groups.

In the cases of data like stock market or other financial data, the groups are not only overlapping but there is a genuine separation function which can be nonlinear hyperplanes or surfaces. The nonlinear separation hyperplane is used to separate the data in the training set with almost no error. For this, the kernel function is adopted. There are four popular kernel functions. They are the linear kernel function, the polynomial kernel function, the radial basis function and the sigmoid kernel function (Hsu *et al*., 2004). These functions can be expressed as shown below:

Linear Kernel Function $K\left(x_i, x_j\right) = x_i^T x_i$

Polynomial Kernel Function $K\left(x_i, x_j\right) = (\gamma x_i^T x_i + r)^d$

Radial Basis Function $K\left(x_i, x_j\right) = \exp(-\gamma \left|\left|x_i - x_j\right|\right|^2)$

Sigmoid Kernel Function $K\left(x_i, x_j\right) = \tanh\{\gamma x_i^T x_j + d\}$

There is no established technique for determining the best kernel function. However, it has been observed by researchers that SVMs with RBF kernel gives better results than those obtained using the linear kernel function and polynomial kernel function (Keerthi and Lin, 2003). Also, the sigmoid kernel behaves like RBF for certain parameters (Lin and Lin, 2003). Hence, the radial basis function (RBF) is adopted over the other kernel functions in the model for classifying the securities into manipulated and non-manipulated categories.

The dual form of the decision function for an SVM (Vapnik, 1998), can be stated as,

$$min_\alpha \quad L_D = \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{n} \alpha_i$$

Such that,

$0 \leq \alpha_i \leq C\,, \quad i = 1, \ldots, n$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$

Solving this equation can help define the maximum margin hyperplane that will separate the data into two categories. This optimal hyperplane that separates one class from the other will help in the classification decision through the following equation:

$$f(y) = sign\,(\sum_{i=1}^{sv} \alpha_i y_i k(x, x_{sv}) + b)$$

In the above equation, $k(x,x_{sv})$ is a kernel function denoting a nonlinear classifier, and 'sv' is the number of support vectors. The RBF kernel function is adopted and the equation is rewritten as explained by Punniyamoorthy and Thoppan (2012c).

$$f(y_j) = sign\,(\sum_{i=1}^{sv} \alpha_i y_i e^{-\gamma||x_j - x_i||^2} + b)$$

The calculated value of α and the corresponding value of γ along with the values of x and b, are substituted to get the resulting value. Depending on the sign of the result we could classify the data as belonging to either the manipulated or non-manipulated category.

Confusion Matrix

Once the stocks are categorized into two groups based on each of the above techniques, their prediction accuracies can be compared by quantifying the error that would creep in. To estimate the classification accuracy of the functions, the result obtained from the models are compared

using the misclassification table or the confusion matrix for each of the models. This gives a method to arrive at the comparable value for the results of each of the models.

The redistribution method is adopted to estimate the misclassification. The proportion of misclassifications that is obtained after redistribution is tabulated in the misclassification table or confusion matrix as shown in Table 1.

Table 1: Misclassification table.

		Predicted Groups		
		Group 0	**Group 1**	**Total**
Actual Groups	**Group 0**	X_1	X_2	$\Sigma X = X_1 + X_2$
	Group 1	Y_2	Y_1	$\Sigma Y = Y_1 + Y_2$
	Total	$X_1 + Y_2$	$X_2 + Y_1$	$\Sigma X + \Sigma Y$

Separate confusion matrices are drawn up to provides a visual representation on the classification efficiency of discriminant analysis, support vector machines and ANN-GA-based model in their ability to classify the data. Based on the values obtained, the most efficient method for classifying the given data set can be identified.

RESULTS

The results that were obtained from carrying out the analysis based on the three techniques mentioned earlier are described in this section. For the quadratic discriminant analysis, MATLAB was used to perform the analysis. For the ANN-GA-based model, a system prototype was developed using Microsoft .Net framework. The application was developed using C#.NET 2.0 and SQL Server 2005 and the experiments are run on PCs with Intel Core 2 Duo 2.53GHz CPU and 4 GB memory. For the support vector machine, DTREG was used to carry out the analysis.

Discriminant Analysis

To ensure that the data is multivariate normal, the Q-Q plot is used to remove the outliers and arrive at a processed dataset. Both the χ^2 and the F approximation test return values rejecting the hypothesis that the variance-covariance matrix of the two groups is the same. Since this is in violation of the assumption for linear discriminant function, the appropriate technique, quadratic discriminant function or the generalized squared distance function is adopted.

In the generalized squared distance function, the sample variance-covariance matrix S_1 and S_2 for each of the two groups, namely manipulated and non-manipulated, are used to form the quadratic equations, $L_1(y)$ and $L_2(y)$. The 'y' values are then substituted in each of the two equations.

The prior probability P_i for each observation is assigned a value 0.5. Since it is assumed that there is an equal probability that a stock be manipulated, the '$\ln P_i$' can be dropped from the above equation. The resultant two equations are as below:

$$L_1(y) = 0.5 \ln|S_1| - 0.5\,(y - \bar{y}_1)' S_1^{-1} (y - \bar{y}_1) \text{ and}$$
$$L_2(y) = 0.5 \ln|S_2| - 0.5\,(y - \bar{y}_2)' S_2^{-1} (y - \bar{y}_2)$$

Here, y represents an array containing data on the price, volume and volatility of the stocks. The value of $L_i(y)$ for i=1, 2 is calculated. 'y' is then allocated to the group for which $L_i(y)$ is maximum.

The attributes of the stocks in this dataset are then substituted into the above equation. Once the results are obtained, a misclassification table is drawn up for the above result. The confusion matrix created for the quadratic discriminant analysis is as shown in Table 2.

Table 2: Misclassification Table – QDF.

		Predicted Groups		
		Group 0	**Group 1**	**Total**
Actual Groups	**Group 0**	81.81%	18.18%	100%
	Group 1	36.36%	63.63%	100%

In the above table, the manipulated stocks are indicated as 'Group 0' and the non-manipulated are marked as 'Group 1'. It is observed that the model is able to identify 81.18% of the manipulated and 63.63% of the non-manipulated stocks correctly, whereas the remaining was misclassified. The error estimate on misclassification of the non-manipulated group is 18.22% and the manipulated group is 36.36%. The combined error rate in the model's ability to classify the stocks as manipulated and non-manipulated, using the quadratic discriminant function is 27.27%.

ANN-GA-based Hybrid Model

For the determination of weights of the neural network, a genetic algorithm-based model is used. The outputs of the genetic algorithm are then directly taken in as the input for the neural network. The chromosomes are initialized using random numbers. Over multiple generations, by evaluating the fitness functions, a solution set with about 95% of the chromosomes being identical is achieved. This is then used to determine the weights of the neural network. These weights are then used to evaluate the data to determine the prediction accuracy of the neural network in classifying stocks as manipulated or non-manipulated. The misclassification table is as below in Table 3.

Table 3: Misclassification Table – ANN-GA.

		Actual Grouping		
		Group 0	**Group 1**	**Total**
Predicted Groups	**Group 0**	80%	20%	100%
	Group 1	26.66%	73.33%	100%

A confusion matrix as shown above provides a visual representation of the classification efficiency of the hybrid model in its ability to classify the data into manipulated and non-manipulated. From the above confusion matrix it is observed that using the ANN-GA-based model, the error estimate of the misclassification of non-manipulated stock is 20% and the manipulated group is 26.66%. The combined error rate in the model's ability to categorize the stock as manipulated and non-manipulated using the ANN-GA model is 23.33%, which is an improvement over the QDF model.

Support Vector Machines

A support vector machine-based learning algorithm model has been explained for the classification of the companies as manipulated and non-manipulated. The training set of stocks from the Indian Capital Market is collected in the form $T = (x_i, y_i)$, where the p dimensional vector $x_i \in R^p$ indicates the stocks and their attributes. Each stock that was considered had three parameters defining them, namely the price, volume and volatility of the securities under study. The variable y_i indicates the category to which every stock x_i belongs, i.e. $y_i \in \{-1, +1\}_{i=1}^{n}$. To identify the manipulated securities, the given data was categorized into two sets, namely manipulated and non-manipulated. The objective is to categorize all data points having $y_i = -1$ as manipulated and all data point having $y_i = +1$ as non-manipulated, thus forming two distinct groups.

The parameters of the maximum-margin hyperplane are derived by solving the optimization. There exist several specialized algorithms for quickly solving the QP problem that arises from SVMs, mostly reliant on heuristics for breaking the problem down into smaller, more manageable chunks.

As mentioned earlier, there are three predictor variables, and the radial basis function is used as the SVM kernel function to categorize the stocks as manipulated and non-manipulated. The search criterion was to minimize the total error. The values for C and γ were to be in the range of from 0 to 1000 and -100 to 100 respectively. Analyzing the data with DTREG software, the following results were found. For the radial basis function, the optimum value of C was arrived at as 322.54 and γ was 8. The minimum error found by the search was 0.2068965. The number of support vectors used in the model was 12.

Table 4: Confusion Matrix – SVM.

		Actual Grouping		
		Group 0	**Group 1**	**Total**
Predicted Groups	**Group 0**	93.33%	6.66%	100
	Group 1	33.33%	66.66%	100

In the above Table 4, the input data consisted of manipulated stocks, indicated as ‘Group 1’ and non-manipulated, marked as ‘Group 0’. It was observed that the SVM-based model is able to identify 93.33% of the non-manipulated and 66.66% of the manipulated correctly, whereas the remaining is misclassified. The error estimate on misclassification of the non-manipulated group is 6.66% and the manipulated group is 33.33%. The combined error rate in the model’s ability to classify the stocks as manipulated and non-manipulated using the SVM is 20.00%, thus providing a better result than both the earlier models.

CONCLUSION

Manipulation of stock prices has adverse impact on investor sentiments and returns. This is detrimental to the viability of the trading venues. Regulators are under pressure to effectively regulate the marketplace. The detection models employed in this paper helps identify stocks that are witnessing activities indicative of potential manipulation, irrespective of the type of manipulation – action-based, information-based or trade-based.

The confusion matrices drawn up for the QDF model, ANN-GA-based composite model and SVM model show that the QDF-based model has a prediction accuracy of 72.73% or a misclassification error of 27.27%. The ANN-GA-based model had a prediction accuracy of 76.66% or a misclassification error of 23.33%, whereas using the SVM-based model, we were able to get the best prediction accuracy among all three models, which was 80.00% or a misclassification error of 20.00%. The result obtained by using SVM is significantly better than the result from the discriminant analysis-based model and the composite ANN-GA-based model.

From the above results it can be observed that the quantitative models can help detect stock price manipulation in the Indian market with a great level of prediction accuracy. These models can be deployed at the regulator or at exchanges which are the frontline regulators to the market. It can be concluded that, SVM is the most efficient of the three techniques that can be incorporated as a part of an effective market surveillance system to help identify proscribed transactions or anomalous trading behavior in the stock exchanges and other trading venues. It helps scrutinize the trading activity by analyzing large amount of stock market data to identify atypical situations indicative of market manipulations. The implementation of such a system will act as a strong deterrent to potential manipulators, improve investor confidence in a particular market and enable the determination of fair prices for different securities.

In a future study, the model's performance can be studied using data from different companies and from different markets to test if the model's prediction accuracy results can be generalized to the global markets and across asset classes. Additional variables like spreads, trade sizes, price reversions, P/E ratio, EPS, free float, liquidity, number of trades etc., can also be identified which could better identify the patterns and thus possibly reduce the misclassification error giving better prediction accuracy. An attempt can also be made to increase the number of stocks taken for the study as the current sample size is relatively small.

REFERENCES

Aggarwal, R.K., & Wu, G. (2006). Stock market manipulation - theory and evidence. *Journal of Business*, 79(4), 1915–1953.

Arefin, J., & Rahman, R.M. (2011). Testing different forms of efficiency for Dhaka Stock Exchange. *International Journal of Financial Services Management, 5*(1), 1–20.

Allen, F., & Gale, D. (1992). Stock price manipulation. *Review of Financial Studies, 5*(3), 503–529.

Allen, F., & Gorton, D. (1992). Stock price manipulation, market microstructure and asymmetric information. *European Economic Review, 36*(2-3), 624–630.

Allen, F., Litov, L., & Mei, J. (2005). Large investors, price manipulation, and limits to arbitrage: an anatomy of market corners. *Western Finance Association, Annual Meeting*, June 18-21, 2005, Portland, Oregon.

Basu, D. & Dalal, S. (2009). *The Scam - from Harshad Mehta to Ketan Parekh*, 3rd ed., Mumbai, India: KenSource Information Services P. Ltd.

BBC News (2001). Guinness Four fail in fight for acquittal. Retrieved from *http://news.bbc.co.uk/2/hi/business/1723136.stm*

Black, F. (1971). Towards a fully automated exchange. *Financial Analysts Journal*, *1*(27), 29-34.

Burges, C.J.C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121-167.

Comerton-Forde, C., & Putnins, T.J. (2011). Measuring closing price manipulation. *Journal of Financial Intermediation*, *20*(2), 135-158.

Gaganis, Ch., P. Sochos & C. Zopounidis (2010). Bankruptcy prediction using auditor size and auditor opinion. *International Journal of Financial Services Management*, *4*(3), 220-238.

Haardle, W., Moro, R., & Schafer, D. (2005). Predicting bankruptcy with support vector machines. *Humboldt University and the German Institute for Economic Research*, Retrieved from http://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2005-009.pdf.

Holland, J.H. (1992), *Adaptation in natural and artificial systems*, 2nd ed. Cambridge. MA: MIT Press.

Hsu, C.W., Chang, C.C. & Lin, C.J. (2004). A practical guide to support vector classification *Technical Report, Department of Computer Science and Information Engineering, National Taiwan University*. Retrieved from http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

Jarrow, R. (1992). Market manipulation, bubbles, corners, and short squeezes. *Journal of Financial and Quantitative Analysis*, *27*(3), 311–336.

Keerthi, S.S. & Lin, C.J. (2003). Asymptotic behaviours of support vector machines with gaussian kernel. *Neural Computing*, *15*(7), 1667–1689.

Kyle, A. & Viswanathan, S. (2008). How to define illegal price manipulation. Retrieved from w4.stern.nyu.edu/finance/docs/pdfs/Seminars/081m-kyleviswanathan.pdf

Lin, K.M. & Lin, C.J. (2003). A study on reduced support vector machines. *IEEE Transactions on Neural Networks*, *14*(6), 1449–1559.

Madhavan, A. (2000). Market microstructure: a survey. *Journal of Financial Markets, 3*(3), 205-258.

Market Surveillance (2007). What is market surveillance? Retrieved from www.marketsurveillance.org on February 26, 2018.

Min, J. H., & Lee, Y. C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*, *28*(4), 603–614.

Palshikar, G.K., & Apte, M.M. (2008). Collusion set detection using graph clustering. *Data Mining and Knowledge Discovery, 16*(2), 135 – 164.

Palshikar, G.K., & Bahulkar, A. (2000). Fuzzy temporal patterns for analysing stock market databases. *Proceedings of the International Conference on Advances in Data Management (COMAD-2000),* at Pune, India, Tata-McGraw Hill, pp. 135-142, December 14-16.

Punniyamoorthy, M., & Thoppan, J.J. (2012a). Detection of stock price manipulation using discriminant analysis. *International Journal of Financial Services and Management, 5*(4), 369-388.

Punniyamoorthy, M. & Thoppan, J.J. (2012b). ANN-GA based model for stock market surveillance. *Journal of Financial Crime, 20*(1), 52-66.

Punniyamoorthy M. & Thoppan, J.J. (2012c). Detecting stock price manipulation. (unpublished)

Rencher, A.C. (2002). *Methods of multivariate analysis, 2nd ed.* Hoboken, NJ: John Wiley & Sons.

SEBI Adjudication Orders [Online], Retreived from http://www.sebi.gov.in/Index.jsp?contentDisp=SAT

TIME (2006). The livedoor scandal: tribe versus tribe. Retrieved from *http://www.time.com/time/world/article/0,8599,1151722,00.html*

Tissaoui, K. & Aloui, C. (2011). Information flow between stock return and trading volume: the Tunisian stock market. *International Journal of Financial Services Management, 5*(1), 52–82.

Vapnik, V. (1998). *Statistical learning theory*, 1st edition. New York, NY: Springer Publishing.

[i] This is based on the response given by Mr. Namo Narain Meena (Minister of State for Finance, Government of India) in response to a question raised by Mr. Sukh Dev Dhindsa in the Rajya Sabha (Upper House of Parliament). This question and response is available at http://164.100.47.4/newrsquestion/Search_QnoWise.aspx, Session 220, Question number 1669, question type – unstarred (10 August 2010), "Trading in NSE".

[ii] This is based on the response given by Mr. Namo Narain Meena (Minister of State for Finance, Government of India) in response to a question raised by Mr. Mohammed Adeeb in the Rajya Sabha (Upper House of Parliament). This question and response is available at http://164.100.47.4/newrsquestion/Search_QnoWise.aspx, Session 220, Question number 1692, question type – unstarred (10 August 2010), "Trading turnover of top companies in NSE".