CHAPTER-1

INTRODUCTION

Fraud Detection has great importance to Financial Institutions. The proposed research work is concerned with the problem of Fraud Detection in Stock Market Using Outlier Analysis.

To find out fraud [62,64-65,86] in the Stock Market is very difficult task when it is being done using traditionally data analysis methods from the huge data. For example suppose we have to study and analyze the whole trading data of a very old company which is being traded in the stock market from last 35 years, then we have to analyze around more than 8700 days trading data of that company and further if we compare the data of this company with 9-10 more companies of similar group of companies for the same period then this will become a very difficult and also a very time consuming job which also require lot of manpower. Now suppose what the critical and cumbersome situation will be there when we are asked to analyze the trading data of all the companies listed in BSE and NSE for last many years and to detect fraudulent activities from this data?

So, to come out from the above stated problem, we can use “Data Mining” [36], a very important and powerful technology of Computer Science, with great potential to extract the hidden predictive information from large databases, for detection of fraud in Stock Market.
1.1 Fraud

The Oxford English Dictionary [92] defines fraud as —wrongful or criminal deception intended to result in financial or personal gain.

According to section 17 and 18(Indian Contract Act, 1872), ‘Fraud’ (Gulshan S.S, Kapoor G.K, 2013) [109] means and includes any of the following acts committed by a party to a contract (or with his connivance or by his agent) with intent to deceive another party thereto or his agent; or to induce him to enter into the contract:

(i) the suggestion, as a fact, of that which is not true by one who does not believe it to be true;

(ii) the active concealment of a fact by one having knowledge or belief of the fact;

(iii) a promise made without any intention of performing it;

(iv) any other act fitted to deceive;

(v) any such act or omission as the law specially declares to be fraudulent.

In our proposed research work we have studied various possibilities of frauds happening in Stock Market like price manipulation, publishing fake balance sheets, false news like bonus, dividend, takeovers, merger, new orders, expansions etc to arouse people to buy their company shares at unfair price.
1.2 Fraud Detection

Fraud detection is known as identifying fraud as quickly as possible when it has been perpetrated. Once fraud prevention fails, Fraud detection comes into play. Fraud detection must be used continually, because one may be unaware that fraud prevention has failed. Fraud detection is a continuously evolving process. Whenever criminals come to know that one detection method is in place, they will change their strategies and try others. Day by day, new criminals are entering in this field. Many of them may not be aware of the fraud detection methods which were successful in the past and will adopt strategies which can lead to identifiable frauds. This means that with latest developments, the earlier detection tools need to be applied as well. This is difficult to develop new fraud detection methods because exchange of ideas in fraud detection is very limited. It may be harmful to describe fraud detection techniques in detail in the public domain, as the criminals can take advantage of it. Data sets are not made available and results are often censored, making them difficult to assess (e.g., Leonard, 1993) [77].

According to G.K. Palshikar (2002) [34], it is very difficult to detect fraud using traditional methods of data analysis which require complex and time-consuming investigations and require knowledge of different domains like financial, economics, business practices and law. Fraud often may consist of many instances or incidents which can involve repeated transgressions
using the same method. Fraud instances may be similar in content and in appearance but usually they are not identical.

According to **Decker (1998)** [19], first of all, the telephony companies, the insurance companies and the banks used data analysis techniques to prevent fraud. One earlier example of successfully implementation of data analysis techniques in the banking industry is the Falcon fraud assessment system, which was based on a neural network shell. Retail industries also suffer from fraud at Point of Sale (POS). Today almost all supermarkets are using digitized Closed-Circuit Television (CCTV) with POS data of most susceptible transactions to fraud. Frauds like cell phones, insurance claims, tax return claims, credit card transactions etc. are creating sever problems for governments and businesses because yet detection and prevention of fraud is not a simple task. Fraud is known as adaptive crime, so it requires some special methods of intelligent data analysis to detect and prevent it. These methods existing in the areas of Knowledge Discovery in Databases (KDD), Data Mining, Machine Learning and Statistics because they are offering applicable and successful solutions in different areas of fraudulent crimes.

Our proposed research work is concerned with Stock Market Fraud Detection. In stock market, fraudster always come up with new ideas to trap unaware or new investors by their various newer fraudulent activities every day so fraud detection is very essential.
1.3 Various Techniques for Fraud Detection

From literature review (Chapter-2) we have studied following techniques for fraud detection-

(i) Logistic Model

(ii) Neural Networks

(iii) Bayesian Belief Networks

(iv) Decision Trees

(v) Fuzzy Logic

(vi) Genetic Algorithms

(vii) Support Vector Machines

(viii) Game Theory

(ix) Peer Group Analysis

(i) Logistic Model

Logistic Model [59,113] is a generalized linear model that is used for binomial regression in which the predictor variables can be either numerical or categorical. It is principally used to solve problems caused by automobile insurance and corporate fraud. Some empirical studies have used Logistic Model as a benchmark in Fraud Detection [10,78,99].

(ii) Neural Networks

The neural network [59,105] is a technique that imitates the functionality of the human brain using a set of interconnected vertices.
It is widely applied in classification and clustering, and its advantages are as follows. First, it is adaptive; second, it can generate robust models; and third, the classification process can be modified if new training weights are set. Neural networks are mainly applied to credit card, automobile insurance and corporate fraud. Some empirical studies have used Neural Networks as a benchmark in Fraud Detection [8,12-13,27-28,30,35,68,70,72-73,78,82,87,96,98,101,117,121,141].

(iii) Bayesian Belief Network

The Bayesian Belief Network (BBN) [61,70] represents a set of random variables and their conditional independencies using a directed acyclic graph (DAG), in which nodes represent random variables and missing edges encode conditional independencies between the variables. The Bayesian belief network is often adopted in credit card, automobile insurance, and corporate fraud detection. Some empirical studies have used Bayesian Belief Network as a benchmark in Fraud Detection [26,70,73,82,117,141].

(iv) Decision Trees

Decision Tree [60,63] is a flowchart like tree structure in which each internal node denotes a test on an attribute, each branch represents an outcome of the test, each leaf node holds a class label and the topmost node in a tree is the root node. Decision Trees are predictive decision support tools that create mapping from observations to
possible consequences. Some empirical studies have used Decision Trees as a benchmark in Fraud Detection [70,73].

(v) Fuzzy Logic

Fuzzy Logic [5,33,67,74,79,93] is a multi-valued logic which is derived from fuzzy set theory to deal with the reasoning that is approximate rather than precise, specifically, fuzzy logic is a superset of conventional Boolean Logic that is extended to handle the concept of partial truth, i.e., the intermediate truth values between conventional evaluations, “completely true” and “completely false” such as yes or no. In Boolean Logic variables may have a membership value of only 0 or 1. Similar to Boolean Logic, Fuzzy Logic has three basic operations: (i) intersection, (ii) union, and (iii) complement. Fuzzy Logic systems are based on set of rules. These rules allow the input to be fuzzy, i.e., more similar to the natural way that human express knowledge. Fuzzy Logic has been used in various applications some of them are automobile and vehicle subsystems, air conditioners, cameras, image processing, elevators, washing machines, rice cookers, dishwashers, video games, speech recognition, and pattern recognition. Some empirical studies have used Fuzzy Logic as a benchmark in Fraud Detection [4,14,21,75-76,97].
(vi) Genetic Algorithms

Genetic Algorithms [63] attempt to include ideas of natural evolution. Generally genetic learning starts as follows- an initial population is created consisting of randomly generated rules and each rule can be represented by a string of bits. In Genetic Algorithms, populations of rules “evolve” via operations of crossover and mutation until all rules within a population satisfy a specified threshold. Some empirical studies have used Genetic Algorithms as a benchmark in Fraud Detection [40, 69, 120].

(vii) Support Vector Machine

A Support Vector Machine [63] is an algorithm which is used for the classification of both linear and non linear data and transforms the original data in a higher dimension from where it can find a hyperplane (i.e. a “decision boundary” separating the tuples of one class from another) for separation of the data using essential training tuples called support vectors. Some empirical studies have used Support Vector Machine as a benchmark in Fraud Detection [22, 98, 101, 121].

(viii) Game Theory

According to Turocy, T. L. (2001) [118], Game Theory is the formal study of conflict and cooperation and concepts of Game Theory are applied whenever the actions of several agents are interdependent.
These agents can be individuals, groups, firms, or any combination of these. The concepts of game theory provide a language to formulate, structure, analyze, and understand strategic scenarios. Some empirical studies have used Game Theory as a benchmark in Fraud Detection [94,116].

(ix) Peer Group Analysis

Peer Group Analysis (Bolton & Hand, 2001) [103], is a term that have been coined to describe the analysis of the time evolution of a given object (the target) relative to other objects that have been identified as initially similar to the target in some sense (the peer group). Peer Group Analysis [134] detects individual objects (target) that begin to behave significantly different from the other objects to which they had previously been similar (peer group). That is, we detect the abnormal behavior of a target by comparing it with its peer group members and measuring the deviation of its behavior from the peer group. The advantage of this approach is that we can find local outliers which cannot otherwise be detected when compared to the whole population. Some empirical studies have used Peer Group Analysis as a benchmark in Fraud Detection [11,134,140].

According to G.K Palshikar (2002) [34], techniques used for fraud detection fall into two primary classes- (i) Statistical Techniques and (ii) Artificial Intelligence. Statistical fraud detection methods may be of two
types-supervised or unsupervised. Samples of both fraudulent and non-fraudulent records are taken to construct models which will allow one to assign new observations into one of the two classes, in the supervised methods. This will require one to be confident about the true classes the original data used to build the models. It will also require that one has examples of both classes. It can only be used to detect those frauds which have previously occurred. In other hand, unsupervised methods find those accounts, customers and so forth which are most dissimilar from the norm. Further these can be examined more closely. Outliers are a basic form of nonstandard observation. Tools which are used for checking data quality can be used, but the detection of accidental errors is different problem from the detection of intentionally falsified data or data which accurately describes a fraudulent pattern.

**Examples of statistical data analysis techniques for fraud detection are:**

(i) Data preprocessing techniques which may includes detection, validation, error correction, and filling up of missing or incorrect data.

(ii) Calculation of various statistical parameters which may include averages, quantiles, performance metrics, probability distributions etc. For instance, the averages may include average length of call, average number of calls per month and average delays in bill payment in detection of telecom fraud.
(iii) Models and probability distributions of various business activities which may be either in terms of various parameters or probability distributions.

(iv) Computing user profiles.

(v) Time-series analysis for time-dependent data.

(vi) Clustering and classification for finding patterns and associations among the groups of data.

(vii) Matching algorithms for detection of anomalies in the behavior of transactions or users as compared to previously known models and profiles.

**Examples of AI techniques for fraud detection are:**

(i) Data Mining to classify, cluster, and segment the data and automatically find associations and rules in the data that may indicate interesting patterns related to fraud.

(ii) Expert Systems for encoding expertise for detection of fraud in the form of rules.

(iii) Pattern Recognition for detection of approximate classes, clusters, or patterns of suspicious behavior either automatically or to match the given inputs.

(iv) Machine Learning techniques for automatically identification of the characteristics of fraud.
(v) Neural Networks which can learn suspicious patterns from samples and can be used later to detect them.

According to E.W.T Ngai et al. (2011)[24], the main data mining techniques used for Financial Fraud Detection are Logistic Models, Neural Networks, the Bayesian Belief Network, and Decision Trees. They specified six classes of data mining techniques for financial fraud detection in their research work which are as follows-

(i) Classification
(ii) Regression
(iii) Clustering
(iv) Prediction
(v) Outlier Detection
(vi) Visualization.

(i) Classification

According to Han and Kamber [63], Classification builds up and utilizes a model to predict the categorical labels of unknown objects to distinguish between objects of different classes. These categorical labels are predefined, discrete and unordered. According to Zhang and Zhou (2004) [17] classification and prediction is the process of identifying a set of common features and models that describe and distinguish data classes or concepts. Some of common classification techniques include neural networks, the naïve Bayes technique,
decision trees and support vector machines. Such classification tasks are used in the detection of credit card, healthcare and automobile insurance, and corporate fraud, among other types of fraud, and classification is one of the most common learning models in the application of data mining in Financial Fraud Detection. Some empirical studies have used classification as a benchmark [6,66].

(ii) Regression

Regression [63] is a statistical methodology used to reveal the relationship between one or more independent variables and a dependent variable (that is continuous-valued). The regression technique is typically undertaken using such mathematical methods as logistic regression and linear regression, and it is used in the detection of credit card, crop and automobile insurance, and corporate fraud. Some empirical studies have used regression analysis as a benchmark [1,6,9,58,91,113-115,138].

(iii) Clustering

Clustering [63,95] is used to divide objects into conceptually meaningful groups (clusters), with the objects in a group being similar to one another but very dissimilar to the objects in other groups. Clustering is also known as data segmentation or partitioning and is regarded as a variant of unsupervised classification. According to Yue et al. [139], “clustering analysis concerns the problem of decomposing
or partitioning a data set (usually multivariate) into groups so that the points in one group are similar to each other and are as different as possible from the points in other groups.” Further, Zhang and Zhou [17] argue that each cluster is a collection of data objects which are similar to one another within the same cluster but dissimilar to those in other clusters. The most common clustering techniques are the K-nearest neighbor, the Naïve Bayes technique and self-organizing map techniques. Some empirical studies have used clustering as a benchmark [15,102,136].

(iv) Prediction

Prediction [81,108] estimates numeric and ordered future values based on the patterns of a data set. Han and Kamber [63] told that, for prediction, the attribute for which the values are being predicted is continuous-valued (ordered) rather than categorical (discrete-valued and unordered). This attribute can be referred to simply as the predicted attribute. Neural networks and logistic model prediction are the most commonly used prediction techniques.

(v) Outlier Detection

Outlier detection [63] is employed to measure the “distance” between data objects to detect those objects that are grossly different from or inconsistent with the remaining data set. According to Agyemang et al. (2006) [83], “Data that appear to have different characteristics than the rest of the population are called outliers”.

14
Yamanishi et al. [135] point out that the problem of outlier/anomaly detection is one of the most fundamental issues in data mining. A commonly used technique in outlier detection is the discounting learning algorithm. Some empirical studies have used outlier detection as a benchmark [2,716,38-39,71,83,135,137,140].

(vi) Visualization

Visualization [23,80] refers to the easily understandable presentation of data and to methodology that converts complicated data characteristics into clear patterns to allow users to view the complex patterns or relationships uncovered in the data mining process. Eick and Fyock [106] report that researchers at Bell and AT&T Laboratories have exploited the pattern detection capabilities of the human visual system by building a suite of tools and applications that flexibly encode data using color, position, size and other visual characteristics. Visualization is best used to deliver complex patterns through the clear presentation of data or functions.

1.4 Outlier Analysis

Outlier detection and analysis [63] are very useful for fraud detection, customized marketing, medical analysis and many other tasks. Computer-based outlier analysis methods typically follow either a statistical distribution-based approach, a distance-based approach, a density-based local outlier detection approach or a deviation-based approach. According to Aggarwal
and Yu (2001) [2] outliers may be considered as noisy points lying outside a set of defined clusters or may be defined as points that lie outside of the set of clusters but are also different from the noise.

In our research work we have used Outlier Analysis to detect Price Manipulation done by operators in the Stock Market, we have detected outliers in Intra-day trading who are buying at high price and selling at low price in bulk and they are repeating this process again and again to hammer the stock price. Using Outlier Analysis, we have also detected a pattern from the combined bulk deal data of BSE and NSE, that total selling quantity of many Trading Entities is more than total buying quantity, these Trading Entities may be QIB or NII but they also did Intra-Day Trading which is a illegal course of action. Further we have also compared mean of Outliers in Peer Group Companies with Non Peer Group Companies of Category A and Category B stocks.

1.5 Sources of data for Fraud Detection

Following are the sources of our data for fraud detection in Stock Market-

(a) BSE (Bombay Stock Exchange) Ltd.

(b) NSE (National Stock Exchange) Ltd.

(c) SEBI (Security Board of India).

Initially we have started our research work by studying and analyzing historical data of BSE (Bombay Stock Exchange) [123], NSE (National Stock Exchange) [128] of various companies and investors complains registered with SEBI (Security Board on India) [132] against the
Companies, Brokers, Promoters, Institutional/Non-Institutional Investors, individuals etc. from their respective official websites, to study and analyze the various possibilities of frauds in Indian Stock Market.

1.5.1 BSE (Bombay Stock Exchange) Ltd

BSE was established in 1875 which is the first ever stock exchange in Asia and the first in India which was granted permanent recognition under the Securities Contract Regulation Act, 1956, over the past 137 years. BSE Ltd is also known as “Dalal Street”, now a days. In the 1850s, the earliest stock broker meetings venue was under banyan trees, in front of the Town Hall in the natural environment, where Horniman Circle is now situated. After a decade, the brokers again moved their place to under banyan trees at the junction of Meadows Street, now a days called “Mahatma Gandhi Road”. The brokers had to shift from one venue to another venue as the number of brokers increased. Lastly in 1874, the brokers shifted to a permanent venue, now a days called Dalal Street (Broker’s Street).

BSE Ltd.’s drive is very exciting and attractive. Almost every leading corporate in India now a days has been sourcing BSE Ltd. services for the raising of capital and has been listing itself with BSE Ltd. BSE Ltd., which has became an institutional brand now a days is also known as the capital market in India. BSE’s SENSEX has now become the scale of equity index which is reflecting the health of the Indian economy.
Following are the snapshots of collecting historical data from BSE website-

<table>
<thead>
<tr>
<th>Quick Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Notices and Circulars</td>
</tr>
<tr>
<td>Media Releases</td>
</tr>
<tr>
<td>Trading Holidays</td>
</tr>
<tr>
<td>Settlement Calendar</td>
</tr>
<tr>
<td>Markets Statistics</td>
</tr>
<tr>
<td>Charting</td>
</tr>
<tr>
<td>Get Quote</td>
</tr>
<tr>
<td>Historical Data</td>
</tr>
<tr>
<td>Trade Check</td>
</tr>
</tbody>
</table>

Figure 1.5.1(a) Quick links provided on BSE website.
Figure 1.5.1(b)  Historical Data of Stock Prices in BSE

Figure 1.5.1(c)  Historical Data of Bulk Deals/Block Deals
1.5.2 NSE (National Stock Exchange) Ltd

In November 1992, NSE was promoted by top Financial Institutions at the order of the Government of India, which was incorporated as a tax-paying company.

NSE Ltd. is India's peak stock exchange which is jacketing various cities and towns from all over the country. To offer up to date, fully computerized, screen-based trading system with reach in all over India, NSE Ltd. was established by leading institutions.

NSE Ltd. has played a key role in reforming the Indian Stock Market. Today market is using information technology to provide well organized and transparent trading, compression of settlement cycles, clearing and settlement, stocks lending and borrowing, screen based trading, professionalization of trading members, emergence of clearing corporations to assume counterparty risks, well-tuned risk management systems, screen based trading, market of debt and derivative instruments and demanding use of information technology.

Except Capital Market, NSE Ltd. offers trading in Derivatives Market and Currency Derivatives segments including equities, equities based derivatives, equity based ETFs, Gold ETF, Currency futures and options and Retail Government Securities. Today NSE network spread to more than 1,500 locations in all over country and has been supporting more than 2, 30,000 terminals.
Following are the snapshots of collecting historical data from NSE website:

Figure 1.5.2(a) Historical Data of Stock Prices in NSE

Figure 1.5.2(b) Historical Data of Bulk Deals/Block Deals/Short Selling in NSE
1.5.3 SEBI

The Securities and Exchange Board of India was established on April 12, 1992 in accordance with the provisions of the Securities and Exchange Board of India Act, 1992. SEBI is working as a regulator to control Indian Capital Market. The main role of SEBI [53] is as follows:-

(a) Regulation of Stock Exchanges and Subsidiaries.

(b) Registration and Regulation of the working of Intermediaries.

(c) Registration and regulation of collective investment schemes including mutual funds

(d) Promotion and regulation of self-regulatory organizations.

(e) To control Fraudulent and unfair trade practices.

(f) Investors Education and the training of intermediaries.

(g) Prohibition of Insider Trading

(h) Substantial acquisition of shares and Take-Overs.

(i) Inspections and Enquiries

(j) To receive fees and other charges

(k) Research and international relations.

(l) Grievance Redressal in Mutual Funds, Security Market and Investors Associations.
Following is the snapshot of SEBI’s Investors complain website -

Figure 1.5.3  To register complains and view complains status in SEBI
1.6 Softwares Used

There are lots of softwares available for data analysis in the context of Fraud Detection but due to limitations, we have used following softwares-

(i) Excel

(ii) IBM-SPSS-20

(iii) Rapid Miner-5.2

(iv) Mini Tab-15

(i) Excel

Excel [44] is an application software developed by Microsoft Corporation which is an electronic spreadsheet program, that is used for storing, organizing and manipulating data. Using formulae and functions in Excel, we can perform various required calculations. Spreadsheets are generally used to store financial data. We can use graphs, charts to display the data to assist users to identify data trends. We can use Sorting, Filtering options to find out specific information. We have used Excel in our research work for storing our data and data preprocessing.

(ii) IBM-SPSS-20

SPSS stands for Statistical Package for the Social Sciences. It was developed by IBM (International Business Machine). It is basically used
for Statistical Analysis. IBM-SPSS-Statistics 20 [32] is a package for analyzing data and can take data from almost any file type and can use them to generate reports in tabulated form, provides charts, and plots of distributions and trends, descriptive statistics, and complex statistical analyses. By SPSS Statistics statistical analysis is more accessible for the beginner and it is more convenient for the experienced user. It provides Simple menus and dialog box selections to make it possible to perform complex analyses without typing a single line of command syntax. It’s Data Editor offers us a simple and efficient spreadsheet-like facility to enter data and working data file browsing. We have used SPSS in our research work for Outlier Analysis and various statistical tests.

(iii) Rapid Miner- 5.2

Rapid Miner- 5.2 [43] is a leading open-source system for data mining which is available as a stand-alone application for data analysis. Rapid Miner is easy to use and provides visual environments for predictive analysis. In Rapid Miner no programming required. It can use various data sources like Excel, Access, Oracle, IBM DB2, Microsoft SQL, Sybase, Ingres, MySQL, Postgres, SPSS, dBase, Text files etc. It can analyze every size data source and breaks away from the limitations of traditional data analysis tools and allows us to work with large data sources. It is free to download and use from it’s official website http://www.rapidminer.com. We have used Rapid Miner in our research work for Outlier Analysis.