CHAPTER-5

METHODOLOGY

5.1 Data Collection and Organization

The datasets we have used for this research work is secondary data which was taken from Indian Stock Market- National Stock Exchange (NSE) and Bombay Stock Exchange (BSE). The following sections describe how the data was collected, the structure of the data, and how the data was processed.

5.1.1 Data Description

The current study utilizes the historical data of bulk deals and historical data of stock prices available on BSE and NSE websites, www.bseindia.com and www.nseindia.com for various fields (Attributes). The observations, utilized in the study were collected from the sites in the month of March-2012, July-2012 and December-2012.

In Historical Data of Bulk Deal available fields (attributes) are-

1. Deal Date

2. Script Code

3. Company Name

4. Trading Entity Name
5. Deal Type (Buy/Sell)

6. Quantity

7. Price

8. Remarks

In Historical Data of Stock Prices fields (attributes) are-

1. Symbol

2. Series

3. Date

4. Previous Close

5. Open Price

6. High Price

7. Low Price

8. Last Price

9. Close Price

10. Average Price

11. Total Traded Quantity

12. Turnover
13. Deliverable Qty

14. % Dly Qty to Traded Qty

5.1.2 Data Preprocessing

The following section details various aspects of the data preprocessing, including data integration, cleaning, transformation, and reduction. The data was downloaded from official websites of NSE and BSE, which was available in CSV (comma separated values) format. As a result, some preprocessing was needed. We have converted CSV files into excel file.

(i) Data Integration

This step of data preprocessing was one of the most difficult part. We have combined bulk deal data of NSE and BSE for the same company for the same period, which were stored in 2 different files, converted into a single file. There were different date formats in both files, except this the field names were different for the same field. We have kept common date format and field names after merging the data.

(ii) Data Cleaning

The data cleaning involves removing records that had too many missing values or improperly formatted values. As we had taken secondary data from official websites of NSE and BSE, there was no missing value in the data, so we had no need for data cleaning.
(iii) Data Transformation

The numeric field was transformed from a textual format into numerical numeric formats which allowed for easier mathematical calculations using functions and formulae in Excel.

(iv) Data Reduction

The overall dataset was reduced for the purposes of outlier detection by selecting only the attributes that were the most relevant. Although there are many attributes in the available dataset but there are only a few, critical ones, that can contribute the most in fraud detection. Among all the attributes available on these sites, the current study utilizes only the data with respect to following fields:

From **Historical Data of Bulk Deal** we have taken only following fields (attributes)-

1. Deal Date
2. Trading Entity Name
3. Deal Type(Buy/Sell)
4. Quantity
5. Price
From **Historical Data of Stock Prices** we have taken only following fields (attributes)-

1. Date
2. Open Price
3. High Price
4. Low Price
5. Close Price

### 5.2 Methodologies Used

There are many statistical and computer based tools applicable for Fraud Detection as we have discussed in Chapter-1(1.3). Among all these methods following methods are applied in the proposed research work:

1. **(i) Outlier Analysis**
2. **(ii) Multiple Linear Regression Analysis**
3. **(iii) Peer Group Analysis**

#### 5.2.1 Outlier Analysis

According to **Han and Kamber (2010) [63]**, a database may contain data objects which do not comply with the general behavior or model of the data. These data objects are outliers. Most data mining methods discard outliers as noise or exceptions. In some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring ones.
The analysis of outlier data is known as outlier mining. Outliers can be detected using statistical tests that assume a distribution or probability model for the data, or using distance measures where objects that are substantial distance from any other cluster are considered outliers. Rather than using statistical or distance measures, deviation based methods can identify outliers by examining differences in the main characteristics of objects in a group. Outlier detection and analysis are very useful for fraud detection, customized marketing, medical analysis etc. Computer-based outlier analysis methods generally follow either a statistical distribution-based approach, a distance-based approach, a density-based local outlier detection approach or a deviation-based approach. Yamanishi et al. [135] point out that the problem of outlier/anomaly detection is one of the most fundamental issues in data mining. To find outliers is a very important task in Data Mining. In Stock Market, outlier detection plays very important role in detecting fraud. The main objective of Data Mining is to search for a general pattern for the input data. There is lack of research work done using Data Mining Techniques of Outlier Detection for Financial Fraud Detection, may be it is very difficult to detect outliers as according to Agyemang et al.(2006) [83], Outlier Detection is a very complex task akin to finding a needle in a haystack. According to Barnett and Lewis (1994) [7] Novelty detection, or so-called outlier detection, is the identification of “novel” or “unknown” events that an expert system is not aware of during training or testing. Outliers may indicate abnormal running conditions and lead to significant performance degradation.
An outlier is one that appears to obviously deviate from the others of the sample in which it occurs or an observation which appears to be inconsistent with the remainder of the dataset. Knorr E. and Ng. R. (1998) [71] studied finding of outliers (exceptions) in large, multidimensional datasets. The identification of outliers can lead to the discovery of truly unexpected knowledge in areas such as electronic commerce, credit card fraud, and even the analysis of performance statistics of professional athletes. According to them existing methods for finding outliers in large datasets can only deal efficiently with two dimensions/attributes of a dataset. They studied the notion of DB- (Distance- Based) outliers. While they provided formal and empirical evidence of showing the usefulness of DB-outliers, they focused on the development of algorithms for computing such outliers. 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. According to He, Xu, Huang, Deng, 2004 [38], Coderre, 2009 [18], Aggarwal & Guojun, 2003 [100], outlier analysis attempts to find the rare class whose behavior is very exceptional when compared to the rest of input data. Many techniques are proposed to detect outliers, drawn from Statistics, Computer Science or Machine Learning. Hodge and Austin (2004) [39], reviewed some fundamental approaches to solve the problem of Outlier Detection. This technique is usually named as Novelty Detection since it aims to define the boundary of normality instead of estimating the density of the dataset. In
addition to the surveillance of stock price changes, Anomaly Detection Techniques have been applied to various fields such as Network Intrusion Detection (Naiman, 2004 [90]), Financial Fraud Detection (Juszczak, Adams, Hand, Whitrow, & Weston, 2008 [66]), Fault Detection (Chen, Martin, & Montague, 2009 [16]; Martins, Pires, & Amaral, 2011 [85]; Yiakopoulos, Gryllias, & Antoniadis, 2011 [136]). Some empirical studies have used outlier detection as a benchmark [2,716,38-39,71,83,135,137,140].

In our research work using Outlier Analysis, we have detected a rare pattern of trading 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 [110], so we have considered those trading entities as outliers who’s Total Sell Qty is greater than Total Buy Qty and did Intra Day Trading(i.e. on the same day buy and sell) (Chapter-4, Algorithm-II).

5.2.2 Multiple Regression Analysis

According to Margaret H. Dunham (2005) [84], Multiple Linear Regression assumes that a linear relationship exists between the input variables (Predictors) and the output variables (Response Variable). The common formula for a linear relationship is used in this model:

\[ y = c_0 + c_1x_1 + \ldots + c_nx_n \]
In the above equation there are \( n \) input variables \((x_1, x_2, \ldots, x_n)\), which are predictors or regressors and one output variable \((y, \text{ the response variable})\). The constants \( c_0, c_1, \ldots, c_n \) are the regression coefficients which is computed by the method, known as Principle of Least Square, and computed during the modeling process when processed on a statistical software like Minitab-15 or IBM-SPSS-20 etc.

Suppose we are dealing with investigation on multivariate data. Let \( X_1, X_2, \ldots, X_p \) be a set of \( p \) random variables. Suppose the variable \( X_1 \) is of primary interest while the remaining element \( X_2, X_3, \ldots, X_p \) of secondary interest. Suppose variable \( X_1 \) is dependent to \( X_2, X_3, \ldots, X_p \). Let us suppose that we want to see the notion of independence of \( X_1 \) on \( X_2, X_3, \ldots, X_p \). To do this, we deal with the conditional distribution or more simplify the conditional expectation of \( X_1 \) for given \( X_2, X_3, \ldots, X_p \). Let expectation of \( X_1 \) for fixed \( X_2, X_3, \ldots, X_p \) is given as-

\[
E[X_1 | X_2, X_3, \ldots, X_p] = E[X_1 | X_2 = x_2, X_3 = x_3, \ldots, X_p = x_p] = \Psi(x_2, x_3, \ldots, x_p) \ldots \ldots \ldots \ldots \ldots \ldots (1)
\]

The above conditional expectation is the conditional mean of \( X_1 \) as a function of \( X_2, X_3, \ldots, X_p \). The function \( \Psi \) may be linear or non linear. The conditional mean given by equation (1) is known as the “Multiple Regression Plane” of \( X_1 \) on \( X_2, X_3, \ldots, X_p \). The variable \( X_1 \) is referred to as the regressor(dependent) and the variables \( X_2, X_3, \ldots, X_p \) as regressands (independent).

If \( \Psi \) is linear then equation (1) may be written as-
\[ \Psi(x_2, x_3, \ldots, x_p) = \alpha + \beta_{12}x_2 + \beta_{13}x_3 + \ldots + \beta_{1p}x_p \]  \hspace{1cm} (2) 

Where \( \beta_{ij} = \beta_{ij.2,3,\ldots,p} \) is the regression coefficient of \( X_1 \) on \( X_j \) while keeping other variable fixed. \( \beta_{ij} \) gives the amount of change in conditional mean or a unit change in the value of variable \( X_j \) while the variable \( X_2, X_3, \ldots X_{j-1}, X_{j+1}, \ldots, X_p \) are kept fixed. \( \beta_{ij} \) is called the partial regression of \( X_1 \) on \( X_j \), when \( X_2, X_3, X_{j-1}, X_{j+1}, \ldots, X_p \) are fixed.

Regression Analysis [63] is a statistical methodology that is most often used for numeric prediction and 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].

ANOVA in Multiple Regression Analysis

ANOVA (Han and Kamber, 2010) [63], is a statistical method used to test differences between two or more means. These techniques analyze experimental data for two or more populations described by a numeric response variable and one or more categorical variables (factors).

Let \( \sum x_1^2, \sum x_1 x_2, \sum x_1 x_3, \ldots, \sum x_1 x_p \) are the corrected sum of squares and cross products and \( b_{12}, b_{13}, \ldots, b_{1p} \) are the least square estimate of \( \beta_{12}, \beta_{13}, \ldots \)
β_{1p} (multiple regression coefficients). If R is a population multiple correlation coefficient of X₁ and linear combination of X₂,X₃…Xₚ then we can test the null hypothesis H₀: \( \bar{R} = 0 \) by the help of ANOVA table given as follows-

**Table 5.2.2- ANOVA in Multiple Regression Analysis**

<table>
<thead>
<tr>
<th>Sources of Variation</th>
<th>D.F</th>
<th>S.S</th>
<th>M.S.S</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression of X₁ on X₂, X₃... Xₚ.</td>
<td>p-1</td>
<td>( b_{12}\sum X_1 X_2 + b_{13}\sum X_1 X_3 + \ldots + b_{1p}\sum X_1 X_p = A (\text{say}) )</td>
<td>( \frac{A}{p-1} )</td>
<td>( \frac{A}{V^2 X \frac{N-p}{p-1}} )</td>
</tr>
<tr>
<td>Error</td>
<td>N-p</td>
<td>( \sum X_1^2 - A = V^2 )</td>
<td>( \frac{V^2}{N-p} )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>N-1</td>
<td>( \sum X_i^2 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**F** is calculated as:

\[
F = \frac{b_{12}\sum X_1 X_2 + b_{13}\sum X_1 X_3 + \ldots + b_{1p}\sum X_1 X_p}{\sum X_1^2 - (b_{12}\sum X_1 X_2 + b_{13}\sum X_1 X_3 + \ldots + b_{1p}\sum X_1 X_p)} \left( \frac{N-p}{p-1} \right)
\]

\[
R^2 = \frac{\sum_{j=2}^{p} b_{1j}X_1 X_j}{\sum_{j=1}^{p} X_1^2}
\]

In our research work using Multiple Regression and ANOVA, we have detected outliers. We have analyzed and studied 15 stocks by collecting historical data of a certain period from BSE website. Using Multiple Regression Analysis, in our model, we have taken Close Price as Response Variable and two predictors (i) Open Price (ii) Difference of High and Low Price. The resultant Outliers are data points that are more than some
appropriate distance from a regression line that is estimated using all the other data points in the sample.

### 5.2.3 Peer Group Analysis

According to Bolton & Hand (2001) [103], Peer Group Analysis, 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). Using Peer Group Analysis in Time Series Financial Data, Zakia et al. (2006) [140] demonstrated the experimental results of PGA tool in an unsupervised problem over real stock market data sets with continuous values over regular time intervals. The visual evidences have been shown through graphical plots that Peer Group Analysis can be useful in detecting observations that deviate from their peers. They also applied t statistics to find the deviations effectively. Their aim was to proceed by incorporating other information, other than simply the quantity sold, into the outlier detection process (PGA) to increase the effectiveness of the fraud detection system. They told following cases of possible outliers have to be investigated. Using Peer Group Analysis Y. Kim et. al. (2012) [134], proposed a method to detect suspicious patterns of stock price manipulation. This technique detects abnormal behavior of a target by comparing it with its peer group and measuring the deviation of its behavior from that of its peers. They proposed a method to improve the
general peer group analysis by incorporating the weight of peer group members into summarizing their behavior, along with the consideration of parameter updates over time. Using real time series data of Korean stock market, they showed the advantage of the peer group analysis in detecting abnormal stock price change. **Bolton, Hand and David J.H (2001) [11]** studied that Credit Card Fraud falls broadly into two categories: (i) Behavioral Fraud and (ii) Application Fraud. They told that Application Fraud occurs when individuals obtain new credit cards from issuing companies using false personal information and then spend as much as possible in a short space of time. However, most credit card fraud is behavioral and occurs when details of legitimate cards have been obtained fraudulently and sales are made on a 'Cardholder Not Present' basis. These sales include telephone sales and e-commerce transactions where only the card details are required. They concerned with detecting Behavioral Fraud through the analysis of longitudinal data. These data usually consist of credit card transactions over time, but can include other variables, both static and longitudinal. Statistical methods for fraud detection are often classification (supervised) methods that discriminate between known fraudulent and non-fraudulent transactions; however, these methods rely on accurate identification of fraudulent transactions in historical databases - information that is often in short supply or non-existent. They were particularly interested in unsupervised methods that do not use this information but instead detect changes in behavior or unusual transactions.
They discussed two methods for unsupervised fraud detection in credit data in this paper and applied them to some real data sets. Peer group analysis is a new tool for monitoring behavior over time in data mining situations. In particular, the tool detects individual accounts that begin to behave in a way distinct from accounts to which they had previously been similar. Each account is selected as a target account and is compared with all other accounts in the database. Some empirical studies have used Peer Group Analysis as a benchmark in Fraud Detection [11,103,134,140].

In our research work using **Peer Group Analysis** we have made peers of similar group companies that is the companies which are from same sector and non peers that is the companies which are from different-different sectors. For this we have made peers of companies of Nutek India Ltd., from same sector- Other Telecom Service, and compared the outliers in Nutek India Ltd. with its Peer Group companies of type-Other Telecom Services, to find out whether the results are almost similar or different, and further compared the Peer Group companies with Non Peer Group companies of Category A stocks and Category B stocks and found that category A stocks has lesser outliers in comparison of category B stocks. As all the stocks from Peer Group are from category B stocks, so we have further compared Peer Group with Non Peer Group of category B stocks, and found still Peer Group stocks have more outliers.
5.3 Procedure for Outlier Analysis

In our research work we have used Outlier Analysis, Multiple Regression Analysis, Peer Group Analysis and ANOVA methodologies to detect possibilities of fraud (outliers) in stock market.

In Experiment-1(Chapter-6), using Time Series financial data from official websites of BSE and NSE, we have studied how share price of a company ‘Nutek India Ltd’ fallen from its IPO listing price of Rs. 192 to cheap dirt price of Rs 0.50. We have studied monthly price variation from IPO month (August, 2008) to February, 2012. We have studied that how manipulators/operators/trading entities hammered the price from peak level to bottom level by doing the malpractice of buying at higher price and selling at lower price in bulk in Intraday-Day Trading repeatedly. In this study, we have consider those trading entities as outliers who’s Weighted Average Buy Price was higher than Weighted Average Sell Price or Weighted Average Price of Still Holding Qty was higher than Current Market Price (Chapter-4, Algorithm-I).

In Experiment-2(Chapter-6), using Outlier Analysis, we have detected a rare pattern of trading 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 [110], so we have considered those
trading entities as outliers who’s Total Sell Qty is greater than Total Buy Qty and did Intra Day Trading(i.e. on the same day buy and sell) (Chapter-4, Algorithm-II).

In Experiment-3 (Chapter-6), using Multiple Regression, ANOVA and Peer Group Analysis, we have detected outliers. We have analyzed and studied 15 stocks by collecting historical data of a certain period from BSE website. Using Multiple Regression Analysis, in our model, we have taken Close Price as Response Variable and two predictors (i) Open Price (ii) Difference of High and Low Price. The resultant Outliers are data points that are more than some appropriate distance from a regression line that is estimated using all the other data points in the sample. We have detected outliers stock wise and then using Peer Group Analysis we have compared these outliers with its Peer Group to find out whether the results are almost similar or different, and further compared the Peer Group with Non Peer Groups of Category A and Category B stocks and found that category A stocks has lesser outliers in comparison of category B stocks. As all the stocks from Peer Group are from category B stocks, so we have further compared Peer Group with Non Peer Group of category B stocks, and found still Peer Group stocks have more outliers. Statistical tests like “t-Test”, “Analysis of Variance (ANOVA)” are tested. All the Statistical Tests are tested at 5% level of significance. The level of significance is the risk value at which the null hypothesis is presumed to be true.