MRP

by Ajay Kumar

Submission date: 28-May-2019 03:59PM (UTC+0530)

Submission ID: 1136869002

File name: Ajay_Chauhan.docx (2.96M)

Word count: 10426

Character count: 61202

Report on
Financial Statements Fraud Detection Model Based on Hybrid Data Mining Methods: Proposing an Optimized Detection Model

Submitted By :
Ajay Singh Chauhan
2K17/MBA/703

Under the Guidance of :
Mrs. Kusum Lata
(Assist. Prof., Delhi Technological University)

UNIVERSITY SCHOOL OF MANAGEMENT & ENTREPRENEURSHIP

Delhi Technological University
MAY 2019

CERTIFICATE

This is to certify that **Mr. Ajay Singh Chauhan(2K17/MBA/703)**, a student of MBA in Finance and Information Technology has successfully completed the project entitled, **"Financial Statement Fraud Detection Model Based on Hybrid Data Mining Methods: Proposing an Optimized Financial Fraud Detection Model"** under the guidance of **Assist. Prof. Mrs. Kusum Lata** (Supervisor/ Mentor) in the year 2019 in partial fulfilment of end semester examination conducted at the University School of Management and Entrepreneurship, New Delhi – 110095.

Assist. Prof. Mrs. Kusum Lata **Date**
University School of Management and Entrepreneurship
Delhi Technological University
New Delhi – 110095

DECLARATION

This research project is my original work and has not been submitted for examination to any other university.

Ajay Singh Chauhan
Date

This research project has been submitted for examination with my approval as the University Supervisor/ Mentor.

Assist. Prof. Mrs. Kusum lata
University School of Management & Entrepreneurship
Delhi Technological University
New Delhi - 110095
Date

ACKNOWLEDGEMENT

The completion of this study would have been impossible without the material and moral support from various people. It is my obligation, therefore, to extend my gratitude to them. First of all, I thank the Almighty God for giving me good health, and guiding me through the entire course. I am greatly indebted to Assist. Prof. Mrs. Kusum Lata who was my supervisor for his effective supervision, dedication, availability and professional advice. I extend my gratitude to my lecturers who taught me in the MBA programme, therefore enriching my research with knowledge. The data sourced from various publications, journals, and their authors, deserve my appreciation for their work and findings for providing the required information during my study. My appreciation finally goes to my classmates, with whom I weathered through the storms, giving each other encouragement and for their positive criticism.

Ajay Singh Chauhan **Date**

(2K17/MBA/703)

EXECUTIVE SUMMARY

Financial statement fraud has been a difficult problem for both the public and government regulators, so various data mining methods have been used for financial statement fraud detection to provide decision support for stakeholders. The purpose of this study is to propose an optimized financial fraud detection model combining feature selection and machine learning classification. I used feature selection to reduce the dimensionality following PCA and Xgboost to do it. Principal component analysis (PCA) is a statistical method. By orthogonal transformation, a set of observations of possibly correlated variables converted into a set of linearly independent variables, which called principal component. I used machine learning methods to explore the variables with Support Vector Machine (SVM), Random Forest(RF), Decision Tree(DT), Artificial Neural Network(ANN), and Logistic Regression(LR) for PCA and then with Xgboost. The study indicated that random forest outperformed the other four methods. As to two feature selection methods, Xgboost performed better. And according to our research, 2 or 5 variables are more acceptable for models in this research.

TABLE OF CONTENTS

CHAPTER - 1
INTRODUCTION

1. Introduction

Association of Certified Fraud Examiners (ACFE) defines fraud as "A kind of misrepresentation or deception that an entity or individual makes knowing that it could result in some unauthorized benefit." According to a study conducted by the ACFE, financial statement fraud accounnts for about 10% of whitecollar crime. Once fraudulent accounting practices happened, various actions willbe taken to maintain a sustainable appearance.

Considering financial statement fraud brinng huge property damage to investors, a large number of researches have been connducted on this area usinng machine learning methodssuch as ANN, DT, SVM, and text mining. Meanwhile, other fraud problems like credit card fraud, internal transaction fraud and insurance fraudhave also been investiigated. Given the different characteristics of each type of financil fraud, specific methods have been developed. The research putsforward a hybrid detection model for financal statemennt fraud, and the model has the advantages of (1) combining the financal and non-financal data, (2) using two feature selection methods, and (3) easy to explain.

1.1. The objective of the study

The study is conducted for the detection of fraud in the Financial Statements. There are various methods of data mining methods that are used in the study. The purpose of this study is to propose an optimized financial fraud detection model combining feature selection and machine learning classification.

The objective of this study is to examine the performance of two advancd data mining techniques, random forests and support vector machines, together with the well-known logistic regression, for credit card fraud identification. We also want to compare the effect of the extent of data under sampling on the performance of these techniques. This section describes the data used for training and testing the models and performance measures used.

For the comparative evaluation, parameters for the techniques were set from what has been found generally useful in the literature and as determined from the preliminary tests on the data. No further fine-tuning of parameters was conducted. While fine-tuning of parameters to specific datasets can be beneficial, consideration of generally accepted settings is more typical in practice. The need for significant effort and time for parameter fine-tuning can often be a deterrent to practical use, and can also lead to issues of overfitting to specific data.

For SVM, I use Gaussian radial basis function as the kernel function which is a general-purpose kernel with good performance results.

1.2. About Finance

Finance is a broad term that describes activities associated with banking, leverage or debt, credit, capital markets, money, and investments. Basically, finance represents money management and the process of acquiring needed funds. Finance also encompasses the oversight, creation, and study of money, banking, credit, investments, assets, and liabilities that make up financal systems.

Many of the basic concepts in finance originate from micro and macroeconomic theories. One of the most fundamental theories is the time value of money, which essentially states that a dollar today is worth more than a dollar in the future.

Since individuals, businesses, and government entities all need funding to operate, the field includes three main sub-categories: personal finance, corporate finance, and public (government) finance.

There are three main types of finance:

1. personal,
2. corporate, and
3. public/government.

Personal Finance

Financial planning involves analyzing the current financal position of individuals to formulate strategies for future needs within financial constraints. Personal finance is specific to every individual's situation and activity; therefore, financial strategies depend largely on the person's earnings, living requirements, goals, and desires.

For example, individuals must save for retirement, which requires saving or investing enough money during their working lives to fund their long-term plans. This type of financal management decision falls under personal finance.

Personal finance includes the purchasing of financial products such as credit cards, insurance, mortgages, and various types of investments. Banking is also considered a component of personal finance including checking and savings accounts and online or mobile payment services like PayPal and Visa.

Corporate Finance

Corporate finance refers to the financial activities related to running a corporation, usually with a division or department set up to oversee the financal activities.

For example, a large company may have to decide whether to raise additional funds through a bond issue or stock offering. Investment banks may advise the firm on such considerations and help them market the securities.

Startups may receive capital from angel investors or venture capitalists in exchange for a percentage of ownership. If a company thrives and decides to go public, it will issue shares on a stock exchange through an initial public offering (IPO) to raise cash.

In other cases, a company might be trying to budget their capital and decide which projects to finance and which to put on hold in order to grow the company. These types of decisions fall under corporate finance.

Public Finance

Public finance includes tax, spending, budgeting, and debt issuance policies that affect how a government pays for the services it provides to the public.

The federal government helps prevent market failure by overseeing the allocation of resources, distribution of income, and economic stability. Regular funding is secured mostly through taxation. Borrowing from banks, insurance companies, and other nations also help finance government spending.

In addition to managing money in day-to-day operations, a government body also has social and fiscal responsibilities. A government is expected to ensure adequate social programs for its tax-paying citizens and to maintain a stable economy so that people can save and their money will be safe.[1]

Below is a list of the most common examples:

1. Investing personal money in stocks, bonds, or guaranteed investment certificates (GICs)
2. Borrowing money from institutional investors by issuing bonds on behalf of a public company
3. Lending money to people by providing them a mortgage to buy a house with
4. Using Excel spreadsheets to build a budget and financial model for a corporation
5. Saving personal money in a high-interest savings account
6. Developing a forecast for government spending and revenue collection

There is a wide range of topics that people in the financal industry are concerned with. Below is a list of some of the most common topics you should expect to encounter in the industry.

1. Interest rates and spreads
2. Yield (coupon payments, dividends)
3. Financial statements (balance sheet, income statement, cash flow statement)
4. Cash flow (free cash flow, other types of cash flow)
5. Profit (net income)
6. Cost of capital (WACC)
7. Rates of return (IRR, ROI, ROA)
8. Dividends and return of capital

9. Shareholders
10. Creating value
11. Risk and return [2]

Corporate finance will be our area of interest. **Fraud**, in a general sense, is an intentionally deceptive action designed to provide the perpetrator with an unlawful gain or to deny a right to a victim.

Corporate fraud consists of activities undertaken by an individual or company that aredone in a dishonest or illegalmanner, and are designedto give an advantage to the perpetratingindividual or company. Corporate fraud schemes go beyond the scope of an employee's stated position and are marked by their complexity and economic impact on the business, other employees and outside parties.

Corporate fraud can be difficult to prevent and tocatch. Bycreating effective policies, a system of checks and balances and physical security, a company may limit the extent to which fraud can take place. It is considered a **white-collar crime**.

Though it may be conducted in avariety of ways, corporate fraud frequently is performed by taking advantage of confidential information or access tosensitive assets and then leveraging thoseassets for gain. The fraud is often hidden behind legitimate business practices or exchanges in order to disguise the illicit activity. For example, accounting for a company may be altered topresent an image of high revenue and profits compared with the actual financial results. These actionsmight be taken to hide shortcomings such as a net loss, slow revenue, declining sales, or hefty expenses. Falsified accounting might be done to make the company more attractive to potential buyers or investors.

Other forms of corporate fraud may aim to disguise or misrepresent a service or product the company is developing or has in service, hiding its flaws or defects. Rather than invest inrepairing, refurbishing, or redesigning the product, those responsible for the productattempt to deflect or disguise these issues. This might be doneif the department or company does not havethe finances to correct the problem or if revealing the issue might drive away customers and investors.

If a company or individual claims it is using some of its funds to put towards investments or other types of monetary eserves that are intended to gain in value, but in actuality, those funds have been expended or diverted elsewhere, which is a type of corporate fraud.

The deceptive accounting and business practices that led to the downfall of Enron is an example of corporate fraud. Due to the widespread use of loopholes and other disguising tactics, the company hid debt from failed deals, the sum reaching into the billions of dollars. In order to maintain the charade, those responsible ressured their auditors to hide their deception, which included the destruction of financial documents. [3]

The most common financial statement fraud red flags:

- Accounting anomalies, such as growing revenues without a corresponding growth in cash flows. Sales are much easier to manipulate than cash flow but the two should move more or less in tandem over time.
- Consistent sales growth while established competitors are experiencing periods of weak performance. Of course, this may be due to efficient business operations rather than fraudulent activity.
- A rapid and unexplainable rise in the number of day's sales in receivables in addition to growing inventories. This suggests obsolete goods for which the firm records fictitious future sales.
- A significant surge in the company's performance within the final reporting period of the fiscal year. The company may be under immense pressure to meet analysts' expectations.
- The company maintains consistent gross profit margins while its industry is facing pricing pressure. This can potentially indicate failure to recognize expenses or aggressive revenue recognition.
- A large buildup of fixed assets. An unexpected accumulation of fixed assets can flag the usage of operating expense capitalization, rather than expense recognition.
- Depreciation methods and estimates of assets' useful life that do not correspond to those of the overall industry. An overstated life of an asset will decrease the annual depreciation expense.

- A weak system of internal control. Strong corporate governance and internal controls processes minimize the likelihood that financial statement fraud will go unnoticed.
- Outsized frequency of complex related-party or third-party transactions, many of which do not add tangible value (can be used to conceal debt off the balance sheet).
- The firm is on the brink of breaching their debt covenants. To avoid technical default, management may be forced to fraudulently adjust its leverage ratios.
- The auditor was replaced, resulting in a missed accounting period. Auditor replacement can signal a dysfunctional relationship while missed accounting period provides extra time to "fix" financials.
- A disproportionate amount of management compensation is derived from bonuses based on short term targets. This provides incentive to commit fraud.
- Something just feels off about the corporation's business model, financial statements or operations[4]

1.3. About Data Mining

We live in a world where vast amounts of data are collected daily. Analyzing such data is an important need. Looks at how data mining can meet this need by providing tools to discover knowledge from data. We observe how data mining can be viewed as a result of the natural evolution of information technology.

Data mining can be viewed as a result of the natural evolution of information technology. The database and data management industry evolved in the development of several critical functionalities: *data collection* and *database creation*, *data management* (including data storage and retrieval and database transaction processing), and *advanced data analysis* (involving data warehousing and data mining). The early development of data collection and database creation mechanisms served as a prerequisite for the later development of effective mechanisms for data storage and retrieval, as well as query and transaction processing. Nowadays numerous database systems offer query and transaction processing as common practice. Advanced data analysis has naturally, become the next step.

Many people treat data mining as a synonym for another popularly used term, knowledge discovery from data, or KDD, while others view data mining as merely an essential step in the process of knowledge discovery. The knowledge discovery process is an iterative sequence of the following steps:

1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (where multiple data sources may be combined)
3. Data selection (where data relevant to the analysis task are retrieved from the database)
4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
5. Data mining(an essential process where intelligent methods are applied to extract data patterns)
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on *interestingness measures*)
7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

Machine learning investigates how computers can learn (or improve their performance) based on data. The main research area is for computer programs to *automatically* learn to recognize complex patterns and make intelligent decisions based on data. For example, a typical machine learning problem is to program a computer so that it can automatically recognize handwritten postal codes on mail after learning from a set of examples.

Machine learning is a fast-growing discipline. Here, I illustrate classic problems in machine learning that are highly related to data mining.

1. *Supervised learning* is basically a synonym for classification. The supervision in the learning comes from the labelled examples in the training data set. For example, in the postal code recognition problem, a set of handwritten postal code images and their corresponding machine-readable translations are used as the training examples, which supervise the learning of the classification model.
2. *Unsupervised learning* is essentially a synonym for clustering. The learning process is unsupervised since the input examples are not class labelled. Typically, I may use clustering to discover classes within the data. For example, an unsupervised learning method can take, as input, a set of images of handwritten digits. Suppose that it finds 10 clusters of data. These clusters may correspond to the 10 distinct digits of 0 to 9,

respectively. However, since the training data are not labelled, the learned model cannot tell us the semantic meaning of the clusters found.

3. *Semi-supervised learning* is a class of machine learning techniques that make use of both labelled and unlabeled examples when learning a model. In one approach, labelled examples are used to learn class models and unlabeled examples are used to refine the boundaries between classes. For a two-class problem, I can think of the set of examples belonging to one class as the *positive examples* and those belonging to the other class as the *negative examples*. If I do not consider the unlabeled examples, the dashed line is the decision boundary that best partitions the positive examples from the negative examples. Using the unlabeled examples, I can refine the decision boundary to the solid line. Moreover, I can detect that the two positive examples at the top right corner, though labelled, are likely noise or outliers.
4. *Active learning* is a machine learning approach that lets users play an active role in the learning process. An active learning approach can ask a user (e.g., a domain expert) to label an example, which may be from a set of unlabeled examples or synthesized by the learning program. The goal is to optimize the model quality by actively acquiring knowledge from human users, given a constraint on how many examples they can be asked to label. [5]

There are various data mining techniques which are used to analyze and detect fraudulent companies. The data mining techniques like SVM(Support Vector Machines) , RF(Random Forests), DT(Decision Trees), ANN (Artificial Neural Networks) and LR(Logistic regression). These methods can be used for a comprehensive and multi-dimensional analysis of the data.
A brief description of all these algorithms is given below:

(A) *Support Vector Machines(SVM)*:

This is a type of discriminative classifier, which formally uses a hyperplane to distinguish the data plane i.e. the algorithm outputs an optimal hyperplane that helps in categorizing the new data points. This optimal hyperplane should be one such that the number of outliers present should be minimized. But there exists a tradeoff, more the perfect hyperplane, it is required that a million of training points are trained which takes a lot of time. So to reduce the time, I introduce a regularization parameter. These are the tuning parameters in the SVM classifier. The regularization parameters tell how much do we want to misclassify a training point. Mor the regularization parameter value, a smaller margin hyperplane will be chosen so tht the all the training points are classified correctly . and with a largr regularzation paramters means that a higher margin hyperplane will be chosen, such that more points can be misclassified by that hyperplane. Another important factor is the gamma factor, if the value is low then the far away points are also used to calculate the decision boundary and if the value of gamma is less, then nearby points are only used to evaluate the decision boundary. Krnel is also an imortant factor, the learning in the SVM classifer becomes linr if the linear kerel is used, ie the decsion bondary chosn is a linear decision boundary.

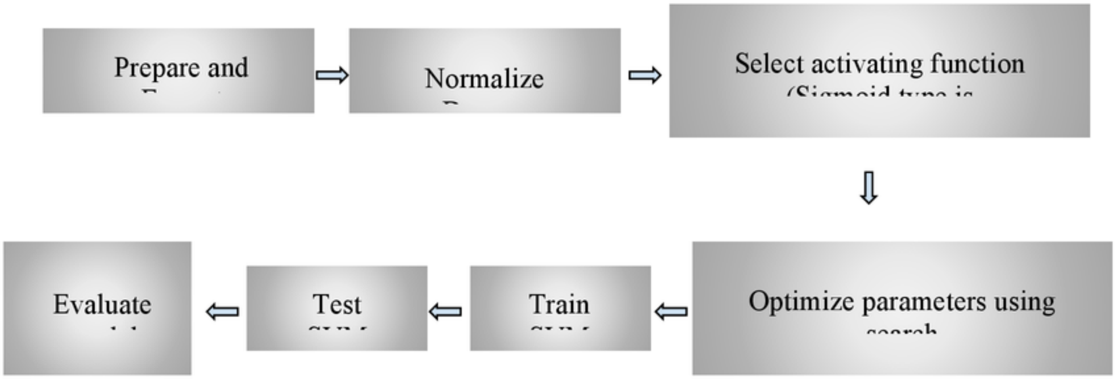

Fig. 1.1 SVM algorithmic flow

B) *Random Forest(RF)*:

In this machine learning algorithm , it is an ensemble algorithm of Decision Tress. In ths algorithm it builds decision tress and merges all them together to get a better and accurate and stable predictios. With Random Foest it can be used with both the classification and regression problems . Its parameter are similar to bagging classifier . In random forest only a random subset of features is determined by taking the algoithm for splitting the node. The trees are more random by using random threholds for all the features rather than searching the best possible thresholds. The parameters of random forests are used to increase the predctive power of the model and makes the model aster. Few important parameters are the number of trss required to build the algorthm before taking the maxmum vote. Higher the number of trees mre the performance and stable the prediction. The jobs telss the engine how many processsors to use. Randm state parameter tells the model outputs replicable. N_sample leaf parameeter determinned the minimum number of leas that a required to split the internal node. Randonm Forest is easy to use and a goof prediction resultant but the problem is overfitting. Also the more number of tress can take more time to execute So a much more accurate number of tress are required to get a best ideal results and classifiers.

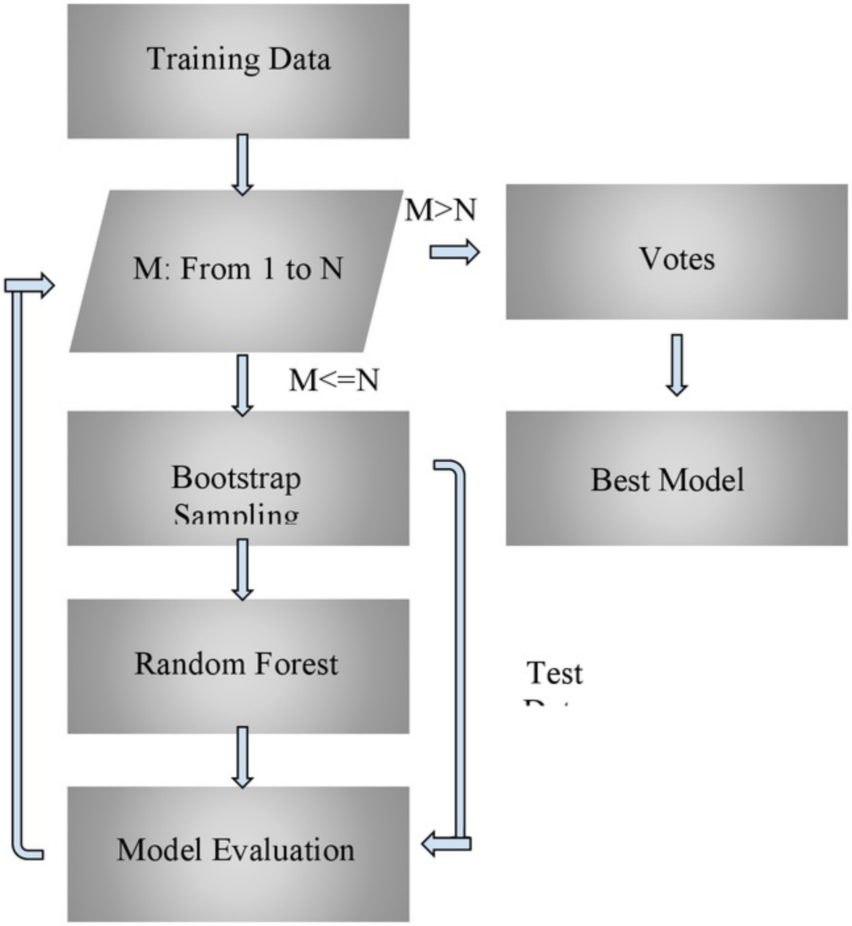

Fig. 1.2 Random Forest general algorithmic flow

C) *Decision Trees(DT)*:

Decision Trees can be used for both classification and regression problem. The advantage of this classification algorithm is it mimics the human level thinking and can make a good level of interpretation. In the decision tree every node exhibits a feature and every link represents a decision and the leaf will represent the outcome that can be categorical or continues values.

Overfitting is a problem in the decision tree , overfitting is more when the trees go deeper and deeper. This happens when build with many branches due to outliers and irregularities in data. The approaches that we can use for reducing the overfitting are , *Pre-Pruning* and *Post-Pruning,* in former the decision tree constructions it is preferred not to split a node if its a goodness measures is below a threshold value whreas in the latter , the tree goes deeper and deeper. If the tee shows a overfitting problem then pruning is done as a post - step. The advantages of using a decision tree is it follows a similar approach as humn approach. The number of hyperparamaters are also few. The disadvantages are the problm of overfitting is high in this algorithm and generally it gives low accuracy as compared to the other ML algorithms.

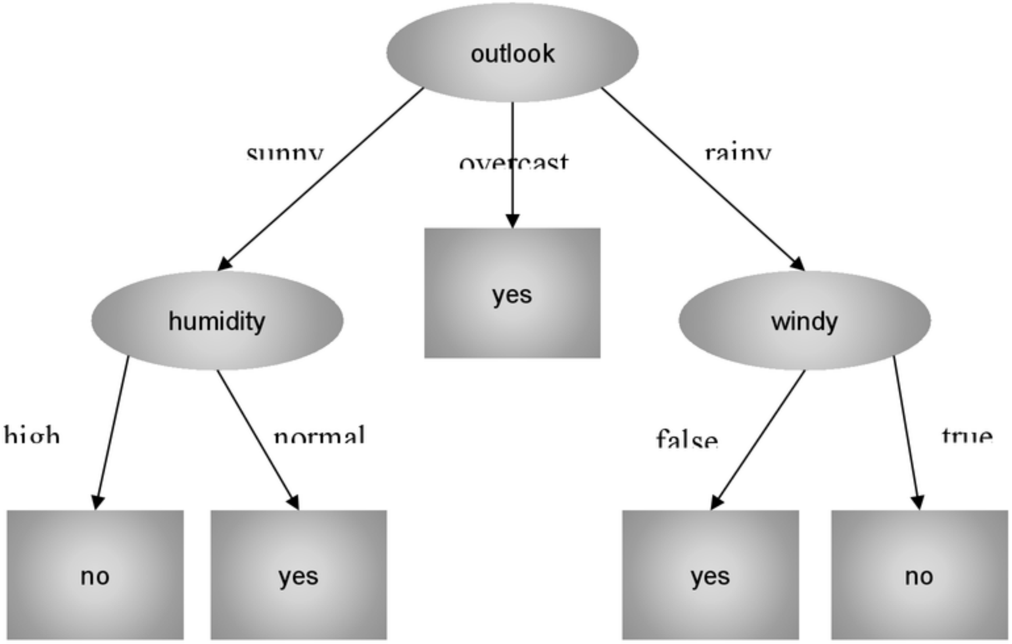

Fig. 1.3 Sample Decision Tree problem

D) *Artificial Neural Networks(ANN)*:

The term “neural” is inspired by the brain system .Neural networks consists of a input layer , output layers and in majority cases a set of hidden layers. They are excellent in finding patterns and teach and lean the machine to recognize. The back-propagation algorithm is a major part which helps in adjusting the hidden layers and all the other parameters. To train a neural network three types of data are used the training data , the validation data and the testing data. The different tasks a nueral network can do are recognizing the faces, recognizing different patterns etc .The can basically spot different patterns and study them. Tasks include classification , clustering and also prediction and patten mining. One major limitaion of the artificial nueral networks is the computation power and the time taken to execute.the layers act like a black boxes and keep on fine tuing the prameters as in more sets od data set is used.

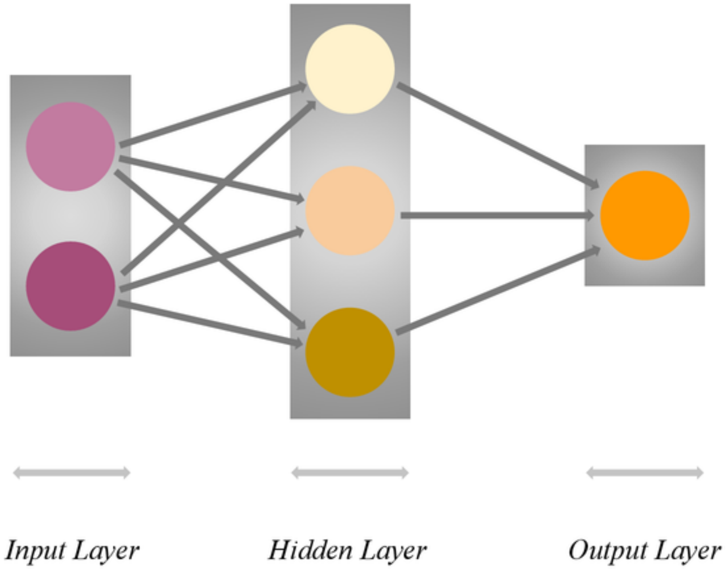

Fig. 1.4 A neural network model

E) *Logistic Regression(LR):*

With this machine learning algorithm, we can assign observations to the discrete classes.It is used for the classification problems and predictive analysis using the concept of probability.

Logistic regression uses much more complex cost functions like the sigmoid function, which maps value between (0,1).In this classification, the classifier should return aa set of outputs or classes based on the probability score given by the algorithm. Which ever class has higher score that data point is classified as that class data point.The cost function act as optimization objective and the aim are to minimize it. And redice the error. the cost is reduced by using the GrdientDescent algorithm, with the aim to reduce the cost function. It moves in the direction which largent descent so as to reach a minimum valueof the function.

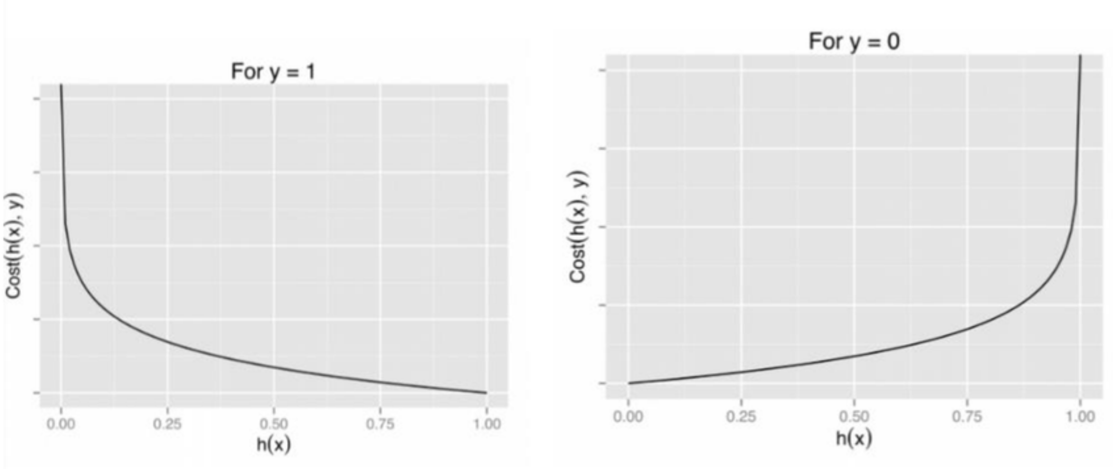

Fig. 1.5 Graph of logistic regression

CHAPTER - 2
LITERATURE REVIEW

2. Literature review

In the past, people used to use expert analysis to find fraudulent financial statements. In this way, people may not fully analyze the report data for its huge amount and wild range, which caused many shortcomings in judgment.

In the paper, Bermúdez L., Perez J.M., Ayuso M., Gomez E, Vazquez F.J. introduced a simulation-based approach by applying a Monte Carlo Bayesian Gibbs sampling for fitting an insurance fraud database using a dichotomous model. Their approach identified the likelihood of the data by using an asymmetric logit model and then assuming a proper prior distribution of the parameters of the model. These consideratios, combined with the Gibbs sampler, allow us to simulate on the basis of the exact posterior distribution of these parameters.
Comparing the standard and the Bayesian logit (with symmetric links) estimation results, they see that the Bayes lgistic model gives posterior estimations for the coefficients that are quite similar to the classical ones. A logistic model with classical inference gives almost the same estimations as a Byesian inference with noninformative normal priors. However, the Bayesian methodology is unable to signal the significance of the parameter associated with the variable *proximity*. Moreover, both models present a lack of fit due to the incorrect classification of zero cases.
A large diffrence can be observed in the confusion matrix when they consider a Bayesian skewed logit modl, which is more suitable for fitting data than a classical logistic model if the zero response is observed more often than the one response. Still, a nonsignificant parameter for *Proxim* has been found. Finally, the Bayesian procedure presents standard errors for the parametes slightly higher than those obtained with the classical logic.
These conclusions are analyzed in the context of insurance fraud, where the poor classifiction of hoens and fraudent claims implies high auditing costs for the inurance company. The auditng cost related to honest claims incorrectly classified as fraud will be notably reduced with the application of a skewed logit model. Since the Bayesian skewed logit model presented here is only used for fitting purposes, it is necessary to search for an asymmetric link function which would model the insurance fraud database so that the best predictive model would be obtained. In this

way, they have already used the most common asymmetric link functions (log-log and complementary log-log), but they have not found a relevant improvement in terms of predicting or classification performance. Therefore, a natural extension of this paper is looking for skewed link functions which may help them to obtain better predictions.[6]

In this paper Cecchini Mark, Aytug Haldun, Koehler J. Gary, Pathak Praveen develop a methodology for automatically analyzing financial text. They make several interrelated contributions. They develop a methodology to detect financial events. They extend technical research in computational linguistics via the dictionary creation method. This method extends the VSM model to include WordNet and tunes the weights to discriminate between documents from two different domains (event and nonevent). To validate the methodology empirically, they test it on two financial events, bankruptcy and fraud. They find that the dictionaries they create are able to discriminate fraud from non-fraud MD&As 75% of the time and bankrupt from non-bankrupt MD&As 80% of the time. They compare the results with quantitative prediction methods. Their methodology achieved superior results using the same data. They tested for complementarities by merging the quantitative data with the text data. They achieved the best results for both bankruptcy (83.87%) and fraud (81.97%) with the combined data. This shows that the text of the MD&A contains information that is complementary to the quantitative information. Using both quantitative and textual information improved the prediction of both financial events. The methodology could be used to help investors screen out companies at risk for bankruptcy. It could also be used by government and regulatory bodies as a tool to help determine the firms that may be committing fraud, and thus be considered for auditing. The methodology can be applied to available text for any financial problem where the goal is to create a dictionary (ontology) of discriminating concepts.

Future research can include gaining insight into the keywords or phrases in the financial text that lead to future changes in company valuation (big gains and losses). Finding the firms that were correctly classified by one method (quantitative or text) but not another could give accounting/financial researchers valuable information (e.g., the characteristics of firms that report financial and qualitative information differently). This could be explored more thoroughly in future research. The ability to quantify text information and place alongside financial information allows researchers to extend previous econometric models that only includd quantitative financial informaton.

Improvemets to the methodology could be made to reduce noise in the ontology and improve prediction accuracy. A natural methodological extension would be to further develop the methodology so that it can be useful in analyzing continuous variables (as opposed to binary outcomes) such as are commonly researched in accounting. A possible addition to the methodology would be the inclusion of a temporal component to track changes to the structure of text over time for individual firms. Such an extension would allow researchers to better understand how firm changes are manifest in financial texts. The method could be further developed to learn the context of multi-word phrases via WSD.[7]

Ngai E.W.T., Hu Yong, Wong Y.H. , Chen Yijun, Sun Xin determine the main algorithms used for FFD, they present a simple analysis of FFD and the data mining techniques identified in the articles. Twenty-six data mining techniques have been applied to the detection of financial fraud. Mortgage fraud, securities and commodities fraud, and mass marketing fraud are not listed because the techniques identified in their research have not been applied to these problems. The most frequently used techniques are logistic models, neural networks, the Bayesian belief network, and decision trees, all of which fall into the “classification” category. Of these techniques, logistic models are the most popular.[8]

In this paper Bhattacharyya Siddhartha, Jha Sanjeev, Tharakunnel Kurian examined the performance of two advanced data mining techniques, random forests and support vector machines, together with logistic regression, for credit card fraud detection. A real-life dataset on credit card transactions from the January 2006–January 2007 period was used in their evaluation. Random forests and SVM are two approaches that have gained prominence in recent years with noted superior performance across a range of applications. Till date, their use for credit card fraud prediction has been limited. With the typically very low fraud cases in the data compared to legitimate transactions, some form of sampling is necessary to obtain a training dataset carrying an adequate proportion of fraud to non-fraud cases. We use data undersampling, a simple approach which has been noted to perform well and examines the performance of the three techniques with varying levels of data undersampling. The study provides a comparison of performance considering various traditional measures of classification performance and certain measures related to the implementation of such models in practice. For performance assessment, they use a test dataset with much lower fraud rate (0.5%) than in the training datasets with different levels of undersampling. This helps provide an indication of the performance that may be expected when models are applied for fraud detection where the proportion of fraudulent transactions are typically low.

Encouragingly, all techniques showed adequate ability to model fraud in the considered data. Performance with different levels of undersampling was found to vary by technique and also on different performance measures. While sensitivity, G-mean and weighted- accuracy decreased with lower proportions of fraud in the training data, precision and specificity were found to show an opposite trend; on the F-measure and AUC, logistic regression maintained similar performance with varying proportions of fraud in the training data, while RF and SVM showed a decreasing trend on AUC and an increasing trend on F. Perhaps more informative from an application and practical standpoint is the fraud capture rate performance at different file depths. Here, random forests showed much higher performance at the upper file depths. They thus capture more fraud cases, with fewer false positives, at the upper depths, an important consideration in real-life use of fraud detection models. Logistic regression maintained similar performance with different levels of undersampling, while SVM performance at the upper file depths tended to increase with a lower proportion of fraud in the training data.

Random forests demonstrated overall better performance across performance measures. Random forests, being computationally efficient and with only two adjustable parameters which can be set at commonly considered default values, are also attractive from a practical usage standpoint. Logistic regression has over the years been a standard technique in many real-life data mining applications. In their study, too, this relatively simple, well-understood and widely available technique displayed good performance, often surpassing that of the SVM models. As noted earlier, no deliberate attempts at optimizing the parameters of the techniques were made in this study. Parameter tuning can be important for SVM, and balanced sampling has been noted to be advantageous in using Random Forests on imbalanced data. These carry the potential for better performance over that reported here and present useful issues for further investigation.

A factor contributing to the performance of logistic regression is possibly the carefully derived attributes used. Exploratory data analysis and variable selection is, of course, a time-consuming step in the data mining process, and where such effort is in doubt, the performance of logistic regression may be uncertain. For ease of comparison, models from all techniques in their study were developed using the same derived attributes. Random forests and SVM carry natural variable selection ability and have been noted to perform well with high dimensional data. Their potential for improved performance, when used on the wider set of available attributes, is an interesting issue for further investigation.[9]

In this paper Pai Ping-Feng, Hsu Ming-Fu, Wang Ming-Chieh focus on the detection of FFS as important and challenging issue that has been rigorously investigated in recent years, as the number of management fraud cases has increased. Many different kinds of technology have been introduced to deal with the audit-related risk, and the attempt to improve on these models continues. The current investigation presents an SVMFW model for analyzing financial statement data. The empirical results indicate that the proposed model is an effective and efficient alternative in detecting top management fraud. Rather than presenting complicated mathematical functions, the SVMFW model provides a set of comprehensible decision rules for auditors, who must allocate limited audit resource. In the future, other input features such as audit committees, boards with a significant proportion of outside members, CPA tenure and CEO duality might include in the SVMFW model to further enhance its effectiveness in analyzing financial statement data. Sequential forward selection (SFS), a typical heuristic searching scheme, identifies important features from unselected features and places them into a selected feature subset in each iteration. The support vector machine (SVM) proposed by Vapnik uses classification techniques based on statistical learning theory. SVM produces a binary classifier, the so-called optimal separating hyperplanes, through an extremely non-linear mapping of the input vectors into a high-dimensional feature space. SVM constructs a linear model to estimate a decision function using non-linear class boundaries based on support vectors. If the data are linearly separated, SVM trains linear machines for an optimal hyperplane that separates the data without error and into the maximum distance between the hyperplane and the closest training points. The training points that are closest to the optimal separating hyperplane are called support vectors. The knowledge acquisition problem has been identified as a major bottleneck in the expert system development process. Recent empirical research reveals that a classification and regression tree (CART) is significant in helping to extract certain types of knowledge within specific problem domains. CART, a statistical procedure introduced by Breiman et al., is a robust, easy-to-use decision tree tool that automatically shifts through large, complex databases in searching for and isolating significant patterns and relationships. CART adopted a recursive partitioning, a combination of exhaustive searches and intensive testing techniques to identify an informative tree structure in the data also referred to as a decision tree. The decision tree is then used to generate reliable, easy-to-grasp predictive models in the form of ''if-then" rules. Auditors can use these rules to allocate limited audit resources.[10]

Olszewski Dominik's approach is based on the user profiling technique utilizing LDA, and detecting fraudulent behaviour on the basis of the threshold- type classification with use of the KL-divergence. Consequently, their method requires the computation of the KL-divergence between two LDAs, which is an unsolved problem. Therefore, this paper focuses also on the issue of approximation of the KL-divergence between two LDAs, introduces three approximation methods, and chooses the most effective one. The fraudulent activity is indicated by crossing a previously defined threshold that causes the fraud alarm. In this paper, also a method for automatic threshold computation is proposed.

The method, proposed in this paper, is based on a classification algorithm that could be applied to any kind of detection problem (not only fraud detection), however, the relation to fraud lies in the fact of using the LDA probabilistic model, and in the fact that the problem of insufficient training data, which is particularly impeding in case of fraud detection, is overcome here. The LDA model provides an accurate description of a user profile, and as it is shown in, it can be successfully applied to fraud detection in a telecommunications problem. The issue of insufficient training data is overcome, because their method does not require a training process, like it is, for example, in vthe case of the neural-network- based approaches.

Our technique strongly relies on user profiling with LDA probabilistic model. Employing LDA for fraud detection in telecommunications, however, the difference between and their paper is that they detect whole fraudulent accounts, in contrast to, where single fraudulent calls are detected. Consequently, they apply a different classification algorithm with the original automatic threshold setting method. This kind of approach is also useful in real-world fraud detection problems. Furthermore, the approach proposed in requires a training phase, while their method is independent of this constraint.

The LDA model is an example of a probabilistic mixture model, i.e., a model described with a combination (linear combination or product) of certain probability distributions. A well-known example of such a model is the GMM, being a linear combination of Gaussian distributions. The estimation of the LDA model's parameters is described in, where the model was introduced and is described, itself.[11]

v

vKirkos Efstathios, Spathis Charalambos, Manolopoulos Yannis threw light on auditing practices as nowadays have to cope with an increasing number of management fraud cases. Data Mining techniques, which claim they have advanced classification and prediction capabilities, could facilitate auditors in accomplishing the task of management fraud detection. The aim of this study has been to investigate the usefulness and compare the performance of three Data Mining techniques in detecting fraudulent financial statements by using published financial data. The methods employed were Decision Trees, Neural Networks and Bayesian Belief Networks.

vThe results obtained from the experiments agree with prior research results indicating that published financial statement data contains falsification indicators. Furthermore, a relatively small list of financial ratios largely determines the classification results. This knowledge, coupled with Data Mining algorithms, can provide models capable of achieving considerable classification accuracies.

The present study contributes to auditing and accounting research by examining the suggested variables in order to identify those that can best discriminate cases of FFS. It also recommends certain variables from publicly available information to which auditors should be allocating additional audit time. The use of the proposed methodological framework could be of assistance vto auditors, both internal and external, to taxation and other state authorities, individual and ivnstitutional, investors, the stock exchange, law firms, economic analysts, credit scoring agencies and to the banking system. For the auditing profession, the results of this study could be beneficial in helping to address its responsibility for detecting FFS.

vIn terms of performance, the Bayesian Belief Network model achieved the best performance managing to correctly classify 90.3% of the validation sample in a 10-fold cross-validation procedure. The accuracy rates of the Neural Network model and the Decision Tree model were 80% and 73.6%, respectively. The Type I error rate was lower for all models.

The Bayesian Belief Network revealed dependencies between falsification and the ratios debt to equity, net profit to total assets, sales to total assets, working capital to total assets and Z score. Each of these ratios refers to a different aspect of a firm's financial status, i.e., leverage, profitability, sales performance, solvency and financial distress, respectively. The Decision Tree model primarily associated falsification with financial distress, since it used Z-score as a first level splitter.[12]

For the case of the data mining domain field, it took some decades before the application of this revsearch domain was projected from the academic world into the business environment (and more precisely as a fraud detection mean and as a market segmentation aid). As for the case of provcess mining, Jans Mieke, van der Werf Jan Martijn, Lybaert Nadine, Vanhoof Koen wish to accelerate this step and recognize already in this quite early stage which opportunities process mining offers to business practice. Process mining offers the ability to objectively extract a model out of transactional logs, so this model is not biased towards any expectations the researcher may have. In light of finding flaws in the process under investigation, this open mind set is a very important characteristic. Also, the ability to monitor internal controls is very promising.
vIn this paper, Jans Mieke, van der Werf Jan Martijn, Lybaert Nadine, Vanhoof Koen presented a case study in which they applied process mining in the context of transaction fraud. Given the procurement process of an organization using SAP as an ERP system, they applied the process diagnostics approach to discover the real process and to analyze flaws, i.e., to discover cases that are not compliant. This enables the explicit possibility of checking internal controls and business rules in more general. This way, process mining enables auditing by not only providing theory and algorithms
to check compliance but also by providing tooling that helps the auditor to detect fraud or other flaws in a much earlier stage. However, the case study also shows that, although tools are available, they are still quite premature. Therefore, they need to enhance tools like ProM to better automate the audit process and to visualize results for management.[13]

Kim Yeon kook J., Baik Bok, Cho Sung zoon undertake classification with three multi-class classifiers, multinomial logistic regression, support vector machine (with a linear kernel), and Bayesian networks, using stratified ten-fold cross-validation. As a sanity check, they test how the classifiers classify Enron Corporation, a notorious example of financial fraud. All three models classify the instance correctly, assigning highest class probability levels to the intentional-misstatement class.

To assess the classification performance, they select evaluation measures from prior studies. For example, used accuracy and the G-mean to compare the performance of cost- sensitive boosting algorithms in multi-class classification problems with imbalanced class distribution. On the other hand, used the total misclassification costs to evaluate the performance of sampling and threshold-moving methods in training cost-sensitive neural networks for both binary and multi-class classification problems. Since each measure used in the past studies has advantages and disadvantages, they use these three measures as a more balanced set of measures of the classification performance. Moreover, since the first objective in building misstatement detection models is to detect material misstatements, they use a measure that shows how well each model detects material misstatements, whether intentional or unintentional, as their fourth measure.[14]

In this paper Yeh Ching-Chiang, Chi Der-Jang, Lin Tzu-Yu & Chiu Sheng-Hsiung describe Rough set theory (RST) as a machine learning method that was introduced by Pawlak (1991) in the early 1980s. It has proven to be a powerful tool for uncertainty and is usually applied to data reduction, rule extraction, data mining, and granularity computation. Here, they illustrate only the basic ideas of RST that are relevant to contemporary work Support vector machines (SVM) are theories based on statistical learning theory and have become an increasingly popular nonparametric methodology for developing classification models. They realize the theory of VC dimension (Vapnik–Chervonenkis dimension) and the principle of structural risk minimum (SRM). The objective of this study is to increase the accuracy of FFS detection; they believe that non-financial ratios are the key factors in a corporation's FFS and provide invaluable information in FFS detection. For verifying the applicability of methodology, they also designed RSTþNN, stepwise regression þSVM, stepwise regression þNN, SVM and NN as the benchmarks.

The proposed RSTþSVM model was applied to a dataset in Taiwan. First, this study found that most prior studies adopted only financial ratios as independent variables. Although non-financial ratios are generally acknowledged to be a key factor in a corporation's FFS, they are usually excluded from early detection models. Therefore, they integrated both financial and nonfinancial ratios as potential predictive variables. Second, the related works did not pay much attention to finding and selecting important independent variables in developing FFS detection.

Third, a data mining approach (SVM and NN) is more accurate in predicting FFS than other multivariate statistical models. SVM finds the maximal margin (hyperplane) between two classes and gives good generalization performance on many business classification problems. In order to ensure a good classification process, the data inputs must be subjected to special treatment during the preparation. One of the reasons for this is that after getting data from experiments and many variables, it cannot, of course, be direcvtly inputted into the classifier because it will decrease the performance therein. Finally, the RST has become very popular among scientists worldwide and is now one of the most developed techniques in intelligent data analyses. The RST approach, by which redundant attributes in a multi-attribute information system can be removed without any information loss, is utilized as a preprocessor to improve FFS detection capability by SVM. Based on these reasons, they can conclude that the proposed RSTþSVM model outperforms the other FFS detecting models.

In order to verify the feasibility of this proposed RSTþSVM model, FFS detection tasks were vperformed using public companies' data between 1996 and 2007 in Taiwan. The contribution of this study can be summarized as follows. First, based on rough set reduction, some important variables for FFS detection were discovered according to their experiment results. Especially, non-financial ratios, which do provide valuable information in FFS detections. Second, the proposed RSTþSVM model provides better classification results than other FFS detecting models. Hence, they can conclude that the RSTþSVM model is a better alternative for conducting FFS detection tasks. In addition, the RSTþSVM model not only has better classification accuracy, but also the lowest incidence of Types I and II errors. Thus, the forecasting technique (RSTþSVM) can provide a decision support system for investors and governments.[15]

In this paper of Kotsiantis Sotiris, Tzelepis D., Tampakas V., purpose of the study was to use a vrepresentative algorithm for each described learning technique. The most commonly used C4.5 algorithm was the representative of the decision trees in their study. RBF algorithm - was the representative of the ANNs. The K2 algorithm was the representative of the Bayesian networks in their study. The 3-NN algorithm that combines robustness to noise and less time for classification than using a larger k for KNN was also used. Ripper was the representative of the rule-learners. Finally, the Sequential Minimal Optimization (or SMO) algorithm was the representative of the SVMs as one of the fastest methods to train SVMs.[16]

CHAPTER - 3

RESEARCH METHODOLOGY

3. Research methodology

3.1. About Data

The dataset is taken from the free public website Kaggle, which includes 22 variables among which 17 are financial andremaining are the non-financial variables. The intention of using non-financial variablevs is to improve the accuracy of detecction of fraud. In order to identify the variious types of financial reporting fraud, the seleected collection of financial variables should cover as many aspeectsv as possible. There are total of 240 companies listed in the datvaset among which 120 are classified as fraudulent and remaining as non-fraudulent.

A. Financial Variables

Variables	**Variable Description**
x1	Quick ratio
x2	Sales Growth
x3	Liquidity Ratio
x4	Operation revenue / average account receivable
x5	Rate of return on total assets
x6	Inventory turnover
x7	Operating profit / income before tax
x8	Net cash content of operating profit
x9	NCFPS
x10	Turnover of total capital
x11	Return on assets
x12	Operating profit ratio
x13	EPS

x14	Asset quality index
x15	Accounts receivable/ current assets
x16	Growth rate net profit
x17	Growth rate of net cash flow of operating activites

B. Non-financial Variables

Variable	Variable Description
x18	The proportion of the largest shareholder
x19	Board of directors
x20	Board of supervisors
x21	The proportion of independent directors
x22	LHSR(Behavioural aspect)

C. Descriptive Statistics

	Min	Max	Mean ± S.D.
x1	0.02704	8.78993	1.06034 ± 1.05680
x2	-81.28303	1160.93152	24.42644 ± 105.28405
x3	0.16113	8.79113	1.54670 ± 1.16503
x4	0.36027	3324.42767	54.67923 ± 244.50699

x5	7.63125	220.50679	54.05857 ± 24.27745
x6	0.00365	974.98648	11.94133 ± 70.19266
x7	-20.83488	4.02214	0.47371 ± 2.30085
x8	-1405.53898	95.37726	-7.93551 ± 118.54589
x9	-3.33460	4.32000	0.27264 ± 0.81329
x10	0.00431	9.10804	0.77142 ± 0.86311
x11	-46.80306	46.18315	5.42923 ± 7.95734
x12	-475.60620	58.76700	-0.74138 ± 49.67481
x13	-475.60620	2.32000	0.20917 ± 0.43805
x14	-1.20000	0.86654	0.22006 ± 0.29058
x15	0.00008	0.65354	0.16174 ± 0.13355
x16	-5534.31950	3809.30900	32.96119 ± 554.23800
x17	-4077.52524	2177.45421	-65.03040 ± 465.20060
x18	1.27555	78.97457	33.39915 ± 15.25724
x19	1.00000	11.00000	5.15000 ± 1.56932
x20	1.00000	5.0000	2.04583 0.73892
x21	0.12500	0.80000	0.38819 0.08750
x22	0.00000	78.84930	22.01327 20.89372

The above table shows basic descriptive statistics of these chosen companies, from which I can see that the range of several indicators is large, such as x8, x16 and x17.Due to the reasons of high dimension, high repetition, and having noise, original data need to be preprocessed, which includes data cleaning, data transformation and data reduction. In this paper I removed all the records which contain missing value and used sigmoid function shown below

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

to solve the problem of the inconsistent orders of magnitude. A *confusion matrix* is a tabular form of description that is often used to describe as how well a classification model (or “classifier”) performs on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix.

The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a huge set of variables into a short smaller one that still contains most of the information than in the larger set. Reduction of the number of variables of a dataset originally comes at the expense of accuracy, but the benefit in dimensionality reduction is to trade a little accuracy for simplicity, because smaller datasets are really simpler to explore and visualize and make analyzing data so much easier and faster for machine learning algorithms without extraneous variables to process.

3.2 Methodology

First, I used feature selection to reduce its dimensionality. As I knew, feature selection plays an important role in pre-processing. In this paper, I used PCA and Xgboost to do it. Principal component analysis (PCA) is a statistical method. By orthogonal transformation, a set of observations of possibly correlated variables converted into a set of linearly independent variables, which called principal component. Six variables among all are of great importance.

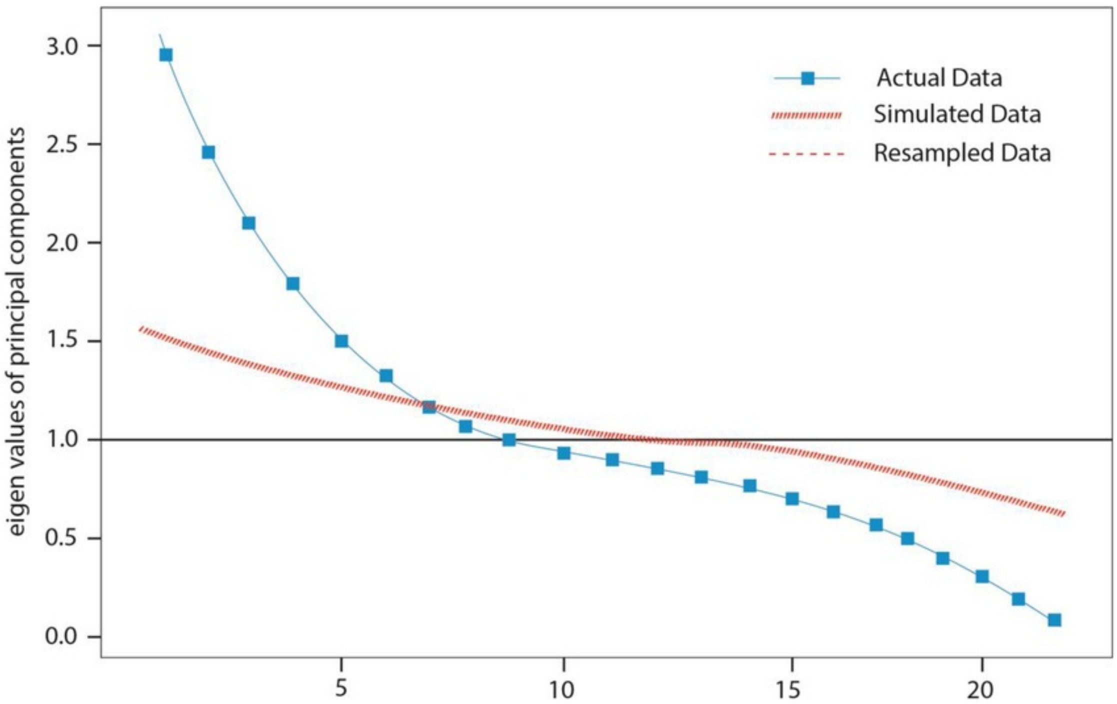

Fig 3.1 The result of feature selection using PCA.

Xgboost usually known as a trump card in a variety of competitions in the area of data mining. It also has the function of feature selection as shown in fig 3.2.

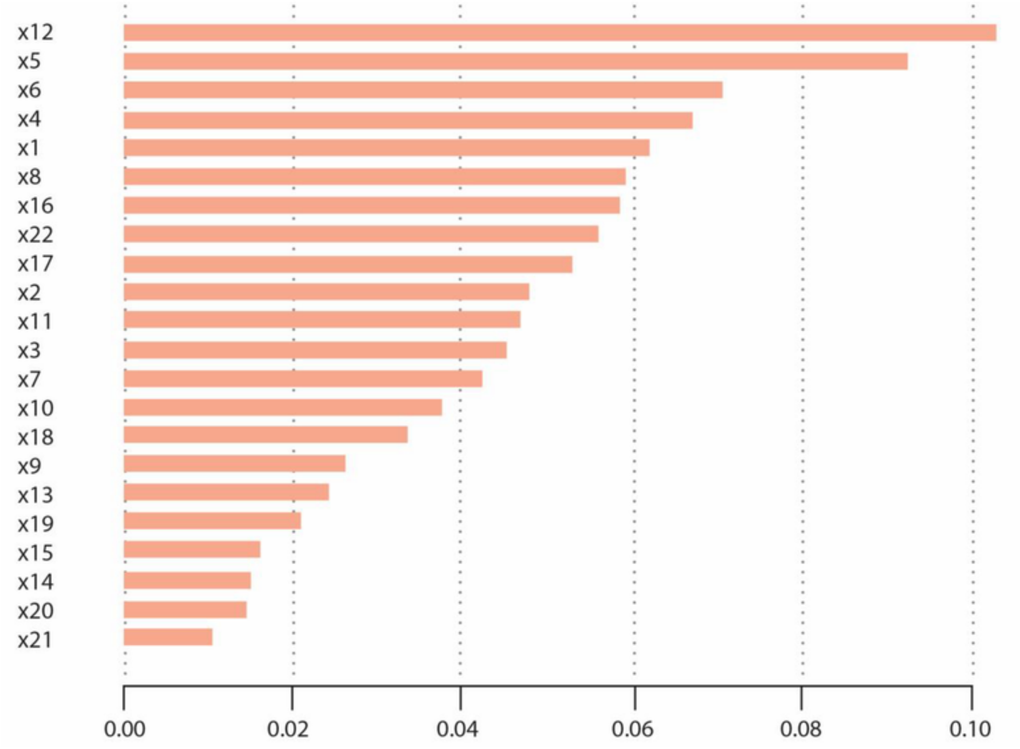

Fig 3.2 The result of feature selection using XGBoost

Second, I applied the five classification methods to two feature subsets selected by PCA and Xgboost, therefore ten models were created. At last, the results were compared and analyzed.

There are many classifiers available in data mining. Different methods can be used for the more comprehensive and multi-angle analysis of data.

Logistic regression is the regression analysis for binary variables. From the angle of implementation, logic regression is simple, easy to understand and realize. Its run fast and the computational cost is relatively low. SVM puts the original data into a higher dimension one using nonlinear mapping, and then an optimal decision hyperplane is established to maximize the distance between the two closest samples of the plane on both sides of the plane. Compared with other non-linear methods, SVM requires relatively few samples, and its goal is to minimize structural risk. SVM performs well in noisy financial data. Decision tree model is a tree structure that describes the classification of an instance, consisting of nodes and directed edges. The biggest advantage of decision trees is their interpretability to the model. ANN is inspired by the central nervous system of animals, which is a model of information processing of applied neurons. The distributed structure makes it have the same robustness as the human brain, while ANN is also good at self-learning, self-organizing. Random forest improves the decision tree by using the voting mechanism of multiple decision tree, it is an ensemble algorithm in essence. For example, if there are N samples and M variables (dimensions), the specific process of random forest is as follows: (1) Determine a value m, which is used to indicate how many variables eacth tree classifier chooses. (2) Collect k samples from the data set and use them to create k tree classifiers. In addition, k bags of external data are generated to be used for testing later.(3) After entering the classified sample, each tree classifier will classify it, and then all classifiers will determine the classification result according to the majority rule.

CHAPTER - 4

RESULTS

4. Results

At the beginning, I use machine learning methods to explore all the variables, the accuracy of SVM, RF,DT, ANN, and LR are shown. It is obviously that SVM was better than others in this condition. Random forest has the lowest accuracy among these methods.

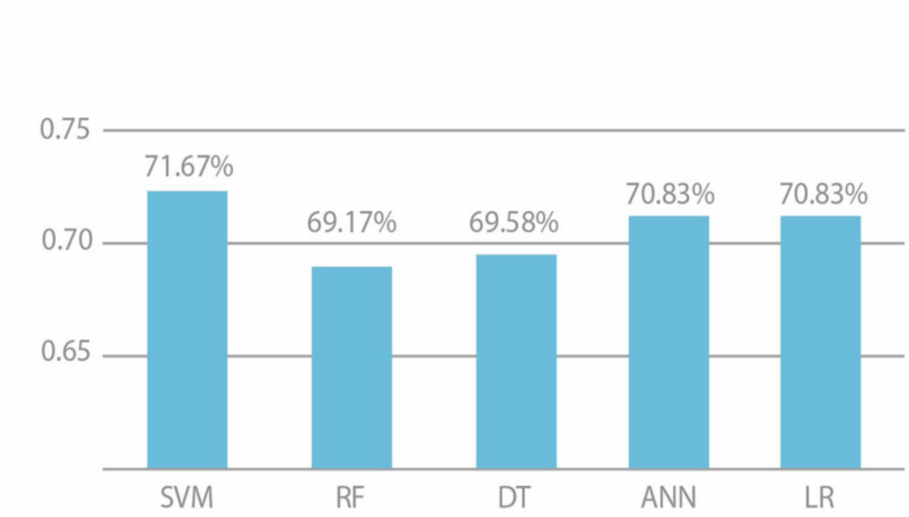

Fig 4.1 Accuracy with all the variables.

Second, in order to test five methods, I added the variables to the model one by one follow the order of importance from high to low. Thus, I do experiments with the most two important variables first, and then added the third, and go on. The results based on the importance of variables provided by PCA below indicate that with variables' growing, RF performs better and more stable.

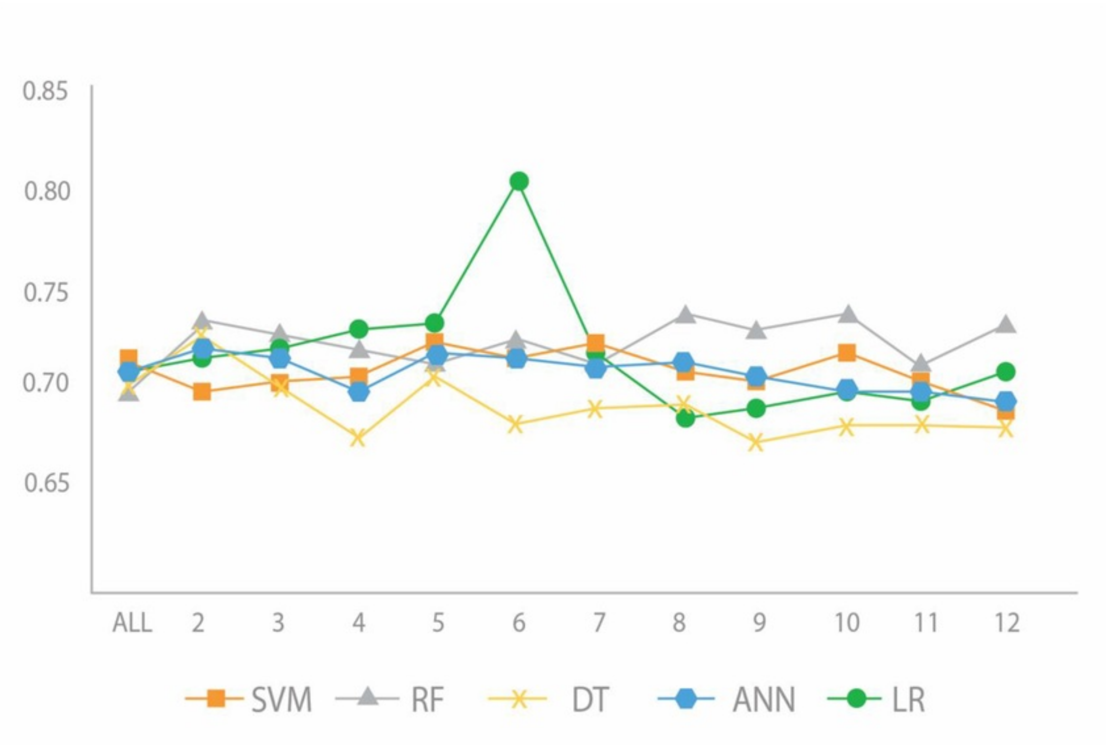

Fig 4.2 Results based on PCA

In the same way, I tested data based on the variables provided by RF.

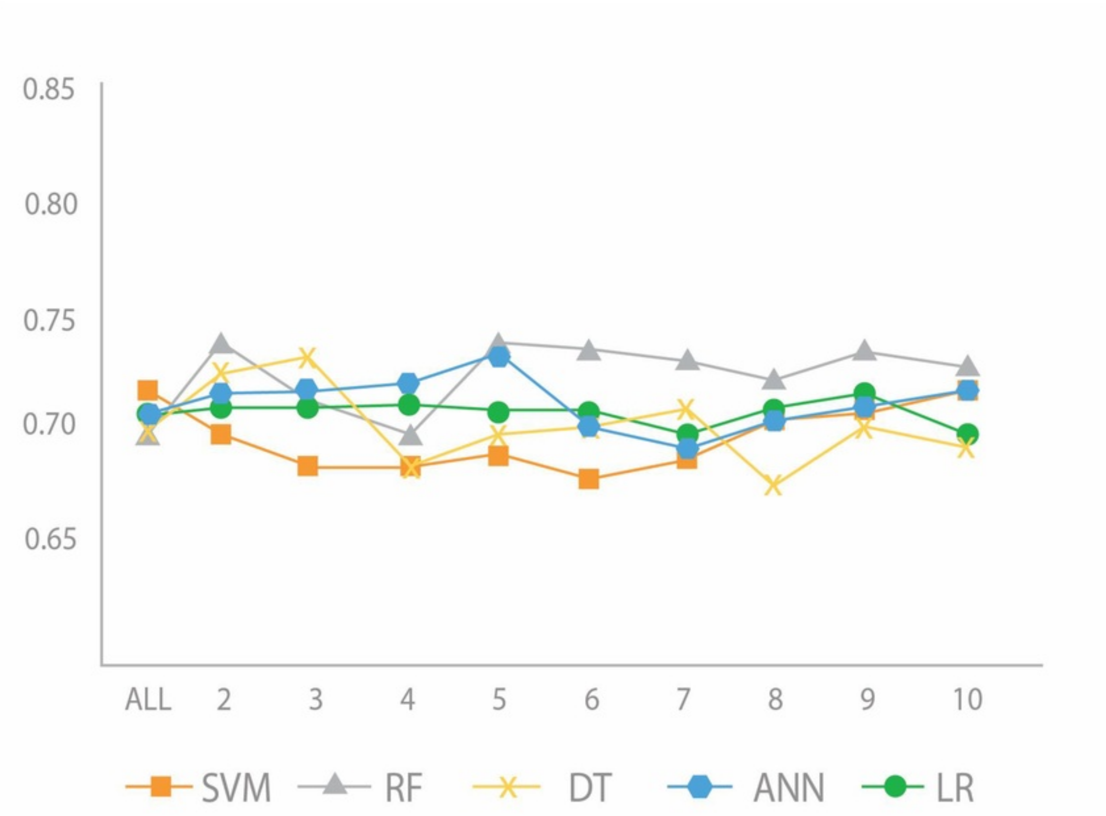

Fig 4.3 Results based on Xgboost

Random forest still performs good the number of selected variables get bigger. LR reach the highest accuracy when the number of selected variables is 6, but it's not stable.

CHAPTER - 5

FINDINGS AND RECOMMENDATIONS

5. Findings and Recommendations

In order to test two feature selection functions and to find which variables are of great importance in machine learning, I plot the average value of five accuracy. We can find that when the number is 2 or 5, We may get a satisfied result.

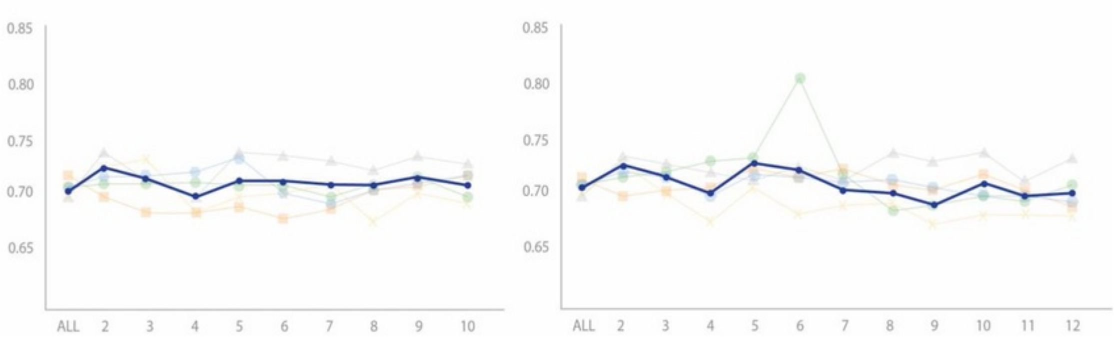

Fig 5.1 Average values of fig 4.2 and fig 4.3

At last, to compare these five methods more intuitively, I put all variables' combination into figure 5.1, from which I find that RF do the best performance.

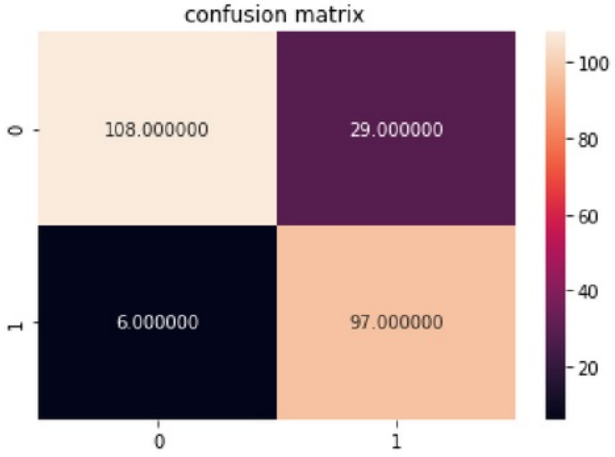

Fig 5.2 For Support Vector Machine(SVM)

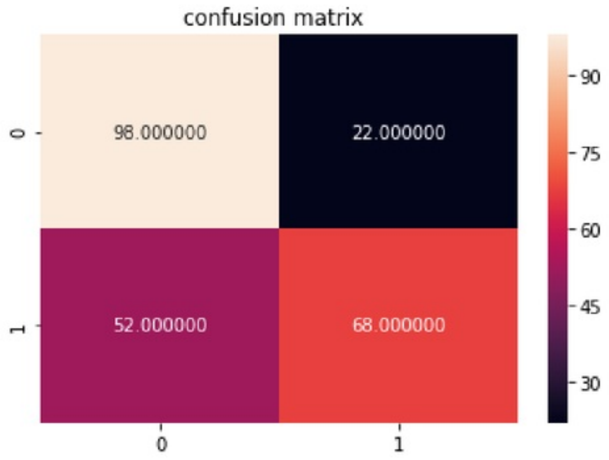

Fig 5.3 For Random Forest(RF)

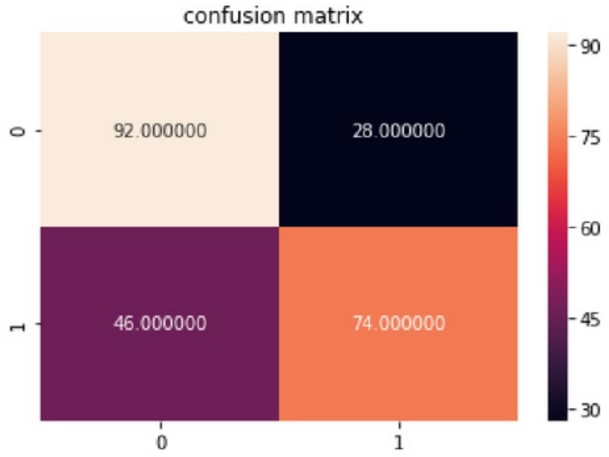

Fig 5.4 For Decision Tree(DT)

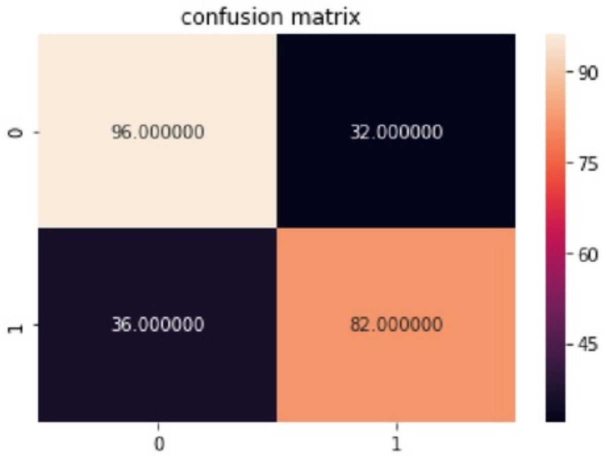

Fig 5.5 For Artificial Neural Network(ANN)

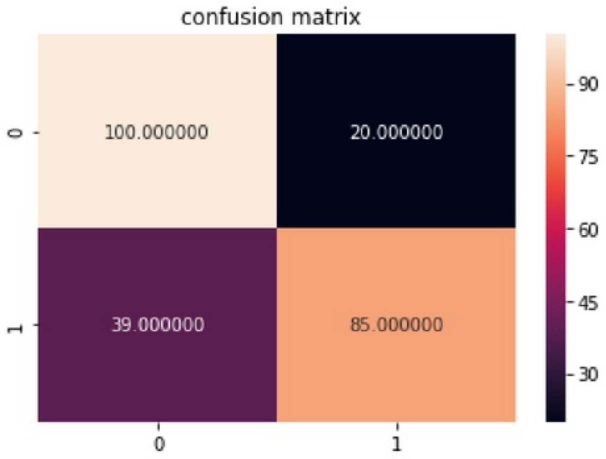

Fig 5.6 For Logistic(Regression)

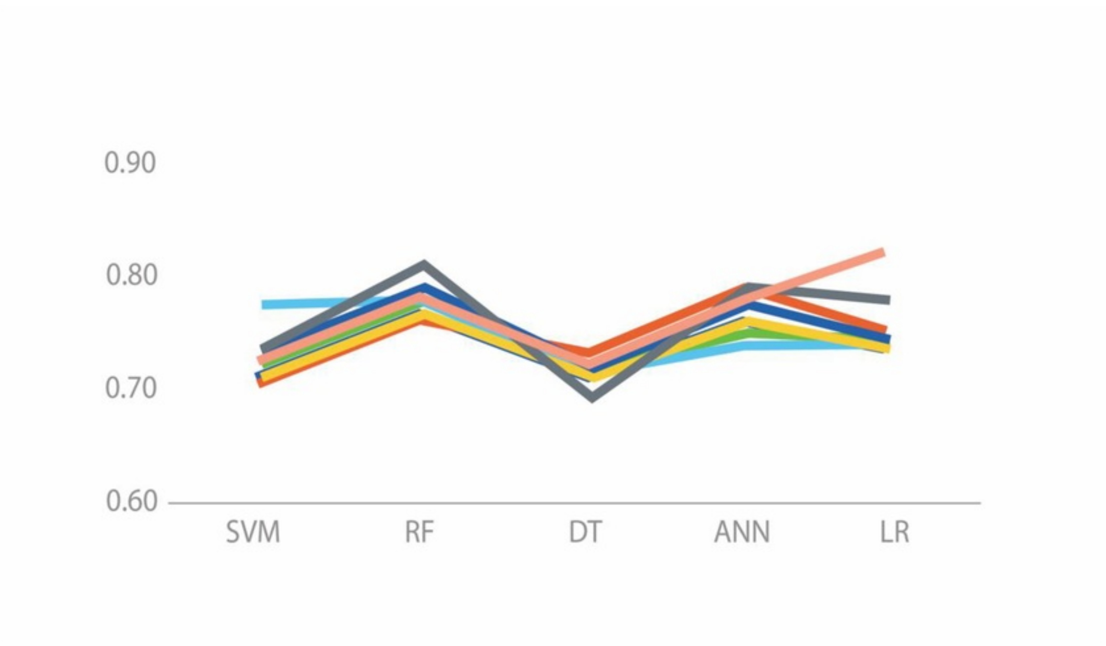

Fig 5.7 Results of all variables' combination

The research has three contributions:First,to calculate and analyze the factors that affect the fraud behavior. According to our research, the effect of x12 and x13 on the model is confirmed by both two feature selection methods. Operating profit margin is the ratio of an enterprise's operating profit and operating income. There's a big risk of financial fraud when there's a big loss in the firm and a lower profit margin, it would be reflected from indicators related to profit. EPS is generally used to measure common profit level and investment risk, and to reflect the operating results of an enterprise. In order to cover losses or exaggerate earnings, the companies have a tendency to increase EPS. Second, consider the influence of the number of variables on the model. The results indicate that 2 or 5 variables are better than the others. Third, I compared the performance of five machine learning methods and found that among them, random forest has the following advantages: (1) it is good at processing high-dimensional data; (2) it may avoid overfitting to some extent; (3) it has good robustness and stable results.

CHAPTER - 6

LIMITATIONS OF THE STUDY

6. Limitations of the study

There limitations in the research are:

- the data is not large enough and variables can be more varied and more innovative.
- The practical problems are far more complicated than the ones in the research and more factors should be taken into consideration when designing variables
- One limitation of this study is that the transactions in the financial statements , the derived attributes can have the same value for both the transactions.So the non-availability of the time stamp makes these derived attributes less viable and precise.Thus it is the limitation of the study.
- There are problems in the audit and the the insufficiency in the auditing procedure can prevent the detection of the fraud from appearing in the statements.
- Another limitation of the proposed approach is the case associated with cases which may fall outside the threshold circle of the classification algorithm and not necessarily fraudulent or may not be either.
- The standard that respects the auditor's consideration of the risk then that fraud and/or error can exist, and that clarifies the arguments on the inherent limitations of an auditor's ability to detect an error and fraud, particularly management fraud. Moreover, it should also be emphasized that the distinction between management and employee fraud and elaborates on the discussion concerning fraudulent financial reporting.

CHAPTER - 7

BIBLIOGRAPHY/ REFERENCES

7. Bibliography/References

[1] *"What is finance"*. Retrieved from https://www.investopedia.com/ask/answers/what-is-finance/

[2] *"Finance Definition"*. Retrived from https://corporatefinanceinstitute.com/resources/knowledge/finance/what-is-finance-definition/

[3] *"Corporate fraud"*. Retrieved from https://www.investopedia.com/terms/c/corporate-fraud.asp

[4] *"Detecting financial fraud"*. Retrieved from https://www.investopedia.com/articles/financial-theory/11/detecting-financial-fraud.asp

[5] Jiawei Han, Micheline Kamber, Jian Pei, "*Data mining Concepts and Techniques*". Retrieved from a text book with the given name and authors of the book.

[6] Bermúdez L., Perez J.M., Ayuso M., Gomez E, Vazquez F.J.(August 2007),*"A Bayesian dichotomous model with asymmetric link for fraud in insurance"*. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167668707000947

[7] Cecchini Mark, Aytug Haldun, Koehler J. Gary, Pathak Praveen(2010), *"Making words work: Using financial text as a predictor of financial events"*. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167923610001181

[8] Ngai E.W.T., Hu Yong, Wong Y.H. , Chen Yijun, Sun Xin(August 2010), *"The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature"*. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167923610001302

[9] Bhattacharyya Siddhartha, Jha Sanjeev, Tharakunnel Kurian, Westland J. Christopher(August 2010), *"Data mining for credit card fraud: A comparative study"*. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167923610001326

[10] Pai Ping-Feng, Hsu Ming-Fu, Wang Ming-Chieh(October 2010), *"A support vector machine-based model for detecting top management fraud"*. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0950705110001632

[11] Olszewski Dominik(August 2011), *"A probabilistic approach to fraud detection in telecommunications"*. Retrieved from https://www.sciencedirect.com/science/article/pii/S0950705111001973

[12] Kirkos Efstathios, Spathis Charalambos, Manolopoulos Yannis(2007), *"Data Mining techniques for the detection of fraudulent financial statements"*. Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417406000765

[13] Jans Mieke, van der Werf Jan Martijn, Lybaert Nadine, Vanhoof Koen(2011), *"A business process mining application for internal transaction fraud mitigation"*. Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417411006865

[14] Kim Yeon kook J., Baik Bok, Cho Sung zoon(February 2016), *"Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning"*. Retrieved from https://www.ieee.org

[15] Yeh Ching-Chiang, Chi Der-Jang, Lin Tzu-Yu & Chiu Sheng-Hsiung(May 2016), *"A Hybrid Detecting Fraudulent Financial Statements Model Using Rough Set Theory and Support Vector Machines"*. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/01969722.2016.1158553

[16] Kotsiantis Sotiris, Tzelepis D., Tampakas V.(November 2005), *"Forecasting fraudulent financial statements using data mining"*. Retrieved from https://www.researchgate.net/publication/228084523_Forecasting_fraudulent_financial_statements_using_data_mining

CHAPTER - 9

PLAGIARISM REPORT

8. Plagiarism Report

MRP

ORIGINALITY REPORT

17%	9%	14%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Jianrong Yao, Jie Zhang, Lu Wang. "A financial statement fraud detection model based on hybrid data mining methods", 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD), 2018 Publication	14%
2	ethesis.nitrkl.ac.in Internet Source	2%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		