ANALYSIS OF BANKRUPTCY USING DATA MINING APPROACH

ONG AI PING

UNIVERSITI UTARA MALAYSIA

2009

ANALYSIS OF BANKRUPTCY USING DATA MINING APPROACH

A project submitted to the Faculty of Information Technology in partial fulfillment of the requirement for the degree

Master of Science (Information Technology)

Universiti Utara Malaysia

By

ONG AI PING

PERMISSION TO USE

In presenting this project in partial fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for copying of this project in any manner, in whole or in part, for scholarly purposes may be granted by my supervisor, in her absence, by the Dean of the Faculty of Information Technology. It is understood that any copying or publication or use of this project or parts thereof for financial gain should not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my project.

Request for permission to copy or to make use of material in this project, in whole or in part should be addressed to:

Dean of the Faculty of Information Technology
Universiti Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman

ABSTRAK

Kajian ini berkaitan dengan pembangunan model ramalan rangkaian neural untuk syarikat yang ingin membuat ramalan tahap syarikat sama ada akan menghadapi kebankrapan. yang Dalam pada itu, Jumalah 367 data set adalah diperolehi daripada *Kuala Lumpur Stock Exchange* (KLSE) and Bank Negara Malaysia. Data ini seterusnya dianalisis dengan menggunakan asas statistic, *frequency* dan *cross tabulation* untuk mendapatkan lebih banyak maklumat berkaitan data. Pada peringkat awal, data adalah diklasifaikan dengan menggunakan *logistic regression*. Seterusnya ianya ditrain dengan rangkaian neural untuk mendapatkan model kebankrupan. Dimana, capaian menunjukkan adalah lebih sesuai dengan model yang mengandungi 12 nod *input*, 6 nod *hidden layer* dan 1 nod untuk *output*. Model yang dipilih menunjukkan generalisasi 100%. Metodologi ini sepatutnya memperolehi pendekatan baru kepada paten yang wujud dalam data ini. Oleh itu, rangkaian neural amat berpotensi untuk menyokong ramalan kebankrupan ini.

ABSTRACT

This study involves the development of neural network prediction model to predict the stage of bankruptcy of a company. A total of 367 data was attained from the Registrar of Business and Companies, Kuala Lumpur Stock Exchange (KLSE) and Bank Negara Malaysia (Central Bank of Malaysia). The data was then analyzed by considering the basic statistics, frequency and cross tabulation in order to get more information about the data. Initially, the data was classified using logistic regression. In addition, it was also trained using neural network in order to obtain the bankruptcy model. The findings show that the most suitable prediction model consist of 12 nodes of input , hidden layer 6 node and one output layer. The generalization performance of the selected model is100%. This methodology should be able to provide some new insight into the type of pattern that exists in the data. Thus, neural network has a great potential in supporting for predicting bankruptcy.

ACKNOWLEDGEMENT

Sincere thanks are due to my both supervisors, Associate Professor Fadzilah Siraj and Puan Nur Azzah Binti Abu Bakar for their patiently navigating and generously sharing their rich source and knowledge with me.

Evenly thanks to my course mate that zealously advice me and helping me all over this project.

Further gratitude must go to my beloved family members who have supporting and encourage me during completing this project.

Last and not least, I would like to express my deep appreciation to all who are helped me to completed this project whether in direct or indirect way.

Thanks

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Networks
BPN	Back-Propagation Neural Network
CRISP	Cross Industry Standard Process
GA-BP	Genetic Algorithm And Back Propagation
KDD	Knowledge Discovery In Databases
KLSE	Kuala Lumpur Stock Exchange
LRA	Logistic Regression Analysis
LSA	Latent Semantic Analysis
MDA	Multivariate Discriminant Analysis
MLP	Multi-Layer Perceptron
PLSA	Probabilistic Latent Semantic Analysis
RBAC	Role-Based Access Control
ROB	Registrar Of Business
ROC	Registrar Of Companies
RST	Rough Set Theory
SVM	Support Vector Machines

CHAPTER 1

INTRODUCTION

This study focuses on using data mining approach for analysis of bankruptcy. The aim of the study is to alert and give the warning signs in earlier stage to the company's that facing financial problem and almost to bankruptcy.

1.0 Background

Bankruptcy refers to the firms which are unable to pay debts and are either declared bankrupt in terms of Commercial Code (1857, Part 111, Title 1) or dissolved and wound under the Companies Act 1995 (Section 35(c) and Section 214 subsection 2 paragraph (a)(ii)). Bankruptcy is a proceeding in which a court administers the estate (i.e. the property and other assets) of the debtor for the benefit of the creditors. A debtor (i.e. a person or business who owes money to others) may choose to file a bankruptcy proceeding to resolve a hopeless financial situation, or to stave off the collection of debts for a period of time to allow for financial re-organization.

Recently, the number of companies declared as bankrupt due to the recession has increased. Therefore, the development of the bankruptcy prediction model has been considered as important, as bankruptcy prediction can have major impact on lending decisions and profitability of financial institutions. Bankruptcy prediction method has been developed by Altman and Beaver since 1960s. Early studies of bankruptcy prediction used statistical techniques such as multiple discriminant analysis (MDA) (Altman, 1968; 1983), logit (Ohlson, 1980) and probit (Zmijewski, 1984). Recently, many studies have established that Artificial Intelligence technique such as neural networks (NN) can be used as an alternative method to traditional statistical method (Kyung et al., 2004)

In recent times analyzing data with different methods has gained more popularity as new learning paradigms, as a result multi-strategy learning and data mining have emerged. For example multi-method analyses can lead to new approaching of the data and to better understanding suitable methods for the problem. Different induction methods have frequently been analyzed in various domains, e.g. Michie et al. (1994), Weiss and Kapouleas (1989). The results from different comparisons have varied. However, in most cases there does not seem to be a constant overall best method, but different domains have their own characteristics which are met by different methods. (Hekanaho et al, 1998).

The emergence of advances technology have increased the availability of data for analysis purposes up to a point that it exceeds the human capacity to manually analyze them. Proceeding to 1990, data collected by banks, credit card companies, department stores and others financial institutions are rarely used advance of its intended usage. With the emergence of data mining, this large set of data can be used to discover patterns and trends that will be useful for decision-making. The technique, that integrates with the ideas form statistics, machine learning, database technology and high performance computing, is useful for extracting hidden patterns from data in order to find the nuggets of knowledge. It is used in a wide range of applications, such as marketing, fraud detection and scientific discovery. As for business applications, it is used to find new market opportunities from the data stored in operational data bases which formerly are not apparent to managers because the volume of data is too large or they are generated too quickly by experts.

This study focuses on using data mining approach for analysis of bankruptcy. The aim of the study is to alert and give the warning signs in earlier stage to the company's that facing financial problem and almost to bankruptcy. Data mining approach are saves resources while at the same time maximizes the efficiency and increases productivity without increases the cost.

1.2 Problem Statement

Bankruptcy is an important apparatus to provide significant with debt problems. Bankruptcy often stated as a “last resort” for financially troubled consumers. Bankruptcy should be neither the first option nor the last resort. To predict bankruptcy, there have a number of difficulties to analyzing bankruptcy model that reflects the company financial situation. For example, identify the factor analysis to predict bankruptcy. There are a lot of reasons due to bankruptcy. One of the reasons is liquidity or cash-burning startups that run out of money.

Others problem that contributed to afford this study include given a set of parameters (mainly of financial nature) that describes the situation of a company over a given period, predict the probability that the company may become bankrupted during the following year

1.3 Research Objective

The main objective of the study is to find the hidden pattern of dataset of bankruptcy company data.

i. To identify bankruptcy model using data mining approach.

ii. To evaluate bankruptcy models in (i).

1.4 Scope of the Study

This study uses descriptive data mining approach as well as descriptive statistic and correlation to uncover the hidden information from the business insolvency or bankruptcy company data.

The datasets are from Kuala Lumpur Stock Exchange (KLSE), Registrar of Business (ROB) and Registrar of Companies (ROC) are chosen to develop the propose model. The datasets are preprocessed and transformed into appropriate values in order to ensure better classification model development takes place.

1.5 Research Questions / Hypotheses

The research question is:

How to identify and evaluate bankruptcy model based on bankruptcy company data.

1.6 Significance of the Study

This study will uncover the hidden information within the bankruptcy company's data. The model been used to predict the bankruptcy or insolvency of the company.

1.7 Conclusion

This study focuses on using data mining approach for analysis of bankruptcy model which aim to alert the company's that facing financial that almost bankrupt in earlier period. Fundamentally, data mining is concerned with analysis of data by using software techniques such as SPSS for finding patterns and regularities in sets of data. Whilst it is refers to the application of algorithms for extracting patterns from data.

This study uses neural network and descriptive data mining approach as well as descriptive statistic, correlation and logistic regression to uncover the hidden information from the bankruptcy company data.

CHAPTER 2

LITERATURE REVIEW

This chapter describes the review of the relevant literature on bankruptcy. The initial bankruptcy modeling started with financial ratios analysis and followed by statistical analysis or mathematic analysis and then emerged of artificial intelligent technique.

2.1 Factor Analysis of Finance and Banking

To analyze bankruptcy, first of all should identified factor analysis and financial ratios. In a simple firm value model the impact of the bankruptcy probability on the valuation of equity and debt, which are assumed to be not publicly traded. Schmidt, (2009) find out for the distressed company, which usually has high debt and low equity, the impact becomes increasingly important. Thus, disregarding this yields an overvaluation of debt and an undervaluation of equity.

In other hand, Mierzejewski, (2006) find out since the capital structure affects the performance of financial institutions confronted to liquidity constraints, the Economic Capital is determined by the maximisation of value. Schied and Schoeneborn, (2008) consider the infinite-horizon optimal portfolio liquidation problem for a von Neumann-Morgenstern investor in the liquidity model. Using a stochastic control approach, they characterize the value function and the optimal strategy as classical solutions of nonlinear parabolic partial differential equations. Furthermore analyze the sensitivities of the value function and the optimal strategy with respect to the various model parameters.

Besides, Cipollinia and Missagliab, (2005) focus on the measurement of the capital charges of a bank against expected and unexpected losses affecting the bank loan portfolio. Therefore, Angelidis and Lyroudi, (2006) had examines the productivity of the 100 larger Italians banks for the period 2001-2002. Inputs and outputs are used as nominal values (millions of euros) and as the natural logarithms of these values. The mean error between the actually total factor productivity and the estimated one is calculated according to both approaches. Moreover, the weighted arithmetic mean of the Malmquist productivity index is calculated in addition to the geometric mean. Also, the correlation coefficient and the ranking correlation coefficient are computed to shed more lights to the relationship between bank' size and its performance. The empirical results revealed that the use of natural logarithms and neural networks regression reduces the errors in the estimates.

Park, (2008) examines how banking market concentration affects small businesses finance. Using the Survey of Small Business Finance, the empirical model show that bank concentration may unfavorably affect the amount of credit that supplied to small businesses. Result show that bank concentration lowers the overall debt-to-asset ratio of small firms. Therefore, suggesting that credit from non-bank institutions do not fully make up the effect of bank concentration. While, He, (2002) reviews the nature of non-performing assets in the Indian banking system and discusses the key design features that would be important for the Asset Reconstruction Companies to play an effective role in resolving such non-performing assets.

Topaloglou et al, (2005) consider the alternative means for controlling currency risk exposure in actively-managed international portfolios. They extend multi-stage stochastic programming models to incorporate for optimal selection of either forward contracts or currency options for equivocation purposes and get used to valuation procedure to price currency options consistently with discrete distributions of exchange rates that are used in the context of stochastic programming model. They find that optimally selected currency forward contracts yield superior results in comparison to single protective puts per currency. However, option-trading strategies with suitable payoffs can improve performance in terms of higher portfolio returns.

2.2 Bankruptcy Prediction Modeling and Technique

Neural network has been proved in many ways and in a number of publications to be a real challenger to statistical methods, especially in logit analysis for predicting failures. Hekanaho *et al.*, (1998) compare rule-based learning with neural networks and logit analysis using a larger data set consisting of 570 companies. The study reveals neural networks and rule-based learning perform better than logit analysis, but there is substantial variation in the results depending on the sample size and time period. Lei and Chan (2003) study the efficacy of applying rule-based classifiers to bankruptcy prediction problem. They introduce an inference engine that applies rule sets generated by inductive learning programs based on rough set theory for binary classification problems. Experimental results show that the proposed classifier has better performance than the classic Altman's Z-score model for the bankruptcy prediction problem in five out of six testing data sets.

In addition, Yeung *et al*., (1998) propose a multiple classifier system which is embedded in a multiple intelligent agent system to predict the financial health of a company. In the model, each individual agent, or classifier, makes a prediction on the likelihood of bankruptcy based on only partial information of the company. Each of the agents is an expert, having certain part of the knowledge which is represented by features of the company. The decisions of all agents are combined together to form a final bankruptcy prediction.

The experiments conducted by Abdelwahed and Amir (1998) reveal promising results of model for the forecasting of firms insolvency, in terms of predictive accuracy and predictability. The combination of the two sub-models which is genetic algorithms and neural networks had improves the results as well in training as in test. Sai *et al*. (2007) suggest a hybrid approach to Chinese listed company bankruptcy prediction, using a GA-BP (genetic algorithm and back propagation) model to construct a bankruptcy prediction model with variables derived by rough set theory (RST). An example is given to validate this model. The results show our hybrid model has higher prediction accuracy and less execution time in bankruptcy prediction when compared against GA-BP algorithm.

Furthermore, Charalambous and Martzoukos (2005) carry out a hybrid evaluation methodology to propose and test (test what) for improving the efficiency of contingent claims pricing by combining Artificial Neural Networks (ANN) and conventional parametric option pricing techniques. With one application on financial derivatives and one on real options the method's superiority is demonstrated. The resulting efficiency is instrumental for real time applications.

In the past, various statistical techniques, such as univariate and multivariate discriminant analysis have been used in the modeling of corporate bankruptcy prediction. Nasir *et al.* (2005) use domain expert knowledge to select and organize data in the modular neural network architecture constructed for their study. There are three sub-networks representing the periods, 1994, 1995, and 1996. Each sub-network is made of five adjacent networks representing the Balance Sheet network, the Profit and Loss network, the Financial Summary network, the Key Financial Ratios network, and the Economic and Political factors network. These adjacent networks although coupled but not linked at the input level represents five facets of failure in predicting corporate bankruptcy. The training sets represents data for 2500 companies selected randomly from a population of 270,000 sample. The trained neural network will access 435,000 data records before making a prediction for the particular company. The results obtained shows that neural networks outperform statistical techniques in modeling corporate failure prediction.

Additionally, Nakaoka (2006) describes a new bankruptcy prediction system available even in such cases by adopting "Cash Flows" that ought to be more essential indices than the profit, sales and so on in the bankruptcy. First, the system selects cash flow indices by factor analysis to huddle the money changes. Second, the system makes a SOM map and predict bankruptcies by evaluating not individual company indices but interrelationships between companies indices by SOM. Third, prediction errors and their accuracies are discussed. Results show that the method with those of the discriminant analysis including Altman's Z-score and show the usefulness of their method.

In order to develop the model, requests to derive some variables and analyze the relationship between good and bad credits. Yoon *et al.* (2007) classify twelve variables that are significant in predicting good or bad risk for small and micro business, which are categorized into the business period, scale for sale, fluctuation in sales, sales pattern and business category's bankruptcy ratio. They utilize the new statistical technique to support vector machines (SVM) as a classifier. The grid search technique been used to find out better parameter for SVM. The result shows that credit card sales information could be have good substitute for financial data on business credit risk in predicting the bankruptcy of small micro business. They find out SVM performs best compare with other classifiers such as neural networks, CART, C5.0, multivariate discriminant analysis (MDA) and logistic regression analysis (LRA).

Concurrently, Shin *et al*. (2004) investigate the efficacy of applying support vector machines (SVM) to bankruptcy prediction problem. Since SVM captures geometric characteristics of feature space without deriving weights of networks from the training data, it is capable of extracting the optimal solution with the small training set size. The proposed classifier of SVM approach outperforms back-propagation neural network (BPN) to the problem of corporate bankruptcy prediction. The results demonstrate that the accuracy and generalization performance of SVM is better than that of BPN as the training set size gets smaller.

Predicting the financial health of companies is a problem of great importance to various stakeholders in the increasingly globalized economy. Vieira *et al*. (2004) apply several learning machines methods to the problem of bankruptcy prediction of private companies. Financial data obtained from Diana, a database containing 780,000 financial statements of French companies, are used to perform experiments. Classification accuracy is evaluated with respect to Artificial Neural Networks, Linear Genetic Programming and Support Vector Machines. They analyze both type I (bankrupted companies misclassified as healthy) and type II (healthy companies misclassified as bankrupted) errors on three datasets containing balanced and unbalanced class distribution. Linear Genetic Programming has the best accuracy in the balanced data while Support Vector Machines is more stable for the unbalanced dataset. As the results, though preliminary in nature, demonstrate the tremendous potential of using learning machines in solving important economics problems such as predicting bankruptcy with accuracy.

Whereas Foster and Stine, (2001) develop and illustrate a methodology for fitting models to large, complex data sets. The methodology uses standard regression techniques that make few assumptions about the structure of the data. They accomplish this with three small modifications to stepwise regression: First, add interactions to capture non-linearities and indicator functions to capture missing values then exploit modern decision theoretic variable selection criteria; and estimate standard error using a conservative approach that works for heteroscedastic data. Omitting any one of these modifications leads to poor performance.

2.3 Data Mining

The explosive growth in data has generated an urgent need for techniques and tools for extracting useful hidden information from the data. Marchiori *et al*. (2002) describe that information from the data is used to analyze the user behavior, to improve the services provided and to increases the business opportunities. Data mining refers to discovery process of hidden information from the data.

Ikizle and Guvenir (2002) carry out most of the data mining algorithms to produce a long list of rules in which it is up to user to find the ones that are really important and profitable. They introduced rule interestingness measures and a new rule selection mechanism. An interesting rule selection mechanism is proposed by combining those criteria in an effective manner. The important action to take will be showing the obtained results to users of the domain and get feedback about the rules' real interestingness values.

McLaren (1999) describes that data mining is slowly gaining acceptance as a method for obtaining business intelligence from corporate data banks. An issue which is arises from the problems, namely the effective design of a database to support data mining, and the integration of these designs with existing data warehouse designs. The requirements of decision support applications, including data mining, are entirely different from traditional transaction processing systems and therefore require new database design methodologies. a review of the database design techniques employed by established decision support applications are structured and highlighted to appropriateness data mining solutions. Then introduces a number of advanced design techniques which have particular significance to supporting enterprise data mining.

Data warehousing and data mining are technologies that deliver critical and optimally useful information to facilitate performance analysis of business organizations. These technologies are not only an emerging trend in information technology but also a booming market in a range of industries. Fang and Tuladhar (2004) describes the key components that comprise a course which would introduce both data warehousing and data mining technologies to a graduate program of information technology. Even though, Ahmaed et al, (1998) carried out an effective data mining system for mining multiple-level knowledge from data warehouse, database and of raw data is proposed. The data warehouse represents the backbone of the proposed architecture. The mining and OLAP kernel includes generic analysis modules for performing a wide spectrum of applications. Active data mining is adopted to support knowledge-driven business processes. Continuously gathered business data is partitioned according to application-dependent time periods.

Many organizations have their digital information stored in a distributed systems structure scheme, From a classical data mining view, where the algorithms expect a denormalised structure to be able to operate on, heterogeneous data sources, such as static demographic and dynamic transactional data are to be manipulated and integrated to the extent commercial association rules algorithms can be applied. Zhao *et al*. (2007) find out the rules produced are interesting, valuable, complete and understandable, which shows the applicability and effectiveness of the new method.

Besides, Molloy *et al*. (2008) find out the growing adoption of role-based access control (RBAC) in commercial security and identity management products, how to facilitate the process of migrating a non-RBAC system to an RBAC system has become a problem with significant business impact. Researchers have proposed to use data mining techniques to discover roles to complement the costly top-down approaches for RBAC system construction.

Kuhlmann *et al*. (2003) discover role-finding to implement Role-Based Security Administration. Results show stem from industrial projects, where large-scale customers wanted to migrate to Role-Based Access Control (RBAC) based on already existing access rights patterns in their production IT-systems. The core of this paper creates a link between the use of well established data mining technology and RBAC. They present a process for detecting patterns in a data base of access rights and for deriving enterprise roles from these patterns. Moreover, a tool (the SAM Role Miner) is described. The result allows an organized migration process to RBAC with the goal of building a single point of administration and control, using a cross-platform administration tool.

2.4 Business using Data Mining

Shi *et al*. (2001) emphasis that the Internet, ecommerce and e-business undoubtedly hold an important key to every organization's future. A web data mining provides solution to e-businesses to discover hidden patterns and business strategies from their customer and web data. A three-layer virtual e-business framework is as well as the web mining technique to personalize e-services, increase cross-selling, and improve the customer relationship management. They find out data mining can help e-business to improve their customer relationship, make intelligent business strategies, and sharpen competitive edge. Lin and Wu (2005) find out to inaugurate a new era of e-business development and face both opportunity and challenge of audit, it is necessary to introduce some new technology in traditional audit. Based on the analysis of the environment of e-business in China, the solutions and key technology.

Meanwhile, Apte *et al*. (2002) pointed out that the traditional approach to data analysis for decision making has been to couple business and scientific expertise with statistical modeling techniques in order to develop hand-crafted solutions for specific problems. They find out KDD techniques that emphasize scalable, reliable, fully-automated, and explanatory structures are demonstrating that such techniques can step up to the data analysis challenge.

In addition, Ettl *et al*. (2005), describes a software solution that combines business performance management with data mining techniques to provide a powerful combination of performance monitoring and proactive customer management in support of the new telesales business processes. They present a case study on the development and implementation of a data mining system in order to mine customer knowledge for electronic catalog marketing. Thus, it is begun to design and implement a data mining system using the relational database approach, including conceptual, logical, and physical database design as a data mining methodology. The data mining results and electronic catalog design from customer knowledge.

In addition, Liao and Chen (2004) find out catalogs for retailing firms are presented to customers in the format of paper catalogs without strategic segmentation design and implementation. In this regard, electronic catalog design and marketing could be a method to integrate the Internet and catalog marketing using market segmentation in order to enhance the effectiveness of direct marketing and sales management in retailing. Data mining has been used based on association rules from relational database design and implementation for mining customer knowledge. As a result, marketing knowledge patterns and rules are extracted for the electronic catalog marketing and sales management of a retailing mall in Taiwan. Besides, Chen *et al*. (2005) classifies the selected customers into clusters using RFM model to identify high-profit, gold customers. Subsequently, they are delineate data mining using association rules algorithm. Association rules algorithm been used to mine the data to detect the association rules in each time period.

In another study, Bloemer (2002) carried out the customer satisfaction be an important topic in the financial services industry. From a theoretical point of view, the problem of identifying latently dissatisfied customers. The descriptive data mining technique of characteristic rules helps to discover typical characteristics or properties of the target group. The results of this study indicate that characteristic rules provide an efficient and effective instrument to identify latently dissatisfied customers.

In database marketing, data mining has been used extensively to find the optimal customer targets so as to maximize return on investment. Lo (2002) uses marketing campaign data, models are typically developed to identify characteristics of customers who are most likely to respond. While these models are helpful in identifying the likely responders, they may be targeting customers who have decided to take the desirable action or not regardless of whether they receive the campaign contact. Ghani and Soares (2006), on the other hand, find out even though data mining has been successful in becoming a major component of various business processes as well as in transferring innovations from academic research into the business world, the gap between the problems that the research community works on and real-world ones is still significant.

Weblogs, or blogs, have rapidly gained in popularity over the past few years. In particular, the growth of business blogs written by or providing commentary on businesses and companies opens up new opportunities for developing blog-specific search and mining techniques. (Chen *et al*., 2007) propose probabilistic models for blog search and mining using two machine learning techniques, Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis (PLSA). They implement the models in their database of business blogs, with the aim of achieving higher precision and recall. The probabilistic model is able to segment the business blogs into separate topic areas, which is useful for keywords detection on the blogosphere. From the study, they can uncover domain-driven data mining techniques that can better strengthen business intelligence in complex enterprise applications. While (Cumby *et al*., 2004) find out their results show that can predict a shopper's shopping list with high levels of accuracy, precision, and recall. They believe that this impacts both the data mining and the retail business community. The formulation of shopping list prediction as a machine learning problem results in algorithms that should be useful beyond retail shopping list prediction.

2.5 Loan using Data Mining

Loan level modeling of prepayment is an important aspect of hedging, risk assessment, and retention efforts of the hundreds of companies in the US. Goodarzi *et al.* (1998), investigate different aspects of modeling customers who have taken jumbo loans in the US. They findings that one factor that was crucial in deriving insight quickly is having a data mining environment that supports modeling, drill-downs, drill-through, and integration of different tools and visualization.

Moreover, Agarwal *et al*. (2005) finds out loaners offer a menu of mortgage interest rate and point combinations in an effort to learn private information about borrowers' potential mobility. They use a unique dataset of individual automobile loan performance to assess whether borrower consumption choice reveals information about future loan performance. Results indicate that the automotive make and model a consumer selects provides information about the loan's performance. Besides, Mody and Patro (1996) show that guarantees are extremely valuable the value of a guarantee increases with the risk of the underlying asset or credit, the size of the investment, and the time to maturity. They describe methods of guarantee valuation, reports estimates of guarantee values in different settings, and summarize methods of guarantee accounting and their implications.

2.6 Banking using Data Mining

Dass (2007) finds out the huge size of data sources make it impossible for a human analyst to come up with interesting information or patterns that will help in decision making process. Application areas such as risk management, portfolio management, trading, customer profiling and customer care, can be used data mining techniques in banks and other financial institutions to enhance their business performance. Meanwhile Hunziker *et al.* (1999) summarizes experiences and results of productively using knowledge discovery and data mining technology in a large retail bank. They present data mining as part of a greater effort to develop and deploy an integrated IT-infrastructure for loyalty based customer management, combining data warehousing, and campaign management together with data mining technology. Data mining is not a stand-alone technology, but can be an important piece in many business processes. They focused on application of data mining for marketing campaigns, which is one of the most useful and promising applications.

In addition, Scott *et al*. (1999) recognize that, to effectively compete in increasingly competitive global markets, banks must better understand and profile their customers. Knowledge Discovery in Databases (KDD), often called data mining, is the inference of knowledge hidden within large collections of operational data. Their paper reports on experiences of applying the KDD process in a banking domain. A number of data mining techniques have been used, within the KDD process, and the results obtained have influenced the business activities of the banks. The procedures used are analyzed with respect to the domain knowledge they utilize, in order to evaluate the input from a domain expert during the KDD process. Business models should accurately represent the domain. The data mining results themselves are a source of domain knowledge and therefore can be used to validate and possibly enrich the business models. This cross validation highlights how the integration of KDD with business modeling will mutually benefit both areas.

2.7 Other Area using Data Mining

Kalos and Rey (2005) describe the experience of introducing data mining to a large chemical manufacturing company. The multinational nature of doing business with multiple business units, presents a unique opportunity for the deployment of data mining. While each business unit has its own objectives and challenges, which may be at odds with those of other units, they also share many common interests and resources. In this environment, data mining can be used to identify potential value-creating opportunities, through large site integration of multiple assets and synergies from the use of common assets, such as site-wide manufacturing facilities, and world-wide supply-chain, purchasing and other shared services. In the addition, they describe about the approach for launching a data mining capability within this framework, the strategy for securing upper management support, drawing from internal modeling, statistical, and other communities, and from external consultants and universities.

Hueglin and Vannotti (2001) find out that predictive models developed by applying data mining techniques are used to improve forecasting accuracy in the airline business. In order to maximize the revenue on a flight, the number of seats available for sale is typically higher than the physical seat capacity as overbooking. To optimize the overbooking rate, an accurate estimation of the number of no-show passengers means passengers who hold a valid booking but do not appear at the gate to board for the flight is essential. The forecasting approach is more accurate than the currently used method. In addition, the selected models lead to a deepened insight into passenger behavior.

Cavaretta (2006) finds out that automotive companies, such as Ford Motor Company, have no shortage of large databases with abundant opportunities for cost reduction and revenue enhancement. The Data Mining Group at Ford has worked in the areas of Quality, Customer Satisfaction and Warranty Analytics for close to ten years. At the time, they have developed a number of methods for building systems to help the business. One area of particular success has been in warranty analysis.

2.8 Conclusion

From the literature review, it is concluded that data mining approach has been used in various areas, especially the business area including loan, e-commerce, accounting, banking, financial and others. By the way, there are many researchers studying in prediction bankruptcy. They are using several of approach and technique such as Cash Flow, Rule-based learning, neural network, support vector machines and hybrid approach. However, hybrid technique and neural network had shown the potential in improving classification results. Hence, this study only focuses on data mining approach and neural network as bankruptcy model.

CHAPTER 3

METHODOLOGY

This study is carried out according to Cross Industry Standard Process (CRISP) methodology (Chapman *et al.,* 2000). The methodology is described in terms of a hierarchical process model, consisting of sets of tasks express at four levels of abstraction from general to specific for example phase, generic task, specialized task, and process instance. At the top level, the data mining process is organized into a number of phases; each phase consists of several second-level generic tasks which are general enough to cover all possible data mining situations; the tasks are also intended to be as complete and stable as possible. Complete means covering both the whole process of data mining and all possible data mining applications. Stable means that the model should be valid for yet unforeseen developments like new modeling techniques.

The specialized task level describes how actions in the generic tasks should be carried out in certain specific situations. For example, the generic task might have clean data; this level describes how this task differs in different situations, such as cleaning numeric values versus cleaning categorical values, or whether the problem type is clustering or predictive modeling. The description of phases and tasks as discrete steps performed in a specific order represents an idealized sequence of events. Many of the tasks can be performed in different order, and it often necessary to be repeatedly backtrack to previous tasks and repeat certain actions.

The process instance level is a record of the actions, decisions, and results of an actual data mining engagement. Process instance is organized according to the tasks defined at the higher levels, but represents what actually happened in a particular engagement, rather than what happens in general.

Fig. 3.1 shows the life cycle of data mining project using CRISP. It contains of project respective tasks and the relationships between these tasks.

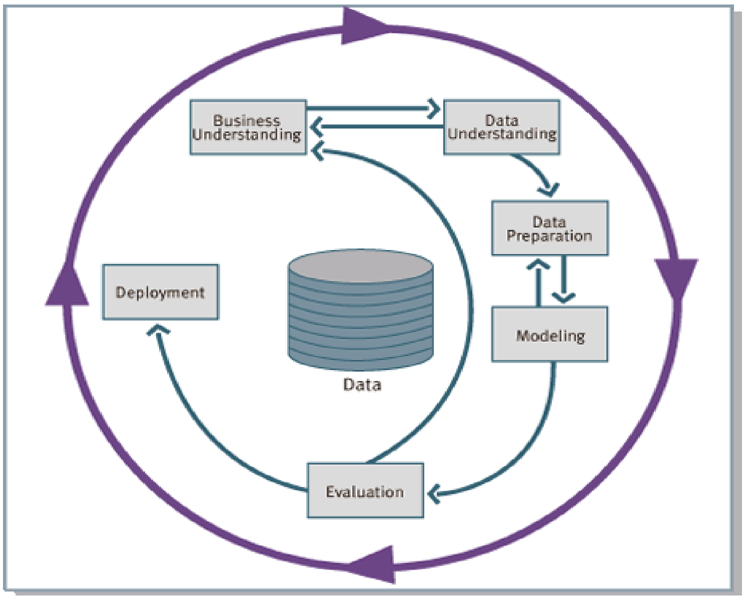


Figure 3.1: Phases of CRISP-DM Reference Model

As shown in Fig. 3.1, there are six phases in this method, i.e. business understanding, data understanding, data preparation, modeling, evaluation and deployment. The sequence of the phases, however, is not rigid. The outcome of each phase determines which phase has to be executed next. The arrows indicate the dependencies between phases. The following sections describe each of the phases in detail.

3.1 Business understanding

This first phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives. In the business understanding phase, the scope and objective of the project are determined.

3.2 Data understanding

The data understanding phase begins with initial data collection Data collected from the respondents need to be checked and understood at data understanding phase. Afterwards, it proceeds with activities to familiar with the data, identify data quality problems, discover first insights into the data, and detect interesting subsets to form hypotheses regarding hidden information.

3.3 Data preparation

The data preparation phase covers all activities needed to construct the final dataset where data that will be fed into the modeling tools from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools. Data preparation includes data collection, data cleansing, data selection and data preprocessing, as described in the following sections.

3.3.1 Data Preparation

Data Preparation involves checking or logging the data in, checking the data for accuracy, entering the data into the computer, transforming the data, and developing and documenting a database structure that integrates the various measures. Data preparation roughly divided into data collection, data cleansing, data selection and data preprocessing.

3.3.2 Data Collection

Two types of data are used in this study, i.e. data from internal financial indicators and data based on macro-economic features. A total of 367 data was attained from the Registrar of Business and Companies, Kuala Lumpur Stock Exchange (KLSE) and Bank Negara Malaysia (Central Bank of Malaysia). The data was from 150 data of bankrupt companies and 200 from healthy companies. The data covers a wide range of financial ratios and macro-economic data. Table 3.1 shows the input data parameters and the type of each. Table 3.1 shows the input data parameters.

Table 3.1: Input Features

Parameter	Data Type
Working Capital/ Total Assets	Double
Retained Earnings/ Total Assets	Double
Earnings Before Income Tax/ Total Assets	Double

Total Sales/ Total Assets	Double
Total Debts/ Total Assets	Double
Type of Industries	Integer
Gross Domestic Product	Double
Age	Integer
Size	Integer
Bank Rate	Double
Inflation Rate	Double
Target	Double

3.3.3 Data cleansing

Data cleansing is the act of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. Used mainly in databases, the term refers to identifying incomplete, incorrect, inaccurate, irrelevant etc. parts of the data and then replacing, modifying or deleting this dirty data.

After cleansing, a data set will be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by different data dictionary definitions of similar entities in different stores, may have been caused by user entry errors, or may have been corrupted in transmission or storage. In this study, the cleansing was done manually by vetting throughout all the data, whereby any detected incorrect or missing value were corrected accordingly. The denominator of some ratio is zero because some of final report has value of either zero or “NA”. Therefore, some ratios are finite. The parameter which is with finite ratios is infeasible and deleted. The occurrence of non-feasible observation is relatively infrequent.

3.3.4 Data selection

Data selection is defined as the process of determining the appropriate data type and source, as well as suitable instruments to collect data. Data selection goes before the actual practice of data collection. This definition make a distinction that data selection from selective data reporting (selectively excluding data that is not supportive of a research hypothesis) and interactive/active data selection (using collected data for monitoring activities/events, or conducting secondary data analyses). The process of selecting suitable data for a research project can impact data integrity. The most important thing is to select the right variables that have strong effects on the output results. Table 3.2 shows the ratio of data and Table 3.3 illustrates the relationship between internal and external variables with risk of bankruptcy.

Table 3.2: Ratios with the data sign

No.	Ratio	Data Sign
1	Working Capital / Total Assets	
2	Retained Earning / Total Assets	
3	Earning Before Income Tax / Total Assets	Ratio less then 1
4	Total Sales / Total Assets	Ratio less then 1
5	Total Debts / Total Assets	Ratio greater then 1

Table 3.3: Relationship between internal and external variables with risk of bankruptcy

No	Type of Internal Variables	
1	Working Capital/ Total Assets	
2	Retained Earnings/ Total Assets	
3	Earnings Before Income Tax/ Total Assets	

4	Total Sales/ Total Assets	
5	Total Debts/ Total Assets	
	Type of External Variables	Risk
6	Type of Industries 1.Manufacturing 2.Utility 3.Mining 4.Construction 5.Finance 6.Retailing 7.Services 8.Food and Beverages	 Highest Highest Highest Highest High High Medium Low
7	Gross Domestic Product 1. less than 4 % 2. more than 4% and less than 6% 3. more than 6%	 High Medium Low
8	Age 1. less than 1 year 2. more than 1 year and less than 5 year 3. more than 5 year and less than 10 year 4. more than 10 year	 Highest High Medium Low
9	Size 1. less than 100 staff 2. more than 100 staff and less than 500 staff 3. more than 500 staff	 High Medium Low
10	Bank Rate 1. more than 9% 2. more than 7% and less than 9% 3. more than 6% and less than 7% 4. less than 6%	 Highest High Medium Low
11	Inflation Rate 1. more than 10% 2. more than 6% and less than 10% 3. more than 2% and less than 6% 4. less than 2%	 Highest High Medium Low
12	Target	

3.3.5 Data Preprocessing

Every single data is normalized in order to distributed the data evenly and scale it into acceptance range. There are three data preprocessing techniques:

- Binarization

To transform symbolic and numeric value into binary

- Symbolization

  To transform symbolic value into integer representation

- Normalization

  A continuous scaling technique to convert continuous numeric value into a certain range (between 0 and 1)

Binarization is a technique to represent nominal or categorical data into binary value. Table 3.4 shows the binarization output for each particular ratio.

Table 3.4: Binary and Symbolic Representation

Type of Internal Variables	Binary	
Working Capital/ Total Assets	1	
Retained Earnings/ Total Assets	1	
Earnings Before Income Tax/ Total Assets	1	
Total Sales/ Total Assets	1	
Total Debts/ Total Assets	1	
Type of External Variables	Symbolic	Risk
Type of Industries		
1.Manufacturing	4	Highest
2.Utility	4	Highest
3.Mining	4	Highest
4.Construction	4	Highest
5.Finance	3	High
6.Retailing	3	High
7.Services	2	Medium
8.Food and Beverages	1	Low
Gross Domestic Product		
1. less than 4 %	3	High
2. more than 4% and less than 6%	2	Medium
3. more than 6%	1	Low
Age		
1. less than 1 year	4	Highest
2. more than 1 year and less than 5 year	3	High
3. more than 5 year and less than 10 year	2	Medium
4. more than 10 year	1	Low
Size		
1. less than 100 staff	3	High
2. more than 100 staff and less than 500 staff	2	Medium

3. more than 500 staff	1	Low
Bank Rate 1. more than 9% 2. more than 7% and less than 9% 3. more than 6% and less than 7% 4. less than 6%	 4 3 2 1	 Highest High Medium Low
Inflation Rate 1. more than 10% 2. more than 6% and less than 10% 3. more than 2% and less than 6% 4. less than 2%	 4 3 2 1	 Highest High Medium Low
Target		

3.4 Modeling

In this phase, various modeling techniques are selected and applied, and the parameters are standardizing to optimal values. Usually, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, going back to the data preparation phase is often necessary. At this point, modeling technique will be selected and applied to the problem domain. Here, neural network technique from data mining approach has been selected and applied to the problem domain. Neural networks have seen an explosion of interest over the last few years, and are being successfully applied across an extraordinary range of problem domains, in areas as diverse as finance, medicine, engineering, geology and physics. Indeed, anywhere that there are problems of prediction, classification or control, neural networks are being introduced and used.

Neural Networks or Multi-layer Perceptron (MLP) have architecture that allows multiple coefficients per input variable. The processing at each node is functionally equivalent to logistic regression. Multiple nodes that organized into layers (input layer, hidden layer and output layer) allows the model to represent complex, nonlinear, interactions between variables. MLP neural networks have been described as “universal approximators” because they are capable of accurately approximating any functional relationship. Fig. 2 shows relationship of input layer, hidden layer and output layer.

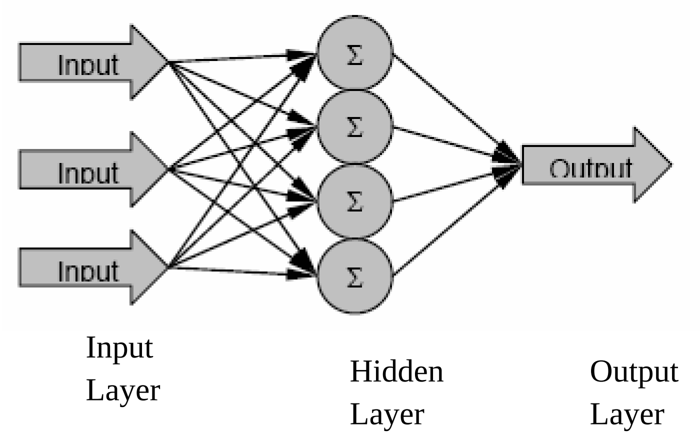


Figure 3.2: Node and Layers of Neural Network

3.5 Evaluation

Before proceeds to final deployment phase of the model, it is important to systematically evaluate it and review the steps executed to create the model. This intends to be convinced the model properly achieves the project’s objectives. At the end of this phase, a decision on the use of the data mining results should be reached. Afterward the modeling technique will be evaluated.

3.6 Deployment

The best model will be selected and use to develop a modeling. Establishment of the model is in general not the end of the project. Yet if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented. To tool that been used is Neural Connection to have the accuracy of the prediction. While doing the correlation, SPSS had been used to have the statistical value of the prediction from the dataset. By the way, the deployment phase is depending on the requirements and to identify with actions that need to be performs with the purpose of actually use of the created models.

3.7 Conclusion

Cross Industry Standard Process (CRISP) methodology is a methodology that suitable for data mining approach. The methodology is described in terms of a hierarchical process model, consisting of sets of tasks express cover all possible data mining situations. The tasks are proposed to complete both the whole process of data mining and all possible data mining applications.

CHAPTER 4

RESULT

In this chapter, conveys about the results and finding of this study. The results of using neural network and descriptive statistics including cross tabulation, scatter plot and correlation as well as logistic regression are described in the following subsections.

4.1 Attribute of Bankruptcy Model

Ratios can help predict bankruptcy before it's too late for a business to take corrective action and for creditors to reduce potential losses. With careful planning, predicted futures can be avoided before they become reality. The first five bankruptcy ratios in this section can detect potential financial problems up to three years prior to bankruptcy. The sixth ratio, Cash Flow to Debt, is known as the best single predictor of failure.

This data set can be used to predict the bankruptcy of the company. The attributes that used in this study include working capital, retained earning, earning before income tax, total sales, total debts, type of industries, gross domestic product, age, size, bank rate, inflation rate, target.

Working capital is a financial metric that signify operating liquidity of the company. It is calculated as current assets minus the current liabilities. When currents assets less then current liabilities, means the entity has a working capital deficiency. The positive working capital needed to ensure the company is able to continue to operate. Working capital is the most valuable sign of an alarming business failure.

Retained earnings are net income which is retained by the corporation before distributed to its owners as dividends. Likewise, if the company makes a loss, then that loss is retained and identified as variously retained losses. Earning before income tax is a company's profitability that excludes income tax expenses.

Ratio of total debt that indicates the percentage of a company's assets is provided via debt. Companies with high debt ratios are not actually bankruptcy candidates if they willing to produce sufficient income to cover their interest payments and other normal expenses.

While, ratio of total sales that show the percentage of a company's assets is provided via sales. The companies that have high

The data set comprises of 367 instances with 12 attributes, including the target attribute. The detail description about each attribute is shown in Table 4.1.

Table 4.1: Type of Attributes

No.	Attribute	Type	Representation
1	Working Capital	Double	
2	Retained Earning	Double	
3	Earning Before Income Tax	Double	
4	Total Sales	Double	
5	Total Debts	Double	
6	Type Of Industries	Integer	1.Manufacturing 2.Utility 3.Mining 4.Construction

			5.Finance 6.Retailing 7.Services 8.Food and Beverages
7	Gross Domestic Product	Double	1.Less than 4% 2.More than 4% and less than 6% 3.More than 6%
8	Age	Integer	1.Less than 1 year 2.More than 1 year and less than 5 year 3.More than 5 year and less than 10 year 4.More than 10 year
9	Size	Integer	1.Less than 100 staff 2.More than 100 staff and less than 500 staff 3.More than 500 staff
10	Bank Rate (%)	Double	1.More than 9% 2.More than 9% year and less than 7% 3.More than 7% and less than 6% 4.Less than 6%
11	Inflation Rate (%)	Double	1.More than 10% 2.More than 6% and less than 10% 3.More than 2% and less than 6% 4.Less than 2%
12	Target	Double	

Attributes of Working Capital, Retained Earning, Earning Before Income Tax, Total Sales and Total Debts are the financial ratio which is divided to Total Assets. The detail description about each attribute is shown in Table 4.2.

Table 4.2: Ratio of Attributes

No.	Attribute	Ratio	Data Sign	Binary
1	Working Capital	Working Capital / Total Assets		1
2	Retained Earning	Retained Earning / Total Assets		1
3	Earning Before Income Tax	Earning Before Income Tax / Total	Ratio less than 1	1

		Assets		
4	Total Sales	Total Sales / Total Assets	Ratio less than 1	1
5	Total Debts	Total Debts / Total Assets	Ratio greater than 1	1

Missing Attribute Values: No

Table 4.3: The missing attributes

Attribute	Missing Value
Working Capital	No
Retained Earning	No
Earning Before Income Tax	No
Total Sales	No
Total Debts	No
Type Of Industries	No
Gross Domestic Product	No
Age	No
Size	No
Bank Rate	No
Inflation Rate	No
Target	No

The number of missing values is computed using SPSS and the results are exhibited in Table 4.3 and Table 4.4which represent that no missing value for each attribute.

Table 4.4: Number of Attributes

	Working Capital	Retained Earning	Earning Before Income Tax	Total Sales	Total Debts	Type Of Industries	Gross Domestic Product	Age	Size	Bank Rate	Inflation Rate	Target
N Valid	367	367	367	367	367	367	367	367	367	367	367	367
Missing	0	0	0	0	0	0	0	0	0	0	0	0

4.1.1 Frequency of Attribute

The frequency of attributes is shown as below:

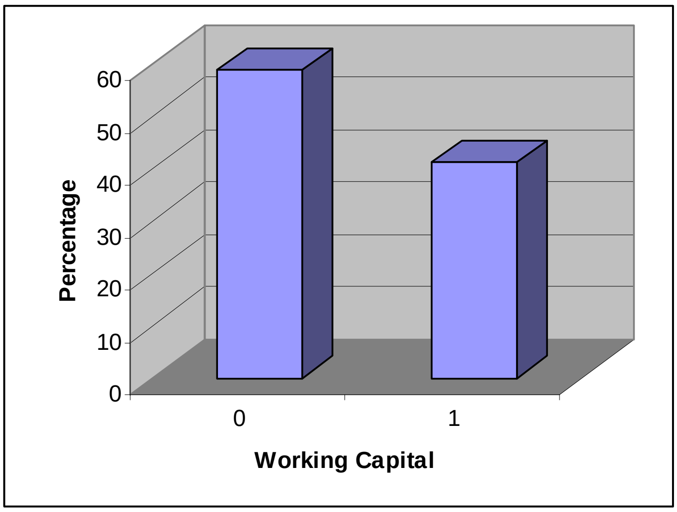


Figure 4.1: Working Capital

The values of working capital are represented by number 0 to 1 from the ratio working capital divide total assets. However, fig. 4.1 shows that most of the company obtained value of 0 rather than value of 1.

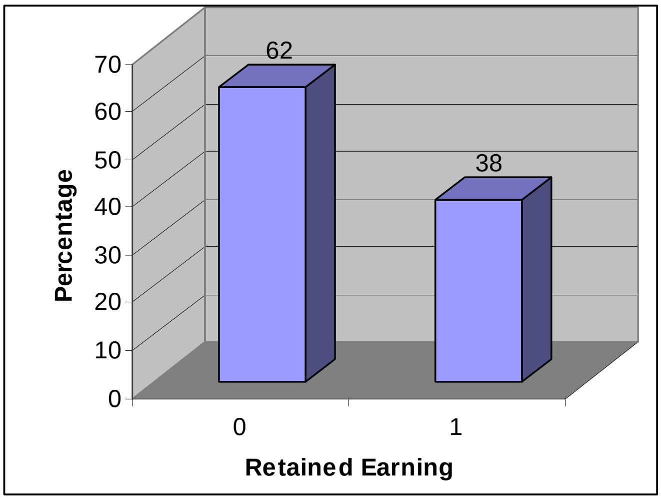


Figure 4.2: Retained Earning

The values of Retained Earning are represented by number 0 to 1 where ratio calculate from retained earning divide total assets.. However, fig. 4.2 shows that most of the company obtained value of 0 therefore 62% and 38% by value of 1

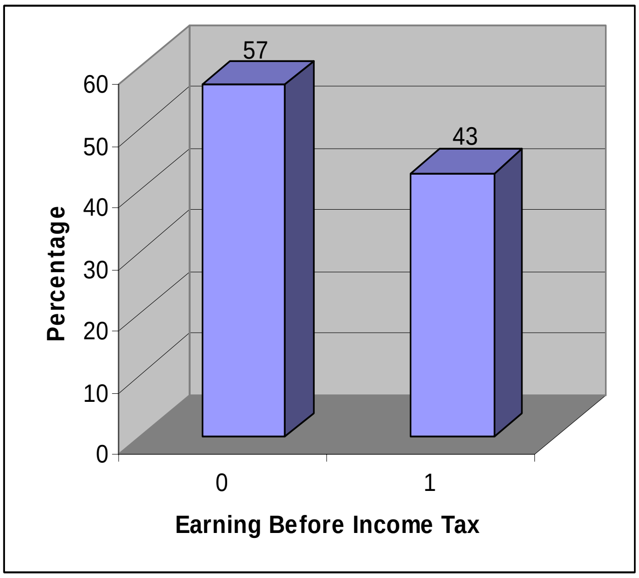


Figure 4.3: Earning before Income Tax

The values of earning before income tax calculated by divide earning before income tax with total assets therefore the value are represented by 0 to 1. The fig. 4.3 shows that most of the company obtained value of 0 rather than value of 1 with only 43%.

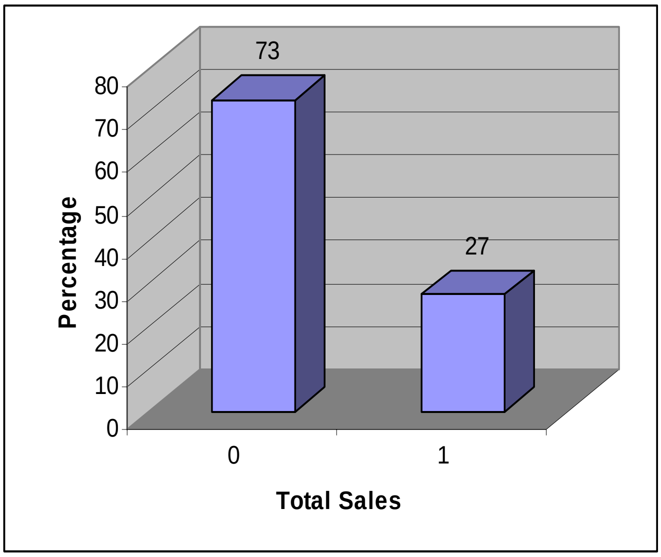


Figure 4.4: Total Sales

The ratio of total sales calculated by total sales divided total assets therefore the values are within the range of 0 and 1. The results also indicate that more than 73% of the companies have total sales less than total assets. The results refer on fig. 4.4.

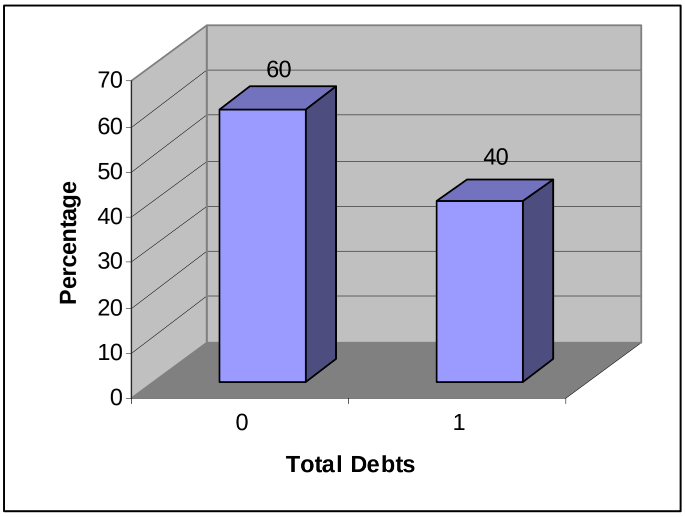


Figure 4.5: Total Debts

Fig. 4.5 shows total debt's ratio which is total debts divide by total assets, the values be supposed in the range of 0 to 1. Whereby, nearly 60 percent of the companies have total debts less than total assets.

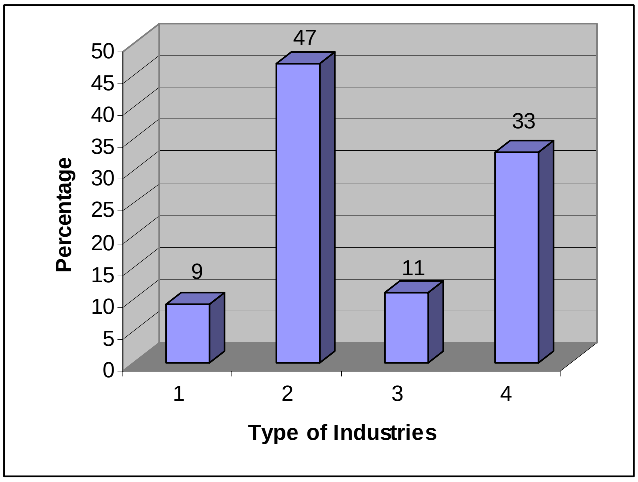


Figure 4.6: Type of Industries

Fig. 4.6 shows that the highest Type of Industries is in Utility Industry therefore 47%. The second is Construction followed by Mining and Manufacturing

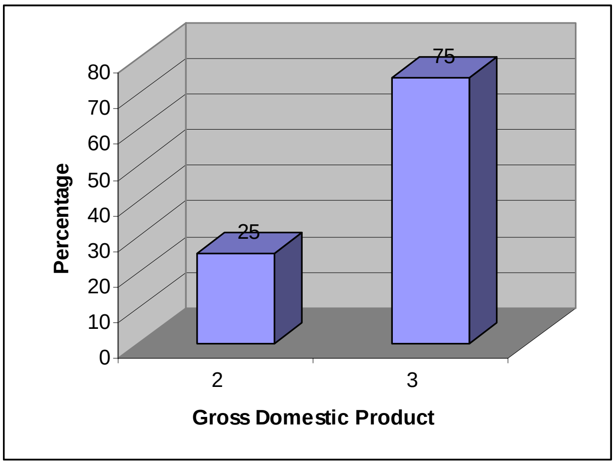


Figure 4.7: Gross Domestic Product

Results on fig. 4.7 shown that 75% of Gross Domestic Product stated as 3 which is in the range of more than 6% and 25% of Gross Domestic Product stated as 2 which is in the range of more than 4% and less than 6%

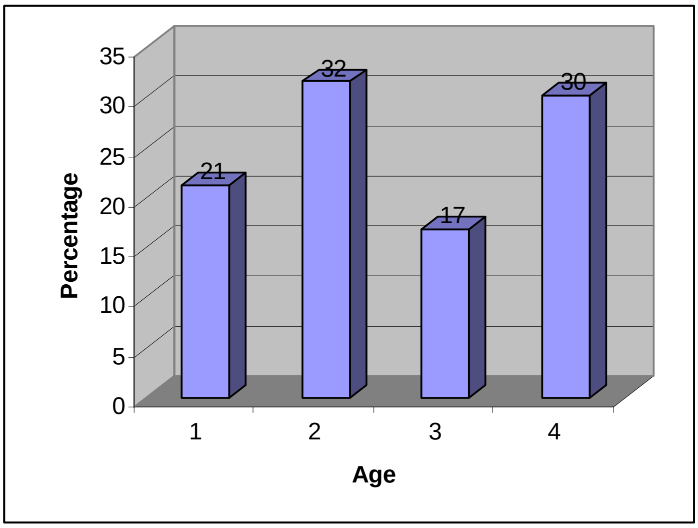


Figure 4.8: Age

Based on the fig.4.8, the highest age of the company which is 32% are in the range of more than 1 year and less than 5 year which is denoted as 2. Followed by 30% more than 10 year (denoted by 4), 21% less than 1 year (denoted by 1) and 17%more than 5 year and less than 10 year (denoted by 3).

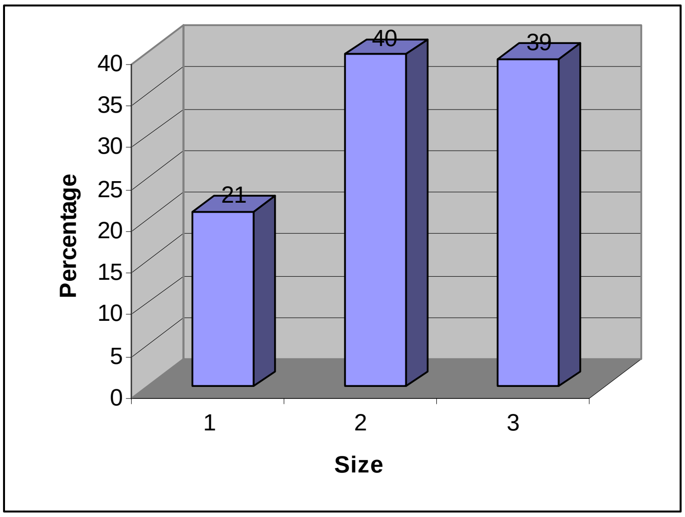


Figure 4.9: Size

Results illustrate in fig.4.9 that the highest range of size of the company which is 40% are denoted by 2 which is in the range of more than 100 staffs and less than 500 staffs and followed by more than 500 staff (denoted as 3) and less than 100 staff (denoted as 1)

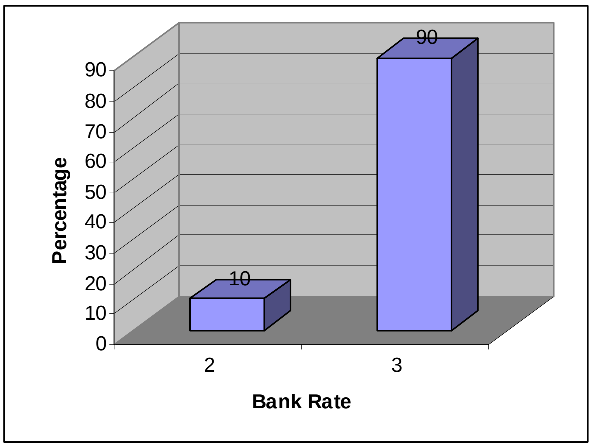


Figure 4.10: Bank Rate

Results on fig. 4.10 shown that the highest range of bank rate which is 90% are represent by 3 whereby in the range of more than 6% and less than 7%. Followed by 10% of bank rate within more than 7% and less than 9%.

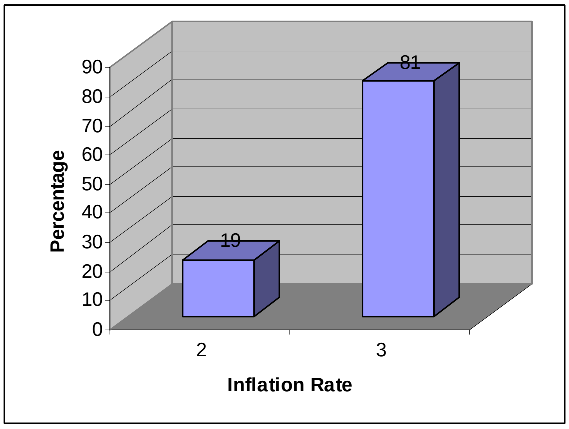


Figure 4.11: Inflation Rate

Results shown at fig. 4.11, the highest range of inflation rate which is 81% denoted by 3 where are in the range of more than 2% and less than 6%. Followed by inflation rate that more than 6% and less than 10%.

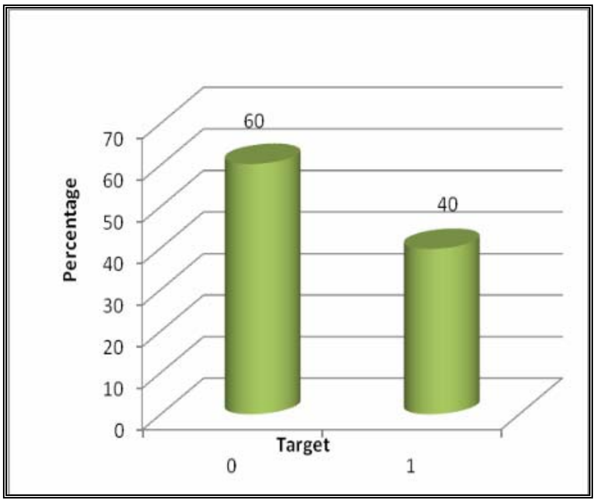


Figure 4.12: Target

Fig. 4.12 shows The percentage of target of Class 0 and 1 is 60-40. Therefore, it is anticipated that the model build based on these values of targets would not be biased.

Table 4.5: Target

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	221	60.2	60.2	60.2
	1	146	39.8	39.8	100.0
	Total	367	100.0	100.0	

Table 4.5 shows the frequency of target from SPSS. Where 60% of target 0 signify as Failure Company and 40% of target one for healthy company.

4.2 Descriptive Data Mining Approach

Descriptive data mining approach describes the concepts or task relevant dataset in concise, summarize, informative and discriminate form.

4.2.1 Cross Tabulation

Cross tabulation is mapping results into cross tabulation form. Cross Tabulation between independent variables (Working Capital, Retained Earning, Earning Before Income Tax, Total Sales, Total Debts, Gross Domestic Product, Age, Size, Bank Rate, Inflation Rate and Type Of Industries) and dependent variable (Target) is carried out and the results are presented in the following sections.

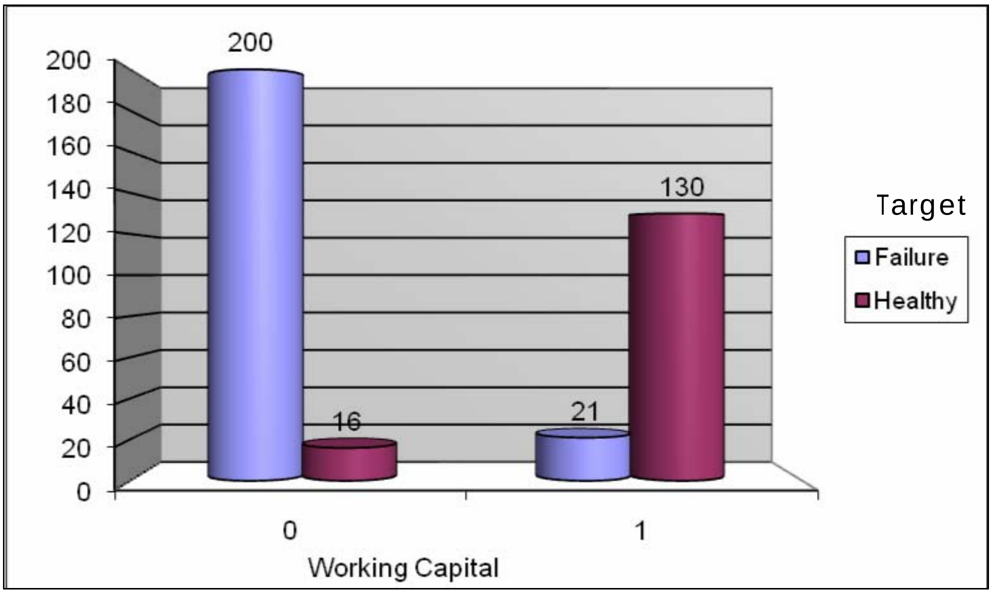


Figure 4.13: Cross Tabulation of Working Capital

Working capital is a ratio of working capital divided by total assets. Once the value of working capital is 0, meaning that the liquidity of the company are low rather than value 1. As a comparison with the target, the working capital which attains 0 which is an unhealthy company that nearly to bankrupt. For the healthy company mostly get hold of working capital value of 1 that have better liquidity. Results show as fig. 4.13.

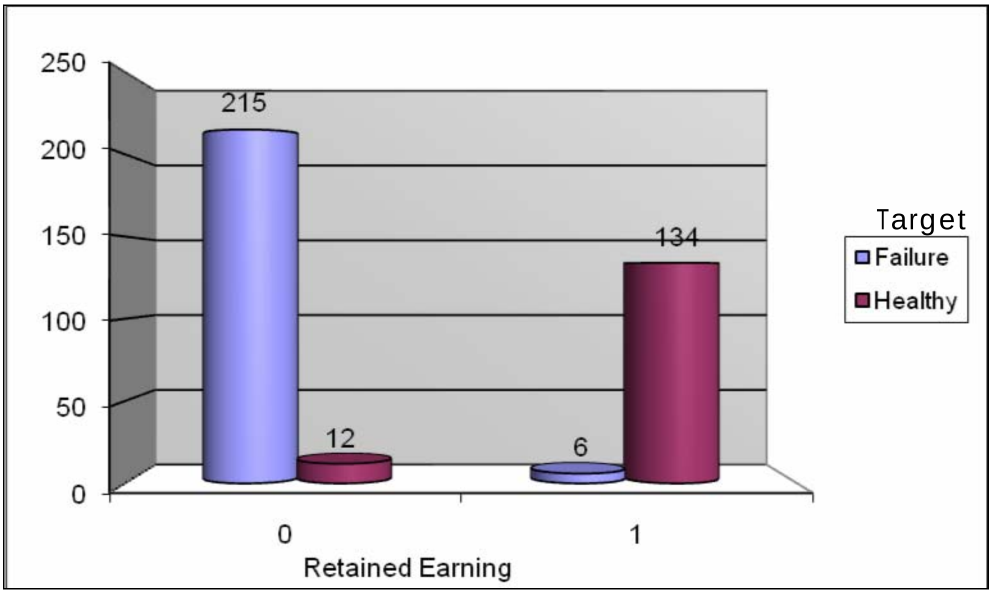


Figure 4.14 : Cross Tabulation of Retained Earning

Retained Earning is a ratio of Retained Earning divided by total assets. From the graph shown that, the Retained Earning which obtains value 0 mostly is an unhealthy company that nearly to bankrupt. For the healthy company mostly obtain Retained Earning value of 1 rather than value 0. Results show as fig. 4.14.

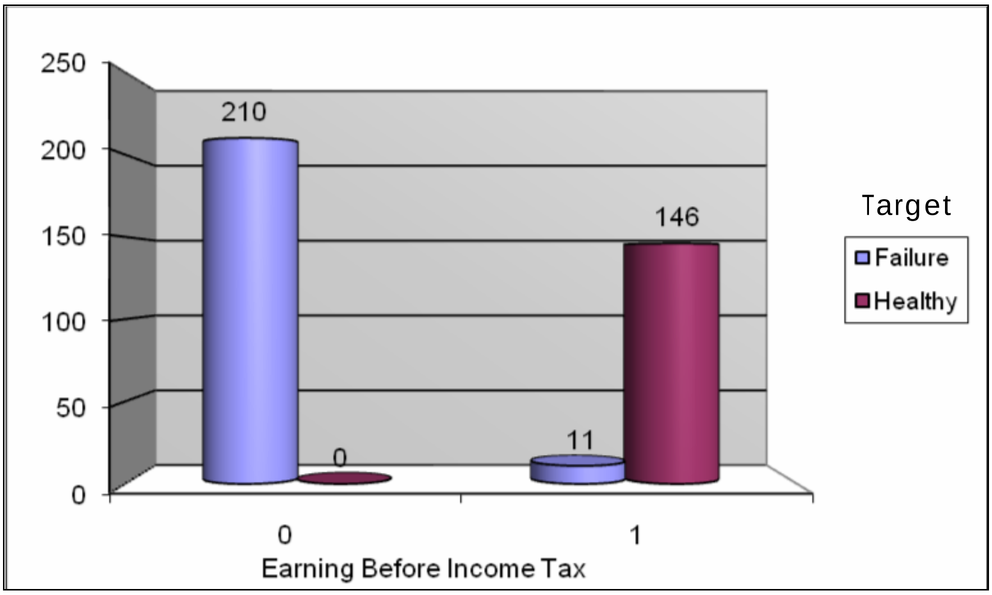


Figure 4.15 : Cross Tabulation of Earning before Income Tax

While, Earning Before Income Tax is a ratio of Earning Before Income Tax divided by total assets. The Earning Before Income Tax which carry out value 0 generally is an unhealthy company. There is no healthy company obtain Earning Before Income Tax values of 0. All of the healthy company obtain Earning Before Income Tax value of 1 as illustrated at fig. 4.15.

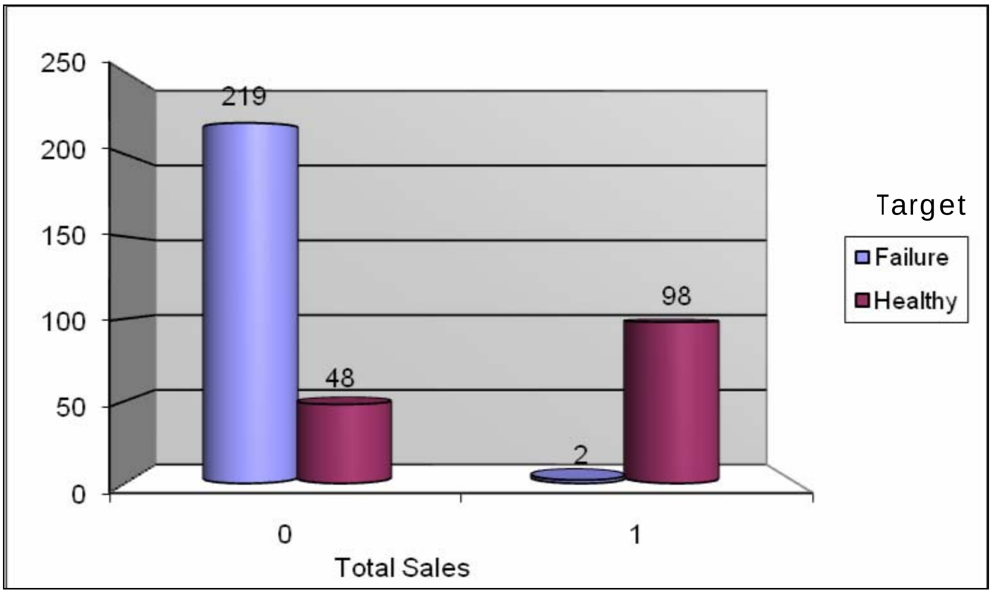


Figure 4.16 : Cross Tabulation of Total Sales

Whilst, Total Sales is a ratio of Total Sales divided by total assets. Fig. 4.16 clarify unhealthy company attains Total Sales with value of 0 rather than value of 1. There are only 2 company which is obtaining the ratio with value of 0 are considering as healthy company. Consequently, the healthy company mostly obtains Total Sales value of 1 meaning that have Total Sales more than total assets.

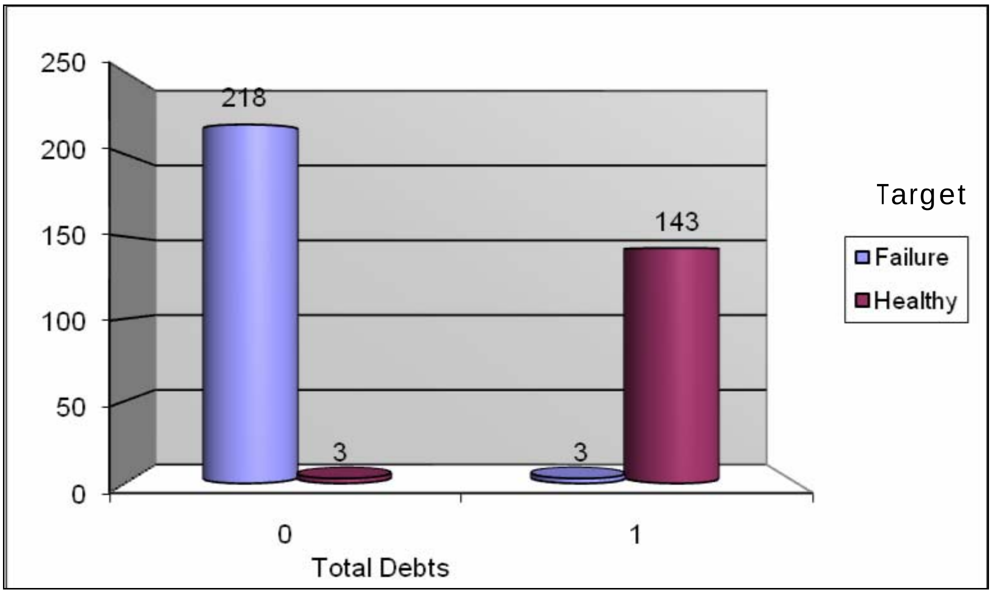


Figure 4.17 : Cross Tabulation of Total Debts

At the same time, Total Debt is a ratio of Total Debt divided by total assets. Frequently, the Total Debt which attains value of 0 is an unhealthy company. For the healthy company mostly obtain Total Debt value of 1 as fig. 4.17.

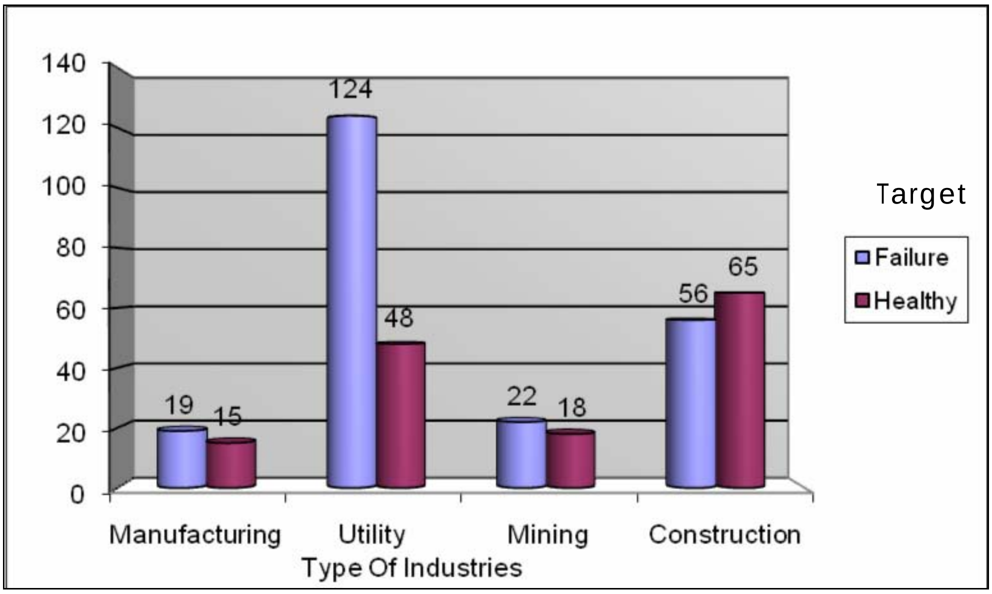


Figure 4.18 : Cross Tabulation of Type of Industries

As shown by fig. 4.18, Utility Industry is a highest Type of Industry that considers as unhealthy company and followed by Construction Industry, Mining Industry and Manufacturing Industry. For the healthy company initiate from Construction Industry than followed by Utility Industry, Mining Industry and Manufacturing Industry.

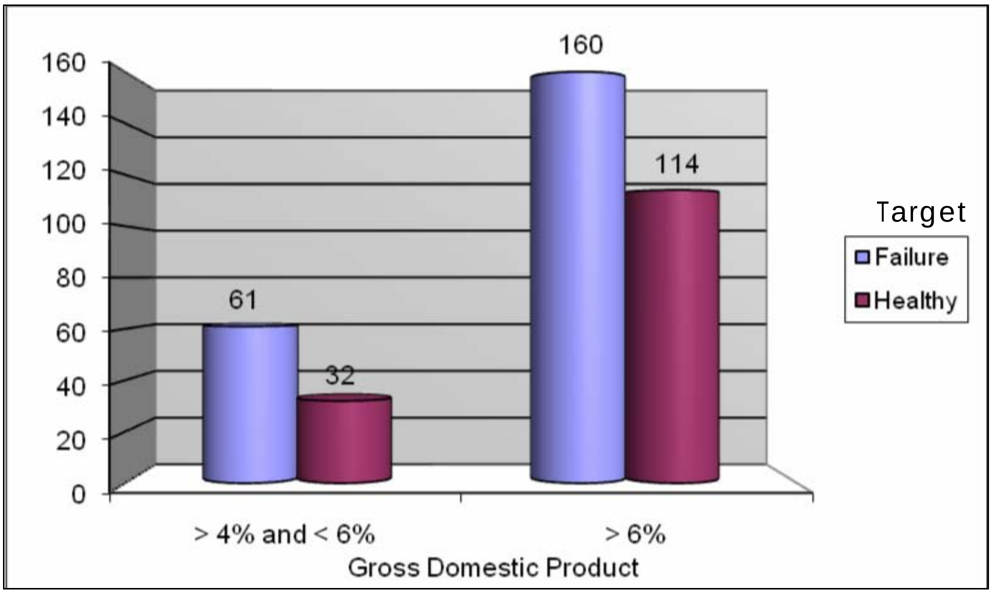


Figure 4.19 : Cross Tabulation of Gross Domestic Product

Fig.4.19 show healthy company and unhealthy company, highest Gross Domestic Product are in the range of more than 6% and followed by the range of more than 4% and less than 6%.

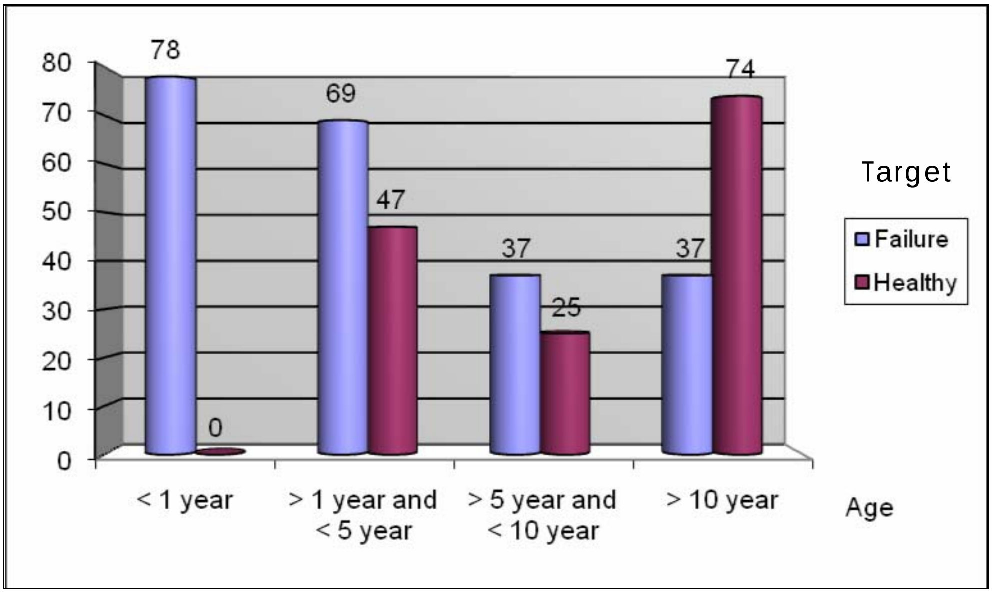


Figure 4.20 : Cross Tabulation of Age

The highest Age of unhealthy company in range of less than 1 year subsequently more than 1 year and less than 5 year and than more than 5 year and less than 10 year and lastly more than 10 year. That’s mean, the less experience company are higher risk to bankrupt compare to veteran company. The healthiest company typically growth more than 10 year as show as fig. 4.20.

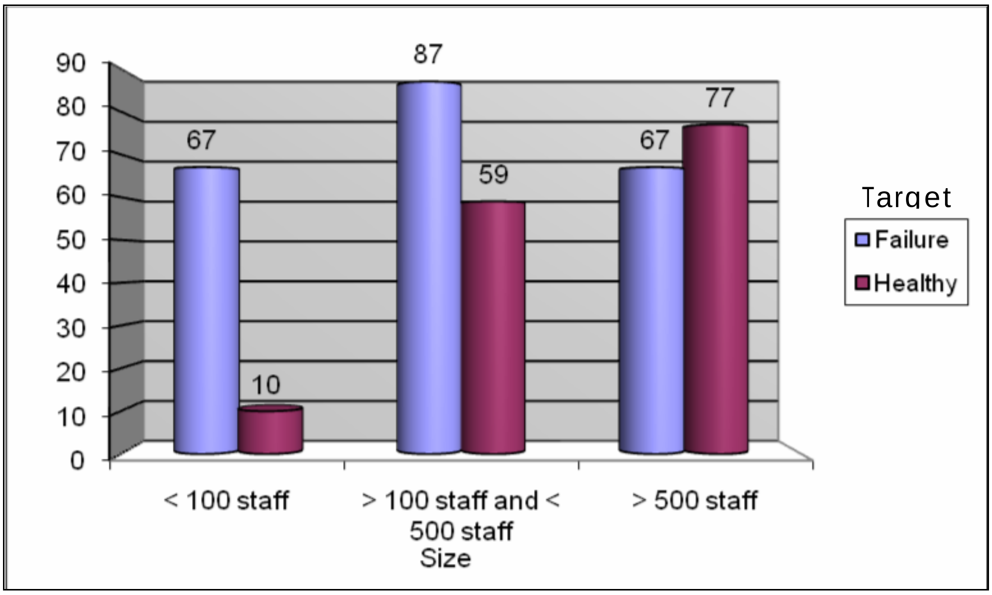


Figure 4.21 : Cross Tabulation of Size

Whilst, unhealthy company in Size of range more than 100 staff and less than 500 staff and subsequently for the size of less than 100 staff and the size of more than 500 staff . Incidentally, the healthy company mostly obtains large number of staff rather than less number of staff as shows fig 4.21.

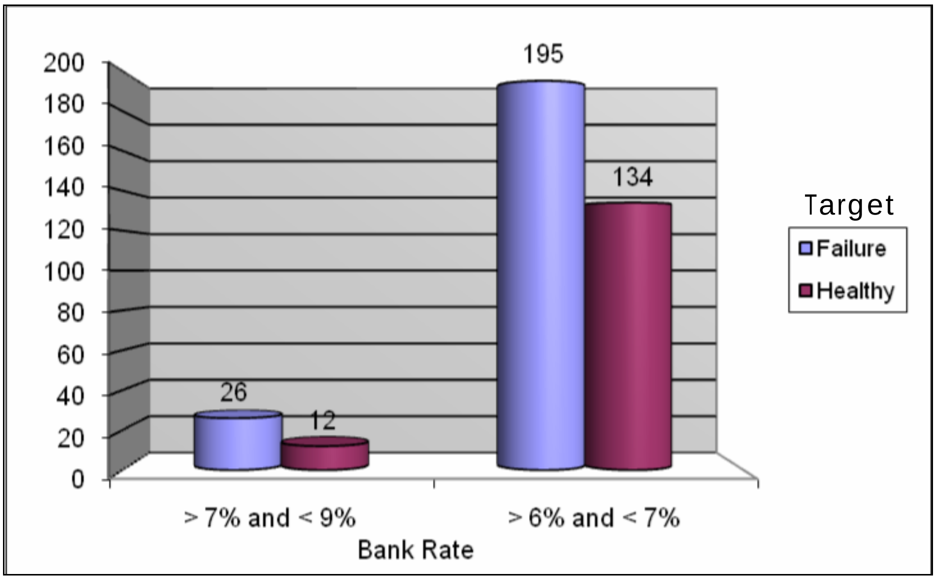


Figure 4.22 : Cross Tabulation of Bank Rate

Similarly for the unhealthy company and healthy company which acquire Bank Rate from the range of more than 6% and less than 7% followed by more than 7% and less than 9%. However the majority of unhealthy company acquires Bank Rate from the range of more than 6% and less than 7% showed as fig. 4.22.

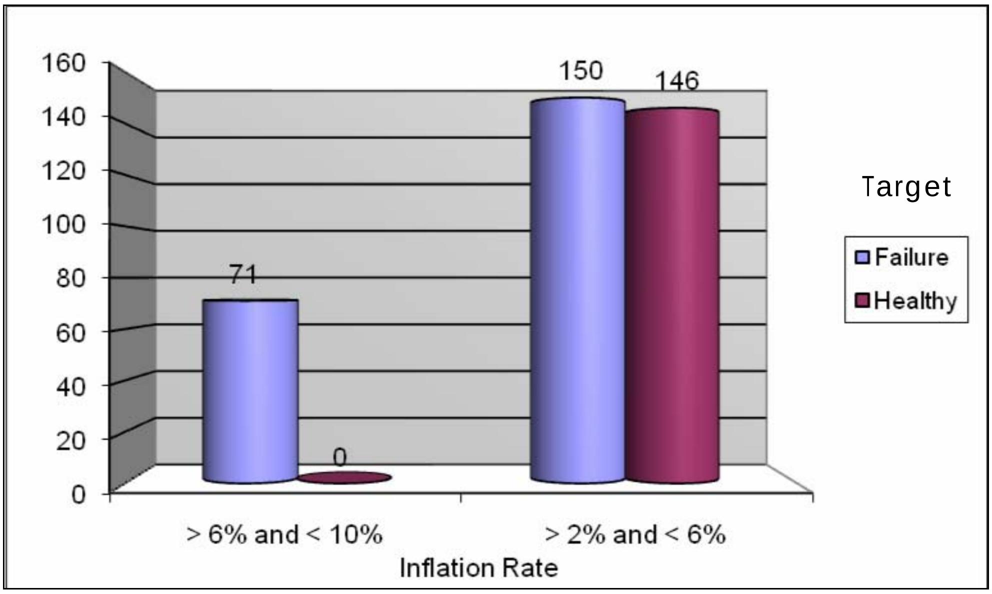


Figure 4.23 : Cross Tabulation of Inflation Rate

At the same time, the highest Inflation Rate of unhealthy company is in range of more than 2% and less than 6% and followed by more than 6% and less than 10%. For the healthy company only obtains Inflation Rate in range of 2% and less than 6% as fig. 4.23.

4.2.2 Scatter Plot

Scatter plot indicates the relationship between the dependent variable (Target) and the independent variables (Working Capital, Retained Earning, Earning Before Income Tax, Total Sales, Total Debts, Type Of Industries, Gross Domestic Product, Age, Size, Bank Rate and Inflation Rate).

Clearly, the relationship between the dependent and independent variables is nonlinear (see red oval line) since the dependent variable is of categorical value shown in Fig. 4.24.

.

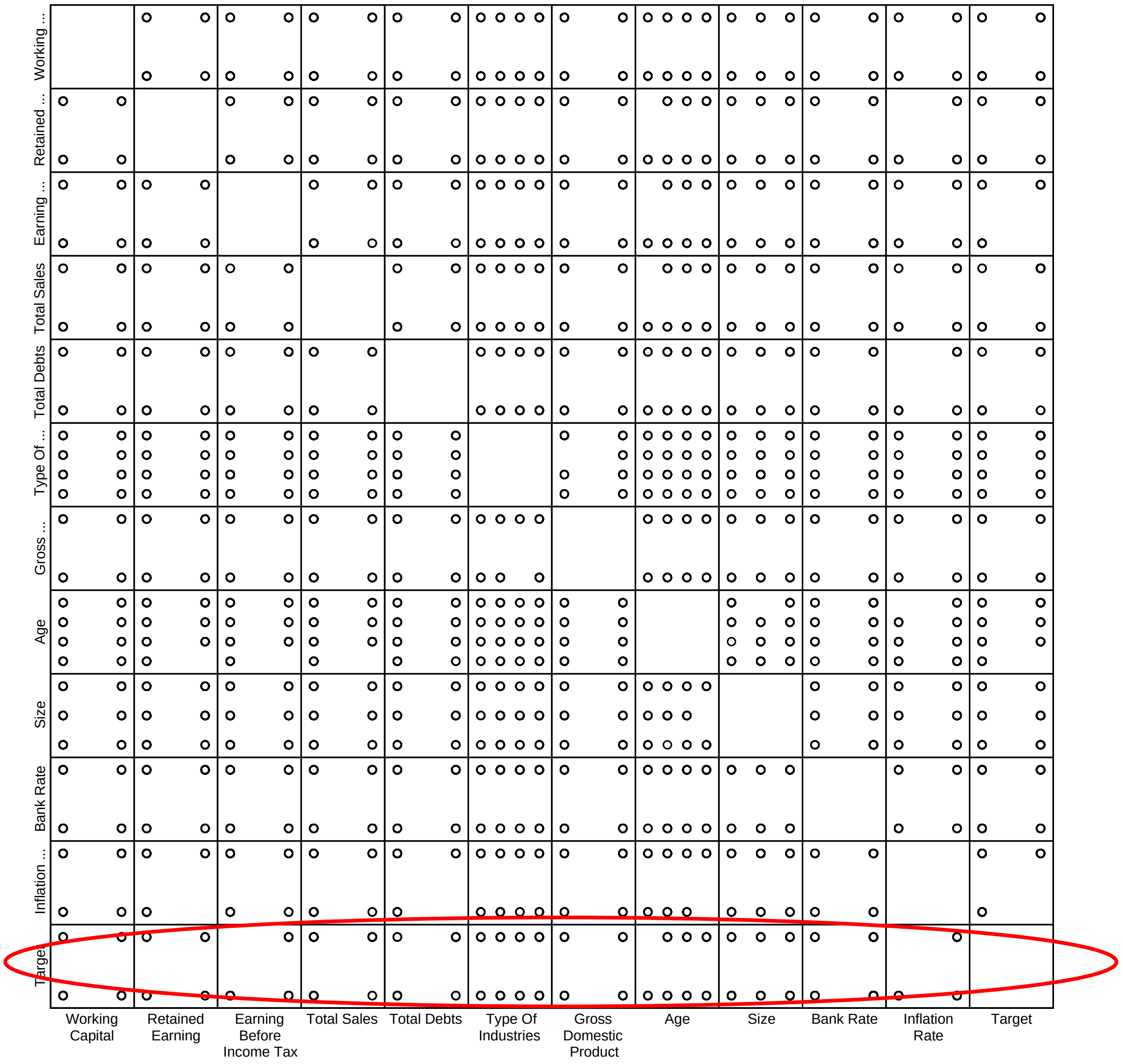


Figure 4.24 : Scatterplot Diagram

4.3 Correlation

Correlation is a statistical technique that shows strength of pairs of variables is related. Correlation or usually measured by correlation coefficient illustrate that the intensity and direction of the linear relationship between two random variables.

Multiple Correlation measures the amount of linear association between one dependent (response) variable and more than one independent (explanatory) variables. The correlation results between these variables are shown in Table 4.6.

Table 4.6 Correlations

N=367			Target	
Spearman's rho	Working Capital	Correlation Coefficient Sig. (1-tailed)	.791(**) .000	Very Strong
	Retained Earning	Correlation Coefficient Sig. (1-tailed)	.897(**) .000	Very Strong
	Earning Before Income Tax	Correlation Coefficient Sig. (1-tailed)	.940(**) .000	Very Strong
	Total Sales	Correlation Coefficient Sig. (1-tailed)	.728(**) .000	Very Strong
	Total Debts	Correlation Coefficient Sig. (1-tailed)	.966(**) .000	Very Strong
	Type Of Industries	Correlation Coefficient Sig. (1-tailed)	.182(**) .000	Weak
	Gross Domestic Product	Correlation Coefficient Sig. (1-tailed)	.064 .111	Weak
	Age	Correlation Coefficient Sig. (1-tailed)	.452(**) .000	Medium
	Size	Correlation	.291(**)	Medium

		Coefficient Sig. (1-tailed)	 .000	
	Bank Rate	Correlation Coefficient Sig. (1-tailed)	.057 .138	Weak
	Inflation Rate	Correlation Coefficient Sig. (1-tailed)	.398(**) .000	Medium
	Target	Correlation Coefficient Sig. (1-tailed)	1.000 .	

** Correlation is significant at the 0.01 level (1-tailed).

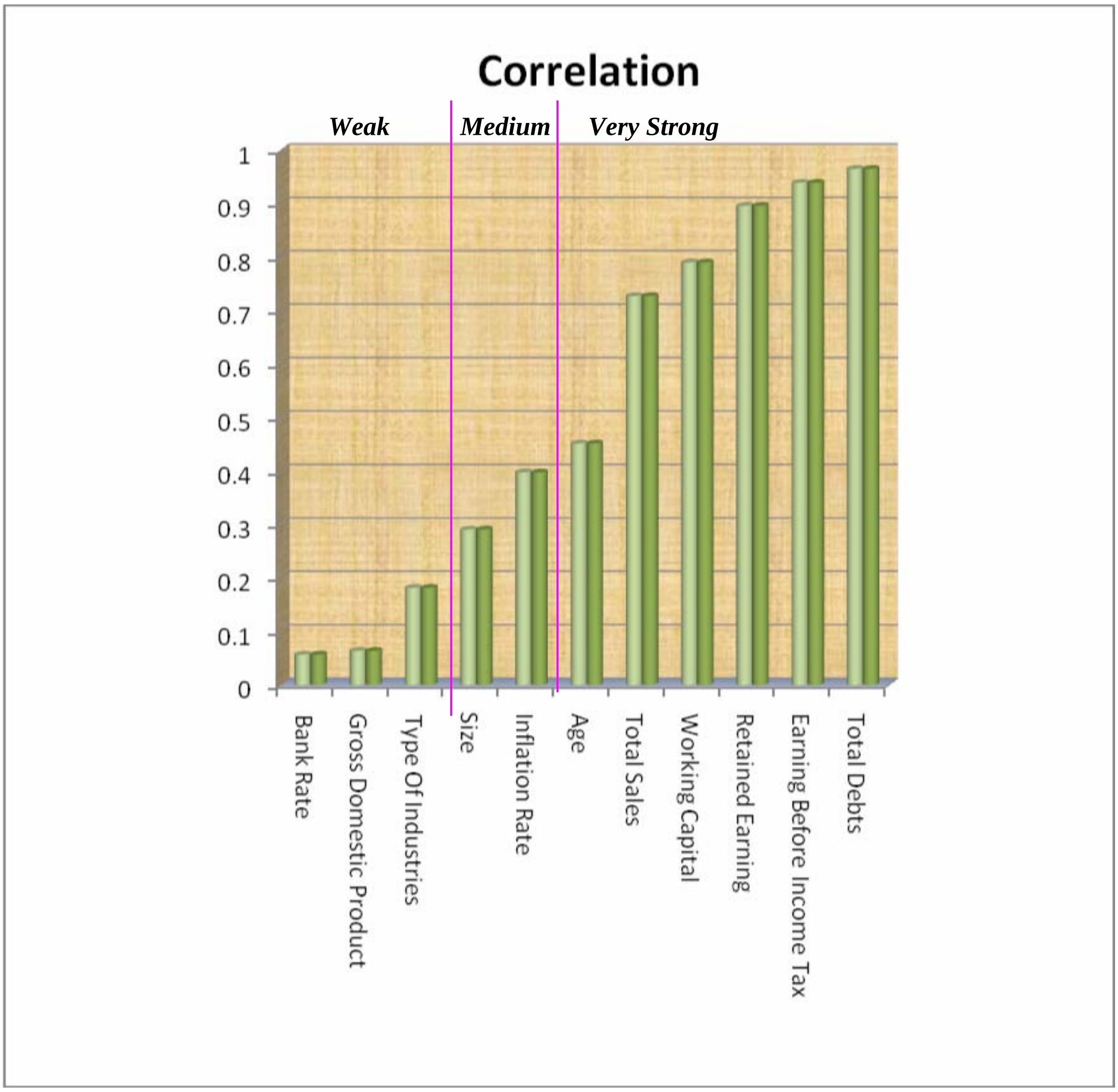


Figure 4.25 : Correlation

The results displayed in Table 4.6 and fig. 4.25 indicates that independent variables such as Working Capital, Retained Earning, Earning Before Income Tax, Total Sales, Total Debts have significantly very strong correlation with Target. The independent variable, Age, Size and Inflation Rate has medium correlation. however variable Type Of Industries, Bank Rate, Gross Domestic Product shows weak correlation with Target.

4.4 Logistic Regression

Logistic regression is a variant of nonlinear regression that proper when the dependent variable (target) has only two possible values (e.g., 0 and 1).

4.4.1 Examining the Variables

First of all to get a regression model, the scatterplot diagram (as fig. 4.24) was drawn to indicate the relationship between independent variables (Working Capital, Retained Earning, Earning Before Income Tax, Total Sales, Total Debts, Type Of Industries, Gross Domestic Product, Age, Size, Bank Rate and Inflation Rate) and dependent variables (Target). The scatterplot show the relationship between independent variables and dependent variables is not linear. Therefore, logistic regression analysis is more suitable for this data. Besides, since the dependent variable has two class values (benign and malignant), specifically binary logistic regression has to be applied on this data.

The analysis for bankruptcy dataset was progress by using SPSS version 12.0 to accomplish the experiment. To analyze data using binomial logistic regression, click ***Analyze*** and the drop down menu will be appear so select ***Binary Logistic*** then as show at Fig.4.26. the dialog box showed select Target as the dependent variable and Working Capital, Retained Earning, Earning Before Income Tax, Total Sales, Total Debts, Type Of Industries, Gross Domestic Product, Age, Size, Bank Rate and Inflation Rate as the covariate as show at Fig.4.27.

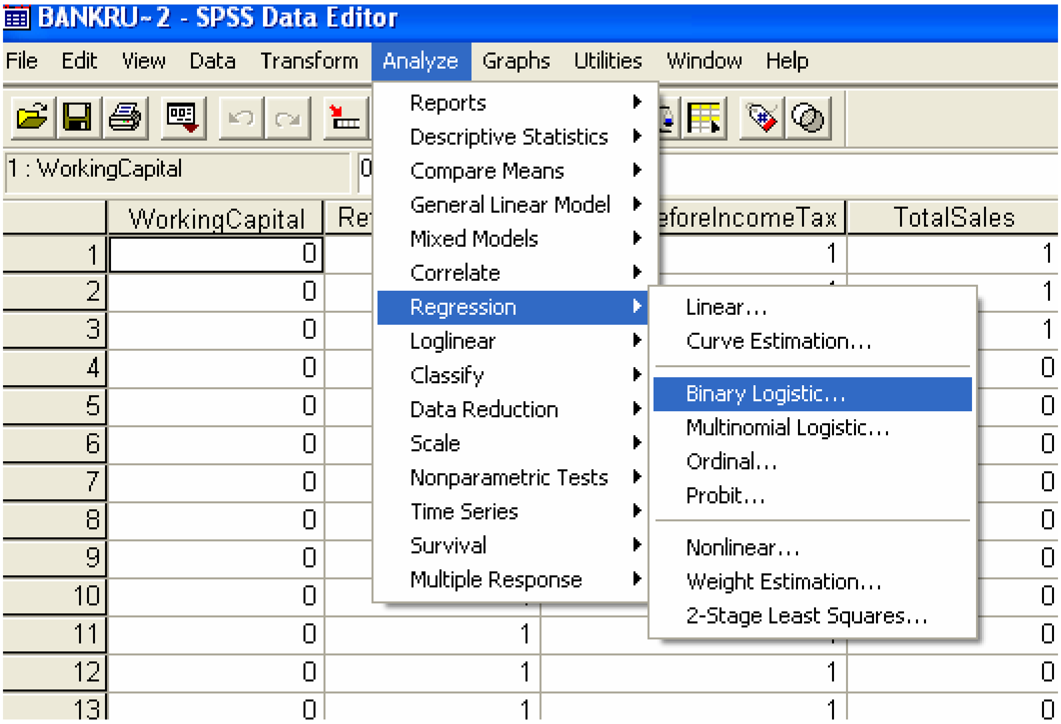
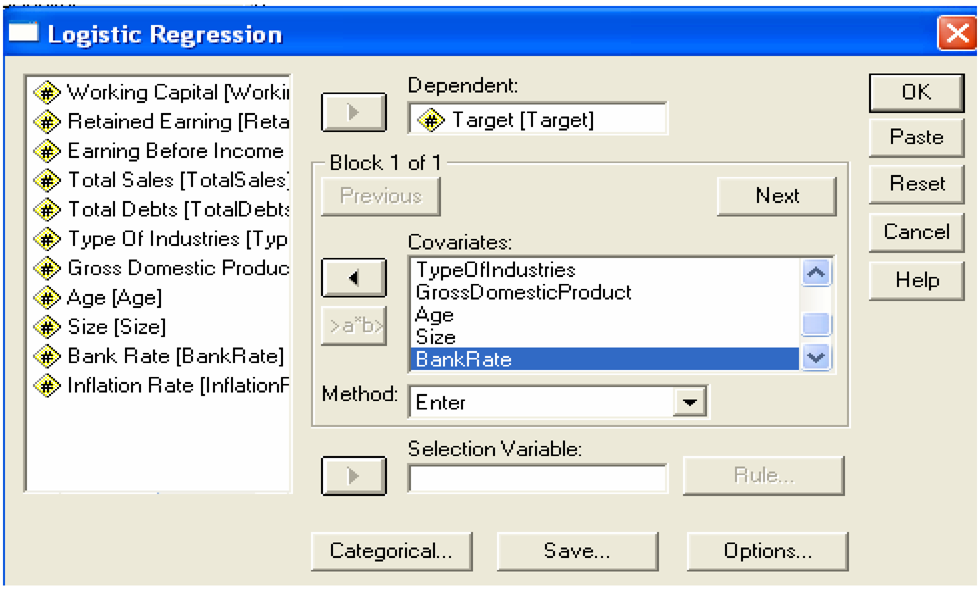


Figure 4.26: Logistic Regression using SPSS



Working Capital [Worki
Retained Earning [Reta
Earning Before Income
Total Sales [TotalSales
Total Debts [TotalDebts
Type Of Industries [Typ
Gross Domestic Produc
Age [Age]
Size [Size]
Bank Rate [BankRate]
Inflation Rate [InflationF

Dependent: Target [Target]

Covariates:
TypeOfIndustries
GrossDomesticProduct
Age
Size
BankRate

Method: Enter

Selection Variable:

Figure 4.27: Dialog Box of Logistic Regression

4.4.2 Case Processing Summary

Referring to table 4.7 case processing summary have 367 valid cases and 11 independent variables. Thus, the ratio of cases is 33.36 (367/11) to 1 independent variable in this analysis that greater than 10 to 1 (the minimum ratio of valid case). In additional, the ratio of 33.36 to 1 satisfies recommend ratio of 20 to 1. Incidentally, there is no missing value in the dataset, thus all cases are used in constructing logistic regression model.

Table 4.7: Case Processing Summary

Unweighted Cases(a)		N	Percent
Selected Cases	Included in Analysis	367	100.0
	Missing Cases	0	.0
	Total	367	100.0
Unselected Cases		0	.0
Total		367	100.0

a If weight is in effect, see classification table for the total number of cases.

4.4.3 Omnibus Tests of Model Coefficients

In this analysis, table 4.8 shows the probability of the model chi-square (493.335) that less than 0.001 and less than or equal to the level of significant of 0.05. The null hypothesis is the independent variables are not linearly related to the log odds of the dependent variable. Since the significant level shown in table 4.8 is less than 0.05, the null hypothesis has been rejected. Hence, it can be concluded that the independent variables are linearly related to the log odds of the dependent variable.

Table 4.8: Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	493.335	11	.000
	Block	493.335	11	.000
	Model	493.335	11	.000

4.4.3 Variables in the Equation

Table 4.9: Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	WorkingCapital	6.968	9207.108	.000	1	.999	1062.438
	RetainedEarning	8.234	7902.995	.000	1	.999	3765.785
	EarningBeforeIncome Tax	11.473	8494.730	.000	1	.999	96067.308
	TotalSales	13.971	8778.172	.000	1	.999	1168118.550
	TotalDebts	20.238	6761.194	.000	1	.998	615544512.475
	TypeOfIndustries	2.634	2566.785	.000	1	.999	13.928
	GrossDomesticProduct	7.498	11194.034	.000	1	.999	1803.890
	Age	4.765	3838.694	.000	1	.999	117.343
	Size	1.008	3122.000	.000	1	1.000	2.741
	BankRate	-2.167	10648.506	.000	1	1.000	.115
	InflationRate	8.734	14129.070	.000	1	1.000	6207.552
	Constant	-91.002	82993.593	.000	1	.999	.000

The independents (Working Capital, Retained Earning, Earning Before Income Tax, Total Sales, Total Debts, Type Of Industries, Gross Domestic Product, Age, Size, Bank Rate and Inflation Rate) are linearly related to the log odds for the independent variable (Target).

If significant value less than (<) 0.05 for a particular logit, then reject hypothesis. In this case, as table 4.9 non since significant value are less than (<) 0.05. Whereas Working Capital, Retained Earning, Earning Before Income Tax, Total Sales, Total Debts, Type Of Industries, Gross Domestic Product, Age, Size, Bank Rate and Inflation Rate with significant value greater than (>) 0.05.

Evaluate of significant greater than (>) 0.05, the correlation analysis indicated that all independent value has significant relationship between dependent variables, therefore, all of the independent variable all include in the prediction model so equation for bankruptcy or business insolvency can be written as:

Probability = 1 / 1 + exp (-(-91.002 + 6.968 * Working Capital+ 8.234 * Retained Earning + 11.473 * Earning Before Income Tax+ 13.971 * Total Sales+ 20.238 * Total Debts + 2.634 * Type Of Industries + 7.498 * Gross Domestic Product + 4.765 * Age + 1.008 * Size +-2.167 * Bank Rate + 8.734 * Inflation Rate)

4.4.5 Model Summary

According to Table 4.10, the Cox & Snell R Square obtained 73.9% of accuracy. Besides, Nagelkerke R. Square achieved 100%. That is mean Nagelkerke R. Square represented the highest accuracy. Further analysis on classification accuracy should be referred to Nagelkerke Model.

Table 4.10:Model Summary of accuracy

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000(a)	.739	1.000

a Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

4.4.6 Classification

According to table 4.11, the relative by chance accuracy rate was computed by calculating the number of correct cases for each group from complete number of cases in each group in the classification table. The proportion in the "Healthy" group is 221/367 = 0.602. The proportion in the "Failure" group is 146/367 = 0.398. in the addition, we can gain the proportional by chance the accuracy rate by squaring and sum both values in each group ($0.602^2 + 0.398^2 = 0.521$). Therefore the classification accuracy rate is 52.1% which is greater than 25% the proportional by chance accuracy rate.

Table 4.11: Classification Table(a,b)

	Observed		Predicted		
			Target		Percentage Correct
			Healthy	Failure	
Step 0	Target	0	221	0	100.0
		1	146	0	.0
	Overall Percentage				60.2

a Constant is included in the model.
b The cut value is .500

The accuracy rate computed in Table 4.12 was 100% which is greater than or equal to the proportional by chance accuracy rate 65.125% (52.1% * 1.25). Therefore, it is satisfied the criteria for classification accuracy.

Table 4.12: Classification Table (a)

	Observed		Predicted		
			Target		Percentage Correct
			0	1	
Step 1	Target	0	221	0	100.0
		1	0	146	100.0
	Overall Percentage				100.0

a The cut value is .500

4.4.7 Final Accuracy

From this experiment the best accuracy achieved from the bankruptcy data set is 100%.

4.5 Neural Networks

Neural networks or multilayered perceptron endow with models of data relationships throughout well interconnected, simulated “neurons” that accept inputs, apply weighting coefficients and feed their output to other “neurons” and continue the process through the network to the final output. Some neurons may send feedback to earlier neurons in the network. Neural networks act as “trained” that deliver the desired result by an iterative process where the weights applied to every input at every hidden unit then it is adjusted to optimize the desired output.

4.5.1 Neural Network Tool

Neural Connection is software that offers a unique technology that recognizes complex patterns and trends in data, much like the human brain. As a result, users get a more accurate predictive model in a shorter amount of time.

Just as the brain learns from past experience, neural networks apply knowledge from past experience to new problems. Neural networks acquire the knowledge by learning patterns in a set of data. After the network has been trained and validated, the resulting model may be applied to data it has not seen previously for prediction, classification, and time series analysis or data segmentation.

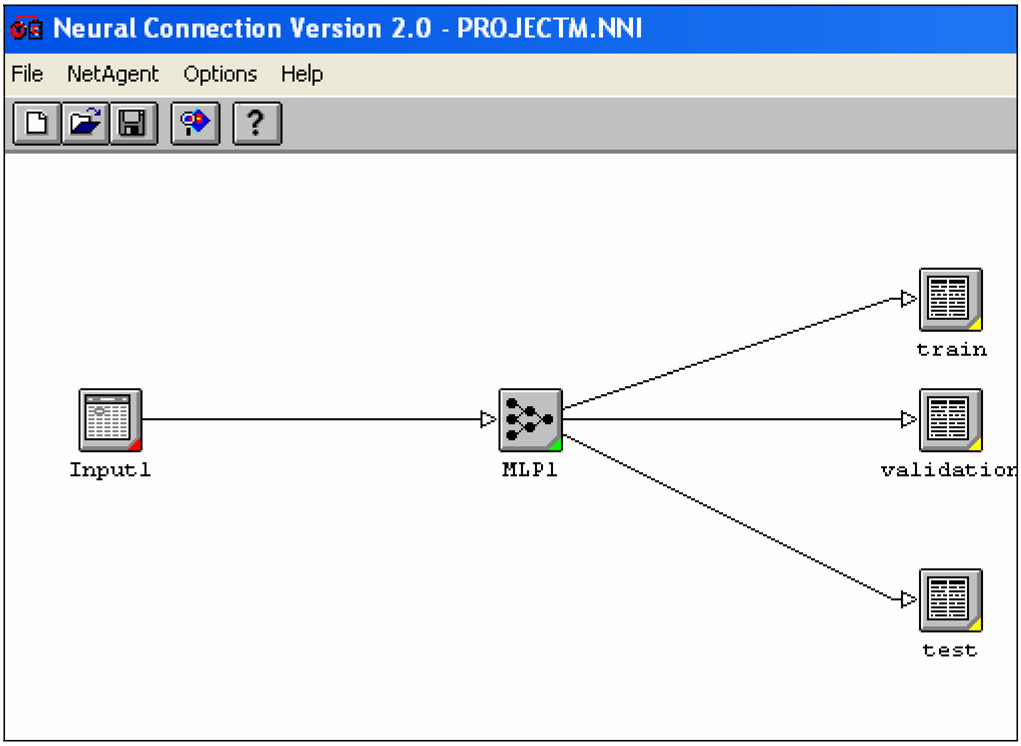
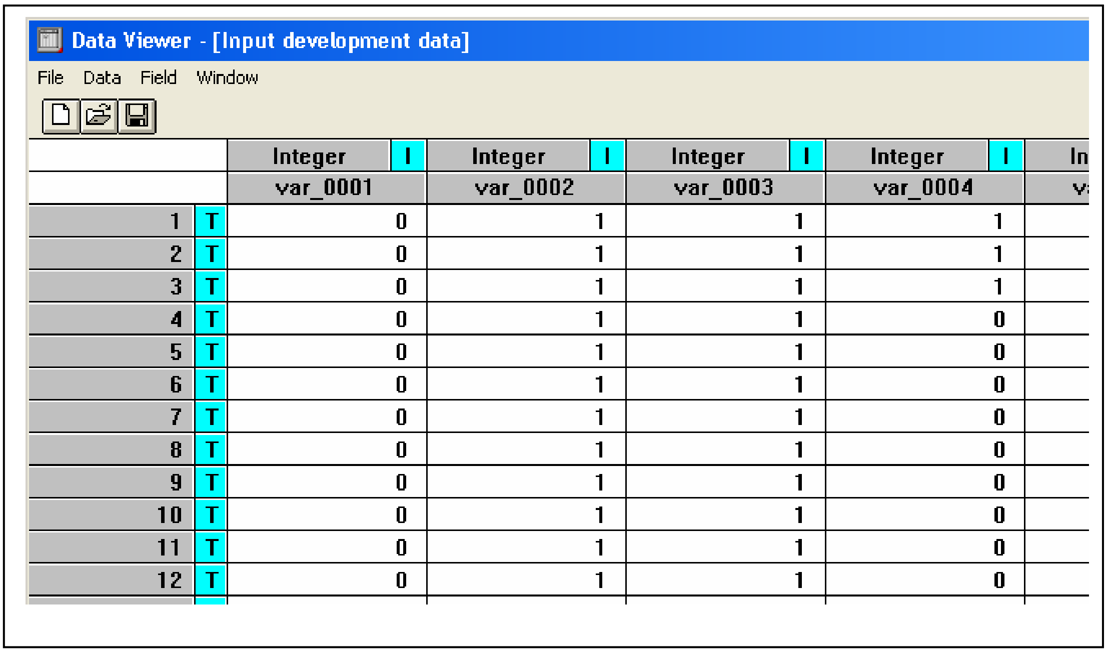


Figure 4.28: Neural Connection

The Neural Connection Version 2.0 is the tools that been used to predict the bankruptcy company. Firstly, drag the input icon to the stage and then MLP1 icon and followed by the three text icon renamed as train, validation and test as illustrated at fig. 4.28.

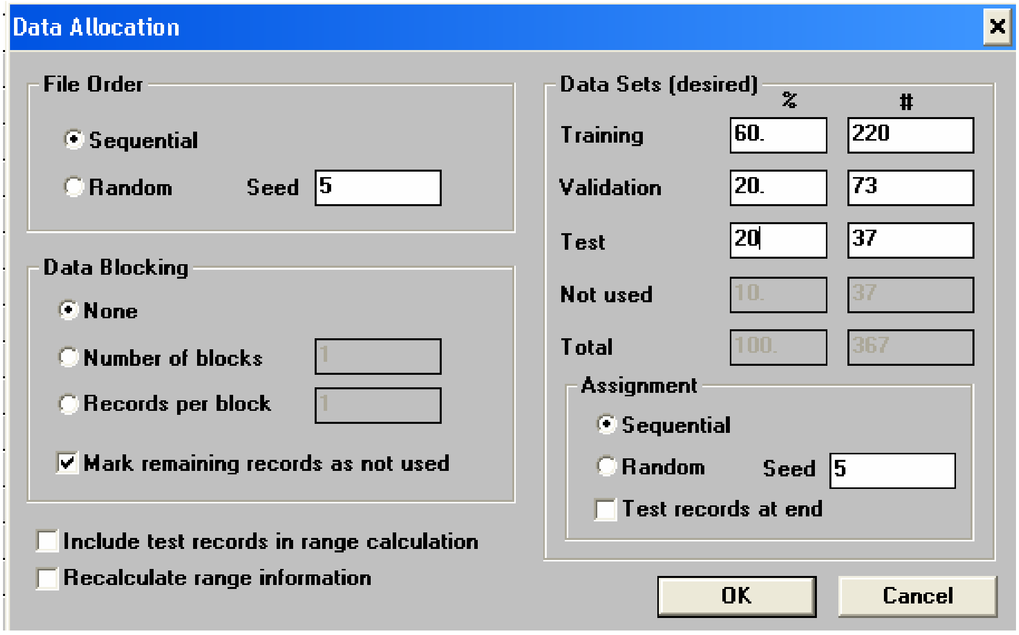
Next, open the data viewer page by clicking the input icon and then right click to view. The next step is to load the data set into Neural Connection with open the file and choosing the data set with Flat file type. Then the illusion as fig. 4.29 will be appeared. Distinguish that the selection of patterns (records) is sequential by default. By default the data set is allocated to training (80%), validation (10%) and testing (10%) sequentially.



Data Viewer - [Input development data]

		Integer I	Integer I	Integer I	Integer I	In
		var_0001	var_0002	var_0003	var_0004	v
1	T	0	1	1	1	
2	T	0	1	1	1	
3	T	0	1	1	1	
4	T	0	1	1	0	
5	T	0	1	1	0	
6	T	0	1	1	0	
7	T	0	1	1	0	
8	T	0	1	1	0	
9	T	0	1	1	0	
10	T	0	1	1	0	
11	T	0	1	1	0	
12	T	0	1	1	0	

Figure 4.29: Data Viewer



Data Allocation

File Order

● Sequential

○ Random Seed 5

Data Blocking

● None

○ Number of blocks 1

○ Records per block 1

☑ Mark remaining records as not used

□ Include test records in range calculation

□ Recalculate range information

Data Sets (desired)

	%	#
Training	60.	220
Validation	20.	73
Test	20	37
Not used	10.	37
Total	100.	367

Assignment

● Sequential

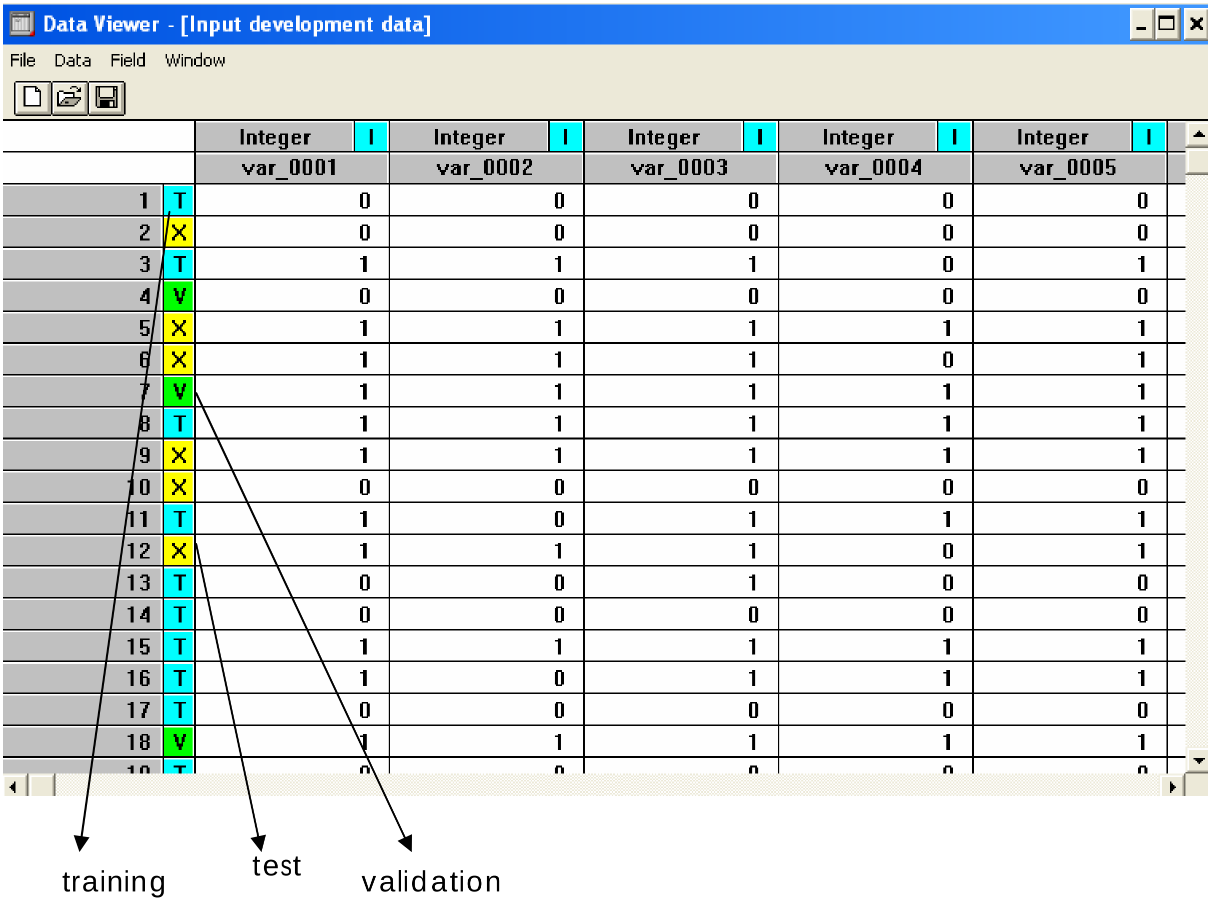
○ Random Seed 5

□ Test records at end

Figure 4.30: Data Allocation

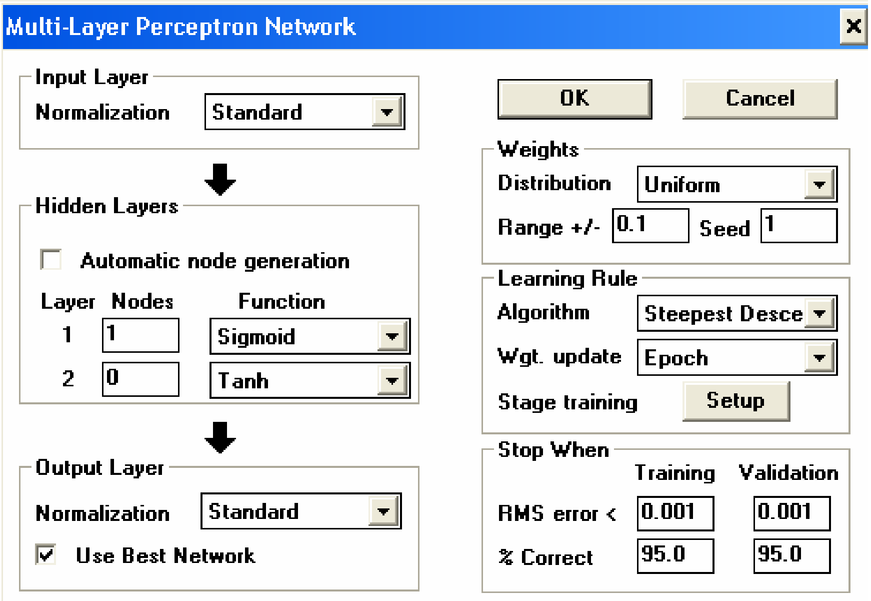
The percentage for training, validation and test sets can be changed as desired. The data allocation option in Neural Connections allows a user to change the data allocation percentage. Such facility is represented in Fig. 4.30. This means that the data is assigned to the first 60% of data training, 20% as validation and 20% as test. However data allocation can be changed to other percentage such as 70%: 15%: 15% or 60%: 20%: 20% for most real application. For scientific and engineering application, data allocation can vary to 40%:30% or 50%:25%:25%

The data is allocated to 60% of data training set (written as T), 20% as validation set (written as V) and 20% as test set (written as X). To illustrate how neural connection allocates data, refer to fig 4.31.



		Integer I	Integer I	Integer I	Integer I	Integer I
		var_0001	var_0002	var_0003	var_0004	var_0005
1	T	0	0	0	0	0
2	X	0	0	0	0	0
3	T	1	1	1	0	1
4	V	0	0	0	0	0
5	X	1	1	1	1	1
6	X	1	1	1	0	1
7	V	1	1	1	1	1
8	T	1	1	1	1	1
9	X	1	1	1	1	1
10	X	0	0	0	0	0
11	T	1	0	1	1	1
12	X	1	1	1	0	1
13	T	0	0	1	0	0
14	T	0	0	0	0	0
15	T	1	1	1	1	1
16	T	1	0	1	1	1
17	T	0	0	0	0	0
18	V	1	1	1	1	1

Figure 4.31: Denoted of Data Allocation



Multi-Layer Perceptron Network

Input Layer
Normalization: Standard

Hidden Layers
□ Automatic node generation

Layer	Nodes	Function
1	1	Sigmoid
2	0	Tanh

Output Layer
Normalization: Standard
☑ Use Best Network

Weights
Distribution: Uniform
Range +/- 0.1 Seed 1

Learning Rule
Algorithm: Steepest Desce
Wgt. update: Epoch
Stage training

Stop When

	Training	Validation
RMS error <	0.001	0.001
% Correct	95.0	95.0

Figure 4.32: Multilayer Perceptron Network (MLP)

The multilayer perceptron (MLP) icon allows the user to control the parameters for training neural networks using backpropagation learning. When applying neural network to any data sets, the range of values for each attribute should lie between 0 and 1, or between -1 and 1. The window shown should look like in Fig. 4.32.

From the Multilayer Perceptron dialog, a user could select whether to normalize the attribute values using the standard normalization technique or do nothing to the data set. In this case, choose standard normalization technique.

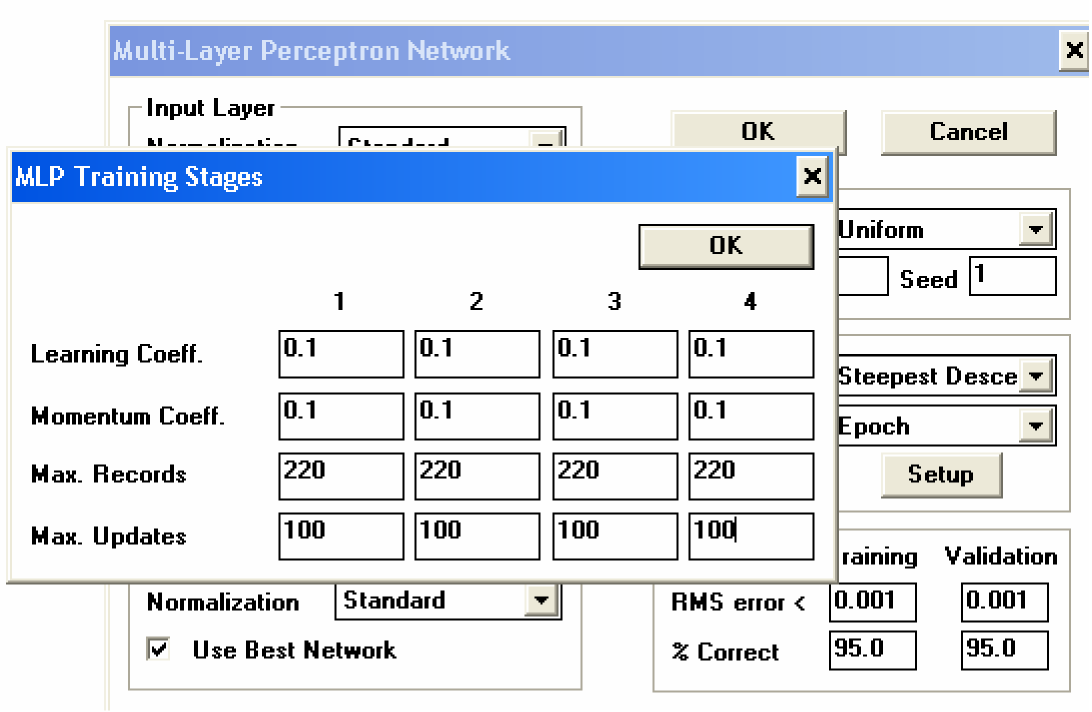
At the hidden layer, a few options are provided such as automatic node generation, number of layer, number of nodes and the activation function for backpropagation learning. In this case, automatic node generation is disable, hence the options can be changed as desired.

While the output layer, a normalization required to select here standard normalization technique selected and USE BEST NETWORK button is enable.

By default the weight is initialized using seed 1. The seed number for the weights can be selected to produce a network with different weight and seed. Normally, the weight seed does not affect the overall performance of the network.

The learning rate dialog box contain of options for learning algorithm, weight update and stage training. Two types of learning algorithm available, which are the Conjugate Gradient and the Steepest Descent.

Once the gradient descent is selected, weight update is disabling. This indicates that a user cannot control the parameters for learning algorithm. In this project, a steepest descent is chosen so that other parameters such as the learning and momentum rates can be controlled.



	1	2	3	4
Learning Coeff.	0.1	0.1	0.1	0.1
Momentum Coeff.	0.1	0.1	0.1	0.1
Max. Records	220	220	220	220
Max. Updates	100	100	100	100

Input Layer

Normalization: Standard

☑ Use Best Network

Uniform

Seed: 1

Steepest Desce

Epoch

	raining	Validation
RMS error <	0.001	0.001
% Correct	95.0	95.0

Figure 4.33: MLP Training Stages

Here we can choose the suitable learning rate, momentum rate, max records and maximum updates. Maximum records are taken from data training that consists of 60% of the data set or 220 record (60%*367). While maximum updates is number of epoch that been used. In this project, the max update of 150 epoch is set. To control the learning rate and momentum rate, a selected rate is required to be used for all training stages (i.e the same for stage 1, 2, 3 and 4). The initial window for MLP training stages is illustrated in Fig. 4.33.

To stop the training, a few options are provided by Neural Connection such as setting the number of Epoch, Root Mean Squared Error (RMS) or % Correct. In this project, the number epoch is controlled so that if this condition is first satisfied, the training will stop.

4.5.2 The Experiments

A few experiments was conducted to determine the multilayer perceptron's architecture as well as the backpropagation learning parameters. The experiments and the results are explained in details in the following subsections. The following parameters are fixed for initial experiments, but the number of hidden unit varies.

Input Layer:

Data Allocation: 60% for training, 20% for validation and 20% for testing

Read the data as Sequential File.

Assigned the data sequentially.

Standard normalization technique is necessary at the input layer.

Hidden Layer

No. of Hidden Layer: 1

No. of Hidden Unit: 1

Activation Function: Sigmoid

Output Layer

Standard normalization technique is necessary at the output layer.

Learning Parametes

Weight seed no.: 1 (by default)

Learning algorithm: Steepest Descent

For weight update:

Learning rate= 0.1

Momentum rate =0.1

Number of epoch: 150

4.5.2.1 To determine the most suitable number of hidden units

The different number of hidden units is explored and the results are displayed in Table 4.13.

Table 4.13: The training, validation and test results using various number of hidden units

No of hidden unit	Accuracy		
	Training	Validation	Test
1	100	100	97.26
2	100	100	97.26
3	100	100	97.26
4	100	100	100
5	100	100	100
6	100	100	100
7	100	100	100
8	100	100	100
9	100	100	100
10	100	100	100

The criterion of choosing most suitable ***Hidden Unit*** based on training and testing results. The most suitable ***Hidden Unit*** was chosen from higher percentage of testing and lower percentage of training. Referring to the test results exhibited in table 4.13, ***Hidden Unit*** 4 to 10 obtains the highest percentage of test accuracy (100%). therefore the ***Hidden Unit*** 3, 4 and 6 been chosen. To confirm most suitable ***Hidden Unit*** for bankruptcy data set, several experiments involving different weight seed were conducted. The results were shown in table 4.14.

Table 4.14: The Weight seed using various number of hidden units

Weight seed	**Hidden unit**					
	3		**4**		**6**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
1	100	97.26	100	100	100	100
2	100	100	100	100	100	100
3	100	100	100	100	100	100
4	100	100	100	100	100	100
6	100	100	100	100	100	100
7	100	100	100	100	100	100
8	100	97.26	100	100	100	100
9	100	97.26	100	100	100	100
10	100	97.26	100	100	100	100
11	100	100	100	100	100	100
Average	**100**	**98.904**	**100**	**100**	**100**	**100**

The experiment for ***Hidden Unit*** 3, 4, 6 was conducted using different weight seeds. The average test accuracy result for ***Hidden Unit*** 4 and 6 are slightly higher then ***Hidden Unit*** 3. Although ***Hidden Unit*** 4 and 6 shows more accuracy but the differences in average train and test results with ***Hidden Unit*** 3 are negligible. In this case, ***Hidden Unit*** 4 and 6 shows the same results, thus we need to mull over number of epoch.

Table 4.15: The number of hidden units using various number of epoch

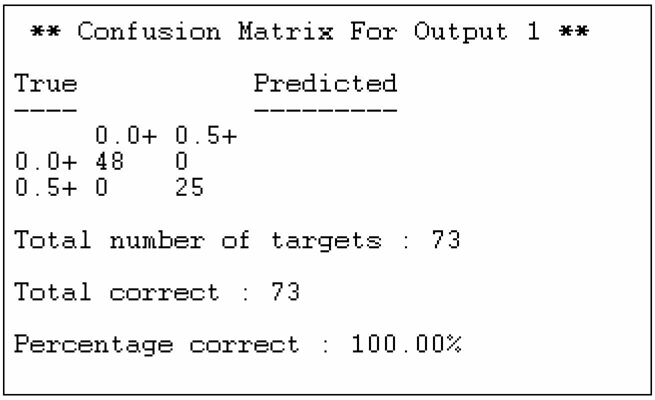
Epoch	**Hidden unit**					
	3		**4**		**6**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
150	100	97.26	100	100	100	100
200	100	97.26	100	100	100	100
250	100	97.26	100	100	100	100
300	100	97.26	100	100	100	100
350	100	97.26	100	100	100	100
400	100	97.26	100	100	100	100
450	100	97.26	100	100	100	100
500	100	97.26	100	100	100	100

The condition of choosing most suitable ***Hidden Unit*** based on number of epoch. The most suitable ***Hidden Unit*** was chosen from higher percentage of number of epoch. Referring to the test results exhibited in table 4.15, number of epoch 150 highest percentage of test accuracy (100%). therefore the number of epoch 150 been chosen. The results verification was shown in table 4.16.

Table 4.16: The weight seed using various number of hidden units

Weight seed	**Hidden unit**			
	4		**6**	
	Train	**Test**	**Train**	**Test**
1	100	100	100	100
2	100	100	100	100
3	100	100	100	100
4	100	100	100	100
7	100	100	100	100
10	100	100	100	100
11	100	100	100	100
12	100	100	100	100
16	100	100	100	100
19	100	97.26	100	100
Average	**100**	**99.726**	**100**	**100**

The experiment for ***Hidden Unit*** 4 and 6 was conducted using different weight seeds. The average test accuracy result for ***Hidden Unit*** 6 are slightly higher then ***Hidden Unit*** 4. Although ***Hidden Unit*** 4 and 6 shows more accuracy but the differences in average train and test results with ***Hidden Unit*** 3 are negligible. Thus ***Hidden Unit*** 6 been chosen.



```
 ** Confusion Matrix For Output 1 **

True           Predicted
----           ---------
     0.0+ 0.5+
0.0+ 48   0
0.5+ 0    25

Total number of targets : 73

Total correct : 73

Percentage correct : 100.00%
```

Figure 4.34: Confusion matrix for output

From Neural Connection, the result was tested. Fig. 4.34 show the Confusion matrix for output means the hidden unit 6 is suitable to use this study.

4.5.2.2 To determine the most suitable learning rate

Learning rate used to control parameter of some training algorithms. For example, learning rate controls the step size when weights are iteratively adjusted. The learning procedure requires the change in weight is proportional to True gradient descent requires in nitesimal steps. The constant of proportionality is the learning rate. For practical, choose a learning rate that is as large as possible without leading to oscillation. Make the change in weight dependent of the past weight change by adding a momentum term is the way to avoid oscillation at large.

The following parameters are fixed but number of learning rate will vary:

Hidden unit= 6	Momentum rate =0.1
Activation function = sigmoid	Number of epoch = 150

Table 4.17: The training and test results using various number of learning rate

Learning rate	**Accurancy**	
	Training	**Test**
0.1	100	100
0.2	100	100
0.3	100	100
0.4	100	100
0.5	100	100
0.6	100	100
0.7	100	100
0.8	100	100
0.9	100	100
1.0	100	100

The condition of choosing most suitable ***Learning Rate*** based on training and testing results. The most suitable ***Learning Rate*** was chosen from higher percentage of testing and lower percentage of training. Referring to the test results exhibited in table 4.17, ***Learning Rate*** 0.1 to 1.0 obtains the highest percentage of test accuracy (100%). Therefore the ***Learning Rate*** 0.1 been chosen. To confirm most suitable ***Learning Rate*** for bankruptcy data set, several experiments involving different weight seed were conducted. The results were shown in table 4.18.

Table 4.18: The weight seed using various number of learning rate

Weight seed	**Learning rate**					
	0.1		**0.3**		**0.4**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
1	100	100	100	100	100	97.26
2	100	100	100	100	100	100
3	100	100	100	100	100	100
4	100	100	100	100	100	100
5	100	100	100	100	100	100
6	100	100	100	100	100	100
7	100	100	100	100	100	100
8	100	100	100	100	100	100

9	100	100	100	100	100	100
10	100	100	100	100	100	100
Average	**100**	**100**	**100**	**100**	**100**	**99.726**

The experiment for ***Learning Rate*** 0.1, 0.3, 0.4 was conducted using different weight seeds. The average test accuracy result for ***Learning Rate*** 0.1 and 0.3 are slightly higher then ***Learning Rate*** 0.4. Although ***Learning Rate*** 0.1 and 0.3 shows more accuracy but the differences in average train and test results with ***Learning Rate*** 0.4 are negligible. In this case, ***Learning Rate*** 0.1 been chosen.

4.5.2.3 To determine the most suitable momentum rate

Table 4.19: The training and test results using various momentum rate

momentum rate	**Accuracy**	
	Training	**Test**
0.1	100	100
0.2	100	100
0.3	100	100
0.4	100	100
0.5	100	100
0.6	100	100
0.7	100	100
0.8	100	100
0.9	100	100
1.0	100	100

The criterion of choosing most suitable ***Momentum Rate*** based on training and testing results. The most suitable ***Momentum Rate*** was chosen from higher percentage of testing and lower percentage of training. Referring to the test results exhibited in table 4.19, ***Momentum Rate*** 0.1 to 1.0 obtains the highest percentage of test accuracy (100%). therefore the ***Momentum Rate*** 0.1, 0.3 and 0.4 been chosen.

To confirm most suitable ***Momentum Rate*** for bankruptcy data set, several experiments involving different weight seed were conducted. The results were shown in table 4.20.

Table 4.20: The weight seed using various number of momentum rates

Weight seed	**Momentum rate**					
	0.1		**0.3**		**0.4**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
1	100	100	100	100	100	100
2	100	100	100	100	100	100
3	100	100	100	100	100	100
4	100	100	100	100	100	100
5	100	100	100	100	100	100
6	100	100	100	100	100	100
7	100	100	100	100	100	100
8	100	100	100	100	100	100
9	100	100	100	100	100	100
10	100	100	100	100	100	100
Average	**100**	**100**	**100**	**100**	**100**	**100**

The experiment for ***Momentum Rate*** 0.1, 0.3, 0.4 was conducted using different weight seeds. The average test accuracy for that ***Momentum Rate*** is slightly shows the same results. Thus ***Momentum Rate*** 0.1 been chosen.

4.5.2.4 To determine the best Activation Function

The activation function acts as a squashing function, such that the output of a neuron in a neural network is between certain values (usually 0 and 1, or -1 and 1). In general, there are three types of activation functions. First, Threshold Function which takes a value of 0 if the summed input is less than a certain threshold value (v), and the value 1 if the summed input is greater than or equal to the threshold value.

Second, is the Piecewise-Linear function. This function can take on the values of 0 or 1, but can also take on values between that depending on the amplification factor in a certain region of linear operation. Third, is the sigmoid function that in range between 0 and 1, but it is also sometimes useful to use the range -1 to 1. An example of the sigmoid function is the hyperbolic tangent function.

Hidden unit = 6 Momentum rate = 1.0

Learning rate = 0.1

Table 4.21: Result to determine the best Activation Function

Weight seed	**Linear**		**Sigmoid**		**Tanh**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
1	100	100	100	100	100	97.26
2	100	100	100	100	100	100
3	100	100	100	100	100	100
4	100	100	100	100	100	100
5	100	100	100	100	100	100
6	100	100	100	100	100	100
7	100	100	100	100	100	100
8	100	100	100	100	100	100
9	100	100	100	100	100	100
10	100	100	100	100	100	100
Average	**100**	**100**	**100**	**100**	**100**	**100**

The experiment for Active Function of Linear, Sigmoid and Tanh was conducted using different weight seeds. The average test accuracy result for Linear, Sigmoid and Tanh are slightly same.

4.6 Conclusion

The results show that the neural network is a most suitable data mining approach in this study compare with logistic regression.

CHAPTER 5

CONCLUSION

This chapter concludes the findings of the study. The limitations of the study are also mentioned in this chapter.

5.1 Conclusion

This study carried out analysis in order to build a bankruptcy model with financial factor by using data mining techniques. The bankruptcy model is generally used to alert the company for business insolvency and precisely handle and control the financial problem from the various types of risk. In general, data mining methods such as neural networks and logistic regression could be useful techniques to the financial analyst. The results show that the neural network obtains 100% accuracy for predicting the bankruptcy of a company. However, the data mining techniques tend to require more historical data than the standard models and, in the case of neural network, it can be difficult to interpret.

5.2 Limitation

Predictive modeling as well as neural network is concerned with analyzing patterns and trends in historical and operational data in order to transform data into actionable decisions. This enables analysis and modeling the dynamics of the application can be carried out. In its raw form, this data is of limited value and is mainly used for reporting what has happened.

Logistic regression fits an S-shaped logistic function to the data. As with general nonlinear regression, logistic regression cannot easily handle categorical variables nor is it good for detecting interactions between variables.

5.3 Recommendation

Further, the system of the prediction can be constructed based on the model. The improvements of the system should include integrating neural network technique with another technique such as hybrid techniques. Besides, the study should perform on financial data parameter in tern of global features that precise to business insolvency.

REFERENCE

Ahmed, K. M., El-Makky, N. M. and Taha, Y. (1998). Effective data mining: a data warehouse-backboned architecture

Abdelwahed, T. and Amir, M. (2005). New evolutionary bankruptcy forecasting model based on genetic algorithms and neural networks

Agarwal, S., Ambrose, B. W., & Chomsisengphet, S. (2005). Asymmetric information and the automobile loan market.

Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Preceding of The Journal of Finance, 23, 589-609.

Altman, E. (1983). Corporate financial distress - a complete guide to predicting avoiding and dealing with bankruptcy. New York: Willey.

Angelidis, D. and Lyroudi, K. (2006). Efficiency in the Italian banking industry: data envelopment analysis and neural networks

Apte, C., Liu, B., Pednault, E. P.D., Smyth, P. (2002). Business applications of data mining. Proceeding of Communications of The ACM Vol. 45, No. 8

Bloemer, J., Brijs, T., Swinnen, G., Vanhoof, K. (2002). Identifying latently dissatisfied customers and measures for dissatisfaction management. Proceeding of International Journal of Bank Marketing. pp. 27-37.

Cavaretta, M. (2006). Data Mining Challenges in the Automotive Domain

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Thomas, R., Shearer, C., & Wirth, R., (2000). CRISP-DM 1.0 Step-by-step data mining guide.

Charalambous, C. and Martzoukos, S. H. (2005). Hybrid Artificial Neural Networks for Efficient Valuation of Real Options and Financial Derivatives

Chen, R. S., Wu, R. C. & Chen, J. Y. (2005). Data Mining Application in Customer Relationship Management Of Credit Card Business. Proceedings of the 29th Annual International Computer Software and Applications Conference. Institute of Information Management, National Chiao Tung University, Taiwan.

Chen, Y., Tsai, F. S., and Chan, K. L. (2007). Blog Search and Mining in the Business Domain proceeding of ACM SIGKDD Workshop on Domain Driven Data Mining

Cipollini, A. and Missaglia, G. (2005). Business cycle effects on capital requirements: scenario generation through Dynamic Factor analysis

Cumby, C., Fano, A., Ghani, R. and Krema, M. (2004). Predicting Customer Shopping Lists from PointofSale Purchase Data

Dass, R. (2007).Data Mining In Banking And Finance: A Note For Banker. Indian Institute of Management Ahmedabad

Ettl, M., Zadrozny, B., Chowdhary, P., & Abe, N. (2005). Business Performance Management System for CRM and Sales Execution. International Workshop on Business Process Monitoring & Performance Management. IBM T.J. Watson Research Center, USA.

Foster, D. P. and Stine, R. A. (2001). Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy

Fang, R. and Tuladhar, S. (2004). Teaching Data Warehousing And Data Mining In Graduate Program Of Information Technology. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)

Ghani, R. and Soares, C. (2006). Data Mining for Business Applications. Proceeding of KDD-2006 Workshop

Goodarzi, A., Kohavi, R., Harmon, R., Senkut, A. (1998). Loan Prepayment Modeling. American Association for Artificial Intelligence.

He, D. (2002): Resolving Non-performing Assets of the Indian Banking System.

Hekanaho, J., Back, B., Sere, K. and Laitinen, T. (1998). Analysing Bankruptcy Data with Multiple Methods

Hueglin, C. and Vannotti, F. (2001). Data Mining Techniques to Improve Forecast Accuracy in Airline Business

Hunziker, P., Maier, A., Nippe, A., Tresch, M., Weers, D., & Zemp, P. (1999). Data mining at a major bank: Lessons from a large marketing application. Credit Suisse.

Ikizle, N., & Guvenir, H. A. (2002). Mining Interesting Rules in Bank Loans Data. Bilkent University,Department of Computer Engineering.

Kalos, A. and Rey, T. (2005) Data Mining in the Chemical Industry

Kuhlmann, M., Shohat, D. and Schimpf, G. (2003). Role Mining - Revealing Business Roles for Security Administration using Data Mining Technology

Lei, H., Chan, C.C.(2003). Rule-Based Classifier for Bankruptcy Prediction

Liao, S. H., & Chen, Y. J. (2004). Mining customer knowledge for electronic catalog marketing. Expert Systems with Applications. 27. pp.521–532. Department of Management Sciences, Decision Making, Tamkang University.

Lin, Z. and Wu, J. (2005) Research on Audit Informatization under the Environment of E-business.

Lo, V. S.Y. (2002) The True Lift Model - A Novel Data Mining Approach to Response Modeling in Database Marketing Proceeding of SIGKDD Explorations, 4(2), p.p. 78-86

Marchiori, E. (2002). Data Mining. Free University Amsterdam.

McLaren, I. (1999).Designing the Data Warehouse for Effective Data Mining. Searchspace Limited.

Mierzejewski, F. (2006): Economic capital allocation under liquidity constraints. Published in: Proceedings of the 4th Actuarial and Financial Mathematics Day (2006): pp. 107-116.

Mitra, S., Pal, S. K., Mitra, P. (2002). Data Mining in Soft Computing Framework:A Survey. Proceedings of IEEE Transactions on Neural Networks. v.13. pp3-14.

Mody, A., & Patro, D. (1996). Methods of Loan Guarantee Valuation and Accounting.

Molloy, I., Chen, H., Li, T., Wang, Q., Li, N. and Bertino, E. (2008) Mining Roles with Semantic Meanings

Nakaoka, I., Tani, K., Hoshino, Y.and Kamei, K. (2006). A Bankruptcy Prediction Method Based on Cash flow Using SOM

Nasir, M.L., John,R.I., Bennett, S.C..(2005). Predicting Corporate Bankruptcy Using Modular Neural Networks

Park, Y. (2008): Banking Market Concentration and Credit Availability to Small Businesses

Sai, Y., Zhong, C. J., Nie, P. Y.(2007). A Hybrid RST and GA-BP Model for Chinese Listed Company Bankruptcy Prediction

Schied, A. and Schoeneborn, T. (2008): Risk aversion and the dynamics of optimal liquidation strategies in illiquid markets.

Schmidt, F. (2009): The Undervaluation of Distressed Company's Equity.

Scott, R. I., Svinterikou, S., Tjortjis, C., Keane, J. A. (1999). Experiences of using Data Mining in a Banking Application. Department of Computation, UMIST, Manchester, UK.

Shi, A., Long, A. & Newcomb, D. (2001). Enhancing e-Business Through Web Data Mining.

Shin K. S. , Lee, T. S., Kim, H. J. (2004). An application of support vector machines in bankruptcy prediction model

Topaloglou, N., Vladimirou, H., and Zenios, S. A. (2005) Controlling Currency Risk with Options or Forwards

Vieira, A. S., Ribeiro, B., Mukkamala, S.,. Neves, J. C and Sung, A. H. (2004). On the Performance of Learning Machines for Bankruptcy Detection

Yeung, D. S., Ng, W. W. Y., Chan, A. P. F., Chan, P. P. K., Firth, M., Tsang, E. C. C. (1998). Bankruptcy Prediction Using Multiple Intelligent Agent System via a LocalizedGeneralization Error Approach

Yoon, J. S., Kwon, Y. S., Roh, T. H. (2007). Performance Improvement of Bankruptcy Prediction using Credit Card Sales Information of Small & Micro Business

Zhao, Y., Zhang, H., Figueiredo, F., Cao, L., Zhang, C. (2007). Mining for Combined Association Rules on Multiple Datasets. Proceeding of ACM SIGKDD Workshop on Domain Driven Data Mining