DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DETERMINATION OF THE KEY FINANCIAL RATIOS IN THE SUCCESS OF FIRMS IN DIFFERENT SECTORS THROUGH DATA MINING

by
Müge AVŞAR

May, 2018
İZMİR

DETERMINATION OF THE KEY FINANCIAL RATIOS IN THE SUCCESS OF FIRMS IN DIFFERENT SECTORS THROUGH DATA MINING

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Master of Science of Philosophy in Statistics, Applied Statistics Program

by
Müge AVŞAR

May, 2018
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"DETERMINATION OF THE KEY FINANCIAL RATIOS IN THE SUCCESS OF FIRMS IN DIFFERENT SECTORS THROUGH DATA MINING"** completed by **MÜGE AVŞAR** under supervision of **PROF. DR. GÜÇKAN YAPAR** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

.....

Prof. Dr. Güçkan YAPAR

Supervisor

.....

Doç. Dr. Deniz ÜNAL

(Jury Member)

.....

Dr. Öğr. Üyesi İdil Yavuz

(Jury Member)

Prof.Dr. Latif SALUM

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Dr. Güçkan YAPAR for the continuous support of my thesis and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

My sincere thanks also goes to Dr. Hanife TAYLAN SELAMLAR, who provided me an advice to begin master and who gave me a motivation. And I thank to Ali Sabri TAYLAN for great helps to my thesis and his encouragements.

Besides my manager, I would like to thank Hüseyin Avni AKTULGA for his insightful comments and permission.

Last but not the least; I would like to thank Mesut AVŞAR and Hatice AVŞAR. They are my life. I thank my family for supporting me spiritually throughout writing this thesis and my my life in general. Besides, I would like to thank to Berker BAYAZIT for supporting, encouragement and faithy.

Müge AVŞAR

DETERMINATION OF THE KEY FINANCIAL RATIOS IN THE SUCCESS OF FIRMS IN DIFFERENT SECTORS THROUGH DATA MINING

ABSTRACT

The aim of the study is to determine the significant financial ratios for the companies operating in different sectors in the year 2016. Within the scope of the research, the financial ratios of the companies operating in the retail and manufacturing sectors, which were traded at the BIST in 2016, were calculated and a mix model was created to help calculate the credit score by using both traditional and new models in determining the success and failure of the firms. In the estimation of the financial situation of the firms, discriminant analysis from the traditional methods is used and the z-score model developed by Altman is taken into consideration. Decision trees are applied from the new models for determination of the effects of rates on financial condition. As a result of this study, the financial ratios in the sectors have revealed the importance of determining the financial situation of firms.

Keywords: Altman z-score, data mining, financial ratios

FARKLI SEKTÖRLERDEKİ FİRMALARIN BAŞARI TAHMİNİNDE VERİ MADENCİLİĞİ METOTLARI KULLANILARAK ANAHTAR FİNANSAL ORANLARIN TESPİTİ

ÖZ

Çalışmada, 2016 döneminde BİST de işlem gören, değişik sektörlerde faaliyet gösteren firmalar için anlamlı finansal oranların tespiti amaçlanmıştır. Araştırma kapsamında, 2016 yılında BİST'de işlem gören, toptan ve imalat sektörlerinde faaliyet gösteren firmaların mali oranları hesaplanmış ve firmaların başarı ve başarısızlığının belirlenmesinde hem geleneksel hem de yeni modeller kullanılarak kredi skorlamasına yardımcı olacak karışık bir model yaratılmıştır. Firmaların finansal durumlarının tahmininde geleneksel yöntemlerden discriminant analizi kullanılmış olup Altman'ın geliştirdiği z-skor modeli dikkate alınmıştır. Oranların finansal durum tespitindeki etkilerinin belirlenmesinde ise yeni modellerden karar ağaçları uygulanmıştır. Bu çalışma sonucunda sektörlerdeki finansal oranların, firmaların finansal durumlarının tespitindeki önemi ortaya konmuştur.

Anahtar Kelimeler: Altman z-skor, veri madenciliği, finansal oranlar

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

The speed and integration of information has reached astonishing levels, in today's globalizing and rapid technological developments. Therefore, the response to the information is very fast and frequent. Fast information sharing and instant responses to information constitute the main reason for volatility in financial markets. So, markets are responding to an event that is likely to happen anywhere in the world instantly. The main reason for the reaction is that investors want to protect themselves from the consequences of the expected or actual events. However, the basis of the risk phenomenon constitutes this sense of protection. In this context, risk can be defined as events that develop beyond expectations and result in positive or negative situations. Financial risk is factors that affect the likelihood that an investment will be made available or possibility of positive and negative results to the financial liabilities.

Businesses are living beings with unique dynamics like humans. Socio-economic, political, structural changes where businesses operate constitute a financial risk factor in the markets. The correct measurement of risk contributes to the correct decision making of the risk structure and of the strategies for hedging. Companies takes more risk than occur, because measure the risk wrongly. If the expected risk occurs, an error in this way results in unexpected losses. In this case, investors, company managers and the markets have negative effects. In other case, the more measurement of risk, the more risk protection action is taken by companies. So, the abandonment of expected potential gains creates a cost increase.

The risk protection policy is important not only for businesses, investors, and managers, but also for the banks that provide the financing of the enterprises at the same time and make profit from this situation. As is the case with many other fields, the effects of globalization in the banking sector are increasing rapidly, and it is observed that the competition power is increasing in the periods. The effects of increasing competitive power and practices between banks and strategically-based

power distributions and balances are changing. For this reason, especially large banks are aware of the added value created by risk management to prevent possible losses by taking the necessary precautions against the risks that are carried out. Credit risk is at the forefront of the most important risks assumed by commercial banks, and it is not possible for banks to undertake banking activities without undertaking the risk of credit and managing it. In order to manage credit risks, it is necessary to define measure, apply, evaluate and watch risks. In addition to conducting a profitable credit policy to calculate the creditworthiness of the companies to which the banks lend or give credit, it is also necessary to make use of the financial statements to determine the financial structures and the solvency of the applicants when examining loan claims. In this context, certain scorings and risks are becoming necessary.

For the estimation of the risk factor and the early detection of the risk, it is necessary to examine and analyze the financial structures of the companies and it is possible to utilize a large number of indicators showing the financial status. However, the main argument is to determine the variables that are most significant in predicting the risk from a large number of ratios. Furthermore, it should not be overlooked that the financial structures of companies operating in sectors with different dynamics should be different from each other. In this respect, it is clear that the significance levels of the ratios in determining sectoral risks will vary. Estimation of the ratio to be made by considering the sectoral differences and risk estimation are taken as the main scope. For this reason, it is necessary to determine the significant rates as pioneer and early warning signals that will help to measure the risk.

The economic dynamics of the countries changes brought by globalization cause changes in the financial situations of firms. Financial resources are needed for a firm falling into distress, especially by banks in Turkey. In this direction, the credit risk of the companies that come to the credit demand of banks is required to be healthy. Because, if the loans given to the firm are seen as borrowing, the prospect of following their return can be understood. For this reason, the banks implement specific credit risk policies by making separate regulations. For calculation of credit

risk, some mathematical and statistical calculations were needed and scoring was developed. With the analysis made by the scoring, the risk situation of the firms is classified and credit is provided according to the risk group they are in. When it is thought that the purpose of the establishment and existence of the banks is to profit by giving credit, it seems that the healthy process of the credit processes given to the firm is important.

Credit scoring process is statistical calculations made on balance sheets and income tables of firms. Numerous studies have been carried out daily from the past, these studies include traditional models such as logit models, logistic regression, discriminant analysis; and new models such as decision trees, association rules, and neural networks. Our aim in this study is to bring a new perspective to credit scoring by doing a mixed modeling with both traditional and new models. In this process, discriminant analysis was applied in predicting financial failure. Decision trees have been applied for the classification when the rates effective in predicting failure have been determined.

In the current period, there is BIST, which is available for the companies' financial statements. When the company received the assets and equity structure are taken into account, the financial data of companies with significant market power in Turkey can be achieved is remarkable. However, the random feature on the stock market does not make it difficult to identify the risk with a limited number of functions, and does not respond much to the need for risk measurement. For this reason, the nonparametric model which perceives the change in the prediction of risk and which is able to renew itself will respond more to the need than a low-level model with many factors. The best solution to this problem is considered data mining. Data mining has become the most popular approach with the expansion of today's technologies and it is used in many sectors, especially in banking and finance. Data mining manages computer, machine learning, database or data warehouse management using mathematical algorithms and statistical techniques in the process of generating previously unknown meaningful data.

Using data from decision trees is the most plausible result when considering the fact that ratios are the main factor in determining the risk factors in data mining which have many fields and method. Decision trees are estimators, tree-like structures. Every branch in the tree is a classification question, and the leaves are a part of the data set. It is very useful technique to create understandable models since it allows easy rule removal of the tree (Berson et al., 1999). The CHAID algorithm, a special case of decision trees, was developed by Kaas. In the target variable mapping pairs where the calculation of the best partition is not statistically significant, the predictor variable is formed by combining possible pairs of categories. The chi-square test is used to select the most appropriate divisions in CHAID algorithm. The predictor variables are combined with a statistically significant difference in a pair that matches the target variable for calculate the best profile. As a result, the variable which having the highest F value among the independent variables with statistically significant effect on the dependent variable is in the first place in the CHAID tree. And so the level of significance is determined by the profiling method for the estimating variables when the target is changed.

Manufacturing and retail sectors were taken into consideration, during the period of 2016 in BIST. 33 retail and 158 manufacturing companies were used whose financial statements were supplied in the relevant period. 44 ratios that are considered to be important in predicting the financial significance are calculated from the balance sheet and income statements for all of the companies. These ratios were used as predictor variables of CHAID analysis application of decision tree methods to achieve target variance. Estimates of good and bad financial conditions of firms are taken as target variables. Making a clear assessment of the financial status of the companies is a difficult process and it is impossible to include a judgment of precision. The financial situation of the companies varies with market conditions, social and economic factors. For this reason, only estimation can be done. In order to make a healthy forecast, it is necessary to make comparisons for healthy firms with firms that declared bankruptcy in the same periods, same financial size and same market conditions. In this context, Altman(1968) has set a rate for the determination of the financial failure situation in 33 studies in the American stock exchange that

were traded in the manufacturing sector and 33 studies on healthy financial firms. In the study, 66 firms used the multi-discriminant analysis for 20 year balance sheet and income tables to determine the rates that determine the financial status of firms. Subsequently, the model has been revised many times and has been applied in various sectors. However, the determination of the validity of Turkey's market of Altman's method is deemed necessary because of the changing market conditions, and dynamics of countries. In this context, many studies are made in many different estimation methods. Altman's period of validity is proven to sustain the market in Turkey. Kulalı (2016) used 19 companies from the BIST between 2000 and 2013 in his study. After the calculations, it has been revealed that the z-score model estimated 95% before one year and 90% before two years ago. Considering the studies carried out in this context, the validity of the Altman model is accepted and Altman's financial condition determination model is used to estimate the target variable. The financial situation is estimated with z-score values and the target variable is assigned.

CHAPTER TWO
MANAGEMENT OF CREDIT RISK

Before talking about credit rating and credit scoring definitions, it is necessary to explain what the risk is. For this, risk definitions should be explained on the companies and risk elements should be determined. Status of the financial situation should be decided in order to identify companies' risks. For this reason, financial distress should be defined first. After that, it is better to make comments on credit scoring and rating for firms located in financial distress. Credit scoring and credit rating are the most important, same reasoning, but different processes in measuring credit risk.

2.1 Financial Distress and Measuring the Risk

Companies are living assets. Every living thing has a healthy life, sickness and death. There are cases where the business is successful, unsuccessful, bankrupt and close, such as being alive. Financial failure has been a subject of discussion and does not contain definite judgment for hundreds of years. Many scientists have studied the possibilities of failure from the past to the present. The definition of financial failure causes different studies to be made relative to different judgments. Firstly, it is necessary to define the failure of a business, to be able to make these judgments. Failures of businesses can be defined as having insufficient resources to pay their debts, bankruptcy, as they may not be able to pay their debts at their expiration dates or pay their debts at all. Bankruptcy means that the operation of the business is stopped, it is a legal process. The bankruptcy of a company in terms of investors and company executives are undesirable. Therefore it is important to develop an early warning system in terms of bankruptcy have come. Bankruptcy is considered as the last digit of financial failure detection by the failure of the company would be considered an early warning system for bankruptcy. Companies failing states can be divided into economic and financial failure.

Economic failure can be defined as costs incurred by a company in carrying out its activities can not be covered by income earned as a result of its activities. Companies can be called economically successful that can generate income from their activities.

Financial failure in general is the inability of companies to meet their financial obligations. Financial failure occurs when the company's debts are delayed or their sources of debt are inadequate. The final state of financial failure is the bankruptcy process, which is the beginning of a legal process.

2.2 Sector Outlook In Terms of Risk

Companies with different activity areas continue their activities with different dynamics. This emphasizes that different ratios are important when firms are analyzed. For example, firms operating in the manufacturing sector perform regular inventory production, while stock volumes and inventory turnover periods are important variables in the interpretation of asset structures. On the other hand, the period of stock transfer for the companies operating in construction sector is insignificant because of the project-based work and the determination of the working period on the date of the work, the absence of stock requirement and the absence of a certain production process. Therefore, it is necessary to first decide which sectors to examine. The financial statements of the companies are not easy to find. Therefore, it has been decided to use the balance sheets and income statements published in Istanbul in the scope of the study. Sectoral distinctions have been made for companies in BIST. It is determined that the highest quality results will be achieved in the manufacturing and retail sectors when the number of firms and healthy financial data are taken into consideration. Therefore, it is important to evaluate the change in Turkey as the years studied sectors.

2.2.1 Retail Sector

The retail sector is distinguished from other sectors by its unique structure, whose dynamics have similar characteristics all over the world. Customer demands and customer behaviors, which are the two most influential forces directing the sector, are transformed into globall behavior in a short period of time by becoming independent of the geography and demography they emerge. Firms operating in the retail sector are struggling with very similar effects and problems created by their customers' demands and behaviors. The main problem in the sector around the world is that customer behaviors and journeys have changed. At the point reached in communication channels, customers can access all kinds of information and all kinds of shopping using many channels. Online shopping, then rising mobile shopping and social shopping, which is frequently mentioned in the present, greatly influence the effects of traditional physical merchandising.

Retail sales worldwide are estimated at $ 23.445 billion at the end of 2017, including e-commerce. Every year, the e-commerce industry plays about 1 percent of the traditional retail sector. Retail e-commerce sales will reach $ 2.350 billion worldwide by the end of 2017. This is 10.1 percent of total retail sales. At this point, it should be noted that the share of global e-commerce in total retail sales was 3.6 percent in 2011 and rose to 8.7 percent in 2016. Global e-commerce sales are expected to increase by 20 percent annually on average annually and reach $ 4.048 billion by 2021. Total retail sales are showing Figure 2.1 and driving force of the digitalizing world economy are showing Figure 2.2 between the years of 2015-2020.

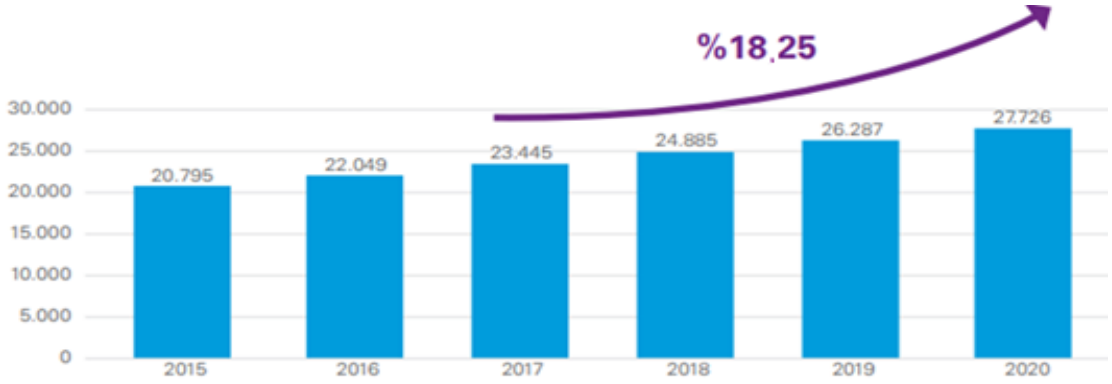


Figure 2.1 Total retail sales worldwide, 2015-2020 (billion USD) (KPMG, 2017)

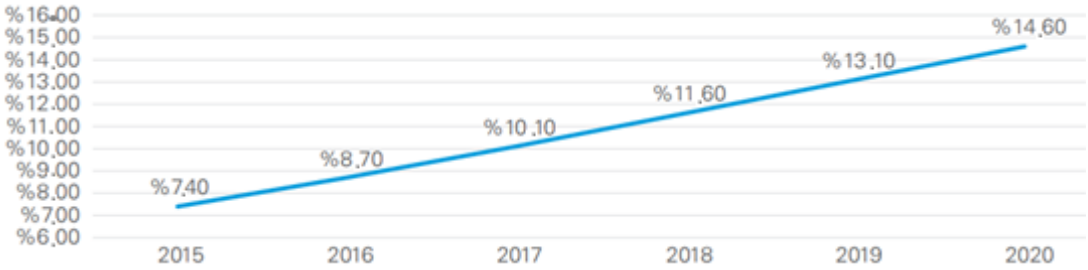


Figure 2.2 Driving force of the digitalizing world economy: E-Trading (TÜSİAD, 2017)

2.2.2 Manufacturing Sector

The world has entered the era of the rise of unmanned intelligent factories that can manage and manage themselves with the developing technology. In factories with self-regulated digital intelligence, the fact that large data is analyzed correctly

Systems form the basis of the manufacturing industry where objects communicate with each other and with people. Estimates of industrial production growth rates in selected countries are in Figure 2.3.

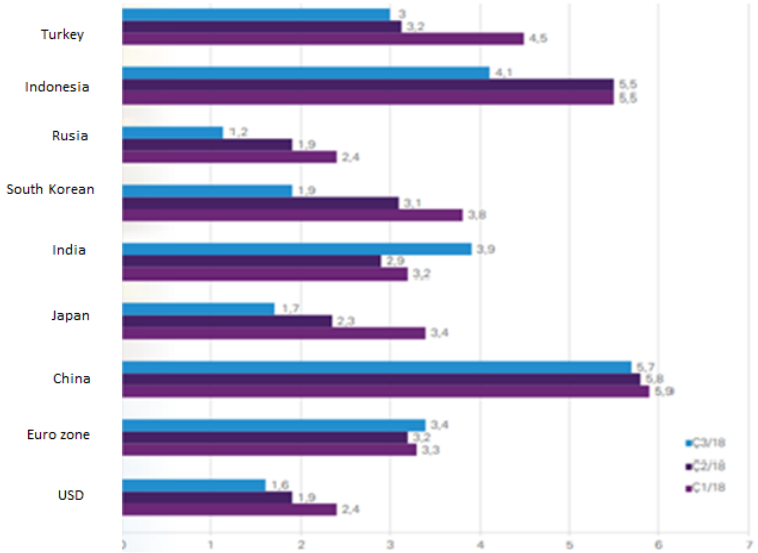


Figure 2.3 Estimates of industrial production growth rates in selected countries (TÜSİAD, 2017)

2.3 Credit Rating

Credit rating can be defined as opinions created by credit rating agencies on the credit risk of borrowers in general. A definition should be made by introducing the concept of risk, which is related to credit rating. In summary, Rating is defined as an assessment that raises the risk that the investor will undertake. Credit ratings can be made by private agencies; banks can also do to customers to prevent their losses. The purpose of credit rating calculations is to improve the early warning system to reduce the likelihood of defaults for banks. Summarizing the credit rating;

- Credit Rating is an independent view of a borrower's debt repayment request and capacity.
- It is not equal to the likelihood of a failure to produce a result of statistical methods.
- Rating provides transparency to the markets, facilitating the measurement and management of credit risk and risk-based pricing.

2.4 Credit Scoring

Banks use statistical / mathematical models based on various qualitative and quantitative data sets to measure credit risk. As a result of these analyzes, a certain grade is formed and this grade is used in the credit decision. Summarizing the credit score;

- They are mathematical models that measure credit default risk
- Mathematical probabilities can be converted to calibrated notes.
- Scoring provides transparency to the banking system, facilitating the measurement and management of credit risk and risk-based pricing.

2.4.1 Historical Background of Credit Scoring

One of the pioneers of modern financial failure prediction is Beaver (1966). Beaver applied univariate statistical model in his study and aimed to determine the financial ratios of 79 firms that have failed in the past five years and 30 financial

ratios belonging to the pre-default period. As a result of the study, the most important ratio for financial failure for Beaver was determined as "cash flow / debt sum" (Sayılgan, 2008).

However, Beaver's work has undergone a lot of criticism from different angles. The one-dimensional nature of the model leads to a high correlation between the data used (Outecheva, 2007).

Altman (1968)'s failure model emerged as the development of Beaver's work. Altman z-score model is one of the most common models that use discriminant analysis for financial failure detection. Primarily Altman has identified 22 financial ratios in 33 bankruptcy firms and 33 financially healthy firms. As a result of the statistical analysis, the model was developed by reducing the number of variables to be used in the model to 5 (Miller, 2009).

Altman has used multiple discriminant statictical methodology instead of traditional methods in corporate bankruptcy estimation. The data used in the study are taken from manufacturing companies.

In general, the ratios that measure profitability, liquidity and debt repayment are the most important indicators, and in each study made, a different ratio is determined as the most effective indicator of the oncoming problems in traditonal studies. It is far from showing the importance of rates. Although previous studies have formed some important generalizations about the performance and trends of certain metrics, they both become theoretically and practically questionable to assess firms' bankruptcy potential. At the beginning of these discussions is the definition of financial performance. Despite emphasizing the approaching problems with the single variable methods employed, the ratio analysis leads to erroneous interpretations. For example, a firm with insufficient profitability and / or paying power may be considered a potential bankruptcy, but it may not be required to be considered bankrupt with the liquidity ratio. For this reason, Altman aimed to obtain estimation models that yielded meaningful results instead of univariate ratio

analyzes. But in this case, the question becomes how to determine which percentages are most important without determining the potential for bankruptcy, what weights should be added to selected ratios, and how weights can be established objectively.

2.4.2 Types of Credit Scoring

Certain financial analysis techniques need to be applied in order for credit scoring to be done.Financial analysis techniques derived from the financial statements used in determining the financial condition of the company. After studies with various analytical techniques, investors try to decide on the investment decision, and business owners and managers decide on future plans. Financial analysis techniques are performed with ratios derived from the financial data of the companies. The information obtained with the help of the financial ratios clearly shows the status of firms and allows comparison with other firms on the market. In fact, financial analysis techniques aim to detect financial distress by comparing it with a single financial ratio or a combination of ratios of multiple financial ratios. With the approach of making use of financial ratios, estimated shortage of accounting-based approach is its simplicity and easy availability of financial information has led to gain widespread use.

A wide variety of models for the prediction of financial failure, such as artificial neural networks and multivariate statistical models, multiple regression models are used. The one-dimensional models formed by using a ratio are easier to use than the multidimensional models formed by using multiple ratios, but are less preferred because of the low estimation power. Numerous statistical methods are used in multidimensional models, such as multiple regression, discriminant analysis and logit modeller.

2.4.3 Multiple Discriminant Analysis (MDA) Usage in Credit Scoring

MDA was first used in the 1930s and is not as popular as regression analysis (Fisher, 1936). In previous years, MDA has been used in biological and behavioral

sciences and has been applied by Cochran (1964). Later it was used in finance by Durand (1941). In the Durand study, assessed the credit worthiness of the used car loan application. Myers and Forgy (1963) analyzed various techniques, including MDA, in evaluating good and bad installment loans. Myers and Forgy analyzed several techniques, including MDA, in the evaluation of good and bad installment loans. In the upcoming years, Walter (1959) used an MDA model to categorize companies with high and low price-earnings ratios. Smith (1965) applied the technique in the classification of companies into standard investment categories.

MDA is a statistical technique used to classify an observation into one of several priori groups, depending on individual characteristics. Firstly, it is used to make classification and / or prediction in problems where the dependent variable emerges in a qualitative form. For this reason, it is the first step to establish explicit groupings of which the number of original groups can be two or more.

After the groups are created, data is collected for the objects within the groups. Then MDA tries to achieve a linear combination of these properties that makes the "best" discriminates between groups. For example, if an institution has measurable characteristics (financial ratios) for all companies involved in the analysis, the MDA sets a number of discriminant coefficients. When these coefficients are applied to actual ratio, there is a basis for classifying one from another in mutually exclusive groups. The MDA technique has the advantage of taking into account all the common features of the respective firms and the interaction of these properties. While a univariate study can only assess the measures used for group assignments one at a time, the MDA technique has the advantage of taking into account all of the common features of the respective firms and their interaction.

A feature in Bryan (1951) and Rao's (1952) work can be among other advantages of MDA, this property says that MDA is the reduction of the analyst's space dimensionality, i.e., from the number of different independent variables to G - 1 dimensions, where G equals the number of original a priori groups. But the Altman

operation takes place in a one dimension that includes bankrupt and non-bankrupt firms. The discriminant function of the form

$$Z = x_1 v_1 + x_2 v_2 + ... + x_n v_n \quad (2.1)$$

transforms individual variable values to a single discriminant score or z value which is then used to classify the object

where $v_1, v_2, ...v_n$ = Discriminant coefficients,

$x_1, x_2, ...x_n$ =Independent variables.

The MDA computes the discriminant coefficients, v_j, while the independent variables x_j are the actual values

where j=1, 2,... n.

2.4.4 Altman's Credit Scoring Model

In Altman's work, he used 66 executive statements, bankrupt and non-bankrupt. The bankrupt group consists of firms operating in the manufacturing sector between 1946 and 1965 and under the National Bankruptcy Act. Selected companies have $6.4 million mean asset size, between $0.7 million and $25.9 million range. It is important to note that this group is not homogeneous due to the industry and size differences. On the other hand, second group has $1-$25 million asset size range, which consisted of a paired sample of manufacturing firms chosen on a stratified random basis. After the determination of the groups, the necessary rate calculations were made from the balance sheets and income tables of the companies. The ratios used in the study were selected from the ratios found to be significant in previous studies and relevant to the study.

Through the ratios, five ratios are selected as doing the best overall job together in the prediction of corporate bankruptcy with the MDA computer program which was developed by W. Cooley and P. L. In order to arrive at a final profile of variables the following procedures are applied:

(1) Observation of the statistical significance of various alternative functions including determination of the relative contributions of each independent variable;

(2) Evaluation of inter-correlations between the relevant variables;

(3) Observation of the predictive accuracy of the various profiles;

(4) Judgment of the analyst.

After many iterations, the final discriminant function is as follows:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.06X_4 + X_5 \quad (2.2)$$

Where X_1 = Working capital/Total assets

X_2 = Retained Earnings/Total assets

X_3 = Earnings before interest and taxes/Total assets

X_4 = Market value equity/Book value of total debt

X_5 = Sales/Total assets

Z = Overall Index

X_1 : Working capital/Total assets: It is a measure of the net liquid assets according to the firm's total capitalization. Working capital is defined as the difference between current assets and current liabilities. Liquidity and size characteristics are clearly accepted. A living company with consistent business losses is expected to shrink existing assets in relation to total assets. The current ratio, quick ratio and working capital / total assets were used as the liquidity ratios in the study. The working capital / total assets were found to be the most significant variable in liquidity ratios.

X_2 : Retained earnings/Total assets: Retained earnings are the ratio that reports the total amount of reinvested earnings or losses of a firm over its entire life. It can be known that earned surplus. Retained earnings can be manipulated through corporate restructurings and stock dividend notices. The age of the firms are effective in calculating this rate. For example, it is expected that the RE / TA ratio will be low because new firms will have fewer profits. For this reason, it can be said that this ratio is more meaningful in the old firms.

X_3 : Earnings before interest and taxes/Total assets: This ratio is calculated by dividing the total assets of a firm into its earnings before interest and tax reductions. This rate allows the firm to calculate the true productivity independent of the tax and leverage factor. The net assets of the companies are the earnings power of their assets. Bankruptcy occurs when total liabilities exceed the value determined by the earnings power.

X_4 : Market value of equity/Book value of total debt Equity is measured by the combined market value of all shares of stock which is preferred and common, while debt includes both current and long-term. This ratio shows how much the assets of the company are depreciated before the liabilities exceed their assets and the firm goes bankrupt. This ratio shows how much the assets of the company are depreciated before the liabilities exceed their assets and the firm goes bankrupt. This ratio adds a market value dimension which other failure studies did not consider. It also appears to be a more effective predictor of bankruptcy than a similar, more commonly used ratio: Net worth/total debt (book values).

X_5 : Sales/Total assets The capital-turnover ratio shows that illustrating the sales generating ability of the firm's assets. It is an important factor in terms of competition. Despite the least significant ratio in the study done, it was considered the most important second rate due to its contribution to the discrimination ability of the model and its unique relationship with the other variables in the model.

The value obtained as a result of the values included in the equation, called Altman z-score, is compared with the limit value determined by Altman. If the z-score value is below 1.81, the company is in compliance with the bankruptcy profile. If the value of z-score exceeds 2.99, the firm has no financial difficulties. While the z-score value is among these two values, it is difficult to determine the success profile of the company although it is a grey area for Altman. When he wanted to determine the border within the grey area, Altman stated that 2.67 was the border value and that it separated the success and failure of the firms.

In an environment where traditional rate analysis is inadequate, Altman's discriminant analysis technique has been introduced in the determination of bankruptcy of firms. In the study, the model was proved extremely accurate in predicting bankruptcy correctly in 94 percent of the initial sample with 95 per cent of all firms in the bankrupt and non-bankrupt groups assigned to their actual group classification. In addition, the validity of the discriminant function is proven in two of the applied examples to test the reliability of the model. Investigation of the individual ratio movements prior to bankruptcy corroborated the model's findings that bankruptcy can be accurately predicted up to two years prior to actual failure with the accuracy diminishing rapidly after the second year.

Altman has developed a model after criticism of how to determine the market value of stocks in non-public companies. In the new model, X_4 the value is calculated as follows.

$$Z = 0.717X_1 + 0.847X_2 + 3.107X_3 + 0.420X_4 + 0.998X_5 \qquad (2.3)$$

where

X_1 = Working capital/Total assets

X_2 = Retained earnings/Total assets

X_3 = Earnings before interest and taxes/Total assets

X_4 = Book value equity/Book value of total debt

X_5 = Sales/Total assets

Z = Overall Index

The equation now looks different than the earlier model; for instance, the coefficient for X_1 went from 1.2 to 0.7. But, the model looks similar to the one using Market Values. The actual variable that was modified,X_4 , showed a coefficient change to 0.42 from 0.6001; that is, it now has less of an impact on the z-score. X_3 and X_5 are virtually unchanged. The univariate F-test for the book value of X_4 (25.8) is lower than the 33.3 level for the market value but the scaled vector results show that the revised book value measure is still the third most important contributor. This model is called Zeta (Altman, 2000).

The Z value of the new model is

Z'<1.21 = Zone I (no errors in bankruptcy classification)

Z'>2.99 = Zone II (no errors in nonbankruptcy classification)

grey area = 1.23 to 2.99.

Altman has made some changes in the formula for bankruptcy of service sector companies not listed in the manufacturing sector. The model developed for these companies is as follows:

$$Z^{"} = 6.56X_1 + 3.26X_2 + 6.72X_3 + 1.05X_4 \quad (2.4)$$

All of the coefficients for variables X1 to X4 are changed as are the group means and cutoff scores. This particular model is also useful within an industry where the type of financing of assets differs greatly among firms and important adjustments, like lease capitalization, are not made. In the emerging market model, we added a constant term of +3.25 so as to standardize the scores with a score of zero (0) equated to a D (default) rated bond.

In this study, the first z-score model developed by Altman was used. The second model (ZETA) developed to evaluate the financial situation of non-public companies and the latest model study developed to cover sectors other than the manufacturing sector have not been examined.

Determination of the validity of Turkey's market Altman's method is deemed necessary because of the changing market conditions, and dynamics of countries. In this context, many studies are made in many different estimation methods Altman's period of validity is proven to sustain the market in Turkey. Kulalı (2016) used 19 companies from the BIST between 2000 and 2013 in his study. After the calculations, it has been revealed that the z-score model estimated 95% before one year and 90% before two years ago. Considering the studies carried out in this context, the validity of the Altman model is accepted and Altman's financial condition determination model is used to estimate the target variable. The financial situation is estimated with z-score values and the target variable is assigned.

CHAPTER THREE
THE USE OF DATA MINING TECHNIQUES IN CREDIT SCORING

3.1 Knowledge Discovery in Database (KDD)

Data mining, knowledge discovery and knowledge discovery in databases are mixed by some researchers. Many researchers and practitioners use the concepts of data mining and information discovery in a similar way. However, Data Mining is a phase of the information discovery process. If we briefly define the discovery of information in the database, it is the process of revealing patterns that are meaningful, useful, and original that have a certain value in the data (Cios et al., 2007).

The data is based on manual analysis and interpretation of the traditional method of computing. For example; in the health industry, after the experts periodically analyze the changes in the health care datas and elaborate the analysis on the health care institution, these past datas provide the basis for future decision-making and health management. Analyzes of such future estimates are used in a wide variety of industries such as science, marketting, finance, health care, retail, or any other field. With the increase of information stacks over the years, it has become compulsory to automate the analytical subject datas.

The ability to perform each field analysis leads to the need to deal with a large number of bytes. After analyzes made, productivity increase for the enterprises, better service to the customers, competitive advantage is provided. We use to build the theories and models of the universe what we live in is the basic evidence. Computers have facilitated computing techniques in this process and have stored more datas than we can store. It also contributed to significant results from stored large volumes of datas. For this reason, KDD is an attempt to solve a problem that the age of digital information creates a life reality for all of us. In summary, KDD

process was defined by Fayyad (1996) as “the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.”

The details of the process are given below;

***1-Developing an understanding of the application domain*:** This is the first step. The target of the step is, KDD process is defined by the point of view of the customer. At this stage, it is tried to understand what to do with many decision options (transformations, algorithms, representation, etc.). So, responsible for the data mining project should understand and define the goals of the end user and the environment in which the information discovery process will take place (including relevant preliminary information). As the process progresses, this step can also be done during revisions and adjustments. After the targets are understood, the preliminary processing of the data starts as described in the next steps.

2-Creating a target data set: After the objectives have been determined, the data to be used for the discovery of the information should be decided. This step involves integrating into a single set of data, including what to look for, what additional data is needed, and what to consider handling all data related to information discovery. This process is important because data mining is very important as it learns and explores new models from existing data. It is the basis of model building process. For a successful process, as many features as possible should be considered at this stage. However, the collection, organization and operation of complex data repositories are expensive and challenging process.

3-Data cleaning and preprocessing: This step is the step of data security. Data cleaning features such as handling missing values, removing noise or outliers are done at this stage. Complex statistical methods and specific data mining algorithms can be used to clean up. For example, if you suspect that a particular feature is not reliable enough or too many missing bits, this feature are implemented using a data mining control algorithm. With these methods, a prediction model can be developed and the missing value can be estimated and changed. The importance that people give to this process due to many factors can change. But, whatever the level of

importance, it is important to examine these aspects of the data and to understand the general corporate information systems.

4-Data reduction and projection: In this step, better data production is prepared and improved for the data mining. One of the methods that can be used here is size reduction, such as feature selection, subtraction, and registration sampling. Another method that can be used in this step is the transformation of qualities such as decomposition of numerical qualities and functional transformations. This step can be specific to the project and can be considered the most important step for the success of the project. For example, in medical examinations, individual characteristics that make a difference often consist of features considered as the most important factors. In marketing, which is another sector, it is necessary to investigate the effect of advertising accumulation and external influences on the detection of marketing success, adding temporal issues to account. Even if the correct conversion is not used at the beginning, it can have a surprising effect, giving hints about the conversion needed in the model. After completing the four basic steps described above, the following step is applied. The steps described below depend on the data mining sections that focus on the algorithmic directions used for each project.

5-Choosing the appropriate data mining task: This step matches the targets of the KDD process (step 1) to a particular data mining method. It is decided that data mining will be most appropriate for the needs of the task, such as, classification, regression or clustering. This usually depends on the goals and previous steps. Data mining has two objectives: prediction and explanation. Data mining based on estimates is generally referred to as supervised data mining, often involving the classification and visualization aspects of data mining. Most data mining techniques are based on inductive learning where a model is created explicitly or indirectly by generalizing a sufficient number of training examples. The inductive approach ensures that the trained model can be applied in the future. The strategy of the approach also takes into account the level of meta-learning for a given data set.

6-Exploratory analysis and hypothesis selection: In this step the specific method to be used for search models is selected (including multiple inducers). For example, given sensitivity to clarity, this model works better with older neural networks, while others work better with decision trees. For each meta-learning strategy, there are several possibilities how to achieve it. Meta learning focuses on explaining the factors that are effective in the success or failure of a data mining algorithm. So, this approach tries to find the conditions in which a data mining algorithm is most appropriate.

7- Employing the data-mining algorithm: After the selection and implementation of the data-mining algorithm, the algorithm is used several times until a satisfactory result is obtained in this step. For example, we must use the algorithm several times until a satisfactory result is obtained by adjusting the control parameters of the algorithm such as the minimum number of samples for a single leaf of a decision tree.

8-Evaluation: At this stage, mined patterns (rules, reliability, etc.) are evaluated and interpreted in relation to the objectives defined in the first step. In this step, the preprocessing steps are dealt with regarding the effects of the Data Mining algorithm on the results. (For example, by adding features from step 4 and repeating from that step). This step focuses on the intelligibility and usability of the applied model. In this step, the discovered information is documented for further use.

9- Using the discovered knowledge: After the above steps have been followed, the information at the latest stage is ready to be transferred to another system. Changes can be made to the information system after this process. This step determines the overall success of the process. In this step, there are many difficulties such as losing the "laboratory conditions" in which we operate. For example; information is obtained from the snapshot of the information, but now the information has become dynamic. Data structures and data fields may change (Fayyad, 1996). An overview of the steps that compose the KDD process is shown in Figure 3.1.

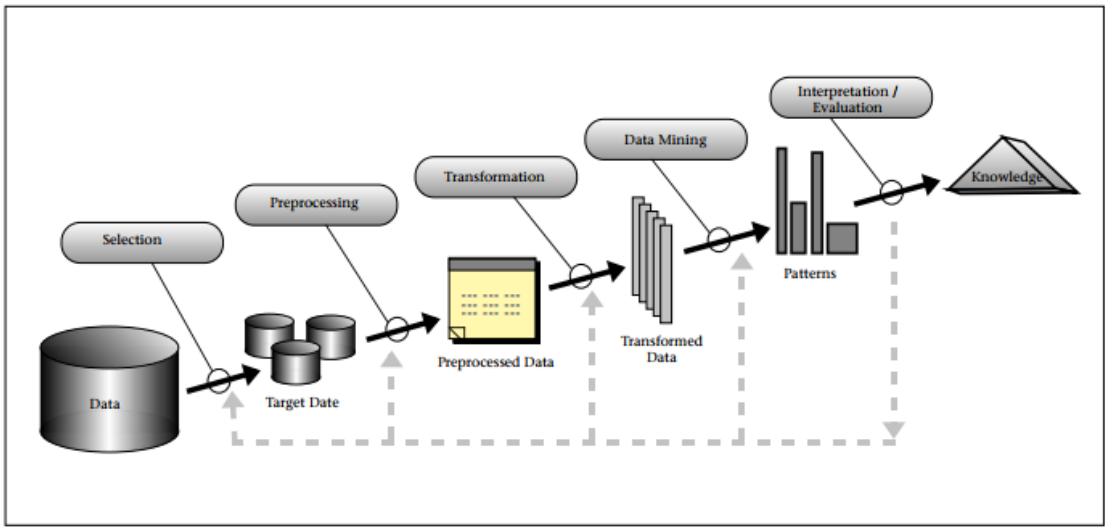


Figure 3.1 An overview of the steps that compose the KDD process (Fayyad, 1996)

3.2 Definition of Data Mining

The data is unprocessed raw material to obtain the information and is worthless alone. The data is processed for a specific purpose with computer systems and the information is obtained. It is important to reveal valuable, usable information that is hidden from large amounts of data due to the increase in data volume nowadays. Data mining is a collection of techniques used to uncover this information. This process is carried out by using machine learning, database or data warehouse management, mathematical algorithms and statistical techniques. Data mining is a process that uses many analysis tools to explore patterns and relationships for making valid predictions within the data. The purpose of data mining is to create decision-making models that based on analysis of past activities, for forecasting future behavior. Data mining supports the definition of information discovery '... disclosure of non-trivial confidential, previously unknown and potentially useful information in the data...' by William Frawley and Gregory Piatetsky-Shapiro (1992) (Berson et al., 1999).

Statisticians used to try to find meaningful and important patterns from very large data sets. Data mining has become a powerful technology that produces meaningful information and patterns from large databases nowadays. Although this means that the basic logic of data mining is considered classical statistical techniques, it is not

true to express any statistical method or software as data mining. Data mining is known to have differences in classical statistical techniques. The following Table 3.1 is a list of details that Moss and Atre (2003) shared about the comparison of statistical analyses and data mining, its different points.

Table 3.1 The comparison of statistical analysis and data mining

Statistical Analysis	**Data Mining**
Statisticians usually start with a hypothesis	Data mining does not require a hypothesis
Statisticians have to develop their own equality in order to match their hypothesis.	Data mining improves algorithms and equations automatically.
Statistical data only uses numeric data.	Data mining uses different types of data
Statisticians find and filter dirty data during their analysis.	Data mining is based on clean data

3.3 Classification of Data Mining Methods

Data mining methods can be divided into two categories;

- Supervised
- Unsupervised

We can define supervised and unsupervised methods in general as follows.

Supervised: This expression is used when it is a well-defined or precise target. This method used to extract information and results from the data.

Unsupervised: This expression is used when a special definition for the desired result situations that are not done or in the case of uncertainty (Hastie, 2001). This method used to understand, discover and recognize data mostly, and they aim to give an idea of the methods to be applied in the future.

In machine learning, prediction methods are called supervised learning. Unsupervised learning models the distribution of samples in the typical, high-dimensional input space, and supervised learning is against it. According to Kohavi and Provost (1998), the term "unsupervised learning" refers to "learning techniques that group instances without a prespecified dependent attribute". Supervised methods are methods that try to discover the relationship between input attributes (independent variables) and a target attribute (dependent variables). The relationship between these variables forms a model. Often, they disclose a phenomenon that can be used to estimate the value of a target attribute if it is hidden in the model data set and the input attributes values are known. The main datamining methods are listed below as supervised and unsupervised;

Supervised:

- K-nearest-neighbor
- K-means clustering
- Regression models
- Rule induction
- Decision trees
- Neural networks

Unsupervised:

- Hierarchical clustering
- Self-organized maps (Hastie et al., 2001; Thearling, 2004)

3.4 Data Mining Methods

New methods and algorithms are added in addition to many methods used in data mining almost every day. Some of them are statistical methods which have been used for decades and defined as classical techniques. Other methods are new generation methods that based on statistic, supported with machine learning and artificial intelligence mostly.

These are the main classical methods used in data mining (Berson et al., 1999):

- Regression
- K-nearest-neighbor
- Clustering

This is the first of the new generation methods (Berson et al., 1999):

- Decision trees
- Association Rules
- Neural networks

Moreover, other methods of data mining are as follows (Chen, 2001):

- Basic Components Analysis
- Discrimination Analysis
- Factor Analysis
- Kohonen Networks
- Fuzzy Logic-Based Methods
- Genetic Algorithms
- Bayesian Networks
- Rough Set Theory Methods

In addition to the above mentioned methods, hybrid methods involving multiple techniques and time series methods are also used as data mining methods. Thus, every method for discovering information can be used as a data mining method (Kovalerchuk & Vityaev, 2002).

It is possible to examine data mining models according to their usage areas under three main headings; classification and regression models, clustering models, and association rules. Classification and regression models are predictive models, association models are descriptive models.

3.4.1 Classification and Regression

Classification, which is one of the most widely applied data mining techniques, leveraging the class with defined existing assets, is a data-mining model. Classification has a two-step. A model is created to use for prediction in step one. In the second step, the classes are estimated by applying this model to data that is not class specific (J.Han et al., 2001). It is used in estimating your future, using the available data.

Basic classification techniques are Artificial Neural Networks, Genetic Algorithms, K–Nearest Neighbor, Memory Based Reasoning, Naive – Bayes, Logistic Regression, Decision Trees. The decision tree method will be used in this study.

3.4.2 Decision Trees

Decision trees are used to classify an object or an instance in a predefined group based on attribute values. Decision trees are frequently used in applied areas such as finance, marketting, engineering and medicine.

While not attempting to modify existing conventional statistical methods, it includes many other techniques that can be used to classify or predict associations with a predefined set of classes, such as artificial neural networks and support vector machines.

In data mining, a decision tree is a prediction model that can be used to represent both classifiers and regression models. Decision trees express a hierarchical model of decisions and their results in an operational research. Decision trees are most commonly used among classification models. When used for regression analysis, it is defined as regression trees.

The decision tree is a predictable model that appears as a tree. Each branch of the tree is a classification question and leaves are part of the data set. It is a very useful technique to establish understandable models since the tree allows easy rule extraction. Decision tree technology is used for the discovery of data sets and business problems. This is usually done by looking at the estimators and their values in each section of the tree. These estimators can often provide usable content or suggest questions that need to be answered. If the tree continues to grow until it has a single record, it can be thought that many questions and branches will be created. The growth of the tree in this way is both expensive and unnecessary in terms of calculation (Berson et al., 1999).

Decision trees are one of the most common methods in classification problems in the tree view because of its ease of installation, interpretation and integration with databases (Yılmaz, 2009). In principle, the task of a decision tree algorithm is to divide the data recursively by branching into sub-data groups. Every new branch that occurs during this division process refers to a rule. Decision trees are basically two steps. The first step is to create the tree. The other step is to categorize the data by applying each record in the database to Figure 3.2.

In Decision Trees, each attribute is represented by a node. In this way, the nodes are expressed with the letter A and separated by the next branch on each node. In this separation process, a question is asked about node A and a branch is followed according to the answer given in the database. If a classification can be obtained as a result of the branch, the leaves indicated by C will be obtained. Each of the obtained leaves represents a class. Therefore, the top structure is called root, the last structure is called leaf and the remaining structures between them are called branches (Silahtaroğlu, 2008).

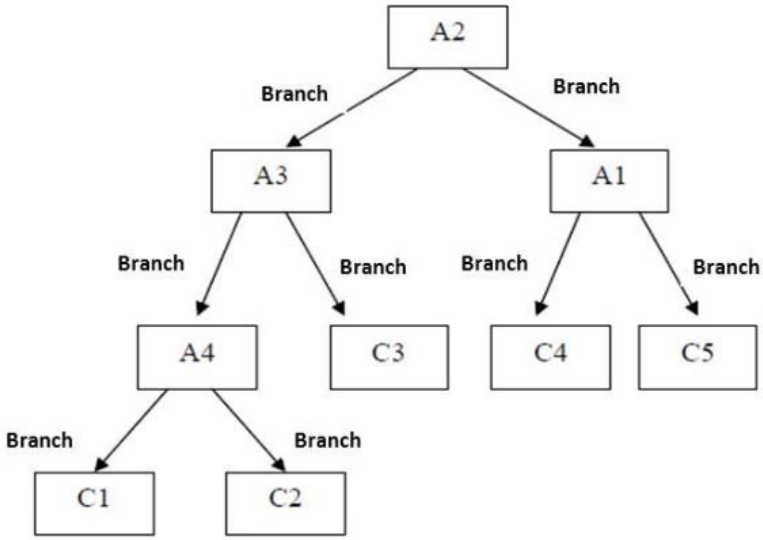


Figure 3.2 Decision trees

According to the algorithm to be applied in decision trees, the shape of the tree changes, different tree structures create different classifications. Different root node changes the path and classification to follow when reaching the end leaf. Because of this reason, it is very important that the root node is divided into branches. If the answer in the database is yes/no, it is asked to be separated into two branches; if the answer in the database is yes/no/maybe, it is asked to be separated into three branches. The concept of entropy is used to find the nature of branching in Decision Trees. Entropy is the quantification of data in the database and the measurement of uncertainty in the data. It gets a value from 0 to 1. For each variable, the uncertainty of the variables is measured by taking advantage of the entropy concept. In the next step, the gains are calculated for each variable and the branching is done from that variable, whichever is higher (Koyuncugil, 2007).

Strengths of decision trees methods can be listed as follows:

- Classification can be done with a very small calculation.
- It can be easily applied to complex data sets, as well as working with categorical and continuous variables.
- It can clearly indicate which variables are more important for classification and prediction.

- Its assumptions are limited because it is a nonparametric model.
- The model results can be easily interpreted without the need for much statistical information, since the relationship between dependent and independent variables are in the form of trees.
- Makes the most accurate classification possible by modeling all the independent variables associated with the identified dependent variable and all their combinations. Variable combinations also allow for the evaluation of all interactions.
- It is not affected by missing or extreme values for dependent and independent variables.

Decision tree analyses are widely used as follows:

- Identification of expected elements that are members of a specific class (Segmentation),
- Separation of cases into various categories such as high, moderate, low risk groups(Stratification),
- Establishing rules for estimating future events,
- Selection of a large number of variables and data sets to be used in the establishment of parametric models
- Identification of relationships that is unique only to certain subgroups,
- Consolidation of categories and conversion of continuous variables to decimals,

A few applications related to decision trees are as follows:

- Determining the variables that cause product errors by examining the production data,
- Identifying variables affecting sales,
- Determination of the most effective treatment methods by using medical observation data,
- Giving credit decisions by using the credit histories of individuals and firms (credit scoring),

- In marketing applications with letters, which demographic groups have high determination of the answer rate (direct mail)

There are different algorithms which developed to create a decision tree like CHAID (Chi-Squared Automatic Interaction Detector; G.V. Kaas; 1980), C&RT (Classification and Regression Trees; Breiman, Friedman, Olshen ve Stone; 1984), ID3 (Quinlan; 1986), Exhaustive CHAID (Biggs, de Ville ve Suen; 1991), C4.5 (Quinlan; 1993), MARS (Multivariate Adaptive Regression Splines; Friedman), QUEST (Quick, Unbiased, Efficient Statistical Tree; Loh ve Shih, 1997), C5.0 (Quinlan), SLIQ (Supervised Learning in Quest; Mehta, Agarwal ve Rissanen), SPRINT (Scalable Parallelizable Induction of Decision Trees; Shafer, Agrawal ve Mehta).

The most widely used algorithms are C&RT, C5.0 and CHAID. Details of the most commonly used algorithms are described below. In the study CHAID, C5.0 and C&RT algorithms have been studied and the best presentations are provided by the CHAID algorithm. The results from other algorithms have been added to the appendices.

3.4.2.1 C&RT Algorithm

It is processed by binary division based on gini. So that data in each sub-cluster is more homogeneous than in the previous cluster. The branching continues by renewing, so that it stops when the homogeneity measure for each branch is reached. The model uses certain impurity measures to measure homogeneous branches. There are 3 types of impurity measures used in the model. For the symbolic target fields, the gini and for the continuos targets, the least-squared deviation method is used. The same predictor field can be used in different levels of the network; so that missing data can be used in the best way by making surrogate splits (Breiman, 1984).

The C&RT tree steps are as follows:

Finding each predictor's best split: For each predictor field, the best possible splits can be finding as follows;

Numeric fields: After sorting the field values of the records in the node from the smallest to the largest; each point is selected as a split point. The impurity statistics are calculated for the child nodes that appear after split. The best split point is selected for the area showing the greatest change in pollution according to the impurity of the split node.

Categorical fields: Each possible value combination is examined as two subsets. The impurity of the child nodes is calculated for each combination. The best split point is selected for the area that gives the greatest drop in impurity according to the splitting nodes.

Finding the best split for the node: The best split of the areas providing the greatest decreas in impurity of the nodes found is chosen as the best split for the node.

Checking stopping rules and recursivig: If the stopping rules are not triggered, the split is applied to create two child nodes. And this rule is repeated for every child node.

Pruning is done according to the complexity of the tree. Target variables are continuous. For this reason, it is necessary to prepare the data.

3.4.2.2 Quest Algorithm

The QUEST algorithm is an algorithm developed after the C&RT algorithm and similar to the C&RT algorithm. The QUEST algorithm makes all of the nodes separate, as opposed to C&RT, which makes variable selection and division simultaneously during the creation of the tree. The only variant split in QUEST

makes approximately unbiased selection. That is, if all the predictor fields are equally informative about the target field, QUEST selects from the estimated fields of any of the predicted probability fields. The QUEST algorithm has been developed to make the prejudiced selection of the branch of the tree more generalized and to reduce the calculation cost. C&RT may become bulky as it is. Automatic cost-complexity pruning can be done to the algorithm to get rid of the bulkiness (Loh & Shih, 1997).

3.4.2.3 C5.0 Algorithm

The algorithm which is improved version of the ID3 algorithm is used especially for large data sets. In general, if we talk about the structure of algorithm; trees are formed by multiple branches from each node; the number of branches equals the number of categories of the estimator. The decision tree combines a single classifier; the information gain is utilized for the merging process. Pruning is done according to the error rate in each field (IBM SPSS Modeler).

3.4.2.4 CHAID Algorithm

The Chaid (Chi-squared-Automatic-Interaction-Detection) method was developed by Kaas in 1980 by combining the possible category pair of the prediction variable with the pairs that fit the target variable, where there was no statistically significant difference to calculate the best division (IBM SPSS Modeler). Chi-square test is used instead of entropy or Gini metrics to select the most appropriate sections in Chaid analysis (Mamaoğlu, 2005). To compute the best pane, prediction variables are combined until there is no statistically significant difference in a pair that matches the target variable for calculating the best division. As a result of this analysis, the variable having the highest F value in the independent variables whose effect on the dependent variable is statistically significant is ranked in the Chaid diagram firstly (Koyuncugil, 2007). If the relationship between CHAID analysis variables is more complex than the linear structure, it is used by eliminating certain parts of the data to find this relation hidden in the data. Algorithm is based on the use of many cross-

tabulations and statistical significance, and the study is named "Chi-square" (Hoare, R., 2004).

CHAID analysis summarizes a large data set, where multiple categories of variables are involved, by dividing them by similar categories and dividing them by the values considered important. After combining the categories for each explanatory variable in a meaningful way, the contingency tables for the dependent variable are created and the Bonferroni p values and χ^2 statistics are calculated. Data are divided into subgroups according to the categories of explanatory variables with the smallest Bonferroni p value, after the explanatory variables are compared with each other.

In CHAID analysis, after finding the best partition for each descriptive variable, the best of the descriptive variables is compared to the selected one and the repetations are performed according to the best selected descriptive variable. In this way all subdivisions are re-analyzed independently. Make possible divisions allowed for each explanatory variable category to create contingency tables according to the significance level in the χ^2 test. In summary, CHAID analysis uses chi-square statistics, the Bonferroni approach and the category combining algorithm, and allows the researcher to obtain the most important explanatory variables and interactions with the dependent variable with the tree diagram.

For each selected pair of values, the CHAID checks whether the resulting value of p is greater than a certain merging threshold. If the answer is positive, merge the values and search for an additional potential pair to join. The process is repeated until a significant pair is found (Kaas, 1980).

3.4.2.4.1 Merging Categories for Predictors in CHAID Algorithm. All predictor fields are merged to combine categories that are not statistically different with respect to the target field for determine each split. If X is used to split the node, X, each final category of a predictor field, will represent a child node. The following steps are applied to each predictor field X:

1- If X has one or two categories; no more categories are merged, so proceed to node splitting below.
2- Find the appropriate pair of the least significant (most similar) X categories determined by the p value of the statistical test associated with the target field. For more information is given below. For original fields, only adjacent categories are suitable for merging but all pairs are suitable for nominal areas.
3- For the pair with the largest p value, if the p value is greater than α_{merge}, then combine the pair of categories into a single categorization. Otherwise, proceed to step.
4- If the user chooses to allow splitting of merged categories option and the newly created compound category has three or more original categorizations, then find the best binary split in the compound category (the smallest p-value of the statistical test). If this p value is less than or equal to $\alpha_{split-merge}$, apply the split to form two categories from the compound category.
5- Continue to merge the categories in step 1 for this predictor field.
6- Any category that is less than the user-specified minimum segment size records are combined with the most similar category (gives the largest p value when compared with small category) (IBM SPSS Modeler).

3.4.2.4.2 Splitting Node. When the categories are combined for all predictive fields, each field is evaluated to be associated with the target field, depending on the regulated p value of the statistical test of association, as described below.

The predictor with the strongest correlation shown by the smallest p value is compared to the α_{split} threshold. If the P value is less than or equal to α_{split}, then that field is selected as the split area for the current node. Each of the merged categories of the split field defines a child node of the split.

After the current node has been applied, the child node is examined to see if the child nodes guarantee splitting by applying the merge/split operation, respectively. For each unsplit node, the process proceeds consecutively until one or more stops

have been triggered by the rule, and no further splits can be made (IBM SPSS Modeler).

3.4.2.4.3 Stopping Rules. The stopping rules control how the algorithm decides when the tree stops splitting nodes. When tree growth proceeds, every leaf node in the tree triggers at least one stopping rule. When any of the following conditions are made, stopping rules apply:

- The node is pure (all records have the same value for the target field)
- All records on the node have the same value for all the predictor fields used by the model.
- The tree depth for the current node (the number of recurrent nodes that define the current node) is the maximum tree depth (default or specified by the user).
- The number of records in the node is less than the minimum parent node size (default or user-specified)
- The number of records in any of the child nodes resulting from the best splitting of the node is less than the minimum child node size (default or specified by the user).
- The best split for a node gives a larger p-value than the splited size (default or user specified) (IBM SPSS Modeler).

3.4.2.4.4 Statistical Tests Used. The type of the destination field changes the calculation of unset p values. During the consolidation phase, the categories are compared with pairwise. For such comparisons, only records belonging to one of the comparison categories in the current node are considered. During the split step, the categories are considered when computing the p-value, so that all records in the current node are used (IBM SPSS Modeler).

3.4.2.4.5 Target Field. For models with a scale-level target field, p-value is calculated based on a standard ANOVA F-test that compares the target means across the predictor domain categorizations. The F statistic is calculated as

$$F = \frac{\sum_{i=1}^{I} \sum_{n \in D} w_n f_n I(x_n = i)(\bar{y}_i - \bar{y})^2 / (I-1)}{\sum_{i=1}^{I} \sum_{n \in D} w_n f_n I(x_n = i)(\bar{y}_i - \bar{y})^2 / (N_f - 1)} \tag{3.1}$$

and the p-value is $p = \Pr(F(I-1, N_f - 1) > F)$, N_f is the sum of frequency weights for all records in node t, where

$$\bar{y}_i = \frac{\sum_{n \in D} w_n f_n y_n I(x_n = i)}{\sum_{n \in D} w_n f_n I(x_n = i)} \tag{3.2}$$

$$\bar{y} = \frac{\sum_{n \in D} w_n f_n y_n}{\sum_{n \in D} w_n f_n} \tag{3.3}$$

$$N_f = \sum_{n \in D} f_n \tag{3.4}$$

$F(I-1, N_f - 1)$ is a random variable following an F-distribution with $(I-1)$ and $(N_f - 1)$ degrees of freedom.

For models with a nominal target field, the p-value is calculated based on the chi-sqared statistics. If the target field Y is a categorical field, the null hypothesis of independence of X and Y is tested. For doing the test, a contingency table is formed using classes of Y as columns and categories of the predictor X as rows. Expected cell frequencies under the null hypothesis are estimated. Observed and expected cell frequencies are used to calculate the Chi-square statistic, and the value of p is based on the calculated statistic (IBM SPSS Modeler).

For the ordinal target field Y, the null hypothesis of independence of X and Y is tested against the row effects model, which the rows being the categories of X and the columns the categories of Y (Goodman, 1979).

3.4.2.4.6 Likelihood-ratio Chi-squared test. The likelihood-ratio chi-square is calculated based on the expected and observed frequencies. The likelihood ratio chi-square is calculated as

$$G^2 = 2\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ij} \ln(n_{ij} / \tilde{m}_{ij}) \tag{3.5}$$

And the p-value is calculated as $p = \Pr(\chi_d{}^2 > G^2)$.

3.4.2.4.7 Bonferroni Adjustment. The adjusted p value is calculated by the p value of a Bonferroni multiplier. The bonferron multiplier controls the overall p value in multiple statistical tests.

If we suppose that a predictor field originally has I categories, and it is reduced to r categories after the merging step, the Bonferroni multiplier B is the number of possible ways that I categories can be merged into r categories. For $r = I$, $B = 1$. For $2 \leq r < I$, (IBM SPSS Modeler).

$$B = \binom{I-1}{r-1} \quad \text{ordinal predictor} \tag{3.6}$$

$$B = \sum_{v=0}^{r=1} (-1)^v \frac{(r-v)^r}{v!(r-v)!} \quad \text{nominal predictor} \tag{3.7}$$

$$B = \binom{I-2}{r-2} + \binom{I-2}{r-1} \quad \text{ordinal with a missing value} \tag{3.8}$$

In summary, the CHAID algorithm is as follows.

1-For each estimator variable X, find a category pair that is least important to consider the X's Y-target variable. (This has the largest p value). The method will calculate the p value depending on the measure level of Y.

a) If Y is continuous, use the F test.
b) If Y is a categorical, edit a two-way cross table with X's categories in rows and Y's categories in columns. Use the likelihood ratio test or Pearson ki-square test.
c) If y is sequential, fit Y union model (Goodman, 1979). Use the likelihood ratio test.

2- Compare to the p value with the predefined alpha level α_{merge} for the category pair of X which has the largest p value.

a) If p is greater than α_{merge} , combine this pair under a single category. Start the process from step 1 for X's new category set.
b) b) If p value is less than α_{merge}, go to step 3.

3-Calculate the corrected p value by using the appropriate Bonferroni correction for the category set of X and Y.

4-Select the X estimator variable with the smallest corrected p value(the most important). Compare the p value with the pre-defined alpha level α_{split}.

a) If the p value is less than or equal to α_{split} , split the node based on the category set of X.
b) b) If the p value is greater than α_{split}, do not split the node. This node is the end node.

5- Continue the tree growth process until you see the stop rules (IBM SPSS Modeler).

One of the most important differences between CHAID and other methods is tree derivation. CHAID derives non-binary multiple trees when the other methods derive binary multiple (Berson et al., 2000). CHAID can work continuously and categorically with all types of variables. In addition, continuous predictive variables are automatically categorized according to the purpose of the analysis. CHAID categorizes groups that differ according to the level of relationship through Chi-square metric. Because of this reason, the leaves of the tree are branched not in pairs but in the number of different structures in the data.

Decision trees have been decided to be used in the profile of the significance ratios in the data mining techniques. Among many decision tree methods, CHAID analysis, which makes the most accurate determinations in terms of tree growing, has been applied. CHAID analysis can generate multiple trees differently from other methods. At the same time it works for continuous and categorical data. The continuous predictor variables in the analysis are automatically categorized according to the purpose of the analysis. Groups that differ according to the level of the relationship are also classified using the chi-square metric. Classifications are divided according to the number of different structures in the given tree. In this study, the profiles of companies operating in manufacturing and retail market were determined by CHAID.

While many multivariate statistical techniques such as regression analysis, discriminant analysis and logistic regression analysis have been used to date, our study has benefited from decision tree algorithms which have started to be accepted in our country recently. In this study, the reasons for the use of decision tree algorithms can be explained as; be used in other multivariate techniques because of being in nonparametric methods, no need for statistical assumptions, visualization of the relationship between dependent and independent variables.

CHAPTER FOUR
STUDY: MANUFACTURING AND RETAIL SECTORS ANALYSIS

4.1 Identification of the Problem

It is necessary to identify factors that affect financial performance in order to comment on companies when changes in the market are excluded. Therefore it is necessary to make an evaluation based on the financial data of the companies and this is a long and difficult process. So, it is necessary to determine which of these ratios used in interpretation is more effective. In addition, it should be normal that the effects of the ratios from sector to sector. In this context, how can the companies respond to the question of which ratios are more effective in interpreting the financial performance of firms?

The main objective of this study is to determine the most important variables that investors, business owners and other individuals in decision-making positions should consider in evaluating the financial performance of enterprises and to reveal the most significant ratios in determining the financial performance on a sectoral basis. In this way, it is aimed to facilitate the identification of sector-based financial performance by the ratios that are important on the basis of sectors.

4.2 Selection of the Data

In this study, the balance sheet and income statement of companies operating in manufacturing and retail marketting sectors, which were traded in BIST during 2016 periods. They were obtained from the web browser of the stock exchange Istanbul. The study was conducted with 33 retail marketing and 158 manufacturing firms traded on the stock exchange during the 2016 period.

The companies that are traded in the BIST and whose financial account period is January 1-December 31 are evaluated, ın order to reach a healthy conclusion in calculating and evaluating the ratios. The ratios in the four main headings are

significant when examining the financial situations of firms. These are liquidity ratios, turnover ratios, financial structure ratios, profitability ratios.

The concept of liquidity can be expressed as the period of cash flow of an asset. If the companies are closed today, it will be possible to determine whether all the assets in the hands and the debts can be closed in a period of 1 year.

Turnover ratios are the ratios of company's revenues and assets. These ratios can also be defined as productivity ratios.

Financial structure ratios are the ratios of company's assets and resources, which are calculated from the companies balance sheets. It is known that companies have financed their assets in two ways: borrowing and own funds. Borrowing helps firms to grow while raising financing costs. It will increase the risk of firms. Therefore, it is preferred that the assets be financed with own funds.

Profitability ratios are the result of the work carried out companies in certain periods. These ratios shape the decisions of firms and investors.

Calculated rates are listed on the Table 4.1. Financial firms enable investors to meet savings owners. Financial firms do not include the production process such as non-financial firms. These companies operate in the form of a fund transfer using a variety of tools and are different from other production and service companies (Erken, 1998). So, financial firms have been excluded from the study in order to achieve consistent and meaningful results.

Table 4.1 Financial ratios

A	LIQUIDITY RATIOS	C	TURNOVER RATIOS
1	Current Ratio	1	Inventory Turnover
2	Acid-Test Ratio	2	Receivables Turnover
3	Quick Ratio (Liquidity Ratio)	3	Working Capital Turnover
4	Inventories to Current Assets	4	Net Working Capital Turnover
5	Inventories to Total Assets	5	Tangible Fixed Assets Turnover
6	Inventory Dependency Ratio	6	Own Funds Turnover
7	Current Account Receivables to Current Assets	7	Total Assets Turnover
8	Current Account Receivables to Total Assets		
B	**FINANCIAL STRUCTURE RATIOS**	**D**	**PROFITABILITY RATIOS**
1	Total Loans to Total Assets	1	Net Profit To Own Funds
2	Own Funds to Total Assets	2	Profit Before Tax To Own Funds
3	Own Funds to Total Loans	3	Profit Before Interest And Tax+Financing Expenses To Total Liabilities
4	Short-Term Liabilities to Total Liabilities	4	Net Profit To Total Assets
5	Long-Term Liabilities to Total Liabilities	5	Operating Profit To Assets Used In Carrying Out Of The Operations
6	Tangible Fixed Assets to Own Funds	6	Cumulative Profitability Ratio
7	Tangible Fixed Assets to Long-Term Liabilities	7	Operating Profit To Net Sales
8	Fixed Assets to Total Loans	8	Gross Profit To Net Sales
9	Fixed Assets to Own Funds	9	Net Profit To Net Sales
10	Fixed Assets to Long Term Loans+Own Funds	10	Cost Of Goods Sold To Net Sales
11	Short-Term Liabilities to Total Loans	11	Interest Expenses To Net Sales
12	Bank Loans to Total Assets	12	Profit Before Interest And Tax + Financing Expenses To Financing Expenses
13	Bank Loans to Short-Term Liabilities	13	Net Profit And Interest Expenses Net Profit + Financing Expenses To Interest Expenses
14	Bank Loans to Total Loans		
15	Current Assets to Total Assets		
16	Tangible Fixed Assets to Total Assets		

It is observed that the ratios used are different in terms of sectors and discrimination is made. The details of which rates are used in the study are given in Appendix- 1 and Appendix-6. At the same time, the basic statistical concepts of the rates used are included.

4.3 Prediction of Financial Situation

In addition, Altman z-score calculations are made for each company and companies are classified in two groups. The classification was made according to the Altman z-scores. The first ratio, which was found by Altman, was used in the study. According to Altman's ratio scoring, good and bad profiles were estimated for the companies. The forecast interval was chosen to be 2.67. After the calculations, poor financial situation estimates have been made for companies with a score below 2.67; good financial situation estimates have been made for companies with a score above 2.67.

Altman z-score values have been calculated and listed Figure 4.1 below for tehe retail sector.

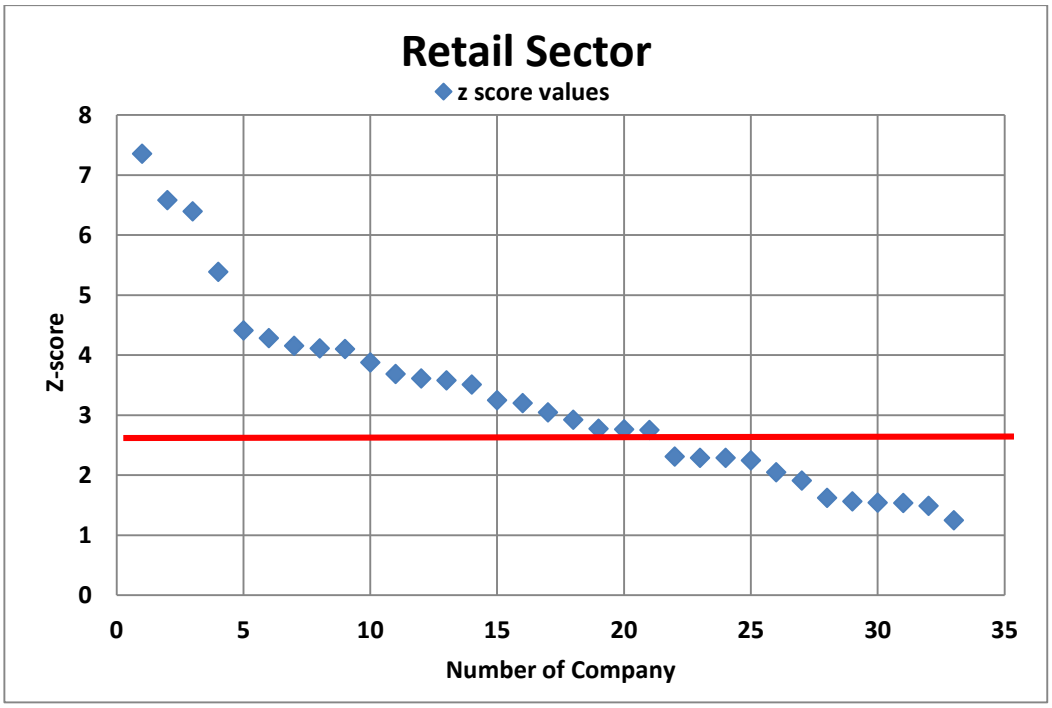


Figure 4.1 Retail Sector-Z Score

Altman z-score values have been calculated and listed Figure 4.2 below for the manufacturing sector.

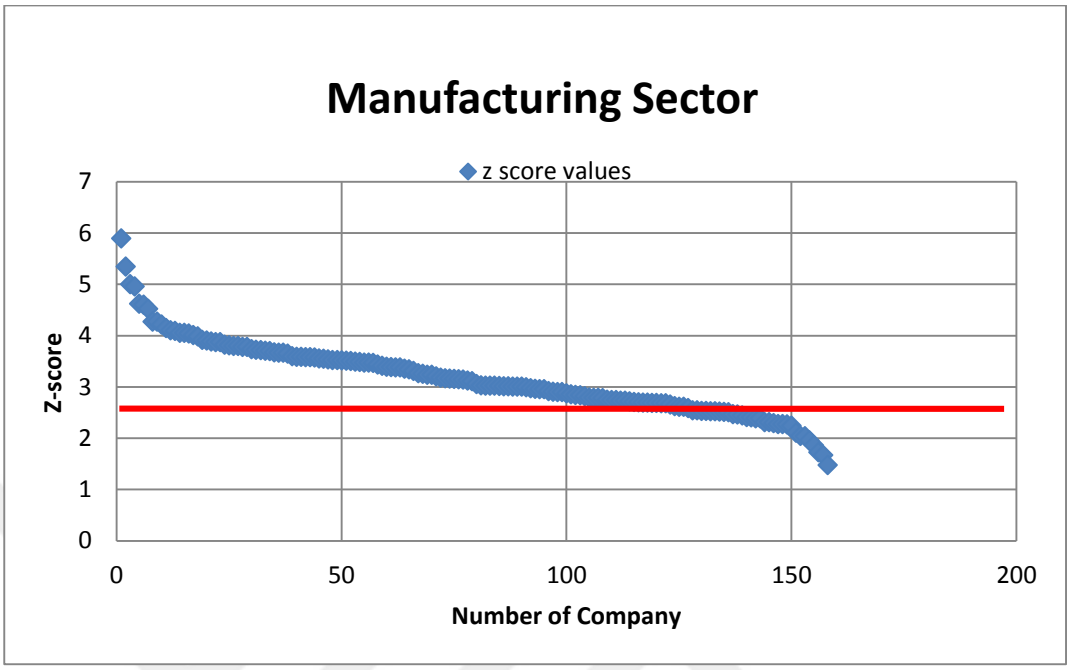


Figure 4.2 Manufacturing Sector Z-Score

4.4 Classification of Ratios

In this study, the determination of the effective rates in the determination of the financial profiling was done using CHAID analysis application from data mining techniques. After profiling, the most significant ratios were determined for the finding profiles. It is aimed to find the most effective rates for determining the financial profiles of the companies.

Decision Trees algorithms consist of dependent variables and independent variables. Analysis was performed using the IBM SPSS Modeler program. $\alpha_{split-merge}$ =0,05 and decision trees were obtained separately for each sector. During the implementation of the model, the cross - validation rule was selected and different node numbers were attempted to follow the algorithm's stopping rules to monitor the different tree structures. CHAID analysis was used to determine individual ratios in the each sector, and total industry data were not studied. Studies on the significance of ratios over total data, allow CHAID to perform the most

appropriate division, creating different profiles and investigating the importance of ratios. But, the aim of this study is to determine the rates that differ in the sectors in terms of success and failure. Because of this, each sector has been calculated separately. The results obtained as a result of the classification are described below with distinguishing sectors.

4.4.1 Chaid Algorithm in Retail Sector

There are 33 retail sector companies that are traded at BIST in 2016 period. 44 rates were calculated for the first 33 companies traded in the stock market. And significant ratios were included in the study.

The comparison of the z-scores obtained in the classification of the CHAID algorithm and the true z-score values is shown Figure 4.3. CHAID algorithm is used in our application, because C&RT and C5.0 algorithms do not perform as accurate classification as the CHAID analysis in comparison. The results of comparison with real z values of C&RT and C5.0 algorithms are given in Appendix-7 and Appendix-9. Decision trees which are derived from C&RT and C5.0 algorithms are also included in Appendix-8 and Appendix-10.

Good-bad	default	non-default
default	12	0
non-default	0	21

Figure 4.3 Cross tabulation at CHAID algorithm in retail sector

The decision tree obtained in the appliction of the CHAID algorithm is also included in the Figure 4.4.

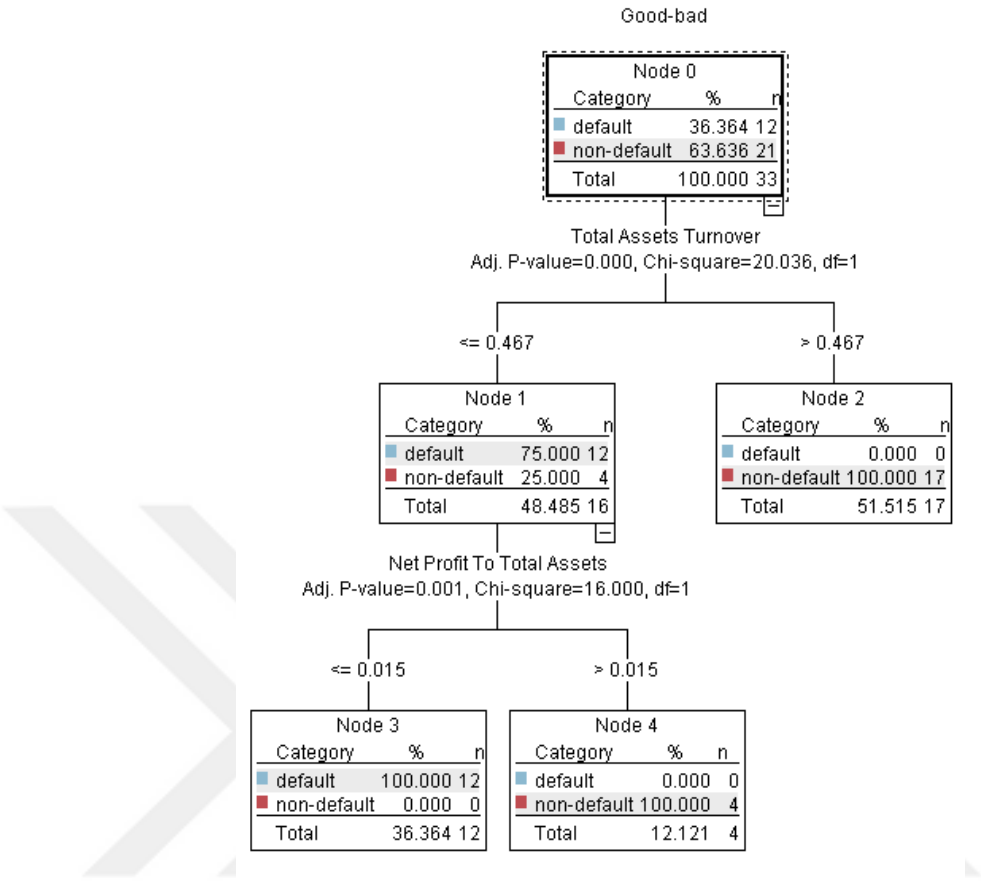


Figure 4.4 Decision tree of retail sector

Total assets turnover is the statistically strongest correlation with the financial performances of 33 firms in total. According to the total assets turnover values, the nodes are gathered between 1-4. The other nodes that make up these 5 nodes are creating profiles.

Profile-1: (Node 2) the profile consists of total assets turnover ratio. Profile-1 has a total of 17 companies, all of which have good financial stuation companies. If total assets turnover ratio is greater than 0.467, then it says that companies are expected to have good financial conditions.

Profile-2: (Node 3) the second profile consists of total assets turnover ratio and net profit to total assets ratio. If total assets turnover ratio is less than or equal to 0.467,

net profit to total assets is greater than 0.015, then it says that companies are expected to have good financial conditions. 4 companies providing this profile is monitored and financial performance is weak.

Profile-3: (Node-4) The third profile consists of total assets turnover ratio and net profit to total assets ratio. If total assets turnover ratio is less than or equal to 0.467, net profit to total assets is less than or equal to 0.015, then it says that companies are expected to have bad financial conditions. There are 4 companies which have the succesfull financial performans.

4.4.2 Chaid Algorithm in Manufacturing Sector

There are 158 manufacturing sector companies that are traded at BIST in 2016 period. 44 rates were calculated for the companies traded in the stock market. And significant ratios were included in the study.

The comparison of the z-scores obtained with the classification of the CHAID algorithm and the true z-score values is shown Figure 4.5 CHAID algorithm is used in our application, because C&RT and C5.0 algorithms do not perform as accurate classification as the CHAID analysis in comparison. The results of comparison with real z values of C&RT and C5.0 algorithms are given in Appendix-2 and Appendix-4. Decision trees which are derived from C&RT and C5.0 algorithms are also included in Appendix-3 and Appendix-5.

Good-bad	default	non-default
default	34	2
non-default	2	118

Figure 4.5 Cross tabulation at CHAID algorithm in manufacturing sector

The decision tree obtained in the appliction of the CHAID algorithm is also included in the Figure 4.6.

Current assets to total assets ratio is the statistically strongest correlation with the financial performances of 158 firms in total. According to the total assets turnover

values, the nodes are gathered between 1-18. The other nodes that make up these 18 nodes are creating profiles.

Profile-1: (Node-10) the profile consists of current assets to total assets, total loans to total assets ratios. If the current assets to total assets ratio is greater than 0.592, total loans to total assets ratio is greater than 0.731; then there are 3 companies which has the bad financial performance and there are 7 companies which has the good financial performance. The detailed view of the profile is also included in the Figure 4.7.

Profile-2: (Node-9) the profile consists of current assets to total assets, total loans to total assets ratios. If the current assets to total assets ratio is greater than 0.592, total loans to total assets ratio is less than or equal to 0.731; then there are 53 companies which has the good financial performance. The detailed view of the profile is also included in the Figure 4.7.

Profile-3: (Node-8) the profile consists of current assets to total assets, net profit to total assets ratios. If the current assets to total assets ratio is between (0.317, 0.592] and net profit to total assets ratio is greater than 0.033; then there are 36 companies which has the good financial performance. The detailed view of the profile is also included in the Figure 4.7.

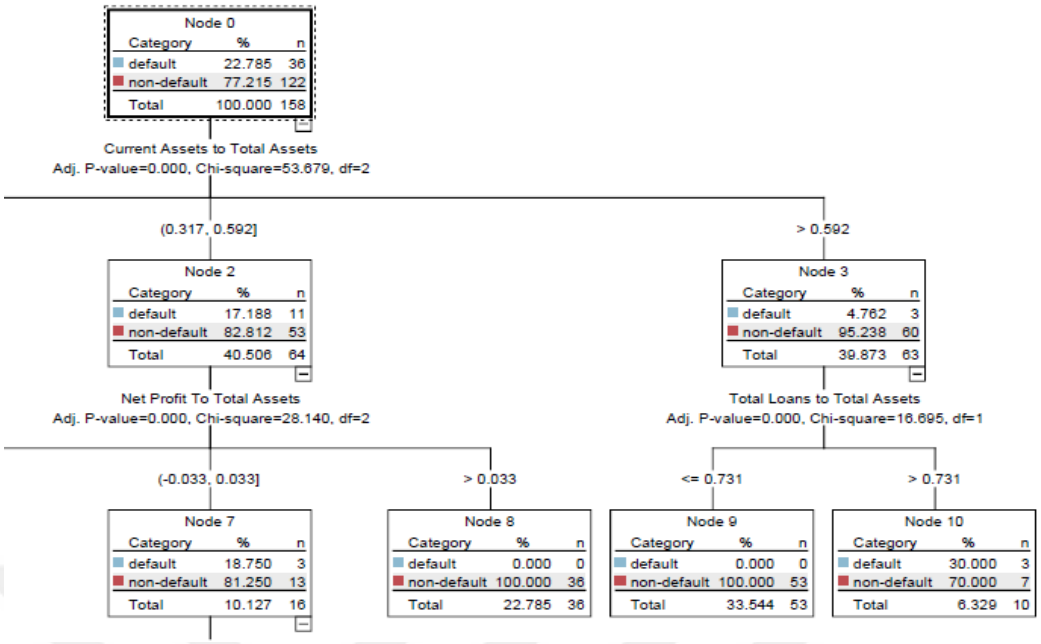


Figure 4.6 Decision tree nodes 2-3-7-8-9-10

Profile-4: (Node-16) the profile consists of current assets to total assets, net profit to total assets and operating profit to net sales ratios. If the current assets to total assets ratio is between (0.317, 0.592], net profit to total assets ratio is between (-0.033, 0.033] and operating profit to net sales ratio is greater than 0.102; then there are 3 companies which has the bad financial performance. The detailed view of the profile is also included in the Figure 4.7.

Profile-5: (Node- 15) the profile consists of current assets to total assets, net profit to total assets and operating profit to net sales ratios. If the current assets to total assets ratio is between (0.317, 0.592], net profit to total assets ratio is between (-0.033, 0.033] and operating profit to net sales ratio is less than or equal to 0.102; then there are 13 companies which has the good financial performance. The detailed view of the profile is also included in the Figure 4.7.

Profile-6: (Node- 14) the profile consists of current assets to total assets, net profit to total assets and own funds turnover ratios. If the current assets to total assets ratio is between (0.317, 0.592], net profit to total assets ratio is less than or equal to -0033, and own funds turnover ratio is greater than 0.762; then there are 8 companies which

has the bad financial performance and there are 1 companies which has the good financial performance. The detailed view of the profile is also included in the Figure 4.7.

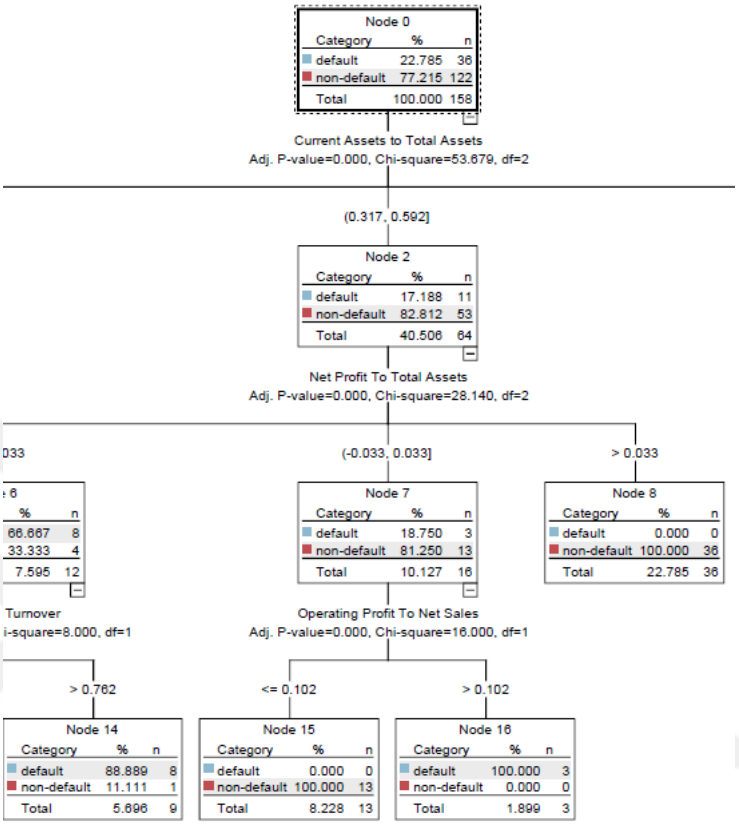


Figure 4.7 Decision tree nodes 14-15-16

Profile-7: (Node- 13) the profile consists of current assets to total assets, net profit to total assets and own funds turnover ratios. If the current assets to total assets ratio is between (0.317, 0.592], net profit to total assets ratio is less than or equal to -0033, and own funds turnover ratio is less than or equal to 0.762; then there are 3 companies which has the good financial performance. The detailed view of the profile is also included in the Figure 4.8.

Profile-8: (Node-5) the profile consists of current assets to total assets, tangible fixed assets turnover. If the current assets to total assets ratio is less than or equal to 0.317 and tangble fixed assets turnover is greater than 1.879; then there are 4 companies

which has the good financial performance. The detailed view of the profile is also included in the Figure 4.8.

Profile-9: (Node-18) The profile consists of current assets to total assets, tangible fixed assets turnover, interest expenses to net sales, quick ratio. If the current assets to total assets ratio is less than or equal to 0.317 and tangible fixed assets turnover is less than or equal to 1.879, interest expenses to net sales ratşo is greater than 0.035 and quick ratio is greater than 0.344; then there are 20 companies which has the bad financial performance. The detailed view of the profile is also included in the Figure 4.8.

Profile-10: (Node-17) the profile consists of current assets to total assets, tangible fixed assets turnover, interest expenses to net sales, quick ratio. If the current assets to total assets ratio is less than or equal to 0.317 and tangible fixed assets turnover is less than or equal to 1.879, interest expenses to net sales ratşo is greater than 0.035 and quick ratio is less than or equal to 0.344; then there are 1 companies which has the bad financial performance and there are 1 companies which has the good financial performance. The detailed view of the profile is also included in the Figure 4.8.

Profile-11: (Node-11) the profile consists of current assets to total assets, tangible fixed assets turnover, interest expenses to net sales, quick ratio. If the current assets to total assets ratio is less than or equal to 0.317 and tangible fixed assets turnover is less than or equal to 1.879, interest expenses to net sales ratio is less than or equal 0.035; then there are 1 companies which has the bad financial performance and there are 4 companies which has the good financial performance. The detailed view of the profile is also included in the Figure 4.8.

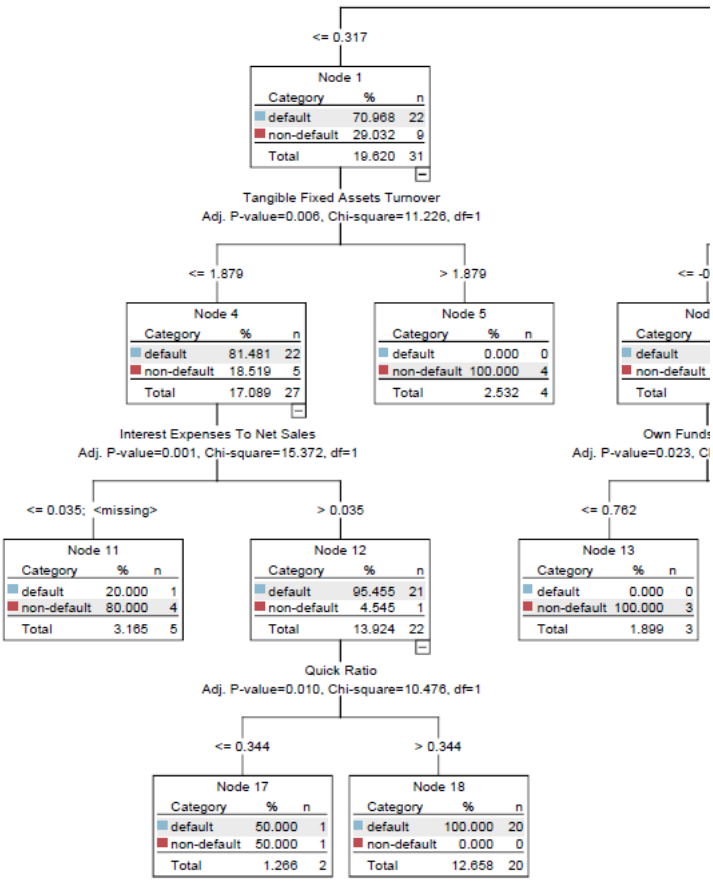


Figure 4.8 Decision tree nodes 5-11-13-17-18

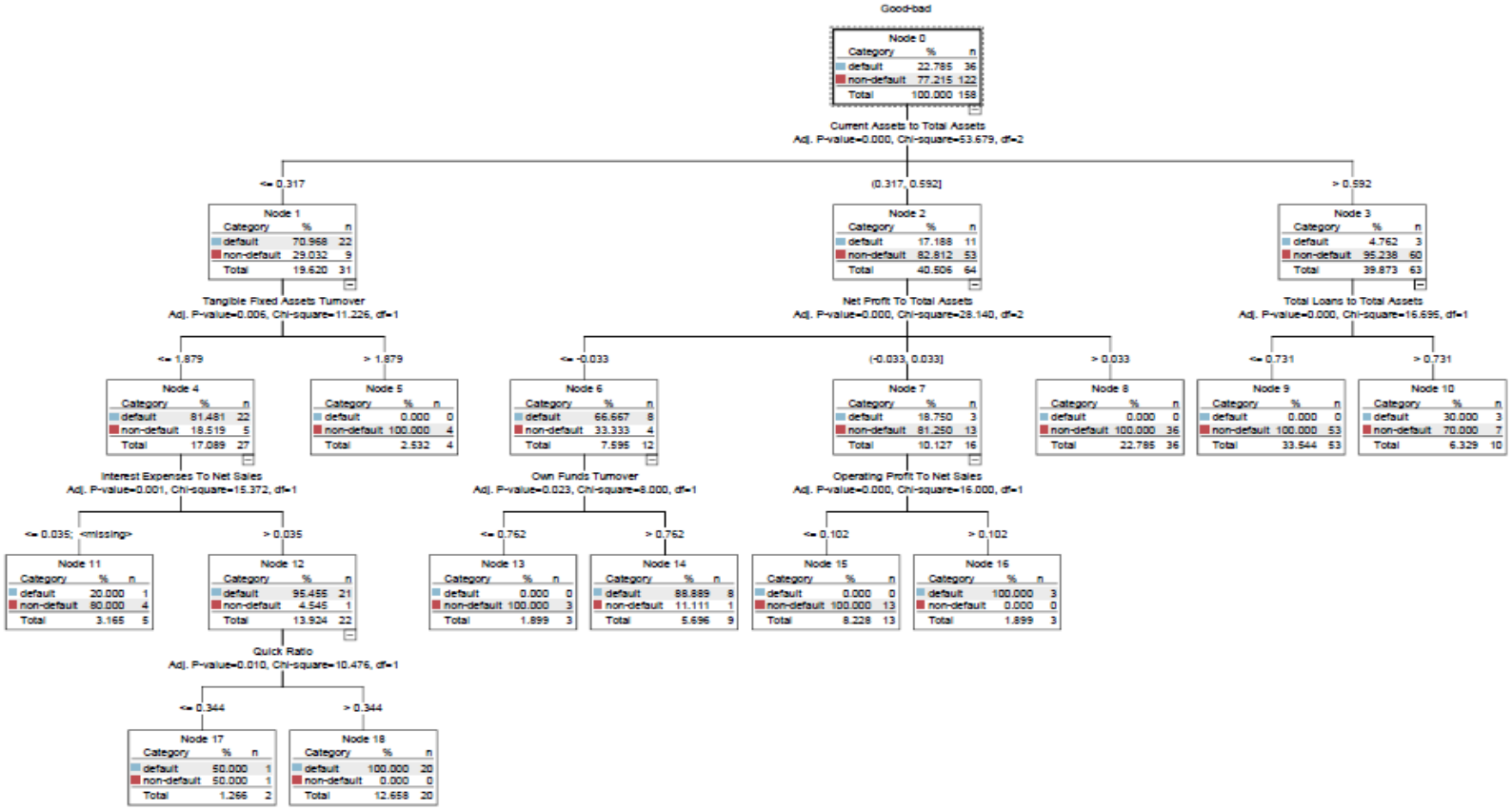


Figure 4.9 Decision tree in manufacturing sector

CHAPTER FIVE
CONCLUSION

The increase in the amount of data makes traditional methods inadequate and time consuming in the process of converting data into information at these days in the technology age. Data mining has been widely used as a result of the discovery of confidential information by appropriate techniques. It is widely used in many areas like health, education, trade, military areas, shopping, government and private sector.

Ratios, affecting financial performance, were determined with decision trees method of data mining techniques in different sectors. Financial situation of companies forecasts were found by multiple discriminant analysis. Thus, a mixed breath has been introduced into the credit score according to both traditional and new methods. For this purpose, balance sheets and income statements for 2016 period of firms operating in retail and manufacturing sectors were obtained from BİST. 44 financial ratios were calculated over the financial data provided. The data obtained by using financial information from companies have been applied to the CHAID algorithm, which has been found to be significant in determining the financial performance of the sector. In the algorithm, the dependent variable was selected as financial performance and independent variable was selected as 44 ratios. Financial performance, which is dependent variable, is calculated according to Altman z-score.

In the study, according to the most appropriate tree structure examined and interpreted, the following issues have been revealed as a result of the evaluation of the findings obtained.

12 of 33 companies operating in the retail trade sector have failed financially and 21 have been successful. The most significant variables were total assets turnover, net profit to net sales with the CHAID algorithm applied in financial status detection. In the retail sector, it is observed that the operating financing is carried out mainly through bank borrowings. When the retail sector is considered, it is known that the companies in the sector have a fast stock and credit cycle. In this context, the

operational profitability is directly related to the periods. Equity and fixed asset structures are considered to be negligible and are shaped by the status of active asset and financing-oriented activities. As can be seen in the results obtained in this direction, the rate of active turnover and the ratio of net sales to asset size is determined as the most effective rate in determining the financing profile. It can be suggested that the speed of the cash returns of stocks and receivables should be paid attention to the companies in the retail sector.

36 of 158 companies operating in the manufacturing sector have been financially successful and 122 have failed. The most significant variables were current assets to total assets, total loans to total assets, net profit to total assets, tangible fixed assets turnover, interest expenses to net sales, own funds turnover, and operating profit to net sales, quick ratio the CHAID algorithm applied in financial status detection. When the manufacturing sector is considered, it is known that it is made up of firms with certain resources for production. In the case of insufficient funds, financial debts and stocks in the active structure are financed. In this context, a robust equity structure is expected in stock and credit financing, and if the equity is insufficient, the financial debt structure is expected to be long term.

REFERENCES

Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23* (4), 589-609.

Altman, E. (2000). Predicting financial distress of companies: Revisiting the Z-score and zeta models. Retrieved Feb 2, 2018, from https://pdfs.semanticscholar.org/3a40/ad1e6e88fc05ae19564fbd90bccae48accd1.pdf .

Beaver, W.H. (1966). Financial ratios as predictors of failure. *The Journal of Accounting Research*, (4).

Berson, A., Smith S. & Thearling K. (1999). *Building data mining applications for CRM*. USA: McGraw-Hill.

Breiman, L. (1984). *Classification and regression trees.* New York: Taylor and Francis Group.

Bryan, J.G. (1951). The generalized discriminant function, Mathematical foundation & computational routine. *Harvard Educational Review, 22* (2), 90-95.

Chen, Z. (2001). An integrated approach. In *Data mining and uncertain reasoning* (370). Canada: John Wiley & Sons.

Cios K., Swiniarski J., Roman W, Witold, Pedrycz W., Kurgan & Lukasz A. (2007). *Data mining knowledge discovery approach.* USA:Espringer Science, Businesss Media.

Cochran, W. G. (1964). On the performance of the linear discriminant function. *The Journal of Technometrics*, (6), *179-190.* Retrieved June, 04, 2018, from https://amstat.tandfonline.com/doi/abs/10.1080/00401706.1964.10490162#.WvHukIiFM2w .

Durand, D. D. (1941) Risk elements in consumer installment financing, Studies in consumer installment financing. *National Bureau of Economic Research*, 105-142. Retrieved April 5, 2018, from https://EconPapers.repec.org/RePEc:nbr:nberbk:dura4.

Erken, A. (1998). *Başlıca fiyat bazlı oranların hisse senedi getirisi üzerindeki etkileri ve İstanbul menkul kıymetler borsası'nda bir uygulama*. (Unpublished Master's Thesis). Ankara: University Institute of Social Sciences.

Fayyad, U., Pıatetsky-Shapıro, G. & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17* (3).

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179-188.

Frawley W. J. , Piatetsky G., Shapiro G., and Matheus C. J., (1992). Knowledge discovery in databases: An overview. *AI Magazine 13* (3).

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, (74), 537-552.

Han J. & Kamber M.(2001). *Data mining: concepts and techniques.* USA: Morgan Kaufmann Publishing .

Hastie, T., Tibshirani, R. & Friedman, J.(2001). *Springer series in statistics. The elements of statistical learning; data mining, ınference and prediction*. USA: Springer Publisher.

Hoare, R. (2004). *Using CHAID for classification problems.* Paper presented at the New Zealand Statistical Association Conference, Wellington, New Zealand.

IBM SPSS Modeler 18.1 algorithms guide. (n.d.). Retived May 5, 2018 from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.1/en/AlgorithmsGuide.pdf.

Kaas,G.V. (1980) An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Journal of the Royal Statistical Society, 29* (2).

Kohavi R., & Provost F. (1998). Glossary of terms, *Machine Learning 30* (2/3), 271–274.

Kovalerchuk, B. & Vityaev E. (2000). *Data mining in finance.* USA: Kluwer Academic Publisher, Hingham MA.

Koyuncugil, A.S. (2007). *Borsa şirketlerinin sektörel risk profillerinin veri madenciliğiyle belirlenmesi*, (1-17). Ankara: Capital Market Board Report, 1-17.

KPMG Sektörel Bakış. Retrived May 8, 2018, from https://assets.kpmg.com/content/dam/kpmg/tr/pdf/2018/01/sektorel-bakis-2018-perakende.pdf

Kulalı, İ. (2016) Altman Z-skor modelinin BIST şirketlerinin finansal başarısızlık riskinin tahmin edilmesinde uygulanması. *Journal of Management Economics and Business, 12* (27).

Loh, W.& Shih, Y.(1997) Split selection methods for classification trees. *Institute of Statistical Science, Academia Sinica, 7* (4).

Mamaoğlu, T. (2005). *Veri madenciliginde karar agaçları ile bir ögrenci ders basarısı tahmin aracı*. (Unpublished Master's Thesis). Kocaeli University Graduate School of Natural and Applied Sciences, Kocaeli.

Miller, W. (2009). *Comparing models of corporate bankruptcy prediction: Distance to default vs. z-score.* Morningstar, Inc. Retived May 5, 2018, from https://ssrn.com/abstract=1461704.

Moss, L.T. & Atre, S. (2003). *Bussiness intelegence roadmap*. USA: Addison Wisley.

Myers H. & Forgy E. W. (1963) Development of numerical credit evaluation systems. *Journal of American Statistical Association*, (50), 797-806.

Outecheva, N. (2007). *Corporate financial distress: An empirical analysis of distress risk.* The University of St.Gallen Graduate School Of Business Administration, Economics, Law and Social Sciences.

Rao, C. R. (1952). *Advanced statistical methods in biometric research.* New York: John Wiley & Sons, Inc.

Sayılgan, G. (2008). *İşletme finansmanı*. Ankara: Turhan Publishing.

Silahtaroğlu, G. (2008). *Kavram ve algoritmalarıyla veri madenciliği*. İstanbul: Papatya Publishing.

Smith, K. V. (1965). *Classification of investment securities using MDA.* Purdue University. Institute for Research in the Behavioral, Economic, and Management Sciences.

Thearlıng, K. (2004). *Information about analytics and data science.* Retrieved Feb 02, 2018, from http://www.thearling.com/.

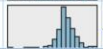
TÜSİAD 2017. Retrived May 8, 2018, from http://www.eticaretraporu.org/wp-content/uploads/2017/04/TUSIAD_E-Ticaret_Raporu_2017.pdf

Walter, J. E. (1959) A Discriminant function for earnings price ratios of large industrial cor- porations. *Economics and Statistics Review*, (41), 44-52.

Yılmaz, Koltan, Ş., Albayrak, S.A. (2009), Veri madenciliği: karar ağacı algoritmaları ve imkb verileri üzerine bir uygulama. *The journal of Süleyman Demirel University Faculty of Economics and Administrative Sciences*, *14* (1), 31-52.

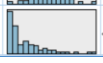
APPENDICES

Appendix-1 Ratios used in manufacturing sector

Field	Graph	Measurement	Min	Max	Range	Mean	Std. Dev	Skewness	Kurtosis
z score		Continuous	1.476	5.895	4.420	3.173	0.707	0.606	1.174
Current Ratio		Continuous	13.726	1141.457	1127.731	210.928	179.659	2.647	9.096
Acid-Test Ratio		Continuous	0.071	340.047	339.976	28.862	64.969	3.028	9.576
Quick Ratio		Continuous	0.020	515.005	514.985	37.681	72.581	4.018	19.331
Inventories to Current Assets		Continuous	0.000	0.822	0.822	0.307	0.166	0.524	-0.082
Inventories to Total Assets		Continuous	0.000	0.638	0.638	0.156	0.105	1.504	3.319
Current Account Receivables/ Current Assets		Continuous	0.061	0.912	0.850	0.456	0.185	0.330	-0.466
Current Account Receivables/ Total Assets		Continuous	0.015	0.846	0.831	0.245	0.146	1.106	1.711
Total Loans to Total Assets		Continuous	0.052	4.395	4.343	0.521	0.390	6.310	61.883
Own Funds to Total Assets		Continuous	-3.395	0.948	4.343	0.479	0.390	-6.310	61.883
Own Funds to Total Loans		Continuous	-0.772	18.202	18.974	1.851	2.405	3.311	15.549
Short-Term Liabilities to Total Liabilities		Continuous	0.045	3.947	3.902	0.458	0.360	5.887	55.618
Long-Term Liabilities to Total Liabilities		Continuous	0.000	0.778	0.778	0.285	0.179	0.557	-0.352
Tangible Fixed Assets to Long-Term Liabilities		Continuous	0.083	52.234	52.151	6.204	8.664	3.040	10.819
Fixed Assets to Total Loans		Continuous	0.063	13.021	12.958	1.347	1.456	4.252	27.144
Fixed Assets to Long Term Loans+Own Funds		Continuous	-0.155	8.111	8.267	0.829	0.794	6.186	50.653
Total Assets Turnover		Continuous	0.000	4.380	4.380	0.874	0.561	2.581	11.464
Profit Before Interest And Tax+Financing Expenses To Total Liabilities		Continuous	-0.197	0.465	0.662	0.089	0.103	0.359	1.557
Net Profit To Total Assets		Continuous	-0.407	0.358	0.765	0.036	0.095	-0.562	3.548
Operating Profit To Assets Used In Carrying Out Of The Operations		Continuous	-0.173	0.387	0.560	0.072	0.088	0.261	1.346
Cumulative Profitability Ratio		Continuous	0.000	0.770	0.770	0.047	0.090	5.164	33.928
Operating Profit To Net Sales		Continuous	-0.416	0.362	0.778	0.084	0.115	-0.735	2.720
Gross Profit To Net Sales		Continuous	-0.218	0.567	0.786	0.218	0.113	-0.130	1.057
Net Profit To Net Sales		Continuous	-0.728	0.496	1.224	0.044	0.151	-0.981	5.604

Short-Term Liabilities to Total Loans		Continuous	0.222	1.000	0.778	0.715	0.179	-0.560	-0.351
Current Assets to Total Assets		Continuous	0.122	0.994	0.872	0.528	0.200	0.102	-0.732
Tangible Fixed Assets to Total Assets		Continuous	0.006	0.878	0.872	0.472	0.200	-0.102	-0.732
Inventory Turnover		Continuous	0.000	137.991	137.991	7.725	14.643	6.748	52.049
Working Capital Turnover		Continuous	0.000	8.085	8.085	1.729	1.006	2.471	10.509
Net Working Capital Turnover		Continuous	-73.849	854.746	928.595	11.421	72.073	10.557	121.649
Tangible Fixed Assets Turnover		Continuous	0.000	118.042	118.042	5.613	13.861	6.684	49.871
Own Funds Turnover		Continuous	-77.166	500.001	577.167	4.809	40.237	11.970	148.866

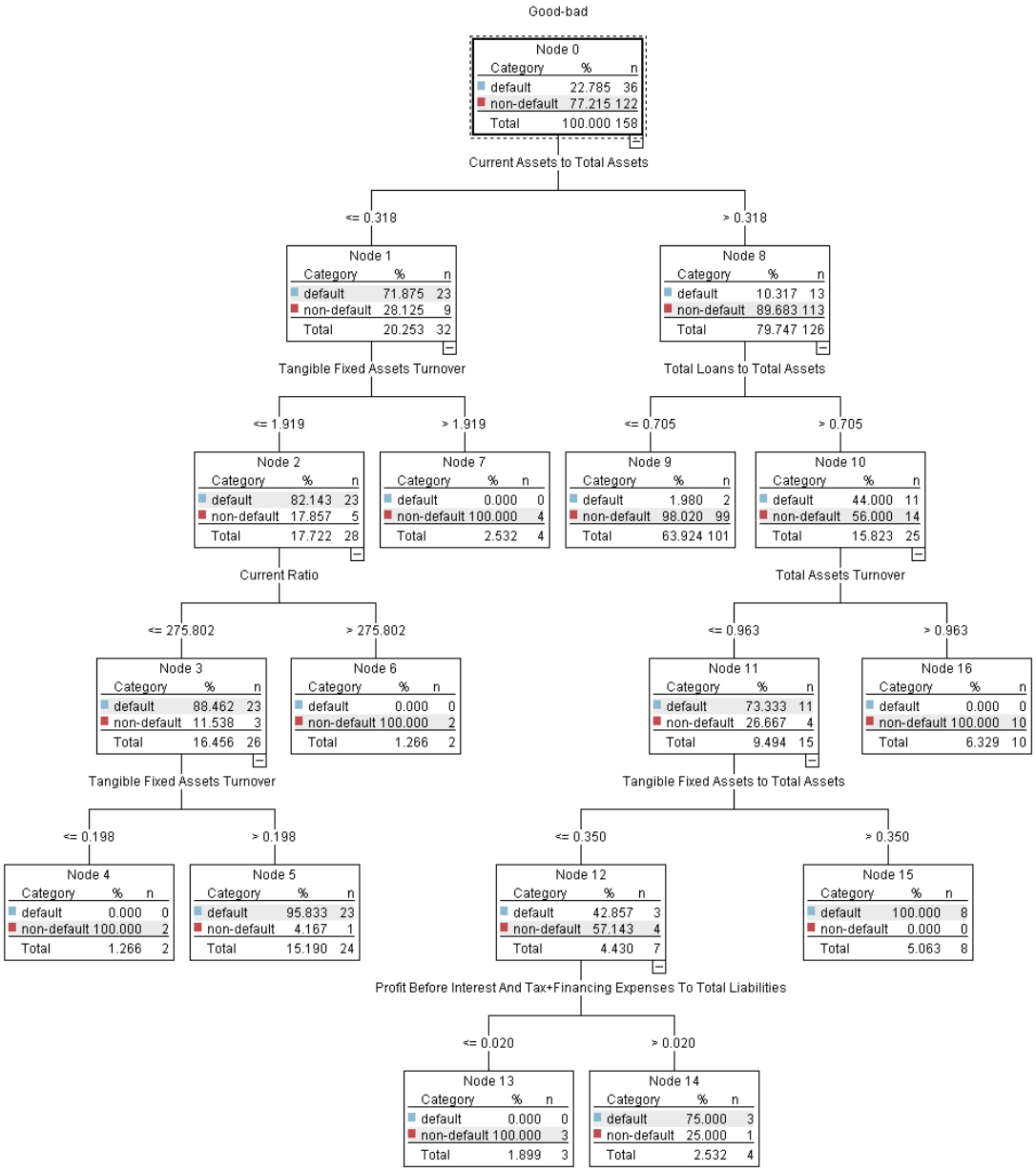
Total Assets Turnover		Continuous	0.000	4.380	4.380	0.874	0.561	2.581	11.464
Profit Before Interest And Tax+Financing Expenses To Total Liabilities		Continuous	-0.197	0.465	0.662	0.089	0.103	0.359	1.557
Net Profit To Total Assets		Continuous	-0.407	0.358	0.765	0.036	0.095	-0.562	3.548
Operating Profit To Assets Used In Carrying Out Of The Operations		Continuous	-0.173	0.387	0.560	0.072	0.088	0.261	1.346
Cumulative Profitability Ratio		Continuous	0.000	0.770	0.770	0.047	0.090	5.164	33.928
Operating Profit To Net Sales		Continuous	-0.416	0.362	0.778	0.084	0.115	-0.735	2.720
Gross Profit To Net Sales		Continuous	-0.218	0.567	0.786	0.218	0.113	-0.130	1.057
Net Profit To Net Sales		Continuous	-0.728	0.496	1.224	0.044	0.151	-0.981	5.604

Cost Of Goods Sold To Net Sales		Continuous	0.433	1.218	0.786	0.782	0.113	0.125	1.041
Interest Expenses To Net Sales		Continuous	0.000	0.533	0.533	0.083	0.095	1.995	5.000

Appendix-2 Cross-tabulation at C5.0 algorithm in manufacturing sector

Good-bad	default	non-default
default	32	4
non-default	2	118

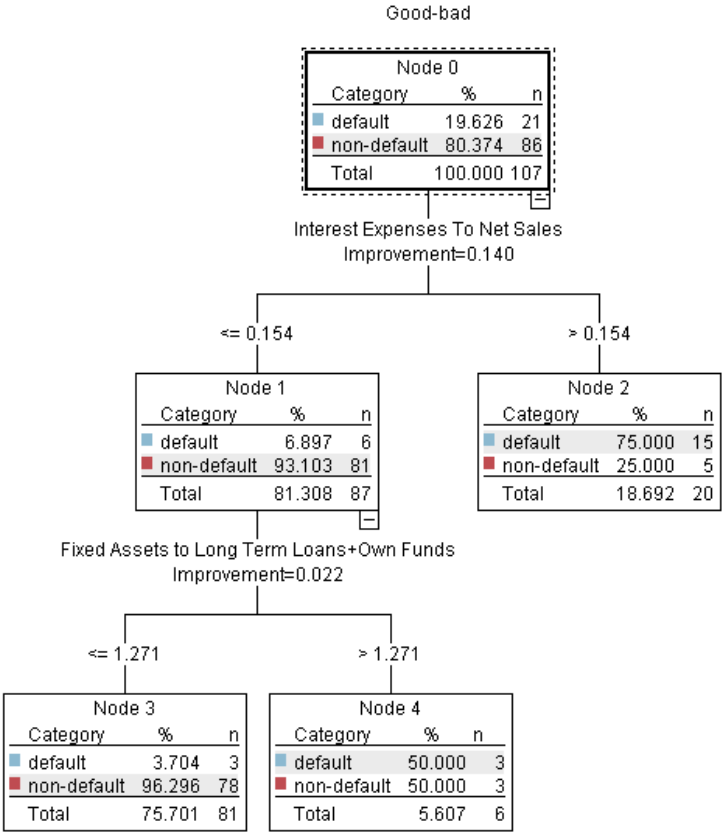
Appendix-3 Decision tree at C5.0 algorithm in manufacturing sector



Appendix-4 Cross tabulation at C&RT algorithm in manufacturing sector

Good-bad	default	non-default
default	24	12
non-default	12	108

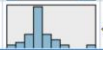
Appendix-5 Decision tree at C&RT algorithm in manufacturing sector



Appendix-6 Ratios used in retail sector

z score		Continuous	1.249	7.354	6.106	3.268	1.524	1.010	0.824
Current Ratio		Continuous	29.704	686.269	656.565	147.611	159.878	2.544	6.600
Acid-Test Ratio		Continuous	0.162	6.423	6.262	1.120	1.286	2.655	8.580
Quick Ratio		Continuous	0.117	621.370	621.253	42.288	120.059	4.234	18.592
Current Account Receivables/ Current Assets		Continuous	0.006	0.937	0.931	0.471	0.310	-0.056	-1.486
Current Account Receivables/ Total Assets		Continuous	0.002	0.821	0.819	0.220	0.243	1.345	0.654
Total Loans to Total Assets		Continuous	0.051	1.081	1.030	0.575	0.319	-0.170	-1.291
Own Funds to Total Assets		Continuous	-0.081	0.949	1.030	0.425	0.319	0.170	-1.291

Short-Term Liabilities to Total Liabilities		Continuous	0.031	1.075	1.043	0.452	0.312	0.418	-1.001
Long-Term Liabilities to Total Liabilities		Continuous	0.000	0.780	0.779	0.232	0.248	0.903	-0.520
Tangible Fixed Assets to Own Funds		Continuous	-1.595	12.335	13.930	1.205	2.344	3.684	16.347
Fixed Assets to Own Funds		Continuous	-3.048	24.199	27.247	3.003	5.221	2.837	8.854
Fixed Assets to Long Term Loans+Own Funds		Continuous	-3.292	4.323	7.616	1.044	1.176	-0.669	6.609
Short-Term Liabilities to Total Loans		Continuous	0.220	1.000	0.779	0.765	0.247	-0.911	-0.498
Bank Loans to Short-Term Liabilities		Continuous	0.000	0.698	0.698	0.143	0.195	1.440	1.146
Bank Loans to Total Loans		Continuous	0.000	0.751	0.751	0.160	0.244	1.462	0.559

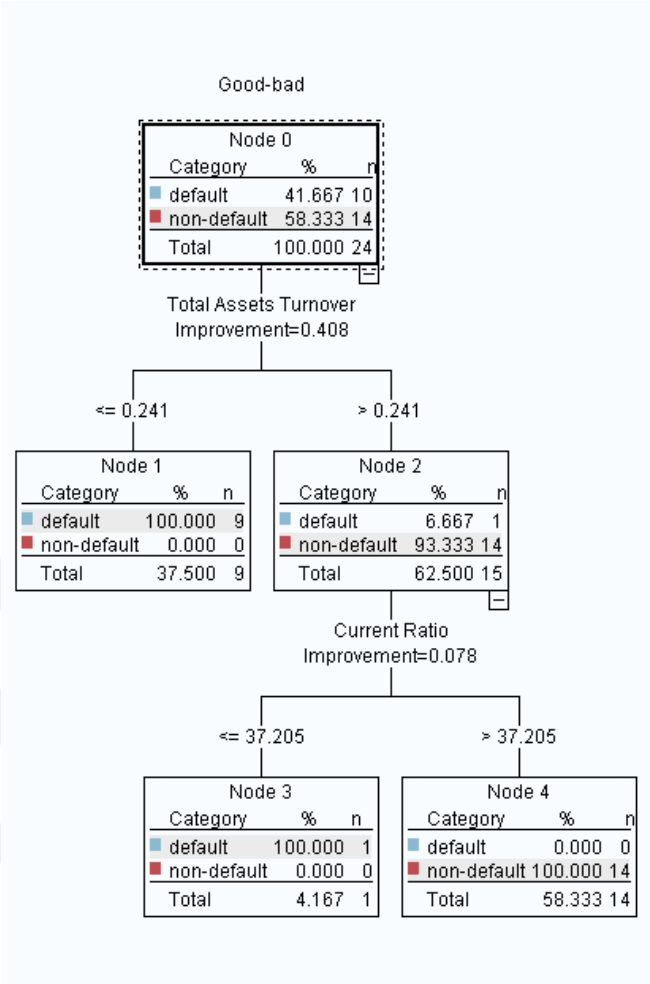
Current Assets to Total Assets		Continuous	0.020	1.000	0.980	0.448	0.294	0.363	-1.096
Tangible Fixed Assets to Total Assets		Continuous	0.000	0.980	0.980	0.552	0.294	-0.363	-1.096
Working Capital Turnover		Continuous	0.000	7.561	7.561	2.377	2.161	0.965	0.190
Net Working Capital Turnover		Continuous	-64.083	102.970	167.053	-1.673	23.509	2.170	13.572
Own Funds Turnover		Continuous	-49.650	241.848	291.498	12.279	45.189	4.388	22.144
Total Assets Turnover		Continuous	0.000	4.822	4.822	1.134	1.307	1.324	1.221
Net Profit To Own Funds		Continuous	-29.665	2.594	32.259	-0.995	5.196	-5.570	31.654
Profit Before Interest And Tax+Financing Expenses To Total Liabilities		Continuous	-0.221	0.527	0.748	0.058	0.129	1.372	5.008

Net Profit To Total Assets		Continuous	-0.268	0.401	0.669	-0.002	0.134	0.596	1.605
Operating Profit To Assets Used In Carrying Out Of The Operations		Continuous	-0.224	0.616	0.840	0.017	0.152	1.941	6.861
Cumulative Profitability Ratio		Continuous	0.000	0.092	0.092	0.016	0.025	1.865	2.528
Operating Profit To Net Sales		Continuous	-2.410	8.121	10.531	0.049	1.557	4.429	24.226
Gross Profit To Net Sales		Continuous	-0.267	1.000	1.267	0.205	0.299	1.606	2.148
Net Profit To Net Sales		Continuous	-4.318	6.411	10.729	0.114	1.660	1.542	8.212
Cost Of Goods Sold To Net Sales		Continuous	0.000	1.267	1.267	0.764	0.327	-1.418	1.164

Appendix-7 Cross tabulation at C&RT algorithm in retail sector

Good-bad	non-default
default	12
non-default	21

Appendix-8 Decision tree at C&RT algorithm in retail sector



Appendix-9 Cross tabulation at C5.0 algorithm in retail sector

Good-bad	default	non-default
default	12	0
non-default	0	21

Appendix-10 Decision tree at C5.0 algorithm in retail sector