STATISTICAL METHODS IN CREDIT RATING

ÖZGE SEZGİN

SEPTEMBER 2006

STATISTICAL METHODS IN CREDIT RATING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF APPLIED MATHEMATICS
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

ÖZGE SEZGİN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER
IN
THE DEPARTMENT OF FINANCIAL MATHEMATICS

SEPTEMBER 2006

Approval of the Graduate School of Applied Mathematics

Prof. Dr. Ersan AKYILDIZ
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master.

Prof. Dr. Hayri KÖREZLIOĞLU
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master.

Assist. Prof. Dr. Kasırga YILDIRAK
Supervisor

Examining Committee Members

Prof. Dr. Hayri KÖREZLIOĞLU

Assoc. Prof Dr. Azize HAYVAFİ

Assoc. Prof. Dr. Gül ERGÜN

Assist. Prof. Dr. Kasırga YILDIRAK

Dr. C. Coşkun KÜÇÜKÖZMEN

“I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.”

Name, Lastname : ÖZGE SEZGİN

Signature :

Abstract

STATISTICAL METHODS IN CREDIT RATING

SEZGİN Özge
M.Sc., Department of Financial Mathematics
Supervisor: Assist. Prof. Dr. Kasırga YILDIRAK

September 2006, 95 pages

Credit risk is one of the major risks banks and financial institutions are faced with. With the New Basel Capital Accord, banks and financial institutions have the opportunity to improve their risk management process by using Internal Rating Based (IRB) approach. In this thesis, we focused on the internal credit rating process. First, a short overview of credit scoring techniques and validation techniques was given. By using real data set obtained from a Turkish bank about manufacturing firms, default prediction logistic regression, probit regression, discriminant analysis and classification and regression trees models were built. To improve the performances of the models the optimum sample for logistic regression was selected from the data set and taken as the model construction sample. In addition, also an information on how to convert continuous variables to ordered scaled variables to avoid difference in scale problem was given. After the models were built the performances of models for whole data set including both in sample and out of sample were evaluated with validation techniques suggested by Basel Committee. In most cases classification and regression trees model dominates the other techniques. After credit scoring models were constructed and evaluated, cut-off values used to map probability of default obtained from logistic regression to rating classes were determined with dual objective optimization. The cut-off values that gave the maximum area under ROC curve and minimum mean square error of regression tree was taken as the optimum threshold after 1000 simulation.

Keywords: Credit Rating, Classification and Regression Trees, ROC curve, Pietra Index

Öz

KREDİ DERECELENDİRMEDE İSTATİSTİKSEL TEKNİKLER

SEZGİN Özge
Yüksek Lisans, Finansal Matematik Bölümü
Tez Yöneticisi: Yrd. Doç. Dr. Kasırga YILDIRAK

Eylül, 2006 95 sayfa

Kredi riski, bankalar ve finansal kuruluşların karşılaştıkları başlıca risklerden biridir. Yeni Basel Sermaye Uzlaşısıyla birlikte, bankalar ve finansal kuruluşlar iç derecelendirmeye dayanan yaklaşımla risk yönetimi yöntemlerini geliştirme olanağına sahiptirler. Bu tezde iç derecelendirme yöntemi üzerinde durulmuştur. İlk önce, kredi skorlama teknikleri ve geçerlilik testleri hakkında kısa bir tanıtım verilmiştir. Daha sonra, imalat sanayi firmaları hakkında Türkiye'deki bir bankadan elde edilen gerçek veri seti kullanılarak borcu ödememe tahmini, lojistik regresyon, probit regresyon, ayırma (diskriminant) analizi ve sınıflandırma ve regresyon ağaçları modelleri oluşturulmuştur. Modellerin performanslarını geliştirmek için, lojistik regresyon için en iyi örneklem tüm veri kümesi içinden seçilmiştir ve modellerin kurulması için kullanılacak örneklem olarak alınmıştır. Ayrıca, değişkenlerin ölçü farklılıkları problemini engellemek için, sürekli ölçekli verinin nasıl sıralı ölçekli veriye dönüştürüldüğü hakkında bilgi verilmiştir. Modeller kurulduktan sonra modellerin performansları örneklem içi ve dışı tüm veri seti için Basel Komitesi tarafından önerilen geçerlilik testleriyle değerlendirilmiştir. Tüm durumlarda klasifikasyon ve regresyon ağaçları modeli diğer yöntemlerden üstündür. Kredi skorlama modelleri oluşturulduktan ve değerlendirildikten sonra, lojistik regresyon sonucu elde edilen ödememe olasılıklarını, derece sınıflarına atayan kesim noktaları iki amaçlı optimizasyon ile belirlenmiştir. 1000 simülasyondan sonra ROC eğrisi altında kalan maksimum alanı veren ve regresyon ağacı için minimum hata kareler ortalamasını veren kesim noktaları alınmıştır.

Anahtar Kelimeler: Kredi Derecelendirme, Sınıflandırma ve Regresyon Ağaçları, ROC eğrisi, Pietra Endeksi

To my family

Acknowledgments

I appreciate my supervisor, Assist. Prof. Dr. Kasırga YILDIRAK for his great guidance, his support and providing me a suitable data set.

I deeply thank the members of the Baskent University Statistics and Computer Sciences Department for encouraging me and sharing me their experience about statistical techniques.

I am grateful to my family for their patience and support.

Lastly I am indebted to my friend Sibel KORKMAZ that she shared her latex files and to all my friends for their understanding.

Table of Contents

CHAPTER

List of Tables

List of Figures

Chapter 1

Introduction and Review of Literature

Managing credit risk becomes one of the main topics of modern finance with the recent dramatic growth in consumer credit. Credit risk is the risk of financial loss due to the applicants' failure to pay the credit back. Financial institutions and banks are trying to deal with the credit risk by determining capital requirements according to the risk of applicants and by minimizing the default risk with using the statistical techniques to classify the applicants into "good" and "bad" risk classes. By taking into account these facts Basel Committee on Banking Supervision put forward to use risk based approaches to allocate and charge capital. According to the Committee credit institutions and banks have the opportunity to use standard or internal rating based (IRB) approach when calculating the minimum capital requirements [1].

The standard approach is based on the ratings of external rating agencies such as Standard and Poors (S&P) and Moody's whereas IRB is based on institutions' own estimates. IRB system can be defined as a process of assessing creditworthiness of applicants. The first step is to determine the probability of default of the applicant by means of statistical and machine learning credit scoring methods such as discriminant analysis, logistic regression, probit regression, non-parametric and semi-parametric regression, decision trees, linear programming, neural networks and genetic programming.

The results of credit scoring techniques can be used to decide whether to grant or not to grant credit by assessing the default risk. Since 1941 beginning with the Durand's [2] study most of the studies in literature has been concentrated on using qualitative methods for default prediction. Less attention has been given to the second step of IRB approach. After default probability is estimated, observations are classified into risk levels by cut-off values for default probabilities. By this way credit scoring results not only used to decide to give credit, it can be also applied to credit risk management, loan pricing and minimum capital requirement estimation.

This thesis is not only concentrated on credit scoring models but also the applicants were mapped to the rating grades. This thesis is organized as follows:

Firstly, future works of default prediction are summarized, then short overview about classification and New Basel Capital Accord [3] is given in Chapter 2 and Chapter 3. Chapter 4 and Chapter 5 give the technical details about statistical credit scoring techniques and validation techniques. In Chapter 6 data set and the sample selected are described, the model parameters are estimated, performances of models are compared and optimal scale determination is explained. Concluding remarks are given in Chapter 7.

1.1 REVIEW OF LITERATURE

Credit assessment decision and the default probability estimation have been the most challenging issues in credit risk management since 1930's. Before the development of mathematical and statistical models, the credit granting was based on judgemental methods. Judgemental methods have many shortcomings. First of all, the methods are not reliable since they depend on creditors' mode. The decisions may change from one person to another, so they are not replicable and difficult to teach. They are unable to handle a large number of applications [4]. By the development of classification models and ratio analysis, these methods took the place of judgemental methods.

The studies using ratio analysis generally use the potential information of financial statements to make decision about the firm's profitability and financial difficulties. One of the most important studies about ratio analysis was conducted by Beaver in 1966 [5]. The aim of the study was not only to predict the payment of loans but also to test the ability of accounting data to predict by using likelihoods. To avoid sample bias, a matched sample of failed and non-failed firms was used in univariate ratio analysis. Additionally, by profile analysis the means of ratios were compared. In 1968, Beaver [6] expanded his study to evaluate whether market prices were affected before failure. The conclusion shows that investors recognize the failure risk and change their positions of failing and so the price decline one year before failure.

Beaver's study [5] was repeated and compared with linear combination of ratios in 1972 by Deakin [7].

The earliest study about statistical decision making for loan granting was published by Durand in 1941 [2]. Fisher's discriminant analysis was applied to evaluate the creditworthiness of individuals from banks and financial institutions. After this study, the discriminant age of credit granting was started. This study followed by Myers and Forgy [8], Altman [9], Blum [10] and Dombolena and Khoury [11].

In 1963, Myers and Forgy [8] compared discriminant analysis with stepwise multiple linear regression and equal weighted linear combination of ratios. In this study, both financial and non-financial variables were used. Firstly, the variables in nominal scale were scaled into a "quantified" scale from best to worst. Surprisingly, they found that equal weighted functions' predictive ability is as effective as other methods.

In 1968, Altman [9] tried to assess the analytical quality of ratio analysis by using the linear combination of ratios with discriminant function. In the study, the discriminant function with ratios was called as Z-Score model. Altman concluded that with the Z-Score model that was built with matched sample data, 95 % of the data was correctly predicted.

In 1974, Blum [10] reported the results of discriminant analysis for 115 failed and 115 non-failed companies with liquidity and profitability accounting data. In the validation process, the correctly predicted percentages were evaluated. The results indicates that 95 % of observations classified correctly at one year prior to default but prediction power decreases to 70 % at the third, fourth and fifth years prior to default.

Dombolena and Khoury in 1980 [11] added the stability measures of the ratios to the model of discriminant analysis with ratios. The standard deviation of ratios over past few years, standard error of estimates and coefficient of variations were used as stability measures. The accuracy of ratios was found as 78 % even five years prior to failure and standard deviation was found to be the strongest measure of stability.

Pinches and Mingo [12] and Harmelink [13] applied discriminant analysis by using accounting data to predict bond ratings.

Discriminant analysis was not the only technique in 1960's, there was also the time varying decision making models built to avoid unrealistic situations by modelling the applicant's default probability varying overtime. The first study on time varying model was introduced by Cyert et al. [14]. The study followed by Mehta [15], Bierman and Hausman [16], Long [17], Corcoran [18], Kuelen [19], Srinivasan and Kim [20], Beasens et al. [21] and Philosophov et al. [22].

In 1962, Cyert et al. [14] by means of total balance aging procedure built a decision making procedure to estimate doubtful accounts. In this method, the customers were assumed to move among different credit states through stationary transition matrix. By this model, the loss expectancy rates could be estimated by aging category.

In 1968, Mehta [23] used sequential process to built a credit extension policy and established a control system measuring the effectiveness of policy. The system continues with the evaluation of the acceptance and rejection costs alternatives. The alternatives with minimum expected costs were chosen. In 1970, Mehta [15] related the process with Markov process suggested by Cyert et al. to include time varying states to optimize credit policy. Dynamic relationships when evaluating alternatives were taken into account with Markov chains.

In 1970, Bierman and Hausman [16] developed a dynamic programming decision rules by using prior probabilities that were assumed to distributed as beta distribution. The decision was taken by evaluating costs not including only today's loss but also the future profit loss.

Long [17] built a credit screening system with optimal updating procedure that maximizes the firms value. By screening system, scoring had decaying performance level overtime.

Corcoran in 1978 [18] adjusted the transition matrix by adding dynamic changes by means of exponential smoothing updated and seasonal and trend adjustments.

Kuelen 1981 [19] tried to improve Cyert's model. In this model, a position between total balance and partial balance aging decisions was taken to make the results more accurate.

Srinivasan and Kim [20] built a model evaluating profitability with Bayesian that updates the profitability of default overtime. The relative effectiveness of other classification procedures was examined.

In 2001, the Bayesian network classifier using Markov chain Monte Carlo were evaluated [21]. Different Bayesian network classifiers such as naive Bayesian classifier, tree arguments naive Bayesian classifier and unrestricted Bayesian network classifier by means correctly classified percentages and area under ROC curve were assessed. They were found to be good classifiers. Results were parsimonious and powerful for financial credit scoring.

The latest study on this area was conducted by Philosophov et al. in 2006 [22]. This approach enables a simultaneous assessment to be made of prediction and time horizon at which the bankruptcy could occur.

Although results of discriminant analysis are effective to predict, there are difficulties when the assumptions are violated and sample size is small. In 1966, Horrigan [24] and in 1970, Orgler [25] used multiple linear regression but this method is also not appropriate when dependent variable is categorical. To avoid these problems, generalized linear models such as logistic, probit and poisson regression were developed. This is an important development for credit scoring area. In 1980, Ohlson [26] used the new technique logistic regression that is more flexible and robust avoiding the problems of discriminant analysis. By using logistic and probit regression, a significant and robust estimation can be obtained and used by many researchers: Wihinton [27], Gilbert et al. [28], Roshbach [29], Feelders et al. [30], Comoes and Hill [31], Hayden [32] and Huyen [33].

Wiginton's [27] compared logistic regression with discriminant analysis and concluded that logistic regression completely dominates discriminant analysis.

In 1990, Gilbert et al. [28] demonstrated that in bankruptcy model developed with bankrupt random sample is able to distinguish firms that fail from other financially distressed firms when stepwise logistic regression is used. They found that variables distinguished bankrupt and distressed firms are different from bankrupt and nonbankrupt firms.

In 1998, Roszbach [29] used Tobit model with a variable censoring threshold proposed to investigate effects of survival time. It is concluded that the variables with increasing odds were of decreasing expected survival time.

In 1999, Feelders et al. [30] included reject inference to the logistic models and parameters estimated with EM algorithms. In 2000, Comoes and Hill [31] used logit, probit, weibit and gombit models to evaluate whether the underlying probability distribution of dependent variable really affect the predictive ability or not. They concluded that there are no really difference between models.

Hayen in 2003 [32] searched univariate regression based on rating models driven for three different default definitions. Two are the Basel II definitions and the third one is the traditional definition. The test results show that there is not much prediction power is lost if the traditional definition is used instead of the alternative two ones.

The latest study about logistic regression was by Huyen [33]. By using stepwise logistic regression, a scoring model for Vietnamese retail bank loans prediction was built.

Since credit scoring is a classification problem, neural networks and expert systems can also be applied. Beginning of 1990's and ending of 1980's can be called as the starting point of intelligent systems age. By the development of technology and mathematical sciences, systems based on human imitation with learning ability were found to solve decision making problem. In 1988, Shaw and Gentry [34] introduced a new expert system called MARBLE (managing and recommending business loan evaluation). This system mimics the loan officer with 80 decision rules. With this system, 86.2 % of companies classified and 73.3 % of companies predicted accurately. The study of Odom and Sharda' study in 1990 [35] is the start of neural network age. Backpropogation algorithm was introduced and was compared with discriminant analysis. Bankrupt firms found to be predicted more efficiently with neural networks. In 1992, Tam and Kiang [36] extended the backpropogation by incorporating misclassification costs and prior probabilities. This new algorithm compared with logistic regression, k nearest neighbor and decision tress by evaluating robustness, predictive ability and adoptability. It was concluded that this extended algorithm is a promising tool. In 1993, Coats and Fants [37] presented a new method to recognize financial distress patterns. Altman's ratios were used to compare with discriminant analysis and algorithms is found to be more accurate.

Kiviloto's [38] research included self organizing maps (SOM) a type of neural network and it was compared with the other two neural network types learning vector quantization and radial basis function and with linear discriminant analysis. As a result like in previous researches, neural network algorithm performed better than discriminant analysis especially the self organizing maps and radial basis functions. Also Charalombous et al. [39] aimed to compare neural network algorithms such as radial basis function, feedforward network, learning vector quantization and backpropogation with logistic regression. The result is similar as Kivilioto's study, the neural networks has superior prediction results.

Kaski et al. [40] extended the SOM algorithm used by Kivilioto by introducing a new method for deriving metrics used in computing SOM with Fisher's information matrix. As a result, Fisher's metrics improved PD accuracy.

The genetic programming intelligent system was used in many research. In 2005, Huang et al. [41] built a two stage genetic programming method. It is a sufficient method for loan granting.

In credit scoring, the object of banks or financial institutions is to decrease the credit risk by minimizing expected cost of loan granting or rejecting. The first study of such an mathematical optimization problem was programmed by Wilcox in 1973 [42]. He utilized a dynamic model that is relating bankruptcy in time t with financial stability at $t-i$. In 1985, Kolesar and Showers [43] used mathematical programming to solve multicriteria optimization credit granting decision and compared with linear discriminant analysis. Although the results of mathematical modelling were violated, linear discriminant analysis gave effective results. In 1997, a two stage integer programming was presented by Geherline and Wagner [44] to build a credit scoring model.

The parametric techniques such as logistic regression and discriminant analysis are easily calibrating and interpretable methods so they are popular but non-parametric methods has the advantage of not making any assumptions about the distribution o variables although they are difficult to display and interpret so there are also researches using non-parametric and semiparametric methods. Hand and Henley 1996 [45] introduced k nearest neighbor technique that is a non-parametric technique used for pattern preconization. They extended the model with Euclidian metric adjustment. In 2000, Hardle and Müller [46] used a semiparametric regression model called generalized partially linear model and showed that performed better than logistic regression.

1980's new method for classifying was introduced by Breiman et al. [47] which is splitting data into smaller and smaller pieces. Classification and regression tree is an appropriate method for classification of good and bad loans. It is also known as recursive partitioning.

In 1985, Altman, Frydman and Kao [48] presented recursive partitioning to evaluate the predictively and compared with linear discriminant analysis and concluded that performs better than linear discriminant analysis. In 1997, Pompe [49] compared classification trees with linear discriminant analysis and Neural Network. The 10-fold cross validation results indicates that decision trees outperform logistic regression but not better than neural networks. Xiu in 2004 [50] tried to build a model for consumers credit scoring by using classification trees with different sample structure and error costs to find the best classification tree. When a sample was selected one by one, this means that the proportion of good loans is equal to the proportion of bad loans and type I error divided by type II error is equals to the best results were obtained.

Chapter 2

CLASSIFICATION

2.1 CLASSIFICATION

The first step of a rating procedure is to build the scoring function to predict the probability of default. The credit scoring problem is a classification problem.

Classification problem is to construct a map from input vector of independent variables to the set of classes. The classification data consist of independent variables and classes.

$$X = \{x_i, ..., x_n\} \qquad (i = 1, ..., n), \tag{2.1}$$

$$xi = \{x_{11}, ..., x_{1p}\}, \tag{2.2}$$

$$\Omega = \{w_i, ..., w_n\} \qquad and \tag{2.3}$$

$$L = \{(x_1, w_1), ..., (x_n, w_n)\}. \tag{2.4}$$

Here,

X is the independent variable matrix,

x_i is the observation vector,

Ω is the set of classes vector, and

L is the learning sample.

There is a function c(x) defined on X that assigns an observation x_i to the numbers $w_1, ..., w_n$ by means of post experience of independent variables. It is called as classifier.

$$X \qquad \overrightarrow{c(x)} \qquad \Omega \tag{2.5}$$

The main purpose of classification is to find an accurate classifier or to predict the classes of new observations. Good classification procedure should satisfy both . If the relation between independent variables and classes is consistent with the past, a good classifier with high discriminatory power can be used as an good predictor of new observations.

In credit scoring, the main problem is to build an accurate classifier to determinate default and non-default cases and to use the scoring model to predict new applicants classes.

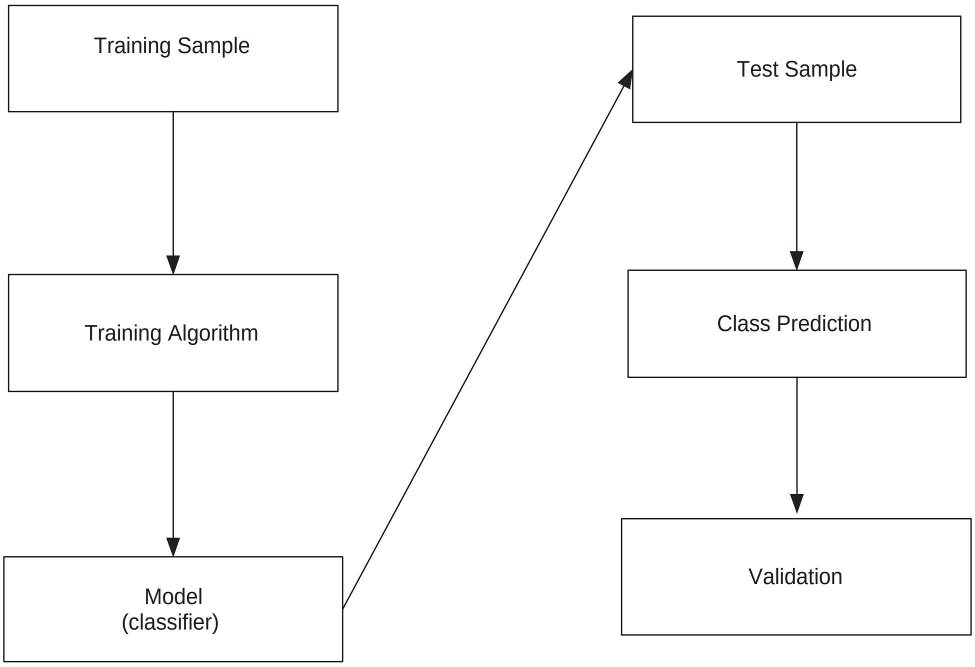


Figure 2.1: Classification flowchart

The classification procedure is implemented by the following steps:

1. The learning sample is divided into two subsamples. The first one is the training sample used to built the classifier. The second one is the test sample used to evaluate the predictive power of the classifier.
2. By using the training sample, the classifier is built by mapping X to Ω
3. The classifier is used to predict class labels of each observation in the test sample.
4. After new class labels are assigned with validation tests discriminatory power of the classifier is evaluated.
5. The classifier with high discriminatory power is used to predict the classes of new observations which are not in the learning sample.

The main goal of a classifier is to separate classes as distinct as possible.

2.1.1 Classification Techniques

There are three types of classification techniques mostly used [51]:

Statistical Techniques

During 1960s and 1970s, the mostly used technique was the linear discriminant analysis invented by Fisher. As statistical techniques and computer science has been improved, modern techniques have been started to be used. Generally, statistical techniques have underlying assumptions about their probability model and independence of variables sometimes, these can be seen as shortcomings of the models. The most popular models models are: logistic regression, probit regression, kernel regression, k nearest neighbor estimation method, etc.

Machine Learning Techniques

They are computing procedures based on computer logic. The main aim is to simplify the problem to be understood by human intelligence. The methods such as decision trees and genetic algorithms are kinds of machine learning techniques.

Neural Network Techniques

Neural networks are the combination of statistical and machine learning techniques. It combines the complexity of statistical methods with the machine learning human intelligence imitations. They consist of layers of interconnected nodes, each node producing non-linear function of its inputs. The popular ones are: backpropagation, radial basis functions and support vector machines.

2.1.2 The Difficulties in Classification

As mentioned before the fundamental aim of discriminating is to build classifiers that separate groups as well as possible. There are difficulties in building classifiers. Sometimes classifiers with high discriminatory power can not be achievable. The basic reasons causing such problems are:

i *To access the data is difficult:* As the number of sample size increases, the model assumptions such as normality are achieved more easily. If the assumptions of models are not achieved the discriminatory power of the classifier will be low. The most important factor that affects the model is the quality of the sample.

ii *The representative characteristic of independent variables are not successful to explain the difference between classes:* If the representative ability of independent variables are low, there will be overlapping problem. That means, observations with identical attributes may fall into different classes. This problem can be also defined as not including relating variables. If the sample can not be discriminated well by the independent variables means, they have low representative power. The reason is that the variables with good predictive power are omitted. To solve this problem, first all possible variables should be used to build the model, then by using variable selection or dimension reduction techniques the unnecessary ones can be eliminated.

iii *There could be mismeasurement problems of class labels:* Since the default definition changed both developed model and predictive structure. It should be consistent with the aim of the research.

Chapter 3

BASEL II ACCORD AND LIMITATIONS FOR PROBABILITY OF DEFAULT ESTIMATION

In 2001, the Banking Committee on Banking Supervision issued a new revisited Capital Accord [1] on capital requirement standards to respond to the deficiencies in 1988 accord. The fundamentals of the Accord is to protect the economy from negative signals caused by banks risks and to avoid the value of banks to drop below of depositors claims.

It has new rules for calculating the risk weights and the supervision of financial institutions. The most important difference from the viewpoint of credit risk consists in the estimation of minimum capital requirements estimation. The 1988 Accord states that banks should hold minimum capital that is the 8% of credits, since in Basel II Accord the estimation is more closely to its rating grades.

3.1 PRINCIPLES OF BASEL II ACCORD

Basel II consists of three pillars [3]:

Pillar 1

It sets principles for minimum capital requirements to cover both credit and operational risks. Capital requirement is a guarantee amount against unexpected losses.

It is taken as equity in banks accounts. To determine minimum capital requirements, a bank can either use external sources or an internal rating base approach. There are three fundamental components to calculate the minimum capital requirement according to Basel II.

a **Probability of Default (PD):** It is the likelihood that an applicant will default in one year time period.

b **Loss Given Default (LGD):** It is the proportion of the exposure that will be lost if the spllicant defaults.

c **Exposure at Default (EAD):** The nominal value of loan granted.

The minimum capital requirement (MCR) estimation is shown in (3.1) with respect to Basel II:

$$\text{MCR} = 0.08^*\text{RW}^*\text{EAD} = 0.08\ \text{RWA} \tag{3.1}$$

Here

RW is the risk weight calculated by using PD, LGD and remaining maturity of exposure.

It has specific formulas for each asset type. RWA is the risk weighted asset.

EL = PD*EAD*LGD

MCL=EAD*LGD*PD-b*EL

Where

EL is the expected loss and

b is the proportion of expected loss of loan covered by minimum capital requirement.

Pillar 2

It defines principles for supervisors to review assessments to ensure adequate capital.

The rating system and risk management activities are checked by supervisors. Supervisors review process, to be sure that banks have adequate and valid techniques for capital requirements. Accurate and valid techniques lead to better credit risk management for the banks. Banks are expected to manage their internal capital assessments.

According to Basel Committee, there is a relation between capital required and banks risk. Banks should have a process for assessing overall capital adequacy in relation to their risk profile. Supervisors are responsible for the review and evaluation of the assessment procedure. When supervisors think the validity of the rating process is not adequate, they can take appropriate actions. They can take early stage actions to prevent capitals from falling below the minimum levels required to support the risk characteristic.

Pillar 3

It sets principles about banks disclosure of information concerning their risk. Its purpose is to maintain the market discipline by completing pillar 1 and pillar 2. The Basel Committee encourages market discipline by developing sets of disclosure requirements. According to the new accord, banks should have a disclosure policy and implement a process to evaluate the appropriateness of the disclosure. For each separate risk areas banks must describe their risk management objectives and policies.

3.1.1 PD Dynamics

Probability of default is one of the challenging factors that should be estimated while determining the minimum capital requirement. New Accord has sets principles in estimating PD. According to Basel II, there are two definitions of default:

a) The bank considers that the obligor is unlikely to pay its credit. There are four main indicators that bank considers the obligor is unlikely to pat the obligation:

- The bank puts the obligation on an non-accrued stratus
- The bank sells the credit obligation at a material credit related economic loss.
- The bank consents to a distressed restriction of credit obligation.
- The obligor sought or has been placed in bankruptcy.

b) The obligor past due more than 90 days on credit obligation to the bank.

Banks should have a rating system of its obligor with at least 7 grades having meaningful distribution of exposure. One of the grades should be for non-defaulted obligor and one for defaulted only. For each grade there should be one PD estimate common for all individuals in that grade. It is called as pooled PD. There are three approaches to estimate pooled PD.

Historical experience approach:

In this approach, PD for the grade is estimated by using the historical observed data default frequencies. In other words, the proportion of defaulted obligers in a specific grade is taken as pooled PD.

Statistical Model Approach

In that approach, firstly predictive statistical models are used to estimate default probabilities of obligor's. Then, for each grade the mean or median of PDs are taken as pooled PD.

External Mapping Approach

In this approach, firstly a mapping procedure is established to link internal ratings to external ratings. The pooled PD of external rating is assigned to internal rating by means the mapping established before.

Basel II allows the banks to use simple averages of one year default rates while estimating pooled PD.

While establishing the internal rating process, the historical data should be at least 5 years, and the data used to build the model should be representative of the population. Where only limiting data are available or there are limitations of assumptions of the techniques, banks should add the margins of conservatism in their PD estimates to avoid over optimism. The margin of conservatism is determined according to the error rates of estimates depending on the satisfactory of the models. There should be only one primary technique used to estimate PD, the other methods can be used just for comparison. Therefore, the best model should be taken as the primary model representing the data.

After the estimation of PDs, the rating classes are needed to be built. The banks are allowed to use the scale of external institutions.

In the PD estimation process, just building the model is not enough supervisors need to know not only the application also the validity of the estimates. Banks should guarantee to the supervisor that the estimates are accurate and robust and the model has good predictive power. For this purpose, a validation process should be built.

The scoring models are built by using a subset of available information. While determining the variables relevant for the estimation of PD, banks should use human judgment. Human judgment is also needed when evaluating and combining the results.

Chapter 4

STATISTICAL CREDIT SCORING TECHNIQUES

4.1 GENERALIZED LINEAR MODELS

Generalized linear models (GLM) are the class of parametric regression models which are the generalization of linear probability models. These kind of models serve to describe, how the expected values of the dependent variable varies according to the changes in values of independent variables. In such models, the main aim is to find the best fitting parsimonious model that can represent the relationship between a dependent variable and independent variables.

By means of GLM we can model the relationship between variables when the dependent variable has a distribution other than normal distribution. It also allows to include non-normal error terms such as binomial or poisson.

GLM is specified with three components [52]:

1. **The Random Component**

   In a random component the dependent variable and its conditional distribution are identified. Dependent variable can be in the form of nominal, ordinal, binary, multinomial, counts or continuous. The distribution would change according to the scale of dependent variable. Generally, the distribution of dependent variable comes from exponential family such as: normal, binomial, poisson ... etc. The general form of exponential family probability distribution function is given in (4.1):

$$fy(y;\theta,\phi) = exp\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\} \tag{4.1}$$

where

ϕ is the dispersion parameter,

θ is the canonical parameter and

$a(.), b(.), c(.)$, are real valued functions [52] with

$$E[Y] = b'(\theta), \tag{4.2}$$

$$var[Y] = b"(\theta)a(\phi). \tag{4.3}$$

2. **The Systematic Component**

   The systematic component of the model consists of a set of independent variables. It is also known as linear predictor function. It is identified as 4.4:

   $$\eta i = \beta_0 + \beta_1 * x_{i1} + ... + \beta_p * x_{ip} \tag{4.4}$$

   In systematic component's can be quantitative and qualitative independent variables.

3. **The Link Function**

   The link function is the function $g(.)$ that links the random and the systematic components:

   $$g(E[Y/X]) = \eta \tag{4.5}$$

   The most common link functions are shown in Table 4.1:

   Table 4.1: The most commonly used link functions

Dist. of Y	Scale of Y	Link	η	g^{-1} (η)	E [Y/X] Range
Normal	Numeric	Identity	μ	η	(-∞,+∞)
Binomial	Binary or Multinomial	Logit	log $\frac{\mu}{1-\mu}$	$\frac{1}{1+e^{-\eta}}$	$\frac{0,1,....,n}{n}$
Binomial	Binary or Multinomial	Probit	$\phi^{-1}(\mu)$	$\phi(\eta)$	$\frac{0,1,....,n}{n}$
Poisson	Count	Log	log(μ)	e^{η}	0,1,....+∞

   The function $g(.)$ is a monotone and invertible, and it transforms the expectation of the dependent variable to the linear predictor:

$$g^{-1}(\eta) = E[Y/X]. \tag{4.6}$$

Like the other model fitting processes, the GLM fitting also includes three steps [52].

1. **Model Selection** The choice of dependent variables's scale is so important in model selection. As mentioned before, the scale of dependent variable can be nominal, ordinal, binary, numerical, multinomial or counts. According to the scale of the dependent variable the link function and the model changes. The common assumption of GLM is the independence of the observations of the dependent variable before selection a model of GLM this assumption should be satisfied.

2. **Estimation** After the selection of the model, it is required to estimate the unknown parameters. In GLM, generally maximum likelihood estimation (MLE) method is used instead of the ordinary least square (OLS) method. Then, the normality assumption of the independent variables is no more required.

   In MLE the values of the unknown parameters are obtained by maximizing the probability of the observed data set [53]. To obtain this estimates we need to identify the log-likelihood function.

   If f(y;θ) is the probability function of the observations of the dependent variable. Then log-likelihood function is as (4.7):

   $$l(\mu; y) = log(f(y; \theta)). \tag{4.7}$$

   This function shows the probability of the observed data by means of a function of unknown parameters. The unknown parameters can be estimated by maximizing the log-likelihood function or briefly by equalizing the score vector to zero.

3. **Prediction**

   Prediction means that the value of dependent variable could be at some time t in the future. After calibrating the model by using historical data, we can predict future values of the dependent variable if the independent variables at t are known.

4.1.1 Binary Choice Models

In binary GLM model, the dependent variable takes only two possible values. In credit scoring the dependent variable is identified as follows:

$$y = \begin{cases} y_i & \text{if} y_i = 1, i.e., \text{the firm defaults} \\ y_i & \text{if} y_i = 0, i.e., \text{the firm non-defaults} \end{cases}$$

There are discrete or continuous independent variables; the model is:

$$E[Y/X] = P\{Y = 1/X\} = P\{X\beta + \varepsilon > 0/X\} = F(X\beta) = \pi, \tag{4.8}$$

here

F is the cumulative distribution function (inverse link function),

β is unknown parameter vector of the model and

π is the probability that the dependent variable takes the value 1.

In binary response models, since the dependent variable takes only two possible values with probability π, it can be assumed that the distribution of the dependent variable is Bernoulli probability distribution.

The Bernoulli probability function is:

$$f(y/\pi) = \pi^y (1-\pi)^{1-y} \qquad (y = 0, 1), \tag{4.9}$$

$$E[y] = \pi, \tag{4.10}$$

$$var[y] = \pi(1-\pi). \tag{4.11}$$

Maximum likelihood estimation

As mentioned before to estimate unknown parameters we require to write the likelihood function. The likelihood function through the observed data is defined by (4.12):

$$L(x_i) := \prod_{i=1}^{n} \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i}, \tag{4.12}$$

where

$\pi(x_i)$ is the probability that each observation with x_i independent variable vector takes the value one as dependent variable.

Since mathematically it is easier to maximize the natural logarithm of the likelihood function and monotonic transformation does not change the results when finding the optimum points generally we are working with log-likelihood functions when using MLE. The log-likelihood for binary data is defined by (4.13):

$$l(x_i) = \sum_{i=1}^{n} \{y_i \text{ln}(\ p_i(x_i)) + (1 - y_i)\text{ln}(1 - p_i(x_i))\}. \tag{4.13}$$

The estimate of unknown parameter $\widehat{\beta}$ is obtained by solving 4.14

$$\frac{\partial \text{ln} L(\widehat{\beta})}{\partial \beta} = 0. \tag{4.14}$$

Goodness of fit measures

1. **Deviance**

   In regression models for binary dependent variables, the comparison of the predicted and observed models is depend on the log-likelihood function. The model is called saturated if all independent variables are used in the model. The current model is the fitted model which we want to compare with other models.

   Deviance is a measure of deviation of the model from realized values. The deviance measure is defined as:

   $$y = -2\text{ln}(\frac{\text{likelihood of the current model}}{\text{likelihood of the saturated model}}). \tag{4.15}$$

   When models are compared, we can use deviance as a measure to determine which one to choose. The model with lower deviance will be choosen.

2. **Pearson Chi-Square Goodness of Fit Statistic**

   It is a simple non-parametric goodness of fit test which measures how well an assumed model predicts the observed data. The test statistic is:

   $$\chi^2 = \sum_{i=1}^{n} \frac{(\text{observed frequency-fitted frequency})^2}{\text{fitted frequency}}; \tag{4.16}$$

χ^2 is assumed to be chi-square with $n-p$ degrees of freedom.

3. **G Likelihood Ratio Chi-Square Statistic**

   G statistic is a goodness of fit test depends on log-likelihood function. The purpose of this test is to compare the models with and without independent variables. The test statistic is:

$$\mathrm{G} = -2\ln(\frac{L_0}{L_1}) = -2(\ln L_0 - \ln L_1). \tag{4.17}$$

   Here

   L_0 is the likelihood function value of the model without any independent variables and

   L_1 is the likelihood function value of the model with independent variables.

   G is assumed to be distributed as chi-square with p-1 degrees of freedom.

4. **Pseudo R^2**

   As in linear regression, pseudo R^2 measures the explained percentage of dependent variables. It also can be called as the determination coefficient. The statistic is:

$$\mathrm{pseudo}R^2 = \frac{\mathrm{G}}{\mathrm{G}+n}, \tag{4.18}$$

   where

   G is the value estimated in equation (4.17).

   Pseudo R^2 ranges between 0 and 1. When comparing the models, the model with higher pseudo R^2 will be preferred as it is the determination coefficient.

5. **Wald Statistic**

   To assess the significance of all coefficients we can use Wald stratistic as a significance test. It is also known as pseudo t statistic. The statistic is:

$$W = \frac{\widehat{\beta}_i}{Se(\widehat{\beta}_i)} \qquad (i = 1, ..., p+1), \tag{4.19}$$

   where

   $\widehat{\beta}_i$ is the maximum likelihood estimate of ith, and regression coefficient.

$Se(\widehat{\beta_i})$ is the standard error of i_{th} regression coefficient identified as:

$$Se(\widehat{\beta_i}) = \sqrt{covii} \tag{4.20}$$

The result of Wald statistic is assumed to be normally distributed. The result is asymptotic since the normal distribution provides a valid approximation for large n.

Binary logistic regression

Binary logistic regression is a type of GLM binary choice models. In logistic regression as the other binary choice models the dependent variable can take only two possible values and the distribution is assumed to be Bernoulli.

The link function of the logit model is:

$$\eta(\pi(x)) = \ln\frac{\pi(x)}{1-\pi(x)} = \beta X. \tag{4.21}$$

The link function in logistic regression is called logit. To predict the unknown parameters the cumulative logistic distribution function is needed:

$$F(x) = \eta^{-1}(x) = \Lambda(x) = \frac{1}{1-exp(-\beta X)} = \pi(x). \tag{4.22}$$

The score vector for logistic regression is:

$$\frac{\partial \ln L(\widehat{\beta})}{\partial \beta} = \sum_{i=1}^{n} x_i(y_i - \Lambda(x_i)). \tag{4.23}$$

By using iterative optimization methods the unknown parameters can be estimated. By Wald test and goodness of fit tests, the significance of the model can be checked. The significant logistic regression model can be applied to predict future values of observations.

Variable Selection in Logistic Regression

The main goal of statistical models is to build a parsimonious model that explains the variability in dependent variable. With less independent variables a model is generalized and interpreted more easily. Since the model with more independent variables may give more accurate results for within sample observations, the model will become specific for the observed data. For this purposes variable selection is needed.

In variable selection, the first thing to do is to check the significance of each coefficients. For binary choice models Wald statistic can be used for testing the significance. After estimating the test statistic we can conclude that if the significance $p<0.05$ for any coefficient of the variable with a 95% confidence level, then the contribution of the variable to the model is important. There is an important point that if the observations are inadequate, the model could be unstable and Wald statistic would be inappropriate [53].

After the significance is determined, the insignificant variables are eliminated and models without these variables are compared to the model with these variables by means of G likelihood ratio test. For the new model, the significance of variables should also be checked since the estimated values of coefficients are changed. To investigate the variables more closely the linearity of relation between logits and an independent variable can be checked via graphs.

After the significant variables are selected if the model includes much variables the variable selection methods such as stepwise variable selection can be used.

Stepwise Selection Method

The stepwise selection method is a variable selection method that is used to include and exclude a significant variable to the model by means of decision rules. It is also used in linear regression.

(a) **Forwardation**

Forward variable selection begins with including only the constant term and evaluation of log-likelihood value of this model. Then log-likelihoods of models for each variable are estimated. By these estimates, the value of likelihood ratio tests that the model containing constant term versus an independent variable is estimated:

$$G_j^{(0)} = 2(L_j^0 - L_0) \qquad (j = 1, \dots p), \tag{4.24}$$

where
$G_j^{(0)}$ is the likelihood ratio test in step 0 statistic and
L_j^0 is the log-likelihood of the model with j^{th} independent variable in step 0.

The significance value of G likelihood test is estimated as:

$$Pr(\chi_1^2 > G_j^{(0)}). \tag{4.25}$$

The most important variable is selected as the variable with smallest significance level. The most important variable is included to the model. If the significance level is smaller than α we stop in step 0, and otherwise the process continues.

If the process continues in the next step, the model with the variable in step 0 is taken as the reference model and second important variable that could be included to the model is tried to be selected. The likelihood ratio is estimated for the model with the most important variable versus the model with both the most important variable and another independent variable. In this step, the significance value is estimated for $p-1$ variables and the variable with minimum significance is included into the model. Then, the significance level is compared to the α; if it is smaller than α is stops. This process continues until all variables that are important by means of alpha criteria are included to the model.

The meaning of α significance value is different than in general since it determines the number of independent variables. It is recommended to take α between 0.15 and 0.20 [53].

(b) **Backwardation**

Backwardation begins with including all variables in the model. In the first step, one variables deleted and G is estimated for the models with all variables versus one variable deleted and also the significance value is estimated as in forwardation method. The variable with the maximum significance is deleted. This process is also continued until all variables with significance estimate higher than α are deleted from the model.

Binary probit regression

The probit regression is also a GLM model. As binary logistic regression the dependent variable can take only two possible values with Bernoulli distribution.

The link function for probit regression is,

$$\eta(\pi(x)) = \phi^{-1}(\pi(x)) = \beta X, \tag{4.26}$$

where $\phi^{-1}(.)$ is the inverse standard normal distribution function.

The link function in probit regression is called probit or normit. To estimate the unknown parameters, again the cumulative probit function is needed. It is identified as:

$$F(x) = \eta^{-1}(x) = \phi(x) = \pi(x). \tag{4.27}$$

Here $\phi(.)$ is the standard normal distribution function.

The score vector for probit is:

$$\frac{\partial logL(\widehat{\beta})}{\partial\beta} = \sum_{i=1}^{n} x_i\phi(x_i)\frac{y_i - \phi(x_i)}{\phi(x_i)(1-\phi(x_i))}. \tag{4.28}$$

After the significant probit model is found, it can be also used to predict future values of dependent variable.

Properties of logit and probit maximum likelihood estimators

i The maximum likelihood estimator $\widehat{\beta}$ is a consistent estimator for β. Consistency means that $\widehat{\beta}$ converges in probability to β:

$$\lim_{n\to\infty} P\{\|\widehat{\beta} - \beta\| > \varepsilon\} = 0, \tag{4.29}$$

where ε>0.

ii The $\widehat{\beta}$ is approximately normally distributed with mean vector is β and variance matrix is equal to the information matrix:

$$\widehat{\beta} \sim N(\beta, I(\beta)^{-1}). \tag{4.30}$$

The information function is:

$$I(\beta) = -E[\frac{\partial^2 l}{\partial\beta_i\beta_j}]. \tag{4.31}$$

iii The inverse information matrix is the Crammer Rao lower bound. Then $\widehat{\beta}$ is also asymptotically efficient which means that it is an unbiased estimator with minimum variance.

4.2 CLASSIFICATION AND REGRESSION TREES

The classification and regression trees (CART) model was first introduced by Breiman et al. (1984) [47]. It is a nonparametric technique alternative to regression type fitting that is used to split the observations into different classes by building binary decision trees depending on independent variables. Binary decision trees split sample into classes by starting from root node and by ending with homogenous sub samples. Unlike other classification techniques in CART, the decision rules are represented by tree.

When a decision tree is used to classify the data into classes the tree is called the classification tree. In classification tree there is a connection between categorical random variable and discrete, continuous or categorical random variables. If the dependent variable that we want to predict is a continuous random variable, then the decision tree is called the regression tree. For both trees the main goal is to produce accurate set of classifiers, to present the predictive structure of the classification problem [47]. The only difference is the scale of dependent variable and so the splitting rules are different. After trees are build they are used, to classify or predict new observations.

4.2.1 Classification Tree

To build a tree, historical data is required. The data used to build a tree are called the learning sample. In a classification tree, the learning sample consists of the measurement space (independent variable matrix) and classes (dependent variable vector):

$$\delta = \{(x_{11}, ..., x_{i1}, ..., x_{p1}, c_1), ..., (x_{1n}, ..., x_{in}, ..., x_{pn}, c_n)\}. \tag{4.32}$$

Here,

$i = 1, ..., p \; j = 1, ..., n$,

p is the number of independent variables,

n is the number of observations,

X is the measurement space:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{p1} \\ \vdots & \vdots & \dots & \vdots \\ x_{1n} & x_{2n} & \dots & x_{pn} \end{pmatrix}$$

and ζ is the vector of all possible classes:

$$\zeta = \{c1, ..., ck\}, \tag{4.33}$$

k is the number of all possible values the dependent variable can take.

By using the learning sample, we can construct the tree. The first step of the classification tree is to split the learning sample into classes up to the last observation. This process is called building T_{max} that is the maximum size tree.

After constructing the tree, we have to check whether the size of tree is optimal or not. If the tree is too large in size, although it has low misclassification rates, it can provide inaccurate results when applied to new data and it can be so complicated to interpret. If the tree is too small, then it may not use some information in the learning sample that leads to higher misclassification rates.

In the last step of the analysis, after building the right sized tree, we use the tree to classify new observations.

Constructing the tree

In classification tree, binary partitioning is used to construct a tree with nodes. In binary partitioning the sample is divided into two subsets for each node. It is a recursive process so that it is also known as recursive partitioning. For partitioning there is a function called classifier that is used for predicting the class of observation. It is required to determine how to split a data to produce a classifier.

Building the maximum tree means splitting the learning sample absolutely into homogeneous classes. The main problem of growing the tree is how to split the learning sample. For this purpose it is required to answer following questions:

i. How to split the data in to nodes.

ii. How to terminate a node.

iii. How to assign a class to a terminal node.

The tree starts by dividing the learning sample into two subsamples or the root node is divided into two sub nodes. Firstly, all observations of an independent variables are placed in the root node. It is assumed that all observations of that independent variable divide the learning sample and this is called as split. For all splits the learning sample is divided into two nodes as seen in Figure 4.1. The node t_p is the parent node and the nodes t_l and t_r are the left and right child nodes respectively.

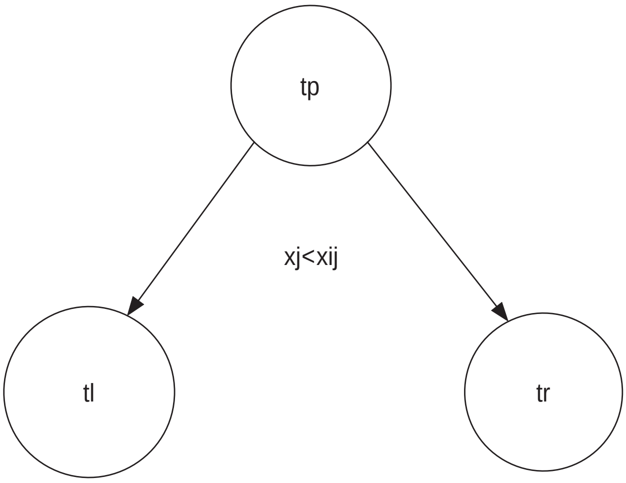


Figure 4.1: Splitting node

At each splits the standard questions $x_i \leq x_{ij}$ are asked and each observation with an answer "yes" are sent to the left child node and with answer "no" are sent to the right child node. Then for each split goodness of split criteria are estimated and according to the criteria a best split is selected. This process is repeated for all independent variables. After that all variables are observed according to their goodness of split criteria that reduces the impurity. According to the terminating rule, each node is determined as terminal or non-terminal and for each terminal node a class is assigned. This process continues until each observation in the learning sample is assigned to a class.

Splitting Rules

The splitting rule is used to split the data into smaller pieces with homogeneity. To determine the best split is to measure the goodness of split criteria. To calculate goodness of split it is required to define an *impurity function.* Impurity function measures the purity of any node. To estimate impurity function it is needed to know the proportion of class j in a node t:

$$\pi(j) := \frac{N_j}{N}, \tag{4.34}$$

where

N_j is the number of observations belong to the class j,

N is the number of observations and

$\pi(j)$ is the prior probability of class j.

We put,

$$p(j,t) = \pi(j)\frac{N_j(t)}{N_j}, \tag{4.35}$$

where

N_j(t) is the number of observations in node t belonging to class j.

$p(j,t)$ is the probability of an observation both in class j and in node t.

Furthermore,

$$p(t) = \sum_{\substack{J \\ j=1}} p(j,t), \tag{4.36}$$

where

$p(t)$ is the probability of any observation belonging to node t.

$$p(j/t) = \frac{p(j,t)}{t} = \frac{N_j(t)}{N(t)}, \tag{4.37}$$

where

$p(j/t)$ is the probability of class j in node t.

The impurity function is a function of $p(j/t)$ having the following properties [47]:

i When $p(1/t) = p(2/t) =, ..., = p(j/t) = \frac{1}{\mathrm{j}}$ the impurity function $\phi(\frac{1}{\mathrm{j}}, ..., \frac{1}{\mathrm{j}})$ takes its maximum value.

ii In a node t when there is only one majority of class,then the the impurity function $\phi(p(1/t), ..., p(j/t))$ takes its minimum value.

iii ϕ is a symmetric function of probability of class j in node t.

The impurity function is defined as follows:

$i(t){=}\phi(p(1/t), ..., p(j/t))$

When determining the best split we need to measure homogeneity of the nodes. The best splitting rule can be chosen by maximizing the change in impurity. The change in impurity is identified in (4.38) :

$$\Delta i(t) = i(t_p) - E[i(tc)] = i(t) - p_L i(t_L) - p_R i(t_R), \tag{4.38}$$

where

P_L is the proportion of observations in the left child node and

P_R is the proportion of observations in the right child node.

There are many splitting rules introduced but two of them Gini index and towing rule which are first used by Breiman et al. [47], they are the most frequently used ones.

1. **The Gini Index**

   Gini index is a node impurity index. It is the most commonly used rule to determine the best splitting rule. The impurity function in Gini criterion is as:

$$i(t) = \sum_{j \neq i} p(j/t)p(i/t); \tag{4.39}$$

   It can also be represented as:

$$i(t) = 1 - \sum_{j}^{J} p^2(j/t). \tag{4.40}$$

   For a binary class problem it can be represented as:

$$i(t) = 2p(1/t)p(2/t). \tag{4.41}$$

Since it is linear in proportion and quadratic, it gives more weight to the purer nodes.

The change in impurity is:

$$\Delta i(t) := i(t) - p_L i(t_L) - p_R i(t_R)$$

$$= 1 - \sum_j^J p^2(j/t) - p_L(1 - \sum_j^J p^2(j/t_L)) - p_R(1 - \sum_j^J p^2(j/t_R)),$$

or since p_L+p_R=1

$$\Delta i(t) = -\sum_j^J p^2(j/t) + p_L \sum_j^J p^2(j/t_L) - p_R \sum_j^J p^2(j/t_R). \tag{4.42}$$

As mentioned before, the best splitting rule is determined by maximizing the change in impurity. Then, the splitting problem is to select the split which maximizes (4.42).

2. **The Twoing Rule**

   Unlike Gini index *twoing criteria* for each node separate classes into two subclassses and each time estimates the impurity as if it is a binary class problem.

   For any node the best split is determined by maximizing the change in impurity. In twoing criteria there is no specific measure of impurity. The change in impurity is identified as:

$$\Delta i(t) = \frac{p_L p_R}{4}[\sum_{j=1}^J p(j/t_L)p(j/t_R)]^2. \tag{4.43}$$

   In Twoing rule, the two classes that will make up together the fifty percentage of the data is searched. If there are J classes, the algorithm searches for 2^{J-1} possible splits [47].

Terminating Rule

Terminating a node is a simple process. A splitting process stops when it is impossible to decrease the impurity of the tree. The change in impurity of the tree is defined as:

$$\Delta I(T) = \Delta i(t)p(t). \tag{4.44}$$

When drawing the tree, the terminal nodes are represented by rectangles and non-terminal nodes are represented by circles.

Class Assignment Rule

After determining the terminal nodes, there are two ways to assign a class to each terminal node.

i The first rule is known as the *plurality rule.* It is a rule of assigning a class to a node when if the majority of observations are belonging to that class. In other words, the class which has the highest probability is assigned. If (4.45) holds then j is assigned to node t.

$$j = \max p(j/t) \tag{4.45}$$

When the cost of misclassification is the same then we can use plurality rule.

ii The second rule assigns a class to a node which minimizes the expected misclassification cost. If the cost of misclassification of classes are different, we cannot use the plurality rule.

The expected misclassification cost is defined as:

$$r(t) := \sum_{j=1}^{J} c(i/j)p(j/t), \tag{4.46}$$

where

$c(i/j)$ is the cost of assigning a class i to a class j observation.

Determining the optimum size of tree

The basic problem of trees is their complexity and non-accuracy. Then by using an accuracy criterion, it is required to determine the right sized tree which is accurate with less complexity.

In determining the size of tree, it does not work well to use stop splitting rules with determining thresholds. It is assigned a node as terminal if it decreases the change in impurity by a small amount but the child nodes may decrease the impurity with large amounts. By terminating the node, the good splits t$_L$ and t$_R$ can be lost. Since then pruning can be used to form the subtrees.

Minimal Cost Complexity Pruning

The minimal cost complexity is an effective way to find a right sized tree. In this method, subsequences of the T_{max} are formed.

Breiman et al. [47] stated that a minimal cost complexity measure compares the complexity and accuracy rate of trees. The complexity of a tree is determined by the number of terminal nodes.

The cost complexity measures is identified as:

$$R_\alpha(T) = R(T) + \alpha\tilde{T}, \tag{4.47}$$

where

$\alpha \geq 0$, T$\leq T_{max}$,

$R(T)$ is the misclassification cost of tree,

α is the complexity parameter; it is also known as the penalty of additional terminal nodes and

$\tilde{T}$ is the number of terminal nodes of the tree.

The cost complexity measure is the linear combination of misclassification cost and complexity measure. If α=0, the optimal tree is the T_{max}, if $\alpha > 0$, the optimal tree is a sub-sequence of T_{max}.

Firstly, for each value of α the subtrees are formed as T_{max},T_1,...,t_0 with decreasing number of terminal nodes. "t_0" is the root node of the tree. Secondly, the cost complexity measure is determined for each sub-trees. Then, subtree from these sequence that minimizes the cost complexity measure is selected as the right sized tree.

When we are using this measure, we are also finding a tree with optimum misclassification rate which is very important for the future predictions. The selection of misclassification rate can be obtained from both test sample and cross-validation approaches.

1. **Test Sample Technique**

   In test sample approach the learning sample is divided into two sub-samples L_1 and L_2. Generally L_2 is taken as the one third of the observations in the learning sample [47]. In the fist step, by L_1 is used to construct T_{max} and its subsequences. For each subsequence the test sample L_2 is used to predict the values of the dependent variable. The realized values of the dependent variables are known before, so we can count the misclassified classes for each node and so for each tree. Also the probability of classifying a class j observation as class i for each tree can be estimated as follows:

$$p(i/j,T) = \frac{N_{ij}^{(2)}}{N_j^{(2)}}, \tag{4.48}$$

where

$N_{ij}^{(2)}$ is the number ith class observations that are assigned as class i in L_2,

$N_j^{(2)}$ is the number of jth class observations in L_2.

The misclassification cost of a tree is estimated as:

$$R^{ts}(T) := \frac{1}{N^{(2)}} \sum_{j=1}^{J} c(i/J) N_{ij}^{(2)} \tag{4.49}$$

The misclassification cost is a measure of tree accuracy. The tree which has minimum misclassification cost can be used as the right sized tree. Also the test sample of misclassification cost estimate can be used in minimum cost complexity measure.

2. **V-Fold Cross Validation Technique**

   In v-fold cross validation technique the learning sample is divided into v sub-samples nearly equal sizes as: L_1,..., L_v v=(1,...,V)

   Each sub-samples are used as the test samples and the sample without vth sample is used as the learning sample. $L - L_v$ is used to construct T_{max} and the sub-sequences. Now, L_v is used to predict the dependent variable. For each tree the total number of observations that are miss-classified as class i can be counted as:

$$p(i/j,T) = \frac{N_{ij}}{N_j}, \tag{4.50}$$

$$N_{ij} = \sum_{v=1}^{V} N_{ij}^{v}. \tag{4.51}$$

Here,

N_{ij}^{v} is the number of class j observation in L_v classified as i,

N_j is the number of jth class observations.

The misclassification cost of any tree is identified as:

$$R^{cv}(T) := \frac{1}{N}\sum_{j=1}^{J} c(i/J)N_{ij} \tag{4.52}$$

The tree with minimum misclassification cost can be taken. Also as test sample approach the cross validation estimate of misclassification cost can be used in cost complexity measure.

4.2.2 Regression Tree

As mentioned before when a dependent variable is continuous, the decision tree is called a regression tree. The construction of the tree is similar, but there are some differences. The main difference is the scale of dependent variable, so instead of classes, numerical predicted values of dependent variable are tried to be assigned to each terminal node by means of independent variables. For this reason the splitting rule also differs.

Constructing the tree

The main steps in construction of regression tree is the same as the classification tree. In the first step, the questions to split learning sample are asked and based on goodness of splitting rule; the best splits are determined. As in classification tree, binary partitioning is used to split. After a maximum tree is constructed by using cost-complexity, cross validation or test sample, approaches the optimal sized tree is constructed.

Splitting Rule

In a regression tree the splitting rule depends on the within node sum of squares. The best split is selected as the one that most reduces the average within the node sum of squares or the re-substitution estimate identified as:

$$R(t) := \frac{1}{N}\sum_{xn\epsilon t}(y_n - \overline{y(t)})^2. \tag{4.53}$$

Here,

$\overline{y(t)})$ is the average of observations in node t,

y_n are the dependent variables in node t.

The best splitting rule is defined similar to the classification tree that maximizes the change in the re-substitution estimate [47]:

$$\Delta R(t) := R(t) - R(t_L) - R(t_R). \tag{4.54}$$

The alternative way is to use weighted variances. In this method, weights are the proportions of observations that are in the right and left child nodes. The variance of the node t is:

$$S^2(t) = \frac{1}{N(t)} \sum_{i=1}^{N(t)} (y_n - \overline{y(t)})^2. \tag{4.55}$$

The change in weighted variance is defined as :

$$\Delta S^2(t) := S^2(t) - p_L S^2(t_L) - p_R S^2(t_R). \tag{4.56}$$

The split that maximizes the change in weighted variance is used as the best split.

Terminating Rule

As in classification tree in regression tree to stop a node again terminating rule is needed. Different from classification tree a node terminates when following condition is satisfied:

$N(t) < N(min)$

$N(min)$ is generally taken as 5 [47].

Assigning a Value

The assigned value in the regression tree is the one that minimizes the misclassification estimate in a node. The estimate of predicted value that minimizes the sum-of-square is the mean of the dependent variable in that node:

$$\overline{y(t)} = \frac{1}{N(t)} \sum_{n=1}^{N(t)} y_n. \tag{4.57}$$

The mean of the observed dependent variable in a terminal node is assigned as the predicted value.

Determining the right-sized tree

After building the maximum tree to determine the optimum sized tree, the pruning algorithm can be used. For regression tree an error-complexity measure is defined instead of cost-complexity measure, since the accuracy in regression tree is identified by means of mean square error estimates. The measure is defined as:

$$R_\alpha(T) = R(T) + \alpha\tilde{T}, \tag{4.58}$$

where

$$R(T) = \sum_{t=1}^{T} R(t). \tag{4.59}$$

As in classification tree for each α the sub-sequence T_{max}, $T_1, ..., t_0$ is formed and for each subsequence error complexity measure is estimated and the tree that minimizes the measure is used as the right-sized tree.

1. **Test Sample Technique**

   Test sample is estimated in the same manner in classification tree. The learning sample is divided into two sub-samples. L_1 is used to form the sub-sequence and L_2 is used to estimate the accuracy such as:

$$R^{ts} = \frac{1}{N_2}\sum_{n=1}^{N_2}(y_n - \overline{y})^2. \tag{4.60}$$

2. **Cross Validation**

   Randomly learning sample is divided into v subsets. Each time L_v is used to estimate the performance measure and $L - L_v$ is used to prune the tree. The cross validation estimate is:

$$R^{cv}(T) = \frac{1}{N}\sum_{v=1}^{V}\sum_{n=1}^{N_v}(y_n - \overline{y})^2. \tag{4.61}$$

4.3 DISCRIMINANT ANALYSIS

Discriminant analysis (DA) is a multivariate statistical technique that uses the independent variables to separate the observations into groups. The main goal of the analysis is to assign predefined groups for observations. It can be used in two or more than two group separation problems. It has two main steps [54]:

i To define a function that is used to discriminate the groups (discrimination).

ii To classify the out-of-sample observations to groups by minimizing the classification error (classification).

"In credit the scoring problem, the response variable is binary; for this reason "Fisher's Linear Discriminant Analysis" can be used to classify.

4.3.1 Linear Discriminant Analysis for Two Group Seperation

In Linear Discriminant analysis (LDA] the problem is to identify the linear surface into two groups. It is a statistical decision making based on the differences in means. The independent variables are linearly combined to form the dependent variables of two groups. The groups are tried to separated as well as possible.

Discrimination

In LDA, for each group there is a linear discrimination function. In this way, the analysis has the advantage of dimension reduction when interpreting the results. By discrimination functions the analysis is transformed to one dimensional simplest form.

The discrimination function is defined as given in (4.62) and (4.63):

$$y^{(1)} = w_1 x_1^{(1)} + \ldots\ldots + w_p x_p^{(1)}, \tag{4.62}$$

$$y^{(2)} = w_1 x_1^{(2)} + \ldots\ldots + w_p x_p^{(2)}, \tag{4.63}$$

where

$w_{p\times 1}$ is the weight vector of dimension p,

$x^{(1)}$ is the independent variables matrix for group 1 observations and

$x^{(2)}$ is the independent variable matrix for group 2 observations.

Fisher introduce a criteria used in estimation of weight vector. The criterion which has the property of differences in groups is defined as:

$$F := \frac{w^T B w}{w^T W w}, \tag{4.64}$$

where

B is the between group sum of square and

W is the within group sum of square.

We put:

$$B := \sum_{k=1}^{2} n_k (\overline{x_k} - \overline{\overline{x}})(\overline{x_k} - \overline{\overline{x}})^T, \tag{4.65}$$

$$W := \sum_{k=1}^{2} \sum_{i=1}^{n_k} (x_i - \overline{x_k})(x_i - \overline{x_k})^T, \tag{4.66}$$

$$\overline{x_k} = \frac{1}{n_k} \sum_{i=1}^{n_k} x_i \qquad \mathit{and} \tag{4.67}$$

$$\overline{\overline{x}} = \frac{1}{n} \sum_{i=1}^{n} n_k \overline{x_k}. \tag{4.68}$$

The optimal weight vector is found as the one that maximizes the Fisher's criteria. This is a maximization problem. The criteria is constant with respect to the establishment of the weight vector on a new scale $w \rightarrow \alpha w$, so for simplicity we can be take the weight vector that satisfies the denominator $w^T W w$=1 as a scalar. Thus, the problem becomes a minimization problem if defined as follows:

$$\min_{w} \qquad -\frac{1}{2} w^T B w \tag{4.69}$$

$$\text{st} \qquad w^T W w = 1 \tag{4.70}$$

The Lagrangien function can be written as:

$$L(w, \lambda) := -\frac{1}{2} w^T B w + \frac{1}{2} \lambda (w^T W w - 1), \tag{4.71}$$

where λ is called lagrange multiplier vector.

The first-order optimality conditions, state the existence of w and λ are such that:

$$Bw + \lambda W w = 0 \qquad and \tag{4.72}$$

$$(W^{-1} B - \lambda I) w = 0. \tag{4.73}$$

The problem becomes an eigenvalue decomposition problem. The eigenvector of an eigenvalue that maximizes the criteria is the optimal weight vector. The solution of the problem is given by:

$$\widehat{w} := d S_{pooled}^{-1}, \tag{4.74}$$

where

$$S_{pooled} := \frac{1}{n_1 + n_2 - 2} W \tag{4.75}$$

an d is the mean difference vector defined as:

$$\mathbf{d} := \begin{pmatrix} \overline{x_1^{(1)}} - \overline{x_1^{(2)}} \\ \vdots \\ \overline{x_p^{(1)}} - \overline{x_p^{(2)}} \end{pmatrix}$$

The linear discrimination function then defined as

$$\widehat{y} = [x^1 - x^2] S_{pooled}^{-1} X. \tag{4.76}$$

The significance of discrimination function can be tested by Mahalanobis distance measure. It is a separation criterion of the model defined as:

$$D^2 = [x^1 - x^2]^T S_{pooled}^{-1} [x^1 - x^2] \tag{4.77}$$

or

$$D^2 = d^T S_{pooled}^{-1} d. \tag{4.78}$$

To test the significance we can define Hotelling T^2 statistic:

$$T^2 = \frac{n_1 n_2}{n} D^2. \tag{4.79}$$

Here,

n_1 is the number of observations in group 1,

n_2 is the number of observations in group 2 and

n is the total number of observations (n_1+n_2).

To test the significance of Hotelling T^2 statistic, it is needed to convert it into known test statistic to check the table value. The F distribution derived from Hotelling T^2 is defined as:

$$F = \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} T^2. \tag{4.80}$$

The significance of F statistic is evaluated by comparing it with the F distribution table value with p, (n_1+n_2-p-1) degrees of freedom. If the significance value holds $p<0.05$, with 95 % of confidence, then it is concluded that the model is efficient in separating the groups, in other words weight vector is significant.

Classification

As mentioned before, the next step, after the determination of the discriminant function and significance check consists of the assignment of new observations to the groups. In Fisher's problem, the independent variables are converted to $\widehat{y}$ as follows:

$$\widehat{y} = w^T X. \tag{4.81}$$

The mid-point estimation of the predicted value $\widehat{y}$, which is used as a tool in assignment rule, is defined as:

$$m = \frac{1}{2}[x^1 - x^2]^T S_{pooled}^{-1}[x^1 + x^2] \tag{4.82}$$

Suppose new observation has an independent variable vector represented as x_0. The estimated dependent variable for that observation can be estimated as

$$y_0 = [x^1 - x^2]^T S_{pooled}^{-1} x_0. \tag{4.83}$$

The assignment rule is defined as:

$y_0 - m \geq 0$: classify the observation as group 1,

otherwise: classify the observation as group 2.

4.4 NONPARAMETRIC AND SEMIPARAMETRIC REGRESSION

Density estimation is a tool to examine the structure of data. When building a model, to capture the features such as skewness or kurtosis, visualizing the density of observations is needed. It can be a preferable way to summarize the outcome of Monte Carlo simulation [55]

Univariate Kernel Density Estimation

The aim of kernel smoothing is to estimate approximately the probability density function of any random variable without any known functional form. It is a flexible method when summarizing or constructing a model. Histogram is another a summary tool that visualizes the density, but since it is a very rough method, kernel estimation is preferable. The approximate probability density function with kernel smoothing is defined as

$$\widehat{f(x)} := \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x_i - x}{h}) = \frac{1}{nh}\sum_{i=1}^{n} K(\psi_i), \tag{4.84}$$

where

x is a random variable with unknown density function,

$\{x_i : i = 1, ..., n\}$ the iid sample of random variable x,

h is the smoothing parameter, bandwidth or window-width and

$K(.)$ is the kernel function which has the following properties:

i Generally it is symmetric around zero,

ii $\int_{-\infty}^{\infty} K(\psi)d\psi = 1,$

iii $\int_{-\infty}^{\infty} \psi K(\psi)d\psi = 0 \qquad and$

iv $\int_{-\infty}^{\infty} \psi^2 K(\psi)d\psi < \infty,$

Most popular kernel functions are:

1. **Standard Normal Kernel**

$$K(\psi) = (2\pi)^{\frac{-1}{2}} exp(\frac{-1}{2}\psi^2). \tag{4.85}$$

2. **Uniform Kernel**

$$K(\psi) = \frac{1}{2a} \qquad (-a < \psi < a). \tag{4.86}$$

3. **Epanechnikov Kernel**

$$K(\psi) = 0.75(1 - \psi^2)I(\mid \psi \mid \leq 1); \tag{4.87}$$

here,

$I(.)$ is the indicator function that takes the value 1 when $|\psi|$≤1,and it takes the value 0 otherwise.

Asymptotic Properties of Kernel Density Estimation

1. **Asymptotic Unbiasedness**
For finite samples the density estimation is biased but it is unbiased asymptotically.

$$\lim_{n\to\infty} E[\widehat{f}] = f \tag{4.88}$$

and

$$\sup_{x} \mid E[\widehat{f} - f] \mid = 0 \qquad as \qquad n \to \infty \tag{4.89}$$

here n is the number of observations.

2. **Asymptotic Consistency**

   Estimator $\widehat{f}$ converges in mean square to f, since it is asymptotically unbiased

$$MSE(\widehat{f}) \to 0 \qquad as \qquad n \to \infty \tag{4.90}$$

where

$$MSE(\widehat{f}) := Bias(\widehat{f})^2 + variance(\widehat{f}), \tag{4.91}$$

$$Bias(\widehat{f}) := E[\widehat{f} - f] = \int K(\psi)[f(h\psi + x) - f(x)]d\psi, \tag{4.92}$$

$$Variance(\widehat{f}) := \frac{1}{nh}\int K^2(\psi)f(h\psi + x)d\psi - \frac{1}{n}[\int K(\psi)f(h\psi + x)d\psi]^2 \qquad and \tag{4.93}$$

n is the number of observations.

3. **Asymptotic Normality**

   If there exist any δ such that $\int K(\psi)^{2+\delta}d\psi < \infty$. Then,

$$\sqrt{nh}[\widehat{f} - E[\widehat{f}]] \sim N(0, f(x)\int K^2(\psi)d\psi) \qquad as \qquad n \longrightarrow \infty \tag{4.94}$$

The Selection of Smoothing Parameter

Selection of the smoothing parameter is a challenging problem in kernel density estimation since it changes the range of variability of the estimation. When the smoothing parameter h is too small the bias of the estimator will be small but variance will be large. When h is too large then bias will be large although variance is small. Bias and variance are undesirable properties of any statistical estimator. For this purpose, the selection of h minimizing any criteria that depends both on bias and variance would be preferable. The most popular way used in selection is to minimize the criteria MISE (Mean Integrated Squared Error) or the AMISE (Approximate Mean Integrated Squared Error):

$$MISE := \int [Bias(\widehat{f}^2) + Variance(\widehat{f})]dx, \tag{4.95}$$

$$AMISE := \frac{h^4}{4}(\int \psi^2 K(\psi)d\psi)^2(\int (f''(x))^2 dx) + \frac{1}{nh}\int f(x)dx \int K^2(\psi)d\psi. \tag{4.96}$$

The smoothing parameter that minimizes AIMSE is given in [55]

$$h := (\frac{\int K^2(\psi)d\psi}{(\int \psi^2 K(\psi)d\psi)^2(\int (f''(x))^2 dx)})^{\frac{1}{5}} n^{\frac{-1}{5}}. \tag{4.97}$$

Multivariate Kernel Density Estimation

When there is not single independent variable, there are independent variables the kernel smoothed multivariate densities can also been constructed. Suppose $\widehat{f(x), H}$ is the multivariate density estimation [56]:

$$\widehat{f(x,H)} = \frac{1}{n}\sum_{i=1}^{n} K_H(x - x_i), \tag{4.98}$$

where

$x = (x_1, ..., x_d)^T$ independent random variables,

$x_i = (x_{i1}, ..., x_{id})$ (i=1,...,n) values of random variables,

H is invariant $d \times d$ symmetric smoothing parameter matrix and

$K_H(.)$ is the multivariate kernel function defined as

$$K_H(x) = |H|^{\frac{-1}{2}} K(H^{\frac{-1}{2}} x) \tag{4.99}$$

When determining the smoothing parameter, for multivariate kernel regression it is also needed to minimize criterion at (4.100):

$$MISE := E[\int [\widehat{f(x,H)} - f(x)]^2]dx = \int variance(\widehat{f(x,H)})dx + \int Bias^2\widehat{f(x,H)}, \tag{4.100}$$

where

$$Bias\widehat{f(x,H)} = (K_H(x)of(x)) - f(x), \tag{4.101}$$

$$variance\widehat{f(x,H)} = n^{-1}[(K_H^2(x)of(x))(K_H(x)of(x))^2] \tag{4.102}$$

The most commonly used smoothing parameter matrix is derived by Wand and Jones [56] defined as:

$$H = \text{diagonal}(h_1^2, h_2^2, ..., h_d^2), \tag{4.103}$$

where $h_1, ..., h_d$ are the univariate smoothing parameters.

The more specific way to estimate the kernel density estimation is to use the multiplicative kernel [46]

$$f(x,H) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_1, ..., h_d} K(\frac{x_i - x}{h_1}, ..., \frac{x_i - x}{h_d}); \tag{4.104}$$

here,

H=dioganal$(h_1, ..., h_d)^T$,

K_H^{ψ}=$k(\psi_1), ..., k(\psi_d)$ and

$k(.)$ is the univariate kernel function.

4.4.1 Non-Parametric Regression by Multivariate Kernel Smoothing

The weakness of parametric regression consists in its assumptions. When estimating the regression parameters, it is assumed that the independent variables comes from a multivariate normal distribution. Sometimes this assumption cannot been satisfied for such cases; non-parametric regression is preferred. In non-parametric regression, the joint distribution functions of independent variables are not known as parametric function. The regression model defined by means of the conditional mean of dependent variable is

$$Y = E[Y|X = x] + \varepsilon = m(x) + \varepsilon, \tag{4.105}$$

where ε is the error term which has the following properties:

$E[\varepsilon|X] = 0 \qquad and$

$E[\varepsilon^2|X] = \sigma^2(X).$

Non-parametric estimator of m(x)

The non-parametric estimator of the conditional moment is

$$\widehat{m} = \sum_{i=1}^{n} v_n(x_i, x)y_i; \tag{4.106}$$

here,

$v_n(x_i, x)$ is the weight for the observations has the following properties:

$v_n(x_i, x) \geq 0and$

$\sum_{i=1}^{n} v_n(x_i, x) = 1.$

The most commonly used estimator for the weight of any observations is "Nadaraya Watson Kernel Estimator" defined as

$$v_n(x_i, x) = \frac{K\frac{x_i - x}{h}}{\sum_{i=1}^{n} K\frac{x_i - x}{h}} \tag{4.107}$$

The multivariate kernel estimator of conditional mean estimated by using Nadaraya Watson Kernel estimator based on locally weighted averages is:

$$\widehat{m}(x, H) := \frac{\sum_{i=1}^{n} K_H(x_i - x)yi}{\sum_{i=1}^{n} K_H(x_i - x)}. \tag{4.108}$$

The properties of the estimator defined by Ruppert and Watson 1994 [57] are

$$Bias(\widehat{m}) := \frac{H^2}{4}\{m^{"}(x) + 2\frac{m^{'}(x)f^{'}(x)}{f(x)}\}\mu_2(k), \tag{4.109}$$

where $\mu_2(k)$ is the second moment for kernel density estimation and

$$Variance(\widehat{m}) = \frac{1}{nH}\frac{Var(Y|X)}{f(x)}||K||_2^2. \quad (4.110)$$

The solution of Naraya Watson estimator is an least square minimization problem. Least square of moment estimation is [57]:

$$\min_{\widehat{\beta}}(Y - X\beta)^T W(Y - X\beta), \quad (4.111)$$

where $w = \text{diagonal}(K_H(x_i - x, ..., K_H(x_n - x))$.

The solution is:

$$\widehat{\beta} = (X^T W X)^{-1} X^T W Y. \quad (4.112)$$

The conditional moment estimator is:

$$\widehat{m}(x) = e^T (X^T W X)^{-1} X^T W Y, \quad (4.113)$$

where $e^T = (1, ..., 0)$,

$\widehat{m}(x) = (\widehat{m}(x_1), ..., \widehat{m}(x_n))^T$.

The conditional mean is also known as weighted average since kernel function is defined as the weight of observations.

4.4.2 Semiparametric Regression

Semiparametric regression is a tool that concern with flexible non-linear functional relations. In semiparametric regression, there are assumptions between non-parametric and parametric models. In other words, they are models in which there is an unknown function and unknown finite dimensional functions. There are a lot of semiparametric models and the models for binary response variable are generally the extensions of generalized linear models. The one mostly used in credit scoring area is called a *generalized partially linear model* that is separating the explanatory variables into two.

Generalized partially linear models

Generalized partially linear model (GPLM) is the extension of generalized linear models. In GPLM the independent variables are separated into two [58]. Here, Z indicates p dimensional discrete independent variables and T indicates the q dimensional continuous independent variables. The model is the conversion of $Z'\beta+T'\alpha$ to a partially linear form $Z'\beta + m(T)$. The conditional mean function is as:

$$E[Y/Z,T] = GZ'\beta + m(T), \tag{4.114}$$

where,

G is the logistic link function,

$m(.)$ is the unknown parametric function and

β is the unknown finite dimensional parameter vector.

The unknown β and $m(.)$ should be estimated. The estimation of unknown function and parameters are based on quasi-likelihood function. A quasi-likelihood function is defined as:

$$Q(\mu; y) = \int_{\mu}^{y} \frac{(s-y)}{V(s)} ds, \tag{4.115}$$

where

$E[Y|Z,T] = GZ'\beta + m(T)$

$var[Y|Z,T] = \sigma^2 V(\mu)$,

σ is the scale parameter estimated as:

$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i-\widehat{\mu_i})^2}{V(\widehat{\mu_i})}$.

Since both $m(.)$ and β are unknown, by optimization methods $\widehat{\beta}$ is obtained for any specific $m(.)$ and $\widehat{m(.)}$ is obtained for known β. For the estimation of $m(.)$, a locally weighted kernel quasi-likelihood function is used:

$$L_l(m(.), \beta) = \Sigma_{i=1}^{n} K_H(t_i - t).Q[GZ'\beta + m(T); y]dt \tag{4.116}$$

Here,

$K_H(.)$ is the kernel function used as local weights.

For the estimation of β, a quasi-likelihood function is used

$$L(m(.), \beta) = \Sigma_{i=1}^{n} Q[GZ'\beta + m(T); y] \tag{4.117}$$

Firstly β is fixed and assumed to be known, the function $m_\beta(.)$ is estimated depending on β by maximizing the local quasi-likelihood function

$$\widehat{m_\beta} = \arg\max_{m} L_l(m, \beta) \tag{4.118}$$

Then, by using $\widehat{m_\beta}(.)$ the quasi-likelihood function for β is constructed and optimized to form β:

$$\beta = \arg\max_{\beta} L(m, \beta). \tag{4.119}$$

To find optimum $m(.)$ and β the following should be satisfied.

$$\Sigma_{i=1}^{n} Q_i^{'}\{Z_i^{'}\beta + m(T_i)\}K_H(t_i - t_j) = 0 \tag{4.120}$$

and

$$\Sigma_{i=1}^{n} Q_i^{'}\{Z_i^{'}\beta + m(T_i)\}\{Z_i + m(Ti)^{'}\} = 0 \tag{4.121}$$

To solve this problem Newton Raphson type algorithm can be used.

Chapter 5

VALIDATION TECHNIQUES

5.1 CUMULATIVE ACCURACY PROFILE CURVE

A *cumulative accuracy profile* (CAP) curve is a visual tool used by Moody's to asses the scoring model performance. Moody's uses the term Cumulative Accuracy Profile since it represents the cumulative probabilities of default over the entire population, as opposed to the non-defaulting population" [59, 60]. This is also known as "Gini Curve", "Power Curve" or "Lorenz Curve". It measures how a scoring model is successful in assessing the individuals with low creditworthiness in bad classes. A high scoring class represents the low default probability of default.

In building CAP curve, the first step is to order all the observations from high rating class to low rating class. For each scoring class, the following probabilities are estimated:

$$P_D^s = \frac{N_D^s}{N_D}, \tag{5.1}$$

$$P_{ND}^s = \frac{N_{ND}^s}{N_{ND}}, \tag{5.2}$$

$$P_T^s = \pi P_D^s + (1-\pi) P_{ND}^s \qquad and \tag{5.3}$$

$$\pi = \frac{N_D}{N}, \tag{5.4}$$

where

P_D^s is the probability of defaulter has a score s

P_{ND}^s is the probability of defaulter has a score s

P_T^s is the probability of an individual has a score s

π is the probability of defaulters in the sample

N_D^s is the number of defaulter in score s

N_{ND}^s is the number of non-defaulter in score s

N_T^s is the number of all observations

Then cumulative probabilities can be estimated for each score as follows:

$$C_D^s = \sum_{s=1}^{S} P_D^s, \tag{5.5}$$

$$C_{ND}^s = \sum_{s=1}^{S} P_{ND}^s \qquad \mathit{and} \tag{5.6}$$

$$C_T^s = \sum_{s=1}^{S} P_T^s, \tag{5.7}$$

where

C^s_D is the probability of defaulter has a score at least s,

C^s_{ND} is the probability of defaulter has a score at least s and

C^s_T is the probability of an individual has a score at least s.

The CAP curve is a graph of all points of C^s_D and C^s_T for each scoring class. The random model with no discriminatory power is represented by a straight line. A perfect model exactly assigns all defaulters to the lowest class. Generally, the performance of a real scoring model is in between these two models.

When comparing two or more scoring models, CAP curve can only be a visualization tool and can be misleading. Then, summary measure is needed to evaluate the performance. The success of any model can be summarized and evaluated by accuracy ratio (AR). It includes both Type I and Type II errors defined as [59]

$$AR = \frac{a_r}{a_p}, \tag{5.8}$$

where

a_r area between scoring model and random model and

a_p area between perfect model and random model.

AR ranges between 0 and 1; the higher the accuracy ratio the more successful the scoring model. The model with AR closest to 1 is the best model. It is designed to determine whether the model is better than the random model with zero information or not. It can be roughly estimated as

$$AR = 1 - 2\sum_{s=1}^{S} P^s_D [\frac{C^{s-1}_{ND} + C^s_{ND}}{2}]. \tag{5.9}$$

5.2 RECEIVER OPERATING CHARACTERISTIC CURVE

The receiver operating characteristic (ROC) curve is a visual tool that represents the possible distributions of scores for defaulting and non-defaulting applicants. Suppose a decision maker tries to predict a new observation's future behavior in the next period, he or she should determine a cut-off value and classify the observations as a potential defaulter when the score is lower than that cut-off value, or classify the observation as a potential non-defaulter when a score is higher than the cut-off value. There will be four possible scenarios summarized in Table 5.1

Table 5.1: Possible scenarios for payment

Observations	Score<x	Score>x
Default	true (A)	misclassified (B)
Non-Default	misclassified (C)	true (D)

Hit rate is defined as:

$$HR(x) := \frac{H(x)}{N_D}, \tag{5.10}$$

where,

$HR(x)$ is the hit rate,

$H(x)$ is the number of defaulters predicted correctly with the cut-off value x and

N_D is the total number of defaulters in the sample.

False alarm rate is defined as:

$$FAR(x) := \frac{FR(x)}{N_{ND}}, \tag{5.11}$$

where,

$FAR(x)$ is the false alarm rate,

$FR(x)$ is the number of non-defaulters predicted incorrectly as defaulters with the cut-off value x and

N_{ND} is the total amount off non-defaulters.

The ROC curve is a graph of $HR(x)$ versus $FAR(x)$ drawn for each cut of value. To compare the scoring models, the model with steeper ROC curve is the better one but make a decision only by means of ROC curve may be misleading since it may be difficult to visualize the difference between curves. For this purpose, summary statistics are needed. The most common summary statistic is the area under curve (AUC). The worst model would have AUC equals to 0.5, the best model has AUC equals to 1. It is estimated as follows [59, 60]:

$$AUC = \int_0^1 HR(FAR)d(FAR) \tag{5.12}$$

It can be approximately calculated as:

$$AUC = 1 - \sum_{s=1}^{S} P_D^s[\frac{C_{ND}^{s-1} + C_{ND}^s}{2}]. \tag{5.13}$$

There is a connection between the AUC and AR defined as:

$$AR = 2AUC - 1. \tag{5.14}$$

Pietra Index

Pietra index is also a summary measure of the ROC curve. Geometrically, it represents two times the maximum area of the triangle can be drawn between ROC curve and unit square's diagonal. It can be estimated as follows:

$$Pietra - Index := \frac{\sqrt{2}}{4} max_s \mid HR(s) - FAR(s) \mid \tag{5.15}$$

The index is between 0 and 1. The model with Pietra index closest to 1 is the best model, the model with index equal to 0 have no discriminatory power. Sometimes it is approximately estimated as the maximum distance between cumulative frequency distributions of default and non-default observations. To evaluate the performance of Pietra index, the Kolmograv Simirnow non-parametric test is used.

The approximate estimate of Pietra Index is defined as:

$$P = max_s \{ \mid CP_D(s) - CP_{ND}(s) \mid \} : \tag{5.16}$$

and the hypothesis is considered as:

H_0:The difference between defaults and non-defaults is insignificant.

H_1:The difference between defaults and non-defaults is significant.

The test statistic is defined as

$$KS = \frac{D_q}{\sqrt{Np(1-p)}}, \tag{5.17}$$

where

N is the number of observation in the sample,

p is the default probability in the sample and

D_q is the significance value KS statistic.

When the difference is greater than or equal to the table value then H_0 hypothesis is rejected with q significance level. When we reject H_0, we conclude that the difference between default and non-default observations is significant and the model is successful in discriminating the observations.

Bayesian Error Rate

It is also a discriminatory power measure of ROC curve depending on both Type I and Type II errors. It is defined as:

$$\text{Error Rate} := \min_s \{ \mid p(1 - HR(s)) + (1-p)FAR(s) \mid \}, \tag{5.18}$$

where

$1 - HR(s)$ is the α (Type I) error,

$FAR(s)$ is the β (Type II) error and

p is the default probability of the sample.

Error rate depends on the samples' default and non-default probabilities; so it is weaker than the AUC and Pietra index. The smaller the error rate, the more accurate the model is.

5.3 INFORMATION MEASURES

Information measure used for all kind of models without any assumption about the distribution of the variables. It is a measure of uncertainty information [59, 60].

Information of default probability without taking into account the scores is:

$$I(p) = -p\ln(p) + (1-p)\ln(1-p). \tag{5.19}$$

Information takes into account the model and it is defined by,

$$I(s) = -P(D|s)\ln(P(D|s)) + P(ND|s)\ln(P(ND|s)). \tag{5.20}$$

The conditional information measure is the expected value of information of the model:

$$I(score) = -E[P(D|s)\ln(P(D|s)) + P(ND|s)\ln(P(ND|s))] =$$

$$-\sum_{s=1}^{S} P(s)(P(D|s)\ln(P(D|s)) + P(ND|s)\ln(P(ND|s))).$$

When evaluating the predictive power of the model, any interpreting by means of information measures is not adequate. Therefore, summary measures such as conditional entropy ratio (CIER) and Kullback Leibler Distance are used for comparison.

5.3.1 Kullback Leibler Distance

Kullback Leibler (KL) distance is the measure of difference between information with and without any model:

$$K - L \qquad distance := I(p) - I(score). \tag{5.21}$$

The model with maximum distance measure is the best one since the distance is interpreted as the information added to the prediction by the model.

5.3.2 Conditional Information Entropy Ratio

Conditional Information Entropy Ratio(CIER) is the normalized measure of the Kullback Leibler distance. It represents the reduction in uncertainty defined as [59, 60]:

$$CIER = \frac{I(p) - I(score)}{I(score)}. \tag{5.22}$$

Since it represents the reduction in uncertainty, the higher the ratio the more accurate the model is. The model with CIER equal to 0 is the model with no discriminatory power; this means that this model does not add any information to the default event.

5.4 BRIER SCORE

Brier Score is the measure of residual sum of square of non-linear models. It is also known as the *mean square error.* It measures the estimation accuracy of the default probability defined as

$$Brier = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{p_i})^2, \tag{5.23}$$

where

y$_i$ is the dependent variable of the i^{th} observation and

$\widehat{p_i}$ is the default probability estimation of the model for the i^{th} observation.

The Brier Score is between 0 and 1. The smaller the score is, the more accurate is the model. The model with lowest mean square error will be the best one. The expectation of the Brier score is the variance of the dependent variable. Since the dependent variable is binary, it is assumed to distribute as binomial. The binomially expectation is:

$$E[Breier] = p(1-p), \tag{5.24}$$

where

p is the default probability of the sample and

$(1-p)$ is the non-default probability of the sample.

Chapter 6

APPLICATION AND RESULTS

6.1 DATA

In this section data used in this study is described. The data used was provided by a Turkish bank. It consists of 29 independent variables; 13 are ordered independent variables and 16 are the financial ratios obtained from manufacturing industry firms. The response variable has two categories 0 indicates non-defaulted and 1 indicates defaulted firms. There are 1649 observations including 61 defaulted and 1578 non-defaulted firms. The variables in data not only include applicant's previous credit history but also the external source information about operations and market.

The first 13 variables have four levels on ordinal scale. In 1989, Hosmer and Lemeshow stated that traditionally, variables of this type have either been analyzed if they were continuous or categorical [53]. In addition, In 2002, Agresti [61] stated that:

"The position of ordinal variables in the qualitative and quantitative classification is fuzzy. Analysts often treat them as qualitative using this methods for nominal variables. But in many respects, ordinal variables more closely resemble interval variables than they resemble nominal variables"

By taking into account these concepts, we used the qualitative nature of our ordered variables instead of treating them as if they were nominal variables. Therefore, we did not need to include them to the models as dummy variables.

When the independent variables are on different scales, the variables can be transformed to the same scale to get more efficient results. In this study, the ratios were transformed to ordered form having four levels representing an order from low performance to higher performance. By this process, we transformed all our variables to common nature. Some ratios representing high performance when they are high, some are desirable when they are low. By transforming the ratios to ordered form, we also attained to rank our variables.

When transforming, the limits were determined after interviewing with the credit experts from banks. The exact limits and number of groups were determined after several simulation trials. The limits for each ratio that gives the best determination power was taken as the mapping limit.

6.1.1 Variables

1. Credit in follow-up period,
2. Rediscount,
3. Payment routine,
4. Firm's and shareholders' prestige,
5. Relations with other firms and banks,
6. Firm's and shareholders' assets,
7. Financial and managerial risk,
8. Plant and building ownership,
9. Relations with financial institutions,
10. The maturity of the credit,
11. Demand quality of the product,
12. Purchasing and selling maturity structure and
13. Utilization capacity.
14. *Current Ratio*:

    $$\frac{\text{Current Assets}}{\text{Current Liabilities}} \tag{6.1}$$

    Current Ratio is an indicator of a firms' ability to meet its short term debt obligations. If the current assets of a firm are more than twice of the current liabilities then the firm is considered to be successful in meeting its short term obligations. If the current ratio is equal to 1 this means that the firm could theoretically survive for one year.

15. *Quick Ratio*:

$$\frac{\text{QuickAssets}}{\text{CurrentLiabilities}} \tag{6.2}$$

QuickAssets = CurrentAssets − Inventories = Accounts Receivable + Cash

Quick Ratio is an indicator of a firms' liquidity and ability to meet short term obligations. It is also known as *acid-test ratio*. The higher the ratio is, the better the liquidity position of the firm is.

16. *Net Working Capital Ratio*:

$$\frac{\text{Net Working Capital}}{\text{Total Assets}} \tag{6.3}$$

Net Working Capital = Current Assets − Current Liabilities

Net working capital is an good indicator of firms' liquidity. It is widely used for bankruptcy prediction. A low ratio may indicate a higher risk of bankruptcy.

17. *Total Assets Turnover Ratio*:

$$\frac{\text{Net Sales}}{\text{Total Assets}} \tag{6.4}$$

Total Assets turnover Ratio is an indicator of firms' ability in using its assets to generate sales and earn profits. A high ratio means that the firm is using it's assets efficiently.

18. *Fixed Assets Turnover Ratio*:

$$\frac{\text{Net Sales}}{\text{Fixed Assets}} \tag{6.5}$$

Fixed assets turnover ratio is an indicator that shows how effectively the firm uses its buildings, plant, machines and equipment to generate sales. In other words, it shows how productive the firm is.

19. *Debt Ratio*:

$$\frac{\text{(Current + Long Term) Debt}}{\text{Total Assets}} \tag{6.6}$$

Debt ratio indicates the percentage of assets that have been financed by borrowing. The lower the ratio is, the less risky the firm is since debt may have high interest payments.

20. *Current Liabilities to Net Worth Ratio*:

$$\frac{\text{Current Liabilities}}{\text{Net Worth}} \tag{6.7}$$

The current liabilities to net worth ratio is an indicator of creditor's security. The higher the ratio is the less secure the creditors are. A ratio of 0.5 or higher may indicate inadequate owner investment.

21. *Fixed Assets to Total Assets Ratio*:

$$\frac{\text{Fixed Assets}}{\text{Total Assets}} \tag{6.8}$$

Fixed assets to total assets is a measure extent which fixed assets financed with shareholder's equity. A high ratio indicates an insufficient usage of working capital. In other words, it indicates a low cash reserve that may limit the firms' ability to respond to a increased demand for its products.

22. *Current Liabilities to Net Sales Ratio*:

$$\frac{\text{Current Liabilities}}{\text{Net Sales}} \tag{6.9}$$

Current liabilities to net sales ratio is an indicator of firms' activity and riskiness. The higher the net sales is the lower the ratio is and less risky the firm is. Default probability would decline, as the ratio becomes smaller.

23. *Net Operating Margin*:

$$\frac{\text{Operating Income}}{\text{Net Sales}} \tag{6.10}$$

Net operating margin is the amount of margin generated by operations. It is an indicator of profitability of firm. The higher the ratio is the more profitable the firm is in operating. It is the amount of margin generated by operations.

24. *Net Profit Margin*:

$$\frac{\text{Net Profit}}{\text{Net Sales}} \tag{6.11}$$

Net profit margin is an indicator of effectiveness of the firm at cost control. The higher the ratio is, the more effective the firm is to cover its revenue from sales to profit. It tell us how much profit generated for every unit of revenue.

25. *Return on Assets Ratio (ROA)* :

$$\frac{\text{Net Profit}}{\text{Total Assets}} \tag{6.12}$$

Return on assets is an indicator of profitability of the firm. It shows how much a firm generates by using its assets. It is also known as *return on investments*.

26. *Financial Expense to EBITDA Ratio*:

$$\frac{\text{Interest Expense}}{\text{EBITDA}} \tag{6.13}$$

EBITDA = Earnings Before Interest, Taxes, Depreciation and Amortization

Financial expense to EBITDA is an indicator of expenditure of borrowing.

27. *Return on Equity Ratio (ROE)*:

$$\frac{\text{Net Profit}}{\text{Shareholders' Equity}} \tag{6.14}$$

Return on equity is an indicator which shows how much a firm generates with the shareholders' investments. With higher ROE, it is more likely to generate cash with internal sources.

28. *Net Sales Increase*:

$$\frac{\text{Current Year's Net Sales- Prior Year's Net Sales}}{\text{Prior Year's Net Sales}} \tag{6.15}$$

29. *Total Assets Growth Rate*:

$$\frac{\text{Current Year's Total Assets}}{\text{Prior Year's Growth Rate}} \tag{6.16}$$

6.1.2 Data Diagnostic

The most important part of modelling is to get information from data. The process usually gives what you expect, but often it also provides clues and hints. The role of data diagnostic of time series data is to get clues and hints to have powerful idea about what has happened in the past. To get an idea, it is very important to look at the data by means of visualizing and evaluating descriptive statistics. In data analysis, the theoretical assumptions should be achieved and model specifications should capture the main feature of data.

An average is one of such a feature; it is the location parameter of distribution. When taken alone, it tells us little about data. To interpret an average, we need to have a good idea about the variability. Standard deviation, which is measures the distance from the mean, is the right measure for variability. The smaller the variance, square of standard deviation, the more representative the mean is. In Table 6.1, there are descriptive statistics of ratios .

Table 6.1: Descriptive statistics for ratios

	N	Min	Max	Mean	Std. Dev.	Skewness	Kurtosis
x14	1649	0.00	96.82	2.2932	4.33662	13.886	249.464
x15	1649	0.01	55.14	1.2904	2.46420	15.056	289.465
x16	1649	-307.11	709.08	7,5211	30.87337	14.593	310.937
x17	1649	0.00	21.91	1.8204	1.36298	4.196	39.433
x18	1649	-78.29	1128.37	28.5556	70.56754	7.001	70.483
x19	1649	-29.23	338.57	3.6267	13.49042	16.081	330.875
x20	1649	0.00	4.35	0.4786	0.28894	3.921	41.282
x21	1649	-10.64	60.56	0.7183	2.04891	19.656	500.945
x22	1649	0.00	1450.34	1.2467	35.71187	40.590	1648.042
x23	1649	-838.40	0.55	-0.5325	20.66393	-40.502	1643.133
x24	1649	-865.01	0.34	-0.5529	21.32062	-40.497	1642.867
x25	1649	-2.60	0.46	0.0305	0.16199	-6.299	72.895
x26	1649	-189.01	121.28	0.4523	6.35912	-11.919	580.968
x27	1649	-62.97	41.73	0.0809	1.95309	-14.619	784.933
x28	1649	-0.98	874.25	1.6223	29.51309	26.572	737.372
x29	1649	0.00	51.58	0.4060	1.69212	20.822	559.016

In addition to the mean and the standard deviation, there are minimum and maximum values in data. The variability in x_{18}, x_{16}, x_{22} and x_{28} are the highest since the range between minimum and maximum values are high, so using the mean is less representative for the sample.

The distribution of the data is very important for assumptions of future estimations. We can test normality by means of skewness and kurtosis. One way to observe the normality property is to look at the kurtosis of the distributions. *Kurtosis* is the degree of peakness of a distribution. It is a measure of the extent to which observed data fall near the center of a distribution or in the tails. Standard normal distribution has a kurtosis of zero. Positive kurtosis indicates a ”peaked” distribution and negative kurtosis indicates a ”flat” distribution. *Skewness* defines the degree of asymmetry of the distribution. A negative skewness value indicates that the data have a distribution skewed to left. A positive skewness value indicates a right skewed distribution. A zero skewness value indicates that the data has a symmetric distribution.

In Table 6.1, there are skewness and kurtosis statistics. The statistics indicates that x23, x_{24}, x_{25}, x_{26} and x_{27} are skewed to left and other ratios skewed to right. The kurtosis statistic for all ratios are all positive and high implies that the distribution of ratios are peaked. Bar graphs are a very common type of graph suited for a qualitative independent variable. It is used to compare the amounts or frequency of occurrence of different characteristics of data. The graph allows to compare groups of data, and to make generalizations about the data. The Figure 6.1 shows the bar graph of ordered variables.

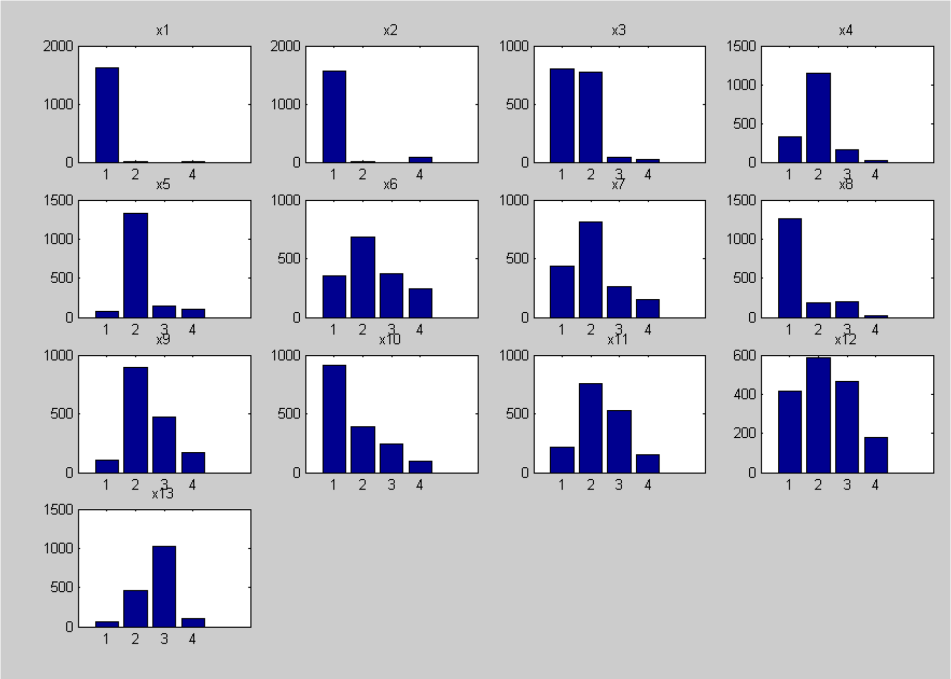


Figure 6.1: Bargraphs of ordered variables

6.1.3 Sample Selection

In credit rating, the problem mostly faced is rather technical than theoretical. The credit scoring procedure requires a lot of high quality data. For this purpose, sample selection is the most challenging problem in credit scoring.

In 1984, Zmijewski [62] studied two types of sample bias in financial distress prediction. One is the bias results when researcher firstly observes the dependent variable and then selects a sample based on this knowledge. Secondly, there is the result when only observations with complete data are used to estimate the model. The sample bias is observed clearly and probit model is extended by bias adjustment to avoid bias. However, when statistical inference is investigated not much significance difference was found.

In 1998, Greene [63] tried to determine how sample selection affects the predictors bias. He claimed that when the model consists of the individuals who were accepted, the sample is arguably random. A conditional model taking into account the condition of only accepted individuals is used to predict probability of default to avoid this problem. This conditional models predict much higher default rate and this model could distinguish more sharply the ones who defaulted before.

Geert [64] analyzed the impact of sample bias on consumer credit scoring performance profitability and concluded that when sample size increases the sample bias decreases. The possible impact of the effect of sample bias on credit scoring is limited and does not appear to have a negative influence on scoring performance and profitability.

Since the results of previous studies indicates that the sample bias caused by using only accepted individuals has no significant effect on discriminatory power and there is no such data is available in Turkey, in this research only the observations of accepted applicants are used. On the other hand, the unbalanced valued sample also affects the reliability of models such as logistic regression and probit regression. In 2005, Işcanoğlu [65] checked the conditions that logistic regression performs more accurate and concluded that unbalanced size of defaulted and undefaulted individuals affect the accuracy of the model. Additionally, she concluded that when the number of defaulted observations is too small, the accuracy is low, increasing the size of defaulted observations bias decreases and when sample size increases the bias decreases.

By taking into account her results to obtain more reliable and accurate results, a sample selection procedure is used. While selecting the sample, logistic regression was used as the basic model. By fixing the defaulted observations with random numbers, 1000 different samples were selected independently. The samples, which were used to construct logistic regression and the significance and explanatory power of models were compared. The sample giving the significance model with highest explanatory power was taken as the sample to compare credit scoring models.

6.2 CREDIT SCORING MODEL RESULTS

6.2.1 Classification and Regression Trees Results

CART is a widely used nonparametric technique for feature selection. The purpose is to determine a set of if-then else conditions to classify observations. In other words, it is a sequence of questions which can be answered as yes or no concerning whether the condition satisfied or not.

It is a form of binary recursive partitioning, it splits the sample into classes with root node to ending with homogenous sub-samples. This process is called building T_{max}. A desirable tree is the one having a relatively small number of branches and high predictive power. The optimum tree is built by evaluating the cost and complexity measure.

In Figure 6.2, the decision tree with optimum cost and complexity is represented. It can also be called the *right-sized tree* for credit scoring problem. The right-sized tree for our problem is built with x_1, x_4, x_{14}, x_{17}, x_{18}, x_{19}, x_{20}, x_{22}, x_{28} and x_{29}.

The root node is the one in the top of the tree that is containing all the applicants in the sample. This node's split is based on x_{14}; if variable x_{14} <2.5, the applicants put into x_4; and if x_4 <1.5, the applicant is classified as defaulted otherwise classified as non-defaulted. When the applicants in the child node x1 have the property that x_1 <1.36, then the applicants classified as defaulted and the applicants are put into x_{17}, otherwise. This process continues to the last terminal node. The first split is based on x_{14}, the second based on x_{14} and x_{17}, the third based on x_{18}, the fourth x_{29}, the fifth x_{20}, sixth x_{22}, seventh x_{19} and the last split based on x_{28}.

At the upper levels of the tree there are more significant variables, less significants are at the bottom of the tree. Therefore, x14 is the most significant variable for CART since it is located at the root node.

The CART results not only consist of the tree, they also include a *tree report.* The report consists of tree sequence, terminal nodes, child nodes, assigned classes , splitting criteria, variables included to model, cutoff values of the variables, misclassification rates and cost of complexity parameter.

The right sized tree was obtained after 5 prunings with optimum cross-validation result. Like the other classification tools the results of CART also depend on the sample, so this model was built with the sample used by the other analysis.

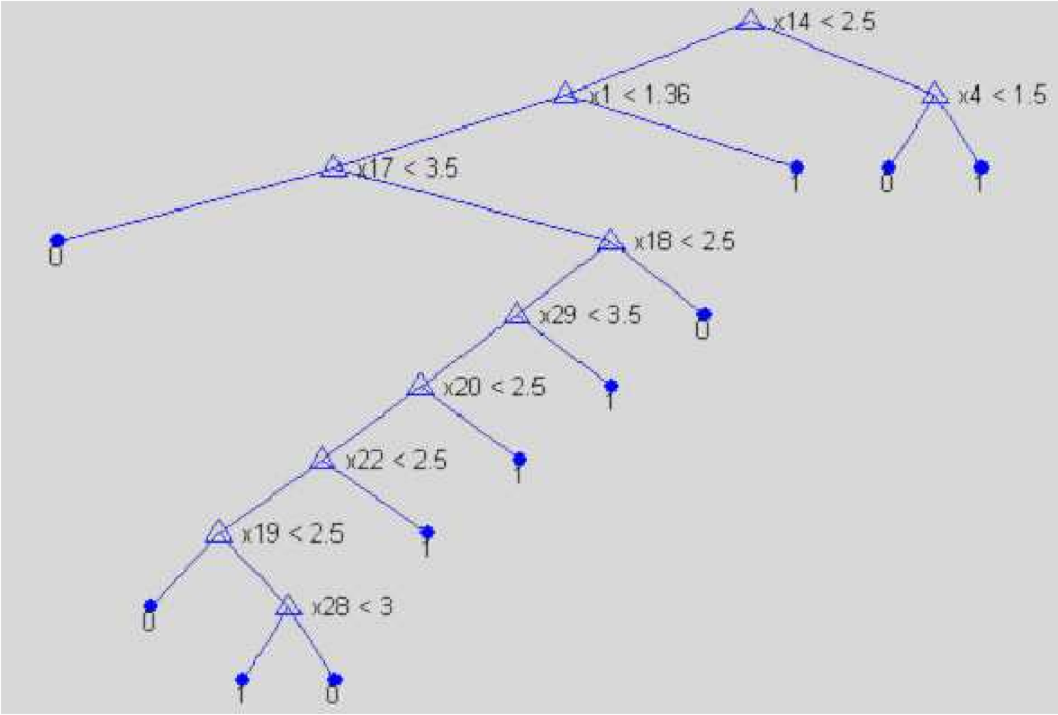


Figure 6.2: Classification tree

The CART applications are aimed to achieve the best predictive accuracy with minimum expected cost. The expected misclassification cost depends both on cost of assigning a class i observation as class j observation and prior probabilities of classes. Priors that are the proportions of classes that minimizing the expected cost may affect the classification of cases. CART has a direct way to incorporating such prior information to the tree model.

For this purposes, trees including different combinations of misclassification costs, prior probabilities and splitting rules were constructed to determine the classification tree with highest predictive ability. The tree in Figure 6.2 includes equal prior probability and equal cost of misclassification assumption with Gini splitting rule. In Table 6.2, the cross validation estimates of the all combinations with Gini and Twoing splitting rule are shown. The trees giving the best cross-validation estimates are the ones constructed with Gini splitting rule, having misclassification of defaulted applicants as non-defaulted is 3 unit cost and having misclassification cost of non-defaulted case as defaulted is 1 unit cost for both equal probabilities and when prior probability of default is taken as 0.1 and non-defaulted 0.9.

Table 6.2: Cross-validation results for alternative classification trees

	p=0.5,q=0.5					p=0.1,q=0.9				
	(1,1)	(1,2)	(1,3)	(2,1)	(3,1)	(1,1)	(1,2)	(1,3)	(2,1)	(3,1)
Gini	2.3%	3.4 %	3.82%	2.18%	1.82%	2.3%	3.46%	3.4%	2.12%	1.82%
Twoing	2.3%	3.09 %	3.4 %	2.67 %	2.12 %	2.3%	3.40%	3.40%	2.73%	2.12%

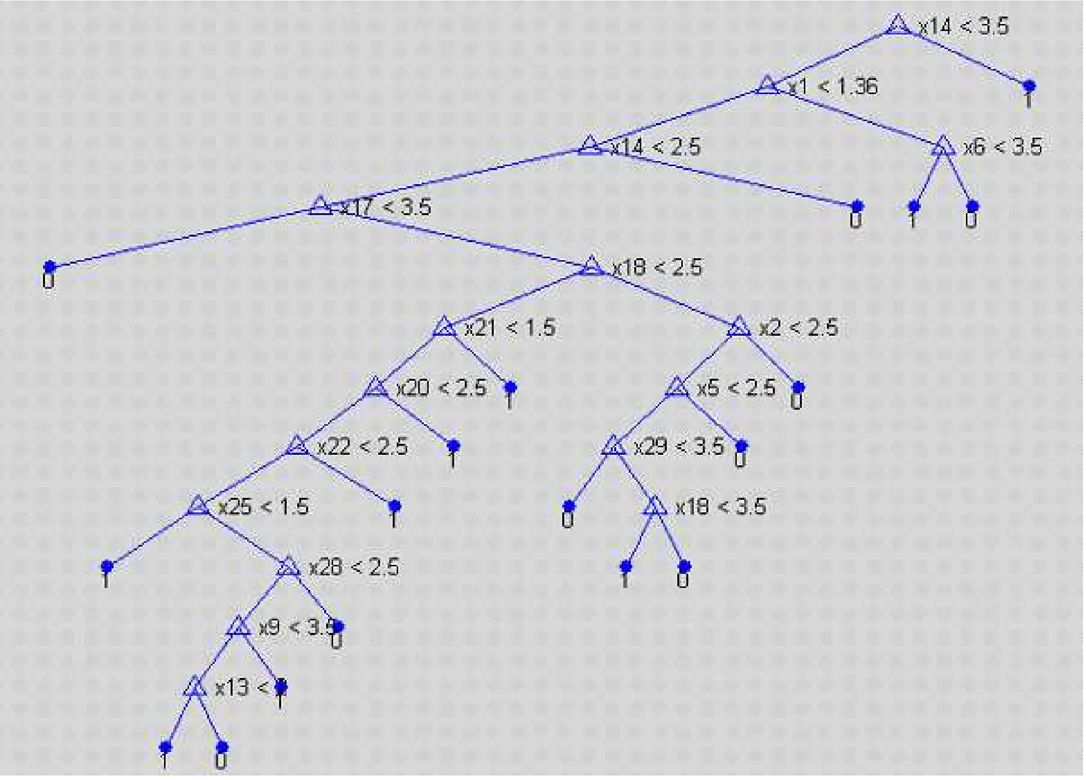


Figure 6.3: The best classification tree

The prior probabilities did not change the result. In both cases the two trees have 0.0182 cross-validation result. Both trees can be used in prediction. The prior probability p=0.1 and p=0.9 is taken since the sample is assumed to come from Bernoulli so the predictions can be taken as class proportions. In this thesis the tree with Gini splitting rule, equal prior probabilities and having misclassification of defaulted applicants as non-defaulted is 3 unit cost and having misclassification cost of non-defaulted case as defaulted is 1 unit cost is taken as optimal tree. It is shown in Figure 6.3.

6.2.2 Logistic Regression Results

Like linear regression, logistic regression is also a method to produce a prediction equation with regression weights. But unlike linear regression there are no assumptions about the linearity relation between independent and dependent variables and about the normality of independent variables. Therefore, there is no need to check the normality of independent variables before analyzing.

Logistic regression applies maximum likelihood estimation after transforming the dependent variable. In this study, to select the variables to use in regression equation backward selection method was used. Firstly, all variables were included to the model that significance of coefficients and overall significance are evaluated. The insignificant variables with 95 % confidence level were thrown away, with the other variables the model is again constructed and significance of variables checked and insignificant ones are again thrown away. This procedure continued until the significant model with significant parameters is found. In this study, after three steps the significant model was found. The coefficients are summarized in Table 6.3.

Table 6.3: Logistic regression model parameters

Variables	Coeff.	Std. Errors	Wald	P-Values	Lower 95 %	Upper 95 %
x1	2.604538	0.680633	3.83	0	1.270522	3.93856
x10	1.81057	0.456525	3.97	0	0.915798	2.70534
x14	3.029178	0.835939	3.62	0	1.390767	4.66759
x15	-1.47571	0.549624	-2.68	0.007	-2.55295	-0.3985
x16	1.373796	0.501039	2.74	0.006	0.391778	2.35582
x17	7.28283	1.437741	5.07	0	4.46491	10.1008
x18	-1.94684	0.400815	-4.86	0	-2.73242	-1.1613
x20	0.789897	0.37517	2.11	0.035	0.054578	1.52522
x22	1.376298	0.382063	3.6	0	0.627468	2.12513
x24	3.885938	1.163319	3.34	0.001	1.605875	6.166
x25	-2.90873	0.984381	-2.95	0.003	-4.83808	-0.9794
constant	-44.8438	8.451162	-5.31	0	-61.4078	-28.28

Although, the dependent variable takes values 0 and 1, the regression equation does not give prediction values of 0 and 1. The regression equation of linear combinations of independent variables gives the log odds, and log odds are used to compute the predicted values of probabilities of default. The parameter coefficients are called *logits* of explanatory variables used to estimate log odds. One unit of increase in a variable with β_1 logit is associated with a β_1 change in log odds of the dependent variable. It does not directly affect the change in dependent varaible.

The regression equation of our model was found as:

$$\ln(\frac{P(Y=1)}{1+P(Y=1)}) = \text{log-odds} = -44.84381 + 2.604538x_1 + 1.81057x_{10} + \\ 3.029178x_{14} - 1.47571x_{15} + 1.373796x_{16} + 7.28283x_{17} - 1.94684x_{18} \\ +0.789897x_{20} + 1.376298x_{22} + 3.885938x_{24} - 2.90873x_{25}.$$

By using transformations, log odds are transformed to probabilities of default. A firm is classified as "defaulted" if it has a probability of default greater than 0.5, and "non-defaulted" otherwise. There is no theoretical reason why 0.5 is chosen generally as the cutoff point but it is implicitly chosen by taking into account the assumption of loss function that is assumed to be symmetric across two types of error.

The output of logistic regression includes logits, Wald statistics, log likelihood, chi-square statistic and pseudo-R^2 statistic. Our results are summarized in Table 6.4

Table 6.4: Logistic regression statistics

Log-likelihood	-33.791208
Pseudo R^2	0.8311
Chi-square statistic	332.55
P-Value	0.000

Log-likelihood is the natural logarithm of likelihood. Likelihood indicates how likely the observed values of dependent variable are predicted with observed values of independent variables. In our study, likelihood was found to be -33.791208. Log likelihoods are negative since likelihoods is a probability between 0 and 1. Likelihood and log likelihood can be used when comparing logistic models with different variable combinations.

The overall significance of model is evaluated by goodness of fit test. In logistic regression, generally, the chi-square test derived from log likelihood value is used. The hypothesis for overall significance is defined as:

H_0: $\beta_1 = \beta_2$=...=β_p=0.

H_1: At least one of the coefficient is different than 0.

The rejecting or accepting can be determined by comparing chi-square value with chi-square table value or by comparing the p value with α significance level. In our study the chi-square statistic was found as 332.55 and p value was found to be 0.000. The significance of the model was proved since p=0.000 $< \alpha = 0.05$ with 95 % confidence level. This means H_0 was rejected and the equation of log-odds not only consist of constant coefficient.

In logistic regression, like other regression types there is also a chance to test the significance of individual model parameters. For this purpose Wald statistic is a commonly used tool. It is the square of coefficients divided by the standard errors. In Table 6.3, the coefficients, standard errors, Wald statistics, p values and confidence intervals are summarized. The hypothesis tested by Wald statistic is defined as:

H_0: β_i=0 (i=1,...,p).

H_1: $\beta_i \neq 0$.

The decision about the hypothesis is taken by comparing Wald statistic with standard normal table value or p value compared with α significance level. Since all p values are smaller than $\alpha = 0.05$ all these 11 variables contribution to the model is significant.

Pseudo R^2 is the determination coefficient, interpreted like the R^2 in linear regression. In our model it is 83,11 % that means 83.11 % variability is dependent variable can be explained by independent variables. But generally it is not suggested to use pseudo R^2 as a comparison tool in logistic regression.

6.2.3 Probit Regression Results

Probit regression is an alternative method for logistic regression. They are both generalized linear models. Their underlying assumptions and evaluation tests are same. Like logistic regression a transformation of probability of Y such that it equals to 1 is obtained before maximum likelihood estimation. While logit transformation is conducted to obtain odds, the function used in probit regression is the standard normal cumulative distribution function.

Generally, both models come to the same conclusion, but the meanings and magnitudes of coefficients are different. In this study, like logistic regression the variables were selected with a backward selection algorithm. After two steps, the variables in

Table 6.6 are selected. They are the same variables the ones were selected in logistic regression. The probit regression results are shown in Table 6.5 and Table 6.6.

Table 6.5: Probit regression statistics

Log-likelihood	-33.050788
Pseudo R^2	0.8348
Chi-square statistic	332.03
P-Value	0.000

Table 6.6: Probit regression model parameters

Variables	Coeff.	Std. Errors	Wald	P-Values	Lower 95 %	Upper 95 %
x1	1.467358	0.366954	4	0	0.7481402	2.18658
x10	1.026047	0.250032	4.1	0	0.5359937	1.5161
x14	1.678499	0.449107	3.74	0	0.7982667	2.55873
x15	-0.79628	0.292105	-2.73	0.007	-1.368794	-0.2238
x16	0.764288	0.280438	2.73	0.006	0.2146386	1.31394
x17	4.030123	0.773951	5.21	0	2.513207	5.54704
x18	-1.08406	0.218014	-4.97	0	-1.511361	-0.6568
x20	0.443471	0.205909	2.15	0.035	0.0398973	0.84705
x22	0.78544	0.209872	3.74	0	0.3740992	1.19678
x24	2.190696	0.647229	3.38	0.001	0.9221511	3.45924
x25	-1.64083	0.54457	-3.01	0.003	-2.708164	-0.5735
constant	-25.0479	4.600916	-5.44	0	-34.06557	-16.03

The regression coefficients in probit regression corresponds to the logits of logistic regression. But the pobit coefficients indicate the effect of the independent variables on the dependent variable. They indicate the change in the cumulative normal probability of the dependent variable when independent variable changes by one unit. While logit model expressed log odds the probit model expressed z-scores.

The Z-score equation of this study is obtained as follows:

$$\begin{aligned}\Phi^{-1}(P(Y=1)) = -25.0479 + 1.467358x_1 + 1.026047x_{10} + 1.678499x_{14}\\ -0.79628x_{15} + 0.764288x_{16} + 4.030123x_{17} - 1.084064x_{18} +\\ 0.443471x_{20} + 0.78544x_{22} + 2.190696x_{24} - 1.64083x_{25}\end{aligned}$$

To predict probabilities of default by using the Z-score values to the standard normal distribution table is needed. As in logistic regression, after the probabilities of defaults predicted, the individuals were classified by using the value 0.5 as a threshold.

The logistic regression coefficients are approximately 1.8 times the probit regression coefficients. These two models can be converted to each other.

Like logistic regression, the output includes log likelihood, chi-square test statistic, pseudo R^2, Wald statistic and coefficient confidence intervals summarized in Table 6.6

The first step in interpretation of the model is to test the overall significance. The result of the goodness of fit statistic is 334.03; it can be compared with chi-square table value with 11 degrees of freedom or by comparing the p value with α significance level. Since p=0.000 $<$ 0.05 with 95 % confidence null hypothesis that includes only the constant term in the Z-score model was rejected, and we concluded that the model is significant.

Pseudo R^2 was found as 83.48 % that gives the explained part of the variability in dependent variable. Like logistic regression all the variables were found to be significant when Wald statistic was used. All the p values in Table 6.6 are smaller than $\alpha = 0.05$; then, with 95 % confidence level, all the 11 variables are important in predicting the probability of Y taking the value 1.

6.2.4 Linear Discriminant Analysis Results

Before using discriminant analysis, to assess assumptions it is helpful to visualize the data and check descriptive statistics. Figure 6.1 show the distributions of data and Table 6.1 shows the descriptive statistics. As mentioned before skewness and kurtosis indicate that the data are not suitable for discriminant analysis. Although the assumptions could not be achieved, discriminant analysis was used in this study by taking into account the previous researches which were claimed that discriminant analysis is an effective tool for credit scoring even if the assumptions does not hold.

For discriminating, Fisher's linear discriminant analysis was used in this study. The purpose of LDA is to establish linear combinations of variables which discriminates the groups best. The most common use of discriminant analysis is to include many variables to the model and using a stepwise procedure to select optimal set of variables. In this study, a forwardation variable selection was used by taking Wilk's lamda as a criterion. This variable was then paired with other variables and the pair giving the best result of criterion was taken. This procedure continues untill the criterion reaches to the value given before or until all variables are selected. When two variable combinations give the same performance measure, the more parsimonious model is chosen.

In this study, after 8 steps the optimal variable combination was found including x_1, x_5, x_{14}, x_{17}, x_{18}, x_{19}, x_{21}, x_{22}. From 29 variables these 8 variables were found to be optimal to built the discriminant function.

The output of discriminant analysis consists of discriminant function weights, Wilk's lambda statistic and standardized canonical discriminant function.

Table 6.7: Discriminant analysis model parameters

Variables	coefficients for Y=0	coefficients for Y=1
x1	2.677	5.748
x5	7.215	9.544
x14	8.652	12.856
x17	9.850	14.565
x18	-0.497	-2.319
x19	2.226	2.885
x21	-1.539	-0.577
x22	5.118	6.732
constant	-36.845	-75.530

The weights are used to combine linearly the variables that will produce score which maximizes the distance between two groups. The Fisher's discriminant weights are shown in Table 6.7. The LDA function was built as follows:

$$\begin{aligned} D_0 = -36.845 + 2.677x_1 + 7.215x_5 + 8.652x_{14} + 9.850x_{17} \\ -0.497x_{18} + 2.226x_{19} - 1.536x_{21} + 5.118x_{22}, \end{aligned}$$

and

$$\begin{aligned} D_1 = -75.530 + 5.748x_1 + 9.544x_5 + 12.856x_{14} + 14.565x_{17} \\ -2.319x_{18} + 2.885x_{19} - 0.577x_{21} + 6.732x_{22}. \end{aligned}$$

Here, D_0 and D_1 are the discriminant functions for defaulted and non-defaulted individuals respectively. In addition to the weights of discriminant functions, there are standardized beta coefficients shown in Table 6.8, they are given for each variable in the discriminant function. They show the relative magnitude of the contribution of each variable. The variables giving larger standardized coefficients are the ones that contribute more in discrimination. In other words, the larger the coefficient the more important the model is. In this study, x_{17} has the highest coefficient; this means it is the most important one in discriminant analysis.

Table 6.8: Discriminant analysis standardized coefficients

Variables	Standardized coefficients
x1	0.307
x5	0.331
x14	0.581
x17	0.979
x18	-0.591
x19	0.161
x21	0.217
x22	0.578

The significance of any discriminant function is assessed by using Wilk's lambda. The null hypothesis used to conclude the result of Wilk's lambda is that the means of each variables vectors for each group are equal. When the hypothesis is rejected, the discriminant function is found to be significant in discriminating the groups. The smaller the value, the more significant the model is. Table 6.9 shows Wilk's lambda statistics for each steps. To decide whether to reject or accept the hypothesis, p value can be used. As the p values compared to the $\alpha = 0.05$ null hypothesis is rejected, and concluded that these models are significant. The discriminant function is successful to discriminate defaulted and non-defaulted individuals. As seen from the Table 6.9 in steps 8 the Wilk's lambda value is the smallest.

When variables chosen for models are compared, there is a difference in size of variables chosen between models. But there is common 5 variables significant to classify in each model: x_1, x_{14}, x_{17}, x_{18}, x_{22}. These variables are the common variables so we can say that they are important variables in credit scoring. The most significant variable for CART is x_{14} but for discriminant analysis is x_{17} it changes from one

Table 6.9: Discriminant analysis Wilk's lambda statistics

Step	Number of Variables	Wilk's Lambda	F statistic	df1	df2	P-value
1	1	0.716	247,449	1	625	0.000
2	2	0.627	185.779	2	624	0.000
3	3	0.554	167.228	3	623	0.000
4	4	0.497	157.280	4	622	0.000
5	5	0.440	158.101	5	621	0.000
6	6	0.418	143.877	6	620	0.000
7	7	0.402	131.280	7	619	0.000
8	8	0.399	116.553	8	618	0.000

method to another.

6.3 VALIDATION RESULTS

In the previous section, the separate models for each techniques were presented. The dynamics of methods and variables included to the models are different, so the predictive abilities are expected to be different and should be tested to determine the best model.

In this section, model performances were evaluated by means of predictive power and discriminatory power tests. For the prediction power misclassification rates were estimated and compared. For discriminatory power, ROC curve, CAP curve, Pietra index, Bayesian error rate, K-L distance and CIER techniques suggested by Basel committee were used [66].

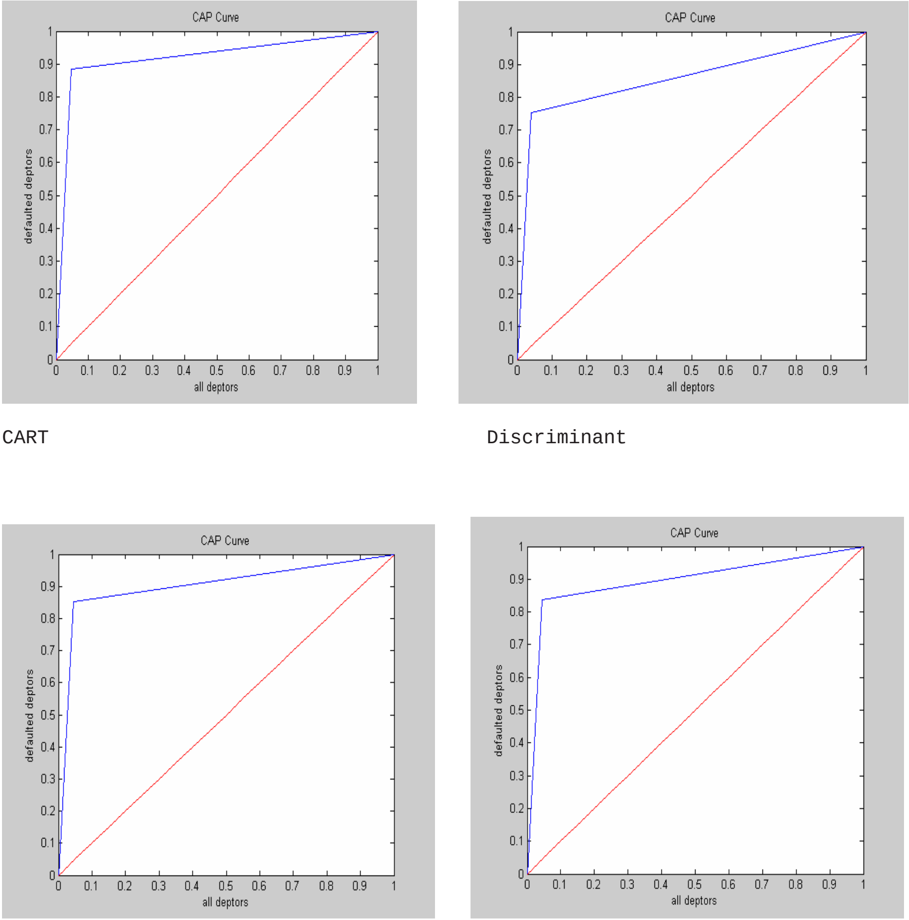
The Table 6.10 represents the misclassification rates. The models were built by a sample, but the accuracy of models were evaluated for whole data. As seen from Table 6.10, CART has the minimum misclassification error rate, 98.18 % percent of data is predicted correctly. The results for all models are nearly equal, approximately 98 % of the observations is correctly predicted.

Table 6.10: Misclassification rates of models

	CART	Logistic Regression	Probit Regression	Discriminant Analysis
Misclassification Rates	1.82 %	2.18 %	1.94 %	1.88%

In Table 6.11, Figure 6.5 and Figure 6.4, the discriminatory power test results are summarized.

In Figure 6.4, there are the CAP curves of models. It is a curve of cumulative proportions of true good risks against the proportions of true bad risks. The steeper the curve, the more successful the models is. As seen from Figure it is not easy to compare the models so; the AR statistic should be used. When AR results are evaluated, CART dominates the other three modes. Discriminant analysis has the lowest performance.



CART

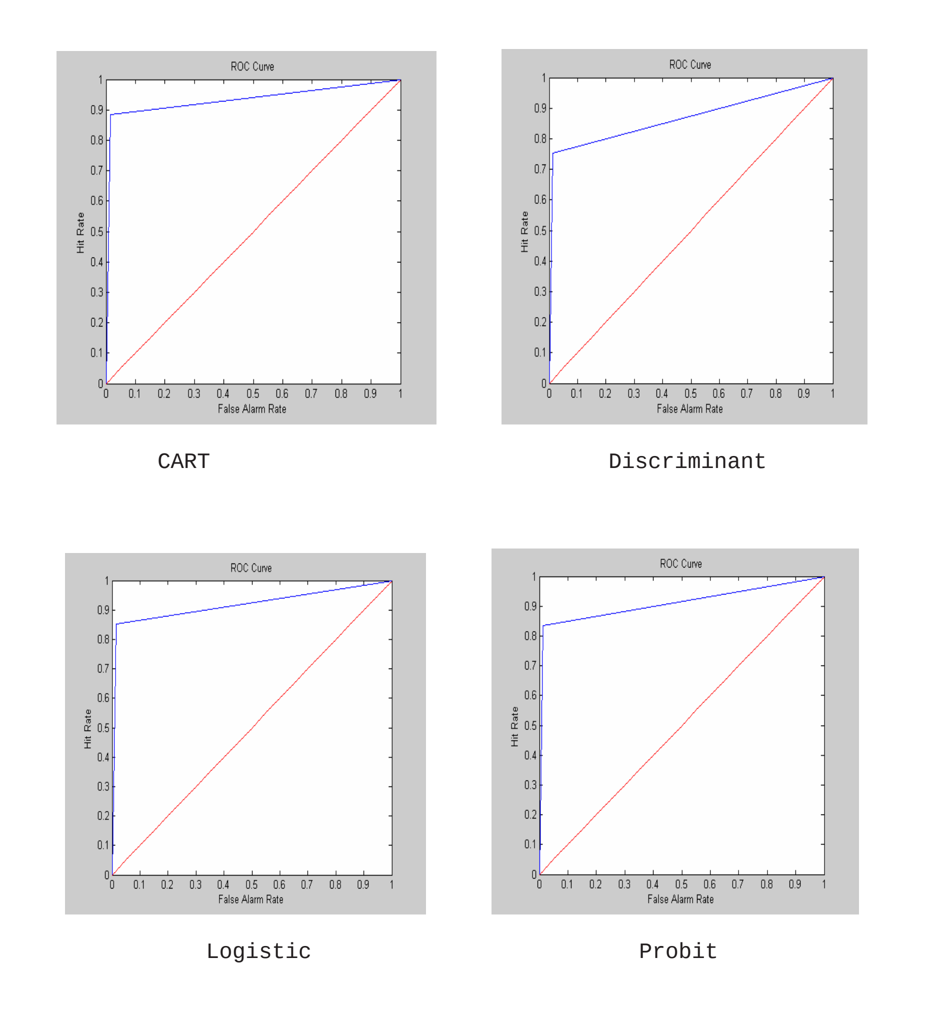
Discriminant

Logistic

Probit

Figure 6.4: CAP curves of models

In Figure 6.5, the ROC curves of models are represented. The ROC curve is a curve of the true positive rate against false positive rate. Like CAP curve, the models with steeper curve has better discriminatory power. The mostly used summary statistic of the curve is AUC. The higher the AUC, the better the models is. The results of AUC is consistent with the results of AR. The performance of CART is the best, it is followed by logistic regression, probit regression and discriminant analysis, respectively.



CART

Discriminant

Logistic

Probit

Figure 6.5: ROC curves of models

Pietra index is also a summary statistic of ROC curve relating with the area between the curve and the unit square diagonal. The model is desirable when it has a higher index and the results are the same with results of AUC.

Bayesian error rate is also known as a summary statistic of ROC, but it is simply the sum of type I and type II errors, so it gives the same result with the misclassification

Table 6.11: Discriminatory power results of models

	CART	Logistic Reg.	Probit Reg.	Discriminant
AR	0.9193	0.9099	0.9035	0.8962
AUC	0.9354	0.9256	0.9190	0.9114
Pietra Index	0.3079	0.3010	0.2963	0.2909
Bayesian Error Rate	0.0182	0.0218	0.0194	0.0188
K-L Distance	-0.1002	-0.1212	-0.1339	-0.1468
CIER	-0.3876	-0.4336	-0.4584	-0.4813

rate.

KL-distance and CIER are information based tests. They give the information added to the discriminatory power by the models. KL-distance is the distance between with and without model information and CIER is the normalized measure of KL-Distance. The model with maximum distance is the best model. In Table 6.11, there are the results of these information measures. The result are again the same, CART added the maximum information to the discrimination and is followed by logistic regression, probit regression and discriminant analysis

In summary, our empirical results show that in most cases CART performs better than others. All the models have different predictive and discriminatory power and also the information included.

The most accurate measures suggested by Basel Committee are AR and AUC. The results of both measure are the same result. It is better to use CART and logistic regression for our data.

6.4 ASSIGNMENT OF RATINGS

When banks or financial institutes use models to estimate probabilities of default, they require to assign the individual applicants to discrete rating classes by determining cutoff thresholds. In other words, the scoring model results are divided into intervals so the applicants of each interval can be associated with a single PD.

There are important points that should be taken into account, while assessing rating classes. Firstly, when assigning a rating class, the bank must assess the risk of the applicant, this means that the number of classes and thresholds should be sufficient to ensure the differentiation of risk in each category. According to Basel II, there must be at least seven classes; the first one represents the non-defaulted applications and the last one represents only the defaulted applicants. The classes in between have monotonically increasing risk profiles.

In the internal rating based approach, there is not only one way to assign rating class. According to Basel II, banks or institutions can apply the cut-off values used by rating agencies such as Fitch, S & P and Moody's. The rating cutoff values of S & P is represented in Table 6.12

Table 6.12: S & P rating scale with cut-off values

Rating Map	Lower Limit	Upper Limit
AAA	0	0.002
AA	0.002	0.01
A	0.01	0.04
BBB	0.04	0.15
BB	0.15	0.42
B	0.42	0.96
CCC	0.96	1

The rating classes can be determined intuitively by experts or a mapping procedure that mimics the assignment rules of the rating agencies can be taken. However, they are all standard methods and may not fit for all situations.

For each application profile the cut-off that is giving optimum risk differentiation and satisfying within group homogeneity is needed. Therefore, it is better to construct the cutoff values special for each sector and each applicant profile and adjust the cut-off values for each adjustment periods of scoring.

In this study a data specific cut-off values were constructed with dual objective optimization. The main issue here is to choose a threshold that satisfies two objectives. First objective is to maximize area under ROC curve. It is needed to improve the classification performance across classes. When the case in which default predicted are default realized, the analysis gives higher AUC statistic. However, we have one more problem, that is the firms which were assigned to a specific class must act alike. In other words, homogeneity within a class should be satisfied. To achieve this objective ROC curve is not enough since it does not represent the statistical properties of the classless in consideration. For this purpose, a criterion that represents the statistical property of a rating class should be optimized. Since the validation results indicated that CART is the best model for our data set, we used mean square error (MSE) of the regression tree as a criterion to be minimized to achieve our second objective.

The process begins with the estimation of PDs of each applicants by using logistic regression. In the high rating classes the PD is lower. By using this information, first PDs were sorted, then they were assigned to classes that in each class there are approximately equal number of applicants. This is an intuitive way of rating assignment. There are 7 classes and AUC higher than 0.9. Since it is just an intuitive way of assignment, it should be improved to get better results by optimizing our two objective. The process continued with the evaluation of two criteria.

To improve the intuitive ranking method, different combinations of cut-off values were obtained by moving the thresholds up and down. Since the cut-off values in intuitive method has AUC high enough, there is no need to move the thresholds much. Firstly, to choose combinations 1000 times 7 random numbers limited by 40 were chosen. Then the thresholds places were moved up and down by these numbers. To evaluate the criterion, firstly for each combination the ROC curve was constructed and the AUC statistic was estimated. Secondly all the applicants were assigned to classes by using all threshold combinations. Thirdly, teh applicants who were ranked in classes were used as an ordered data to construct regression tree. Then the MSEs for regression tree were estimated. There were 1000 AUC and MSE estimates for the combination that had been moved up, 1000 estimates for the combinations that had been moved down.

The last stage is to find a cut-off value that gives the optimum result. Since one criterion should be maximized and the other should be minimized, a sum function combining both was written. Since maximizing AUC is minimizing the negative of AUC, negative weight was given to AUC criterion in the function. The cut-offs giving minimum value of function is taken as the optimum result. The cut-off value and groups are summarized in Table 6.13

Table 6.13: Optimum rating scale

Rating Map	Lower Limit	Upper Limit
A	0	0.000000301
B	0.000000301	0.000002996
C	0.000002996	0.000020061
D	0.000020061	0.000223369
E	0.000223369	0.003878509
F	0.003878509	0.143434960
G	0.143434960	1

Chapter 7

CONCLUSION

In this thesis, technical aspects of internal rating were given. The internal rating process has three steps. The first step is to construct credit scoring models. Since credit scoring is a statistical classification problem, statistical classification techniques were summarized. To improve the models reliability the problems faced in classification is tried to be minimized. Firstly, all possible variables relating with default prediction were obtained from a Turkish bank. Secondly, to minimize the sample bias in logistic regression an optimum sample was chosen by Monte Carlo simulation. The data set was tried to be taken as large as possible to satisfy the asymptotic properties of the models. However, we had no chance to build our own data set so it was impossible to use conditional models including both accepted and unaccepted applicants to avoid sample bias; also we had no chance to compare different default definitions. Since the variables were in different scales, the continuous variables were transformed to ordered scale by determining thresholds after taking the opinions of credit experts of banks. After overcoming such problems summarized above credit scoring models were built. The models were constructed with different significant variables but it was concluded that credit in the follow-up period, current ratio, total asset turnover ratio, fixed asset turnover ratio and current liabilities to net sales were found to be significant for all models. The pseudo R^2 of logistic regression and probit regression is approximately 0.83. This, means that by our variables and data set the 83% variability in default and non-default cases can be explained.

The second stage is the evaluation of the models. The models performances were evaluated with misclassification rates, ROC curve, CAP curve, Pietra index, Bayesian error rate, K-L distance and CIER for whole data set. CART has the minimum misclassification rate, 98.18% of data is predicted correctly with CART. The results are the same for other discriminatory power tests. CART dominates the other techniques. In most cases, logistic regression has the second best performance.

The last stage is the assignment of ratings. Cut-off values used to map PD's to rating classes were assigned by evaluating two criteria: AUC and MSE for regression tree. Seven classes were built after assessing 2000 results of 1000 simulation. PDs estimated by logistic regression were assigned to classes with simulated thresholds. Then these results were used to built a regression tree and a threshold giving optimum result was chosen.

In future works, if it is possible to obtain specific data set, the models with different default definitions for Turkish manufacturing firms can be compared and conditional models including all applicants can be built. The aim of this thesis was to use statistical credit scoring methods; non-statistical techniques such as neural networks can also be applied and compared with statistical techniques. The results of this study can be used take a decision about new applicants whether to grant or not to grant credit. In addition, the results of internal ratings can be applied to construct credit risk measurement models.

References

[1] The internal rating approach. Technical report, Basel Committee on Banking Supervision, http://www.bis.org, 2001.

[2] D. Durand. Risk elements in consumer installment financing. *National Bureau of Economic Research: New York USA*.

[3] The new basel capital accord. Technical report, Basel Committee on Banking Supervision, http://www.bis.org, 2003.

[4] D.J. Hand. Modelling consumer credit risk. *IMA Journal of Management Mathematics*, 12:139–155, 2001.

[5] W.H. Beaver. Financial ratios as predictors of failure. *Journal of Accounting Research*, 4:71–111, 1966.

[6] W.H. Beaver. The information content of annual earnings announcements. *Journal of Accounting Research*, Supplement:67–92, 1968.

[7] E.B. Deakin. A discriminant analysis of predictors of business failure. *Journal of Accounting Research*, March, 1972.

[8] J.H. Myers and E.W. Frogy. The development of numerical credit evaluation systems. *Journal of American Statistics Association*, 58(September):799–806, 1963.

[9] E.I. Altman. Financial ratios, discriminant analysis and prediction of corporate bankruptcy. *The Journal of Finance*, 23:589–609, 1968.

[10] M. Blum. Failing company discriminant analysis. *Journal of Accounting Research*, 12(Spring):1–25, 1974.

[11] I.G. Dambolena and S.J. Khoury. Ratio stability and corporate failure. *Journal of Finance*, 35(4):1017–1026, 1980.

[12] G.E. Pinches and K.A. Mingo. A multivariate analysis of industrial bond ratings. *Journal of Finance*, 28(1):1–18, 1973.

[13] P. Harmelink. Prediction of best's general policyholder's ratings. *Journal of Risk and Insurance*, December:621–632, 1974.

[14] H.J. Davidson R.M. Cyert and G.L. Thompson. Estimation allowance for doubtful accounts by markov chains. *Management Science*, 8:287–303, 1962.

[15] D. Mehta. Optimal credit policy selection: A dynamic approach. *Journal of Financial and Quantative Analysis*, December, 1970.

[16] H.Jr. Bierman and W.H. Hausman. The credit granting decision. *Management Science*, 16(8):519–B532, 1970.

[17] M.S. Long. Credit screening system selection. *Journal of Financial and Quantitative Analysis*, June:313–328, 1976.

[18] A.W. Corcoran. The use of exponentially-smoothed transition matrices to improve forecasting of cash flows from accounts receivable. *Management Science*, 24:732–739, 1978.

[19] J. Spronk J.A.M. Van Kuelen and A.W. Corcoran. On the cyert-davidson-thompson doubtful accounts model. *Management Science*, 27:108–112, 1981.

[20] V. Srinivasan and Y.H. Kim. Credit granting: A comperative analysis of classifivcation procedures. *The Journal of Finance*, 42:665–681, 1987.

[21] R. Castelo B. Baesens, M. Egmont-Petersen and J. Vanthienen. Learning bayesian network classifiers for credit scoring using markov chain monte carlo search. *Proc. International Congress on Pattern Recognition (ICPR02)*, pages 49–52, 2002.

[22] J.A. Batten L.V. Philosophov and V.L. Philosophov. Multi-period bayesian bankruptcy prediction: Using financial ratios and the maturity schedule of long-term debt. 2005.

[23] D. Mehta. The formulation of credit policy models. *Managerial Science*, 15:30–50, 1968.

[24] J. Horrigan. The determination of long-term credit standing with financial ratios. *Journal of Accounting Research*, 4:44–62, 1966.

[25] Y. Orgler. A credit scoring model for commercial loans. *Journal of Money, Credit and Banking*, 2:435–445, 1970.

[26] J.A. Ohlson. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18:109–131, 1980.

[27] J.C. Wiginton. A note on the comparison of logit and discriminant models of consumer credit behaviour. *Journal of Financial and Quantitative Analysis*, 15:757–770, 1980.

[28] K. Menon L.R. Gilbert and K.B. Schwartz. Predicting bankruptcy for firms in financial distress. *Journal of Business, Finance and Accounting*, 17(1).

[29] K. Roszbach. Bank lending policy, credit scoring and the survival of loans. *Working Paper Series in Economics and Finance Stockholm School of Economics*, Working Paper No 261, 1998.

[30] S. Chang A.J. Feelders and G.J. McLachlan. Mining in the presence of selectivity bias and its application to reject inference. *Knowledge Discovery and Data Mining*, pages 199–203, 1998.

[31] F. Cames and M. Hill. Consumer credit scoring models:does the underlying probability distribution matters? *Conference CEMAPRE*, 2000.

[32] E. Hayden. Are credit scoring models sensitive to different default definitions? evidence from the austrian market. *SSRN Working Paper*, 2003.

[33] D.T. Huyen and Thanh. Credit scoring for vietnam's retail banking market : implementation and implications for transactional versus relationship lending. *METEOR, Maastricht research school of Economics of Technology and Organizations*, 2006.

[34] M. Shaw and J.A. Gentry. Using an expert system with inductive learning to evaluate business loans. *Financial Management*, 17 (Autumn):45–56, 1988.

[35] M.D. Odom and R. Sharda. A neural network model for bankruptcy prediction. *Proc. IEEE lnt. Conf. on Neural Networks, San Diego*, pages II163–II168, 1992.

[36] K. Tam and M. Kiang. Managerial application of neural networks: the case of bank failure predictions. *Management Science*, 38(7):926–947, 1992.

[37] Coats and Fant. Recognizing financial distress patterns using a neural network tool. *Financial Management*, Autumn:142–155, 1993.

[38] K. Kiviluoto. Predicting bankruptcies with the self-organizing map. *Neurocomputing*, 21:191–201, 1998.

[39] C. Charalambous and et al. Comparative analysis of artificial neural network models: Application in bankruptcy prediction. *Annals of Operations Research*, 99:403–425, 2000.

[40] J. Sinkkonen S. Kaski and J. Peltonen. Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks*, 12(4):936–947, 2001.

[41] G.H. Tzeng J.J. Huang and C.S. Ong. Two-stage genetic programming (2sgp) for the credit scoring model. *Applied Mathematics and Computation*, articles in press, 2005.

[42] J.W. Wilcox. A prediction of business failure using accounting data. *Journal of Accounting Research*, 11, 1973.

[43] P. Kolesar and J.L. Showers. A robust credit screening model using categorical data. *Management Science*, 31:123–133, 1985.

[44] W.V. Gehrlein and B.J. Wagner. A two-stage least cost credit scoring model. *Annals of Operations Research*, 74:159–171, 1997.

[45] W.E. Henley and D.J. Hand. A k-nn classifier for assessing consumer credit risk. *The Statistician*, 45:77–95, 1996.

[46] M. Muller and W. Hardle. Multivariate ans semiparametric kernel regression. *Wiley*, pages 357–391, 2000.

[47] L. Brieman et al. *Classification and regression trees.* Chapman and Hall, 1998.

[48] E.I. Altman H. Frydman and D.L. Kao. Introducing recursive partitioning for financial classification: The case of financial distress. *The Journal of Finance*, 40:269–291, 1985.

[49] P. Pompe and A. Feelders. Using machine learning, neural networks, and statistics to predict corporate bankruptcy. *Microcomputers in Civil Engineering*, 12:267–276, 1997.

[50] J. Tuo B. Li X. Li, W. Ying and W. Liu. Applications of classification trees to consumer credit scoring methods in commercial banks. *Systems, Man and Cybernetics SMC (5)*, pages 4112–4117, 2004.

[51] D.J. Spiegelhalter D. Michie and C.C. Taylor. *Machine Learning, Neural and Statistical Classification*. Prentice Hall, Ellis Horwood, London, July 1994.

[52] P. Mc Cullagh and J.A. Nelder. *Generalized linear models*. Chapman and Hall, 2 edition, 1989.

[53] D.W. Hosmer and Jr.S. Lemeshow. *Applied logistic regression*. Wiley, 1989.

[54] H. Tatlıdil. *Uygulamalı çok değişkenli istatistiksel analiz*. Cem Web Ofset, 1996.

[55] A. Pagan and A. Ullah. *Nonparametric econometrics*. Cambridge University Press, July 1999.

[56] M.P. Wand and M.C. Jones. Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the american Statistical Association*, 88:520–528, 1993.

[57] W.P. Wand D. Ruppert and R.J. Carroll. *Semiparametric regression*. Springer, 1998.

[58] W. Hardle et al. *Nonparametric and semiparametric models*. Springer, 1 edition, May 2004.

[59] S. Keenan J. Sobehart and R. Stein. Benchmarking quatitative risk models: A validation methodology. *Moody's Risk Management*, 2000.

[60] S. Keenan J. Sobehart and R. Stein. Validation methodologies for default risk models. *Credit*, pages 51–56, 2000.

[61] A. Agresti. *Categorical Data analysis*. Wiley, Canada, 2002.

[62] M.E. Zmijewski. Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, 22:59–82, 1984.

[63] W. Greene. Sample selection in credit-scoring models. *Japan and the World Economy*, 10(3):299–316, 1998.

[64] G. Verstraeten and Van den Poel. The impact of sample bias on consumer credit scoring performance and profitability. *Journal of Operational Research Society*, 56(8), 2005.

[65] A. İşcanoğlu. Credit scoring methods ans accuracy ratio. Master's thesis, METU, 2005.

[66] Studies on the validation of internal rating systems. BIS Working Paper 14, Basel Committee on Banking Supervision, http://www.bis.org, 2005.

[67] H. Benishay. Economic information in financial ratio analysis. *Accounting and Business Research*, 2:174–179, 1971.

[68] R.A. Eisenbeis and R.B. Avery. *Discriminant analysis ans classification procedure: theory and applications.* Heath, 1972.

[69] D.C. Warner V.P. Apilado and J.J. Dauten. Evaluative techniques in consumer finance - experimental results and policy implications. *Journal of Financial and Quantitative Analysis*, March:275–283, 1974.

[70] R. Libby. Accounting ratios and the prediction of failure: some behavioral evidence. *Journal of Accounting Research*, 13(1):150–161, 1975.

[71] C.L. Norton and R.E. Smith. A comparison of general price level and historical cost financial statements in the prediction of bankruptcy: A reply. *Accounting Review*, 55(3):516–521, 1980.

[72] W.R. Klecka. *Classification and regression trees.* Sage Publications, 1980.

[73] M.A. Wolfson M.L. Marais, J.M. Patell. The experimental design of classification models: An application of recursive partitioning and bootstrapping to commercial bank loan classifications. *Journal of Accounting Research*, 22:87–114, 1984.

[74] J.R. Quinlan. Introuction of decision trees. *Machine Leraning*, 1:81–106, 1986.

[75] G.J. McLachlan. *Discriminant analysis ans statistical pattern recognition.* Wiley, 1992.

[76] D.J. Hand and W.E. Henley. Statistical classification methods in consumer credit: a review. *Journal of the Royal Statistical Society, Series A*, 160:523–541, 1997.

[77] J.L. Horowitz. *Semiparametric methods in econometrics.* Springer, 1998.

[78] M. Muller and B. Ronz. Semiparametric credit scoring. *Springer*, pages 357–391, 1999.

[79] A. Charitou and L. Trigeorgis. Option based bankruptcy prediction. *Working paper, University of Cyprus*, 2000.

[80] D.W. Hosmer and Jr.S. Lemeshow. *Applied logistic regression.* Wiley, 1989.

[81] A.F. Atiya. Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks*, 12(4):929–935, 2001.

[82] M. Muller. Estimation and testing generalized partial linear models- a comparative study. *Statistics and Computing*, 11:299–309, 2001.

[83] D.B. Edelman L.C. Thomas and J.N. Crook. *Credit scoring and its applications.* SIAM, Philadelphia, 2002.

[84] K. Komorád. On credit scoring estimation. Master's thesis, humboldt University, 2002.

[85] T. Janosi and R. Jarrow. Estimating default probabilities implicit in equity prices. *Journal of Investment Management*, 1(1), 2003.

[86] E. Hayden B. Engelmann and D. Tasche. Testinf rating accuracy. *Risk*, 16(1):82–86, 2003.

[87] G. Verstraeten and D. Van Den Poel. The impact of sample bias on consumer credit scoring performance and profitability. *Working Papers of Faculty of Economics and Business Administration, Ghent University, Belgium*, 2004.