HYBRID MODEL OF CREDIT SCORING

by

Serdar Akar

B.S., Computer Engineering, Mersin University, 2002

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in System and Control Engineering

Boğaziçi University

2007

ACKNOWLEDGEMENTS

It is pleasure to express my appreciation to those who have influenced this study. I am grateful to Dr. Tamer Şıkoğlu for his encouragement, guidance, and always supportive correspondence.

I am indebted to Prof. Dr. Fikret Gürgen for his help, valuable comments and suggestions.

I am also very grateful to Eng. Ahmet Boyalı for all his efforts.

My special thanks are to my family without whose encouragement and supports this thesis would not be possible.

I acknowledge my debt and express my thanks to Eng. Fatih Kasap for his endless motivation, brotherhood and being with me all the way.

In my deprivation with his hardware help to my foreign friend Sadık.

ABSTRACT

HYBRID MODEL OF CREDIT SCORING

One of the most important innovative concepts is the credit scoring. Today it can be interested in different sectors. Thus the improvement of credit scoring is increasing day by day. The credit scoring with the help of classification techniques provides to take easy and quick decisions in lending. However, no definite consensus has been reached with regard to the best method for credit scoring and in what conditions the meth- ods performs best. Although a huge range of classification techniques has been used in this area, the logistic regression has been seen an important tool and used very widely in studies. This study aims to examine accuracy and bias properties in parameter estimation of the logistic regression (binary logistic) , linear discriminant analysis , linear regression by using German Data which has different variables, data types, real basement and accurately results. Moreover, application of these significant statistical analyzes on German data is provided and the method accuracies are examine for new consumer elements by the software application. Finally, ratings on the results of best method is done by hybrid model by its most reliance comparing and completion all methods.

ÖZET

KREDİ SKORLAMASI HİBRİD MODELİ

Günümüzde yenilikçi, ilerlemeye açık, çok önemli kavramlardan biri haline gelen kredi skorlaması farklı sektörlerce kullanılmaktadr. Böylece gündenden güne kredi skorlaması ilerlemektedir. Kredi skorlama, klasifikasyon metodlarının yardımı ile kredi verme operasyonlarında kolay ve hızlı karar verdirir. Fakat en iyi kredi skorlama metodlarının hangi koşullarda iyi performans gösterdikleri ve bunlardan hangisi en iyisi ise bu durum hakkında kesin bir yargı yoktur. Bu alanda bir çok farklı metot kullanılmasına karşın, lojistik regresyon, lineer diskriminant ve lineer regresyon analizleri önemli bir araç olarak çalışmamızda farklı varyantlar, veri tipleri, reel temelli, net sonuçlu Alman Verisinde kullanıldı . Dahası, Alman Verisi üzerinde kullanılan bu çok önemli istatistiksel analizler kesin netlik için yeni kurumsal unsurlar üzerinde de yazılım uygulaması yapıldı . Nihayetinde en iyi model ; güvenilirlik, kıyaslama ve birbirini tamamlama açısından hibrit modellemesi vasıtası ile sonuçlanır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

a_{ij}	Reliable a variation over coefficients
α	Mathematical value of alpha
$\sum$	Sum of the equion
$\int$	Integral of the equation
$\prod$	Production of the equation
BL	Binary Logistics
CART	Classification and Regression Trees
DA	Discriminant Analysis
LR	Logistic Regression
NN	Neural Networks
Ra.A	Ratio Analysis
SPSS	Statistical Program for Social Sciences
SPR	Semiparametic Regression Trees

1. INTRODUCTION

Throughout the history, financial problems has been the main issues which has to be solved for the human life. However it can not be imagine to live without financial balances. Therefore the historical development of credit scoring occured by itself. In a credit granting procedure, a credit company's main aim is to determine whether a credit application should be granted or refused. The credit scoring procedures in fact measure the risk on lending. From the early civilizations this risk has been assessed by the interest rate on it. However, the studies on the financial situations of the governments, companies and individuals has demon- strated that the interest does not diminish the risk. The credit risk should be assessed separately. The default of a firm is always very costly for both shareholders and credit agencies. Because the credit agencies could lose whatever they give and the shareholders could lose all or nearly all of their value of equity. Here, the problem is to learn default some time before the default in order to be take some precautions. The empirical studies indicated that classification methods gives signals of defaults. However, these methods could act differently according to size, shape, and structure of the data. Therefore, selection of the most suitable method for available data is a much more complex concern.

In literature, as far as we know, the scoring studies shows only accuracy comparisons of two or more models. The close examinations of the methods are gaps in this area. Therefore, this study includes the empirical research on especially linear regression, binary logistic , linear discriminant analysis . These methods because of their environment is the most widely used method in studies of credit scoring. However all these analysis types would be the main hybrid design of our credit scoring system. Moreover, the study presented here also includes the applications of classification methods to the real German credit data. The accuracies of all methods are consummate and compare with each other. Thus, for German data sets which are very volatile in consumer from the different group of customers, the most appropriate model is tried to be selected by the hybrid modelling. So this situation would give us the most current accurent validation for the confidence and potent investment. The instruction of the

thesis is as follows. Chapter 2 gives a brief overview of the development of credit scoring, the related studies and the expository variables in the studies. Chapter 3 provides the fundamentals of statistical methods used in credit scoring. In Chapter 4 we will instruct a new combine multiple base model supporting with the software technology. Thus we will try to reach the right establishing by the hybrid analysis model. In Chapter 5 having the final results of the system provides the accuracy ratio and validations of the methods mentioned in previous episodes concludes this thesis.

2. FUNDAMENTALS OF CREDIT SCORING

As we mentioned before; the research in the area of credit scoring started in the 1930's. After that date, many different works and methods have entered the literature of credit scoring. According to type of methods, we can split the period 1930-2005 into 4 sub periods. We call the first period as a primitive age of credit scoring because this includes very basic applications. In this part, research was based only on a ratio analysis. In those years, scientists compared ratios of default and non-default companies and tried to develop an idea of companies' financial performances. As it can be guessed, these type of methods had no predictive power and so they were not very suitable.

The second period of the credit scoring started at 1966 with the application of discriminant analysis. By this application, research gained predictive power. However, this method has very strong assumptions on variables and so, the pre- diction power is not very high. Moreover, this method does not give the idea of relative performances of the variables. We call this period as discriminant age.

The application of discriminant analysis is a turning point for credit scoring because it opened the door for computer-based methods. After the 1970's, the methods that applied to this area changed rapidly. The main types of methods were the regression based approaches. Therefore, we can call this period as regression age of credit scoring. The linear regression was applied firstly, but it did not give good results. Because the credit default probabilities takes only values between zero and one, but linear regression can give the results between $-\infty$ and ∞. Then, secondly, probit regression came into play. Since it also has strong assumptions of normality, the end of the application of this methods came rapidly. In other words, in the period 1970-1980, the regression type methods were not gone to the fore of discriminant analysis. In the 1980's, the study of logistic regression increased the interest to the regression since it has no normality assumptions on variables, allows predictions, and interpretation of coefficients, and it gives the output on the interval zero and one. Although after

the 1980's many other statistical methods have been also applied such as k-nearest neighborhood, classification and regression trees, survival analysis, etc, this method has kept its importance even nowadays as the most widely used statistical technique in research. The year 1990 is an another turning point for credit scoring. In this year, the statistical methods gave their place to the machine learning type methods with the application neural networks. Therefore, we name this period as machine age .

2.1. Primitive Age Researchers and Expository

The first researchers which we found in the primitive age are Ramser and Foster with their 1931 paper. This was followed by Fitzpatrick in 1932. Fitzpatrick investigated nineteen pairs of failed and non-failed companies. He showed a significant difference in the ratios of failed and non-failed companies at least three years prior to failure. After then, Winekor and Smith in 1935, searched the mean ratios of failed firms ten years prior to failure and detected the breakdown in the mean values when failure was coming. In 1942, Merwin studied mean ratios of the failed and non-failed companies in the period 1926-1936 and his result was only differing form Winekor and Smith' work in the six year before failure .In these years, the primary ratio was the current ratio. However, some other ratios were also used. For example, Ramser and Foster studied with equity / (net sales); Fitzpatrick used equity / fixed-asset and return on stock; Winekor and Smith applied its analysis on (working capital) / (total assets); and Mervin beside current ratio, used total debt/equity, (working capital) / (total assets).

2.2. Discriminant Age Researchers and Expository

The researcher who started the discriminant age is Beaver . In his work, he applied a univariate type discriminant analysis by using: (cash flow) / (total debt), (current assets) / (current liabilities), (net income) / (total assets), (total debt) / (total assets), (working capital) / (total assets) . Then, in 1968, Altman investigated a multivariate discriminant analysis by his famous z-score with 5 variables that are 1. (MV of equity) / (book value of debt), 2. (net sales / total assets), 3. (operating

income) / (total assets), 4. (retained earnings) / (total assets), 5. (working capital) / (total assets). Altman obtained 94% and 97% classification accuracy among default and non-default firms, respectively and 95% overall accuracy . The discriminant analysis applications were also continued after 1970's. In 1972, Deakin applied discriminant analysis. In his analysis, he used: cash / (current liabilities), (cash flow) / (total debt), cash / (net sales), cash / (total assets), current ratio, (current assets) / (net sales), (current assets) / (total assets), (net income) / (total assets), (quick assets) / (current liabilities), (quick assets) / (net sales), (quick assets) / (total assets), (total debt) / (total assets), (working capital) / (net sales) and (working capital) / (total assets) and obtained 97% overall accuracy .

In 1972, Lane and Awh and then, Waters in 1974, used the discrimi- nant analysis. Moreover, in 1974, Blum with ratios: market rate of return, quick ratio, (cash flow) / (total debt), fair market value of net worth, (net quick assets) / (inventory), (book value of net worth) / (total debt), (standard deviation of income), (standard deviation of net quick assets) / inventory, slope of income, (slope of net quick assets) / (inventory), trend breaks of income, (trend breaks of net quick assets) / (inventory) applied discriminant analysis . One year later, Sinkey used: (cash+U.S. treasury security) / (assets), (loans) / (assets), (provision for loan losses) / (operationg expense), (loans) / (capital + reserves), (operating expense) / (operating income), (loan revenue) / (total rev- enue), (U.S. treasury securities' revenue) / (total revenue), (state and local obligations' revenue) / (total revenue), (interest paid on deposits) / (total revenue), (other expenses) / (total revenue). Then, Altman and Lorris (1976) acquire 90% classification accuracy with the help of five financial ratios that are

- (net-income)/(total assets)
- (total liabilities + subordinate loans)/equity
- (total assets)/(adjusted net capital)
- (ending capital-capital additions)/(beginning capital)

For detailed information, we refer to Altman and Lorris.

Another paper with discriminant analysis was published by Altman, Halde- man and Narayan (1977). Here, (retained earnings) / (total assets), (earnings before interest and taxes) / (total interest payments), (operating in- come) / (total assets), (market value of equity) / (book value of debt), (current assets) / (current liabilities) were their ratios. And their results showed a 93% overall accuracy .

The investigations which I were able to reach were from the last researchers who made their studies only with discriminant analysis. They are Dambolena and Khory (1980); their ratios: 1. (working capital) / (total assets), 2. (retained earnings) / (total assets), 3. earning before interest and taxes to total assets, 4. (market value of equity) / (book value of debt), 5. sales to total assets, Altman and Izan (1984) and lastly, Pantalone and Platt (1987) with 95% accuracy.

2.3. Regression Age Researchers and Expository Variables

The first researcher who used regression analysis according to my study was Orgler (1970). In his analysis, Orgler basically used: current ratio, working capital, cash / (current liabilities), inventory / (current assets), quick ratio, (working capital) / (current assets), (net profit) / sales, (net profit) / (net worth), (net profit) / (total assets), net profit $\neq 0$, net profit, (net worth) / (total liabili- ties), (net worth) / (fixed assets), (net worth) / (long-term debt), net worth $\neq 0$, sales / (fixed assets), sales / (net worth), sales / (total assets), sales / inventory and sales / receivables. He, in the hold-out sample, was only able to classify 75 % of the bad loans as bad and 35 % of the good loans as good .

In 1976, Fitzpatrick applied multivarite regression also. This research was followed by Olhson (1980) . In his paper, he investigated a logistic regression analysis. He collected data from the period 1970-76. Besides the basic ratios, that are: (total liabilities) /(total assets), (working capital) / (total assets), (current liabilities) / (current assets), (total liabilities)/(total assets) bigger than zero, it takes one, otherwise it takes zero, (net income) / (total assets), (funds provided by opera- tions) / (total liabilities), dummy; One if net income negative for the last two years, he also used size of the

company as an explanatory variable. He calculated the type one and type two types of errors in different cut points and found for his second model better average error that is 14.4%. Then, Pantalone and Platt (1987) tried logistic regression in their paper. They obtained 98% accuracy in the classification of failed firms and 92% accuracy in that of non-failed firms .

In this time period, the recursive partitioning algorithm was gained to the literature by Altman, Frydman and Kao (1985). Their explanatory variables were as follows:

- (net income)/(total assets)
- (current assets)/(current liabilities)
- log(total assets)
- (market value of equity)/(total capitilazation)
- (current assets)/(total assets)
- (cash flow)/(total debt)
- (quick assets)/(current liabilities)
- (earning before interest and taxes)/(total assets)
- log(interest coverage+15)
- cash/(total sales)
- (total debt)/(total assets)
- (quick assets)/(total assets)

For a closer information please look at Frydman documents.

2.4. Machine Age Researchers and Expository

Odom and Sharda in 1990 made a comparison between the discrimi- nant analysis and neural networks by using Altman's (1968) explanatory variables. They collected the data set in the period 1975 to 1982. Their training sample consisted of 74 companies 38 of which were default and hold-out sample was constituted by 55 companies 27 of which default. They concluded that neural networks performed better with respect to

both training sample and hold-out sample. Moreover, the study proved that neural networks were more robust than discriminant analysis even in small sample sizes .

In 1991, Cadden and Coats and Fant made a comparison of discriminant analysis and neural networks also. After that in the following year, Tam and Kiang applied discriminant analysis, logistic regression and neural networks with eighteen explanatory variables. This was followed by another study of Coats and Fant in 1993. In the study, Coats and Fant obtained the data from the 1970 to 1989.

By using Altman's (1968) variables, they run discriminant analysis and neural networks and get that the discriminant analysis gave the highest misclassification error that is classifying a default as non-default. In 1996, Back, Laitinen, Sere and Wesel worked with a huge set of variables that are: 1. cash /(current liabilities), 2. (cash flow) / (current liabilities), 3. (cash flow) / (total assets), 4. (cash flow) / (total debt), 5. cash / (net sales) 6. cash / (total assets), 7. (current assets) / (current liabilities), 8. (current assets) / (net sales), 9. (current assets) / (total assets), 10. (current liabilities) / equity, 11. equity / (fixed assets), 12. equity / (net sales), 13. inventory / (net sales), 14. (long term debt) / equity, 15. (market value of equity) / (book value of debt), 16. (total debt) / equity, 17. (net income) / (total assets), 18. (net quick assets) / inventory, 19. (net sales) / (total assets), 20. (operating income) / (total assets) 21. (earnings before interest and faxes) /(total interest payments), 22. (quick assets) / (current liabilities), 23. (quick assets) / (net sales), 24. (quick assets) / (total assets), 25. rate of return to common stock, 26. (retained earnings) / (total assets), 27. return on stock, 28. (total debt) / (total assets), 29. (working capital) / (net sales), 30. (working capital) / equity, and 31. (working capital) / (total assets).

In 1998, Kiviluoto tried the discriminant analysis, neural networks by means of operating margin, net income before depreciation, extraordinary items, net income before depreciation, extraordinary items of the previous year, equity ratio. For closer details see. After that in 1999, Laitinen and Kankaanpaa compared discriminant analysis, logistic regression, recursive partitioning, survival analysis and neural networks. The study's ratios were cash to current liabilities, total debt to total assets, operating in-

come to total assets. They examined all the methods from three years prior to failure. Moreover, in the total error, they found neural networks as best one year prior to failure and recursive partitioning as best two and three years prior to failure.In 1999, Muller and Ronz showed a different approach to credit default prediction.They implemented the semi parametric generalized partial linear models to this area. In the paper, the 24 variables were used but not specified .In 2000, recursive partitioning, discriminant analysis and neural networks were attempted by McKee and Greenstein. The ratios were as follows:

- (net income)/(total assets),
- (current assets)/(total assets),
- (current assets)/(current liabilities),
- cash/(total assets),
- (current assets)/sales,
- (long-term debt)/(total assets).

In 2001, Atiya documents used: the 1. (book value) / (total assets), 2. (cash flow) / (total assets), 3. price / (cash flow ratio), 4. rate of change of stock price, 5. rate of change of cash flow per share, 6. stock price volatility and he investigated neural networks.

The time table of the studies and methods can be found in Table (2.1). For a close examination please we refer to it, and we abbreviate:

Table 2.1: The methods and scientists.

Method Researchers	Year	Ra. A.	DA	RA	LA	CART	SPR	NN
Ramser, Foster [BLSW96]	1931	X						
Fitzpatrick [BLSW96]	1932	X						
Winakor, Smith [BLSW96]	1935	X						
Merwin	1942	X						
Beaver	1966		X					

Continued on Next Page...

Method Researchers	Year	Ra. A.	DA	RA	LA	CART	SPR	NN
Mears	1966	X						
Horrigan	1966	X						
Neter	1966	X						
Altman	1968		X					
Orgler	1970			X				
Wilcox	1971	X						
Deakin	1972		X					
Lane	1972		X					
Wilcox	1973	X						
Awh, Waters	1974		X					
Blum	1974		X					
Sinkey	1975		X					
Libby	1975	X						
Fitzpatrick	1976			X				
Altman and Lorris	1976		X					
Altman, Haldeman,	1977		X					
Narayan								
Dambolena,	1980		X					
Khoury								
Olhson	1980				X			
Altman, Izan	1984		X					
Altman, Friedman,	1985					X		
Kao								
Pantalone, Platt	1987		X					
Pantalone, Platt	1987				X			
Odom, Sharda	1990		X					X
Cadden	1991		X					X
Coats, Fant	1991		X					X
Tam, Kiang	1992		X		X			X

Continued on Next Page...

Method Researchers	Year	Ra. A.	DA	RA	LA	CART	SPR	NN
Coats, Fant	1993		X					X
Fletcher, Goss	1993				X			X
Udo	1993				X			X
Chung, Tam	1993							X
Altman, Marco,	1994		X					X
Varetto								
Back, Laitinen, Sere,	1996		X		X			X
Wesel								
Bardos, Zhu	1997		X		X			X
Pompe, Feelders	1997		X			X		X
Kivilioto	1998		X					X
Laitinen,	1999		X		X	X		X
Kankaanpaa								
Muller, Ronz	1999				X		X	
Mckee, Greenstein	2000		X			X		X
Pompe, Bilderbeek	2000		X					X
Yang, Temple	2000				X			X
Neophytou,	2001				X			X
Mar-Molinero								
Atiya	2001							X

Thus;

DA: Discriminant analysis,

RA: Regression analysis,

LA: Logistic regression,

CART: Classification and regression trees,

SPR: Semiparametric regression,

NN: Neural networks.

Ra.A: Ratio analysis,

3. STATISTICAL METHODS OF CREDIT SCORING

3.1. Introduction

A credit institute faces with a prejudgement problem of measuring its customer firms creditworthiness. For this reason, the credit institute primarily collects information about its customer firms' measurable features, or; namely, age of the firm, (current assets) / (current liabilities) ratio, and so on. Let $X_i \subseteq \Re^n$ represents the each feature of the customer firm, e.g., X_1 may be the age of the firm, X_2 may be the current assets/current liabilities ratio and so on. Then, each customer firm can be described by a tuple of p random variables, namely, by the vector $X = (X_l, X_2, ..., X_p)$ which indicates a firm's completely all characteristic properties and market and internal performance features. Let the actual values of the variables for a particular customer firm be $x = (x_l, x_2, ..., x_p) \in X \subseteq \Re^n$. Furthermore, let any different possible value of x_i to variable X_i be called an attribute of that feature.

Let us call the space of X as the input space and denote it by Ω since each customer is represented as a point in this space. According to records of the credit institute, in market, there are two types of firms: default firms (D) and non-default firms (ND). Here, a default firm is a customer firm that did not fulfill its obligation in the past, and a non-default firm is a customer firm that fulfilled its obligation in the past. Moreover, the space of all possible outcomes that has only two elements: D and ND, is called the output space Y .

According to our concern, objective is to find best scoring procedure in order to indicate space(X)→space(Y) which splits the space Ω into two subspaces: Ω_{ND} and Ω_D so that classifying new customer firms whose indicator vector belongs to the set Ω_{ND} D as "non-default firm" and whose indicator vector belongs to the set Ω_D as "default firm". In here, a brief review of credit scoring methods which are used in literature most commonly will be given. The above notations are giving to be throughout this chapter.

3.2. Discriminant Analysis

The discriminant analysis is a standard tool for classification. It is based on maximizing the between-group variance relative to the within-group variance.

3.2.1. Decision Theory Approach

According to the discrete or continuous character of the probability distributions, we refer to discrete or the continuous case in the followings.

3.2.1.1. Discrete Case. Assume the companies which ask for credit feature vector has a finite number of discrete attributes so that Ω is finite and there is only a finite number of different attributes x. Suppose p(x—N D) represent the probability that a non-default firm will has an attribute x. Similarly, p(x—D) represents the probability that a default firm will has an attribute x. These conditional probabilities can be shown to be

$$p(x|ND) = \frac{P(\text{firm is a non-default firm and has an indicator vector } x)}{P(\text{firm is a non-default firm})} \tag{3.1}$$

and

$$p(x|D) = \frac{P(\text{firm is a default firm and has an indicator vector } x)}{P(\text{firm is a non-default firm})} \tag{3.2}$$

Since in a market these conditional probabilities can not be observed directly, they will be obtained by using Bayes rule and directly observable probabilities. To apply Bayes rule, let us define p(N D—x) as the probability that a company an attribute vector x as a non-default company and let us define p(D—x) as the probability that a company an attribute vector x as a default company. Then,

$$p(x|ND) = \frac{P(\text{firm is a non-default firm and has an indicator vector } x)}{P(\text{firm has an indicator vector})} \tag{3.3}$$

$$p(x|D) = \frac{P(\text{firm is a default firm and has an indicator vector } x)}{P(\text{firm has an indicator vector})} \tag{3.4}$$

Let γ(x):= p(firm has an indicator vector x), then Equations (3.3)(3.2.1), (3.2.3), (3.2.2) and (3.2.4), respectively, can be put together in the following formulae:

$$\begin{aligned} &p(\text{firm is a non-default firm and has an indicator vector } x) \\ &= p(ND|x)\gamma(x) = p(x|ND)\delta_{ND} \end{aligned} \tag{3.5}$$

and

$$\begin{aligned} &p(\text{firm is a default firm and has an indicator vector } x) \\ &= p(D|x)y(x) = p(x|D)\delta_D. \end{aligned} \tag{3.6}$$

Then, by Bayes rule,

$$p(ND|x) = \frac{p(x|ND)\delta_{ND}}{\gamma(x)}, \tag{3.7}$$

$$p(D|x) = \frac{p(x|D)\delta_D}{\gamma(x)} \tag{3.8}$$

Suppose a credit institute loses the amount $c(ND|D)$ of money for each firm if it classifies a default firm as non-default and loses $c(D|ND)$ amount of money for per firm if it classifies a non-default firm as default. These misclassification costs are summarized in Table (3.1).

Table 3.1. Misclassification Costs

		Classification Result	
		non−default	default
True value	non−default	0	$c(ND\|D)$
	default	$C(D\|ND)$	0

Since in a market these conditional probabilities can not be observed directly, they will be obtained by using Bayes rule and directly observable probabilities. To apply Bayes rule, let us define p(N D—x) as the probability that a company an attribute

vector x as a non-default company and let us define p(D—x) as the probability that a company an attribute vector x as a default company. Then, Ω_{ND} be

$$p(\text{firm is misclassified as non-default}) = p(x|D)\delta_D, \tag{3.9}$$

and the probability that misclassifying a non-default firm as Ω_D be

$$p(\text{firm is misclassified as default}) = p(x|ND)\delta_{ND}. \tag{3.10}$$

Then, the expected cost of misclassifying firms if the firms with attributes belonging to the set Ω_{ND} are accepted and if the firms with attributes belonging to the set Ω_D are refused is

$$\begin{aligned} & c(D|ND) \sum_{x \in \Omega_D} p(x|ND)\delta_{ND} + C(ND|D) \sum_{x \in \omega_{ND}} P(x|D)\delta_D \\ & = c(D|ND) \sum_{x \in \Omega_D} p(ND|x)\gamma(x) + C(ND|D) \sum_{x \in \omega_{ND}} P(D|x)\gamma(x). \end{aligned} \tag{3.11}$$

At this point, the decision rule that minimizes this expected cost is clear. Let us consider cost of classifying a firm with x = (xl, x2, ..., xp). If a firm puts into Ω_{ND}, then the only cost if it is a default is the expected cost C (N D—D)p(x—D)δ_D. If a firm puts into δ_{ND} if it is non-default, then the expected cost is

$$c(D|ND)p(x|ND)\delta_{ND}. \tag{3.12}$$

Therefore, x can be classified into Ω_{ND} if

$$c(ND|D)p(x|D)\delta_D \leq c(D|ND)p(x|ND)\delta_{ND} \tag{3.13}$$

is satisfied. For this reason, the decision rule that minimizes the expected costs is

$$\begin{aligned}\omega_{ND} &= \{x|c(ND|D)p(x|D)\delta_D \leq c(D|ND)p(x|ND)\delta_{ND}\} \\ &= \left\{x|\frac{c(ND|D)}{c(D|ND)} \leq \frac{p(x|ND)\delta_{ND}}{p(x|D)\delta_D}\right\} \\ &= \left\{x|\frac{c(ND|D)}{c(D|ND)} \leq \frac{p(ND|x)}{p(D|x)}\right\}.\end{aligned} \tag{3.14}$$

That is, we classify a firm as a non-default firm if the above condition is satisfied. Otherwise, classify the firm as a default firm.

3.2.1.2. Continuous Case. Let us assume that the indicator vector has a finite number of continuous type attributes so that is finite and there is only a finite number of different attributes x. The same procedure from the discrete case is applied to the con- tinuous case. Here, the only difference is that the conditional probability mass functions p(x—N D) and p(x—D) are replaced by the continuous probability den- sity functions. Then, the expected cost of misclassifying firms if the firms with attributes belonging to set Ω_{ND} are accepted and if the firms with attributes belonging to set Ω_D are refused, will become

$$c(D|ND)\int_{x\in\omega_D} f(x|ND)\delta_{ND}\,dx + c(ND|D)\int_{x\in\omega_{ND}} f(x|D)\delta_D\,dx \tag{3.15}$$

and the *decision rule* that minimizing this expected cost will become (3.2.14)

$$\begin{aligned}\omega_{ND} &= \{x|c(ND|D)f(x|D)\delta_D \leq c(D|ND)\delta_{ND}\} \\ &= \left\{x|\frac{c(ND|D)\delta_D}{c(D|ND)\delta_{ND}} \leq \frac{f(x|ND)}{f(x|D)}\right\}.\end{aligned} \tag{3.16}$$

That is, we classify a firm as a non-default firm if the above condition is satisfied. Otherwise, we classify the firm as a default Firm.

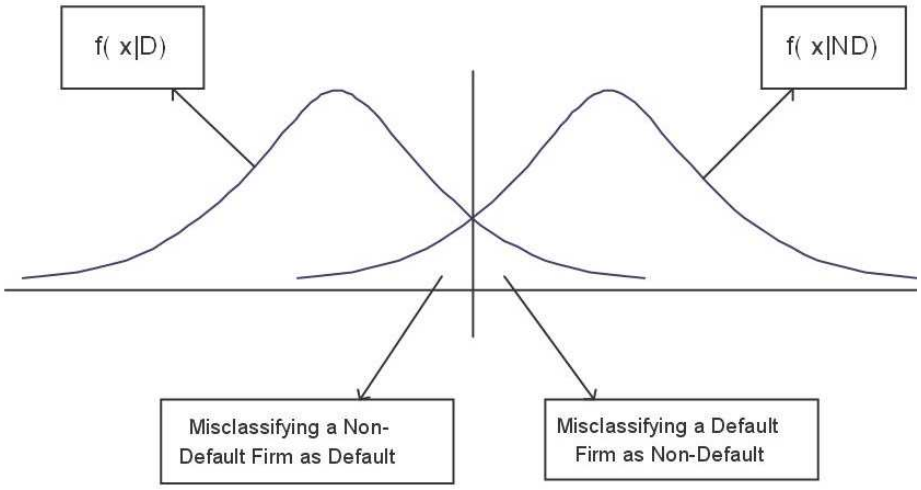

Figure 3.1. Misclassification errors

3.2.2. Fishers discriminant function analysis

This approach is also known as Fisher's discriminant function analysis after Fisher, 1936. In his work, he tried to fit a linear discriminant function of feature variables that best splits the set into two subsets (see Figure 3.2 for an impression). The Fisher's discriminant function consists of combination of feature variables.

Let $Y = w_1X_1 + w_2X_2 + ... + w_pX_p$ be any linear combination of credit performance measure $X = (X_1, X_2, ..., X_p)$. Like analysis of variance (ANOVA), the Fisher's discriminant function analysis uses the differences of the mean values of Y in the two subspaces: the space of default firms and space of non-default firms as a splitting criteria. In this analysis, therefore, the weights of X_i's (i = 1, 2,. . . , p)are such that they minimize the distance between the sample means of default and non-default firms over the square root of the common sample variance. Fisher's Principle is an optimization problem which look as follows:

$$\min J(w) = w^T \frac{(m_{ND} - mD)^T}{(wSw^T)^{1/2}} \tag{3.17}$$

where m_{ND} and m_D are sample means vectors for non-default and default companies, respectively, and S is the common variance-covariance matrix.Differentiating (3.17)

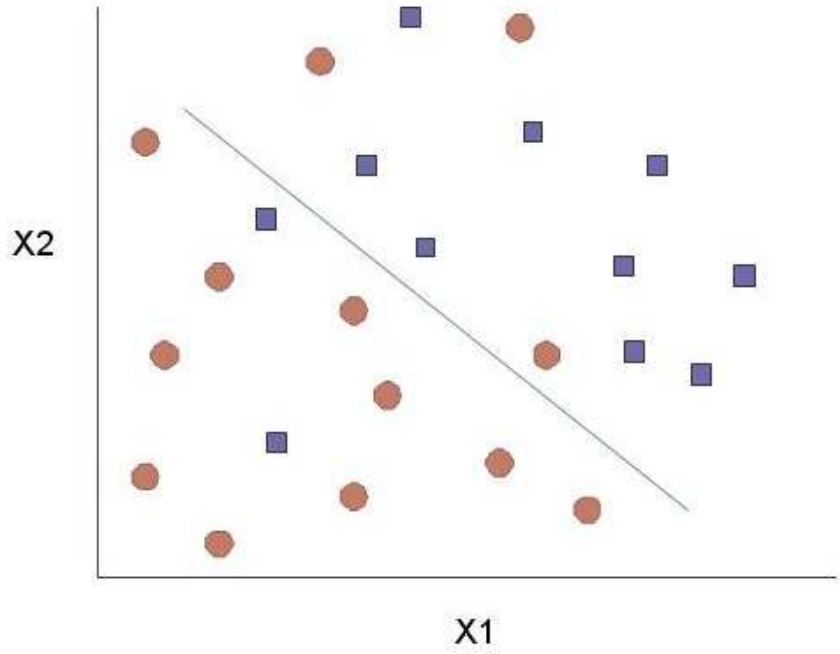

Figure 3.2. Fishers linear discriminant analysis

with respect to w and setting it to 0 derives the following equation:

$$\frac{m_{ND}^T - m_D^T}{(wSw^T)^{1/2}} - \frac{(w(m_{ND} - m_D)^T)(Sw^T)}{(wSw^T)^{3/2}} = 0 \tag{3.18}$$

$$(m_{ND} - m_D)^T(wSw^T) = (Sw^T)(w(m_{ND} - m_D)^T) \tag{3.19}$$

Since $\frac{wSw^T}{w(m_{ND}-m_D)^T}$ is a constant, Equation (3.19) results in

$$w^T \alpha (S^{-1}(m_{ND} - m_D)^T). \tag{3.20}$$

For a detailed information it can be seen in [CET02] [B04] [JW97].

3.2.3. Advantages and Disadvantages of Discriminant Analysis

Discriminant analysis has the following advantages:

- dichotomous response variable,

- easy to calculate,
- yields the input needed for an immediate decision,
- reduced error rates.

Discriminant analysis has the following disadvantages:

- normality assumption on variables,
- approximately equal variances in each group,
- assumption on equivalent correlation patterns for groups,
- problem of multi-collinearity,
- sensitivity to the outliers.

3.3. Linear Regression

3.3.1. Introduction

The linear regression is a statistical technique for investigating and modelling the linear relationships between variables. The probability of default is defined by the following form of linear regression:

$$w_0 + w_1X_1 + w_2X_2 + ... + w_pX_p = w^*X^{*T} \tag{3.21}$$

where $w^* = (w_0, w_1, w_2, \ldots, w_p)$ and $X^* = (X_0, X_1, X_2, \ldots, X_p)$. Suppose $p(x_i)$ defines the probability of default for the i^{th} individual company, then

$$p(x_i) = w_0 + w_1x_{i1} + w_2x_{i2} + \ldots + w_px_{ip} + \epsilon_i. \tag{3.22}$$

The linear regression has some primary assumptions, namely:

- The relationship between probability of default and explanatory variables is linear
- Or at least it is well approximated by a straight line.

- The error term ϵ has zero mean.
- The error term ϵ has constant variance.
- The errors are uncorrelated.
- The errors are normally distributed.

Suppose n_D of the training set are default companies and n_{ND} ones are non-default companies. We denote the default and non-default companies by 1 and 0, respectively. That is, $p(x_i) = 1$ when $i = 1, 2, \ldots, n_D$ and $p(x_i) = 0$ when $i = n_D + 1, n_D + 2, \ldots, n_D + n_{ND}$, and $n = n_D + n_N D$. Then, our aim is to find the best set of weights, i.e., the one which satisfies

$$min_w \sum_{i}^{n} \epsilon_i^2. \tag{3.23}$$

After all equations we reach the matrix display ;

$$\begin{pmatrix} 1_D & X_D \\ 1_{ND} & X_{ND} \end{pmatrix} \begin{pmatrix} w_0 \\ w^T \end{pmatrix} = \begin{pmatrix} 1_D \\ 0 \end{pmatrix} \tag{3.24}$$

or

$$Xw^T = y^T. \tag{3.25}$$

3.3.2. Advantages and Disadvantages of Regression

As very important advantages of regression, we note:

- Estimation under the usual assumptions are used for process modelling.
- Good results can be obtained with relatively small data sets.
- It can be constructed of different types of easily-interpretable statistical intervals.

As the disadvantages of regression we state :

- Outputs of regression can lie outside of the range [0,1].
- It has limitations in the shapes that linear models can assume over long ranges.
- The extrapolation properties will be possibly poor.
- It is very sensitive to outliers.
- It often gives optimal estimates of the unknown parameters.

3.4. Logistic regression

3.4.1. Introduction

Logistic regression is a form of regression which is used when the dependent variable is a binary or dichotomous and the independents are of any type. In any regression analysis, the main feature is to find the expected value of the dependent variable under the known explanatory variables, i.e., $E(Y|x)$, where Y and x denote the dependent and the vector of explanatory variables, respectively. Let us use the notation $p(x_z) = E(Y|x_z)$ being the probability of default for the i^{th} individual company. Then, the form of the logistic regression model is in logit transformation as;

$$ln(\frac{p(x_i)}{1-p(x_j)}) = w_0 + w_1x_{i1} + w_2x_{i2} + \ldots + w_px_{ip} + \epsilon_i = x_iw + \epsilon_i. \tag{3.26}$$

Here, the estimation of the coefficients of a logistic regression model is done with the help of the *maximum likelihood estimation* (MLE). MLE primarily states that the coefficients are estimated in a way in which the likelihood function is minimized. In order to obtain a likelihood function, we should firstly introduce the probability mass function of Y_z. Since Y_z follows a Bernolli distribution, the probability mass function for the i^th company can be written as

$$p(x_i)^{y_i}(1-p(x_i))^{1-y_i}. \tag{3.27}$$

If we assume that all observations are independently distributed, the likelihood function expression will be

$$L(w) := \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i} \tag{3.28}$$

Since these equations are nonlinear in w, it is not possible to solve them directly. Therefore, the solution of it is usually made with the well-known nonlinear optimization method called Gauss-Newton algorithm.

3.4.2. Advantages and Disadvantages of Logistic Regression

Advantages :

- Useful for when appropriate, especially, combined with feature creation and selection.
- Constructed probabilities have chance of being meaningful.
- It is modelled as a function directly rather than as ratio of two densities.
- Scores are interpretable in terms of log odds.

Disadvantages :

- It invites to an over interpretation of some parameters.

4. APPLICATION AND IMPLEMENTATION

4.1. Introduction

In this section we will focus on the designing the software application and its all fundemantal basements to implement variant situations. First of all the using the reliable infrastructure such as Statistical Package for the Social Sciences (SPSS) gives us the powerful and strength model design software. After all method analyzing and critisizing the results integrates on the main programming segment. .NET and Visual Basic Environment is the useful application ambient for our Hybrid Model Design. So we will use the dataset of German Data in order to design our model.

4.1.1. Dataset Description

- Number Of Examples: 1000
- Number Of Classes : 2
- Class 1 is 700 good credit
- Class 2 is 300 bad credit
- Number of Properties : 20 (7 numerical, 13 categorical)

4.1.2. Attributes Description

Attribute 1: (qualitative) Status of existing checking account

- A11 : ... $\leq$ 0 DM
- A12 : 0 $\leq$ ... $\leq$ 200 DM
- A13 : ... $\geq$ 200 DM
- A14 : no checking account

Attribute 2: (numerical) Duration in month

Table 4.1. German Data

Variable #	**Variable Name**	**Variable Type**
1	Status of checking account	Qualitative
2	Duration in month	Numerical
3	Credit history	Qualitative
4	Purpose	Qualitative
5	Credit amount	Numerical
6	Savings account/bonds	Qualitative
7	Present employment	Qualitative
8	Instalment rate	Numerical
9	Personal status and sex	Qualitative
10	Other debtors/guarantors	Qualitative
11	Present residence since	Numerical
12	Property	Qualitative
13	Age in years	Numerical
14	Other instalment plans	Qualitative
15	Housing	Qualitative
16	Number of existing credits	Numerical
17	Job	Qualitative
18	Number of people being liable	Numerical
19	Telephone	Qualitative
20	Foreign worker	Qualitative

Attribute 3: (qualitative) Credit history

- A30 : no credits taken
- A31 : all credits at this bank paid back duly
- A32 : existing credits paid back duly till now
- A33 : delay in paying off in the past
- A34 : critical account/other credits existing (not at this bank)

Attribute 4: (qualitative) Purpose

- A40 : car (new)
- A41 : car (used)
- A42 : furniture/equipment
- A43 : radio/television
- A44 : domestic appliances
- A45 : repairs
- A46 : education
- A47 : (vacation - does not exist)
- A48 : retraining
- A49 : business

Attribute 5: (numerical) Credit amount

Attribute 6: Savings account/bonds

- A61 : ... $\leq$ 100 DM
- A62 :101 $\leq$ ... $\leq$ 500 DM
- A63 :501 $\leq$ ... $\leq$ 1000 DM
- A64 : ... $\geq$ 1001 DM
- A65 :unknown/ no savings account

Attribute 7: (qualitative) Present employment since

- A71 : unemployed
- A72 : $\ldots \leq 1$ year
- A73 : $2 \leq \ldots \leq 4$ years
- A74 : $5 \leq \ldots \leq 7$ years
- A75 : $\ldots \geq 7$ years

Attribute 8: (numerical) Installment rate in percentage of disposable income

Attribute 9: (qualitative) Personal status and sex

- A91 : male : divorced/separated
- A92 : female : divorced/separated/married
- A93 : male : single
- A94 : male : married/widowed
- A95 : female : single

Attribute 10:(qualitative) Other debtors / guarantors

- A101 : none
- A102 : co-applicant
- A103 : guarantor

Attribute 11:(numerical) Present residence since

Attribute 12:(qualitative) Property

- A121 : real estate
- A122 : if not A121 : building society savings agreement/life insurance
- A123 : if not A121/A122 : car or other, not in attribute 6
- A124 : unknown / no property

Attribute 13:(numerical) Age in years

Attribute 14:(qualitative) Other installment plans

- A141 : bank
- A142 : stores
- A143 : none

Attribute 15: (qualitative) Housing

- A151 : rent
- A152 : own
- A153 : for free

Attribute 16:(numerical) Number of existing credits at this bank

Attribute 17:(qualitative) Job

- A171 : unemployed / unskilled
- A172 : unskilled
- A173 : skilled employee / official
- A174 : management/ self-employed

Attribute 18:(numerical) Number of people being liable to provide care for

Attribute 19:(qualitative) Telephone

- A201 : yes
- A202 : no

Attribute 20:(qualitative) foreign worker

- A201 : yes
- A202 : no

4.2. Binary Logistic (Logistic Regression) on SPSS

Logical regression is used in situations where an attribute or the existence or lackness of the outputs from the descriptive variables are wanted to be predicted. It bears a resemblance to linear regression model. Its difference is that it is more suitable when the dependent variable is double-choice. Also our model is double-choice (good and bad). Logical regression is used in a wider area than discriminant analysis.In SPSS integrated environment results will be as;
CRITERIA = PIN(.05) POUT(.10) ITERATE(20) CUT(.5)
LOGISTIC REGRESSION VARIABLES default
/METHOD = ENTER:
checking, duration, history,purpose, Amount_new, savings
employed, installp, marital, coapp, resident, property age_new, other, housing
existcr, job, depends, telephon, foreign.

If weight is in effect, see classification table for the total number of cases.
N is the count of the dataset members. So here is the 1000 the number of total interested in theGerman Data. We havent got any missing cases. So missing case count is zero.

Block 0: Beginning Block

Constant is included in the model.The cut value is ,500. So we can realize that 700 is the number of the estimation correct overall the dataset of German data. Thus the BL Regression is the effective model especially for granting results

Block 1: Method = Enter

In order to find significance dataset is being tested. Having the eminent seperative property significance value must be low. Otherwise binary score would be lowest value. Thus we could select the best variable. f we have much data in German Data it could be subscribed the highest significance values and repeated the alanyze again and then again.

Table 4.2. Logistic Regression's Notes

Output Created		12-JUN-2007 07:52:31
Comments		
Input	Data Active Dataset Filter Weight Split File N of Rows in Working Data File	C:\germandata.sav DataSet1 none none none 1000
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing
Syntax		LOGISTIC REGRESSION VARIABLES default /METHOD = ENTER checking duration history purpose Amount_new savings employed installp marital coapp resident property age_new other housing existcr job depends telephon foreign /CRITERIA = PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
Resources	Processor Time Elapsed Time	00:00,1 00:00,3
	[DataSet1] C:/germandata.sav	

Table 4.3. Case Processing Summary

Unweighted Cases(a)		N	Percent
Selected Cases	Included in Analysis	1000	100
	Missing Cases	0	0
	Total	1000	100
Unselected Cases		0	0
Total		1000	100

Table 4.4. Dependent Variable Encoding

Original Value	**Internal Value**
0	0
1	1

Variable(s) entered on: checking, duration, history, purpose, Amount_new, savings, employed, installp, marital, coapp, resident, property, age_new, other, housing, existcr, job, depends, telephon, foreign. Basement of the system depends on the reality of the variables. So the coefficient can be occured only the accurant operations all over these variables. However the situation on the dataset can not be changed if the fields modify accordingly. Unless it is necessary it will not be offered.

4.3. Linear Discriminant Analysis

Discriminant analysis is useful for forming a model for predicting if any observed attribute/feature belongs to a group or not. The procedure creates a discriminant

Table 4.5. Classification Table(a,b)

	Observed		Predicted		
			default		Percentage Correct
			0	1	0
Step 0	default	0	700	0	100
		1	300	0	0
	Overall Percentage				70

Table 4.6. Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-0,847	0,069	150,76	1	0	0,429

Table 4.7. Variables in the Equation

			Score	df	Sig.
Step 0	Variables	checking	123,09	1	0
		duration	46,193	1	0
		history	52,342	1	0
		purpose	0,227	1	0,634
		Amount_new	6,503	1	0,011
		savings	32,021	1	0
		employed	13,456	1	0
		installp	5,242	1	0,022
		marital	7,776	1	0,005
		coapp	0,632	1	0,427
		resident	0,009	1	0,925
		property	20,338	1	0
		age_new	5,86	1	0,015
		other	12,066	1	0,001
		housing	0,373	1	0,541
		existcr	2,091	1	0,148
		job	1,072	1	0,301
		depends	0,009	1	0,924
		telephon	1,33	1	0,249
		foreign	6,737	1	0,009
	Overall Statistics		236,91	20	0

Table 4.8. Classification Table(a)

	Observed		Predicted		
			default		Percentage Correct
			0	1	0
Step 1	Default	0	628	72	89.7
		1	156	144	48
	Overall Percentage				77.2

Table 4.9. Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	checking	-0.585	0.07	69.476	1	0	0.557
	duration	0.043	0.009	22.836	1	0	1.044
	history	-0.39	0.086	20.353	1	0	0.677
	purpose	-0.037	0.031	1.439	1	0.23	0.963
	Amount_new	-0.086	0.084	1.052	1	0.305	0.917
	savings	-0.232	0.058	15.854	1	0	0.793
	employed	-0.171	0.071	5.856	1	0.016	0.843
	installp	0.183	0.083	4.847	1	0.028	1.2
	marital	-0.253	0.115	4.872	1	0.027	0.776
	coapp	-0.362	0.178	4.145	1	0.042	0.696
	resident	0.003	0.077	0.001	1	0.972	1.003
	property	0.215	0.091	5.54	1	0.019	1.24
	age_new	0.087	0.102	0.732	1	0.392	1.091
	other	-0.315	0.11	8.188	1	0.004	0.73
	housing	-0.298	0.167	3.171	1	0.075	0.743
	existcr	0.275	0.161	2.925	1	0.087	1.316
	job	0.062	0.137	0.202	1	0.653	1.064
	depends	0.148	0.233	0.403	1	0.526	1.16
	telephon	-0.213	0.184	1.35	1	0.245	0.808
	foreign	-1.128	0.598	3.555	1	0.059	0.324
	Constant	3.825	1.123	11.609	1	0.001	45.847

equation which develops the most suitable linear combination provided by the determining variable. In usinf of integrated SPSS environment the results will be as ;
DISCRIMINANT
/GROUPS=default(0 1) /VARIABLES=checking, duration, history, purpose, savings, Amount_new, employed, installp, marital, coapp, resident, property, age_new, other, housing, existcr, job, depends, telephon, foreign.
/ANALYSIS ALL
/PRIORS EQUAL
/STATISTICS=MEAN STDDEV COEFF TABLE CROSSVALID
/PLOT=COMBINED SEPARATE MAP
/CLASSIFY=NONMISSING POOLED .

Pooled within groups correlations between discriminating variables and standardized canonical discriminant functions examined on the system. Thus, variables ordered by absolute size of correlation within function. However, unstandardized canonical discriminant functions evaluated at group means.

We could change the priorities over the German data. However we do not need to change it because of the default values are the same importance for us Default variables gives the result of the data set about the individuals who are granted or rejected. So half and than half is the right choice for analyze.

Fisher's linear discriminant functions

Default values 0 and 1 gives us the sum of the granted or not. So the weight for every variable has the importance over the discriminant eminent exclusivenes about the each variable. Here is the constant comes from the Standard analysis extensions for the design.

Separate-Groups Graphs

Table 4.10. Discriminant Analysis Notes

Output Created		12-JUN-2007 07:36:09
Comments		
Input	Data	C:/germandata.sav
	Active Dataset	DataSet1
	Filter	none
	Weight	none
	Split File	none
	N of Rows in Working Data File	1000
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics are based on cases with no missing values for any variable used.
Syntax		DISCRIMINANT /GROUPS=default(0 1) /VARIABLES=checking variotans listed not crypted below /NOORIGIN /DEPENDENT default checking duration history purpose Amount_new savings employed installp marital coapp resident property

Table 4.11. Additional Notes

Syntax	existcr job depends telephon foreign /ANALYSIS ALL /PRIORS EQUAL /STATISTICS=MEAN STDDEV COEFF TABLE CROSSVALID /PLOT=COMBINED SEPARATE MAP /CLASSIFY=NONMISSING POOLED .
Resources Processor Time	00:01.3
Elapsed Time	00:01.5
[DataSet1] C:/germandata.sav	

Table 4.12. Analysis Case Processing Summary

	Unweighted Cases	N	Percent
Valid		1000	100
Excluded	Missing or out-of-range group codes	0	0
	At least one missing discriminating variable	0	0
	At least one missing discriminating variable	0	0
	Total	0	0
Total		1000	100

Table 4.13. Mean and Std. Deviation Results for 0

default		Mean	Std. Dev.	Valid N (listwise)	
		Unweighted	Weighted	Unweighted	Weighted
0	checking	2.87	1.229	700	700
	duration	19.21	11.08	700	700
	history	2.71	1.045	700	700
	purpose	2.77	2.574	700	700
	Amount_new	2.92	1.372	700	700
	savings	2.29	1.651	700	700
	employed	3.48	1.19	700	700
	installp	2.92	1.128	700	700
	marital	2.72	0.691	700	700
	coapp	1.15	0.5	700	700
	resident	2.84	1.108	700	700
	property	2.26	1.038	700	700
	age_new	2.77	0.887	700	700
	other	2.73	0.659	700	700
	housing	1.94	0.493	700	700
	existcr	1.42	0.585	700	700
	job	2.89	0.647	700	700
	depends	1.16	0.363	700	700
	telephon	1.42	0.493	700	700
	foreign	1.05	0.212	700	700

Table 4.14. Mean and Std. Deviation Results for 1

default		Mean	Std. Dev.	Valid N (listwise)	
		Unweighted	Weighted	Unweighted	Weighted
1	checking	1.9	1.051	300	300
	duration	24.86	13.283	300	300
	history	2.17	1.078	300	300
	purpose	2.85	2.882	300	300
	Amount_new	3.17	1.502	300	300
	savings	1.67	1.303	300	300
	employed	3.17	1.225	300	300
	installp	3.1	1.088	300	300
	marital	2.59	0.738	300	300
	coapp	1.13	0.422	300	300
	resident	2.85	1.095	300	300
	property	2.59	1.045	300	300
	age_new	2.92	0.879	300	300
	other	2.56	0.793	300	300
	housing	1.91	0.611	300	300
	existcr	1.37	0.56	300	300
	job	2.94	0.669	300	300
	depends	1.15	0.361	300	300
	telephon	1.38	0.485	300	300
	foreign	1.01	0.115	300	300

Table 4.15. Group Statistics

Total	Mean	Std.Dev.	Unweighted	Weighted
checking	2.58	1.258	1000	1000
duration	20.9	12.059	1000	1000
history	2.55	1.083	1000	1000
purpose	2.79	2.669	1000	1000
Amount_new	3	1.416	1000	1000
savings	2.11	1.58	1000	1000
employed	3.38	1.208	1000	1000
installp	2.97	1.119	1000	1000
marital	2.68	0.708	1000	1000
coapp	1.15	0.478	1000	1000
resident	2.85	1.104	1000	1000
property	2.36	1.05	1000	1000
age_new	2.81	0.887	1000	1000
other	2.68	0.706	1000	1000
housing	1.93	0.531	1000	1000
existcr	1.41	0.578	1000	1000
job	2.9	0.654	1000	1000
depends	1.16	0.362	1000	1000
telephon	1.4	0.491	1000	1000

Table 4.16. Standardized Canonical Discriminant Function Coefficients

	Function 1
checking	0.6
Duration	-0.453
History	0.36
Purpose	0.071
Amount_new	0.109
Savings	0.261
employed	0.174
nstallp	-0.158
Marital	0.155
Coapp	0.148
Resident	-0.007
Property	-0.206
age_new	-0.042
Other	0.169
Housing	0.142
Existcr	-0.137
Job	-0.029
Depends	-0.055
Telephon	0.101
Foreign	0.108

Table 4.17. Structure Matrix

	Function 1
checking	0.672
History	0.422
Duration	-0.395
Savings	0.326
Property	-0.259
employed	0.21
Other	0.198
Marital	0.159
Foreign	0.148
Amount_new	-0.145
age_new	-0.138
nstallp	-0.13
Existcr	0.082
Telephon	0.065
Job	-0.059
Coapp	0.045
Housing	0.035
Purpose	-0.027
Depends	0.005
Resident	-0.005

Table 4.18. Functions at Group Centroids

default	Function 1
0	0.364
1	-0.85

Table 4.19. Classification Processing Summary

Processed		1000
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0
Used in Output		1000

Table 4.20. Prior Probabilities for Groups

default	Prior	Cases Used in Analysis	
	Unweighted	Weighted	Unweighted
0	1	700	700
1	1	300	300
Total	1	1000	1000

a Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case. **b** 72.7% of original grouped cases correctly classified. **c** 71.5% of cross-validated grouped cases correctly classified.

4.4. Linear Regression

Regression analysis is writing a mathematical function for the relationship between a variable which is dependent and one or more variables which are independent explanatory. This function is called the regression equation. By this equation, the value of the dependent variable is predicted by changing the values of independent variables. Also the determination of independent variables affective on the dependent variable puts through which variables are more important. In using of SPSS integrated environment results will be as ;

GET FILE='C:/germandata.sav'.

DATASET NAME DataSet1 WINDOW=FRONT.

REGRESSION

Table 4.21. Classification Function Coefficients

	default values	
	0	1
checking	2.196	1.578
Duration	-0.119	-0.073
History	0.798	0.384
Purpose	0.84	0.807
Amount_new	1.587	1.493
Savings	0.65	0.446
employed	1.185	1.009
nstallp	2.54	2.712
Marital	3.253	2.986
Coapp	6.316	5.938
Resident	3.051	3.059
Property	0.185	0.426
age_new	6.788	6.845
Other	6.539	6.246
Housing	8.638	8.314
Existcr	3.234	3.523
Job	5.46	5.514
Depends	6.654	6.838
Telephon	2.803	2.553
Foreign	34.361	33.667
(Constant)	-87.06	-82.7

Table 4.22. Classification Results

		default	Predicted Group Membership		Total
			0	1	0
Original Val	Count	0	505	195	700
		1	78	222	300
	%	0	72.1	27.9	100
		1	26	74	100
Cross-validated(a)	Count	0	495	205	700
		1	80	220	300
	%	0	70.7	29.3	100
		1	26.7	73.3	100

/MISSING LISTWISE

/STATISTICS COEFF OUTS R ANOVA

/CRITERIA=PIN(.05) POUT(.10)

/NO ORIGIN

/DEPENDENT default

/METHOD=ENTER checking, duration, history, purpose, Amount_new, savings, employed, installp, marital, coapp, resident, property, age_new, other, housing existcr, job, depends, telephon, foreign.

/PARTIALPLOT ALL

/RESIDUALS HIST(ZRESID) NORM(ZRESID) .

Table 4.23. Regression Notes

Output Created		12-JUN-2007 07:36:09
Comments		
Input	Data	C:/germandata.sav
	Active Dataset	DataSet1
	Filter	none
	Weight	none
	Split File	none
	N of Rows in Working Data File	1000
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics are based on cases with no missing values for any variable used.
Syntax		REGRESSION /MISSING LISTWISE /STATISTICS COEFF OUTS R ANOVA /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT default /METHOD=ENTER checking duration history purpose Amount_new savings employed installp marital coapp resident property

Table 4.24. Regression Notes Extra

		age_new other housing existcr job depends telephon foreign /PARTIALPLOT ALL /RESIDUALS HIST(ZRESID) NORM(ZRESID) .
Resources	Processor Time	00:18.1
	Elapsed Time	00:22.5
	Memory Required	12236 bytes
	Additional Memory Required for Residual Plots	16272 bytes
	[DataSet1] C:/germandata.sav	

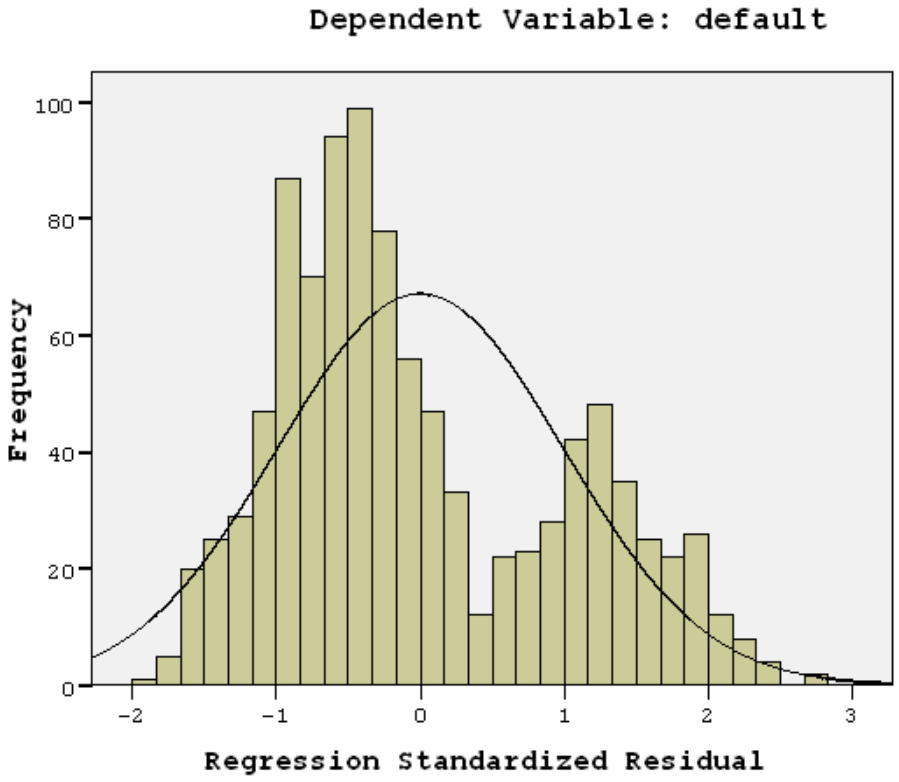

Figure 4.1. Histogram

Table 4.25. Variables on Method

Model	Variables Entered	Variables Removed	Method
1	foreign, savings, other, existcr,, installp, housing, resident, purpose, job, marital, coapp, depends, checking, duration, employed, age_new, telephon, history, property, Amount_new(a).		Enter

Table 4.26. Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	(Constant)	1.048	0.166		6,328	0
	checking	-0.099	0.011	-0.272	-9,146	0
	duration	0.007	0.001	0.197	5,102	0
	history	-0.067	0.014	-0.157	-4,836	0
	purpose	-0.005	0.005	-0.03	-1,049	0,294
	Amount_new	-0.015	0.013	-0.047	-1,122	0,262
	savings	-0.033	0.008	-0.113	-3.854	0
	employed	-0.028	0.012	-0.074	-2.455	0.014
	installp	0.028	0.013	0.067	2.077	0.038
	marital	-0.043	0.019	-0.066	-2.296	0.022
	coapp	-0.061	0.028	-0.063	-2.182	0.029
	resident	0.001	0.012	0.003	0.101	0.919
	property	0.039	0.014	0.089	2.682	0.007
	age_new	0.009	0.016	0.018	0.578	0.563
	other	-0.047	0.019	-0.072	-2.52	0.012
	housing	-0.052	0.027	-0.06	-1.893	0.059
	existcr	0.046	0.025	0.058	1.828	0.068
	job	0.009	0.023	0.012	0.379	0.705
	depends	0.03	0.037	0.023	0.805	0.421
	telephon	-0.04	0.029	-0.043	-1.379	0.168
	foreign	-0.111	0.07	-0.046	-1.586	0.113

Table 4.27. Residuals Statistics(a)

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-0.4	0.9	0.3	0.223	1000
Residual	-0.757	1.108	0	0.401	1000
Std. Predicted Value	-3.116	2.694	0	1	1000
Std. Residual	-1.871	2.739	0	0.99	1000

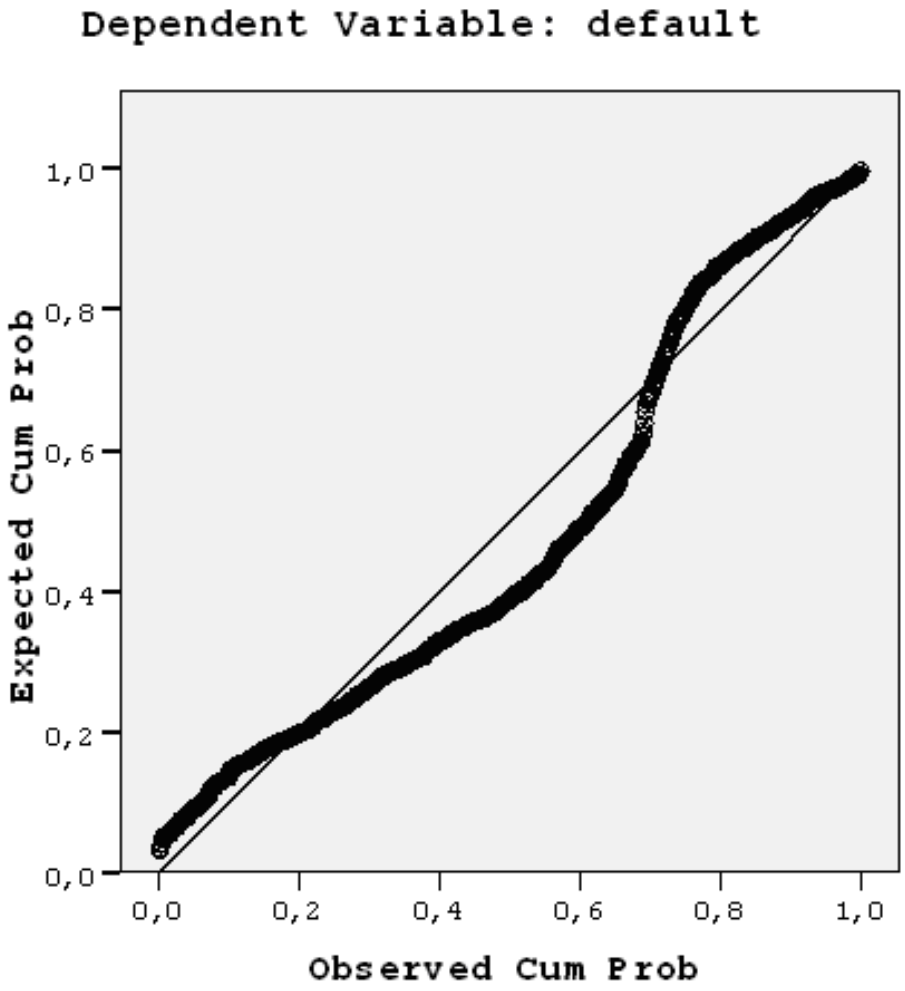

Figure 4.2. Normal P-P Plot

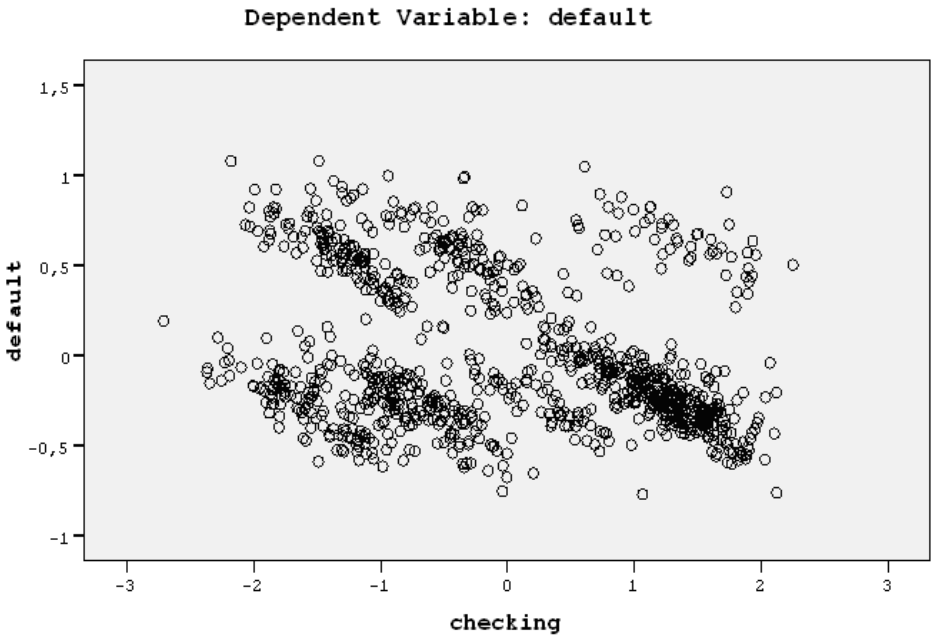

Figure 4.3. PRP Checking

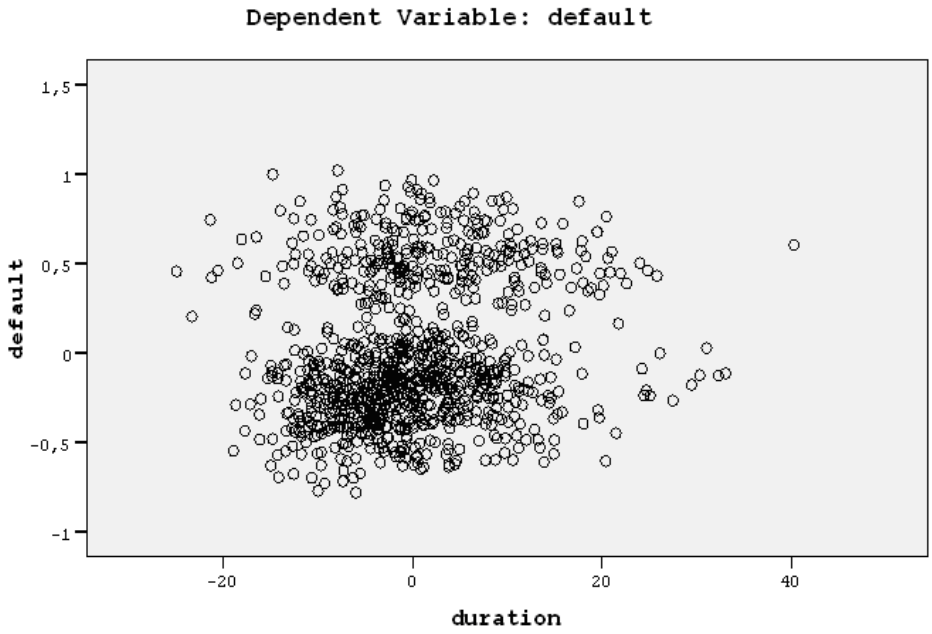

Figure 4.4. PRP Duration

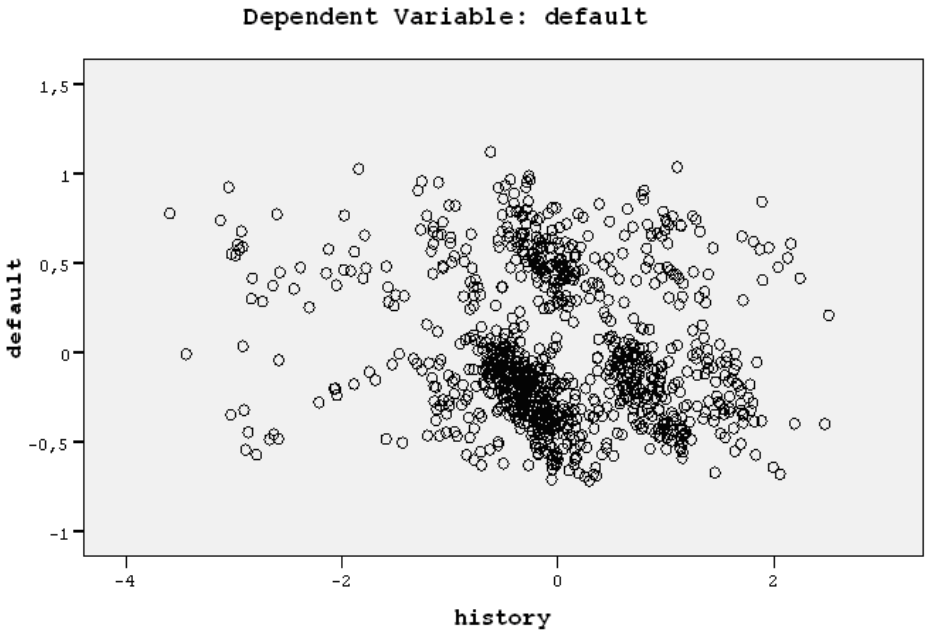

Figure 4.5. PRP History

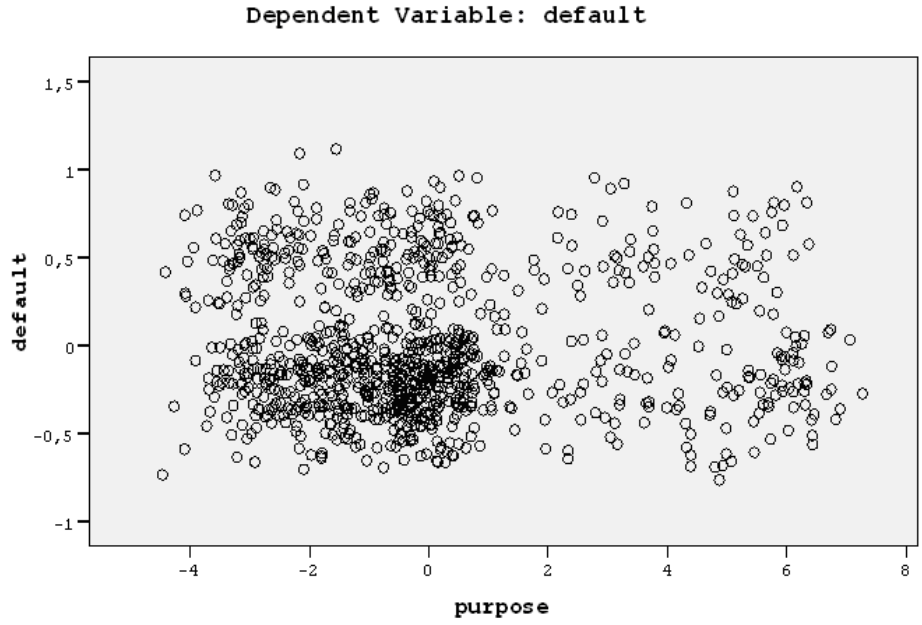

Figure 4.6. PRP Purpose

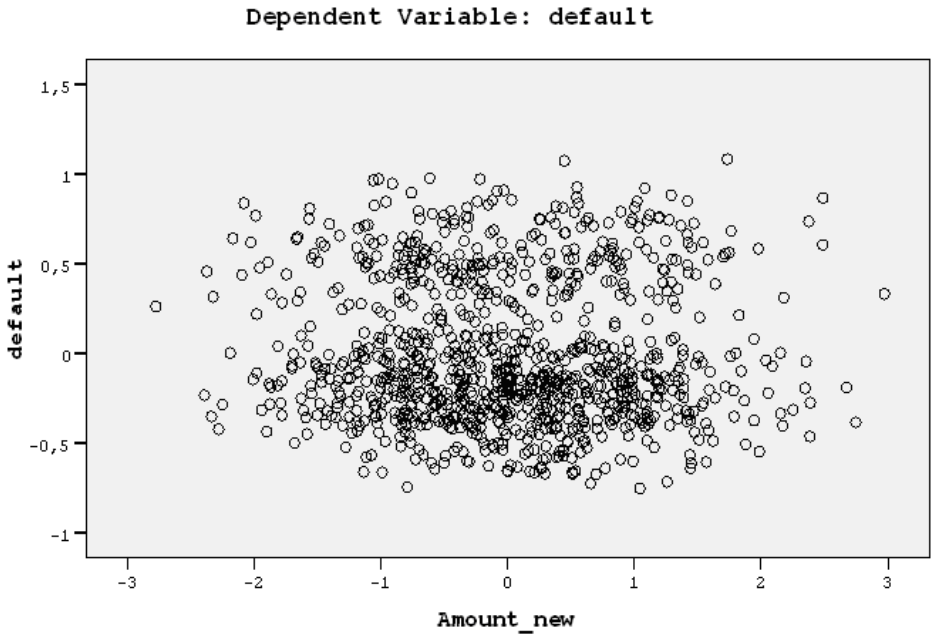

Figure 4.7. PRP Amount New

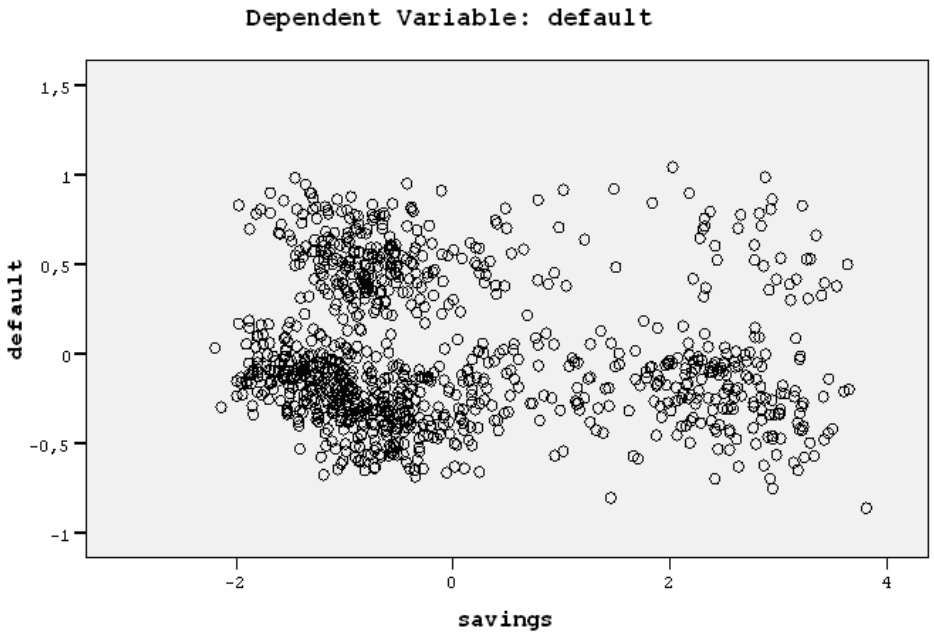

Figure 4.8. PRP Savings

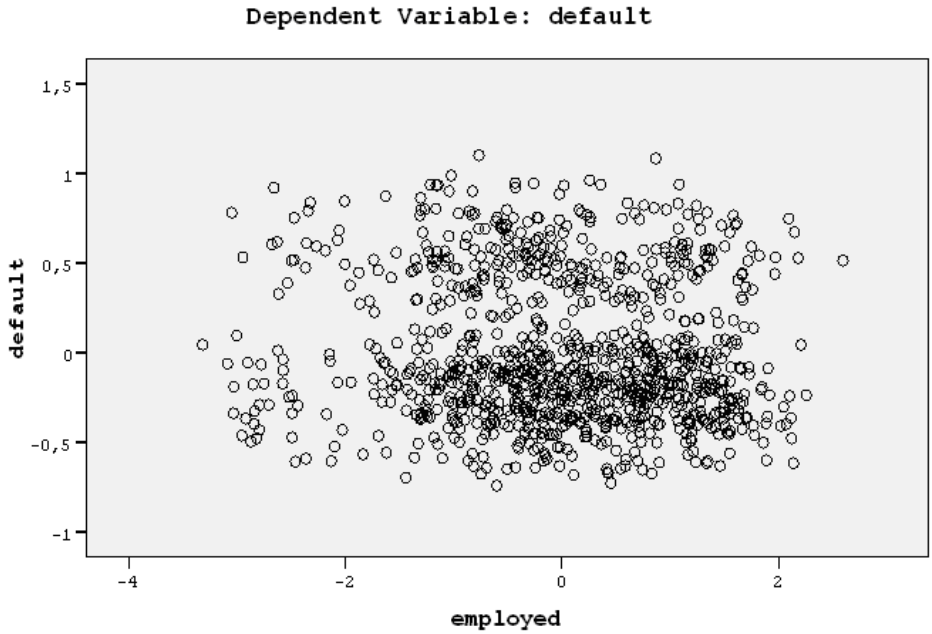

Figure 4.9. PRP Employed

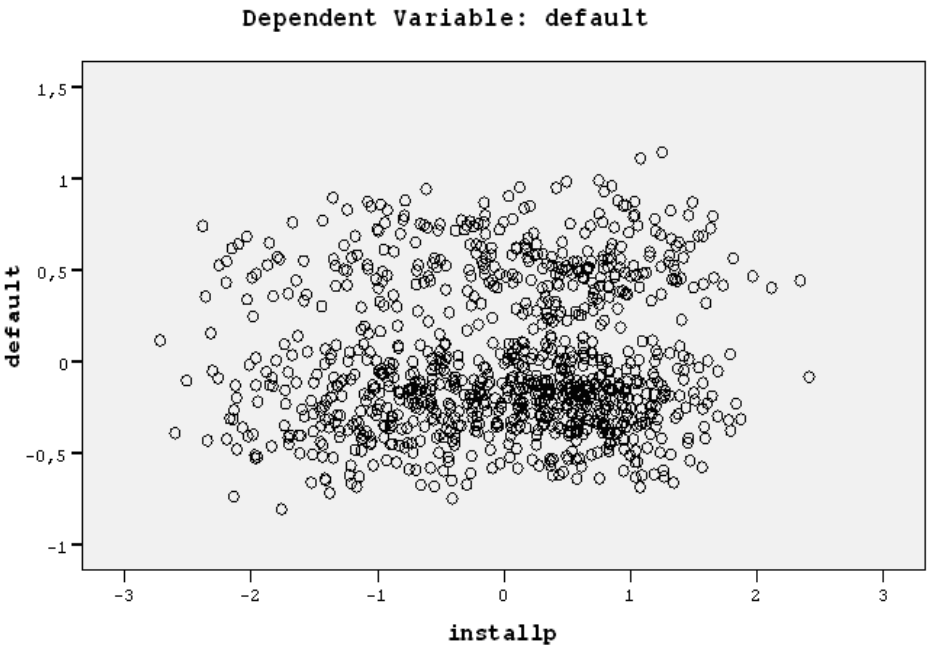

Figure 4.10. PRP Installp

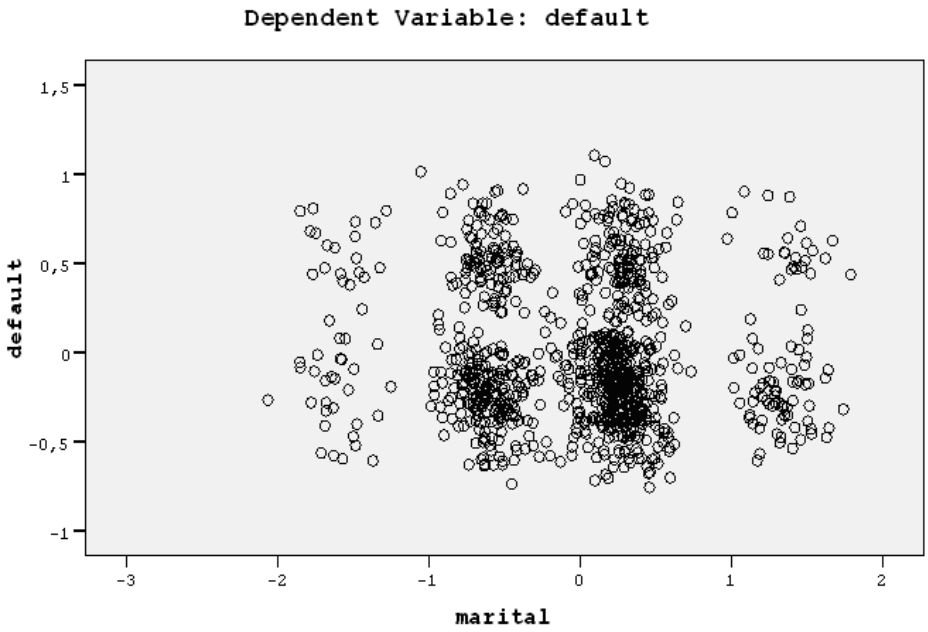

Figure 4.11. PRP Marital

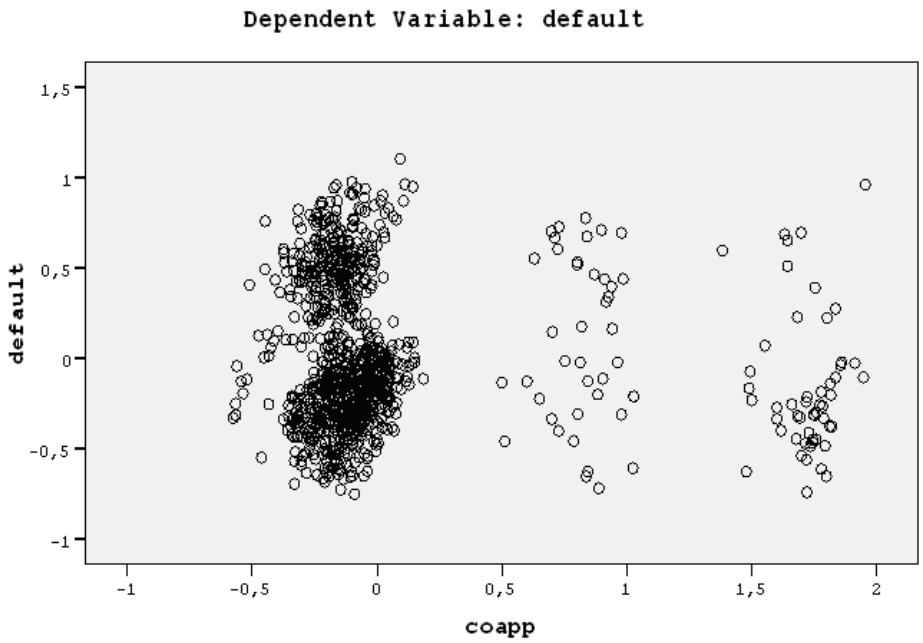

Figure 4.12. PRP Coapp

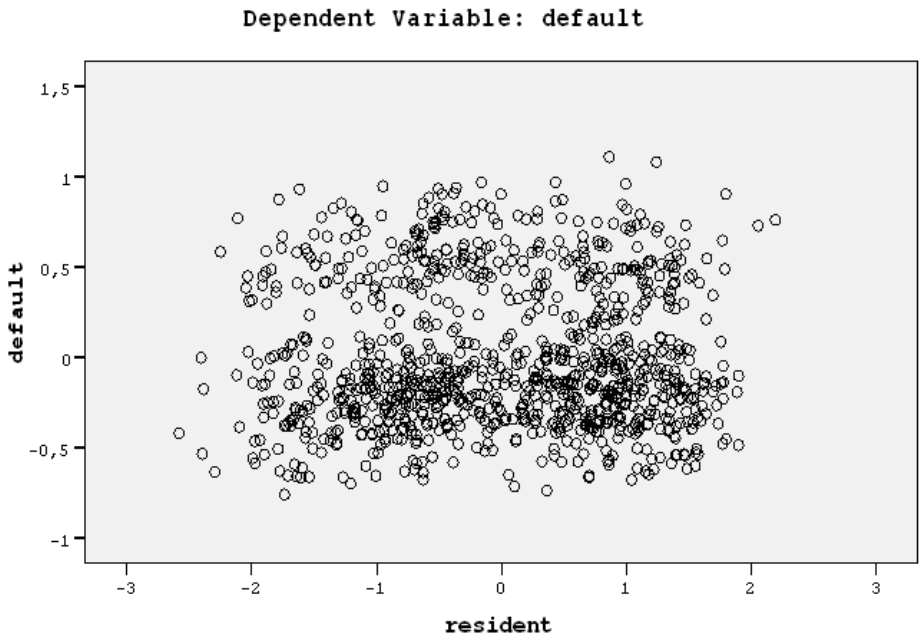

Figure 4.13. PRP Resident

All these binary comparing charts shows the relation between the dependent variable default cross the other variables. Here is the default variable indicates the credit results granted or not by 1 or 0. Thus it can be realized that the seperatable imaginary line always horizontally divides every chart on the cut off point. Therefore over the cut off point results shows us the rejected credit applications and under the cut off point resuls indicates us the granted credit applications.

4.4.1. Analysis of Charts

The values from the SPSS environment we can see that they have symetric form according to each other. Especially the good values and the bad values grouped together. Thus we can realize that the system works efficiently because of all these results. On the other hand the good group has the dominant shape in all the charts. Good results effort to the user grant to the applicant when the new individual score among the zero ones. However the same algorithm is valid for the opposite side which are bad ones. On the other hand there are some insignificant values which is displayed on the charts.But they are negligible.

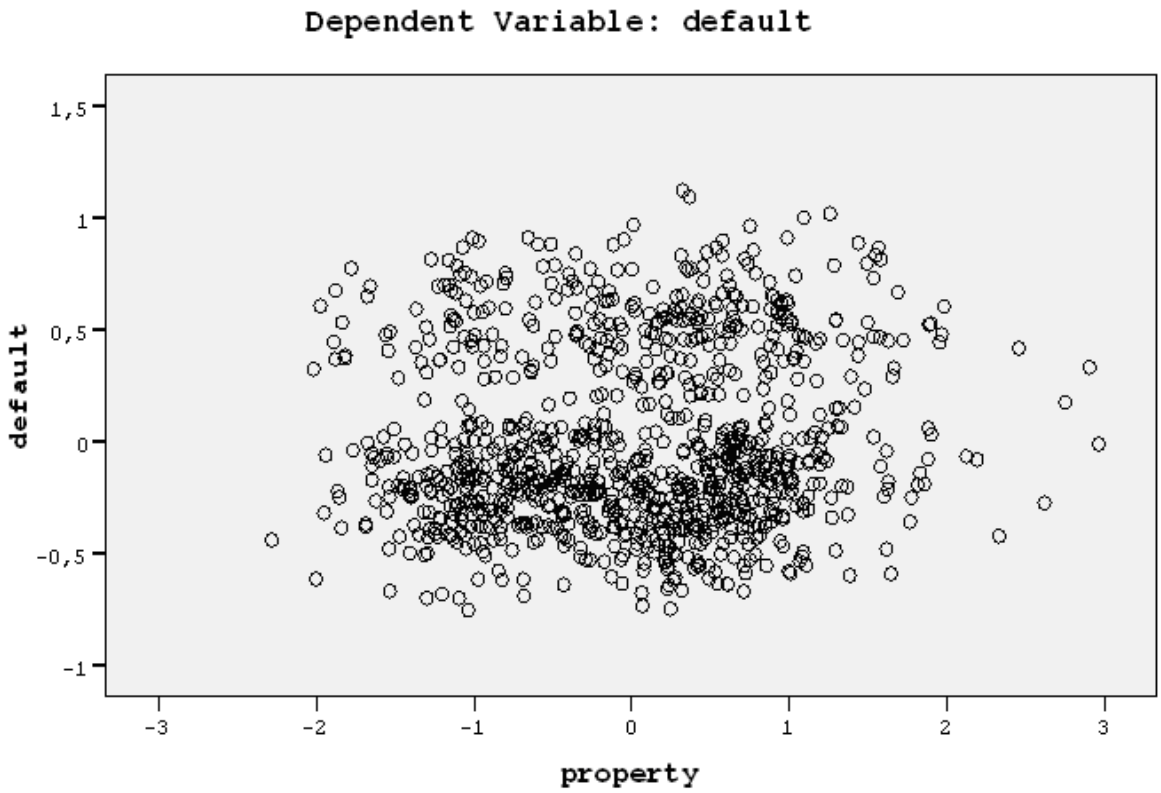

Figure 4.14. PRP Property

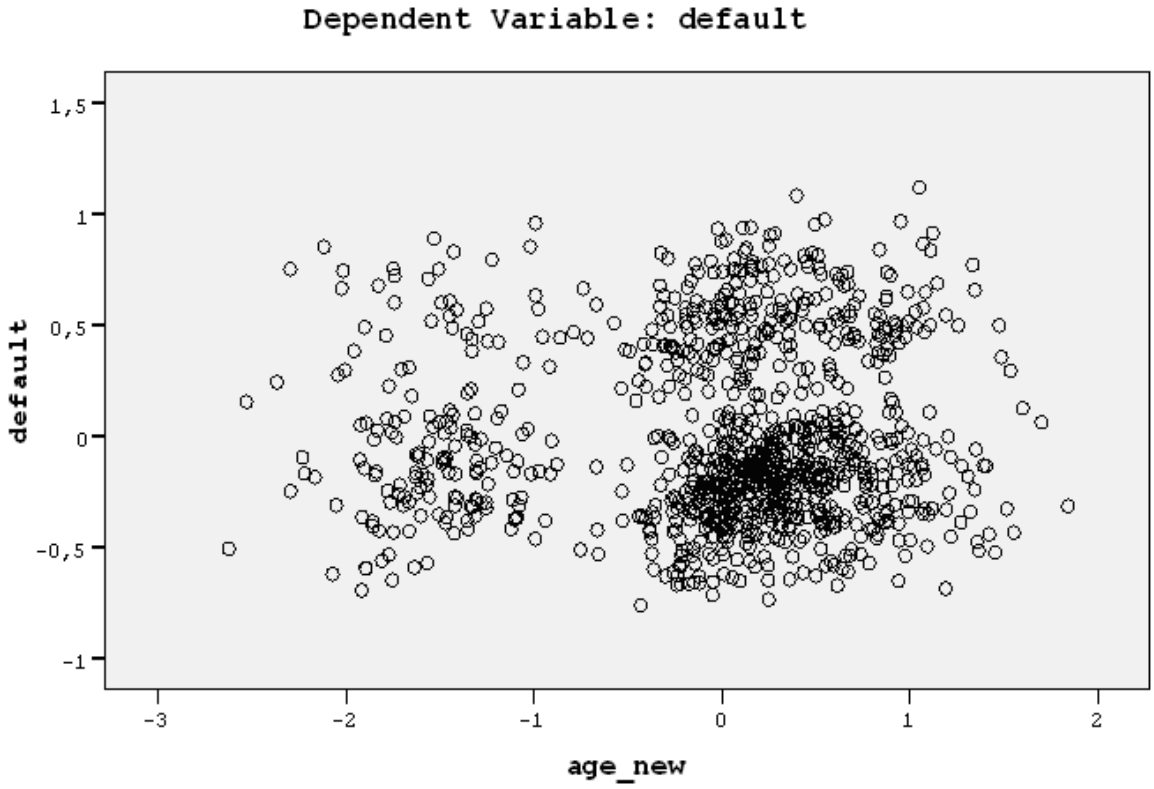

Figure 4.15. PRP Histogram

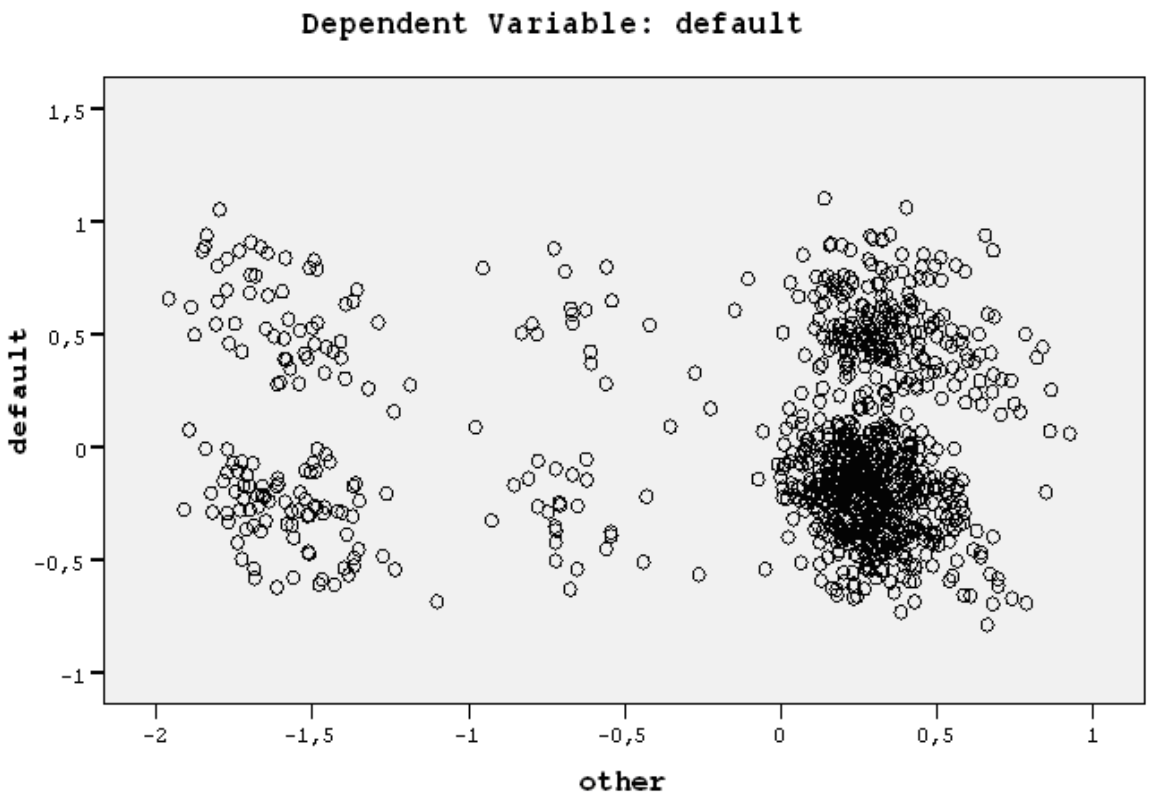

Figure 4.16. PRP Other

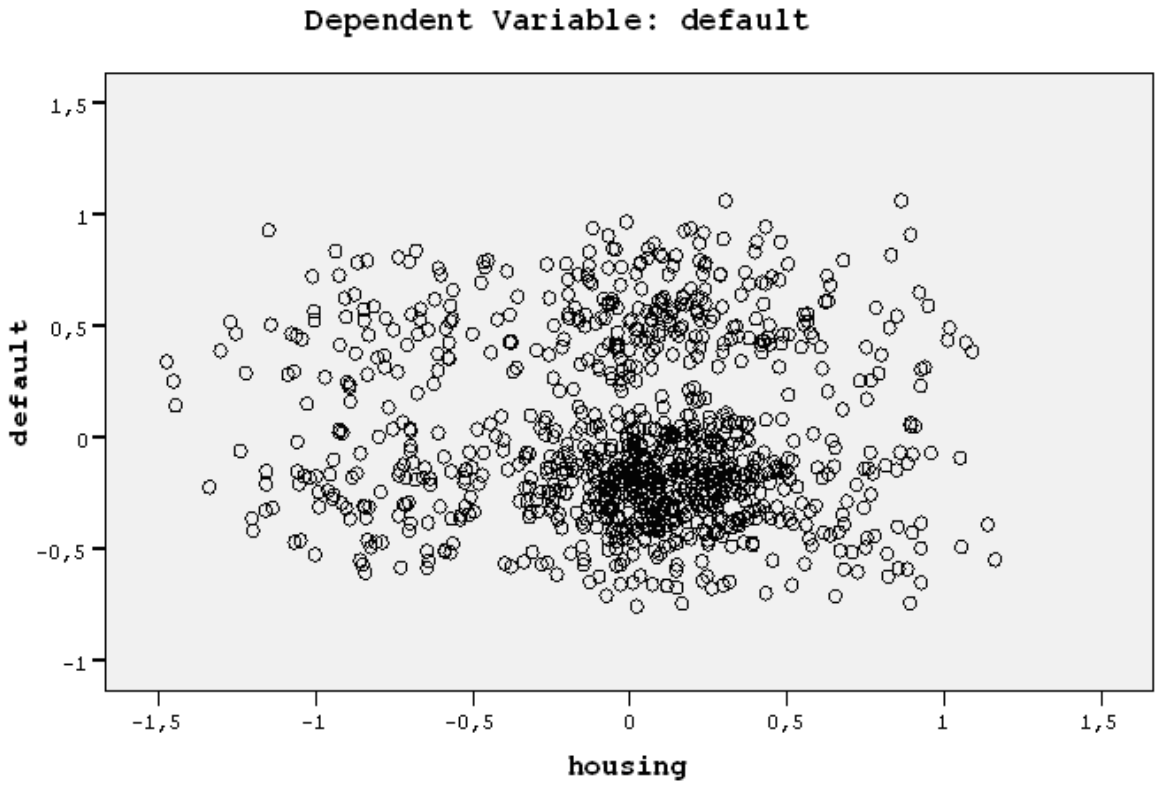

Figure 4.17. PRP Housing

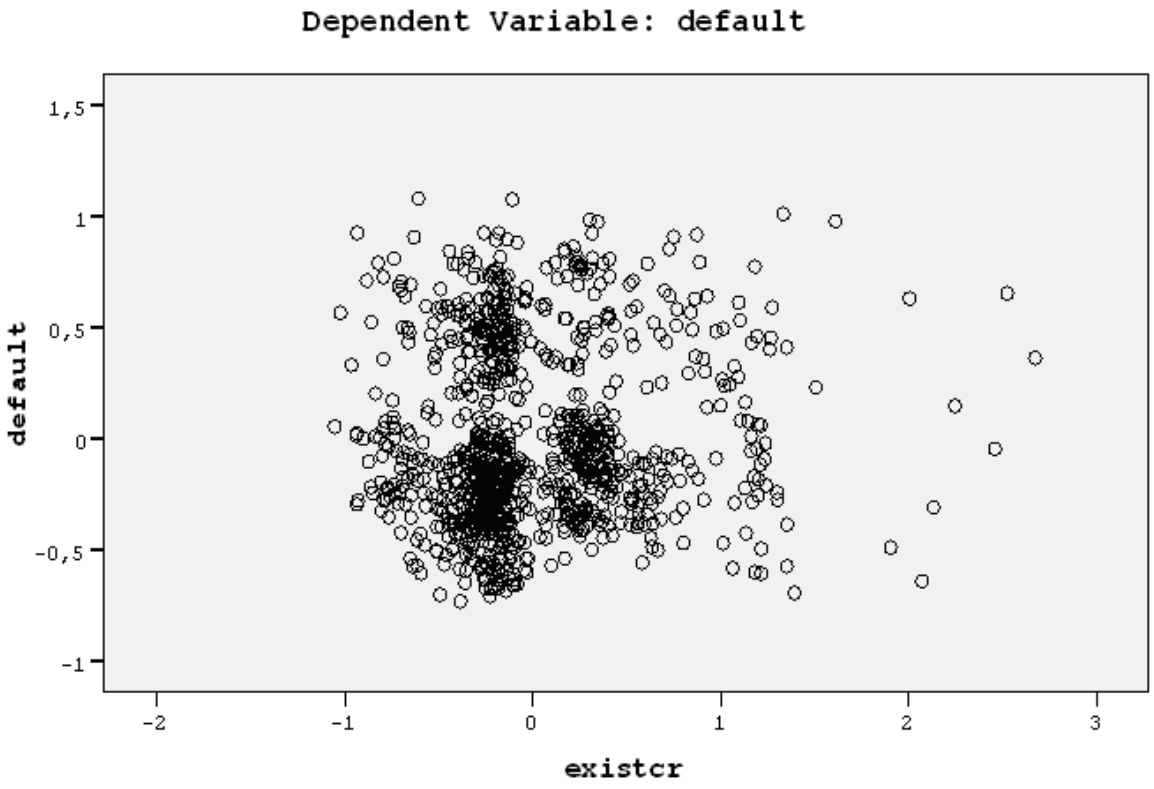

Figure 4.18. PRP Existcr

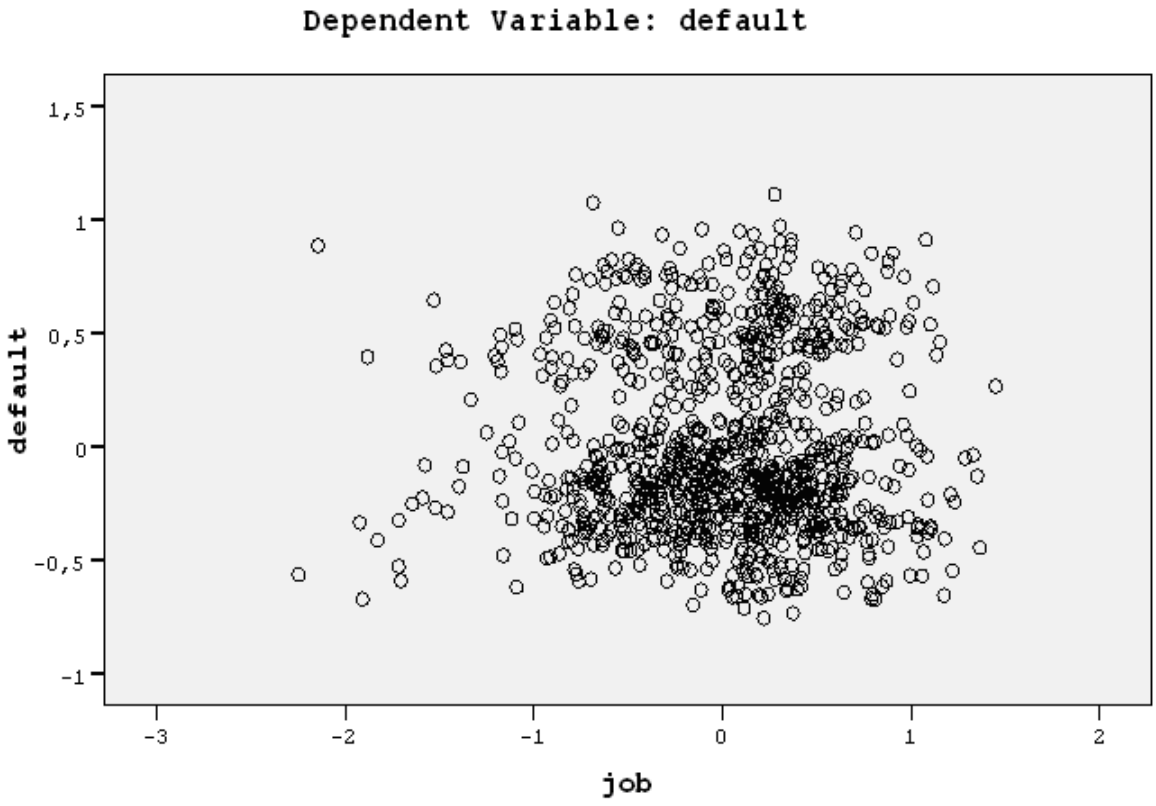

Figure 4.19. PRP Job

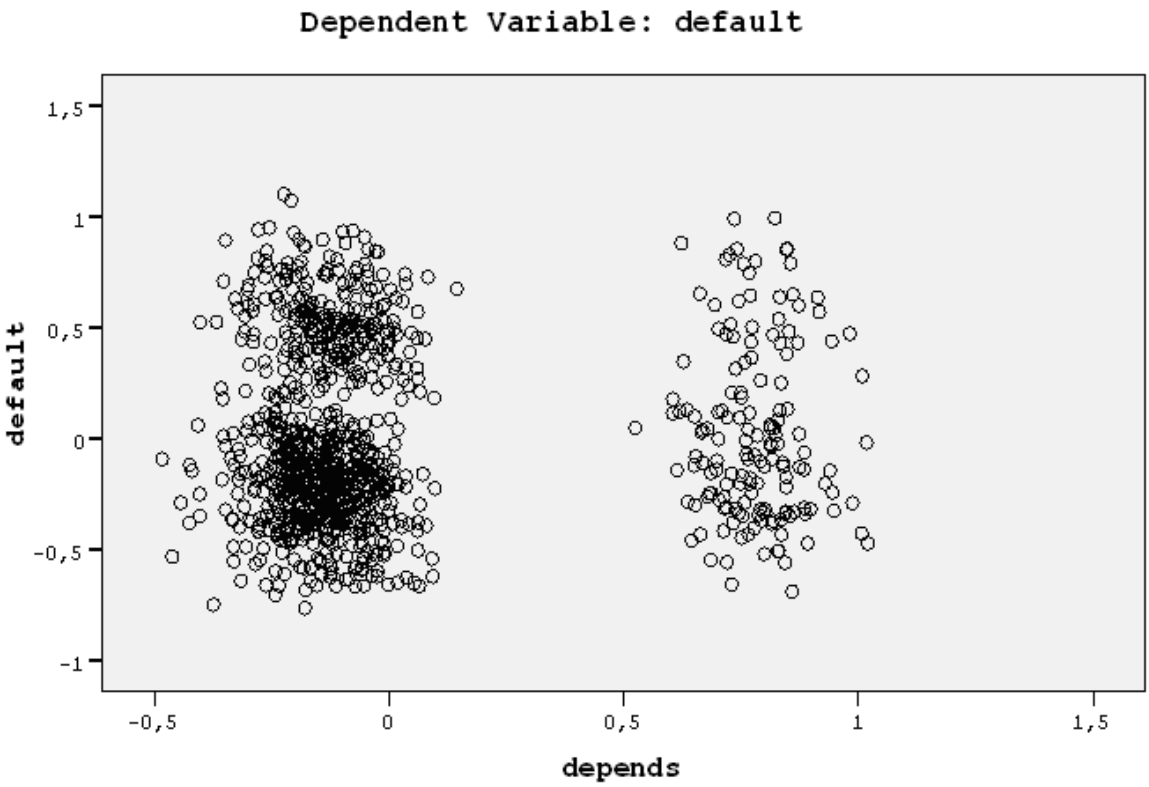

Figure 4.20. PRP Depends

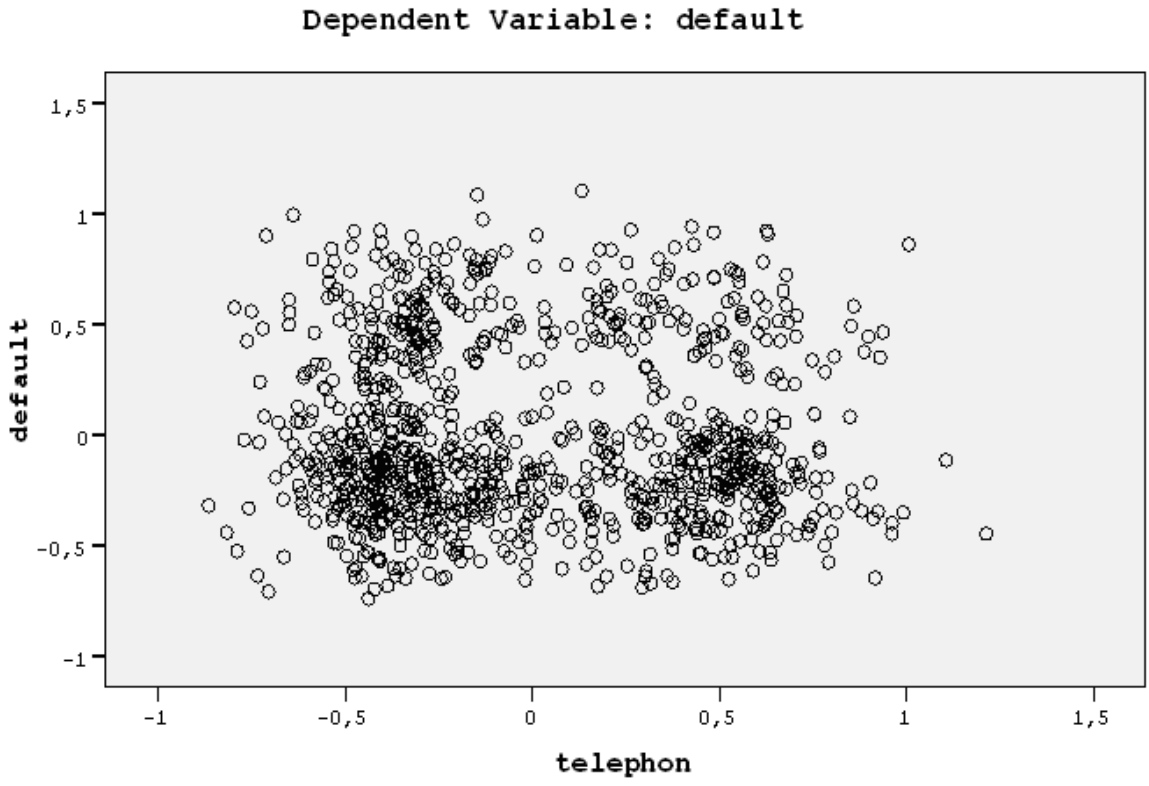

Figure 4.21. PRP Telephon

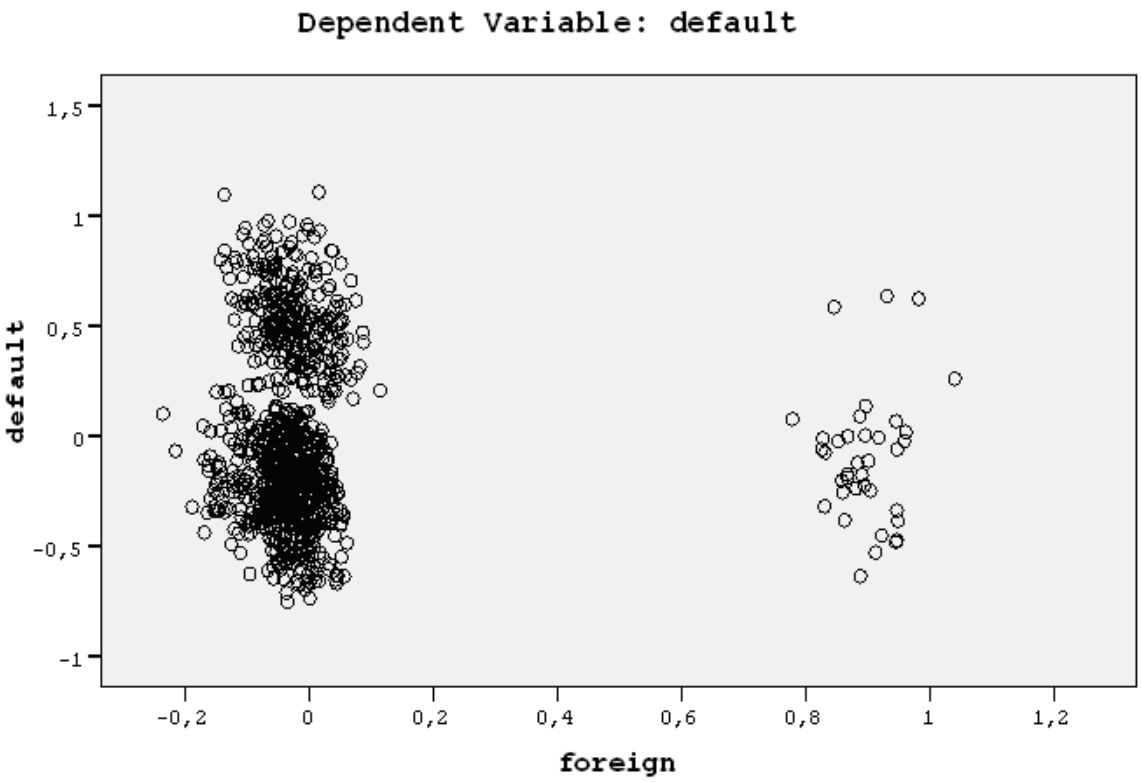

Figure 4.22. PRP Foreign

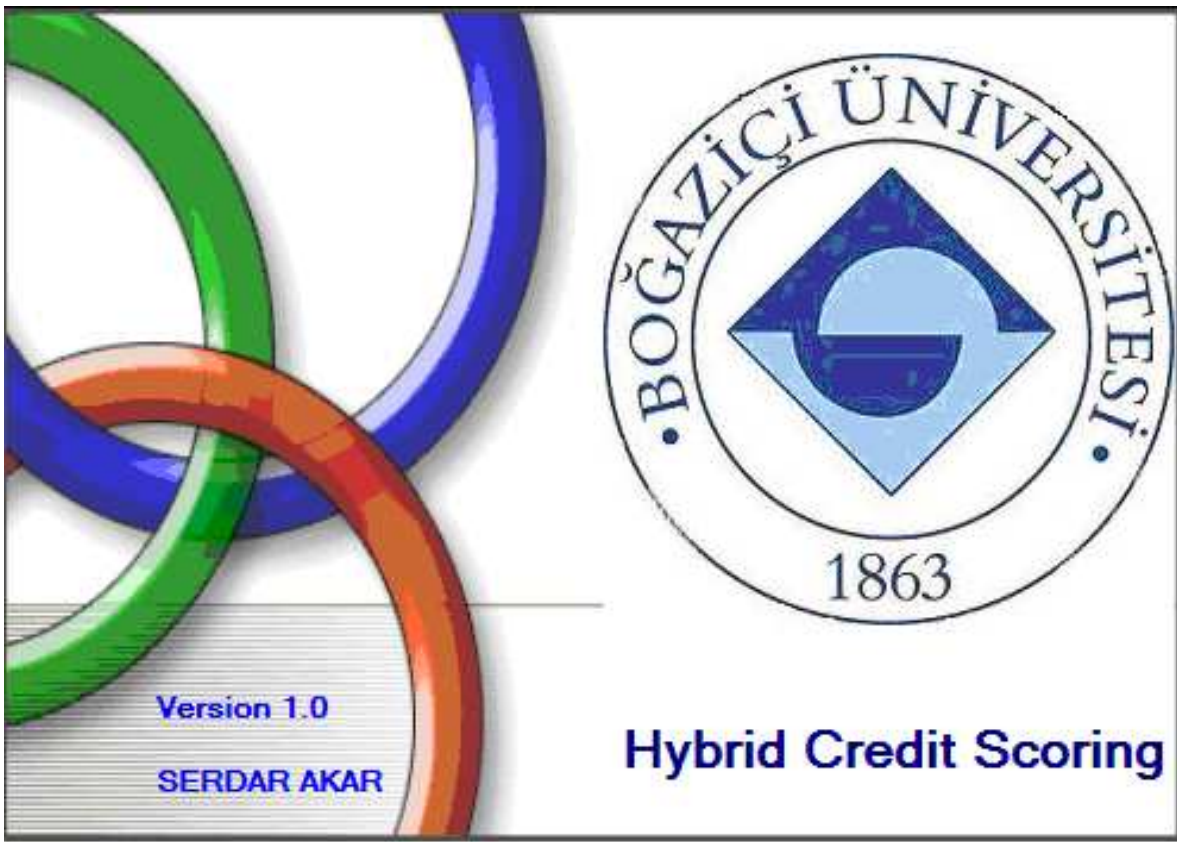

Figure 4.23. Splash Screen

4.5. Integration

When the program firstly open the user encounter the screen which is showed above. In the begining screening hasn't got any filling information or data. At this step if the user try to do anything without filling, there will be a warning.

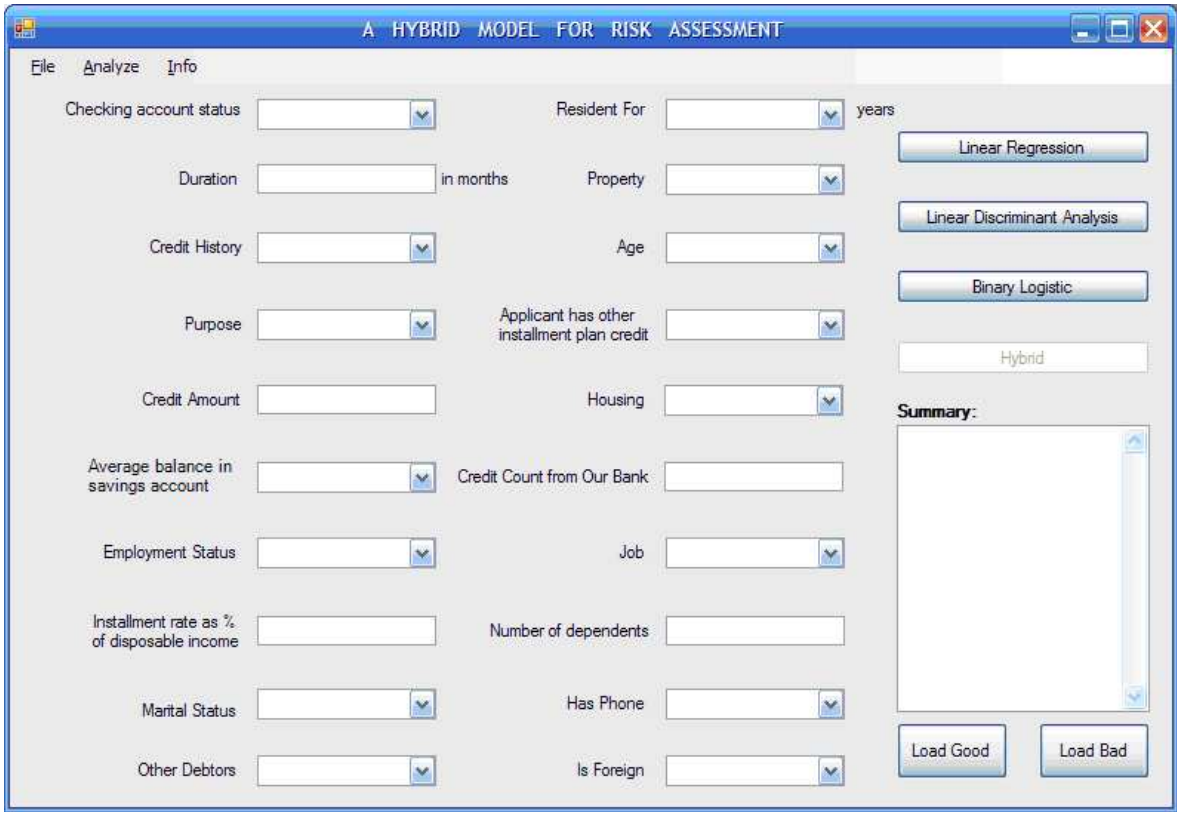

Figure 4.24. Main frame of the program

The aim of this warning enforce the user care about the blanks. Giving the directions like this one provide the caution of the user. Furthermore it must not be forget that there can be any stage kind user on the system.

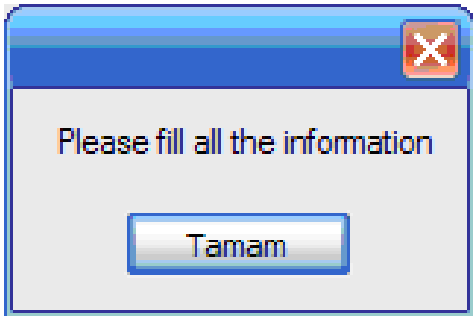

Figure 4.25. Missing Value

On the other hand if the user try to examine any data without full fillment the same caveat will appear again to warn user. So first of all the user have to iniatize the data set and the modelling path to the system immediately. For this operation user must click on the "file" on the menu bar. Later, it would be clicked on initialize unless wishing exit from the program click "exit".

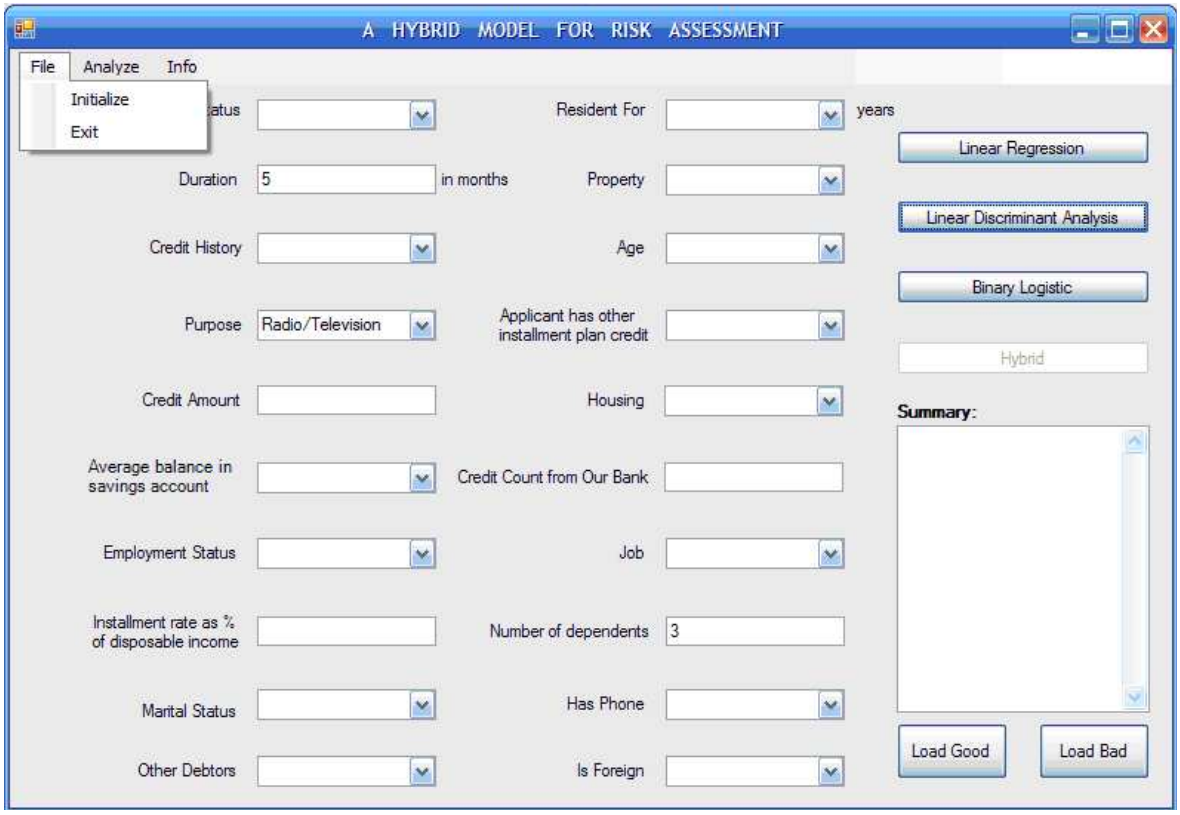

Figure 4.26. Initialization of system

Main frame of the program shows the user of the details of the variable by the names. However all these names refer each variable which has the self properties. Thus the user enter the information one by one to the boxes that space for everyone. Therefore the realization of the system prepare itself ordinary.

Commonly the user habits choose directly the first blanks directly. But at this system it is not necessary. So every stage can be chosen immediately when the program is being used. Integrated .NET Technology support the program more efficiently usage for all level users. Because of the complication and coincidence of the failure situations evaluate before the design provides us the clear and full suppoting system.

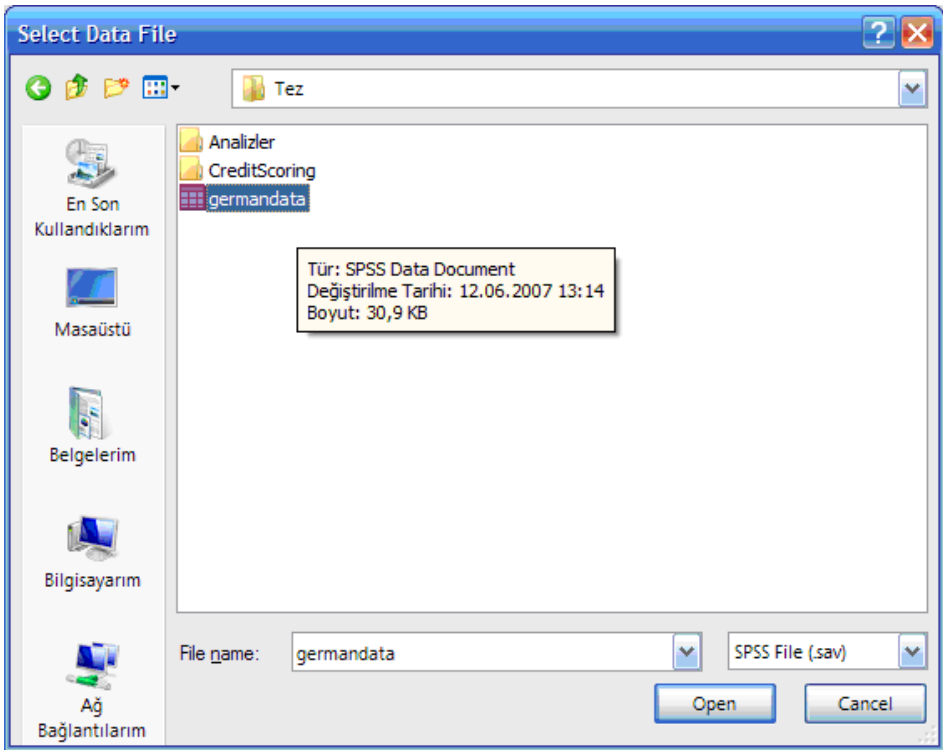

Figure 4.27. Selection of Dataset

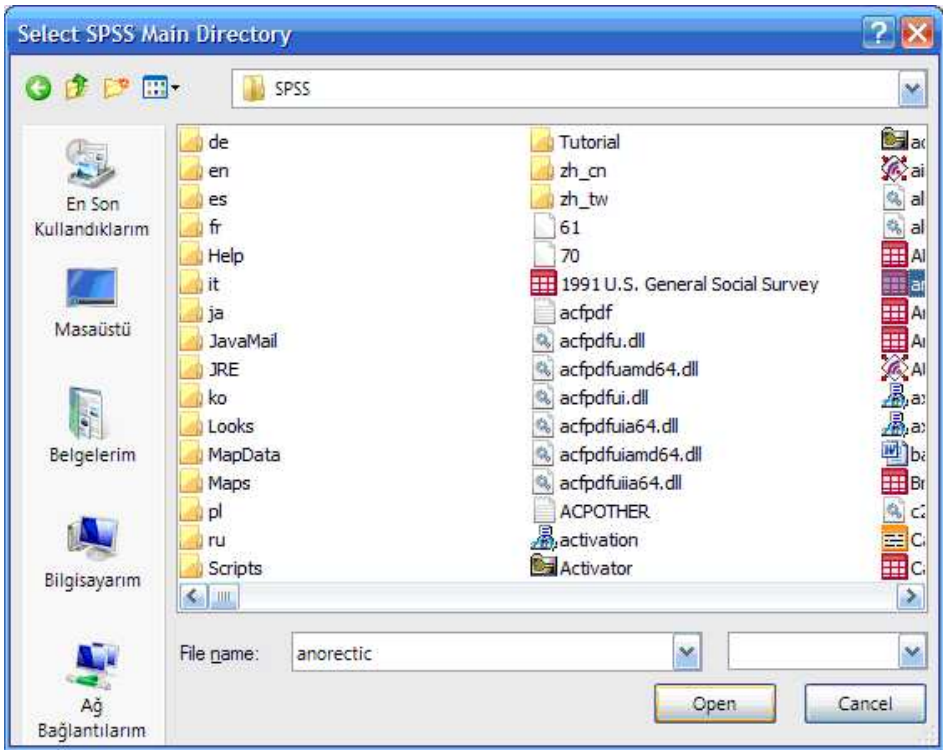

Figure 4.28. Selection of SPSS

Later , if the userwant to chech a good sample to try to find the result of the individual application for credit scoring he or she can click "Good" or "Bad" button.If the Load good buton is clicked then a good record form the German Data will load the program and show its values on the screen to the user. Then, if the analyze buton is clicked for instance "Lineer Regression" exute and the sum show on the screen.

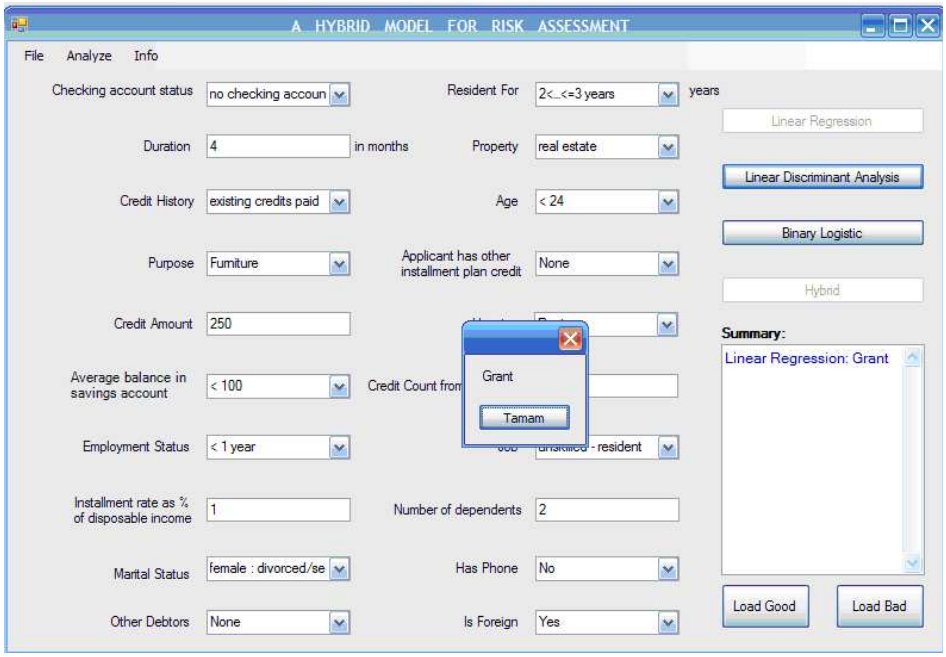

Figure 4.29. Analyzing Result

After the other analysis of Binary Logistic and Linear Discriminant Analysis the result shows as ;

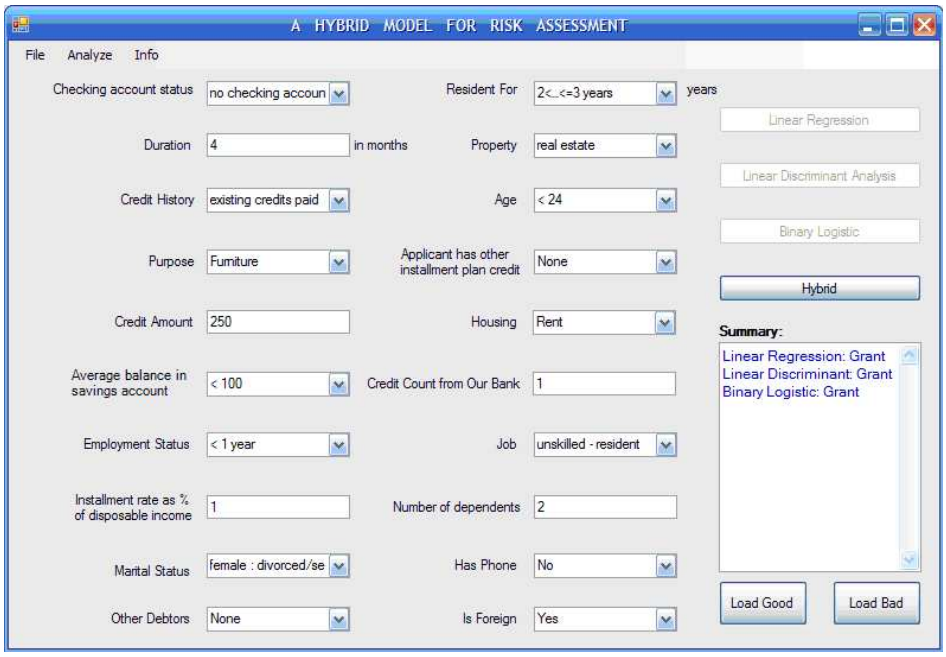

Figure 4.30. Trio Results

When we activate the "Hybrid" button right this time it will give us the final credit scoring "Grant" or "Reject".

If a new credit application examine on the program filling the right information it can be showed as different outputs. But as it can be seen the main idea is the final situation when activating the hyrid modelling.

Additionaly if the user want to see the details of the analysis and their charts it can be possible to see when clicking on the menu bar "Analyze".

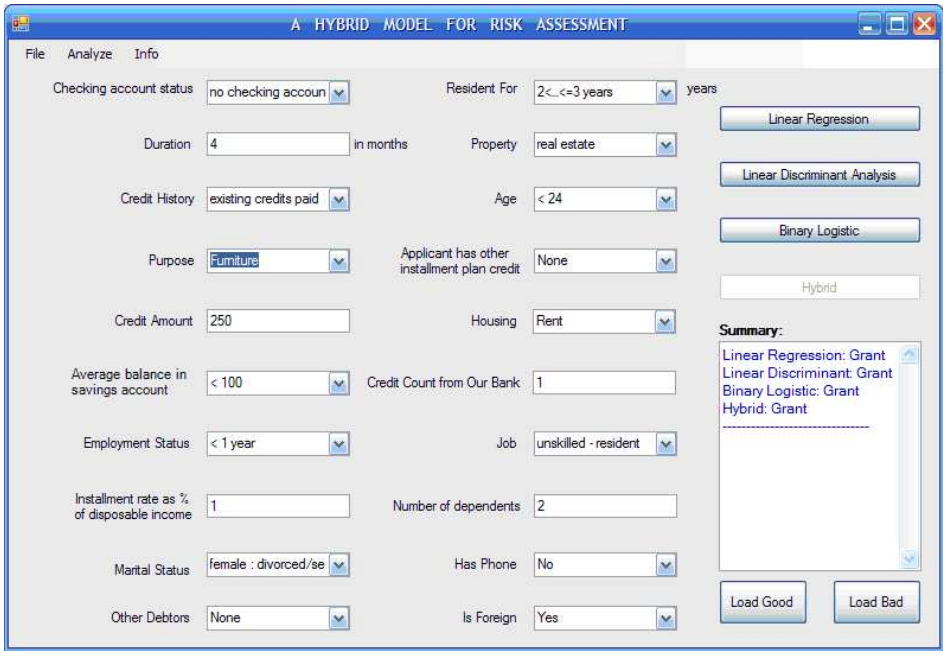

Figure 4.31. Hybrid Analyze

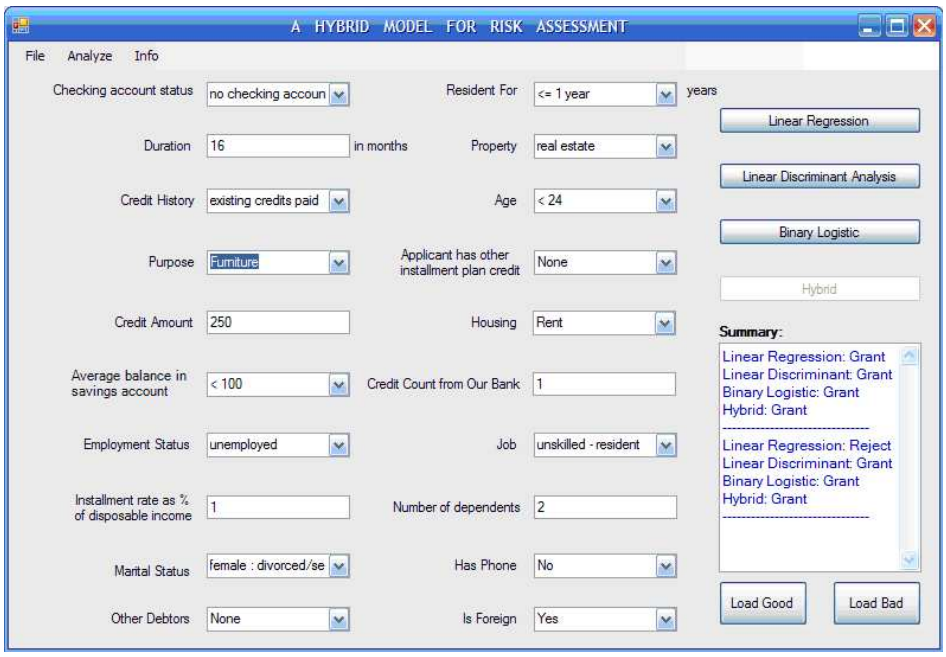

Figure 4.32. Hybrid Grant

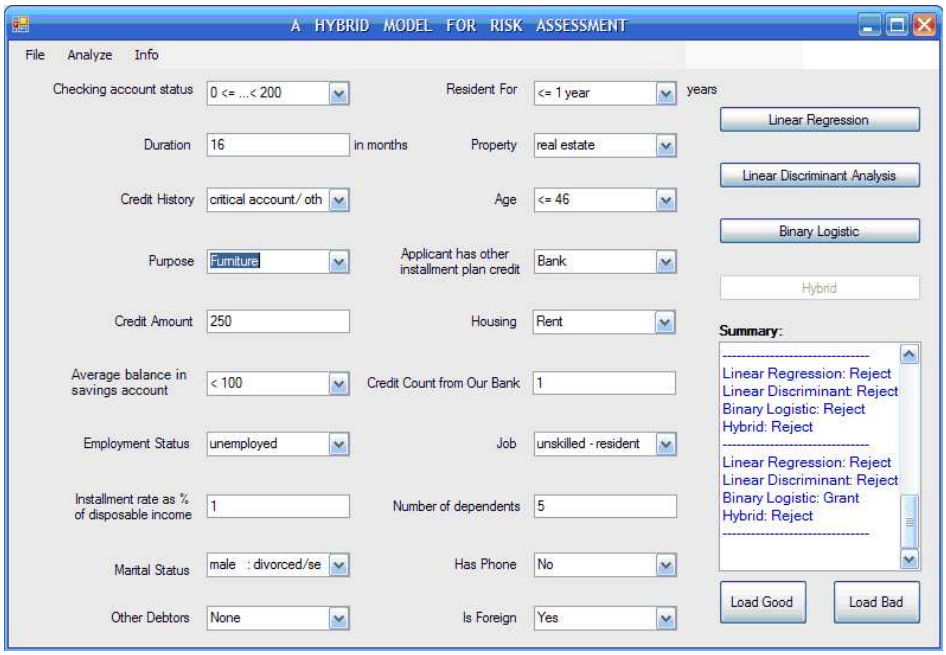

Figure 4.33. Hybrid Reject

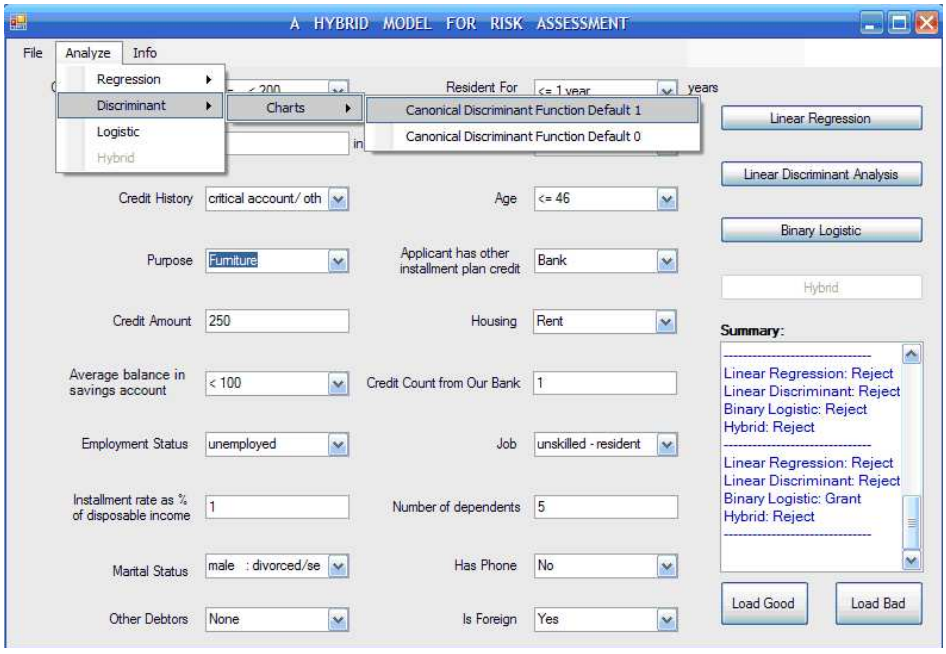

Figure 4.34. Menu Usage

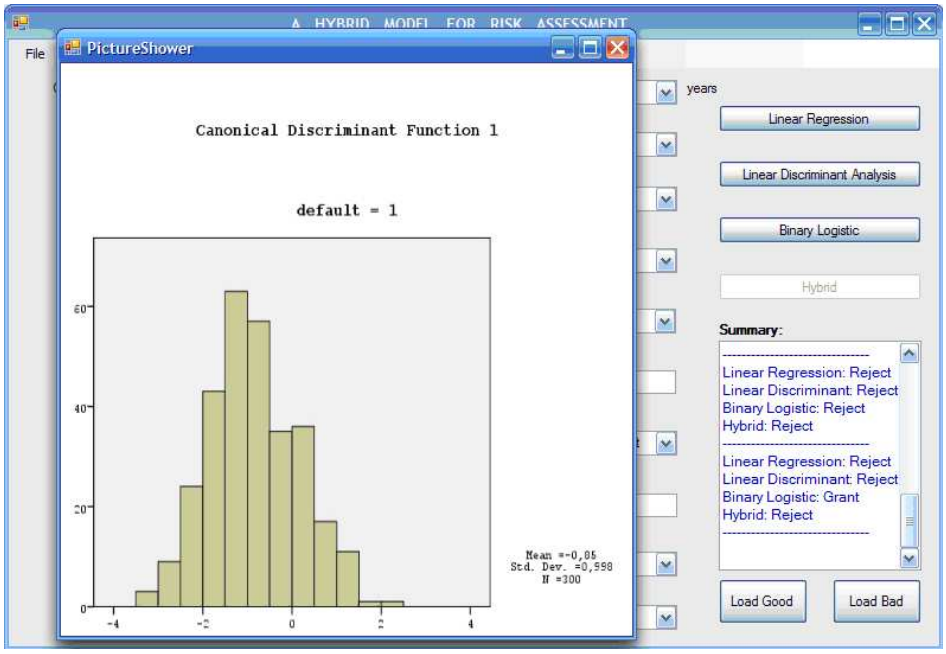

Figure 4.35. Statistical Chart

5. CONCLUSIONS

In this work, we gave an overview about the theoretical aspects of important statistical methods and their applications in credit scoring. In our application part, we firstly focus on Binary Logistic (BL) which is the most widely used method in studies. After all , we reached the Linear Regression (LR) and later is the time of Linear Discriminant Analysis (LDA). By German Data, we examined LR, LDA and BL (Binary Logistic = Logistic Regression) under the aspect of its bias in parameter estimation by using this data set which includes various (%) defaults and different lengths of variables. Our results show that they performs well when the data sets include nearly 30% default cases rejected and 70% default cases granted. Moreover, if the independent variable set would be very large and some variables could have higher effects on dependent variable, the coefficient estimates are much more different from their true values. Secondly, we checked the prediction accuracies of all methods mentioned in this work. In this part, we achieve that the modelling of hybrid design on real German credit data were made. The results of the analysis show that the hybrid model which is combining of logistic regression, linear regression and the discriminant analysis is the best classification technique for German credit data. In future works there will be huge comparing and examining data sets which can be tested and applied over the hybrid and multiple analyzing systems.

REFERENCES

Altman, E.I., and Loris, B., September 1976, "A financial early warning system for over the counter broker-dealers", *Journal of Finance*, Vol. 31,4 pp. 1201-1217.

Atiya, A.F., July 2001, "Bankruptcy prediction for credit risk using neural networks: A Survey & New Results", *IEEE Transactions on Neural Networks*, Vol. 12, pp. 29-35.

Back B., Laitinen T., Sere K. and van Wezel M., September 1996, "Choosing bankruptcy predictors using discriminant analysis, logit analysis, and genetic algorithms", *Turku Centre for Computer Science Technical Report* pp. 40.

Coats P. and Fant L., 1993, "Recognizing financial distress patterns using a neural network toll", *Financial Management* Vol. 22 , pp. 142-155.

Crook J.N., Edelman D.B., and Thomas L.C., 2002, "Credit Scoring and Its Applications", *SIAM*, pp.13-23 .

Frydman H., Altman E.I., and Kao D.L., March 1985, "Introducing recursive partioning for financial classification: The case of financial distress", *The Journal of Finance* Section XI-1 .

Johnson R.A., and Wichern D.W.,1997, "Applied Multivariate Statistical Analysis", *Prentice Hall*, New Jersey.

Libby, R.,Spring 1975, "Accounting ratios and the prediction of failure: some behavioral evidence", *Journal of Accounting Research*, Section 13:1, pp. 150-161.

Misc.,Anonymous, 2000-2007, "Documents Over Issues", *METU*.

Smith,S., 1970, "Journal of Money", *Credit and Bank 2* , pp. 435-445.

Sobehart, J.R., Keenan, S.C., and Stein, M.S., 2000, "Bencmarking quantitative default risk models: a validation methodology", *Moody's Rating Methodology.*

Tam Y. K. and Kiang Y.M., July 1992, "Managerial applications of neural networks: the case of bank failure predictions", *Management Science*, Vol. 38 , pp. 926-947.

Wilcox, J. W., 1971, "A Simple theory of financial ratios as predictors of failure", *Journal of Accounting Research*, Vol. 9.2 , pp. 389-395.

Wilcox, J. W., 1973, "A Prediction of business failure using accounting data", *Journal of Accounting Research, Emprical Research in Accounting: Selected Studies 11* , pp. 163-179.

Wong B.K., Bodnovich T.A., and Selvi Y., 1997, "Neural network applications in business: a review and analysis of the literature" , *Decision Support Systems* Vol. 19 , pp. 301-320.

Yohannes Y., and Webb P., 1999, "Classification and Regression Trees,CART, A User Manual for Identifying Indicators of Vulnerability to and Chronic Food Insecurity", *International Food Policy Research Institute* .