İSTANBUL BİLGİ UNIVERSITY
INSTITUTE OF GRADUATE PROGRAMS
BANKING AND FINANCE MASTER'S DEGREE PROGRAM

PREDICTING CREDIT DEFAULT RISK USING MACHINE LEARNING ALGORITHM

MEHMET TELİMEN
110673019

Assist. Prof. BARIŞ SOYBİLGEN

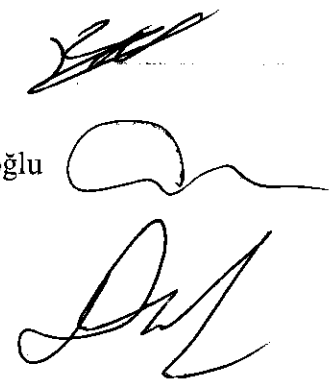
İSTANBUL
2019

PREDICTING CREDIT DEFAULT RISK USING MACHINE LEARNING ALGORITHM

MAKİNE ÖĞRENMESİ ALGORİTMASINI KULLANARAK KREDİ TEMERRÜT RİSKİNİ TAHMİN ETME

Mehmet Telimen
110673019

Tez Danışmanı:	Dr. Öğr. Üyesi Barış Soybilgen İstanbul Bilgi Üniversitesi
Jüri Üyesi:	Dr. Öğr. Üyesi Burak Alparslan Eroğlu İstanbul Bilgi Üniversitesi
Jüri Üyesi:	Dr. Öğr. Üyesi Dinçer Dedeoğlu Bahçeşehir Üniversitesi



Tezin Onaylandığı Tarih: 10.06.2019
Toplam Sayfa Sayısı: 70

Anahtar Kelimeler
1) Tüketici Kredisi
2) Temerrüt Riski
3) Sınıflandırma
4) Makine Öğrenmesi
5) Tahmin

Key Words
1) Consumer Credit
2) Default Risk
3) Classification
4) Machine Learning
5) Prediction

ACKNOWLEDGMENT

Firstly, I would like to thank my supervisor Barış SOYBİLGEN who guided and supported me at every stage of this study.

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

Finally, I would like to thank my dear friends who encouraged me in my hard times.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

NPL: Non-Performing Loans

LDA: Linear Discriminant Analysis

QDA: Quadratic Discriminant Analysis

KNN: K-Nearest Neighbors

SQL: Structured Query Language

KKB: Credit Bureau (Turkey)

NB: Naive Bayesian

DT: Decision Tree

SVM: Support Vector Machines

RF: Random Forest

BNN: Bagged K-Nearest Naihgbors

IMF: International Monetary Fund

SDIF: Savings Deposit Insurance Fund

QR: Quick Response

ATM: Automated Teller Machine

LIST OF CHARTS

LIST OF TABLES

ABSTRACT

In this study, it was aimed to construct the analytical models that predict the probability of default of consumer credit by using machine learning algorithms. The data belonging to the customers of a bank has been used by making anonymity from the bank's test environment. This data set was composed of the lending status of the customers in the bank and the questioned credit bureau data at the credit application stage. Half of the samples in the data set were selected from those who had been in default and half were not.

In the study, four of the widely used techniques of classification based on machine learning have been discussed. Those are Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and K-Nearest Neighbors (KNN). For each model, half of the data set was used for training and the other half was used for testing.

Those models, which were trained with the same training set using the corresponding functions in R studio with R programming language, were tested with the same data set and the accuracy rates of them were compared. As a result of the comparison, with given this data, it is observed that the model of the Logistic Regression estimated the probability of default of the consumer loan with the highest accuracy rate which was 58.30%.

Key Words: Consumer Credits, Default Risk, Classification, Machine Learning, Estimation

ÖZET

Bu çalışmada, makine öğrenmesi algoritmaları kullanılarak, tüketici kredisinin temerrüte düşme ihtimalini tahminleyen analitik modellerin oluşturulması amaçlanmıştır. Çalışmada, bir bankanın müşterilerine ait veriler, bankanın test ortamından anonim hale getirilerek kullanılmıştır. Söz konusu veri seti müşterinin bankadaki kredisinin gecikme durumu ve kredi başvuru aşamasındaki sorgulanmış Kredi Kayıt Bürosu (KKB) verilerinden oluşturulmuştur. Veri kümesindeki örneklemlerin yarısı gecikmişe düşmüş, yarısı da gecikmişe düşmemiş kredi kayıtlarından oluşacak şekilde hazırlanmıştır.

Çalışmada, gözetimli makine öğrenmesine dayalı sınıflandırma tekniklerinden yaygın olarak kullanılan dört tanesi ele alınmıştır. Bunlar, Logistik Regression, Linear Discriminant Analizi, Quadratik Discriminant Analizi ve K-En Yakın Komşuluk Metodudur. Her bir model için veri setinin yarısı eğitim için kullanılırken diğer yarısı da modelin test edilmesi için kullanılmıştır.

R studioda R programlama dilindeki ilgili fonkisyonlar kullanılarak aynı eğitim seti ile eğitilen bu modeller yine aynı test seti ile test edilip tahmin oranları kıyaslanmıştır. Kıyaslama sonucu, bahsedilen veri kümesi üzerinden hesaplandığında, tüketici kredisinin temerrüte düşme ihtimalini en yüksek doğruluk oranı (%58.30) ile lojistik regresyona ait modelin tahmin ettiği gözlemlenmiştir.

Anahtar kelimeler: Tüketici Kredileri, Temerrüt Riski, Sınıflandırma, Makine Öğrenmesi, Tahmin

INTRODUCTION

Nowadays, firms those have big data, especially financial institutions like banks, are changing their business models. In this sense, they are investing heavily in artificial intelligence and machine learning by using available data. The business models supported by artificial intelligence or machine learning algorithms are safer and more sustainable way to survive. We can say that predicting the future from today and taking a decision accordingly, for example, foresee the credit risk for the bank is a key to being able to withstand its competitors and to be one step ahead. At this point of view, it is vital to correctly analyze data, correctly classify data and use the historical data in the right place.

Machine learning is a rapidly spreading approach, where analytical models are created in order to enable machines to learn a subject by using real data and to make the predictions about the future. In this age that we are faced with very large data sets, it is very important to make machines learn something and come to the point where they can predict the future. Machine learning has a wide range of applications. For example, banks or financial institutions need to know what kind of risk factors they need to analyze when lending to their customers. It is clear that the situational or behavioral characteristics of the customers should be taken into account. Most of the time, however, many of these characteristics may have little or no impact on the future collection process of the loan. At this point, it is a very critical issue that banks or financial institutions face when giving credit to determine which features of the customers are effective on the life of the loan and to calculate a score value accordingly (Turkson, Baagyere, and Wenya (2016), p.1).

People receive loans from banks to meet their various needs, and they may have difficulty in paying the debts of those loans. This situation is expressed as the follow-up of the loan, and the ratio of such credits to the total loans is expressed as the NPL ratio which states non-performing loans. The NPL ratio is one of the most important indicators of the asset quality of a bank as well as the risk level of the

bank. NPL rate is an important parameter for banks and economy management. The decrease in the probability of return of all or some of the loans granted by a bank causes a decrease in the asset quality of the bank and adversely affects the profitability of the bank. If the asset quality of a bank falls, it is expected that the bank's credit volume will be negatively affected. From this point of view, it can be said that the increase in the NPL ratio adversely affects both the profitability of the bank and the credit volume, and negatively affects both economic growth and employment (Finansal Göz, 2018, para. 1). It is therefore essential to be able to predict the probability of credit default, which has the potential to affect many areas indirectly. In this study, it is aimed to calculate the probability of follow-up when the related loan is used while a person is still in the loan application process by using machine learning algorithms.

Although it has not reached a very advanced stage, various studies have been carried out in the literature on the models that predict customer credit risk or related issues using machine learning techniques. For example, Khandani, Adlar, and Lo (2010) have created non-linear and non-parametric models to estimate credit default, using customer and credit bureau and transaction data for customers of a large commercial bank via machine learning techniques. By using the linear regression models they could obtained the estimations that highly improve the classification of credit-card defaults.

Kruppa, Schwarz, Arminger, and Ziegler (2013) presented a broad framework for estimating credit risks using machine learning methods. They applied the machine learning algorithms and the optimized logistic regression to a large data set containing the full payment history of short-term installment loans. In the study, an algorithm is also defined for setting the terminal node size regarding the probability forecasting. They have shown that RF (random forest) regression works better than KNN (k-nearest neighbors) and BNN (bagged k-nearest neighbors) with the data set they used in the case.

Turkson, et al. (2016) examined the credit data of a real bank and selected important features to run different machine learning algorithms on the data for comparative analysis and chose which algorithm was the best way to learn bank credit data. The algorithms constructed in the study provided over 80% accuracy in estimations. In addition, they provided a predictive model using significant variables to estimate the credit worthiness of a customer.

Son, Byun, and Lee (2016) conducted a comprehensive study to verify the estimation performance of different maturities and different rating groups on the credit default swap margins of different matrices and different estimation performance of the traditional parametric model. As a result of their study they pointed that artificial neural networks get better score than parametric and nonparametric models.

Based on historical data, Islam, Eberle, and Ghafoor (2018) presented an intuitive approach where the probability of a risk is modeled, and the probability of risk for the subsequent transactions is calculated as soon as the event occurs. In addition to their heuristic method, they also implied a new proposed machine learning approach that was not previously implemented to their targeted data set. Finally, they found that those applied approaches performs better than existing ones.

Kvamme, Sellereite, Aas, and Sjursen (2018) conducted a study to predict mortgage risk by applying convolutional neural networks. Just by using the balances of the accounts of checking and saving and some daily transaction on checking accounts they achieved good point in the manner of classification.

Emil and Sivasankar (2018) worked on the estimation accuracy of data mining techniques to compare the features of the credit card clients with the

selection and removal techniques and to suggest the most effective technique in credit card analysis by applying classification algorithms such as K-nearest neighbor (KNN), Naive Bayesian (NB), Decision Tree (DT) and Support Vector Machines (SVM) to the customer's default data in Taiwan.

In order to estimate the probability of credit default, Addo, Guegan, and Hassani (2018) set up binary classifiers based on machine and deep learning models according to actual data. The most effective variables of those models were selected and used in the modeling process by comparing them in separate data to test the stability of binary data classifiers. They found that models based on multi-layer artificial neural networks are not as stable as tree-based models.

Vanneschi, Horn, Castelli, and Popovic (2018) have developed a credit risk model to replace the pre-risk control of the e-commerce risk management system to estimate the probability of default of customers. The application of genetic programming to credit risk is also presented in the paper. The results showed that genetic programming performs better than the general credit risk model, both in terms of classification accuracy and in pre-risk control.

Wang, Han, Liu, and Lou (2018) have studied on the evaluation of the loans used in the peer-to-peer lending industry, in China. Based on research on natural language processing, they used debtors' online transaction behavior data, and constructed a method of scoring consumer credit based on the one of deep learning algorithms.

Currently, the academic studies conducted on the subject in Turkey is limited. However, it is a well-known fact that all banks around the world have internally established systems to estimate the lag of the loan by internal rating or credit scoring methods in advance (Jia, 2018, para. 2).

In this study, it was aimed to estimate the probability of default of a consumer loan with different machine learning algorithms based on classification approach by using only credit bureau data and compared the accurate rates in estimating credit risk of the models obtained. In the study, a Turkish bank's obscured customer data, and the follow-up status of the loans they had used have been used. It is hoped that this thesis will contribute to the further studies on this area in the literature where we still have not reached a very advanced point.

In the first part of this study, the historical development of Turkish banking sector is examined. In the second part, the data used in the study is explained in detail. In the third chapter, the machine learning algorithms used in the study are explained. In the last part of the study, the models based on the machine learning algorithms are formed and the success rate of the predictions of these models are discussed.

CHAPTER 1

HISTORICAL DEVELOPMENT OF TURKISH BANKING

Banks are profitable economic entities that collect funds as a deposit in the market and provide the funds they collect as a loan to the institutions or persons who need. We can say that the main activity of banks is to take credit or to give credit. Today, in addition to deposits and credit transactions, banks play an important role in the implementation of credit and monetary policy of the government, conduct foreign exchange transactions, intermediate in the purchase and sale of securities, save valuable paper or goods in the cash register, mediate various financial transactions, and facilitate payment transactions through credit and debit card. The position of banks in the national economy is very important in terms of fulfilling these great functions. The power of a country's economic structure is closely related to banks. Banks are a fundamental tool for a healthy economic development (Yetiz, 2016).

When the money was not used as a tool of change in history, it was made in the form of bank transactions, securities and property loans. In history, the importance of banking for the first time required to organize found in the period of Babylonian Empire. It has been determined that some clay plates belonging to this period were issued bonds with interest and borrowing transactions. In this period, land mortgage and bail transactions were also found. The first commercial documents are known to occur in Mesopotamia. In addition to the land mortgage and bail credit transactions, the valuable assets that they gifted to the gods were the basis of the credit system. The fact that these precious beings were borrowed by the clergy to generate income and that the sacred areas were the most valuable asset stores became a starting point in the banking field. The emergence of private banks began with the Romans. When the money was used as a tool of exchange, the borrowers took out deposits from their customers and lent money to others in return for high interest. During the Crusades in the Middle Ages, money transfer

operations were started to meet the needs of the army. This situation enabled institutions to mediate money transactions. These events accelerated the development of the banking system and led to the emergence of commercial law. With the establishment of the Bank of Venice in 1157, today's banking began (Bozdemir, 2007).

While the above mentioned developments were experienced in Europe, money in Anatolia occupied an important place in the lives of Turks in the same period. The fact that Anatolia was located on the east and west trade route caused the development of trade and the need to print money due to facilitating trade. The use of gold and silver coins at the time of the Seljuks showed that there was a real money economy in this period (Bozdemir, 2007).

Since the developments in the field of industry in Western Europe and the outward opening were not in the Ottoman Empire, the banking sector was not developed and encouraged. Therefore, we don't see any bank in the Ottoman Empire until the mid-19th century. Banking transactions were generally in the hands of the bankers and sarrafs. Although the first bank was established in 1847 by the Galata Bankers in the Ottoman Empire, it could not work for a very long time and ended its activities in 1852. For this reason, the establishment of the Ottoman Bank in 1856 is accepted as the beginning of banking in the Ottoman Empire. After the 1839 Tanzimat Edict, a period has started which the state's expenditures exceed their incomes and state's needs were met by borrowing from the sarrafs and bankers in İstanbul. After the Crimean War, the Peace Treaty of Paris in 1856 was made and thus the opportunities of foreign borrowing of the Ottoman Empire increased. This is the most important factor that triggered the establishment of the Ottoman Bank. The Ottoman Bank was established with British capital and mediated between the Ottoman government and foreign capital owners in external borrowing. In 1863, Homeland Chests were established to provide farmers with agricultural loans in more favorable conditions. Homeland

Chests were first funded by the method of imece, then by giving wheat to be proportional to the farmer's assets. Ziraat Bank was established as the first state bank in 1888 because of various problems related to the functioning of the Homeland Chests. Ziraat Bank aimed to open agricultural credit to the state control. The capital of the bank is formed by the receivables of the Homeland Chests (Yetiz, 2016).

When we look at the early years of the republic period in 1923, in Turkey, there were 35 banks in total. 22 of them national and 13 of them were foreigners. In Economic Congress held in Turkey in 1923, the government and society of agriculture, the leading sectors of commerce and industry agreed upon the development of the national bank for economic. Since the private sector did not have the facilities to establish a bank yet, it was stated that the state should contribute to the establishment of banks, in the congress. In 1924, Turkey Business Bank began operations as the first private bank. In 1925, Turkey Industry and Mines Bank was established as the first development bank. Turkey Industry and Mines Bank, whose aim was to give credit to private industrial enterprises, was transferred to Sumer Bank in 1933 because it could not operate in accordance with its purpose. Ziraat Bank, which was established during the Ottoman period, was rearranged and its capital was increased in order to provide a more appropriate loan in favor of the agricultural sector with the proposal of the agricultural sector participating in the congress. Founded in 1927 to support the housing sector, Emlak and Eytam Bank were converted into Emlak ve Credit Bank in 1946. As a result of the work started in the 1920s, the Central Bank of the Republic of Turkey was founded in 1930 (Yetiz, 2016).

Between 1933 and 1945, which is accepted as public banking period in Turkey, Sumerbank was established in 1933 to support industrial development. Iller Bank was established in 1935 in order to support the development of local administrations and the provision of infrastructure services with medium and long-

term loans. Denizbank was established in 1937 to operate regular postal flights between Turkish and foreign ports and to conduct various port operations. Halkbank was founded in 1938 to give credit to small tradesmen and craftsmen.

Between 1945 and 1959, Turkish private banking was developed. In this period, the rapid increase in the population and national income, the increase of investments and enterprises, the fact that the industry started to receive more shares from the national income and the increase in production caused the need for financing. All of this increased the return on investment in banking and private banking had rapidly gained value. In 1946, Garanti Bank, Akbank in 1948, in 1955 Pamukbank and Turkey Industrial Development Bank was established in 1950. Those years, the commission rates and the interest of banking transactions were determined by the government and only the Central Bank was authorized to perform foreign exchange transactions. This provided a valuable understanding of competition based on branch banking and deposit collection. The spread of branch banking has paved the way for the liquidation of local banks (Yetiz, 2016).

The periods in which five-year development plans are made regularly are called planned periods. Although the five-year development plans have come to the present day with pauses, especially the 1960s and the second half of the 1970s are called the planned period in Turkey banking history. In the planned period, a total of 7 new banks were established, 5 of which were development bank, 2 of which were commercial bank. In 1962, the Republic of Turkey Tourism Bank, in 1963 Industrial Investment and Credit Bank, in 1964 State Investment Bank, in 1968 Turkey Mining Bank, in 1976 State Industrial and Labor Investment Bank, in 1964 American-Turkish Foreign Trade Bank and in 1977 Arab-Turkish Bank was established. In spite of the Stability Program in 1958, the inability to achieve economic balances and the unstoppable recession in the economy led to the abandonment of the liberal economic policy implemented in the 1950s and the introduction of a mixed economy in which the state intervened in the economic

field. In the period of 1960-1980, the investments in the development plans, which were started to be implemented in 1963, were realized and an industrialization policy aimed at ensuring the production of imported industrial goods within the country was followed. In this way, the government aimed to take the control of the economy and to improve the Turkish economy that had been under the influence of foreign countries (Bozdemir, 2007).

The Turkish banking sector, which lived under state control until 1980, progressed rapidly towards liberalization since 1980. In accordance with the new development plan adopted by the state, real interest rate and flexible exchange rate was introduced. These developments led to the regulations of development of financial markets. In this period, investment in the banking sector became easier. Bank of Creditand Commerce (1980), Bank of Melland (1981), Turk Bank Ltd (1981), Habib Bank (1982), The First National Bank of Boston (1984), Manufacturers Hanover Trust Company (1984), Saudi American Bank (1984), the Bank of Bahrain and Kuwait BSC (1985), Standard Chartered Bank (1985), Cyprus Credit Bank Ltd. (1988) and Societe Generale SA (1989) 's are allowed to operate in Turkey. In 1982, the necessary legal and institutional structure for the use of Capital Market Law and capital market instruments was established. In 1986, the Istanbul Stock Exchange became operational. Arrangements for the functioning of the free market mechanism and liberalization of financial markets in the economy had important effects on the banking system. Competition in the sector has increased due to allowing the entry of new domestic / foreign banks into the sector and the release of deposit / credit interest rates. Increasing competition led to the adoption of a banking sector where banks had increased both resource and displacement diversity rather than conventional deposit banking. The banks' customers were offered new products and services such as consumer loans, credit cards, foreign exchange deposit accounts, leasing, factoring, forfeiting, swap, forward, and future. Also, productivity was increased in the sector as a result of the use of computer systems and other technological innovations and emphasis on

personnel training. In 1986, Money Market was established to provide more efficient use of resources in banking. The monetary program implemented in 1990 paved the way for the Treasury to finance the budget deficit by transferring funds from private banks. In 1999, the Banking Regulation and Supervision Agency was established in order to start the operation of banks, to monitor and audit their activities and to decide the results of the audit. In the same year, Savings Deposit Insurance Fund (SDIF) was established. In the first half of 2000, as a result of the Stand-by agreement with the IMF in 1999, positive results were achieved for achieving price stability and a sustainable public borrowing system. Domestic borrowing interest rates and inflation declined. However, in the second half of the year, the reasons such as the delay in structural adjustment arrangements, the decline in the inflation rate, the increase in public goods and services as well as the increase in inflation caused a negative situation and the banking sector experienced a concussion in November 2000. During this period, a large number of banks, which suffered losses due to large fluctuations in exchange rates and interest rates, remained out of the market. Some commercial banks with private capital have been transferred to SDIF. In April 2001, transition to a strong economy program was implemented in order to eliminate structural problems in the economy, to strengthen the financial structure of the financial system, to reduce inflation, to decrease public debt and to strengthen the banking system. The program aimed at increasing the resilience of the economy to external shocks, reducing debts, ensuring fiscal discipline, reducing inflation and strengthening the banking system. In 2002, inflation displayed a downward trend, short-term interest rates were cut, and foreign exchange market interventions were made to prevent fluctuations. As a result of restructuring, banks became profitable. The merger, sale and liquidation of the banks transferred to the SDIF were realized. The number of banks operating in 2003 decreased to 50 as compared to 2002. In 2004, the banking system performed well in parallel with the improved economy and the confidence in the banks increased. With the transfer of the banks, whose financial situation deteriorated, to the SDIF,

the financial system has started to work healthily and a competitive environment has arisen (Yetiz, 2016).

Today the banking sector in Turkey has an important place in the financial structure. In Turkish banking system, banks can be classified as deposit banks, participation banks, development and investment banks according to their operational areas. As of March 2019, 53 banks are active in the banking system. There are 34 deposit banks, 13 development and investment banks, and 6 participation banks.

Thanks to computers, mobile phones and other digital devices, the banking sector is now also taking its share of this rapid change in technology. Digital transformation in finance and banking spearheads rapid digitization of individuals and society. For individuals who keep up with the digitizing world, it's tempting to take care of all banking transactions with a click. In our world where all life processes are carried on the web, cost-cutting developments in the banking sector attract everyone's attention and innovations in digital banking affect customer preferences. Industry 4.0 revolution, artificial intelligence and the internet of things stand out, banks are following these developments.

Banks are working hard to meet the needs of their digital customers. It is observed that almost all banks have established units under the name of digital transformation and try to keep up with the new trend. The fact that large technology companies in the world started to provide services given by the banking sector has given rise to a new kind of competitor for banks. This situation causes banks to invest more in technology. In this sense, in Turkey like in the world, the classic branch banking is rapidly decreasing and online banking is gradually increasing. In the digitization process, banks that want to attract more customers through mobile applications are competing in this sense. With the use of digital wallets, all banks bring new alternatives to payment channels and encourage their customers to use

digital applications. Banks encourage their customers to use mobile applications via mobile devices to do their banking transactions. For example, customers are now able to log in faster, more easily and securely via smartphones with eye-scanning technology. Also, with the QR code, customers can use features such as fast withdrawal and direct connection to the call center without going to the ATM.

While banks offer smart solutions, needs-based analyzes also come to the fore. Responding to customer demands becomes easier with analyzes. The customer representation system is gaining a new dimension. Thanks to artificial intelligence, chat boats and voice recognition technologies are remodeling to respond much faster to customers. Using artificial intelligence and machine learning, banks can better serve their customers and minimize their costs and risks.

CHAPTER 2

DATA

2.1. DATA SOURCE

The data used in this study includes the current status of the consumer loans of a bank in Turkey and the queried information from the credit bureau at the application stage. The data set was withdrawn from the bank's test environment. All data specific to the customer such as identity number and address was deleted when the data set was being collected. The data were prepared so that they could not be associated with anyone. The necessary permissions were taken from the relevant units of the bank to use the data in this study.

The data set is taken from the test environment of the bank by querying the anonymized data and the credit bureau data combined with SQL. It was transferred to the data file with CSV extension, which could be used directly in the study.

2.2. TRANSACTION DATA

22.554 loan records belonging to the real customers of the bank were taken. The current follow-up status of each credit record has been received. This data consisted of only two columns: A randomly given credit record number and the information on whether it is overdue or not (0 and 1). Half of the credit records obtained from the system were in default.

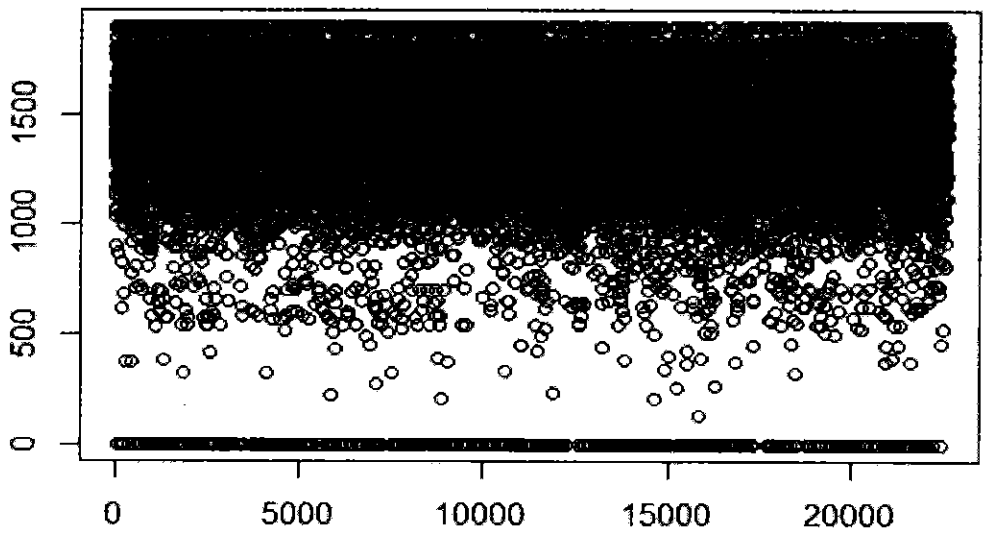
2.3. CREDIT BUREAU DATA

The credit bureau data consisted of 29 significant variables, which were related to the real credit records of the bank, from the questioned credit bureau data of the customer during the application stage. Those variables are described in detail below.

2.3.1. Credit Bureau Score Information

One of the variables used in the study is related to the score section returned from the credit bureau query. It specifies the score by which credit bureau calculates based on the customer's information in the entire sector.

Chart 2.3.1-1: Credit Bureau Score Distribution for Records



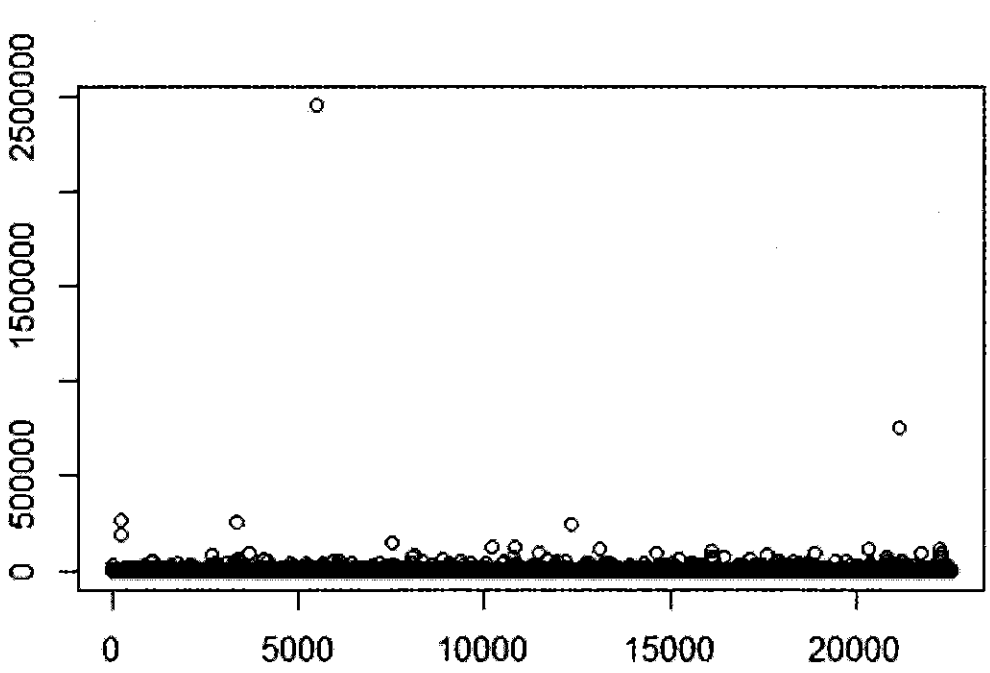
2.3.2. Credit Accounts Summary

The variables used in the study are mostly related to the credit accounts summary section. This section is designed to distinguish between open and closed

accounts. Besides, the distinction is made between the accounts in which the individual in question is the original or joint debtor and not.

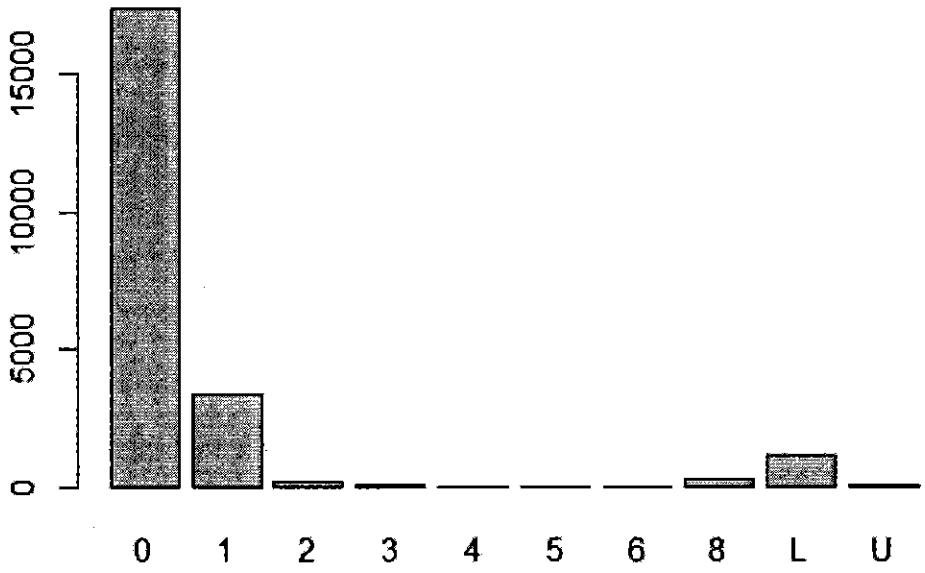
Total monthly liability in actual/joint accounts: Indicates the total amount of installments in all open accounts where the applicant (s) appear to be the principal or shareholder.

Chart 2.3.2-1: Distribution of Monthly Liabilities



Worst current-payment status: Specifies the worst current-payment status for all records returned from the query.

Chart 2.3.2-2: Distribution of Worst Current Payment Status



The descriptions of the codes for the payment status information are given below.

0: Means no delay.

1: Means one payment is delayed.

2: Means two payments are delayed.

3: Means three payments are delayed.

4: Means four payments are delayed.

5: Means five payments are delayed.

6: Means six payments are delayed.

8: Means it is on administrative follow-up.

D: Refers to a stationary account.

L: Indicates that the client is in legal pursuit.

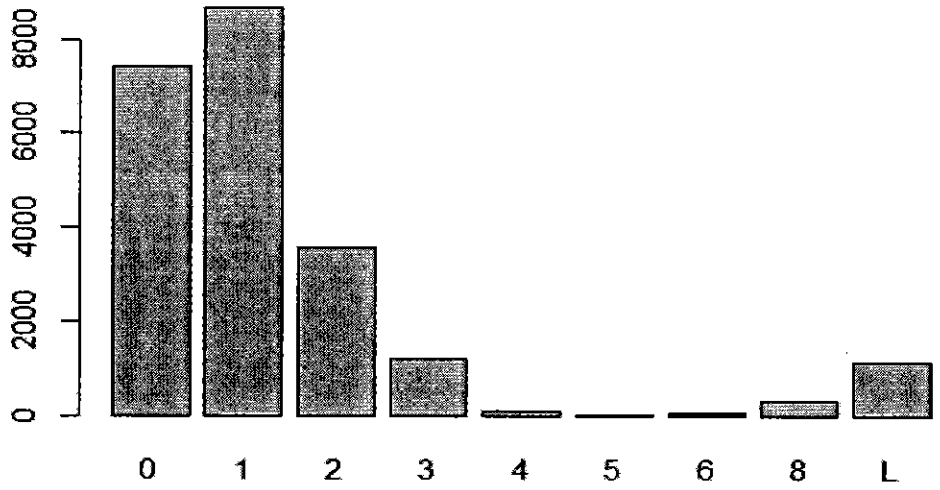
U: Refers to unclassified status.

X: Indicates missing information.

NULL: Refers to empty status.

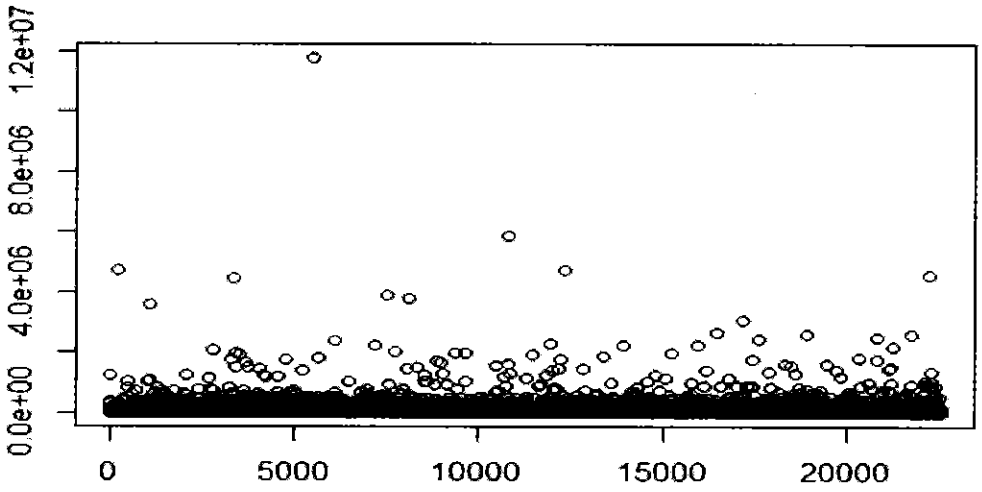
Worst payment status in the system: The worst payment status in the system (takes into account all 36-month old payment terms in the payment histories of all records returned in the query).

Chart 2.3.2-3: Distribution of the Worst Payment Status in the System



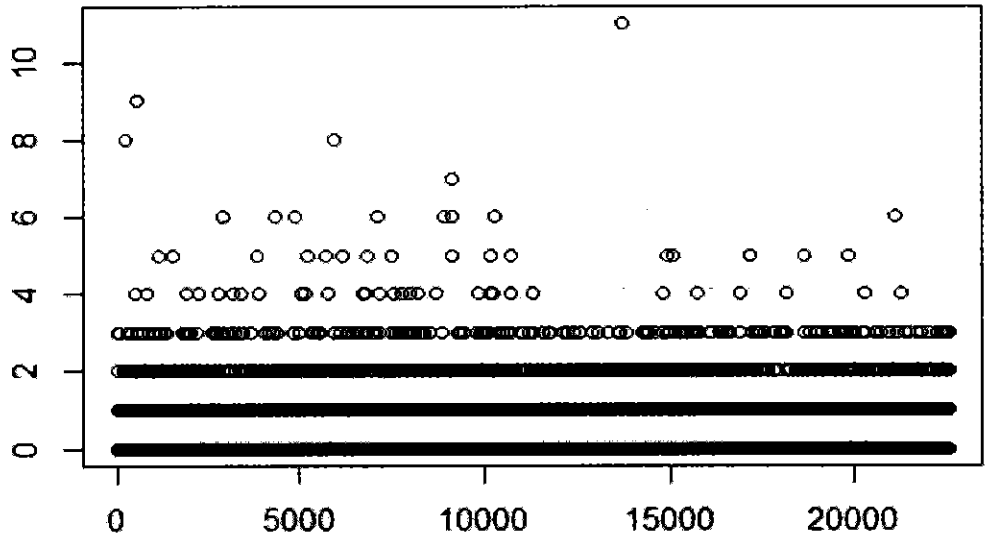
Total debt balance of all records returned as query result: Specifies the total unpaid balances of all records returned as a result of the query.

Chart 2.3.2-4: Distribution of Total Debt Balance of All Records Returned as Query Result



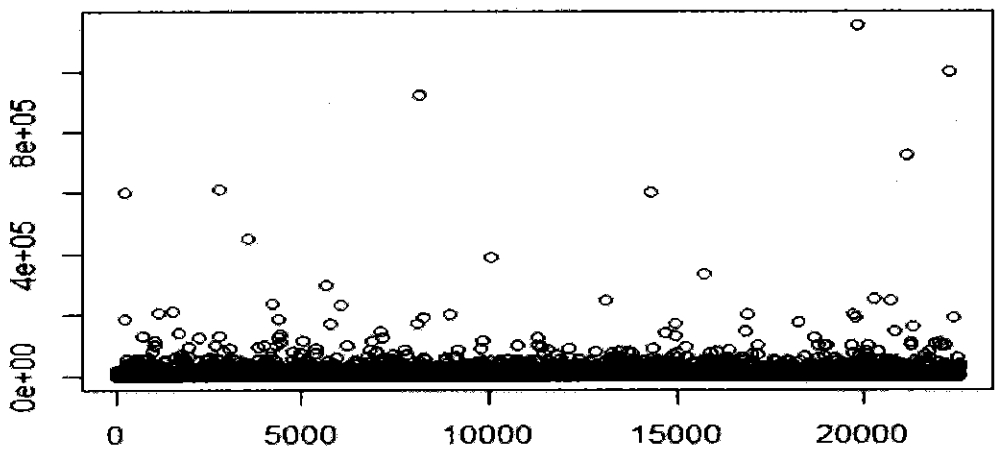
The number of accounts-1: Indicates the number of credit accounts opened in the last three months.

Chart 2.3.2-5: Distribution of Accounts in the Last 3 Months



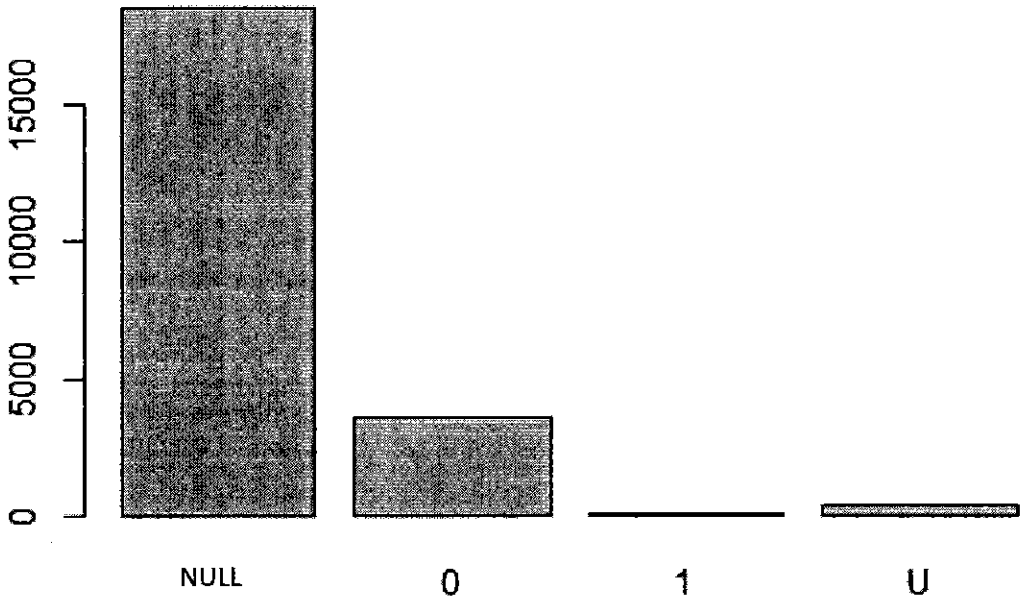
Total debt balance (excluding mortgage loans)-1: Indicates the sum of outstanding balances in the last three months excluding mortgage loans.

Chart 2.3.2-6: Distribution of Total Debt Balance in Last 3 Months



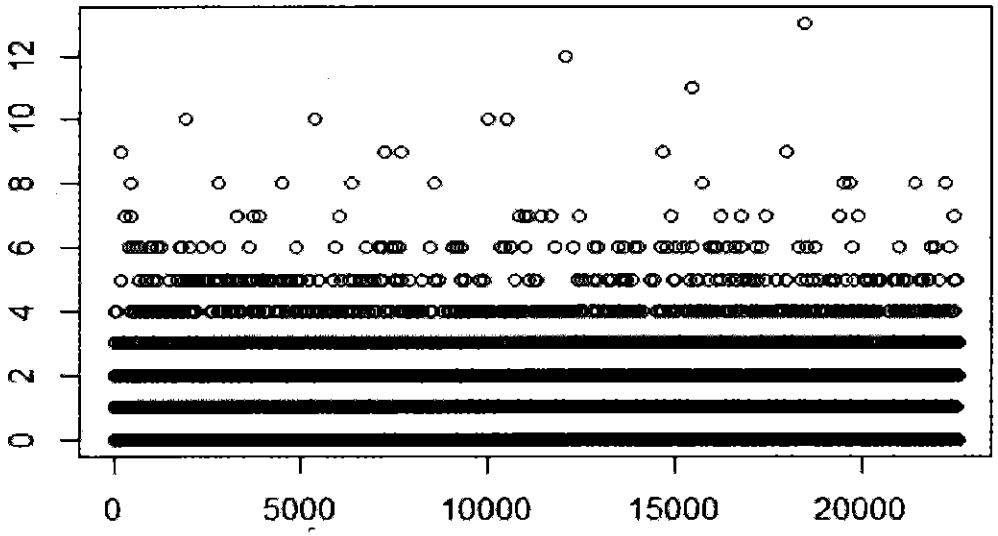
Worst payment status-1: Specifies the worst payment status for records in the last three months.

Chart 2.3.2-7: Distribution of the Worst Payment Status in the Last 3 Months



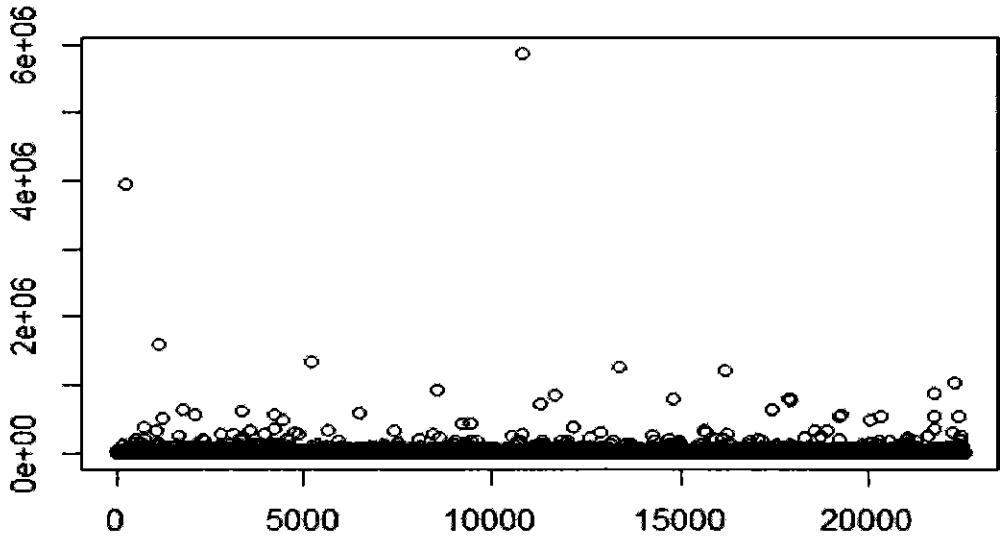
The number of accounts-2: Specifies the number of credit accounts opened in the last 4-12 months.

Chart 2.3.2-8: Distribution of Number of Credit Accounts in 4-12 Months



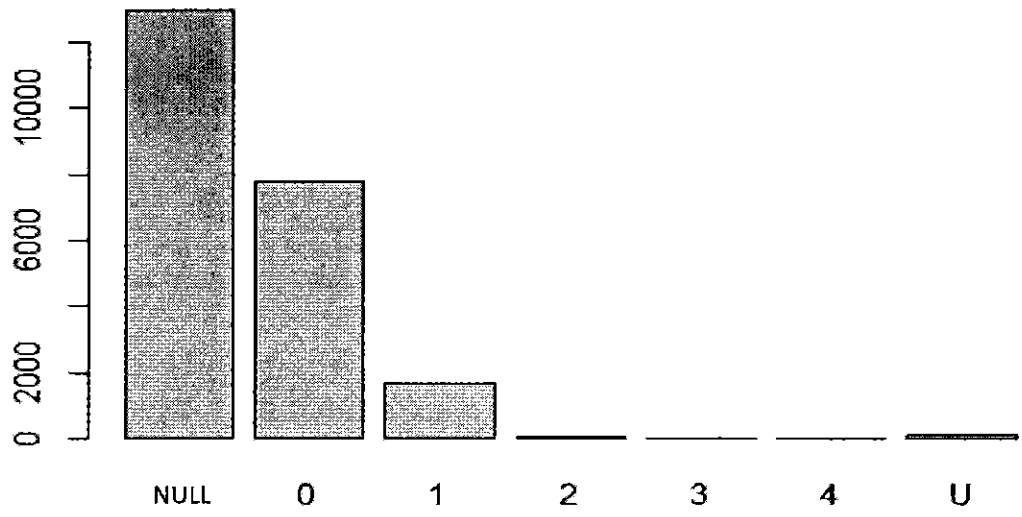
Total debt balance (excluding mortgage loans)-2: Indicates the sum of unpaid balance excluding mortgage loans in the last 4-12 months.

Chart 2.3.2-9: Distribution of Total Debt Balance in 4-12 Months



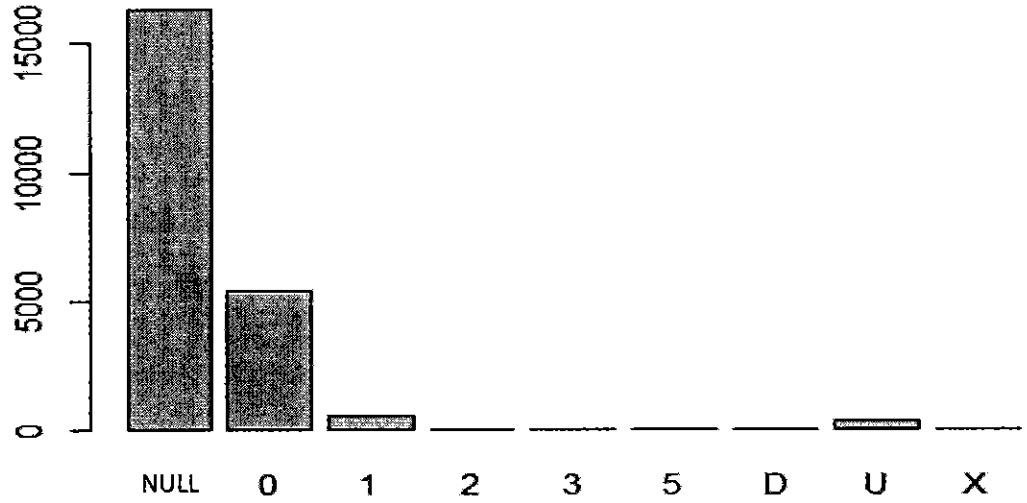
Worst payment status 0-6 months: Specifies the worst payment status in the last six months for the records in the last 4-12 months.

Chart 2.3.2-10: Distribution of the Worst Payment Status in the Last 6 Months in 4-12 Months



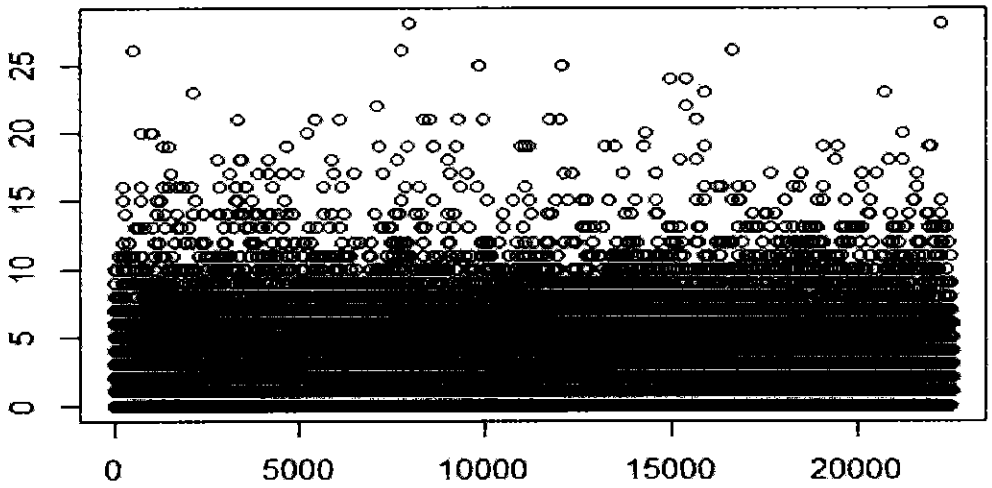
Worst payment status 7-12 months: Indicates the worst payment status in the last 7-12 months for the records in the last 4-12 months.

Chart 2.3.2-11: Distribution of the Worst Payment Status for the Last 7-12 Months in 4-12 Months



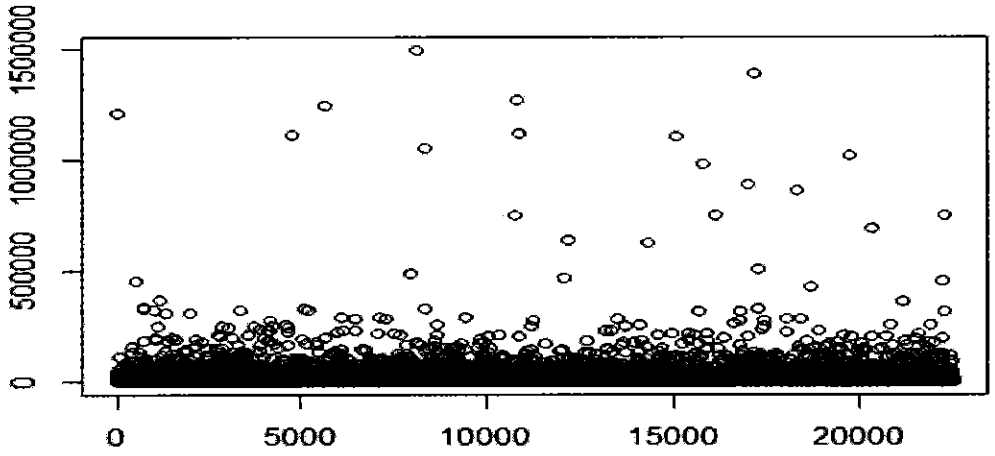
The number of accounts-3: Specifies the number of credit accounts opened during the last 12 months.

Chart 2.3.2-12: Distribution of Number of Credit Accounts Opened During Last 12 Months



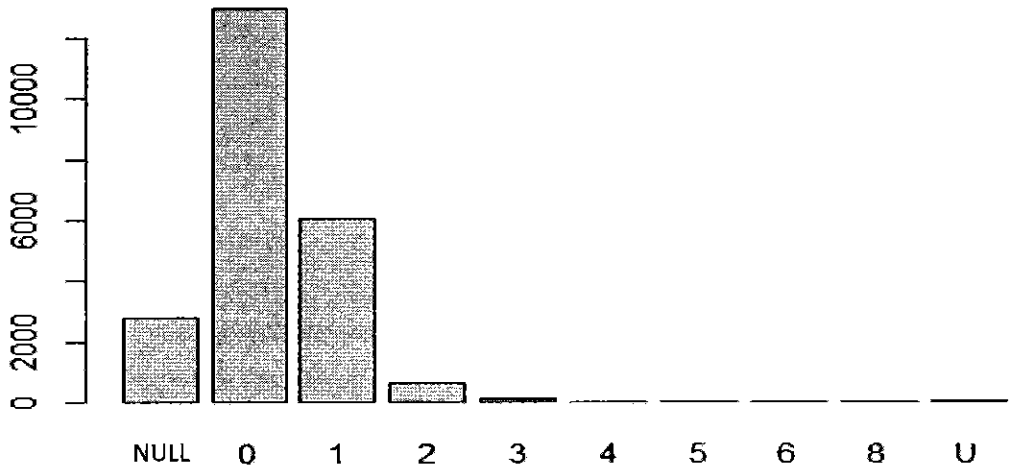
Total debt balance (excluding mortgage loans)-3: Indicates the sum of unpaid balances excluding mortgage loan accounts opened during the last 12 months.

Chart 2.3.2-13: Distribution of Unpaid Balances Regarding Loan Accounts Opening Before 12 Months



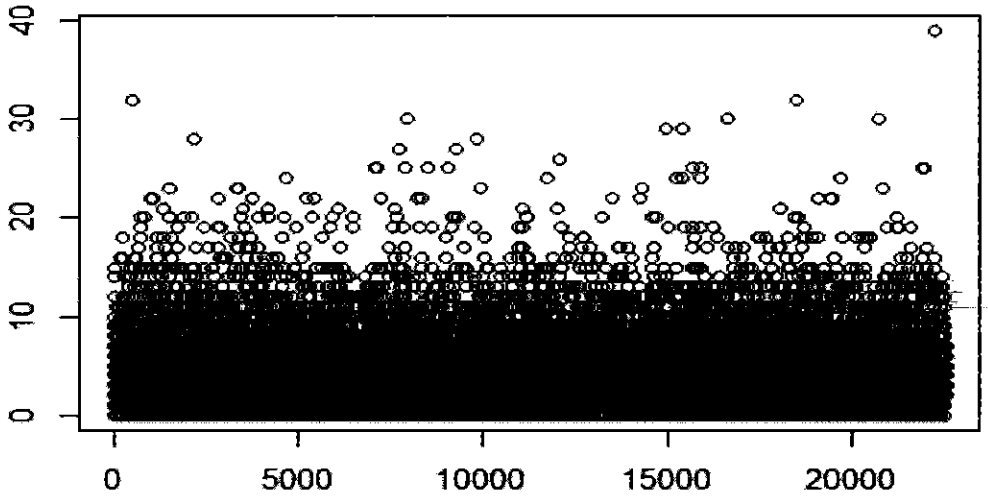
Worst payment status-2: Specifies the worst payment status for credit records opened during the last 12 months.

Chart 2.3.2-14: Distribution of the Worst Payment Status Regarding the Loan Accounts Opening Before the Last 12 Months



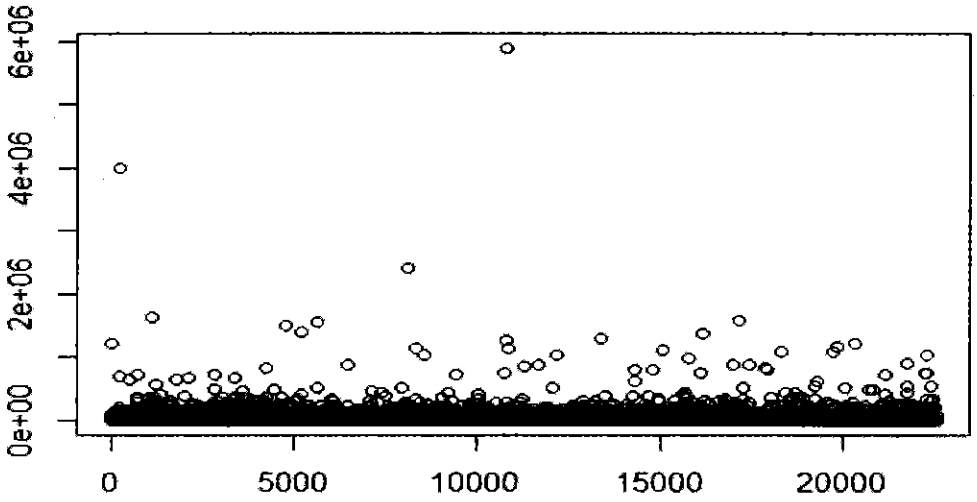
The number of accounts-4: Specifies the number of all open credit accounts.

Chart 2.3.2-15: Distribution of All Open Account Numbers



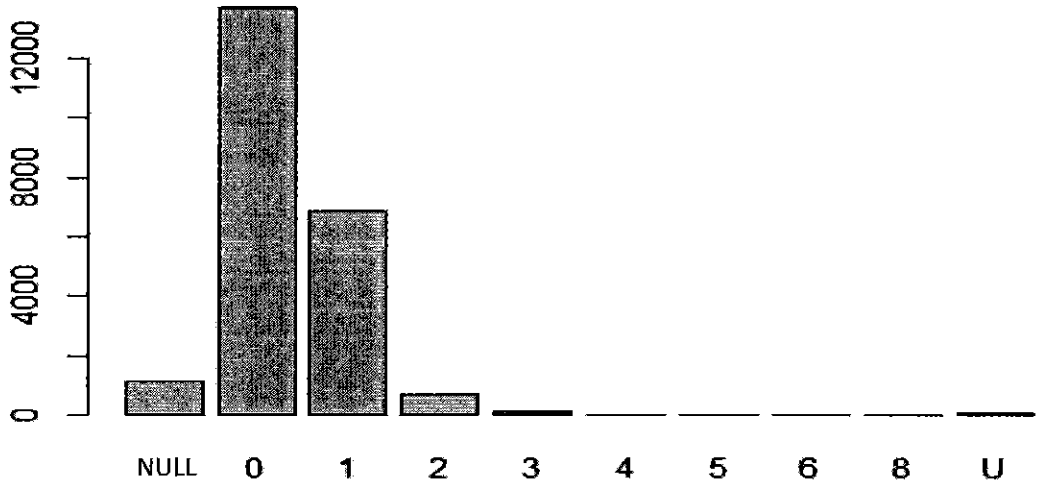
Total debt balance (excluding mortgage loans)-4: Indicates the sum of unpaid balances of all open credit accounts excluding mortgage loans.

Chart 2.3.2-16: Distribution of All Open Account Balances



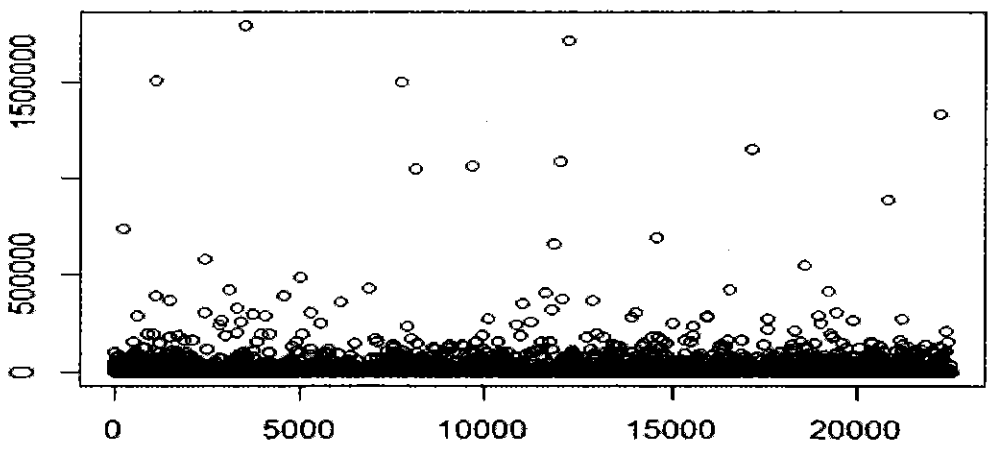
Worst payment status-3: Specifies the worst payment status for all open credit accounts.

Chart 2.3.2-17: Distribution of the Worst Payment Status of All Open Accounts



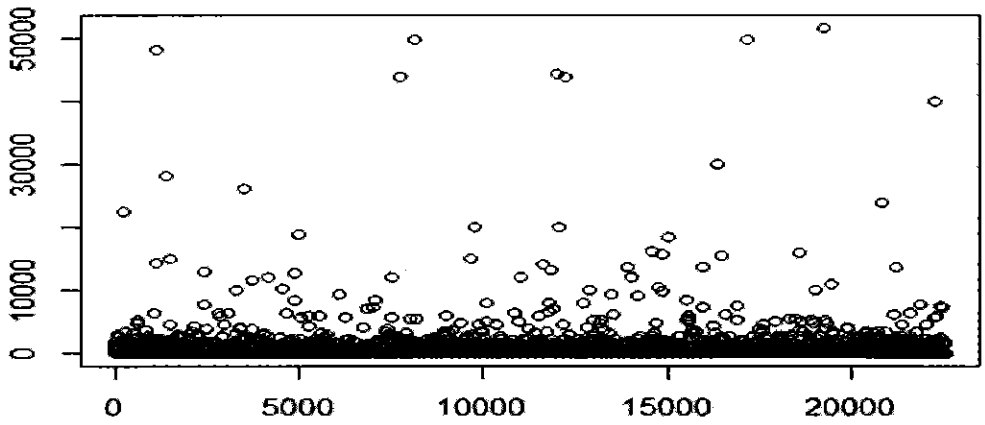
Total debt balance (including mortgage loans): Indicates the sum of outstanding balances of the open accounts where current-payment status is 0 including mortgage loans.

Chart 2.3.2-18: Distribution of Debt Balance of Accounts with Current-Payment Status 0



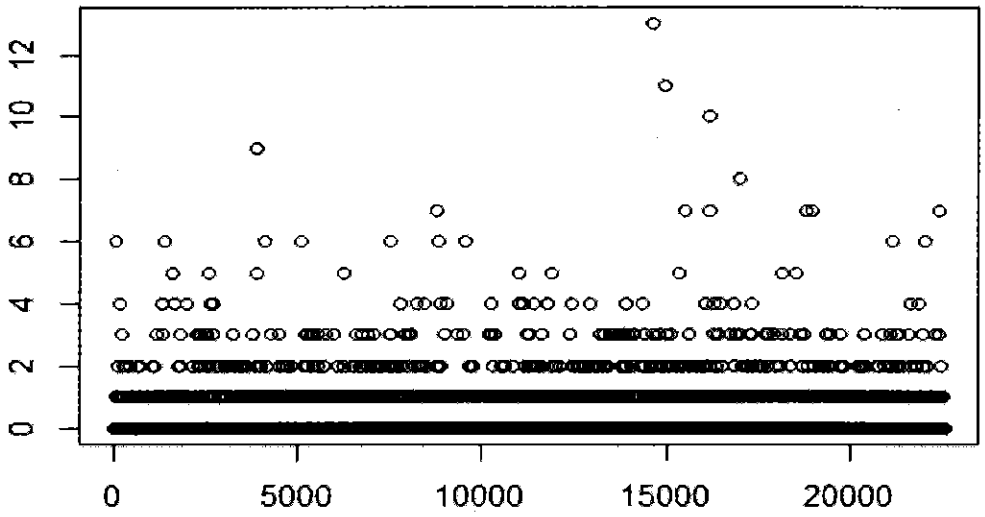
The monthly payment obligation: Specifies the sum of the installment amounts in all loans where current-payment status is 0.

Chart 2.3.2-19: Distribution of Monthly Payment Obligation for Accounts with Current-Payment Status 0



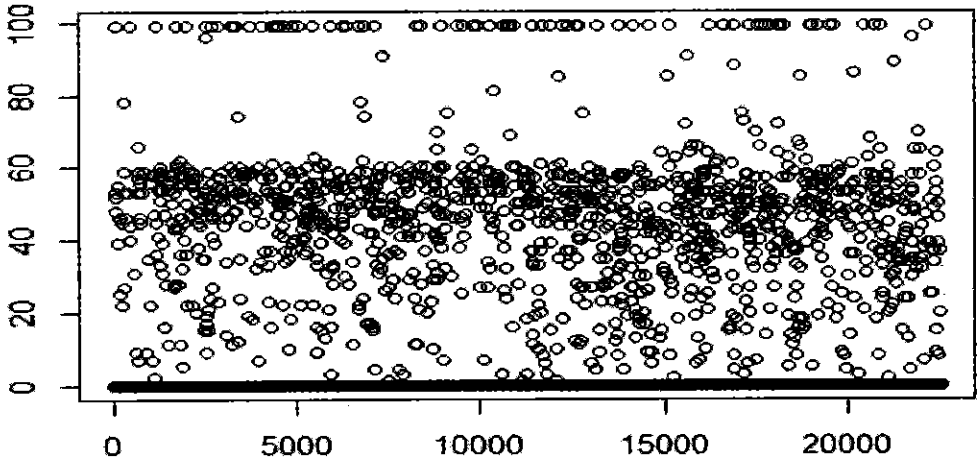
The number of accounts-5: Specifies the number of all administrative and legal follow-up credit accounts.

Chart 2.3.2-20: Distribution of NPL Accounts' Number



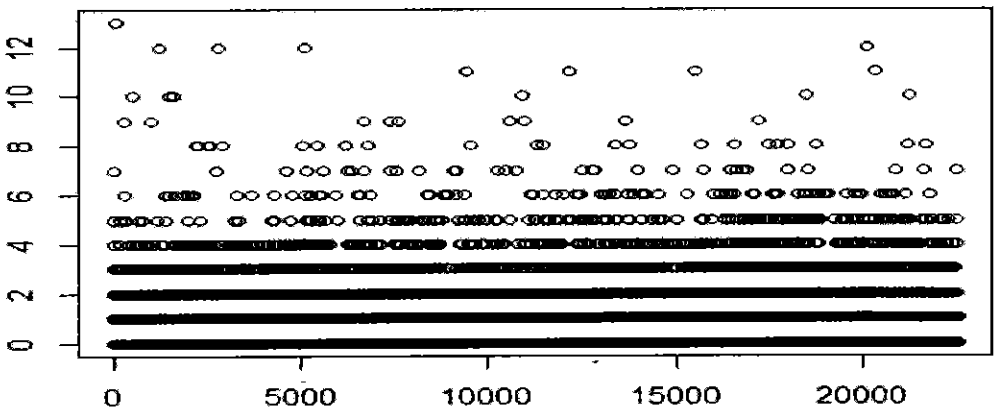
Time elapsed since the last administrative/legal follow-up: It indicates the time elapsed in months since the last administrative/legal follow-up.

Chart 2.3.2-21: Distribution of Time Elapsed Since the Last Follow-Up in Months



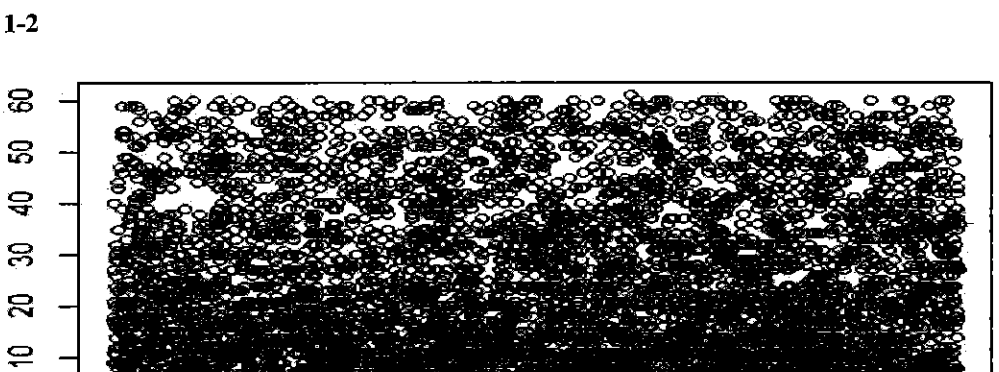
The number of accounts-6: Specifies the number of all closed credit accounts that have a delay in history where the worst payment status in the last 12 months is 1-2.

Chart 2.3.2-22: Distribution of the Numbers of All Closed Accounts That Have Delay in History Where the Worst Payment Status in the Last 12 Months is 1-2



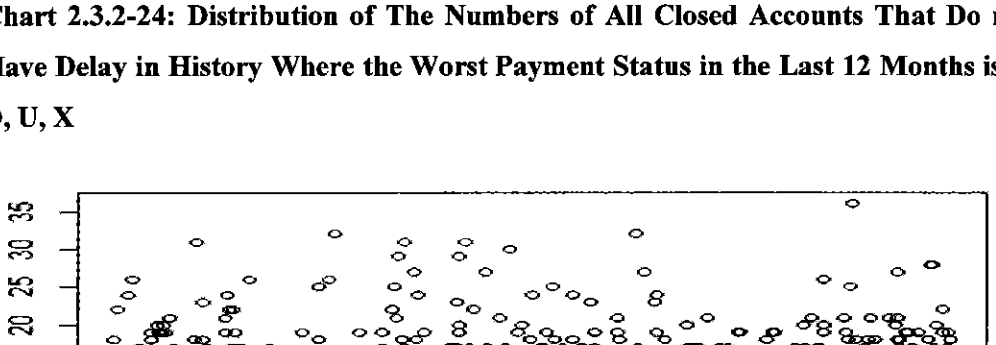
Time elapsed since the last credit closure-1: Specifies the time (in months) elapsed since the last credit closure of the credit accounts that have a delay in history where the worst payment status in the last 12 months is 1-2.

Chart 2.3.2-23: Distribution of the Time (in months) Elapsed Since the Last Credit Closure Accounts That Have Delay in History Where the Worst Payment Status in the Last 12 Months is 1-2



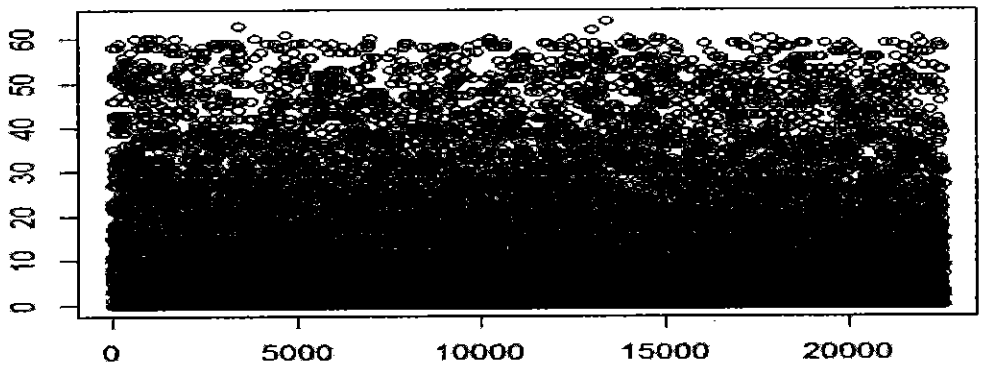
The number of accounts-7: Specifies the number of all closed accounts that do not have a delay in history where the worst payment status in the last 12 months is 0, D, U, X.

Chart 2.3.2-24: Distribution of The Numbers of All Closed Accounts That Do not Have Delay in History Where the Worst Payment Status in the Last 12 Months is 0, D, U, X



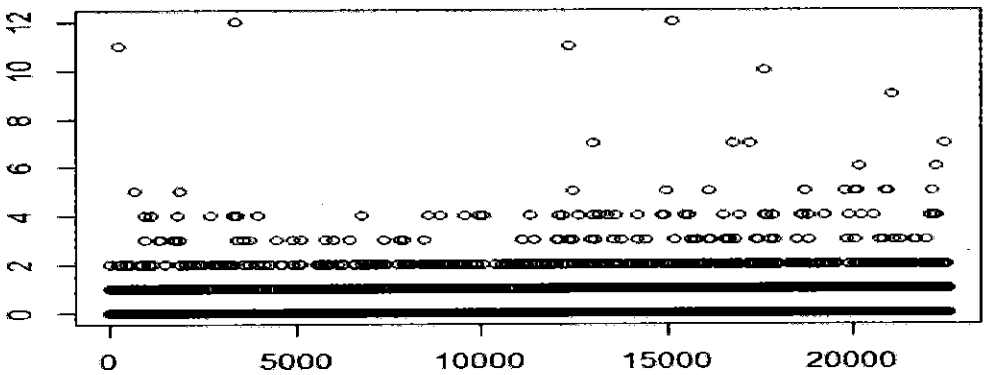
Time elapsed since the last credit closure-2: Specifies the time (in months) elapsed since the last credit closure of all closed accounts that do not have a delay in history where the worst payment status in the last 12 months is 0, D, U, X.

Chart 2.3.2-25: Distribution of the Time (in months) Elapsed Since the Last Credit Closure Of the Accounts That Do not Have Delay in History Where the Worst Payment Status in the Last 12 Months is 0, D, U, X



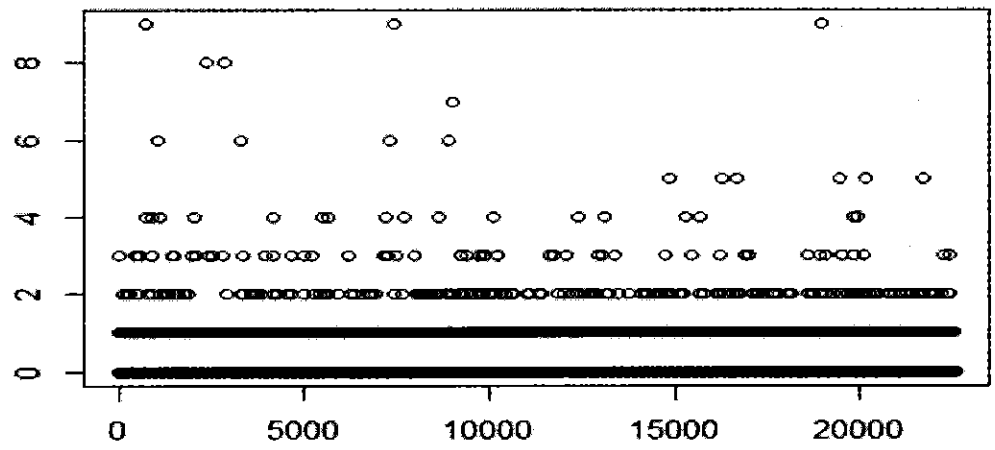
The number of the accounts of the person being questioned: Specifies the number of accounts that are closed with not a significant delay in history, where the person being questioned is the principal debtor.

Chart 2.3.2-26: Distribution of The Numbers of Own Closed Accounts of The Person Being Questioned That Do not Have Delay in History Where the Worst Payment Status in the Last 12 Months is 0, D, U, X



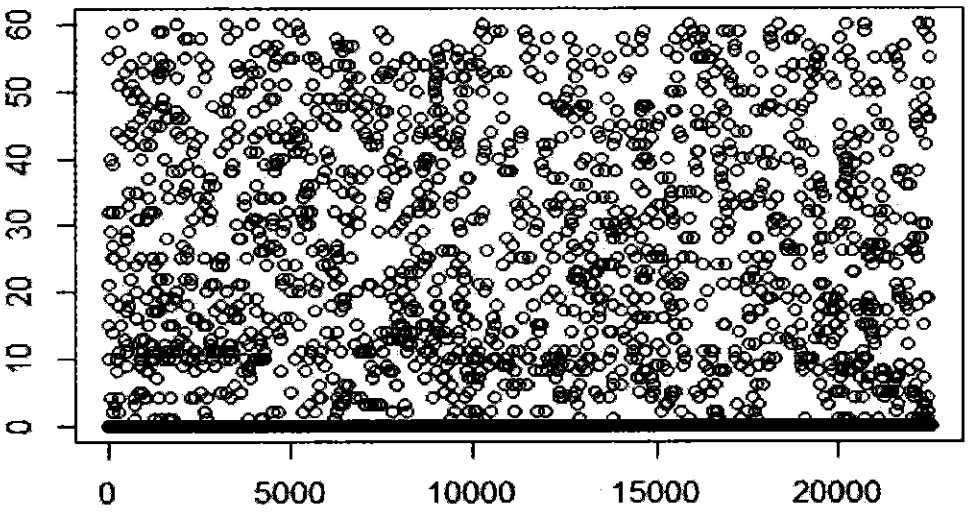
The number of accounts of the people being questioned: Specifies the number of credit accounts that are closed with not a significant delay in the history of the people being questioned.

Chart 2.3.2-27: Distribution of The Numbers of All Closed Accounts of The People Being Questioned That Do not Have Delayed in History Where the Worst Payment Status in the Last 12 Months is 0, D, U, X



Time elapsed since the last credit closure of the person being questioned: Indicates the time elapsed (in months) since the last credit closure of all closed accounts that do not have delay in history where the worst payment status in the last 12 months is 0, D, U, X for the person being questioned.

Chart 2.3.2-28: Distribution of the Time (in months) Elapsed Since the Last Credit Closure of The People Being Questioned That Do not Have Delay in History Where the Worst Payment Status in the Last 12 Months is 0, D, U, X



The numerical data for each variable is given below in the next two tables.

Table 2.3.2-1: The Numerical Value of Variables

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Overdue	0.00	0.00	0.50	0.50	1.00	1.00
Credit Bureau Score	0.00	1287.00	1460.00	1431.00	1656.00	1900.00
Monthly Liability	0.00	0.00	0.00	1408.00	1003.00	2451748.00
Total Debt Balance	0.00	1072.00	6860.00	43532.00	31450.00	11774927.00
Number Of Accounts In Last 3 Months	0.00	0.00	0.00	0.25	0.00	11.00
Total Debt Balance In Last 3 Months	0.00	0.00	0.00	2318.00	0.00	1150000.00
Number Of Credit Accounts In 4 To 12 Months	0.00	0.00	0.00	0.73	1.00	13.00
Total Debt Balance In 4 to 12 Months	0.00	0.00	0.00	6633.00	1357.00	5859930.00

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Number Of Credit Accounts In Last 12 Months	0.00	1.00	3.00	3.26	4.00	28.00
Total Credit Debt Balance Over 12 Months	0.00	77.00	2372.00	11401.00	9247.00	1488468.00
Number Of All Accounts	0.00	2.00	3.00	4.24	6.00	39.00
Total Debt Balance Of All Accounts	0.00	715.00	4456.00	20352.00	17670.00	5876307.00
Total Debt Balance Of Payment Status 0 Accounts	0.00	0.00	0.00	3899.00	0.00	1794657.00
Monthly Obligation Of Payment Status 0 Accounts	0.00	0.00	0.00	127.20	0.00	51524.00
Number Of NPL Accounts	0.00	0.00	0.00	0.09	0.00	13.00
Time Elapsed In NPL	0.00	0.00	0.00	2.72	0.00	99.00
Number Of Closed Accounts Of Payment Status Is 1 Or 2	0.00	0.00	0.00	0.59	1.00	13.00
Time Elapsed Of Credit Accounts Of Payment Status Is 1 Or 2	0.00	0.00	0.00	6.28	5.00	61.00
Number Of Closed Accounts In Last 12 Months	0.00	1.00	2.00	3.16	4.00	36.00
Number Of Closed Credit Accounts In Last 12 Months	0.00	1.00	6.00	11.55	17.00	64.00
Number Of Closed Own Accounts In Last 12 Months	0.00	0.00	0.00	0.12	0.00	12.00
Number Of Closed Joint Accounts In Last 12 Months	0.00	0.00	0.00	0.10	0.00	9.00
Time Elapsed Of Closed Joint Accounts In Last 12 Months	0.00	0.00	0.00	2.12	0.00	60.00

Table 2.3.2-2: The Numerical Value of Payment Status Variables

Variable	0	1	2	3	4	8	L	U	X	Other	NULL
Worst Current Payment Status	17350	3382	230	0	0	311	1150	71	0	60	0
Worst Ever Payment Status	7411	8654	3573	1223	0	340	1151	0	0	202	0
Worst Payment Status In Last 3 Months	3569	122	0	0	0	0	0	363	0	0	18500
Worst Payment Status Last 6 Months	7763	1646	57	6	2	0	0	142	0	0	12938
Worst Payment Status In Last 7 To 12 Months	5401	523	11	0	0	0	0	314	13	9	16283
Worst Credit Payment Status Over 12 Months	12943	6029	647	90	0	0	0	73	0	5	2767
Worst Payment Status Of All Accounts	13679	6859	700	96	0	0	0	90	0	7	1123

CHAPTER 3

MODEL

In this section, the machine learning algorithms are mentioned for the production of the model which predict whether a loan will be delayed in the future by using the lag status of the consumer loans and credit bureau data. At this point of view, the supervised learning technique which has wide usage area in the field of machine learning is considered as the most suitable method for solving this problem. Supervised learning is based on a technique of generating a function that can predict the future by categorizing over past data, training data set (Borkar, 2018, para. 2). In other words, in this technique, a matching function is generated between the inputs and the outputs. The data used in the training process includes both inputs and outputs (Uzun, 2016, para. 1). The function can be determined by the classification and regression methods. If outputs are continuous regression is used; otherwise, classification methods are used. In this study, the outputs are composed of delayed and non-delayed conditions, and they are discrete. That is why the classification methods were chosen in this study. The methods commonly used in the classification are detailed below.

3.1. LOGISTIC REGRESSION

Logistic regression is used to calculate the probability that it belongs to each of an output consisting of only two values, such as 0-1 or yes-no, through the training set was given to it (Akın, 2017, para. 1). In logistic regression, the output is called the dependent variable. The dependent variable may be qualitative or quantitative but should consist of only two states, not continuity. Other variables that are tried to be found in the data set with the result are called independent variables.

Logistic regression name comes from the logistic function. This function always produces a value between 0 and 1 only. There are odds and odds ratio concepts in logistic regression. To explain these concepts: The probability of logistic regression is between 0 and 1. For example; let's say that there are 1 white and 3 red balls in a bag. The probability of coming of the white ball is 1/4, and the probability of not coming is ¾. The ratio of the probability of coming (1/4) to the probability of not coming (3/4) refers to this, which is (1/4) / (3/4) (Şirin, 2016, para. 2).

We can obtain the logistic regression equation from the linear regression equation. A simple linear regression equation is as follows :

$$p(X) = \beta 0 + \beta 1X$$

In this equation, the argument X on the right side of the equation represents the independent variable; β1 represents the coefficient of the independent variable. As can be understood from this, the one-unit change in X affects the value of p (X) as much as β1 (dependent variable). It is clear that the result is a continuous variable with more than two values. The purpose of the logistics function is to calculate the probability that the independent variable is included in one of two classes. In order to achieve this, in the linear regression equation, we apply the logistic function to the right side of the equation so that the result (p (X)) is only between 0 and 1.

$$p(X) = e^{\beta 0 + \beta 1X} / 1 + e^{\beta 0 + \beta 1X}$$

When we leave alone the term $e^{\beta 0 + \beta 1X}$ in the equation, we get the following result.

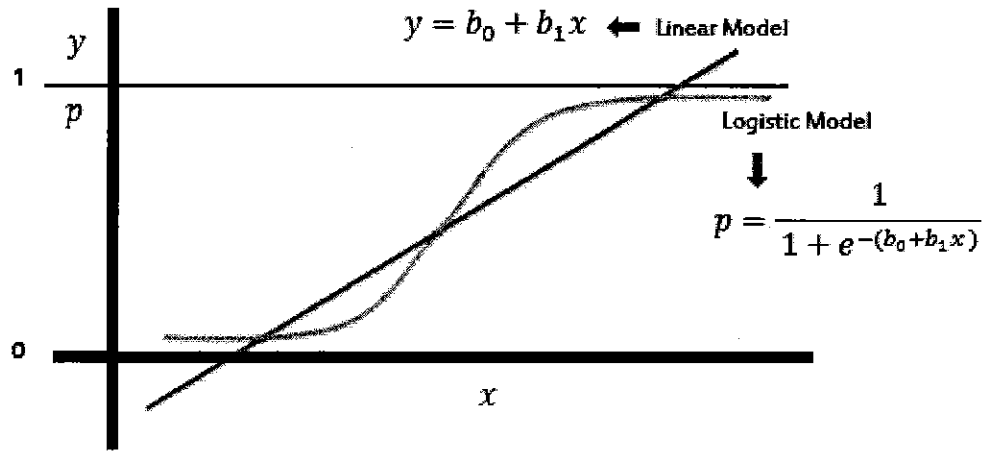
$$p(X) / 1 - p(X) = e^{\beta 0 + \beta 1X}$$

In logistic regression, p(X) / 1 - p(X) gives the odds value, and this value can be between 0 and ∞. When we take the logarithm of both sides of the equation, we get the following equation.

$$\log(p(X) / 1 - p(X)) = \beta 0 + \beta 1X$$

The left side of this equation is called log-odds or logit. In this equation, β1 will not cause the same change in p (X) for each unit change of X, anymore. So there will be no linear relationship between p (X) and x. The value of p (X) will now change according to the current value of x (James, Witten, Hastie, & Tibshirani, 2014, p. 133).

Chart 3.1-1: Comparative Graph of Linear Regression and Logistic Regression



Source : (Akın, 2017, p. 133)

3.2. LINEAR DISCRIMINANT ANALYSIS

In the case of logistic regression with only two different classes, using a logistic function, a direct link between the variable and the result is modeled. LDA is a model in which more than two finite number of different classes are analyzed.

In statistics, the conditional distribution of the output is modeled according to the given independent variables. In order to estimate these possibilities, an approach that is not directly linked as an alternative to logistic regression is considered. In this alternative approach, output classes are modeled separately for each independent variable. Then Bayes' theorem is used for the estimation of these classes. If it is assumed that the distributions present here are reasonable, it is seen that the LDA model is very close to logistic regression (James et al., 2014, p. 138).

There are several reasons why we need LDA in addition to logistic regression:

- When classes are well separated, parameter estimates for the logistic regression model are interestingly unstable. This problem does not occur in LDA.
- If n (training data) is small and the distribution of independent variables is approximately normal for each class, the linear discriminant model is still more stable than the logistic regression.
- In the case of more than two classes, logistic regression is usually not preferred.

3.2.1. Bayes' Theorem

Bayes' theorem is a commonly used mathematical formula found out by British mathematician Thomas Bayes to calculate the conditional probability (Unal, 2018, para. 1). Before we go into the details of this formula, let's talk a little about the possibility.

In order to express the possibility that a random event A will occur depending on a random event B, the notation P (A | B), called conditional probability, is used.

The conditional probability of an event A according to an arbitrary B event is shown below.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

In the event that two events (A and B), which are discrete and random, occur one after another, the probability of occurrence of the second event is denoted as P (A, B) or P (B, A). With the help of the change feature of the multiplication, this situation can be written in two different ways as follows.

$$P(A \cap B) = P(A \mid B)P(B)$$
$$P(A \cap B) = P(B \mid A)P(A)$$

Bayes' theorem refers to the relationship between the contingent probabilities and the marginal probabilities for a randomized event A and another randomized event B, with the following formula:

$$P(A|B) = \frac{P(A)\ P(B|A)}{P(B)}$$

P(A|B) = The probability that event A occurs when event B occurs,
P(A) = The likelihood of occurrence of event A,
P(B|A) = The probability that event B occurs when event A occurs,
P(B) = The likelihood of occurrence of event B.

Let's examine how this formula is obtained. By conditional probability rule :

P(A|B) = P(A and B) / P(B)

Similarly :

$P(B|A) = P(B \text{ and } A) / P(A)$

$P(A \text{ and } B) = P(B \text{ and } A)$ is obtained. From the above two equations, we can obtain the Bayes' rule.

$P(A|B).P(B)=P(B|A).P(A)$

$P(A|B) = P(B|A) . P(A) / P(B)$

3.3. QUADRATIC DISCRIMINANT ANALYSIS

In LDA, the variance is assumed to be equal for each class. The approach of QDA is different. In QDA, each class has its variance. Thanks to QDA, the margin of error in a variable assignment to a group is minimized (Akbaba, 2005). It is recommended to use QDA in cases where the training set is too large, the variance of the class itself is not the main focus or the assumption that the variance for all classes is the same is entirely meaningless (James et al., 2014, p. 149).

3.4. K-NEAREST NEIGHBORS

Linear regression has a parametric approach. Because it provides a linear function for f (X). Parametric methods have some advantageous aspects. In general, they are easy to understand, because only a small number of layers must be estimated. Also, parametric methods have some disadvantages. For example, suppose there is a linear relationship between X and Y. If the actual relationship is not linear, the outputs produced by the resulting model will be very incompatible with the actual data (James et al., 2014, p. 104). In contrast, nonparametric methods can be a more flexible alternative in regression because they do not provide a

parametric form for f (X). The K-Nearest Neighborhood (KNN) algorithm is the simplest and most known of the non-parametric methods.

The KNN algorithm is an easy-to-use method and enters the method of supervised learning. It is widely used in the solution of classification and regression problems. In this algorithm, the distance of the new data to be added to the sample data set from the existing data is calculated, and a k number of nearby neighborhoods are considered. As can be understood, it is a costly process to calculate the distance from each data when working with massive data. This is the disadvantageous aspect of this algorithm (Ulgen, 2017, para. 1).

The KKN algorithm consists of the following steps :

- Firstly, the parameter k is set. This parameter is the number of neighbors closest to a given point. For example, if we accept $k = 3$, then we will classify according to the nearest three neighbors.
- The distance from the new data to be added to the sample data set is calculated from the existing data.
- The nearest k neighbor is taken over. According to the values, k is associated with the class of neighbors.
- The selected class is the estimated class for the newly added data. So the new incoming data is classified.

CHAPTER 4

APPLICATION

In this section, the algorithms of the models described in the previous section have been implemented through the data obtained from the bank. Implementation was done using the R programming language in R studio. Half of the data was used as a training set, while the other half was used to calculate the accuracy of the estimated results produced by the model. All the program codes for construction the models and the definition of the variables are in the Appendix.

The data set consists of 30 columns (variables) and 22.554 lines (observations). In order to construct the models we need to divide our data into two categories those are training and test data sets.

After run the logistic regression function for all variables in the data set we obtained below results.

Table 4-1: The Results of the Logistic Regression Model for All Variables

	Min	**1Q**	**Median**	**3Q**	**Max**
Deviance Residuals	-2.87	-1.11	0.71	1.03	2.05
Coefficients: (6 not defined because of singularities)					
	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**	
(Intercept)	0.76	0.15	4.97	0.07	***
Credit Bureau Score	-0.00	0.06	-2.93	0.00	**
Monthly Liability	-0.09	0.01	-1.86	0.07	.
Worst Current Payment Status Is 1	0.23	0.07	3.31	0.00	***
Worst Current Payment Status Is 2	0.26	0.24	1.09	0.28	
Worst Current Payment Status Is 3	-0.31	0.48	-0.65	0.52	
Worst Current Payment Status Is 4	13.24	535.40	0.03	0.98	
Worst Current Payment Status Is 6	14.00	535.40	0.03	0.98	
Worst Current Payment Status Is 8	-3.06	1.06	-2.88	0.00	**
Worst Current Payment Status Is L	10.78	535.40	0.02	0.98	
Worst Current Payment Status Is U	-0.61	0.39	-1.55	0.12	
Worst Ever Payment Status Is 1	0.72	0.06	12.49	0.00	***
Worst Ever Payment Status Is 2	0.79	0.08	10.38	0.00	***

	Estimate	Std. Error	z value	Pr(>|z|)	
Worst Ever Payment Status Is 3	0.78	0.11	7.24	0.00	***
Worst Ever Payment Status Is 4	0.64	0.33	1.91	0.06	.
Worst Ever Payment Status Is 5	1.33	0.63	2.13	0.03	*
Worst Ever Payment Status Is 6	0.42	0.37	1.14	0.26	
Worst Ever Payment Status Is 8	3.22	1.03	3.11	0.00	**
Worst Ever Payment Status Is L	-10.34	535.40	-0.02	0.99	
Total Debt Balance	0.07	0.00	2.55	0.01	*
Number Of Accounts In Last 3 Months	0.24	0.07	3.38	0.00	***
Total Debt Balance In Last 3 Months	-0.01	0.00	-2.40	0.02	*
Worst Payment Status In Last 3 Months Is 0	0.30	0.11	2.69	0.01	**
Worst Payment Status In Last 3 Months Is 1	0.54	0.41	1.32	0.19	
Worst Payment Status In Last 3 Months Is U	0.12	0.19	0.61	0.54	
Number Of Credit Accounts In 4 To 12 Months	0.16	0.03	5.01	0.01	***
Total Debt Balance In 4 To 12 Months	-0.01	0.00	-2.10	0.04	*
Worst Payment Status In Last 6 Months Is 0	0.05	0.08	0.70	0.48	
Worst Payment Status In Last 6 Months Is 1	-0.04	0.14	-0.31	0.76	
Worst Payment Status In Last 6 Months Is 2	-25.78	439.20	-0.06	0.95	
Worst Payment Status In Last 6 Months Is 3	11.75	281.20	0.04	0.97	
Worst Payment Status In Last 6 Months Is U	-0.37	0.27	-1.36	0.18	
Worst Payment Status In Last 7 To 12 Months Is 0	-0.29	0.07	-4.18	0.02	***
Worst Payment Status In Last 7 To 12 Months Is 1	-0.06	0.16	-0.37	0.72	
Worst Payment Status In Last 7 To 12 Months Is 2	-0.61	0.94	-0.65	0.52	
Worst Payment Status In Last 7 To 12 Months Is 5	11.08	535.40	0.02	0.98	
Worst Payment Status In Last 7 To 12 Months Is D	12.51	228.50	0.06	0.96	
Worst Payment Status In Last 7 To 12 Months Is U	-0.08	0.19	-0.40	0.69	
Worst Payment Status In Last 7 To 12 Months Is X	0.32	0.92	0.35	0.73	
Number Of Credit Accounts In Last 12 Months	-0.09	0.01	-8.99	0.00	***
Total Credit Debt Balance Over 12 Months	0.01	0.00	4.02	0.04	***
Worst Credit Payment Status Over 12 Months Is 0	-0.39	0.09	-4.25	0.02	***
Worst Credit Payment Status Over 12 Months Is 1	-0.17	0.18	-0.98	0.33	
Worst Credit Payment Status Over 12 Months Is 2	-25.67	439.20	-0.06	0.95	

	Estimate	Std. Error	z value	Pr(>|z|)	
Worst Credit Payment Status Over 12 Months Is 3	-0.43	0.38	-1.14	0.25	
Worst Credit Payment Status Over 12 Months Is 4	12.68	373.70	0.03	0.97	
Worst Credit Payment Status Over 12 Months Is 6	-0.97	757.20	-0.00	1.00	
Worst Credit Payment Status Over 12 Months Is 8	9.03	535.40	0.02	0.99	
Worst Credit Payment Status Over 12 Months Is U	0.24	0.44	0.54	0.59	
Number Of All Accounts	NA	NA	NA	NA	
Total Debt Balance Of All Accounts	NA	NA	NA	NA	
Worst Payment Status Of All Accounts Is 0	-0.30	0.13	-2.39	0.02	*
Worst Payment Status Of All Accounts Is 1	-0.10	0.20	-0.48	0.63	
Worst Payment Status Of All Accounts Is 2	25.23	439.20	0.06	0.95	
Worst Payment Status Of All Accounts Is 3	NA	NA	NA	NA	
Worst Payment Status Of All Accounts Is 4	NA	NA	NA	NA	
Worst Payment Status Of All Accounts Is 6	NA	NA	NA	NA	
Worst Payment Status Of All Accounts Is 8	NA	NA	NA	NA	
Worst Payment Status Of All Accounts Is U	-0.12	0.42	-0.29	0.79	
Total Debt Balance Of Payment Status Is 0 Accounts	0.01	0.00	1.98	0.05	*
Monthly Obligation Of Payment Status Is 0 Accounts	-0.00	0.03	-2.38	0.02	*
Number Of NPL Accounts	-0.22	0.08	-2.79	0.01	**
Time Elapsed In NPL	0.01	0.00	1.87	0.06	.
Number Of Closed Accounts Of Payment Status Is 1 Or 2	0.15	0.02	6.17	0.00	***
Time Elapsed Of Credit Accounts Of Payment Status Is 1 Or 2	-0.01	0.00	-4.27	0.01	***
Number Of Closed Accounts In Last 12 Months	-0.05	0.01	-6.53	0.00	***
Number Of Closed Credit Accounts In Last 12 Months	0.00	0.00	0.37	0.71	
Number Of Closed Own Accounts In Last 12 Months	0.12	0.06	2.21	0.03	*
Number Of Closed Joint Accounts In Last 12 Months	-0.20	0.07	-3.00	0.00	**
Time Elapsed Of Closed Joint Accounts In Last 12 Months	-0.00	0.00	-1.02	0.31	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Dispersion parameter for binomial family taken to be 1)					
Null deviance: 15529 on 11276 degrees of freedom					
Residual deviance: 14302 on 11213 degrees of freedom					

AIC: 14430
Number of Fisher Scoring iterations: 12

From the above results table we obtain some indicators that help us to check the statistical significance. Those are briefly summarized below:

Deviance Residuals: For poisson or linear regression, it is desired that they be more or less than normal distribution. It is can be checked whether the absolute value 1Q and 3Q are closed whether the median value is zero. Probably there are some problems in the data if any of them is not like as expected. These are can be ignored for logistic regression.

Coefficients: This is the output.

- Intercept: For logistic regression, this value is far from zero. The difference between the number of observations in each class is greater. The standard error shows how unclear we are about this, so it is better to be lower.
- Inputs: Show us how much the output will change when we increase it by one unit. The larger the estimate, the greater the effect of the input variable on the output. In addition, positive coefficient means increase in the probability of the event when you have a positive change in the input value, negative coefficient means decrease in the probability of the event when you have a positive change in the input value.
- Signif. Codes: Show the significance of each input and intercept.
- (Dispersion parameter for binomial family taken to be 1): This only allow us to know that there are some additional scaling parameter to help us comply with the model. This can be ignored.
- Null deviance: Shows us how well we can forecast our output using the intercept. Smaller value of it is better.
- Residual deviance: Shows us how well we can estimate our output using the intercept and inputs. It's small value is better. The greater the difference between the null deviation and the residual deviation, the more useful it is for out input variables to estimate the output variable.

- AIC: This is a prediction of how well our model defines patterns in the data. It is mainly used to compare the models that trained in the same data set. The lower value of it means the model works better.
- Number of Fisher Scoring iterations: This is a measure of the time it takes to construct our model. It is can be ignored (Interpretation of R's output, 2014, para. 3).

In addition to the coefficients, the p-values are important. We can think that a linear model is statistically significant only if these values are lower than the predetermined statistical significance level, which is 0.05. The smaller the value of $Pr(>|z|)$, the greater the significance of the variable (Prabhakaran, 2017, para. 12). We can understand the significant of each variable in the model by looking at the asterisk it has. The higher the number of asterisk, the higher the significance of the variable. As seen above, the effect of variables are different. Some of them even don't have any meaning statistically, some do not have a significant effect on the model, but some have. Therefore we need to use the effective variables to make our model meaningful.

Table 4-2: The Most Common Metrics in Model Section

Statistic	Criterion
R-Squared	Higher the better (> 0.70)
Adj R-Squared	Higher the better
F-Statistic	Higher the better
Std. Error	Closer to zero the better
t-statistic	Should be greater 1.96 for p-value to be less than 0.05
AIC	Lower the better
BIC	Lower the better
Mallows cp	Should be close to the number of predictors in model
MAPE (Mean absolute percentage error)	Lower the better
MSE (Mean squared error)	Lower the better
Min_Max Accuracy => mean(min(actual, predicted)/max(actual, predicted))	Higher the better

Source: (Prabhakaran, 2017, para. 16)

When we run the logistic regression function with only significant variables, the result is as follows.

Table 4-3: The Results of the Logistic Regression Model

	Min	1Q	Median	3Q	Max
Deviance Residuals	-3.01	-1.16	0.76	1.09	1.97
Coefficients:					
	Estimate	**Std.Error**	**z value**	**Pr(>|z|)**	
(Intercept)	1.31	0.11	11.69	0.00	***
Credit Bureau Score	-0.00	0.05	-10.24	0.00	***
Monthly Liability	-0.01	0.01	-2.17	0.03	*
Total Debt Balance	0.01	0.00	3.02	0.00	**
Number Of Accounts In Last 3 Months	0.35	0.04	9.25	0.00	***
Total Debt Balance In Last 3 Months	-0.01	0.00	-2.27	0.02	*
Number Of Credit Accounts In 4 To 12 Months	0.10	0.02	5.29	0.00	***
Total Debt Balance In 4 To 12 Months	-0.01	0.00	-2.28	0.02	*
Number Of Credit Accounts In Last 12 Months	-0.07	0.01	-7.56	0.00	***
Total Credit Debt Balance Over 12 Months	0.01	0.00	3.75	0.00	***
Total Debt Balance Of Payment Status Is 0 Accounts	0.01	0.00	2.03	0.04	*
Monthly Obligation Of Payment Status Is 0 Accounts	-0.00	0.03	-2.57	0.01	*
Number Of NPL Accounts	-0.22	0.05	-4.43	0.00	***
Number Of Closed Accounts Of Payment Status Is 1 Or 2	0.24	0.02	11.14	0.00	***
Number Of Closed Accounts In Last 12 Months	-0.05	0.07	-6.64	0.00	***
Number Of Closed Own Accounts In Last 12 Months	0.14	0.06	2.44	0.02	*
Number Of Closed Joint Accounts In Last 12 Months	-0.24	0.05	-4.69	0.01	***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Dispersion parameter for binomial family taken to be 1)					
Null deviance: 15529 on 11276 degrees of freedom					
Residual deviance: 14827 on 11260 degrees of freedom					
AIC: 14861					
Number of Fisher Scoring iterations: 4					

Table 4-4: The Predictions of the Logistic Regression Model

Predictions		
	0	1
0	2659	1182
1	3521	3915
Mean		
		0.58

The results and predictions of the new logistic regression model are seen above. Now the variables used in the model are effective and their coefficients are as written above. After finding the logistic regression model we made prediction for all test data set and compared the estimated results with the actual results to find the accuracy rate of the model. By using *mean* function in R, we computed the rate of correct results in the test data. In this case, logistic regression model correctly estimated the credit default with the accuracy rate of 58.30%. This result shows us that logistic regression model works better than the random guessing (50.00%). In addition, the false positive rate is 2659/11277=23.58%, the false negative rate is 1182/11277=10.48%, true negative rate is 3521/11277=31.22% and the true positive rate is 3915/11277=34.72%. In this model, the accuracy rate of overdue cases is higher.

Now we use the same data we used for logistic regression, and we create a model for LDA.

Table 4-5: The Results of the LDA Model

	False	True
Prior probabilities of groups :	0.45	0.55
Group means :		
Credit Bureau Score	1507.26	1402.38
Monthly Liability	1080.62	1730.34
Total Debt Balance	35168.97	50535.96
Number Of Accounts In Last 3 Months	0.17	0.32
Total Debt Balance In Last 3 Months	1591.46	2635.63
Number Of Credit Accounts In 4 To 12 Months	0.62	0.81
Total Debt Balance In 4 To 12 Months	5862.21	7760.76

	False	True
Number Of Credit Accounts In Last 12 Months	3.36	3.10
Total Credit Debt Balance Over 12 Months	9902.90	12679.63
Total Debt Balance Of Payment Status Is 0 Accounts	3287.17	4155.74
Monthly Obligation Of Payment Status Is 0 Accounts	114.82	114.69
Number Of NPL Accounts	0.08	0.08
Number Of Closed Accounts Of Payment Status Is 1 Or 2	0.44	0.65
Number Of Closed Accounts In Last 12 Months	3.36	2.88
Number Of Closed Own Accounts In Last 12 Months	0.08	0.09
Number Of Closed Joint Accounts In Last 12 Months	0.12	0.08
Coefficients of linear discriminants:		
	LD1	
Credit Bureau Score	0.00	
Monthly Liability	-0.02	
Total Debt Balance	0.00	
Number Of Accounts In Last 3 Months	0.61	
Total Debt Balance In Last 3 Months	-0.02	
Number Of Credit Accounts In 4 To 12 Months	0.21	
Total Debt Balance In 4 To 12 Months	0.00	
Number Of Credit Accounts In Last 12 Months	-0.12	
Total Credit Debt Balance Over 12 Months	0.01	
Total Debt Balance Of Payment Status Is 0 Accounts	0.01	
Monthly Obligation Of Payment Status Is 0 Accounts	-0.00	
Number Of NPL Accounts	-0.42	
Number Of Closed Accounts Of Payment Status Is 1 Or 2	0.44	
Number Of Closed Accounts In Last 12 Months	-0.09	
Number Of Closed Own Accounts In Last 12 Months	0.27	
Number Of Closed Joint Accounts In Last 12 Months	-0.01	

Table 4-6: The Predictions of the LDA Model

Predictions		
	False	**True**
False	2598	1171
True	3582	3926
Mean		
		0.58

As seen above, by execution the LDA function we obtained the probabilities of each group and the coefficient of each variables. By using *pred* function we made

estimation for the test data and compared the predicted results with the real results. As of the result, we could see that the accuracy rate of the linear discriminant model in this study was 57.90%, which is not better than logistic regression but not worse so much at the same time. Also, the false positive rate is 23.04%, the false negative rate is 10.38%, true negative rate is 31.76% and the true positive rate is 34.81%. In this model, also, the accuracy rate of overdue cases is higher.

Now we use the same data we used in the previous two models to create a model for QDA.

Table 4-7: The Results of the QDA Model

	False	True
Prior probabilities of groups :	0.45	0.55
Group means :		
Credit Bureau Score	1507.26	1402.38
Monthly Liability	1080.62	1730.34
Total Debt Balance	35168.97	50535.96
Number Of Accounts In Last 3 Months	0.17	0.32
Total Debt Balance In Last 3 Months	1591.46	2635.63
Number Of Credit Accounts In 4 To 12 Months	0.62	0.80
Total Debt Balance In 4 To 12 Months	5862.21	7760.76
Number Of Credit Accounts In Last 12 Months	3.36	3.10
Total Credit Debt Balance Over 12 Months	9902.90	12679.63
Total Debt Balance Of Payment Status Is 0 Accounts	3287.17	4155.74
Monthly Obligation Of Payment Status Is 0 Accounts	114.82	114.69
Number Of NPL Accounts	0.08	0.08
Number Of Closed Accounts Of Payment Status Is 1 Or 2	0.44	0.65
Number Of Closed Accounts In Last 12 Months	3.36	2.88
Number Of Closed Own Accounts In Last 12 Months	0.08	0.09
Number Of Closed Joint Accounts In Last 12 Months	0.12	0.08

Table 4-8: The Predictions of the QDA Model

Predictions		
	False	**True**
False	5530	4089
True	650	1008
Mean		
		0. 58

To build a model of QDA in R studio we used *qda* function. Similar as LDA, by using QDA function we obtained the group probabilities and the coefficients of each variables used in the model. Like other methods, when we make prediction on the test data and compare them with actual results we saw that the accuracy rate of the model in QDA was 58.00%. Using this information with this study we can say that quadratic discriminant analysis works a bit better than linear discriminant analysis, but does not work as good as logistic regression model. In this model, false positive rate is 49.03%, the false negative rate is 36.26%, true negative rate is 5.76% and the true positive rate is 8.93%. We can see that in QDA analysis, the accuracy rate of overdue is smaller.

In KNN analysis, by its nature, not like the other methods, training and forecasting are performed at the same time. We use the data we have used in the previous models with the parameters corresponding to the KNN method and run the *knn* function in the R library.

Table 4-9: The Predictions of the KNN Model

Predictions (k=1)		
	False	**True**
0	2166	2931
1	2931	3249
k	**Mean**	
1		0.48
2		0.48
3		0.48
5		0.47

k	Mean
10	0.46
11	0.46

As can be seen above, different results were obtained for different *k* values when running KNN model. When *k*=15 the accurate rate was 45.73%, when *k*=10 the accurate rate was 46.27%, when *k*=5 the accurate was 47.08% and went on this way. In this case we can see that the greater the *k* value, the lower the accuracy rate. Therefore, the highest accuracy rate was obtained for *k*=1, which was 48.00%. Under those conditions it was obvious that KNN model has produced worse results than all of the other models which are logistic regression, LDA and QDA. It was also worse than the random guessing which is 50.00%. Finally, in the model, false positive rate is 19.20%, the false negative rate is 25.99%, true negative rate is 28.81% and the true positive rate is 28.93%. We can see that in KNN analysis, the accuracy rate of both overdue and normal cases are not so far from each other.

CONCLUSION

The default of loans given by banks is a critical issue due to the wide variety of its impacts. For each credit to be used, predicting whether the loan will default at the application stage will lead to significant gains according to the accuracy rate.

In this study, a data set has been prepared by combining NPL data within a specified period in a bank and the credit bureau data at the time of the credit application stage. Using this data set and supervised machine learning algorithms, models that predict the probability of a credit default were produced, and their results were compared. Four types of models were discussed: Logistic Regression, LDA, QDA, and KNN. The same data set was used for all the models. Half of the data set for each model was used for model testing, and a half was used for the model's test.

The accuracy rate of the estimations of the models produced in the study is given below.

- Logistic Regression = 58.30%
- Linear Discriminant Analysis (LDA) = 57.90%
- Quadratic Discriminant Analysis (QDA) = 58.00%
- K-Nearest Neighbors (KNN) = 48.00%

Accordingly, in the present scenario, the following results were obtained in summary with the mentioned data set:

- The accuracy of Logistic Regression, LDA and QDA models were very close to each other, but Logistic Regression had the most successful results.
- LDA and QDA models gave almost the same results.
- KNN gave the worst results with a significant difference of 10% from the other three models.

REFERENCES

- Addo, P.M., Guegan, D., and Hassani, B. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. *Risks* 2018, *6*, 38. DOI: 10.3390/risks6020038.
- Akbaba, S. Ö. (2005). *Karesel diskriminant analizi ve hata oranları tahmini* (Unpublished Master's Thesis). Ankara University. Ankara.
- Akın, Ç.E. (2017). *Logistic regression (classification) – #8.* Retrieved from (http://cagriemreakin.com/veri-bilimi/logistic-regression-classification-8.html), on (20.04.2019).
- Borkar, J.V. (2018). *Default risk using deep learning.* Retrieved from (https://towardsdatascience.com/default-risk-using-deep-learning-6924cdada04d), on (05.04.2019).
- Bozdemir, T. (2007). *Türk bankacılığının tarihsel gelişimi ve reel sektöre katkısına ilişkin bir araştırma.* (Unpublished Doctoral Thesis). Istanbul University. Istanbul.
- Emil, R. S. B., and Sivasankar, E. (2018). Risk Analysis in Electronic Payments and Settlement System Using Dimensionality Reduction Techniques. *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence) Cloud Computing, Data Science & Engineering (Confluence), 2018 8th International Conference on. :14-19 Jan, 2018*
- Finansal Göz. (2018). *Takibe dönüşüm oranı.* Retrieved from (https://www.finansalgoz.com/2018/12/takibe-donusum-oran.html?m=1), on (01.04.2019).
- Interpretation of R's output, (2014). *Interpretation of R's output for binomial regression.* Retrieved from (https://stats.stackexchange.com/questions/86351/interpretation-of-rs-output-for-binomial-regression), on (20.04.2019).

- Islam, Y. R., Eberle, W., and Ghafoor, S. K. (2018). Credit default mining using combined machine learning and heuristic approach. *ICDATA*. Jul2018. https://arxiv.org/abs/1807.01176.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). An introduction to statistical learning with applications in R. *Springer New York Heidelberg Dordrecht London*. ISSN 1431-875X ISBN 978-1-4614-7137-0 ISBN 978-1-4614-7138-7 (eBook) DOI 10.1007/978-1-4614-7138-7.
- Jia, H. (2018). *Bank loan default prediction with machine learning*. Retrieved from (https://medium.com/henry-jia/bank-loan-default-prediction-with-machine-learning-e9336d19dffa), on (02.04.2019).
- Khandani, A. E., Adlar J. K., and Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance* 34 (2010): 2767-2787
- Kruppa, J., Schwarz, A., Arminger, G., and Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*. Oct2013, Vol. 40 Issue 13, p5125-5131. 7p. DOI: 10.1016/j.eswa.2013.03.019.
- Kvamme, H., Sellereite, N., Aas, K., and Sjursen, S. (2018). Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications*. July 2018. DOI: 10.1016/j.eswa.2018.02.029.
- Prabhakaran, S. (2017). *Linear regression*. Retrieved from (http://r-statistics.co/Linear-Regression.html), on (20.04.2019).
- Şirin, E. (2016). *Sınıflandırma notları-2 (lojistik regresyon)*. Retrieved from (http://www.datascience.istanbul/2016/12/19/siniflandirma-lojistik-regresyon-notlari-2/), on (15.04.2019).
- Son, Y., Byun, H., and Lee, J. (2016). Nonparametric machine learning models for predicting the credit default swaps: An empirical study. *Expert Systems with Applications*. Oct2016, Vol. 58, p210-220. DOI: 10.1016/j.eswa.2016.03.049.
- Turkson, R. E., Baagyere, E. Y., and Wenya, G. E. (2016). A machine learning approach for predicting bank credit worthiness. *2016 Third International*

Conference on Artificial Intelligence and Pattern Recognition (AIPR) Artificial Intelligence and Pattern Recognition (AIPR), International Conference on. :1-7 Sep, 2016.

- Ulgen, E. K. (2017). *Makine öğrenimi bölüm-2 (k-en yakın komşuluk).* Retrieved from (https://medium.com/@k.ulgen90/makine-öğrenimi-bölüm-2-6d6d120a18e1), on (18.04.2019).
- Unal, E. (2018). *Bayes Teoremi.* Retrieved from (https://medium.com/@enginunal/bayes-teoremi-431543ad9a59), on (17.04.2019).
- Uzun, E. (2016). *Sınıflandırma.* Retrieved from (https://www.e-adys.com/makine_ogrenmesi/04-makine-ogrenmesi-siniflandirma/), on (10.04.2019).
- Vanneschi, L., Horn, D. M., Castelli, M., and Popovic, A. (2018). An artificial intelligence system for predicting customer default in e-commerce . *Expert Systems with Applications*. Aug2018, Vol.104, pp. 1-21. DOI: 10.1016/j.eswa.2018.03.025
- Wang, C., Han D., Liu, Q., and Luo, S. (2018). A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism LSTM. *IEEE Access*. Dec2018, Vol.7, pp. 2161-2168. DOI: 10.1109/ACCESS.2018.2887138
- Yetiz, F. (2016). Bankacılığın Doğuşu ve Türk Bankacılık Sistemi. *Niğde Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, Nisan 2016; 9(2).*

APPENDIX

Table A-1: The Description of the Variables Used in the Program

Variable	Description
Overdue	Overdue
Score	Credit Bureau Score
MonthlyCommitmentMainOrJoint	Monthly Liability
TotOutstBalAllRecsRetrieved	Total Debt Balance
NumberOfAccountsL3M	Number Of Accounts In Last 3 Months
TotBalanceExclMortgagesL3M	Total Debt Balance In Last 3 Months
NumberOfAccountsL4_12M	Number Of Credit Accounts In 4 To 12 Months
TotBalanceExclMortgagesL4_12M	Total Debt Balance In 4 to 12 Months
NumberOfAccountsLOver12M	Number Of Credit Accounts In Last 12 Months
TotBalanceExclMortgagesLOver12M	Total Credit Debt Balance Over 12 Months
NumberOfAccounts	Number Of All Accounts
TotBalanceExclMortgages	Total Debt Balance Of All Accounts
TotBalanceForGoodAccAsNonMJ	Total Debt Balance Of Payment Status 0 Accounts
MontlyCommitmentForGoodAccAsNonMJ	Monthly Obligation Of Payment Status 0 Accounts
NumberOfD_LAccounts	Number Of NPL Accounts
TimeFromMostRecentDefaultForD_LAcc	Time Elapsed In NPL
NumberOfClosedAccountsWPS1_2	Number Of Closed Accounts Of Payment Status Is 1 Or 2
TimeFromMostRecentClosedWPS1_2	Time Elapsed Of Credit Accounts Of Payment Status Is 1 Or 2
NumberOfClosedAccountsWPS0_D_U_X	Number Of Closed Accounts In Last 12 Months
TimeFromMostRecentClosedWPS0_D_U_X	Number Of Closed Credit Accounts In Last 12 Months
NumberOfOwnClosedAccountsWPS0_D_U_X	Number Of Closed Own Accounts In Last 12 Months
NumberOfClosedAccountsWPS0_D_U_XNonMJ	Number Of Closed Joint Accounts In Last 12 Months
TimeFromMostRecentClosedWPS0_D_U_XNonMJ	Time Elapsed Of Closed Joint Accounts In Last 12 Months
WorstCurrentPaymentStatus	Worst Current Payment Status
WorstEverPaymentStatus	Worst Ever Payment Status
WorstPaymentStatusL3M	Worst Payment Status In Last 3 Months
WorstPymtStatusL4_12M1	Worst Payment Status Last 6 Months
WorstPymtStatusL7_12M2	Worst Payment Status In Last 7 To 12 Months
WorstPymtStatusLOver12M1	Worst Credit Payment Status Over 12 Months
WorstPymtStatus1	Worst Payment Status Of All Accounts

The Codes of the Application

The following codes were executed to load the data set into memory and show the column counts and total row counts.

```
> data=read.csv2("Data_ENG.csv", sep = ";", dec=".")
> dim(data)
```

The below codes were executed to divide data set into two categories those are training and test data sets.

```
> Data$Overdue<-!as.logical(Data$Overdue)
> TrainData<-Data[1:11277,]
> TestData<-Data[11278:nrow(Data),]
```

To construct the logistic regression model for all variables in the data set, the following codes were executed.

```
> glm.fit <- glm(as.factor(Overdue) ~
Score +
MonthlyCommitmentMainOrJoint +
as.character(WorstCurrentPaymentStatus) +
as.character(WorstEverPaymentStatus) +
TotOutstBalAllRecsRetrieved +
NumberOfAccountsL3M +
TotBalanceExclMortgagesL3M +
as.character(WorstPaymentStatusL3M) +
NumberOfAccountsL4_12M +
TotBalanceExclMortgagesL4_12M +
as.character(WorstPymtStatusL4_12M1) +
as.character(WorstPymtStatusL7_12M2) +
NumberOfAccountsLOver12M +
TotBalanceExclMortgagesLOver12M +
as.character(WorstPymtStatusLOver12M1) +
NumberOfAccounts +
TotBalanceExclMortgages +
as.character(WorstPymtStatus1) +
TotBalanceForGoodAccAsNonMJ +
MontlyCommitmentForGoodAccAsNonMJ +
NumberOfD_LAccounts +
TimeFromMostRecentDefaultForD_LAcc +
NumberOfClosedAccountsWPS1_2 +
TimeFromMostRecentClosedWPS1_2 +
NumberOfClosedAccountsWPS0_D_U_X +
TimeFromMostRecentClosedWPS0_D_U_X +
NumberOfOwnClosedAccountsWPS0_D_U_X +
NumberOfClosedAccountsWPS0_D_U_XNonMJ +
TimeFromMostRecentClosedWPS0_D_U_XNonMJ,
```

```
    family = binomial(link = "logit"), data = TrainData)

> summary(glm.fit)
```

To build the logistic regression model the following codes were executed.

```
> glm.fit <- glm(as.factor(Overdue) ~
     Score +
    MonthlyCommitmentMainOrJoint  +
    TotOutstBalAllRecsRetrieved +
    NumberOfAccountsL3M +
    TotBalanceExclMortgagesL3M +
    NumberOfAccountsL4_12M +
    TotBalanceExclMortgagesL4_12M +
    NumberOfAccountsLOver12M +
    TotBalanceExclMortgagesLOver12M +
    TotBalanceForGoodAccAsNonMJ +
    MontlyCommitmentForGoodAccAsNonMJ +
    NumberOfD_LAccounts +
    NumberOfClosedAccountsWPS1_2 +
    NumberOfClosedAccountsWPS0_D_U_X +
    NumberOfOwnClosedAccountsWPS0_D_U_X +
    NumberOfClosedAccountsWPS0_D_U_XNonMJ,
         family = binomial(link = "logit"), data = TrainData)

> summary(glm.fit)
```

```
> glm.probs=predict(glm.fit, TestData, type="response")
> glm.pred=rep(0,nrow(TestData))
> glm.pred[glm.probs >0.5]=1
> table(glm.pred, as.factor(as.numeric(TestData$Overdue)))
> mean(glm.pred==as.factor(as.numeric(TestData$Overdue)))
```

To construct the LDA model the below codes were executed.

```
> lda.fit=lda(as.factor(Overdue)~
    Score +
    MonthlyCommitmentMainOrJoint  +
    TotOutstBalAllRecsRetrieved +
    NumberOfAccountsL3M +
    TotBalanceExclMortgagesL3M +
    NumberOfAccountsL4_12M +
    TotBalanceExclMortgagesL4_12M +
    NumberOfAccountsLOver12M +
    TotBalanceExclMortgagesLOver12M +
    TotBalanceForGoodAccAsNonMJ +
    MontlyCommitmentForGoodAccAsNonMJ +
    NumberOfD_LAccounts +
    NumberOfClosedAccountsWPS1_2 +
    NumberOfClosedAccountsWPS0_D_U_X +
    NumberOfOwnClosedAccountsWPS0_D_U_X +
    NumberOfClosedAccountsWPS0_D_U_XNonMJ,
        data = TrainData)
```

```
> lda.fit
```

```
> lda.pred=predict(lda.fit , TestData)
> lda.class=lda.pred$class
> table(lda.class ,TestData$Overdue)
> mean(lda.class==TestData$Overdue)
```

To builds the QDA model the following codes were executed.

```
> qda.fit=qda(as.factor(Overdue)~
      Score +
      MonthlyCommitmentMainOrJoint  +
      TotOutstBalAllRecsRetrieved +
      NumberOfAccountsL3M +
      TotBalanceExclMortgagesL3M +
      NumberOfAccountsL4_12M +
      TotBalanceExclMortgagesL4_12M +
      NumberOfAccountsLOver12M +
      TotBalanceExclMortgagesLOver12M +
      TotBalanceForGoodAccAsNonMJ +
      MontlyCommitmentForGoodAccAsNonMJ +
      NumberOfD_LAccounts +
      NumberOfClosedAccountsWPS1_2 +
      NumberOfClosedAccountsWPS0_D_U_X +
      NumberOfOwnClosedAccountsWPS0_D_U_X +
      NumberOfClosedAccountsWPS0_D_U_XNonMJ,
          data = TrainData)

> qda.fit
```

```
> qda.pred=predict(qda.fit , TestData)
> qda.class=qda.pred$class
> table(qda.class, TestData$Overdue)
> mean(qda.class==TestData$Overdue)
```

To construct the KNN model the following codes were executed.

```
library(class)
> Train.X=cbind(Score,
 MonthlyCommitmentMainOrJoint,
 TotOutstBalAllRecsRetrieved,
 NumberOfAccountsL3M,
 TotBalanceExclMortgagesL3M,
 NumberOfAccountsL4_12M,
 TotBalanceExclMortgagesL4_12M,
 NumberOfAccountsLOver12M,
 TotBalanceExclMortgagesLOver12M,
 TotBalanceForGoodAccAsNonMJ,
 MontlyCommitmentForGoodAccAsNonMJ,
 NumberOfD_LAccounts,
 NumberOfClosedAccountsWPS1_2,
 NumberOfClosedAccountsWPS0_D_U_X,
 NumberOfOwnClosedAccountsWPS0_D_U_X,
 NumberOfClosedAccountsWPS0_D_U_XNonMJ)[1:11277,]
```

```
> Test.X=cbind(Score,
  MonthlyCommitmentMainOrJoint,
  TotOutstBalAllRecsRetrieved,
  NumberOfAccountsL3M,
  TotBalanceExclMortgagesL3M,
  NumberOfAccountsL4_12M,
  TotBalanceExclMortgagesL4_12M,
  NumberOfAccountsLOver12M,
  TotBalanceExclMortgagesLOver12M,
  TotBalanceForGoodAccAsNonMJ,
  MontlyCommitmentForGoodAccAsNonMJ,
  NumberOfD_LAccounts,
  NumberOfClosedAccountsWPS1_2,
  NumberOfClosedAccountsWPS0_D_U_X,
  NumberOfOwnClosedAccountsWPS0_D_U_X,
  NumberOfClosedAccountsWPS0_D_U_XNonMJ)[11278:nrow(Data),]

> Train.Overdue=Overdue[1:11277]
> set.seed(1)
> knn.pred=knn(Train.X, Test.X, Train.Overdue ,k=1)
> table(knn.pred ,TestData$Overdue)

> mean(knn.pred==as.numeric(TestData$Overdue))

> knn.pred=knn(Train.X, Test.X, Train.Overdue ,k=2)
> mean(knn.pred==as.numeric(TestData$Overdue))

> knn.pred=knn(Train.X, Test.X, Train.Overdue ,k=3)
> mean(knn.pred==as.numeric(TestData$Overdue))

> knn.pred=knn(Train.X, Test.X, Train.Overdue ,k=5)
> mean(knn.pred==as.numeric(TestData$Overdue))

> knn.pred=knn(Train.X, Test.X, Train.Overdue ,k=10)
> mean(knn.pred==as.numeric(TestData$Overdue))

> knn.pred=knn(Train.X, Test.X, Train.Overdue ,k=15)
> mean(knn.pred==as.numeric(TestData$Overdue))
```