KHAZAR UNIVERSITY

Faculty: Engineering and Applied Sciences

Department: Computer Sciences

Specialty: Mathematics of Computer

MASTER THESIS

Theme: Usage of Artificial Neural Network and Support Vector Machine model for classification of Credit Scores.

Master Student: Lala Gafarova

Supervisor: PhD Associate Professor. Nuru Safarov

BAKU – 2017

CONTENTS

ABREVIATIONS

CS	: Credit Scoring
ANN	: Artificial Neural Network
SVM	: Support Vector Machine
BS	: Behavior Score
RF	: Random Forest
LR	: Logistic Regression
DT	: Decision Tree
PE	: Processor Elements
LVQ	: Linear Vector Quantization
SOM	: Self Organizing Map
MLP	: Multi Layer Perceptron
AUROC	: (AUC – ROC)
KS	: Kolmogorov Smirnov
Gini	: Gini Coefficient

LIST OF FIGURES

LIST OF TABLES

Abstract

In the emerging banking sector, credit is an important product. The decision to give or not to give credit to the customer is a decision that should be taken carefully from the point of view of the bank and as credit requests increase, the evaluation of applicants becomes even more complex. Decisions may be subjective because the evaluators may consider different criteria. In this case, various statistical and non-statistical techniques are used to answer both the increasing number of applications and to make objective decisions without subjective criteria.

In this study, we tried to distinguish between good and bad customers with twenty variables of the german loan data set, and the results of the applications are compared with one another.

Some non-statistical techniques were used in the study: Artificial Neural Network and Support Vector Machine and the practice of these techniques are discussed. Practice presented as theoretical information without their practice are Logistic Regression, Random Forest, Decision Tree and K-Nearest Neighbor Approach. Practice related to these techniques will reprieve to work in the future.

Since there are various advantages and disadvantages in the implementation of models, it can be said that the model with the highest prediction success, according to the data set used is Support Vector Machine -Vanilladot Kernel method.

Keywords : Credit Scoring, Artificial Neural Networks, Support Vector Machines.

Introduction

Credit scores, one of the first developed financial risk management issues, is one of the most successful statistical and operational models used in finance and banking and credit scoring analysts are needed more and more over time. The credit scoring, which is also affected by the increase in credit cards, automatically calculates the risk and the models that make up this account are able to expand the card volumes of credit card issuing banks more easily based on the data in their hand.

Credit scoring has provided extensive user support in economic environments since 1995. That year, major US mortgage agencies, Fannie Mae and Freddie Mac, advised lenders to use FICO score ratings. two agencies had more than two thirds of the US mortgage market, it is not difficult to calculate the effect of this recommendation.

Credit scoring has provided extensive user support in economic environments since 1995. That year,to major US mortgage agencies, Fannie Mae and Freddie Mac, advised lenders to use FICO skor ratings. two agencies had more than two thirds of the US mortgage market , it is not difficult to calculate the effect of this recommendation.

Contain in the first part of this study, the definition of credit scoring, the history and the development of scorecards. The key points of the model, such as sampling selection, data sources, separation of customers as good or bad, and classification of required data in credit card application form, are considered in this chapter. The data set used in the implementation phase contains the information of a bank's customers and assuming that the theoretical background described for the preparation of the data was used to prepare the data beforehand, no changes were made to the data.

In the second part of the study, information on the theoretical backgrounds of some non-statistical techniques used in the classification of customer data in credit scoring is given. Logistic Regression, Random Forest, Conditional Inference Trees, Bayesian Network was examined during non-statistical techniques.

Support Vector Machine (SVM) algorithms such as VanillaDot Kernel and Gaussian RBF Kernel models, as well as techniques such as an Artificial Neural Network, are included in the third chapter. Firstly approaches theoretical knowledge of these non-

static techniques and then given place practically how these techniques are used to classify customers as well-bad in credit scoring.

Finally, models applied at the end are evaluated together and the comparison between the techniques is given.

1. DEFINITION and CREATION of CREDIT SCORING

1.1 What is Credit Scoring?

Calculating the probability that a customer will not be able to repay loans on a loan application is called credit scoring.

CS (credit scoring) is the decision models and techniques that help the lender give the consumer credit. These techniques will help to make decisions about whom and will be given credit how much and what kind of operational strategies will increase the profitability of the borrower.

By credit score refusing to give credit to high-risk customers will reduce the potential harm to the financial institution, will increase the profit by giving loans to low-risk customers, therewithal it will also reduce the inconvenience caused by customers who cannot reimburse for debt.

CS techniques dissipate the risk of giving credit to a particular customer. Credit worthiness is not a personal attribute such as weight, length, or income. It shows the relationship of the debt with the lender and reflects the conditions of both parties and shows the possible future economic scenarios in terms of the lender. Thus, lenders class according to whether an individual is worthy or not worthy of a credit. The biggest long-term danger of CS is that this process is stopped, and some customers are borrowing from all lenders but some customers never get it. Defining a customer as not suitable for a credit leads to reaction. It is best for creditors to show the truth. There is always a risk of non-repayment of debts received, lenders should never forget that.

A lender should decide two kinds: to decide whether to give credit to a new application and to determine how to act against existing customers who want to increase their credit limits. While the techniques that describe the first type of question are called credit scoring, the second type of decision is called behavior scoring.

Whichever technique is used, it is important point in both decision types: it is necessary to sample a lot of detailed information and credit history information from previous clients. All techniques use sampling to describe the relationships between the characteristics of the customers and to make a good-bad distinction based on their

past history. Most of the techniques from a scorecard, on which features a score is given and the sum of these scores allows you to determine whether giving a credit to a person has a bad outcome. Some techniques, such as score cards, directly understand that the customer is not good at giving credit, and these techniques work in parallel with credit and behavioral scoring.

Although scoring is generally used in credit terms, it has been used in many different areas, especially recently time. It is especially useful in direct and other marketing techniques to determine the target customer group. In the finance and retail sectors, many companies need to apply scoring techniques to store data. Similarly, data mining and highly sophisticated information systems are preparing successfully scoring applications.

1.2 History of Credit Scoring

Although the credit history is based on 5000 years, credit scoring is only used for 50 years. KS is the most important way of describing different groups in a main group, based on their interrelated properties. Fisher first introduced a statistical approach to solve such problems in 1936. He tried to distinguish two species of a flower named Iris according to their physical size and structure. In 1941, Durand tried to classify good and bad debts for the first time using the same techniques.

In the 1930s, mail-ordering companies developed a numerical scoring system to eliminate the adverse effects of credit decisions. Along with the beginning of the Second World War, all lenders and postal sales companies suffer from difficulties in credit management. By going to the troops of credit analysts, the number of specialists in this sector has decreased considerably. Thus, companies want their analysts to write down the rules they apply to when deciding whom to give credit (Johnson, 1992). Some of these have led to the establishment of digital scoring systems, while others have created the conditions that make up the satisfaction of needs. These rules have thus led even non-experts to take credit decisions (Thomas et al., 2002).

Soon after the war ended, automatic landing systems, statistical classification models, began to be used in lending decisions. The first consulting firm on this subject was founded in San Francisco by Bill Fair and Earl Isaac in the early 1950s and clients are financial houses, retailers and mail-order companies (Thomas VD., 2002).

With the introduction of credit cards at the end of the 1960s, CS has become very useful for credit card issuers. With the use of computers, this technique can evaluate the application of many people every day. Thus, companies have seen CS as a very good predictor and decision-making tool. CS is a legal technique used in lending with the Equal Loan Opportunity Act introduced in the US in 1975 and 1976. Thus, in the next 25 years KS analysis has become a rapidly growing profession. It has become very popular, especially in the United States and the UK (Thomas et al., 2002).

In the 1980s, with the success of CS's credit cards, the banks began to use this technique in other products such as personal loans, home loans and small investor loans. In the 1990s, the use of scorecards in direct marketing has led to increased returns to advertising campaigns. Developments in the computer have allowed other techniques to be used to generate score cards. In the 1980s, two of the most important techniques used today, logistic regression and linear programming techniques, began to be used. Recently, artificial intelligence and neural network techniques have been used for testing purposes.

Today, the purpose, function is based on how customers can earn more than such customers, rather than to minimize their debt repayments. Significant improvements were made in the risk estimates of customers who did not pay the debt with score cards. Scorecards "How often will customers use a new product and direct sales?", "How often will customers use a product?", "How much time will they use the old product when a new product emerges?", "Will customers submit another loan? "How will customers be able to pay off their debts, and what will be their attitude towards them?" And "How to avoid fraud on applicants" It helps to find answers to questions such as.

1.3 Credit Scoring and Data Mining

Data mining is a data analysis and research technique to identify meaningful relationships and constructs in data. Similar to mining, it is tried to determine where and how to find the necessary data in this technique. In recent years, companies, especially banks and retailers, have been conceptualizing the value of identifying information about their customers. With electronic fund transfer and widespread use of loyalty cards, such companies can easily gather information about their customers. Computer technology also facilitates the analysis of large quantities of collecting

data. Increasing competition, substitute products and easy communication channels such as the internet make customers easily relocate. Thus, understanding and analyzing customer behavior is of great importance. For this reason, companies spend a huge amount of money to create data warehouses and use techniques such as data mining.

When you look at the main techniques of data mining, it is seen that this technique provides very successful results in credit scoring. Basic data mining techniques include data summary, variable reduction, observation clustering, prediction and explanation. Standard descriptive statistics such as frequency, median, variance, and cross tabulation is used to summarize the data. It is also very useful for categorizing continuous variables in discrete classes. Descriptive statistics are rough classification techniques that are widely used in CS. Determining which variables is most important and removing unnecessary ones from the analysis is also used in data mining applications as well as frequently used techniques in CS applications. It is another data mining tool to segment customers into groups according to different products they purchase or other features. KS also creates different groups according to the behaviors of the customers and a separate score card is prepared for each group. This idea implies the segmentation of the sub-masses so that a score card profile is created for each sub-mass.

In fact, techniques developed for use in CS, such as estimating which client will use, which financial instrument next year, are also very important for data mining. In fact, segmentation analysis used in data mining is used to show segments with certain types of behavior. Thus, data mining is an indispensable technique and technology for KS, and it needs to be applied to a wider area. Those who use data mining in combination with KS will achieve much more success and development in their work in order to prevent mistakes in implementation, to prevent deficiencies and to apply them in other areas.

1.4 Sampling Selection

All methods related to Credit Scoring (DS) and Behavioral Score (DS) require customer history and their stories to improve the scoring system. There are two issues to consider when choosing a sample. First; the sampling should represent the applicants in the future as possible. Secondly; the sample should include enough

information to reflect that the payment habits are good or bad. The best example of this is the database where the information of borrowed persons is included in the most recent possible period. In the application scoring these last 12 months and in the behavior scoring the last 18-24 months (Thomas et al., 2002).

Another point to note when choosing a sample is; how much will be the sample size and which elements of the good-bad loans will be separated. Should the good customer-bad customer ratio in the sample is equal or should it be represented as it is in the main mass? When the well-to-bad ratio is determined according to the ratio in the mainstream, it is generally assumed that this ratio is 50:50 since the data are not available in the sample until the bad loans are announced. If the distributions of the good-bad variables in the sample are not the same, the results must be corrected to obtain the sample that will permit it.

For Lewis sampling size and good-to-bad credit ratio; 1500 was good and 1500 was bad enough (Lewis, 1992). In practice, much larger samples are used.

If the sampling is randomly selected from the existing mainstream, you need to make sure that it is really random. If one out of every 10 bones in the list of main mass is selected, this will be 10% of the sample. However, it is important to make sure that the first things to be done when it is necessary to go to the branches randomly select rural and urban areas. Or, at the time the sample was selected, it should be checked whether certain products were marketed or not addressed to specific masses. For example, when a certain month is taken as sampling, one product may be marketed in that month, and if this product appeal to young people, it is inevitable that the sample rate in the sample will be higher.

1.5 Good Customer-Bad Customer Description

In the development of the scorecard, one of the stages is how to do good customer-bad customer classification. Poor identification of some customers does not mean that all other customers are good. There are at least 2 more choices besides customers being good-bad. The first one is "unidentifiable" and the second is "not worth watching".

Generally, those who suffered from 3 period problems among the payments without improvement of CS are described as problem loans. Those who cannot be defined

are those who experience problems for 2 terms and those who do not return for 3 years.

Whatever the good-bad distinction is made, the KS technique is not affected. It is necessary to remove the "unidentifiable" and "insignificant" from the sample and develop a scorecard only for good-bad classification. Of course different classification of good-bad will develop different scorecard results. Another problem stems from the fact that the bad credit is defined as an extremity. In this case, the reliability of the model 8 may be shaken by poor credit (Thomas et al., 2002).

2. CLASSIFICATION METHODS USED IN CREDIT SCORE

2.1 Traditional Approach to Credit Scores

In the traditional approach; the lenders decide on the 5C of the lender. These; character, capacity, capital, collateral and external factors. In this approach, the experience of the person using the credit, taking advantage of his / her past knowledge, and his / her views on the future situation of the person who will use the credit are important. With the credit scoring methods, the risks of the person or company that will use the credit are reduced by certain models.

The greatest benefit of credit scoring methods arises when making decisions that will affect customers. In the decision-making process, the person or institution that will give the loan will be in lending action according to different scenarios and policies; acceptance / non-acceptance, loan interest, duration of the loan. The credit scoring approach leaves place to traditional methods by credit specialists in situations where there is not enough credit for scoring and when the likely profit is too high.

Credit scores have different names depending on their usage. These;

- Application Score: In this scoring technique for newly started customers, the scorecards are being used within the data obtained from past agreements and credit facilities of the customers.
- Behavior Score: Movements in existing customers 'accounts are used to identify gore customers' behaviors and to set limits and allow actions.
- Collection Score: A scoring method used for the collection process.
- Customer Score: It is used to analyze the customer behavior of many accounts and to manage customer's accounts and cross-sell.
- Office Score: A scoring used by the credit bureaus to estimate the office score, delays and bankruptcies.

Scoring studies have several common characteristics in spite of their different nomenclature for their purposes.

The customer uses internal and external data as data.

1. Customer behavior has 4 forms. These; risk, income, reaction and incentive.

2. All these scores can be used in marketing, new business processes, collections, advertising campaigns (Anderson, 2007).

2.2 Classification Methods in Credit Scores

Credit scoring methods use past experiences to predict whether the situation will be good or bad in the future, using predictive methods (algorithms). Despite the use of different algorithms and methods in credit scoring approaches, the most accepted approach is regression, which is a statistical method. Various methods have been used for the development of credit scoring methods to the development of the scorecards. These methods are now divided into parametric and non-parametric. While parametric methods accept some assumptions on the data used, there is no assumption of the data used in non-parametric methods (Anderson, 2007).

- Parametric methods;
    1. Logistic regression
- Non-Parametric methods;
    1. Random Forest
    2. Decision Tree
    3. K nearest neighborhood
    4. Support vector machine
    5. Artificial neural networks

2.2.1 Logistic Regression

Logistic regression (LR) is a widely used model during the development period of credit scoring models. The reason for this is that the target variables in the credit scoring model are binary. Logistic regression uses the maximum likelihood estimation process (Anderson, 2007).

This process;

(1) transforming dependent variables into a logarithmic function,

(2) which coefficients should be,

(3) the determination of coefficient changes and the maximization of the logarithmic likelihood.

$$\ln\left(\frac{p(Good)}{1} - p(Good)\right) = b_0 + b_1 x_1 + b_2 b_2 + \cdots + b_k b_k + e \ (\mathbf{1})$$

The assumptions required to use logistic regression;

1. categorical target variable,
2. linear relationship in logarithmic odds functions,
3. Independent error term,
4. unrelated estimators,
5. the appropriate variables.

Today, credit scoring models are developed and logistic regression is accepted as the most important method. The reasons for this are;

1. design of binary outputs for finalization,
2. the probability of the results remaining between 0 and 1,
3. is to be able to make highly accurate probability estimates with the given information.

Compared with logistic regression, discriminant analysis and linear regression methods, it has the following advantages;

1. Logistic regression requires no assumption of normal distribution of arguments.
2. Logistic regression can work well if there are large differences between group sizes.
3. The models presented by many people are quite understandable.b

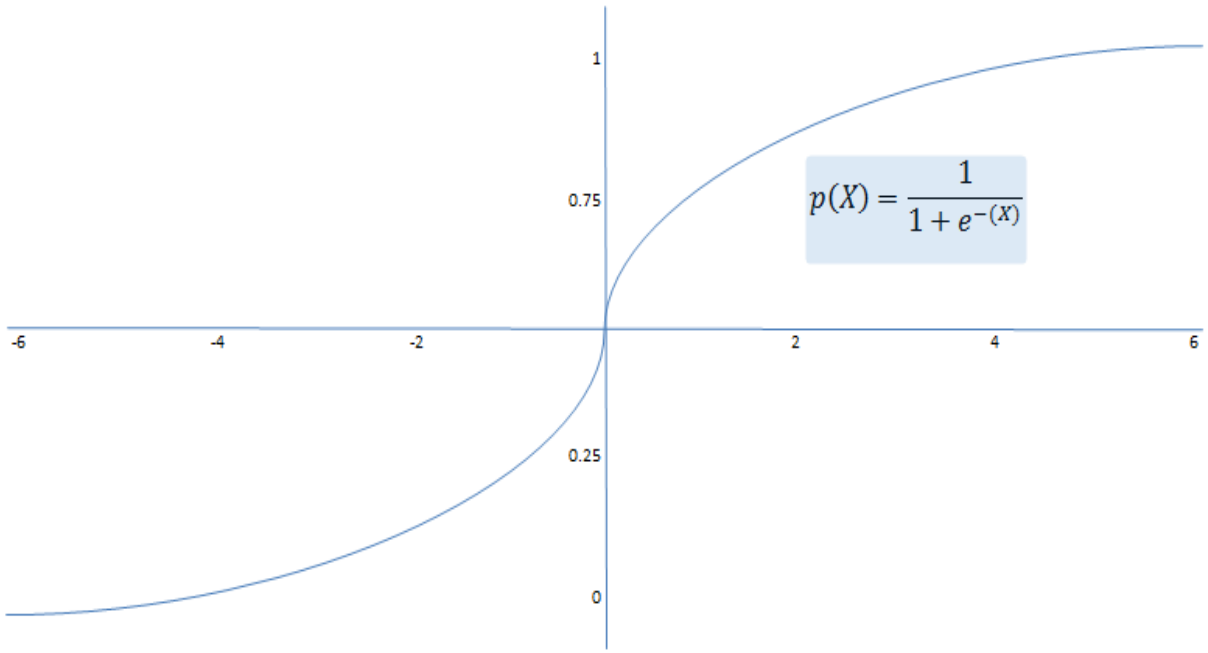

Figure 1. Standard Logistics Regression chart

2.2.2 Random Forest

Random forests or random decision forests are a community learning method for working classifications, regression and other tasks by creating a large number of decision trees at the time of the training and by creating a class (classifying) mode or average estimation (regression) of classes. Unstable decision forests correct the habit of over-fitting decision trees to the educational setting.

The first algorithm for random decision forests was created by Tin Kam Ho using the random subspace method, which is a way of applying a "stochastic discrimination" approach to the classification proposed by Eugene Kleinberg in Ho's formulation.

The Random Forest (RF) is a community classifier that uses the paging mechanism. RF consists of a series of CART classifiers. At each node of a tree, only a small subset of features is selected for the partition; this allows the algorithm to quickly classify for high-dimensional data. The number of randomly selected features (try) should be determined in each section. The default value is secret (p) for the classification of the number of properties of p. The separation criterion is the Gini index, as shown in Equation (1).

$$gini(N) = \frac{1}{2}\left(\sum_j p(\omega_j)^2\right) \quad (2)$$

2.2.3 Decision Tree

Decision tree (DT) is a graphical tool that shows the possible consequences of events to decision makers. Decision trees are also used in classification problems and estimation problems. One of the advanced methods is data analysis.

As an example of decision trees, the following graph can be given;

When this decision is considered downward, the top-tier age is defined as the basic needs of living with the family and becoming professional. Intermediate boxes; age, children, possession of the house, and intermediate nodes and bottom boxes are defined as end nodes. When the decision tree is over, the scores obtained from the finish node are used. A sample decision tree structure is shown in Figure 2.

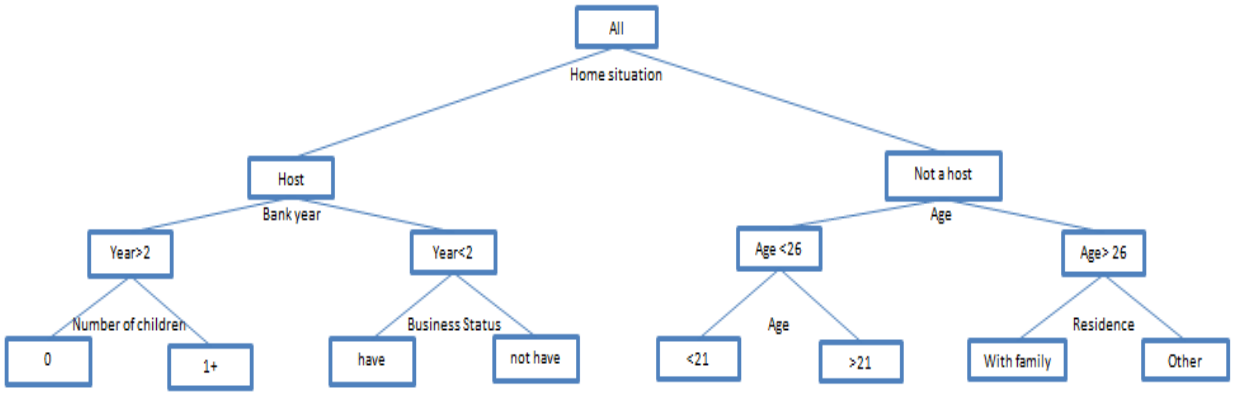

Figure 2. A schematic representation of a sample decision tree structure

Decision trees have several disadvantages and advantages over other techniques;

1. In the case of a set of rules, this technique can identify high or low risk categories that can be quickly and easily understood.

2. It is very simple to use with computer programs. The process ends with the selection of the variables and the creation of the decision tree structure. But it is not a modeling process that has a lot of flexibility. There is not enough

information on how to make a lot of changes in the variables and affect the result.

3. The use of small data sets may lead to doubts about the reliability of the results. In order not to worry about the reliability of the model, it is necessary to work with large data sets.

4. It is quite simple to examine the results in simple decision trees. But it is rather difficult to examine the results when the decision tree structure is complicated (Anderson, 2007).

2.2.4 K-Nearest Neighbour Approach

The nearest neighbors technique is a standard, non-parametric approach to the classification problem developed by Fix and Judges for the first time. This technique K - nearest neighbor approach was first applied by Chatterjee and Barcun, then by Henley and Hand. The logic on which this technique is based on choosing a distance in the application data space to measure how far apart any two applications are from each other. Sampled representation of past applicants is taken as standard. A new applicant is classified as good or bad on the basis of the well-bad ratio between close-up applicants (the nearest neighbor of the new application) in the sample (Thomas, 2002).

To implement this approach, three parameters are needed: the distance, how many reference counts (*k*) constitute the nearest neighbors set, and what should be the best rate of application for an application to be classified as good. Normally, if the majority of the neighbors is good, the referral is classified as good. Otherwise the application is classified as bad. Define the average existing cost *M*, and the average loss profit *K* of rejecting good. If at least *M / M + K* of the nearest neighbors is good, a new reference is classified as good. If the proportion of neighbors who are likely to be good at a new application, this criterion will minimize the expected loss.

The choice of distance is very important. Fukanaga and Flick defined a general distance;

$$d(x_1, x_2) = (x_1 - x_2)A\ (x_a)((x_1 - x_2)^t)^{\frac{1}{2}}\ (\mathbf{3})$$

$A\ (x)$, is a $p \times p$ symmetric end positive matrix. If connected to *x, A (x)* is called local distance, if it is independent of *x* it is called global distance. The lack of local distance often takes into account the characteristics of the unsuitable test set. For this reason, many researchers focus on global distance. The most detailed application of the nearest neighbors approaches in CS was done by Henley and Hand. With this technique, the focus is on a mixture of Euclidean length and good length, which best distinguishes evil. If w; the p-dimensional direction vector is the distance expression of Henley and Hand;

$$d(x_1, x_2) = \{(x_1 - x_2)^T (I + D\ w^T)(x_1 - x_2)\}^{\frac{1}{2}} \ (\mathbf{4})$$

Although the CS is not as frequently used as linear and logistic regression approaches, the nearest neighbors have some important features for real applications. It is very easy to update the experiment set by dynamically adding new events and it can be easily removed from the sample when it is known that the addition is good or bad. Finding a good distance the first time is almost equivalent to creating a scorecard with the regression technique. Thus, most practitioners prefer to stop at this point and use a traditional scorecard. When we compare it with the classification tree approach, the nearest neighbor approach does not produce a score for the future of each applicant. They identify a balance point for practitioners and they enable them to understand what the system actually does.

3.SUPPORT VECTOR MACHINE AND NEURAL NETWORK CREDIT SCORING CLASSIFICATION MODELS

3.1 Data Preprocessing

The R programming language was used to establish the models for classification in Credit Scores.

The German data set is used for modeling. The set consists of 21 columns (variable) and 1000 lines (data).

Variables- *Existing account status, Month period, Credit history, Aim,Amount of credit, Savings account / stock, Since then get a job, Installment rate, Personal status and sex, Other debtors / sureties, Place of residence, Estate, Age, Other plans of installments,Home, Number of existing loans in this bank,Job, The number of persons obliged to provide care,Phone, Foreign employee, Cost Matrix.*

The extensive article attached to this article is given in Table 1.

	Attribute	**Data Type**	**Value**	**Description**
1	Existing account status	qualitative	A11	<0
			A12	0 <= ... < 200
			A13	>= 200
			A14	no checking account
2	Month period	numerical		Duration in month
3	Credit history	qualitative	A30	no credits taken/all credits paid back duly
			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)
4	Aim	qualitative	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
			A410	others
5	Amount of credit	numerical		
6	Savings account / stock	qualitative	A61	<100
			A62	100 <= ... < 500
			A63	500 <= ... < 1000
			A64	>= 1000
			A65	unknown/ no savings account
7	Since then get a job	qualitative	A71	unemployed
			A72	< 1 year
			A73	1 <= ... < 4 years
			A74	4 <= ... < 7 years
			A75	>= 7 years
8	Installment rate	numerical		
9	Personal status and sex	qualitative	A91	male: divorced/separated
			A92	female: divorced/separated/married
			A93	male: single
			A94	male: married/widowed
			A95	female: single
10	Other	qualitative	A101	none

	debtors / sureties		A102	co-applicant
			A103	guarantor
11	Place of residence	numerical		
12	Estate	qualitative	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
13	Age	numerical		
14	Other plans of installments	qualitative	A141	bank
			A142	stores
			A143	none
15	Home	qualitative	A151	rent
			A152	own
			A153	for free
16	Number of existing loans in this bank	numerical		
17	Job	qualitative	A171	unemployed/ unskilled - non-resident
			A172	unskilled - resident
			A173	skilled employee / official
			A174	management/ self-employed/highly qualified employee/ officer
18	The number of persons obliged to provide care	numerical		
19	Phone	qualitative	A191	none
			A192	yes, registered under the customers name

20	Foreign employee	qualitative	A201	yes
			A202	no
21	Cost Matrix	numerical	1	good
			2	bad

Table 1. Information about the variables used in the Credit Scores

For modeling, *Month period, Amount of credit, Installment rate , Place of residence, Age, Number of existing loans in this bank, The number of persons obliged to provide care* variables convert to numeric types.

3.1.1 Analysis of Variables

There are 300 bad and 700 good customers in the dataset. The customer analysis is described with charts using R graph, while the following values are calculated for each change.

- Names- the value of the variable in the data set
- Good- good customer count
- Bad- bad customer count
- Good_pct- good customer count by percentage
- Bad_pct- bad customer count by percentage
- Total-total customer count
- Total_Pct- total customer count by percentage
- Bad_Rate-Bad rate or response rate
- grp_score-score for each group
- WOE-Weight of Evidence for each group
- IV-Information value for each group
- Efficiency- Efficiency for each group

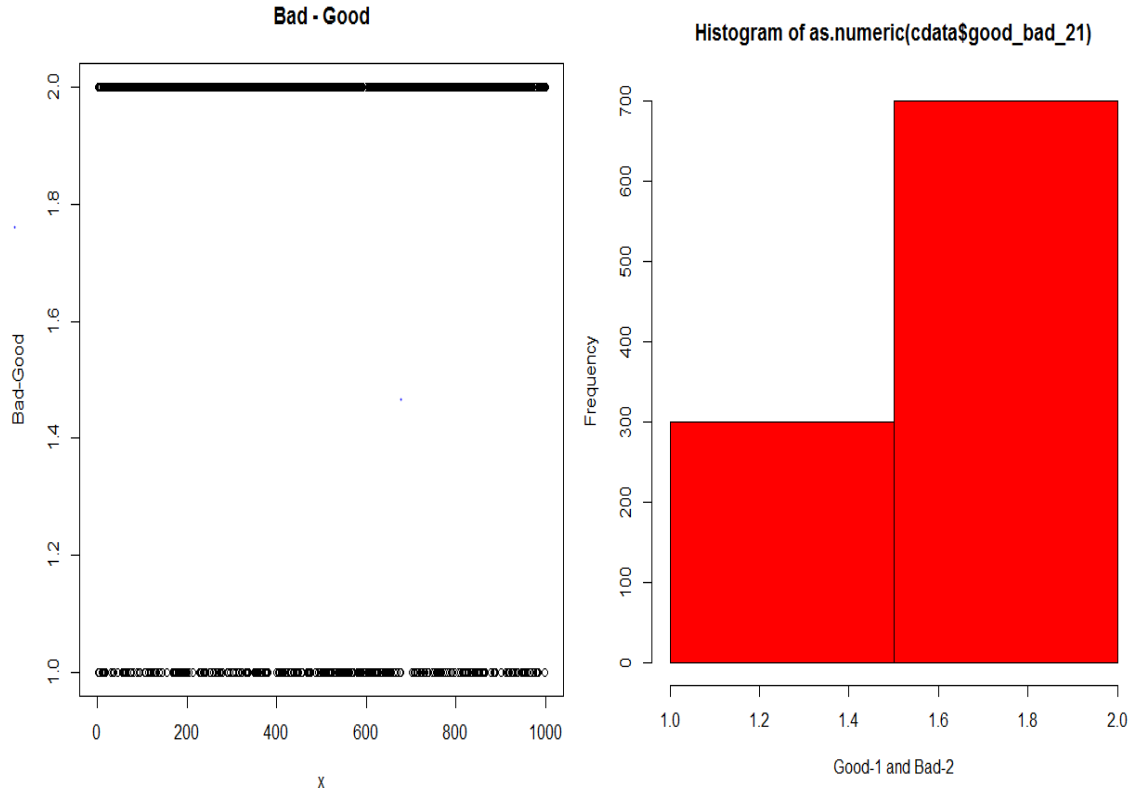

Figure 3. Division of customer data for analysis and representation by histogram chart

Variable- Existing account status :

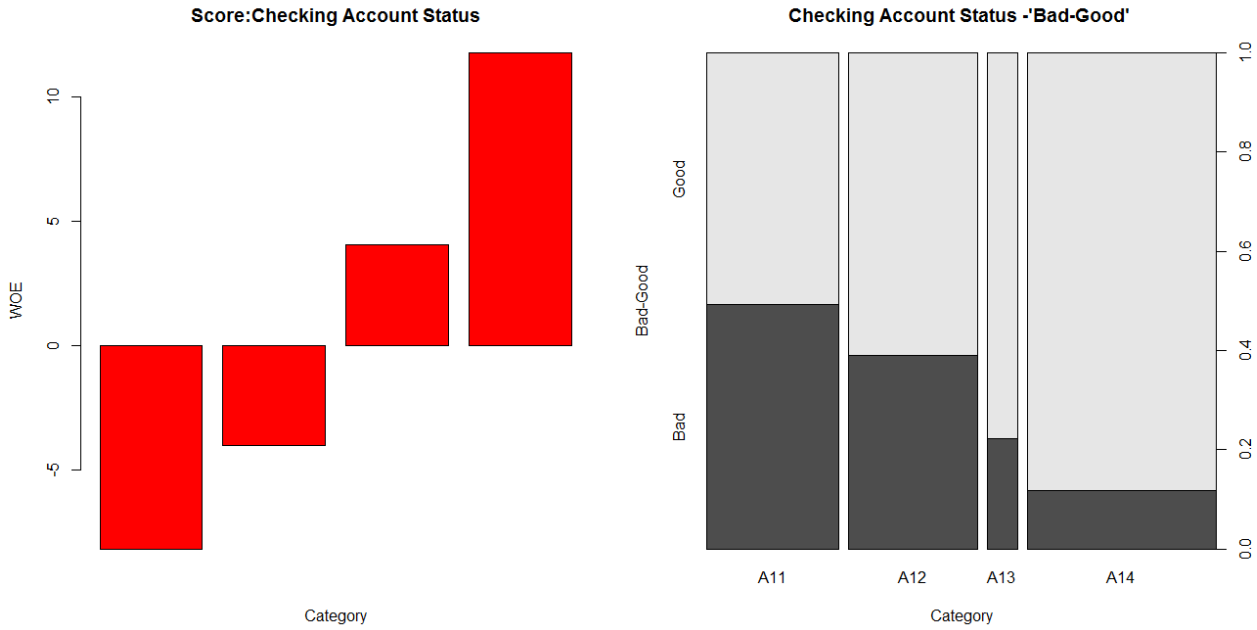

Figure 4. Variable- Existing account status

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
A11	139	135	19.86	45.00	274	27.4	49.27	3.06	-8.18	20.56452	12.570
A12	164	105	23.43	35.00	269	26.9	39.03	4.01	-4.01	4.63957	5.785
A13	49	14	7.00	4.67	63	6.3	22.22	6.00	4.05	0.94365	1.165
A14	348	46	49.71	15.33	394	39.4	11.68	7.64	11.76	40.43088	17.190

Table 2. Value of Variable- Existing account status

Variable- Month period :

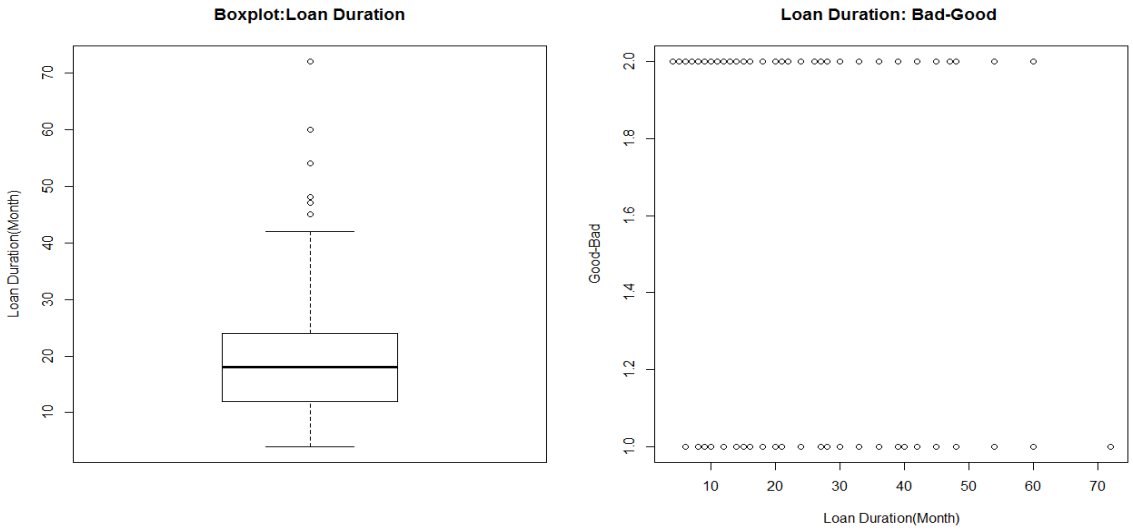

Figure 5. Variable- Month period

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
4	6	0	0.86	0.00	6	0.6	0.00	10.00	Inf	Inf	0.430
5	1	0	0.14	0.00	1	0.1	0.00	10.00	Inf	Inf	0.070
6	66	9	9.43	3.00	75	7.5	12.00	7.59	11.45	7.36235	3.215
7	5	0	0.71	0.00	5	0.5	0.00	10.00	Inf	Inf	0.355
8	6	1	0.86	0.33	7	0.7	14.29	7.23	9.58	0.50774	0.265
9	35	14	5.00	4.67	49	4.9	28.57	5.17	0.68	0.02244	0.165
10	25	3	3.57	1.00	28	2.8	10.71	7.81	12.73	3.27161	1.285
11	9	0	1.29	0.00	9	0.9	0.00	10.00	Inf	Inf	0.645
12	130	49	18.57	16.33	179	17.9	27.37	5.32	1.29	0.28896	1.120
13	4	0	0.57	0.00	4	0.4	0.00	10.00	Inf	Inf	0.285
14	3	1	0.43	0.33	4	0.4	25.00	5.66	2.65	0.02650	0.050
15	52	12	7.43	4.00	64	6.4	18.75	6.50	6.19	2.12317	1.715
16	1	1	0.14	0.33	2	0.2	50.00	2.98	-8.57	0.00000	0.095
18	71	42	10.14	14.00	113	11.3	37.17	4.20	-3.23	1.24678	1.930
20	7	1	1.00	0.33	8	0.8	12.50	7.52	11.09	0.74303	0.335
21	21	9	3.00	3.00	30	3.0	30.00	5.00	0.00	0.00000	0.000
22	2	0	0.29	0.00	2	0.2	0.00	10.00	Inf	Inf	0.145
24	128	56	18.29	18.67	184	18.4	30.43	4.95	-0.21	0.00798	0.190
26	1	0	0.14	0.00	1	0.1	0.00	10.00	Inf	Inf	0.070
27	8	5	1.14	1.67	13	1.3	38.46	4.06	-3.82	0.20246	0.265
28	2	1	0.29	0.33	3	0.3	33.33	4.68	-1.29	0.00516	0.020
30	27	13	3.86	4.33	40	4.0	32.50	4.71	-1.15	0.05405	0.235
33	2	1	0.29	0.33	3	0.3	33.33	4.68	-1.29	0.00516	0.020
36	46	37	6.57	12.33	83	8.3	44.58	3.48	-6.30	3.62880	2.880
39	4	1	0.57	0.33	5	0.5	20.00	6.33	5.47	0.13128	0.120
40	0	1	0.00	0.33	1	0.1	100.00	0.00	-Inf	Inf	0.165
42	8	3	1.14	1.00	11	1.1	27.27	5.33	1.31	0.01834	0.070
45	1	4	0.14	1.33	5	0.5	80.00	0.95	-22.51	2.67869	0.595
47	1	0	0.14	0.00	1	0.1	0.00	10.00	Inf	Inf	0.070
48	20	28	2.86	9.33	48	4.8	58.33	2.35	-11.82	7.64754	3.235
54	1	1	0.14	0.33	2	0.2	50.00	2.98	-8.57	0.00000	0.095
60	7	6	1.00	2.00	13	1.3	46.15	3.33	-6.93	0.69300	0.500
72	0	1	0.00	0.33	1	0.1	100.00	0.00	-Inf	Inf	0.165

Table 3. Value of Variable- Month period

Variable- Credit History :

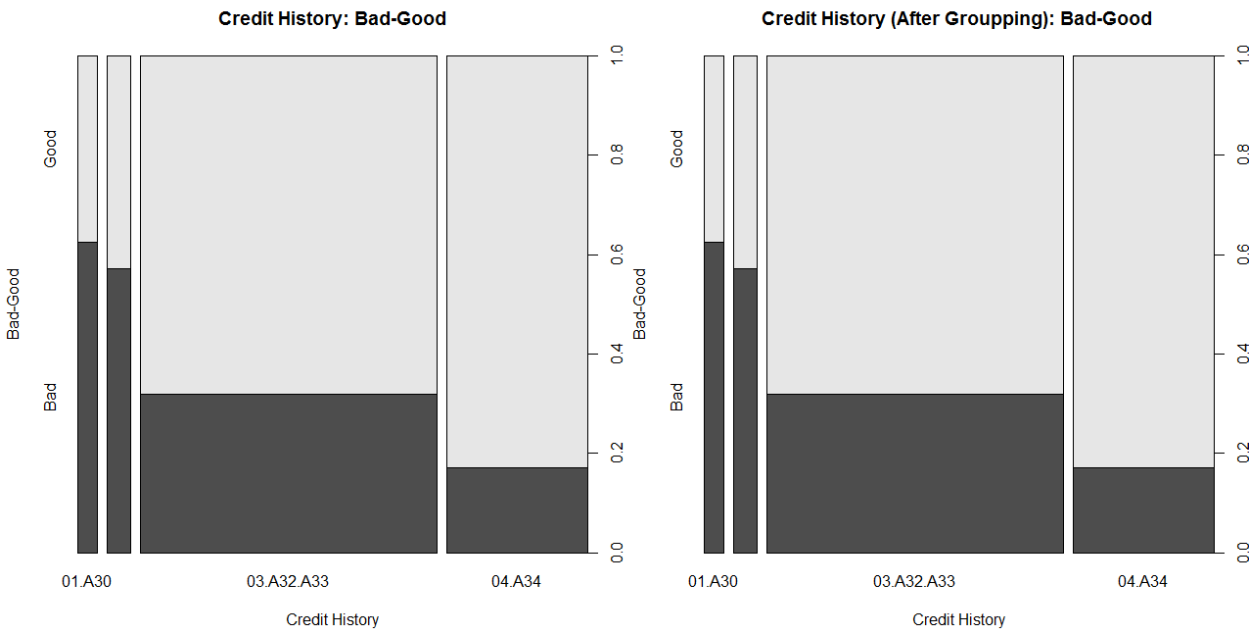

Figure 6. Variable- Credit History

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
01.A30	15	25	2.14	8.33	40	4.0	62.50	2.04	-13.59	8.41221	3.095
02.A31	21	28	3.00	9.33	49	4.9	57.14	2.43	-11.35	7.18455	3.165
03.A32.A33	421	197	60.14	65.67	618	61.8	31.88	4.78	-0.88	0.48664	2.765
04.A34	243	50	34.71	16.67	293	29.3	17.06	6.76	7.33	13.22332	9.020

Table 4. Value of Variable- Credit History

Variable- Aim of the loan :

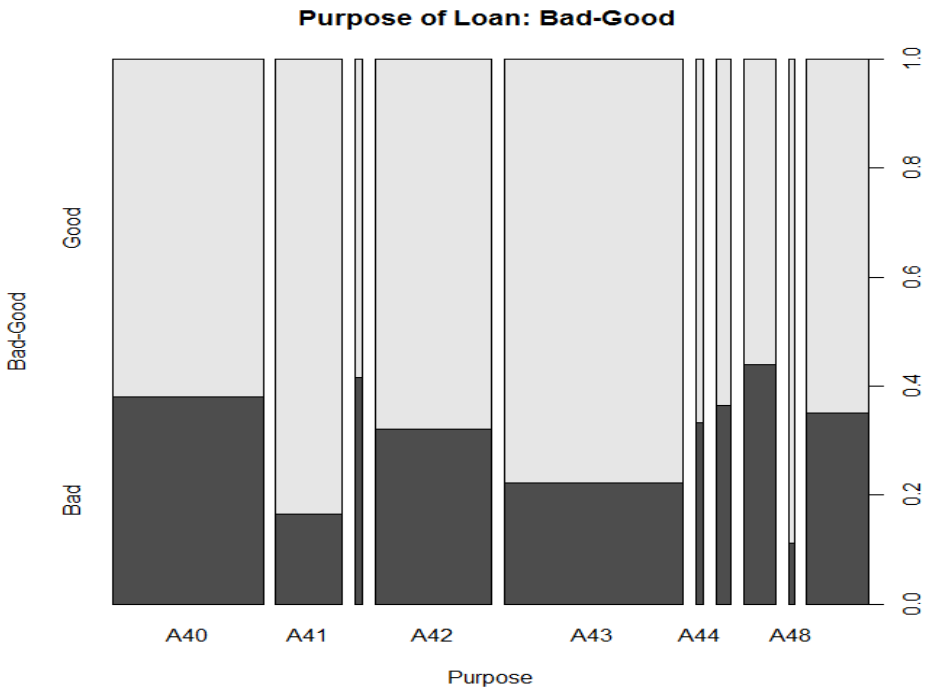

Figure 7. Variable- Aim of the loan

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
A40	145	89	20.71	29.67	234	23.4	38.03	4.11	-3.60	3.22560	4.480
A41	86	17	12.29	5.67	103	10.3	16.50	6.84	7.74	5.12388	3.310
A410	7	5	1.00	1.67	12	1.2	41.67	3.75	-5.13	0.34371	0.335
A42	123	58	17.57	19.33	181	18.1	32.04	4.76	-0.95	0.16720	0.880
A43	218	62	31.14	20.67	280	28.0	22.14	6.01	4.10	4.29270	5.235
A44	8	4	1.14	1.33	12	1.2	33.33	4.62	-1.54	0.02926	0.095
A45	14	8	2.00	2.67	22	2.2	36.36	4.28	-2.89	0.19363	0.335
A46	28	22	4.00	7.33	50	5.0	44.00	3.53	-6.06	2.01798	1.665
A48	8	1	1.14	0.33	9	0.9	11.11	7.76	12.40	1.00440	0.405
A49	63	34	9.00	11.33	97	9.7	35.05	4.43	-2.30	0.53590	1.165

Table 5. Value of Variable- Aim of the loan

Variable- Amount of credit :

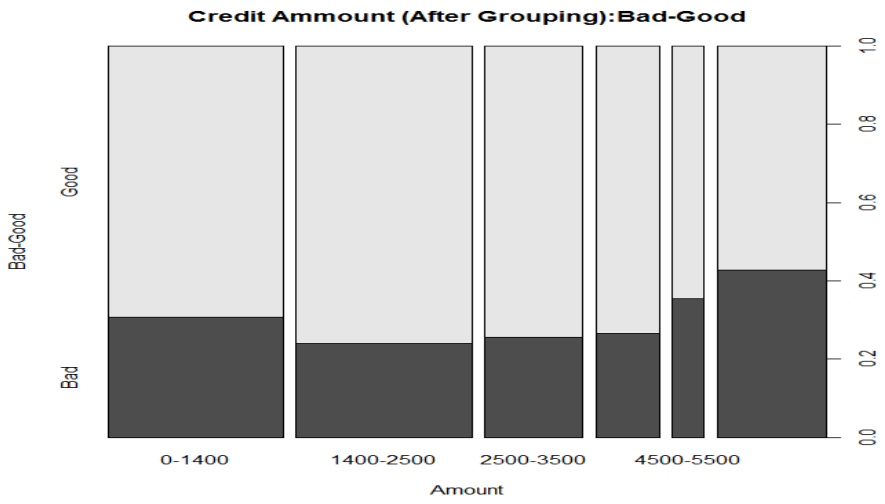

Figure 8. Variable- Amount of credit

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
0-1400	185	82	26.43	27.33	267	26.7	30.71	4.92	-0.33	0.02970	0.450
1400-2500	205	65	29.29	21.67	270	27.0	24.07	5.75	3.01	2.29362	3.810
2500-3500	111	38	15.86	12.67	149	14.9	25.50	5.56	2.25	0.71775	1.595
3500-4500	72	26	10.29	8.67	98	9.8	26.53	5.43	1.71	0.27702	0.810
4500-5500	31	17	4.43	5.67	48	4.8	35.42	4.39	-2.47	0.30628	0.620
5500+	96	72	13.71	24.00	168	16.8	42.86	3.64	-5.60	5.76240	5.145

Table 6. Value of Variable- Amount of credit

Variable- Savings account / stock :

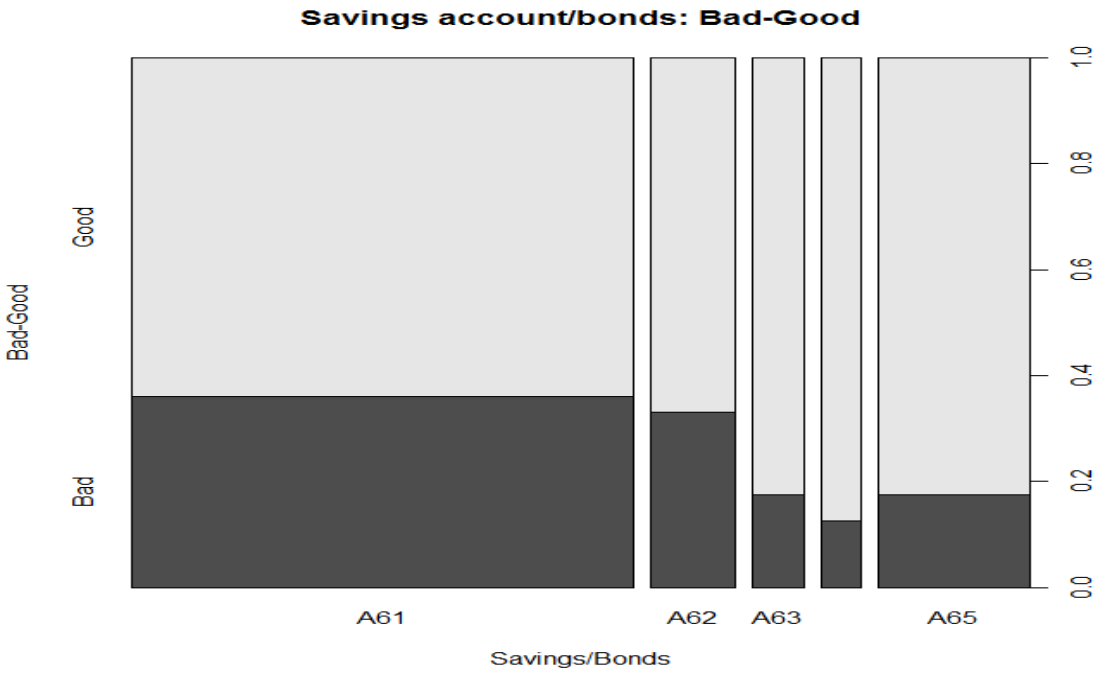

Figure 9. Variable- Savings account / stock

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
A61	386	217	55.14	72.33	603	60.3	35.99	4.33	-2.71	4.65849	8.595
A62	69	34	9.86	11.33	103	10.3	33.01	4.65	-1.39	0.20433	0.735
A63	52	11	7.43	3.67	63	6.3	17.46	6.69	7.05	2.65080	1.880
A64	42	6	6.00	2.00	48	4.8	12.50	7.50	10.99	4.39600	2.000
A65	151	32	21.57	10.67	183	18.3	17.49	6.69	7.04	7.67360	5.450

Table 7. Value of Variable- Savings account / stock

Variable- Since then get a job :

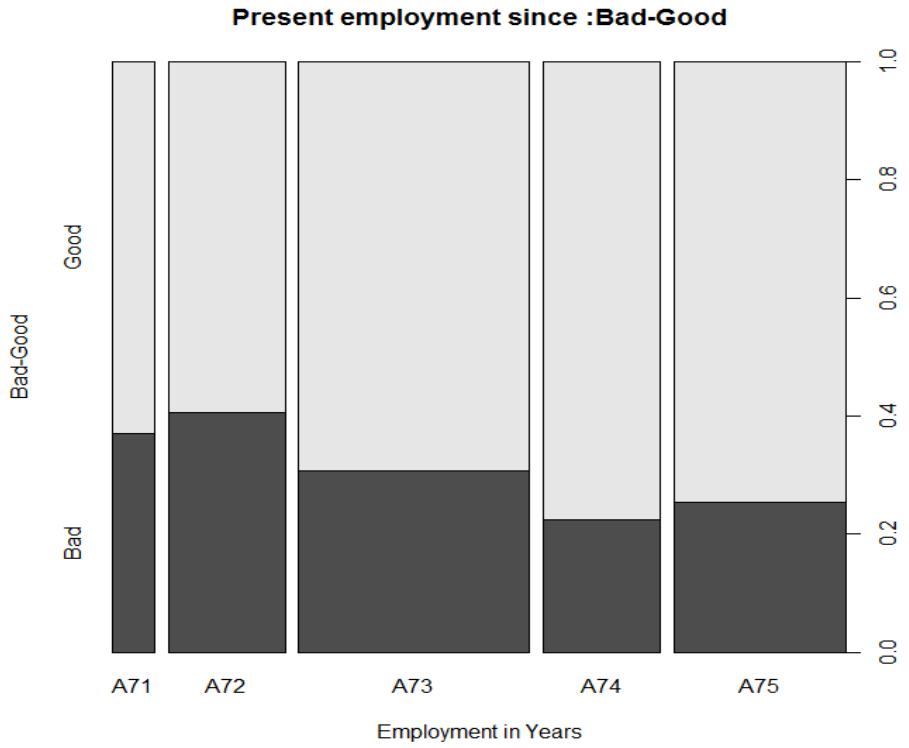

Figure 10. Variable- Since then get a job

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
A71	39	23	5.57	7.67	62	6.2	37.10	4.21	-3.20	0.67200	1.050
A72	102	70	14.57	23.33	172	17.2	40.70	3.84	-4.71	4.12596	4.380
A73	235	104	33.57	34.67	339	33.9	30.68	4.92	-0.32	0.03520	0.550
A74	135	39	19.29	13.00	174	17.4	22.41	5.97	3.95	2.48455	3.145
A75	189	64	27.00	21.33	253	25.3	25.30	5.59	2.36	1.33812	2.835

Table 8. Value of Variable- Since then get a job

Variable- Installment rate :

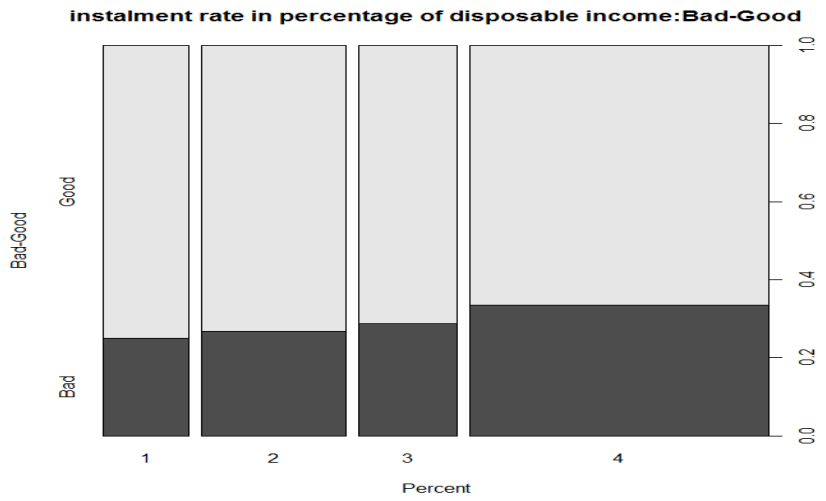

Figure 11. Variable- Installment rate

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
1	102	34	14.57	11.33	136	13.6	25.00	5.63	2.52	0.81648	1.620
2	169	62	24.14	20.67	231	23.1	26.84	5.39	1.55	0.53785	1.735
3	112	45	16.00	15.00	157	15.7	28.66	5.16	0.65	0.06500	0.500
4	317	159	45.29	53.00	476	47.6	33.40	4.61	-1.57	1.21047	3.855

Table 9. Value of Variable- Installment rate

Variable- Personal status and sex :

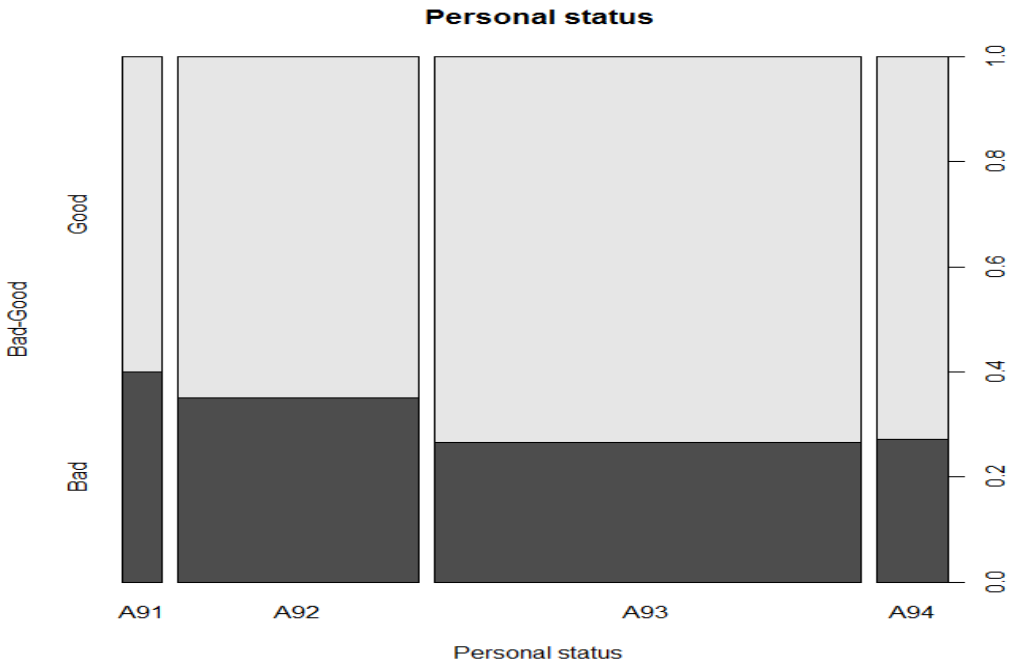

Figure 12. Variable- Personal status and sex

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
A91	30	20	4.29	6.67	50	5.0	40.00	3.91	-4.41	1.04958	1.19
A92	201	109	28.71	36.33	310	31.0	35.16	4.41	-2.35	1.79070	3.81
A93	402	146	57.43	48.67	548	54.8	26.64	5.41	1.66	1.45416	4.38
A94	67	25	9.57	8.33	92	9.2	27.17	5.35	1.39	0.17236	0.62

Table 10. Value of Variable- Personal status and sex

Variable- Other debtors / sureties :

Figure 13. Variable- Other debtors / sureties

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
A101	635	272	90.71	90.67	907	90.7	29.99	5.00	0.00	0.00000	0.020
A102	23	18	3.29	6.00	41	4.1	43.90	3.54	-6.01	1.62871	1.355
A103	42	10	6.00	3.33	52	5.2	19.23	6.43	5.89	1.57263	1.335

Table 11. Value of Variable- Other debtors / sureties

Variable- Place of residence :

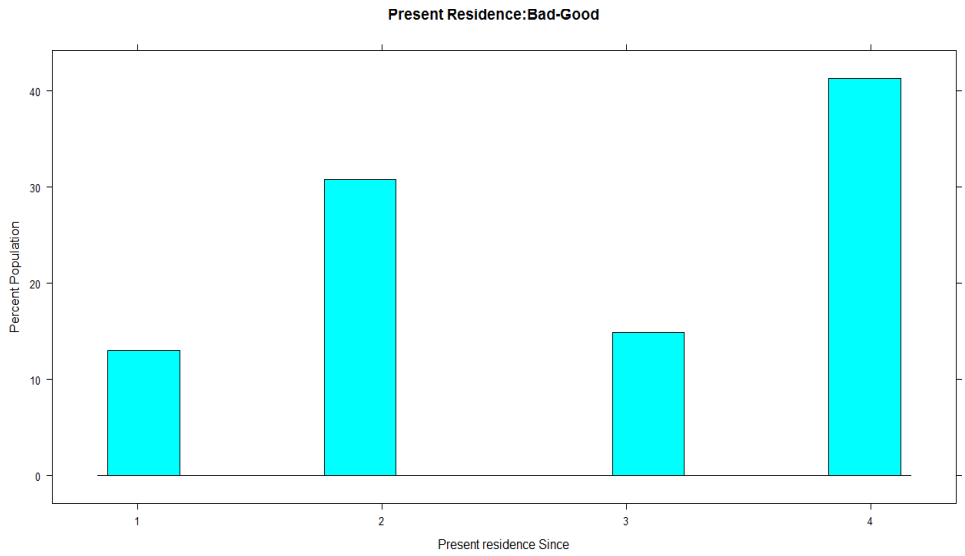

Figure 14. Variable- Place of residence

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
1	94	36	13.43	12.00	130	13.0	27.69	5.28	1.13	0.16159	0.715
2	211	97	30.14	32.33	308	30.8	31.49	4.82	-0.70	0.15330	1.095
3	106	43	15.14	14.33	149	14.9	28.86	5.14	0.55	0.04455	0.405
4	289	124	41.29	41.33	413	41.3	30.02	5.00	-0.01	0.00004	0.020

Table 12. Value of Variable- Place of residence

Variable- Estate :

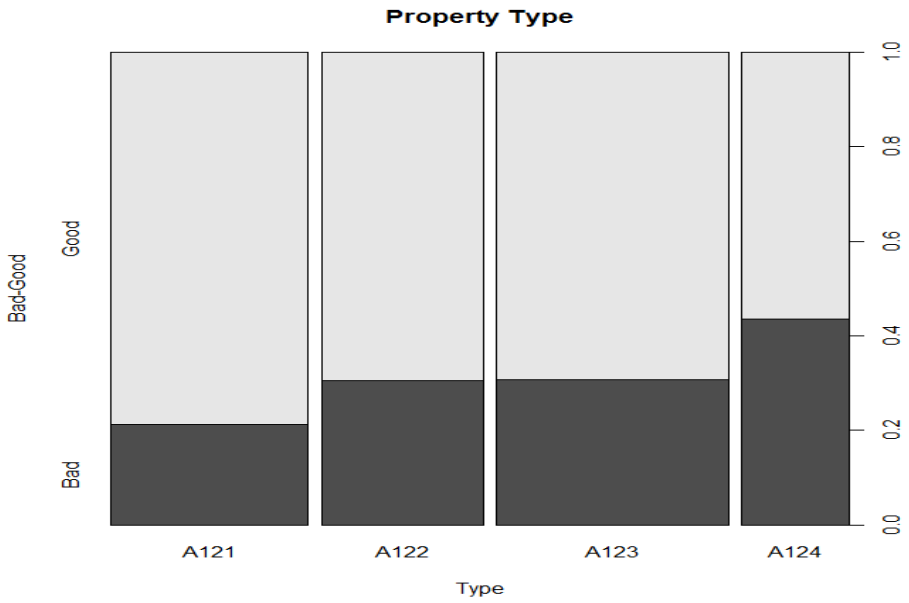

Figure 15. Variable- Estate

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
A121	222	60	31.71	20.00	282	28.2	21.28	6.13	4.61	5.39831	5.855
A122	161	71	23.00	23.67	232	23.2	30.60	4.93	-0.29	0.01943	0.335
A123	230	102	32.86	34.00	332	33.2	30.72	4.91	-0.34	0.03876	0.570
A124	87	67	12.43	22.33	154	15.4	43.51	3.58	-5.86	5.80140	4.950

Table 13. Value of Variable- Estate

Variable- Age :

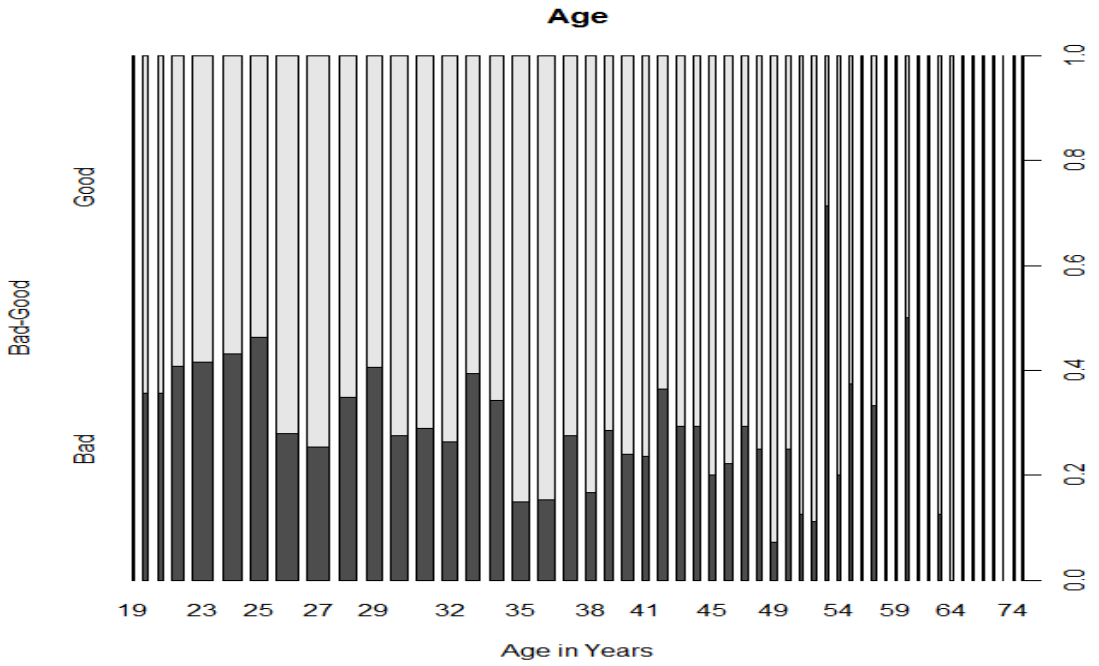

Figure 16. Variable- Age

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
19	1	1	0.14	0.33	2	0.2	50.00	2.98	-8.57	0.00000	0.095
20	9	5	1.29	1.67	14	1.4	35.71	4.36	-2.58	0.09804	0.190
21	9	5	1.29	1.67	14	1.4	35.71	4.36	-2.58	0.09804	0.190
22	16	11	2.29	3.67	27	2.7	40.74	3.84	-4.72	0.65136	0.690
23	28	20	4.00	6.67	48	4.8	41.67	3.75	-5.11	1.36437	1.335
24	25	19	3.57	6.33	44	4.4	43.18	3.61	-5.73	1.58148	1.380
25	22	19	3.14	6.33	41	4.1	46.34	3.32	-7.01	2.23619	1.595
26	36	14	5.14	4.67	50	5.0	28.00	5.24	0.96	0.04512	0.235
27	38	13	5.43	4.33	51	5.1	25.49	5.56	2.26	0.24860	0.550
28	28	15	4.00	5.00	43	4.3	34.88	4.44	-2.23	0.22300	0.500
29	22	15	3.14	5.00	37	3.7	40.54	3.86	-4.65	0.86490	0.930
30	29	11	4.14	3.67	40	4.0	27.50	5.30	1.21	0.05687	0.235
31	27	11	3.86	3.67	38	3.8	28.95	5.13	0.50	0.00950	0.095
32	25	9	3.57	3.00	34	3.4	26.47	5.43	1.74	0.09918	0.285
33	20	13	2.86	4.33	33	3.3	39.39	3.98	-4.15	0.61005	0.735
34	21	11	3.00	3.67	32	3.2	34.38	4.50	-2.02	0.13534	0.335
35	34	6	4.86	2.00	40	4.0	15.00	7.08	8.88	2.53968	1.430
36	33	6	4.71	2.00	39	3.9	15.38	7.02	8.57	2.32247	1.355
37	21	8	3.00	2.67	29	2.9	27.59	5.29	1.17	0.03861	0.165
38	20	4	2.86	1.33	24	2.4	16.67	6.83	7.66	1.17198	0.765
39	15	6	2.14	2.00	21	2.1	28.57	5.17	0.68	0.00952	0.070
40	19	6	2.71	2.00	25	2.5	24.00	5.75	3.04	0.21584	0.355
41	13	4	1.86	1.33	17	1.7	23.53	5.83	3.35	0.17755	0.265
42	14	8	2.00	2.67	22	2.2	36.36	4.28	-2.89	0.19363	0.335
43	12	5	1.71	1.67	17	1.7	29.41	5.06	0.24	0.00096	0.020
44	12	5	1.71	1.67	17	1.7	29.41	5.06	0.24	0.00096	0.020
45	12	3	1.71	1.00	15	1.5	20.00	6.31	5.36	0.38056	0.355
46	14	4	2.00	1.33	18	1.8	22.22	6.01	4.08	0.27336	0.335
47	12	5	1.71	1.67	17	1.7	29.41	5.06	0.24	0.00096	0.020
48	9	3	1.29	1.00	12	1.2	25.00	5.63	2.55	0.07395	0.145
49	13	1	1.86	0.33	14	1.4	7.14	8.49	17.29	2.64537	0.765
50	9	3	1.29	1.00	12	1.2	25.00	5.63	2.55	0.07395	0.145
51	7	1	1.00	0.33	8	0.8	12.50	7.52	11.09	0.74303	0.335
52	8	1	1.14	0.33	9	0.9	11.11	7.76	12.40	1.00440	0.405
53	2	5	0.29	1.67	7	0.7	71.43	1.48	-17.51	2.41638	0.690
54	8	2	1.14	0.67	10	1.0	20.00	6.30	5.32	0.25004	0.235
55	5	3	0.71	1.00	8	0.8	37.50	4.15	-3.42	0.09918	0.145
56	3	0	0.43	0.00	3	0.3	0.00	10.00	Inf	Inf	0.215
57	6	3	0.86	1.00	9	0.9	33.33	4.62	-1.51	0.02114	0.070
58	3	2	0.43	0.67	5	0.5	40.00	3.91	-4.43	0.10632	0.120
59	2	1	0.29	0.33	3	0.3	33.33	4.68	-1.29	0.00516	0.020
60	3	3	0.43	1.00	6	0.6	50.00	3.01	-8.44	0.00000	0.285
61	4	3	0.57	1.00	7	0.7	42.86	3.63	-5.62	0.24166	0.215
62	2	0	0.29	0.00	2	0.2	0.00	10.00	Inf	Inf	0.145
63	7	1	1.00	0.33	8	0.8	12.50	7.52	11.09	0.74303	0.335
64	5	0	0.71	0.00	5	0.5	0.00	10.00	Inf	Inf	0.355
65	4	1	0.57	0.33	5	0.5	20.00	6.33	5.47	0.13128	0.120
66	3	2	0.43	0.67	5	0.5	40.00	3.91	-4.43	0.10632	0.120
67	3	0	0.43	0.00	3	0.3	0.00	10.00	Inf	Inf	0.215
68	1	2	0.14	0.67	3	0.3	66.67	1.73	-15.66	0.82998	0.265
70	1	0	0.14	0.00	1	0.1	0.00	10.00	Inf	Inf	0.070
74	3	1	0.43	0.33	4	0.4	25.00	5.66	2.65	0.02650	0.050
75	2	0	0.29	0.00	2	0.2	0.00	10.00	Inf	Inf	0.145

Table 14. Value of Variable- Age

Variable- Other plans of installments :

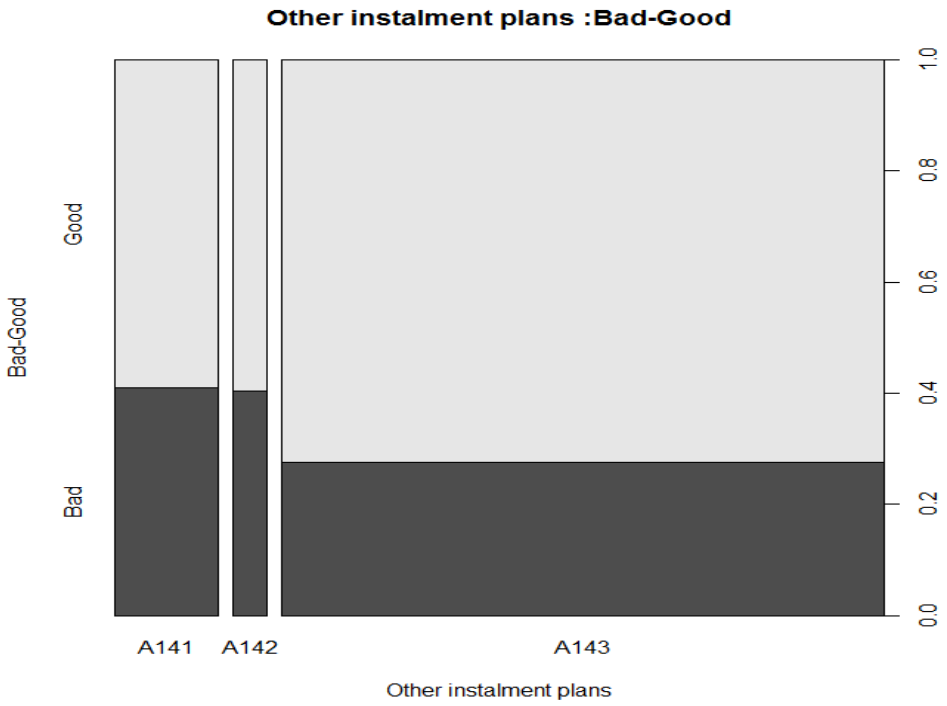

Figure 17. Variable- Other plans of installments

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
A141	82	57	11.71	19.00	139	13.9	41.01	3.81	-4.84	3.52836	3.645
A142	28	19	4.00	6.33	47	4.7	40.43	3.87	-4.59	1.06947	1.165
A143	590	224	84.29	74.67	814	81.4	27.52	5.30	1.21	1.16402	4.810

Table 15. Value of Variable- Other plans of installments

Variable- Home :

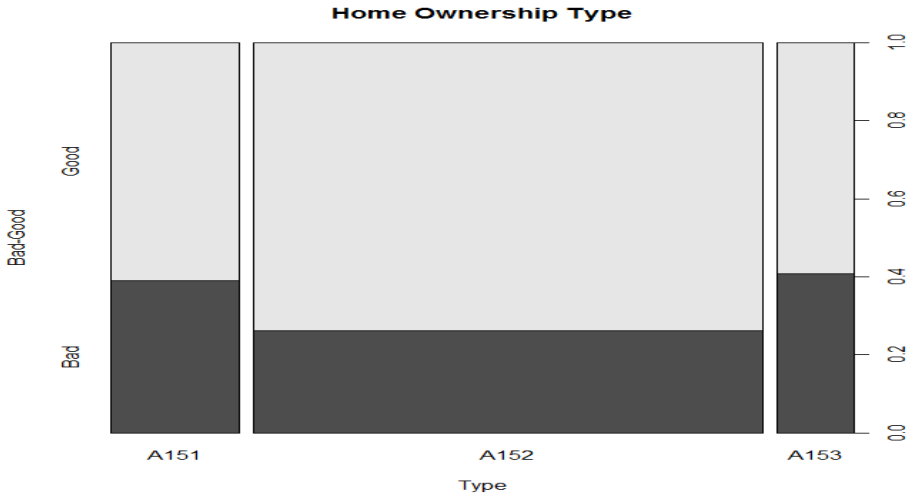

Figure 18. Variable- Home

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
A151	109	70	15.57	23.33	179	17.9	39.11	4.00	-4.04	3.13504	3.880
A152	527	186	75.29	62.00	713	71.3	26.09	5.48	1.94	2.57826	6.645
A153	64	44	9.14	14.67	108	10.8	40.74	3.84	-4.73	2.61569	2.765

Table 16. Value of Variable- Home

Variable- Number of existing loans in this bank :

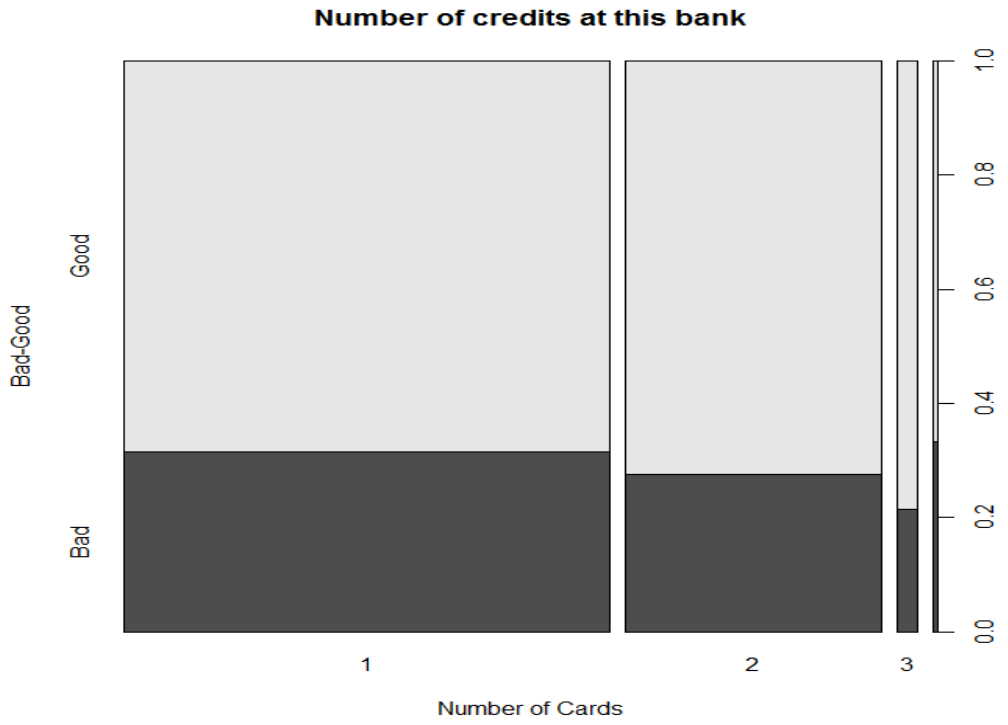

Figure 19. Variable- Number of existing loans in this bank

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
1	433	200	61.86	66.67	633	63.3	31.60	4.81	-0.75	0.36075	2.405
2	241	92	34.43	30.67	333	33.3	27.63	5.29	1.16	0.43616	1.880
3	22	6	3.14	2.00	28	2.8	21.43	6.11	4.51	0.51414	0.570
4	4	2	0.57	0.67	6	0.6	33.33	4.60	-1.62	0.01620	0.050

Table 17. Value of Variable- Number of existing loans in this bank

Variable- Job :

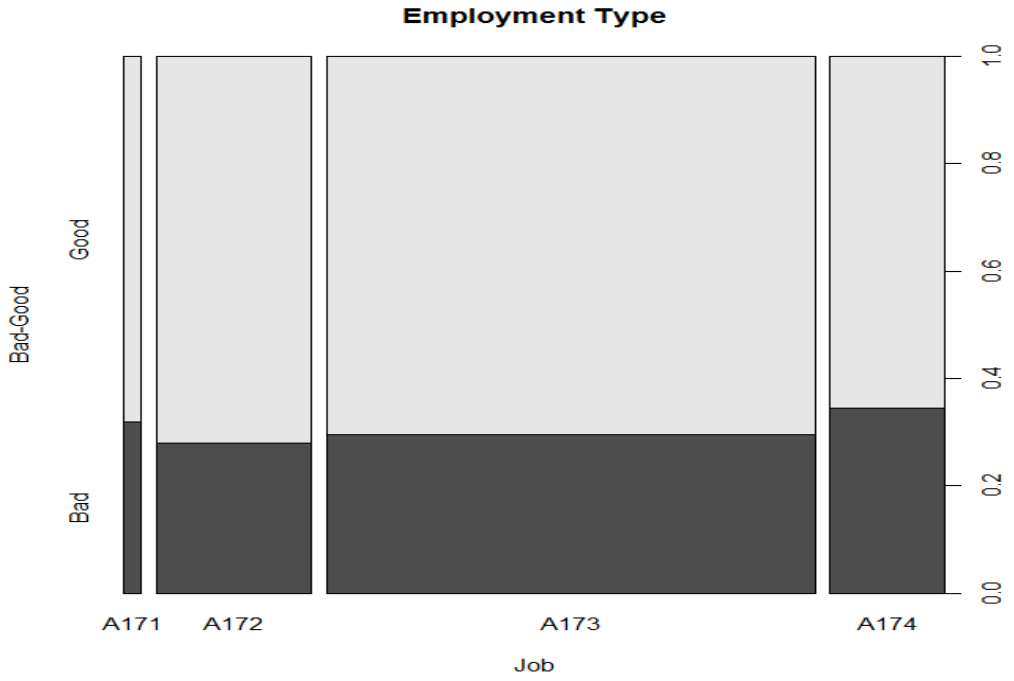

Figure 20. Variable- Job

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
A171	15	7	2.14	2.33	22	2.2	31.82	4.79	-0.85	0.01615	0.095
A172	144	56	20.57	18.67	200	20.0	28.00	5.24	0.97	0.18430	0.950
A173	444	186	63.43	62.00	630	63.0	29.52	5.06	0.23	0.03289	0.715
A174	97	51	13.86	17.00	148	14.8	34.46	4.49	-2.04	0.64056	1.570

Table 18. Value of Variable- Job

Variable- The number of persons obliged to provide care:

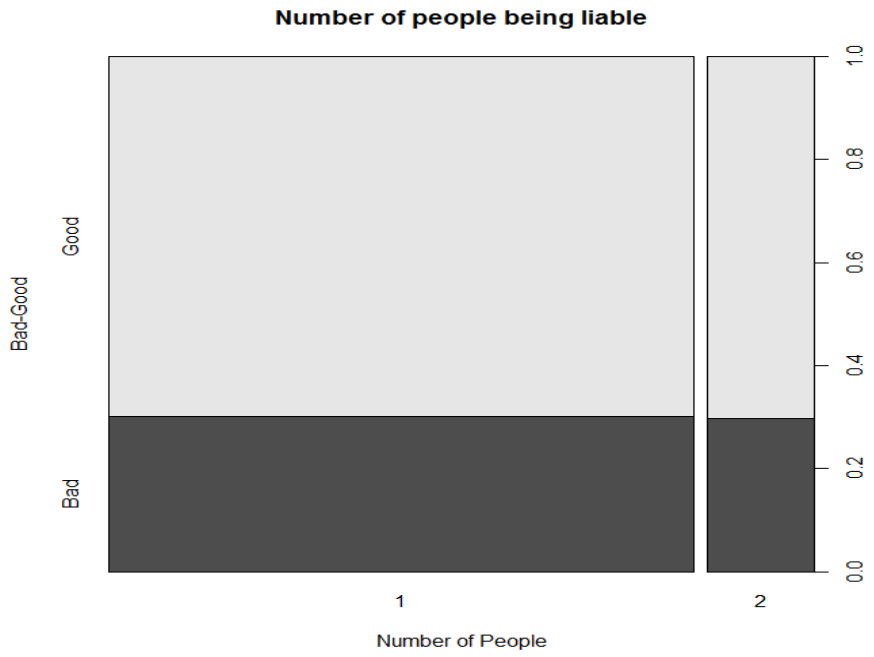

Figure 21. Variable- The number of persons obliged to provide care

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
1	591	254	84.43	84.67	845	84.5	30.06	4.99	-0.03	0.00072	0.12
2	109	46	15.57	15.33	155	15.5	29.68	5.04	0.16	0.00384	0.12

Table 19. Value of Variable- The number of persons obliged to provide care

Variable-Telephone Number (Yes/No) :

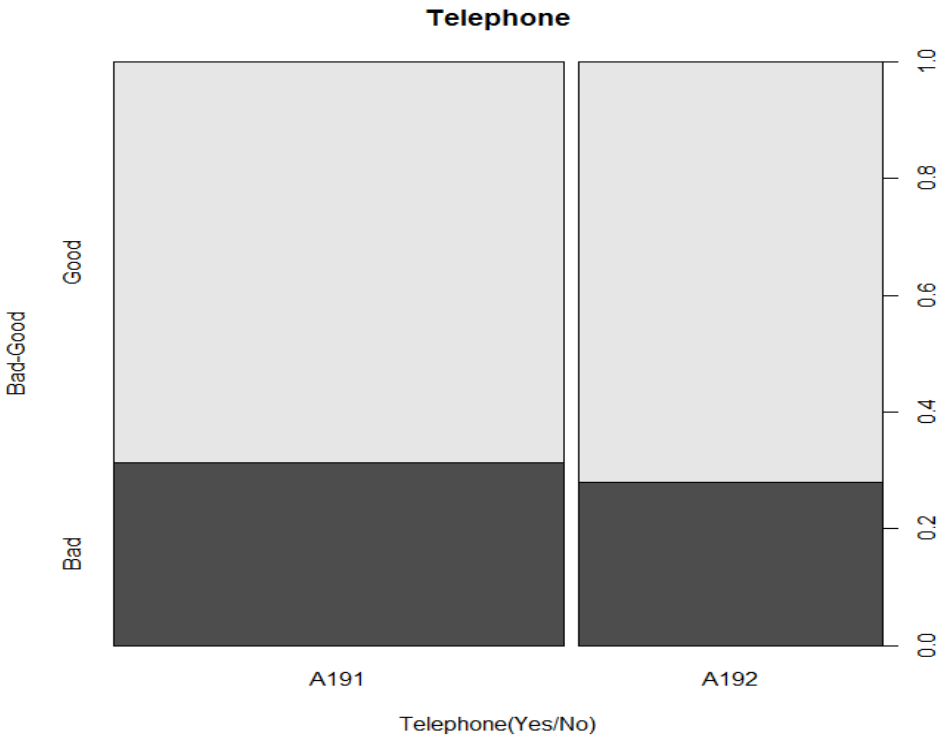

Figure 22. Variable- Telephone Number (Yes/No)

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
A191	409	187	58.43	62.33	596	59.6	31.38	4.84	-0.65	0.2535	1.95
A192	291	113	41.57	37.67	404	40.4	27.97	5.25	0.99	0.3861	1.95

Table 20. Value of Variable- Telephone Number (Yes/No)

Variable-Foreign Employee (Yes/No) :

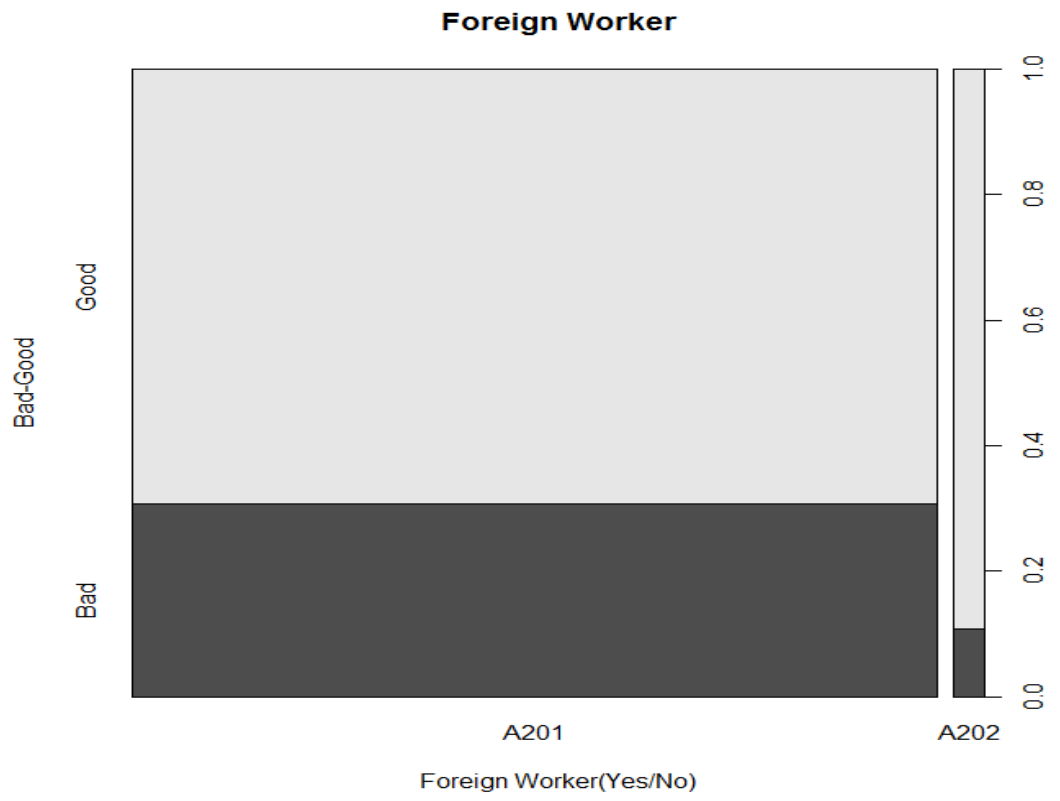

Figure 23. Variable- Foreign Employee (Yes/No)

Names	Good	Bad	Good_pct	Bad_pct	Total	Total_Pct	Bad_Rate	grp_score	WOE	IV	Efficiency
A201	667	296	95.29	98.67	963	96.3	30.74	4.91	-0.35	0.1183	1.69
A202	33	4	4.71	1.33	37	3.7	10.81	7.80	12.65	4.2757	1.69

Table 21. Value of Variable- Foreign Employee (Yes/No)

3.1.2 Partition of Data

After the data is analyzed, the partition is processed in the data set.

We can divide random samples with 50-50, 60-40, or 70-30 ratios for training (Model to be developed or trained), and Test (validation / retention pattern model can be tested according to population size). In the thesis, we will split the sample into 70-30.

Three types of basic sampling strategy:

- Random
- Systematic
- Stratified

Simple random sampling is a sampling technique (example) when selecting a group of subjects to work with a larger group (population). Every individual is chosen purely by chance and every member of the population equals the chance of being included in the sample.

Select train sampling :

	Count	Percentage
Bad	210	30
Good	490	70

Select test sampling :

	Count	Percentage
Bad	90	30
Good	210	70

3.2 Support Vector Machine (SVM) Modeling

3.2.1 Support Vector Machine (SVM)

It is one of the most effective and simple methods used for classification. It is possible to distinguish two groups by drawing a boundary between the two groups in a plane for classification. The place where this border can be drawn is that the two groups should be the farthest place to their members. Here SVM determines how this boundary is drawn.

In order to do this, two near and two parallel border lines are drawn on the two groups and these boundary lines are drawn closer together to produce a common boundary line. Take, for example, the following two groups:

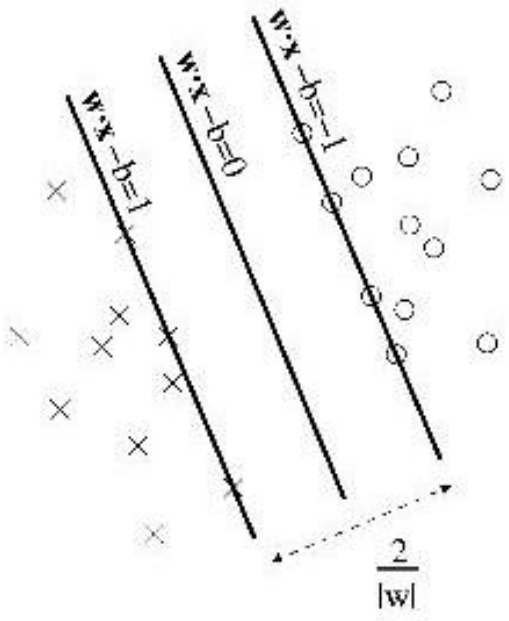

Figure 24. Support Vector Machine algorithm

In this way, two groups are shown on a two-dimensional plane. It is possible to think of these planes and dimensions as properties. In other words, a feature extraction of each input that enters the system in a simple sense has resulted in a different point that shows every input in this two-dimensional plane. The classification of these points is the classification of inputs according to the properties that have been extracted.

It is possible to say the tolerance (offset) between the two classes above. The definition of each point in this plane can be made by the following notation:

$$D = \{(x_i, c_i) | x_i \in R^p, c_i \in \{-1,1\}\}_{i=1}^{n} \ (\mathbf{5})$$

It is possible to read the above formula as follows. For every *x*, *c*, the vector *X* is a point in our space and *c* is the value indicating that this point is -1 or +1. This set of points goes up to *i = 1* 'den n.

So this formula refers to the points of the previous form.

If we think that this formula is on an extreme plane (hyper plane). Every point in this formula:

$$wx - b = 0 \ (\mathbf{6})$$

can be expressed by the equation. Where w is the normal vector perpendicular to the hyper plane and *x* is the varying parameter of the point and *b* is the shear rate. It is possible to compare this equation to the equation for calc *ax* + *b*.

Again, according to the above equation *b* / || *w* || The value gives us the distance difference between the two groups. We have already given the tolerance (offset) value to this distance difference. In order to obtain the highest value of the distance according to this distance difference equation, *2* / || *w* || in the equation giving 3 straight values having the values 0, -1 and +1 shown in the first above, formula is used. That is, the distance between the lines is 2 units.

The two right equations obtained according to this equation are:

$$wx - b = -1 \ (\mathbf{7})$$

$$wx + b = 1 \ (\mathbf{8})$$

In fact, these equations are the result of finding the highest values obtained as a result of shifting the truths. It is also assumed that the problem is linearly separable with these equations.

As expected, it is not possible that the hyper plane between the two groups is one way. Here is an example of this situation:

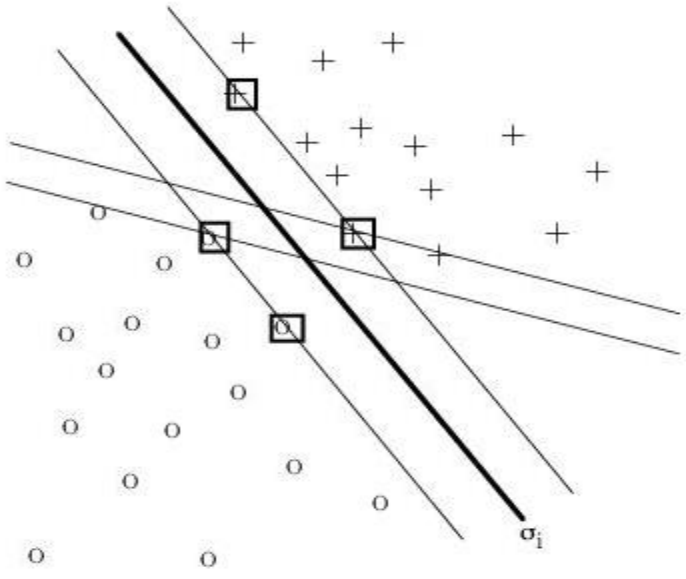

Figure 25. In the SVM, it is not possible for the hyper plane to be unidirectional between the two groups

In case of two different hyper planes (extreme planes) as above, the SVM method takes the one with the greatest possible offset from these possibilities.

3.2.2 SVM - Vanilladot Kernel

Support Vector Machines are the perfect tool for classification, regression and innovation detection. Svc, nu-saver (regression), no-Svc (classification) formulations with the well-known single class Svc (novelty) eps-saver, together with the kSVM, the classification formulations in native multi-species and the borderline SVM formulations. In addition, KSVM supports confidence intervals and class probability output for regression.

KSVM Basic Model:

$$ksvm(x, data, kernel)$$

x - A symbolic description of the model to follow. If you do not use a formula, *x* may be a matrix or vector containing training data, or a list of character vectors (for use with a character string kernel) or a kernel matrix of the class core matrix of training data. Note that regardless of whether the cut point is given in the form, it is always excluded.

$data$ - is the data set containing the training data when using the formula. When no parameter is given, the data are taken from the environment where `ksvm 'is called.

$kernel$ - the core function used in prediction and training. This parameter can be set to any function of the core class that computes the inner product of the property field between the two vector arguments. Kernlab provides the most popular kernel functions that can be used by setting the kernel parameter to the following strings:

- *rbfdot* - Radial Basis kernel "Gaussian"
- *polydot* - Polynomial kernel
- *vanilladot* - Linear kernel
- *tanhdot* - Hyperbolic tangent kernel
- *laplacedot* - Laplacian kernel
- *besseldot* - Bessel kernel
- *anovadot* - ANOVA RBF kernel

- *splinedot*- Spline kernel
- *stringdot* - String kernel

We have chosen the "VanillaDot Kernel" parameter as a kernel in designing our model.

$$ksvm(x = [good\_bad\sim], data = [train\_data\_set], kernel = [Vanilladot])$$

3.2.3 SVM - Gaussian RBF kernel

Radial base function kernel, also called RBF kernel, or Gaussian kernel is a cure in the form of a radial basis function (more specifically a Gaussian function). Gaussian kernel is probably the most used kernel functions. RBF kernel :

$$K(x, y) = exp\left\{-\frac{\|x - y\|^2}{2\partial^2}\right\} \quad (9)$$

Matches the input field to the infinite size property field. This feature is very flexible and can accommodate a wide variety of decision boundaries. The Gaussian kernel is often called the radial basis function (RBF). In some cases, it is parameterized somewhat differently:

$$K_{RBF}(x, y) = exp[-\gamma\|x - y\|^2] \quad (10)$$

The following function is used in our model in the Gauss Kernel model.

$$ksvm(x = [good\_bad\sim], data = [train\_data\_set], kernel = [RBFDot])$$

3.3 Artificial Neural Network Modeling

3.3.1 Artificial Neural Network (ANN)

An artificial neural network is an information processing system arising from the imitation of the nerve cell and the nerve network by a computer. Artificial neural networks are interconnected, connected in parallel, distributed systems, each of which has its own information processing capability and memory. Artificial neural networks are systems of similar structure derived from artificial neurons in the human brain. Artificial neural networks are a mathematical model or measurement model developed based on biological neurons.

Artificial neural networks are artificial intelligence technologies that are multivariate and produce successful results when there is complex, mutual interaction between variables, or when a single set of solutions is not found.

Artificial neural networks have been developed through the mathematical modeling of human nervous systems and have certain assumptions (Fausett, 2004);

- Information processing is through the elements called neurons.
- The signal is propagated through the connections between the neurons.
- Each connection has a certain weight and these weights are multiplied by the signals.
- Each neuron has an activation function that generates the output signal by means of a formula.

Artificial neural networks are characterized by the following forms;

- the connections between neurons,
- (learning, teaching, algorithm) of the connections between the links,
- activation formula (Fausett, 2004).

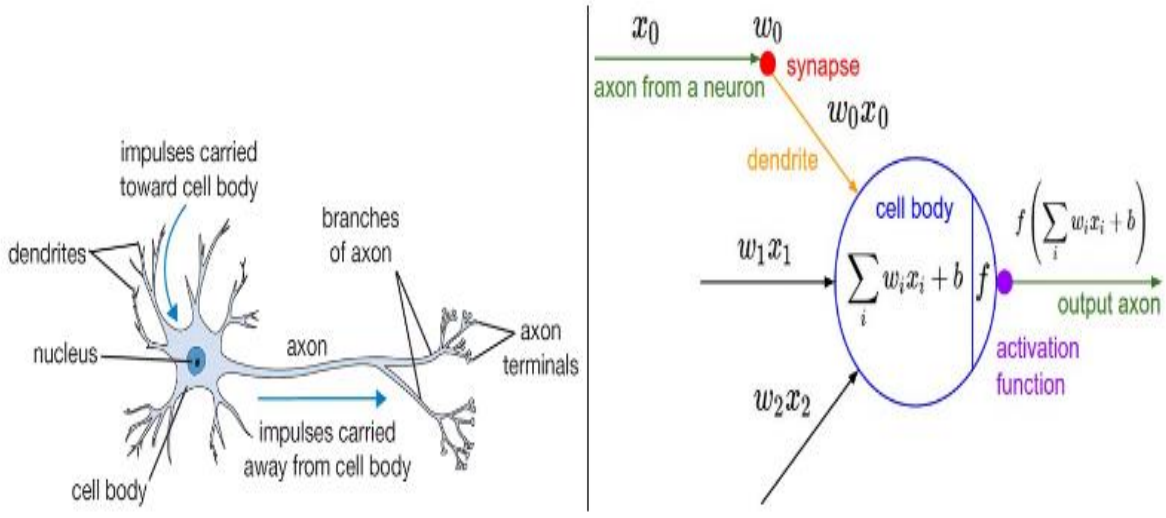

Figure 26. Systematic representation of Artificial Neural Network and biological neural network

Input

Entries are information from an outside world into an artificial neural cell. Entries in artificial neural networks are unprocessed data. Artificial neural network's output according to input data. According to the weights of the inputs, the learning function is performed and the training phase is carried out.

Weights

Weights are the values that determine the effect of incoming information on the cell. The information enters the cell through the weights on the links and the weights show the relative strength (mathematical coefficient) of the values to be used as inputs in the artificial neural network. There are different weight values of all the connections that allow the inputs to be transmitted between cells within the artificial neural network. Thus, weights affect each entry of each processor element. Weights can be variable or fixed.

Activation function

The output value is obtained by using an activation formula summing the weighted inputs. In most cases the output of the weighted input values from the activation process is between -1 and 1 or between 0 and 1. The activation formulas used in most studies are shown in the following chart (Bigus, 1996). The sample activation functions are given in the following chart.

Activation function	Mathematical Definition
Lineer	$f(x) = x$
Lojistik Sigmoid	$f(x) = \frac{1}{1 + exp(-x)}$
Hiperolik Tanjant	$f(x) = tanh(x)$

Gaussion	$f(x) = \exp(-\frac{x^2}{2\partial^2})$

Output

Outputs are final process elements. The output is determined by the activation function. Each neuron sends its output as an input to another neuron. There is only one output value from a neuron. Artificial neural network output, solution of the problem. For example, loan appraisal, credit appraisal; positive or negative. Outputs such as input in artificial neural networks are composed of numerical values. +1 is positive, 0 is negative. It is the calculation of the objective output value of artificial neural networks.

3.3.2 Artificial Neural Networks Structure and Model

Artificial Neural Networks have the following features:

- Feedforward Neural Networks
- Backpropagation Neural Networks
- Multilayer Perceptron

Feedforward Neural Networks-In a Feedforward Neural Networks, the processor elements (PE) are generally divided into layers. The cells in each layer are fed only in the cells of the previous layer. In the forward 40-feed ANN, the cells are arranged in layers and the outputs of the cells are input as weights over the next layer. The entry layer conveys the information from outside the cells in the intermediate layer without any change. Information is processed in the middle and output layer to determine the network output. Signals are transmitted from the input layer to the output layer through one-way connections. When the PEs establish a connection from one layer to another layer, there are no connections within the same layer. Examples of advanced networking include Multi Layer Perceptron (MLP) and Linear Vector Quantization (LVQ) networks (Alavala, 2002).

Feedback Neural Networks - There are dynamic memories of this kind of neural networks, and one output reflects both that input and the previous input. Therefore, they are used especially in forecasting applications. These networks have been quite

successful in predicting various types of problems. As an example of these networks; Hopfield, Self Organizing Map (SOM), Elman and Jordan networks (Alavala, 2002).

Multilayer Perceptron- Multilayer sensor artificial neural networks are one of the most used neural network models in engineering applications.Many layered sensors (perceptron model, consisting of one input, one or more intermediate and one output layer).All the processing elements in a layer depend on all the processing elements in a top layer.The information flow is forward and there is no feedback.For this reason, it is called as a feedforward neural network model.

The following function has been used to train the Neural Network:

$$nnet(formula, data, size, maxit, decay, linout, trace) \quad (11)$$

formula- A formula of the form class : $x_1 + x_2 + \ldots$

data- The data frame to be taken first of the variables specified in the formula

size- number of units in the hidden layer.

maxit- maximum number of iterations.

decay- parameter for weight decay.

linout- switch for linear output units.

trace- switch for tracing optimization.

3.4 Scoring Model with Support Vector Machine and Artificial Neural Network

3.4.1 Scoring with SVM - Vanilladot Kernel Model

We use the ksvm function and the kernel parameter Vanilladot function in the study. We apply this model on the test data.

With the first once predict function, we get a score for the test data, then we look at how well these scores match our target data with the prediction function capability.

Then the performance evaluation of the model is done. The performance evaluation is calculated according to the output values.

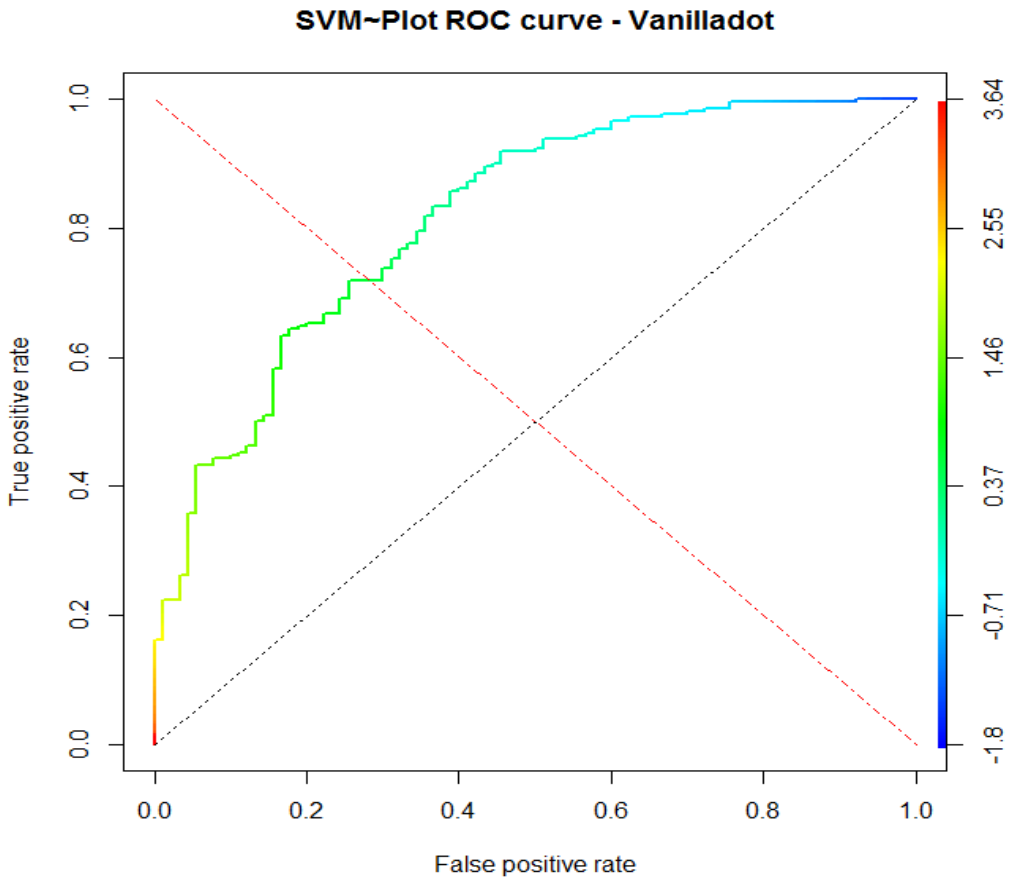

Figure 27. SVM - Vanilladot Kernel Model ROC Curve

The ROC curve (A Receiver Operating Characteristic Curve) shows the performances of the two cluster classifiers at the possible rung intervals. Ideal classifiers are collected on the left and top of the graph under the points where the curve has a value of 1.0. Random classifiers are successful at around 0.5 (classifiers in the area below 0.5 can be developed).

ROC curve classifiers are recommended for benchmarking, not just the arbitrarily chosen decision step, but the performance in all possible decision steps. The ROC curve is used to select the optimum step decision. This step (equal to the wrong classification ratio in both classes) can be used automatically in the step confidence setting.

In SVM-VanillaDot Kernel model, the following results were observed.

Classification			
Observation	Prediction		
	Good	Bad	Rate
Good	183	38	82%
Bad	27	52	65%
Average Rate	70%	30%	78%

Table 22. Outcome from classification made with SVM-VanillaDot Kernel model

As we have seen from the classification chart, the number of well-estimated customers (183) is really good, 82% and the number of poorly estimated customers actually good is 38. In reality, the ratio of badly estimated customers (52) is 65%, and in reality, if it is bad, the number of customers estimated to be good is 27. According to the SVM-Vanilladot Kernel model, the prediction success is 78% for good and bad customers.

In the model, the support vector number is 399 and the training error is 0.225714

And finally, the performance of the model was evaluated by three important Model Evaluation Error Criteria.

Model Evaluation Error Metrics	Performance Metrics
AUROC (AUC – ROC)	81.50
KS (Kolmogorov Smirnov)	46.82
Gini (Gini Coefficient)	63.02

Table 23. Outcome from the evaluation of SVM-VanillaDot Kernel model with Model Evaluation Error Criteria

3.4.2 Scoring with SVM - Gaussian RBF Model

We use the Support Vector Machine classification method to examine a second scoring model. The same KSVM function was used to set up our model, but unlike the Vanilladot model, RBFDot (Gaussian RBF) was chosen as the kernel parameter.

As with the VanillaDot model, new functions are estimated for customers with forecasting functions.

Now let's look at the ROC Curve of this model.

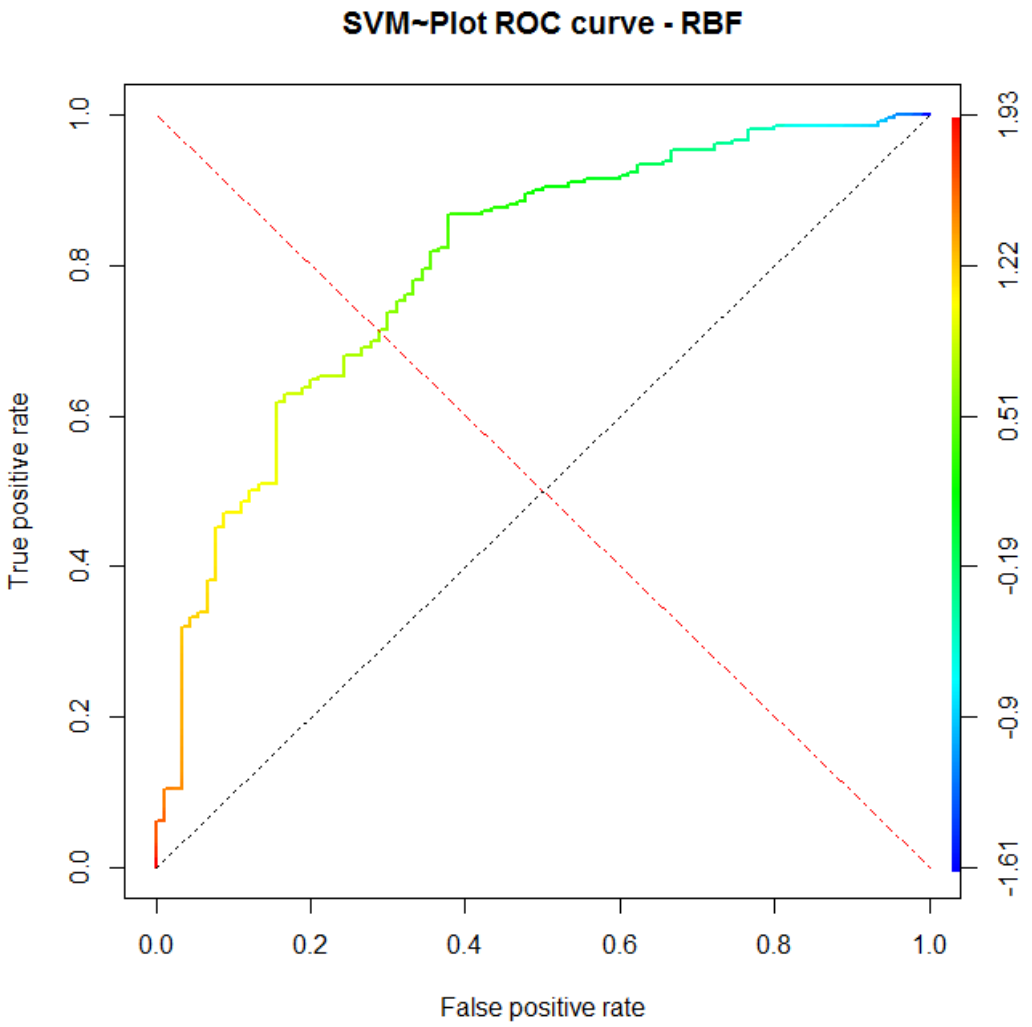

Figure 28. SVM - Gaussian RBF Model ROC Curve

The predicted maximum value (1.93) was the minimum value (-1.61) as seen from the ROC curve.

SVM- Gaussian RBF Kernel model shows the results of classifications realized in the following chart.

Classification			
Observation	Prediction		
	Good	Bad	Rate
Good	199	60	76%
Bad	11	30	73%
Average Rate	70%	30%	76%

Table 24. Outcome from classification made with SVM- Gaussian RBF Kernel model

The above table shows that the number of well-estimated customers is 199 and the rate is 76%. The number of customers is 30 and the ratio is 73%. The number of customers who are really bad and who are good at the estimate is 11, Our model's prediction success rate is 76%.

In the SVM-Gaussian model, support vector number 457 and training error 0.2 were observed.

Performance evaluation of the Scoring Model was measured as follows.

Model Evaluation Error Metrics	Performance Metrics
AUROC (AUC – ROC)	72.89
KS (Kolmogorov Smirnov)	38.42
Gini (Gini Coefficient)	46.02

Table 25. Outcome from the evaluation of SVM- Gaussian RBF Kernel model with Model Evaluation Error Criteria

3.4.3 Scoring with Artificial Neural Networks Model

Neural Networks, NeuralNetTools, e1071 libraries are used in modeling of Artificial Neural Networks with R programming language.

First, the Train and Test data are normalized by the normalize function.

Normalize<- function (x) {

Return((x-min (x)) / (max(x)-min(x)))

}

The model uses a multi-layer sensor network structure, that is, it has a feedforward structure consisting of three layers of input, neural network middle and output layers. The model of the network is as follows:

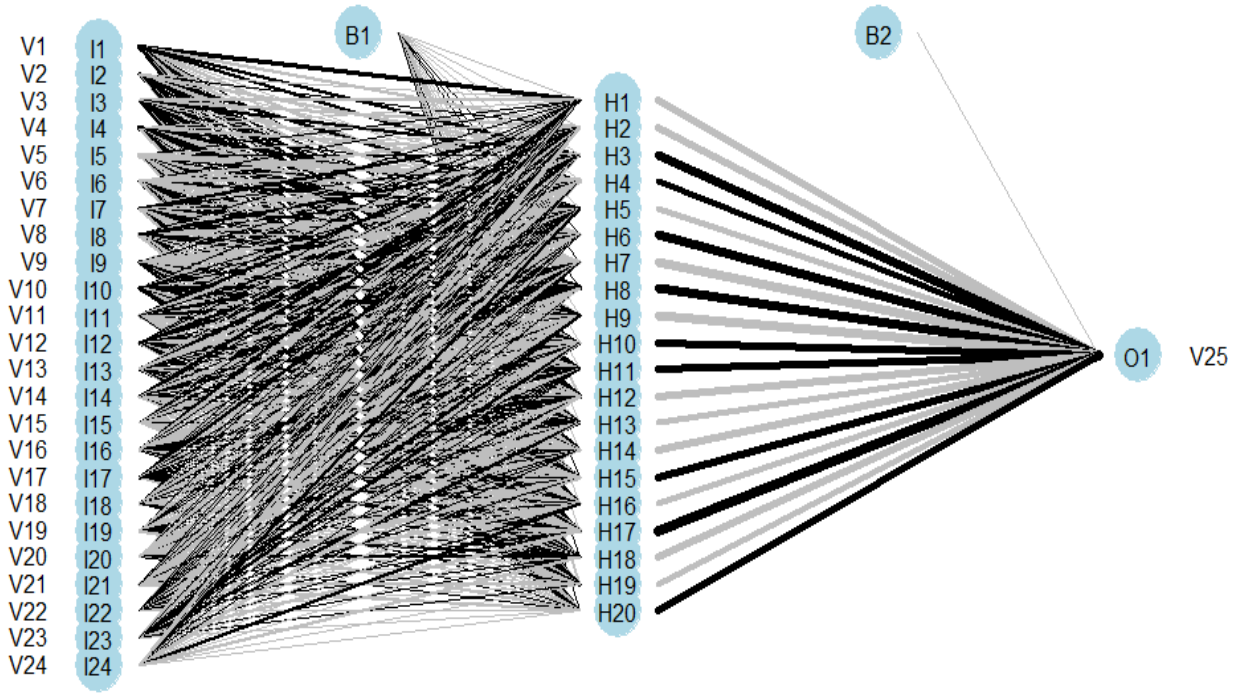

Figure 29.ANN model architecture

As you can see from the picture, 24 input variables are included in the neural network, these inputs are passed to the next layer and the number of inputs in this

intermediate layer is reduced to 20.And the output layer, which is our last layer, consists of 1 variable.The connection weights of the nerve cells are 521.The nnet function is selected for training of neural networks.

And the parameters of this function are given in the appropriate input Table 26.

Parametre	Value
formula	V25 ~ V1+V2+V3+…+V24
data	train_nn (normalize olunmuş train veriler)
size	20
maxit	10000
decay	0.01
linout	F
trace	F

Table 26. Parameter inputs for the ANN training function

The result table of the Artificial Neural Network model configured for classification in scoring model:

Classification			
Observation	Prediction		
	Good	Bad	Rate
Good	162	56	54%
Bad	46	36	12%
Average Rate	69%	31%	66%

Table 27. Outcome from classification made with ANN model

The number of well-anticipated customers is 162 and the ratio is 54%. The number of malicious customers is 36 and the ratio is 12%. The estimated number of clients coming out badly is calculated as 56. And the predictive success of the model is 66% -dir.

The performance curve and values of the model are shown below.

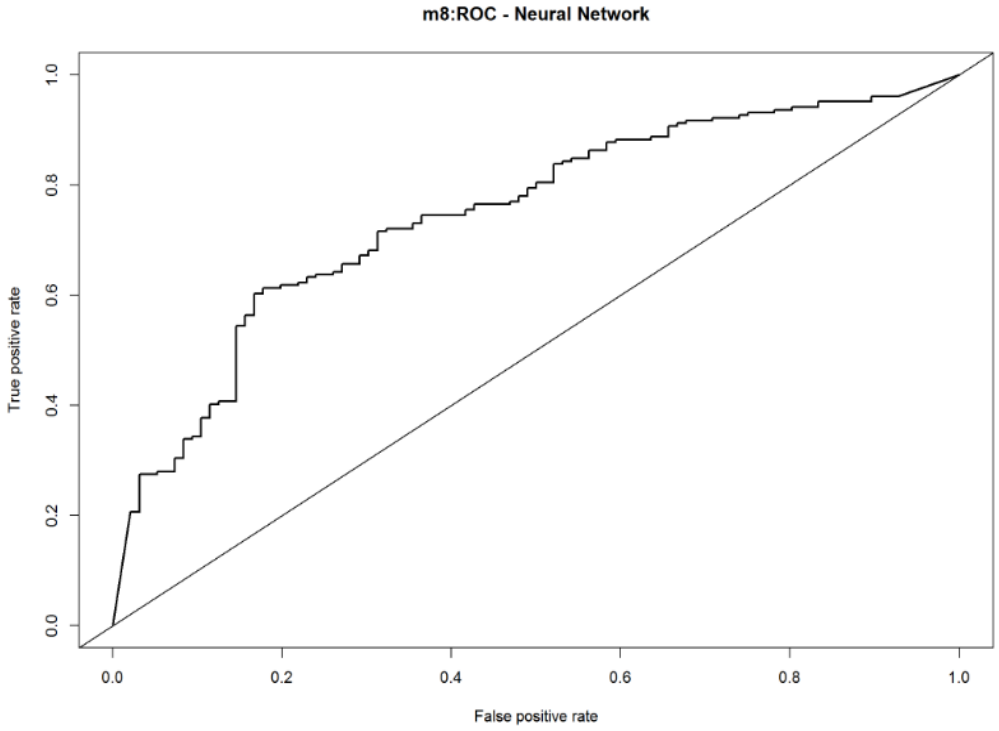

Figure 30. ANN Model ROC Curve

Model Evaluation Error Metrics	Performance Metrics
AUROC (AUC – ROC)	74.60
KS (Kolmogorov Smirnov)	43.61
Gini (Gini Coefficient)	49.30

Table 28. Outcome from the evaluation of ANN model with Model Evaluation Error Criteria

CONCLUSION

The size and importance of the banking sector are closely monitored by governments and supervisory agencies. The most important risk that affects the banking sector is the concept of credit risk. Credit risk can be described as the risk that arises if the credited creditor is not paid. The banks try to make sure that the customers they give credit are well known and that they can repay the loans they give. But nowadays the number of people who use credit from banks reaches to millions, making trusting and customer-based lending process very difficult. Techniques called credit scoring techniques and measuring whether or not to give credit to customers using customer information are now widely used between banks today.

Credit scoring is a technique that uses the customer information to decide whether to give credit to customers who apply for credit. With credit scoring methods, it should be possible to lower the percentage of credit customers who will face difficulties in repayment when choosing lenders to receive higher rates of success and fail to pay back loans. Using statistical and non-statistical methods, the customers pass through the credit scoring and a score is formed in the bank's hands with the 2 customer information they hold.

Some banks in Azerbaijan now use scoring models. In particular, scoring models play an indispensable role in providing consumer loans. Because the information required to give consumer loans in Azerbaijan is almost transparent and the scoring model makes the analysis more accurate and gives more realistic results. There are problems with business loans. It is difficult to assess the financial account deficiency in some Azerbaijani businesses. However, if banks invest in this direction, scoring models can be created for such businesses. Right now, the lack of competition in the credit market is creating wider income sources for banks, and even a decision made by a bad credit scoring model is better than the credit decisions made by some credit analysts at the bank. Within a few years, banks will have to pay more attention to credit risks - because the banks that do not manage the risks will be the least profitable and the closest banks to the bankruptcy.

Data analysis in the day, that is to say, the examination of the data into meaningful information is seen as the key to success for many sectors. Credit scoring is one of the newest applications used in the banking sector in our country for this purpose. As

credit scoring can be used for many purposes in banking applications, only application scoring is covered in this study. Models have been prepared to estimate whether customers applying for loans are at risk and whether they apply. With these techniques, customers are classified as good-bad.

In this study, the data of credit customers of a bank were analyzed by some credit scoring techniques. Support Vector Machine and Artificial Neural Network were used in non-statistical techniques.

The Support Vector Machine algorithm uses the ksvm function, and the function has two different functions as kernel parameters: Vanilladot Kernel and Gaussian RBF Kernel. In the SVM Vanilladot Kernel model, 399 support vectors were used and training error was 0.225714,in the SVM Gaussian RBF Kernel model, 457 support vectors are used and the training error is equal to 0.2.

As the Artificial Neural Network, multilayer feed-forward network is preferred, and the connection weights of the neural networks are specified as 521. The first inputs are 21 variables, 20 of which are passed to the middle layer and 1 output (bad or good) is obtained.

The goal of the study is to compare some of the techniques that can be used for credit scoring by applying the same example.

The first applied SVM is the Vanilladot Kernel Model. The model showed 78% accuracy in determining the good and bad customers. The performance indicators AUROC = 72.89, KS = 38.42, Gini = 46.02.

Secondly SVM - Gaussian RBF Kernel Model was applied. Modeling give bad results in evaluating performance, but the accuracy of prediction indicators is 76%.

The final implementation technique is Artificial Neural Networks Model. The predictor of the model is 66%, performance values are AUROC = 74.60, KS = 43.61, Gini = 49.30.

It is the model SVM - Vanilladot Kernel, which has the best performance estimation among the models for this study and which has the best performance measures at the same time. The SVM - Vanilladot kernel is the SVM - Gaussian RBF kernel and ANN, respectively, which make good estimation after.

Order of the good customers with the highest classification accuracy is; SVM - Gaussian RBF kernel, SVM - Vanilladot Kernel, ANN.

Model ranking with the worst customers with the highest classification success; SVM - VanillaDot Kernel, ANN and SVM - Gaussian RBF kernel.

If we sort by performance evaluation: SVM - VanillaDot Kernel, YSA and SVM - Gaussian RBF kernel.

Reference

1. Ahmad Ghodselahi(2011) A Hybrid Support Vector Machine Ensemble Model for Credit Scoring, Department of Information Technology Management Tarbiat Modares University Tehran, Iran;
2. Alavala, C.R. (2003). Fuzzy logic and neural Networks: Basic concept & appliations, New Age International Publisher;
3. Anderson, R. (2007). The Credit Scoring Toolkit, Oxford University Press,
4. Arminger, G., Enache D. and Bone T., (1997), "Analyzing Credit Risk Data: A Comparison of Logistic Discrimination, Classification, Tree Analysis, and Feedforward Networks", Computational Statistics;
5. Bigus, J. P. (1996). Data Mining with Neural, Networks McGraw-Hill.
6. Brown, I., and Mues, C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications;
7. Cristianini, N., and Shawe-Taylor, J. (2000)An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press;
8. de Souza, C. R.(2010) Kernel functions for machine learning applications, March;
9. Dr. Ir. Tony Van Gestel(1), Bart Baesens(2), Dr. Ir. Joao Garcia(1), Peter Van Dijcke(3), A Support Vector Machine Approach to Credit Scoring;
10. Fausett, L., "Fundamentals of Neural Networks : Architectures, Algorithms and applications";
11. Finlay ,S.,(2012) Credit Scoring, Response Modeling, and Insurance Rating: A Practical Guide to Forecasting Consumer Behavior. Palgrave Macmillan;
12. Ha Van Sang1* , Nguyen Ha Nam2 , Nguyen Duc Nhan3(2016) A Novel Credit Scoring Prediction Model based on Feature Selection Approach and Parallel Random Forest, Indian Journal of Science and Technology;
13. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
14. http://banker.az/
15. https://www.cbar.az/
16. http://cran.r-project.org/
17. https://stackoverflow.com

18. https://www.r-bloggers.com/using-neural-networks-for-credit-scoring-a-simple-example/
19. https://www.rdocumentation.org/
20. Lee, T. S., Chiu, C. C., Chou, Y. C. and Lu C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines;
21. Lewis, E.M., (1992), An Introduction to Credit Scoring, Athena Press, San Rafael, Thomas;
22. Platt, J. C. Fast(1999) training of support vector machines using sequential minimal optimization. In Advances in kernel methods, MIT press;
23. Sunil Bhatia, Pratik Sharma, Rohit Burman , Santosh Hazar, Rupali Hande(2017) , Mumbai University, International Journal of Computer Applications
24. Sustersic, M., Mramor, D. and Zupan J. (2009). Consumer credit scoring models with limited data. Experts Sysems with Application,
25. Thomas, L.C., Edelman, D.B. and Crook, J.N. (2002). Credit Scoring and Its Applications, Society For Industrial and Applied Mathematics;
26. Williams(2008), C. Support Vector Machines. School of Informatics, University of Edinburgh;