Credit Analysis using Data Mining Techniques in Banking Sector

N.A.U.H.Jayarathna

179465N

Master of Science in Information Technology

Faculty of Information Technology

University of Moratuwa

2020

Credit Analysis using Data Mining Techniques in Banking Sector

N.A.U.H.Jayarathna

179465N

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the fulfillment of the requirements of Degree of Master of Science in Information Technology

June 2020

DECLARATION

We declare that is our own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Name of Student   Signature of Student

N.A.U.H..Jayarathna   ………………………

Date: 09th of June 2020

Supervised by

Name of Supervisor   Signature of Supervisor

S.C. Premarathne   …………………………

Date: ………………...

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest sincere gratitude towards my supervisor, Mr. Saminda Premarathne, Senior Lecturer, Faculty of Information Technology, University of Moratuwa for his guidance, supervision, advices and sparing valuable time thorough the research project.

I offer my obeisance to Dr.M.F.M Firdhous who taught Research Methodology and Literature Review and thesis writing subjects which were the basis for this research work.

I would also like to thank all the batch mates of the M.Sc. in IT degree program who gave their valuable feedbacks to improve the results of the research and my family for the support they provided me through my entire life and in particular.

Finally I would like to offer my heartiest gratitude to all the people whose names are not appeared but their untiring effort was very much crucial to make this study success throughout this work.

ABSTRACT

In the financial market, banking sector is one of the major sectors. The main objective of a bank is to maximize their shareholders returns. While maximizing the shareholders returns, they have to bear number of risks. Credit risk is one of their major risks. Credit risk is the risk that the bankers have to bear when they give loan facilities to the customers. Deciding whether the borrower is suitable to get the loan is such a long process. Currently this process is a manual process in the banks and the final decision is based on the credit officers' opinions.

This study has focused on to analyze the credit analysis of businesses using data mining techniques. Basic aim of this study is to sought and to analyze the best data mining techniques which can be used to credit analysis and appraisals of businesses in banking sector in order to get the accurate decisions by minimizing human errors. .

In this study it is empirically evaluated current techniques which are using for credit appraisals and the best data mining techniques which can be used to minimize the human errors in the banking sector. The sample consisted of 1500 records taken from a private bank in Sri Lanka which gives loan facilities to Small and Medium Scale Enterprises.

Keywords: Credit Analysis, Data mining, Banking Sector

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1 Chapter Introduction

This chapter gives an overall introduction about the research project. This chapter has basically four sections. Firstly, background of the study is presented. Secondly, the main research problem is discussed. Thirdly, research questions and objectives of the study are clarified and finally, the organization of this report is presented.

1.2 Background of the Study

Banking sector is one of the major sectors in financial market. When the bank is identified as an institute, it can be identified as a business like any other business companies and the main purpose of a bank is to increase the returns of their shareholders [7].

Due to the effects of globalization as well as the emerging competition of the market, banks always try to gain the competitive advantage over others [5].

In countries with developed financial markets, increasing the market value of common or ordinary shares is something more important. In order to do maximize the values of shares, the bank must be a profitable firm. However, the problem is at the meantime, banks have to bear risks as well. One of the major risks that banks have to bear is, the credit risk.

In banking sector, credit is a major activity and also it can be considered as one of the core sources of their earnings. By giving credit facilities to customers, banks pay salaries to their employees as well as interest to depositors & dividend to shareholders. Recently, competition in the consumer credit market has become severe [1].

Credit facilities in banking sector can be divided into personal loans and business loans. Business loans takes a major role in giving credit facilities. Among the business loans, nowadays banking sector has focused on giving loans to Small and Medium scale Enterprises (SMEs) as SMEs are playing a major role in developing the economy of the Sri Lanka.

In every country, SMEs play a critical role to develop the economy. In Sri Lanka also SMEs play a significant role. There are different parameters used by different countries in order to define SMEs. Some of them are: by referring the number of employees, amount of capital invested or amount of turnover [3]. When considering the parameters taken by Sri Lanka to define SMEs, The National Development Bank (NDB), the Export Development Board (EDB), and Industrial Development Board (IDB) use value of fixed assets as their parameter. Also, numbers of employees have been taken as the criteria to define SMEs by the Department of Census and Statistics (DCS), Small and Medium Enterprise Development (SMED), and the Federation of Chambers of Commerce and Industry (FDCCI) [11].

Nowadays the banking industry has realized the importance and the need of using data mining techniques which help to face the competition in the market. For customer segmentation, fraud detection, credit scoring and appraisals, marketing, data mining techniques are used [5]. Recently, the collection of data have been increased in banking sector. By using existing statistical techniques managing large volumes of data has become a difficult task. These things have become a reason and the need of using new analysis techniques as it supports to find out the hidden patterns in the data [3].

All the above mentioned fact clearly indicates the importance of studying best credit analysis techniques using data mining methods in banking sector in Sri Lanka.

1.3 Research Problem

Credit risk is one of the major risks that banks have to bear. There’s huge risk of deciding whether borrower is suitable for receiving the particular amount of loan and it’s about the trustworthiness. This process is called as credit appraisal or credit analysis. Currently this process is a manual process and whole decision is based on the credit officers’ opinions. This particular process contains around twenty criteria. At the beginning, there’s an initial gate to check all the relevant documents such as business registration, address proofing documents, bank statements, audited accounts, title reports, legal documents (plan, deed, extracts) and invoices. If these documents are completed only, the file goes to credit appraisal process. In that process, Procedure Process Guide (PPG) is used as a manual. All these things are done by credit officer with his or her experience. So human errors can be happen which is caused to reduce the accuracy of the decision. [1].

Early research has done on credit scoring and classifying credit card accounts etc. Although credit appraisal is also a major area where huge amount of human errors are happening currently in banking industry, there is no direct evidence that any researcher is focuses on this matter. As such, the literature has not addressed the best data mining techniques which can be used to minimize the human errors in credit analysis and appraisals in banking sector. Based on the identified gap in the literature and to clear the ambiguity, credit analysis and appraisals using data mining techniques should be analyzed.

1.4 Scope of the research

The credit facilities are given by many financial companies including banking sector. This study is limited to the credit facilities given by only banking sector in order to reduce the unnecessary complexity.

1.5 Aim of the research

The aim of this research is to investigate the ways of using data mining techniques to get the accurate decisions in credit analysis and appraisals in banking sector.

1.6 Research Questions and Objectives

In this study, research problem has been divided into three research questions. Table 1-1 illustrates the research questions and objectives.

Table 1-1 Research Questions and Objectives

Research Questions	Research Objectives
01. What are the limitations on the methods and techniques which are currently used for credit analysis and appraisals?	To identify the limitations on the methods and techniques which are currently used for credit analysis and appraisals.
02. What are the most appropriate data mining technique/s which can be used for credit analysis and appraisals in order to get the accurate decisions by minimizing the human errors?	To assess the most appropriate data mining technique/s which can be used for credit analysis and appraisals in order to get the accurate decisions by minimizing the human errors.

1.7 Proposed Solution of the study

In this study, the proposed solution for the defined problem is to identify the hidden pattern among the parameters and select the most appropriate data mining technique to get the accurate decision making in credit analysis. By identifying hidden patterns and selecting the most appropriate data mining techniques, this study proposed a simulation to get the accurate decision when a new transaction comes. The simulation was built by using several steps. Firstly, 1500 records has been taken from the bank and preprocessed data by identifying missing values. Secondly, data transformation has been done by using discretization. As the third step of preprocessing, data reduction was done by using correlation matrix to select the significant attributes. After selecting the attributes, four models were created by using selected four algorithms in classification. All these four algorithms were tested under cross validation and split validation. Finally, the performance of each model has been compared to select the most appropriate data mining technique which can be used to do the credit analysis in banking sector.

1.8 Organization of the Dissertation

The organization of the dissertation would be as structured as follows. First chapter gives an introduction about the project title, research problem, research questions and objectives, scope of the study and proposed solution. Second chapter is devoted for critically reviewed literature related to banking sector, credit analysis, data mining techniques and use of data mining techniques for credit analysis. Third chapter discusses about the technology adapted in this study. Fourth chapter is about the methodology of the study. It Includes inputs, process and outputs of the study. Fifth chapter is devoted for the detailed description of the research analysis and design. In the six chapter implementation of the proposed solution is discussed. Chapter seven is about the evaluation of the analytical methods by comparing the performances of the models while chapter eight is presented a detailed description about the conclusion of the project, limitations as well as the future works of the study.

CHAPTER TWO

LITERATURE REVIEW

2.1 Chapter Introduction

This chapter has critically reviewed the existing literature which is supporting to raise the research questions. Firstly, the literature of banking sector which is related to credit risk is clarified. Secondly, the current problems of credit analysis and appraisal are defined. Then, the need of using data mining techniques in banking sector is presented. Further, the ways of using data mining data techniques in order to resolve current problems in credit analysis are reviewed. Accordingly, the possible data mining techniques which can be used to minimize the errors in credit analysis and appraisals are identified through theoretical background and literature on banking sector and data mining techniques.

2.2 Credit risk in Banking Sector

When a particular bank needs to acquire an earning asset, the risk is assumed as the borrower will default. That means, borrower will not repay the principle as well as the interest on a timely basis. Credit risk in banking sector can be defined as the potential variation which can be happened in net income and market value of equality as a result of non-payment or delayed payment. Default probabilities are varied based on the types of assets. In banking sector, there is a huge risk of giving loans. Cash flow availability can be changed due to the alterations in general economic conditions and the operating environment of a firm. This type of conditions are difficult to predict. Likewise, the ability of an individual to repay the debt cab be changed due to the changes of employment and personal net worth. Therefore, tha bankers operate a credit analysis process on each and every loan request they receive from the client to assess the capacity of the client to repay it [7].

Possibility or a probability that a borrower will fail to fulfill his or her obligations to the bank during the contract period can be identified as the credit risk. This risk starts from the effective date of the contract agreement between the client and the bank for borrowing or leasing purposes [14]. There are some influential factors such as state of inflation, business cycles, political stability or instability which can effect to the trust or lack of trust in the future of a lease contract.

The concept which the risk of investment should be compatible with the ROI (Return Of Investment) is what one should always bear in mind in making an investment. Therefore, it is very important to acknowledge the need of credit risk analysis in financial institutions {15,16]

Banks are giving loan facilities to customers by verifying different types of details which are related to the loan facility such as period of repayment, rate of lending, an amount of loan, demography, type of property mortgaged as well as the credit history of the borrower. Customers who are dealing with the specific bank and who have relationship with the bank by doing transactions with having higher income, are likely to get loans very easily. Even though, bankers are very cautious in providing loans, there are chances for loans default by customers. Data mining techniques support to identify the borrowers who repay loans from those who don't [2].

When considering the loan facilities in banking sector, there are different types of loans that customers are giving. Customers should consider the type of loans, options and it's important when he or she borrows money. When categorizing the loans, that process can be identified as the evaluation loan collections and assigning loans to groups or grade based on the perceived danger and other related loan properties.

The process of persistent review and classification is important, in order to observe the quality of the loan portfolio and also to take action to counter fall in the credit quality of the portfolio.

The banks should use more complex internal classification schemes without using standardized schemes as bank managers require for reporting reasons which helps to make easy observing and interbank evaluation [17].

Basically banks are giving two types of loans. They are, personal loans and business loans. Banking loans are given for small and medium scale businesses and large scale businesses. In Sri Lanka, Small and Medium Scale businesses plays a major role.

Small and Medium scale Enterprises (SMEs), are considered as the backbone of the economy. [12]. SMEs play an important role in every economy by contributing to the generation of employments, growth GDP, encouraging innovations and stimulating of other activities [13].

Bankers are very enthusiastic in giving loan facilities for Small and medium Scale businesses as they are playing a major role in the economy of the country.

2.3 Issues in credit analysis and appraisal process

Non-performing loans (NPL) are one of the main critical areas in banking sector. It is a major concern in commercial banks. A non performing advance can be defined as a loan or an advance whose principle as well as the interest is in arrears for a period in excess of ninety days. This Non- performing loans is a way of reflecting the performance of commercial banks in the country. High level of non-performing loans can be considered as an indicator which helps to point out many credit defaults that affect the profitability and the net worth of banks and they erode the value of their assets .The continuous growth of non-performing loans caused to reduce the overall profits and returns to shareholders. The issue of non-performing loans has been a major subject which have been already taken higher attention amongst financial systems worldwide. This issue of non-performing loans will affect not only to the banks, but also to the entire economy. This critical issue in banking sector is a kind of reflection about the state of health of the industry and trade in a country [19].

2.4 The need of using data mining techniques in banking sector

With the impact of the globalization of the economy, bankers always keen on gaining the competitive advantage over competitors. Nowadays, banking sector has become a strategic tool for the creation of new knowledge. Recently banks' ability to generate, capture and store data has increased enormously. The information contained in this data can be very important. The wide availability of huge amounts of data and the need for transforming such data into knowledge encourage IT industry to use data mining.

All over the world, the banking industry has undertaken a remarkable change in the way business is conducted. The banking industry has started realizing the need of the techniques like data mining which can help them to compete in the market. Leading banks are using Data Mining (DM) tools for customer segmentation and profitability, credit scoring and approval, predicting payment default, marketing, detecting fraudulent transactions, etc.

Basically, there are number of risks related to bank loans, for the bank as well as for the people those who get the loans from the banks. With the tremendous growth of the customer segmentation, the number of transactions in banking sector has increased. As a result of that, huge data volumes are available which represent the customers' behavior and the risks around loan are increased. Data Mining is one of the most encouraging and dynamic area of research with the aim of extracting information from tremendous amount of accumulated data sets [17].

2.5 Data Mining Techniques

Data mining can be identified as extracting knowledge from huge amounts of data. There are different types of data such as time series data, spatial data, data related to multimedia, text data and data related to web. Moreover, Data mining can be identified as the process of extraction of stimulating, nontrivial, hidden, previously unknown and potentially useful patterns or knowledge from huge amounts of data. Data mining contains set of activities which can be used to search new, hidden or unexpected patterns in data as well as unusual patterns in data. By using information which are contained within data warehouse, data mining can frequently support to provide answers to questions related to an organization that a decision maker has previously not thought to ask [4].

When considering the data mining techniques, there are two important goals. They are prediction and description. The meaning of prediction is, using some variables in data set to predict unknown values of other variables. Description can be clarified as focusses on searching patterns clarifying the data that can be described by human. Data mining is the process of extracting unknown pattern from huge amount of data which can be used to take a correct decisions. The resulting knowledge should be new and not obvious, relevant and can be applied in the field where this knowledge has been obtained. It is also the process of extracting useful information from raw data. [17].

Recently, Data Mining has been identified as one of the most encouraging and important area of research with the aim of extracting information from tremendous amount of accumulated data sets. Moreover, Data Mining has become a popular field in banking sector as there is a way of doing analytical methodology efficiently to detect unidentified and valuable information in banks data. In data mining skills and knowledge are so vital requirement for achieving Data Mining task as the success and failure of Data Mining is greatly needful on the person who is managing the process due to unavailability of standard framework.

Data mining activities can be divided into 6 phases.

Data mining phases:

- Understand the business activity
- Data collection and analyzing
- Data preprocessing
- Modeling

- Evaluation
- Design the model

Data mining techniques support to identify the customers who will pay back the loans at the appointed time from those who don’t. It also support to assume when the customer is at default, whether providing loan to a specific customer will result in bad loans. By using data mining techniques, all processes related to banking sector could be analyzed to detect the customers.

2.6 Use of data mining techniques for credit analysis and appraisal process

Among all the data mining techniques, some advanced techniques such as decision trees, neural networks, and support vector machines have been used to implement the credit scoring models. The reason for using these techniques is their capability of modeling extremely difficult functions and getting better results in accuracy basically. These advanced techniques can be categorized in different ways. They are:

- single classifiers
- ensemble techniques
- hybrid classification techniques

By using single classifiers, credit scoring models have been developed and they are widespread. This single classifier techniques can be divided into multiple groups. They are supervised learning, unsupervised learning and other techniques. When considering the supervised learning under single classifiers, decision trees, support vector machines and neural networks are most commonly used techniques. For unsupervised learning classifiers, self-organized maps and kmeans can be taken as examples which are rarely used. Commonly used techniques for credit scoring is genetic algorithms and other techniques apart from neural networks and support vector machines [20].

Based on data mining, the knowledge discovery process and the use of different models based on the development of data mining may provide the cooperative with practical advantage. Understanding the variables as well as their relationships helps in better categorizing and forecasting cooperative members' behavior. By doing the depth assessment of the variables, support to include variables which can be important and excluding others which are not relevant by using the advantage of offering more succinct and important credit management models, reducing execution time and improving the accuracy of the decision. The analysis of discrepant or outlier cases may be relevant to create a new classification or, on the other hand, to search undesirable patterns [21].

2.7 Chapter Summary

This chapter discusses about the previous researcher’s findings regarding the banking sector, credit analysis, issues in credit analysis, data mining techniques and use of data mining techniques in credit appraisal process. In accordance with the previous research findings classification has been used for credit analysis in banking sector. Based on the identified literature, classification algorithms are used to analyze the credit analysis techniques.

CHAPTER THREE

TECHNOLOGY ADAPTED

3.1 Chapter Introduction

This chapter mainly discuss about the technology adapted in this project. Firstly, it describes the data mining techniques under supervised learning and unsupervised learning. Then, rapid miner tool has been clarified which was used for the data analysis.

3.2 Data Mining Techniques

Data Mining can be defined as a sophisticated data search capability which uses statistical algorithms to discover new patterns as well as the correlations in the data set [4].

Dara Mining also can be defined as extracting new knowledge from the large amounts of data. These data can be in different categories such as text data, wed data, multimedia data, spatial data and time series data. Data mining is also as identified as one of the processes of extraction of nontrivial, interesting, implicit, previously unknown and possibly useful some patterns or knowledge from huge amounts of data.

Data Mining is set of processes which can be used to search new, hidden or unexpected patterns in data or unusual patterns in data. Data Mining can usually provide answers to questions about an organization by using information contained within warehouse.

Data mining can be used to provide the answers for the questions by using information contained within data warehouse, which can be arisen in an organization that a decision maker has previously not thought to ask [6].

There are different data mining techniques and algorithms have been developed and used in data mining like association, classification, clustering, prediction and sequential patterns, Regression, Neural Networks etc [5].

Nowadays, the bankers are realizing the different advantages of data mining. It is so important tool as by using this, banks can recognize potentially useful information from the large amounts of data [5].

Basically, Data mining algorithms can follow three different learning approaches.

They are:

- Supervised learning
- Unsupervised learning
- Semi-supervised learning

Supervised Learning

In this learning approach, labels are known when the algorithm works with a set of examples. These labels can be nominal values in the case of the classification task, or numerical values in the case of the regression task. Under the supervised learning, techniques can be categorized into two parts called classification and regression. According to this study, under classification there are four algorithms namely, Decision Tree, Random Forest, KNN and Naïve Bayes. These techniques can be used when the results are known.

Unsupervised Learning

This is an opposite approach to supervised learning. In this unsupervised learning approach, the labels of the data set are unknown. The algorithms which are used in this approach are grouping examples according to the similarity of their attribute values, characterizing a clustering task. Unsupervised learning can be categorized into two parts namely, association rule and clustering. These techniques should use when the results are unknown.

Semi – supervised Learning

When a small subset of labelled examples is available, semi-supervised learning is usually used, together with a large number of unlabeled examples [10].

Data mining techniques are illustrated in below figure.

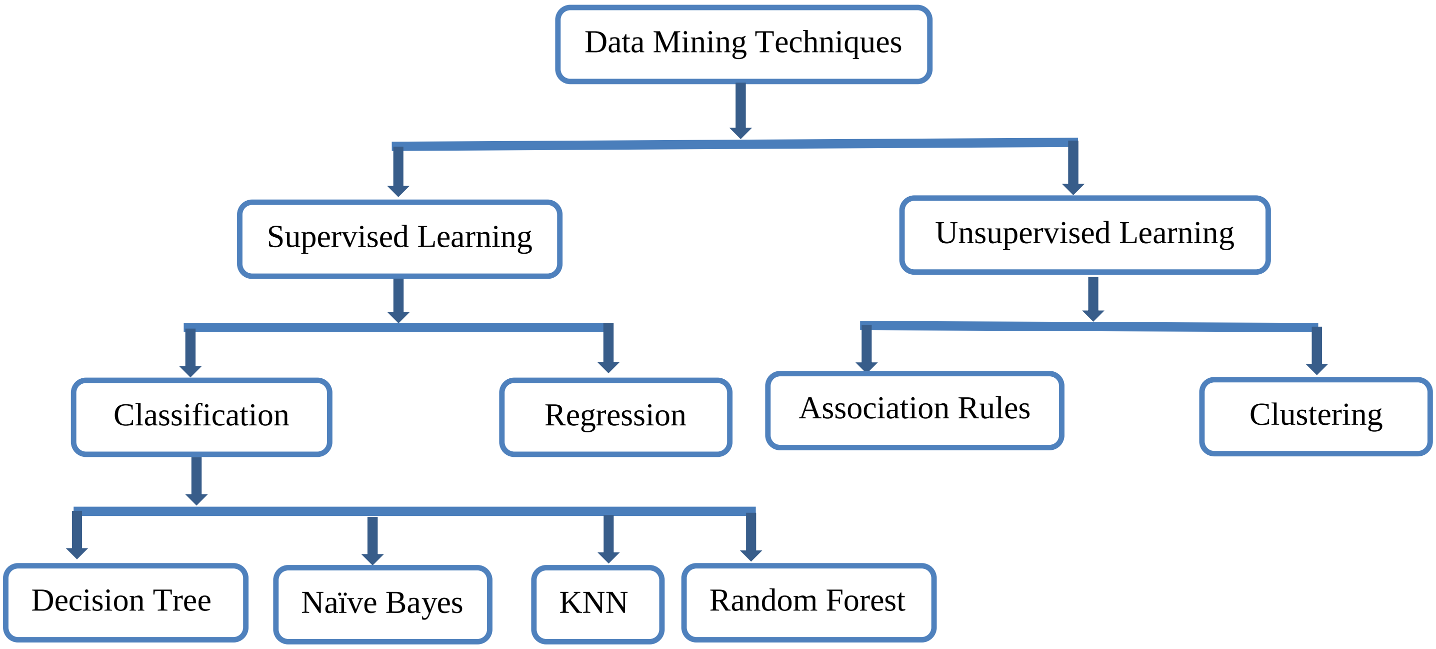


Figure 3-1 – Data Mining Techniques

When considering the classification task, it is coming under supervised technique where each instance belong to a particular class which is specified by the value of a distinctive goal attribute or simply the class attribute. When describing the goal attribute, it is taken in categorical values and each of them corresponding to a class. Each example has mainly two parts. They are:-

- Set of predictor attribute values
- Goal attribute values

The former are used to predict the value of the latter. The predictor attributes should be related for predicting the class of an instance. In the classification task the set of instances being quarried is divided into two commonly exclusive and exhaustive sets. They are called the training set and the test set.

The classification process can be divided into two phases:

- Training -when a classification model is built from the training set
- Testing -when the model is evaluated on the test set

When considering the training phase, the particular algorithm has access to both predictor attributes values as well as the attribute for all examples of the training set. It uses those information in order to build a classification model. This model signifies classification knowledge basically, a relationship between predictor attribute values and classes that permits the forecast of the class of an example given its predictor attribute values.

When considering the testing phase, the test set the class values of the examples is not shown. In the testing phase, the algorithm permitted to see the actual class of the just-classified instance, only after a prediction is made. Maximizing the predictive accuracy obtained by the classification model when classifying examples in the test set hidden during training is, one of the main goals in classification algorithm [10].

3.3 Rapid Miner

Rapid Miner can be considered as a software package which allows text mining, data mining, as well as predictive analytics. In this software, particular user is allowed by program to enter raw data. It includes databases as well as text, which is then repeatedly and logically analyzed on a large scale.

3.4 Chapter Summary

This chapter has discussed about the data mining techniques which was selected to do the analysis. Then, the tool of Rapid Miner is clarified. Rapid Miner is used to analyze the data mining techniques in order to select the most suitable method in credit analysis and appraisals.

CHAPTER FOUR

METHODOLOGY

4.1 Chapter Introduction

Chapter four was discussed about the methodology or the design of the project. This chapter consists with four sections. Firstly, profile of the sample is discussed. Second section is pertaining to discuss input, process and the output of the study. Thirdly, the data preprocessing has been discussed in detail. Fourth section is devoted to the discussion of classification algorithms. Finally summary of the chapter is discussed.

4.2 Profile of the Sample

Profile of this sample includes 1500 records of business loan borrowers.

1500 records include:

- 916 Performing Loans
- 584 Non-performing loans

Class label of the data set has two distinct values

- YES – for performing loans
- NO – for non-performing loans

Attributes of the sample are as below.

Channel, Facility type, Amount, Applicant age, Legal status, business registration, business age, purpose of the business, business sector, PPG ratings, CRIB, Cheque return status, Bank exposure and all bank exposure.

4.3 Hypotheses

Without having a well-structured mechanism to detect non-performing loans in banking sector, Classification algorithms in data mining can be used to sort out this matter. The hypothesis of this study is the classification algorithms which can be used in credit analysis and appraisal process in order to reduce the human errors. In this study, basically four algorithms have been used namely, KNN, Decision Tree, Random Forest and naïve Bayes.

4.4 Input

As the input of this study, data related to Small and Medium scale business loans have been collected from the database in a private bank in Sri Lanka. This project is a special one as it targets to give loan facilities to clients within three working days.

4.5 Output

Main output of this study will be a designed simulation which can be used to identify non-performing loans. Basically, several classification models will be created by using selected classification algorithms. The model which has highest accuracy will be selected to detect the non-performing loans in this selected data set.

4.6 Process

In this study, the main process is focusing on identifying the most appropriate classification algorithm which can be used in credit analysis and appraisal process in order to get the accurate decision by minimizing human errors. Within this main process, several sub processes are included. Those are data preprocessing, data selection and the evaluation.

4.7 Data Preprocessing

Data preprocessing can be defined as a data mining technique which helps to transform raw data into understandable format. Incomplete or inconsistent data are there in the real world. As well as that data can be lack in certain behaviors or trends and also to contain many errors. So, this data preprocessing can be considered as a proven method of resolving these problems. Data preprocessing prepares raw data for further processing.

Data preprocessing is a major step in data mining and it cannot be neglected. The data collection is regularly a process lightly controlled or resulting in out of range values. For an example, there can be impossible data combinations such as Gender: Male; Pregnant: Yes as well as some missing values etc. Analyzing data that has not been carefully screened for such difficulties can make misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. If there are so many data which are not relevant and redundant information present or noisy and unreliable data, then knowledge discovery is more difficult to conduct. Data preparation can take considerable amount of processing time.

4.8 Classification Algorithms

In this study, four algorithms are used to analyze the data.

Classification can be defined as the task of simplifying recognized structure to apply to novel data. For example, an email program might attempt to classify an email as legitimate or spam.

Common algorithms include

□ Decision Tree

□ K-Nearest Neighbor

□ Random Forest

□ Naive Bayesian Classification

□ Decision Tree

Decision Tree is one of the algorithms in classification. This Decision Tree Classifier includes of a decision tree which is generated on the basis of examples. A decision tree is a kind of classifier which can be expressed as a recursive partition of the instance space. The decision tree [4] includes some nodes which support to form a rooted tree. This node is called root and it is a directed tree that has no incoming edges. Apart from that, all the other nodes have precisely one incoming edge. An internal or test node is a node with outgoing edges. All other nodes are called leaves. They are called as terminal or decision nodes as well. In a decision tree, each and every internal node splits the instance space into two or more sub-spaces a certain discrete function of the input attributes values.

□ K-Nearest Neighbor

K-Nearest neighbor is one of the main algorithms which is coming under the classification. It is also called as KNN. This classifiers are based on learning by analogy. By using n dimensional numeric attributes, the training samples are described. Each and every sample represents a point in an n-dimensional space. According to that, all of the training samples are stored in an n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier finds the pattern space for the k training samples which are closest to the hidden sample.

□ Random Forest

Random forest is one of the main algorithms coming under classification technique. It is a supervised learning algorithm which can be used for both classification and regression. However, it is mainly used for classification problems. According to the normal environmental factors a forest is made up of trees and more trees means more robust forest. By using this same scenario, random forest algorithm makes decision trees on data samples and after that, it gets the prediction from each of them and finally selects the best solution by means of voting. It is a kind of cooperative method which is better than a single decision tree as it helps to reduce the over-fitting by averaging the result.

□ Naive Bayesian Classification

Naïve Bayesian classification is one of the main algorithms in classification. A Bayesian network (BN) includes a directed, acyclic graph as well as a probability distribution for each and every node in that graph given its instant predecessors [7]. A Bayes Network Classifier is defined as a algorithms which is based on a bayesian network which represents a joint probability distribution over a set of categorical attributes. This classifier basically includes two parts, the directed acyclic graph G consisting of nodes and arcs and the conditional probability tables. The nodes represent attributes whereas the arcs indicate direct dependencies.

4.9 Chapter Summary

The research methodology were addressed by this chapter. Profile of the sample is 1500 records of loan borrows with multiple attributes. Firstly, sample details have been discussed. Then, four classification algorithms have been clarified clearly.

CHAPTER FIVE

RESEARCH DESIGN AND ANALYSIS

5.1 Chapter Introduction

Chapter five focuses on analysis and design of the project. It discusses the main approach of analyzing different data mining techniques to select the most appropriate data mining technique/s which can be used for credit analysis and appraisals in order to get the accurate decisions by minimizing the human errors.

5.2 Research Design

In this study the researcher has used data mining techniques to identify the hidden patterns among the parameters which is required for credit analysis. Identifying the hidden patterns helps to select the most appropriate data mining technique which is needed for credit analysis in order to minimize the human errors. According to this research design, it helps to identify the non-performing loans at the point of giving loans. In this study, scientific research method has been used to achieve the predetermined research purpose. The main steps of the scientific research is,

- Observation
- Background of study
- Building hypothesis
- Hypothesis testing
- Draw conclusions
- Evaluation

According to the scientific method, this research study starts with observation and the background study of the existing problem of credit analysis and appraisal process in banking sector. Observation was done based on the customer segmentation of the bank, particular project the bank is using to give loan facilities to customers and the attributes which have been used to get the credit decision. The, previous research articles have been thoroughly observed and reviewed in order to find out the solution for this kind of problems. First, main parameters have been identified which is required for credit analysis and appraisals in banking sector. Then the data set has been analyzed using selected classification algorithms. Basically four algorithms have been selected to analyze the data. The results generated by the rapid minor studio using different classification techniques are further analyzed considering the confusion matrix results. Final conclusions were developed using those results.

5.3 Detailed Research Design

The basic research question of this study is to find out the limitations of the existing system which is currently used for credit appraisal process in banking sector. Secondly, by analyzing all the data using classification algorithm, identifying the most suitable data mining technique for credit analysis is the second research question. This has been done by using several steps.

- Data Gathering – Gathered data from a private bank in Sri Lanka related to Small and Medium scale business loans
- Data preprocessing – Preprocessing has been done using three stages namely, data cleaning, data transformation and data reduction.
- Data Analyzing – Analyze the data using classification algorithm
- Model Creation – Created eight models under cross validation and split validation
- Evaluation – Comparison has been done about the performances of each model.
- Draw the conclusion – Conclusion has been made related to the classification models which has the best accuracy.

5.3.1 Sub Research Question One

What are the limitations on the methods and techniques which are currently used for credit analysis and appraisals?

5.3.2 Sub Research Question Two

What are the most appropriate data mining technique/s which can be used for credit analysis and appraisals in order to get the accurate decisions by minimizing the human errors?

5.4 Chapter Summary

This chapter has discussed about of the research design and analysis of this study. Under the detailed research design, every step of the study has been clarified as a summary. Moreover, this chapter focused on how research questions are structured within the research. Basically, there are two main research questions which support to achieve the aim of the study.

.

CHAPTER SIX

IMPLEMENTATION

6.1 Chapter Introduction

This chapter discuss about the implementation of the overall project by giving solutions to each research questions in the study. Firstly, data preprocessing has been done in three stages, data cleaning, data transformation and data reduction. Data Transformation and data reduction have been clarified under discretization and correlation matrix. Then, cross validation and split validation have been discussed.

6.2 Solution for Research Question One

The first research question is to identify the limitations on the methods and techniques which are currently used for credit analysis and appraisals.

The bankers are deciding whether borrower is suitable for receiving the particular amount of loan and it's about the trustworthiness. This process is called as credit appraisal or credit analysis. Currently this process is a manual process and whole decision is based on the credit officers' opinions. This particular process contains around twenty criteria. At the beginning, there's an initial gate to check all the relevant documents such as business registration, address proofing documents, bank statements, audited accounts, title reports, legal documents (plan, deed, extracts) and invoices. If these documents are completed only, the file goes to credit appraisal process. In that process, Procedure Process Guide (PPG) is used as a manual. All these things are done by credit officer with his or her experience. So human errors can be happen which is caused to reduce the accuracy of the decision. [1].

6.3 Solutions for Research Question Two

The second research question is to assess the most appropriate data mining technique/s which can be used for credit analysis and appraisals in order to get the accurate decisions by minimizing the human errors.

As the solution of this research question, several algorithms have been analyzed under classification method using Rapid Miner.

6.3.1 Data Preprocessing

Data preprocessing can be used to transform the raw data in a useful and efficient format.

Data Preprocessing Methods used:-

- Data Cleaning – Filling the missing values
- Data Transformation - Discretization
- Data Reduction – Attribute Selection

Discretization was done to the numerical values of the data set. There were two numerical attributes called CID and the Amount. Under the discretization, values were divided into few ranges. Refer Appendix 9.1

Table 6-1 Discretization

Numerical Data	Number of Bins	Minimum Value	Maximum Value	
CID	5	0	1500	Interval
Amount	4	0	12	Interval

6.3.2 Attribute Selection

In this research study, there are basically 17 attributes. Among those 17 attributes, some special attributes have been selected using correlation matrix by considering the relationship between independent variables and the dependent variable. Attributes of the research are as bellows.

Class Attribute: Paid Loans (Yes – Paid Loans. No – Not paid loans)

- CID - Special number given to each customer
- Amount - Between 5 Laks to 100 Laks
- Channel - Amber (Based on accountant's certification , Green (Based on bank statements & current accounts)
- Facility Type - (POD – Permanent Over Draft), TL – Term Loans, BG – Bank Guarantee)
- Applicant Age -Between 18 to 65
- Legal Status - Individual, Partnership, Proprietorship, Private limited,

- Business Registration - Yes or No
- Business Age - Above or below three years
- Purpose - Working capital, Construction, Purchase
- Business Sector - Retail, Wholesale
- PPG Rating - Product Process Guideline Rate [> 80 – 1, > 60 – 2, < 60 – 3]
- CRIB - Regular or Irregular
- Cheque Return Status - Yes or No
- Bank Exposure - Above 10 Million or Below 10 Million
- All Bank Exposure - Above 25 Million or below 25 Million
- LTV (Loan To asset Value) - Within LTV or Exceed LTV
- FOIR / DSCR / Interest Cover - Acceptable or Not Acceptable

6.3.3 Correlation Matrix

Correlation matrix helps to identify the correlation between all attributes and it can produce a weights vector based on these correlations. Correlation is a kind of statistical technique which can helps to show whether and how strongly pairs of attributes are related.

Correlation Matrix has been used to identify the special parameters which have strong positive relationship towards the label of the data set. According to the results received from the correlation matrix, out of 17 parameters there were 12 special parameters which have a strong positive relationship towards the label. The result of the correlation matrix are as below.

- CID
- Amount
- Facility Type
- Legal Status
- Business Registration
- Business Age
- Business Sector
- PPG Rating
- CRIB

- Bank Exposure
- All Bank Exposure
- FOIR / DSCR / Interest Cover

6.3.4 Model Creation

In this study, the researcher has selected four different algorithm techniques under the classification method. Selected algorithms are KNN, Naïve Bayes, Decision Tree and Random Forest. To analyze the algorithm Rapid Miner tool has been used. Data set has been divided into two parts as training data set and testing data set. Same algorithms are analyzed using cross validation and split validation.

6.3.5 Cross Validation

Cross-validation can be identified as a statistical method which is used to estimate the skill of machine learning models. This statistical technique is commonly used in applied machine learning to compare and also select a specific model for a given predictive modeling problem as it is easy to understand as well as to implement. The results which is obtained in skill estimates which generally have a lower bias when comparing with other methods.

In this procedure of cross validation, there is a single parameter called K, and that the number of groups that a given data sample is to be split into. This procedure is called as k-fold cross-validation. When a particular value for k is selected, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

That method can be used to refer a limited sample in order to evaluate how the model is expected to perform in general when it used to make predictions on data without using during the training of the model. This is a kind of popular method as it is easy and simple to understand since it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.

6.3.6 Split Validation

This split validation operator can be identified as a nested operator. Basically it has two sub processes. They are training sub process and testing sub process. The training sub process is basically used for building or learning a model. After that, the trained model is applied in the testing sub process. During the testing phase, the performance of the model is measured as well.

The input Example Set is partitioned into two subsets. Among those two subsets, one subset is used as the training data set and the other one is used as the test data set. Size of these two subsets are depend on the different parameters used. The model is learned on the training set and is then applied on the test set. This has been conducted in a single iteration, as compared to the Cross Validation operator that iterates a number of times using different subsets for testing and training purposes.

6.4 Chapter Summary

This chapter has discussed about solutions for each research questions. Under that, attribute selection, correlation matrix and model creation of the study have been done. The model is created based on classification algorithms. Moreover, cross validation and split validation have been presented separately.

CHAPTER SEVEN

EVALUATION

7.1 Chapter Introduction

This chapter focuses on the evaluation of the data mining techniques used for credit analysis decisions of this project. Results taken from the data mining techniques have been described here. Then, the comparison have been done between the performances of each algorithm. Finally, the external factors which can be affected to the credit decision have been evaluated.

7.2 Evaluation for Classification

In this study, collected data set have been trained using different classification techniques namely Decision Tree, Naïve Bayes, KNN and Random Forest by the help of Rapid Miner tool. In order to select the special attributes which have positive relationship towards the dependent variable, confusion matrix have been be used. Confusion matrix can be evaluated by using various measurements such as accuracy, recall and precision. These measurements and their definition are given in the below Table.

Table 7-1 – Confusion Matrix

Measure	Formula	Meaning
Precision	TP / (TP + FP)	The percentage of positive predictions that are correctly predicted.
Recall / Sensitivity	TP / (TP + FN)	The percentage of positive labeled instances that were predicted as positive labels.
Specificity	TN / (TN + FP)	The percentage of negative labeled instances that were predicted as negative labels.
Accuracy	(TP + TN) / (TP + TN + FP + FN)	The percentage of predictions those are correct.

Confusion matrix basically gives the prediction results of classification problems. As per the above table, both correct and incorrect predictions have been summarized by count values. By using this confusion matrix, an idea can be taken about the errors and types of errors made by classifier.

7.3 Summary of the Accuracy of Classification Algorithm

The summary of the accuracy level of each algorithm has been presented in below table.

Table 7-2 - Accuracy of Classification Algorithm

Algorithm	Previous accuracy level	Accuracy of special attribute
Cross Validation Decision Tree	63.89%	63.22%
Cross Validation KNN	54%	57.11%
Cross Validation Naïve Bayes	64.33%	64.44%
Cross Validation Random Forest	61.78%	59.44%
Split Validation Decision Tree	68.39%	67.05%
Split Validation KNN	55.91%	55.91%
Split Validation NB	65.89%	66.39%
Split Validation Random Forest	65.89%	63.29%

According to the accuracy levels of classification algorithms, decision tree and Naïve Bayes algorithms have highest accuracy levels and it indicates the significant effect in making credit decisions.

Results of cross validation analysis are as below.

Table 7-3 - Cross Validation Results

Technique	Accuracy	Precision	Recall	Specificity	F-measure
KNN	57.11	63.18	71.45	40.67	67.06
Naïve Bayes	64.44	64.74	91.82	14.06	75.93
Decision Tree	63.22	64.58	88.18	19.63	74.55
Random Forest	59.44	63.54	78.91	31.78	70.39

According to the table 7.3, the model of Naïve Bayes has taken the highest accuracy levels. Results of Rapid Minor analysis is attached in Appendix C.

Results of split validation are as below.

Table 7-4 - Results of Split Validation

Technique	Accuracy	Precision	Recall	Specificity	F-measure
KNN	55.91	62.75	68.39	43.77	65.44
Naïve Bayes	66.39	66.40	91.01	16.33	76.78
Decision Tree	67.05	67.42	89.10	20.20	76.75
Random Forest	63.23	66.74	79.29	34.38	72.47

According to the table 7.4, the model of Decision Tree has taken the highest accuracy levels. Results of Rapid Minor analysis is attached in Appendix C.

According to the results obtained from the tables, the researcher can clearly identified Decision Tree and Naïve Bayes as the algorithms which have highest accuracy levels. Below graph shows the summary of the accuracy levels of algorithms used in this study.

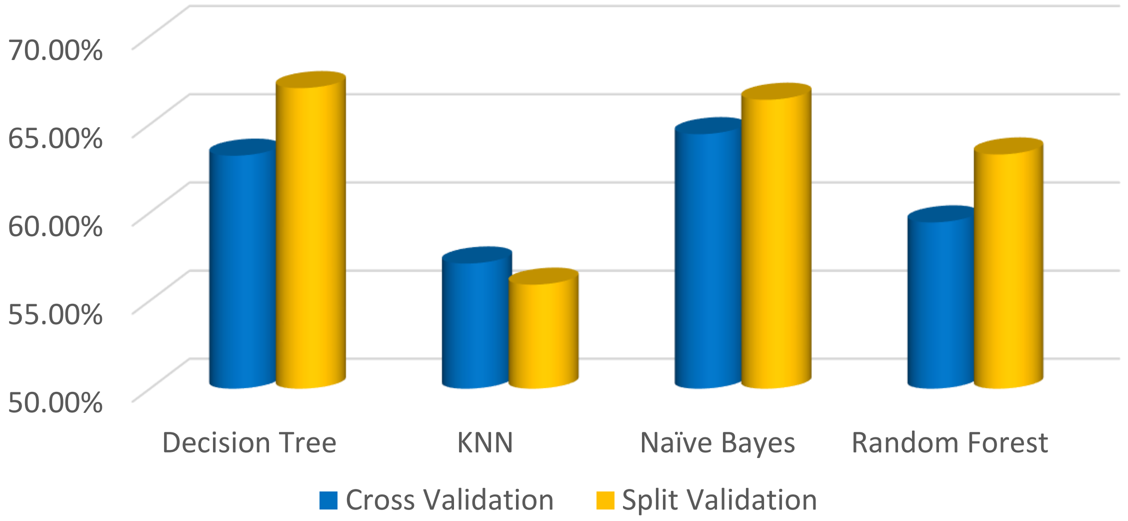


Figure 7-1 – Summary of Accuracy Levels

7.4 External factors affected to the credit decision

When considering the decision making in credit analysis and appraisal process, there are few external factors which may affect to credit decisions.

- Political changes of the country
- Pandemic situations of the country
- Government policies and regulations
- Inflation
- Cultural conflicts of the society

7.5 Chapter Summary

This chapter has discussed the evaluation of the results taken from different algorithms in classification data mining technique. Evaluation results have been described by using the accuracy, recall, precision, specificity and F- measure. All the results have been illustrated by using graphs and tables. Further, external factors are clarified which can be affected to the decisions in credit analysis and appraisal process.

CHAPTER EIGHT

DISCUSSION AND FUTURE WORK

8.1 Chapter Introduction

Chapter eight focuses on presenting a summary of the overall study and future works of this study. Firstly, summary of the study has been presented. Secondly, the limitations of the project is discussed. Finally, the future works which are intended to do have been discussed in this chapter.

8.2 Overview of the research

Banking sector is the one of the main sectors in Sri Lanka. When considering the banking sector, giving loan facilities is one of their main functions, in order to support the individual's as well as the businesses' financial conditions. Therefore, basically they are issuing two types of loans called personal loans and business loans. Bankers giving business loans to small and medium scale businesses and large scale businesses. Giving loan facilities to SME sector is a significant process as SMEs plays a major role in Sri Lankan economy. In this study, loan facilities which have been given to SME businesses have been taken as the sample.

Credit risk is a major risk in the banking sector. Credit risk can be simply defined as the probability or the possibility that a borrower will fail to fulfill his or her obligations to the banker during the period of contract. Credit officer himself has to take this responsibility as all the decision making process is under his or her opinion. In banking sector, credit analysis and appraisal process is a manual process and it is based on the credit officers' opinion. Therefore, human errors can be happen in this process. There are some researches and studies regarding this matter and there is no realistic or adequate solution for this.

When it comes to the new technology in the world, data mining has identified as a novel approach to sort out this kind of issues. Hence, in this study data mining techniques have been used to find out the solution for this problem. Classification algorithms are used to identify the hidden patterns among the attributes of credit analysis and appraisal process. By comparing the accuracy of the algorithms based on the relationships among the attributes, best algorithms are selected in order to get the accurate decisions in credit analysis and appraisals to minimize the human errors.

8.3 Limitations

According to this research study, credit decisions are based on credit officers' opinion. In order to minimize the human errors, data mining techniques can be used as a decision support system. Apart from that, there are some external factors which are affected to the credit decision such as political decisions, government rules and regulations, inflation, pandemic situation etc. Those external factors cannot be tracked or controlled through data mining techniques. Further, this study has conducted using only classification algorithms in data mining techniques.

8.4 Future work of the project

This study has been conducted with special reference to SMEs in business loan category. Hence, further researches can be conducted based on the personal loans or large scale business loans in banking sector. Since the parameters are different, the results can be changed based on that. Further, the researchers can take financial institutes other than banks as their sample, in order to conduct this type of study. Since those institutes have different parameters with different credit analysis process, the results can be altered based on that. Moreover, other data mining techniques can be used for analysis other than classification techniques.

8.5 Chapter Summary

This chapter discussed about the overview of the research. It has given the summary of the research problem and the solutions made from the research study. Then the limitations and the future research works of the project are described.

REFERENCES

[1] C.L Huang a,*, Mu-Chen Chen b, Chieh-Jen Wang c Credit scoring with a data mining approach based on support vector machines Expert Systems with Applications 33 (2007) 847–856

[2] B. Desai and Anita Desai, "The Role of Data mining in Banking Sector", IBA Bulletin, 2004

[3] Dr. M. L. Bhasin, “Data Mining: A Competitive Tool in the Banking and Retail Industries”, The Chartered Accountant October 2006.

[4] H.Newton, Newton’s Telecom Dictionary, CMP Books, http://www.cmpbooks.com

[5] K.I.Moin, Dr. Q.B.Ahmed ,International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 2,Mar-Apr 2012, pp.738-742, Use of Data Mining in Banking

[6] P.Sundari, Dr.K.Thangadurai “An Empirical Study on Data Mining Applications”, Global Journal of Computer Science and Technology, Vol. 10 Issue 5 Ver. 1.0 July 2010.

[7] W.A Grier, Credit analysis for financial institutions, 2007

[8] U.Mishara & Praneetam Naidu, “A Study on Credit Risk Management and Appraisal Process at Punjab National Bank, Nagpur”, International Journal of Multifaceted and Multilingual Studies

[9] S.García , J.Luengo, F.Herrera, “Data Preprocessing in Data Mining”.

[10] S.Neelamegam, Dr.E.Ramaraj, Classification algorithm in Data mining: An OverviewInternational Journal of P2P Network Trends and Technology (IJPTT) - Volume 3 Issue 5 September to October 2013.

[11] White Paper, 2002, National Strategy for Small and Medium sector Development, Task Force for Small and Medium Enterprise Sector Development Program, Sri Lanka

[12] C.C.Williams, (2009) Informal entrepreneurs and their motives: a gender perspective. International Journal of Gender and of Entrepreneurship.

[13] A.Gamage (2003) Small and Medium Enterprise Development in Sri Lanka.

[14] L.Deelen, M.Dupleich, L.Othieno, and O.Wakelin, (2003). Leasing for Small and Micro Enterprises: A guide for designing and managing leasing schemes in developing countries. International Labour Organization.

[15] C.Wilcox, (1971). Public policies toward business. RD Irwin.

[16] Jurgita, B. (2011). An overview of the European leasing market in 2009. World Leasing Yearbook (edited by Adrian Hornbrook); A Euromoney Publication 2011, 46-55.

[17] Aboobyda J.H and Tarig M.A (2016). Developing prediction model of loan risk in banks using data mining, Machine Learning and Applications: An International Journal (MLAIJ).

[18] Tomar, Divya and Sonali Agarwal. “A survey on Data Mining approaches for Healthcare”, International Journal of Bio –Science and Bio –Technology 5.5 (2013

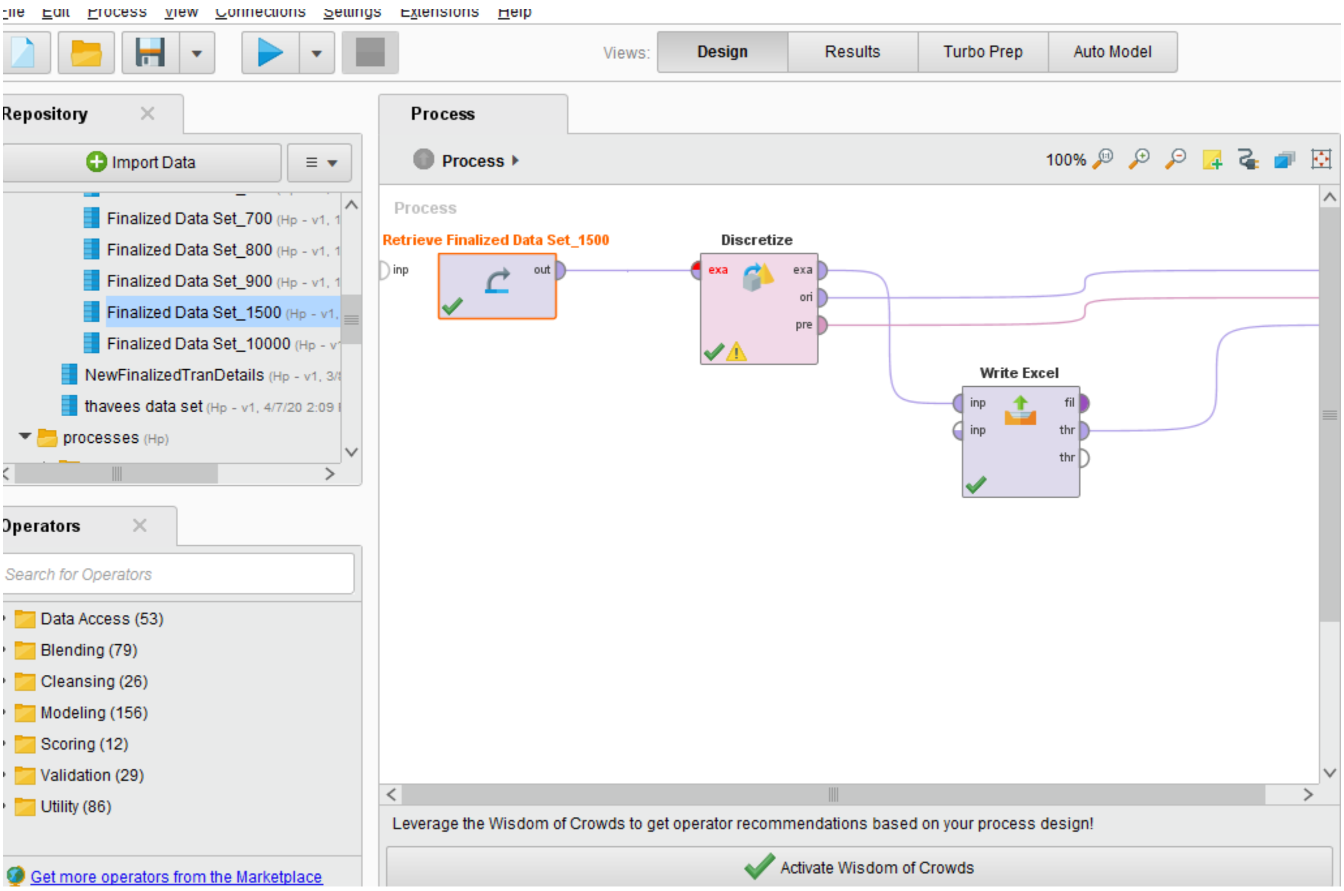
[19] M.W. Joshua, The effect of 7cs credit appraisal model on the level of non-performing advances of commercial banks in Kenya (2010)

[20] A. Dzelihodzic, D. Donko, Data Mining Techniques for Credit Risk Assessment Task.

[21] M.de M.Sousa, R.S.Figueiredo, Credit analysis using data mining: application in the case of a credit union (2014).

APPENDICES

9.1 Appendix A -Data Preprocessing -Discretization



Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Figure 9-1 – Discretization

9.2 Appendix B – Correlation Matrix

Attribute Selection

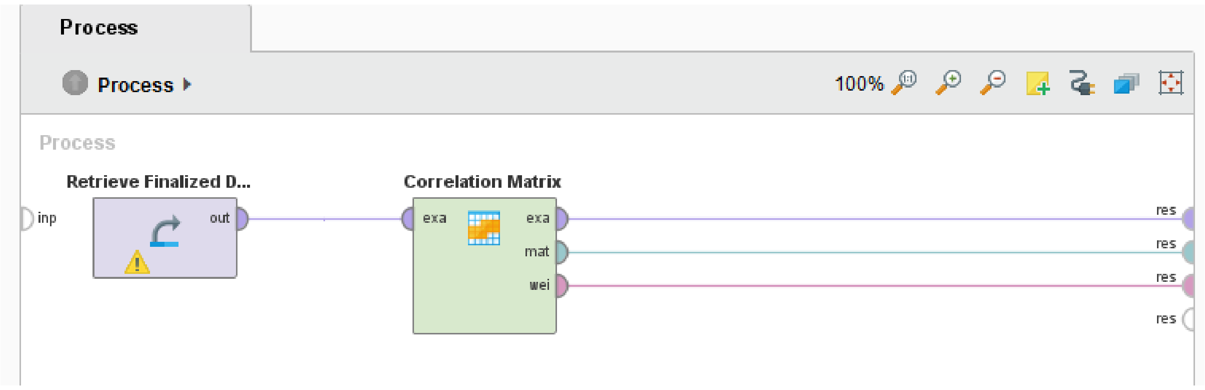
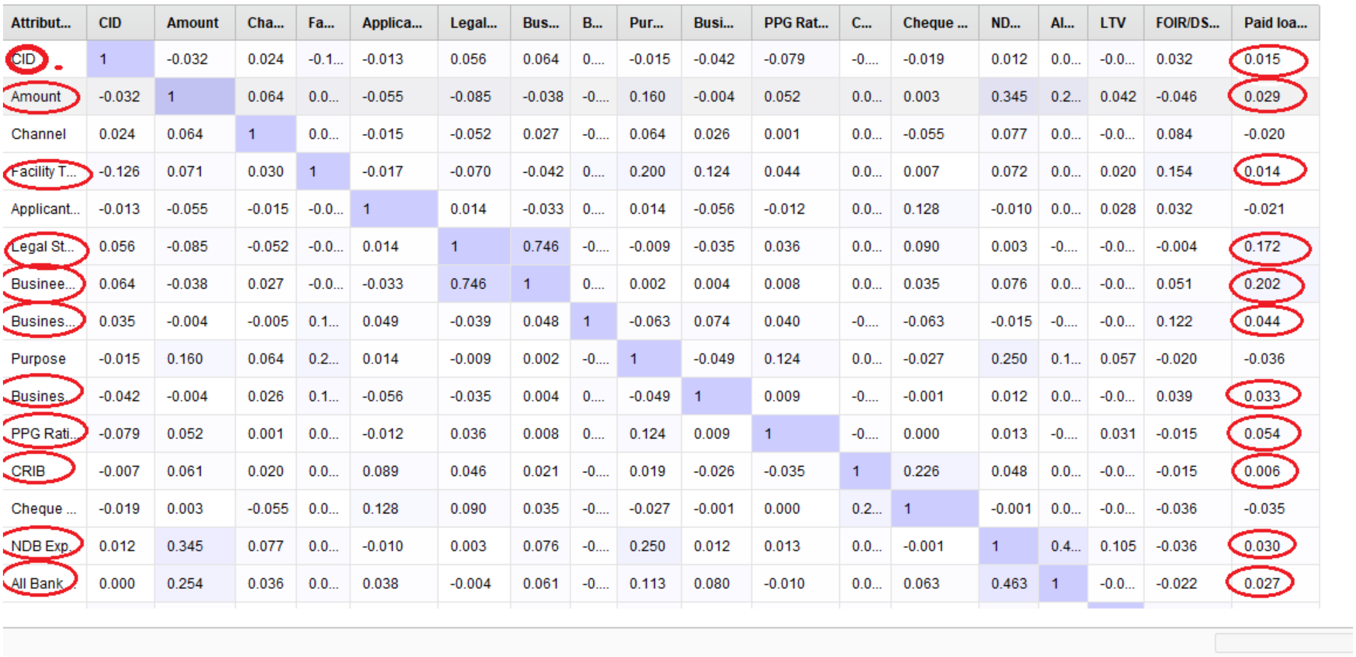
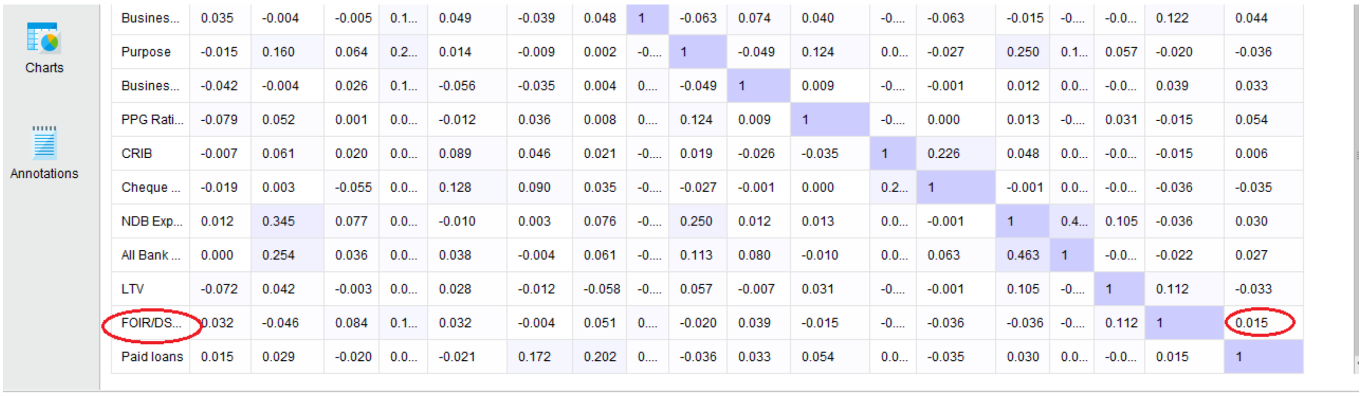


Figure 9-2 - Attribute Selection



Attribut...	CID	Amount	Cha...	Fa...	Applica...	Legal...	Bus...	B...	Pur...	Busi...	PPG Rat...	C...	Cheque ...	ND...	Al...	LTV	FOIR/DS...	Paid loa...
CID	1	-0.032	0.024	-0.1...	-0.013	0.056	0.064	0....	-0.015	-0.042	-0.079	-0....	-0.019	0.012	0.0...	-0.0...	0.032	0.015
Amount	-0.032	1	0.064	0.0...	-0.055	-0.085	-0.038	-0....	0.160	-0.004	0.052	0.0...	0.003	0.345	0.2...	0.042	-0.046	0.029
Channel	0.024	0.064	1	0.0...	-0.015	-0.052	0.027	-0....	0.064	0.026	0.001	0.0...	-0.055	0.077	0.0...	-0.0...	0.084	-0.020
Facility T...	-0.126	0.071	0.030	1	-0.017	-0.070	-0.042	0....	0.200	0.124	0.044	0.0...	0.007	0.072	0.0...	0.020	0.154	0.014
Applicant...	-0.013	-0.055	-0.015	-0.0...	1	0.014	-0.033	0....	0.014	-0.056	-0.012	0.0...	0.128	-0.010	0.0...	0.028	0.032	-0.021
Legal St...	0.056	-0.085	-0.052	-0.0...	0.014	1	0.746	-0....	-0.009	-0.035	0.036	0.0...	0.090	0.003	-0....	-0.0...	-0.004	0.172
Businee...	0.064	-0.038	0.027	-0.0...	-0.033	0.746	1	0....	0.002	0.004	0.008	0.0...	0.035	0.076	0.0...	-0.0...	0.051	0.202
Busines...	0.035	-0.004	-0.005	0.1...	0.049	-0.039	0.048	1	-0.063	0.074	0.040	-0....	-0.063	-0.015	-0....	-0.0...	0.122	0.044
Purpose	-0.015	0.160	0.064	0.2...	0.014	-0.009	0.002	-0....	1	-0.049	0.124	0.0...	-0.027	0.250	0.1...	0.057	-0.020	-0.036
Busines.	-0.042	-0.004	0.026	0.1...	-0.056	-0.035	0.004	0....	-0.049	1	0.009	-0....	-0.001	0.012	0.0...	-0.0...	0.039	0.033
PPG Rati...	-0.079	0.052	0.001	0.0...	-0.012	0.036	0.008	0....	0.124	0.009	1	-0....	0.000	0.013	-0....	0.031	-0.015	0.054
CRIB	-0.007	0.061	0.020	0.0...	0.089	0.046	0.021	-0....	0.019	-0.026	-0.035	1	0.226	0.048	0.0...	-0.0...	-0.015	0.006
Cheque ...	-0.019	0.003	-0.055	0.0...	0.128	0.090	0.035	-0....	-0.027	-0.001	0.000	0.2...	1	-0.001	0.0...	-0.0...	-0.036	-0.035
NDB Exp.	0.012	0.345	0.077	0.0...	-0.010	0.003	0.076	-0....	0.250	0.012	0.013	0.0...	-0.001	1	0.4...	0.105	-0.036	0.030
All Bank.	0.000	0.254	0.036	0.0...	0.038	-0.004	0.061	-0....	0.113	0.080	-0.010	0.0...	0.063	0.463	1	-0.0...	-0.022	0.027

Figure 9-3 - Correlation Matrix I

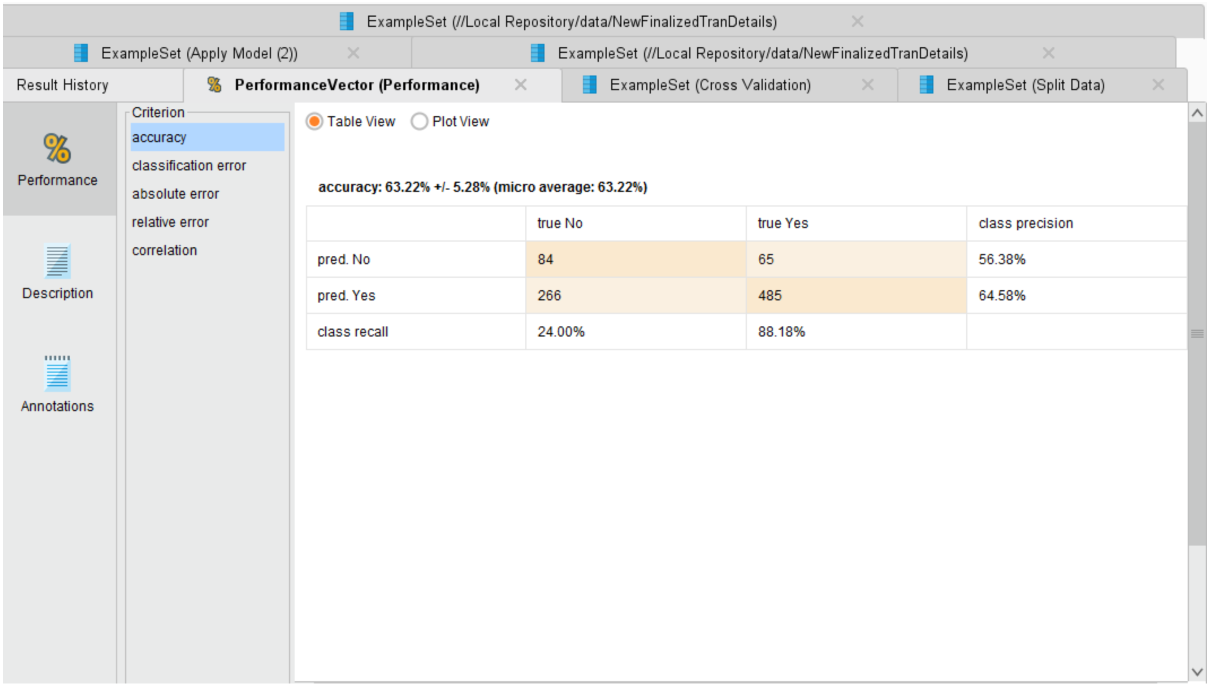


Busines...	0.035	-0.004	-0.005	0.1...	0.049	-0.039	0.048	1	-0.063	0.074	0.040	-0....	-0.063	-0.015	-0....	-0.0...	0.122	0.044
Purpose	-0.015	0.160	0.064	0.2...	0.014	-0.009	0.002	-0....	1	-0.049	0.124	0.0...	-0.027	0.250	0.1...	0.057	-0.020	-0.036
Busines...	-0.042	-0.004	0.026	0.1...	-0.056	-0.035	0.004	0....	-0.049	1	0.009	-0....	-0.001	0.012	0.0...	-0.0...	0.039	0.033
PPG Rati...	-0.079	0.052	0.001	0.0...	-0.012	0.036	0.008	0....	0.124	0.009	1	-0....	0.000	0.013	-0....	0.031	-0.015	0.054
CRIB	-0.007	0.061	0.020	0.0...	0.089	0.046	0.021	-0....	0.019	-0.026	-0.035	1	0.226	0.048	0.0...	-0.0...	-0.015	0.006
Cheque ...	-0.019	0.003	-0.055	0.0...	0.128	0.090	0.035	-0....	-0.027	-0.001	0.000	0.2...	1	-0.001	0.0...	-0.0...	-0.036	-0.035
NDB Exp...	0.012	0.345	0.077	0.0...	-0.010	0.003	0.076	-0....	0.250	0.012	0.013	0.0...	-0.001	1	0.4...	0.105	-0.036	0.030
All Bank ...	0.000	0.254	0.036	0.0...	0.038	-0.004	0.061	-0....	0.113	0.080	-0.010	0.0...	0.063	0.463	1	-0.0...	-0.022	0.027
LTV	-0.072	0.042	-0.003	0.0...	0.028	-0.012	-0.058	-0....	0.057	-0.007	0.031	-0....	-0.001	0.105	-0....	1	0.112	-0.033
FOIR/DS...	0.032	-0.046	0.084	0.1...	0.032	-0.004	0.051	0....	-0.020	0.039	-0.015	-0....	-0.036	-0.036	-0....	0.112	1	0.015
Paid loans	0.015	0.029	-0.020	0.0...	-0.021	0.172	0.202	0....	-0.036	0.033	0.054	0.0...	-0.035	0.030	0.0...	-0.0...	0.015	1

Figure 9-4 - Correlation Matrix II

9.3 Appendix C – Results of Cross Validation and Split Validation

Cross Validation – Decision Tree

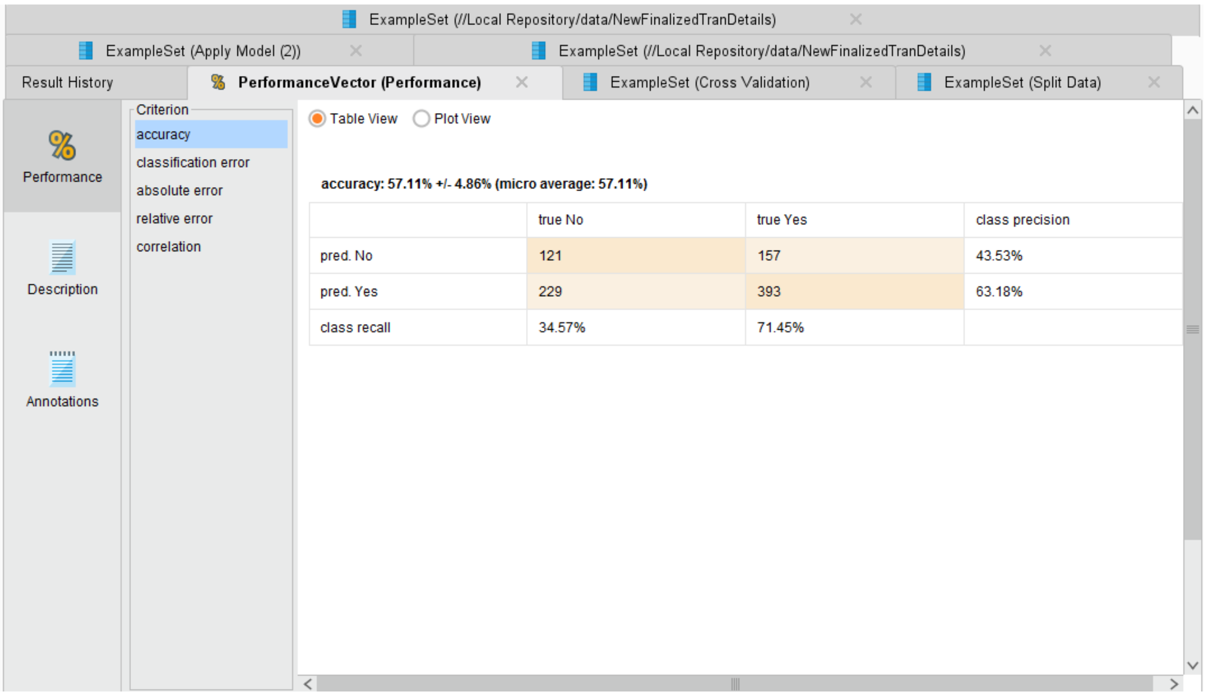

ExampleSet (//Local Repository/data/NewFinalizedTranDetails)
ExampleSet (Apply Model (2))
ExampleSet (//Local Repository/data/NewFinalizedTranDetails)
Result History
PerformanceVector (Performance)
ExampleSet (Cross Validation)
ExampleSet (Split Data)
Performance
Description
Annotations
Criterion
accuracy
classification error
absolute error
relative error
correlation
● Table View
○ Plot View

accuracy: 63.22% +/- 5.28% (micro average: 63.22%)

	true No	true Yes	class precision
pred. No	84	65	56.38%
pred. Yes	266	485	64.58%
class recall	24.00%	88.18%	

Figure 9-5- Cross Validation - Decision Tree

Cross Validation - KNN

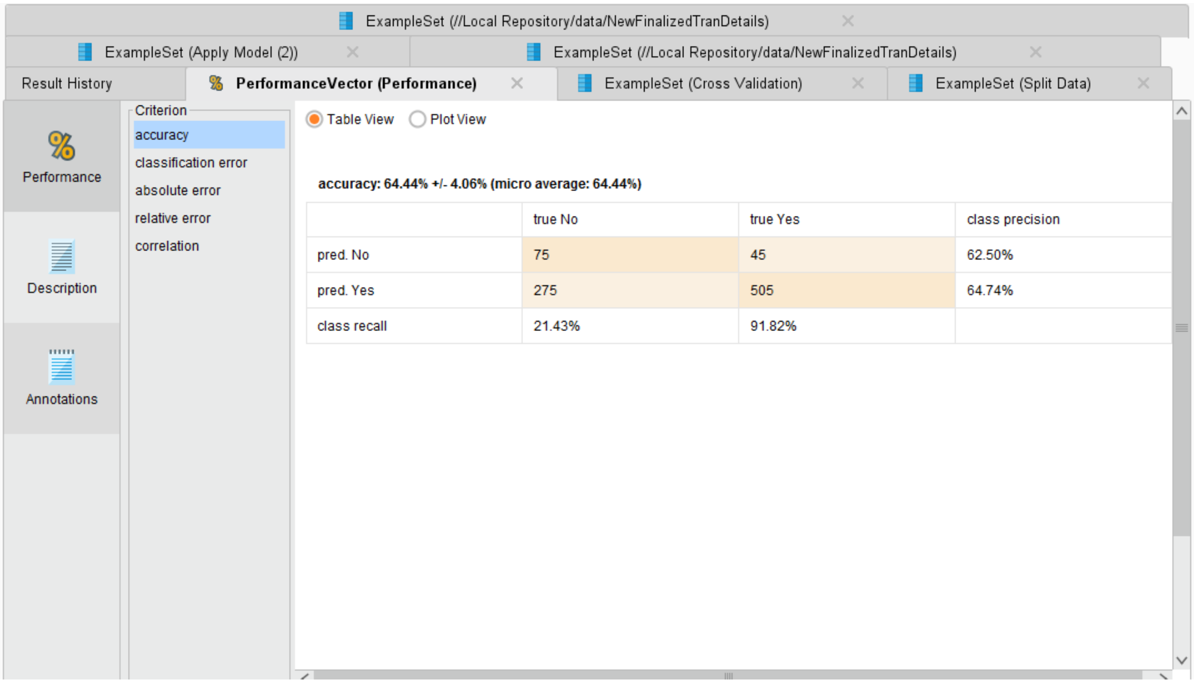

ExampleSet (//Local Repository/data/NewFinalizedTranDetails)
ExampleSet (Apply Model (2))
ExampleSet (//Local Repository/data/NewFinalizedTranDetails)
Result History
PerformanceVector (Performance)
ExampleSet (Cross Validation)
ExampleSet (Split Data)
Performance
Description
Annotations
Criterion
accuracy
classification error
absolute error
relative error
correlation
● Table View ○ Plot View

accuracy: 57.11% +/- 4.86% (micro average: 57.11%)

	true No	true Yes	class precision
pred. No	121	157	43.53%
pred. Yes	229	393	63.18%
class recall	34.57%	71.45%	

Figure 9-6 - Cross Validation - KNN

Cross Validation – Naïve Bayes

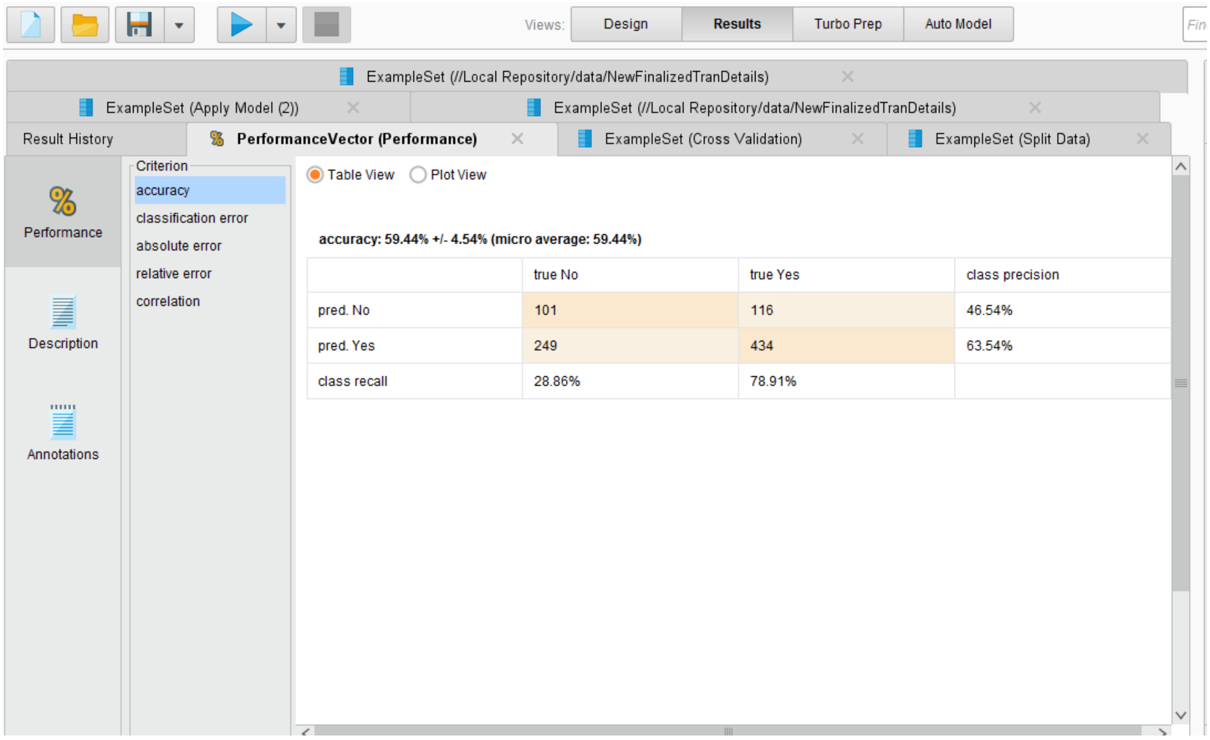

ExampleSet (//Local Repository/data/NewFinalizedTranDetails)
ExampleSet (Apply Model (2))
ExampleSet (//Local Repository/data/NewFinalizedTranDetails)
Result History
PerformanceVector (Performance)
ExampleSet (Cross Validation)
ExampleSet (Split Data)
Performance
Description
Annotations
Criterion
accuracy
classification error
absolute error
relative error
correlation
● Table View ○ Plot View

accuracy: 64.44% +/- 4.06% (micro average: 64.44%)

	true No	true Yes	class precision
pred. No	75	45	62.50%
pred. Yes	275	505	64.74%
class recall	21.43%	91.82%	

Figure 9-7 - Cross Validations - Naive Bayes

Cross Validation – Random Forest

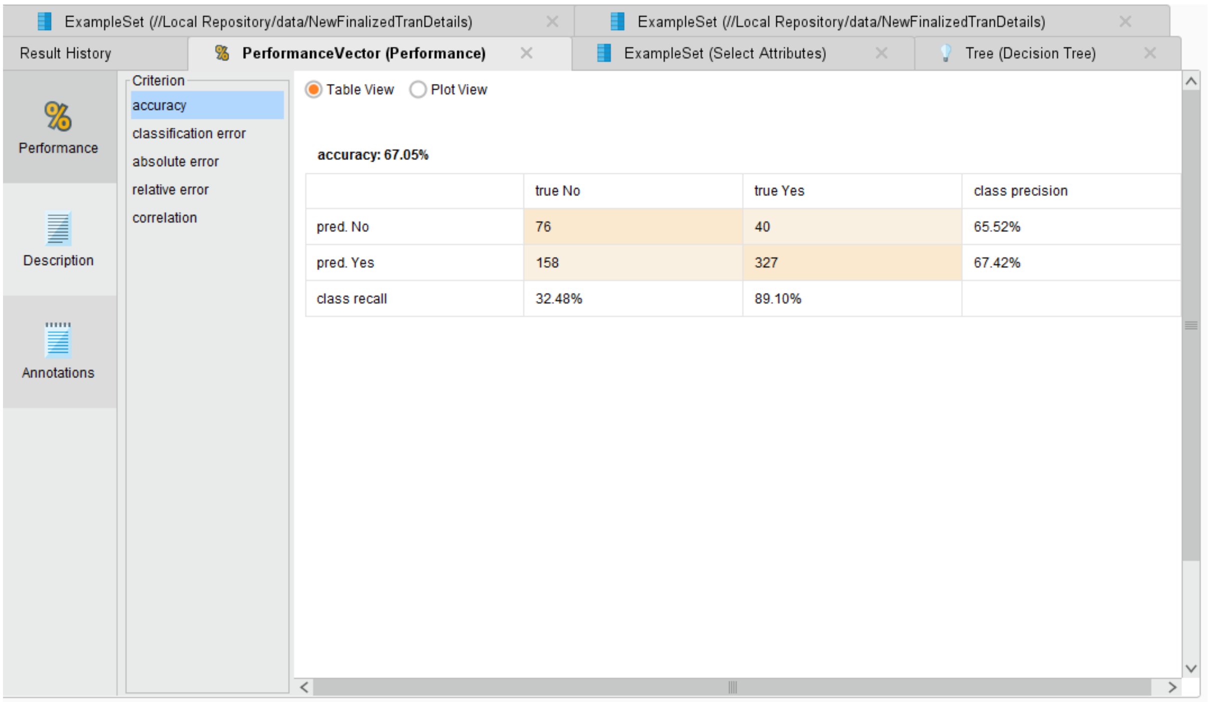

Views:
Design
Results
Turbo Prep
Auto Model
ExampleSet (//Local Repository/data/NewFinalizedTranDetails)
ExampleSet (Apply Model (2))
ExampleSet (//Local Repository/data/NewFinalizedTranDetails)
Result History
PerformanceVector (Performance)
ExampleSet (Cross Validation)
ExampleSet (Split Data)
Performance
Description
Annotations
Criterion
accuracy
classification error
absolute error
relative error
correlation
● Table View
○ Plot View

accuracy: 59.44% +/- 4.54% (micro average: 59.44%)

	true No	true Yes	class precision
pred. No	101	116	46.54%
pred. Yes	249	434	63.54%
class recall	28.86%	78.91%	

Figure 9-8 - Cross Validation - Random Forest

Split Validation – Decision Tree

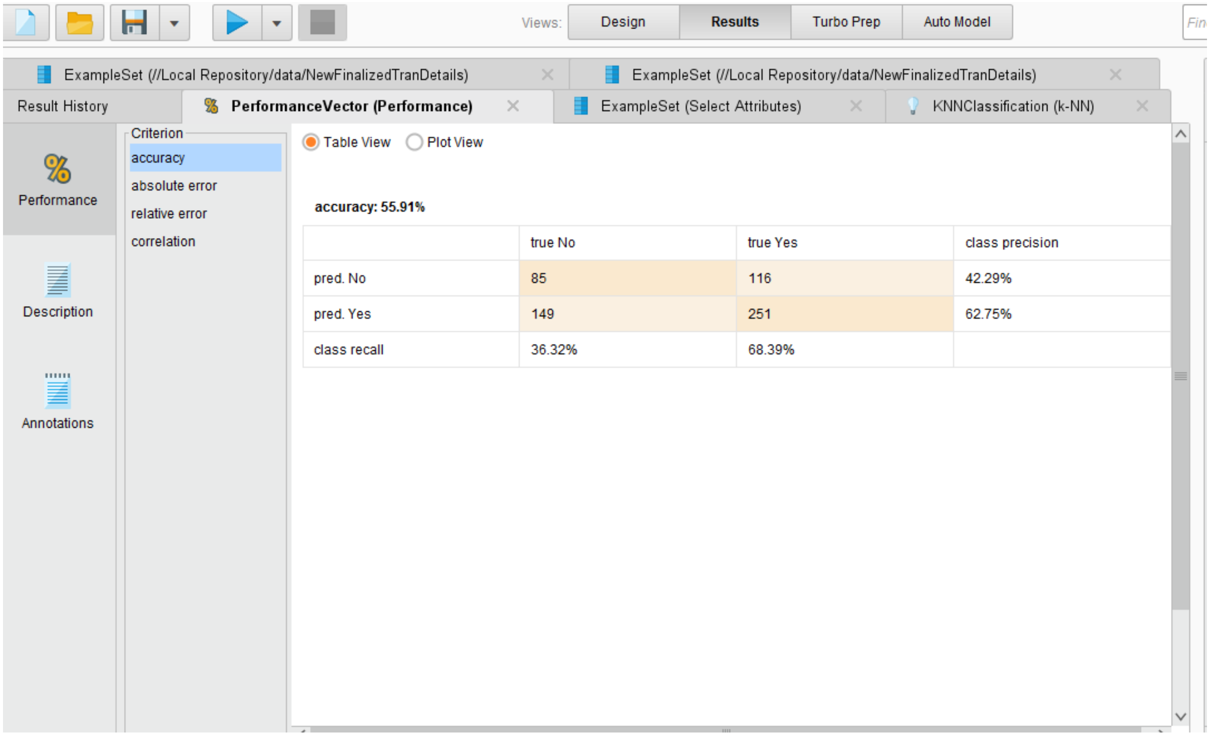

ExampleSet (//Local Repository/data/NewFinalizedTranDetails)
ExampleSet (//Local Repository/data/NewFinalizedTranDetails)
Result History
PerformanceVector (Performance)
ExampleSet (Select Attributes)
Tree (Decision Tree)
Performance
Description
Annotations
Criterion
accuracy
classification error
absolute error
relative error
correlation
● Table View
○ Plot View

accuracy: 67.05%

	true No	true Yes	class precision
pred. No	76	40	65.52%
pred. Yes	158	327	67.42%
class recall	32.48%	89.10%	

Figure 9-9 - Split Validation - Decision Tree

Split Validation – KNN



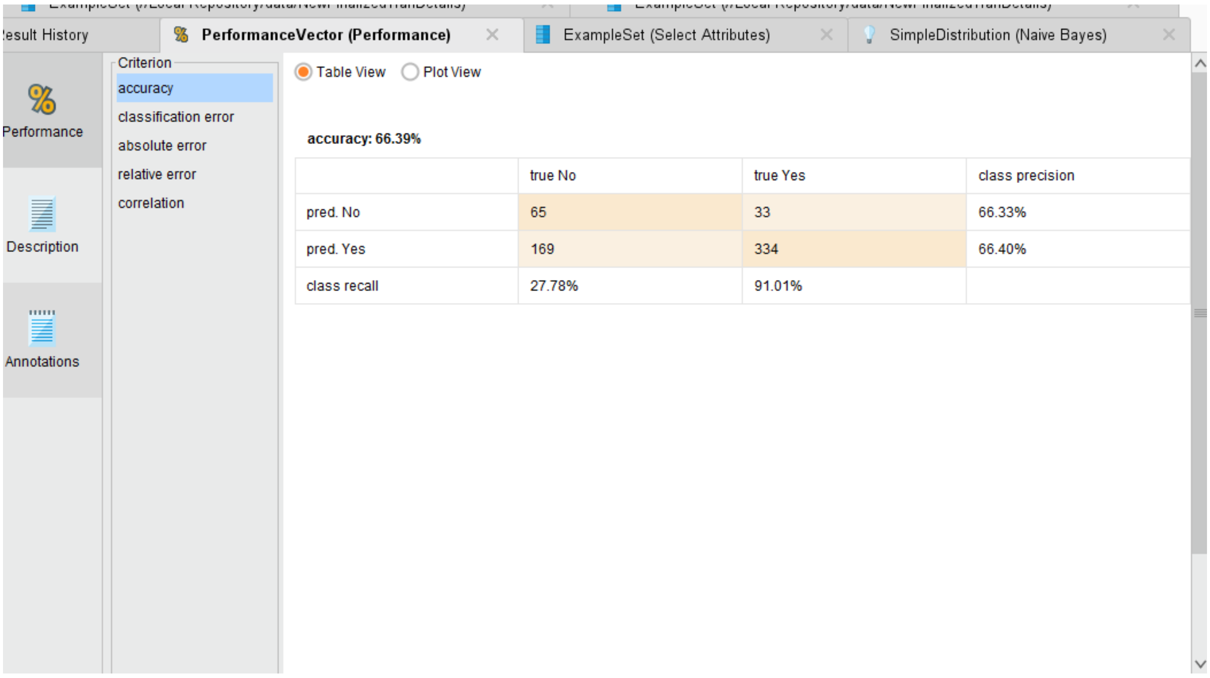
● Table View ○ Plot View

accuracy: 55.91%

	true No	true Yes	class precision
pred. No	85	116	42.29%
pred. Yes	149	251	62.75%
class recall	36.32%	68.39%	

Figure 9-10 Split Validation - KNN

Split Validation – Naïve Bayes


ExampleSet (//Local Repository/data/NewFinalizedTranDetails)
ExampleSet (//Local Repository/data/NewFinalizedTranDetails)
esult History
PerformanceVector (Performance)
ExampleSet (Select Attributes)
SimpleDistribution (Naive Bayes)
Performance
Description
Annotations
Criterion
accuracy
classification error
absolute error
relative error
correlation
Table View
Plot View

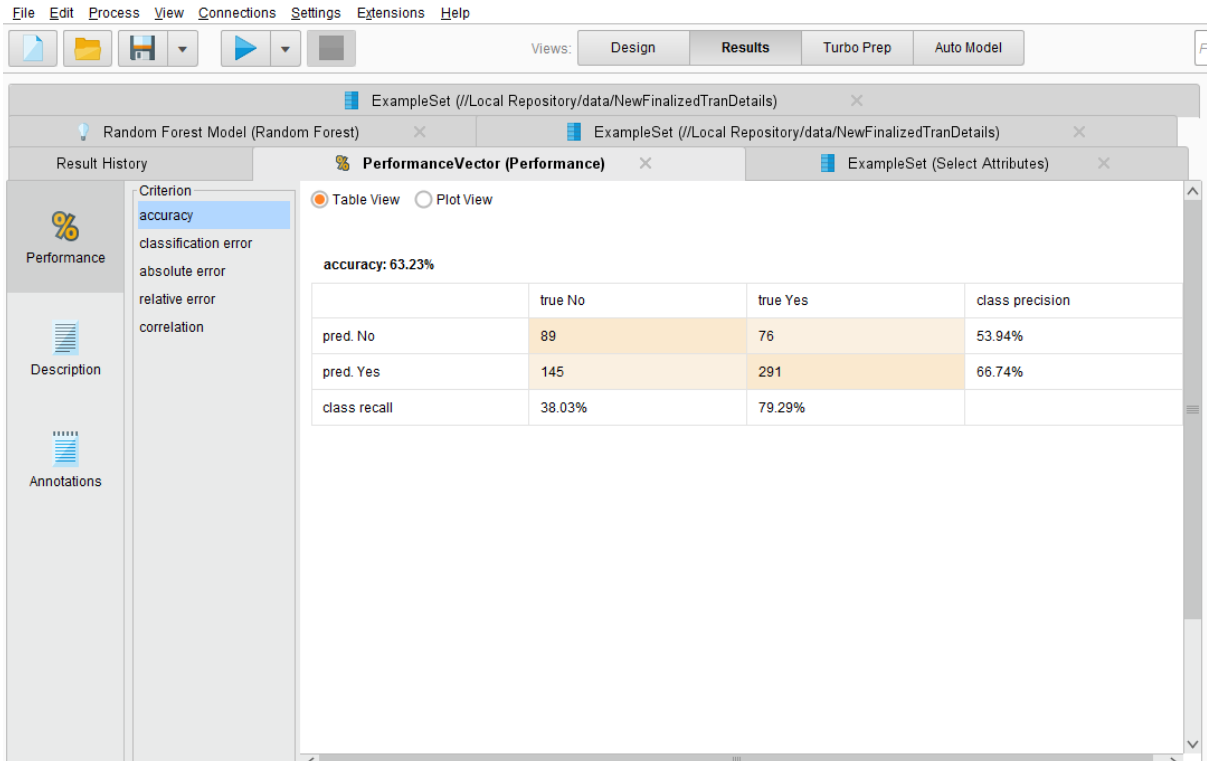
● Table View ○ Plot View

accuracy: 66.39%

	true No	true Yes	class precision
pred. No	65	33	66.33%
pred. Yes	169	334	66.40%
class recall	27.78%	91.01%	

Figure 9-11 - Split Validation - Naive Bayes

Split Validation – Random Forest



File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
ExampleSet (//Local Repository/data/NewFinalizedTranDetails)
Random Forest Model (Random Forest)
ExampleSet (//Local Repository/data/NewFinalizedTranDetails)
Result History
PerformanceVector (Performance)
ExampleSet (Select Attributes)
Performance
Description
Annotations
Criterion
accuracy
classification error
absolute error
relative error
correlation
● Table View ○ Plot View

accuracy: 63.23%

	true No	true Yes	class precision
pred. No	89	76	53.94%
pred. Yes	145	291	66.74%
class recall	38.03%	79.29%	

Figure 9-12-Split Validation - Random Forest