Credit Risk Analysis of Small and Medium-sized Enterprise Loans

G.D.T.D. Chandrasiri
198744F

Faculty of Information Technology
University of Moratuwa
July 2022

Credit Risk Analysis of Small and Medium-sized Enterprise Loans

G.D.T.D. Chandrasiri
198744F

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

July 2022

Declaration

We declare that this thesis 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 Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.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
G.D.T.D Chandrasiri

Signature of Student
UOM Verified Signature

Date:

Supervised by

Name of Supervisor
Dr. S.C. Premaratne

Signature of Supervisor
UOM Verified Signature

Date: 29/07/2022

Dedication

I dedicate my research work affectionately to the following:

My research supervisor, Dr. S.C Premaratne, for providing the valuable guidance on this and for dedicating his valuable time for me,

My parents, for providing me all the courage and their precious love,

My sisters and brother, for helping me as always,

Finally, to all my best friends for supporting me in many ways.

Thank you.

Acknowledgments

First, I would like to convey my sincere gratitude towards my research supervisor, Dr. Saminda Premaratne, Senior Lecturer, Faculty of Information Technology, University of Moratuwa, for his valuable guidance, motivation, and valuable time spent during this research project.

Further, I would like to thank Dr.M.F.M Firdhous who conducted Research Methodology course and Literature Review and Thesis Writing course which laid the foundation for this research and Dr. Chaman Wijesiriwardane who coordinated the research project course.

It is a pleasure to thank all my collegues of the 13th batch of MSc. in IT degree program who encouraged me always.

Importantly, I would like to extend my heartiest thank to my family who gave me their precious love through my entire life. Finally, I must be thankful for all the people whose names are not appeared here, but their diligent effort was very important to make this study a success.

Abstract

Analyzing the credit risk is important in banking systems to ensure that debtors pay the loans regularly, on schedule. The inability of managing the credit risk may lead severe losses in financial institutes. Every financial institute must predict and manage the credit risk to avoid financial crises. Hence, finding an effective method for credit risk analysis is vital. Among various types of loans, Small and Medium-sized Enterprise (SME) loans dominate since SMEs are considered the backbone of any economy. With the higher amount of SME loans, the associated risk also gets increased. SME sector is considered as risky and costly than the large enterprises. The amount of non-performing SME loans has increased at a higher pace throughout the last few quarters in Sri Lanka. Hence, this study analyzes the credit risk of SME loans by using a data set received from a financial institute in Sri Lanka. It recognizes financial and non-financial attributes affecting the credit risk of SME loans. The data mining techniques were used to analyze the data and extract knowledge. It is an emerging technology that provides significant improvements in terms of making accurate decisions. Data mining is used widely for financial analysis since it facilitates knowledge extraction from large data sets and making effective decisions. The study will help financial institutes to predict the credit risk of potential SME borrowers and avoid inefficiencies in the lending process. It identifies the credit risk of debtors in different aspects. Ultimately, it provides valuable insights into effective decision making of an economy.

Keywords: *Credit Risk, SME Loans, Data Mining*

Contents

List of Figures

List of Tables

Chapter 1 : Introduction

1.1 Prolegomena

This chapter introduces the research project on credit risk analysis of small and medium-sized enterprise (SME) loans by using data mining techniques. Credit is the primary source of income in any financial institute around the world. A risk is always associated with the lending process, which is identified as the credit risk. Credit risk poses a substantial exposure to the entire economy of a country. Therefore, it is inevitable to analyze the credit risk and manage it efficiently to avoid economic crisis situations.

Among different types of loans SME loans are considered as more risky and costly than other categories. Simulating SMEs are important since SMEs are a crucial element for any economy. In the Sri Lankan context also, it contributes to a considerable amount of gross domestic production. Though SME provides valuable contribution to the economy, the number of non-performing loans have rapidly increased in the SME sector over the last few years. Hence this study analyzes the credit risk of SME loans by using data mining techniques as an emerging technology. A large data set with nearly 3000 records was obtained from a leading financial institute in Sri Lanka. It was analyzed by applying different data mining techniques. The proposed model will predict the credit risk of potential SME loan borrowers in few different aspects. This study will be helpful in analyzing and managing the credit risk of SME loans in an efficient manner.

1.2 Background of the Study

Credit risk is the most important risk that every financial institute should endure. Analyzing the credit risk of loans is a crucial operation in financial institutes since it ensures their performance and survival. Banks must confirm the lender's ability to pay a loan before granting a loan [1]. The volume of loan facilities has increased daily, making more challenges to banks and financial institutes. The excessive amount of credit and poor credit risk management may lead a financial crisis. Hence, effective credit management is imperative for the stability of the financial system of a country [2]. Finding an effective method for credit risk evaluation is critical for financial

institutions when they are deciding credit facilities and reviewing the facilities granted. Credit risk analysis helps banks to improve their cash flow and make managerial decisions effectively by reducing the risk. Further, the credit rating agencies may need this kind of credit risk analysis mechanism to rate the creditworthiness of lenders of debt agreements [3].

The need for planning, financing and credit management is mandatory for banks with growing loans [4]. Among different types of loans, small and medium entrepreneur (SME) loans are imperative. SMEs play a critical role in economic growth of any country [5]. In Sri Lanka, SMEs are also considered the backbone of the economy. The Sri Lankan policy framework defines SMEs as enterprises which have less than 300 employees and which have less than 750 million annual turnover. It consists with micro, small and medium enterprises. SME contributes 52% of the Gross Domestic Production (GDP) and also it offers 45% of employment [6]. Therefore, stimulating SMEs are necessary for the economic growth in the country. Licensed and specialized banks provide credit support to eligible SMEs to inspire them [6]. The incapability of bankers to foresee the credit risk of potential SME borrowers may adversely affect on the financial system and the economic activities. [7] have found that credit risk is an important determinant of assessing the profitability of the banking sector in Sri Lanka. Moreover, banks perceive the SME sector is risky and costly than the large enterprises [8]. Hence, analyzing the credit risk of SME loans is imperative in the Sri Lankan context. This study aims at analyzing the credit risk of SME loans, by using data mining techniques.

Credit scoring methods are generally applied in banking and financial institutions but can be used in different industries such as telecommunication, insurance, real estate to predict late payment risks [9]. Credit risk scoring has been identified as a predictive tool that can be used to verify the risk associated with a loan borrower [10]. In the first phase, credit scoring was assessed based on the personal experience and afterwards it was based on Character, Collateral and Capacity which is known as 3Cs. Recently, the statistical methods with data mining techniques are adopted in credit scoring process [11]. Although statistical models and techniques are used to analyze the credit risk, the ability to categorize good and bad customers should be further improved with recent technologies [2]. Data mining and machine learning are the most popular and latest techniques in assessing the credit risk. Data mining is a method of extracting knowledge

from huge set of data in order to make successful decisions. It is an essential part of knowledge discovery in databases known as KDD [9]. Data mining is being used in the financial sector, widely for the predictions of credit risks, credit defaults, fraud detections and other important analysis [12]. Thus, the study adopts data mining techniques to analyze data and predict the credit risk. It identifies both the financial and non-financial attributes affecting SME loans' credit risk and analyzes them with data mining techniques to extract knowledge on a large data set obtained from a financial institute in Sri Lanka. Ultimately, the study provides a valuable insight for financial institutes in predicting the credit risk of potential SME borrowers with enhanced technologies. More importantly, an in-depth analysis was done to analyze the characteristics of the borrowers who have not fully paid the loan. It leads the bankers to manage the credits efficiently, minimize the credit risk of SME loans, and reduce the number of non-performing loans which are increasing daily. Further it helps in effective decision making of any financial institute.

1.3 Problem Statement

Credit risk analysis is crucial in any financial institute for their survival [13]. Credit scoring models are widely used in banks to predict the credit risk [9] [14]. Many studies have been conducted on building reliable credit scoring models to analyze the credit risk of new loan applicants. These attempts seem to be not significant due to major changes of the weights around the optimum model have a small impact on its implementation. Some misclassification patterns also could arise, due to the impact of economic conditions and highly nonlinear characteristics of loan data, even the scoring models are accurate [11]. Though financial institutes use different methods in assessing the credit risk, the number of non-performing loans have been increased throughout the past few years in the Sri Lankan context. Specifically, the number of NPLs has been drastically increased in SME loans from the fourth quarter of 2020 to the second quarter of 2021 [15].

The Department of Supervision of Non-Bank Financial Institutions under the Central Bank of Sri Lanka has also recognized that existing directions on credit risk management are outdated and one of the main reasons for non-performing loans in the leasing and finance companies. They have mentioned that the credit risk analysis

methods should be updated regularly in order to cope up with local and international developments, standards and the best practices arising at the Basel Committee on banking supervision [16]. Further, previous studies have recommended that implementing effective tools and techniques in credit risk management will reduce the credit risk [17]. Besides, there are limited research in the Sri Lankan context to predict the credit risk by using data mining techniques. Hence, this study tries to incorporate appropriate factors in assessing the credit risk and analyze the credit risk using data mining techniques. Further, the non-performing loans will be analyzed based on different aspects which will be useful for the financial sector to identify factors affecting the non-payments and to predict the credit risk accordingly.

1.4 Aim and Objectives

1.4.1 Aim

- Predict the credit risk of SME loans by using data mining techniques

1.4.2 Objectives

- Identify the attributes affecting the analysis of the credit risk for SME loans.
- Acquire the knowledge regarding data mining and other necessary technologies in analyzing the credit risk of SME loans.
- Construct models for analyzing SME loans' credit risk in different perspectives through data mining techniques.
- Evaluate models to verify the accuracy of credit risk analysis in different stages.
- Implement the model to analyze the credit risk of potential SME lenders.

1.5 Proposed Solution

Proposed solution for the above study, is to analyze the credit risk of SME loans by using different data mining algorithms and to select the most appropriate data mining technique to get the accurate decision making in credit analysis. Nearly 3000 SME loan data obtained from a leading financial institute in Sri Lanka will be analyzed to build the model and the RapidMiner Studio tool will be used for data analysis. It predicts the credit risk in a few different aspects. First, it predicts whether a particular SME loan

will be fully paid or not. Then, if it is not fully paid, those unpaid loans will be analyzed further. The payment percentage of the non-performing loans will be calculated and the exit point will be predicted in two stages. It identifies whether the borrower will exit the loan by paying installments regularly or irregularly. If they pay in a regular manner, then it will predict the exact point of exiting the loan. If they pay in an irregular manner, it will predict how many stop points may occur throughout their payment duration. In that way this study analyses the probability of non-performing loans occurring and different aspects of those non-performing loans which will be really helpful in decision making for lending companies.

1.6 Structure of the Thesis

The first chapter describes the background of the study, the research problem addressed, aim and objectives and the proposed solution. The second chapter examines the related works of the literature comprehensively. Then the adopted technologies are described in the third chapter while the research methodology of the study is described in the fourth chapter. The fifth chapter emphasized the analysis and the design of the study. The sixth chapter will be the implementation which explains the implementation of the recommended model. In the next chapter the model evaluation will be described as the seventh chapter of the thesis. Finally, the eighth chapter will conclude the study by identifying the limitations of the study and proposing future enhancements. List of references are presented in the next part followed by the appendices.

Chapter 2 : Literature Review

2.1 Introduction

The second chapter critically reviews the existing literature for analyzing the credit risk of SME loans by using data mining techniques. First, it describes the concepts related to credit, credit risk and SME loans. It emphasized the importance of simulating SMEs for any economy. Then it discusses the data mining techniques that have been used to analyze the credit risk in the literature. It also identifies the possible data mining methods that can be applied for analyzing credit risk. This chapter provides thorough insights into the concepts and theoretical models related to credit risk analysis of SME loans and data mining techniques.

2.2 Credit Risk Analysis of SME Loans

Credit is considered as the main source of revenue in any bank around the world [17]. Credit risk analysis is important in the banking sector to confirm the borrowers pay loans on time and adhere to the regulations [1]. Credit scoring methods or judgmental techniques implement the decision of accepting or rejecting a request for a loan. A credit scoring system allocates potential lenders to “good credit” category, if lenders are likely to repay the loan or to “bad credit” category if there is a high possibility of avoiding the commitments [9]. Indeed, the credit accuracy is extremely important in credit scoring. If the accuracy of bad credit candidates grows only 1%, it may lead a huge cost for the bank [11]. [18] have mentioned that if a person or a company fail to make loan payments, it will be highly cost for both the bank and the stockholders. Banks will lose the money they have already given to the borrowers, and stockholders will lose their investments or the equities. Therefore, credit is considered as a fundamental component of the current market economy. An ideal personal credit system is a strong basis for the proper operations of the national market economy in many developed countries [19]. Effective credit management is necessary for the long-term survival of any financial institute [2] [20]. Banks must ensure that customers can pay the installments of a loan, before giving the loan. Credit practitioners in banks get the most crucial decisions based on the data of credit management department rather than stock data or customer information [19]. The recent financial crisis has forced the banks to

consider the credit risk analysis as a crucial task because of severe losses [21]. Due to the harmful effects of credit risk on global economy, the Basel committee allocated a specific component to prudent regulation. The committee proposed measures to redefine the vital facets of the evaluation of customer risk. It confirms the necessity of complex models in credit risk management [21] [22].

Loan categorization is a process of evaluating loans, allocating loans to groups and assigning grades based on the perceived risk and other related properties [14]. Among various types of loans such as commercial loans, SME loans, personal loans, educational and industrial loans, SME loans are imperative since SME financing critically impact on economic growth and it creates employment opportunities. Evidence in SME financing shows that enterprises who are unable to produce adequate operating cash flow are more vulnerable to bankruptcy [8]. [3] indicated that small businesses are more economically risky and have low asset correlation than other companies. SMEs are account to few legal obligations for accounting information, when compared to large companies. Moreover, the author has mentioned that SMEs have a greater impact from external events such as financial fluctuations, strategy changes, changes of managers and so on. SMEs in developing countries are facing many challenges such as higher competition, administrative inefficiencies, economic conditions, the long-lasting threat from globalization, and also financial crisis have initiated distress [23].

Small businesses, on the other hand, are the primary sources of the stability and development of financial systems in most nations, fueling the engine of economic advancement and expansion [24]. SMEs are playing a key role in encouraging economic development and industrial production, all over the world. Specially, SMEs provide the foundation for sustainable economic growth and increasing income in less developed and developing countries [25]. [5] highlighted that SMEs are not only the primary body of the market of China, but also they considered as the pillars of their social economy. SMEs are defined based on the annual turnover and the number of employees in Sri Lanka. The Sri Lankan policy framework defines SME as an enterprise which have an annual turnover below 750 million and which have less than 300 employees. The contribution of SMEs to Sri Lankan economy is more than 75% of total enterprises and contribution to the GDP is 52%. It provides 45% of employment also [6]. Hence, analyzing the credit risk of SME loans is crucial in any economy, to

encourage them and grow the economy [25]. According to [15] bankers' willingness to lend has been increased in 2021, compared to fourth quarter in 2 020, and on the other hand, all the loan categories (corporate, retail, State Owned Enterprises (SOE), SME) has been reported a higher readiness to lend. Further, it emphasizes that the Covid-19 pandemic in the first quarter of 2020 stressed more increase in willingness to lend. With this increasing amount of lending, the number of Non-Performing Loans (NPL) were increased at a higher phase in 2021 [15]. Further, it is expected more growth of NPL, with the expiration of moratorium facilities during uncertainty in income levels. [16] has stressed the need of changing credit risk directions regularly to keep pace with local and international bodies, standards, practices proposed by the Basel Committee in banking supervision. This highlights the necessity for proper management of the credit risk of SME loans in Sri Lankan economy evidently.

It is important to identify attributes that impact on the credit risk [21]. [26] have introduced 6C's analysis for credit evaluation. Those are character, capacity, collateral capital, condition of economy and constraint. Character refers to the lender's credit rating, which evaluates the ability to pay on the agreement. Capital is the amount of funds available at the lender. Capacity is estimated by considering the current income and expected income during the loan period, and it is the ability of lenders to pay the installments. Collateral is a property of a lender which should be considered as a warranty of the loan. Condition of economy represents the social, political, cultural and economic conditions that affect on the sustainability of businesses. Constraint refers to all the barriers and limitations affect on a business, such as scarcity of resources and regulations.

[27] have emphasized different kinds of credit scoring methods such as application credit scoring, collection scoring, behavioral scoring and the fraud detection. The manual credit scoring methods consist of many troubles in the industry. Moreover, the process of credit scoring is not standardized in many banks. Such non-standard models associate with serious problems of analyzing expensive data repeatedly. It is an obstacle for building an optimal model for credit risk analysis [21]. [4] have investigated factors in measuring the credit of bank customers such as current capital turnover, loan interest rate, lender's annual income, customer capital, loan amount and the history of corporate with the bank. According to the results, the average turnover of the account has the highest weightage among the factors. Some researchers have introduced financial early

warning systems (EWS) in reducing risk and predicting the level of success of financial institutes. Financial EWS is a supervising and reporting tool that detects potential problems, hazards, and opportunities before they affect a company's financial accounts. Financial performance, financial risk, and probable bankruptcy are all detected using EWSs. Almost all financial early warning systems are based on the financial accounts. Income statements and balance sheets are the main sources for early warning systems that reflect the financial realities. [23].

The lending databases of the bank consist of diverse variables for customer credit history. It is tricky to categorize clients and recognize the effects of these diverse variables if the number of variables is higher [21]. Financial ratios are also being used to build credit risk predictions. However, it can be problematic because several studies have suggested various ratios as the major indicator of financial issues. Therefore, non-financial factors such as type of enterprise, age, industry, gender etc. should be taken into account with the combination of financial ratios [21] [28]. [28] have also mentioned that similar to the financial indicators, the non-financial indicators are also important for SMEs to achieve the maximum competitiveness. They have suggested that SMEs should improve the internal mechanisms and the environment of their business in order to achieve competitive advantages.

[21] have further identified that size of the company, profitability ratios, creditworthiness, time period of a credit report, repayment capacity, guarantees, ownership structure, loan number and corporate banking relationship duration are the key factors in predicting nonpayment. Furthermore, most credit risk models are static, therefore incapable of functioning effectively in economic crises. Though many financial institutes use credit scoring methods those are not standardized. There are critical issues with these non-standardized models since they analyze repetitive and very expensive data. Indeed, this issue inhibits the development of optimal models in credit scoring. Banks used traditional static modeling methods to assess credit risks, but these models were unproductive due to the shortage of responsiveness to the growing economic environment. Though these traditional static models have performed quite well in stable situations, it fails to work so with political and economic fluctuations [22].

2.3 Data Mining Techniques for Credit Risk Analysis

Thus, usage of data mining methods to assess credit risk is dominant in the field. Data mining is an emerging technology that provides significant advantages in making correct decisions [29]. It contains set of algorithms and methods to extract knowledge and information from large databases. According to [30] data mining is "the process of discovering meaningful new correlations, patterns and trends by examining through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques". It can be supervised or unsupervised based on the data that are going to be analyzed. The rapid development of internet and telecommunication technologies marks the arrival of big data, as structures, semi-structured and unstructured data generated from different sources. Many countries have identified the importance of big data as a global opportunity. In this regard, financial big data research is also critical. It is not only necessary for social growth but also for determining the operational state of the financial sector in order to improve its structure and maintain regional financial stability [31]. The current behavior of data can be described and the future behavior of data can be predicted with data mining [32]. The most widely applied data mining methods are classification, clustering, regression, association rules, visualization, decision trees, support vector machine and neural networks [33]. Implementing the correct data mining technique is extremely important to extract valuable information [4].

Data mining is used in various industries, which are differ in characteristics. Among various purposes, one of the most popular usage of data mining is in the banking sector to perform the financial analysis [12]. It can be used in performing analysis such as credit analysis, optimizing stock portfolios, customer segmentation, predicting payment default, marketing, investments ranking, credit card risk assessments, predicting fraudulent transactions, cross selling and etc. [14]. There are different types of techniques available under the concept of data mining such as classification, association, clustering, anomaly detection and so on. Classification is a predictive method that develops a model which describes and distinguishes classes or concepts for future predictions. It includes machine learning algorithms such as K-Nearest Neighbor, decision tree, artificial neural networks etc. Clustering refers to the separation of objects in a huge dataset into several segments or clusters with similar

characteristics. The detection of groups of objects that appear frequently together, for example, items that are commonly purchased together or genes that are related, is made possible through association analysis. Anomaly detection entails the use of algorithms to detect real anomalies such as unusual diseases, fraud detections, strange weather conditions and environmental issues [34].

[4] have emphasized that it is important to identify the effective factors affecting credit risk, so that the lenders can be evaluated and there would be no issues in returning money to the banks. The clustering technique with a combined data mining approach is beneficial in recognizing the hidden knowledge in the data and use it for productive decision making. [20] has introduced cluster analysis which incorporates both the return and risk of credit bonds to enhance credit portfolios. This combined approach of risk and return reveals clusters with high or low profitability respectively. Among different types of data mining techniques, [12], [21] have identified decision trees as one of the most used methods. They have also mentioned that the success of data mining techniques depends on the data availability, preprocessing (cleaning) and transformation.

Further, different data mining techniques have been stated in the literature for credit risk assessments including K Nearest Neighbors (KNN) [35], [36], Support Vector Machines (SVM) [2] [37], Neural networks [2] , and Logistic Regression analysis [9]. [5] has analyzed the credit risk of SME loans in China by using the Random Forest algorithm and they have ensured the accuracy of the results. Further they have mentioned that analyzing the credit risk of SMEs are very important in this pandemic situation since governments have paid their attention to simulate SMEs. [38] has emphasized that artificial neural networks are more suitable with credit scoring methods as a data mining technique. The author has suggested emotional neural network as a more successful method for pattern recognition when compared to neural networks. Regression models also have been used successfully in predicting credit risk, such as linear regression and logistic regression [20]. Some authors have introduced hybrid or combined data mining techniques by integrating two or more techniques together. [11] have introduced such a hybrid method by combining clustering and classification algorithms. They have used K-Means cluster as a clustering algorithm and support vector machines as a classification algorithm in predicting credit risk of local banks in China. [24] have introduced a data mining approach to predict imbalanced set of small

businesses' loan data by combing weighted SMOTE ensemble algorithms and bagging algorithms. [29] have also proposed a hybrid approach by combining genetic algorithms with support vector machine classifiers.

Several authors have utilized many data mining techniques for credit risk analysis as it is an emerging technology that gives more insights to big data analytics. Since adopting data mining techniques is important in analyzing the credit risk, this study aims to evaluate the credit risk of SME loans using data mining techniques, which will be helpful in taking more reliable decisions based on accurate predictions. Table 2.1 demonstrates the different classification algorithms used in literature, specifically in the credit scoring and risk analysis.

Table 2.1: Summary of Machine Learning Techniques used in Predicting Credit Risk

Machine Learning Technique	**References from Literature**
K Nearest Neighbors (KNN)	[35], [36], [39], [40]
Support Vector Machines (SVM)	[2], [37], [11], [10], [41]
Random Forest	[5], [42], [43], [44]
Decision Tree	[12], [21], [1], [9], [45]
Rule Induction	[46], [47], [48]
Artificial Neural Networks	[19], [2], [21], [38]
Logistic Regression analysis	[9], [2], [10], [49]
Naïve Bayes	[14], [50], [51]
Clustering	[20], [52], [11]

2.4 Summary

This chapter provided a deep understanding about the existing literature on credit risk analysis of SME loans and data mining approaches that have been used. A summary of data mining techniques used by previous researchers for credit risk analysis were presented. It can be concluded that most of the researchers have used classification algorithms for financial predictions. Hence, this study will also adopt classification algorithms in predicting the credit risk of SME loans. The next chapter presents the technologies adopted in the research design of this paper.

Chapter 3 : Adopted Technologies

3.1 Introduction

The previous chapter described the different findings related to credit risk analysis of SME loans as well as usage of data mining techniques for credit risk analysis. This chapter explains the selected technologies which will be used in this study. It will provide a proper understanding about the way of utilizing the above-mentioned data mining technologies in the process of credit risk analysis.

3.2 Data Mining

Data mining is an emerging technology. It consists of algorithms and methods for extract knowledge from large data sets to explain the current behavior and to predict the future behavior of data [21]. DM techniques is being widely used in different domains such as market segmentation, fraud detection, diagnosis, customer service, benchmarking, credit scoring and other applications [53] [54]. Different DM functions are available to cater various requirements including summarization, classification, clustering, regression and so on [53]. In literature, various DM models have been used in assessing the credit risk such as decision trees [1], [9], K nearest neighbors [55], support vector machines [2], artificial neural networks [2], [21] and logistic regression [9]. It has been argued that DM techniques can be successfully used in examining credit risk [21], [53]. Hence, this study aims assessing the credit risk of SME loans by using appropriate data mining techniques.

DM methods can be separated into two categories known as predictive and descriptive methods based on the objective of data mining. Predictive methods use existing variables to forecast the future values of the other variables and it includes methods like classification, anomalies detection, regression etc. Descriptive methods uncover patterns available in data construed by the user and it consists of methods like clustering, sequential patterns, association rules etc. [53]. This study adopts classification models as it is a predictive analysis and consists of five stages of predictions. First it predicts whether a loan will be fully paid or not. Then if it is not fully paid, in the second stage it analyzes those non-paid loans in few different perspectives. The percentage of the payments (up to which extent the borrower will pay

the loan), whether they will stop payments regularly or irregularly and the exit criteria will be further analyzed. If they have exited the loan after set of regular payments, the exact exit point will be predicted and if they have exit in an irregular manner, the number of exit points will be predicted. Different algorithms were used and evaluated in each stage to select the optimal, and they were mentioned under the process of data mining scheme.

3.3 Data Mining Process

The data mining process consists of few stages such as selection of data, preprocessing, transformation, data mining, evaluation and interpreting the results. The process was depicted in figure 02. The way this study has followed these steps was described in the next chapter. Here, in the first phase data should be collected related to the domain to be studied. The exploratory data analysis should be used to get familiarized with data. Then the desired subset of data should be selected. The raw data should be prepared by selecting variables that we want for the analysis and appropriate for the study. Some data may be transformed according to the requirement. As the next step the relevant modeling techniques should be applied and get the optimum results. Different data mining algorithms can be applied on the same data set. In the final stage the applied models should be evaluated to verify the quality and the effectiveness before they deployed. Finally, the results should be interpreted in order to extract the knowledge [56].

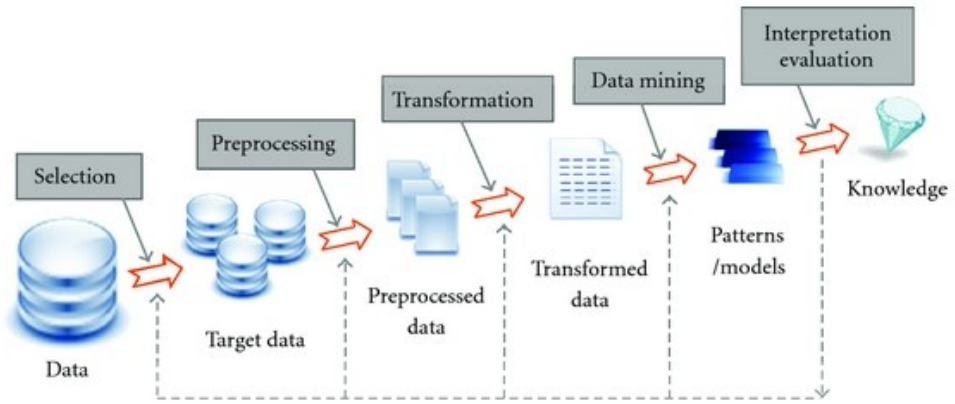


Figure 3.1: Stages in Data Mining Process

3.4 Supervised Learning

Supervised learning is a paradigm of machine learning which acquires knowledge from a set of input-output training samples. Since the output is considered as the label, the training sample is also known as supervised data or labelled training data. The objective of supervised learning is to develop an artificial system that perform the mapping between the input and the output and predict the output by giving new inputs [57]. Simply, if the class label is known or provided in a given training data set, it is recognized as supervised learning or the learning of the classifier is supervised [58]. It finds a model for class attribute or the label, as a function of the other attribute values. The aim of supervised learning is to assign previously unknown examples into a class as much as accurately. Classification belongs to supervised learning since the class label is known.

3.5 Classification Algorithms

Classification refers to the process of developing a model which explains and distinguishes classes or concepts for future prediction. It belongs to supervised learning since the class label is known in the set of training data. For example, countries can be classified based on the climate and motor vehicles can be classified based on gas milage. The resulting model can take many different forms, including classification rules, decision trees, neural networks and mathematical formula [58]. Classification has two main steps including model construction and the model usage. In model construction, a set of determined classes should be described. In the usage of the model, the unknown or future objects can be predicted. Each classification algorithm can be described as follows.

3.5.1 Decision Tree

A decision tree is a tree structure that looks like a flowchart, that a node symbolizes a test on an attribute value, a branch denotes the outcome of the test, and leaves represent class distributions or class labels [58]. Decision tree recognizes a set of variables and association rules and then analyzes the impact of these variables to make decisions. It is capable of processing massive data [1]. Training examples are lied at the very beginning of the tree and in the root examples are partitioned iteratively based on the

attributes. The test attributes can be picked on statistical or heuristic measures such as Gini index, information gain. Decisions trees can be converted into classification rules easily [53]. Researchers have identified advantages of decision tree method such as efficiency, flexibility, easy interpretations, logical relationships, elimination of unwanted calculations etc. [18].

In RapidMiner, decision tree operator creates decision trees with decisions and leaf nodes. This can be used on both numerical and nominal data sets. There are two stages of decision tree development such as tree construction and tree pruning. The generated decision tree can be used to classify unknown examples and test the attribute values of the sample dataset against the decision tree.

3.5.2 K- Nearest Neighbor

This is a non-parametric algorithm which is very useful for decision making hence it relies on the entire training data set and does not comply with the theoretical assumptions made usually. It compares an unknown example to the k training instances that are the unknown example's closest neighbors [59]. And also, this can be used to anticipate a real-valued forecast for a certain unknown row. In that case, the classifier reverts the average of the real-valued labels connected with the k closest neighbors of the unknown row. The main benefit of this method is that it’s not mandatory to determine a predictive model prior to classification. The major disadvantage is that KNN does not yield a simple classification probability formula, and that the measure of distance and the cardinality k of the neighborhood significantly impacts predicted accuracy [18].

In Rapid Miner, the K value can be specified in order to set the number of nearest neighbors that we want to consider for the classification. Then, the measure type can be identified by using the distance functions such as Euclidean Distance, Manhattan Distance, Camberra Distance and so on. The distance functions identify the distance between two points. KNN works on the concept of instance-based learning.

3.5.3 Naïve Bayes

This is based on the Bayes’ theorem and valid for multiple probabilities when the events are independent. It suits for even a small set of data since it is high-bias and low-

variance classifier. It is easy to extend the Naïve Bayes algorithm to cope up with numerical attributes. Nave Bayes can be wrong if a specific attribute value does not appear in the training set with every class value [53]. Further, it has shown high accuracy and speed with large databases as well. It assumes that attributes are independent in real life as well. Since this independence assumption does not exist in most of the real scenarios, Bayesian Belief Network was introduced, which is more flexible [55]. The main benefit of this algorithm is that it needs only a little number of training data to predict the necessary parameters. Simply, it can be used for sentiment analysis, text categorizations, recommender systems and spam detections. RapidMiner models attribute data using Gaussian probability densities [60].

3.5.4 Rule Induction

Information gain and the accuracy are the parameters for specifying the criterion for choosing attributes and numerical splits. With respect to the information gain from the given example set, it learns a pruned set of rules. This results in massive number of association rules which should be pruned down based on the coverage and accuracy. This is identified as quite infeasible [53]. Rule induction algorithm in RapidMiner starts with the less prevailing classes, grows repetitively and prunes rules tilThe rule induction algorithm in RapidMiner starts with the less prevailing classes, grows repetitively and prunes rules until the error rate is higher than 50%t or no positive examples left.

3.5.5 Random Forest

This produces excellent predictors by building randomized decision trees in iterations. This algorithm is an ensemble of a particular set of random trees, which we can specify as a parameter. The trees are trained based on bootstrapped sub-sets of the input data set. As its name implies this algorithm contains enormous number of decision trees that works as an ensemble. Each individual tree generates a class prediction while the class with the most votes becomes the prediction of the classification model [53].

The Random Forest algorithm in rapid miner produces a number of random trees. The resulting forest model includes a specified number of random tree models. This number of trees required, can be given as a parameter of the random forest operator. Pruning can be used to minimize model complexity by replacing sub-trees that have limited predictive potential with their leaves.

3.5.6 Linear regression

This can be used to predict numerical attributes. It finds the best fit linear relationship between the independent and the dependent variables. The idea behind this algorithm is to convey the class as a linear sequence of the attributes, with predetermined weights [53]. LR is considered as a bivariate method since it operates on two variables. In the same manner that classification is applied to predict categorical examples, linear regression is used to predict continuous variables. This can be used in predicting how much certain factors like price of a product, interest rate, impact of price variations of assets and so on.

3.5.7 Logistic Regression

Logistic regression can be used in both regression and classification tasks. It predicts a categorical dependent variable based on a set of independent variables. The main distinction between linear regression and logistic regression is the type of the variable which are going to be predicted. Linear regression is used for continuous variables as specified above [61]. In RapidMiner, this algorithm employs Java implementation of the myKLR introduced by Stefan Rueping. It is a fast algorithm which gives good results in many tasks.

3.5.8 Artificial Neural Networks (ANN)

Neural networks have been used since 1990's for financial analysis as a quantitative forecasting technique. It is a nonlinear optimization tool which comprises of three layers including the input, output layers and one or many hidden layers. Each neuron processes inserts and retrieves a value that is conveyed to the neurons in the layer below it [21]. [38] has identified neural networks as a feasible model in analyzing the credit risk as it allows the characteristics to be interacted in several ways. One characteristic may relate to many characteristics, which making a comprehensive network structure. [18] have emphasized several benefits of ANN such as robustness, ability for generalize and handle large number of data, less statistical assumptions, memory and so on. The drawbacks of the method include longer training times, may be overfit with large datasets, issues with interpretations and may infeasible with small datasets.

3.5.9 Support Vector Machines (SVM)

SVM is more suitable for classification and regression with small number of samples. It is considered as a fact algorithm that generates proper results in many learning tasks. SVM involves three main steps: select the optimal feature, choose the kernel and optimize kernel parameters [11]. It depends on the maximization of the margin for producing a hyperplane that divides the data into two main classes by providing them with great generalization capacity [2]. In RapidMiner, SVM uses a Java implementation of mySVM by Stefan Ruepingthe. SVM operator supports different kernel types such as polynomial, dot, neural, radial, multiquadric, anova, epachnenikov and gaussian combination.

3.6 Data Mining Tools

Different tools are available in machine learning to analyze data, such as Weka, RapidMiner Studio, Oracle Data Mining etc. Weka is a tool which consists of machine learning algorithms and it facilitates data preparation, visualization, classification, clustering, regression and association rule mining. It is an open-source software offered under GNU General Public License. Weka provides scalable and reliable solutions for the AI workloads in modern enterprises.

This study uses RapidMiner Studio which has a rich library of 1500+ algorithms and functions to ensure the best model. It is considered as a visual workflow designer since it has a unique visual representation. It helps speedy and automated creation of visual models in data mining. The scalable architecture of RapidMiner Studio is depicted in Figure 3.2 [60]. It provides an incorporated platform for deep learning, data preparation, machine learning, text mining and predictive analytics. It is more scalable and facilitates better visualizations. Various extensions can be added to RapidMiner according to the requirements and to build more comprehensive models.

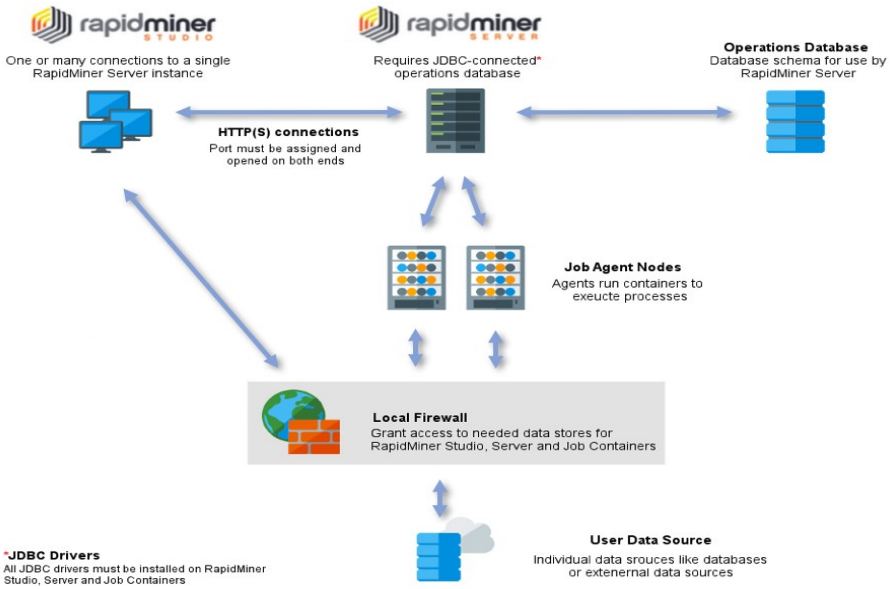


Figure 3.2: Scalable Architecture of RapidMiner Studio

3.7 Summary

This chapter presented data mining techniques as the adopted technology in analyzing the credit risk of SME loan data. It described what data mining is, the process of data mining, the concept of supervised learning, classification algorithms and data mining tools available. RapidMiner Studio was used in this study as a tool in assisting the data mining process. The next chapter will describe the methodology or the proposed approach in analyzing credit risk of SME loans.

Chapter 4 : Methodology for Analyzing the Credit Risk of SME Loans

4.1 Introduction

This chapter describes the research methodology that has been adopted to address the research problem. In this process, data mining techniques were applied as mentioned in the previous chapter. Here, it describes the approach by highlighting the hypothesis and the steps followed in the data mining process to analyze the credit risk of SME loan data.

4.2 Hypothesis

The main hypothesis of this study is classification techniques in data mining can be used for the credit risk analysis of SME loan data. The selection of the hypothesis was affected by the fact that the increasing number of non-performing loans in SME category throughout the last few years and there is no proper mechanism to predict the credit risk of SME loans in the Sri Lankan context. Hence, this study uses classification algorithms such as KNN, Decision Trees, SVM, Random Forest and Naïve Bayes in order to identify the most accurate one for the prediction of credit risk.

4.3 Input

As the input of this study, Small and Medium-sized enterprise loan data have been collected from one of the leading financial institutes in Sri Lanka. Nearly 3000 records were obtained to perform a comprehensive analysis based on data mining techniques. The dataset was identified as a balanced data set since it contains 1535 non-performing loans and 1515 fully paid loans. The attributes available in the data set will be discussed later in this chapter.

4.4 Output

The output of this study will be a designed simulation which can be used to identify non- performing loans. Several classification models will be created by using selected classification algorithms. The model with the highest accuracy will be selected to predict the non-performing loans in this selected data set. In this study, the predictions

will be performed in several stages. First, it will predict whether a particular SME loan will be fully paid or not. It gets arrears, then it will predict the percentage of the arrears amount and further it will identify whether the borrower has exited the loan after a set of regular payments or irregular payments. If it is irregular, then it will predict the number of exit points that can be occurred throughout the loan period. If they exit in a regular way, then it predicts the exact exit point that they will stop the loan. In that way, this study predicts the non-performing loans in different aspects.

4.5 Process

In this study, the main process is to identify the most appropriate and accurate classification algorithm which can be used in analyzing the credit risk of SME loans in order to get the decisions effectively. A thorough literature review was done to identify existing algorithms that can be used in the process. Within this main process, several sub processes are included. Those are data selection, data preprocessing, data transformation and the evaluation. The process will be described further in next few topics.

4.6 Users

The credit officers in all the financial institutes can use the proposed model in predicting the credit risk of SME loans. They can easily identify the potential debtors and can predict their behavior more accurately. Further, this study will provide valuable insights for managers in decision making. On the other hand, this study also indirectly benefits SMEs, since the financial institutions may simulate SMEs by predicting their credit risk.

4.7 Features

The models proposed by this study in predicting credit risk in different stages, can be used to analyze a huge amount of SME loan data in any financial company including state and private banks. It can be applied to analyze any type of data by getting variables related to those types of loans. As the study's main features, this can predict whether a loan will get fully paid or not at the first phase. Then it is capable of analyzing identified non-performing loans in few different perspectives such as exit method, exit point and number of exit points.

4.8 Data Selection

The sample data related to SME loans were collected from one of the leading financial institutes in Sri Lanka. It contains nearly 3000 records of SME loan data pertaining to year 2017 – 2020. Since the Covid 19 pandemic has been adversely affected in financial sector, three years were selected (before the pandemic) for a better prediction of credit risk. The data is in the format of MS Excel thereby the manual preprocessing activities were easy to perform. Among 3050 samples, 1978 samples of loans belong to females. And 2559 of the borrowers are married. The net monthly income has been calculated by deducting their monthly expenditures from the monthly income and it is ranged from approximately 6650-35000 LKR. And the loan amounts were varied from 50000-250000 LKR. The interest rate is varied from 31% to 35% per annum. The loan term is ranged from 12 months to maximum 24 months. Monthly installment is varied between 3000 to 15000 LKR. Further 15 SME types are available in the data set such as plant nursery, weaving, food manufacturing, cane industry, rubber products, batik industry, food manufacturing, garage, craft designing, rubber products, aquarium, wood carving, apparel, mills, diary processing and salon. The descriptions of the variables are presented in Table 4.1.

Table 4.1: The Definitions of Variables

No	Variable	Data Type	Description
1	Gender	Binomial	The gender of the borrower (Male or Female)
2	Age	Numeric	The age of the borrower in years
3	Marital Status	Binomial	Marital Status of the borrower (Consists of two statuses: Married or Single)
4	Net Monthly Income	Numeric	Monthly income - Monthly expenditures
5	Business Type	Polynomial	SME business type (15 types are available as mentioned above)
6	Loan Amount	Numeric	Amount of the loan received
7	Interest	Numeric	Interest rate of the loan

8	Loan Term	Numeric	Time period of the loan
9	Monthly Installment	Numeric	Monthly installment that the borrower should pay
10	Status of Payments	Binomial (Class Variable)	Whether the loan will get arrears or not Yes – Arrears (Non-performing loans) No – Non-arrears (Fully paid loans)
11	Payment percentage	Polynomial	If the loan is not fully paid, the extent to which the payments have been occurred. Very Low – Pay between 0% - 25% Low – Pay between 25% - 50% Medium – Pay between 50% - 75% High – Pay between 75% - 100%
12	Exit Method	Binomial	Whether they exit after a set of regular payments or exit with irregularly payments
13	Exact Exit point	Numeric	If borrowers exit after set of regular payments, in which installment onwards the borrower has stopped the payments.
14	No. of exit points	Numeric	If borrowers exit after set of irregular payments, the number of exit points they have come up with, throughout the loan period.

4.9 Data Preparation

The data set was filtered first, in order to get the loan details only related to 2017-2020 since those three years were selected for the study. It was identified as a balance data set since it has total 3050 records which consists of 1515 of fully paid loans and 1535 of non-performing loans. Some manual preparation of data was done in excel sheets to identify a few attributes such as percentage of non-payments, exit method, the exact

exit point and the number of exit points by analyzing the way the SME borrowers have paid the loan throughout the time period.

4.10 Data Pre-Processing

Data in the real world are incomplete, inconsistent and noisy. If the quality of data is low, then the quality of the results may also low. In data mining, it requires quality data to generate a quality output. Therefore, data preprocessing plays a critical role in the process of data mining. There are major tasks in pre-processing such as data cleaning, integration, transformation, reduction and discretization. Data cleaning refers to filling the missing values, removing inconsistencies and outliers and smoothing noisy data. Data integration involves combining data from several databases, files or data cubes. Data transformation refers to the aggregation and normalization of data. In normalization the data can be scale in manner so that they can be fit into a particular range. Then, data reduction takes reduced representation of data in volume but generates the similar analytical results. Another part of data reduction is discretization which converts some numerical attributes into nominal attributes by categorizing the numerical attribute into a number of bins specified by the user. The way this study follows data preprocessing activities will be further described in the sixth chapter.

4.11 Data Mining Scheme

Among different data mining algorithms, the most accurate one should be selected according to the data set. In this study the credit risk will be predicted by considering the above-mentioned attributes and since a class label is available a classification model is more appropriate. This belongs to supervised learning because the training data set is supplemented by labels showing the class of the observation [53]. Based on the literature, classification algorithms were selected such as Naïve bayes, Decision tree, K-NN, SVM and Random Forest models. Each of these models were used and tested in RapidMiner Studio in order to select the most accurate one for the study.

The above-described algorithms were applied to predict whether the loan will be paid completely or not in the first phase. Then if it is incomplete, it will be analyzed in various perspectives again by using different models. To predict the percentage of the payments, again the same algorithms were applied since the level of percentage was discretized into four nominal categories such as high, medium, low and very low. Then

the exit method was analyzed by applying same set of algorithms. It checks whether a particular borrower will exit the loan in a regular manner or in an irregular manner. If it is regular, then it predicts the exact point of exit (in which installment onwards the borrower has stopped the payments). Linear Regression, KNN and decision tree algorithms were tried for this prediction since they facilitate numerical predictions. Then if the exit method is irregular, it predicts the number of exit points that can be occurred during the loan period. Again LR, KNN and decision tree were used to find the number of exit points. In that way this study provides valuable insights into analyzing the credit risk in various aspects.

4.12 Summary

This chapter presented the approach to analyze the credit risk of SME loans by using data mining techniques. Here, it pointed out how this methodology offers an efficient and accurate solution for credit risk analysis using data mining. It described the input, output, hypotheses, process, users as well as features of the proposed model. The next chapter presents the design and the analysis of presented approach in this chapter.

Chapter 5 : Analysis and Design

5.1 Introduction

This chapter presents the approach to analyze the use of different data mining techniques to predict the credit risk of SME loans. It emphasizes the design of the study including high-level design and sub modules. Then it elaborates the analysis of data by using different data mining algorithms.

5.2 Research Design

This study follows the scientific research methods to achieve its aim of efficiently predicting the credit risk of SME loans. It adopts the quantitative approach since it relies on numeric data. The quantitative approach requires a set of hard data which are then manipulated and analyzed by using statistical methods to check whether the hypotheses are proved or not. This study adopts the data collection methods such as observations and documentations to identify the issues related to credit risk management of financial institutions. Then, a background study was carried out to identify solutions introduced to mitigate the associate risk with credit. Due to the inefficiencies of those solutions and due to the higher growth of non-performing loans in SME sector, the data mining approach was selected to analyze the credit risk of SME loan data. Previous studies have also used data mining techniques in different contexts successfully, especially in the financial sector.

The research questions were identified as mentioned in chapter one and the hypotheses were built. That hypotheses were tested by using a large data set related to SME loans, by applying data mining algorithms. RapidMiner studio was used for the testing process and a few models were developed to identify non-performing loans in different perspectives.

5.3 Detailed Research Design

The main objective of this study is to analyze the credit risk of SME loan data by using data mining techniques. The high-level research design of the study is depicted in Figure 5.1. A sample set of training data will insert as the input and several

classification algorithms will be applied on that. The model will be developed and evaluated by looking at the performance of different machine learning algorithms. Then, the test data set will be inserted to the model and it will predict the credit risk of the given data set.

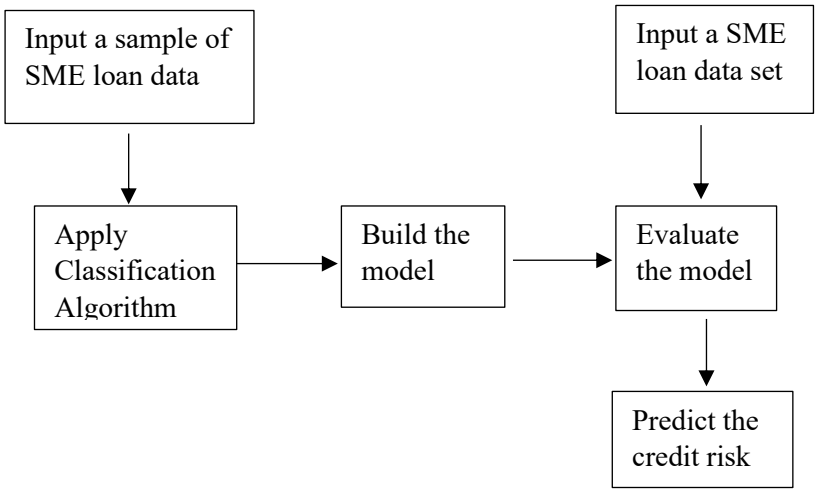


Figure 5.1: The High Level Design of the Proposed Model

The aim of this study is to analyze the credit risk of SME loan data by using data mining techniques. To achieve the aim of the study, there are a few research questions to be addressed as mentioned follows.

5.3.1 Research Question One

What are the factors affecting the credit risk of SME loans?

5.3.2 Research Question Two

How data mining techniques can be used in analyzing the credit risk of SME loan data?

5.3.3 Research Question Three

What is the most appropriate algorithm in analyzing the credit risk of SME loans?

5.4 Analysis

The collected data will be preprocessed and analyzed by using the RapidMiner Studio tool as a data mining tool. RapidMiner is a more robust and user-friendly tool for data mining. It is a data science platform facilitates clear visualizations and a vast range of algorithms, libraries and extensions. Different classification algorithms will be employed on the data set in order to identify the most accurate model. To find the best model, the evaluation will be done by using different performance parameters. The results of the analysis will be presented in the next chapter.

5.5 Summary

This chapter provided details on research design and analysis of the study. Furthermore, this chapter focused on detailed design for the research and how primary and sub research questions are structured within the research. Moreover, RapidMiner was used for analyzing data. Next chapter will explore the implementation details of the research design and analysis.

Chapter 6 : Implementation

6.1 Introduction

This chapter discusses the implementation of the overall project by giving solutions to each research question in the study. Data selection was done first and data preprocessing was done in a few stages. Then the data transformation was done and the patterns were identified by applying different data mining algorithms. Based on the identified patterns the relevant knowledge or the predictions were extracted as the output.

6.2 Solution for Research Question One

The first research question is to identify the factors affecting the credit risk of SME loan data. Here, the different sources are referred such as existing procedures in financial institutes as well as the factors considered by the previous researchers. Both the financial and non-financial factors are considered in this regard. In the literature also, previous scholars have identified the importance of considering non-financial factors for credit risk analysis. Then the most critical factors were identified and obtained from the dataset received from a leading financial institute in Sri Lanka.

6.3 Solutions for Research Question Two

The second research question is to identify the data mining techniques can be used in analyzing the credit risk of SME loan data. A thorough literature review was done in order to understand the data mining and machine learning algorithms that can be used in this scenario. Based on that, the classification method under supervised learning was selected since the class label is available in the obtained data set. The results will be presented under the next few topics of this chapter.

6.4 Solutions for Research Question Three

As the solution for the third research question, identifying the most accurate model in predicting the credit risk, different algorithms were tested and compared in few different levels of predictions as mentioned previously. Several performance criteria

were used in comparing the results such as accuracy, correlation, Kappa statistic, squared correlation and classification error. The results will be discussed in detailed within next chapter.

6.5 Data Preprocessing

In this study, a few missing values were identified under the status, gender and business type attributes. First, those missing values were replaced by using the average values. Then, the duplicate values were removed as it generates an overfit model. Discretization was done only by selecting the percentage attribute to arrange the percentage values into ranges. Here, discretization by user specification method was used with 4 classes. Then normalization was done for the interest and loan term attributes to rescale them. Z-transformation or the statistical transformation was used here in order to verify all the attributes have the similar scale for a reasonable comparison between them. The outliers were also detected with the Euclidian distance function with 10 number of neighbors. Then by using a filter example operator, the outliers were filtered out and only the non-outlier examples were selected.

Then the attribute selection was done. The borrower's birth date, the total receivable, and the total amount paid are not directly relevant to the study, hence not selected for further proceedings. Only the relevant attributes were selected including age, gender, interest rate, loan amount, loan term, monthly installment, net monthly income, marital status, payment percentage and the business type. The class label was set to whether arrears or not attribute by using the set role operator and cross validation was used to split the test and training data set. Afterwards different machine learning algorithms were applied to identify the most accurate predictions.

6.6 Data Mining Algorithms

As discussed in the previous chapters the classification algorithms were selected for the analysis since the class label is available in the data set. Among different classification algorithms, a few were selected by referring the literature. Many researchers have used decision tree, neural networks, SVM, naïve bayes, KNN, linear regression, Rule Induction and Random Forest algorithms in predicting the credit risk. Hence those algorithms were applied and tested for their accuracy in this study. To evaluate each

algorithm the performance operator was used with different parameters such as the accuracy, classification error, correlation, squared correlation and the Kappa statistic.

6.7 Parameter Tuning

Different parameters can be identified under each algorithm. These parameters can be changed to increase each algorithm's performance and reduce the error rate. In decision tree algorithm there are parameters like criterion, maximal depth, confidence, minimal gain, leaf size etc. The criterion and maximal depth was changed to increase the performance. Different criterion can be selected such as gain ratio, information gain, Gini index, accuracy and least square. Gini index shows the highest accuracy here. The depth of a tree differs depending on the size and characteristics of the example set [60]. Further, this can be applied to limit the depth of the decision tree and the default value is 10, that keep as it is.

Under the KNN algorithm the K value was decreased which was 5 by default. It finds the K number of training examples which are closest to the unknown examples. It should be a small, positive and an odd number. Then in SVM, the Kernel type and Kernel cache can be changed to enhance the accuracy. It supports different Kernel types such as Anova, Dot, Polynomial, Neural etc. Different kernel types were tried in order to improve the accuracy of the model. In linear regression algorithm the feature selection can be changed into none, greedy, M5 prime, T-test and iterative T-test. Further, the minimum tolerance ratio can be specified.

In Random Forest algorithm, the number of trees can be specified starting from 1 to infinity. The default value is 100 and it was kept as it is. The criterion can be again set to gain ration, Gini index, information gain, accuracy or least square. Based on the performance, it was set to Gini index. Under LR, feature selection parameter describes the feature selection method to be used in regression such as M5 prime, greedy, T-test and so on [60]. It can be set to none as well.

6.8 Cross Validation

Cross validation operator can be used to assess how accurately a model may perform in the practical usage. It contains two sub parts the training sub process and the testing sub process. A model can be trained by using the training sub process. Then, that trained

model will be used for predictions under the testing sub process. During the Testing phase, the model's performance is evaluated. Here it creates the number of validation subsets internally by itself. The input data set is divided into K number of samples in equal size. Among these K number of samples, one subset is considered as the test data and the rest (K-1) is considered as the training data set. This process is then iterated K number of iterations. Then, the outcomes of K iterations are averaged in order to generate the final output. In this study, the cross validation was used. Figure 6.1 shows the steps followed in RapidMiner up to cross validation.

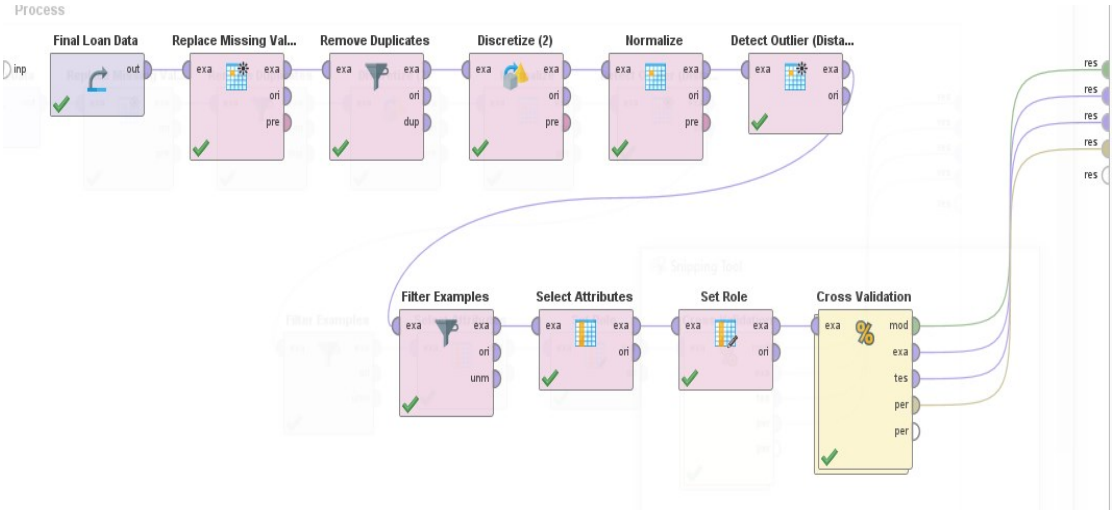


Figure 6.1: Data Mining Process in RapidMiner

6.9 Split Validation

Split validation, on the other hand predict the performance of a model by using a hypothetical test data set when an explicit test data set is not available. Here also the data set can be split into two parts as training data set and the test data set. The size of the subsets can be altered. The model is trained by using the training data set and then will be employed on the test data. This is performed in only one iteration which can be identified as the main difference between cross validation and split validation. In cross validation it repeats K number of times by using different testing and training subsets.

6.10 Summary

This chapter described the process of applying data mining techniques and implementing the models for the purpose of analyzing the credit risk of SME loan data. It explained the way of implementing the proposed solution by using the RapidMiner tool. The next chapter will explain the evaluation phase of the implemented models and the selection of the most suitable model.

Chapter 7 : Evaluation

7.1 Introduction

This chapter focuses on the evaluation of the different data mining techniques used for credit risk analysis of SME loan data. There are five main stages of the analysis as previously explained. The performances of each data mining algorithm were compared to identify the most accurate algorithm under each stage. Several parameters were used in comparing results and finally, based on the highest performance measurements, some data mining algorithms were selected in each stage.

7.2 Model Evaluation

Evaluation of the model is a significant step in the data mining process. It supports making effective decisions on the knowledge extracted from data mining techniques. By using the performance operator in RapidMiner, the models can be evaluated in different perspectives. Several criteria are available in RapidMiner to evaluate different kinds of models. To evaluate binomial classification tasks accuracy, recall, precision and AUC (Area Under the Curve) can be used. Accuracy and Kappa Statistic can be used for polynomial classification tasks. Mean Squared Error and Root Mean Squared Error can be used to evaluation of regression tasks.

7.3 Evaluation Parameters

Different parameters are available to cater different types of analysis as mentioned above. Since this study has five stages of predictions these parameters have been used according to the model. Each parameter can be described as follows.

Accuracy can be used in measuring the correct rate of predictions. The classification error measures the relative number of misclassified examples. Kappa statistic is a more robust measure since it considers the correct prediction occurring by chance rather than a simple percentage of correct prediction. There are various other performance metrics are available such as absolute error, relative error, root mean squared error, weighted mean recall etc. Correlation is also an important aspect to evaluate a model which returns the correlation coefficient between the class label and the independent variables. Another important metric is squared correlation that returns the squared correlation coefficient between the label and the prediction attributes. Absolute error refers to the

prediction's average absolute deviation from the actual value. The actual values are the values of the class label. Class weight is also an excellent parameter which identifies the weights of all the classes. Root mean squared error is also an excellent measurement for numerical predictions. It is referred as the difference between predicted values and the observed values. If it is lower, it indicates a perfect fit in the model [60]. The results of applying the evaluation parameters for each algorithm are shown in the next few subtopics.

7.4 Confusion Matrix

Confusion matrix can be also used to measure the performance of classification tasks. It has two possible values: positive and negative as well as two states of prediction results: right or wrong. It creates a matrix with four entries as depicted in Figure 7.1. Based on the confusion matrix different performance evaluation criteria can be built such as precision, recall, accuracy, positive predicted value, negative predicted value and specificity. It supports in identifying errors of the classifier and types of errors as well. The confusion matrix generated for the first stage of predictions is depicted in Figure 9.7 in appendix.

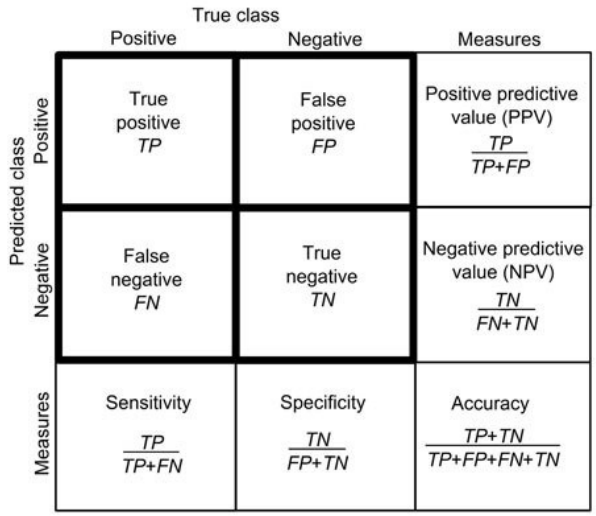


Figure 7.1: Confusion Matrix

True Positive (TP) – Refers to positive examples that are correctly predicted. As an example, a loan is predicted as arrears and actually it is.

True Negative (TN) – Refers to negative examples that are correctly predicted. As an example, a loan is predicted as non-arrears and actually it is.

False Positive (FP) - Refers to positive examples that are incorrectly predicted. As an example, a loan is predicted as arrears and actually it is not.

False Positive (FP) - Refers to negative examples that are incorrectly predicted. As an example, a loan is predicted as non-arrears and actually it is not.

7.5 Prediction 01 – Status of the payments

The first prediction predicts the status of the payment (predicting whether a loan will be get fully paid or not). Five algorithms were selected to apply here, based on the thorough literature review done in chapter 2. A balanced data set with around 3050 records were used with the previously mentioned attributes. The results of different algorithms applied in the first stage were depicted in the below table.

Table 7.1: Prediction 01- Performance Metrics of Different Algorithms

Algorithm	Accuracy	Classification Error	Correlation	Squared Correlation	Kappa Statistic
Naïve Bayes	62.31%	37.69%	25.2%	6.3%	24.8%
Rule Induction	65.28%	34.72%	31.8%	10.1%	30.8%
Decision Tree	65.52%	34.48%	42.8%	18.3%	31.8%
KNN	68.83%	31.17%	37.6%	14.2%	37.6%
Random Forest	72.27%	27.73%	44.7%	20.0%	44.6%

By considering the above performance metrices the Random Forest algorithm was selected for the first stage of the analysis to predict whether a loan will get paid fully or not. Here, Figure 7.2 depicts the performance vector of the Random Forest algorithm. The images of the results generated by RapidMiner was presented in appendices.

PerformanceVector

```
PerformanceVector:
accuracy: 72.27% +/- 44.77% (micro average: 72.27%)
ConfusionMatrix:
True:   No      Yes
No:     1099    464
Yes:    357     1041
classification_error: 27.73% +/- 44.77% (micro average: 27.73%)
ConfusionMatrix:
True:   No      Yes
No:     1099    464
Yes:    357     1041
kappa: 0.446
ConfusionMatrix:
True:   No      Yes
No:     1099    464
Yes:    357     1041
correlation: 0.000 +/- 0.000 (micro average: 0.447)
squared_correlation: 0.000 +/- 0.000 (micro average: 0.200)
```

Figure 7.2: Performance Vector of Prediction 01

7.6 Prediction 02 – Payment Percentage of Non-performing Loans

In the second stage it predicts the non-performing loans in different perspectives. If a particular loan is identified as a non-performing loan, then the model predicts the percentage of the non-payments. The results obtained from the first prediction was imported for this second phase by exporting the results into an excel file. Then the samples are filtered out to get the records which are actually arrears and the prediction is also arrears.

Percentage was discretized into four categories as high (will pay above 75%), medium (will pay between 50%-75%), low (will pay between 25%-50%) and very low (will pay below 25%). Hence, the prediction will be a nominal one. All the algorithms applied in prediction one, were applied here also. Parameters of the algorithms were tuned to get more performance and the KNN algorithm was selected for this stage since it showed the highest performance.

Table 7.2: Prediction 02- Performance Metrics of Different Algorithms

Algorithm	Accuracy	Classification Error	Correlation	Squared Correlation	Kappa Statistic
Naïve Bayes	53.9%	46.11%	19.1%	3.6%	13.2%
Rule Induction	62.06%	37.94%	32.0%	10.3%	29.5%
Decision Tree	62.63%	37.37%	36.2%	13.1%	28.7%
Random Forest	72.81%	27.19%	56.4%	31.8%	50.0%
KNN	73.87%	26.13%	57.3%	32.8%	52.3%

7.7 Prediction 03 – Exit Method

As the next prediction, if a particular loan is identified as a non-performing loan, it will check the exit method of the borrower. There are some borrowers who have exit the loan after paying installments regularly for some period. On the other hand, there are some borrowers who have paid the installments in an irregular way by keeping number of exit points throughout the loan period. The third prediction will predict whether the borrower will exit the loan after set of regular payments or irregular payments. When we get the results data of the first prediction into here, there is an imbalance of the dataset since there are more regularly done payments. Because of that the model will get overfit. Therefore, a balanced dataset was imported here by filtering out 1420 non-performing loans out of 1535. There, 720 samples consist of regularly done payments and 720 samples consist of irregularly done payments. Here also the previously used classification algorithms were applied and according to the highest accuracy, KNN was selected for this stage.

Table 7.3: Prediction 03- Performance Metrics of Different Algorithms

Algorithm	Accuracy	Classification Error	Correlation	Squared Correlation	Kappa Statistic
Rule Induction	94.02%	5.98%	88.4%	78.5%	88%
Naïve Bayes	95.06%	4.94%	90.2%	81.6%	90.1%
Decision Tree	95.29%	4.71%	90.7%	82.5%	90.5%
Random Forest	95.75%	4.25%	91.5%	84%	91.5%
KNN	96.32%	3.68%	92.7%	86%	92.6%

7.8 Prediction 04 – Exact Exit Point

Then, if the borrower exits the loan after set of regular payments it will predict the exact exit point. In other words, it predicts in which installment onwards, they will stop paying the loan. Here, the following algorithms were used as this is a numerical prediction. Based on the highest performance, KNN was selected for this prediction. It has the highest correlation of nearly 74%.

Table 7.4: Prediction 04- Performance Metrics of Different Algorithms

Algorithm	Root Mean Squared Error	Absolute Error	Correlation	Squared Correlation	Squared Error
ANN	3.13%	2.4%	76.7%	59.3%	9.99%
SVM	3.12%	2.51%	75.6%	57.5%	9.83%
Linear Regression	3.05%	2.43%	76.8%	59.2%	9.39%
Decision Tree	2.85%	2.20%	80.4%	64.7%	8.18%
KNN	2.78%	2.14%	81.8%	66.9%	7.75%

7.9 Prediction 05 – Number of Exit Points

In the final step, if the borrower exits the loan after set of irregular payments, the model predicts the number of stop points that can occur during the loan period. In this scenario, it cannot be identified an exact exit point of the loan since the borrowers have done the payments irregularly. For example, they have stopped the loan for few months and then they have paid it again for some time and then they have stopped it again. In such kind of scenarios, the model will predict how many stop points can be occurred throughout the loan period, which will be useful for the institutes to make effective decisions based on that.

As this is a numerical prediction again the same set of algorithms used for prediction 04 were utilized here. The number of exit points are ranged between 2 to 9, hence, there is a high correlation between the attributes. It may be a reason for generating less

percentages of performance at this stage. Based on the highest performance, linear regression was selected in this stage. Table 7.5 shows the results of each algorithm.

Table 7.5: Prediction 05- Performance Metrics of Different Algorithms

Algorithm	Root Mean Squared Error	Absolute Error	Correlation	Squared Correlation	Squared Error
ANN	2.11%	1.75%	5.4%	0.9%	4.53%
Decision Tree	1.90%	1.55%	5.7%	0.6%	3.6%
SVM	1.88%	1.50%	7.1%	2.1%	3.56%
KNN	2.05%	1.67%	4.4%	0.7%	4.20%
Linear Regression	1.86%	1.52%	8.7%	2.2%	3.49%

7.10 Summary

This chapter presents the evaluation of the results of data mining algorithms applied in different stages of predictions. To find the most accurate model, different parameters were used. According to the finding to predict the status of the payments, Random Forest is the most accurate algorithm. To predict the payment percentage of non-performing loans, KNN shows the highest performance. Then in predicting the exit method again KNN is the best. In the fourth stage, the numerical prediction on analyzing the exact exit point, KNN shows the highest performance. In the last stage of predicting the number of exit points, Linear Regression shows the highest performance.

Chapter 8 : Conclusion and Future Works

8.1 Introduction

As the last chapter of the research study, this chapter concludes the study by providing an overview of the study. As in any study, this study also has a few limitations. This chapter discusses those limitations and moreover, it identifies future enhancements which are helpful in improving the performance of the suggested solution.

8.2 Conclusion

Predicting the credit risk is essential in the financial sector for the reduction of non-performing loans and for better decision making. Among various types of loans SME loans are imperative. SMEs laid the foundation of the economic growth of a country. Besides, in Sri Lankan context, the number of non-performing loans have increased drastically throughout the last few years. Banks and other financial institutes use various techniques to predict SME loan credit risk, but there are some inefficiencies in the process. This study adopted a comprehensive model to predict different aspects of SME credit risk by using data mining algorithms to address that issue. The predictions were done in five stages to analyze the non-performing loans in different perspectives. This study helps financial institutes to predict the credit risk of SME borrowers and make effective decisions based on that.

8.3 Limitations

There are a few limitations of this study as in any research. The unavailability of more SME loan data can be a limitation since it focuses on big data analytics. If there are a large amount of data, then the predictions will be more accurate since this study use data mining techniques as a machine learning approach. There were some overfitting issues when comes to the later stages of predictions. It can be solved with a large set of data. Further, if data can be obtained from different financial institutes, including banks, it will give more insights to this study.

Variables were also selected considering only the available dataset from one financial institution. Hence, the results to the research problem can be limited due to the unobtainability of data from different financial institutions. Furthermore, this study has adopted only classification algorithms for the analysis. Though previous studies have

emphasized classification as the most used data mining technique in predicting credit risk, other types of data mining techniques can be applied such as clustering to identify the distinct clusters within that data set. By addressing these limitations, a more accurate and efficient data mining approach can be created to predict SME loan credit risk.

8.4 Future Development

As future enhancements of this study, the model will be finetuned with more examples since this study analyzed only 3000 SME loan data. Data can be obtained from different types of financial institutes such as state and private banks and other institutes. Further, it can be expanded for other types of loans including personal loans, professional loans, property loans, educational loans etc. In addition to that, an integrated model can be developed in predicting different aspects of the credit risk which are not addressed here. A hybrid model can be introduced by combining different data mining techniques such as clustering, regression, classification, association rule mining and so on. It will enhance the overall performance of the model and it will provide a valuable contribution for the financial sector in Sri Lanka.

8.5 Summary

As the final chapter of the study, this chapter concludes the study by describing the proposed solution and identifying its limitations and future enhancements. Finally, this study provides valuable insights into analyzing the credit risk of SME loan data by using data mining techniques as an emerging technology. It will enhance the decision-making process in the financial sector, reduce the credit risk of non-performing loans in SME sector and ultimately it will take the lead for the economic growth of Sri Lanka.

References

[1] I. G. N. N. Mandalaa, C. B. Nawangpalupi and F. R. Praktikto, "Assessing Credit Risk: an Application of Data Mining in a Rural Bank," *Procedia Economics and Finance,* pp. 406-412, 2012.

[2] S. Khemakhem, F. B. Said and Y. Boujelbene, "Credit risk asseessment for unbalanced datasets based on data mining, artificial neural network and support vector machines," *Journal of Modelling in Management,* 2018.

[3] F. Ciampi, "Corporate governance characteristics and default prediction modeling for small enterprises. An empirical analysis of Italian firms," *Journal of Business Research,* vol. 68, no. 5, pp. 1012-1025, 2015.

[4] M. Delivand, M. Moghadam, N. Shamami and M. Taghipour, "Investigating the effective factors in measuring customers' credibility with a combined approach of data mining and multidisciplinary decision making Problem," *Journal of Military Operations Research,* vol. 1, no. 1, 2021.

[5] Liu-yi and Z. Li-Gu, "Research and application of credit risk of small and medium-sized enterprises based on random forest model," *IEEE International Conference on Consumer Electronics and Computer Engineering,* 2021.

[6] M. o. I. a. Commerce, National Policy Framework for SME Development, Ministry of Industry and Commerce, 2015.

[7] H. M. K. S. Bandara, A. L. M. Jameel and H. Athambawa, "Credit Risk and Profitability of Banking Sector in Sri Lanka," *Journal of Economics, Finance and Accounting Studies,* vol. 1, p. 3, 2021.

[8] J. Gupta, N. Wilson, A. Gregoriou and J. Healy, "The value of operating cash flow in modelling credit risk for SMEs," *Applied Financial Economics,* vol. 24, no. 9, pp. 649-660, 2014.

[9] B. W. Yap, S. H. Ong and N. H. M. Husain, "Using data mining to improve assessment of credit worthiness via credit scoring models," *Expert Systems with Applications,* vol. 38, no. 10, 2011.

[10] R. M. Sum, W. Ismail, Z. H. Abdullah, N. Fathihin, M. N. Shah and R. Hendradi, "A new efficient credit scoring model for personal loan using data mining technique toward for sustainability management," *Journal of Sustainability Science and Management,* vol. 17, no. 5, pp. 60-76, 2022.

[11] W. Chen, G. Xiang, Y. Liu and K. Wang, "Credit risk Evaluation by hybrid data mining technique," *Systems Engineering Procedia,* vol. 3, 2012.

[12] M. P. Bach, J. Zoroja, B. Jaković and N. Šarlija, "Selection of variables for credit risk data mining models: preliminary research," *40th International Convention on Information and Communication Technology, Electronics and Microelectronics,* 2017.

[13] J. Aduda and S. Obondy, "Credit Risk Management and Efficiency of Savings and Credit Cooperative Societies: A Review of Literature," *Journal of Applied Finance & Banking,* vol. 11, no. 1, 2021.

[14] A. Hamid and T. Ahmed, "Developing Prediction Model of Loan Risk in Banks Using Data Mining," *Machine Learning and Applications: An International Journal (MLAIJ),* vol. 3, no. 1, 2016.

[15] S. Department, "Credit Supply Survey," Central Bank of Sri Lanka, 2021.

[16] D. o. S. o. N.-B. F. Institutions, "Proposed New Direction on Credit Risk Management for Liscened Finance Companies," Central Bank of Sri Lanka, 2019.

[17] S. Kodithuwakku, "Impact of credit risk management on the performance of commercial banks in Sri Lanka.," 2015.

[18] Hooman, A., Marthandan, G., Yusoff, W. F. W., , Omid, M. and Karamizadeh, S., "Statistical and data mining methods in credit scoring," *The Journal of Developing Areas,* vol. 50, no. 5, pp. 371-381, 2016.

[19] L. Liu, "A Self-Learning BP Neural Network Assessment Algorithm for Credit Risk of Commercial Bank," *Wireless Communications and Mobile Computing,* 2022.

[20] S. Klotz and A. Lindermeir, "Multivariate credit portfolio management using cluster analysis," *The Journal of Risk Finance,* vol. 16, no. 2, pp. 145-163, 2015.

[21] S. Khemakhem and Y. Boujelbene, "Predicting credit risk on the basis of financial and non-financial variables and data mining," *Review of Accounting and Finance,* vol. 17, no. 3, 2018.

[22] S. Moradi and F. M. Rafiei, "A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks," *Financial Innovation,* vol. 5, no. 1, pp. 1-27, 2019.

[23] Koyuncugil, A. S. and Ozgulbas, N., "Financial early warning system model and data mining application for risk detection," *Expert systems with Applications,* vol. 39, no. 6, pp. 6238-6253, 2012.

[24] Abedin, M. Z., Guotai, C., Hajek, P. and Zhang, T. , "Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk," *Complex & Intelligent Systems,* pp. 1-21, 2022.

[25] A. Islam, S. Yousuf and I. Rahman, "SME Financing in Bangladesh: A Comparative Analysis of Conventional and Islamic Banks," *Journal of Islamic Banking and Finance,* vol. 2, no. 1, pp. 79-92, 2014.

[26] V. Rivai, A. P. Veithzal and N. I. Ferry, "Bank dan Financial Institution Management convensional dan Shariah System.," 2007.

[27] S. M. Sadatrasoul, M. R. Gholamian, M. Siami and Z. Hajimohammadi, "Credit scoring in banks and financial institutions via data mining techniques: A literature review," *Journal of AI and Data Mining,* vol. 1, no. 2, 2013.

[28] J. Dobrovic, M. Lambovska, Gallo and Timkova, "Non-financial indicators and their importance in small and medium-sized enterprises," *Journal of Competitiveness,* vol. 10, no. 2, 2018.

[29] C. Huang, M. Chen and C. Wang, "Credit scoring with a data mining approach based on support vector machines," *Expert Systems with Applications,* vol. 33, no. 4, pp. 847-856, 2007.

[30] D. T. Larose, An introduction to data mining, John Wiley & Sons, 2005.

[31] Chen, Tsung-Kang, H.-H. Liao and W.-H. Chen, "CEO ability heterogeneity, board's recruiting ability and credit risk," *Review of Quantitative Finance and Accounting,* vol. 49, no. 4, pp. 1005-1039, 2017.

[32] J. Han and M. Kamber, Data Mining: Concepts and Techniques, San Francisco: Morgan Kaufmann, 2001.

[33] S. Strohmeier and F. Piazza, "Domain driven data mining in human resource management: A review of current research," *Expert Systems with Applications,* vol. 40, no. 7, 2013.

[34] Berry, M. J. and Linoff, G. S., Data mining techniques: for marketing, sales, and customer relationship management, John Wiley & Sons, 2004.

[35] A. K. Abdelmoula, "Bank credit risk analysis with k-nearestneighbor classifier: Case of Tunisian banks," *Accounting and Management Information Systems,* vol. 14, no. 1, 2015.

[36] M. A. Mukid, T. Widiharih, A. Rusgiyono and A. Prahutama, "Credit scoring analysis using weighted k nearest neighbor," *IOP Conf. Series: Journal of Physics,* 2018.

[37] P. Danenasa, G. Garsva and S. Gudas, "Credit Risk Evaluation Model Development Using Support Vector Based Classifiers," *International Conference on Computational Science,* 2011.

[38] A. Khashman, "Credit risk evaluation using neural networks: Emotional versus conventional models," *Applied Soft Computing,* vol. 11, no. 8, pp. 5477-5484, 2011.

[39] Brown, I. and Mues, C., "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Systems with Applications,* vol. 39, no. 3, pp. 3446-3453, 2012.

[40] Mukid, M. A., Widiharih, T., Rusgiyono, A. and Prahutama, A., "Credit scoring analysis using weighted k nearest neighbor," *Journal of Physics: Conference Series ,* 2018.

[41] Yu, L., Yue, W., , Wang, S. and Lai, K. K., "Support vector machine based multiagent ensemble learning for credit risk evaluation," *Expert Systems with Applications,* vol. 37, no. 2, pp. 1351-1360, 2010.

[42] Arora, N. and Kaur, P. D., "A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment," *Applied Soft Computing,* 2020.

[43] Uddin, M. S., Chi, G., Al Janabi, M. A. and Habib, T., "Leveraging random forest in micro-enterprises credit risk modelling for accuracy and interpretability," *International Journal of Finance & Economics,* 2020.

[44] Wang,, Yuelin, , Yihan Zhang, , Yan Lu, and Xinran Yu, "A Comparative Assessment of Credit Risk Model Based on Machine Learning - a case study of bank loan data," *Procedia Computer Science,* pp. 141-149, 2020.

[45] Z. Tian, Xiao, J., , Feng, H. and Wei, Y., "Credit risk assessment based on gradient boosting decision tree," *Procedia Computer Science,* pp. 150-160, 2020.

[46] J. Zurada, "Rule Induction Methods for Credit Scoring," *Review of Business Information Systems (RBIS),* vol. 11, no. 2, pp. 11-22, 2007.

[47] G. T. Albanis, " Implementing neural networks, classification trees, and rule induction classification techniques: an application to credit risk," *Applied quantitative methods for trading and investment,* pp. 213-237, 2003.

[48] Chen, Z., Li, Y. and Li, H. , "Rule Induction in the credit risk management of Inventory financing based on rough set," *7th International Conference on*

Education, Management, Information and Mechanical Engineering, pp. 938-942, 2017.

[49] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. and Vanthienen, J., "Benchmarking state-of-the-art classification algorithms for credit scoring," *Journal of the operational research society,* vol. 54, no. 6, pp. 627-635, 2003.

[50] A. Krichene, "Using a naive Bayesian classifier methodology for loan risk assessment: Evidence from a Tunisian commercial bank," *Journal of Economics, Finance and Administrative Science,* vol. 22, no. 42, pp. 3-24, 2017.

[51] Pandey, T. N., Jagadev, A. K., , Mohapatra, S. K. and Dehuri, S., "Credit risk analysis using machine learning classifiers," *International Conference on Energy, Communication, Data Analytics and Soft Computing ,* pp. 1850-1854, 2017.

[52] Kou, G., Peng, Y. and Wang, G. , "Evaluation of clustering algorithms for financial risk analysis using MCDM methods," *Information sciences,* pp. 1-12, 2014.

[53] I. H. Witten and E. Frank, Data Mining - Practical Machine Learning Tools and Techniques, San Francisco: Morgan Kaufmann, 2005.

[54] G. Paolo, "Bayesian data mining, with application to benchmarking and credit scoring," *Applied,* vol. 17, no. 1, pp. 69-81, 2001.

[55] H.-C. Wu, Y.-H. Hu and Y.-H. Huang, "Two-stage credit rating prediction using machine learning techniques," *Kybernetes,* vol. 43, no. 7, pp. 1098-1113, 2014.

[56] D. T. Larose and Larose, C. D, Discovering knowledge in data: an introduction to data mining, John Wiley & Sons., 2014.

[57] B. Liu, Supervised learning. In Web data mining, Berlin, Heidelberg: Springer, 2011.

[58] Han, J., Pei, J. and Kamber, M, Data mining: concepts and techniques, Elsevier, 2011.

[59] M. A. Mukid, T. Widiharih, A. Rusgiyono and P. A, "Credit scoring analysis using weighted k nearest neighbor," *Journal of Physics,* 2018.

[60] "RapidMiner Documentation," RapidMiner, 2021. [Online]. Available: https://docs.rapidminer.com/. [Accessed 06 12 2021].

[61] Crook, J. N., Edelman, D. B. and Thomas, L. C., "Recent developments in consumer credit risk assessment," *European Journal of Operational Research,* vol. 183, no. 3, pp. 1447-1465, 2007.

Chapter 9 : Appendix

9.1 Statistical Visualizations of the Data Set

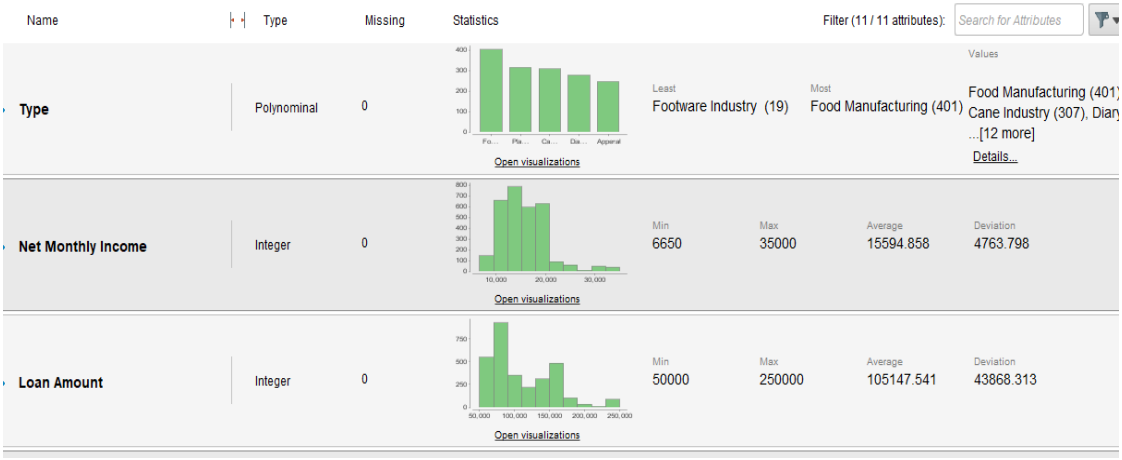


Figure 9.1: Statistical Details of SME Loan Data - 1

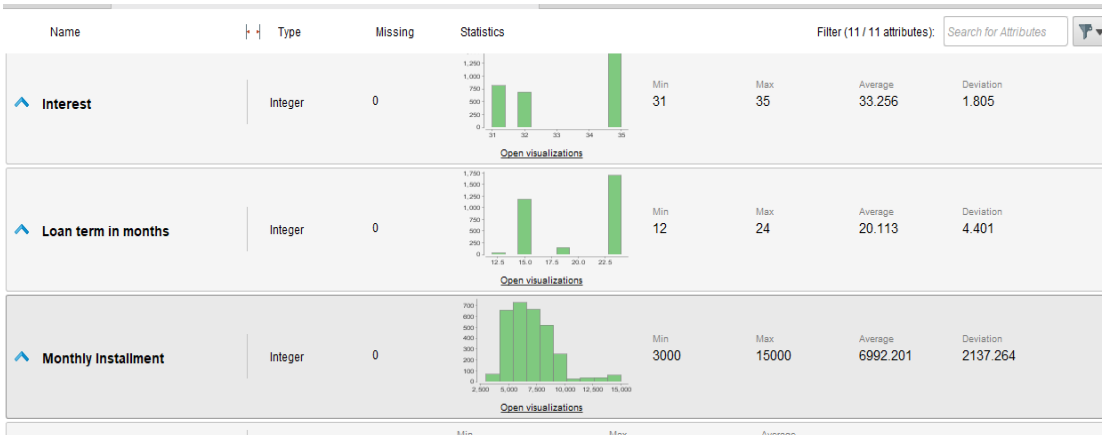


Figure 9.2: Statistical Details of SME Loan Data - 2

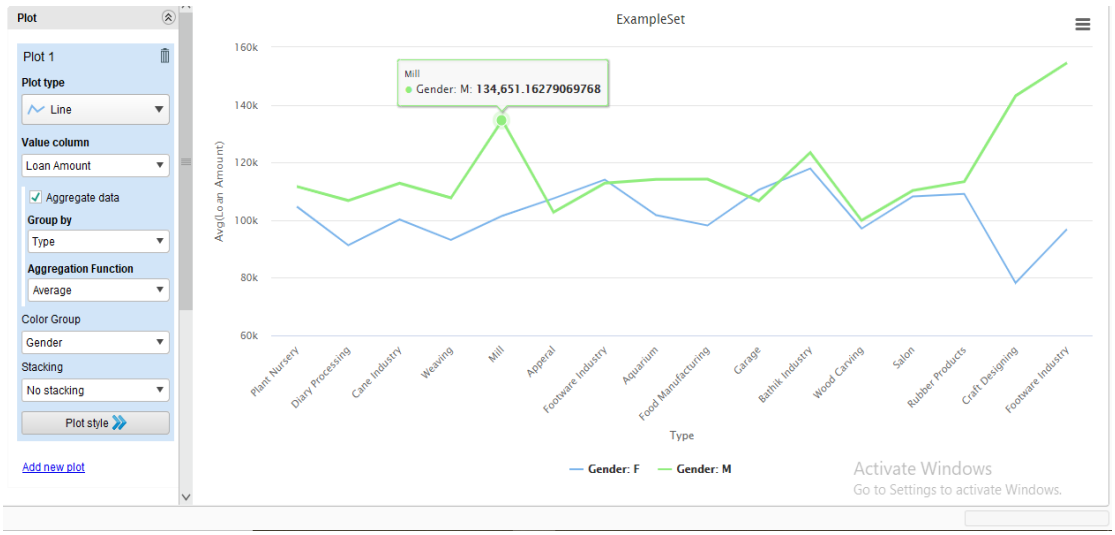


Figure 9.4: Business Type and Gender-wise Representation of Loan Amount

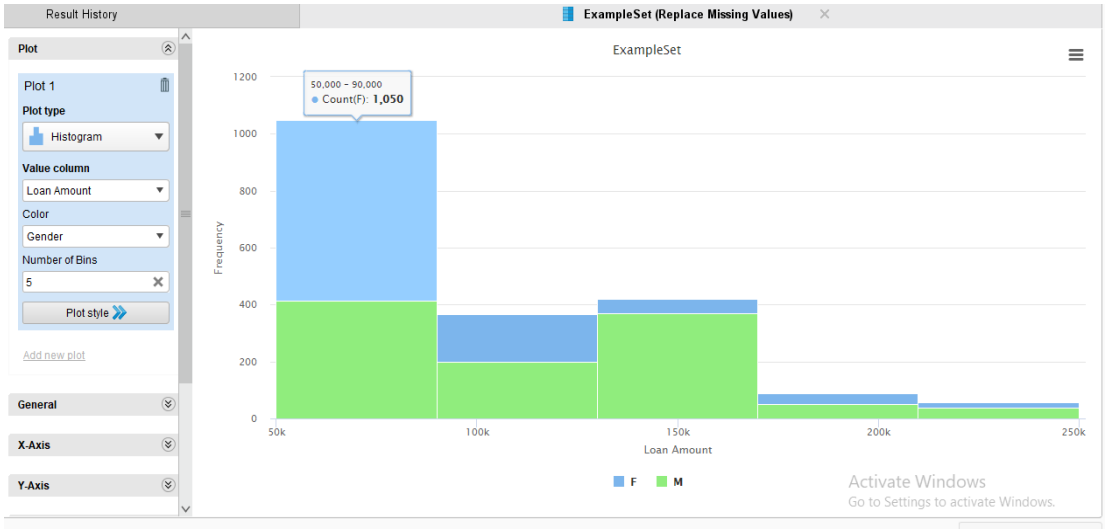


Figure 9.3: Gender-wise Representation of Loan Amount

9.2 Evaluation Results of the Selected Algorithm in Prediction 01

accuracy: 72.27% +/- 44.77% (micro average: 72.27%)

	true No	true Yes	class precision
pred. No	1099	464	70.31%
pred. Yes	357	1041	74.46%
class recall	75.48%	69.17%	

Figure 9.5: Accuracy of Random Forest Algorithm

Open in Turbo Prep Auto Model Filter (2,961 / 2,961 examples): all

Row No.	Whether Arr...	prediction(...	confidence(...	confidence(...	outlier	Net Monthly ...	Loan Amount	Loan term in...	Age	Percentage	Gender
1	Yes	Yes	0.465	0.535	false	-1.100	-1.136	-1.148	range3 [43.5...	High	M
2	No	Yes	0.116	0.884	false	1.090	0.577	-0.467	range1 [-∞ - 3...	High	F
3	No	No	0.606	0.394	false	0.722	0.349	-0.467	range2 [32.2...	High	F
4	No	Yes	0.193	0.807	false	-1.077	-1.136	-1.148	range2 [32.2...	High	M
5	No	No	0.713	0.287	false	0.291	0.577	0.895	range2 [32.2...	High	M
6	No	No	0.920	0.080	false	0.775	1.034	0.895	range1 [-∞ - 3...	Very Low	M
7	No	No	0.941	0.059	false	0.249	0.577	0.895	range2 [32.2...	High	F
8	No	No	0.685	0.315	false	-1.086	-1.136	-1.148	range2 [32.2...	High	F
9	No	No	0.730	0.270	false	0.395	0.120	-0.467	range3 [43.5...	High	M
10	No	No	0.886	0.114	false	-0.014	0.349	0.895	range3 [43.5...	High	M
11	No	No	0.571	0.429	false	0.395	0.120	-0.467	range2 [32.2...	High	F
12	No	Yes	0.216	0.784	false	-0.508	-0.793	-1.148	range4 [54.7...	High	F
13	No	Yes	0.476	0.524	false	-0.907	-1.021	-1.148	range3 [43.5...	High	F
14	No	No	0.810	0.190	false	-0.886	-1.021	-1.148	range1 [-∞ - 3...	High	M

Figure 9.6: Results of the Prediction 01

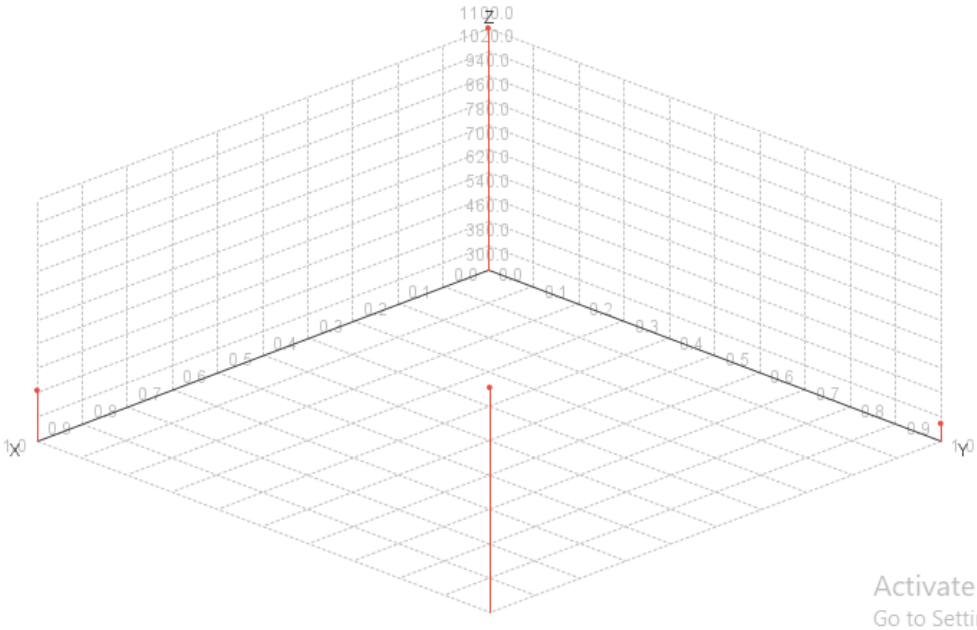


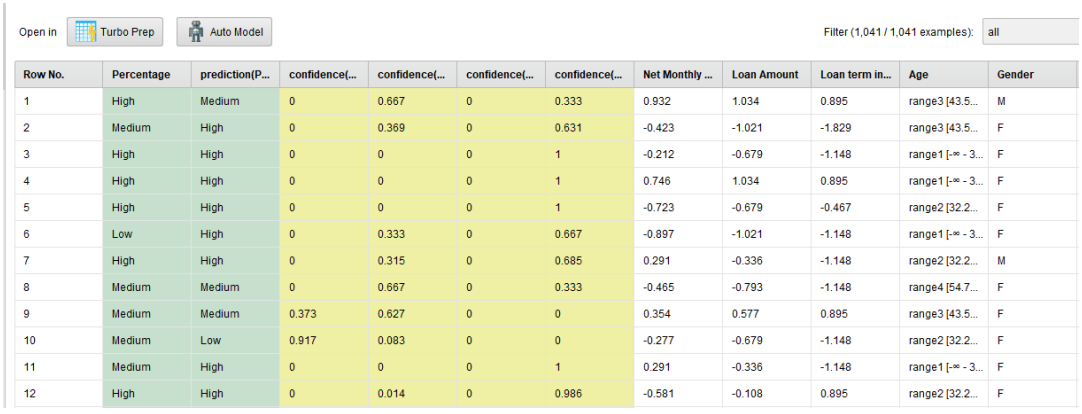
Figure 9.7: Confusion Matrix of Prediction 01

9.3 Evaluation Results of the Selected Algorithm in Prediction 02

PerformanceVector

```
PerformanceVector:
accuracy: 73.87% +/- 3.77% (micro average: 73.87%)
ConfusionMatrix:
True:    High    Very Low        Medium  Low
High:    453     0       126     13
Very Low:        0       2       2       0
Medium: 92       0       274     16
Low:     4       0       19      40
classification_error: 26.13% +/- 3.77% (micro average: 26.13%)
ConfusionMatrix:
True:    High    Very Low        Medium  Low
High:    453     0       126     13
Very Low:        0       2       2       0
Medium: 92       0       274     16
Low:     4       0       19      40
kappa: 0.523 +/- 0.069 (micro average: 0.523)
ConfusionMatrix:
True:    High    Very Low        Medium  Low
High:    453     0       126     13
Very Low:        0       2       2       0
Medium: 92       0       274     16
Low:     4       0       19      40
correlation: 0.573 +/- 0.096 (micro average: 0.573)
squared_correlation: 0.336 +/- 0.110 (micro average: 0.328)
```

Figure 9.8: Performance Vector of Prediction 02



Open in Turbo Prep Auto Model Filter (1,041 / 1,041 examples): all

Row No.	Percentage	prediction(P...	confidence(...	confidence(...	confidence(...	confidence(...	Net Monthly ...	Loan Amount	Loan term in...	Age	Gender
1	High	Medium	0	0.667	0	0.333	0.932	1.034	0.895	range3 [43.5...	M
2	Medium	High	0	0.369	0	0.631	-0.423	-1.021	-1.829	range3 [43.5...	F
3	High	High	0	0	0	1	-0.212	-0.679	-1.148	range1 [-∞ - 3...	F
4	High	High	0	0	0	1	0.746	1.034	0.895	range1 [-∞ - 3...	F
5	High	High	0	0	0	1	-0.723	-0.679	-0.467	range2 [32.2...	F
6	Low	High	0	0.333	0	0.667	-0.897	-1.021	-1.148	range1 [-∞ - 3...	F
7	High	High	0	0.315	0	0.685	0.291	-0.336	-1.148	range2 [32.2...	M
8	Medium	Medium	0	0.667	0	0.333	-0.465	-0.793	-1.148	range4 [54.7...	F
9	Medium	Medium	0.373	0.627	0	0	0.354	0.577	0.895	range3 [43.5...	F
10	Medium	Low	0.917	0.083	0	0	-0.277	-0.679	-1.148	range2 [32.2...	F
11	Medium	High	0	0	0	1	0.291	-0.336	-1.148	range1 [-∞ - 3...	F
12	High	High	0	0.014	0	0.986	-0.581	-0.108	0.895	range2 [32.2...	F

Figure 9.9: Results of the Prediction 02

9.4 Evaluation Results of the Selected Algorithm in Prediction 03

Open in Turbo Prep Auto Model Filter (1,430 / 1,430 examples): all

Row No.	Stop Point	prediction(S...	confidence(I...	confidence(...	outlier	Loan Amount	Monthly Inst...	Age	Net Monthly ...	Gender	Status
1	Irregular	Irregular	0.820	0.180	false	1.078	0.839	range1 [-∞ - 3...	range2 [1333...	F	M
2	Regular	Regular	0.266	0.734	false	-1.097	-0.957	range3 [43 - ...	range1 [-∞ - 1...	F	M
3	Regular	Regular	0.122	0.878	false	-0.734	-0.290	range2 [32 - ...	range2 [1333...	F	M
4	Regular	Regular	0.203	0.797	false	-0.855	-0.495	range1 [-∞ - 3...	range2 [1333...	F	M
5	Regular	Regular	0.383	0.617	false	-1.338	-1.265	range4 [54 - ∞]	range1 [-∞ - 1...	F	M
6	Regular	Regular	0.420	0.580	false	-0.855	-0.495	range3 [43 - ...	range2 [1333...	F	M
7	Irregular	Irregular	0.615	0.385	false	-1.097	-1.726	range2 [32 - ...	range1 [-∞ - 1...	F	M
8	Regular	Regular	0.271	0.729	false	1.078	0.890	range2 [32 - ...	range2 [1333...	F	M
9	Regular	Regular	0.030	0.970	false	0.837	0.541	range2 [32 - ...	range2 [1333...	M	M
10	Regular	Regular	0.171	0.829	false	1.683	1.608	range1 [-∞ - 3...	range3 [2000...	M	M
11	Regular	Regular	0.075	0.925	false	0.837	0.582	range3 [43 - ...	range2 [1333...	M	M
12	Irregular	Irregular	0.922	0.078	false	1.078	0.839	range3 [43 - ...	range2 [1333...	M	M

Figure 9.10: Results of the Prediction 03

PerformanceVector

```
PerformanceVector:
accuracy: 62.10% +/- 3.65% (micro average: 62.10%)
ConfusionMatrix:
True:   Regular Irregular
Regular:        445     271
Irregular:      271     443
classification_error: 37.90% +/- 3.65% (micro average: 37.90%)
ConfusionMatrix:
True:   Regular Irregular
Regular:        445     271
Irregular:      271     443
kappa: 0.242 +/- 0.073 (micro average: 0.242)
ConfusionMatrix:
True:   Regular Irregular
Regular:        445     271
Irregular:      271     443
correlation: 0.243 +/- 0.073 (micro average: 0.242)
squared_correlation: 0.064 +/- 0.035 (micro average: 0.059)
```

Figure 9.11: Performance Vector of Prediction 03

9.5 Evaluation Results of the Selected Algorithm in Prediction 04

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 3.219 +/- 0.286 (micro average: 3.231 +/- 0.000)
absolute_error: 2.484 +/- 0.196 (micro average: 2.484 +/- 2.065)
squared_error: 10.437 +/- 1.799 (micro average: 10.437 +/- 18.416)
correlation: 0.740 +/- 0.055 (micro average: 0.741)
squared_correlation: 0.551 +/- 0.080 (micro average: 0.550)
```

Figure 9.12: Performance Vector of Prediction 04

Open in Turbo Prep Auto Model Filter (710 / 710 examples): all

Row No.	If arrears in ...	prediction(If ...	outlier	Gender = F	Gender = M	Status = M	Status = S	Type = Plant...	Type = Cane...	Type = Appe...	Type = Diary...
2	9	10.133	false	1	0	1	0	0	0	0	0
3	10	12.956	false	0	1	1	0	0	0	0	0
4	15	20.796	false	0	1	1	0	0	0	0	0
5	10	10.257	false	1	0	1	0	0	1	0	0
6	9	10.836	false	1	0	1	0	0	1	0	0
7	10	12.955	false	1	0	0	1	0	0	0	0
8	11	17.942	false	0	1	1	0	0	1	0	0
9	23	21.837	false	1	0	1	0	0	0	0	0
10	22	18.724	false	1	0	1	0	0	1	0	0
11	17	17.055	false	0	1	1	0	0	0	0	0
12	15	16.497	false	1	0	1	0	0	0	1	0
13	11	12.439	false	1	0	1	0	0	0	0	0

Figure 9.13: Results of the Prediction 04

9.6 Evaluation Results of the Selected Algorithm in Prediction 05

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 1.757 +/- 0.092 (micro average: 1.760 +/- 0.000)
absolute_error: 1.435 +/- 0.071 (micro average: 1.435 +/- 1.018)
squared_error: 3.096 +/- 0.316 (micro average: 3.096 +/- 4.406)
correlation: 0.156 +/- 0.093 (micro average: 0.141)
squared_correlation: 0.032 +/- 0.027 (micro average: 0.020)
```

Figure 9.14: Performance Vector of Prediction 05

Open in Turbo Prep Auto Model Filter (710 / 710 examples): all

Row No.	No.of Exit Po...	prediction(N...	outlier	Gender = F	Gender = M	Status = M	Status = S	Type = Appe...	Type = Salon	Type = Wea...	Type = Aqua...
3	3	2.959	false	1	0	1	0	0	1	0	0
4	2	4.091	false	1	0	1	0	0	0	0	0
5	3	3.023	false	1	0	1	0	0	1	0	0
6	2	3.155	false	1	0	1	0	0	0	0	0
7	4	3.350	false	1	0	1	0	0	0	0	0
8	3	2.695	false	1	0	1	0	0	0	0	0
9	3	3.707	false	0	1	1	0	0	0	0	0
10	2	3.659	false	0	1	1	0	0	0	0	0
11	2	3.185	false	1	0	1	0	0	1	0	0
12	3	3.937	false	0	1	1	0	0	0	0	0
13	2	3.866	false	1	0	1	0	1	0	0	0
14	4	3.924	false	0	1	1	0	0	0	0	0

Figure 9.15: Results of the Prediction 05