

# **Developing a Tool to Manage the Credit Risk using Data Mining**

Shamli Rajapakshe

139175M

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



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
March 2016  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

## 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 Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Shamli Rajapakshe

Date:

Signature of Student

Supervised by

Mr. Saminda Premaratne  University of Moratuwa, Sri Lanka.  
Senior Lecturer [www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)  
Electronic Theses & Dissertations

Faculty of Information Technology

University of Moratuwa

Date:

Signature of Supervisor

## **Acknowledgement**

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

I would like to express my profound sense of gratitude and respect to all those who helped me in various ways throughout the duration of this research project.



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

## Abstract

Ceasing vehicles are affecting to the credit liquidity of the leasing company. This research has been conducted to develop a tool to manage credit risk in leasing companies using data mining. This tool will predict the ability of recoverability of the loan and determine the most suitable plan for the customer. It is hypothesis that, using data mining technology, the credit risk of leasing companies can be managed. Past dataset from the leasing company has been used to create the data mining model. When a customer comes to lease a vehicle, decision maker will get the information from the customer and enter to the system as inputs then the system will predict the tendency of recoverability of the loan and will give the suitable plans for the customer after evaluating with the previously generated model. This system generated details will support the decision maker to take his decision. The overall design includes frontend software and it is connected to the WEKA API which issued under the GNU General Public License.

The data model that is used in this tool to manage credit risk in leasing companies has been tested by considering a data collected from the medium scale leasing company in Sri Lanka.



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)



# Table of Content

Declaration.....	i
Acknowledgement .....	ii
Abstract.....	iii
Table of Content .....	iv
List of Figures .....	vii
List of Tables .....	vii
Chapter 1.....	1
Introduction.....	1
1.1 Prolegomena.....	1
1.2 Background and Motivation.....	2
1.3 Problem Definition.....	2
1.4 Objectives.....	2
1.5 Summary.....	2
Chapter 2.....	3
Literature Review.....	3
2.1 Introduction.....	3
2.2 Credit Risk.....	3
2.3 Credit Risk Management.....	3
2.4 Comparative Study of Data Mining Model for Credit Card Application Scoring in Bank .....	4
2.5 Feature Selection in Credit Scoring Model for Credit Card Applicants .....	4
2.6 Predict the usage of laptops among students in rural areas using WEKA tool....	4
2.7 Data Mining: A prediction Technique for the workers in the PR Department of Orissa.....	5
2.8 comparative study of classification algorithms for credit card approval using Waikato Environment of Knowledge Analysis (WEKA).....	5
2.9 Summary .....	6
Chapter 3.....	7
Technology Adopted.....	7
3.1 Introduction .....	7
3.2 Data Mining.....	7

3.3 Java.....	7
3.4 WEKA / WEKA API .....	8
3.5 SPSS .....	8
3.6 Summary .....	8
Chapter 04.....	9
Data mining approach to manage credit risk .....	9
4.1 Introduction .....	9
4.2 Hypothesis.....	9
4.3 Inputs.....	9
4.4 Outputs .....	9
4.5 Process and features .....	9
4.7 Users.....	10
4.8 Summary .....	10
Chapter 5.....	11
Design of the Prediction Tool.....	11
5.1 Introduction .....	11
5.2 Data Input Procedure.....	11
5.3 Risk Assessment Process.....	12
5.5 Top level Design .....	13
5.5.1 Gather data in the relevant field and Pre preparation .....	13
5.5.2 Load data to the WEKA .....	13
5.5.3 Model building by using suitable classification technology .....	13
5.5.4 Requirement gathering for the application .....	13
5.5.5 Design the application .....	14
5.5.6 Implement a front-end application by using JAVA.....	15
5.5.7 Integration.....	15
5.5.8 Testing .....	15
5.6 User Interfaces.....	16
5.7 Summary .....	18
Chapter 6.....	19
Implementation of the Tool .....	19
6.1 Introduction .....	19
6.2 Implementation of CRMT (Credit Risk Management Tool) .....	19

6.2.1 Pre preparation.....	19
6.2.2 Best Model Selection.....	21
6.2.3 Risk Assessment Process.....	23
6.2.4. Plan Prediction Process .....	25
6.2.5 Requirement gathering for the application .....	26
6.2.7 Implement a front-end application by using JAVA.....	26
6.2.8 Integration.....	26
6.2.9 Testing .....	26
6.2.10 Building an Evaluation Model.....	26
6.3 Summary .....	26
Chapter 7.....	27
Evaluation .....	27
7.1 Introduction .....	27
7.2 Data Model Testing.....	27
7.3 Test the system for randomly selected values .....	277
7.4 Summary .....	30
Chapter 8.....	31
Conclusion and further work .....	31
8.1 Introduction .....	31
8.2 Conclusion of CRMT (Credit Risk Management Tool).....	31
8.3 Limitations of CRMT (Credit Risk Management Tool) .....	31
8.4 Further work.....	31
8.5 Summary .....	31
References.....	32
Appendix A: Categorization .....	34
Appendix B: Variable Selection Procedure (Technology: Chi – Squared) .....	36
Appendix C: Best Model Selection.....	44



## List of Figures

Figure 5.1: Data Input Procedure.....	11
Figure 5.2: Risk Assessment Procedure.....	12
Figure 5.3: Plan Prediction Procedure .....	12
Figure 5.4: Top level design .....	13
Figure 5.5: Use case.....	14
Figure 5.6: Frontend application with other connected components.....	15
Figure 5.7: Data Input Interface.....	16
Figure 5.8: Risk Analysis Interface .....	16
Figure 5.9: Model Prediction .....	17
Figure 5.10: Requested amount Prediction.....	17
Table 6.1: Variable Selection Summary table .....	20
Figure 6.1: Accuracy levels of algorithms.....	23
Figure 6.2: Error rates of algorithms.....	23
Figure 6.3: Building the Classifier.....	23
Figure 6.4: Using Classifier for Classification .....	24
Figure 7.1: Test the selected model with testing data set .....	217
Figure 7.2: Model evaluation with the test data set .....	218
Figure 7.3: Customer Details.....	28
Figure 7.4: Most Suitable vehicle model for particular customer.....	29
Figure 7.5: Most suitable loan ranges for particular user .....	29

## List of Tables

Table 6.1: Variable Selection Summary .....	21
Table 6.2: Algorithm Accuracy Summary.....	21

# Chapter 1

## Introduction

### 1.1 Prolegomena

Credit is an important in the making of investments which measures the whole performance of economy all around the world. Lending industry is playing major role when funds are made available for various investment purposes. This lending is most of the time getting trouble with a number of risks which include risk of default and risk of recovery of the defaulted loan.

Risk management today is in the spotlight due to the tightened regulatory supervision followed by the volatility in financial markets. To identify the unavoidable uncertainty associated with business as an integral part of Corporate Governance, risk management can be used as a tool. [6].



University of Moratuwa, Sri Lanka.

Electronic Theses & Dissertations

[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

Credit risk can in turn cause cash flow problems and affect the company's liquidity. Minimize the risk and maximize company's risk adjusted rate of return can be considered as the objective of credit risk management [7].

The process requires that sufficient information be gathered to enable a comprehensive assessment of true risk profile of the borrower [8].

In order to give a more support to identify the risk profile of the borrower, the company can use their past data in efficient manner to predict the recoverability of the loan. As well as the customer would be able to get the information about, what will be the most suitable plan for them according to the previous customers behavior.

Data mining and warehousing methodologies can be used to implement a tool to predict the risk and, to suggest selecting the most suitable plan by using previous customer data of the company.

Data mining techniques can be used in the range from really complex to simple. Each and every technique is dedicated to a slightly different purpose or goal. As well as their behaviors are different in various situations, such as dataset, output, predictions etc. In essence, data mining helps organizations analyze incredible amounts of data in order to detect common patterns or learn new things. It would not be possible to process all this data without automation.

The proposing tool would be a tool which can be used in leasing companies. An interface has been provided to insert the data when the arrival of new customer. After that they can simply generate the predictions, just using this tool as a decision support system.

## **1.2 Background and Motivation**

At present analysis are done with particular dataset and those data sets are being used to check for the most accurate results and algorithm that can be used. However, there is no any tools have been developed to use these technologies in real world situations.

## **1.3 Problem Definition**

Ceasing vehicles are affecting to the credit liquidity of the leasing company.

## **1.4 Objectives**

Developing a tool to manage credit risk can be used in leasing industry and to get a support to the decision maker of the company as a decision support tool.

## **1.5 Summary**

Credit Risk management is highly important in the lending industry and present analysis has been conducted to check the most suitable model for the particular data set.

## Chapter 2

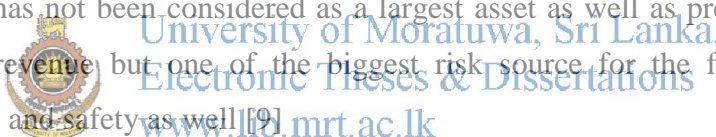
### Literature Review

#### 2.1 Introduction

There are some published materials related to the credit risk and credit risk management. Other than that some researches have been conducted to manage credit risk in related domains, Use of data mining for prediction, selection of best model for predictions for a particular data set.

#### 2.2 Credit Risk

Sufi Faizan Ahmed and Qaisar Ali Malik are mentioned that, Risk faced by investor to lose money from borrower who fails to make payments is known as Credit risk. This may outcome in default or default risk. Investors may lose interest and principal that can outcome in increased cost of gatherings and decreased cash flows. Loan portfolio has not been considered as a largest asset as well as predominate source to generate revenue but one of the biggest risk source for the financial institutions soundness and safety as well [9].



Generally banks are focused on three main types of risk: Credit, Operational and Market [10]. Credit risk is the single largest risk most banks face and arises from the possibility that loans or bonds held by a bank will not be repaid either partially or fully. As well as this is the potential loss a bank would suffer if a bank borrower, also known as the counterpart, fails to meet its obligations pay interest on the loan and Repay the amount borrowed in accordance with agreed terms [11]. Credit risk is typically represented by means of three factors: default risk, loss risk and exposure risk. Credit and default risk are often synonymous.

#### 2.3 Credit Risk Management

Credit risk management is a method that involves the identification of potential risks, the measurement of these risks, the appropriate treatment, and the actual implementation of risk models [12]. Credit risk assessment was the first tool developed in financial services 60 years ago. Establishing a standardized and practical assessment system for commercial

banks is of positive and practical significance to comprehensively improve the bank's management level and to effectively reduce and prevent credit risks [13].

#### **2.4 Comparative Study of Data Mining Model for Credit Card Application Scoring in Bank**

Evaristus Didik Madyatmadja and Mediana Aryuni have discussed Comparative Study of Data Mining Model for Credit Card Application Scoring in Bank. They stated that the growth of credit card application needs to be balanced with the anticipation of bad credit risk because it does not use security collateral as warranty. Credit scoring can be used to help the credit risk analysis in determining the applicant's eligibility. Data mining has been proven as a valuable tool for credit scoring. The proposed model applies classification using Naïve Bayes and ID3 algorithm. In this research, they have concluded that Naïve Bayes classifier has better accuracy. However, they have compared only two methods of data mining. Also they have mentioned that the data set that has been selected for the research is not representing the whole category, constructing credit scoring models is the change of patterns overtime, and more time may be needed to construct the final model [1].

#### **2.5 Feature Selection in Credit Scoring Model for Credit Card Applicants**

Evaristus Didik Madyatmadja and Mediana Aryuni have published an another research paper under topic of Feature Selection in Credit Scoring Model for Credit Card Applicants in XYZ Bank: A Comparative Study. In this research, they have compared four data mining technologies (Information Gain, Gain Ratio, GINI Index, Chi-Squared Statistics) and concluded that GINI Index and Information Gain feature selection methods performed relatively better. After feature selection applied, the model accuracy was increased. Furthermore, the training time was decreased and the final model became more simple because the reduction in the number of features. There is some limitation such as the data set that has been selected for the research was not representing the whole category [2].

#### **2.6 Predict the usage of laptops among students in rural areas using WEKA tool**

Nithya.M,Suba.S and Vaishnavi.B have conducted a research to analyze for what purpose the students using laptops more whether for entertainment or education or internet usage. In this Research, three classes have been divided to separate the



student's variety how they spend the precious time in laptops. Data have been collected from the students who are using laptops. The process was done by classification techniques in data mining with Waikato Environment of Knowledge Analysis (WEKA) tool. The goal of classification was to build a set of model that can correctly be predicted the classification of different objects. This research has concluded that laptops have been used for education, entertainment and internet usage purposes and more number of students was using their laptops for entertainment, chatting and access social networks. In this research, they have used only one classification method (Decision tree) for analysis [3].

### **2.7 Data Mining: A prediction Technique for the workers in the PR Department of Orissa**

Research by Neelamadhab Padhy and Rasmita Panigrahi, they have discussed the method of data mining which contains the large information about the Panchayat Raj Department (PR), a worker intensive organizations of Orissa. They have focused on some of the techniques, approaches and different methodologies of the demand forecasting. Here, they have designed a tool with a smart selection function to help users to make a judgement on the information. This system also provides calculation function to help users to work out a predication result. In their approach, they developed an automated system for attribute classification based on the algorithm with a very sound practical application of Linear Regression technique and also mentioned; no approaches or tools can guarantee to generate the accurate prediction in the organization. They have analyzed the different algorithm and prediction technique. Result has shown that the least median squares regression is known to produce better results than the classifier linear regression techniques from the given set of attributes. As comparison they have found that Linear Regression technique which takes the lesser time as compared to Least Median Square Regression. They have concluded that linear regression analysis can't handle the large data sets [4].

### **2.8 comparative study of classification algorithms for credit card approval using Waikato Environment of Knowledge Analysis (WEKA)**

Devendra Kumar Thiwary has discussed a comparative study of classification algorithms for credit card approval using Waikato Environment of Knowledge

Analysis (WEKA). In this Research paper, four classification algorithms (Decision Tree, Naive Bayes, Artificial Neural Network and Support Vector Machine) have been comparatively tested to find the optimum algorithm for classification and credit card application has been used for experimental purpose. Performance of the different classification algorithms using WEKA tool have been investigated by this research. As well as same experiment procedures have been followed in all four classification algorithms. They have concluded that Decision Tree classifier is the optimum algorithm with higher accuracy for the credit card data. As they mentioned, a major problem in the financial analysis is to build an ultimate model that successful in certain given information and a single data mining model is unable to fulfill all business requirements vice versa a business is also need more than one model to depend on [5].

## 2.9 Summary

It is evidenced from the literature that unrecoverable loans are affecting to the credit liquidity of the company. It is a research challenge and no adequate researches are done to solve this problem in leasing domain.



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations

This research has been carried out to solve this problem using data mining technology to develop a tool to support decision making system to reduce the percentage of cease vehicles by using past data of the company.

# Chapter 3

## Technology Adopted

### 3.1 Introduction

Chapter 2 presented that how the data mining can be used in predictions. This chapter discussed the technologies that can be used to develop a tool to manage credit risk in a leasing company.

### 3.2 Data Mining

Data mining can be used for the process of extraction of interesting, nontrivial, implicit, previously unknown and potentially useful patterns or knowledge from huge amounts of leasing customers' data, as well as for the prediction of future tendencies of customer payments. Extracted knowledge from the Data warehouse can be used for giving suggestions as planes for customers when they are going to lease a vehicle. It is the set of activities used to find new, hidden or unexpected patterns in data or unusual patterns in data. Using information contained within data warehouse, data mining can often provide answers to questions about an organization that a decision maker has previously not thought to ask. Data mining tools can answer business questions that traditionally were time consuming too to resolve.

### 3.3 Java

When we consider about the advantages of Java is its capability to move effortlessly from one computer system to another. The ability to run the same program on many different systems is important to World Wide Web software, and Java is doing well this by being platform-independent at both the source and binary levels. As part of its design, Java considers security. The Java language, compiler, interpreter, and runtime environment were each developed with security in mind. Java puts a lot of weight on early checking for possible errors, as Java compilers are able to detect many problems that would first show up during execution time in other languages [16]. As well as WEKA API is developed by using and Java language therefore, Using Java for accessing the WEKA API, is most compatible than using other languages.

### **3.4 WEKA / WEKA API**

WEKA can be used for the process of analyzing data, it contains tools for data preprocessing, regression, association rules, clustering, classification, and visualization. It is also well-suited for developing new machine learning schemes. It is open source software issued under the GNU General Public License. And it is possible to apply WEKA to big data.

### **3.5 SPSS**

Statistical Package for the Social Sciences (SPSS) has been used in the process of variable selection. It was acquired by IBM in 2009. SPSS is a commonly used program for statistical analysis in social science. It is also used by health researchers, education researchers, survey companies, government, market researchers, marketing organizations, data miners and others.

### **3.6 Summary**

Data mining technology can be used to analyze large data set, WEKA is open source software which supports the analysis and WEKA API contains the set of functions which are used in WEKA open source software. Java language can be used to communicate with the API. SPSS is used for statistics analysis



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
www.lib.mrt.ac.lk

## **Chapter 04**

### **Data mining approach to manage credit risk**

#### **4.1 Introduction**

Chapter 3 discussed the technology to develop a tool to manage the credit risk of the leasing company. This chapter presents the approach to develop a tool to manage credit risk in a leasing company using data mining technology under several headings, namely Hypothesis, Input, Output, Process and Features.

#### **4.2 Hypothesis**

Using data mining technology, the credit risk of leasing companies can be managed.

#### **4.3 Inputs**

Past data from a particular leasing company is using as an input through the user Interface of the tool. Data should be preprocessed before entering. As well as, selected attributes of new customer's details should be entered. (vehicle details, personal details, lease details etc.) by using user interface of the tool

#### **4.4 Outputs**

Predict tendency to cease or close the leasing agreement and suitable leasing plan has been predicted according to the given details of the customer (Profession and Monthly Income).

#### **4.5 Process and features**

Past data that have been collected from the particular party are subjected for preprocessing and are loaded into the proposed tool using its interface, then data mining algorithm is applied the model is built with the cooperation of WEKA API. After that, customer data is entered as an input data to the provided interface which is needed to evaluate the risk profile. Then prediction can be made whether there is tendency to cease the vehicle or close the agreement after completing the payments. As well as there is a feature to generate suitable leasing plans to the particular customer based on their Profession and Monthly Income. Finally decision maker can be used these system generated output to support their decision.

#### **4.7 Users**

Decision makers are the users who are getting benefit from this tool.

#### **4.8 Summary**

The chapter highlights how the novel approach offers a tool to manage credit risk in the process of decision making.



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

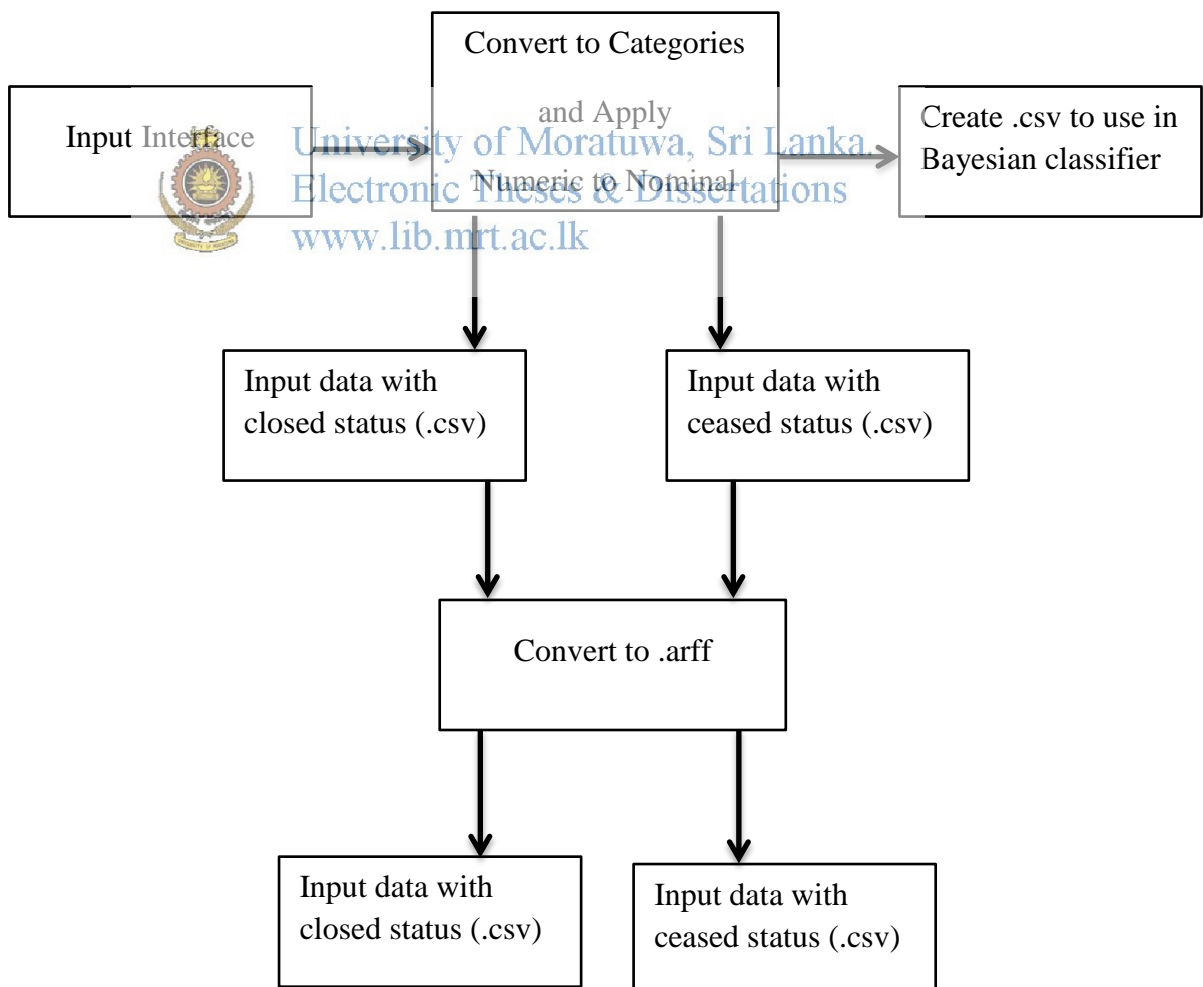
# Chapter 5

## Design of the Prediction Tool

### 5.1 Introduction

Chapter 4 presented the approach to develop a tool to manage credit risk in a leasing company by using data mining. This chapter elaborates the approach and describes the design of the solution. The top level design includes classifier, data set, and frontend application to call WEKA API and User Interface to input data and output results. In this tool, Design can be divided in to three parts. Namely, Data Input Procedure, Risk Assessment Process and Plan Prediction Process.

### 5.2 Data Input Procedure



**Figure 5.1: Data Input Procedure**

### 5.3 Risk Assessment Process

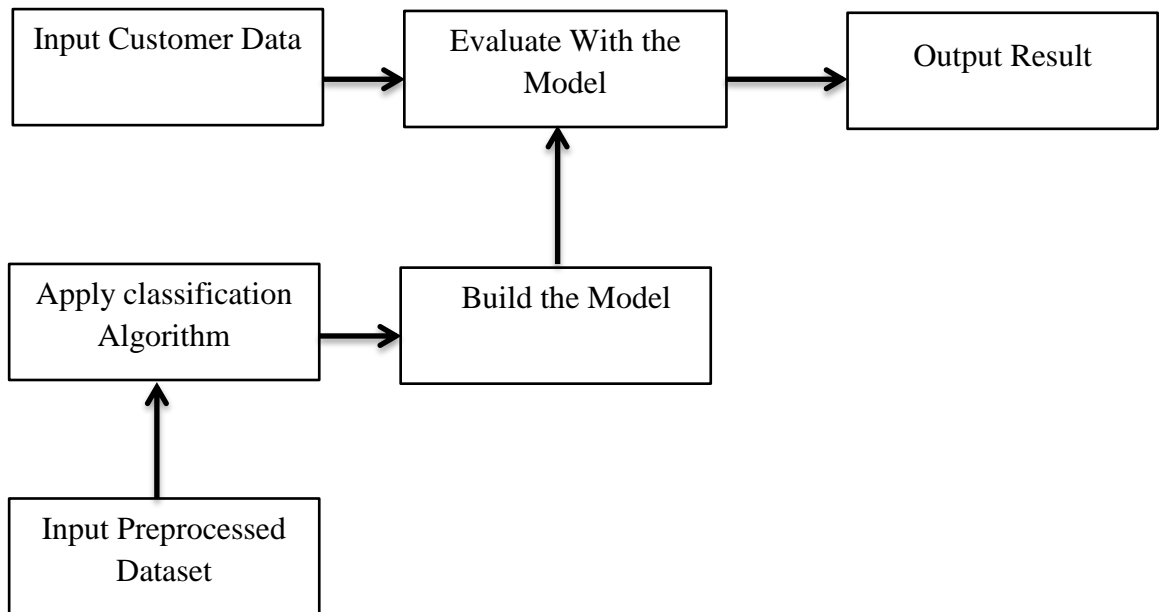


Figure 5.2: Risk Assessment Procedure

### 5.4 Plan Prediction Process

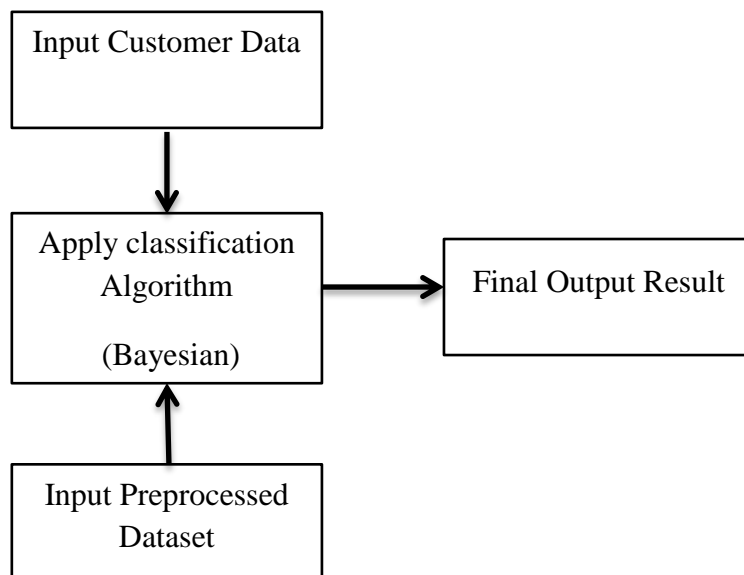


Figure 5.3: Plan Prediction Procedure



## 5.5 Top level Design

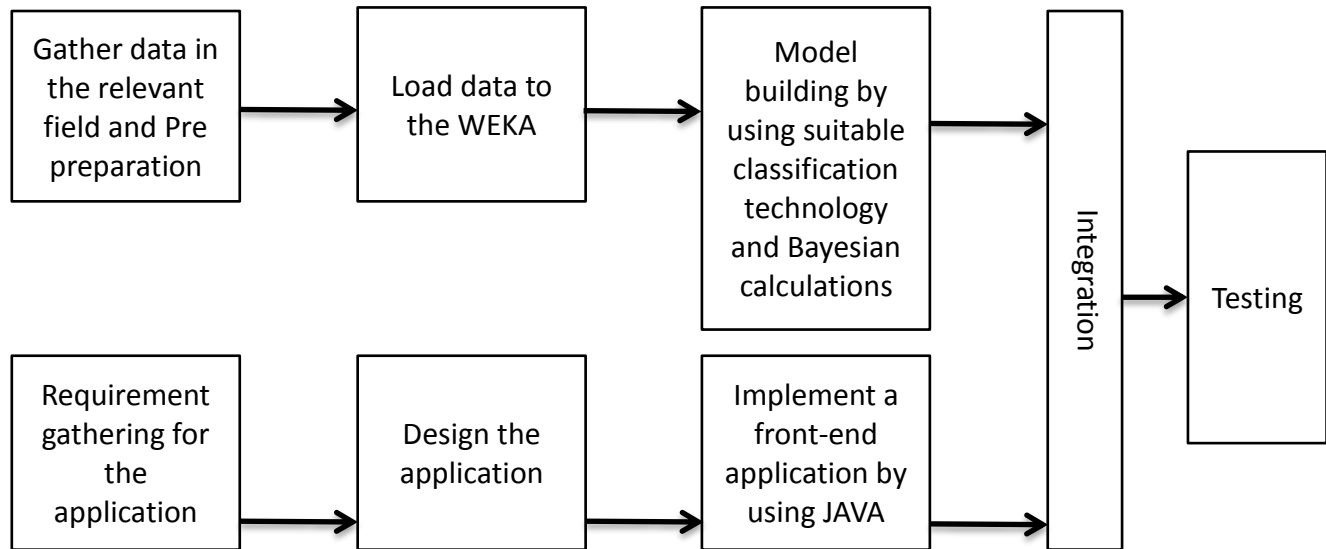


Figure 5.4: Top level design

### 5.5.1 Gather data in the relevant field and Pre preparation

After gathering the required information, collected data is subjected to the preprocessing process. Then, variables which are only affected to the dependent variable are selected and the final dataset is prepared. After that the application is built and that application is integrated with the final model with the help of WEKA API and required calculations are done for the plan prediction part. Finally whole system is tested.

### 5.5.2 Load data to the WEKA

Preprocessed dataset is loaded with no errors by using WEKA explorer.

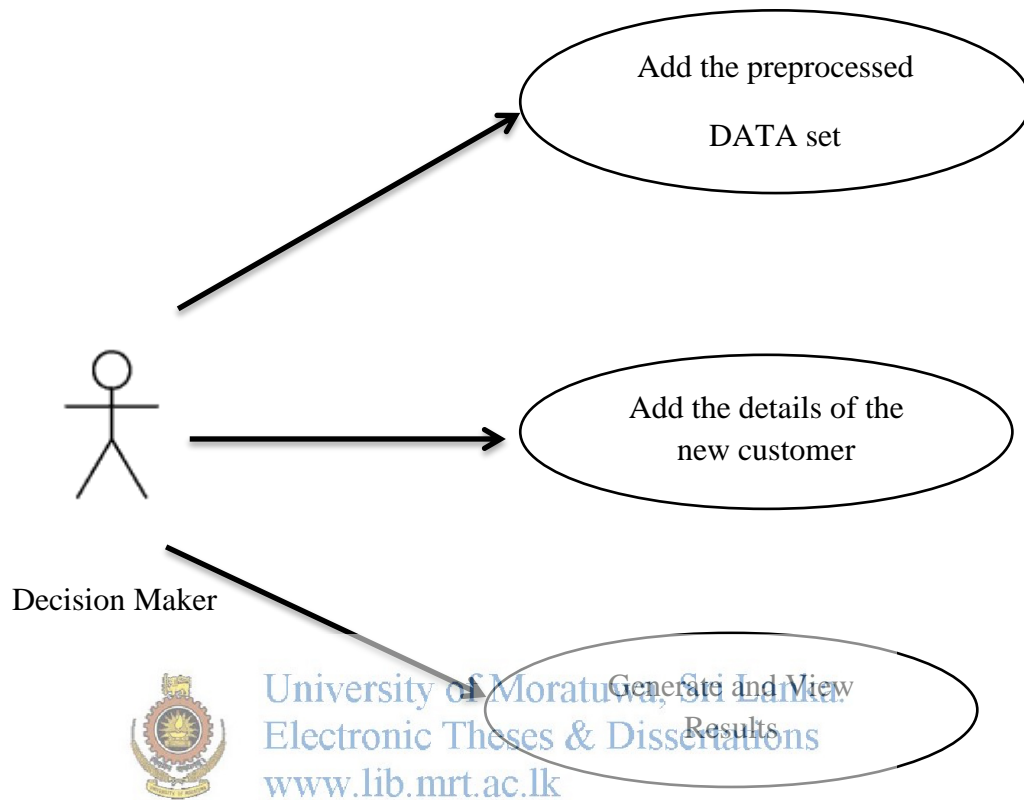
### 5.5.3 Model building by using suitable classification technology

Then various classification techniques are applied in data mining to the 2/3 portion of the whole data set which is selected as Training Dataset and create a model by using that data. these steps are repeated until get the good result.

### 5.5.4 Requirement gathering for the application

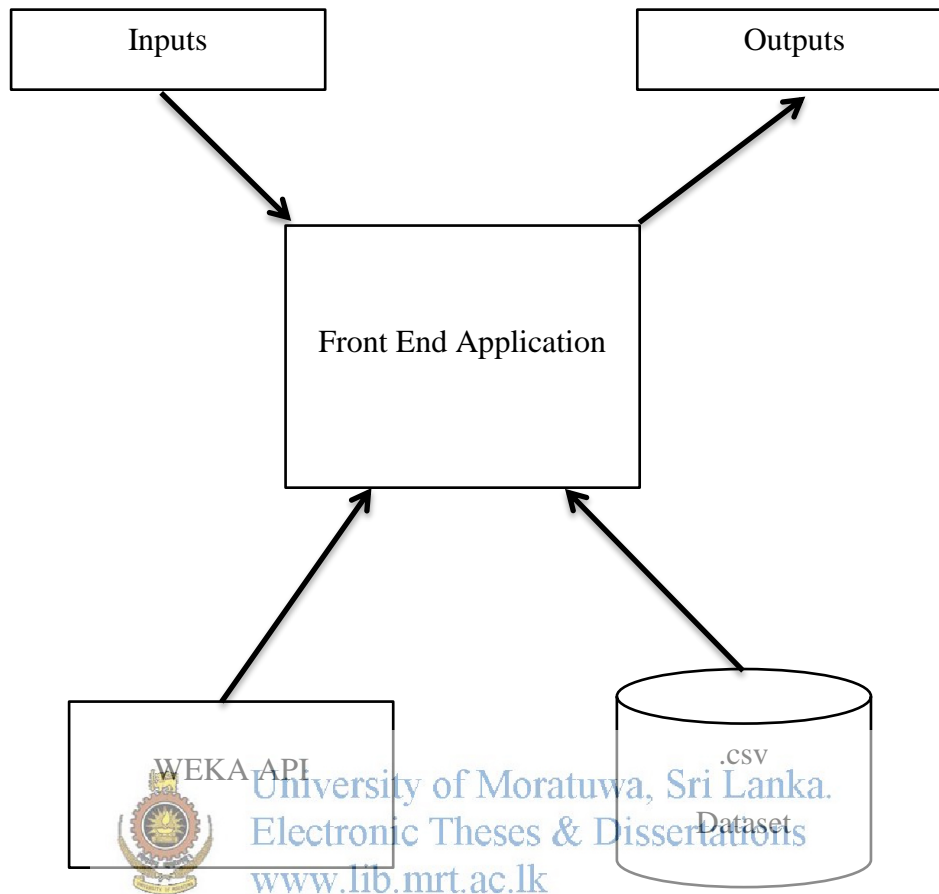
The requirements for the front end application are collected, namely what are the inputs for the tool, what are the outputs from the tool. Who is going to access this tool etc.

### 5.5.5 Design the application



**Figure 5.5: Use case**

### 5.5.6 Implement a front-end application by using JAVA



**Figure 5.6: Frontend application with other connected components**

### 5.5.7 Integration

For the risk assessment part, Integration part is done by using WEKA API and for the plan prediction, WEKA API has not been involved. It is done with normal calculations that are used in Bayesian classification. Same input data has been used for the both Risk Profile Assessment Process and Plan prediction Process.

### 5.5.8 Testing

Data model that has been selected as a final model is evaluated by using test data set which is a 1/3 portion of the whole dataset that has been collected from the leasing company. Testing of the tool is done by using various inputs.

## 5.6 User Interfaces

User interface offers facilities to interact with the system for the users. It retrieves the tendency to recoverability of the loan. It can also give the suitable leasing plan to the customer.

Profes	Brand	Model	Manuf	Vehical	Fuel Ty	Actual	Lease	Req A	Month	Interest	No of	Install	Status
Busine	Toyota	TOWN	2001	Dual p	Petrol	R	100	r	8	12	48	4	Ceased
Farmer	Bajaj	BAJAJ	2010	Motor T	Petrol	P	80	p	1	15	24	1	Ceased
Busine	Bajaj	AUTO	2012	Motor T	Petrol	P	90	p	1	15	12	1	Ceased
Busine	Isuzu	ELF 350	2001	Heavy	Diesel	R	100	r	8	12	60	8	Ceased
Busine	Suzuki	LA-HAZ	2005	Motor	Petrol	R	60	r	8	12	36	4	Ceased
Busine	Nissan	ATLAS	2011	Motor L	Diesel	R	80	r	8	12	12	4	Ceased
Busine	Toyota	TOWN	2007	Dual p	Petrol	R	90	r	8	12	24	4	Ceased
Busine	Nissan	VAHAT	2006	Dual p	Diesel	U	90	r	8	9	48	4	Ceased
Small	Bajaj	AUTO 4S	2007	Motor T	Petrol	P	60	p	1	15	24	1	Ceased
Busine	Bajaj	BAJAJ	2002	Motor T	Petrol	R	100	p	1	15	36	1	Ceased
Busine	Toyota	TOWN	2001	Dual p	Petrol	R	60	r	8	12	24	4	Ceased
Busine	Bajaj	BAJAJ	2001	Motor T	Petrol	P	70	p	1	15	24	1	Ceased
Busine	Tata M	ACE	2008	Light M	Diesel	G	60	p	8	15	12	4	Ceased
Self E	Bajaj	AUTO	2004	Motorcy	Petrol	P	90	p	1	15	24	1	Ceased
Farmer	Bajaj	AUTO	2010	Motor T	Petrol	P	70	p	1	15	24	1	Ceased
Busine	Isuzu	ELF 350	2003	Heavy	Diesel	R	80	r	8	12	48	4	Ceased
Busine	Maruti	800	2009	Motor	Petrol	R	80	r	8	12	24	4	Ceased
Busine	Mazda	TITAN	1999	Dual p	Diesel	T	70	r	8	12	36	4	Ceased
Busine	Toyota	DYNA	2002	Motor L	Diesel	T	70	r	8	12	60	4	Ceased
Busine	Nissan	VAHAT	2003	Dual p	Diesel	U	70	r	8	12	36	4	Ceased
Busine	Bajaj	AUTO 4S	2004	Motor T	Petrol	P	60	p	1	15	36	1	Ceased
Small	Bajaj	AUTO	2006	Motorcy	Petrol	P	100	p	1	15	12	1	Ceased

Figure 5.7: Data Input Interface

Figure 5.8: Risk Analysis Interface

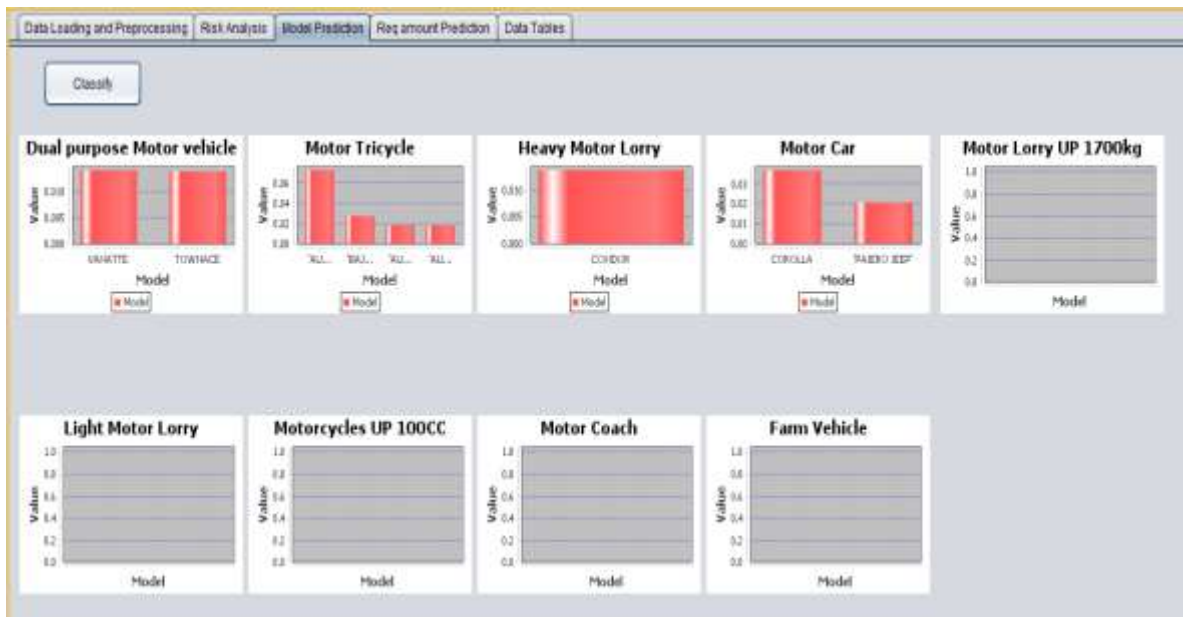


Figure 5.9: Model Prediction

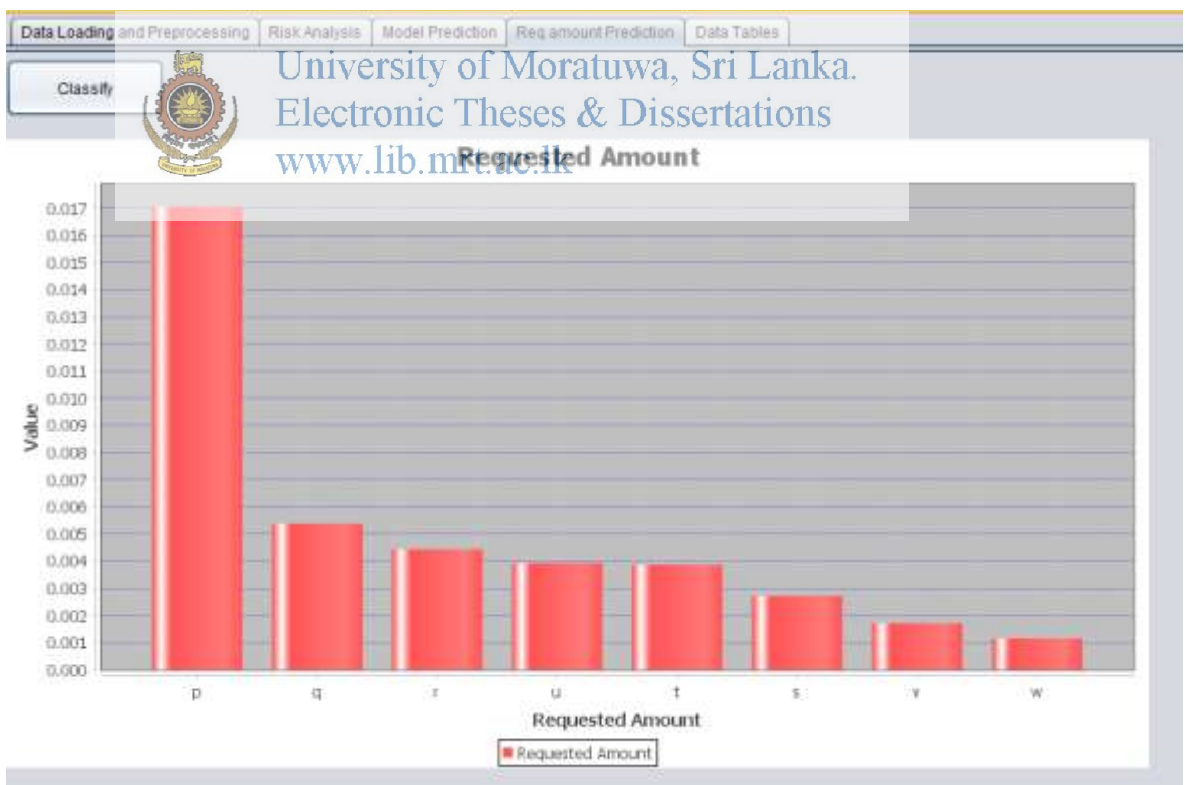


Figure 5.10: Requested amount Prediction

## 5.7 Summary

When a customer comes to lease a vehicle, decision maker get the details from the customer and enter to the system. Then system will predict, whether there is a tendency to cease the vehicle or close the leasing agreement after repaying the loan.

As well as system will predict the most suitable vehicle model for the customer according to the previous behavior of the company customers and predict the most recoverable loan amount for the customer, based on Profession and Monthly income of the borrower.

Prediction is done, based on the company's previous transaction data therefor relevant data have been gathered and create the model using 2/3 data and test that model using 1/3 data.

All the sections has been interconnected using front end application



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

# Chapter 6

## Implementation of the Tool

### 6.1 Introduction

In chapter 5, the design of the solution has been described in terms of what each component does. This chapter described implementation of each component regarding Software/Algorithms etc. In that sense this chapter is about how the system is implemented.

### 6.2 Implementation of CRMT (Credit Risk Management Tool)

CRMT has been developed with Java, WEKA API with Data mining technologies.

After identifying the problem domain, data gathering should be done in the relevant party; therefore according to the problem definition of this research there was a requirement to find the data set which relevant to the credit risk management and decision support. as a result, large number of historical data have been collected from the medium level leasing company which contains both ceased and closed transaction details. Then 2/3 is selected as a training dataset and 1/3 is selected as a testing dataset

#### 6.2.1 Pre preparation

##### Categorization

The data are categorized in to groups (Appendix A).

##### Variable Selection for the model

The chi-square test for independence, also called Pearson's chi-square test or the chi-square test of association, is used to discover if there is a relationship between two categorical variables (Appendix B).

**Table 6.1: Variable Selection Summary table**

Variable	P- Value	Selected / Not Selected
Gender	.101	Not Selected
Age	.693	Not Selected
Profession	.000	Selected
District	.766	Not Selected
Brand Name	.000	Selected
Model Name	.000	Selected
Manuf. Year	.002	Selected
Vehicle Class	.000	Selected
Fuel Type	.000	Selected
Actual Value	.000	Selected
Lease Percentage	.000	Selected
Req. Amount	.000	Selected
Monthly Income	.048	Selected
Interest	.000	Selected
No.of Rentals	.000	Selected
Installment	.000	Selected



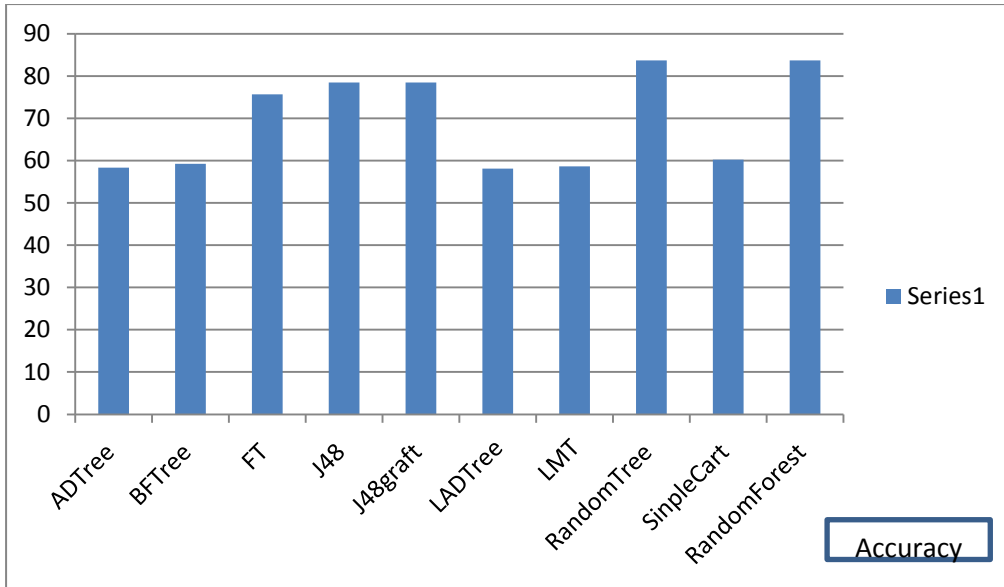
According to the results, Age, Gender, District cannot be selected to the final model as variables.

### 6.2.2 Best Model Selection

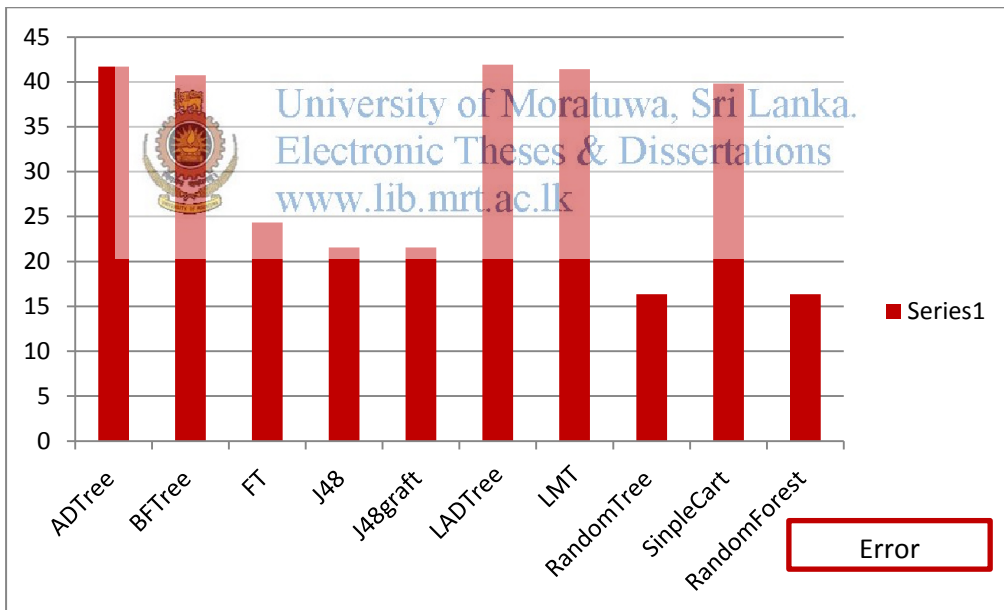
After the testing for the data set to identify which independent variables are highly affect to the dependent variable, then the data set with selected variables is loaded to the WEKA to select most suitable algorithm for the model. In this process, various algorithms are applied to the data set (Appendix c).

**Table 6.1: Algorithm Accuracy Summary**

Algorithm	Accuracy
ADTree	58.3
BFTree	59.25
FT	75.65
J48	78.43
J48graft	78.43
LADTree	58.1
LMT	58.6
RandomTree	83.65
SimpleCart	60.2
RandomForest	83.65



**Figure 6.1: Accuracy levels of algorithms**



**Figure 6.2: Error rates of algorithms**

For the particular data set, there are two highest accuracies are shown. However the RandomTree has the lowest execution time. Therefore this algorithm has been used for the evaluation of the particular input values.

In the situation of the loan officer wants to analyze the data in order to know which customers are risky or which are safe, classification can be used.

### 6.2.3 Risk Assessment Process

#### Building the Classifier (classification model)

In Classification algorithm, the classifier and is built the classifier is built from the training set made up of database tuples and their associated class labels.

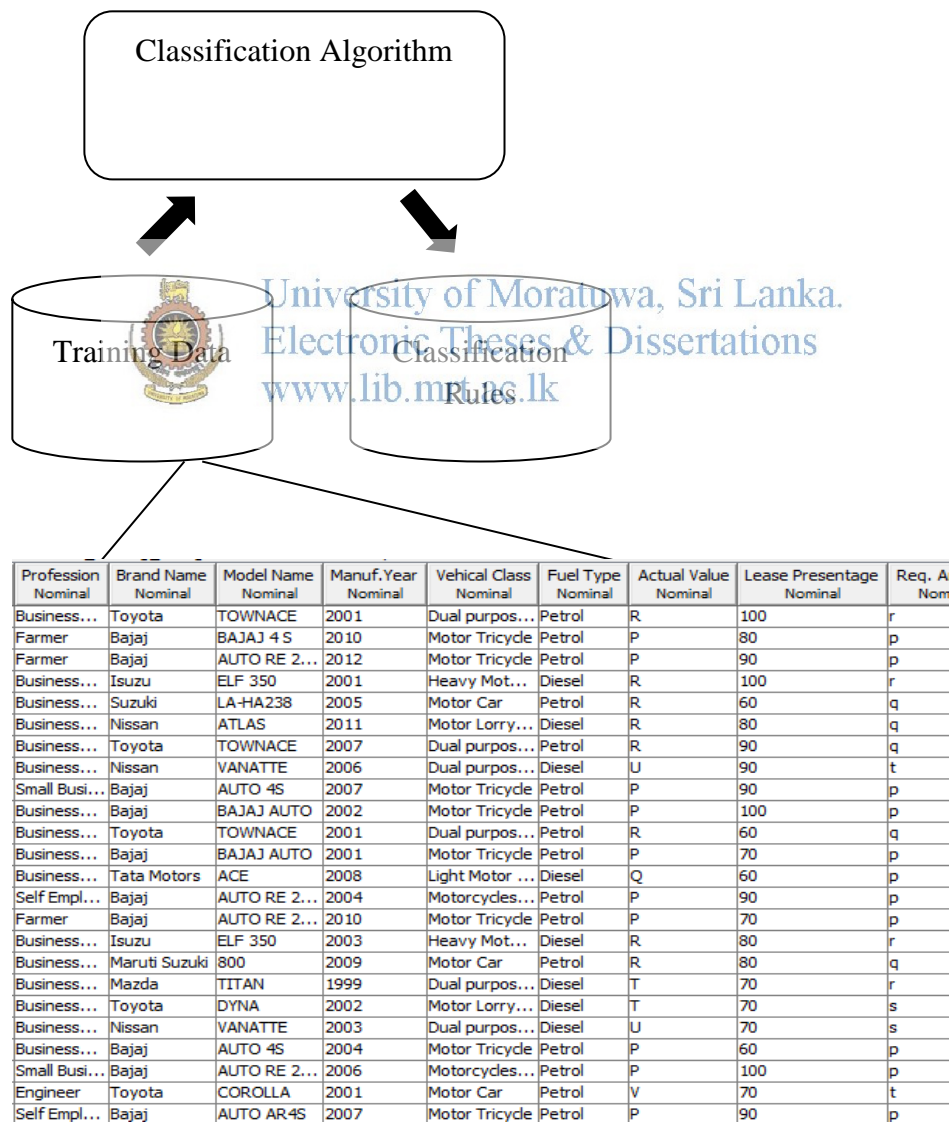


Figure 6.3: Building the Classifier

## Using Classifier for Classification

The classifier is used for classification. Customer data which is collected from the user interface are saved as two .csv files (one with status: Ceased and another one with status: Close) before applying categorization to the input data to match with the created model format. Then WEKA functions are used and apply numeric is applied to nominal and saved it as an .arff (Attribute-Relation **File** Format) format. After that, pass that files to the created model (classifier) as test Instances. Then 100% accuracy is given with the model by matching instances. Therefore decisions can be made as tendency to ceased or closed.

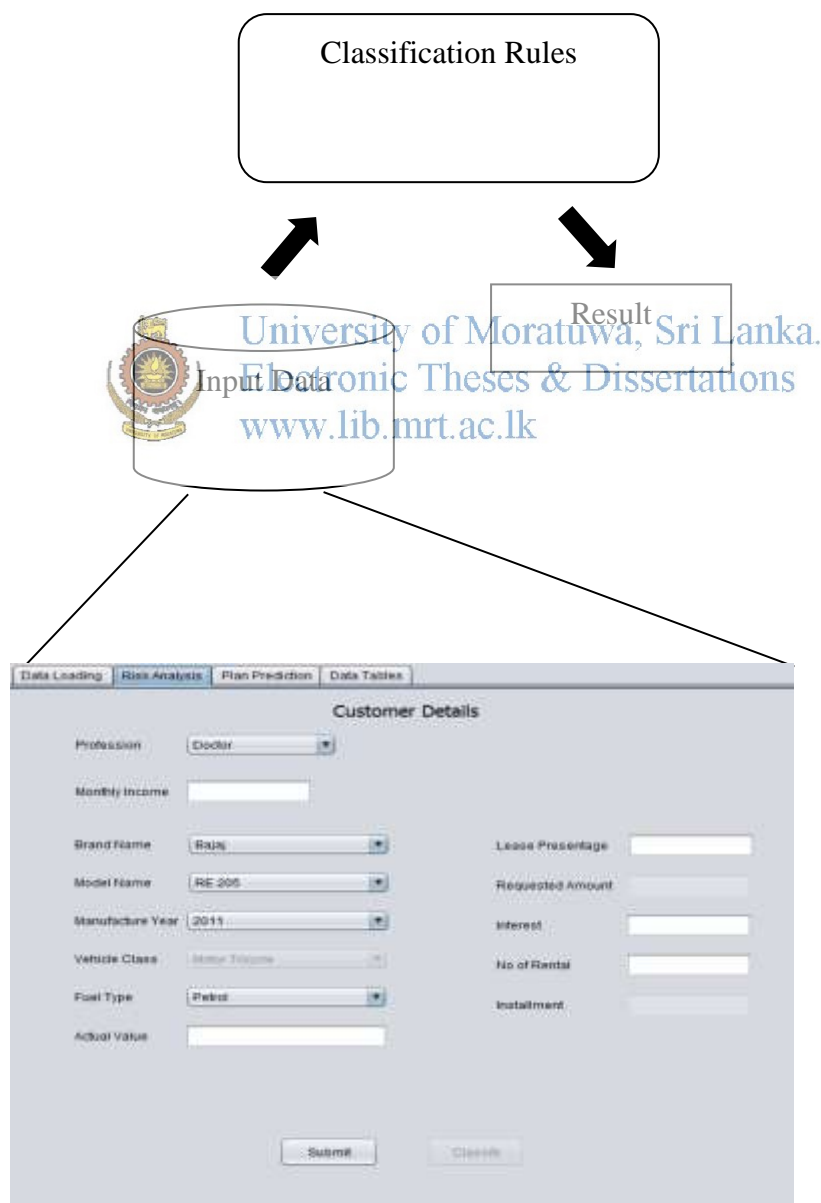


Figure 6.4: Using Classifier for Classification

## 6.2.4. Plan Prediction Process

### Bayesian classification

Bayesian classification has been used in this tool to predict suitable vehicle model and requested amount for the particular customer based on their profession and monthly income.

#### Equation:

$$p(C_k|\mathbf{x}) = \frac{p(C_k) p(\mathbf{x}|C_k)}{p(\mathbf{x})}$$

#### 01. Tendency to Close

$$P(\text{Status} = \text{Closed}) = P((\text{Profession} = \text{Doctor}) / (\text{Status} = \text{Closed}))$$

$$* P((\text{Monthly Income} = \text{IV}) / (\text{Status} = \text{Closed}))$$

$$* P((\text{Vehicle model} = \text{COROLLA}) / (\text{Status} = \text{Closed}))$$

$$* P((\text{Closed}) / (\text{Ceased} + \text{Closed}))$$

#### 02. Tendency to Cease

$$P(\text{Status} = \text{Ceased}) = P((\text{Profession} = \text{Doctor}) / (\text{Status} = \text{Ceased}))$$

$$* P((\text{Monthly Income} = \text{IV}) / (\text{Status} = \text{Ceased}))$$

$$* P((\text{Vehicle model} = \text{COROLLA}) / (\text{Status} = \text{Ceased}))$$

$$* P((\text{Ceased}) / (\text{Ceased} + \text{Closed}))$$

The largest value has been selected and it is the tendency to close/close. the same calculation is done by changing vehicle model and the closed vehicle models are selected and displayed in the graph.

In order to find the requested amount requested amount variable is applied instead of the vehicle model. Then the same procedure is followed.

### **6.2.5 Requirement gathering for the application**

Determine the requirements like; who are the user of the system? How they are going to use the system? What are the inputs into the system? What are the outputs from the system? These are common questions that get answered during a requirements gathering. After requirement gathering these requirements should be analyzed for their validity and the possibility of incorporating the requirements in the system to be development is also studied.

### **6.2.6 Application Design**

Front end application has been designed using java language and NetBeans software. Graphs have been generated with the help of JFreeChart open source library.

### **6.2.7 Implement a front-end application by using JAVA**

Front end application has been implemented by using NetBeans 8.01. All the interfaces and coding have been done by using this software.

### **6.2.8 Integration**

After identifying the particular classification algorithm for the data set, that algorithm is used for evaluate the input data which input from the interface of the CRMT. As well as, the suitable plan for the customer is predicted using Bayesian Classifier. Evaluation with the model is done the help of WEKA API.

### **6.2.9 Testing**

After identifying the suitable model, it has been evaluated by the test dataset. Then the model in integrated with the front end application and the whole system is tested for the compatibility.

### **6.2.10 Building an Evaluation Model**

The evaluation model is built by using WEKA open source software with the help of data mining algorithms.

## **6.3 Summary**

prepreparation has been done of the data set including categorization and variable selection, then research for the most suitable model for the data set and RandomTree algorithm has been selected. Risk Assessment process and Plan Prediction process has been done as the next step. Finally test the tool.

# Chapter 7

## Evaluation

### 7.1 Introduction

This chapter describes how the CRMT is tested in terms of users, and how the selected model tested by using testing data set.

### 7.2 Data Model Testing

The model has been created by using 2/3 of the whole dataset and rest of the 1/3 has been used for testing the model. This gives 59.29% accuracy for the test data set with 0.06 execution time.

All other algorithms have been given the less accuracy with the test data set than RandomTree except the RandomForest (around 60%) but it takes more time to execute than RandomTree.

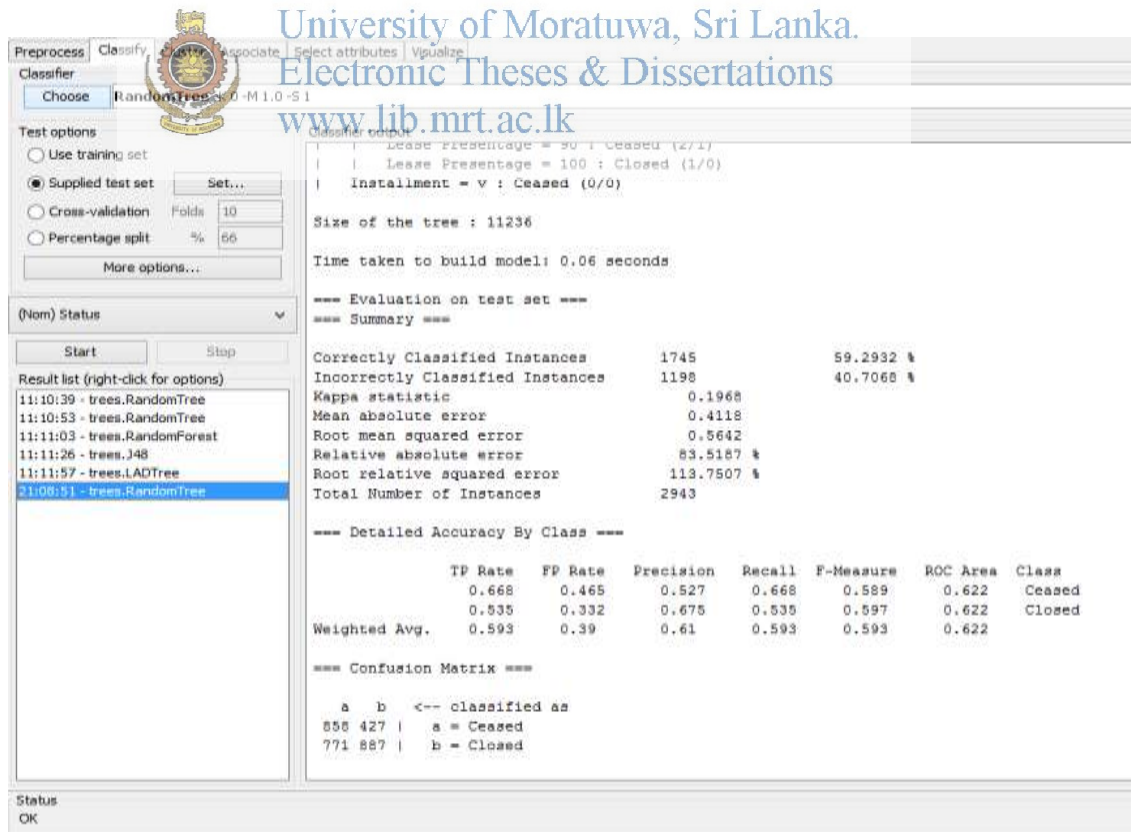


Figure 7.1: Test the selected model with testing data set

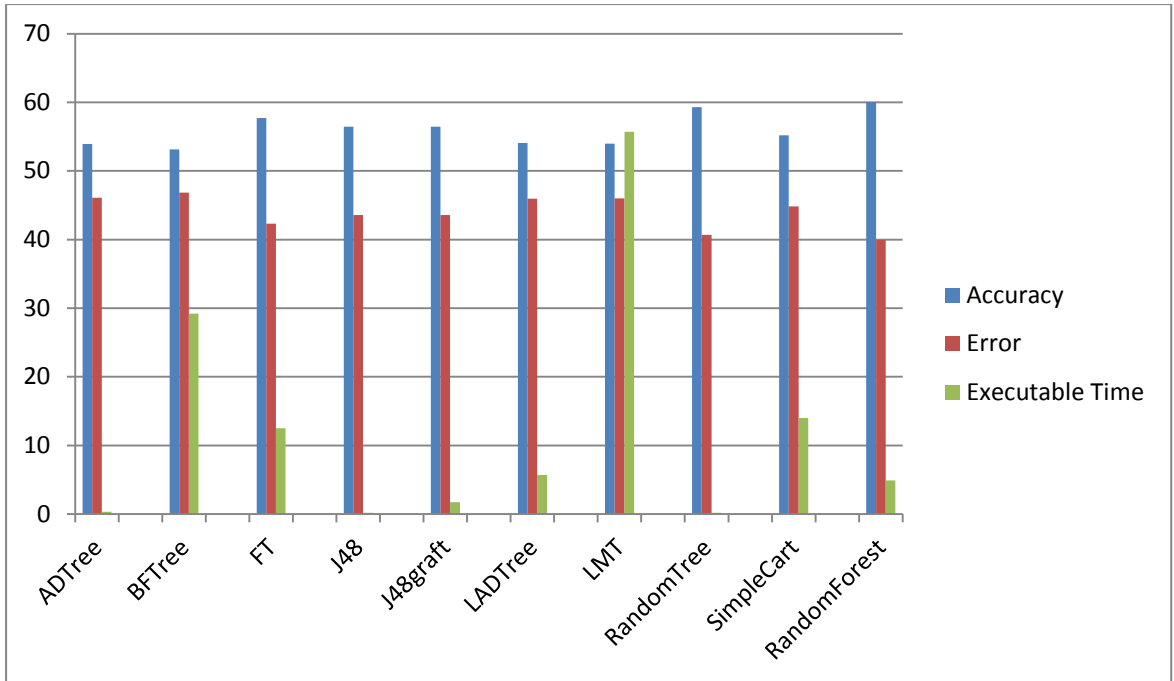


Figure 7.2: Model evaluation with the test data set

7.3 Test the System for Randomly selected values

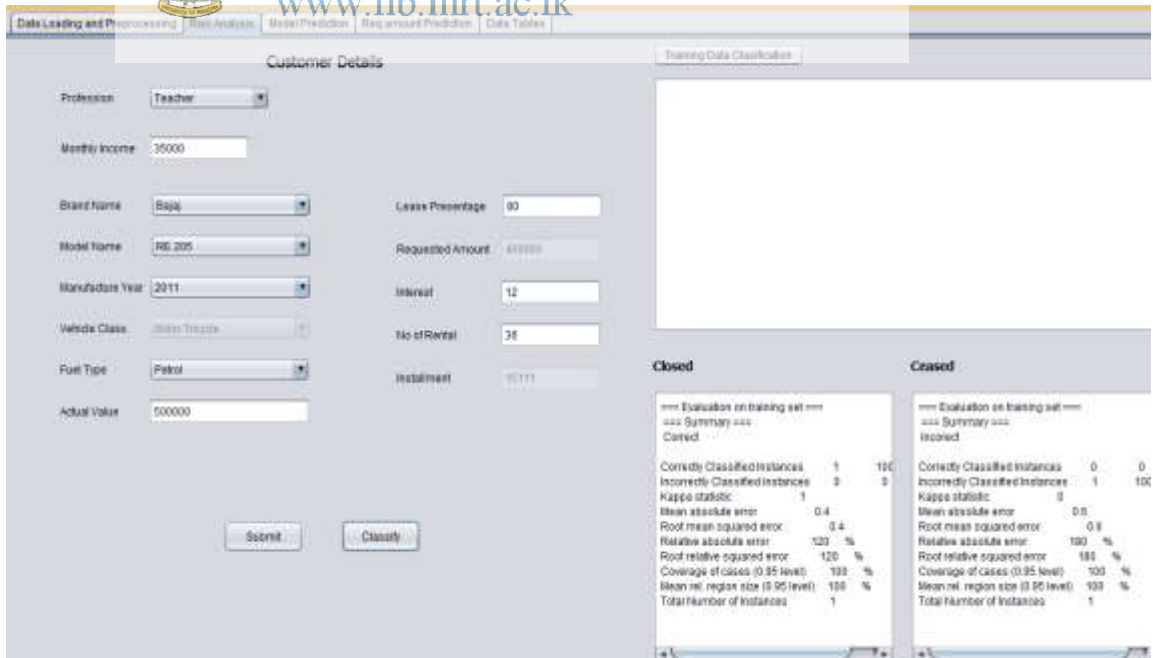


Figure 7.3: Customer Details





Figure 7.4: Most Suitable vehicle model for particular customer

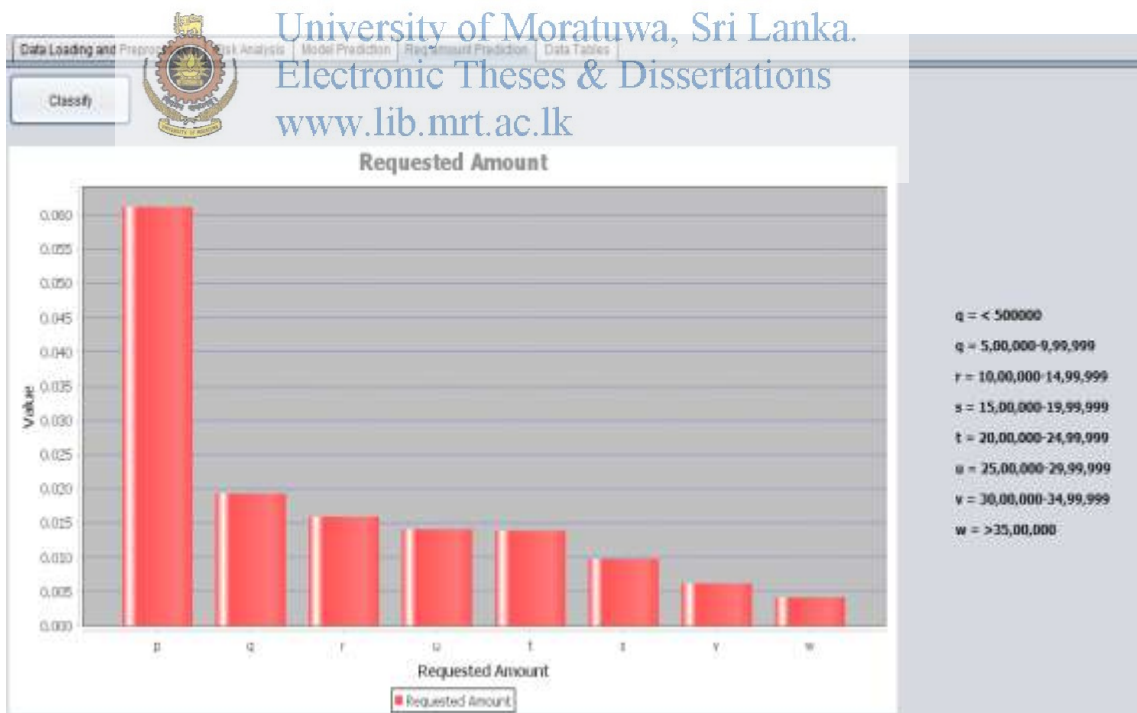


Figure 7.5: Most suitable loan ranges for particular user

#### **7.4 Summary**

This testing is ensured the objectives being tested in relation to input output, Process and features as mentioned in the approach chapter. As well as, accuracy of the selected model has been tested.



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

# Chapter 8

## Conclusion and further work

### 8.1 Introduction

Credit risk assessment is very important research field, and it is hard to find that decision support system for manage credit risk in leasing domain in sri lanka. And even in worldwide there are no such systems has been developed for the leasing domain.

### 8.2 Conclusion of CRMT (Credit Risk Management Tool)

It is concluded that, proposed system has predicted the tendency to cease or close the leasing agreement of the customers with the help of RandomTree classification algorithm and predict the suitable vehicle model and most compatible Requested Amount to the customer based on previous records of the company with the help of Bayesian classification, according to the Profession of the customer and Monthly income of the customer



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
www.lib.mrt.ac.lk

### 8.3 Limitations of CRMT (Credit Risk Management Tool)

There are some identified limitations of the tool such as when we input customer details; it should be in the area of previous customer's data in order to evaluate with the previous data. And accuracy of the tool is highly depending on completeness of the dataset.

### 8.4 Further work

As a further work, we can suggest that Pre perpetration part such as data cleaning, filling missing values, identification of the suitable categorization, variable selection and identification of most suitable classification techniques according to the data set can be integrated to the tool.

### 8.5 Summary

This section concluded about the tool's functions, limitations of the tool and further work.

## References

- [1] E. D. Madyatmadja and M. Aryuni, "Comparative Study of Data Mining Model for Credit Card Application Scoring in Bank," *J. Theor. Appl. Inf. Technol.*, vol. 59, no. 2, 2014.
- [2] M. Aryuni and E. D. Madyatmadja, "Feature Selection in Credit Scoring Model for Credit Card Applicants in XYZ Bank: A Comparative Study," *Int. J. Multimed. Ubiquitous Eng.*, vol. 10, no. 5, pp. 17–24, May 2015.
- [3] M. Nithya, S. Suba, and B. Vaishnavi, "Predict the usage of laptops among students in rural areas using weka tool," *Int. J. Adv. Technol. Eng. Sci.*, vol. Vol. No.3, no. 01 September 2015, p. 6.
- [4] N. Padhy, "Data Mining: A prediction Technique for the workers in the PR Department of Orissa (Block and Panchayat)," *Int. J. Comput. Sci. Eng. Inf. Technol.*, vol. 2, no. 5, pp. 197–36, Oct. 2012.
- [5] D. Kumar Thiwary, "A Comparative Study Of Classification Algorithms For Credit Card Approval Using WEKA," *GALAXY Int. Interdiscip. Res. J.*, vol. Vol.2, Mar. 2014.
- [6] Plc.lk, 'People's Leasing Company PLC - Annual Report 2013 - 2014', 2015. [Online]. Available: [http://www.plc.lk/inpages/annual\\_report\\_2013\\_14/risk-management.html](http://www.plc.lk/inpages/annual_report_2013_14/risk-management.html). [Accessed: 03- Jan- 2015].
- [7] Annual Report 2013/14, 1st ed. colombo: Swarnamahala Financial Services PLC, 2015, pp. 23,24.
- [8] Annual Report 2012/13, 1st ed. colombo: Swarnamahala Financial Services PLC, 2014, p. 25

- [9]S. Faizan Ahmed and Q. Ali Malik, 574 *International Journal of Economics and Financial Issues* | Vol 5 • Issue 2 • 2015 *Credit Risk Management and Loan Performance: Empirical Investigation of Micro Finance Banks of Pakistan*, 1st ed. International Journal of Economics and Financial Issues, 2015.
- [10]D. Foust and A. Pressman, 'Not-So-Magic Numbers', *Business Week*, 2008.
- [11]*Foundations of Banking Risk: An Overview of Banking, Banking Risks, and Risk-Based Banking Regulation.*, 1st ed. New Jersey: John Wiley& Sons, 2009.
- [12]T. Gestel and B. Baesens, *Credit Risk Management: Basic Concepts: Financial Risk Components, rating analysis,models,economic and regulary capital*, 1st ed. Oxford University Press, 2008.
- [13]G. Yingjian and W. Chong, *Research on Credit Risk Assessment in Commercial Bank Based on Information Integration*, 1st ed. .



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

## Appendix A: Categorization

Data sets have been categorized on to following order

•Dependent Variable: Status - Ceased, Closed

•Other Variables:

Factor 1 - Profession - Businessmen, Doctor, Engineer, Executive, Farmer, Lawyer, Lecturer, Manager, Other, Self Employee, Small Businessmen, Teacher

Factor 2 - Brand Name - Bajaj, Chevrolet, Isuzu, Mahindra & Mahindra, Maruti Suzuki, Mazda, Mitsubishi Motors, Nissan, Suzuki, Tafe, Tata Motors, Toyota

Factor 3 - Model Name - 800, 45D, ACE, ALTO, AR4S-UG, ATLAS, AUTO 4S, AUTO AR4S, AUTO RE 2 STROKE, BAJAJ 4 S, BAJAJ AUTO, BOLERO MAXI TRUCK, CONDOR, COROLLA, DBA-NZE141, DYNA, ELF 350, HIACE, KF-GM70-HALF-BODY, KG-LH172, KG-VWE25, LA-HA238, LP7155, PAJERO JEEP, RE 205, TITAN, TOWNACE, UA-HR528, VANATTE

Factor 4 - Manufacture Year - 1993 to 2012

Factor 5 - Vehicle Class - Dual purpose Motor vehicle, Farm vehicle, Heavy Motor Lorry, Light Motor Lorry, Motor Car, Motor Coach, Motor Lorry UP 1700kg, Motor Tricycle, Motorcycles UP 100CC

Factor 6 - Fuel Type - Diesel, Petrol

Factor 7 - Actual Value - <5,00,000(P), 5,00,000-9,99,999(Q), 10,00,000-14,99,999(R), 15,00,000-19,99,999(S), 20,00,000-24,99,999(T), 25,00,000-29,99,999(U), 30,00,000-34,99,999(V), >35,00,000(W)

Factor 8 - Lease Percentage - 60% to 100%

Factor 9 - Req. Amount - <5,00,000(p), 5,00,000-9,99,999(q), 10,00,000-14,99,999(r), 15,00,000-19,99,999(s), 20,00,000-24,99,999(t), 25,00,000-29,99,999(u), 30,00,000-34,99,999(v), >35,00,000(w)

Factor 10 - Monthly Income - <50,000(I), 50,000-99,999(II), 1,00,000-1,49,999(III),  
1,50,000-1,99,999(IV), >2,00,000(V)

Factor 11 - Interest - 9 to 15

Factor 12 - Number of rentals - 12 to 60

Factor 13 - Installment - <20,000(i), 20,000-39,999(ii), 40,000-59,999(iii), 60,000-  
79,999(iv), 80,000-99,999(v), >1,00,000(vi)



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

## Appendix B: Variable Selection Procedure (Technology: Chi – Squared)

### Status \* Gender

#### Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2.691 <sup>a</sup>	1	.101		
Continuity Correction <sup>b</sup>	2.572	1	.109		
Likelihood Ratio	2.700	1	.100		
Fisher's Exact Test				.107	.054
N of Valid Cases	6000				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 393.38.

b. Computed only for a 2x2 table

### Status \* Age

#### Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	3.878 <sup>a</sup>	6	.693
Likelihood Ratio	3.877	6	.693
N of Valid Cases	6000		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 16.91.



## Status \* Profession

### Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	679.354 <sup>a</sup>	11	.000
Likelihood Ratio	955.310	11	.000
N of Valid Cases	6000		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 29.82.

## Status \* District

### Chi-Square Tests



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4.115 <sup>a</sup>	7	.766
Likelihood Ratio	4.109	7	.767
N of Valid Cases	6000		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 44.06.

## Status \* BrandName

### Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	157.472 <sup>a</sup>	11	.000
Likelihood Ratio	167.321	11	.000
N of Valid Cases	6000		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 13.35.

## Status \* ModelName

### Chi-Square Tests



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	392.574 <sup>a</sup>	28	.000
Likelihood Ratio	457.831	28	.000
N of Valid Cases	6000		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 12.02.

## Status \* Manuf.Year

### Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	41.545 <sup>a</sup>	19	.002
Likelihood Ratio	41.803	19	.002
N of Valid Cases	6000		

a. 1 cells (2.5%) have expected count less than 5. The minimum expected count is 4.90.

## Status \* Vehicle Class

### Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	252.964 <sup>a</sup>	8	.000
Likelihood Ratio	269.882	8	.000
N of Valid Cases	6000		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 13.35.

## Status \* Fuel Type

### Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	37.032 <sup>a</sup>	1	.000		
Continuity Correction <sup>b</sup>	36.705	1	.000		
Likelihood Ratio	36.967	1	.000		
Fisher's Exact Test				.000	.000
N of Valid Cases	6000				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 987.90.

b. Computed only for a 2x2 table



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

## Status \* ActualValue

### Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	311.275 <sup>a</sup>	7	.000
Likelihood Ratio	367.826	7	.000
N of Valid Cases	6000		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 99.68.

## Status \* Lease Percentage

### Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	27.255 <sup>a</sup>	4	.000
Likelihood Ratio	27.262	4	.000
N of Valid Cases	6000		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 494.40.

## Status \* Req.Amount

### Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	197.350 <sup>a</sup>	7	.000
Likelihood Ratio	208.573	7	.000
N of Valid Cases	6000		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 86.33.



## Status \* Monthly Income

### Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.921 <sup>a</sup>	1	.048		
Continuity Correction <sup>b</sup>	3.571	1	.059		
Likelihood Ratio	3.892	1	.049		
Fisher's Exact Test				.057	.030
N of Valid Cases	6000				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 56.07.

b. Computed only for a 2x2 table



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

## Status \* Interest

### Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	134.384 <sup>a</sup>	2	.000
Likelihood Ratio	138.243	2	.000
N of Valid Cases	6000		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 561.59.

## Status \* No.of Rentals

### Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	88.824 <sup>a</sup>	4	.000
Likelihood Ratio	90.264	4	.000
N of Valid Cases	6000		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 425.42.

## Status \* Installment

### Chi-Square Tests



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	85.487 <sup>a</sup>	4	.000
Likelihood Ratio	87.334	4	.000
N of Valid Cases	6000		

# Appendix C: Best Model Selection

Classifier

Choose **LMT -I -1 -M 15 -W 0.0**

Test options

Use training set

Supplied test set

Cross-validation Folds

Percentage split %

(Nom) Status

Result list (right-click for options)

- 20:57:11 - trees.J48
- 20:58:49 - trees.ADTree**
- 20:58:55 - trees.BFTree
- 20:59:39 - trees.FT
- 21:00:45 - trees.J48graft
- 21:02:45 - trees.LADTree
- 21:03:24 - trees.SimpleCart
- 21:05:46 - trees.RandomForest
- 21:06:17 - trees.RandomTree
- 21:12:43 - trees.LMT

Classifier output

Legend: -ve = Cease, +ve = Close

Tree size (total number of nodes): 25

Leaves (number of predictor nodes): 17

Time taken to build model: 0.22 seconds

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances	3502	58.3667 %
Incorrectly Classified Instances	2498	41.6333 %
Kappa statistic	0.1548	
Mean absolute error	0.4395	
Root mean squared error	0.4678	
Relative absolute error	88.9747 %	
Root relative squared error	94.1275 %	
Total Number of Instances	6000	

=== Detailed Accuracy By Class ===


	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.518	0.364	0.533	0.518	0.526	0.633	Cease
	0.636	0.482	0.622	0.636	0.629	0.633	close
Weighted Avg.	0.584	0.429	0.583	0.584	0.583	0.633	

=== Confusion Matrix ===

a b <-- classified as

1212 2118 | a = Cease

1212 2118 | b = close



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
www.lib.mrt.ac.lk

Status  
OK

ADTree

Correctly Classified Instances	Time Taken to Build the Model
58.3%	0.22 Seconds



Classifier

Choose LMT -I -1 -M 15 -W 0.0

Test options

Use training set

Supplied test set Set...

Cross-validation Folds 10

Percentage split % 66

More options...

(Nom) Status

Start Stop

Result list (right-click for options)

- 20:57:11 - trees.J48
- 20:58:49 - trees.ADTree
- 20:58:55 - trees.BFTree
- 20:59:39 - trees.FT
- 21:00:45 - trees.J48graft
- 21:02:45 - trees.LADTree
- 21:03:24 - trees.SimpleCart
- 21:05:46 - trees.RandomForest
- 21:06:17 - trees.RandomTree
- 21:12:43 - trees.LMT

Classifier output

Size of the tree: 7

Number of Leaf Nodes: 4

Time taken to build model: 31.04 seconds

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances	3555	59.25 %
Incorrectly Classified Instances	2445	40.75 %
Kappa statistic	0.2096	
Mean absolute error	0.4362	
Root mean squared error	0.467	
Relative absolute error	88.3177 %	
Root relative squared error	93.9777 %	
Total Number of Instances	6000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.763	0.544	0.529	0.763	0.625	0.637	Cease
	0.456	0.237	0.706	0.456	0.554	0.637	close
Weighted Avg.	0.593	0.374	0.627	0.593	0.586	0.637	

=== Confusion Matrix ===

a b <-- classified as

2037	633	a = Cease
1812	1518	b = close

Status OK



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

BFTree

Correctly Classified Instances	Time Taken to Build the Model
59.25%	31.04 Seconds

Classifier

Choose LMT -I-1 -M 15 -W 0.0

Test options

Use training set

Supplied test set Set...

Cross-validation Folds 10

Percentage split % 66

More options...

(Nom) Status

Start Stop

Result list (right-click for options)

- 20:57:11 - trees.J48
- 20:58:49 - trees.ADTTree
- 20:58:55 - trees.BFTree
- 20:59:39 - trees.FT
- 21:00:45 - trees.J48graft
- 21:02:45 - trees.LADTree
- 21:03:24 - trees.SimpleCart
- 21:05:46 - trees.RandomForest
- 21:06:17 - trees.RandomTree
- 21:12:43 - trees.LMT

Classifier output

```
[Installation=i1] * -0.34 +
[Installation=v] * 0.5
```

Time taken to build model: 12.61 seconds

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances	4539	75.65 %
Incorrectly Classified Instances	1461	24.35 %
Kappa statistic	0.5082	
Mean absolute error	0.3044	
Root mean squared error	0.419	
Relative absolute error	61.6239 %	
Root relative squared error	84.3042 %	
Total Number of Instances	6000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.738	0.229	0.721	0.738	0.73	0.821	Cease
	0.771	0.262	0.786	0.771	0.779	0.821	close
Weighted Avg.	0.757	0.247	0.757	0.757	0.757	0.821	

=== Confusion Matrix ===

```

a   b  <-- classified as
1971 699 | a = Cease
762 2568 | b = close

```

Status OK



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

FT Tree

Correctly Classified Instances	Time Taken to Build the Model
75.65 %	12.61 Seconds

Classifier

Choose **LMT -I -1 -M 15 -W 0.0**

Test options

Use training set

Supplied test set

Cross-validation Folds

Percentage split %

(Nom) Status

Result list (right-click for options)

- 20:57:11 - trees.J48
- 20:58:49 - trees.ADTree
- 20:58:55 - trees.BFTree
- 20:59:39 - trees.FT
- 21:00:45 - trees.J48graft
- 21:02:45 - trees.LADTree
- 21:03:24 - trees.SimpleCart
- 21:05:46 - trees.RandomForest
- 21:06:17 - trees.RandomTree
- 21:12:43 - trees.LMT

Classifier output

Size of the tree : 4826

Time taken to build model: 86.52 seconds

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances	4706	78.4333 %
Incorrectly Classified Instances	1294	21.5667 %
Kappa statistic	0.5682	
Mean absolute error	0.2798	
Root mean squared error	0.374	
Relative absolute error	56.6426 %	
Root relative squared error	75.2614 %	
Total Number of Instances	6000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.813	0.239	0.732	0.813	0.77	0.879	Cease
	0.761	0.187	0.836	0.761	0.797	0.879	close
Weighted Avg.	0.784	0.21	0.789	0.784	0.785	0.879	

=== Confusion Matrix ===

a b <-- classified as

2172	498	a = Cease
796	2534	b = close

Status OK



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

J48

Correctly Classified Instances	Time Taken to Build the Model
78.4%	86.52 Seconds

Classifier

Choose **LMT -I -1 -M 15 -W 0,0**

Test options

Use training set  
 Supplied test set Set...  
 Cross-validation Folds 10  
 Percentage split % 66

More options...

(Nom) Status

Start Stop

Result list (right-click for options)

- 20:57:11 - trees.J48
- 20:58:49 - trees.ADTTree
- 20:58:55 - trees.BFTree
- 20:59:39 - trees.FT
- 21:00:45 - trees.J48graft**
- 21:02:45 - trees.LADTree
- 21:03:24 - trees.SimpleCart
- 21:05:46 - trees.RandomForest
- 21:06:17 - trees.RandomTree
- 21:12:43 - trees.LMT

Status OK

Classifier output

Size of the tree : 5684

Time taken to build model: 98.1 seconds

=== Evaluation on training set ===  
 === Summary ===

Correctly Classified Instances	4706	78.4333 %
Incorrectly Classified Instances	1294	21.5667 %
Kappa statistic	0.5682	
Mean absolute error	0.2798	
Root mean squared error	0.374	
Relative absolute error	56.6426 %	
Root relative squared error	75.2614 %	
Total Number of Instances	6000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.813	0.239	0.732	0.813	0.77	0.879	Cease
	0.761	0.187	0.836	0.761	0.797	0.879	close
Weighted Avg.	0.784	0.21	0.789	0.784	0.785	0.879	

=== Confusion Matrix ===

```

a    b  <-- classified as
2172 498 | a = Cease
 796 2534 | b = close

```



University of Moratuwa, Sri Lanka.  
 Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

J48graft

Correctly Classified Instances	Time Taken to Build the Model
78.4%	98.1 Seconds

Classifier

Choose **LMT -I -1 -M 15 -W 0.0**

Test options

Use training set

Supplied test set

Cross-validation Folds

Percentage split %

(Nom) Status

Result list (right-click for options)

- 20:57:11 - trees.J48
- 20:58:49 - trees.ADTree
- 20:58:55 - trees.BFTree
- 20:59:39 - trees.FT
- 21:00:45 - trees.J48graft
- 21:02:45 - trees.LADTree**
- 21:03:24 - trees.SimpleCart
- 21:05:46 - trees.RandomForest
- 21:06:17 - trees.RandomTree
- 21:12:43 - trees.LMT

Classifier output

#expanded nodes: 60

#Processed examples: 199524

#Ratio e/n: 2494.05

Time taken to build model: 5.35 seconds

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances	3486	58.1 %
Incorrectly Classified Instances	2514	41.9 %
Kappa statistic	0.1851	
Mean absolute error	0.4405	
Root mean squared error	0.468	
Relative absolute error	89.1866 %	
Root relative squared error	94.1653 %	
Total Number of Instances	6000	

=== Detailed Accuracy By Class ===


	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.736	0.543	0.521	0.736	0.61	0.625	Cease
	0.457	0.264	0.683	0.457	0.548	0.625	close
Weighted Avg.	0.581	0.388	0.611	0.581	0.575	0.625	

=== Confusion Matrix ===

a b <-- classified as

1965	705	a = Cease
1809	1521	b = close

Status  
OK



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

### LAD Tree

Correctly Classified Instances	Time Taken to Build the Model
58.1%	5.35 Seconds

Classifier

Choose **LMT -I -1 -M 15 -W 0.0**

Test options

Use training set

Supplied test set

Cross-validation Folds

Percentage split %

(Nom) Status

Result list (right-click for options)

- 20:57:11 - trees.J48
- 20:58:49 - trees.ADTree
- 20:58:55 - trees.BFTree
- 20:59:39 - trees.FT
- 21:00:45 - trees.J48graft
- 21:02:45 - trees.LADTree
- 21:03:24 - trees.SimpleCart
- 21:05:46 - trees.RandomForest
- 21:06:17 - trees.RandomTree
- 21:12:43 - trees.LMT**

Classifier output

[Actual value=v] \* 0.63 +  
[Monthly Income=II] \* 0.09

Time taken to build model: 89.46 seconds

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances	3516	58.6 %
Incorrectly Classified Instances	2484	41.4 %
Kappa statistic	0.1925	
Mean absolute error	0.4391	
Root mean squared error	0.4684	
Relative absolute error	88.8885 %	
Root relative squared error	94.2469 %	
Total Number of Instances	6000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.726	0.526	0.525	0.726	0.61	0.634	Cease
	0.474	0.274	0.683	0.474	0.559	0.634	close
Weighted Avg.	0.586	0.386	0.613	0.586	0.582	0.634	

=== Confusion Matrix ===

a	b	<-- classified as
1939	731	a = Cease
1753	1577	b = close

Status OK



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

LMT

Correctly Classified Instances	Time Taken to Build the Model
58.6%	89.46 Seconds

Classifier

Choose **LMT -I -1 -M 15 -W 0.0**

Test options

Use training set

Supplied test set

Cross-validation Folds

Percentage split %

(Nom) Status

Result list (right-click for options)

- 20:57:11 - trees.J48
- 20:58:49 - trees.ADTree
- 20:58:55 - trees.BFTree
- 20:59:39 - trees.FT
- 21:00:45 - trees.J48graft
- 21:02:45 - trees.LADTree
- 21:03:24 - trees.SimpleCart
- 21:05:46 - trees.RandomForest
- 21:06:17 - trees.RandomTree**
- 21:12:43 - trees.LMT

Classifier output

Monthly income = v : Cease (0/0)

Size of the tree : 10900

Time taken to build model: 0.07 seconds

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances	5019	83.65 %
Incorrectly Classified Instances	981	16.35 %
Kappa statistic	0.6757	
Mean absolute error	0.1892	
Root mean squared error	0.3076	
Relative absolute error	38.304 %	
Root relative squared error	61.8904 %	
Total Number of Instances	6000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.92	0.23	0.762	0.92	0.834	0.941	Cease
	0.77	0.08	0.923	0.77	0.839	0.941	close
Weighted Avg.	0.837	0.147	0.851	0.837	0.837	0.941	

=== Confusion Matrix ===

a	b	-- classified as	
2456	214	a = Cease	
767	2563	b = close	

Status OK



University of Moratuwa, Sri Lanka.  
 Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

RandomTree

Correctly Classified Instances	Time Taken to Build the Model
83.65%	0.07 Seconds

Classifier

Choose **LMT -I -1 -M 15 -W 0.0**

Test options

Use training set

Supplied test set    Set...

Cross-validation    Folds

Percentage split    %

More options...

(Nom) Status

Start    Stop

Result list (right-click for options)

- 20:57:11 - trees.J48
- 20:58:49 - trees.ADTree
- 20:58:55 - trees.BFTree
- 20:59:39 - trees.FT
- 21:00:45 - trees.J48graft
- 21:02:45 - trees.LADTree
- 21:03:24 - trees.SimpleCart
- 21:05:46 - trees.RandomForest
- 21:06:17 - trees.RandomTree
- 21:12:43 - trees.LMT

Classifier output

```

number of leaf nodes: 7
Size of the Tree: 13
Time taken to build model: 12.02 seconds

=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances      3612      60.2 %
Incorrectly Classified Instances    2388      39.8 %
Kappa statistic                    0.2105
Mean absolute error                 0.4346
Root mean squared error             0.4661
Relative absolute error             87.976 %
Root relative squared error         93.7957 %
Total Number of Instances          6000

=== Detailed Accuracy By Class ===
                TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
                0.657   0.442   0.544     0.657   0.595     0.651   Cease
                0.558   0.343   0.67     0.558   0.609     0.651   close
Weighted Avg.   0.602   0.387   0.614     0.602   0.603     0.651

=== Confusion Matrix ===
      a    b  <-- classified as
1753  917 |  a = Cease
1471 1859 |  b = close

```

Status  
OK



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

### SimpleCart

Correctly Classified Instances	Time Taken to Build the Model
60.2 %	12.02 Seconds



Classifier

Choose **LMT -I -1 -M 15 -W 0.0**

Test options

Use training set

Supplied test set

Cross-validation Folds

Percentage split %

(Nom) Status

Result list (right-click for options)

- 20:57:11 - trees.J48
- 20:58:49 - trees.ADTTree
- 20:58:55 - trees.BFTree
- 20:59:39 - trees.FT
- 21:00:45 - trees.J48graft
- 21:02:45 - trees.LADTree
- 21:03:24 - trees.SimpleCart
- 21:05:46 - trees.RandomForest
- 21:06:17 - trees.RandomTree
- 21:12:43 - trees.LMT

Classifier output

```

out of bag error: 0.4525

Time taken to build model: 5.42 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances      5019      83.65 %
Incorrectly Classified Instances    981      16.35 %
Kappa statistic                    0.6696
Mean absolute error                 0.2543
Root mean squared error             0.327
Relative absolute error              51.476 %
Root relative squared error         65.7902 %
Total Number of Instances          6000

=== Detailed Accuracy By Class ===

              TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
              0.826   0.155   0.81       0.826   0.818     0.939   Cease
              0.845   0.174   0.858     0.845   0.852     0.939   close
Weighted Avg.  0.837   0.166   0.837     0.837   0.837     0.939

=== Confusion Matrix ===

      a   b  <-- classified as
2205 465 |  a = Cease
 516 2814 |  b = close

```

Status  
OK



University of Moratuwa, Sri Lanka.  
Electronic Theses & Dissertations  
[www.lib.mrt.ac.lk](http://www.lib.mrt.ac.lk)

Correctly Classified Instances	Time Taken to Build the Model
83.65 %	5.42 Seconds