

LB / DON / 106 / 2016

IT 01 / 138

# Comparison of data mining techniques for the predictive accuracy of payments of leasing customers in Sri Lanka.

H.A.P.L.Perera

139174J

LIBRARY  
UNIVERSITY OF MORATUWA, SRI LANKA  
MORATUWA

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 degree in  
Information Technology.

April 2016

004 "16"  
004 (043)

University of Moratuwa



TH3173

TH3173

+ DVD ROM

(TH3160 - TH3180)

TH3173

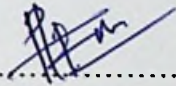
## Declaration

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 Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.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

Signature of Student

Mr. H.A.P.L.Perera

  
.....

Date: 26/04/2016

Supervised by

Name of the Supervisor

Signature of Supervisor

Mr.Saminda Premarathne

***UOM Verified Signature***  
.....

Date: 26/04/2016

## **Dedication**

**This thesis is dedicated to my parents**

**Mr. Peter Henry Perera**

**Mrs. L.M.Subadra**



## **Acknowledgement**

First, I would like to acknowledge my deepest gratitude to my supervisor, Mr.Saminda Premarathne for the continuous support and guidance of my MSc and final thesis. Without his guidance and persistent help this thesis would not have been possible. I would also like to thank all the lecturers who conducted lectures for us during last couple of years. Special thank should go to, Mr. D.K.Withanage, former Dean of the Faculty of Information Technology, University of Moratuwa for his kind advices and Professor Asoka S Karunananda, for his wisdom, knowledge and guidance to the highest standards; which inspired and motivated me a lot.

My sincere thanks also should go to Dr.Anura Karunaratne, Head, Department of Accountancy, University of Kelaniya and all the other department members for their encouragement and guidance.

I am extremely grateful to my parents, who raised me and taught me to study hard and to give priority in my life to the quest for knowledge. Also I wish to express my gratitude to my loving two sisters for supporting me throughout my life.

Last but certainly not least, I would like to thank all my friends who supported me during this period and those who were of direct and indirect help in undertaking and completing this thesis.

## Abstract

The Data mining is the area which helps to uncover hidden patterns and identify correlations from massive amount of structured and unstructured data. With the advent of improved and modified prediction techniques in data mining, there is a need for an analyst to know which tool performs best for a particular type of data set. Hence, selecting the best technique among many techniques at the correct time is very much important and that will save enormous amount of valuable time of decision makers.

This research has been conducted to construct a model, which can be used to measure the predictive accuracy of the credit risk of leasing customers in Sri Lanka and to compare different data mining techniques in the finance domain for the purpose of selecting the best and adequate technique. It is hypothesized that, using Logistic Regression, Naïve Bayes algorithm, Decision Tree-J48 and Neural Networks, credit risk prediction can be addressed in the leasing industry of Sri Lanka.

The dataset employed in this study was obtained from one of the leading finance/leasing companies in Sri Lanka. All the agreements, which were matured on December 2015 were considered for the study under 24 variables. Altogether 8235 customers/ data instances have been considered for the analysis. The variable refining process conducted using the Statistical Package for Social Sciences software. Since the dependent variable is categorical and dichotomous, backward elimination method in the logistic regression was employed. There were nine independent variables and one dependent variable have been selected from the refining process. The data set was divided in to two different datasets, training (60%) and test (40%) data sets. The Waikato Environment for Knowledge Analysis machine learning software was the major software tool used for the entire model construction process. Four (4) main data mining techniques (Logistic Regression, Naïve Bayes, Decision Tree – J 48 and Neural Networks) were used to construct models and results from each model were obtained and compared with other techniques. According to the results of the study, we can conclude that, with healthy classification accuracy, kappa statistic, Area under the curve (AUC) value and F-Measure, a model constructed using the neural network as the best model to predict the payment accuracy of leasing customers in Sri Lanka.

**Keywords:** Data Mining, Credit Risk, Logistic Regression, Naïve Bayes, Decision Tree, Neural Networks



# Table of Contents

Chapter 01 .....	1
Introduction.....	1
1.1 Prolegomena .....	1
1.2 Background and Motivation .....	2
1.3 Significance of the Study.....	3
1.4 Problem Definition .....	5
1.5 Aim and Objectives .....	5
1.5.1 Aim.....	5
1.5.2 Objectives.....	5
1.6 Hypothesis .....	5
1.7 Structure of the thesis .....	5
1.8 Summary.....	6
Chapter 02.....	7
Developments in Data Mining .....	7
2.1 Introduction.....	7
2.2 Background and Essentials of Data Mining.....	7
2.3 What is Data Mining? – Definition.....	7
2.4 Data Mining Paradigms .....	9
2.5 Data Mining Techniques used for Financial Data Analysis .....	10
2.6 Financial Fraud Detection.....	10
2.7 Bankruptcy Prediction .....	10
2.8 Customer Credit Risk Analysis.....	11
2.9 Summary.....	13
Chapter 3 .....	14
Technology Adopted.....	14
3.1 Introduction.....	14
3.2 Technology - Data Mining.....	14
3.2.1 Problem Definition.....	15
3.2.2 Data Gathering and Preparation.....	15
3.2.3 Model Building and Evaluation .....	17
3.2.4 Knowledge Deployment .....	17
3.3 Computing Technique.....	17
3.3.1 WEKA Machine Learning Software.....	17
3.3.2 Statistical Package for Social Sciences (SPSS) .....	17

3.4 Summary .....	18
Chapter 4.....	19
Predictive data mining approach for the credit risk .....	19
4.1 Introduction .....	19
4.2 Hypothesis .....	19
4.3 Input.....	19
4.3.1 Data Introduction.....	19
4.3.2 Variable Selection .....	21
4.4 Process .....	32
4.4.1 Prediction Model .....	32
4.4.2 Data Sampling.....	33
4.5 Output .....	35
4.6 Users .....	35
4.7 Summary.....	35
Chapter 05.....	36
Construction of the Model .....	36
5.1 Introduction.....	36
5.2 Data mining algorithms for classification and prediction.....	36
5.2.1 Naïve Bayes Classifier.....	36
5.2.2 Logistic Regression.....	36
5.2.3 Decision Tree.....	37
5.2.4 Artificial Neural Networks (ANNs).....	37
5.3 Construction of the model using Logistic Regression .....	38
5.3.1: The Classification Accuracy.....	39
5.3.2: Inter-rater Agreement: The Kappa Statistic.....	39
5.3.3: The Confusion Matrix for Logistic Regression .....	40
5.3.4 The ROC Curve / Receiver Operating Characteristic curve .....	41
5.3.5: Mean Absolute Error(MAE) and Root Mean Squared Error(RMSE) ...	44
5.4 Construction of the model using the Naïve Bayes Algorithm .....	45
5.5 Construction of the model using the Decision Tree (J48) Algorithm.....	48
5.6 Construction of the model using the Neural Network (NN) Algorithm.....	52
5.7 Comparison and model selection for the intended solution.....	55
5.8 Summary .....	58
Chapter 6.....	59
Evaluation of the Model.....	59
6.1 Introduction.....	59
6.2 Data Mining Engine.....	59

6.3 Integrating a Data Mining System with a Database/Data Warehouse System .....	59
6.3 Prediction of unseen data using WEKA .....	60
6.4 Summary .....	61
Chapter 7 .....	62
Conclusion and Further Work.....	62
7.1 Introduction.....	62
7.2 Methodology Used and Findings .....	62
7.3 Comparison of data mining models for the predictive accuracy .....	63
7.4 Limitations and Recommendations for Future Work .....	65
7.5 Summary .....	66
References.....	67
Appendices.....	73



# List of Figures

	Page No
<b>Figure 1.1:</b> Product-wise accommodation of the Licensed Finance Companies and Specialized Leasing Companies (Source: Central Bank Annual Report)	04
<b>Figure 2.1:</b> Data mining as a step in the process of knowledge discovery (Source: Han & Kamber, 2012, p.6)	08
<b>Figure 2.2:</b> Paradigms of Data Mining (Source: Maimon and Rokach (2010; p.7))	09
<b>Figure 3.1:</b> Steps of a Data Mining Project (Source: Oracle Database Online Documentation 12c Release 1)	14
<b>Figure 4.1:</b> Categorical variables, after been encoded	29
<b>Figure 4.2:</b> Numerical variables, after been encoded	29
<b>Figure 4.3:</b> Prediction Model: From historical data to new data	32
<b>Figure 4.4:</b> Predictive Model: Current Study	33
<b>Figure 4.5:</b> The process of using the training and testing data sets to measure the predictive accuracy	34
<b>Figure 4.6:</b> Model Construction	34
<b>Figure 4.7:</b> Use the Model in Prediction	34
<b>Figure 5.1:</b> Visualization of all refined variables in WEKA (Training Dataset)	38
<b>Figure 5.2:</b> Calculation of the Kappa statistic	39
<b>Figure 5.3:</b> Calculation of Random Accuracy	40
<b>Figure 5.4:</b> The ROC curve for Logistic regression for active customers (Training Dataset)	42
<b>Figure 5.5:</b> The ROC curve for Logistic regression for sink customers (Training Dataset)	42
<b>Figure 5.6:</b> The ROC curve for Naïve Bayes for active customers (Training Dataset)	46
<b>Figure 5.7:</b> ROC curve for Naïve Bayes Algorithm - Training Data (Sink Customers)	46
<b>Figure 5.8:</b> The ROC curve for the Decision Tree (J48) Algorithm	49

for active customers (Training Dataset)	
<b>Figure 5.9:</b> The ROC curve for the Decision Tree (J48) Algorithm	50
for sink customers (Training Dataset)	
<b>Figure 5.10:</b> The ROC curve for the Neural Network Algorithm	53
for active customers (Training Dataset)	
<b>Figure 5.11:</b> The ROC curve for the Neural Network Algorithm	53
for sink customers (Training Dataset)	



# List of Tables

	Page No
<b>Table 2.1:</b> Limitations of the studies conducted under credit risk predictions	13
<b>Table 4.1:</b> Assigned codes for each categorical variable	23
<b>Table 4.2:</b> Model Summary for the Logistic Regression	24
<b>Table 4.3:</b> Output for the Hosmer and Lemeshow Test	24
<b>Table 4.4:</b> Classification Table	25
<b>Table 4.5:</b> Variables in the equation	26
<b>Table 4.6:</b> Final list of refined variables in the model	27
<b>Table 4.7:</b> Output - Collinearity Diagnostic Test	31
<b>Table 5.1:</b> WEKA output - Evaluation Summary for logistic regression – Training Data Set	39
<b>Table 5.2:</b> The confusion matrix for Logistic regression (Training Dataset)	40
<b>Table 5.3:</b> WEKA output for detailed accuracy by class for the Logistic Regression (Training Data)	43
<b>Table 5.4:</b> Evaluation summary on training data set for the Naïve Bayes Algorithm	45
<b>Table 5.5:</b> The confusion matrix for the Naïve Bayes algorithm (Training Dataset)	45
<b>Table 5.6:</b> WEKA output for detailed accuracy by class for Naïve Bayes (Training Data)	47
<b>Table 5.7:</b> Evaluation summary on training data set for the Decision Tree (J48) Algorithm	48
<b>Table 5.8:</b> The confusion matrix for the Decision Tree (J48) Algorithm (Training Dataset)	48
<b>Table 5.9:</b> WEKA output for detailed accuracy by class for Decision Tree (J48) (Training Data)	50
<b>Table 5.10:</b> Evaluation summary on training data set for the Neural Network Algorithm	52



<b>Table 5.11:</b> The confusion matrix for the Neural Network Algorithm (Training Dataset)	52
<b>Table 5.12:</b> WEKA output for detailed accuracy by class for the Neural Network Algorithm (Training Data)	54
<b>Table 5.13:</b> Comparison of classification accuracy rates of four data mining techniques	55
<b>Table 5.14:</b> Comparison of Kappa statistics and AUC values of four data mining techniques for training data.	56
<b>Table 5.15:</b> Performance of data mining techniques based on the average of training and test results: recall, precision, F-measure, Mean Absolute Error (MAE) and Root Mean Absolute Error (RMAE).	57
<b>Table 7.1:</b> Summary of the performance of four classifiers	63
<b>Table 7.2:</b> The best model to predict the payment accuracy of leasing customers in Sri Lanka.	64