
# ENHANCING ACCURACY OF CREDIT SCORING CLASSIFICATION WITH IMBALANCE DATA USING SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE – SUPPORT VECTOR MACHINE (SMOTE-SVM) MODEL

MUHAMMAD TOSAN BINGAMAWA

MASTER OF COMPUTER SCIENCE
(SOFTWARE ENGINEERING AND INTELLIGENCE)

2017


## Faculty of Information and Communication Technology

# ENHANCING ACCURACY OF CREDIT SCORING CLASSIFICATION WITH IMBALANCE DATA USING SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE – SUPPORT VECTOR MACHINE (SMOTE-SVM) MODEL

**Muhammad Tosan Bingamawa**

**Master of Computer Science in Software Engineering and Intelligence**

**2017**

# ENHANCING ACCURACY OF CREDIT SCORING CLASSIFICATION WITH IMBALANCE DATA USING SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE – SUPPORT VECTOR MACHINE (SMOTE-SVM) MODEL

**MUHAMMAD TOSAN BINGAMAWA**

**A thesis submitted**
**in fulfilment of the requirements for the degree of Master of Computer Science**
**in Software Engineering and Intelligence**

**Faculty of Information and Communication Technology**

**UNIVERSITI TEKNIKAL MALAYSIA MELAKA**

**2017**

## DECLARATION

I declare that this thesis entitled “Enhancing Accuracy of Credit Scoring Classification with Imbalance Data Using Synthetic Minority Oversampling Technique – Support Vector Machine (SMOTE-SVM) Model” is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : .....

Name : .....

Date : .....

**APPROVAL**

I hereby declare that I have read this dissertation / report and in my opinion this dissertation / report is sufficient in terms of scope and quality as a partial fulfilment of Master of Computer Science (Software Engineering and Intelligence).

Signature : .....

Supervisor Name : .....

Date : .....

**DEDICATION**

To my beloved mother and father

"Ibu Sukini and Bapak Ferry Priyono"

To my beloved brother and sister

"Muhammad Dhiyaa'us Zaman and Triesa Aprilia Cahayani"

To my beloved family

"Keluarga Besar Alm. Bp. Sagi Sastrorejo and Keluarga Besar Alm. Bp. Parno Widodo"

To all my beloved friends

To my beloved country

"Indonesia"

**ABSTRACT**

Credit is one of the business models that provide a significant growth. With the growth of new credit applicants and financial markets, the possibility of credit problem occurrence also become higher. Thus, it becomes important for a financial institution to conduct a preliminary selection to the credit applicants. In order to do that, credit scoring becomes one of the models used by a financial institution to perform a preliminary selection of potential customer. One of the most common techniques used to develop a credit scoring model is data mining classification task. However, this technique provides difficulties in classifying imbalanced data distribution. It is because imbalanced data problem may lead the classifier to perform misclassification by classified all of the data into majority class and perform poorly on minority class. In the case of credit scoring, credit data also have imbalanced data distribution. Therefore, classifying a credit data with imbalanced data distribution using unappropriated technique may lead the classification provides a wrong decision result for a financial institution. In this study, several methods for handle imbalanced data problem are identified. Moreover, an improvement of credit scoring model with imbalanced data problem in a financial institution using SMOTE-SVM model is also proposed in this study. This study is conducted in five phases which are data collection, data pre-processing, feature selection, classification, validation, and evaluation. For the experiments using SMOTE-SVM model, the experiments are conducted by taking a consideration in different data ratio and nearest neighbours used in SMOTE. The result of experiments provides that the accuracy and performance result are improved along with the balanced data using SMOTE-SVM model. The performance measurement using 10-fold cross validation and confusion matrix shows that SMOTE-SVM model can correctly classify most of the data in each class with the good result of accuracy, class precision, and class recall. Based on this result, an SMOTE-SVM model is believed to be effective in handling imbalanced data for credit scoring classification.

***ABSTRAK***

*Kredit adalah salah satu model perniagaan yang menyediakan satu pertumbuhan yang sangat pesat. Dengan pertumbuhan pemohonan kredit baru dan pasaran kewangan, kemungkinan besar masalah kredit juga akan meningkat. Justeru itu, sangat penting untuk institusi kewangan menjalankan satu pemilihan awal bagi pemohonan kredit. Oleh itu, permarkahan kredit dijadikan salah satu model yang digunapakai oleh syarikat kewangan dalam menjalankan satu pemilihan awal bagi pelanggan yang berpotensi. Salah satu teknik yang sering digunakan dalam pemarkahan model kredit adalah data melombong dalam klasifikasi. Walau bagaimanapun, teknik ini mempunyai masalah dalam mengklasifikasikan pengagihan data yang tidak seimbang. Masalah ini terjadi kerana ketidakseimbangan data yang boleh membawa pengelas melaksanakan kesalahan klasifikasi bagi kekelasan data oleh semua pengklasifikasi dalam kelas majoriti dan berfungsi dengan sangat teruk terhadap kelas minoriti. Dalam kes skor kredit, data kredit juga mempunyai taburan data yang tidak seimbang. Oleh itu, mengklasifikasikan data kredit dengan pengagihan data yang tidak seimbang menggunakan teknik yang tidak tepat boleh memberikan keputusan yang salah bagi institusi kewangan. Dalam kajian ini, beberapa kaedah untuk menyelesaikan masalah data yang tidak seimbang telah dikenal pasti. Selain itu, peningkatan model bagi pemarkahan kredit dengan masalah data yang tidak seimbang dalam institusi kewangan menggunakan model SMOTE-SVM juga dicadangkan dalam kajian ini. Kajian ini dijalankan dalam lima fasa iaitu pengumpulan data, pemprosesan data, pemilihan rencana, pengelasan, pengesahan, dan penilaian. Dalam eksperimen menggunakan model SMOTE-SVM, eksperimen dijalankan dengan mengambil satu pertimbangan dalam nisbah data yang berbeza dan jiran-jiran terdekat digunakan dalam SMOTE. Keputusan eksperimen menunjukan bahawa keputusan lebih tepat dan prestasi diperbaiki dengan data yang ditunjukan adalah seimbang dengan menggunakan model SMOTE-SVM. Pengukuran prestasi menggunakan 10 kali ganda pengesahan silang dan kekeliruan matriks menunjukkan bahawa model SMOTE-SVM dapat mengklasifikasikan dengan betul kebanyakan data dalam setiap kelas dengan mendapat keputusan baik dari segi ketepatan, ketepatan kelas, dan penarikan balik kelas. Berdasarkan keputusan ini, model SMOTE-SVM boleh dipercayai dan sangat berkesan dalam mengendalikan data yang tidak seimbang untuk pengelasan permarkahan kredit.*

## ACKNOWLEDGEMENTS

First of all, I am grateful to Allah SWT, The Almighty God for establishing me to complete this master project.

Second, I would like to take this opportunity to express my sincere acknowledgement to my supervisor Prof. Dr. Burairah Bin Hussin from Faculty of Information and Communication Technology University Technical Malaysia Malacca (UTeM) for his essential supervision, support, and encouragement towards the completion of this thesis.

Third, I would like to express my greatest gratitude to Ministry of Education and Culture Indonesia, Ministry of Research Technology and Higher Education Indonesia, and also Dian Nuswantoro University that provides an opportunity and financial support to continuing master study in UTeM Malaysia.

Last but not least, I would also like to express my deepest gratitude to my beloved parents, Ibu Sukini and Bapak Ferry Priyono, my beloved brother and sister, Muhammad Dhiyaa'us Zaman and Triesa Aprilia Cahayani, and my beloved family, Keluarga Besar Alm. Bp. Sagi Sastrorejo and Keluarga Besar Alm. Bp. Parno Widodo, for their moral support in completing this degree.

Special thanks to all my peers in UTeM and Indonesia for their support during the study. Lastly, thank you to everyone who had been become parts of realisation of this project.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS

| | | |
|---|---|---|
| ANN | - | Artificial Neural Network |
| BN | - | Bayesian Network |
| BS | - | Beam Search |
| CRISP-DM | - | Cross-Industry Standard Process for Data Mining |
| DA | - | Discriminant Analysis |
| DTM | - | Decision Tree Method |
| GA | - | Genetic Algorithm |
| GR | - | Gain Ratio |
| IG | - | Information Gain |
| KDD | - | Knowledge Discovery in Database |
| KNN | - | k-Nearest Neighbours |
| LDA | - | Linear Discriminate Analysis |
| LG | - | Logistic Regression |
| LR | - | Linear Regression |
| MAR | - | Missing at Random |
| MBND | - | Missing by Natural Design |
| MCAR | - | Missing Completely at Random |
| MNAR | - | Missing not at Random |
| ROC Curve | - | Receiver Operating Characteristic Curve |
| RST | - | Rough Sets |

| | | |
|---|---|---|
| SMOTE | - | Synthetic Minority Over-sampling Technique |
| SOM | - | Self Organization Map |
| SPSS | - | Statistical Package for the Social Science |
| SVM | - | Support Vector Machine |

## LIST OF APPENDICES

# CHAPTER 1

# INTRODUCTION

## 1.1. Background of Study

Credit is one of the business models that provide a significant increase of growth. It is proved by the significant increase of the credit card users in the last decade. However, the type of credit is not only available in a credit card form. Recently, the form of credit can be various, such as automotive loans (car, motorcycle, and any other vehicles), home loans, and also business loans. With many various types of credit and the ease of requirement in applying credit loan, makes many people want to apply new credit loans to a financial institution. Previously, a bank is the most common parties used to apply for a credit loan. But recently, there are a lot of relevant parties that are concerned about credit activity. With the growth of credit applicants and financial markets, the possibility of the credit problem occurrence is also become higher (Lang and Sun, 2014). An example of the problem that frequently occurs in the credit activity is lost credit or bad credit. That problem might be happening due to several reasons. One simple example of the reason is regarding the financial aspect of the credit customer during the credit repayment period. Because of that reason, the customer might not fulfil the obligation to repay the loan instalment. Therefore, it is very important for a financial institution that offers credit services to consider the prospective credit applicants who deserve to get the credit loan.

In order to overcome a credit problem, credit scoring model is developed to help the financial institution in managing the credit risk problem, especially for lost credit problem. Credit scoring is one of the models used to perform a preliminary selection of potential

customer at the earlier phase in requesting credit loans. The basic concept of credit scoring model is to predict the credit score of new credit applicants by comparing their data with the performance of past credit customer data that has been analysed (Wang and Huang, 2009; Marikkannu and Shanmugapriya, 2011). From this credit scoring result, the financial institution can decide whether the new credit applicants will be accepted or rejected. With the effective use of credit scoring implementation, credit scoring model is widely implemented as a decision support system in the financial institution.

Along with the development of the technology, credit scoring model and technique is also constantly improved. Various method can be implemented in the development of credit scoring models such as statistical model, machine learning, and artificial intelligent (Zhang et al., 2008; Ping, 2009). Besides that, there is also data mining technique used in the implementation of credit scoring model. According to Zhang and Wang (2011), the use of data mining for developing credit scoring model can perform effective classification. Data mining is a technique used to extract hidden knowledge from the set of data. It provides several performances for data analysis such as classification, clustering, association, estimation, prediction, and description (Larose, 2005). In the implementation in credit scoring model, classification is a technique used to differentiate and classify new credit applicant whether it is categorised as a bad or good credit. Several classification techniques like a neural network, decision tree, naïve bayes, k-nearest neighbour, and support vector machine are often used by the researchers to propose an effective credit scoring model.

The research about credit scoring model itself is increasingly important. It is because the development of credit scoring model can help the financial institution to reduce the occurrence of credit risk problem. From several techniques of classification that is available, Support Vector Machine (SVM) are believed to be less prone to the class imbalance problem than other classifications learning algorithms (Sun et al., 2009). Moreover, SVM also has

highly accurate, owing to their ability to model complex nonlinear decision boundaries. However, there are several problems that might occur in the implementation of credit scoring using classification technique. Misclassification is the most significant problem that needs to be concerned. Moreover, in the credit scoring model, the most important task is to classify the credit applicants whether it is categorised as good credit or bad credit. With a large amount of real world data, it is generated that the data may have a significant problem in the data distribution. This problem arose because of the numbers of different classes are differ greatly each other. This problem is known as an imbalanced data problem. Customer credit data also have this imbalance data problem, where the differences of class with "good" category have become the majority data. While customer credit with "bad" category only has a few percentage of the whole credit data. With high differences in the data amount between each class, the classification task may perform misclassification result. This problem may cause performance error for the credit scoring model. That is why building classification model in the case of credit scoring with the enhancement of handling imbalance credit data becomes a new challenge that still interesting to be discussed.

## 1.2. Statement of the Purpose

The purpose of this research is to propose an enhancement technique in order to provide a good credit scoring model used by a financial institution. The proposed model will be effective in handling imbalance data credit and also provide a good classification result in terms of accuracy and performance.

## 1.3. Problem Statement

With the demand of new credit applicants, it becomes important for a financial institution to conduct a preliminary selection to the credit applicants. The selection can be performed using the credit scoring model in order to differentiate between good credit and bad credit. With many various techniques that can be implemented in credit scoring model, data mining has become one of the common technique used to develop a credit scoring model. By using data mining classification approach, it would be possible to perform an early selection of the credit customer before granting them credit loans. However, classification approach provides difficulties in classifying the imbalance data distribution. Classify imbalance data using unappropriated technique may lead the classification result into the wrong result. In the case of credit scoring data, credit data often have imbalanced data distribution. Which means the distribution portion of good credit compare with bad credit is not well distributed. Since credit scoring model is used as decision support system in the financial institution, it is become important to have a credit scoring model that can perform well in terms of performance and accuracy with the ability on handling imbalance credit data. In this study, the analysis to provide an improvement on credit scoring model with imbalance dataset in the financial institution is conducted.

## 1.4. Research Question

Based on the problem stated, research question for this research can be derived as follows:

1. What is the suitable method to handle imbalance data?
2. How to improve classification technique for credit scoring model that has imbalanced data distribution in the credit data?

## 1.5. Research Objectives

Based on the research questions, this research will pursue three objectives, which are:

1. To propose Synthetic Minority Oversampling Technique for imbalance data problem.
2. To enhance credit scoring classification with SMOTE-SVM model.
3. To validate credit scoring classification performances with SMOTE-SVM model.

## 1.6. Research Scope and Limitation

The scope and limitation of this research will be as follows:

1. The data used in this study is a credit data provided by Multindo Auto Finance, Semarang, Indonesia.
2. The algorithm used to handle imbalance data is SMOTE.
3. The classification technique used to model the credit scoring is SVM.
4. The experiment follows CRISP-DM as a standard data mining procedure.
5. The experiment is performed by using Weka and SPSS application.

## 1.7. Organization of the Thesis

This study provides six chapter for this project report. Chapter 1 is the introduction part of this study. This chapter provides information about the origins of the research. Moreover, this chapter also provides a brief outlining of a background of the study, and problem statement that is followed by the research question, research objective, the scope of the study, as well as project report review and chapter summary.

Chapter 2 is the literature review part of this study. This chapter will provide related literature and studies on the research problem. Moreover, this chapter also describes some of the theoretical aspects such as data, data mining, credit scoring, imbalance data, classification task, feature selection, validation and evaluation method, and chapter summary.

Chapter 3 is the methodology. This chapter discusses the methodology used to achieve the research objectives, including research design, step by step process, and proposed model.

Chapter 4 will explains how the experiments are conducted in this study. In this chapter, step by step process of the experiments is performed. Moreover, experiments using proposed technique which is SMOTE-SVM model to handle imbalance data problem in classification task is done.

Chapter 5 will be discuss about the experimental result in the chapter 4. The discussion will be including validation and evaluation used to validate and evaluate the experiments result.

Chapter 6 will discuss the general summary of the project including research background, problem study, proposed method used, experiments and discussion result of this study. Moreover, conclusion of this study is provided in chapter 6. Furthermore, the limitations as well as the suggestion for future work is also provided.

## 1.8. Chapter Summary

In the conclusion, this chapter describes the background of the study that related to the credit scoring implementation in bank or finance company. The problem statement describes in this research is the improvement of classification technique for credit scoring model which has imbalance dataset problem. The objective of this study is to propose an

improvement method of credit scoring that specifies the problem of this research in the problem statement. This chapter also provides the outline of the objective, scope of the study and also the statement of the purpose.