UNIVERSITI TEKNOLOGI MARA

AUTOMATED MODEL SELECTION FOR CORPORATION CREDIT RISK ASSESSMENT USING MACHINE LEARNING

ZULKIFLI BIN HALIM

Thesis submitted in fulfillment
of the requirements for the degree of
Philosophy of Doctorate
(Computer Science)

Faculty of Computer and Mathematical Sciences

May 2023

ABSTRACT

Credit risk assessment is the procedure by the investors or lenders to predict the chances of loan default to measuring the risk. A wrong decision places the institution at risk. Corporation Credit Risk Assessment (CCRA) depends on the financial indicators representing the companies' status at a given time. Nowadays, machine learning is a significant field used in various applications, including the financial domain. The global trend in the CCRA study shows that implementing machine learning and deep learning techniques is expanding rapidly. These techniques have demonstrated their superiority over traditional approaches in many CCRA studies. Machine learning model selection is an iterative process of exploring, evaluating, and improving algorithms. Selecting an optimal model for a particular domain is rigid, challenging, and complicated. No free lunch theorem implies that no particular algorithm or combination of features will always produce considerably superior outcomes to others. Hence, the question arises about selecting the optimal model: the characteristic data for CCRA and the best practice machine learning pipelines. The characteristic data for CCRA includes the features used and data dimension. This study used thirteen features, including ten financial ratios, two macroeconomic variables, and the company's age. The features are selected based on the extensive literature on CCRA studies worldwide. This study also investigates the significance of data dimension in CCRA: single or multi-dimensional, and the correlation of the features. For the best practice machine learning pipelines, various machine learning models are used to discover the best model for CCRA study. This study has proposed an automated model selection based on the exhaustive search algorithm—that caused the timeout and memory leak issues. The proposed automated model selection has solved the timeout and memory leak issue by automatically writing all results in CSV files to reduce memory consumption. The samples of the study are the PN17 status companies. Through the automated model selection, 176 models are created across the experiment settings. The models are based on the four machine learning algorithms: logistic regression, support vector machine, decision tree, and neural network; two ensemble techniques: adaptive boost and bootstrap aggregation; three deep learning algorithms: recurrent neural network, long short-term memory (LSTM), and gated recurrent unit (GRU). Besides that, this study proposed two hybrid LSTM-GRU based models. The hybrid models were LSTM-GRU Double Stack (LGDS) and LSTM-GRU Alternate Double Stack (LGADS). As a result, the proposed automated model selection has found that the LGADS model on multi-dimensional data of FR-only features and without a features correlation setup has outperformed the other models with the highest accuracy and Fl score. The LGADS model achieved 84.2% for both measurements. This study contributed to the body of knowledge by proposing an automated machine learning model selection for the CCRA study. This study might be expanded with extensive scope. The scope can be extended by adding more financial ratios since the FR features are significant for the CCRA study and adding more samples to produce better results.

ACKNOWLEDGEMENT

Alhamdulillah, I express my heartfelt gratitude to ALLAH SWT for granting me the opportunity to pursue and successfully complete my Ph.D. journey. I would like to extend my sincere appreciation and thanks to my supervisor, Dr. Shuhaida Mohamed Shuhidan, for her unwavering guidance and support throughout this challenging undertaking. I am also grateful to my co-supervisor, Dr. Sharifalillah Nordin, whose advice has been invaluable in the completion of my studies. Furthermore, I would like to acknowledge the support of Dr. Nordin Abu Bakar, my former supervisor, during the initial stages of my Ph.D. endeavor.

I dedicate this thesis to my beloved father, , and my mother, whose vision, determination, and prayers have played a crucial role in my education. I would like to express my deep appreciation to my wife, who has been my greatest supporter throughout the arduous journey. I am also immensely grateful to my dear sons, and for their unwavering support and boundless love. I am grateful for their patience during the demanding times of my Ph.D. pursuit. May the knowledge gained from this journey benefit us in this world and the hereafter.

I would like to extend my thanks to my father-in-law, ^ my mother-in-law, , as well as my family members and in-law family members, for their unwavering support and prayers. This triumph is dedicated to all of you. Lastly, I would like to express my gratitude to all my friends, colleagues, and superiors who have provided assistance and facilitated my progress, ultimately enabling me to attain this level of accomplishment.

Alhamdulillah, all praise and thanks are only for Allah in all circumstances.

TABLE OF CONTENT

		Page
CONFIRMATION BY PANEL OF EXAMINERS		**ii**
AUTHOR'S DECLARATION		**iii**
ABSTRACT		**iv**
ACKNOWLEDGEMENT		**v**
TABLE OF CONTENT		**vi**
LIST OF TABLES		**x**
LIST OF FIGURES		**xii**
LIST OF ABBREVIATIONS		**xviii**
CHAPTER ONE INTRODUCTION		**1**
1.1	Research Background	1
1.2	Preliminary Study	5
	1.2.1 Malaysian Credit Risk Management Review	5
	1.2.2 Findings	7
1.3	Research Motivation	9
1.4	Problem Statement	10
1.5	Research Questions	12
1.6	Research Objectives	13
1.7	Research Scope and Limitations	13
1.8	Research Significance and Contribution	15
1.9	Thesis Outline	16
1.10	Summary	18
CHAPTER TWO LITERATURE REVIEW		**19**
2.1	Introduction	19
2.2	Malaysian Corporation Review	19
	2.2.1 Overview	19
	2.2.2 Corporation Credit Risk in Malaysia	22
2.3	Corporation Credit Risk Assessment	24

2.3.1 CCRA Data Characteristics 24
2.3.2 CCRA Techniques 30
2.4 Machine Learning Overview 45
2.4.1 Types of Machine Learning 46
2.4.2 Machine Learning Techniques 47
2.4.3 Advantages and Disadvantages of Machine Learning Techniques 57
2.4.4 Machine Learning Pipelines 61
2.4.5 Automated Machine Learning 64
2.5 CCRA Studies using Machine Learning and Deep Learning 65
2.6 Summary 68

CHAPTER THREE RESEARCH METHODOLOGY 69
3.1 Introduction 69
3.2 Research Framework 69
3.2.1 Conceptual Framework 71
3.3 Research Process 74
3.3.1 Phase 1: Requirement Analysis 74
3.3.2 Phase 2: Data Collection 79
3.3.3 Phase 3: Data Pre-Processing 89
3.3.4 Phase 4: Machine Learning Pipelines 91
3.3.5 Phase 5: Performance Measurement 101
3.4 Summary 103

CHAPTER FOUR MACHINE LEARNING PIPELINES 105
4.1 Introduction 105
4.2 Descriptive, Mean Difference & Significant Analyses 105
4.3 Machine Learning Models Settings 111
4.3.1 Import Python Libraries 111
4.3.2 Define Required Functions 112
4.3.3 Load Data & Set Iteration Values 136
4.4 Automated Machine Learning Pipelines 137
4.4.1 Pipeline 1: Single-Dimensional - FR 138
4.4.2 Pipeline 2: Single-Dimensional-FRME 146