



بسم الله الرحمن الرحيم

∞∞∞∞

تم رفع هذه الرسالة بواسطة / مني مغربي أحمد

بقسم التوثيق الإلكتروني بمركز الشبكات وتكنولوجيا المعلومات دون أدنى

مسئولية عن محتوى هذه الرسالة.

ملاحظات: لا يوجد





Computer Science Department
Faculty of Computer & Information Sciences
Ain Shams University

Developing an Intelligent System Based on Knowledge Engineering Paradigms for Bankruptcy Prediction

Thesis submitted to the Department of Computer Science
Faculty of Computer and Information Sciences
Ain Shams University, Cairo, Egypt

In partial fulfillment of the requirements for the master's degree of Computer and
Information Sciences

Submitted by:

Samar Aly Mohamed Taha Khalifa
B.Sc. of Computer Science,

Teaching Assistant, Department of Computer Science,
Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt.

Under Supervision of:

Prof. Dr. Abdel-Badeeh Mohamed Salem
Professor of Computer Science
Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt.

Dr. Marco Alfonse Tawfik
Lecturer in the Department of Computer Science,
Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt.

May 2022, Cairo, Egypt

Table of Contents

	Page No.
Acknowledgment.....	iii
ABSTRACT	iv
List of Publications	vi
List of Tables	vii
List of Figures.....	viii
List of Algorithms	ix
List of Abbreviations	x
Chapter 1 Introduction	1
1.1 Thesis Motivation	2
1.2 Thesis Objective	2
1.3 Thesis Importance	3
1.4 Thesis Methodology.....	3
1.5 Thesis Contributions	4
1.6 Thesis Organization.....	4
Chapter 2 Overview of Bankruptcy Prediction	6
2.1 What is Bankruptcy?	6
2.2 Bankruptcy Prediction	6
2.3 Statistical Techniques for Predicting Bankruptcy	7
2.3.1 Multiple Discriminant Analysis (MDA)	7
2.3.2 Logistic Regression (LR)	7
2.4 Machine Learning in Predicting Bankruptcy	8
2.4.1 Overview of Machine Learning.....	8
2.4.2 Machine Learning Techniques for Predicting Bankruptcy	10
2.5 Summary	17
Chapter 3 Related Works	18
3.1 The most common experimental Datasets Features	18
3.2 Statistical Based Models for Predicting Bankruptcy	20
3.3 Machine Learning Techniques in Predicting Bankruptcy	21
3.3.1 Machine Learning Techniques for Predicting Bankruptcy without Pre-processing Step...21	
3.3.2 Machine Learning Techniques for Predicting Bankruptcy with Pre-processing Step	29
3.4 Summary	39
Chapter 4 Developing an Intelligent System for Predicting Bankrupt	40
4.1 System Overview	40

4.2 Datasets Pre-processing	41
4.2.1 Handling Missing Values	42
4.2.2 Pre-processing Techniques to Balance Datasets	42
4.3 Machine Learning for Data Classification	45
4.3.1 AdaBoost	46
4.3.2 XGBoost	47
4.3.3 CatBoost	48
4.4 Performance Evaluation	49
4.4.1 Datasets Used	49
4.4.2 Metrics	50
4.5 Results and Discussion	51
4.6 Conclusion	62
Chapter 5 Intelligent System for Enhancing the Bankruptcy Prediction with Imbalanced Data Using Oversampling and CatBoost	64
5.1 System Overview	64
5.2 Feature Selection	66
5.2.1 Filter-based Feature Selection	66
5.2.2 Wrapper-based Feature Selection	66
5.3 Performance Evaluation	68
5.3.1 Polish Enterprises Dataset	68
5.3.2 The Used Metrics	68
5.4 Results and Discussion	69
5.5 Conclusion	73
Chapter 6 Summary, Conclusions, and Future Work	75
6.1 Summary	75
6.2 Conclusions	75
6.3 Future Work	76
References	77

Acknowledgment

Thanks, God, for guiding me through my life. I would like to thank my supervisors Prof.Dr. Abdel-Badeeh Mohamed Salem and Dr. Marco Alfonse for providing me with knowledge and experience during the writing of this work. I would also like to thank my family for their patience and support during the journey of working on my masters. Especial thanks to my son Yassin for accompanying me through this journey. And finally, thanks to my colleagues and my friends who gave me support.

ABSTRACT

Early prediction of bankruptcy events is one of the most important topics in finance and investment decision making. Bankruptcy can lead to severe consequences on both micro and macroeconomics. Preventing financial institutions from going bankrupt requires efficient models for the prediction of bankruptcy. For predicting bankruptcy, some studies have applied traditional statistical techniques, while other studies have applied Artificial intelligence (AI) techniques. This research conducted a comparison among various studies in predicting bankruptcy. These studies applied different techniques on several datasets.

The main objective of this research is to determine the strength and weakness of the models applied to predict bankruptcy and their effect. This research proved that AI techniques especially machine learning are more efficient than statistical techniques in predicting bankruptcy. Machine learning based classifiers have been heavily utilized in predicting bankruptcy. In terms of machine learning, predicting bankruptcy with imbalanced dataset is a very big challenge. Despite the variety of the existing models for predicting bankruptcy, it is an interesting topic of research to find a model that achieves a high-performance measurement with working on imbalanced datasets. This is because imbalanced dataset misleads the classification results.

Therefore, this research attempted to design two efficient machine learning based systems to predict bankruptcy while solving the imbalanced dataset problem. The datasets used were selected from the University of California, Irvine (UCI) machine learning repository. The data used with the first system is a selection of three different imbalanced datasets, which are the Polish, the Australian and the German datasets. The first system consists of three main stages: pre-processing the dataset, re-sampling the dataset and applying varied machine learning classifiers to predict bankruptcy.

The first system has applied four varied re-sampling techniques to balance the input training dataset for more reliable performance. The first system also applied six machine learning classifiers for predicting bankruptcy. The best experimental results of the first system showed that the performance measures of accuracy and Area Under the Curve (AUC) are 97% and 95.4%, respectively with the Polish dataset. Moreover, the best experimental results of the first system showed that the performance measures of accuracy and AUC are 88.4% and 92%, respectively with the Australian dataset. The best experimental results of the first system showed that the performance measures of accuracy and AUC are 81.5% and 83.4%, respectively with the German dataset.

The second system consists of four main steps: pre-processing the dataset, re-sampling the training dataset, selecting the most relevant feature from dataset and then applying CatBoost classifier to predict bankruptcy. It used the oversampling as a good re-sampling technique. It used the Categorical Boosting (CatBoost) classifier to classify between bankrupt and non-bankrupt classes. Moreover, the main objective of the second system was to reduce dimensionality of the used dataset for increasing classification performance. The second system applied three varied feature selection methods. It was evaluated on the imbalanced Polish dataset. The experimental results of the second system showed the effectiveness of the developed system according to the accuracy measure. For predicting bankruptcy on the Polish five years datasets, the performance measures of the second system in terms of accuracies are 98%, 98%, 97%, 97% and 95%, respectively.

List of Publications

This research includes the following three publications:

1. Samar Aly, Marco Alfonse and Abdel-Badeeh M. Salem, “Bankruptcy Prediction Using Artificial Intelligence Techniques: A survey”, Digital Transformation Technology in Proceeding of Internet of Things: Applications & Future (ITAF 2020) Online conference, Cairo, Egypt, Springer Nature, pp. 335-360, ISSN: 2367-3370, E-ISSN: 2367-3389, Q4, 0.17 SJR. (Published 2021, Scopus, doi:10.1007/978-981-16-2275-5_21).
2. Samar Aly, Marco Alfonse, Mohamed I. Roushdy and Abdel-Badeeh M. Salem, “Developing an Intelligent System for Predicting Bankruptcy”, Journal of Theoretical and Applied Information Technology, Q4 journal, April 2022, pp. 2068-2088, Vol.100. No 7, Scopus, ISSN: 1992-8645.
3. Samar Aly, Marco Alfonse and Abdel-Badeeh M. Salem, “Intelligent Model for Enhancing the Bankruptcy Prediction with Imbalanced Data Using Oversampling and CatBoost”, International journal of Intelligent Computing and Information Sciences journal. (Accepted and pending publication).

List of Tables

Page No.

Table 3.1: Represents the most common used datasets for predicting bankruptcy and their features.....	18
Table 3.2: presents the tuned hyper-parameters and default setting for RF, GBoost, XGBoost, LBoost and CatBoost by [90].	25
Table 3.3: Analysis of machine learning-based models presented for predicting bankruptcy without pre-processing step.	26
Table 3.4: Analysis of machine learning-based models presented for predicting bankruptcy with pre-processing step.	37
Table 4.1: Meta-parameters for each applied machine learning classifier.....	49
Table 4.2: The description of the used datasets for evaluating the first developed system.	50
Table 4.3: The performance measurements of the first developed system across the selected datasets after applying oversampling strategy.	52
Table 4.4: The performance measurements of the first developed system across the selected datasets after applying SMOTE strategy.	54
Table 4.5: The performance measurements of the first developed system across the selected datasets after applying under-sampling strategy.....	55
Table 4.6 The performance measurements of the first developed system across the selected datasets after applying SMOTETomek link strategy.....	57
Table 4.7 : Performance evaluation of the first developed system across the Australian dataset.	60
Table 4.8: Performance evaluation of the first developed system across the German dataset.	60
Table 4.9: The performance measurement of the first developed system across the three used datasets.....	62
Table 5.1: Meta-parameters of CatBoost technique.....	67
Table 5.2: The basic information of the Polish dataset.....	68
Table 5.3: The indicators of the performance measurements	69
Table 5.4: The performance evaluation of the second developed system on the Polish dataset with the three used feature selection methods	69
Table 5.5: The performance evaluation of the previous studies on the Polish dataset	72

List of Figures

	Page No.
Figure 2.1: Binary classification of Logistic regression model by log-likelihood method [44].	8
Figure 2.2: The common phases of machine learning [46].	9
Figure 2.3: The separation between two classes using	10
Figure 2.4: Artificial neural networks layers using multi-layer perceptron to	11
Figure 2.5: Six 1s and three 0s, model's prediction is 1 (sub-trees)	12
Figure 2.6: Linear classification between 2 classes by the SVM	14
Figure 2.7: The building structure of the DT technique [57].	15
Figure 2.8: The structure of the parallel and sequential ensemble classifiers.	16
Figure 3.1: The flow chart diagram of XGBoost proposed model by [15].	31
Figure 3.2: presents the steps of the proposed model by [24].	34
Figure 3.3: presents the framework of the proposed model by [27].	36
Figure 4.1: Framework of the first developed system.	40
Figure 4.2: SMOTE technique based on the Euclidian distance [95].	43
Figure 4.3: The under-sampling strategy to balance training dataset [98].	44
Figure 4.4: The SMOTE oversampling technique is applied either alone or followed by Tomek link to balance imbalanced dataset.	45
Figure 4.5: The confusion matrix of the presented performance measurements (accuracy, precision, AUC, and recall).	51
Figure 4.6: Comparison between the presented techniques in terms of average accuracy.	59
Figure 4.7: Comparison between the presented techniques in terms of average AUC ratio.	59
Figure 5.1: The overall steps of the second developed system	65
Figure 5.2: Accuracy performance measure of the second developed system.	71
Figure 5.3: AUC performance measure of the second developed system.	71
Figure 5.4: F-score performance measure of the second developed system.	72

List of Algorithms

	Page No.
Algorithm 2.1. Boosting [26].....	17
Algorithm 3.1. Pseudo-code of Borderline SMOTE [40]	35
Algorithm 4.1. SMOTE [51].....	44
Algorithm 4.2 The pseudo code of the AdaBoost [13].....	46

List of Abbreviations

Abbreviation	Definition
Acc	Accuracy
ACO-DC	ACO based Data Classification
ACO-FS	ACO based Feature Selection
AdaBoost	Adaptive Boosting
AP	Affinity Propagation
ACO	Ant Colony
AUC	Area Under the Curve
AI	Artificial intelligence
ANN	Artificial Neural Network
BA	Bagging
BLOF	Bagging-based Local Outlier Factor
BO	Boosting
BSC	Business Source Complete
CBR	Case-Based Reasoning
CatBoost	Categorical Boosting
CART	Classification and Regression Tress
CBoost	Cluster-based Boosting
CNNs	Convolutional Neural Networks
CFS	Correlation Feature Selection
DT	Decision Tree
D.CatBoost	Default CatBoost
EV	Engineering Village
XGBoost	Extreme gradient boosting
FM	Failure pattern-based Models
FN	False Negative
FRs	Financial Ratios

List of Abbreviations (Continued)

Abbreviation	Definition
FP	False Positive
FSCGACA	Fitness-Scaling Chaotic Genetic Ant Colony Algorithm
GP	Gaussian Process
GA	Genetic Algorithm
CGIs	Governance Indicators
GBoost	Gradient Boosting
HM	Hybrid ensemble-based Model
IHT	Instance Hardness Threshold
K-NNs	K-Nearest Neighbors
LDA	Linear Discriminate Analysis
LOF	Local Outlier Factor
LR	Logistic Regression
μ	mean
MDA	Multiple Discriminant Analysis
NB	Naïve Base
NNs	Neural Networks
OCHE	Overfitting-Cautious Heterogeneous Ensemble
RBF	Radial Basis Function
RF	Random Forest
RFE	Recursive Feature Elimination
RSs	Rough Sets
SFS	Sequential Feature Selection
δ	standard deviation
SVM	Support Vector Machine
SMOTE	Synthetic Minority Oversampling Technique
TN	True Negative
TP	True Positive

List of Abbreviations (Continued)

Abbreviation	Definition
T.CatBoost	Tuned CatBoost
TSCM	Two-Step Classification Method
UV	Unanimous Voting
U.S.	United States
UCI	University of California, Irvine
WoS	Web of Science

Chapter 1 Introduction
