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

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 companying 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.

List of Figures

Page No.

List of Algorithms

Page No.

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