A Machine Learning Approach to assist the Prediction of Loan Characteristics

Name: C. L. Perera

Index No: 198767D

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of Degree of Master of Science in Information Technology

July, 2022

DECLARATION

I declare that this thesis is my 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 Information derived from the published or unpublished work of others has been acknowledged. Due references have been provided on all supporting literature and resources.

C. L. Perera

UOM Verified Signature

Date: 25/07/2022 Signature of Student:

Supervised By:

Mr. Saminda Premarathne

Senior Lecturer

Faculty of Information Technology

University of Moratuwa

UOM Verified Signature

Date: 25/07/2022 Signature of Supervisor:

ACKNOWLEDGEMENT

At the outset, I would wish to acknowledge everyone who contributed for the success of the research project and would wish to appreciate their efforts.

First, I am deeply grateful for the supervision throughout my research and the helps received from my supervisor Mr. S. C. Premarathne, Department of Information Technology, Faculty of Information Technology, University of Moratuwa. I have learned so much from our project discussions and his willingness to motivate me contributed tremendously to the study.

Besides, the entire academic staff of the Faculty of Information and technology, who shared their vast awareness throughout, providing me with a good environment that influenced a lot to achieve this goal, is greatly appreciated.

It is with great pleasure that I thank the staff of the University of Moratuwa, Sri Lanka, for all its efforts and facilities that it has contributed towards successful accomplishment of this postgraduate programme. My colleagues, who supported to complete this research successfully, are also greatly appreciated.

Also, my thank goes to the finance institute, which provided me with the dataset and guided throughout the study.

I am as ever, especially obligated to my parents and brother for their affection, inspiration and backing throughout my life to improve my career.

ABSTRACT

The business environment in Sri Lanka has become complex and competitive with the development of the financial sector and the spread of the Covid-19 pandemic. The number of business organizations and individuals applying for loans has increased. The practices that are being used to predict financial allocation for loans of future periods are based on previous experiences and rough estimates. The most challenging risk faced during this process is the credit risk, which is the risk of lending money to unsuitable loan applicants. Lengthy authentication procedures are being followed by financial institutes prior to approving loans. However, there is no assurance whether the chosen applicant is the right applicant or not. Also, predicting the risks of credit loans prior to becoming non-performing is essential as the outcomes are unbearable except provisions are arranged for anticipated downsides. Thus, this study focused on analyzing the historical data of loans and evaluating customer profiles based on the demographic, geographical, and behavioral data of the customers to enable the prediction of future loan amounts, evaluation of the credit risks of loans and prediction of Non-Performing Loans using Machine Learning (ML) algorithms, in order to help make appropriate choices in the future. An exploratory data analysis was first performed to provide insights on developing marketing strategies based on loan types and to identify the type of customers who can be approached. Thus, three models were devised to predict the identified loan characteristics. Model 1 was devised to predict the future loan amounts with the highest R-squared score of 0.9967 using Light Gradient Boosting Regression. Model 2 was devised to evaluate the credit risk with the highest training and test accuracy of 0.9960 and 0.7842, respectively, using Stacking Ensemble Classification. Model 3 was devised to predict the Non-Performing Loans with the highest training and test accuracy of 0.9999 and 0.9522, respectively, using Random Forest Classification. Finally, the study illustrated a remarkable approach in predicting loan characteristics which ideally suits the ever changing economy. It achieved outstanding results which could enable any financial institute in the country, in minimizing the expected risks.

Keywords: *loan characteristics, loan amount, credit risk, Non-Performing Loans, Machine Learning, exploratory data analysis, Random Forest, Boosting Algorithms, Ensemble Learning*

Table of Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
NPL	Non-Performing Loan
EDA	Exploratory Data Analysis
ANN	Artificial Neural Network
DT	Decision Tree
NB	Naïve Bayes
RF	Random Forest
KNN	K-Nearest Neighbor
LGB	Light Gradient Boosting
SVM	Support Vector Machine
EML	Ensemble Machine Learning