

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 Information derived from the published or unpublished work of others has been acknowledged.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

DECLARATION	i
ACKNOWLEDGEMENT	ii
ABSTRACT.....	iii
List of Figures	viii
List of Tables	xi
Abbreviations	xii
Chapter 1	1
Introduction.....	1
1.1 Problem Background.....	1
1.2 Problem Statement	4
1.3 Aim, Objectives and Research Questions	6
1.3.1 Aim and Objectives.....	6
1.3.2 Research Questions.....	7
1.4 Overview of the dissertation	7
1.5 Summary	7
Chapter 2.....	8
Literature Review.....	8
2.1 Introduction	8
2.2 Findings of previous studies.....	8
2.3 Summary	14
Chapter 3.....	15
Technologies adopted in Predicting Loan Characteristics.....	15
3.1 Introduction	15
3.2 Involvement of Data Mining and Machine Learning.....	15
3.1.1 Data Mining	15
3.1.2 Machine Learning	16
3.3 ML Technologies used to develop models.....	17
3.3.1 Regression.....	17
3.3.2 Naïve Bayes (NB).....	17

3.3.3	Decision Tree (DT)	18
3.3.4	Random Forest (RF)	18
3.3.5	Artificial Neural Network (ANN)	19
3.3.6	Boosting Algorithms	20
3.3.7	Ensemble Learning	21
3.4	Implementation Technologies used to develop models	22
3.5	Summary	23
Chapter 4		24
A Novel Approach for Predicting Loan Characteristics		24
4.1	Introduction	24
4.2	Overview of the Novel Approach for Predicting Loan Characteristics	24
4.3	Conceptual Design	24
4.4	Research Methodology	25
4.5	Hypothesis	26
4.6	Inputs to the models	26
4.7	Outputs from models	26
4.8	Process in brief	26
4.9	Users	27
4.10	Features	27
4.11	Significance of the study	27
4.12	Summary	27
Chapter 5		28
Research Analysis and Design for Prediction of Loan Characteristics		28
5.1	Introduction	28
5.2	Overview of the Proposed Model Design	28
5.3	Attributes Selected for modeling	29
5.4	Data Preprocessing	30
5.5	Devising Prediction Models	30
5.6	Evaluation of Model Performance	32
5.7	Summary	32
Chapter 6		33

Implementation	33
6.1 Introduction	33
6.2 Data Collection.....	33
6.3 Loading of Data and Libraries	33
6.4 Data Preprocessing.....	34
6.4.1 Feature Extraction.....	34
6.4.2 Removing outliers.....	35
6.4.3 Label Encoding of Categorical features.....	35
6.4.4 Feature scaling	36
6.5 Exploratory Data Analysis (EDA)	36
6.6 Prediction Modeling.....	38
6.6.1 Defining the input and target variables.....	38
6.6.2 Splitting training and test sets.....	38
6.6.3 Devising of Models.....	38
6.6.4 Hyperparameter tuning with RandomizedSearchCV.....	39
6.6.5 Evaluation of Model Performance	39
6.7 Deployment of Models.....	40
6.8 Summary	40
Chapter 7.....	41
Research Findings and Evaluation.....	41
7.1 Introduction	41
7.1 Exploratory Data Analysis (EDA)	41
7.1.1 Univariate Analysis.....	41
7.1.2 Multivariate Analysis.....	45
7.1.3 Correlation Analysis	46
7.2 Findings of Developed Models	47
7.2.1 Model 1: Prediction of the loan amount for future periods	47
7.2.2 Model 2: Prediction of the default risks of Loan customers.....	62
7.2.3 Model 3: Prediction of Non-Performing Loans for future periods	74
7.3 Deployment of Models.....	86
7.4 Summary	87

Chapter 8.....	88
Conclusions and Future Works.....	88
8.1 Introduction	88
8.2 Conclusions	88
8.3 Limitations of the Study.....	89
8.4 Recommended Future Studies.....	90
8.5 Summary	90
Chapter 9.....	91
References.....	91
APPENDIX A.....	95
APPENDIX B	98
APPENDIX C	100

List of Figures

Figure 1.1: Process for Loan Sanction	3
Figure 2.1: Neural Network Structure	8
Figure 2.2: Devising of Models	9
Figure 2.3: ROC Curve	9
Figure 2.4: Accuracies provided by different algorithms	10
Figure 2.5: ROC curves for the algorithms.....	10
Figure 2.6: Evaluation of Forecast Accuracy	11
Figure 2.7: Comparison of Accuracy and Execution Time	12
Figure 2.8: ROC curves of (a) Baseline (b) Ada Boost	12
Figure 2.9: Training of the ensemble model.....	13
Figure 2.10: ROC Curve Exploration	13
Figure 2.11: Assessment of the performance of the models.....	14
Figure 3.1: Stages of KDD process.....	16
Figure 3.2: Steps of ML Algorithm	16
Figure 3.3: Decision Tree	18
Figure 3.4: Random Forest	18
Figure 3.5: Illustration of an ANN.....	19
Figure 3.6: Functioning of Boosting Algorithms.....	20
Figure 3.7: Ensemble Learning.....	21
Figure 3.8: Bagging Ensemble learning.....	21
Figure 3.9: Voting Ensemble learning.....	22
Figure 3.10: Anaconda Navigator.....	23
Figure 3.11: Power BI.....	23
Figure 4.1: CRISP-DM Approach	25
Figure 4.2: Research Design	26
Figure 5.1: The Architecture Diagram of the Proposed Model Design	28
Figure 6.1: Loading of Data and Libraries.....	34
Figure 6.2: Feature Extraction	34
Figure 6.3: Removing outliers by Interquartile range.....	35
Figure 6.4: Box plot after removing outliers	35
Figure 6.5: Label Encoding of Facility Type.....	35
Figure 6.6: Feature scaling.....	36
Figure 6.7: Univariate Analysis	36
Figure 6.8: Multivariate Analysis	37
Figure 6.9: Correlation Analysis.....	37
Figure 6.10: Defining the input and target variables	38
Figure 6.11: Splitting training and test sets	38
Figure 6.12: Devising of Models	38
Figure 6.13: Hyperparameter tuning with RandomizedSearchCV	39

Figure 6.14: Evaluation of Model Performance	39
Figure 6.15: Saving of Models.....	40
Figure 6.16: POST methods to receive the request and post back.....	40
Figure 7.1:Log_FacilityAmount Distribution and Probability plots	41
Figure 7.2: Distribution, Violin and Box plots	42
Figure 7.3: Loan Count vs. Loan Status	42
Figure 7.4: Loan Count vs. Facility Type	43
Figure 7.5: Loan Count vs. Month.....	43
Figure 7.6: Average interest rate vs. Month	43
Figure 7.7: Loan Count vs. Year.....	44
Figure 7.8: Loan Count vs. District	44
Figure 7.9: Loan Count vs. Province	44
Figure 7.10: Loan Count vs. Marital Status.....	45
Figure 7.11: Distribution of Age according to Gender	45
Figure 7.12: Facility Amount vs. Occupation.....	46
Figure 7.13: Correlation Analysis.....	46
Figure 7.14: Actual and Predicted Loan Amounts of Linear Regression	48
Figure 7.15: Density plot of Actual and Predicted Loan Amounts of Linear Regression	48
Figure 7.16: Feature importance values obtained using Linear Regression	49
Figure 7.17: Model 1 statistics obtained using Ridge and Lasso Regression.....	49
Figure 7.18: Actual and Predicted Loan Amounts obtained using Decision Tree Regression.....	50
Figure 7.19: Residual plot obtained using Decision Tree Regression	51
Figure 7.20: Random Forest Regression.....	51
Figure 7.21: Best parameters obtained using Random Forest Regression.....	52
Figure 7.22: Feature importance values obtained using Random Forest Regression	52
Figure 7.23: AdaBoost Regression	53
Figure 7.24: Feature importance values obtained using AdaBoost Regression.....	53
Figure 7.25: Gradient Boosting Regression.....	54
Figure 7.26: Feature importance values obtained using Gradient Boosting Regression ..	55
Figure 7.27: Light Gradient Boosting Regression	55
Figure 7.28: Feature importance values obtained using LGB Regression.....	56
Figure 7.29: Hyperparameter tuning of ANN.....	56
Figure 7.30: Network Layer Structure and Model Configuration	57
Figure 7.31: Training and Validation MAE and loss.....	57
Figure 7.32: MAE values of each fold.....	58
Figure 7.33: Validation MAE per epoch.....	58
Figure 7.34: Actual and Predicted Loan Amounts obtained using ANN	59
Figure 7.35: Ensemble Stacking Regression	60
Figure 7.36: Performance Evaluation Statistics of Model 1	61

Figure 7.37: The R-squared scores of Model 1.....	61
Figure 7.38: Model 2 statistics obtained using Logistic Regression	62
Figure 7.39: Model 2 statistics obtained using Naïve Bayes Classification	63
Figure 7.40: Model 2 statistics obtained using Decision Tree Classification.....	64
Figure 7.41: Model 2 statistics obtained using Random Forest Classification.....	65
Figure 7.42: Model 2 statistics obtained using Extra Tree Classification	66
Figure 7.43: Model 2 statistics obtained using Ada Boost Classification	67
Figure 7.44: Model 2 statistics obtained using Gradient Boosting Classification.....	68
Figure 7.45: Model 2 statistics obtained using Light Gradient Boosting Classification ..	69
Figure 7.46: Hyperparameter Tuning of ANN	70
Figure 7.47: Loss and Accuracy values during training	70
Figure 7.48: Test accuracy and confusion matrix.....	71
Figure 7.49: Stacking ensemble classification.....	71
Figure 7.50: Model 2 statistics obtained using stacking ensemble classification.....	72
Figure 7.51: Performance Statistics of Model 2	73
Figure 7.52: Accuracy scores of Model 2.....	73
Figure 7.53: Model 3 statistics obtained using Logistic Regression	74
Figure 7.54: Model 3 statistics obtained using Naïve Bayes Classification	75
Figure 7.55: Model 3 statistics obtained using Decision Tree Classification.....	76
Figure 7.56: Model 3 statistics obtained using Random Forest Classification.....	77
Figure 7.57: Model 3 statistics obtained using Extra Trees Classification.....	78
Figure 7.58: Model 3 statistics obtained using Ada Boost Classification	79
Figure 7.59: Model 3 statistics obtained using Gradient Boosting Classification.....	80
Figure 7.60: Model 3 statistics obtained using Light Gradient Boosting Classification ..	81
Figure 7.61: Hyperparameter Tuning of ANN	82
Figure 7.62: ANN layer structure	82
Figure 7.63: Loss and accuracy values during training	83
Figure 7.64: Test accuracy and confusion matrix	83
Figure 7.65: Stacking ensemble classification.....	83
Figure 7.66: Model 3 statistics obtained using Stacking ensemble classification	84
Figure 7.67: Performance Statistics of Model 3	85
Figure 7.68: Accuracy scores of Model 3.....	85
Figure 7.69: Login	86
Figure 7.70: UI for the Prediction of Loan Amount in future periods.....	86
Figure 7.71: UI for the Prediction of credit or default risk in future periods	87
Figure 7.72: UI for the Prediction of Non-Performing Loans in future periods.....	87

List of Tables

Table 5.1: Definition of Variables	29
Table 7.1: Model 1 statistics obtained using Linear Regression	47
Table 7.2: Model 1 statistics of Decision Tree Regression	50
Table 7.3: Model 1 statistics obtained using Random Forest Regression	52
Table 7.4: Model 1 statistics obtained using AdaBoost Regression	53
Table 7.5: Model 1 statistics obtained using Gradient Boosting Regression	54
Table 7.6: Model 1 statistics obtained using LGB Regression	55
Table 7.7: Model 1 statistics obtained using ANN	59
Table 7.8: Model 1 statistics obtained using Stacking Ensemble Regression	60

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