

ABSTRACT

Felicia Adeline Setiawan

Undergraduate Thesis

A Comparative Performance Assessment Of Single Classifier And Ensemble Learning
For Credit Card Default Prediction

The credit card debt crisis has become a significant concern, impacting card-issuing institutions, despite the continuous global rise in credit card customers since 2018. The rise of e-commerce platforms encourages consumerist behavior, with credit cards becoming a preferred and convenient payment method, leading to increased transactions and affecting the risk of customer default. Consequently, banks, as issuers, should avoid this behavior to prevent costly defaults. Machine learning is a recent tool used to predict credit card defaults due to its ability to handle large datasets, explored in previous research. The prevalence of imbalanced data is a common challenge in practice and can significantly impact prediction performance if the existence of imbalanced data is neglected. Hence, this study aims to investigate the impact of imbalanced data by comparing several classification algorithms and identifying features with significant contributions to predictions. Results indicate that Random Forest stands out as the most effective algorithm due to the highest F1-Score compared to others. Additionally, implementing SMOTE to address data imbalances enhances model performance across various imbalance ratios. Certain features, such as payment status in the most recent one to two months and credit card limit balances, play crucial roles in predicting default payments.

Keywords : Credit card debt crisis, machine learning, imbalanced data, classification algorithms, Random Forest, SMOTE, F1-Score, payment status, credit card limit balances, customer default.

TABLE OF CONTENT

APPROVAL SHEET.....	i
AGREEMENT LETTER FOR THE PUBLICATION OF SCIENTIFIC WORKS FOR ACADEMIC PURPOSES.....	ii
FOREWORD.....	iii
ABSTRACT.....	v
TABLE OF CONTENT.....	vi
LIST OF FIGURE.....	viii
LIST OF TABLE.....	ix
LIST OF FORMULA.....	x
LIST OF APPENDIX.....	xi
1. INTRODUCTION.....	1
1.1 Background.....	1
1.2 Research Question.....	3
1.3 Objective.....	3
1.4 Scope and Limitations.....	3
2. LITERATURE REVIEW.....	4
2.1 Credit Card Default.....	4
2.2 Machine Learning Prediction.....	5
2.3 Machine Learning Procedure.....	6
2.3.1 Machine Learning Steps.....	7
2.4 Machine Learning Classification.....	8
2.4.1 Single Classifier.....	8
2.4.2 Ensemble Learning.....	11
2.4.3 Hyperparameter Tuning and Model Validation.....	14
2.4.4 Machine Learning Performance Measurements.....	15
2.5 Imbalance Classification.....	16
3. RESEARCH METHOD.....	20
3.1 Research Step.....	20
3.2 Problem Identification.....	21
3.3 Review of Literature.....	21
3.4 Data Collection.....	21
3.5 Preparing Data.....	21
3.6 Splitting Training and Testing Data.....	21
3.7 Oversampling the Minority.....	22
3.8 Training the Model.....	22
3.9 Testing The Model.....	23
3.10 Performance Evaluation.....	23

3.11 Comparing The Models.....	23
3.12 Sensitivity Analysis.....	23
3.13 Features Importance.....	24
4. RESULT AND ANALYSIS.....	25
4.1 Data Description.....	25
4.2 Descriptive Statistics.....	28
4.2.1 Imbalanced Data Set.....	34
4.3 Data Preparation.....	34
4.4 Training the Model.....	35
4.5 Testing the Model and Evaluation.....	36
4.6 Comparing the Models.....	46
4.7 Sensitivity Analysis.....	47
4.7.1 Sensitivity Analysis Summary.....	56
4.8 Features Importance.....	58
4.8.1 Features Importances Insight.....	64
5. CONCLUSION.....	67
5.1 Conclusion.....	67
5.2 Recommendation.....	68
REFERENCES.....	70
APPENDIX.....	74

LIST OF FIGURE

Figure 2.1 Machine Learning's Algorithms Types.....	6
Figure 2.2 Machine Learning Steps.....	7
Figure 2.3 Logistic regression.....	9
Figure 2.4 Naive bayes classifier.....	10
Figure 2.5 Decision tree model.....	11
Figure 2.6 AdaBoost Algorithms.....	12
Figure 2.7 Flow Chart of Random Forest.....	13
Figure 2.8 XG Boost Flowchart.....	13
Figure 2.9 K-Fold Cross Validation Systematic.....	14
Figure 2.10 Confusion Matrix.....	16
Figure 2.11 Nearest Neighbor Computation Demonstrate.....	18
Figure 3.1 Research Method Flowchart.....	20
Figure 4.1 Histogram of Limit Balances and Percentage Default.....	28
Figure 4.2 Customer's Repayment Status Every Month.....	29
Figure 4.3 Customer's Bill Statement Histogram.....	30
Figure 4.4 Customer's Pay Amount Diagram.....	31
Figure 4.5 Graph of Customer's Age.....	32
Figure 4.6 Bar Plot The Correlation Between Sex, Marriage, and Education with Percentage of Default.....	32
Figure 4.7 The Correlation Between Demographic Factors And The Percentage Of Default.....	33
Figure 4.8 The Number Of Non-Default And Default In The Dataset.....	34
Figure 4.9 Logistic Regression F1-Score in Several Imbalance Ratio.....	48
Figure 4.10 Logistic Regression G-Mean in Several Imbalance Ratio.....	49
Figure 4.11 Decision Tree F1-Score in Several Imbalance Ratio.....	50
Figure 4.13 Decision Tree G-Mean in Several Imbalance Ratio.....	50
Figure 4.14 Naive Bayes F1-Score in Several Imbalance Ratio.....	51
Figure 4.15 Naive Bayes G-Mean in Several Imbalance Ratio.....	51
Figure 4.16 AdaBoost F1-Score in Several Imbalance Ratio.....	52
Figure 4.17 AdaBoost G-Mean in Several Imbalance Ratio.....	53
Figure 4.18 Random Forest F1-Score in Several Imbalance Ratio.....	54
Figure 4.19 Random Forest G-Mean in Several Imbalance Ratio.....	54
Figure 4.20 XGBoost F1-Score in Several Imbalance Ratio.....	55
Figure 4.21 XGBoost G-Mean in Several Imbalance Ratio.....	55
Figure 4.22 The 25th Tree from the Random Forest Model.....	66

LIST OF TABLE

Table 2.1 Review of Previous Research.....	5
Table 2.2 Comparison SMOTE and Random Oversampling.....	19
Table 4.1 Initial Data Attributes.....	25
Table 4.2 Modified Data Attributes.....	27
Table 4.3 Best Hyperparameter for Each Algorithms.....	35
Table 4.4 Best Threshold for Oversampling Data.....	37
Table 4.5 Logistic Regression Performance Evaluation.....	38
Table 4.6 Decision Tree Performance Evaluation.....	39
Table 4.7 Naive Bayes Performance Evaluation.....	41
Table 4.8 AdaBoost Performance Evaluation.....	42
Table 4.9 Random Forest Performance Evaluation.....	43
Table 4.10 XGBoost Performance Evaluation.....	45
Table 4.11 Performance Evaluation of Six Algorithms.....	47
Table 4.12 Sensitivity Analysis Summary.....	57
Table 4.13 Features Importance of Logistic Regression.....	58
Table 4.14 Features Importance of Decision Tree.....	60
Table 4.15 Features Importance of AdaBoost.....	61
Table 4.16 Features Importance of Random Forest.....	62
Table 4.17 Features Importance of XGBoost.....	63

LIST OF FORMULA

(2.1).....	15
(2.2).....	15
(2.3).....	15
(2.4).....	16
(2.5).....	16
(2.6).....	16

LIST OF APPENDIX

Appendix 1 : Logistic Regression Table.....	74
Appendix 2 : Percentage Importance of Decision Tree.....	77
Appendix 3 : Percentage Importance of AdaBoost.....	79
Appendix 4 : Percentage Importance of Random Forest.....	80
Appendix 5 : Percentage Importance of XGBoost.....	82
Appendix 6 : Decision Tree Cross Validation Hyperparameter Test.....	84
Appendix 7 : AdaBoost Cross Validation Hyperparameter Test.....	85
Appendix 8 : XG Boost Cross Validation Hyperparameter Test.....	86
Appendix 9 : Random Forest Cross Validation Hyperparameter Test.....	87
Appendix 10 : AUC-ROC Curve.....	88