Using Machine Learning to Predict Customers' Ability to Repay Loans

Bạch Lương Phúc
Thai Nguyen University of Information and Communication Technology
Thai Nguyen University, Thai Nguyen, Vietnam
student492@ictu.edu.vn

***Abstract*—This undergraduate thesis explores the application of machine learning techniques to predict the ability to repay a loan of customers. The author discusses the challenges and opportunities in the banking sector, reviewing various machine learning algorithms, including Logistic Regression, Decision Trees, and XGBoost. The study involves data preprocessing, model training, and evaluation, but the exact performance metrics are not provided. The model was trained using a dataset, but the results did not achieve high accuracy. The author acknowledges the need for further research to improve the model's predictive power. The limitations include the need for more data and better optimization, as well as the importance of understanding the data and continuous learning to enhance the model's performance.**

***Index Terms*—Machine Learning, Credit Risk, Logistic Regression, Decision Trees, XGBoost, Banking**

I. Introduction

The thesis introduces the problem of loan repayment and the challenges faced by banks, such as credit risk and competition. It also provides an overview of the banking industry and the application of machine learning in this field. The author aims to develop a model to predict the ability to repay a loan of customers, which could help banks make better lending decisions. The study focuses on the application of machine learning algorithms, particularly Logistic Regression, Decision Trees, and XGBoost, to address the issue of credit risk and improve the efficiency of the lending process [1]. Earlier ICTU student work on related designs is included here for context [5], [6], [7]. Related published studies provide context for the subject and methods [8], [9], [10]. These contextual works were selected for this English summary; they are not asserted to appear in the original thesis bibliography.

II. Methods

The author reviews several machine learning algorithms, including Logistic Regression, Decision Trees, and XGBoost. The study involves data preprocessing, model training, and evaluation. The data preprocessing steps include cleaning, normalization, and feature selection. The model training process uses a dataset to train the algorithms, but the exact performance metrics are not provided. The evaluation process involves testing the model on a separate dataset to assess its predictive accuracy. The author acknowledges the need for further research to improve the model's predictive power [1].

Supervisor: Nguyễn Tuấn Anh, Thai Nguyen University of Information and Communication Technology (anhnt@ictu.edu.vn).

III. Results

The model was trained using a dataset, but the results did not achieve high accuracy. The author acknowledges the need for further research to improve the model's predictive power. The limitations include the need for more data and better optimization, as well as the importance of understanding the data and continuous learning to enhance the model's performance. The study did not provide specific performance metrics, such as accuracy, precision, recall, or F1 score, to evaluate the model's effectiveness [1].

IV. Discussion

The author discusses the limitations of the current model, such as the need for more data and better optimization. They also mention the importance of understanding the data and continuous learning to enhance the model's performance. The study highlights the challenges in applying machine learning to the banking sector, including the complexity of financial data and the need for robust data preprocessing. The author suggests that further research is needed to address these challenges and improve the model's predictive accuracy [1].

V. Conclusion

The author concludes that while the project has been completed, the model's accuracy is still low. They plan to continue learning and improving the model to better predict customer ability to repay a loan. The study emphasizes the importance of continuous learning and data understanding in developing effective machine learning models for the banking sector. The author acknowledges the need for more data and better optimization to improve the model's predictive power [1].

References

[1] Bạch Lương Phúc, "Using Machine Learning to Predict Customers' Ability to Repay Loans," internal undergraduate thesis (original in Vietnamese), Thai Nguyen University of Information and Communication Technology, 2024.
[2] Jose, J. (2022). Introduction to Time Series Analysis and its Applications. Christ University, Bangalore
[3] W. C. Apgar, with R. Peng and J. Olson. Recent Trends in Real Rents. Working Paper W87-5, Joint Center for Housing Studies of Harvard University, 1987
[4] "Grid Search for model tuning", Medium, 2022. [Online]. Available: https://towardsdatascience.com/grid-search-for-model-tuning-3319b259367e. [Accessed: 10- Jan- 2022]

[5] Đỗ Thị Thu Hằng, ”Loan Interest Calculation Program for Agribank’s Gang Thep Branch in Thai Nguyen,” internal undergraduate thesis, Thai Nguyen University of Information and Communication Technology, 2014.
[6] Trần Thị Vui, ”Development of a Loan File Management Program for the Agricultural and Rural Development Bank Branch in Song Cau, Thai Nguyen,” internal undergraduate thesis, Thai Nguyen University of Information and Communication Technology, 2013.
[7] ICTU Student 188, ”An Overview of Some Algorithms in Data Mining,” internal undergraduate thesis, Thai Nguyen University of Information and Communication Technology, 2010.
[8] Lkhagvadorj Munkhdalai et al., ”An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments,” Sustainability, 2019. doi:10.3390/su11030699.
[9] Nicolas Suhadolnik et al., ”Machine Learning for Enhanced Credit Risk Assessment: An Empirical Approach,” Journal of Risk and Financial Management, 2023. doi:10.3390/jrfm16120496.
[10] Sunil Bhatia et al., ”Credit Scoring using Machine Learning Techniques,” International Journal of Computer Applications, 2017. doi:10.5120/ijca2017912893.