Predicting Loan Default: A Comparative Analysis of Multiple Machine Learning Models

Yuelin Jiang

Department of Computer Science, University of Toronto, ON M5R 0A3, Toronto, Ontario, Canada

yuelin.jiang@mail.utoronto.ca

Abstract. Financial decision-making, particularly in loan approval, requires precise risk prediction. To enhance the prediction accuracy, this study utilizes various machine learning models, namely Logistic Regression, XGBoost, an Artificial Neural Network (ANN), and a hybrid XGBoost + Logistic Regression (XGB+LR). These models were selected based on their unique capacities to capture complex patterns and relationships within the data, thereby potentially improving the loan default prediction task. The training and validation of these models were performed on a meticulously prepared dataset, following crucial preprocessing steps such as one-hot encoding, feature selection, and scaling. To ensure the models' optimal performance, intensive hyperparameter tuning was conducted. The application of these techniques resulted in a robust set of models. Each model's performance was rigorously evaluated through established metrics, including the Area Under the ROC Curve (AUC) and Accuracy (ACC). Among these models, the XGBoost model demonstrated superior predictive power, achieving an AUC of 0.798 and an ACC of 0.861 on the validation set. A detailed feature importance analysis using the XGBoost model further revealed that Credit_Score and Loan_Amount were the primary factors impacting loan approval decisions. Despite slight overfitting observed in the models, the results confirm the potential of machine learning in improving financial decision-making processes. This study sets the foundation for future advancements, which may include the application of advanced regularization techniques, further hyperparameter optimization, and the inclusion of a broader feature set.

Keywords: Machine learning, feature importance, financial debt.

1. Introduction

Modern life is characterized by a complex interplay of ambitions and challenges, where individuals strive for stability through various means such as homeownership, vehicle ownership, entrepreneurial endeavors, or substantial investments. In pursuit of these goals, individuals commonly resort to borrowing money, especially from banks. However, such behaviors are not devoid of cost. It is a loan that they need to repay, with additional interest. This financial obligation often ends up adding to their stress, as they are now under constant pressure to repay. This is more common than people would like to admit, and a real problem arises when a borrower cannot repay the loan, resulting in a default. This situation benefits no one - the borrower's financial situation is strained, and the lending bank faces losses. Therefore, how to predict which borrowers are most likely to default on their loans is important. It will provide banks and lenders with a tool to help them avoid risky loans. This way, banks can save money and the overall financial health of society can be improved. But this model is not just about preventing losses for banks. It is also about creating a more stable economic environment. If banks can predict which loans are likely to default, they can be more responsible in their lending practices. This can help ensure that loans are only given to those who can realistically repay them, thus promoting financial sustainability.

Over the years, the application of machine learning has witnessed a notable surge, particularly within the domains of banking and economics. Algorithms are being used to analyze large data sets to help predict trends and behavior in different areas. However, a significant research void exists concerning the prediction of loan defaults, necessitating the present study to address this gap by leveraging machine learning methodologies to forecast the likelihood of such defaults. There have been several key studies using machine learning to analyze and predict bank customer behavior. For example, research conducted by Moro et al. provides a solid foundation and guidance for this study

[1]. They highlight the potential of machine learning in the banking industry, particularly in understanding customer behaviour and risk assessment.

In contrast to these previous studies, this research focuses specifically on predicting loan defaults. It will examine a range of borrower data, such as loan amount, gender, loan type, loan purpose, creditworthiness, and more, to determine the likelihood of loan default. By focusing on this specific aspect, the work will distinguish itself from previous studies and add new value to the field.

The decision-making processes used by banks for loan approvals become more crucial as the current global economy faces increasing uncertainties. The importance of loan default issues is on the rise, and there is a clear need for a predictive model that can help banks make better lending decisions. This type of model could result in significant financial savings for businesses and individuals by preventing potential financial issues. Additionally, the insights gained from this study could foster a more secure and sustainable financial environment, encouraging responsible lending behavior and making a significant contribution to a stronger economy.

This study tackles this complex issue using machine learning. Through the use of advanced algorithms and large data sets, a model that can predict which borrowers are more likely to default on their loans can be created. This model could give banks the information they need to make informed decisions before giving out loans, which could help to reduce defaults. More specifically, this study uses several models—Logistic Regression, XGBoost, and Neural Networks—to make solid predictions. These models were chosen for their ability to deal with large data sets and their ability to capture complex relationships between variables. They also make it possible to assess feature importance, providing valuable insights into the factors that greatly influence the probability of loan defaults. These models take into account various factors, including individual demographics, credit history, and loan size, which all contribute to a comprehensive understanding of the loan approval process.

2. Method

2.1. Data Preparation

The dataset used for estimating the probability of borrowers not paying back their debt was sourced from [2]. It consisted of multiple features and a corresponding label indicating whether borrowers defaulted on their loans. Prior to model implementation, the dataset underwent several preparation steps.

Initially, the presence of missing values was determined, and a missing rate was observed to quantify the extent of missingness. Subsequently, Variables deemed irrelevant, such as 'year' and 'ID', were dropped from the dataset. The remaining variables were categorized as categorical or continuous. To handle categorical variables, one-hot encoding was applied. This process converted categorical variables into binary columns, representing different categories. The resulting one-hot encoded variables were then concatenated with the continuous variables and the target variable to create the final dataset for modelling.

2.2. Logistic Regression

Logistic Regression is a widely used classification algorithm that models the relationship between independent variables and a binary outcome [3-5]. In this study, Logistic Regression was employed using the LogisticRegression class from the scikit-learn library. After the dataset underwent the necessary preparation steps, including one-hot encoding and feature selection, the Logistic Regression model was trained on the prepared dataset. Hyperparameters such as the regularization parameter (C), penalty type, and solver were carefully selected. The model was then trained and the parameters were optimized to best fit the training data.

The coefficients of the Logistic Regression model were extracted, providing valuable insights into the importance and directionality of each feature's impact on the probability of borrowers not paying

back their debt. Additionally, an intercept term was included in the model, allowing for the baseline prediction.

The performance of the Logistic Regression model was evaluated using evaluation metrics such as the area under the Area Under Curve (AUC) and accuracy. The AUC provides a measure of the model's ability to discriminate between defaulters and non-defaulters. Accuracy, comparing the predicted labels to the true labels, determined the overall classification performance. These metrics were computed for both the training and validation datasets, offering a comprehensive understanding of the model's performance.

2.3. XGBoost

XGBoost is a powerful gradient boosting algorithm known for its high performance and effectiveness in various machine learning tasks [6-8]. In this analysis, XGBoost was implemented using the XGBClassifier class from the XGBoost library. Prior to training the XGBoost model, the dataset underwent preprocessing steps, including the encoding of categorical variables and feature selection. Hyperparameters such as the number of estimators, learning rate, maximum depth, and regularization parameters were carefully tuned to optimize the model's performance.

The XGBoost model was then trained on the prepared dataset. During the training process, XGBoost iteratively built an ensemble of weak decision trees, optimizing a specific objective function to minimize the loss. By combining the predictions of multiple weak learners, the model improved its predictive capability.

To gain insights into the relative importance of each feature in the XGBoost model, the feature importance scores were subsequently visualized. This visualization aided in identifying the most influential features for predicting the probability of borrowers not paying back their debt.

The performance of the XGBoost model was evaluated using evaluation metrics similar to those used for Logistic Regression, including AUC and accuracy. These metrics were calculated for both the training and validation datasets, providing an assessment of the model's generalization ability.

2.4. Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) are computational models inspired by the structure and function of biological neural networks [9-11]. They excel at capturing complex patterns and relationships from data. In this analysis, an ANN was implemented using the Keras library with a TensorFlow backend.

The architecture of the ANN model consisted of multiple layers of interconnected neurons, with each layer having a specified number of units and an activation function. The number of layers, number of neurons, and activation functions were determined based on experimentation and model performance.

Before training the ANN model, the dataset underwent preprocessing steps, including feature scaling using techniques like Min-Max scaling. This ensured that the input features were on a similar scale and improved the convergence of the model during training. The prepared dataset was then divided into training and validation sets.

The ANN model was trained on the training dataset. During training, the model adjusted the weights and biases of the neurons to minimize the loss function and improve its predictive ability. The performance of the model was evaluated on both the training and validation datasets using metrics such as loss and accuracy.

2.5. XGBoost + Logistic Regression (XGB+LR)

In addition to the individual implementation of Logistic Regression and XGBoost models, a hybrid approach called XGBoost + Logistic Regression (XGB+LR) was also employed. This approach combines the strengths of both models by using the predictions of the XGBoost model as features for training a Logistic Regression model.

Following the preprocessing and preparation of the dataset, as described earlier in this study, the XGBoost model was trained on the meticulously processed dataset. The resulting predictions generated by the XGBoost model for both the training and validation datasets were extracted. Subsequently, a one-hot encoding technique was applied to the obtained XGBoost predictions. This encoding step facilitated the transformation of the continuous predictions into binary columns, each representing distinct categories. The resulting one-hot encoded predictions were then concatenated with the original dataset, excluding the target variable. Finally, a Logistic Regression model was trained on the concatenated dataset, incorporating the one-hot encoded predictions from the XGBoost model as supplementary features. The performance of the Logistic Regression model was meticulously assessed using the same evaluation metrics previously mentioned in this study.

3. Results and Discussion

3.1. Performance Comparison of Different Machine Learning Models

This analysis examines the performance of four distinct machine learning models, namely Logistic Regression, XGBoost, a Neural Network and XGBoost + Logistic Regression. Each model was trained and validated using the same dataset. The models' performance was measured using two metrics, namely Area Under the ROC Curve (AUC) and Accuracy (ACC).

3.1.1. Logistic Regression.

The first model, Logistic Regression, was used to predict whether a loan application would be approved or not. The performance of the Logistic Regression model was notable, with an AUC score of 0.791 and accuracy of 0.862 on the training set. However, upon validation, the model's performance demonstrated a slight drop, with an AUC of 0.787 and an ACC of 0.860 shown in Table 1. This marginal performance decline suggests the model may be overfitting on the training data, potentially due to its high dimensional feature space.

Table 1. The performance of logistic regression.

	Model	Set	AUC	ACC
0	Logistic Regression	Train	0.791	0.862
1	Logistic Regression	Valid	0.787	0.860

3.1.2. XGBoost.

The second model, XGBoost, demonstrated the ability that can effectively handle the diverse feature set of the dataset and manage to outperform the Logistic Regression model on both training and validation sets. With an AUC of 0.823 and an ACC of 0.865 on the training set, and an AUC of 0.798 and an ACC of 0.861 on the validation set shown in Table 2, the XGBoost model exhibited a strong balance between bias and variance, indicating a robust model capable of generalizing well on unseen data.

Table 2. The performance of XGBoost.

	Model	Set	AUC	ACC
0	XGBoost	Train	0.823	0.865
1	XGBoost	Valid	0.798	0.861

3.1.3. Neural Network.

The third examined is a Neural Network, which is versatile, capable of learning complex patterns and dependencies in data. In this context, the Neural Network model achieved an accuracy of 86.15% on the training set and 85.97% on the validation set shown in Table 3. Although it didn't outperform

the XGBoost model, the Neural Network demonstrated substantial prediction power, given its relative simplicity.

Table 3. The performance of ANN.

	Model	Set	AUC	ACC
0	ANN	Train	0.790	0.861
1	ANN	Valid	0.784	0.859

3.1.4. XGBoost + Logistic Regression.

Additionally, this research also employed a hybrid model that merges the strengths of XGBoost and Logistic Regression, referred to as XGBoost + Logistic Regression (XGB+LR). This ensemble model generated promising outcomes on the training set, achieving an AUC of 0.845 and an ACC of 0.869 shown in Table 4. Nevertheless, its performance saw a slight dip on the validation set, yielding an AUC of 0.773 and an ACC of 0.855. This points to a need for further evaluation and optimization to enhance its generalization capabilities and manage potential overfitting issues.

Table 4. The performance of XGB+LR.

	Model	Set	AUC	ACC
0	XGB+LR	Train	0.845	0.869
1	XGB+LR	Valid	0.773	0.855

To summarize, the XGBoost model outperformed the other models across the board, achieving the highest accuracy of 0.861 on the validation set. This superior performance can be ascribed to a variety of reasons. First, XGBoost, being a robust ensemble model, proficiently handles a wide array of features and captures intricate interrelationships. By uniting weak decision trees, it produces a potent predictive model. XGBoost also integrates regularization methods that deter overfitting and bolster generalization. The model's optimization algorithm enhances performance measures, such as the area under the ROC curve, augmenting its discriminatory power. Additionally, XGBoost's ability to estimate feature importance helps in pinpointing influential variables for precise predictions. Taken together, these facets contribute to the XGBoost model's high predictive accuracy on the validation set, cementing its place as the most efficient model among those assessed.

3.2. Feature Importance

Feature importance is an invaluable technique for interpreting the output of a machine learning model, providing insights into the relationship between predictors and the target variable. For the model in this analysis, the feature importance was evaluated using the XGBoost model shown in Figure 1.

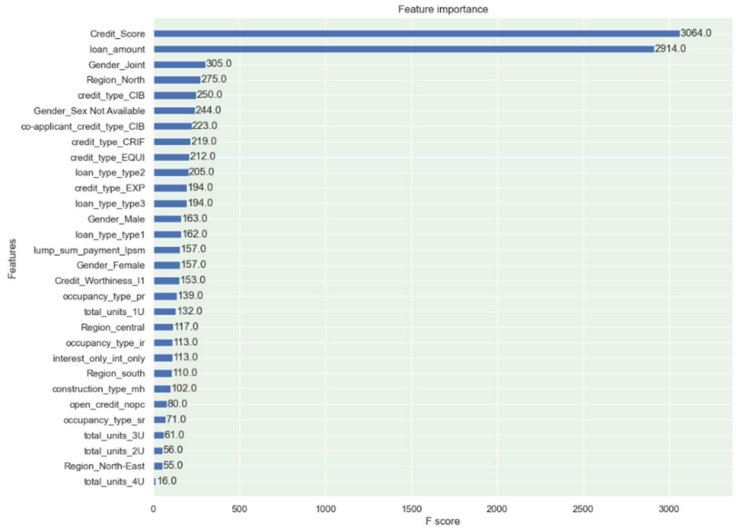

Figure 1. The feature importance based on XGBoost (Photo/Picture credit: Original).

The 'plot_importance' function is an inbuilt method in XGBoost, and it directly leverages the fitted model to calculate the F-Scores for each predictor. These scores reflect the number of times each feature is split on, weighted by the improvement it provides to the model, and averaged over all trees in the ensemble. High F-Score values indicate features that contribute significantly to the model's predictive power.

In this analysis, the most impactful predictors were Credit_Score, Loan_Amount, Gender_Joint, Region_North, and Credit_Type_CIB. The importance values (F-Scores) for these features were 3064, 2914, 305, 275, and 250, respectively. The analysis showed that Credit_Score was the most influential feature in predicting loan approval, which aligns with the commonly acknowledged fact that credit scores are a primary factor considered by financial institutions. Similarly, Loan_Amount was identified as the second most critical feature, indicating that the amount requested significantly impacts loan approval decisions. Gender_Joint and Region_North were found to be less impactful, suggesting a more secondary role in the decision-making process. The presence of these features highlights the model's capacity to capture more nuanced and contextual information.

4. Conclusion

The principal goal of this research was to develop a robust predictive model for determining if a borrower would repay their loan. This objective was achieved by deploying three distinctive machine learning models: Logistic Regression, XGBoost, XGBoost + Logistic Regression and an Artificial Neural Network (ANN). While all models performed commendably, the XGBoost model emerged as the most reliable due to its superior balance of bias and variance. However, the models showed a mild tendency towards overfitting, indicating the scope for improvement. For future enhancements, advanced regularization techniques, further hyperparameter optimization, and refined feature selection methods could be applied. The inclusion of a more comprehensive feature set could also aid in improving the models' predictive power and interpretability. Ultimately, these efforts aim to further the utility of machine learning in financial decision-making, contributing towards an efficient, data-driven loan approval process.

Reference

[1] Moro S Cortez P Rita P 2014 A data-driven approach to predict the success of bank telemarketing Decision Support Systems 62 22–31

[2] Dorman Kaggle Loan Default Dataset 2022 Retrieved from

[3] https://www.kaggle.com/datasets/yasserh/loan-default-dataset

[4] Menard S 2002 Applied logistic regression analysis Sage

[5] Field A 2009 Logistic regression. Discovering statistics using SPSS 264 315

[6] Sperandei S 2014 Understanding logistic regression analysis. Biochemia medica 24(1) 12–18

[7] Chen T He T Benesty M et al 2015 Xgboost extreme gradient boosting R package version 0 4–2 1(4) 1–4

[8] Chen T Guestrin C 2016 Xgboost A scalable tree boosting system Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 785-794

[9] Ogunleye A Wang Q G 2019 XGBoost model for chronic kidney disease diagnosis IEEE/ACM transactions on computational biology and bioinformatics 17(6) 2131–2140

[10] Qiu Y Wang J Jin Z et al 2022 Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training Biomedical Signal Processing and Control 72 103323

[11] Zhou Z H Jiang Y 2003 Medical diagnosis with C4 5 rule preceded by artificial neural network ensemble IEEE Transactions on information Technology in Biomedicine 7(1) 37–42

[12] Abdelatief M A Zamel A A Ahmed S A 2019 Elliptic tube free convection augmentation: an experimental and ANN numerical approach. International Communications in Heat and Mass Transfer 108 104296