Credit Default Analysis and Prediction Based on Machine Learning

Yuxiang Lai

South China Agricultural University, Guangzhou, China

* Corresponding Author Email: 202223220214@stu.scau.edu.cn

Abstract. The Credit Default Prediction project aims to develop an efficient machine learning model to accurately predict loan default risk on the Lending Club platform. The project is based on historical loan data, including borrower personal information, financial metrics, and credit history records, with the goal of building a robust predictive model. The project's workflow encompasses several critical steps, including data preprocessing, exploratory data analysis, feature engineering, model selection, and performance evaluation. In terms of model selection, the project employs two primary machine learning algorithms, namely Artificial Neural Networks (ANN) and Random Forest. These algorithms are renowned for their outstanding performance in handling extensive borrower data and providing reliable risk predictions. Model training and evaluation are conducted using a substantial amount of historical data to ensure accurate predictions across various scenarios. Furthermore, the project conducts feature importance analysis to identify factors that significantly influence loan default risk. These insights contribute to enhancing Lending Club's risk assessment process and supporting more informed decisions regarding loan approval and pricing strategies. By combining data-driven predictive modeling with in-depth data analysis, this project aims to enhance the efficiency of Lending Club's loan operations and elevate its risk management capabilities, ultimately providing more dependable financial services to investors and borrowers.

Keywords: Credit Fraud; Random Forests; Artificial Neutral Networks.

1. Introduction

Credit default risk is one of the significant challenges faced by financial institutions in lending operations. When borrowers fail to repay loans on time or refuse to repay, financial institutions are exposed to potential financial losses. This risk has a significant impact on the profitability and stability of financial institutions. Therefore, accurately predicting credit default risk has become an urgent need for financial institutions in loan approval and risk management.

In the past, financial institutions mainly relied on traditional risk assessment methods, such as evaluation models based on credit scores and financial indicators, to assess borrowers' credit default risk. However, these methods often fail to comprehensively consider factors such as borrowers' personal circumstances, behavioral patterns, and market conditions, resulting in limited accuracy of prediction results.

With the rapid development of big data and machine learning technologies, an increasing body of research and practice has shown that utilizing large-scale data and complex models can improve the predictive ability of credit default risk. By analyzing borrowers' personal information, financial data, and behavioral data, combined with the application of machine learning algorithms, potential risk factors and patterns can be discovered, leading to more accurate predictions of borrowers' default probabilities.

In the current wave of financial technology, more and more fintech companies and financial institutions are adopting credit default risk prediction models based on big data and machine learning. These models can consider multiple factors comprehensively, such as personal information, credit history, financial conditions, and behavioral patterns, to more holistically evaluate borrowers' credit default risk. The application of this technology can not only help financial institutions improve the accuracy and efficiency of loan approvals but also reduce potential credit losses, promoting stability and sustainable development of financial markets.

Therefore, researching and developing credit default risk prediction models based on big data and machine learning have significant theoretical and practical value. By conducting in-depth studies on borrowers' characteristics, behavioral patterns, and market environments, combined with advanced data analysis techniques and machine learning algorithms, we can more accurately predict credit default risk, enhance financial institutions' risk management capabilities, and promote stability and sustainable development of financial markets.

2. Literature References

Research conducted on credit risk assessment has started abroad for a long time. Currently, foreign countries have developed a complete and comprehensive credit assessment system, which is widely used in banks and other financial institutions. The first researcher in this particular area was Fisher, who highlighted that the core aspect of forecasting individual credit scores is to categorize a specific customer population into various groups using diverse characteristic criteria [1]. David Durand also proposed the categorization of personal credit into favorable loans and unfavorable loans to segment customers, subsequently employing this approach for the assessment of individual credit. [2]. Currently, foreign research on credit assessment system achievements and data preprocessing techniques remains advanced and more mature.

Studying the credit scoring system model holds significant significance in today's society. Nowadays, there exists a wide array of credit scoring models, with the majority of mainstream models being developed using insights from statistics, operations research, or artificial intelligence. Among the various algorithms available, the most frequently employed ones include decision trees, logistic regression, random forests, discriminant analysis, neural networks, support vector machines, and genetic algorithms. Among these methods, the decision tree algorithm utilizes non-parametric statistical research techniques. Due to the heterogeneity of the data processed with prior information, the decision tree algorithm has the capability to substantially enhance and augment the efficiency and precision of data classification. Makowski was the first to use the decision tree algorithm in personal credit evaluation, and his fundamental idea revolved around partitioning diverse customer groups based on similar feature variables, aiming to minimize the disparity within similar customer groups, while having significant differences between different groups [3]. Logistic regression is similar in principle to linear regression; it just takes another form. Logistic regression uses the iterative method of maximum likelihood to find the estimates in the coefficients that are closest to the truth. This method's effect is more pronounced when the explanatory variables are qualitative indicators in problems encountered. Wiginton was the first to use this method in credit scoring research [4] and believed that this method does not restrict the distribution assumption of the explanatory variables when performing discriminant analysis. Durand was among the early adopters of discriminant analysis in credit analysis systems. In his research, he highlighted the key advantage of discriminant analysis, which lies in its efficacy in analyzing and predicting a company's ability to repay loans. Durand's work emphasized the utility of discriminant analysis as a valuable tool for assessing and forecasting loan repayment capacity. Westgaard successfully increased the discriminant analysis model's accuracy to over 90% [5]. Coats and Fant [6] conducted an empirical study to showcase the superiority of neural networks in terms of prediction accuracy in credit scoring. They developed a credit scoring model based on neural networks, leveraging real data from diverse countries. Their research aimed to substantiate the effectiveness of neural networks as a powerful methodology for credit prediction by utilizing actual data collected from various geographical locations; Odom and Sharda [7] highlighted the substantial benefits of neural networks when it comes to handling nonlinear problems.; Chen and Singh [8] conducted a study to assess the prediction accuracy of neural network models in credit prediction tasks, which can be widely promoted. Support Vector Machines (SVM) are relatively novel; they are a machine learning method based on statistical theory. SVMs are efficient and accurate in solving nonlinear problems comparing to other methods, local minimum points, and model selection. Today, SVMs are a focus and heated topic of research as well. SVM was

first applied to the credit scoring process by Baesens and Gestel, and according to the conclusion, the SVM algorithm's prediction accuracy is significantly superior to that of neural networks [9]. Genetic algorithms were first proposed by Holland as an optimization space search method. They are capable of both simulating biological reproduction processes and guiding the evolution of biological populations in a positive direction [10].

3. Methodology

3.1. Pearson Correlation Analysis

Pearson Correlation Analysis is used to measure the linear correlation between two variables.

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - Y)^2}} \quad (1)$$

For detail, X_i and Y_i are the values of each pair of borrowers' characteristics. $\bar{X}$ and $\bar{Y}$ are the respective means of X_i and Y_i.

The numerator represents the covariance between each pair of two characteristics.

The denominator represents the product of the standard deviations of the two variables.

3.2. Random Forests Brief Introduction

The structure of a decision tree is primarily divided into three parts: decision nodes, state nodes, and result nodes.

Firstly, decision nodes are used to select the most probable option among numerous possibilities. However, when a decision belongs to multiple levels, the final choice is determined by the decision node at the root of the decision tree, as there can be a large number of decision nodes.

On the other hand, the selection of state nodes involves comparing multiple alternative options using specific decision criteria to provide the optimal solution, aiming to achieve maximum economic benefits.

The profit or loss values marked on the right side of the result nodes represent the outcome of the decision tree. Decision trees are commonly used predictive models that involve constructing a decision tree tailored to the sample categories for predicting new samples. The core of this process is the binary separation of the predicted sample variables.

4. Experiment

4.1. Data Source and Comparative Advantages

This dataset originates from Kaggle [11], a famous machine learning community. It offers personal information from borrowers across 25 different aspects, providing a multidimensional depiction of their living conditions, shown as Table1. In comparison to other datasets, it is more closely aligned with the interactive information that can occur in real-world lending scenarios. Hence, it possesses a higher degree of authority.

4.2. Data Progressing

I converted categorical data, which is in string format, into numerical values using one-hot encoding. This method ensures that each categorical value is correctly transformed into a binary vector, where only one element is 1, representing the category, and the rest are 0s. This allows machine learning models to handle categorical data correctly without introducing erroneous numerical relationships.

We also noticed that the dataset contains some columns that are irrelevant or unnecessary for my analysis or modeling tasks. In order to reduce data dimensionality and improve model efficiency, I

removed these unused columns, which may include unique identifiers, timestamps, or other information unrelated to the analysis task.

Regarding missing values, I addressed them by taking appropriate measures. Typically, I filled missing values with means, medians, modes, or other suitable methods to ensure data integrity.

Table 1. Loan Borrower Attributes

No.	Loan Statistic	Elaboration on the Variables
1	Loan Status	Current status of the loan
2	Verification Status	Whether the income source has been verified by the lending platform, which can be either verified or unverified.
3	Term	The loan term（36 months or 60 months）, which is the period within which the borrower needs to repay the entire loan amount and interest.
4	Open Accounts	The number of currently active credit lines that the borrower has, typically referring to credit cards and other revolving credit products they hold.
5	Installment	The installment amount for each payment in the installment plan.
6	Grade	Representing the borrower's overall credit risk, typically denoted by a letter such as "A," "B," "C," "D," etc.
7	Sub Grade	Within each major credit grade, there are more detailed subgrades, often represented as combinations of numbers and letters, such as "A1," "A2," "B1," "B2," etc.
8	Loan amount	The total amount borrowed by the borrower this time.
9	Annual Inccome	The total annual income that the borrower provided during registration.
10	Public Record Bankruptcies	The number of times the individual has applied for bankruptcy protection as recorded in the borrower's credit report.
11	Initial Listing Status	In the lending marketplace or lending platform, the initial status of a borrower's loan when it is first listed and made available to potential investors is referred to as the "listing status."
12	Purpose	The information regarding the purposes for which the borrower will use the loan funds.
13	Title	The information regarding the purposes for which the borrower will use the loan funds.
14	Debt-to-Income Ratio	A ratio calculated based on the borrower's total monthly debt compared to their monthly income, used to assess the borrower's debt burden.
15	Earliest Credit Line	The time when the borrower's earliest credit line was opened
16	Interest rate	Interest Rate on the loan
17	Home Ownership	This variable represents the borrower's home ownership status, indicating whether the borrower owns their own home. Values include RENT, MORTGAGE, OWN, OTHER.
18	Employment Title	The borrower's job title provided during the loan application.
19	Employment Length	The variable indicating the borrower's years of work experience, typically used to describe the number of years the borrower has been employed by their current employer. Values vary from 0 to 10.
20	Revolving Balance	The total amount of outstanding debt that the borrower has across all credit card accounts.
21	Application Type	The type or nature of the loan application submitted by the borrower, which can be either an "individual loan" or a "joint loan."
22	Mortgage Accounts	Indicates the quantity of mortgage accounts opened by borrower.
23	Public Records	Negative events or defaults appeared in the borrower's credit history in the past. The record here is the count or number of occurrences.
24	Revolving Utilization	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
25	Total Number of Credit Lines	The ratio between the borrower's current credit card balance and their total credit card credit limit.
26	Issue Date	The time at which the loan was approved, specified down to the month.

4.3. Exploratory Data Analysis (EDA)

We start by conducting a Pearson analysis on the data, showcasing the relationships between various characteristics of the borrowers, shown as Figure1.

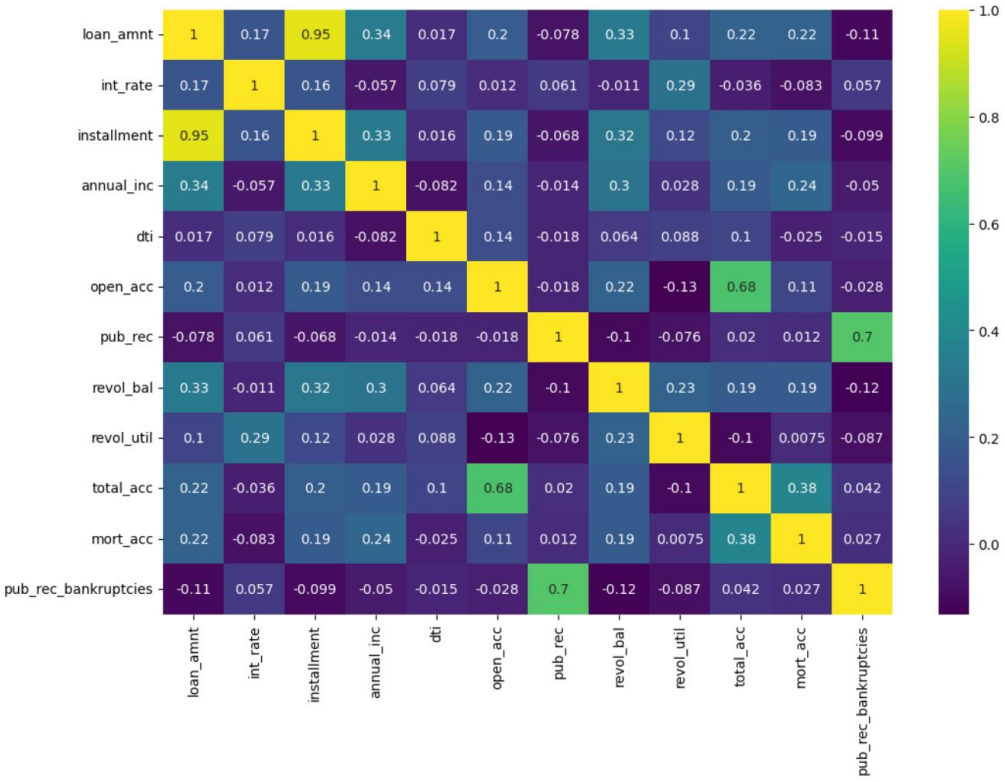
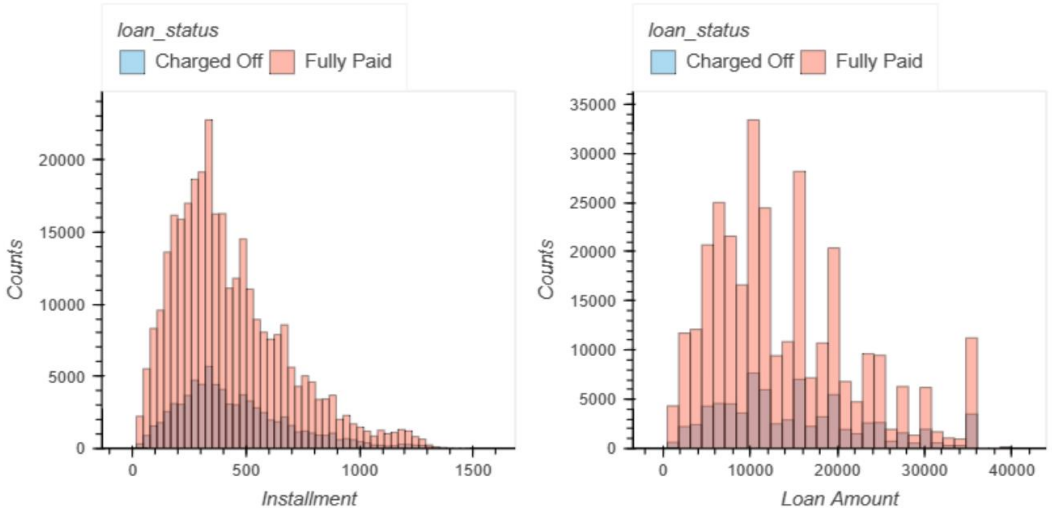


Figure 1. Pearson Analysis on Numeric Data

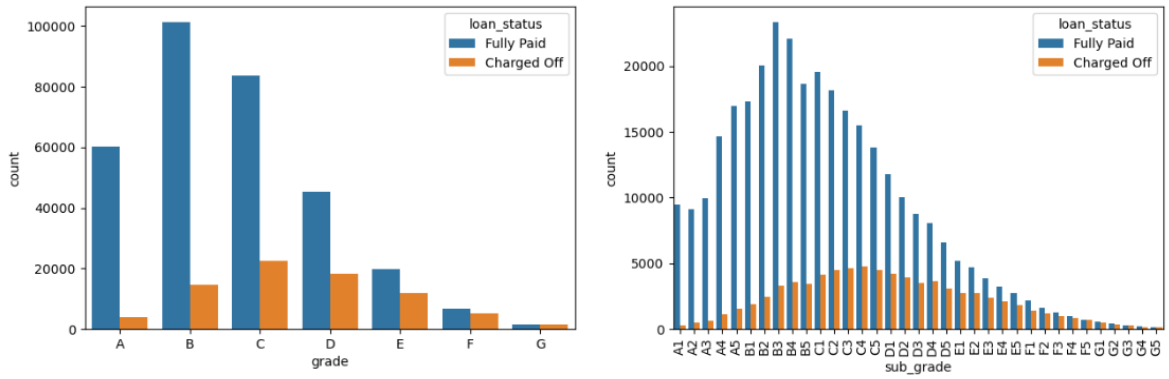
The "loan_amnt" feature and the "installment" feature have a almost perfect correlation, as I have observed. We will investigate these features further. Print out their descriptions and perform a scatterplot between them.



(a) (b)

Figure2. (a) Installment by Loan Status, (b) Loan Amount y Loan Status

Grade and subgrade help assess the borrower's repayment ability and credit status, making them crucial for lending institutions in determining loan approval, as well as loan rates and terms, shown as Figure 2.



(a) (b)

Figure 3. (a) Loan Status by Grade, (b)Loan Status by Sub_Grade

From the perspective of default rates, the lower the grade, the higher the number of defaults, indicating a greater likelihood of default, shown as Figure 3.

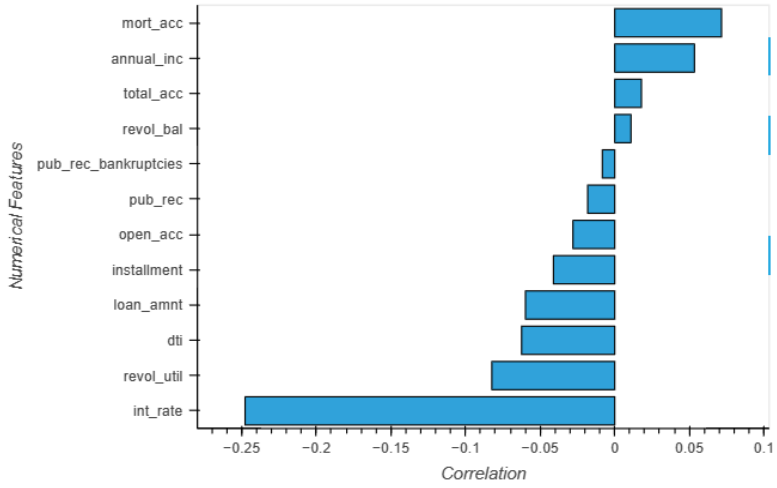


Figure 4. Correlation Between Loan Status and Numeric Features

In our analysis, we transformed the "loan status" variable, categorizing "Defaults" and Fully Paid" as 0 and 1, respectively, to determine the correlation with various features. From the results, it appears that "mort_acc," "total_acc," "annual_inc," and "revol_bal" exhibit positive correlations, while "revol_util," "int_rate," and "loan_amnt" show negative correlations. However, none of these correlations demonstrate an exceptionally strong relationship, shown in the Figure4. Therefore, it is essential to consider the borrower's overall repayment capacity comprehensively.

4.4. Model Training

4.4.1 ANNs Model

This model constructs an Artificial Neural Network (ANN) designed to address binary classification problems. The model's objective is to classify data by learning the features of the input data. The number of input features is determined by `num_columns`, while the output label is set to 1, indicating a binary classification problem. The model employs multiple hidden layers, and the number of neurons and dropout rates for each hidden layer can be customized to adapt to different datasets and tasks. Specifically, the model comprises three hidden layers, each consisting of 150 neurons and using the ReLU activation function to introduce non-linear relationships. The output layer utilizes the Sigmoid activation function, producing probability values between 0 and 1 for binary classification probability estimation.

During the model construction process, batch normalization is applied to expedite training and mitigate gradient vanishing issues. Additionally, the model incorporates Dropout layers to randomly deactivate a portion of neurons, enhancing generalization while reducing the risk of overfitting.

To train this ANN model, the model utilizes the Adam optimizer and selects binary cross-entropy as the loss function. The learning rate is controlled by the `learning_rate` parameter. Throughout the training process, the model divides the training data into small batches, each containing 32 samples, and conducts 20 training epochs. Furthermore, the model includes a validation set to monitor its performance on unseen data.

Lastly, the model's code includes several auxiliary functions for evaluating performance and visualizing training progress. The `print_score` and `evaluate_nn` functions are employed to display performance evaluations for both training and testing, including accuracy, classification reports, and confusion matrices. Additionally, the `plot_learning_evolution` function visualizes the evolution of loss and AUC metrics during training, aiding in monitoring convergence and performance improvement.

The construction and training process of this ANN model are valuable for addressing binary classification problems, allowing the model to adjust its structure and parameters based on different datasets and task requirements to achieve optimal performance and generalization.

4.4.2 Random Forests

I created a Random Forest Classifier object and set the number of estimators to 100. Then I trained the Random Forest Classifier using the training data along with their corresponding labels. The model constructs an ensemble of decision trees by learning from the features and labels of the training data in order to perform classification.

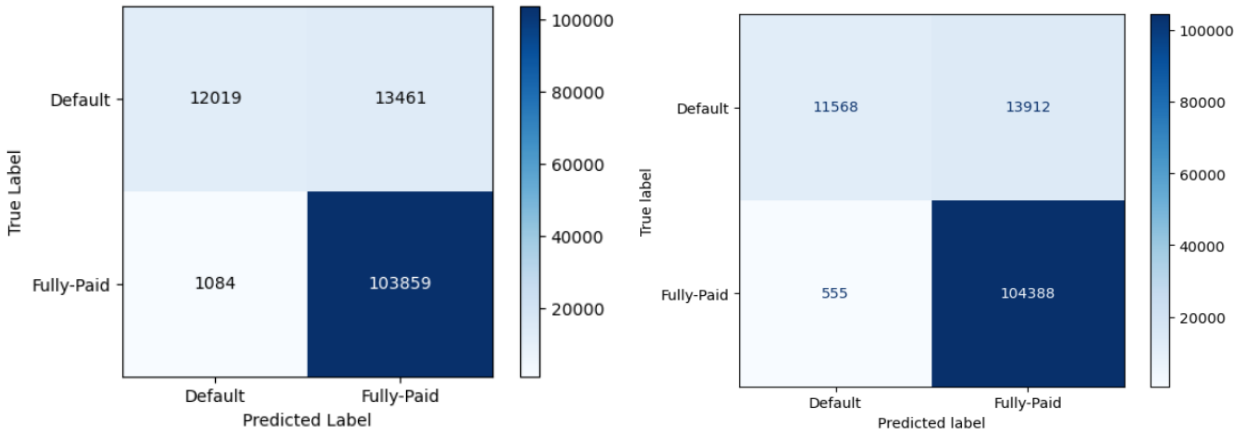
5. Result

From the classification report based on confusion matrix, shown in Tabel 2 and Figure 5, it can be observed that there isn't a significant difference in the performance of the Random Forest model and the ANN model. However, there are some distinctions to note. The ANN model exhibits better performance in terms of negative class recall, which is at 0.5, but it comes at a slightly lower precision. On the other hand, the Random Forest model achieves a better overall balance in its performance.

The ROC curve, shown in Figure 6, is a graphical representation employed to illustrate the classification model's performance across various thresholds. The X-axis of the ROC curve represents the False Positive Rate (FPR), while the Y-axis represents the True Positive Rate (TPR) or recall. A ROC curve positioned closer to the upper-left corner signifies superior model performance. AUC is the area under the ROC curve and is used to quantify the performance of a model. The AUC value ranges between 0 and 1, where a higher AUC value indicates better model performance. If AUC is equal to 0.5, the model's performance is equivalent to random guessing, while an AUC greater than 0.5 signifies that the model performs better than random guessing. Despite the excellent performance of the Random Forest model on the training set, achieving a 100% accuracy, its performance on the test set is only around 70%, far lower than the ANN model's 90%.

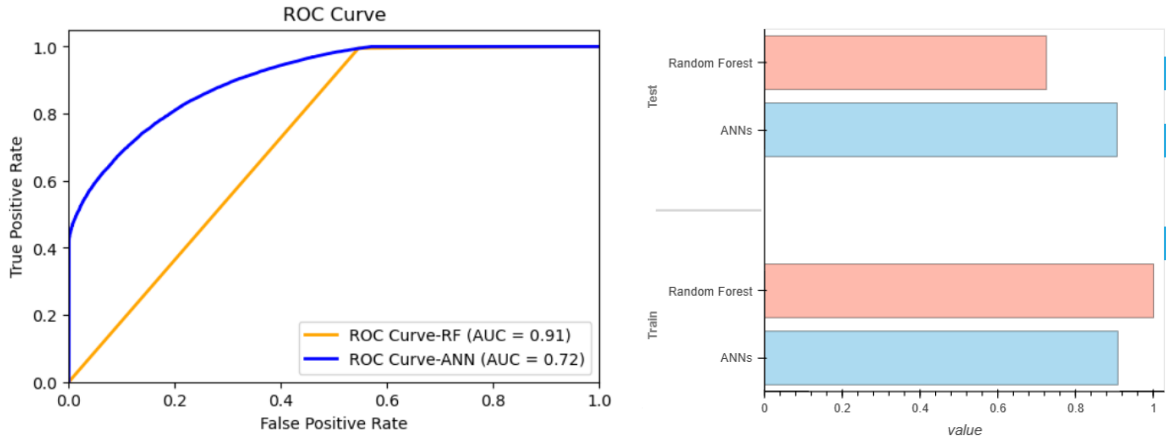
Tabel 2. Classification Report

CLASSIFICATION REPORT	Classifier	Default	Fully Paid	Accuracy	Macro Average	Weighted Average
Precision	ANN	0.88	0.89	0.89	0.88	0.89
	Random Forest	0.95	0.88		0.92	0.90
Recall	ANN	0.50	0.98		0.74	0.89
	Random Forest	0.45	0.99		0.72	0.89
F1-score	ANN	0.63	0.93		0.78	0.88
	Random Forest	0.62	0.94		0.78	0.87
Support	ANN	25480	104943		130423	130423
	Random Forest					



(a) (b)

Figure 5. Confusion Matrix (a) ANN Model (b) Random Forest



(a) (b)

Figure 6. ROC Evaluation (a) ROC-AUC Curves Comparison (b) ROC Scores of ML Models

6. Conclusion

Over the past few years, the banking industry's reliance on personal credit operations has been steadily increasing. Alongside the fast growth of the socio-economic landscape and the requirement for commercial transactions, financial institutions such as banks have faced the pressing challenge of improving their ability to predict individual credit defaults. This has led to the exploration of the application of machine learning-based predictive models in real-world financial transactions.

This project primarily focuses on key steps such as data preprocessing, feature engineering, and model selection. It utilizes algorithms like artificial neural networks and random forests to handle extensive borrower data. Additionally, by conducting feature importance analysis, the project can identify factors that significantly impact loan default risk, providing a more comprehensive risk assessment. This will help banks enhance loan operation efficiency and risk management capabilities, ultimately offering more reliable financial services to investors and borrowers.

In future work, there are several avenues for further optimization and improvement of machine learning models. The performance of machine learning algorithms often depends on parameter settings. Systematic parameter tuning can be employed to enhance model performance. Techniques such as cross-validation can help find the best parameter combinations, and methods like grid search or Bayesian optimization can be used for parameter adjustment to improve prediction accuracy and stability.

Dealing with imbalanced data, where default samples are often much less frequent than non-default ones, is a common challenge. Various approaches to address this issue can be explored, such as undersampling, oversampling, or class-balancing techniques within ensemble learning, to enhance

the model's ability to recognize default samples. Additionally, consider incorporating more data sources and feature engineering techniques to obtain comprehensive and accurate borrower information. Furthermore, integrating this model with other risk assessment methods can create a comprehensive risk management framework.

In summary, through ongoing research and practical implementation, the Credit Default Prediction project can continuously evolve and provide financial institutions with more reliable solutions for credit default risk prediction and risk management.

References

[1] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, 179-188.

[2] Durand, D. (1941). Risk elements in consumer installment financing. National Bureau of Economic Research, New York, 60-129.

[3] Makowski, P. (1985). Credit scoring branches out. Credit World, 75, 30-37.

[4] Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. The Journal of Financial and Quantitative Analysis, 15(3), 757-770.

[5] Westgaard, S., & Van der Wijst, N. (2001). Default probabilities in a corporate bank portfolio: A logistic model approach. European Journal of Operational Research, 135(2), 338-349.

[6] Coats, P. K., & Fant, L. F. (1993). Recognizing financial distress patterns using a neural network tool. Financial Management, 22(3), 142-155.

[7] Odom, M. D., & Sharda, R. (1990). A neural network model for bankruptcy prediction. IJCNN International Joint Conference on Neural Networks, 5, 163-168.

[8] Chen, C. C., Singh, J. P., Poland, W. B., et al. (1994). Parallel protein structure determination from uncertain data. Supercomputing 94: Proceedings. IEEE, 570-579.

[9] Van Gestel, I. T., Baesens, B., Garcia, I. J., et al. (2003). A support vector machine approach to credit scoring. Forum Financier Revue Bancaire Et Financiere Bank En Financiewezen, 6, 73-82.

[10] Holland, J. L. (1985). The self-directed search. Psychological Assessment Resources, 5, 11-45.

[11] https://www.kaggle.com/datasets/wordsforthewise/lending-club