## Loan default risk prediction and borrower profiling

## using machine learning models

## Summary

This study, based the credit risk dataset collected from Kaggle, aims to predict default risk by analyzing borrowers and profile borrowers to enhance the efficiency of lending decisions. Machine learning models, including CART, Random Forest, Logistic Regression, and XGBoost, are employed to classify and predict loan status. Additionally, K-means clustering is applied to analyze the characteristics of users with different loan statuses, helping to profile user segments.

Firstly, this study finds that XGBoost model beat Logistic Regression, CART and Random Forest in the field of classification accuracy evaluation. XGBoost model demonstrates superior performance in both the training and testing sets and achieves accuracy rates of 0.9461 and 0.8713, with F1-scores of 0.9460 and 0.8708, respectively. In contrast, CART and Random Forest models exhibit high precision in the training set, but their recall rates are relatively low. The Logistic Regression model performs the weakest across all metrics, particularly in recall, which is only 0.7679.

Secondly, in the feature importance analysis section, financial factors(including percentage of income represented by the loan amount and individual’s income) and low attributes(including grade assigned to loans, interest rate and intention of loan application) significantly impact the prediction of loan status across all models, while stability and background attributes(including type of home ownership,employment length and age), historical credit attributes(including length of credit history, historical default record) follow behind. And percentage of income represented by the loan amount is the most significant feature compared with other financial factors, especially in the XGBoost model.

Thirdly, the study conducts a clustering analysis of customer groups, categorizing them into four clusters ranked from high to low default risk as Cluster 1, Cluster 2, Cluster 3, and Cluster 4. The study finds that customer groups with higher income, lower loan-to-income ratios, and lower interest rates generally exhibit lower default rates.

Overall, the XGBoost model excels at identifying borrower default risk, and K-means clustering can identify the characteristics of different borrower groups. This paper provides empirical support for financial institutions in selecting risk models to predict default risk and in formulating risk management strategies tailored to different groups.

1. **Introduction**

In today's financial industry, managing loan default risk is a critical issue. Loan defaults not only directly affect the profitability of financial institutions but can also contribute to systemic financial risks. As global financial markets become increasingly complex and uncertain, effectively managing default risk has become a major challenge for financial institutions. A loan default can lead to substantial economic losses and undermine trust and stability in the financial market. Therefore, the timely and accurate prediction of borrowers' default risk is essential for reducing bad debt, optimizing credit decisions, and maintaining the healthy operation of the financial system.

However, managing default risk presents several challenges. First, borrowers' personal characteristics and financial conditions are highly complex and multidimensional, making them difficult for traditional evaluation methods to handle effectively. Relying solely on manual judgment or simple statistical models cannot manage large-scale, diverse data or capture potential nonlinear relationships. Second, as data volumes increase, extracting meaningful insights and building robust prediction models with strong generalization abilities become significant hurdles for financial institutions in risk management. Furthermore, borrower default behavior is often influenced by external economic conditions and market fluctuations, further complicating risk prediction. As a result, traditional judgment-based and rule-driven approaches are no longer sufficient to meet the demands of modern credit risk management. In light of these challenges, this study aims to introduce advanced machine learning techniques to address these issues.

The objectives of this study are twofold: first, to identify the most accurate machine learning model for predicting default behavior by classifying borrowers' default risk; and second, to analyze differences among borrower groups. To achieve these goals, a variety of machine learning algorithms were employed, including decision trees, random forests, Logistic regression, and XGBoost for loan status classification. In addition to predictive modeling, this study utilizes k-means clustering to analyze user characteristics based on loan status, revealing potential high-risk borrower groups.

The motivation for this study lies in providing a more comprehensive risk assessment framework by machine learning models to meet the business needs of financial institutions. This study predicts default risk and provides financial institutions with multi-layered decision-making support. Furthermore, the use of cluster analysis to profile borrower groups offers valuable insights into the significant differences in default risk, enabling better risk management and market positioning strategies.

In summary, this study distinguishes itself by combining multiple machine learning techniques to address a key issue faced by financial institutions: how to accurately predict borrowers' default risk. This study aims to provide financial institutions with a more effective risk assessment tool and support future risk management strategies with data-driven insights.

1. **Literature Review**

In recent years, with the continuous development of the financial market and the increasing application of big data technology, credit risk management has become a major challenge for banks and financial institutions. Traditional statistical models have been widely used in credit default prediction. However, with the increase in data complexity and size, machine learning models have been gradually introduced and their application has improved the accuracy of credit risk prediction. It proved that AI methods are superior to statistical methods in dealing with corporate credit risk evaluation problems(Ghodselahi et al.,2011)).The following section provides an overview of the previous uses, challenges, and prospects of machine learning in the field of credit risk prediction.

**2.1 Application of Machine Learning Models in Credit Default Prediction**

Both supervised and unsupervised learning are widely used in the field of credit risk assessment. Supervised learning is one of the most commonly used methods, which relies on labeled data and makes predictions by learning known input-output relationships. A study by Khandani et al. (2010) combines consumer transaction data and credit bureau data to develop nonlinear predictive models using a machine learning model, which significantly improves accuracy in credit card default prediction.

While unsupervised learning is mainly used to explore the underlying structures and patterns in data. In credit risk management, it is well suited for data clustering and anomaly detection. Wang et al. (2019) proposed a credit risk assessment method that combines SOM with supervised learning, and it significantly outperformed a single model in credit scoring.

In some applications, we also combine multiple machine learning models to improve the accuracy and robustness. Ghodselahi et al. (2011) constructed an integrated model based on support vector machines (SVMs), neural network models, and decision trees, demonstrating the classification accuracy of the model. Plawiak et al. (2019) proposed an integrated learning model based on genetic algorithm optimization that performs well in handling complex credit scoring tasks.

**2.2 Challenges in the use of machine learning**

Although machine learning has demonstrated significant advantages in credit default prediction, it still faces some challenges. Shi et al. (2022) identified data imbalance, dataset consistency, model transparency, and under-application of deep learning models as the main challenges in the current field .

In credit default risk prediction, the impact of data quality and feature selection on model performance is critical and can both contribute to enhancing the performance of supervised models in accurately predicting credit risk, which could be advantageous for the lending industry.( Wattanakitrungroj et al. ,2024)

**2.3 Data-processing issues**

Data imbalance is one of the common challenges in credit default prediction. In credit risk data, usually only a small portion of customers will default while most of them pay on time, which will lead the model to prefer predicting non-defaulting customers. To deal with this problem, Wattanakitrungroj et al. (2024) proposed the use of sampling strategies (e.g., undersampling and oversampling to balance the data. By redistributing the data during model training, it is possible to improve the model's prediction performance for defaulting customers. In addition,SMOTE(Niu et al.,2020) is one of the most widely used approach to address this problem.

**2.4 Feature selection and feature engineering**

Feature selection is an important step in model optimization, by selecting the most representative and predictive features, noise can be reduced and the generalization ability of the model can be improved.A study by Plawiak et al. (2019) indicated that using genetic algorithms for feature selection can further optimize the model performance. Wang et al. (2019) used SOM to discover hidden patterns in the data during the unsupervised learning phase and apply these patterns to optimize feature selection.

Future research directions should focus on addressing these challenges while further exploring the combination of integrated learning and deep learning to improve the accuracy and interpretability of credit default prediction.

In summary, although the combination of machine learning, big data, and multiple advanced algorithms has significantly improved the accuracy and efficiency of credit risk assessment in past studies, challenges still exist. Therefore, our research will apply multiple machine learning models ,improving the performance and interpretability of credit risk prediction through comparisons and to provide more reliable decision support for financial institutions.

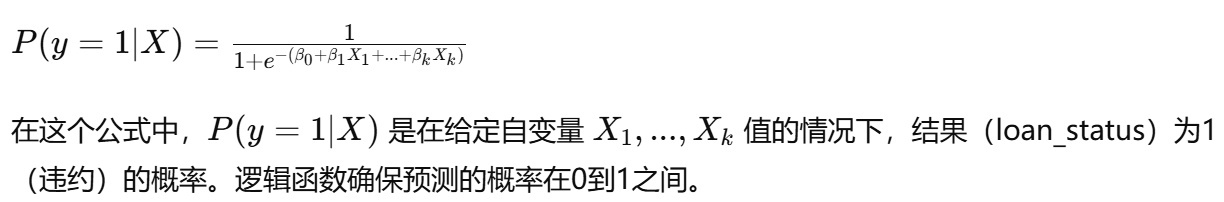
1. **Methodology**

**3.1 Data source**

The Credit Risk Dataset comes from Kaggle, which contains columns simulating credit bureau data. It contains 11 features used to predict whether a borrower will default, including personal information such as age, income, homeownership, length of employment, loan intent, whether there is a history of default, as well as the approved loan grade, loan amount, and loan interest rate. Among the 11 features, 4 are categorical variables, and 7 are continuous variables.

**3.2 Model selection**

**3.2.1 Logistic regression**

Logistic regression is a powerful tool for classification tasks, particularly suited for predicting binary outcomes such as loan status (default or non-default).Unlike linear regression, which predicts continuous values, logistic regression predicts the probability of an event occurring, expressed through a mathematical formula：

Logistic regression is highly interpretable and can predict the conditional probability of a sample belonging to a certain category. Therefore, logistic regression is well-suited for credit risk prediction, as it not only provides a classification (default or non-default) but also estimates the probability of default, helping financial institutions to reasonably assess borrower risk.

**3.2.2 Decision tree**

Decision tree is a common supervised learning method for classification and regression problems. It helps the user to make decisions by segmenting the data through a tree structure. The basic steps are: first, the best features in the current dataset are selected for segmentation by some criterion (e.g., Gini index, mean square error). Then, divide the dataset into subsets based on different values of the selected features, after that, repeat previous steps for each subset until the stopping condition is satisfied. Finally, the value of each leaf node indicates the predicted category or value. In our research, we use a decision tree that traverses downward from the root node until it reaches a leaf node to obtain the predicted status (default or non-default) of a new borrower based on their features.

**3.2.3 Random Forest**

Random Forest is an integrated learning method widely used for classification and regression tasks. It consists of multiple decision trees, and combines the predictions of each tree by voting or averaging to improve the accuracy and robustness of the model. Its basic principle is to reduce the risk of overfitting by introducing randomness. In our research, based on the features of new borrowers, the random forest model generates predictions about whether they will default or not.

**3.2.4 XGboost**

XGBoost (Extreme Gradient Boosting) is an efficient and powerful integrated learning algorithm that belongs to an extension of GBM, which employs a combination of multiple decision trees for prediction, and enhances the performance of the model by incrementally adding new trees to correct the errors of the previous trees. In credit loan prediction, the XGBoost model is trained with inputs from a series of features such as borrower-related historical data, which is ultimately used to predict both the loan status and the loan amount of a new borrower.

**3.2.5 K-means**

K-means is a widely used clustering algorithm that belongs to unsupervised learning. Its core objective is to divide the dataset into K clusters so that the data points within the same cluster have high similarity. The basic process of K-means clustering is as follows: Randomly select k cluster centers and assign each data point to its nearest cluster center,after that, update the cluster center to the center-of-mass position of all the current clusters, and then reassign each data point to the nearest new cluster, finally repeat the updating and reassignment steps until no data point undergoes a position change. In the application in our research, K-means is able to analyze the borrower's characteristic information and help to construct user profiles, so as to understand the behavioral patterns and risk characteristics of the customer groups more deeply.

**3.3 Data cleaning and descriptive analysis**

During the data cleaning process, this chapter focuses on handling missing values, duplicate entries, and outliers, along with a comprehensive descriptive analysis. Firstly, this study carries out a preview of structure of the dataset. As shown in Table 3.1, the dataset comprises 12 variables, of which 4 are categorical and 8 are continuous.

|  |  |  |
| --- | --- | --- |
| Variable | Label Encoding | Explanation |
| Categorical Variables |  |  |
| loan\_status | Non-default:0, Default:1 | Loan status, where 0 indicates non-default and 1 indicates default. |
| person\_home\_ownership | Mortgage:0, Other:1, Own:2, Rent:3 | Type of home ownership of the individual. |
| loan\_intent | Debt.:0, Edu.:1, Homei.:2, Medical:3, Personal:4, Venture:5 | Intent behind the loan application. |
| loan\_grade | A:0, B:1, C:2, D:3, E:4, F:5, G:6 | The grade assigned to the loan based on creditworthiness of the borrower. |
| cb\_person\_default\_on\_file | N:0, Y:1 | Historical default of the individual as per credit bureau records. |
| Numerical Variables |  |  |
| person\_age |  | Age of the individual applying for the loan. |
| person\_income |  | Annual income of the individual. |
| Person\_emp\_length |  | Employment length of the individual in years. |
| loan\_amnt |  | The loan amount requested by the individual. |
| loan\_int\_rate |  | The interest rate associated with the loan. |
| loan\_percent\_income |  | The percentage of income represented by the loan amount. |
| cb\_preson\_cred\_hist\_length |  | The length of credit history for the individual. |

Table 3.1 Variable Defination and Label Encoding

**3.3.1 Data cleaning**

In preparation for analysis, the loan\_status variable is recoded as a categorical variable, with 1 as "Default" and 0 as "Non-default." Additionally, categories in loan\_intent—"DEBTCONSOLIDATION," "EDUCATION," and "HOMEIMPROVEMENT"—are shortened to "DEBT," "EDU," and "HOMEI" for clarity in visualizations. These preprocessing steps are undertaken to streamline the dataset and facilitate a more interpretable analysis.

Secondly,this paper addresses missing values and duplicates in the dataset, which originally contained 165 duplicate entries and 3,982 missing values, including 887 in \*\*person\_emp\_length\*\* and 3,095 in \*\*loan\_int\_rate\*\*. A direct deletion method is applied to resolve these issues.

The table below summarizes the dataset after removing missing values.However, something abnormal still exists. However, outliers remain, as seen in Figure 3.3, where person\_age and person\_emp\_length have maximum values of 144 and 123, respectively. Entries with person\_age over 80 and person\_emp\_length over 40 are classified as outliers and removed. After processing, the dataset consists of 28,493 samples, including 22,306 non-default samples and 6,187 default samples.

Figure 3.1 Dataset Summary

Figures 3.2 and 3.3 depicts the distribution of loan\_status across individual variables, while Table 3.2 summarizes the proportions of Default and Non-default across categorical variables. The stacked bar chart highlights the loan default risk across various demographic groups. Specifically, Figure 3.2(a) shows that renters make up the largest proportion of loan applicants, with the highest default rate of 31.27%. Figure 3.2(b) shows that the highest proportion of loans is allocated for education, medical expenses, and venture capital, with the lowest default rates among those pursuing venture capital and education. Figure 3.2(c) reveals that most loans are concentrated in grades A and B, indicating that individuals with higher ratings are more likely to secure loans, which underscores the effectiveness of loan grading in reducing credit risk. Figure 3.2(d) shows that individuals without prior defaults are more likely to obtain loans, with a lower default rate of 18.21% compared to 37.83% for those with past defaults.

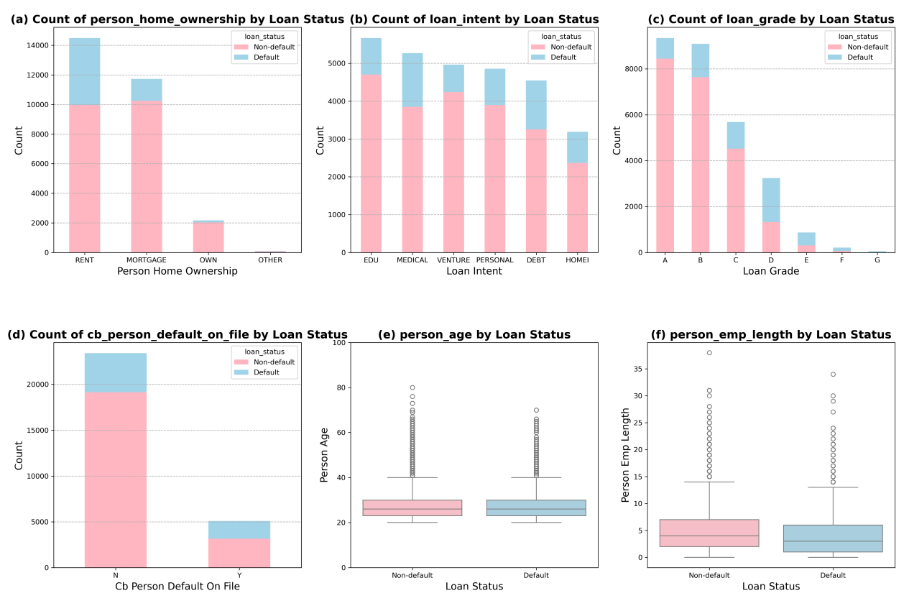
The box plots in Figures 3.2 and 3.3 show the distribution of continuous variables by loan default status. The distribution of borrowers' ages and credit history length are largely consistent across the two default status groups, suggesting these may not be significant in default risk assessment. Notably, Figure 3.2(f) demonstrates that individuals with shorter employment histories are more prone to default. Figure 3.3(a) displays income distribution by loan status, where the log-transformed income (Figure 3.3(b)) shows a more uniform distribution with fewer outliers, and the median income is slightly higher for the non-default group.As is shown in Figure 3.3(c) , the default group has a slightly higher median loan amount and a broader range. Figure 3.3(d) reveals that the default group faces higher interest rates, with more extreme outliers, while Figure 3.3(e) indicates a higher loan-to-income ratio for default applicants.

Figure 3.2 Distribution of Variables Across Different Loan Statuses 1

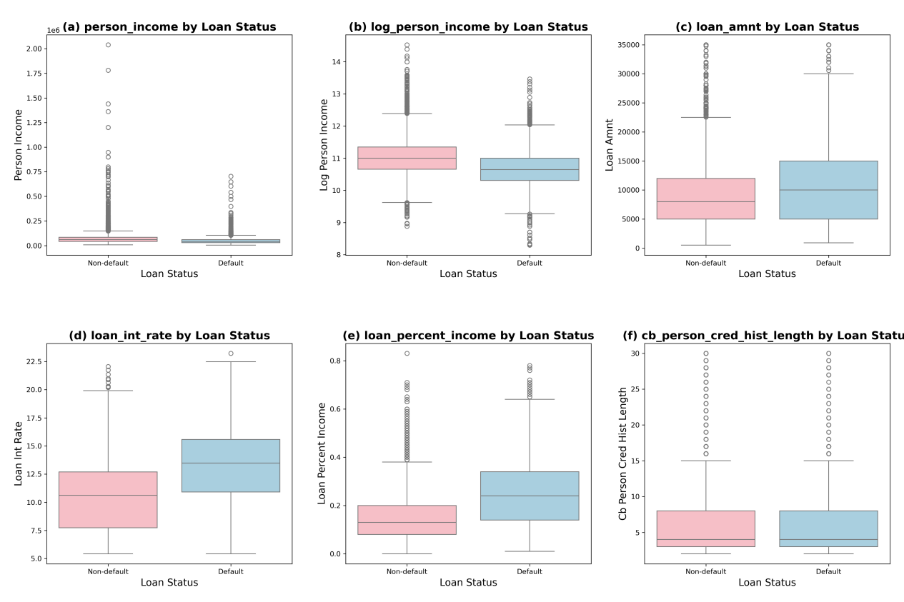


Figure 3.3 Distribution of Variables Across Different Loan Statuses 2

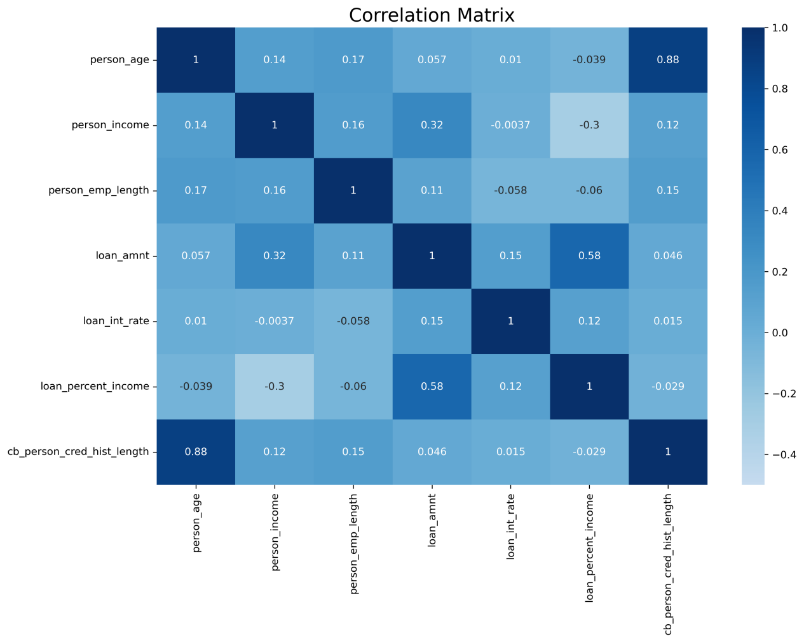
Table 3.2: Proportions of Default and Non-default across Categorical Variables

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Default | Non-default |
| person\_home\_ownership | RENT | 31.27% | 68.73% |
| MORTGAGE | 12.64% | 87.36% |
| OWN | 6.67% | 93.33% |
| OTHER | 29.03% | 70.97% |
| loan\_intent | EDU | 17.06% | 82.94% |
| MEDICAL | 26.92% | 73.08% |
| VENTURE | 14.66% | 85.34% |
| PERSONAL | 19.77% | 80.23% |
| DEBT | 28.46% | 71.54% |
| HOMEI | 25.73% | 74.27% |
| loan\_grade | A | 9.61% | 90.39% |
| B | 15.93% | 84.07% |
| C | 20.33% | 79.67% |
| D | 59.22% | 40.78% |
| E | 64.67% | 35.33% |
| F | 69.86% | 30.14% |
| G | 98.31% | 1.69% |
| cb\_person\_default\_on\_file | N | 18.21% | 81.79% |
| Y | 37.83% | 62.17% |

In summary, default applicants typically exhibit the following characteristics: Renters, engaged in debt consolidation, lower loan grades, history of defaults, lower income, higher loan amounts, higher interest rates, larger proportion of income allocated to loan repayments.

**3.3.3 Correlation Analysis**

After excluding log\_person\_income, a correlation analysis of continuous variables is conducted using a heatmap. Figure 3.4 reveals a strong correlation(0.88) between the borrower's age and their credit history length. Additionally, the correlation coefficient between the loan amount and income percentage allocated to the loan is 0.58, while the correlation between the loan amount and the borrower's income is 0.32. These findings indicate that borrowers with higher income levels are more likely to receive approval for larger loan amounts from financial institutions.

Feature 3.4 Correlation Heatmap

1. **Feature Engineering**

To address class imbalance and convert categorical variables into a suitable format for modeling, this chapter implemented a data preprocessing step.

**4.1 Random Undersampling and Label Encoding**

The dataset is initially divided into feature variables and the target variable, loan status (loan\_status), categorized as "Default" and "Non-default." To address the imbalance in the target variable, random undersampling is applied to reduce the majority class samples, aligning them more closely with the minority class and balancing class proportions.

Label encoding is applied to categorical features, including person\_home\_ownership, loan\_intent, loan\_grade, and cb\_person\_default\_on\_file. This process transforms these categorical variables into numerical formats compatible with machine learning algorithms. The target variable loan\_status is also encoded to ensure that "1" denotes default and "0" indicates non-default, maintaining consistency in model training and evaluation.As shown in Table 3.1 ,the mapping of all encoded variables ensures traceability of the original categorical information.

**4.2 Data Preprocessing**

To facilitate effective model training and evaluation, the encoded dataset is split into training and testing sets, with the latter comprising 30% of the total data.

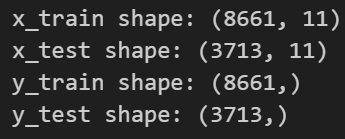
Subsequently, feature variables and the target variable are extracted from each set, where 'loan\_status' served as the target variable. Finally, the shapes of the feature and label sets for both the training and testing groups are shown in Figure 4.1.

Figure 4.1 Results of Train-Test Split

1. **Machine Learning Model**

**5.1 Loan default risk prediction**

* + 1. **logistic regression**

**5.1.1.1 logistic regression Model Establishment and Training**

Table 5.1. Range of logistic regression parameters in Grid Search

|  |  |
| --- | --- |
| **Parameter** | **Grid Search** |
| logspace | (-4, 4, 20) |
| penalty | (l1, l2) |

Table 5.1 delineates the parameters and their respective ranges used in the grid search for optimizing the logistic regression model. The logspace parameter is varied from -4 to 4 with 20 intervals to identify the best regularization strength, while the penalty parameter tests both l1 (Lasso regression) and l2 (Ridge regression) to ascertain the most effective method for reducing overfitting and improving model prediction accuracy.

The logspace parameter in table 5.1 of logistic regression commonly refers to the range of values for the regularization strength parameter, often denoted as C in logistic regression formulations. The regularization strength inversely controls the influence of the regularization term, which helps in preventing overfitting by penalizing larger coefficients in the model.The values are typically distributed on a logarithmic scale, which in this case ranges from 10−410^{-4}10−4 to 10410^4104, spaced into 20 intervals. This broad range allows for extensive testing to determine the optimal level of regularization needed to achieve the best generalization performance outside of the training dataset.

The penalty parameter specifies the type of regularization used in the logistic regression model. Regularization is a technique used to reduce the model complexity and prevent overfitting which might result from simple linear regression models. l1 (Lasso regularization) type of regularization adds an absolute value penalty to the model coefficients, which can lead to zero coefficients for some variables, effectively performing variable selection. It is particularly useful when we suspect that only a subset of predictors are actually informative.l2 (Ridge regularization) adds a squared magnitude penalty to the model coefficients. Unlike l1, it does not reduce coefficients to zero but distributes the coefficient shrinkage more evenly across all variables. It is useful when many variables contribute small effects towards the response variable.

**5.1.1.2 Model evaluation**

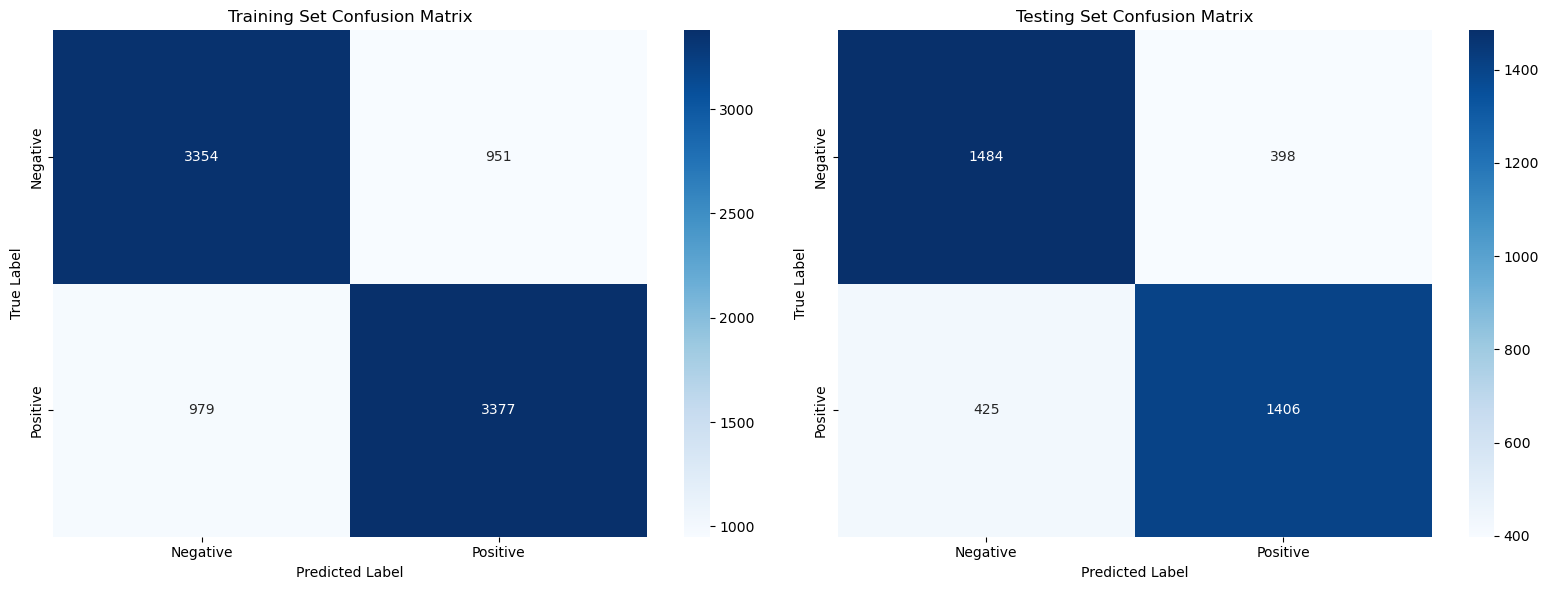
The evaluation of our logistic regression model is depicted using the confusion matrix for both the training and testing sets, as illustrated in Figure 1. Based on these matrices, we calculated several performance metrics which are summarized in Table 2, including accuracy, precision, recall, F1 score, and ROC AUC.

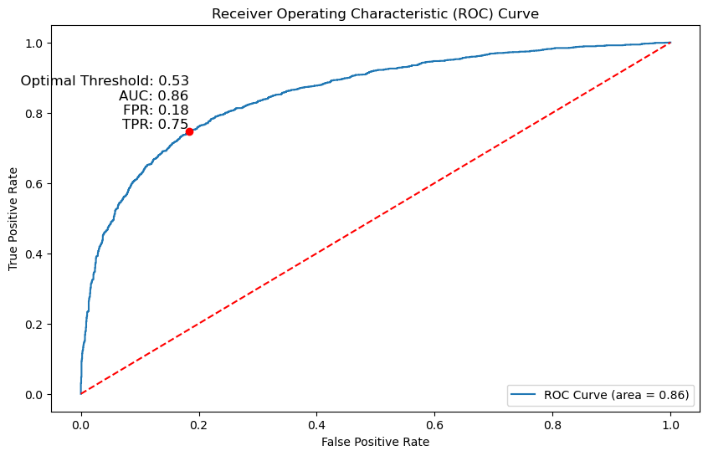
Figure 5.1: The confusion matrix of the logistic regression model

The confusion matrices demonstrate the model's ability to classify the positive (default) and negative (non-default) cases effectively. From the matrices, we observed a relatively balanced performance in terms of positive and negative classifications, with both training and testing sets showing similar patterns of true positives and true negatives.

Table 5.2. Empirical Results of the Logistic Regression Model

| **Metrics** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **ROC AUC** |
| --- | --- | --- | --- | --- | --- |
| **Training Set** | 77.72% | 78.03% | 77.53% | 77.78% | 77.72% |
| **Testing Set** | 77.83% | 77.94% | 76.79% | 77.36% | 77.82% |

As detailed in Table 5.2, the logistic regression model achieves comparable accuracies of 77.72% for the training set and 77.83% for the testing set, reflecting its effectiveness across both datasets. The precision, recall, and F1 scores consistently underscore the model's robustness in balancing both the accuracy of predictions and the rate at which positive cases are correctly identified.

Figure 5.2

Additionally, the Receiver Operating Characteristic (ROC) curve for the test dataset, as shown in Figure 5.2, further establishes the model’s capability. The AUC score of 0.86, combined with an optimal threshold of 0.53, highlights the model's strong discriminative power in distinguishing between 'Default' and 'Non-default' cases.

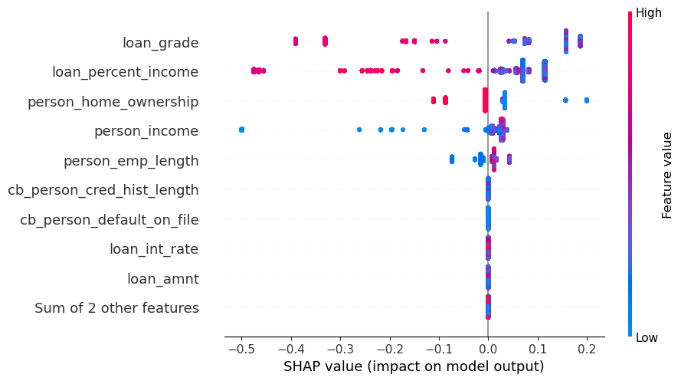
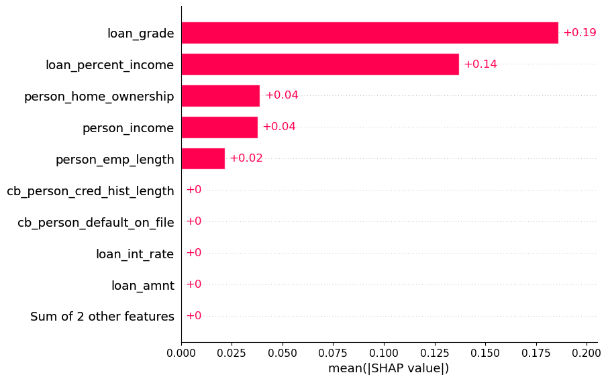
**5.1.1.3 Feature Importance Analysis**

To understand the impact of different customer attributes on the credit risk classification of loan statuses, we utilized SHAP values to conduct a detailed feature importance analysis. Figure 5.3(a) provides an exhaustive overview of feature importance pertaining to loan status, computed as the mean of the absolute Shapley values across the dataset. This visual representation clearly delineates the most influential features. Meanwhile, Figure 5.3(b) offers a SHAP summary plot that illustrates the distribution of SHAP values for each feature, indicating their respective impact on model output.

(1) **Financial Factors:** Attributes such as loan\_percent\_income and person\_income are crucial in evaluating default risk. These factors underscore the significance of a borrower's financial capacity relative to the loan amount. Notably, higher loan\_percent\_income ratios and lower personal income levels are commonly linked to increased default risks. Conversely, the total loan amount (loan\_amnt) tends to have a comparatively minor influence on default probability.

(2) **Loan Attributes:** The loan\_grade, loan\_int\_rate, and loan\_intent are pivotal attributes. The loan\_grade is particularly critical as it reflects the borrower's creditworthiness and substantially affects the likelihood of repayment. The SHAP values indicate that loans intended for purposes like debt consolidation carry a higher risk, and elevated interest rates are typically correlated with increased default risks.

(3) **Stability and Background Attributes:** Variables such as person\_home\_ownership, person\_emp\_length, and person\_age also play significant roles. The model suggests that the stability provided by homeownership (person\_home\_ownership) influences credit risk assessments, with renters showing a higher propensity towards defaulting. Moreover, employment duration (person\_emp\_length) appears to be a more decisive factor than age, suggesting that shorter employment periods may enhance default risks.

(4) **Historical Credit Attributes:** The history of credit length (cb\_person\_cred\_hist\_length) and defaults (cb\_person\_default\_on\_file) are among the less impactful attributes. While they contribute to the overall risk assessment, their effects are relatively minor compared to the ****previously discussed factors.

**Figures** 5.3 : Global interpretations of logistic regression model by SHAP values: (a) SHAP Feature Importance (b) SHAP Summary Plot

These insights derived from SHAP value analysis are instrumental in refining the predictive model by emphasizing significant predictors of loan default, thus enhancing the accuracy and reliability of credit risk evaluations.

**5.1.2 CART**

**5.1.2.1 Model Training**

This chapter utilizes a decision tree classification model to enhance the accuracy of loan default predictions. To systematically tune the hyperparameters, GridSearchCV is utilized for cross-validation to identify the optimal parameter configuration, with value ranges listed in Table 5.3. The final model features a maximum depth of 6, a maximum of 7 features, a minimum leaf node count of 2, and a minimum sample split count of 6.

Once the model is trained, predictions are generated for both the training and testing sets. Model performancel is assessed using key metrics, including accuracy, precision, recall, F1 score, and the area under the ROC curve.

Table 5.3 Range of DecisionTree Parameters in GridSearch

|  |  |
| --- | --- |
| Parameter | GridSearch |
| max-depth | (1,30,5) |
| min\_samples\_split | (2, 10, 2) |
| max\_features | (1, x\_train.shape[1] + 1) |
| min\_samples\_leaf | (2, 10, 2) |

**5.1.2.2 Model evaluation**

The confusion matrices for the training and testing datasets based on the proposed Decision Tree model are presented in Figure 5.4. Performance metrics are summarized in Table 5.4.

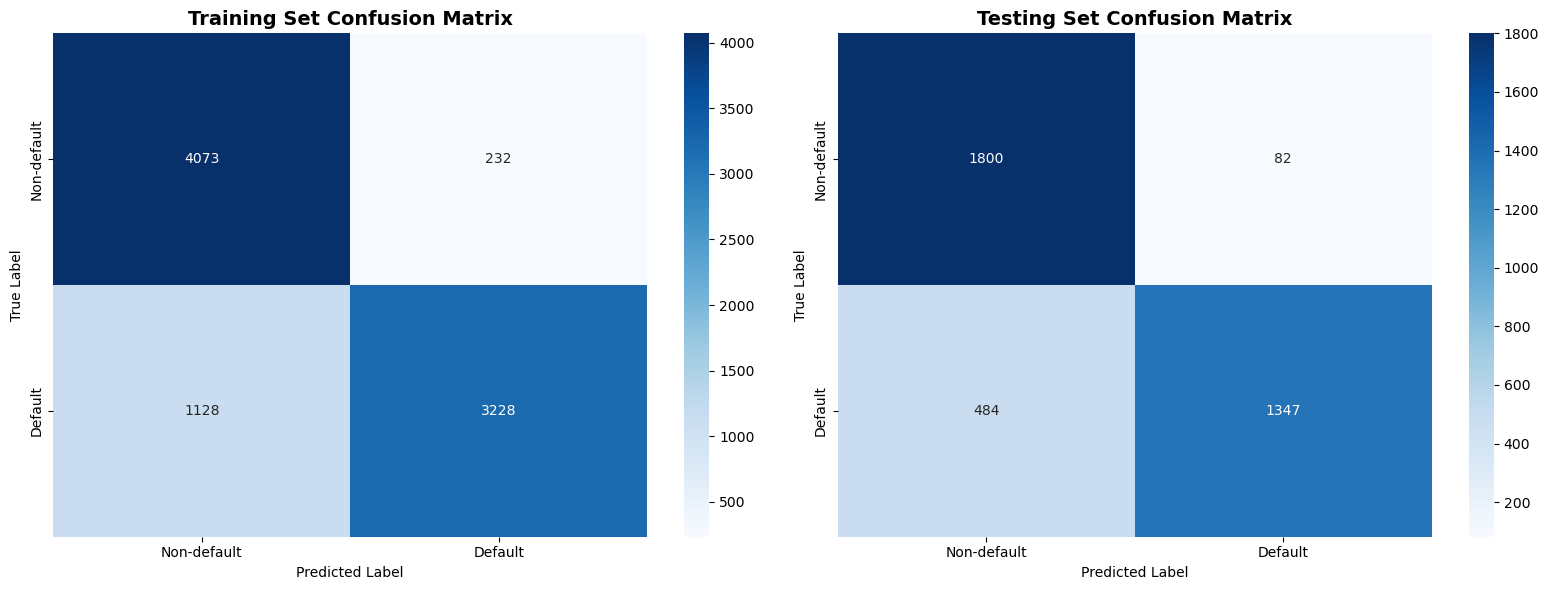


Figure 5.4. The confusion matrix of the Decision Tree model

According to the table, the model achieved training and testing accuracies of approximately 84%, with high precision values exceeding. Recall values were approximately 74% for both datasets, resulting in consistent F1 scores around 82%. These results collectively indicate the model's effectiveness and reliability in classifying loan\_status.

Table 5.4 Empirical results of DecisionTree classification algorithm

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| Train dataset | 0.8430 | 0.9329 | 0.7410 | 0.8260 |
| Test dataset | 0.8476 | 0.9426 | 0.7357 | 0.8264 |

Additionally, the ROC curve for the test dataset, shown in Figure 5.5, reveals an AUC of 0.90, indicating strong performance in distinguishing between default and non-default loans. With a false positive rate (FPR) of 0.04, only 4% of non-default loans are misclassified as defaults, highlighting the model's effectiveness in reducing misclassification. The true positive rate (TPR) of 0.74 shows that the model accurately identifies 74% of default loans. The optimal threshold is set at 0.90, reflecting a reasonable balance between false positive and false negative rates.

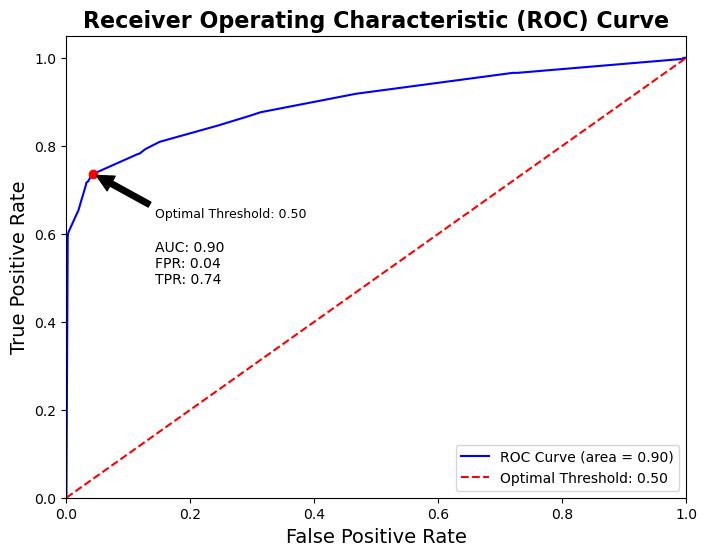
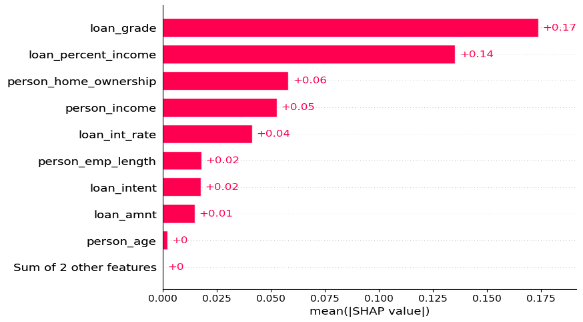
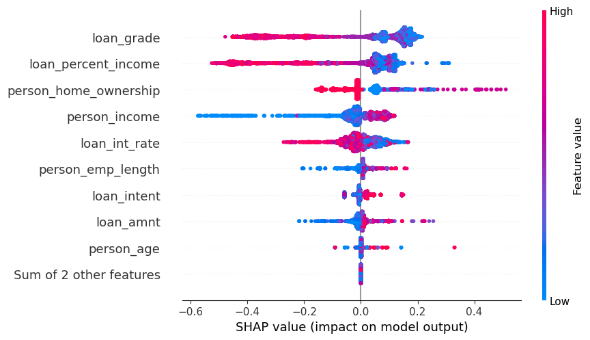
 In summary, the model performs well in terms of AUC and false positive rate and shows a decent effectiveness in identifying default loans. However, for practical applications, optimizing recall is recommended, to ensure its reliability in predicting loan defaults.

Figure 5.5 The ROC curve of the test data

**5.1.2.3 Feature importance analysis**

To analyze the impact of customer attributes on credit risk classification, SHAP values were employed for feature importance analysis. Figure 5.6(a) shows the average absolute Shapley values, highlighting the most significant features related to loan status. Additionally, Figure 5.6(b) presents a SHAP summary plot, illustrating the distribution of SHAP values for each feature and their corresponding influence trends.

The analysis of the figure below categorizes the variables influencing loan status into three main groups. Financial factors include person\_income and loan\_percent\_income. Loan attributes consist of loan\_grade and loan\_int\_rate. In terms of borrower background, person\_home\_ownership indicates that renters face significantly higher default risks compared to homeowners, while a shorter person\_emp\_length is also associated with an increased risk of default. However, the bee warm plot indicates that the Decision Tree model's predictions for certain groups do not totally align with the visualization results, particularly for variables like person\_income and loan\_amnt. This discrepancy may be due to the Decision Tree's low recall rate, which limits its capacity to effectively identify the defaulting population.

In conclusion, the decision tree model effectively assesses credit risk, primarily relying on loan\_grade and loan\_percent\_income. To enhance model performance, advanced modeling techniques are explored in the following chapters, such as random forests or XGboost, in order to improve accuracy by leveraging complex interactions among features.

1. (b)

Figure 5.6 Global interpretations of Decision Tree model by SHAP values: (a) SHAP feature importance (b) SHAP summary plot.

**5.1.3** **Random Forest**

The Random Forest model is utilized to predict a binary outcome, likely a classification task related to credit scoring or loan default prediction. The model is trained and evaluated on a dataset, and various metrics are calculated to assess its performance.

**5.1.3.1 Random Forest Model Establishment and Training**

In establishing and training the Random Forest model, a systematic approach was adopted to optimize the model parameters using a grid search method. This process is critical to refining the model's ability to generalize well to unseen data while effectively managing the trade-offs between bias and variance. The parameters adjusted during the grid search, along with the range of values considered for each, are summarized in Table 5.5 below:

Table 5.5 Range of Random Forest Parameters in Grid Search

|  |  |
| --- | --- |
| **Parameter** | **Grid Search** |
| n\_estimators | (50, 100, 200) |
| max\_depth | (1, 10, 5) |
| max\_features | (1,12) |
| min\_samples\_leaf | (2, 6, 2) |

Each parameter plays a vital role in the performance of the Random Forest model.

**n\_estimators**: Specifies the number of trees in the forest. Increasing the number of trees can improve the model's accuracy but also increases computational load.

**max\_depth**: Controls the maximum depth of each tree. Deeper trees can model more complex patterns but might lead to overfitting.

**max\_features**: Determines the number of features to consider when looking for the best split at each node. Using more features can increase the predictive power of the model, but too many features might cause overfitting.

**min\_samples\_leaf**: The minimum number of samples required to be at a leaf node. Setting this parameter can help in smoothing the model, especially in regression.

The grid search explores various combinations of these parameters to identify the set that offers the best performance, typically evaluated using cross-validation techniques. This meticulous tuning ensures that the Random Forest model not only captures the underlying trends in the training data but also maintains a robust performance on new, unseen data. This approach helps in achieving a balanced model that is neither too simplistic nor overly complex, providing reliable predictions that are critical for decision-making processes in credit risk evaluation.

**5.1.3.2 Model evaluation**

Table 5.6 Performance Metrics of the Random Forest Model

| Dataset | Accuracy | Precision | Recall | F1-Score | ROC AUC |
| --- | --- | --- | --- | --- | --- |
| Train Dataset | 0.7772 | 0.7803 | 0.7753 | 0.77 | 0.8485 |
| Test Dataset | 0.8484 | 0.9522 | 0.7291 | 0.8259 | 0.8468 |

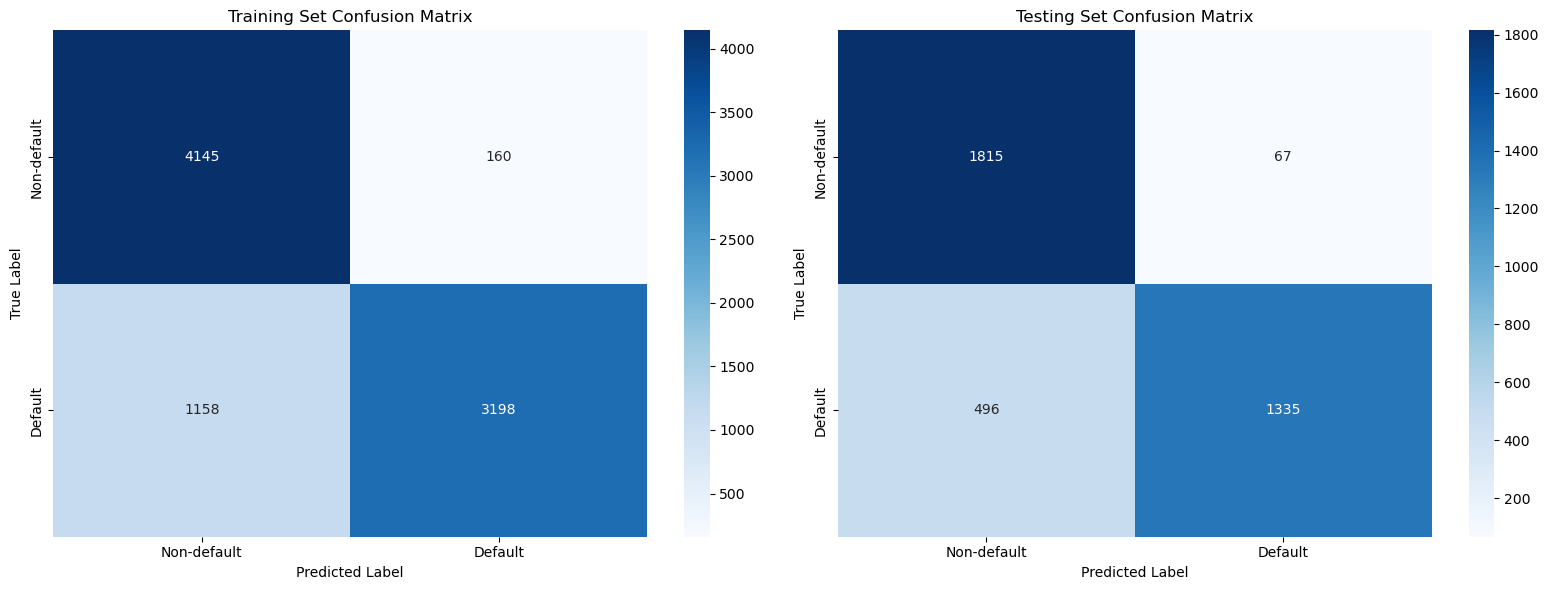
The confusion matrix for the training dataset (in Table 5.7) provides a visual breakdown of the model's performance, showcasing the true positives, true negatives, false positives, and false negatives. This matrix demonstrates the model's ability to distinguish between the default and non-default cases, with a significant number of true positives and true negatives, indicating a relatively balanced classification capability.

Figure 5.7Testing Set Confusion Matrix

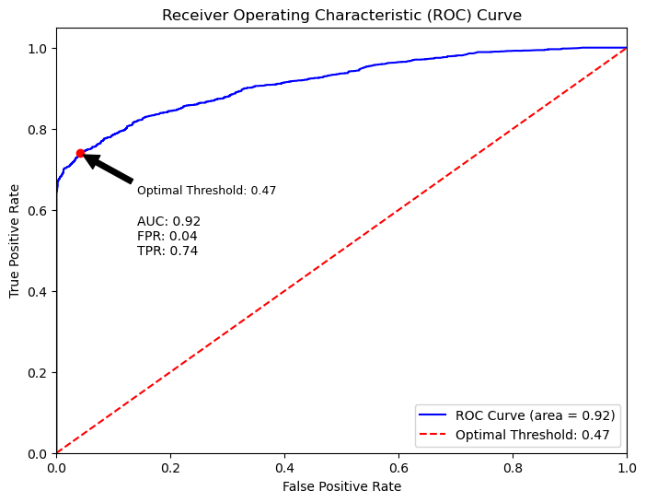
Similar to the training set, the testing set confusion matrix (in figure 1) illustrates the model's performance on unseen data. It highlights the precision of the model in predicting non-default cases and its relatively weaker performance in accurately identifying default cases, as indicated by the lower number of true positives compared to true negatives.

Figure 5.8 Receiver Operating Characteristic (ROC) Curve

The ROC curve and the area under the curve (AUC) metric (in figure 5.8) provide insights into the model's diagnostic ability. With an AUC of 0.92, the model demonstrates excellent discriminative ability between the default and non-default classes. The ROC curve's steep rise and optimal threshold point suggest a strong trade-off between the true positive rate and false positive rate, affirming the model's effectiveness in classifying the outcomes at an optimal threshold of 0.47.

**5.1.3.3 Decision Tree Visualization**

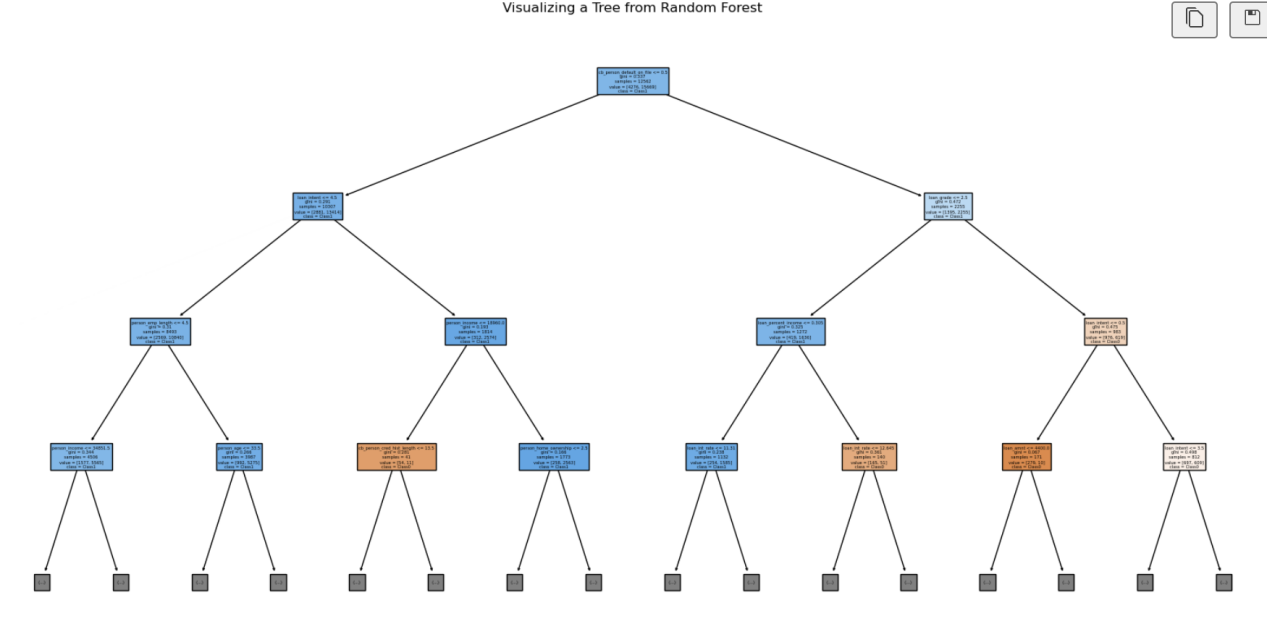
The tree visualization from the Random Forest shows the decision-making process (in Figure 5.9):The root node and subsequent splits are based on certain feature thresholds, indicating the features that most significantly affect the model's predictions. Each node specifies a condition on a feature that leads to either a further split or a terminal node, which represents a class prediction based on the majority of target values in that node.

Figure 5.9: Decision-making process

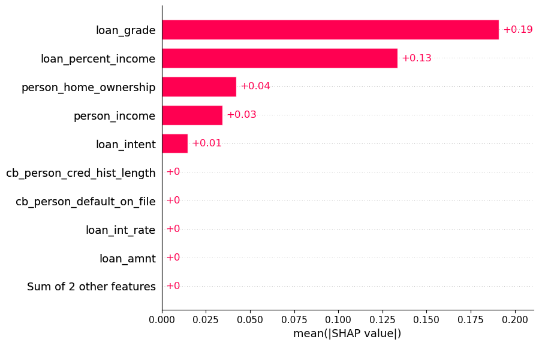
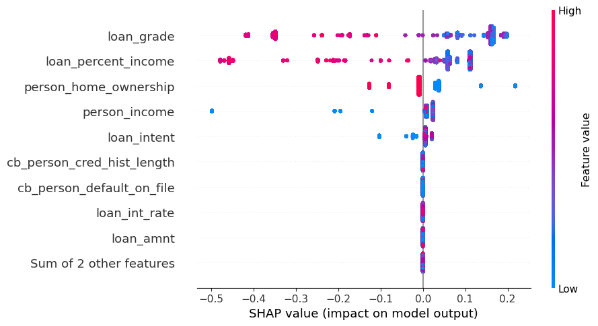
**5.1.3.4 Feature importance analysis**

To gain insights into how various customer attributes influence the credit risk classification of loan statuses, SHAP values were employed to conduct a feature importance analysis. Figure 3(a) presents a comprehensive overview of feature importance related to loan status, calculated as the average of the absolute Shapley values across the dataset. This visualization provides a clear representation of the most impactful features. Figure 3(b) displays a SHAP summary plot, illustrating the distribution of the SHAP values for each feature and indicating their corresponding influence trends.

**Financial Factors:** Notably, financial factors such as 'loan\_percent\_income' and 'person\_income' are paramount in assessing default risk. This emphasizes the importance of a borrower’s financial capacity relative to the loan size. Higher values of 'loan\_percent\_income' and lower income levels are generally associated with an increased risk of default. Conversely, the absolute amount of 'loan\_amnt' appears to have a relatively lesser impact on influencing default risk.

**Loan Attributes:** 'Loan\_grade', 'loan\_int\_rate', and 'loan\_intent' are significant attributes. The 'loan\_grade', which indicates the creditworthiness of the borrower, plays a crucial role in predicting the likelihood of repayment. Both 'loan\_intent' and 'loan\_int\_rate' are mid-level influential features. According to the SHAP values, loans intended for debt consolidation are perceived as higher risk, while higher loan interest rates generally correlate with a higher risk of default.

**Stability and Background Attributes:** 'Person\_home\_ownership' along with 'person\_emp\_length', and 'person\_age' significantly contribute to the model's predictions. This reflects the impact of housing stability on the risk assessment. Renters are associated with a higher risk of default, while homeowners show a lower risk. Additionally, 'person\_emp\_length' is identified as a more impactful factor than 'person\_age', indicating that borrowers with shorter employment durations are at a higher risk of default.

**Historical Credit Attributes:** 'Cb\_person\_cred\_hist\_length' and 'cb\_person\_default\_on\_file' are the least influential, implying that while they contribute to the overall risk assessment, their effects are minimal compared to the other factors discussed. These attributes, while essential for a holistic view, do not weigh as heavily in the predictive model for default risk.

1. (b)

Figure 5.10 Global interpretations of Random Forest model by SHAP values: (a) SHAP feature importance (b) SHAP summary plot.

**5.1.4 XGboost**

**5.1.4.1 Model establishment and training**

XGBoost combines multiple weak learners, such as CART trees, into a robust model by iteratively adding predictors that improve upon the predictions of previous iterations. Given the complex relationships between various borrower attributes and loan outcomes in the banking sector, we employ an XGBoost model to determine whether borrowers will default on their loans. To fine-tune the model's performance, we employ a grid search combined with 10-fold cross-validation for hyperparameter optimization. The range of values for the hyperparameters is presented in Table 5.7.

Table 5.7 Range of XGBoost Parameters in Grid Search

|  |  |
| --- | --- |
| **Parameter** | **Grid Search** |
| n\_estimators | (50, 100, 150, 200) |
| max\_depth | (4, 5, 6, 7) |
| learning\_rate | (0.01, 0.1, 0.2, 0.3) |
| reg\_lambda | (0.2, 0.4, 0.6, 0.8, 1) |

Following parameter optimization, the final model structure consists of 150 CART trees, with a maximum depth of 5 and a learning rate of 0.2. To mitigate the risk of overfitting due to XGBoost's strong fitting capability, we add an L2 regularization penalty term to balance prediction error and model complexity. The optimized value for the reg\_lambda parameter is determined to be 0.4.

**5.1.4.2. Model evaluation**

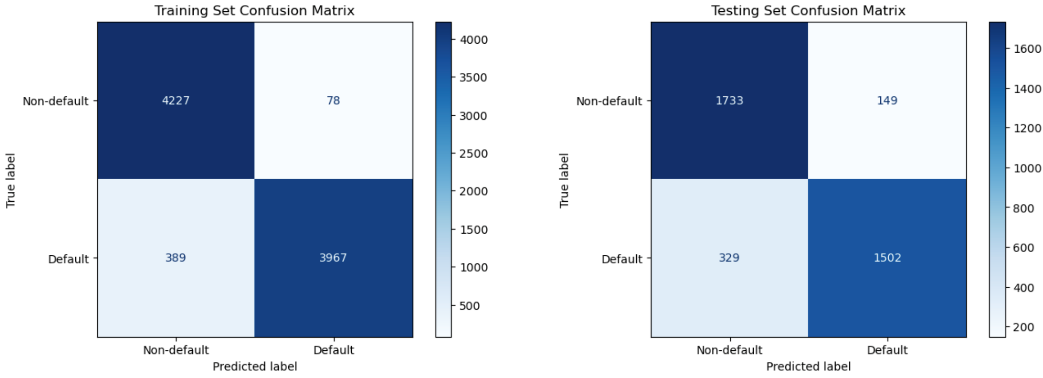
The confusion matrix of the train and test dataset based on the proposed XGBoost model is shown in Figure 5.11. Based on it, various performance metrics were calculated and are summarized in Table 5.8. According to the table, we can see that the model achieved training and testing accuracies of 94.61% and 87.13%, respectively, indicating its effectiveness in classifying loan statuses. Both precision and recall scores for the classes are well-balanced. In the test dataset, the F1 scores of 87.08% reflect the model's strong performance, effectively balancing both precision and recall.

Figure 5.11. The confusion matrix of the XGBoost model.

Table 5.8. Empirical results of XGBoost classification algorithm

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| Train dataset | 0. 9461 | 0.9482 | 0.9463 | 0.9460 |
| Test dataset | 0.8713 | 0.8751 | 0.8706 | 0.8708 |

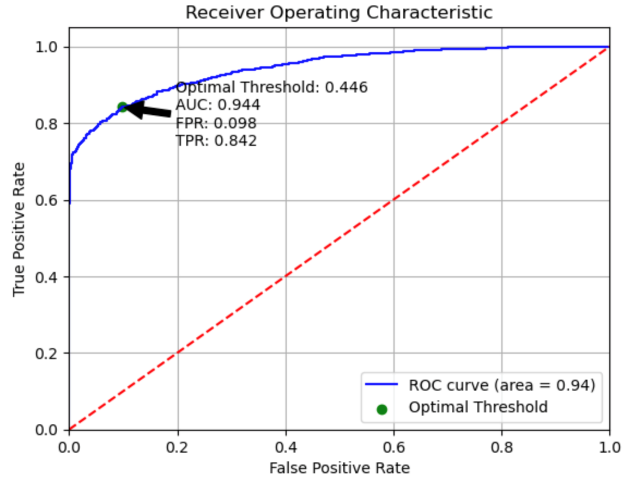
Additionally, the ROC curve for the test dataset is shown in Figure 5.12. The AUC score of 0.944 for the test dataset, along with an optimal threshold of 0.446, demonstrates the model's exceptional capability in distinguishing between 'Default' and 'Non-default' loans.

Figure 5.12 The ROC curve of the test data

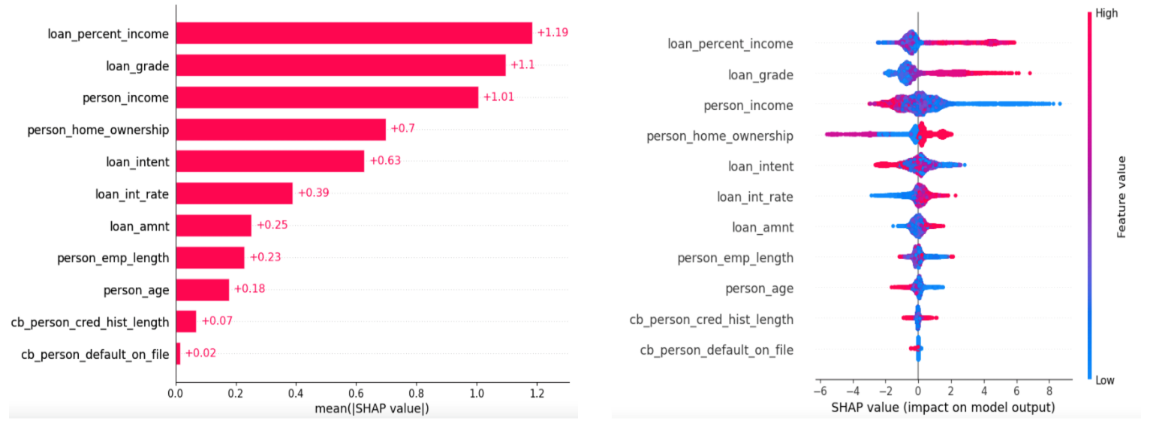
**5.1.4.3. Feature importance analysis**

To analyze how various customer attributes influence the credit risk classification of loan statuses, we employ SHAP values to conduct a feature importance analysis. Figure 5.13(a) presents a comprehensive overview of feature importance related to loan status, calculated as the average of the absolute Shapley values across the dataset. Figure 5.13(b) displays a SHAP summary plot, illustrating the distribution of the SHAP values for each feature and indicating their corresponding influence trends.

**Financial factors** includesloan\_percent\_income and person\_income, are paramount in assessing default risk, emphasizing a borrower’s financial capacity relative to the loan size. Higher loan\_percent\_income and lower income levels are generally associated with an increased risk of default. In contrast, the total loan amount has a lesser impact on default risk.

**Loan attributes** include loan\_grade, loan\_int\_rate and loan\_intent. Loan\_grade is a significant attribute, indicating that the grade assigned to the loan based on the borrower's creditworthiness plays a key role in predicting the likelihood of repayment. Loan\_intent and loan\_int\_rate are mid-level influential features; SHAP values indicate that loans intended for debt consolidation (0) are perceived as higher risk, while higher loan interest rates are generally associated with a higher risk of default.

**Stability and background attributes** include person\_home\_ownership, person\_emp\_length and person\_age. Person\_home\_ownership also contributes to model's predictions, with renters (3) associated with higher default risk and homeowners (2) with lower risk. Additionally, person\_emp\_length is a more impactful factor than person\_age, indicating that borrowers with shorter employment durations are at a higher risk of default.

**Historical credit attributes**, including cb\_person\_cred\_hist\_length and cb\_person\_default\_on\_file, are the least influential factors. While they contribute to the overall risk assessment, their effects are minimal compared to the other factors discussed.

(a) (b)

Figure 5.13 Global interpretations of XGBoost model by SHAP values: (a) SHAP feature importance; (b) SHAP summary plot.

**5.2 Clustering Analysis for Loan Status Profiling**

Cluster analysis is an important unsupervised learning method that automatically groups sample data based on their characteristics. In this section, we aim to understand the profiles of borrowers based on their loan statuses to reduce overall default risk. We apply the K-means clustering algorithm to analyze historical loan data, categorize customers into distinct groups, and provide personalized credit loan services tailored to the default risk and characteristics of each group.

**5.2.1 Data preparation and preprocessing**

Firstly, we remove the feature loan\_amnt, as it is strongly correlated with the combination of person\_income and loan\_percent\_income. Secondly, we exclude factor variables from the dataset including loan\_intent, loan\_grade, person\_home\_ownership and cb\_person\_default\_on\_file. This decision was made to ensure the effectiveness of the clustering process, as K-means relies on calculating distances between data points. Including categorical variables could distort these distance calculations and lead to less meaningful clusters. By focusing on numerical features, we aim to enhance the interpretability and accuracy of the clustering results. Lastly, min-max normalization is applied to scale all features, ensuring that varying scales do not disproportionately affect the clustering outcomes.

**5.2.2 Cluster Selection and Results analysis**

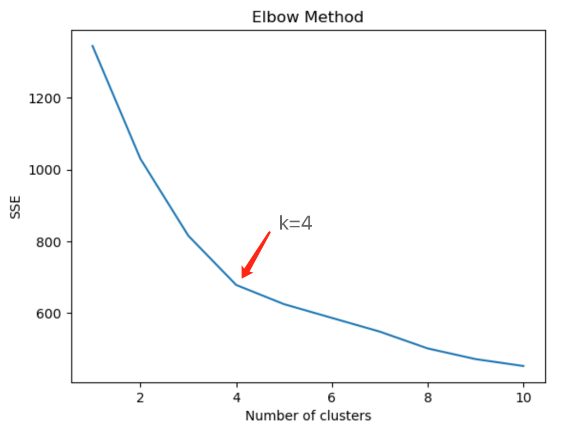
To determine the optimal number of clusters K, we use the Elbow Method. This method plots the Within-Cluster Sum of Squares (WCSS) for a range of K values to identify the "elbow" point, where the reduction in WCSS begins to slow significantly. Based on the results displayed in Figure 5.14, we select K = 4 as the optimal number of clusters for our dataset.

Figure 5.14: Elbow Method for optimal K-value selection in K-Means clustering

After determining the optimal K, we analyze the cluster centers generated by the K-means algorithm, shown in Table 5.9. These centers represent the average feature values for each cluster, providing insights into borrower profiles associated with different loan statuses.

Table 5.9 K-means cluster centers: average feature values

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **cluster** | **person\_age** | **person\_income** | **person\_emp\_length** | **loan\_int\_rate** | **loan\_percent\_income** | **cb\_person\_cred\_hist\_length** | **loan\_status** | **count** |
| **1** | **25.80** | **40862.35** | **3.91** | **11.29** | **0.37** | **4.64** | **0.79** | **2585** |
| **2** | **26.03** | **63275.70** | **4.07** | **14.94** | **0.16** | **4.72** | **0.62** | **4354** |
| **3** | **41.30** | **70953.18** | **5.90** | **12.10** | **0.19** | **14.73** | **0.47** | **1322** |
| **4** | **26.04** | **66686.77** | **4.96** | **8.78** | **0.13** | **4.64** | **0.20** | **4113** |

The clustering results categorize customers into four distinct groups. Based on the characteristics of these groups, we outline the following key observations.

(1) Cluster 1: This cluster includes the youngest borrowers with the lowest incomes and shortest employment durations. These borrowers exhibit the highest loan-percent-income, indicating the highest risk profile, reflected in the highest default rate.

(2) Cluster 2: This cluster consists of slightly older borrowers with higher incomes and longer employment histories. Despite facing relatively high loan\_int\_rate, the higher person\_income and lower loan\_percent\_income suggest a stronger financial position, indicating a lower risk profile compared to Cluster 1.

(3) Cluster 3: Comprising oldest individuals with the highest average income and longest employment histories among all clusters, Cluster 3 benefits from lower loan\_int\_rate, lower loan\_percent\_income and longer cb\_person\_cred\_hist\_length. These factors indicate lower credit risk and a more favorable borrower profile compared to Clusters 1 and 2.

(4) Cluster 4: Including younger borrowers with average incomes and employment lengths, this cluster shows the lowest loan\_status default rate. This may be attributed to the lowest loan\_percent\_income and lowest loan\_int\_rate, aligning with the negative SHAP values for these features, which suggest a lower risk.

### 6. Model Comparison

### 6.1 Evaluation of classification accuracy

This section primarily presents the evaluation metrics for different machine learning models. In this study, the dataset is divided into training and testing sets with a ratio of 30%, resulting in a total of 28,493 samples, with 8,661 samples in the training set and 3,713 samples in the testing set. Various machine learning models are trained using grid search for hyperparameter tuning and 10-fold cross-validation.

Tables 6.1-6.2 and Figure 6.1 present a comparative analysis of four machine learning models—Logistic Regression, Decision Tree, Random Forest, and XGBoost—across four critical performance metrics: Accuracy, Precision, Recall, and F1-score.

According to Figure 6.1 and Table 6.1-6.2,XGBoost demonstrates superior performance, achieving the highest scores in both training and testing sets, with an Accuracy of 0.9461 (training) and 0.8713 (testing), alongside an F1-score of 0.9460 (training) and 0.8708 (testing). This performance highlights its capability to effectively balance precision and recall. Conversely, both the Decision Tree and Random Forest models exhibit high precision, particularly in the training set; however, their lower Recall suggests challenges in capturing all default instances, which may limit their practical application. Lastly, Logistic Regression consistently shows the weakest performance across all metrics, particularly in Recall, with a score of only 0.7679 in the testing set, which underscores its limitations in identifying positive cases.

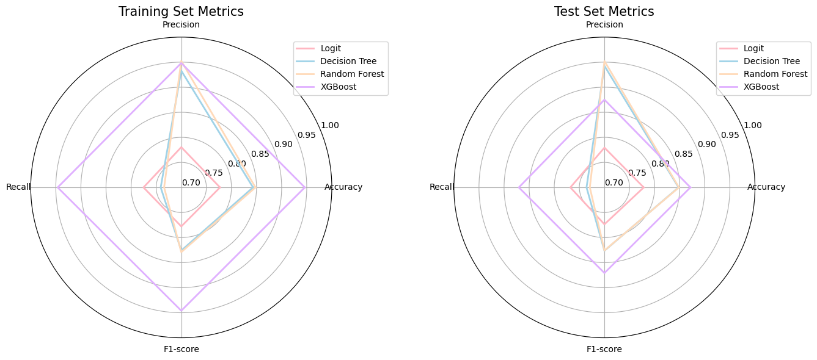
In summary, while XGBoost is identified as the most robust model, the inherent strengths and weaknesses of the Decision Tree and Random Forest models warrant further consideration, particularly in their Recall capabilities.

Table 6.1 The performance of Training Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| Logit | 0.7772 | 0.7803 | 0.7753 | 0.7778 |
| Decision Tree | 0.843 | 0.9329 | 0.741 | 0.826 |
| Random Forest | 0.8478 | 0.9524 | 0.7342 | 0.8291 |
| XGBoost | 0. 9461 | 0.9482 | 0.9463 | 0.946 |

Table 6.2 The performance of Testing Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithm** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| Logit | 0.7783 | 0.7794 | 0.7679 | 0.7736 |
| Decision Tree | 0.8476 | 0.9426 | 0.7357 | 0.8264 |
| Random Forest | 0.8484 | 0.9522 | 0.7291 | 0.8259 |
| XGBoost | 0.8713 | 0.8751 | 0.8706 | 0.8708 |

Figure 6.1 The radar chart of training and testing evaluation metrics

**6.2 Feature importance analysis among models**

Table 6.3 presents SHAP feature importance across four predictive models. Loan\_grade and loan\_percent\_income consistently rank as the most influential factors in predicting loan default risk, indicating their essential roles in assessing borrower creditworthiness and financial leverage. Additionally, person\_income and person\_home\_ownership rank highly across all models, highlighting the importance of personal stability and income levels in credit evaluation. Notably, loan\_intent serves as a top predictor in both XGBoost and Random Forest, indicating its relevance in specific contexts. This consistent ranking across diverse modeling approaches reinforces the robustness of these features as key indicators of credit risk, demonstrating their utility in developing more accurate credit scoring systems.

Table 6.3 The summary of SHAP Feature Importance of 4 Machine Learning Methods

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rank** | **Logistic Regression** | **CART** | **Random Forest** | **XGBoost** |
| 1 | loan\_grade | loan\_grade | loan\_grade | loan\_percent\_income |
| 2 | loan\_percent\_income | loan\_percent\_income | loan\_percent\_income | loan\_grade |
| 3 | person\_home\_ownership | person\_home\_ownership | person\_home\_ownership | person\_income |
| 4 | person\_income | person\_income | person\_income | person\_home\_ownership |
| 5 | person\_emp\_length | Loan\_int\_rate | loan\_intent | loan\_intent |

In summary, this study emphasizes the significance of loan\_grade, loan\_percent\_income, person\_home\_ownership and person\_income in predicting loan default risk. It also acknowledges the critical role of personal stability factors in credit evaluation. Furthermore, the variations in feature importance rankings across models illustrate the significance of model selection and feature engineering in effectively managing credit risk.

### Conclusion and recommendation

**7.1 Conclusion**

This study is based the credit risk dataset collected from Kaggle, covering personal information, loan details and the loan status. It employs various machine learning models to classify and assess borrower default risk. In the classification accuracy evaluation section, the XGBoost model demonstrates superior performance in both the training and testing sets, achieving accuracy rates of 0.9461 and 0.8713, with F1-scores of 0.9460 and 0.8708, respectively. In contrast, the Decision Tree and Random Forest models exhibit high precision in the training set, but their recall rates are relatively low. The Logistic Regression model performs the weakest across all metrics, particularly in recall, which is only 0.7679.

In the feature importance analysis section, financial factors(including percentage of income represented by the loan amount and individual’s income) and low attributes(including grade assigned to loans, interest rate and intention of loan application) significantly impact the prediction of loan status across all models, while stability and background attributes(including type of home ownership,employment length and age), historical credit attributes(including length of credit history, historical default record) follow behind. And percentage of income represented by the loan amount is the most significant feature compared with other financial factors, especially in the XGBoost model.

The study also conducts a clustering analysis of customer groups, categorizing them into four clusters ranked from high to low default risk as Cluster 1, Cluster 2, Cluster 3, and Cluster 4. The study finds that customer groups with higher income, lower loan-to-income ratios, and lower interest rates generally exhibit lower default rates.

Overall, the XGBoost model excels at identifying borrower default risk, and K-means clustering can identify the characteristics of different borrower groups. This paper provides empirical support for financial institutions in selecting risk models to predict default risk and in formulating risk management strategies tailored to different groups. However, the study still has limitations such as data constraints and model selection constraints, which need to be addressed in future research.

**7.2 Recommendations**

**7.2.1 Focusing on feature importance**

**7.2.1.1 Strengthening Income Assessment Mechanisms**

Given the importance of income in predicting loan status, financial institutions should establish stricter income assessment standards to ensure that borrowers' income levels maintain a reasonable proportion to their loan amounts. Introducing additional income verification documents may improve the accuracy of loan approvals.

**7.2.1.2 Optimizing Loan Product Design**

Based on the significance of loan grades, financial institutions can design various loan products to cater to borrowers' credit ratings. For example, offering more flexible repayment options or lower initial interest rates for borrowers with lower credit ratings can help alleviate their financial pressure, thereby reducing the risk of default.

**7.2.1.3 Implementing Risk Monitoring Mechanisms**

Establish a dynamic risk monitoring system based on features such as person\_home\_ownership and person\_income to regularly assess borrowers' financial situations. Timely awareness of changes in borrowers' income and family circumstances can assist banks in making quicker risk management decisions.

**7.2.2 Focusing on Cluster-centers**

**7.2.2.1 Focusing on financial health assessment with loan-to-income ratio**

Banks should enhance the evaluation of borrowers' financial health, particularly focusing on the loan-to-income ratio. For clients with higher loan-to-income ratios, a comprehensive assessment of their overall financial situation, including income, expenses, and liabilities, is essential to mitigate default risk. Regular communication with clients to understand their financial changes will help maintain their financial stability. Additionally, for higher-risk groups, such as Cluster 1 borrowers, encouraging them to provide additional collateral or guarantors can increase the security of the loans.

**7.2.2.2 Designing personalized loan product**

Financial institutions should develop diverse loan products tailored to different customer groups. For Cluster 1 borrowers—young individuals with lower incomes and high default risks—small loans with flexible repayment options are essential. For Cluster 3 borrowers, who are older with long employment histories, larger loans with longer repayment periods can help reduce monthly payment pressure, facilitating timely repayments. For Cluster 4 borrowers with lower default rates, banks should offer favorable terms like reduced interest rates or relaxed approval to attract these quality clients.

**7.2.2.3 Enhancing financial counseling and educational service**

Establish a comprehensive financial counseling and education platform, particularly targeting Cluster 1 and Cluster 2 borrowers. This platform could offer training on financial literacy, budget management, and credit assessment. By enhancing the financial management skills of these borrowers, it is possible to effectively reduce default risk and help them better understand and manage their loans.

## References

[1] Bao, W., Lianju, N., & Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. Expert Systems with Applications, 128, 301-315.

[2] Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11), 2767-2787.

[3] Ma, X., & Lv, S. (2019). Financial credit risk prediction in internet finance driven by machine learning. Neural Computing and Applications, 31(12), 8359-8367.

[4] Niu, A., Cai, B., & Cai, S. (2020). [Retracted] Big Data Analytics for Complex Credit Risk Assessment of Network Lending Based on SMOTE Algorithm. Complexity, 2020(1), 8563030.

[5] Oualid, A., Hansali, A., Balouki, Y., & Moumoun, L. (2022). Application of machine learning techniques for credit risk management: a survey. In Advances in Information, Communication and Cybersecurity: Proceedings of ICI2C’21 (pp. 180-191). Springer International Publishing.

[6] Pławiak, P., Abdar, M., & Acharya, U. R. (2019). Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring. Applied Soft Computing, 84, 105740.

[7] Seitshiro, M. B., & Govender, S. (2024). Credit risk prediction with and without weights of evidence using quantitative learning models. Cogent Economics & Finance, 12(1), 2338971.

[8] Shi, S., Tse, R., Luo, W., D’Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: a systemic review. Neural Computing and Applications, 34(17), 14327-14339.

[9] Wattanakitrungroj, N., Wijitkajee, P., Jaiyen, S., Sathapornvajana, S., & Tongman, S. (2024). Enhancing Supervised Model Performance in Credit Risk Classification Using Sampling Strategies and Feature Ranking. Big Data and Cognitive Computing, 8(3), 28.

[10] Ghodselahi, A., & Amirmadhi, A. (2011). Application of artificial intelligence techniques for credit risk evaluation. International Journal of Modeling and Optimization, 1(3), 243.

[11] Hu, Y., & Su, J. (2022). Research on credit risk evaluation of commercial banks based on artificial neural network model. Procedia Computer Science, 199, 1168-1176.

[12] Yanenkova, I., Nehoda, Y., Drobyazko, S., Zavhorodnii, A., & Berezovska, L. (2021). Modeling of bank credit risk management using the cost risk model. Journal of Risk and Financial Management, 14(5), 211.

[13] Lopez, J. A., & Saidenberg, M. R. (2000). Evaluating credit risk models. Journal of Banking & Finance, 24(1-2), 151-165.

[14] Crouhy, M., Galai, D., & Mark, R. (2000). A comparative analysis of current credit risk models. Journal of Banking & Finance, 24(1-2), 59-117.