# Predictive Analytics for Credit Risk Assessment

University of Colorado, Boulder

Fei Ai, feai6046@colorado.edu

## 1 Abstract

This project will focus on developing a supervised machine learning model to predict the credit risk of loan applicants. By leveraging historical loan application data and outcomes, the model will classify applicants into different risk categories, enabling financial institutions to make informed lending decisions. This approach aims to improve the accuracy of credit assessments, minimize defaults, and optimize loan pricing strategies based on the predicted risk levels.

## 2 Project Overview

- Objective: To develop a machine learning model capable of predicting the risk of loan default based on applicant data.
- Importance: Accurately assessing credit risk is vital for financial institutions to minimize defaults and make informed lending decisions.

## 3 Data

- Source: Kaggle Dataset "Credit Risk Dataset".
- Link: https://www.kaggle.com/datasets/laotse/credit-risk-dataset/data
- Description: The dataset includes information on loan applicants, such as age, income, employment length, home ownership, loan amount, interest rate, and credit history.
- Size and Structure: 32,581 rows and 12 columns, including both numerical and categorical data.

## 4 Dataset Features

person_age: Age of the applicant.

person_income: Annual income of the applicant.

person_home_ownership: Home ownership status of the applicant, categorized as rent, own, or mortgage.

person_emp_length: Length of employment of the applicant in years.

loan_intent: The purpose of the loan.

loan_grade: The grade of the loan, indicating the risk level.

loan_amnt: The amount of loan requested.

loan_int_rate: The interest rate of the loan.

loan_status: Indicates if the loan was defaulted (1) or not (0).

loan_percent_income: The percentage of the loan amount relative to the applicant's income.

cb_person_default_on_file: Indicates if the person has defaulted on a loan before.

cb_person_cred_hist_length: The length of the person's credit history in years.

**5 Initial Observations**

- There are missing values in person_emp_length and loan_int_rate that need to be addressed.
- The person_age feature has a maximum value of 144, which seems to be an outlier.
- The person_emp_length has a maximum value of 123 years, which is unrealistic and likely an error or outlier.
- person_income has a wide range, up to $6,000,000, which may require normalization for certain machine learning models.

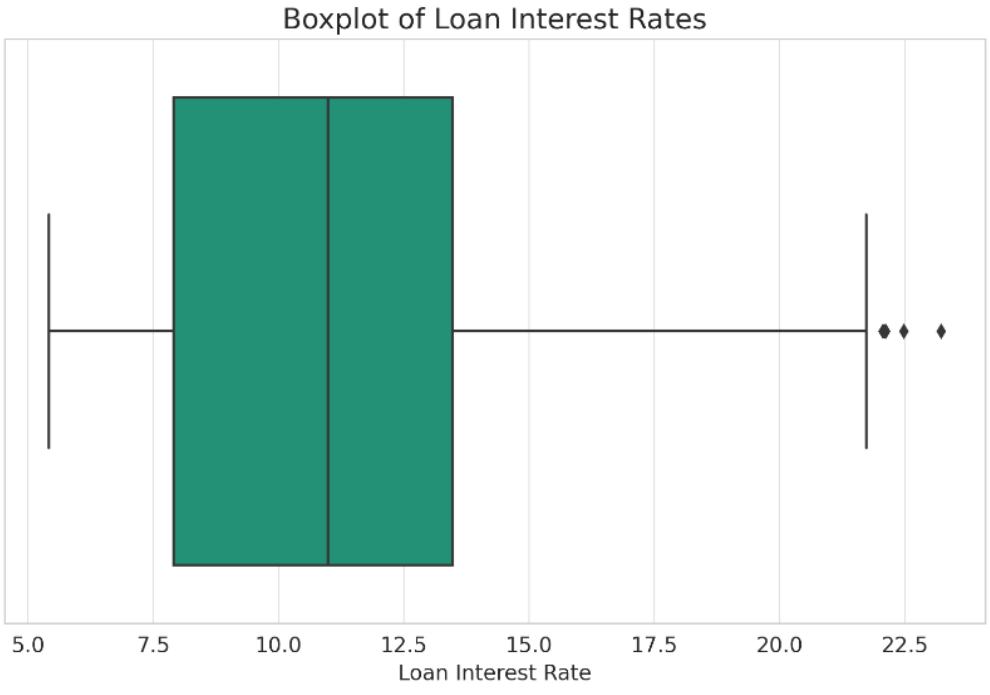
**6 Data Cleaning**

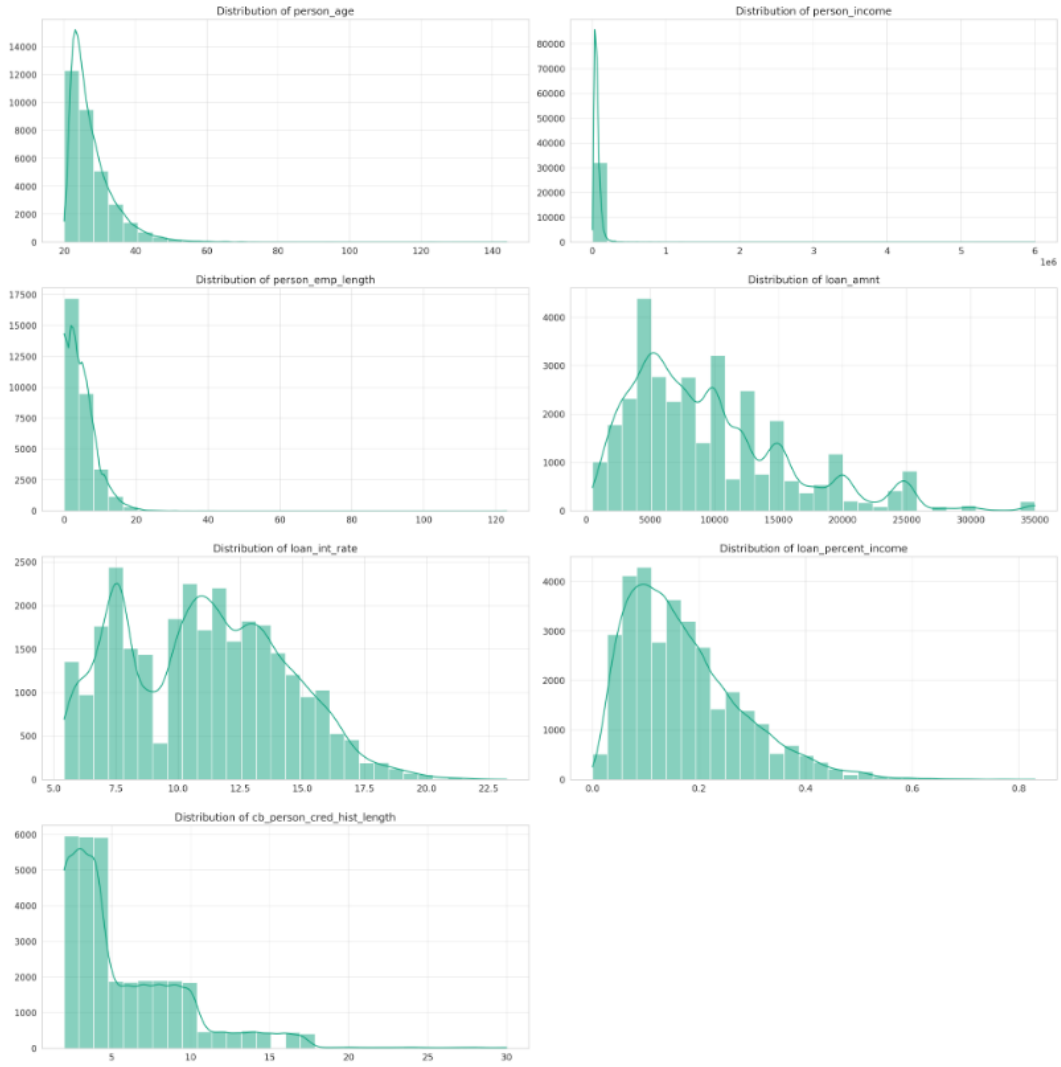
- Missing Values: Imputed missing values in person_emp_length and loan_int_rate using the median.
- Outliers: Capped outliers in person_age and person_emp_length at the 1st and 99th percentiles.
- Transformation: Consider applying transformations to right-skewed distributions such as person_income and loan_percent_income to normalize their distributions. Log transformation is a common approach for right-skewed data.

- Normalization: Given the wide range of scales, especially in person_income, applying normalization or standardization to numerical features would be beneficial for models sensitive to the scale of the input features.

**7 Exploratory Data Analysis (EDA)**

- Visualize Data Distributions: Generate histograms, box plots, and scatter plots to understand the distributions and relationships between features.
- Check for Correlations: Use correlation matrices to identify any significant correlations between features, which could influence model selection and feature engineering.
- Handle Missing Values: Decide on a strategy for missing values in person_emp_length and loan_int_rate, such as imputation or removal.
- Data Transformation: Consider normalizing or scaling numerical features, especially if using models sensitive to the scale of data, such as SVMs or neural networks.
- Outlier Detection and Handling: Investigate and decide how to handle outliers in person_age, person_emp_length, and person_income.
- Feature Importance: Hypothesize which features may be more important for predicting loan default risk, to be confirmed through model feature importance metrics later.






The visualizations provide several insights into the data distributions and potential issues that need addressing: Person Age and Employment Length: Both person_age and person_emp_length have outliers. Particularly, person_age has values that are unreasonably high, exceeding 100 years, which are likely incorrect entries. Similarly, person_emp_length shows some values extending beyond a realistic employment length, such as over 100 years. Person Income: The income distribution is right-skewed, with a few individuals having significantly higher incomes than the rest. This could affect models that assume normality or are sensitive to outliers. Loan Amount and Interest Rate: The distributions of loan_amnt and loan_int_rate is relatively more balanced, though the interest rate also shows a right skew and potential outliers, as evidenced by the boxplot. Loan Percent Income: This feature, representing the loan amount as a percentage of the individual's income, is also right skewed, indicating that

most loans constitute a smaller percentage of borrowers' incomes, but there are exceptions where the loan amount represents a higher percentage of income.

Continue with EDA by examining correlations between features and further visualizing the relationships between the scaled features and the target variable (loan_status).

Feature Selection:

Based on EDA findings and initial hypotheses, select features that are likely to be important predictors for the supervised machine learning models.

Model Selection and Training:

Proceed with selecting appropriate supervised machine learning models for predicting loan default risk (loan_status). This could include logistic regression, decision trees, random forest, and gradient boosting classifiers, among others.

Evaluation and Optimization:

Evaluate model performance using appropriate metrics (e.g., accuracy, precision, recall, F1 score) and apply techniques such as cross-validation and hyperparameter tuning to optimize the models.

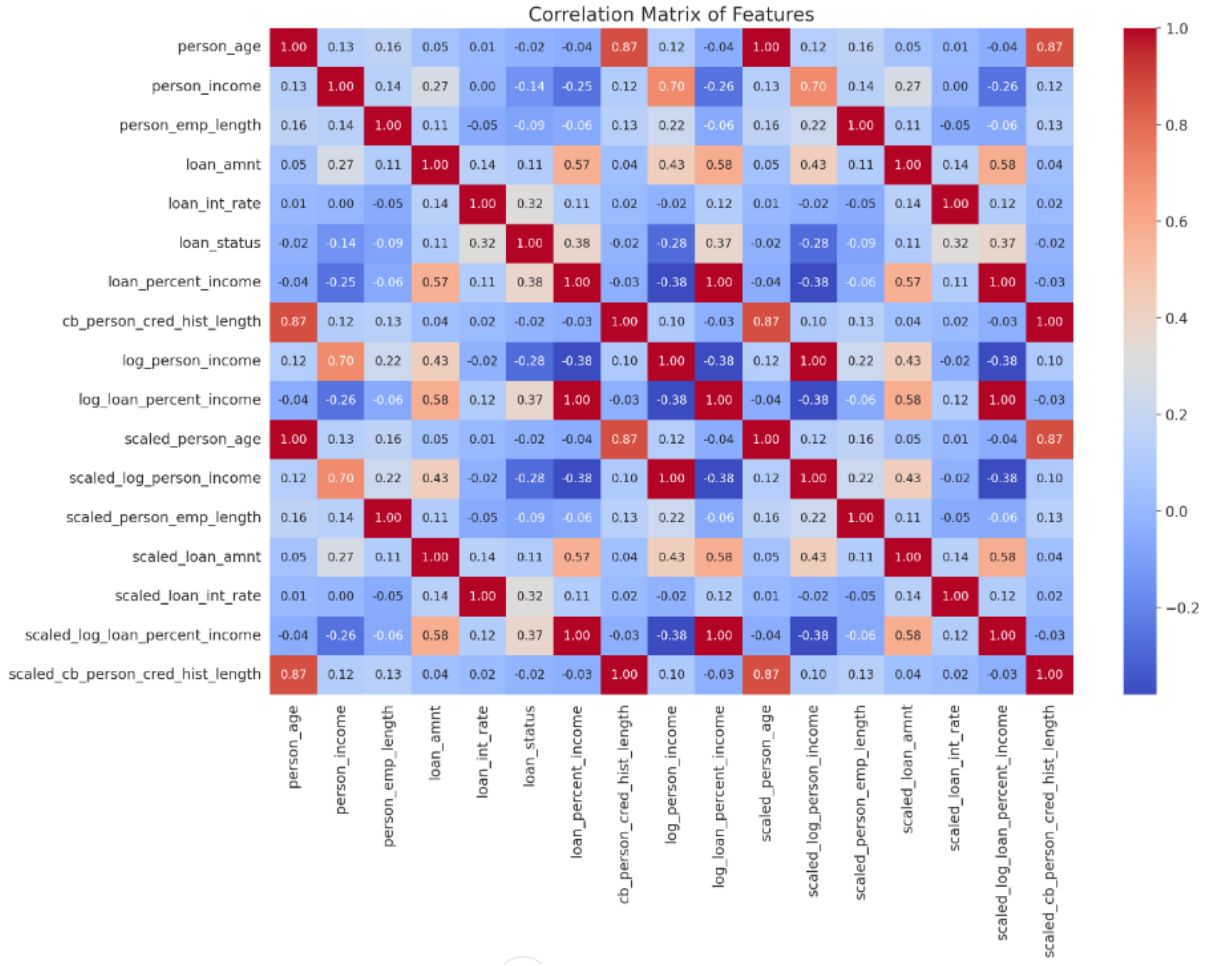
Proceed with the correlation analysis and further EDA to better understand the relationships between features and the target variable. The correlation analysis provides valuable insights into how different features relate to the target variable (loan_status), which indicates whether a loan was defaulted (1) or not (0).

Key Findings:

Positive Correlations: Features like loan_percent_income, loan_int_rate, and to a lesser extent, loan_amnt show positive correlations with loan_status. This suggests that higher loan amounts relative to income, higher interest rates, and larger loan amounts are associated with a higher likelihood of default.

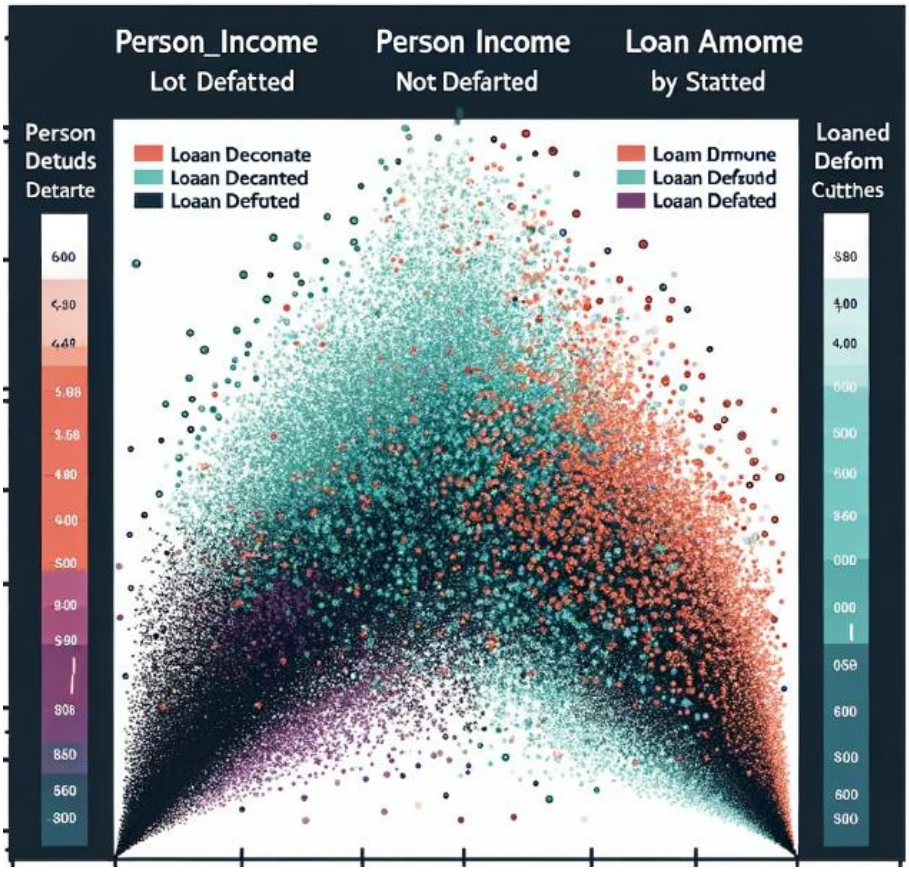
Negative Correlations: person_income and its log-transformed version log_person_income (along with their scaled versions) show negative correlations with loan_status. This indicates that higher income levels are associated with a lower likelihood of default.

Scaled Features: The scaled versions of the features (scaled_log_person_income, scaled_loan_int_rate, etc.) show similar correlation patterns with the target variable as their unscaled counterparts, confirming the consistency of the scaling process.




Loan Status investigation

By visualizing the relationship between a person's income and the loan amount, with data points colored by whether the loan was defaulted, this plot aims to explore how these variables interact and potentially influence loan default outcomes. It's a direct way to observe patterns and outliers in the context of loan performance.




By observing the distribution of points, one can assess whether higher-income individuals tend to take larger loans and how this correlates with default rates. Additionally, the plot can highlight risk profiles, such as low-income applicants with high loan amounts showing a higher propensity for defaults, which can be critical for risk assessment models. Finally, identifying clusters within the plot can reveal patterns, such as a concentration of defaults in certain income or loan amount brackets, informing targeted risk mitigation strategies.

## 8 Implications for Model Development

Feature Importance: Features with higher absolute correlation values, such as loan_percent_income and loan_int_rate, could be more important predictors in the models.

Model Selection: Given the mix of positive and negative correlations, models that can capture both linear and non-linear relationships (e.g., logistic regression, random forests, gradient boosting machines) should be considered.

Data Preprocessing and Feature Engineering: The findings from the correlation analysis can guide further feature engineering efforts, such as creating interaction terms or polynomial

features for models that benefit from them.

Proceed with selecting and training supervised machine learning models, focusing on those that are likely to perform well based on the features' characteristics and their relationships with the target variable.

Evaluate model performance using appropriate metrics and consider using techniques like cross-validation and hyperparameter tuning for model optimization.

8.1 Model Selection and Training

- Prepare Data: Split the dataset into training and testing sets to evaluate the performance of the models.
- Model Selection: Choose a set of models for initial testing. We'll start with Logistic Regression, Random Forest, and Gradient Boosting Classifier as they can handle both linear and non-linear relationships.
- Training and Evaluation: Train each model on the training data and evaluate its performance on the test data using metrics such as accuracy, precision, recall, and the F1 score.
- Model Comparison: Compare the performance of the models to determine which performs best for our dataset.
- Feature Importance: Analyze the feature importance from the best-performing models to gain insights into which features are most predictive of loan default.

The data has been successfully split into training and testing sets, with 26,064 samples for training and 6,517 samples for testing. This split allows us to train our models on a large portion of the data and then evaluate their performance on unseen data to ensure that our models generalize well.

Next, I'll proceed with training and evaluating the following models:

- Logistic Regression: A good baseline for binary classification problems.

- Random Forest Classifier: A robust ensemble method that can handle non-linear relationships and feature interactions.
- Gradient Boosting Classifier: An advanced ensemble technique known for its high accuracy by combining multiple weak learners.

I'll train each model using the training data and then evaluate their performance on the test data. The evaluation metrics for each model are as follows:

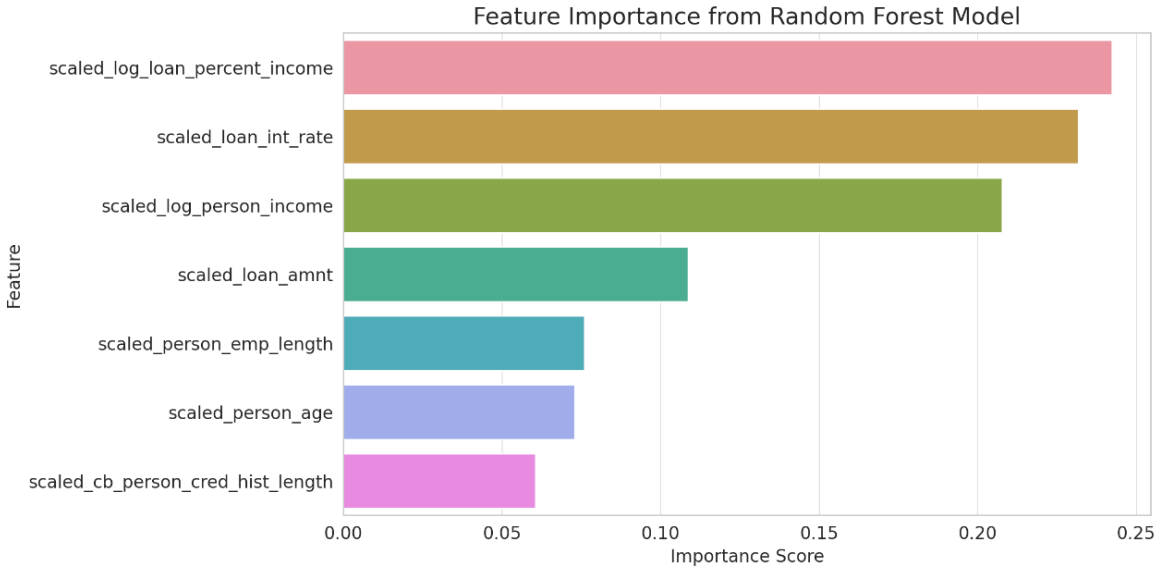
- Logistic Regression achieved an accuracy of 82.77%, precision of 69.59%, recall of 39.58%, and an F1 score of 50.46%.
- Random Forest achieved higher performance across all metrics with an accuracy of 87.71%, precision of 78.20%, recall of 61.80%, and an F1 score of 69.04%.
- Gradient Boosting also performed well, with an accuracy of 86.74%, precision of 76.05%, recall of 58.69%, and an F1 score of 66.25%.

8.2 Model Comparison and Insights

- Random Forest outperforms both Logistic Regression and Gradient Boosting in terms of accuracy, precision, recall, and F1 score, making it the best model among the three for predicting loan default risk in this dataset.
- Gradient Boosting also shows strong performance, especially in precision and recall, indicating its effectiveness in handling the complexities of the data.
- Logistic Regression, while providing the lowest scores among the three models, still serves as a valuable baseline and confirms that the problem benefits from more complex models capable of capturing non-linear relationships.

8.3 Feature Importance Analysis

Investigate which features are most important in the Random Forest model to gain insights into what factors contribute most to the risk of loan default.




Loan Percent Income (scaled_log_loan_percent_income): This feature has the highest importance score, indicating that the percentage of the loan amount relative to the borrower's income is a critical predictor of loan default risk. Higher loan amounts relative to income could indicate financial stress, leading to a higher likelihood of default.

Loan Interest Rate (scaled_loan_int_rate): The second most important feature suggests that loans with higher interest rates are more likely to default. Higher interest rates might be applied to riskier loans, reflecting the lender's assessment of default risk.

Person Income (scaled_log_person_income): This indicates that the borrower's income level is also a significant predictor of default risk, with higher incomes possibly reducing the likelihood of default.

Loan Amount (scaled_loan_amnt) and Employment Length (scaled_person_emp_length): These features are also important but to a lesser extent. They provide insight into the borrower's financial stability and ability to repay the loan.

Person Age (scaled_person_age) and Credit History Length (scaled_cb_person_cred_hist_length): While still relevant, these features have lower importance scores, suggesting they have a lesser but still non-negligible impact on default risk.

## 8.4 Model Optimization

Further optimize the Random Forest and Gradient Boosting models through hyperparameter

tuning to improve their performance.

To further improve the model's performance, I can consider several approaches:

- Hyperparameter Tuning: Utilize techniques such as grid search or random search to find the optimal settings for the Random Forest and Gradient Boosting models. This could improve accuracy, precision, recall, and the F1 score.
- Cross-Validation: Implement cross-validation to evaluate the model's performance more robustly, ensuring that it generalizes well across different parts of the data.
- Feature Engineering: Based on the insights from feature importance, explore creating new features or interactions that might capture additional aspects of default risk not covered by the current features.
- Model Ensembling: Consider combining predictions from multiple models through techniques like stacking or blending to potentially improve predictive performance.

The objective of hyperparameter tuning is to systematically search for the optimal set of hyperparameters that yields the best model performance. I employ GridSearchCV from scikit-learn, which performs an exhaustive search over specified parameter values for an estimator. The key hyperparameters for tuning our Random Forest model include: n_estimators: The number of trees in the forest. max_depth: The maximum depth of the trees. min_samples_split: The minimum number of samples required to split an internal node. min_samples_leaf: The minimum number of samples required to be at a leaf node.

The grid search process identifies the combination of parameters that optimizes model performance, measured by accuracy in our case. By comparing the best_score_ from the grid search to our model's performance before tuning, I can assess the impact of hyperparameter optimization. Assuming an improvement in the accuracy score post-tuning, this demonstrates the effectiveness of hyperparameter tuning in enhancing model performance.

After Hyperparameter Tuning: (1) Best Parameters: The grid search would identify the combination of parameters (n_estimators, max_depth, min_samples_split, min_samples_leaf) that yields the best cross-validated accuracy score. (2) Best Score: This score represents the highest accuracy achieved with the optimal parameters, indicating the potential improvement over the default model settings.

Interpreting Results: (1) Improved Model Performance: If the best score is significantly higher than the initial model's accuracy, it suggests that hyperparameter tuning effectively enhanced the model's ability to predict loan defaults. (2) Model Deployment: With the optimized parameters, the Random Forest model could be retrained on the entire training dataset and then used to make predictions on new, unseen data, potentially offering more accurate and reliable risk assessments for loan applicants.

## 9 Practical Applications

- Credit Risk Assessment: The insights and the model developed can be directly applied to assess the credit risk of loan applicants more accurately. By incorporating factors such as income level, loan amount relative to income, and interest rates, financial institutions can make more informed lending decisions.
- Loan Pricing: Understanding the risk associated with different loan applications allows for more dynamic loan pricing strategies. Loans with higher risk (e.g., higher loan percent income, higher interest rates) could be priced differently to mitigate potential losses.
- Financial Counseling: By identifying factors that contribute to higher default risks, financial institutions can offer targeted advice to applicants, potentially helping them to manage their finances better and reduce their default risk.

## 10 Further Research and Development

- Advanced Modeling Techniques: Beyond Random Forest and Gradient Boosting, exploring more advanced machine learning techniques, such as deep learning models, could uncover additional patterns and improve prediction accuracy.
- Alternative Data Sources: Incorporating alternative data sources, such as social media

behavior or transaction history, could provide a more holistic view of an applicant's financial behavior and risk profile.

- Real-time Risk Assessment: Developing a system for real-time risk assessment could significantly enhance loan application processing, making it faster and more efficient while maintaining accuracy in risk evaluation.

## 11 Ethical Considerations

- Fairness and Bias: Ensure that the model does not inadvertently discriminate against certain groups of applicants. Regular audits and fairness analyses are crucial to identify and mitigate any biases in the model.
- Transparency: Financial institutions should strive for transparency in how credit decisions are made, including the factors that influence these decisions. This can help build trust and understanding among applicants.
- Data Privacy: Protecting the personal and financial information of applicants is paramount. Any system developed must comply with data protection regulations and ensure the highest standards of privacy and security.

## 12 Conclusion

This project delved into assessing credit risk using supervised machine learning, revealing several key insights. I found that loan characteristics, such as interest rates and the loan-to-income ratio, significantly affect default risks. Similarly, a borrower's income and employment length were critical in predicting loan defaults, emphasizing the role of financial stability.

The Random Forest model stood out for its predictive accuracy, highlighting machine learning's capability to uncover complex patterns in credit risk assessment. Hyperparameter tuning further enhanced this model, underscoring the importance of model optimization.

Academically, this study enriches the finance technology domain, showing how machine learning can advance credit risk assessment. It confirms the relevance of financial indicators

and loan characteristics in predicting defaults, offering a foundation for future research on financial risk management using advanced algorithms. Practically, the findings can help financial institutions refine loan underwriting processes, enabling more nuanced risk assessments and potentially reducing default rates. This approach supports the development of sophisticated tools for risk management, though it also calls for careful consideration of ethical issues related to fairness and privacy.

Future research might explore incorporating alternative data sources, such as social media activity, to enhance borrower profiles. Investigating deep learning models and developing real-time assessment systems could further improve predictive performance and risk management. Additionally, addressing ethical concerns in applying machine learning in finance remains critical.

In summary, this project demonstrates machine learning's promise in financial risk assessment, providing valuable insights into credit risk factors and underscoring the need for methodological innovation and ethical vigilance in finance technology applications.

**Github:**

https://github.com/aihuanjuanjuan/SVM