Predictive Analytics for Credit Risk Assessment

(final version)

Fei Ai
University of Colorado, Boulder
feai6046@colordo.edu

Outline

1 Introduction

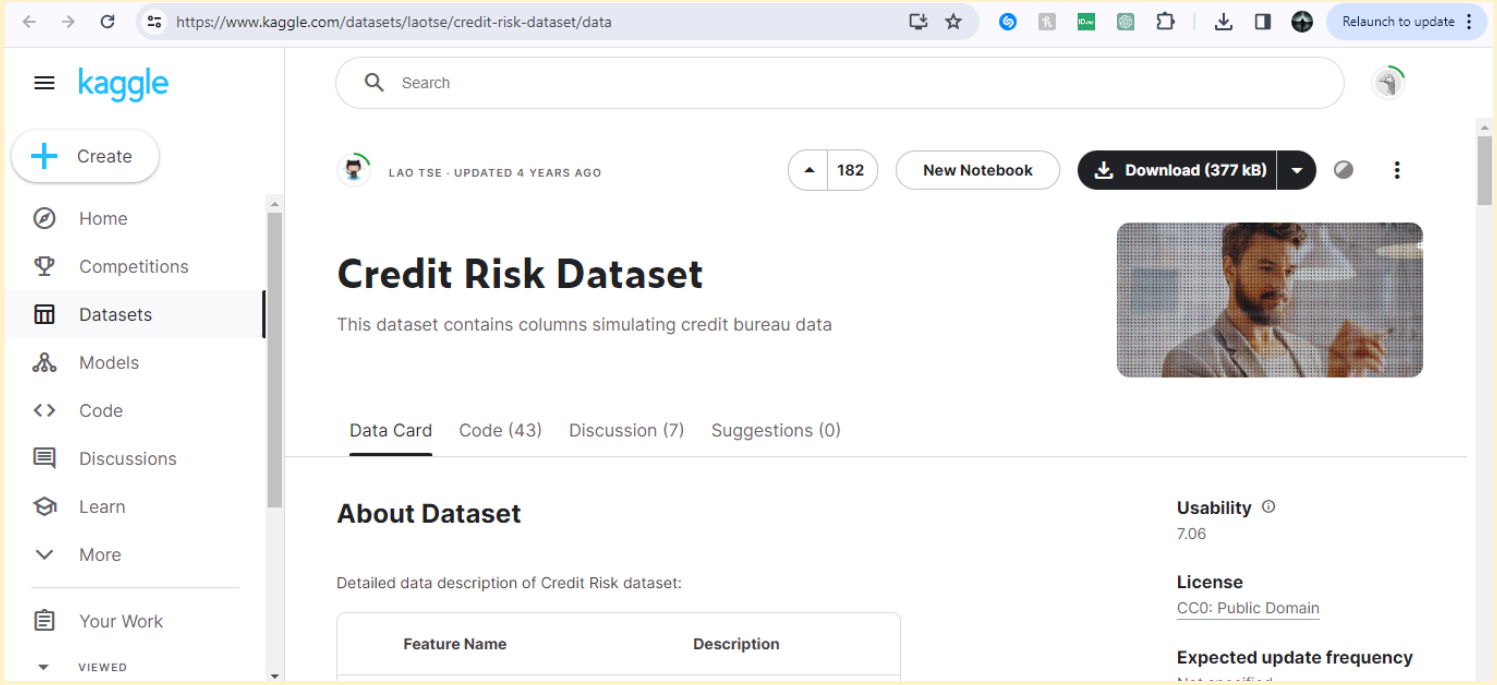
This project will focus on developing a supervised machine learning model to predict the credit risk of loan applicants. By leveraging historical loan application data and outcomes, the model will classify applicants into different risk categories, enabling financial institutions to make informed lending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

To develop a machine learning model capable of predicting the risk of loan default based on applicant data.

Accurately assessing credit risk is vital for financial institutions to minimize defaults and make informed lending decisions.

2 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.



Data Cleaning

Missing Values: Imputed missing values in person_emp_length and loan_int_rate using the median.

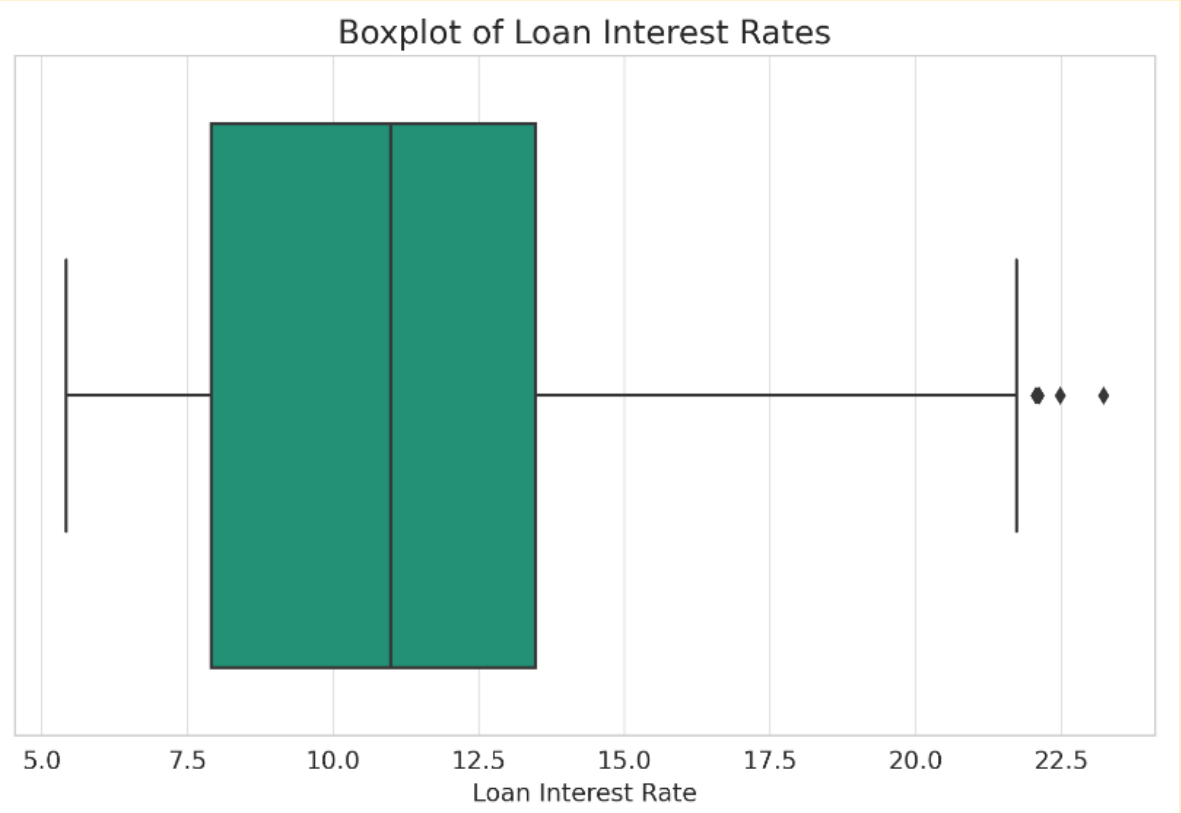
Outliers: Capped outliers in person_age and person_emp_length at the 1st and 99th percentiles.

Transformation: to right-skewed person_income and loan_percent_income to normalize their distributions

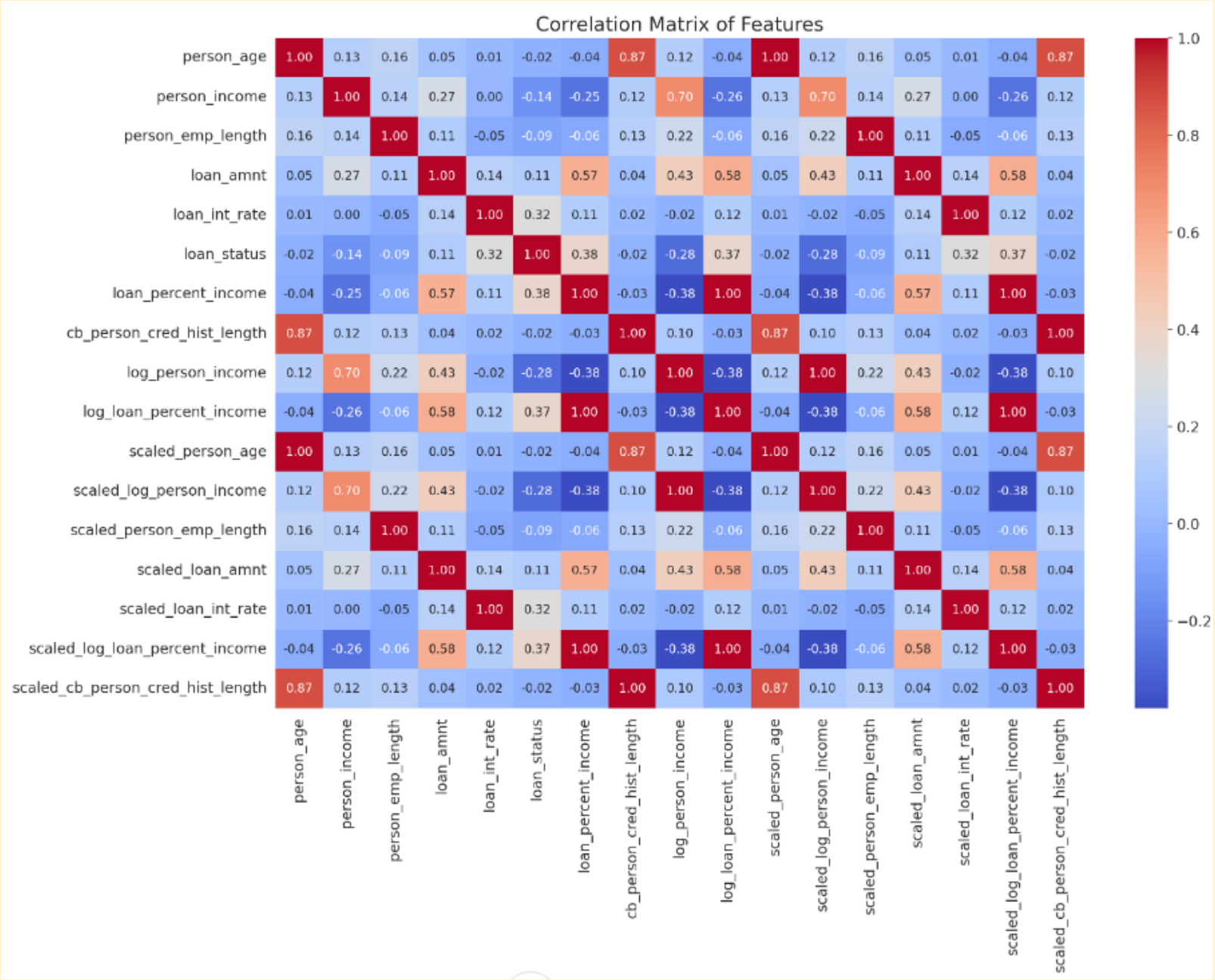
Normalization: person_income, applying normalization to numerical features

3 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.


Boxplot of Loan Interest Rates
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
Loan Interest Rate


Distribution of person_age
Distribution of person_income
1e6
Distribution of person_emp_length
Distribution of loan_amnt
Distribution of loan_int_rate
Distribution of loan_percent_income
Distribution of cb_person_cred_hist_length

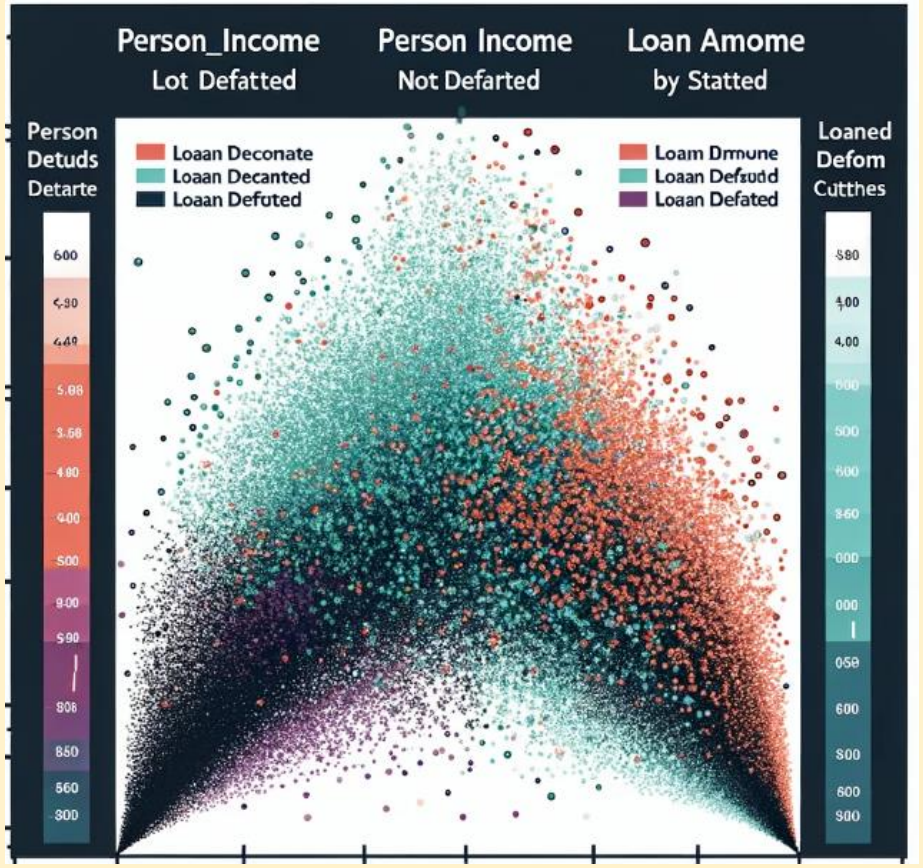

Correlation Matrix of Features
person_age 1.00 0.13 0.16 0.05 0.01 -0.02 -0.04 0.87 0.12 -0.04 1.00 0.12 0.16 0.05 0.01 -0.04 0.87
person_income 0.13 1.00 0.14 0.27 0.00 -0.14 -0.25 0.12 0.70 -0.26 0.13 0.70 0.14 0.27 0.00 -0.26 0.12
person_emp_length 0.16 0.14 1.00 0.11 -0.05 -0.09 -0.06 0.13 0.22 -0.06 0.16 0.22 1.00 0.11 -0.05 -0.06 0.13
loan_amnt 0.05 0.27 0.11 1.00 0.14 0.11 0.57 0.04 0.43 0.58 0.05 0.43 0.11 1.00 0.14 0.58 0.04
loan_int_rate 0.01 0.00 -0.05 0.14 1.00 0.32 0.11 0.02 -0.02 0.12 0.01 -0.02 -0.05 0.14 1.00 0.12 0.02
loan_status -0.02 -0.14 -0.09 0.11 0.32 1.00 0.38 -0.02 -0.28 0.37 -0.02 -0.28 -0.09 0.11 0.32 0.37 -0.02
loan_percent_income -0.04 -0.25 -0.06 0.57 0.11 0.38 1.00 -0.03 -0.38 1.00 -0.04 -0.38 -0.06 0.57 0.11 1.00 -0.03
cb_person_cred_hist_length 0.87 0.12 0.13 0.04 0.02 -0.02 -0.03 1.00 0.10 -0.03 0.87 0.10 0.13 0.04 0.02 -0.03 1.00
log_person_income 0.12 0.70 0.22 0.43 -0.02 -0.28 -0.38 0.10 1.00 -0.38 0.12 1.00 0.22 0.43 -0.02 -0.38 0.10
log_loan_percent_income -0.04 -0.26 -0.06 0.58 0.12 0.37 1.00 -0.03 -0.38 1.00 -0.04 -0.38 -0.06 0.58 0.12 1.00 -0.03
scaled_person_age 1.00 0.13 0.16 0.05 0.01 -0.02 -0.04 0.87 0.12 -0.04 1.00 0.12 0.16 0.05 0.01 -0.04 0.87
scaled_log_person_income 0.12 0.70 0.22 0.43 -0.02 -0.28 -0.38 0.10 1.00 -0.38 0.12 1.00 0.22 0.43 -0.02 -0.38 0.10
scaled_person_emp_length 0.16 0.14 1.00 0.11 -0.05 -0.09 -0.06 0.13 0.22 -0.06 0.16 0.22 1.00 0.11 -0.05 -0.06 0.13
scaled_loan_amnt 0.05 0.27 0.11 1.00 0.14 0.11 0.57 0.04 0.43 0.58 0.05 0.43 0.11 1.00 0.14 0.58 0.04
scaled_loan_int_rate 0.01 0.00 -0.05 0.14 1.00 0.32 0.11 0.02 -0.02 0.12 0.01 -0.02 -0.05 0.14 1.00 0.12 0.02
scaled_log_loan_percent_income -0.04 -0.26 -0.06 0.58 0.12 0.37 1.00 -0.03 -0.38 1.00 -0.04 -0.38 -0.06 0.58 0.12 1.00 -0.03
scaled_cb_person_cred_hist_length 0.87 0.12 0.13 0.04 0.02 -0.02 -0.03 1.00 0.10 -0.03 0.87 0.10 0.13 0.04 0.02 -0.03 1.00
person_age
person_income
person_emp_length
loan_amnt
loan_int_rate
loan_status
loan_percent_income
cb_person_cred_hist_length
log_person_income
log_loan_percent_income
scaled_person_age
scaled_log_person_income
scaled_person_emp_length
scaled_loan_amnt
scaled_loan_int_rate
scaled_log_loan_percent_income
scaled_cb_person_cred_hist_length
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

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.

4 Implications for Model Development

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.

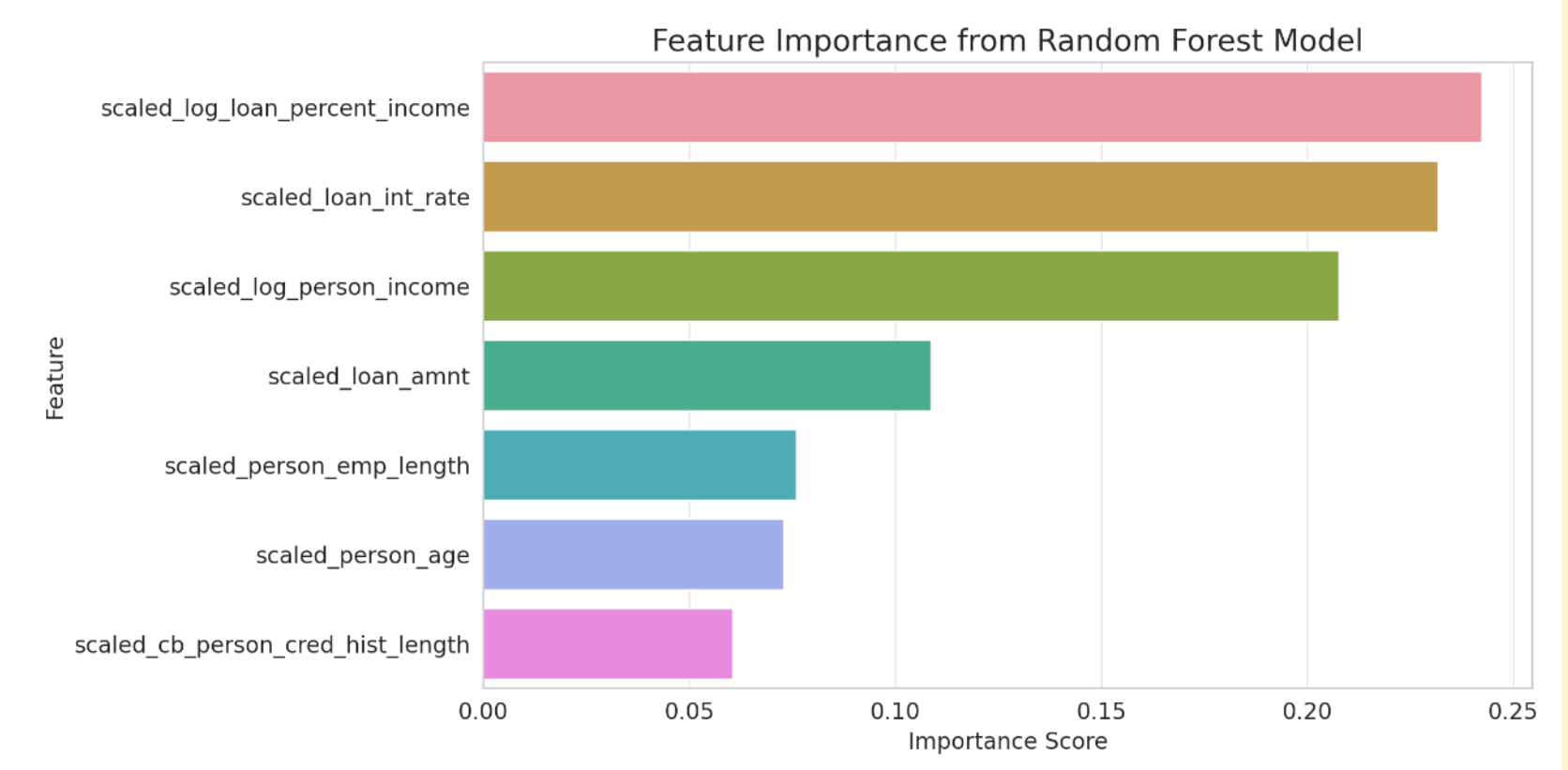
- Logistic Regression achieved an accuracy of 82.77%, precision of 69.59%, recall of 39.58%, and an F1 score of 50.46%.
- Random Forest achieved higher performance across all metrics with an accuracy of 87.71%, precision of 78.20%, recall of 61.80%, and an F1 score of 69.04%.
- Gradient Boosting also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

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.

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.



Model Optimization

- 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.

5 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.

6 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.

7 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.

5 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.