Loan Default Prediction for Financial Loan Services

Jing Xu
Brown DSI

Github repository: https://github.com/Jing-Xu1223/DATA1030-Project

I.Introduction

1.Motivation:

Financial loan services are essential for supporting individuals and corporations. They are fundamental in a variety of industries, from commercial banks to government-backed lending programs. However, a persistent challenge for them is managing and mitigating loan defaults, as those can significantly impact profitability and overall financial stability. To overcome this difficulty, institutions are increasingly leveraging data-driven strategies to evaluate credit worthiness and forecast default risk.

By analyzing diverse data sources, such as personal demographics, credit histories, and other financial indicators, models can identify patterns and make accurate predictions on potential defaulters. These insights allow institutions to proactively implement interventions, such as adjusting loan terms or deploying risk-based pricing strategies.

2.Dataset Overview:

The dataset originates from Coursera, which offers an exceptional opportunity to address a highly relevant machine learning problem in the financial industry. The dataset comprises 255,347 rows and 18 columns, representing a substantial amount of data that reflects the complexity encountered in industry. Each row corresponds to an individual's loan applicant, capturing various attributes that are critical for predicting default risk.

This dataset is a binary classification framework, where the target variable “Default”, indicates whether an individual caused a loan default (class1) or did not (class0). It contains 16 features that cover categorical, ordinal, and continuous data types. In addition, this dataset has no missing values, and is independently and identically distributed.

3.Previous research:

Previous work has analyzed the performance of accuracy, precision and recall for different machine learning classifiers. Random Forest stood out with an accuracy of 0.79, offering a reliable balance across metrics; while the Logistic Regression achieved the highest accuracy at 0.86, despite limited recall, highlighting its struggle to identify true positives. Models like Naive Bayes and KNN prioritized recall, with Naive Bayes achieving the highest recall 0.51, suitable for scenarios where sensitivity is critical. This analysis highlights the trade-offs between accuracy, precision, and recall. These findings provide valuable benchmarks and insights for further exploration and optimization in this study.

II.Exploratory Data Analysis

The distribution of the target variable “Default” is highly imbalanced, with the majority class0(repayment) significantly outweighing the minority class1(default).

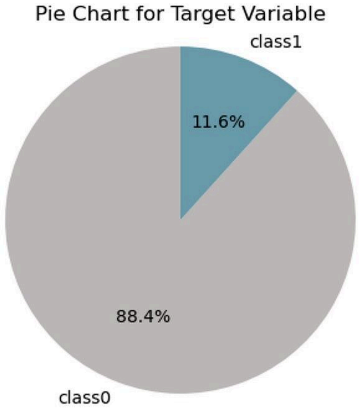


Fig 1. Pie Chart of Default Outcome

Categorical and ordinal features displayed approximately uniform distributions across their categories, indicating no dominant category that skews the dataset.

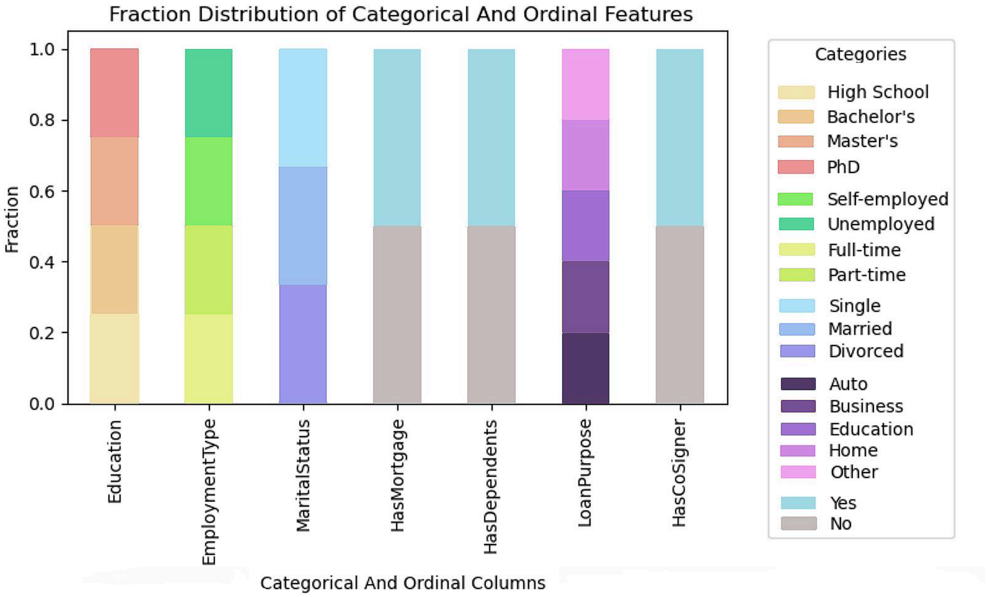
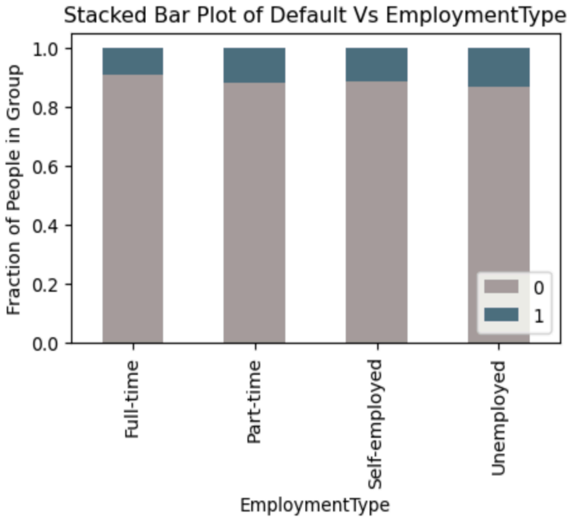


Fig 2. Stacked Bar Plot of Categorical&Ordinal Features

Three continuous features:Employment Type, Annual Income and Age, were selected to establish pairwise relationships between the target variable in attempt to uncover insightful patterns:

1.Employment Type: Although there didn't appear to be much of a difference within each employment type, borrowers who are employed full-time are less likely to default compared to those who are unemployed, reflecting the importance of stable employment in financial health.



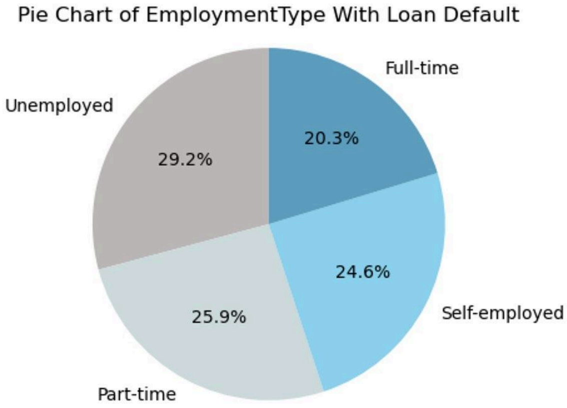


Fig 3. Stacked Bar Plot of Default vs Employment Type

Fig 4. Pie chart of Employment Type with Loan Default

2.Annual Income: It suggests that the distribution is slightly more concentrated at lower incomes for Class1, whereas Class0 shows a more even distribution. There seems to be a pattern where individuals in the lower-income bracket are more likely to default (Class 1), while individuals in higher-income brackets are evenly spread between default and no default.

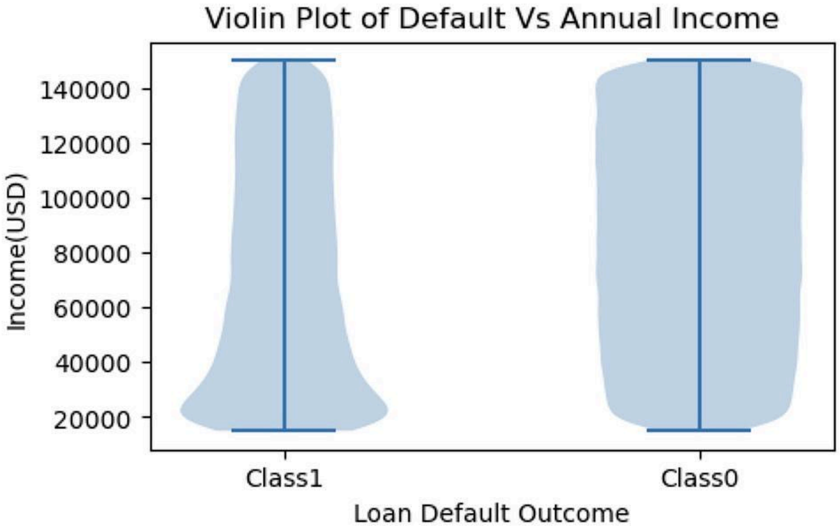


Fig 5. Violin Plot of Default vs Annual Income

3.Age: Younger individuals appear more prone to loan defaults, which could be attributed to lower financial stability or income. As age increases, the likelihood of default decreases, indicating that financial responsibility and stability improve with age.

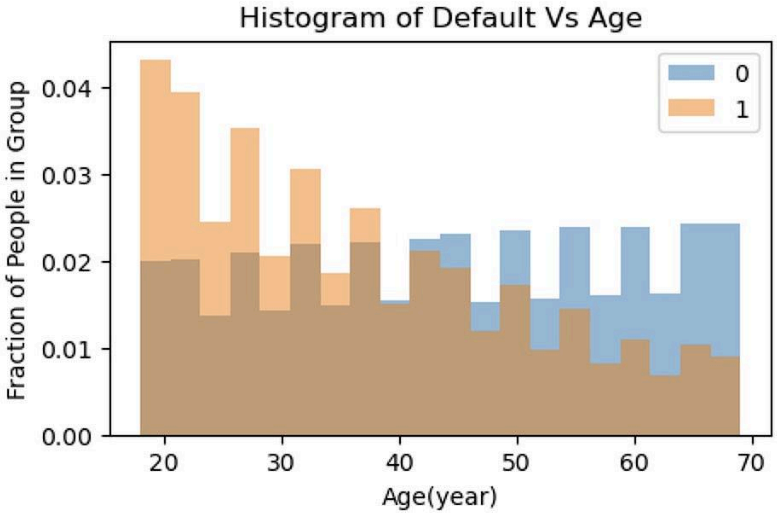


Fig 6. Histogram of Default vs Age

III.Methodology

1.Splitting Strategy: Stratified Kfold Split
To address the highly imbalanced nature of the dataset and ensure consistency, stratified splitting was employed. Initially, the data was stratified into two subsets: 10% was allocated as the test set, while the remaining 90% formed the training and validation set. Subsequently, the training and validation set underwent a 3-fold cross-validation using stratification to maintain the class balance across all folds.

2.Data Preprocessing
- 6 categorical features: EmploymentType: ,MaritalStatus, LoanPurpose, HasMortgage, HasDependents, and HasCoSigner, were transformed using One-Hot Encoder.
- 1 ordinal feature, Education, was encoded using an Ordinal Encoder with a predefined order (High School < Bachelor's < Master's < PhD).
- 9 continuous features: Age, Income, LoanAmount, CreditScore, MonthsEmployed, NumCreditLines, InterestRate, LoanTerm,DTIRatio, were standardized using Standard Scaler.

As a result, 12 new columns were added to the dataset, increasing the feature set from 16 to 28 columns, thereby capturing all necessary information in a model-ready format.

3. Evaluation Metric: F1 Score
By focusing on the harmonic mean of precision and recall, the F1 score effectively accounts for both FPs and FNs, ensuring the model's performance is not biased towards the majority class. This is crucial in loan default prediction, where accurately identifying defaults (class1) is as important as avoiding false alarms.

4.ML Algorithms And Hyperparameter Tuning:

Four models: Logistic Regression with elastic net regularization, Random Forest, Support Vector Classification, XGBoost, were selected in cross-validation and corresponding hyperparameters are tuned.

Model	Hyperparameter Tunning	Class weight
Logistic Regression (elastic net)	C: [0.01, 0.1, 1, 10, 100] l1_ratio: [0.1, 0.3, 0.5, 0.7, 0.9]	balanced
Random Forest Classification	max_depth: [1, 3, 5, 10, 20, 30] max_features: [0.01, 0.1, 0.3, 0.5, 0.7, 0.9]	balanced
SVC (support vector classification)	C: [0.001, 0.01, 0.1, 1, 10, 100] kernel': ['linear', 'rbf'] gamma': ['scale', 'auto', 0.01, 0.1, 1]	balanced
XGBoost	n_estimators: [50, 100, 200, 300], learning_rate: [0.001, 0.01, 0.1, 0.2], max_depth: [1, 3, 5, 7, 10], subsample: [0.8, 1.0], colsample_bytree: [0.8, 1.0]	scale_pos_weight = (class0/class1)

Table 1. Models trained and hyperparameters tuned

To address imbalance, the parameter *class_weight='balanced'* was employed automatically for Logistic Regression, Random Forest, and Support Vector Classifier. For XGBoost, the parameter *scale_pos_weight* was set as the *imbalance ratio*: counts of class0 over class1.

5. ML Pipeline:

After determining splitting strategy, setting up preprocessor, choosing F1 as metric, as well as listing 4 ML models and corresponding parameter grids:

- Cross-Validation Function (MLpipeline_StratifiedKFold):

- For each of 5 random states, the dataset was stratified into a test set (10%) and a train-validation set (90%).
- Within the train-validation set, three-fold cross-validation was conducted, employing grid search to optimize hyperparameters.
- The pipeline was trained on the training subset of each fold and validated on the corresponding validation subset.
- The test F1 score and the best hyperparameters were recorded for each iteration.

- The process was repeated with 5 different random states to measure splitting uncertainty, each model was trained multiple times on the same split using different random seeds, and the mean and standard deviation of test F1 scores were recorded.

- The final evaluation involved comparing these metrics to identify the best-performing model, determined by the highest mean test F1 score.

IV.Results

1.Baseline F1 Score:

The baseline F1 score, calculated on the class distribution of over 88% no loan defaulter and 12% loan defaulter, was 0.2081. All 4 ML models demonstrated substantial improvements over this score, with their test F1 scores consistently higher and exhibiting minimal standard deviations that were thousands of times above the baseline.

2. ML Model Performance Summary:

Model	Best Hyperparameters	Mean F1	Std F1
Logistic regression (elastic net)	C: 0.1, l1_ratio: 0.3	0.3581	0.0209
Random Forest Classification	max_depth: 5, max_features: 0.1	0.3312	0.0260
SVC (support vector classification)	C: 0.01, Gamma: scale, Kernel: rbf	0.3633	0.0226
XGBoost	N_estimators: 200, Learning_rate: 0.01, Max_depth: 5, Subsample: 0.8, Colsample_bytree: 0.8	0.3529	0.0421

Table 2. F1 Test Score Summary

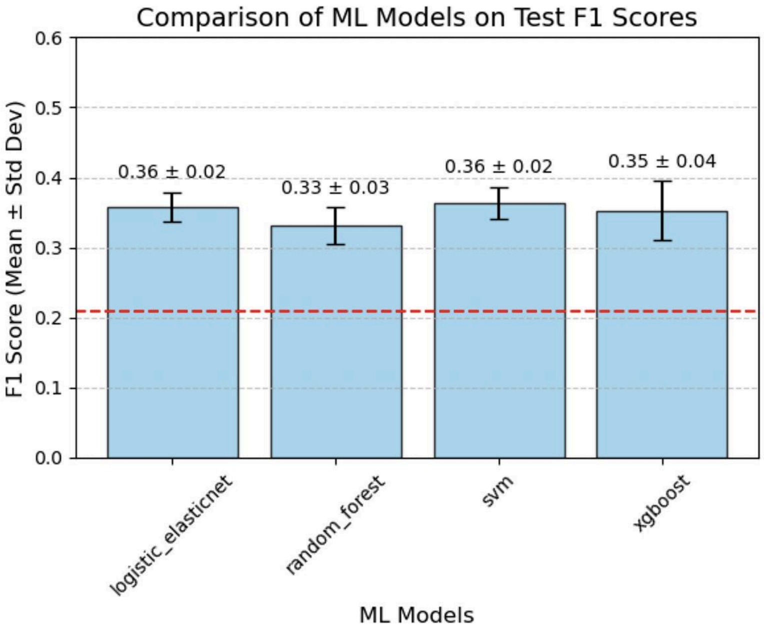


Fig 7. Mean and Std Test F1 Score Distribution For Different Models

All 4 ML models achieved similar test F1 scores, ranging from 0.33 to 0.36, with very low standard deviations of approximately 0.02. Among them, the Support Vector Classifier emerged as the best-performing model with its optimal hyperparameters yielding the highest mean F1 score. Given its insight, SVC was selected as the best model for subsequent interpretations.

The confusion matrix for the best-performing SVC model demonstrates an improvement in detecting the minority class1(defaulter). However, due to the significant class imbalance, the model still struggles with a high number of true negatives and false positives. The imbalance in the dataset continues to impact overall performance, highlighting the challenges of achieving balanced predictions for both classes.

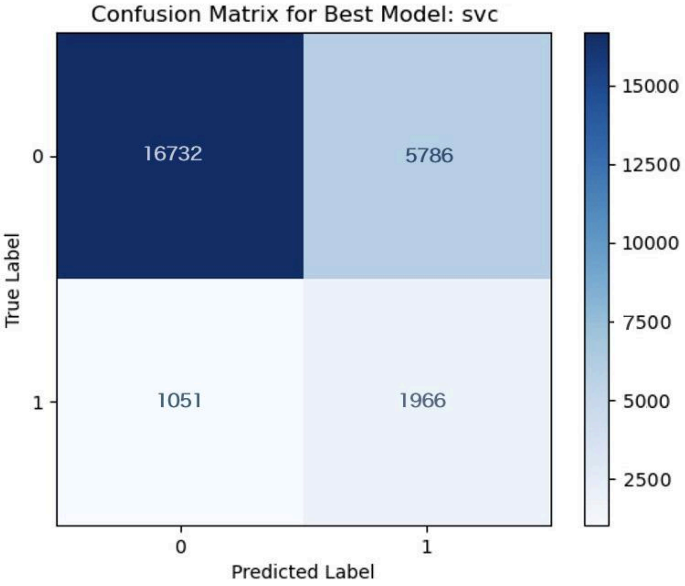


Fig 8. Confusion Matrix For Best Model: SVC

3.Global Feature Importance:

To identify the most influential features in predicting loan defaults, three global feature importance methods were employed: ***Permutation Feature Importance, SVC Model Weights,*** and ***SHAP Global Importance***. These complementary approaches provided insights into the role of each feature in the model's performance.

The TOP5 significant features from three plots below consistently highlighted that *Age, Interest Rate, Income, and CreditLine* are the most significant features influencing the prediction of loan defaults. These features play a critical role in the model’s decision-making process, emphasizing their importance in assessing an individual's likelihood of defaulting on a loan.

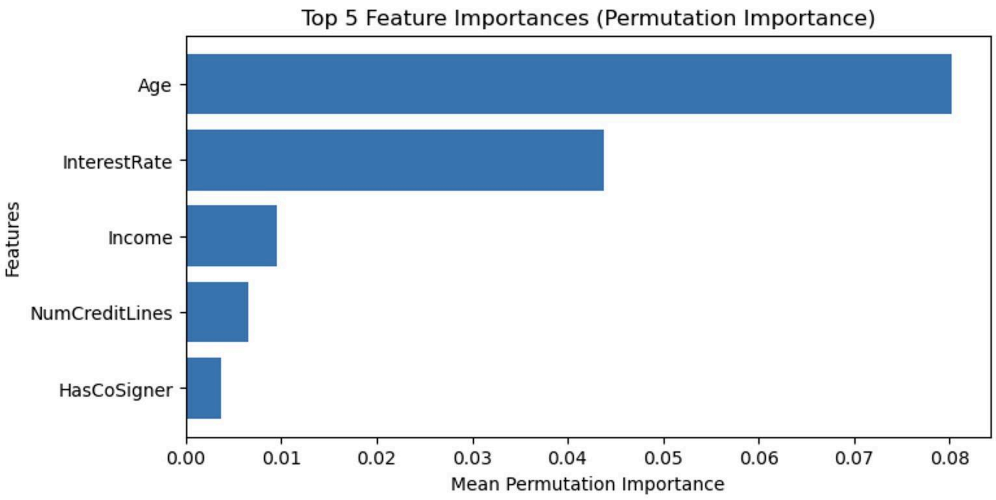


Fig 9. Permutation Feature Importance: SVC

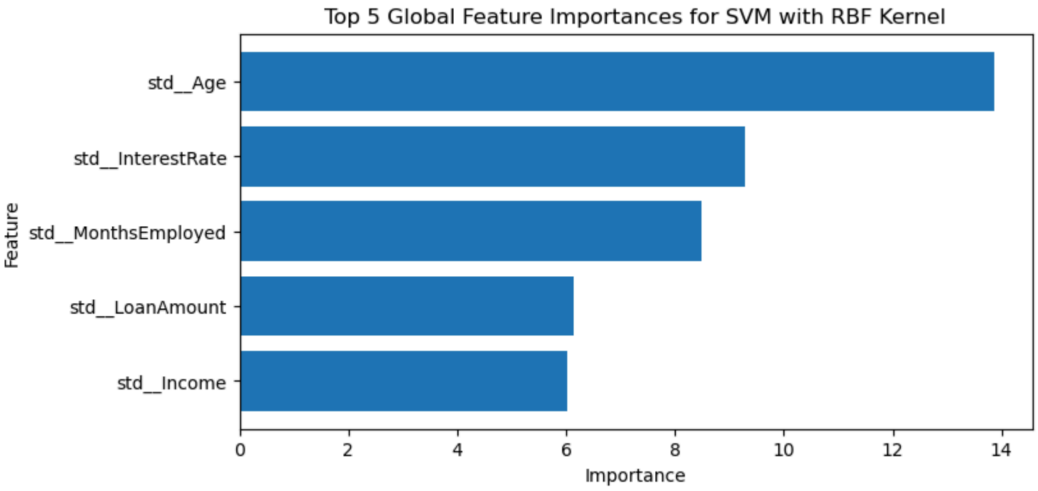


Fig 10. Model Weights Importance: SVC

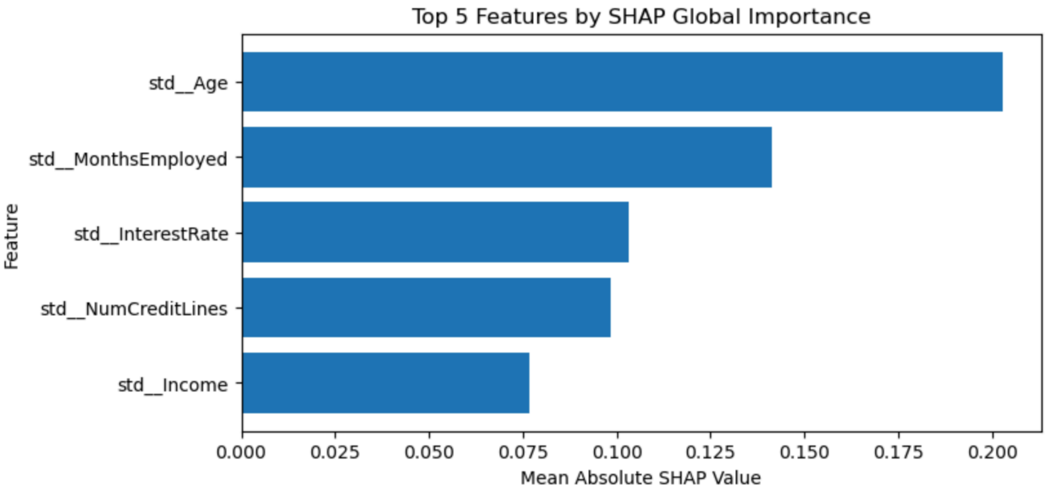


Fig 11. SHAP Global Importance: SVC

4. SHAP Local Importance:

The SHAP values for local feature importance provide deeper insights into how individual features influence the predictions of the SVC model.

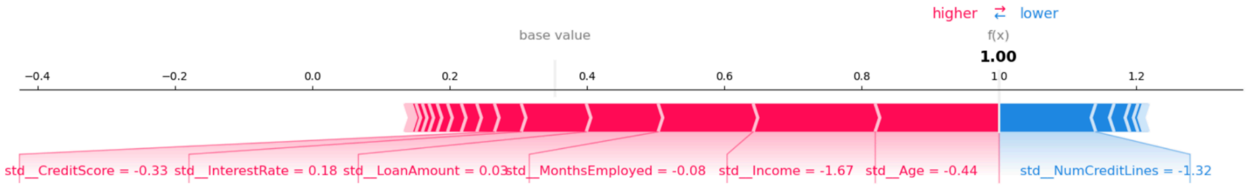


Fig 12. SHAP Force Plot: Index 50

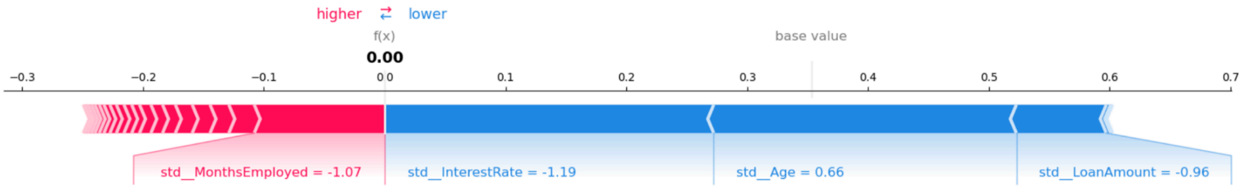


Fig 13. SHAP Force Plot: Index 500

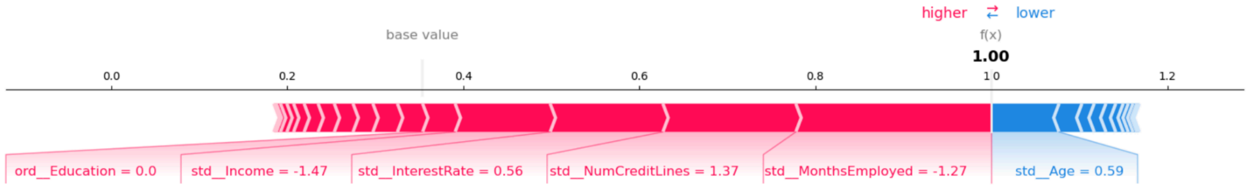


Fig 14. SHAP Force Plot: Index 5000

The force plots revealed that Age, Interest Rate, Income are consistently the most important features influencing loan default predictions. These findings align well with real-world expectations, where financial indicators such as low credit scores, high interest rates often signal higher default risk.

One unexpected finding is the limited contribution of Education, as it is often assumed that higher education levels would correlate with lower default risk. However, the SHAP values suggest that education alone does not significantly impact default prediction, emphasizing that financial behaviors and credit indicators are more predictive.

V.Outlook

To further improve the predictive power and interpretability of the model, several enhancements can be considered:

- First, hyperparameter tuning could be made more explicit by utilizing advanced optimization techniques such as Bayesian optimization, which allows for a more efficient and targeted search of the hyperparameter space.

- Feature engineering presents another opportunity for improvement. Removing the least important features, as identified through feature importance scores, could reduce noise and enhance the model's generalization capability. Additionally, further investigation into feature interactions may uncover significant relationships that can boost the model’s predictive power.

- Addressing the class imbalance remains a key challenge. Beyond adjusting class weights, alternative techniques such as oversampling the minority class and undersampling the majority class could be explored. Combining these sampling techniques with ensemble methods may further improve the model's ability to capture patterns in the minority class while mitigating bias toward the majority class.

- Finally, collecting additional data, such as detailed borrower financial histories or macroeconomic indicators, could enrich the feature set and provide more context for predicting loan defaults.

These improvements would enhance the evaluation and interpretability of the model, addressing its current limitations and providing a more comprehensive solution to each problem.

VI.References

[1] Loan Default Prediction Dataset. *Kaggle*. Available at: https://www.kaggle.com/datasets/nikhil1e9/loan-default/data.

[2] Data Science Coding Challenge: Loan Default Prediction. *Coursera*. Available at: https://www.coursera.org/projects/data-science-coding-challenge-loan-default-prediction.

[3] Previous Work: Loan Default Classification & Prediction. *Kaggle*. Available at: https://www.kaggle.com/code/adekunlesolomon/loan-default-classification-prediction.

VII.Github Repository

https://github.com/Jing-Xu1223/DATA1030-Project