**Predicting Loan Default Risk Using Machine Learning**

Tiffany Delacruz, Ewere Alizomor, Uyenmy Nguyen

University of Texas – Arlington

DATA 4380: Data Problems

Prof. Masoud A. Rostami

October 17, 2025

1. **Introduction**

Financial default and bankruptcy are critical concepts in credit risk modeling, each representing distinct stages of financial distress. Default occurs when a borrower fails to meet payment obligations, while bankruptcy is a legal process that may follow prolonged default, allowing individuals or businesses to restructure or discharge debt under court supervision.

Default prediction is not only a technical challenge but also a socially consequential task. Misclassifying high-risk borrowers can lead to financial exclusion, while failing to identify defaulters may result in systemic losses. Therefore, predictive models must balance accuracy with fairness, especially when applied to diverse populations with varying financial histories.

This report explores the predictive modeling of loan default, a precursor to bankruptcy, using supervised learning techniques. Given the inherent class imbalance, where defaults are rare but consequential, we apply resampling strategies such as SMOTE to improve model sensitivity to minority outcomes. Ensemble methods and hyperparameter tuning were used to optimize recall and F1-score for defaulters, with threshold adjustments and fairness audits included to ensure equitable performance across demographic subgroups.

The modeling pipeline includes baseline classifiers, ensemble learners, and hypertuned variants; each evaluated using precision, recall, and F1-score. Subgroup metrics across income and age are logged to assess fairness, and threshold tuning is explored to improve sensitivity without sacrificing reliability.

By distinguishing between default and bankruptcy, this report emphasizes the importance of early intervention and ethical modeling practices in financial decision-making. The goal is not only to maximize predictive performance, but also to support responsible lending, regulatory transparency, and inclusive financial access.

1. **Background & Related Work**

Accurate loan default prediction is a critical component of financial risk management, enabling institutions to minimize losses and maintain long-term stability. Traditional credit scoring models, such as logistic regression and discriminant analysis, primarily rely on static financial and demographic factors, such as income, credit score, and debt-to-income ratio—to assess borrower risk. While effective for basic credit evaluation, these conventional methods often fail to capture complex, non-linear relationships between borrower behavior and repayment outcomes (Khandani, Kim, & Lo, 2010).

In recent years, the rapid growth of digital financial data and advances in computational power have led to increased adoption of machine learning techniques in credit risk modeling. ML algorithms can uncover intricate patterns within large datasets, enabling more accurate and adaptive default predictions. However, one persistent challenge is class imbalance, where non-default borrowers significantly outnumber defaulting ones. This imbalance often causes models to overfit to the majority class, leading to high accuracy but low recall for defaulters. Addressing this issue requires specialized techniques, such as resampling and threshold optimization, to ensure balanced model performance and reliability in real-world financial decision-making.

Previous research has demonstrated that machine learning algorithms substantially improve predictive accuracy and risk differentiation in credit scoring compared to traditional statistical approaches. Khandani et al. (2010) were among the first to show that consumer credit risk models built with ML could adapt more effectively to changing borrower behavior, outperforming conventional rule-based systems. Building upon this foundation, Garg et al. (2024) conducted a comparative analysis of multiple machine learning algorithms, including Random Forest, Gradient Boosting, and XGBoost, and concluded that ensemble-based methods achieve superior predictive performance and robustness for loan default classification tasks.

1. **Methodology:**

The methodology for this project followed a structured machine learning workflow consisting of data understanding, preprocessing, model development, and evaluation. The dataset used was the Loan Default Prediction Dataset from Kaggle, comprising 255,347 records and 18 features representing borrowers’ demographic information, financial attributes, and loan characteristics. The target variable, Default, was binary-coded, where 1 represented borrower who defaulted and 0 represented those who did not.

**3.1 Data Understanding:**

The Loan Default Prediction dataset contains 255,347 records and 18 features describing borrower demographics, financial characteristics, and loan attributes. The target variable Default is binary, indicating whether a borrower defaulted (1 = Defaulted, 0 = Not Defaulted).

Data Overview:

* Numerical variables: Age, Income, LoanAmount, CreditScore, MonthsEmployed, NumCreditLines, InterestRate, LoanTerm, and DTIRatio.
* Categorical variables: Education, EmploymentType, MaritalStatus, HasMortgage, HasDependents, LoanPurpose, and HasCoSigner.

The dataset contains no missing values and exhibits consistent data types across all variables. The absence of missingness and outliers simplifies preprocessing and allows the team to focus on modeling and addressing class imbalance.

Target Distribution:

The dataset is highly imbalanced, with 88.39% non-defaults and 11.61% defaults. This imbalance can cause predictive models to favor the majority class, leading to poor recall for actual defaulters. Consequently, resampling methods such as SMOTE and class weighting were planned to counteract bias.

Key Patterns:

* Income vs. Default: Lower-income borrowers have a higher chance of default.
* Age vs. Default: Younger individuals tend to default more often than older borrowers.
* LoanAmount vs. Default: Default frequency increases slightly with larger loan amounts, but non-defaults remain dominant.
* Feature Correlation: Most numerical features show weak linear correlation with default, the highest being Age (−0.17) and InterestRate (0.13).

Overall, the dataset is clean and well-structured, though the imbalance requires deliberate mitigation before modeling.

**3.2 Data Preprocessing**

Preprocessing was conducted to prepare the dataset for machine-learning models and ensure numerical compatibility across all features.

Steps Taken

1. Feature Segregation:  
   The data was split into numerical and categorical groups to apply transformations accordingly.
2. Handling Missing Values:  
   Since no missing data existed, no imputation was necessary.
3. Encoding Categorical Variables:  
   All categorical variables (Education, EmploymentType, MaritalStatus, etc.) were converted to numeric using one-hot encoding, creating binary columns for each category.
4. Feature Scaling:  
   Numerical features such as Income, LoanAmount, and DTIRatio were standardized to improve model convergence and prevent bias toward features with larger numeric ranges.
5. Data Integration:  
   The cleaned numerical and encoded categorical data were merged into one unified numeric matrix suitable for machine-learning algorithms.
6. Train-Test Split:  
   The dataset was divided into 80% training and 20% testing subsets to evaluate model generalization on unseen data.
7. Handling Class Imbalance:  
   SMOTE (Synthetic Minority Over-Sampling Technique) was applied to the training set, increasing the proportion of default cases and improving the model’s ability to detect minority-class patterns.

**3.3 Modeling:**

The modeling phase aimed to build, compare, and evaluate several classification algorithms to predict loan default. The focus was to observe how each model performed before and after applying SMOTE (Synthetic Minority Over-Sampling Technique) to handle class imbalance. Three algorithms were selected: Logistic Regression, Random Forest, and Gradient Boosting.

**Logistic Regression:**

Logistic Regression served as the baseline model because of its simplicity and interpretability. Before applying SMOTE, the model achieved a relatively high accuracy of about 88%, but this result was misleading due to the imbalance in the dataset. The model correctly identified most non-default cases but failed to capture a large portion of true defaulters, resulting in very low recall.

After applying SMOTE, the model’s recall improved significantly, showing a stronger ability to detect defaulting borrowers. Precision decreased slightly, which means the model produced more false positives, but this was an acceptable trade-off since the goal was to reduce missed defaults. Adjusting the decision threshold to 0.346 helped balance this trade-off between recall and precision, resulting in an overall more reliable risk detection model.

**Random Forest:**

The Random Forest model performed strongly even before SMOTE, reaching an accuracy near 88.6%. However, like Logistic Regression, it tended to favor the majority class (non-default), making it less effective for identifying borrowers likely to default.

Once SMOTE was applied, the Random Forest model’s recall improved noticeably while maintaining nearly the same overall accuracy. This demonstrated that the model became more balanced detecting more default cases without sacrificing overall performance. Its ROC-AUC score also increased, indicating better discrimination between default and non-default borrowers. Overall, Random Forest proved to be a robust and stable model that balanced predictive power and generalization.

**Gradient Boosting:**

Gradient Boosting showed similar behavior to Random Forest but offered slightly higher model discrimination after SMOTE was applied. Before balancing, it exhibited high accuracy but low recall, underperforming on the minority class. After introducing SMOTE, recall improved substantially, allowing the model to detect more defaulters while still maintaining reasonable precision.

Although Gradient Boosting required more computational time to train compared to Random Forest, it provided consistent results with a smoother trade-off between precision and recall. Its ROC-AUC improvement reflected better capability in distinguishing between risky and safe borrowers, making it a strong candidate for deployment in a credit risk context.

**Overall Comparison and Insights:**

Before SMOTE, all models appeared accurate but were biased toward predicting non-defaults, meaning they underperformed in identifying actual defaulters. After SMOTE, recall and F1-scores improved across all models, proving that balancing the dataset was essential for fair and reliable predictions. Logistic Regression showed the most dramatic recall improvement, Random Forest maintained the best balance between accuracy and recall, and Gradient Boosting demonstrated the highest overall discriminative power.

From a business perspective, high recall is preferred in loan risk prediction because failing to identify a potential defaulter (a false negative) is far costlier than incorrectly flagging a safe borrower. Therefore, the post-SMOTE model, especially Random Forest and Gradient Boosting, offered a better foundation for minimizing financial losses while still maintaining decision efficiency.

1. **Challenges & Solutions**

A primary challenge encountered in this project was achieving an optimal balance between recall and precision in predicting loan defaults. The dataset exhibited a significant class imbalance, with non-default cases constituting the majority of observations. This imbalance led the models to favor the majority class, resulting in high accuracy but poor detection of true defaulters. To address this issue, both oversampling and undersampling techniques were employed to improve the model’s ability to learn from the minority class. Specifically, the SMOTE (Synthetic Minority Oversampling Technique) method was used to generate synthetic samples for the minority class, allowing the model to recognize a broader range of default patterns, while undersampling helped reduce bias from the majority class by balancing the dataset. However, these methods introduced trade-offs, some techniques would increase recall but lower precision or vice versa. To mitigate these effects, threshold tuning was implemented to adjust the probability cutoff used for classification, optimizing the trade-off between precision and recall. This adjustment helped enhance the F1-score, which represents the harmonic mean of precision and recall, thereby improving the model’s overall performance. Ultimately, the team decided to prioritize F1, as correctly identifying potential defaulters and non-default was deemed more critical from a business and risk management aspect.

1. **Key Results & Insights**

From a data science perspective, the final model Logistic Regression with SMOTE and threshold tuning achieved an accuracy of 0.802, precision of 0.286, recall of 0.477, F1-score of 0.360, and ROC–AUC of 0.75. These results demonstrate that while overall accuracy decreased slightly compared to models such as Random Forest (which achieved approximately 0.89), the tuned Logistic Regression model provided a more balanced performance across both classes. The use of SMOTE improved the model’s sensitivity to the minority class by generating synthetic examples of defaults, while threshold tuning adjusted the decision boundary to achieve a better balance between precision and recall. This combination led to a higher F1-score, indicating improved harmonic performance between the two key metrics. Ultimately, this approach enhanced the model’s ability to identify default cases that were previously misclassified, improving recall without overly compromising precision.

From a business standpoint, these results have meaningful implications for financial risk management. A higher recall rate (0.477) means the model is better at identifying borrowers who are likely to default, thereby helping institutions take preventive actions such as adjusting loan terms, requiring additional documentation, or implementing stricter credit checks. Although precision was moderate, the focus on recall aligns with the strategic goal of reducing financial losses associated with missed defaulters. Misclassifying a defaulter as a safe borrower (a false negative) poses a much greater financial risk than incorrectly flagging a reliable borrower as risky (a false positive). Therefore, prioritizing recall allows the organization to manage credit risk more effectively, even if it results in a higher rate of unnecessary loan rejections. This balance between predictive accuracy and financial prudence reflects a data-driven approach to decision-making in lending operations. After addressing the identified challenges, the final models were evaluated to assess their predictive performance and business implications.

1. **Conclusion & Next Steps**

This study aimed to develop predictive models capable of identifying borrowers at risk of loan default based on their financial and demographic characteristics. Through the application of various machine learning algorithms, it was observed that traditional models such as Random Forest and Gradient Boosting achieved high accuracy but performed poorly in identifying defaulters due to the pronounced class imbalance within the dataset. To address this limitation, the Logistic Regression model was enhanced through the use of SMOTE for oversampling and threshold tuning for optimizing the decision boundary. This combination improved the model’s ability to detect true defaults, resulting in a more balanced performance across evaluation metrics. The final model achieved an accuracy of 0.802, recall of 0.477, precision of 0.286, and an F1-score of 0.360, reflecting a meaningful improvement in the detection of high-risk borrowers. These findings underscore the importance of addressing data imbalance and adjusting classification thresholds to align predictive modeling objectives with the risk mitigation priorities of financial institutions.

Moving forward, the next phase of this project will focus on enhancing both the predictive performance and practical applicability of the model. Future work will involve experimenting with advanced ensemble algorithms, such as XGBoost, LightGBM, and CatBoost, to capture more complex, non-linear relationships in the data and further improve recall without substantially compromising precision. Additional feature engineering will be conducted to include behavioral and temporal variables, such as debt-to-income ratio trends, payment history consistency, and credit utilization patterns, which may provide deeper insights into borrower risk. To address potential model bias and improve transparency, explainable AI (XAI) techniques, such as SHAP and LIME, will be explored to interpret the influence of key features on prediction outcomes. Moreover, incorporating cost-sensitive learning can ensure that the model reflects real-world lending priorities by assigning higher penalties to false negatives.

**Citations**

Garg, Ashish and Maheshwari, Ayush and Kapoor, Rameshwari and Bhanot, Nilesh, A Comparative Analysis for Predicting Loan Default Risks using Machine Learning Algorithms (November 15, 2024). Proceedings of the 3rd International Conference on Optimization Techniques in the Field of Engineering (ICOFE-2024), Available at SSRN: https://ssrn.com/abstract=5065367 or <http://dx.doi.org/10.2139/ssrn.5065367>

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. <https://doi.org/10.1016/j.jbankfin.2010.06.001>

Zhang X, Zhang T, Hou L, Liu X, Guo Z, Tian Y, Liu Y. Data-Driven Loan Default Prediction: A Machine Learning Approach for Enhancing Business Process Management. Systems. 2025; 13(7):581. <https://doi.org/10.3390/systems13070581>