Predicting Loan Default Risk Using Machine Learning

Geremy Paige
Colorado School of Mines
gpaige@mines.edu

December 2024

1 Introduction

Financial institutions face significant challenges with loan defaults, which result in major losses. Predicting the likelihood of loan default is crucial to help mitigate risks and make informed lending decisions for everyone. Traditional credit scoring methods often fail to capture complex relationships in data, which leads to suboptimal risk assessments. This project aims to develop a machine learning model capable of predicting loan default using historical data and leveraging advanced algorithms to improve accuracy and recall.

Using random forest, gradient boost, and decision tree models, this study explores the potential of machine learning in predicting credit risk. The primary objective is to achieve high recall, minimizing the risk of missing high-risk borrowers while maintaining a balance between precision and accuracy. This ensures that financial institutions can manage risk proactively and improve decision-making processes.

2 Dataset Description

The dataset used for this study contains 1,000 records with 17 features related to loan applications. These include both numerical and categorical variables, such as:

- **Checking Balance:** Current account balance categorized as less than 0 DM, between 1 and 200 DM, or greater than 200 DM. Here, **DM** refers

to *Deutsche Mark*, the former official currency of Germany before the introduction of the Euro.

- **Loan Duration:** Loan duration in months.
- **Credit History:** Historical credit performance.
- **Loan Amount:** The amount of the loan.
- **Default Status:** Target variable indicating whether the borrower defaulted.

2.1 Exploratory Data Analysis

Initial analysis revealed that most of the borrowers had low account balances (less than 0 DM) and a significant portion of the loans had durations between 12 and 24 months. The default rate was approximately 30%, highlighting an imbalance in the dataset.

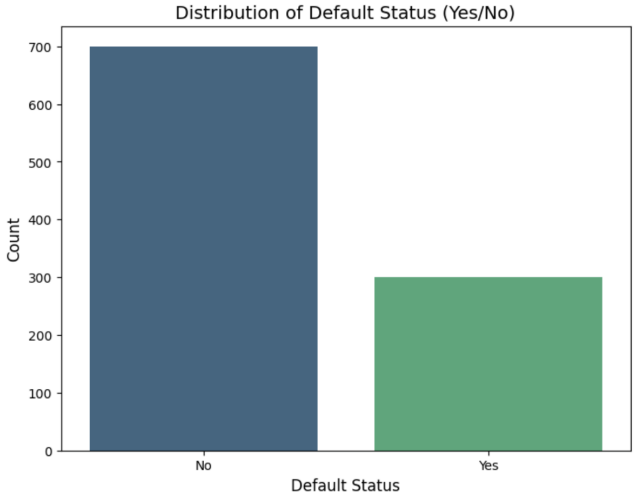

Figure 1: Distribution of Default Status (Yes/No).

2.2 Preprocessing Steps

To prepare the data for modeling, I utilized the following steps:

- **Encoding:** Categorical features were encoded using label encoding.

- **Scaling:** Numerical features were standardized to ensure uniformity.
- **Balancing:** The Synthetic Minority Oversampling Technique (SMOTE) was used to address class imbalance, improving the model's ability to detect defaults.

3 Methodology

3.1 Model Selection

Three machine learning models were tested:

- **Decision Tree:** Known for interpretability but prone to overfitting, particularly on imbalanced datasets.
- **Random Forest:** Handles non-linear relationships and reduces overfitting by combining multiple decision trees.
- **Gradient Boosting:** Focuses on minimizing errors iteratively, often achieving high accuracy on structured data.

3.2 Training and Validation

The dataset was split into training (70%) and testing sets (30%) to evaluate the performance of the model. Hyperparameter tuning was performed using grid search to optimize parameters such as tree depth and learning rate. Cross-validation ensured that the models generalized well to unseen data.

3.3 Evaluation Metrics

Performance was evaluated using:

- **Accuracy:** Overall correctness of the predictions.
- **Precision:** Ratio of correctly predicted positives to total predicted positives.
- **Recall:** Ability to capture all positive cases (high recall is critical for this task).
- **F1-Score:** Balance between precision and recall, providing a single metric for performance evaluation.

4 Results

The Random Forest model achieved the best performance among the tested models:

- **Accuracy:** 82.38%
- **Precision:** 78.57%
- **Recall:** 87.13%
- **F1-Score:** 82.63%

4.1 Confusion Matrices

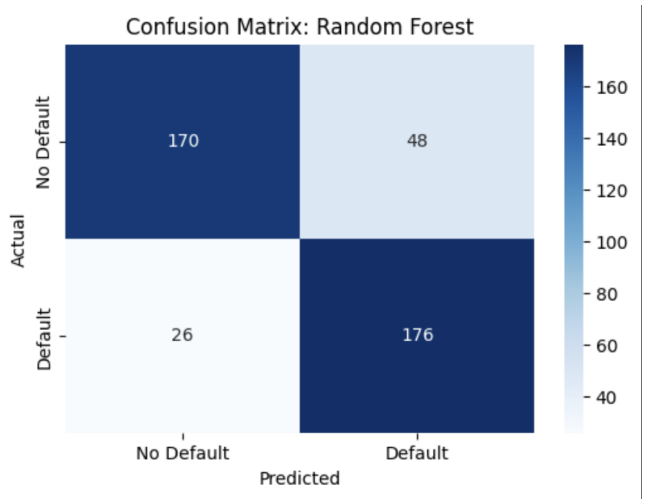

Figure 2: Confusion Matrix: Random Forest.

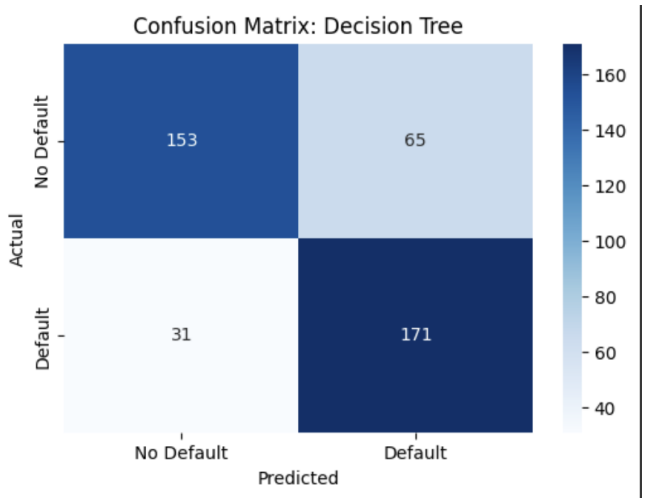

Figure 3: Confusion Matrix: Decision Tree.

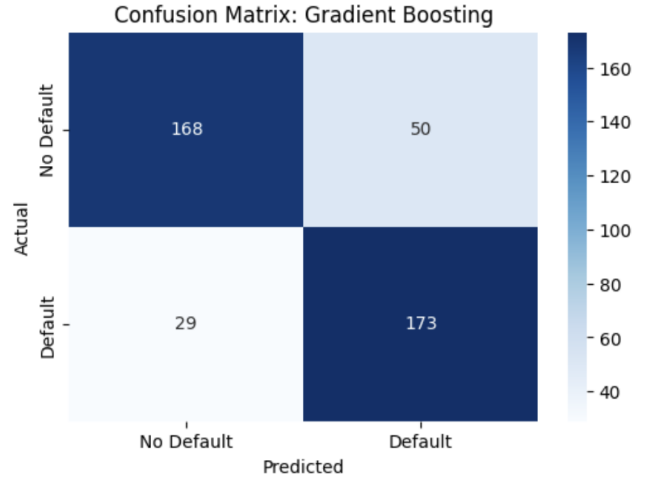

Figure 4: Confusion Matrix: Gradient Boosting.

4.2 Model Comparison

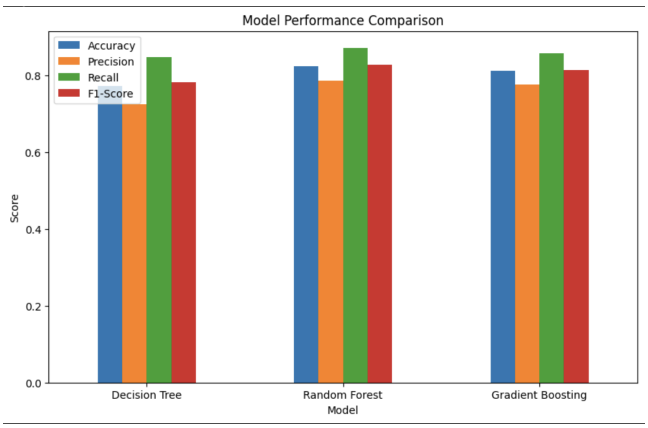

Figure 5: Model Performance Comparison.

4.3 Feature Importance

The feature importance analysis highlighted the most influential predictors, including:

- *Checking Balance:* Strong indicator of financial stability.
- *Loan Duration:* Longer durations correlated with higher default risk.
- *Loan Amount:* Higher loan amounts increased default likelihood.

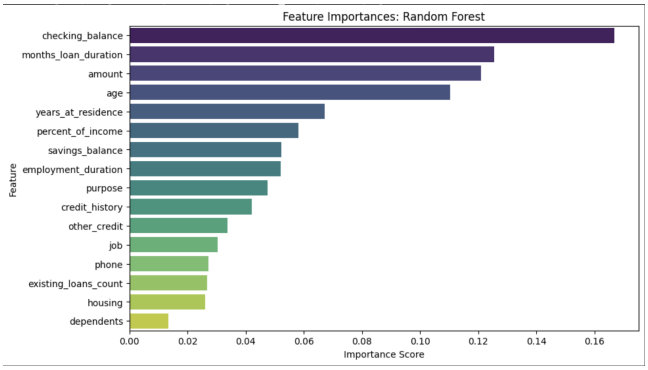

Figure 6: Feature Importances: Random Forest.

5 Conclusion

The Random Forest model was identified as the best-performing algorithm for predicting loan default, achieving high recall (87.13%) and balanced performance across all metrics, including accuracy (82.38%) and F1-Score (82.63%). This makes it a suitable choice for financial institutions aiming to minimize the risk of loan defaults while maintaining a balance between identifying high-risk borrowers and avoiding false positives.

This report highlights a foundation for using machine learning to enhance traditional credit scoring methods, making them more predictive and actionable for financial institutions. The insights gained from this study can be applied to real-world financial scenarios to reduce losses and improve lending strategies.

Future Work:

- **Incorporate external data sources:** Future studies could integrate external credit bureau data, macroeconomic indicators, or other financial metrics to improve model predictions. These additional data points could provide a broader context for credit risk assessment and enhance the robustness of the model.
- **Real-world deployment:** Deploying the Random Forest model in a real-world financial setting would provide an opportunity to evaluate its performance in live scenarios. Continuous monitoring and retraining of the model with fresh data can help maintain accuracy and adapt to changing borrower behavior and economic conditions.
- **Explore advanced modeling techniques:** Future research could experiment with deep learning models, such as neural networks, to capture more complex patterns in the data. Techniques like Long Short-Term Memory (LSTM) networks or transformer models could be used to account for temporal trends in credit risk over time.
- **Enhance interpretability:** While Random Forest offers feature importance, more transparent methods like SHAP (Shapley Additive Explanations) values or LIME (Local Interpretable Model-agnostic Explanations) could be explored to provide financial institutions with actionable insights and greater confidence in the model's decisions.
- **Cost-sensitive learning:** Future iterations of this study could incorporate cost-sensitive learning techniques to minimize the financial implica-

tions of false positives and false negatives, tailoring the model's performance to align with institutional risk tolerance.

Acknowledgment

The dataset used for this study was sourced from Kaggle, a valuable platform for data science and machine learning projects. The availability of this well-structured dataset helped with the development and evaluation of predictive models for the loan default risk. I sincerely acknowledge Kaggle for providing a rich repository of datasets and resources that greatly contributed to the success of this study.