Predicting Loan Defaults: A Comparative Analysis of Random Forest and Gradient Boosting Models

1. Introduction:

In the lending industry, assessing loan default is a significant concern that directly influences financial stability and institutional decision-making. Traditionally, criteria such as credit scores, income, and employment status are utilized to evaluate applicants. However, accurately predicting loan repayment versus default remains a challenging and typically imbalanced classification issue. This project seeks to enhance loan risk assessment accuracy using machine learning techniques applied to publicly available LendingClub data. We employed extensive preprocessing steps, including handling missing data and class imbalance via weighted training. Random Forest (RF) and Gradient Boosting (GB) models were developed to predict loan outcomes, aiming for interpretability and effective classification.

2. Methods

We utilized LendingClub data comprising over two million loan applications, including borrower income, loan amount, term, interest rate, grade, employment length, verification status, home ownership, loan purpose, debt-to-income (DTI) ratio, and FICO scores. The analysis specifically focused on predicting whether loans would be fully paid or charged off (defaulted).

Data preprocessing involved removing ambiguous records, retaining only clear outcomes (Fully Paid or Charged Off), handling missing values by discarding incomplete records, and numerically encoding categorical and textual information.

Initially, we trained a baseline Random Forest model due to its capability of handling diverse feature types, robustness against overfitting, and interpretability. Subsequently, we trained a Gradient Boosting model known for capturing intricate patterns through iterative refinement. The dataset exhibited significant imbalance, with approximately 80% fully paid loans versus 20% charged-off loans (figure 1). To address this imbalance, class-weight adjustments, oversampling, and undersampling were tested. Undersampling emerged as the most effective strategy, significantly improving the models' predictive power.

3. Results and Discussion

Despite similar overall performance metrics, both models experienced challenges, particularly in achieving high precision (Table 2 and 3). Though recall was satisfactory, precision was notably low, indicating many false-positive predictions.

Importantly, the feature importance differed significantly between the two models:

- **Random Forest:** Highlighted DTI, loan amount, annual income, and interest rate as critical. These findings are consistent with practical banking experience, recognizing that high DTI and larger loan amounts correlate with repayment difficulties.
- **Gradient Boosting:** Dominantly emphasized interest rates as the critical predictor, significantly more than other factors. This aligns well with financial risk management strategies, given that higher interest rates generally reflect elevated lending risk.

Given the disparity, logistic regression was conducted as an additional validation step. Results from logistic regression closely mirrored Random Forest findings, lending further credibility to its broader and more nuanced perspective.

From a practical banking viewpoint, the model selection depends upon institutional risk management priorities. Gradient Boosting offers higher recall, making it particularly suitable for risk-averse banks, and provides a clear focus on interest rates, thus facilitating straightforward adjustments in pricing and lending strategies. In contrast, Random Forest provides multi-dimensional insights into loan defaults, which is advantageous for comprehensive risk management, supporting nuanced borrower evaluations that consider multiple financial attributes.

Considering the models' limitations (particularly low precision), a combined strategic approach is advisable. Initial filtering by Gradient Boosting based on interest rate thresholds could efficiently highlight potentially risky loans. Detailed reviews using Random Forest insights would subsequently provide comprehensive borrower assessments. From a management perspective, the significant importance of interest rates identified by Gradient Boosting emphasizes the need for careful evaluation of loan pricing policies. Conversely, the multi-faceted risk profile offered by Random Forest underscores detailed evaluations of borrower income stability and debt management practices.

4. Conclusion

This comprehensive analysis illustrates that Random Forest and Gradient Boosting, when combined with undersampling techniques, offer valuable but complementary insights into predicting loan defaults. Random Forest’s multi-factorial evaluation provides nuanced risk assessments, whereas Gradient Boosting's focused prediction on interest rates can direct clear-cut strategic policy adjustments. A hybrid model leveraging both strengths could optimally balance precision and recall, enhancing practical application. Future research could further improve model precision by exploring hybrid methodologies and advanced feature engineering approaches.

References

Wordsforthewise. 2020. *Lending Club Loan Data*. Kaggle. https://www.kaggle.com/datasets/wordsforthewise/lending-club.

Appendix

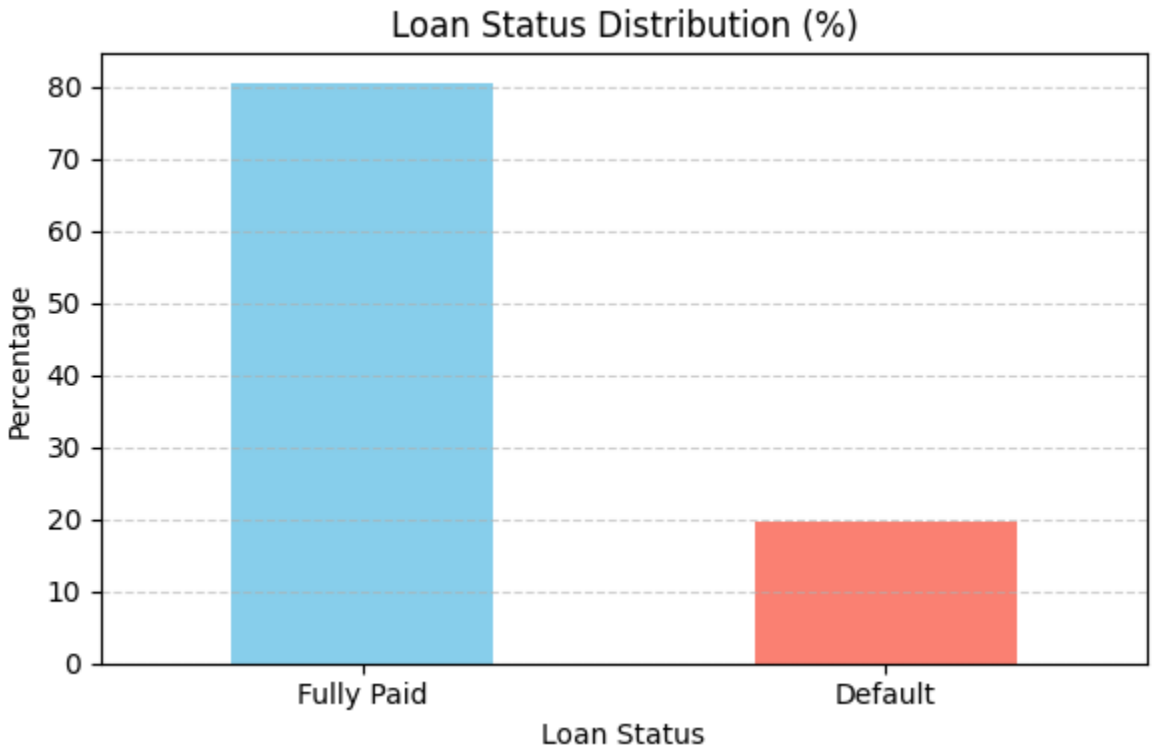

Figure 1: Imbalance Distribution Between Fully-paid and Default Cases

Metric	Precision	Recall	F1-score	Support
Class 0 (Fully Paid)	0.82	0.98	0.89	204,097
Class 1 (Default)	0.48	0.09	0.15	49,260
Accuracy	–	–	0.8	253,357
macro avg	0.65	0.53	0.52	253,357
weighted avg	0.75	0.8	0.75	253,357

Table 1: Random Forest Classifier

Metric	Precision	Recall	F1-score	Support
Class 0 (Fully Paid)	0.88	0.63	0.74	204,097
Class 1 (Default)	0.30	0.65	0.41	49,260
Accuracy	–	–	0.64	253,357
macro avg	0.59	0.64	0.58	253,357
weighted avg	0.77	0.64	0.67	253,357

Table 2: Random Forest Classifier (Undersampling)

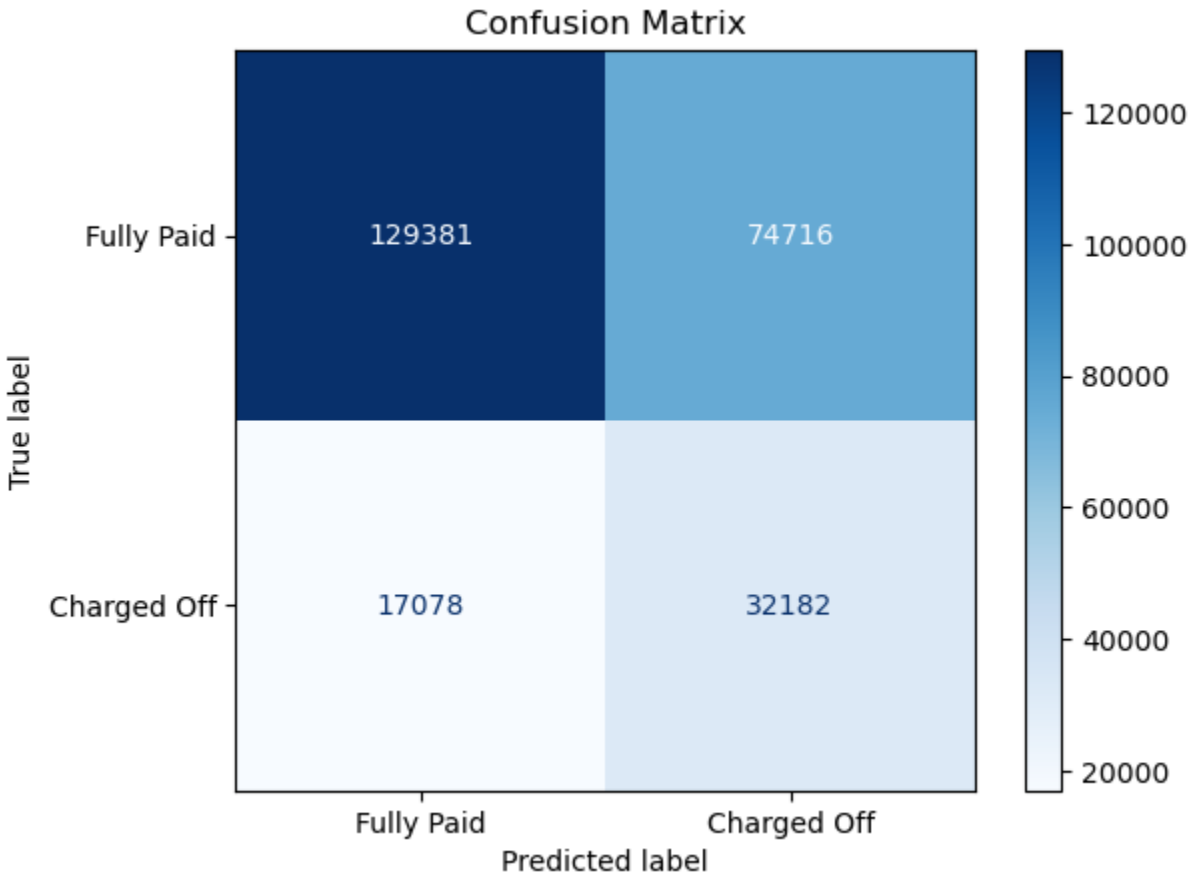

Figure 2: Confusion Matrix by Random Forest (Undersampling)

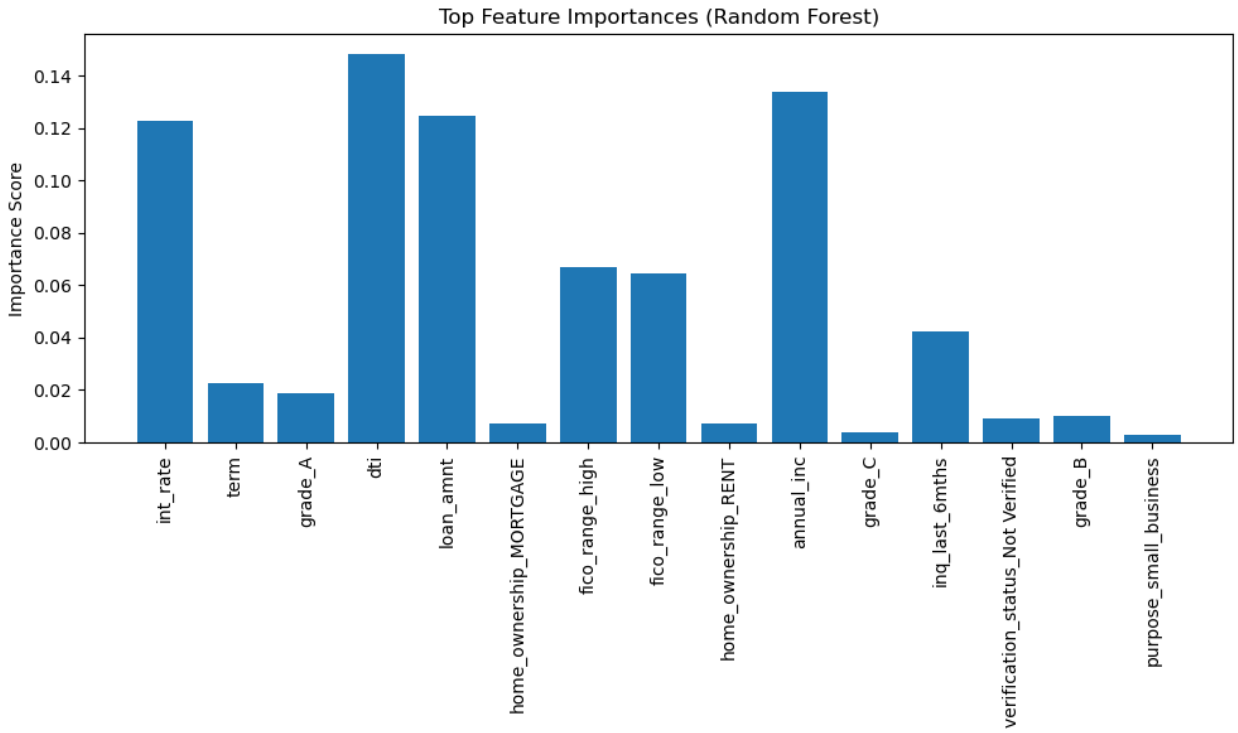

Figure 3: Top Important Factors to Default by Random Forest (Undersampling)

Metric	Precision	Recall	F1-score	Support
Class 0 (Fully Paid)	0.89	0.64	0.74	204,097
Class 1 (Default)	0.31	0.67	0.42	49,260
Accuracy	–	–	0.64	253,357
macro avg	0.6	0.65	0.58	253,357
weighted avg	0.78	0.64	0.68	253,357

Table 3: Gradient Boosting Classifier

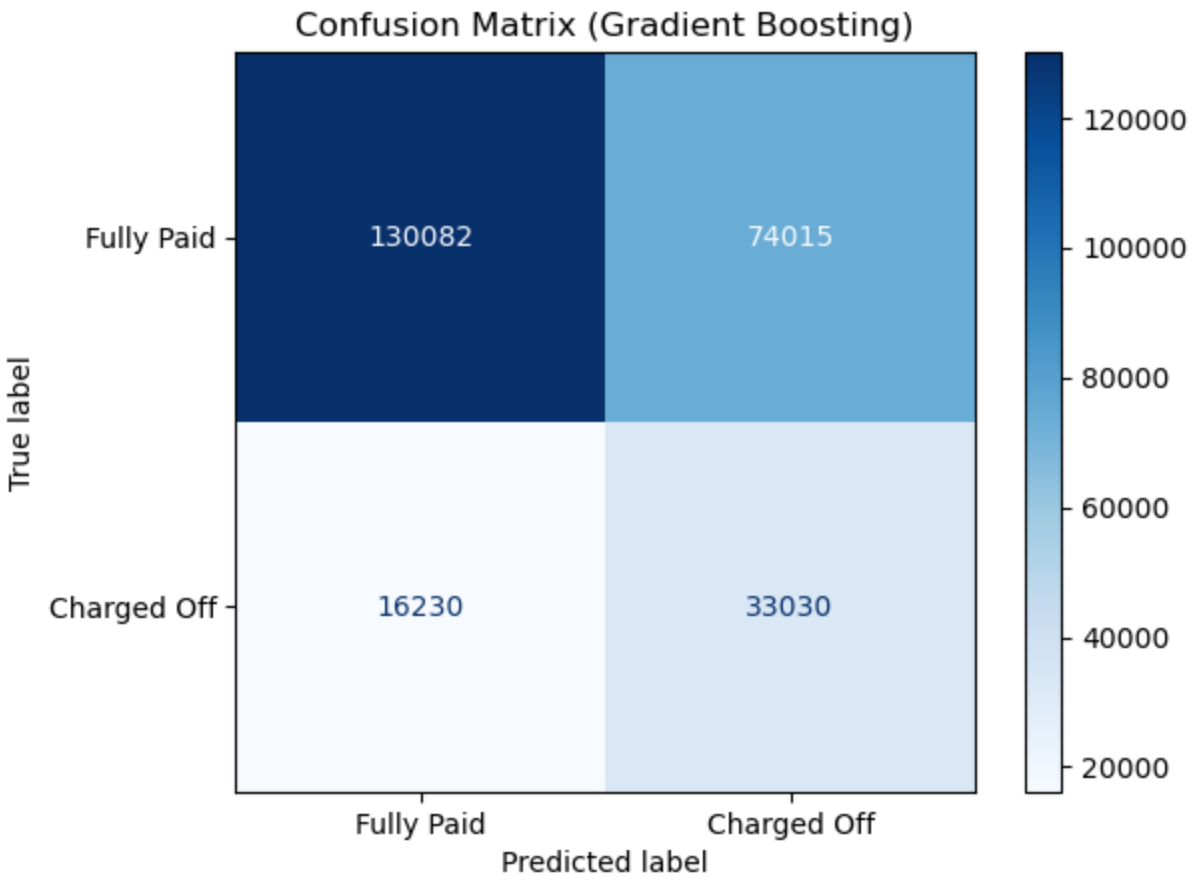

Figure 4: Confusion Matrix by Gradient Boosting

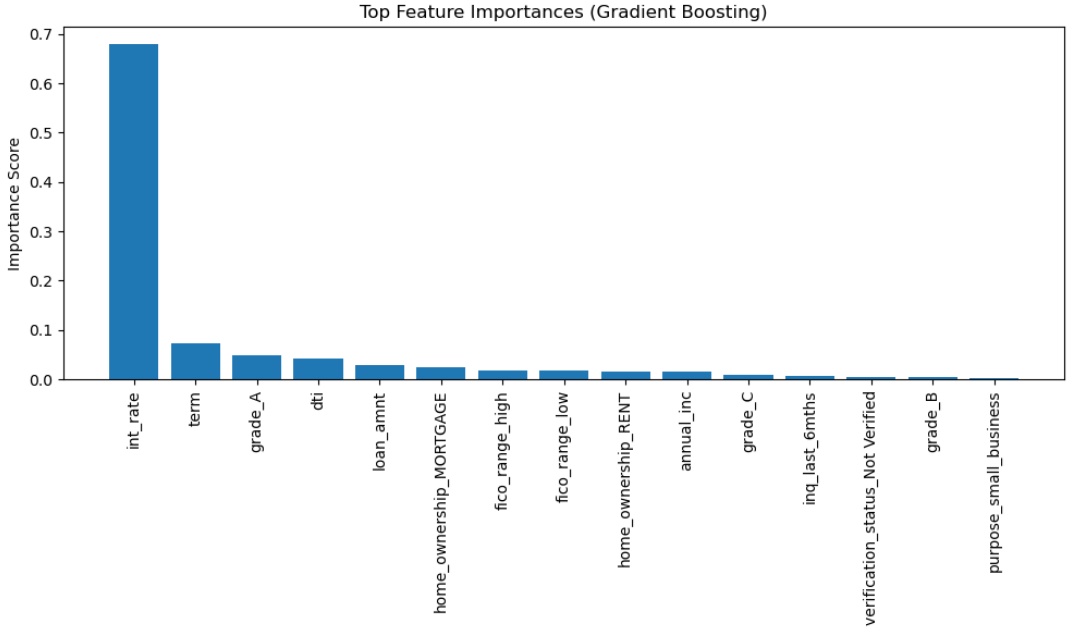

Figure 5: Top Important Factors to Default by Gradient Boosting