Loan Delinquency Analysis

Business Analysts

Israel Davidson, Fiona Huang, Melanie Jalbert, Ying Lin Zhao

Abstract

With rising interest rates, inflation, and post-pandemic economic uncertainty, loan delinquency has become an increasingly urgent issue in the U.S. credit landscape. This project aims to identify key predictors of 90-day loan delinquency using both macroeconomic and microeconomic factors across multiple loan types—including credit card, auto, mortgage, student, and personal loans. We applied and compared three supervised categorical models: Logistic Regression, Random Forest Classifier, and HistGradientBoostingClassifier. Given the imbalanced nature of our dataset, we applied resampling methods and then evaluated the effect of downsampling and upsampling techniques on model performance. Among the evaluated models, the upsampled HistGradientBoostingClassifier achieved the best trade-off between precision and recall. With an accuracy of 97% and an F1-score of 35%, this model outperformed the others and had a more consistent performance, solidifying it as the best model.

Introduction

With the rise of living costs due to inflation and higher interest rates after the COVID-19 Pandemic, loan delinquency rates, the ratio of the dollar amounts of loans with late payments to the total dollar amount of all loans, have also been on the rise. Borrowers struggle with lowered credit scores, legal actions, and increased financial stress from the higher loan interest rate and lenders face liquidity problems and increased management costs. Consequently, understanding which factors predict loan delinquency is crucial for making better lending decisions and minimizing risk. Therefore, we raised a few questions: What are some characteristics of consumers that have a high delinquency rate? Given both consumer credit data and macroeconomic data, can we predict the probability of having a 90-day loan delinquency?

The Center for Microeconomic Data of the Federal Reserve Bank of New York conducts the Survey of Consumer Expectations (SCE), which collects monthly online surveys from a random sample of about 1,300 household heads about their credit situation, expectations for income changes, and expectations for household spending changes. The Credit Access Survey contains information about the types of loans the consumers possessed, their current balances, and their credit scores. The types of loans considered in this study include: credit, auto, mortgage, student, home-based (loans used towards renovation or fixing) and personal loans. From this dataset, we

know whether or not the consumer has a 90-day loan delinquency. We found that most consumers do not have a 90-day loan delinquency, which implies that the data is highly unbalanced. We also have data about macroeconomic factors such as inflation rate, interest rate, and GDP per capita. We web-scraped these data from other sources to join them into the dataset as we would like to explore the effect of macroeconomic factors on the 90-day loan delinquency rate.

To explore this question, we implemented and compared three supervised categorical models: Logistic Regression, Random Forest Classifier, and HistGradientBoostingClassifier. We used these models to predict the probability of 90-day late loan delinquency based on a set of borrowers' credit profile features, such as credit card balance, credit card score, and student loan balance, and macroeconomic variables like inflation rate, interest rate, and GDP per capita. Logistic regression, Random Forest and HistGradientBoosting are all popular statistical techniques for predicting binary outcomes such as yes or no. HistGradientBoostingClassifier performs the best, achieving an accuracy of 97%. This model can help identify customers who are likely to become delinquent.

However, our analysis has several limitations. For instance, the model is better at predicting non-delinquent loans than identifying delinquent ones. It is not recommended to use this model to identify customers who will pay on time. Therefore, while the model can help identify more low-risk loans, we advise caution in relying solely on it to detect high-risk customers.

The rest of this report is structured as follows: we begin by describing the dataset and the exploratory approaches we performed in Data Description and Preprocessing. Next, we detail the construction and evaluation of our model in Analysis/Methodology. In Results, we assess the model's performance and interpret its implications for loan risk assessment. We conclude with a summary of findings and recommendations.

Literature Review

With the rising concern about loan delinquency in the US, many studies have explored the macroeconomic factors contributing to it, its consequences, and the broader implications for borrowers and lenders. For instance, during times of high interest rates, lenders charge more for loans, increasing the total debt value and making it more difficult for borrowers to repay their loan payments. When inflation rates are high, the return value decreases for lenders, and the higher cost of living hinders borrowers' ability to repay, leading to higher delinquency rates (Nigmonov et al., 2021). Furthermore, a high unemployment rate, which indicates widespread job loss and lack of stable income, is also correlated with high loan delinquency. In addition to macroeconomic factors, studies have shown that demographic factors such as age, marital status, income level, and credit score are significant predictors of financial stress and loan delinquency rates, with younger, lower-income, and lower credit score borrowers being associated with higher delinquency rates (Chong, 2021).

These previous studies motivated our topic and our study, for we used both theory and study findings to narrow the survey questions we used in our analysis. Looking to build on these findings, our study considers both macroeconomic factors such as inflation and interest rates with microeconomic factors such as credit scores, types of loans owned, and whether an individual maxed out their credit allocation. By considering both perspectives, it could better inform loaner decision-making.

Data Description & Exploration

Our dataset contains information from 34,362 survey responses to a questionnaire consisting of 25 credit- and loan-related questions. Responses were collected between 2015 and 2023, and participants were randomly selected to participate. Each record in the dataset represents an individual's responses regarding their loan balances, types of loans held, credit score, and other credit-related information for a specific month and year.

Based on our literature review, we reduced the original 140 variables– which included dummy variables derived from multi-select questions– to 14 variables in order to minimize noise and focus on factors likely to influence delinquency. The selected variables include, for example, whether an individual believes they could access $2,000 in an emergency, whether they used their credit allocation for the month, and their credit score.

The dataset includes two delinquency-related variables: one indicating whether the individual had a payment 30+ days late in the past 12 months, and another indicating whether they had a payment 90+ days late. Since a 90-day or more late payment is considered more serious and provides deeper insight into financial distress, we chose the 90-day late payment variable as our response variable.

In addition to the survey data, we web-scraped macroeconomic indicators– GDP, GDP per capita, interest rate, and inflation rate– from sources such as the U.S. Bureau of Labor Statistics. These indicators were merged with the survey responses using the corresponding month and year. Table 1 in the appendix provides a complete list of the covariates used in this analysis.

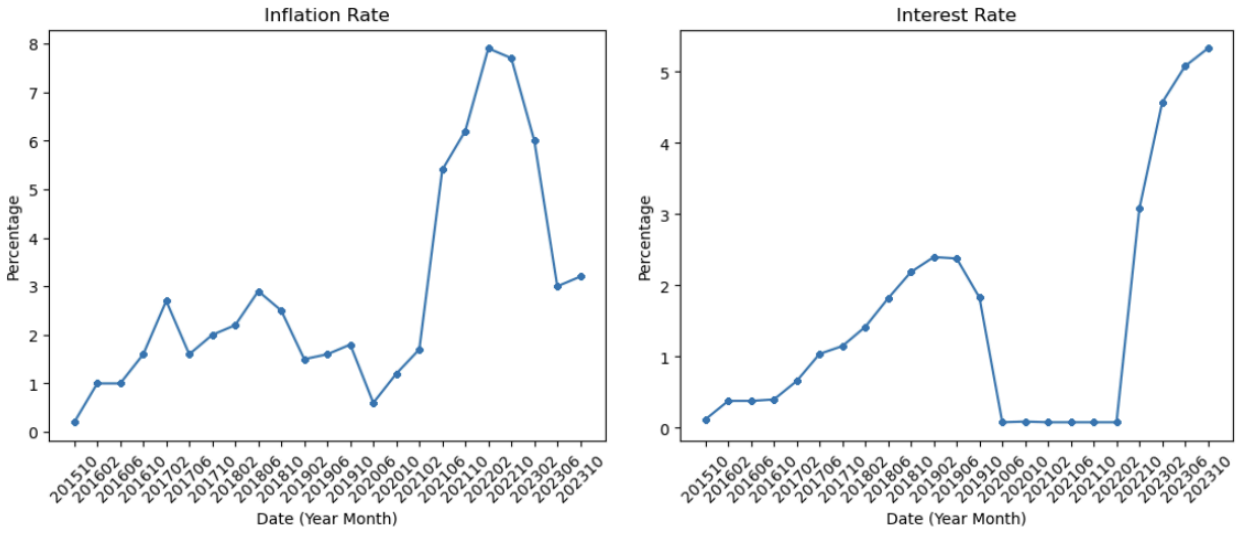

Figure 1a and 1b. Line plot of interest rates and inflation rates by month. Notice the sudden spike in the inflation rate in October 2020, likely a consequence of COVID-19 relief programs aimed at supporting individuals who lost their jobs or were struggling financially. In addition, the Federal Reserve (Fed) lowered the federal funds rate range to approximately 0%–0.25%, which reduced interest rates and further fueled inflation. Throughout 2022, the Fed raised interest rates in an effort to combat high inflation, leading to a sharp increase in interest rates beginning in February 2022.

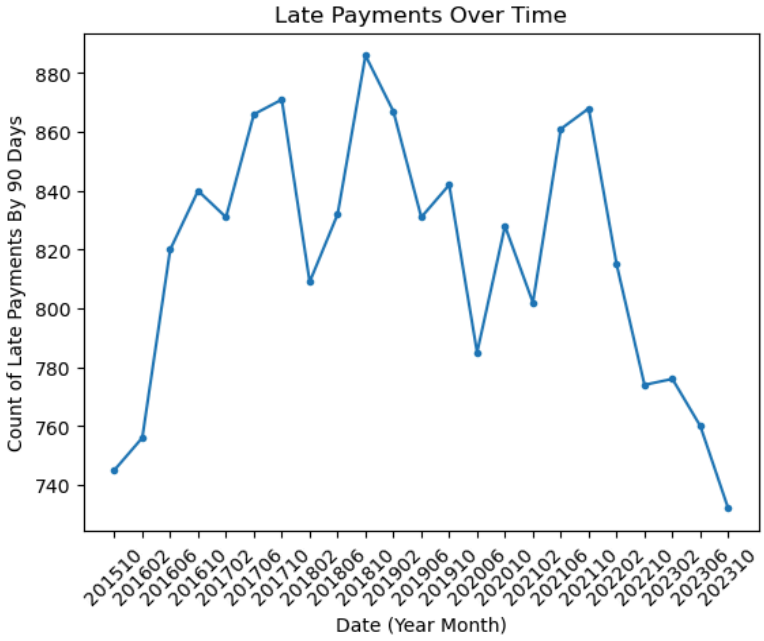

Figure 2. Line plot of the number of individuals with a 90+ day late payment, aggregated by month and year. Notice how the delinquency count correlates with interest and inflation rates. Soon after the rapid increase in the inflation rate in October 2020, we observe a spike in

delinquency counts. In 2022, when the Fed raised interest rates, a decreasing trend in delinquency counts was observed.

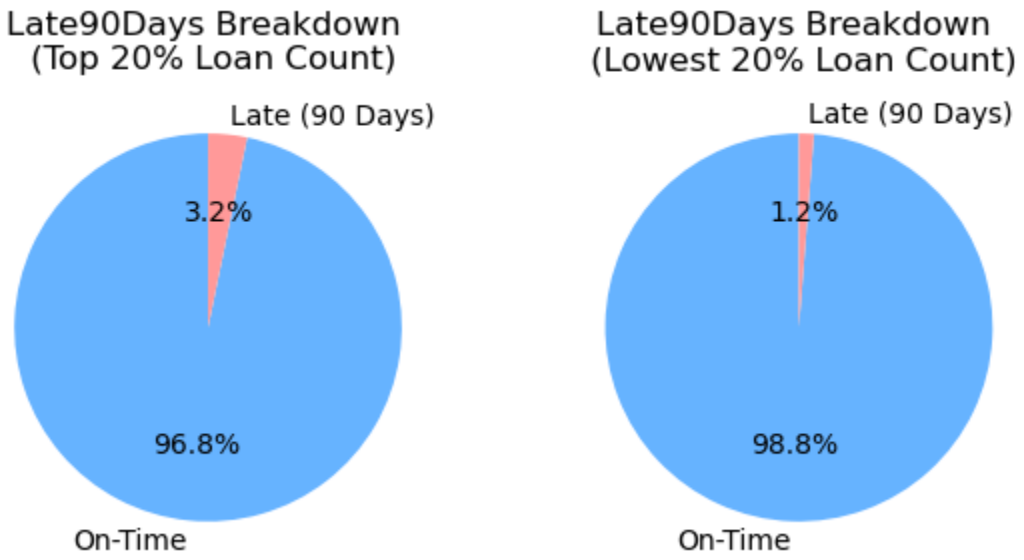

Figure 3. Pie chart showing the percentage of individuals with a 90+ day late payment, comparing two groups: those with three or more loans (top 20th percentile of loan count) and those with one loan (bottom 20th percentile of loan count). Notice that individuals with more loans have a higher percentage of late payments.

Exploring the data, we observe similar trends to the studies we reviewed, fluctuations in the macroeconomic factors impact delinquency in addition to personal spending habits. We also discovered that relatively few respondents report delinquent behavior (only 2.64% of the observations report delinquent behavior), making classification tasks more difficult. While we touch on these challenges here, a fuller discussion of modeling limitations and preprocessing steps is included in the next section.

Analysis/Methodology

For any given individual at a specific point in time, we aim to use information about their loan and spending habits, along with macroeconomic factors, to predict serious delinquency and identify the most significant predictors. To predict 90-day loan delinquency, which is a binary variable, we implemented classification models, commonly used for categorical prediction tasks.

We compared three popular supervised classification models: Logistic Regression, Random Forest Classifier, and HistGradientBoostingClassifier using grid search with 5-fold stratified cross-validation to compare the robustness and model efficacy. Logistic Regression is the most interpretable model and assumes a linear relationship between the dependent and independent variables. In contrast, Random Forest Classifier and HistGradientBoostingClassifier are less

interpretable, as they combine multiple small decision trees to capture non-linear relationships more effectively. These three models vary in interpretability and model complexity, allowing us to explore the trade-off between model transparency and predictive power in our analysis.

Given the highly imbalanced nature of the dataset– approximately 97% of cases labeled as "not late" and only 3% as "late"– we applied resampling techniques to better train models that can generalize to the minority class.

A 70/30 stratified train-test split was used to maintain the original class distribution in both subsets. To address class imbalance in the training set, we applied two separate sampling strategies:

- **Random Downsampling**: Reducing the number of majority class examples to match the minority class.
- **Random Upsampling**: Duplicating minority class examples to balance the dataset.

Feature scaling using StandardScaler was applied to all continuous predictors for the Logistic Regression models, ensuring that variables with different scales– such as mortgage loan balance and total number of loans– contribute equally to the model and prevent any single feature from disproportionately influencing the results due to its scale. Scaling was fitted on the training data and applied to both training and test sets to prevent data leakage. In addition to scaling the data, we also used different regularization methods, which prevent overfitting by adding a penalty to the model for having large or complex coefficients and can improve the robustness of the models.

Logistic Regression

Logistic Regression provides interpretable coefficients and is suitable as a baseline linear model. We evaluated both L1 (Lasso) and L2 (Ridge) regularization. Hyperparameter tuning was conducted using the following grid, on the following parameters: penalty and regularization. These methods are used to ensure the model does not overfit to the random noises of the data and performs better for unseen data, which is significant given we want to create an efficient model that could predict the probability of delinquency.

The best model was selected based on cross-validated ROC AUC scores, which evaluate the model's ability to distinguish between response classes– late or not late– across all classification thresholds. Feature importance was assessed using permutation importance on the held-out test set; in other words, we randomly drop one variable and observe how much the model's performance drops. For instance, if we dropped the total loan balance and the model's ability to correctly predict the outcome drops significantly, then it would have a high importance score.

Random Forest Classifier

Random Forest is a tree-based ensemble model capable of capturing nonlinear relationships and interactions between features. It is also robust to multicollinearity and outliers. Our grid for this model included: the number of estimators, the max amount of features to be used as well as the maximum depth that it should also have.

Like our Logistic Regression, the grid search selected the model with the highest ROC AUC. Feature importance was based on the model's built-in impurity-based importance measure.

HistGradientBoostingClassifier

This model is a fast, regularized implementation of gradient boosting that uses histogram-based binning for efficiency and scalability. It tends to outperform traditional boosting algorithms on large or complex datasets. Our grid for this model included: the learning rate for the model, its max iterations, max depth, and L2 regularization.

Evaluation Metrics

All metrics were calculated on the same held-out test set to ensure comparability across sampling strategies and models. Models were evaluated using the following metrics:

- **ROC AUC** : Measures separability between classes regardless of threshold.
- **Recall (Sensitivity)**: Especially important for identifying delinquent borrowers.
- **Precision, F1-score**: To balance false positives and false negatives.
- **Confusion Matrix**: For class-specific performance insight.
- **Calibration Curves**: To assess how well predicted probabilities reflect actual risk. The closer the plot is to the diagonal dotted line the better the calibration

Results

Below talks about our performance metrics for the 3 different kinds of models that we used, along with both the upsampled and downsampled versions of them. Something to note is that since there is such a large imbalance in our classes for classification, we chose the best model based on the accuracy of the model and most importantly the Precision for our "late" class. You see, as precision is the percentage that our model would predict the class correctly, it would stand right to focus more on the precision of our minority class, as there is major skewness in both the testing and training data.

1. **Logistic Regression**
    - **Downsampled Logistic Regression**:
      The downsampled Logistic Regression model achieved an overall accuracy of 88%. While it classified the majority "not late" class with very high precision

(1.00) and strong recall (0.88), its performance on the minority “late” class was modest, with a recall of 0.91 but a low precision of 0.17, resulting in a weak F1-score of 0.28. The low precision for the "late" class indicates the model's difficulty in identifying true positives without significant false positives.

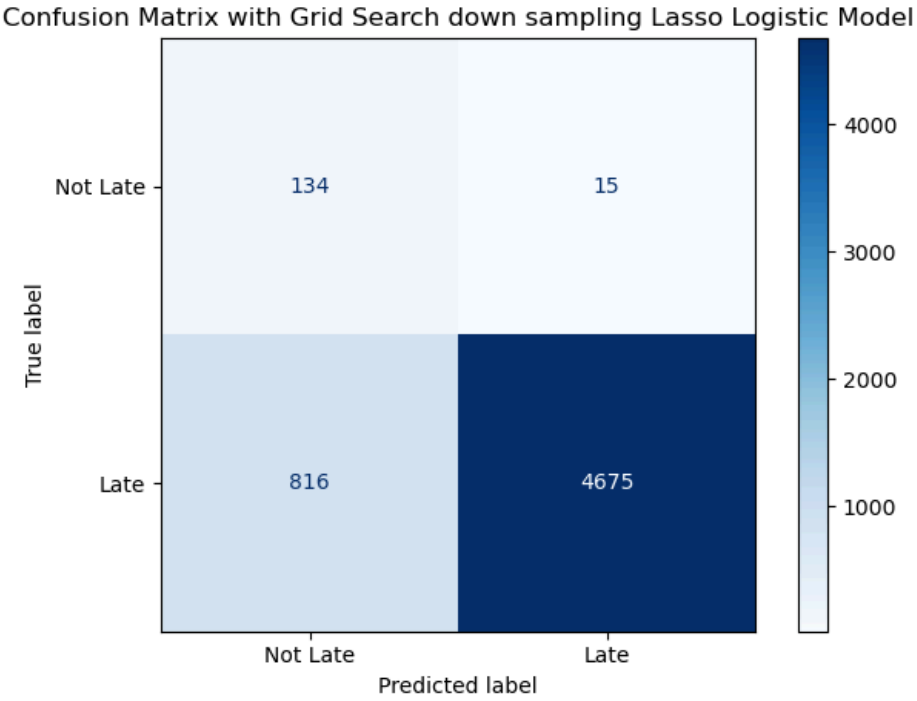

Figure 4: Confusion matrix for the Lasso Logistic Regression model with downsampling and hyperparameter tuning via Grid Search. The model shows high accuracy in identifying late payments (True Positives = 4675), but a notable number of actual late cases are misclassified as not late (False Negatives = 816). The matrix illustrates model performance on a heavily imbalanced dataset, highlighting the trade-off between sensitivity, correctly identifying individuals who display delinquent behavior, and specificity, correctly identifying individuals who do not display delinquent behavior.

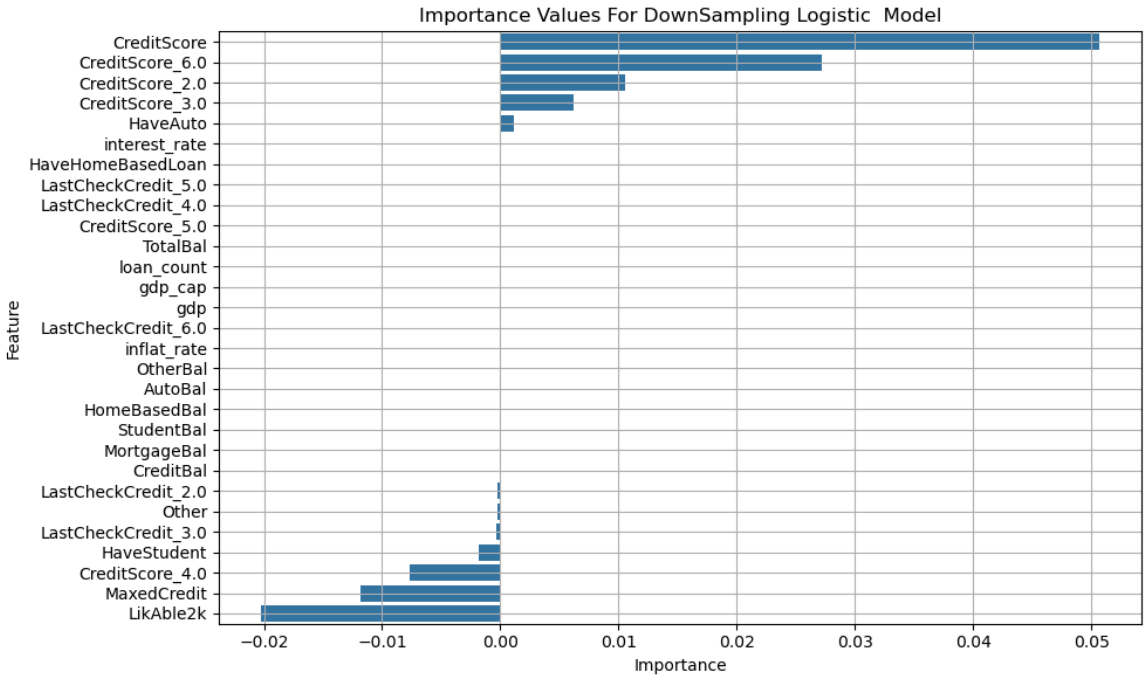

Figure 5: Feature importance values from the downsampled Lasso Logistic Regression model. The plot shows the magnitude and direction of each variable's influence on predicting loan delinquency. The Lasso regularization has shrunk many coefficients to near zero, emphasizing feature selection and sparsity. CreditScore and its categorical variants have the highest positive contribution, indicating their strong predictive power. Features like LikAble2k and MaxedCredit have negative coefficients, suggesting an inverse relationship with the likelihood of being late. For instance, if an individual maxed out their credit card for the month, then the model expected it to decrease the likelihood that they will display delinquent behavior. However, this does not match our expectations, given that if one's self-prescription of their financial circumstances is negative, their ability to make payments on time should be lower. This may be indicative of a possible confounding variable such as income, that we were not able to obtain.

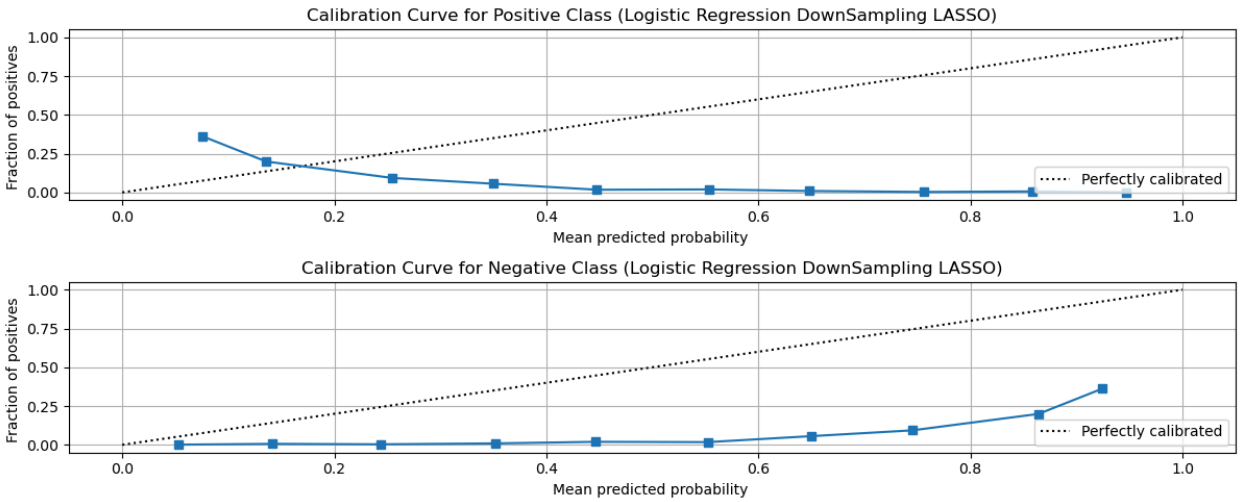

Figure 6: Calibration curves for the downsampled Lasso Logistic Regression model. The top panel shows the calibration curve for the positive class (late payments), while the bottom panel depicts the negative class (not late payments). Both curves deviate significantly from the diagonal line of perfect calibration, indicating that the model tends to overestimate probabilities for both classes. The poor alignment suggests that although the model may distinguish between classes reasonably well, its predicted probabilities are not well-calibrated.

- **Upsampled Logistic Regression**:
  The upsampled Logistic Regression model showed a modest improvement in overall performance, increasing accuracy to 88%. The precision for the “late” class rose slightly to 0.17, with a recall of 0.91 and an F1-score of 0.28. The upsampling strategy improved the recall for the "late" class, though the precision remained low, indicating that the model still struggled with false positives.

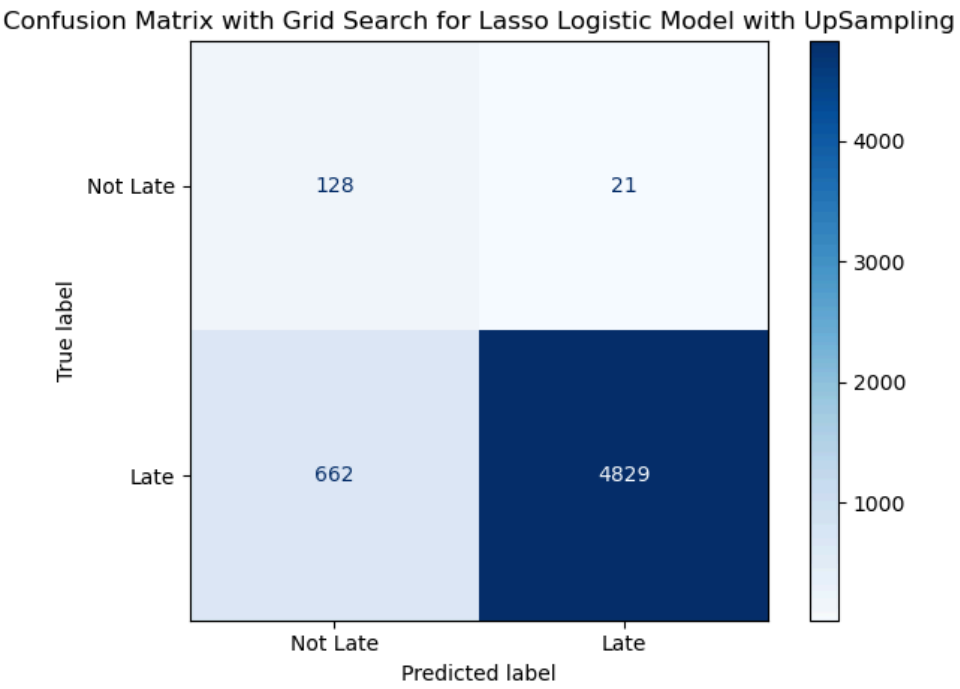

Figure 7: Confusion matrix for the Lasso logistic regression model with upsampling and hyperparameter tuning via grid search. The model shows strong performance in identifying late instances (True Positives = 4829) but has a relatively high number of false negatives (662), indicating room for improvement in detecting late cases.

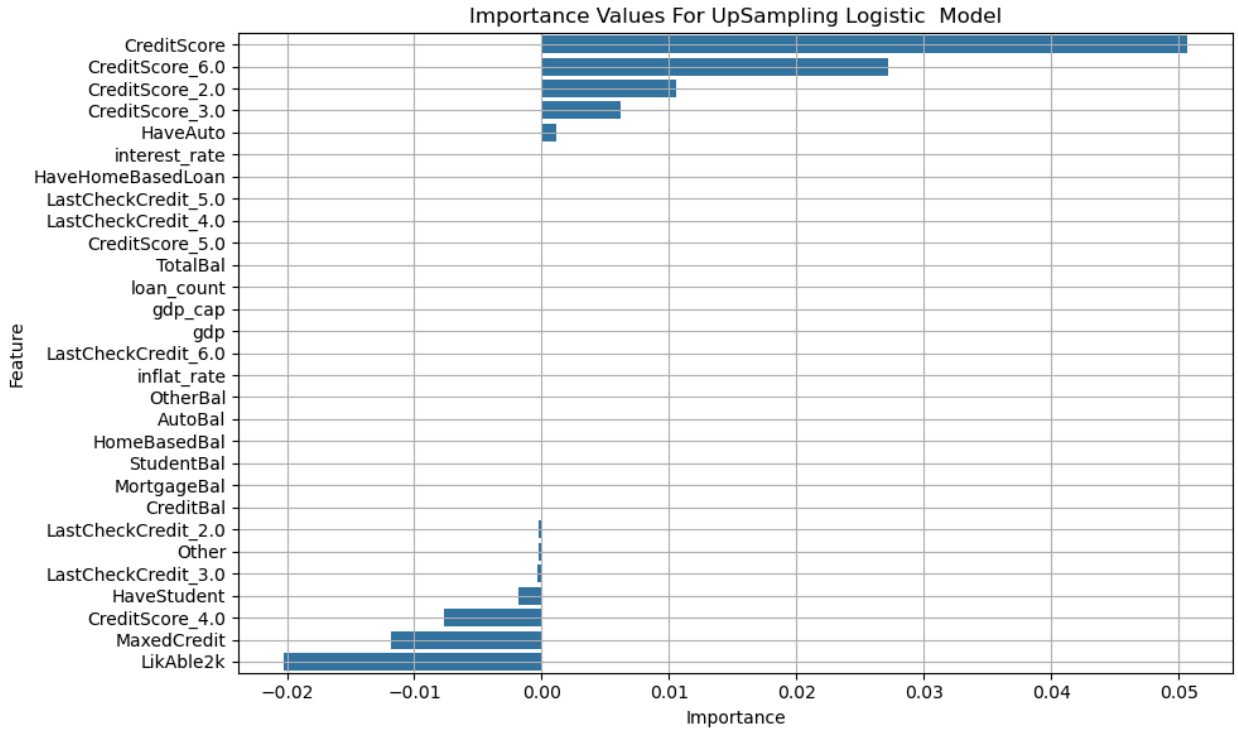

Figure 8: Feature importance values for the upsampled logistic regression model. The most influential features include CreditScore, CreditScore_6.0, and CreditScore_2.0, indicating their significant contribution to the model's predictive performance. Features like MaxedCredit and LikAble2k have smaller positive impacts, while others contribute minimally or negatively.

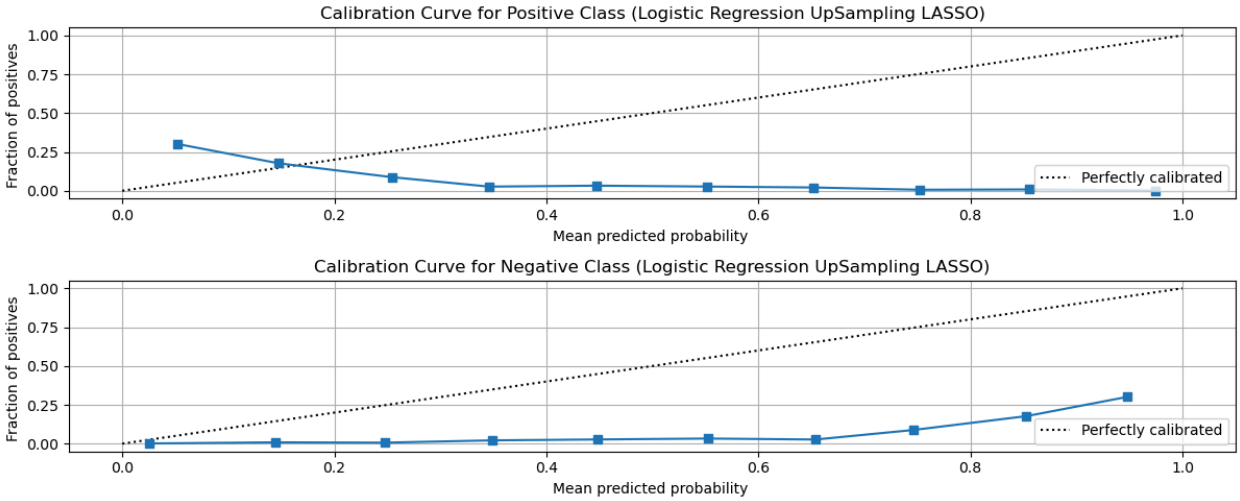

Figure 9: Calibration curves for the upsampled Lasso logistic regression model. The top panel shows the calibration for the positive class (Late), where predicted probabilities consistently overestimate the actual fraction of positives. The bottom panel shows the calibration for the negative class (Not Late), which also exhibits overconfidence, especially at higher predicted probabilities. Both plots indicate that the model is not well-calibrated.

2. **Random Forest Classifier**
    - **Downsampled Random Forest**:
      The downsampled Random Forest model achieved an accuracy of 86%. It preserved high recall for the “late” class (0.93) but suffered from low precision (0.15), leading to an F1-score of 0.26. The model performed strongly on the “not late” class with a high recall of 0.86 and precision of 1.00, but the imbalance between classes remained a challenge.

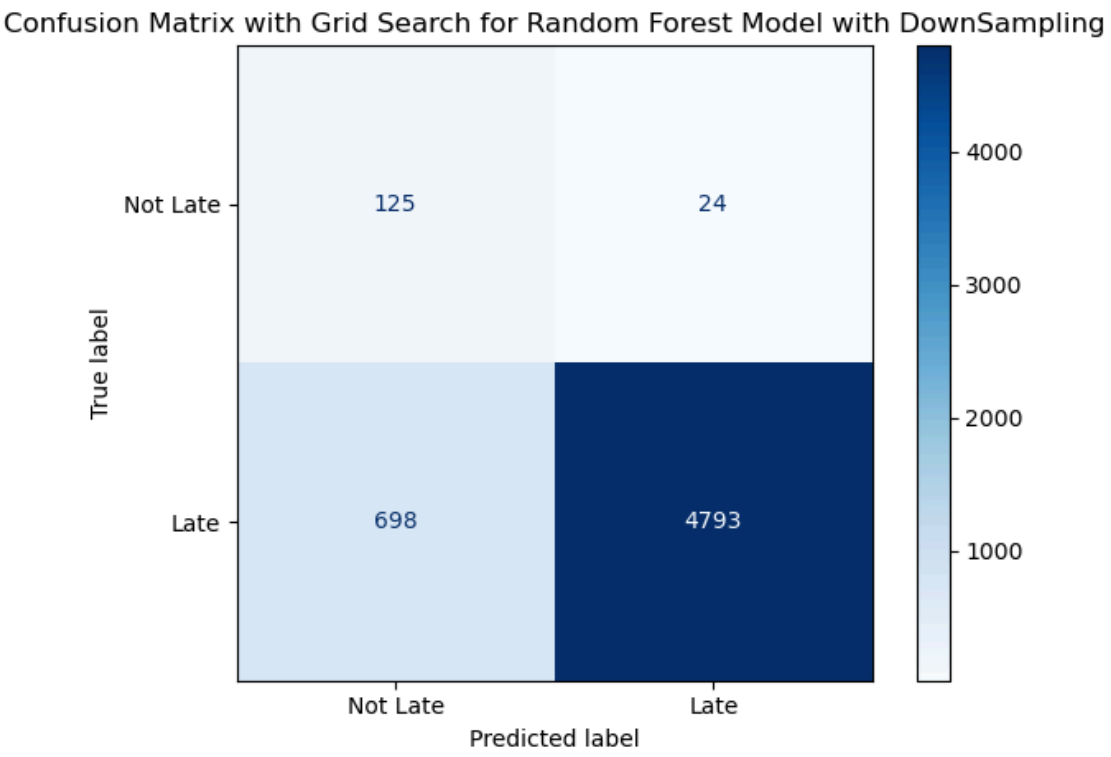

Figure 10: Confusion matrix for the Random Forest model with downsampling and hyperparameter tuning via grid search. The model shows strong performance in identifying late instances (True Positives = 4793) but has a relatively high number of false negatives (698), indicating room for improvement in detecting late cases. The number of false positives is low (24), suggesting good specificity in identifying "Not Late" instances.

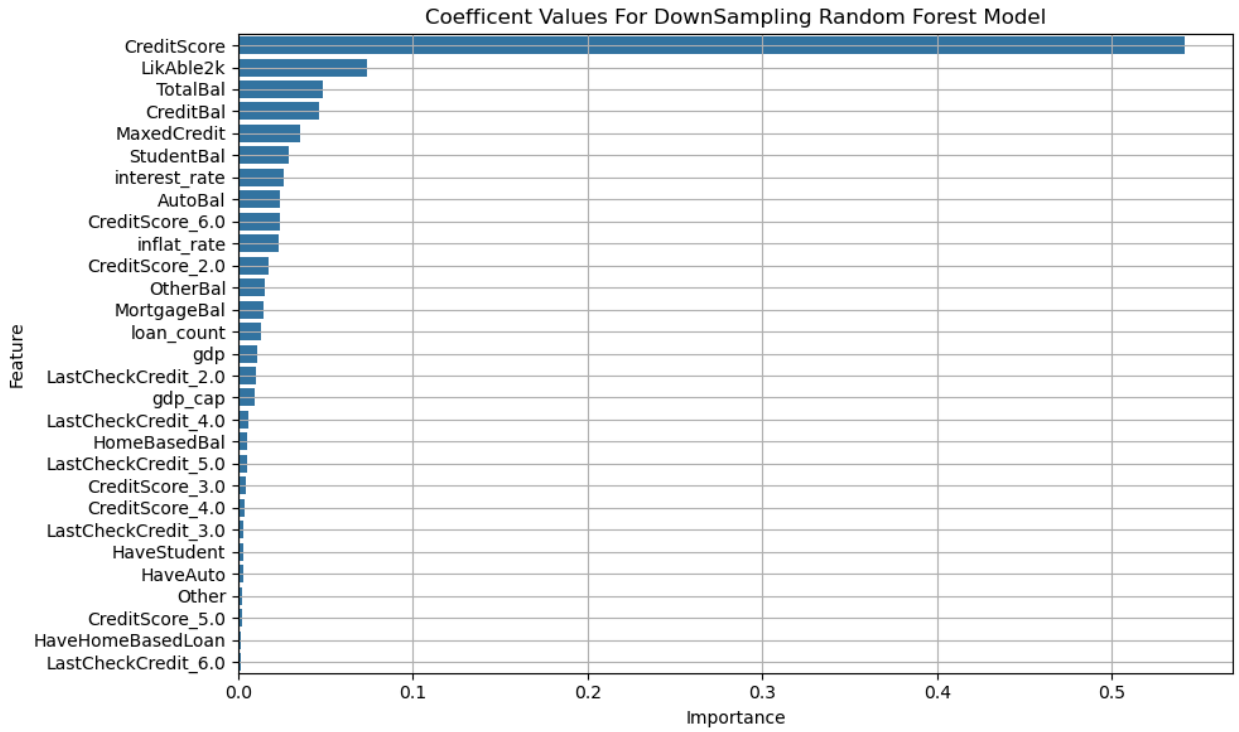

Figure 11: Feature importance plot for the Random Forest model with downsampling. The plot highlights the top contributing features based on their importance scores, with CreditScore being the most influential predictor by a wide margin, followed by LikAble2k, TotalBal, and CreditBal. For instance, assuming all else is equal, as total loan balance or credit card balance increases, we expect an increase in the likelihood of a user being 90+ late on their payment. These features play a critical role in the model's decision-making process for predicting lateness.

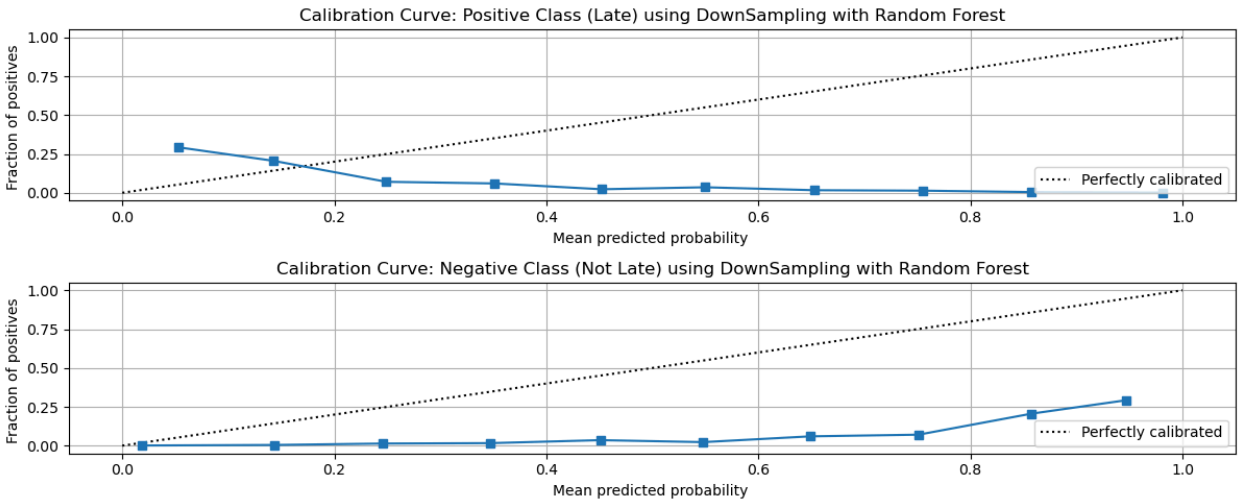

Figure 12: Calibration curves for the downsampled Random Forest model. The top panel shows the calibration curve for the positive class (late payments), while the bottom panel depicts the negative class (not late payments). Both curves deviate substantially from the diagonal line of perfect calibration, indicating that the model consistently overestimates the likelihood of positive outcomes. The misalignment suggests that while the model may effectively differentiate between classes, its predicted probabilities lack proper calibration.

- **Upsampled Random Forest**:
  The upsampled Random Forest model achieved the highest accuracy among all models at 96%. However, this came at the cost of very low recall (0.32) and precision (0.29) for the "late" class, yielding an F1-score of just 0.26 for that group. Despite high accuracy, the model was biased toward the majority class, making it less effective for identifying delinquent borrowers.

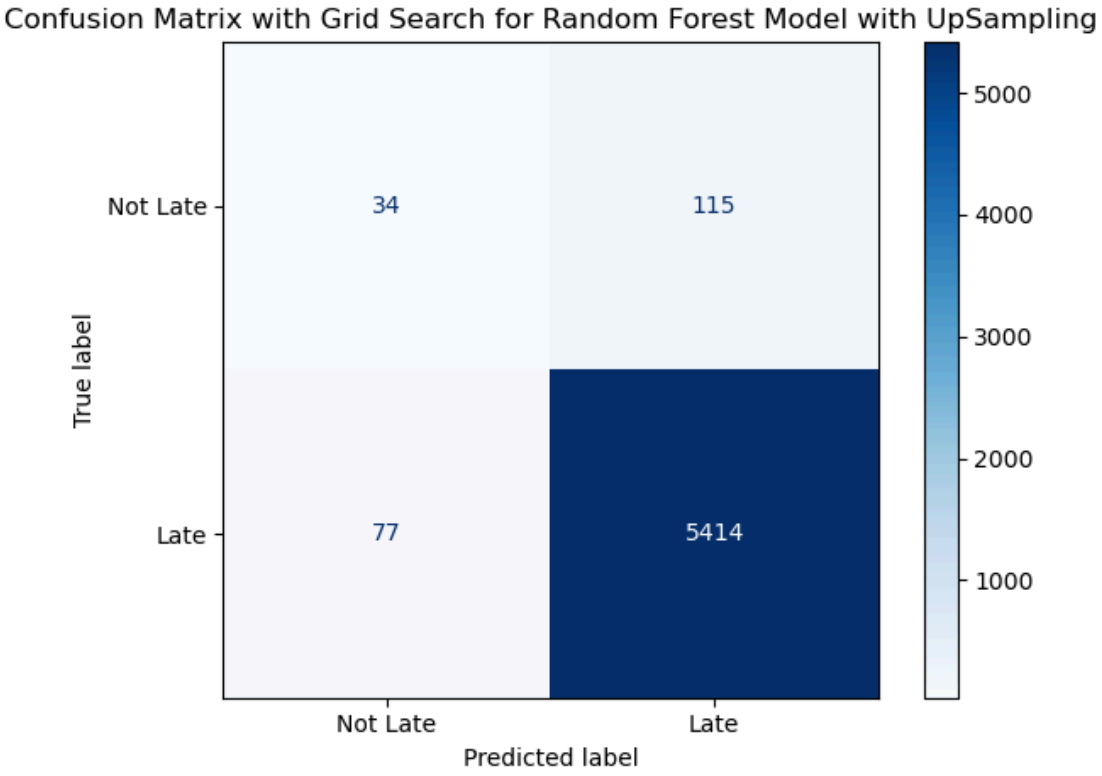

Figure 13: Confusion matrix for the Random Forest model with upsampling and hyperparameter tuning via grid search. The model demonstrates strong performance in identifying late instances (True Positives = 5414) with a relatively low number of false negatives (77), indicating high sensitivity. However, the model misclassified a notable number of "Not Late" instances as "Late" (False Positives = 115), suggesting reduced specificity.

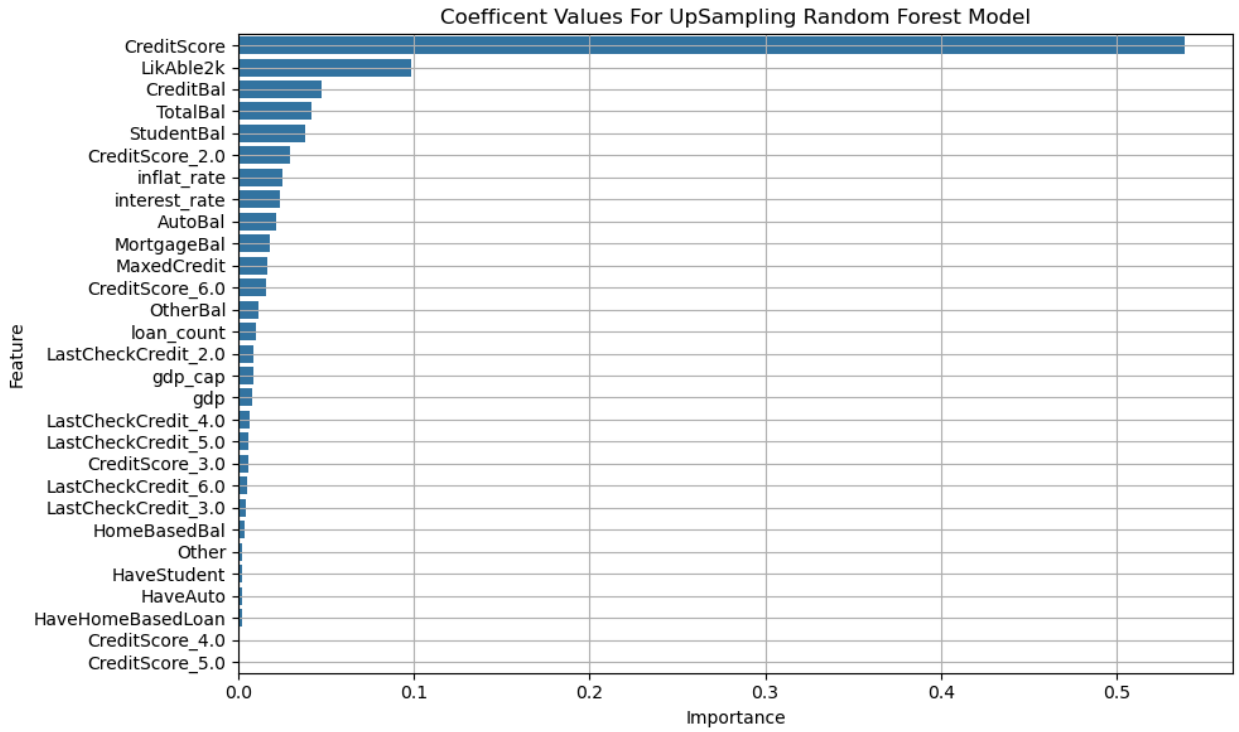

Figure 14: Feature importance plot for the Random Forest model with upsampling. The plot highlights the top contributing features based on their importance scores, with CreditScore being the most influential predictor by a wide margin, followed by LikAble2k, CreditBal, and TotalBal. These features play a critical role in the model's decision-making process for predicting lateness.

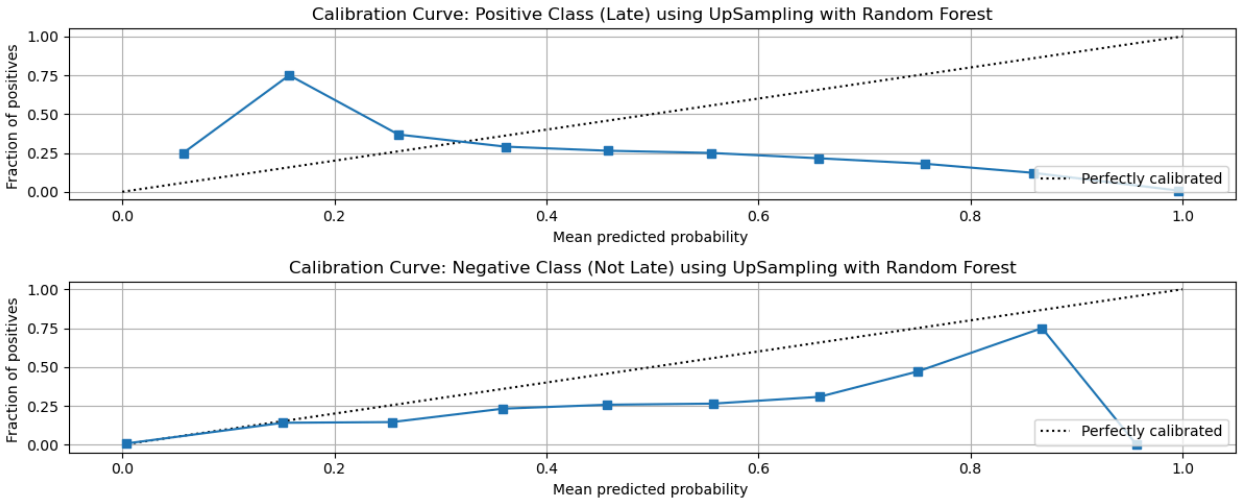

Figure 15: Calibration curves for the upsampled Random Forest model. The top panel shows the calibration curve for the positive class (late payments), while the bottom panel depicts the negative class (not late payments). Both curves deviate significantly from the diagonal line of

perfect calibration, indicating that the model tends to overestimate probabilities for both classes. The poor alignment suggests that although the model may distinguish between classes reasonably well, its predicted probabilities are not well-calibrated.

3. **HistGradientBoostingClassifier**
    - **Downsampled HistGradientBoostingClassifier**:
      The downsampled HistGradientBoostingClassifier performed similarly to the downsampled Logistic Regression and Random Forest models, with an accuracy of 87%. It maintained a high recall for the "late" class (0.92), but its precision remained low (0.15), yielding an F1-score of 0.25. While the model identified most late payments, it struggled to minimize false positives.

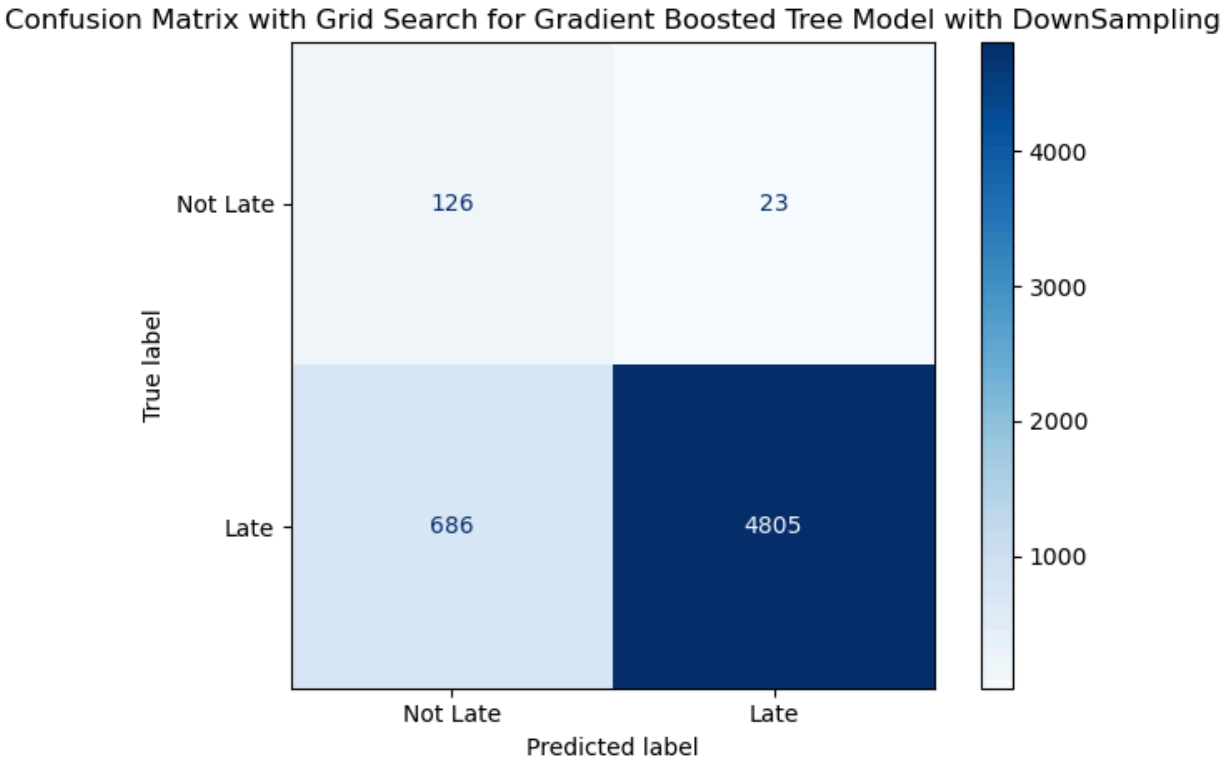

Figure 16: Confusion matrix for the Gradient Boosted Tree model with downsampling after hyperparameter tuning via grid search. The matrix shows strong predictive performance for the "Late" class (4,805 true positives) with relatively few false negatives (686). The model demonstrates a higher misclassification rate for the "Not Late" class, with 126 true negatives versus 23 false positives, indicating greater sensitivity toward predicting late payments while maintaining reasonable specificity.

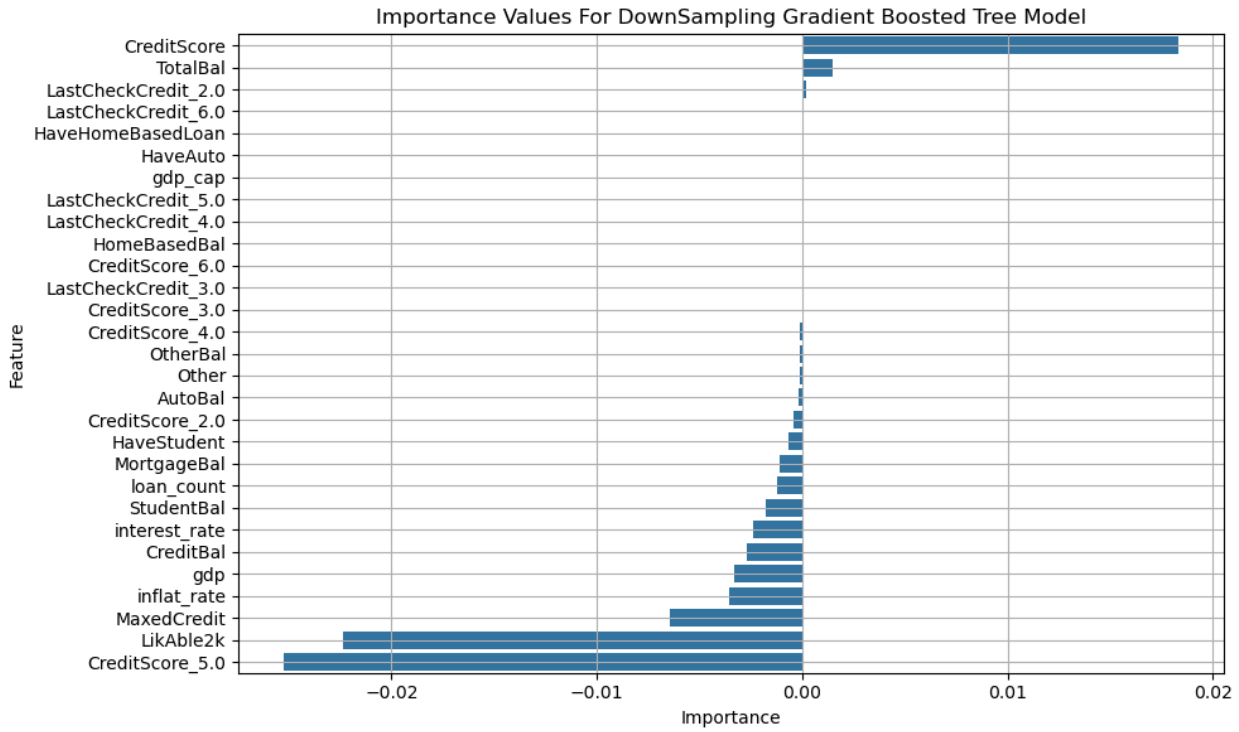

Figure 17: Feature importance plot for the Gradient Boosted Tree model with downsampling. The plot highlights CreditScore as the overwhelmingly dominant predictor with an importance value approaching 0.04, significantly outweighing all other features. Secondary contributors include inflation_rate and AutoBal with modest importance values, while the remaining features exhibit minimal influence on the model's predictions. This clear disparity in feature importance suggests that the model's decision-making process for predicting lateness relies heavily on credit score information, with limited contributions from other financial indicators such as LikAble2k and TotalBal, which show slight negative importance values.

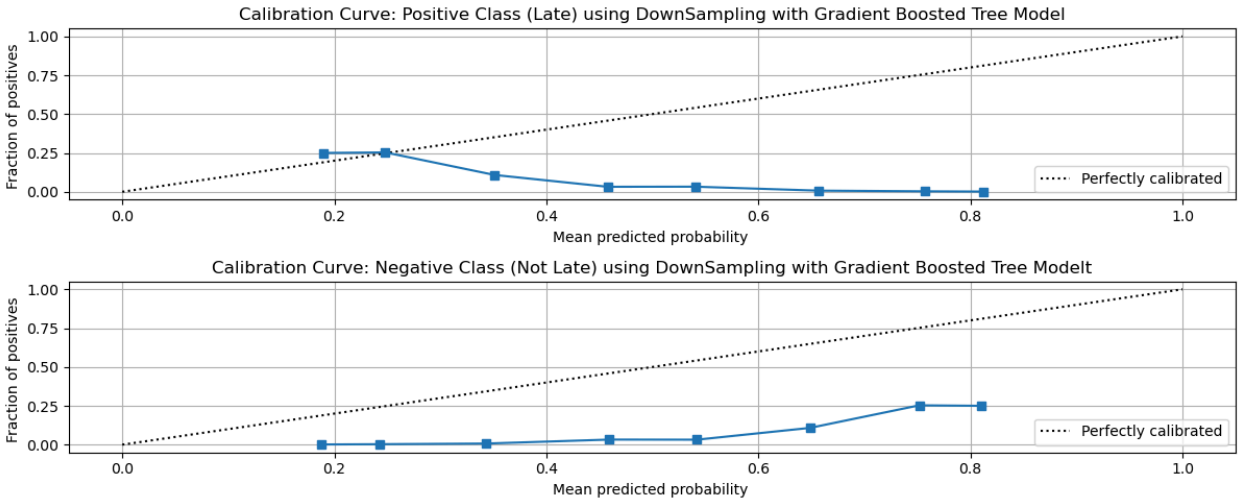

Figure 18: Calibration curves for the downsampled Gradient Boosted Tree model. The top panel shows the calibration curve for the positive class (late payments), while the bottom panel depicts the negative class (not late payments). Both curves show notable deviation from the diagonal line of perfect calibration, especially in the lower probability ranges. This indicates that the model tends to overestimate the likelihood of late payments and underestimates not late payments in certain intervals. The miscalibration suggests that while the model may provide good classification performance, its probability estimates are not well-calibrated.

- **Upsampled HistGradientBoostingClassifier**:
  The upsampled HistGradientBoostingClassifier emerged as the best-performing model overall. It achieved 97% accuracy and recorded the highest F1-score for the "late" class (0.35), with a recall of 0.34 and precision of 0.38. It also delivered the highest macro-averaged F1-score (0.67) and recall (0.66) among all models, indicating improved class balance and the model's ability to detect loan delinquencies effectively. This model demonstrated the best overall trade-off between precision and recall.

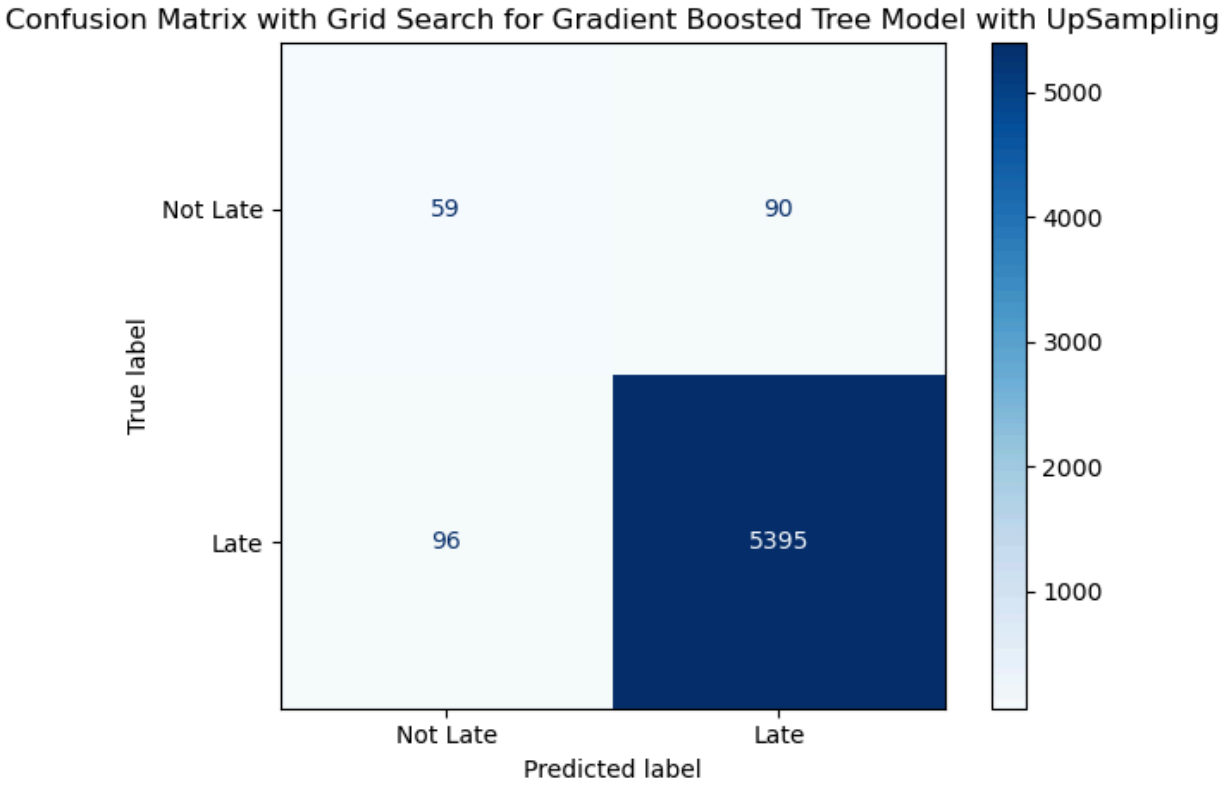

Figure 19: Confusion matrix for the Gradient Boosted Tree model with upsampling after hyperparameter optimization via grid search. The matrix reveals strong performance in correctly identifying late payments (5,395 true positives) with relatively few false negatives (96). However, the model shows reduced accuracy for the "Not Late" class with only 59 true negatives

compared to 90 false positives. This imbalance suggests the upsampling approach has optimized the model for high sensitivity to late payments (98.2%) at the expense of specificity (39.6%), reflecting a priority on identifying potential late payments even if it results in some false alarms.

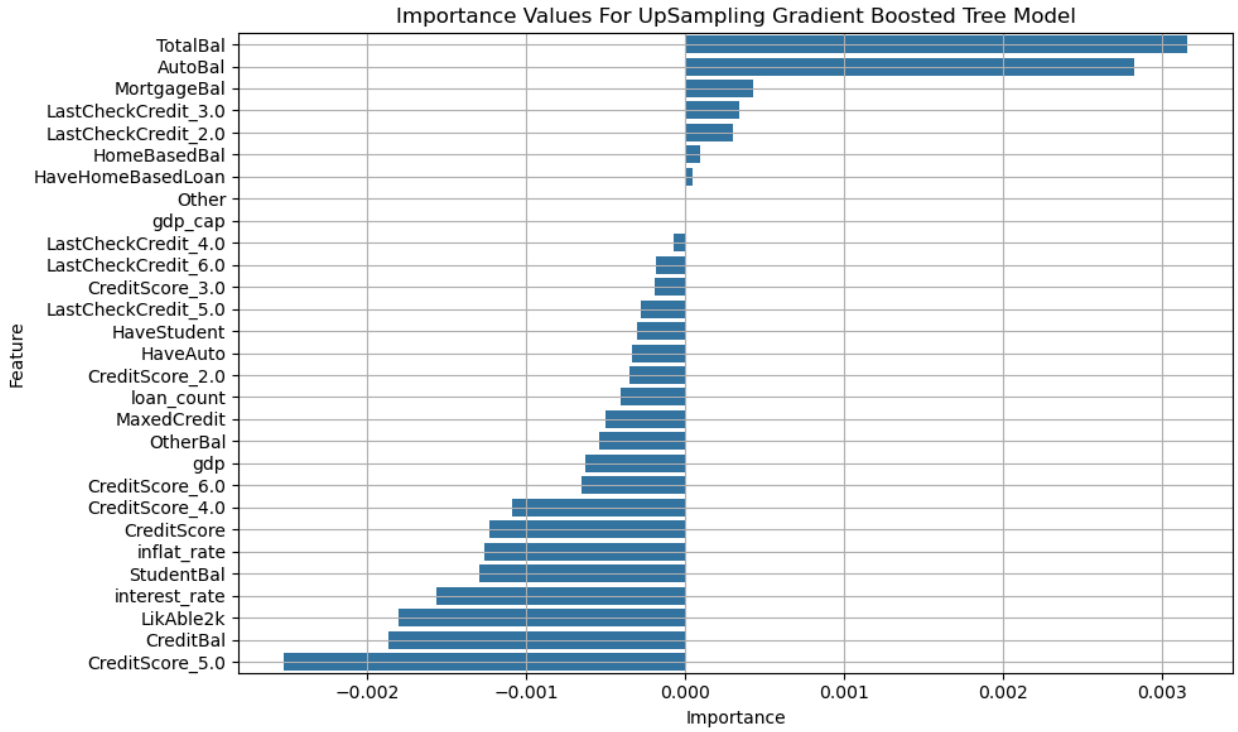

Figure 20: Feature importance plot for the Gradient Boosted Tree model with upsampling. The plot illustrates the relative contribution of each feature to the model's predictions, with CreditScore showing the highest positive importance value, followed by AutoBal with moderate positive importance. Notably, several features like CreditBal and MaxCredit display negative importance values, indicating they may have an inverse relationship with the target variable, which means that as the credit balance increases, we expect a smaller chance of delinquency. This could indicate possible behaviors of payment for loans using credit cards, which is an issue that researchers have raised as well about using credit cards to pay loans (Carrns 2024).

This distribution of importance values across both positive and negative ranges provides insight into the complex dynamics influencing the prediction of late payments.

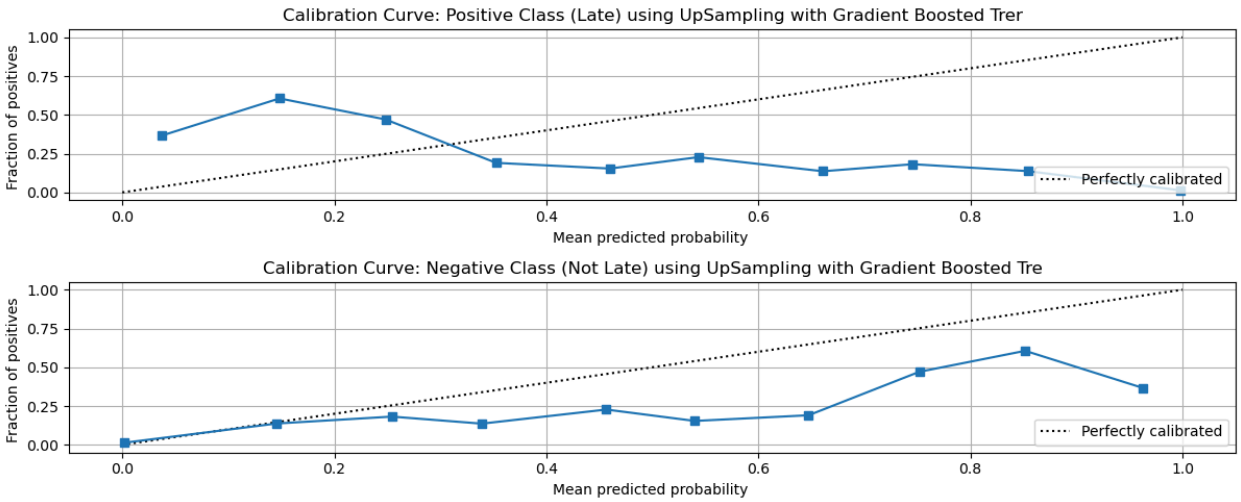

Figure 21: Calibration curves for the Gradient Boosted Trees model using upsampling. The top plot shows the calibration curve for the positive class ("Late"), while the bottom plot shows the curve for the negative class ("Not Late"). The x-axis represents the mean predicted probability, and the y-axis represents the observed fraction of positives. The dashed diagonal line indicates perfect calibration. Deviations from this line indicate overconfidence (below the line) or underconfidence (above the line) in predicted probabilities.

Robustness

To test how reliable and consistent our models were, we looked at a few different factors that could impact performance. We first used 5-fold stratified cross-validation to check how each model handled different data splits. This gave us insights into how the models hold up with more varied data, giving us a better idea of their performance in real-world settings rather than one specific data sample. This also reduced the likelihood of overfitting, ensuring consistent performance with different data sets, increasing our confidence in the models' real-world applicability.

We also trained each model using both upsampled and downsampled versions of the data to address class imbalance. Random undersampling balanced the classes by reducing the majority class (decreasing users who do not display delinquent behavior), while random oversampling balanced the classes by selecting more of the minority class (increasing users who displayed delinquent behavior). This helped our models learn from the minority class and improved their ability to detect potential late payments. We also used a grid search with various hyperparameters on our models, ensuring that performance metrics like ROC AUC, recall, precision, and F1-score were performing consistently across various hyperparameter combinations.

After employing these approaches for all three models, we noticed that the upsampled HistGradientBoostingClassifier was our best-performing model. It achieved the highest overall accuracy, highest F1-score, and had the best trade-off between precision and recall among the

three models. Its consistent performance across different sampling methods and folds also highlights its potential as a powerful tool for loan delinquency predictions by loaner companies.

Conclusion

Based on our findings, we believe our model can serve as a valuable tool for the Office of Federal Student Aid and other major loaning agencies in identifying high-risk borrowers before loan disbursement. Specifically, the upsampled HistGradientBoostingClassifier identifies balances for other loans, users' frequency of checking their credit score, and interest rate as significant covariates. As a result, loaning organizations could look to consider those covariates when determining whether to approve a loan application or not. In addition, the model demonstrated strong performance predicting individuals likely to become delinquent on their student loans, with relatively few false negatives. This means that most of the borrowers flagged as likely to become delinquent were actually likely to miss payments. These predictions will help prevent risky loan disbursements, which increases the financial security of major loaner companies. However, we do not recommend using the model to identify borrowers who are expected to repay on time. The high rate of misclassifying reliable borrowers as delinquent could lead to unnecessary barriers for otherwise secure applicants.

Looking ahead, we aim to improve our model's precision, especially by reducing the number of false positives. We plan to explore how adjusting the classification threshold can help balance out identifying delinquent borrowers and avoiding misclassification of reliable ones. However, we recognize this may lead to a higher false negative rate, so the effects will need to be evaluated before committing to the change. Additionally, we want to conduct more feature engineering, likely by creating features based on ratios or combinations of existing features, to increase the model's predictions. If we can access more socioeconomic data, like age and marital status, we would also like to explore how these personal factors play into loan delinquency. We believe these changes could really help the performance of our model and make it an even more powerful tool for our clients.

References

Carrns, Ann. “Chase to Bar Customers from Using Credit Cards for ‘Pay Later’ Loans - The New York Times.” New York Times, 26 July 2024, www.nytimes.com/2024/07/26/your-money/chase-pay-later-loans-credit-cards.html.

Chong F. Loan Delinquency: Some Determining Factors. Journal of Risk and Financial Management. 2021; 14(7):320. https://doi.org/10.3390/jrfm14070320

Federal Reserve Bank of New York. (n.d.). Center for Microeconomic Data: Data Bank. https://www.newyorkfed.org/microeconomics/databank.html

Nigmonov, Asror and Shams, Syed and Alam, Khorshed, Macroeconomic Determinants of Loan Delinquencies: Evidence from the US Peer-to-Peer Lending Market (April 29, 2021). ADB-IGF Special Working Paper Series “Fintech to Enable Development, Investment, Financial Inclusion, and Sustainability”, Available at SSRN: https://ssrn.com/abstract=3836404 or http://dx.doi.org/10.2139/ssrn.3836404

U.S. Bureau of Labor Statistics. (n.d.). Survey Output Data Retrieval Tool. https://data.bls.gov/pdq/SurveyOutputServlet

U.S. Inflation Calculator. (n.d.). Historical Inflation Rates: 1914–2024. https://www.usinflationcalculator.com/inflation/historical-inflation-rates/

Appendix

Table 1. List of variable names, descriptions, and types (quantitative or categorical). We have 20 variables that we can use to predict y, a binary variable indicating whether or not an individual is delinquent for a loan.

Variable Name	Description	Type
Late90Days	Whether an individual has a 90+ day late payment in the last 12 months	Categorical (binary)
HaveCredit	Whether an individual owns a credit card	Categorical (binary)
HaveMortgage	Whether an individual possess mortgage loan(s)	Categorical (binary)
HaveStudent	Whether an individual possess student loan(s)	Categorical (binary)
HaveHomeBasedLoan	Whether an individual possess home-based loan(s)	Categorical (binary)
HaveAuto	Whether an individual possess auto loan(s)	Categorical (binary)
HaveOther	Whether an individual possess other loans	Categorical (binary)
CreditBal	Credit card balance	Quantitative (numeric)

Variable Name	Description	Type
MortgageBal	Mortgage balance	Quantitative (numeric)
StudentBal	Student loan balance	Quantitative (numeric)
HomeBasedBal	Home-based loan balance– includes loans taken to renovate homes and purchase furniture	Quantitative (numeric)
AutoBal	Auto loan balance	Quantitative (numeric)
OtherBal	Balance of other personal loans	Quantitative (numeric)
TotalBal	Total balance of all loans	Quantitative (numeric)
MaxedCredit	Whether an individual max out (borrow up to the limit) on any of their credit cards	Categorical (binary)
LikAble2k	Percent change an individual could come up with $2,000 if an unexpected need arose within the next month?	Quantitative (numeric)
inflat_rate	Inflation rate	Quantitative (numeric)

Variable Name	Description	Type
interest_rate	Interest rate	Quantitative (numeric)
gdp	Gross Domestic Product	Quantitative (numeric)
gdp_cap	Gross Domestic Product per Capita	Quantitative (numeric)

For credit score, there are 6 levels and the corresponding values for each are listed in order: Below 620, 620-679, 680-719, 720-760, Above 760, Don’t Know.

For last time the user check their score, there are 6 levels and the correspond values for each of those values are listed in order: less than 1 month ago, between 1 and 6 months ago, between 6 to 12 months ago, between 1 to 2 years ago, more than 2 years ago, and never checked their score.