Credit Risk Prediction Using Machine Learning

Yuyao (Olivia) Wang

April 2023

yuyaow@bu.edu

Project Overview

This project aims to predict credit risk using structured data such as financial and demographic variables. The primary objective is to develop machine learning models that estimate the likelihood of a customer defaulting on their loan. Techniques include:

- XGBoost algorithm for prediction
- SMOTE (Synthetic Minority Over-sampling Technique) to address imbalanced data
- Visualization of important risk factors

Dataset

The dataset used is from the *Home Credit Default Risk* competition on Kaggle. Key components include:

- **application_train.csv**: Training data with loan default status (TARGET).
- **application_test.csv**: Test data used for predictions (TARGET absent).
- **TARGET Variable**: Binary indicator where 1 represents a default, and 0 represents no default.

Key Tasks and Steps

1. **Data Preprocessing:** Handle missing values, feature engineering, and data transformation.
2. **Feature Extraction:** Use techniques like TF-IDF for text data and feature importance analysis for structured data.
3. **Model Building:** Develop baseline models (e.g., Logistic Regression), apply advanced models like XGBoost.
4. **Model Evaluation:** Assess models using ROC-AUC, accuracy, and SHAP values for feature interpretation.

Expected Results

The outcomes of this project include:

- **Accurate Predictions:** Models that predict loan defaults effectively.
- **Feature Insights:** Identification of critical factors affecting default risk.
- **Improved Decision-Making:** Enhanced risk assessments for loan applicants.
- **Explainable AI:** Use of SHAP values for transparent and interpretable models.

ROC Curve Analysis - Logistic Regression

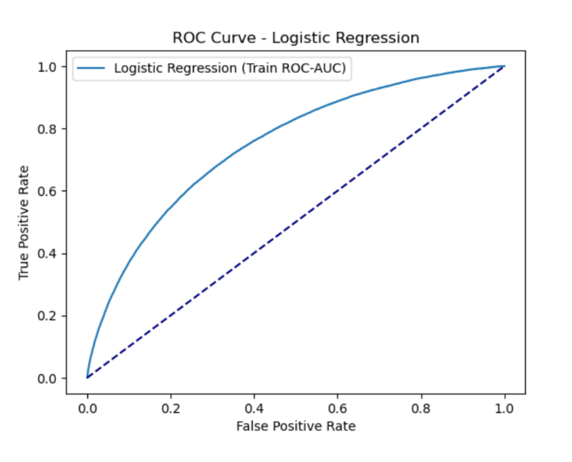


Figure 1: ROC Curve for Logistic Regression Model

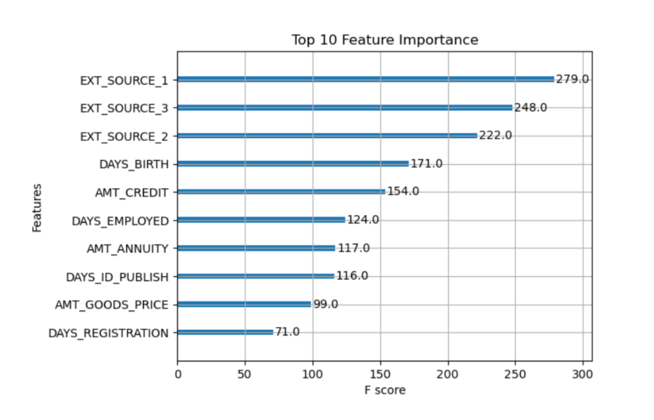
ROC Curve Analysis

- **True Positive Rate (TPR)** vs. **False Positive Rate (FPR)**: The ROC curve shows the trade-off between TPR (recall) and FPR at various thresholds.
- **Model Performance**: The blue curve indicates the logistic regression model's ability to distinguish between positive and negative classes. A curve closer to the top-left corner indicates better performance.
- **Random Classifier**: The dashed diagonal line represents the performance of a random classifier (TPR = FPR), indicating a 50% chance of accuracy.

Analysis and Conclusion

- The ROC curve suggests the model performs significantly better than random guessing, with an estimated AUC between 0.7 and 0.85.
- While the model shows moderate-to-good classification ability, further improvements could be made through advanced models like XGBoost or hyperparameter tuning.

Feature Importance Analysis



Feature Importance Analysis

- **Key Features**: The XGBoost model identifies external credit scores as the most significant predictors of default risk. This indicates that third-party evaluations of financial behavior are crucial in assessing customer creditworthiness.
- **Demographic and Financial Variables**: Age, credit amount, and employment history also rank highly in importance, contributing to the model's predictive power.
- **Insights**: This feature importance plot helps to highlight the critical variables driving the model's decision-making process, suggesting areas for potential improvement or deeper exploration.

Conclusion

- The emphasis on external credit scores suggests that financial behavior evaluations from external sources are integral to the model's success in predicting default risk.
- Demographic and financial variables such as age and credit amount remain crucial, informing further analysis and possible model refinement.

Confusion Matrix - Logistic Regression

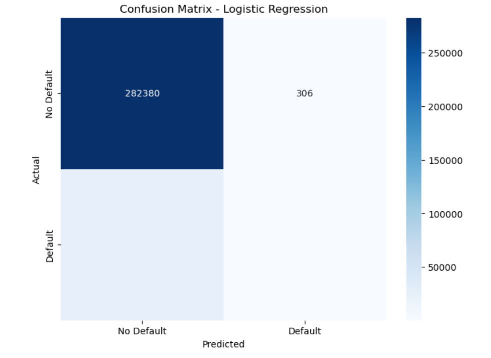


Figure 2: Logistic Regression Confusion Matrix

Confusion Matrix Observations

- **Imbalance in Predictions:** The model predicts ”No Default” almost exclusively, with 282,380 predictions for ”No Default” and only 306 for ”Default.”
- **False Positives but No True Positives:** There are 306 false positives but no true positives, indicating difficulty in identifying actual defaults.
- **Class Imbalance:** This issue is likely caused by the dataset’s class imbalance, where the ”No Default” class dominates the data.

XGBoost Confusion Matrix

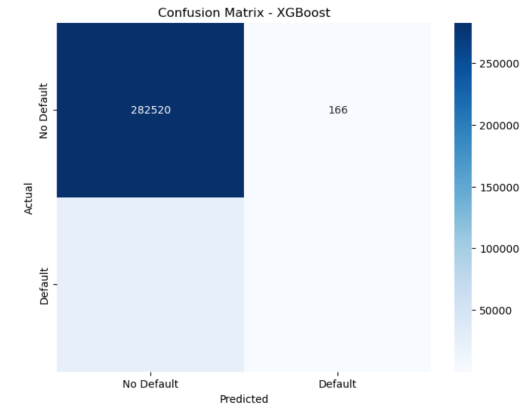


Figure 3: Confusion Matrix - XGBoost

XGBoost Confusion Matrix

- **Improved Performance:** XGBoost has reduced false positives from 306 (Logistic Regression) to 166, showing better classification of "No Default."
- **No True Positives:** Despite improvements, the model still fails to identify any default cases, indicating persistent difficulty in detecting the minority class.
- **Class Imbalance Issue:** The confusion matrix shows a continued bias towards the "No Default" class, which may require addressing the class imbalance more directly.

Implications and Conclusion

- **Class Imbalance Handling:** Consider techniques such as SMOTE or undersampling to address the imbalance.
- **Evaluation Metrics:** Use metrics like Precision, Recall, and AUC-ROC for better evaluation of the minority class.
- **Threshold Tuning:** Adjust the probability threshold to increase sensitivity to defaults.
- **Advanced Models:** Consider more complex models or ensemble techniques to further improve performance on imbalanced data.
- **Conclusion:** While XGBoost reduced false positives, it still struggles with predicting true positives, highlighting the need for addressing class imbalance and refining the model further.

XGBoost Confusion Matrix with SMOTE and Threshold Tuning

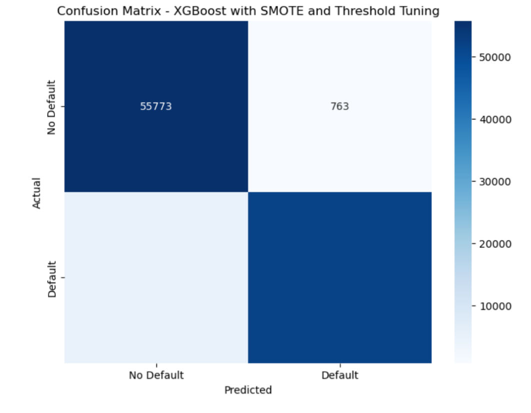


Figure 4: Confusion Matrix - XGBoost with SMOTE and Threshold Tuning

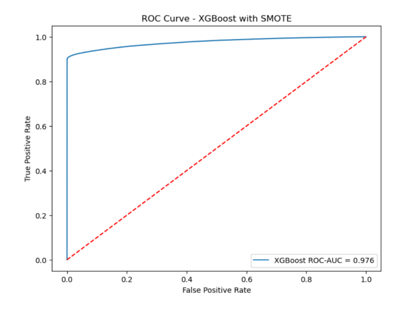
XGBoost Confusion Matrix with SMOTE and Threshold Tuning

- **Improved True Positives:** The model now correctly identifies more default cases, indicating that SMOTE and threshold tuning have increased its sensitivity to the minority class.
- **Increased False Positives:** There is a trade-off, as false positives have increased to 763, a common outcome when lowering the threshold to boost recall.
- **False Negatives Remain:** Despite improvements, some defaults are still being misclassified as "No Default," leaving room for further enhancement.

Interpretation and Conclusion

- **Improved Recall:** The model's ability to detect defaults has significantly improved, which is critical in contexts where capturing all potential defaults is important, even if it leads to more false positives.
- **Precision-Recall Trade-off:** The increase in false positives highlights the trade-off between precision and recall. While recall has improved, precision may have decreased slightly.
- **False Negatives:** The remaining false negatives suggest there is still room for optimization, such as through further tuning or more complex models.
- **Conclusion:** SMOTE and threshold tuning have enhanced the model's ability to detect defaults, but future steps should aim to strike a balance between precision and recall, possibly through cost-sensitive learning or additional threshold adjustments.

ROC Curve - XGBoost with SMOTE



- **AUC-ROC Score:** The model achieved an AUC-ROC score of 0.976, indicating excellent discriminatory power between the "Default" and "No Default" classes.
- **ROC Curve:** The curve shows the model's performance across varying thresholds, with a shape close to the top-left corner, signaling strong performance.

Classification Report - Class 0 (No Default)

- **Precision:** 0.92 - 92% of the predicted ”No Default” instances were correct.
- **Recall:** 0.99 - The model successfully captured 99% of actual ”No Default” cases.
- **F1-Score:** 0.95 - Balances precision and recall, showing consistent performance on the ”No Default” class.

Classification Report - Class 1 (Default)

- **Precision:** 0.99 - 99% of the predicted "Default" instances were correct.
- **Recall:** 0.91 - The model successfully captured 91% of actual "Default" cases.
- **F1-Score:** 0.95 - Indicates a balanced performance between precision and recall on the "Default" class.

Accuracy and Support Metrics

- **Overall Accuracy:** 95% - The model correctly predicted 95% of all cases.
- **Support:** Both classes have approximately equal cases (56,536 for "No Default" and 56,539 for "Default"), indicating balanced data after applying SMOTE.

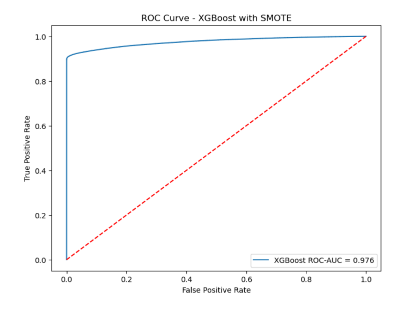
Key Observations

- **Balanced Performance:** F1-scores of 0.95 for both classes reflect the model's balanced precision and recall.
- **High Recall for Defaults:** The recall of 0.91 for the "Default" class shows that the model captures most default cases, a significant improvement over models without SMOTE.
- **Trade-off in Precision and Recall:** The model slightly sacrifices precision for the "No Default" class to achieve higher recall for the "Default" class, a necessary trade-off for imbalanced data.

Conclusion

- **Strong Balance:** The model maintains a strong balance between precision and recall, with both classes achieving F1-scores of 0.95.
- **Excellent Discrimination:** The AUC-ROC of 0.976 demonstrates the model's ability to effectively differentiate between default and non-default customers.
- **Well-Calibrated:** SMOTE and threshold tuning have helped the model handle imbalanced data effectively, as shown by the high recall for the ”Default” class.
- **Future Steps:** Fine-tuning thresholds or exploring cost-sensitive learning could further improve the trade-off between precision and recall depending on business needs.

ROC Curve and Key Metrics



- **AUC-ROC:** 0.976 - Indicates excellent performance in distinguishing between "Default" and "No Default."
- **Overall Accuracy:** 95% - 95% of predictions are correct.
- **Balanced Performance:** F1-scores of 0.95 for both classes indicate well-balanced precision and recall.

Classification Report - Precision, Recall, F1-Score

- **Class 0 (No Default):**
  - Precision: 0.92, Recall: 0.99, F1-Score: 0.95
- **Class 1 (Default):**
  - Precision: 0.99, Recall: 0.91, F1-Score: 0.95
- **Support:** Balanced dataset (56,536 "No Default" and 56,539 "Default").

Key Observations and Conclusion

- **High Recall:** 91% recall for the "Default" class shows significant improvement in identifying defaults.
- **Precision-Recall Trade-off:** Slight decrease in precision for "No Default" to achieve better recall for "Default."
- **Conclusion:** XGBoost with SMOTE delivers strong performance (AUC-ROC 0.976) and handles imbalanced data well. Future tuning could balance precision and recall further.

Precision-Recall Curve Analysis

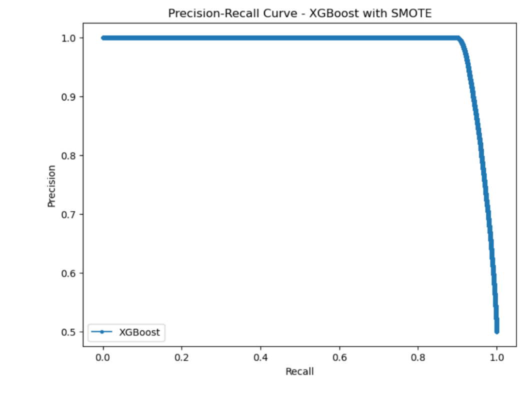


Figure 5: Precision-Recall Curve - XGBoost with SMOTE

Precision-Recall Curve Analysis

- **High Precision and Recall:** The model maintains near-perfect precision and recall across most thresholds, reflecting its accuracy in identifying defaults while minimizing false positives.
- **Sharp Drop at High Recall:** As recall approaches 1.0, precision drops, indicating a trade-off between capturing more true positives and introducing false positives.
- **Excellent Performance:** The curve's shape indicates a well-calibrated model, performing strongly across various thresholds.

Conclusion

- **Overall Performance:** The Precision-Recall curve shows that XGBoost with SMOTE performs exceptionally well in identifying defaults with minimal false positives.
- **Trade-off:** While precision slightly declines at maximum recall, this trade-off is expected when aiming to capture all positive instances.
- **Effectiveness:** The model demonstrates a strong balance between precision and recall, making it highly effective for credit risk prediction.

Cumulative Gains Chart Analysis

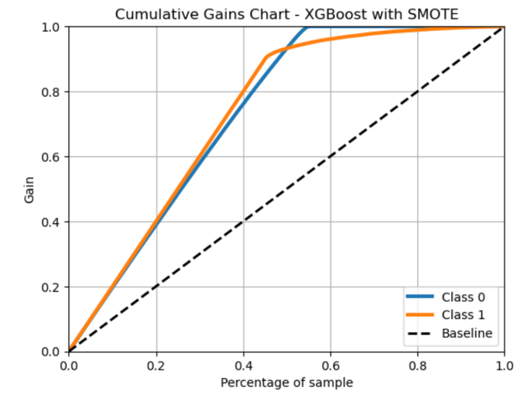


Figure 6: Cumulative Gains Chart - XGBoost with SMOTE

Cumulative Gains Chart Analysis

- **Class 1 (Default):** The model identifies nearly 80% of defaults by analyzing only 40% of the sample, effectively prioritizing high-risk customers.
- **Class 0 (No Default):** The model also performs well for "No Default" cases, although it is slightly less focused compared to "Default" cases.
- **Baseline:** The model significantly outperforms the baseline (random guessing), highlighting its strong predictive power in identifying positive cases.

Conclusion

- **Effectiveness:** The Cumulative Gains Chart shows that the XGBoost model with SMOTE is highly effective in identifying defaults early in the sample.
- **Optimized for Credit Risk:** The model is optimized for credit risk prediction, identifying a large percentage of defaults by focusing on a smaller portion of the population.
- **Strong Gains:** The model offers substantial gains over random selection, making it a valuable tool for prioritizing high-risk customers.

Lift Curve Analysis - XGBoost with SMOTE

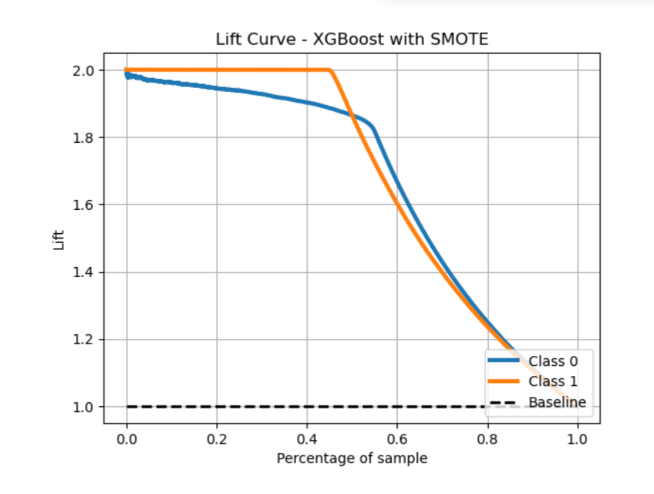


Figure 7: Lift Curve - XGBoost with SMOTE

Lift Curve Analysis - XGBoost with SMOTE

- **Lift at Start (Class 1):** The model achieves a lift of 2.0, meaning it is twice as effective as random selection in identifying defaults early on.
- **Decline in Lift:** As the sample size increases, lift decreases but remains above random selection for a significant portion of the sample.
- **Conclusion:** The model provides strong gains in identifying defaults early, critical in credit risk management.

Distribution of Predicted Probabilities - XGBoost

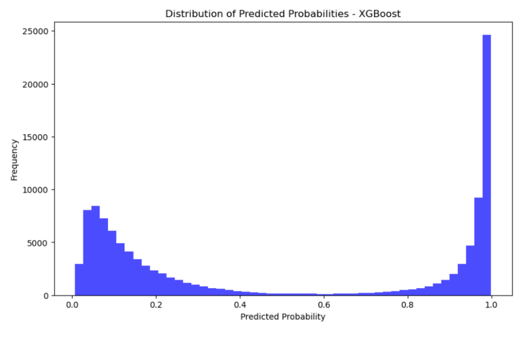


Figure 8: Distribution of Predicted Probabilities - XGBoost

Distribution of Predicted Probabilities - XGBoost

- **Bimodal Distribution:** Most probabilities cluster around 0 and 1, showing the model is confident in its predictions.
- **High Confidence:** The model assigns very high or very low probabilities, indicating reliable decision-making.
- **Few Mid-Range Predictions:** The lack of mid-range probabilities suggests the model rarely produces uncertain predictions.

Conclusion

- **Confident Model:** The XGBoost model confidently separates the two classes, providing reliable credit risk predictions.
- **Potential for Tuning:** Few mid-range probabilities could benefit from threshold tuning or further calibration to refine ambiguous decisions.

Calibration Curve - XGBoost with SMOTE

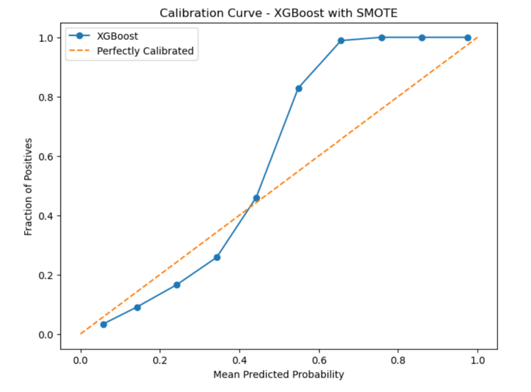


Figure 9: Calibration Curve - XGBoost with SMOTE

Calibration Curve - XGBoost with SMOTE

- **Perfect Calibration Line:** The orange dotted line represents a perfectly calibrated model where predicted probabilities match actual outcomes.
- **XGBoost Calibration:** The blue line deviates from the perfect line, suggesting overconfidence in high probability predictions and underconfidence in low probabilities.
- **Improvement Needed:** The model could benefit from calibration techniques like Platt Scaling or Isotonic Regression to better align probabilities with actual outcomes.

Conclusion

- **Calibration for Risk Management:** Well-calibrated probabilities ensure accurate risk assessment in credit modeling. Overconfidence may lead to riskier loan approvals, while underconfidence may result in overly conservative decisions.
- **Actionable Steps:** Implementing calibration techniques can help improve the model's accuracy in probability prediction.

Decision Function Plot - XGBoost with SMOTE (2D Projection)

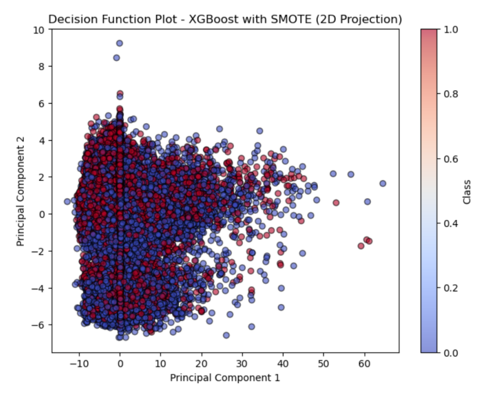


Figure 10: Decision Function Plot - XGBoost with SMOTE (2D Projection)

Decision Function Plot - XGBoost with SMOTE (2D Projection)

- **Class Separation:** Red indicates high probability of default (Class 1), blue indicates high probability of no default (Class 0), with some overlap.
- **Class Distribution:** Default instances cluster in specific areas but also mix with non-defaults, indicating areas of uncertainty.
- **Impact of SMOTE:** SMOTE has helped achieve better balance between the two classes in this projection.

Conclusion

- **Complex Decision Boundary:** The non-linear decision boundary is typical of tree-based models like XGBoost, showing both separability and challenges in classification.
- **SMOTE Effectiveness:** SMOTE helped in improving the representation of the minority class, though overlapping features still pose challenges.

SHAP Summary Plot Analysis - XGBoost

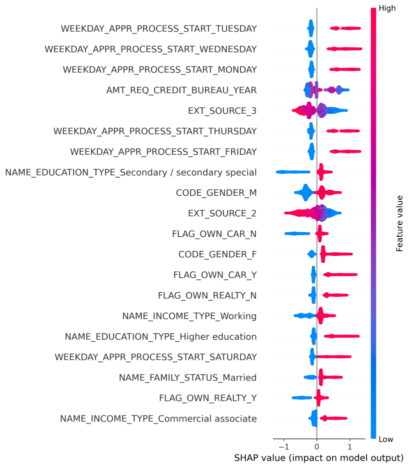


Figure 11: SHAP Summary Plot - XGBoost Model

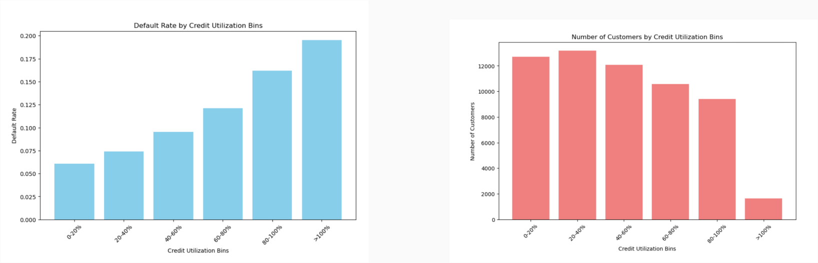
SHAP Summary Plot Analysis - XGBoost

- **Feature Importance:** The plot shows key features like external sources, application process weekdays, and income type significantly impacting model predictions.
- **Feature Value Impact:** Red indicates higher feature values, while blue indicates lower values, showing how different feature values push predictions towards or away from default.
- **Insights for Feature Engineering:** The insights from this plot can guide further analysis and improvements in feature engineering.

Conclusion

- **Valuable Insights:** The SHAP summary plot provides detailed insights into feature importance, offering a basis for refining the model.
- **Next Steps:** Potential improvements in feature engineering can be explored based on SHAP analysis, particularly around operational and behavioral data.

Credit Utilization and Default Risk Analysis



(a) Default Rate by Credit Utilization Bins

(b) Number of Customers by Credit Utilization Bins

Figure 12: Analysis of Default Rates and Customer Distribution by Credit Utilization

Key Observations and Risk Implications

- **Default Rates:** Higher credit utilization correlates with higher default rates, reaching nearly 20% for utilization above 100%.
- **Customer Distribution:** The majority of customers have utilization under 60%, with fewer customers in the higher utilization bins, but these higher bins have significantly increased risk.
- **Risk Implications:** Credit institutions should monitor high utilization customers closely, as they present a higher risk of default, particularly beyond 80% utilization.

Default Rate by Income and Credit Utilization

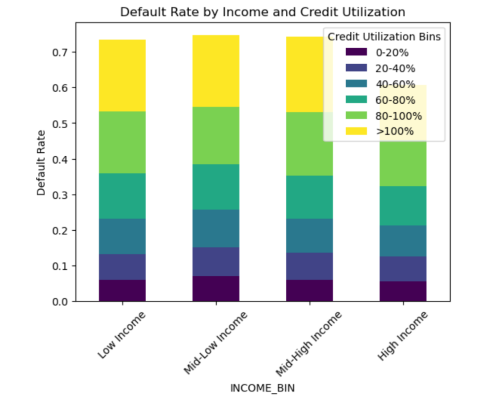


Figure 13: Default Rate by Income and Credit Utilization Bins

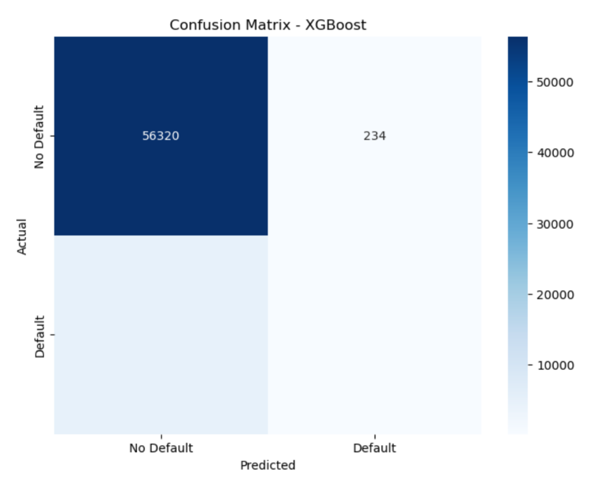
Key Insights

- **Default Rate Consistency:** Default rates are consistent across all income levels, indicating that income alone is not a major factor for default risk.
- **High Credit Utilization:** Higher credit utilization, especially above 80%, increases default risk across all income levels.
- **Lower Credit Utilization:** Customers with lower credit utilization ($0 - 40\%$) have a smaller impact on default rates.
- **Risk Thresholds:** Utilization above 60% significantly raises default risk, emphasizing the need to monitor customers with high credit usage.

Aggregating Credit Card Balance, Installments Payments, and Previous Applications

- **Credit Card Balance Data:** Aggregation involves calculating the total credit card balance, average balance, credit card utilization, and overdue amounts for each customer over time.
- **Installments Payments Data:** Aggregation includes summing total installments paid, overdue installments, and calculating average payment delays across different credit applications.
- **Previous Applications Data:** This includes summarizing past application statuses (approved, refused, canceled) and their features such as loan amounts, durations, and the overall outcome of the application process.

Confusion Matrix Analysis - XGBoost



Confusion Matrix - XGBoost Model (Merged Data)

Confusion Matrix Analysis - XGBoost

- **Model Performance:** The model performs well in identifying the majority class ("No Default"), with only 234 cases misclassified. This suggests that the model is highly accurate for the majority class.
- **Imbalanced Data:** The significant imbalance between the "No Default" and "Default" classes is evident, with a much smaller number of defaults being correctly predicted.

ROC Curve Comparison - XGBoost with and without SMOTE

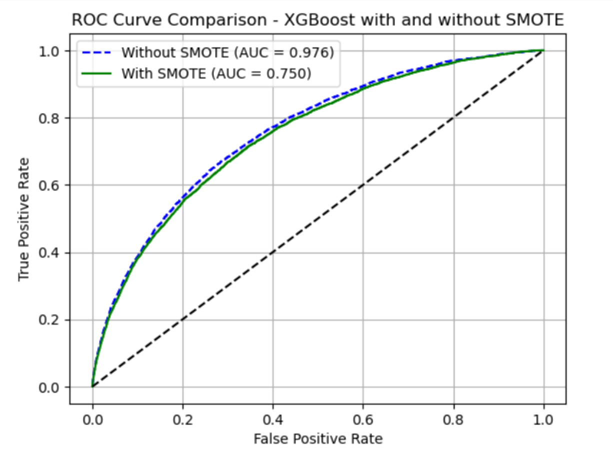


Figure 14: ROC Curve Comparison - XGBoost with and without SMOTE

Key Insights

- **Without SMOTE:** The model without SMOTE achieves a higher AUC score of 0.976. This indicates stronger discrimination between the ”Default” and ”No Default” classes, as the model is better at making correct predictions with the imbalanced dataset.
- **With SMOTE:** SMOTE improves class balance, but the model's performance decreases slightly, resulting in an AUC of 0.750. Synthetic oversampling may introduce noise, weakening the model's performance for the majority class.
- **Implications:** While SMOTE helps balance the dataset, it may reduce overall model performance. Threshold tuning and alternative oversampling techniques may be explored to improve performance.

ROC Curve - XGBoost with SMOTE

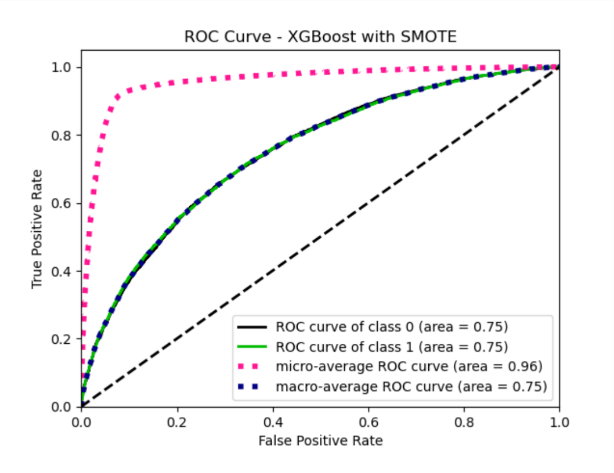


Figure 15: ROC Curve - XGBoost with SMOTE

Key Insights

- **Class ROC Curves:** Both Class 0 and Class 1 have an AUC of 0.75, reflecting moderate classification performance.
- **Micro-Average ROC Curve:** With an AUC of 0.96, the model shows strong overall performance across all instances.
- **Macro-Average ROC Curve:** The AUC of 0.75 for the macro-average indicates consistent performance across classes.
- **Takeaway:** While overall performance is strong, distinguishing between defaulters and non-defaulters needs improvement. Further tuning could enhance class-specific accuracy.

Precision-Recall Curve - XGBoost with SMOTE

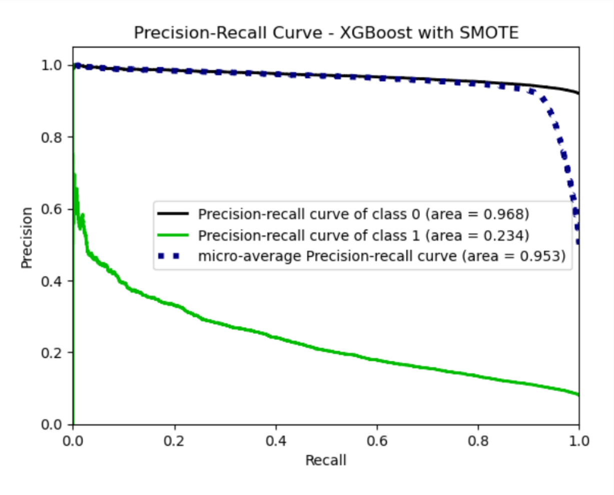


Figure 16: Precision-Recall Curve - XGBoost with SMOTE

Key Insights

- **Class 0 Performance:** The precision-recall area for class 0 (No Default) remains high at 0.97, indicating excellent accuracy in predicting non-defaults.
- **Class 1 Performance:** Despite applying SMOTE, the precision-recall area for class 1 (Default) is lower, around 0.234, showing that default predictions are still challenging.
- **Micro-Average Curve:** The micro-average precision-recall curve shows an area of 0.95, reflecting a strong overall balance between precision and recall across all classes.
- **Conclusion:** SMOTE improves class balance but only slightly affects the trade-offs between precision and recall, especially for predicting defaults. The model remains very effective at predicting non-defaults, while further improvement is needed for default predictions.