Loan Default Prediction Model Evaluation (SMOTE)

1. Introduction

This report presents the evaluation of three machine learning models trained on an imbalanced dataset using SMOTE (Synthetic Minority Over-sampling Technique) and Threshold Tuning. The objective is to identify the most effective model for predicting loan defaults within the SACCO.

2. Model Performance Summary

Model	Precision (Defaulters)	Recall (Defaulters)	F1-Score (Defaulters)	AUC-ROC
Logistic Regression	0.41	0.99	0.58	0.99
Random Forest	0.51	0.96	0.67	0.99
Gradient Boosting	0.45	0.99	0.62	0.99

3. Interpretation of Results

All three models were evaluated with class balancing (SMOTE) and a tuned decision threshold to improve classification performance on the minority class (defaulters).
- Logistic Regression achieved the highest recall (0.99) but with the lowest precision (0.41), indicating a high false positive rate.
- Random Forest showed the best balance between recall (0.96) and precision (0.51), yielding the highest F1-score (0.67).
- Gradient Boosting also achieved high recall (0.99) with moderate precision (0.45), placing it between the other two in overall performance.
All models recorded an excellent AUC-ROC of 0.99, demonstrating strong discriminatory ability.

4. Conclusions

1. Random Forest stands out as the most balanced and effective model for identifying loan defaulters, combining high recall with acceptable precision and the highest F1-score.
2. Logistic Regression is suitable when model transparency and interpretability are critical, although it may over-flag non-defaulters.
3. Gradient Boosting offers a middle ground with competitive metrics.

5. Recommendations

- We recommend deployment of the Random Forest model in the Banks/SACCO's loan processing system.
- Gradient Boosting could be used when conservative default identification is acceptable.
- Logistic Regression might be useful when regulatory interpretability is essential.
- There is a need to Continuously monitor model performance and update the threshold based on business needs.
- Consider feature importance tools like SHAP for explainability.

6. Next Steps

1. Integrate the Random Forest model into the Loan management system.
2. Conduct quarterly model reviews and retraining.
3. Set up dashboards for tracking precision, recall, and default detection rates.
4. Evaluate additional ensemble techniques or cost-sensitive methods.
5. Provide training to credit officers on model interpretation and usage.