Loan Default Prediction

Report 2: Findings & Recommendations
Project Title: Loan Default Prediction for GhanaLoanConnect

Organized by Thrive Africa
Machine Learning 2025 Cohort

Submitted by: Patrick Ayeh Ayisi (PM)
Date: June 02, 2025

TEAM (ALPHA-2)

Patrick Ayeh Ayisi

Umar Zaidu Yakubu

David Wilson

Owiredu-Amoh Baffuor Etto Jnr.

Marvin Dawson

Jemima Ama Fiapemetsi

Iwikotan Oreofe Gloria

Maya Leotina Agebedekpui

Key Insights from the Model

Target Variable
`not.fully.paid` was used to represent loan default (1 = defaulted).

Data Characteristics

- Total records: 9,578 borrowers
- No missing values detected
- Class imbalance present in the target variable (~84% not defaulted)

Features That Influence Default

Top contributing features from feature importance (Random Forest & Gradient Boosting):
1. `int.rate` – Higher interest rates correlated with higher default risk
2. `fico` – Lower FICO scores were strong predictors of default
3. `revol.util` – High revolving credit usage increased risk
4. `long_credit_history` – Borrowers with long credit history were less likely to default
5. `purpose` – Certain loan purposes (e.g., small business) showed higher default rates

Model Performance

Model	Accuracy	Precision	Recall	| F1 Score	| ROC AUC
Logistic Regression	0.90	0.91	0.97	0.94	0.95
Gradient Boosting	0.99	0.99	0.997	0.99	0.995

Gradient Boosting proved to be the best-performing model on all evaluation metric

Addressing Class Imbalance

- SMOTE was applied to oversample minority class (defaults)
- Resulted in better recall and reduced false negatives

Confusion Matrix (Baseline Model)

Actual / Predicted	0 (No Default)	1 (Default)
0	325	233
1	59	2257

The baseline model struggled with identifying non-defaults. This was improved significantly after using Gradient Boosting.

Key Findings

- High interest rates and low FICO scores are the most critical indicators of default.
- Borrowers with shorter credit history and high revolving credit utilization also present a higher risk.
- The loan purpose has a significant impact—business and educational loans tend to carry more default risk.
- Class imbalance significantly affects prediction, and applying SMOTE greatly improves model fairness.

Recommendations for GhanaLoanConnect

1. Use the Gradient Boosting model in production for more accurate loan risk predictions.
2. Implement credit history-based risk segmentation to tailor interest rates.
3. Encourage higher-FICO-score borrowers through customized offers to reduce defaults.
4. Regularly retrain the model with new data to adapt to changing borrower behavior.
5. Consider excluding or tightening criteria for high-risk purposes like small business loans unless additional guarantees are secured.

Challenges & Solutions

1. Class Imbalance:
    - The dataset had a heavy skew toward non-defaulters.
    - Solution: Used SMOTE to balance classes and reduce bias.
2. Feature Interpretability:
    - Tree-based models are hard to interpret.
    - Solution: Used feature importance plots to explain key drivers of predictions.
3. Model Overfitting:
    - Some models (e.g., Random Forest) initially overfit the training data.
    - Solution: Tuned hyperparameters using GridSearchCV and evaluated with cross-validation.
4. Target Variable Confusion:
    - Initially used `credit.policy` instead of `not.fully.paid`.
    - Solution: Corrected to use the proper label for modeling default risk.

Conclusion

The project successfully built a robust model that can predict loan defaults with high accuracy. Key risk indicators have been identified, and the model can be used to reduce GhanaLoanConnect’s exposure to risky borrowers. With further refinements, this system can enhance the company’s data-driven lending strategy.