Loan Default Prediction Analysis

Leveraging Machine Learning to Mitigate Credit Risk

Group 7

Betty Koila
Alex Irungu
Susan Wanjiru
James Kosgei
Michael Arita
Brian Amani

Business Problem

Financial Losses

Institutions lose millions annually due to defaults.

Risk Management

Need to optimize lending decisions.

Improve Lending Decisions

Lenders struggle with selecting the right criteria for lending, leaving many Kenyans unbanked

Objective

To build a loan prediction model that can be used scalably across multiple financial institutions to help stakeholders:

- Identify high-risk loans early.
- Optimize lending strategies.
- Reduce financial losses due to defaults

Data Preprocessing

Data Understanding

1. 115,893 loan records with 18 features (demographics, credit scores, loan details)
2. 19.3% of the borrowers in the dataset are defaulters.

Our final cleaned dataset comprised of:

Demographic Variables

- Gender, Age, Marital Status
- Employment Status

Financial Variables

- Credit Score, Net Income
- EMI, Principal Disbursed
- Overdraft Amount

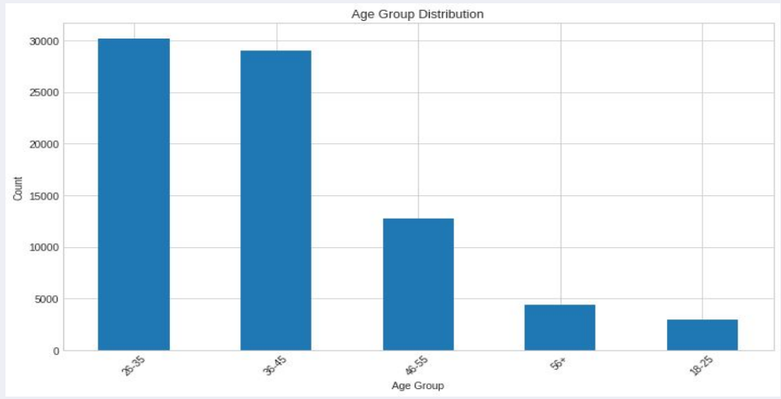
Key Steps in Data Cleaning

- Handled missing values
- Created DEFAULT_BINARY
- Added AGE_GROUP categories
- Undersampled majority class

Demographic Analysis

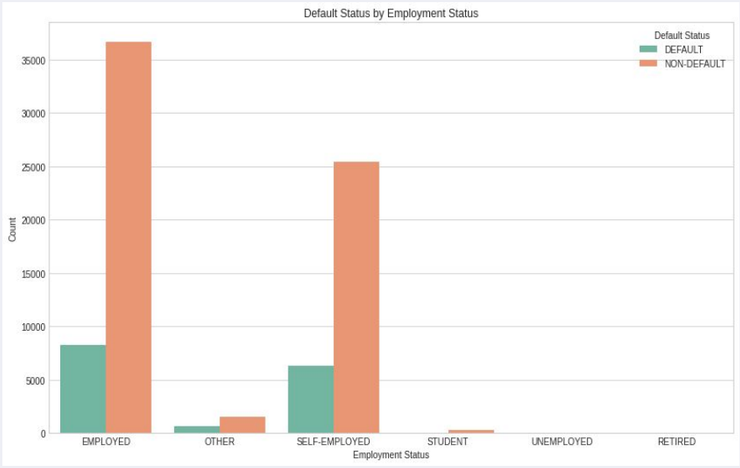
Age Groups

36-45 and 26-35 age groups comprise most borrowers.



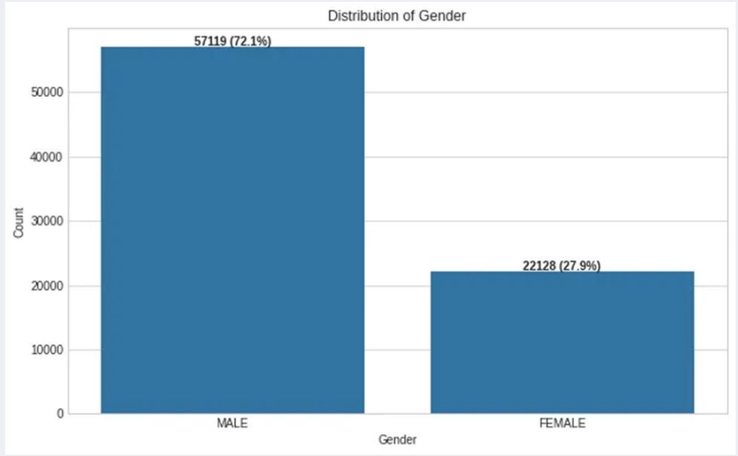
Employment

Self-employed borrowers show higher default rates.



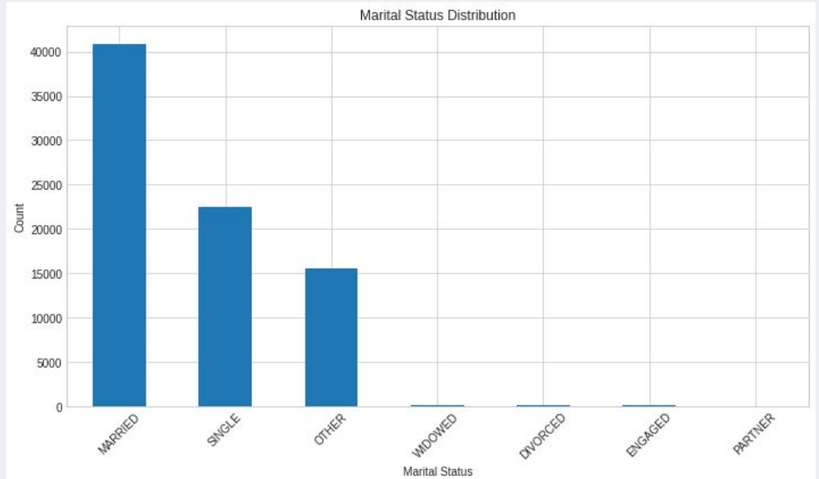
Gender

72.1% male borrowers vs 27.9% female borrowers.

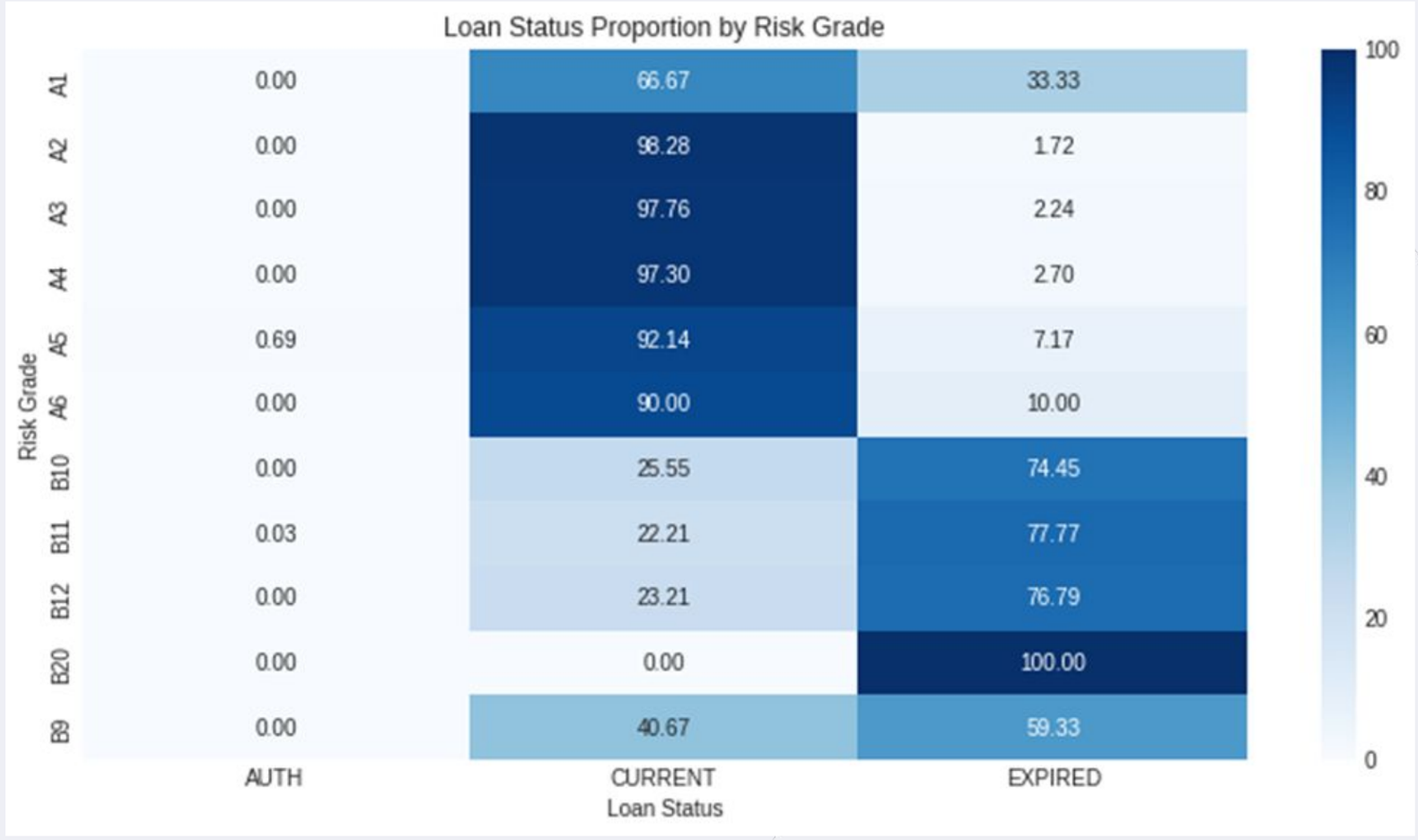


Marital Status

Majority of borrowers are married (40,000+).



Risk Analysis



Low Risk (A2-A5)

Mostly current loans

Medium Risk (B1-B9)

Mixed loan status

High Risk (B10-B12)

Significant expired loans

Critical Risk (B20)

100% expired loans

Variable Correlations

Variables	Correlation	Significance
EMI & Principal	0.39	Moderate
Financial Variables	Low	Independent predictors
Net Income	Minimal	Stands alone
Credit Score & OD_AMOUNT	-0.068	Negative

Model Development

Performance Metrics:

Model	Accuracy	F1 Score	ROC AUC
XGBoost (Tuned)	97.58%	0.976	0.995
Random Forest	97.58%	0.976	0.995
Gradient Boosting	97.58%	0.976	0.995
Decision Tree	96.10%	0.961	0.961
Logistic Regression	92.01%	0.925	0.915

Logistic Regression

Baseline model for interpretability.

Decision Tree

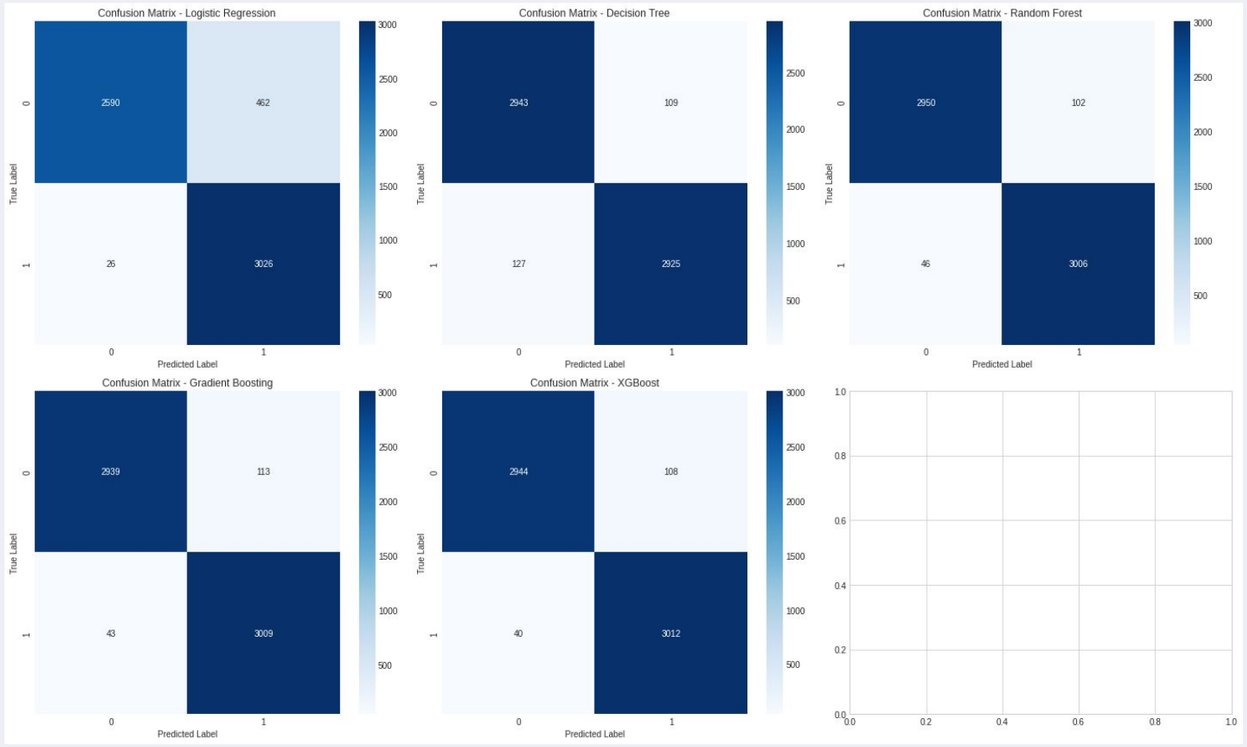
For transparent decision rules.

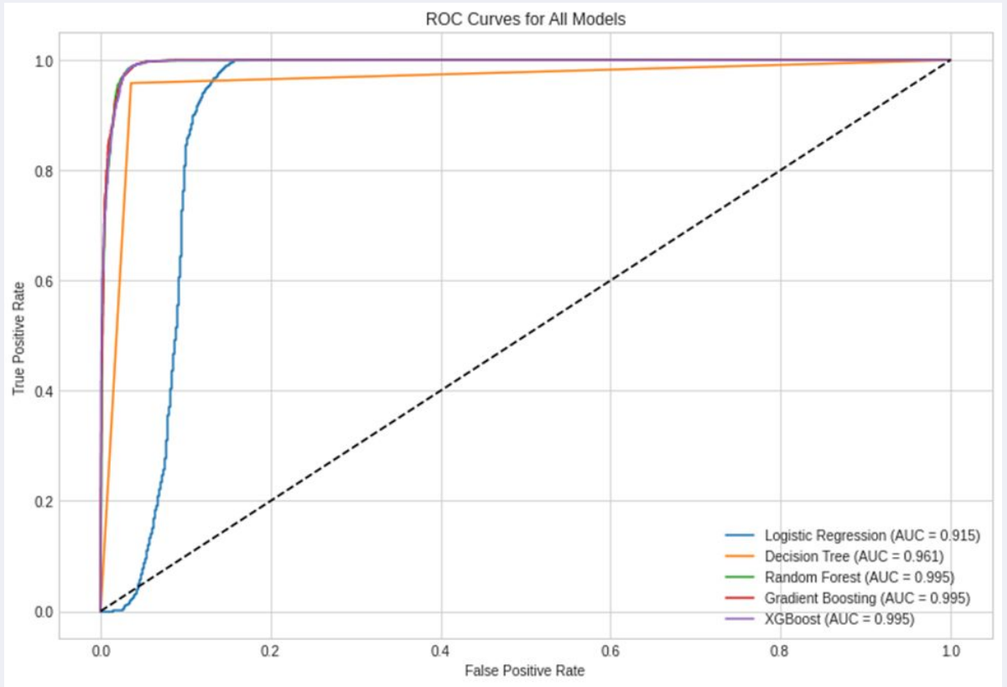
Random Forest

For ensemble learning power.

XGBoost

For optimized gradient boosting.





Model Performance & Business Impact

97.58%

XGBoost Accuracy

Best overall performance.

0.995

ROC AUC

Exceptional discrimination ability.

3,010

Defaults Predicted

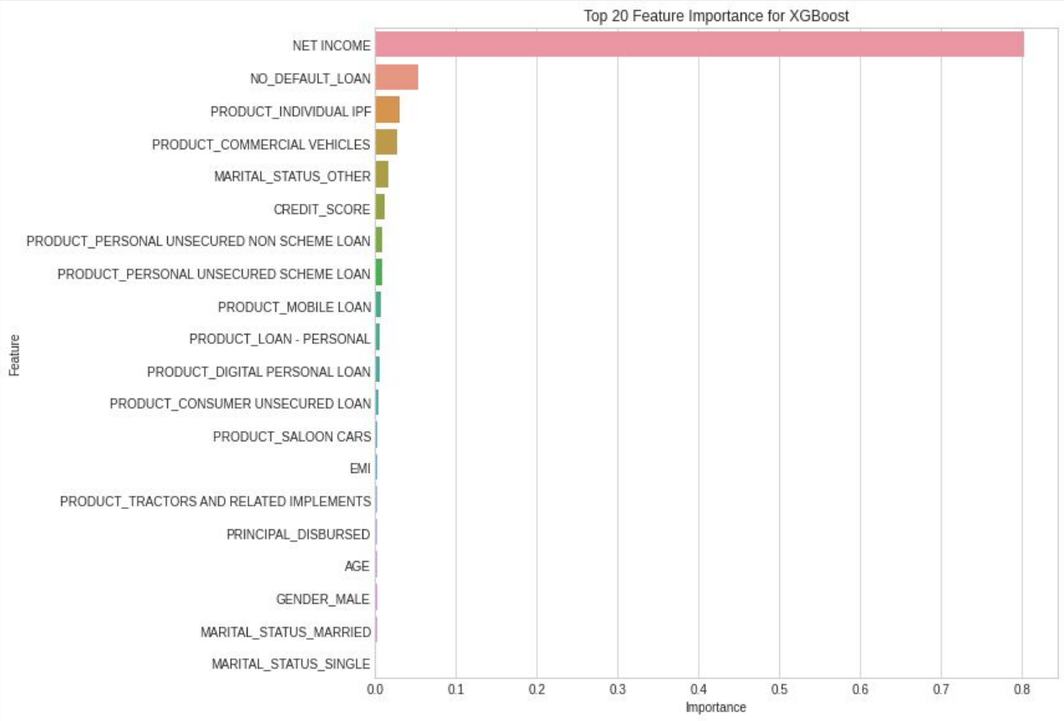
Out of 6,104 test cases.

155.97M

Net Value (Ksh)

Total business impact.

Key Insights



Top 10 Important Features:

1. **NET INCOME** (0.803083) - By far the most influential predictor
2. **NO_DEFAULT_LOAN** (0.053243) - Previous loan performance
3. **PRODUCT_INDIVIDUAL IPF** (0.029323)
4. **PRODUCT_COMMERCIAL VEHICLES** (0.027311)
5. **MARITAL_STATUS_OTHER** (0.015945)
6. **CREDIT_SCORE** (0.011784)
7. **PRODUCT_PERSONAL UNSECURED NON SCHEME LOAN** (0.008060)
8. **PRODUCT_PERSONAL UNSECURED SCHEME LOAN** (0.008030)
9. **PRODUCT_MOBILE LOAN** (0.007030)
10. **PRODUCT_LOAN - PERSONAL** (0.004991)

Income is Paramount

Net income is the strongest predictor (80.3% importance).

Product Risk Varies

Commercial vehicles and mobile loans need close monitoring by the financial institutions that provide them

Employment Matters

Self-employed borrowers show higher risk profiles.

Prior Behavior Predicts

Previous defaults strongly indicate future risk (5.3% importance).

Recommendations & Next Steps

1. **Include Enhanced Features:** Incorporate macroeconomic indicators and consumer behavior metrics.
2. **Explainability:** Implement SHAP values for transparent risk explanations to customers
3. **Default Prevention:** Develop early intervention programs for high-risk customers
4. **Deployment:** Integrate model into loan approval workflows with A/B testing.