CREDIT RISK CLASSIFICATION ANALYSIS AND MODELING USING MACHINE LEARNING

by patrice okoiti

Date: 04/06/2024

BUSINESS PROBLEM

Problem Statement:

Mambo Leo commercial bank seeks to improve its ability to assess the creditworthiness of loan applicants to reduce default risk. The bank wants to use historical loan data to identify key risk indicators and build a predictive model that can accurately classify applicants as likely to default or not.

Target:

- Loan officers, risk analysts, and credit managers who need tools to assess risk more accurately.
- Regulatory teams who require transparent and interpretable credit models.
- Data science teams responsible for model development and monitoring.

Challenges:

- Understanding applicants' profile.
- Identifying most influential credit risk features.
- Developing model with metrics over 80%.
- Balancing predictions and ethical considerations.

Solutions:

- Analyze historical applications features against loan status for patterns.
- Train and evaluate multiple classification models.
- Provide actional insights and recommendations.

Objectives

- Identify key drivers of loan default
- Predict likelihood of default with ≥80% accuracy.
- Prioritize recall to minimize missed defaulters (false negatives)
- Provide interpretable outputs for regulatory and policy alignment

Analysis Methodology

Data Split

25% training, 75% testing dataset.

Models Tested

- Logistic Regression
- Decision Trees
- XGBoost Classifier

Optimization

- Hyperparameter tuning
- SMOTE for class imbalance

Performance Metrics

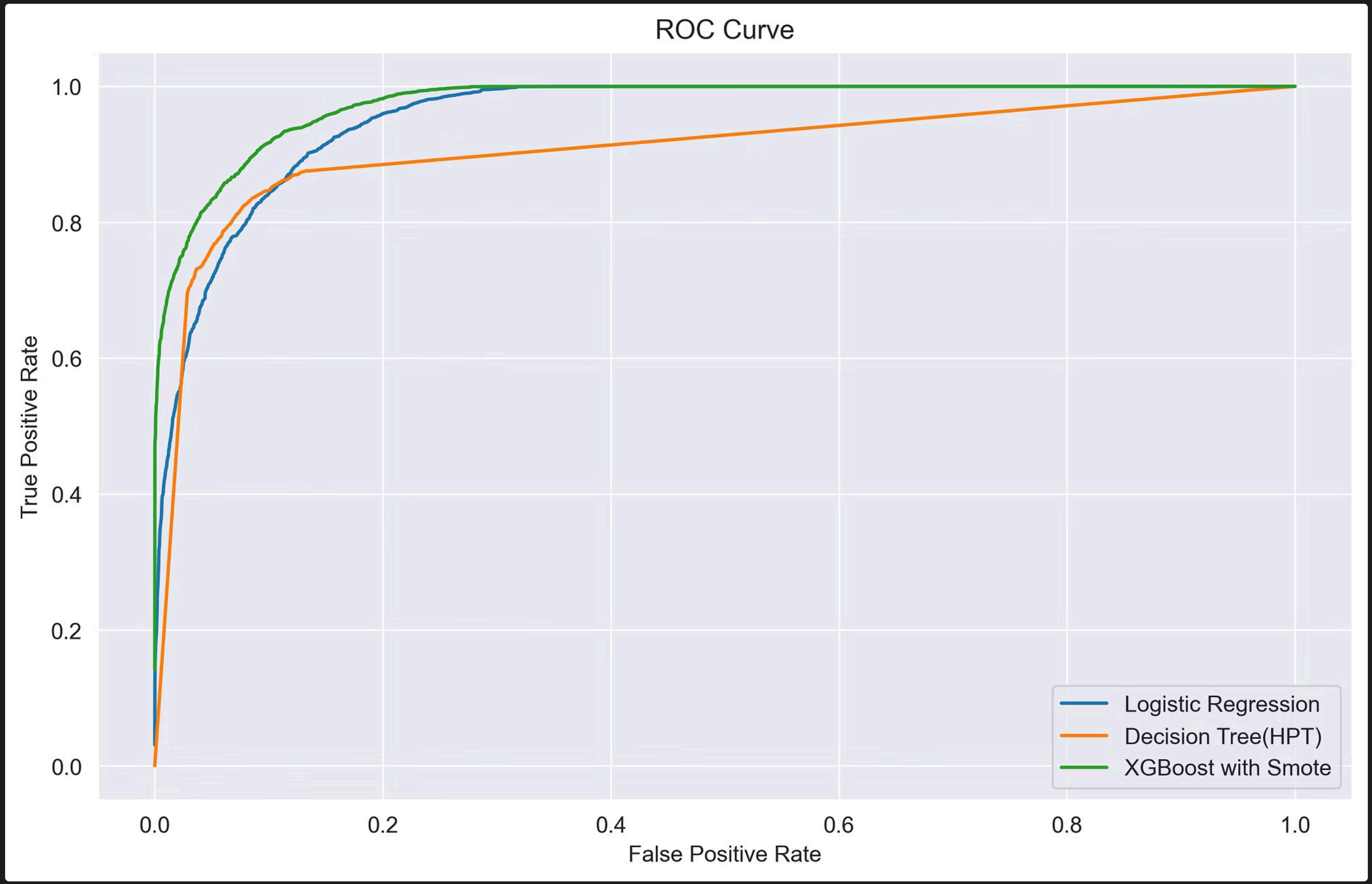
- Accuracy, recall, precision, F1-score, ROC-AUC

Model Evaluation

Model	Precision	Recall	F1-Score	Accuracy	AUC
1. Logistic Regression	78.48	74.76	76.57	89.87	95.60
1. Decision Tree (HPT)	81.85	75.44	78.51	90.85	90.77
1. XGBoost + Smote	83.02	82.78	82.90	92.43	97.57

Best Model: XGBoost has best recall, auc and overall metrics

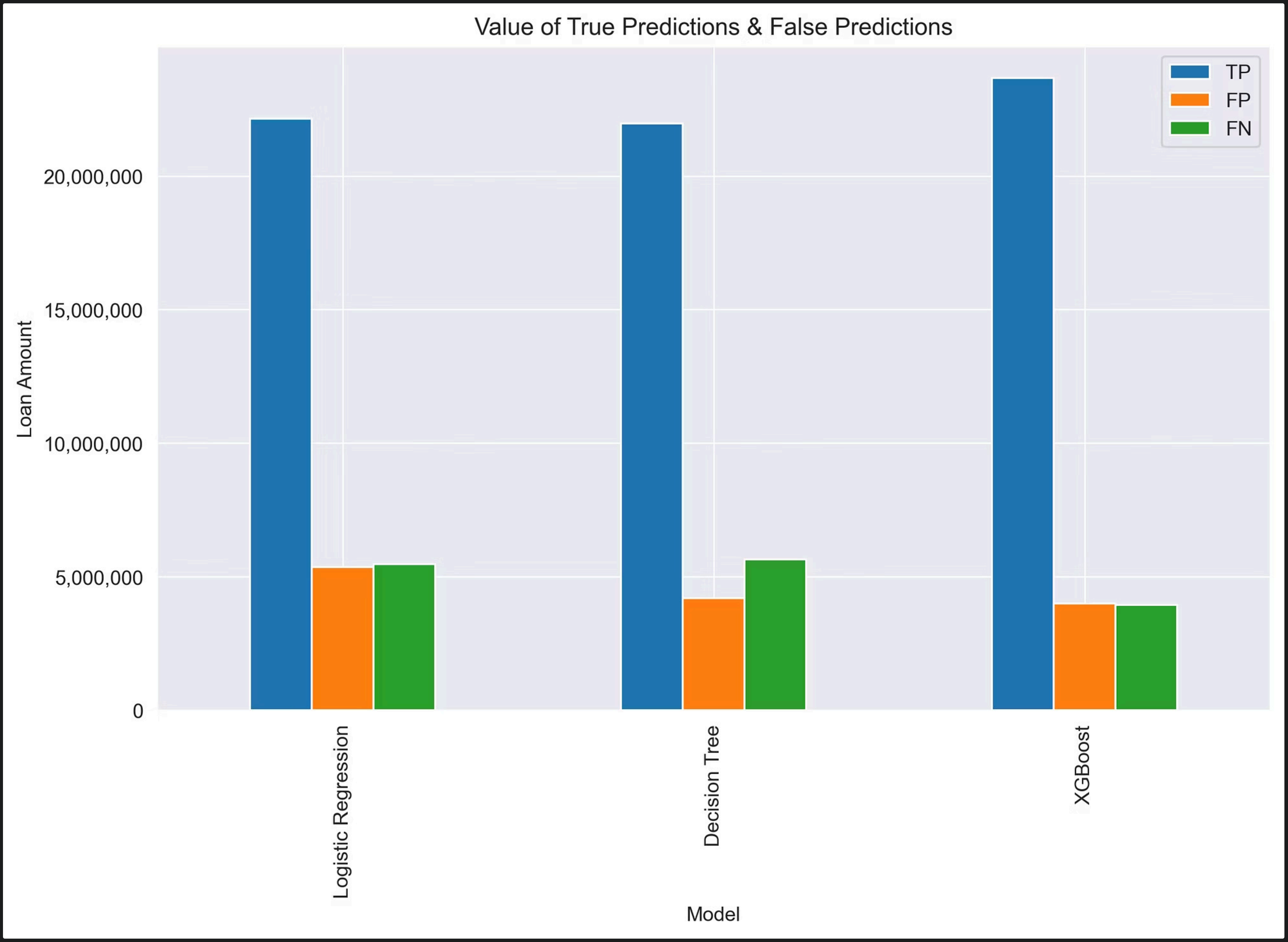
Roc Curve Model Performance



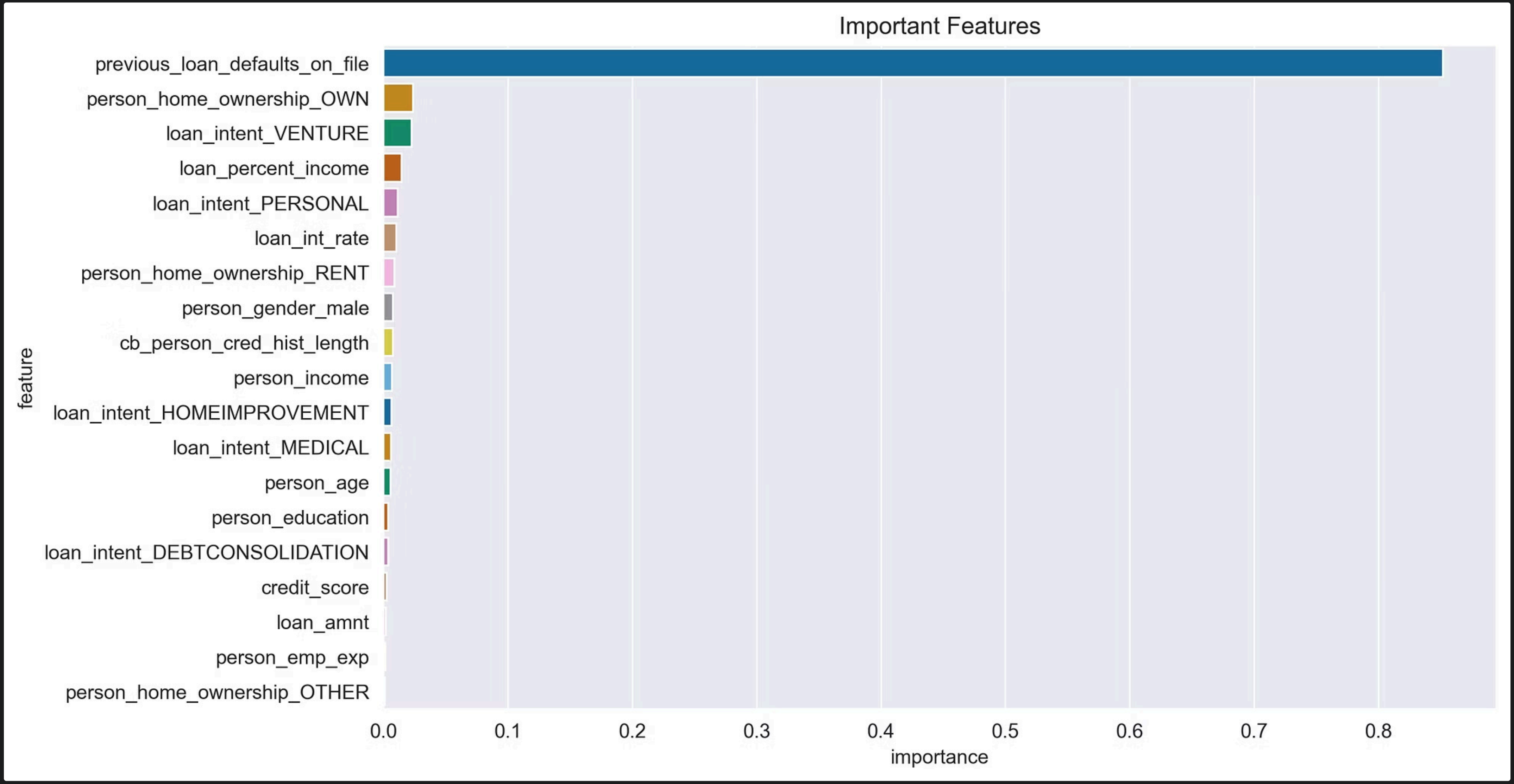
XGBoost hugs the top left corner- Indicates high true positive rates and low false positive rate.

Value of Predictions

Name	TP	FP	FN
Logistic Regression	22150488.0	5370141.0	5466519.0
Decision Tree	21965863.0	4192675.0	5651144.0
XGBoost	23663097.0	3997891.0	3953910.0


Value of True Predictions & False Predictions
TP
FP
FN
20,000,000
15,000,000
10,000,000
5,000,000
0
Loan Amount
Logistic Regression
Decision Tree
XGBoost
Model

Influential Features Contributing to Credit Risk



Key Predictors of Default

- Default History
- Home Ownership
- Loan Purpose
- Loan-to-Income Ratio
- Interest Rate
- Credit History

Risk Segmentation Strategy

1

Low Risk (0-10%)

Pre-approved for premium products.

2

Moderate Risk (11-30%)

Standard offers with verification required.

3

High Risk (31-60%)

Needs collateral or guarantor.

4

Very High Risk (61%+)

Alternative products or loan decline.

Implementation Recommendations

1. ### XGBoost + SMOTE
   Use as primary credit-risk scoring model for loan risk assessment.

2. ### Integration
   Incorporate with current loan approval workflows.

3. ### Ongoing Optimization
   Retrain quarterly and use A/B testing to compare performance.

Next Steps

- Approve Segmentation Thresholds
  - IT Platform Integration
    - Staff Training

      Interpret model outputs confidently.
      - Regulatory Compliance

        Complete documentation for audits.
    - Phase 1 Rollout

      Pilot with small personal loans.

CONCLUSION

- All three models surpassed the 80% accuracy threshold, but XGBoost with SMOTE clearly leads across every key metric, especially recall and AUC, which are paramount in credit-risk contexts. By catching more actual defaulters while keeping false positives relatively low, it offers the best risk-mitigation payoff for the bank. Continuous monitoring and periodic retraining will ensure it remains reliable as market conditions and borrower profiles evolve.