Katherine Gabriana – MIT PE Applied Data Science Program

Driving Data-Driven Lending Decisions

LOAN DEFAULT PREDICTION

Financial Innovation Through Machine Learning

Problem Definition

Loan defaults cause significant financial losses to lenders.

Traditional credit scoring lacks nuance in borrower behavior.

Early prediction of high-risk applicants can reduce bad debt.

Problem to Solve

Can we build a predictive model to identify loan applicants likely to default?

Which borrower features are the strongest indicators of risk?

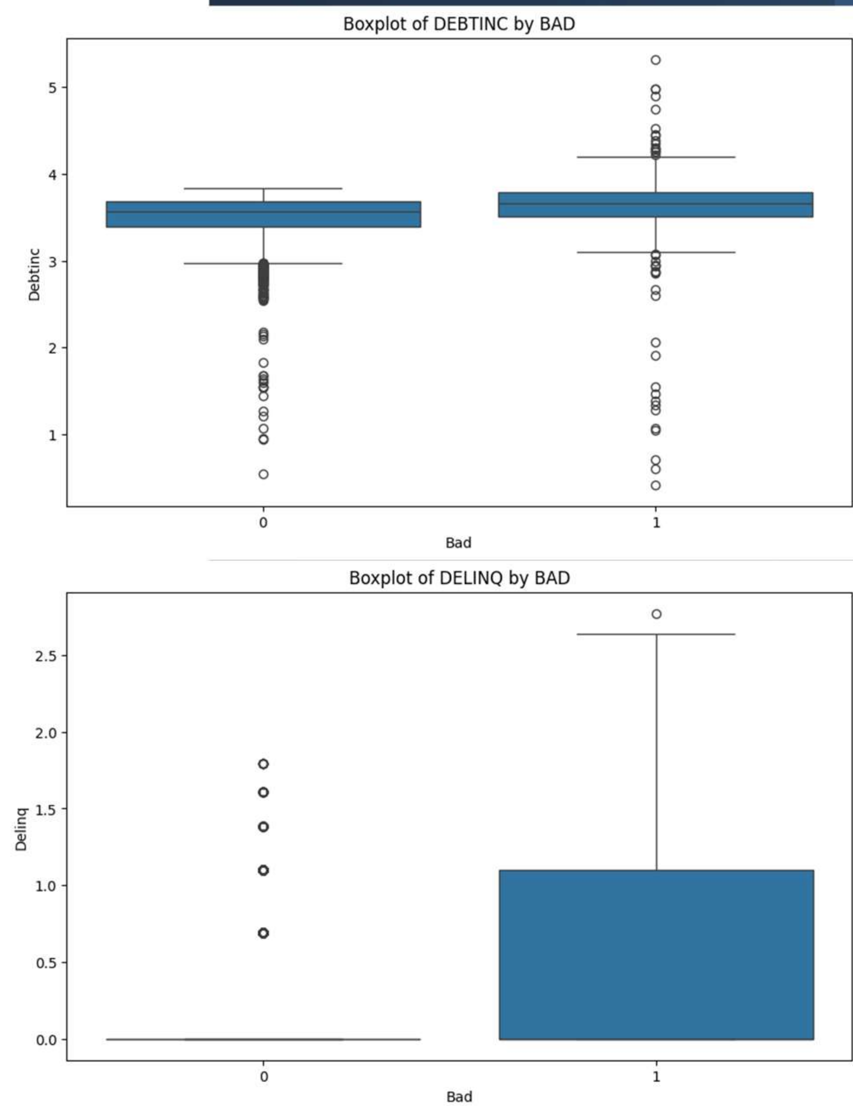
How should we implement this model?

Exploratory Data Analysis (EDA)

Missing Value Treatment

Addressing Skewness

Scaling



Financial Engineering

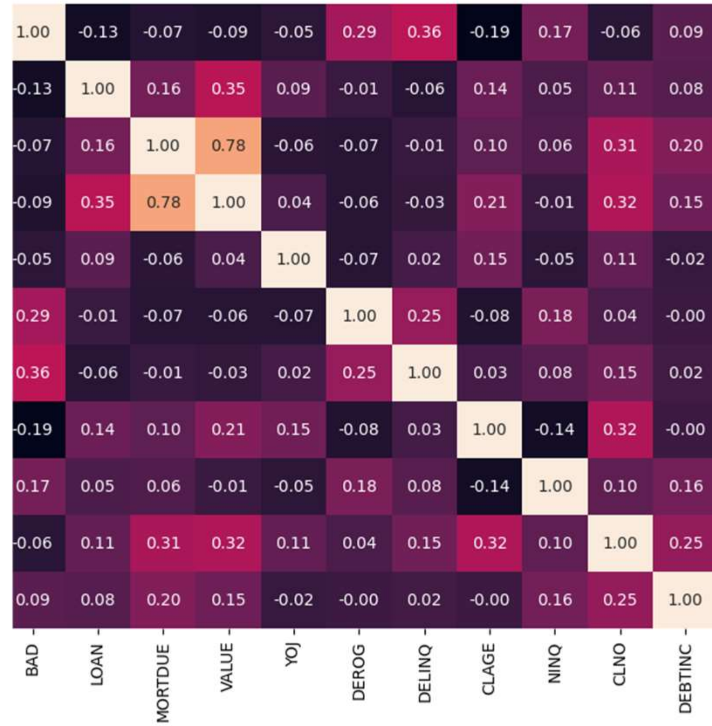
Best Recall and Balanced F1 for Loan Default Detection

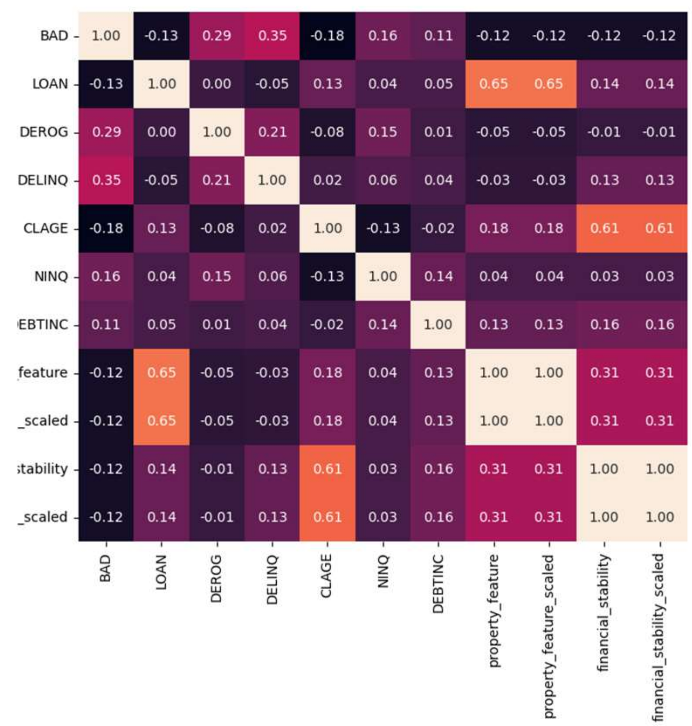
Property Feature

A feature combining effects of loan, value, and mortgage due.

Financial Stability Score

A measure of borrower stability, calculated by multiplying account age and the number of credit lines.





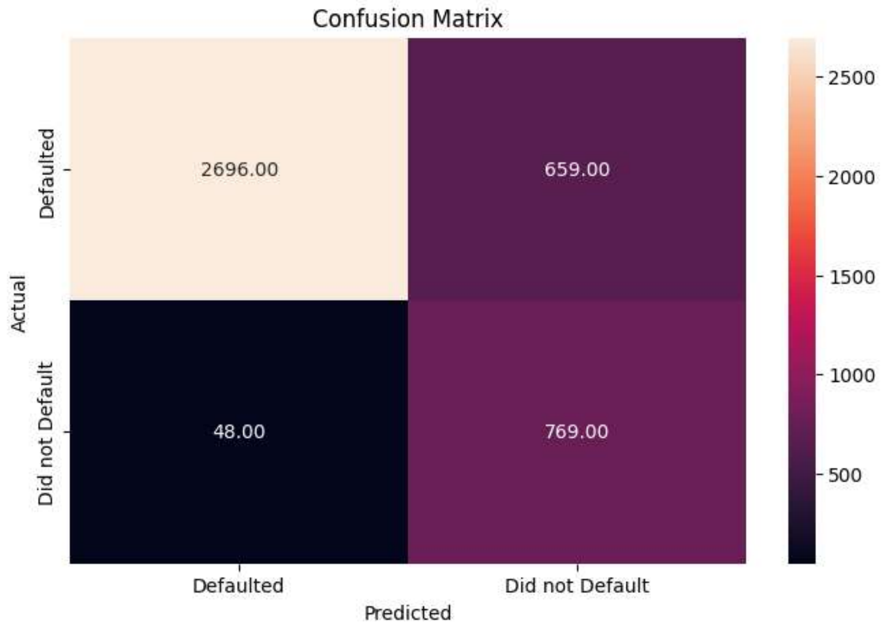
Solution Approach

XGBoost Wins on Default Recall

MODELS: Decision Trees, Random Forest, XGBoost.

OPTMIZE FOR RECALL: Defaults are Expensive

ANALYZE METRICS: XGBoost had best results



XGBoost Tuned - Test Results:

	precision	recall	f1-score	support
0	0.97	0.80	0.87	1416
1	0.53	0.89	0.67	372
accuracy			0.82	1788
macro avg	0.75	0.84	0.77	1788
weighted avg	0.88	0.82	0.83	1788

XGBOOST Selected

Best Recall and Balanced F1 for Loan Default Detection

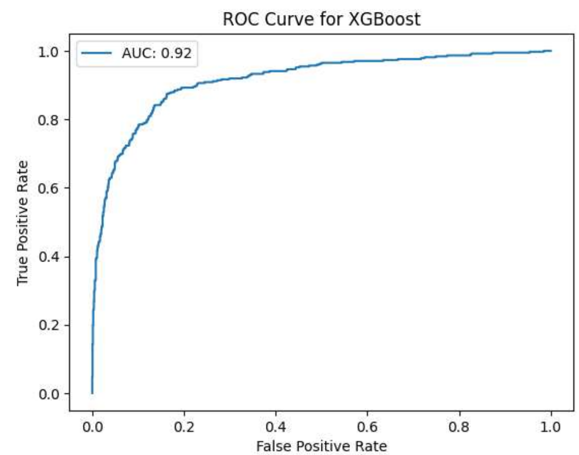
RECALL ON DEFAULTERS:

89% - BEST AMONG MODELS

BUSINESS LOGIC:
DEFAULTS ARE COSTLIER THAN FALSE POSITIVES

F1 SCORE (WEIGHTED):
83% - BALANCES PRECISION AND RECALL

SHAP EXPLAINABILITY:
ENABLES TRUST + COMPLIANCE



Metric	Decision Tree Tuned	XGBoost Tuned	Random Forest Tuned
Accuracy	82.4%	81.5%	85.2%
Weighted F1 Score	83.4%	82.9%	85.7%
Class 0 F1 Score	88.2%	87.2%	90.4%
Class 1 F1 Score	65.3%	66.8%	67.8%
Recall (Class 0)	83.0%	79.5%	87.9%
Recall (Class 1)	79.8%	89.2%	74.7%
Precision (Class 0)	94.0%	97.0%	94.0%
Precision (Class 1)	55.0%	53.0%	55.0%

SHAP Analysis

Identifies Key Drivers of Default Risk, Enabling Transparent Decisions

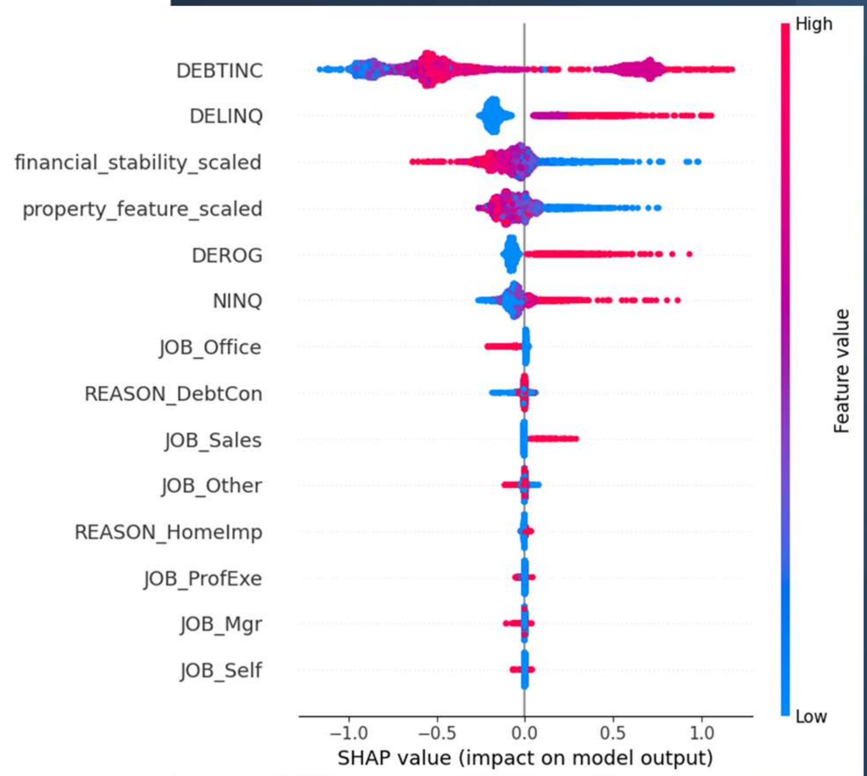
Key Outcomes:

Debt-to-Income Ratio (DEBTINC):
Strong positive correlation with default risk.

Delinquency History (DELINQ):
Past delinquencies are predictive of future defaults.

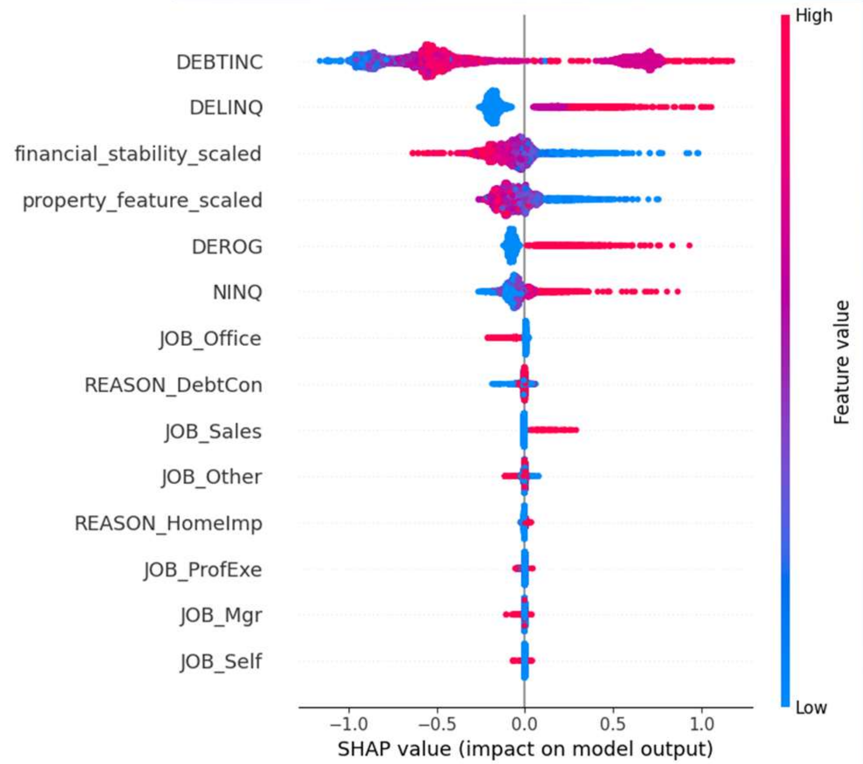
Financial Stability Score:
Lower scores align with higher risk profiles

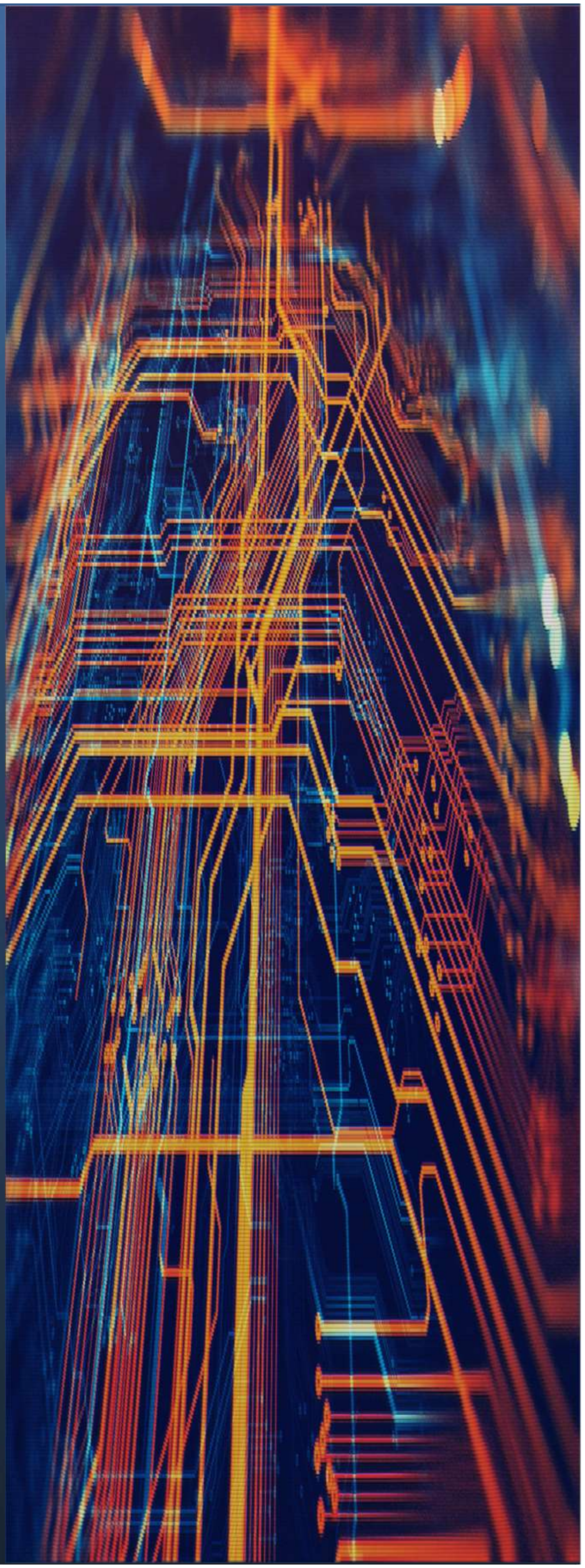
Property Feature:
Borrowers who own property and have an active mortgage are often associated with greater stability



SHAP Analysis
Recommended Action

Introduce	**Introduce** targeted interventions such as stricter lending limits for applicants with DEBTINC exceeding a certain threshold.
Develop	**Develop** loan products tailored to individuals with minor delinquencies but good repayment patterns (e.g., higher interest rates, shorter tenures, or smaller loan amounts).
Identify	**Identify** Offer more favorable terms for people with greater financial stability. Help those without much credit history by offering secured credit cards or small loans.
Consider	**Consider** lower interest rates or extended loan tenures for applicants with property ownership.





Proposed Business Solution and Execution

Deploy Model: Integrate with loan origination systems using calibrated thresholds.

Flag High-Risk Applicants: Manual review or alternative credit scoring.

SHAP Dashboard: Support transparent lending decisions.

Integrate with Existing Systems: Ensure seamless integration.

Define Score Thresholds: Set for auto-decisioning and manual review.

Train Staff: Equip with SHAP-based explainability tools.

Monitor and Retrain: Establish quarterly cadence.

Business Impact

Data Summary

- **Total Loan Amount**: $110,903,500
- **Loans in Default**: $20,120,400
- **Recall (Class 1):** 89.2%
  - XGBoost model correctly identifies 89.2% of defaults.

Defaults Detected

- **Formula**: Recall × Loans in Default
- **Result**: $17.96M

Missed Defaults

- **Formula**: Loans in Default × (1 - Recall)
- **Result**: $2.16M

Savings from Detection

- **Assumption**: 50% mitigation rate
- **Result**: $8.98M

Risks and Challenges

Data drift or economic shifts may affect model accuracy.

FICO, employment trends, and additional data can enhance accuracy.

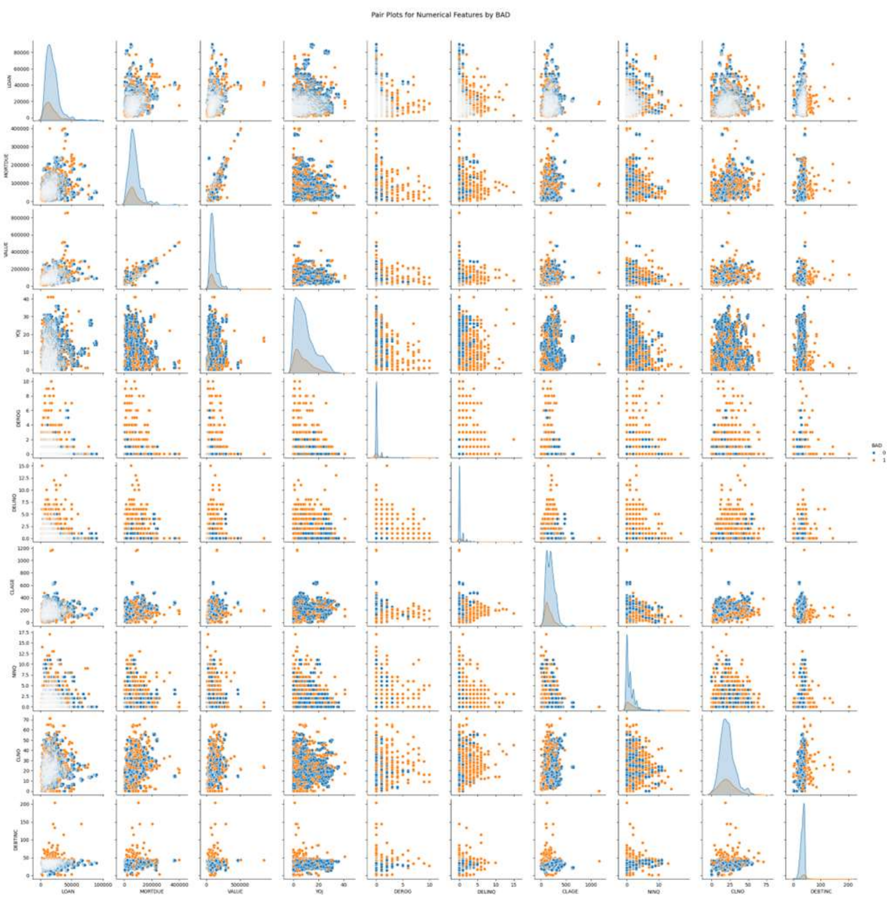
Potential bias in features must be monitored continuously. Make sure model adheres to regulatory requirements.

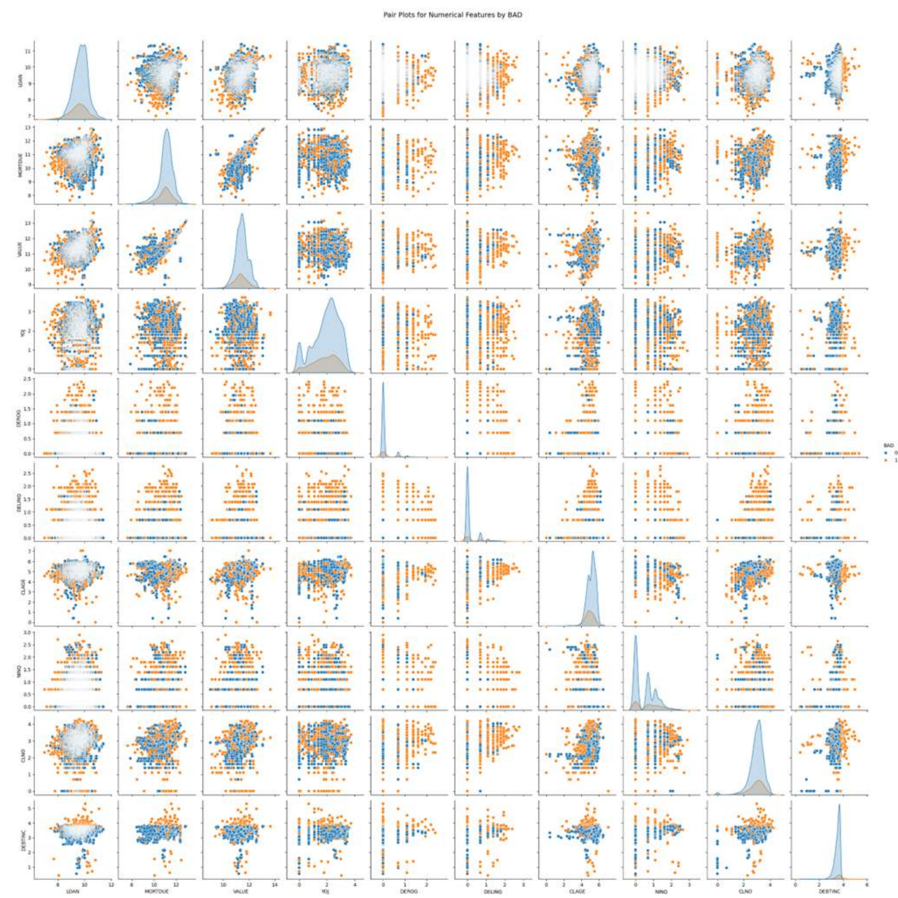
Periodic retraining is essential for sustained performance.

Appendix

Bivariate Analysis

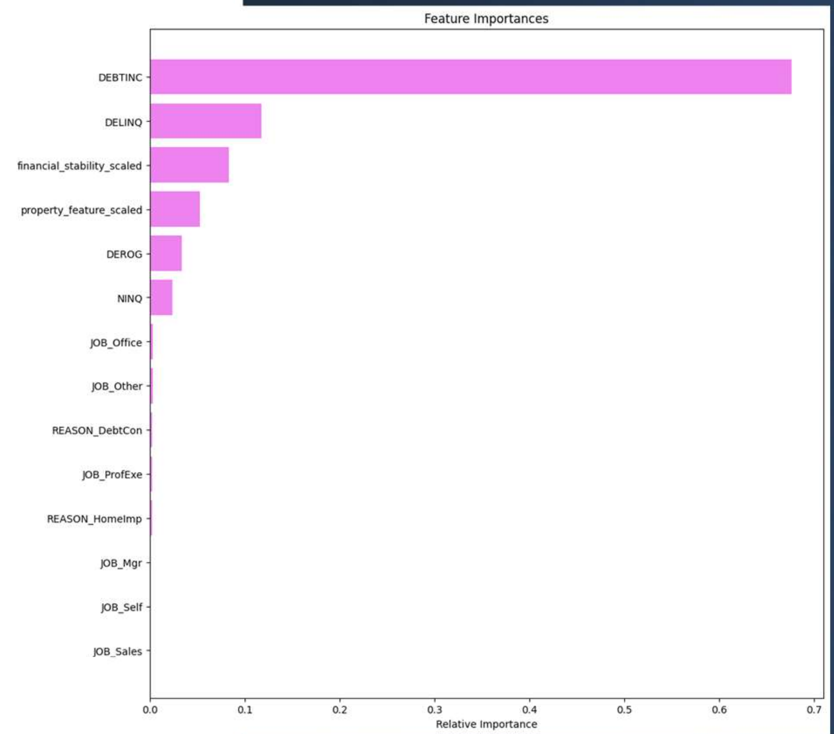
Pre- and Post-Transformation for Skewness and Scale

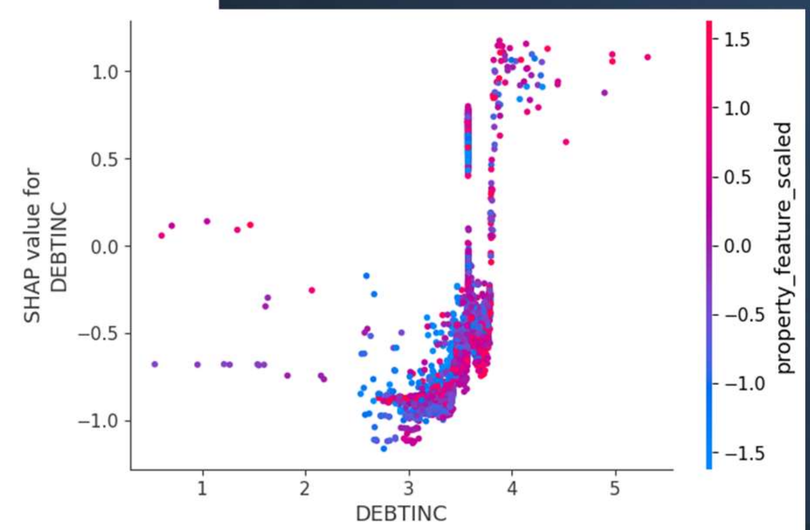




Model Insights with SHAP

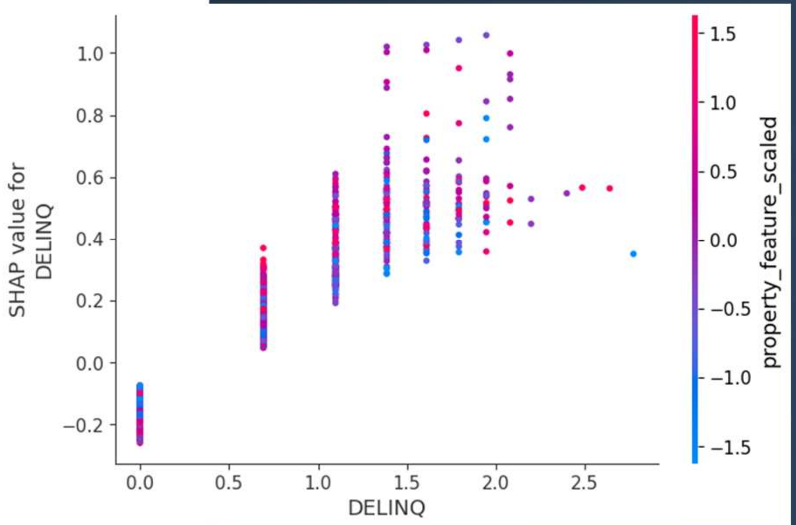
- **DEBTINC (Debt-to-Income ratio)**:
    - **Lower values**: Negatively impact predictions.
    - **Higher values**: Positively affect predictions.
- **Interaction with "property feature scaled"**:
    - Higher "property feature scaled" values amplify DEBTINC's positive influence.
- **Insights**:
    - DEBTINC is a crucial feature.
    - Highlights its interplay with "property feature scaled."





Model Insights with SHAP

- **Delinquencies and Inquiries**:
    - Higher delinquencies (DELINQ) correlate with increased inquiries (NINQ).
    - Indicates possible financial stress or credit-seeking behavior.
- **Low Delinquencies Concentration**:
    - Majority have few delinquencies.
    - Shows varied inquiry levels.
- **Risk Identification**:
    - Use this trend to enhance risk models.
    - Develop targeted financial solutions.



Model Insights with SHAP

- **DELINQ (number of delinquencies)**:

  As DELINQ increases, its SHAP value varies, indicating its significant role in shaping model predictions.

- **Color Gradient (NINQ - number of inquiries)**:

  Transition from blue (lower values) to red (higher values) suggests that higher delinquencies and inquiries correlate with notable changes in predictions.

- **Implications**:

  Offers insights into risk assessment and customer segmentation.