Loan Default Prediction

Zorana Loncar

June 2025

Agenda

1. Problem Overview
2. Solution Approach
3. ML Models in Scope
4. Key Findings & Data Insights
5. Financial Analysis
6. Risks & Challenges
7. Benefits of Implementing the Solution
8. Recommendations & Next Steps

Problem Overview

> *Predictive analytics also play a crucial role in identifying potential defaulters before they miss a payment.*

Forbes

The New Era Of Financial Freedom

Home Loan Default Rate

~20%

Context:

Banks heavily reply on home loan profits, where defaults cause significant financial losses. Regulatory pressure is growing to make credit practices more transparent, explainable and equitable, especially in rejecting or/and approving loan applications.

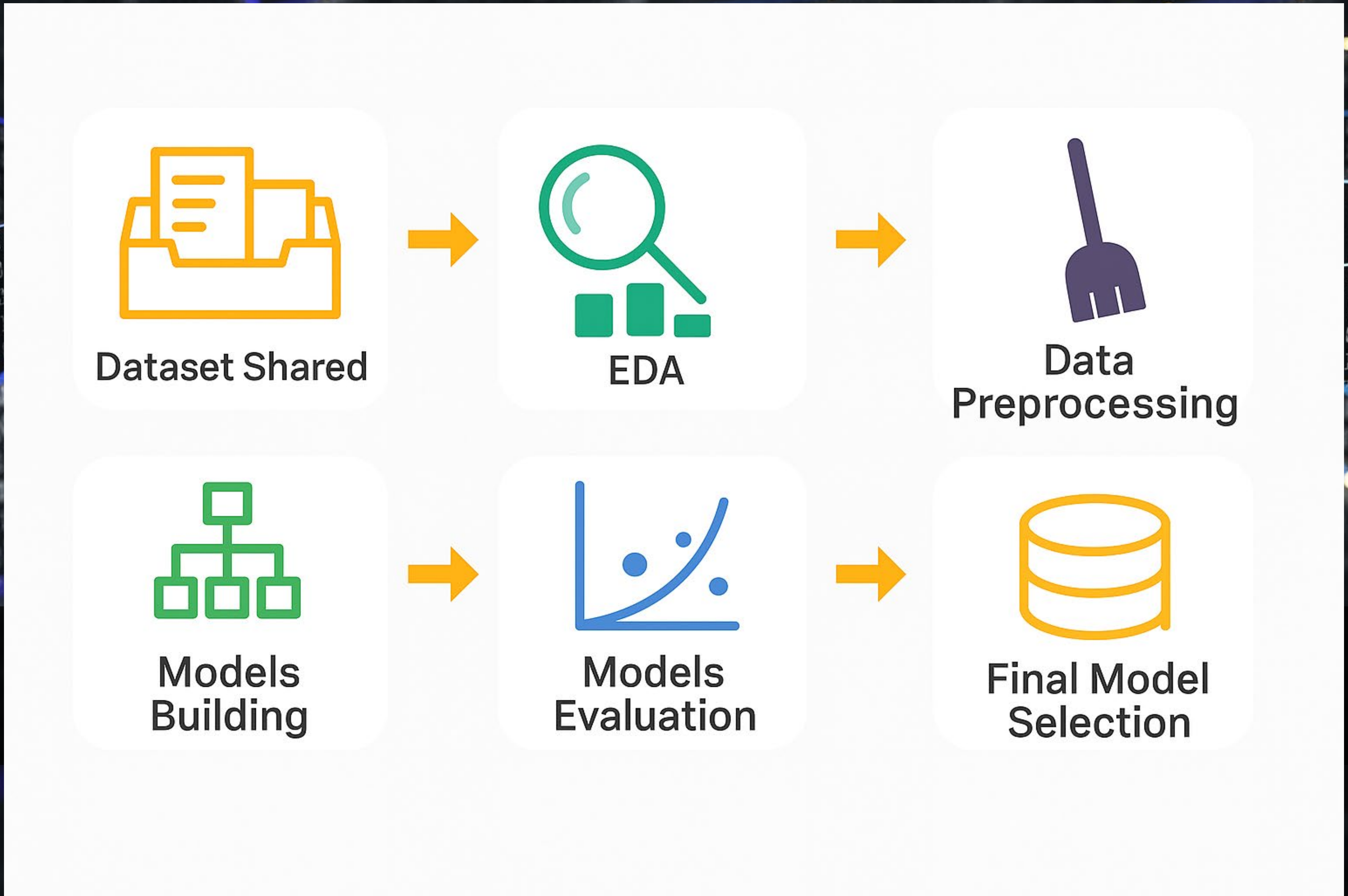
Problem:

Traditional loan approval process is manual, prone to human bias and errors. There is a critical need to enhance loan risk assessment using data-driven approach. The problem has direct implications on profit margins, customer experience, regulatory compliance, risk mitigation strategy.

Objective:

Develop a machine learning model to accurately predict loan defaults, minimizing risk and improving operational efficiency.

Solution Approach



ML Models in Scope

	Logistic Regression	Decision Tree	Random Forest	XGBoost
Model Recall (Defaults)	~57% Defaults correctly identified by the model	~81% Defaults correctly identified by the model	~82% Defaults correctly identified by the model	~82% Defaults correctly identified by the model
Model Precision (Defaults)	~70% Flagged high-risk loans that actually defaulted	~82% Flagged high-risk loans that actually defaulted	~91% Flagged high-risk loans that actually defaulted	~94% Flagged high-risk loans that actually defaulted
Model Accuracy	~81% Correct loan default predictions on test data	~88% Correct loan default predictions on test data	~92% Correct loan default predictions on test data	~92% Correct loan default predictions on test data

Key Findings

Debt-to-Income Ratio (DEBTINC)

1. **DEBTINC** is one of the strongest predictors for loan default
2. High **DEBTINC** increases default risk
3. Applicants with high **DEBTINC** should be flagged for additional evaluation

Delinquencies (DELINQ)

Derogatory Marks (DEROG)

1. **DELINQ** and **DEROG** are strong predictors for loan default
2. **DELINQ > 2** and **DEROG > 1** significantly increase default risk

Loan Amount (LOAN)

1. Loan amount alone is not a strong predictor
2. **LOAN > 40000** combined with high DEBTINC increases default risk

Credit Age (CLAGE)

1. Credit age is protective
2. **CLAGE > 300 months** in associated with low default risk

Years at Present Job (YOJ)

1. Years at present job have an impact on default risk
2. **YOJ < 3** increases default risk

Mortgage Amount (MORTDUE)

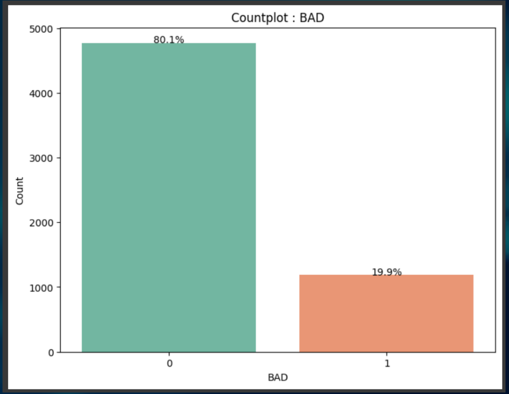
1. **MORTDUE** correlates with loan **LOAN**
2. MORTDUE > 200000 increases default risk

XGBoost Classifier

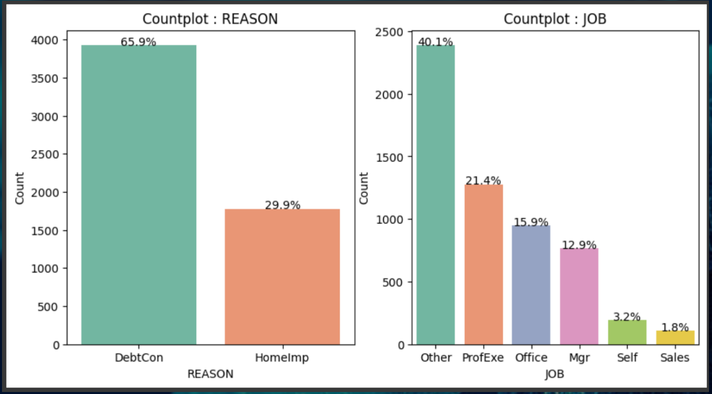
1. **XGBoost** outperforms **Random Forest in predictive power**
2. Recommended for production deployment

Data Insights

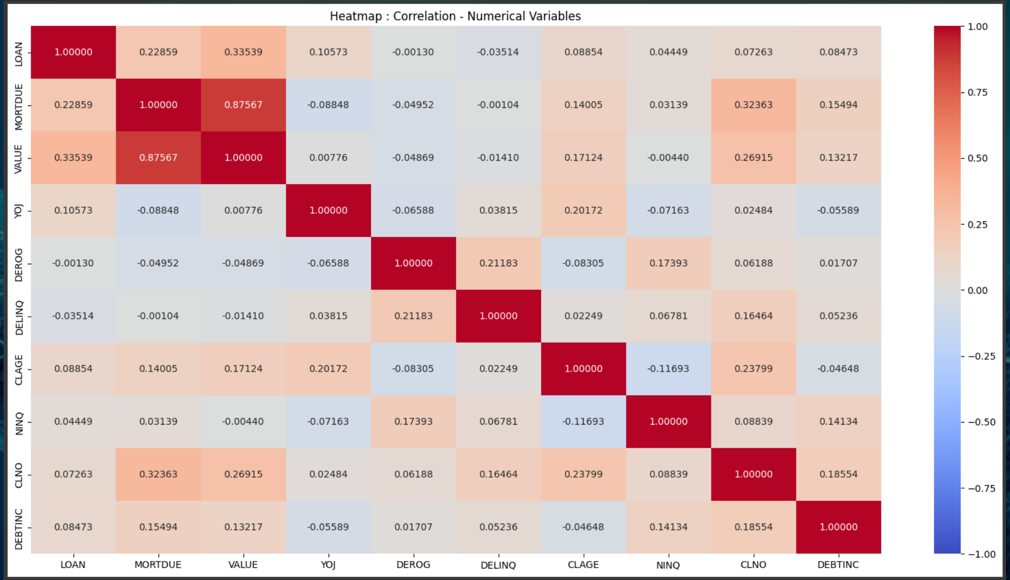
20% Defaulted



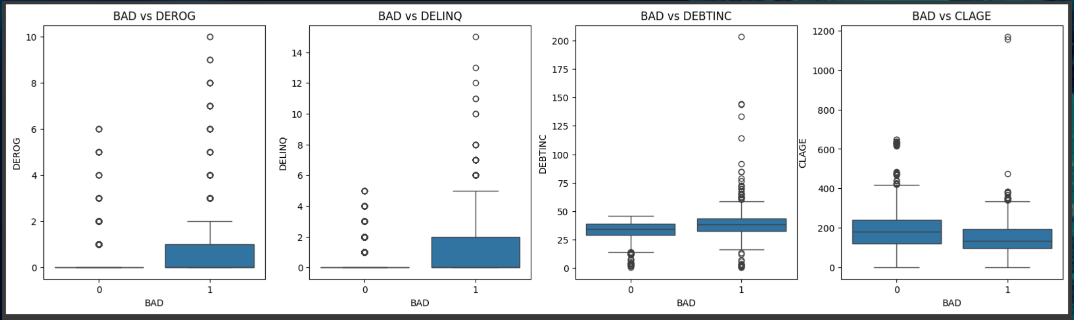
Categorical Variables: Class Imbalance



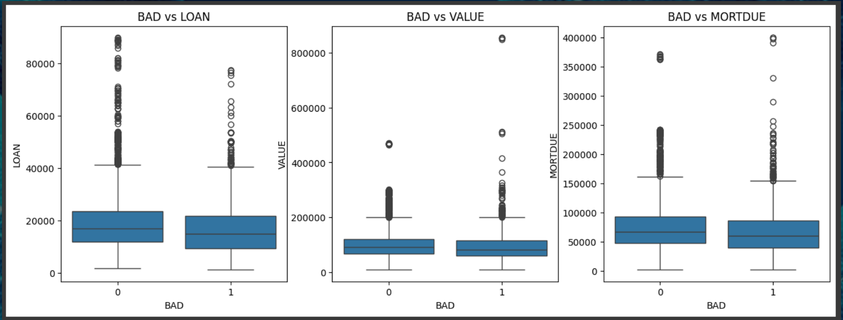
Heatmap: VALUE <> MORTDUE Strong Positive Correlation



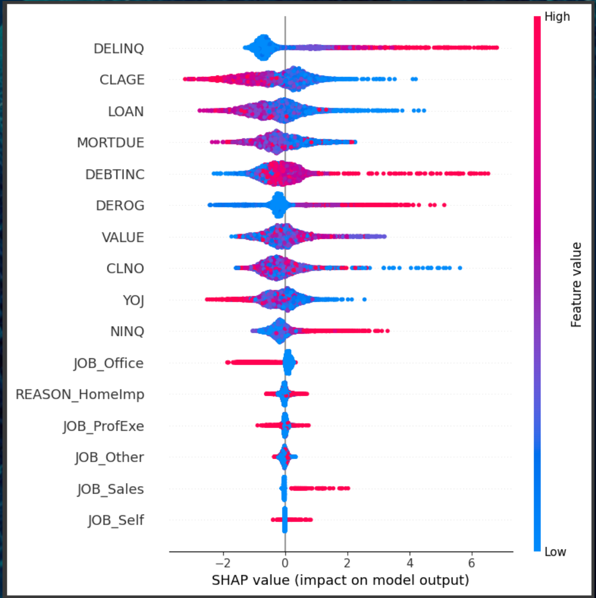
Strong Predictors

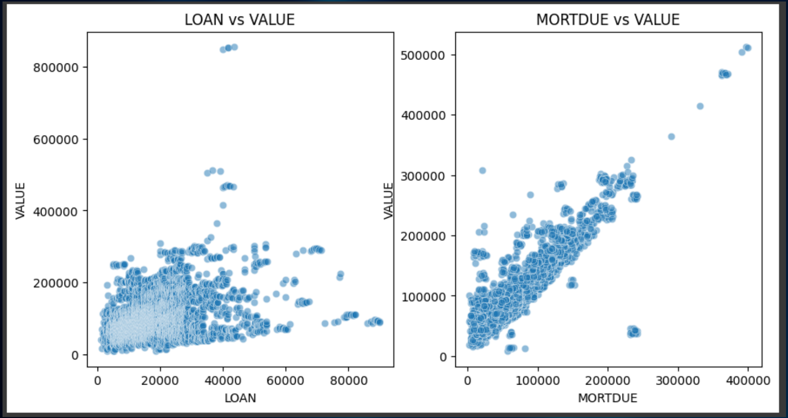


Weak Predictors



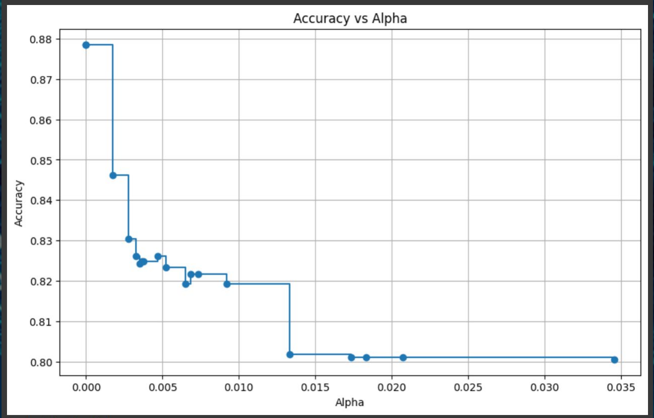
SHAH Summary Plot: DELINQ & DEBTINC Strongest Predictors





VALUE <> MORTDUE
Strong Collinearity

Best alpha = 0.0
No Pruning



Financial Analysis

Model	TN	FP	FN	TP	FN Loss ($)	FP Loss ($)	Total Estimated Loss ($)	Total Projected Gain ($)	Estimated Revenue ($)
Logistic Regression	848	106	130	108	$2,419,036.00	$197,244.00	$2,616,280.00	$3,390,372.00	$1,577,956.00
Decision Tree	593	361	58	180	$1,079,262.00	$671,748.00	$1,751,010.00	$3,781,140.00	$1,103,453.00
Random Forest	935	19	81	157	$1,507,246.00	$35,355.00	$1,542,601.00	$4,625,941.00	$1,739,845.00
XGBoost	949	5	85	153	$1,581,677.00	$9,304.00	$1,590,981.00	$4,603,612.00	$1,765,896.00

Legend

TN = True Negative

FP = False Positive

FN = False Negative

TP = True Positive

FN Loss = Loans that are wrongly approved

FN Loss Rate = 100%

FP Loss = Loans that are wrongly rejected

FP Loss Rate = 10%

ER = Total income earned from correctly approved loans

Calculation

Average Loan Value (AVG) = $18,607.97

FN Loss = FN * AVG * 100%

FP Loss = FP * AVG * 10%

FN Baseline Loss = (FN + TP) * AVG * 100%

Total Estimated Loss (TEL) = FN Loss + FP Loss

Total Projected Gain (TPG) = FN Baseline Loss – TEL + ER

Estimated Revenue (ER) = TN * AVG * 10%

Risks & Challenges

Trust and Adoption Resistance

Slower adoption rate due to trust in automated decisions over traditional judgment, by bank officers.

Market Sensitivity

Economic changes, like unemployment or change in interest rates, can shift applicant behaviour, making previously accurate models outdated.

Overdependence on Historical Data

If pact credit practices were biased or inconsistent, model may unknowingly replicate those patterns.

Customer Experience Impact

Poorly explained rejection or perceived unfairness from algorithmic decision may lead to dissatisfaction and churn.

Benefits of Implementing the Solution

Reduced Financial Losses

Early detection of high-risk applicants minimizes defaults.

Operational Efficiency

Automates large parts of the loan approval process, reducing costs.

Scalability

Adaptable to new markets, customer segments and economic conditions.

Regulatory Compliance

Data-driven and auditable decisions help meet transparency standards.

Increased Approval Accuracy

Minimizes both **False Negatives** and **False Positives**

Improved Customer Experience

Faster processing of loan applications and more objective evaluation.

Recommendations & Next Seps

Recommended Model: **XGBoost**

ML Model Training & Deployment

Train ▶ Validate ▶ Deploy ▶ Monitor

Integrate ML model into existing loan approval process to flag high-risk applicants.

Continuous model monitoring adapt to changing applicant's behavior.

Incorporate additional data sources, for example transaction data, social media signals, to enhance predictive accuracy.

Human-in-the-Loop System: Allow manual review for borderline predictions.

Thank you!