LOAN DEFAULT PREDICTION USING MACHINE LEARNING

ADONIS BOYD

MIT APPLIED DATA SCIENCE PROGRAM

2025

EXECUTIVE SUMMARY

- Banks face losses from loan defaults.
- Goal: Predict loan default risk before approval.
- Business Impact: Reduce losses, automate credit screening, ensure fair lending.

PROBLEM & SOLUTION SUMMARY

Data Science Objectives

- Predict if a customer will default (BAD = 1).

- Identify key drivers of default.

- Compare different models.

- Recommend a model ready for deployment.

BUSINESS GOALS

Improve	Reduce	Prevent
Improve loan underwriting	Reduce manual workload.	Prevent losses before they happen

DATA OVERVIEW

Dataset: 5,960 rows, 13 columns.

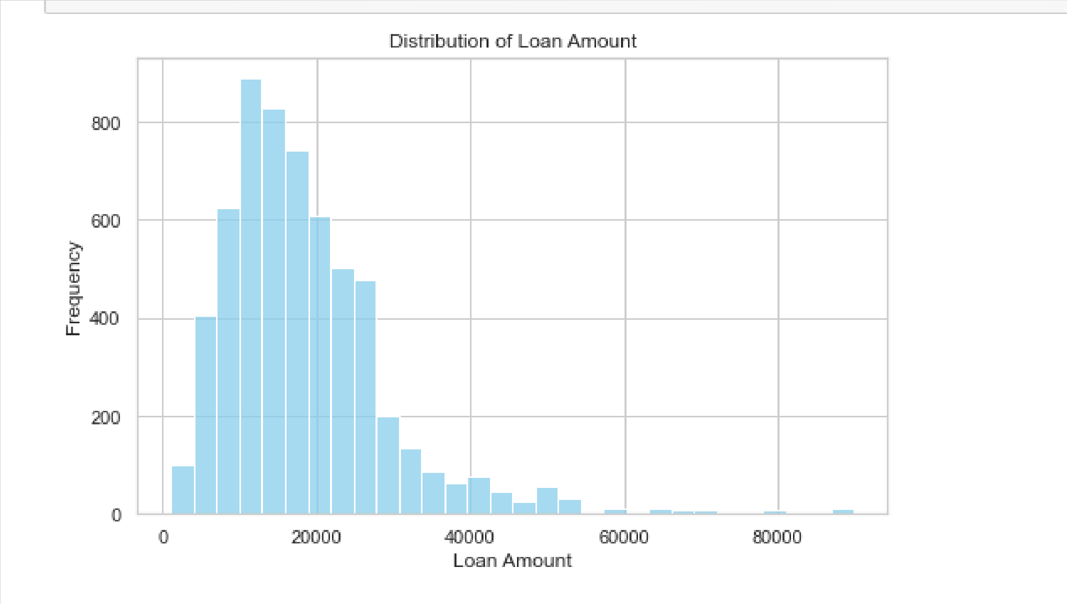
Target: BAD (1 = default, 0 = repaid).

Features: Loan amount, credit history, employment, etc.

Challenge: Slight Class Imbalance + missing values in key columns like (YOJ), DEBTINC.

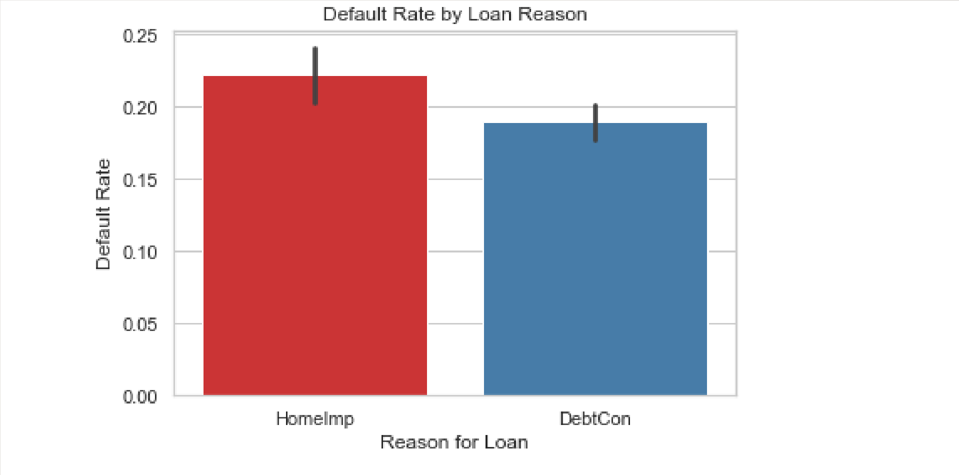
DATA CLEANING

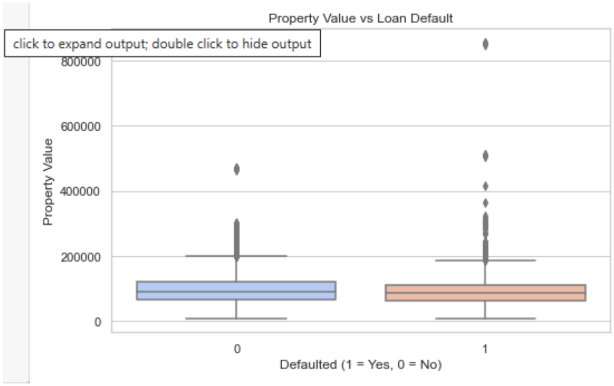
- • Missing values handled using median (numerical) and mode (categorical).
- • Outliers capped using IQR method.
- • Categorical variables one-hot encoded.



- Right-skewed, with most applicants requesting between 10,000 and 25,000.
- Very few applicants request amounts above 50,000 which may be considered outliers.

- Observation: People who are seeking Home Improvement typically have high default rates than those seeking Debt Consolidation.
- Defaulters tend to own properties with lower market value.

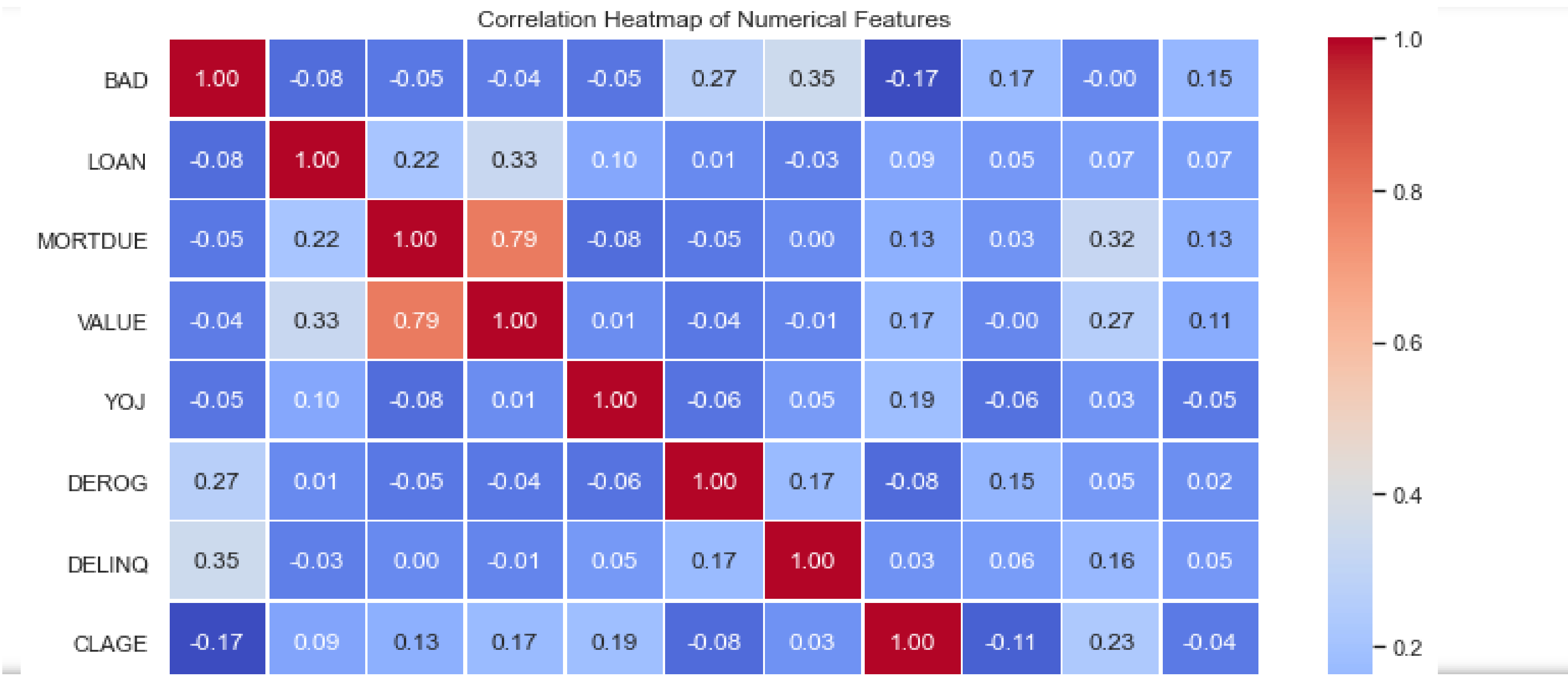




- Defaulters tend to own properties with low market value compared to non-defaulters.

EXPLORATORY DATA ANALYSIS

- Higher default rates in Home Improvement loans.
- Credit behavior (delinquencies, derogatory marks) linked to default.
- Long job tenure and credit history = lower risk.


Correlation Heatmap of Numerical Features
BAD 1.00 -0.08 -0.05 -0.04 -0.05 0.27 0.35 -0.17 0.17 -0.00 0.15
LOAN -0.08 1.00 0.22 0.33 0.10 0.01 -0.03 0.09 0.05 0.07 0.07
MORTDUE -0.05 0.22 1.00 0.79 -0.08 -0.05 0.00 0.13 0.03 0.32 0.13
VALUE -0.04 0.33 0.79 1.00 0.01 -0.04 -0.01 0.17 -0.00 0.27 0.11
YOJ -0.05 0.10 -0.08 0.01 1.00 -0.06 0.05 0.19 -0.06 0.03 -0.05
DEROG 0.27 0.01 -0.05 -0.04 -0.06 1.00 0.17 -0.08 0.15 0.05 0.02
DELINQ 0.35 -0.03 0.00 -0.01 0.05 0.17 1.00 0.03 0.06 0.16 0.05
CLAGE -0.17 0.09 0.13 0.17 0.19 -0.08 0.03 1.00 -0.11 0.23 -0.04
1.0
0.8
0.6
0.4
0.2

CORRELATION HEATMAP INSIGHTS

- • Strongest predictors of default:
- - DELINQ, DEROG, NINQ, DEBTINC VALUE
- • Loan size and home value have weak correlation with default.
- BAD is mostly strong correlated with credit behavior features (DELINQ, DEROG, NINQ, DEBTINC while LOAN & VALUE have weak correlations.
- The map suggest that credit behavior matters more than the size of a loan when predicting risk.

MODEL COMPARISON

- Logistic Regression: ~85% Accuracy, ROC-AUC ~0.77
- Decision Tree: ~86% Accuracy, ROC-AUC ~0.76
- Random Forest: 90% Accuracy, ROC-AUC 0.96 (Best Model)

COMPARISON

DECISION TREE

- Captures non-linear relationships
- Slighty prone to overfitting
- Recall improves (~61%), but not the best overall
- ROC-AUC: ~0.76
- Great for interpretability, not the strongest performer

LOGISTIC REGRESSION

- Simple and interpretable
- Struggles with complex patterns in the data
- Lower recall (~50%) for defaulters
- ROC-AUC: ~0.77 → baseline model

RANDOM FOREST

Delivered highest overall performance with 90% accuracy and 0.96 ROC-AUC.

```
Confusion Matrix:
 [[923  31]
 [ 90 148]]

Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.97      0.94       954
           1       0.83      0.62      0.71       238

    accuracy                           0.90      1192
   macro avg       0.87      0.79      0.82      1192
weighted avg       0.89      0.90      0.89      1192

ROC-AUC Score: 0.9582496520620827
```

FINAL MODEL – RANDOM FOREST

- Tuned using cross-validation (GridSearchCV)

Optimized for best accuracy and recall

Chosen for its robust performance and reliability

Final ROC-AUC: 0.96 – strongest of all models

Tuned with 200 trees, full depth and fine-grained leaf splits

FINAL MODEL

- High accuracy and ROC-AUC
- Balanced precision and recall
- Robustness to overfitting
- Interpretability through feature importances
- It aligns well with the bank's goal of minimizing risk while maintaining fair and efficient loan decisions.

```
Confusion Matrix:
 [[919  35]
 [ 90 148]]

Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.96      0.94       954
           1       0.81      0.62      0.70       238

    accuracy                           0.90      1192
   macro avg       0.86      0.79      0.82      1192
weighted avg       0.89      0.90      0.89      1192

ROC-AUC Score: 0.9602866303754207
```

KEY BUSINESS INSIGHTS

- Behavioral credit history is more predictive than loan amount.
- Home Improvement loans carry higher risk.
- Random Forest can support smarter, fairer loan approvals.

RECOMMENDATIONS FOR IMPLEMENTATION

- Deploy the Random Forest model in the loan approval system
- Use it to flag high-risk applicants early.
- Regularly monitor performance and retrain as needed
- This model enables proactive loan risk management, reduces manual review, and supports scalable underwriting.


THANK YOU