Loan Risk Assessment – Final Report

Author: *Aakriti Agrawal*
Date:

Executive Summary

This project builds a Machine Learning model to predict loan approval outcomes using a synthetic but realistic loan applicant dataset. After extensive preprocessing, exploratory analysis, model development, tuning, and fairness checks, the final model selected was **XGBoost**, which demonstrated strong performance on unseen test data.

On the test set, the model achieved:

- **AUC:** 0.978
- **Accuracy:** *0.93*
- **Precision:** 0.974
- **Recall:** *0.68*
- **Threshold used:** *0.770*

Using this model with an optimized threshold improves risk discrimination and can significantly reduce expected credit losses while maintaining acceptable approval volumes.
Explainability (via SHAP values) confirms the business logic: **credit score, loan percent income, and payment-to-income ratio** are the strongest predictors.

---

1. Introduction

Lending institutions must balance risk and growth. This project builds a model to determine the likelihood of loan approval using applicant demographics, financial data, credit history, and loan characteristics.

Goals:

- Predict loan approval
- Understand drivers of approval
- Segment risk levels
- Test fairness
- Recommend policy improvements

---

2. Dataset Overview

- Total rows: **[19,999]**
- Total features: **14 (raw)** → **20+ (after feature engineering)**
- Target variable: **loan_status**
  - 1 = Approved
  - 0 = Denied

Key Columns:

- `person_income` (numeric)
- `loan_amnt`
- `loan_percent_income`
- `credit_score`
- `person_home_ownership`, `loan_intent`, etc.

Feature Engineering:

- `monthly_income`
- `monthly_payment_est`
- `payment_to_income`
- `credit_bucket`
- `age_bucket`

---

3. Exploratory Data Analysis

Key Findings:

- Higher **credit scores** correlate with higher approval probability.
- High **loan_percent_income** → strong decline in approval.
- Applicants with **previous loan defaults** have significantly lower approval rates.
- Certain loan intents (e.g., personal loans) have higher risk.

(Insert plots: income distribution, loan amount distribution, approval rate by intent, etc.)

---

4. Modeling Approach

Models Attempted:

- Logistic Regression
- Random Forest
- XGBoost (final winner)

XGBoost had the highest validation and test performance.

Final Test Performance

Fill these after running your evaluation:

- **AUC:** *[Your Value]*
- **Precision:** *[Your Value]*
- **Recall:** *[Your Value]*
- **F1-score:** *[Your Value]*
- **Confusion Matrix:** *(insert heatmap image)*

---

5. Explainability (SHAP)

SHAP analysis identifies the most influential variables:

Top Features:

1. credit_score
2. loan_percent_income
3. payment_to_income
4. loan_amnt
5. person_income

Include these visuals:

- SHAP summary bar plot
- SHAP global beeswarm plot
- Two individual SHAP waterfall plots

Key Interpretation:

- Higher **credit score** → strong positive approval influence
- Higher **loan_percent_income** → strong negative influence
- Higher **income** increases approval probability
- Previous defaults significantly reduce approval likelihood

---

6. Risk Segmentation

Based on XGBoost predicted probabilities:

Segment	Probability Range	Description
High Risk	0% – 35%	Very low likelihood of repayment
Medium Risk	35% – 70%	Requires manual review
Low Risk	70% – 100%	Safe approvals

Insert this table:

Segment	Count	Avg Credit Score	Avg Income	Approval Rate
Low Risk	X	X	X	X
Medium Risk	X	X	X	X
High Risk	X	X	X	X

(These values come directly from Day 7 code.)

---

7. Policy Simulation

Example:

- Increasing approval threshold from **0.50** → **0.65**

    - Reduces approvals by **XX%**
    - Reduces defaults by **YY%**
    - Increases expected portfolio profit by **ZZ%**

(Insert your results from simulation)

---

8. Fairness & Robustness

Fairness checks:

- Gender approval rates
- FPR and FNR by subgroup
- Subgroup AUC

Findings:

- No extreme disparity between gender or education groups.
- Model is reasonably calibrated.

Robustness:

- Bootstrap AUC showed stable performance (± small variance)
- Permutation importance validated SHAP top features

---

9. Conclusion & Recommendations

- Deploy XGBoost model with threshold = **[your selected threshold]**

- Use Medium Risk segment for **manual underwriting review**
- Monitor fairness metrics periodically
- Collect more data for under-represented groups
- Improve credit score features for higher accuracy