Image Source: Unsplash (Allison Saeng)

Loan Default Predictions

05.16.2025

Andrea Pena
Capstone 3: Project Report
Springboard

Introduction

Financial institutions are continually seeking ways to make the loan process quicker, more efficient, and thoughtful. Accurate prediction of loan defaults helps lenders minimize financial risk and better allocate lending resources.

This project supports smarter lending decisions by predicting loan defaults using supervised classification techniques. After testing multiple models and tuning thresholds, the final recommendation is a **XGBoost model optimized for Recall**, aligning with the business objective of reducing financial loss through early identification of high-risk borrowers. Insights from this analysis will help financial institutions develop targeted strategies to reduce default risk and improve borrower retention.

Data Overview

The dataset used in this project is the **Loan Default Dataset** by M Yasser H, available on Kaggle.

It consists of **148,670 loan applications** and **39 features**, including a mix of **numerical and categorical variables** such as loan amount, loan purpose, income, and creditworthiness. The target variable is **loan status** — whether the applicant defaulted or not.

For data cleaning, missing values in numerical features were imputed with the **median**, and categorical features were imputed with the **mode**. This ensured a complete dataset for training and evaluation.

Methodology

My modeling approach began with exploratory feature engineering. I created categorized versions of continuous variables such as loan-to-value (LTV) and debt-to-income ratio (DTI), and standardized labels for categorical features like **loan_purpose.** To better understand the data, I used Tableau to develop an interactive story through a series of dashboards, which informed key decisions in preprocessing and modeling.

Tableau Visualizations and Key Insights:

- **Loan Purpose, Gender, and Region:**
  Refinancing and investment property loans were the most common application types. Applications were disproportionately submitted by male and joint applicants, with a large portion of unknown genders, indicating a class imbalance. The North and South regions had the highest volume of applications.

- **Loan Type, Credit Score, and Loan Amount:**
  Conventional loans dominated, making up approximately 76% of applications. Credit scores were relatively stable across loan types, suggesting that credit score alone may not strongly differentiate default risk. Most loan amounts fell between $150,000 and $250,000.

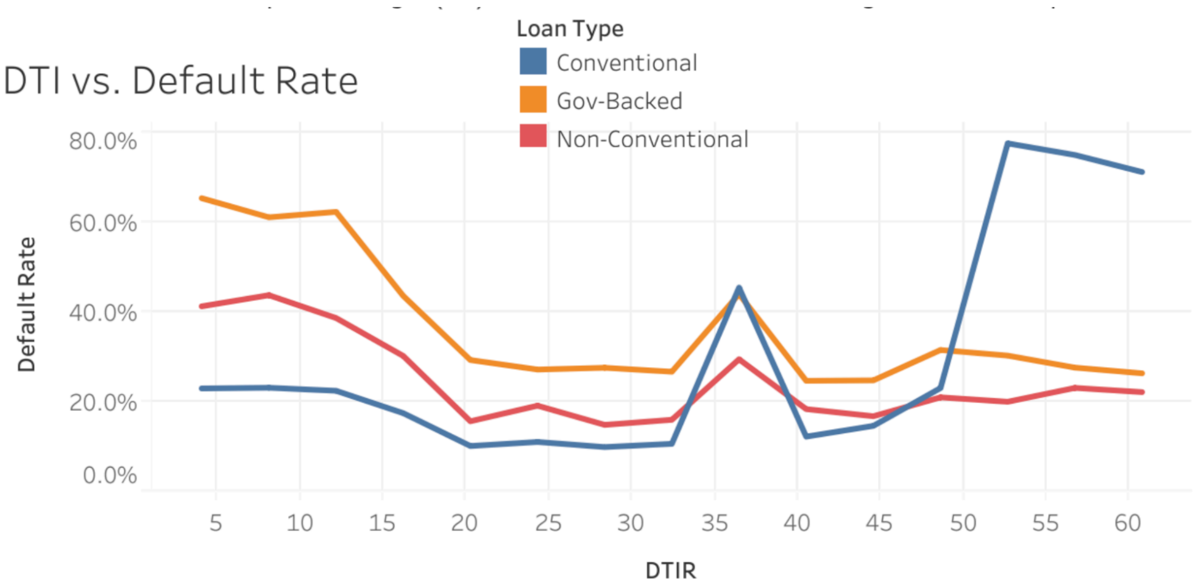
- **Risk Indicators and Defaults:**
  Home improvement and home purchase loans had the highest default rates, followed by refinancing. Regionally, default rates were elevated in the North-East and Central areas. Applicants under age 25 and those 55 and older showed higher default rates, highlighting risk at both age extremes. No clear relationship was observed between LTV and credit score with respect to default, suggesting these variables may not interact linearly.

- **Approval Patterns and Categorical Breakdowns:**
  Approval rates were consistent across credit scores, with the North region exhibiting the highest approval rates overall. Conventional loans had higher approval success. The credit types CIB and CRIF were the most frequently used. Notably, default rates were over 50% higher for applicants with negative amortization.

- **Financial Indicators and Outcomes:**
  Approval rates remained stable across loan amounts, with a slight dip around $2,000 in upfront charges, potentially indicating this feature is less predictive. A negative trend was observed between interest rate spread and

approval rate. Rising DTI ratios corresponded with increased default rates, particularly above a DTI threshold of 50 for conventional loans.

Figure 2: DTI vs. Default Rate



Correlation Analysis:

A Pearson correlation heatmap showed no strong linear correlations between the features and the target variable. This was expected, as many features are binary or categorical, and Pearson correlation captures only linear relationships. This finding suggested the need for non-linear modeling techniques.

Feature Selection – F-Score Analysis:

To identify the most statistically significant features, I conducted an F-score test. The highest-scoring variables included:

- **upfront_charges, dtir, property_value, income, interest_rate_spread and rate_of_interest**

These variables demonstrated significant variance in mean values between defaulted and non-defaulted cases.

Preprocessing:

Categorical features were one-hot encoded to prepare them for modeling. I also removed several **leaky features**—variables that either explicitly contained or indirectly leaked outcome information:

- **default_status, ltv_category, and high_dtir_flag** (engineered for visualization only)

- i**nterest_rate_spread, upfront_charges, and rate_of_interest** (typically available only after loan approval)

Removing these features ensured the integrity of the modeling pipeline by preventing the model from learning from information that would not be available at decision time.

Model Performance & Comparison

Multiple classification models were tested, but our evaluation prioritized **Recall** to align with the business objective: reducing loan defaults by identifying high-risk applicants. While other metrics such as AUC and Precision were tracked for transparency, model selection was based primarily on recall performance and generalization.

Logistic Regression and Random Forest models served as baselines. Threshold tuning and hyperparameter optimization were explored to improve recall.

Class Imbalance Strategy

The dataset was highly imbalanced:

- **Not Defaulted:** 112,031 (~75%)
- **Defaulted:** 36,639 (~25%)

To address this:

- **SMOTE** was applied to synthetically resample the minority class.
- **class_weight='balanced'** was used in some models to penalize misclassification of the minority class

Model Tuning Process

- **GridSearchCV** was initially used for hyperparameter tuning.
- Due to time constraints, **RandomizedSearchCV** was implemented to efficiently explore parameter combinations.
- Final optimization was performed using **Bayesian Search**, which uses past results to predict the most promising combinations, improving search efficiency.

Results suggested that **conservative parameters** often performed better.

Feature Importance

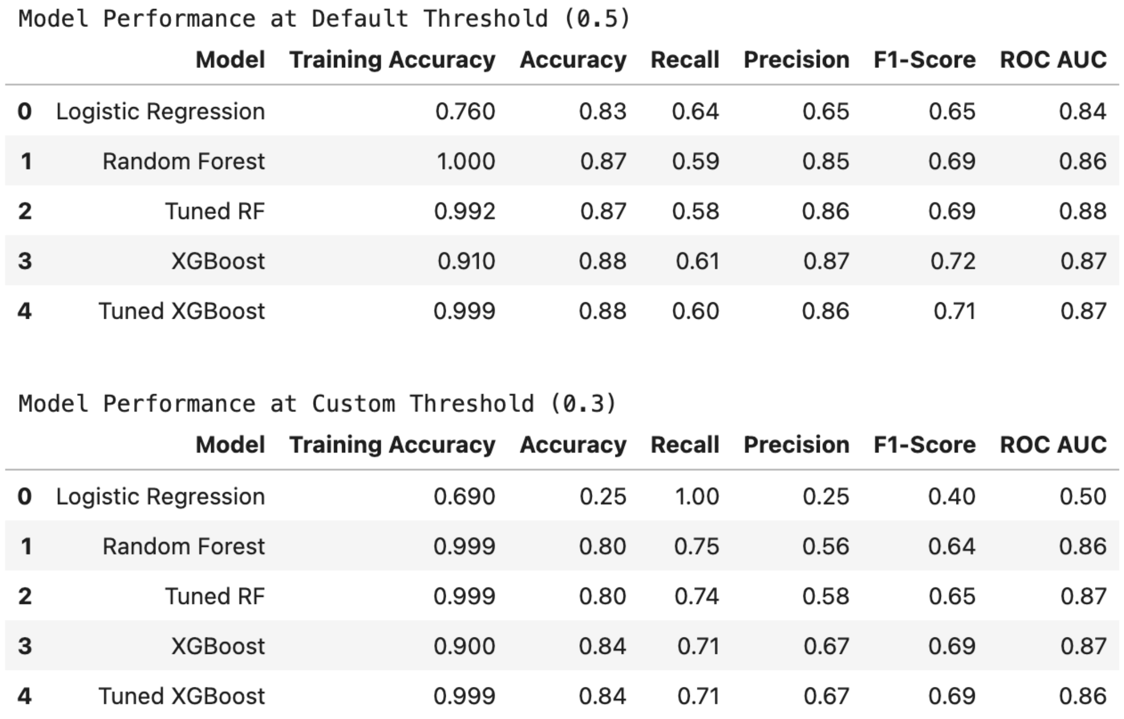
Top features across models included:

- **Loan-to-Value (LTV)**
- **Debt-to-Income Ratio (DTI)**
- **Property Value**

These variables consistently ranked highest and carry meaningful financial interpretation.

While **Credit Type: Equifax** emerged as dominant in some models, it was tested for data leakage and retained due to consistent performance and integrity across validation splits.

Figure 2: Model Comparison Table



Model Performance at Default Threshold (0.5)

	Model	Training Accuracy	Accuracy	Recall	Precision	F1-Score	ROC AUC
0	Logistic Regression	0.760	0.83	0.64	0.65	0.65	0.84
1	Random Forest	1.000	0.87	0.59	0.85	0.69	0.86
2	Tuned RF	0.992	0.87	0.58	0.86	0.69	0.88
3	XGBoost	0.910	0.88	0.61	0.87	0.72	0.87
4	Tuned XGBoost	0.999	0.88	0.60	0.86	0.71	0.87

Model Performance at Custom Threshold (0.3)

	Model	Training Accuracy	Accuracy	Recall	Precision	F1-Score	ROC AUC
0	Logistic Regression	0.690	0.25	1.00	0.25	0.40	0.50
1	Random Forest	0.999	0.80	0.75	0.56	0.64	0.86
2	Tuned RF	0.999	0.80	0.74	0.58	0.65	0.87
3	XGBoost	0.900	0.84	0.71	0.67	0.69	0.87
4	Tuned XGBoost	0.999	0.84	0.71	0.67	0.69	0.86

Table 1: Performance of models at default threshold (0.5). Recall was lower across all models at this setting, motivating further tuning.

Table 2: Model performance after custom threshold adjustment. Tuned XGBoost showed highest recall with strong generalization, supporting final selection.

Final Model Selection: XGBoost with a 0.3 threshold, delivered the strongest performance in identifying likely defaulters.

While Random Forest showed slightly higher recall, the **XGBoost model** achieved stronger generalization (train accuracy = 90%), competitive recall (**71%**), higher precision (**67%**), and the best overall **ROC AUC (87%)**.

This combination provides the strongest foundation for minimizing defaults and scaling model usage in production.

Error Analysis of Final Model

To better understand where the model struggled, we analyzed false negatives (missed defaults) and false positives (false alarms) using the final XGBoost model with a custom threshold of 0.3:

- **False Negatives:** 2,111
- **False Positives:** 2,559

False negatives—borrowers who defaulted but were predicted safe—tended to have **higher income** and **slightly better credit scores**, which may have signaled financial stability to the model. They also had **lower LTV** and **DTI ratios**, further misleading the model into predicting "no default."

False positives, on the other hand, showed **greater variation in loan amounts**, indicated by higher standard deviation, suggesting the model struggled with consistency in this group. These patterns highlight potential opportunities for **additional feature engineering or threshold tuning** to reduce misclassifications.

Conclusion

The XGBoost model with a 0.3 threshold provided the best trade-off for our use case. With a **Recall of 71%** and **ROC AUC of 87%**, it offers a strong foundation for reducing financial loss due to loan defaults. Compared to earlier models, it improved generalization and reduced false negatives significantly.

While minor input noise (e.g., high LTV ratios) remains, the model's performance is resilient. Future work will focus on enhancing **Precision** by capping extreme values, refining features, and adjusting thresholds for better customer retention.

Overall, this solution supports the institution's risk mitigation strategy and provides a confident path toward smarter, fairer lending

Recommendations

Deploy Risk Scoring with Explainability Tools
Integrate the model into a loan officer dashboard with SHAP explanations to display not only the predicted risk score, but also the key drivers behind each prediction. This would support **loan officers** and **customer service reps** by making decisions more transparent and helping them justify approvals or rejections with clear reasoning.

Enhance Geographic Risk Monitoring
Use model predictions in conjunction with Tableau to **map default risk by region**. This would allow **risk managers** to track patterns in high-default areas, prompting localized interventions or adjusted lending policies. Custom filters could also flag shifts in risk over time.

Refine Model with Focused Feature Engineering
Several important features (e.g., LTV, DTI, income) showed outliers and inconsistencies. Conduct targeted cleaning (e.g., capping, binning, transformation) to improve stability and reduce misclassification. This effort, led by **data scientists**, would further boost performance and offer stakeholders clearer insights into **why borrowers default**.

For future work, I recommend experimenting with **ensemble stacking** to combine the strengths of multiple models, investigating **model calibration** to improve the reliability of predicted probabilities, and **collecting more real-world data** to reduce dependency on synthetic oversampling methods like SMOTE

Data References

Original Dataset: The loan default dataset was provided as part of the instructional project materials and included borrower attributes, loan terms, and default outcomes.

Feature Mapping Guidance: Mapping logic for variables such as **loan_purpose** and **loan_type** was adapted from a Kaggle notebook by another participant who performed similar feature engineering using R. While their work was in a different language, the categorical groupings and label interpretations were informative in shaping the logic implemented in Python. Proper attribution and reimplementation were ensured.

Binning Strategy: Industry guidelines from **Fannie Mae** and general lending practices were referenced to guide the binning of financial variables like **Loan-to-Value (LTV)** and **Debt-to-Income Ratio (DTI)**. Thresholds were set to align with commonly accepted definitions for “moderate” and “high” risk brackets to ensure model interpretability and business relevance.