Loan Classification Using Logistic Regression and SMOTE

S.Vardhan

July 20, 2025

Abstract

This project presents a binary classification approach to identify good and bad loans using machine learning techniques. The business goal emphasizes reducing false negatives—cases where a bad loan is predicted as good—due to the financial implications of loan defaults. A Logistic Regression model with SMOTE (Synthetic Minority Oversampling Technique) was chosen after evaluating multiple classifiers, prioritizing interpretability, generalization, and class imbalance handling.

1 Introduction

In the financial domain, accurately predicting loan defaults is critical. A bad loan wrongly classified as good (false negative) can result in significant financial loss for the lender. Hence, this project focuses on reducing false negatives in a supervised binary classification problem using historical loan data.

2 Problem Statement

Given a dataset of historical loans, the goal is to classify whether a loan is "good" or "bad". Good loans are defined as those with a `loan_status` of `Fully Paid` or `Current`, and bad loans as those with a `loan_status` of `Charged Off`. The primary metric of concern is the minimization of false negatives to avoid approving loans likely to default.

3 Data Preprocessing

- Merged "Fully Paid" and "Current" into one label: `0 (Good Loan)`.
- Labeled "Charged Off" as: `1 (Bad Loan)`.
- Performed null-value treatment and selected relevant numerical and categorical features.

- One-hot encoding was used for categorical variables.
- Features were scaled using `StandardScaler`.

4 Class Imbalance Handling

The dataset showed significant class imbalance. To counter this, we applied **SMOTE** (Synthetic Minority Over-sampling Technique) to synthetically generate examples from the minority class (bad loans). This allows the classifier to better learn decision boundaries and reduce bias toward the majority class.

5 Model: Logistic Regression with SMOTE

- Used Logistic Regression with balanced classes and L2 regularization.
- Applied SMOTE to the training data only.
- Evaluated model on untouched test data.

Classification Report

```
              precision    recall  f1-score   support

           0       0.95      0.90      0.92      6656
           1       0.90      0.95      0.92      6642

    accuracy                           0.92     13298
   macro avg       0.92      0.92      0.92     13298
weighted avg       0.92      0.92      0.92     13298
```

Confusion Matrix

$$\begin{bmatrix} 5961 & 695 \\ 341 & 6301 \end{bmatrix}$$

False Negatives (Bad loans predicted as good): 341 — This is significantly low, aligning with our objective.

6 Discussion

Why Minimizing False Negatives Matters:

- False negatives correspond to high-risk loans incorrectly predicted as low-risk.
- Lending to such applicants may lead to charge-offs or financial losses.
- This metric is prioritized even at the cost of increasing false positives.

7 Comparison with Other Models

While several classifiers were tested, including XGBoost, Random Forest, and Decision Trees, Logistic Regression was selected due to:

- **Interpretability**: Clear probabilistic outputs and decision boundaries.
- **SMOTE Compatibility**: Logistic Regression works well with oversampled datasets.
- **Performance**: Comparable F1-scores and recall with fewer false negatives than more complex models.
- **Generalization**: Simpler models like Logistic Regression are less prone to overfitting and easier to monitor in production.

8 Conclusion

This project demonstrates a robust and interpretable loan classification pipeline that prioritizes reduction of false negatives. Logistic Regression with SMOTE offers a strong balance of recall, precision, and interpretability, making it suitable for deployment in high-stakes financial environments.