Credit Risk Classification Analysis

By Carlos Ruiz

Overview of the Analysis

The purpose of this analysis is to evaluate the effectiveness of a linear regression machine learning (ML) model when predicting healthy vs high-risk loans from financial data.

Credit Lenders would benefit from this type of analysis, as they like to manage their loan portfolios more effectively by being able to predict the likelihood of defaulting loans.

Dataset

Lending_data.csv dataset is characterized by several financial features that describe each loan, including information about the size of the loan, interest rate, as well as income, assets and history of the borrower. The dataset offered credit risk as a target variable where healthy loans were indicated as 0, while high-risk loans were indicated as 1.

Insights into the Data:

Credit Risk data was imbalanced. Only 3.33% of the loans were healthy loans, while 96.6% of the loans were high-risk loans. This would have a significant impact on the performance of the model.

Basic Distribution of credit risk (value_counts):

0: Healthy Loans = 75,036 instances

1: High-risk Loans = 2,500 instances

Stages of Machine Learning Process

1. Data Preprocessing: After loading the data, no imputation or encoding were needed; however, although, scalar was not performed, it is highly recommended. After performing a correlation analysis, it became visible that multi-collinearity is a factor. Regular linear regression might not be optimal.

2. Train-Test Split: The dataset was split into training and testing sets to evaluate the model performance.
3. Modeling: Applied a logistic regression model with the original data.
4. Model Evaluation: The model predicts high-risk loans quite well, with 99% accuracy, 94% recall and 89% f1-score.

Results

Machine Learning Model 1: Logistic Regression

Accuracy: 99%

Precision (Healthy Loans - 0): 100% - no healthy loans were misclassified as high risk

Recall (Healthy Loans - 0): 100% - all healthy loans were classified correctly

F1-Score Healthy Loans - 0: 99%

Precision High-Risk Loans - 1: 87% - 13% of the high-risk loans were healthy loans

Recall High-Risk Loans - 1: 95% - properly identified 95% of the high-risk loans

F1-Score High-Risk Loans - 1: 91%

Summary

As seen from these evaluation metrics, Logistic Regression does quite an excellent job in predicting healthy loans with a very high precision and recall for the healthy loans. There is, however, slightly reduced performance in the precision regarding the prediction of the high-risk loans, meaning there are instances of healthy loans being mislabeled as high risk.

Recommendation:

This model is appropriate if the objective function would be to minimize the misclassification of healthy loans; however, it does need a scaler.

Further model tuning or the use of models such as Random Forest or SVM will give better precision and recall for the high-risk class.