Credit Risk Analysis Using Machine Learning

Name: Pranitha Sree Kurmeti

Abstract

Credit risk analysis plays a critical role in the financial sector by helping lenders assess the probability of loan default. This project presents a machine learning–based approach to evaluate credit risk using applicant financial and credit history. A synthetic dataset was created to simulate real-world loan application data. After data preprocessing and feature engineering, a Logistic Regression classification model was trained to predict loan default risk. Model performance was evaluated using a confusion matrix and ROC–AUC score. Finally, applicant-level risk scores were generated to categorize borrowers into different risk groups.

1. Introduction

Financial institutions rely on credit risk analysis to make informed lending decisions and minimize financial losses due to loan defaults. Traditional credit evaluation methods are increasingly being supplemented by data-driven and machine learning techniques. These methods allow for more accurate and consistent risk assessment by analyzing multiple financial indicators simultaneously.

The objective of this project is to develop a credit risk prediction system using machine learning that classifies loan applicants as potential defaulters or non-defaulters and assigns a risk score to each applicant.

2. Dataset Description

Since real banking data is confidential, a **synthetic dataset** was generated for academic purposes. The dataset consists of **1000 loan applicants** with the following attributes:

- Age
- Annual Income
- Credit Score
- Loan Amount
- Loan Term (in months)
- Employment Years
- Number of Existing Loans

The target variable is:

- **Default**
  - 1 → Loan default
  - 0 → No default

The dataset was designed to reflect realistic financial patterns observed in credit lending scenarios.

---

3. Data Preprocessing and Feature Engineering

Data preprocessing steps were performed to ensure data quality and suitability for modeling. Invalid or negative financial values were corrected. No missing values were present in the dataset.

Feature engineering was conducted to enhance predictive power:

- **Loan-to-Income Ratio** was created to capture borrower repayment capacity.
- Credit scores were grouped into risk categories (Poor, Fair, Good, Excellent).
- Categorical features were encoded using one-hot encoding.

The dataset was then split into **training (70%)** and **testing (30%)** subsets to evaluate model generalization.

---

4. Model Development

A **Logistic Regression** model was selected for this project due to its suitability for binary classification problems and its interpretability. Logistic Regression estimates the probability of loan default based on input financial features.

The model was trained using the training dataset and optimized to converge efficiently using an increased iteration limit.

---

5. Model Evaluation

Model performance was evaluated using the following metrics:

- **Confusion Matrix**:
  Displays true positives, true negatives, false positives, and false negatives, providing insight into classification accuracy.
- **Classification Report**:
  Includes precision, recall, and F1-score for both default and non-default classes.
- **ROC–AUC Score**:
  The Area Under the Receiver Operating Characteristic Curve measures the model's ability to distinguish between defaulters and non-defaulters. A higher AUC score indicates better model performance.

The evaluation results indicate that the model has good predictive capability and is effective in identifying high-risk applicants.

---

6. Risk Scoring and Categorization

After model training, probability estimates of default were generated for each applicant. These probabilities were converted into **risk scores (0–100)**.

Applicants were categorized into:

- **Low Risk**
- **Medium Risk**
- **High Risk**

This risk-based segmentation can help financial institutions tailor lending decisions, interest rates, or credit limits accordingly.

---

7. Conclusion

This project demonstrates the application of machine learning techniques in credit risk analysis using a synthetic dataset. The Logistic Regression model successfully classified loan applicants and generated meaningful risk scores. Such models can support financial institutions in making data-driven lending decisions and reducing default risk.

Future enhancements may include the use of advanced models such as Random Forests or Gradient Boosting, incorporation of additional behavioral features, and model validation using real-world financial datasets.

---

8. References

1. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
2. Scikit-learn Documentation – Classification Models
3. Credit Risk Modeling Concepts – Basel Committee Guidelines

Code used to generate the analysis-
https://colab.research.google.com/drive/1swYsVUI0fLAmqrIY8zN4DEhHWS1rXsKW?usp=sharing