Loan Repayment Prediction Using Machine Learning

Welcome to the Loan Repayment Prediction Project! This repository hosts a comprehensive machine learning project designed to predict the likelihood of successful loan repayment by borrowers.

Project Goals and Objectives

- **Identify Key Factors:** To identify key factors that have a significant impact on the loan repayment behavior of borrowers.
- **Evaluate Predictive Models:** To compare and evaluate the performance of different predictive models for loan repayment prediction.
- **ROC Curve Analysis:** Compare ROC curves of different predictive models and determine the model with the highest area under the curve (AUC) for loan repayment prediction.
- **Select Best Model:** To identify and select the best predictive model that provides accurate and reliable loan repayment predictions.

This project aims to leverage machine learning techniques to predict the probability of borrowers successfully repaying their loans. Accurate loan repayment prediction is vital for financial institutions to make informed lending decisions, reduce risks, and promote responsible lending practices.

Data Description

The dataset used in this project was obtained from Lending Club and comprises the following attributes: loan_data.csv

1. **credit.policy:** A binary variable indicating whether the customer meets LendingClub.com's credit underwriting criteria (1 for meeting the criteria, 0 otherwise).
2. **purpose:** The purpose of the loan, taking values such as "credit_card," "debt_consolidation," "educational," "major_purchase," "small_business," and "all_other."
3. **int.rate:** The interest rate of the loan, represented as a proportion (e.g., 0.11 for 11%).
4. **installment:** The monthly installment amount owed by the borrower if the loan is funded.
5. **log.annual.inc:** The natural log of the self-reported annual income of the borrower.

6. **dti:** The debt-to-income ratio of the borrower, calculated as the amount of debt divided by annual income.
7. **fico:** The FICO credit score of the borrower.
8. **days.with.cr.line:** The number of days the borrower has had a credit line.
9. **revol.bal:** The borrower's revolving balance, which represents the amount unpaid at the end of the credit card billing cycle.
10. **revol.util:** The borrower's revolving line utilization rate, indicating the amount of the credit line used relative to the total credit available.
11. **inq.last.6mths:** The number of inquiries by creditors made on the borrower's credit history in the last 6 months.
12. **delinq.2yrs:** The number of times the borrower had been 30+ days past due on a payment in the past 2 years.
13. **pub.rec:** The number of derogatory public records, including bankruptcy filings, tax liens, or judgments.
14. **Not.fully.paid:** The dependent variable, where "0" indicates that the loan was fully paid by borrowers, and "1" indicates that it was not fully paid.

This dataset consists of over 9578 records, making it a valuable resource for building predictive models for loan repayment behavior. It is used in this project to train and evaluate machine learning models for accurate loan repayment predictions.

Analysis and Conclusion

In this project, we conducted a comprehensive analysis of a loan repayment dataset to gain insights into the factors influencing loan repayment behavior. The analysis included the following key steps:

- **Descriptive Statistics:** We performed descriptive statistics to understand the distribution of variables in the dataset and identify any notable patterns or outliers.
- **Exploratory Data Analysis (EDA):** EDA was conducted to explore relationships between variables, uncover correlations, and visualize data trends. This step helped us identify potential predictors of loan repayment behavior.

Machine Learning Models

We employed various machine learning models to predict loan repayment outcomes, including:

- Logistic Regression
- Decision Tree
- Bagging

- AdaBoosting
- Random Forest
- Support Vector Machine
- Naïve Bayes Classifier
- K-NN Classification
- Artificial Neural Network

Model Performance Evaluation

We compared the performance of these models using a range of evaluation metrics, including accuracy, precision, recall, F1 score, and ROC curves. The results showed that the Random Forest model consistently outperformed other models across all metrics.

Feature Importance

A feature importance plot from the Random Forest model revealed significant predictors influencing loan repayment decisions, including variables such as inq.last.6mths, credit policy, interest rate, installment amount, FICO score, purpose, days with a credit line, revolving balance, revolving utilization rate, debt-to-income ratio, and log annual income.

Conclusion

In conclusion, the Random Forest model emerged as the most suitable and effective model for predicting loan repayment behavior. Its consistent high performance, ability to handle complex relationships between variables, and strong predictive power make it a reliable tool for lenders to make well-informed decisions when approving loans.

By leveraging insights from feature importance, lenders can focus on key factors that contribute to loan repayment behavior, improving their risk assessment strategies and optimizing loan approval processes. Additionally, analyzing loan repayment patterns over time can provide lenders with a deeper understanding of borrower behavior, enabling them to implement targeted strategies for ensuring successful loan repayment.