Predicting Loan Repayments

Abstract:

Predicting loan repayments is a crucial task in financial risk management. In this study, we present an anonymized loan default dataset comprising various features such as loan amount, term, interest rate, installment, employment length, home ownership, annual income, loan status, purpose, and others, alongside a binary indicator for repayment failure. Leveraging this dataset, we aim to develop predictive models to anticipate loan repayment behavior. With a comprehensive exploration of the data and the application of various prediction algorithms, our goal is to create a robust model capable of accurately predicting loan repayment outcomes. This research holds significant implications for improving lending practices and mitigating financial risks for both borrowers and lenders.

Data source and Description:

We have taken the data source from Kaggle (https://www.kaggle.com/datasets/joebeachcapital/loan-default).

The following table shows a description of the dataset.

No	Attribute	Description
1	id	This is a unique identifier assigned to each loan record in the dataset.
2	member_id	An identifier associated with the member (borrower) who applied for the loan.
3	loan_amnt	The total amount of the loan applied for by the borrower.
4	funded_amnt	The total amount of the loan funded by the lending institution.
5	funded_amnt_inv	The total amount of the loan funded by investors.
6	term	The length of the loan term, expressed as the number of months.
7	int_rate	The annual interest rate charged on the loan as a percentage.
8	installment	The monthly payment amount due for the loan, including both principal and interest.
9	emp_length	Length of employment in years for the borrower at the time of applying for the loan.
10	home_ownership	Indicates the type of home ownership status of the borrower (own, mortgage, or rent).
11	annual_inc	The annual income reported by the borrower during the loan application.
12	verification_status	Indicates whether the borrower's income was verified by the lending institution.
13	issue_d	The date when the loan was issued.
14	loan_status	Indicates the current status of the loan (fully paid, charged off, or does not meet the credit policy).
15	purpose	Describes the purpose of the loan as stated by the borrower (e.g., debt consolidation, medical, credit card, etc.).
16	zip_code	The first three digits of the borrower's zip code.
17	addr_state	The state where the borrower resides.
18	dti	Debt-to-Income ratio, representing the borrower's total monthly debt payments divided by their gross monthly income.
19	delinq_2yrs	The number of times the borrower has been delinquent on payments in the past 2 years.
20	earliest_cr_line	The date when the borrower's earliest credit line was opened.
21	inq_last_6mths	The number of inquiries made by creditors on the borrower's credit report within the last 6 months.
22	mths_since_last_delinq	The number of months since the borrower's last delinquency on payments.

23	open_acc	The number of open credit lines in the borrower's credit file.
24	pub_rec	The number of derogatory public records (e.g., bankruptcies, tax liens) on the borrower's credit report.
25	revol_bal	The total revolving balance (i.e., credit card balance) owed by the borrower.
26	revol_util	The revolving line utilization rate, or the percentage of available credit being used by the borrower.
27	total_acc	The total number of credit lines in the borrower's credit file.
28	total_pymnt	The total amount of payments received on the loan.
29	total_pymnt_inv	The total amount of payments received on the loan by investors.
30	total_rec_prncp	The total amount of principal received to date.
31	total_rec_int	The total amount of interest received to date on the loan.
32	last_pymnt_d	The date of the most recent payment received from the borrower.
33	last_pymnt_amnt	The amount of the most recent payment received from the borrower.
34	next_pymnt_d	The date when the next payment is due from the borrower.
35	last_credit_pull_d	The date when the last credit report inquiry was made by the lending institution.
36	repay_fail	A binary indicator representing loan repayment failure, where 1 indicates failure to repay the loan and 0 indicates successful repayment.

Goal of the Project:

The goal of the project is to predict loan repayment behavior and identify key factors influencing repayment success or failure. Through data exploration and model development, we aim to address the following objectives:

Data Exploration:

1. Investigate trends in loan repayment outcomes across different categories to discover which are the most impactful on loan repayment.
2. Analyze the relationship between loan repayment status and a variety of attributes such as loan amount, interest rate, employment length, annual income, and purpose of the loan.

Prediction Model:

1. Data Understanding and Preparation: Perform data preprocessing tasks including cleaning, merging, exploration, normalization, handling missing values, and feature engineering.
2. Modeling: Apply various Probit or Cumulative Link Models to predict loan repayment outcomes based on relevant attributes identified during data exploration (we are using these models because y is binary and regressors are a mix of ordinal factors and continuous variables).
3. Validation: Assess the performance of the classification models on a separate test dataset and generate evaluation metrics to measure predictive accuracy.
4. Conclusion: Summarize insights gained from the predictive models and discuss implications for lending institutions. Explore how these models can be used to optimize lending practices, identify high-risk borrowers, and enhance decision-making processes related to loan approvals and monitoring.