Predicting Loan Repayment Outcomes During Global Financial Crisis

Risk Modeling with Random Forest and Logistic Regression

Created by Nailya Alimova, Simmons University

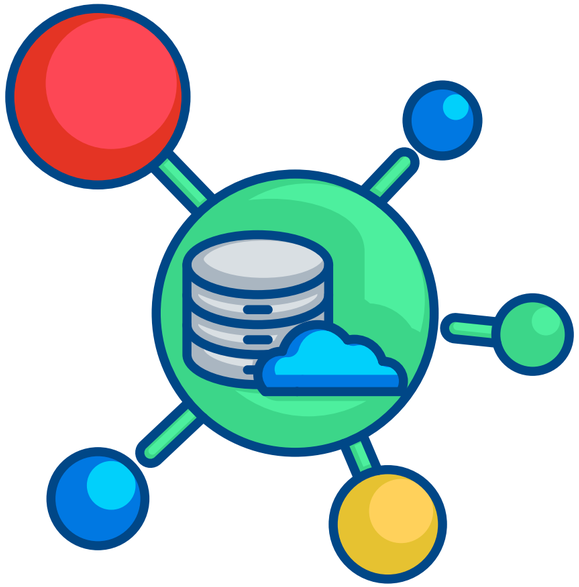
Date: May 1st, 2025

Kaggle Dataset

Source:
Lending Club Loan Data from Kaggle

Objective:
Predict loan repayment outcomes
using borrower and loan features.

Dataset Size:
Over 2.2 million U.S. loans
issued between 2007 and 2017



Why is assessing loan repayment outcomes critically important?

Not long ago, the world was shaken by the Global Financial Crisis (GFC) — a seismic event largely triggered by a massive wave of loan defaults, especially in the mortgage and consumer lending sectors.

Lenders and investors underestimated borrower risk, mispriced loans, and built financial products on top of vulnerable, poorly understood debt.

Before the crisis, financial institutions relied heavily on flawed or overly simplistic risk models that underestimated borrower risk, failed to account for rising default correlations, and ignored critical signals like unsustainable debt-to-income ratios (dti),
as well as full borrower credit histories with various metrics.

Why is assessing loan repayment outcomes critically important?

By accurately predicting which borrowers are more likely to default or repay lenders can:

- Avoid issuing high-risk loans that destabilize their balance sheets.
- Set appropriate interest rates and loan terms to reflect true risk.
- Protect financial systems from systemic risk by reducing exposure to highly correlated, poorly understood debt pools.

Better predictive modeling directly addresses one of the root failures of the GFC: mispricing of credit risk. also, predictive modeling helps prevent future crises by ensuring that lending practices are grounded in robust, data-driven risk assessment.

Let’s explore the variables of interest!

Number of variables in the original dataset: 145

Response Variable:
- "loan_status" - borrower's loan status with four categories: “Fully Paid”, “Charged Off”, “Does not meet the credit policy. Status: Fully Paid", "Does not meet the credit policy. Status: Charged Off".

Explanatory Variables:

- loan_amnt - amount the borrower applied for (USD);
- funded_amnt - actual amount funded by investors (USD);
- grade - lending Club–assigned credit grade (A–G);
- int_rate - annual interest rate set on the loan (%);
- installment - borrower’s fixed monthly repayment amount (USD);
- emp_length - borrower’s reported employment length (e.g., “10+ years”)
- annual_inc - borrower’s self-reported annual income (USD);

Let’s explore the variables of interest!

Number of variables in the original dataset: 145

Response Variable:
- "loan_status" - borrower's loan status with four categories: “Fully Paid”, “Charged Off”, “Does not meet the credit policy. Status: Fully Paid", "Does not meet the credit policy. Status: Charged Off".

Explanatory Variables:

- dti - debt-to-income ratio, showing how leveraged the borrower is (%);
- open_acc - number of currently open credit lines;
- revol_bal - total revolving balance (mainly credit cards) (USD);
- revol_util - revolving credit utilization rate (%);
- mths_since_last_delinq - months since the last delinquency (missing if none recorded)
- - verification_status - whether the borrower’s income was verified (e.g., “Verified”, “Not Verified”)

Let’s explore the variables of interest!

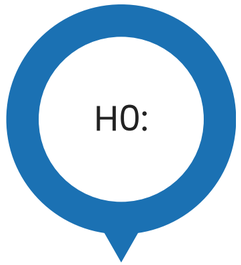
Number of variables in the original dataset: 145

Response Variable:
- "loan_status" - borrower's loan status with four categories: “Fully Paid”, “Charged Off”, “Does not meet the credit policy. Status: Fully Paid", "Does not meet the credit policy. Status: Charged Off".

Other Variables of Interest:

- issue_d - loan issue date (Month-YYYY)
- term - length of the loan term (e.g., 36 or 60 months);
- home_ownership - borrower’s homeownership status (e.g., “RENT”, “OWN”);
- purpose - reason for the loan (e.g., “debt consolidation”)
- total_acc - the total number of credit accounts the borrower has.

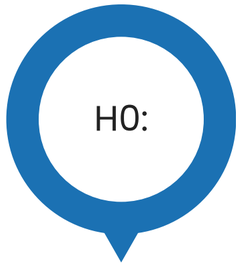
Project Hypothesis:
Borrower and loan characteristics significantly influence loan repayment outcomes.



Null Hypothesis (H_0) 1:
The logistic regression model's prediction accuracy, measured by the AUC (area under the curve) metric, is 0.5 or lower, indicating performance no better than random chance.

Alternative Hypothesis (H_1) 1:
The logistic regression model's prediction accuracy, measured by the AUC metric, exceeds 0.5, demonstrating meaningful predictive power.

Project Hypothesis:
Borrower and loan characteristics significantly influence loan repayment outcomes.



Null Hypothesis (H_0) 2:
The random forest model's prediction accuracy, measured by the AUC metric, is 0.5 or lower, indicating performance no better than random chance.

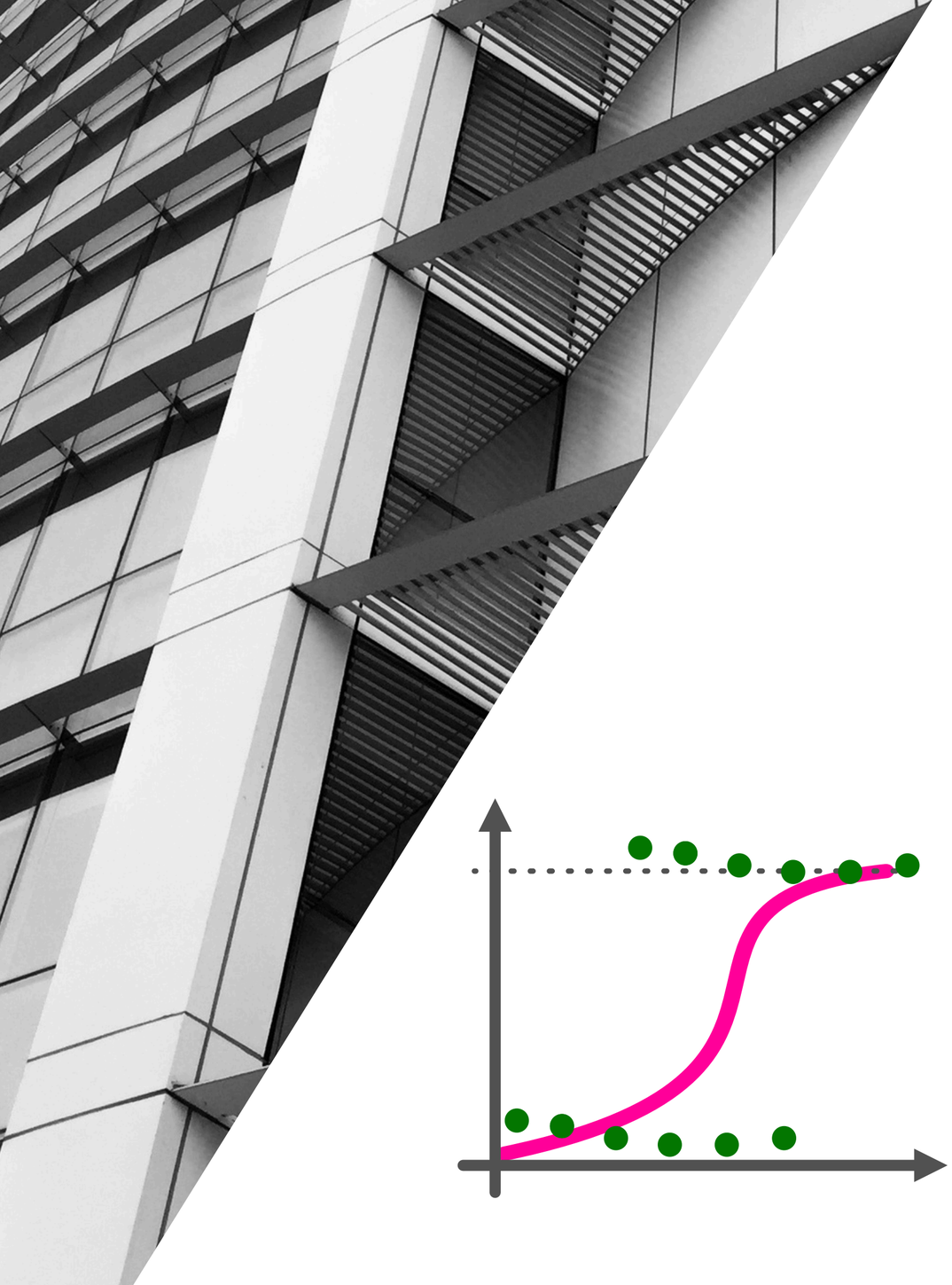
Alternative Hypothesis (H_1) 2:
The random forest model's prediction accuracy, measured by the AUC metric, exceeds 0.5, demonstrating meaningful predictive power.

Methodology – Random Forest

Random Forest is an ensemble learning method using multiple decision trees to capture nonlinear patterns and interactions.

- Advantages:
  - Captures complex interactions between variables.
  - Robust to outliers and missing values.

- Disadvantages:
  - Less interpretable than logistic regression.
  - Computationally intensive with very large datasets.
  - Requires careful tuning to avoid bias towards dominant features.



Methodology – Logistic Regression

Logistic Regression is a statistical model estimating the probability of a binary outcome based on linear relationships between predictors and the log-odds of the event.

- Advantages:
  - Highly interpretable coefficients (odds ratios).
  - Efficient computation, suitable for large datasets.
  - More effective when linear relationships dominate.

- Disadvantages:
  - Assumes linearity in the logit, potentially missing complex interactions.
  - Sensitive to outliers and multicollinearity.

Data Manipulation Process

Data Cleaning

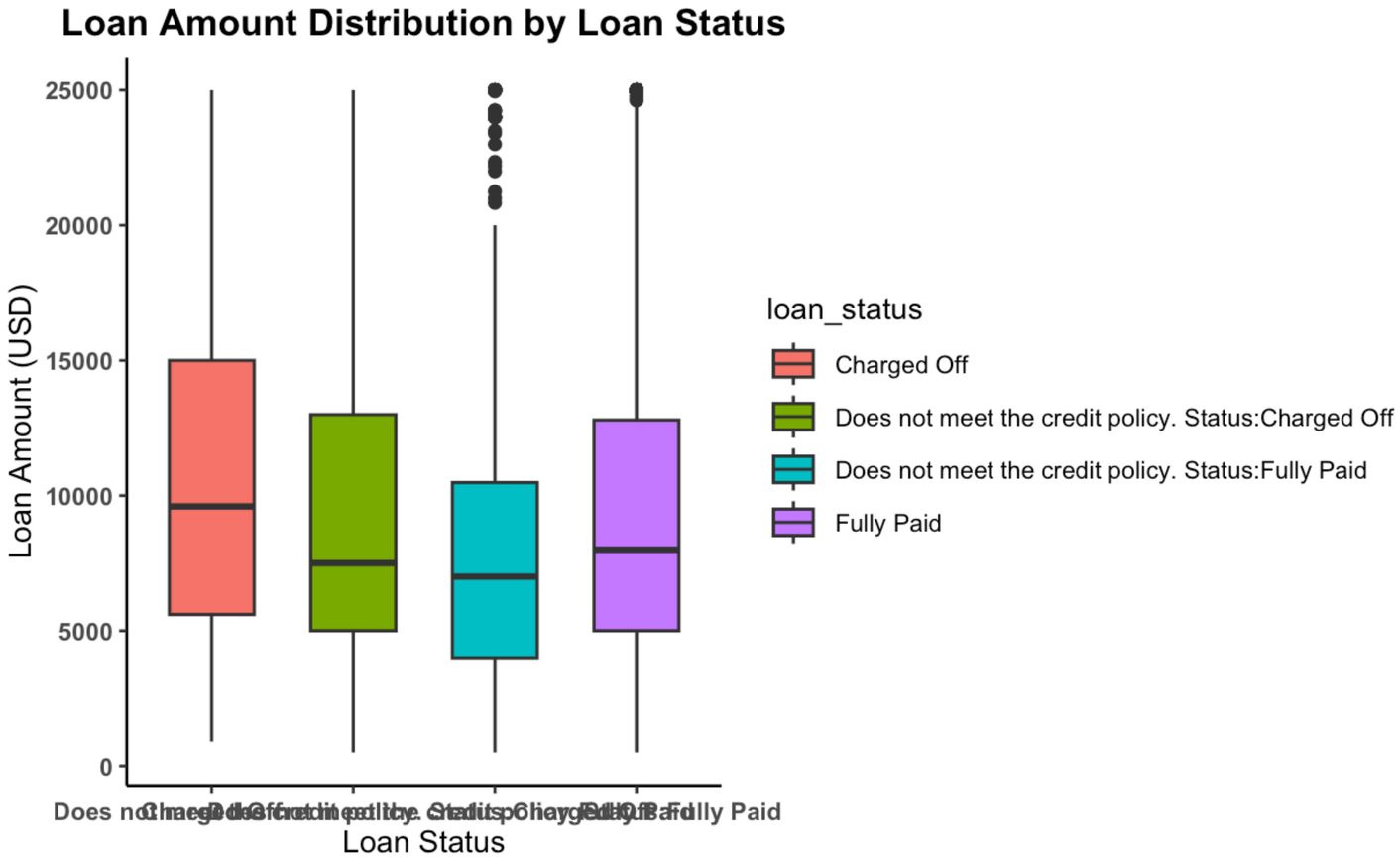
Data Modification

Model 1 - Random Forest

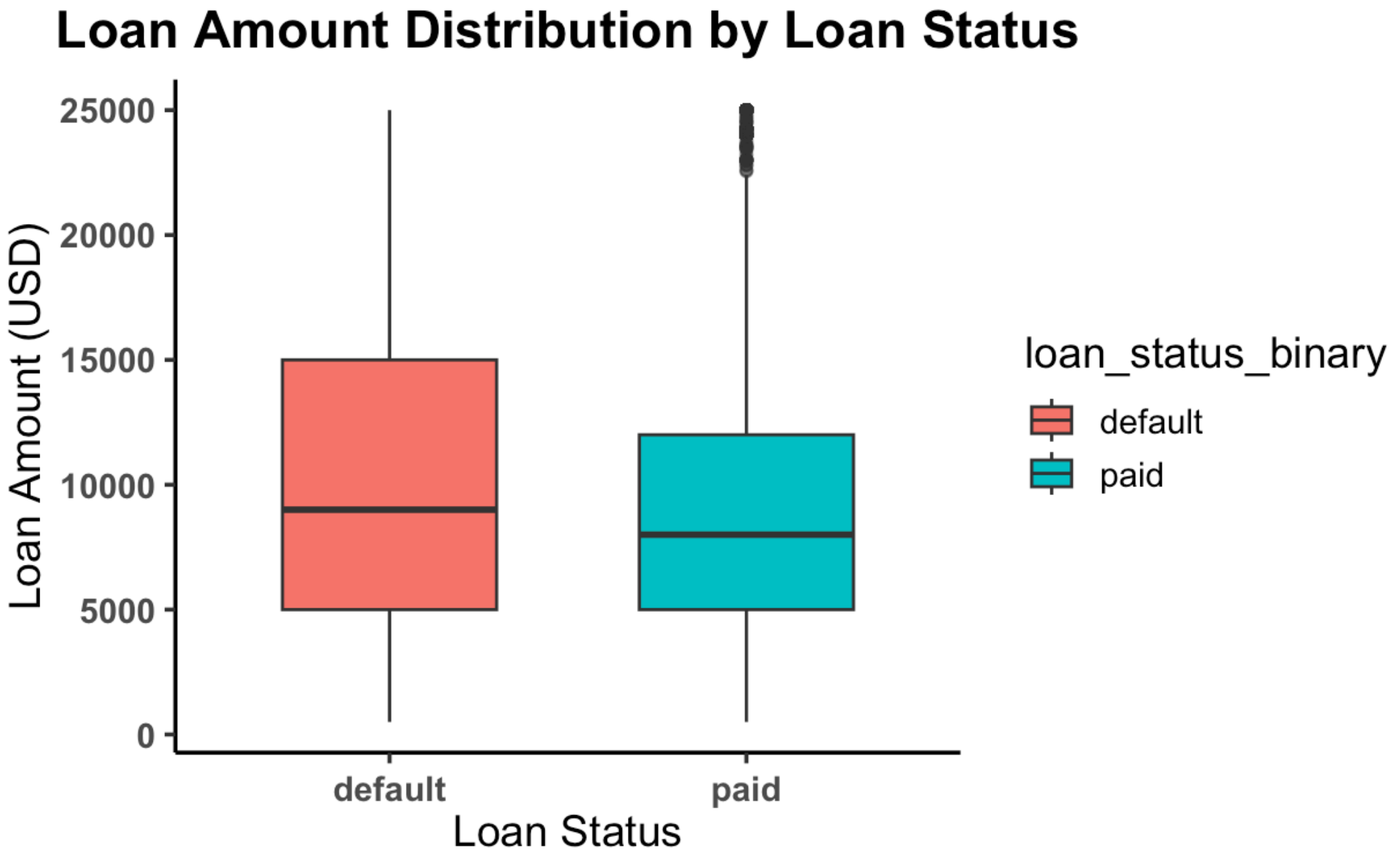
Model 2 - Logistic Regression

Models' Evaluation

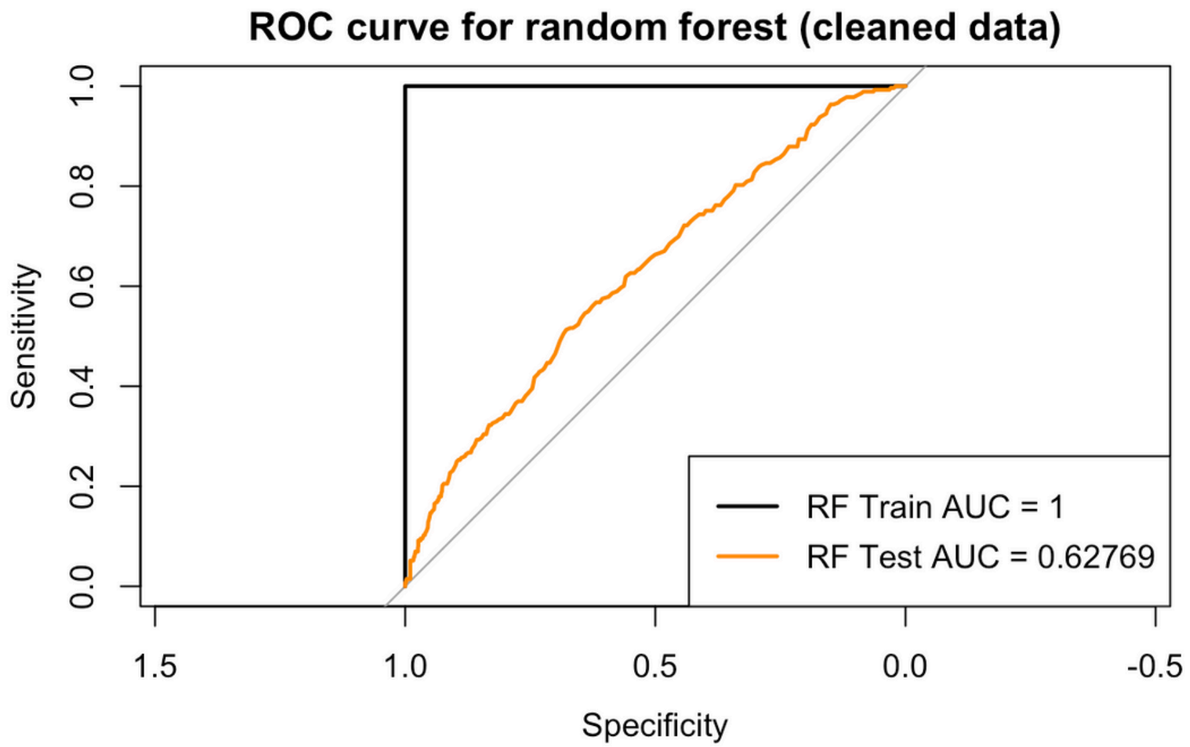
Explorartory Data Analysis

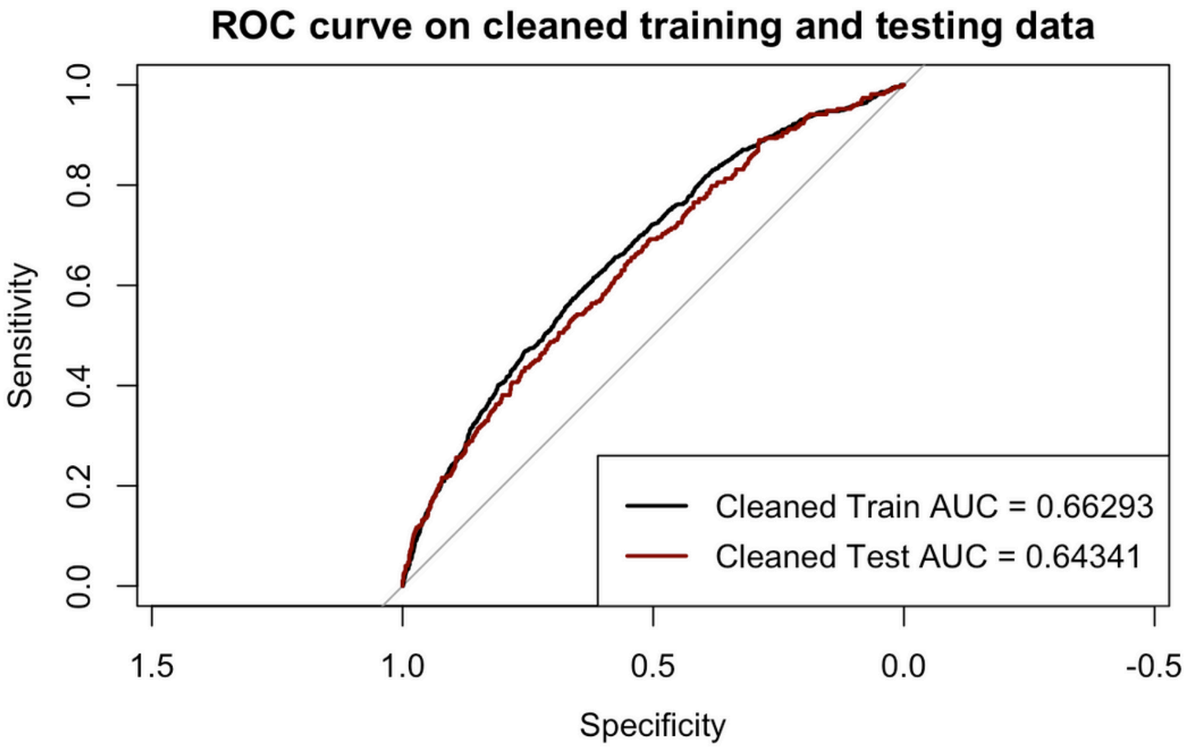


Explorartory Data Analysis



Data Modeling





Conclusion

In conclusion, this project applied both logistic regression and random forest models to predict loan repayment outcomes using Lending Club data from the global financial crisis period.

Despite the random forest's complex ensemble approach, logistic regression performed slightly better, achieving a testing AUC of approximately 0.64, compared to the random forest's testing AUC of approximately 0.63.

Hypotheses Results: we can reject the first null hypothesis, because our logistic regression model with AUC = 0.64 is above random chance (0.5); and we can also reject the second null hypothesis, because our random forest model with AUC = 0.63 is above random chance (0.5).

References: Adarsh S. (2021). Lending Club Loan Data (CSV). Kaggle : https://www.kaggle.com/datasets/adarshsng/lending-club-loan-data-csv

Thank You!

Questions?
alimova@simmons.edu

Let’s connect
on LinkedIN!