[**Loan Data | Kaggle**](https://www.kaggle.com/datasets/itssuru/loan-data)

[**https://raw.githubusercontent.com/mrbarron3/stat451/main/loan\_data.csv**](https://raw.githubusercontent.com/mrbarron3/stat451/main/loan_data.csv)

* **Link so you can directly read in the data without downloading the file first**

Let’s focus on **debt consolidation**. That specific purpose has 3957 entries, whereas the next highest purpose has 1262 rows (credit card), and educational only has 343

**Write-up:**

Question(s): Will a borrower repay their loan in full? How can we classify whether a person will pay back their loan based on various features like FICO score, debt-to-income ratio, and interest rate?

Lenders earn money from people who repay their loans, but lose money from people who do not repay their entire loan. So our goal will be to predict, with as much accuracy as possible, whether someone will repay their loan based on a range of variables. In application, this model could help determine whether or not a person’s loan application should be approved or denied.

**Dataset and chosen variables:** This dataset was found on Kaggle, but the data itself came from LendingClub.com where it is publicly available.

The dataset has lending data from 2007-2010 and includes 14 variables, but we will be using a subset of those variables for our purposes. We will be predicting the not.fully.paid variable for debt consolidation loans based on the variables int.rate, installment, log.annual.inc, dti, fico, days.with.cr.line, revol.bal, and delinq.2yrs, which we will describe below.

* purpose:
  + Str
  + Describes the purpose of the loan
* int.rate
  + Float (between 0 and 1)
  + Shows the interest rate of the loan as a proportion
* installment
  + Float
  + Describes the monthly installments to be paid
* log.annual.inc
  + Float
  + Describes the natural log of the annual income of the borrower
* dti
  + Float
  + Describes the debt-to-income ratio for the borrower (debt/annual income)
* fico
  + Int
  + Shows the FICO credit score of the borrower
* days.with.cr.line
  + Float
  + Describes how many days the borrower has had a credit line
* revol.bal
  + Int
  + Shows the amount unpaid at the end of the credit card billing cycle
* delinq.2yrs
  + Int
  + Within the past 2 years, describes the number of times the borrower has been 30+ days past due on a payment
* not.fully.paid
  + Int (0 or 1)
  + Describes whether the loan was fully paid back (0) or not (1)
  + This will be the value we are predicting using the other columns

**Lines to read in code:**

import pandas as pd

df = pd.read\_csv('https://raw.githubusercontent.com/mrbarron3/stat451/main/loan\_data.csv')[['int.rate', 'installment', ‘log.annual.inc’, 'dti', 'fico', 'days.with.cr.line', 'revol.bal', 'delinq.2yrs', 'not.fully.paid','purpose']]

**Methods:**

We will be using a decision tree classifier to model this problem. Our feature names will be int.rate, installment, log.annual.inc, dti, fico, days.with.cr.line, revol.bal, and delinq.2yrs. Our y value that we are predicting is the not.fully.paid variable. At each node, we will choose the feature and threshold on which to split by minimizing the cost associated with the split. The best subset pair is the one that minimizes the weighted average entropy of the split.

**Other Suggestions**

* 750 word max for final report
* Try several classifiers/parameters with hyperparameter tuning
* Find more questions/more depth
  + Exploratory data analysis with graphs
  + Feature importance/analysis
    - Lasso(log regression) implicitly does this?
* Consider using purpose as a feature variable
* Since we are using feature selection, keep as many variables as possible