

Data Analysis:

Predicting Loan Chargeoff

In this problem, you will forecast the outcome of a portfolio of loans. Each loan is scheduled to be repaid over 3 years and is structured as follows:

- First, the borrower receives the funds. This event is called origination.
- The borrower then makes regular repayments, until one of the following happens:
 - The borrower stops making payments, typically due to financial hardship, before the end of the 3-year term. This event is called chargeoff, and the loan is then said to have charged off.
 - The borrower continues making repayments until 3 years after the origination date. At this point, the debt has been fully repaid.

In the attached CSV, each row corresponds to a loan, and the columns are defined as follows:

- The column with header `days since origination` indicates the number of days that elapsed between origination and the date when the data was collected.
- For loans that charged off before the data was collected, the column with header `days from origination to chargeoff` indicates the number of days that elapsed between origination and chargeoff. For all other loans, this column is blank.

We would like you to estimate what fraction of these loans will have charged off by the time all of their 3-year terms are finished. Please include a rigorous explanation of how you arrived at your answer, and include any code you used. You may make simplifying assumptions, but please state such assumptions explicitly.