Lending Club Case Study

MOHAN RAM BHADRAVATI

Problem statement

Analyze the past loan applicant data set

Identify driver variables which affect the likelihood of loan default

Provide actionable insights for future loan applications

Analysis approach

Understand the details of each column in the dataset from data_dictionary.csv file and identify variables which appear relevant for the problem

Extract past loan application data from the file ‘loan.csv’ and clean the data for the identified variables

Derive additional metrics from the data as needed for analysis

We compute **percentage delinquencies** which is selected as the target variable

We **generate derived categorical variables from selected continuous variables**

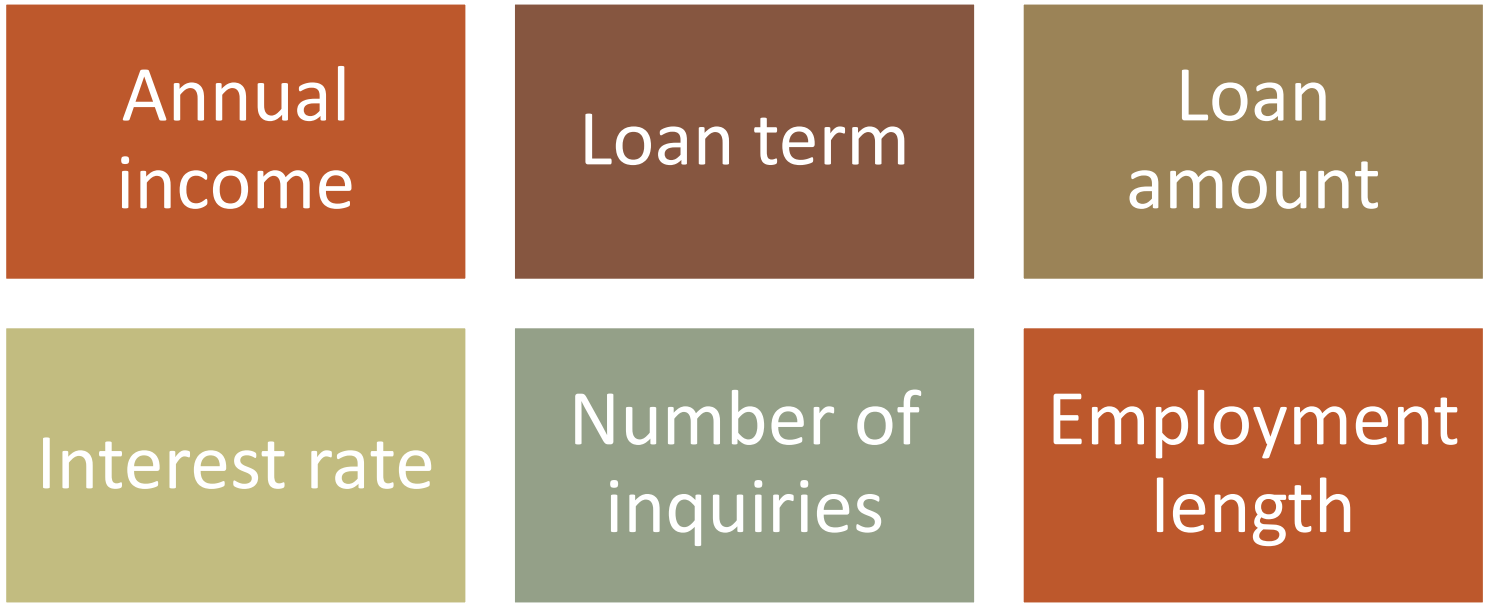
Perform univariate, segmented univariate and bivariate analysis with visualizations

Primarily use **bar charts and seaborn plots** to understand the correlation

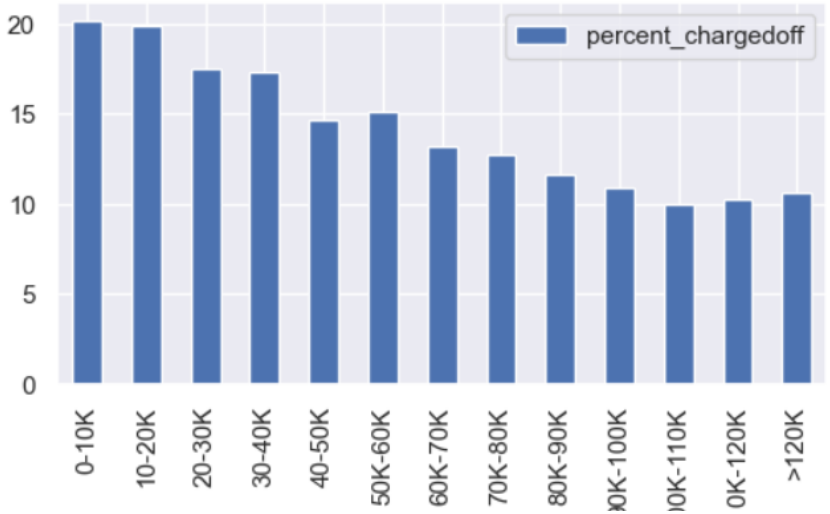
Additionally, **heatmaps** are generated for understanding the correlation against combination of some variables

Capture and present the insights from analysis

Key variables affecting loan delinquencies



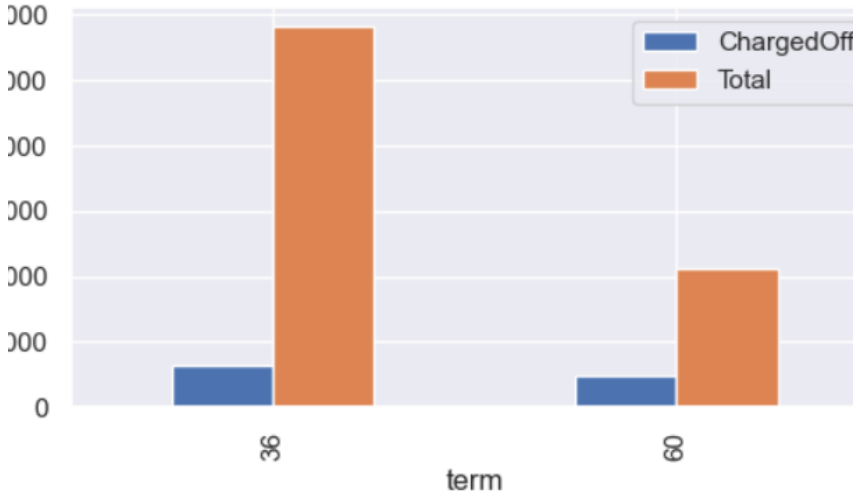
- *In the following slides we present the details for each of these*
- *Subsequent slides will present some key observations from combination of variables and present some recommendations*

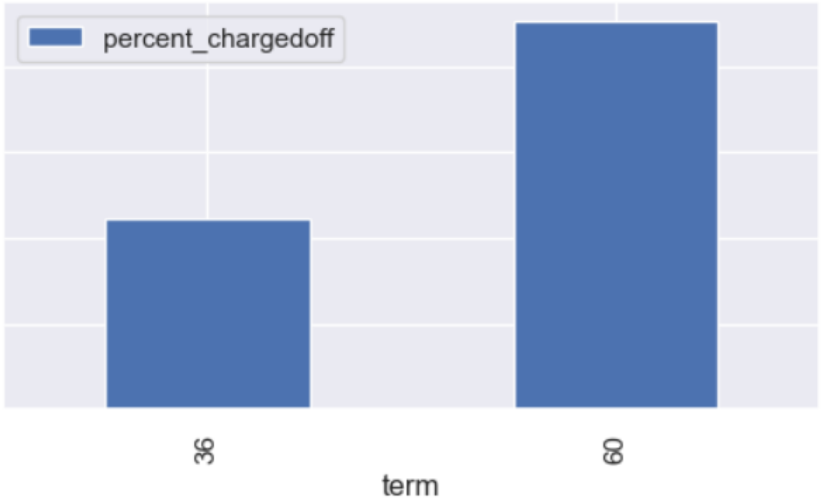


Annual income

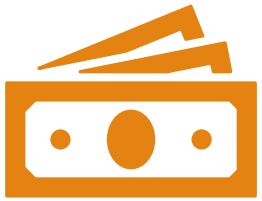
Most applicants have annual income in the range of 20K to 100K

% of delinquencies decline linearly with increasing income

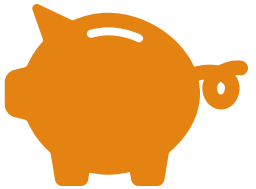




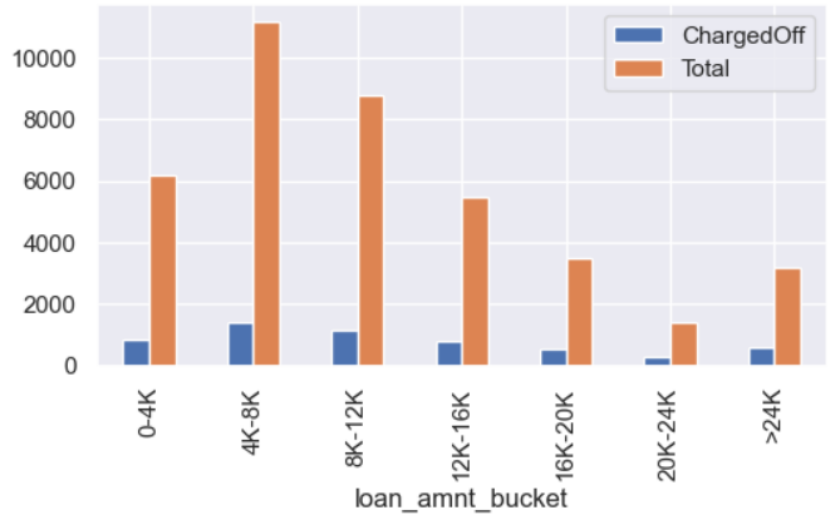
Loan term

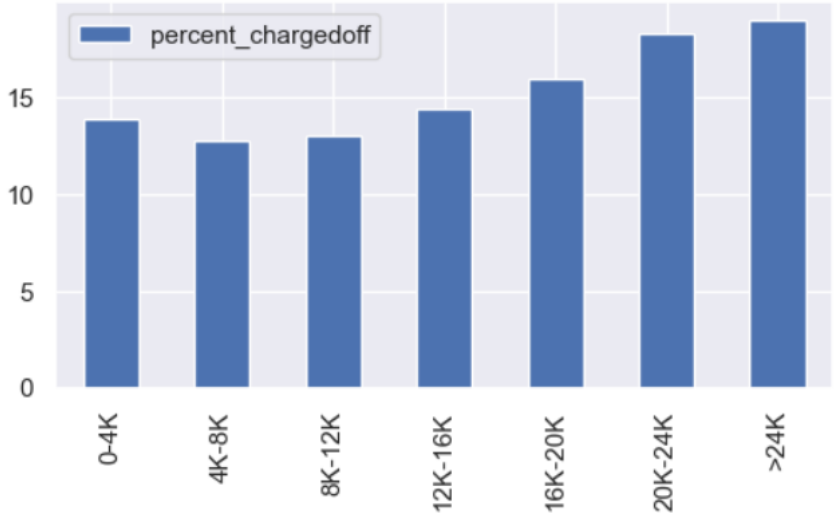


% delinquencies for loan term of 60 months is almost two times the % for loan term of 36 months



Majority of loans are 36-month term

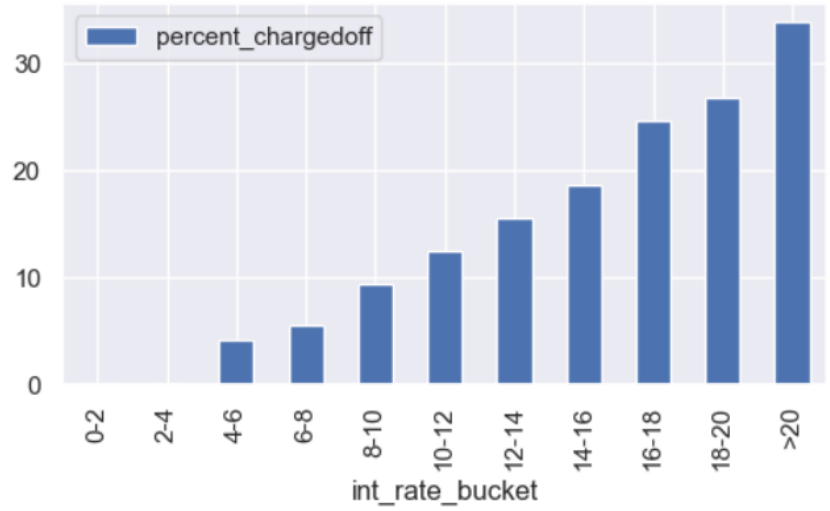




Loan amount

Major proportion of loan amounts are 0-20K

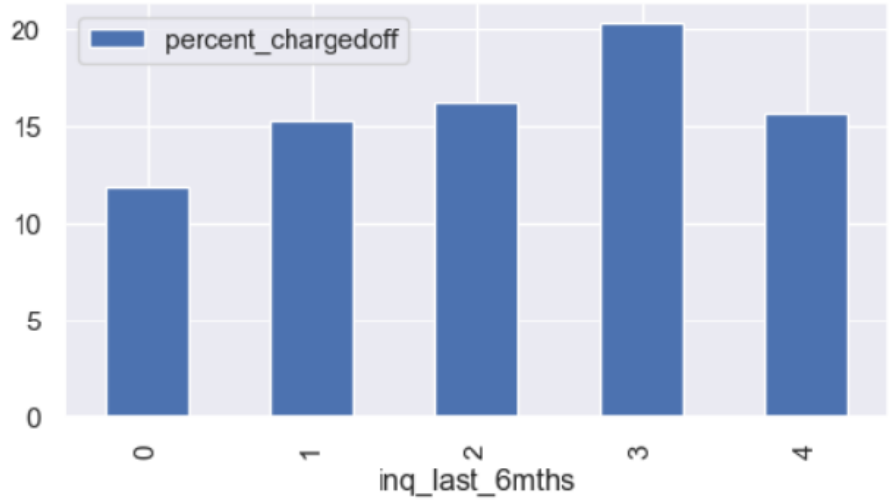
% delinquencies are seen rising with loan amount almost at a uniform rate

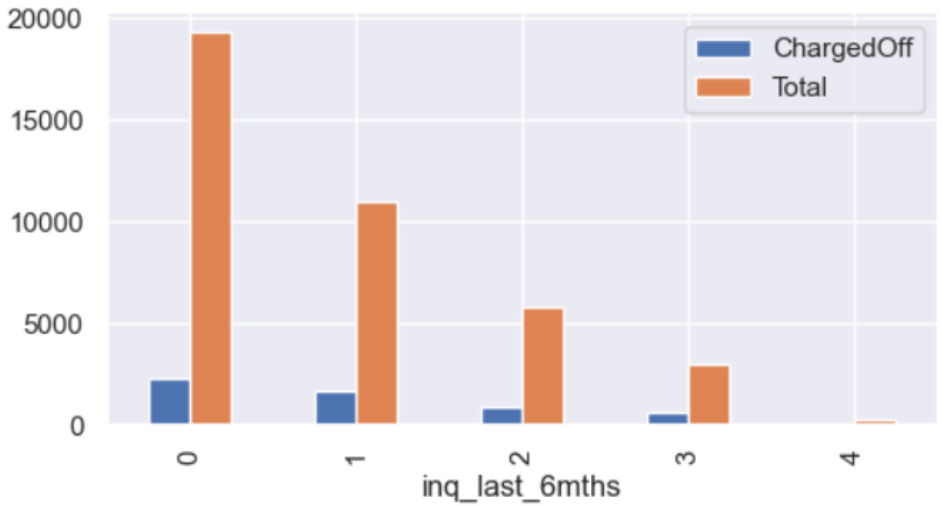


Interest rate

% delinquencies are seen increasing with interest rate linearly

The interest rates are mostly between 6% to 18%

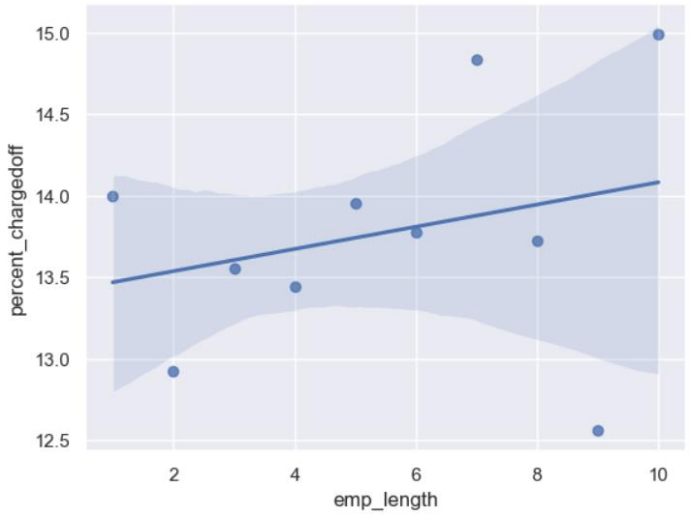




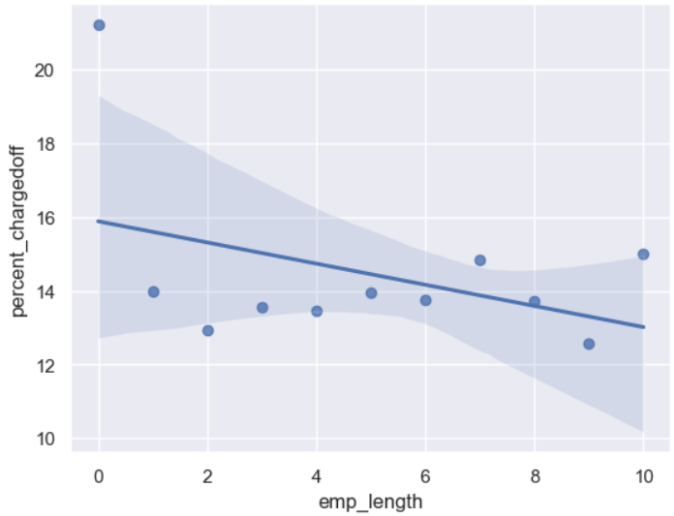
Number of inquiries

Most loans are issued on first or second inquiry

%delinquencies are higher for loans issued after first inquiry

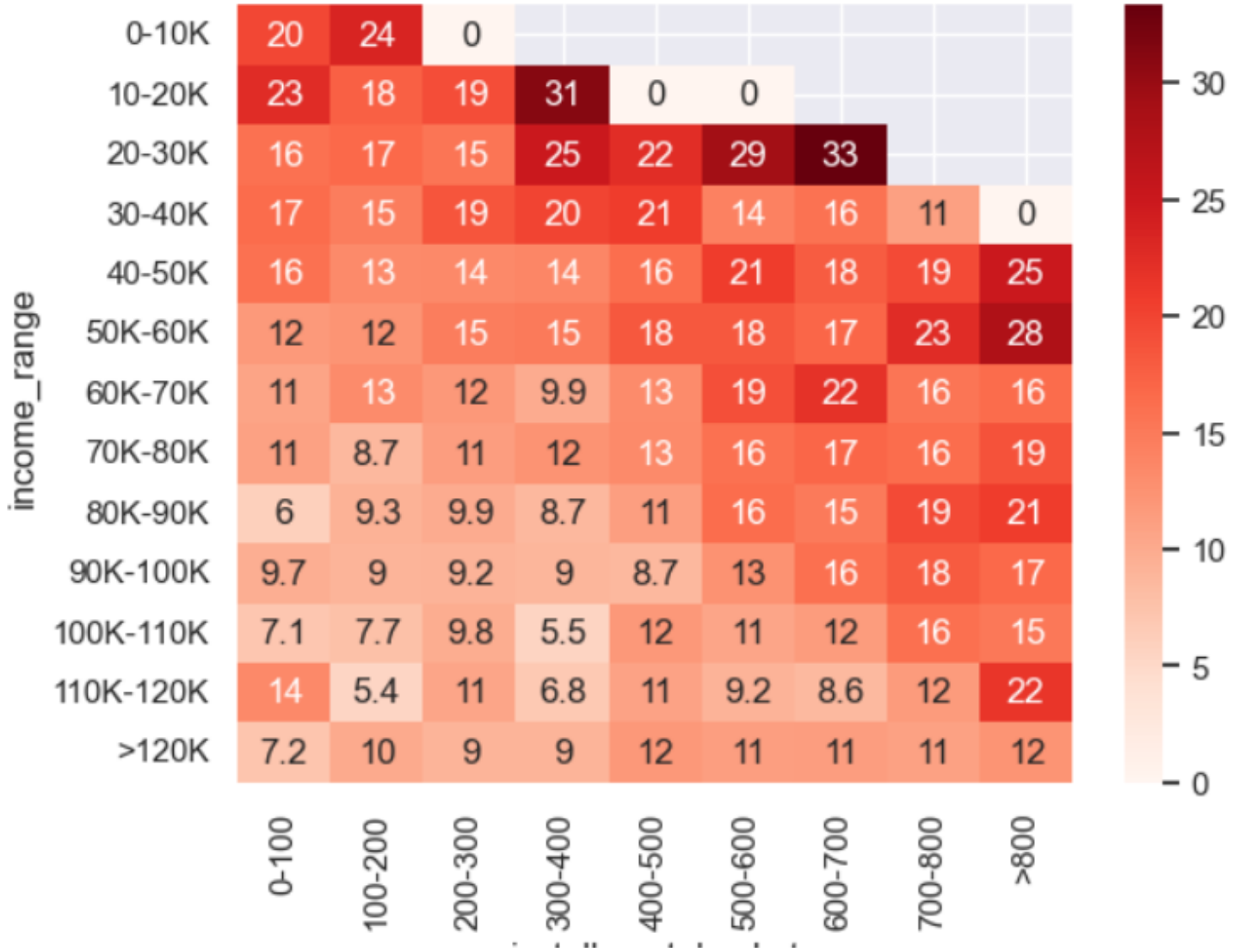


(top)



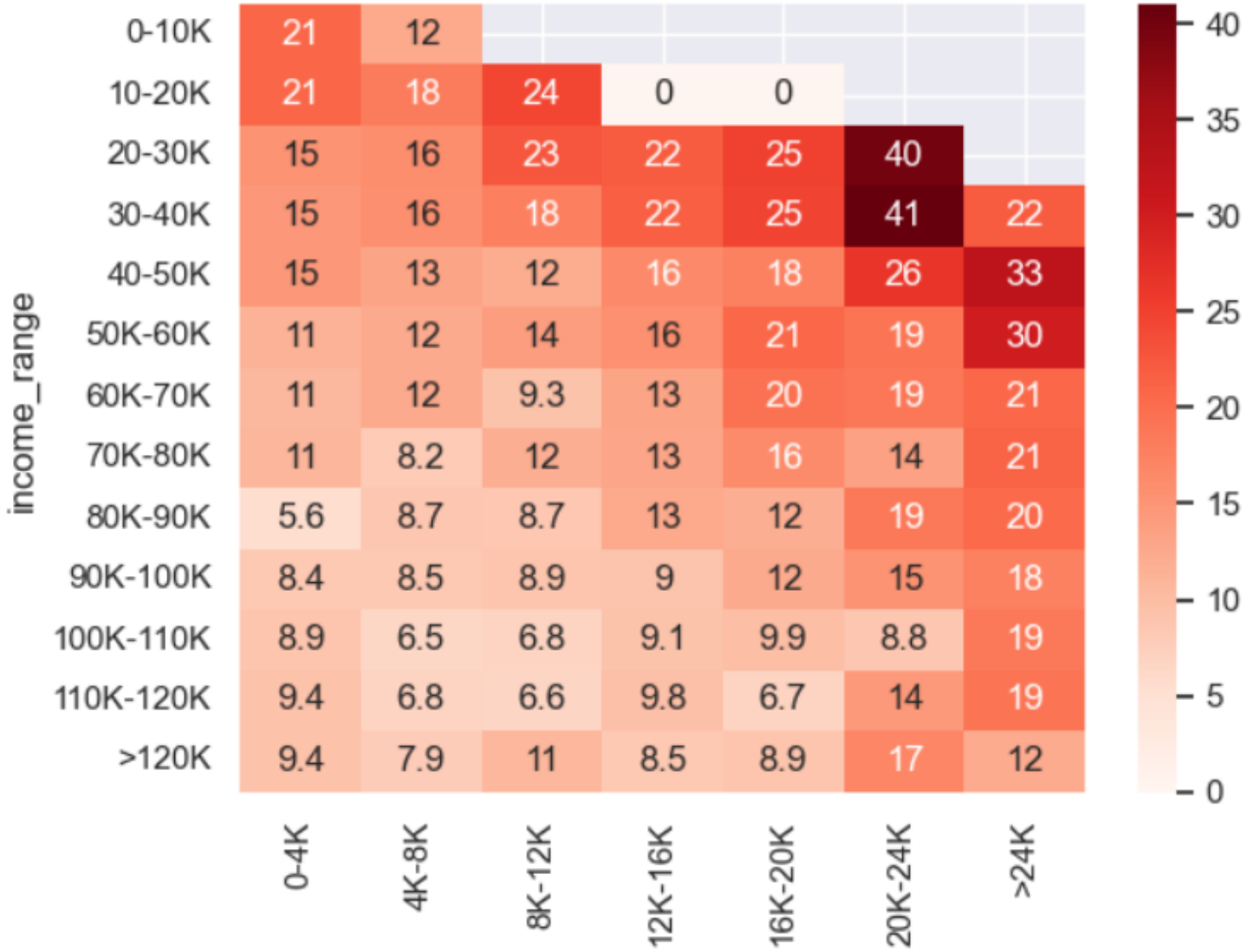
Employment length

- We see two charts showing the %delinquencies plotted against employment duration
- The bottom chart includes the %delinquencies for **employment** duration **less than one year**. The chart on the top does not have the same.
- We see a significantly high delinquency rate for employment duration <1 (22%). When we exclude it (top plot) we see that for all other values there is no correlation with %delinquency.



Annual income and Installment amount

- This heatmap shows the values of %delinquencies for combination of different values of income range and different values of installment bucket
- For income range from 10-60K, reducing the maximum installment amount by 200-400 could reduce default rate significantly

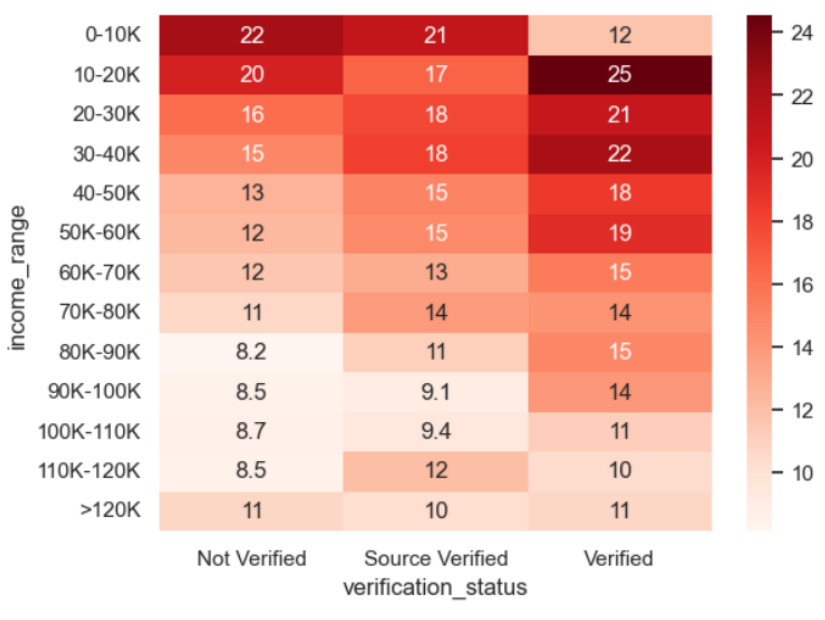


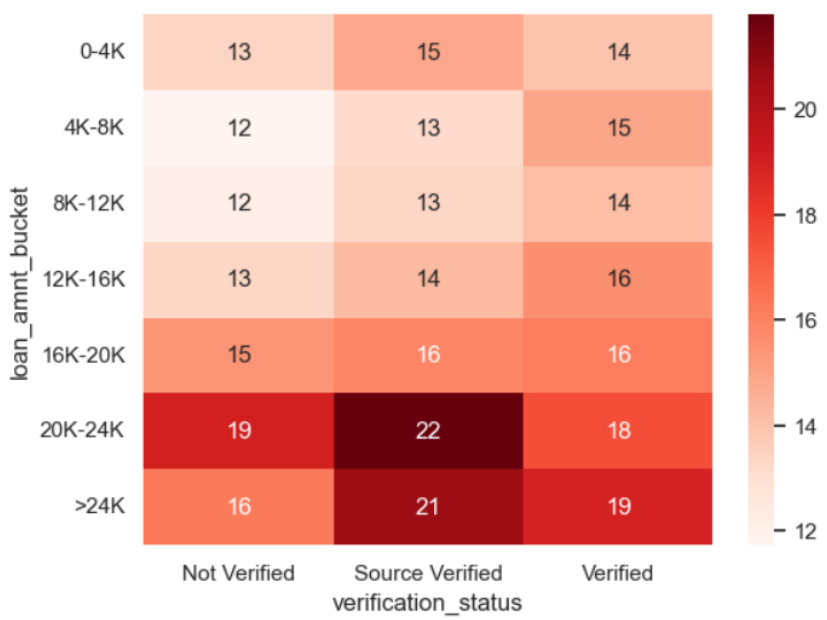
Annual income and Loan Amount

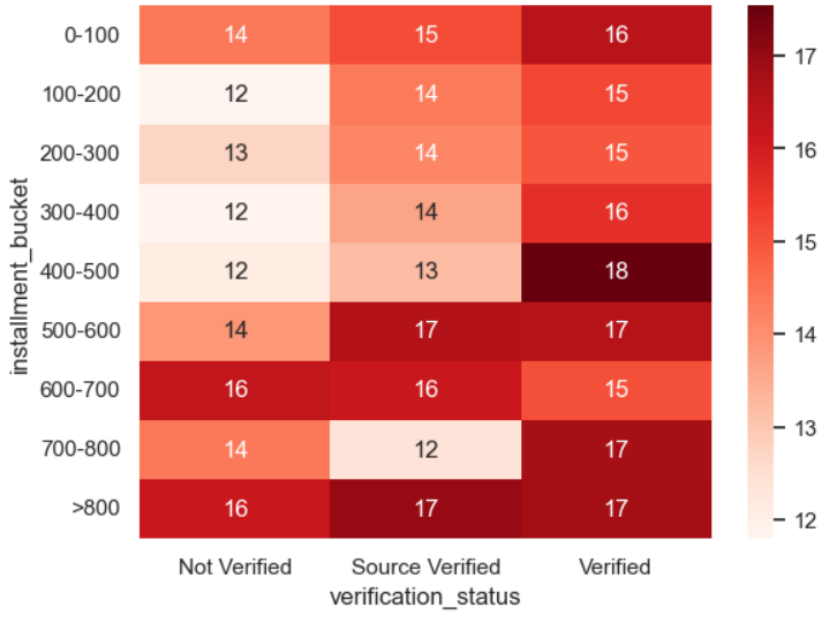
This heatmap shows the values of %delinquencies for combination of different values of income range and different values of loan amount.

For income ranges up to 60K, reducing the maximum loan amount by 4-8K could reduce default rates significantly

Income verification







The above graphs show that income verification does not help reduce delinquencies for any values of annual income, or loan amount or installment amount.

Recommendations

- Applicants who are employed with annual income above 60K with short term (36 months) loan are the ones who are least likely to default on loan
- For applicants with annual income up to 60K, reducing the monthly installment amount by 200-400 from the current maximum could reduce default rate significantly
- For applicants with annual income up to 60K, reducing the loan amount by 4000-8000 from the current maximum could reduce default rates significantly
- Income verification is not useful and can be avoided
- Recommend collecting data for fields such as “Number of accounts ever 120 or more days past due” and others for better analysis. Perhaps age group of applicants could be another useful metric

Thank you!