Lending Club case study

Submitted by: Konda Reddy Golamaru

AGENDA

- Problem Statement
- Goal
- Data understanding
- Data cleaning
- Data Analysis
- Plotting and Insights
- Recommendations

Problem STATEMENT

- Lending Club is a marketplace for personal loans that matches borrowers who are seeking a loan with investors looking to lend money and make a return.
- When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company

Goal

- The goal of analysis is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

Data understanding

- Read through the data dictionary provided as part of use case. Understood each column and the description about column
- Understood which attributes could be categorical by looking at sample data
- Understood which columns are significant for analysis by looking at % null values each column contains and if the column contains distinct values or not

Data cleaning

- Checked the percentage of missing values
- Removed all those columns with very high missing percentage i.e. 80% and above
- For columns with less missing percentage, dropped the rows (As imputation is not in scope for this use case)
- Created new derived columns
- Type conversion as needed e.g. String to Int
- Cleaning of data i.e. removing special characters like “%”

Data Analysis

- Analyzed the data and identified that the attributes fall into 3 categories.

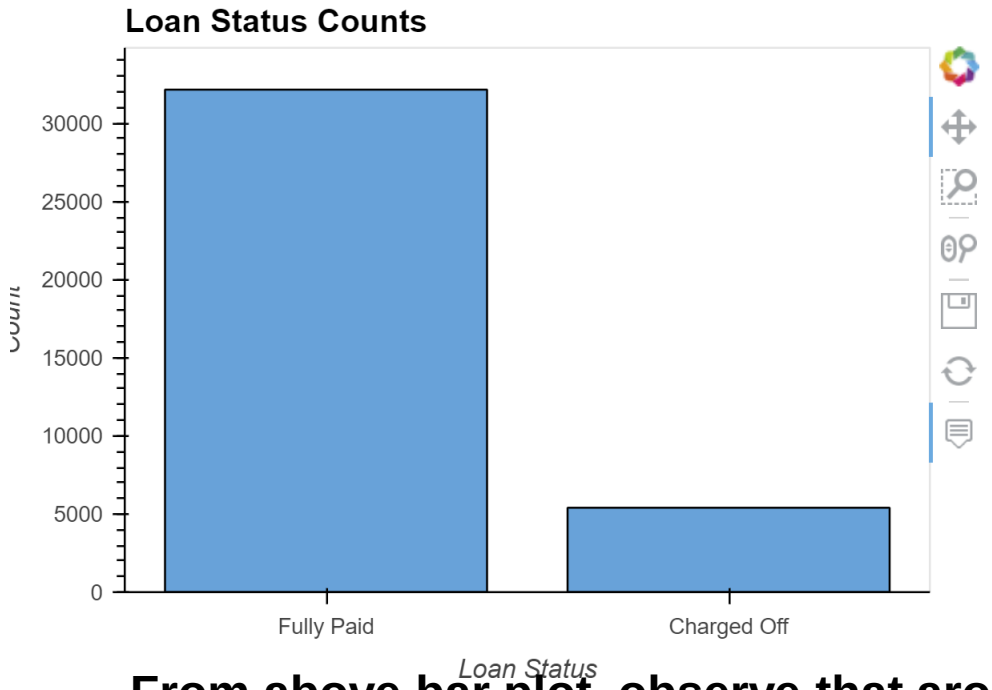
  1. demographic variables such as age, occupation, employment details etc.

  2. Loan characteristics (amount of loan, interest rate, purpose of loan etc.)

  3. Customer behavior variables (those which are generated after the loan is approved such as delinquent 2 years, revolving balance, next payment date etc.)

- Understood that the customer behavior variables are not available at the time of loan application, and thus they cannot be used as predictors for credit approval.
- Identified that the rows with loan status with "current" is neither fully paid nor defaulted. So, these rows cannot be used for prediciton
- Came up with various plots to do univariate variate, bivariate analysis, binning, grouping the data as needed.

Plotting and insights

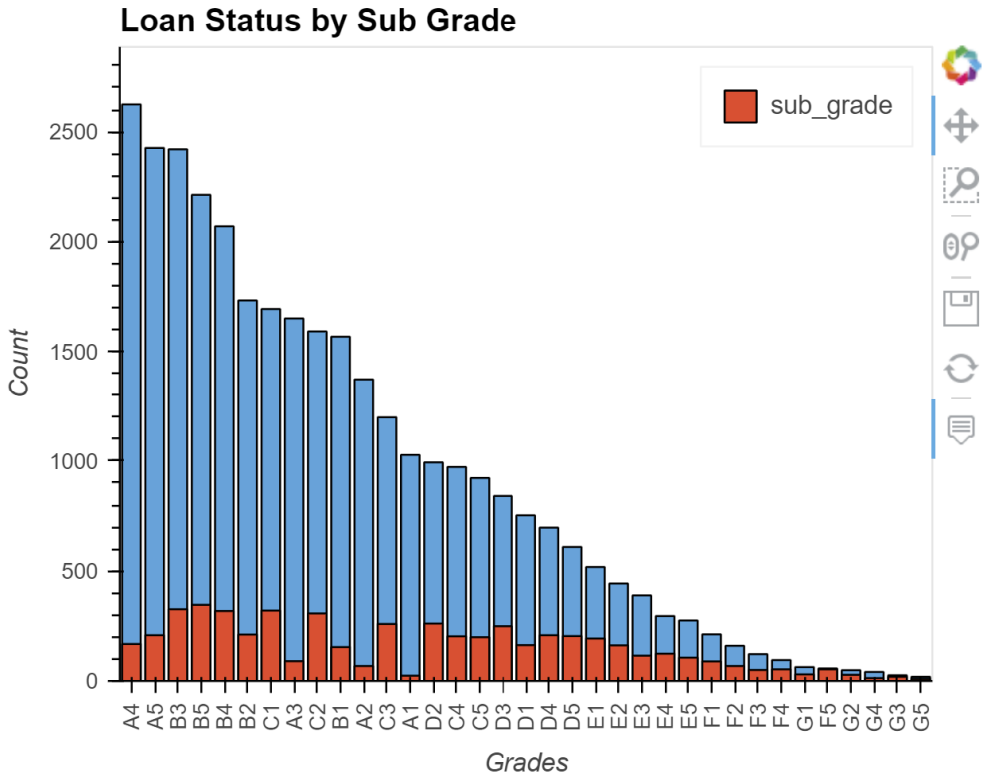


From above bar plot, observe that around 14% of the loan applicants are defaulted on the loan.

Plotting and insights – continued…

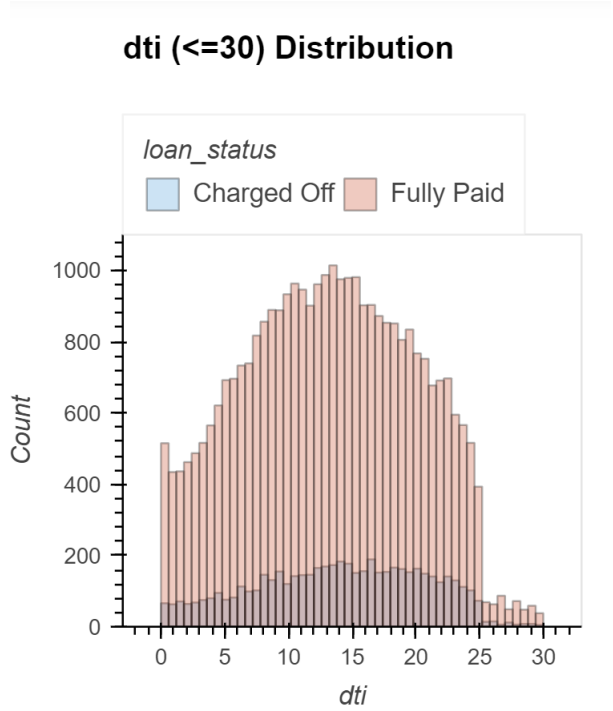
From above plot, observe that high correlation between 'loan_amnt' and 'installment' attributes.

Plotting and insights-continued…



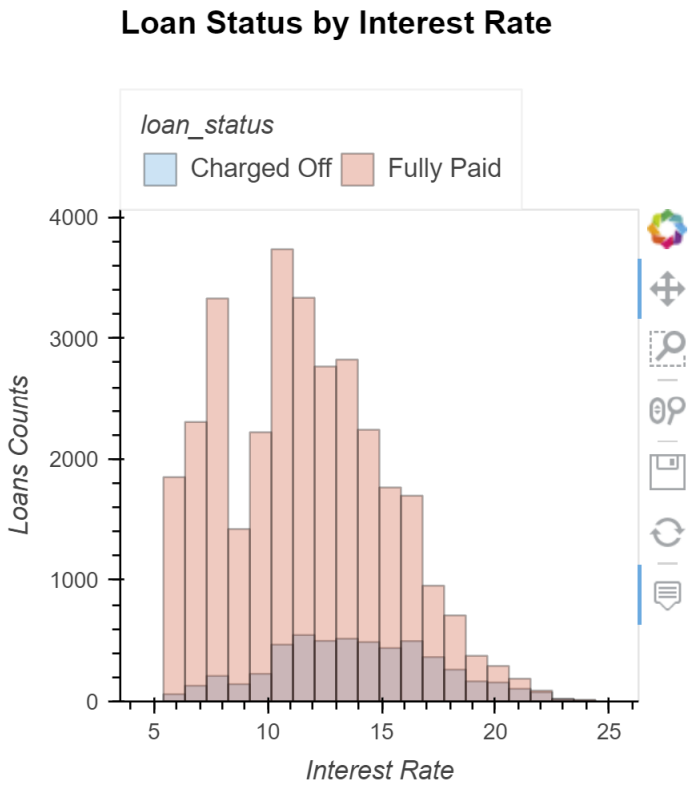
F and G subgrades don't get paid back that often

Plotting and insights-continued…



Observation is that smaller the dti the more likely that the loan will not be paid.

Plotting and insights-continued…



Observation is that loans with high intersest rate are more likely to be defaulted.

Recommendations

- Stop approving loans where income is >30%
- Reduce number of approvals if loan purpose is small business
- Stop approving high value loans when revolving line utilization is > 75%
- Start charging higher interest rate for risk loans i.e. dti > 20

Thank You!