Lending Club Case Study
Group Project

SUBMISSION

Team Members: Rajaram Somanath
Sandeep Dutta
Shaiju Janardhanan
Srinivas Sadagopan

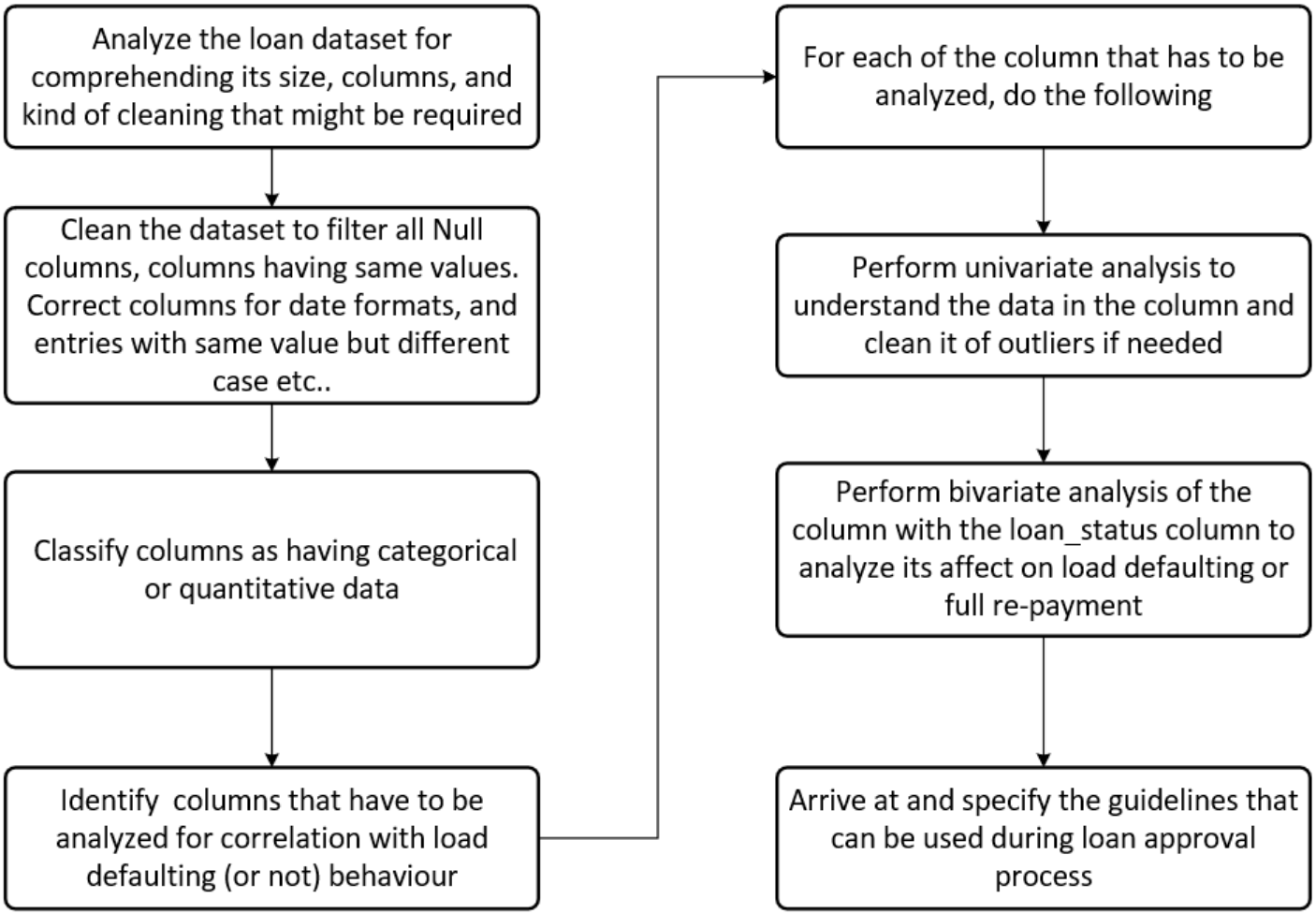
Abstract

- The assignment is to analyse a given data set that has details of loans disbursed through the years 2007 to 2011 and infer the key attributes that are indicative of loan defaulting. The parameters of those attributes is then expected to provide guidelines in deciding to grant or decline loans to prospective borrowers in the future.

- The Guidelines provided must help the bank reject risky loans as well as approve more safer loans

- One dataset with 111 columns and 39717 entries has been provided for analysis. The intent is to apply EDA methodologies and arrive at the guidelines

Problem solving methodology



Data cleaning

- Following data cleaning steps were applied on the data set
  - Filter out columns that have same values.. Or only one value. Example: all '0's / 'NA's/ ‘n’s
  - Filter out columns with more than 50% data
  - Clean up emp_title column to account for same value with different cases and other vagaries
  - Convert columns with date values to pandas datetime format
  - Convert values in column having %age values to float (strip %age symbol)

- After the data cleaning steps, there were 50 columns remaining in the data set.

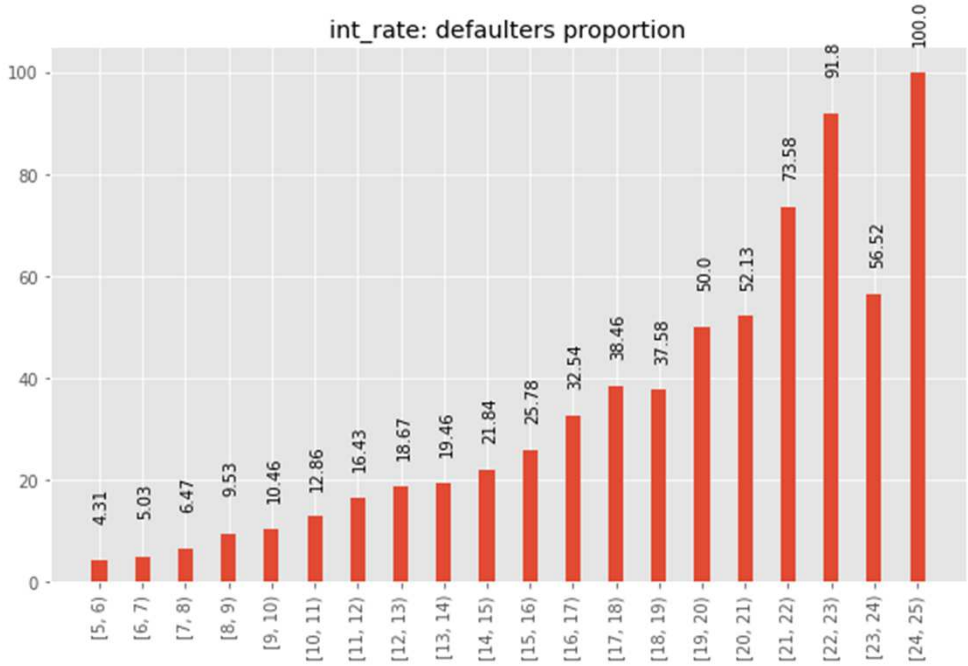
Column selection

- The data set was split into three, based on the loan_status values - "Fully Paid", "Current", and "Charged-Off"
- From the 50 columns after the data cleaning step, following types of columns were filtered out of analysis
  - Categorical columns having tight correlation between the "Fully Paid" and "Charged-Off" data sets
  - Columns like 'id', 'member_id', 'loan_status', 'url' that would very obviously not have a bearing on loan defaulting
  - Also, all the columns describing attributes of active borrower like "collection_recovery_fee" need not be analyzed as those attributes will not be known for a prospective borrower (unless an active customer is asking for additional loan. Current assignment focus is for "new" prospective borrowers
- Following columns are selected for analysis as part of this assignment.

int_rate	*addr_state*	*funded_amnt*
installment	*inq_last_6mths*	*funded_amnt_inv*
grade	*total_acc*	*pub_rec*
sub_grade	*loan_amnt*	*pub_rec_bankruptcies*
verification_status	*purpose*	*revol_util*
zip_code	*emp_length*	
dti	*home_ownership*	
earliest_cr_line	*term*	
annual_income	*open_acc*	

Analysis: int_rate

- "int_rate" is the interest rate on the loan
- Analysis methodology:
  - ➢ Univariate analysis on the int_rate column to determine range and spread of values:
    - ▪ Bin the column entries into 1 percentage point ranges from 5 to 25 for uniformity
  - ➢ Find the number of defaulters and fully paid entries in each of the bins
  - ➢ Plot the proportion of defaulters to fully paid
- From the plot shown it is clear that higher the interest rate higher the risk of loan default

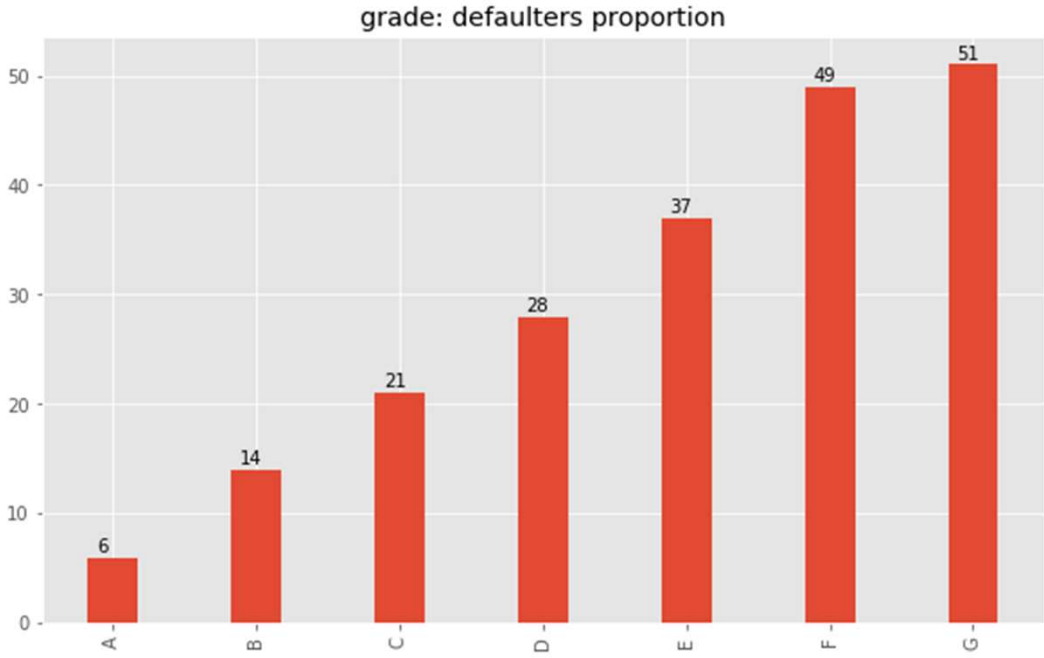


Guideline

Higher the interest rate higher the risk of loan default.
Loans with Interest rates greater than 20% are especially risky
Loans with Interest rate less than 14% are very safe

Analysis: grade

- "grade" is the LC assigned loan grade.
- Analysis methodology:
  - ➢ Univariate analysis on the grade column to determine range and spread of values:
    - Categorical data with 7 categories
  - ➢ Find the number of defaulters and fully paid entries in each of the categories
  - ➢ Plot the proportion of defaulters to fully paid
- From the plot shown it is clear that risk of default consistently increases across the grades from A to G



Guideline

Loans graded A and B have very low risk
Loans graded F and G carry high risk

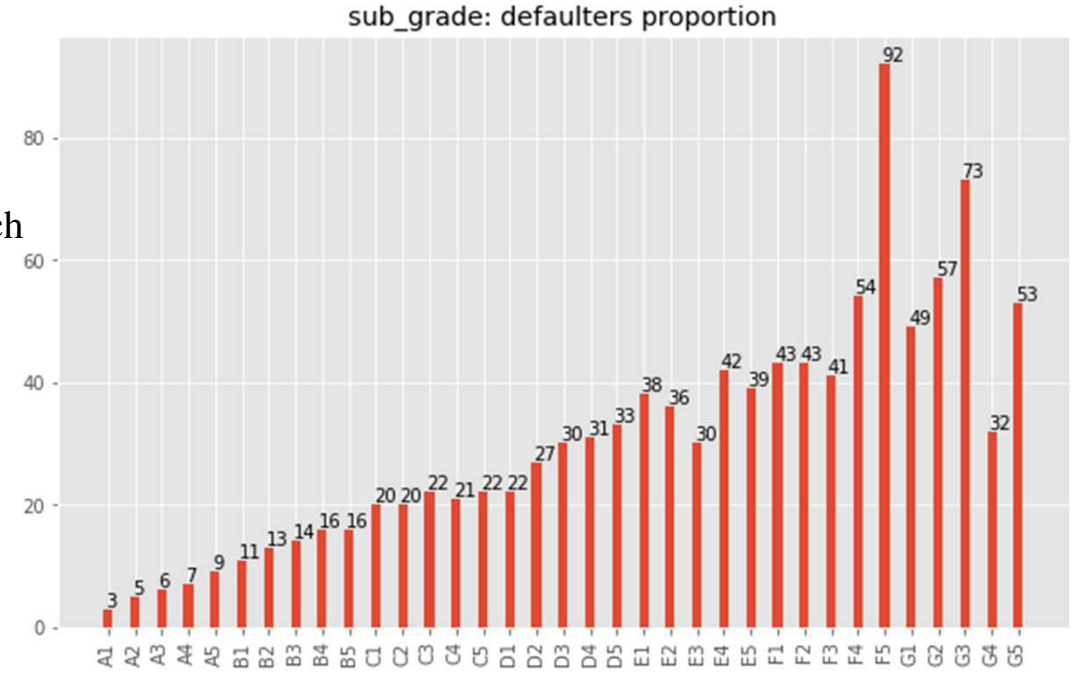
Analysis: sub_grade

- "sub_grade" is the LC assigned loan sub_grade.
- Analysis methodology:
  - ➢ Univariate analysis on the sub_grade column to determine range and spread of values:
    - ▪ Categorical data with 35 categories. Each grade is subdivided into 5 sub grades
  - ➢ Find the number of defaulters and fully paid entries in each of the categories
  - ➢ Plot the proportion of defaulters to fully paid

- From the plot shown it is clear that risk of default follows the trend seen with loan grades.
- Within the grade risk of default generally increases across subgrades 1-5 for grades A, B, C, and D. No such pattern in grades E, F, and G

Guideline

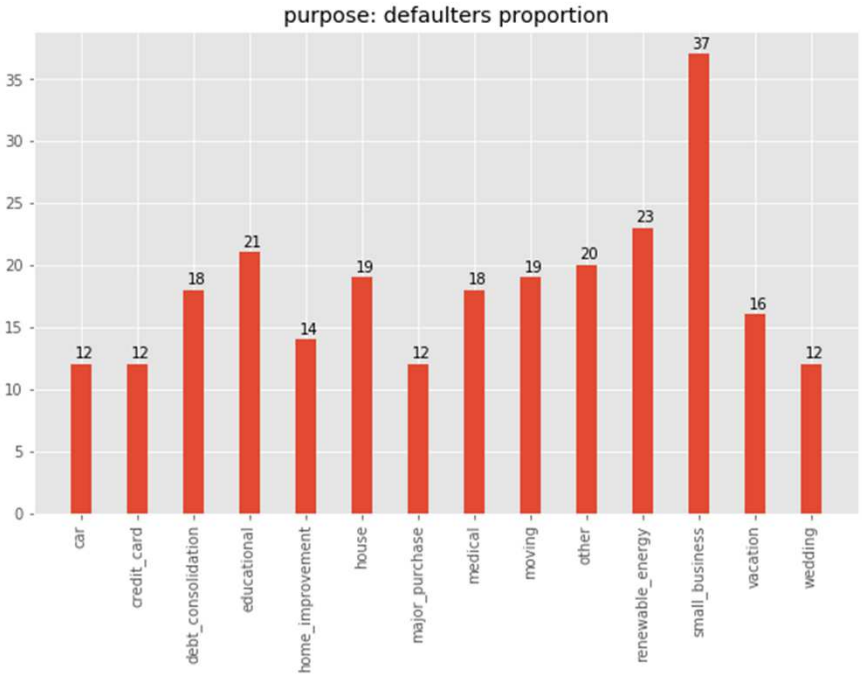
Loans graded at F5 and G3 are especially risky



Analysis: purpose

- "purpose" is a category provided by the borrower for the loan request
- Analysis methodology:
  - ➢ Univariate analysis on the purpose column to determine range and spread of values:
    - Categorical data with 14 categories
  - ➢ Find the number of defaulters and fully paid entries in each of the categories
  - ➢ Plot the proportion of defaulters to fully paid

- From the plot shown it is clear that loans taken under small business category has significantly higher proportion of defaulters



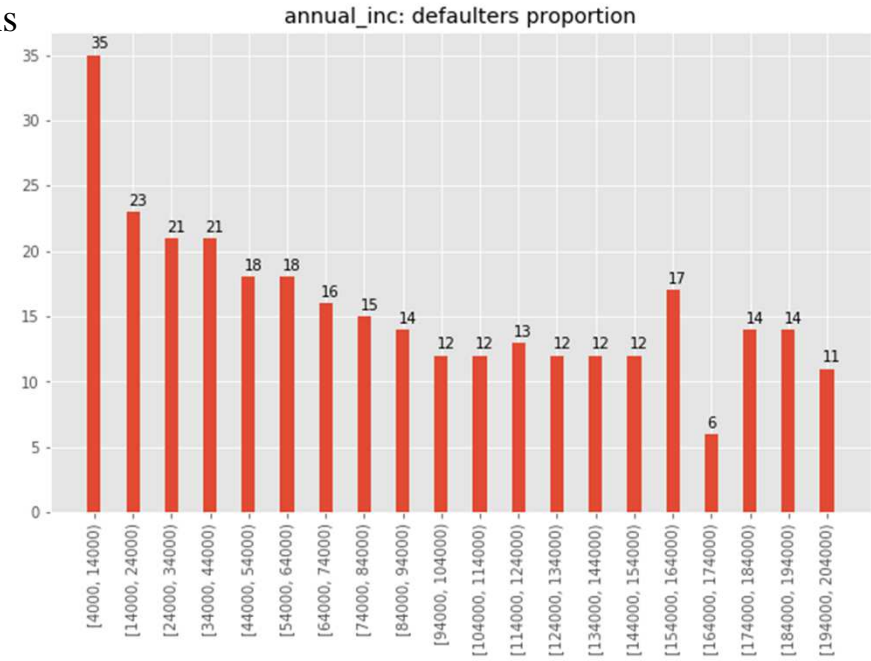
Guideline

Loans for small business purposes carry significantly high risks

Analysis: annual_inc

- "annual_inc" is the self-reported annual income provided by the borrower during registration.
- Analysis methodology:
  - ➢ Univariate analysis on the annual_inc column to determine range and spread of values:
    - Needed outlier value removal at 99th percentile for getting a good subset of values for analysis
    - Bin the column entries into ranges of 10000 from 4000 to 250000 for uniformity
  - ➢ Find the number of defaulters and fully paid entries in each of the bins
  - ➢ Plot the proportion of defaulters to fully paid in each of the bins

- From the plot, it is clear that lower income applicants tend to default more



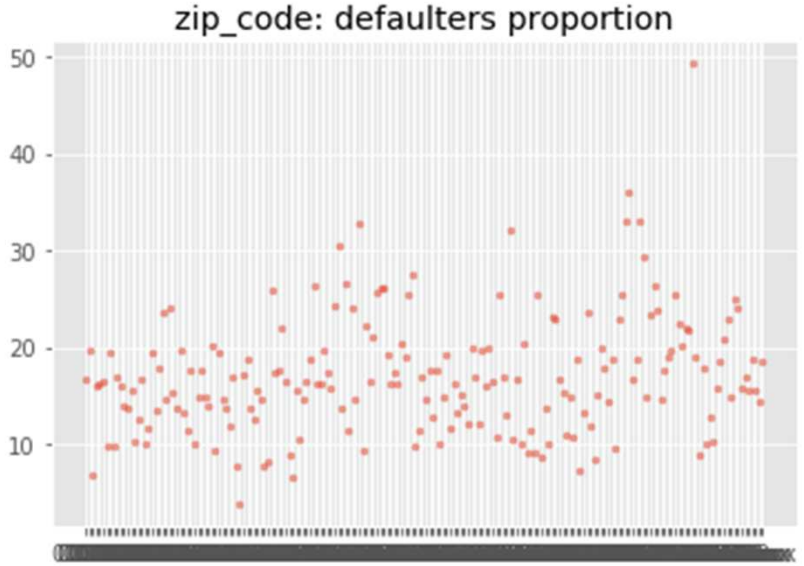
Guideline

Applicants with lower annual income especially in the range of 4000 to 14000 are more likely to default on the loan

Analysis: zip_code

- "zip_code" is the first 3 numbers of the zip code provided by the borrower in the loan application.
- Analysis methodology:
  - ➢ Univariate analysis on the zip_code column to determine range and spread of values:
    - Categorical data with more than 600 zip codes.
  - ➢ Find the number of defaulters and fully paid entries in each of the categories
  - ➢ Plot the proportion of defaulters to fully paid as a scatter plot to see if defaulter proportions in any of the zip codes bucks the trend

- From the plot, if defaulters proportion of 10 to 30 can be considered as an average cluster then, loan by applicants from zip codes below the cluster can be considered safe
- Loan by applicants from zip codes above the normal cluster can be considered riskier. One zip code particularly seems far above the cluster – 935XX and could be subject of further analysis

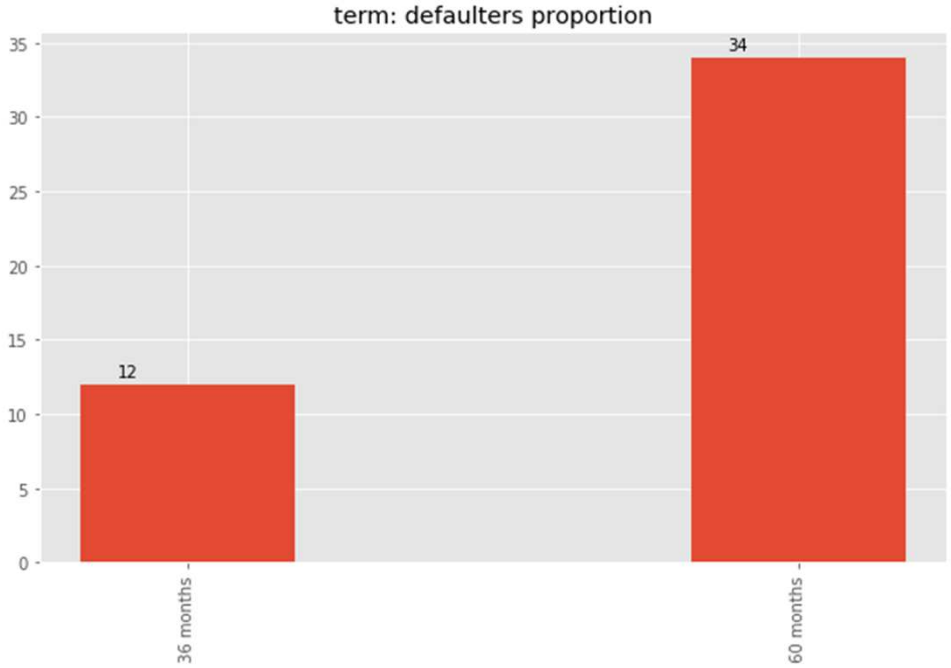


Guideline

Loan applications from certain zip codes can be classified risky based on further research

Analysis: term

- "term" is the number of payments on the loan. Values are in months.
- Analysis methodology:
  - ➢ Univariate analysis on the term column to determine range and spread of values:
    - Categorical data with 2 categories
  - ➢ Find the number of defaulters and fully paid entries in each of the categories
  - ➢ Plot the proportion of defaulters to fully paid
- From the plot shown it is clear that risk of default is higher for the 60 months tenure

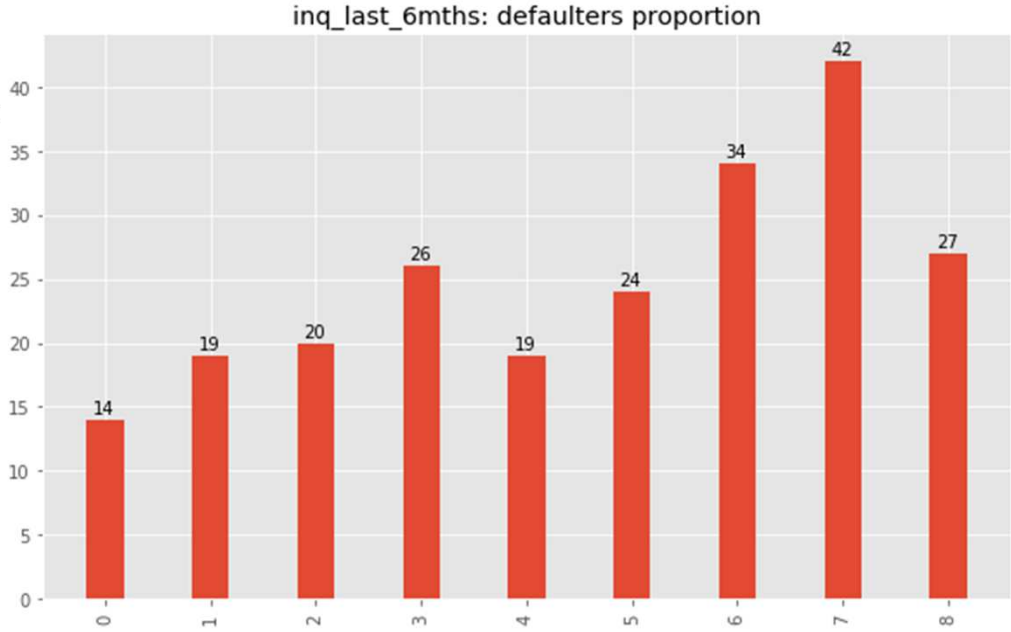


Guideline

Loans having higher repayment tenure are riskier

Analysis: inq_last_6mths

- "inq_last_6mths" is the number of inquiries in past 6 months (excluding auto and mortgage inquiries)
- Analysis methodology:
  - ➢ Univariate analysis on the inq_last_6mths column to determine range and spread of values:
    - Takes a value between 0 and 8
  - ➢ Find the number of defaulters and fully paid entries for each value of inquiries
  - ➢ Plot the proportion of defaulters to fully paid
- From the plot shown it is clear that there is an increasing trend of loan default with increasing numbe of inquiries

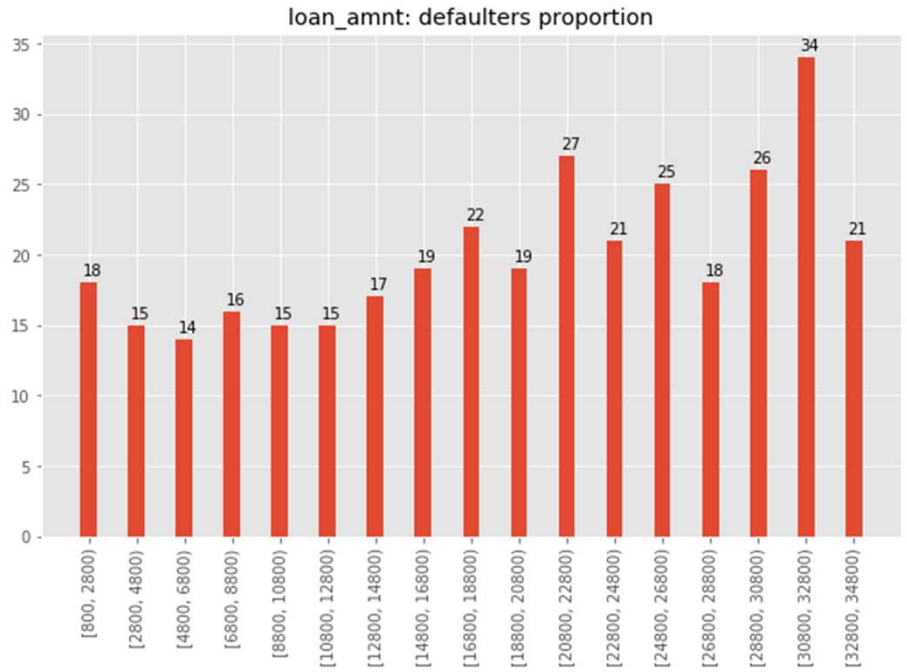


Guideline

Higher number of recent inquiries make a loan riskier

Analysis: loan_amnt

- "loan_amnt" is the listed amount of the loan applied for by the borrower
- Analysis methodology:
  - ➢ Univariate analysis on the loan_amnt column to determine range and spread of values:
    - Bin the column entries into ranges of 2000 from 800 to 36000 for uniformity
  - ➢ Find the number of defaulters and fully paid entries for each of the bins
  - ➢ Plot the proportion of defaulters to fully paid



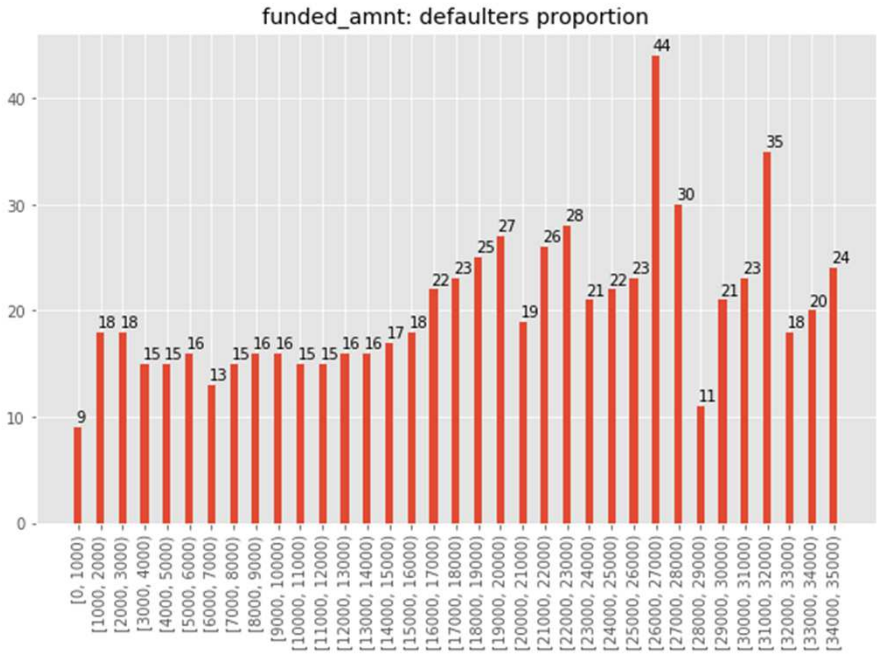
- From the plot shown it appears that there is an upward trend to default loan as the loan amount increases. However this does not appear to be a very clear trend

Guideline

Loans with higher amount are comparatively riskier

Analysis: funded_amnt and funded_amnt_inv

- "funded_amnt" is the total amount committed to that loan at that point in time.
- Analysis methodology:
  - ➢ Univariate analysis on the funded_amnt column to determine range and spread of values:
    - Bin the column entries into ranges of 1000 from 0 to 36000 for uniformity
  - ➢ Find the number of defaulters and fully paid entries for each of the bins
  - ➢ Plot the proportion of defaulters to fully paid



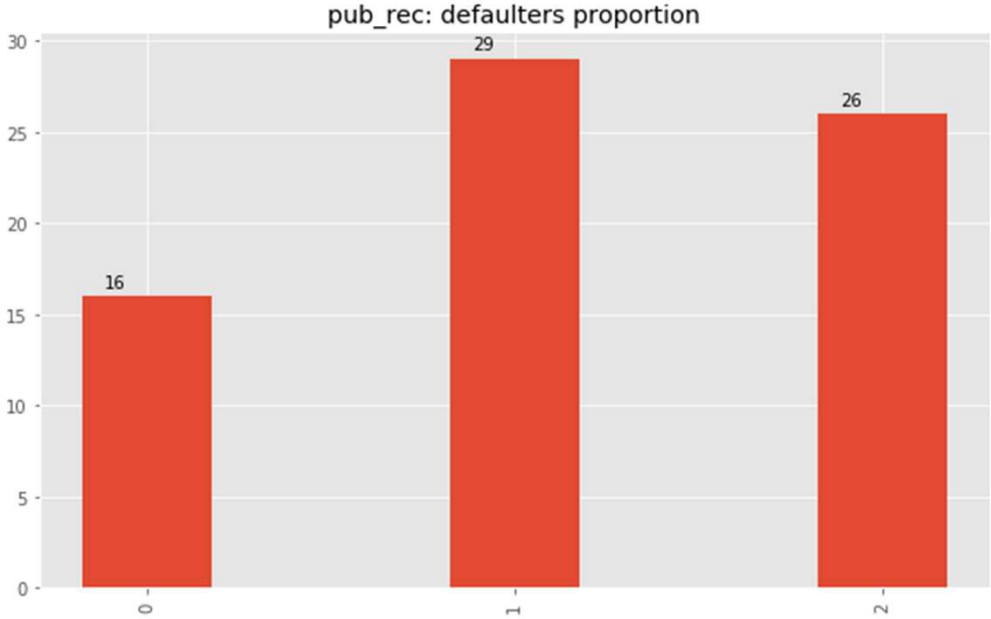
- From the plot shown it can be seen that funded amounts in the right side half have higher proportions of defaulters compared to the left side half
- "funded_amnt_inv" shows the same trends as "funded_amnt"

Guideline

Loans with higher funded_amount tend to default more

Analysis: pub_rec

- "pub_rec" is the number of derogatory public records
- Analysis methodology:
  - ➢ Univariate analysis on the pub_rec column to determine range and spread of values:
    - Takes a value between 0, 1, or 2
  - ➢ Find the number of defaulters and fully paid entries for each value of records
  - ➢ Plot the proportion of defaulters to fully paid
- From the plot shown it is clear that higher number of derogatory record indicates higher is the proportion of loan default

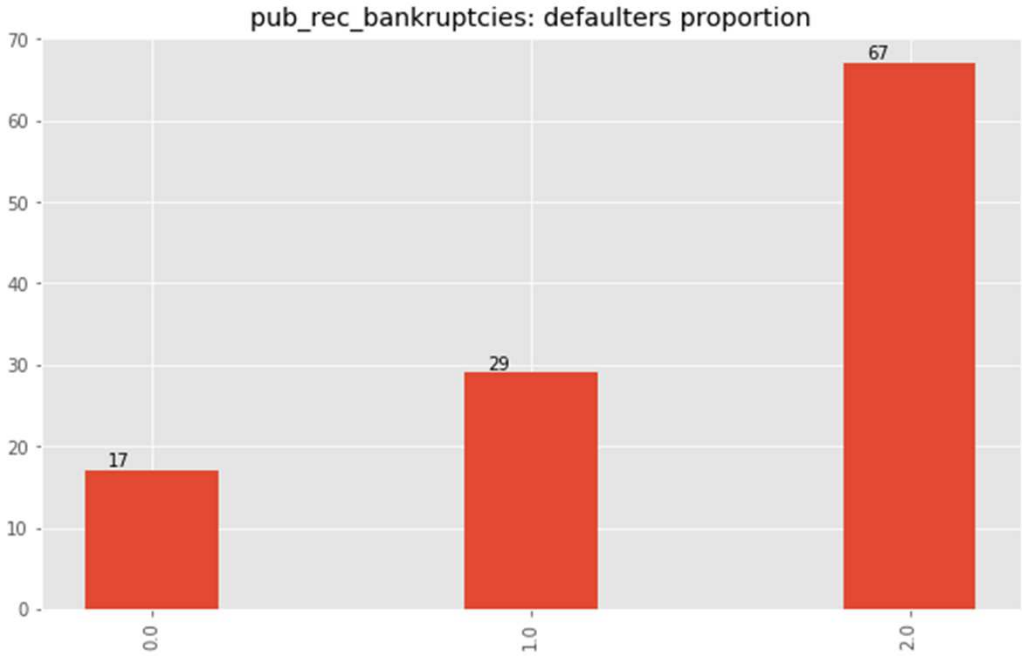


Guideline

Loan to applicants with more number of derogatory record would be riskier

Analysis: pub_rec_bankruptcies

- "pub_rec_bankruptcies" is the number of public record bankruptcies
- Analysis methodology:
  - ➢ Univariate analysis on the pub_rec_bankruptcies column to determine range and spread of values:
    - Takes a value between 0, 1, or 2
  - ➢ Find the number of defaulters and fully paid entries for each value of records
  - ➢ Plot the proportion of defaulters to fully paid
- From the plot shown it is clear that higher number of bankruptcy record indicates higher is the proportion of loan default

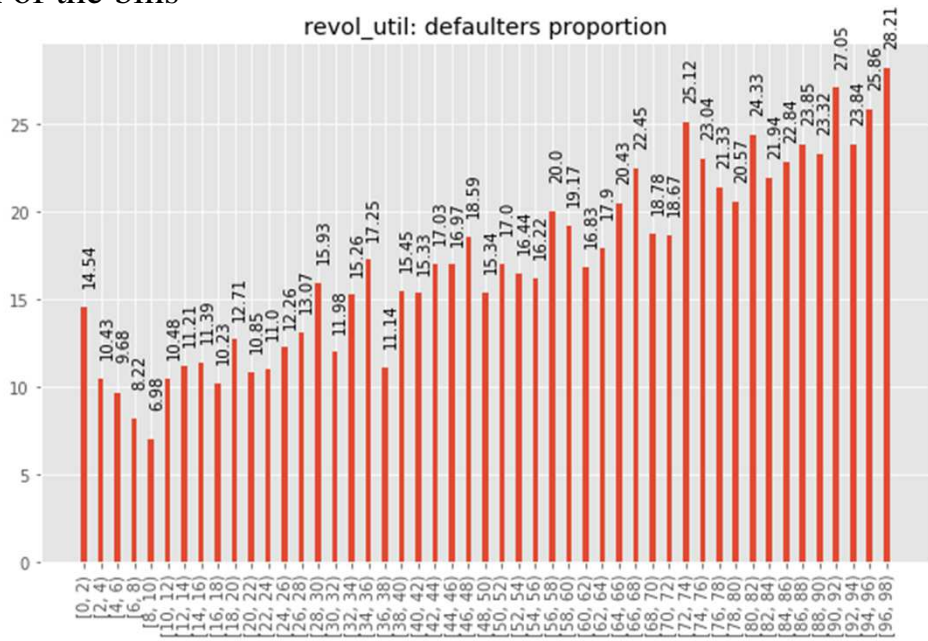


Guideline

Loan to applicants with more number of bankruptcy record would be riskier

Analysis: revol_util

- "revol_util" is the revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
- Analysis methodology:
  - ➢ Univariate analysis on the revol_util column to determine range and spread of values:
    - ▪ Bin the column entries into 2 percentage point ranges from 0 to 100 for uniformity
  - ➢ Find the number of defaulters and fully paid entries in each of the bins
  - ➢ Plot the proportion of defaulters to fully paid
- From the plot shown it is clear that higher the rate of revol_util, higher is the proportion of loan default

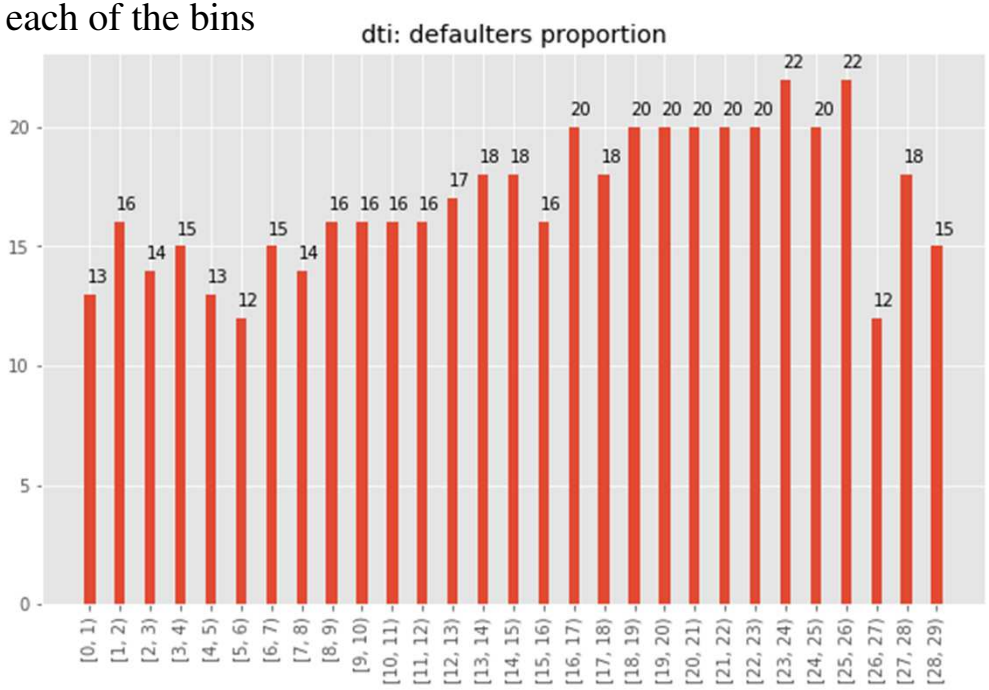


Guideline

A trend of higher loan default is seen with Higher values of revol_util

Analysis: dti

- "dti" is a ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
- Analysis methodology:
  - ➢ Univariate analysis on the dti column to determine range and spread of values:
    - ▪ Bin the column entries into ranges of 1 from 0 to 30 for uniformity
  - ➢ Find the number of defaulters and fully paid entries in each of the bins
  - ➢ Plot the proportion of defaulters to fully paid



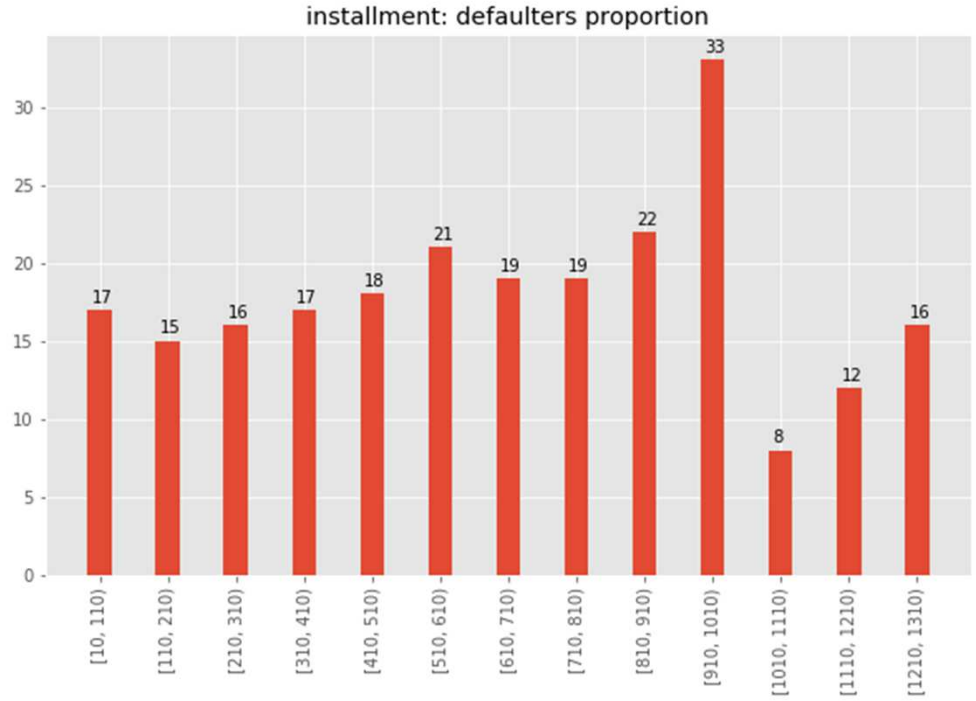
- From the plot, there is an ever so slight increase in defaulter's proportion with higher dti ratios.
- However, the lower defaulters proportions at the highest values of dti ratios would point to very minimal correlation of dti ratios to loan default
- This is little surprising as one would expect higher dti to result in higher defaults – Needs further sudy

Guideline

None based on dti ratios

Analysis: installment

- "installment" is the monthly payment owed by the borrower if the loan originates.
- Analysis methodology:
  - ➢ Univariate analysis on the installment column to determine range and spread of values:
    - Bin the column entries into a range of 100s from 10 to 13350 for uniformity
  - ➢ Find the number of defaulters and fully paid entries in each of the bins
  - ➢ Plot the proportion of defaulters to fully paid



- From the plot shown there is no clear pattern of defaulters proportion based on installment amount
- There is a spike in defaulters proportion for installment amount of 910-1010 that can be investigated further..

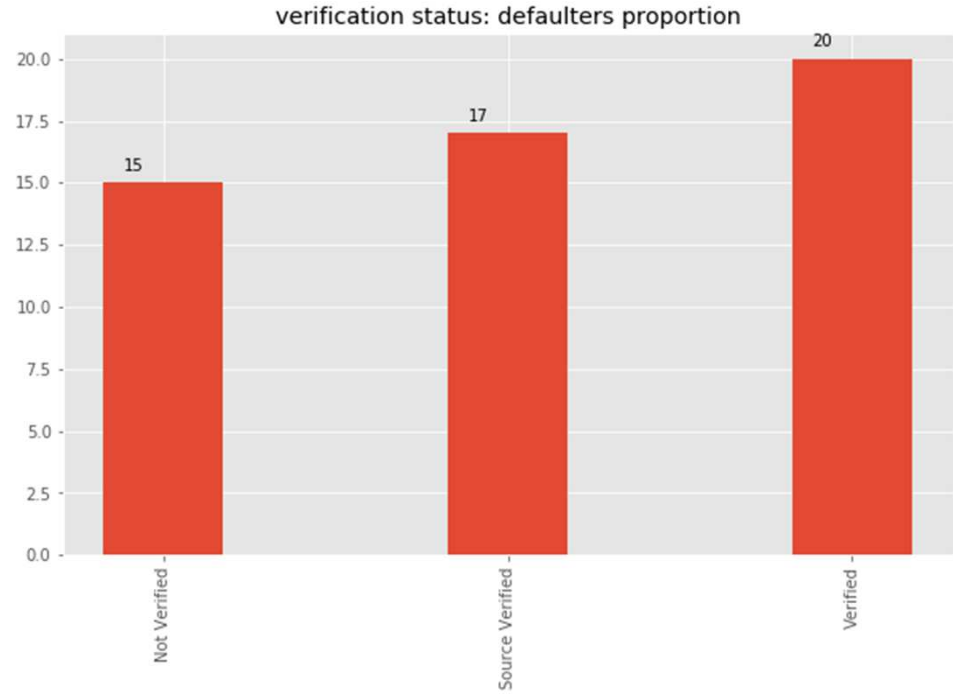
Guideline

None based on the installment amount column

Analysis: verification_status

- "verification_status" indicates if income was verified by LC, not verified, or if the income source was verified.
- Analysis methodology:
  - ➢ Univariate analysis on the verification_status column to determine range and spread of values:
    - Categorical data with 3 categories
  - ➢ Find the number of defaulters and fully paid entries in each of the categories
  - ➢ Plot the proportion of defaulters to fully paid

- Even though there is a slight increasing trend in defaulters proportion across categories, it is not significant
- Interestingly verified loans are showing higher tendency of defaulting. Something to study further closely



Guideline

None based on verification_status column

Column analysis note

- Following columns that were analyzed but did not show a trend to loan default behavior have not been shown in this presentation

earliest_cr_line
total_acc
open_acc
emp_length
addr_state
home_ownership

Conclusions

- The loans data set has been analysed. Inferences and guidelines from the analysis have been presented
- Further analysis on the data set is possible:
  - Study patterns in descriptive text based columns like ‘desc’ for patterns correlation to loan default behavior
  - Some observations in the preceding slides have been marked for further study with domain experts