Lending Club Case Study

- Saravanan Mathivanan

Assumptions

- Since the loan data set includes variables which are customer behaviour variables hence they are removed
- Also only primary variables which are essential for the analysis are taken and split accordingly

Unordered Categorical Variables:

- member_id
- emp_title
- home_ownership
- verification_status
- issue_d
- loan_status
- purpose

Ordered Categorical Variables:

- grade
- sub_grade
- loan_term
-

Assumptions..

Continuous Variables:

- loan_amnt
- funded_amnt
- funded_amnt_inv
- int_rate
- installment
- emp_length
- annual_inc
- dti

Data Clean-up

2. Checklist for Fixing Rows I) Delete summary rows - None

- II) Delete incorrect rows - None
- III) Delete extra rows – There are no blank rows in the given dataset

3. Checklist for Fixing Columns I) Merge columns for creating unique identifiers if needed - NA

- II) Split columns for more data - NA
- III) Add column names - All column names are unique and its relevant info is available in Data Dict.
- IV) Rename columns consistently - Please refer to point # 1 on the column/variable name which is not aligning with Data Dict.
- V) Delete columns - All column names are valid and unnecessary columns which are not applicable for the analysis would be removed as part of Data cleanup.
- VI) Align misaligned columns - NA

4. Standardise Numbers I) Non-standard units - None

- II) Values with varying Scales - None
- III) Over-precision a) 'funded_amnt_inv' - some of the amount is available as Float and that can be rounded off to whole no/INT b) 'annual_inc' - some of the values are available as Float and can be rounded off to whole no/INT c) 'total_pymnt' - Though most of the values are available as Float but it can be rounded off to 2 decimal places
- IV) Outliers a) 'annual_inc' - This Variable has user income ranging from 4000 to 6000000 hence this can be figured out through box plot

Data Clean-up

4. Standardise Text I) Extra characters. Standardise Text - None

- II) Different cases of same words a) 'emp_title' variable can be standardised from All uppercase to Title Case but currently this variable is not going to be used for analysis so no changes done.
- III) Non-standard formats a) All Date fields are given in standard format as 'mm-dd-yyyy' b) 'term' is also mentoined in months so no changes

5. Variables classification

- Categorical Variables a) grade b) sub_grade c) home_ownership d) verification_status e) loan_status f) purpose h) addr_state
- Continous Variables a) loan_amnt, funded_amnt, funded_amnt_inv b) term c) int_rate, installment e) annual_inc f) issue_d, earliest_cr_line g) dti h) emp_length

Univariate Analysis

Driver Variables which are strong indicators of defaulters -

- a. verification_status - Most of the defaulters are either not verified or verified by the LC agency.
- b. purpose - This is a key variable to determine if person would be defaulted after getting the loan as 'debt consolidaiton plays major role here
- c. home_ownership - Most of the Defaulters do either fall under Rent or Mortgage category so this variable is very crucial.
- d. issue_d - Surprisingly there is huge amout of loan been availed during Dec/Holidays
- e. emp_length - There is huge difference between the % of loan purchased by person who has stayed longer than rest.

Bivariate Analysis

Bivariate analysis

Based upon the heatmap and scatterplot for the Continuous variable, below is the finding

a. Positive Correlation -

when the loan amount increases, there is increase in the installment.

When the annual anount/income of the loan applicant is high, the loan amount requested is also high

b. Constant correlation

When the dti varies and scattered evenly across for all the continuous variables

When the emp length increases the amout of loan or salary of employee are evenly distributed so that is constant/zero correlation

Bivariate Analysis

Variables	Q1	Q2	Q3	Q4
loan_amnt	5600.00	10000.00	16500.00	32850.00
funded_amnt	5575.00	10000.00	16000.00	31637.50
funded_amnt_inv	5000.00	9401.00	15000.00	30000.00
int_rate	11.31	13.61	16.40	24.03
installment	168.55	293.87	457.84	891.76
emp_length	2	5	9	10
annual_inc	37000.00	53000.00	75000.00	132000.00
dti	9.05	14.29	19.29	29.85
Variables	Mean	Med	SD	
loan_amnt	12045.71	10000.00	7925.65	
funded_amnt	11674.40	10000.00	7537.65	
funded_amnt_inv	10758.64	9401.00	7357.80	
int_rate	13.82	13.61	3.65	
installment	334.38	293.87	211.48	
emp_length	5	5	3	
annual_inc	59203.46	53000.00	29627.20	
dti	14.00	14.29	6.58	

Recommendation to LC

Based upon Analysis done with the customer variables and past loan data history

1. It is likely that loan applicant comes with purpose of ‘debt consolidation’ then the applicant details should be scrutinized since the % of this same purpose had very bad history on not paying the loan amount
2. Also while checking the past history of the customer, it is likely to check if the customer is still on ‘Rent’ or ‘Mortgage’ on house property is still running. There are high chances of customer not paying the debts
3. Also the % of loan application received during Dec is more likely to be spent on Vacation or holidays so Lending club needs to investigate the same and avoid recommending application during that period.

Thank you