Lending Club Case Study

Contributors :
Monika Siluveru, SAP
Ramya Devarajan, SAP

Month 07, 2023

Introduction and Goal of Analysis

Lending loans to ‘risky’ applicants is the largest source of financial loss(called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed.

The main objective is to be able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

Perform an analysis to understand the driving factors (or driver variables)behind loan default, i.e.the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Process Steps

Step 1: Understanding the Data

Step 2: Data Cleaning

Step 3: Univariate Analysis

Step 4: Bivariate/Multivariate Analysis

Step 5: Results

Step 1 : Understanding the Data

Check shape of the dataset

Have 39717 rows and 111 columns

Check datatype of the dataset columns

Have int64, float64, object types

Get stats for numeric data

Insights : Understanding the Data

1. There are 39717 rows and 111 columns in dataset

2. There are lot of columns with high percentage of missing data or null values

3. There are instances where the loan was rejected by the investor even after the lending club agent had approved

4. The max DTI that was accepted is ~30

5. There is a lot of distribution in the data for loan amount, funded amount, funded amount invested, dti as the mean and median values shows a difference

Step 2 : Data Cleaning

- Drop Duplicate Data
- Check for Null Values
- Remove columns which are related to customer behaviors
- Removing the single valued columns
- Removing columns with unique values
- Missing Values Check
- Standardizing the data

Data Cleaning

Drop Duplicate Data

No duplicate data found

Check for Null Values

There are lot of columns with null values. Removed all those columns.

Removing the Columns related to Customer Behavior

delinq_2yrs
earliest_cr_line
inq_last_6mths
open_acc
pub_rec
revol_bal
revol_util
total_acc
out_prncp
out_prncp_inv
total_pymnt
total_pymnt_inv
total_rec_prncp
total_rec_int
total_rec_late_fee
Recoveries
collection_recovery_fee
last_pymnt_d
last_pymnt_amnt
next_pymnt_d
last_credit_pull_d
application_type
installment

Removing the single valued columns

pymnt_plan

initial_list_status

collections_12_mths_ex_med

policy_code

acc_now_delinq

chargeoff_within_12_mths

tax_liens

delinq_amnt

Removing columns with unique values

There are several columns which have only unique values

They would be able to contribute to the analysis in any which way. Hence removing them.

id

member_id

url

We are left with 39717 rows and 23 columns after this.

Missing Values Check

We have around 23 columns now check if there is a huge percentage of missing values and remove the columns

Drop Duplicate data

No duplicates found

Remove columns which has null values with a threshold of 50% as this does not contribute to the analysis

Remove rows which has high percentage data missing

Remove rows with null values

After this we have 39717 rows and 21 columns

Missing Values Check

For the missing value of the below variables below imputation metrics can be considered

- emp_title 2459 -> mode can be used for imputation on missing values
- emp_length 1075 -> mode can be used for imputation on missing values
- desc 12940 -> mode can be used for imputation on missing values
- title 11 -> mode can be used for imputation on missing values

Removing the records with current loan status as the current loan status cannot talk about or give inferences about defaulters. It is to be considered only for either fully paid or charged off loans

After this we have around 38577 rows and 21 columns

Insights from Data Cleaning

Removing customer behavior variables as these are not relevant when an applicant applies for loan

Removing variables which just has only one value as this will not contribute to our analysis

Removing variables which have only unique values as this will not contribute to our analysis

Removing rows and columns which has missing percentage value of 50%

Imputation methods that can be used for missing values , here the important variable seems to be emp_length for analysis

- **emp_title 2459 -> mode can be used for imputation on missing values**
- **emp_length 1075 -> mode can be used**
- **desc 12940 -> mode can be used for imputation on missing values**
- **title 11 -> mode can be used for imputation on missing values**

Remove rows which contain ‘current’ loan status as this will not be helpful for our analysis

Remove suffix/prefix from int_rate, terms and emp_length and retaining numerical values for analysis

Step 3 : Data Analysis steps

Converted loan status to numeric , 0 as fully paid and 1 as charged off as this is a ordered category

Encoded grade to numeric data as this is a ordered category

Checked for Correlation

Insights for above steps :

- Here there is positive correlation between loan_amt, funded_amnt, funded_amnt_inv so we could use only one for analysis.
- Here there is a positive correlation between interest rate and grade.

Step 3 : Data Analysis : Univariate analysis on charged off cases

Here there are several categorical variables like term,grade,subgrade,emp_length,home_ownership,verification_status,loan_status,purpose,addr_state,pub_rec_bankruptcies

Performed segmented univariate analysis based on loan status

Step 3 : Data Analysis : Univariate analysis Observations

The above analysis with respect to the charged off loans for each variable suggests the following.

There is a more probability of defaulting when :

- Applicants having house_ownership as 'RENT'
- Applicants who use the loan to clear other debts and Grade is 'B' And a total grade of 'B5' level.
- Applicants who receive interest at the rate of 13-17%
- Applicants who have an income of range 31201 - 58402
- Applicants with employment length of 10
- When funded amount by investor is between 5000-10000
- DTI is between 10-20
- Term of 36 months
- When the loan status is Not verified
- when from CA,FL address state
- Applied in the month of Dec and year 2011 had max defaulting and the defaulting increases with year on year.

Step 3 : Data Analysis : Bivariate analysis on charged off cases

1. Annual income vs loan purpose

Observation :

Though the number of loans applied and defaulted are the highest in number for "debt_consolation", the annual income of those who applied isn't the highest.

Applicants with higher salary mostly applied loans for "home_improvment", "house", "renewable_energy" and "small_businesses"

Applicants with lower salary applied for educational loan

Step 3 : Data Analysis : Bivariate analysis on charged off cases cont…

2. Annual Income vs funded amount inv

    - Loan amount higher the defaulting is higher

3. Annual income vs int_rate

    - Irrespective of the annual income , the defaulting is higher in higher interest rate group 21-25%

4. Annual Income vs term

    - with higher term the defaulting seems to increase

5. Annual Income vs grade

    - irrespective of the annual income as the grade increases the defauling increases

6. Annual Income vs emp length

    - annual income increases as emp length increases also ppl with higher income tend to default more

Step 3 : Data Analysis : Bivariate analysis on charged off cases cont..

7.funded_amnt_inv_group vs Interest Rate

Interest rate increases with funded_amnt_inv and higher interest more defaulting

8.Funded Amount vs Loan purpose

In the lower segment below 7k defaulting is more for education loan and vacation and in higher amount it is more for small business, debt consolidation and credit card

9.Funded Amount vs House Ownership

Applicants living in mortgage took higher amount and defaulting is also higher in mortgage group

10. Funded amount vs month issued and year issued

Max defaulting that is issued in the December month is for debt consolidation and credit card

Step 3 : Data Analysis : Bivariate analysis on charged off cases cont…

11. Funded amount inv vs Grade

Grade 'F' and 'G' have highest defaulting compared to other groups

12. Emp length vs funded amnt inv

funded amnt inv increases as emp length increases but the max defaulting is with 10 + years. Employees with longer working history got the loan approved for a higher amount.

13. Funded amnt inv vs Verification Status

Looking at the verification status data, verified loan applications tend to have higher loan amount which might indicate that the firms are first verifying the loans with higher values.

14. Purpose vs int_rate

Higher interest group 17%-21%

21%-25% max defaulters are for debt consolidation and small business

Step 3 : Data Analysis : Bivariate analysis on charged off cases cont…

15 . grade vs interest rate

As the grade increases interest rate increases

16. Funded amnt inv group vs int rate

The interest rate for charged off loans is pretty high than that of fully paid loans in all the loan_amount groups. This can be a possible strong driving factor for loan defaulting.

17. Funded amnt inv group vs term

Applicants with higher term tend to default more for higher loan amount which means that applicants applying for long term has applied for more loan.

Step 3 : Data Analysis : Bivariate analysis on charged off cases cont…

18. DTI vs Funded Amount Invested Group

    Applicants with higher DTI tend to default more and the 15k to 20k group have higher DTI

19. Annual income vs funded amnt inv

    As the annual amount increases , funded amount inv also increases

20. Term vs grade

    for charged off cases the loan defaulting increases from grade D to G when the term is higher

Step 3 : Insights from Data Analysis : Bivariate analysis on charged off cases

Bivariate analysis for charged off cases is high

1. When in lower income for education loan and higher income for house improvement, house, small business
2. Higher income group is on mortgage and mortgage tend to default more
3. High loan amount group tend to default more
4. Irrespective of the annual income , the defaulting is higher in higher interest rate group 21-25%
5. With higher term the defaulting seems to increase
6. irrespective of the annual income as the grade increases the defaulting increases
7. Annual income increases as emp length increases also ppl with higher income tend to default more
8. Interest rate increases with funded amount invested and higher interest more defaulting
9. In the lower segment below 7k defaulting is more for education loan and vacation
   and in higher amount it is more for small business, debt consolidation and credit card
10. Applicants living in mortgage took higher amount and defaulting is also higher in mortgage group
11. Irrespective on the amount max defaulting in dec.
12. Max defaulting that is issued in the December month is for debt consolidation and credit card
13. Grade 'F' and 'G' have highest defaulting
14. funded amnt inv increases as emp length increases but the max defaulting is with 10+ years
15. verified status gets higher loan and higher loan defaults since the amount is higher.
16. Higher interest group 17%-21% 21%-25% max defaulters are for debt consolidation and small business
15. As the grade increases from A to G interest rate increases and max defaulting in F and G
    21-25% group have the highest defaulting in grade F and G

Step 3 : Data Analysis : Bivariate analysis on charged off cases

Bivariate analysis for charged off cases is high

16. As the funded amount increases , the interest rate also increases and in the same funded amount inv group defaulting is higher when the interest rate is higher .

17. Applicants with higher term tend to default more for higher loan amount which means that applicants applying for long term has applied for more loan amount.

18. Applicants with higher DTI close to 16 tend to default more and the funded amount inv group of 15k to 20k group have higher DTI

19. As the annual income amount increases , funded amount inv also increases

20. For charged off cases the interest rate for 60 months duration is higher.¶

21. For charged off cases the loan defaulting increases from grade D to G when the term is higher

Step 3 : Insights from Data Analysis

Multivariate analysis from pair plot:

1. Funded_amnt_inv V/s interest rate: Defaulting more as the interest rate increases irrespective of the funded amount inv

2.Funded_amnt_inv v/s term : Defaulting seems to be more in higher term group

3.Funded_amnt_inv v/s dti: defaulting higher as funded amount invested increases irrespective of the DTI

4. Interest rate: defaulting higher in higher interest rate and with a bit of lesser income.

Step 3 : Summary Data Analysis : Bivariate analysis on charged off cases

Observations

The above analysis with respect to the charged off loans. There is a more probability of defaulting when :

- Members taking loan for 'home improvement' and have income of 60k -70k

- Members whose home ownership is 'MORTGAGE and have income of 60-70k

- Members who receive interest at the rate of 21-24% and have an income of 70k-80k

- Members who have taken a loan in the range 30k - 35k and are charged interest rate of 15-17.5 %

- Members who have taken a loan for small business and the loan amount is greater than 14k

- Members whose home ownership is 'MORTGAGE and have loan of 14-16k

- grade is F and loan amount is between 15k-20k

- employment length is 10yrs and loan amount is 12k-14k

- the loan is verified and loan amount is above 16k

- For grade G and interest rate above 20%

Thank you.

Contact information:

Monika Siluveru, SAP
Ramya Devarajan, SAP