Lending Club Case Study

Overview

Data Understanding and Quality check

Data Cleaning

Data Analysis

- ➢ Uni Variate analysis results
- ➢ Bi Variate analysis results

Recommendations based on analysis

Summary

Overview

Loan Dataset has below customers:

1. Full Paid -> Completed Loan
2. Current -> Loan In Progress
3. Charged off -> Defaulted loan

In the above list we will only consider “Charged Off” customers for this analysis as the interest is mainly on defaulted loan attributes and customer attributes

Expected Analysis Outcomes

Factors influencing loan default means as part of this analysis identify variables which are strong indicators of default loan

Data Understanding and Quality check

- Checking the loan dataset shape shows 39717 rows and 111 columns

  ```
  <class 'pandas.core.frame.DataFrame'> RangeIndex: 39717 entries, 0 to 39716 Columns: 111 entries, id to
  total_il_high_credit_limit dtypes: float64(74), int64(13), object(24) memory usage: 33.6+ MB
  ```

- Checking for null values shows there are many columns with more null/na/NaN values

  ```
  id 0 member_id 0 loan_amnt 0 funded_amnt 0 funded_amnt_inv 0 ... tax_liens 39 tot_hi_cred_lim 39717
  total_bal_ex_mort 39717 total_bc_limit 39717 total_il_high_credit_limit 39717 Length: 111, dtype: int64
  ```

- Checking for these loan attributes in given Data Dictionary shows these are not very much important attributes needed for our analysis and so cleaning these will be better

DATA CLEANING

Step1:

Dropping all rows with missing values – Couldn’t find any rows with complete missing values and so nothing to be dropped

Step2:

Dropping columns with at least 1 or max null values -> after this step the loan dataset shape becomes 39717 rows and 43 columns

Step3:

Making sure no null values present after these cleaning

Step4:

Listing the columns remaining in the loan dataset after cleaning

```
Index(['id', 'member_id', 'loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'term', 'int_rate', 'installment', 'grade', 'sub_grade',
'home_ownership', 'annual_inc', 'verification_status', 'issue_d', 'loan_status', 'pymnt_plan', 'url', 'purpose', 'zip_code',
'addr_state', 'dti', 'delinq_2yrs', 'earliest_cr_line', 'inq_last_6mths', 'open_acc', 'pub_rec', 'revol_bal', 'total_acc',
'initial_list_status', '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_amnt', 'policy_code', 'application_type',
'acc_now_delinq', 'delinq_amnt'], dtype='object’)
```

Step5:

Filtering for only defaulted loan customers (with loan_status as “Charged Off”) resulted in dataset shape of 5627 rows and 43 columns

DATA ANALYSIS – UNI VARIATE ANALYSIS

Considering below 8 variables for univariate analysis from this defaulted loan dataset

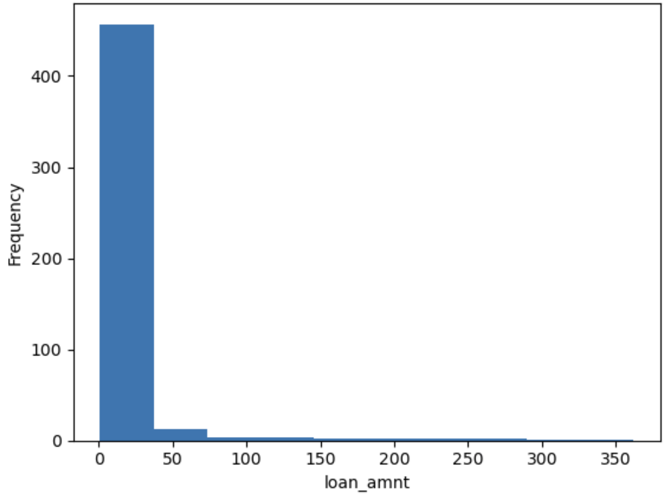
1. loan_amnt
2. term
3. int_rate
4. grade
5. home_ownership
6. annual_inc
7. verfification_status
8. purpose

Variable 1: Loan_amnt

Analysis Outcome:

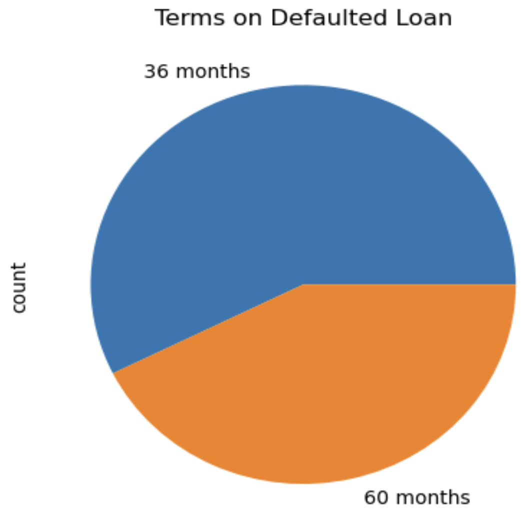
Defaulted loan customers took

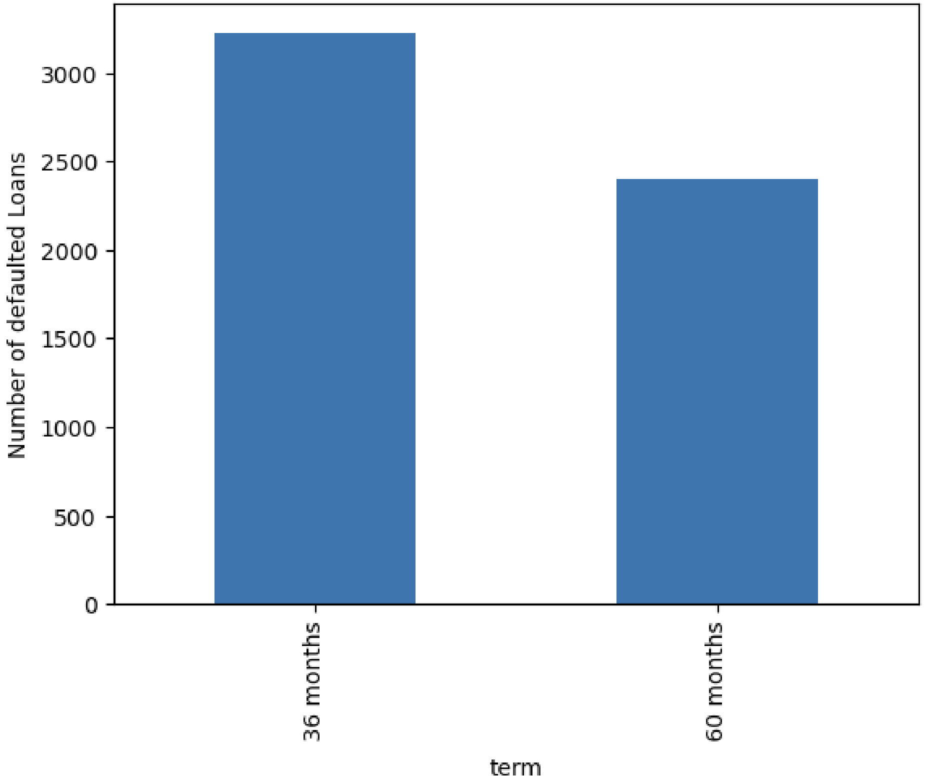
Loan amount < 50,000rs



DATA ANALYSIS – UNI VARIATE ANALYSIS

Variable 2: Term



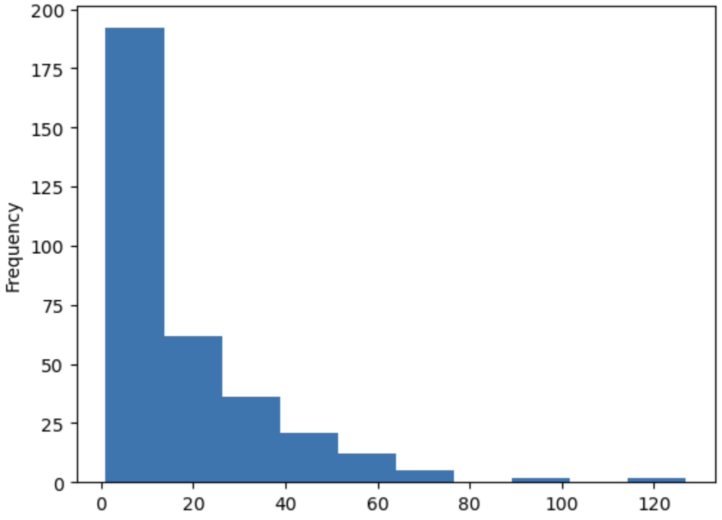


Analysis Outcome:

Loans with tenure of 36 months defaulted more than the loans with tenure of 60 months might be due to HIGH EMI per month due to less tenure

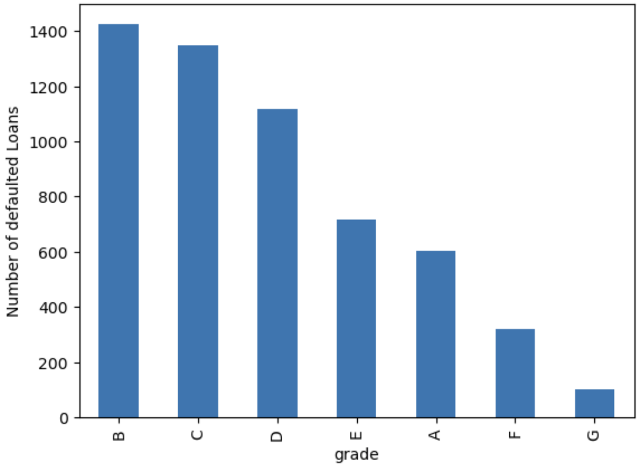
DATA ANALYSIS – UNI VARIATE ANALYSIS

Variable 3: int_rate



Analysis Outcome:
Interest rates for most of the defaulted loan customers are less than 20%

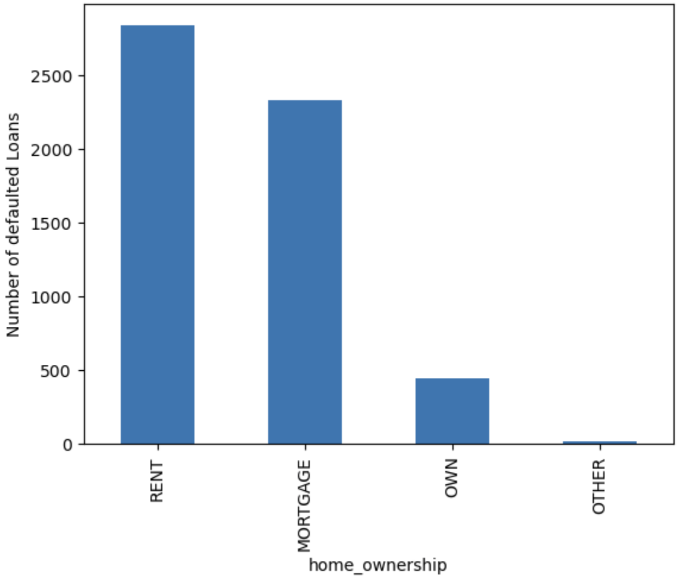
Variable 4: grade



Analysis Outcome:
Grades - B, C, D are the top 3 loan grades which got defaulted

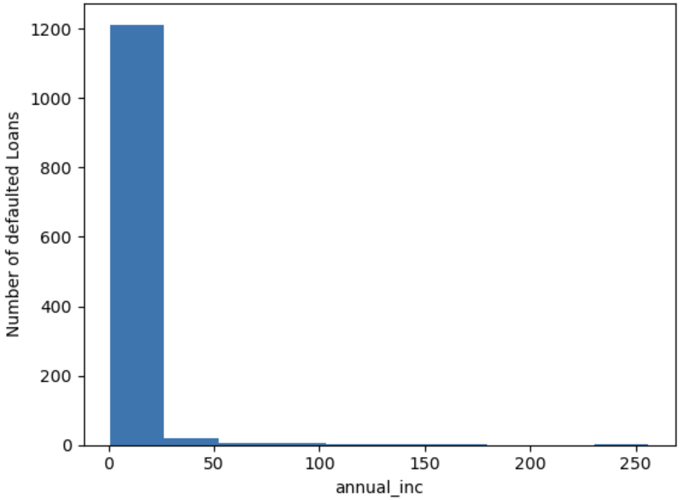
DATA ANALYSIS – UNI VARIATE ANALYSIS

Variable 5: home_ownership



Analysis Outcome:
Customers who are either in RENTED or MORTGAGED Homes defaulted loans more times

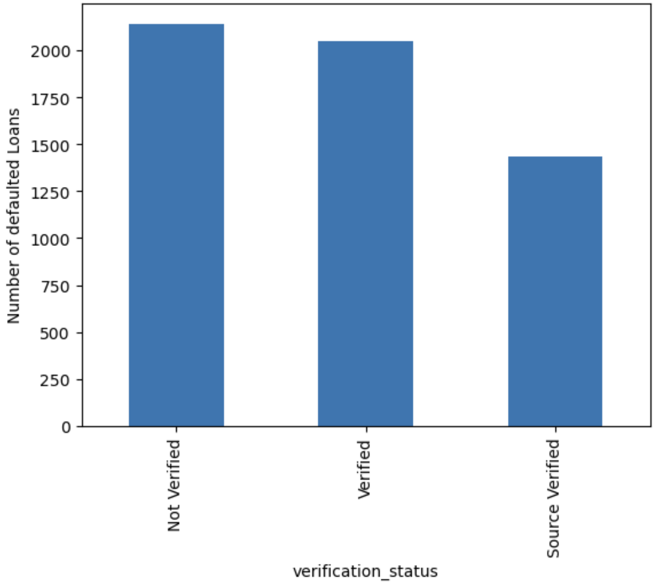
Variable 6: annual_inc



Analysis Outcome:
Customers who defaulted loans have annual income < 50k

DATA ANALYSIS – UNI VARIATE ANALYSIS

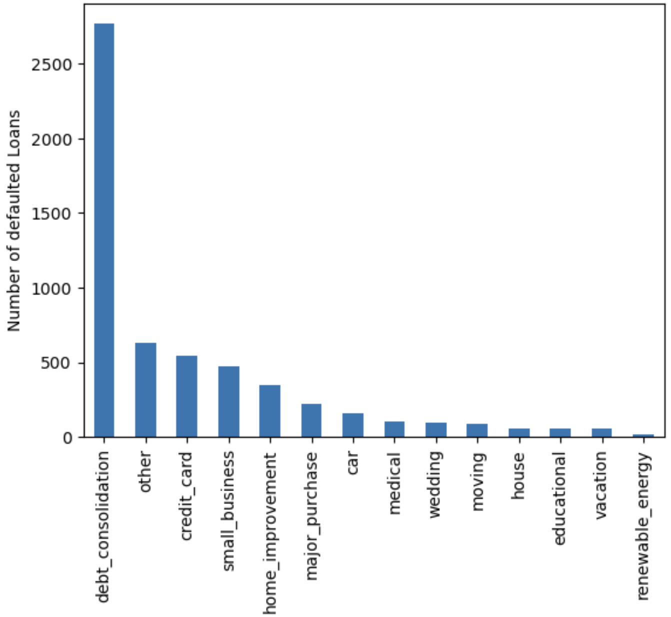
Variable 7: Verification Status



Analysis Outcome:

Sum of Verified and Source Verified > Not Verified which means more verified loans defaulted than not verified. This looks weird and need to find the root cause for the same, it could be lapse in the process of verification

Variable 8: Purpose

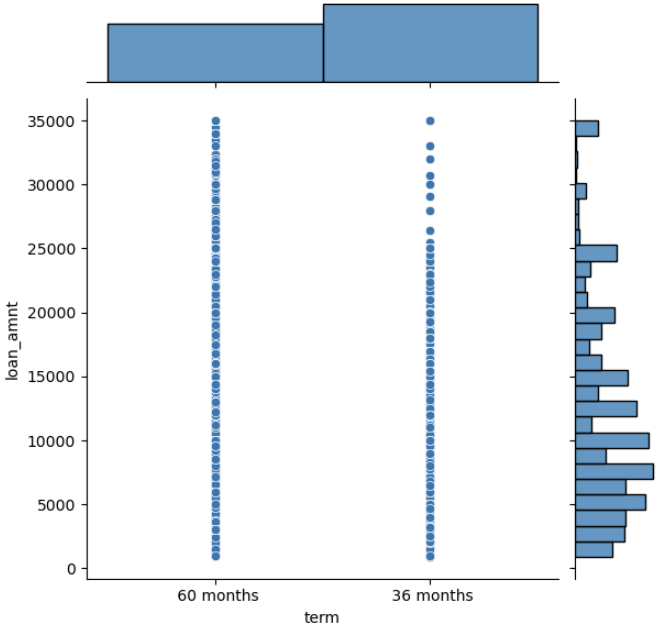


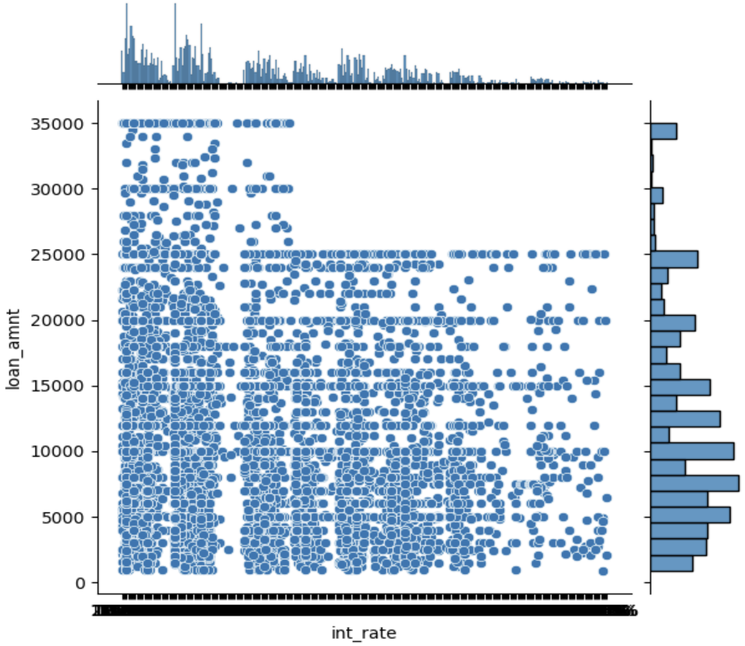
Analysis Outcome:

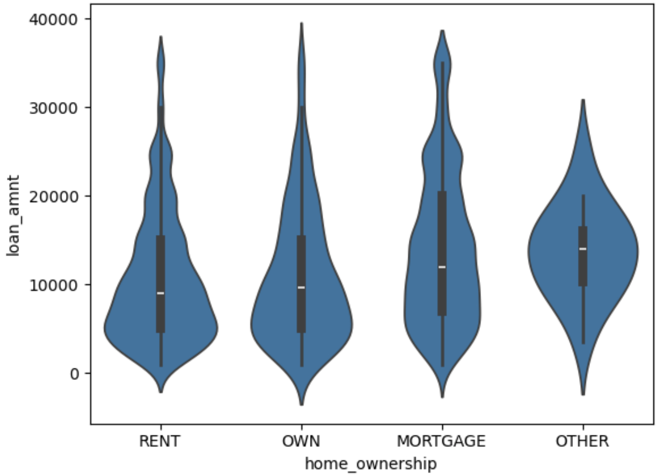
Customer who took loan for the purpose of debt consolidation defaulted more than any other purpose

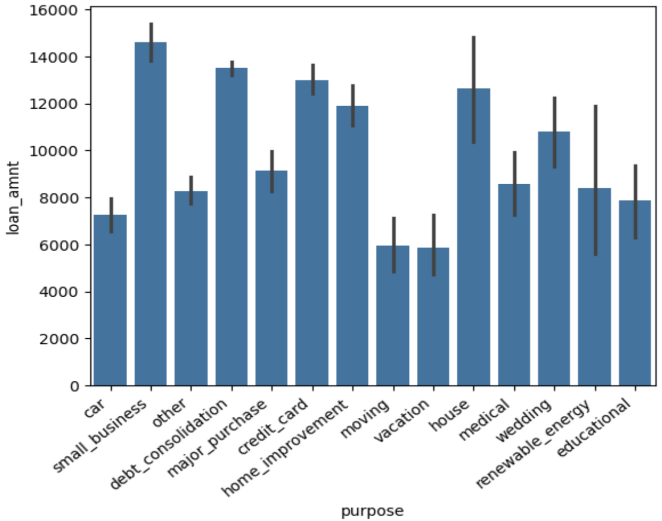
DATA ANALYSIS – BI VARIATE ANALYSIS

Comparison of Loan amnt with different other variables like term, home_ownership, int_rate, purpose shows some interesting information as shown below





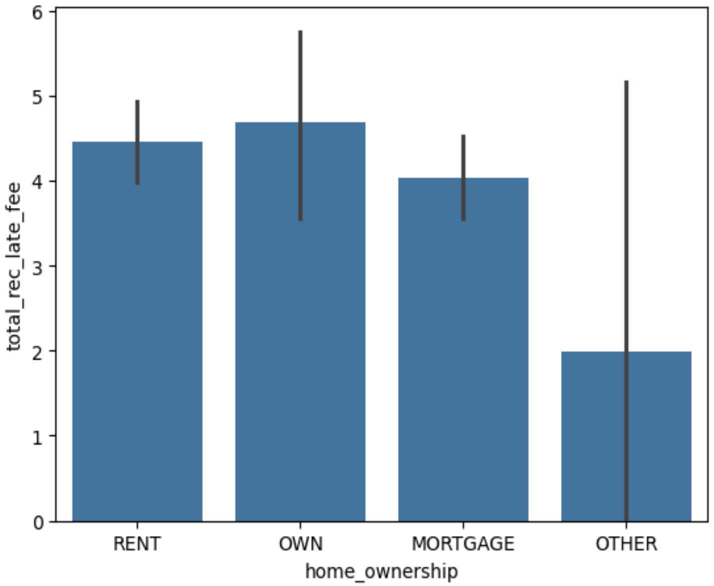


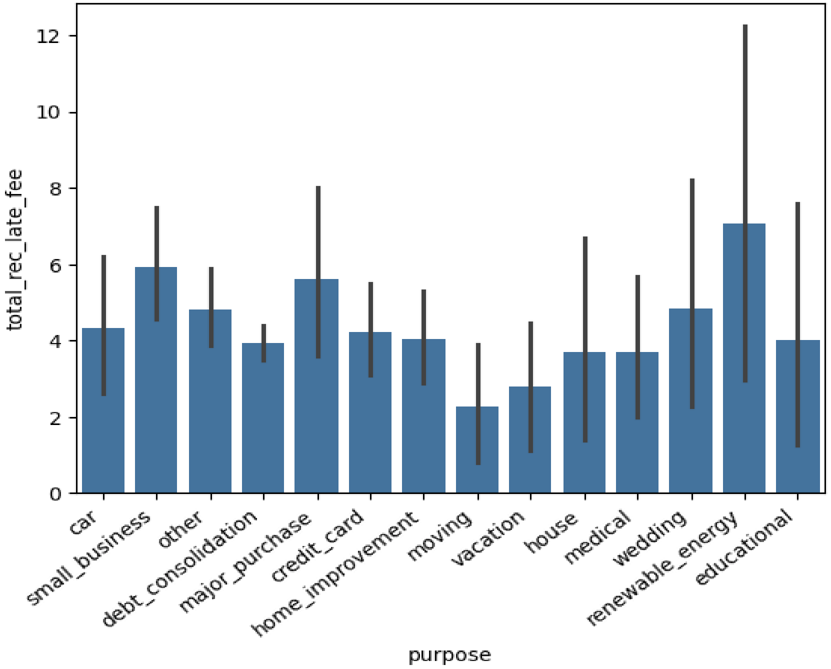


DATA ANALYSIS – BI VARIATE ANALYSIS

Other few Bi Variate analysis results shown below

1. Homeowner ship vs Total recovery late fee
2. Purpose vs Total recovery late fee





Recommendations based on Analysis

Some Interesting outcomes of this analysis are

- Loans with lesser tenure of 36 months have a high chance of defaulting loan might be due to High EMI per month
- Customers who have Rented house or Mortgaged their home tends to default loan
- Customers with annual income of less than 50k has a high chance of defaulting loan might be due to less salary
- Verified customers defaulted loan than non-verified customers - this shows there could be a lapse in verification process or might be due to corruption involved in this process of verifying customers. Needs further investigation
- Customers who got loan for Debt Consolidation purpose defaulted loan heavily, this could be due to heavy loan amount as they were consolidating multiple loans. Need to be careful in lending those customers as they might tend to get heavy loan which they couldn’t pay back

SUMMARY

For future loan grants it is good to consider these recommendations which came out of this analysis seriously and needs further investigation to streamline the process of loan verification. Doing these we can reduce the loan default count in future

Please refer below GitHub Link for more details

https://github.com/muthuvadivel/LendingClubCaseStudy