Lending Club Case Study

Ver 0.1

Problem Statement

- Problem statement of the case study is that the company wants 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.
- 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
- The data given in the case study contains information about past loan applicants and whether they 'defaulted' or not. The aim 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.
- When a person applies for a loan, there are **two types of decisions** that could be taken by the company:

1. **Loan accepted:** If the company approves the loan, there are 3 possible scenarios described below:
    1. **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
    2. **Current**: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
    3. **Charged-off**: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has **defaulted** on the loan
2. **Loan rejected**: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

Analysis Approach

- We have followed the very simple analysis approach as below:-
  1. First we have uploaded the csv file to the python library.
  2. Then we have opened the new python notebook.
  3. We imported the numpy, seaborn, matplotlib and pandas library.
  4. We used the pandas library to **read the CSV file data**.
  5. We first determined the number of rows and columns of the data set.
  6. Then we checked number columns in the data set which has the null values.
  7. We found that there are many columns which have the null values hence started the **data cleaning**.
  8. We used the dropna to drop the columns which have the null values.
  9. After dropping the columns with null values, we found that there are various columns which have very minimal information for the analysis hence dropped the same.
  10. There are various columns which correspond to the post loan approval hence we dropped then same.

      Post Approval features are as below:-

      #recoveries

      #total_rec_late_fee

      #total_rec_int

      #total_rec_prncp

#last_pymnt_d

#last_pymnt_amnt

#next_pymnt_d

#last_credit_pull_d

- There are some columns which will not help in the analysis hence we dropped the same.

#id

#member_id

#url

#zip_code

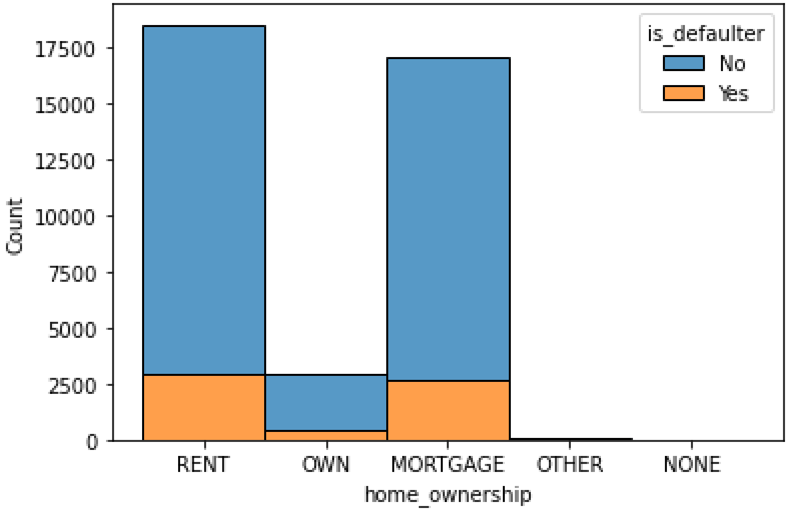
#addr_state

#title

#emp_title

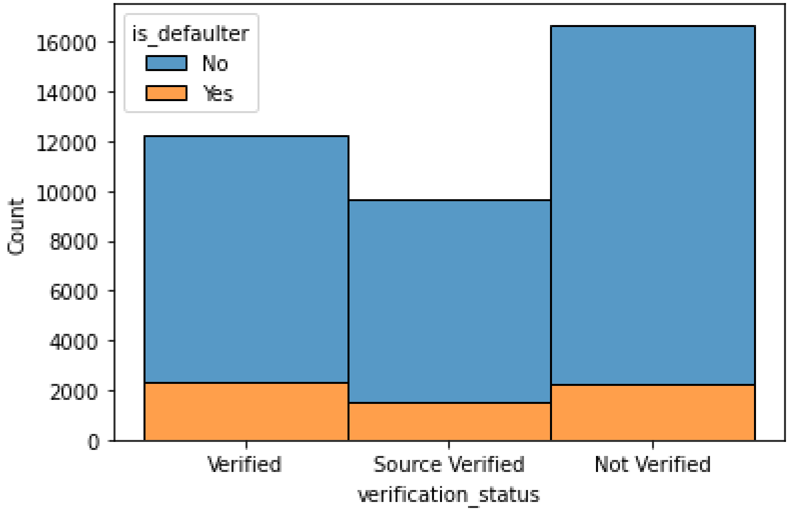
#addr_state

11. There are some columns some more columns which are post approval related hence we dropped the same.
12. We also removed the data related to **loan status as current** as that would not help in the analysis.
13. After removing the data related to current loan status, we have only data related to fully paid and charged off loans.
14. We created a function as defaulter and added that as a column where the **total amount is equal to or greater than the loan amount**.
15. We replaced the revol_util, int_rate and term from string to the float values and integer.
16. Then we renamed the **term** to **term in months.**
17. We encoded the data in the columns 'home_ownership', 'verification_status','loan_status','is_defaulter’ to get the values in 0 and 1 for easy analysis.
18. We determined the correlation between the different varaibles in the data set.
19. We did the bivariate analysis of the cleaned and structured data on basis of **home_ownership** and below is the result as per the histogram.



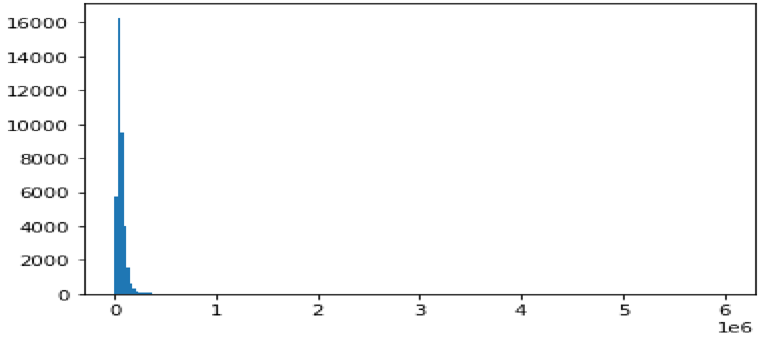
- **From the histogram it is clear that there is very high risk that the people who stay on rent or have mortgaged there house have more chances to default than the ones who have their own houses.**

20. We did the bivariate analysis of the cleaned and structured data on basis of **verification_status** and below is the result as per the histogram.



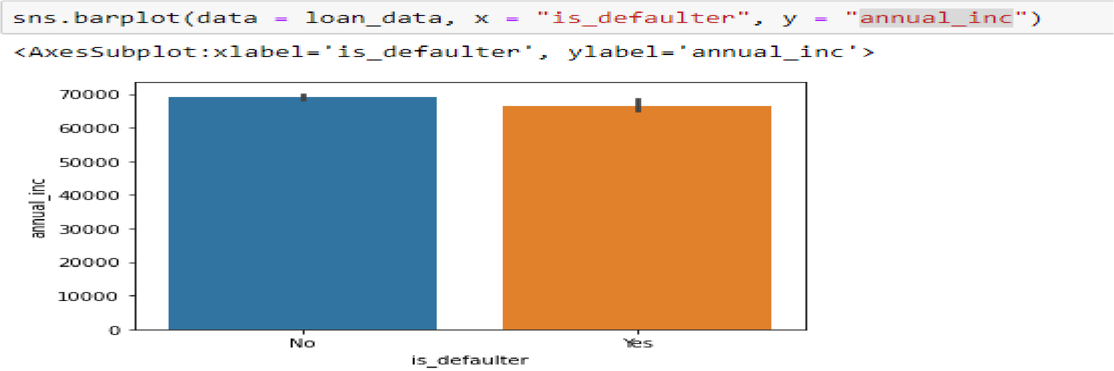
- **From the histogram it is clear that there is very high risk that the people who are not verified have more chances to default than the ones who have been verified or source verified.**

21. We did the univariate analysis of the cleaned and structured data on basis of **annual_inc** and below is the result as per the histogram.



- From the above histogram we reached to the conclusion that the annual income is less for most of the applicants and only one range of people have more annual range.

22. We did the bivariate analysis of the cleaned and structured data on basis of **annual_inc** against the **defaulter as yes** and below is the result as per bar chart.



```
sns.barplot(data = loan_data, x = "is_defaulter", y = "annual_inc")
```

```
<AxesSubplot:xlabel='is_defaulter', ylabel='annual_inc'>
```

Summary

- From the analysis of the loan data set it is found that the people with own houses have less chances of defaulting the loan than the people who live in the rented or mortgaged house.
- From the analysis it is clear that the verified applicants have less chances of defaulting than the applicants who are not verified.
- From the analysis it is clear that the applicants with good annual income have less chances of defaulting then the applicants who have less annual income.