Lending Case Study

Presented By
Mayuri Kumar

Problem Statement

As an online lending marketplace, company aims to :

- Identify 'risky' loan applicants who are likely to default and cause credit loss.
- Analyzing the driving factors behind loan defaults using exploratory data analysis (EDA).
- Understanding the key indicators of loan defaults, to minimize financial loss and for better risk assessment.

Approach

1. Understanding and Data cleaning
    a. Understanding what each columns represents
    b. Dropping columns which cannot be used, having null values, no variation etc.
    c. imputing missing values, removing duplicates, and correcting inconsistencies.
2. Analysing the data
    a. Univariate Analysis
    b. Bivariate Analysis
    c. Multivariate and segmented Analysis
3. Conclusion
    a. Drawing observations and recommendations based on the data

Analysis

Upon data cleaning we narrowed down our data to these columns, these will undergo further analysis to figure out if they are relevant:

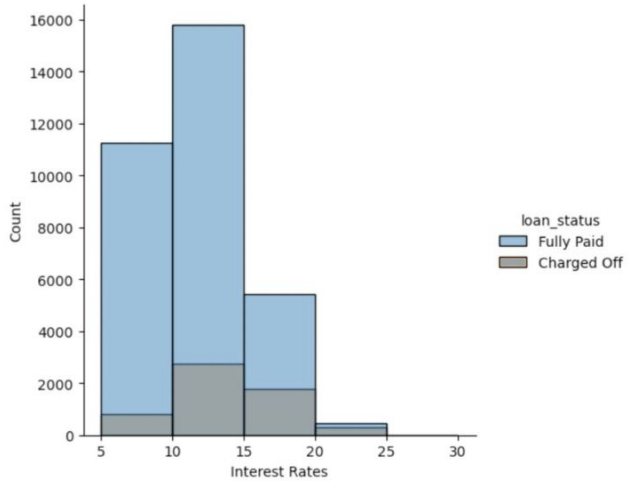
- loan_amnt
- funded_amnt
- funded_amnt_inv
- term
- int_rate
- installment
- grade
- emp_length
- home_ownership
- annual_inc
- verification_status
- issue_d
- loan_status
- purpose
- zip_code
- addr_state
- dti
- delinq_2yrs
- open_acc
- pub_rec
- revol_bal
- revol_util
- total_acc
- total_rec_prncp
- last_credit_pull_d
- pub_rec_bankruptcies
- earliest_cr_line
- inq_last_6mths

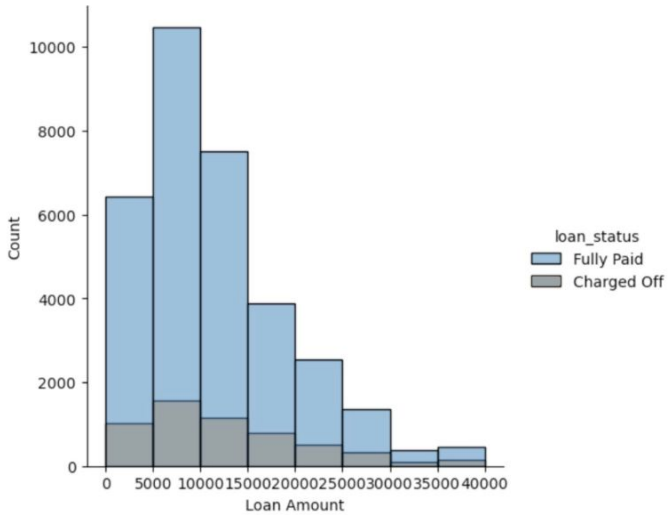
Data cleaning

Analysing each table using value_counts has been found

- These can be used after modification
  - emp_length, revol_util, last_credit_pull_d, pub_rec_bankruptcies
- Cannot be used
  - emp_title dropping as it has a lot of unique values
  - desc dropping this has no standard format and 12k values missing
  - title dropping as it has same data as 'purpose'
  - mths_since_last_delinq dropped too large values missing
  - mths_since_last_record dropped too large values missing
  - next_pymnt_d dropped due to very large values missing
  - last_pymnt_d dropped too large values missing
  - collections_12_mths_ex_med all values are 0
  - chargeoff_within_12_mths all values are 0
  - tax_liens all values are 0

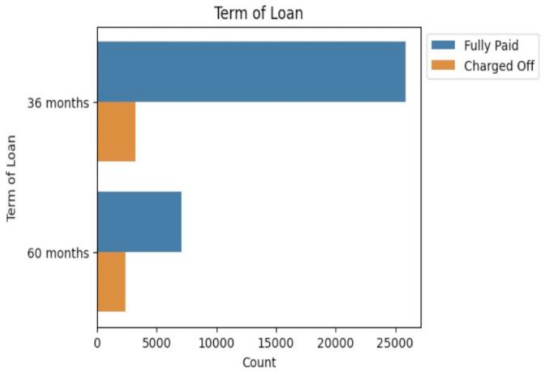
Observation via graphs

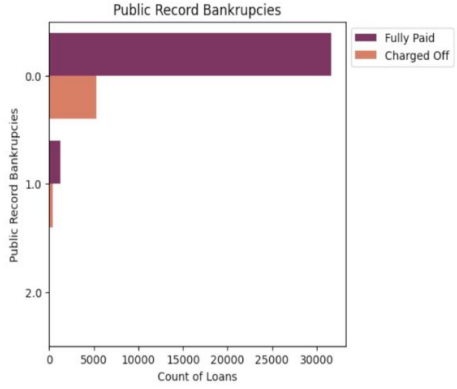


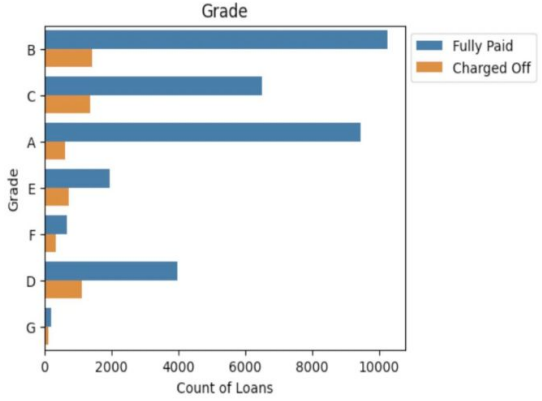


Observation via graphs

- Term of loan has a high probability of getting defaulted if term 60 Months
- Better grades(and sub grades) are observed to have positive impact on loan getting fully paid.

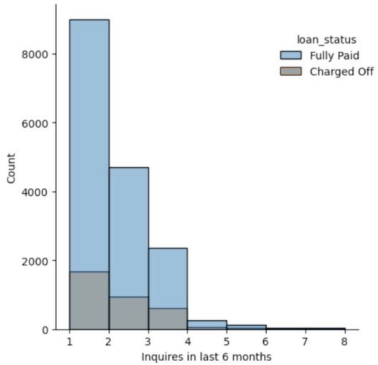


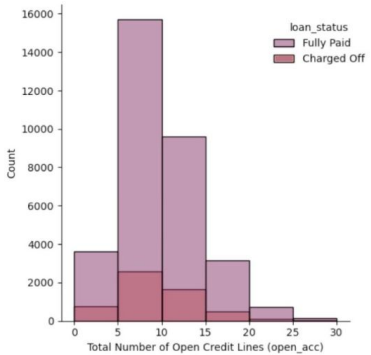




Observation via graphs

- If inquires made by person are more in last 6 months are increasing - so is his probability of defaulting
- Better grades(and sub grades) are observed to have positive impact on loan getting fully paid.





Observation via graphs

- Majority of the interest rate is in the range of 5% to 15%
- Most loan applicant fall in the range of 5k to 10k loan amount
- Grade B has majority of loan applicants
- Most loan application count are in terms of 36 months

Heatmap of numerical values

Negative Correlation

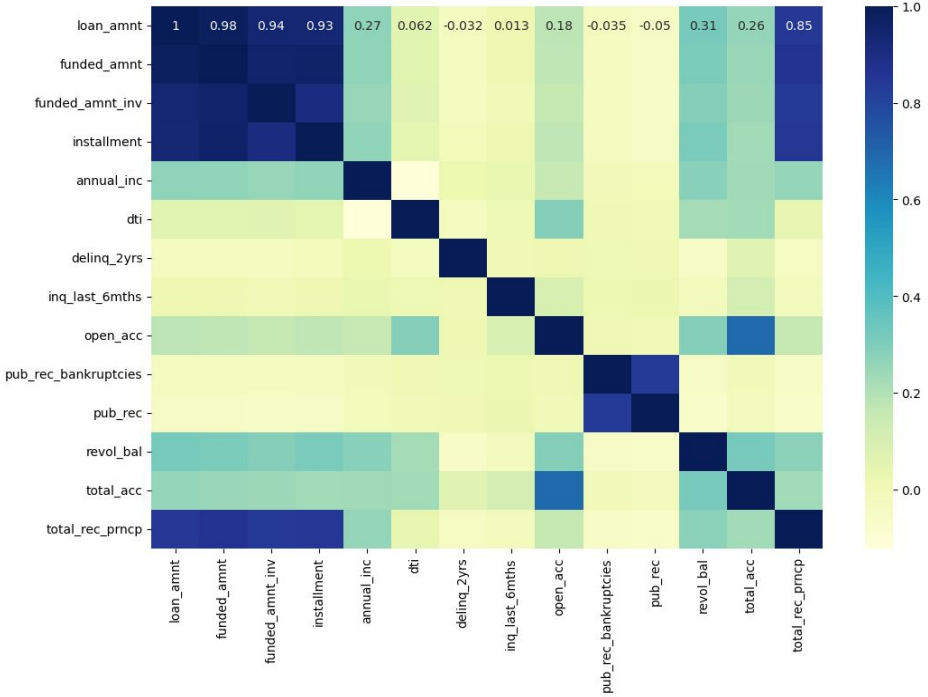
- loan_amnt has negative correlation with pub_rec_bankrupticies
- loan_amnt,funded_amnt, funded_amnt_inv and installment has negative correlation with delinq_2years
- annual income has a negative correlation with dti
- pub_rec_bankrupticies has negative correlation with all of the fields.

Strong Correlation

- annual income has a strong correlation with loan_amount
  term has a strong correlation with loan amount and interest rate
- loan_amnt, funded_amnt,installment and total_rec_prncp has strong correlation with each other.
- revol_bal has positive correlation with loan_amnt, funded_amnt, funded_amnt_inv, installment

Weak Correlation

- Pub_rec_bankruptcies and dti has weak correlation with most of the fields



Recommendations & Conclusion

The following segments seem to have strong influence on defaults :

1. **purpose** (small business is more likely to default)
2. **grade/sub_grade** (lower grades are more likely to default)
3. **addr_state** (some states are more likely to have high defaults)
4. **delinq_2yrs** (higher values seem to have more)
5. **inq_last_6_months** (probability of default increases with number of inq)
6. **term** (36 month loan has high probability of full payment)
7. **pub_rec_bankrupcies/pub_rec** (individuals reporting higher values can lead to defaulting)
8. **loan_amnt** (Higher rates can increases the probability of person defaulting)
9. **int_rates**(majority of the interest rate is in the range of 5% to 15% and the Very High interest rates are in more likely to default which is greater than 15%)

Recommendations & Conclusion

The following segments seem to have **Low/No** influence on loan defaults :

1. Verfication_status
2. Emp_length
3. Home_ownership
4. Installments
5. Issue_d
6. Total_acc
7. Total_rec
8. Dti
9. Home_ownership