Lending Club Case Study

Business Understanding

This case study is for a loan lending Finance Company which specialises in lending various types of loans.

Company makes profit when customers pay loan on time.

The company cannot give loan to all , as some applicants may default which may be a loss for company.

They need some mechanism to identify Risk Factors of applicants based on their profile and decide whether to lend loan or not.

Problem Statement

Identify whether to approve loan for a customer or not.

Chances are that some customers may default. So we are trying to solve problem associated with loan lenders and identify risks associated with.

Company can lend loan with risks , based on high interest rates.

We need to identify such risks also.

Necessary Steps

1.Collecting Data

2.Understanding Data

3.Data Cleanup(Keep Necessary and relevant Data , remove Null Values)

4.Analyse different features

5.Visualise Data through different Graphs

6.Study Correlation between features

7.Identify main features which contribute to further case study

Data Set

Loan Case Study Data set provided by Upgrad

Technologies

1.Python

2.Jupyter Notebook

3. Pandas

4.Seaborn for graphs

5.Google Colab

Data Set Information

1.In data set we had 39717 rows and 111 features

2.There were some columns which have missing values

3. All columns fall in different data types of float,int,object

4.54 Empty Columns are present

EDA Steps and Observations

1.We identified Columns which have empty values and dropped them

2.We identified Columns which have one unique value and dropped them

3.Based on further information of dataset, we dropped columns with high number of null values

4.Based on loan status , calculated loan paid and charged off percentage as additional feature

Further Observations

1.Based on loan loan status, we found out -

85.41% applicants have Fully Paid their load, while 14.59% have been Charged Off.

2.Observed some outliers in interest rate data

3.There were large number of null values in title column

4. Saw some more outliers in other columns also

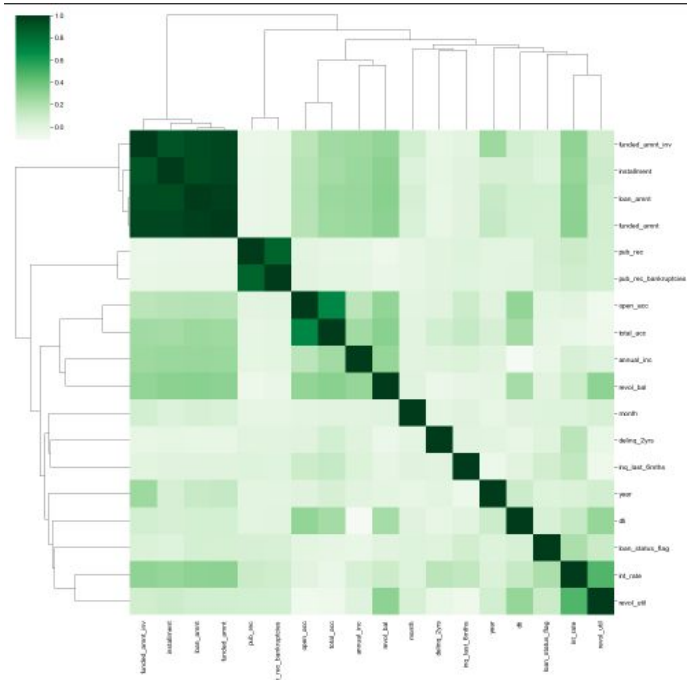
5.Extracted some features from existing data like month and year based on loan issue date

Observations in Correlation Matrix

1.Strong correlation visible between loan_amnt, funded_amnt, funded_amnt_inv and installment. These fields are proportional to each other.

2.pub_rec & pub_rec_bankrupcies are correlated. We also see a positive correlation between public records and number of accounts opened.

3.Positive correlation is visible between int_rate and revol_util

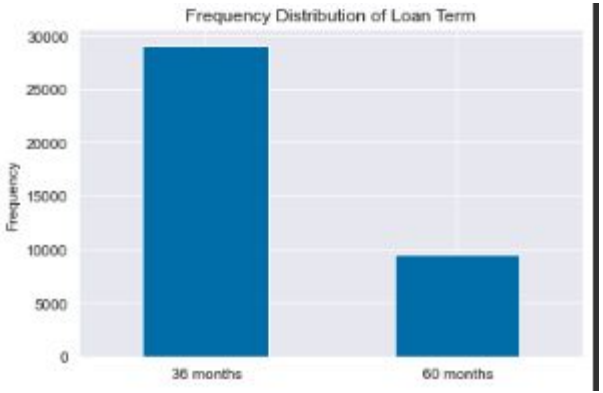


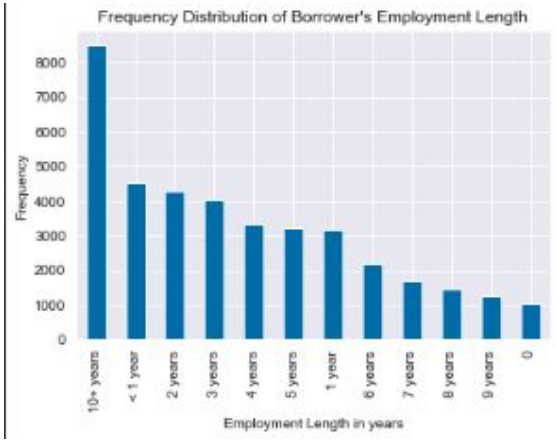
Univariate Analysis

1.Loan Amount Distribution(5000 to 15000 variation is more)

2.Term(30 months is more)

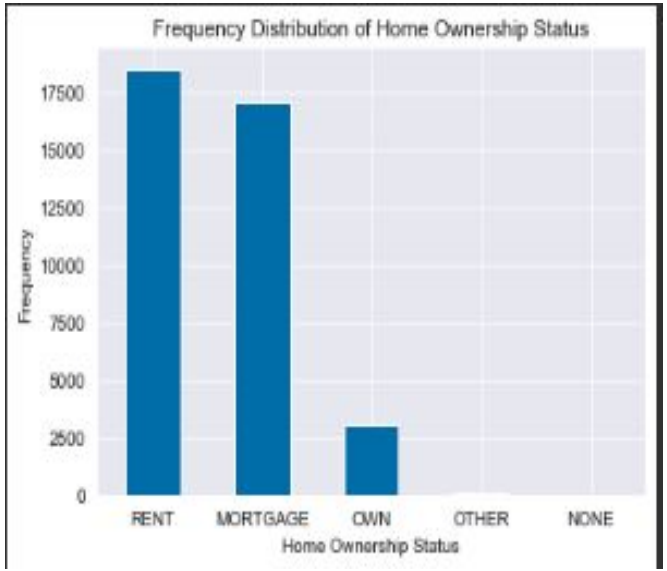
3.Borrowers Employment Length(10 years is highest)

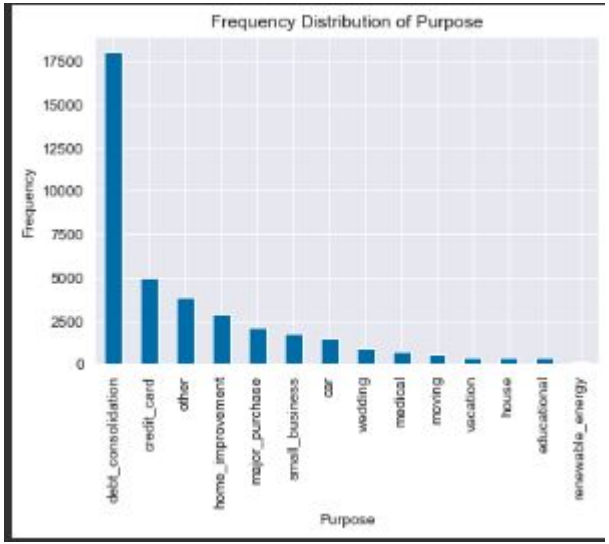




1.Home Ownership(Rented House Applicants are more)

2.Frequency Distribution of Purpose(debut payment loans are more)

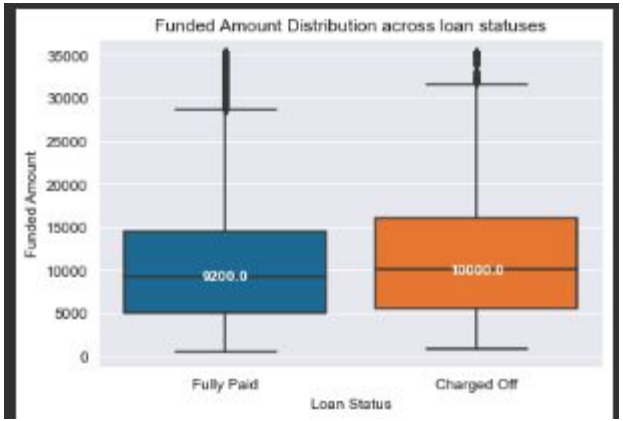


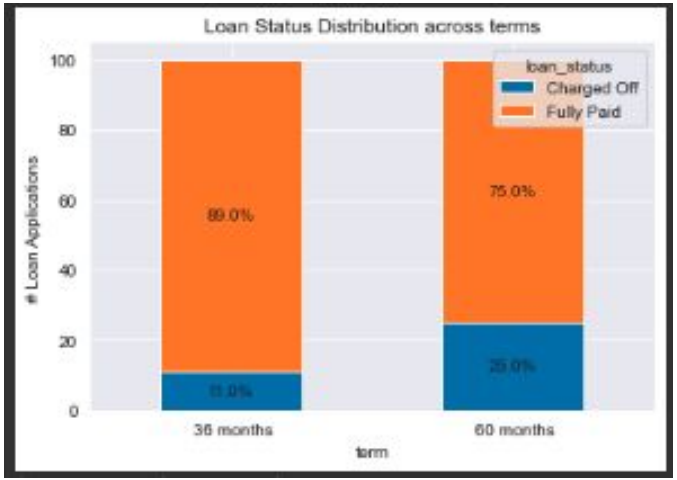


Analysis of Variables

1.Here we can see clearly Loan applicants with higher amount have tend to default than lower amount

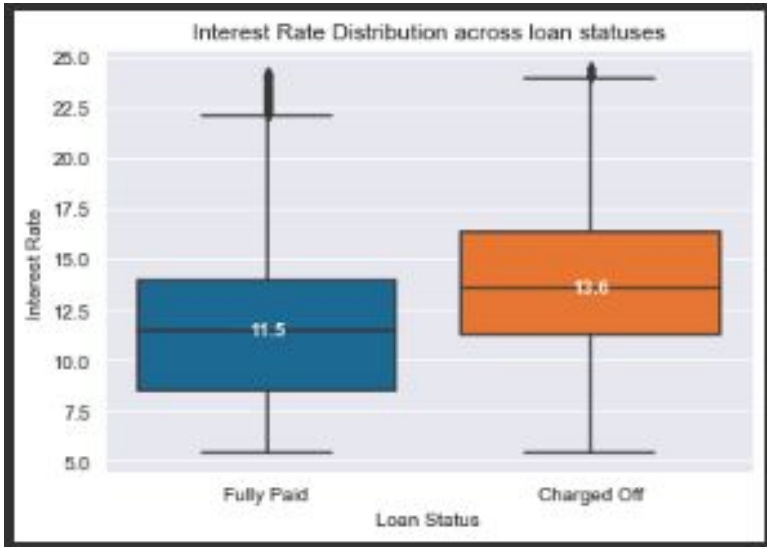
2.Loan Applicants with 60 months are likely to default than that of 36 months

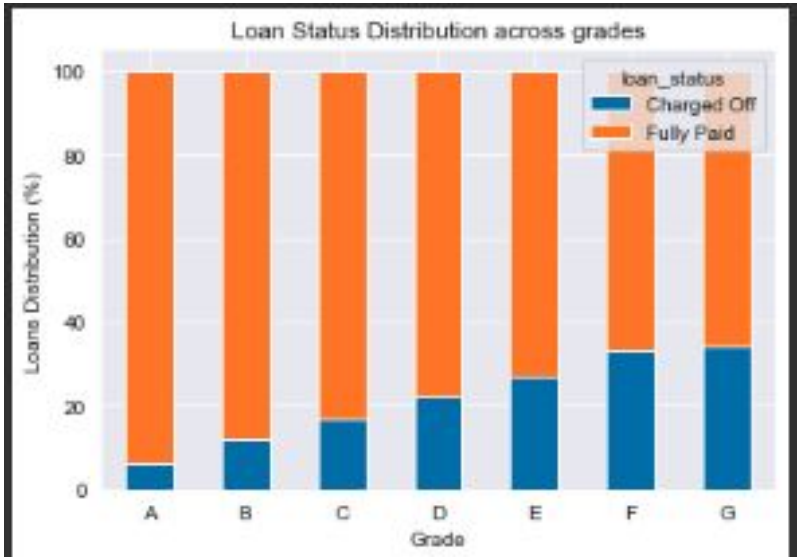




1.Loan applicants with higher interest rates more likely to default than of lower interest rates

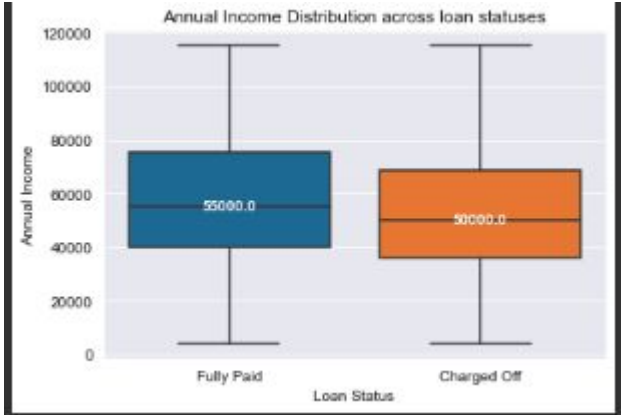
2.Loan applicants with E,F,G are more likely to default than other graders, This correlates with interest rates since E,F,G has higher interest rates

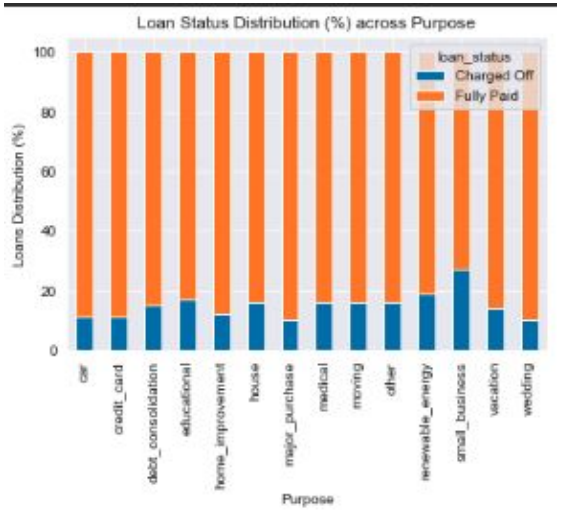




1.Loan Applicants who has lesser Annual Income are likely to default than that of higher annual income

2.Loan Applicants for the purpose of Small Business tend to default more than other purposes by a huge difference between 8 % - 17 %

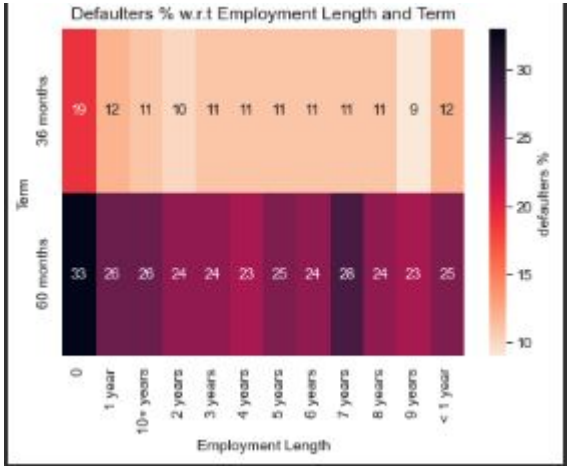




Bivariate Analysis:

1.Loan Amount vs Funded Amount-There are very less defaulters when Funded Amount is same as the Loan Amount(linear relation)

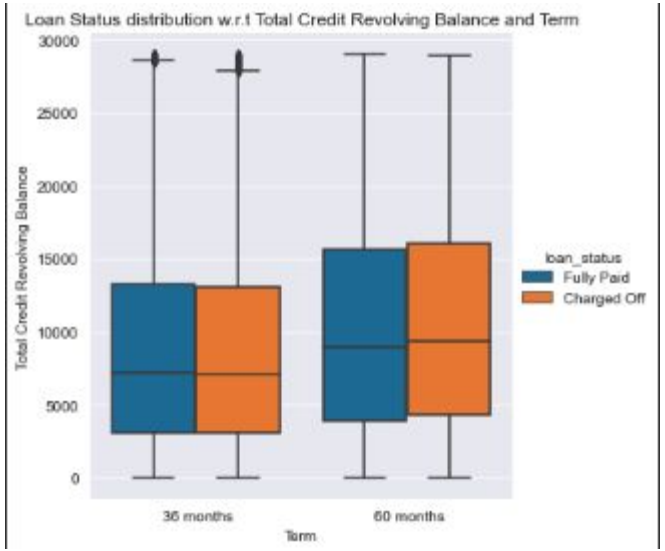
2.Term vs Employment Length-Loan Applicants with employment length of 7 years and NA values tend to default more when the term of the loan is 60 months

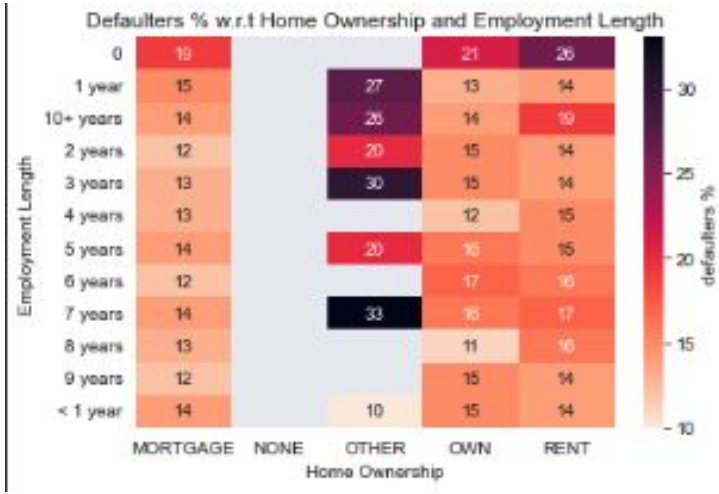


Bivariate Analysis

1.Term vs revol_bal - Loan Applicants with higher Total Credit Revolving Balance tend to default likely when the term is 60 months

2.Emp Length vs Home Ownership - Loan Applicants with OTHER home ownership tend to default more when their employment length is 7 years or 3 years





Conclusion

We have identified additional parameters which are contributing to identifying our problem statement.But to keep presentation short,I am concluding with 5 key factors for identifying Risk -

1) Loan Purpose

2) Grade

3) Higher Interest Rate (Above 13.6%)

4) Higher Revolving line utilization rate

5) Home ownership

Combined Impact

1) Home ownership Vs loan purpose

2) Term Vs Verification Status