LendingClub Case Study

Submitted By:
P Paul Jonathan (ppauljonathan@gmail.com)
Pradeep Kumar R (pradeep.kumarr10@gmail.com)

General Information

What is the background of your project?

The background of this project is to develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.

General Information

What is the business problem that your project is trying to solve?

The Business Problem that this project aims to solve is how consumer attributes and loan attributes influence the tendency of default and 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.

General Information

What is the dataset that is being used?

We are using a dataset of the complete loan data for all loans issued through the time period 2007 to 2011 by LendingClub - a consumer finance company which specialises in lending various types of loans to urban customers.

Driving Variables behind loan defaults

We have done Univariate and Bivariate analysis on the data and have identified the following variables as having a high impact on whether or not the customer will default on their loan

- grade
- sub_grade
- emp_length
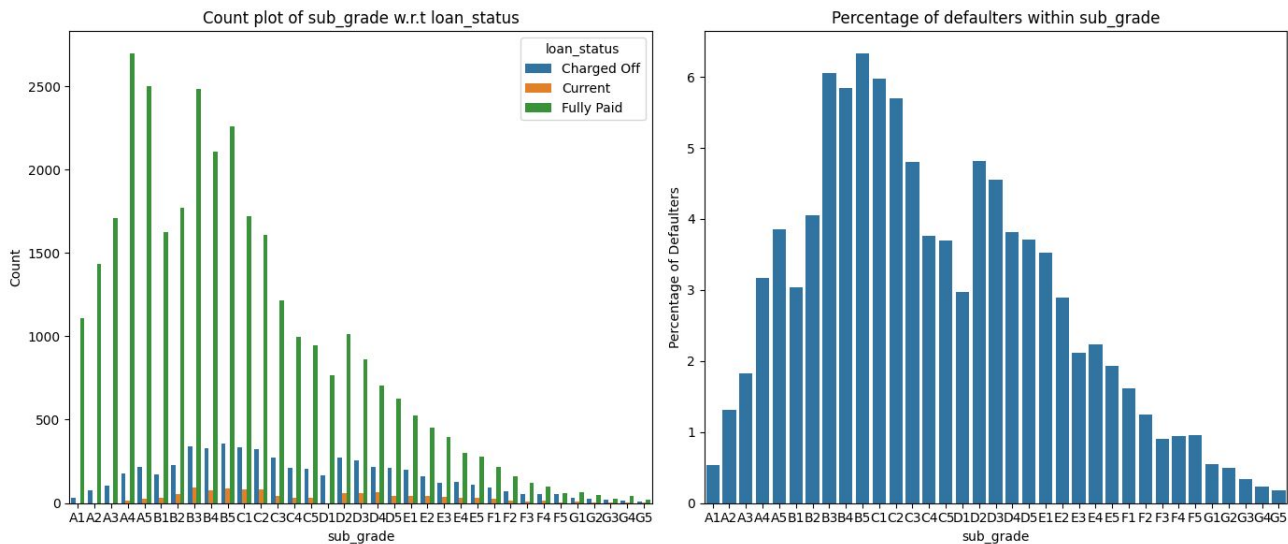- home_ownership
- verification_status
- purpose

grade as Driving Variable for Defaulters

We can see that loan applicant with grade B are the highest defaulters.
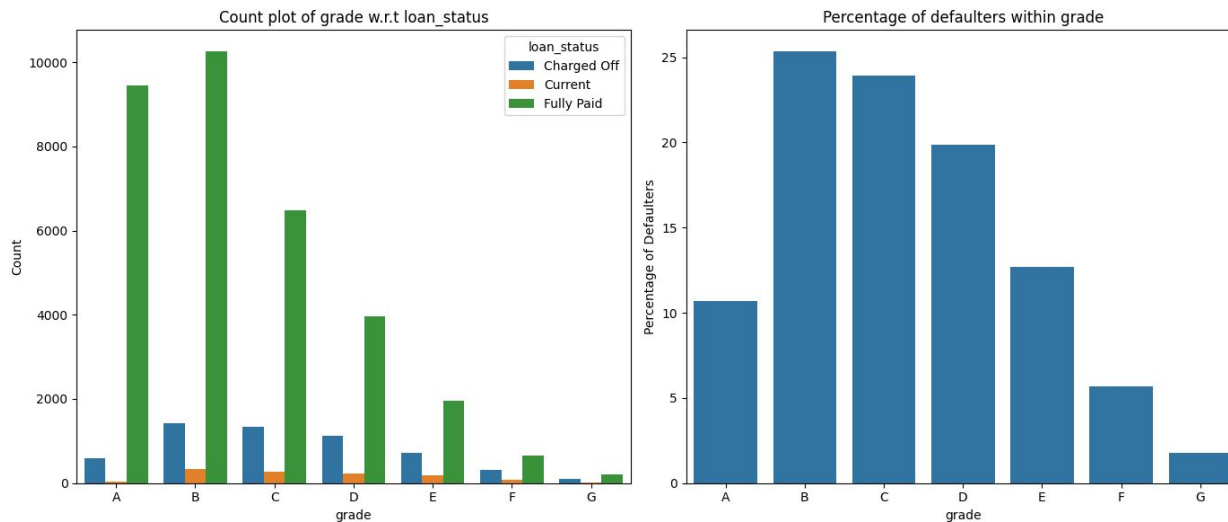
sub_grade as Driving Variable for Defaulters

sub-grades B and C being the highest defaulters suggests that these borrowers might have moderate credit risk. However, the fact that sub-grade B5 has the highest percentage of defaulters indicates that within the B grade, B5 is particularly risky.
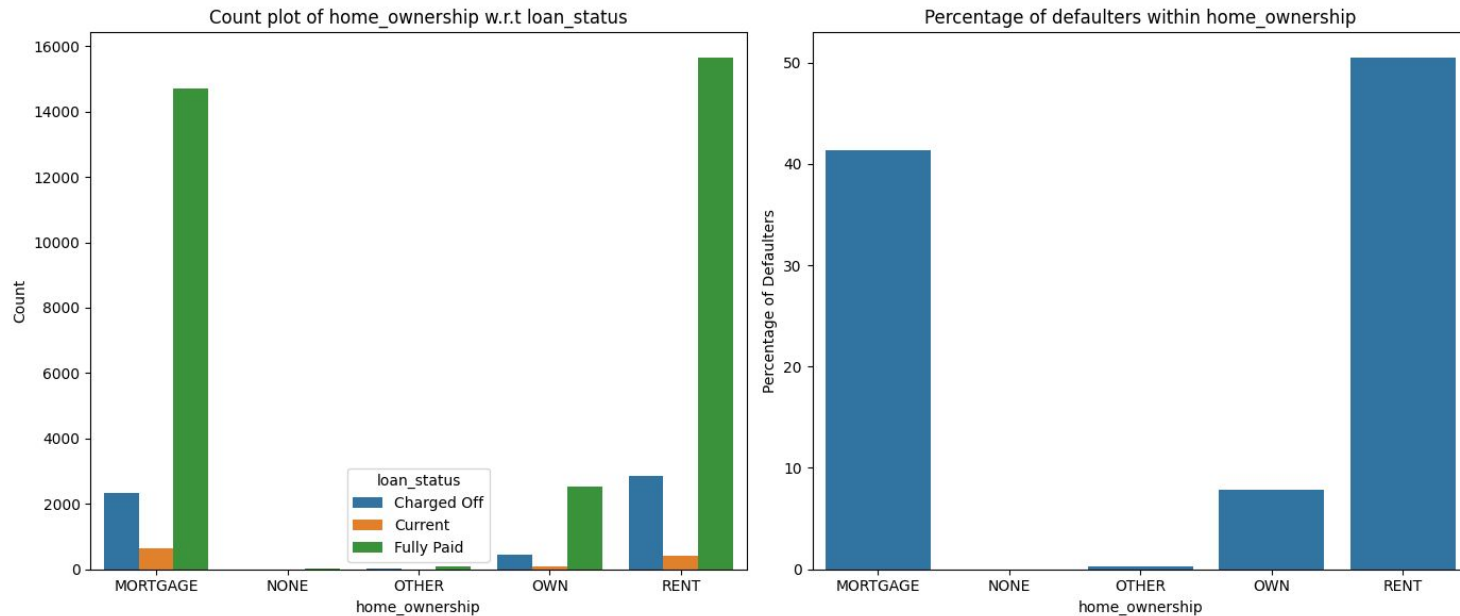
emp_length as Driving Variable for Defaulters

We can see loan applicant with 10+ years of experience are the highest defaulters.
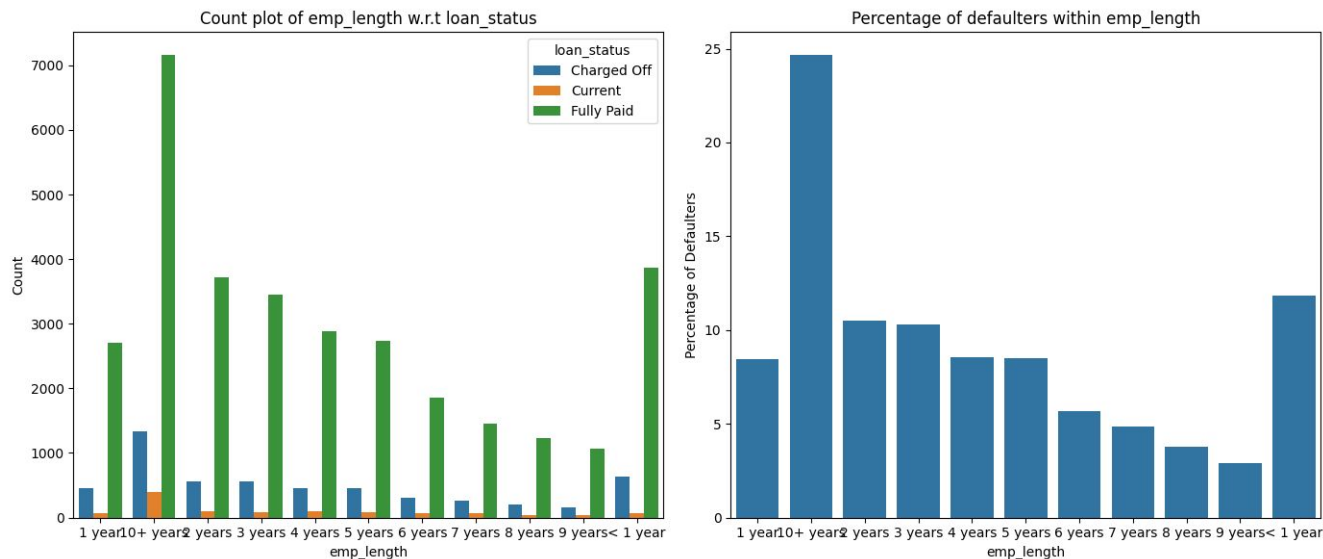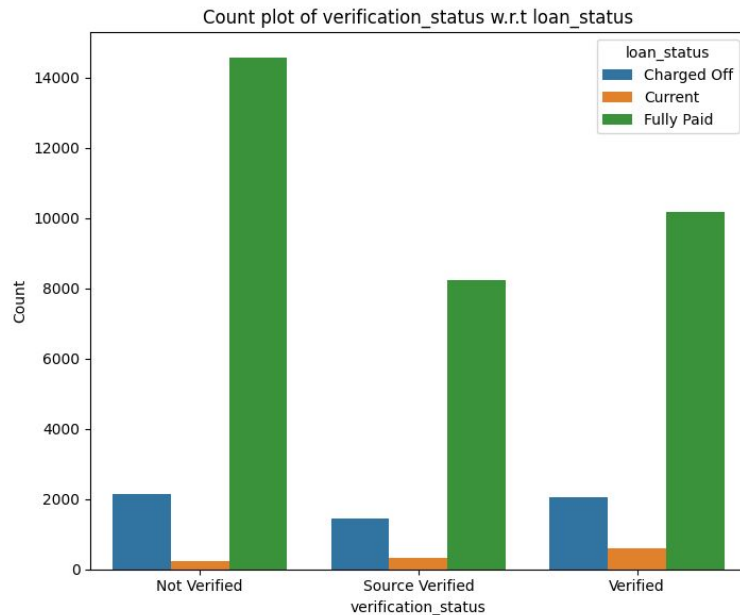
home_ownership as Driving Variable for Defaulters

We can see loan applicant living in rented/mortgage house are the highest defaulters.

Count plot of home_ownership w.r.t loan_status

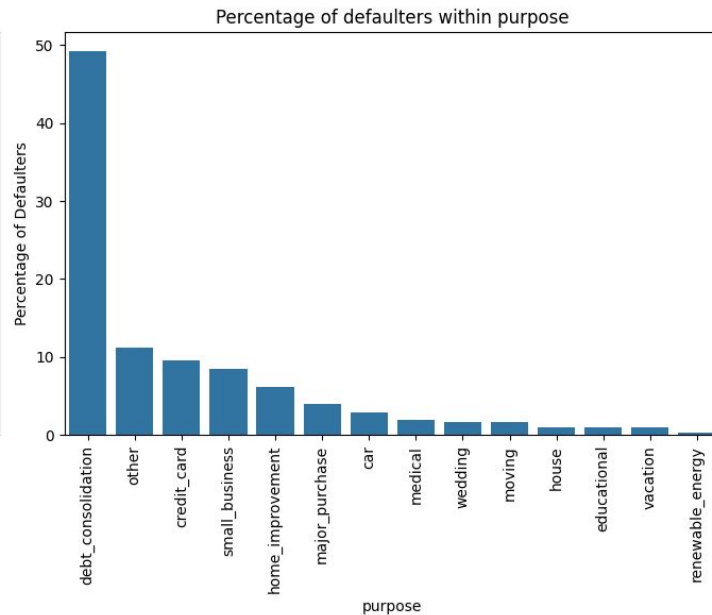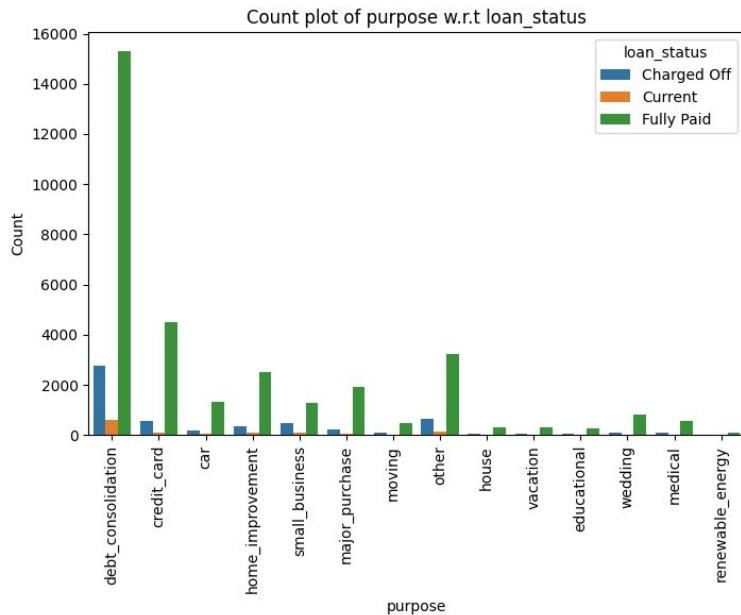Percentage of defaulters within home_ownership

verification_status as Driving Variable for Defaulters

We can see loan applicant who are 'not verified' are the highest defaulters.
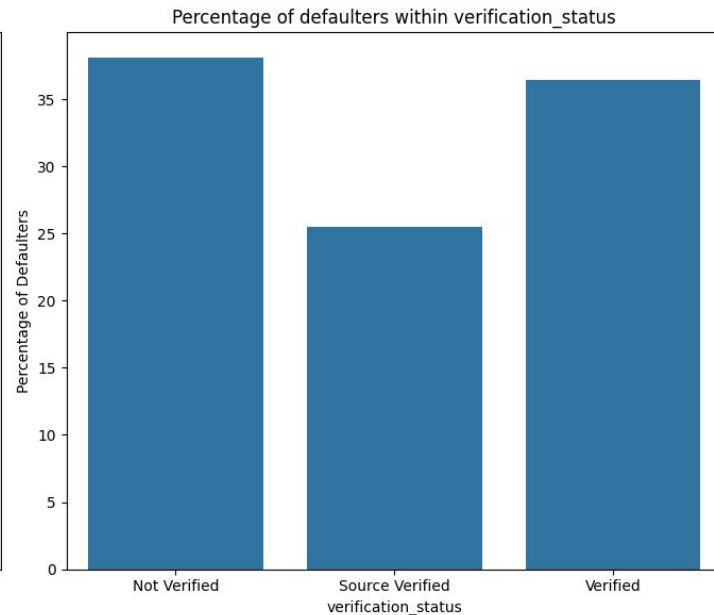
purpose as Driving Variable for Defaulters

We can see loan applicant whose loan purpose is related to 'debt_consolidation' are the highest defaulters.

Bivariate Analysis

We have created a Correlation matrix for all the meaningful variables of the dataset and found that the following variables have a very high correlation (> 80%)

- loan_amnt, funded_amnt, funded_amnt_inv and installment is highly correlated with total_pymnt, total_pymnt_inv, total_rec_prncp, total_rec_int
- total_acc is highly correlated with open_acc

Bivariate Analysis
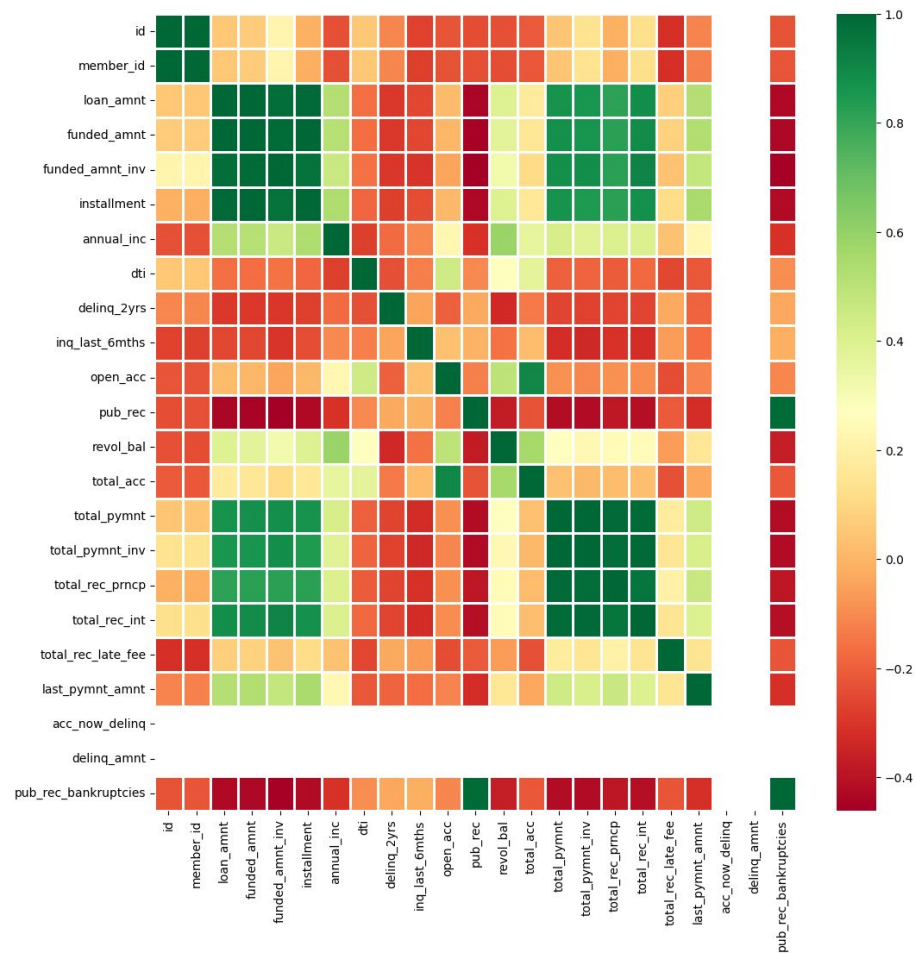
Conclusion

Decisive Factor whether an applicant will be Defaulter:

The following attributes indicate that people from these category tend to defaulters but then due to the number of people and the amount of loan, the bank could provide loan with higher interest to mitigate any default risk thus preventing business loss:

1. grade: Grade B has a relatively higher default rate.
2. sub_grade: among the grades, we see that sub_grade B5 has a high default rate.
3. emp_length: we can see loan applicants with 10+ years of experience are the highest defaulters
4. home_ownership: we can see loan applicants living in rented/mortgage house are the highest defaulters.
5. verification_status: we can see loan applicants who are 'not verified' are the highest defaulters.
6. purpose: we can see loan applicants whose loan purpose is related to 'debt_consolidation' are the highest defaulters.