Lending Club Case Study

Submitted by Mithun R

Problem Statement

Lending club is a consumer finance company providing loan to urban customers. Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss).

They need to identify these risky loan applicants to reduce their credit loss

Objective

Identify 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 taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

Data Understanding

Some of the important columns in the data set are annual income, purpose of the loan, loan amount, loan term, interest rate, grade etc.

Based on the given data, we identified ‘loan_status’ as our target variable.

We will see how loan status is related to other variables to identify patterns and associations that might help identify driving factors. The strategy is to compare the average default rates across various independent variables and identify the ones that affect default rate the most.

Data Cleaning

Step 1:

Our first step was to deal with null values.

We removed all those columns that had more than 90% null values (90% was chosen based on the current data set and not any predefined threshold)

Step 2:

We removed those columns that had 0 variance.
As all values in these columns were same and keeping them does not help with our analysis.

Step 3:

We removed ‘desc’ column too.
Reason: desc is essentially text entries explaining reasons for borrowing. To analyze these data, we would need NLP or add logics to make sense out of it. We decided against using it for our analysis.

Step 4:

We then looked for those entries (rows) which had more than 5 missing values. However, there were none.

Step 5:

Next step was to ensure all columns were in the right format

‘int_rate’ was in string format. This was converted to number

‘emp_length’ was in string format. This was converted to number

Data Analysis

IMPORTANT

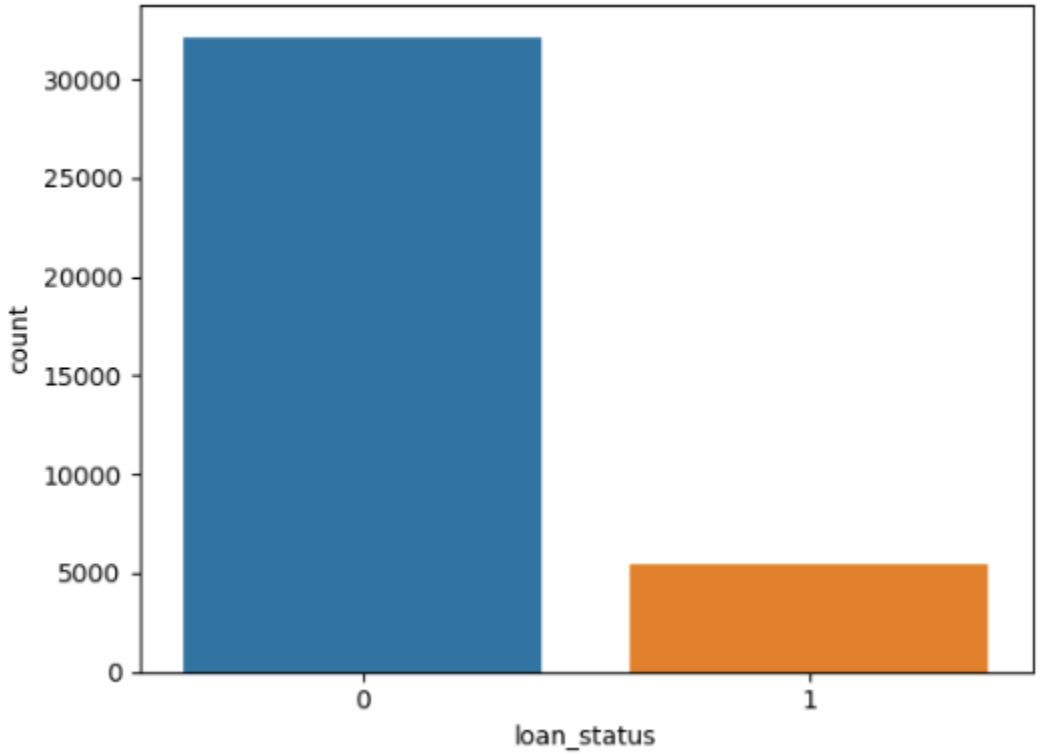
- In our analysis, we decided to focus only on those attributes which are available for all users including first time loan applicants. Hence, we decided to drop all those columns which are characteristics of loan repayment

  E.g.:- delinq_2yrs, last_payment_d, recoveries etc.

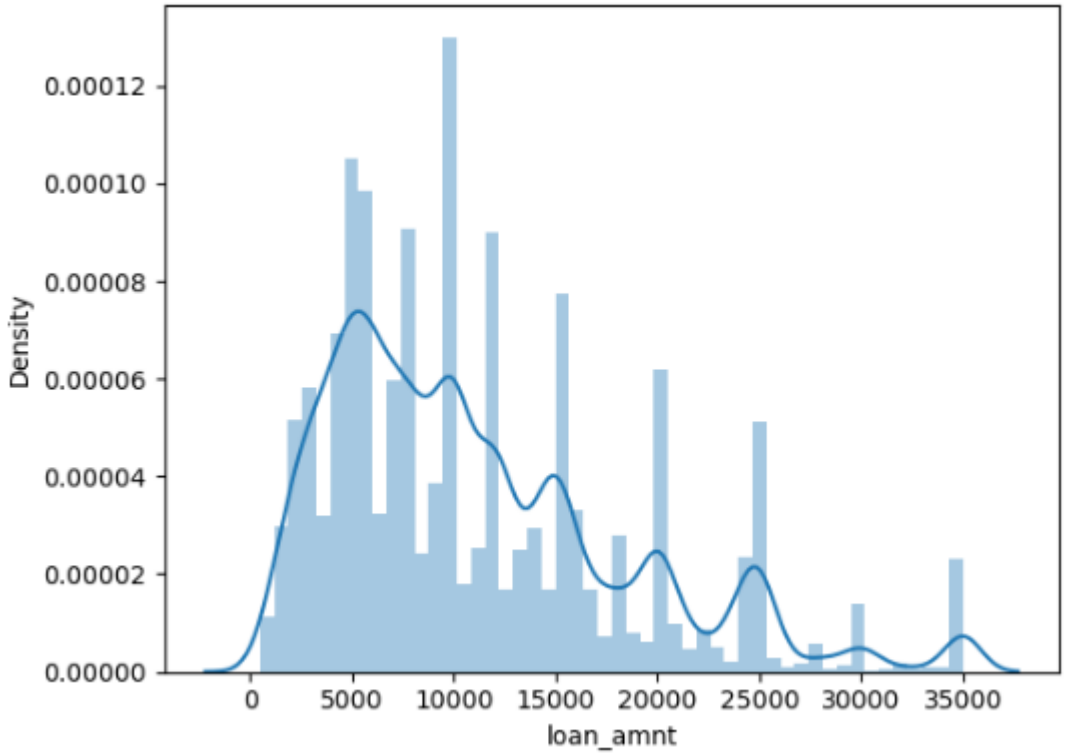
- Non influential customer characteristics like id , member_id, url, title, zip code and addr_state were removed too
- ‘loan_status’ had 3 values of which ‘Current’ is not of importance to our analysis, since they are currently paying their loan and can go either ways. Hence, we had decided to filter out these entries
- We then converted loan_status to a numerical variable so that we can calculate mean etc across driver variables

Univariate Analysis

Loan Status: 85% total customers had completely paid off their loan against 15% who defaulted

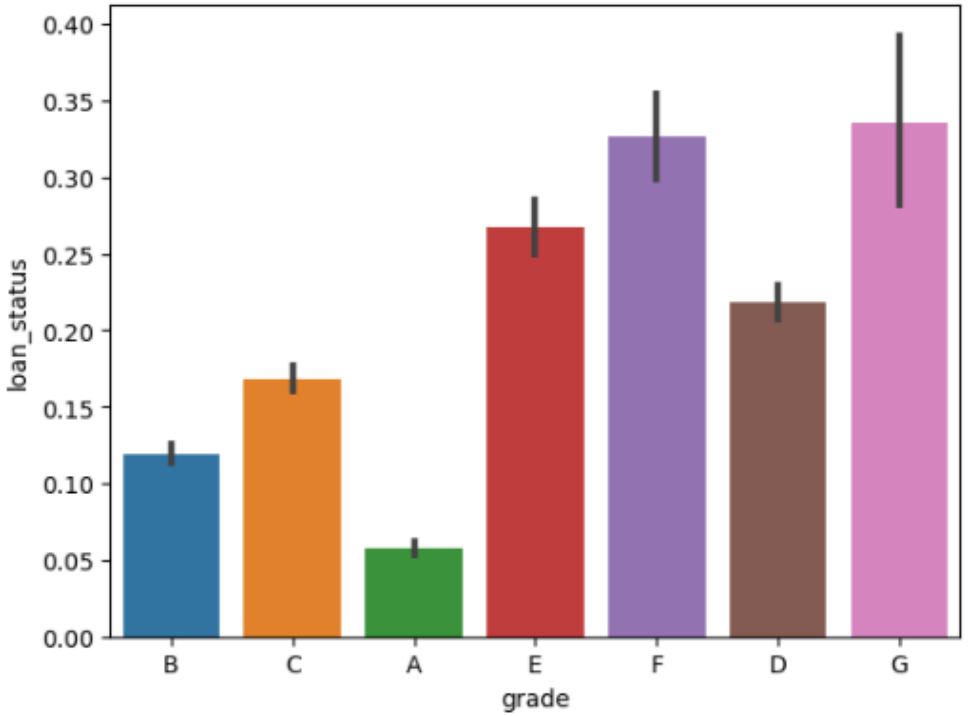


Loan Amount: Lending club has disbursed lower loan amounts in higher density



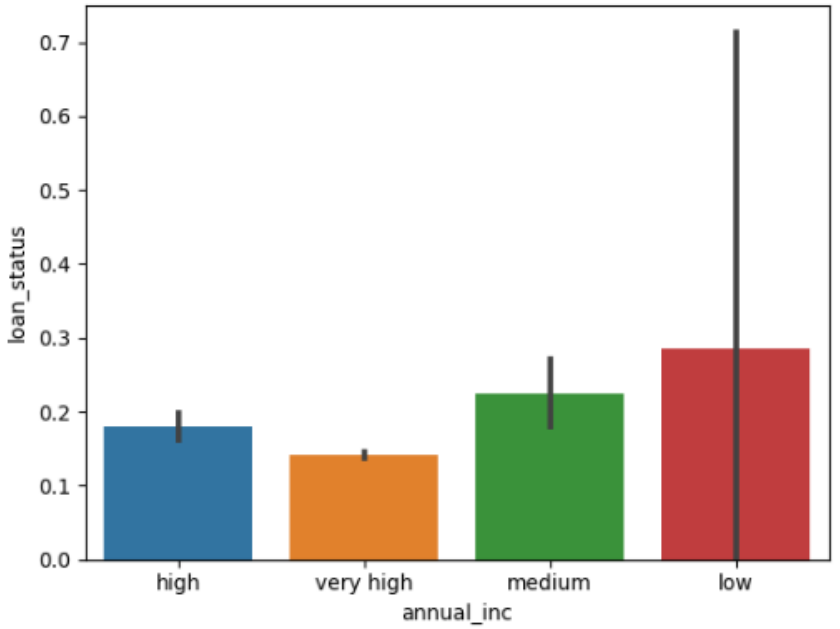
Bivariate Analysis

GRADE



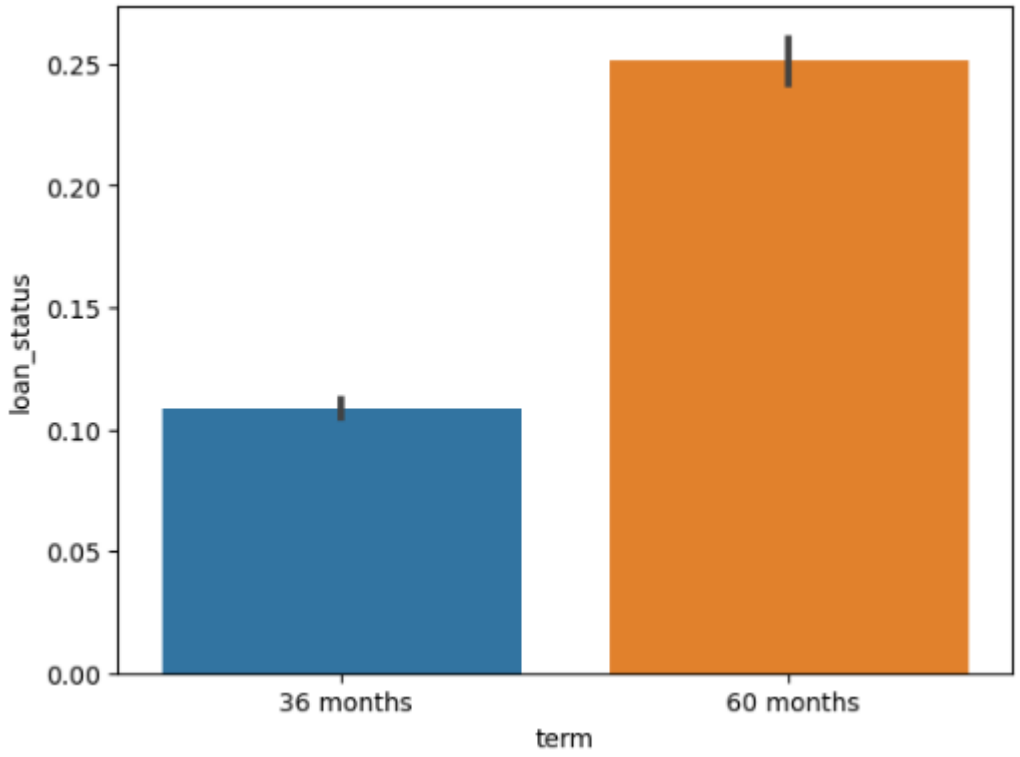
There seems to be a direct correlation here, where the default rate increases as the grade moves from A to G. This is expected as the grade is decided based on the riskiness of the loan

ANNUAL INCOME



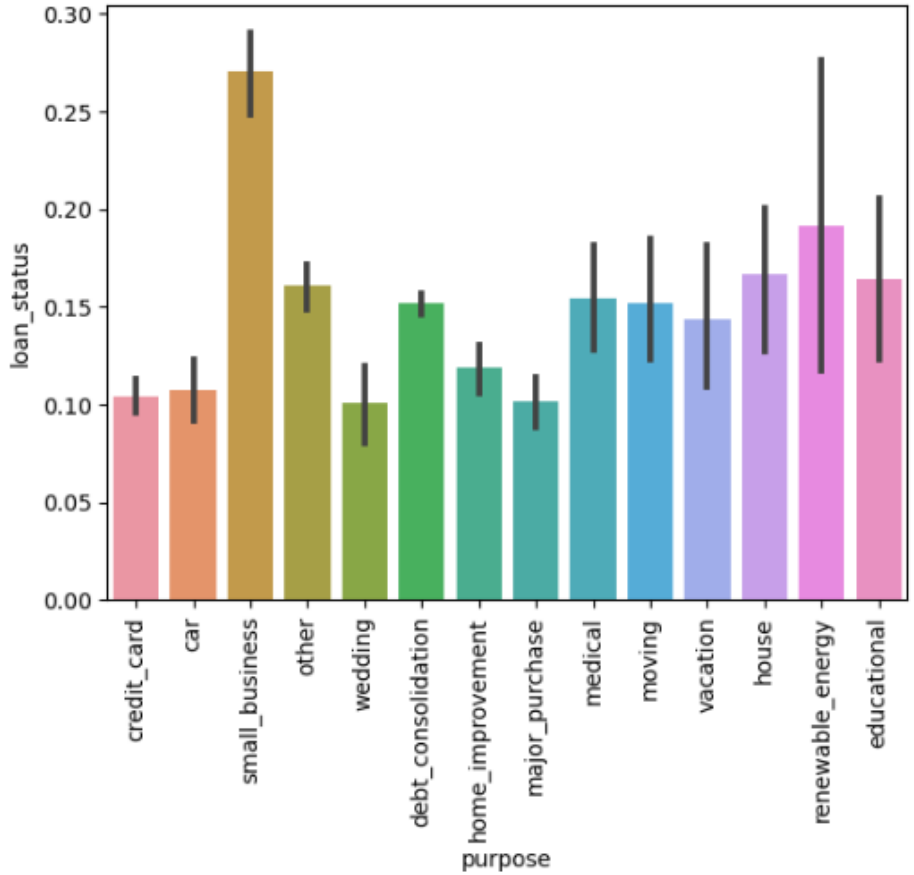
As we can see here higher annual income is linked to lower default rate. There is ~6% decrease in default rate when we move from low to high income range

TERM



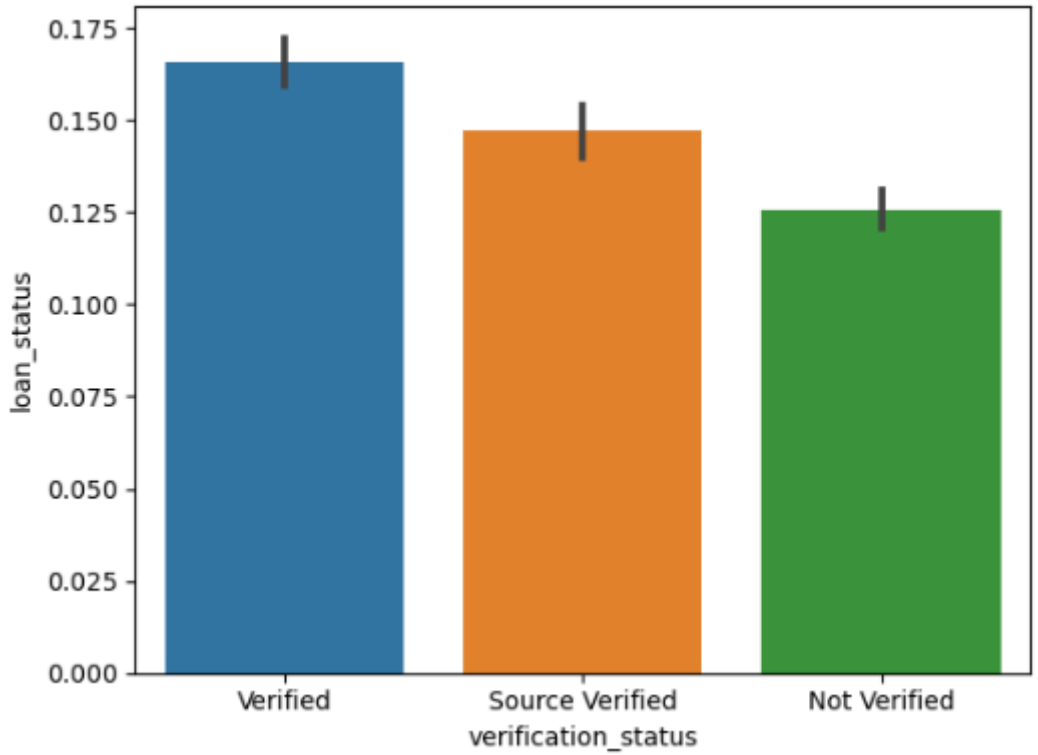
Higher terms have higher default rate

PURPOSE



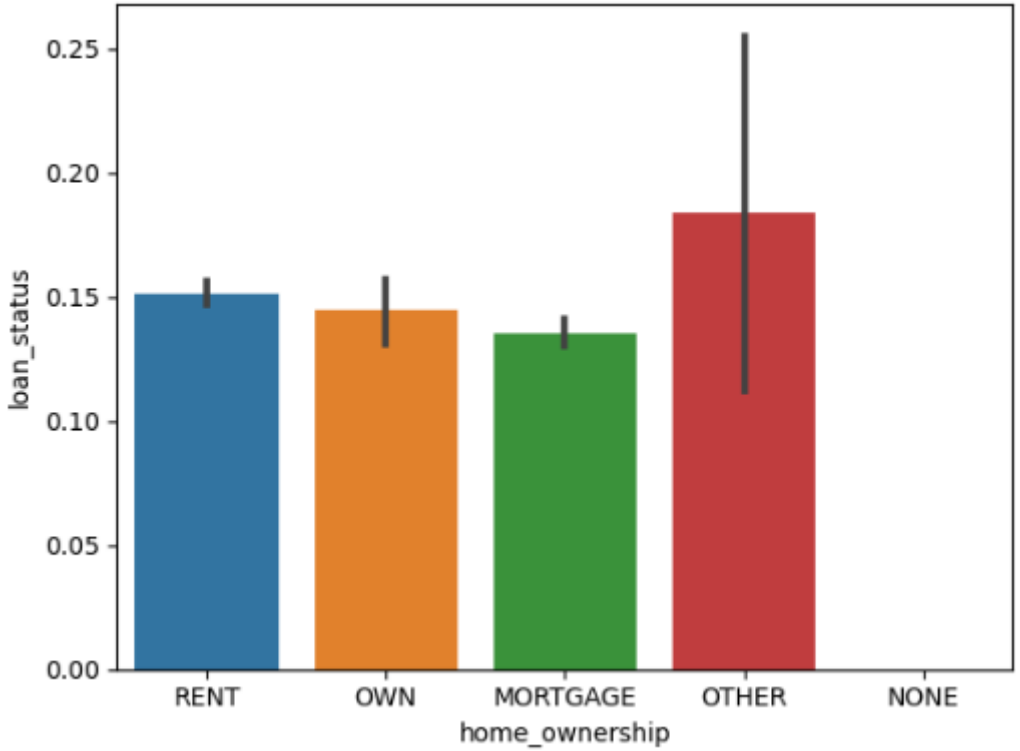
Personal purpose loans seems to be having lower default than business purpose loans

VERIFICATION STATUS



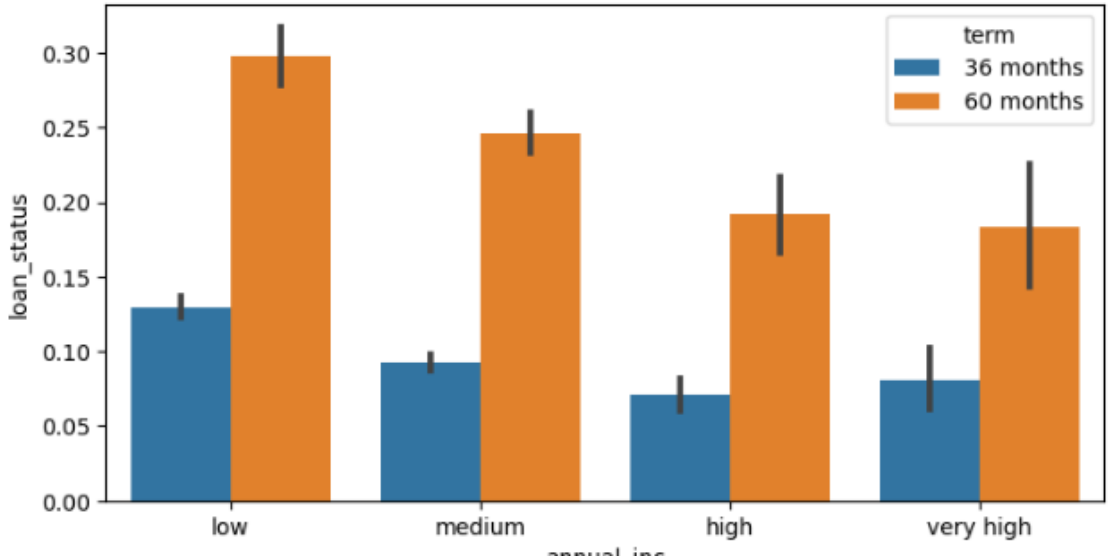
Though we would assume verification would increase the chance of repayment, data shows otherwise. Default rate is higher amongst verified profile vs non verified

HOME OWNERSHIP



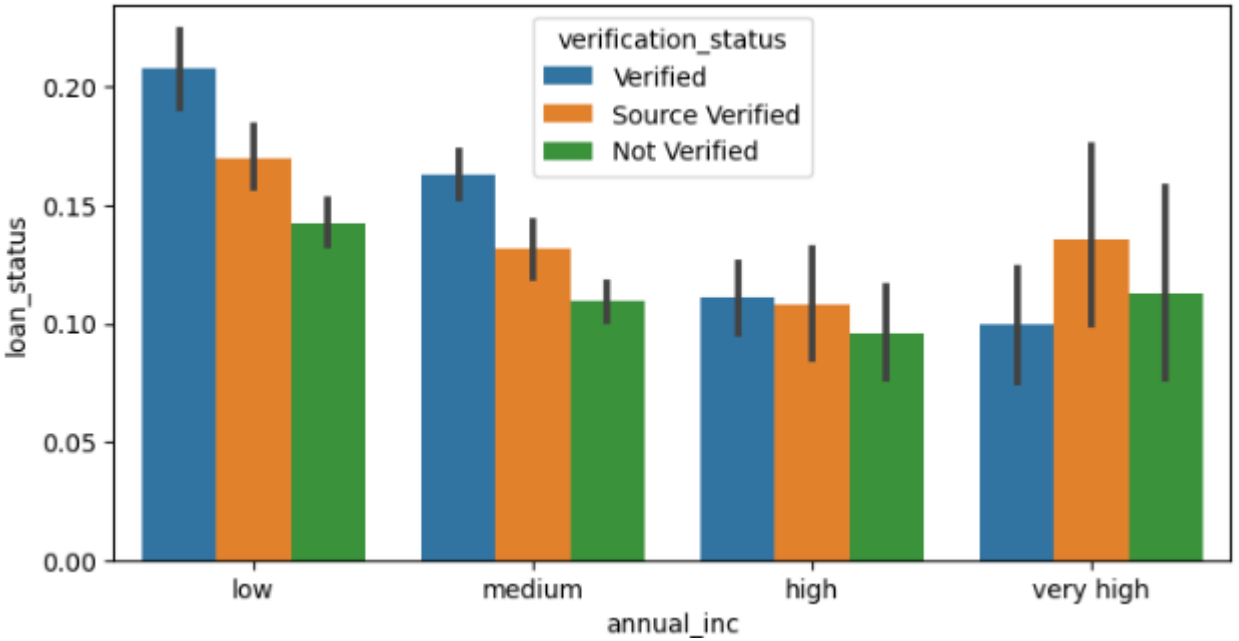
Home ownership does not have much affect on the loan status

Segmented Bivariate Analysis



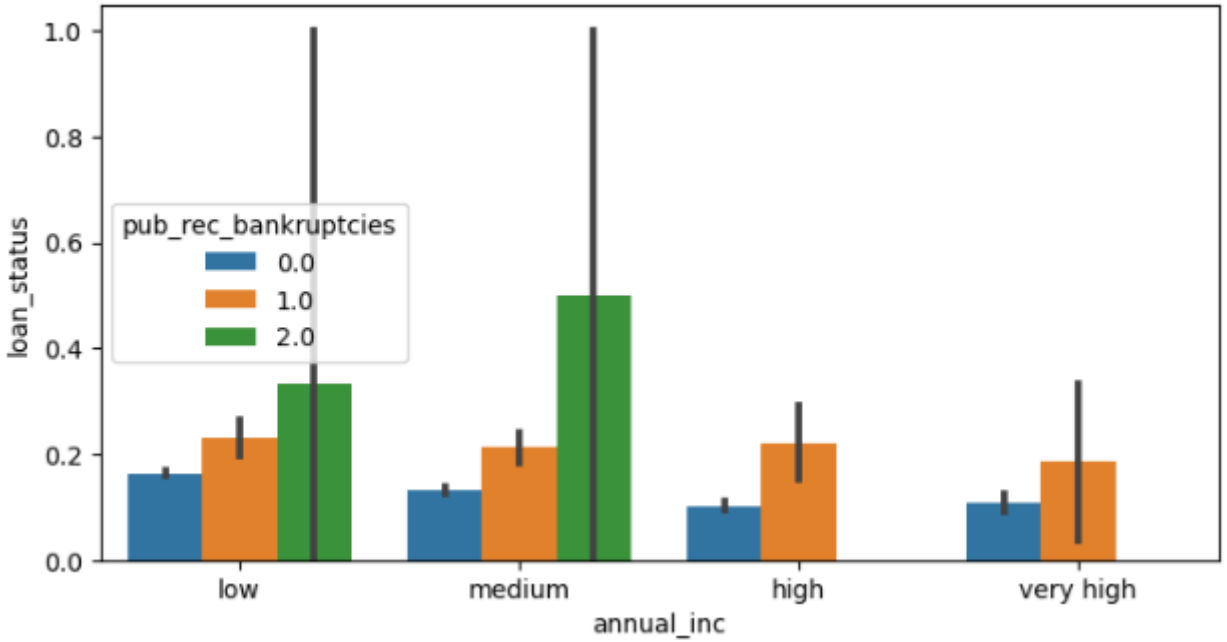
ANNUAL INCOME & LOAN TERM vs LOAN STATUS

Across all income groups, there is a sharp increase in default rate from 36 to 60-month loan term



ANNUAL INCOME & VERIFICATION STATUS vs LOAN STATUS

Except for very high-income users, there is an antipattern where the verified users are having higher default rate than non verified users

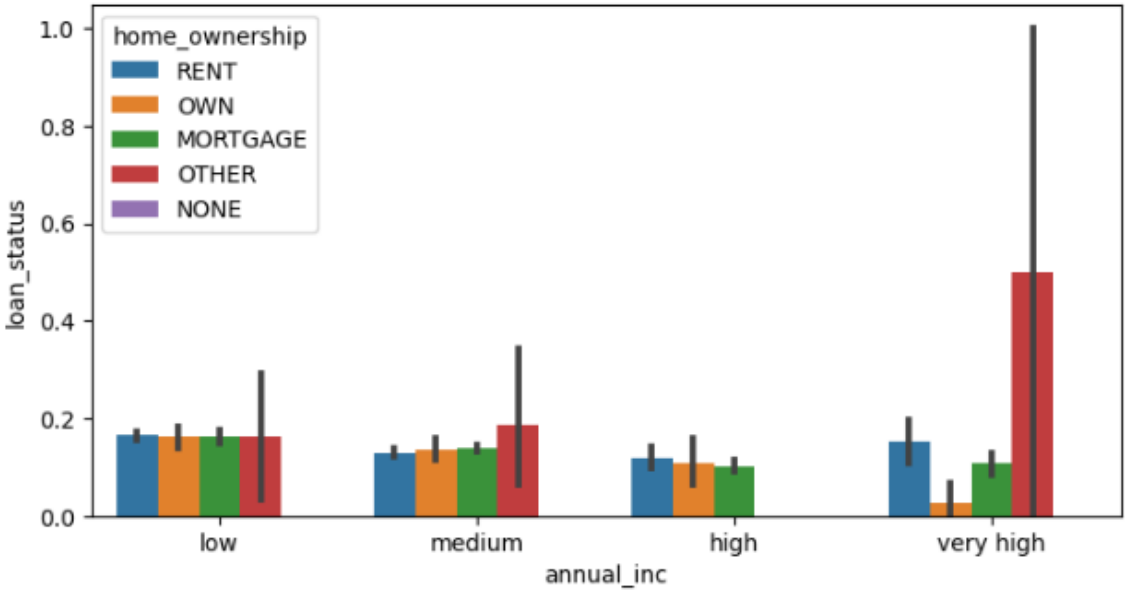


INCOME & BANKRUPTCY vs LOAN STATUS

Those users with 2 or more public record bankruptcy has higher than normal default rate

INCOME & HOME OWNERSHIP vs LOAN STATUS

Except for very high-income group home ownership does not reveal any patterns. However for very high income group, ‘Other’ category is of importance



Recommendations

- There is a sharp increase in default rate from 36 to 60 months loan terms. Introducing an intermediate loan term may be useful
- Improve customer verification process as data shows that verified customers are defaulting more than the non verified except for very high-income groups.
- Focus on verticals with better repayment status like wedding, credit card, home improvement
- Users with 2 or more public record bankruptcies are high risk. Such customers should be handled with due diligence
- For very high-income group, there is a high concentration of defaulting for ‘other’ type of home ownership. This may reveal something useful