Loan Lending Club

Problem Statement

Consumer finance company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

The company wants to understand the driving factors behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

Analysis Approach

- To identify the driving factors, we have analyzed the past loan data for all loans issued through the time period 2007 to 2011.
- To start with the data analysis, the data set had to be cleaned and prepared for the analysis, then analyzed and visualized.
- Below steps were performed to draw the conclusion

    1. Columns with more than 50% of null values were dropped.

    2. Columns like emp_title, url etc were dropped as these columns do not add value to the data analysis.

    3. Missing values were imputed so that all the NULL and NAN values were removed.

    4. Once the data set was cleaned, univariate, bivariate and mulitvariate analysis were performed

    5. Visualization techniques were used to draw charts and graphs

Data Cleaning in detail

- Multiple columns with more than 30% of null or missing values were dropped as imputing values to high percentage of missing values is not advisable. Columns like tot_hi_cred_lim,total_bal_ex_mort,total_bc_limit,total_il_high_credit_limit etc. which had more that 50% of null values were dropped.
- Descriptive columns like url, employee_title, loan description and idenitity columns like id, employee_id were dropped as these columns do not add value to data analysis. Only columns like loan_status, interest_rate, loan_amount etc were retained using which driving factors could be determined.
- Default value was imputed to few of the numeric column where values were NAN. Values were imputed keeping in mind that the new values inserted do not adversely impact the data analysis.

Univariate Analysis

- Univariate Analysis is analyzing a single column of data to infer outcomes.
- In our analysis we are interested only in defaulted loans.
- Loan_status column was considered to check how many loans were fully paid and how many were defaulted.
- From this analysis, it was inferred that over 30,000 loans were fully paid and just over 5000 loans were defaulted.

Bivariate Analysis

- Two columns are analyzed in Bivariate Analysis to visualize the effect of one value on another.
- Multiple columns were analyzed against loan_status column to visualize the effect of different parameters on defaulting a loan.
- Columns like loan_grade, loan_sub_grade, Term or the duration of the loan, employee_experience, purpose of the loan etc. were analyzed against the loan_status to visualize what factor drive the loan towards defaulting.
- Bar graphs, Histogram, line graphs were used to visualize the data

Bivariate Analysis (Continued..)

- Interest rate Vs Loan_status
  I. The effect of Interest rate of the loan, on loan defaulting was analyzed
  II. Interest rates were divided in 4 buckets as 5-10, 10-15, 15-20, 20-25
  III. It was inferred that loan with interest rate between 20-25 has the highest chances of defaulting
  IV. Loans with lower rate of interest between 5-10 has less chances of defaulting.
  V. There is a steady growth in the changes of defaulting with increase in interest rate. So higher the interest rate, the more chances of loan getting defaulted.

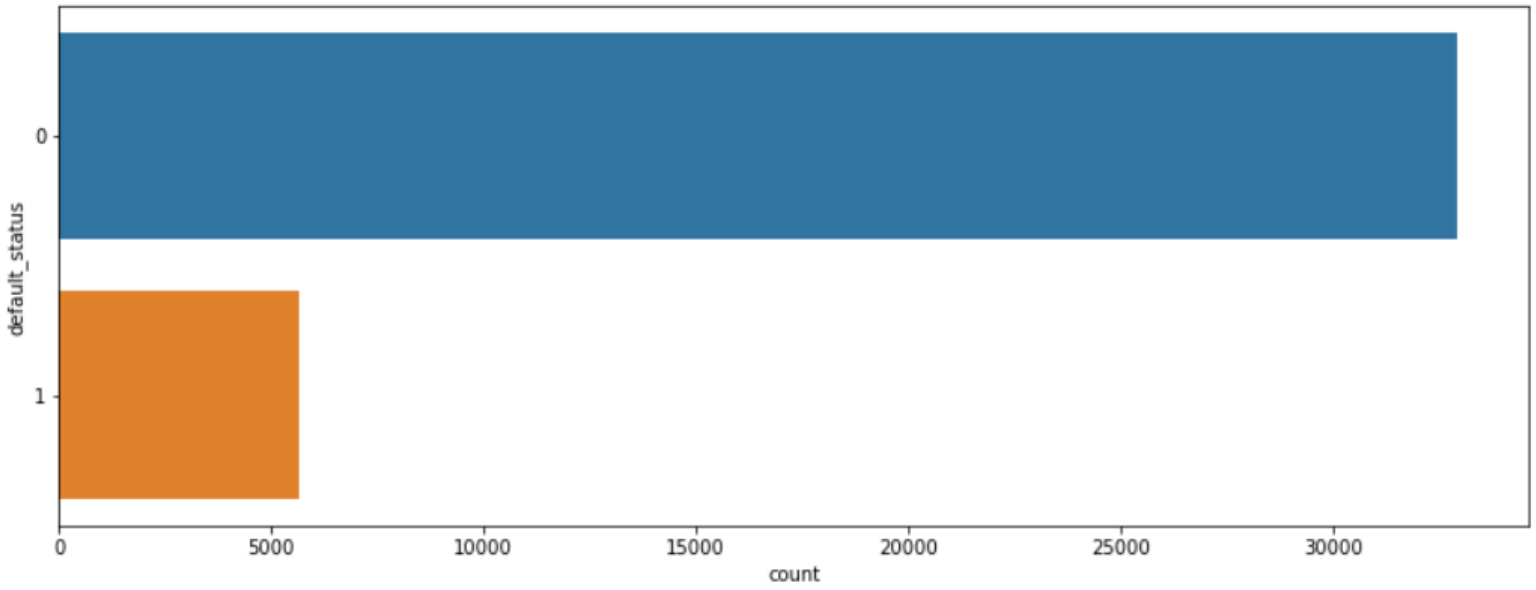
Bivariate Analysis (Continued..)

- purpose of loan Vs loan_status
  i. Purpose of loan column was analyzed against loan_status column to visualize loan taken for what purpose are defaulted the most
  ii. Loans that are taken for 'Small Business' has the highest changes of defaulting
  iii. There is no significant difference in the chances of defaulting between the other purpose of loans
- Loan_grade Vs loan_status
  i. Loan_grade was analyzed against loan_status column to visualize what grade of loans are bound to default.
  ii. Loan_grade 'A' is least to default
  iii. Loan_grade 'F' and 'G' has highest chances of defaulting

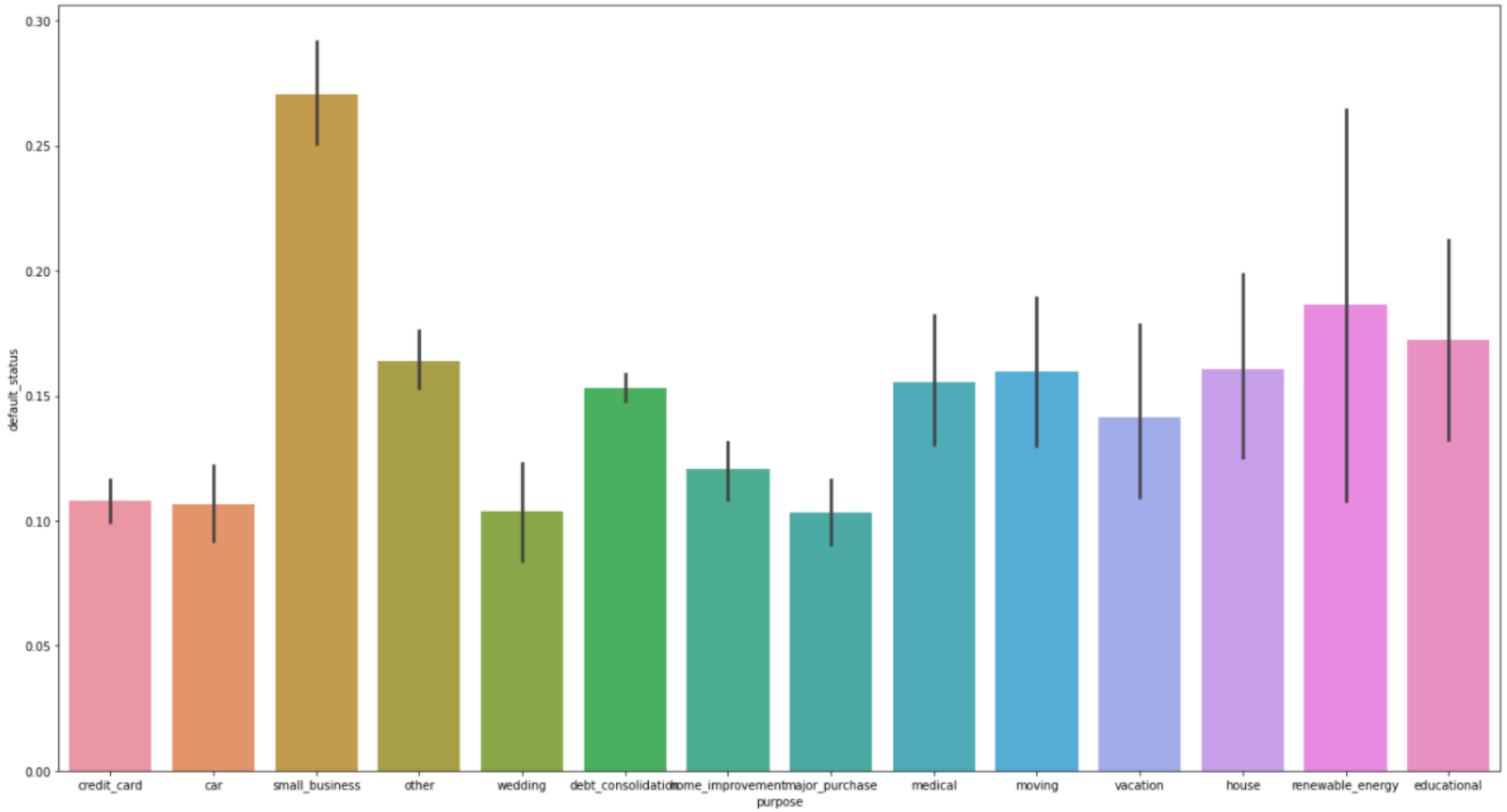
Multivariate Analysis

- The effect of multiple columns were analyzed against loan_status.
- The influence of loan amount and purpose of loan on the status of loan was visualized .
- The loan amounts were divided into different buckets of 0-5000, 5000-10000, 10000-15000, 15000-20000, 20000-25000, 25000+
- All the loan that fell into the each bucket were further divided based on the purpose of loan and were plotted
- Loan with 'Other' Purpose and loan amount greater than 25000 has the highest ration of defaulting.

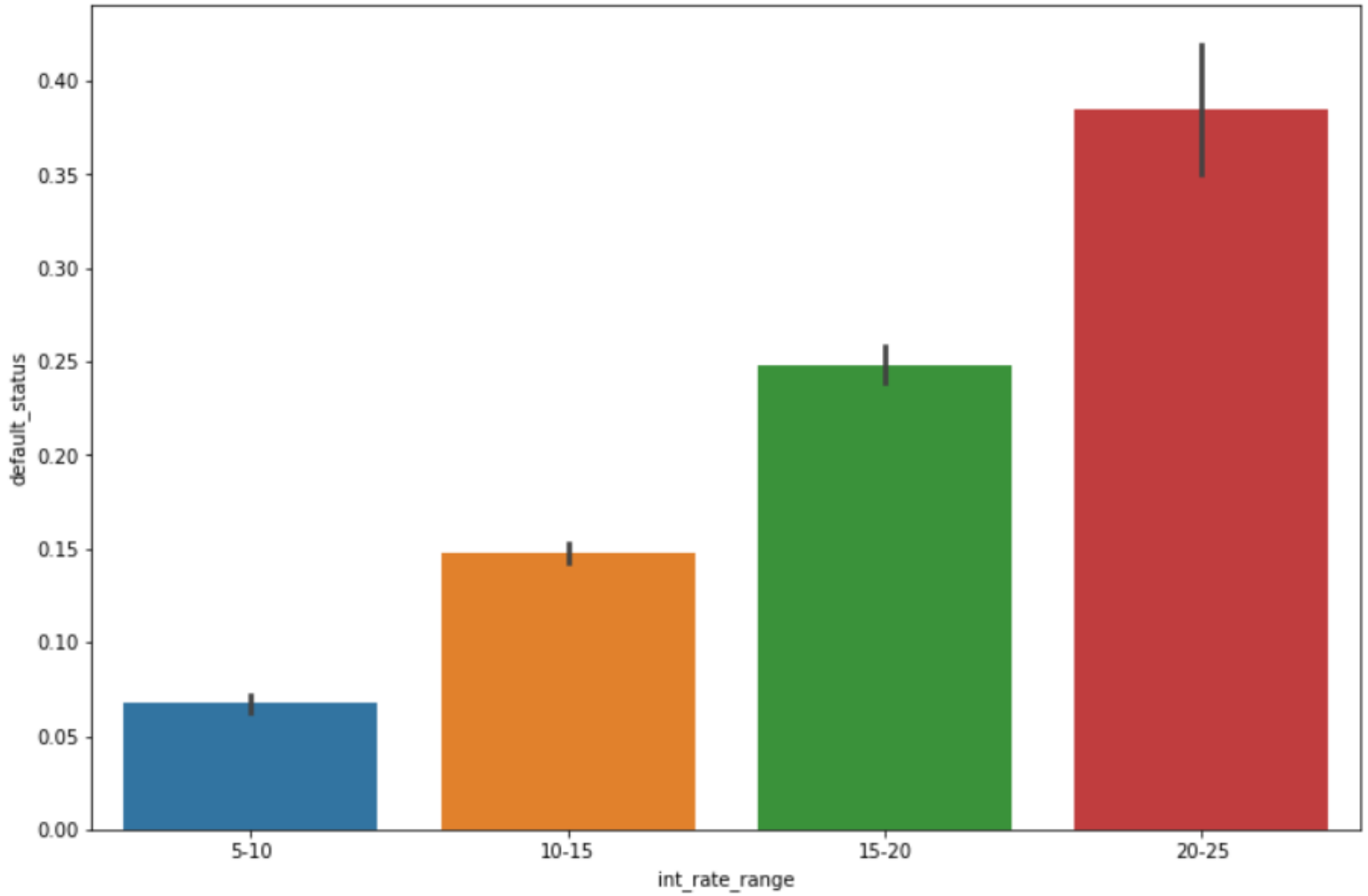
Visualization



14.6% of loans are Charged off loans
1- Fully Paid 0-Defaulters

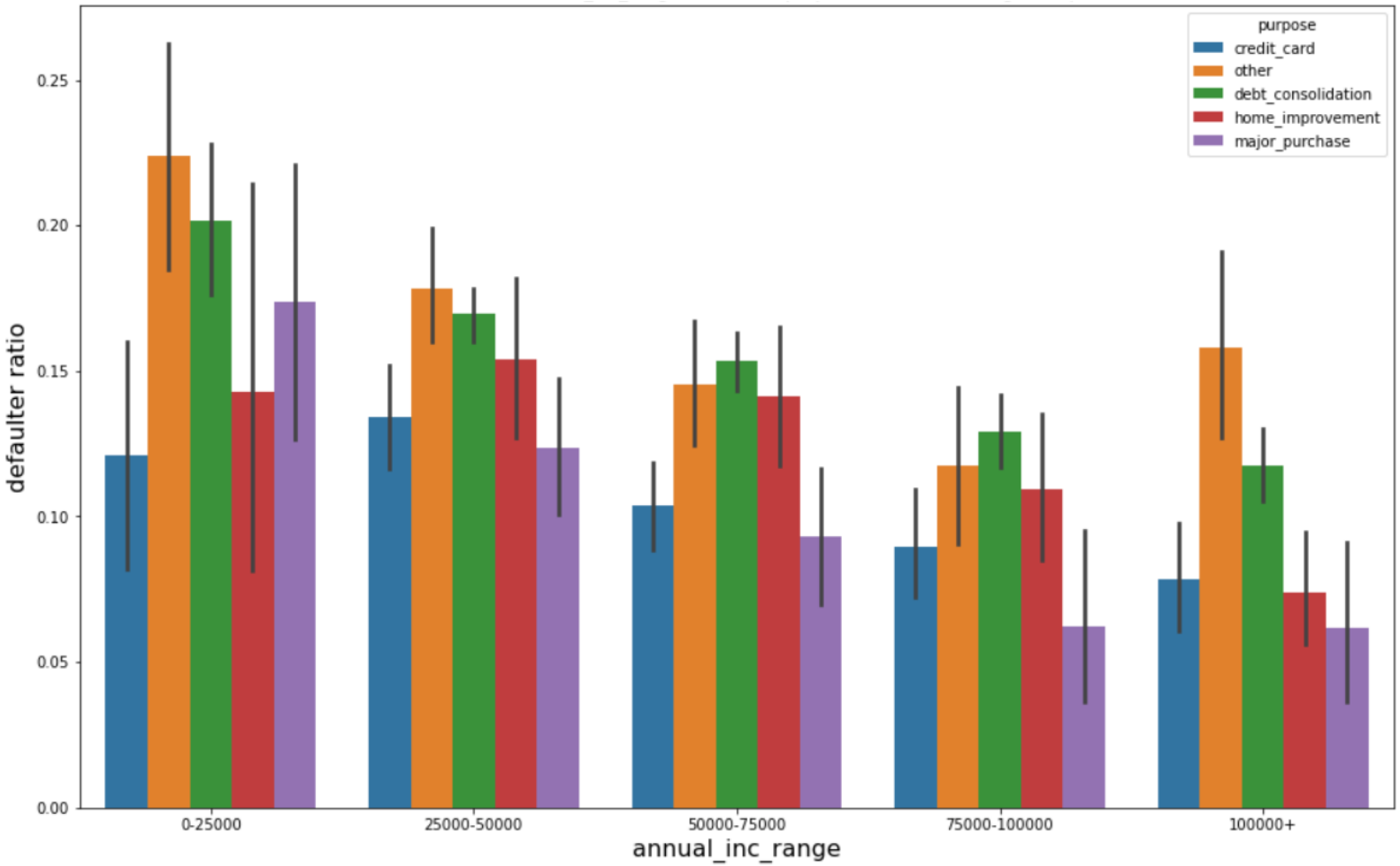


Loans that are given for ‘Small Business’ have high chances of getting Charged off



Higher the interest rate of the loan, higher the chances of getting Charged off

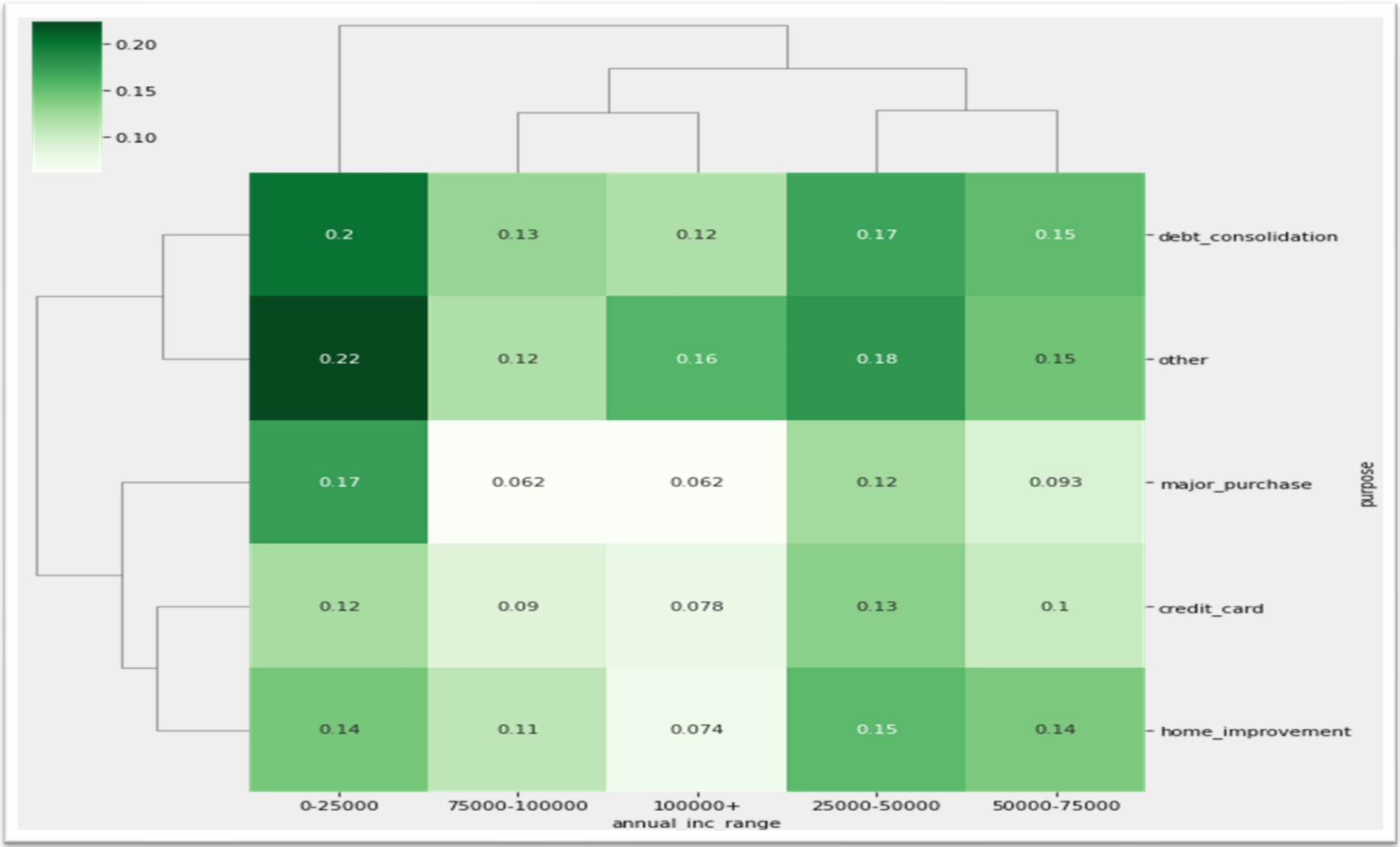
Debt_consolidation ,Credit_card ,Home_improvement ,Other ,Major_purchase cover upto 80% of loan purposes to the Lending Club Business, analysing on these purposes will be helpful in reducing the number of defaulted loans.



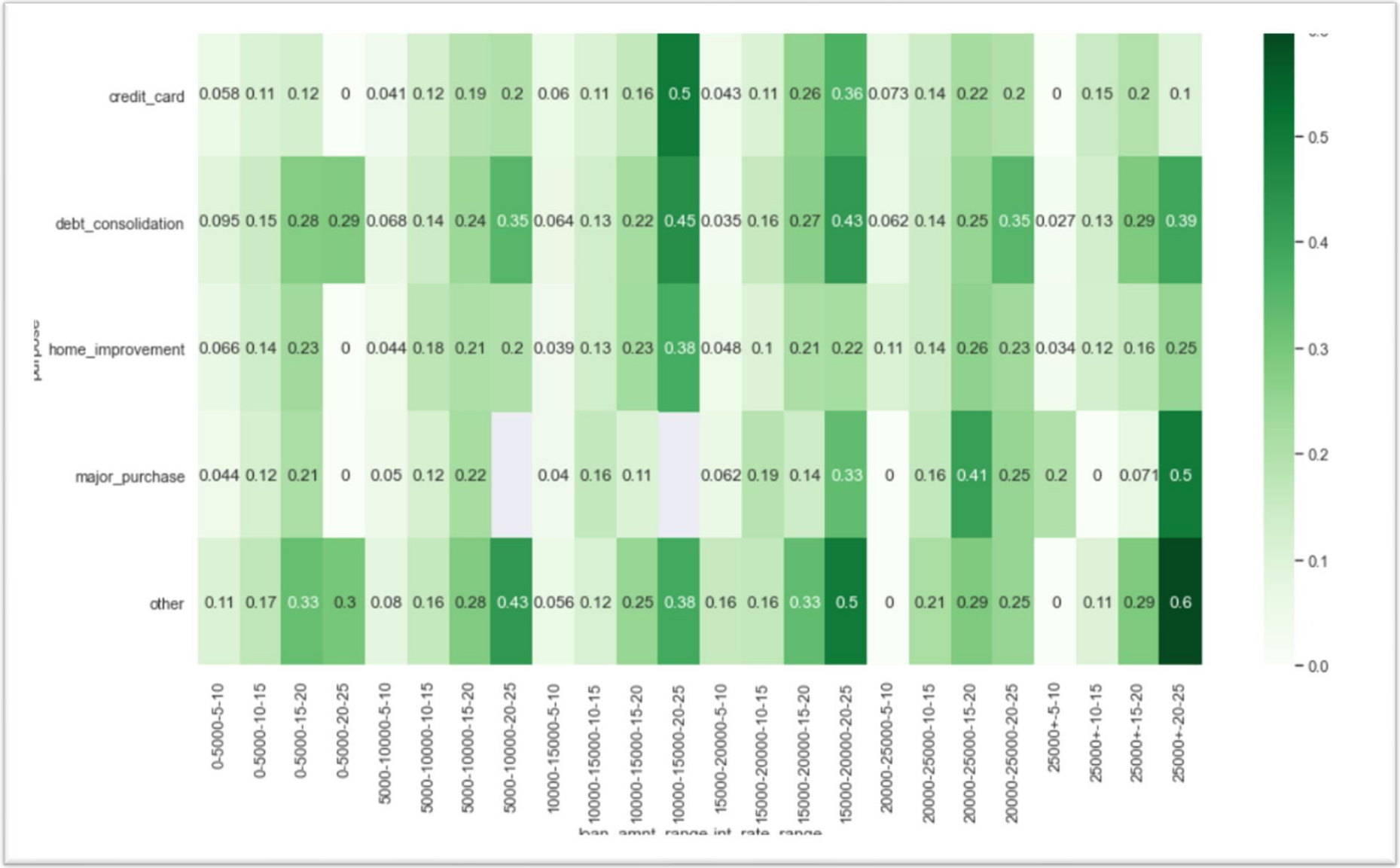
Loans taken for 'Other' purpose by the borrowers with annual income of less than 25000, default the most

Corelation between purpose,annual_income_range with default status is well distributed from the above Chart

We see Debt_consolidation and Other are corelated , so we use same methods to avoid risky customers for both



Other loan purpose with loan_amnt - range 20000+ and 20-25 interest range have the highest defaulter correlation and followed by other combinations



Conclusion

- After the data clean up and analysis, below are few of the insights that were visualized
  - Around 14% of all the loans get defaulted
  - Higher the loan interest rate, higher the chance of loan getting defaulted
  - There is higher chance of loan getting defaulted if the loan is taken for the purpose of ‘Small Business’ or ‘Debt Consolidation’
  - Chances of loan default is more if the income of the borrower is less than 25000 and the loan purpose is ‘Other’.
  - Longer the term higher the Defaulters
  - Dti_Range, defaulter rate increases with increase in dti ratio.
  - Higher the Grade , higher the Defaulters
  - Lower Income customers are likely to get defaulted