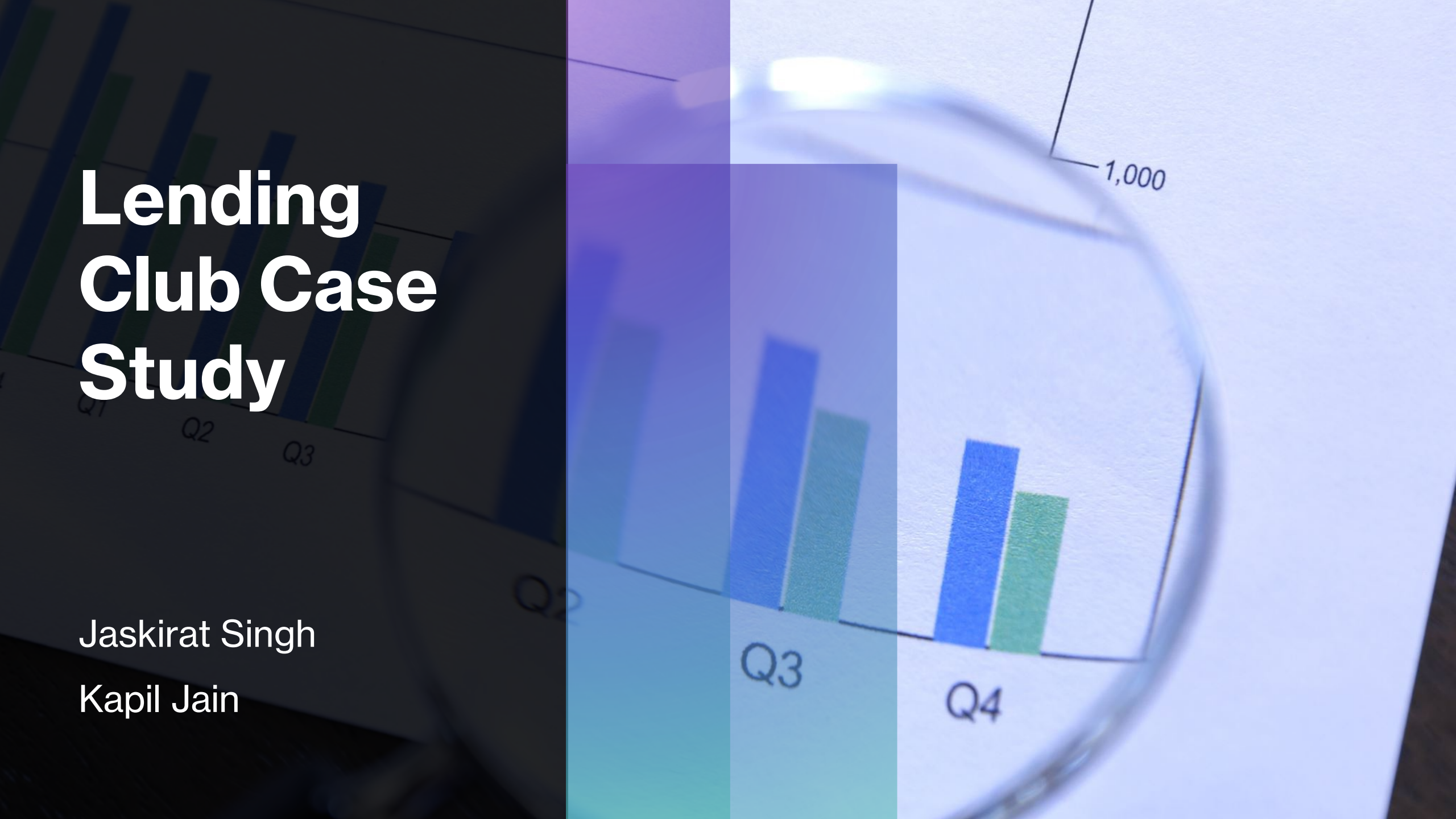
Lending Club Case Study



Jaskirat Singh

Kapil Jain

Lending Club: EDA Case Study

- What is Lending Club?
- Lending Club is a marketplace for personal loans that matches borrowers who are seeking a loan with investors looking to lend money and make a return.

When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company

- 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.
- In other words, the company wants to understand the 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 its portfolio and risk assessment.

When a person applies for a loan, there are two types of decisions that could be taken by the company:

Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:

1. **Fully paid:** Applicant has fully paid the loan (the principal and the interest rate)
2. **Current**: Applicant is in the process of paying the installments, i.e. the tenure of the loan is not yet completed. These candidates are not labeled as 'defaulted'.
3. **Charged-off**: Applicant has not paid the installments in due time for a long period of time, i.e. he/she has **defaulted** on the loan

Objective:-

- We have a loan dataset of the lending club, in which there are details of people who took the loan.
- The objective is to identify predictors of default so that at the time of loan application, we can use those variables for approval/rejection of the loan.
- So that we can reduce the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

Case Study Approach:-

We will approach this in **4 stages**:

1. Data understanding
2. Data cleaning
3. Data Analysis
4. Recommendations

Data Understanding

- There are 111 columns (or variables) out of which 74 are of datatype float, 64 of int, and 24 are objects.
- In total we have 39717 rows.
- This is what we have at the start of this exercise. As we clean and analyze the dataset, the columns and rows will vary based on the operations applied.
- Many columns have no values and some rows also have null values.
- There are broadly three types of variables –
    1. Those which are related to the applicant (demographic variables such as age, occupation, employment details, etc.),
    2. Loan characteristics (amount of loan, interest rate, purpose of loan etc.) and
    3. Customer behavior variables (those which are generated after the loan is approved such as delinquent 2 years, revolving balance, next payment date etc.).
- The objective is to identify predictors of default so that at the time of loan application, we can use those variables for approval/rejection of the loan.
- Now, the customer behavior variables are not available at the time of loan application, and thus they cannot be used as predictors for credit approval.
- The ones marked 'current' are neither fully paid not defaulted, so get rid of the current loans. Also, tag the other two values as 0 or 1 to make your analysis simple and clean.

Data cleaning

- Based on previous stage outcome and generic data cleaning guidelines we will be performing the following set of operations on the dataset:

1. Remove Customer behavior Variables like delinq_2yrs, earliest_cr_line, Inq_last_6mths, etc.
2. Remove columns that are not relevant like url, zip_code, title, etc.
3. Filter out rows having loan_status as current.
4. Remove all columns which have a higher percentage of missing values.
5. Fix Datatypes.
6. Remove Outliers.

After doing all the above steps, now we have data that can be analyzed. Now we have 21 columns and 38191 rows.

Data Analysis

We have cleaned up the data to a level and hence will move on to the Analysis stage. We did the following analysis over the dataset:-

1. Univariate Analysis: In this, we are going to analyze a single column and derive patterns/insights which will help further in our analysis.
2. Segmented Univariate Analysis: In this, we are going to analyze a single column in subsets which will be very useful as it can show the change metric in pattern across the different segments of the same variable.
3. Bivariate Analysis: In this, we are going to check the correlation among a group of two variables at a time to root out similar variables and also extract combinations that help us solve the business problem

- **Outcome - Univariate Analysis:-**
- Since each of these variables has some value range or specific categories which have a higher number of applications or variables like int_rate which are pretty common.
- As the data itself is for approved applicants hence to conclude these categories are the highest applicants will be inconclusive but we can say that once the existing approval mechanism filters out the applicants then among those these categories have the highest application submission-approval.
- We observed that:
- Loan amounts 10000, 5000 are the most sought ones
- Among interest rates 7.5-8% is given to a large number of applicants and then there is a drop and then >10% interest again it picks up. It seems some category is having resulting lower interest rates granted.
- Majority loans have installments in the range of 175-200
- Quite a major chunk of loan applicants are in the range of 40k-70k as annual income
- We will be using these to further target our analysis in segmented univariate and bivariate analysis so that we can discover patterns for applicant default and share recommendations.

Outcome - Segmented Univariate Analysis:-

We observed that :

- Majority applicants have loan grades among A,B, and C
- US Army, Bank of America employees are major applications or they are the ones that get approved better than others as these are trusted organizations.
- Majority applicants tend to be either newly employed or employed >= 10 years.
- Majority of applicants have mortgaged or rented accommodations.
- Income status is not a significant driving variable for a loan. Maybe since application approval also takes other factors into account.
- Quite a majority of applications have debt_consilation as the purpose of a loan.
- December seems to be a month with the highest loan approvals.
- 36 months is the most offered/selected loan repayment term.
- We will be using these to further target our analysis in segmented univariate and bivariate analysis so that we can discover the patterns for applicant default and share recommendations.

Outcome - Bivariate Analysis:-

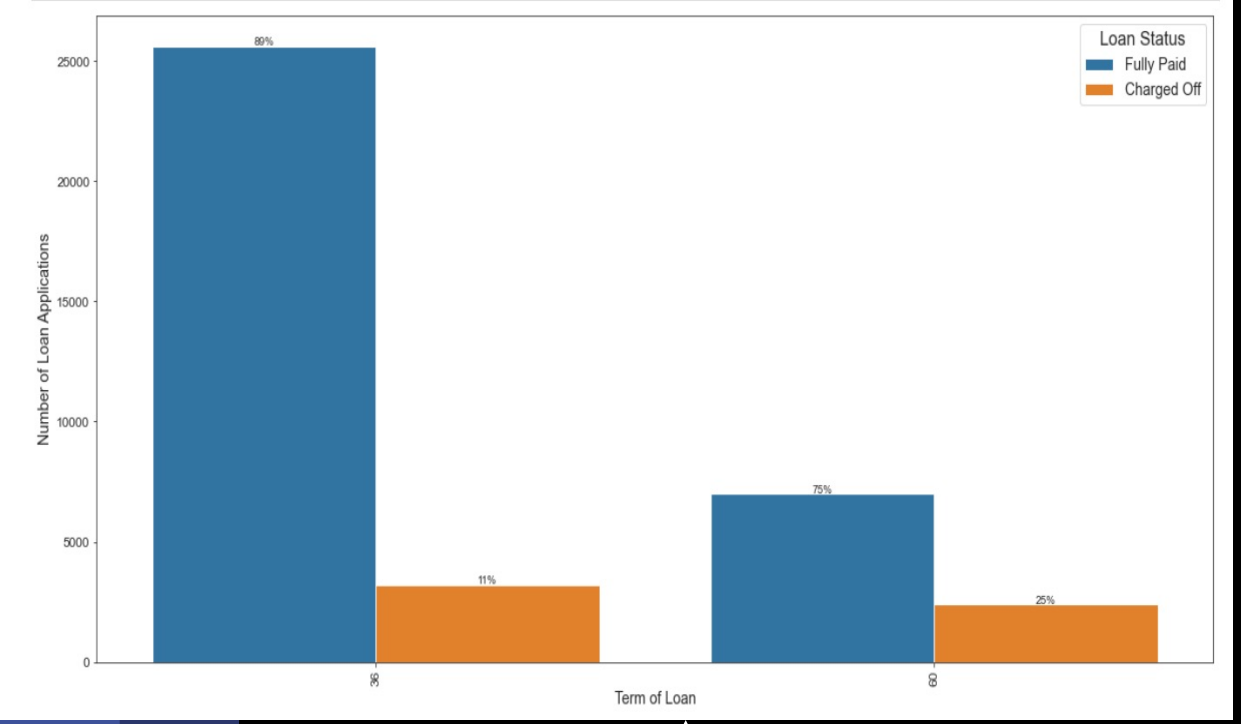
We observed that :

- Loans are taken for Small Business and Renewable Energy purpose has higher chances of defaulting
- Applicants having 60 months as term length have higher chances of defaulting on a loan
- 18% and greater interest rate results in more loan default as applicants seems to have a hard time paying higher interest rates
- Applicants with dti >20 have higher chances of default which is logical since they have more debt payments to be paid for a given amount of monthly income.
- Higher the amount, more the chances of default
- Higher loan grades (>C) although have less approved applications but have very high default especially G
- December seems to be the month of max approvals and hence the default.
- Chances of default Decreases if the applicant has higher income i.e. >= 100000
- Interest rate(int_rate) and grade are highly co-related hence only one can be considered when trying to screen whether a given application will default or not.

After doing Univariate, Segmented Univariate, and bivariate analysis. We found that the following are needed to be considered while approving loan applications:-

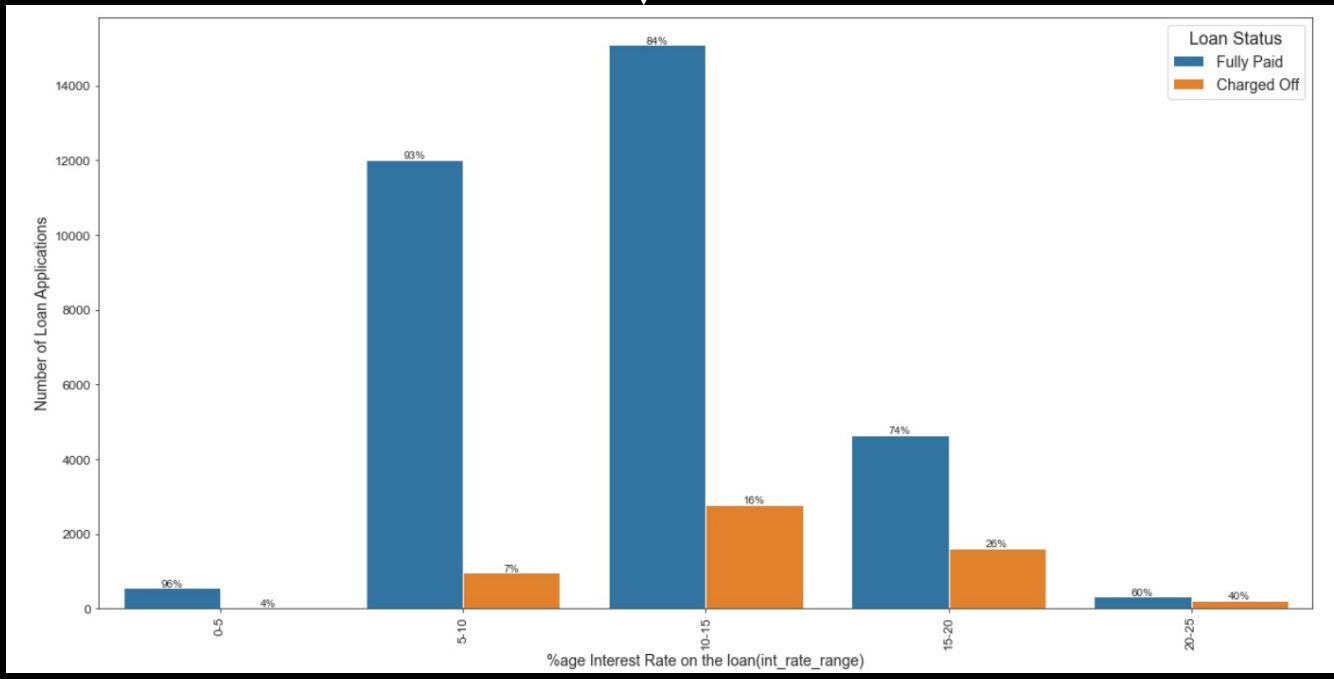
1. dti
2. int_rate
3. purpose
4. term
5. loan_amt
6. purpose
7. Grade
8. issue_d
9. annual_inc
10. emp_title
11. home_ownership
12. Installment
13. emp_length

Some charts of analysis



loan_status vs term
Applicants with **60 months** as term length have a higher chance of ***defaulting*** on a loan

loan_status vs int_rate
On binning, we are able to clearly see a range of **>18%** of interest rate resulting in more loan ***default***



Recommendations

We have following observation :

- More amount of loan should be given for wedding purpose since default rate is very low and a number of loans are quite less in this category, same is true for car loans.
- Renewable Energy loans applications are already less they should be completely avoided.
- Higher interest rate should be given to applicants with higher annual income as they have the capability to pay off loans as this has a negative correlation with dti since monthly income is greater
- The employer reputation is also a driver, hence applicants from reputed organizations like US Army, etc. should be given preference.
- Applications approved in December should be scrutinized more since they are having higher defaults.
- Higher interest rate and dti is also one driver of default
- Number of payments (term) is also a factor for default. The higher term means high default.
- Higher grades mean grade greater than C are applicants who are more defaulter in loan.