Lending Club Case Study

By Chaitanya Singh & Arjun Menon

Problem Statement

You work for a **consumer finance company** which specialises in lending various types of loans to urban customers. 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

The aim is 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.

Understanding the data

The dataset loan.csv contains information about past loan applicants and whether they ‘defaulted’ or not.

The aim is 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.

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:

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

Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

Data Cleaning

- Several Columns contained All-Null Values, since they provide no information, they were removed
- Other Columns with significant number of null values were handled individually
  - emp_title and emp_length specify the job title supplied by the borrower when applying for loan and the employment length respectively.
  - desc privides a description of the nature of the loan provided by the borrower. **However, since deriving any insights from this data would require sentiment analysis, we will drop it for now**
  - mths_since_last_delinq provides the number of months since the borrowers last delinquecy.
    - note that this is usually null for a large majority of any given sample.
    - This also makes it a good metric to keep an eye on, as past delinquents tend to regress.
    - **However, due to the lack of data in this field, we cannot come to accurate conclusions that best describe the given dataset. Hence, we drop this column as well.**
  - mths_since_last_record is the number of months since the borrowers last public record.
    - **This field has over 90% missing values, we may not be able to derive relevant information. We'll drop this column.**

The objective is to determine the predictive variables. Those that may impact the target variable Loan_Status

We notice that there are several types of variables

- those that are related to the borrower that can influence their overall probability of paying off the loan. These include employment details, age, etc.
- those that describe the nature of the loan, such as amount, interest, purpose, etc.
- and those that are generated after the loan is processed. Something like the next payment date, etc. Since these cannot be derived until after the loan is processed, they can be dropped.

Data Manipulation

- **int rate** defines the interest rate of the applied loan. This is expected to be a numerical value, yet it is defined as an Object.

  Since the the values are suffixed with a '%' symbol, we remove the '%' and convert it to the type float

- **term is defined as an object but it consists of numerical values**
- **Home_ownership** consists of several values of values called NONE, We can categorize them under the type OTHER
- We can derive the months and years of **issued month**

Identifying the target variable¶

The problem statement provides us with a dataset consisting of only approved loans in the past. They are then categorized as either *Fully Paid*, *Charged off* or *Current*.

So we can assume that users who have Fully Paid off their debts are preferrable and those that have been Charged off are red flags.

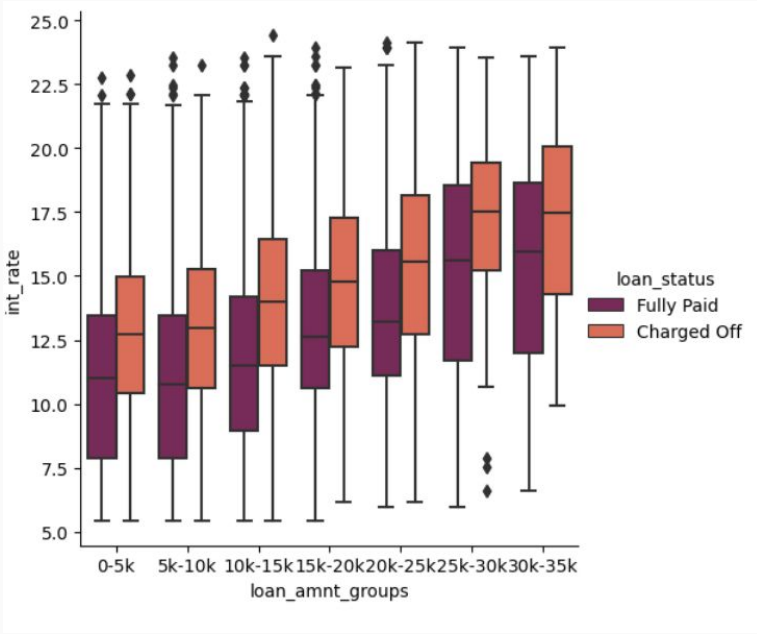
This gives us enough reason to choose this field as our target variable to identify potential defaulters.

Note. since borrowers who are currently paying off their loans (i.e. in the 'current' category) can't be classified as a potential defaulter or not, we can choose to ignore this field and drop it.

Variables of interest (Results of Analysis)

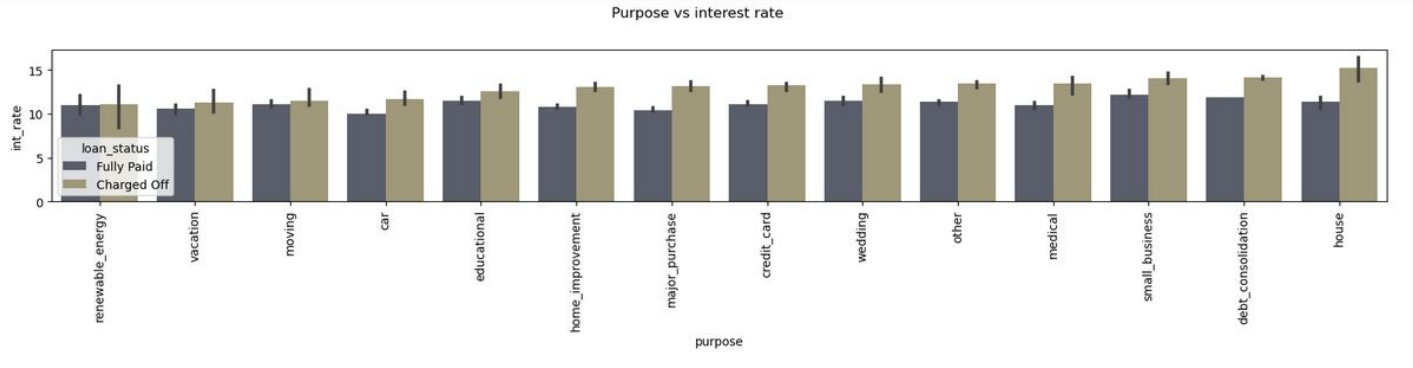
1. **Interest Rate**

   Generally, borrowers who apply for loans with a high interest rate tend to find it difficult to pay it back. This can be a major cause for defaulting.



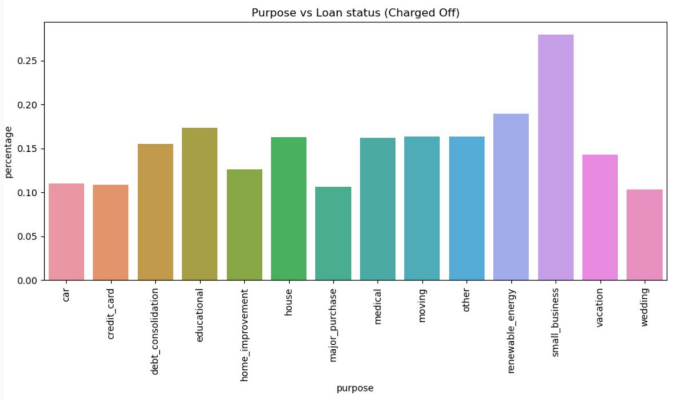
This can be observed when the purpose is pitted against the interest rate.

We notice that home loans with high interest rates tend to mostly default



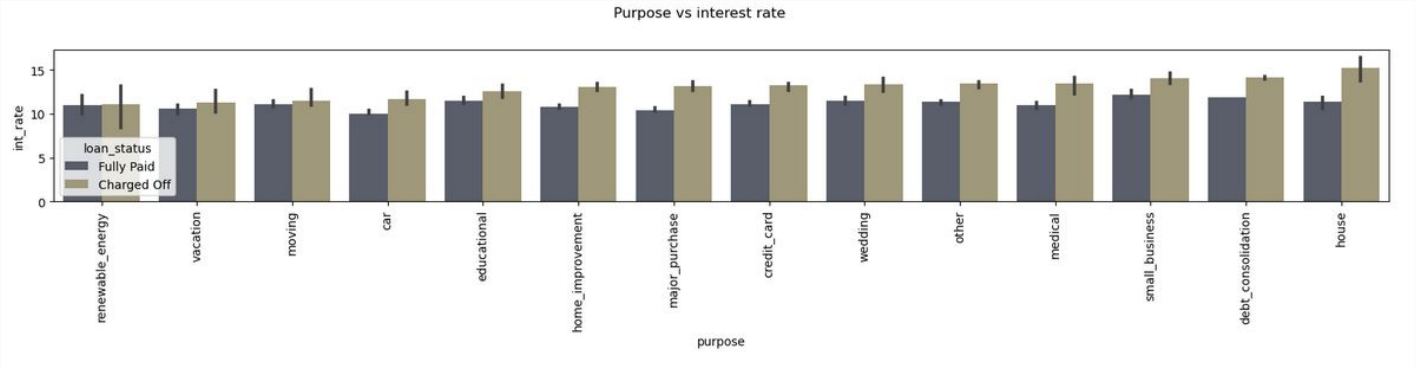
2. Purpose

The purpose of the loan is an important variable to determine if the risk is worth it.



Generally, a higher percentage of borrowers apply for loans in the small businesses category

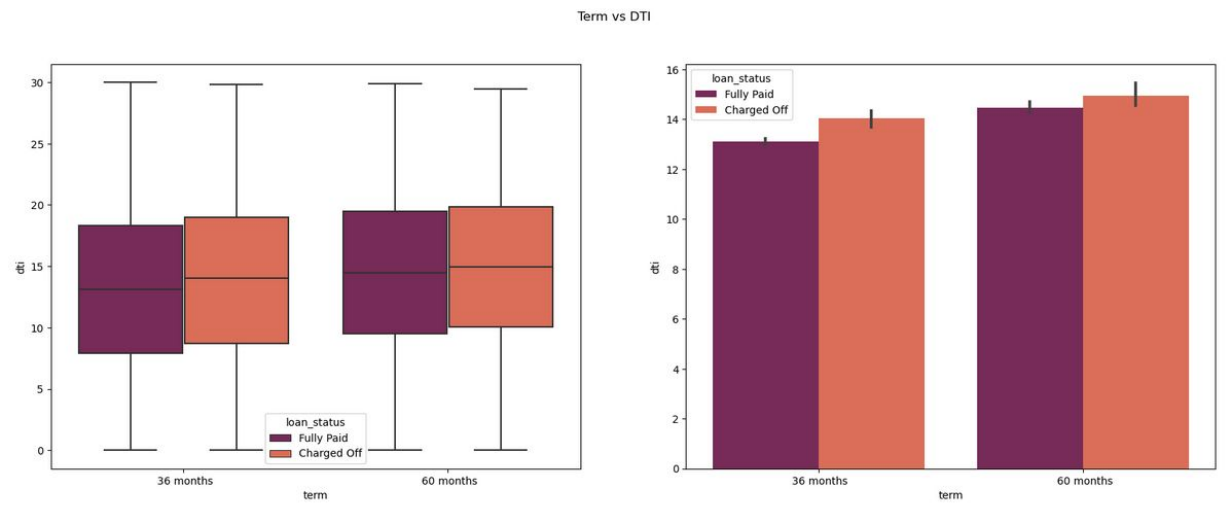
- Small Businesses also tend to default more often.



3. DTI

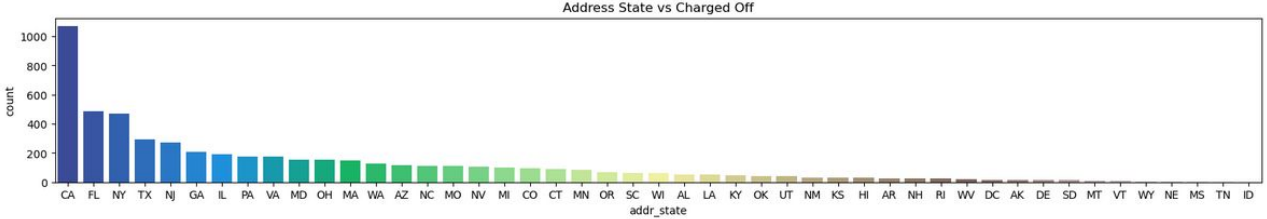
Borrowers in the terms

- 36 months: With a dti of over 13, tend to get charged off
- 60 months: With a dti of over 14, tend to get charged off



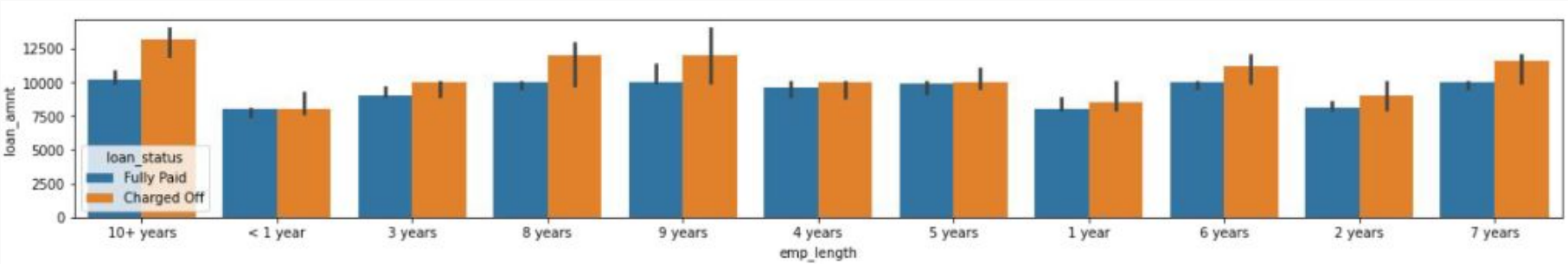
4. Address

More number of borrowers are observed in the states of CA, FL, NY and TX.

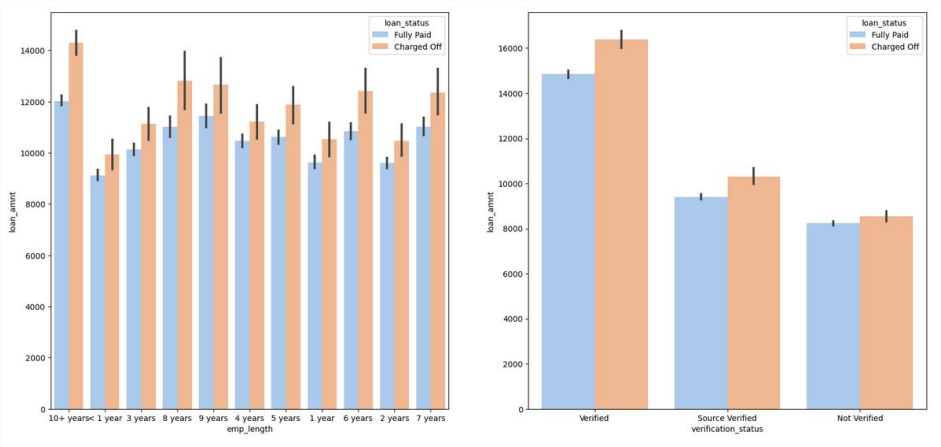


5. Loan Amounts

- Borrowers with higher loan amounts tend to default more often



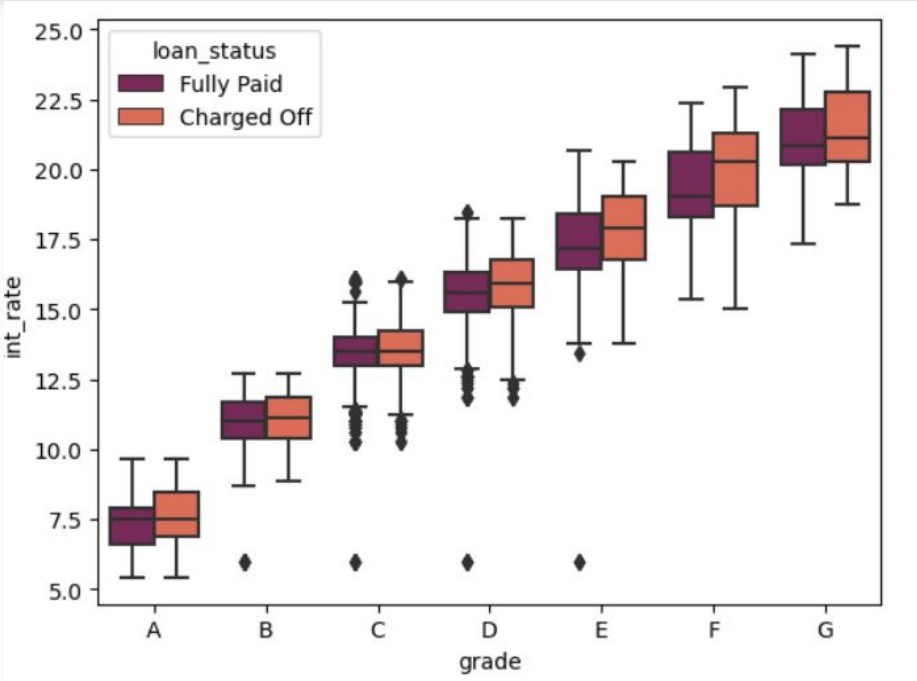
- Borrowers with more than 5 years of employment duration are preferred. They are given higher loan amounts



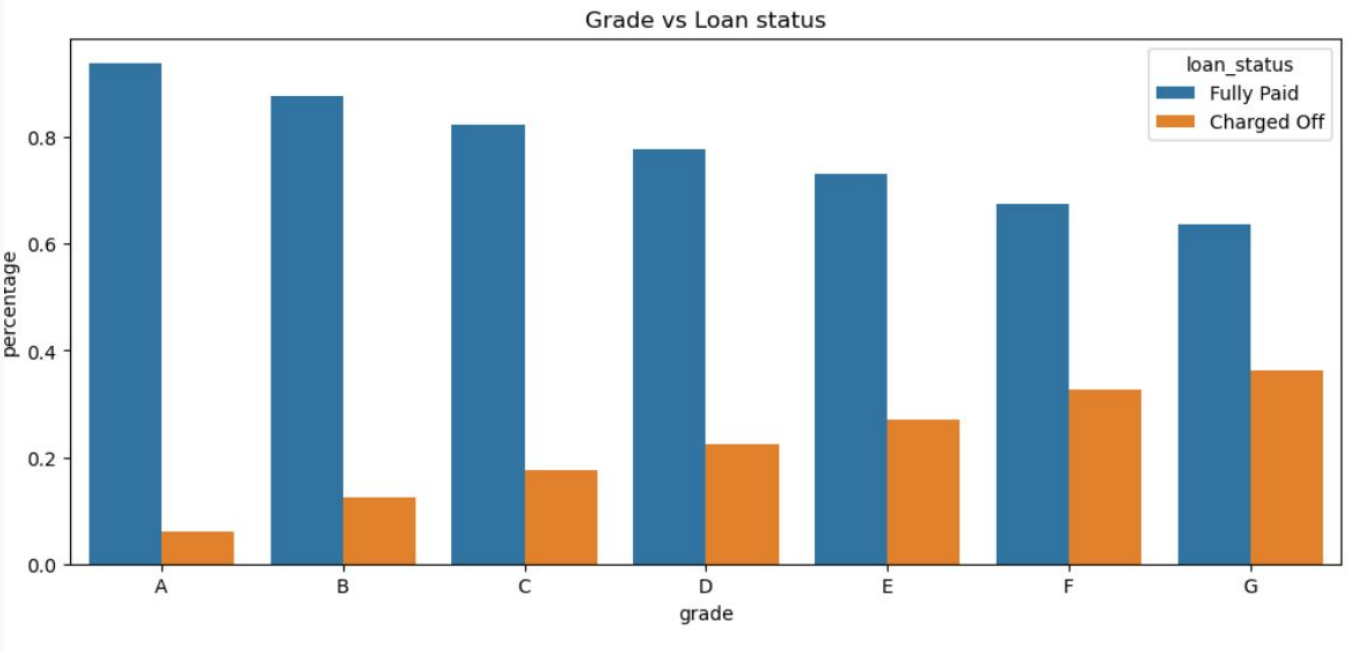
- Also, verified loan applications tend to have higher loan amount. Which might indicate that the firms are first verifying the loans with higher values

6. Grades

- Lower grades have higher interest values.



- we notice that as the grade decreases, the percentage of charged off borrower increases.



Conclusion

Key Observations:

- Borrowers who have request for higher loan amounts have to pay off their debts at a higher interest rate tend to default
- Generally borrowers who apply for lower grade loans tend to default.
- Small Businesses tend to get charged off more often these are generally applicants who have taken a loan for small business and the loan amount is greater than 14k
- When employment length is 10yrs and loan amount is 12k-14k, they tend to default.
- Borrowers belonging to grade lower than E, are more likely to default.
- Purposes involving categories other than ‘home’ or ‘renewable energy’, show more signs to default. Hence, extra caution must be taken.

Other Observations:

- Borrower from CA, FL or NY states have a higher chance to default
- Borrowers opting for 60 month term are more likely to default than 30 month terms
- Employees with longer working history got the loan approved for a higher amount

Thanks!