Lending Club Case Study

By :-

Kunal Chitale

Eshwar B

Introduction

- Solving this assignment will give you an idea about how real business problems are solved using EDA. In this case study, apart from applying the techniques you have learnt in EDA, you will also 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.

Business Understanding

- 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 data given below 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.

- In this case study, you will use EDA to understand how **consumer attributes** and **loan attributes** influence the tendency of default.

Business Understanding (Contd.)

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

1. **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 instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
    3. **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
2. **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)

Business Objectives

- This 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). 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.
- 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 utilise this knowledge for its portfolio and risk assessment.
- To develop your understanding of the domain, you are advised to independently research a little about risk analytics (understanding the types of variables and their significance should be enough).

Data Understanding

- The dataset contains the complete loan data for all loans issued through the time period 2007 t0 2011.

Results Expected

1. Write all your code in one well-commented Python file; briefly mention the insights and observations from the analysis
2. Present the overall approach of the analysis in a presentation:
    1. Mention the problem statement and the analysis approach briefly
    2. Explain the results of univariate, bivariate analysis etc. in business terms
    3. Include visualisations and summarise the most important results in the presentation

- You need to submit one Ipython notebook which clearly explains the thought process behind your analysis (either in comments of markdown text), code and relevant plots.

Approach

- The assignment is divided in the two parts
  - ETL – Format the data from CSV with the correct data types
  - Visualization – Plot the Charts to understand the nature of data and outcome

ETL Tasks

- In the CSV file the below fields are mentioned as String, out of these columns some areincorrectly mentioned as String, for ex : Term, int_rate, revol_util etc. , these need to be converted to Number
  - Term
  - int_rate
  - Grade
  - sub_grade
  - emp_title
  - emp_length
  - home_ownership
  - verification_status
  - revol_util
  - initial_list_status
  - last_credit_pull_d
  - application_type
  - issue_d
  - loan_status
  - pymnt_plan
  - url
  - Desc
  - Purpose
  - Title
  - zip_code
  - addr_state
  - earliest_cr_line
  - last_pymnt_d
  - next_pymnt_d

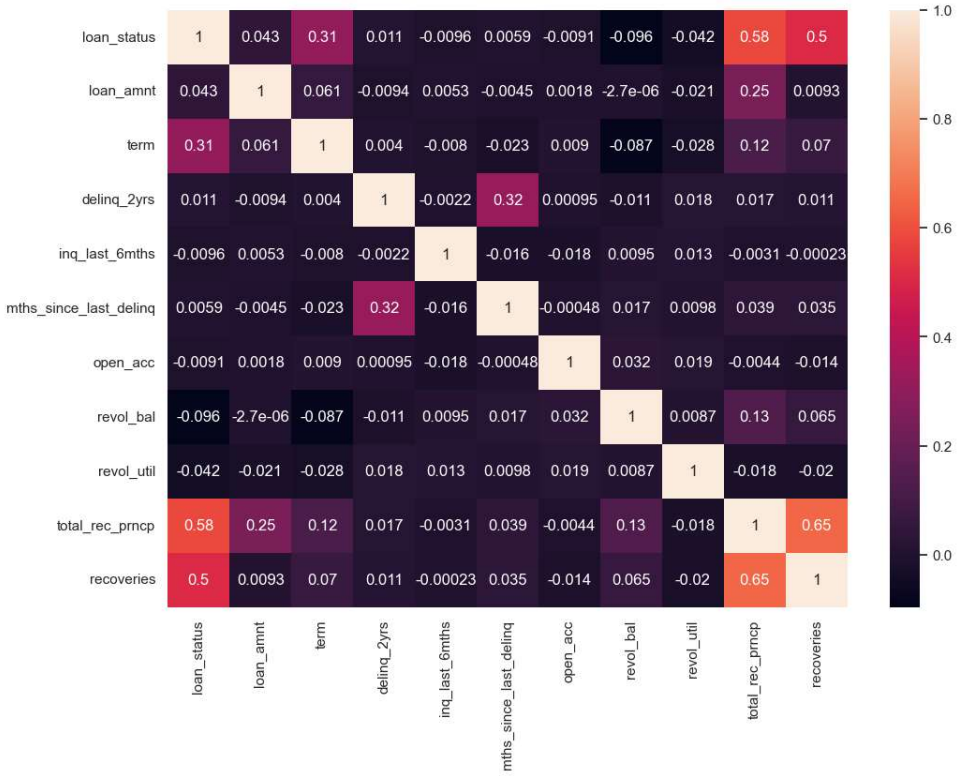
ETL Tasks (Contd.)

- Columns Charge Off, Fully Paid, Current added to Dataframe
- Loan Status Column bifurcated to 3 different columns

Visualization

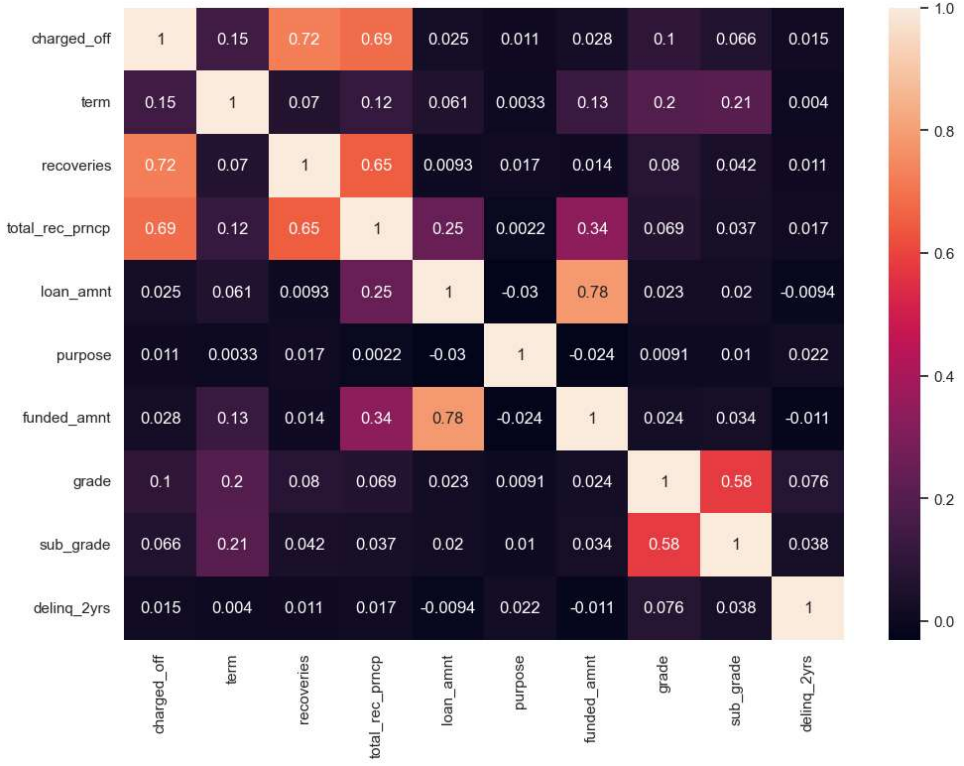
- Find the relationship between each attribute amongst itself
- Factorize the data to find the relation
- Plot the Factorized Data on Heat Map

- Strong Relation is seen in between Loan Status and
  - Recoveries
  - Total Principal Received till Date
  - Term



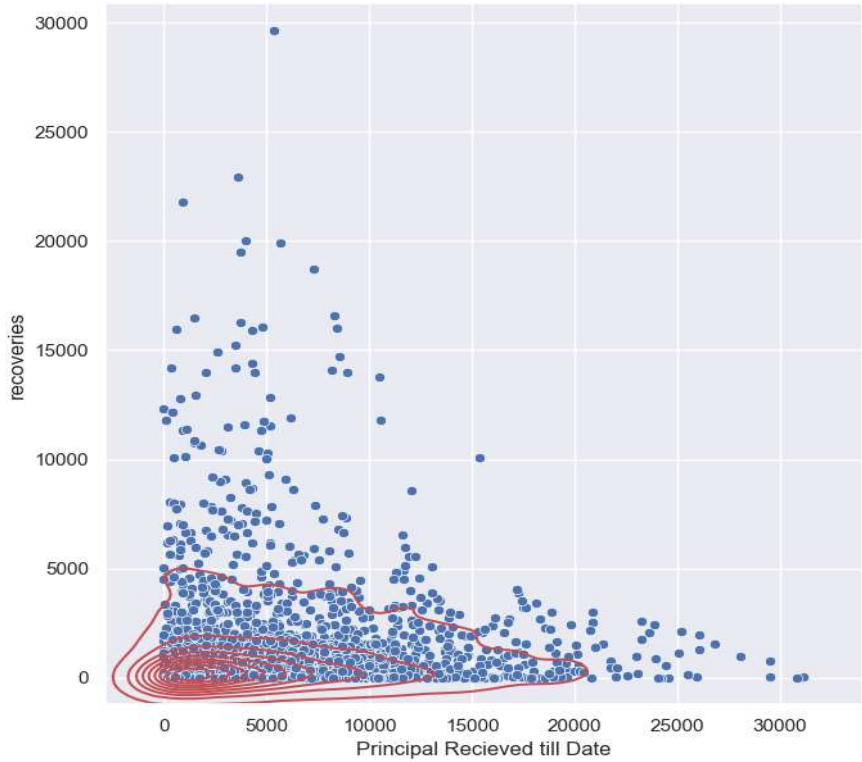
Visualization (Contd.)

- Data Relation between Charged Off Loan and other Attributes Shows
  - ➢ High Relation between Recovery and Paid Principal Amount
  - ➢ Medium Relation between Grade and Term



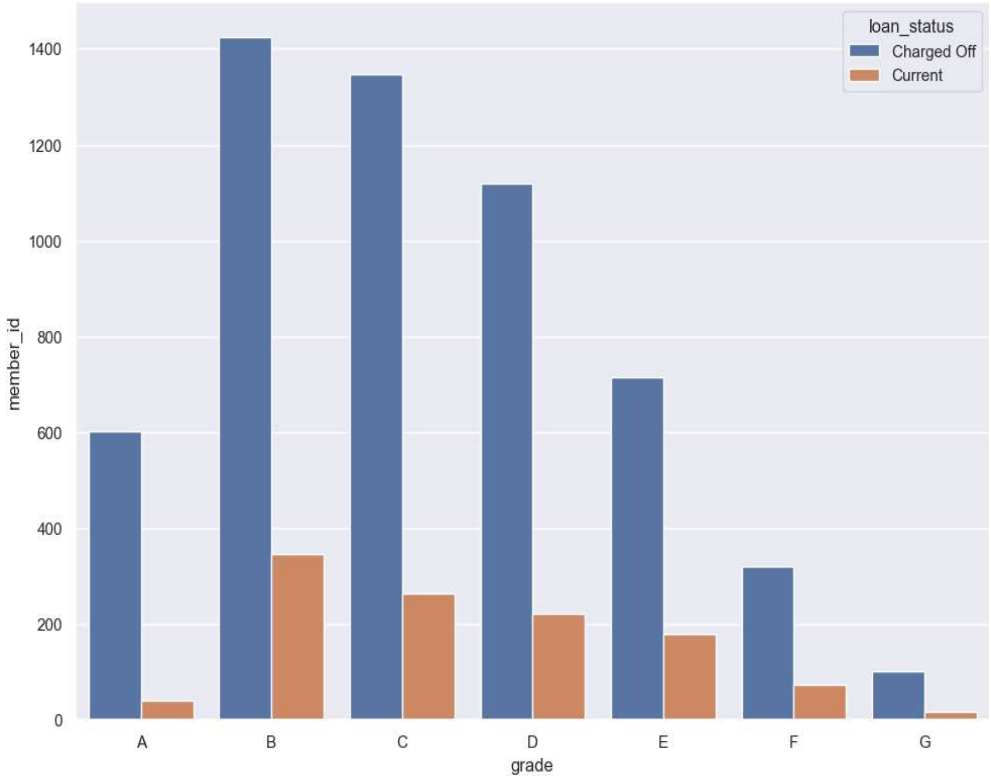
Visualization (Contd.)

- Borrower tend to Default when
  - Recoveries less than 5000
  - Principal Received till Date less than 20000
- This is the stage that occurs when the Loan is accepted
- Need some other identifiers



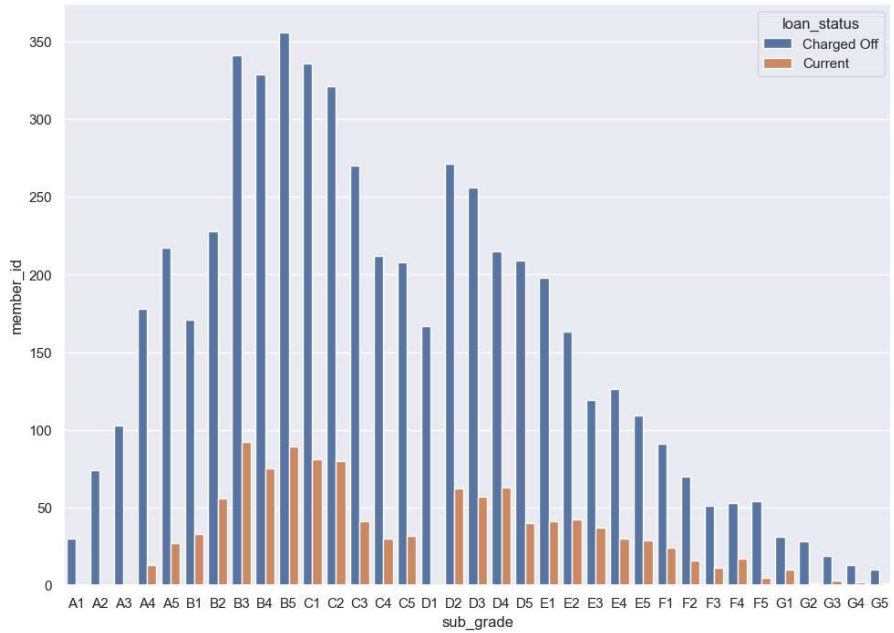
Visualization (Contd.)

- Heat Map shows that the co-relation between the grade and Chargeoff is less
- The Bar Plot shows the same
- Difficult to identify if the loan will be Fully Paid
- Chart shows insight about the Grade B having max defaulters amongst other grades



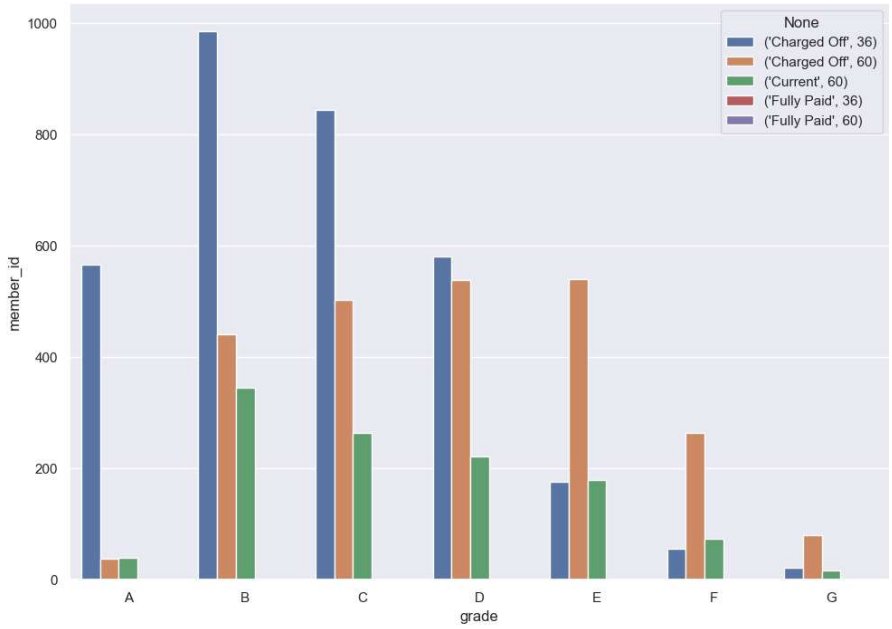
Visualization (Contd.)

- The Further Grade level Drill Down shows that B3, B5, C1, C2 Defaults more than any other Grade



Visualization (Contd.)

- As the Grade increase the Borrower tends to default the higher term loans
- Grade Level D irrespective of Loan Term tends to Default
- Grade Level B has high Default Rate for 36 Months Loan
- Grade Level B has Low Default Rate for 60 Months Loan



Visualization

- Total Recovered Principal amount till date is around 4000 for the tenure of 60 days are defaulted more