Lending Club Case Study

Table to Contents

Background

Lending club is the largest peer-to-peer marketplace connecting borrowers with lenders. Borrowers apply

through an online platform where they are assigned an internal score. Lenders decide 1) whether to lend and 2) the terms of loan such as interest rate, monthly instalment, tenure etc. Some popular products are credit card loans, debt consolidation loans, house loans, car loans etc.

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’.

This Data Analysis project 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 minimize the risk of losing money while lending to customers.

Business Objective

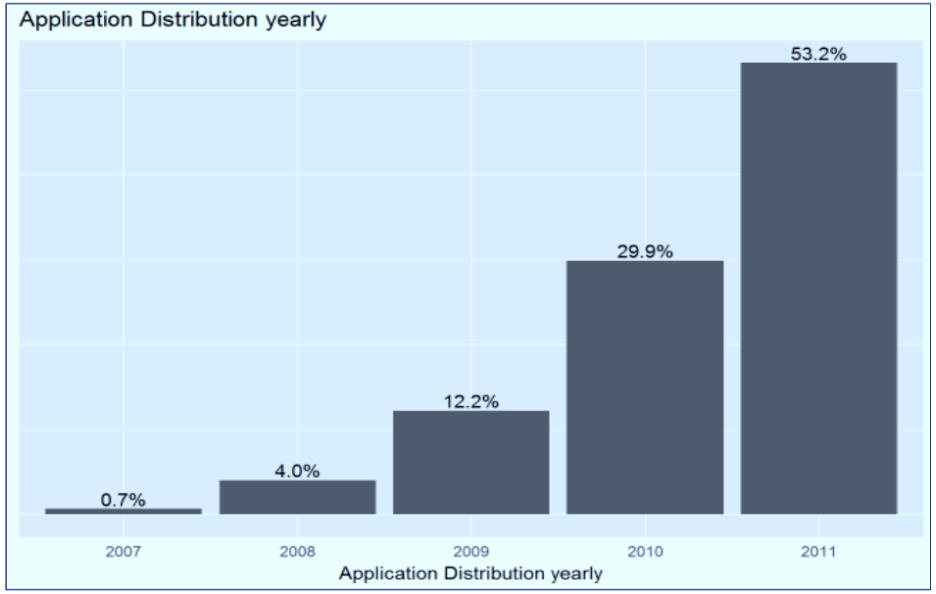
To identify variables which are strong indicators of default and potentially use the insights in approval / rejection decision making.

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.

The **target** variable is “Loan_Status” , which we want to compare across the independent variables, is loan status. The strategy is to figure out compare the average default rates across various independent variables and identify the ones that affect default rate the most.

Data Understanding

- Dataset provided/used is a related to Loan Historical dataset which contains the complete loan data for all loans issued through the time period 2007 to 2011
- Dataset contains various features related to customer demographic, characteristics , customer behavior
- Dataset contains 39717 rows and 111 features



Types of Features

- Customer (applicant) demographic – Employment Length , Employment Title, Annual Income, Zip Code, Description
- Loan related info Types of variables information & characteristics – Loan Amount, Funded Amount, Funded Amount investment, Interest Rate, Loan Status , Loan Grade
- Customer behavior (if the loan is granted) – Delinquency year, earliest credit line, Revolving Balance, Recoveries, Application type, Loan Purpose
- Some of the important columns in the dataset are loan_amount, term, interest rate, grade, sub grade, annual income, purpose of the loan etc.

Data Cleaning

Fixing Rows & Columns

- Features which have 90% and more of N/A are removed/dropped, so resultant dataset has 55 columns
- Additional columns which were dropped - "delinq_2yrs","earliest_cr_line","inq_last_6mths","open_acc","pub_rec","revol_bal",

"revol_util","total_acc","out_prncp","out_prncp_inv","total_pymnt","total_pymnt_inv","total_rec_prncp","total_rec_int","total_rec_late_fee", "recoveries","collection_recovery_fee","last_pymnt_d","last_pymnt_amnt", "last_credit_pull_d","application_type"

Standardize Values

- Converting the column type to string – emp_title
- Int_rate column has % in text, which need has been removed
- Set int_rate column dtype to float
- Term Column has values '36 months' and '60 months', We can remove ' months' and convert the column type to integer.

Identify Invalid/Outliers

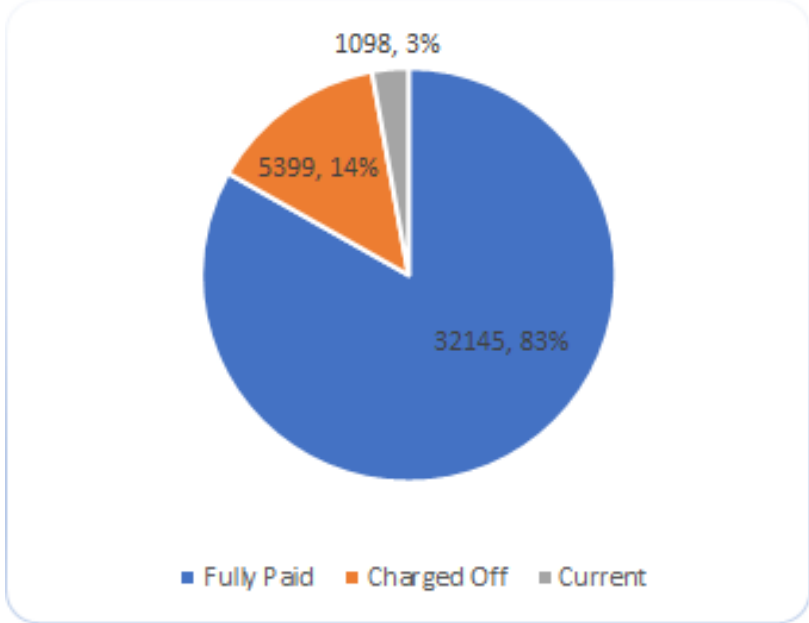
- All id and member_id values are unique; these columns are not useful for us in the analysis
- delinq_2yrs, pub_rec_bankruptcies and pub_rec has very low variance

Data Analysis

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, 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.).

- We have removed the behavior variables from analysis, as these will not have any impact on the Target variable
- In Final resultant dataset, we see that majority of the applicants has fully paid (83%), while 14% have deferred/Charged off
- We have changed **Target Variable "Loan_Status"** to represent binary form - 0 or 1, 1 indicating that the person has defaulted and 0 otherwise, instead to categorical variable



Univariate Analysis

- Dataset provided/used is a related to Loan Historical dataset which contains the complete loan data for all loans issued through the time period 2007 to 2011

- Dropped 'id','member_id','collections_12_mths_ex_med', 'acc_now_delinq', 'chargeoff_within_12_mths','delinq_amnt','tax , as these variable does not have any significance on the target variable

- Analysis on relationship between loan status and int_rate, determines that when the interest rates are increased, changes of defaulters is more

- Analysis on relationship between loan status and Annual _Income, determines that when the annual income is more, changes of defaulters is less as they are able fully payback the loans

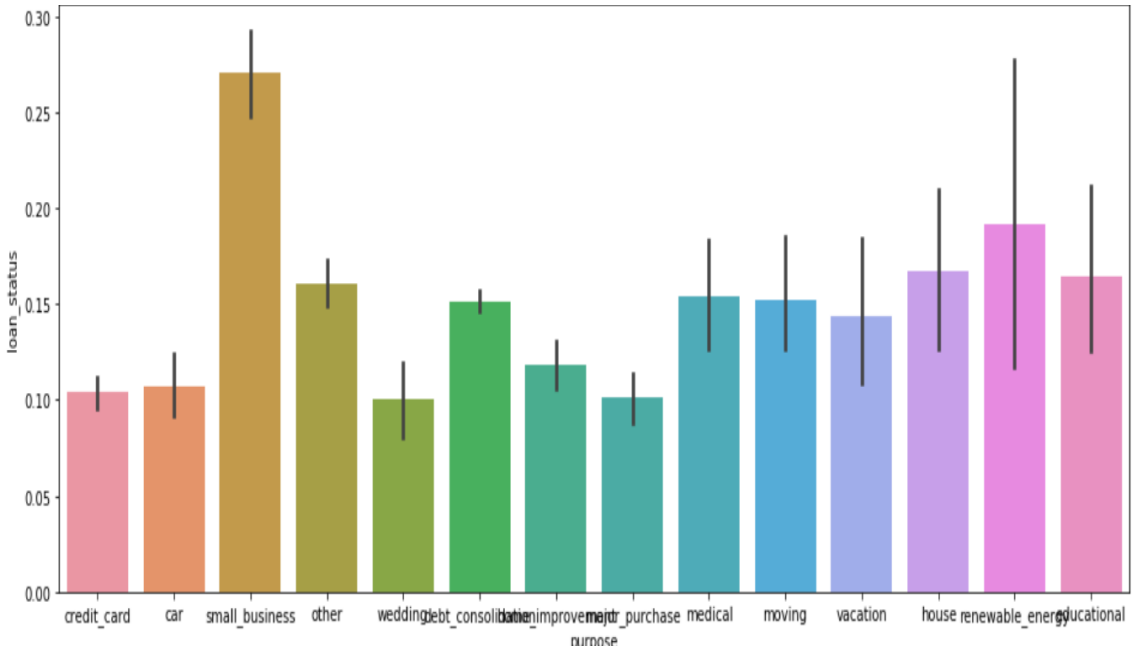
- Dataset is having 39717 rows and 111 columns, with Overall Default Rate being 14%

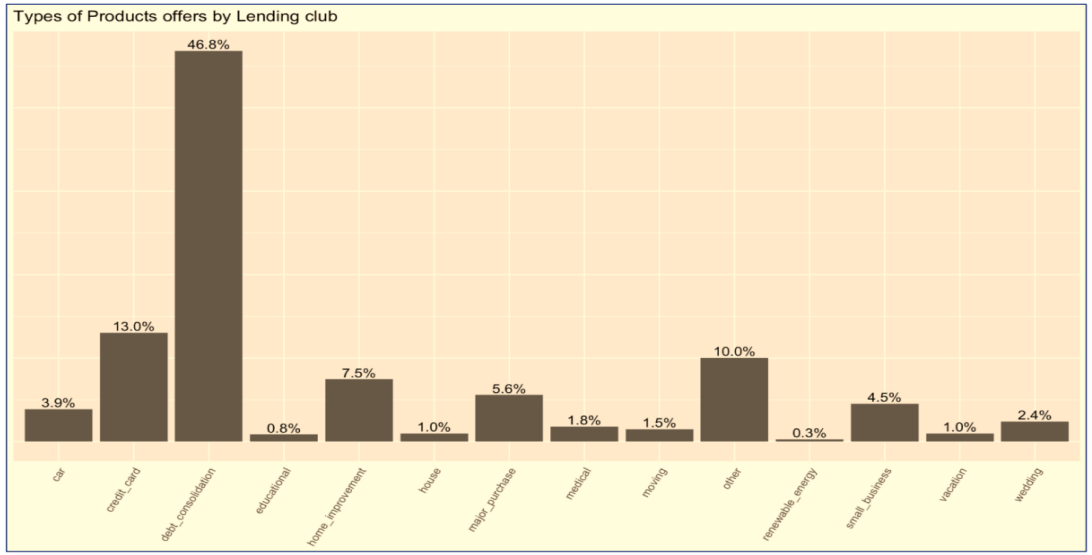
- It is also clear, as the grade of loan goes from A to G, the default rate increases. This is expected because the grade is decided by Lending Club based on the riskiness of the loan

- It is also observed that the applications which were verified have defaulted more than un-verified, which leads need to validate current verification process.

Bi-Variate Analysis

- It is observed that when compared the default rates across various variables, and some of the important predictors are purpose of the loan
- It is observed from the graph, small business loans default the most, then renewable energy and education
- Debt Consolidation Loan is the most popular
  - 46.8% debt consolidation loans
  - 13% credit card loans
  - 7.5% home improvement loans
  - 5.6% major purchase loans
- Small business loans default the most, then renewable energy and education





Observations

Target Variable which is “Loan_Status” highly depends on the following features

- Annual Income
- Purpose of the Loan
- Duration of the Loan
- Interest Rate

If we segregate the different products offered by the company, we see that there a few important features determining the defaulter status, as follows:

1. Credit Card: Grade , Term , Bin interest rate , Year and Home Ownership

2. Debt Consolidation: Bin loan amount, Grade , Term , Bin interest rate , Year and Home Ownership

3. Home Improvement: Bin loan amount, Grade , Term , Bin interest rate , Year and Home Ownership

4 . Major Purchase : Grade, Bin interest rate, Term, Home Ownership and Year