LENDING CLUB CASE STUDY: EXPLORATORY DATA ANALYSIS

Naresh Negi

PROBLEM STATEMENT

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

It specializes in lending various types of loans to urban customers. When the company receives a loan application, the company must decide for loan approval based on the applicant's profile.

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'.

The core objective of the exercise is to help the company minimize the credit loss. There are two potential sources of credit loss are:

Applicant likely to repay the loan, such an applicant will bring in profit to the company with interest rates.** Rejecting such applicants will result in loss of business**.

Applicant not likely to repay the loan, i.e. and will potentially default, then approving the loan may lead to a financial loss* for the company

DECISION?

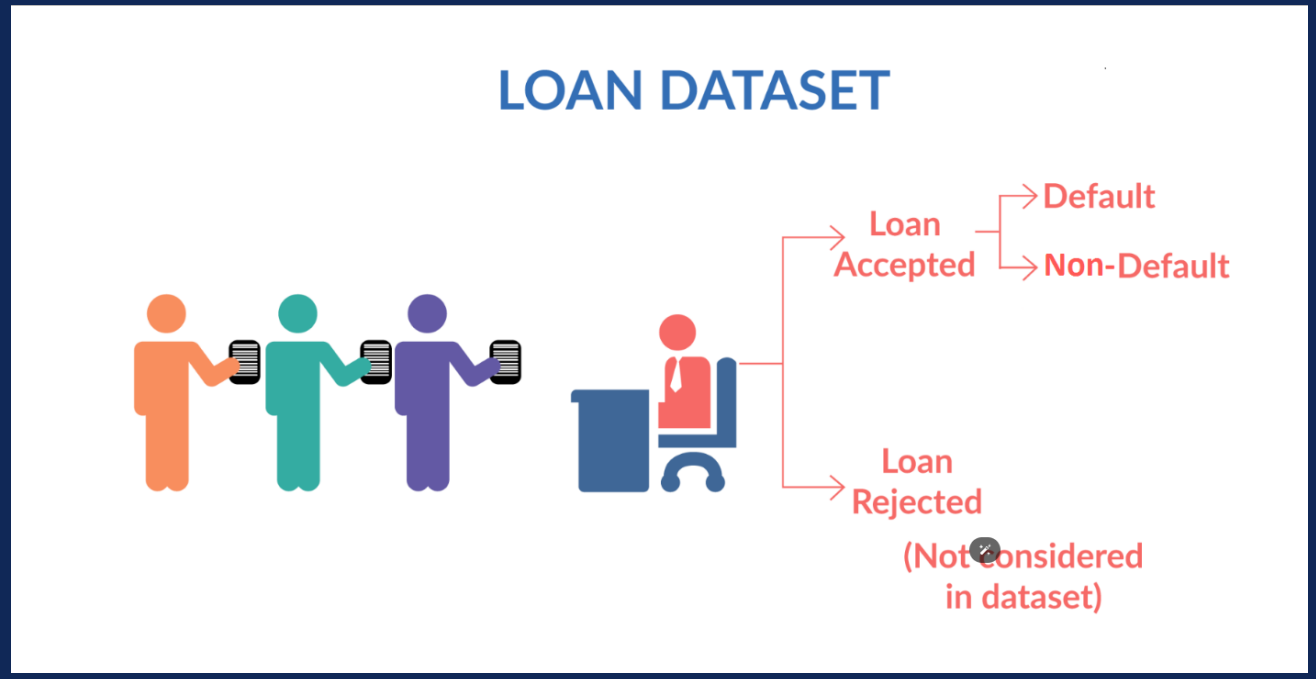
Loan acceptance decisions can result in three scenarios:

Fully paid: The applicant has repaid the loan entirely.

Current: The applicant is currently paying instalments, indicating an ongoing loan tenure.

Charged-off: The applicant has defaulted on the loan by failing to repay instalments for an extended period.

Loan rejection occurs when the company decides not to approve the loan due to various reasons, resulting in no transactional history available for those applicants in the dataset.



OBJECTIVES

- The goal is 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 using the given dataset, is the aim of this case study.

- 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.

DATA UNDERSTANDING

- The data given below contains the 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.

- The dataset reflects loans post approval, thus does not represent any information on the rejection criteria process
- Overall objective will be to observe key leading indicaters (driver variables) in the dataset, which contribute to defaulters
- Use the analysis as a the foundation of the hypothesis
- The overall loan process is represented by three steps
- Potential borrower requests for loan amount (loan_amnt)
- The approver approves/rejects an amount based on past history/risk (funded_amnt)
- The final amount offered as loan by the investor (funded_amnt_inv)

DATA ANALYSIS APPROACH

Leading Attribute

- Loan Status - Key Leading Attribute (loan_status). The column has three distinct values
- Fully-Paid - The customer has successfuly paid the loan
- Charged-Off - The customer is "Charged-Off" ir has "Defaulted"
- Current - These customers, the loan is currently in progress and cannot contribute to conclusive evidence if the customer will default of pay in future
- For the given case study, "Current" status rows will be ignored

Decision Matrix

- Loan Accepted - Three Scenarios
- 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 ANALYSIS APPROACH

Important Columns

The given columns are leading attributes, or predictors. These attributes are available at the time of the loan application and strongly helps in prediction of loan pass or rejection. Key attributes Some of these columns may get dropped due to empty data in the dataset

- Customer Demographics
  - Annual Income (annual_inc) - Annual income of the customer. Generally higher the income, more chances of loan pass
  - Home Ownership (home_ownership) - Wether the customer owns a home or stays rented. Owning a home adds a collateral which increases the chances of loan pass.
  - Employment Length (emp_length) - Employment tenure of a customer (this is overall tenure). Higher the tenure, more financial stablity, thus higher chances of loan pass
  - Debt to Income (dti) - The percentage of the salary which goes towards paying loan. Lower DTI, higher the chances of a loan pass.
  - State (addr_state) - Location of the customer. Can be used to create a generic demographic analysis. There could be higher delinquency or defaulters demographicaly.
- Loan Attributes
  - Loan Ammount (loan_amt)
  - Grade (grade)
  - Term (term)
  - Loan Date (issue_date)
  - Purpose of Loan (purpose)
  - Verification Status (verification_status)
  - Interest Rate (int_rate)
  - Installment (installment)
  - Public Records (public_rec) - Derogatory Public Records. The value adds to the risk to the loan. Higher the value, lower the success rate.
  - Public Records Bankruptcy (public_rec_bankruptcy) - Number of bankruptcy records publocally available for the customer. Higher the value, lower is the success rate.

DATA ANALYSIS APPROACH

Ignored Columns

- The following types of columns will be ignored in the analysis. This is a generic categorization of the columns which will be ignored in our approach and not the full list.
  - Customer Behaviour Columns - Columns which describes customer behaviour will not contribute to the analysis. The current analysis is at the time of loan application, but the customer behaviour variables generate post the approval of loan applications. Thus, these attributes will not be considered towards the loan approval/rejection process.
  - Granular Data - Columns which describe next level of details which may not be required for the analysis. For example, grade may be relevant for creating business outcomes and visualizations, sub grade is be very granular and will not be used in the analysis

DATA ANALYSIS APPROACH

Rows Analysis

- Summary Rows: No summary rows were there in the dataset
- Header & Footer Rows - No header or footer rows in the dataset
- Extra Rows - No column number, indicators etc. found in the dataset
- Rows where the loan_status = CURRENT will be dropped as CURRENT loans are in progress and will not contribute in the decision making of pass or fail of the loan. The rows are dropped before the column analysis as it also cleans up unecessary column related to CURRENT early and columns with NA values can be cleaned in one go
- Find duplicate rows in the dataset and drop if there are

Columns Analysis

Drop Columns

- There are multiple columns with NA values only. The columns will be dropped.
- There are multiple columns where the values are only zero, the columns will be dropped
- There are columns where the values are constant. They dont contribute to the analysis, columns will be dropped
- There are columns where the value is constant, but the other values are NA. The column will be considered as constant. columns will be dropped
- There are columns where more than 50% of data is empty (mths_since_last_delinq, mths_since_last_record) - columns will be dropped

DATA ANALYSIS APPROACH

Convert Column Format

- (loan_amnt, funded_amnt, funded_amnt_inv) columns are Object and will be converted to float
- (int_rate, installment, dti) columns are Object and will be converted to float
- strip "month" text from term column and convert to integer
- Percentage columns (int_rate) are object. Strip "%" characters and convert column to float
- issue_d column converted to datetime format

Standardise values

- All currency columns are rounded off to 2 decimal places as currency are limited to cents/paise etc only.

Convert Column Values

- loan_status column converted to boolean Charged Off = False and Fully Paid = True. This converts the column into ordinal values
- emp_length converted to integer with following logic. Note < 1 year is converted to zero and 10+ converted to 10.
  - < 1 year: 0,
  - 2 years: 2,
  - 3 years: 3,
  - 7 years: 7,
  - 4 years: 4,
  - 5 years: 5,
  - 1 year: 1,
  - 6 years: 6,
  - 8 years: 8,
  - 9 years: 9,
  - 10+ years: 10

DATA ANALYSIS APPROACH

Added new columns

- verification_status_n added. Considering domain knowledge of lending = Verified > Source Verified > Not Verified. verification_status_n correspond to {Verified: 3, Source Verified: 2. Not Verified: 1} for better analysis
- issue_y is year extracted from issue_d
- issue_m is month extracted from issue_d

Missing Data Rules

- Missing Data Rules
- Columns with high percentage of missing values will be dropped (50% above for this case study)
- Columns with less percentage of missing value will be imputed
- Rows with high percentage of missing values will be removed (50% above for this case study)

Column Dropping Rules

- Approach taken here in this analysis, if total number of rows (for all columns) which are blank is less than 5% of the dataset, we are dropping the rows. If the total rows are greater than 5% we will impute
- If the dataset of blanks is considerably small, dropping the rows will possible be more accurate approach withoug impacting the dataset and the outcomes
- If the dataset of blanks are considerably large, dropping the rows will skew the analysis and impute approach will be taken
- In the current dataset, combined row count of blanks for emp_length and pub_rec_bankruptcies is 1730, which is 4.48% of the total rows thus dropping the rows will be the more accurate approach
- If imputing, we will correlate emp_length with annual_inc, with the logic that higher the length of employment, higher the salary potentially. With this approach, the outliers can potentially introduce noise.

DATA ANALYSIS APPROACH

Outlier Treatment Rules

- Approach taken in this analysis to drop all outlier rows
- The following columns were evaluated for outliers loan_amnt,funded_amnt,funded_amnt_inv,int_rate,installment,annual_inc,dti
- Total rows dropped due to outlier treatment: 3791
- Percentage of rows dropped due to outlier treatment: 10.29 %