Lending Club Case Study Exploratory Data Analysis

upGrad and IIITB Machine Learning and AI Program - Jan 2024

1. Introduction
2. Problem statement & objective
3. Primary goals & approach
4. Exploratory data analysis (EDA)
5. Conclusion - Recommendations

Introduction

A consumer finance company which specializes 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

Problem statement

- The 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’.

Business Objectives

Objective 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.

Primary goal

To identify the driving factors (or driver variables) behind loan default such that, the company can utilize this knowledge for its portfolio and risk assessment.

Approach

- Utilize the dataset provided which contains the complete loan data for all loans issued through the time period 2007 to 2011.
- Using EDA for risk assessment

Exploratory Data Analysis

Data understanding

LOAN Data Set

- Number of Columns – 111
- Number of Rows – 39717

Observations -

- Many Columns have null value
- Data type mismatch for certain columns
- Certain columns can be converted to categorical column

Data handling

- Dropped the columns with null value
- Dropped the rows which have more than 30% of null value
- Dropping single value columns and irrelevant columns
- Excluding loan_status = 'Current’. These candidates are not labelled as 'defaulted’.
- Replacing missing values to default values
- Correcting the data types and added derived columns
- Defined Category columns

Data Visualization

Visualization approach used:

- Univariate
- Segmented Univariate Analysis
- Bivariate Analysis

Visualizations used:

- Box plot
- Count plot
- Dist plot
- Bar plot
- Pie chart
- Heat Map

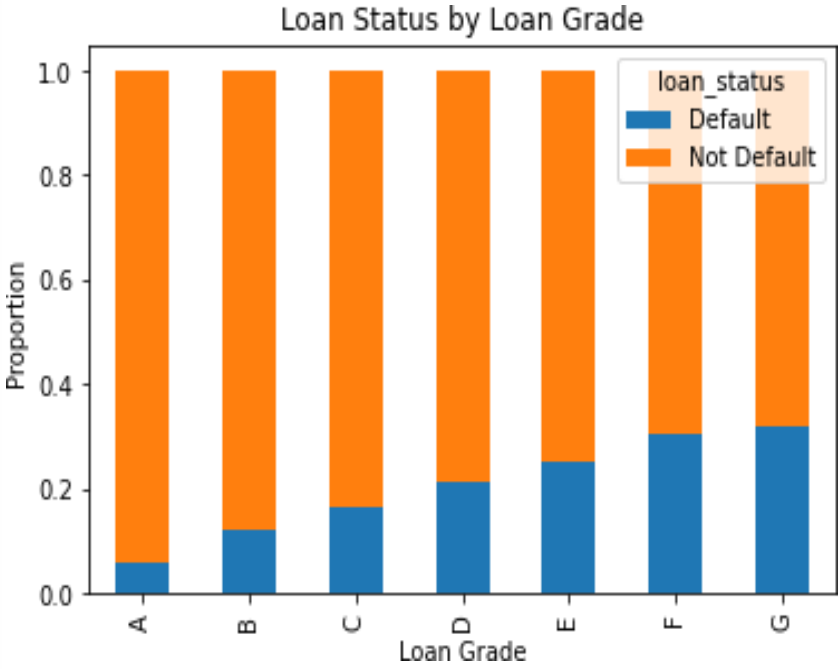
Observations

Univariate Analysis

- More than 70% of the loans are sanctioned for the term of 60 months
- Loan Amount varies from 500 to 35K and approx.. 80% of the loans are sanctioned for loan amount in range of 500 – 15K
- 73% of loans defaulted are in the loan amount range of 500 – 15K.
- Around 50% of Charged Off Loans are in 13% - 21% interest rate range.
- Loans taken for Debt Consolidation purpose are major defaulters, followed for Credit Card & Other purpose. This is similar for Charged Off Loans too.

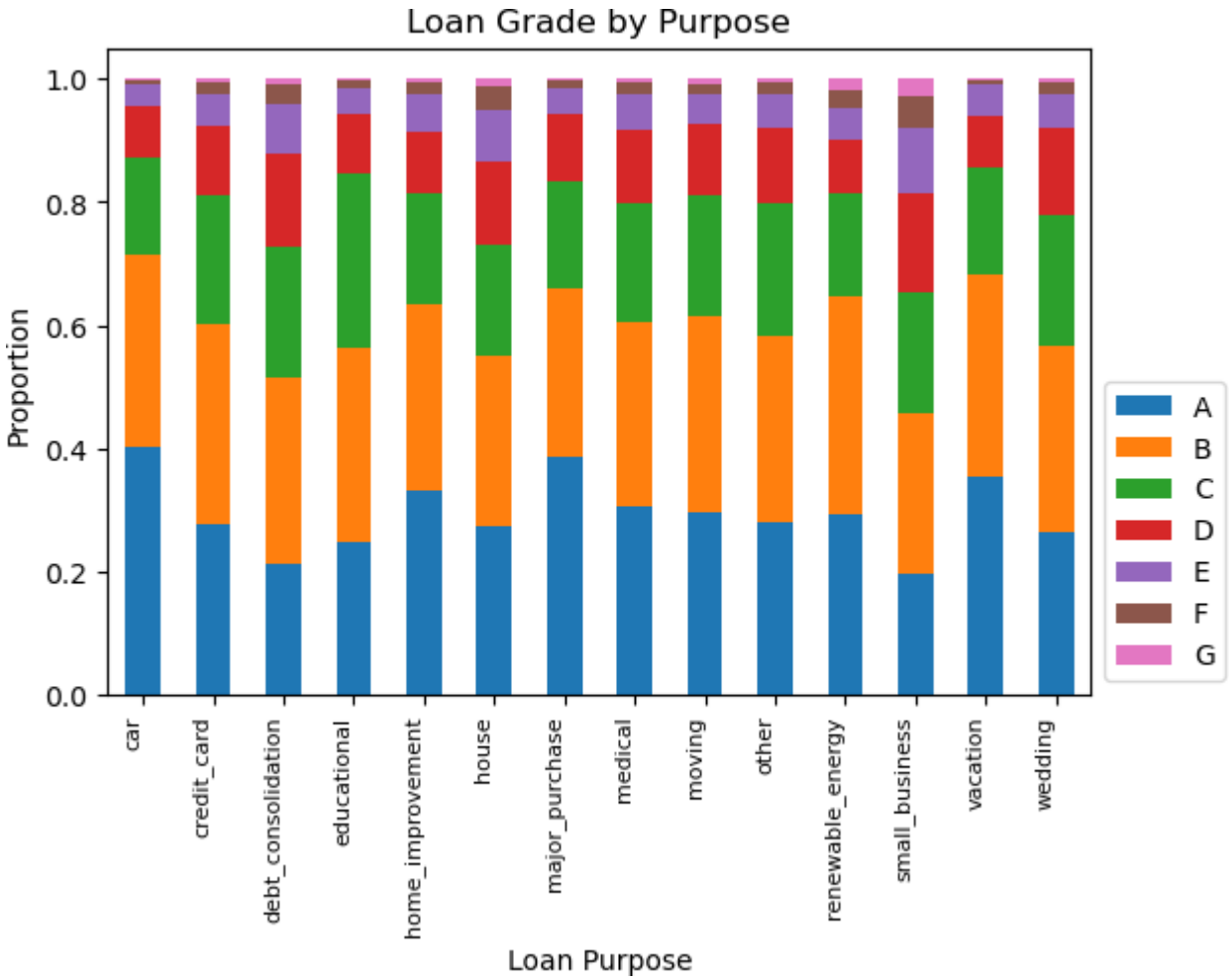
Univariate Analysis

Loan status indicates loan outcome and default risk, essential for identifying patterns, building models, and assessing portfolio performance. This chart shows the proportion of loans in each grade that are either defaulted or not defaulted and helps us understand how loan grade and loan status are related. Loans with lower grades have higher default rates, and loans with higher grades are more likely to be fully paid off.



Univariate Analysis

Loan grade by purpose



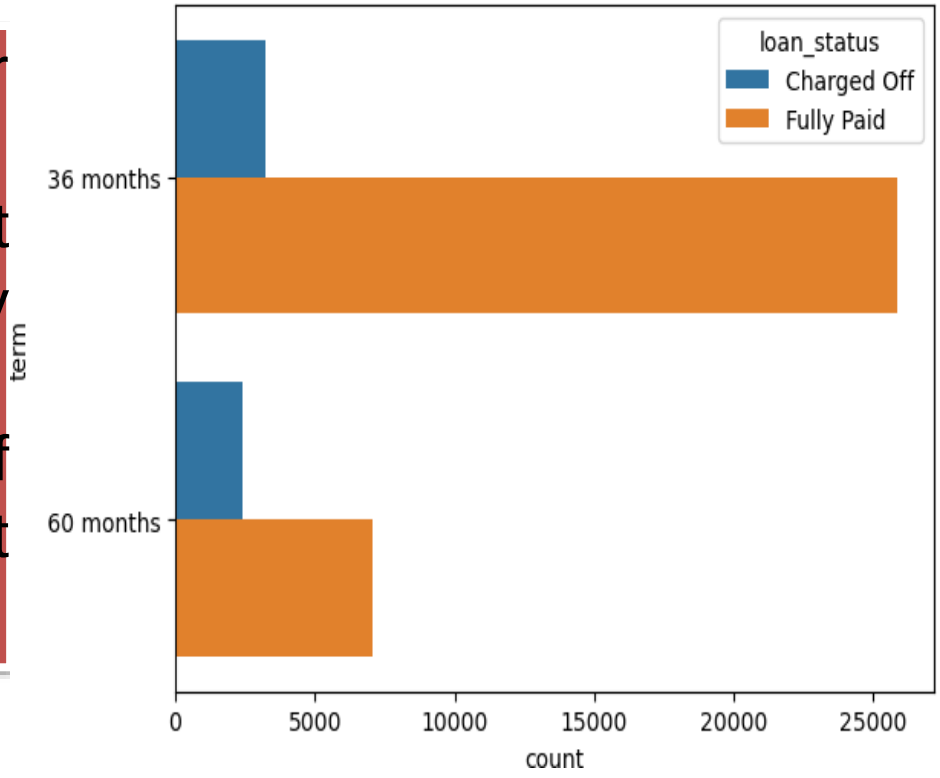
Univariate Analysis

The stacked bar chart shows how loan purpose and loan grade are related:

- Debt consolidation is the most common loan purpose across all loan grades, while educational loans are the least common.
- The proportion of loans for debt consolidation is highest in the A and B loan grades, while the proportion of loans for small business and renewable energy is highest in the C and D loan grades.
- The proportion of loans for credit card refinancing is highest in the E and F loan grades, while the proportion of loans for home improvement is highest in the G loan grade.
- The highest proportion of defaulted loans is in the D and E loan grades, with the highest default rates for small business and renewable energy loans.

Segmented Univariate Analysis

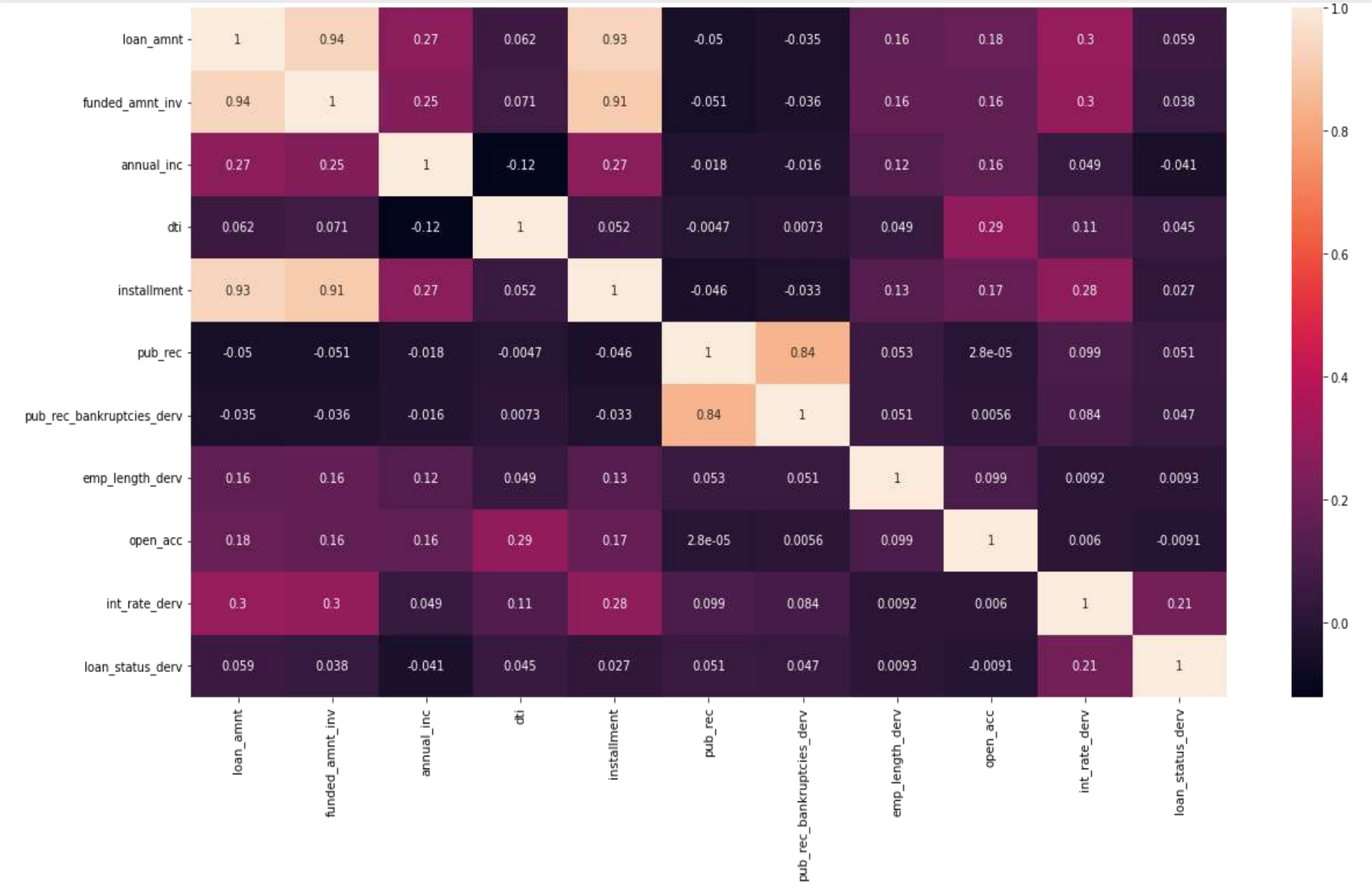
- 36 months terms has higher number of defaulters
- Applicants with employment length of 10+ years are mostly defaulters
- Around 50% of Charged Off Loans are in 13% - 21% interest rate range.



Bivariate Analysis

- Interest Rate for Charged Off Loans is higher than the Fully Paid loans for both 36 months and 60 months term. Indicates that loans with higher interest rate are more likely to be defaulted.
- Loan applicants with mortgage have high loan amount
- Loan applicants with mortgage have high defaulters

Correlation Analysis



Correlation Analysis

The heat map shows the correlation coefficients between different variables related to loan amounts, income, credit score, and loan status. Some notable observations include:

- Loan amount and funded amount are highly correlated, which is expected as they represent the same information.
- There is a positive correlation between income and loan amount, indicating that borrowers with higher income tend to borrow more.
- Interest rate has a strong negative correlation with credit score, indicating that borrowers with higher credit scores tend to get lower interest rates.
- Debt-to-Income (DTI) ratio has a weak negative correlation with credit score, suggesting that borrowers with higher credit scores tend to have lower DTI ratios.
- Loan status is negatively correlated with interest rate, indicating that loans with higher interest rates are more likely to default or be charged off.
- The number of public records and bankruptcies are positively correlated, indicating that borrowers with more public records tend to have more bankruptcies as well.

Recommendations

The Probability of defaulting is high when:

- Loan Applicants not owing home (Mortgage or Rent) and have high Annual Income Range (60K - 70K)
- Loans with interest rates between 9% - 17%
- Applicants with Annual Income range between 35k - 70k and availing loan for Debt Consolidation.
- Loan Applicants with >10 years of experience and with loan amount > 10K or with interest rate (>10%)

Indicators for loan defaulters

- Annual Income, Home Ownership, Purpose of Loan, Loan Amount, Interest Rate

Thank you!!!