Lending Club Case Study

Exploratory Data Analysis

Pawan Mani Teja Kuppili

Jatan Porecha

Topics

- ❏ Problem Statement
- ❏ Exploratory Data Analysis - Activities
- ❏ Exploratory Data Analysis - Insights
- ❏ Summary

Problem Statement

We are given the loans data from consumer finance company which provide the information attributes about historical and current loans.

We are required to apply exploratory data analysis learnings and activities to cleanse and analyze the data to derive the insights from this dataset.

Our focus would be mostly on analyzing the risk of loans going bad (i.e. default) and corroborate or correct the conventional understanding about some financial data.

Exploratory Data Analysis - Activities

As part of this exercise, we have performed following activities to derive the insights,

- ❏ Data Cleanup and Conversion
- ❏ Derived Metrics
- ❏ Outliers Removal
- ❏ Data Analysis

In next pages, we will term Exploratory Data Analysis as EDA for brevity.

EDA - Data Cleanup Activity

The loans data provided is cleaned up horizontally i.e. removing the empty, unimportant and lower cardinality columns

Following are the major steps taken up,

- Finding Columns having all Null or 90% Null data
- Remove columns which have constant as a value.
- Inspect and remove unwanted columns

After the cleanup, the dataset has been reduced from 111 to 43.

We also removed some of the rows which had null/empty values in most of the columns as they are not providing any valuable information at dataset level.

EDA - Data Conversion

Some of the columns in the dataset contain the date values but they are not properly converted to datetime type. The percentage columns were converted to numeric data like 10.5% is converted to 0.105

There are columns which contain the no. of years / months but they are type of string strings, e.g. 4 years, 10+ years, 36 months, etc…

Following columns are converted,

- term
- emp_length
- int_rate
- revol_util
- issue_d
- last_pymnt_d
- last_credit_pull_dy
- earliest_cr_line

EDA - Derived Metrics

The measures provided in the dataset can be combined or “binned” to do meaningful univariate / bivariate segmented analysis.

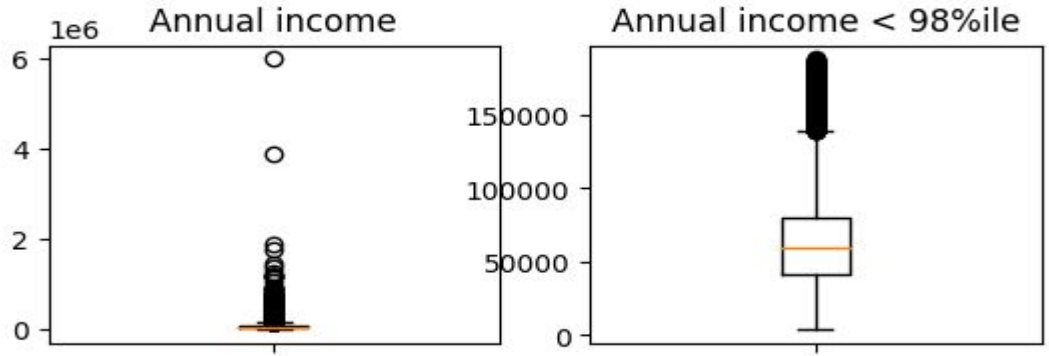
Some of the derived columns are,

- **installment_to_month_inc** : Ratio of installment to monthly income (derived from annual income)
- **open_to_total_acc :** ratio of open account to total account with binned into 0 to 10 bins.
- **annual_inc_range** : binning of annual income to 6 bins
- **loan_amnt_range** : binning of loan amounts to 5 bins
- **Int_rate_range** : binning of interest rates to 5 bins

EDA - Outliers Removal

The loan data has some of the borrowers who have a large annual income and they are clearly the outliers as per the box plot.

Therefore, we removed the records which had annual income more than USD 2 million.



EDA - Analysis

We performed various kinds of analysis on the cleaned up dataset to derive the insights.

Some of the important analysis activities are,

- Loan Amount vs Funded Amount vs Funded Amount by investors
- Find relation among all the amount columns
- Distribution of loans and their status per state
- Distribution of loans status on various parameters
- Analysis of credit history of borrowers with respect to defaults
- Analysis of DTI ratios with respect to defaults.

EDA - Insights

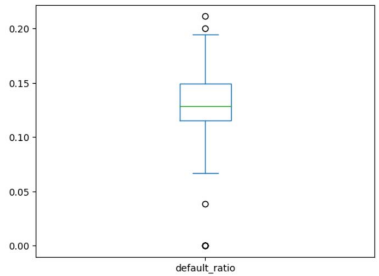
In this section we will discuss the various insights which we could gather from the analysis which we performed the loan dataset.

Following are the insights at high level,

- Probability of loan defaults based on the US states residents
- Impact of Loan verification status
- Probability of loan defaults based on borrowers' delinquency data
- Impact of loan term on payback probability
- Impact of amount of principal received on loan turning default.
- Impact of DTI and monthly income
- Borrowers' credit history and interesting finding on generational gap.

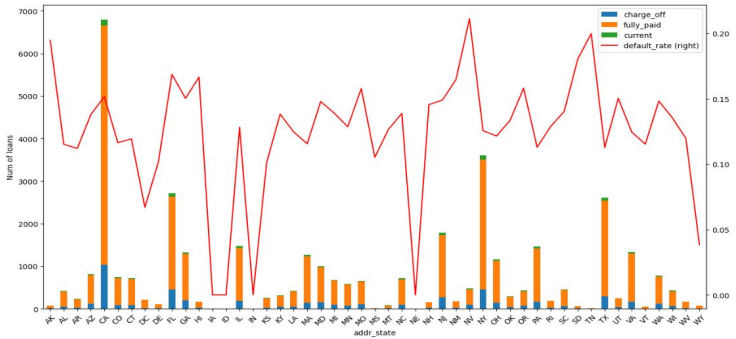
Insights - Loan Default based on US States

75 percentile of states are having default rate of around 0.15, so combining both visualizations,



Following states are having higher default rates

- FL
- HI
- NV
- TN
- SD

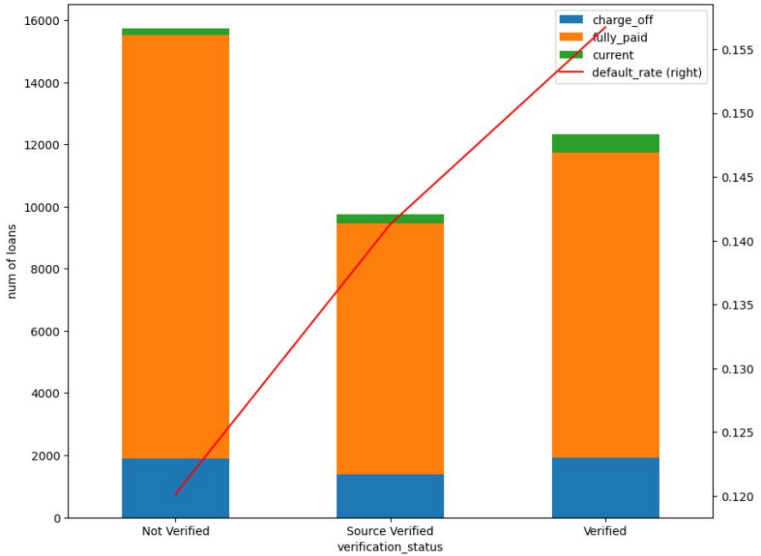


Especially the residents of Florida (FL) and Nevada (NV) have the higher tendencies to go for default.

Insights - Impact of loan verification status

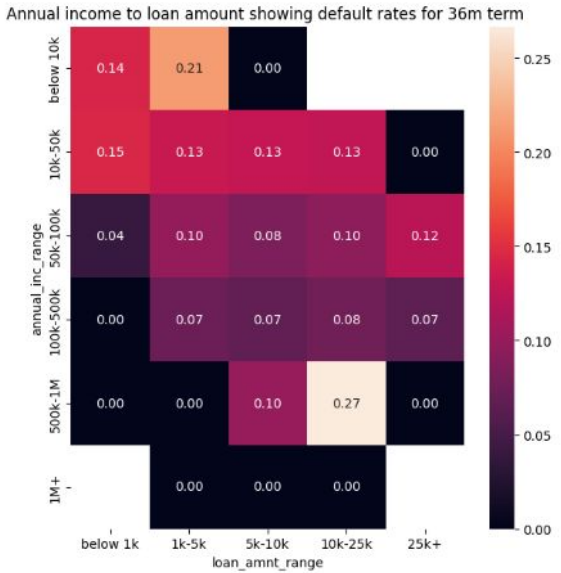
The data is counter intuitive to conventional wisdom of loan status “Verified” is safer.

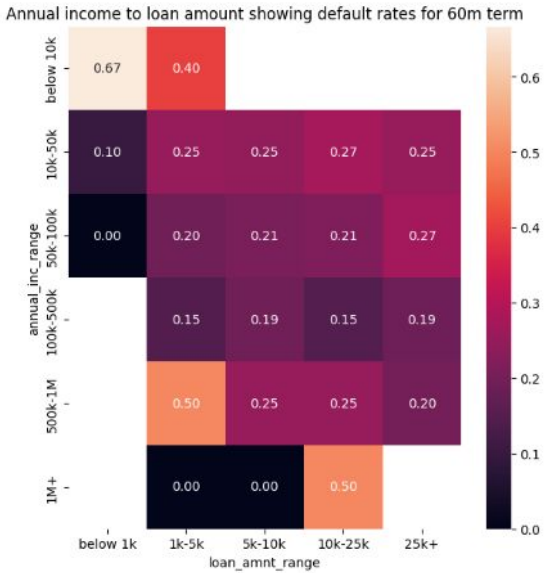
We found that default rate went up for verified loans. One of the possibility is that the loan is verified when the loan amount is higher and if the borrower is going for such high value loan, the borrower could be a risky one and may have chances of default.



Insights - Impact of loan amount and annual income

By correlating the data of defaulted loans, on loan amount vs annual income of the borrowers and focusing on default loans, what we found that the borrowers who borrow the loan of amount 30% to 40% to their annual income, have higher chances of default, regardless of verification status

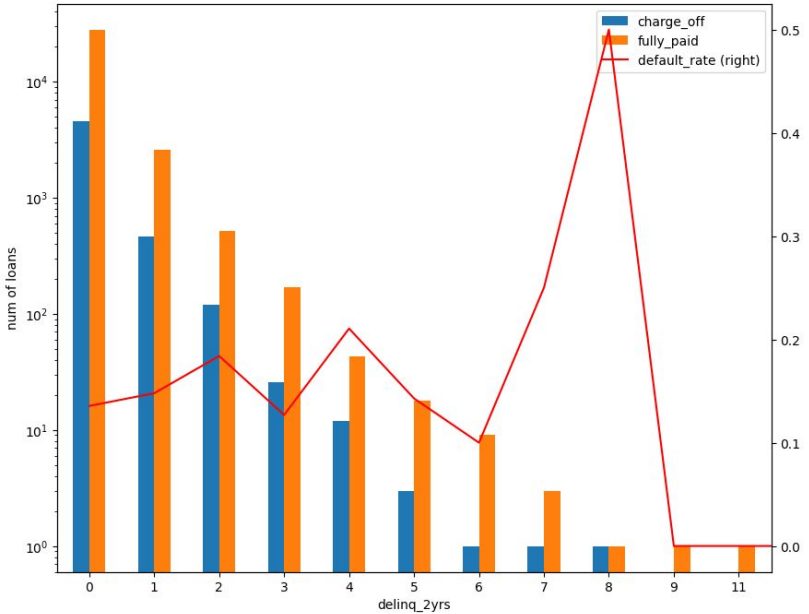




Insights - Based on delinquency data

The delinquency instances (payments delayed by 30 days) of the borrowers don't have much impact on loan default status.

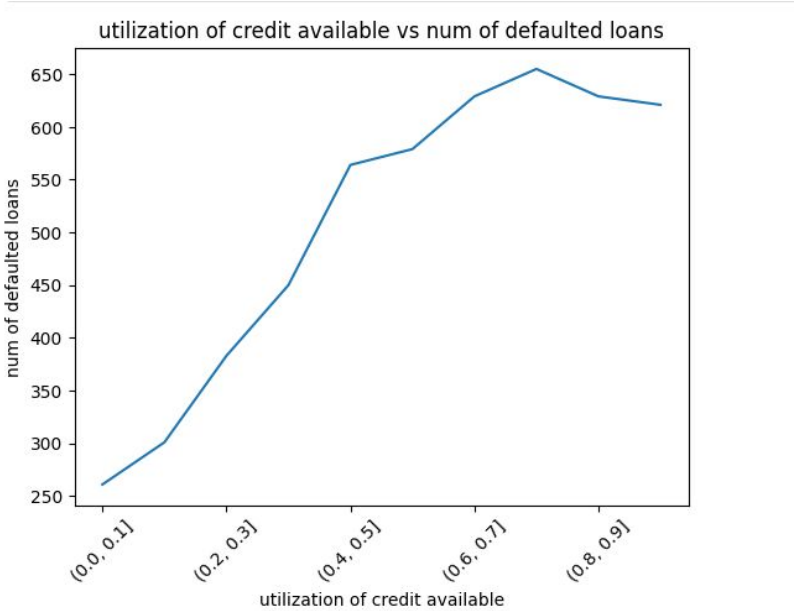
However it seems that the institution is considering the delinquency data before giving out the loans and hence the no. of loans go down drastically after 4 instances of delinquency.



Insights - Based on utilization of credit

There is a positive correlation between the utilization of credit (revolve balance) and the default rates.

The borrowers who have penchant of delaying the payments in their ongoing credit lines may also default on their loans payback.

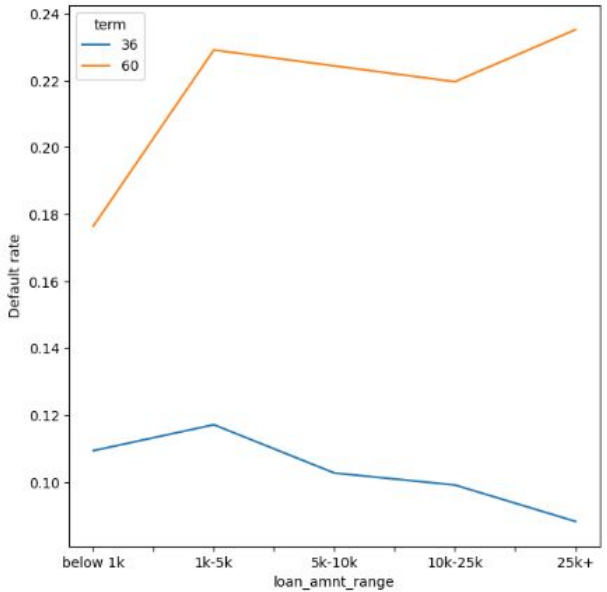


Insights - Impact of loan term on payback

The impact of loan term on probability of loan going default is very clear.

The longer the loan term, there is a higher probability of loan default regardless of the loan amount.

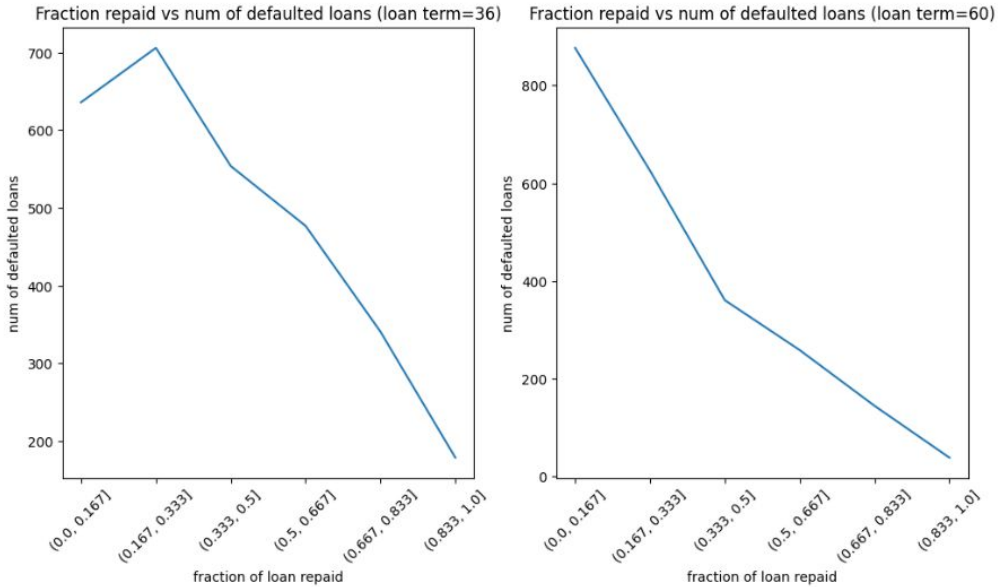
Possible reasons could be the longer term loans have higher interest component and hence less principal received in first 50% of tenure.



Insights - Principal received to loan default

There is a very clear **inverse trend** between principal received and loan going default regardless of loan tenure, loan amount and annual income.

Fraction of loan repaid, is a ration between principal received and total loan amount.

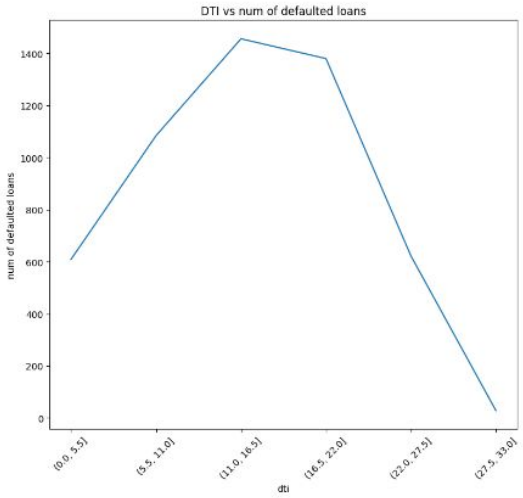


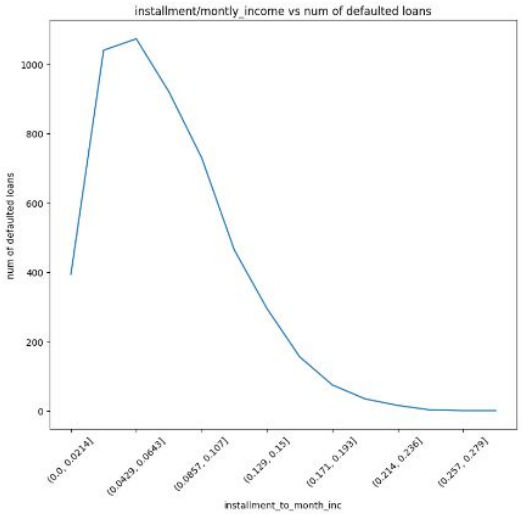
Insights - Impact of DTI and Monthly Income

Theoretically lower the DTI ratio is better however the data given in the dataset suggests opposite, wherein the rate of defaulted loans go down for the higher DTI borrowers.

However there is an interesting insight coming out of plotting loan defaults against ratio : (installment / monthly income)

Conventionally lower the ratio would be a better borrower given better paying capability but the loan defaults are higher for lower ratios. This could be because of borrower being riskier and overestimating his paying capability considering his annual income .



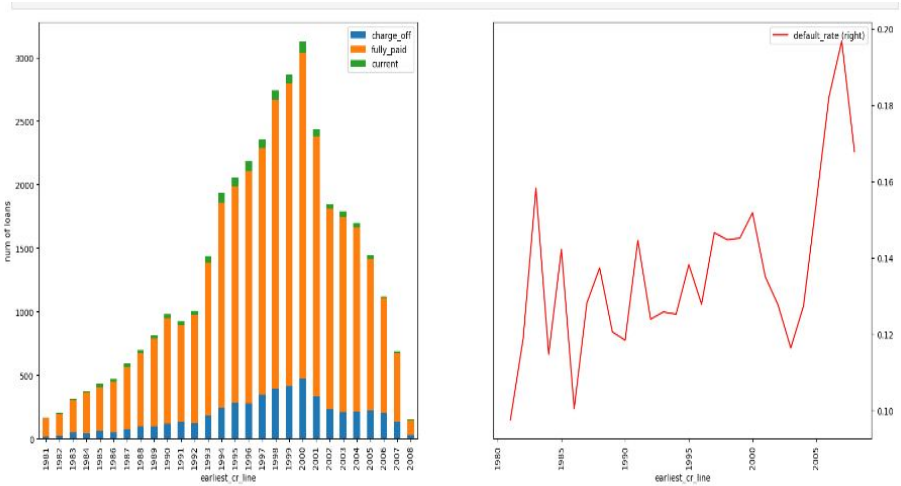


Insights - Borrower's credit history

The plots show the borrower's earliest credit line year to the loans taken and default rates.

It is interesting to note that the default rates picked up for the borrowers who started their credit taking history 2005 and later.

This would be due to easy money available during that time which resulted into 2008 meltdown.



Summary

This exercise of analysing the loan data helped us to apply various techniques learnt in EDA module, and derive interesting insights on, borrower's credit history, loan terms, loan amounts vs borrower incomes and their locations.

It also surprised us and challenged some of our perceptions (e.g. verified loans would be more secure than unverified)