Lending Loan Case Study

- Soumya Swaroop Gupta

Problem statement

For a consumer finance company,
identify patterns which indicate if a person is likely to default.

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.

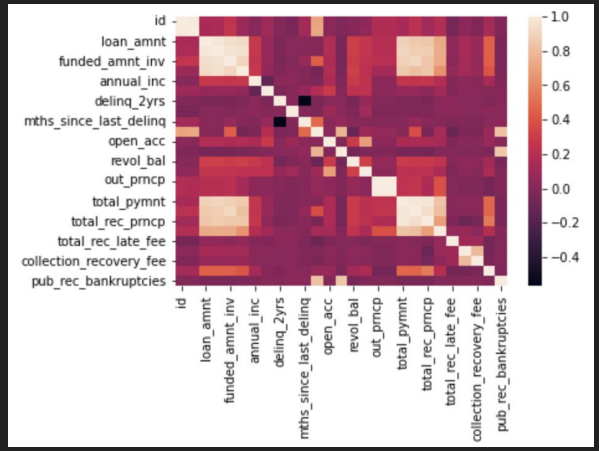
Approach

- Clean the data and understand it.
- Identify the data/sub-set of data to work on.
- Identify risky loan applicants. To do that, identify the slice of data
  - By loan amounts.
- With this slice of data, we can find the driving factors. These factors can be used to come up with methods to identify applicants risky borrowers. Thus, we can reduce Charged off loans and thereby cut down the amount of credit loss.
- To Validate using the Driving factors, we should be able to get a criteria where
  - The sum of the loan amounts of ‘Fully Paid’ loans is zero and
  - The sum of the loan amounts of ‘Charged Off’ loans is significantly high

Data cleansing

Reasoning behind Dropping columns

- There were many columns with only null values, dropped them.
  - loan = loan.drop(loan.columns[loan.isna().all()].tolist(),axis=1)
- There were a few columns with constant value, dropped them.
  - Example:- application_type is 'individual', initial_list_status and so on
- There were text columns that will not help in analysis, dropped them
  - Example:- desc,emp_title
- Found out the columns with high correlation, retained only one of them.
  - Example:- out_prncp, out_prncp_inv and so on
- There were columns with repeated information, retained only one of them.
  - Example:- state and first 3 character of zipcode are related. So retained only state.



- Soumya Swaroop Gupta

Handling missing data

- Revol_util had null data, but Cannot derive revol_util from revol_bal. Cannot set it to the mode of the column as it doesn't make sense functionally. So slicing those rows that have revol_util_cat as Nan
- Functionally, Pub_rec_bankruptcies filling with a common value does not make sense, hence filling it with 0
- There are many rows with mths_since_last_delinq as null hence filling it with 0
- Since pub_rec for the records are 0 where mths_since_last_record is null. Setting the value to 0
- When revol_bal==0, revol_util is set to zero. Since there is only one record of revol_bal not 0 when revol_util is nan, removing that record

Converted mths_since_last_delinq,mths_since_last_record,revol_util to integers

Reasoning behind type conversion

- Emp_length is a numeric type. But it had 10+, <1 and Nan. Based on the definition given in the dictionary, changed 10+ to 10 and <1 year to 0.
- Term has only 2 unique values ' 36 months', ' 60 months' and was of type object. Retained only the numbers and changed the data type to int
- Int_rate and revol_util had %. Converted to float after cleansing it.
- Extracted the month and years from issue_d, last_pymnt and other dates. Dropped the original column after extracting the information to new columns.

Reasoning behind slicing rows

- For analysing the required information for this project, we need only ‘Fully paid’ and ‘Charged Off’. So retained only the rows with the given loan_status.
  - Also, out_prncp was highly correlated to everything else. So, removed the column.
- Sliced rows that have revol_util as null
- Sliced the outliers and retained only annual_inc<=9.000000e+04

Data Understanding

From the data - About the company

This company is based in US, which lends loans only to individuals. The number of loan issued by the company has increased from 2007 to 2011.

The loan amount seems to be capped at around 40000(dollars). Assuming the currency to be dollars as it is based in US. The company offers only 2 loan durations, it is either 3 years or 5 years. Though many of the loan applicant's income was not verified, most of the payments are happening on time. However, the loan payments have been less in 2015 and 2016.

PS - Using univariate analysis on application_type, issue_d, loan_amnt, term, verification_status, total_rec_late_fee, last_pymnt_year, loan_status

From the data About the loan

The commonly borrowed loan amount is around 5,000(dollars). Most borrowers opted for a loan duration of 3 years. The purpose listed most often is debt consolidation.

In the loans that are issued in the 3 years(2007-2011), about 5000 loans that are charged off.

PS - Using univariate analysis on loan_amnt, emp_length, title, issue_d, loan_status

From the data About the borrowers

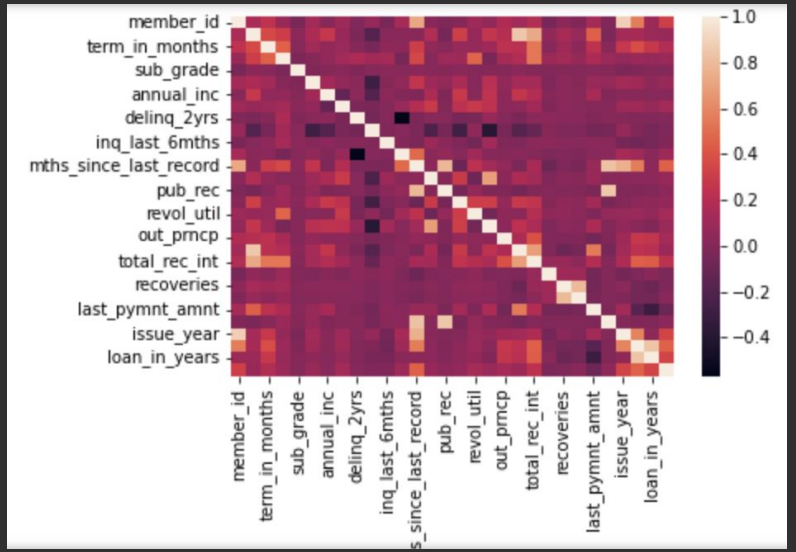
Most of the borrowers are from California area. It is observed that most borrowers are applying for individual loans either early(<=5 years) or late(10+years) in their carriers. People with LC assigned grade A/B have more chances of getting their loan approved.

According to the data given here, the probability of people in rented or mortgaged house applying for loans is higher.

In the data given, the borrowers have open credit lines between 6 and 12. Most of them have at max of 1 inquiry in the last 6 months. Though there are applicants who have been delinquent in the past, it has been 18 to 52 months since their delinquency. Also, most of the borrowers do not have derogatory public records.

PS - Using univariate analysis on addr_state, emp_length, grade, home_ownership

Derived metric



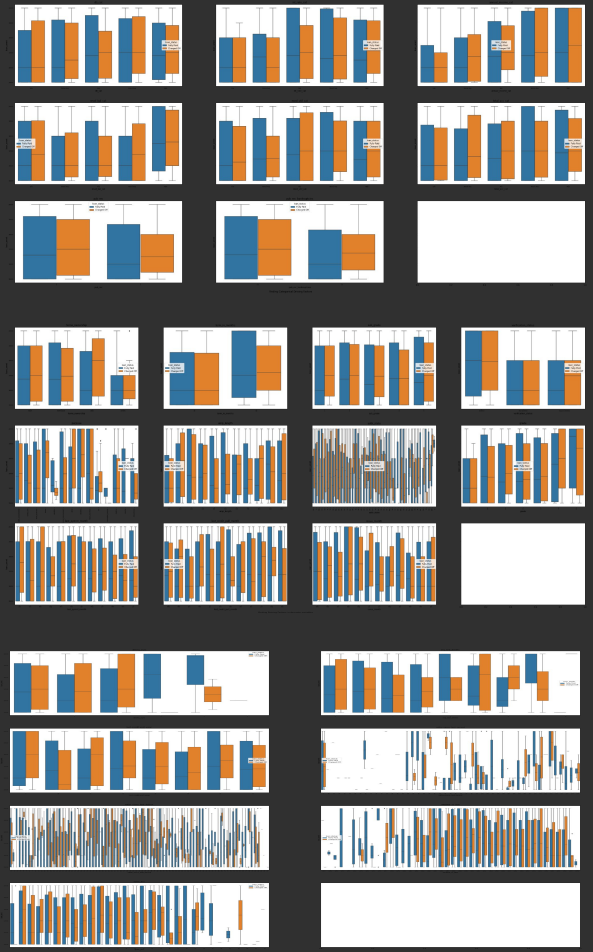
- Months and years were derived out of dates. Categorised continuous variables dti, int_rate, revol_util,revol_bal, total_acc, annual_inc.

Finding Driving Factors

Segment the loan data on Loan amount. The theory is that charged off loans pattern would be similar in a given loan amount range. So to begin with taking the segment with 25000 and above loans.

Public bankruptcies is a good place to start. By definition, probability of charged off loans are high when there is a bankruptcies.

Started by plotting all the categorical, and derived categorical(from continuous variables). By examining them for the segments for the loan amount segments, was able to find the driving factors.



Driving Factors

By definition given in Data Dictionary, the driving factors should be

- Loan Amount - Segment the loan details for analysis
- Pub_rec_bankruptcies
- Annual income
- Interest rate
- total_acc_cat
- inq_last_6mths
- Revol_bal_cat - derived from revol_bal
- Dti_cat - derived from dti
- Grade, subgrade
- Open_acc
- pub_rec
- Last credit pull year/month
- 'Delinq_2yrs'
- 'mths_since_last_delinq'

Proof of concept

By using the driving factors, was able to find the slice of data where fully paid loans are zero and charged off loans are significant.

This is just a POC and not the full implementation. It suggests that we can find segments of data by loan amount and investigate it with variables by definition are known to impact loan status and find the driving factors.

These driving factors are used to find subset of charged off loans where the fully paid loans are 0.