Gramener Case Study

By

Raghunandan Vellanki

Vyom Bhatt

Vasanthkumar Subramanian

Abstract

A consumer finance company, the largest online loan market place, has many risk applicants who default the loans and thereby causing credit loss to the company and its investors.

The primary objective of the study is to identify potential applicants who may cause credit loss to the company by defaulting the loan.

The analysis that we have done is

- Using EDA, understanding how the consumer and loan attributes influence the tendency of loan default
- Identifying the driving factors which cause the applicants to default the loans

Stages involved in the Analysis

- Data Cleansing and Manipulation
- Exploratory Data Analysis
- Summarizing the output and answering the business problem

Data Cleaning

For the analysis, we have taken into consideration the following **assumptions**:

- Nothing outside the scope of the analysis is to be considered for the desired output
- All the monetary figures in the dataset are in the same currency

In order to get the data prepared for our EDA, we went ahead and cleansed the data in the following manner:

- Excluded all the columns that were filled with over 90% null values
- On further observation, we got rid of columns that didn't seem fit for analysis based on the data description
- Dropping all columns that had only 1 unique value – as these won't be of any impact to our final output
- Corrected the format for the date columns to a date format for processing in Python
- Cleaning the columns such as:
  - Term column to exclude the ' months' string and change it to numeric
  - Employee length column to change it to numeric
  - Interest rate column to change it to decimal format
- Bucketing the interest rates, income levels, DTI for understanding trends in these variables later

Deriving New Metrics

Based on the data provided, we went ahead and derived new variables to assist our analysis further. These include:

- Extracting the year from issue date column
- Extracted the month from the issue date column

We went ahead and created a new metric that would give us a holistic view of how every variable is impacting the rate at which loans are getting charged off i.e **Charged-Off Index.**

This index takes into account the distribution of users that are charged off and the distribution of the total population i.e. pct_charged_off / pct_total_population. The reason for doing this is to ensure that we don't ignore the entire population set and just consider the charged_off/fully_paid users, as this will create a bias in case the population is skewed towards one entity

In addition to the above, we also created a business driven metric to understand the performace of each variable with regard to the nature of the business. The metric **payback percentage** which is essentially (total_rec_prncp / loan_amnt)*100

This metric will tell us what percentage of the loans are getting paid back – the lower it is, the riskier the loan gets as we haven't received the amount.

EDA

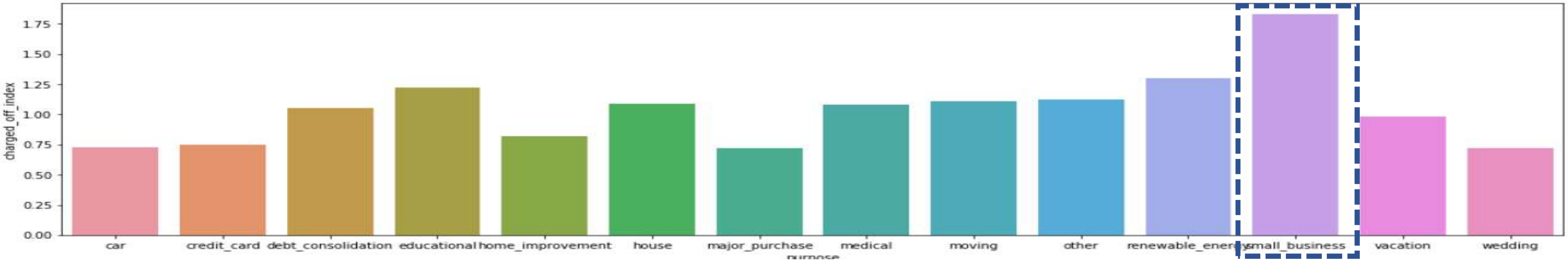
Running through the code submitted, we have gone through an extensive piece of data cleaning that involves:

- Data Cleaning and Preparation (as discussed previously)
- Univariate analysis
  - Outlier Analysis and treatment
  - Analyzing distributions for independent variables
- Segmented Univariate analysis
  - Analyzing independent variables by segmenting them with other variables to observe trends and impact on the dependent variable
- Bivariate Analysis
  - Analyzing multiple variables to identify how they impact the dependent variable
  - Prepared a correlation matrix for all the variables that we considered
- Using the Dervied Metrics to perform a deep-dive analysis on important variables identified

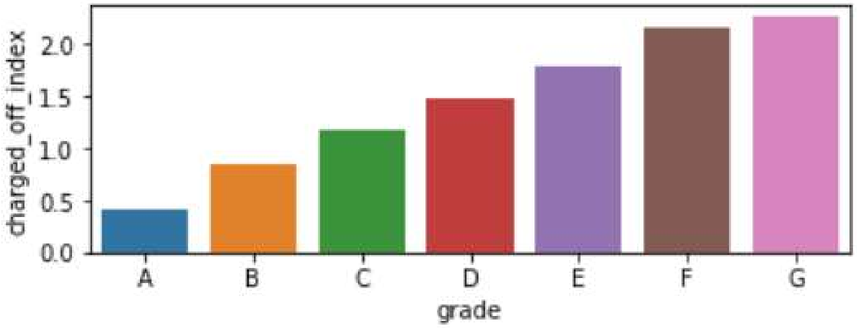
In the upcoming sections of this presentation, we will present the major findings of our EDA, and present the shortlisted variables that we identified as the most important in terms of impacting whether a loan would get defaulted or not

EDA Findings - 1

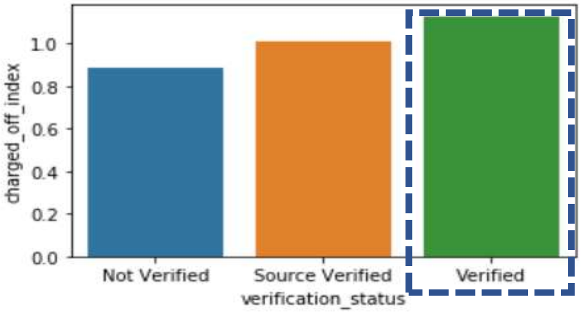
1. Among all the purposes, the 'Small Businesses' loan purpose is the riskiest with the highest Charged-off index, however the most loans are taken for the Debt Consolidation purpose



2. We observe a trend in the 'Grade' variable, as the grade reduces, the risk of the loan getting defaulted increases



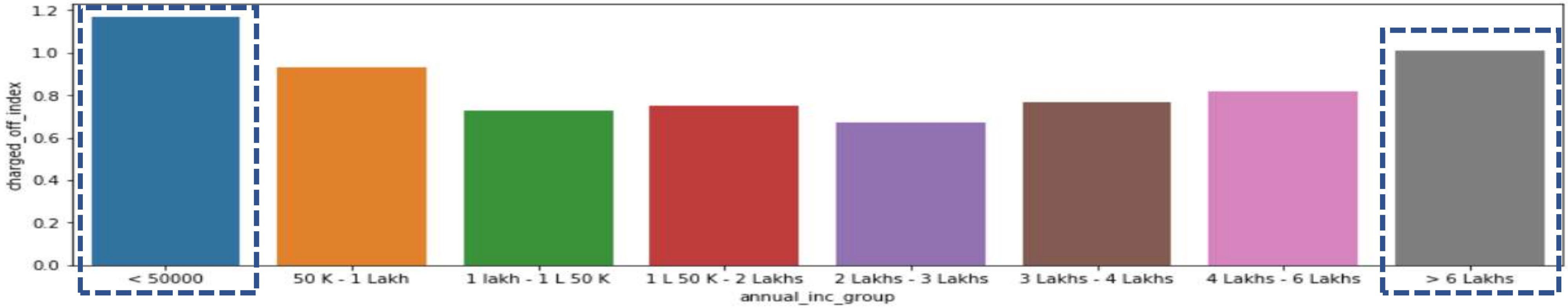
3. the data for verification is one that stands out as it goes against what one would expect: the loans with a verified status have a higher risk compared to not-verified ones.



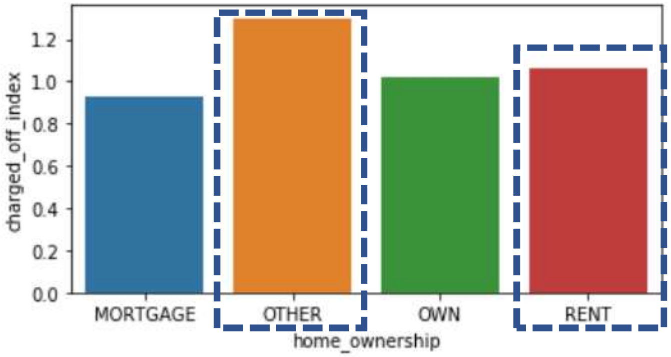
Perhaps the firm should look to make the verification process more stringent to improve this so that the verified applicants are actually ones that don't end up defaulting

EDA Findings - 2

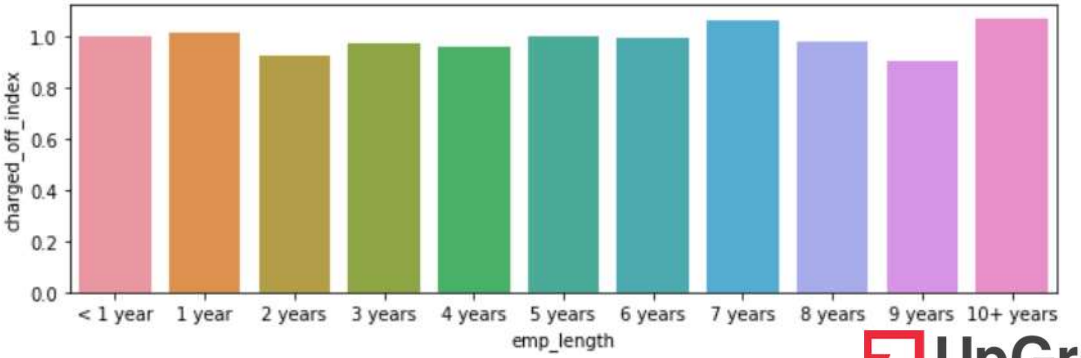
4. The income levels is **not** the most important variable in determining the chances of the loan getting defaulted as it doesn't follow any trend



5. For the home ownership variable, the **other** and **rent** are the riskiest home ownership types



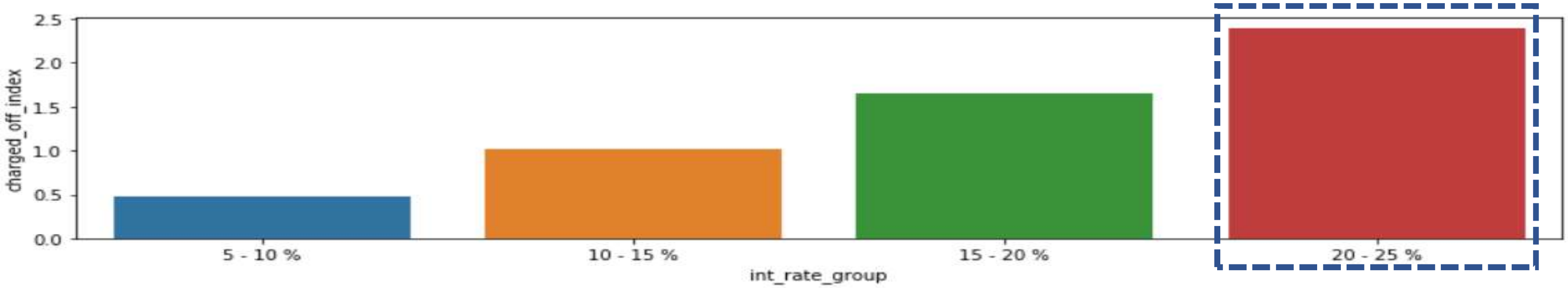
6. The employment length in years variable does not show to have any major impact on the chances of a loan getting charged-off



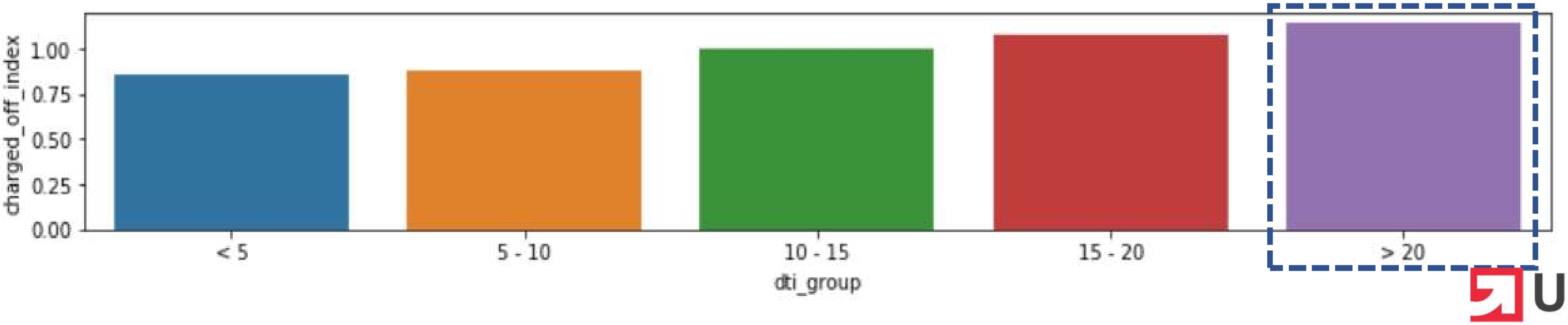
EDA Findings - 3

7. Analysis on the public bankruptcy level tells us that there is a high rate of loans getting defaulted for applicants who have a high number of bankruptcy. However, **94%** of the data provided was for users that had 0 numbers of public bankruptcy, hence actioning based on this wouldn't be advised due to population skewness

8. As the interest rate increases, the rate at which the users default also increases. This, however, could be as a result of the action that the bank takes for charging higher interest rates to users that are assumed to be of high risk



9. The DTI groups also have a direct impact on the rate at which loans get charged off. As the DTI rate increases, the loan gets more riskier

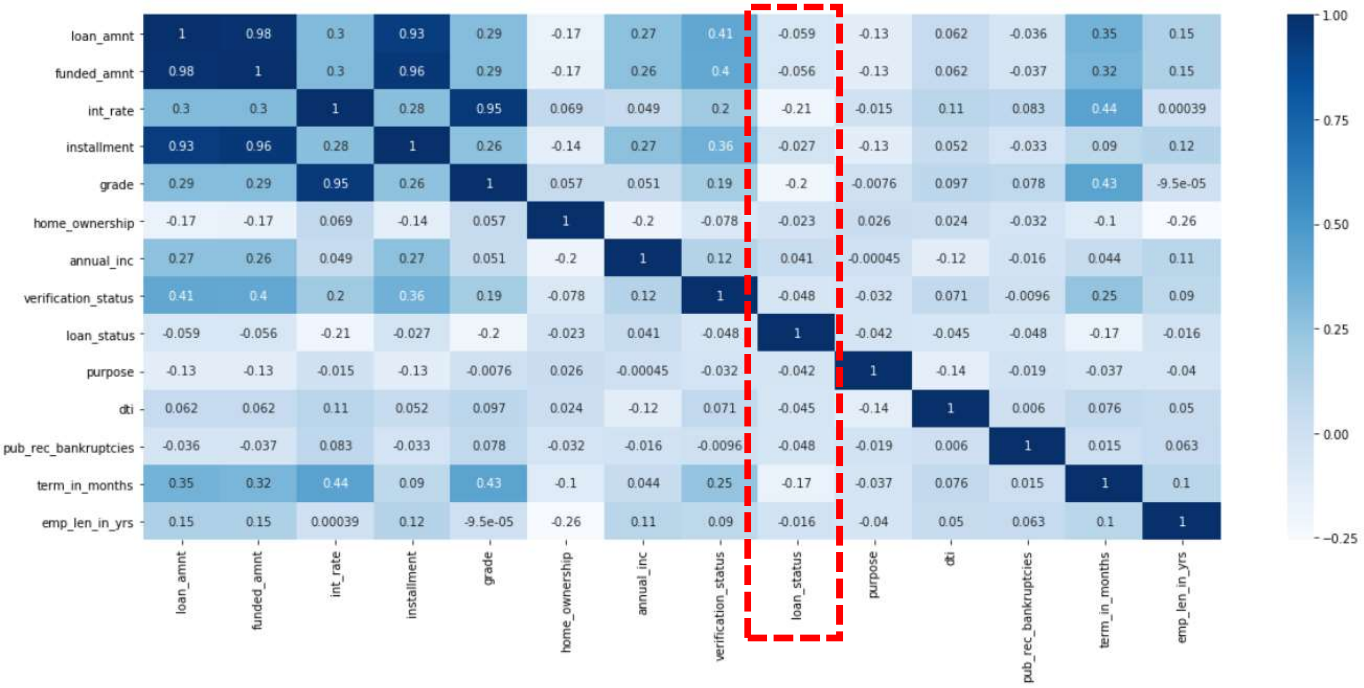


UpGrad

EDA Findings - 4

On performing Bi-variate analysis for every variable against each other, we observed correlations between the variables to see what variables are representing string relationships against each other.

The dataset with the dependent variable as loan status was filtered out only for loans that had status as ‘charged off’ or ‘fully paid’ to observe the relationship every other variable was having with respect to this dependent variable. The categorical variables here were fist converted to factors before the correlations were found.



	loan_amnt	funded_amnt	int_rate	installment	grade	home_ownership	annual_inc	verification_status	loan_status	purpose	dti	pub_rec_bankruptcies	term_in_months	emp_len_in_yrs
loan_amnt	1	0.98	0.3	0.93	0.29	-0.17	0.27	0.41	-0.059	-0.13	0.062	-0.036	0.35	0.15
funded_amnt	0.98	1	0.3	0.96	0.29	-0.17	0.26	0.4	-0.056	-0.13	0.062	-0.037	0.32	0.15
int_rate	0.3	0.3	1	0.28	0.95	0.069	0.049	0.2	-0.21	-0.015	0.11	0.083	0.44	0.00039
installment	0.93	0.96	0.28	1	0.26	-0.14	0.27	0.36	-0.027	-0.13	0.052	-0.033	0.09	0.12
grade	0.29	0.29	0.95	0.26	1	0.057	0.051	0.19	-0.2	-0.0076	0.097	0.078	0.43	-9.5e-05
home_ownership	-0.17	-0.17	0.069	-0.14	0.057	1	-0.2	-0.078	-0.023	0.026	0.024	-0.032	-0.1	-0.26
annual_inc	0.27	0.26	0.049	0.27	0.051	-0.2	1	0.12	0.041	-0.00045	-0.12	-0.016	0.044	0.11
verification_status	0.41	0.4	0.2	0.36	0.19	-0.078	0.12	1	-0.048	-0.032	0.071	-0.0096	0.25	0.09
loan_status	-0.059	-0.056	-0.21	-0.027	-0.2	-0.023	0.041	-0.048	1	-0.042	-0.045	-0.048	-0.17	-0.016
purpose	-0.13	-0.13	-0.015	-0.13	-0.0076	0.026	-0.00045	-0.032	-0.042	1	-0.14	-0.019	-0.037	-0.04
dti	0.062	0.062	0.11	0.052	0.097	0.024	-0.12	0.071	-0.045	-0.14	1	0.006	0.076	0.05
pub_rec_bankruptcies	-0.036	-0.037	0.083	-0.033	0.078	-0.032	-0.016	-0.0096	-0.048	-0.019	0.006	1	0.015	0.063
term_in_months	0.35	0.32	0.44	0.09	0.43	-0.1	0.044	0.25	-0.17	-0.037	0.076	0.015	1	0.1
emp_len_in_yrs	0.15	0.15	0.00039	0.12	-9.5e-05	-0.26	0.11	0.09	-0.016	-0.04	0.05	0.063	0.1	1

Looking at the correlation matrix above, we observe that the variables correlating the most with our dependent variable "loan status" are: interest rate, grade, and month.

However, all of these are at most mildly correlated with the dependent variable

The Driving Factors

Based on the exhaustive exploratory data analysis done above, the factors that we identified as the most crucial ones that impact the chances of loan defaults are:

- ❑ Grade
- ❑ Purpose
- ❑ Verification Status (This process needs to be looked at and ,made more stringent by the firm)
- ❑ Home Ownership
- ❑ Interest Rate
- ❑ DTI