upGrad

#LifeKoKaroLift

Lending Club Case Study

Learners:

Karishma Kumari

Vinay Karandi

upGrad

Content

- Problem statement and the analysis approach
- Results of univariate, bivariate analysis
- Visualizations and Summary

Problem Statement

What is Lending Club?

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

How Lending Club Works

Borrowers apply for loans.
Investors open an account.

Borrowers get funded.
Investors build a portfolio.

Borrowers repay automatically.
Investors earn & reinvest.

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:

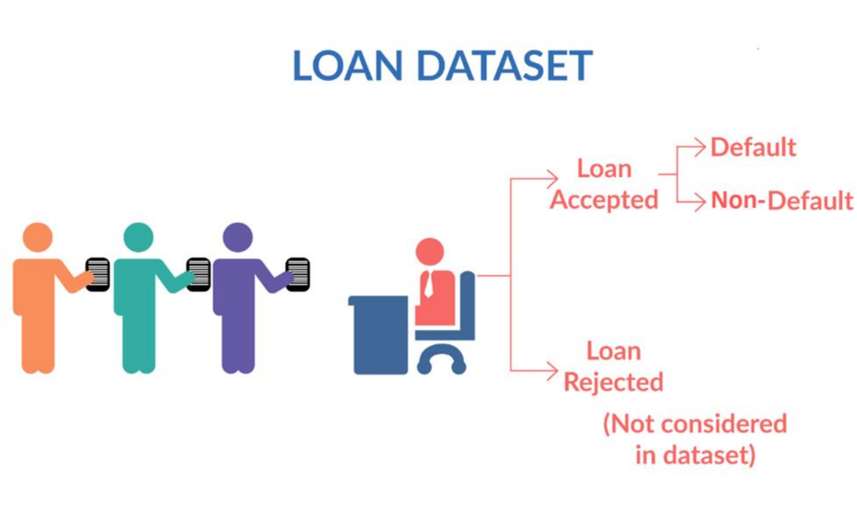
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company

Problem Statement

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 utilize this knowledge for its portfolio and risk assessment.

Problem Statement



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

Analysis Approach

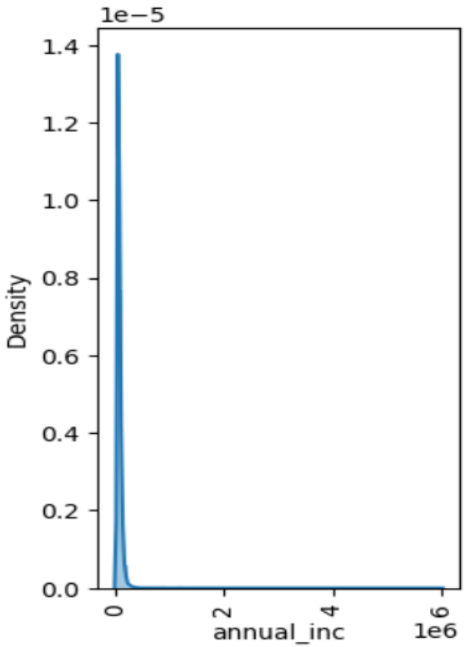
There are four major steps while doing the case study:

1. Data understanding
2. Data cleaning (cleaning missing values, removing redundant columns etc.)
3. Data Analysis- Univariate and bivariate analysis
4. Summary and conclusion

Data Cleaning

1. Checked the duplicate values
2. Checked the percentage of missing values – the columns having more than 20% NAN values are removed
3. Dropped the rows having NAN values
4. Fixed the datatype of numerical and date columns
5. Removed the columns having just 1 unique value from the analysis

Data Analysis – Univariate Analysis – Continuous variables



From Univariate analysis of annual_inc variable, it is observed that the data for this is skewed. Hence, creating buckets for this variable.

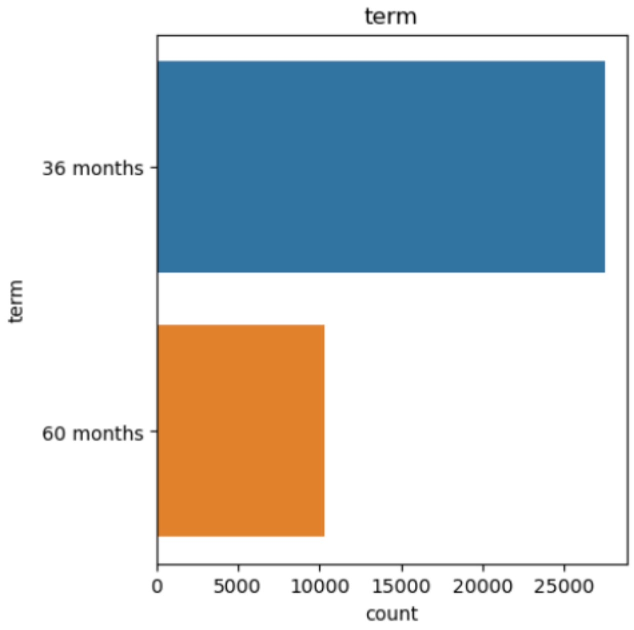
Similarly, buckets have been created for the following variables:

1. Annual income
2. delinq_2yrs
3. pub_rec
4. out_prncp_inv
5. total_rec_late_fee
6. recoveries
7. collection_recovery_fee
8. pub_rec_bankruptcies

Lending Club: EDA Case Study

upGrad

Data Analysis – Univariate Analysis – Categorical variables

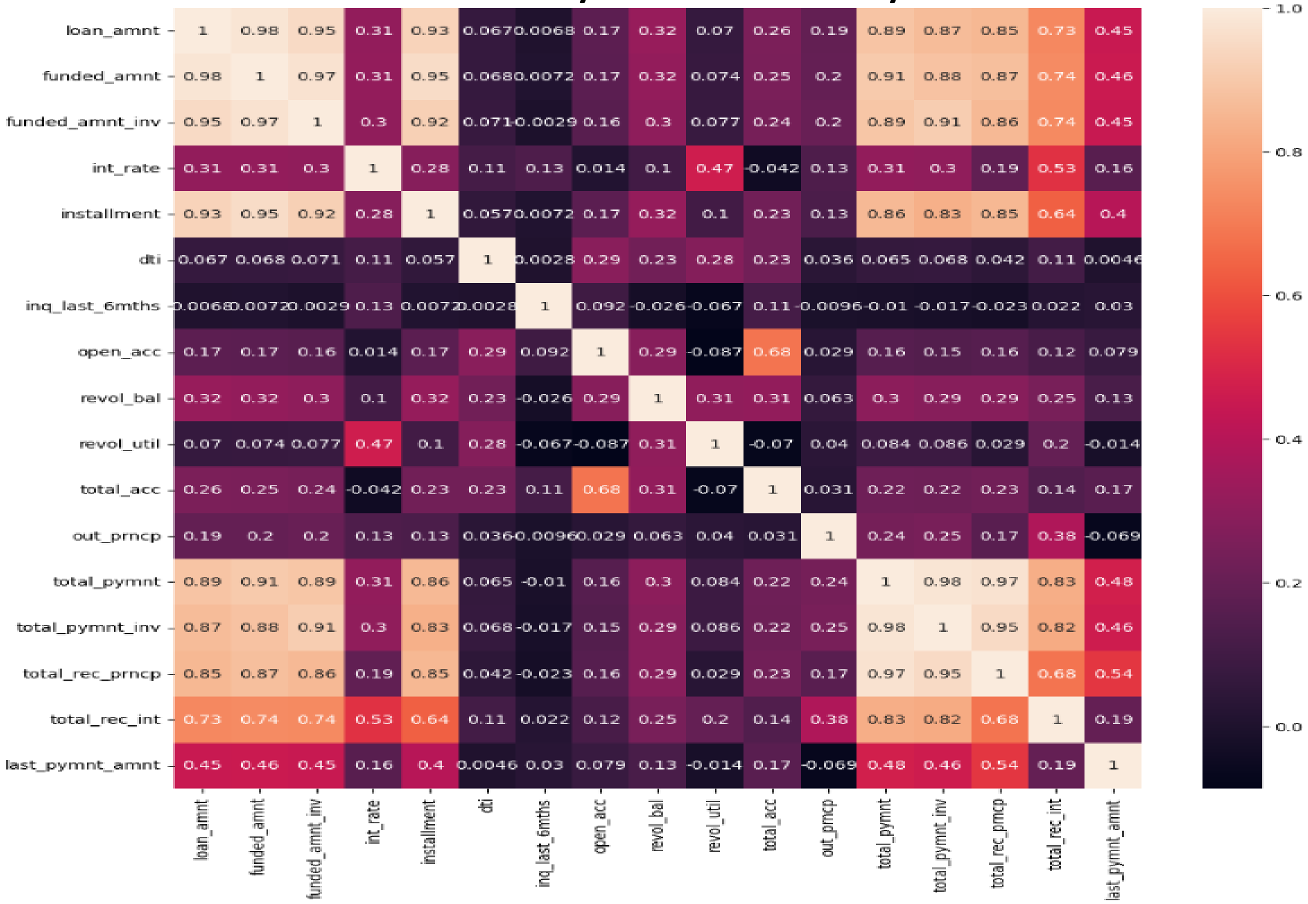


To analyse the categorical variables, countplot has been used to see the frequency of each variable.

Lending Club: EDA Case Study

upGrad

Data Analysis – Bivariate Analysis – Continuous variables

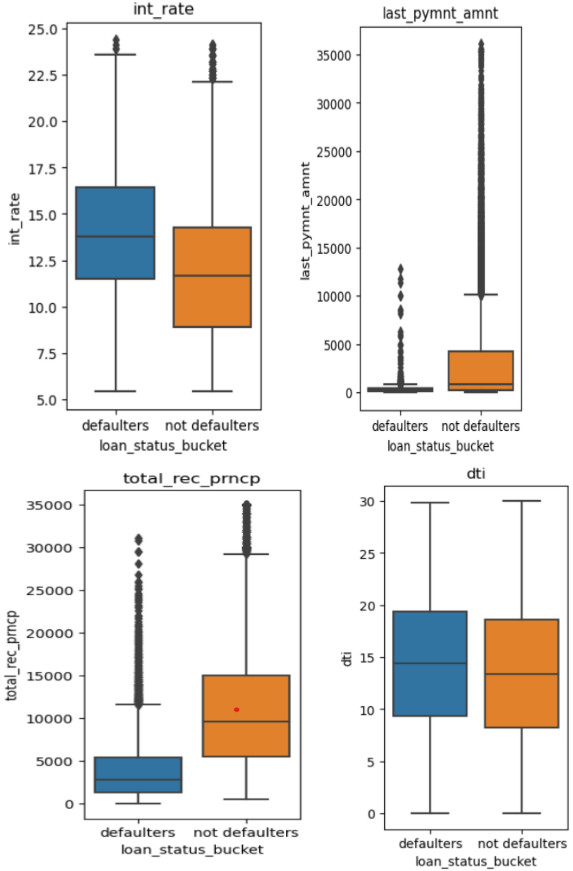


	loan_amnt	funded_amnt	funded_amnt_inv	int_rate	installment	dti	inq_last_6mths	open_acc	revol_bal	revol_util	total_acc	out_prncp	total_pymnt	total_pymnt_inv	total_rec_prncp	total_rec_int	last_pymnt_amnt
loan_amnt	1	0.98	0.95	0.31	0.93	0.067	0.0068	0.17	0.32	0.07	0.26	0.19	0.89	0.87	0.85	0.73	0.45
funded_amnt	0.98	1	0.97	0.31	0.95	0.068	0.0072	0.17	0.32	0.074	0.25	0.2	0.91	0.88	0.87	0.74	0.46
funded_amnt_inv	0.95	0.97	1	0.3	0.92	0.071	-0.0029	0.16	0.3	0.077	0.24	0.2	0.89	0.91	0.86	0.74	0.45
int_rate	0.31	0.31	0.3	1	0.28	0.11	0.13	0.014	0.1	0.47	-0.042	0.13	0.31	0.3	0.19	0.53	0.16
installment	0.93	0.95	0.92	0.28	1	0.057	0.0072	0.17	0.32	0.1	0.23	0.13	0.86	0.83	0.85	0.64	0.4
dti	0.067	0.068	0.071	0.11	0.057	1	0.0028	0.29	0.23	0.28	0.23	0.036	0.065	0.068	0.042	0.11	0.0046
inq_last_6mths	0.0068	0.0072	0.0029	0.13	0.0072	0.0028	1	0.092	-0.026	-0.067	0.11	-0.0096	-0.01	-0.017	-0.023	0.022	0.03
open_acc	0.17	0.17	0.16	0.014	0.17	0.29	0.092	1	0.29	-0.087	0.68	0.029	0.16	0.15	0.16	0.12	0.079
revol_bal	0.32	0.32	0.3	0.1	0.32	0.23	-0.026	0.29	1	0.31	0.31	0.063	0.3	0.29	0.29	0.25	0.13
revol_util	0.07	0.074	0.077	0.47	0.1	0.28	-0.067	-0.087	0.31	1	-0.07	0.04	0.084	0.086	0.029	0.2	-0.014
total_acc	0.26	0.25	0.24	-0.042	0.23	0.23	0.11	0.68	0.31	-0.07	1	0.031	0.22	0.22	0.23	0.14	0.17
out_prncp	0.19	0.2	0.2	0.13	0.13	0.036	-0.0096	0.029	0.063	0.04	0.031	1	0.24	0.25	0.17	0.38	-0.069
total_pymnt	0.89	0.91	0.89	0.31	0.86	0.065	-0.01	0.16	0.3	0.084	0.22	0.24	1	0.98	0.97	0.83	0.48
total_pymnt_inv	0.87	0.88	0.91	0.3	0.83	0.068	-0.017	0.15	0.29	0.086	0.22	0.25	0.98	1	0.95	0.82	0.46
total_rec_prncp	0.85	0.87	0.86	0.19	0.85	0.042	-0.023	0.16	0.29	0.029	0.23	0.17	0.97	0.95	1	0.68	0.54
total_rec_int	0.73	0.74	0.74	0.53	0.64	0.11	0.022	0.12	0.25	0.2	0.14	0.38	0.83	0.82	0.68	1	0.19
last_pymnt_amnt	0.45	0.46	0.45	0.16	0.4	0.0046	0.03	0.079	0.13	-0.014	0.17	-0.069	0.48	0.46	0.54	0.19	1

To analyse which all Continuous variables are correlated, correlation matrix has been used. It can be observed that the following set of variables are high correlated:

- Loan Amnt, Funded Amnt and Funded Amnt_inv are highly correlated
- 'total_pymnt', 'total_pymnt_inv' and 'total_rec_prncp' are highly correlated
- funded_amnt and total_pymnt are highly correlated
- total_pymnt_inv and funded_amnt_inv are highly correlated

Data Analysis – Bivariate Analysis – Continuous variables



To analyse the Continuous variables' impact on loan_status, box plots for each level of loan_status_bucket (categorical variable) has been drawn.
It can be observed that the following Continuous variables are influencing loan_status_bucket:

- int_rate, dti, inq_lat_6mths, revol_util, total_pymnt, total_pymnt_inv, total_rec_prncp and last_pymnt_amnt

Since from the data dictionary definitions, we understood that the following variables are the customer behavior variables which are not available at the time of loan application, and thus they cannot be used as predictors for credit approval. Hence, they can not be taken into consideration to identify the chances of defaults and the risky loan applicants:

- total_pymnt, total_pymnt_inv, total_rec_prncp and last_pymnt_amnt

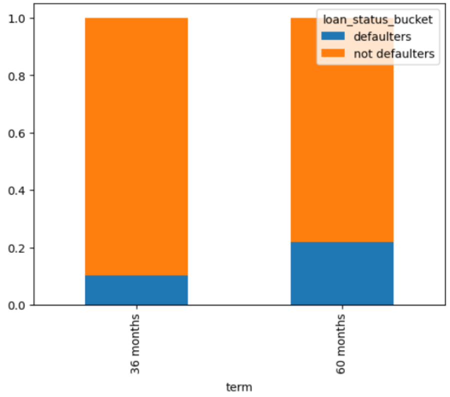
Hence, the final list of continuous variables influencing loan_status_bucket are:

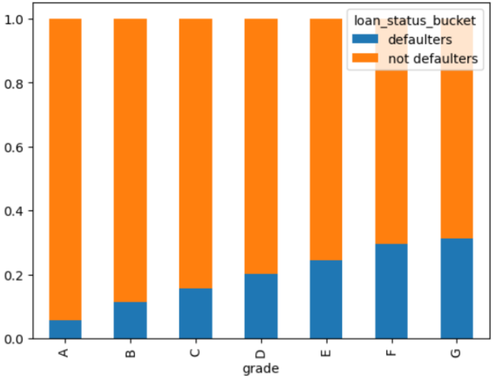
- **int_rate, dti, inq_lat_6mths, revol_util**

Lending Club: EDA Case Study

upGrad

Data Analysis – Bivariate Analysis – Categorical variables





To analyse the Categorical variables’ impact on loan_status, stacked bar charts for different categories have been created.

It can be observed that the following Categorical variables are influencing loan_status_bucket:

- Term, grade, sub grade, purpose, addr_state, zip code, ann_inc_bucket, 'out_prncp_cat’, 'out_prncp_inv_cat’, 'total_rec_late_fee_cat’, 'recoveries_cat’, 'collection_recovery_fee_cat'

Since from the data dictionary definitions, we understood that the following variables are the customer behavior variables which are not available at the time of loan application, and thus they cannot be used as predictors for credit approval. Hence, they can not be taken into consideration to identify the chances of defaults and the risky loan applicants:

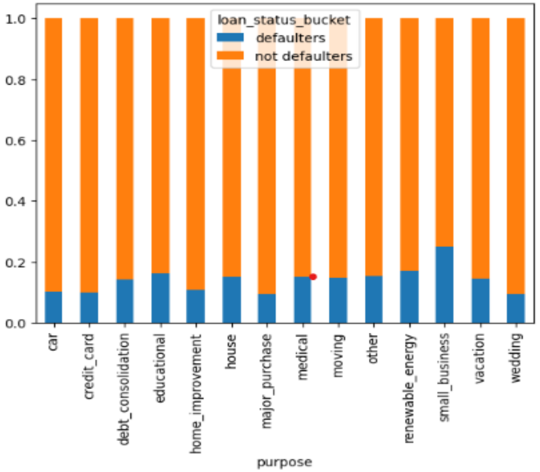
- 'out_prncp_cat’, out_prncp_inv_cat, total_rec_late_fee_cat, recoveries_cat, collection_recovery_fee_cat

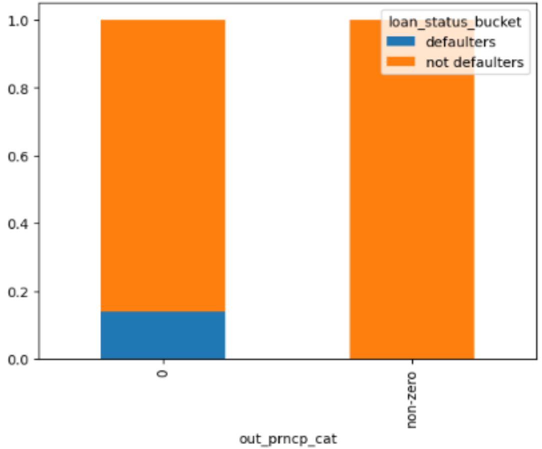
Also, since grade and sub-grade are related, taking the most granular info into consideration i.e. Sub-Grade. Similarly, since addr_state and zip code are geographic related variables, taking the most granular info into consideration i.e. zip code.

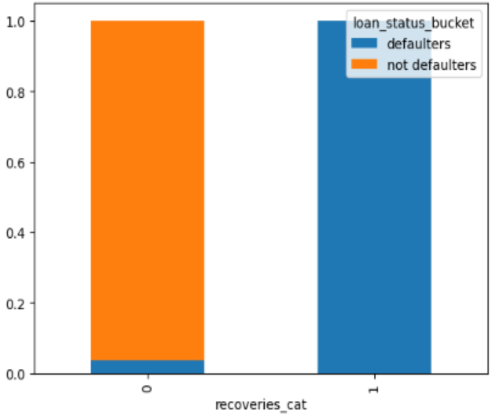
Hence, the final list of categorical variables influencing loan_status_bucket are:

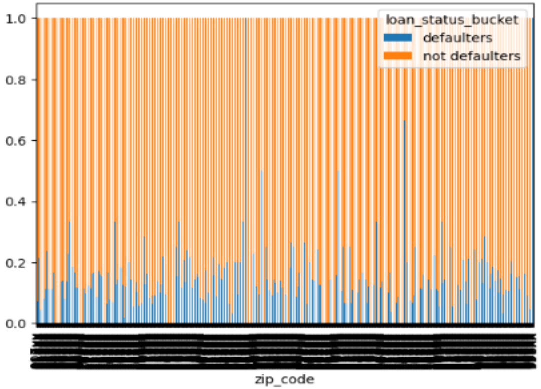
- **Term, sub grade, purpose, zip code, ann_inc_bucket**

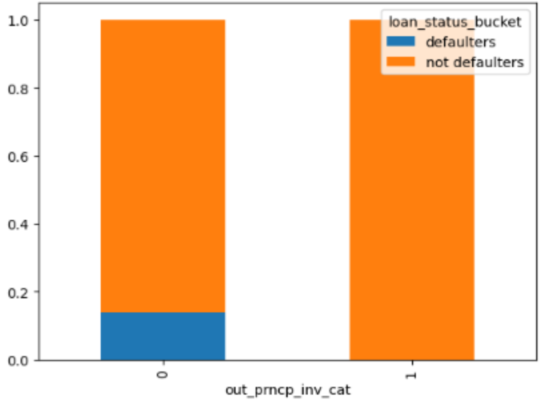
Data Analysis – Bivariate Analysis – Categorical variables

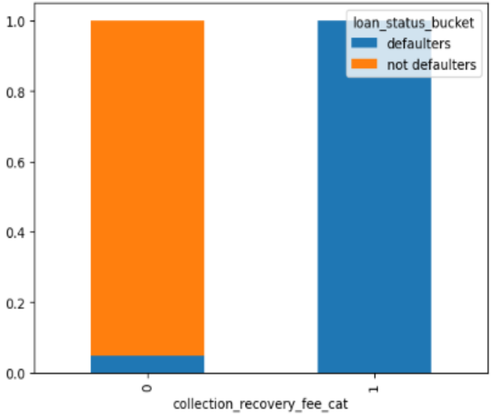












Conclusion

Final list of categorical variables influencing loan_status_bucket are:

- term
- sub grade
- purpose
- zip code

Final list of continuous variables influencing loan_status_bucket are:

- int_rate
- dti
- inq_lat_6mths
- revol_util
- ann_inc