upGrad

Lending Club Loan data analysis for preliminary recommendation to investors as observed in sample data set

– EDA based solution approach

02-07-2022

upGrad

Course : Machine Learning

Batch : ML C41

Team : Manish Mandal

upGrad

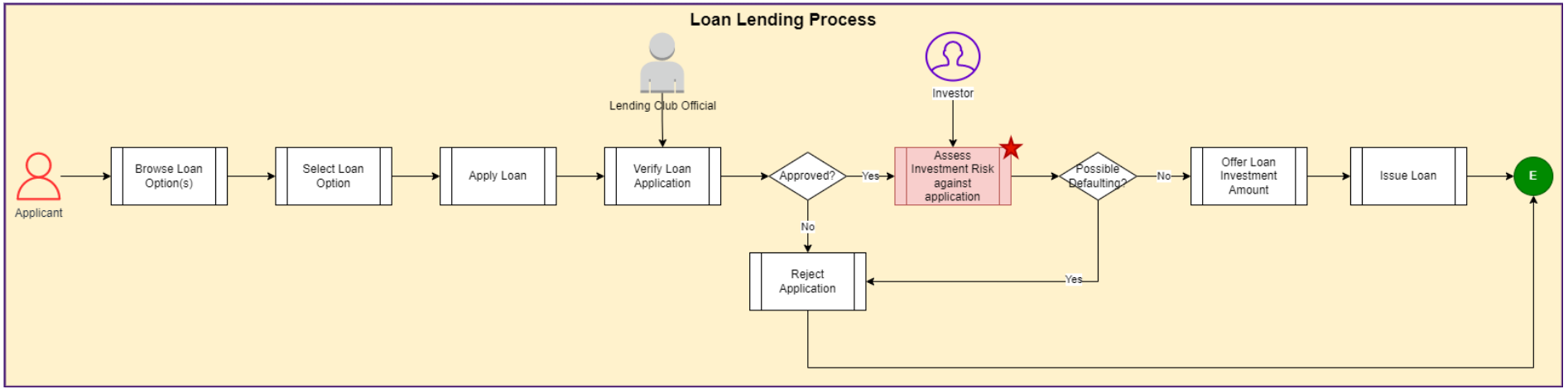
- Overview
- Understanding and scope of the assignment
- Analysis
- Preliminary observation and recommendation

Lending Club: Overview

upGrad

What is Lending Club?

It is a marketplace bank for loans that matches borrowers who are seeking a loan to achieve their goals with investors looking to lend money and make a return.



Problem domain of interest to assess for best judgement of investment and issue loan to appropriate applicants with minimal risk as much as possible

Lending Club: Scope of the assignment

upGrad

Overview and Scope

- To assess the risk and identify applicants a data driven approach is needed to establish a model that can segregate with defined confidence level between possible defaulter and non-defaulter cases
- As a first step towards building a model from a representative dataset (population) – the data needs to be understood, analysed and made ready for subsequent step of model building
- The current scope of assignment is limited to perform data cleaning and analysis, which follows a preliminary recommendation based on data analysis
- An Exploratory Data Analysis (EDA) needs to be performed to preliminary screening of risky loan application from sample data set

Lending Club: Sample Data Set

upGrad

Data Summary

- A representative data set in comma separated file format consisting of past sanctioned loan details
- Each loan record contains representative loan information containing key information like:
  - Applicant's demographic info (e.g., Residential status, Employment info, Salary etc.)
  - Applied loan details (e.g., Type, Description, Interest rate, Terms, Grade, Applied Amount etc.)
  - Sanctioned amount (e.g., Loan invested amount etc.)
  - Repayment status
- There are 111 columns and 39717 rows available in provided dataset
- The data definition dictionary is mentioned in the attached document

Lending Club: Sample Data Set Definition

upGrad

Data Definition

#	Data Element	Description
1	acc_now_delinq	The number of accounts on which the borrower is now delinquent.
2	acc_open_past_24mths	Number of trades opened in past 24 months.
3	addr_state	The state provided by the borrower in the loan application
4	all_util	Balance to credit limit on all trades
5	annual_inc	The self-reported annual income provided by the borrower during registration.
6	annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration
7	application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
8	avg_cur_bal	Average current balance of all accounts
9	bc_open_to_buy	Total open to buy on revolving bankcards.
10	bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.
11	chargeoff_within_12_mths	Number of charge-offs within 12 months
12	collection_recovery_fee	post charge off collection fee
13	collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
14	delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
15	delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
16	desc	Loan description provided by the borrower
17	dti	A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income.
18	dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income
19	earliest_cr_line	The month the borrower's earliest reported credit line was opened
20	emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.

Lending Club: Sample Data Set Definition

upGrad

Data Definition contd.

#	Data Element	Description
21	emp_title	The job title supplied by the Borrower when applying for the loan.*
22	fico_range_high	The upper boundary range the borrower's FICO at loan origination belongs to.
23	fico_range_low	The lower boundary range the borrower's FICO at loan origination belongs to.
24	funded_amnt	The total amount committed to that loan at that point in time.
25	funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
26	Grade	LC assigned loan grade
27	home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
28	id	A unique LC assigned ID for the loan listing.
29	il_util	Ratio of total current balance to high credit/credit limit on all install acct
30	initial_list_status	The initial listing status of the loan. Possible values are – W, F
31	inq_fi	Number of personal finance inquiries
32	inq_last_12m	Number of credit inquiries in past 12 months
33	inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
34	installment	The monthly payment owed by the borrower if the loan originates.
35	int_rate	Interest Rate on the loan
36	issue_d	The month which the loan was funded
37	last_credit_pull_d	The most recent month LC pulled credit for this loan
38	last_fico_range_high	The upper boundary range the borrower's last FICO pulled belongs to.
39	last_fico_range_low	The lower boundary range the borrower's last FICO pulled belongs to.
40	last_pymnt_amnt	Last total payment amount received
41	last_pymnt_d	Last month payment was received

Lending Club: Sample Data Set Definition

upGrad

Data Definition contd.

#	Data Element	Description
42	loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
43	loan_status	Current status of the loan
44	max_bal_bc	Maximum current balance owed on all revolving accounts
45	member_id	A unique LC assigned Id for the borrower member.
46	mo_sin_old_il_acct	Months since oldest bank installment account opened
47	mo_sin_old_rev_tl_op	Months since oldest revolving account opened
48	mo_sin_rcnt_rev_tl_op	Months since most recent revolving account opened
49	mo_sin_rcnt_tl	Months since most recent account opened
50	mort_acc	Number of mortgage accounts.
51	mths_since_last_delinq	The number of months since the borrower's last delinquency.
52	mths_since_last_major_derog	Months since most recent 90-day or worse rating
53	mths_since_last_record	The number of months since the last public record.
54	mths_since_rcnt_il	Months since most recent installment accounts opened
55	mths_since_recent_bc	Months since most recent bankcard account opened.
56	mths_since_recent_bc_dlq	Months since most recent bankcard delinquency
57	mths_since_recent_inq	Months since most recent inquiry.
58	mths_since_recent_revol_delinq	Months since most recent revolving delinquency.
59	next_pymnt_d	Next scheduled payment date
60	num_accts_ever_120_pd	Number of accounts ever 120 or more days past due
61	num_actv_bc_tl	Number of currently active bankcard accounts
62	num_actv_rev_tl	Number of currently active revolving trades

Lending Club: Sample Data Set for Analysis

upGrad

Data Definition contd.

#	Data Element	Description
63	num_bc_sats	Number of satisfactory bankcard accounts
64	num_bc_tl	Number of bankcard accounts
65	num_il_tl	Number of installment accounts
66	num_op_rev_tl	Number of open revolving accounts
67	num_rev_accts	Number of revolving accounts
68	num_rev_tl_bal_gt_0	Number of revolving trades with balance >0
69	num_sats	Number of satisfactory accounts
70	num_tl_120dpd_2m	Number of accounts currently 120 days past due (updated in past 2 months)
71	num_tl_30dpd	Number of accounts currently 30 days past due (updated in past 2 months)
72	num_tl_90g_dpd_24m	Number of accounts 90 or more days past due in last 24 months
73	num_tl_op_past_12m	Number of accounts opened in past 12 months
74	open_acc	The number of open credit lines in the borrower's credit file.
75	open_acc_6m	Number of open trades in last 6 months
76	open_il_12m	Number of installment accounts opened in past 12 months
77	open_il_24m	Number of installment accounts opened in past 24 months
78	open_il_6m	Number of currently active installment trades
79	open_rv_12m	Number of revolving trades opened in past 12 months
80	open_rv_24m	Number of revolving trades opened in past 24 months
81	out_prncp	Remaining outstanding principal for total amount funded
82	out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors
83	pct_tl_nvr_dlq	Percent of trades never delinquent
84	percent_bc_gt_75	Percentage of all bankcard accounts > 75% of limit.
85	policy_code	1. publicly available policy_code=1 2. new products not publicly available policy_code=2

Lending Club: Sample Data Set Definition

upGrad

Data Definition contd.

#	Data Element	Description
86	pub_rec	Number of derogatory public records
87	pub_rec_bankruptcies	Number of public record bankruptcies
88	purpose	A category provided by the borrower for the loan request.
89	pymnt_plan	Indicates if a payment plan has been put in place for the loan
90	recoveries	post charge off gross recovery
91	revol_bal	Total credit revolving balance
92	revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
93	sub_grade	LC assigned loan subgrade
94	tax_liens	Number of tax liens
95	term	The number of payments on the loan. Values are in months and can be either 36 or 60.
96	title	The loan title provided by the borrower
97	tot_coll_amt	Total collection amounts ever owed
98	tot_cur_bal	Total current balance of all accounts
99	tot_hi_cred_lim	Total high credit/credit limit
100	total_acc	The total number of credit lines currently in the borrower's credit file
101	total_bal_ex_mort	Total credit balance excluding mortgage
102	total_bal_il	Total current balance of all installment accounts
103	total_bc_limit	Total bankcard high credit/credit limit
104	total_cu_tl	Number of finance trades
105	total_il_high_credit_limit	Total installment high credit/credit limit
106	total_pymnt	Payments received to date for total amount funded

Lending Club: Sample Data Set Definition

upGrad

Data Definition contd.

#	Data Element	Description
107	total_pymnt_inv	Payments received to date for portion of total amount funded by investors
108	total_rec_int	Interest received to date
109	total_rec_late_fee	Late fees received to date
110	total_rec_prncp	Principal received to date
111	total_rev_hi_lim	Total revolving high credit/credit limit
112	url	URL for the LC page with listing data.
113	verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
114	verified_status_joint	Indicates if the co-borrowers' joint income was verified by LC, not verified, or if the income source was verified
115	zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.

Lending Club: Sample Data Set for Analysis

upGrad

Raw Data Profile without any cleaning

#	Parameter	Value
1.	Row	39717
2.	Columns	111
3.	Float	74
4.	Int	13
5.	Object	24
6.	Number of columns with 80% or more missing values	56

Lending Club: Sample Data Set for Analysis

upGrad

Step 1: Data Cleansing/ Sanitization

#	Task	Action	Rational	Outcome
1.	Missing value treatment	Columns having data set more than or equal to 80% are dropped from data set	Columns having more than 80% missing value will not be worth considering for analysis as these neither can be imputed nor collected from source for the given scenario	56 out of 111 columns are removed from the data set
2.	Record Cleaning	In the data set there are three categories of data set based on status “Fully Paid”, “Current” and “Charged Off”. Only loan records having status “Fully Paid” and “Charged Off” are of interest for EDA	Loan records having status “Current” is not considered for the analysis as these are ongoing loans and do not provide any conclusive decision whether applicant has defaulted or fully paid	After filtration there are 38577 out of 39717 records are considered for subsequent studies
3.	Removal of behavioural attribute	Attributes such as "delinq_2yrs","earliest_cr_line","inq_last_6mths” etc. are removed from the data set. There are 21 such data elements in the population.	The customer behavior variables are not available at the time of loan application, and thus they cannot be used as predictors for credit approval.	34 columns are retained for subsequent analysi

Lending Club: Sample Data Set for Analysis

upGrad

Step 1: Data Cleansing/ Sanitization

#	Task	Action	Rational	Outcome
4.	Remove redundant columns	Data elements such as "id","member_id","url","desc","initial_list_status","emp_title“ are removed from data set	It is assumed that there are few other attributes that do not have any role in modelling	6 more data elements have been removed resulting in 28 data elements remaining

Data Profile after cleansing

#	Parameter	Value
1.	Row	38577
2.	Columns	28
6.	Number of columns with missing values	6

Lending Club: Sample Data Set for Analysis

upGrad

Step 2: Data Conversion/ Scaling

#	Task	Action	Rational	Outcome
1.	Data Conversion	Interest are converted to numeric values	Loan interest rate is an important attribute and it is a continuous value that is currently in object format sue to '%' character and it needs to be converted to float	Converted float amount

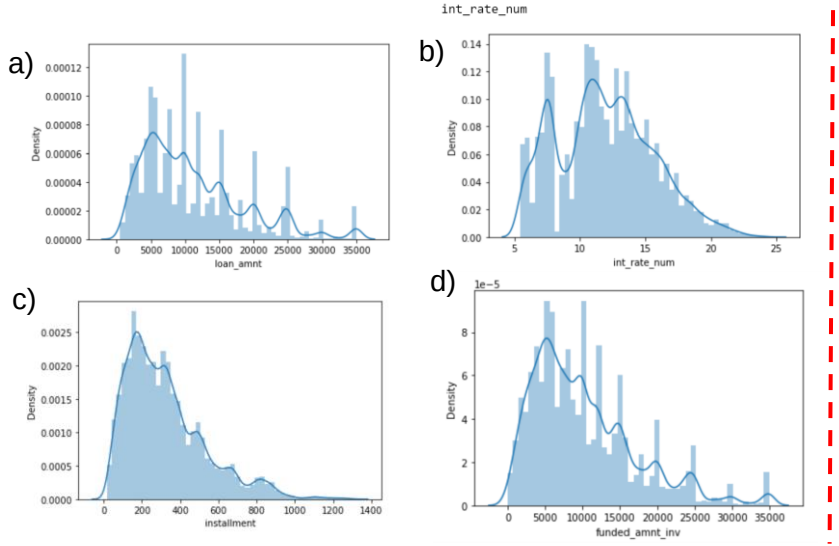
Data Profile after cleansing

#	Parameter	Value
1.	Row	38577
2.	Columns	28
6.	Number of columns with missing values	6

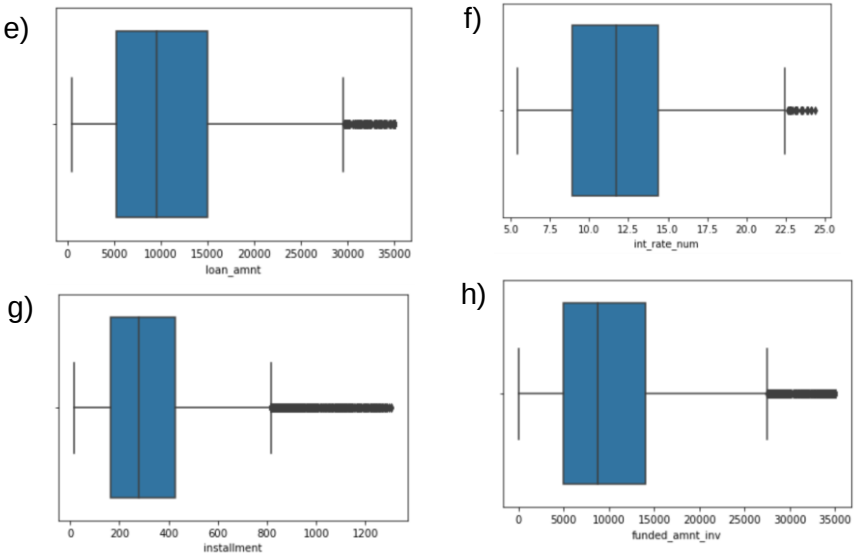
Lending Club: Sample Data Set for Analysis

upGrad

Step 3: Univariate Analysis of continuous variables



Plotting distribution for different key continuous variables



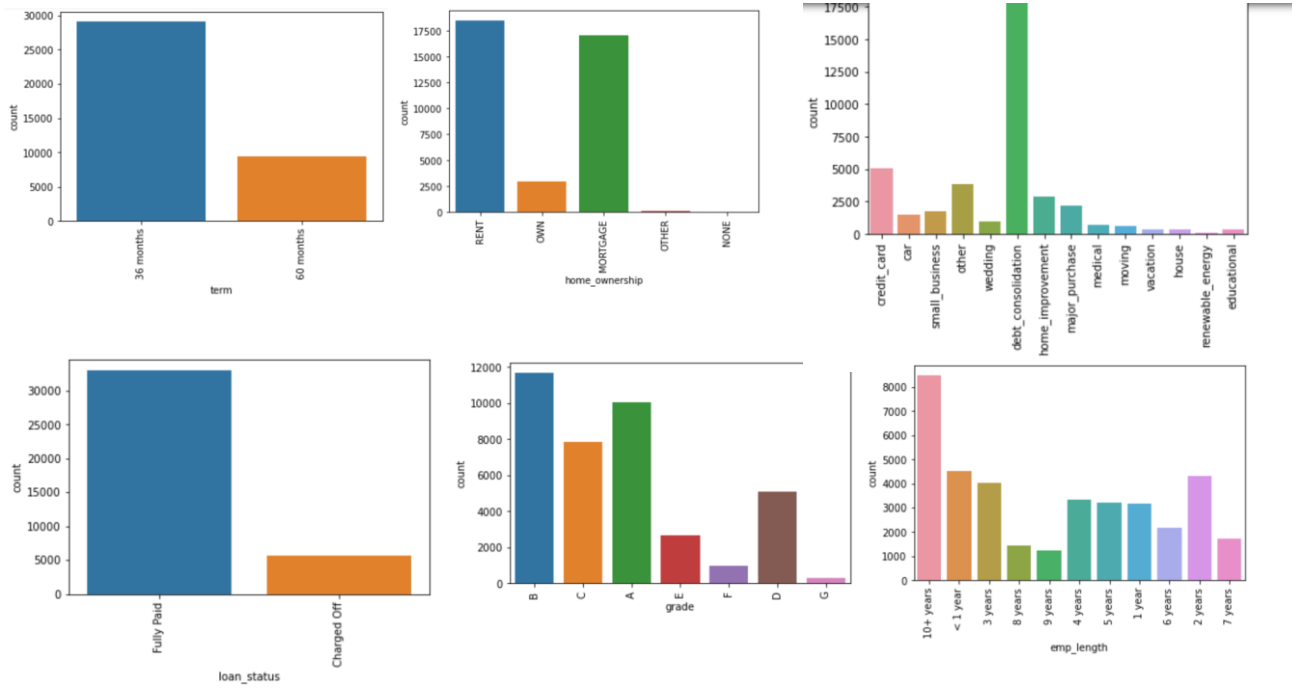
Plotting box for different key continuous variables

- Majority of the applicants applied loan between 5000 unit to 15000 unit as observed in (a)
- Most of the investors felt safe to invest in loan amount between 5000 unit to 15000 unit as observed in (d)
- Majority of applicants opted for loans of interest rate between 10% to 15% and repayment term as ~200 to 400 as observed in (b) and (c) respectively
- The above observations are also depicted through box plots (e), (f), (g) and (h)

Lending Club: Sample Data Set for Analysis

upGrad

Step 3: Univariate Analysis of categorical variables



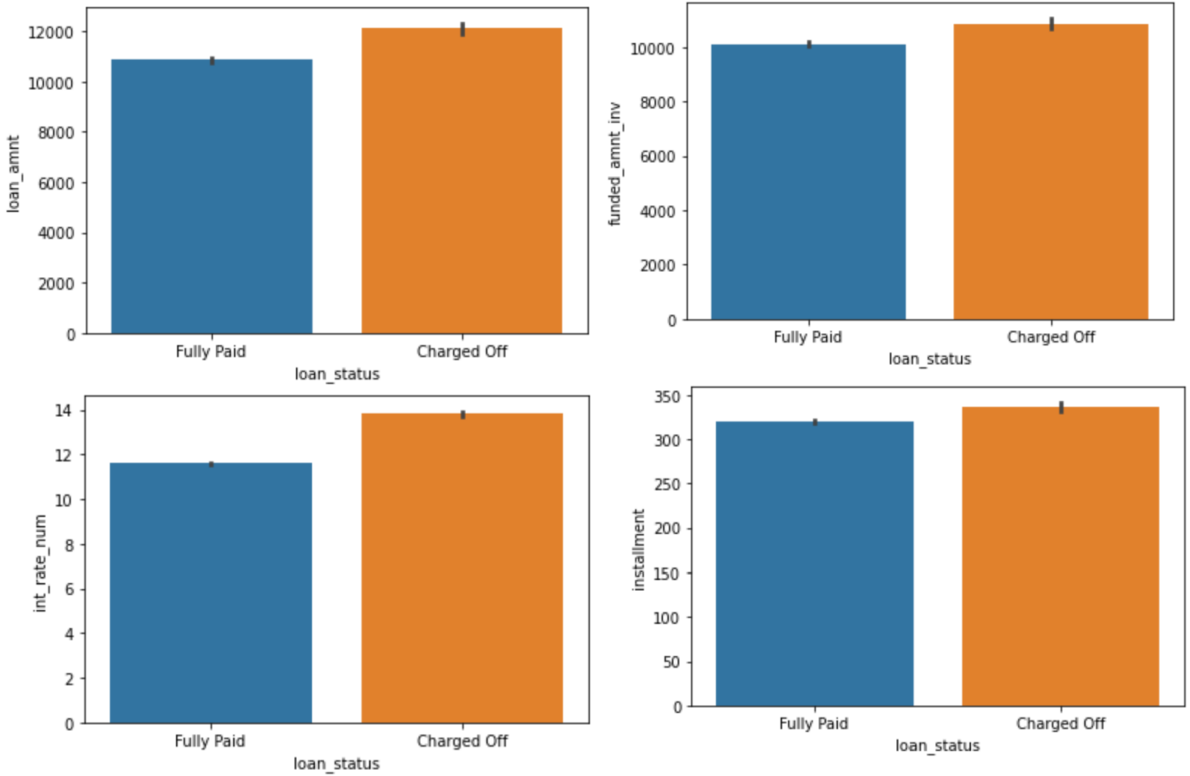
Plotting bar for different key categorical variables

- Applicants preferred shorter term length based loan
- It has been observed that loan has been applied or issued to applicants having home ownership as “RENT” or “MORTGAGE”.
- Most of the applicants have fully paid loan
- Grade “B” loan has been observed for most applications
- Loan is applied and/or issued to applicant’s having more than 10 years of employment length

Lending Club: Sample Data Set for Analysis

upGrad

Step 3: Bivariate Analysis between Loan Status and Categorical variables



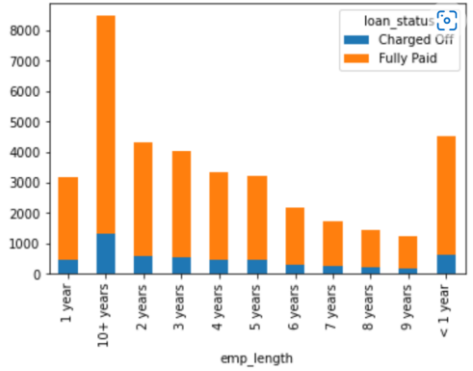
- Number of applicant's defaulting is higher with larger value of applied loan
- Investment in higher amount of loan may result in more chances of defaulting
- Chances of defaulting with high interest rate is more
- Chances of defaulting increases with increase in installment number

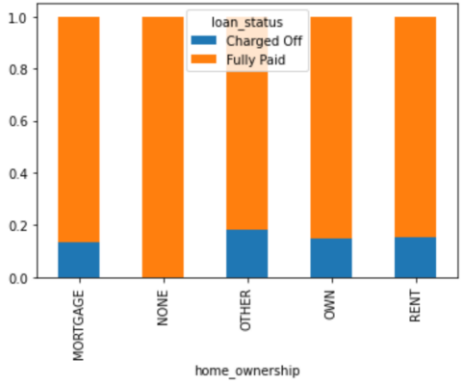
Plotting between loan status and different key categorical variables

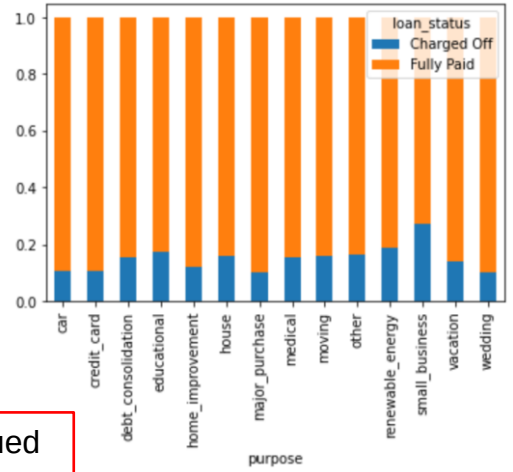
Lending Club: Sample Data Set for Analysis

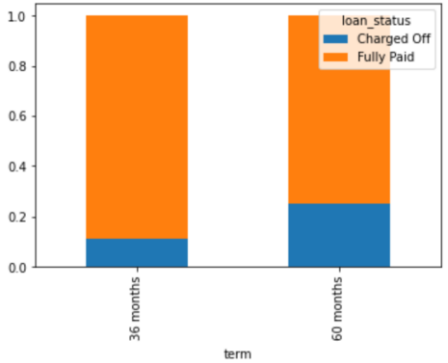
upGrad

Step 3: Bivariate Analysis between Loan Staus and Categorical variables









- Chances of full payment is more with loan issued to applicant's of longer employee length
- Chances of defaulting is high for loan applicant's whose home ownership is of type "Other"
- Loan taken for Business purpose is having higher chances of defaulting
- Chances of defaulting increases as number of term increases

Lending Club: Preliminary Recommendation

upGrad

Data Analysis based recommendation based on sample populaton

- Applicant applying for lower interest rate scheme - chances of defaulting less
- Applicant without defined ownership defaulting - hence risk is high such applicants
- Chances of Applicant issued with longer term period of loan is high compared to lower term period - hence should invest in loan with less term period
- Higher the loan amount greater the risk of defaulting for applicant
- Better chances of full repayment is loan amount between 5000 to 14000. Hence should investing in this range less risky
- Investing in loan applied for business is more risky than loan applied for any other purpose

upGrad

#LifeKoKaroLift

Thank You!