Group Facilitator: Pankaj Suryavanssi

Team Member: Harish Kumar

Lending Club Case Study

SUBMISSION

Lending Club Problem Statement and Requirements

Problem Statement:

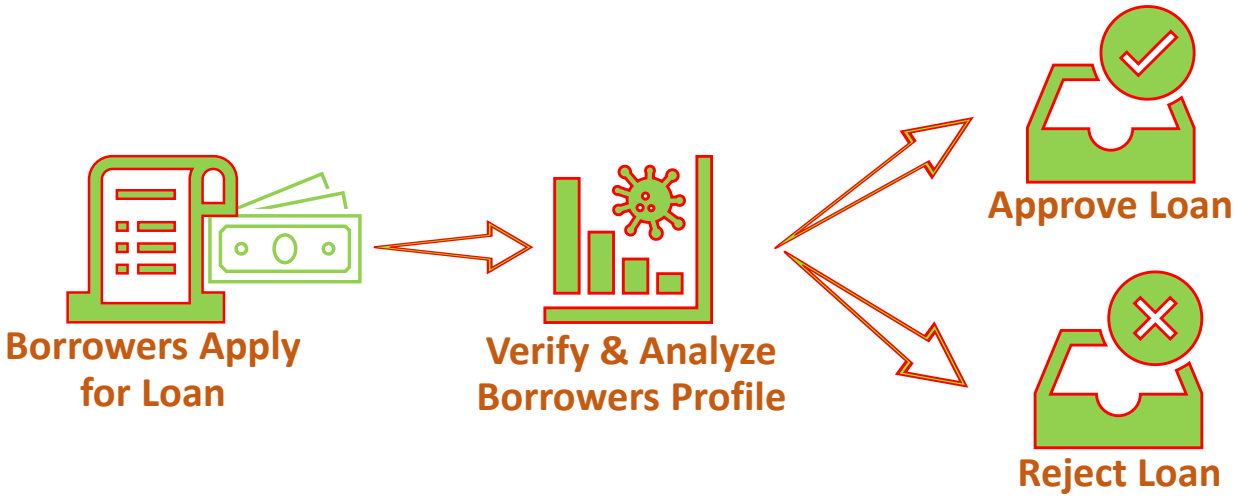
When Lending Club receives a loan application, the company has to make a decision for loan approval based on the applicant's profile.

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

Requirements:

- Lending club wants to understand the driving factors behind loan default.
- Lending club wants to understand driver variables which are strong indicators of default to utilize this analysis for its portfolio and risk assessment Funds



Approach to solve the problem

Understand the sample loan data provided for period 2007 to 2011

Clean the loan data with below steps

1.Check the percentage of missing values

2.Remove all those with very high missing percentage

3.For columns with less missing percentage: perform Imputations

4.Drop rows where the missing percentage is quite high

Perform Data Analysis on below variables

1.Variable related to the applicant (demographic variables such as age, occupation, employment details etc.),

2.Loan characteristics (amount of loan, interest rate, purpose of loan etc.) and

3.Customer behavior variables (those which are generated after the loan is approved such as delinquent 2 years, revolving balance, next payment date etc.).

4.Since customer behavior variables are not available at the time of loan application, and thus they cannot be used as predictors for credit approval.

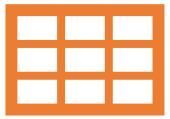
5.Remove data with loan status as 'current' and tag the Charged off as 0 and Fully Paid as 1.

Recommendations

1.Mention the finding of Univariate, Segmented Univariate and Multivariate Data analysis

2.Draw conclusion based on above analysis and provide recommendation to Lending club on key factors identified for loan default

Data Understanding and Data Cleaning



Data Set Details:

1.There are 39717 Rows and 111 Columns

2.Out of 111 Columns, 74 are float, 13 are int, and 24 are object

3.There are 54 Columns of which data is completely Null, which will require data cleaning

Unique Fields:

1.Id column is unique

2.member_id is unique

Field with null, NA and redundant values which will requirement data cleaning

1.Application_type - INDIVIDUAL

2.Collections_12_mths_ex_med - This column got 0 and NA value - 100%

3.Chargeoff_within_12_mths - Missing or 0 Value Percentage 100%

4.tax_liens - Missing or O value Percentage 100%

5.mths_since_last_record - Missing Value Percentage 93%

6.mths_since_last_delinq - Missing Value Percentage 65%

7.initial_list_status - f

8.pymnt_plan -n

9.id - Since id and member_id columns are having unique values, we can maintail only member_id column and drop the id column

10.next_pymnt_d - 98% Data is null

11.url, desc, title, zip_code, addr_state - We can't derive any co-relation from these columns

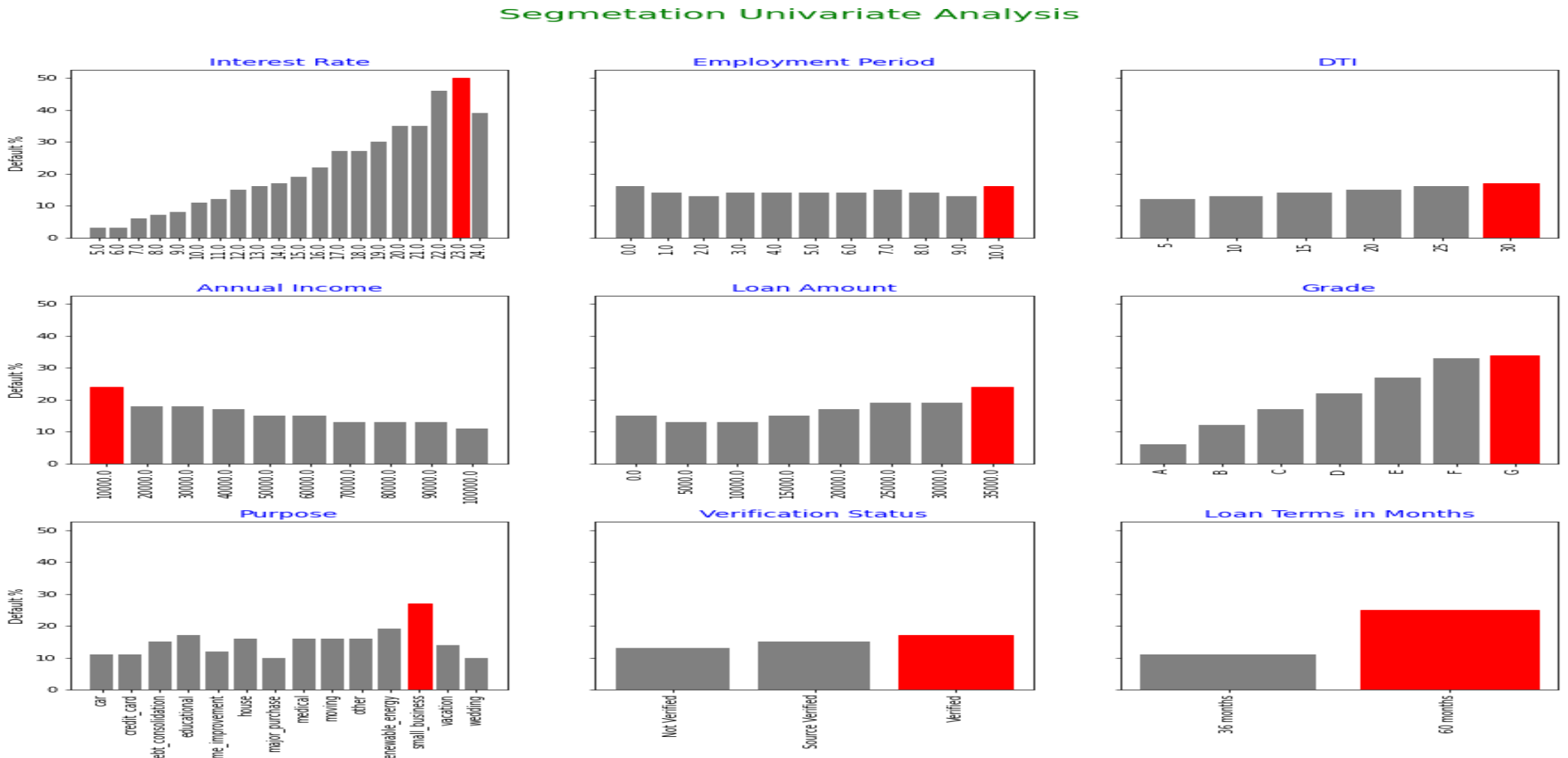
Univariate data Analysis

Univariate analysis: Below analysis has been performed on Loan data after cleaning

Variable	Top1	Top2	Top3	Remarks
int_rate	22	21	20	The Default Probability Ranges between **35% to 49%**, if the applicant given loan at 19% and above. Assumption: Considering 23% & 24% as outliers
grade	G	F	E	The Default Probability is **upto 34%** for Applicants with lower Grade(lower credit history)
purpose	small_business	other	medical	The Default Probability is **27%** if the applicatant taking loan for small business. The debt_consolidation category also very close to medical (15%+)
loan_amnt_5k-derived from Loan amount	35000	25000	30000	The Default Probability is **upto 25%** if the given loan amount is above 20,000.
terms	60 months	36 months		The Default Probability is **upto 25%** for Applicants with loan term is 60 months
annual_inc_10k-derived from annual income	10000	20000	30000	The Default Probability is **upto 24%** if the applicatant Annual income below 40,000. Assumption: Income has been segmented into nearest 10K, and outliers has been removed
Dtiq-Derived from DTI	30	25	20	The Default Probability **gradually increases from 12 to 16%** as DTI percetange increases.
emp_length	10	0	7	**No specific pattern** found with the Employement Period of applicant
verification_status	Verified	Source Verified	Not Verified	**No specific pattern** found with the Employement Period of applicant.

Univariate Data Analysis

- Univariate analysis: Below are the graphical representation of Univariate analysis.



Univariate data Analysis-Top 3 Findings

The probability of default is **upto 49%** for the loans with **Higher Interest Rate(19% to 23%)**

The probability of default is **upto 34%** for the application with **Lower Credit History/Grade(E,F,G)**

The probability of default is **upto 37%** for the loans taken for the purpose of **Small Business**

Assumption: Grade are considered in decreasing order, A is considered as higher grade(customer with highest credit history) and Grade G is considered as lowest grade(customer with worst credit history)

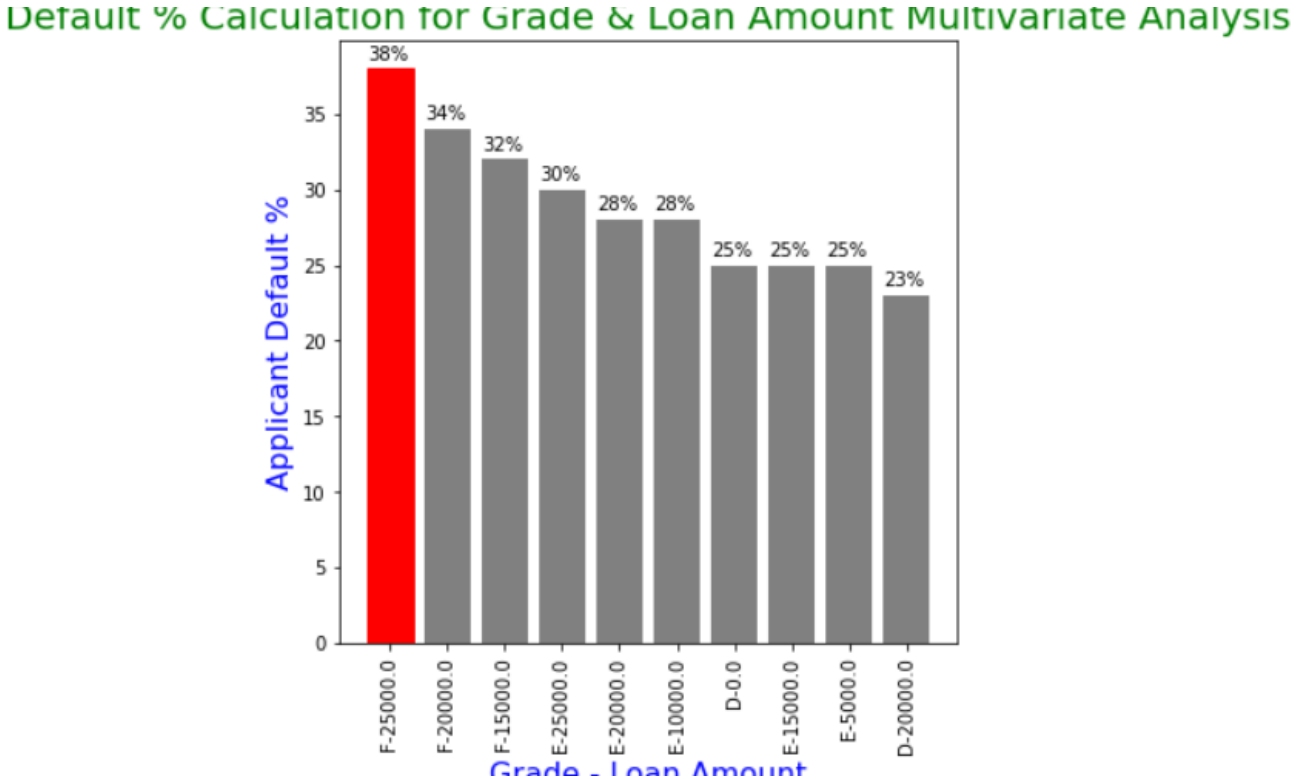
Multivariate data Analysis

'Default % Calculation for Grade & Loan Amount

Grade	Loan amount	# Charged Off	# Fully Paid	% of Default
F	25000	69	111	38
F	20000	62	121	34
F	15000	52	112	32
E	25000	92	219	30
E	20000	113	287	28
E	10000	157	413	28
D	0	66	203	25
E	15000	126	377	25
E	5000	118	363	25
D	20000	117	395	23

Note: Multivariate analysis has been performed where number of Charged off applicants are more than 50 for given combination of variables in all 3 multivariate data analysis

Multivariate data Analysis(Graphical representation and Conclusion)



- **Conclusion:** Up to **38% Applicants with lower Grade(lower credit history),** and if given an higher loan amount having high probability of getting defaulted
- **Assumption:** Grade are considered in decreasing order, A is considered as higher grade(customer with highest credit history) and Grade G is considered as lowest grade(customer with worst credit history)

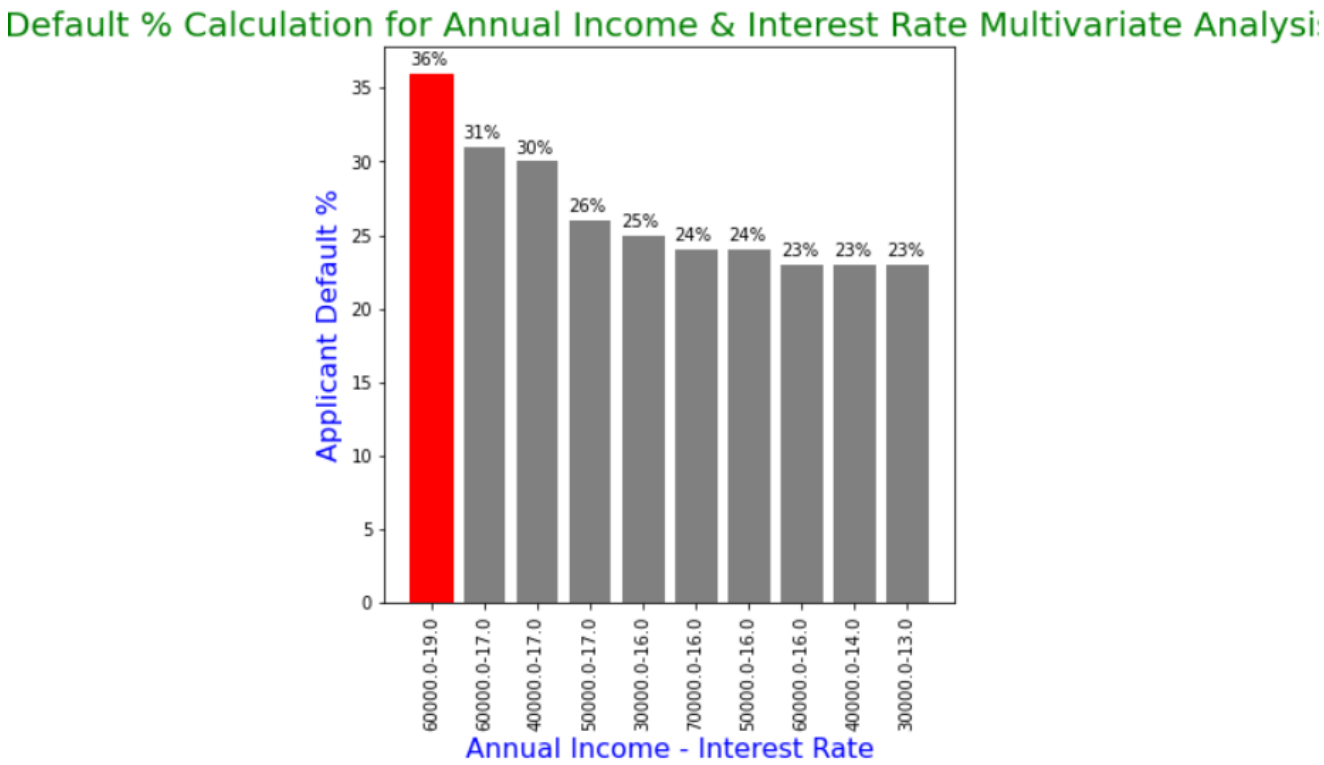
Multivariate data Analysis

Default % Calculation for
Annual Income & Interest Rate

Annual_Income	int_rate	# Charged Off	# Fully Paid	% of Default
60000	19	53	93	36
60000	17	67	147	31
40000	17	61	141	30
50000	17	51	143	26
30000	16	56	172	25
70000	16	54	169	24
50000	16	94	298	24
60000	16	81	275	23
40000	14	103	348	23
30000	13	102	343	23

Note: Applicants with annual income between 10000 and 100000 are analyzed to avoid outliers interfering to analysis.

Multivariate data Analysis(Graphical representation and Conclusion)



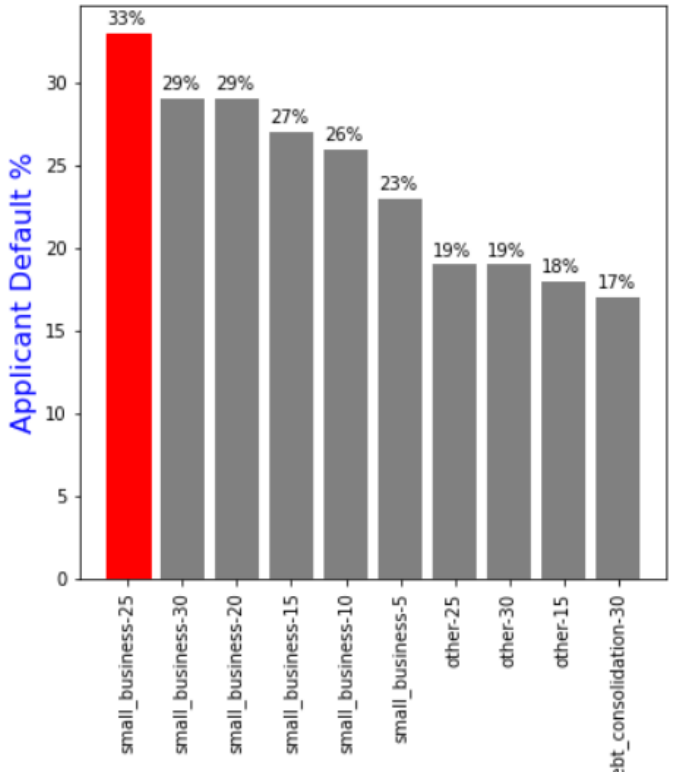
- **Conclusion:** Up to **36% applicants with mid-tier income, and if given loan at above 15% interest rate**, having high probability of getting defaulted.

Multivariate data Analysis

Default % Calculation for Purpose & Debt To Income

Purpose	DTIQ	# Charged Off	# Fully Paid	% of Default
small_business	25	74	148	33
small_business	30	58	140	29
small_business	20	63	154	29
small_business	15	75	200	27
small_business	10	98	277	26
small_business	5	107	360	23
other	25	108	471	19
other	30	108	454	19
other	15	101	469	18
debt_consolidation	30	618	2925	17

Multivariate data Analysis(Graphical representation and Conclusion)



- **Conclusion:** Up to **33% applicants with higher DTI, and if loan taken for the purpose of Small Business**, having high probability of getting defaulted

Recommendations to Investors

- Applicant with Lower Credit history/ Grade(E,F,G), should not be funded for higher loan amount(more than 25K).
- Applicant with Mid Tier annual income(30K to 60K) should be funded with reduced loan amount.
- Investors should be cautious while investing loan amount on Applicant with Higher Debt to Income ratio (more than 15%) with purpose of Small Business.
- Applicants charged with interest rate more than 19% are riskier candidate so Investors should be watchful while investing on such applicants.