LENDING CLUB CASE STUDY

GOKUL NARAYANAN
SAURABH PUROHIT

OVERVIEW

Lending Club is a consumer finance company which specializes in lending various types of loans to urban customers. 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

BUSINESS OBJECTIVE

- As Data Scientists for Lending Club our goal is to analyze the data obtained from the past loan applicants to help the company understand the patterns which indicate if a person is likely to Default for a Loan.

- Following Data driven analysis is performed for the same

  1) Understand the provided data set

  2) Clean the data set

  3) Data Analysis (Univariate, Segmented Univariate , Bivariate etc.)

  4) Report on factors resulting to default after the analysis

UNDERSTANDING THE DATA

Sample Data	Applicants who were granted loan between 2008-2011
Shape	Rows -39717 , Columns -111
Major Columns	loan_amnt, funded_amount_inv ,term,int_rate ,installment ,grade ,emp_length, home_ownership ,annual_inc,loan_status
Abbreviations	Explanation on each abbrevations used in data set is provided in Data_Dictionary.xlsx
Key Observations	1) Many columns have NaN values 2) Many columns have single values 3) Some columns have long string descriptions

DATA CLEANING – STAGE1

- Out of the total 111 columns , total of 54 columns were removed as they were having Null values . The left-out columns after removing null columns were 57
- Out of the remaining 57 columns some of the columns really don't matter for Loan approval stage. Total of 31 columns were removed resulting into remaining 26 columns

  eg: Columns removed are as follows

"id","member_id","emp_title","url","title","zip_code","addr_state","pymnt_plan","desc","delinq_2yrs“,“last_credit_pull_d","collections_12_mths_ex_med","policy_code","application_type","acc_now_delinq","chargeoff_within_12_mths","delinq_amnt","pub_rec_bankruptcies","tax_liens","initial_list_status“,"revol_bal","out_prncp","total_pymnt","total_rec_prncp","total_rec_int","total_rec_late_fee","recoveries","collection_recovery_fee","last_pymnt_d","last_pymnt_amnt","next_pymnt_d“

- mths_since_last_delinq and mths_since_last_record had 25682 and 36931 null values respectively and hence those were removed
- total_pymnt_inv and out_prncp_inv were also removed as they don't significantly contribute to loan approval. **So, this results to a total columns of 22 left out for analysis**

DATA CLEANING – STAGE2

- Null sum is taken on the remaining 22 columns and it is found that emp_length and revol_util have 1075 and 50 null values respectively
- Emp_length null values are filled with the mode values of emp_length and revol_util 50 rows are removed from the analysis

```
#Chekcing full columns to see if any of them having any null value
loan_data.isnull().sum()
```

```
loan_amnt               0
funded_amnt             0
funded_amnt_inv         0
term                    0
int_rate                0
installment             0
grade                   0
sub_grade               0
emp_length              0
home_ownership          0
annual_inc              0
verification_status     0
issue_d                 0
loan_status             0
purpose                 0
dti                     0
earliest_cr_line        0
inq_last_6mths          0
open_acc                0
pub_rec                 0
revol_util              0
total_acc               0
dtype: int64
```

DATA STANDARDIZATION

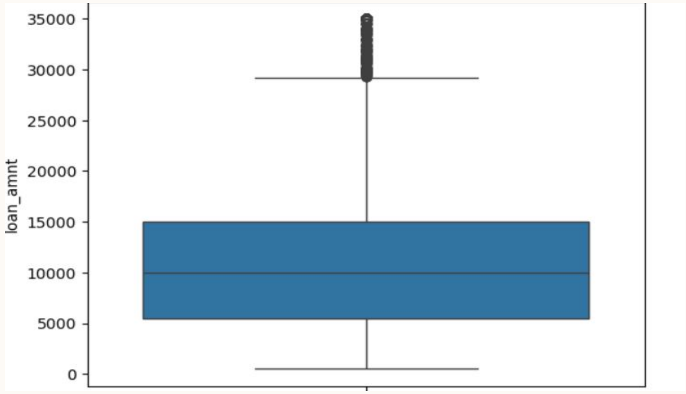
- % is removed from int_rate and revol_util columns and this was converted to float for analysis
- changed emp_length to numeric value by extracting the integer part

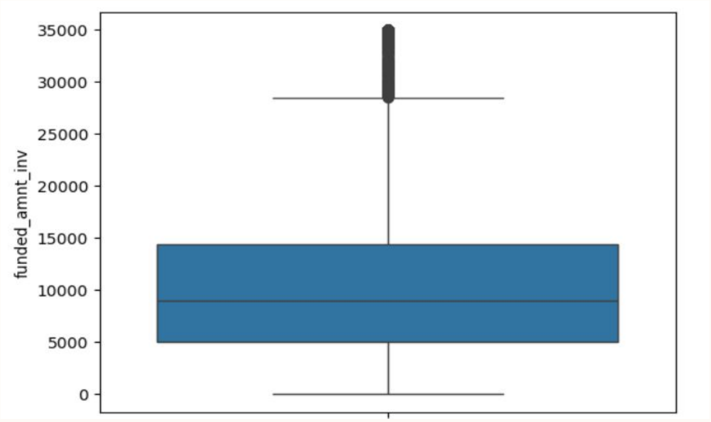
```
[386]: #checking the dataTypes after conversion
       loan_data.info()

<class 'pandas.core.frame.DataFrame'>
Index: 39667 entries, 0 to 39716
Data columns (total 22 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   loan_amnt             39667 non-null  int64
 1   funded_amnt           39667 non-null  int64
 2   funded_amnt_inv       39667 non-null  float64
 3   term                  39667 non-null  object
 4   int_rate              39667 non-null  float64
 5   installment           39667 non-null  float64
 6   grade                 39667 non-null  object
 7   sub_grade             39667 non-null  object
 8   emp_length            39667 non-null  int64
 9   home_ownership        39667 non-null  object
 10  annual_inc            39667 non-null  float64
 11  verification_status   39667 non-null  object
 12  issue_d               39667 non-null  object
 13  loan_status           39667 non-null  object
 14  purpose               39667 non-null  object
 15  dti                   39667 non-null  float64
 16  earliest_cr_line      39667 non-null  object
 17  inq_last_6mths        39667 non-null  int64
 18  open_acc              39667 non-null  int64
 19  pub_rec               39667 non-null  int64
 20  revol_util            39667 non-null  float64
 21  total_acc             39667 non-null  int64
dtypes: float64(6), int64(7), object(9)
memory usage: 7.0+ MB
```

OUTLIER DETECTION -1

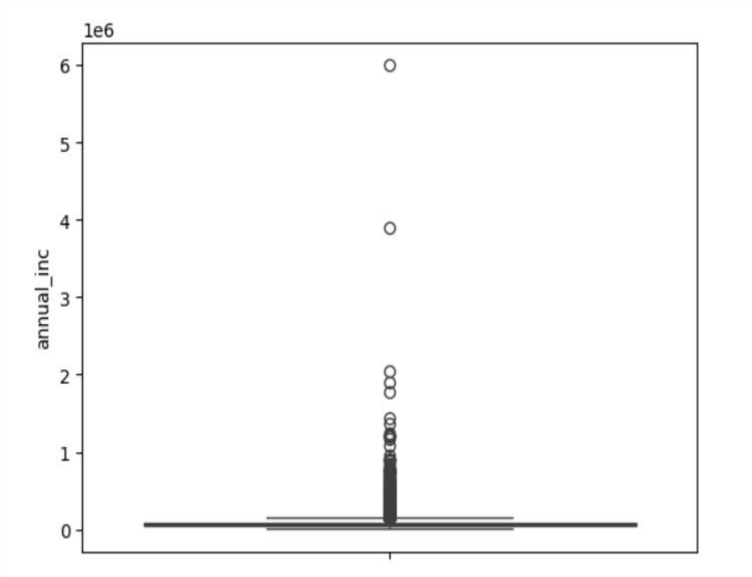
- Performed outlier detection on loan_amnt and funded_amnt_inv and it was observed that distribution was consistent and hence no outliers were removed

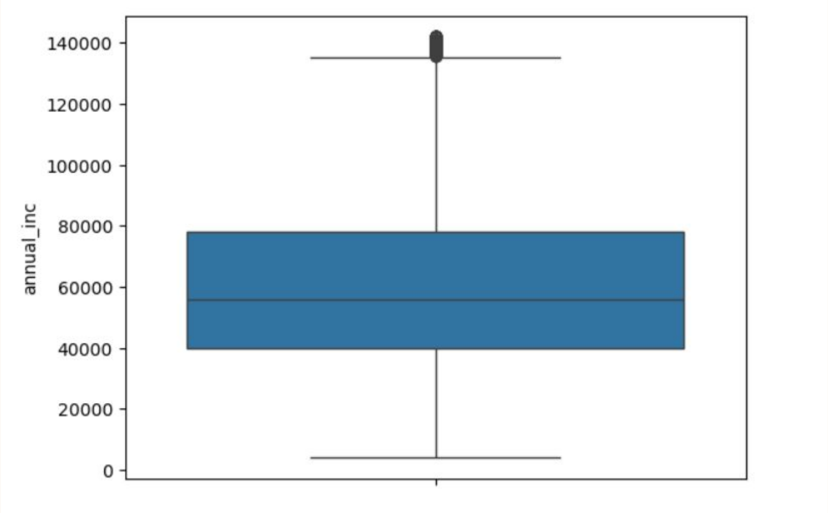




OUTLIER DETECTION -2

- Cleaned annual_inc column which had some outliers(fig1). 95 percentile was taken as a benchmark and data above 95 percent was removed (fig2)





UNIVARIATE ANALYSIS

- Since the Defaulting is not applicable for current loan_status . Univerate analysis was started by removing applicants with loan_status = “current” . i.e, the applicants with loan_status Fully Paid and Charged Off is only considered for analysis
- Initially a count plot was plotted to see how many Fully Paid and Charged Off applicants are present. This indicated around 30K of Fully paid and around 5K of Charged off users were present in data
- Below counterplots were done to analysis the pattern for Charged Off applicants against different columns

  1) grade -LC assigned Loan grade.

  2) home_ownership -The home ownership status provided by the borrower during loan request

  3) purpose – Category provided by the borrower for the loan request

  4) term – Number of payments on Loan – value is either 36 or 60

  5) emp_length - Employment length in years

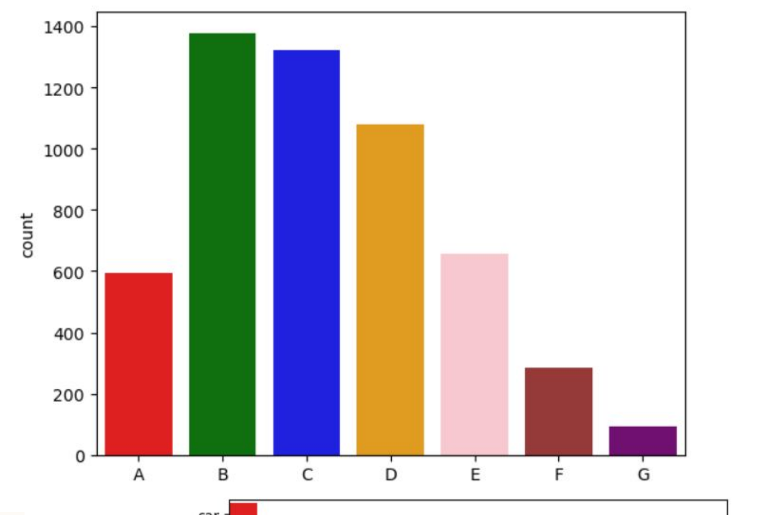
  6)verification_status – Income source was verified

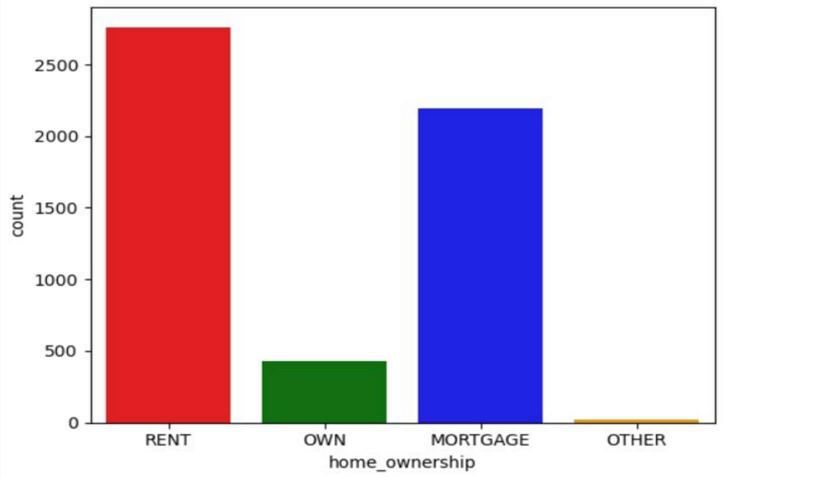
  7) inq_last_6_moths – Number of inquiries in last 6 months

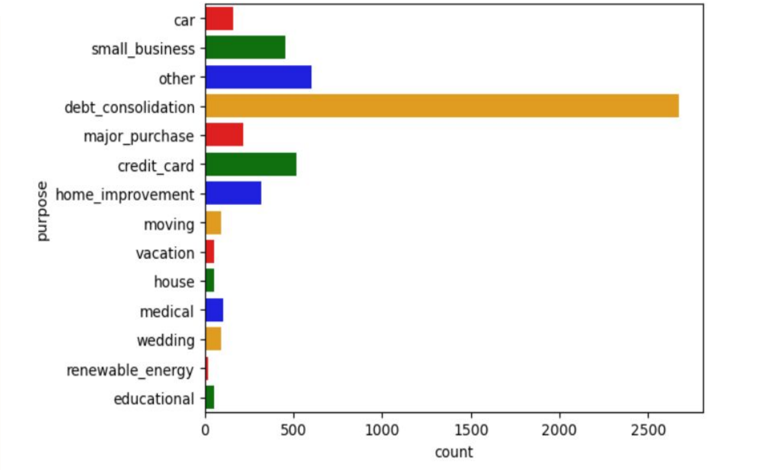
  8) pub_rec – Number of Derogatory public records

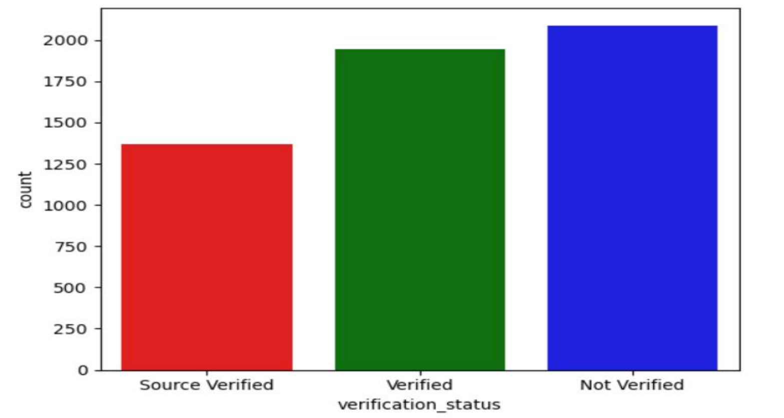
Note * Some of the plots for the univariate analysis is given in next slide

PLOTS –UNIVARIATE ANALYSIS









SEGMENTED ANALYSIS

- Segmented analysis for the Charged off Applicants were mainly performed on below attributes by plotting the counter plots

1) int_rate – This attribute was divided into 5 bins ranging from 5% to 25%

2) loan_amnt – This attribute was divided into 9 bins ranging from 0k -36K

3) funded_amnt_inv – This attribute was divided into 9 bins ranging from 0k-36K

4) installment - The installment attribute was divided into 10 bins ranging from 15-1500 for analysis

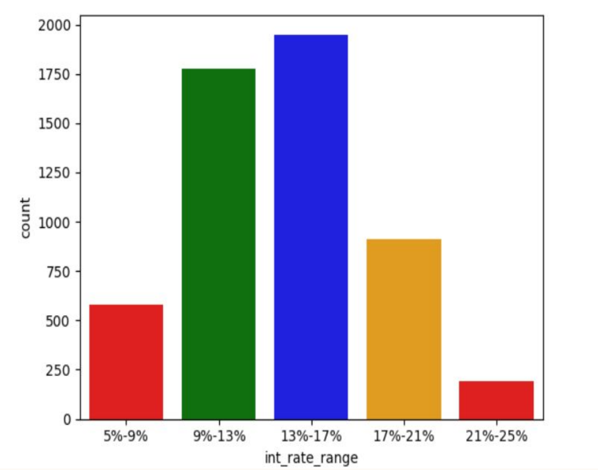
5) Annual income – Divided into 9 bins starting from 4K to 148K

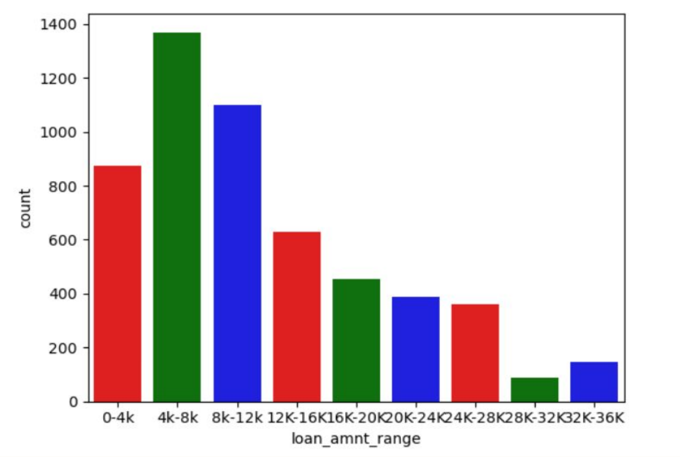
6) DTI range – dti range was plotted ranging from 0-30

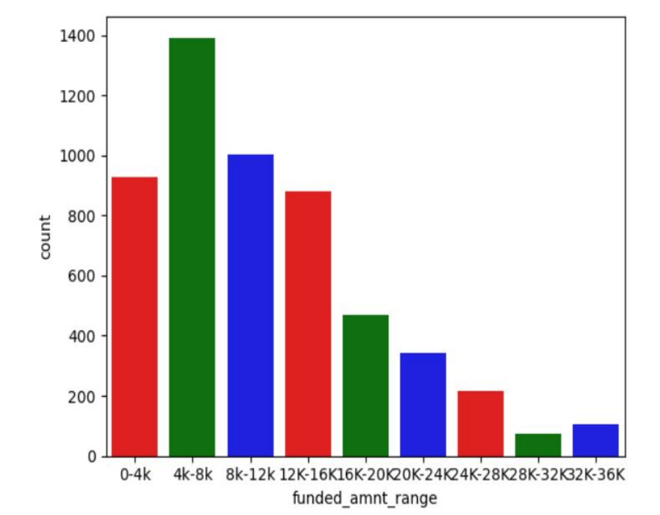
7) open_acc_range – open accound range was plotted for a range of 0-45

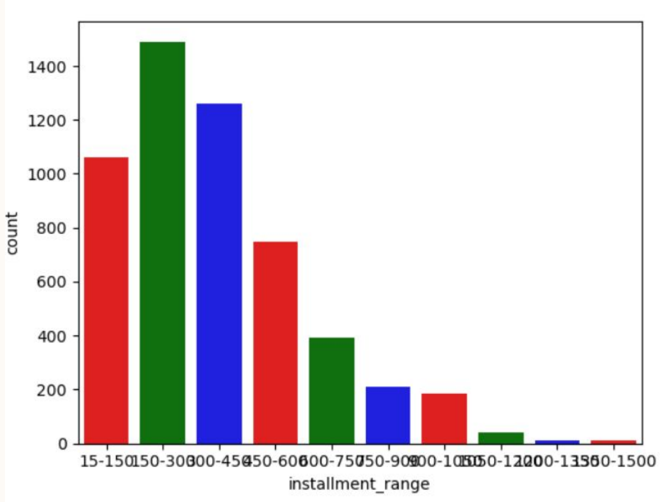
Note* - Some of the plots for segmented analysis is given in next slide

PLOTS –SEGMENTED ANALYSIS









OUTCOME FROM UNIVARIATE AND SEGMENTED ANALYSIS

For the Charged off loans the more chance of Applicant being default is as follows

1) Applicants with grade B have a high chance of Loan Default
2) Applicants with RENT as home ownership have the high chance of Default
3) Applicants with debt consolidation , ie using a new loan to close the existing loans have a high chance of being defaulted
4) Applicants with interest range between 13%-17% have a high chance of being Defaulted
5) Applicants with term 36 months have high chance of being defaulted
6) Applicants with employee length 10 have the high chance of being defaulted
7) Not Verified Applicants have a high chance of becoming Defaulted
8) Loan amount between 4K-8K have the high range of being defaulted
9) Funded_amount_inv in the range of 4K-8K have the high chance of being defaulted
10) Installment in the range if 150-300 have the high chance of being defaulted
11) dti range of 12-15 have the high chance of being defaulted
12) Applicants having a annual_income of range 36k-52K is having high chance of being defaulted
13) Those applicants who have zero inquiry in last 6 months have high chance of being defaulted
14) Applicants having open accounts in the range of 5-10 have the high chance of being defaulted
15) Applicants having Derogatory public records zero have a high chance of being defaulted

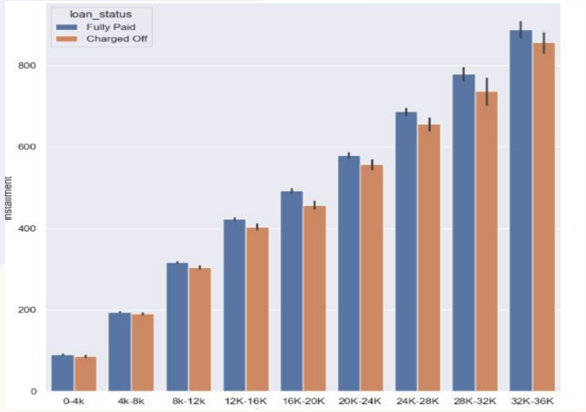
BIVARIATE ANALYSIS

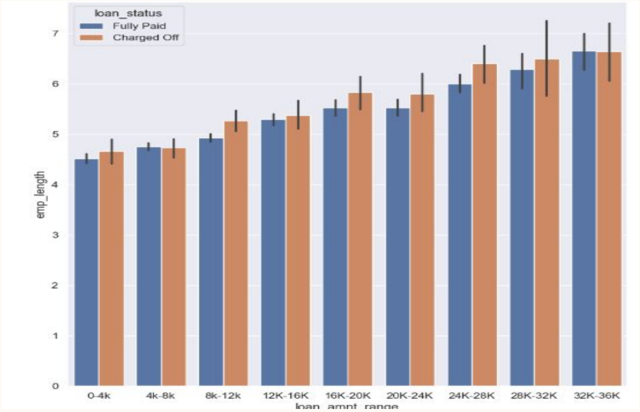
- Bivariate analysis for the Charged off Applicants were mainly performed on below attributes by plotting the bar plots

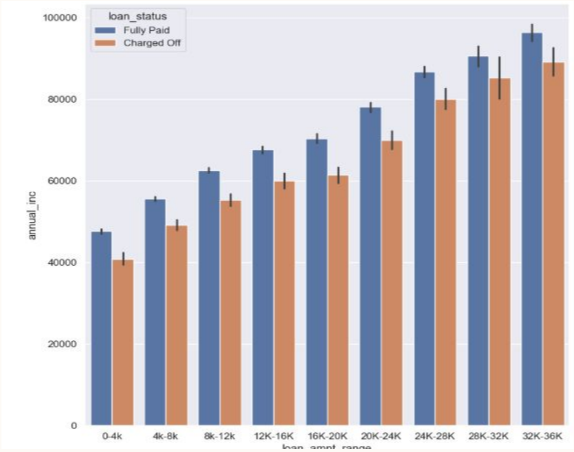
1) loan_amnt - This attribute was divided into 9 bins ranging from 0k -36K
2) term - Number of payments on Loan – value is either 36 or 60
3) int_rate - This attribute was divided into 5 bins ranging from 5% to 25%
4) installment - The installment attribute was divided into 10 bins ranging from 15-1500 for analysis
5) grade - LC assigned Loan grade
6) emp_length - Employment length in years
7) home_ownership - The home ownership status provided by the borrower during loan request
8) annual_inc- Divided into 9 bins starting from 4K to 148K
9) purpose - Category provided by the borrower for the loan request

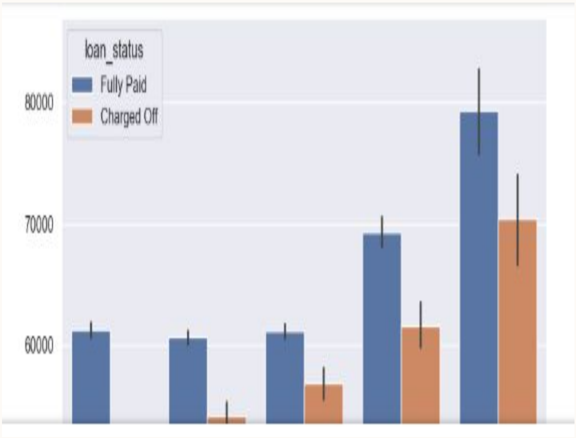
Note* - Some of the plots for Bivariate analysis is given in next slide

PLOTS –BIVARIATE ANALYSIS









OUTCOME FROM BIVARIATE ANALYSIS

For the Charged off loans the more chance of Applicant being default is as follows

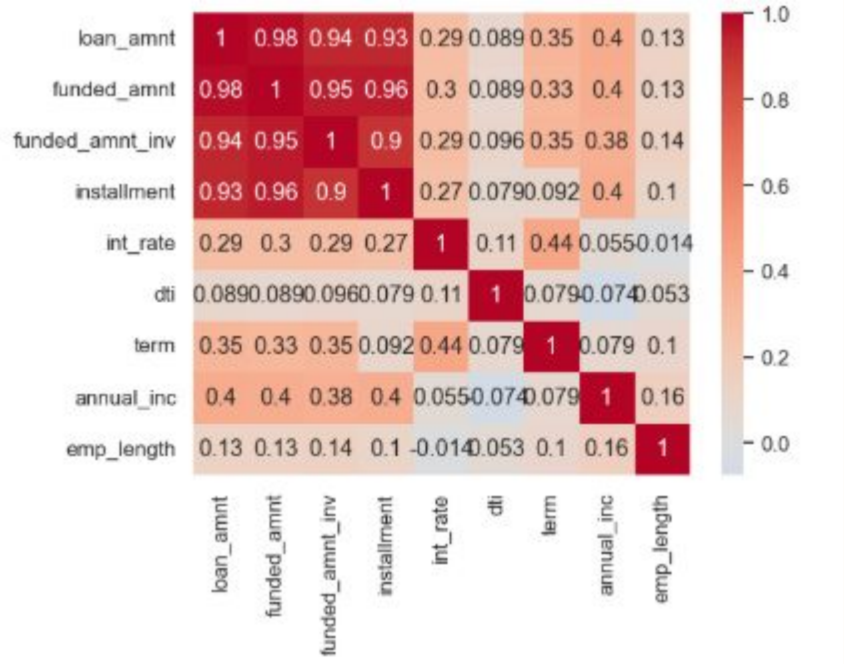
1) Applicants having income in the range of 60K-80K and home is mortgaged
2) Applicants having income in the range of 84K-100K and employee length between 5 and 6
3) Applicants having income in the range of 70K-80K and grade G and F
4) Applicants having income in the range of 116K-132K and who has installments greater than 500
5) Applicants having income in the range of 70K-80K and interest rates in range of 21% -25%
6) Applicants having loan amount in the range of 12k-14k and taken a loan for small business
7) Applicants having loan amount in the range of 32k-36k and annual income in range of 70k-80k
8) Applicants having loan amount in the range of 12k-14k and who are not owning the home
9) Applicants having loan amount in the range of 32k-36k and Interest rates in the range of 15-17.5
10) Applicants having loan amount in the range of 15k-17.5k and Grade is F

CORRELATION ANALYSIS

- Correlation analysis for the Charged off Applicants were mainly performed on below attributes by plotting the heat map
    1) loan_amnt
    2) funded_amnt
    3) int_rate
    4) funded_amnt_inv
    5) installment
    6) dti
    7) term
    8) annual_inc
    9) emp_length

Note* - Plots for Correlation analysis is given in next slide

PLOTS –CORRELATION ANALYSIS



OUTCOME FROM CORRELATION ANALYSIS

Insights from Correlation Metrics

Strong Correlation

- installment is strongly correlated with funded_amnt, loan_amnt, and funded_amnt_inv.
- term shows a strong correlation with the interest rate.
- annual_inc is strongly correlated with loan_amount.

Weak Correlation

- dti has a weak correlation with most fields.
- emp_length also shows a weak correlation with most fields.

Negative Correlation

- annual_inc has a negative correlation with dti.

THANK YOU

Gokul Narayanan

Saurabh Purohit