Lending Club Case Study

- Suma Santosh

Aim of the study

In this case study, Using EDA to understand how consumer attributes and loan attributes influence the tendency of default.

The aim is to identify patterns which indicate if a person is likely to default for given data set

Flow of the Analysis

1. Data sourcing
2. Data cleaning
3. Univariate analysis
4. Bivariate analysis
5. Derived metrics
6. Conclusion

Data Sourcing

Loading data from given data set

```
#Import required libraries

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

```
#load data set for inspection
lending_data_records = pd.read_csv('loan.csv')
```

```
lending_data_records.shape
```

```
(39717, 111)
```

Data Sourcing (Contd.)

Columns for the analysis and inspection of columns for null value

```
['id', 'member_id', 'loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'term', 'int_rate', 'installment', 'grade', 'sub_grade', 'emp_title',
'emp_length', 'home_ownership', 'annual_inc', 'verification_status', 'issue_d', 'loan_status', 'pymnt_plan', 'url', 'desc', 'purpose', 'title',
'zip_code', 'addr_state', 'dti', 'delinq_2yrs', 'earliest_cr_line', 'inq_last_6mths', 'mths_since_last_delinq', 'mths_since_last_record',
'open_acc', 'pub_rec', 'revol_bal', 'revol_util', 'total_acc', 'initial_list_status', 'out_prncp', 'out_prncp_inv', 'total_pymnt',
'total_pymnt_inv', 'total_rec_prncp', 'total_rec_int', 'total_rec_late_fee', 'recoveries', 'collection_recovery_fee', 'last_pymnt_d',
'last_pymnt_amnt', 'next_pymnt_d', 'last_credit_pull_d', 'collections_12_mths_ex_med', 'mths_since_last_major_derog', 'policy_code',
'application_type', 'annual_inc_joint', 'dti_joint', 'verification_status_joint', 'acc_now_delinq', 'tot_coll_amt', 'tot_cur_bal',
'open_acc_6m', 'open_il_6m', 'open_il_12m', 'open_il_24m', 'mths_since_rcnt_il', 'total_bal_il', 'il_util', 'open_rv_12m', 'open_rv_24m',
'max_bal_bc', 'all_util', 'total_rev_hi_lim', 'inq_fi', 'total_cu_tl', 'inq_last_12m', 'acc_open_past_24mths', 'avg_cur_bal', 'bc_open_to_buy',
'bc_util', 'chargeoff_within_12_mths', 'delinq_amnt', 'mo_sin_old_il_acct', 'mo_sin_old_rev_tl_op', 'mo_sin_rcnt_rev_tl_op', 'mo_sin_rcnt_tl',
'mort_acc', 'mths_since_recent_bc', 'mths_since_recent_bc_dlq', 'mths_since_recent_inq', 'mths_since_recent_revol_delinq',
'num_accts_ever_120_pd', 'num_actv_bc_tl', 'num_actv_rev_tl', 'num_bc_sats', 'num_bc_tl', 'num_il_tl', 'num_op_rev_tl', 'num_rev_accts',
'num_rev_tl_bal_gt_0', 'num_sats', 'num_tl_120dpd_2m', 'num_tl_30dpd', 'num_tl_90g_dpd_24m', 'num_tl_op_past_12m', 'pct_tl_nvr_dlq',
'percent_bc_gt_75', 'pub_rec_bankruptcies', 'tax_liens', 'tot_hi_cred_lim', 'total_bal_ex_mort', 'total_bc_limit', 'total_il_high_credit_limit']
```

Data Cleaning

1. Fixing Column and Rows
    - Checking for Null Columns/Rows and deleting them
    - Checking for unique value column and deleting them
    - Identifying columns which are not helpful for analysis and delete them
2. Fix missing values
    - Deleting Significant percentage missing value Columns/Rows
3. Standardise values
    - Format the columns values to help in analysis
4. Fix Invalid values
    - Checking data types and setting it right for the column
5. Filter out duplicate rows

Data Cleaning(Contd.)

Output : Clean Rows and columns without null/invalid/useless columns

After keeping only delinquency history present records

```
lending_data.shape
```

```
(11795, 22)
```

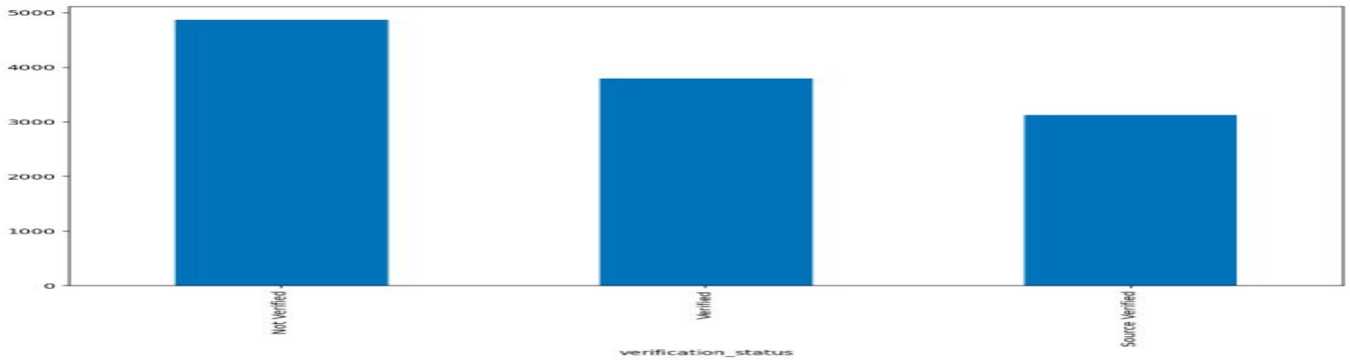
Univariate Analysis

Analysing one driver variable at a time, to find out influencing factors for risky loan/default customer

Variable : dti

dti less than 36 percent is considered good , median of dti value is in valid range, this variable is not influencing factor

Variable : verification_status



Univariate Analysis(Contd.)

Variable : annual_inc, loan_status

Gave some insights to take it to next stage analysis, based on distribution

Variable : revol_util

Below data shows revolving line utilization rate above 49 percent may be one of the indicator of default tendency

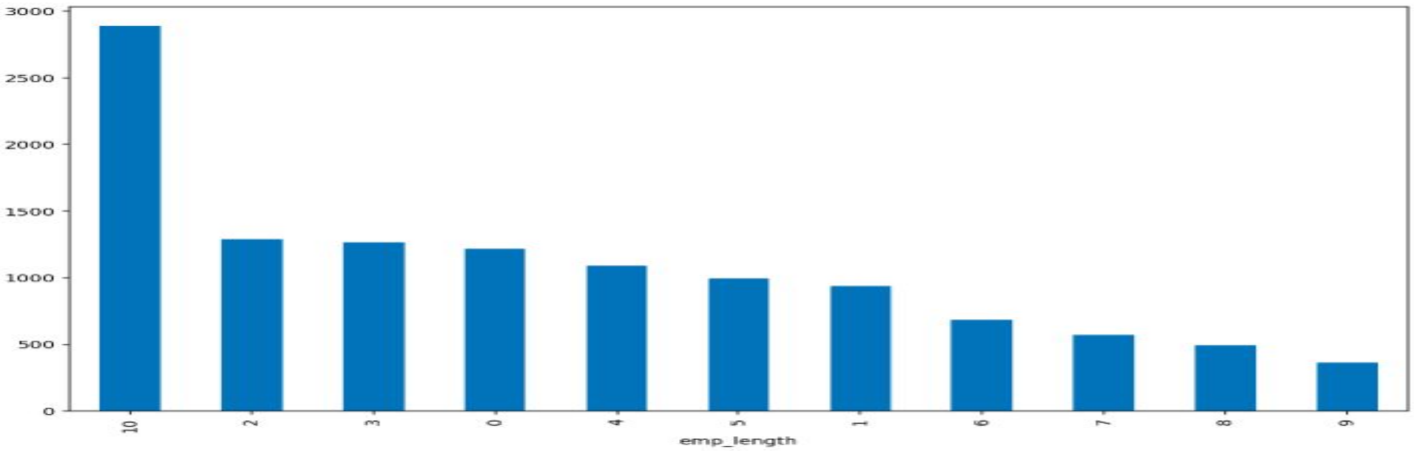
```
lending_data['revol_util'].describe()
```

```
count    11795.000000
mean        49.227798
std         27.718508
min          0.000000
25%         27.100000
50%         49.600000
75%         72.000000
max         99.900000
Name: revol_util, dtype: float64
```

Univariate Analysis(Contd.)

Variable : emp_length

With higher employee length, delinquency cases have increased



Univariate Analysis(Contd.)

Segmented Univariate Analysis (refer tables in next slide with respective to below number)

1. Analysis of revolving credit utilisation percent based on loan status

   From data we can derive that high revolving credit utilisation can influence loan defaulting

2. Analysis of employee length based on loan status

   Employee length is not giving significant insights w.r.to delinquency

3. Analysis of number of months since last delinquency based on verification status

   Data clearly show around 30 percent people who are not verified, have contributed to delinquency

4. Analysis of number of months since last delinquency based on loan status

   From data , number of Delinquency is high, those customer who tend to be charged off

5. Based on Grade

   Grade A has high delinquency rate

6. Based on Home ownership, segmented variable analysis

   Based on data e we can derive that home_ownership with MORTGAGE and RENT have high tendency towards delinquency

Univariate Analysis(Contd.)

1

	revol_util							
	count	mean	std	min	25%	50%	75%	max
loan_status								
Charged Off	1748.0	52.837208	27.898078	0.0	32.300	54.20	76.20	99.9
Current	316.0	51.863956	26.574725	0.0	30.500	53.55	72.25	99.1
Fully Paid	9731.0	48.493827	27.669224	0.0	26.365	48.80	71.30	99.9

2

	emp_length							
	count	mean	std	min	25%	50%	75%	max
loan_status								
Charged Off	1748.0	5.184211	3.615831	0.0	2.0	5.0	10.0	10.0
Current	316.0	6.155063	3.550722	0.0	3.0	6.0	10.0	10.0
Fully Paid	9731.0	5.121879	3.528237	0.0	2.0	5.0	9.0	10.0

3

	mths_since_last_delinq							
	count	mean	std	min	25%	50%	75%	max
verification_status								
Not Verified	4868.0	36.551767	18.912503	6.0	21.0	35.0	50.0	75.0
Source Verified	3129.0	36.346437	18.712819	6.0	21.0	35.0	50.0	75.0
Verified	3798.0	36.321485	18.798521	6.0	21.0	35.0	50.0	75.0

4

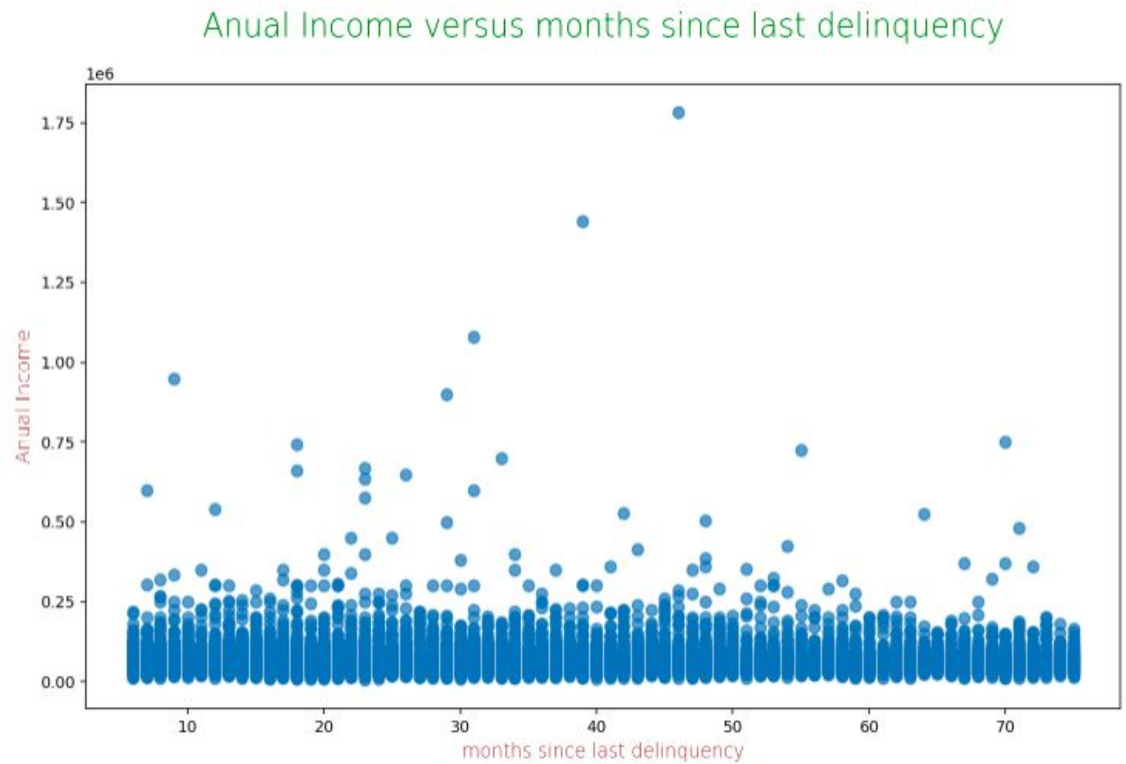
	mths_since_last_delinq							
	count	mean	std	min	25%	50%	75%	max
loan_status								
Charged Off	1748.0	36.593822	19.456078	6.0	20.0	35.0	51.00	75.0
Current	316.0	35.591772	18.913197	6.0	19.0	34.0	47.25	75.0
Fully Paid	9731.0	36.419484	18.703770	6.0	21.0	35.0	50.00	75.0

Univariate Analysis(Contd.)

5	mths_since_last_delinq							
	count	mean	std	min	25%	50%	75%	max
grade								
A	1452.0	40.457989	18.924247	6.0	25.75	39.0	56.0	75.0
B	3294.0	37.269885	18.842250	6.0	22.00	36.0	52.0	75.0
C	2968.0	36.303235	18.798356	6.0	21.00	35.0	50.0	75.0
D	2220.0	34.901351	18.594356	6.0	19.00	33.0	48.0	75.0
E	1224.0	34.508170	18.320262	6.0	19.00	33.0	47.0	75.0
F	486.0	32.462963	18.355978	6.0	17.00	30.0	45.0	75.0
G	151.0	32.152318	18.923970	6.0	15.50	29.0	46.5	74.0

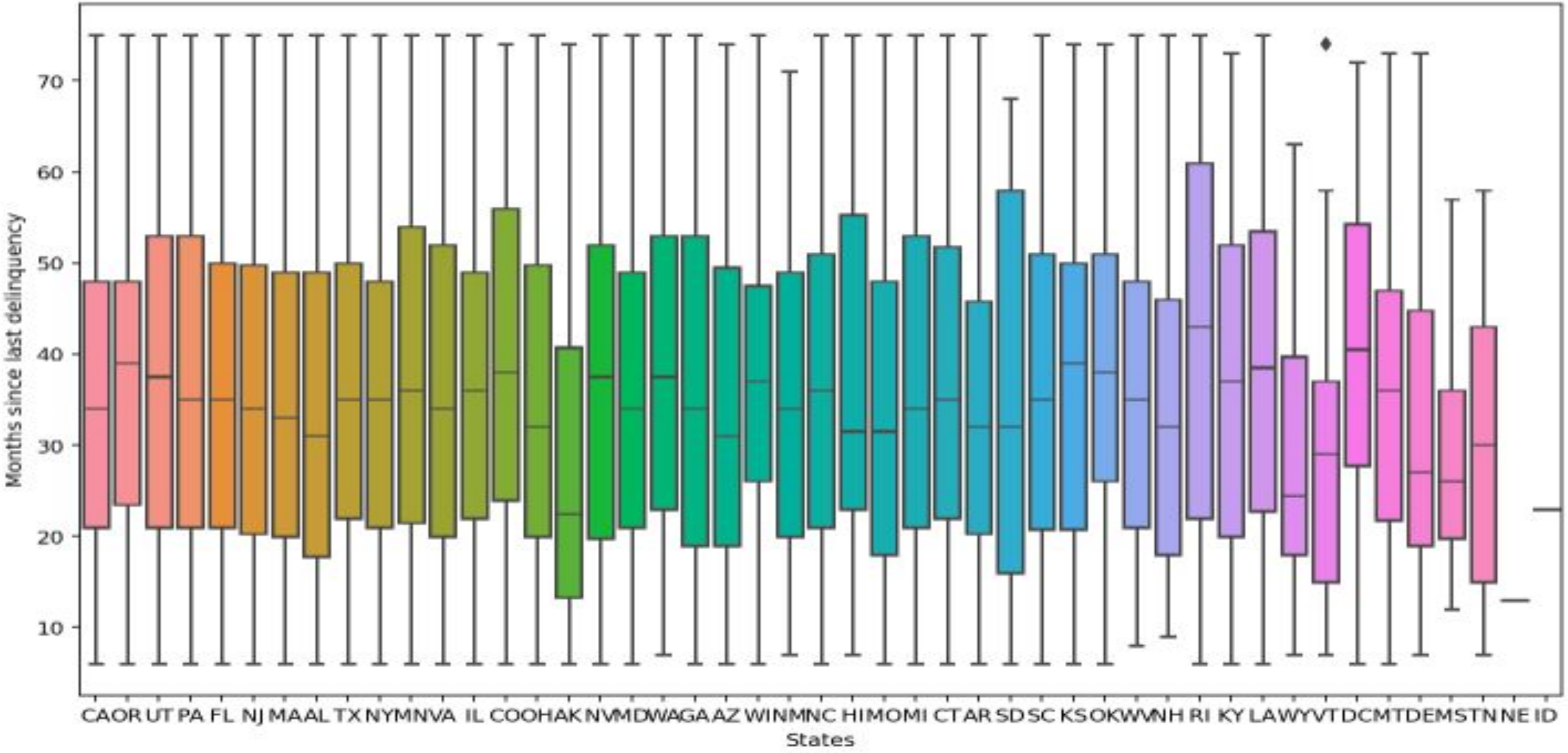
6	mths_since_last_delinq							
	count	mean	std	min	25%	50%	75%	max
home_ownership								
MORTGAGE	5330.0	36.588368	19.316015	6.0	20.00	35.0	51.0	75.0
OTHER	32.0	28.812500	20.648264	6.0	12.75	22.5	43.0	70.0
OWN	868.0	35.665899	19.316137	6.0	20.00	33.0	49.0	75.0
RENT	5565.0	36.426774	18.237457	6.0	21.00	35.0	49.0	75.0

Bivariate Analysis



Bivariate Continuous Variable Analysis

Bivariate Analysis(Contd.)



Bivariate Variable Analysis

Bivariate Analysis(Contd.)

Bivariate categorical variable analysis

Analysing Grade variable against months since last delinquency and annual income

Inference : From below table Grade A with lower income have more tendency towards delinquency

	mths_since_last_delinq								annual_inc							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
grade																
A	1452.0	40.457989	18.924247	6.0	25.75	39.0	56.0	75.0	1452.0	69310.246198	58499.576660	8500.0	43925.00	60000.0	82000.0	1440000.0
B	3294.0	37.269885	18.842250	6.0	22.00	36.0	52.0	75.0	3294.0	68428.345024	48079.419682	9600.0	42000.00	60000.0	82000.0	948000.0
C	2968.0	36.303235	18.798356	6.0	21.00	35.0	50.0	75.0	2968.0	67134.148181	52386.054930	9600.0	41985.25	57600.0	80000.0	1782000.0
D	2220.0	34.901351	18.594356	6.0	19.00	33.0	48.0	75.0	2220.0	71238.464248	48216.305833	6000.0	43000.00	60000.0	85000.0	648000.0
E	1224.0	34.508170	18.320262	6.0	19.00	33.0	47.0	75.0	1224.0	80122.899828	57318.184527	13920.0	48000.00	65000.0	95109.0	750000.0
F	486.0	32.462963	18.355978	6.0	17.00	30.0	45.0	75.0	486.0	88783.980185	54860.764227	15600.0	57000.00	76900.0	105000.0	600000.0
G	151.0	32.152318	18.923970	6.0	15.50	29.0	46.5	74.0	151.0	95923.966358	73969.351995	24000.0	60000.00	80000.0	112500.0	725000.0

Derived Metrics

To derive new data from existing data , pivot table is created to get insights analysis based on important variables

```
loan_data_subset =lending_data.pivot_table(values=['delinq_2yrs','pub_rec_bankruptcies','pub_rec','revol_util'],index=
['loan_status'],aggfunc='mean')
print(loan_data_subset)
```

```
             delinq_2yrs   pub_rec  pub_rec_bankruptcies  revol_util
loan_status
Charged Off     0.433638  0.101259              0.068078   52.837208
Current         0.427215  0.053797              0.041139   51.863956
Fully Paid      0.393793  0.064639              0.042544   48.493827
```

Conclusion

From above all analysis we can infer that borrower who has

- more number of derogatory public record
- bankruptcies
- history of default instances in last 2 year
- Is on Home Mortgage or Rent
- Revolving credit line utilization rate above 49 percent

will tend to default and they are more riskier loan applicants

Thank You