Loan Lending Case Study

RAVI SHEKHAR | RAHUL ALKARI

EPGP_AIML_C53_MAY23

Problem Statement

Company XYZ is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

- Company wants to identify ‘Risky’ applicants who probably turn into defaulter

- Identify key driving factors which are strong identifier of loan defaulter based on data provided

Analysis Approach

We use Exploratory Data Analysis methodologies to identify variable or group of variables which are strong indicator of default rate based on available data

Analysis is done step wise as mentioned below :

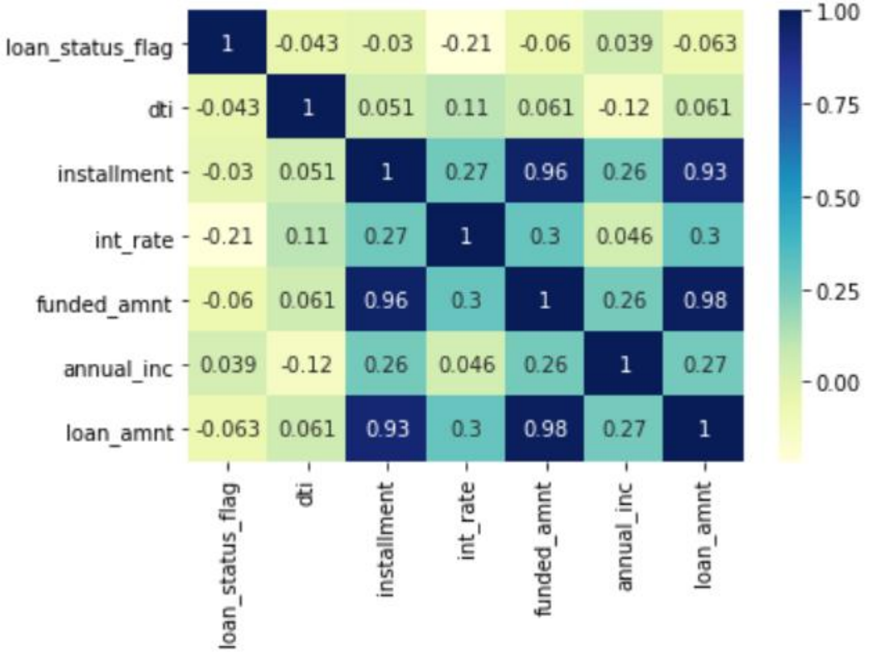
1. Select only relevant columns for analysis having enough data values to support analysis
2. Clean data wherever possible by filtering out columns, rows, removing nulls and imputing null values
3. Check default average Default Rate % based in cleaned data
4. Do univariate analysis to get mean, max , standard deviation and distribution of variable.
5. Based on higher or lower percentile numbers of relevant variables (can be selected based on business judgement) filter out the dataset and check Defaulter rate, if it increases or goes down
6. Apply bivariate analysis to further identify the driving factors
7. Repeat steps 1 to 6 based on observations till key variable and their appropriate values are identified.

Data Cleanup

1. Removed columns with null values
2. Removed unnecessary columns based on calculated judgment based on unique values of each column
3. Removed unnecessary rows , For ex Rows having all null values. , Loan = current
4. Imputed null values in desc title columns with ‘Not Available’ text. Numerical column is imputed with mean
5. Added binned columns for Loan_amnt and interest rate
6. Added derived columns from loan_status as Loan_status_flag
7. Derived final dataframe with not a single null values in any column. Dataframe : 37747 Rows X 43 columns

Check Correlation Matrix

Created a Correlation matrix for Loan Status Flag to check which variables show either positive or negative correlation.



Few variables like int_rate shows relatively strong negative correlation. So it can be analyzed along with other variables

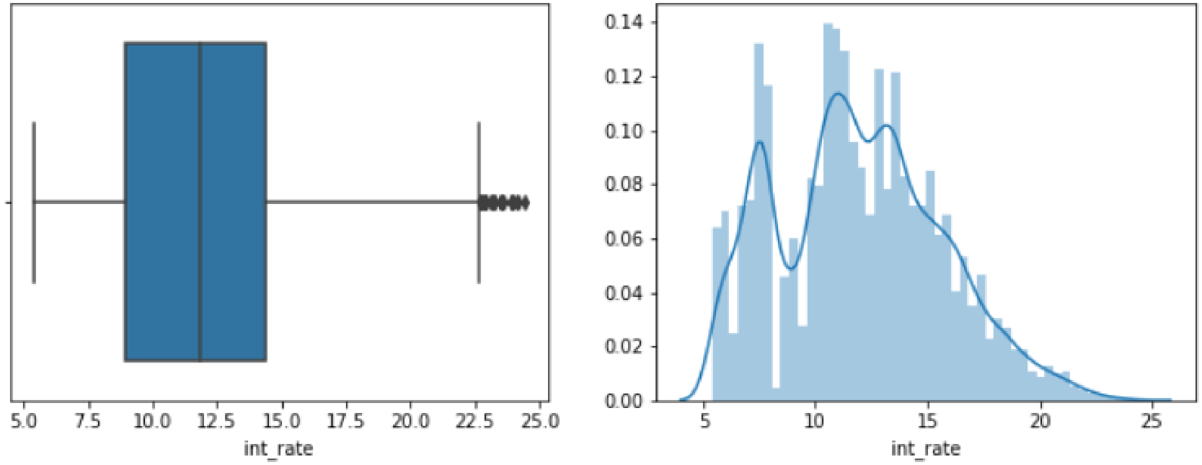
Univariate Analysis on Interest Rate

It gives us mean value of around 12 , while 90 percentile value is 16.5% after removing Outliers.

Values above 22.5 are outliers. Further analysis is done by removing outliers.

```
print(loan_redord_anyls4['int_rate'].describe(percentiles=[.25,.50,.75,.80,.90,.95]))
## 95 % data is below
#plt.boxplot(loan_redord_anyls4['loan_amnt'])
plt.figure(figsize=(18,4))
plt.subplot(1,2,1)
sns.boxplot(loan_redord_anyls4['int_rate'])
plt.subplot(1,2,2)
sns.distplot(loan_redord_anyls4['int_rate'])
plt.show()
```

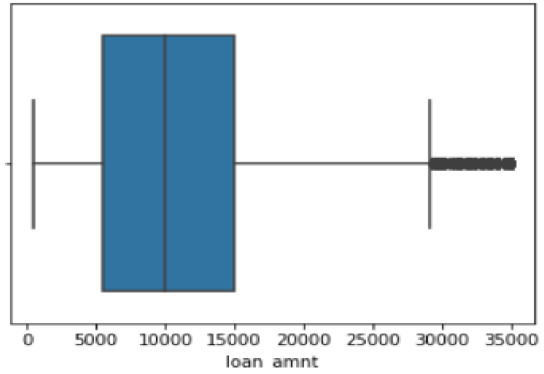
```
count    37542.000000
mean        11.963470
std          3.683023
min          5.420000
25%          8.940000
50%         11.830000
75%         14.420000
80%         15.230000
90%         16.820000
95%         18.390000
max         24.400000
Name: int_rate, dtype: float64
```

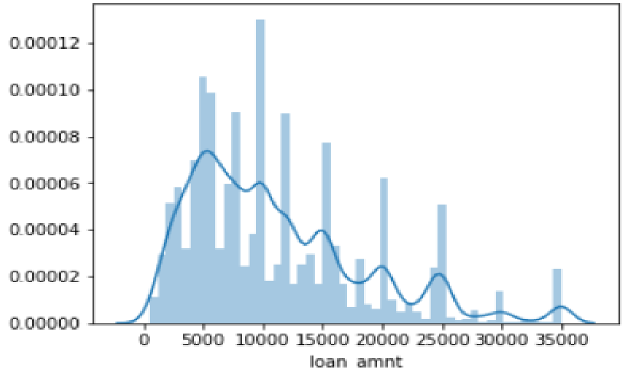


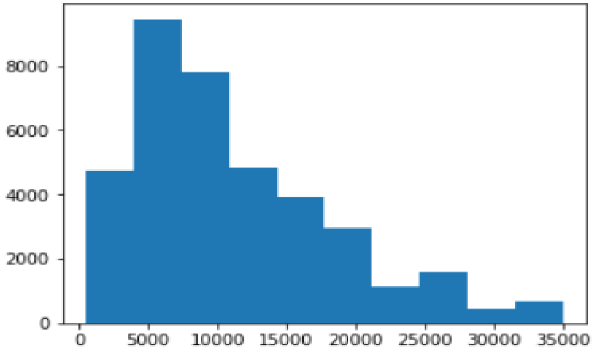
Observations : Most customers receive loan at rates 9% to 15%

Univariate Analysis on further variables

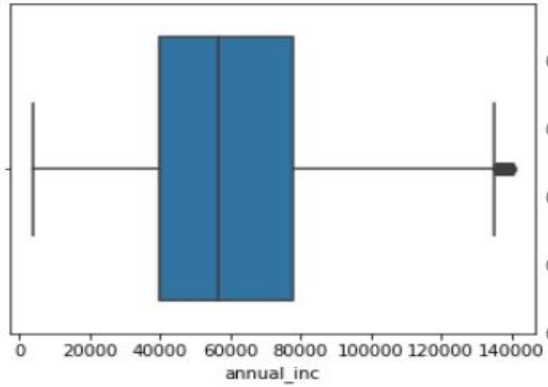
Further box plots are created for loan_amnt , annual income and installments to understand mean , max ,mean and outliers values.

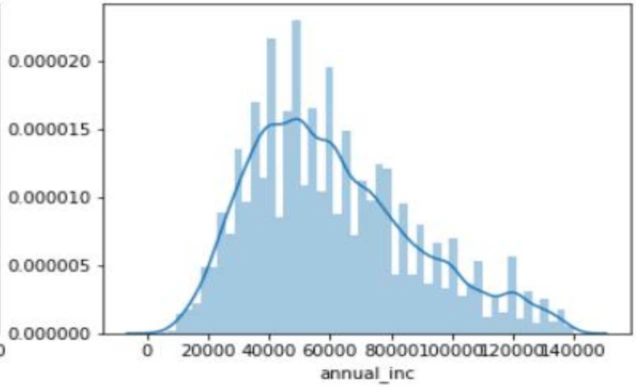


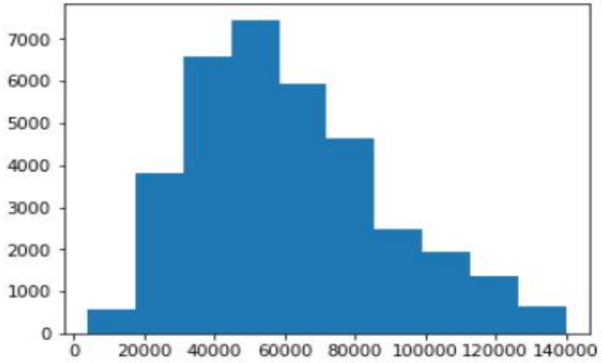




Observations : Most customers received loans between 5000 to 20000 USD







Observations : Annual income of most customers is between 30000 to 90000USD

Univariate Analysis Scenario -1

Filter data based on mean , max , min , 90 percentile , 75 percentile values received as part of univariate analysis and se the impact o default rate %

```
crosstab = pd.crosstab(Annuac_inc['int_rate_bucket'], Annuac_inc['loan_status'],margins=True)
#print(crosstab)
crosstab['Default Rate']  = 100* crosstab['Charged Off']/ crosstab['All']

print(crosstab)
```

```
loan_status      Charged Off  Fully Paid    All  Default Rate
int_rate_bucket
10-15                   2545       14623  17168     14.824091
15-20                   1651        4931   6582     25.083561
20-25                    261         397    658     39.665653
5-10                     717       10341  11058      6.483993
All                     5174       30292  35466     14.588620
```

Observations : Loans given at interest rate less than Mean interest rate (12.5%) decreases defaulter rate to 9.34% (Average defaulter rate is 14.32%)

Bivariate Analysis Scenario -1

When interest rate is further analyzed with purpose of loan taken, it shows the safest loan types for interest less than or equal to mean interest rates.

```
check_int_rate = loan_redord_anyls4[loan_redord_anyls4['funded_amnt'] <35000 ]
check_int_rate = check_int_rate[check_int_rate['installment'] <500 ]
check_int_rate = check_int_rate[check_int_rate['int_rate'] <12.5 ]

crosstab = pd.crosstab(check_int_rate['purpose'], check_int_rate['loan_status'],margins=True)
#print(crosstab)
crosstab['Default Rate']  = 100* crosstab['Charged Off']/ crosstab['All']

print(crosstab)
```

loan_status purpose	Charged Off	Fully Paid	All	Default Rate
car	82	918	1000	8.200000
credit_card	179	2402	2581	6.935296
debt_consolidation	743	6957	7700	9.649351
educational	21	156	177	11.864407
home_improvement	121	1423	1544	7.836788
house	9	154	163	5.521472
major_purchase	84	1208	1292	6.501548
medical	38	336	374	10.160428
moving	46	274	320	14.375000
other	217	1803	2020	10.742574
renewable_energy	10	47	57	17.543860
small_business	125	531	656	19.054878
vacation	30	197	227	13.215859
wedding	31	442	473	6.553911
All	1736	16848	18584	9.341369

```
check_int_rate = loan_redord_anyls4[loan_redord_anyls4['funded_amnt'] <35000 ]
check_int_rate = check_int_rate[check_int_rate['installment'] <500 ]
check_int_rate = check_int_rate[check_int_rate['int_rate'] <12.5 ]

crosstab = pd.crosstab(check_int_rate['emp_length'], check_int_rate['loan_status'],margins=True)
#print(crosstab)
crosstab['Default Rate']  = 100* crosstab['Charged Off']/ crosstab['All']

print(crosstab)
```

loan_status emp_length	Charged Off	Fully Paid	All	Default Rate
1	393	3512	3905	10.064020
2	190	1991	2181	8.711600
3	165	1841	2006	8.225324
4	150	1464	1614	9.293680
5	147	1439	1586	9.268600
6	94	978	1072	8.768657
7	86	741	827	10.399033
8	71	650	721	9.847434
9	50	561	611	8.183306
10	390	3671	4061	9.603546
All	1736	16848	18584	9.341369

Observations : House Loan , Home improvements , Wedding loan and educational loans are safest when loan at interest rate < mean interest rate (12.5) is provided to customers while small_business loan is riskiest

No significant impact observed due to employment length

Univariate Analysis Scenario -2

Filter data based on 90 percentile , values received as part of univariate analysis and see the impact of default rate %

```
check_int_rate = loan_redord_anyls4[loan_redord_anyls4['funded_amnt'] > 0 ]
check_int_rate = check_int_rate[check_int_rate['installment'] < 900 ]
check_int_rate = check_int_rate[check_int_rate['int_rate'] > 16 ]

crosstab = pd.crosstab(check_int_rate['purpose'], check_int_rate['loan_status'],margins=True)
#print(crosstab)
crosstab['Default Rate']  = 100* crosstab['Charged Off']/ crosstab['All']

print(crosstab.sort_values('Default Rate', ascending = False))
```

```
loan_status         Charged Off  Fully Paid   All  Default Rate
purpose
small_business              141         211   352     40.056818
educational                   5           8    13     38.461538
house                        22          40    62     35.483871
other                       133         281   414     32.125604
debt_consolidation          827        1983  2810     29.430605
All                        1443        3591  5034     28.665077
medical                      22          56    78     28.205128
car                          27          74   101     26.732673
vacation                      8          25    33     24.242424
wedding                      23          77   100     23.000000
major_purchase               40         134   174     22.988506
moving                       13          45    58     22.413793
credit_card                 118         414   532     22.180451
home_improvement             62         231   293     21.160410
renewable_energy              2          12    14     14.285714
```

Observations : Loans given for Vacations, small business, moving, medical at higher interest rate (>16%) increases defaulter rate% considerably to more than 35 % for few of these categories. (14 % is average)

Bivariate Analysis Scenario -2

When interest rate is further analyzed with purpose of loan taken with high interest rates, it shows the riskiest loan types.

```
check_int_rate = loan_redord_anyls4[loan_redord_anyls4['funded_amnt'] > 0 ]
check_int_rate = check_int_rate[check_int_rate['installment'] < 900 ]
check_int_rate = check_int_rate[check_int_rate['int_rate'] > 16 ]

crosstab = pd.crosstab(check_int_rate['emp_length'], check_int_rate['loan_status'],margins=True)
#print(crosstab)
crosstab['Default Rate']  = 100* crosstab['Charged Off']/ crosstab['All']

print(crosstab.sort_values('Default Rate', ascending = False))
```

```
loan_status  Charged Off  Fully Paid   All  Default Rate
emp_length
10                   385         873  1258     30.604134
8                     54         125   179     30.167598
3                    147         365   512     28.710938
All                 1443        3591  5034     28.665077
6                     89         222   311     28.617363
7                     72         181   253     28.458498
1                    256         644   900     28.444444
5                    128         330   458     27.947598
4                    127         337   464     27.370690
2                    142         389   531     26.741996
9                     43         125   168     25.595238
```

Observations : When analyzed with emp length no conclusion can be drawn on the pattern of loan defaulters based on employee length. Definitely Defaulter is high since int_rate is on hihger side.

END