Lending Club Case Study

Prepared by

Sudha Velpuri and Moka

Contents

Objective

The aim of this case study is to identify the risky loan applicants, so that loans sanctioned to these applicants can be reduced thereby cutting down the amount of credit loss. Identification of such risky applicants will be done using Exploratory Data Analysis.

Business Understanding

This is a Consumer Finance Company that lends various types of loans to urban customers.

Based on the applicant's profile company will decide to approve the loan application or not.

There are **two types of risks** associated with the bank's decision:

1. Loss of business to company if applicant is likely to repay loan and is not sanctioned the loan approval.

2. Approving the loan to an applicant who is not likely to repay loan may bring financial loss for the company.

Two types of decisions taken by the company:

1. **Loan accepted:**

If the company approves the loan, there are 3 possible scenarios described below:

• Fully paid: Applicant has fully paid the loan (the principal and the interest rate)

• Current: Applicant is in the process of paying the instalments, i.e., the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.

• Charged-off: Applicant has not paid the instalments in due time for a long period of time, i.e., he/she has defaulted on the loan.

2. **Loan rejected:**

If company did not approve the loan because of not meeting the requirements of the bank rules.

Executive Summary

   1. **Univariate Analysis**

• Juniors who have 1-4 years of experience are the highest loan seekers.

• A large proportion of applicants are applying loan for debt consolidation.

• A large portion of applicants have applied for loan grade B.

• A large portion of applicants have applied for B3-C1 loan subgrade followed by A1-B2

• Most of the loans are of 36 months tenure.

• Applicants who rent or mortgage home are the highest loan seekers.

2. Segmented Univariate Analysis

• Most of the charged off applicants have loan amount less than 15K.

• Applicants who are professionals (Above 10 years) have taken highest average loan amount.

• Charged off applicants report higher percentage of income going to debts (DTI) compared.

• Most of the applicants who took the loan for debt consolidation have paid fully

• Loan was paid by all applicants whose average annual income is high compared to charged off applicants

• Interest rate is higher, the chance for charged off is more.

• Most of the applicants with higher income have their home mortgaged.

• Applicants with highest average income for Loan grade F are being charged off for highest average loan amount.

3. Bivariate Analysis

• There is high correlation between annual income to loan amount (0.42), there 40 % chance that higher income individuals apply to higher loan amounts.

• Very low correlation between annual income to interest rate (0.068), there is positive but negligible influence on higher income individuals to apply for loan based on interest rate.

• Annual income to DTI has very minor but negative correlation (-0.11).

• Loan amount to DTI has positive but little positive correlation (0.071).

Conclusion

1. It is safe to provide loans to employees with higher income, low DTI value for 36 months tenure.

2. It is also safe to provide loans to applicants whose home is mortgaged, and purpose of the loan is debt consolidation.

Steps followed:

Step 1: **Data Understanding:**

Tried to understand the complete data set.

Step 2: **Data Cleaning:**

Removed all unnecessary columns that is All columns and rows that contain 'NaN' values are removed as these does not contribute to the analysis.

Step 3: **Univariate Analysis**:

Tried to analyze the data in individual variables to infer its value towards identifying potential

defaults.

Step 4: **Segmented Univariate Analysis:**

This was done by grouping categorical data with quantitative variables to make analysis

more insightful compared to Univariate analysis.

Step 5: **Bivariate/Multivariate Analysis:**

Here we used two or more variables to draw a better conclusion to understand correlation between them.

Step 6: **Recommendations/Results:**

Here, we made an effort to provide insightful recommendations to ensure that Lenders will

not incur loss by offering bad loans to the applicants who has no intent to payback.

Data Cleaning Steps:

• Delete columns: Deleted unwanted columns.

• Remove outliers: Removed high and low values that would disproportionately affect the results of our analysis.

• Missing values: Treat missing values with appropriate approach.

• Duplicate data: Remove identical rows, remove rows where some columns are identical.

• Filter rows: Filter by segment to get only the rows relevant to the analysis.
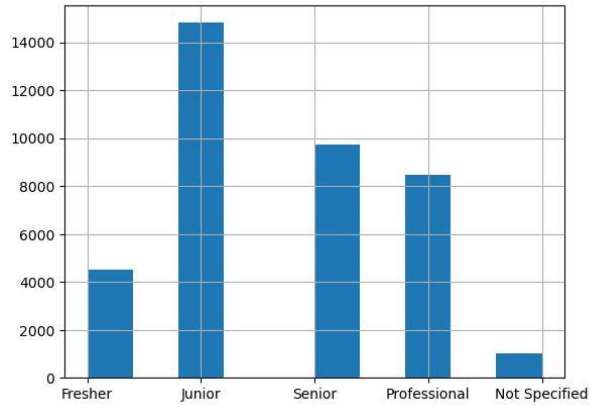
Analysis:

In this project we have analysed and understood how consumer and loan attributes are influencing the tendency of Charge off. Performed Bivariate analysis to understand how different variables interact with each other and how it's affecting the Charge off% and fully paid%. During this analysis used the different plots.

1. **Histograms / Bar charts**, used to check out the distribution of all the variables
2. **Scatter plots** to detect the outliers
3. **Box plots** to see the quantile percentages distribution across variables.
4. **Heat map** for correlation between quantitative variables

Univariate Analysis:

```
In [102]: #Binning on applicants employment duration/length

          df["emp_exp"].sort_values(ascending = True).hist()

          #Observation: Juniors who have 1-4 years of experience are the highest loan seekers.

Out[102]: <Axes: >
```
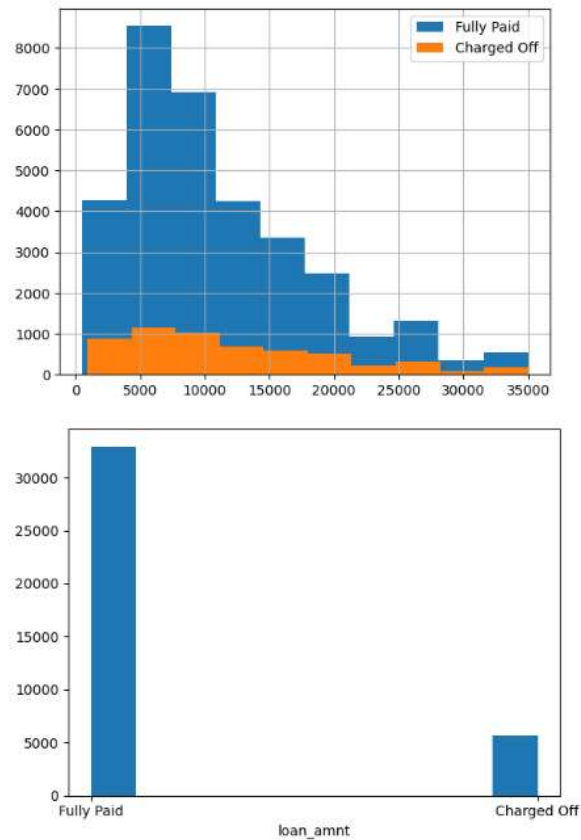

From the graph above, we can infer that _Juniors are the most loan seekers._

Segmented Univariate Analysis

```
In [110]: df[df["loan_status"] == "Fully Paid"]["loan_amnt"].hist()
          df[df["loan_status"] == "Charged Off"]["loan_amnt"].hist()
          plt.legend(["Fully Paid", "Charged Off"])
          plt.show()

          n, bins, patches = plt.hist(x = df['loan_status'])
          plt.xlabel("loan_amnt")
          plt.show()

          #Observation: Charged off applicants have loan amount below 10k
```


From the second graph above, we can infer that *Charged off applicants have loan amount below 10k*

```
In [111]: pd.crosstab(df['emp_exp'], df['loan_status'], margins = True)

          #Observation: Juniors are highest applicants for loan which has highest applicants being charged off and fully paid.
```

Out[111]:

loan_status	Charged Off	Fully Paid	All
emp_exp			
Fresher	639	3869	4508
Junior	2040	12774	14814
Senior	1389	8345	9734
Professional	1331	7157	8488
Not Specified	228	805	1033
All	5627	32950	38577

Juniors who have 1-4 years of experience have more applicants being charged off and fully paid.

```
In [114]:  #Data Dictionary: 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.

           def quantile25(x):
               return np.percentile(x, 25)

           def quantile50(x):
               return np.percentile(x, 75)

           def quantile75(x):
               return np.percentile(x, 75)

           print(df.pivot_table(values = "dti",
                           index = "loan_status",
                           fill_value = 0,
                           aggfunc=[np.mean, np.median, np.std, np.min, np.max, quantile25, quantile50, quantile75]).round(2))

           df.boxplot(column = ['dti'], by = 'loan_status')
           plt.show()

           #Observation : charged off applicants report higher percentage of income going to debts compared to fully paid
```
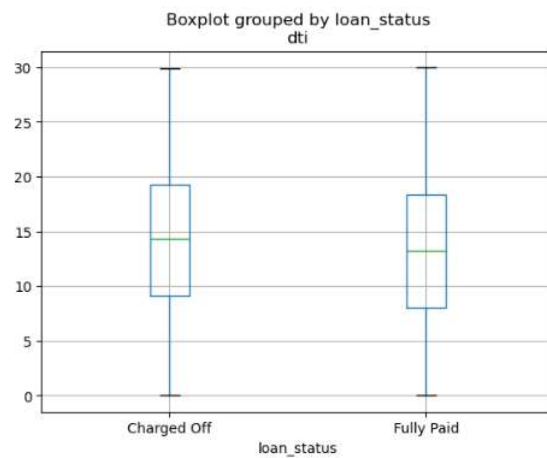
```
                   mean median   std amin   amax quantile25 quantile50 quantile75
                    dti    dti   dti  dti    dti        dti        dti        dti
loan_status
Charged Off  14.00  14.29  6.59    0  29.85       9.05      19.29      19.29
Fully Paid   13.15  13.20  6.68    0  29.99       7.98      18.39      18.39
```


Boxplot grouped by loan_status
dti

Charged off applicants report higher percentage of income going to debts compared to fully paid

```
In [119]:  df.pivot_table(values = ["annual_inc","loan_amnt"],
                           index = "grade",
                           columns = "loan_status",
                           fill_value = 0,
                           aggfunc= np.mean).round(0)

           #Observation: Applicants with highest average income for Loan grade F are being charged off for highest average loan amount.
```

Out[119]:

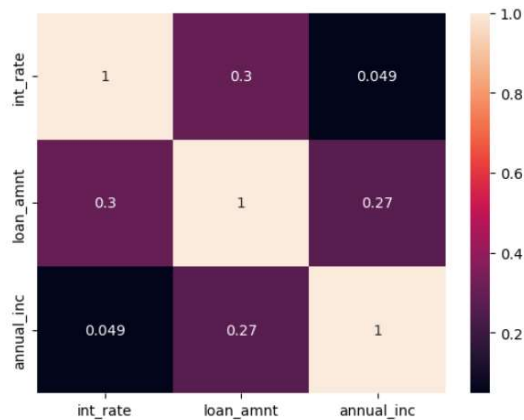| | annual_inc | | loan_amnt | |
loan_status	Charged Off	Fully Paid	Charged Off	Fully Paid
grade				
A	52510.0	67585.0	7800.0	8671.0
B	59638.0	68663.0	10906.0	10939.0
C	58642.0	69619.0	11044.0	10768.0
D	62356.0	69954.0	12205.0	12119.0
E	73261.0	79454.0	15850.0	15618.0
F	80347.0	85004.0	19273.0	17524.0
G	78869.0	102131.0	18543.0	21125.0

Loan grade F have highest average income being charged off for highest average loan amount.

Bivariate and Multi-Variate Analysis

Bivariate & MultiVariate Analysis

```
In [120]: df1 = df[['int_rate','loan_amnt','annual_inc']]
          df1.corr()
          corrMatrix = df1.corr()
          sns.heatmap(corrMatrix, annot=True)
          plt.show()

          #observation: There is relatively high correlation between annual income compared to loan amount (0.27)
          #              and very low correlation between annual income to interest rate (0.049)
```


There is high correlation between annual income to loan amount (0.27, this implies that higher income individuals apply to higher loan amounts.

Very low correlation between annual income to interest rate (0.049), there is positive but negligible influence on higher income individuals to apply for loan based on interest rate.

```
In [121]: df1 = df[['int_rate','loan_amnt','annual_inc','dti','acc_now_delinq','chargeoff_within_12_mths']]
df1.corr()
corrMatrix = df1.corr()
sns.heatmap(corrMatrix, annot=True)
plt.show()

#Observation: Loan amount & interest rate has higest correlation (0.3) out of all variables correlation.
#            dti and annual income has negative correlation (-0.12)
#            Loan amount and dti has positive but very little correlation (0.062)
#            Interest rate to annual income has positive but negligible correlation (0.049).
```
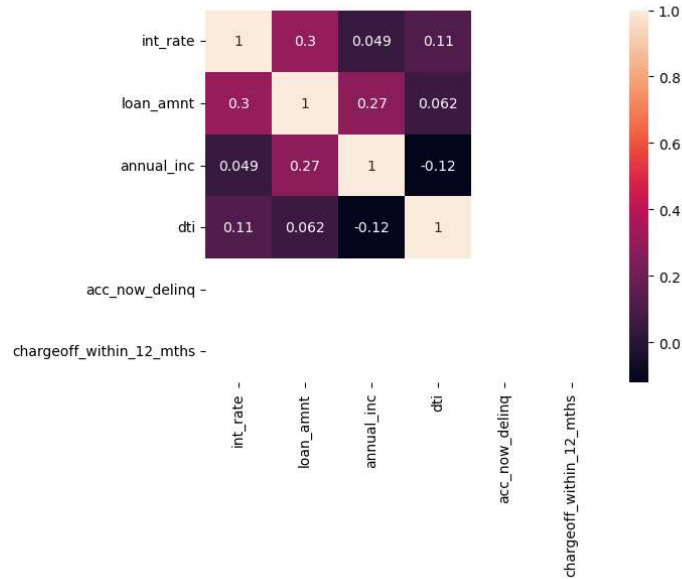

Annual income to DTI has very minor but negative correlation (- 0.12).

Loan amount to DTI has positive but little positive correlation (0.062).

Thank You