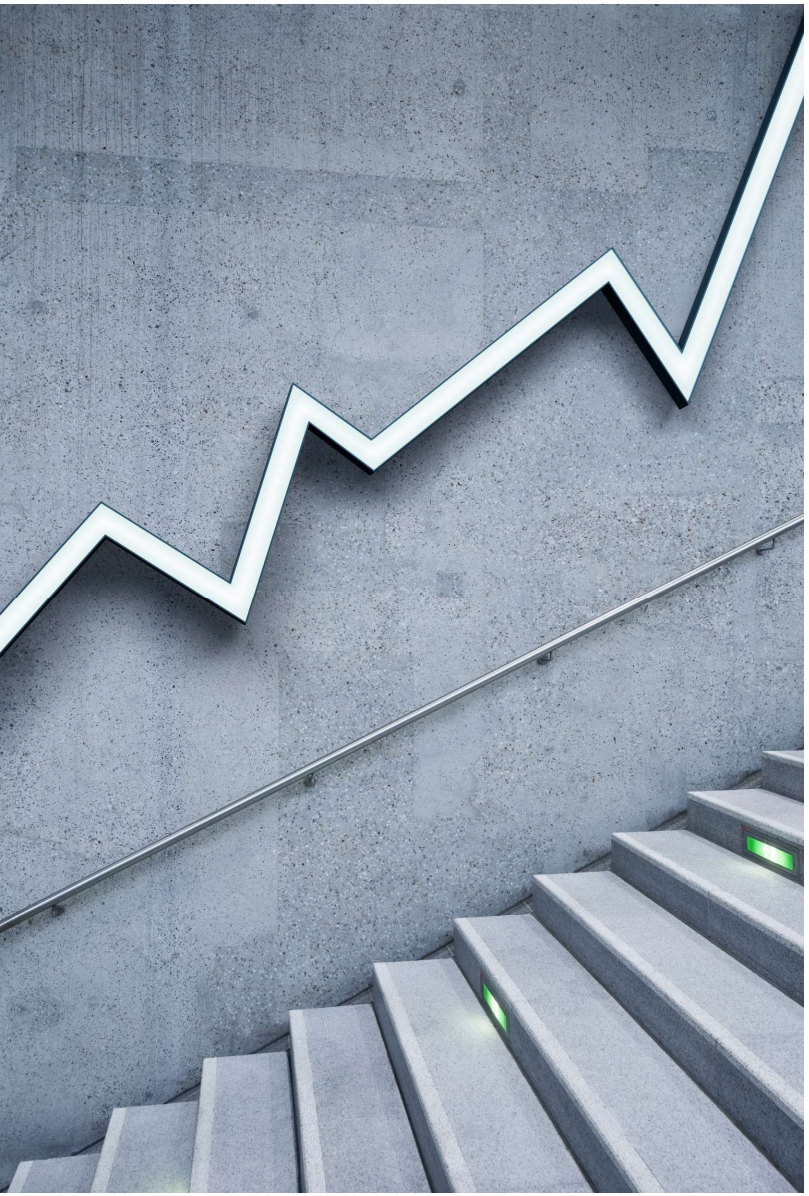


# Lending Club Case Study

ABHIROOP BHATTACHARYA

ANINDITA DEB

# Introduction

Lending Club connects borrowers with investors. Borrowers can get loans for a variety of purposes whereas investors can earn interest on their loans by investing in Lending Club notes

# Problem Statement

Lending club wants to enhance its ability to predict which borrowers are likely to default on their loans. This will help Lending Club to make better lending decisions and to protect its investors.

# Analysis method

The analysis is broken down into 4 steps:

1. Data understanding
2. Data cleaning
3. Segmented univariate and bivariate analysis
4. Conclusions

# Data understanding

Includes information about borrowers

EG: employment history, annual income, address.

It also includes information about loans

EG: amount, term, interest rate, installment.

# Data cleaning

## The following steps will be taken for this part of the analysis

- Converting date columns to Pandas DateTime type and percentage columns to numerical type after removing the % symbol.
- Removing columns which contain more than 10% of missing values or only a single unique value since they will not be of use in analysis.
- Dropping missing values where imputing them with mean/median/mode will heavily skew the data.
- Imputing missing values where feasible.
- Converting numeric columns to categorical where feasible.
- Removing outliers from all the numerical columns.

# Segmented univariate analysis

This analysis has been done on the numerical columns, segmented by loan status

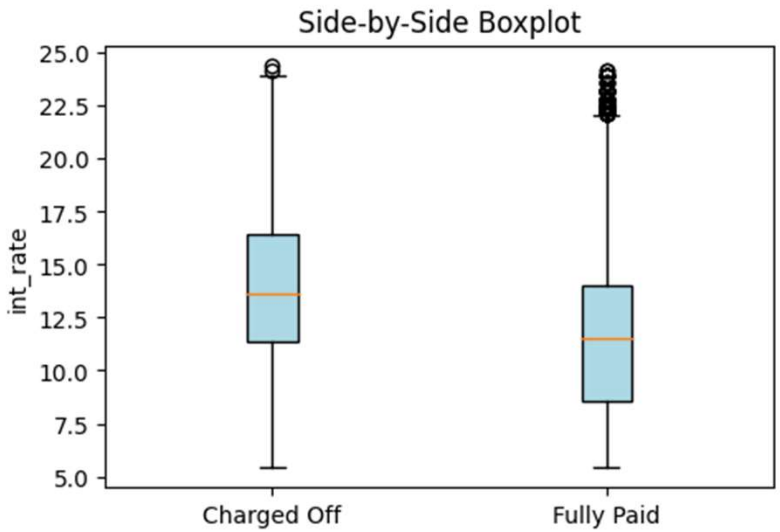
The purpose behind it is to discover the numerical variables which have the most stark differences between loans which were charged off and those which were fully paid

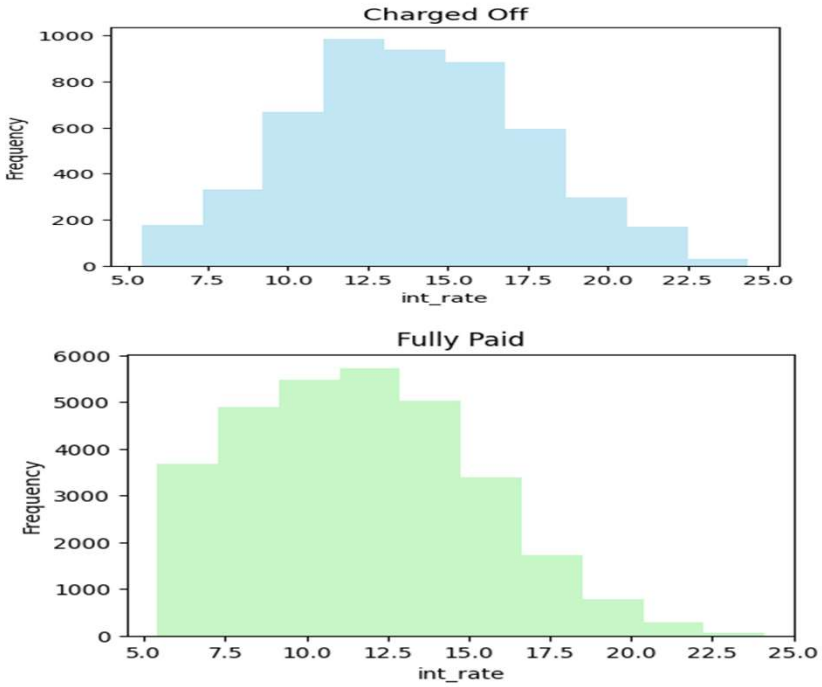
The differences were discovered via descriptive stats, side-by-side boxplots and histograms

The largest differences were in the variables

- Interest rate
- Annual income
- Revolving credit utilization

# Segmented univariate analysis – Interest Rate



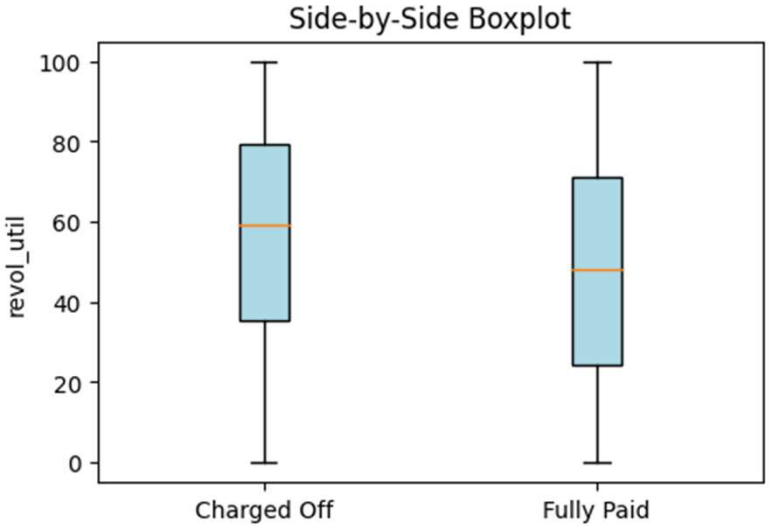
# Segmented univariate analysis – Annual Income

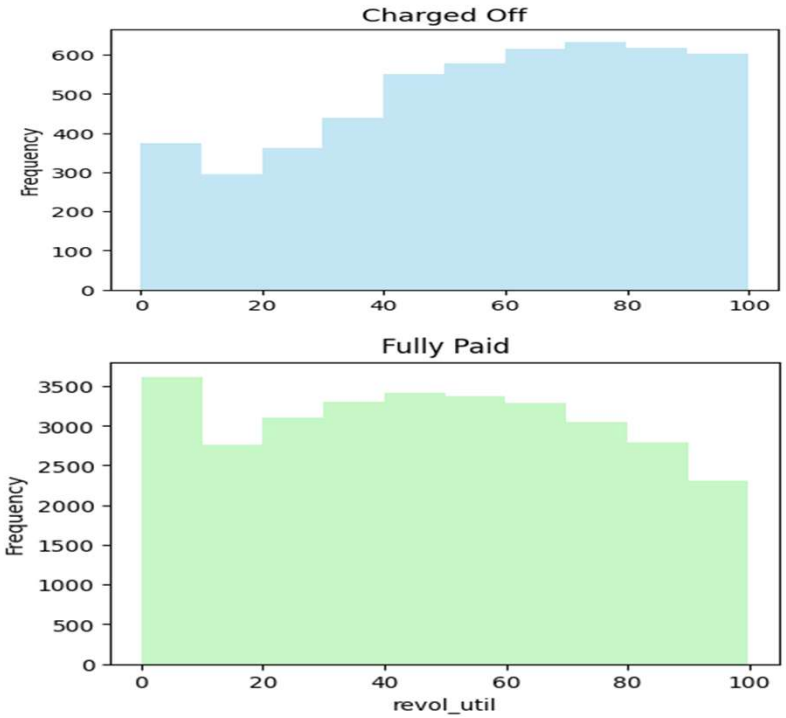
```
Descriptive Stats for annual_inc
Min:
df1:4080.0
df2:4000.0
difference = -1.9607843137254901
Max:
df1:1250000.0
df2:6000000.0
difference = 380.0
Mean:
df1:63121.02501381215
df2:70031.023523874
difference = 10.947221640570323
25th percentile:
df1:38400.0
df2:42000.0
difference = 9.375
Median:
df1:54000.0
df2:60000.0
difference = 11.11111111111111
75th percentile:
df1:75000.0
df2:84000.0
difference = 12.0
Plots for annual_inc
```

Df1 here refers to loans which were charged off

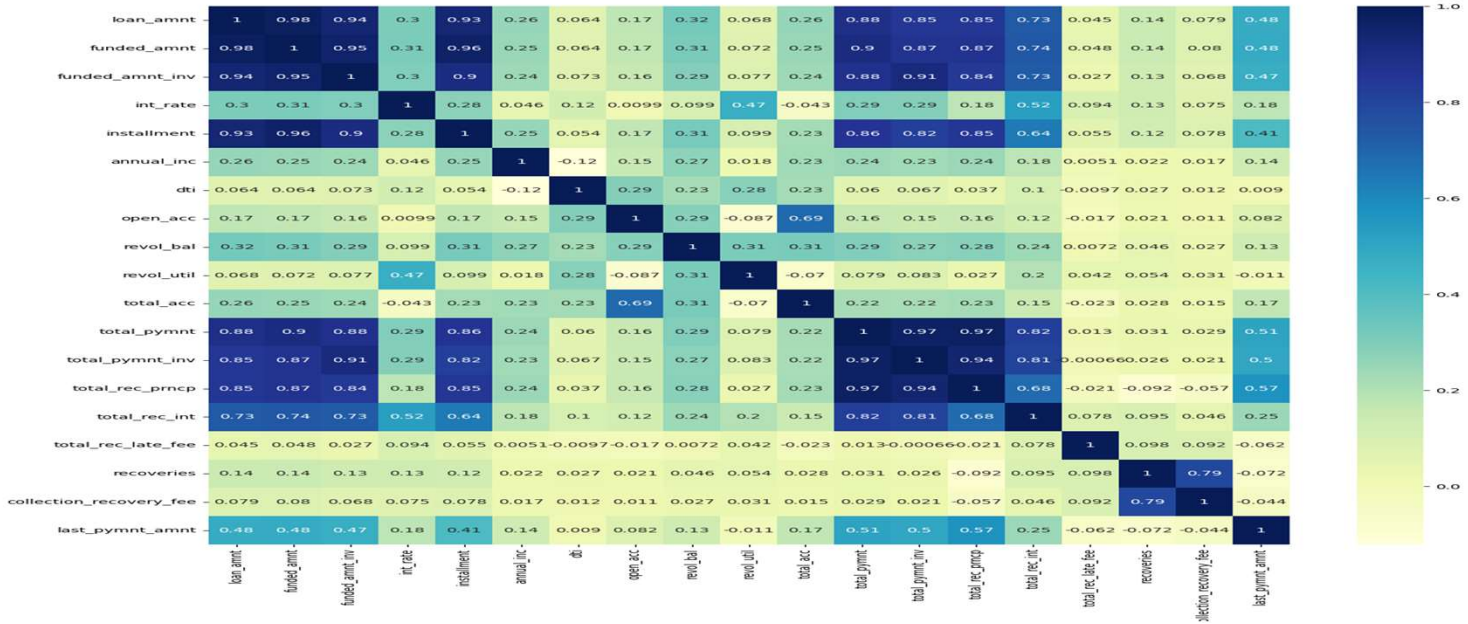
Df2 refers to loans which were fully paid

# Segmented univariate analysis – Revolving Credit Utilization



# Bivariate analysis - Numerical



# Bivariate analysis - Categorical

Here, the purpose is to compare how the values are spread among the various categorical columns

This is achieved through pivot tables with a count applied as the aggregate function

By weighing the percentage of values within loans which were charged off vs those that were fully paid, we can then compare how these values are spread among various categories

The columns which showed the largest differences are:

- Grade
- Public record bankruptcy
- Derogatory public records

# Bivariate analysis – Grade and Subgrade by loan status

Bivariate analysis of spread of values of grade vs loan_status

| loan_status | Charged Off | Fully Paid |
|---|---|---|
| grade | | |
| A | 10.280189 | 28.236129 |
| B | 25.355170 | 31.128091 |
| C | 23.796369 | 19.872973 |
| D | 19.988161 | 12.141729 |
| E | 13.003157 | 5.987039 |
| F | 5.741910 | 2.018248 |
| G | 1.835043 | 0.615791 |

Bivariate analysis of spread of values of sub_grade vs loan_status

| loan_status | Charged Off | Fully Paid |
|---|---|---|
| sub_grade | | |
| A1 | 0.453828 | 3.204694 |
| A2 | 1.243094 | 4.255731 |
| A3 | 1.657459 | 5.110101 |
| A4 | 3.176796 | 8.163265 |
| A5 | 3.749013 | 7.502337 |
| B1 | 2.920284 | 4.881194 |
| B2 | 3.946330 | 5.390592 |
| B3 | 6.136543 | 7.479769 |
| B4 | 5.682715 | 6.444853 |
| B5 | 6.669298 | 6.931683 |
| C1 | 5.899763 | 5.245511 |
| C2 | 5.682715 | 4.958571 |
| C3 | 4.794791 | 3.749557 |
| C4 | 3.729282 | 3.033820 |
| C5 | 3.689818 | 2.885514 |
| D1 | 2.900552 | 2.343876 |
| D2 | 4.853986 | 3.111197 |
| D3 | 4.617206 | 2.614695 |
| D4 | 3.926598 | 2.163330 |
| D5 | 3.689818 | 1.908631 |
| E1 | 3.709550 | 1.647484 |
| E2 | 2.959747 | 1.366992 |
| E3 | 2.111287 | 1.221911 |
| E4 | 2.308603 | 0.922075 |
| E5 | 1.913970 | 0.828578 |
| F1 | 1.677190 | 0.677048 |
| F2 | 1.262826 | 0.499726 |
| F3 | 0.927388 | 0.373988 |
| F4 | 0.947119 | 0.293387 |
| F5 | 0.927388 | 0.174098 |
| G1 | 0.513023 | 0.203114 |
| G2 | 0.532755 | 0.151530 |
| G3 | 0.355170 | 0.074153 |
| G4 | 0.256511 | 0.132186 |
| G5 | 0.177585 | 0.054809 |

# Bivariate analysis – Public record bankruptcy and derogatory public records vs loan status

Bivariate analysis of spread of values of pub_rec vs loan_status

| loan_status | Charged Off | Fully Paid |
|---|---|---|
| pub_rec | | |
| 0 | 91.969219 | 95.257439 |
| 1 | 7.833465 | 4.597479 |
| 2 | 0.197316 | 0.116065 |
| 3 | NaN | 0.022568 |
| 4 | NaN | 0.006448 |

Bivariate analysis of spread of values of pub_rec_bankruptcies vs loan_status

| loan_status | Charged Off | Fully Paid |
|---|---|---|
| pub_rec_bankruptcies | | |
| 0.0 | 93.764799 | 96.266563 |
| 1.0 | 6.195738 | 3.723764 |
| 2.0 | 0.039463 | 0.009672 |

# Analysis Findings

1. Those who default tend to get loans at a higher interest rate as compared to those that do not
2. Those who have fully paid off their loans tend to have a higher income that those who default on their loans
3. Revol_util is higher for defaulters, indicating that they tend to use up more of their available lines of credit as compared to those who have not defaulted
4. The percentage of defaulters with one public record bankruptcy is twice that of non-defaulters with one public record bankruptcy
5. There are no defaulters with more than 2 derogatory public records, however non-defaulters have more
6. As far as grade is concerned, very few defaulted loans are grade A (of those, the majority are A4/A5). However, among non-defaulters, more than a quarter of the loans are grade A

# Conclusions

1. Defaults are more common with those loans which get higher interest rates, lower annual income, higher debt to income and higher revolving credit utilization
2. A grade loans are less likely to be default
3. A single public record bankruptcy is more likely to cause default
4. During model creation, all columns with over 90% correlation become redundant and either one or a ratio of one to the other is required (ex. Ratio of loan amount invested to loan amount)

# Thank You