**Summary of Lending Club EDA**

1) Total dataset contains 39717 rows and 111 columns.

2) We found 54 such columns where all the values were null and removed those columns.

3) We are left with 57 columns now.

4) Dropped few more columns:

- Cannot predict next\_payment\_date, so it's better to drop the column which has 97.129693% of missing data.

- Months since last record, this field has 92.985372% of missing data. We do not know what NA means here. So dropped it.

- mths\_since\_last\_delinq has 65% of missing data, and this field isn't that relevant to work with only 35% available data.

- desc, its just description/reason for the loan.

- emp\_title is not relevant here so dropping it.

5) Considered emp\_length nulls as separate values for analyzing them seperately.

6) Replaced NAs where they have fully paid , with 0 bankrupties as we saw that majority of them with 0 bankrupties have paid their loans.

- We dropped the missing NAs in pub\_rec\_bankruptcies if they have been charged\_off for their loans. This is because, we didn't want to deal with such data as there are no evident proofs to impute right values in pub\_rec\_bankruptcies.

7) Removed all the rows with missing last payment due

8) Removed columns with single value throughout rows.

- chargeoff\_within\_12\_mths, collections\_12\_mths\_ex\_med

9) Removed nulls for revol\_util column which constitued 0.1% nulls

10) Dropped tax\_liens columns as it contained only 0's and nulls

11) Now there are 0 nulls in all columns.

12) List of columns to keep based on assumptions below:

- id : loan\_id

- pub\_rec: We can analyze effect of loan status on derogatry public record.

- revol\_bal: we can check credit effect on loan\_defaulters

- revol\_util: we can check whether deafulters behaviour toward credit utilization

- total\_acc: we are expecting positive co-relation between defaulting and total number of credit lines

- out\_prncp : co-relation between out\_prncp\_inv and out\_prncp is very high. So keeping out\_prncp.

- total\_pymnt : co-relation between total\_pymnt\_inv and total\_pymnt is very high. So keeping total\_pymnt.

- term : loans are given in two terms i.e. 3years or 5years. So, we can analyse effect of loan terms on defaulters

- int\_rate : it is key factor in any loans purchase.

- installment : it is also key factor in any loan process

- total\_rec\_int : it is also key factor in any loan process

- total\_rec\_late\_fee : it is also known key factor in any loan process

- last\_pymnt\_d : we can derive another column using issue\_d and last\_pymnt\_d to get probable time period of defaulting

- issue\_d : important for above requirement

-open\_acc : to check if credit lines are affecting default rate

- earliest\_cr\_line : We can find how quick a person is to default after he is given a credit line

- delinq\_2yrs : We can see behaviour of a defaulter in the last two years over his alloted credit

- loan\_amnt: Main criteria to decide the defaulting rate

- grade: We can check the relation between level of loan given and the defaulting behaviour

- sub\_grade: We can check the relation between subgrade of LC and the defaulting behaviour. And also which subgrades within a grade are more susciptable for defaulting

- emp\_length : Will help in work experience v/s defaulting comparision

- home\_ownership: The home ownership usually gives confidence to get a loan. Good metric for comparision

- annual\_inc : Good metric to evaluate on this front.

- verification\_status : income verification plays a key role in defaulting behaviour usually.

- loan\_status: A kind of target variable for this EDA

- purpose: good metric to see the reason of loan and eventual defaulting behavior if any.

- state: origin state behaviour on defaulting can be checked

- dti - debt to income is a ratio if more , chances of defaulting are more. But we will indulge later based on data for this.

- pub\_rec\_bankruptcies: relation between this and defaulting is an intersting insight.

13) List of columns to drop:

- inq\_last\_6mths : Since person's enquiry process is what we are not concerned about.

- initial\_list\_status : There is only one value in entire column.

- out\_prncp\_inv : co-relation between out\_prncp\_inv and out\_prncp is very high. So dropping out\_prncp\_inv.

- total\_pymnt\_inv : co-relation between total\_pymnt\_inv and total\_pymnt is very high. So dropping total\_pymnt\_inv

- funded\_amnt : amount of loan amount committed when approved at that time, we are focussing more on what is approved.

- total\_rec\_prncp : its the difference of loan approved and outstanding principal. We are choosing out\_prncp

- recoveries : since this is post loan application process, we are not concerned about it.

- collection\_recovery\_fee : this column values amount to recovery fees required for collection through third party agencies or any other means. So, we can ignore this column as it is post charge off process.

- last\_pymnt\_amnt: As this amount refers to last month's payment, no insight can be drawn from this column. It is better to drop it.

- last\_credit\_pull\_d : This colum values are last month's credit line pulled for this loan. So, not much meaningful insight can be drawn from it.

- policy\_code : There is only one value throught column. Hence we can drop it.

- application\_type : There is only one value throught column. Hence we can drop it.

- acc\_now\_delinq : There is only one value throught column. Hence we can drop it.

- delinq\_amnt : There is only one value throught column. Hence we can drop it.

- member\_id: we can eliminate, just keeping id is enough because both were unique

- funded\_amnt\_inv : co-relation between funded\_amnt\_inv and funded\_amnt is very high. So dropping funded\_amnt\_inv

- pymnt\_plan: There is only one value throught column. Hence we can drop it.

- url: We can remove this as its the url of the loan which looks like https://lendingclub.com/browse/loanDetail.action?loan\_id=1077501 and we are anyways keeping the id column.

- title: we can ignore this column, as we are looking at purpose column. Going indepth for this column requires more processing like clustering etc..

- zip\_code : Doesn't have significant value to decide defaulter behaviour based on his origin place, doing on state level is better

- ["inq\_last\_6mths", "initial\_list\_status" , "out\_prncp\_inv" , "total\_pymnt\_inv" , "recoveries", "collection\_recovery\_fee" ,"last\_pymnt\_amnt", "last\_credit\_pull\_d" , "policy\_code" , "application\_type","member\_id","funded\_amnt\_inv","pymnt\_plan","url","title","zip\_code","acc\_now\_delinq","delinq\_amnt","funded\_amnt","total\_rec\_prncp"]

14) dropped rows for "current" loan\_status as they would not add much value to defaulting behaviour.

- Marked "charged off" as 1 and "Fully Paid" as 0 for simplifying our analysis.

15) After all the above processes we are now left with 38329 rows and 29 columns.

- Number of rows dropped: 1388(3.4947251806531208%)

- Number of columns dropped: 82(73.87387387387388%)

**Univariate Analysis**

1) loan\_amnt

- Divided loan amount into 7 buckets for better analysis.

- Above 15,000 i.e., from bucket 4 we see that defualting rate is increasing.

- In the buckets 0-5k , 5-10k and 10-15k we see similar defaulting rate around 13%

- We can infere here that , if the loan amounts are high, defaulting rate is increasing and paying fully is decreasing.

2) annual\_inc

- since there are no outliers on the lower side we can focus on the outliers above upper extreme.

- mean and standard deviation are very near to each other. It signifies there are lot of outliers on the upper side as seen above.

- Above 12,50000 we don't see any borrowers defaulting when seen at the box plot of 0 and 1

- We see similar normal distribution for both the types within 3 standard devialtions data(till upper whisker length)

- Both of them have salaries ranging below 150000 itself.

- split annual\_inc bracket into 7 buckets of 25000 values each.

**-** We can observe above that defaulting rate decreases as the income increases. The number went down from 18.13% to 10.12% accross 6 buckets of annual income til 150000.

3) int\_rate

- We see that mean rate is around 12%

- We see more of a near to normal distribution of rates , which might be based on other factors like reason, amount\_loan,policy of company etc..

- Most of the loans in IQR are between 8.9% - 14.4% which is generally seen in most of the real life loans.

- Defaulters have higher interest rates with 13.5 and non-defaulters having 11.5% median in the data

4) term

- we observe based on above plot, that defaulting rate has increased around 14.35% from 10.55 to 25.15 when the term changed from 3years to 5years.

- Giving loans for lesser tenure sounds better!

5) installment

- It contains lot of outliers above upper whisker. It might be due to the people with higher salary taking higher loans

- IQR range is very less

- Installments are similar to both the categories as its obvious that they will receive installements based on the loan amount approved.

- Both def and non-def have significant outliers due to high std.dev

- no significant difference in IQR as well for these both

- Similar behaviour to that of loan\_amnt as installment is dependent on loan amnt and interest rate

6) verification\_status

- Based on data we see that non-verified loans end up getting less defaulted than verified and source verified

- Even within only defaulters verified and non-verified are very close.

7) grade

- We can see that A type loan has very less defaulting rate of 5.8%

- As we go from A to G , the defaulting rate is decreasing

-When looked only at defaulters, there are less defaulters in G and highest in B

8) sub\_grade

- We see that sub grades in F and G have high default rates above 25%

- F5 records the highest default rate in all of the sub grades

- all the subgrades in A are having very less default rates

9) home\_ownership

- we can see that defaulting percentage is independent of house\_ownership type.

- Above data can be read as, among rented 14.83% are defaulters, so are other categories similarly

- Among the defaulters 50% of them are renting a home

- just 7.8% are owning one

- we can also see that 41.7% of the people have homes at mortage, which means its already kept as surity for something else.

10) emp\_length

- defaulting percentage in 10 out of 12 categories is almost near to 13%.

- nothing insightful above

- When looked at % of categories within defaulters only, we see that 24% of the borrowers have more than 10 years of experience.

- Something quite interesting here is the borrowers within 1 year of experience are defaulting more than people with higher experience.

11) purpose

- the above plot shows percentage of defaulters within a given purpose.

- we can see that defaulting percentage is higher for debt\_consolidation borrowers i.e. 26%

- purpose field seems to be the key factor in determining loan defaulter

- we can see that people buying loan for debt\_consolidation are very much likely to default which accounts to almost 50% of purpose for loan buying

12) dti

- we can see that non-defaulters have good spread on both the sides of the mean

- defaulters have less spread on the left, which says that they have little more dti in general than non-defaulters

- even non defaulters have high dti on the right of mean, similar to defaulters. So not a great indicator for defaulting.

13) open\_acc

- Even non-defaulters have similar number of credit lines. If looked together for defaulters and non-defaulters, then the number of credit lines doesn't make much value.

- When looked numerically, both the categories are distributed very similarly.

- The values of mean and median are also almost close.

- Its better, we donot work with this column and drop it

14) pub\_rec\_bankruptcies

- As number of bankruptcies are increasing, the defaulting rate is also increasing

- Among the defaulters, there are 93% with 0 bankruptcies

### 15) revol\_utilization

- There are no outliers present

- defaulters have higher credit utilization with a mean utilization of 55% around, than non-defaulters who have around 47%

16) total\_acc

- We can see that both the categories have same kind of distribution in box plot.

- We cannot conclude anything on defaulting here due to its similar distributions on non defaulting

- Let's drop this column

17) pub\_rec

- We see that 91.6% of them within the defaulters are with 0 records.

- Nothing insightful out of this column, because non-defaulters also will have the similar prop

18) delinq\_2yrs

- We can see that within defaulters highest are 0 delinquacies like in non-defaulters.

- We can above that if we consider all the borrowers with a given number of delinquecies, we can see that as the delinquecies are increasing , the defaulting is also increasing. We donot see that trend beyond 2 delinquecies as the datapoints are less for more than 2 delinq.

**Derived Metrics**

1) tenure\_paid

- Formula = (((last paid date - last received date) - 1(because he will pay from next month))/ term) x 100

- we can clearly see that defaulter has very less paid tenure percentage(36.7%) as compared to non-defaulters(83%) in terms of median.

- This variable is a strong indicator if we are analysing post a borrowers payment cycle

2) total\_prncp\_paid\_perc

Formula:(total payment received / loan amount) x 100

- We can see that there are few borrowers who are shown as fully paid, but the principal percentage is not 100

- This might be due to some other factors like surity, loan\_waiver, house\_disposal etc.. which is out of the scope of dataset

### -Its clearly evident that the median of defaulters is near 36% prncp paid, but the non defaulters have paid 100% of the principal

- This is not a strong factor as its a post loan insight.

### 3) revol\_bal\_mult

- We can see that both defaulters and non defaulters have the same spread of values in terms of the amount revolving to their emi

Bivariate Analysis

1) Grade and Subgrade

- We can observe here that Grade F has the highest average defaulting percentage accross all the grades

- G is also close by F in avg defaulting percentage

- When seen at subgrade levels A has its sub grades recording the least defaulting percentage

- F5 records highest defaulting rate of 47.7%

- As the grade increases, any subgrade we might choose within that grade, we will end up with higher defaulting rate than the previous grade

2) Purpose Analysis

### i) loan amount bucket v/s defaulting in each purpose

- The definiton of Average defaulting, we will get the percentage of defaulters per loan\_amnt\_bucket and further average for all of them in each purpose.

- there are overall less probability of defaulting in major\_purchase.

- small\_businesses tend to default more than any other purpose. Even Medical looks a good area to catch defaulters.

- One interesting insight is there are 100% defaulters in who take loan in range of 30-35k (bucket 7) under "moving" purpose

- In wedding, loan\_amnt bucket 6(25000-30000) we can see a 0% defaulting

- When we did univariate analysis, we observed that half the defaulters were from debt\_consolidation. But when done along with loan bucket we see a different trend.

### ii)  income v/s defaulting in each purpose

- The definiton of Average defaulting, we will get the percentage of defaulters per income\_bucket and further average for all of them in each purpose.

- small businesses tend to default more on an average accross all the income buckets with 23% average.

- car loan borrowers default less on an average on all the income ranges

- In the field of Medical, we see that the least income bucket people that is 0-5K INR have the highest defaulting percentage

- In small businesses we see that the defaulting percentage increases for income 0-15K and then goes on decreasing in further income buckets. This might be due to higher income.

- In renewable energy we observe that borrowers with income in 5-10K range default way more than other income brackets.

- Even here observe that debt\_consolidation way below few other purposes in defaulting rate based on income.

### iii) interest v/s defaulting in each purpose

- We see that among the defaulters, median interest rate for house was 15.46% but amongst the non-defaulters it was around 11%, which significantly says due to high interest rate on house loans, there were more defaulters.

- Lets plot the median interest rates below for more insight

- We can clearly see above that defaulters are usually having higher median interest rates in almost all the purpose categories, which can be a deciding factor as well. Especially in house loans, we can see it more evidently.

### iv) verification v/s defaulting in each purpose

- In car,debt\_consolidation,credit\_card,small\_businesses,educational we see more default percentage in verified borrowers than non-verified borrowers.

- On an average, defaulting percentage is less in wedding purpose for all the types of verification\_status and highest for small businesses.

- In moving purpose, we see there is lot of defaulting perecentage in Source Verified type.

## 3) Income Bucket analysis

### i) Income v/s loan bucket on defaulting

- We see that under each income bucket, as the loan amount increases generally in most of them the defaulting percentage is also increasing. Especially is observed in income brackets 2,3,4,5

- If we look at income bucket 1(0-5000 INR annual income) , we see that most of the defaulters lie in loan-amount-bucket of 3(10000-15000 INR).

- On an average, we see that income bracket 6(125000-150000) has the least average defaulting percentage done over all loan\_amount brackets.

- Borrowers with income 5000-10000 are having the highest defaulting percentage of 23.4%

- As the income bracket increases after 125000 INR, the trend in the defaulting among the loan\_buckets is not so evidently pleasing. But based on the heatmap we can see that thicker colours are there in dark blue around income bucket of two and loan\_amnt\_bucket of 5,6,7

- Recorded 37% defaulting in income bucket of 50-75K and loan\_amnt in range of 30-35K

### ii) tenure paid across all the income buckets V/S Defaulting Behaviour

- We can clearly see that Median tenure paid percentage of defaulters is almost same across all income brackets, somewhere near 37%, but where as in non-defaulters its greater than 75%

- Among the non-defaulters the income bracket 1(0-25000) has higher median tenure paid percentage.

- We are seeing whiskers above 100% because people would have paid for more months than their tenure, due to lot of factors.

### iii) principal\_amount percentage across all the income buckets V/S Defaulting Behaviour[¶](http://localhost:8888/notebooks/LendingClubEDA_Final.ipynb" \l "iii)-principal_amount-percentage-across-all-the-income-buckets-V/S-Defaulting-Behaviour)

- We can clearly see that median of all the non defaulters in all the income buckets is 100% principal paid

- In non-defaulters we can observe that, apart from income bracket 5(41%) and 6(49%) who are usually paying more than other income brackets, all others have their median paid percentage near ~35%, which is lot lesser to what non-defaulters are paying.

## 3) employment analysis

### i) employment length v/s loan\_bucket analysis for defaulting

- Borrowers with one year experience taking the loan amount in the highest bucket seem to default less around 6.7%

- We can observe more defaulting in 30000-35000 loan bracket(7) across all emp\_length

- Amongst the emp\_length NA are defaulting more across all the loan\_amnt\_buckets.

### ii) employment length v/s income\_bucket analysis for defaulting

- borrowers with income in 1st bucket i.e., 0-5000 are having high default percentage across most of the employment years.

- Since 7th income bracket is >150000 and if we look only below 150000, we observe that, as income increases the defaulting rate is decreasing. Redness of the heatmap is evidently gradient

- Highest defaulting rate is seen in NA, interestingly we see 0% defaulting for income brackets 5 and 6 .