**Problem Statement**- For XYZ Corp we have to predict whom to issue loan in future based on different indicators.

**Dataset-** Dataset has 73 variables with 855969 rows these are different financial indicator based on which financial institutions decide whether to issue Loan or not.

**Variable Treatment-**

Since dataset has many variables initially we tried to exclude some variable which logically does not contribute much to model building.

|  |  |  |
| --- | --- | --- |
| **Variables** | **Descriptions** | **Reason to remove** |
| addr\_state | The state provided by the borrower in the loan application |  |
| desc | Loan description provided by the borrower |  |
| emp\_title | The job title supplied by the Borrower when applying for the loan. |  |
| id | A unique assigned ID for the loan listing. |  |
| member\_id | A unique Id for the borrower member. |  |
| policy\_code | publicly available policy\_code=1 new products not publicly available policy\_code=2 |  |
| purpose | A category provided by the borrower for the loan request. |  |
| title | The loan title provided by the borrower |  |
| verified\_status\_joint | Indicates if the co-borrowers' joint income was verified by XYZ corp., not verified, or if the income source was verified |  |
| zip\_code | The first 3 numbers of the zip code provided by the borrower in the loan application. |  |

There are two type of applicant in our data set

1) Individual

2) Joint (442 Entries)

|  |  |  |
| --- | --- | --- |
| **Variables** | **Descriptions** | **Comment** |
| annual\_inc | The self-reported annual income provided by the borrower during registration. | Merged |
| annual\_inc\_joint | The combined self-reported annual income provided by the co-borrowers during registration |
| dti | A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested loan, divided by the borrower’s self-reported monthly income. | Merged |
| dti\_joint | A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested loan, divided by the co-borrowers' combined self-reported monthly income |

Then there are few Variables which we have modified to define levels in them.

|  |  |  |
| --- | --- | --- |
| **Variables** | **Descriptions** | **Comment** |
| emp\_length | Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years. | It has entries like <1 year,10+ year, then many entries between 1-10 and n/a. |

In this we have defined level as

|  |  |
| --- | --- |
| < 1 | Low |
| 2-9 | Medium |
| 10+ | High |

After this we checked missing values in dataset, and imputed them with mean or mode value based on their data types.

And recognized few variables which have less than 14000 entries (less than 2%), so we have removed them.

|  |  |  |
| --- | --- | --- |
| **Variables** | **Descriptions** | **Counts** |
| open\_acc\_6m | Number of open trades in last 6 months | 13289 |
| open\_il\_6m | Number of currently active installment trades | 13289 |
| open\_il\_12m | Number of installment accounts opened in past 12 months | 13289 |
| open\_il\_24m | Number of installment accounts opened in past 24 months | 13289 |
| mths\_since\_rcnt\_il | Months since most recent installment accounts opened | 12935 |
| total\_bal\_il | Total current balance of all installment accounts | 13289 |
| il\_util | Ratio of total current balance to high credit/credit limit on all install acct | 11610 |
| open\_rv\_12m | Number of revolving trades opened in past 12 months | 13289 |
| open\_rv\_24m | Number of revolving trades opened in past 24 months | 13289 |
| max\_bal\_bc | Maximum current balance owed on all revolving accounts | 13289 |
| all\_util | Balance to credit limit on all trades | 13289 |
| inq\_fi | Number of personal finance inquiries | 13289 |
| inq\_last\_12m | Number of credit inquiries in past 12 months | 13289 |
| tot\_cur\_bal | Total current balance of all accounts | 13289 |

**Defining X and Y –**

Y –



From Above Pi chart it is evident that our data is biased, but it is in sync with real scenario, where defaulter customers are less.

Remaining Variables are our X.

**Analytical Understanding-**

As our Y is in binary form we recognized it as a Classification Problem.

And tried, implementing different Classification algorithms.