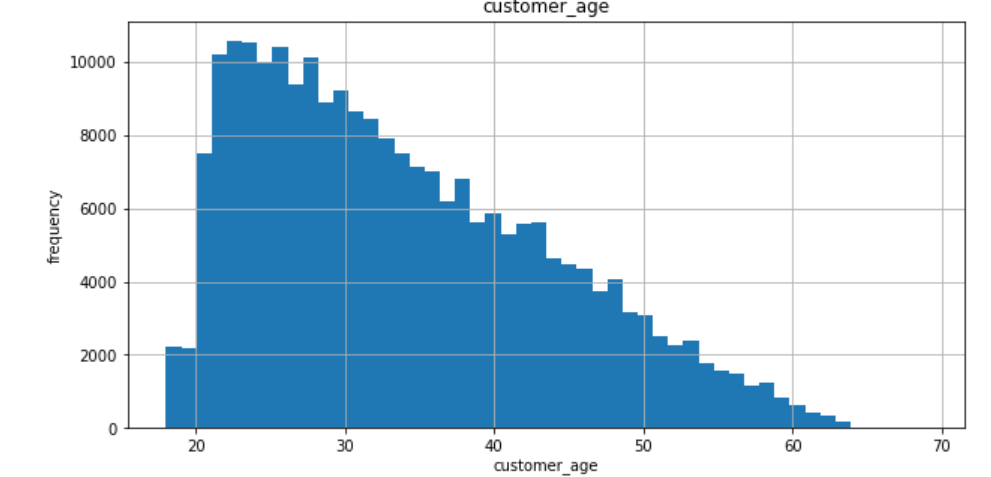
Loan Default EDA

# About the data

* The dataset contained near about 2,30,000 observation which was characterized by 41 different features consist of some Categorical and many oradinal and numerical data.
* Only one feature had missing values i.e. in Employment type which has been replaced by “No\_record”.
* Data cleaning has to be done to keep meaningful features in a proper format for our analysis.

# EXPLORATORY DATA ANALYSIS

In our dataset, we have many samples whose Bureauo history wasn’t available and are in the age of 20-40 contributing the majority sample and rest of the majority of the variable are highly skewed and had outliers samples.



Chart, waterfall chart

Description automatically generated

**Dependent vs Independent variable Correlation plot:**

In our correlation matrix with respect to target variable i.e. loan\_deafult , loan to asses ratio and disbursed\_amount seemed to be little correlated.

Which mean Higher the value of ltv or disbursed amount, the higher the chance of loan default and lower the CNS score higher the chance of loan default which is little negative correlated with the target variable. So lets see some digging in these variable ahead.

Text

Description automatically generated with low confidenceA picture containing company name

Description automatically generated

**Correlation plot within independent variables:**

Further we tried to look at the correlation plot within the independent variable which if important to get an idea of multicollinearity in our dataset.

And hence we found :

- Strong Correlation between sec\_current\_balance , sec\_santioned\_balance , sec\_disbursed\_balance

- Strong Correlation between pri\_santioned\_amount , pri\_disbursed\_amount

Hence only onle should remain between the highly correlated group.

A picture containing chart

Description automatically generated

**Checking target variable spread:**

As a rule in data science that we should always set the target to hit on before forging a tool. Hence in a similar manner we should always select the valuation metric on which our model will be valuated, which can be understand by looking the distribution of our target variable.

As we can see our dataset is highly imbalanced dataset hence learning a model will be challenging and deciding a valuation metric will be depend on the requirement of the project.

Chart, pie chart

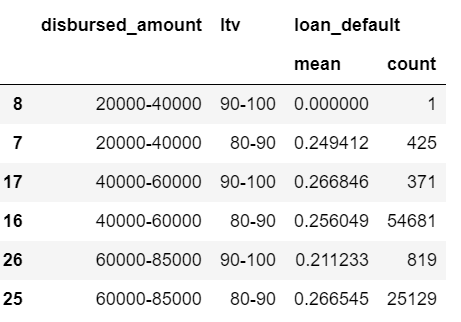
Description automatically generated

**Evaluating variable ltv :**

As we suspected the high the value of ltv tends to have a higher rate of default compared to less ltv.

With a good sample size it is proven.

Table

Description automatically generated Table

Description automatically generated

**Evaluating Performance CNS score:**

As we can observe visually major default occurs having low CNS score especially before 600-700 range.

Chart, bar chart

Description automatically generated

**Evalulating delinquent\_accts\_in\_last\_six\_months:**

As a logic we can think about someone who has been delinquent 1 or more than one time in recently months must be finacially weak or do not want to return a loan. Hence we get the fraud rate of upto **27%**.

Graphical user interface, text, application, chat or text message

Description automatically generated

**Combining the finding:**

So the evaluation with delinquent\_accts\_in\_last\_six\_month >0 and CNS score < mean\_cns\_value and ltv <mean\_ltv\_value we have evaluated the highest default rate on a small sample with the loan\_default upto **30%** .

**Conclusion from Initial EDA :**

- The combinations above tends to give higher default rates

- so delinquent\_accts\_in\_last\_six\_months, perform\_cns\_score & ltv has proven to be important factor at our initial Analysis