CREDIT RISK ANALYSIS

- Data Source – From Upgrad

Agenda

- ❑ Overall Approach
- ❑ Missing Data Treatment
- ❑ Outliers in Dataset
- ❑ Data Imbalance
- ❑ Univariate & Bivariate Analysis
- ❑ Correlation

Overall Approach

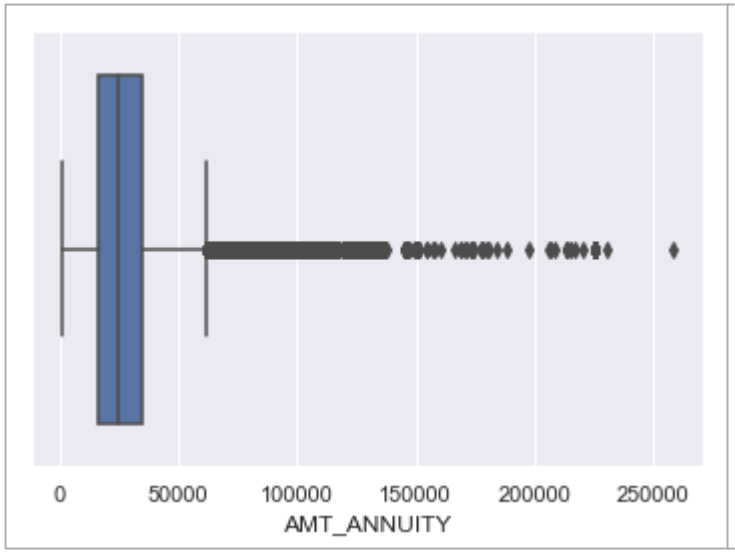
As the dataset is having large number of Null values. If we will directly delete the rows or columns then we might lost some features, variation which would be very helpful for analysis. In order to deal with null values, we have imputed the mean. The same process is repeated for previous application dataset. After imputing mean we have perform outliers analysis which will tell us how our feature varies and identify the values that are not suitable for analysis. After finding outliers in both the dataset, we have analyzed that our dataset is highly imbalance means defaulter are less as compared non defaulter. We have analyzed due to what client is defaulted.

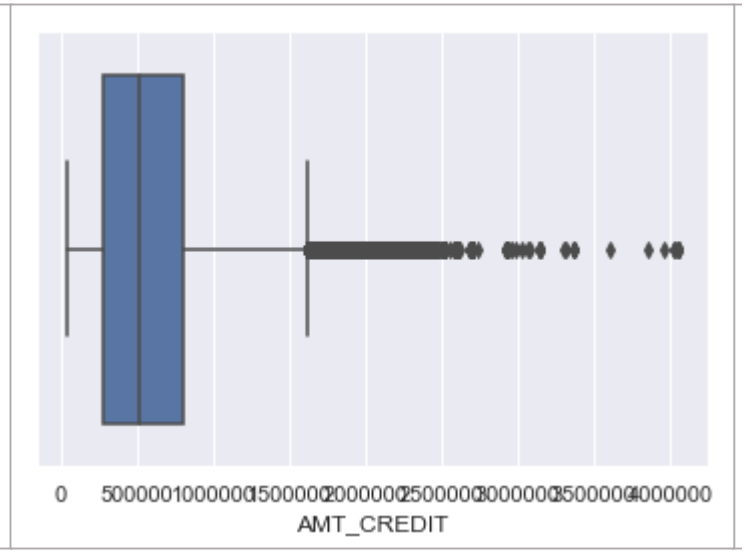
We have also perform all kinds of analysis on both the dataset and draw useful insights through it. We had used different kinds of plots like distplot, countplot, barplots. In the end we performed correlation between all the variables to check which features are highly correlated and plot heatmap for top 5 correlated features.

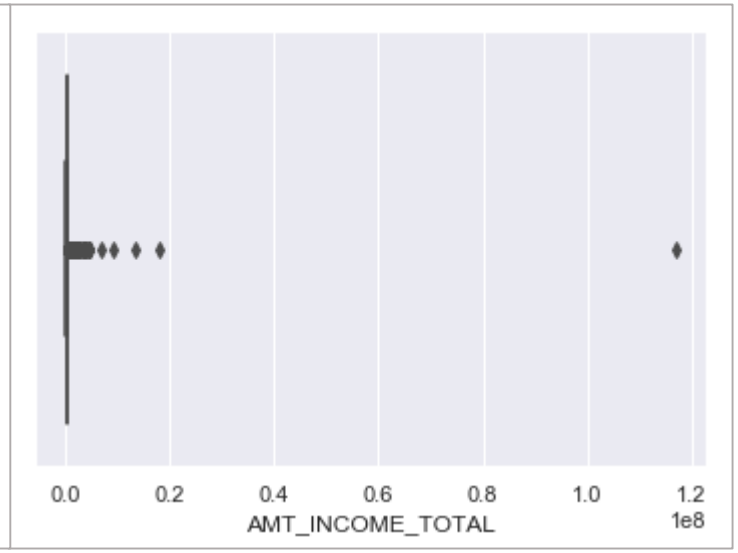
Missing Data Treatment

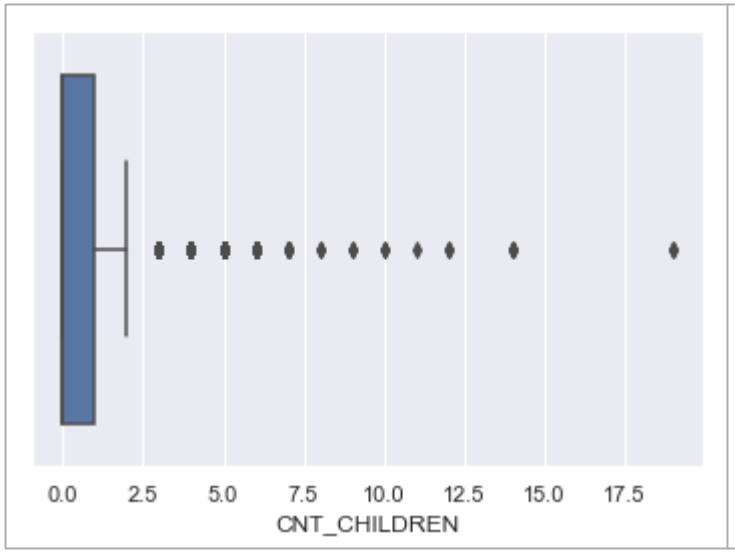
As we can see the amount of data and also treat our outliers correctly we have we considered mean to impute our null value.
As application data is having large number of feature. Inorder to make the dataset simple we have find that to remove one of the column which are higly correlated. Here the threshold is 0.95

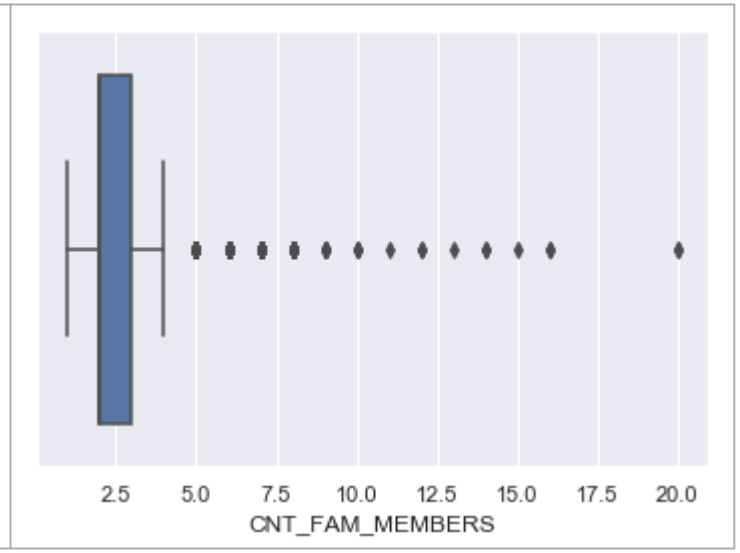
Outliers in Dataset





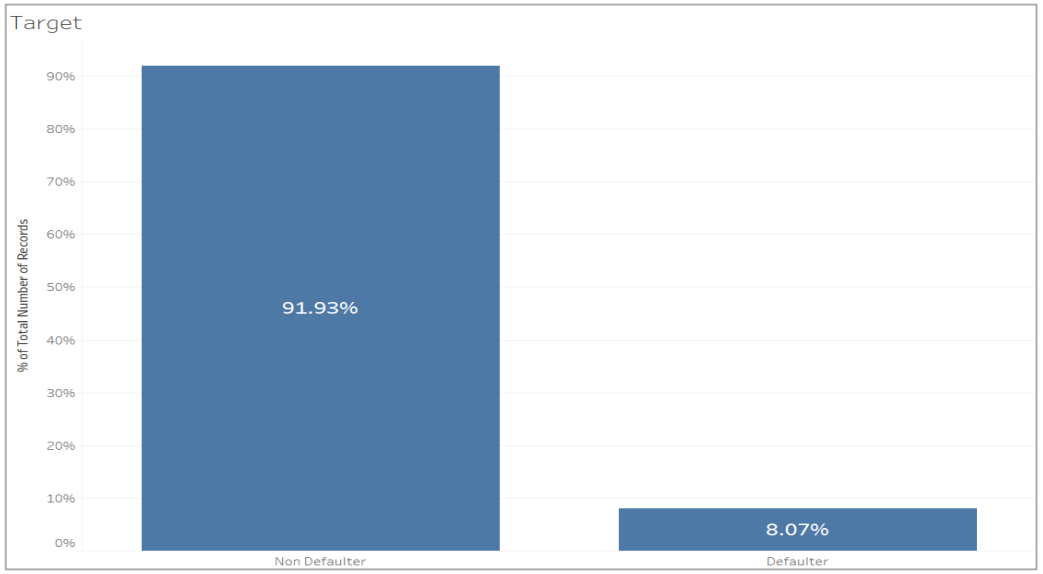


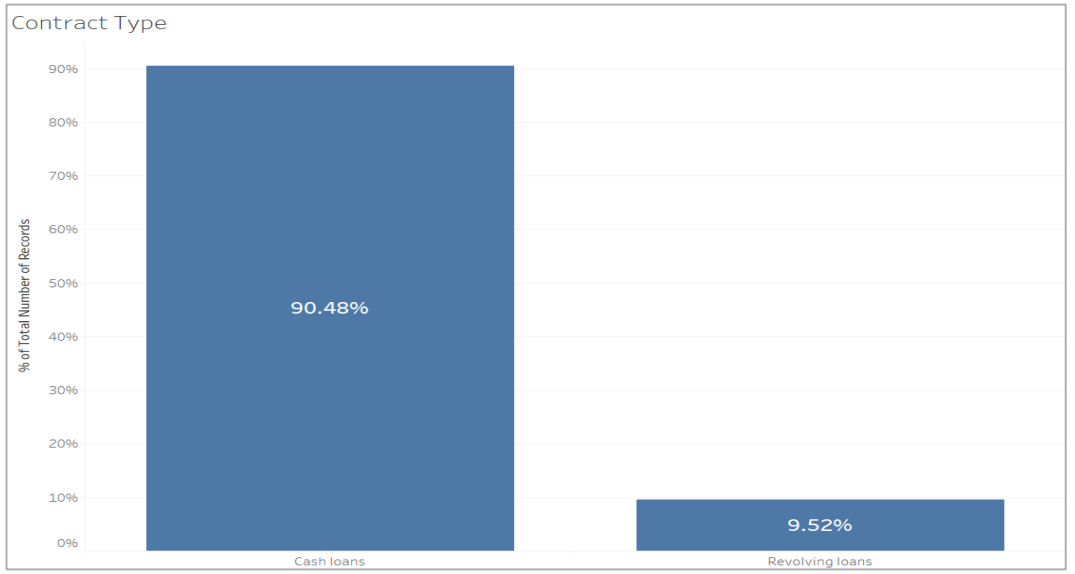


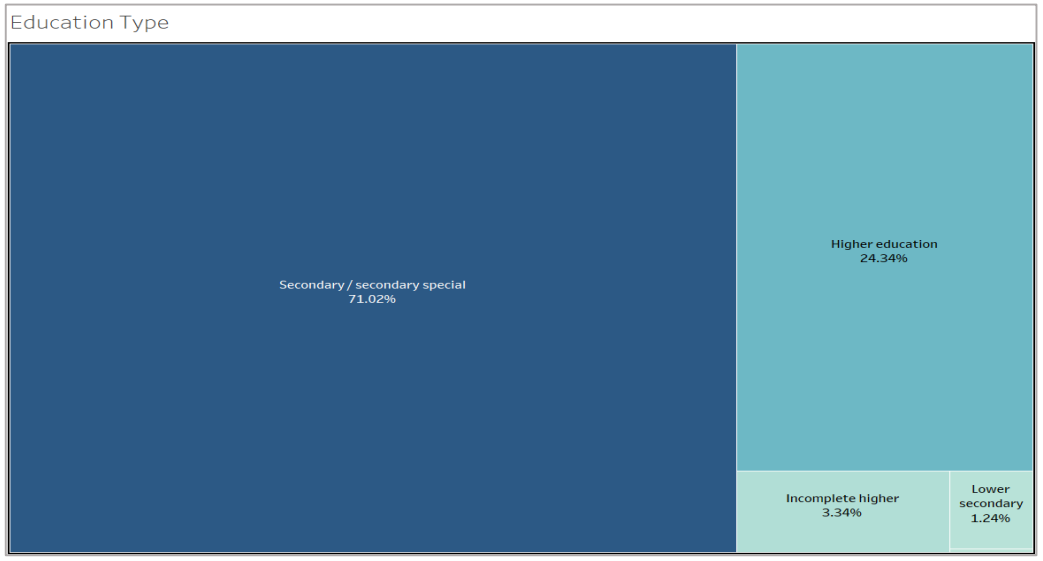


These five information is vital for our analysis and we can see rom the graph that some data is much farther form mean and as well as from 3rd Quartile.

Data Imbalance







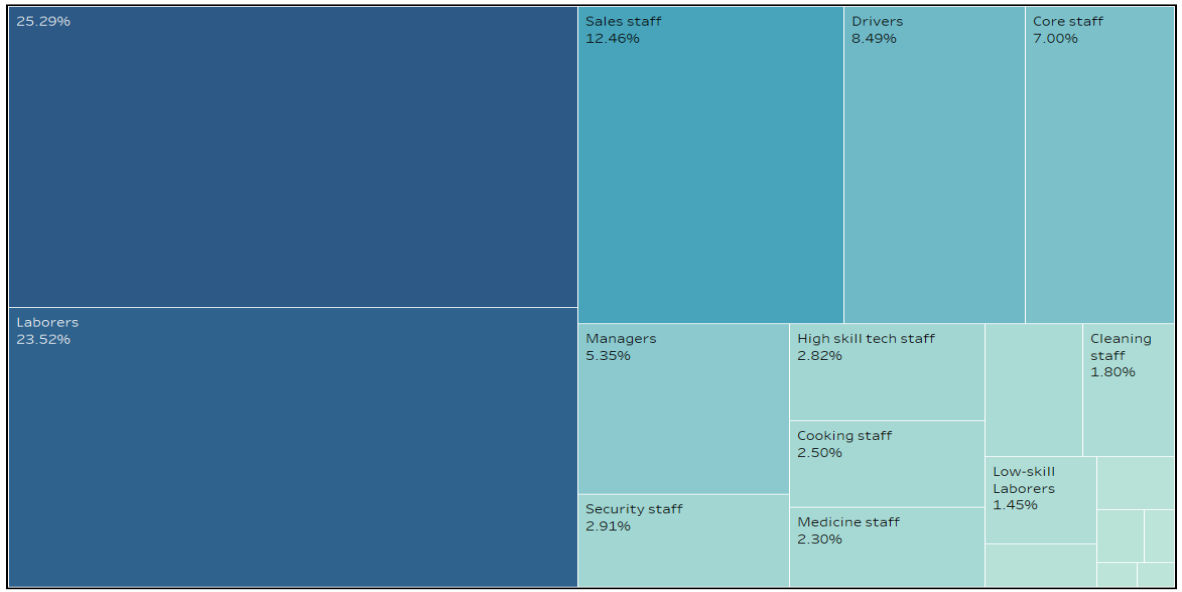
Inputs

As per the Chart, we can see three different column which are imbalanced.

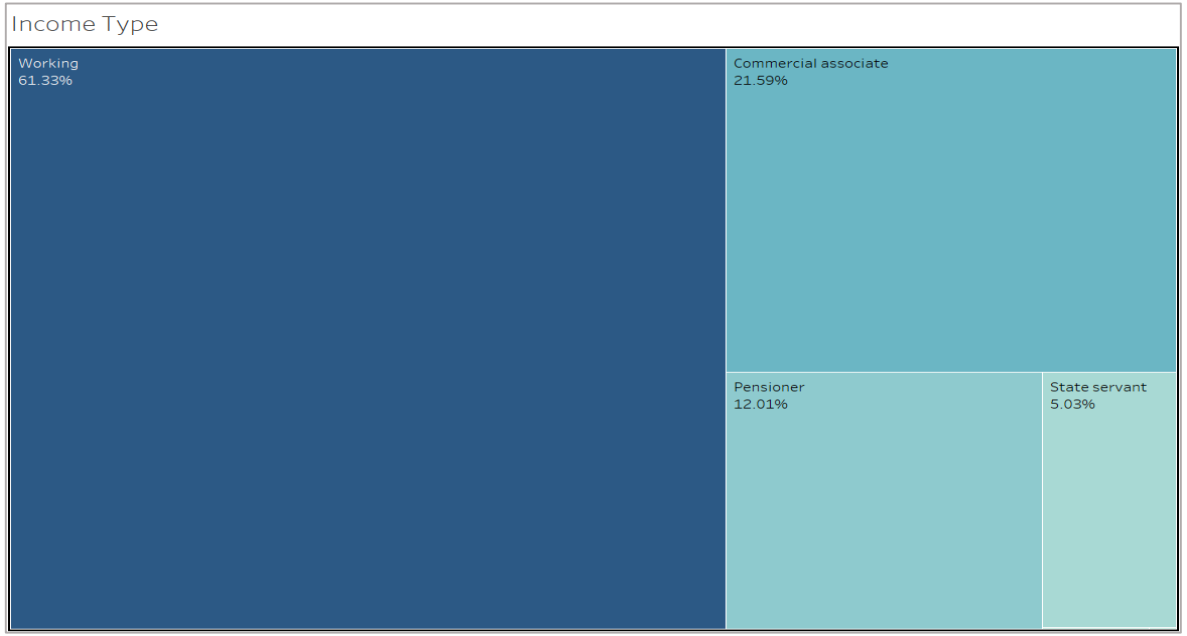
- In Target Column we can see that Non-defaulter percentage is equal to 91.93
- In Type of loan columns, cash loans is at 90.48%
- And in education type, secondary alone is at 71.02%

Univariate Analysis

Ocupation Type



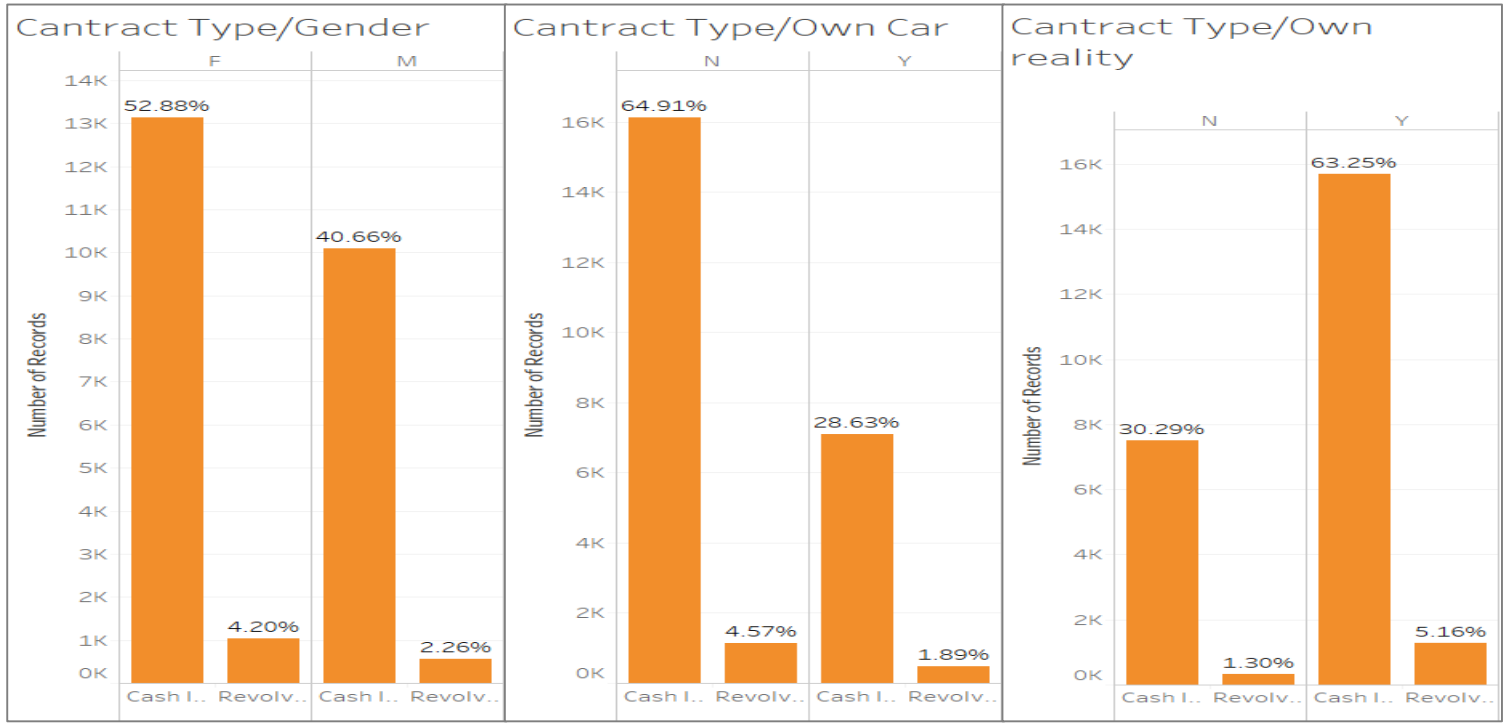
Income Type

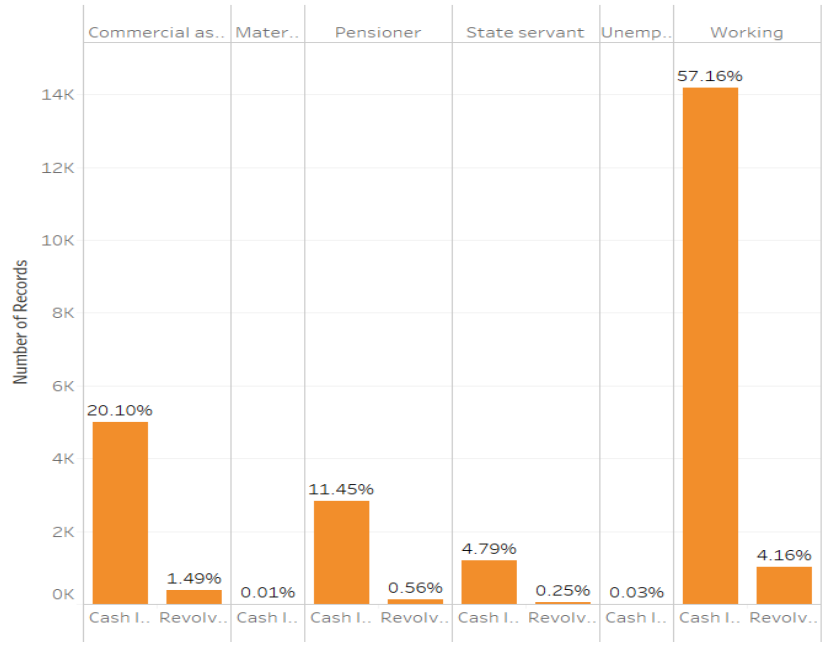


Family Status Type

From these 3 different univariate chart we can analyse the majority of the defaulter belongs to Married, Labours and working group

Bivariate Analysis





Form the above bivariate chart, conclusion we have below

- Comparing gender with contract type we can see women having cash loan have high chance of defaulting
- People who does not own a car and having cash loan have high chance of defaulting
- People having land/apartments and having cash loan have high chance of defaulting
- Clients with working income type and apply for cash loans have high chance of default.
- Combined we can see people with cash loans have high chance of defaulting

Correlation

Top Variable that correlated to defaulter are included in chart,

- If defaulters have external source of income then there will be less chance of defaulting
- If the loan has given to older person there is a high chance of default
- More number of Employed days means less chance that the person will not able to pay the loan
- But there we can see older person have more number of employed days so here we need to see if the person is older than he/she should have some kind of external source of income to do safe business.