**Financing the un-financed (Social Financial lending) - Milestone Report**





**THE QUESTION ?**

Financing or loans are essential for any community. Today Banks are desperate to provide loans to individuals having good tracking financial/credit history. But many people are not part of this system, they don't have a trackable financial history and thus Banks are reluctant to offer loans to them. And these people are often exploited by local untrustworthy lenders.The idea is to predict borrowers' repayment capability by using other easily available dataset like telco and transactional information. The problem is to predict loan pay-ability using various statistical and machine learning methods.

**DATASET USED :**

Below link contains the datasets for both training and testing. This dataset was a part of an online challenge given by Publicis Sapient for their Hiring challenge hackathon.As this challenge was given on techgig, the dataset was presented by techgig only. Since the dataset size is really big, I have given the link below to access it.

<https://www.techgig.com/files/DataScienceFullData/265810/Complete-Data-Set.zip>

**CHALLENGES IN THE DATASET :**

1. Data may also be corrupt (some of mandatory fields are missing),need to filter/impute those records.

2. Data was collected from multiple sources hence it may contain duplicate records, needed to clean up duplicate records.

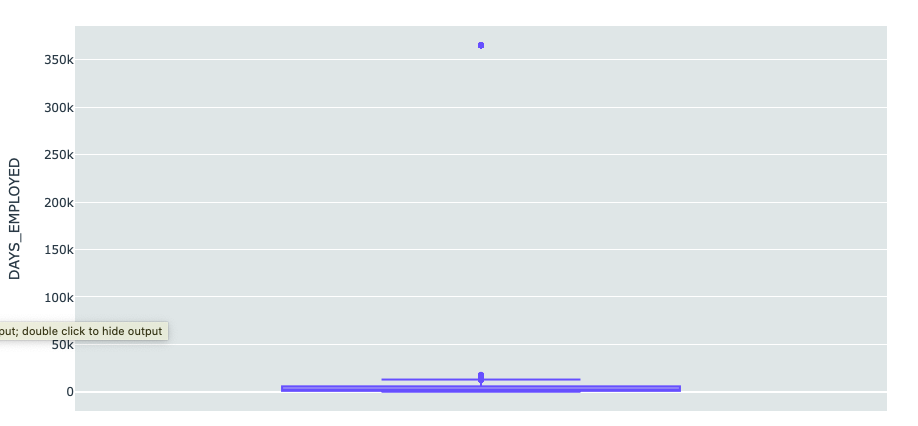
3. Highly imbalanced dataset for the target values.

**STEPS USED IN DATA WRANGLING :**

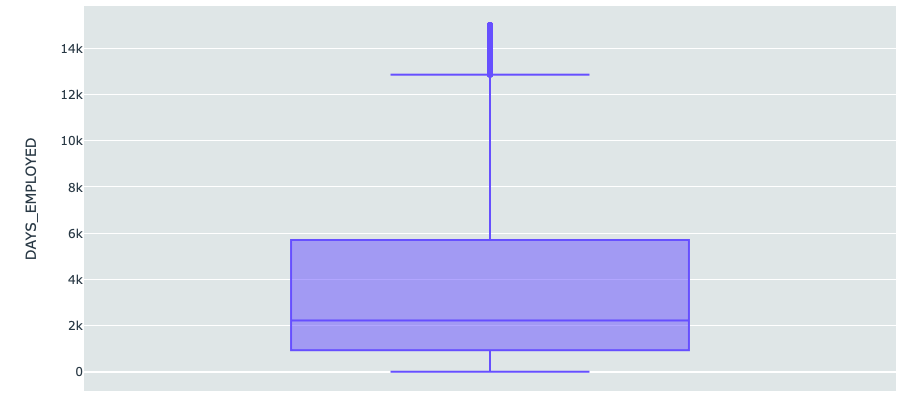
* Converted the categorical columns into binary format (0,1) using get\_dummies method.
* Some of the values for days were in negative,so changed them to positive using the .abs method
* Found out the outliers and removed them by defining a proper ranges for column values(e.g. Maximum Employment years = 45 )
* Changed the time period values from days to years for better visualization.
* Defined ranges for the amount related columns in order to standardize the value ranges.
* Found out the NaN values in the different columns and tried to impute them using different imputation techniques (Sklearn Imputer)

EXAMPLE:

DAYS employed range box plot before data wrangling looks like below:

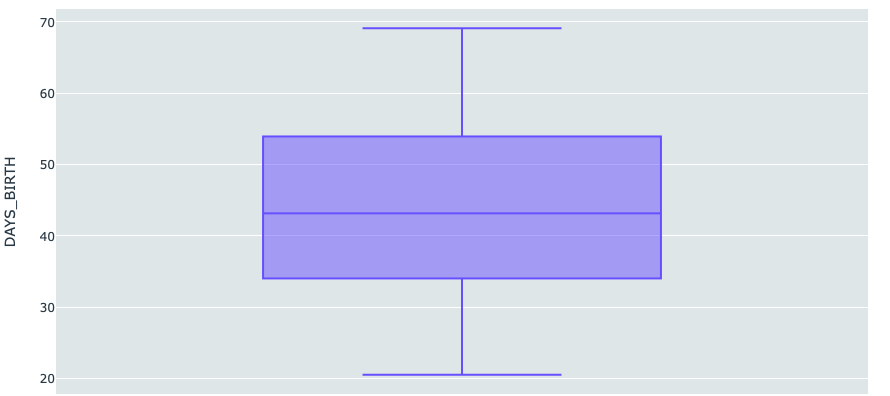


After data wrangling,standardizing the values and removing the outliers, the box plot looks as below :

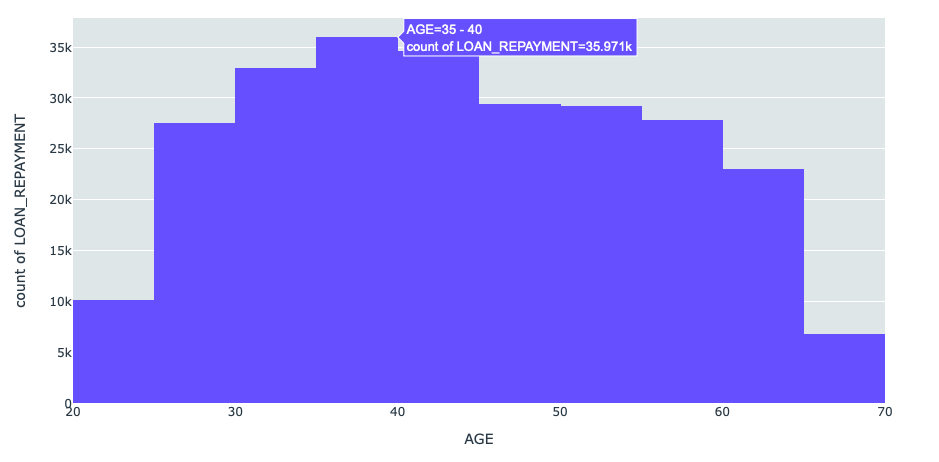


**EXPLORATORY DATA ANALYSIS :**

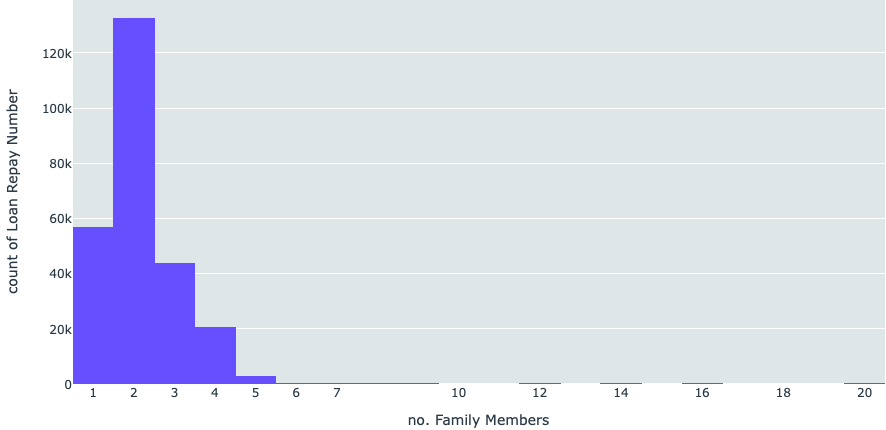
Box plots were used to get the ranges for the columns representing time. E.g. Below box plot shows the range for age of the borrower

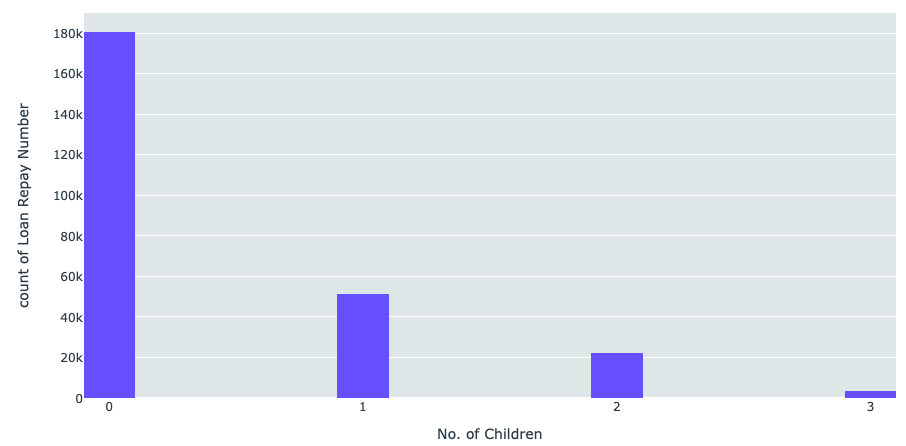


The age of the borrower was a big factor that was responsible for loan repayment by them. The below histogram shows the variation of the number of people paying the loan back as per their age.



Similarly, no. of children and family members also inversely affect the ability of repayment of loans of a borrower.





The financial status of a borrower is directly correlated to the ability of repaying the loans. Below plots show the different types of earnings/credits of a borrower compared to the number of people repaying the loans in that range of earnings.

Similarly, the inflation is also responsible for affecting a person’s ability to repay the loans taken by them as we can see that Goods prices also affect the number of people able to pay back the loans leased to them.

