Data Pre-processing for credit card approval dataset

General Information about csv files.

- There were two csv files .
- **Credit_card.csv** had all the information about the customers like

  Features name: (Credit_Card.csv)

  Ind_ID: Client ID

  Gender: Gender information

  Car_owner: Having car or not

  Propert_owner: Having property or not

  Children: Count of children

  Annual_income: Annual income

  Type_Income: Income type

  Education: Education level

  Marital_status: Marital_status

  Housing_type: Living style

  Birthday_count: Use backward count from current day (0), -1 means yesterday.

  Employed_days: Start date of employment. Use backward count from current day (0)Positive value means, individual is currently unemployed.

  Mobile_phone: Any mobile phone

  Work_phone: Any work phone

  Phone: Any phone number

  EMAIL_ID: Any email ID

  Type_Occupation: Occupation

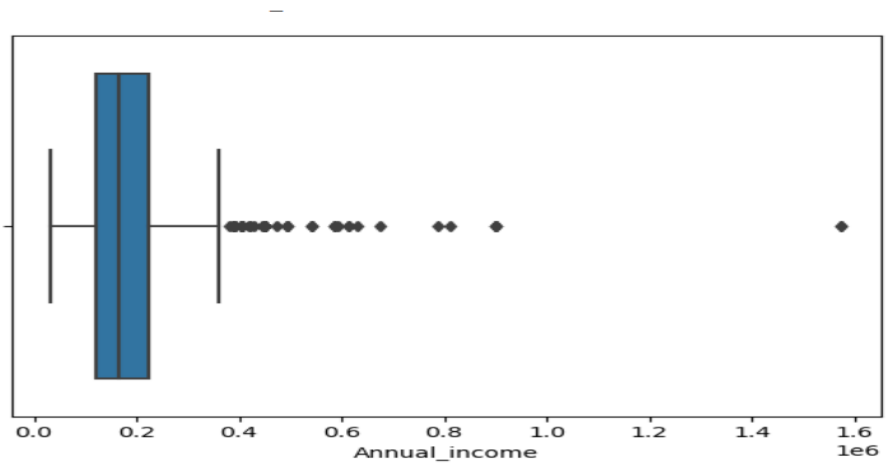
  Family_Members: Family size

- Another data set **(Credit_card_label.csv**) contains two key pieces of information

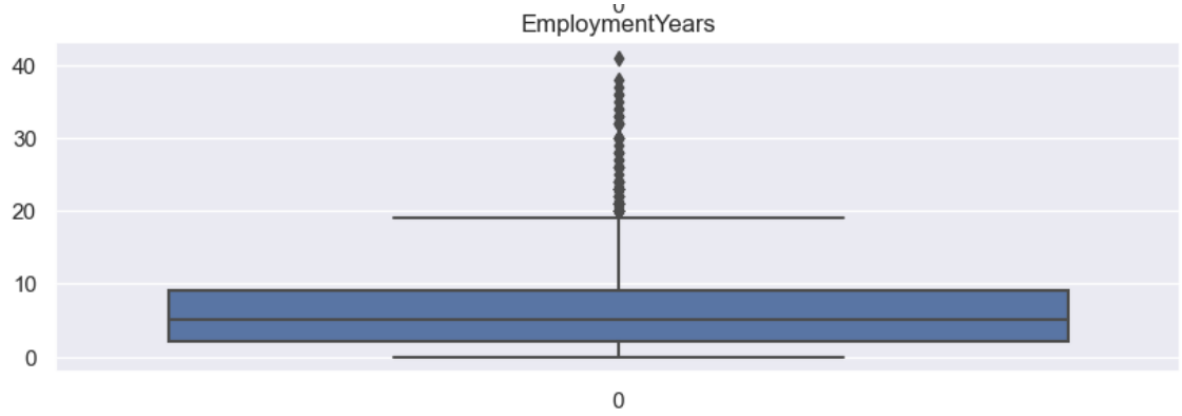
- **ID:** The joining key between application data and credit status data, same is Ind_ID
- **Label:** 0 is application approved and 1 is application rejected.

Steps followed in data preprocessing

- The two datasets have a column that is Id of the customers. We have merged the datasets on the basis of this column.
- We replaced all the positive values in Employed days column with zero
- The Birthday_count and Employed_days columns are in days so we added two new columns representing age and employment years .
- We filled the missing values in age column with mean and converted it into integer type.
- The missing values in Type_occupation column were around 30% so we removed the whole column.
- We filled the missing values in Gender column with mode i.e ‘female’.
- We dropped the id column,Employed_days,Birthday_count columns since we have added two new columns age and EmploymentYears.
- Mobile-phone column has all values as 1 so we dropped that column also.



- The Annual income has 4 percent data as outliers so we removed those data points.



- The employment years column has 5% data as outliers so we removed those data points.
- The final data that we got from cleaning has 1367 rows and 16 column.

Data Transformation Techniques used

- First we split the data into train and test .
- We applied Standard scaler (fit and transform) to numerical train features and one hot encoding (fit and transform) to categorical train features.
- Similarly we applied Standard scaler (only transform) to numerical test features and one hot encoding(only transform) to categorical features.
- Since our target feature was heavily imbalanced so we applied SMOTE() to balance the data.