**EDA ASSINGMENT ON BANK DEFAULTER**

Business Objective:

The driving factors behind the loan defaulter, i.e the variable which are strong indicator of loan default from the data provided we have to find out the useful informations. From the information company will utilize the knowledge for its portfolio and risk management.

THE ASSINGMENT IS DONE IN VARIOUS STEPS:

1. **Data understanding**
2. **Data Cleaning and Manipulation**
3. **Data analysis**
4. **Recommendations**

DATA UNDERSTANDING:

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from google.colab import files

import io

%matplotlib inline

Firstly imported all the library which are going to be used.

After that we are going to upload the two files provided which are

1. Application.csv
2. Previous\_application.csv

Then we read the file one by one.

Step 1

Here application\_data.csv file and previous\_app\_data.csv which have shape as (307511,122) and(1670214,37).

DATA CLEANING AND MANIPILATION

Now there are many columns so firstly I removed the columns which have 40% nan values.

By which our data set become more easy to be analyse.

By removing the columns by 40% we get application\_data 73 column and previous\_app\_data 26 columns.

The application\_data columns:

0 SK\_ID\_CURR 307511 non-null int64

1 TARGET 307511 non-null int64

2 NAME\_CONTRACT\_TYPE 307511 non-null object

3 CODE\_GENDER 307511 non-null object

4 FLAG\_OWN\_CAR 307511 non-null object

5 FLAG\_OWN\_REALTY 307511 non-null object

6 CNT\_CHILDREN 307511 non-null int64

7 AMT\_INCOME\_TOTAL 307511 non-null float64

8 AMT\_CREDIT 307511 non-null float64

9 AMT\_ANNUITY 307499 non-null float64

10 AMT\_GOODS\_PRICE 307233 non-null float64

11 NAME\_TYPE\_SUITE 306219 non-null object

12 NAME\_INCOME\_TYPE 307511 non-null object

13 NAME\_EDUCATION\_TYPE 307511 non-null object

14 NAME\_FAMILY\_STATUS 307511 non-null object

15 NAME\_HOUSING\_TYPE 307511 non-null object

16 REGION\_POPULATION\_RELATIVE 307511 non-null float64

17 DAYS\_BIRTH 307511 non-null int64

18 DAYS\_EMPLOYED 307511 non-null int64

19 DAYS\_REGISTRATION 307511 non-null float64

20 DAYS\_ID\_PUBLISH 307511 non-null int64

21 FLAG\_MOBIL 307511 non-null int64

22 FLAG\_EMP\_PHONE 307511 non-null int64

23 FLAG\_WORK\_PHONE 307511 non-null int64

24 FLAG\_CONT\_MOBILE 307511 non-null int64

25 FLAG\_PHONE 307511 non-null int64

26 FLAG\_EMAIL 307511 non-null int64

27 OCCUPATION\_TYPE 211120 non-null object

28 CNT\_FAM\_MEMBERS 307509 non-null float64

29 REGION\_RATING\_CLIENT 307511 non-null int64

30 REGION\_RATING\_CLIENT\_W\_CITY 307511 non-null int64

31 WEEKDAY\_APPR\_PROCESS\_START 307511 non-null object

32 HOUR\_APPR\_PROCESS\_START 307511 non-null int64

33 REG\_REGION\_NOT\_LIVE\_REGION 307511 non-null int64

34 REG\_REGION\_NOT\_WORK\_REGION 307511 non-null int64

35 LIVE\_REGION\_NOT\_WORK\_REGION 307511 non-null int64

36 REG\_CITY\_NOT\_LIVE\_CITY 307511 non-null int64

37 REG\_CITY\_NOT\_WORK\_CITY 307511 non-null int64

38 LIVE\_CITY\_NOT\_WORK\_CITY 307511 non-null int64

39 ORGANIZATION\_TYPE 307511 non-null object

40 EXT\_SOURCE\_2 306851 non-null float64

41 EXT\_SOURCE\_3 246546 non-null float64

42 OBS\_30\_CNT\_SOCIAL\_CIRCLE 306490 non-null float64

43 DEF\_30\_CNT\_SOCIAL\_CIRCLE 306490 non-null float64

44 OBS\_60\_CNT\_SOCIAL\_CIRCLE 306490 non-null float64

45 DEF\_60\_CNT\_SOCIAL\_CIRCLE 306490 non-null float64

46 DAYS\_LAST\_PHONE\_CHANGE 307510 non-null float64

47 FLAG\_DOCUMENT\_2 307511 non-null int64

48 FLAG\_DOCUMENT\_3 307511 non-null int64

49 FLAG\_DOCUMENT\_4 307511 non-null int64

50 FLAG\_DOCUMENT\_5 307511 non-null int64

51 FLAG\_DOCUMENT\_6 307511 non-null int64

52 FLAG\_DOCUMENT\_7 307511 non-null int64

53 FLAG\_DOCUMENT\_8 307511 non-null int64

54 FLAG\_DOCUMENT\_9 307511 non-null int64

55 FLAG\_DOCUMENT\_10 307511 non-null int64

56 FLAG\_DOCUMENT\_11 307511 non-null int64

57 FLAG\_DOCUMENT\_12 307511 non-null int64

58 FLAG\_DOCUMENT\_13 307511 non-null int64

59 FLAG\_DOCUMENT\_14 307511 non-null int64

60 FLAG\_DOCUMENT\_15 307511 non-null int64

61 FLAG\_DOCUMENT\_16 307511 non-null int64

62 FLAG\_DOCUMENT\_17 307511 non-null int64

63 FLAG\_DOCUMENT\_18 307511 non-null int64

64 FLAG\_DOCUMENT\_19 307511 non-null int64

65 FLAG\_DOCUMENT\_20 307511 non-null int64

66 FLAG\_DOCUMENT\_21 307511 non-null int64

67 AMT\_REQ\_CREDIT\_BUREAU\_HOUR 265992 non-null float64

68 AMT\_REQ\_CREDIT\_BUREAU\_DAY 265992 non-null float64

69 AMT\_REQ\_CREDIT\_BUREAU\_WEEK 265992 non-null float64

70 AMT\_REQ\_CREDIT\_BUREAU\_MON 265992 non-null float64

71 AMT\_REQ\_CREDIT\_BUREAU\_QRT 265992 non-null float64

72 AMT\_REQ\_CREDIT\_BUREAU\_YEAR 265992 non-null float64

And the previous\_app\_data columns are:

0 SK\_ID\_PREV 1670214 non-null int64

1 SK\_ID\_CURR 1670214 non-null int64

2 NAME\_CONTRACT\_TYPE 1670214 non-null object

3 AMT\_ANNUITY 1297979 non-null float64

4 AMT\_APPLICATION 1670214 non-null float64

5 AMT\_CREDIT 1670213 non-null float64

6 AMT\_DOWN\_PAYMENT 774370 non-null float64

7 AMT\_GOODS\_PRICE 1284699 non-null float64

8 WEEKDAY\_APPR\_PROCESS\_START 1670214 non-null object

9 HOUR\_APPR\_PROCESS\_START 1670214 non-null int64

10 FLAG\_LAST\_APPL\_PER\_CONTRACT 1670214 non-null object

11 NFLAG\_LAST\_APPL\_IN\_DAY 1670214 non-null int64

12 RATE\_DOWN\_PAYMENT 774370 non-null float64

13 RATE\_INTEREST\_PRIMARY 5951 non-null float64

14 RATE\_INTEREST\_PRIVILEGED 5951 non-null float64

15 NAME\_CASH\_LOAN\_PURPOSE 1670214 non-null object

16 NAME\_CONTRACT\_STATUS 1670214 non-null object

17 DAYS\_DECISION 1670214 non-null int64

18 NAME\_PAYMENT\_TYPE 1670214 non-null object

19 CODE\_REJECT\_REASON 1670214 non-null object

20 NAME\_TYPE\_SUITE 849809 non-null object

21 NAME\_CLIENT\_TYPE 1670214 non-null object

22 NAME\_GOODS\_CATEGORY 1670214 non-null object

23 NAME\_PORTFOLIO 1670214 non-null object

24 NAME\_PRODUCT\_TYPE 1670214 non-null object

25 CHANNEL\_TYPE 1670214 non-null object

26 SELLERPLACE\_AREA 1670214 non-null int64

27 NAME\_SELLER\_INDUSTRY 1670214 non-null object

28 CNT\_PAYMENT 1297984 non-null float64

29 NAME\_YIELD\_GROUP 1670214 non-null object

30 PRODUCT\_COMBINATION 1669868 non-null object

31 DAYS\_FIRST\_DRAWING 997149 non-null float64

32 DAYS\_FIRST\_DUE 997149 non-null float64

33 DAYS\_LAST\_DUE\_1ST\_VERSION 997149 non-null float64

34 DAYS\_LAST\_DUE 997149 non-null float64

35 DAYS\_TERMINATION 997149 non-null float64

36 NFLAG\_INSURED\_ON\_APPROVAL 997149 non-null float64

In the data we have some error as there are negative values in some of the columns. The negative value columns in the application\_data are as follows:

['DAYS\_BIRTH', 'DAYS\_EMPLOYED', 'DAYS\_REGISTRATION', 'DAYS\_ID\_PUBLISH', 'DAYS\_LAST\_PHONE\_CHANGE']

And the negative value column in previous\_app\_data is ‘DAYS\_DECISION’

So we have to remove the negative value and make all the column positive.

Now the data is ready for analysis.

DATA ANLYSIS

We have two data set:

1. Final\_app\_data
2. Pre\_app\_data

* First we will analyse final\_app\_data
* Then we will analyse pre\_app\_data
* Then we will merge the two data and then analyse the merged data.

So from final\_app\_data:

* We have analysed different columns and take out mean and median of them.
* Then with the help of bar graph and histogram we see the outliners and the get the facts relate to it.
* Some of the important columns of final\_app\_data are ATM\_CREDIT,AMT\_ANNUITY,AMT\_GOODS\_PRICE,AMT\_REQ\_CREDIT\_BUREAU\_YEAR,SK\_ID\_CURR,TARGET,DAYS\_EMPLOYED,NAME\_INCOME\_TYPE,NAME-FAMILY\_STATUS,OCCUPATION\_TYPE etc.

So now from pre\_app\_data:

* We have analysed different columns of pre\_app\_data through mean and median.
* Then with the help of different graphs we have seen outliners.
* Some important columns of pre\_app\_data are

AMT\_ANNUITY,AMT\_APPLICATION,AMT\_CREDIT,AMT\_GOODS\_PRICE,CNT\_PAYMENT,SK\_ID\_PREV,SK\_ID\_CURR,NAME\_CASH\_LOAN\_PURPOSE,NAME\_PORTFOLIO etc.

After that merged data:

* From the merged data univariate and bivariate analysis is done.
* From this we get the relation between previous loan default condition.
* The important columns of merged data is AMT\_ANNUITY,AMT\_GOODS\_PRICE,SK\_ID\_CURR,AMT\_CREDIT,NAME\_CONTRACT\_TYPE,WEEKAY\_APPR\_PROCESS\_START,HOUR\_APPR\_PROCESS\_START etc.

RECOMMENDATION:

* From the data imbalance we get that only 8.1% on the previous data have defaulted and rest 91.9% have not defaulted.
* The data was imbalance on higher level.
* With the help of univariate and bivariate analysis we see that there are few important columns which give us clear idea about the data such as NAME\_INCOME\_TYPE,AMT\_ANNUITY,TARGET,SK\_ID\_CURR etc.
* As there was 121,73 columns (~40%) of the columns have missing values.
* We have analysed the outliers and handled it to get out relative required information.
* The data consists of numeric and categorical columns through which we get information.