CREDIT EDA ASSIGNMENT

Importing and Reading Datafiles

- Imported libraries
- Imported warnings
- Read Datafile - Reading 'application_data' in ‘ad’ variable

Checkpoints for basic – ‘ad’ Data frame

- Checked ad data frame
- Shape- (307511, 122)
- Gone through columns
- Checked null values
- Checked categorical and numerical columns

Data Cleaning

Null values

- Findings -

It seems there are lots of missing values in first 50 columns in this order above.

- Solution-

With the given missing values i will drop the columns having larger values than 50% of length of the data frame in spite of imputing. It may good idea to drop them as it will save the large data surfing, cleaning of that columns.

- Action-

Dropped the columns having missing values larger than 50% of length of the data frame

Null values

- **Result after dropping–**

Shape of ad – (307511, 81) , 81 columns present In new data frame

Treating missing values

Missing Values Imputation
Numerical Columns

Columns having missing values close to 50% -

- If missing values of columns are normally distributed then it can be imputed by 'mean’
- If skewed normal distribution will impute them by ‘median’.

Missing Values Imputation
Numerical Columns

- Column-FLOORSMAX_AVG

  Missing Values- 49.76%

  Distplot of normal distribution

  Imputed missing values by 'mean’

- Column- FLOORSMAX_MODE

  Missing Values- 49.76%

  Distplot of normal distribution

  Imputed missing values by 'mean’

Missing Values Imputation
Numerical Columns

- Column -FLOORSMAX_MEDI

  Missing Values- 49.76%

  Distplot of normal distribution

  Imputed missing values by 'mean’

- Column - YEARS_BEGINEXPLUATATION_AVG

  Missing Values- 49.78%

  Distplot of normal distribution

  Imputed missing values by 'mean’

Missing Values Imputation
Numerical Columns

- Column -YEARS_BEGINEXPLUATATION_MODE

  Missing Values- 49.78%

  Distplot of normal distribution

  Imputed missing values by 'mean’

- Column - YEARS_BEGINEXPLUATATION_MEDI

  Missing Values- 49.78%

  Distplot of normal distribution

  Imputed missing values by 'mean’

Missing Values Imputation
Numerical Columns

- Column - TOTALAREA_MODE

  Missing Values- 48.26%

  Distplot of skewed normal distribution

  Imputed missing values by ‘median’

- Column - EMERGENCYSTATE_MODE

Looking at column description we are not getting to much insights from it. Hence we have dropped this column.

Missing Values Imputation
Numerical Columns

- Column - EXT_SOURCE_3

  Missing Values- 48.26%

  Distplot of skewed normal distribution

  Imputed missing values by ‘median’

Missing Values Imputation
Categorical Columns

- We are imputing categorical columns by mode value otherwise keep it blank if it affects the analysis.

Missing Values Imputation
Categorical Columns

- **Column - OCCUPATION_TYPE**

Missing values – 31%

Plotted count plot to get mode value – Labour

Replacing occupation of any person is not good enough hence we will keep it as 'Unknown'

- **Column - AMT_REQ_CREDIT_BUREAU_QRT**

Missing values – 13.50%

Mode value – 0.0

Imputed missing values by 'mode'

Missing Values Imputation
Categorical Columns

- Column - AMT_REQ_CREDIT_BUREAU_HOUR

Missing values – 13.50%

Mode value – 0.0

Imputed missing values by ‘mode’

- Column - AMT_REQ_CREDIT_BUREAU_DAY

Missing values – 13.50%

Mode value – 0.0

Imputed missing values by ‘mode’

Missing Values Imputation
Categorical Columns

- Column - AMT_REQ_CREDIT_BUREAU_WEEK

Missing values – 13.50%

Mode value – 0.0

Imputed missing values by ‘mode’

- Column - AMT_REQ_CREDIT_BUREAU_MON

Missing values – 13.50%

Mode value – 0.0

Imputed missing values by ‘mode’

Missing Values Imputation
Categorical Columns

- **Column - AMT_REQ_CREDIT_BUREAU_YEAR**

Missing values – 13.50%

Mode value – 0.0

Imputed missing values by ‘mode’

- **Column- NAME_TYPE_SUITE**

Missing values – 0.42%

Mode value – Unaccompanied (Observed through count plot)

Imputed missing values by ‘mode’

Missing Values Imputation
Categorical Columns

- Column- OBS_30_CNT_SOCIAL_CIRCLE

Missing values –%

Mode value – 0.0

Imputed missing values by ‘mode’

- Column - DEF_30_CNT_SOCIAL_CIRCLE

Missing values –%

Mode value – 0.0

Imputed missing values by ‘mode’

Missing Values Imputation
Categorical Columns

- Column- OBS_60_CNT_SOCIAL_CIRCLE

Missing values –%

Mode value – 0.0

Imputed missing values by ‘mode’

- Column - DEF_60_CNT_SOCIAL_CIRCLE

Missing values –%

Mode value – 0.0

Imputed missing values by ‘mode’

Missing Values Imputation
Categorical Columns

- **Format - > Columns – Missing value**

  CNT_FAM_MEMBERS – 0.0006%

  DAYS_LAST_PHONE_CHANGE – 0.0003%

  EXT_SOURCE_2 – 0.214%

  AMT_GOODS_PRICE – 0.09%

  AMT_ANNUITY – 0.0039%

- **Solution:**

  All these columns has missing value less that 1% hence we will keep them as it is.

Imbalance

- TARGET Column -

- Calculated the ratio between count of peoples having difficulties vs count of peoples who paid easily

- Finding -

Imbalance ratio says us that out of every 11 peoples, 1 person is having payment difficulty.

Analysis of data

Data frame ad is divided into two parts based on

- TARGET column having 0- Peoples have paid loan easily as ad_0

- TARGET column having 1-Peoples having difficulties while paid loan as ad_1

Analysis of Categorical Columns

Analysis of Categorical Columns

we are using 3 different plots for analysis:

- Pie plot: For plotting the all the values present in a column in terms of percentage. So, the sum of those data types will be 100.

- count plot 1: Here, plotted the count of the different categories. So, Target=0 will have higher count than Target=1.

- count plot 2: To plot this dataset, we have first divided the dataset into 2 subsets, Target=0 and Target=1. Then again divided the individual Target=0 and Target=1 into different categories. Then, plotted these categories in terms of percentage. So, we can find that the values for Target=0 and Target=1 are mostly equal.

Univariate Analysis of the Categorical Columns Findings

NAME_CONTRACT_TYPE

- **Pie plot**- it gives us better understanding about peoples are likely to take maximum Cash loans(90%) rather than Revolving loans(10%).
- **count plot** - Both the count plot says that peoples having maximum difficulties in paying Cash loans than revolving loans. Around 2,50,000 peoples having payment difficulties for cash loan out of that around 2500 peoples having paid them easily. Also around 3000 peoples having difficulties during paying revolving loans rather around 700 are more likely to pay it easily.
- **count plot in %** Similarly in terms of percentage- 90% peoples failed to pay cash loans while 80% peoples paid it easily. Also in case of Revolving loans around 20% peoples having payment difficulties and 10% peoples paid them easily.

CODE_GENDER

- **Pie plot**- it shows that maximum females(66%)are going to take loans in spite of males(34%).
- **count plot** - count plot says that 2,00,000 females having payment difficulties than males(90000). Out of total males and females 20000 females and 15000 males paid it easily
- **count plot in %** - We can clearly see that 65% females have bounce there EMIs and 55% paid them easily 35% males having difficulty and 45% males paid them easily.

FLAG_OWN_CAR

- **Pie plot** - 34% peoples had own car who had taken loans. and other than that 66 % are haven't any car but taken a loan.
- **Count plot** -Here we can see 29% peoples having payment difficulties who owns car and 69% haven't any car still they have difficulties to pay EMI. Also 65% peoples haven't any car and paid smooth Emi whereas 35 % peoples own a car and paid Emi easily.
- So we can conclude that peoples who owned car having less difficulties while paying loans rather than who have it.

FLAG_OWN_REALTY
if client owns a house or flat

- 69% owns a flat/house
- 31% have not
- Who own house - 69% having had paid easily and 31% had difficulties
- Who haven't any house - 61% having difficulties in payment and 32% had paid it easily.
- So we can conclude that peoples who had own homes are less difficulties rather than who haven't it.

NAME_TYPE_SUITE

- 81% peoples have no companion
- 13% are living with family, 4% are living with there spouse and remaining 2 % are not considerable as divided into 2-3 groups.

- From count plot we can say that - single crowd around 2,30,000 are more likely to pay EMI easily than other groups. as may the have less responsibilities we can say.

- - From this we can catch a insight that we are more likely to prefer single crowd to give a loan than other crowd.

NAME_INCOME_TYPE

- We can see from pie plot that crowd of loan taken is divided as follows:

  - 52% crowd is working,

  - 23% commercial associates

  - 18% pensioner

  - 7% state servant

  - remaining crowd goes close to zero that are - Students, Unemployed, Businessman, Maternity leave category.

- Also we can see from count plot that even if maximum loan taken by the working crowd still this crowd has maximum difficulties having paying EMI's. that is 60%
- Also this category has huge crowd of paying EMIs that is 50%

- Remaining crowd from commercial associated pensioner and state servant had maximum percentage of paying EMIs easily than having difficulties. So we can take these crowd in consideration to give a loan.

NAME_EDUCATION_TYPE

- Maximum peoples from secondary education category % have taken loans.
- otherwise 24 % have Higher education and all others in remaining category
- From count plot we can observe that the maximum crowd who likely to take EMI also more likely t repay loans easily. that is crowd from secondary education category (around 2,00,000)
- Also Higher education category peoples also have maximum percentage of repaying loans easily that are 70000 than having difficulties.

- So we can take this secondary and higher secondary education type crowdinto consideration while giving loans.

NAME_FAMILY_STATUS

- 64% married peoples have taken a loan and 15% are single, remaining are in other categories

- Married peoples have maximum loan taken as well as paid it without difficulties than other categories (67% married peoples)

NAME_HOUSING_TYPE

- Almost 89% crowd having house/ apartment had taken a loan maximum than other categories

- Also out of that 85% peoples have paid loan without any difficulties.

- Conclusion - We can take this crowd in consideration while giving loans rather than other categories.

Top 10 correlations for Selectors

FLAG_EMP_PHONE	DAYS_EMPLOYED	0.999758
DAYS_EMPLOYED	FLAG_EMP_PHONE	0.999758
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998510
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998510
FLOORSMAX_MEDI	FLOORSMAX_AVG	0.997019
FLOORSMAX_AVG	FLOORSMAX_MEDI	0.997019
YEARS_BEGINEXPLUATATION_AVG	YEARS_BEGINEXPLUATATION_MEDI	0.993582
YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.993582
FLOORSMAX_MODE	FLOORSMAX_MEDI	0.988153
FLOORSMAX_MEDI	FLOORSMAX_MODE	0.988153

Top 10 correlations for defaulters

DAYS_EMPLOYED	FLAG_EMP_PHONE	0.999702
FLAG_EMP_PHONE	DAYS_EMPLOYED	0.999702
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998270
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998270
FLOORSMAX_MEDI	FLOORSMAX_AVG	0.997233
FLOORSMAX_AVG	FLOORSMAX_MEDI	0.997233
YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.996125
YEARS_BEGINEXPLUATATION_AVG	YEARS_BEGINEXPLUATATION_MEDI	0.996125
FLOORSMAX_MODE	FLOORSMAX_MEDI	0.989370
FLOORSMAX_MEDI	FLOORSMAX_MODE	0.989370

Analysis for outliers
Using Scatter plot

- Following columns had single outliers present 1.OBS_30_CNT_SOCIAL_CIRCLE
  2.DEF_60_CNT_SOCIAL_CIRCLE
  3.OBS_60_CNT_SOCIAL_CIRCLE
  4.DEF_30_CNT_SOCIAL_CIRCLE
  5.AMT_REQ_CREDIT_BUREAU_QRT
- Following has some outliers are at top
  1.FLAG_DOCUMENT_2
  2.FLAG_DOCUMENT_4
  3.FLAG_DOCUMENT_7
  4.FLAG_DOCUMENT_2
  5.FLAG_DOCUMENT_10
  6.FLAG_DOCUMENT_12
- All other plots not giving any insights.

Univariate Analysis for numerical data

- TARGET - more than 275000 crowd has no difficulties while doing repayments of Emi
- CNT_CHILDREN - around 250000 peoples have children count between 0 to 2
- Most of peoples have income below 1 lakh and only few peoples are getting income in mores of lakhs.
- Amt Annuty - has skewed normal distribution. 1,60000 amount is getting paid at regular intervals.
- 3 lakhs of peoples have provided their contact number and similar count for work phone number also.
- Emails provided around 2,50000 peopls.
- Region Rating of client is good for maximum 2.5 lakh peoples.
- Also for address of living and registration is not same we know by REG_REGION_NOT_WORK_REGION
- Flag documents shows that the given list of documents provided by peoples

- Read another Data frame Previous_application in ‘pa’ variable
- Found duplicates in SK_ID_PREV
- Merged ‘ad’ and ‘pa’ Data frames in previous_train variable
- Shape of previous_train - (1413701, 116)
- Divided previous_train data frame according TARGET 0 and 1 as pa_0 and pa_1

Bivariate Analysis

- We used to plot 4 subplots
- **Pie plot**: For plotting the all the values present in a column in terms of percentage. So, the sum of those data types will be 100.

- **Count plot 1**: Here, plotted the count of the different categories. So, we can analyze count of Target=0 and Target=1.

- **Count plot 2**: To plot this dataset, we have first divided the dataset into 2 subsets, Target=0 and Target=1. Then again divided the individual Target=0 and Target=1 into different categories. Then, plotted these categories in terms of percentage. Also selected a hue as 'NAME_CONTRACT_STATUS' so, we can find that the values for Target=0 and Target=1

NAME_EDUCATION_TYPE

- From pie plot we can see that we had maximum crowd that is 71% from secondary education category. and 21% from higher secondary education had taken loans.
- from count plot of second subplot we get that the whatever the loans taken by secondary education category had also aced the difficulties while repaying the loan that is 75% and 70% repay it easily.
- From 3rd subplot of TARGET 0 (Who had paid Emis easily) says that maximum that is 575000 people got approved loans from secondary education category and around 1.9Laks peoples from higher secondary also got approved for the loans.
- 4th subplot for TARGET=1 (Who had difficulties for repaying loans) had crowd of 55000 who failed to repay from secondary education category. Also 24000 peoples from same category got refused the loans.
- **From this analysis we can look forward to the secondary education category type peoples most with higher secondary also.**

CODE_GENDER

- From pie plot we already got to know that 66% females and 34% males applied for the loans.
- Out of all, 68% females repay it easily and 55% had difficulties while repaying the same and 32% males repay it easily and 45% had difficulties
- Now from Target 0 subplot we got to know that around 550000 females and 250000 males got loan approval and alsowho repaid the loan easily.
- And around 38000 females and 27000 males got approved loan but had a difficulties while repaying the same.
- **We can look forward to female category for providing loans.**

NAME_INCOME_TYPE

- From pie plot we got that 52% peoples from working category with 23% commercial associates and 18% pensioner had taken loans.
- From count plot of TARGET we get that 61% peoples from working category 21% from commercial associates and remaining others got difficulties to repay the loan.
- From TARGET 0 count plot 400000 peoples got approved as well as repaid loan easily. Also we look forward to commercial associates(175000) and pensioner(155000) got approved and repaid loans easily.
- TARGET 1 count plot says that 40000 working crowd, 14000 commercial associates, and around 8000 pensioner got approved but had difficulties having repaying loans.
- **We can look forward to working, commercial associates and pensioner category for loan distribution.**

NAME_FAMILY_STATUS

- From pie plot we got to know that 64% married peoples have taken a loan and 15% are single, remaining are in other categories.
- From count plot of TARGET we got that Married peoples (65%) have maximum loan taken as well as paid it without difficulties than other categories (67% married peoples)
- From TARGET 0 count plot we found that 520000 married crowd got loans approved as well as repaid easily.
- Also TARGET 1 count plot says that around 40000 married peoples repaid with difficulties of payments.
- **From this we can look forward to married category**

Thank You