CREDIT EDA ASSIGNMENT

By Bonani Pramanik

Problem Stated

➢ We are required to find out those clients which help in gaining business profit after lending loan.

➢ Conditions:-

➢ If the applicant is facing any problem in repaying Loan amount then it will not a wise decision to lend money to them as this may lead to Business Loss

➢ If the applicant is not facing any problem in repaying Loan amount or any minor issue then it can be considered a good decision to lend money to them as this may not lead to Business Loss.

Steps Followed

- ▶ After going through the data set provided ie, Application Data & Previous Data and the use of Python Libraries.

- ▶ Identified the columns which can be useful in analysing the data in the better way.

- ▶ With the help of 'describe' function get to know about the mean, minimum, maximum and standard deviation values.

Data Cleaning

▶ Calculate the Null Values present in the different columns and the percentage of Null values present in the Data Set.

▶ Then, identified those % of Null Values which are more than 50 %.

▶ 122 columns were there before cleaning the Data.

▶ After the dropping of columns 81 columns were remaining.

Outliers Identifying & Imputing Values

▶ After the dropping of Columns , We have identified the Outliers and tried to impute them.

▶ Columns imputed:

FLOORSMAX_AVG
FLOORSMAX_MODE
FLOORSMAX_MEDI
YEARS_BEGINEXPLUATATION_AVG
YEARS_BEGINEXPLUATATION_MODE
YEARS_BEGINEXPLUATATION_MEDI
TOTALAREA_MODE
EMERGENCYSTATE_MODE
OCCUPATION_TYPE
EXT_SOURCE_3
AMT_REQ_CREDIT_BUREAU_YEAR

AMT_REQ_CREDIT_BUREAU_HOUR
AMT_REQ_CREDIT_BUREAU_DAY
AMT_REQ_CREDIT_BUREAU_WEEK
AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_QRT
NAME_TYPE_SUITE
DEF_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
OBS_30_CNT_SOCIAL_CIRCLE

| |
|---|
| EXT_SOURCE_2 |
| AMT_GOODS_PRICE |
| AMT_ANNUITY |
| CNT_FAM_MEMBERS |
| DAYS_LAST_PHONE_CHANGE |

Dropped Non Required Data

```
'FLAG_MOBIL', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE','FLAG_PHONE', 'FLAG_EMAIL',
'REGION_RATING_CLIENT','REGION_RATING_CLIENT_W_CITY','FLAG_EMAIL','CNT_FAM_MEMBERS', 'REGION_RATING_CLIENT',
'REGION_RATING_CLIENT_W_CITY','FLAG_DOCUMENT_2', 'FLAG_DOCUMENT_3','FLAG_DOCUMENT_4',
'FLAG_DOCUMENT_5', 'FLAG_DOCUMENT_6','FLAG_DOCUMENT_7', 'FLAG_DOCUMENT_8', 'FLAG_DOCUMENT_9','FLAG_DOCUMENT_10',
'FLAG_DOCUMENT_11','FLAG_DOCUMENT_12','FLAG_DOCUMENT_13', 'FLAG_DOCUMENT_14', 'FLAG_DOCUMENT_15',
'FLAG_DOCUMENT_16', 'FLAG_DOCUMENT_17', 'FLAG_DOCUMENT_18','FLAG_DOCUMENT_19', 'FLAG_DOCUMENT_20',
'FLAG_DOCUMENT_21','EXT_SOURCE_2','EXT_SOURCE_3']
```

Analysis of Application Data
Target Variable

▶ **Analysed the Issues coming from the Client's side ie , Repaying Issues, this shows the High range of Defaulters are with other reasons and few people are in the category of High Defaulter.**


Repay Issues
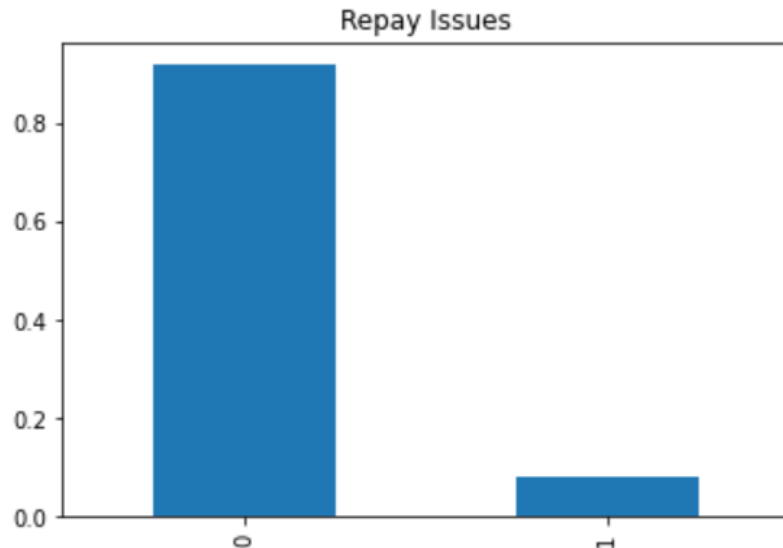
Application Process

▶ Analysis of the Application process we can conclude that the Application process were having low density on Saturdays & Sundays.- Target 0 & 1
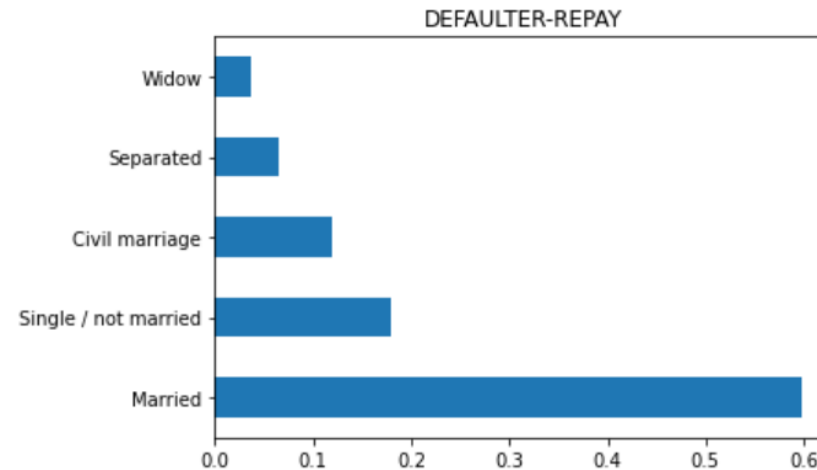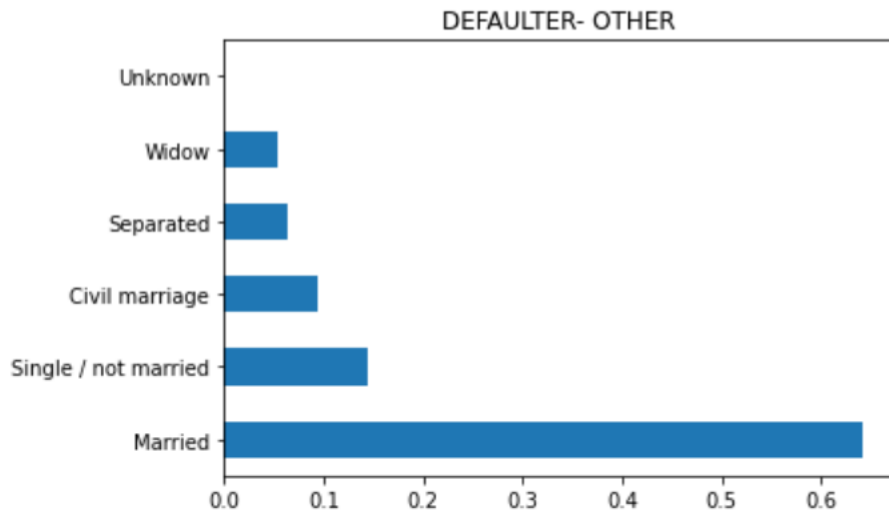

DEFAULTER- OTHER

DEFAULTER-REPAY
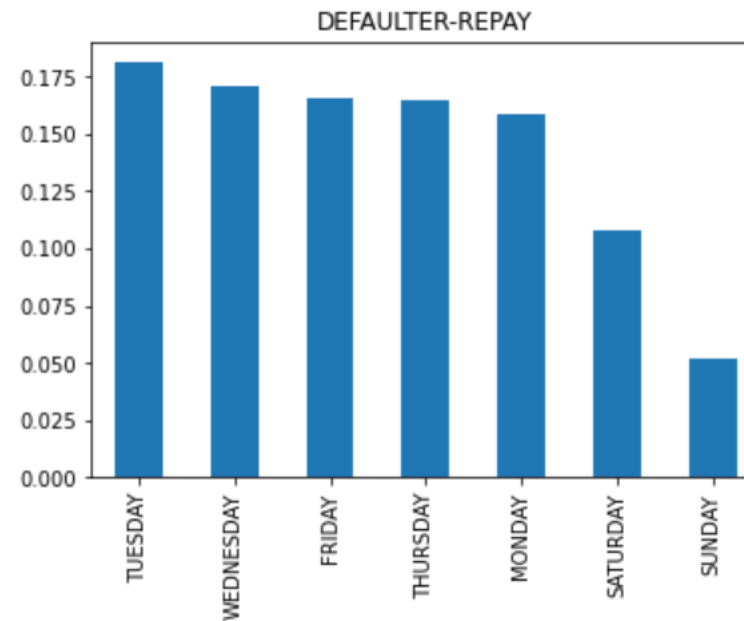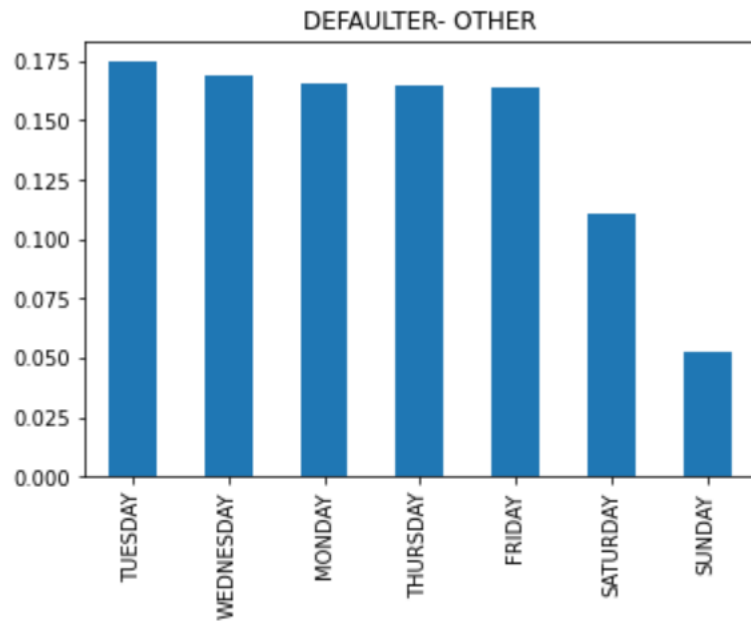
Education Type

▶ we can conclude that secondary/special educated people are applying loans high in number and Academic degree educated people are applying loan in least count for both target= 0 and 1.
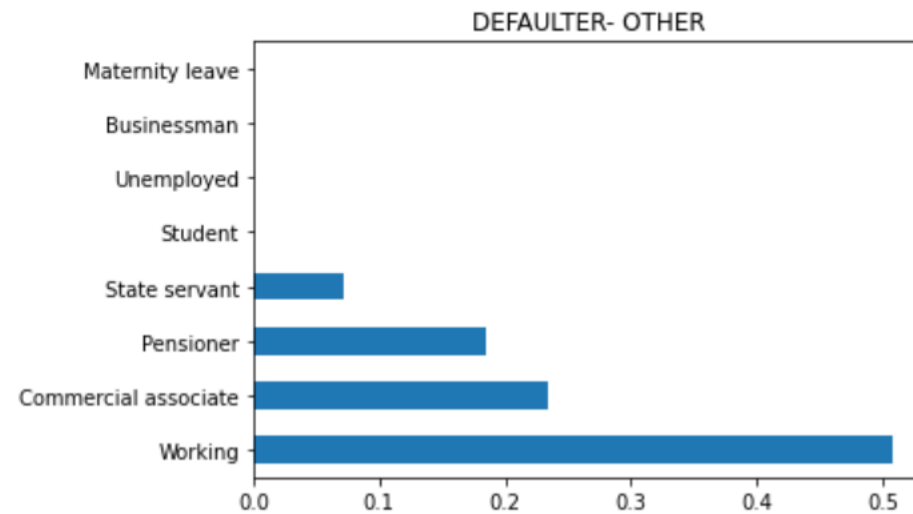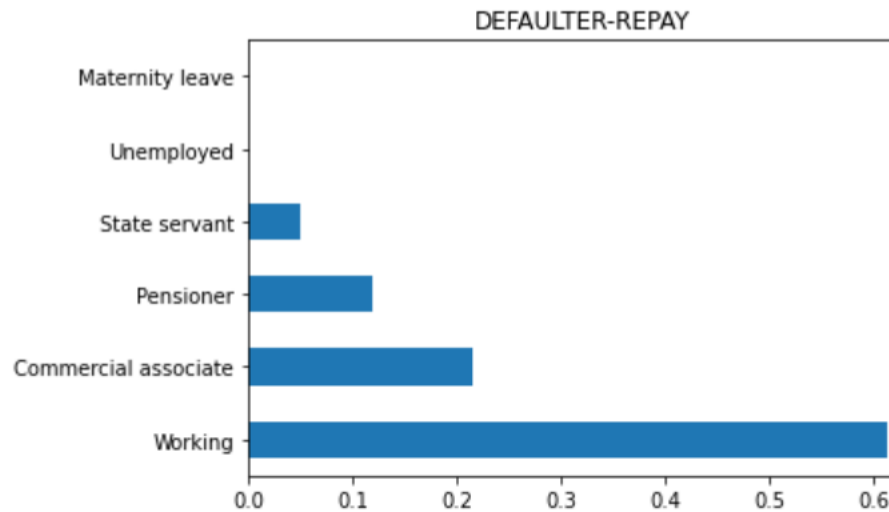

DEFAULTER- OTHER

DEFAULTER-REPAY
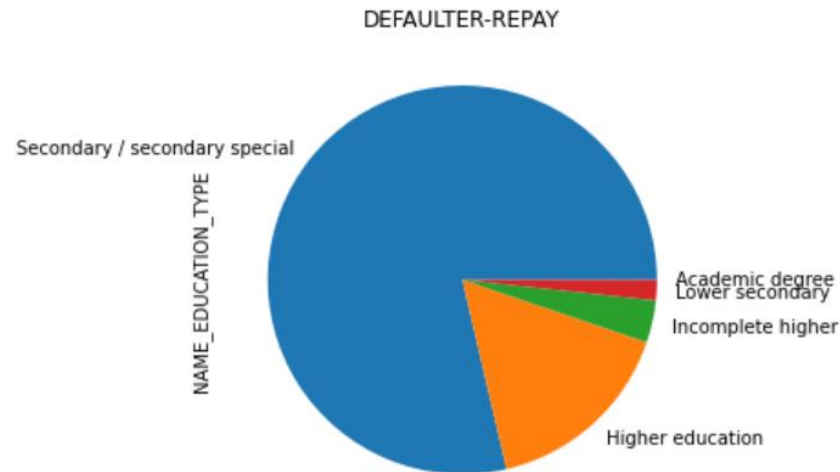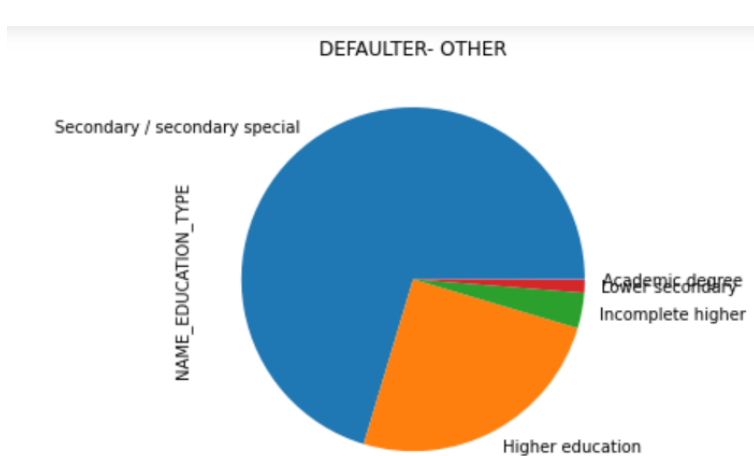
Family Status Analysis

▶ **We can say more married people tend to take more Loan as compared to other categories as compared to the others.**

Income Category

▶ It shows that majority of defaulters income type is working and at the same time there is good income to bank from working people.

Numerical Columns Observations

▶ There exists more clients who changed their their registration details after 4000 days of approval of loan.

▶ For few not default clients, time taken to publish id's are higher than default clients.

▶ The application process start hours taken for default and not default cases are similar.

▶ In non default cases, people keep their phone numbers for greater time.

▶ People with greater number of days born count are less likely to default.

▶ In non default case AMT_GOODS PRICE contains more outliers than default case.

▶ In default case, most of the clients amount annuity tends to be greater than 25000(median value).

▶ Whose credit amount is greater than 50000 tends to be less default than compared to default cases and vice versa.

▶ people with higher no of employment days are less likely to default.

▶ Majority of defaulting people are having less total income.

Correlation Observations – Target Variable

- When defaulted, major portion of decision is taken by-
- (YEARS_BEGINEXPLUATATION_MEDI AND YEARS_BEGINEXPLUATATION_AVG)
- (OBS_60_CNT_SOCIAL_CIRCLE AND OBS_30_CNT_SOCIAL_CIRCLE)
- (FLOORSMAX_MEDI AND FLOORSMAX_AVG)

- When not defaulted, major portion of decision is taken by-
- (FLOORSMAX_MEDI AND FLOORSMAX_AVG)
- (OBS_60_CNT_SOCIAL_CIRCLE AND OBS_30_CNT_SOCIAL_CIRCLE)
- (FLOORSMAX_MEDI AND FLOORSMAX_MODE)

- The values - YEARS_BEGINEXPLUATATION_MEDI and YEARS_BEGINEXPLUATATION_AVG are more correlated in default case than non-default

Analysis After merging Previous Data- Numerical Analysis

- High number of applications are filed in 9 AM to 2 PM for both Current and Previous data.

- So busiest hours for bank are form 9 AM to 2 PM.

- Nuclear family tends to take more loans.

- Previously bank had high unused offers but currently refused is high incase of AMT_GOODS_PRICE.

- Previously bank had high unused offers and currently cancelled/refused offers are similar for AMT_ANNUITY.

- Previously bank had high unused offers and currently high number of refused offers for AMT_CREDIT.

Numerical Analysis

- AMT_CREDIT_Previous has highest refused cases and AMT_CREDIT_Current is similar for all 4 cases.

- time spent in unused offer is higher as compared to other categories.

- So bank should reduce time spent on unused offer.

- nuclear family(2-3 people in family) get highest approval.

- Previously most of the applications were cancelled or refused

- but now Refused/Cancelled/Approved/Unused all four have similar situation for AMT_GOODS_PRICE.

- Previously most of the applications were cancelled or refused

- but now Refused/Cancelled/Approved/Unused all four have similar situation for AMT_ANNUITY.

CONCLUSION

- Target/focused variable for Application dataset - TARGET
- Target/focused variable for Previous dataset - NAME_CONTRACT_STATUS

- Variables for Loan Prediction
- NAME_EDUCATION_TYPE
- AMT_INCOME_TOTAL
- DAYS_BIRTH
- AMT_CREDIT
- DAYS_EMPLOYED
- AMT_ANNUITY
- NAME_INCOME_TYPE
- CODE_GENDER
- NAME_HOUSING_TYPE
- The above mentioned variables are to be considered before approving application to minimize risk of loss.

THANKYOU