Exploratory Data Analysis

Assignment 1

By

Kiran Yashawant Awari

Data reading - ‘application_data.csv’

- Imported warnings
- Imported important libraries – Numpy, Pandas, Matplotlib, Seaborn etc.
- Read CSV file ‘application_data.csv’ for first analysis as df
- Understood shape of file by using df.shape

Data Cleansing

- Identified missing values in columns using function isnull()
- Removed columns with more than 40 % missing values
- For column ‘OCCUPATION_TYPE’ replaced all missing values with mode
- Replaced missing values in column ‘EXT_SOURCE_3’ with median value
- Replaced missing values in right skewed 6 columns with respective median values
- Replaced missing values in columns 'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'EXT_SOURCE_2’ with median value
- Replaced missing values in column ‘NAME_TYPE_SUITE’ with mode

Univariate Analysis on Categorical Columns

1. application for cash loans is more than that of revolving loan
2. number of female applicants are more than that of male applicants
3. most of applicants do not own a car
4. maximum number of applicants own a real estate property
5. number of Unaccompanied applicants is more
6. maximum number of working professionals applied for loan
7. Most of applicants have secondary education
8. marital status of maximum number of applicants is married
9. Housing type of maximum applicants is House/ Apartment
10. Most of applicants are laborers
11. Applications are carried on all week days where on weekends applications are less as compared to other weekdays

Correlation Analysis at Target = 0

There is a strong correlation between the following TOP 10 numeric variables for TARGET=0:

- DAYS_EMPLOYED & FLAG_EMP_PHONE
- OBS_60_CNT_SOCIAL_CIRCLE & OBS_30_CNT_SOCIAL_CIRCLE
- FLOORSMAX_AVG & FLOORSMAX_MEDI
- YEARS_BEGINEXPLUATATION_AVG & YEARS_BEGINEXPLUATATION_MEDI
- FLOORSMAX_MEDI & FLOORSMAX_MODE
- AMT_GOODS_PRICE & AMT_CREDIT
- FLOORSMAX_MODE & FLOORSMAX_AVG
- YEARS_BEGINEXPLUATATION_MODE & YEARS_BEGINEXPLUATATION_AVG
- YEARS_BEGINEXPLUATATION_MEDI & YEARS_BEGINEXPLUATATION_MODE
- REGION_RATING_CLIENT_W_CITY& REGION_RATING_CLIENT

Correlation Analysis at Target = 1

There is a strong correlation between the following TOP 10 numeric variables for TARGET=1

1 FLAG_EMP_PHONE & EMP_YRS

2 DAYS_EMPLOYED & FLAG_EMP_PHONE

3 OBS_30_CNT_SOCIAL_CIRCLE & OBS_60_CNT_SOCIAL_CIRCLE

4 FLOORSMAX_AVG & FLOORSMAX_MEDI

5 YEARS_BEGINEXPLUATATION_AVG & YEARS_BEGINEXPLUATATION_MEDI

6 FLOORSMAX_MEDI & FLOORSMAX_MODE

7 FLOORSMAX_AVG & FLOORSMAX_MODE

8 AMT_GOODS_PRICE & AMT_CREDIT

9 YEARS_BEGINEXPLUATATION_AVG & YEARS_BEGINEXPLUATATION_MODE

10 YEARS_BEGINEXPLUATATION_MEDI & YEARS_BEGINEXPLUATATION_MODE

Reading CSV file ‘previous_application.csv’

- Read csv file ‘previous_application.csv’ as inp0
- Understood data by using functions describe() and info()

Data cleansing of ‘previous_application.csv’

- Identified missing values by using isnull() function
- Removed columns having missing values more than 40 %
- For columns AMT_ANNUITY, AMT_GOODS_PRICE & CNT_PAYMENT replaced missing values with median values
- Replaced missing values in column AMT_CREDIT with median value
- Replaced missing values of PRODUCT_COMBINATION with Cash as it is mode of the column

Univariate analysis on previous_application.csv

- Most number of loans were Cash Loans
- Maximum number of applications are on Monday, Tuesday and Wednesday
- Most of applications had one application per loan contract
- Most had "XAP" purpose of loan application (Analysis cannot be performed)
- Maximum number of applications were approved
- Most of loans were repaid by "Cash through the Bank"
- Maximum number of loans were rejected because of reason "XAP" (Analysis cannot be performed)
- Most previous loans were applied by people who were unaccompanied
- Most applicants were applied by Repeater clients
- Maximum number of XNA goods observed (Analysis cannot be performed)
- Most of them were for POS
- Most were product type "XNA" (not applicable) (Analysis cannot be performed)
- Most applications had clients acquired from Credit and cash offices
- Most were from "XNA" (not applicable) seller industry (Analysis cannot be performed)
- Most of applications fell under "XNA" (not applicable) interest group (Analysis cannot be performed)
- Most applications had the Product Combination as "Cash"

Merging 2 dataframes

- Merged application_data.csv and previous_application.csv by pd.merged function