Team 16

Advanced analysis using statistics module end exam

Name: Sumit Bansod Roll no :220940325081
Name: Gauri Pandey Roll no :220940325033
Name: Keshav Yawale Roll no :220940325036
Name: Dipti Patil Roll no :220940325028
Name: Shubham Mane Roll no :220940325073

Case study:-

Problem Statement:- This analysis is done to identify the patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

- ***Information about the data***

- ☐ The given Application data set has 307511 rows and 122 columns
- ☐ The given previous application data set has 1670214 rows and 37 columns
- ☐ Total null values in application dataset is 9152465
- ☐ Total null values in previous dataset is 11109336
- ☐ Information about the data type, non-null values and memory usage in application dataset

#	Column	Dtype
0	SK_ID_CURR	int64
1	TARGET	int64
2	NAME_CONTRACT_TYPE	object
3	CODE_GENDER	object
4	FLAG_OWN_CAR	object
5	FLAG_OWN_REALTY	object
6	CNT_CHILDREN	int64
7	AMT_INCOME_TOTAL	float64
8	AMT_CREDIT	float64
9	AMT_ANNUITY	float64
10	AMT_GOODS_PRICE	float64
11	NAME_TYPE_SUITE	object
12	NAME_INCOME_TYPE	object
13	NAME_EDUCATION_TYPE	object

Summarised description of application data set with central tendency of the
variable, their dispersion, the presence of empty values and their shape

	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATI
count	307511.000000	307511.000000	307511.000000	3.075110e+05	3.075110e+05	307499.000000	3.072330e+05	307511.0000
mean	278180.518577	0.080729	0.417052	1.687979e+05	5.990260e+05	27108.573909	5.383962e+05	0.0208
std	102790.175348	0.272419	0.722121	2.371231e+05	4.024908e+05	14493.737315	3.694465e+05	0.0138
min	100002.000000	0.000000	0.000000	2.565000e+04	4.500000e+04	1615.500000	4.050000e+04	0.0002
25%	189145.500000	0.000000	0.000000	1.125000e+05	2.700000e+05	16524.000000	2.385000e+05	0.0100
50%	278202.000000	0.000000	0.000000	1.471500e+05	5.135310e+05	24903.000000	4.500000e+05	0.0188
75%	367142.500000	0.000000	1.000000	2.025000e+05	8.086500e+05	34596.000000	6.795000e+05	0.0286
max	456255.000000	1.000000	19.000000	1.170000e+08	4.050000e+06	258025.500000	4.050000e+06	0.0725

8 rows × 106 columns

To achieve the model following methods were followed:-

1.Cleaning the data-

- ☐ columns with more than 20.% and 50% were dropped. We had removed External source column as they were no linear correlation

2. Cleaning The Data Application dataset has overall 9152465 null values in diff col.

- ☐ Some column have more than 50% and 20% of total count rows.

- ☐ A→application data

- ☐ same for previous data total count of null values=11109336. It also has more than 50% and 20% of total count rows in some column so we drop the column which have more than 50% null values

- ☐ After dropping column which have 50% null values 81 columns are remaining in dataset. After that we observe for more than 20% null values only 2 column are relevant to target so we drop that

  column also except 2 column i.e.
  'OCCUPATION_TYPE','EXT_SOURCE_3
  after that we check the correlation between EXT_SOURCE_3, EXT_SOURCE_2 withTARGET as they have normalised value using heatmap. These seems to be no linear correlation so we drop that 2 column also. So we left with 72 column

- ☐ Next we check columns with FLAGS and their relation with TARGET columns to remove irrelevant ones and using group by command we found that Columns (FLAG_OWN_REALTY, FLAG_MOBIL ,FLAG_EMP_PHONE, FLAG_CONT_MOBILE, FLAG_DOCUMENT_3) have more repayers than defaulter and from these keeping FLAG_DOCUMENT_3,FLAG_OWN_REALTY, FLAG_MOBIL more sense thus we can include these columns and remove all other FLAG columns for further analysis.
- ☐ After that we remove all the unnecessary columns and input values for relevant missing columns wherever required

- ☐ Then we found outliers in column that columns are= AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN , AMT_INCOME_TOTAL, DAYS_EMPLOYED

- ☐ Then we convert the desired columns from object to categorical column

- ☐ After all, cleaning and imputing we left with 53 columns
  B→previous data

- After removing column which have 50% null values 33 columns remaining in dataset some of them are not needed so we drop that .then 29 columns are remaining

2)Missing values-

- Total Number of columns having missing values more than 50% in application dataset:41
- After dropping 41 columns we are left with 81 columns
- After removing columns having missing value more than 20 percent except the occupation_type , we are left with 73 columns

3)<u>Study of relationships between variables:</u>

To find the relationship between influence of one variable (EXT_SOURCE_3) over the other variable (Target) heatmap is used . Outcome was that there is no linear correlation between target and ext_source_3.
After dropping this column we are left with 72 columns in application dataset

4)Removing irrelevant columns

All the flag variable data are stored in array variable and it is removed as it has more number of repayers than defaulters
Total count of flag variable is 25

72-25 =47 relevant columns

5)Graph analysis

Firstly, null values are replaced with Unknown in Occupation_Type
Occupation_type is plotted and number of labourer has more default values.

6)Tackling data imbalance and standardising the values-

Some columns had negative , positive values which included days.
Columns DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH, DAYS_LAST_PHONE_CHANGE
which counts days that have negative values. thus will correct those values
so we convert negative values in positive value by arp function and convert days in range by bin and slot

convert DAYS_BIRTH to AGE in years , DAYS_EMPLOYED to YEARS EMPLOYED

7) Converting the data into categorical form

for AMT_INCOME_TOTAL, AMT_CREDIT, AMT_GOODS_PRICE
Binning Numerical Columns to create a categorical column
for AMT_INCOME_TOTA
Creating bins for income amount in term of Lakhs
convert DAYS_BIRTH, DAYS_EMPLOYED columns in terms of Years

8)Identifying Outliers

from describe we could find all the columns those who have high difference between max and 75 percentile and
the ones which makes no sense having max value to be so high.
Observation–
AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have some number of outliers.
AMT_INCOME_TOTAL has huge number of outliers which indicate that few of
the loan applicants have high income when compared to the others.
DAYS_BIRTH has no outliers which means the data available is reliable.
DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which is impossible and
hence this has to be an incorrect entry.

9)Unique value analysis

Checking the number of unique values each column possess to identify categorical columns
Converting Desired columns from Object to categorical column
Total number of categorical columns= 21
Observation–
After removing the null values we have 54 columns

10)DATA VISUALISATION

After removing the correlated columns we plot the heatmap
For further analysis,dateset is divided into two dataset of target=1(client with payment difficulties) and target=0(all other)

a) Univariate analysis-doing Categorical Univariate Analysis
   Plotting income range,income type,contract type,organisation type
   Observation–
   Income range—**Men are at relatively higher default rate**
   Income type-working are highest repayer
   Organisation type-Bussiness entity type 3 highest repayer

b) bivariate

- ***<u>Previous Application Dataset Observations</u>***

- ☐ There are columns having negative, positive values which includes days in previous application datset.
- ☐ There only 4 columns with missing valus(Null values) more than 50% in previous application dataset.
- ☐ There are null values in columns 'DAYS_FIRST_DUE', 'DAYS_TERMINATION', 'DAYS_FIRST_DRAWING','DAYS_LAST_DUE_1ST_VERSION', 'DAYS_LAST_DUE' and these columns count days for the installment thus will keeping null values as they are.
- ☐ There are few negative values present which are then converted to absolute values.
- ☐ Number of Days can be converted into years .
- ☐ ***34.35% loan applicatants have applied for a new loan within 1 year of previous loan decision.***