UPGRADE EXPLORATORY DATA ANALYSIS ASSIGNMENT 1

SUBMITTED BY: KANISHK JAIN

BATCH: DS C59

COURSE: ADVANCED CERTIFICTION IN DATA SCIENCE

AGENDA

- THIS CASE STUDY AIMS TO IDENTIFY 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 (TOO 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 APPLICANT'S USING EDA IS THE AIM OF THIS CASE STUDY.

DATA CLEANING

THE FIRST STEP AFTER ANALYZING THE DATASET FOR EXPLORATORY DATA ANALYSIS IS TO CLEAN THE DATA. SOME BASIC STEPS TO CONSIDER FOR CLEANING DATA ARE:

1. REMOVING ALL THE IRRELEVANT COLUMNS FROM THE DATASET.
2. ANALYZE THE NULL VALUES IN THE DATASET. IT THE NULL VALUES ARE MORE THEN 30 OR 40%, IT IS ADVISABLE TO DROP THOSE COLUMNS FROM THE DATASET.
3. FILL THE NULL VALUE WITH MEAN, MEDIAN AND MODE. IF THE COLUMN'S DATATYPE IS IN INT, IT IS ADVISABLE TO FILL THE NULL VALUES WITH MEAN OR MODE AND IF THE COLUMN'S DATATYPE IS IN CHAR, THE NULL VALUES CAN BE FILL WITH MODE OF THE COLUMN.

DATA SET DESCRIPTION

- THE DATASET PROVIDED IS A BANK DATASET WHICH CONSIST THE DATA OF THE APPLICANTS APPLIED FOR THE LOAD AND THE AIM OF THE ANALYSIS IS TO FIND THE APPLICANTS WHICH MAY FACE DIFFICULTY TO PAYBACK THE LOAN AND THE APPLICANTS WHO WILL NOT HAVE DIFFICULTY TO PAY THE LOAN WITH INTEREST.
- THE DATASET CONSIST OF 122 COLUMN AND THE DATA IS OF 3LAKH+ APPLICANTS.
- THERE ARE A LOT OF NULL COLUMN WHERE THE NULL VALUES ARE MORE THEN 30-40%
- IT CAN ALSO BE OBSERVED THAT THE DATES COLUMN ARE NOT IN THE CORRECT FORMAT

Dealing with Null Values:

1. There were many columns in the dataset which were having more then 30% null values, using the drop function those columns were removed form the dataset.
2. Next, there were 6 columns whose percentage of null values were less then 0. To deal these null values, one can either fill the null values with mean, median, mode or drop those rows. For those 6 column, the null values were dropped from the dataset as it may not create any big impact while analysis.
3. Moving further, there were many categorical as well as numerical columns with null values more then 1%, to deal with those null values, categorical column's null values were filled with mode of the column and numerical column's null values were filled with median.
4. It can also be observed that the dates column are not in the correct format which was coded in the correct that is in years instead of days and also converted into positive values.
5. There are many NA values in columns like for occupation the NA columns were dropped and for the gender column the NA values was filled with median of the column that is "F"
6. After dealing with all the null values, the shape of the dataset was (305545, 52)

STEPS OF DATA CLEANING IN DATASET

- After dealing with all the null values of the dataset, the next step was to drop all the column from the dataset which were irrelevant and of no use for the analysis.
- There were around 10 which were needed to be drop and were executed using drop function
- The new shape of the dataset now is (305545, 42)
- Also it was important to track all the NA values in the dataset. After locating there were 2 column which were having NA values. ORGANIZATION_TYPE and CODE_GENDER. The NA organization column were dropped from the dataset and NA values of gender column was filled with median of the column.
- Next, the data types of the column were checked and the columns were converted into correct datatype like there were 3 different data types of columns, int, float and object. It would be appropriate to convert float to int for analysis. These changes were made in data types of the columns.

OBJECT COLUMNS ANALYSIS

- After plotting bar plot for object columns, it is observed that:
- 1. There are more cash loan then revolving loans
- 2. Females take more loans then men
- 3. Maximum people does not own any car
- 4. But maximum people do own a reality
- 5. The maximum loan seekers are the one who are working
- 6. The maximum people taking loans have completed secondary degree
- 7. The people are in demand for loan are the one who are married
- 8. Maximum people who have demand for loan have their own house
- 9. Also, people from business entity are the one more in demand for seeking loans

INT COLUMN ANALYSIS

- After making a list of all the integer columns and with that this after plotting a boxplot for that column using loop it can be observed that:
- There are a lot of outliners in the column
- The applicants have outliners for having children, there are more then 7 children going all the way up to 19 children in a family which is not usually a case and should be considered as outliners also having more then 6 or 7 children can result in delay of loan payment.
- Columns having outliners are:
- Amt_income_total
- Amt_credit
- Amt_annuity
- Amt_goods_price
- Years_employed
- Years_registration
- Own_car_age
- Days_last_phone_change

UNIVARIATE ANALYSIS

- After cleaning the data the first step taken to analyze the data is done by plotting a count plot for the target column. It is observed that more then 80% applicants from the dataset are eligible and capable of paying back loan and only less then 20% applicants have issues for the loan payment.
- Moving further a list of column for int and object datatypes individually is made to simplify the analysis and further plotting are performed on them
- Further a hist-plot and boxplot IS plotted for the better understanding of the data which gave the information that the applicants have outliners for having children, there are more then 7 children to a family which is not usually a case and should be considered as outliners also having more then 6 or 7 children can result in delay of loan payment also there are around 10 more columns having outliers.
- After plotting a bar plot for object column, it came to know that:

1. There are more cash loan then revolving loans
2. There are more female customers then men
3. Maximum people does not own any car
4. But maximum people do own a reality
5. The maximum loan seekers are the one who are working
6. The maximum people taking loans have completed secondary degree
7. The people are in demand for loan are the one who are married
8. Maximum people who have demand for loan have their own house
9. Also, people from business entity are the one more in demand for seeking loans

NEW SEPARATE DATASETS

- After the previous task, next two new separate columns for the values of target columns are formed. As this column provide the information of two different types of customers of the band, it is a wise move to make a two different dataset for both types of customers to find the difference in the information and actions of the individuals.
- After separation, we formed two dataset df1_1 (with people who may have difficulty to pay back the loan) with shape (21742, 44) and df1_0(the rest) with shape (228777, 44)

```
df1_1=df1[df1['TARGET']==1]
df1_1.head()
```

```
df1_1=df1[df1['TARGET’]==0]
df1_1.head()
```

RESULT OF TARGET DATASETS

- Target = 1

PERFORMING UNIVARIATE, BI-VARIATE AND MULTIVARIATE ANALYSIS USIING VARIOUS PLOTTINGS

1. Here we can observe that people with secondary/ secondary special education are maximum in the dataset who may not be able to replay the loan

2. Maximum application are from the customers having medium income

3. Most of the clients are from average rating of their living area

4. it can be observed that for the low income category there are lot of female applicants also as got to know before there are more female applicants for the loan application

5. Also the maximum income of female is in 30% of the range of all applications

6. This also defines there are more females then men

7. Most of the income of the application lands in the medium range of the income group

8. We can observe that most of the applicants are working professionals and in which the females are in majority

9. Also lowest are state servants in comparison

10. There are no applicants on maternity leave

11. We can observe here is that, lower the age, the credit amount of loan gets a little higher side

12. we can observe that people with very high salary are demanding for a bigger laon as compared to others and visa versa

OUTLIERS FOR TARGET COLUMN

TARGET = 1

1. We can see there are many outliers in many column

2. Firstly, marriage, civil marriage and separated are having extreme outliers or extreme credit

3. Secondly, married from incomplete higher education are having extreme values or credit

4. Similarly, Single, married and civil married with higher education are having multiple extreme values or credit

TARGET = 0

speaking for the target=0

1. we can clearly observe all the categories of married status with higher education are having extreme credits which are outliers

2. married with secondary and secondary special education have higher values for credit.

Also, after Bivariate analysis on the numerical columns int_new_col it can be observed that AMT_CREDIT and AMT_GOODS_PRICE are highly correlated which means one increases then the other increases as well and visa versa.

PREVIOUS APPLICATION DATABASE

- Here the second provided database comes to work. The database was already imported. Now the next step for this dataset is to check the percentage of missing values in each column of this dataset which are relative to the application dataset df1. With this analysis it came to know that there are no null values to work on for merging the dataset in future.
- Next previous application is check for NA values and removed to reduce any type of misunderstanding and error in the further analysis.
- After cleaning the important column of this dataset, next step is to merge this previous application dataset with application dataset with respect to a common column SK_ID_CURR.

UNIVARIATE ANALYSIS ON MERGED DATASET

- Plotting a count plot for NAME_CASH_LOAN_PURPOSE

we got to know from the above graph:

1. Most approved loans are for the repairing purpose

2. Also most rejections are also for the repair purpose

3. There are nearly none who have used offers

4. Least visible applications are for buying garage so there afe least activities of status visible bit it can be seen that rejections are in majority

5. Only everyday expenses and purchase of electronic equipments have higher approval then rejections

6. Payments on other loan, buying a new car and a house has significantly higher rejections then approval.

BIVARIATE ANALYSIS ON MERGED DATA

- After plotting a bar graph for AMT_CREDIT_left credit amount and NAME_HOUSING_TYPE as these two columns are the best categories to judge the background and the living of a person for their ability to payback the loan and also for recovery of any repay failure we got to know that

1. The Co-op apartment have the highest credit of Target = 1 which indicates that the bank should avoid giving loan to Co-op apartments as they have a higher risk that they may not be able to payback the loan

2. Office apartment and house has the highest credit for Target = 0. These are the housing categories where bank can have a confidence of getting back their loan amount with interest which will give profit to the bank

3. Also bank should avoid giving loans to the people with rented apartments and co-op apartments as they have the higher risk of paying back the loan

4. The bank should consider giving loan to people who have house, municipal apartment and the people who live with parents, as they have lower risks and can payback the loan with interest and make profit for the bank.

CONCLUSION OF THE CASE STUDY

- 1. The bank should target the female customers as they are in the majority for the loan applicants
- 2. The bank or the company should target business entity type 3 and self employed as they are at the lowest risk and highest in count for the target 0 which represents the people who have no issue of paying back the loan
- 3. The bank should land approve loan for the reason of repair and other purposes
- 4. The bank or the company should choose customers with residence as, municipal apartment, apartment, office apartments, municipal apartments which has low risk of losses.
- 5. Best targets are female who are self employed, have reason of loan for other purposes and live in an apartment.

THANK YOU