Credit Analysis of Data for Loan

EDA CASE STUDY

DONE BY
SAMANYU GHOSE

Introduction

❑ The case study is for analysing the data in real time scenario and for understanding the concepts of Exploratory Data Analysis (EDA)

❑ Uses of different tools and techniques to get maximum insights from the data which were covered under the chapter

❑ In the case study we have been assigned to evaluate two kinds of risk (ie; 1)Credit Loss 2)Interest Loss) pertaining to Banking Sector and to minimise these risk further. Thus, helping us to improve business more specifically "Profit"

Continued….

- So when a loan application is received by the Bank, the authorised person has to decide whether to give loan to the applicant or not. It consists of two types of Risks

- First one is the risk associated with **Credit Loss**. It means if the applicant is not likely to repay the loan and the loan has been sanctioned to such person, It leads to **Credit Loss**

- Secondly , when the customer is likely to pay the loan and loan is not sanctioned to such an applicant ,It leads to **Interest Loss** for the Bank

Data Cleaning

- The data that we get in the real world, is in an unstructured format which is fixed by using data cleaning.

Process :-

- Firstly, We import data in Python data frames.
- Secondly, We verify and handle the missing values by looking at the no of rows and columns. The columns having a high percentage of null values (greater than 50%) are dropped from the data frame.
- Also we have dropped certain columns where the column is meaningless such as FLAG_DOCUMENT_2, FLAG_DOCUMENT_3, YEARS_BEGINEXPLUATATION_AVG', 'FLOORSMAX_AVG','YEARS_BEGINEXPLUATATION_MODE','FLOORSMAX_MODE', etc

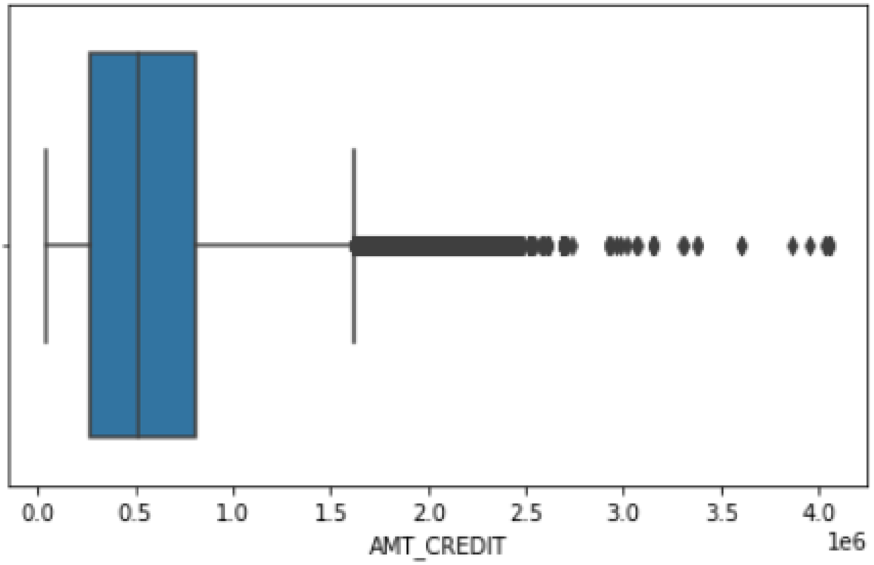
Imputation :-

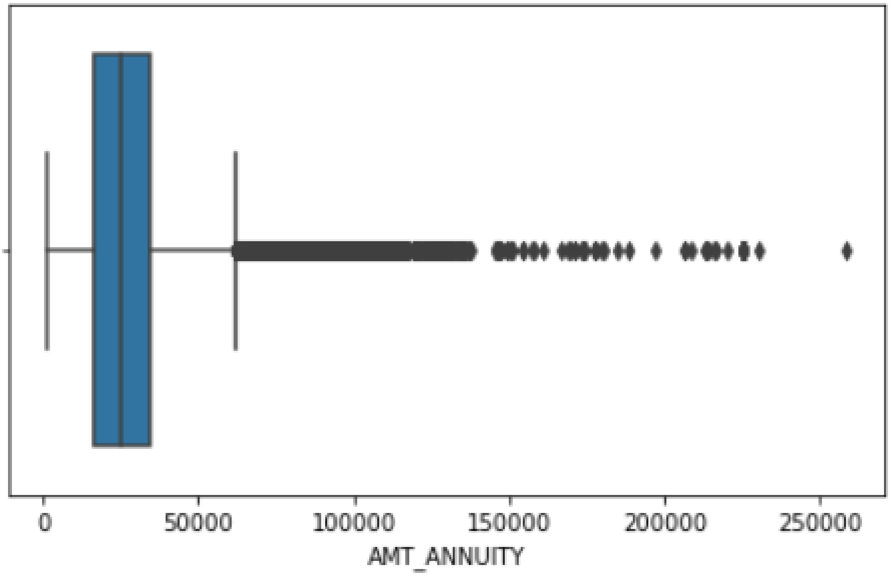
- Similarly, where the null values in columns is less than say 13%, they may be imputed by using zeroes/mean/median in case of numerical data and in case of categorical data, mode can be used. For e.g columns such as AMT_ANNUITY, AMT_GOODS_PRICE, AMT_REQ_CREDIT_BUREAU_HOUR, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_YEAR can be imputed by median values where as OCCUPATION_TYPE can be imputed by Mode values
- Also we have changed the data types of certain columns from Float to Integer as these values cannot be in Float such as DAYS_REGISTRATION, CNT_FAM_MEMBERS
- We have also added a new column AGE, here we have put in Date of Birth in years by dividing it by 365 to draw analysis in the latter part
- Also we have converted certain columns into years such as DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH

Outliers

Checking for the Outliers for atleast 5 variable

- We have checked at least 5 Outliers namely AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, DAYS_EMPLOYED





AMT_CREDIT - Here the 99th percentile is 1854000 whereas max is 4050000 which implies that there is no significant outlier in this column

AMT_ANNUITY -Here 99th percentile is 70006.5 whereas max is 258025 which shows that there is an Outlier in this column

Binning

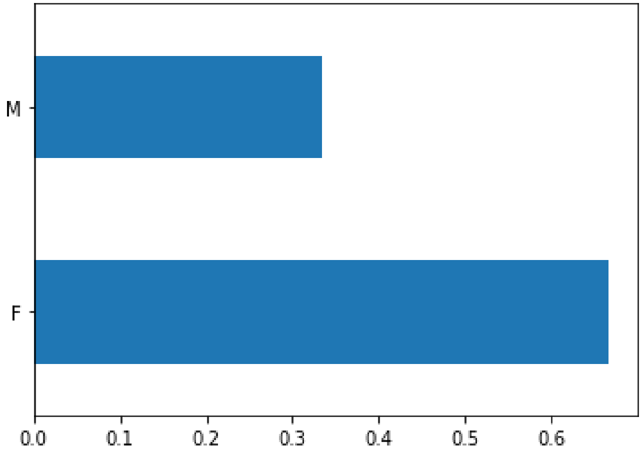
- We have performed Binning on AGE column for Application data and another Binning after merging the Application data and Previous Data on AMT_INCOME_TOTAL

- This is to find that in a given spread of data, how another variable is related

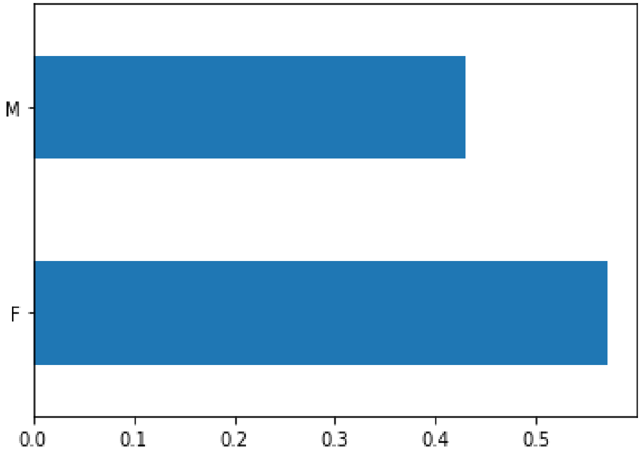
Data Analysis through UNIVARIATE : Categorical

- We have first checked the Imbalance percentage of TARGET columns which is 0 and 1 respectively (0 for non-defaulters and 1 for customers with payment difficulties)
- Now we have done Univariate analysis by dividing the data frames into Tar_0(for 0 values) and Tar_1(for 1 values)
- For Univariate categorical, for e.g we have taken CODE_GENDER column and both the graphs for two data frames as shown below

Tar_0



Tar_1

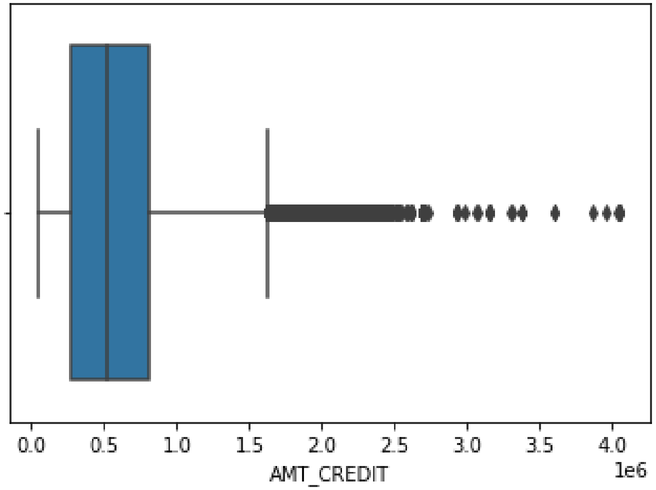


- It shows that for Male gender, the number of customer with payment difficulties is higher in comparison to non-defaulters
- The number of customers with payment difficulties in Female gender is less in comparison to non defaulters

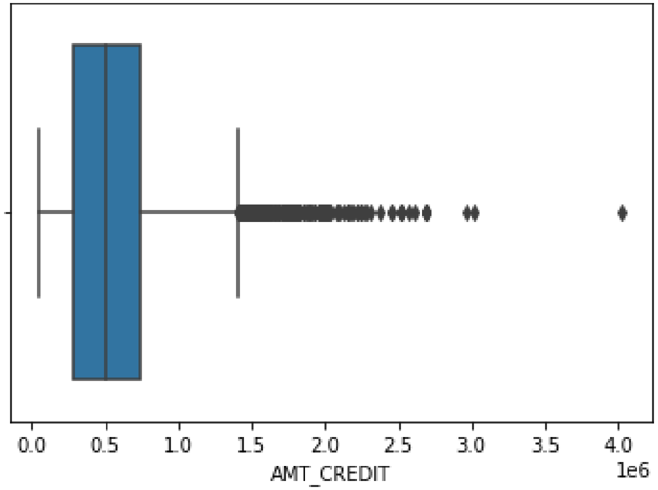
Data Analysis through UNIVARIATE : Continuous

- For Univariate Continuous, for e.g we have taken AMT_CREDIT column and both the graphs for two data frames as shown below

Tar_0



Tar_1

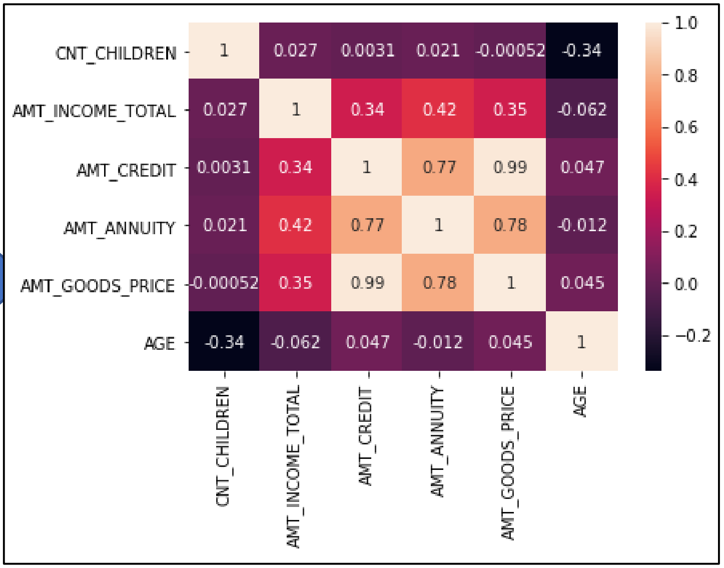


- The loan credit given to the customers for non-defaulters is uniform and the outliers are contiuous in nature and the maximum mass is slightly more in comparison to the customers having difficulties in making payments, median values are almost same in both the graphs
- For the customers having some difficulties in repayment, the outliers are continuous upto an extent still we find a point which is far from rest of points which can be considered as an outlier

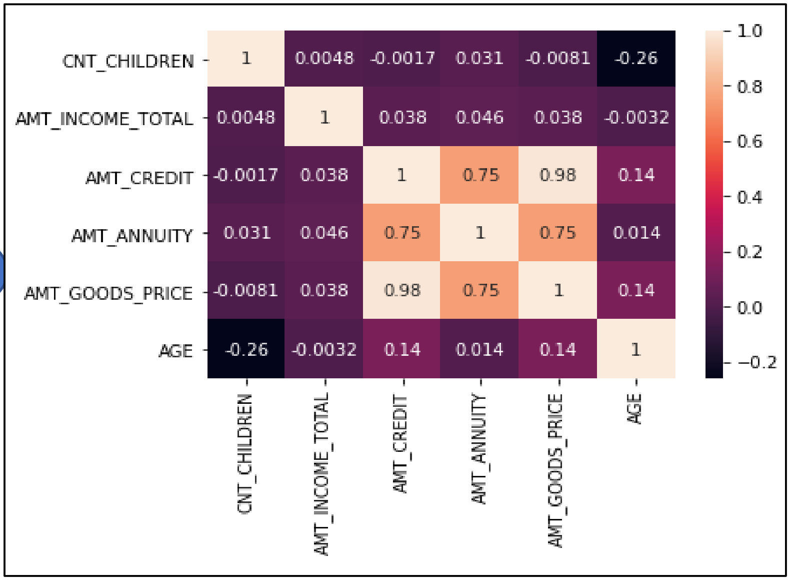
Data Analysis through BIVARIATE : Continuous TO CONTINUOUS

- We have created a new data set for Tar_0 and Tar_1 namely num_0 and num_1 for columns such as CNT_CHILDREN','AMT_INCOME_TOTAL','AMT_CREDIT','AMT_ANNUITY','AMT_GOODS_PRICE',’AGE and created a Heat map to come to find a correlation.

num_0



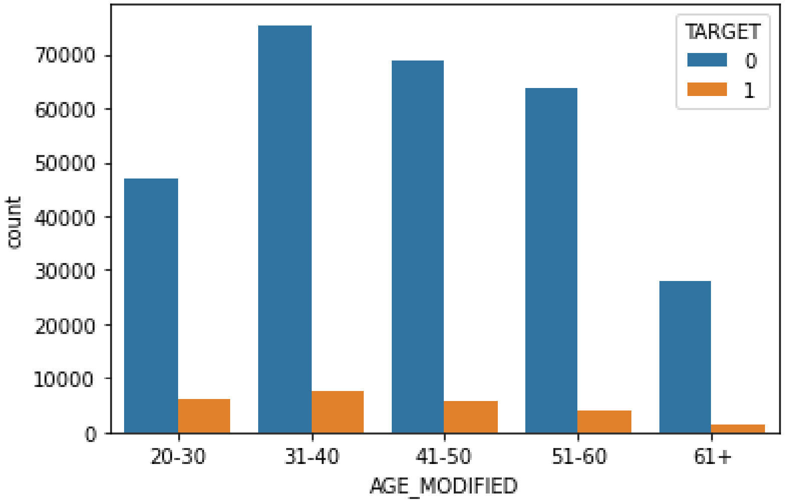
num_1

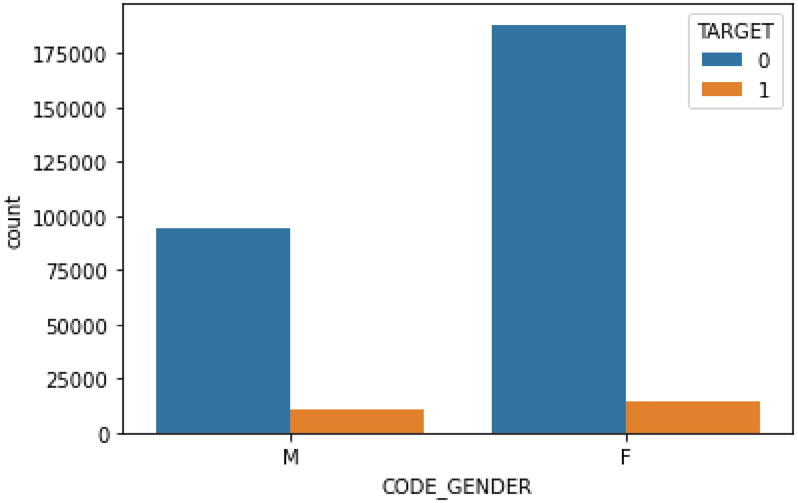


- The lighter shade corelates to higher correlation. For e.g Higher the Goods price, the higher would be the credit amount in both the Heat maps.
- The darker shade correlates to negative correlation. That means if value of one variable increases, the value of other variable will decrease and vice versa. For e.g AGE and CNT_CHILDREN

Data Analysis through BIVARIATE : Categorical TO Categorical

- Drawing count plot between 2 categorical variables AGE_MODIFIED (Binned variable) and TARGET(0 as well as 1) and Drawing count plot between 2 categorical variables CODE_GENDER and TARGET (0 as well as 1)

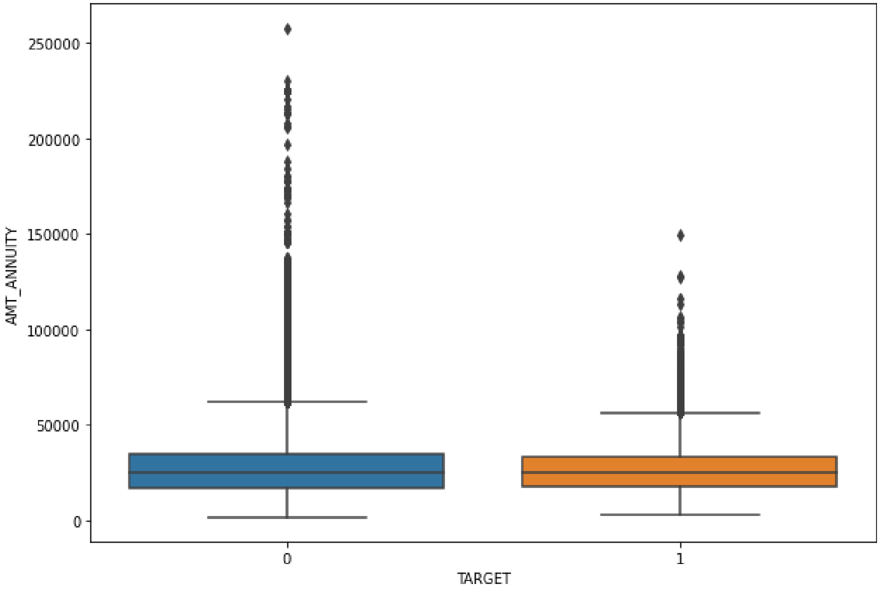


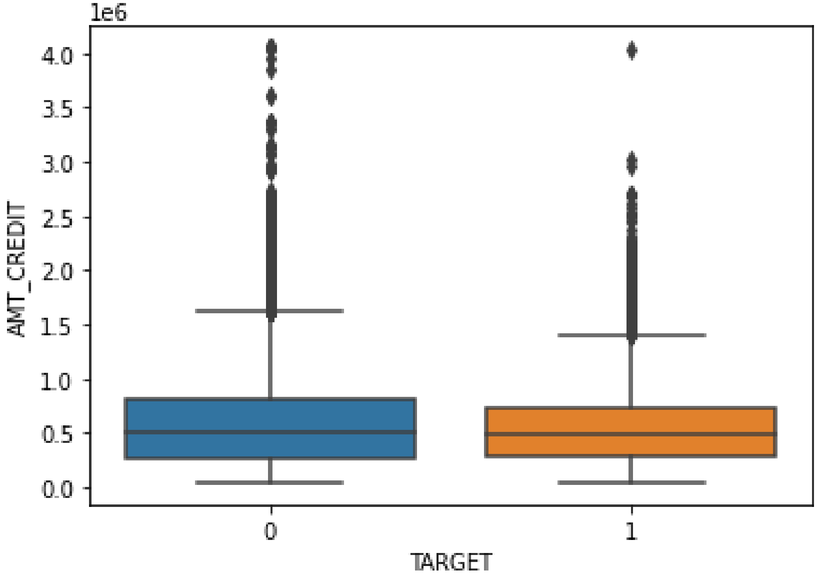


- When comparing with AGE_MODIFIED column, we can see that in the age bracket of 31-40, the number of customers who are non defaulters is the highest, same is the situation for customers who are having difficulties in making payments. Also the lowest number of people are in 61+ age bracket where there is less difficulties in making payment
- When comparing with Male and Female gender, the number of Female non-defaulters are more in comparison to Male and similarly

Data Analysis through BIVARIATE : Categorical TO Continuous

- Drawing box plot between 2 categorical variables AMT_ANNUITY (Binned variable) and TARGET(0 as well as 1) and Drawing count plot between 2 categorical variables AMT_CREDIT and TARGET (0 as well as 1)

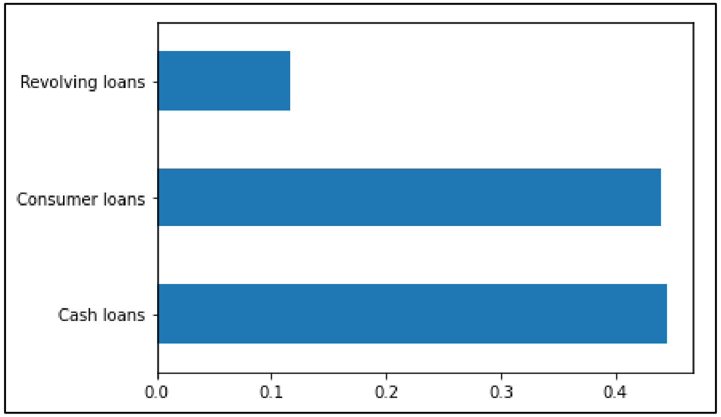


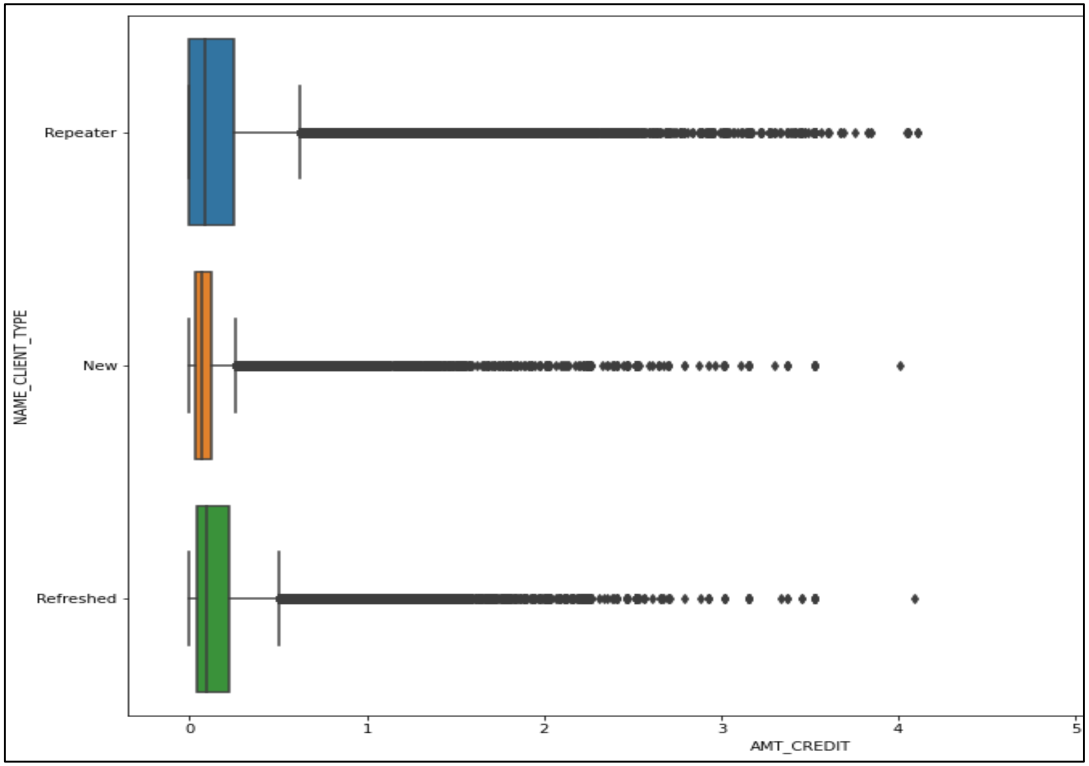


- When comparing with AGE_ANNUITY column, for both types of Customers with non-defaulters and customers having payment difficulties, the mean,median, 25th and 75th percentile is almost at the same level but the number of outliers for non defaulters are much higher. It means people facing problems in making payments have not such high amount of annuity which outcasts a general idea that higher the amount of annuity more difficult is for the customers to pay
- In the Second graph we can see that the median is almost the same for both cases while the upper hinge is higher in case of people not having any difficulties in making repayments. Moreover, the outliers in case of people facing no problems in repayment is of continuous nature while there is a clear outliers in case of defaulters

Conclusion on Previous Application Data

- We have taken out the Data cleaning activity dropping the unnecessary columns where the NULL values is more than 50%
- For e.g We have done Univariate Categorical analysis on NAME_CONTRACT_TYPE and Bivariate Continuous with Categorical analysis (NAME_CLIENT_TYPE with AMT_CREDIT) as given below

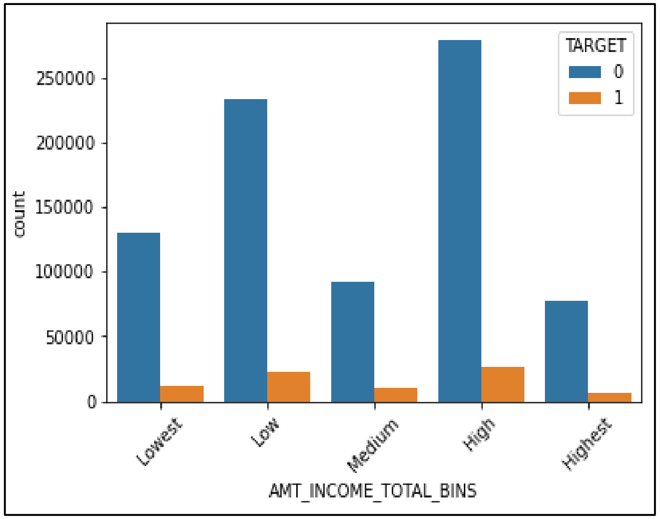


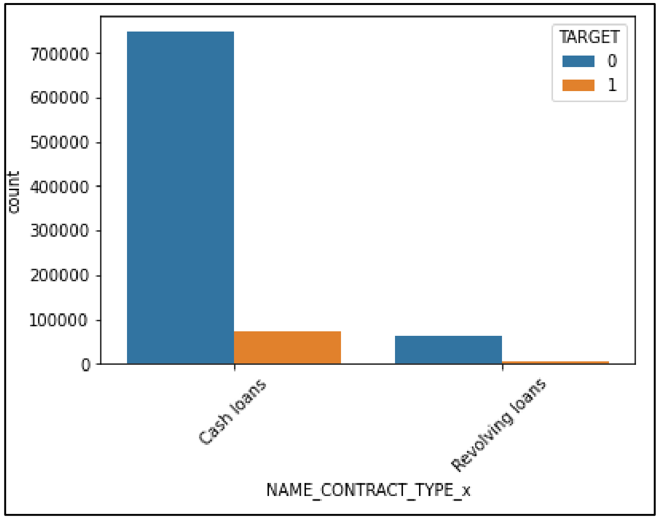


- In Univariate Categorical Analysis, Contract type of Cash Loans is the highest that means cash loans should be targeted more
- In Bivariate Continuous to Categorical Analysis, the spread of data is highest in case of Client type as Repeater it implies that they were given more amount of credit.

Merged Data Conclusion

- Merging both the Application and Previous application data set we have analyzed that the income level of different customers and their ability to pay



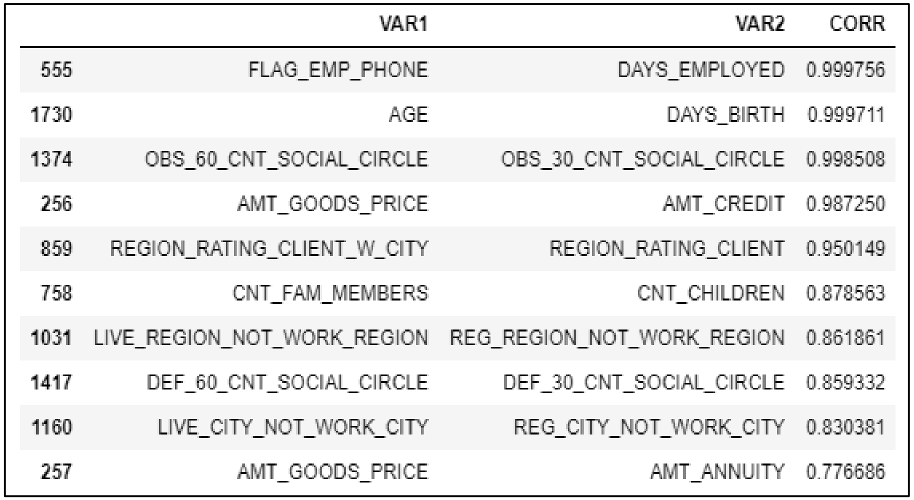


- The Customer in the High income group tends to repay the loan without any difficulty and interestingly Customers from same group represents Highest number defaulters amongst all categories
- The credit department has more focus on giving Cash Loans to the customer as the number of default is much less as compared to the customers who have defaulted

TOP 10 Correlations on Target data frame

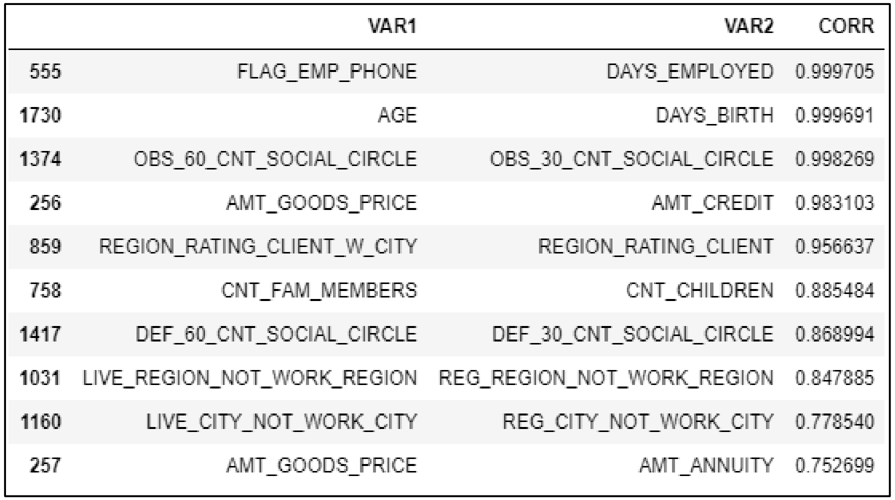
- Tar_0 and Tar_1 Data frame

Tar_0



	VAR1	VAR2	CORR
555	FLAG_EMP_PHONE	DAYS_EMPLOYED	0.999756
1730	AGE	DAYS_BIRTH	0.999711
1374	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998508
256	AMT_GOODS_PRICE	AMT_CREDIT	0.987250
859	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950149
758	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878563
1031	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861
1417	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859332
1160	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.830381
257	AMT_GOODS_PRICE	AMT_ANNUITY	0.776686

Tar_1



	VAR1	VAR2	CORR
555	FLAG_EMP_PHONE	DAYS_EMPLOYED	0.999705
1730	AGE	DAYS_BIRTH	0.999691
1374	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998269
256	AMT_GOODS_PRICE	AMT_CREDIT	0.983103
859	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637
758	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
1417	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.868994
1031	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.847885
1160	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.778540
257	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699

- When we have compared the most correlated variables, in case of Tar_0 and Tar_1, we came to the conclusion that the most highly correlated variables in the given data set is as follows :-
  1. It is most likely that the more number of years a person work in the same organization , the more are the chances of getting his phone no of work place and vice versa. The more connections he will provide for business too.
  2. Obviously , the age of the customer is directly related to the birth date so they are highly correlated.
  3. The areas having more no of people with observation of 30 DPDs surrounding the customer and The areas having more no of people with observation of 60 DPDs surrounding the customer are highly correlated as because the people in the areas who tends to default for 30 days are more likely to default for 60 Days too.
  4. The more the price of the goods, the more the amount of loan will be given to the customer.

Final Summary

From the entire data set , we can conclude that the following factors are responsible for the people to repay the loan more than the other categories

1. Age Group of 31-40
2. Gender Type - Females
3. Customers opting for Cash Loans
4. Working Class
5. Married Profile Customers
6. Customers living in Owned House/Apartment
7. Customers type as Repeaters that is who have already taken loan or previous relationship with the bank
8. High Income group

Thank You