Credit EDA Case Study

By Jay Thakur

Credit EDA Case Study

Business Understanding

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

Problem Statement

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicantis likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e., he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Business Objectives

The company wants to understand the driving factors (or driver variables) behind loan default, i.e., the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Approach & Action:

UNIVARIATE & BIVARIATE ANALYSIS - 2

1. Correlation Analysis for Numeric columns using Target variable
2. Univariate and Bivariate analysis and share insight

OPTIONAL: DATA FILE MERGE

- Merge of Data file : Application Data and Previous Application

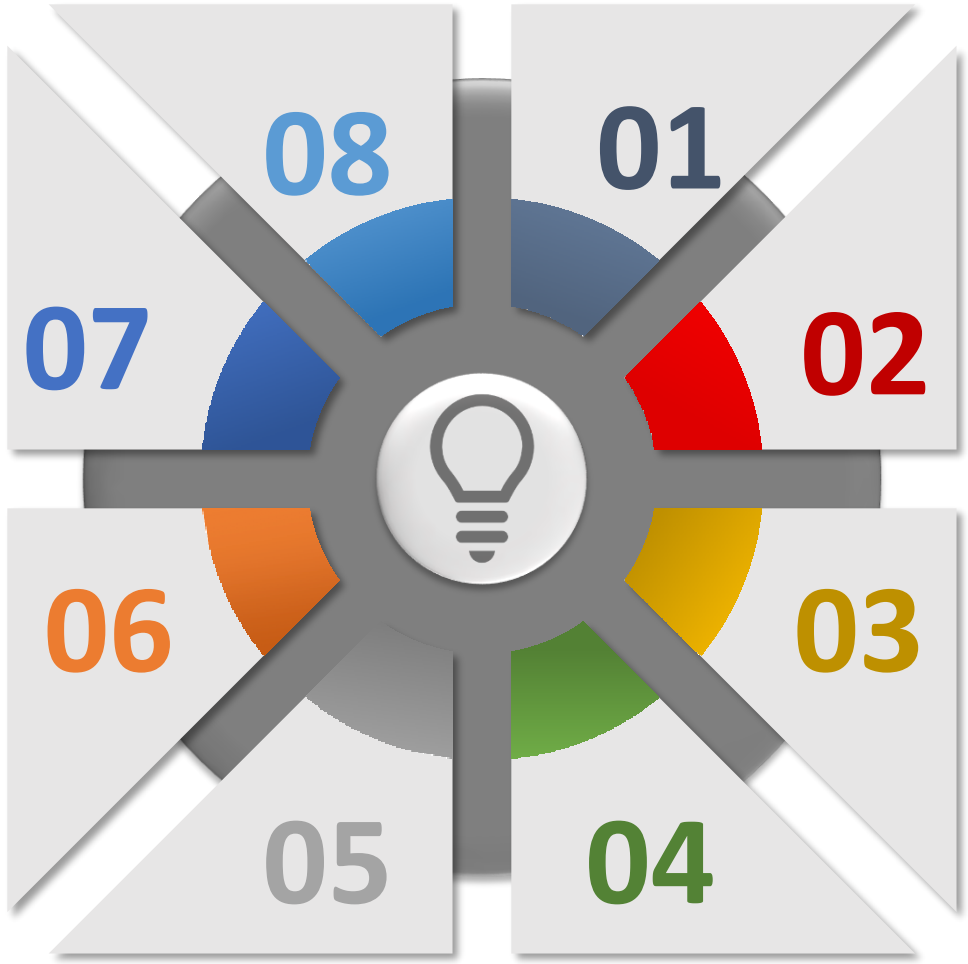
UNIVARIATE & BIVARIATE ANALYSIS -1

- Univariate analysis on Categorical variable & Numeric continuous variable.
- Bivariate Analysis for Categorical – Categorical, Categorical – Continuous & Continuous – Continuous variable
- Share insights

CORRELATION ANALYSIS

1. Correlation Analysis for Numeric columns using Target variable
2. Identity top 10 variable with highest correlation

2. Additional heat map analysis



DATA QUALITY CHECK

- Import Libraries
- Load Data file
- Check Data file Shape and info
- Null value for columns

DATA CLEANUP

- Find missing value percentage for columns
- Drop Columns with more the 50% missing values
- Observation and recommend / impute columns with missing value (5 Columns)

IDENTIFY OUTLIER

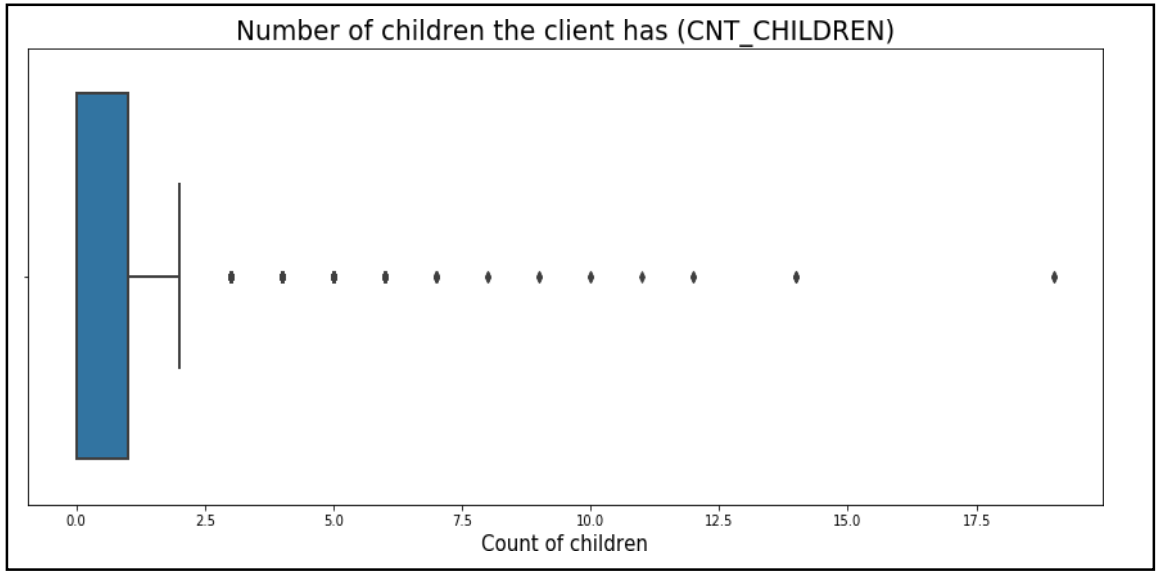
- Identify outlier for 5 columns
- Share observation and reasoning
- Use of Box plot & other relevant plots

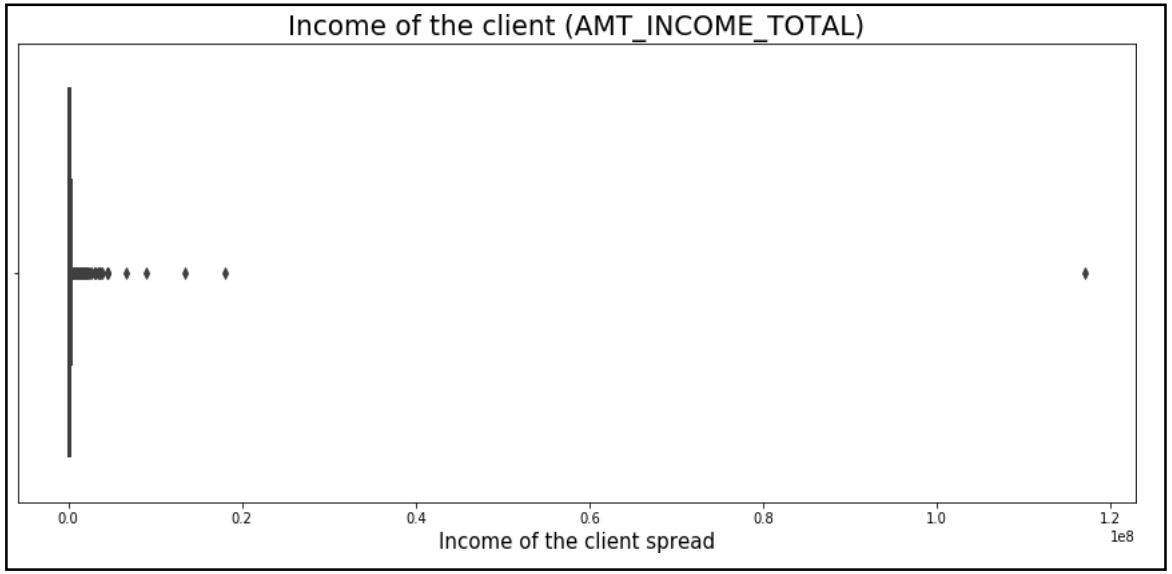
BINNING & IMBALANCE PROCESS

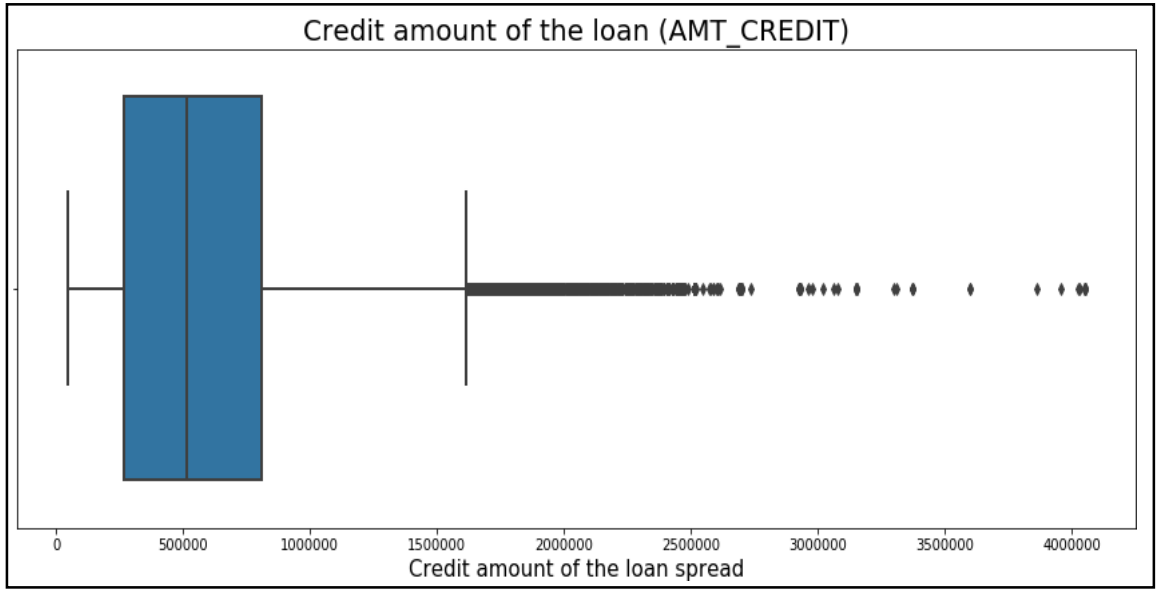
- Binning of 2 Continuous variable columns
- Select relevant columns for analysis
- Check Imbalance percentage with column: Target (0 & 1)
- Share insight

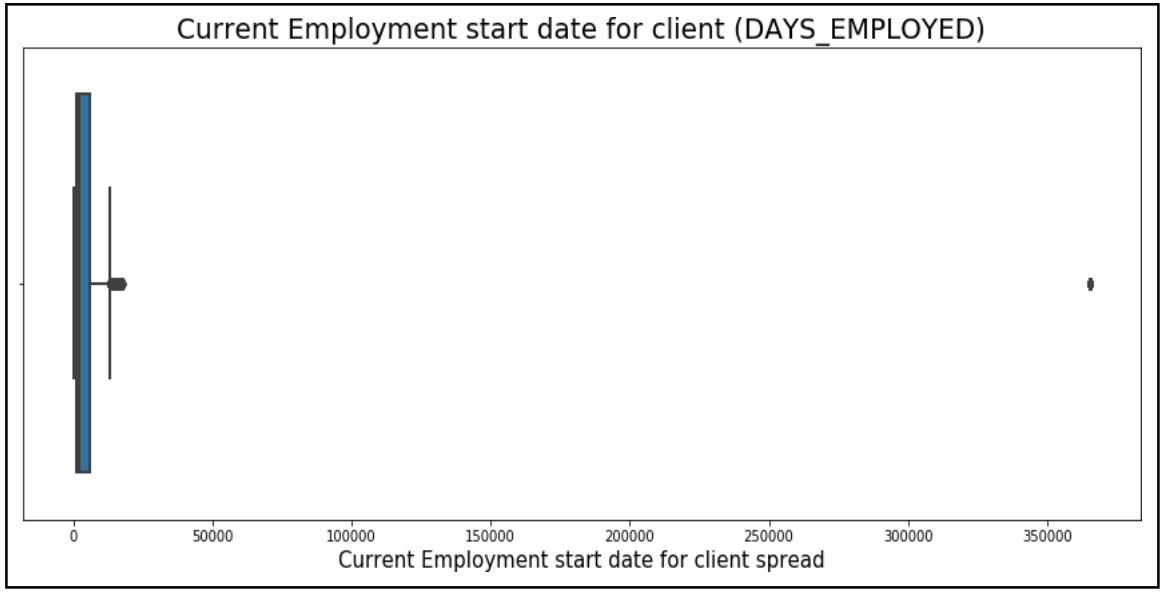
Correlation Analysis:

Identify Outlier





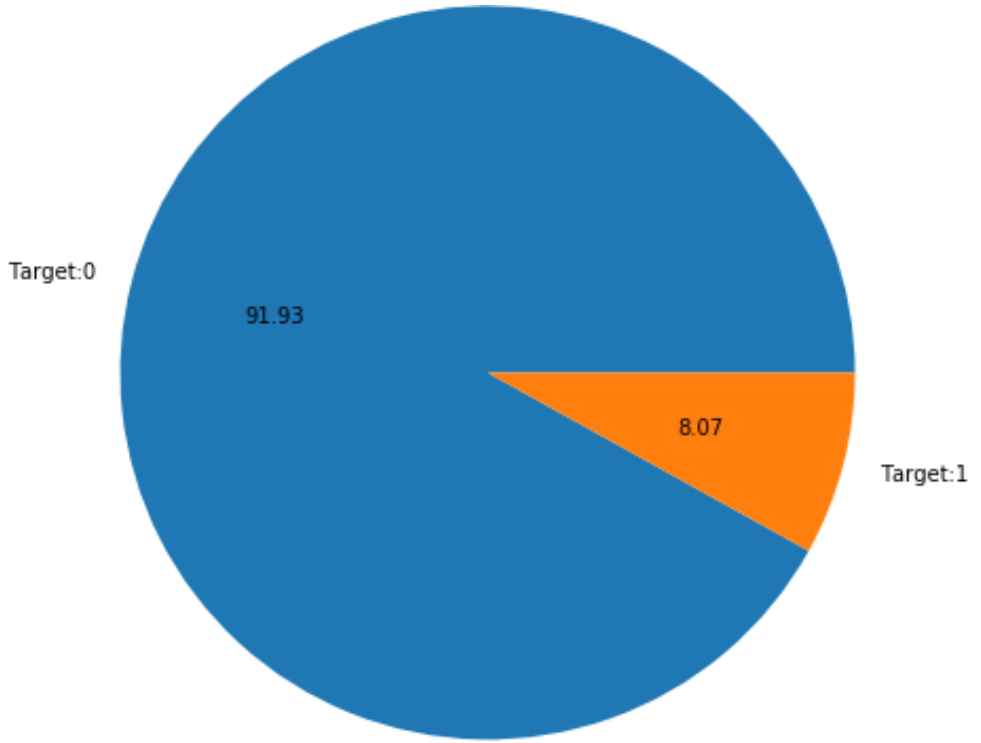




Imbalance Analysis:

Imbalance percentage on Target Column

Target Column:
1 - client with payment difficulties:
0 - all other cases



Observation:

1. It is Imbalance data.

2. 91.93% data is for Target = 0 (Client without payment difficulties)

3. 8.07% data is for Target = 1 (Client with payment difficulties)

Approach & Action:

Correlation Analysis on Application Data file:

Data file Target Column: 0

	Variable 1	variable 2	Correlation	Abs_Correlation
130	AMT_GOODS_PRICE	AMT_CREDIT	0.987250	0.987250
233	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571	0.878571
373	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859332	0.859332
131	AMT_GOODS_PRICE	AMT_ANNUITY	0.776686	0.776686
109	AMT_ANNUITY	AMT_CREDIT	0.771309	0.771309
175	DAYS_EMPLOYED	DAYS_BIRTH	0.626114	0.626114
108	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418953	0.418953
129	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349462	0.349462
87	AMT_CREDIT	AMT_INCOME_TOTAL	0.342799	0.342799
307	REG_CITY_NOT_LIVE_CITY	REG_REGION_NOT_LIVE_REGION	0.341571	0.341571

Data file Target Column: 1

	Variable_1	variable_2	Correlation	Abs_Correlation
130	AMT_GOODS_PRICE	AMT_CREDIT	0.983103	0.983103
233	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484	0.885484
373	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.868994	0.868994
131	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699	0.752699
109	AMT_ANNUITY	AMT_CREDIT	0.752195	0.752195
175	DAYS_EMPLOYED	DAYS_BIRTH	0.582185	0.582185
307	REG_CITY_NOT_LIVE_CITY	REG_REGION_NOT_LIVE_REGION	0.322628	0.322628
196	DAYS_REGISTRATION	DAYS_BIRTH	0.289114	0.289114
149	DAYS_BIRTH	CNT_CHILDREN	-0.259109	0.259109
217	DAYS_ID_PUBLISH	DAYS_BIRTH	0.252863	0.252863

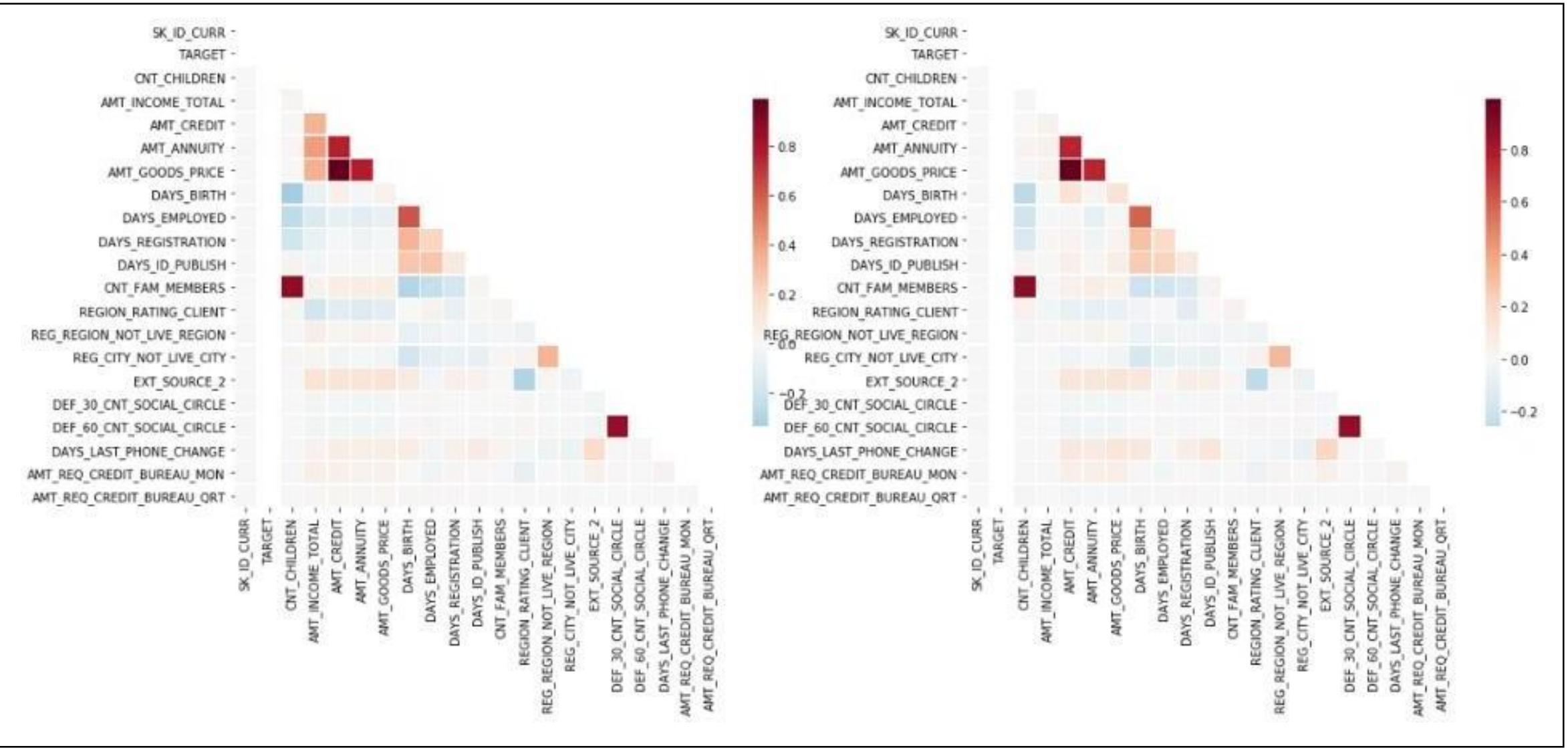
Observation

1. Top 6 variable with highest correlation is same in both Dataset (df0 & df1)
2. In df1, it is interesting to see how:

- REG_CITY_NOT_LIVE_CITY & REG_REGION_NOT_LIVE_REGION has high correlation.
- DAYS_REGISTRATION, DAYS_ID_PUBLISH & DAYS_BIRTH high correlated.
- We will check these correlations in further analysis to find recommendation

Approach & Action:

Correlation Analysis on Application Data file using heat map:



Finding & Recommendation:

- **AMT_CREDIT vs AMT_INCOME_TOTAL**

There is not much of correlation between client income and loan credited. Ideally, loan amount should be linked with client Income. Higher the income, higher the loan eligibility. Lot of high value loan has been credited to low client income and this is not positive move.

- **AMT_INCOME_TOTAL**

There are outliers in client income. Some of the income are unrealistic. Proper data management need to be setup.

- **AMT_CREDIT vs AMT_ANNUITY**

There is a positive linear pattern and high data correlation but loan annuity for the clients having difficulties with payment is high. This could be because of tenure period. Loan tenure period should be based on client income.

- **AMT_INCOME_TOTAL vs AMT_ANNUITY**

There is no major correlation between Income and loan Annuity. Ideally, there should be positive linear correlation with higher the client income, higher the loan annuity. In most of the cases, annuity proportion is high vs client income.

- **AMT_INCOME_TOTAL vs AMT_GOODS_PRICE**

There is no major correlation between Income and loan for good price. Ideally, there should be positive linear correlation with higher client, higher the loan for goods. In most of the cases, loan for good price is increasing without increase in client income .

- **AMT_CREDIT vs AMT_ANNUITY**

There is a positive linear pattern and data has high correlation between Loan credit and Annuity. There seems to be some outlier in Approved status with High loan credit amount vs low Annuity and vice versa.

- **AMT_CREDIT vs AMT_GOODS_PRICE**

There is a positive linear pattern between Loan created and price of good with high data correlation.

Recommendation: Proper data management & balance between Client income vs Loan Credited should be followed to overcome any risk associated with loan replay.

Finding & Recommendation:

- Majority of clients didn't request for Insurance during previous approved applications**. Better insurance option should be provided by the bank.**
- **In Unused offer status, all of these loans are rejected by clients.** Bank should further investigate and follow-ups. Consumer Loan has higher number of Unused offer status. Majority of Unused offer are for Mobile (Goods). High probability of unused Offer for small amount loan (Less than 250000).
- In approved status, Medium and High yield hold the maximum count but in refused status, Low & Medium yield hold the maximum count. **Proper yield and loan interest benefit to be setup for repeating clients.**
- Cash Through Bank is a preferred option for client. **There is payment history for Canceled and Refused loan. This needs further investigation.**
- **Car dealer Channel & Car portfolio type has maximum mean, IQR & outlier across all Approved, Canceled and refused status.** Proper evaluation to be done by bank.

Majority of previous loan & repeater loan were canceled, refused and Unused offer by clients with Target 0 (No difficulty to payment). This is not positive for bank.

- **Income Type:** Businessman has no payment difficulties. They should be promoted for new or repeat loans. Businessman has high client income.
- **Income Type: In** commercial Associates, the loan amount credit has high outliers. Proper loan evaluation to be performed.
- **OCCUPATION_TYPE:** Clients with occupation like Managers and Accountants has high mean and IQR which resulted in difficulties with payment.
- **EDUCATION TYPE:** Clients with Academic Degree having difficulties in payment has higher mean. Proper loan evaluation to be performed.
- **EDUCATION TYPE:** Female clients with Higher Education & Secondary Education has high difficulties in payment. Proper loan evaluation to be performed.
- **EDUCATION_TYPE:** Clients with Lower secondary education loan count is less but percentage of clients with having difficulties in payment is highest.
- **Maternity Leave:** Clients on maternity leave has high mean and resulting to difficulties in payment. Seems like, high value loan is created. proper loan evaluation to be performed.
- **HOUSING TYPE:** Clients having housing type as 'With parents' status has highest percentage with difficulties with payment.
- **FAMILY_STATUS:** Married clients has higher mean, IQR and outliers. We should have proper evaluation.
- **OWN_REALTY:** Own realty could be factor for safer loan. Clients in Target 0 has more % of own house or flat then clients in Target 1 (Difficulties in Payment).
- **OWN_CAR:** Own car could be factor for safer loan. Clients in Target 0 has more % of cars then clients in Target 1 (Difficulties in Payment).
- **FAMILY STATUS:** Clients with 'Married' & 'Separated' status are safe area for loans. As per analysis, they have high volume of loan count vs clients having difficulties with payment.
- **FAMILY STATUS:** Clients with 'Single / not married' status are having highest percentage of having difficulties with payment. Proper loan evaluation to be performed.

Thank You