



CREDIT EDA CASESTUDY - 1

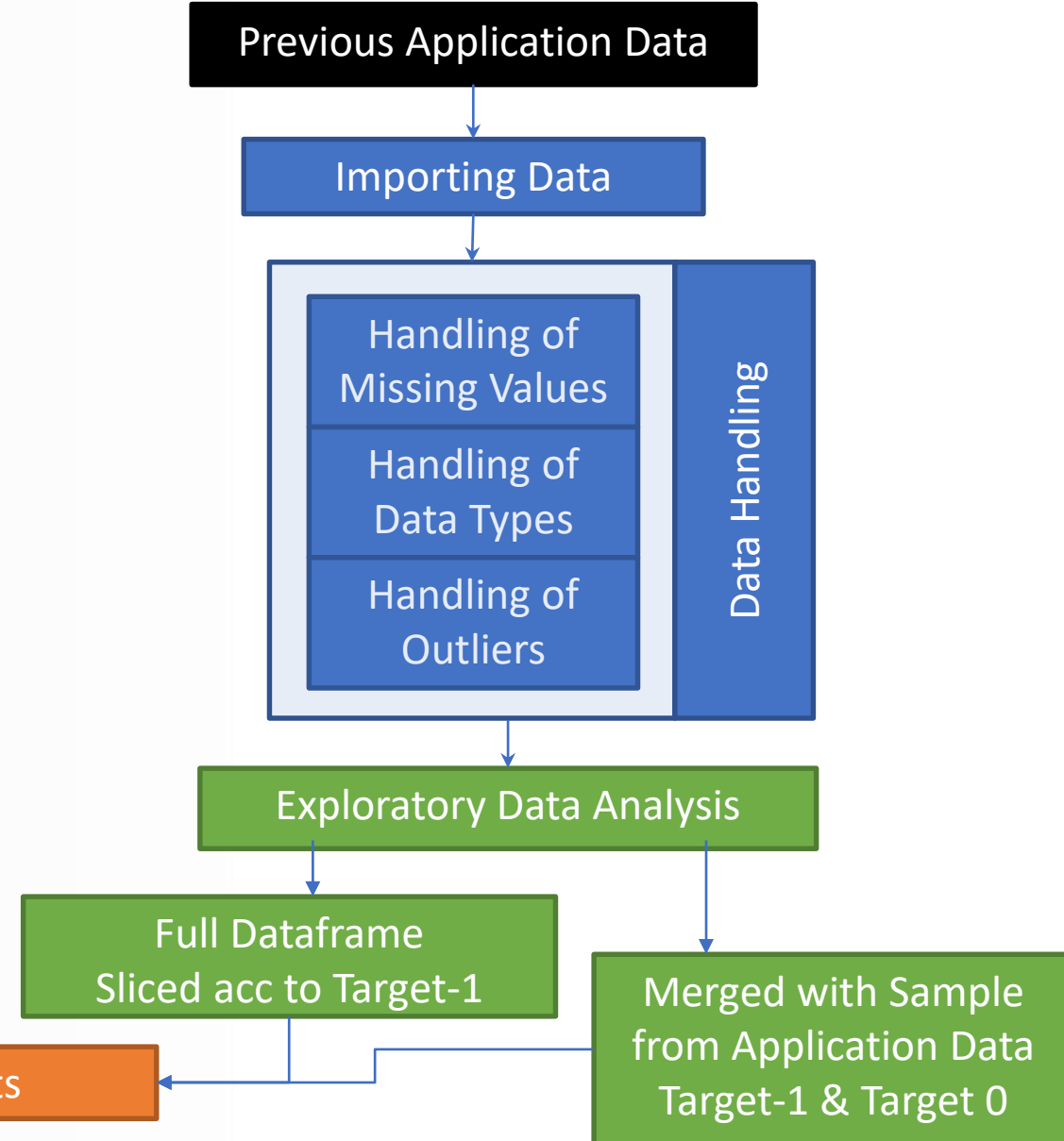
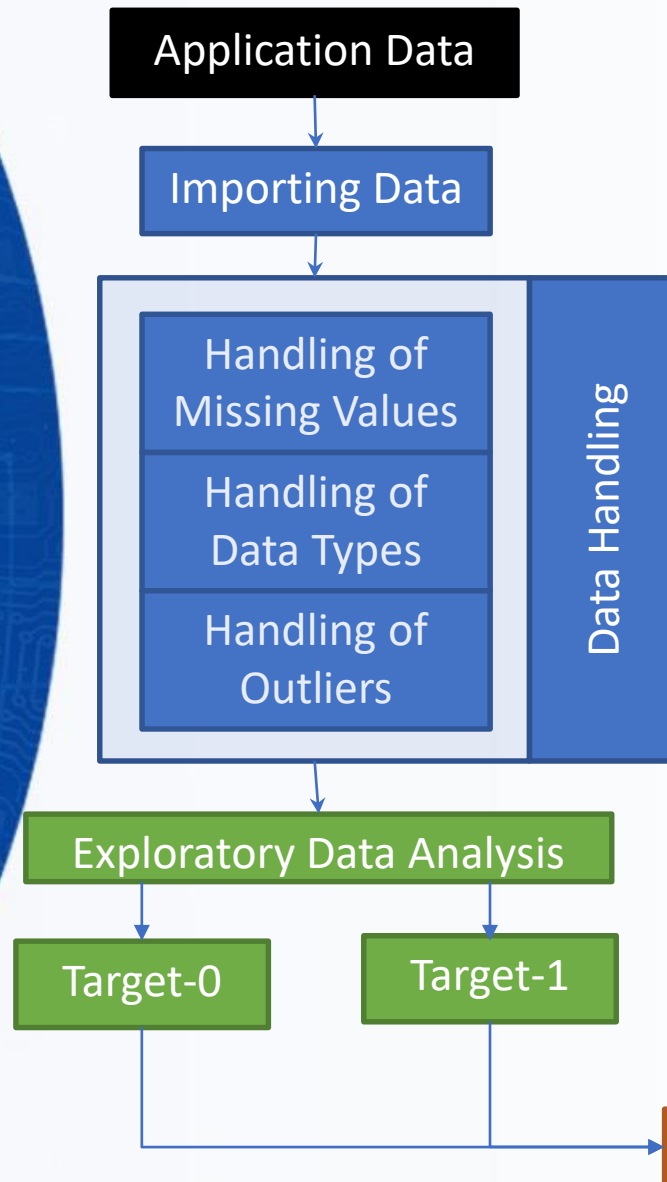
Ganesh Nagappa Shetty
Harish Dave



Introduction & Problem Statement

- **Case Study:** Exploratory Data Analysis on the Credit Applications of customers of a Finance Company
- **Objectives:** Identify patterns which indicate if a client has payment difficulty with their instalments. Understand the driving factors behind loan default, i.e. the variables that are strong indicators of default. Provide a set of insights which may then be used for actions such as denying the loan, reducing the amount of loan etc.
- **Input:** 2 Datasets in csv format + 1 variable description file
 - **Dataset-1:** application_data.csv (Target-1: client with payment difficulty and Target-0: all other cases)
 - **Dataset-2:** previous_application.csv
 - columns_description.csv
- **Output:** Jupyter notebook performing EDA + PowerPoint Report focussing on Insights drawn & Top-10 correlations

EDA Approach



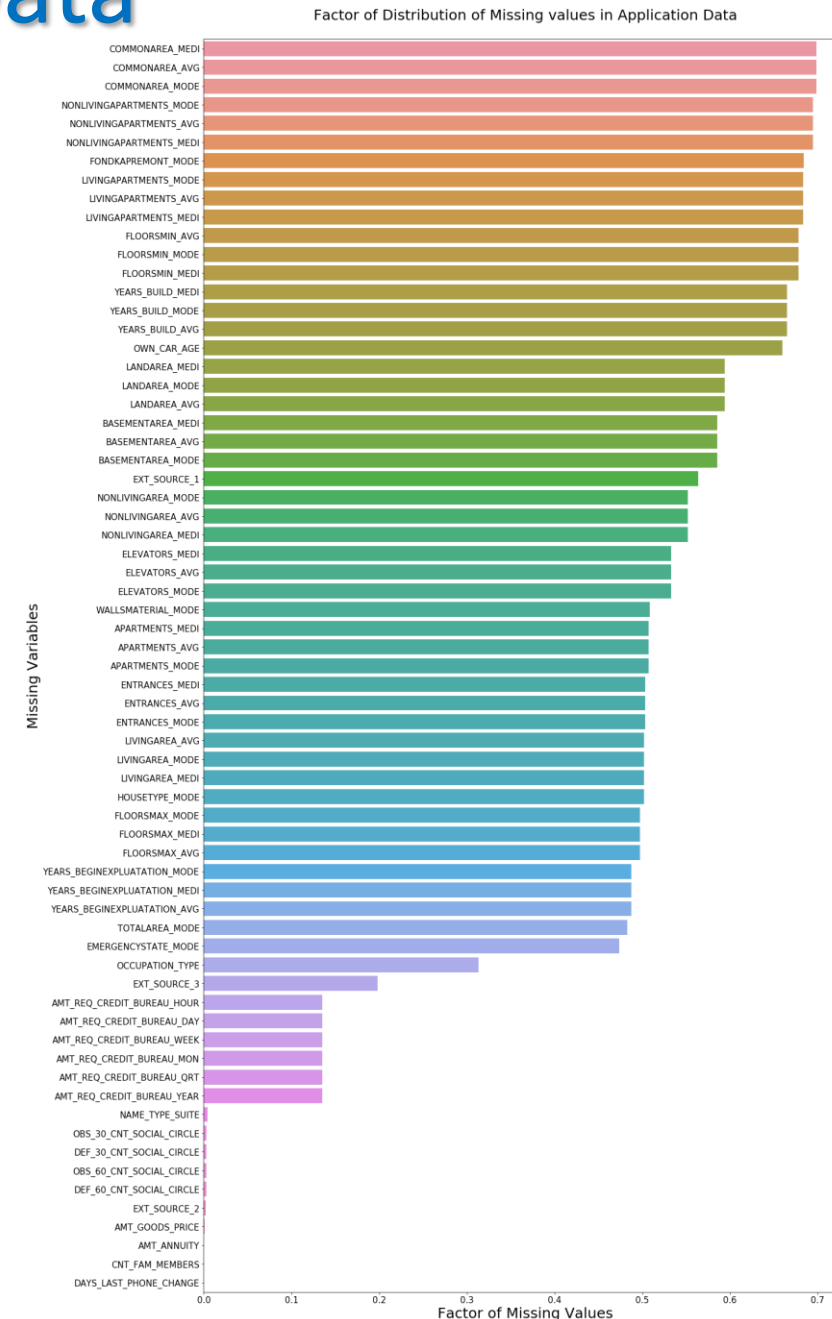
Data Handling – Applications Data

Filename: application_data.csv

Contains information about Loan application
Data of clients

307511 rows and 122 columns

- 41 columns which were identified with more than 50% missing values
- These variables are considered ineffective for data analysis and hence are not considered for further analysis
- Variable “CODE_GENDER” has missing value entered as “XNA”. These instances are dropped from further analysis



Data Handling – Applications Data

	% Missing Values
FLOORSMAX_AVG	49.760822
FLOORSMAX_MODE	49.760822
FLOORSMAX_MEDI	49.760822
YEARS_BEGINEXPLUATATION_AVG	48.781019
YEARS_BEGINEXPLUATATION_MODE	48.781019
YEARS_BEGINEXPLUATATION_MEDI	48.781019
TOTALAREA_MODE	48.268517
EMERGENCYSTATE_MODE	47.398304
OCCUPATION_TYPE	31.345545
EXT_SOURCE_3	19.825307
AMT_REQ_CREDIT_BUREAU_YEAR	13.501631
AMT_REQ_CREDIT_BUREAU_HOUR	13.501631
AMT_REQ_CREDIT_BUREAU_DAY	13.501631
AMT_REQ_CREDIT_BUREAU_WEEK	13.501631
AMT_REQ_CREDIT_BUREAU_MON	13.501631
AMT_REQ_CREDIT_BUREAU_QRT	13.501631
NAME_TYPE_SUITE	0.420148
DEF_30_CNT_SOCIAL_CIRCLE	0.332021
OBS_60_CNT_SOCIAL_CIRCLE	0.332021
DEF_60_CNT_SOCIAL_CIRCLE	0.332021
OBS_30_CNT_SOCIAL_CIRCLE	0.332021
EXT_SOURCE_2	0.214626
AMT_GOODS_PRICE	0.090403
AMT_ANNUITY	0.003902
CNT_FAM_MEMBERS	0.000650
DAYS_LAST_PHONE_CHANGE	0.000325

→ Replaced with mean of the data

→ Variable is Dropped from the analysis

→ Replaced with mean of the data

→ Variable is Dropped from the analysis

→ Replaced with mean of the data

→ Replaced with mode of the data

→ Replaced the missing values as “Unavailable”

→ Replaced with mode of the data

→ Replaced with median of the data

→ Imputed the missing value with mode - “Unaccompanied”

→ Replaced with median of the data

→ Replaced with mode of the data

→ Replaced with mean of the data

→ Missing Data is dropped from analysis

Handling of Data Types & Outliers – Application Data

DAYS_BIRTH: Converted to positive values

DAYS_EMPLOYED: Converted to positive values

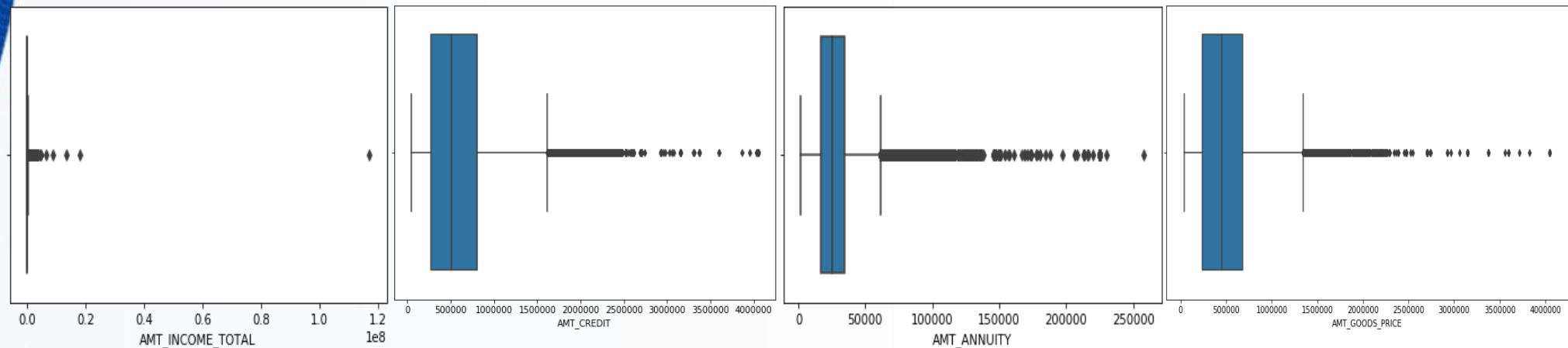
DAYS_REGISTRATION: Converted from Float Data Type to positive integer

DAYS_ID_PUBLISH: Converted to positive values

DAYS_LAST_PHONE_CHANGE: Converted from Float Data Type to positive integer

Handling of Outliers

Variables 'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUIITY' and 'AMT_GOODS_PRICE' have substantial outliers which were handled by binning of the data



Data Handling – Previous Dataframe

Filename: previous_application.csv

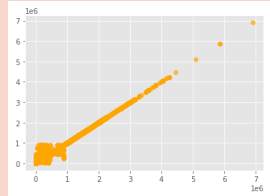
Contains information about previous Loan Data of clients

1.67 million rows and 37 columns

Big Dataset → step by step structured approach needed to navigate

Columns	Step Taken	Reasoning
RATE_INTEREST_PRIVILEGED RATE_INTEREST_PRIMARY	Drop Column	> 95% data missing
AMT_DOWN_PAYMENT RATE_DOWN_PAYMENT	Keep-Out Columns	> 50% Data missing (not dropped but kept them in seperate DF)
WEEKDAY_APPR_PROCESS_START HOUR_APPR_PROCESS_START FLAG_LAST_APPL_PER_CONTRACT NFLAG_LAST_APPL_IN_DAY	Not dropped but kept them in seperate DF	Not so relevant in current analysis
NFLAG_INSURED_ON_APPROVAL	Straightforward Impute NA = „Unknown“	Categorical Variables with Yes, No. So NA likey to mean not known.
DAYS_TERMINATION DAYS_LAST_DUE DAYS_LAST_DUE_1ST_VERSION DAYS_FIRST_DUE DAYS_FIRST_DRAWING	Sliced Impute Impute with Median taken from the respective column but computed after slicing based on the respective NAME_CONTRACT_TYPE	Example for DAYS_FIRST_DUE Grand Median: -831.0 Median when NAME_CONTRACT_TYPE = Cash loans: -708.0 Median when NAME_CONTRACT_TYPE = Consumer loans: -1065.0 Median when NAME_CONTRACT_TYPE = Revolving loans: -254.0 If we had replaced with grand median, it might have led to incorrect skew.

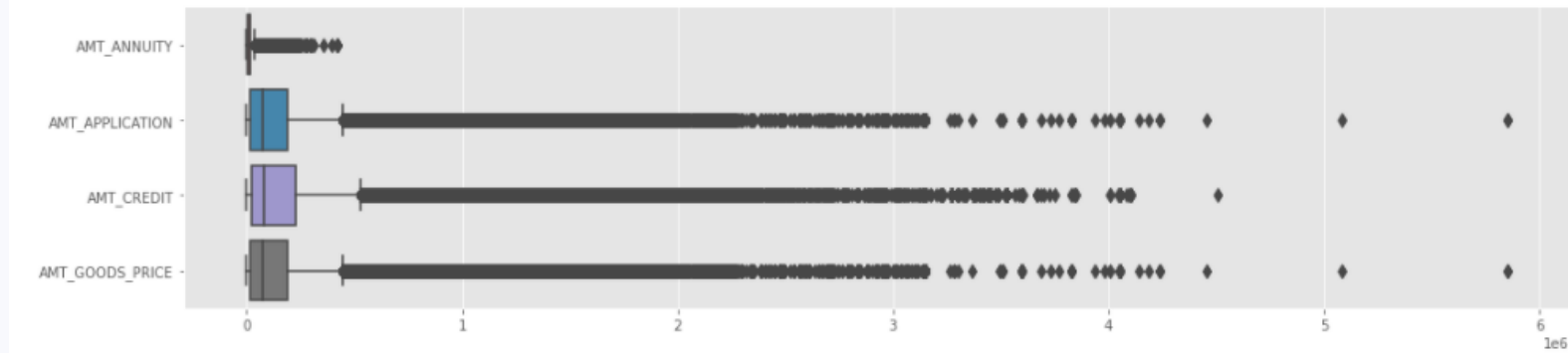
Data Handling – Previous Dataframe

Columns	Step Taken	Reasoning																
AMT_ANNUITY	Conditional Impute Impute with 0, if AMT_APPLICATION is 0	Majority of the missing values in AMT_ANNUITY have AMT_APPLICATION of 0. And strong correlation factor. <div><table><tr><th></th><th>AMT_ANNUITY</th><th>AMT_APPLICATION</th><th>AMT_CREDIT</th></tr><tr><th>AMT_ANNUITY</th><td>1.000000</td><td>0.808872</td><td>0.816429</td></tr><tr><th>AMT_APPLICATION</th><td>0.808872</td><td>1.000000</td><td>0.975822</td></tr><tr><th>AMT_CREDIT</th><td>0.816429</td><td>0.975822</td><td>1.000000</td></tr></table></div>		AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_ANNUITY	1.000000	0.808872	0.816429	AMT_APPLICATION	0.808872	1.000000	0.975822	AMT_CREDIT	0.816429	0.975822	1.000000
	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT															
AMT_ANNUITY	1.000000	0.808872	0.816429															
AMT_APPLICATION	0.808872	1.000000	0.975822															
AMT_CREDIT	0.816429	0.975822	1.000000															
AMT_CREDIT	Drop Row	Only one row missing																
AMT_GOODS_PRICE	Conditional Impute Impute with 0 if conditions in reasoning column met.	1. All missing values are coming from NAME_GOODS_CATEGORY XNA 2. All missing values have AMT_APPLICATION 0 3. AMT_APPLICATION and AMT_GOODS_PRICE have correlation factor 1																
CNT_PAYMENT	Conditional Impute Impute with 0 if conditions in reasoning column met.	If the AMT_APPLICATION & AMT_CREDIT is 0, then CNT_PAYMENT (term of loan repayment) is 0																
DAYS_FIRST_DRAWING	Drop Column Outlier Handling	DAYS_FIRST_DRAWING has 88% data which is 365243. Not reliable col.																
DAYS_LAST_DUE_1ST_VERSION DAYS_LAST_DUE	Drop Columns	3 Columns are correlated with factor 1. Keeping 1 of these 3 is sufficient. <div><table><tr><th></th><th>DAYS_LAST_DUE_1ST_VERSION</th><th>DAYS_LAST_DUE</th><th>DAYS_TERMINATION</th></tr><tr><th>DAYS_LAST_DUE_1ST_VERSION</th><td>1.0</td><td>1.0</td><td>1.0</td></tr><tr><th>DAYS_LAST_DUE</th><td>1.0</td><td>1.0</td><td>1.0</td></tr><tr><th>DAYS_TERMINATION</th><td>1.0</td><td>1.0</td><td>1.0</td></tr></table></div>		DAYS_LAST_DUE_1ST_VERSION	DAYS_LAST_DUE	DAYS_TERMINATION	DAYS_LAST_DUE_1ST_VERSION	1.0	1.0	1.0	DAYS_LAST_DUE	1.0	1.0	1.0	DAYS_TERMINATION	1.0	1.0	1.0
	DAYS_LAST_DUE_1ST_VERSION	DAYS_LAST_DUE	DAYS_TERMINATION															
DAYS_LAST_DUE_1ST_VERSION	1.0	1.0	1.0															
DAYS_LAST_DUE	1.0	1.0	1.0															
DAYS_TERMINATION	1.0	1.0	1.0															

Various data handling strategies applied
From 1.67 million rows and 37 columns
We now have 1.64 milliion rows and 26 columns

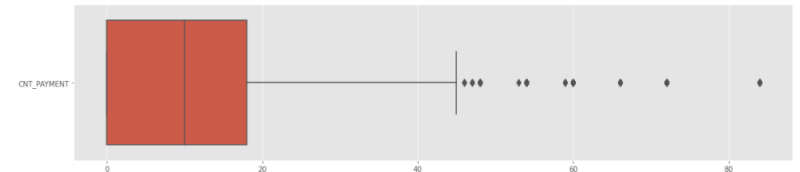
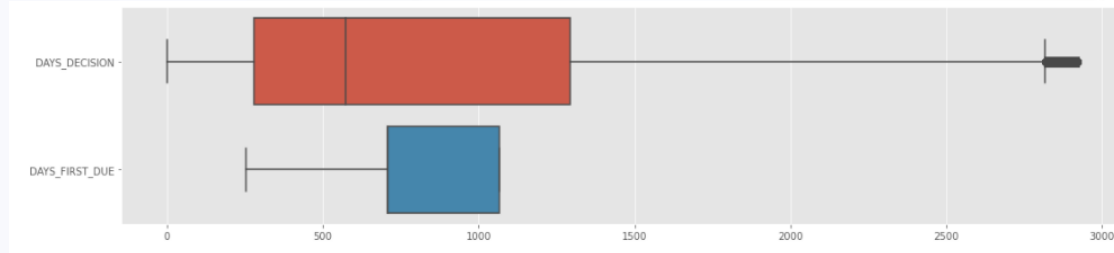
Data Handling – Previous Dataframe

Outlier Handling → 4 Numeric Variables with Amounts are binned



4 New columns added due to category & binning.

Outlier Handling → Few columns had no outliers, no action needed



Data Type Handling → Columns with 'DAYS_' data Type converted to INT

Data Leakage Analysis → Analysed by comparing the common rows in the two provided datasets based on column SK_ID_CURR:

- Received Data = 94.65% SK_ID_CURR common
- Cleaned Data = 94.6% SK_ID_CURR common

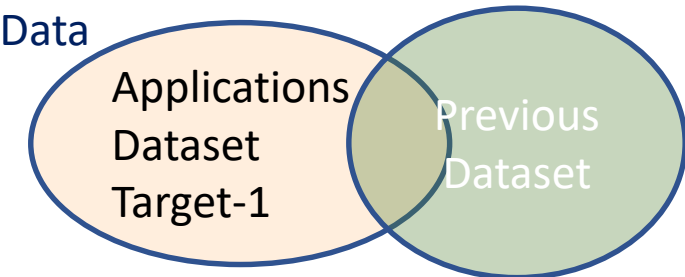
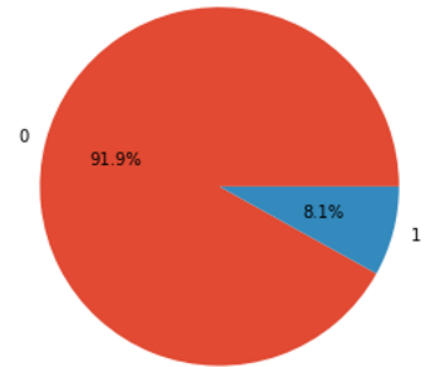
Data Analysis

- Data Analysis is performed on the selected variables of the application data based on the judgement of usefulness of data
- The data is further segmented into 2 DF: One each of the Target Classes 0 and 1
- Categorical Variables are further segmented into Ordered and Unordered variables
- On these data Categorical Univariate Analysis has been performed
- On numerical Variables Univariate, Bivariate and Multivariate Analysis has been performed

Data imbalance in favor of TARGET-0

- Only 8.1% of the applications data is for client with Target-1
 - Only 7.3% of previous data is for SK_ID_CURR corresponding to Target-1.
 - Conclusions on Target-1 should be interpreted carefully
 - Data imbalance needs to be investigated during machine learning stage
- Due to large dataset, the data analysis of Previous Dataframe done in **2 separate steps**:
 1. **Sliced** with SK_ID_CURR for Target-1 (see venn diagram) & used full dataset without sampling
 2. **Merged** with random sample of 3000 rows from Applications Data
 - Univariate Analysis with bar charts
 - Bi-variate Analysis with scatter plots & heatmap

Balance of Target Class



Insight – 1

The following appear more frequent or popular based on applications data. This is reported in case the bank wants to target clients to increase loan applications:

Cash Loans

Female applicants

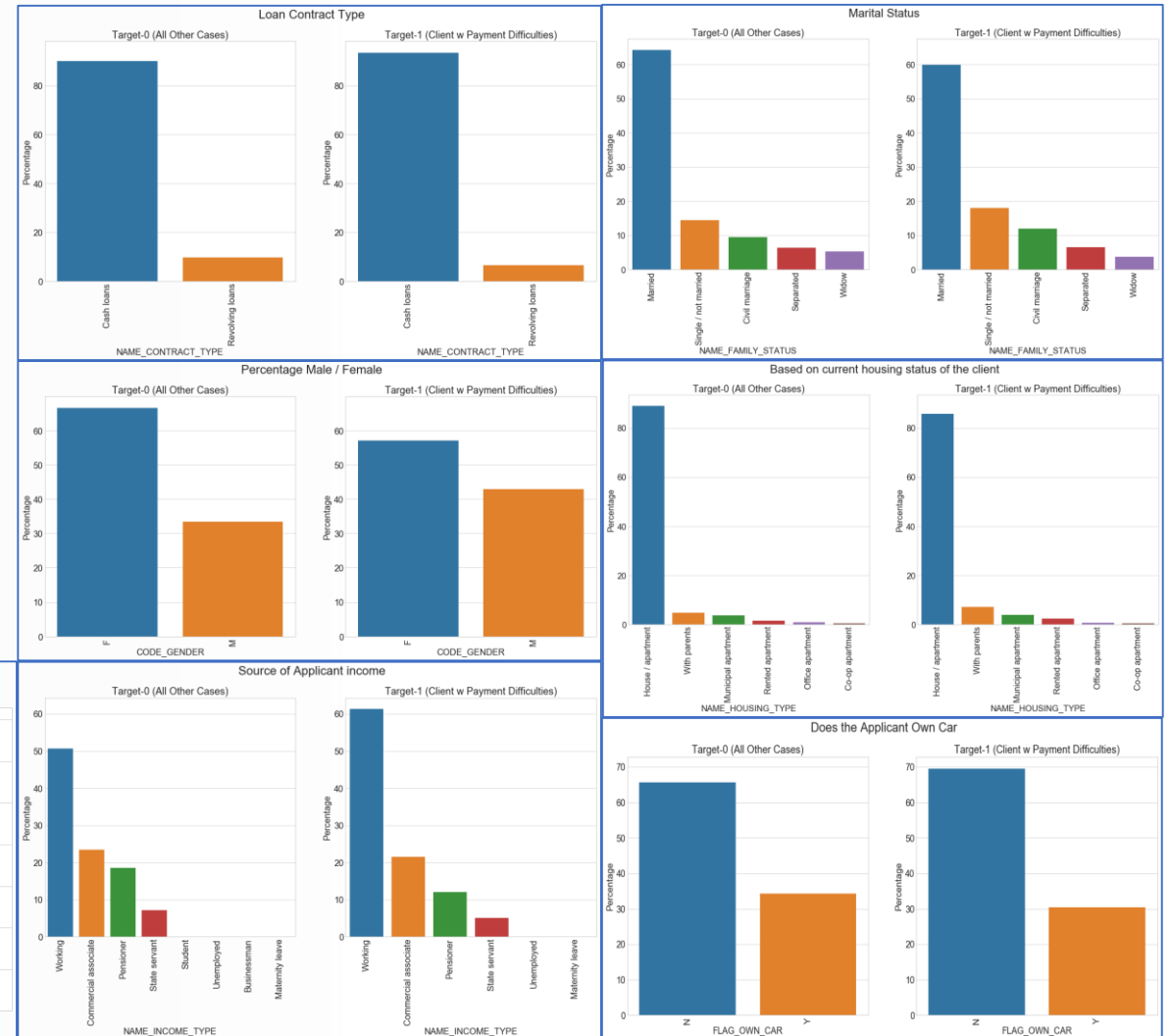
People working

Married

Not living with parents

Not owning car

Owning real estate



Insight - 1 Contd..

The following appear more frequent or popular based on applications data. This is reported in case the bank wants to target clients to increase loan applications:

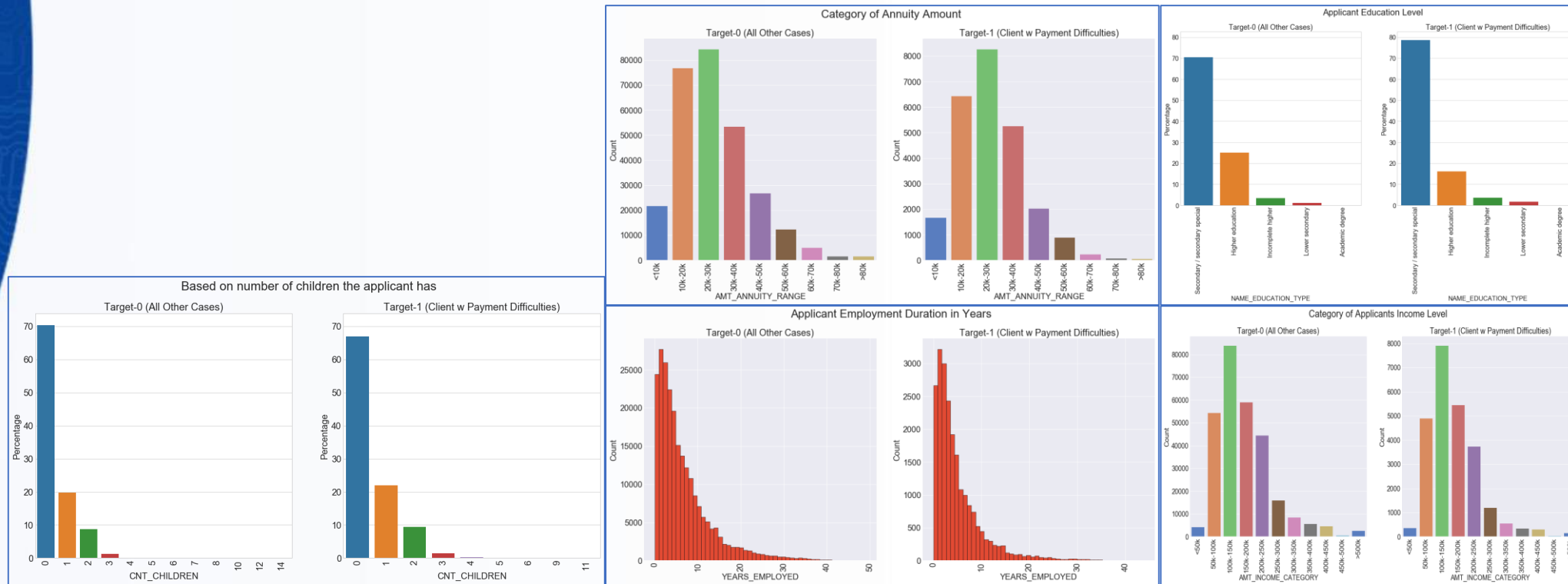
People without children

Secondary Education level

People with income below 200k

People with annuity below 30k

Employment duration less than 5 years





Insight – 2 & 3

Univariate analysis of application data suggests that Target-1 has increasing trend when compared to Target-0 in terms of % (not favorable to bank, but doesn't mean these category will default !).

Male applicant (despite median income of male applicants being slightly higher compared to female applicants)

Working people

Laborers

People who are single

People who live with parents

People with secondary education level

Credit amount in the range of 500k-600k range

Goods Price in the range 500k

People in the age group 28 to 40 years (about 30 years)

Univariate analysis of application data suggests that Target-1 has decreasing trend when compared to Target-0 in terms of % (favorable to bank, but doesn't mean this category will not default !).

Female applicant

People who are pensioners or state servants

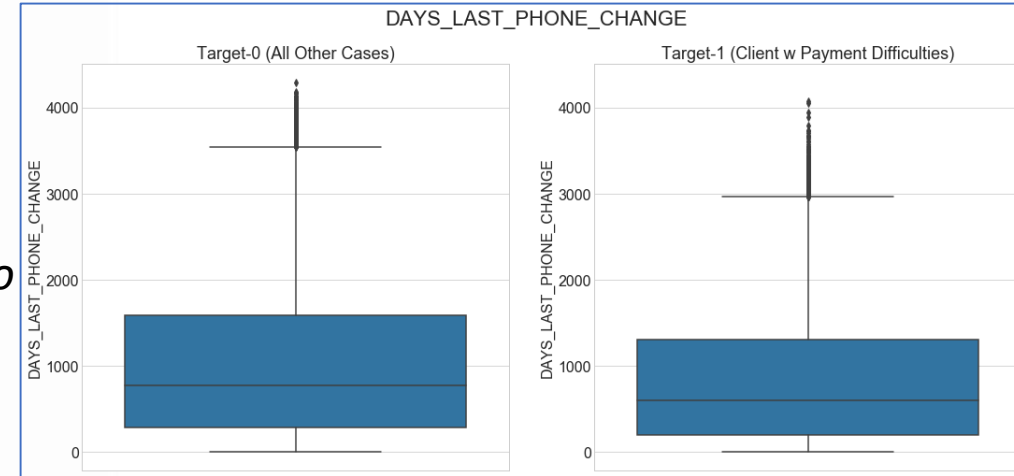
Widowers

People with higher education

Insight 4 & 5

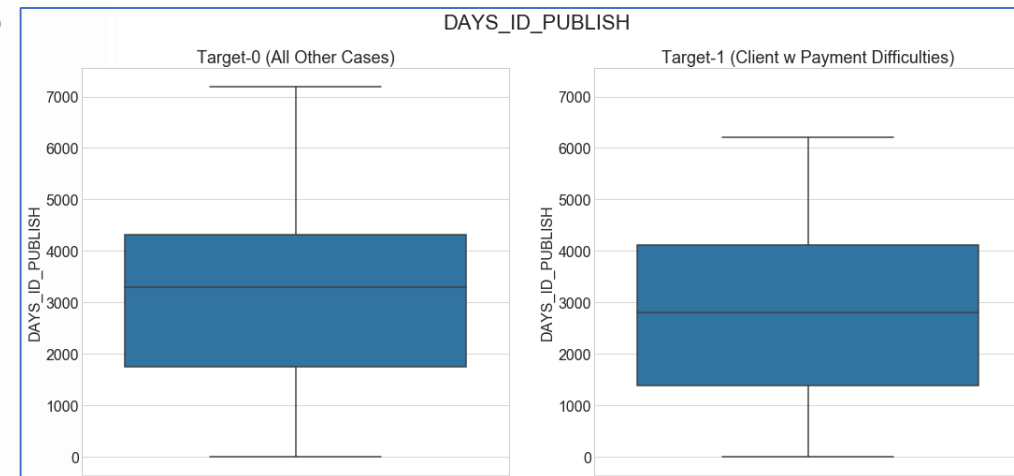
DAYS_LAST_PHONE_CHANGE has a trend with Target-1. This is the parameter that says how many days before application did client change phone ?

Median value reduces from 3300 days (target-0) to 2800 days (target-1)



DAYS_ID_PUBLISH has a trend with Target-1. This is the parameter that says how many days before the application did client change the identity document with which he applied for the loan ?

Median value reduces from 780 days (target-0) to less than 600 days (target-1)



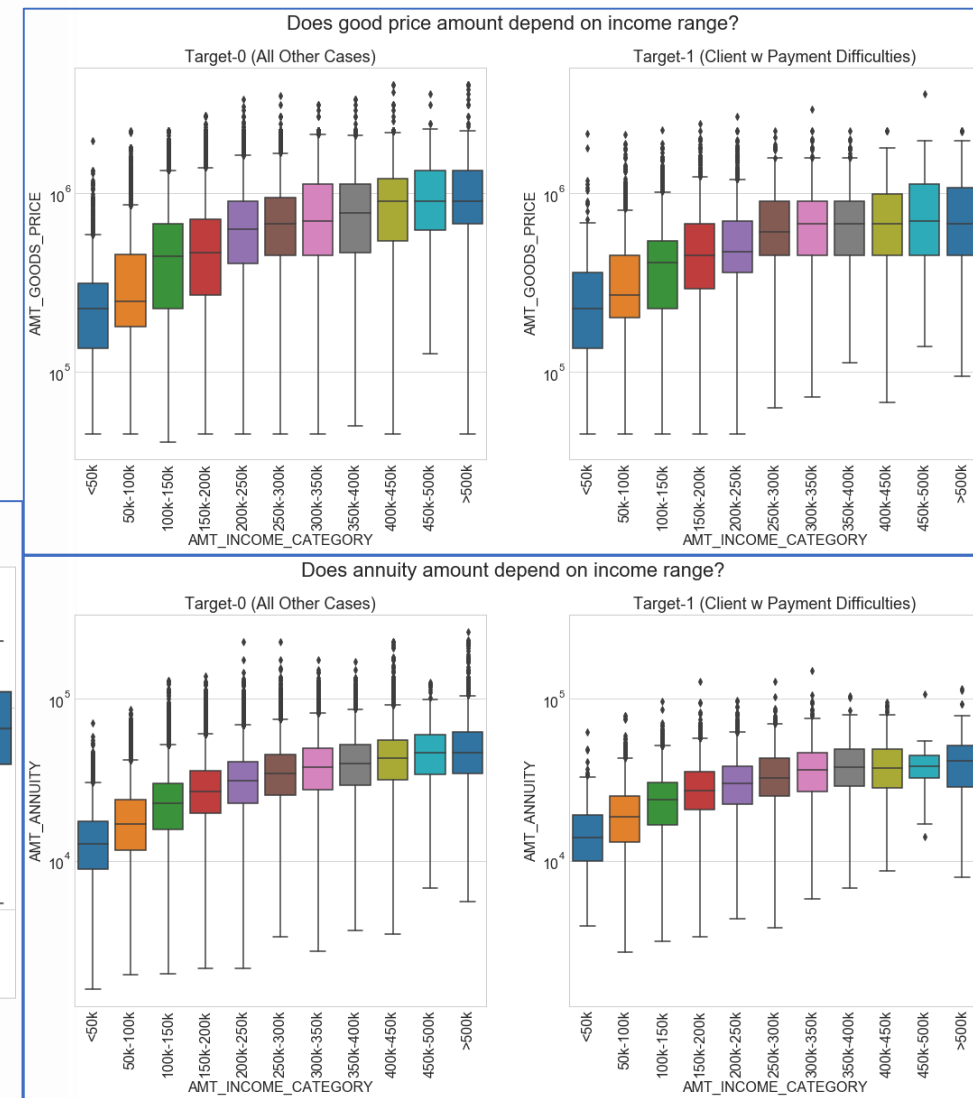
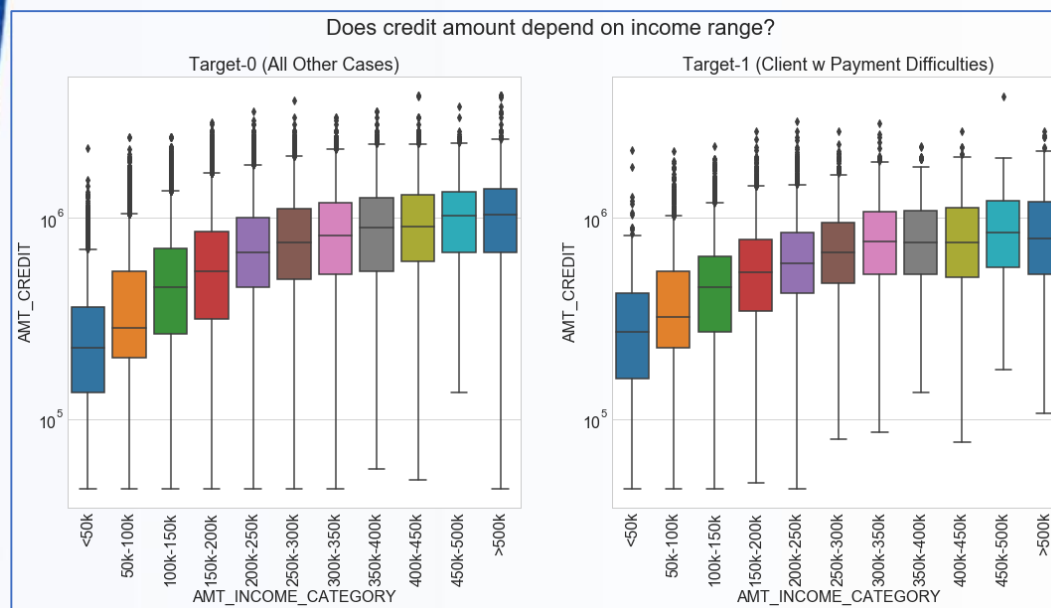
Insight - 6

Good Price, Annuity Amount and Credit Amount have following trend to income range:

Median value of all these variables are increasing with income range, no matter for target0 or target1

Interesting trend is: for target1 the median value of goods price, annuity and credit amount is lower than target0 across all the income ranges

Insight is: people taking lower credit amount are at higher chance to default compared to those taking higher amounts of loan

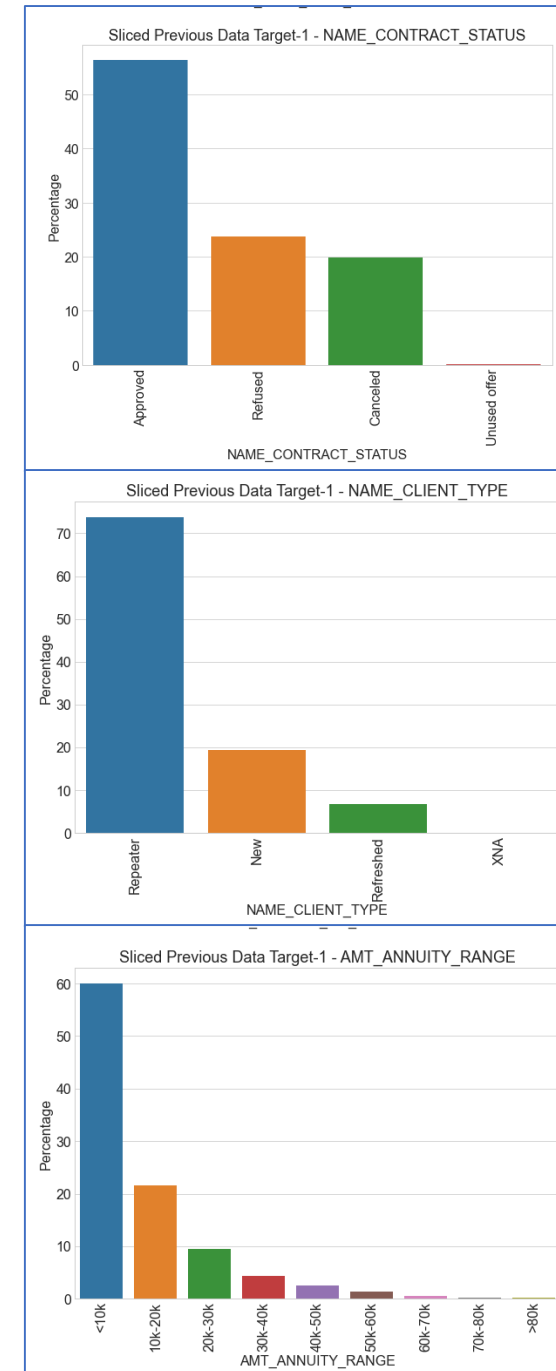


Insight - 7

Previous applications data can be an useful resource for the bank. Therefore univariate analysis was performed for the previous application data sliced based on SK_ID_CURR belonging only to target-1 from applications data. This gives sort of a previous history of target-1 clients with payment difficulty.

Target-1 clients in the past have:

- a) more often applied for cash loans or consumer loans.
- b) more often not mentioned loan application purpose and if recorded it was for repairs.
- c) a loan approval rate of 55% and refusal rate less than 25% (rejection reason XAP).
- d) been repeat applicant (strong indicator with greater than 70% instances !).
- e) often purchased mobile or consumer electronics with loan amount.
- f) 45% of times approached via the credit and cash offices.
- g) been in high or middle yield group.
- h) often opted for top-3 product combination: Cash, POS Household, Pos Mobile.
- i) less than 10k of annuity.
- j) applied for less than 200k of credit amount.



Insight - 8

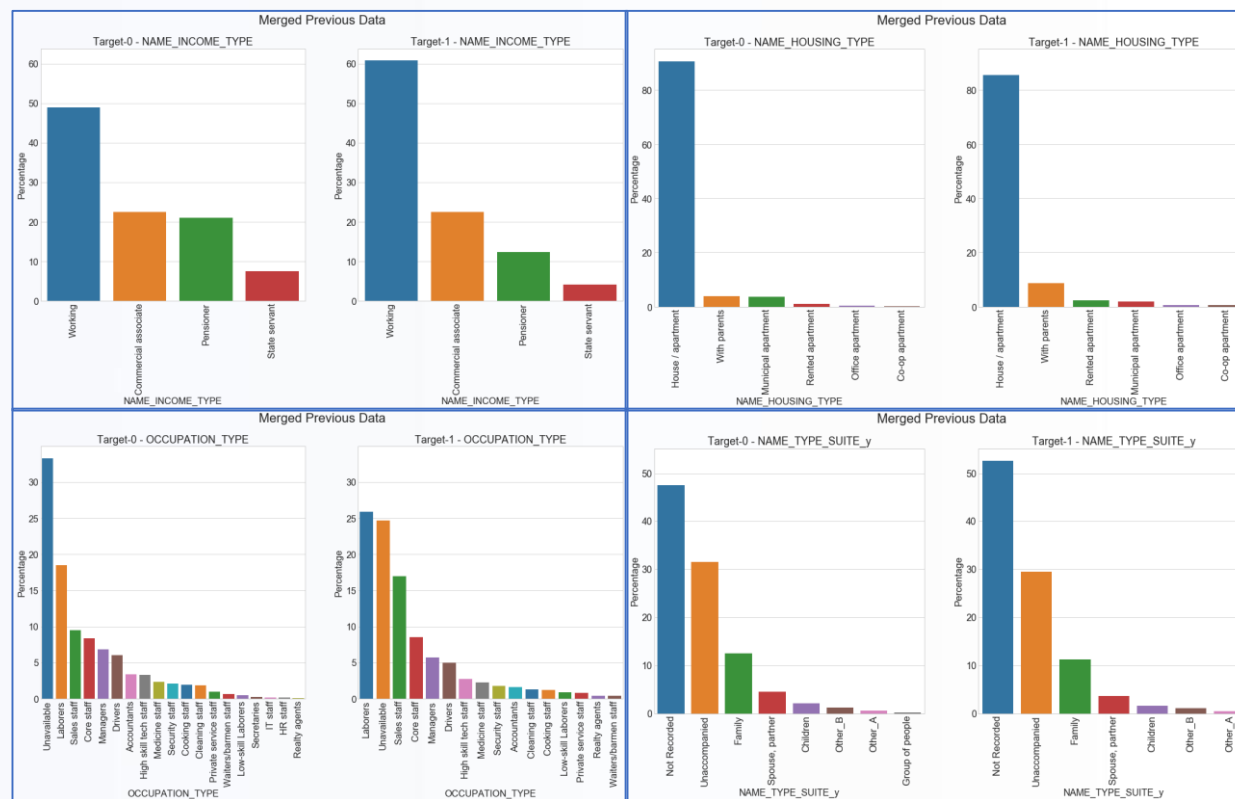
Univariate analysis of merged previous data suggests that Target-1 has increasing trend when compared to Target-0 in terms of % (not favorable to bank, but doesn't mean these category will default !).

Working people

Live with parents*

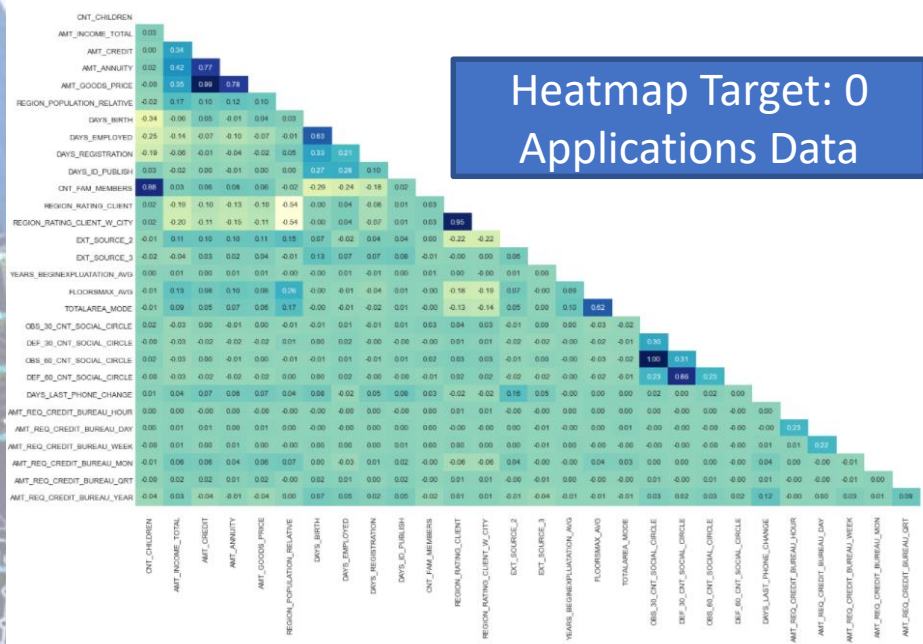
Laborers*

Less often the clients have been accompanied by family while applying for loan*



Insight - 9 & 10

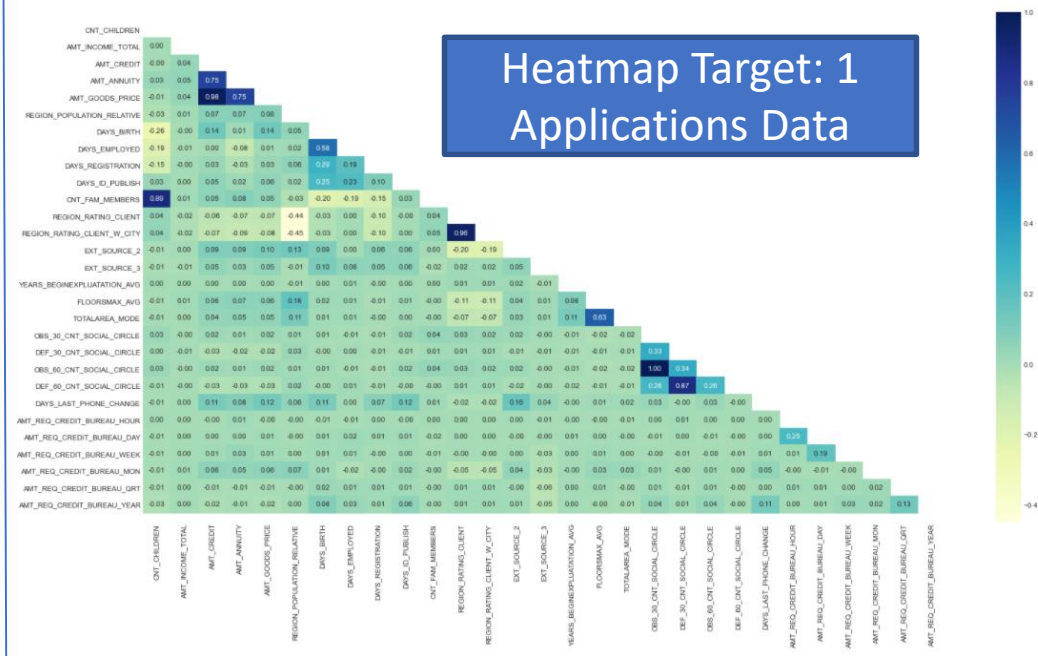
Application Dataframe Heatmap : Target-0



Insight#9: Top-15 Correlations for Target-0

OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998390
AMT_CREDIT	AMT_GOODS_PRICE	0.986882
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950148
CNT_CHILDREN	CNT_FAM_MEMBERS	0.878570
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.856848
AMT_GOODS_PRICE	AMT_ANNUITY	0.776252
AMT_ANNUITY	AMT_CREDIT	0.771296
DAYS_BIRTH	DAYS_EMPLOYED	0.626116
FLOORS_MAX_AVG	TOTALAREA_MODE	0.623806
REGION_POPULATION_RELATIVE	REGION_RATING_CLIENT	0.539015
REGION_RATING_CLIENT_W_CITY	REGION_POPULATION_RELATIVE	0.537312
AMT_INCOME_TOTAL	AMT_ANNUITY	0.418954
AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349367
AMT_CREDIT	AMT_INCOME_TOTAL	0.342805
DAYS_BIRTH	CNT_CHILDREN	0.336980

Application Dataframe Heatmap : Target-1



Insight#10: Top-15 Correlations for Target-1

OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998270
AMT_GOODS_PRICE	AMT_CREDIT	0.982566
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.956637
CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.869016
AMT_CREDIT	AMT_ANNUITY	0.752195
AMT_GOODS_PRICE	AMT_ANNUITY	0.752022
TOTALAREA_MODE	FLOORS_MAX_AVG	0.634193
DAYS_BIRTH	DAYS_EMPLOYED	0.582185
REGION_RATING_CLIENT_W_CITY	REGION_POPULATION_RELATIVE	0.446977
REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	0.443236
OBS_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.337389
OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.334035
DAYS_REGISTRATION	DAYS_BIRTH	0.289114
OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.264357

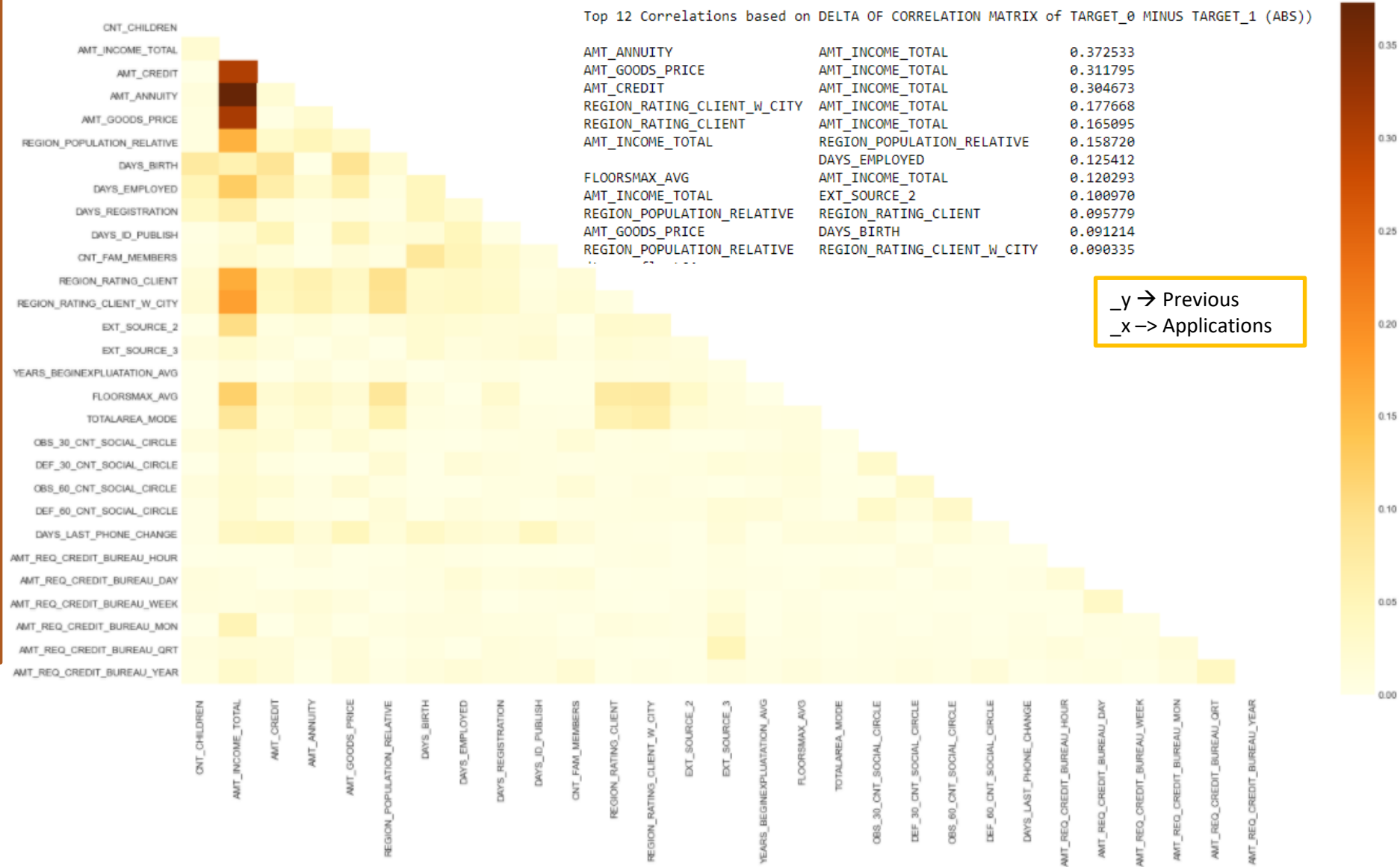
Insights - 11

Insight#12: There is change in correlation factor i.e. $\Delta(\text{Corr}_{\text{Target0}} - \text{Corr}_{\text{Target1}})$:

Total income and few parameters shown in table have different correlation value based on Target-0 or Target-1.

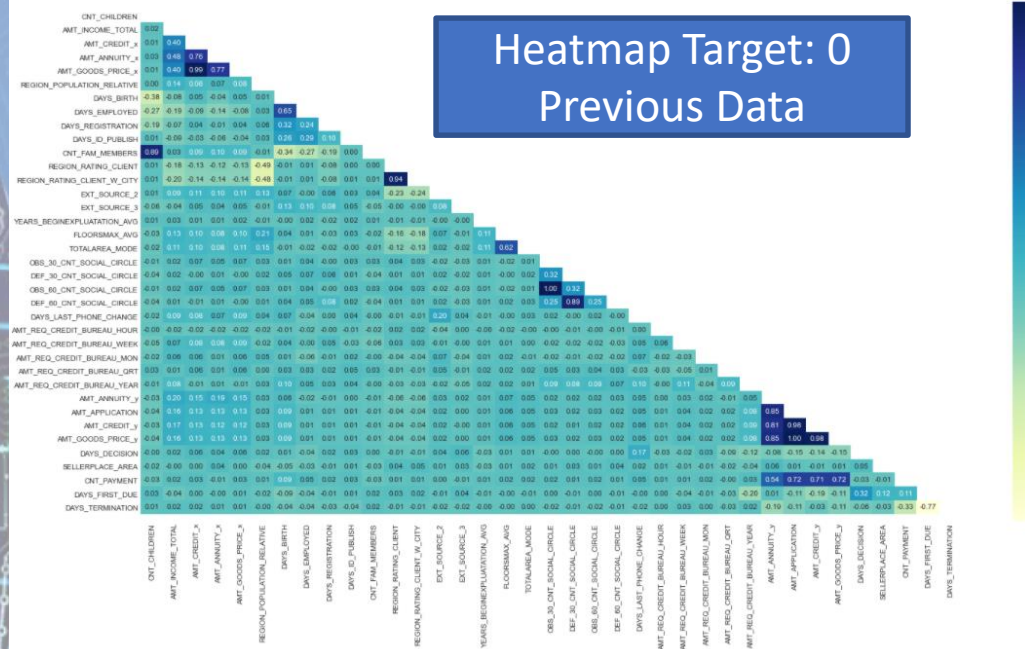
This can be useful to detect possible default and as differentiators.

Applications Dataframe Heatmap with Delta of Target0 minus Target-1



Insights - 12 & 13

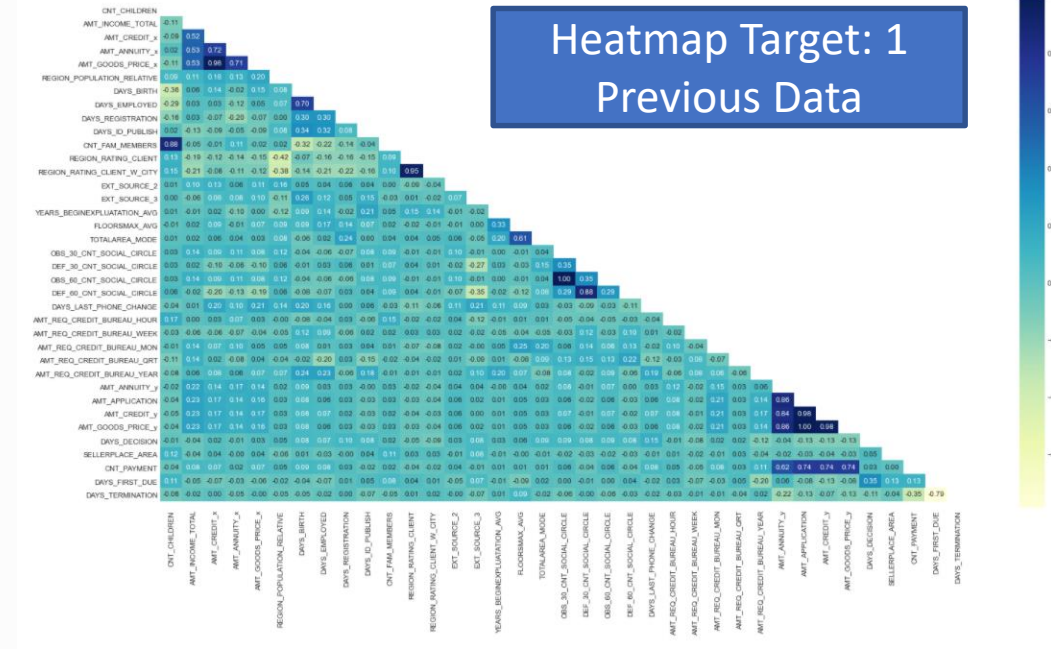
Previous Dataframe Heatmap (Merged with Sample from Applications): Target-0



Insight#12: Top-15 Correlations for Target-0, Previous Data

AMT_APPLICATION	AMT_GOODS_PRICE_y	0.999999
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998656
AMT_CREDIT_x	AMT_GOODS_PRICE_x	0.987874
AMT_GOODS_PRICE_y	AMT_CREDIT_y	0.975444
AMT_APPLICATION	AMT_CREDIT_y	0.975442
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.941436
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.893904
CNT_FAM_MEMBERS	CNT_CHILDREN	0.892506
AMT_GOODS_PRICE_y	AMT_ANNUITY_y	0.848116
AMT_APPLICATION	AMT_ANNUITY_y	0.848114
AMT_ANNUITY_y	AMT_CREDIT_y	0.814335
DAYS_FIRST_DUE	DAYS_TERMINATION	0.767455
AMT_ANNUITY_x	AMT_GOODS_PRICE_x	0.766560
AMT_CREDIT_x	AMT_ANNUITY_x	0.764688
CNT_PAYMENT	AMT_GOODS_PRICE_y	0.720503

Previous Dataframe Heatmap (Merged with Sample from Applications): Target-1



Insight#13: Top-15 Correlations for Target-1, Previous Data

OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.999124
AMT_CREDIT_x	AMT_GOODS_PRICE_x	0.984914
AMT_GOODS_PRICE_y	AMT_CREDIT_y	0.982010
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950673
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.882320
CNT_FAM_MEMBERS	CNT_CHILDREN	0.880135
AMT_ANNUITY_y	AMT_APPLICATION	0.860208
	AMT_CREDIT_y	0.836007
DAYS_TERMINATION	DAYS_FIRST_DUE	0.787395
AMT_CREDIT_y	CNT_PAYMENT	0.738365
AMT_GOODS_PRICE_y	CNT_PAYMENT	0.735742
AMT_CREDIT_x	AMT_ANNUITY_x	0.720493
AMT_GOODS_PRICE_x	AMT_ANNUITY_x	0.706889
DAYS_EMPLOYED	DAYS_BIRTH	0.703710
CNT_PAYMENT	AMT_ANNUITY_y	0.623280

_y → Previous
_x → Applications

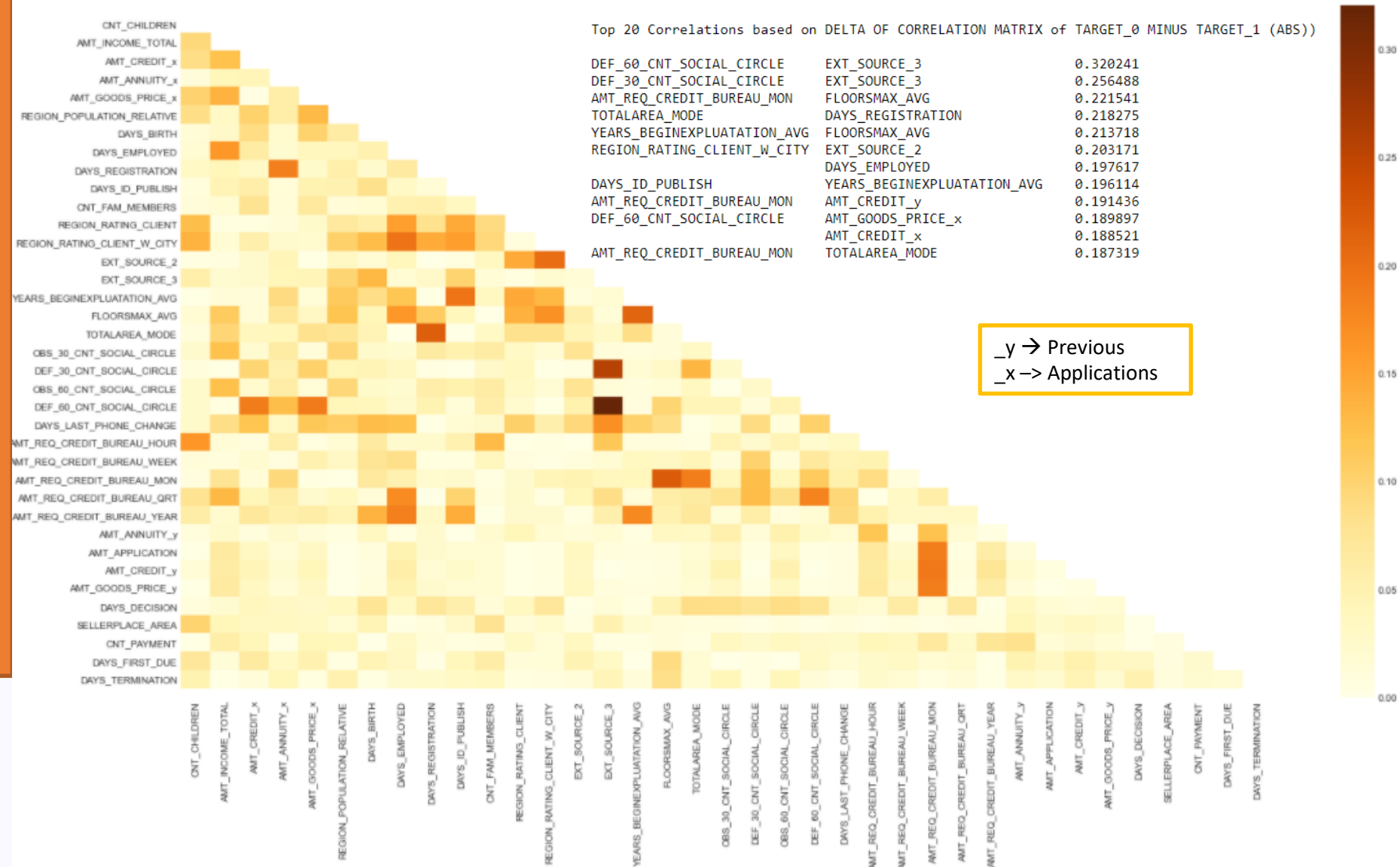
Insight - 14

Insight#15: There is change in correlation factor i.e. $\Delta(\text{Corr}_{\text{Target0}} - \text{Corr}_{\text{Target1}})$:

Good Price & Social Circle Observations have different correlation value.

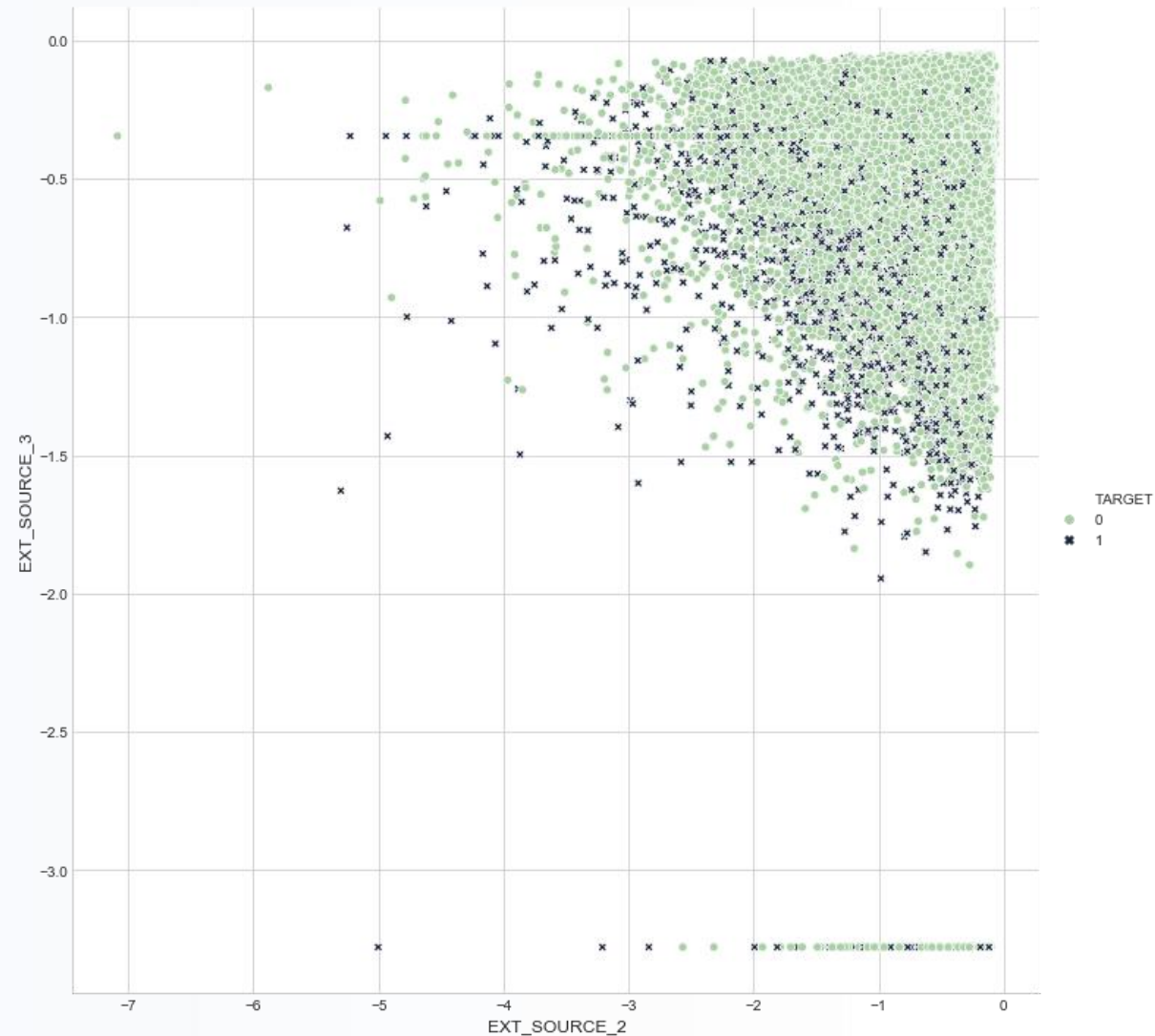
This can be useful to detect possible default and as differentiators.

Merged Previous & Applications Dataframe Heatmap with Delta of Target0 minus Target-1



Insight - 15

Bivariate analysis of Scores from External Sources shows following trend :
Clients with higher negative scores from external sources are more likely to default





Thank you

*Ganesh Nagappa Shetty
Harish Dave*