

# EDA group Case Study

UTKARSH KUMAR

ASHISH JOSHI

# Problem Statement

- ▶ This case study aims to identify patterns which indicate if a client has difficulty paying their installments 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.
- ▶ In other words, 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 utilise this knowledge for its portfolio and risk assessment.

# The DATA

- ▶ The data set provided contains a set of loan applicants with other data about the applicants like: gender, age, income, family size, and so on.
- ▶ Specifically the column/Variable “Target” in the data set identifies applicants with payment difficulties with 1 and other applicants as 0. The variable is binary coded.


# Our Approach

- ▶ Our first step was to look at the data dictionary and see what columns are in the data set and decide what columns would be irrelevant with respect to our analysis.
- ▶ For example: columns with details on the documents submitted were rendered irrelevant to the analysis.
- ▶ We decided that the simplest approach to reach the objective would be to do a comparative analysis of different variables

# Data Cleaning – we retained the following variables

SK\_ID\_CURR,  
TARGET,  
NAME\_CONTRACT\_TYPE,  
CODE\_GENDER,  
FLAG\_OWN\_CAR,  
FLAG\_OWN\_REALTY,  
CNT\_CHILDREN,  
AMT\_INCOME\_TOTAL,  
AMT\_CREDIT,  
AMT\_ANNUITY,  
AMT\_GOODS\_PRICE,  
NAME\_INCOME\_TYPE,

NAME\_EDUCATION\_TYPE,  
NAME\_FAMILY\_STATUS,  
NAME\_HOUSING\_TYPE,  
DAYS\_BIRTH,  
OCCUPATION\_TYPE,  
CNT\_FAM\_MEMBERS,  
ORGANIZATION\_TYPE,  
OBS\_30\_CNT\_SOCIAL\_CIRCLE,  
DEF\_30\_CNT\_SOCIAL\_CIRCLE,  
OBS\_60\_CNT\_SOCIAL\_CIRCLE,  
DEF\_60\_CNT\_SOCIAL\_CIRCLE

- 
- ▶ We further separated the Target = 1 and Target = 0, data points into different data frames.
  - ▶ We found the data imbalance ratio to be 10.55
  - ▶ Notes: When the observation in different classes exist, that is data imbalance. In our case, Target = 0 was a larger data set than Target = 1

# Outlier/Multivariate analysis

- ▶ We used pair plotting to look at how the data from various columns vary with each other. We did this for Target = 0 as well as Target = 1
- ▶ From this analysis we realized that the outliers are at a minimum. Had the outliers been significant, we would have probably treated the outliers as necessary by doing one of the following: deleting the outlier data, looking at the cause of the data being an outlier and then treating accordingly.

# Correlation

- ▶ Pearson Correlation : Pearson correlation measure strength of linear relationship.
- ▶ Spearman Correlation: Spearman correlation measures the strength and direction of a monotonic relationship.
- ▶ For our analysis we have used spearman correlation to gain better insights of variable interaction with each other.



# Correlation continued: comparison

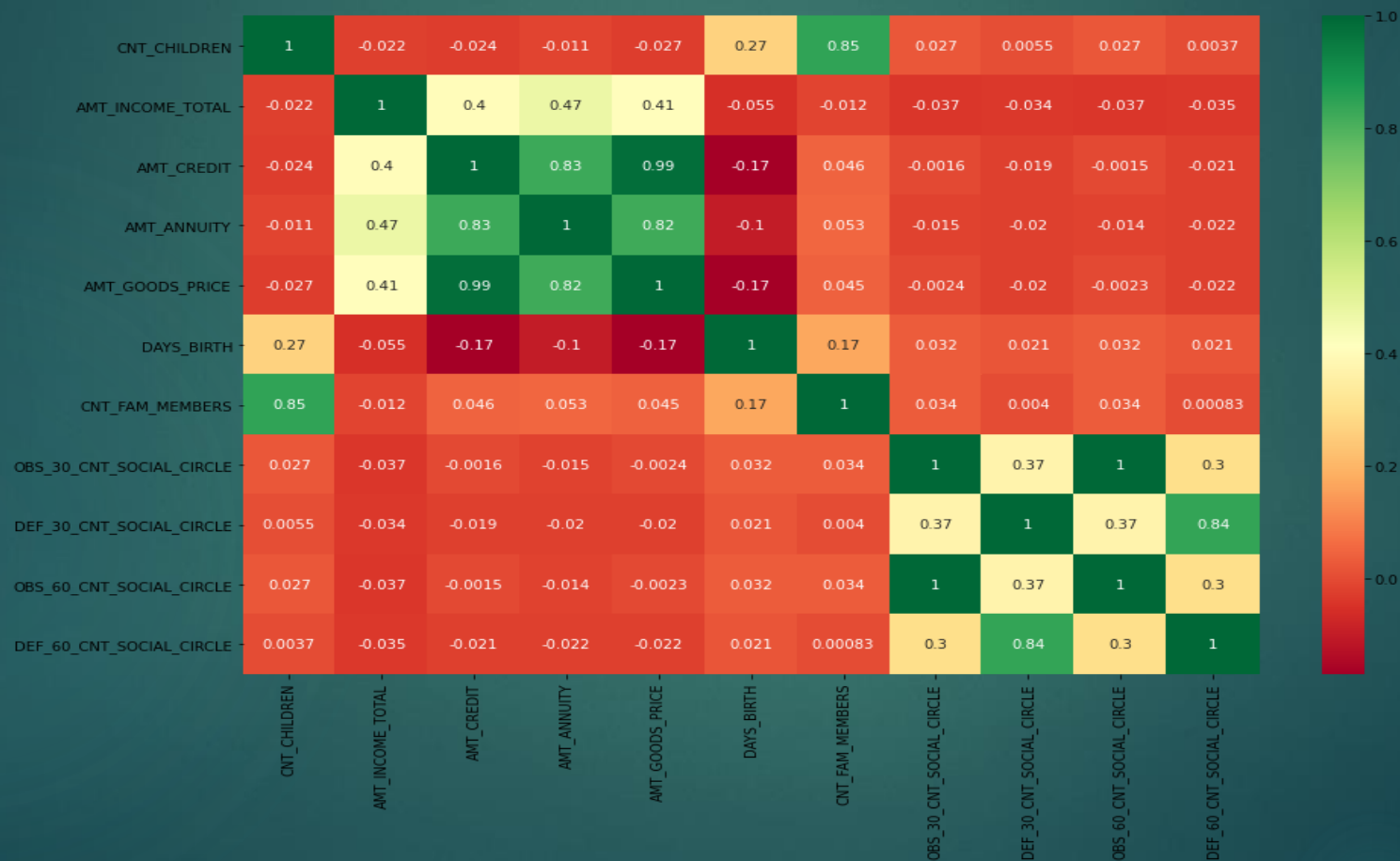
Variables	Target = 0	Target = 1
income/credit	0.403876	0.364559
income/annuity	0.472217	0.428947
income/goods amout	0.408533	0.369419
age/income	-0.05467	-0.10303
age/credit	-0.16903	-0.20072
age/annuity	-0.10029	-0.1002
age/faamily count	0.172363	0.080812

The table shows the values of correlation for different pairs for both the target variables. It is found that:

- Lower the correlation between income and credit, greater are the chances of difficulties in paying back the loan
- Lower the correlation between income and annuity, greater are the chances of difficult in paying the loan
- Lower the correlation between incme and good price, higher the chances of difficulties in paying back the loan
- Lower the correlation between age and income higher the chances of difficulties in paying back the loan

# Heat maps for Correlation: Target = 0

Correlation for target 0



# Heat maps for correlation: Target = 1

Correlation for target 1

