EDA ASSIGNMENT REPORT

HOANG NGOC TIEN – MASTER OF DATA IN DATA SCIENCE PROGRAME

PROBLEM

Based on consumer's application data, using EDA to identify patterns which indicate if a client has difficulty paying their instalments (is likely to default). In other words, using EDA to find out the driving factors (or driver variables) which lead to loan default.

ANALYSIS APPROACH

Overall approach:

- **Data understanding and preparation:**
  - Familiarize yourself with the dataset's structure, variables and their meaning, target variable
  - Handle missing values: Impute missing values using appropriate methods.
- **Exploratory Data Analysis (EDA)**
  - Visualize the distribution of the target variable: Understand the proportion of defaulted vs. non-defaulted loans
  - Examine distribution of features: Analyze the distributions of numerical and categorical features with respect to default status.
  - Identify outliers and anomalies.
  - Explore correlations: Check correlations between variables in each target variable’s segment.
  - Identify trends: Observe any patterns or trends that might be associated with default behavior

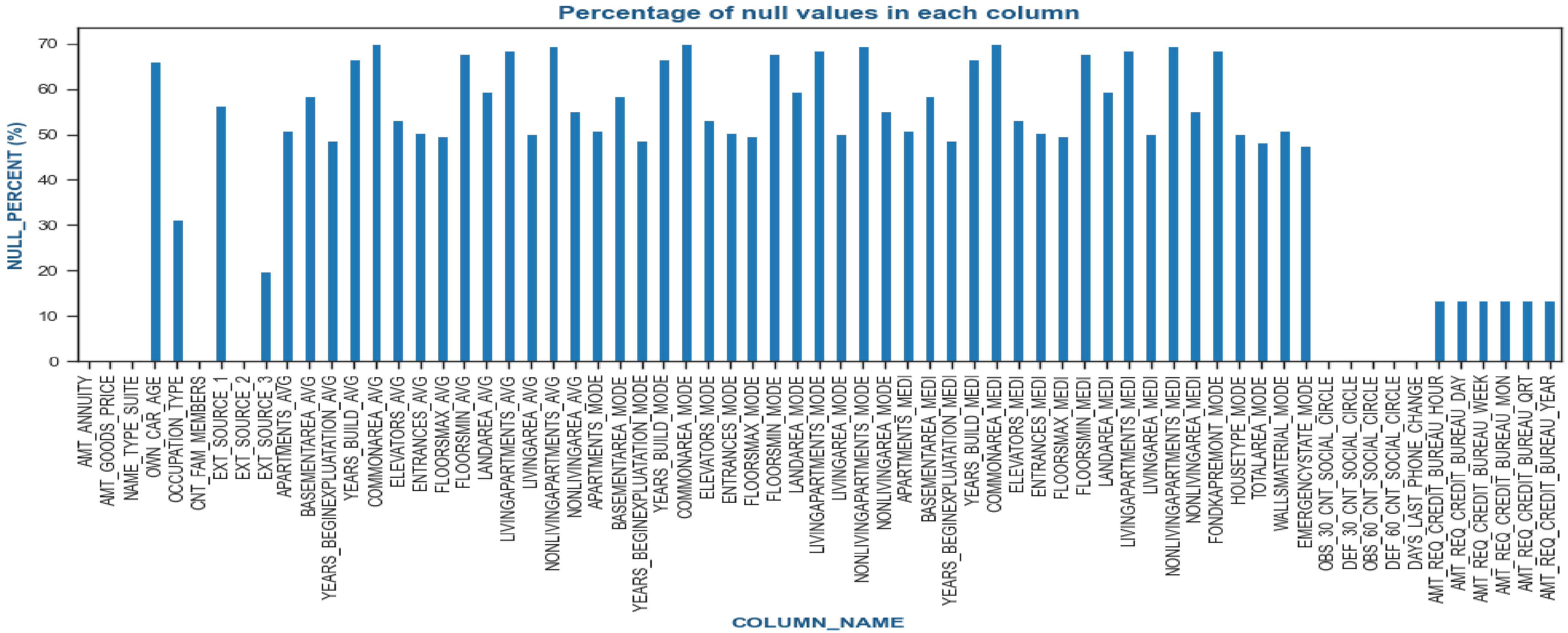
DATA UNDERSTANDING AND PREPARATION

Examine Datasets structure:

- Structure of each dataset
- Overview data in each column
- Identify column’s datatype and variable type

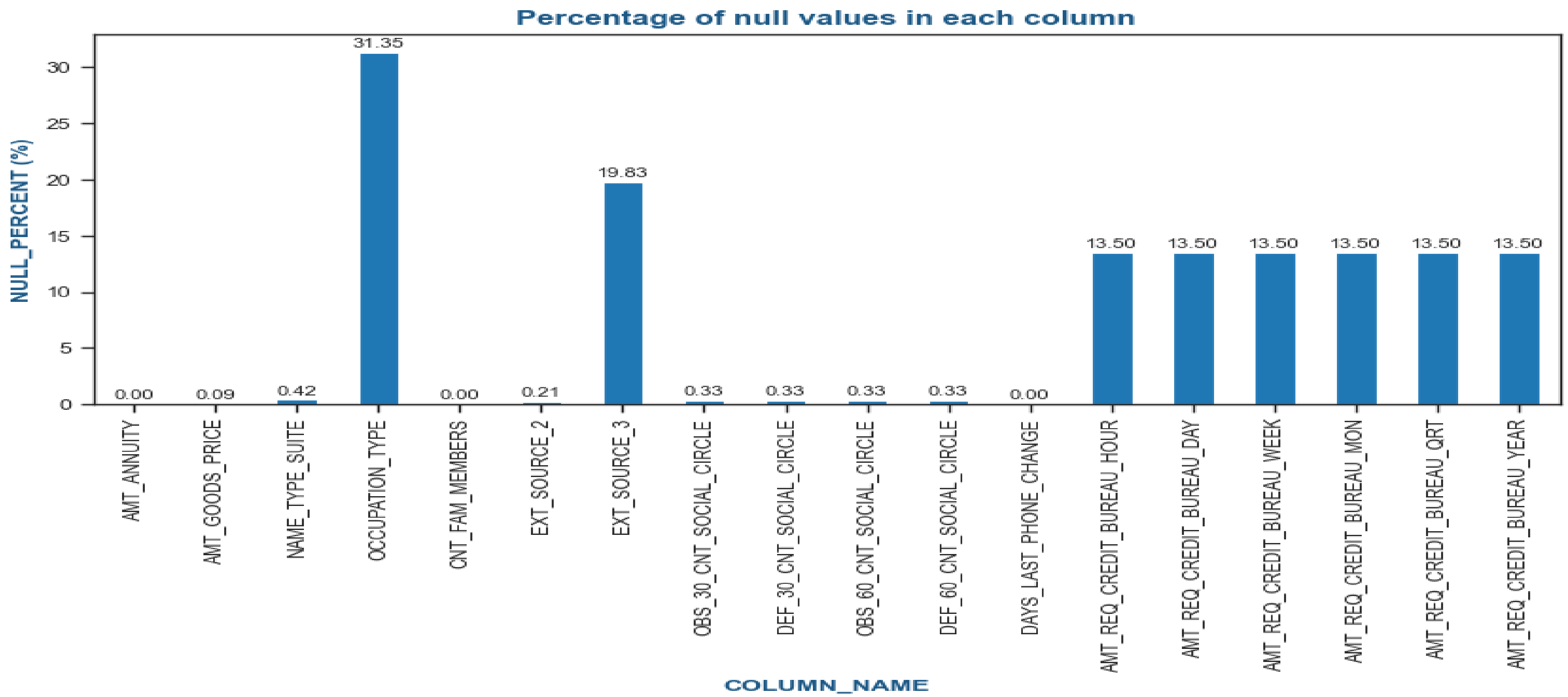
DATA UNDERSTANDING AND PREPARATION

Handling missing values: application_data



DATA UNDERSTANDING AND PREPARATION

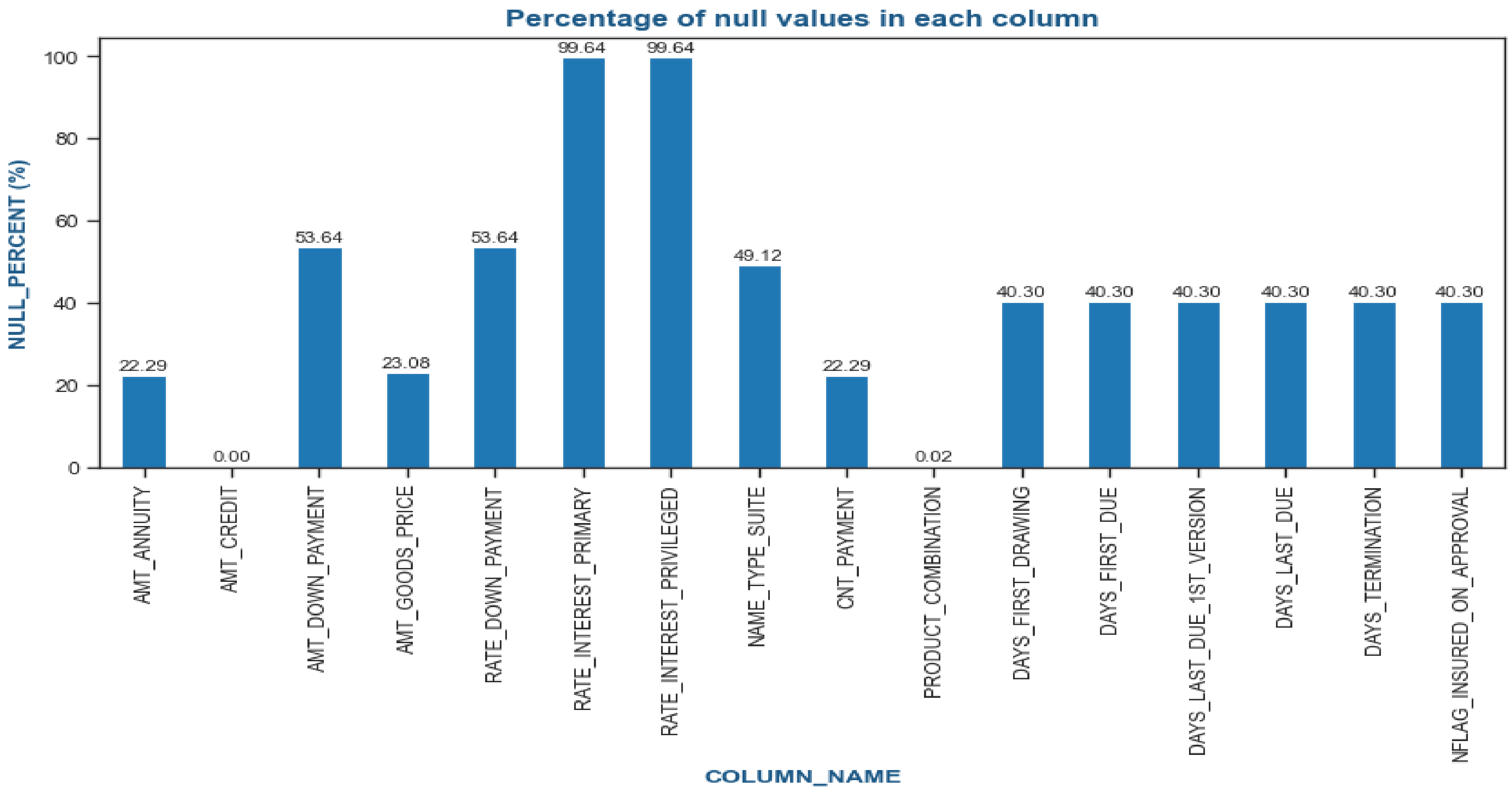
Handling missing values: application_data



Drop columns which have >40% null values:

DATA UNDERSTANDING AND PREPARATION

Handling missing values: previous_application



Drop columns which have >50% null values:

DATA UNDERSTANDING AND PREPARATION

Handling missing values

Imputation methods:

- Categorical variables: Impute with mode value
- Numerical variables:
  - Have lots of outliers: impute with median value
  - Others: impute with mean value

DATA UNDERSTANDING AND PREPARATION

Handling missing values

COLUMN_NAME	NULL_PERCENT	HANDLING_METHOD
AMT_ANNUITY	0.004	median
AMT_GOODS_PRICE	0.090	median
NAME_TYPE_SUITE	0.420	mode
OCCUPATION_TYPE	31.346	mode
CNT_FAM_MEMBERS	0.001	median
EXT_SOURCE_2	0.215	mean
EXT_SOURCE_3	19.825	mean
OBS_30_CNT_SOCIAL_CIRCLE	0.332	median
DEF_30_CNT_SOCIAL_CIRCLE	0.332	median
OBS_60_CNT_SOCIAL_CIRCLE	0.332	median
DEF_60_CNT_SOCIAL_CIRCLE	0.332	median
DAYS_LAST_PHONE_CHANGE	0.000	median
AMT_REQ_CREDIT_BUREAU_HOUR	13.502	median
AMT_REQ_CREDIT_BUREAU_DAY	13.502	median
AMT_REQ_CREDIT_BUREAU_WEEK	13.502	median
AMT_REQ_CREDIT_BUREAU_MON	13.502	median
AMT_REQ_CREDIT_BUREAU_QRT	13.502	median
AMT_REQ_CREDIT_BUREAU_YEAR	13.502	median

COLUMN_NAME	NULL_PERCENT	HANDLING_METHOD
AMT_ANNUITY	22.2867	median
AMT_CREDIT	0.0001	median
AMT_GOODS_PRICE	23.0818	median
NAME_TYPE_SUITE	49.1198	mode
CNT_PAYMENT	22.2864	median
DAYS_FIRST_DRAWING	40.2981	median
DAYS_FIRST_DUE	40.2981	median
DAYS_LAST_DUE_1ST_VERSION	40.2981	median
DAYS_LAST_DUE	40.2981	median
DAYS_TERMINATION	40.2981	median
NFLAG_INSURED_ON_APPROVAL	40.2981	mode

DATA UNDERSTANDING AND PREPARATION

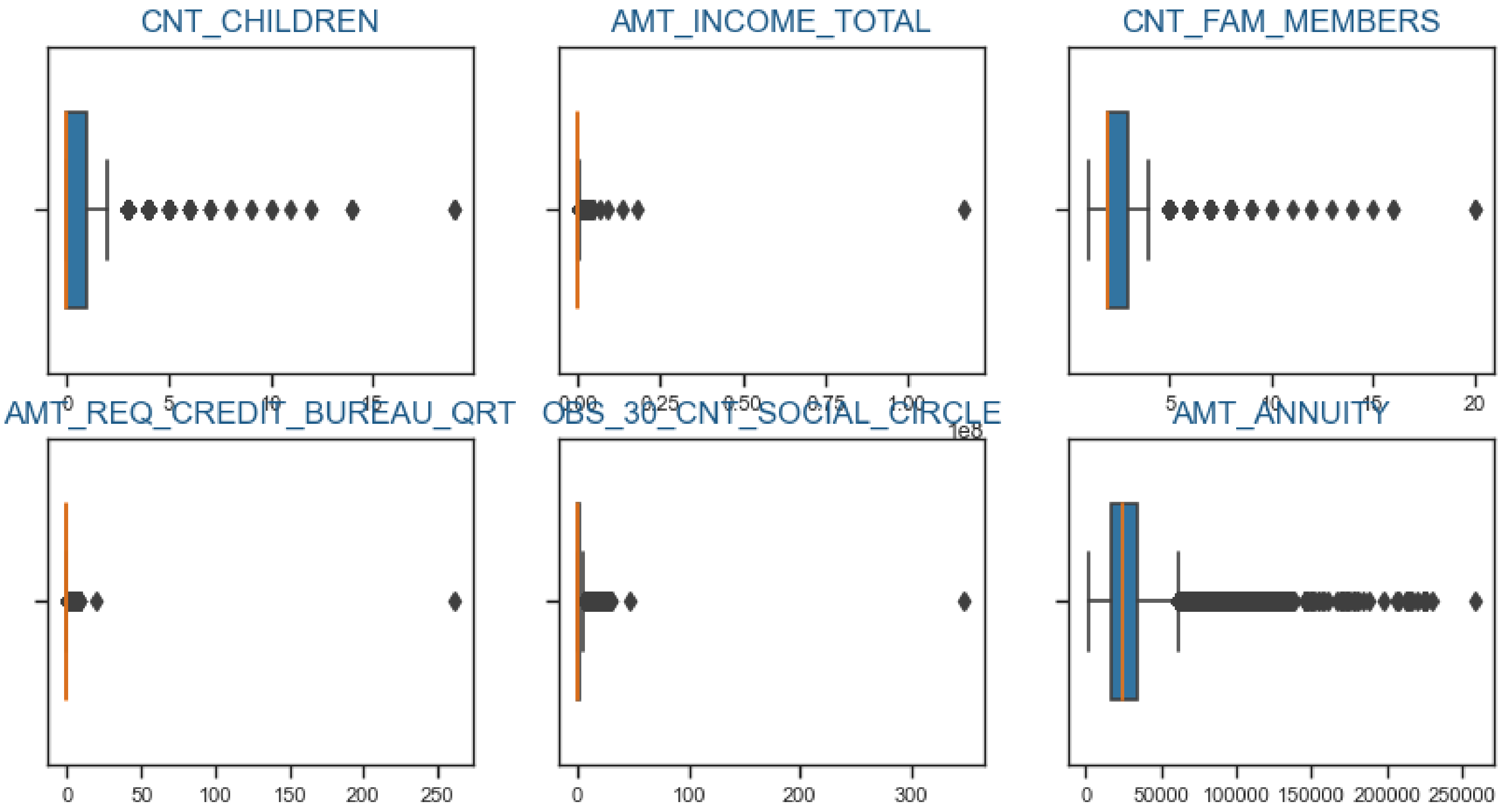
Standardising data

- Application_data: convert DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH, DAYS_LAST_PHONE_CHANGE to positive values
- Previous_application: convert DAYS_DECISION, DAYS_FIRST_DUE, DAYS_LAST_DUE, DAYS_TERMINATION to positive values

EXPLORATORY DATA ANALYSIS

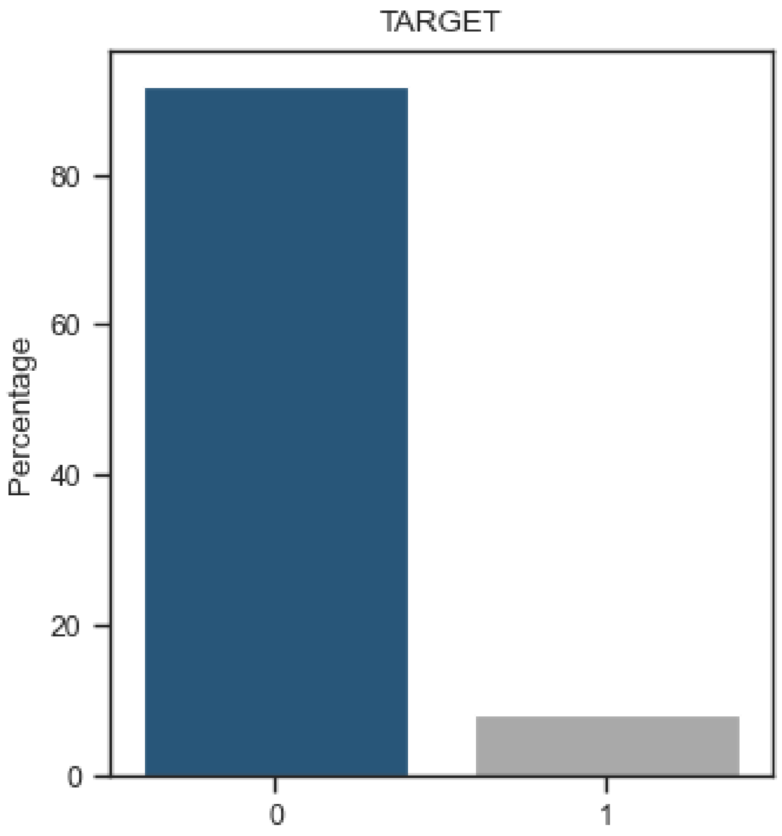
- Outliers
- Data imbalance
- Examine distribution of features: Analyze the distributions of numerical and categorical features with respect to default status.
- Explore correlations: Check correlations between features and the target variable as well as among features themselves.
- Identify trends: Observe any patterns or trends that might be associated with default behavior.

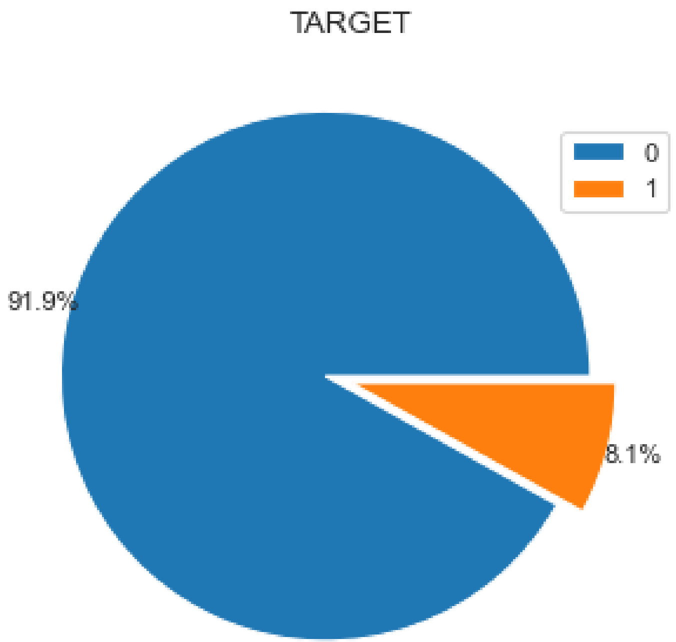
EDA - OUTLIER ANALYSIS



- Some columns contain values much bigger than 95th percentile values => outliers
- Those outliers are normal in most cases, sometimes representing unidentified values

EDA - DATA IMBALANCE



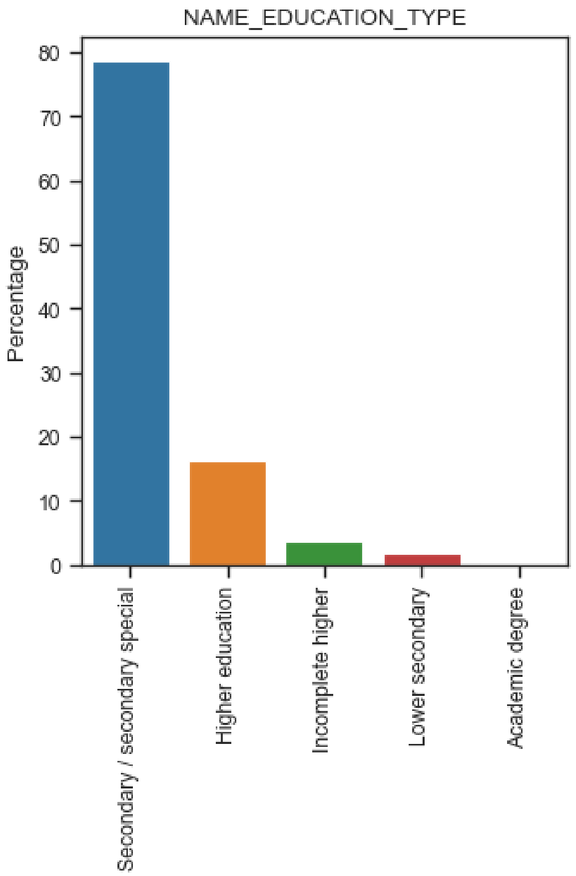


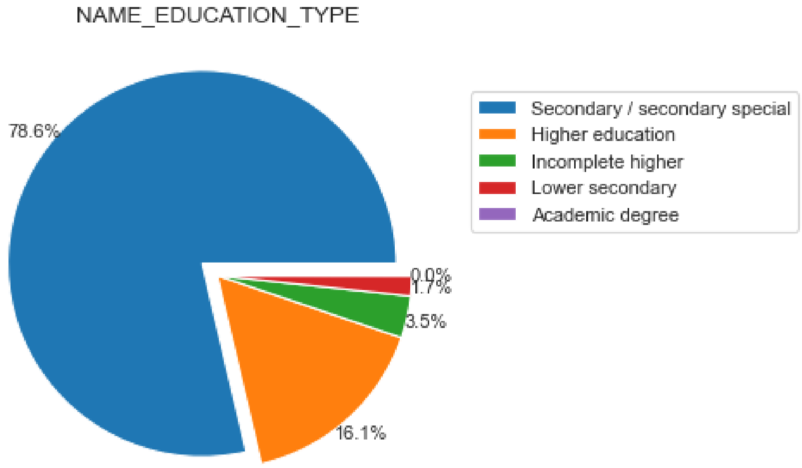
Target Variable:

- Only 8.1% of observations are defaulter's applications, whereas 91.9% are repayer's
- Imbalance ratio ~11.5

EDA - DATA IMBALANCE

Some imbalance cases in TARGET=1 segment:





- NAME_CONTRACT_TYPE: 93.5% of observations are cash loans
- NAME_TYPE_SUITE: 82.2% of observations are Unaccompanied
- NAME_INCOME_TYPE: 61.3% of observations are Unaccompanied Working
- NAME_EDUCATION_TYPE: 78.6% are Secondary / secondary special
- NAME_HOUSING_TYPE: 85.6% are House / apartment ..

EDA - CORRELATION

Top 10 correlation for the **Client with payment difficulties**:

Var1	Var2	Correlation
DAYS_EMPLOYED	FLAG_EMP_PHONE	-1.00
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	1.00
AMT_CREDIT	AMT_GOODS_PRICE	0.98
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.96
CNT_CHILDREN	CNT_FAM_MEMBERS	0.89
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.87
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.85
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.78
AMT_ANNUITY	AMT_GOODS_PRICE	0.75
AMT_CREDIT	AMT_ANNUITY	0.75

EDA - CORRELATION

Top 10 correlation for **all other cases** (Target variable):

Var1	Var2	Correlation
DAYS_EMPLOYED	FLAG_EMP_PHONE	-1.00
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	1.00
AMT_CREDIT	AMT_GOODS_PRICE	0.99
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.95
CNT_CHILDREN	CNT_FAM_MEMBERS	0.88
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.86
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.86
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.83
AMT_ANNUITY	AMT_GOODS_PRICE	0.78
AMT_CREDIT	AMT_ANNUITY	0.77

EDA – INSIGHTS OF DATA

- Identify trends: Observe any patterns or trends that might be associated with default behavior.
- Analyze categorical variables: Examine how different categories within categorical features relate to loan default.

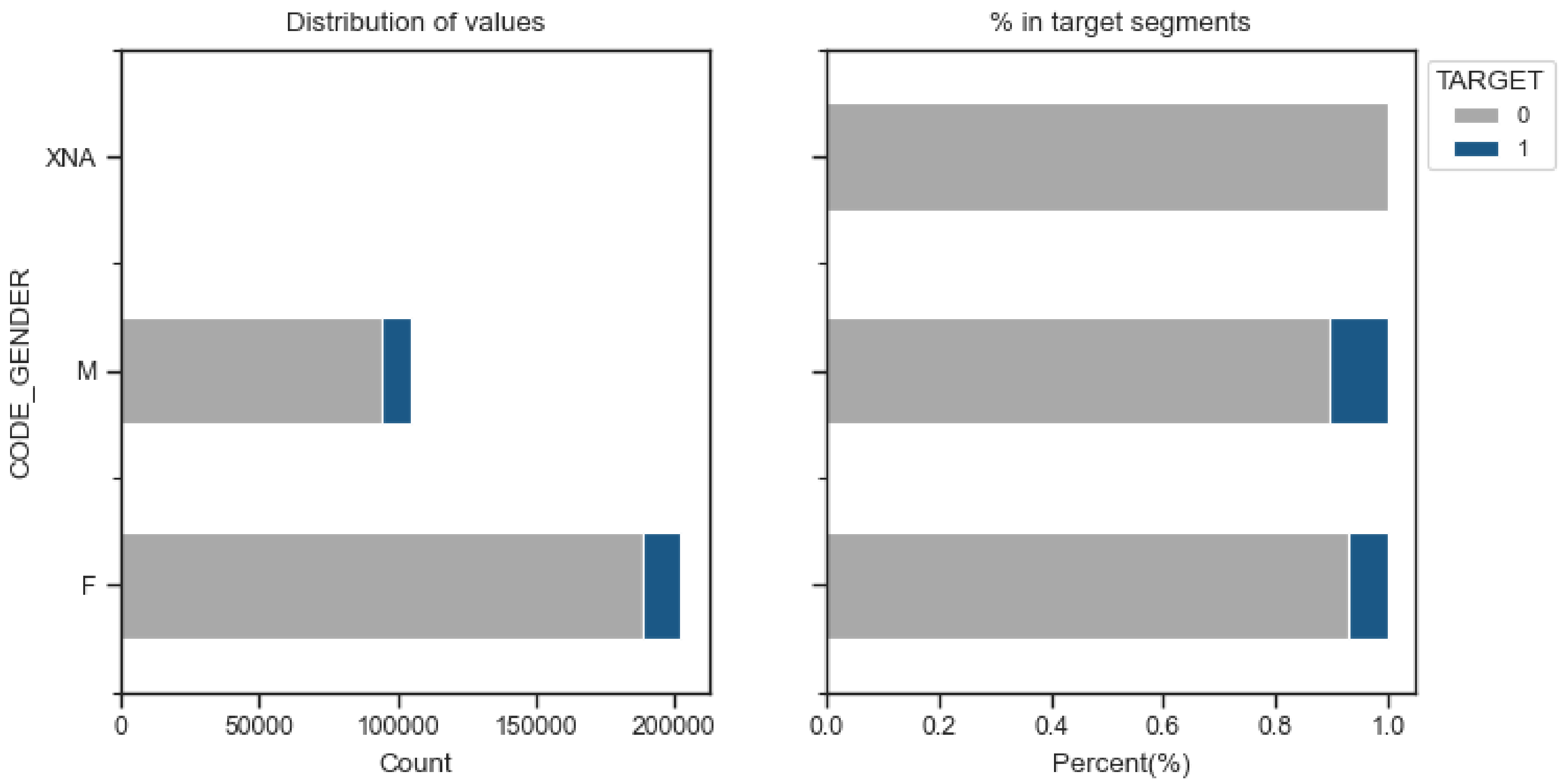
Explain the results of **univariate, segmented univariate, bivariate analysis, etc.** in business terms

Include visualisations and summarise the most important results in the presentation

Insights should explain why the variable is important for differentiating the **clients with payment difficulties with all other cases.**

EDA – INSIGHTS OF DATA

Demographic factors

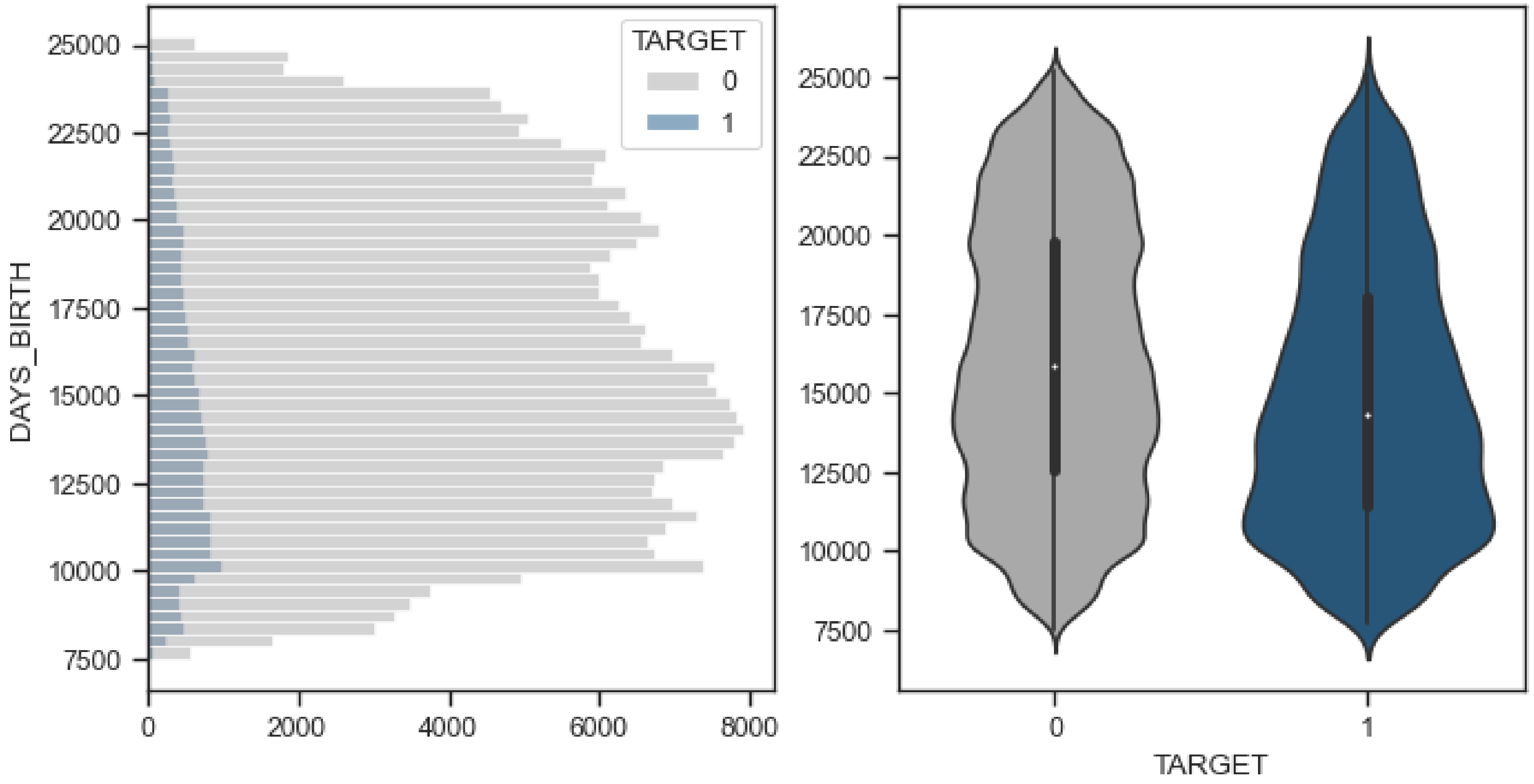


Inference:

- Most of applicants are female
- Proportion of defaulters in group male is higher than that in group female

EDA – INSIGHTS OF DATA

Demographic factors

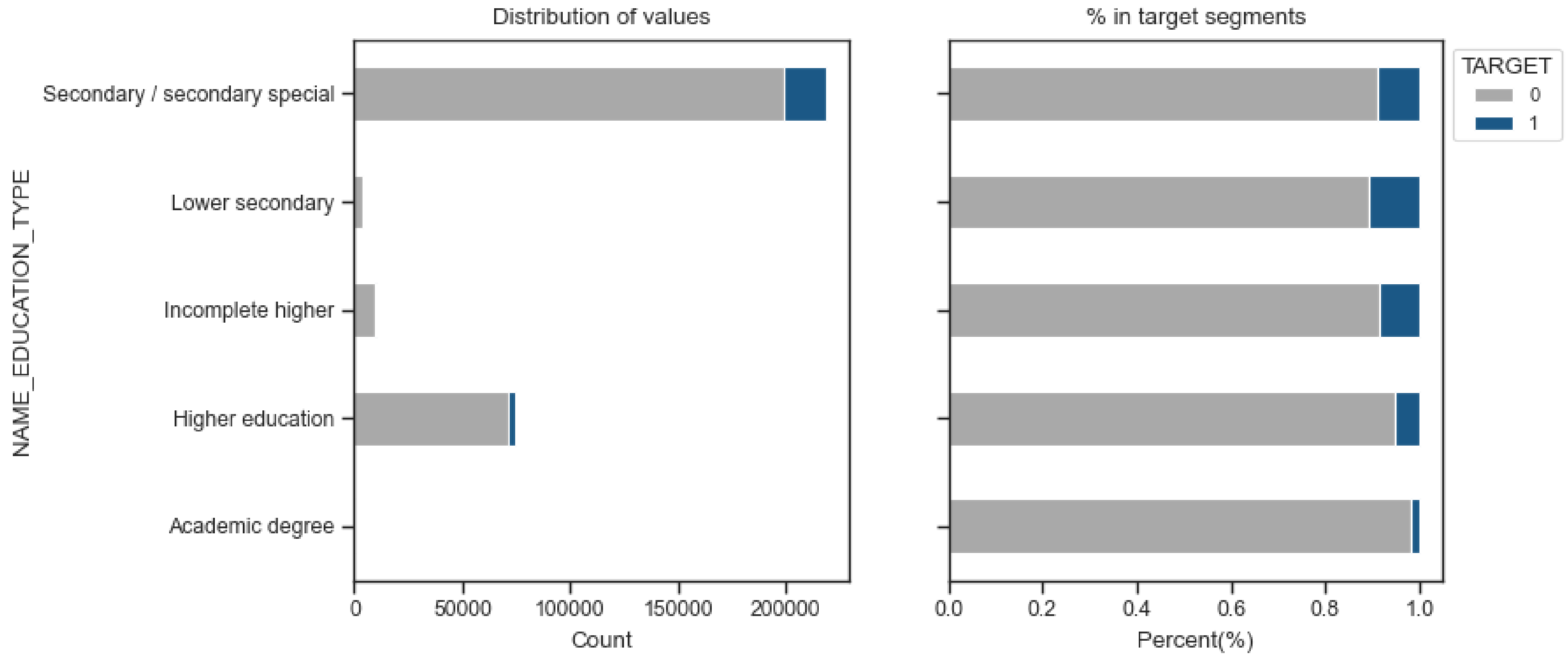


Inference:

- Younger people are more likely to default

EDA – INSIGHTS OF DATA

Education factors

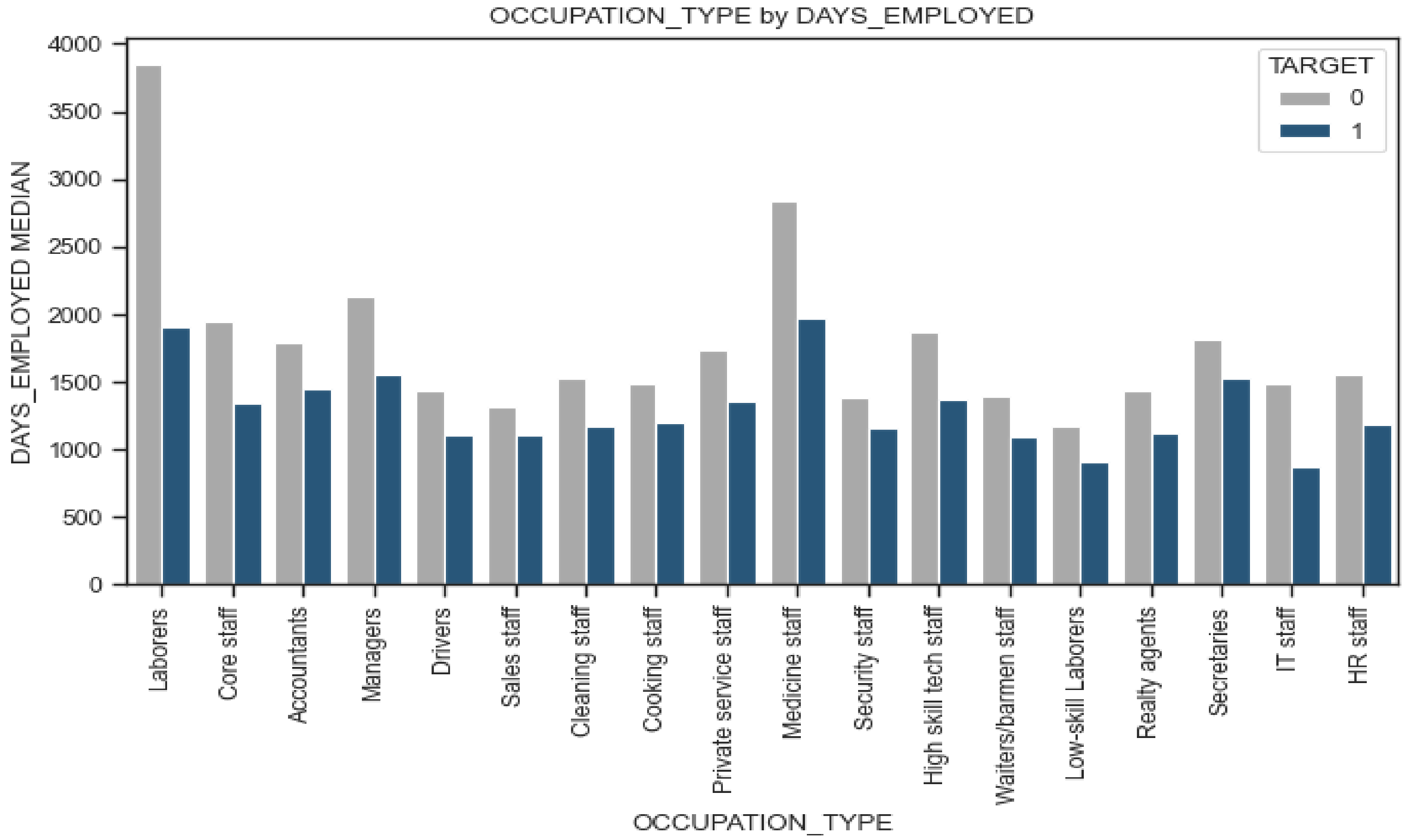


Inference:

- Most of applicants have Secondary Education_Type
- In lower education, proportion of defaulter is higher than in high education

EDA – INSIGHTS OF DATA

Employment factors

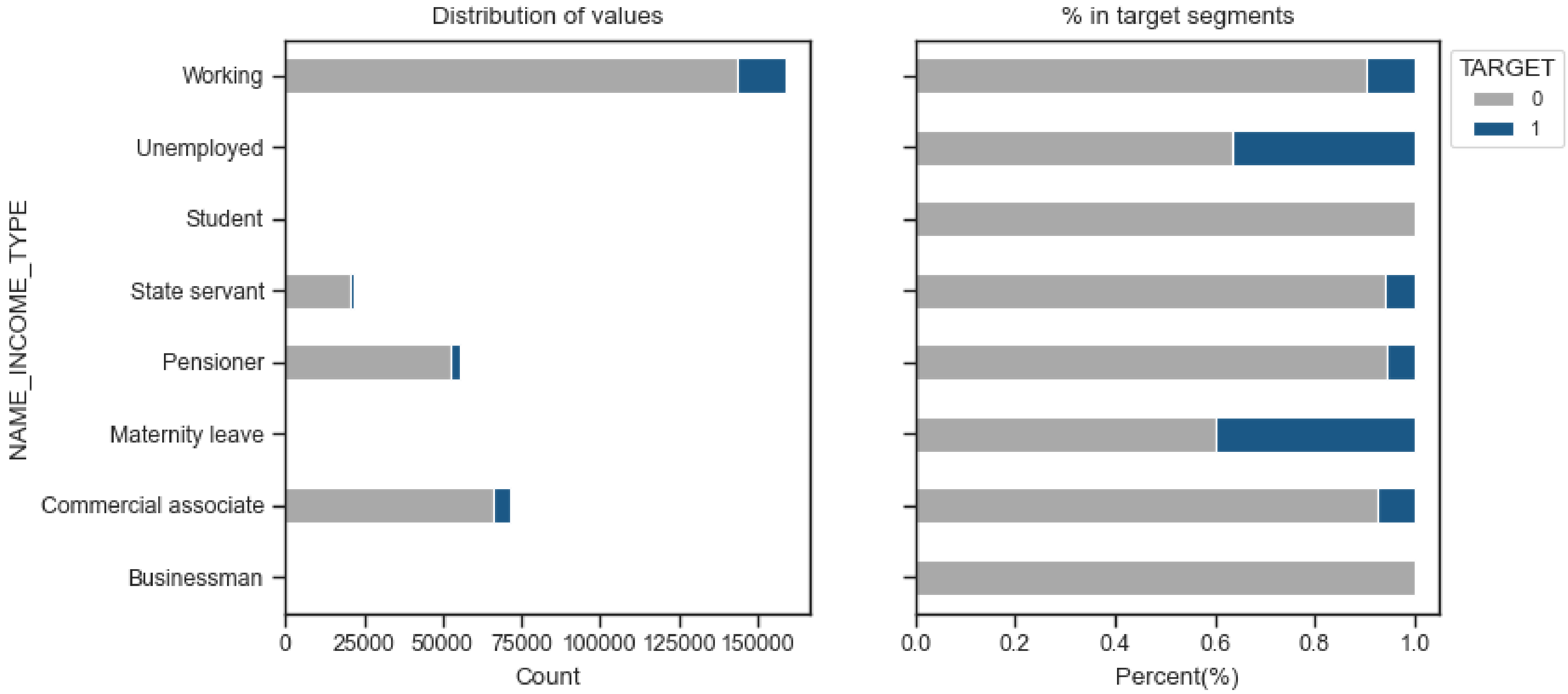


Inference:

- In most cases, defaulters tend to have lower median of DAYS_EMPLOYED than that applicants of other cases.

EDA – INSIGHTS OF DATA

Income factors

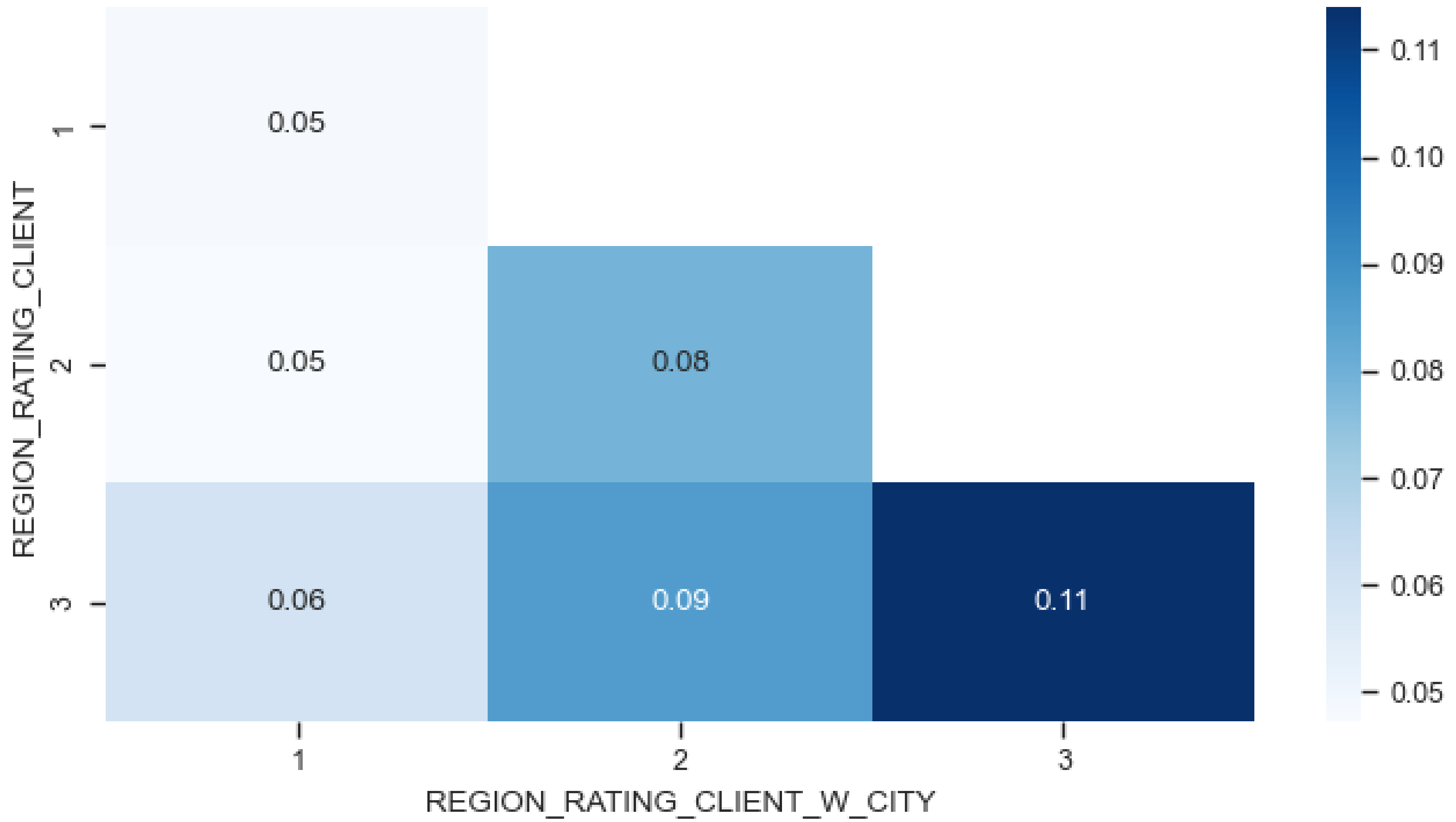


Inference:

- Proportion of Defaulters in group Unemployed and Maternity leave is much higher than in other groups.

EDA – INSIGHTS OF DATA

Geographic factors

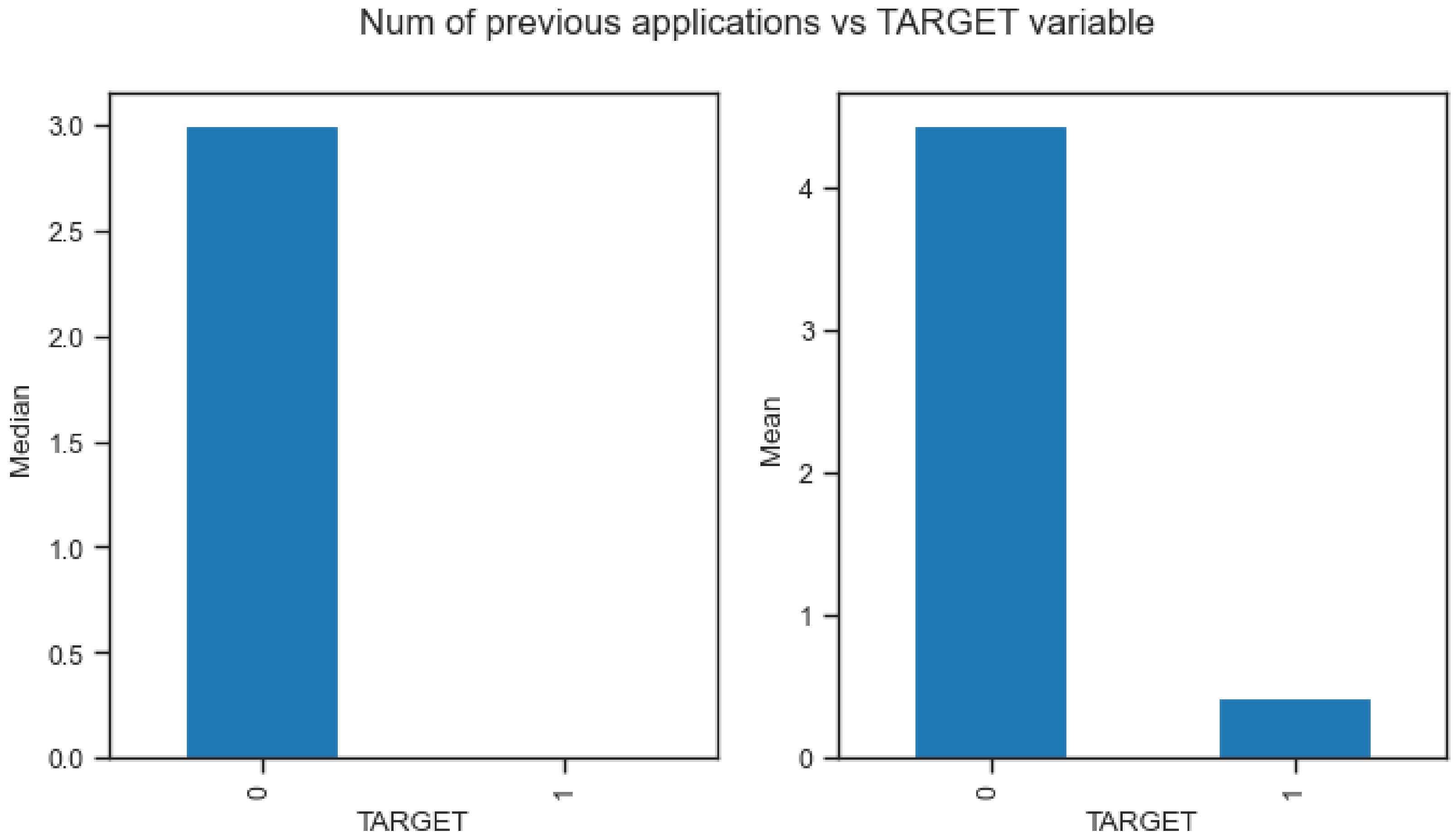


Inference:

- Applicants in REGIONs with higher rating are more likely to be default.

EDA – INSIGHTS OF DATA

Loan history

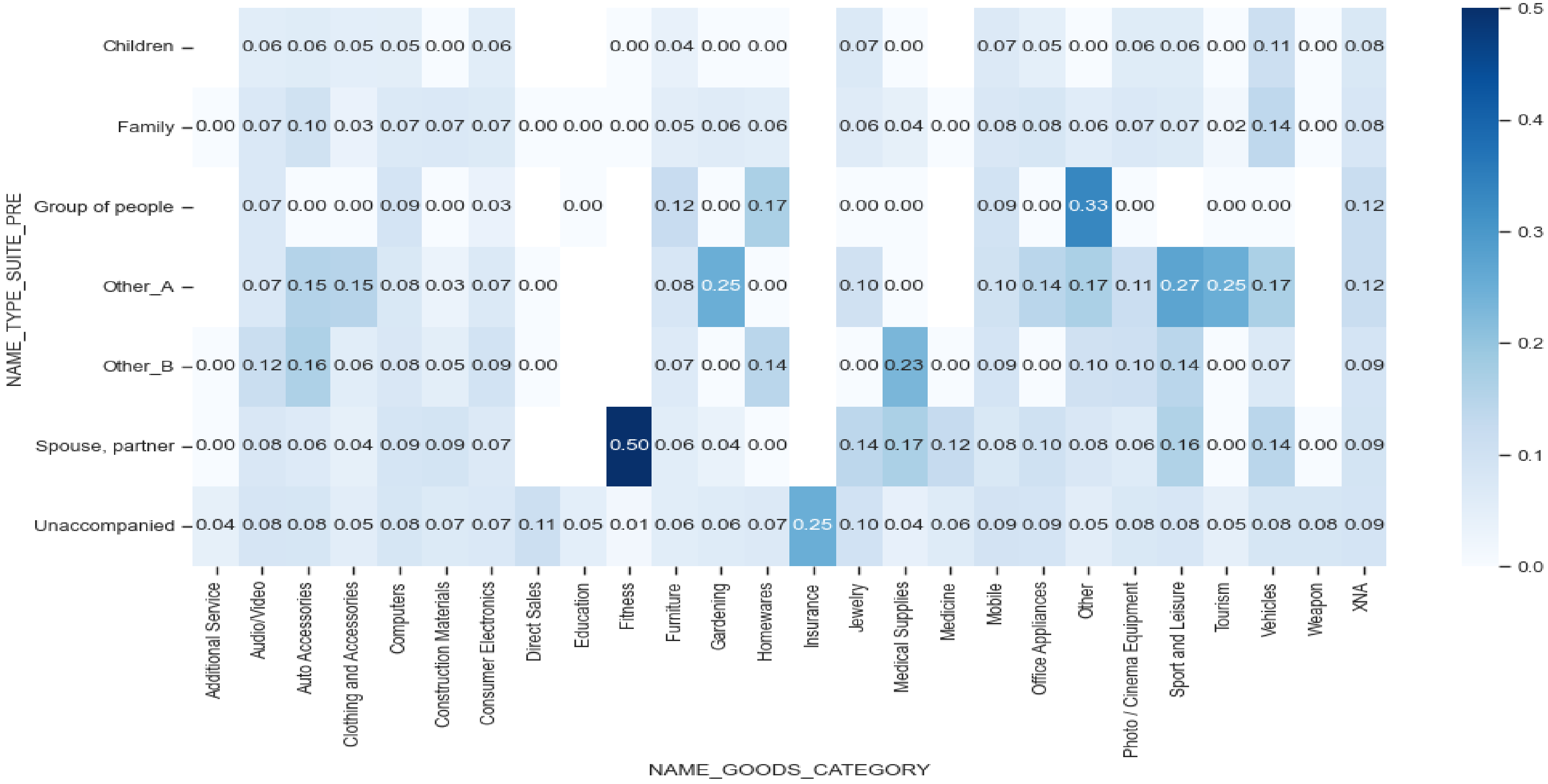


Inference:

- Most of defaulters haven't applied for loan in history.

EDA – INSIGHTS OF DATA

Loan history

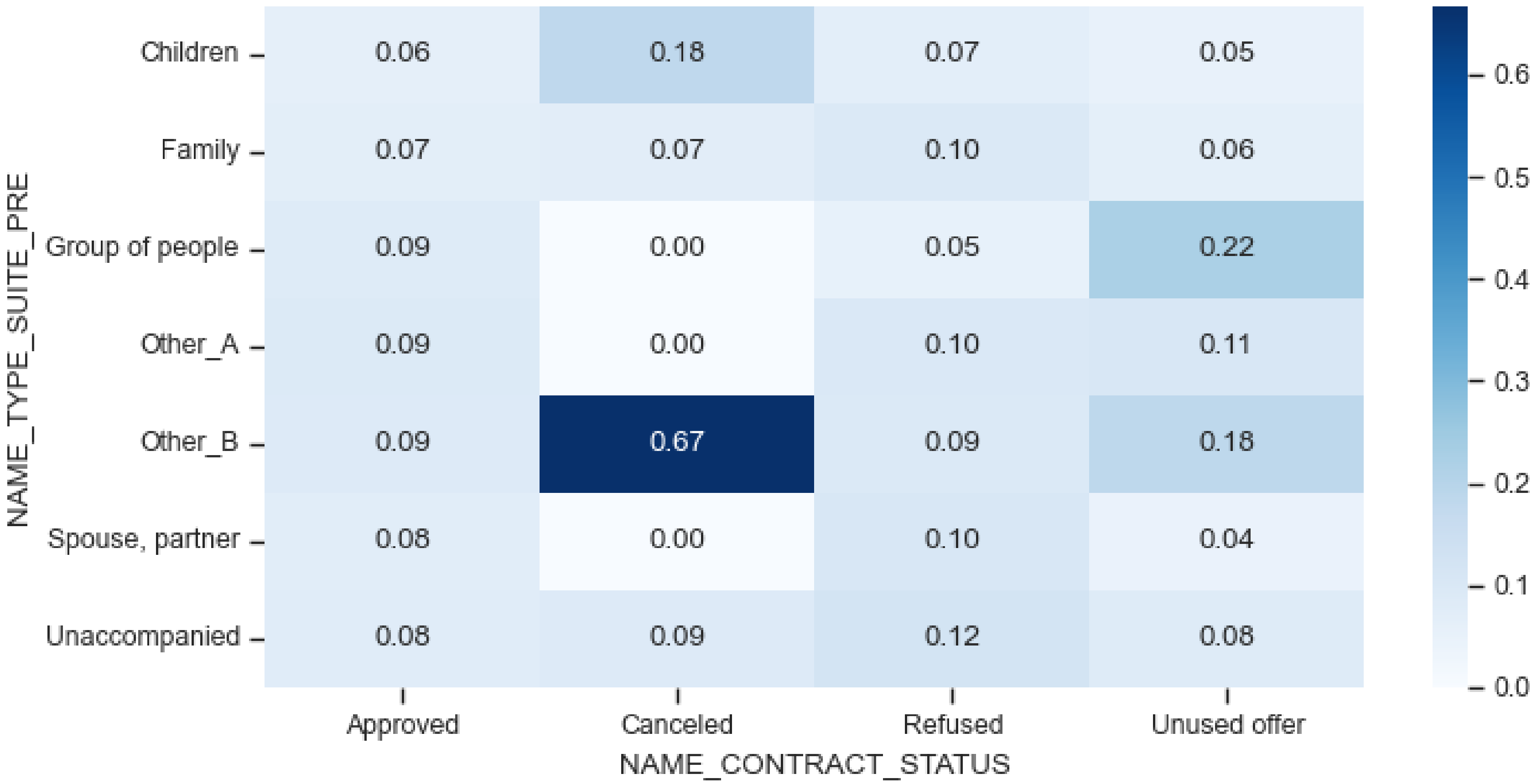


Inference:

- People who go with spouse/partner for Fitness services have higher probability to default

EDA – INSIGHTS OF DATA

Loan history

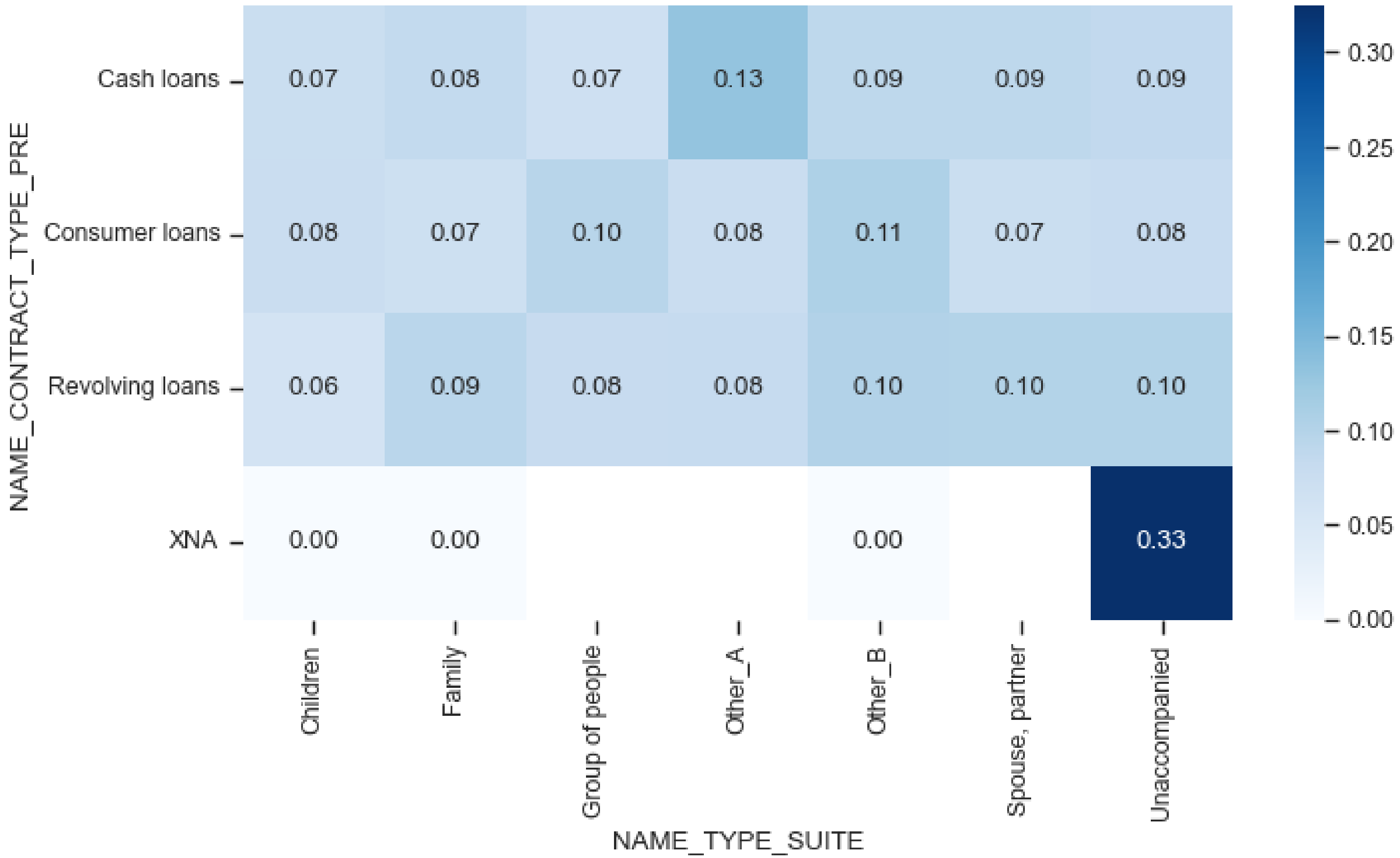


Inference:

- 67% applicants with NAME_TYPE_SUITE= Other_B, Contract status = Canceled are defaulters in current applications.

EDA – INSIGHTS OF DATA

Loan history

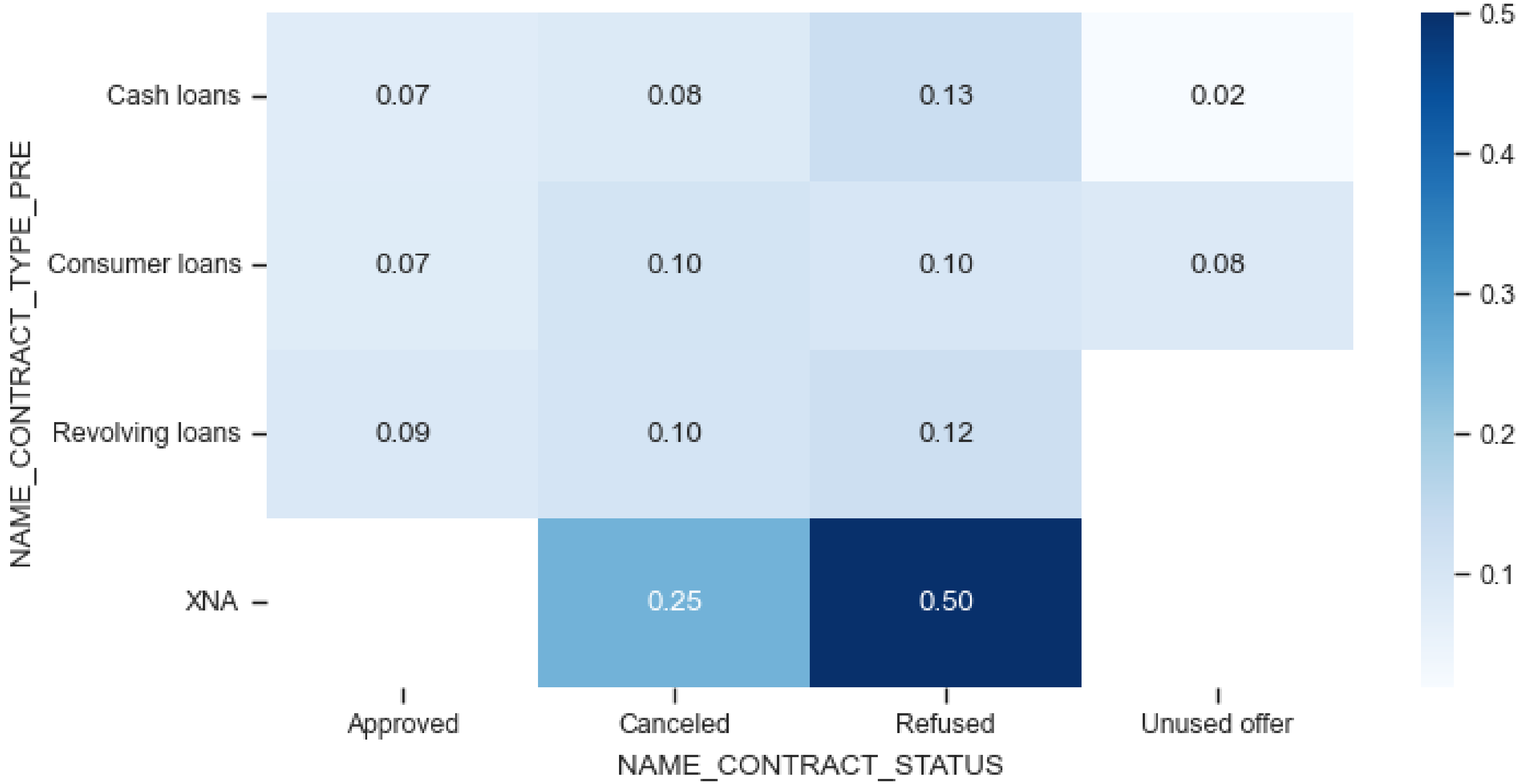


Inference:

- Proportion of default applicants with NAME_TYPE_SUITE= Uncompanied, CONTRACT_TYPE = XNA are significantly higher in comparation with others.

EDA – INSIGHTS OF DATA

Loan history

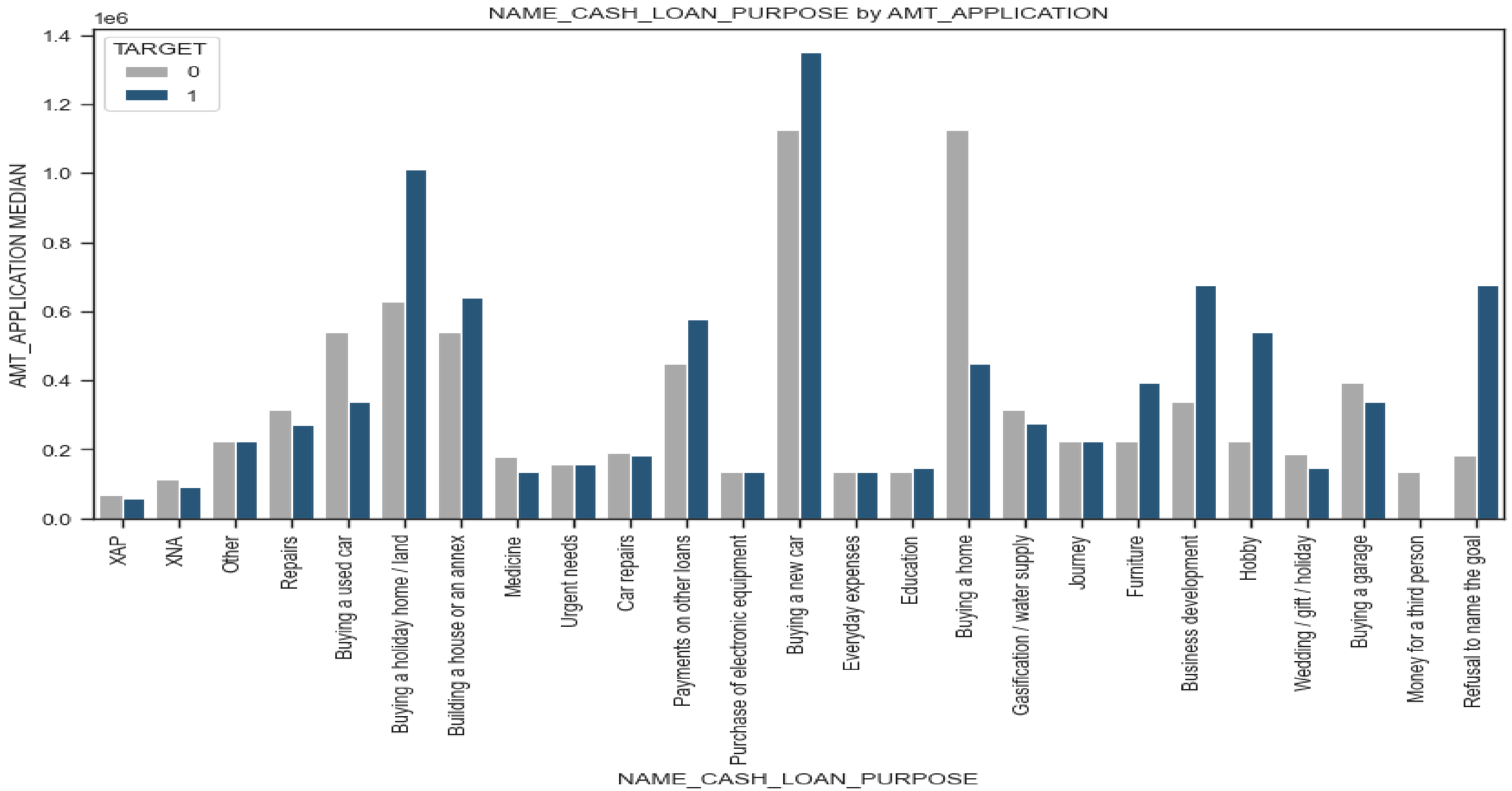


Inference:

- Proportion of default applicants with refused contract status, CONTRACT_TYPE = XNA are significantly higher in comparation with others.

EDA – INSIGHTS OF DATA

Loan history



Inference:

- With purpose of Buying Holiday home/land, Hobby, refusal to tell, medians of AMT_ANNUITY, AMT_CREDIT, AMT_APPLICATION of defaulters are significantly higher than repayers's

CONCLUSION

Significant insights from the data:

- Younger people are more likely to default.
- In lower education, proportion of defaulters is higher than high education's
- In most cases, defaulters tend to have lower median of DAYS_EMPLOYED than applicants of other cases.
- Most of defaulters haven't applied for loan in history
- With some kind of purpose like Buying Holiday home/land, Hobby, refusal to tell, higher in AMT_CREDIT, AMT_APPLICATION, AMT_ANNUITY tend leading to default