Credit EDA Case Study

Problem Statement

The case study involves a consumer finance company, which is in the business of providing loans to consumers and earning from interest payments. An important goal of the company is to identify consumers who are likely to repay their loans and those that are likely to default in order to address two business risks:

(i) risk of business loss by rejecting consumers who are likely to repay loans, and

(ii) risk of financial loss by approving consumers who are likely to default.

The given datasets contain information regarding (i) applicant profiles and payment difficulties, and (ii) information about previous loans.

The goal of EDA is to identify the factors which indicate whether an applicant is likely to repay or default on a loan.

Approach to Analysis

Analysis is done in the following manner:

1. Inspection, cleaning and analysis of application_data.csv
2. Inspection and cleaning of previous_application.csv followed by merging with application_data.csv and subsequent analysis

Steps of Analysis

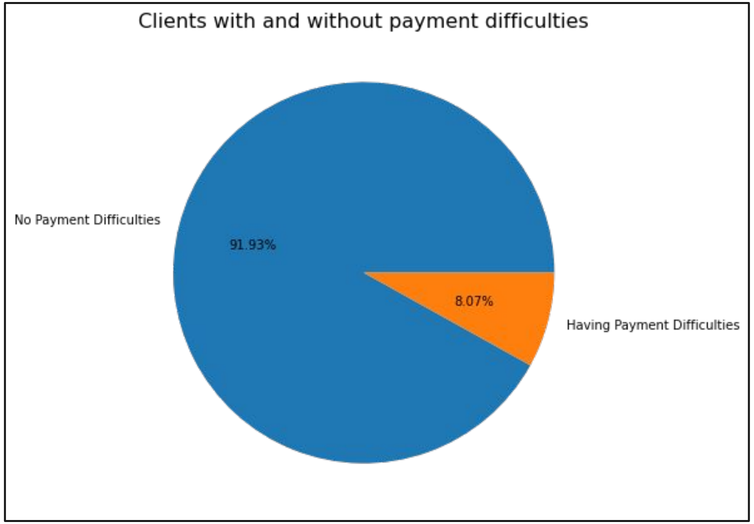
The following steps are followed for the two datasets.

1. Data Loading & Inspection
2. Data Cleaning
    2.1. Checking and fixing header and footer rows
    2.2. Fixing missing values
    2.3. Removing duplicates
    2.4. Validating and standardising data
    2.5. Binning of continuous variables
3. Data Merging (in case of previous_application.csv only)

Steps of Analysis

4. Data Analysis
    4.1. Analysis of Outliers
    4.2. Segmentation by the variable TARGET
    4.3. Univariate Analysis of Select Categorical Variables
    4.4. Univariate Analysis of Select Numeric Variables
    4.5. Bivariate & Multivariate Analysis
        4.5.1. Numeric - Numeric Analysis
        4.5.2. Numeric - Categorical Analysis
        4.5.3. Categorical - Categorical Analysis
    4.6. Correlations

Insights: *Imbalance in data*



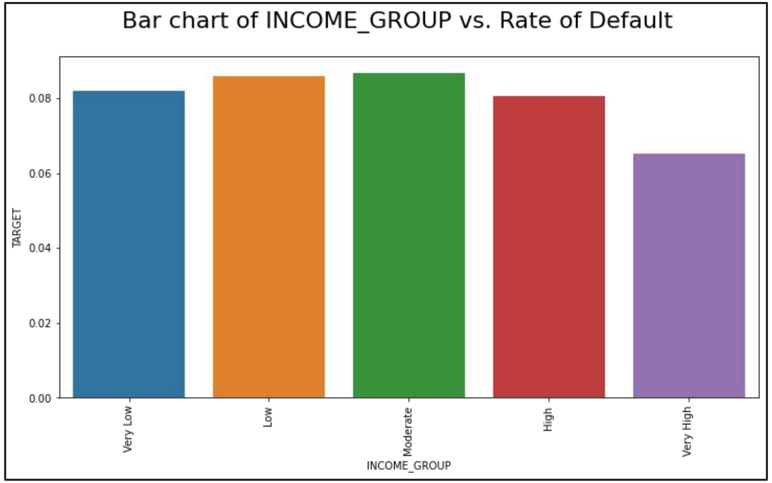
- As shown in the adjacent figure, 91.93% of applicants have no payment difficulties while 8.07% have payment difficulties
- The percentage of imbalance is 8.78%

Insights: *Driver Variables*

Analysis shows that the important driver variables are as follows.

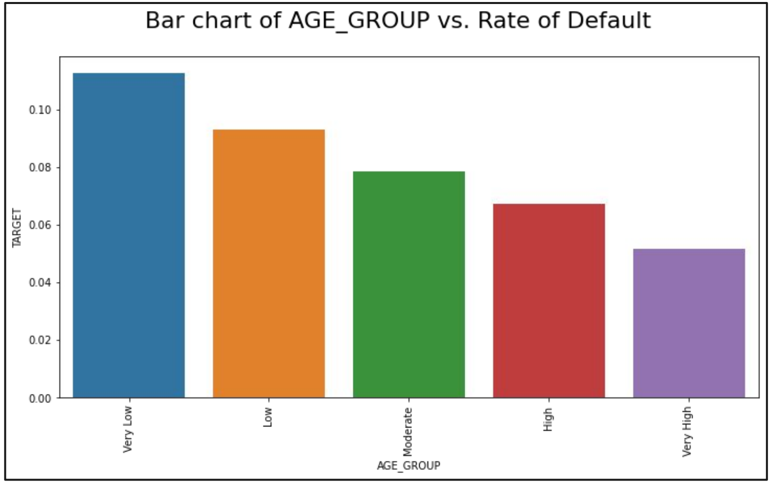
Income	Occupation Type
Age	External Scores
Contract Type	Contract Status
Gender	Product Type
Income Type	DAYS-DECISION
Education Type	Housing Type

Insights: *Driver Variable - Income*



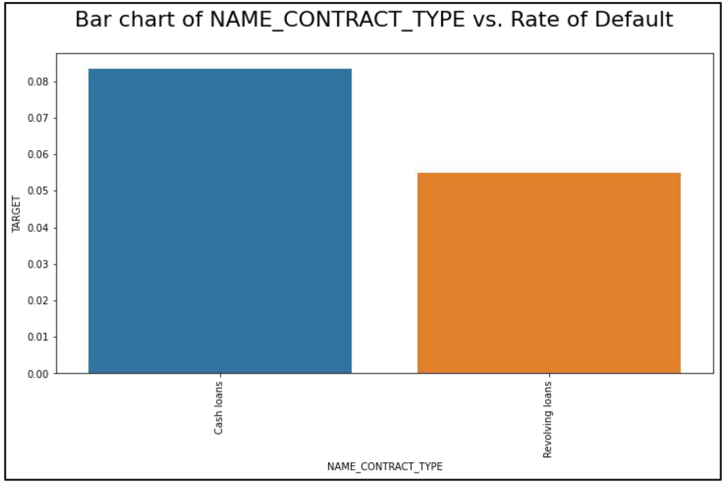
- Income is an important driver variable with low rates of default for individuals with 'Very High' and 'High' incomes.
- Other income groups have relatively higher rates of default.

Insights: *Driver Variable - Age*



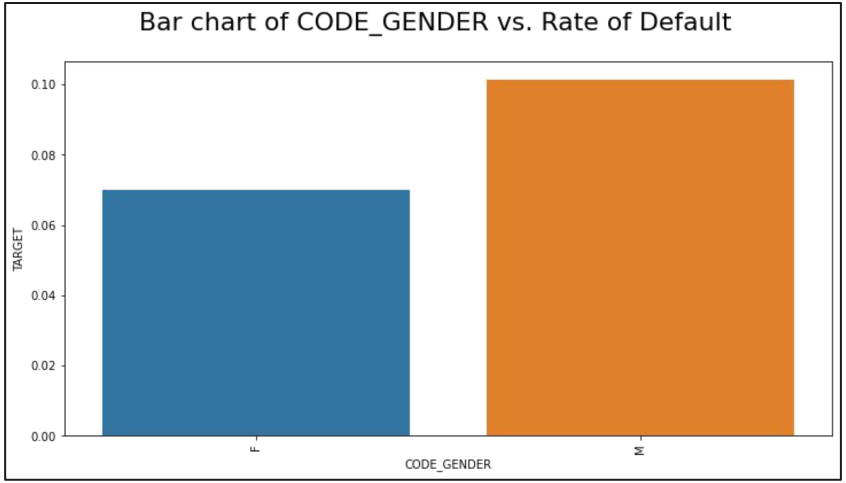
- Age is an important driver variable with lower rates of default for older individuals.
- As shown in the adjacent chart, rate of default decreases with age.

Insights: *Driver Variable - Contract Type*



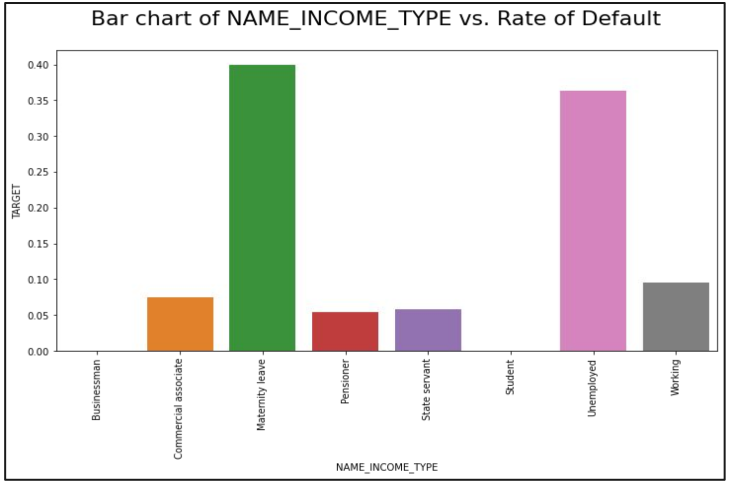
- Contract type is an important driver variable with a higher rate of default for ‘Cash loans’ compared to ‘Revolving loans’.
- Thus, the company should make more revolving loans.

Insights: *Driver Variable - Gender*



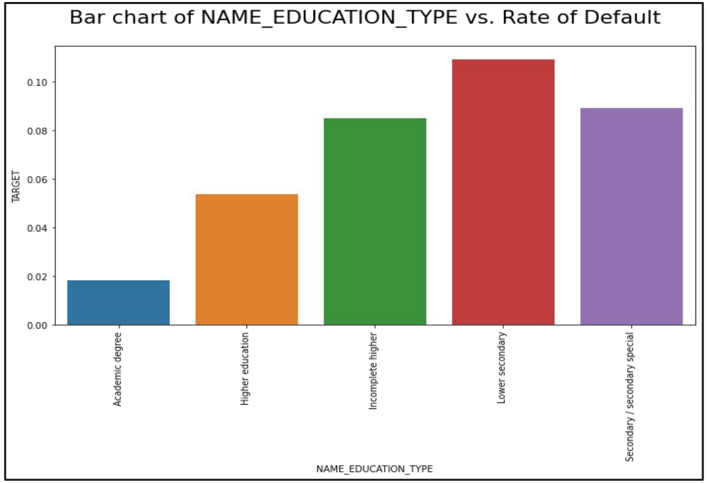
The rate of default is higher for males and lower for females, which indicates that females are better customers for the bank.

Insights: *Driver Variable - Income Type*



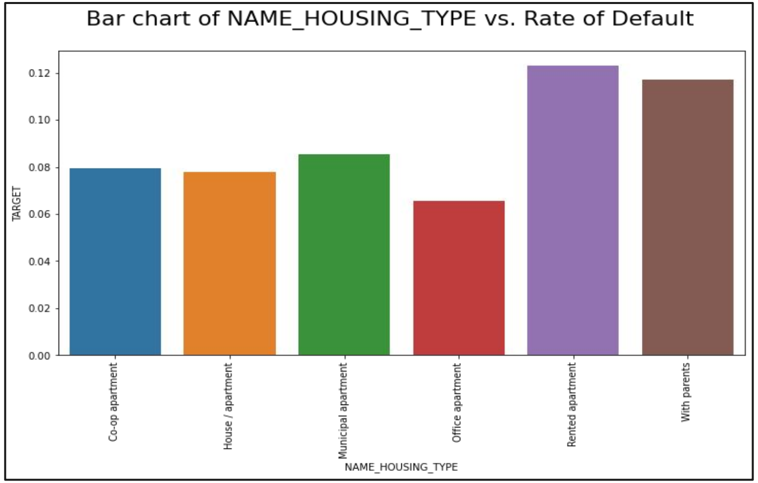
Type of income is an important predictor of default, with high rates of default for ‘Unemployed’ and ‘Maternity leave’.

Insights: *Driver Variable - Education Type*



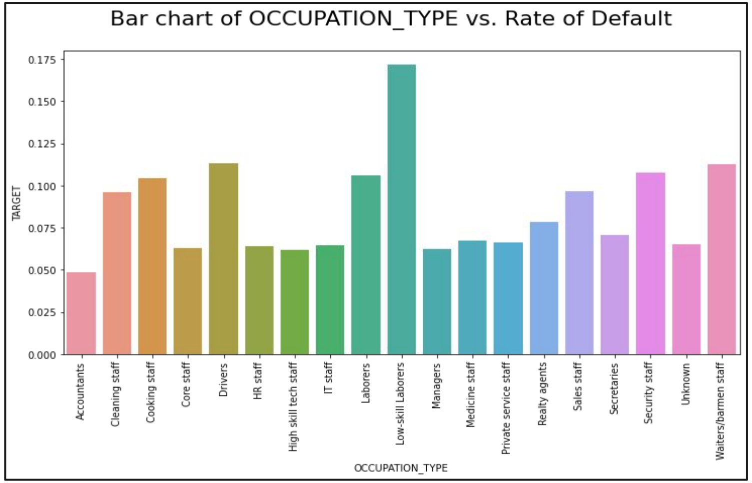
Education is an important predictor of default, with higher education indicating lower rates of default. Thus, the company should try to lend more to applicants having 'Academic degrees' or 'Higher education'.

Insights: *Driver Variable - Housing Type*



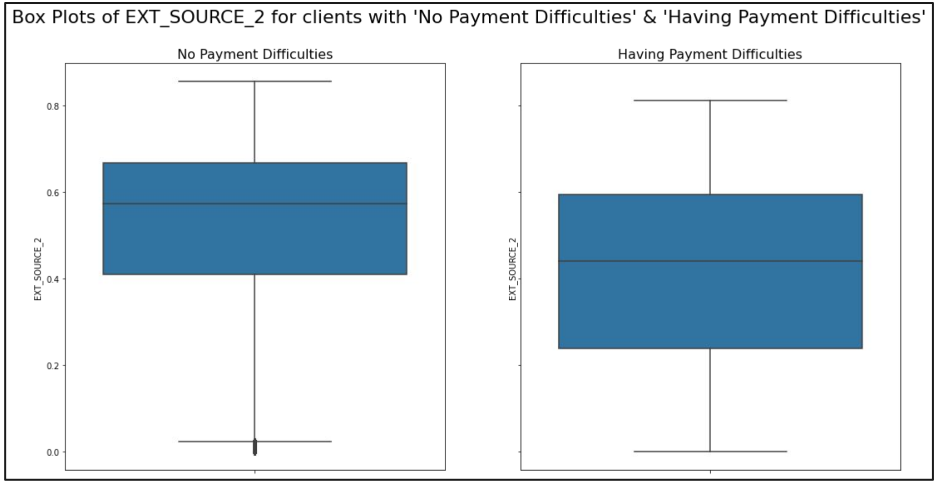
Type of housing is an important predictor of default, with clients living with parents or in rented apartments having higher rates of default. Thus, the company should avoid such clients or offer higher prices to address potential default.

Insights: *Driver Variable - Occupation Type*



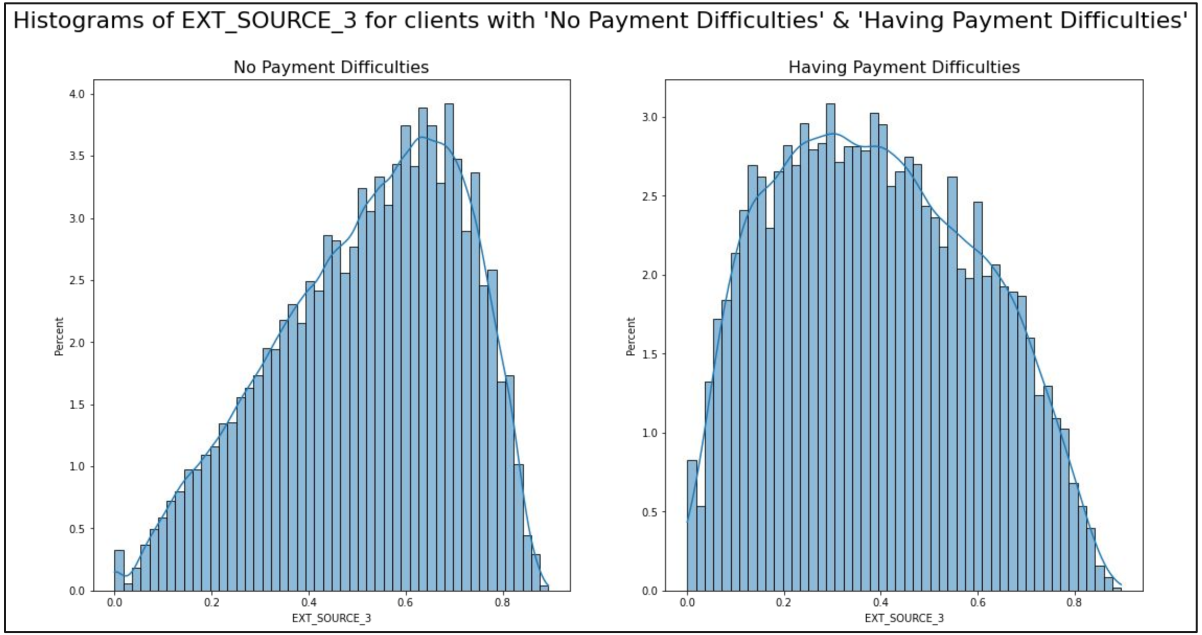
Occupation type is an important driver variable, with low skill workers such as cleaning staff, cooking staff, drivers, laborers and security staff having relatively higher rates of default.

Insights: *Driver Variable - External Scores*

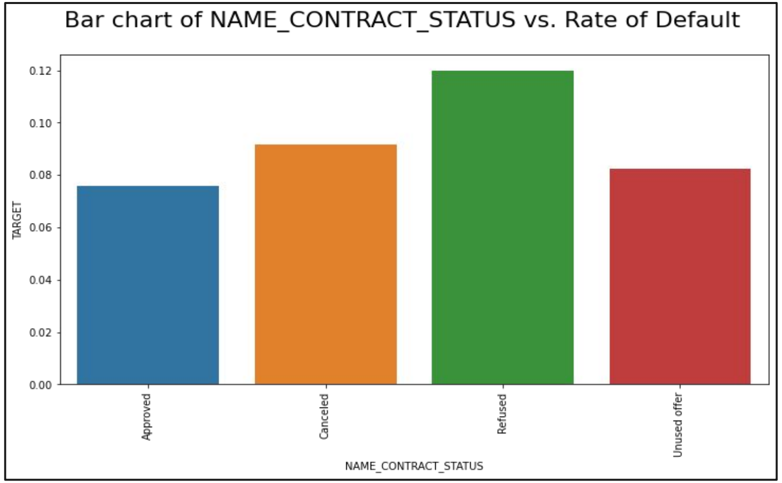


- Scores from external sources (such as EXT_SOURCE_2 & EXT_SOURCE_3) are important indicators of default.
- As shown in the adjacent figure, clients with lower scores tend to have more payment difficulties.

Insights: *Driver Variable - External Scores*

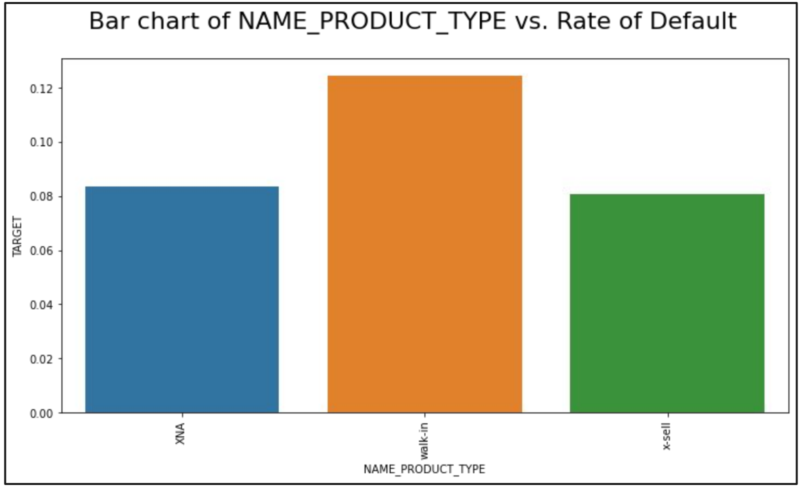


Insights: *Driver Variable - Contract Status*



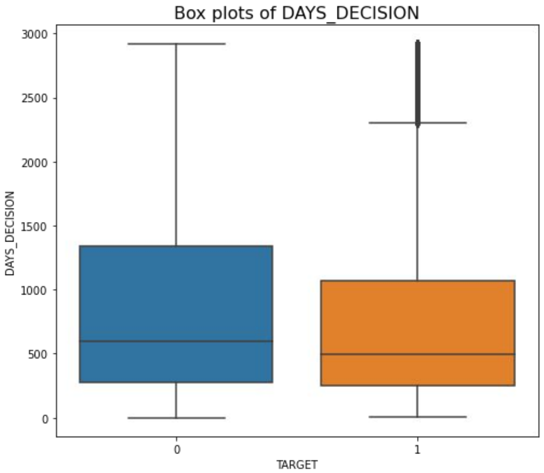
- Contract status is an important driver variable.
- As shown in the adjacent figure, clients whose previous loan applications were cancelled or refused are more likely to default.

Insights: *Driver Variable - Product Type*



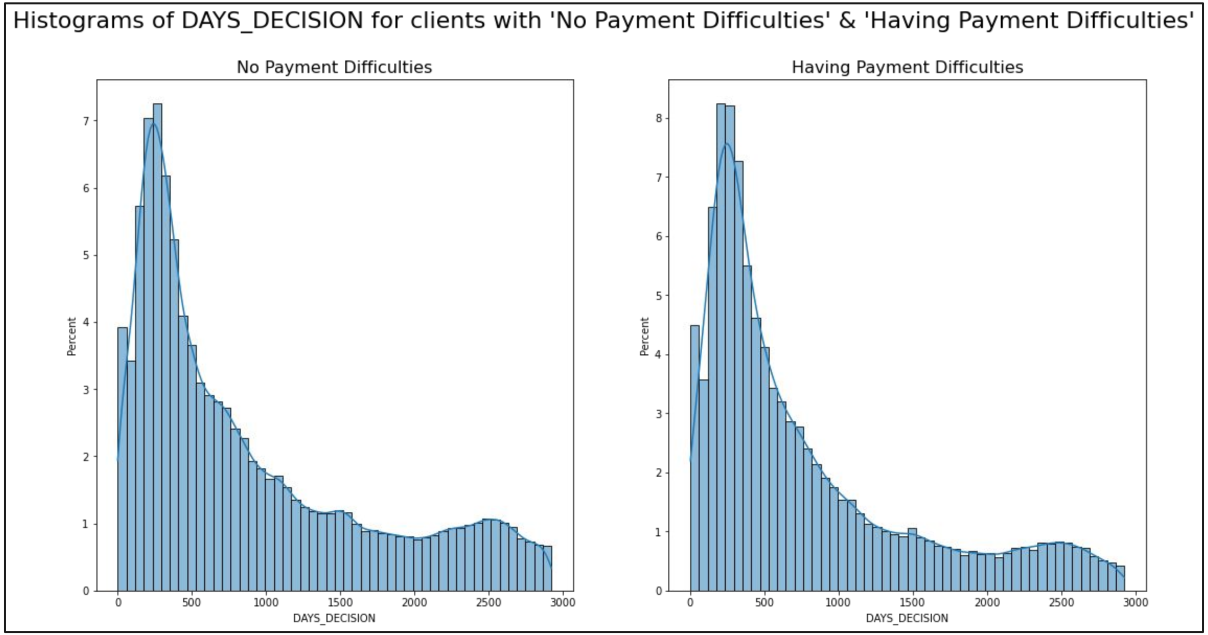
Product type is an important indicator of default with ‘walk-in’ sales having a much higher rate of default than other types.

Insights: *Driver Variable - DAYS_DECISION*



- DAYS_DECISION is the number of days prior to the current application that the decision about the previous application was made.
- As shown in the adjacent figure, clients who apply for loans without much time in between them tend to have higher payment difficulties.

Insights: *Driver Variable - DAYS_DECISION*



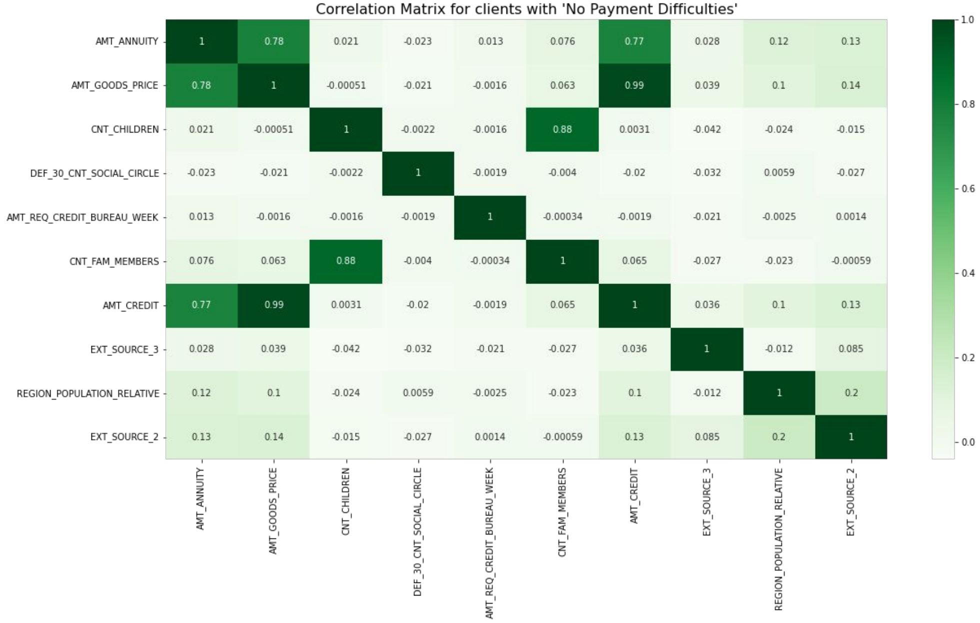
Insights: *Top 10 correlations for clients with 'No Payment Difficulties'*

	level_0	level_1	0
2246	OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998508
1852	FLOORSMAX_AVG	FLOORSMAX_MEDI	0.997018
2042	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.993582
1982	FLOORSMAX_MODE	FLOORSMAX_MEDI	0.988152
328	AMT_GOODS_PRICE	AMT_CREDIT	0.987253
1850	FLOORSMAX_AVG	FLOORSMAX_MODE	0.985602
1912	YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.971032
2044	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.962064
1057	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950148
977	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878570

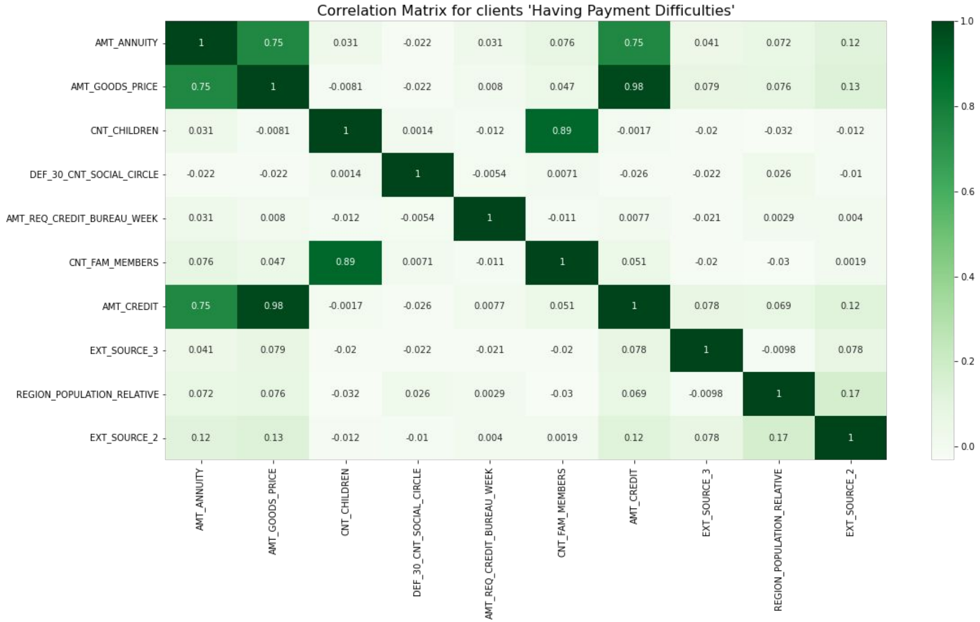
Insights: *Top 10 correlations for clients 'Having Payment Difficulties'*

	level_0	level_1	0
2246	OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998269
1852	FLOORSMAX_AVG	FLOORSMAX_MEDI	0.997187
1786	YEARS_BEGINEXPLUATATION_AVG	YEARS_BEGINEXPLUATATION_MEDI	0.996124
2110	FLOORSMAX_MEDI	FLOORSMAX_MODE	0.989195
1978	FLOORSMAX_MODE	FLOORSMAX_AVG	0.986594
200	AMT_CREDIT	AMT_GOODS_PRICE	0.983103
1784	YEARS_BEGINEXPLUATATION_AVG	YEARS_BEGINEXPLUATATION_MODE	0.980466
2044	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.978073
1057	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.956637
145	CNT_CHILDREN	CNT_FAM_MEMBERS	0.885484

Insights: *Correlation matrix for clients with 'No Payment Difficulties'*



Insights: *Correlation matrix for clients 'Having Payment Difficulties'*



Recommendations

Variable	Good Customers	Risky Customers
Income	Very High, High	Moderate, Low
Age	Very High, High	Low, Very Low
Contract Type	Revolving loans	Cash loans
Gender	Female	Male
Income Type	State servant, Pensioner	Maternity Leave, Unemployed
Education Type	Academic degree, Higher education	Lower secondary, Secondary / secondary special

Recommendations

Variable	Good Customers	Risky Customers
Occupation Type	Accountants, IT Staff, Managers	Laborers, Low skill-laborers, Drivers
External Scores	High	Low
Contract Status	Approved, Unused	Cancelled, Refused
Product Type	walk-in	x-sell, XNA
DAYS_DECISION	High	Low
Housing Type	Office apartment, House / apartment	Rented apartment, With parents