EDA Case Study-Final Report

Group Members:

ANKIT KUMAR SINGH

ARJUN SUCCENA

Business Understanding

Business Objective:

- This case study aims that 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**.**

Goal of Analysis:

- To find out the relation between the different variables and their impact on loan default. And suggest which attributes contributes a significant difference in Loan Default.

Data to be analysed

- **’application_data.csv‘** contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.

- **’previous_application.csv‘** contains information about the client’s previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

- **’columns_description.csv’** is data dictionary which describes the meaning of the variables

Data Understanding

The company has come across some important attributes in order to understand behavior of their approved loan customers w.r.t. loan default. Thus, the lending company has decided to work only on these variables to mitigate the future risk. The driver variables consider for this case study are:

- NAME_CONTRACT_TYPE
- CODE_GENDER
- FLAG_OWN_CAR
- FLAG_OWN_REALTY
- CNT_CHILDREN
- AMT_INCOME_TOTAL
- AMT_CREDIT
- AMT_ANNUITY
- AMT_GOODS_PRICE
- NAME_TYPE_SUITE
- NAME_INCOME_TYPE
- NAME_EDUCATION_TYPE
- NAME_FAMILY_STATUS
- NAME_HOUSING_TYPE
- DAYS_BIRTH

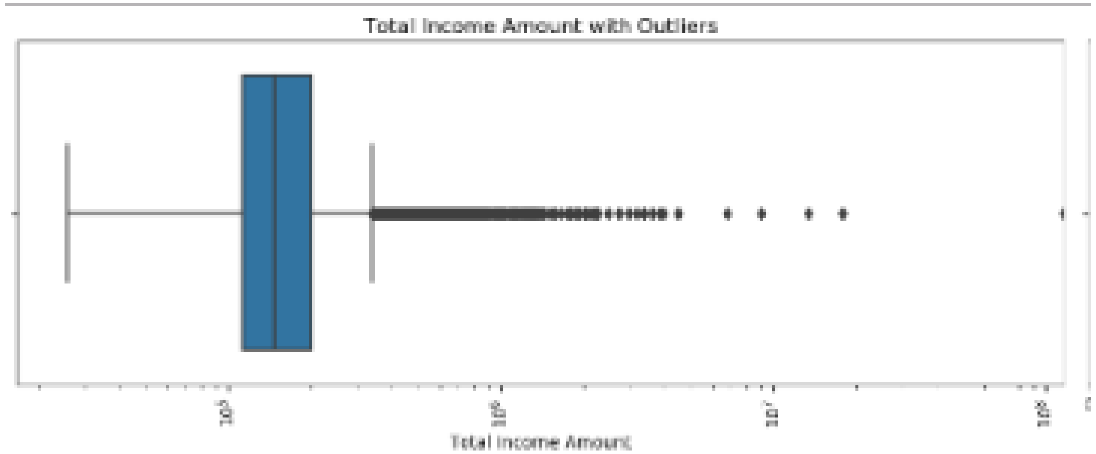
Data Cleaning and Preparation

- We loaded the file application.csv in Jupyter Notebook and found that there are 307511 observation and 122 variables, since in our data understanding stage we have already identified our driver variable we will be focusing our analysis on those variables.
- We found that SK_ID_CURR is the unique variable, so we have check for any duplicate in our data sets.
- We will look out for columns which have 10 -13% of missing values(Suggest metrics(no imputation) using Mean or Median based on the Outliers)
- We will look out for columns which have more that 50% of missing values(Drop the columns).
- We have removed the all columns with more than 50% of missing values and have made a subset of approx. 30 columns which will be useful for the analysis.
- We have then divided the whole dataset into two categories-Target 0 (Defaulters) and Target 1 (Non Defaulters).

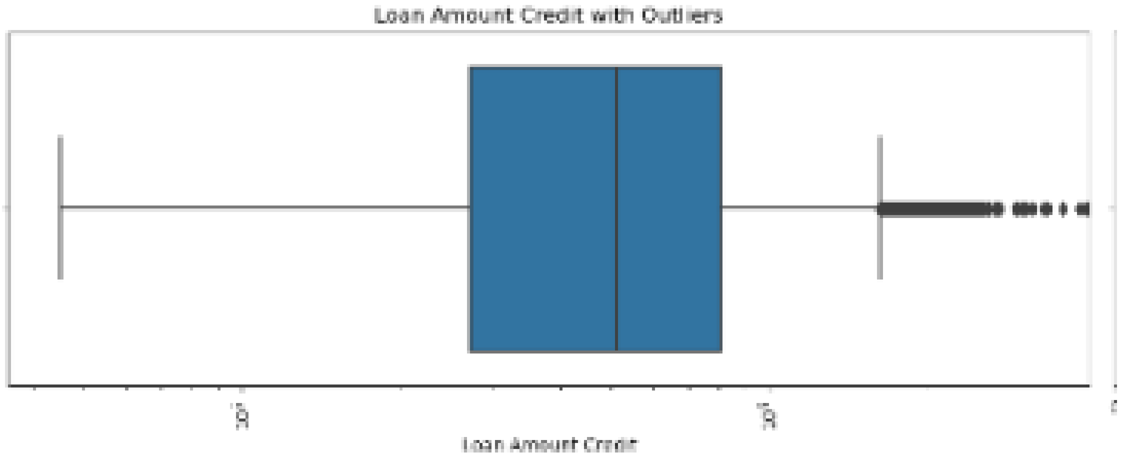
Exploratory Data Analysis: Outlier Detection

The variables Total Income Amount, and loan amount credit have outliers present.

The plot shows the presence of outliers in the data set.



Total Income Amount with Outliers

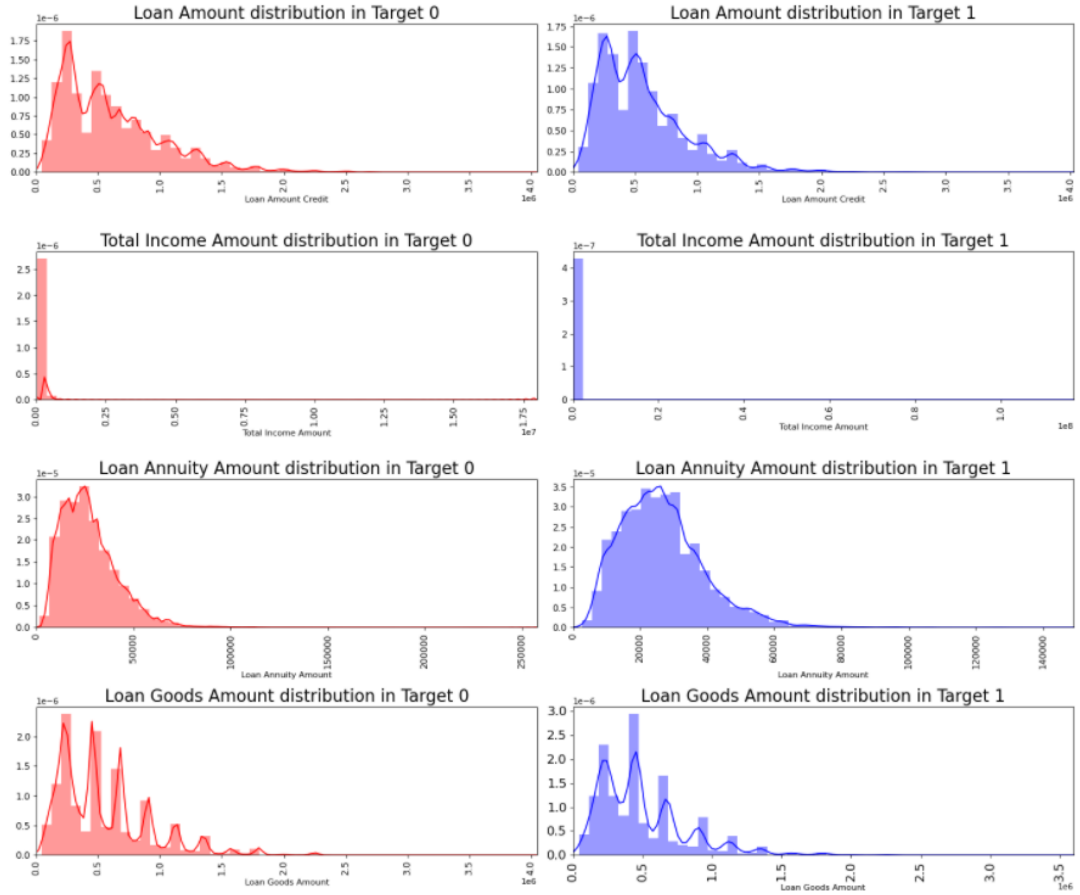


Loan Amount Credit with Outliers

Exploratory Data Analysis: Univariate Analysis(Numerical Variables)

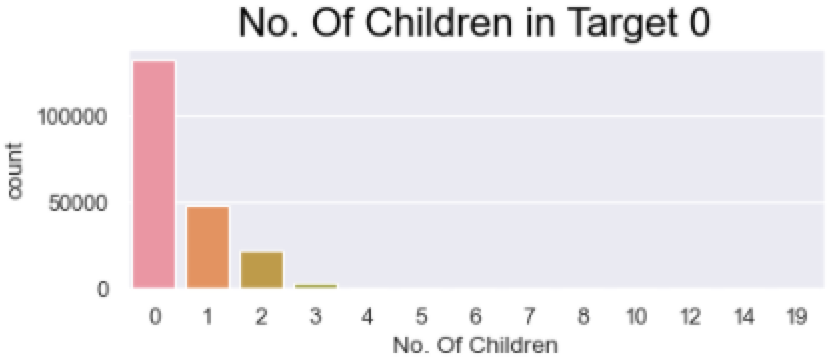
We have plotted the Loan amount credit for target 0 and 1 as well as total income amount for target 0 and 1 both

As we can see in the plots, there is no significant difference in plot of loan amount and total income for both target 0 and target 1 data set.

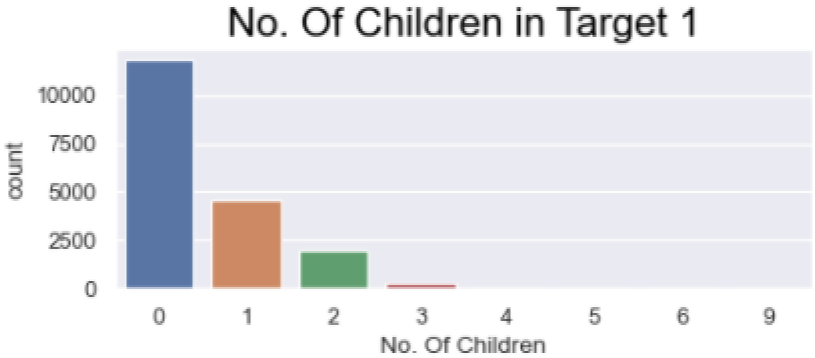


Exploratory Data Analysis: Univariate Analysis (Categorical Variables)

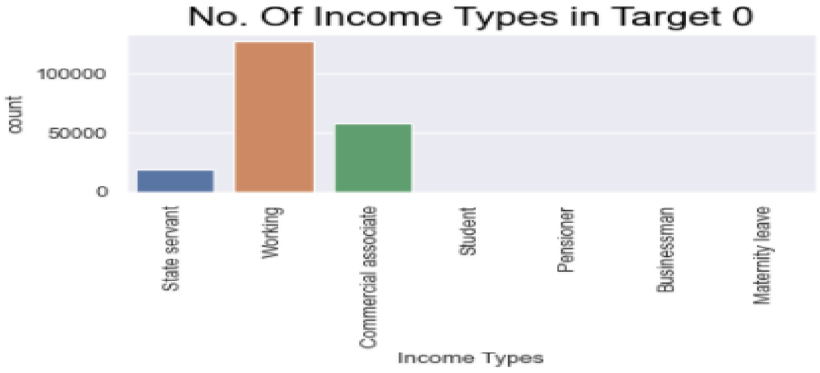
We have plotted bar graph of count of children and count of family members where we can see no significant observation which we can draw from it.



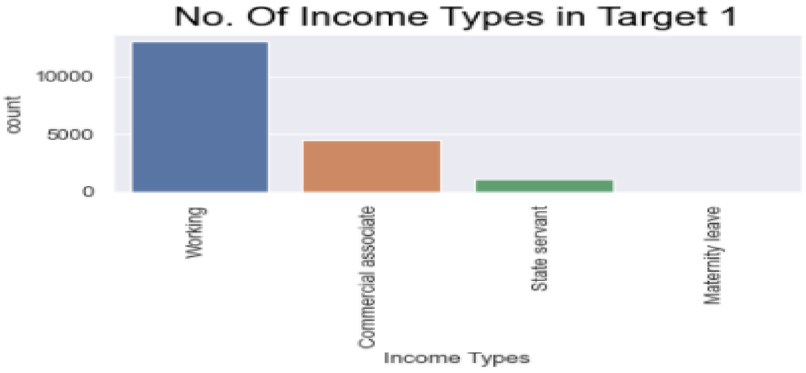
No. Of Children in Target 0



No. Of Children in Target 1



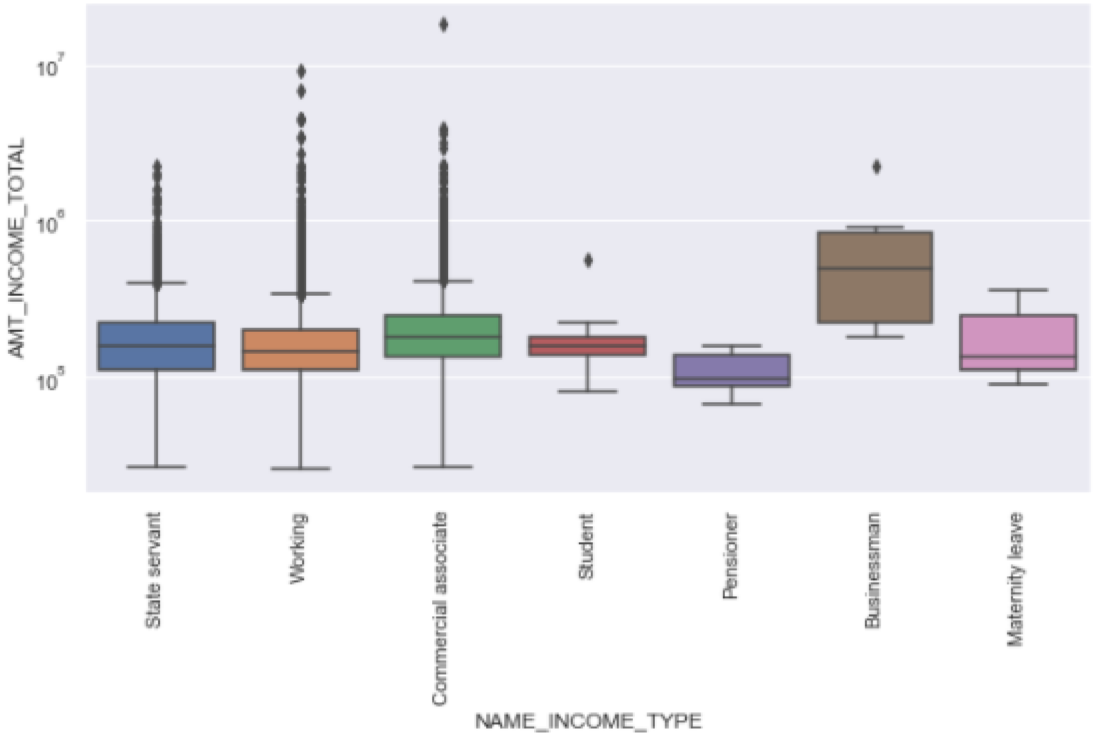
No. Of Income Types in Target 0

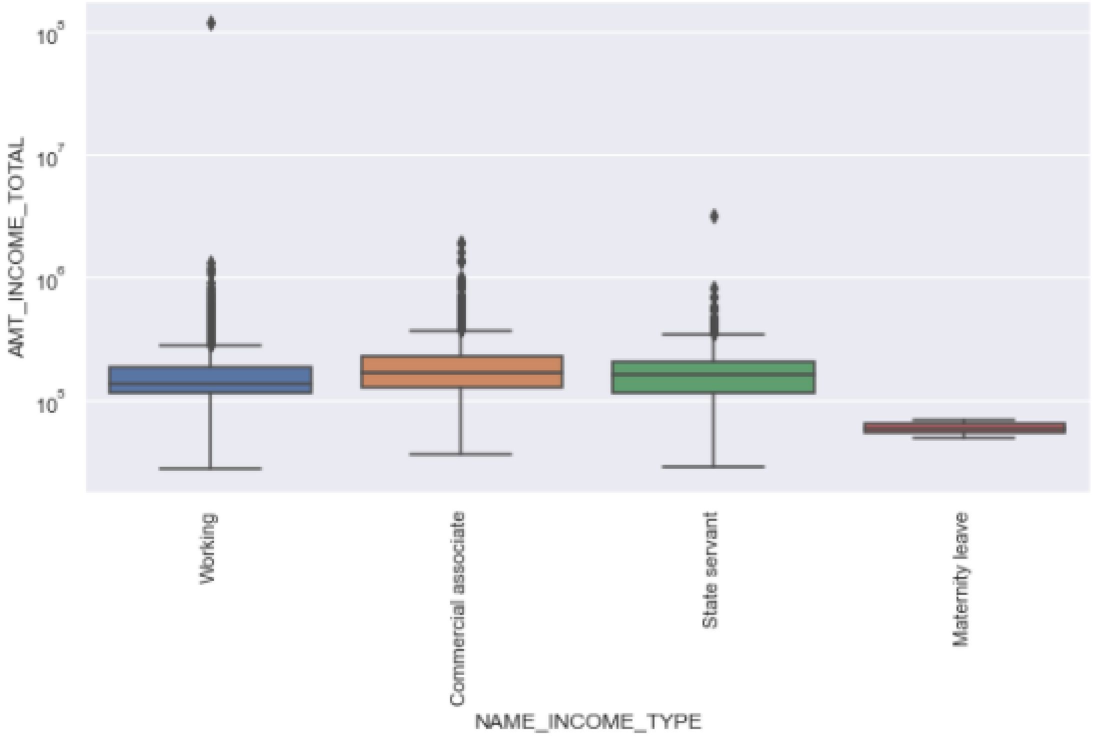


No. Of Income Types in Target 1

Exploratory Data Analysis: Bivariate Analysis

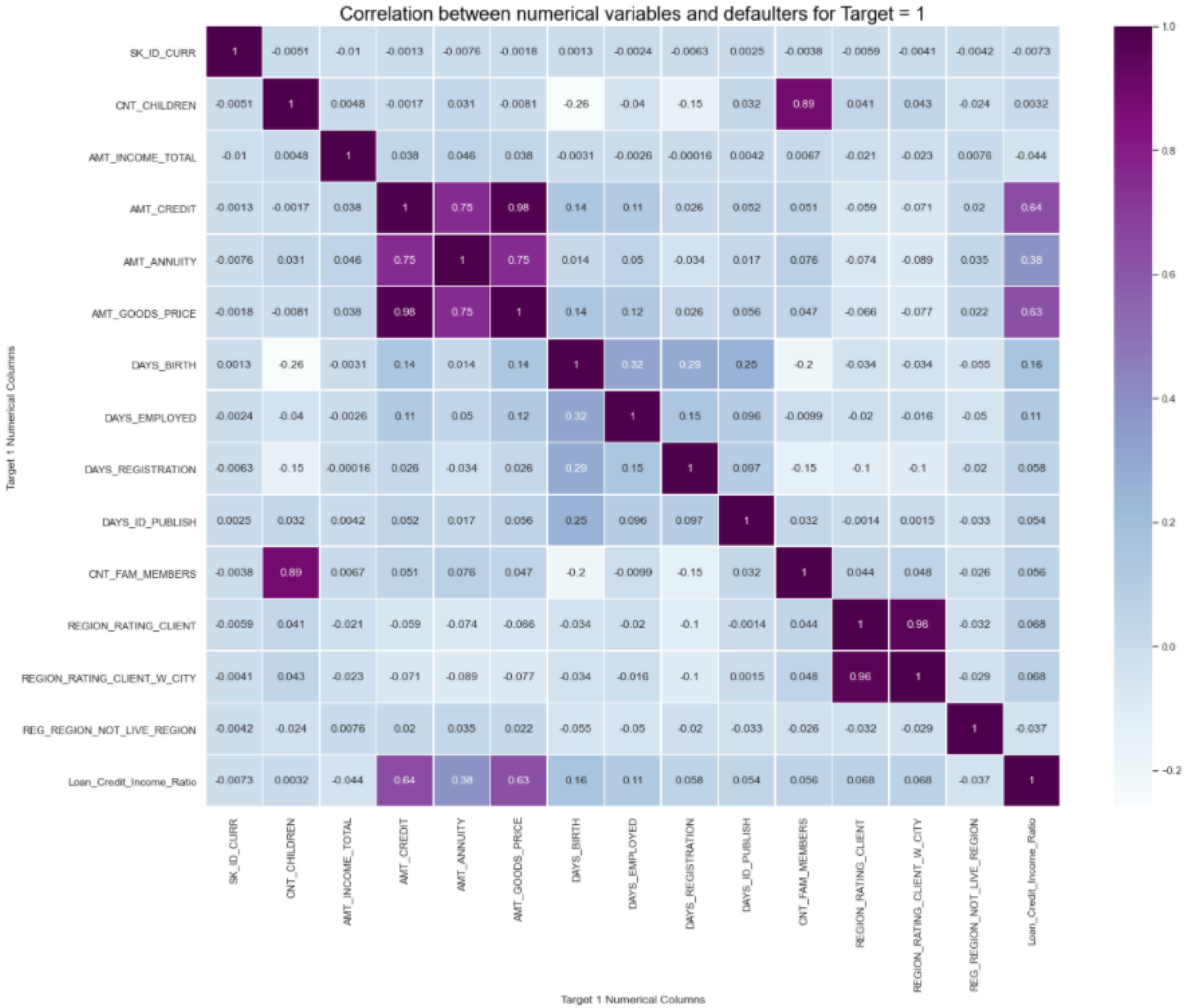
Comparing the Defaulters list and the Non-Defaulters list, we can see that the Pensioners and Students are less likely to default.





Exploratory Data Analysis: Correlation

- The maroon boxes show the most correlated sub-sectors. Specifically, (0.98), (0.96), (089), (0.75) and (0.74) are the most correlated pairs.
- As inferred from the Correlation: Amount Credit and Amount of Goods are highly correlated.
- Also, Amount credit and Amount Annuity are highly correlated pairs as inferred from the naming convention

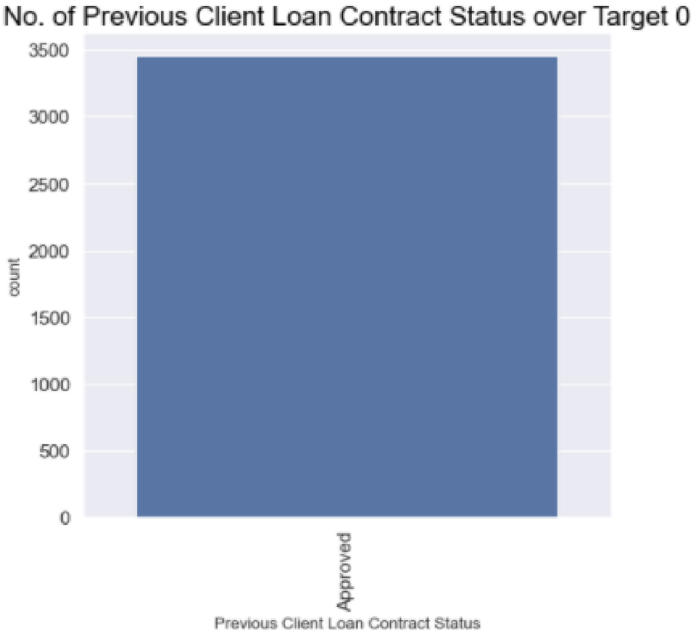


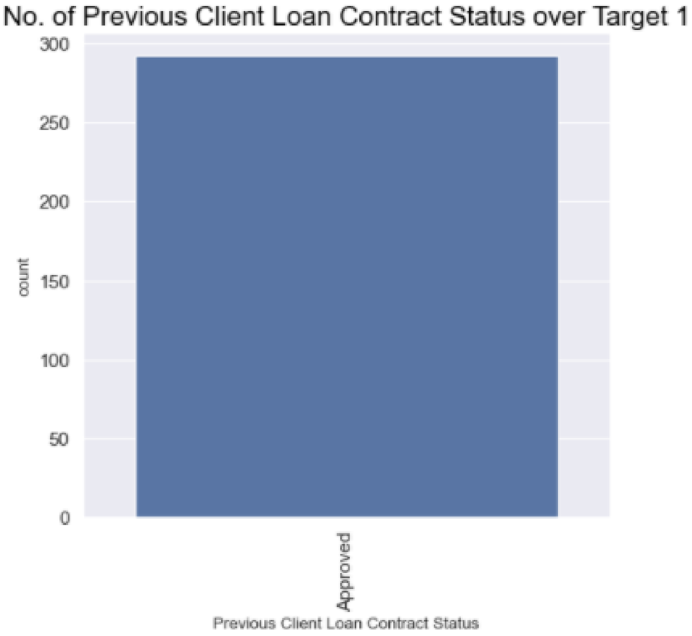
Correlation between numerical variables and defaulters for Target = 1

	SK_ID_CURR	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	REG_REGION_NOT_LIVE_REGION	Loan_Credit_Income_Ratio
SK_ID_CURR	1	-0.0051	-0.01	-0.0013	-0.0076	-0.0018	0.0013	-0.0024	-0.0063	0.0025	-0.0038	-0.0059	-0.0041	-0.0042	-0.0073
CNT_CHILDREN	-0.0051	1	0.0048	-0.0017	0.031	-0.0081	-0.26	-0.04	-0.15	0.032	0.89	0.041	0.043	-0.024	0.0032
AMT_INCOME_TOTAL	-0.01	0.0048	1	0.038	0.046	0.038	-0.0031	-0.0026	-0.00016	0.0042	0.0067	-0.021	-0.023	0.0076	-0.044
AMT_CREDIT	-0.0013	-0.0017	0.038	1	0.75	0.98	0.14	0.11	0.026	0.052	0.051	-0.059	-0.071	0.02	0.64
AMT_ANNUITY	-0.0076	0.031	0.046	0.75	1	0.75	0.014	0.05	-0.034	0.017	0.076	-0.074	-0.089	0.035	0.38
AMT_GOODS_PRICE	-0.0018	-0.0081	0.038	0.98	0.75	1	0.14	0.12	0.026	0.056	0.047	-0.066	-0.077	0.022	0.63
DAYS_BIRTH	0.0013	-0.26	-0.0031	0.14	0.014	0.14	1	0.32	0.29	0.25	-0.2	-0.034	-0.034	-0.055	0.16
DAYS_EMPLOYED	-0.0024	-0.04	-0.0026	0.11	0.05	0.12	0.32	1	0.15	0.096	-0.0099	-0.02	-0.016	-0.05	0.11
DAYS_REGISTRATION	-0.0063	-0.15	-0.00016	0.026	-0.034	0.026	0.29	0.15	1	0.097	-0.15	-0.1	-0.1	-0.02	0.058
DAYS_ID_PUBLISH	0.0025	0.032	0.0042	0.052	0.017	0.056	0.25	0.096	0.097	1	0.032	-0.0014	0.0015	-0.033	0.054
CNT_FAM_MEMBERS	-0.0038	0.89	0.0067	0.051	0.076	0.047	-0.2	-0.0099	-0.15	0.032	1	0.044	0.048	-0.026	0.056
REGION_RATING_CLIENT	-0.0059	0.041	-0.021	-0.059	-0.074	-0.066	-0.034	-0.02	-0.1	-0.0014	0.044	1	0.96	-0.032	0.068
REGION_RATING_CLIENT_W_CITY	-0.0041	0.043	-0.023	-0.071	-0.089	-0.077	-0.034	-0.016	-0.1	0.0015	0.048	0.96	1	-0.029	0.068
REG_REGION_NOT_LIVE_REGION	-0.0042	-0.024	0.0076	0.02	0.035	0.022	-0.055	-0.05	-0.02	-0.033	-0.026	-0.032	-0.029	1	-0.037
Loan_Credit_Income_Ratio	-0.0073	0.0032	-0.044	0.64	0.38	0.63	0.16	0.11	0.058	0.054	0.056	0.068	0.068	-0.037	1

Exploratory Data Analysis: Univariate analysis on Previous Application Data

Based on the graph, we cannot see any significant impact on defaulters and non-defaulters list when we consider the previous variables for analysis.


No. of Previous Client Loan Contract Status over Target 0


No. of Previous Client Loan Contract Status over Target 1

Conclusion

We can see that the Credit Amount and Amount of Goods are highly correlated.

We see a similar correlation between Credit Amount and Annuity Amount.

After analysis, we can conclude that pensioners and students are less likely to default on loan repayment.

We also see that the females are relatively higher in the defaulters list than males.

And we can also see that the defaulters are usually lying between 0-30 years of age.

Thank you