Credit EDA
Case Study

- Dhairyasheel Jadhav

Introduction

- This assignment aims to give you an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Problem Statement

- ➢ When the company receives a loan application, the company has to decide for loan approval based on the applicant’s profile. Two types of risks are associated with the bank’s decision:
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.
- ▪ Use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

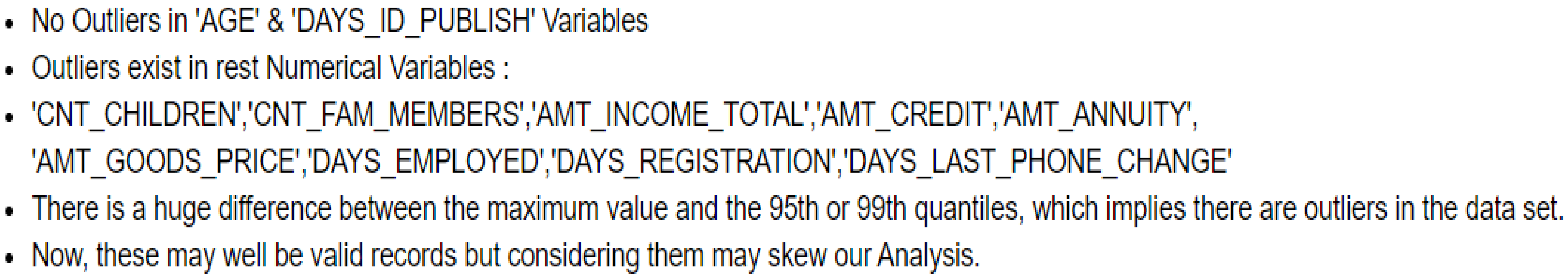
Business Objectives

- 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. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.
- 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.

Steps :

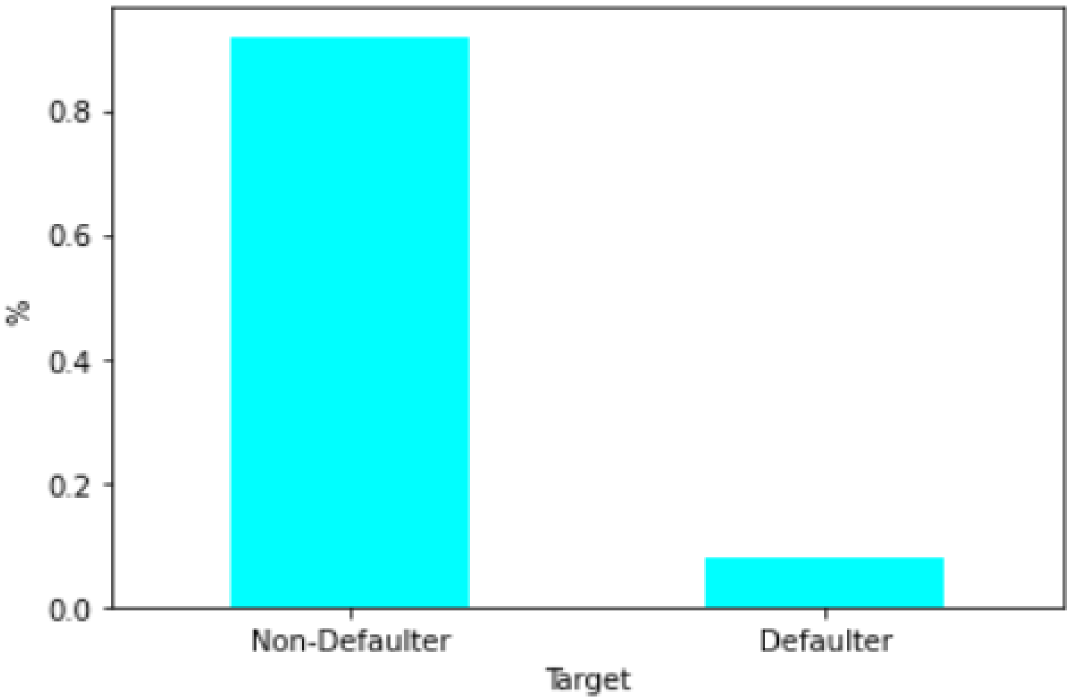
- Check the Structure of the Data
- Data Quality Check & Missing Values
- Removing Insignificant Columns
- Handling Missing Values
- Binning
- Check the Imbalance percentage
- Splitting 'Data' Data frame into two Sets
- Univariate Analysis
- Numerical Columns (Outliers Analysis)
- Categorical Columns
- Segmented Univariate Analysis
- Categorical Columns w.r.t Target Variable
- Top 10 correlation w.r.t Target variable
- Bivariate Analysis
- Observations & Recommendations

Outliers Analysis

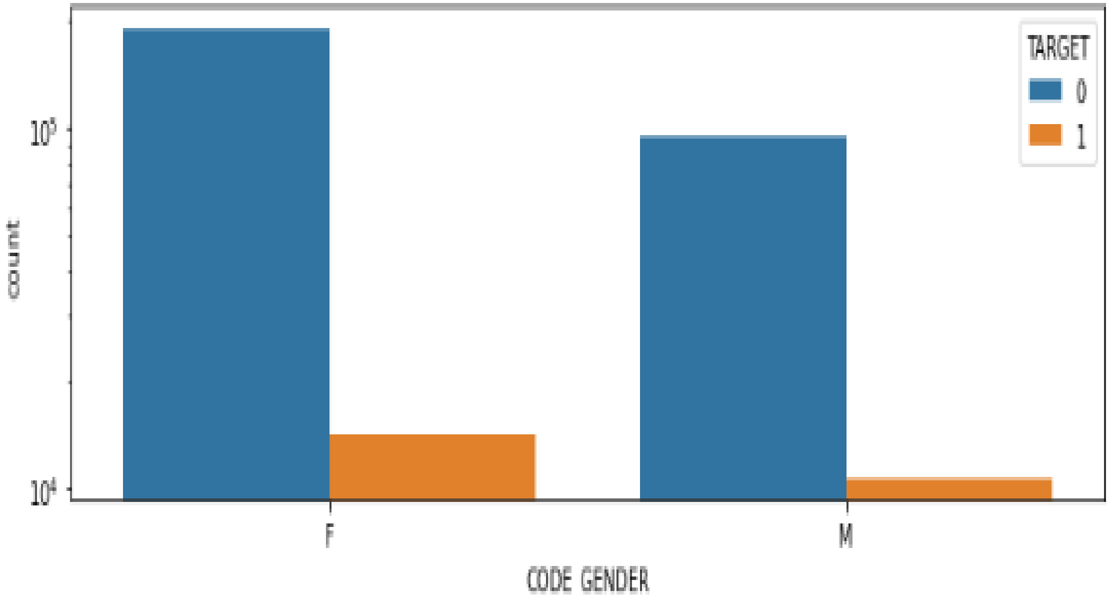


- No Outliers in 'AGE' & 'DAYS_ID_PUBLISH' Variables
- Outliers exist in rest Numerical Variables :
- 'CNT_CHILDREN','CNT_FAM_MEMBERS','AMT_INCOME_TOTAL','AMT_CREDIT','AMT_ANNUITY', 'AMT_GOODS_PRICE','DAYS_EMPLOYED','DAYS_REGISTRATION','DAYS_LAST_PHONE_CHANGE'
- There is a huge difference between the maximum value and the 95th or 99th quantiles, which implies there are outliers in the data set.
- Now, these may well be valid records but considering them may skew our Analysis.

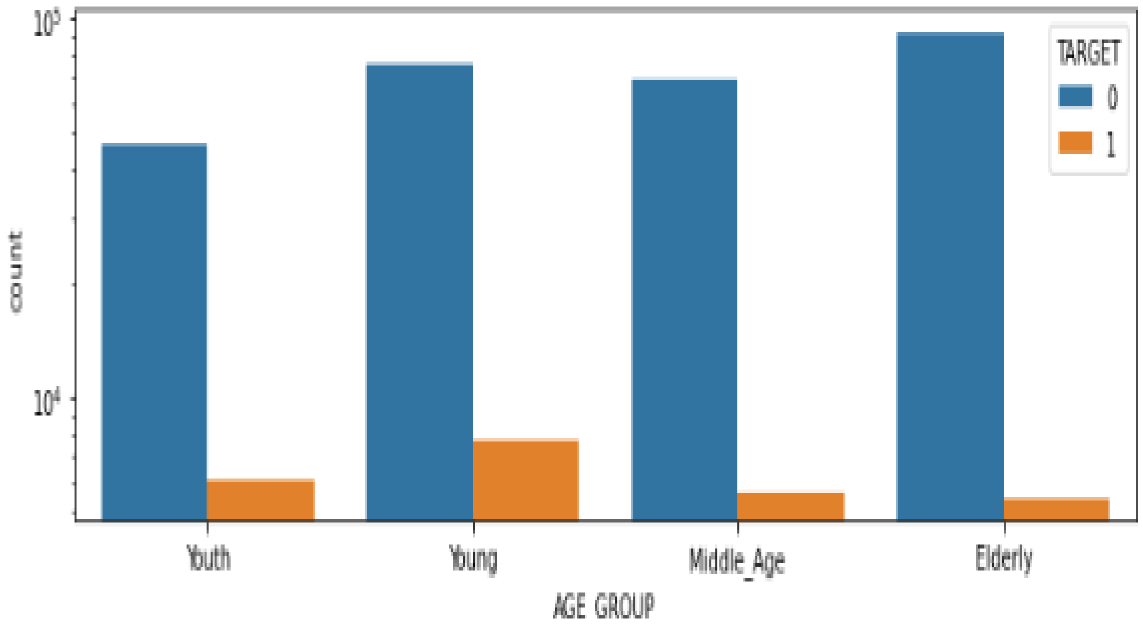
Imbalance percentage



- 0 : Non-Defaulter | 1 : Defaulter
- 91.93% Clients Paid Installments on time.
- 8.07% Clients Defaulted on Installment Payments.
- Distribution concludes this is Highly Imbalanced Dataset.



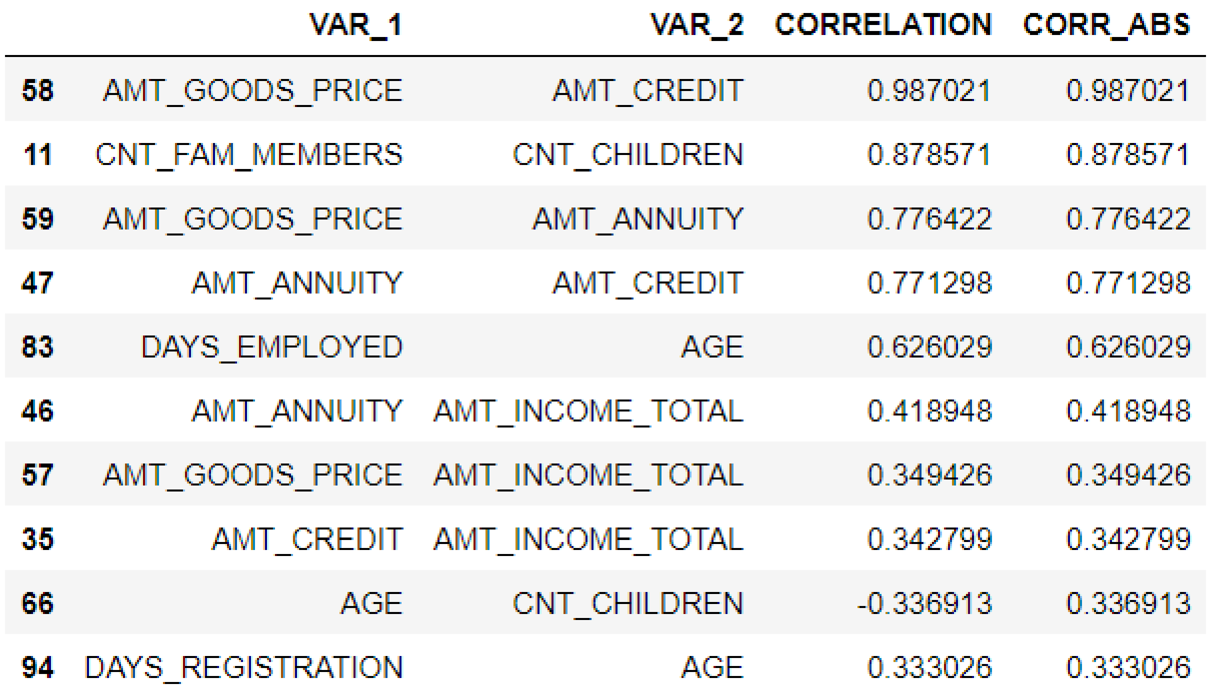
- Proportion of Females Applying for Loans & Defaulting is High compared to Males



- Elderly are taking most loans, but least likely to Default. Its reasonable since they might be having high savings
- Young(30-40) come second, but they're most likely to Default

All Relevant Charts are Attached in Notebook with in-depth Analysis

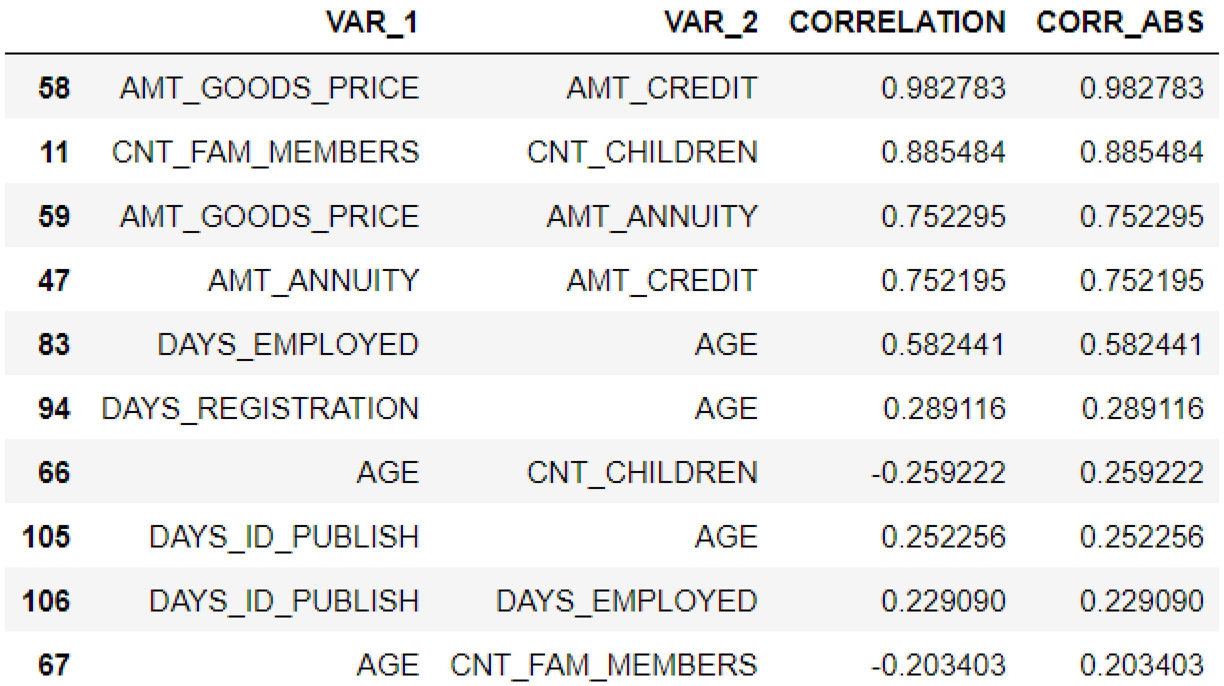
Top 10 correlation w.r.t Target variable



	VAR_1	VAR_2	CORRELATION	CORR_ABS
58	AMT_GOODS_PRICE	AMT_CREDIT	0.987021	0.987021
11	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571	0.878571
59	AMT_GOODS_PRICE	AMT_ANNUITY	0.776422	0.776422
47	AMT_ANNUITY	AMT_CREDIT	0.771298	0.771298
83	DAYS_EMPLOYED	AGE	0.626029	0.626029
46	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418948	0.418948
57	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349426	0.349426
35	AMT_CREDIT	AMT_INCOME_TOTAL	0.342799	0.342799
66	AGE	CNT_CHILDREN	-0.336913	0.336913
94	DAYS_REGISTRATION	AGE	0.333026	0.333026

Non Defaulting Clients (Target = 0)

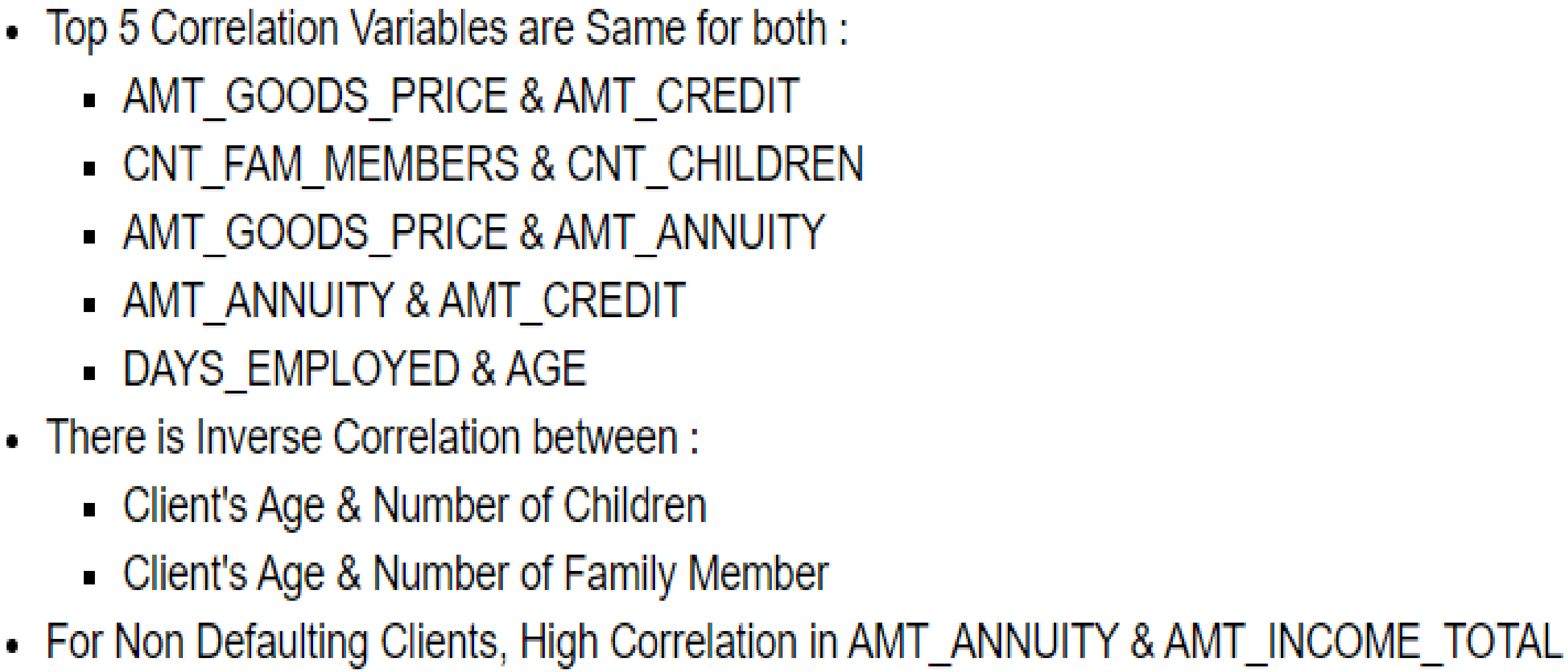
Top 10 correlation w.r.t Target variable



	VAR_1	VAR_2	CORRELATION	CORR_ABS
58	AMT_GOODS_PRICE	AMT_CREDIT	0.982783	0.982783
11	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484	0.885484
59	AMT_GOODS_PRICE	AMT_ANNUITY	0.752295	0.752295
47	AMT_ANNUITY	AMT_CREDIT	0.752195	0.752195
83	DAYS_EMPLOYED	AGE	0.582441	0.582441
94	DAYS_REGISTRATION	AGE	0.289116	0.289116
66	AGE	CNT_CHILDREN	-0.259222	0.259222
105	DAYS_ID_PUBLISH	AGE	0.252256	0.252256
106	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.229090	0.229090
67	AGE	CNT_FAM_MEMBERS	-0.203403	0.203403

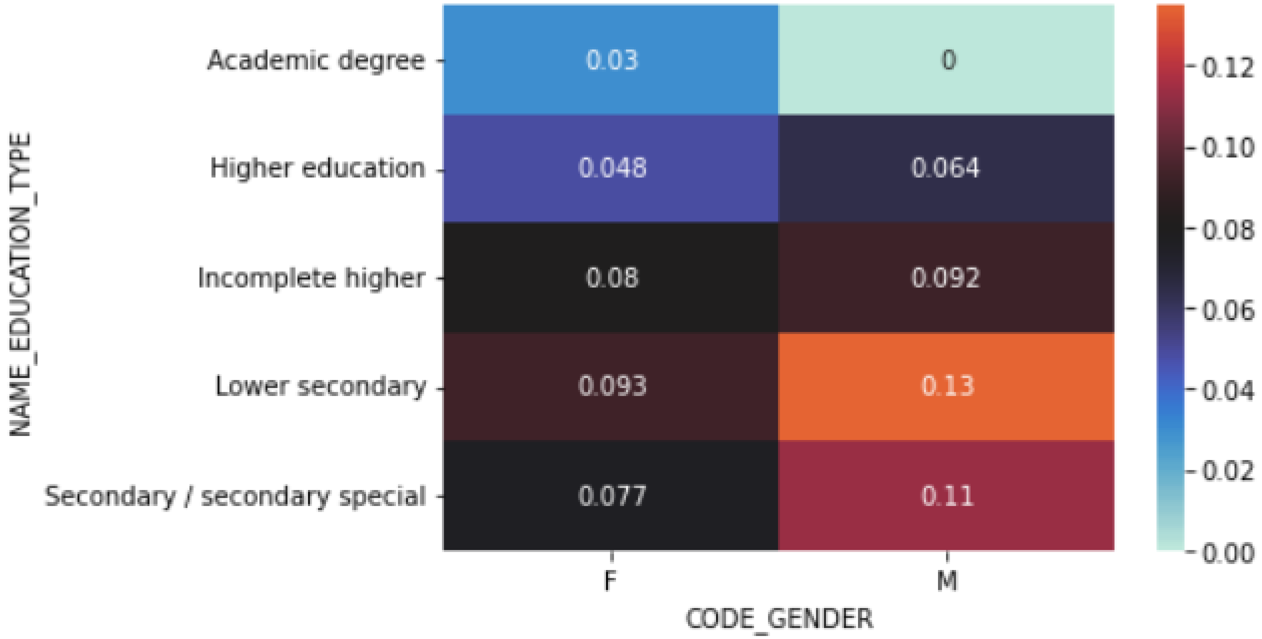
Defaulting Clients (Target = 1)

Correlation Insights



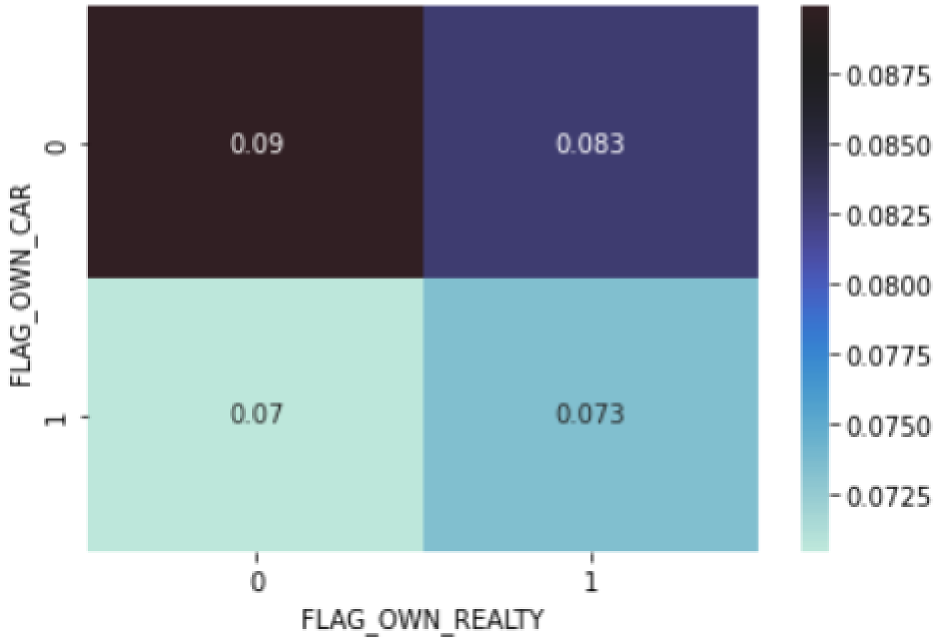
- Top 5 Correlation Variables are Same for both :
  - AMT_GOODS_PRICE & AMT_CREDIT
  - CNT_FAM_MEMBERS & CNT_CHILDREN
  - AMT_GOODS_PRICE & AMT_ANNUITY
  - AMT_ANNUITY & AMT_CREDIT
  - DAYS_EMPLOYED & AGE
- There is Inverse Correlation between :
  - Client's Age & Number of Children
  - Client's Age & Number of Family Member
- For Non Defaulting Clients, High Correlation in AMT_ANNUITY & AMT_INCOME_TOTAL

EDUCATION_TYPE, GENDER with TARGET



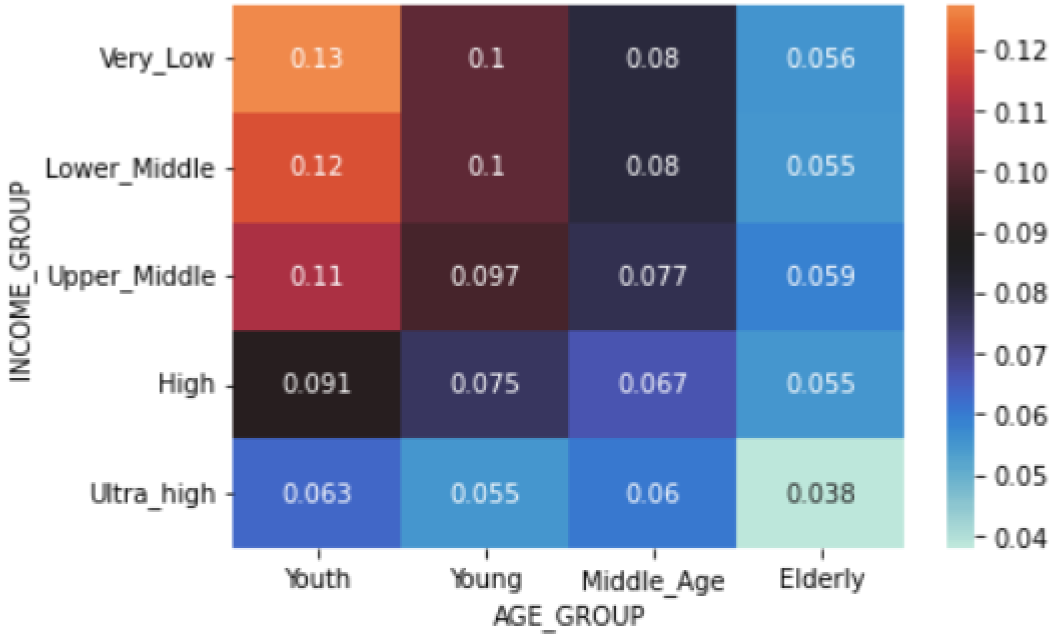
- Lower Secondary Educated are Defaulting Very much
- Academic Degree Holders are Defaulting Very Less

Car & House Ownership with TARGET



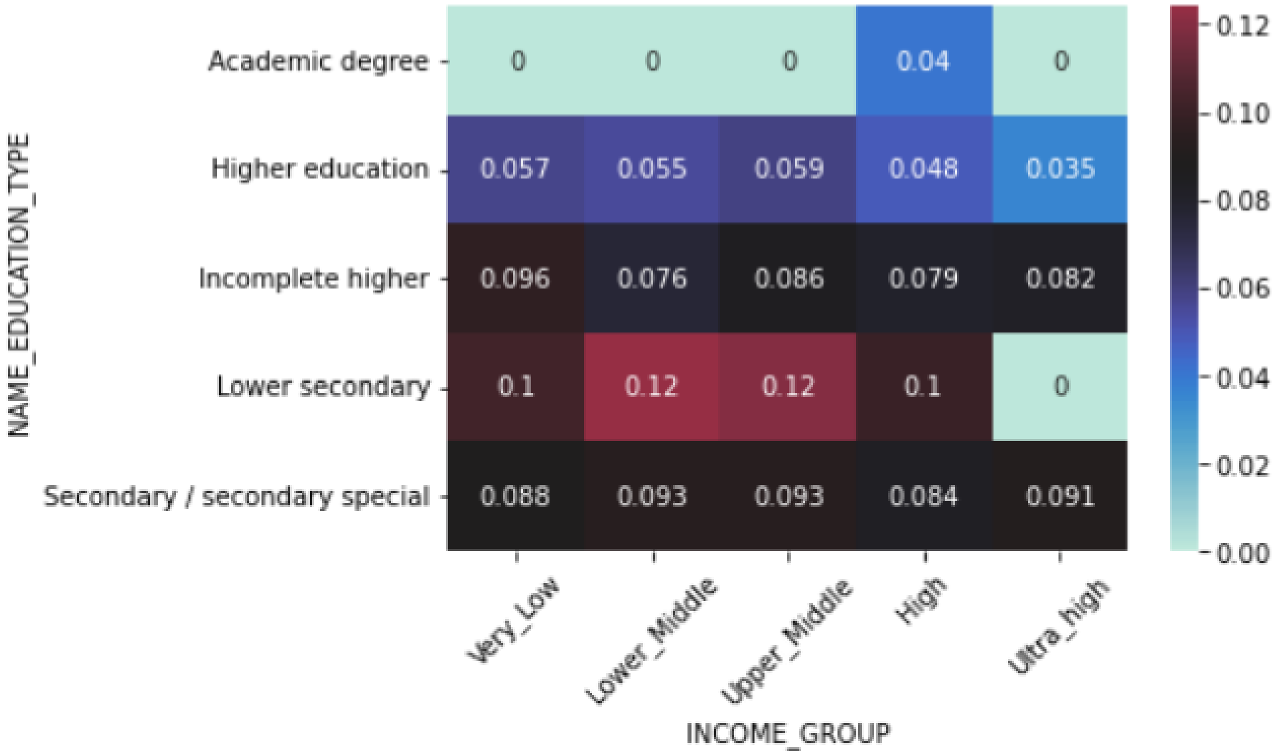
- Default Rate in Clients Not Owning Car & House is very High
- Default Rate in Clients Owning Car & Not Owning House is Low

INCOME_GROUP & AGE_GROUP with TARGET



- Default rate in Low Income Youth is very High
- Default Rate in Ultra Rich Elderly is Very Low

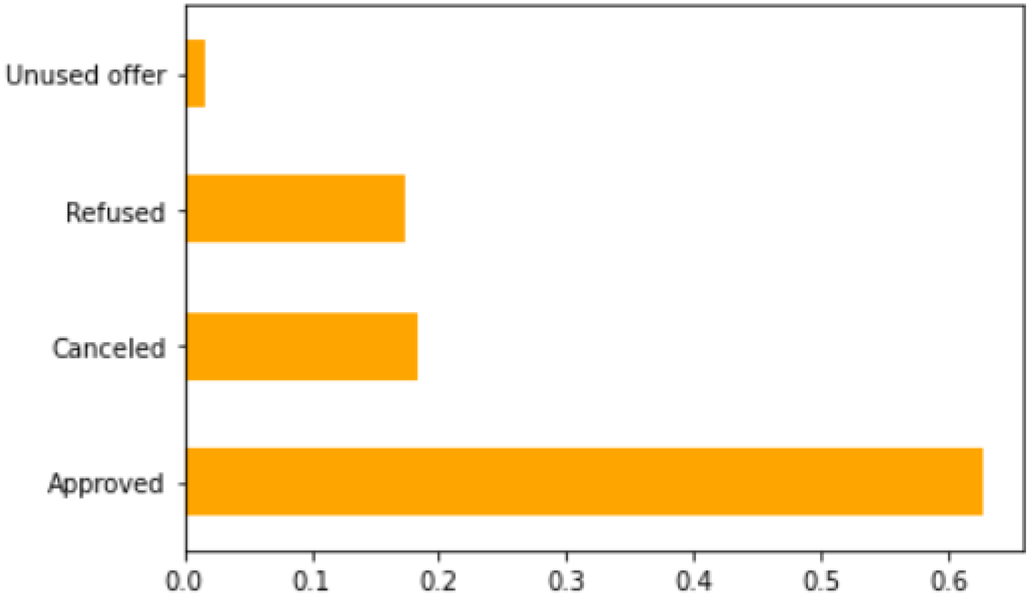
EDUCATION_TYPE & INCOME_GROUP with TARGET



	Very_Low	Lower_Middle	Upper_Middle	High	Ultra_high
Academic degree	0	0	0	0.04	0
Higher education	0.057	0.055	0.059	0.048	0.035
Incomplete higher	0.096	0.076	0.086	0.079	0.082
Lower secondary	0.1	0.12	0.12	0.1	0
Secondary / secondary special	0.088	0.093	0.093	0.084	0.091

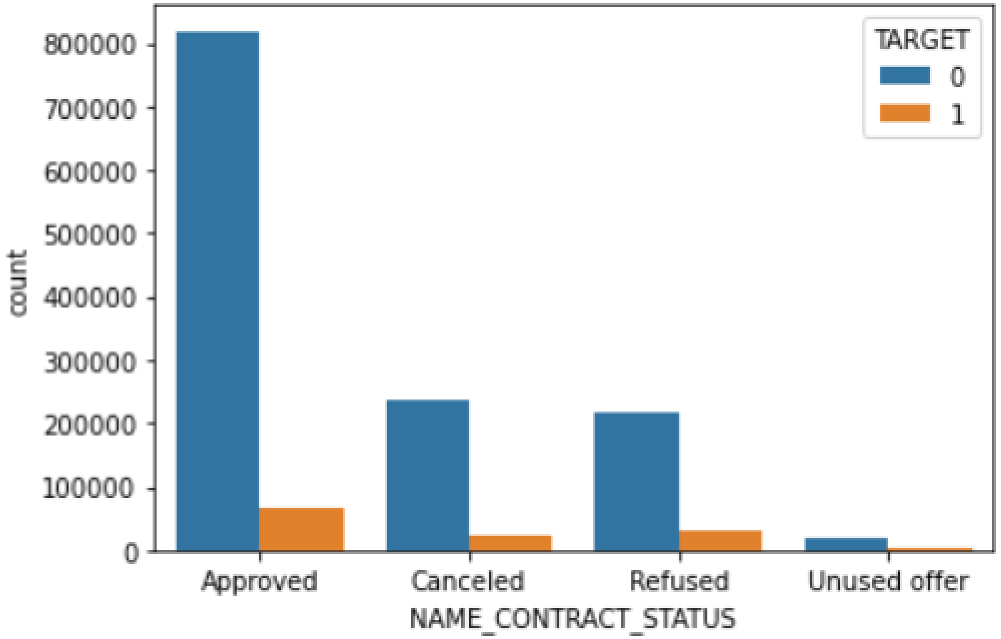
- Very High Default Rate in Low Income Class with Lower Secondary Education
- Very Low Default Rate in Clients having Academic Degree

'NAME_CONTRACT_STATUS' Against 'TARGET' Variable



- Approval Rate is 62.68%
- Rejection Rate is 17.36%
- Cancellation Rate is 18.35%

'NAME_CONTRACT_STATUS’ Vs 'TARGET'



- Loans Previously Refused have High Default Rate

Observations & Recommendations

- 8.07% Clients Defaulted on Installment Payments.
- Distribution concludes this is Highly Imbalanced Dataset.
- Proportion of Females Applying for Loans & Defaulting is High. So, Targeting Must be focused on Males.
- Elderly are taking most loans, but least likely to Default. Young(30-40) come second, but they're most likely to Default. So, this Group must be Scrutinized well before Loan Approval.
- Proportion of Clients not Owning Car & Defaulting is High & Clients Owning Car & Defaulting is Low.
- Proportion of clients Owning House & Defaulting is Very High. Clients Not Owning House & Defaulting is Low.
- Clients Not Owning Car & House must be Avoided. Clients Owning Car & Not Owning House must be Targeted.
- Lending is done more in Cash Loans which has very High Defaults. Cash loans must be Scrutinized well & Revolving loans must be Approved as it has Low Defaults.

- Defaults amongst Unaccompanied Clients is very High & must be Scrutinized well. Clients Accompanied by Group of People be Preferred.
- Very High Defaults amongst Working Class, Commercial associates, Pensioners, on Maternity leave & Unemployed Class & must be Scrutinized well.
- Very Low Income Clients has High Defaults & must be Scrutinized well. Low Income Youth must be Avoided & Ultra Rich Elderly must be Targeted.
- Clients with Lower Secondary Education must be Avoided & Academic Degree Holder Clients must be Targeted.
- Very High Defaults in Married Clients
- Defaults amongst Clients staying in their own House or with Parents is Very High. Clients staying in Office or Co-Op Apartments must be preferred.
- Defaults amongst Laborers & Sales Staff is very High & must be avoided. IT & HR Staff clients be targeted.
- Defaults amongst Business Entities & Self-employed is very High. So, must be Scrutinized thoroughly. Trade & Industry Type Occupations preferred.

- Outliers exist in many Numerical Variables & must be treated.Now, these may well be valid records but considering them may skew our Analysis.
- Most Numerical Variables follow same Distribution & Pattern across Defaulting & Non Defaulting Clients
- Median Age of Defaulting Clients is lower compared to Non Defaulting Clients
- Defaulting clients Change their Identity Document, Registration & Phone little sooner before application
- Defaulting clients started application little early in their current employment

- Variables with Highest Correlation are the same in both Defaulting & Non Defaulting Clients
- There is Inverse Correlation between :

  - Client's Age & Number of Children

  - Client's Age & Number of Family Member
- For Non Defaulting Clients, High Correlation in Loan Annuity & Income of Client

Thank You