CREDIT EDA CASE STUDY

BY

VAISHNAVI SAMPATH KUMAR

Problem Statement

Finance companies decide for loan approvals based on the applicant's profile.

Two types of risks are associated with the bank's decision:

•**Business loss**: Rejecting the loan for a capable client

•**Financial loss**: Approving the loan for a potential defaulter

<u>This case study aims to analyze the data patterns and ensure that the applicants capable of repaying the loan are not rejected</u>

Problem Solving Approach

1. Business Understanding (covered in problem statement)

2. Data Understanding

3. Data Preparation

4. Data Analysis

5. Observations

6. Recommendations

Data Understanding

Application data: Client information at the time of current application.

- Huge dataset with 300k rows and 122 attributes
- Many attributes(40+) with high missing values
- Data imbalance for Target attribute- (92% data for non defaulters, and 8 % defaulters)
- For many attributes, 75 percentile of the data is 0
- Negative values for DAYS attribute

Previous application data- Information on clients previous loan data.

- Very huge dataset (16.7MB rows and 13 attributes)
- RATE_DOWN_PAYMENT incorrectly has negative values
- All days columns have negative values
- The dataset will greatly help to draw inferences from clients historic application

We will be merging the 2 datasets to make observations based on correlation and statistical data between the datasets.

Data Preparation

Below steps taken to ensure data quality:

1. Attributes from application and previous data sets with over 45% and 50% missing values removed respectively.
2. Attributes with around 13% missing values imputed :
   - Attributes with outliers can be imputed with median as they are safer approximation.
     Example: AMT_ANNUITY, AMT_REQ_CREDIT_BUREAU_YEAR, AMT_CREDIT, AMT_GOODS_PRICE imputed with median , right skewed data with outliers on the right side of the mean.
   - Attributes that are normally distributed around mean, imputed with mean/median
     Example: EXT_SOURCE_2
   - Categorical attributes can be imputed based on observations
     Example: PRODUCT_COMBINATION : Although it has 17 unique categories, at a high level they correspond to Cash, POS and others, choosing Cash as it has the highest frequency)
3. Derived Attributes: Created below derived attributes which might simplify the analysis:
   - AGE from DAYS_BIRTH , INCOME_RANGE from AMOUNT_INCOME_TOTAL, Corresponding Year attributes for days attributes
   - ORGANIZATION (high level info) from ORGANIZATION_TYPE (holds sub types)
4. Data & types corrected:
   - Attributes having invalid negatives corrected (made positive)
     Example: DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION
   - Datatypes of attributes corrected & Binned continuous variables (AMT_INCOME_TOTAL, Ext_Rating_3, AGE)
     Example: All Flags are categorical data➔ modified data type to object
     All attributes corresponding to days, count, observations converted to integers

Data Analysis Approach

- Based on our personal loan experiences and research finance companies usually check the below factors to make a decision. So we will be analyzing these attributes to confirm the standards.
  1. Credit score (EXT_RATING_3 in our dataset)
  2. Current income (AMT_INCOME_TOTAL)
  3. Employment history (we don't have exact but will be checking YEARS_EMPLOYED)
  4. Repayment history (NAME_CONTRACT_TYPE)

- From the Univariate & Bivariate analysis, below are the attributes that seem to be correlated and might influence the clients payment capability. We will be looking at these attributes more closely.
  1. Clients Age (derived from DAYS_BIRTH)
  2. Loan application amount (AMT_APPLICATION)
  3. Clients owning car/realty (FLAG_OWN_CAR, FLAG_OWN_REALTY)

Assumptions

The loan process is influenced by a variety of factors, and these factors can vary depending on the type of loan, the lender, and the borrower's financial situation.

Here are some of the key factors that can impact the loan process:

1. Loan amount

2. Loan type requested

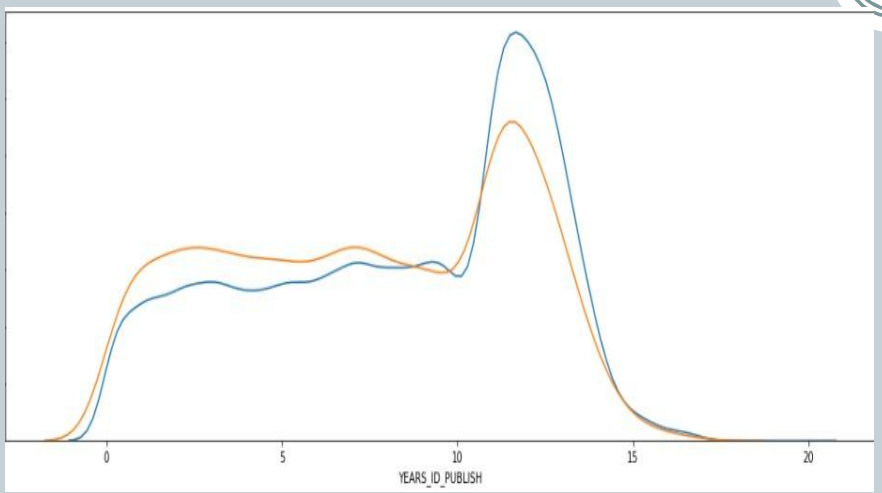
3. Applicants financial standing (Car, Housing type, credit score etc)

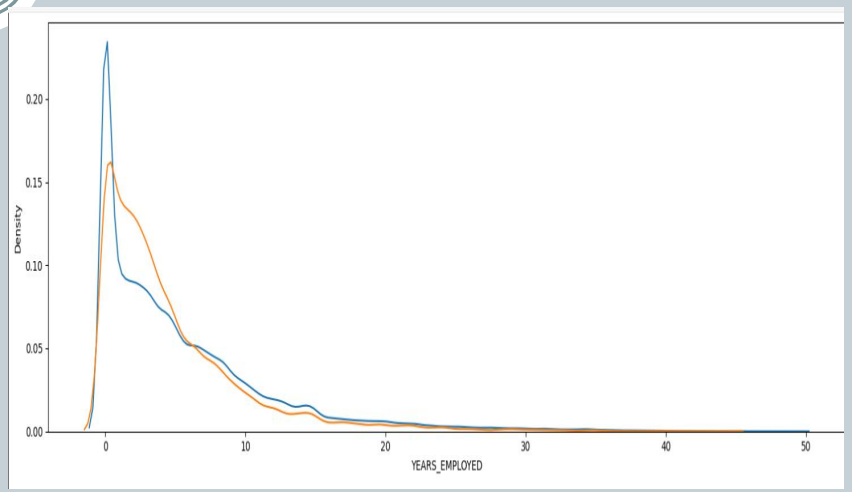
4. Applicant liabilities (Family,children, employment type,Age)

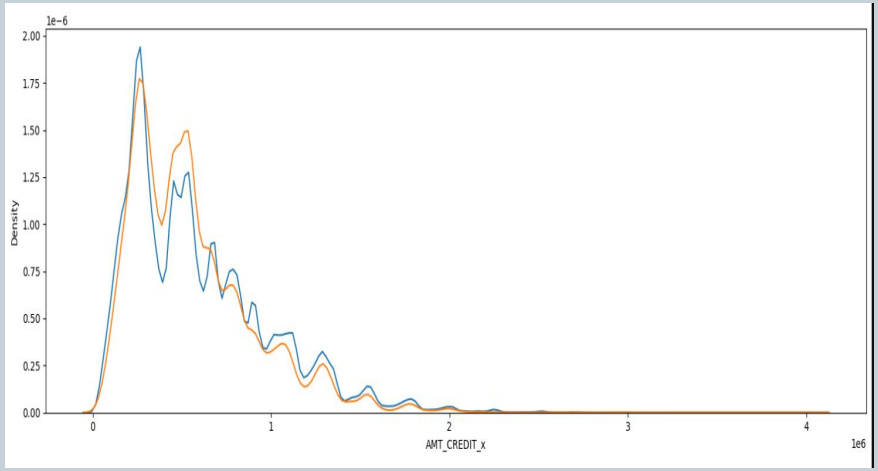
We will inspect the above factors in particular along with other attributes in the dataset to validate our assumptions

Note: The data imbalance in our datasets, we have 92% data of non defaulters so our learnings and analysis might be biased

Univariate Data Analysis (wrt Target)



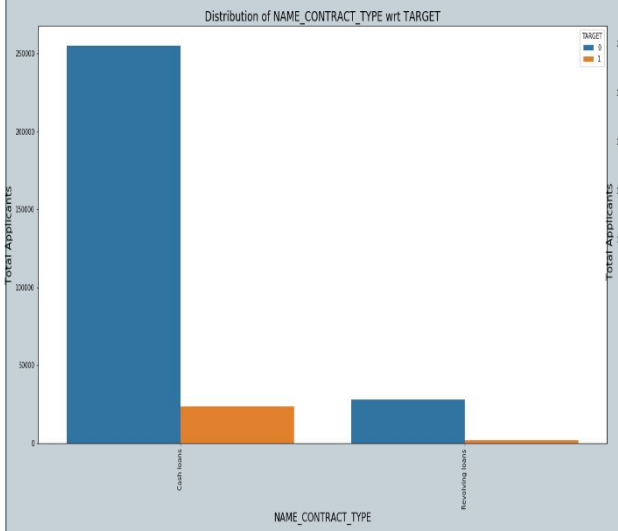


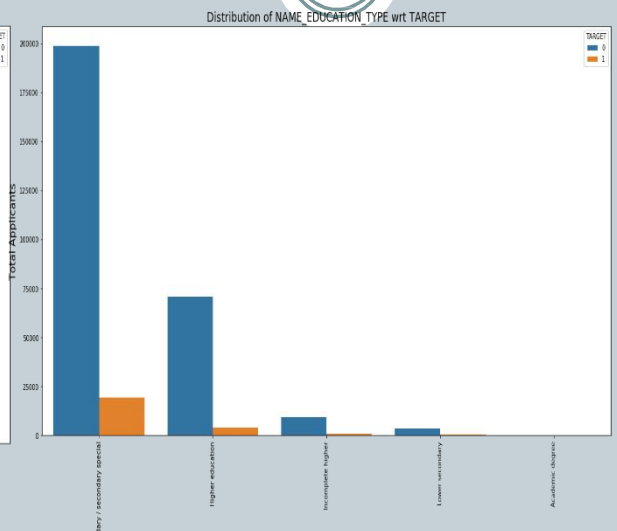


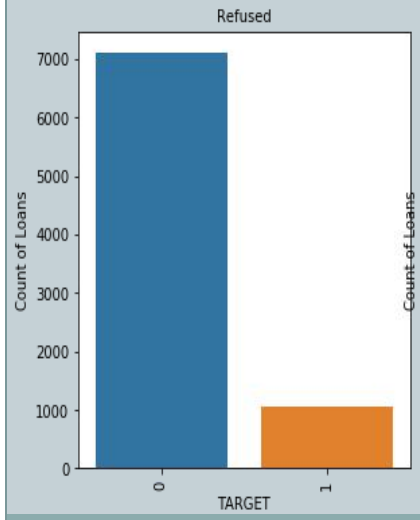
Observations:

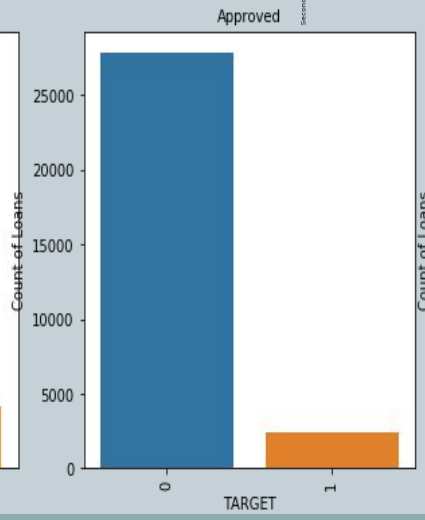
- More defaulters for clients who have been employed between 2-6 years, less defaulter after 40
- More defaulters for those who recently got ID changed/New - may be suspicious activity need to be cautious
- Defaulters for AMT_Credit 3L-5L
- High defaulters for low region rating

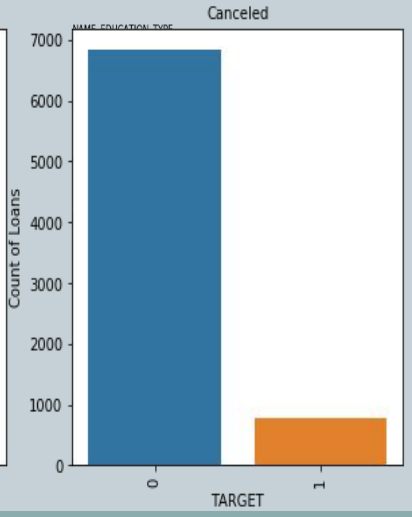
Impact of contract type & Education and Previous loan status on defaults







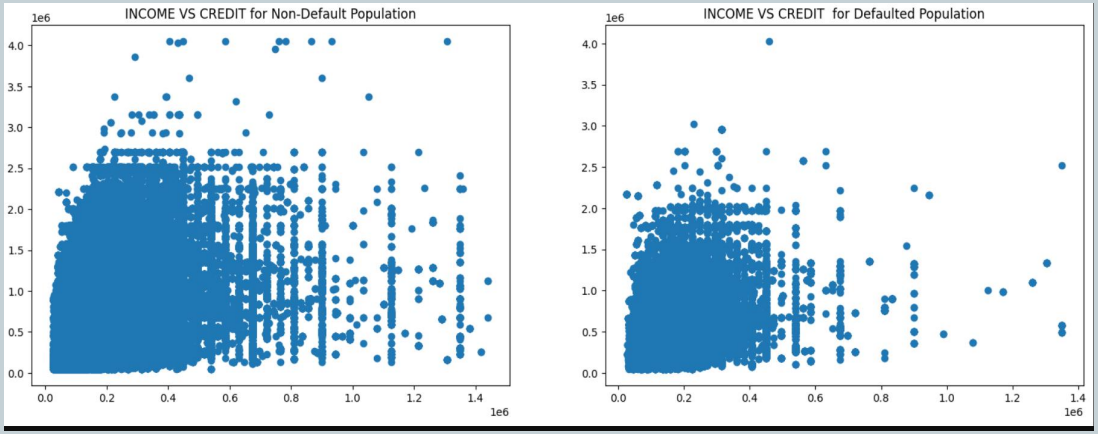




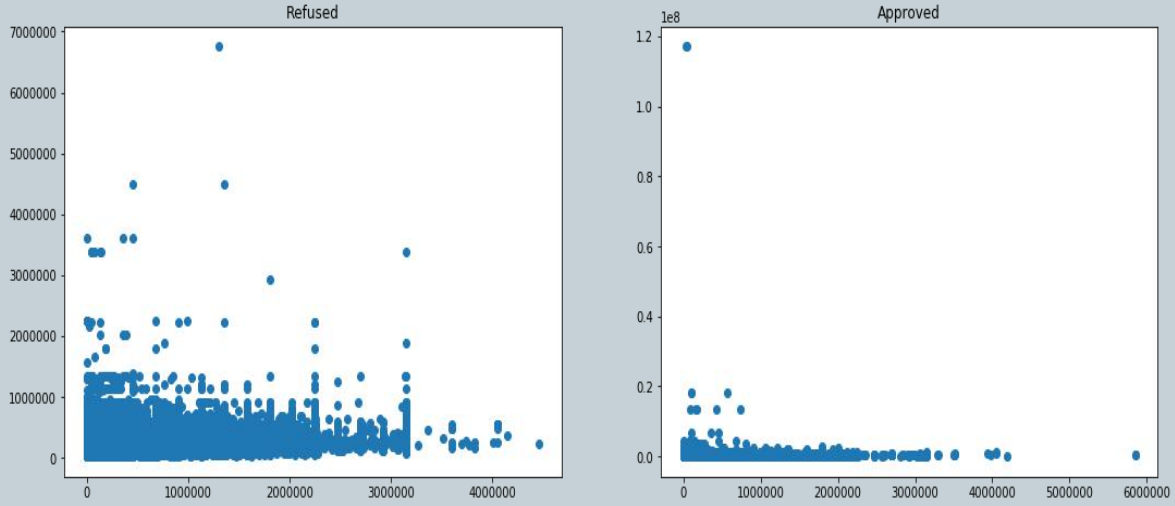
- Defaulted population has 6.7% of revolving loans as compared to 10.6% in Non Defaulted Population. Hence we can infer that **revolving loans are comparatively safer**.
- This may be attributed to the Nature of revolving loan as it is considered a flexible financing tool due to its repayment and re-borrowing flexibility.
- **Loans to clients with Higher Education are comparatively safer.** (16% of clients with higher education as compared to 27% in Non Defaulted). Probably due to the fact that higher educated people would be earning more.
- 13% of current clients whose loans were previously rejected have defaulted. While only 8% of current clients whose loans were previously accepted have defaulted.
- Hence default rate for clients whose applications were previously rejected is higher.

Bivariate Analysis Income Vs Credit (wrt TARGET)

Income Vs Loan Amount

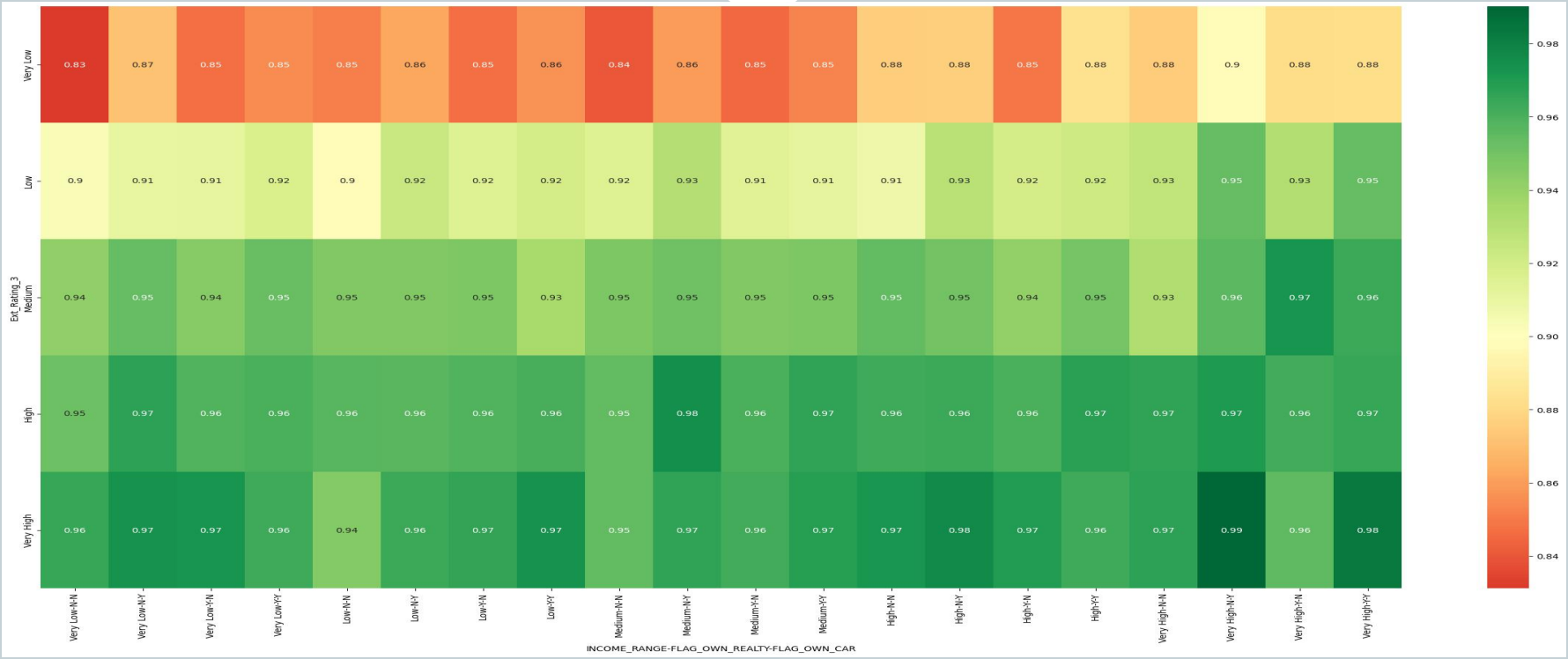


- Lower density of defaults where income is higher than 300k or credit is lower than 200k.



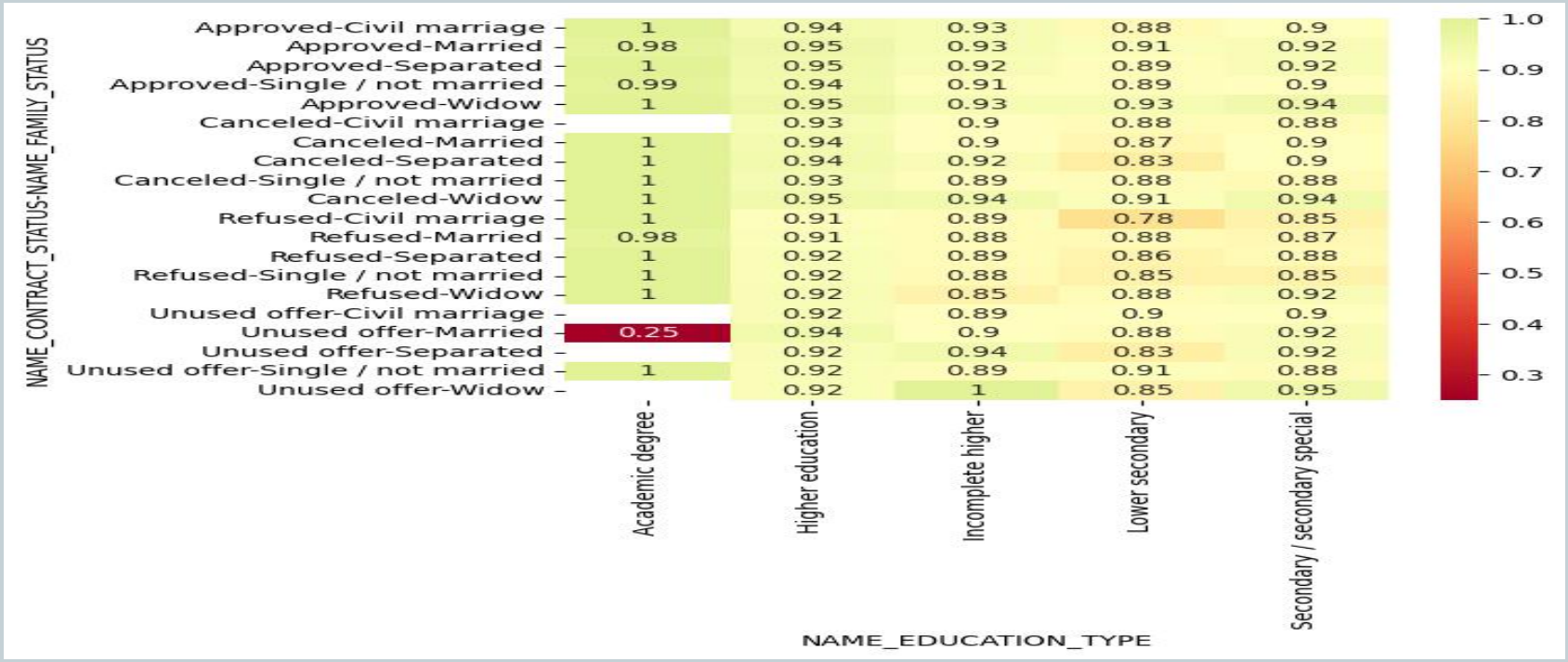
- Loan request higher than 200k had a higher rejection rate. Also loan rejection rate was much lower if the income was higher than 500k.
- Loan request higher than 200k had a higher rejection rate.
- Lower income groups had higher rejection rate

Impact of EXT_SOURCE_3



- Ext_source_3 is a very good predictor of Defaults.
- Clients rated Low have most defaults while those rated very high have least defaults. Hence **Ext_source_3 is predicting defaults with high accuracy.**

Education Type Vs Loan status & Family Status



Target:

- All clients with academic degree except married clients with unused offers
- Widows with incomplete higher education and unused offers

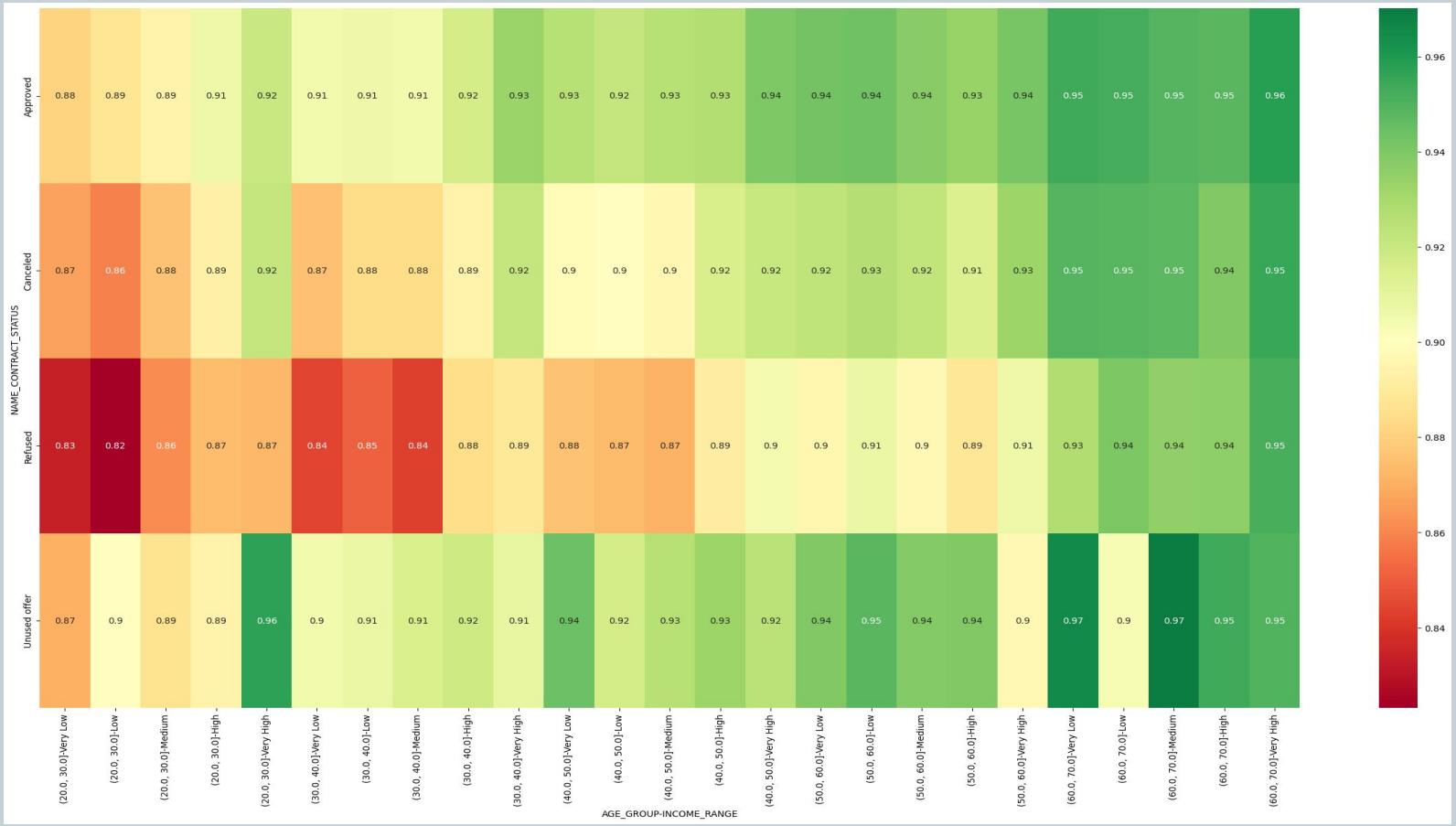
Need to avoid/take caution:

- Married Clients with academic degree who had an unused offer in the past
- Lower secondary degree holders in civil marriage whose previous application was refused

Channel Type Vs Loan status

- No unused offers for car dealers, channel of corporate sales and contact center channels, and they are good clients- Bank needs to extrapolate the same strategy across other channels

Loan status Vs Age group & Income Range



- Refused applications had higher default rate across the board
- Higher income group clients over 60 years had very low deafult rate

Recommendations

- NAME_CONTRACT_STATUS, AGE_GROUP, EDUCATION_TYPE,CHANNEL_TYPE,INCOME_RANGE,AMT_CREDIT,AMT_APPLICATION,DAYS_EMPLOYED EXT_RATING_3,YEARS_ID_PUBLISH are good variables and take a loan decision

- Target:
  1. Senior clients with age over 50
  2. Young clients(20-30) in very high income range whose applications were approved & unused in the past
  3. All clients with academic degree except married clients with unused offers
  4. Widows with incomplete higher education and unused offers
  5. Clients in a stable job for over 40 years
  6. Clients with credit amts up to 13K over application amount and unused offers in the past as they are absolute non defaulters in the past

- Avoid:
  1. Very low income group clients whose applications were rejected in the past
  2. Married Clients with academic degree who had an unused offer in the past
  3. Lower secondary degree holders in civil marriage whose previous application was refused
  4. Clients who have been employed between 2-6 years
  5. Unused offers with that have very low credit amount than application amount in previous application

- Be Cautious with:
  1. Clients whose applications were rejected in the past
  2. Clients who recently got ID changed/New - may be suspicious activity need to be cautious
  3. Clients living in low region rating
  4. AMT_Credit 3L-5L

- Car dealers, channel of corporate sales and contact center channels who are capable clients don't have unused offers, meaning they are taking utilizing the loan offers. Bank needs to extrapolate the same strategy across other channels.

Thank You