Credit EDA Case Study

By: Mehak

Objective

Identify the **key driving factors** that increase the likelihood of person defaulting. This will ensure that the loan application of person, capable to repay the loan is not rejected and the person who is incapable of replaying the loan / has more chances of defaulting is Rejected.

Approach

Following steps have been followed for analysis:

1. Data Cleaning
   - Missing values
   - Incorrect value handling
   - Checking datatype of columns
   - Checking Outliers
   - Checking duplicate rows
   - Creating new columns
2. Univariate Analysis
3. Merging the data
4. Bivariate and Multivariate Analysis

Data Cleaning

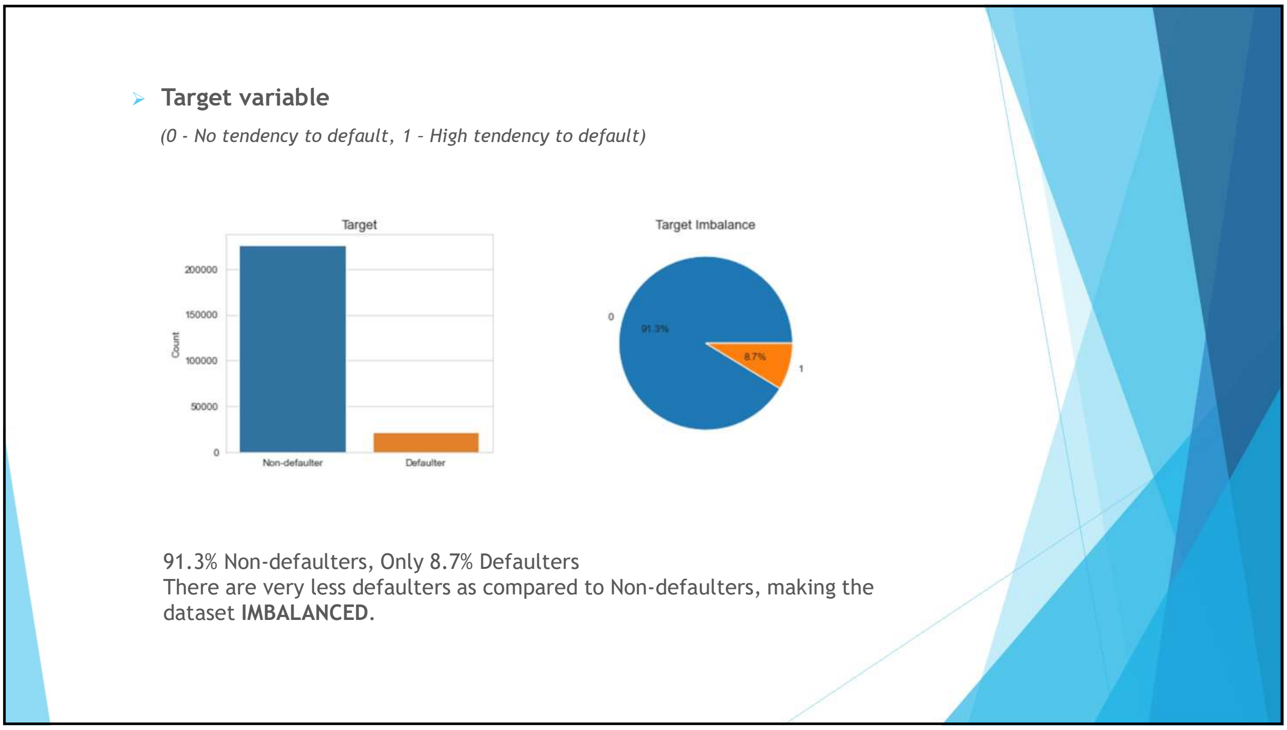
1. Missing values
   - For categorical variables, imputed with **Mode** of the column. If the count of NA values is more than frequency of Mode, created a new ‘**Missing**’ category
   - For numerical variables, if the number of rows with NA values for a column is very less, dropped the rows.
   - For records with meaningful missing values indicating missing due to a reason, imputed -999 / 999.
2. Incorrect value handling
   - Fixed columns with Invalid XNA values by dropping the records (if very less number of records) or creating a separate new category ‘Invalid’.
   - There were some columns (DAYS_FIRST_DUE, DAYS_EMPLOYED, etc.) that when converted to Years had values greater than equal to 1000 years. Since, it is not possible to be employed for 1000 years. These records were dropped.

Data Cleaning

3. Checking datatype of columns
   - The columns are changed to appropriate data type
4. Checking Outliers
   - Dropped, only the records with very extreme values, and retained other outliers.
   - For visual analysis. The **75th percentile** is considered for these columns (mostly price and amount columns)
5. Checking duplicate rows
6. Creating new columns
   - **Applications data :** AGE, AGE_GROUP NUM_EMI, INCOME_CATEGORY, TIME_APPR_PROCESS_START (hour of application in AM PM format) , Year columns (day columns converted to years)
   - **Prev. Applications data :** AMT_CREDIT_APPLY_DIFF (diff in Application and Credit Amount), DAYS_DECISION modified (all negative values converted to positive)

Applications Data Analysis

Univariate Analysis

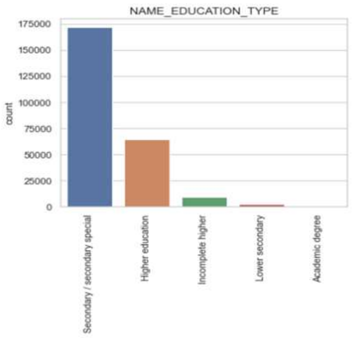


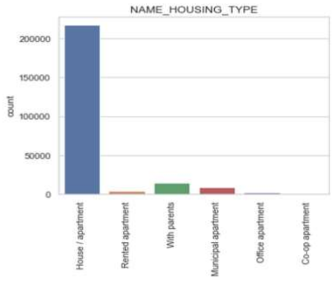
- **Target variable**

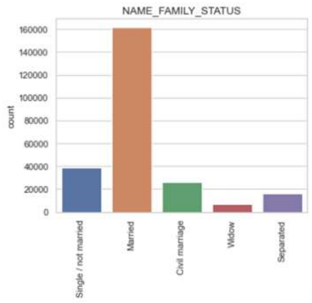
(0 - No tendency to default, 1 - High tendency to default)

91.3% Non-defaulters, Only 8.7% Defaulters
There are very less defaulters as compared to Non-defaulters, making the dataset **IMBALANCED.**

Education Level, Marital Status, Housing Type

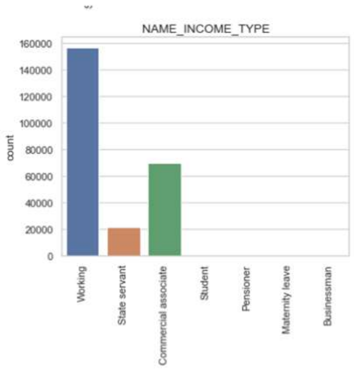


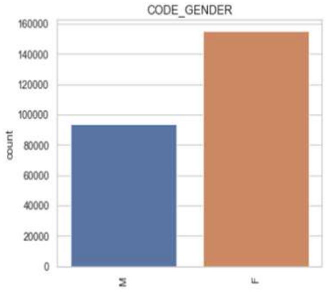


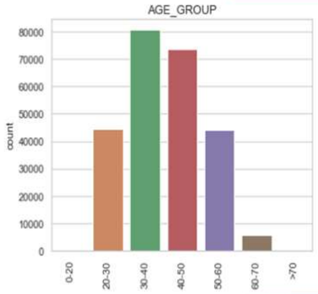


- ➢ Most of the clients have *Secondary/ secondary special* education level
- ➢ In term of Marital Status, majority of them are *Married*.
- ➢ Many Clients have their own House/ Apartment to live.

Income Type, Gender, Age Group

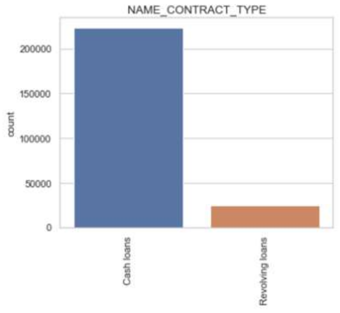


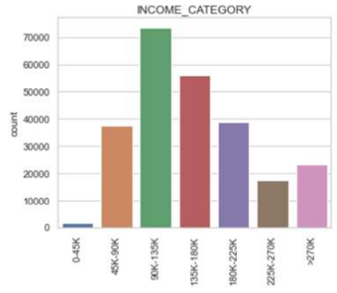


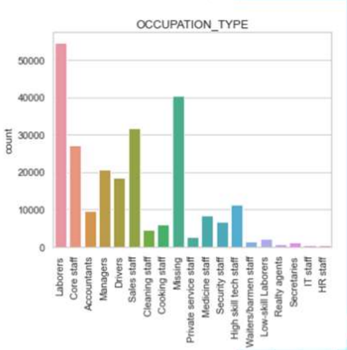


- ➢ Most of the clients are Working.
- ➢ In term of Gender, majority of them are *Females*.
- ➢ Many Clients come from the medium age group 30-40 and 40-50 years.

Contract Type, Gender, Income Category, Occupation

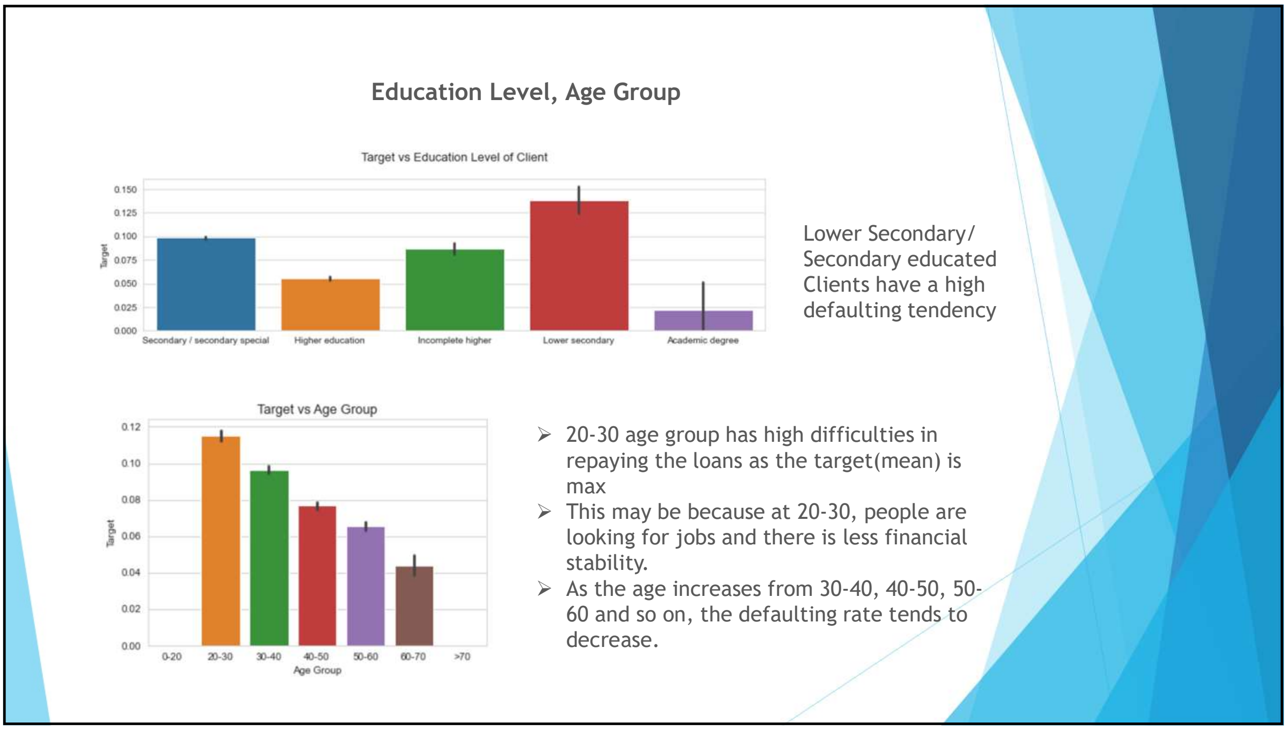






- Most of the clients prefer Cash Loans over Revolving loans
- In term of Income, majority of clients earn 90K- 135K
- Many Clients are Labourers, Sales Staff or Core Staff.

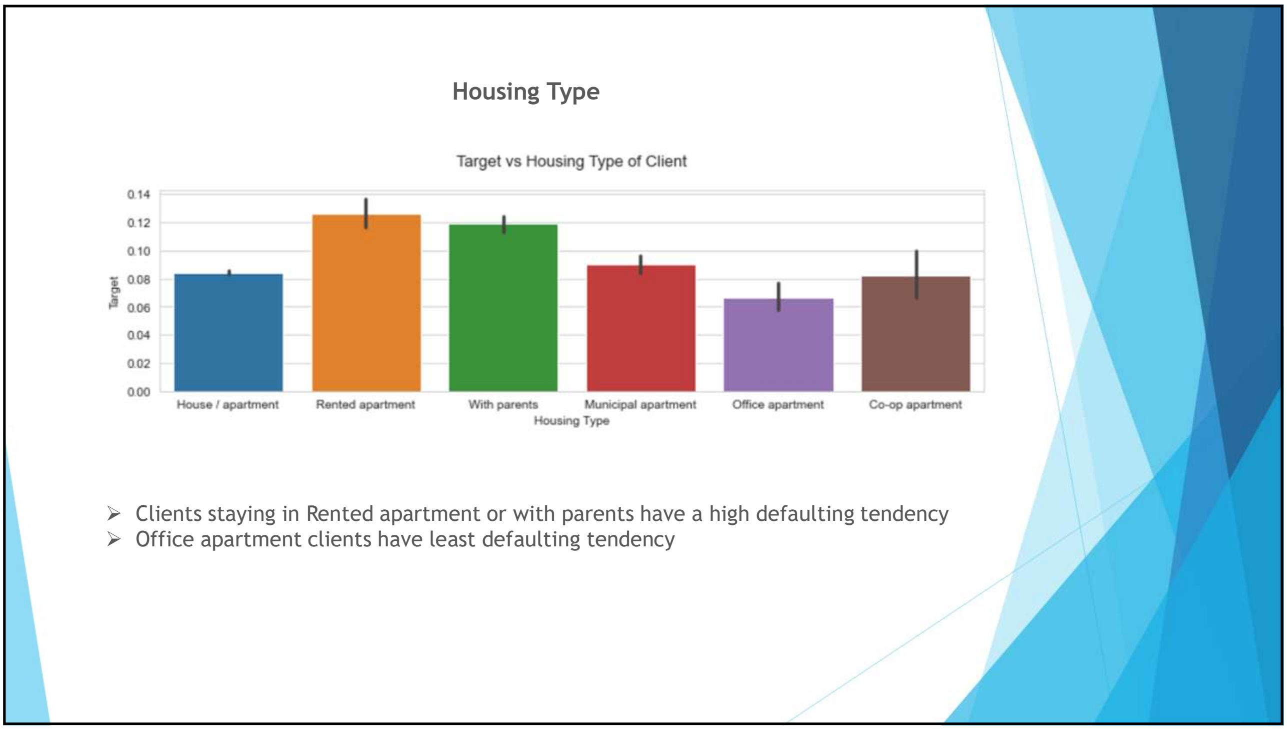
Bivariate and Multivariate Analysis



Education Level, Age Group

Lower Secondary/ Secondary educated Clients have a high defaulting tendency

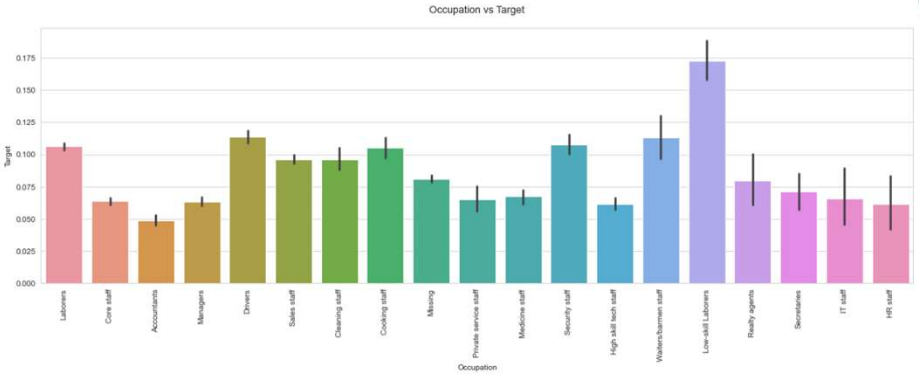
- 20-30 age group has high difficulties in repaying the loans as the target(mean) is max
- This may be because at 20-30, people are looking for jobs and there is less financial stability.
- As the age increases from 30-40, 40-50, 50-60 and so on, the defaulting rate tends to decrease.



Housing Type

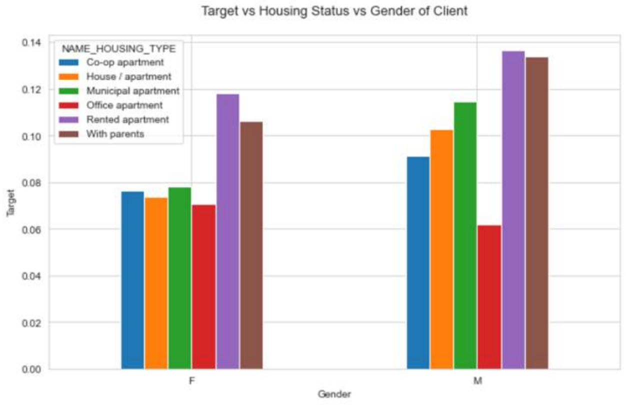
- Clients staying in Rented apartment or with parents have a high defaulting tendency
- Office apartment clients have least defaulting tendency

Occupation



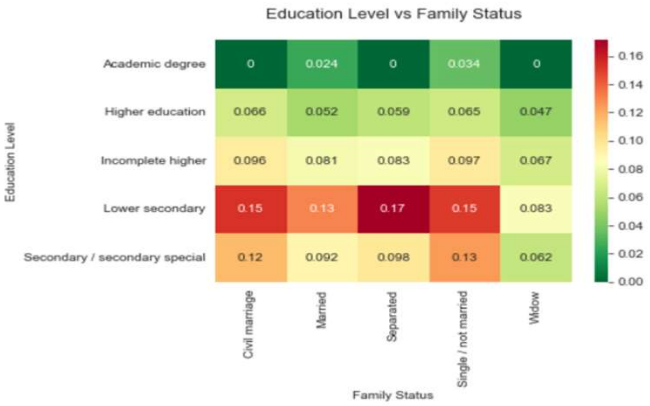
- Low skill labourers, labourers waiters/ barmen staff, Security, Drivers and Cooking staff are the people who have difficulties in paying loans.
- Accountants, HR staff, IT Staff and High skill tech staff are the occupations that have people with least defaulting tendency.

Target vs Housing Status vs Gender of Client



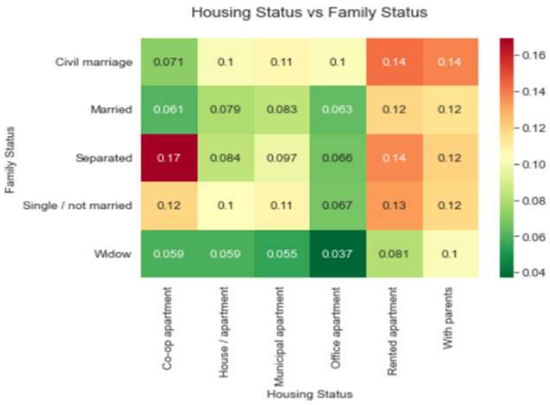
- Males have more difficulty repaying loans as compared to females.
- Females / Males living in Rented apartment, or with Parents have maximum defaulting tendency.
- Females / Males living in Office apartment have the least defaulting tendency.

Education Level, Family Status



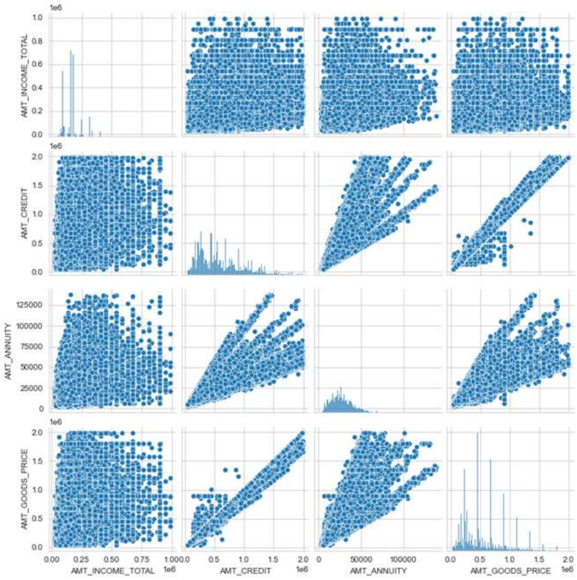
- People who have an Academic Degree or have higher education are BEST people to give loans to
- Lower secondary / Secondary folks are overall a bad segment to give loans
- Widows irrespective of education, are the BEST segment to give loans

Family Status, Housing Status



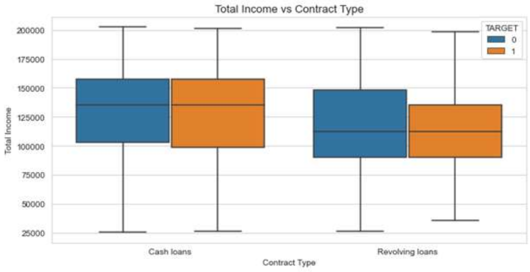
- **Separated Clients – Co-op apartment** is the *worst* segment to give loans. They have high defaulting tendency.
- Married and Widows are the BEST clients to target for giving loans
- People staying in Rented Apartment or With Parents have higher chances or defaulting than others.

Total Income vs Credit Amount vs Annuity Amount vs Goods Price

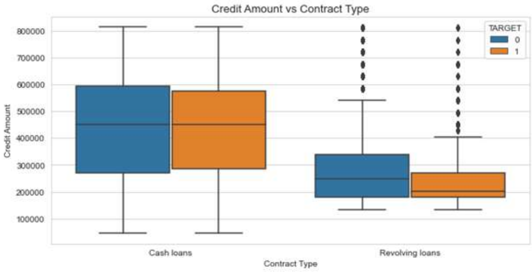


- ➢ Credit Amount, Annuity Amount and Goods Price, all 3 of them have a positive trend with each other.
- ➢ Total Income does not have much of a relation with any other amount

Total Income, Credit Amount vs Contract Type

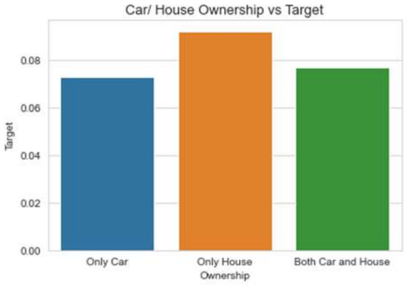


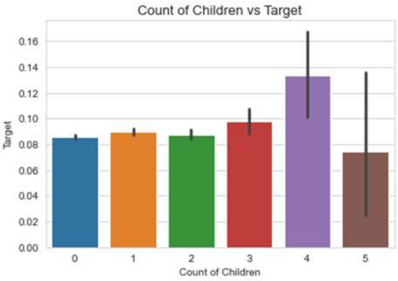
- ➢ **Cash loans:** The income distribution for people applying for both defaulter and non-defaulter is similar.
- ➢ **Revolving loans:** Defaulter have less income and less IQR w.r.t to Non defaulters.



- ➢ The Credit amount for Cash loans is much higher than Revolving loans.
- ➢ Defaulters have less credit amount relative to Non-defaulters for Revolving Loans

Car/ House Ownership, Count of Children



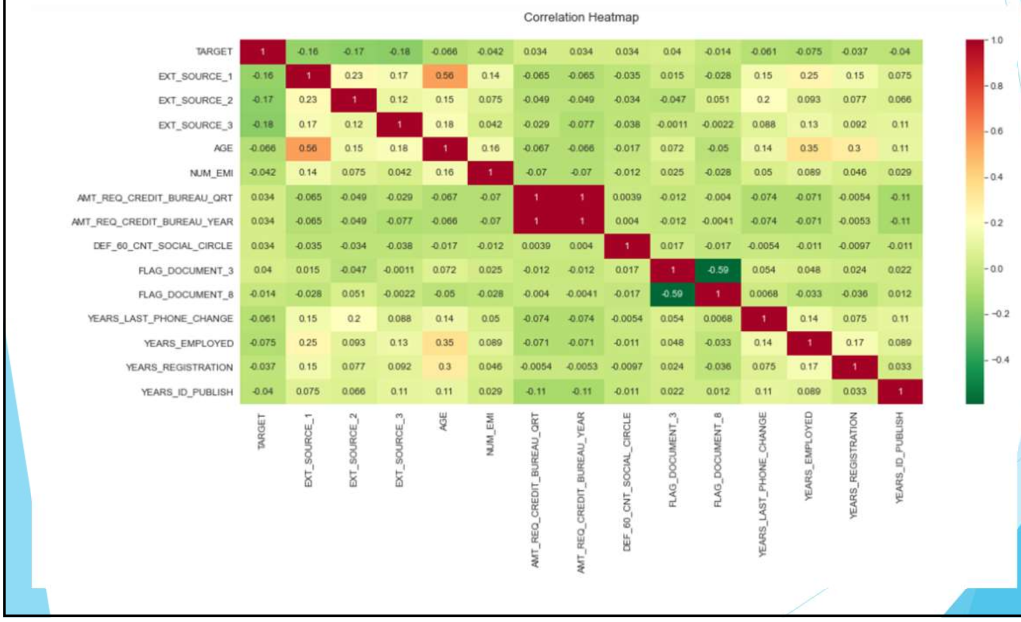


Clients owning **Only a house**, have high chances of defaulting.

- People with more number of Children have high defaulting tendency.
- As the count of Children increases till 4, the defaulting chances also increase.

Correlations

Correlation Heatmap



Correlation Heatmap

	TARGET	EXT_SOURCE_1	EXT_SOURCE_2	EXT_SOURCE_3	AGE	NUM_EMI	AMT_REQ_CREDIT_BUREAU_QRT	AMT_REQ_CREDIT_BUREAU_YEAR	DEF_60_CNT_SOCIAL_CIRCLE	FLAG_DOCUMENT_3	FLAG_DOCUMENT_8	YEARS_LAST_PHONE_CHANGE	YEARS_EMPLOYED	YEARS_REGISTRATION	YEARS_ID_PUBLISH
TARGET	1	-0.16	-0.17	-0.18	-0.066	-0.042	0.034	0.034	0.034	0.04	-0.014	-0.061	-0.075	-0.037	-0.04
EXT_SOURCE_1	-0.16	1	0.23	0.17	0.56	0.14	-0.065	-0.065	-0.035	0.015	-0.028	0.15	0.25	0.15	0.075
EXT_SOURCE_2	-0.17	0.23	1	0.12	0.15	0.075	-0.049	-0.049	-0.034	-0.047	0.051	0.2	0.093	0.077	0.066
EXT_SOURCE_3	-0.18	0.17	0.12	1	0.18	0.042	-0.029	-0.077	-0.038	-0.0011	-0.0022	0.088	0.13	0.092	0.11
AGE	-0.066	0.56	0.15	0.18	1	0.16	-0.067	-0.066	-0.017	0.072	-0.05	0.14	0.35	0.3	0.11
NUM_EMI	-0.042	0.14	0.075	0.042	0.16	1	-0.07	-0.07	-0.012	0.025	-0.028	0.05	0.089	0.046	0.029
AMT_REQ_CREDIT_BUREAU_QRT	0.034	-0.065	-0.049	-0.029	-0.067	-0.07	1	1	0.0039	-0.012	-0.004	-0.074	-0.071	-0.0054	-0.11
AMT_REQ_CREDIT_BUREAU_YEAR	0.034	-0.065	-0.049	-0.077	-0.066	-0.07	1	1	0.004	-0.012	-0.0041	-0.074	-0.071	-0.0053	-0.11
DEF_60_CNT_SOCIAL_CIRCLE	0.034	-0.035	-0.034	-0.038	-0.017	-0.012	0.0039	0.004	1	0.017	-0.017	-0.0054	-0.011	-0.0097	-0.011
FLAG_DOCUMENT_3	0.04	0.015	-0.047	-0.0011	0.072	0.025	-0.012	-0.012	0.017	1	-0.59	0.054	0.048	0.024	0.022
FLAG_DOCUMENT_8	-0.014	-0.028	0.051	-0.0022	-0.05	-0.028	-0.004	-0.0041	-0.017	-0.59	1	0.0068	-0.033	-0.036	0.012
YEARS_LAST_PHONE_CHANGE	-0.061	0.15	0.2	0.088	0.14	0.05	-0.074	-0.074	-0.0054	0.054	0.0068	1	0.14	0.075	0.11
YEARS_EMPLOYED	-0.075	0.25	0.093	0.13	0.35	0.089	-0.071	-0.071	-0.011	0.048	-0.033	0.14	1	0.17	0.089
YEARS_REGISTRATION	-0.037	0.15	0.077	0.092	0.3	0.046	-0.0054	-0.0053	-0.0097	0.024	-0.036	0.075	0.17	1	0.033
YEARS_ID_PUBLISH	-0.04	0.075	0.066	0.11	0.11	0.029	-0.11	-0.11	-0.011	0.022	0.012	0.11	0.089	0.033	1

Correlations w.r.t TARGET

- TARGET has negative correlation with all 3 external sources.
- It has a negative correlation with AGE indicating that as the age increases, the target (defaulting tendency) decreases
- It has a negative correlation with YEARS_LAST_PHONE_CHANGE, YEARS_EMPLOYED meaning that if the phone or employment or ID Proof is changed recently, the defaulting tendency increases.
- It has negative correlation with Number of EMI.
- It has positive correlation with AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_YEAR indicating as the number of Enquiries about a client increase, the client is more likely to default.
- It has a positive correlation with DEF_60_CNT_SOCIAL_CIRCLE indicating that as the number of observations of Clients social surroundings that defaulted in past 30 days increases, the client is more likely to default.

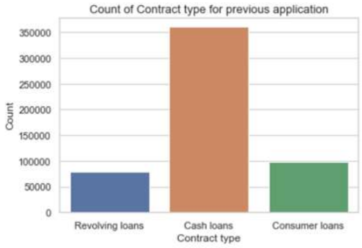
Other Correlations

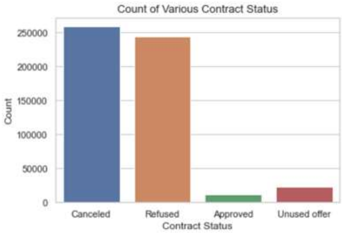
- AGE and EXT_SOURCE_1 have a very high positive correlation
- AGE and YEARS_EMPLOYED have a high positive correlation with YEARS_EMPLOYED and YEARS_REGISTRATION. As the age increases, the Years Employed and Years registered also increase
- EXT_SOURCE_1,EXT_SOURCE_2,EXT_SOURCE_3, all 3 of them have a very high negative correlation with Target
- FLAG_DOCUMENT_3 and FLAG_DOCUMENT_8 have a very high negative correlation with each other.
- YEARS_ID_PUBLISH is negatively correlated to AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_YEAR. It means that if the ID is changed recently, there are likely to be more enquiries. For clients, who changed their IDs long back, the number of enquiries will be less.

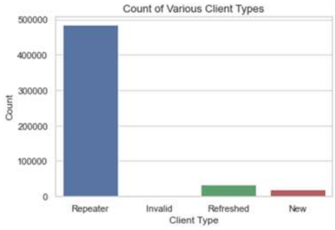
Previous Applications Data Analysis

Univariate Analysis

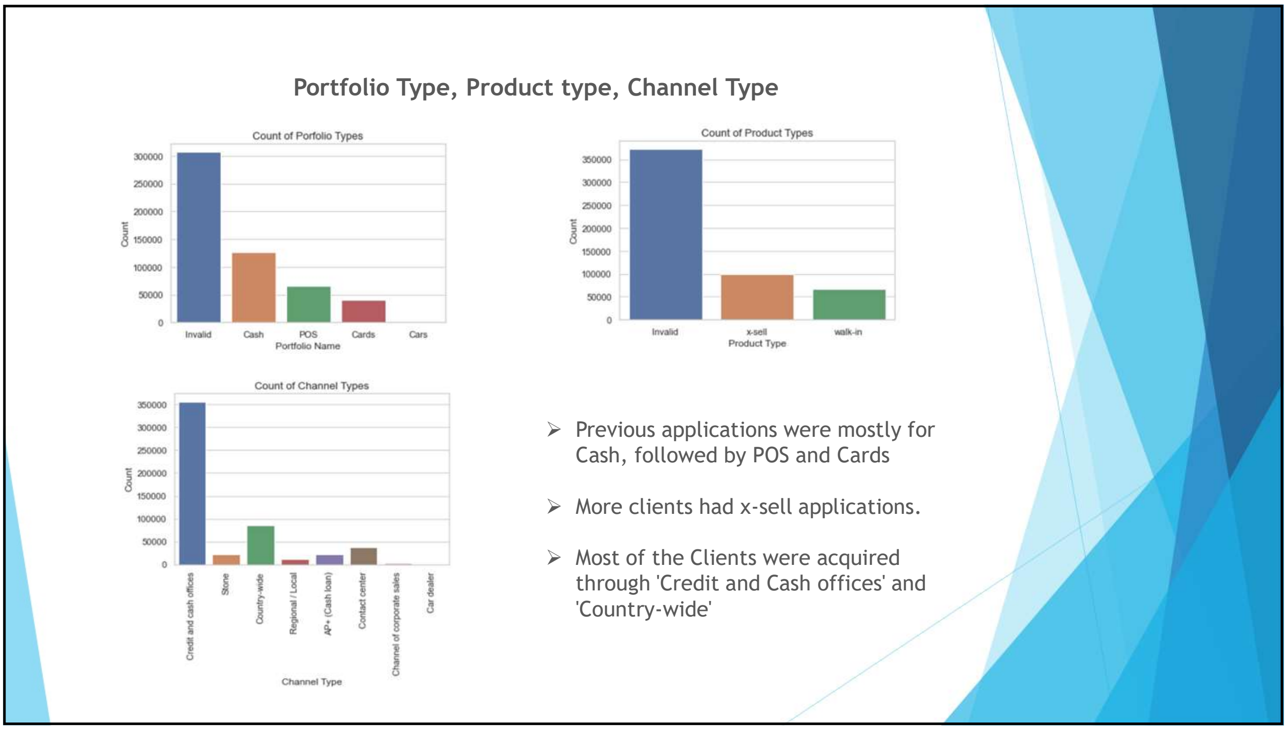
Contract Type, Contract Status, Client Type





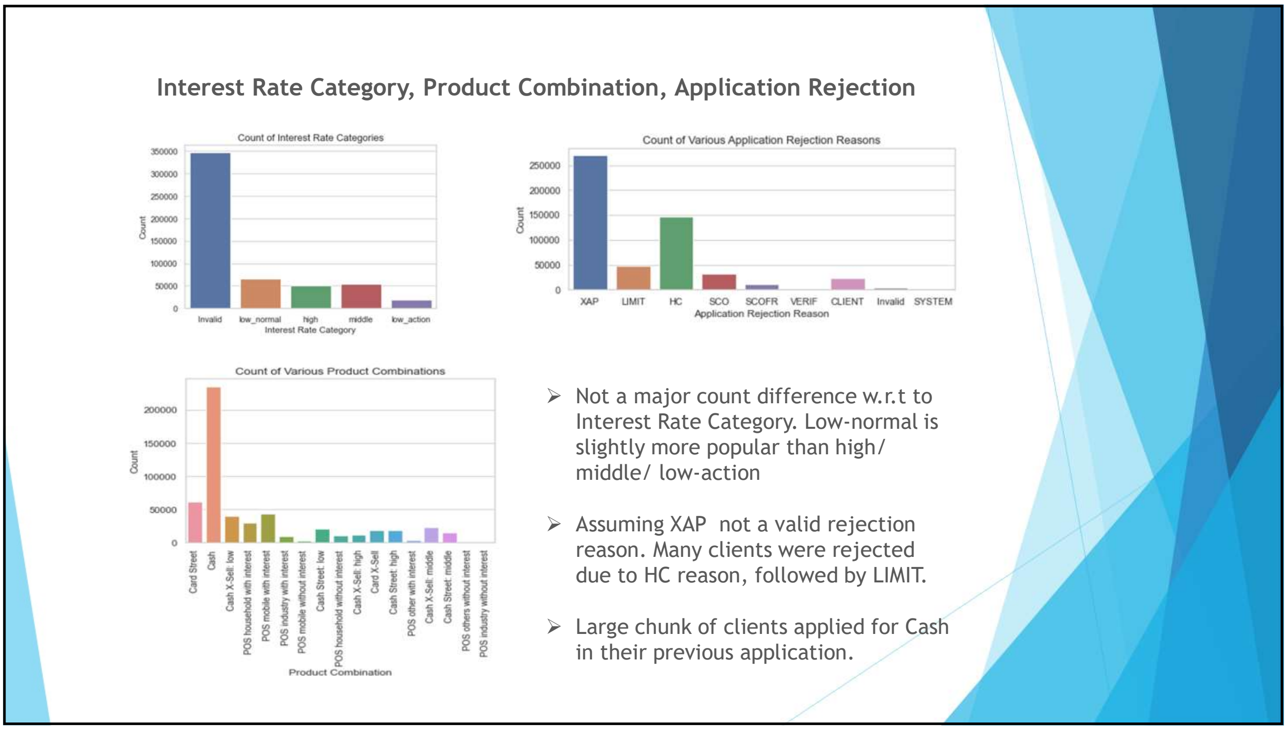


- Most of the clients had applied for Cash loans
- Majority of the applications were Cancelled or Refused. Very few were Approved
- Most of the Clients for previous applications were Repeaters.



Portfolio Type, Product type, Channel Type

- Previous applications were mostly for Cash, followed by POS and Cards
- More clients had x-sell applications.
- Most of the Clients were acquired through 'Credit and Cash offices' and 'Country-wide'



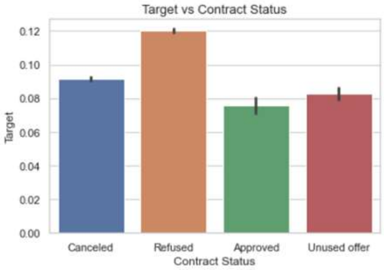
Interest Rate Category, Product Combination, Application Rejection

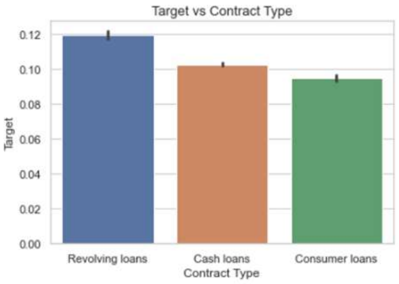
- Not a major count difference w.r.t to Interest Rate Category. Low-normal is slightly more popular than high/ middle/ low-action
- Assuming XAP not a valid rejection reason. Many clients were rejected due to HC reason, followed by LIMIT.
- Large chunk of clients applied for Cash in their previous application.

Bivariate and Multivariate Analysis

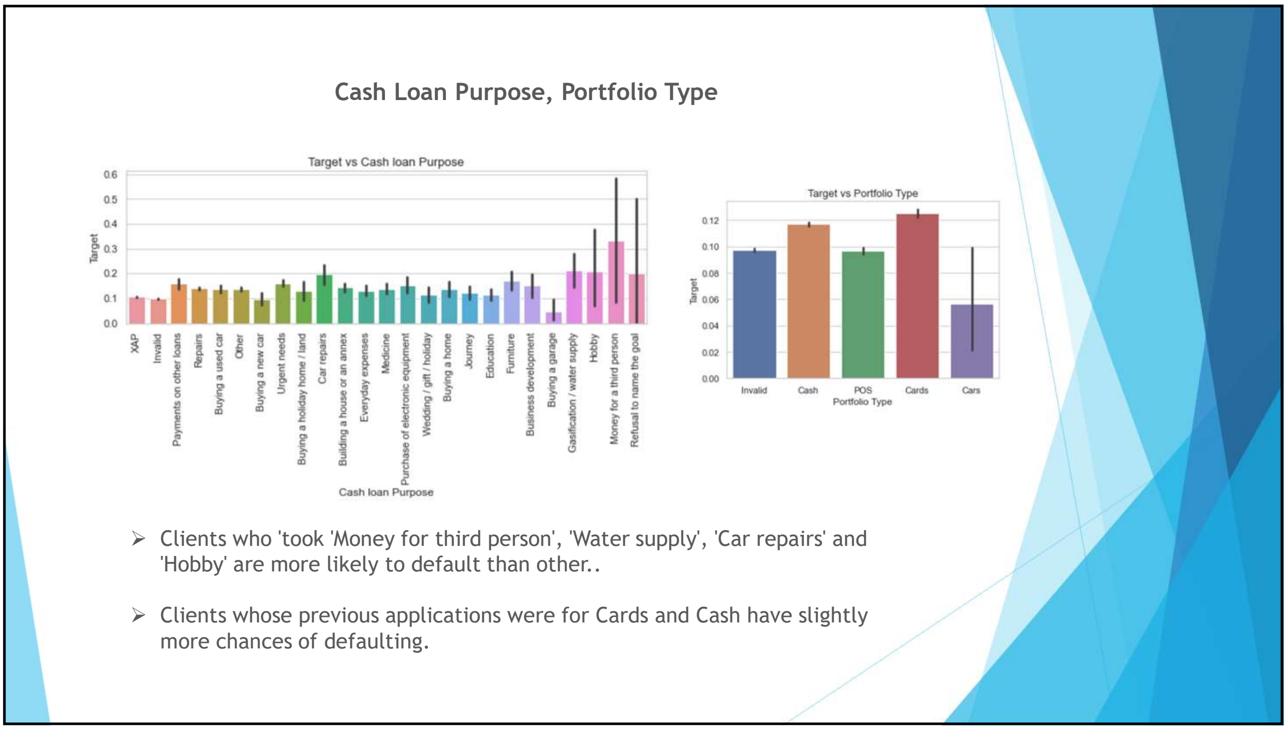
Note: This analysis is performed after merging Previous Applications data with Applications data.

Contract Status, Contract Type



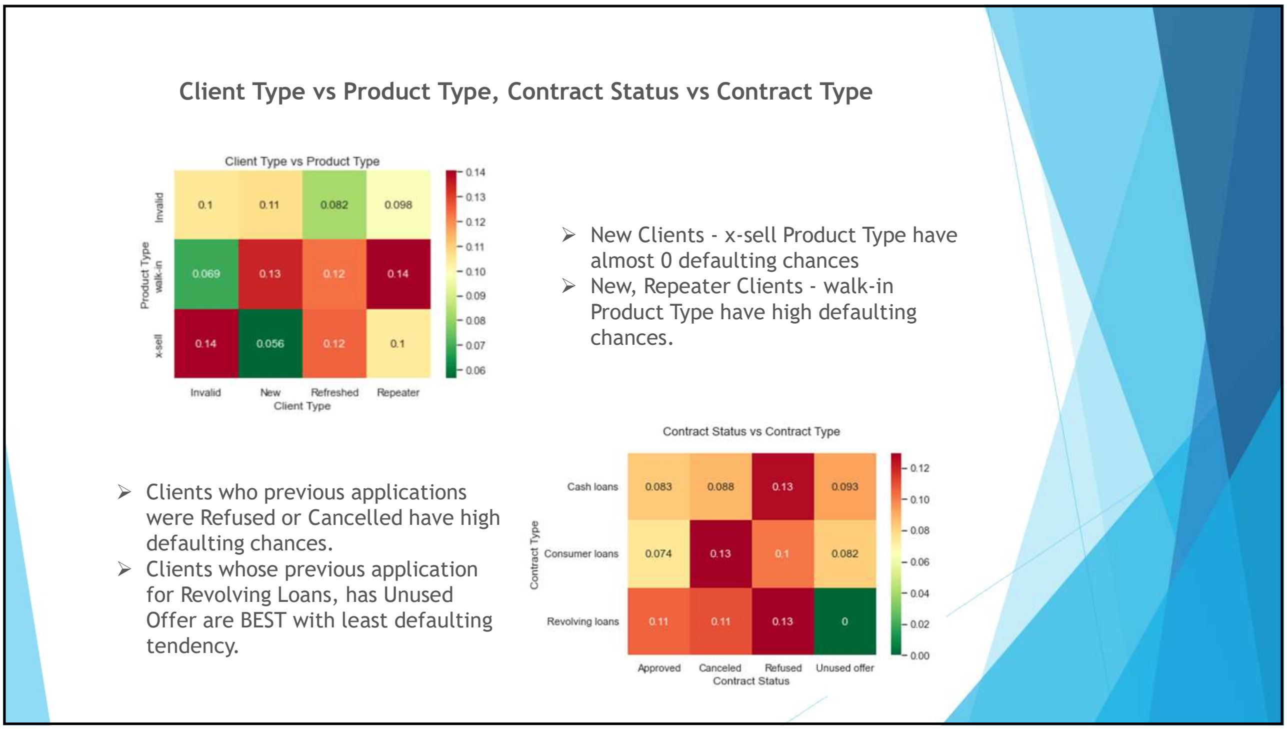


- As expected, the people whose previous application was Refused have highest defaulting chances.
- Clients who had Revolving loans have high defaulting chances, followed by Cash Loans and Consumer Loans



Cash Loan Purpose, Portfolio Type

- Clients who 'took 'Money for third person', 'Water supply', 'Car repairs' and 'Hobby' are more likely to default than other..
- Clients whose previous applications were for Cards and Cash have slightly more chances of defaulting.

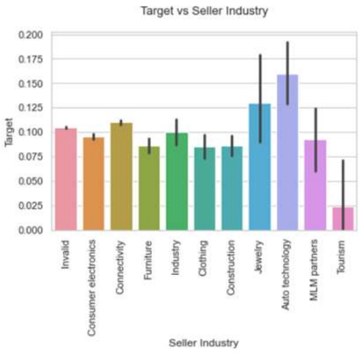


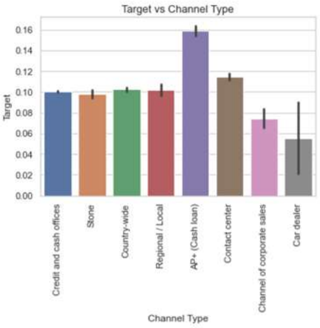
Client Type vs Product Type, Contract Status vs Contract Type

- New Clients - x-sell Product Type have almost 0 defaulting chances
- New, Repeater Clients - walk-in Product Type have high defaulting chances.

- Clients who previous applications were Refused or Cancelled have high defaulting chances.
- Clients whose previous application for Revolving Loans, has Unused Offer are BEST with least defaulting tendency.

Seller Industry, Channel Type

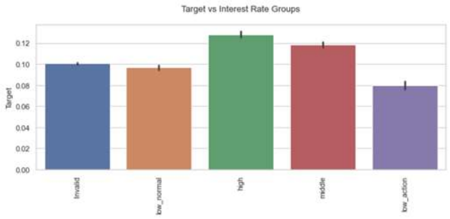


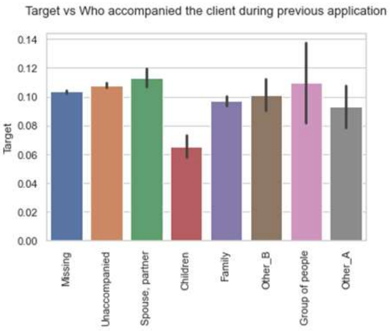


- If the seller belonged to Auto technology, Jewelry or Connectivity Industry, there was a high chance of defaulting. Tourism Industry had least chance of defaulting

- Clients who were acquired through AP+ (Cash loan) and Contact center have high defaulting tendency. Clients acquired through Channel of corporate sales and Car dealer have low defaulting tendency.

Interest Rate Group, Who accompanied the client during application

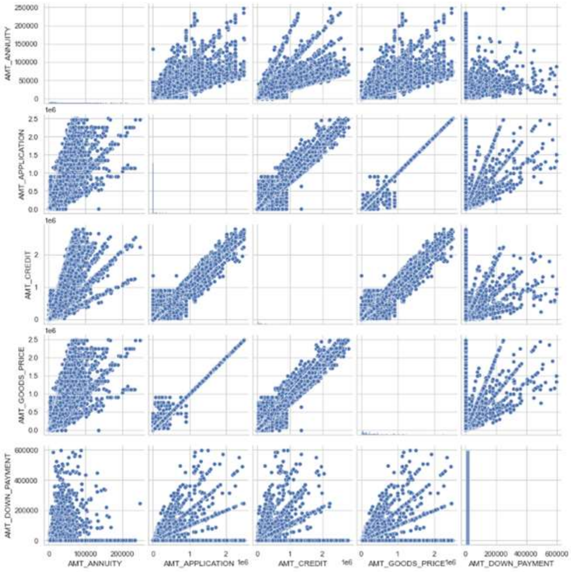




- High and middle interest Rates had higher defaulting chances.

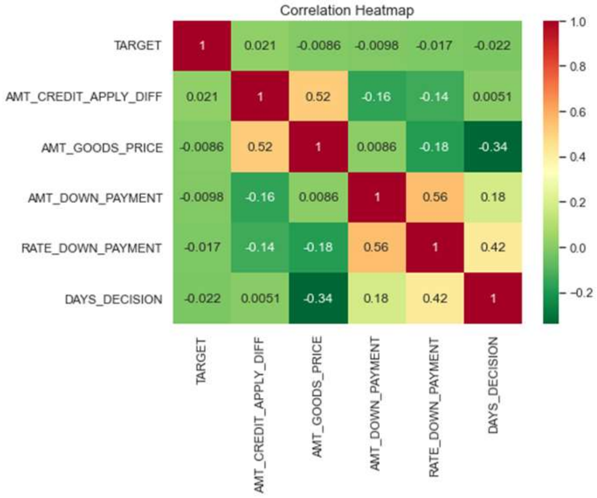
- People who were accompanied by Children had lowest defaulting tendency.

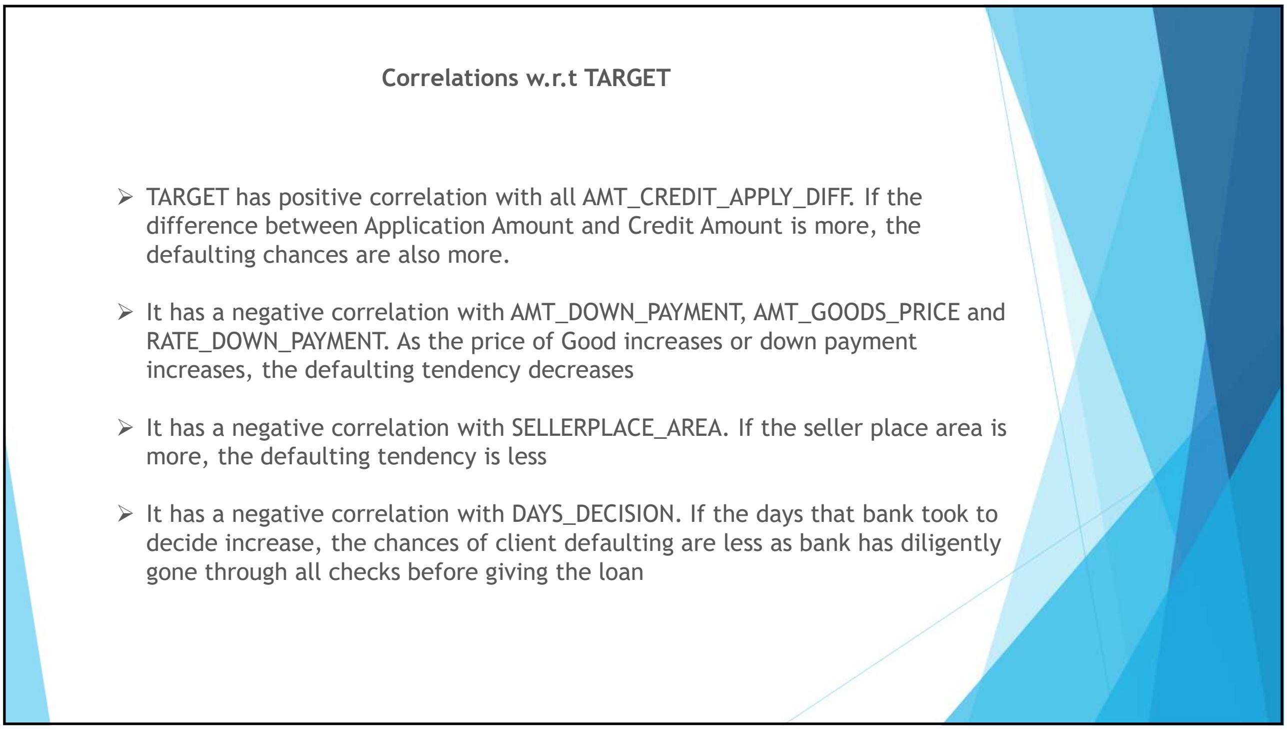
Annuity Amount vs Application Amount vs Credit Amount vs Goods Price vs Down Payment



- ➢ All of these amounts have a positive trend. If any increases, others will also increase.

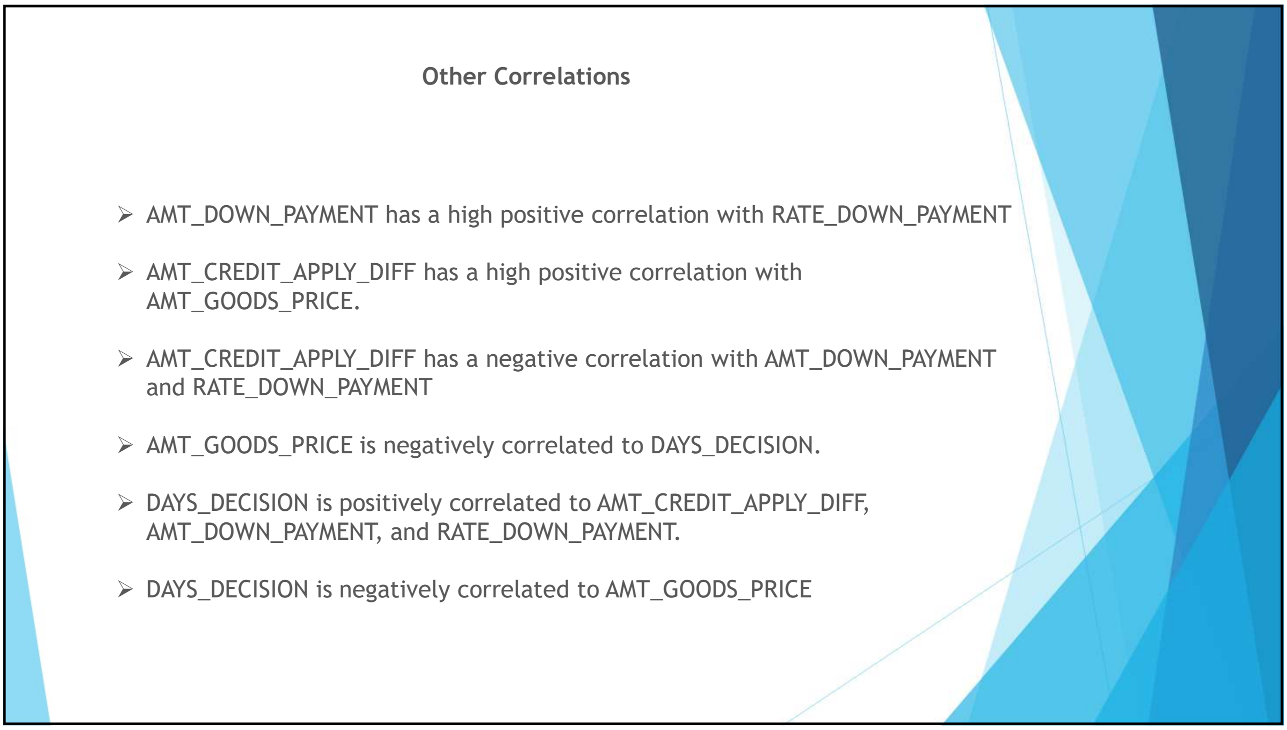
Correlation Heatmap





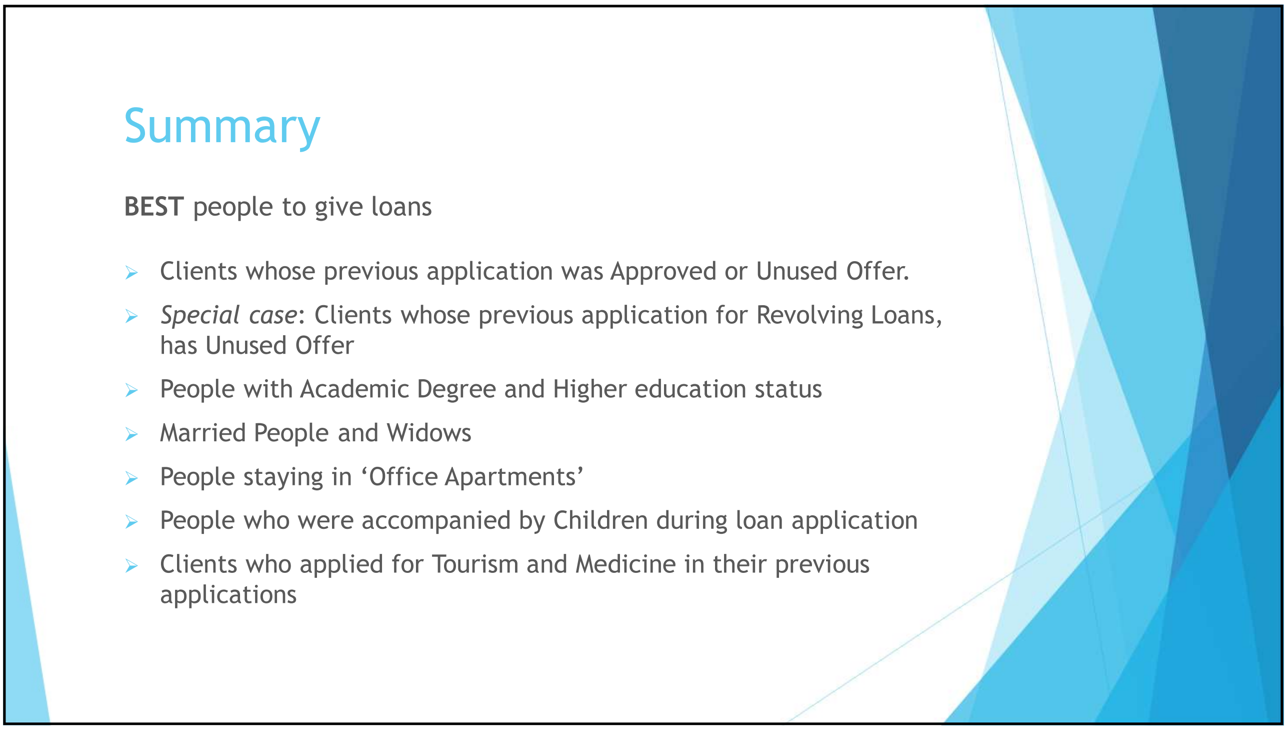
Correlations w.r.t TARGET

- TARGET has positive correlation with all AMT_CREDIT_APPLY_DIFF. If the difference between Application Amount and Credit Amount is more, the defaulting chances are also more.
- It has a negative correlation with AMT_DOWN_PAYMENT, AMT_GOODS_PRICE and RATE_DOWN_PAYMENT. As the price of Good increases or down payment increases, the defaulting tendency decreases
- It has a negative correlation with SELLERPLACE_AREA. If the seller place area is more, the defaulting tendency is less
- It has a negative correlation with DAYS_DECISION. If the days that bank took to decide increase, the chances of client defaulting are less as bank has diligently gone through all checks before giving the loan



Other Correlations

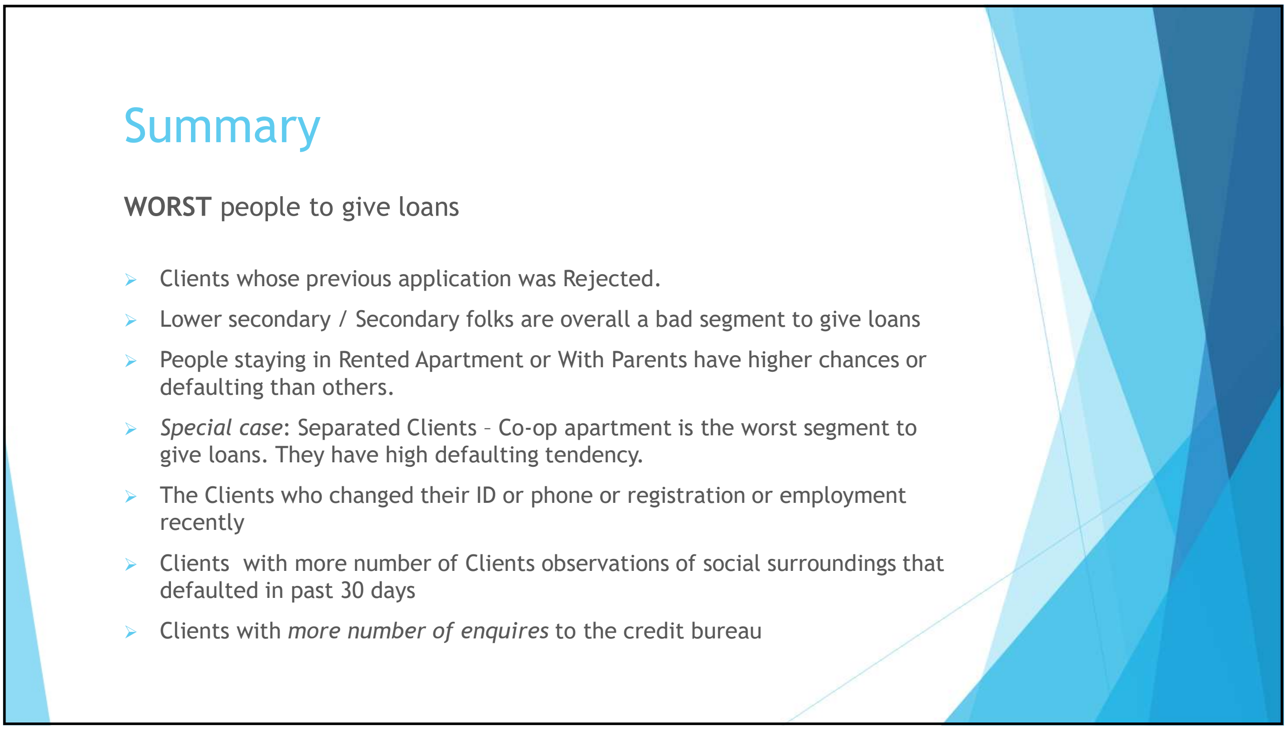
- AMT_DOWN_PAYMENT has a high positive correlation with RATE_DOWN_PAYMENT
- AMT_CREDIT_APPLY_DIFF has a high positive correlation with AMT_GOODS_PRICE.
- AMT_CREDIT_APPLY_DIFF has a negative correlation with AMT_DOWN_PAYMENT and RATE_DOWN_PAYMENT
- AMT_GOODS_PRICE is negatively correlated to DAYS_DECISION.
- DAYS_DECISION is positively correlated to AMT_CREDIT_APPLY_DIFF, AMT_DOWN_PAYMENT, and RATE_DOWN_PAYMENT.
- DAYS_DECISION is negatively correlated to AMT_GOODS_PRICE



Summary

BEST people to give loans

- Clients whose previous application was Approved or Unused Offer.
- *Special case*: Clients whose previous application for Revolving Loans, has Unused Offer
- People with Academic Degree and Higher education status
- Married People and Widows
- People staying in 'Office Apartments'
- People who were accompanied by Children during loan application
- Clients who applied for Tourism and Medicine in their previous applications



Summary

WORST people to give loans

- Clients whose previous application was Rejected.
- Lower secondary / Secondary folks are overall a bad segment to give loans
- People staying in Rented Apartment or With Parents have higher chances or defaulting than others.
- *Special case*: Separated Clients – Co-op apartment is the worst segment to give loans. They have high defaulting tendency.
- The Clients who changed their ID or phone or registration or employment recently
- Clients with more number of Clients observations of social surroundings that defaulted in past 30 days
- Clients with *more number of enquires* to the credit bureau

Thank You.