

Upgrad (IIIT Bangalore) MS Data Science

Presented by Smriti Pradhan for the
submission of Credit EDA Assignment





Analyzing Loan Applications

Problem Statement:

A Consumer Finance Company is a business that specializes in giving loans to clients. They want to comprehend the trends among clients who have trouble making installment payments. These characteristics can assist in recognizing these loan applications, which can then result in loan denial, loan reduction, lending (at a higher interest rate), and, ultimately, an improvement in the borrower's portfolio and risk assessment.



Types of Decision Taken:

Approved: Company's approval of the loan application

Cancelled: Cancelled application by the Client during approval of the loan.

Refused: Company's rejection of the loan application.

Unused offer: Cancelled loan by the client but at different stages of the process.

Business Objective:

The company wants to understand the driving factors or driver variables behind loan default ie. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

CATEGORICAL COLUMNS:

- 'NAME_CONTRACT_TYPE',
- 'FLAG_OWN_CAR',
- 'FLAG_OWN_REALTY',
- 'CODE_GENDER',
- 'NAME_EDUCATION_TYPE',
- 'AMT_CATEGORY',
- 'AGE_GROUP',
- 'NAME_FAMILY_STATUS',
- 'NAME_HOUSING_TYPE',
- 'NAME_TYPE_SUITE',
- 'NAME_INCOME_TYPE',
- 'OCCUPATION_TYPE',
- 'ORGANIZATION_TYPE',
- 'REGION_RATING_CLIENT_W_CITY',
- 'REGION_RATING_CLIENT',
- 'AMT_REQ_CREDIT_BUREAU_HOUR',
- 'DEF_60_CNT_SOCIAL_CIRCLE',
- 'AMT_REQ_CREDIT_BUREAU_WEEK',
- 'AMT_REQ_CREDIT_BUREAU_DAY',
- 'DEF_30_CNT_SOCIAL_CIRCLE',
- 'AMT_REQ_CREDIT_BUREAU_QRT',
- 'CNT_CHILDREN',
- 'CNT_FAM_MEMBERS',
- 'AMT_REQ_CREDIT_BUREAU_MON',
- 'AMT_REQ_CREDIT_BUREAU_YEAR',
- 'OBS_30_CNT_SOCIAL_CIRCLE',
- 'OBS_60_CNT_SOCIAL_CIRCLE',

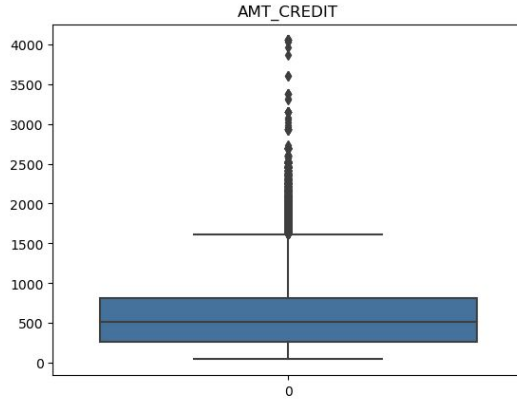
NUMERICAL COLUMNS:

- 'AMT_GOODS_PRICE',
- 'DAYS_LAST_PHONE_CHANGE',
- 'DAYS_ID_PUBLISH',
- 'AMT_INCOME_TOTAL',
- 'DAYS_EMPLOYED',
- 'DAYS_REGISTRATION',
- 'DAYS_BIRTH',
- 'AMT_CREDIT',
- 'AMT_ANNUITY'

There are two types of data presented for analysis. They are as follows:

1. CATEGORICAL : Object Type and Less than < 40 null values
2. NUMERICAL : Float or Int Type

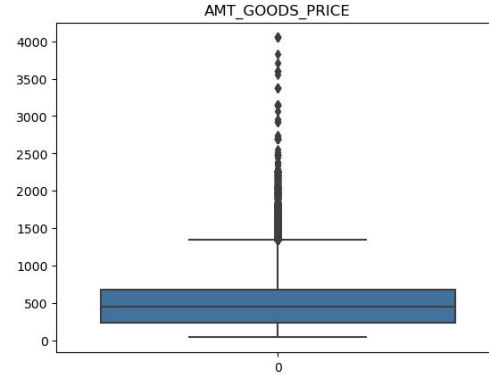
Outlier Analysis



FG 1

From the boxplot, we can see there are outliers in AMT_CREDIT.

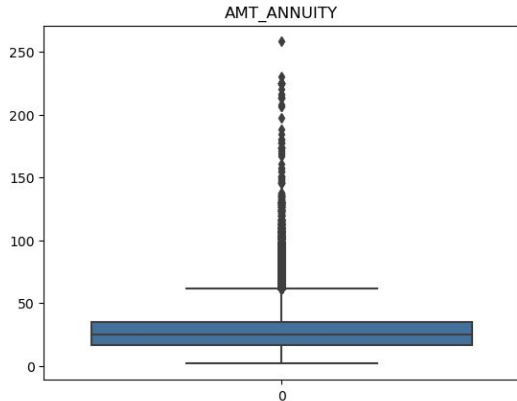
The percentage of outliers stands at 2.13%



FG 2

From the boxplot, we can see there are outliers in AMT_GOODS_PRICE.

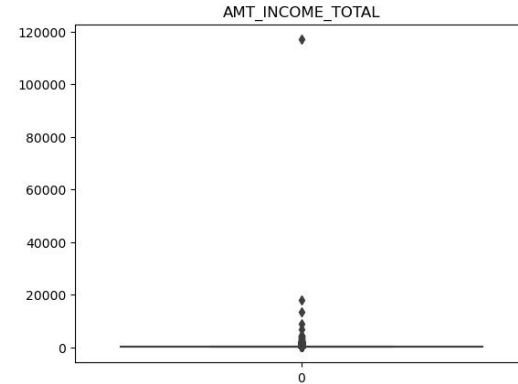
The percentage of outliers stands at 4.79%



FG 3

From the boxplot, we can see there are outliers in AMT_ANNUITY.

The percentage of outliers stands at 2.44%



FG 4

From the boxplot, we can see there are outliers in AMT_INCOME_TOTAL.

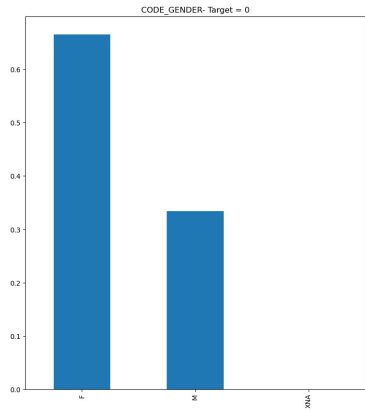
The percentage of outliers stands at 4.56%

Univariate Analysis: Categorical

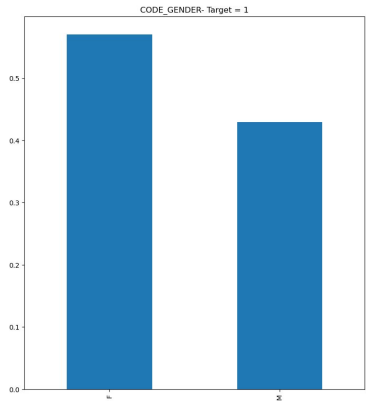
Key interpretation from univariate analysis of categorical variables highlighting variables having significant difference for target = 0 and target = 1

- **CODE_GENDER:** Defaulters (Target = 1) has a higher percentage of male customers in comparison to non-defaulters (Target = 0)
- **NAME_EDUCATION_TYPE:** Defaulters (Target = 1) has a higher percentage of customers with Secondary/Secondat Special education
- **AGE_GROUP :** Defaulters (Target = 1) has a higher percentage of customers in the age group of 30s
- **NAME_INCOME_TYPE:** Defaulters (Target = 1) has a higher percentage of working customers whereas percentage of defaulting pensioners is lesser in comparison to non-defaulters (Target = 0)
- **OCCUPATION_TYPE:** Laborers contribute a higher percentage in defaulters (Target = 1) in comparison to non-defaulters(Target = 0)
- **REGION_RATING_CLIENT_W_CITY:** Customers with rating 3 constitutes a higher percentage of defaulters in comparison to non-defaulters
- **REGION_RATING_CLIENT:** Customers with rating 3 constitutes a higher percentage of defaulters in comparison to non-defaulters

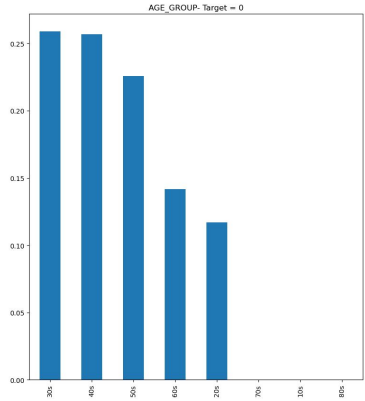
FG 1



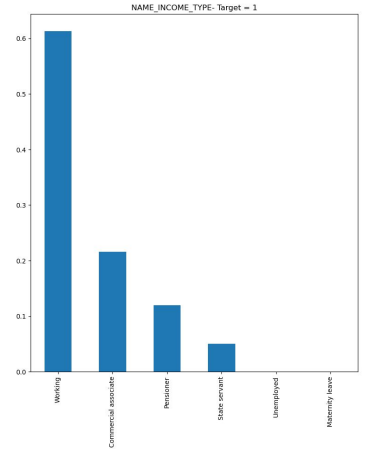
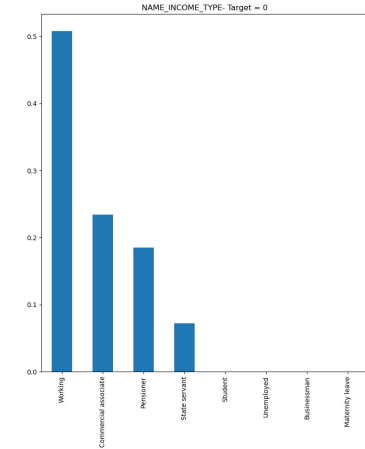
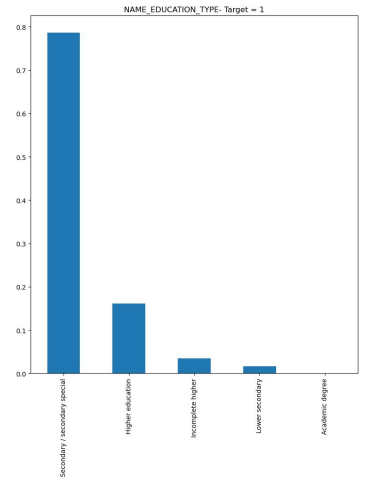
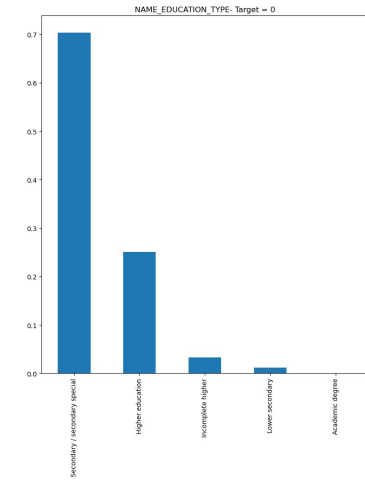
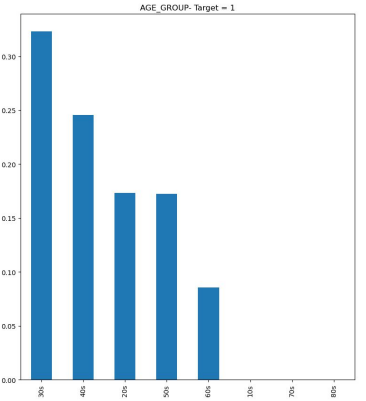
FG 2



FG 3

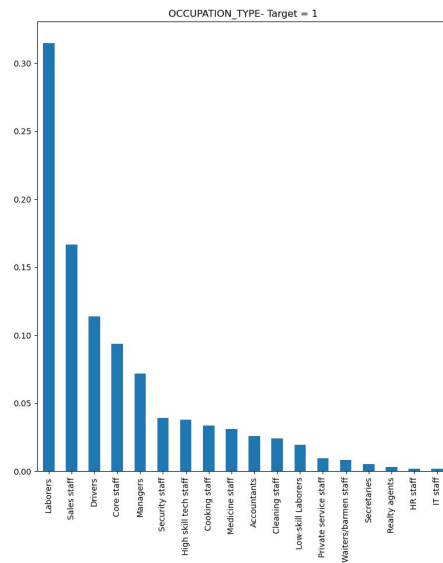
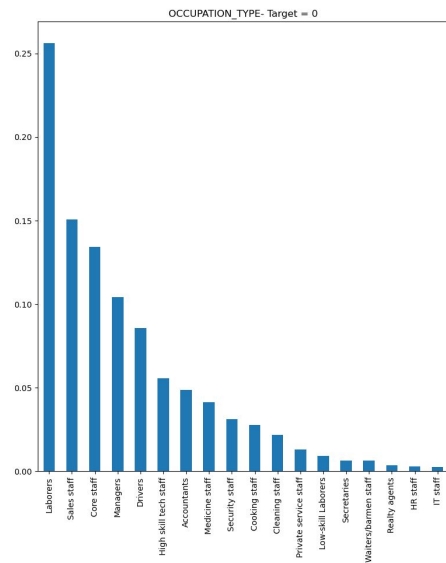


FG 4

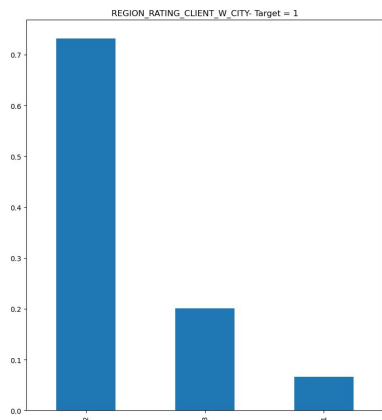
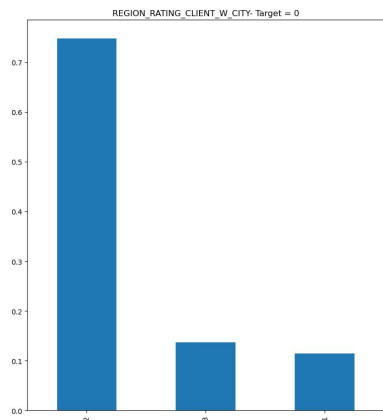




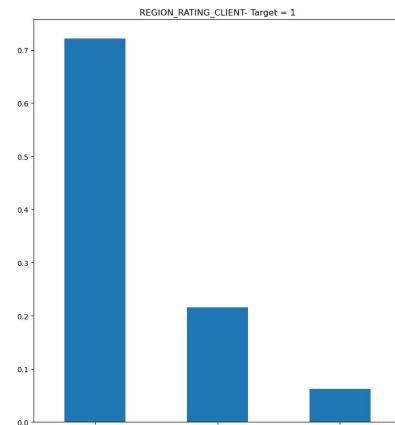
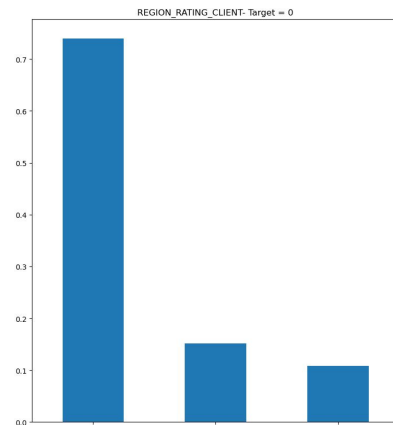
FG 5



FG 6



FG 7

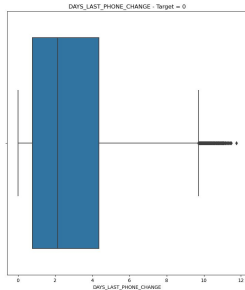


Univariate Analysis: Numerical

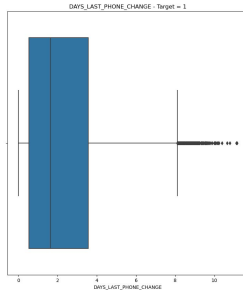
Key interpretation from univariate analysis of numerical variables highlighting variables having significant difference for target = 0 and target = 1

- DAYS_LAST_PHONE_CHANGE: Median value and 75 percentile value for Defaulters (Target = 1) is lesser than non-defaulters. It implies defaulter more often change phone number before application
- DAYS_ID_PUBLISH: Defaulters seem to change IDs more frequently than non-defaulters
- DAYS_BIRTH: 25 percentile, median, and 75 percentile for the age of defaulter applicants are smaller than younger applicants. This means defaulter population is younger than non-defaulter.

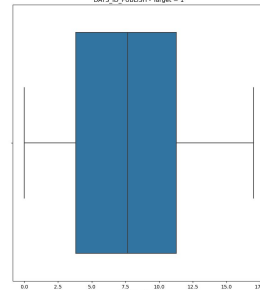
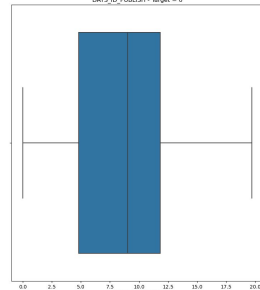
FG 1



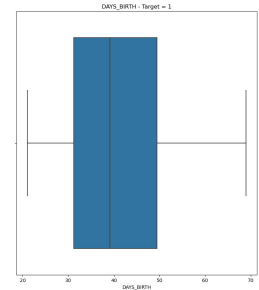
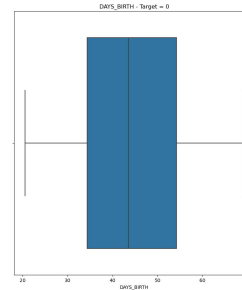
FG 2



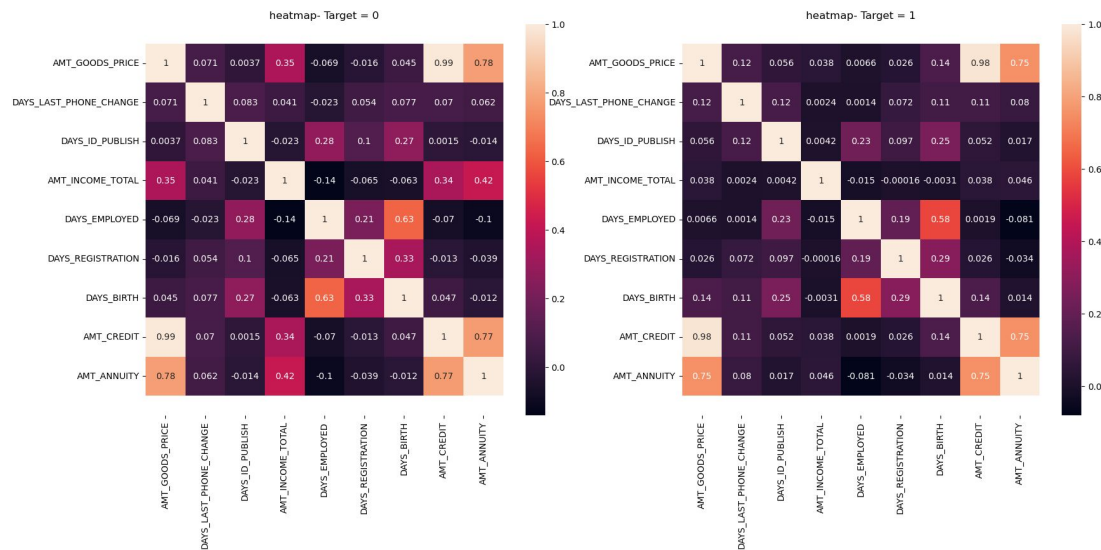
FG 2



FG 3



Correlation for Numerical Columns



Through the heatmap we can see same set of columns seem to have a high correlation across all three data sets. Top correlate columns are:

- AMT_GOOD_PRICE vs AMT_CREDIT
- AMT_GOOD_PRICE vs AMT_ANNUITY
- AMT_CREDIT_AMT_ANNUITY

Top 10 high correlation variables common across Target = 0 and Target = 1 (in tabular format)

| | VAR1 | VAR2 | Correlation | Correlation_abs |
|-----|-----------------------------|--------------------------|-------------|-----------------|
| 414 | OBS_60_CNT_SOCIAL_CIRCLE | OBS_30_CNT_SOCIAL_CIRCLE | 1.00 | 1.00 |
| 154 | AMT_GOODS_PRICE | AMT_CREDIT | 0.99 | 0.99 |
| 337 | REGION_RATING_CLIENT_W_CITY | REGION_RATING_CLIENT | 0.95 | 0.95 |
| 277 | CNT_FAM_MEMBERS | CNT_CHILDREN | 0.88 | 0.88 |
| 440 | DEF_60_CNT_SOCIAL_CIRCLE | DEF_30_CNT_SOCIAL_CIRCLE | 0.86 | 0.86 |
| 155 | AMT_GOODS_PRICE | AMT_ANNUITY | 0.78 | 0.78 |
| 129 | AMT_ANNUITY | AMT_CREDIT | 0.77 | 0.77 |
| 207 | DAYS_EMPLOYED | DAYS_BIRTH | 0.63 | 0.63 |
| 128 | AMT_ANNUITY | AMT_INCOME_TOTAL | 0.42 | 0.42 |
| 153 | AMT_GOODS_PRICE | AMT_INCOME_TOTAL | 0.35 | 0.35 |

Target 0

| | VAR1 | VAR2 | Correlation | Correlation_abs |
|-----|-----------------------------|--------------------------|-------------|-----------------|
| 414 | OBS_60_CNT_SOCIAL_CIRCLE | OBS_30_CNT_SOCIAL_CIRCLE | 1.00 | 1.00 |
| 154 | AMT_GOODS_PRICE | AMT_CREDIT | 0.98 | 0.98 |
| 337 | REGION_RATING_CLIENT_W_CITY | REGION_RATING_CLIENT | 0.96 | 0.96 |
| 277 | CNT_FAM_MEMBERS | CNT_CHILDREN | 0.89 | 0.89 |
| 440 | DEF_60_CNT_SOCIAL_CIRCLE | DEF_30_CNT_SOCIAL_CIRCLE | 0.87 | 0.87 |
| 129 | AMT_ANNUITY | AMT_CREDIT | 0.75 | 0.75 |
| 155 | AMT_GOODS_PRICE | AMT_ANNUITY | 0.75 | 0.75 |
| 207 | DAYS_EMPLOYED | DAYS_BIRTH | 0.58 | 0.58 |
| 415 | OBS_60_CNT_SOCIAL_CIRCLE | DEF_30_CNT_SOCIAL_CIRCLE | 0.34 | 0.34 |
| 389 | DEF_30_CNT_SOCIAL_CIRCLE | OBS_30_CNT_SOCIAL_CIRCLE | 0.33 | 0.33 |

Target 1

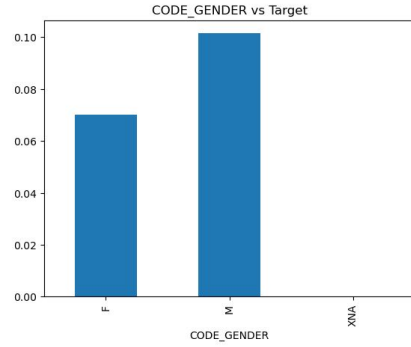
8 out of top 10 pair of high correlated variables are same for both TARGET 0 and TARGET 1

Bivariate Analysis: Categorical

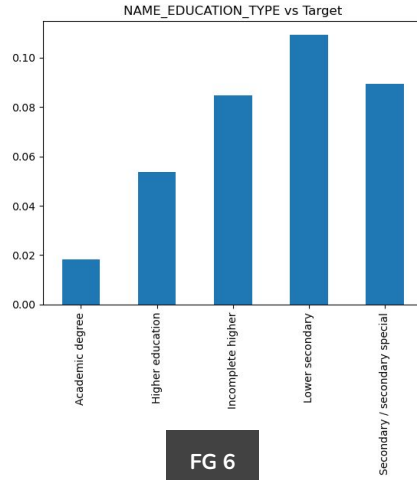
Key interpretation from bivariate analysis of categorical variables

- CODE_GENDER: Male customers have a higher probability of defaulting
- NAME_EDUCATION_TYPE: Customers with lower secondary education have a higher risk of default
- AGE_GROUP: Customers in 20s and 30s have higher chances of defaulting
- NAME_HOUSING_TYPE: Customers living in rented apartments and living with parents seem to default more
- NAME_INCOME_TYPE: Unemployed and Customers on maternity leave have higher
- OCCUPATION_TYPE: Low-skill laborers default more
- REGION_RATING_CLIENT and REGION_RATING_CLIENT_W_CITY: Customers with rating 3 have higher risk of defaulting

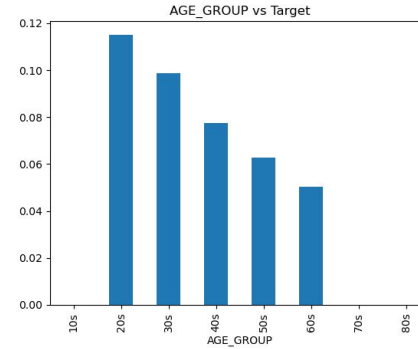
FG 1



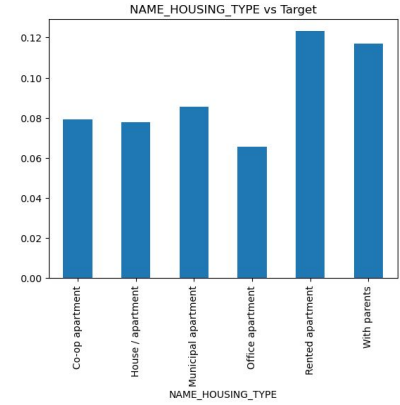
FG 2



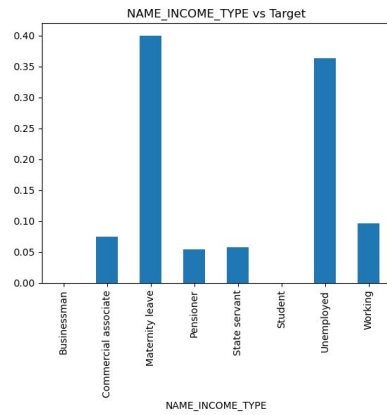
FG 3



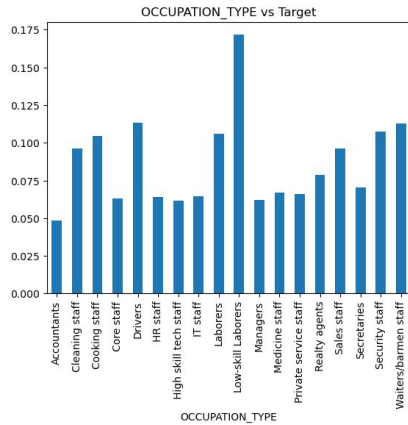
FG 4



FG 5



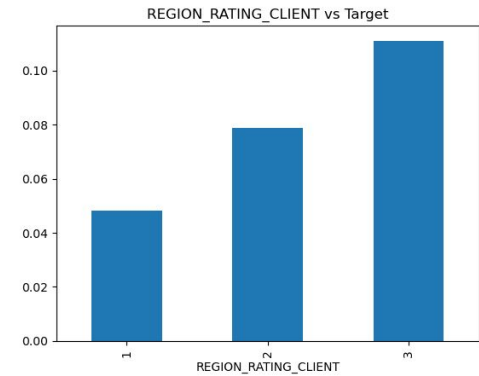
FG 6



FG 7



FG 8

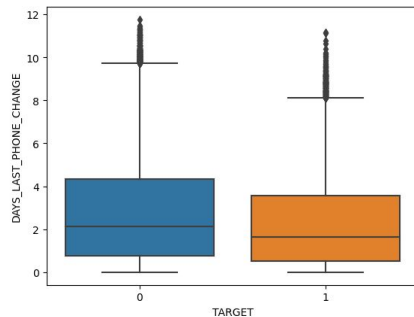


Bivariate Analysis: Numerical

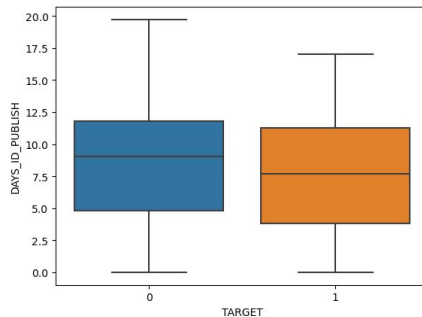
Key interpretation from bivariate analysis of categorical variables

- DAYS_LAST_PHONE_CHANGE: Defaulter customers change phone closer to the submission of application
- DAYSID_PUBLISH: Defaulter customers changes id closer to submission of application
- DAYS_REGISTRATION: Defaulter customers changes registration on a date closer to submission of application
- DAYS_BIRTH: Defaulter customers are relatively younger than non-defaulters

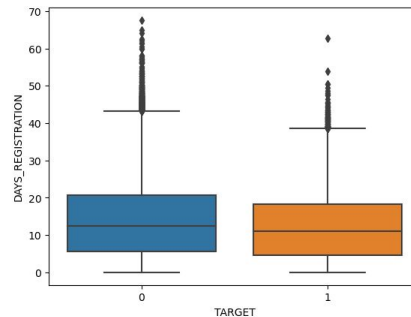
FG 1



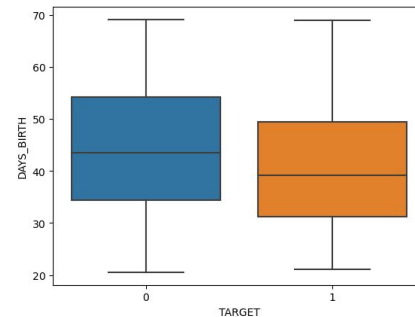
FG 2



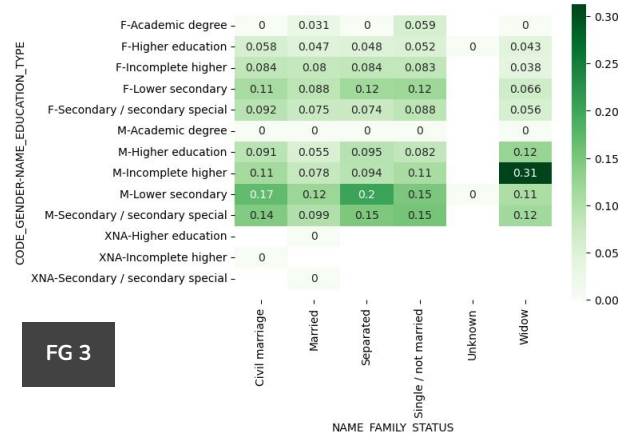
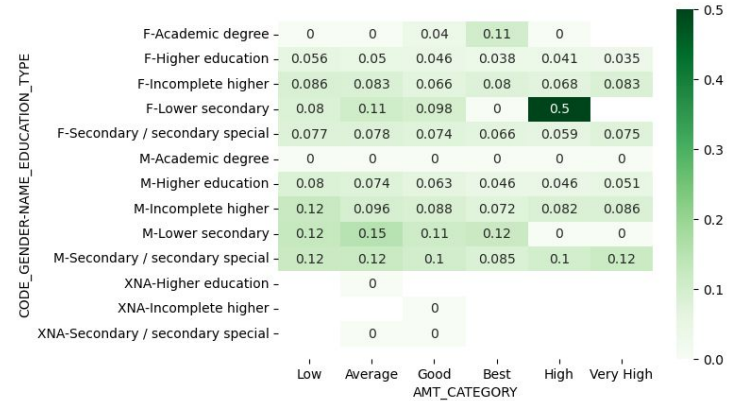
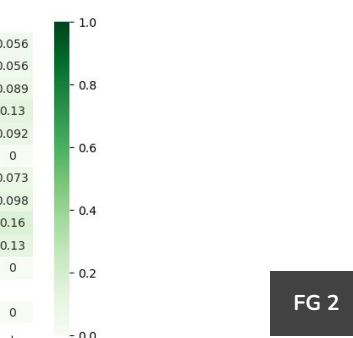
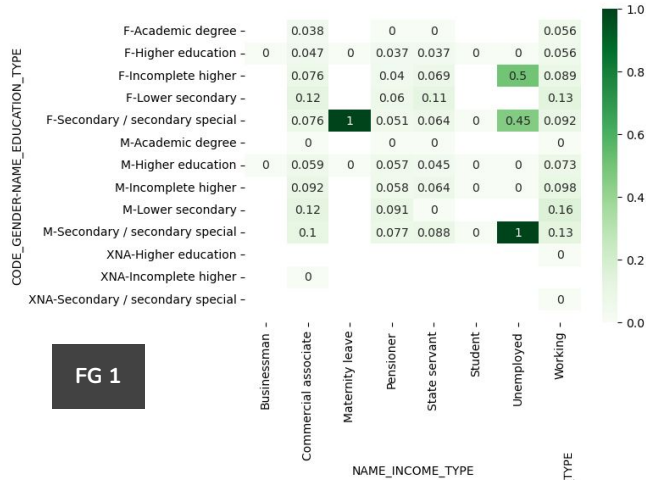
FG 3



FG 4



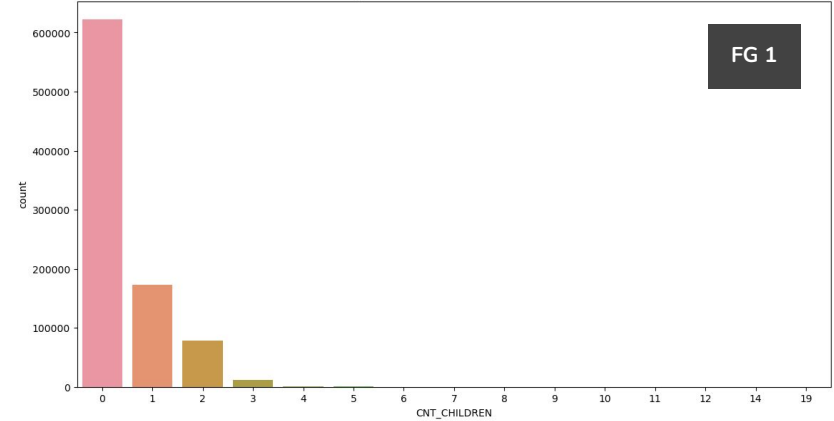
Multivariate Analysis: Categorical



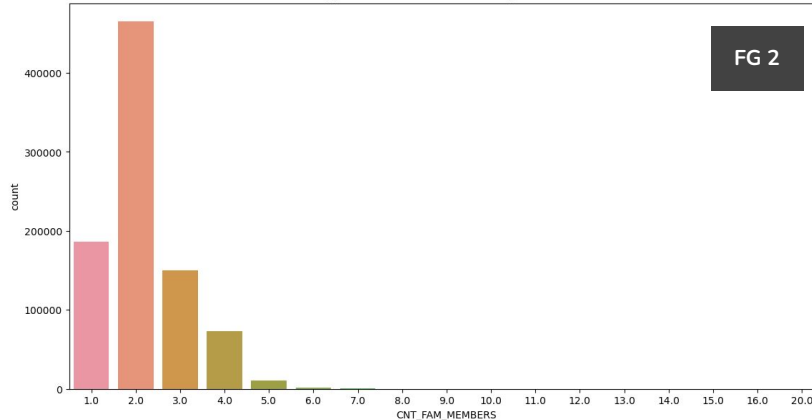
Inference from Bivariate analysis for numerical data:

1. Approved status v/s No. of children: People with 0 children are more likely to get loan approved
2. Approved status v/s No. of family members: If the number of people in a family is 2 they are more likely to get loan approved.
3. Approved status v/s Age: People with age in between 30 -50 years are more likely to get loan approved compared to the people in 20s and 60s

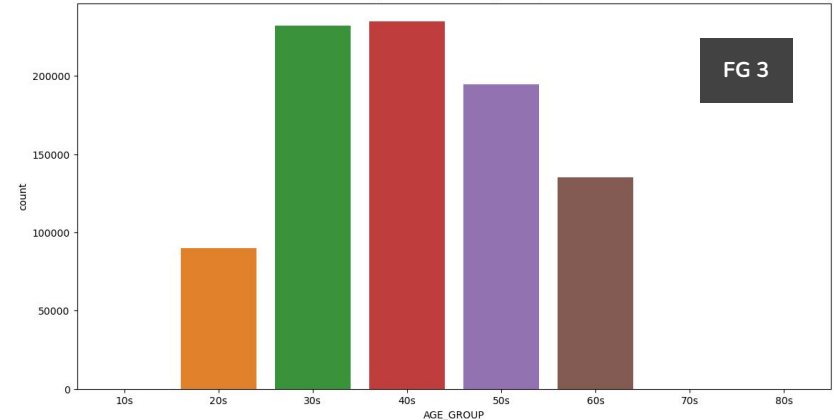
Approved status v/s No. of children



Approved status v/s No. of Family Members



Approved status v/s Age Group

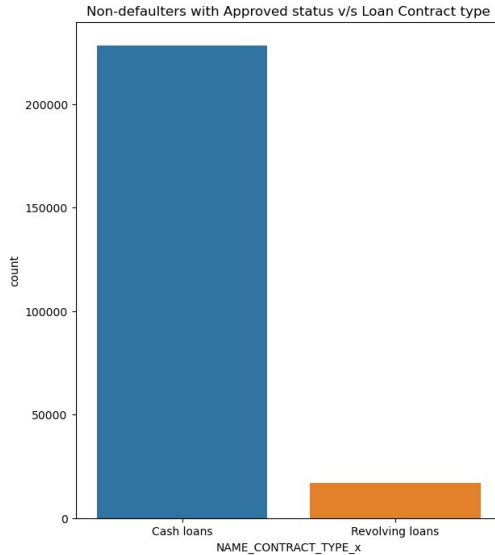


Inference from Bivariate analysis for categorical data:

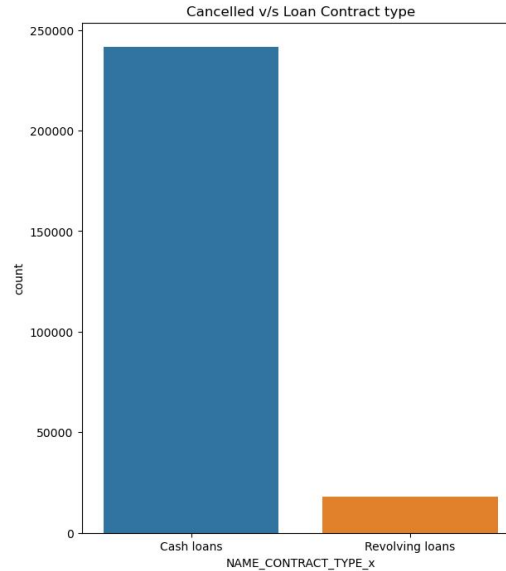


1. Cash loans: Are more likely to be re-applied
2. Cash loans: Are more likely to be cancelled
3. Cash loans: Are more likely to be re-applied

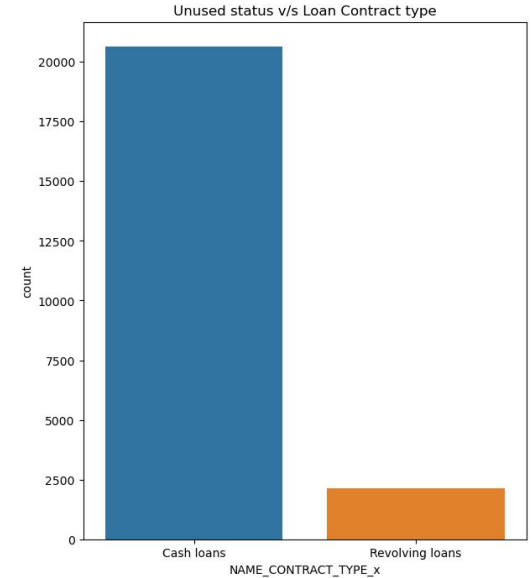
FG 1




FG 2



FG 3





As per analysis made, below is the summary table of the variables that can help in predicting which customer can account to default.

| S.No. | Variable | Variable Type |
|-------|-----------------------------|---------------|
| 1 | CODE_GENDER | Categorical |
| 2. | NAME_EDUCATION_TYPE | Categorical |
| 3. | AGE_GROUP | Categorical |
| 4. | NAME_HOUSING_TYPE | Categorical |
| 5. | NAME_INCOME_TYPE | Categorical |
| 6. | OCCUPATION_TYPE | Categorical |
| 7. | REGION_RATING_CLIENT | Categorical |
| 8. | REGION_RATING_CLIENT_W_CITY | Categorical |
| 9. | DAYS_LAST_PHONE_CHANGE | Numerical |
| 10. | DAYS_ID_PUBLISH | Numerical |

End