EDA Case Study

Identify Driving Factor For Loan Default

By: Deepak Kumar, Rishik Patel

Email :- kdeepak@Hotmail.com , rishik.ptl@gmail.com

Problem Statement

Analyze the loan data to identify factors leading to loan defaults. The outcome of this study will highlight the patterns to distinguish between applicant which are likely to pay the loan & applicants which are likely to default.

Loan Providing company can use this analysis to

1. Approve the loan
2. Reject the loan
3. Approve the loan at high rate
4. Reduce the loan amount

Available Data Set

To Complete this study, we are given 2 data sets.

1. Application Data
    1. This data set has information about the applicant & loan at the time of application and whether this client has payment difficulty or not
    2. There are 307511 records in this data set along with 122 attributes
2. Previous loan Data
    1. This data set has information about applicant's previous loan applications & status for the same (e.g. Approved, Refused ,Cancelled, Unused offer)
    2. There are 1670214 records in this data set along with 37 attributes

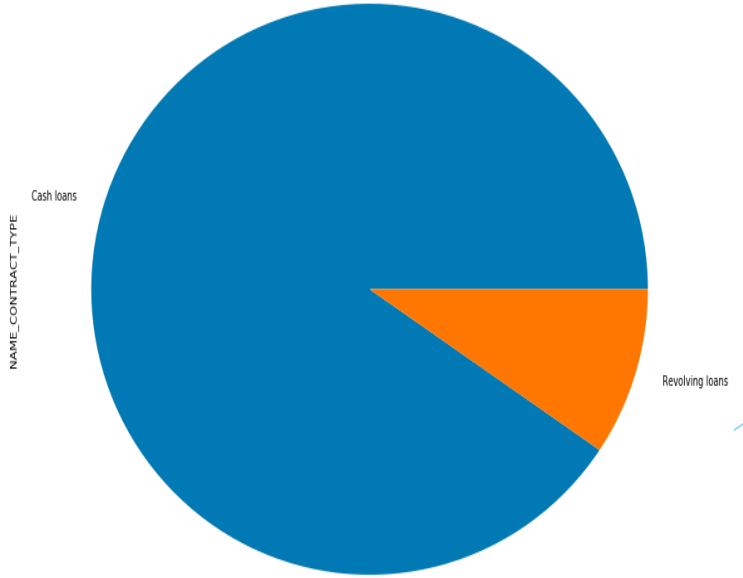
Data Distribution

Data Distribution for Application data sets is as

1. Application Data
    1. 91.9% of the client had no difficulty in payment & 8.1% has difficulty
    2. Distribution of application data by Loan Type

        Cash loans 90.5% |
        Revolving loans 9.5%

    3. Data distribution is imbalanced



Data Distribution

Distribution for Previous Lona data sets is as

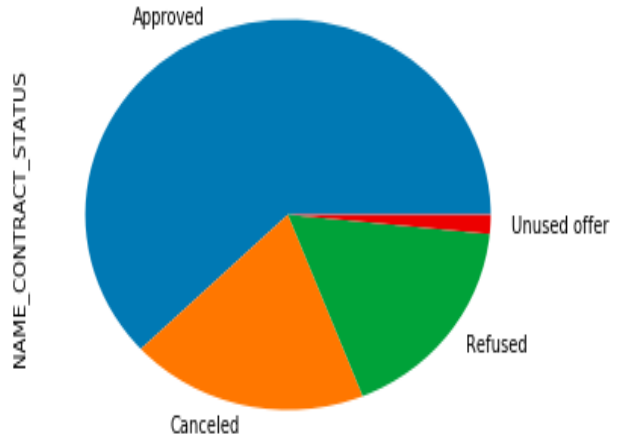
1. Previous loan Data
    1. Distribution of Previous applications By Status

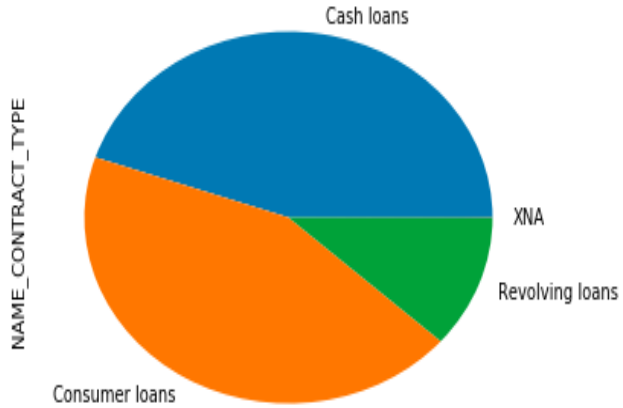
        Approved 62.1% | Refused 17.4%

        Unused offer 1.6% | Canceled 18.9%
    2. Distribution of Previous applications by Loan Type

        Cash loans 44.8% | Consumer loans 43.7%

        Revolving loans 11.6% | XNA 0. 02%





Analysis Approach

1. Data Cleaning
    1. Irrelevant Data Handling
    2. Missing Value handling
2. Univariate Analysis
    1. Identify Outliers in relevant columns
    2. Identify patterns using one attribute at a time
    3. Segmented Analysis using one column at a time
3. Bivariate & Multivariate Analysis
    1. Identify patterns using 2 or more attributes at a time
    2. Identify correlation between attributes
4. Top Correlations
5. Recommendation
    1. Build recommendation for identifying the applicants with high risk of default

Data Cleaning – Irrelevant Column Treatment

1. Application data
    i. We identified & dropped the columns which have 30% or more null values
    ii. There were 50 such columns in application data set
    iii. Few columns were excluded from analysis based on their relevance
2. Previous Application data
    1. We dropped 2 columns with 99% null values (['RATE_INTEREST_PRIVILEGED','RATE_INTEREST_PRIMARY']).
    2. Rest of the column with null values looked genuine case in given context.
    3. Few columns were excluded from Analysis based on their relevance

Data Cleaning – Missing Value Treatment

1. Application data
    i. Columns with 1% or less null values were identified & records containing these nulls were filtered out
    ii. Check for duplicate records was done, but no duplicates found in this data set.
2. Previous Application data
    1. There were records with loan application amount as zero; these records were deleted.
    2. Check for duplicate records was done, but no duplicates found in this data set.

Data Cleaning –Data Correctness & Formatting

1. Identification of Categorical & Numerical variable was done.
2. Data Type for all columns were checked & ensured that they match with values in the column
3. Columns having bigger number such as total income or goods price etc. were converted in 100K unit for better visualization
4. Values in column having days were negative, changed them to positive before analysis.
5. Numeric column values were rounded to 2 decimal places

Univariate Analysis – Approach

1. Application Data set was segmented by TARGET
2. Previous application data was filtered so that only records matching with current loan application id are analyzed against TARGET
3. Previous application data was also segmented
4. Univariate analysis was performed for full data & segmented data.
5. Only relevant columns were included in analysis

Univariate Analysis (Numerical)
DAYS_LAST_PHONE_CHANGE

This column tells that how many days before, applicant changed his phone number

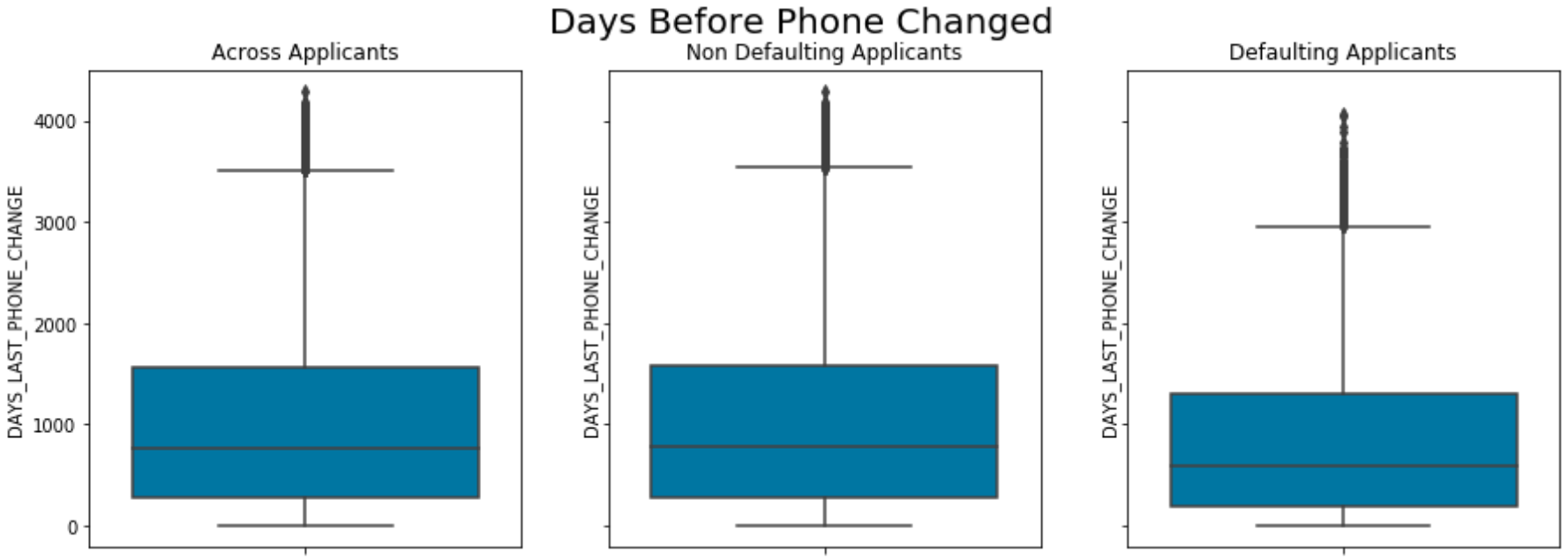
```
ad_df.DAYS_LAST_PHONE_CHANGE = ad_df.DAYS_LAST_PHONE_CHANGE*-1
ad_df.DAYS_LAST_PHONE_CHANGE.describe()
```

```
count    305545.000000
mean        963.776884
std         827.163502
min          -0.000000
25%         274.000000
50%         758.000000
75%        1571.000000
max        4292.000000
Name: DAYS_LAST_PHONE_CHANGE, dtype: float64
```

Observation :-

1. There is difference between mean & median.
2. Max values is also too far from the 75 percentile.
3. This data may have outliers, we will confirm it using the Box plot

Univariate Analysis – DAYS_LAST_PHONE_CHANGE



Observations :-

1. There are higher values for days but they are very adjacent to the upper fence, so not considering them outlier.
2. Defaulting applicants seems to be changing their phone number in more recent timeline compared to not defaulting applicants.

Univariate Analysis (Numerical)
EXT_SOURCE_2

This column tells about the applicants normalized score from external data source

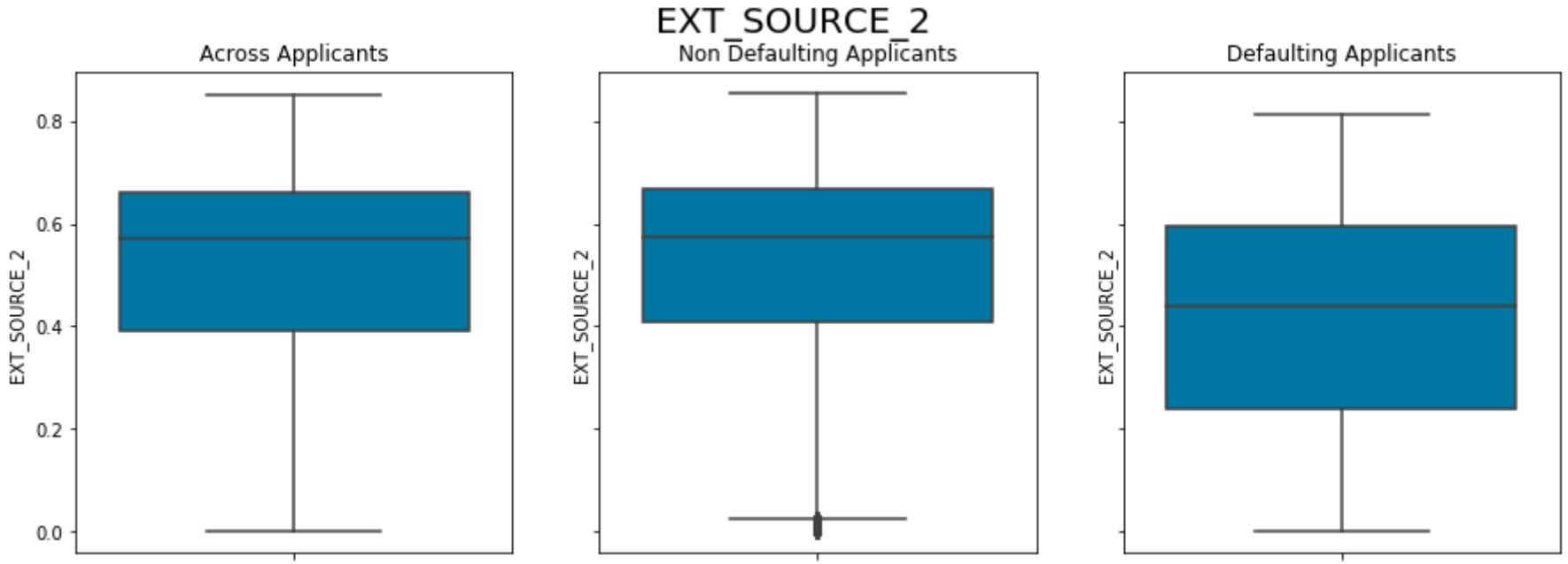
```
ad_df.EXT_SOURCE_2.describe()
```

```
count    305545.000000
mean          0.514265
std           0.191127
min           0.000000
25%           0.390000
50%           0.570000
75%           0.660000
max           0.850000
Name: EXT_SOURCE_2, dtype: float64
```

Observations :-

1. Mean & Median are close to each other.
2. Max Value is also not too far from 75 percentile.
3. There are no outliers in this column.

Univariate Analysis - EXT_SOURCE_2



Observations :-

1. Applicants with higher mean score have no difficulty in payments.
2. On the other hand defaulting candidates have lower mean score.

Univariate Analysis (Categorical)

Gender

This column tells about Applicants Gender

```
ad_df.CODE_GENDER.value_counts(normalize=True)
```

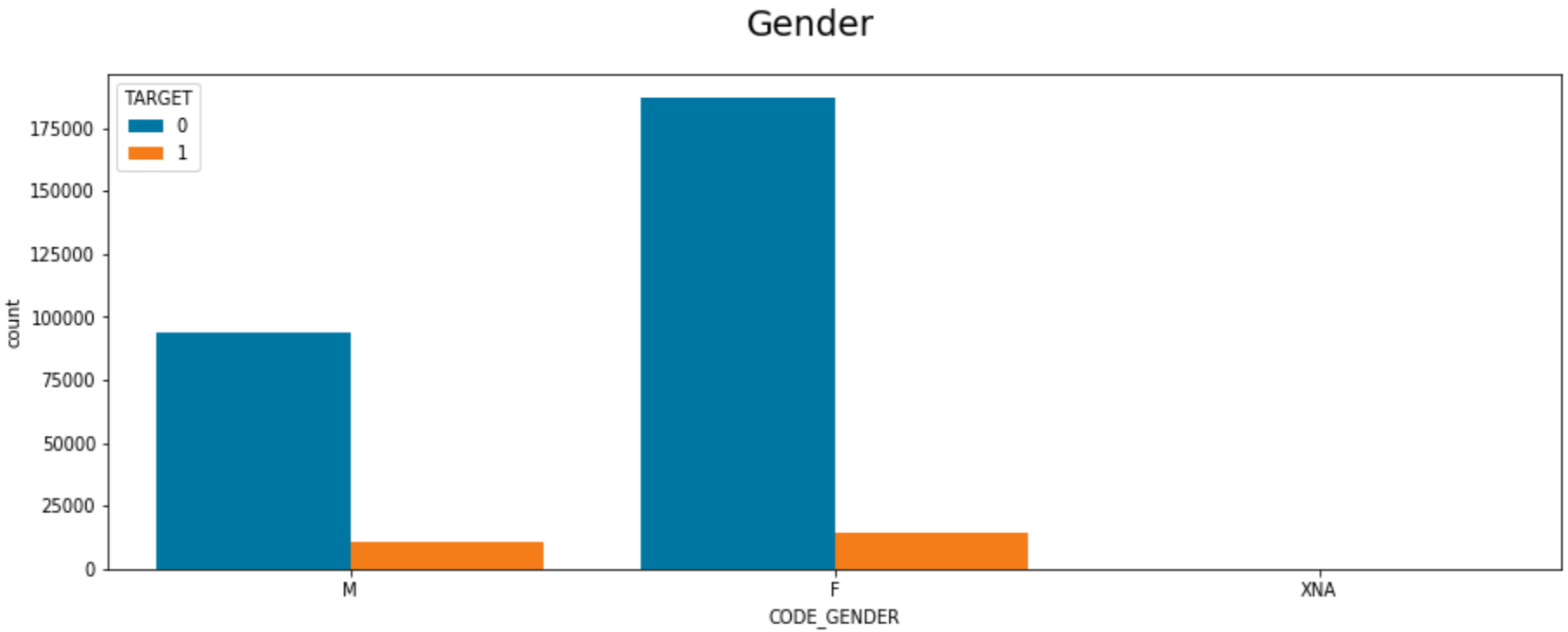
```
F      0.658162
M      0.341825
XNA    0.000013
Name: CODE_GENDER, dtype: float64
```

Observations :-

1. 65.8% applicants are Female
2. 34.2% candidates are Male

Univariate Analysis - Gender

This column tells about Applicants Gender



Observations :-

1. Application data has more number of Female applicants
2. Proportionally Male borrowers are more prone to defaults

Univariate Analysis (Categorical)
FLAG_OWN_REALTY

This columns tells about whether applicant owns a house or flat.

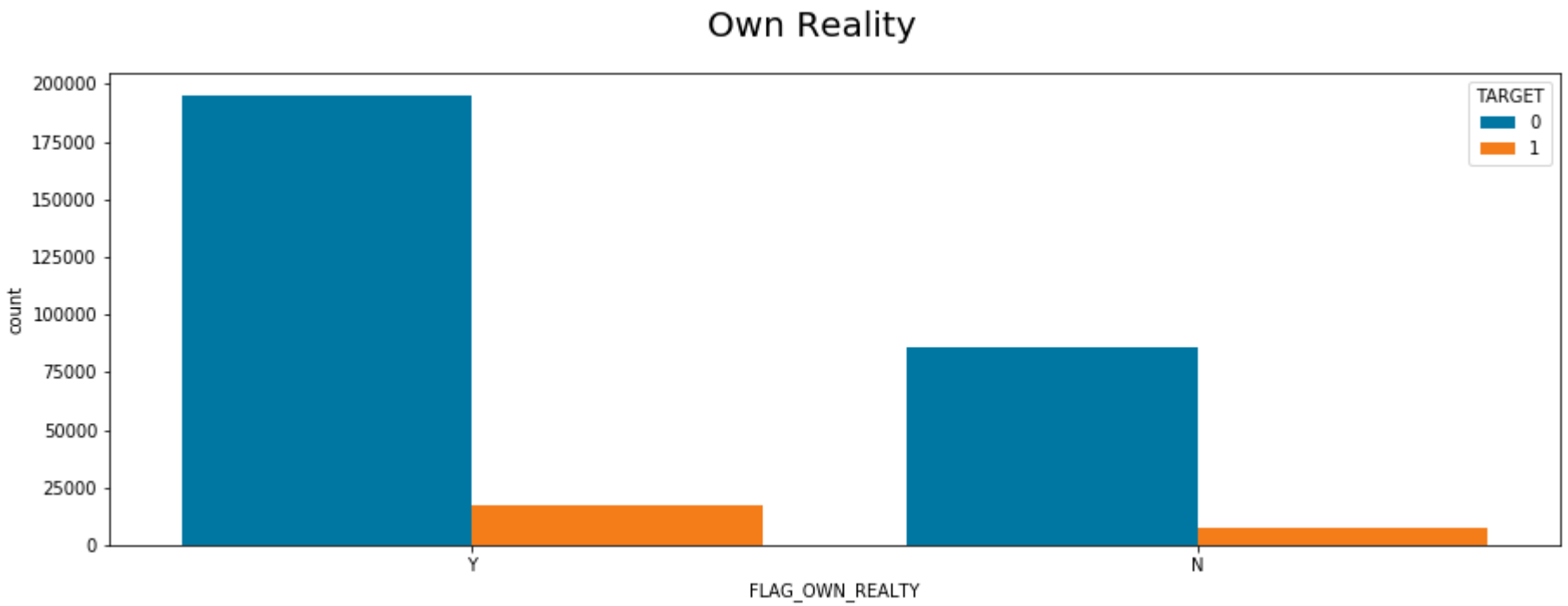
```
ad_df.FLAG_OWN_REALTY.value_counts(normalize=True)
```

```
Y    0.693463
N    0.306537
Name: FLAG_OWN_REALTY, dtype: float64
```

Observation :-

1. Majority of applicants (69.3%) owns house or flat.
2. 30.6% applicants don’t own house or flat

Univariate Analysis - FLAG_OWN_REALTY



Observations :-

1. Applicants owning a house or flat are less prone to defaults

Univariate Analysis (Categorical)
NAME_FAMILY_STATUS

This column tells about the family status (Married, Single, Separated etc.) of the applicants

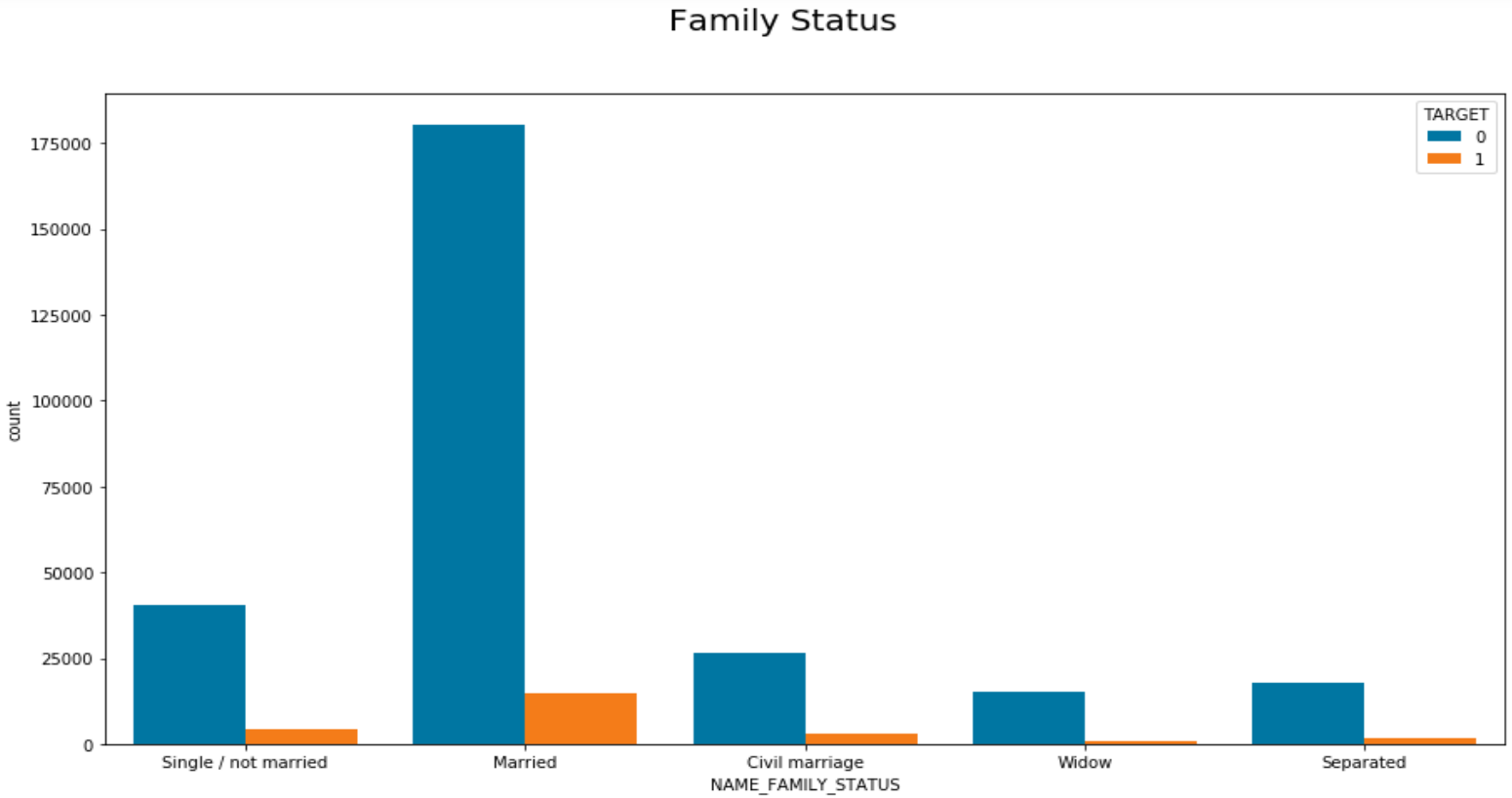
```
ad_df.NAME_FAMILY_STATUS.value_counts(normalize=True)
```

```
Married                 0.639110
Single / not married    0.147481
Civil marriage          0.096784
Separated               0.064311
Widow                   0.052313
Name: NAME_FAMILY_STATUS, dtype: float64
```

Observations :-

1. Majority of the applicants are married (63.9%)
2. Single candidates are 14.7%
3. Rest of the candidates are in separated, widow & civil marriage category

Univariate – NAME_FAMILY_STATUS



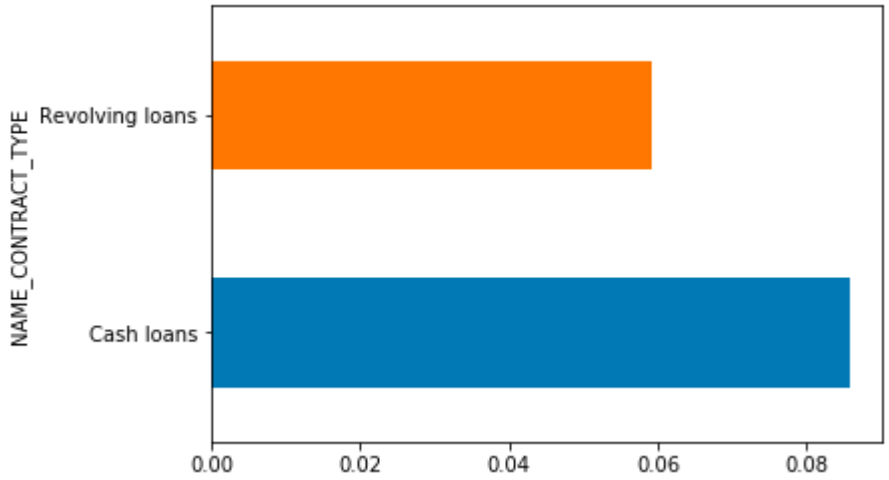
Observation :-

1. Proportionally married applicants are less likely to default.

2. Applicants in category ‘Civil Marriage’, ‘Widow’ and ‘Separated’ are relatively more prone to default

Bivariate Analysis

NAME_CONTACT_TYPE vs TARGET

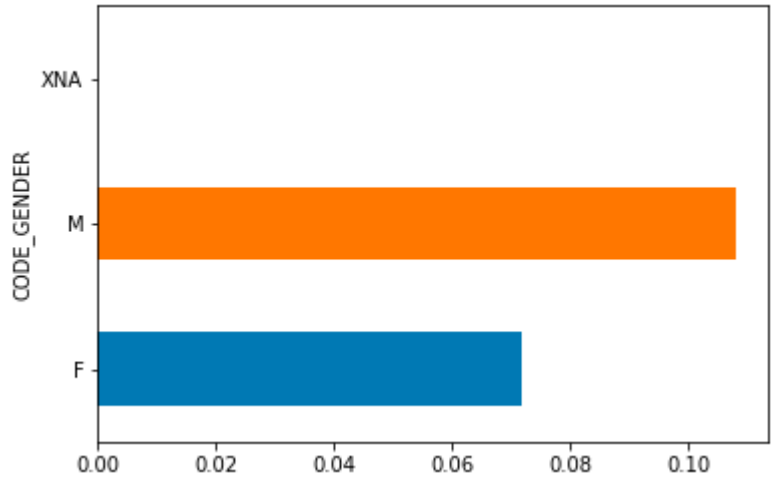


Observations :-

1-Cash loan segment has higher defaulting applicants

Bivariate Analysis

CODE_GENDER vs TARGET

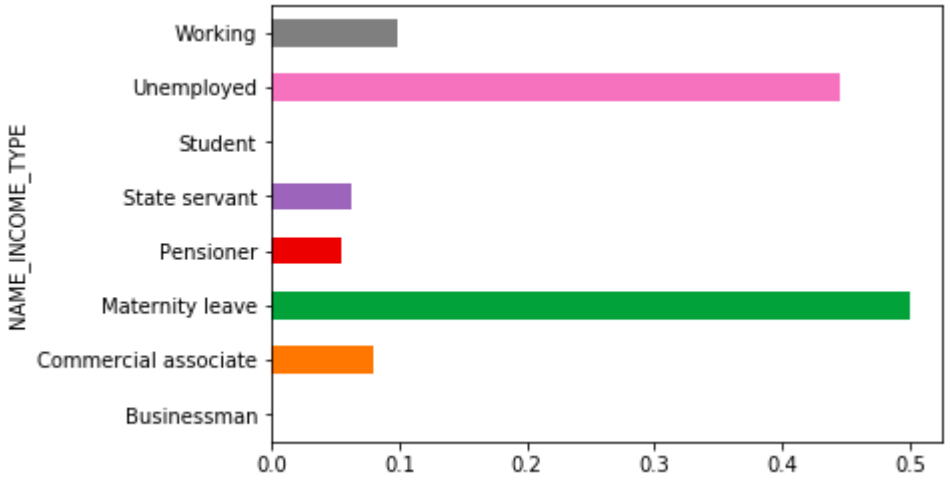


Observations :-

1-Male applicants are more likely to default than Female applicants

Bivariate Analysis

NAME_INCOME_TYPE vs TARGET

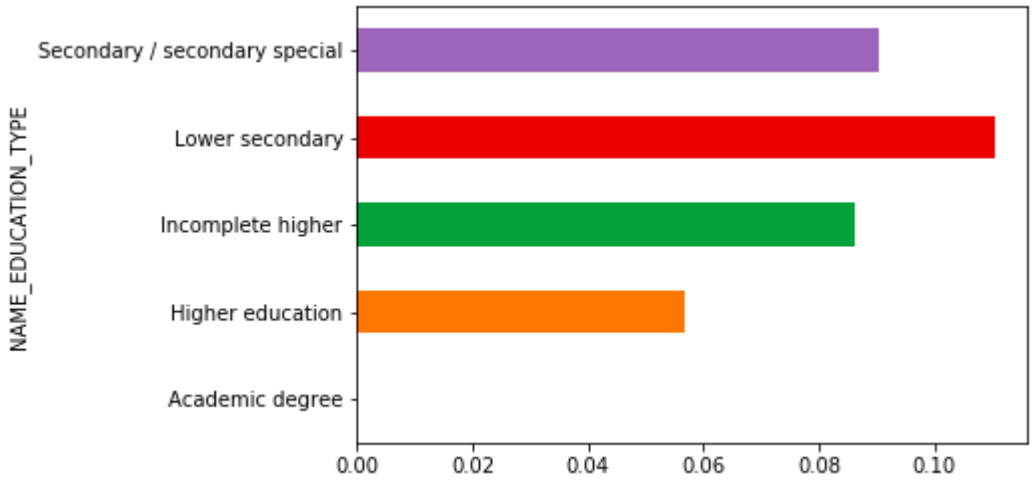


Observations :-

1-Unemplloyed applicants or applicants on Maternity leave have high chances of default

Bivariate Analysis

NAME_EDUCATION_TYPE vs TARGET

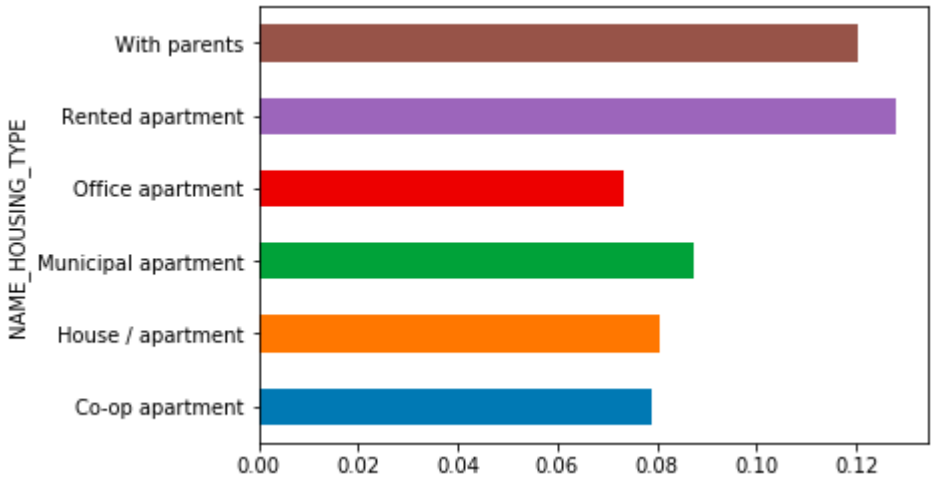


Observations :-

1-Applicants with education level ‘Lower Secondary’, ‘Secondary/Secondary Special’ or ‘Incomplete Higher’ are more likely to default

Bivariate Analysis

NAME_HOUSING_TYPE vs TARGET

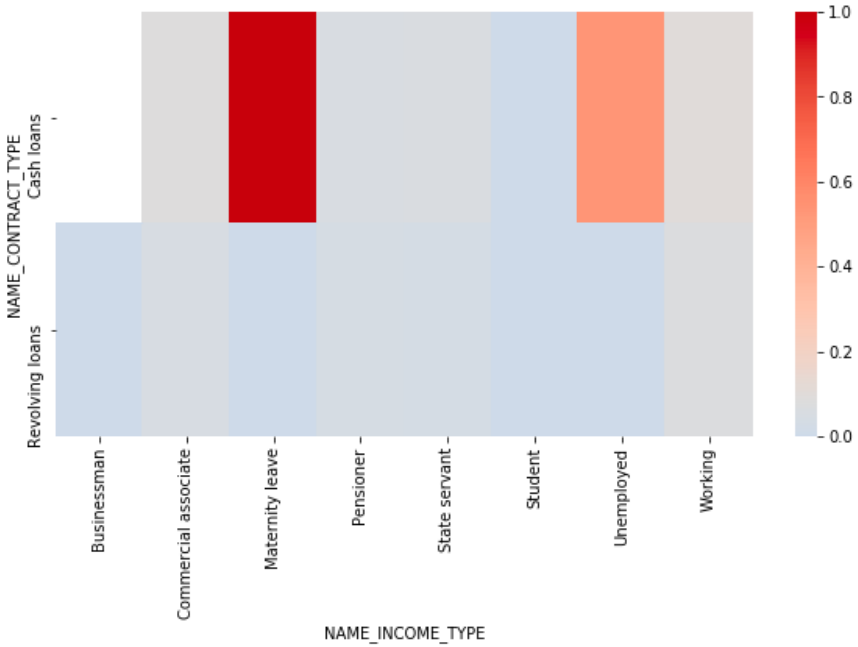


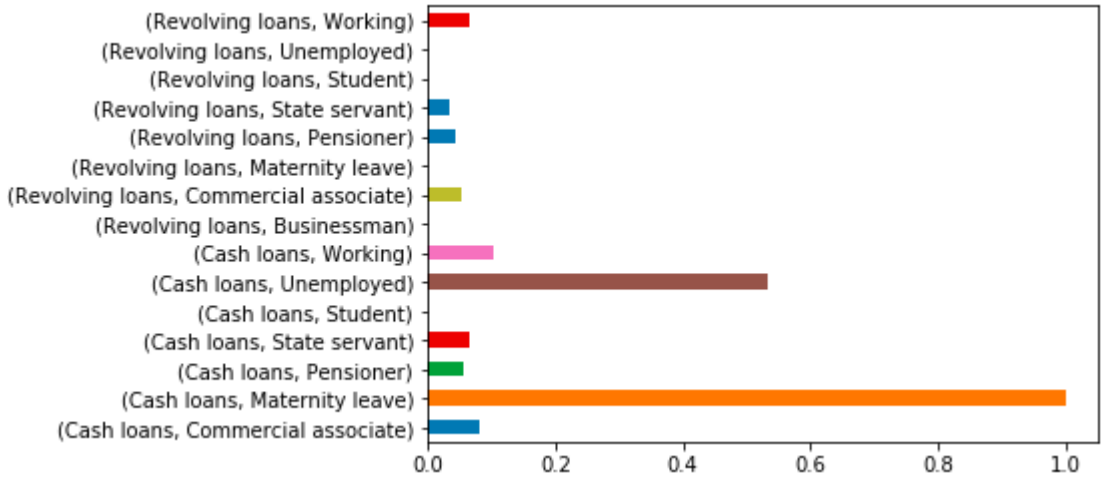
Observations :-

1-Applicants living with parents or in rented apartment are more likely to default.

Multivariate Analysis

NAME_CONTACT_TYPE vs NAME_INCOME_TYPE vs TARGET





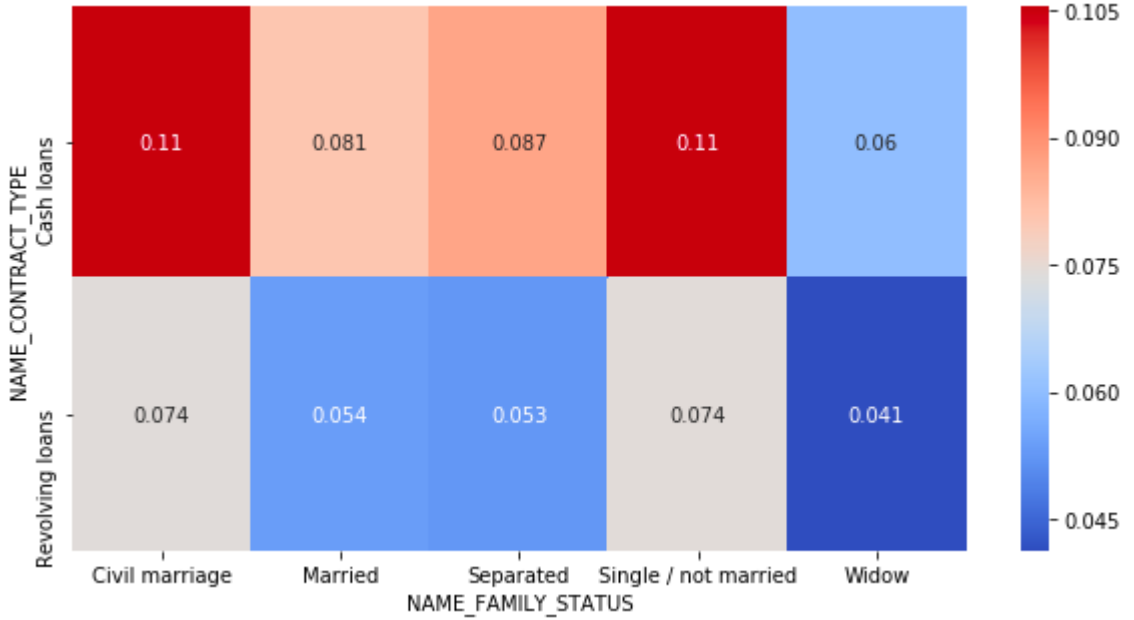
Observations :-

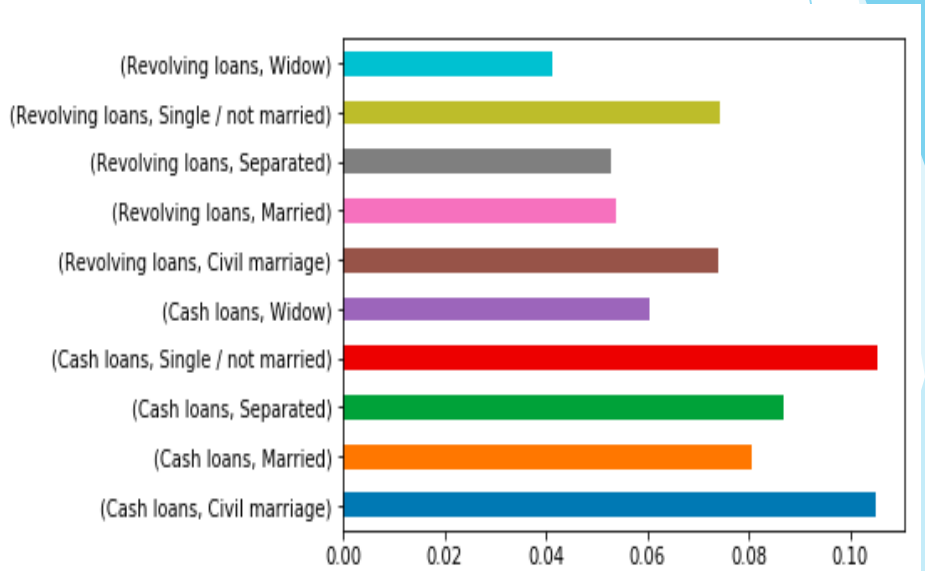
1-Unemployed applicant or applicant on Maternity leave are highly likely to default in cash loans

2-Businessman seems most suitable applicant for Revolving loans.

Multivariate Analysis

NAME_CONTACT_TYPE vs NAME_FAMILY_STATUS vs TARGET



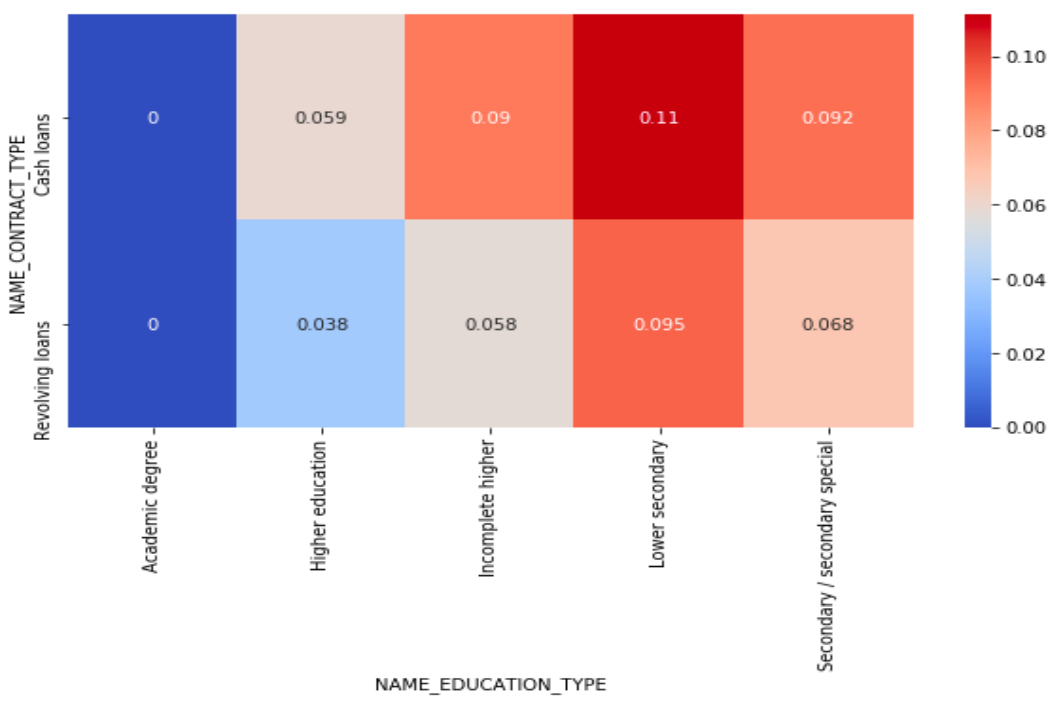


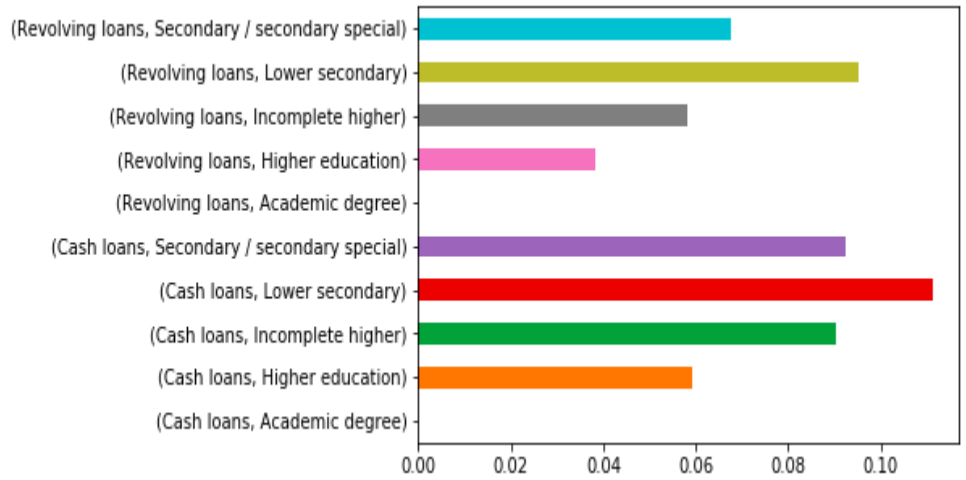
Observations :-

1. Applicants with marital status ‘Single’ or ‘Civil marriage’ are highly likely to default in cash loans
2. Applicants with marital status as ‘widow’ are most suitable (i.e. least likely to default) for Revolving loans.

Multivariate Analysis

NAME_CONTACT_TYPE vs NAME_EDUCATION_TYPE vs TARGET





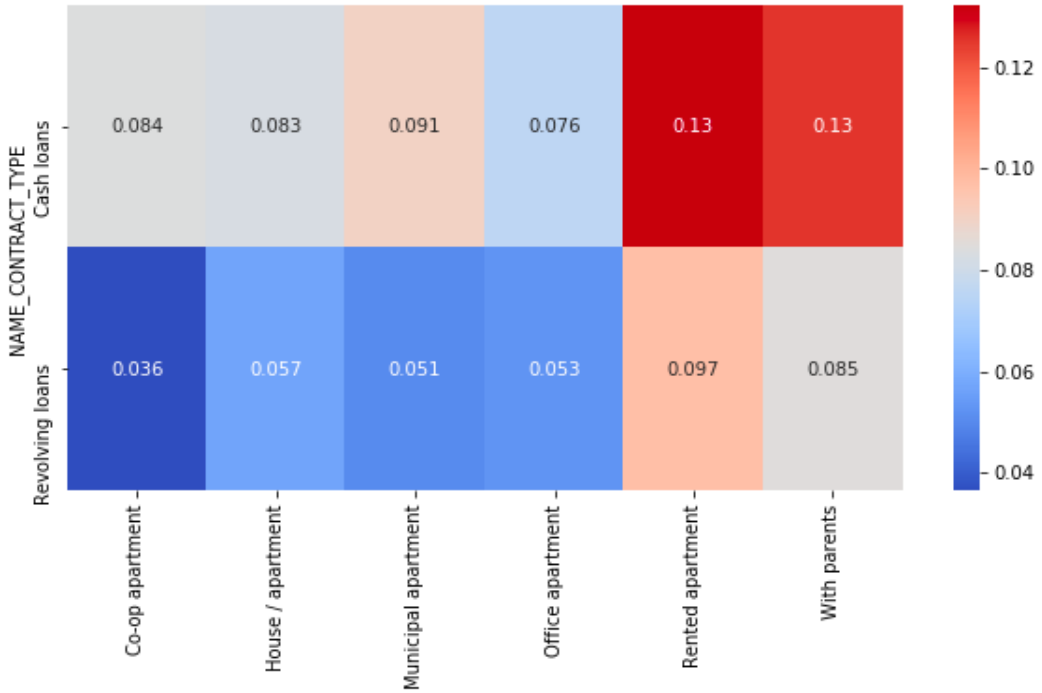
Observations :-

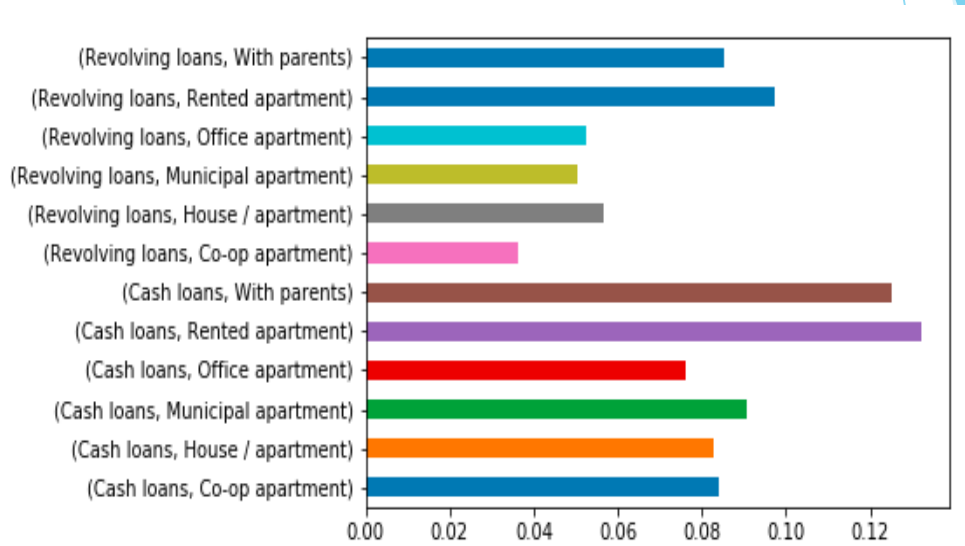
1. Lower the education, higher the default: Applicants with education status as ‘Lower Secondary’, ‘Secondary/Secondary Special’, ‘Incomplete higher’ & ‘Higher education’ are most likely to Credit default in their given order of their sequence, with greater vulnerability in Cash loans followed by Revolving loans.

2. Applicants with Academic degree are not likely to default in cash as well as revolving loan.

Multivariate Analysis

NAME_CONTACT_TYPE vs NAME_HOUSING_TYPE vs TARGET



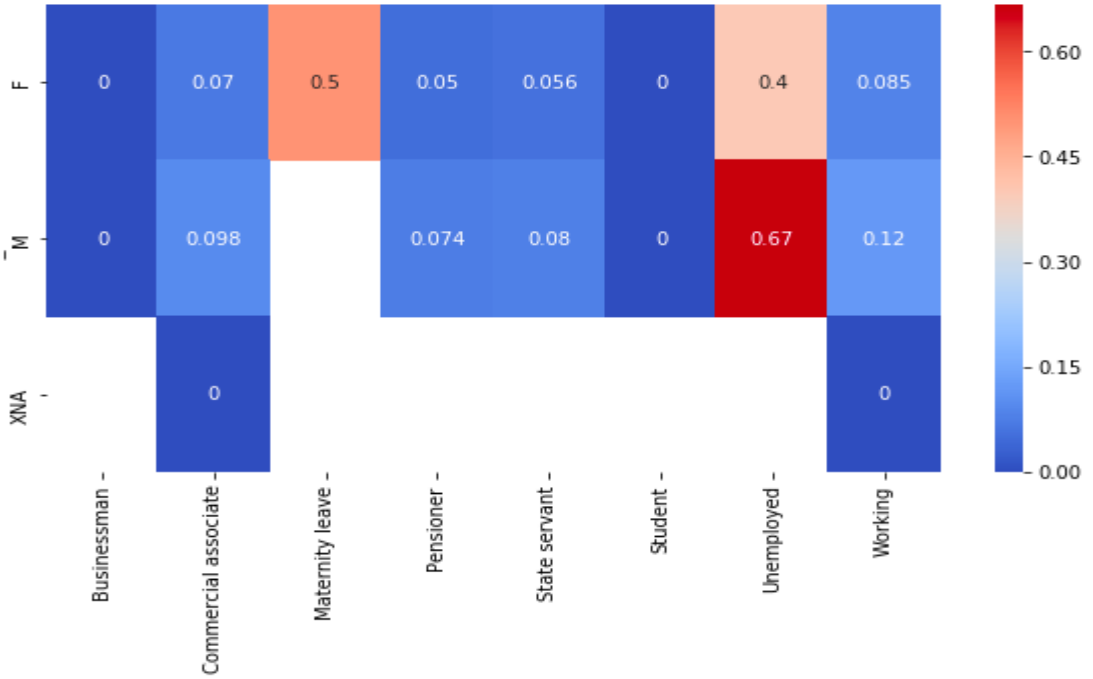


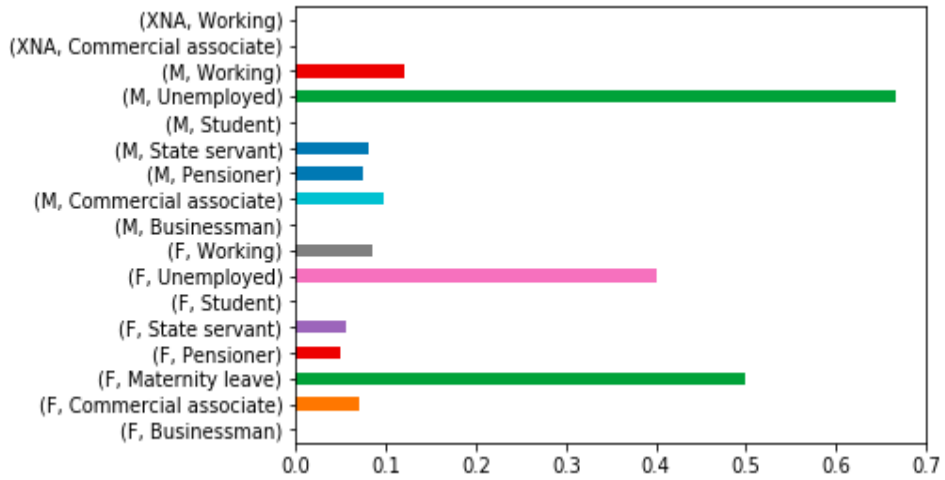
Observations :-

1. Applicants living with parents or in rented apartment are more likely to default in cash loans
2. Applicants living in rented apartment are more likely to default in Revolving loans
3. Applicants living in ‘Co-op apartment’, ‘own house/Apartment’, ‘Municipal Apartment’ or ‘Office Apartment’ are less likely to default in Revolving loans

Multivariate Analysis

CODE_GENDER vs NAME_INCOME_TYPE vs TARGET





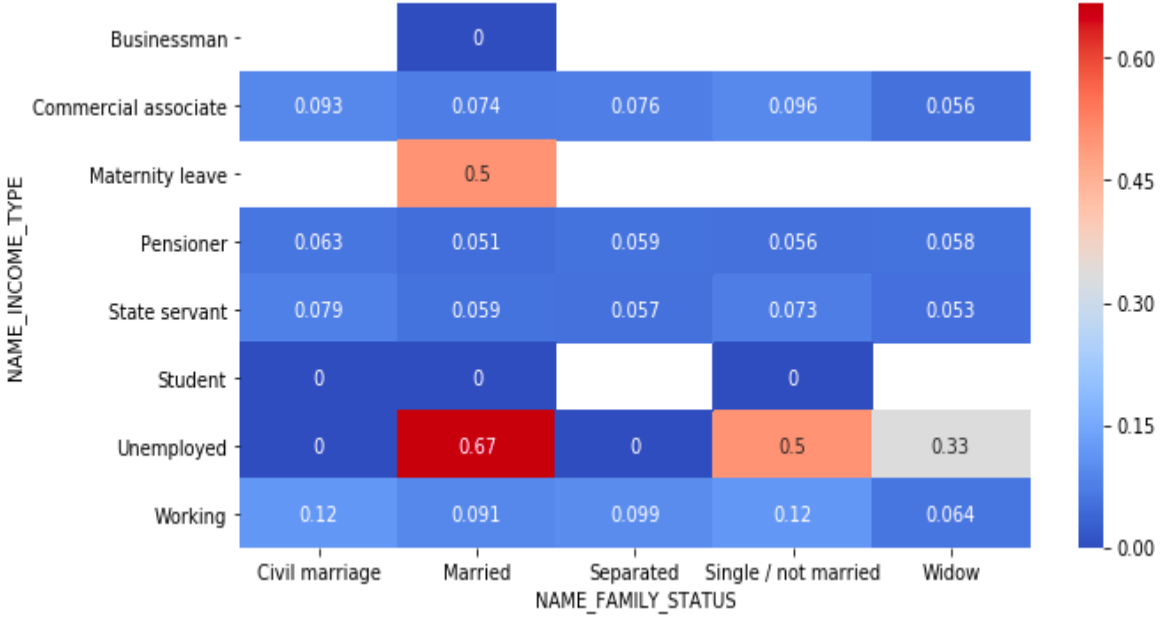
Observations :-

1-Unemployed applicants are more likely to default irrespective of gender

2-Female candidate on maternity leave are also more likely to default

Multivariate Analysis

NAME_INCOME_TYPE vs NAME_FAMILY_STATUS vs TARGET

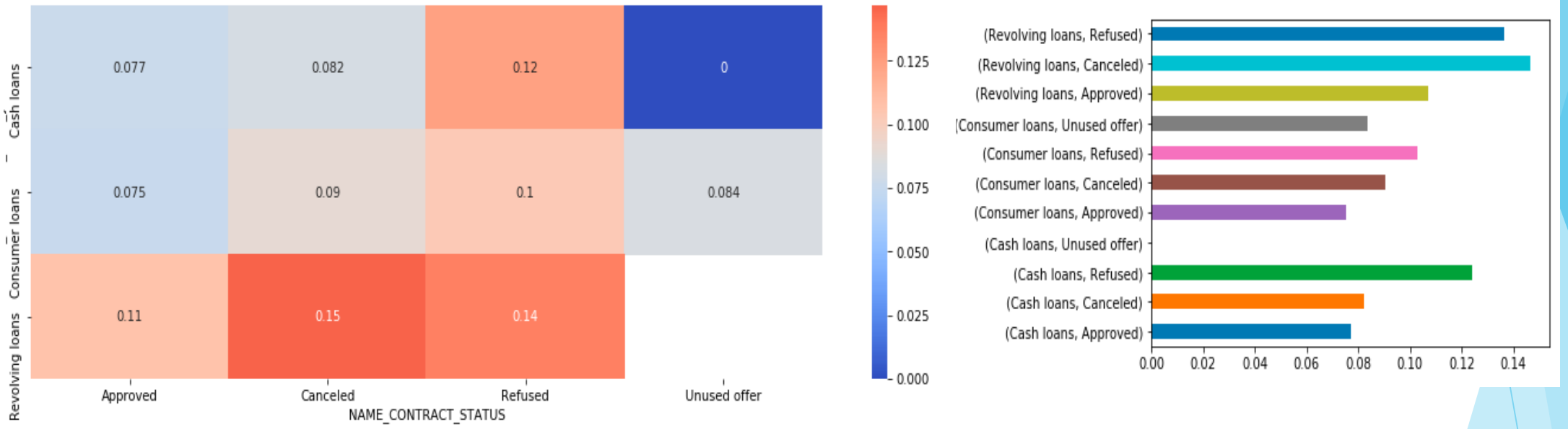


Observations :-

1. Married Unemployed applicants are more likely to default, followed closely by Unemployed Singles & Married client on Maternity leave.

Multivariate Analysis

NAME_CONTRACT_TYPE vs NAME_CONTRACT_STATUS vs TARGET

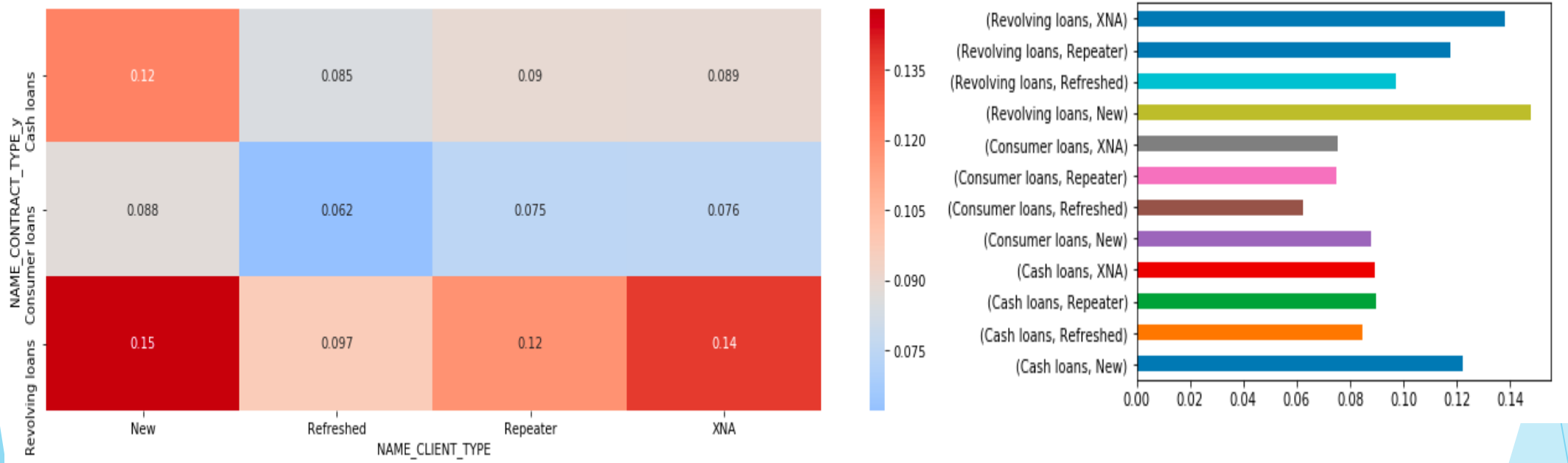


Observations :-

1. Applicants whose previous loan cancelled are more likely to default in case of revolving loan
2. Applicants whose previous loan application was refused are more likely to default across all loan type

Multivariate Analysis

NAME_CONTRACT_TYPE vs NAME_CLIENT_TYPE vs TARGET

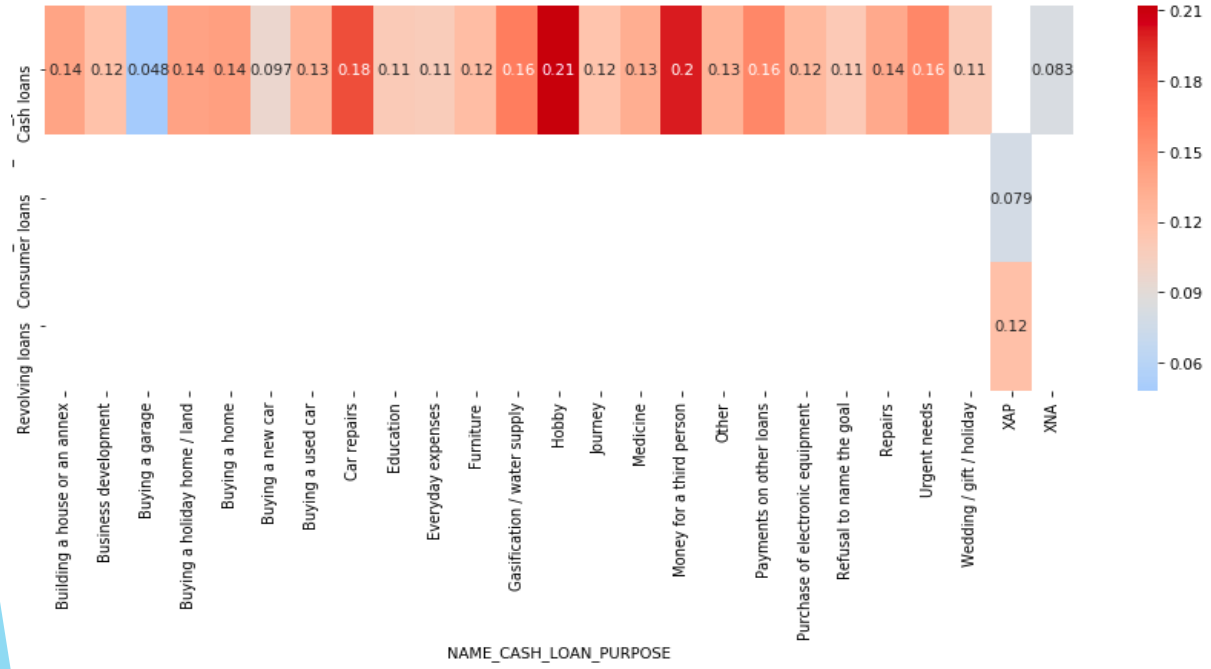


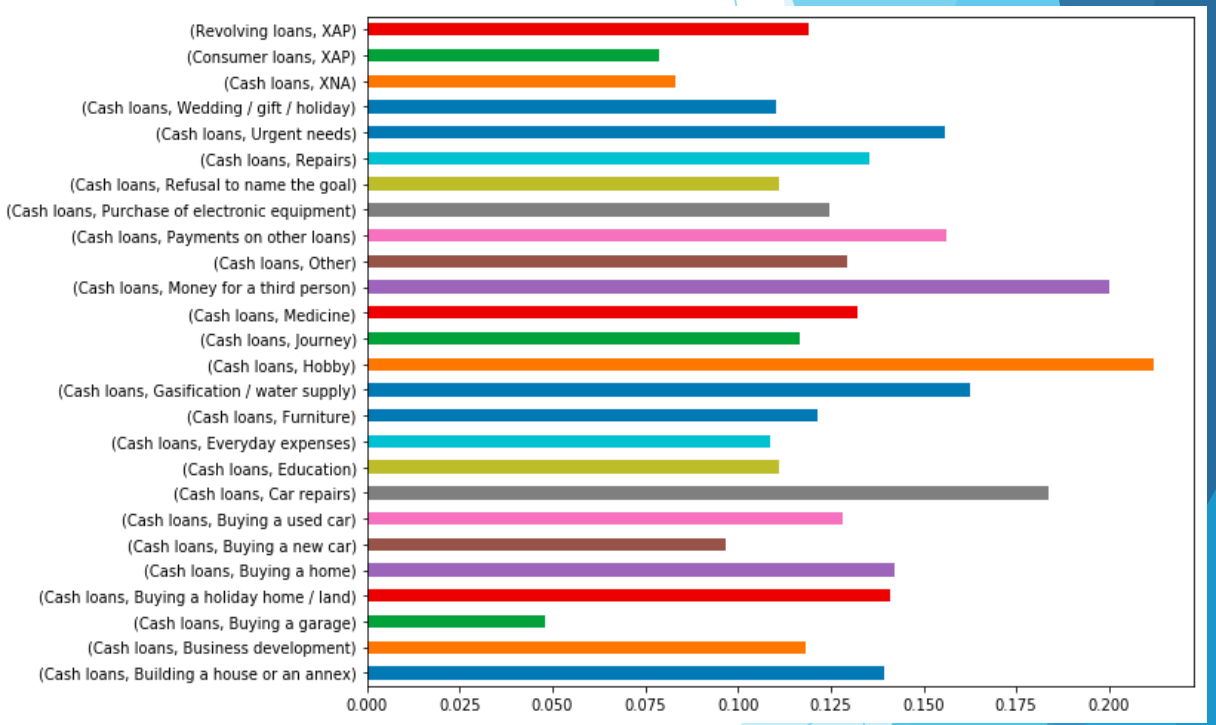
Observations :-

1. In revolving loan segment NEW clients are more likely to default followed closely by a Repeater
2. In cash loan segment new applicants are more likely to default
3. Refreshed client applying for Consumer loans are least likely to default.

Multivariate Analysis

NAME_CONTRACT_TYPE vs NAME_CASH_LOAN_PURPOSE vs TARGET





Observations :-

1. Applicants taking cash loan for Hobby, Car Repairs & for Third person are highly likely to default.
2. Applicants taking Cash loans for Buying a garage are least likely to default.

Top 10 correlations

#	Variable #1	Variable #2	Correlation index
1	DAYS_TERMINATION	DAYS_LAST_DUE	1
2	OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	1
3	DAYS_LAST_DUE_1ST_VERSION	DAYS_TERMINATION	0.97
4	DAYS_LAST_DUE_1ST_VERSION	DAYS_LAST_DUE	0.97
5	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.94
6	REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.87
7	DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.86
8	AMT_ANNUITY_y	AMT_GOODS_PRICE_y	0.76
9	REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.44
10	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.43

Note: Analyzed from the merged data set 'pa_df_temp'

1. 'DAYS_TERMINATION' - 'DAYS_LAST_DUE' has the highest correlation in the dataset along with 'OBS_30_CNT_SOCIAL_CIRCLE' - 'OBS_60_CNT_SOCIAL_CIRCLE' with index value 1.
2. 'REG_REGION_NOT_WORK_REGION' – 'REG_REGION_NOT_LIVE_REGION' has the 10th highest correlation with index value of 0.43.

Recommendation

Gender	Loan Type	Frequency of Phone Change	External Normalized Score	Owns House	Family Status	Income Type	Education	Housing Type	Pervious Loan Status	Client Type	Loan Pupose
M/F	Cash	High	below 0.5	No	Single Civil Marriege Married	Unemployed	Lower Secondary Secondry Special Incomplete Higher	With Parents Rented	Refused	New	Hobby Car Repairs For Third Person
M/F	Revolving	High	below 0.5	No				Rented	Cancelled	New Repeater	
F	Revolving	High	below 0.5	No	Widow	Maternity Leave		Rented	Cancelled	New Repeater	

1. As per the pattern found in data set, loan applicants having above characteristics are high likely to default the loan.
2. Banks should keep a watch on such candidates & extra precaution should be taken before approving loan for such candidates.
3. Banks can reject the loan if majority of conditions satisfy.
4. Banks can approve a reduced loan amount at higher interest rate if not all but some of these characteristics are found in applicants

Thank You