Cairo University
Faculty of Engineering
Computer Engineering Department

Bank Loan Default Risk Analysis

Project Report

Problem Description

The project aims to analyze risk factors in loan applications to help banks minimize losses that happen due to defaults. As there are some applications with no credit history or missing guarantees this leads to high risks to banks and lending companies. This results in increased default rates.
By analysing previous and current loan applications, this project will aim to provide insights to how banks and lending companies can assess risk and improve lending strategies.

Pipeline

Process/Clean Data

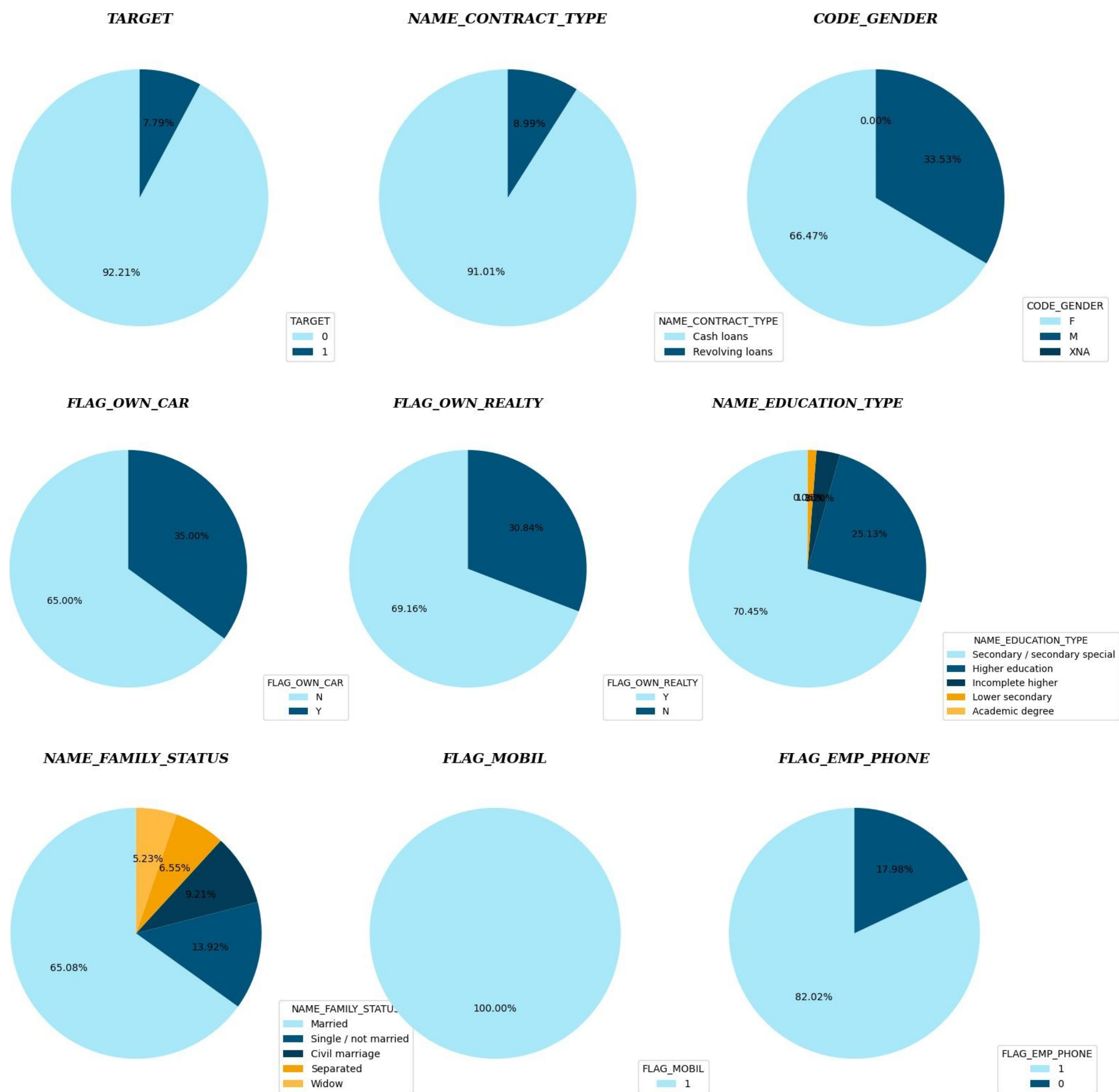
In this phase, the raw dataset was cleaned and preprocessed to ensure data quality and consistency.

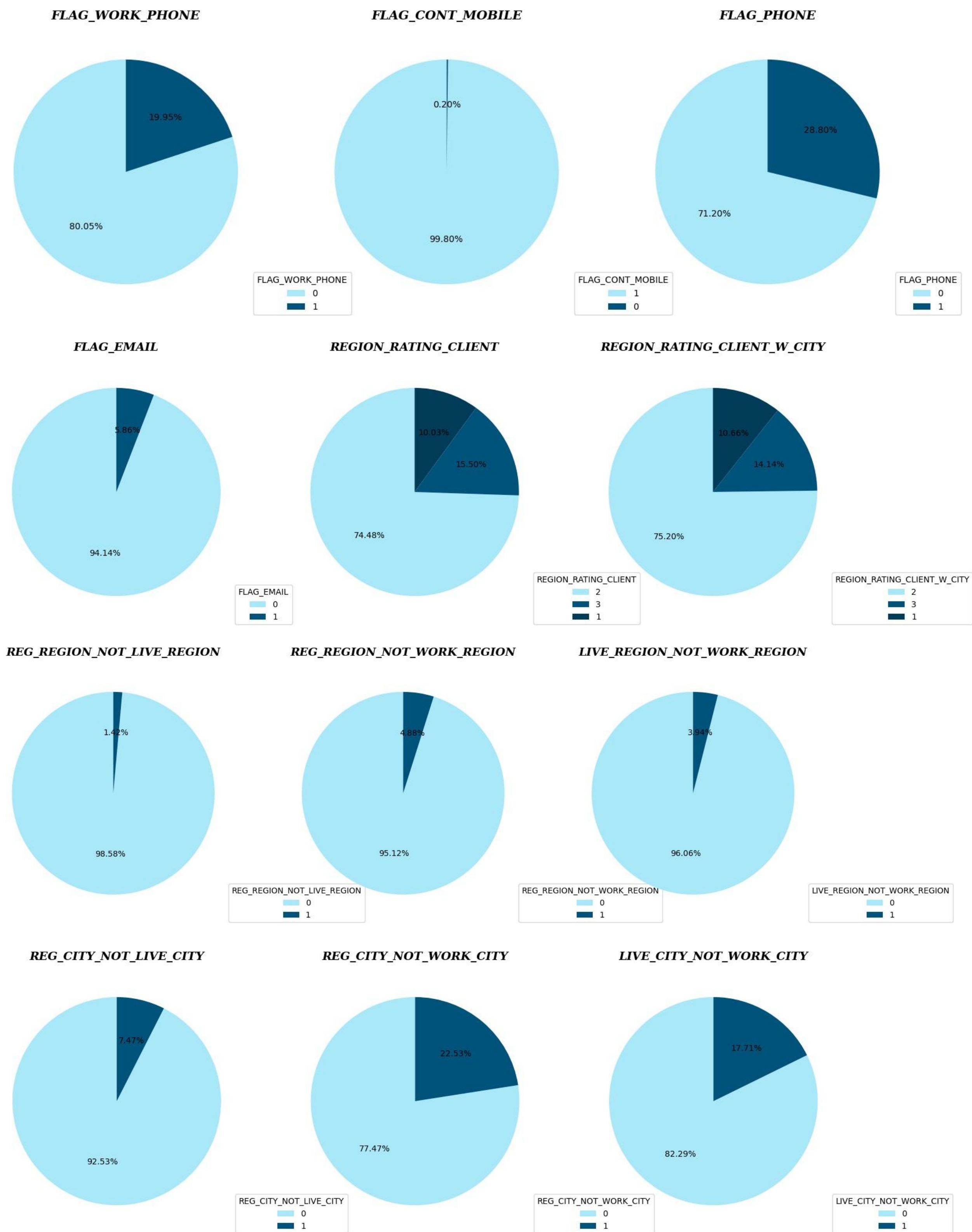
1. **Initial Exploration:**
   Exploring the shape of the dataset.
   Basic information about the dataset was displayed, including column names, data types.
   The number of unique values for each column was calculated to understand the diversity of categorical and numerical features.
2. **Data Cleaning:**
   Null Value Handling: Columns with more than 40% null values were identified and removed to ensure data quality.
   Duplicates Removal: check for duplicate rows and drop it.
   Row Removal: Rows containing any null values were dropped to maintain consistency in the dataset.

EDA

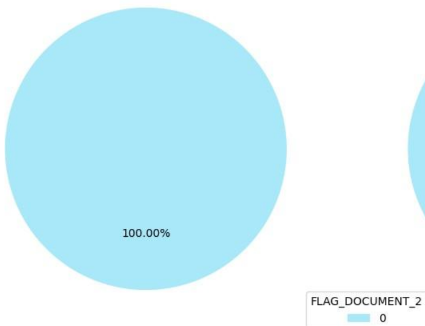
Univariate Analysis (Applications)

Categorical Variables

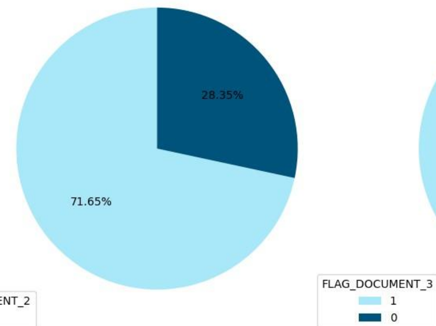



FLAG_WORK_PHONE
19.95%
80.05%
FLAG_WORK_PHONE
0
1
FLAG_CONT_MOBILE
0.20%
99.80%
FLAG_CONT_MOBILE
1
0
FLAG_PHONE
28.80%
71.20%
FLAG_PHONE
0
1
FLAG_EMAIL
5.86%
94.14%
FLAG_EMAIL
0
1
REGION_RATING_CLIENT
10.03%
15.50%
74.48%
REGION_RATING_CLIENT
2
3
1
REGION_RATING_CLIENT_W_CITY
10.66%
14.14%
75.20%
REGION_RATING_CLIENT_W_CITY
2
3
1
REG_REGION_NOT_LIVE_REGION
1.42%
98.58%
REG_REGION_NOT_LIVE_REGION
0
1
REG_REGION_NOT_WORK_REGION
4.88%
95.12%
REG_REGION_NOT_WORK_REGION
0
1
LIVE_REGION_NOT_WORK_REGION
3.94%
96.06%
LIVE_REGION_NOT_WORK_REGION
0
1
REG_CITY_NOT_LIVE_CITY
7.47%
92.53%
REG_CITY_NOT_LIVE_CITY
0
1
REG_CITY_NOT_WORK_CITY
22.53%
77.47%
REG_CITY_NOT_WORK_CITY
0
1
LIVE_CITY_NOT_WORK_CITY
17.71%
82.29%
LIVE_CITY_NOT_WORK_CITY
0
1

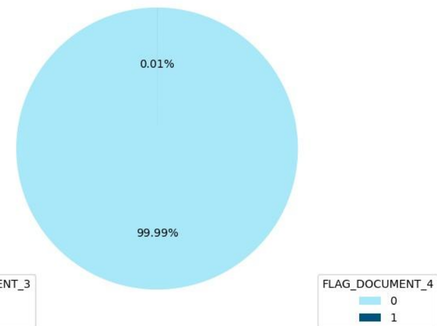
FLAG_DOCUMENT_2



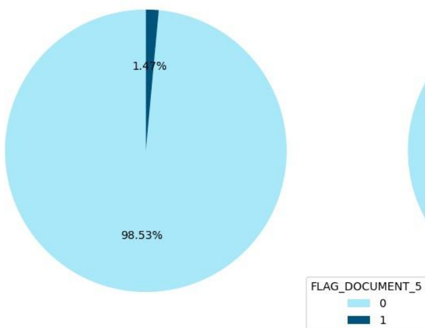
FLAG_DOCUMENT_3



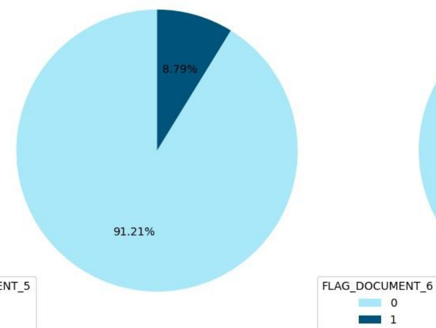
FLAG_DOCUMENT_4



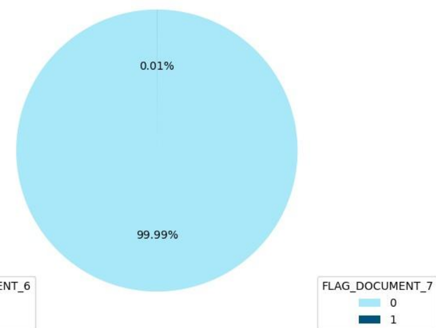
FLAG_DOCUMENT_5



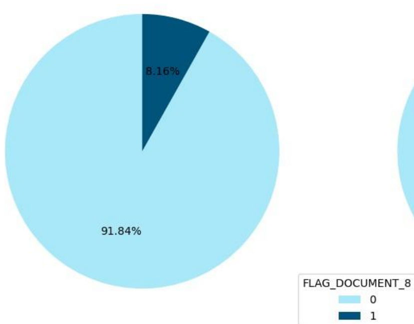
FLAG_DOCUMENT_6



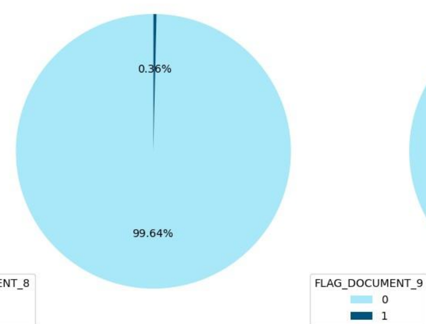
FLAG_DOCUMENT_7



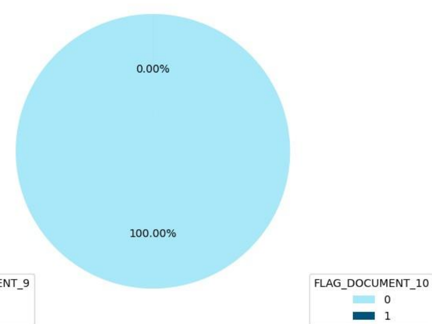
FLAG_DOCUMENT_8



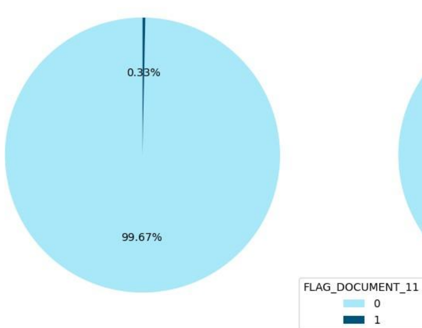
FLAG_DOCUMENT_9



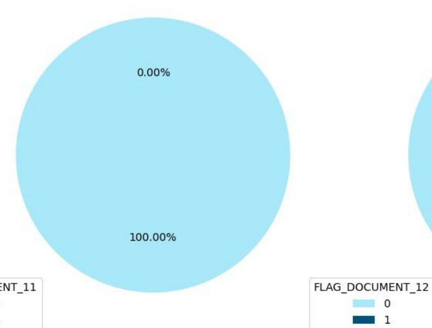
FLAG_DOCUMENT_10



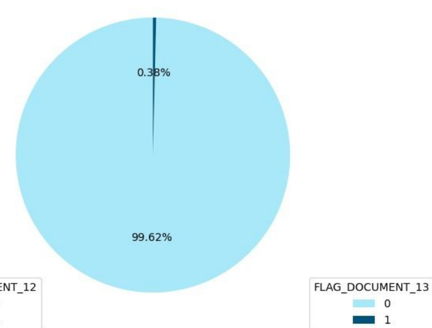
FLAG_DOCUMENT_11

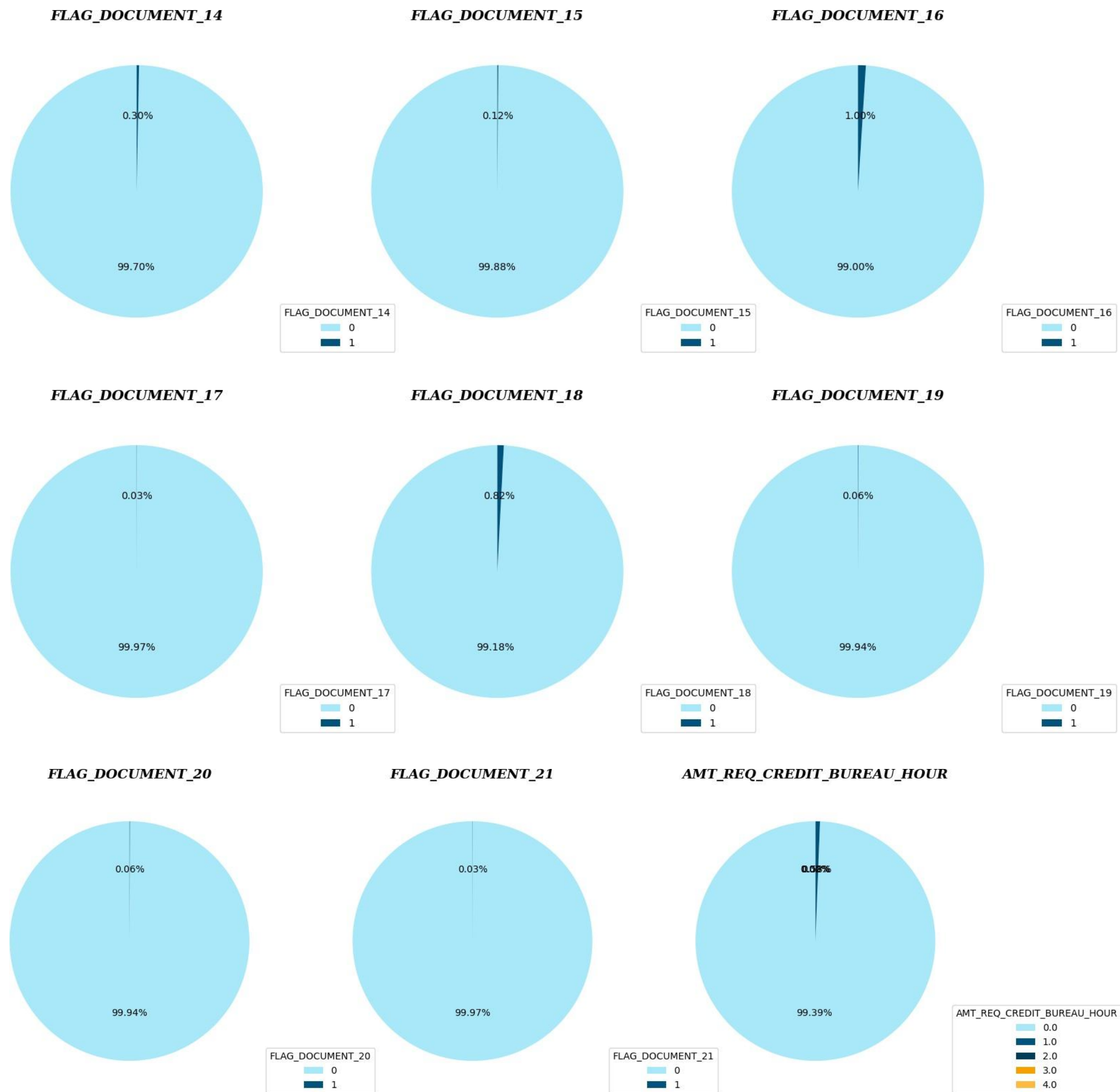


FLAG_DOCUMENT_12



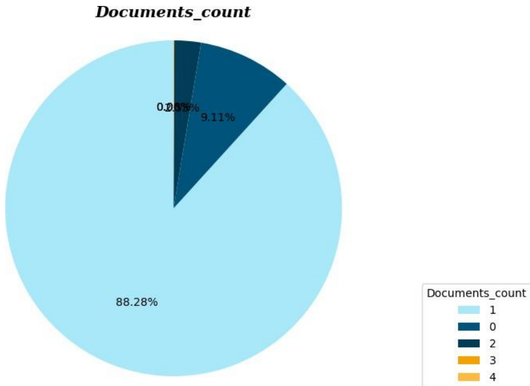
FLAG_DOCUMENT_13

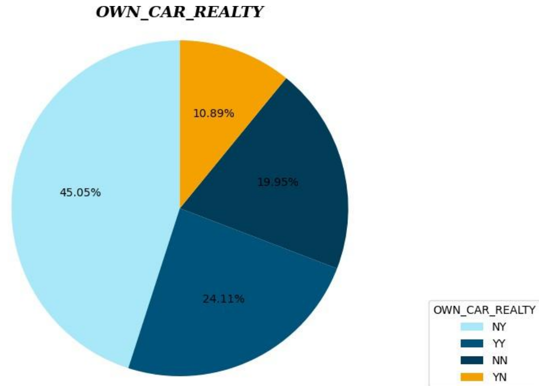


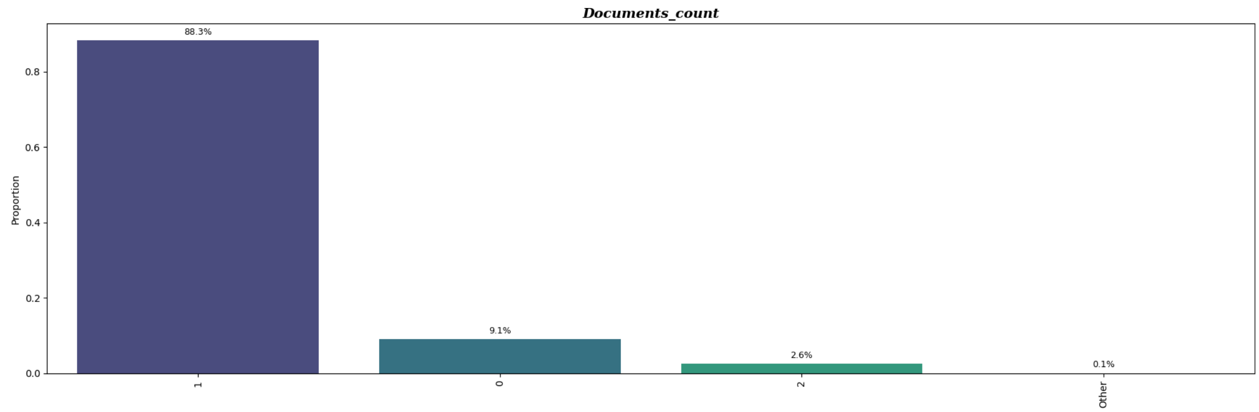


Insights:

- 7.79% of the clients face defaults.
- There are more females in the dataset. (~66.5%)
- Most of the clients (are married/don't own cars/own realty). (~65% - 70%)
- Most of the clients live in regions of middle rating.
- Most of the clients are with education ("Secondary / secondary special", "Higher education") level.
- Most clients don't deliver less than 2 documents.
- Most of the loans are cash (~91%).
- Most of the clients live in the city they work in.

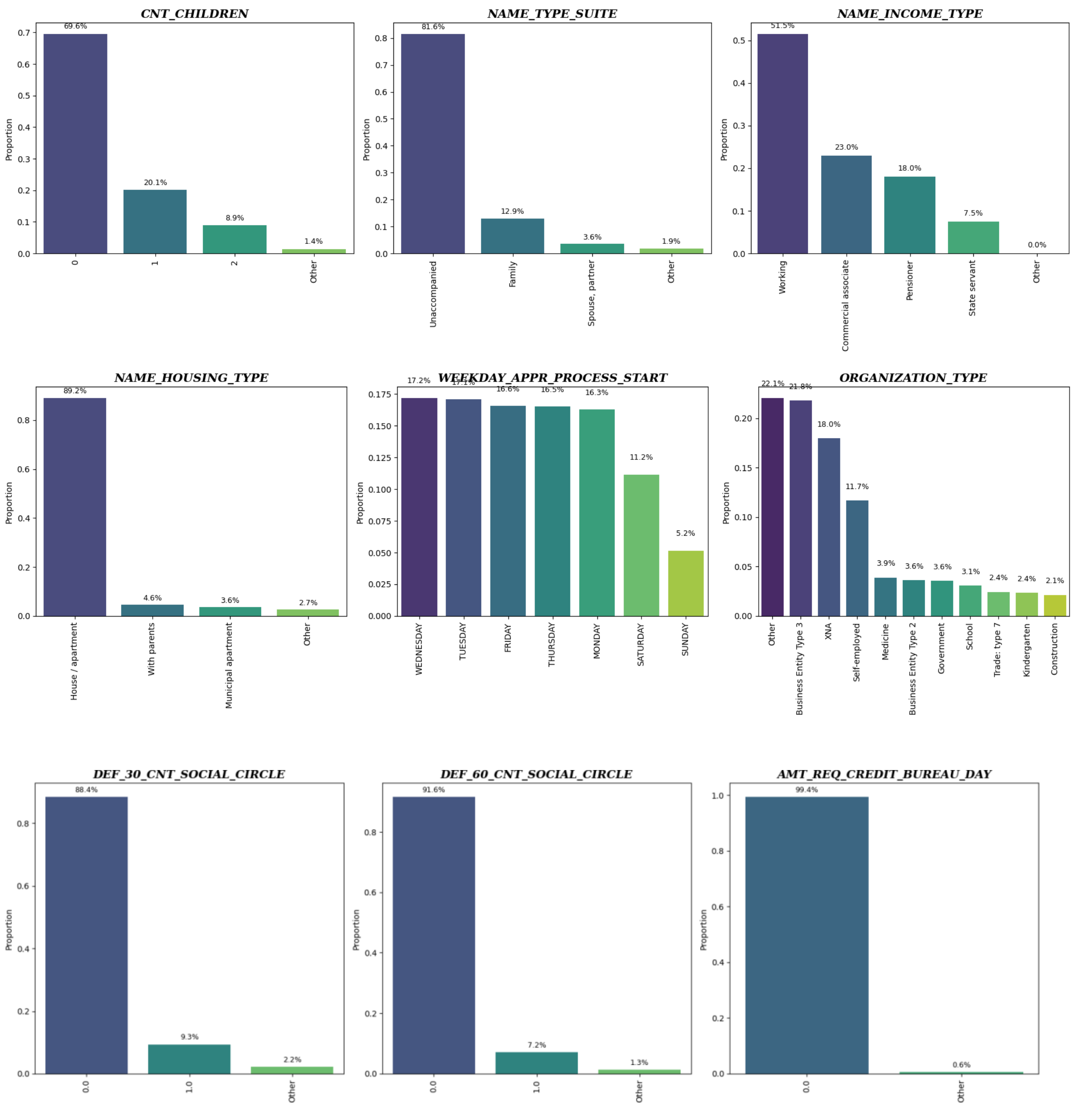


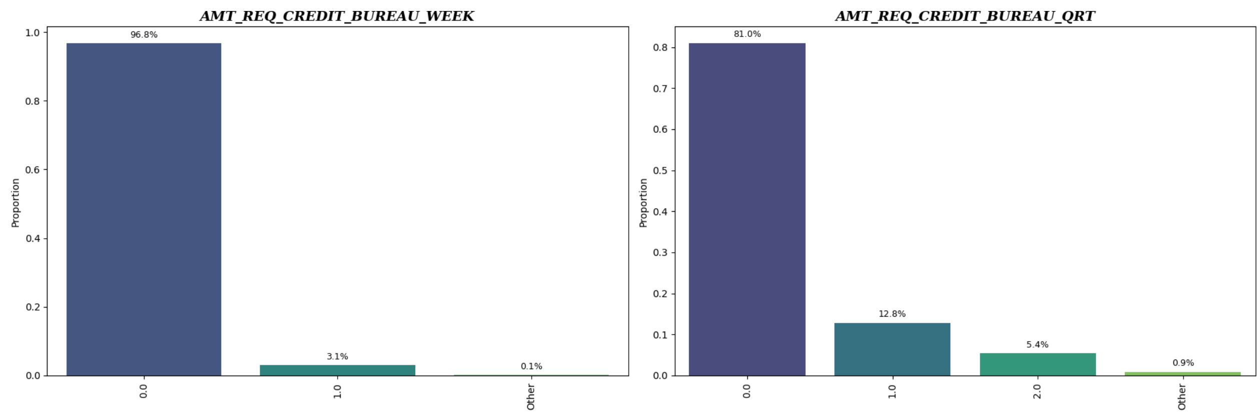




Insights:

- From all the documents 88% of clients deliver only 1.
- Nearly 45% of clients have a realty but don't have a car.

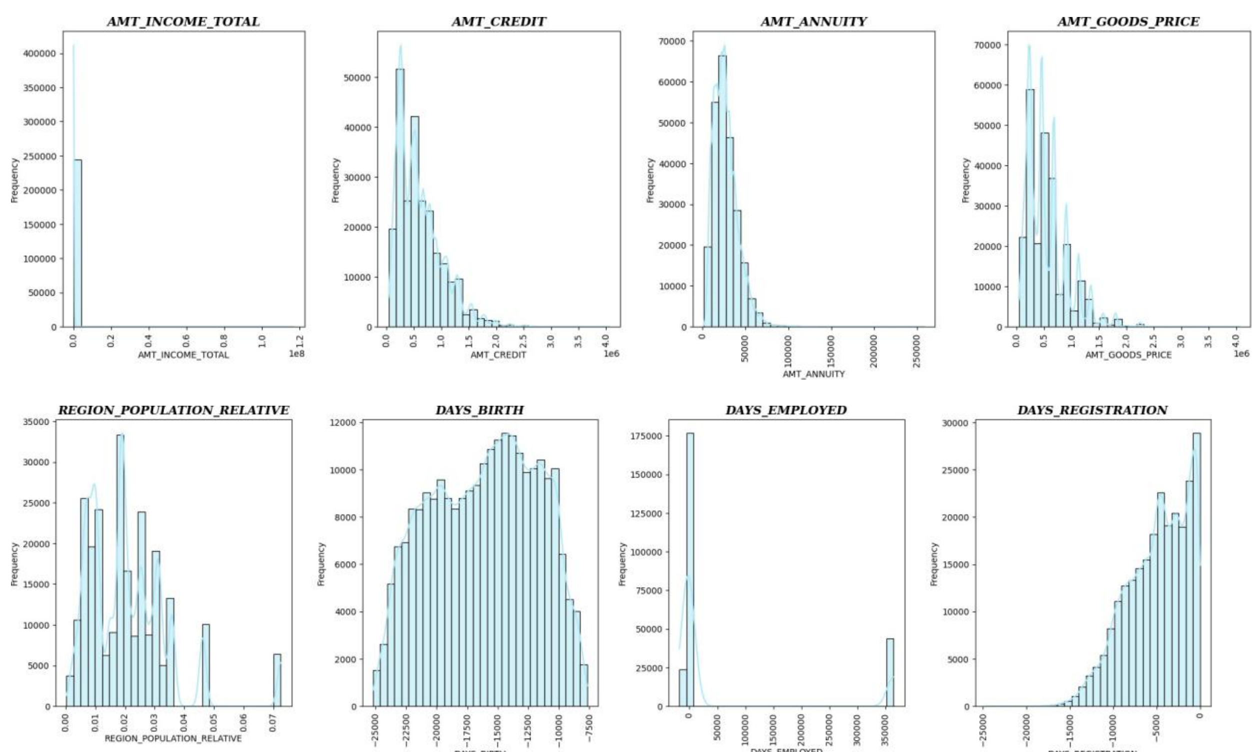

CNT_CHILDREN
Proportion
69.6%
20.1%
8.9%
1.4%
0
1
2
Other
NAME_TYPE_SUITE
Proportion
81.6%
12.9%
3.6%
1.9%
Unaccompanied
Family
Spouse, partner
Other
NAME_INCOME_TYPE
Proportion
51.5%
23.0%
18.0%
7.5%
0.0%
Working
Commercial associate
Pensioner
State servant
Other
NAME_HOUSING_TYPE
Proportion
89.2%
4.6%
3.6%
2.7%
House / apartment
With parents
Municipal apartment
Other
WEEKDAY_APPR_PROCESS_START
Proportion
17.2%
17.1%
16.6%
16.5%
16.3%
11.2%
5.2%
WEDNESDAY
TUESDAY
FRIDAY
THURSDAY
MONDAY
SATURDAY
SUNDAY
ORGANIZATION_TYPE
Proportion
22.1%
21.8%
18.0%
11.7%
3.9%
3.6%
3.6%
3.1%
2.4%
2.4%
2.1%
Other
Business Entity Type 3
XNA
Self-employed
Medicine
Business Entity Type 2
Government
School
Trade: type 7
Kindergarten
Construction
DEF_30_CNT_SOCIAL_CIRCLE
Proportion
88.4%
9.3%
2.2%
0.0
1.0
Other
DEF_60_CNT_SOCIAL_CIRCLE
Proportion
91.6%
7.2%
1.3%
0.0
1.0
Other
AMT_REQ_CREDIT_BUREAU_DAY
Proportion
99.4%
0.6%
0.0
Other

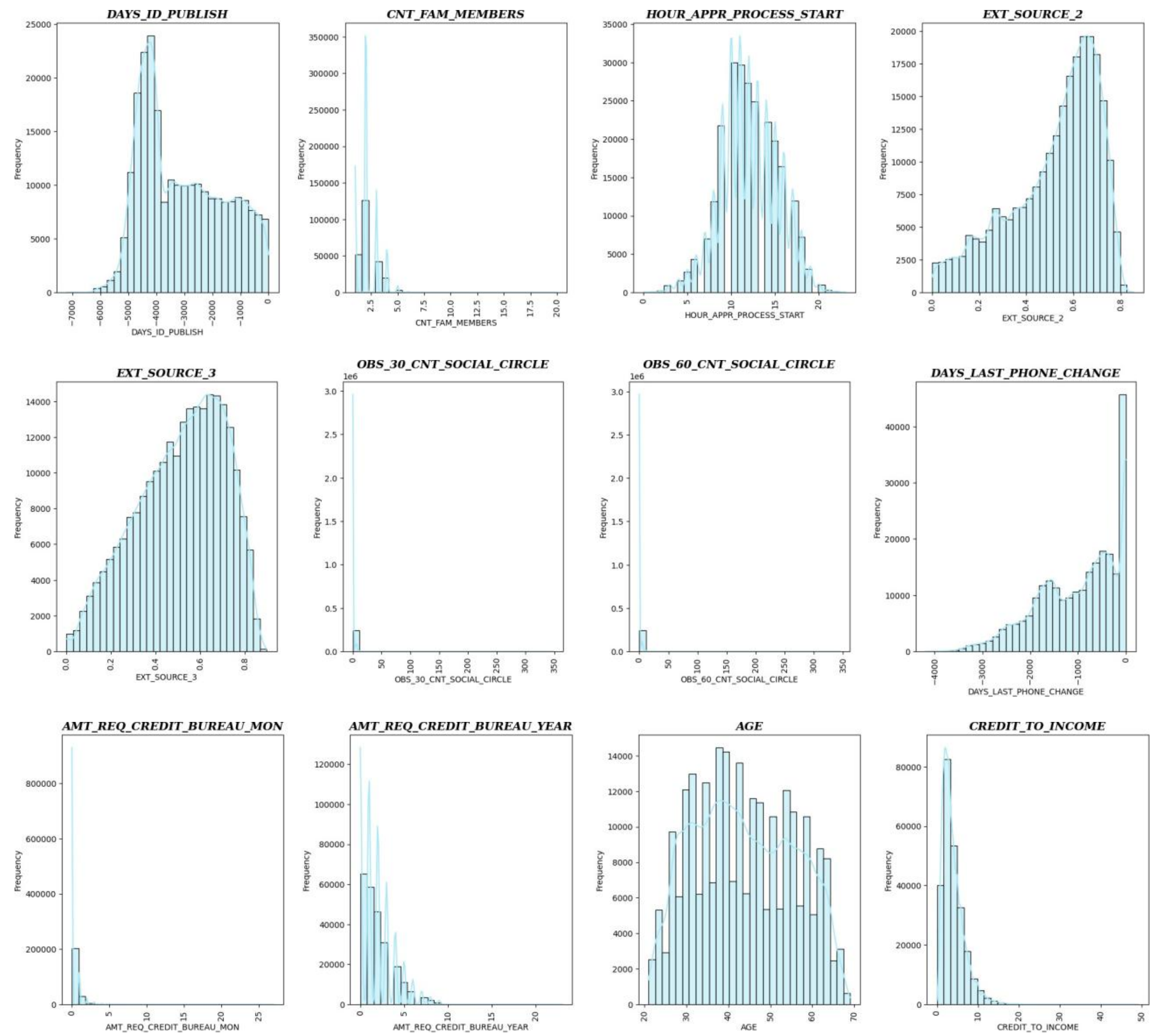


Insights:

- Most clients don't have children. (~70%)
- Most clients apply for the loan alone then with family in 2nd proportion.
- Nearly half of the clients are working in a standard job with fixed income.
- Most of clients have privately owned apartment or flat (~90%)
- ~90% of clients don't have any defaults in their social network in the last 60 days
-

Numerical Variables

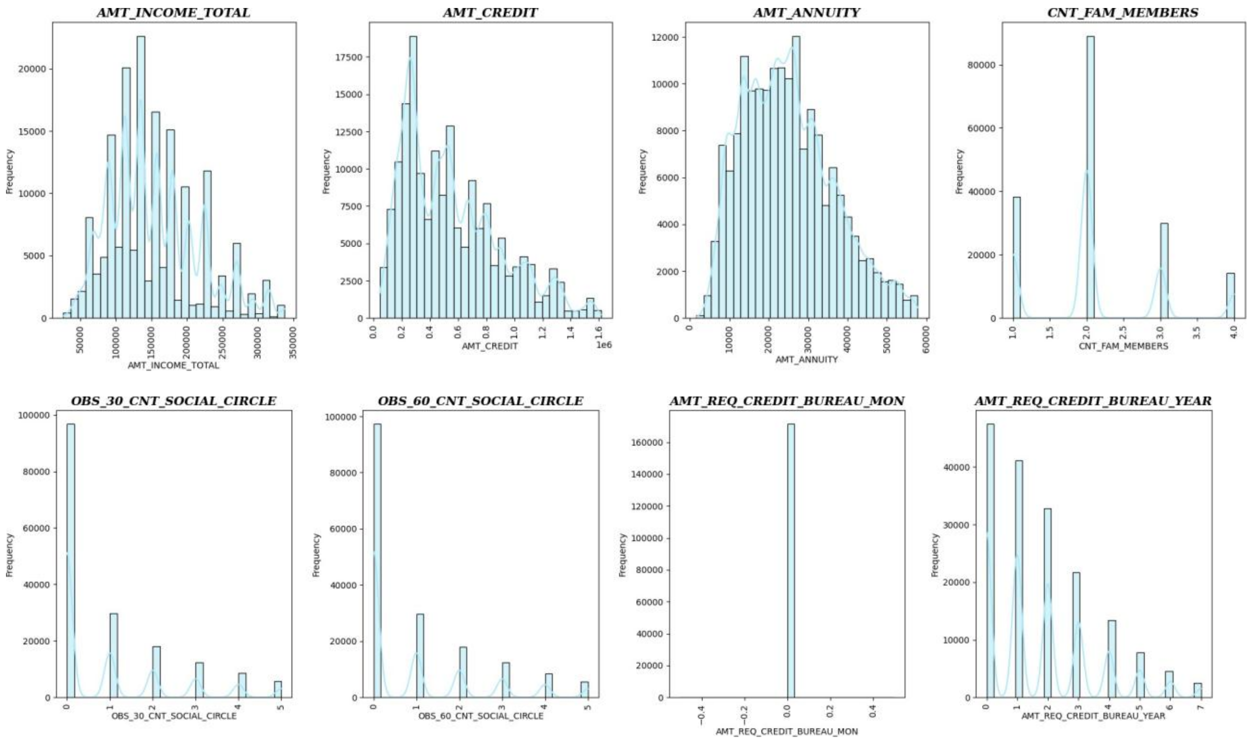




<u>Insights</u>:

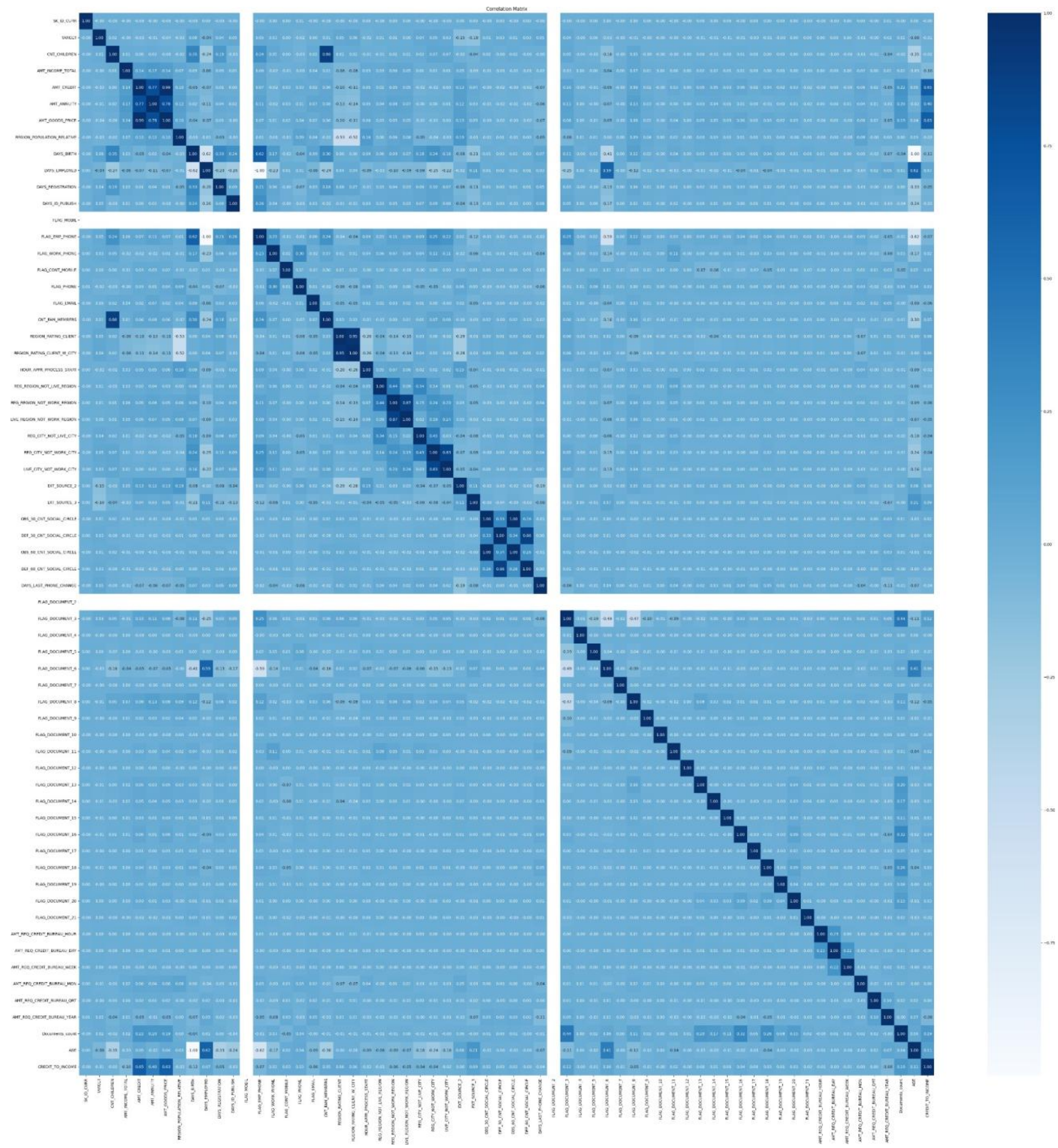
- A large portion of clients modified their phone records fewer than 125 days before applying.

As we saw that some histograms have outliers that made the histogram unreadable, we plotted again with removing outliers



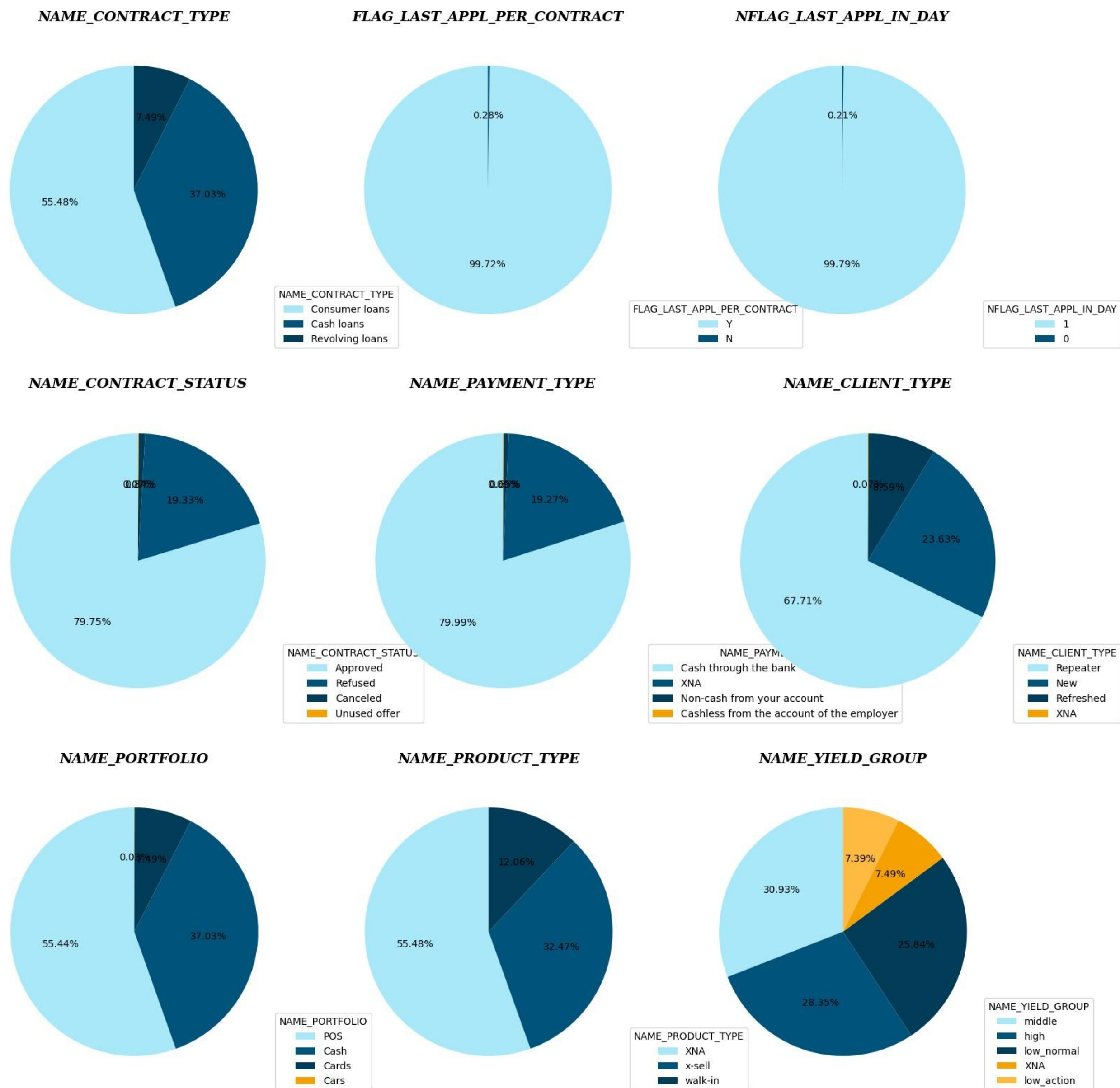
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
CNT_FAM_MEMBERS
OBS_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_YEAR
Frequency
1e6

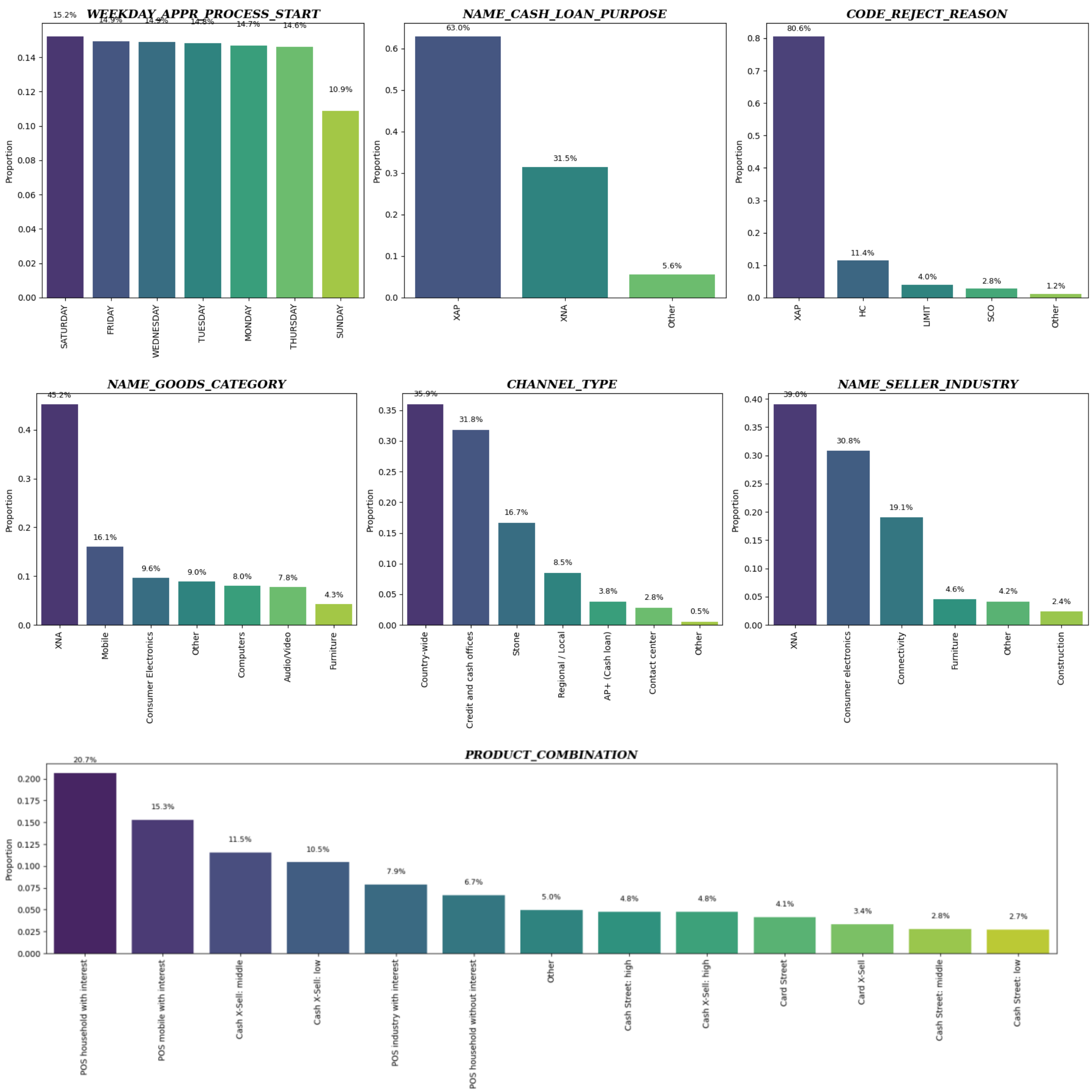
Multivariate Analysis (Applications)



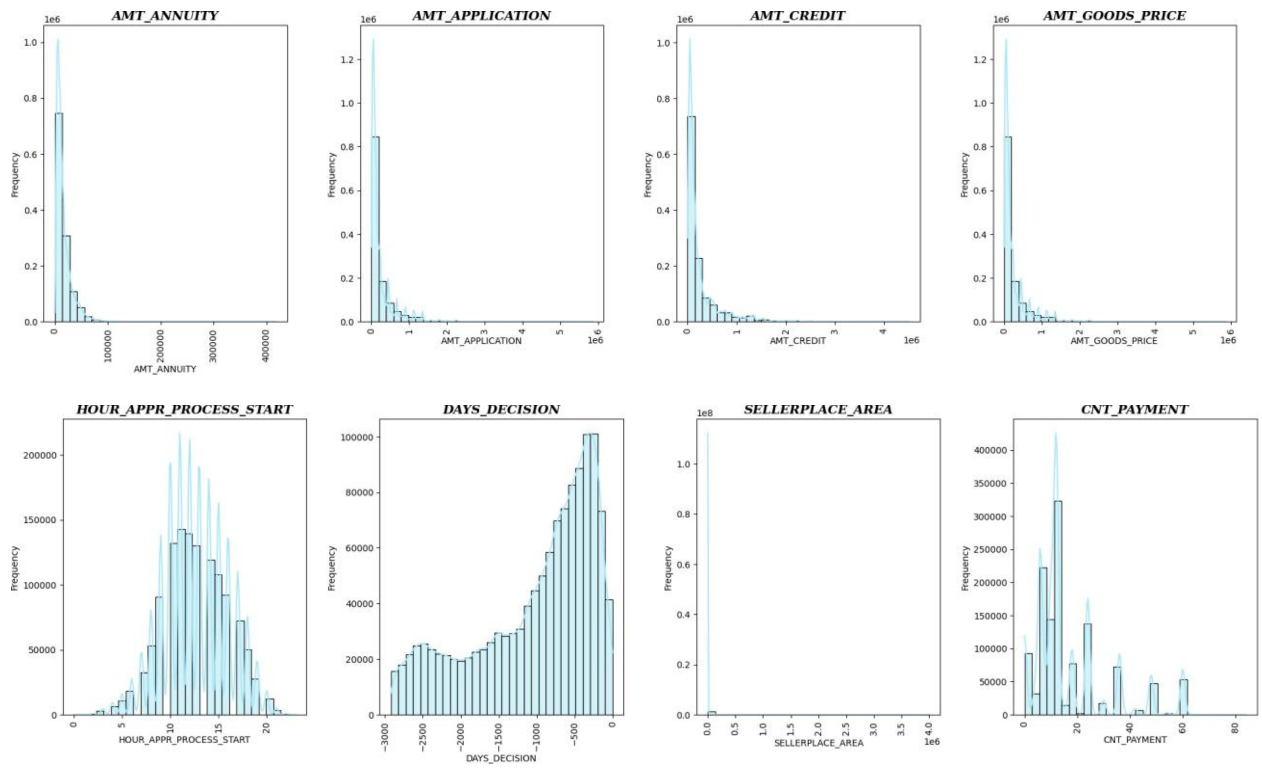
Univariate Analysis (Previous Applications)

Categorical Variables




WEEKDAY_APPR_PROCESS_START
Proportion
15.2%
14.9%
14.9%
14.8%
14.7%
14.6%
10.9%
SATURDAY
FRIDAY
WEDNESDAY
TUESDAY
MONDAY
THURSDAY
SUNDAY
NAME_CASH_LOAN_PURPOSE
63.0%
31.5%
5.6%
XAP
XNA
Other
CODE_REJECT_REASON
80.6%
11.4%
4.0%
2.8%
1.2%
XAP
HC
LIMIT
SCO
Other
NAME_GOODS_CATEGORY
45.2%
16.1%
9.6%
9.0%
8.0%
7.8%
4.3%
XNA
Mobile
Consumer Electronics
Other
Computers
Audio/Video
Furniture
CHANNEL_TYPE
35.9%
31.8%
16.7%
8.5%
3.8%
2.8%
0.5%
Country-wide
Credit and cash offices
Stone
Regional / Local
AP+ (Cash loan)
Contact center
Other
NAME_SELLER_INDUSTRY
39.0%
30.8%
19.1%
4.6%
4.2%
2.4%
XNA
Consumer electronics
Connectivity
Furniture
Other
Construction
PRODUCT_COMBINATION
20.7%
15.3%
11.5%
10.5%
7.9%
6.7%
5.0%
4.8%
4.8%
4.1%
3.4%
2.8%
2.7%
POS household with interest
POS mobile with interest
Cash X-Sell: middle
Cash X-Sell: low
POS industry with interest
POS household without interest
Other
Cash Street: high
Cash X-Sell: high
Card Street
Card X-Sell
Cash Street: middle
Cash Street: low

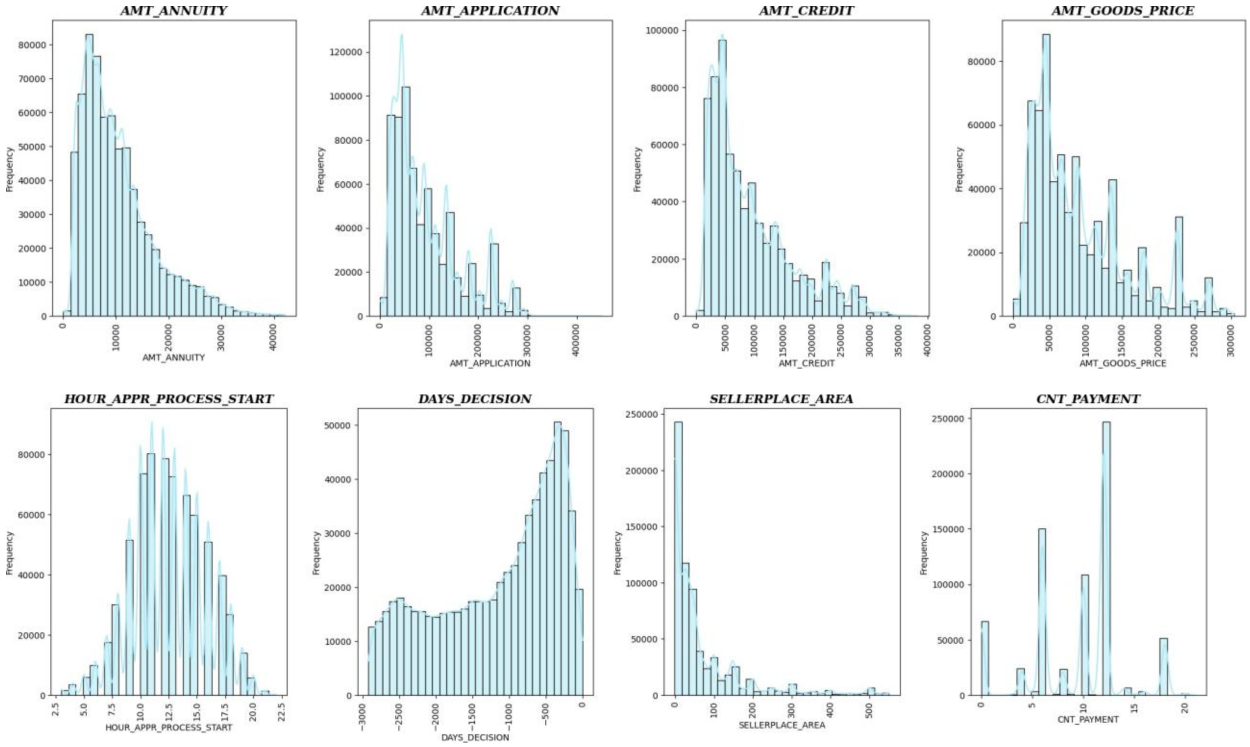
Numerical Variables



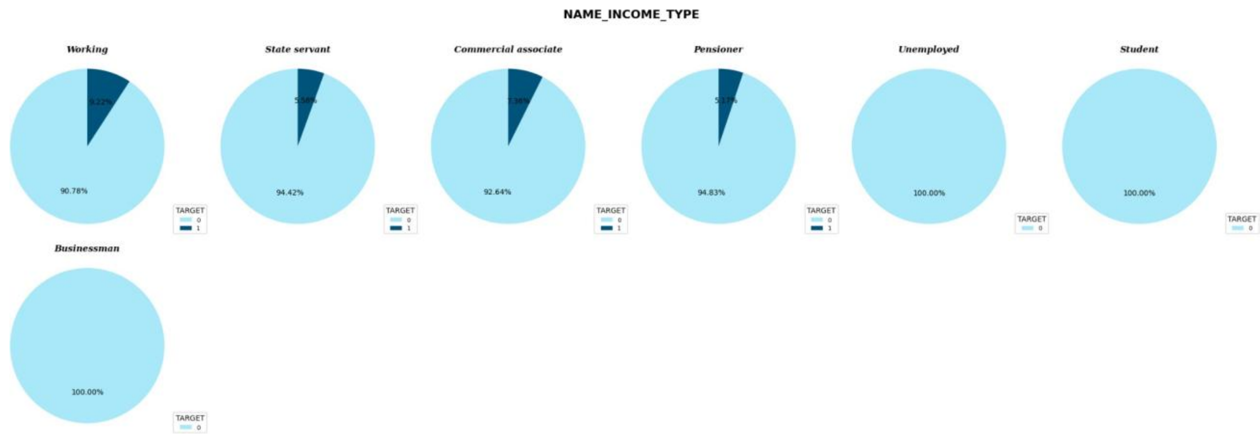
<u>Insights:</u>

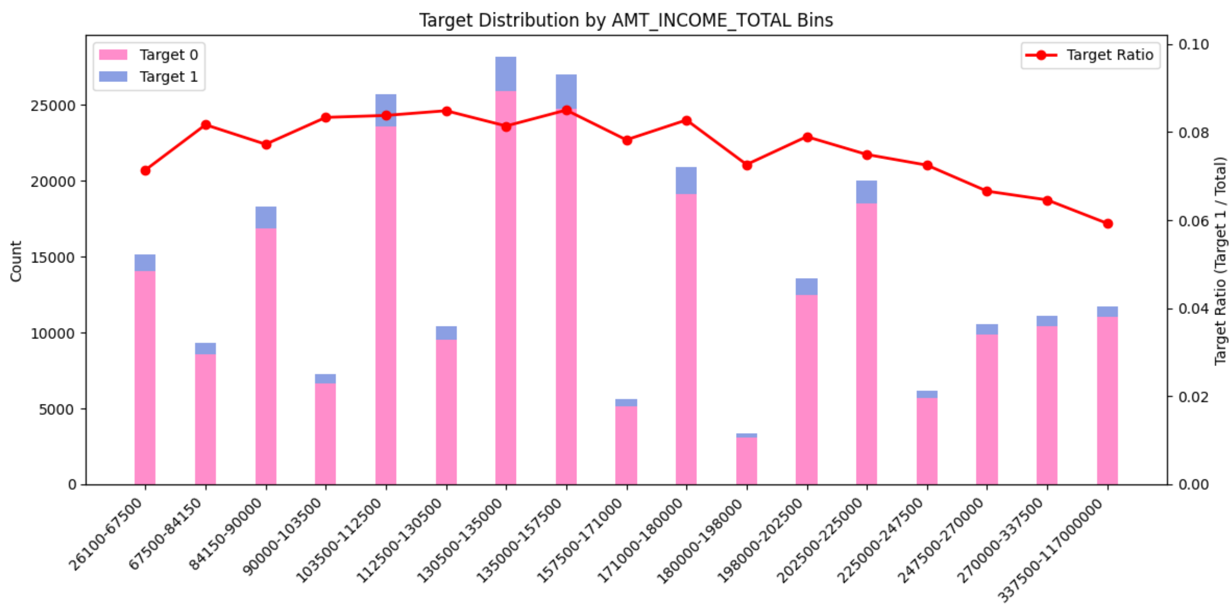
- A large portion of clients schedule loan repayment under 20 terms (ex. months).

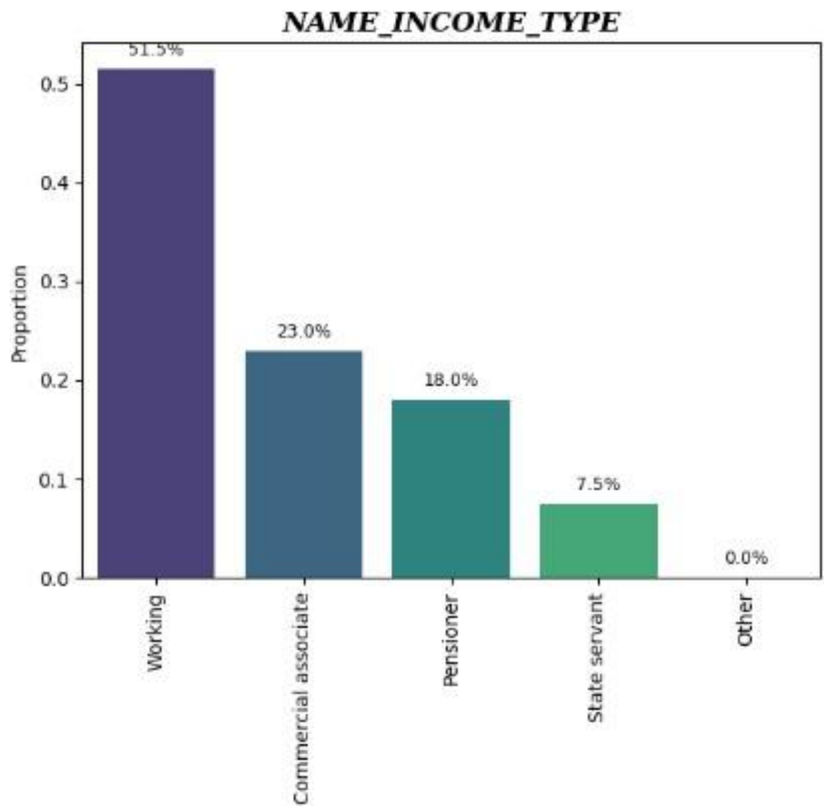
As we saw that some histograms have outliers that made the histogram unreadable, we plotted again with removing outliers

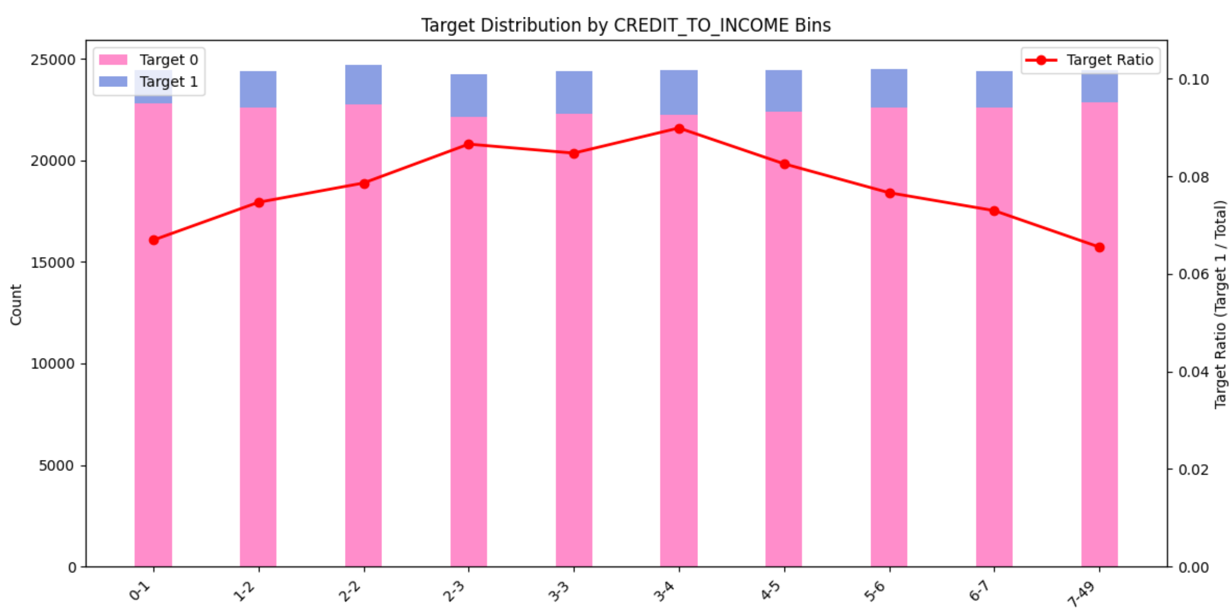

AMT_ANNUITY
Frequency
AMT_ANNUITY
AMT_APPLICATION
Frequency
AMT_APPLICATION
AMT_CREDIT
Frequency
AMT_CREDIT
AMT_GOODS_PRICE
Frequency
AMT_GOODS_PRICE
HOUR_APPR_PROCESS_START
Frequency
HOUR_APPR_PROCESS_START
DAYS_DECISION
Frequency
DAYS_DECISION
SELLERPLACE_AREA
Frequency
SELLERPLACE_AREA
CNT_PAYMENT
Frequency
CNT_PAYMENT

Bivariate Analysis (Applications)



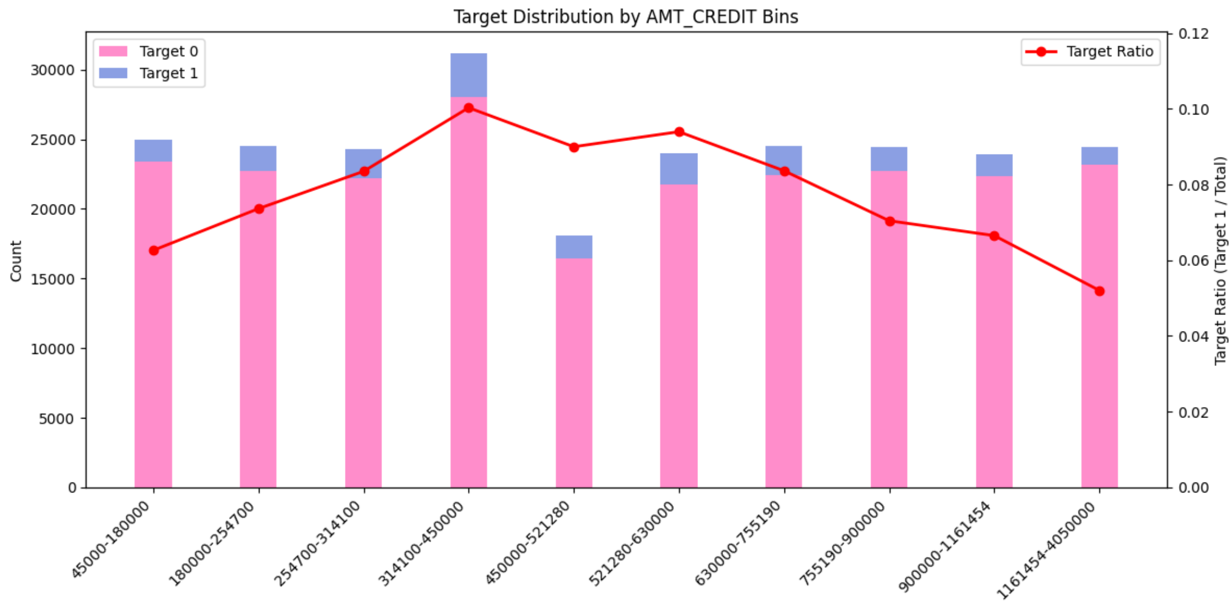






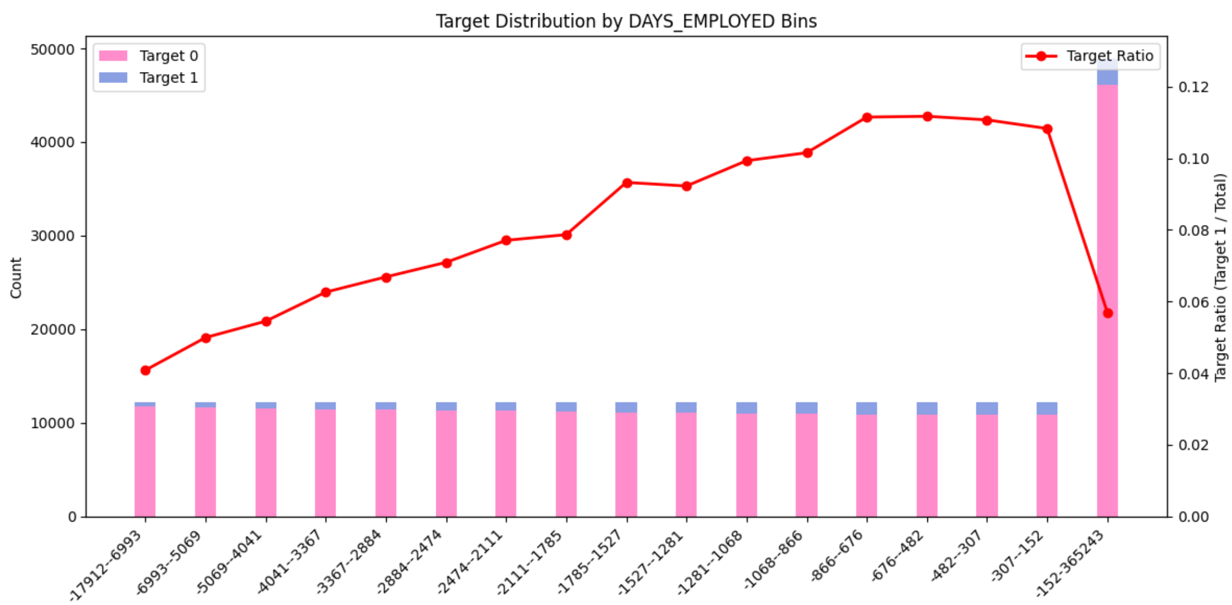
<u>Insights:</u>

- From the working and commercial associate clients there is (7% to 10%) who face defaults.
- As total income increase, default rates decreases
- Clients that ask for credit 2-5 x their income have a higher default rate.



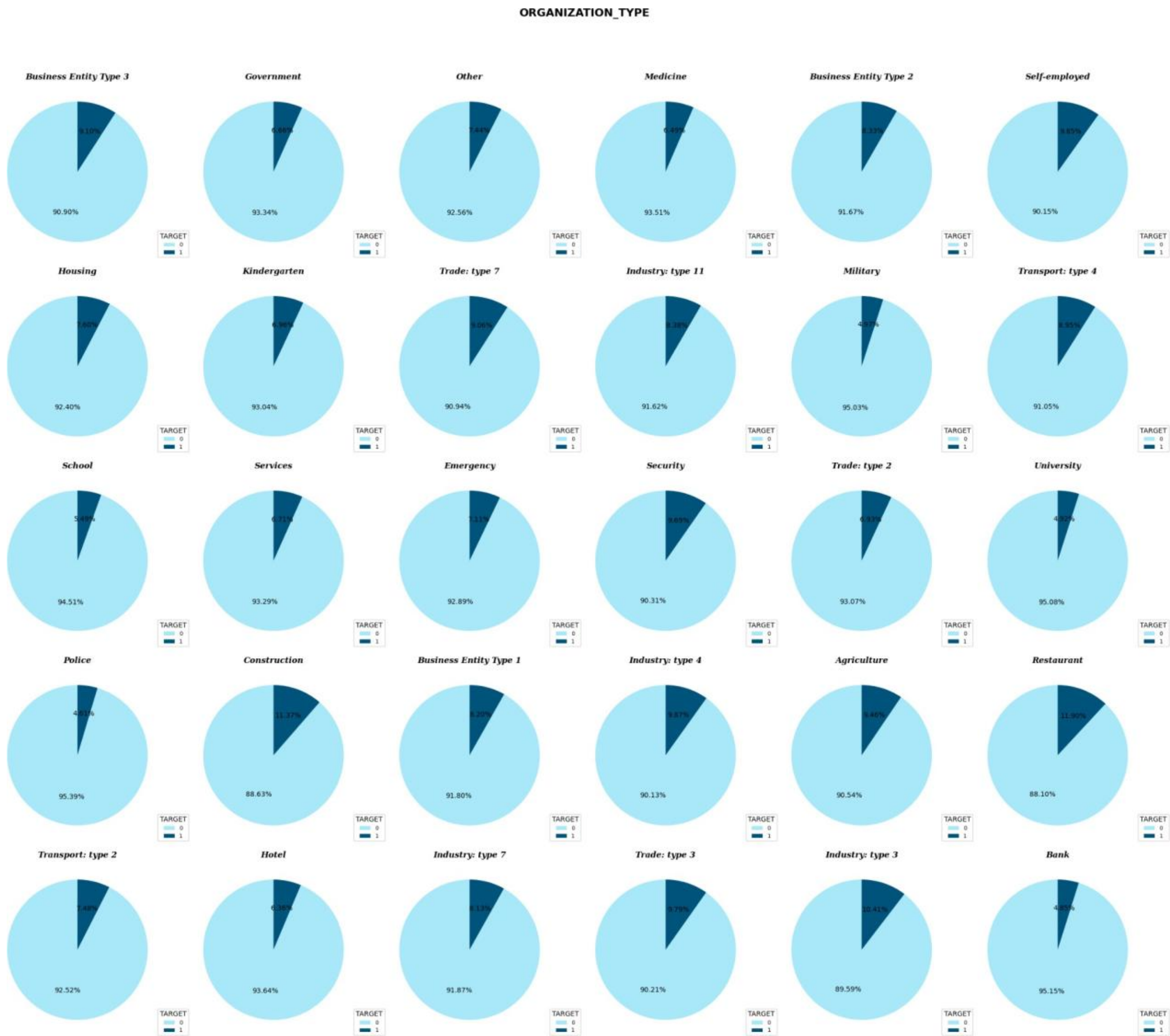
Insights:

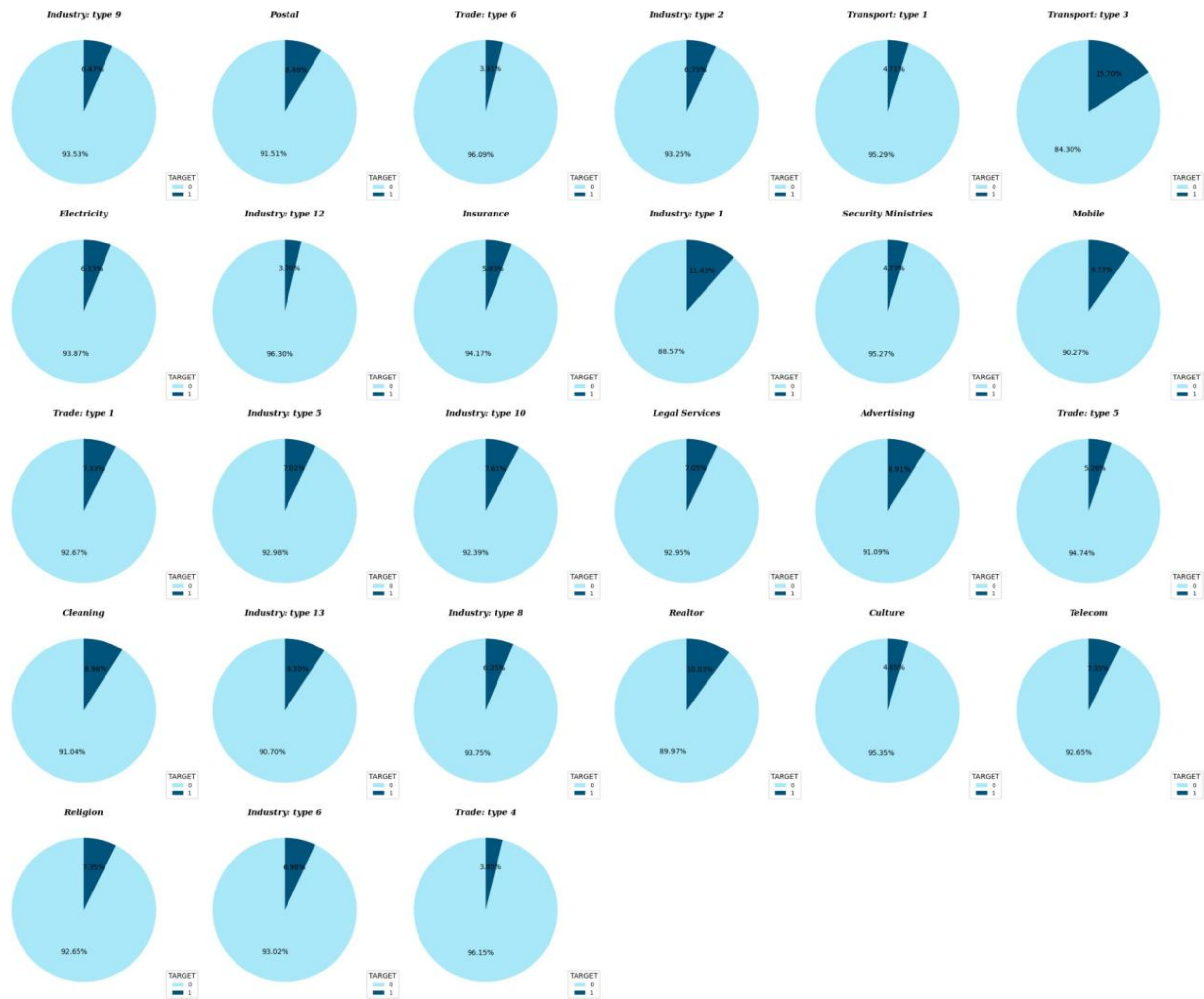
- The risk of default seems to peak in the middle ranges of loan amounts. Specifically, the highest default ratio appears to be around the 314,100-450,000 range, where the red line reaches its maximum at approximately 0.10 (10% default rate).



Insights:

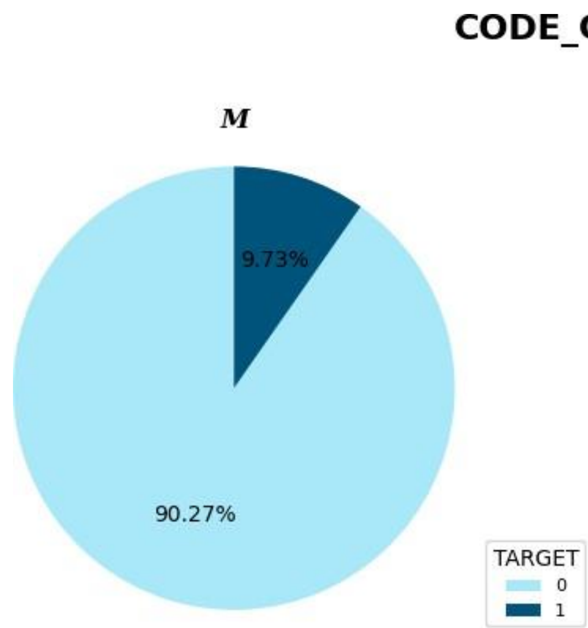
- Default risk increases as employment duration decreases. The highest default rates (~11%) appear in the 676-482, 482-307, and 307-152 bins, representing people with relatively short employment durations. (for the last bin, it likely contains anomalous values)

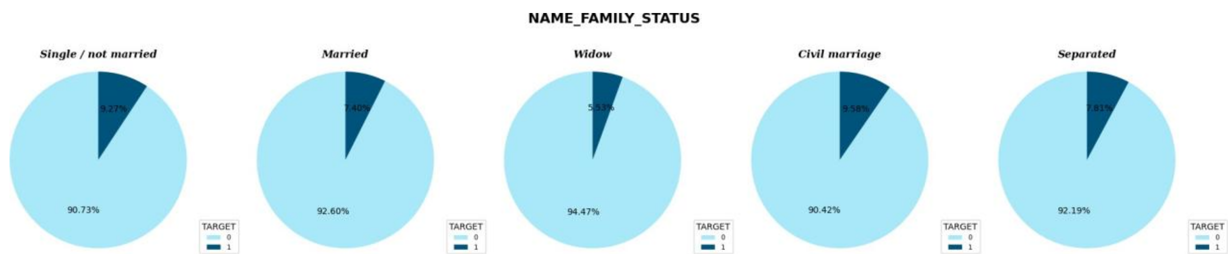

ORGANIZATION_TYPE
Business Entity Type 3
90.90%
Government
93.34%
Other
92.56%
Medicine
93.51%
Business Entity Type 2
91.67%
Self-employed
90.15%
Housing
92.40%
Kindergarten
93.04%
Trade: type 7
90.94%
Industry: type 11
91.62%
Military
95.03%
Transport: type 4
91.05%
School
94.51%
Services
93.29%
Emergency
92.89%
Security
90.31%
Trade: type 2
93.07%
University
95.08%
Police
95.39%
Construction
88.63%
Business Entity Type 1
91.80%
Industry: type 4
90.13%
Agriculture
90.54%
Restaurant
88.10%
Transport: type 2
92.52%
Hotel
93.64%
Industry: type 7
91.87%
Trade: type 3
90.21%
Industry: type 3
89.59%
Bank
95.15%
TARGET
0
1

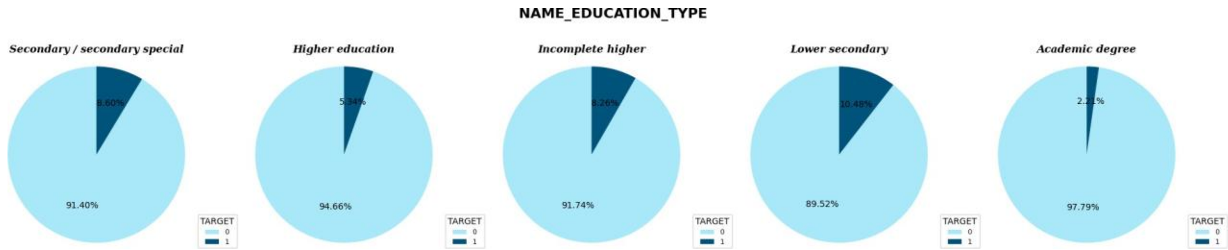


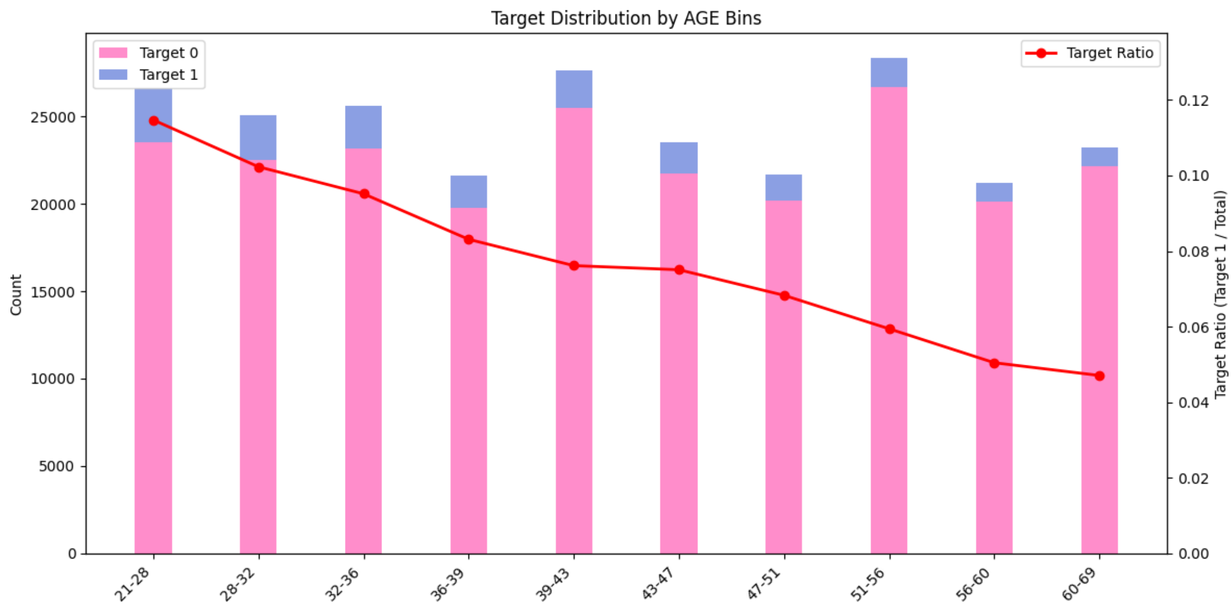
Insights:

- Borrowers employed in the Mobile (16.10%) and Hotel (13.65%) sectors show the highest default rates, while Government, Military, and Banking sectors maintain lower, more stable default rates (~9%).


CODE_GENDER
M
9.73%
90.27%
TARGET
0
1
F
6.81%
93.19%
TARGET
0
1

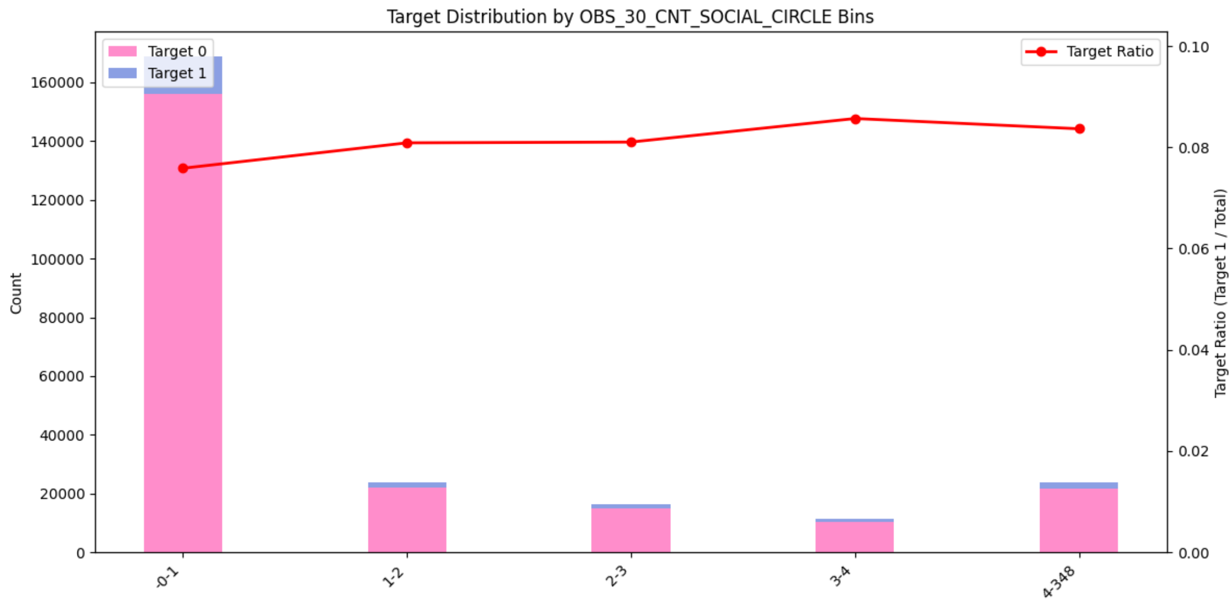

NAME_FAMILY_STATUS
Single / not married
9.27%
90.73%
Married
7.40%
92.60%
Widow
5.53%
94.47%
Civil marriage
9.58%
90.42%
Separated
7.81%
92.19%
TARGET
0
1

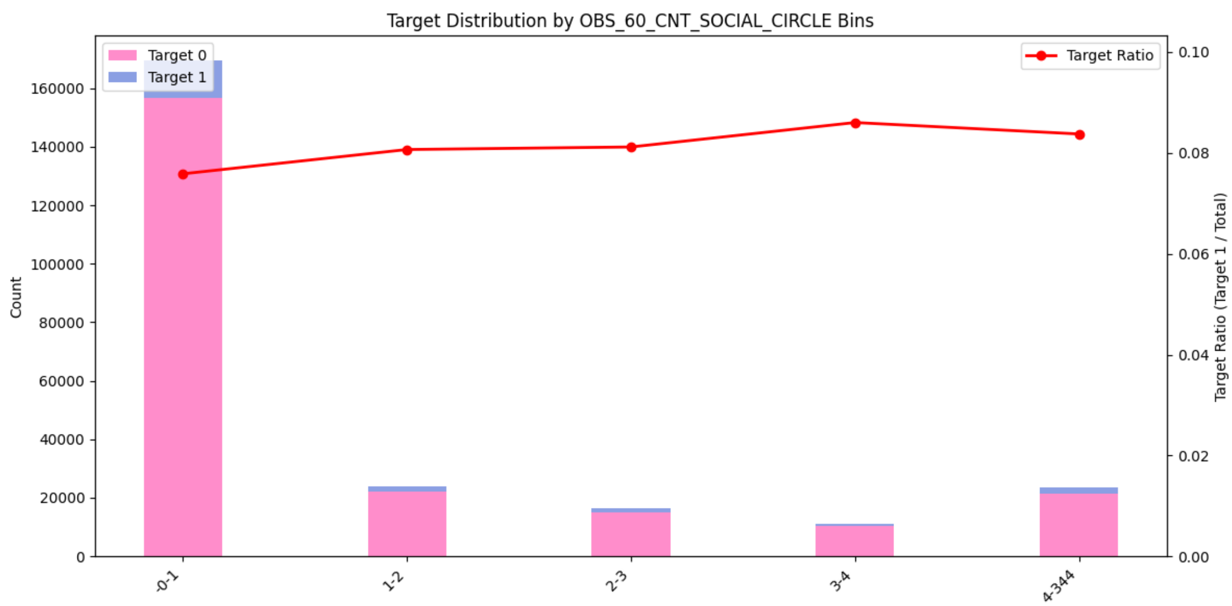




Insights:

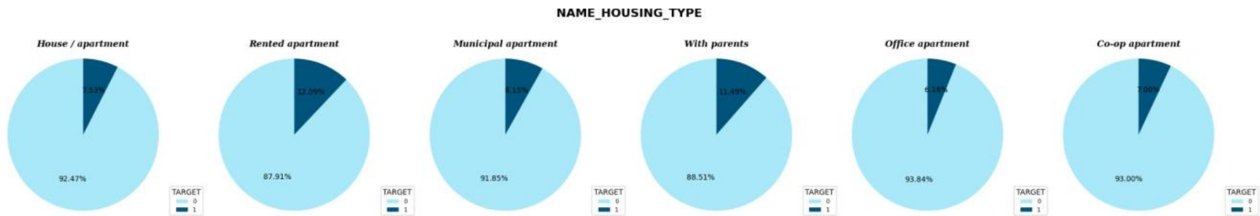
- Borrowers who are men (9.73%) show higher default rates, compared to women (~6.81%).
- Borrowers who are widowed show the lowest default rate (5.53%). Meanwhile, those in a civil marriage (9.58%) and single/not married (9.27%) categories have the highest default rates.
- There is an inverse relationship between education level and default risk. The lower the education level, the higher the chance of default.
- Younger clients have higher default rates.





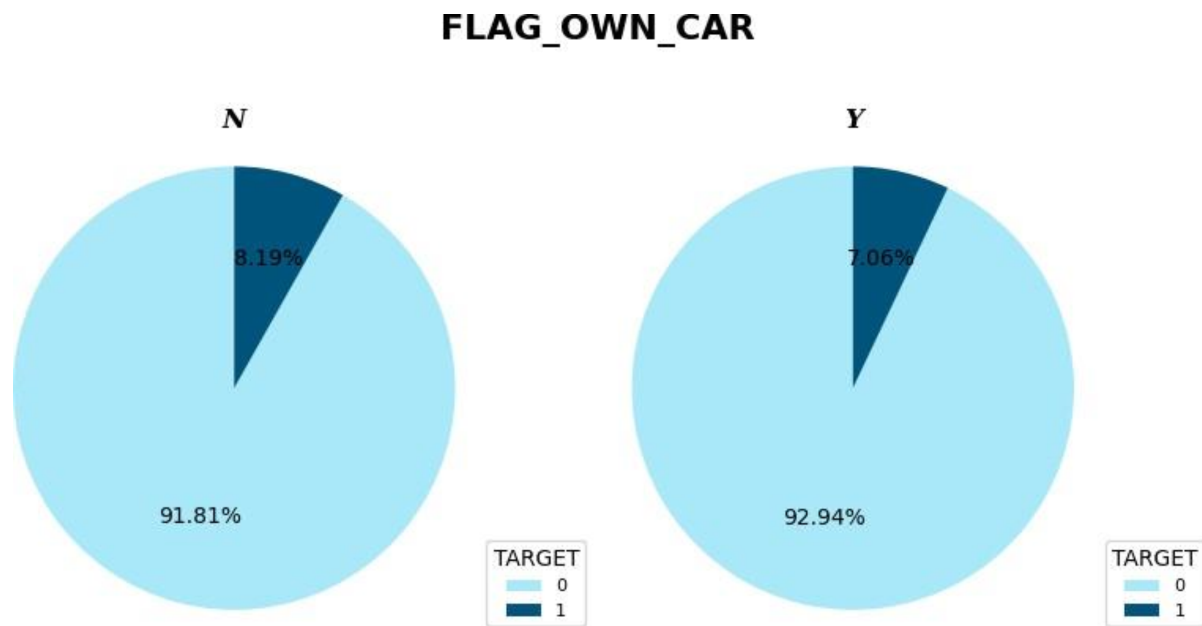
Insights:

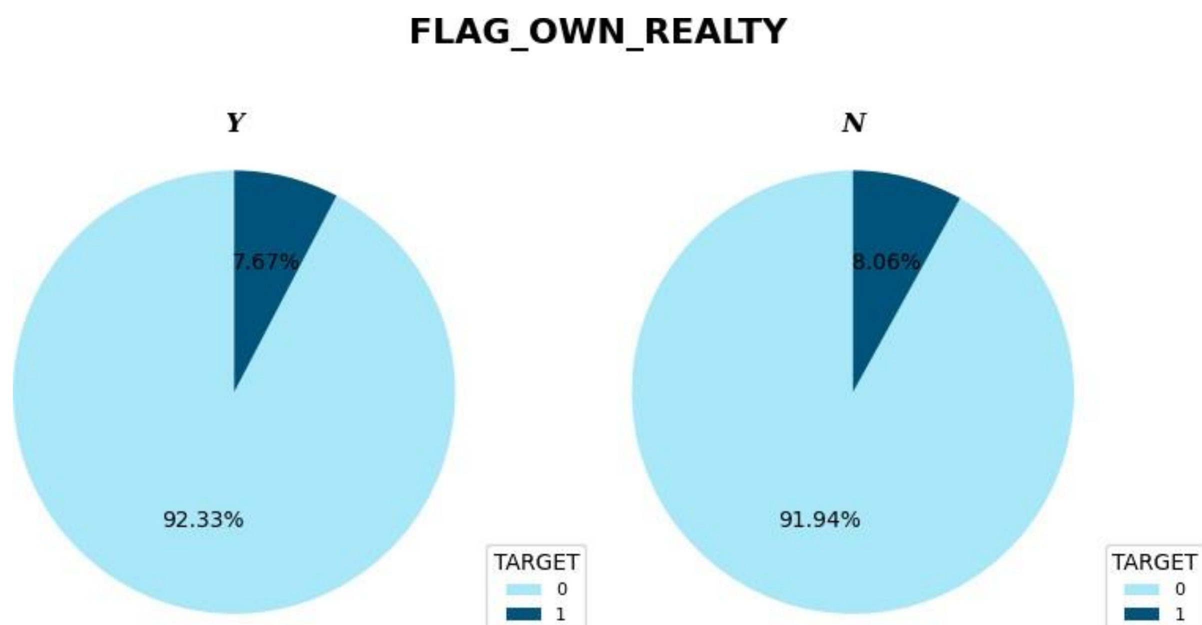
- Most clients (~85%) have 0-1 default in their social circle.
- As the number of acquaintances with defaults increases, the client's default rate increases slightly. The effect is small, but positive.

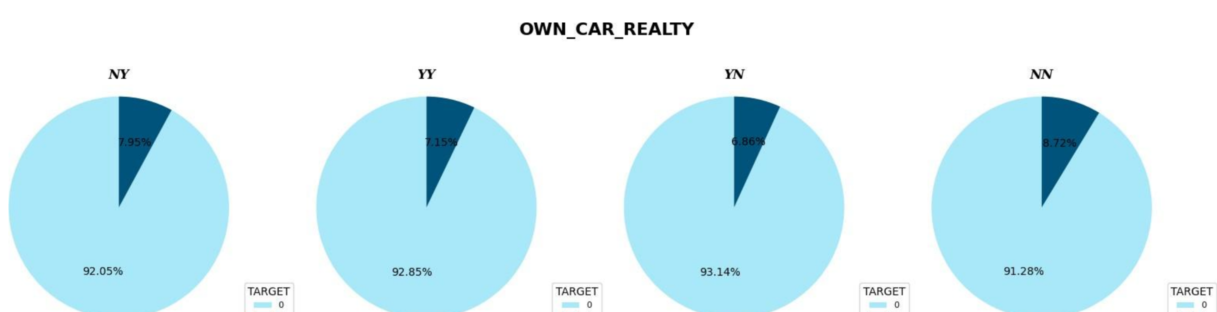


Insights:

- Ownership (house, co-op) is a positive indicator for financial stability.
- Renting or living with parents indicates a higher risk, possibly due to less financial independence or lower income levels.

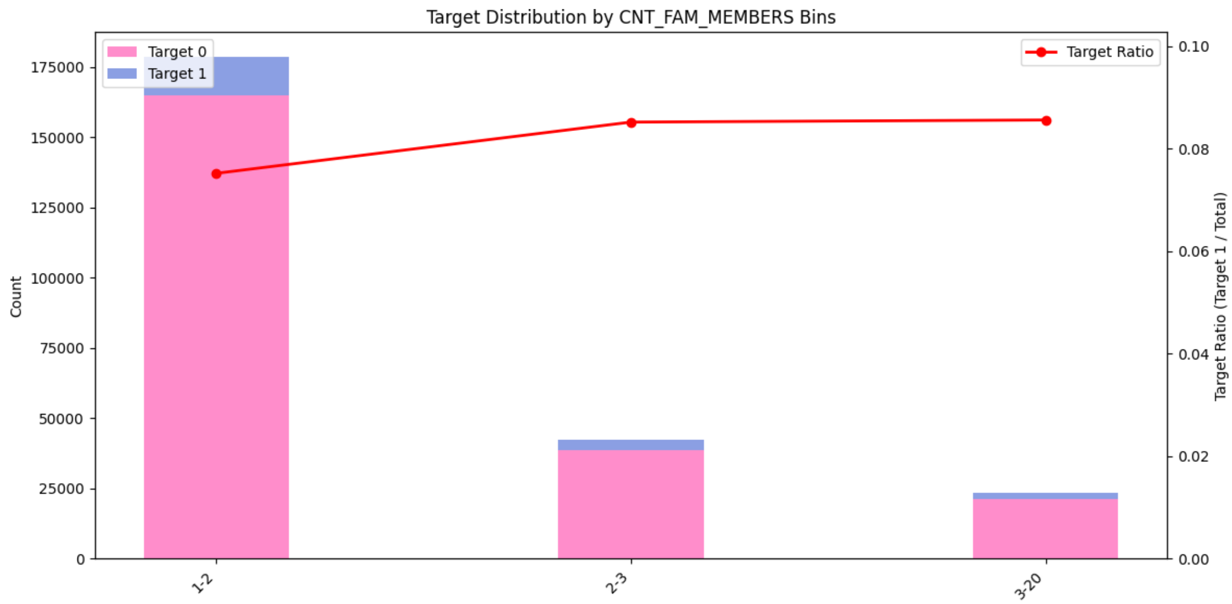

FLAG_OWN_CAR
N
8.19%
91.81%
TARGET
0
1
Y
7.06%
92.94%
TARGET
0
1


FLAG_OWN_REALTY
Y
7.67%
92.33%
TARGET
0
1
N
8.06%
91.94%
TARGET
0
1


OWN_CAR_REALTY
NY
7.95%
92.05%
TARGET
0
1
YY
7.15%
92.85%
TARGET
0
1
YN
6.86%
93.14%
TARGET
0
1
NN
8.72%
91.28%
TARGET
0
1

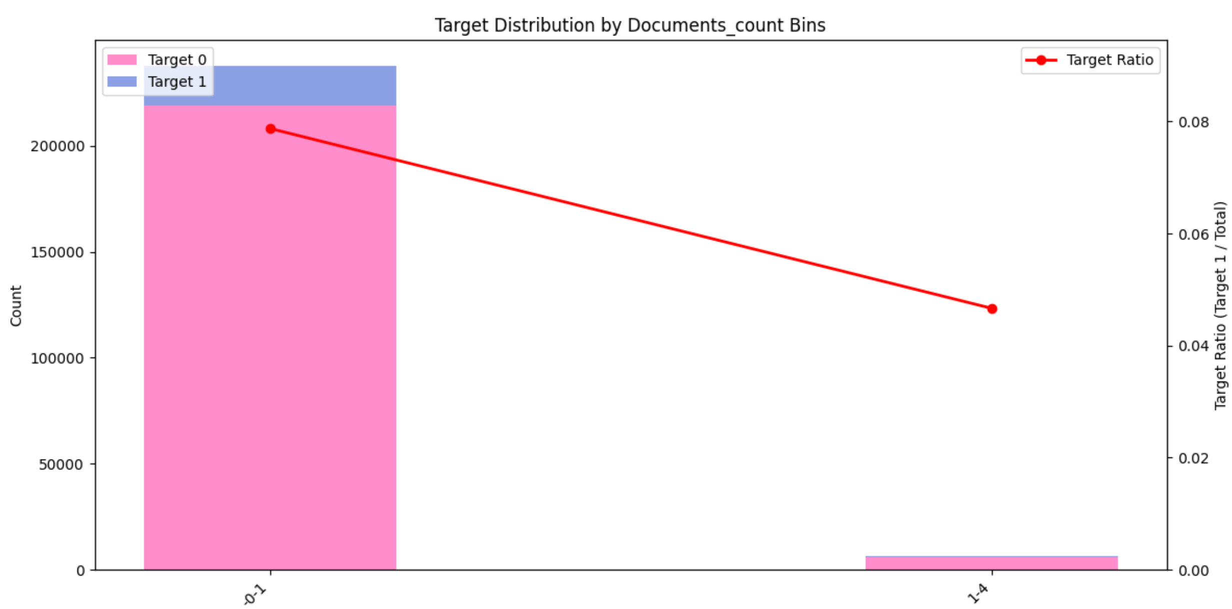
Insights:

- Clients who doesn't have a car, or realty have higher defaults.



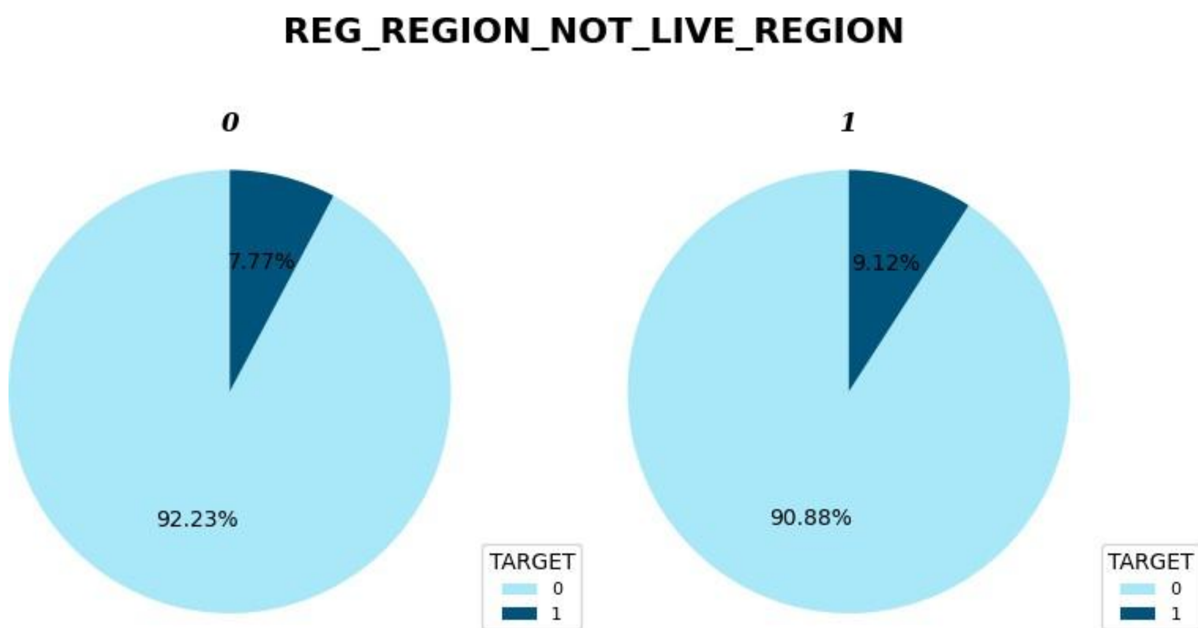
Insights:

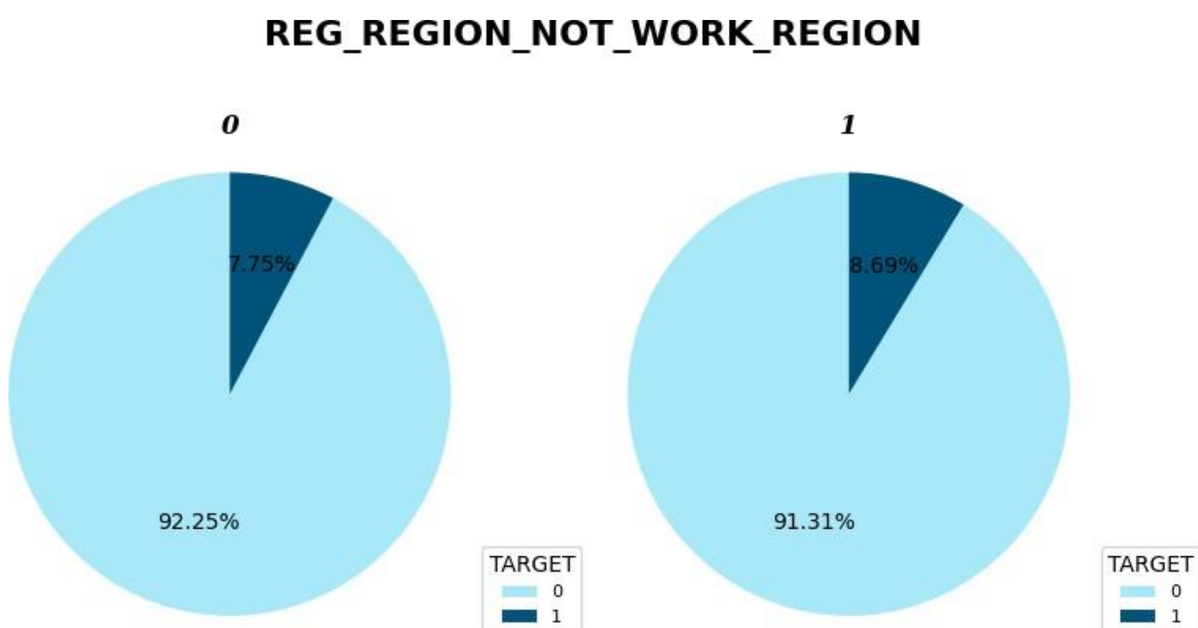
- As the number of family members increases, the client's default rate increases slightly. The effect is small, but positive.

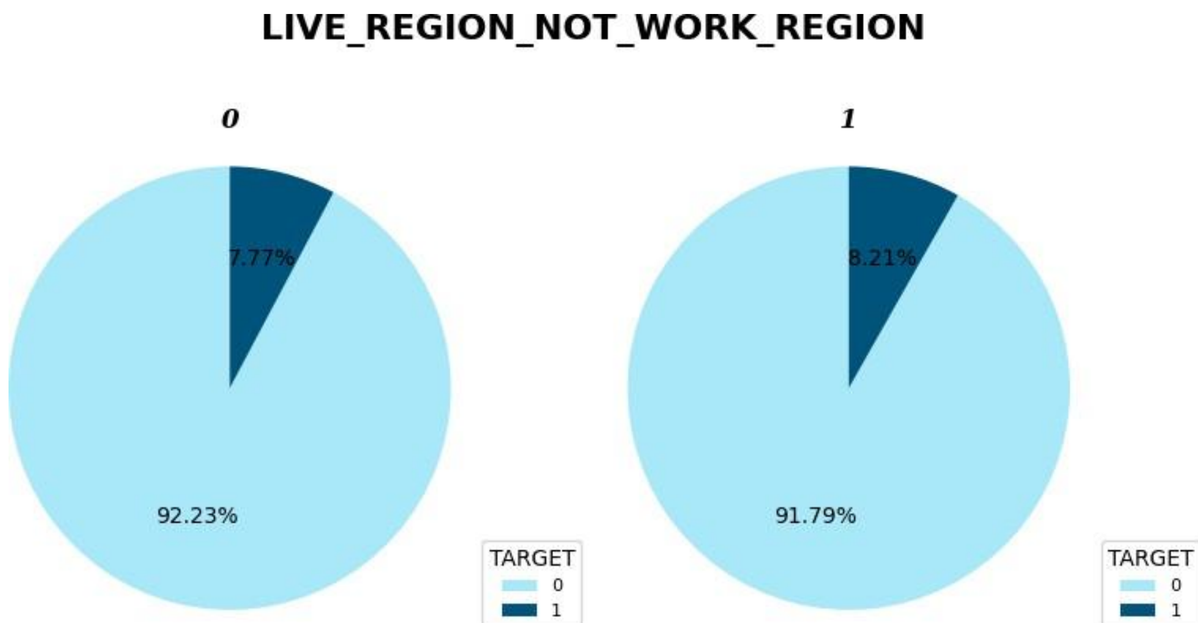


Insights:

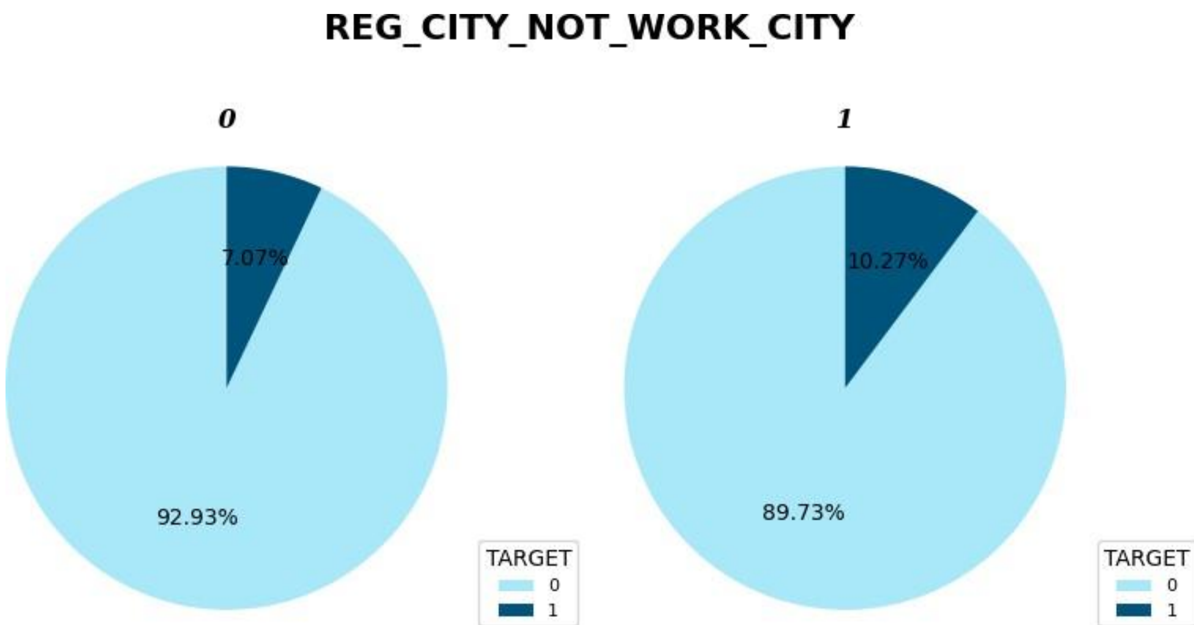
- Most clients submit only one or no documents. Submitting more documents is rare but slightly reduces the risk of default.

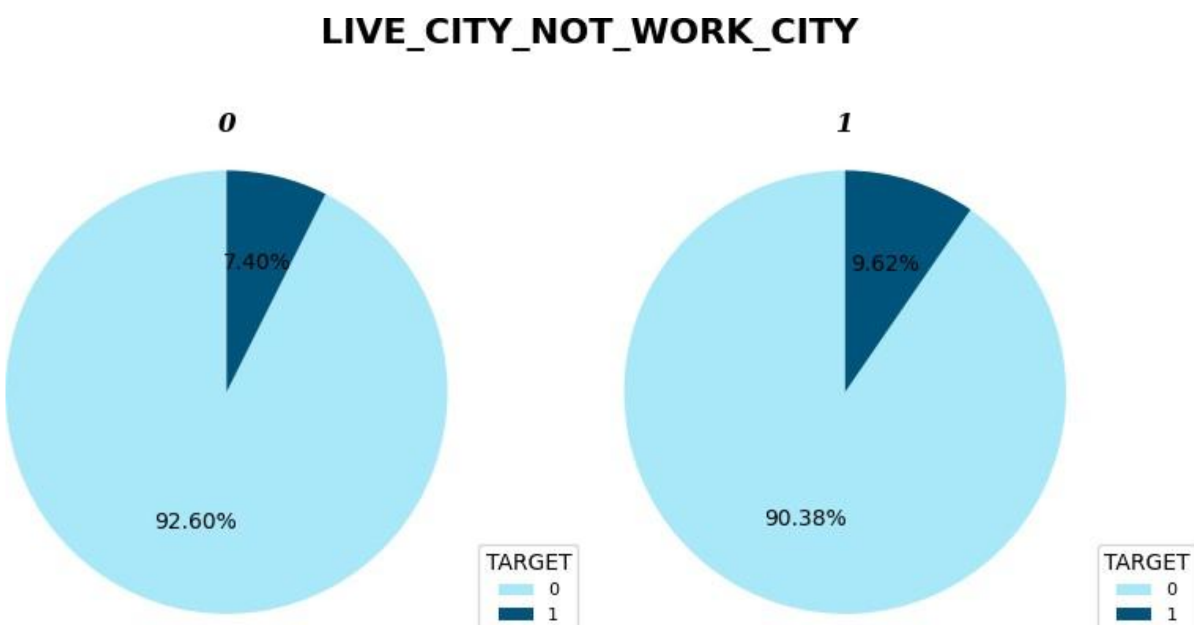





LIVE_REGION_NOT_WORK_REGION
0
7.77%
92.23%
TARGET
0
1
1
8.21%
91.79%
TARGET
0
1

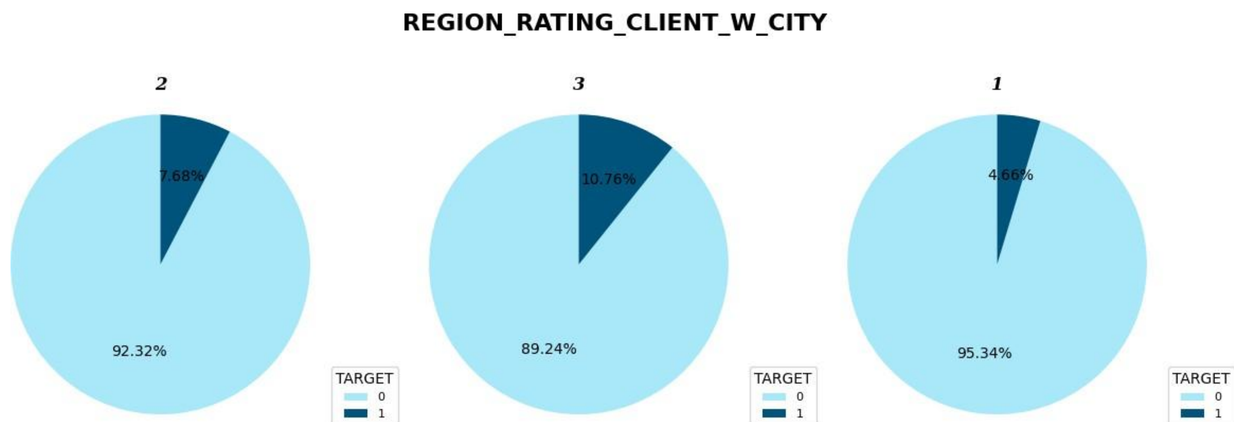

REG_CITY_NOT_LIVE_CITY
0
7.46%
92.54%
TARGET
0
1
1
11.93%
88.07%
TARGET
0
1

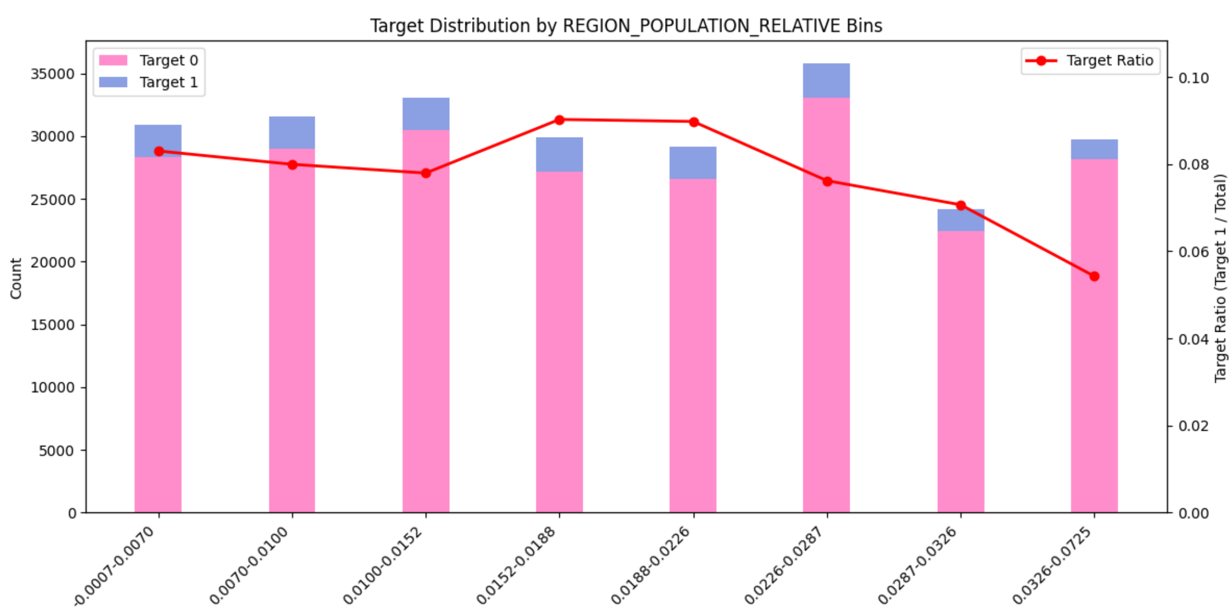




Insights:

- Clients’ whose work region is not the same as live region have more risk of default.

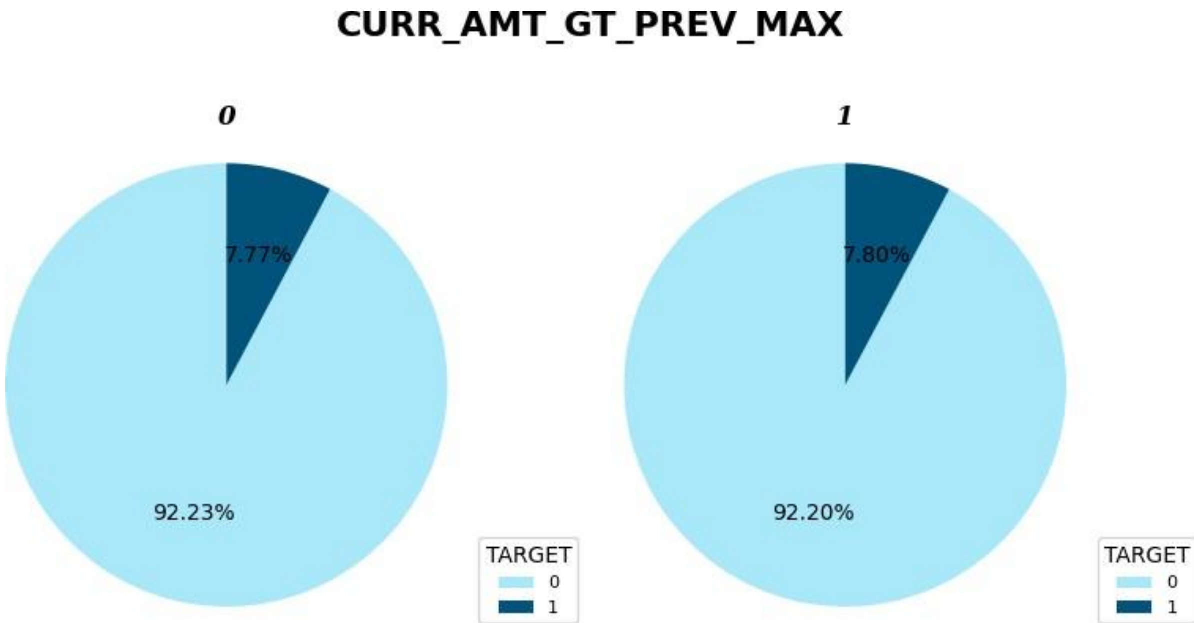


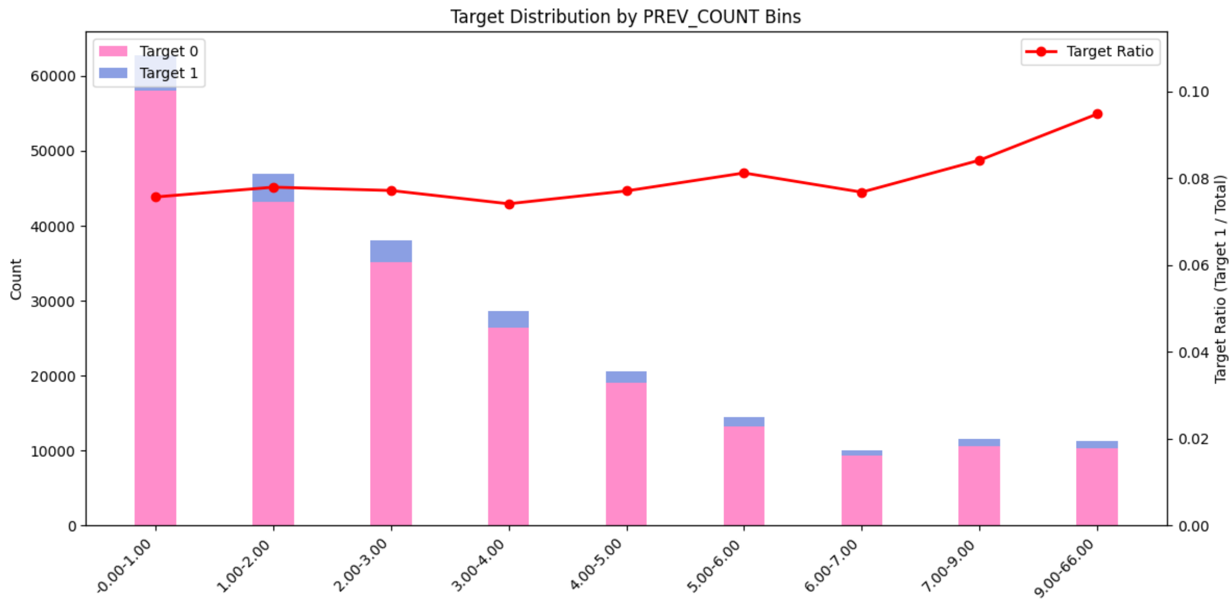


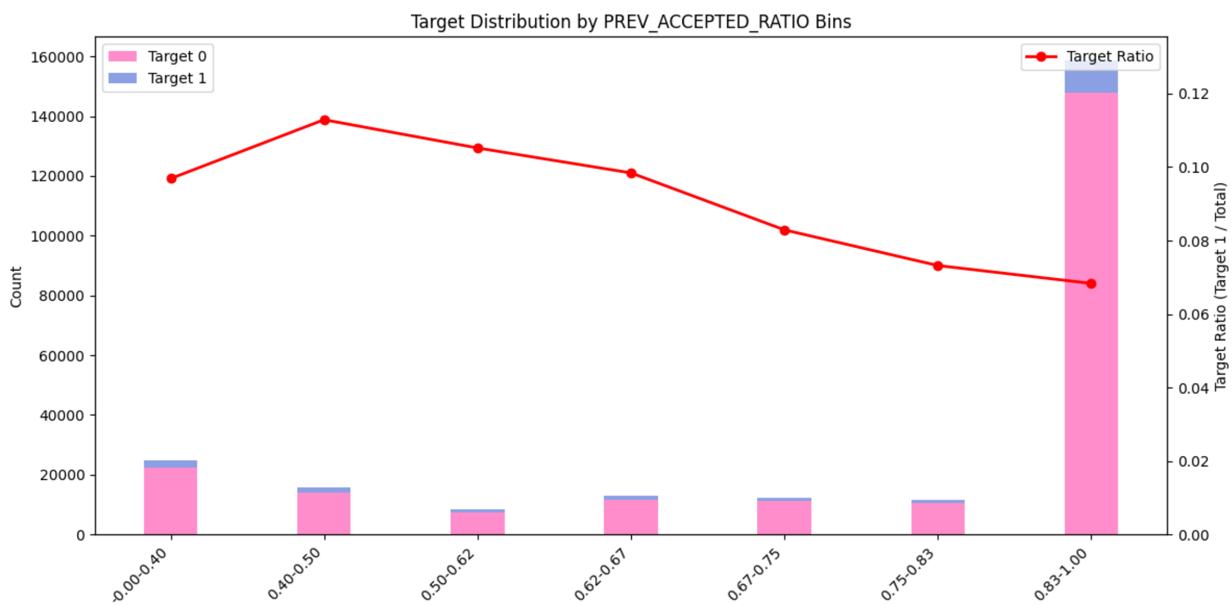
Insights:

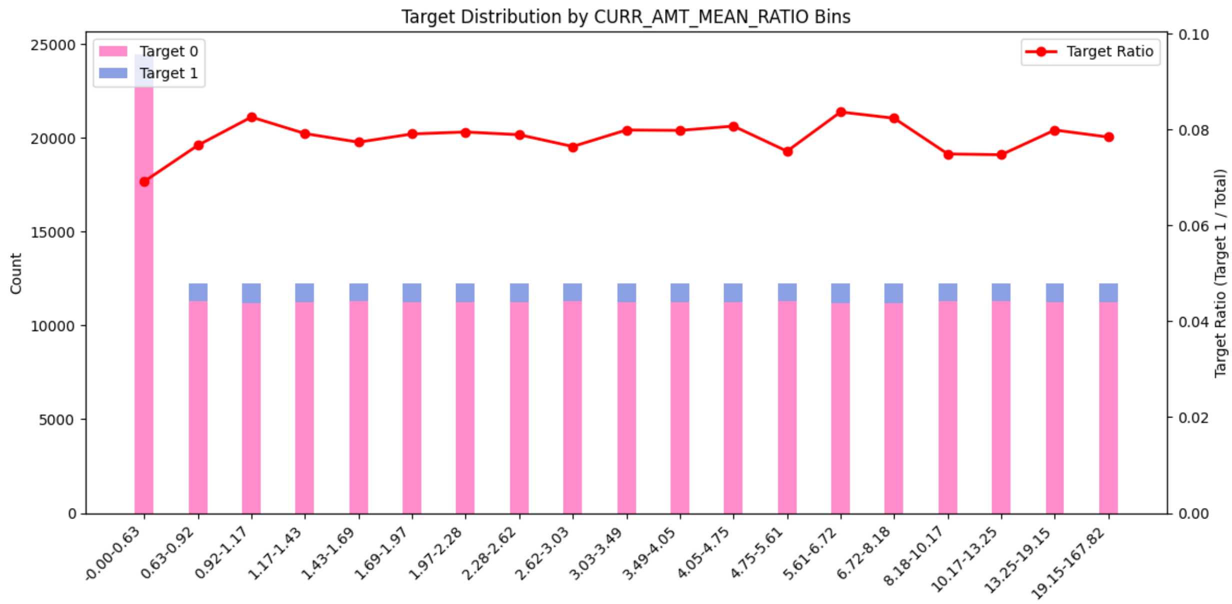
- Clients from regions with higher ratings (Rating 3) have more the default rate compared to those with lower rating (Rating 1).

- Clients who live in regions with higher populations have fewer default risk.




Target Distribution by PREV_COUNT Bins
Target 0
Target 1
Target Ratio
Count
Target Ratio (Target 1 / Total)
0
10000
20000
30000
40000
50000
60000
0.00
0.02
0.04
0.06
0.08
0.10
-0.00-1.00
1.00-2.00
2.00-3.00
3.00-4.00
4.00-5.00
5.00-6.00
6.00-7.00
7.00-9.00
9.00-66.00


Target Distribution by PREV_ACCEPTED_RATIO Bins
Target 0
Target 1
Target Ratio
Count
Target Ratio (Target 1 / Total)
0
20000
40000
60000
80000
100000
120000
140000
160000
0.00
0.02
0.04
0.06
0.08
0.10
0.12
-0.00-0.40
0.40-0.50
0.50-0.62
0.62-0.67
0.67-0.75
0.75-0.83
0.83-1.00



Insights:

- Asking for a new loan larger than any one of the previous doesn't mean higher default risk.
- As the client has more previous loans, the default risk increases.
- As the count of previously accepted loans over the total ratio increases, the default risk decreases.
- Borrowers whose current loan is smaller than their historical mean (ratio < 0.63) have a lower default rate (~7%).
- Above this threshold, default risk seems to stabilize (around 7.5-8.5%).

Preprocessing

1- Process Curr Data

Application_data.csv –> processed_current_application.csv

col with Threshold for null & row contains nulls

2- Process Prev Data

previous_application.csv → processed_previous_application.csv

1- col with Threshold for null & row contains nulls

2- Generate New Features using (prev_features.sh) file

```
NAME_CONTRACT_TYPE  -> 3 features (all features)
AMT_ANNUITY -> avg(AMT_ANNUITY)
AMT_APPLICATION -> avg(AMT_APPLICATION)
AMT_CREDIT -> avg(AMT_CREDIT)
AMT_GOODS_PRICE -> avg(AMT_GOODS_PRICE/AMT_APPLICATION)
NAME_CASH_LOAN_PURPOSE -> XAP, other
NAME_CONTRACT_STATUS -> 2 features (approved , refused)
DAYS_DECISION -> avg(DAYS_DECISION)
FLAG_LAST_APPL_PER_CONTRACT -> sum(0)
NFLAG_LAST_APPL_IN_DAY-> sum(0)
NAME_PAYMENT_TYPE -> 1 features (cash payment) -> XX
CODE_REJECT_REASON -> 3 features (XAP, HC , Limit)
NAME_CLIENT_TYPE -> 2 features (repeater , refreshed)
NAME_PORTFOLIO -> 3 features (POS , Cash , Cards)
CNT_PAYMENT -> mean(CNT_PAYMENT)
NAME_YIELD_GROUP -> avg(encoded (NAME_YIELD_GROUP))
SK_ID_CURR -> count(SK_ID_CURR)
HOUR_APPR_PROCESS_START -> mean(HOUR_APPR_PROCESS_START)
CHANNEL_TYPE -> top 3 (Credit and cash offices, Country-wide, Stone)
PRODUCT_COMBINATION -> top 3 (Cash, POS household with interest, POS mobile with interest)
```

3- Merge Data

Processed_current_application.csv & processed_current_application.csv –> merged_application.csv

Merge on SK_ID_CURR

4- Encoding_Outliers_FeatureSelection

merged_application.csv –> encoded_merged_application.csv

Encode the Features using (curr_application_features_encoding_methods.txt) file

Encoded_merged_application.csv → df_no_outliers.csv

Remove outliers using Zscore

Df_no_outliers.csv → high corr features (two files before and after)

→ featureSelected_encoded_merged_application.csv

→ X_train.csv , y_train.csv & X_test.csv , y_test.csv

1- Split data (train_test)

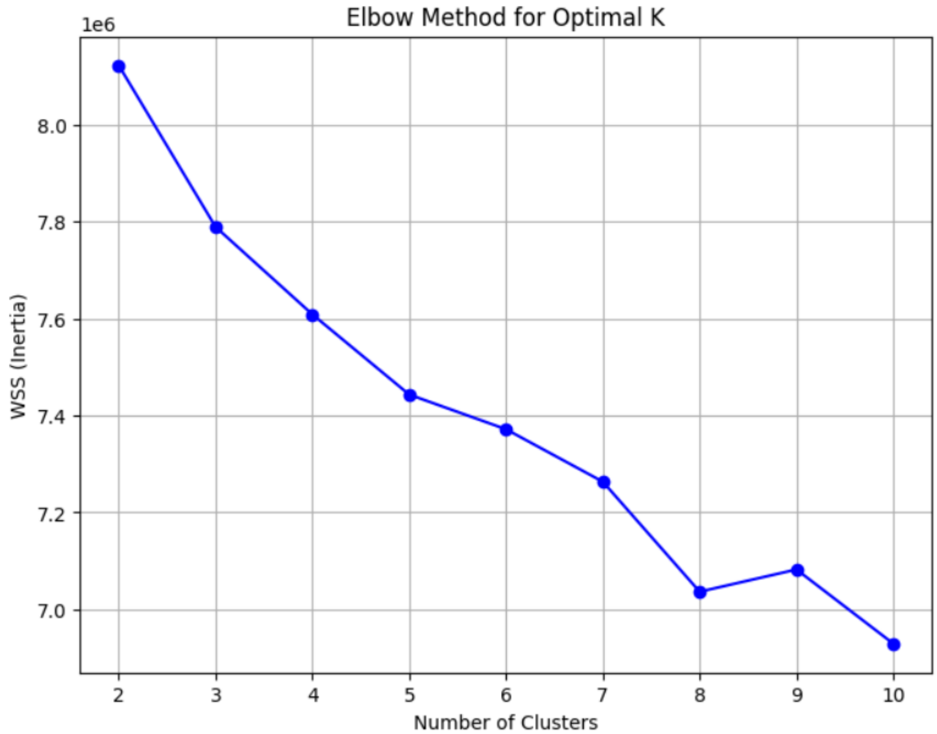
2- Study High Correlated Features (MulCorr)

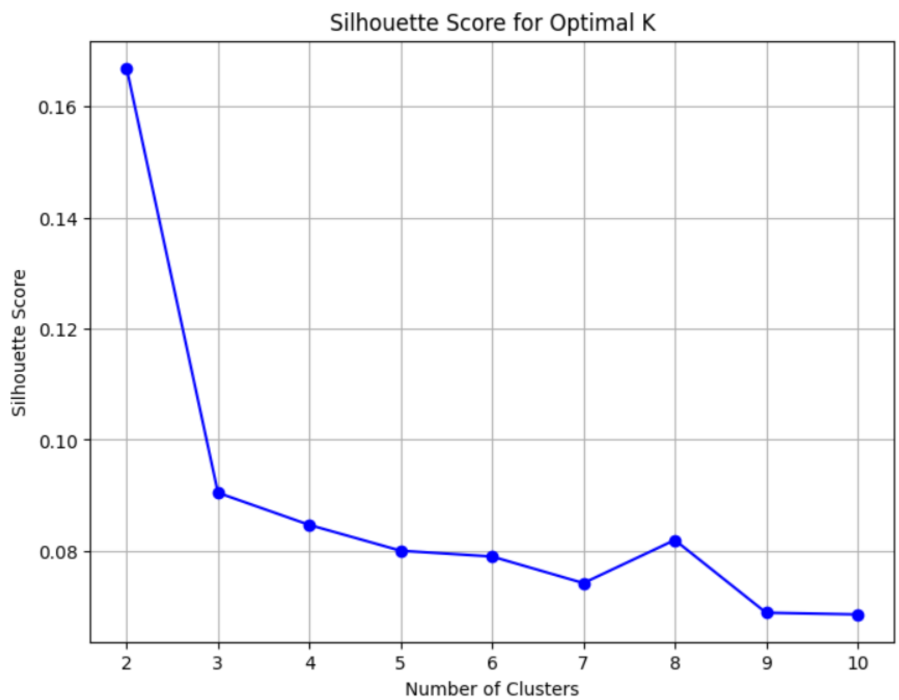
3 -Low Correlated Features with target

Clustering

1- StandardScaler

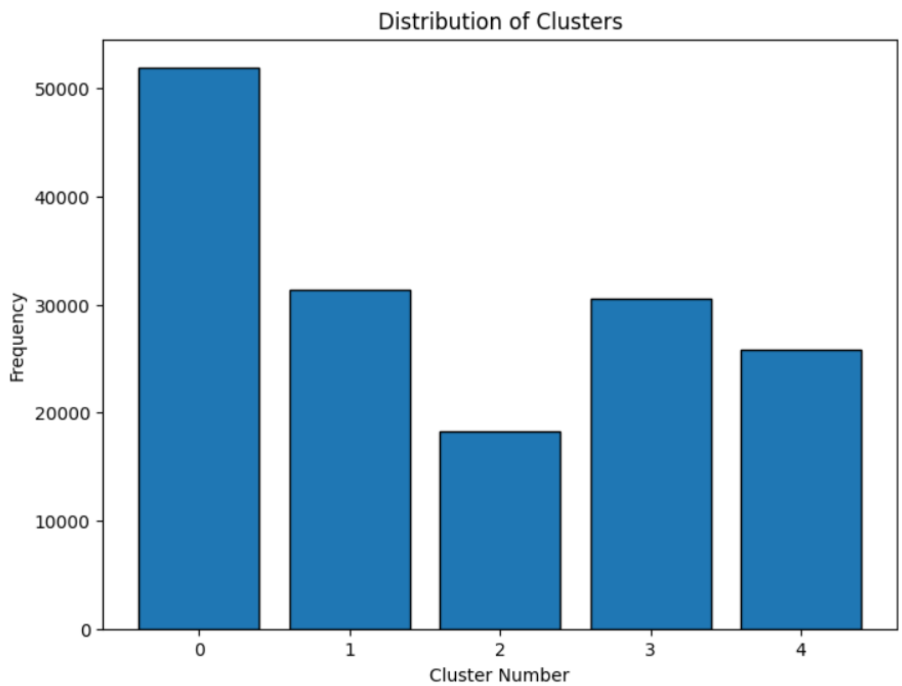
2- choose best #clusters (WSS , Silhouette)





the best number is not clear & we do not have info from the field --> we will choose 5

3 -Kmeans



Classification

1- Split with stratification (biasedData)

2- TreeBased Models

RandomForest Results

```
RandomForest AUC - Train: 0.6565381114570914, Validation:
0.6476129960178587, Test: 0.6468449418542753
```

```
RandomForest Accuracy - Train: 0.9205224339274739, Validation:
0.921956926810244, Test: 0.9232210948103148
```

```
RandomForest F1 Score - Train: 0.882428173078826, Validation:
0.8845199005620704, Test: 0.8863642273919851
```

Gradient Boosting Results

```
Gradient Boosting AUC - Train: 0.7073215529220184, Validation:
0.6771473990499939, Test: 0.6761514600234316
```

```
Gradient Boosting Accuracy - Train: 0.9205777504609711, Validation:
0.9219135936213546, Test: 0.923199551907624
```

```
Gradient Boosting F1 Score - Train: 0.8825761677401183, Validation:
0.8844982692049035, Test: 0.8863534729873809
```

Decision Tree Results

```
Decision Tree AUC - Train: 0.4504909005748263, Validation: 0.44214254129043434, Test: 0.45441470539669887
```

```
Decision Tree Accuracy - Train: 0.9205593116164721, Validation: 0.921956926810244, Test: 0.923199551907624
```

```
Decision Tree F1 Score - Train: 0.8825549170153326, Validation: 0.8845199005620704, Test: 0.8863534729873809
```

XGBoost Results

```
XGBoost AUC - Train: 0.5033, Validation: 0.5010, Test: 0.5010
```

```
XGBoost Accuracy - Train: 0.9210, Validation: 0.9220, Test: 0.9232
```

```
XGBoost F1 Score - Train: 0.8837, Validation: 0.8849, Test: 0.8867
```

3- LogReg

Logistic Regression Results

```
Logistic Regression AUC - Train: 0.6942, Validation: 0.6905, Test: 0.6859
```

```
Logistic Regression Accuracy - Train: 0.9204, Validation: 0.9219, Test: 0.9231
```

```
Logistic Regression F1 Score - Train: 0.8826, Validation: 0.8848, Test: 0.8865
```

4- K-fold CV (XGBoost)

paramGrid

```
paramGrid = (ParamGridBuilder()

    .addGrid(xgb.max_depth, [6, 10])

    .addGrid(xgb.learning_rate, [0.1, 0.3])

    .addGrid(xgb.n_estimators, [100,200])

    .build())
```

Best Model Results

```
Cross-Validated XGBoost F1 - Train+Val: 0.9985, Test: 0.8875
```

Enhancements and future work

Applying DL model and Deployment