**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG


BÁO CÁO ĐỒ ÁN MÔN HỌC

# NHẬP MÔN HỌC MÁY VÀ KHAI PHÁ DỮ LIỆU

# ĐỀ TÀI: DỰ ĐOÁN MỨC ĐỘ RỦI RO CỦA HỒ SƠ VAY TÍN DỤNG

**Nhóm sinh viên thực hiện:**

| | |
|---|---|
| Trần Hữu Huy | 20183557 |
| Lê Dương Long | 20183582 |
| Hoàng Đức Việt | 20183666 |
| Nguyễn Minh Hiếu | 20183532 |
| Nguyễn Đức Khánh | 20183563 |

Giảng viên hướng dẫn: **TS. Nguyễn Nhật Quang**

**Hà Nội, tháng 6 năm 2021**

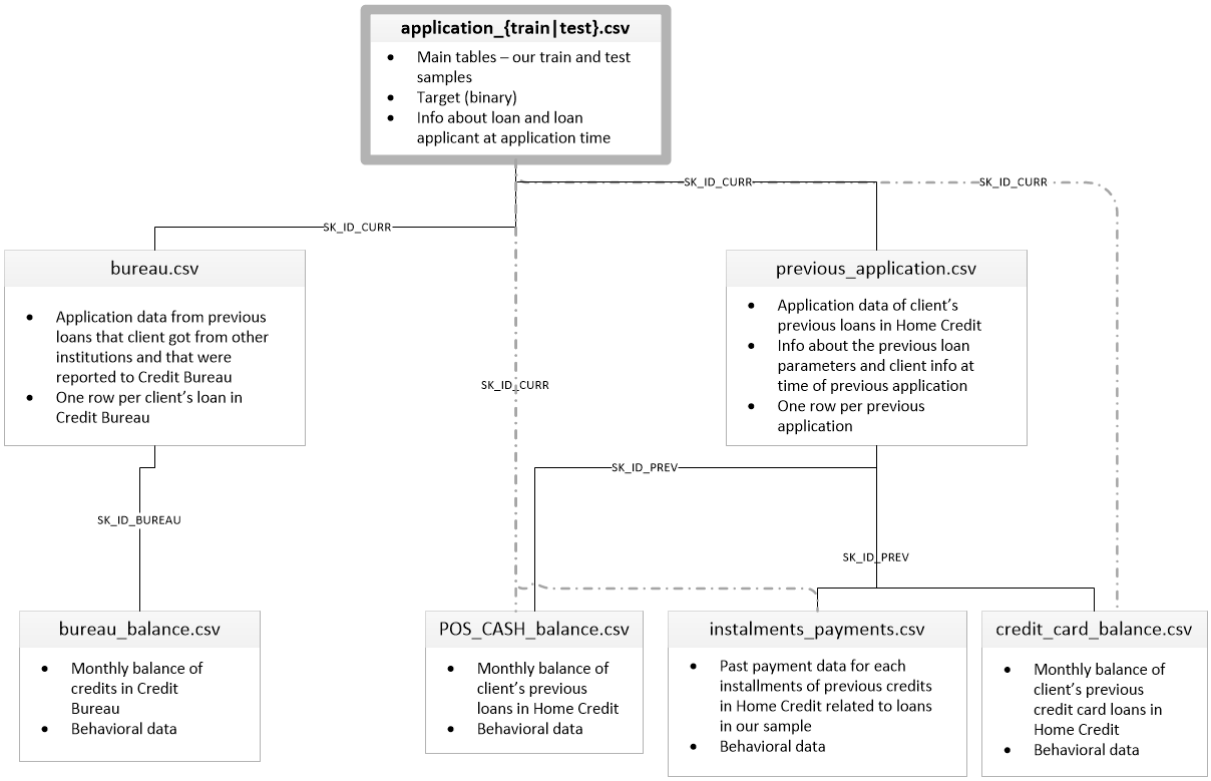
# MỤC LỤC



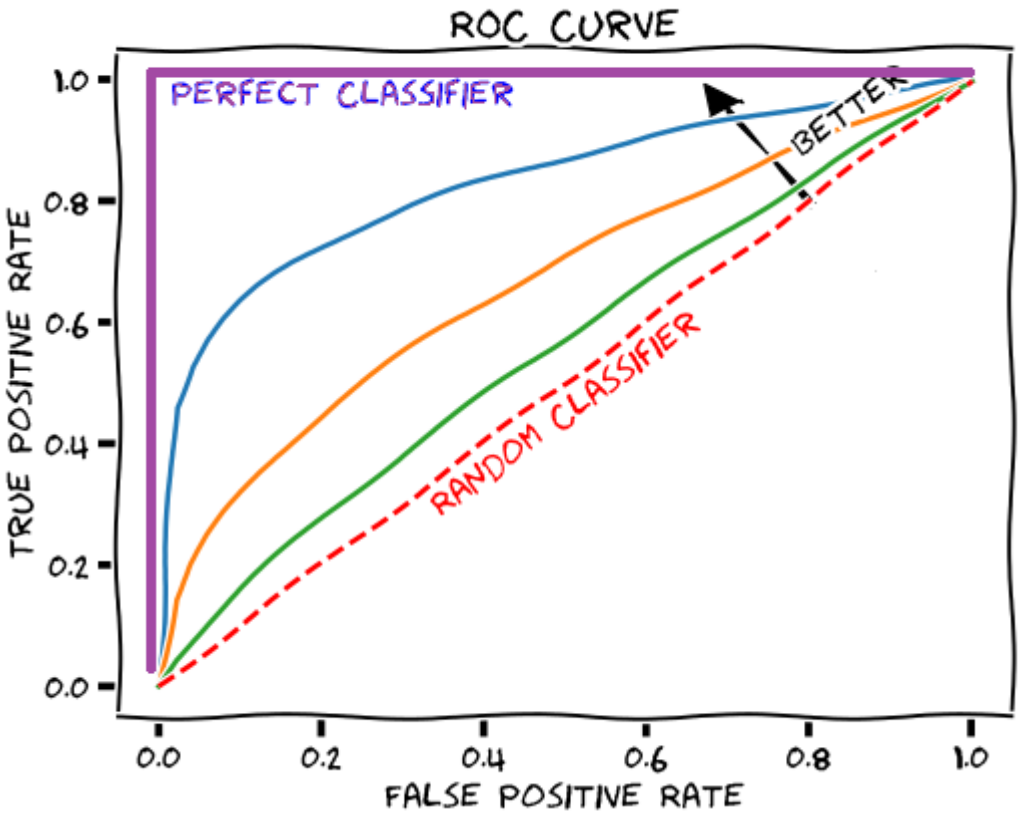
## Phần I. Giới thiệu đề tài

1. Đặt vấn đề
   - Nhiều người gặp khó khăn trong việc vay vốn do lịch sử tín dụng không đủ hoặc không tồn tại. Để giải quyết vấn đề này, Home Credit (Tổ chức hàng đầu trong lĩnh vực vay tiêu dùng trả góp) cố gắng mở rộng phạm vi tài chính cho những người trường hợp trên
   - Home Credit sử dụng nhiều phương pháp thống kê và học máy để dự đoán mức độ rủi ro trả nợ của khách hàng. Xử lý tốt bài toán đó sẽ đảm bảo rằng khách hàng có khả năng trả nợ không bị từ chối và cũng giúp tối đa hóa lợi nhuận công ty khi từ chối những khách hàng không có khả năng trả nợ
   - Home Credit cung cấp bộ dữ liệu thực của họ trên Kaggle vào năm 2018
2. Dữ liệu
   - Bộ dữ liệu có 9 file csv (do dịch ra nghe không sát nghĩa nên chúng em xin phép để như bản gốc):
     - (1) application_{train|test}.csv
       - This is the main table, broken into two files for Train (with TARGET) and Test (without TARGET)
       - Static data for all applications. One row represents one loan in our data sample
     - (2) bureau.csv

        - All client's previous credits provided by other financial institutions that were reported to Credit Bureau (for clients who have a loan in our sample)
        - For every loan in our sample, there are as many rows as number of credits the client had in Credit Bureau before the application date
    - (3) bureau_balance.csv
        - Monthly balances of previous credits in Credit Bureau
        - This table has one row for each month of history of every previous credit reported to Credit Bureau – i.e the table has (#loans in sample * # of relative previous credits * # of months where we have some history observable for the previous credits) rows
    - (4) POS_CASH_balance.csv
        - Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit
        - This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e.the table has (#loans in sample * # of relative previous credits * # of months in which we have some history observable for the previous credits) rows
    - (5) credit_card_balance.csv
        - Monthly balance snapshots of previous credit cards that the applicant has with Home Credit
        - This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e the table has (#loans in sample * # of relative previous credit cards * # of months where we have some history observable for the previous credit card) rows
    - (6) previous_application.csv
        - All previous applications for Home Credit loans of clients who have loans in our sample
        - There is one row for each previous application related to loans in our data sample
    - (7) installments_payments.csv
        - Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample
        - There is a) one row for every payment that was made plus b) one row each for missed payment
        - One row is equivalent to one payment of one installment OR one installment corresponding to one payment of one previous Home Credit credit related to loans in our sample
    - (8) HomeCredit_columns_description.csv
        - This file contains descriptions for the columns in the various data files




3. Chỉ số sử dụng để đánh giá
   - Do đây là một bài toán Imbalanced Data nên chỉ số sử dụng sẽ là ROC AUC



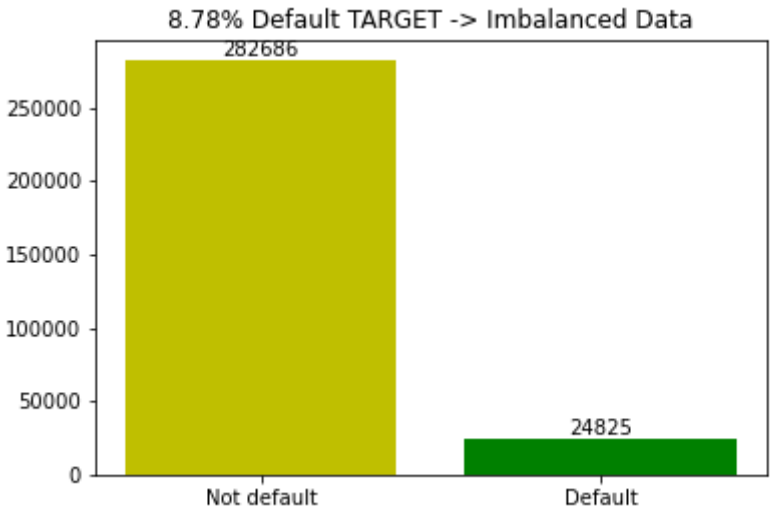

   - Các mô hình được sử dụng:

- Logistic Regression
- Naïve Bayes Classifier
- XGBoost Classifier
- Neural Network

4. Tài liệu tham khảo
   - Giáo trình Nhập môn Học Máy và Khai phá Dữ liệu của TS. Nguyễn Nhật Quang
   - Kaggle: https://www.kaggle.com/c/home-credit-default-risk

## Phần II. EDA và Feature Engineering

1. application_{train|test}.csv
   - Kích thước:
     - application_train: (307511, 122)
     - application_test: (48744, 122)
   - Phân bố của label (train):
     - Default – Nợ **1**
     - Not default – Không nợ **0**




   - Số lượng các kiểu dữ liệu:
     - float64 65
     - int64 41
     - object 16
   - Số lượng missing data (train):

| | Missing Count | Missing Count Ratio | Missing Count % |
|---|---|---|---|
| COMMONAREA_MEDI | 214865 | 0.698723 | 69.9 |
| COMMONAREA_AVG | 214865 | 0.698723 | 69.9 |
| COMMONAREA_MODE | 214865 | 0.698723 | 69.9 |
| NONLIVINGAPARTMENTS_MODE | 213514 | 0.69433 | 69.4 |
| NONLIVINGAPARTMENTS_MEDI | 213514 | 0.69433 | 69.4 |
| NONLIVINGAPARTMENTS_AVG | 213514 | 0.69433 | 69.4 |

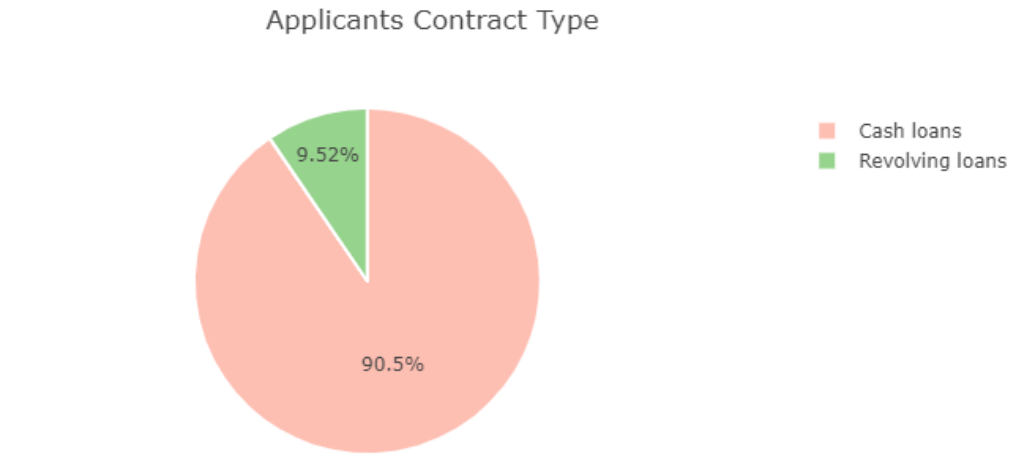
| | | | |
|---|---|---|---|
| FONDKAPREMONT_MODE | 210295 | 0.683862 | 68.4 |
| LIVINGAPARTMENTS_MEDI | 210199 | 0.68355 | 68.4 |
| LIVINGAPARTMENTS_MODE | 210199 | 0.68355 | 68.4 |
| LIVINGAPARTMENTS_AVG | 210199 | 0.68355 | 68.4 |
| FLOORSMIN_MEDI | 208642 | 0.678486 | 67.8 |
| FLOORSMIN_MODE | 208642 | 0.678486 | 67.8 |
| FLOORSMIN_AVG | 208642 | 0.678486 | 67.8 |
| YEARS_BUILD_MEDI | 204488 | 0.664978 | 66.5 |
| YEARS_BUILD_AVG | 204488 | 0.664978 | 66.5 |
| YEARS_BUILD_MODE | 204488 | 0.664978 | 66.5 |
| OWN_CAR_AGE | 202929 | 0.659908 | 66 |
| LANDAREA_MODE | 182590 | 0.593767 | 59.4 |
| LANDAREA_AVG | 182590 | 0.593767 | 59.4 |
| LANDAREA_MEDI | 182590 | 0.593767 | 59.4 |
| BASEMENTAREA_MEDI | 179943 | 0.58516 | 58.5 |
| BASEMENTAREA_AVG | 179943 | 0.58516 | 58.5 |
| BASEMENTAREA_MODE | 179943 | 0.58516 | 58.5 |
| EXT_SOURCE_1 | 173378 | 0.563811 | 56.4 |
| NONLIVINGAREA_MEDI | 169682 | 0.551792 | 55.2 |
| NONLIVINGAREA_AVG | 169682 | 0.551792 | 55.2 |
| NONLIVINGAREA_MODE | 169682 | 0.551792 | 55.2 |
| ELEVATORS_MODE | 163891 | 0.53296 | 53.3 |
| ELEVATORS_AVG | 163891 | 0.53296 | 53.3 |
| ELEVATORS_MEDI | 163891 | 0.53296 | 53.3 |
| WALLSMATERIAL_MODE | 156341 | 0.508408 | 50.8 |
| APARTMENTS_MODE | 156061 | 0.507497 | 50.7 |
| APARTMENTS_AVG | 156061 | 0.507497 | 50.7 |
| APARTMENTS_MEDI | 156061 | 0.507497 | 50.7 |
| ENTRANCES_MEDI | 154828 | 0.503488 | 50.3 |
| ENTRANCES_MODE | 154828 | 0.503488 | 50.3 |
| ENTRANCES_AVG | 154828 | 0.503488 | 50.3 |
| LIVINGAREA_MEDI | 154350 | 0.501933 | 50.2 |
| LIVINGAREA_MODE | 154350 | 0.501933 | 50.2 |
| LIVINGAREA_AVG | 154350 | 0.501933 | 50.2 |
| HOUSETYPE_MODE | 154297 | 0.501761 | 50.2 |
| FLOORSMAX_MODE | 153020 | 0.497608 | 49.8 |
| FLOORSMAX_MEDI | 153020 | 0.497608 | 49.8 |
| FLOORSMAX_AVG | 153020 | 0.497608 | 49.8 |
| YEARS_BEGINEXPLUATATION_MEDI | 150007 | 0.48781 | 48.8 |
| YEARS_BEGINEXPLUATATION_AVG | 150007 | 0.48781 | 48.8 |
| YEARS_BEGINEXPLUATATION_MODE | 150007 | 0.48781 | 48.8 |
| TOTALAREA_MODE | 148431 | 0.482685 | 48.3 |
| EMERGENCYSTATE_MODE | 145755 | 0.473983 | 47.4 |
| OCCUPATION_TYPE | 96391 | 0.313455 | 31.3 |
| EXT_SOURCE_3 | 60965 | 0.198253 | 19.8 |
| AMT_REQ_CREDIT_BUREAU_QRT | 41519 | 0.135016 | 13.5 |

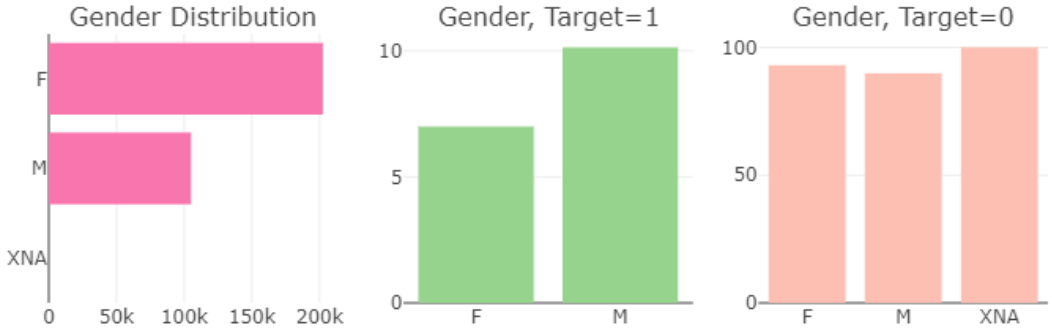
| | | | |
|---|---|---|---|
| AMT_REQ_CREDIT_BUREAU_YEAR | 41519 | 0.135016 | 13.5 |
| AMT_REQ_CREDIT_BUREAU_WEEK | 41519 | 0.135016 | 13.5 |
| AMT_REQ_CREDIT_BUREAU_MON | 41519 | 0.135016 | 13.5 |
| AMT_REQ_CREDIT_BUREAU_DAY | 41519 | 0.135016 | 13.5 |
| AMT_REQ_CREDIT_BUREAU_HOUR | 41519 | 0.135016 | 13.5 |
| NAME_TYPE_SUITE | 1292 | 0.004201 | 0.4 |
| OBS_30_CNT_SOCIAL_CIRCLE | 1021 | 0.00332 | 0.3 |
| OBS_60_CNT_SOCIAL_CIRCLE | 1021 | 0.00332 | 0.3 |
| DEF_60_CNT_SOCIAL_CIRCLE | 1021 | 0.00332 | 0.3 |
| DEF_30_CNT_SOCIAL_CIRCLE | 1021 | 0.00332 | 0.3 |
| EXT_SOURCE_2 | 660 | 0.002146 | 0.2 |
| AMT_GOODS_PRICE | 278 | 0.000904 | 0.1 |
| AMT_ANNUITY | 12 | 3.90E-05 | 0 |
| CNT_FAM_MEMBERS | 2 | 6.50E-06 | 0 |
| DAYS_LAST_PHONE_CHANGE | 1 | 3.25E-06 | 0 |
| AMT_CREDIT | 0 | 0 | 0 |
| FLAG_OWN_CAR | 0 | 0 | 0 |
| FLAG_EMAIL | 0 | 0 | 0 |
| TARGET | 0 | 0 | 0 |
| FLAG_PHONE | 0 | 0 | 0 |
| FLAG_CONT_MOBILE | 0 | 0 | 0 |
| FLAG_WORK_PHONE | 0 | 0 | 0 |
| FLAG_EMP_PHONE | 0 | 0 | 0 |
| FLAG_MOBIL | 0 | 0 | 0 |
| NAME_CONTRACT_TYPE | 0 | 0 | 0 |
| CODE_GENDER | 0 | 0 | 0 |
| FLAG_OWN_REALTY | 0 | 0 | 0 |
| AMT_INCOME_TOTAL | 0 | 0 | 0 |
| DAYS_ID_PUBLISH | 0 | 0 | 0 |
| DAYS_REGISTRATION | 0 | 0 | 0 |
| DAYS_EMPLOYED | 0 | 0 | 0 |
| DAYS_BIRTH | 0 | 0 | 0 |
| REGION_POPULATION_RELATIVE | 0 | 0 | 0 |
| REGION_RATING_CLIENT | 0 | 0 | 0 |
| NAME_FAMILY_STATUS | 0 | 0 | 0 |
| NAME_EDUCATION_TYPE | 0 | 0 | 0 |
| NAME_INCOME_TYPE | 0 | 0 | 0 |
| CNT_CHILDREN | 0 | 0 | 0 |
| NAME_HOUSING_TYPE | 0 | 0 | 0 |
| REG_REGION_NOT_LIVE_REGION | 0 | 0 | 0 |
| REGION_RATING_CLIENT_W_CITY | 0 | 0 | 0 |
| WEEKDAY_APPR_PROCESS_START | 0 | 0 | 0 |
| FLAG_DOCUMENT_2 | 0 | 0 | 0 |
| FLAG_DOCUMENT_3 | 0 | 0 | 0 |
| FLAG_DOCUMENT_4 | 0 | 0 | 0 |
| FLAG_DOCUMENT_5 | 0 | 0 | 0 |

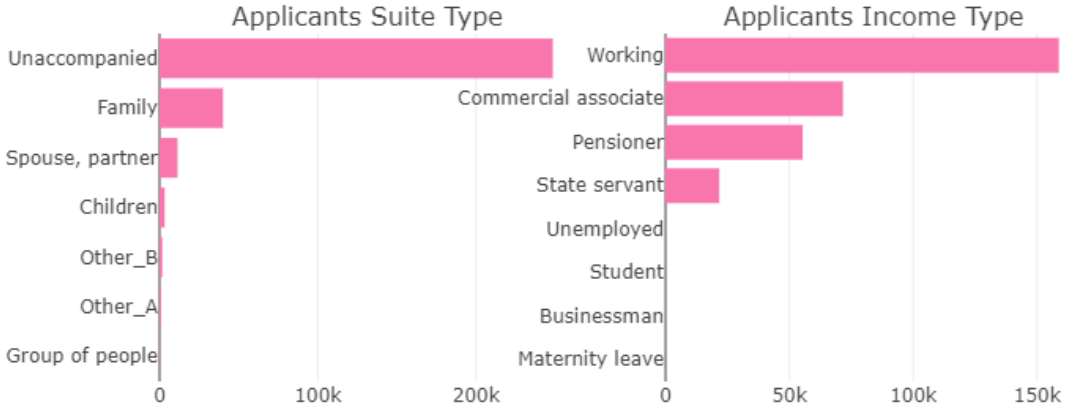
| | | | |
|---|---|---|---|
| FLAG_DOCUMENT_6 | 0 | 0 | 0 |
| FLAG_DOCUMENT_7 | 0 | 0 | 0 |
| FLAG_DOCUMENT_8 | 0 | 0 | 0 |
| FLAG_DOCUMENT_9 | 0 | 0 | 0 |
| FLAG_DOCUMENT_10 | 0 | 0 | 0 |
| FLAG_DOCUMENT_11 | 0 | 0 | 0 |
| FLAG_DOCUMENT_12 | 0 | 0 | 0 |
| FLAG_DOCUMENT_13 | 0 | 0 | 0 |
| FLAG_DOCUMENT_14 | 0 | 0 | 0 |
| FLAG_DOCUMENT_15 | 0 | 0 | 0 |
| FLAG_DOCUMENT_16 | 0 | 0 | 0 |
| FLAG_DOCUMENT_17 | 0 | 0 | 0 |
| FLAG_DOCUMENT_18 | 0 | 0 | 0 |
| FLAG_DOCUMENT_19 | 0 | 0 | 0 |
| FLAG_DOCUMENT_20 | 0 | 0 | 0 |
| FLAG_DOCUMENT_21 | 0 | 0 | 0 |
| ORGANIZATION_TYPE | 0 | 0 | 0 |
| LIVE_CITY_NOT_WORK_CITY | 0 | 0 | 0 |
| REG_CITY_NOT_WORK_CITY | 0 | 0 | 0 |
| REG_CITY_NOT_LIVE_CITY | 0 | 0 | 0 |
| LIVE_REGION_NOT_WORK_REGION | 0 | 0 | 0 |
| REG_REGION_NOT_WORK_REGION | 0 | 0 | 0 |
| HOUR_APPR_PROCESS_START | 0 | 0 | 0 |
| SK_ID_CURR | 0 | 0 | 0 |

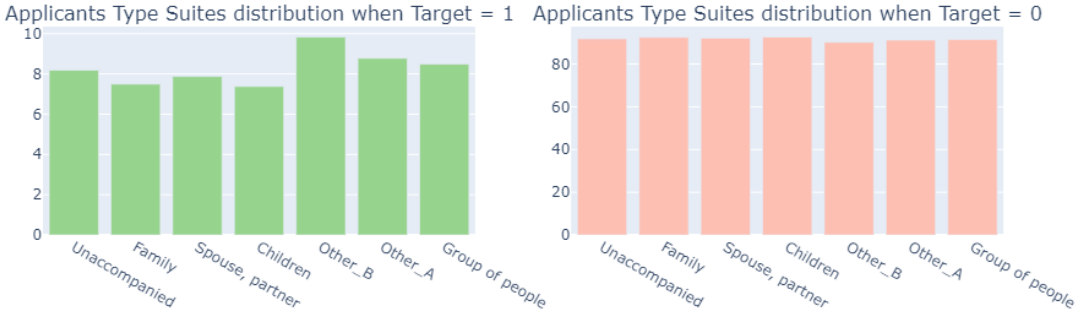
- Các dữ liệu Categorical:
    - NAME_CONTRACT_TYPE: ['Cash loans', 'Revolving loans']
    - CODE_GENDER: ['F', 'M', 'XNA']
    - FLAG_OWN_CAR: ['N', 'Y']
    - FLAG_OWN_REALTY: ['Y', 'N']
    - NAME_TYPE_SUITE: ['Unaccompanied', 'Family', 'Spouse, partner', 'Children', 'Other_B', 'Other_A', 'Group of people']
    - NAME_INCOME_TYPE: ['Working', 'Commercial associate', 'Pensioner', 'State servant', 'Unemployed', 'Student', 'Businessman', 'Maternity leave']
    - NAME_EDUCATION_TYPE: ['Secondary / secondary special', 'Higher education', 'Incomplete higher', 'Lower secondary', 'Academic degree']
    - NAME_FAMILY_STATUS: ['Married', 'Single / not married', 'Civil marriage', 'Separated', 'Widow', 'Unknown']

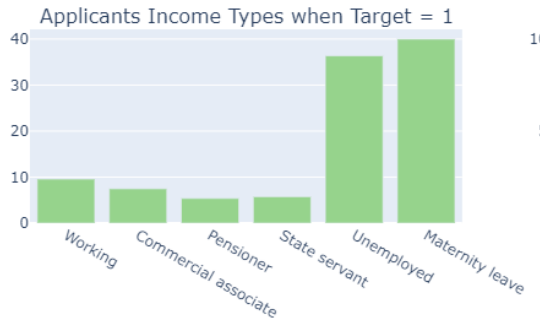
- NAME_HOUSING_TYPE: ['House / apartment', 'With parents', 'Municipal apartment', 'Rented apartment', 'Office apartment', 'Co-op apartment']
- OCCUPATION_TYPE: ['Laborers', 'Sales staff', 'Core staff', 'Managers', 'Drivers', 'High skill tech staff', 'Accountants', 'Medicine staff', 'Security staff', 'Cooking staff', 'Cleaning staff', 'Private service staff', 'Low-skill Laborers', 'Waiters/barmen staff', 'Secretaries', 'Realty agents', 'HR staff', 'IT staff']
- WEEKDAY_APPR_PROCESS_START: ['TUESDAY', 'WEDNESDAY', 'MONDAY', 'THURSDAY', 'FRIDAY', 'SATURDAY', 'SUNDAY']
- ORGANIZATION_TYPE: ['Business Entity Type 3', 'XNA', 'Self-employed', 'Other', 'Medicine', 'Business Entity Type 2', 'Government', 'School', 'Trade: type 7', 'Kindergarten', 'Construction', 'Business Entity Type 1', 'Transport: type 4', 'Trade: type 3', 'Industry: type 9', 'Industry: type 3', 'Security', 'Housing', 'Industry: type 11', 'Military', 'Bank', 'Agriculture', 'Police', 'Transport: type 2', 'Postal', 'Security Ministries', 'Trade: type 2', 'Restaurant', 'Services', 'University', 'Industry: type 7', 'Transport: type 3', 'Industry: type 1', 'Hotel', 'Electricity', 'Industry: type 4', 'Trade: type 6', 'Industry: type 5', 'Insurance', 'Telecom', 'Emergency', 'Industry: type 2', 'Advertising', 'Realtor', 'Culture', 'Industry: type 12', 'Trade: type 1', 'Mobile', 'Legal Services', 'Cleaning', 'Transport: type 1', 'Industry: type 6', 'Industry: type 10', 'Religion', 'Industry: type 13', 'Trade: type 4', 'Trade: type 5', 'Industry: type 8']
- FONDKAPREMONT_MODE: ['reg oper account', 'reg oper spec account', 'not specified', 'org spec account']
- HOUSETYPE_MODE: ['block of flats', 'specific housing', 'terraced house']
- WALLSMATERIAL_MODE: ['Panel', 'Stone, brick', 'Block', 'Wooden', 'Mixed', 'Monolithic', 'Others']
- EMERGENCYSTATE_MODE: ['No', 'Yes']

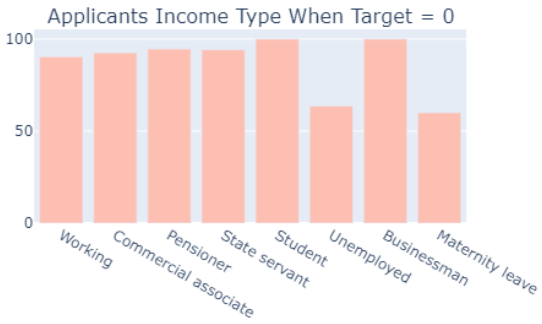

Applicants Contract Type
9.52%
90.5%
Cash loans
Revolving loans

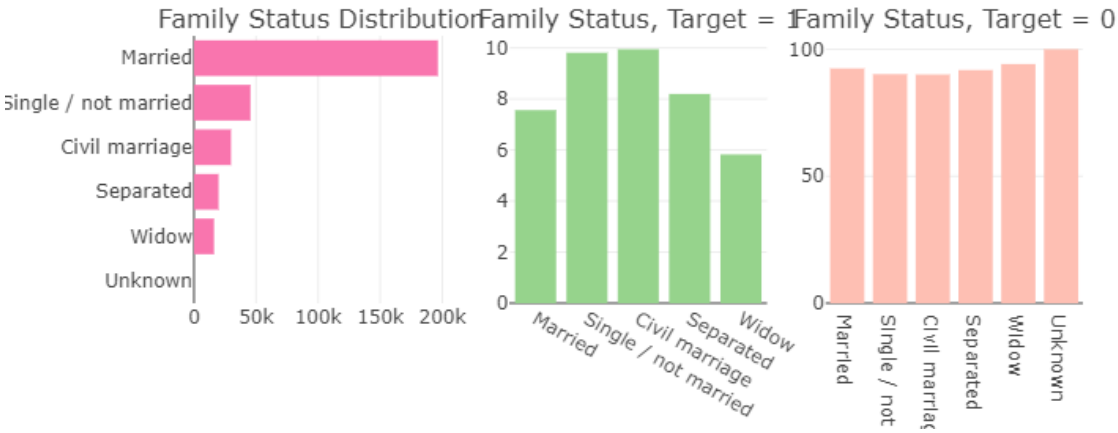

Gender Distribution
F
M
XNA
0
50k
100k
150k
200k
Gender, Target=1
10
5
0
F
M
Gender, Target=0
100
50
0
F
M
XNA

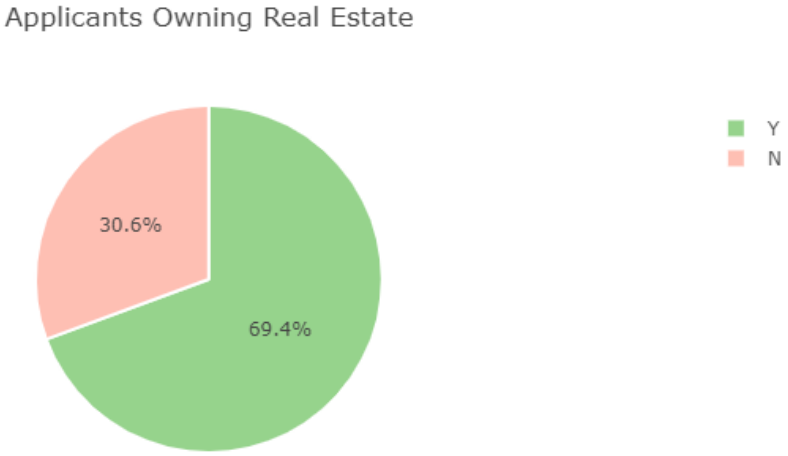

Applicants Suite Type
Unaccompanied
Family
Spouse, partner
Children
Other_B
Other_A
Group of people
0
100k
200k
Applicants Income Type
Working
Commercial associate
Pensioner
State servant
Unemployed
Student
Businessman
Maternity leave
0
50k
100k
150k

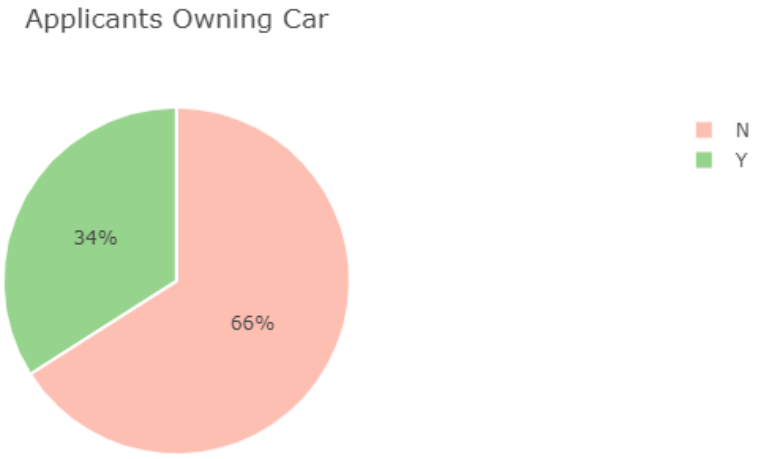

Applicants Type Suites distribution when Target = 1
10
8
6
4
2
0
Unaccompanied
Family
Spouse, partner
Children
Other_B
Other_A
Group of people
Applicants Type Suites distribution when Target = 0
80
60
40
20
0
Unaccompanied
Family
Spouse, partner
Children
Other_B
Other_A
Group of people

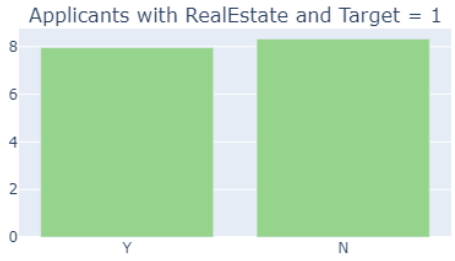

Applicants Income Types when Target = 1
40
30
20
10
0
Working
Commercial associate
Pensioner
State servant
Unemployed
Maternity leave

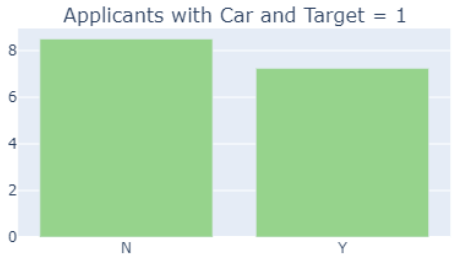

Applicants Income Type When Target = 0
100
50
0
Working
Commercial associate
Pensioner
State servant
Student
Unemployed
Businessman
Maternity leave

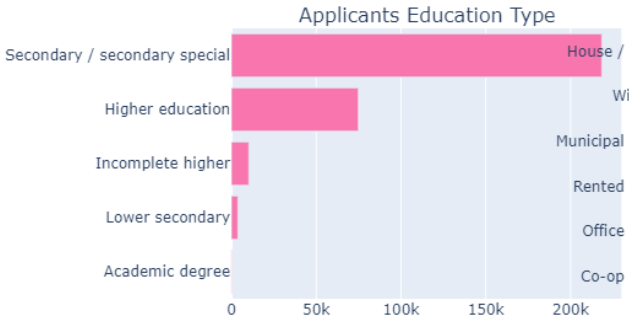

Family Status Distribution
Married
Single / not married
Civil marriage
Separated
Widow
Unknown
0
50k
100k
150k
200k
Family Status, Target = 1
10
8
6
4
2
0
Married
Single / not married
Civil marriage
Separated
Widow
Family Status, Target = 0
100
50
0
Married
Single / not
Civil marriag
Separated
Widow
Unknown

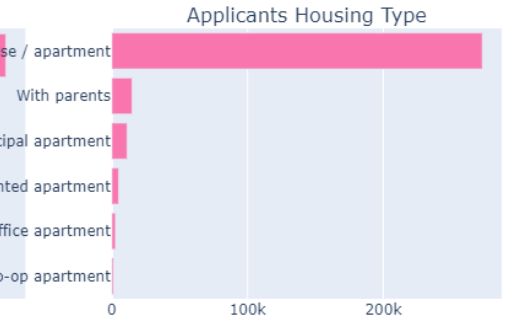

Applicants Owning Real Estate
30.6%
69.4%
Y
N

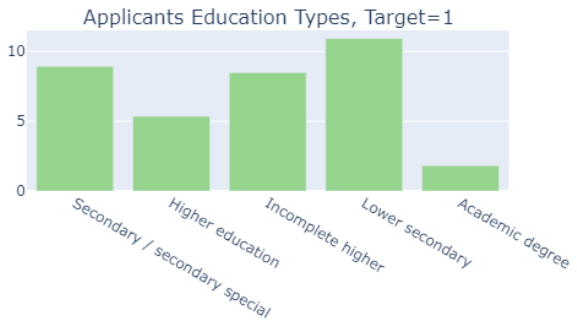

Applicants Owning Car
34%
66%
N
Y

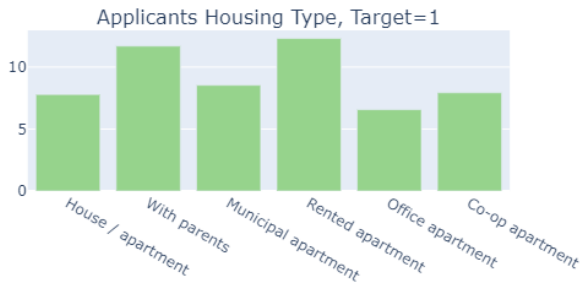

Applicants with RealEstate and Target = 1
8
6
4
2
0
Y
N

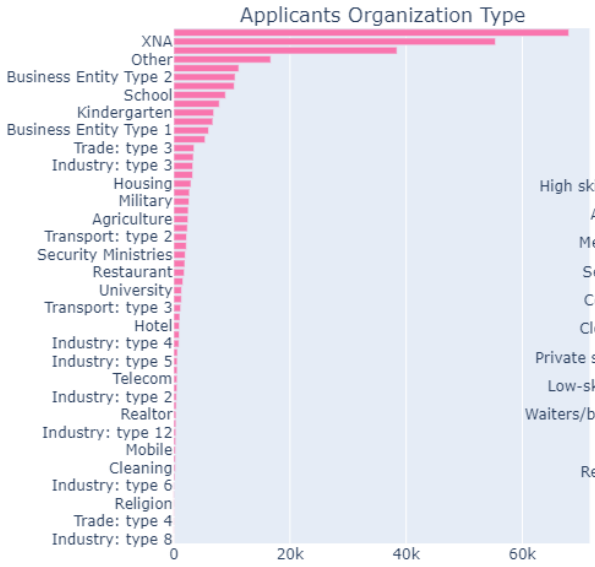

Applicants with Car and Target = 1
8
6
4
2
0
N
Y

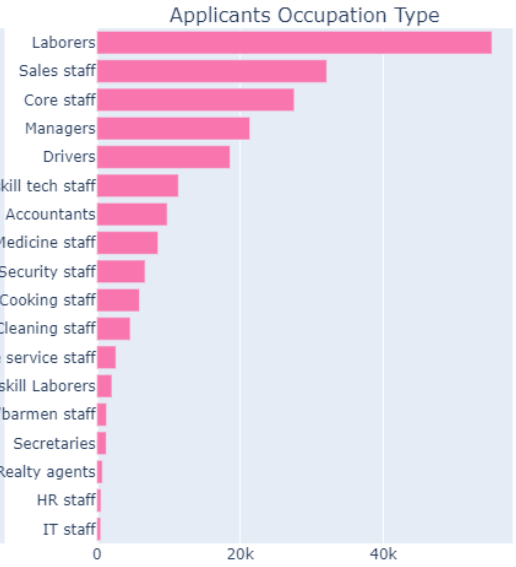

Applicants Education Type
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
0
50k
100k
150k
200k

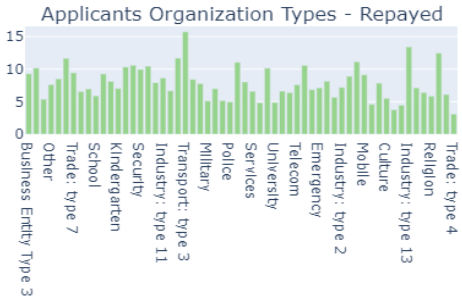

Applicants Housing Type
House / apartment
With parents
Municipal apartment
Rented apartment
Office apartment
Co-op apartment
0
100k
200k

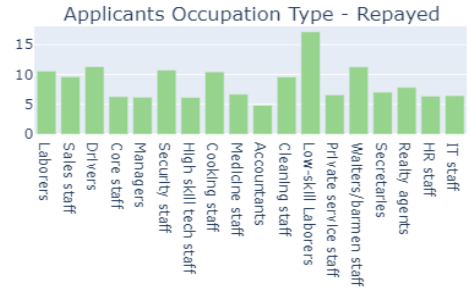

Applicants Education Types, Target=1
10
5
0
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree

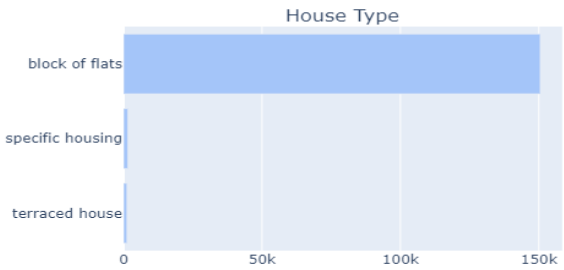

Applicants Housing Type, Target=1
10
5
0
House / apartment
With parents
Municipal apartment
Rented apartment
Office apartment
Co-op apartment

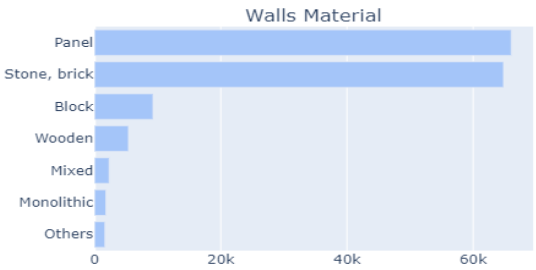

Applicants Organization Type
XNA
Other
Business Entity Type 2
School
Kindergarten
Business Entity Type 1
Trade: type 3
Industry: type 3
Housing
Military
Agriculture
Transport: type 2
Security Ministries
Restaurant
University
Transport: type 3
Hotel
Industry: type 4
Industry: type 5
Telecom
Industry: type 2
Realtor
Industry: type 12
Mobile
Cleaning
Industry: type 6
Religion
Trade: type 4
Industry: type 8
0
20k
40k
60k

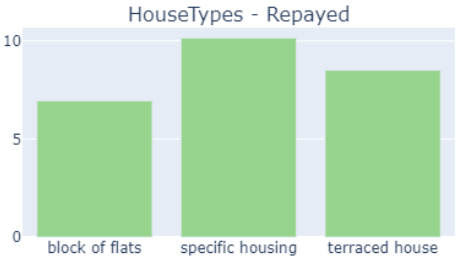

Applicants Occupation Type
Laborers
Sales staff
Core staff
Managers
Drivers
High skill tech staff
Accountants
Medicine staff
Security staff
Cooking staff
Cleaning staff
Private service staff
Low-skill Laborers
Waiters/barmen staff
Secretaries
Realty agents
HR staff
IT staff
0
20k
40k

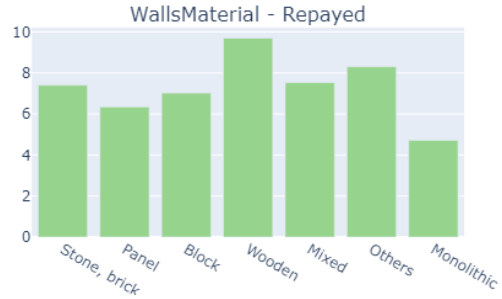


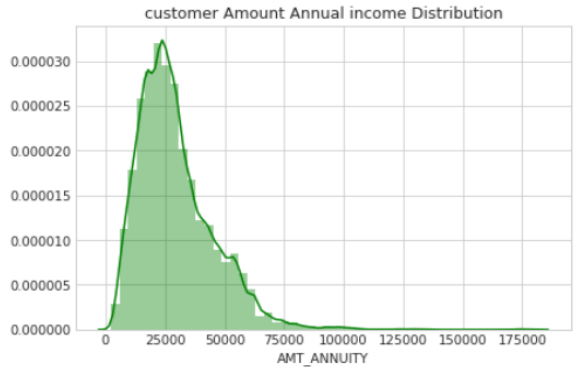



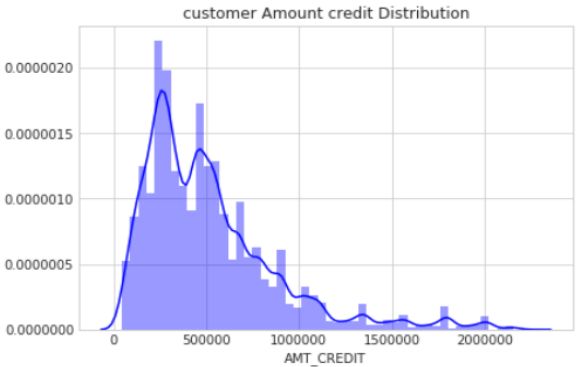



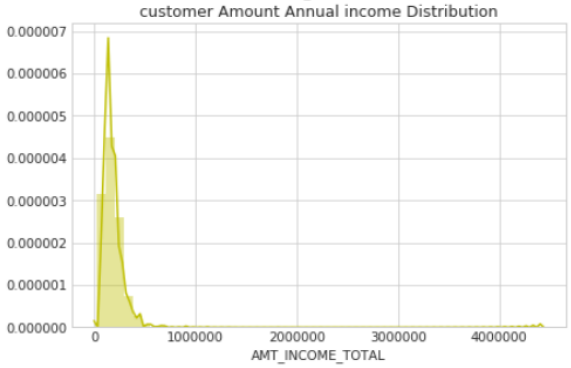



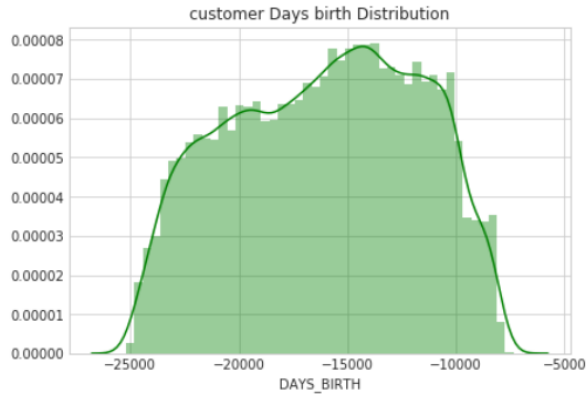

- Các dữ liệu Numerical:

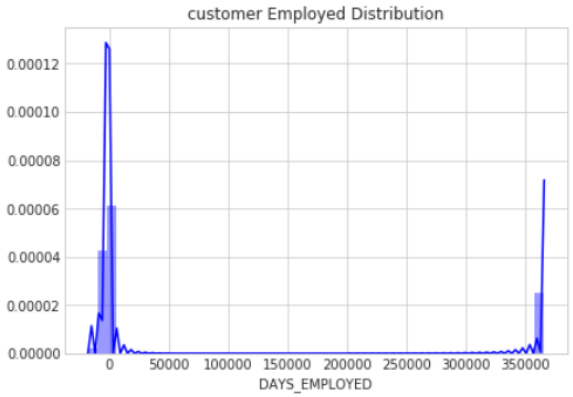


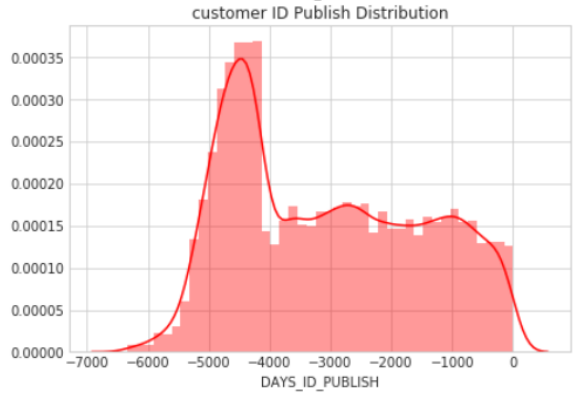



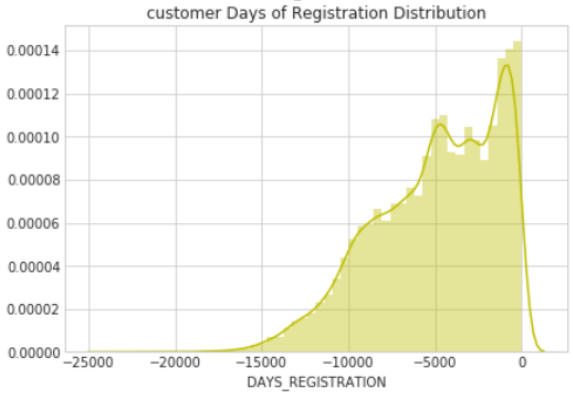


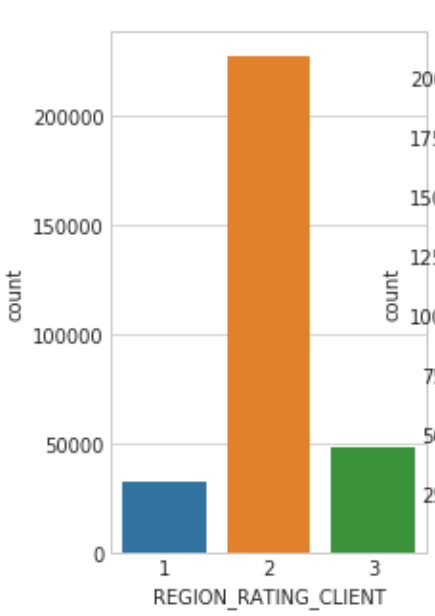


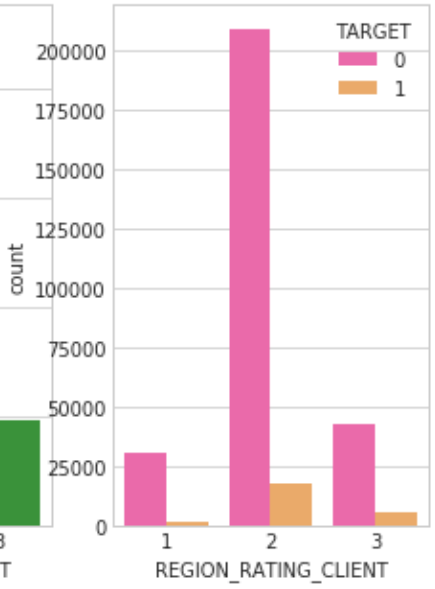

customer Days birth Distribution
DAYS_BIRTH

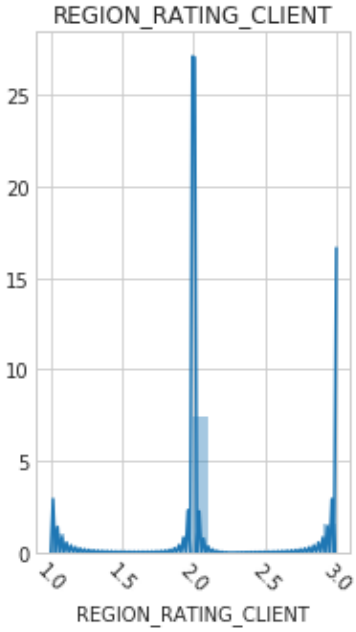

customer Employed Distribution
DAYS_EMPLOYED

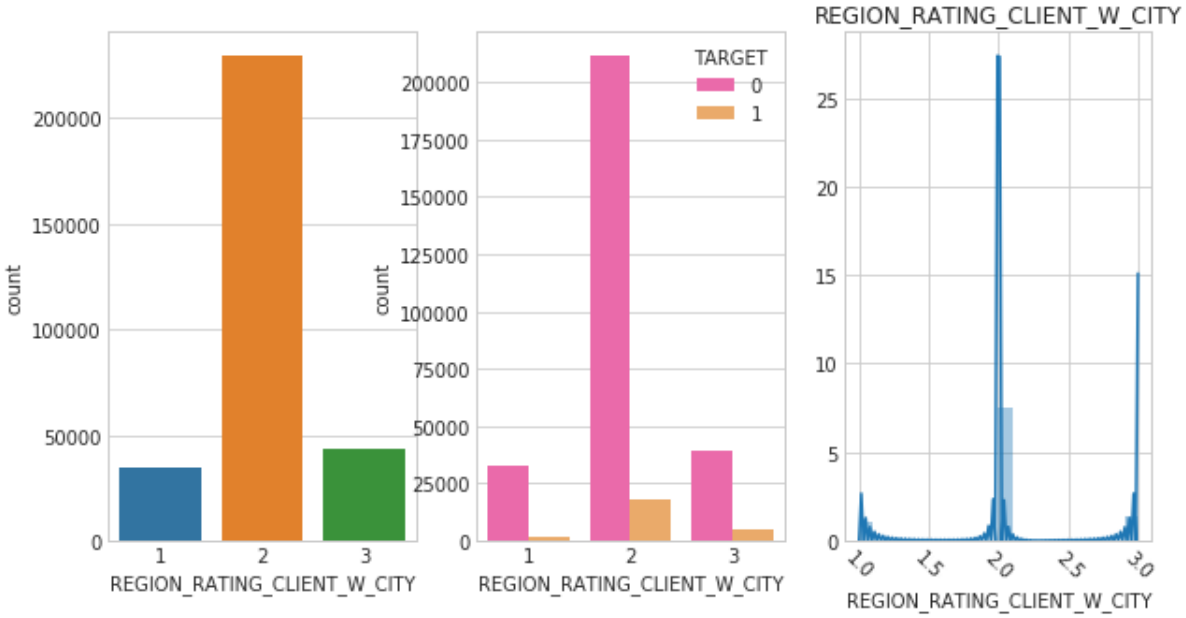

customer ID Publish Distribution
DAYS_ID_PUBLISH

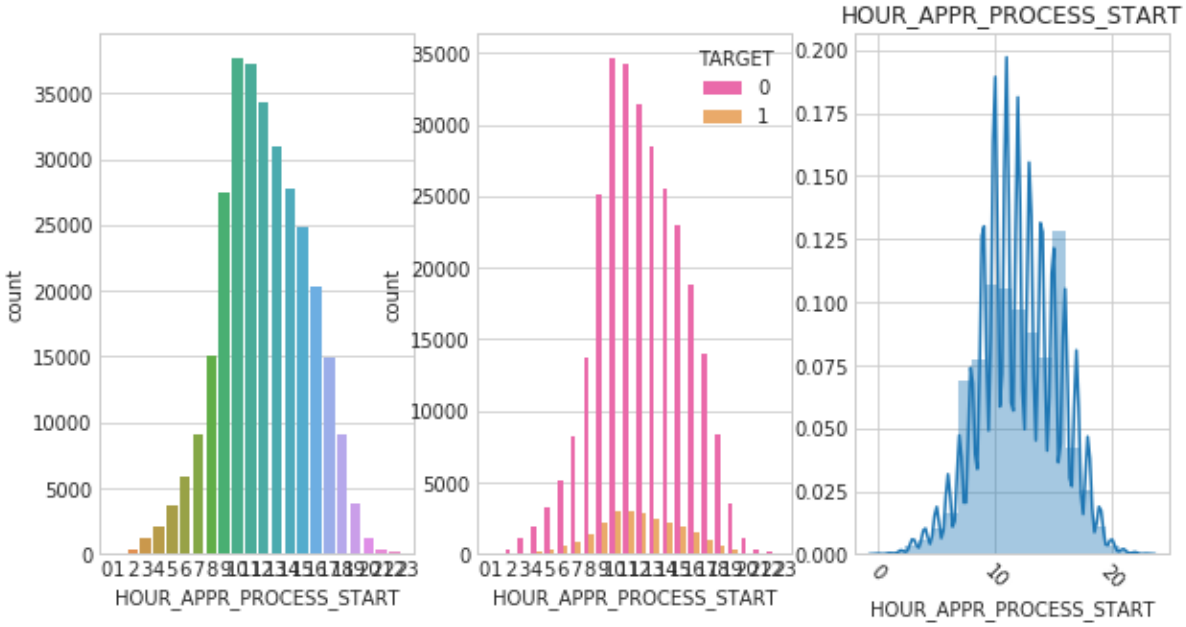

customer Days of Registration Distribution
DAYS_REGISTRATION

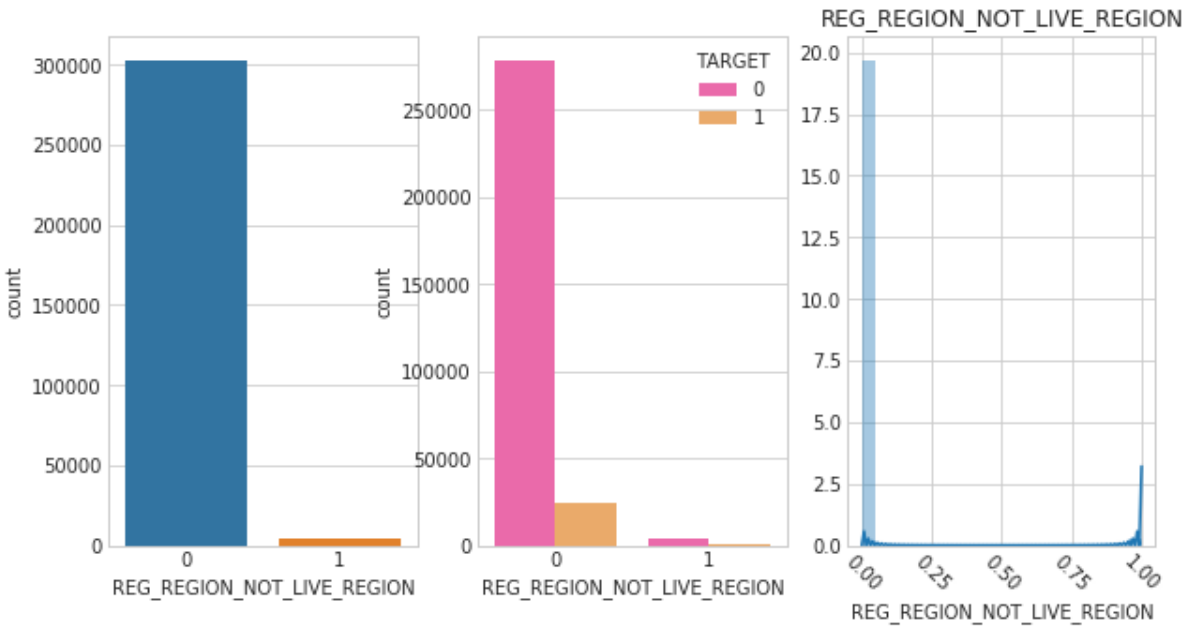

count
REGION_RATING_CLIENT


TARGET
0
1
count
REGION_RATING_CLIENT

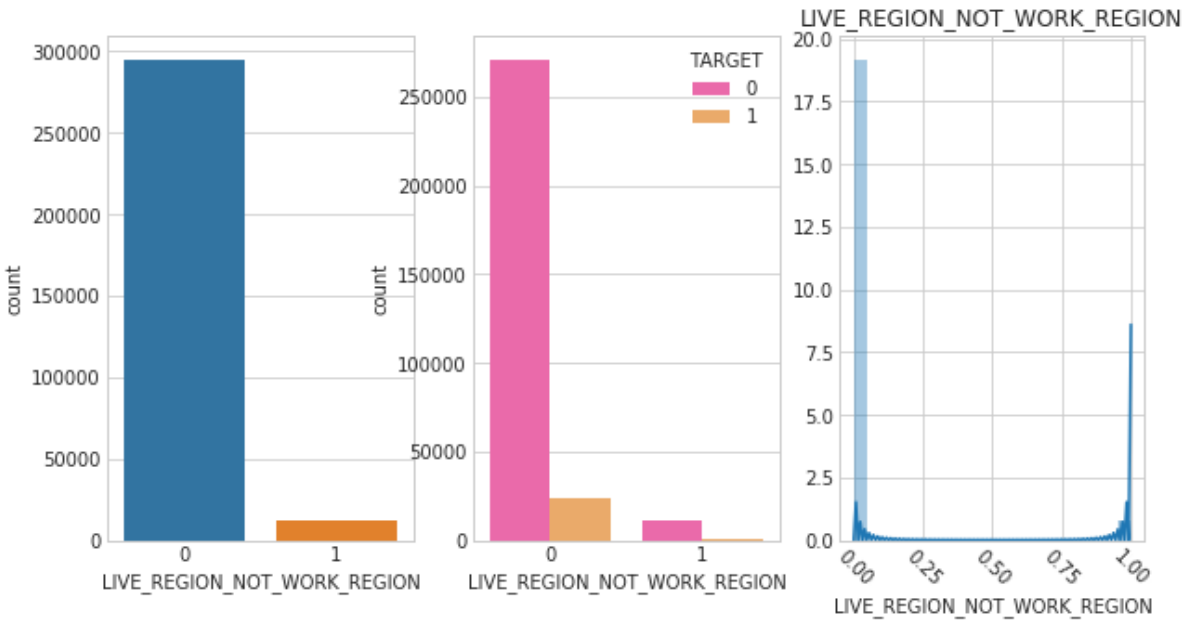

REGION_RATING_CLIENT
REGION_RATING_CLIENT

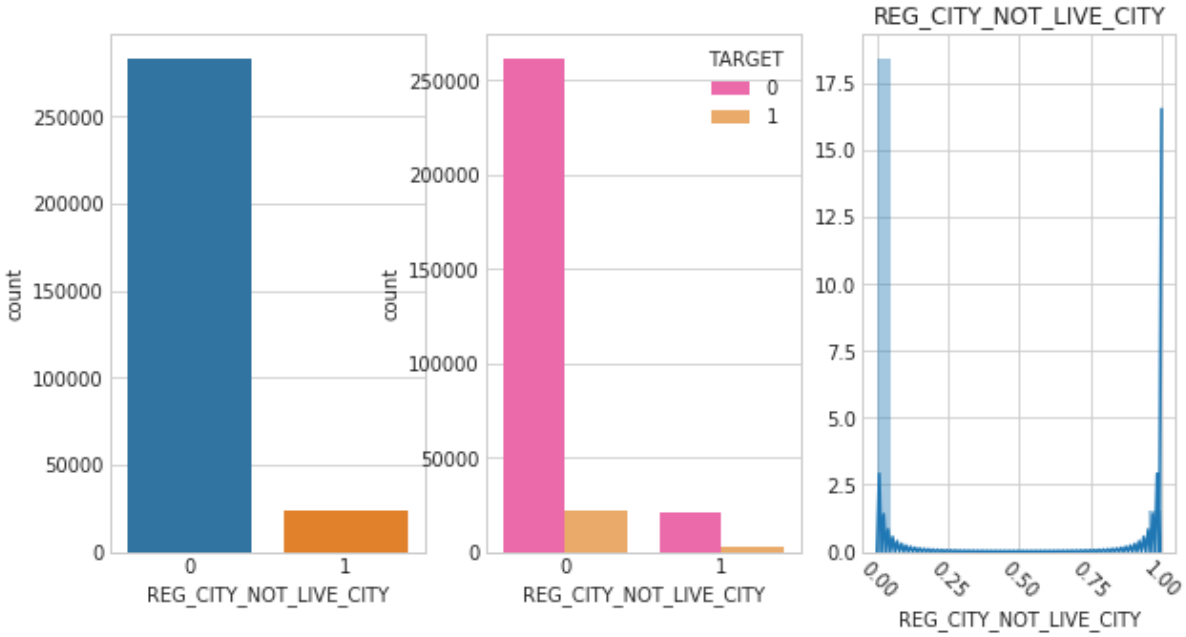

REGION_RATING_CLIENT_W_CITY
count
REGION_RATING_CLIENT_W_CITY
TARGET
0
1
count
REGION_RATING_CLIENT_W_CITY
REGION_RATING_CLIENT_W_CITY


HOUR_APPR_PROCESS_START
count
HOUR_APPR_PROCESS_START
TARGET
0
1
count
HOUR_APPR_PROCESS_START
HOUR_APPR_PROCESS_START


REG_REGION_NOT_LIVE_REGION
count
REG_REGION_NOT_LIVE_REGION
TARGET
0
1
count
REG_REGION_NOT_LIVE_REGION
REG_REGION_NOT_LIVE_REGION

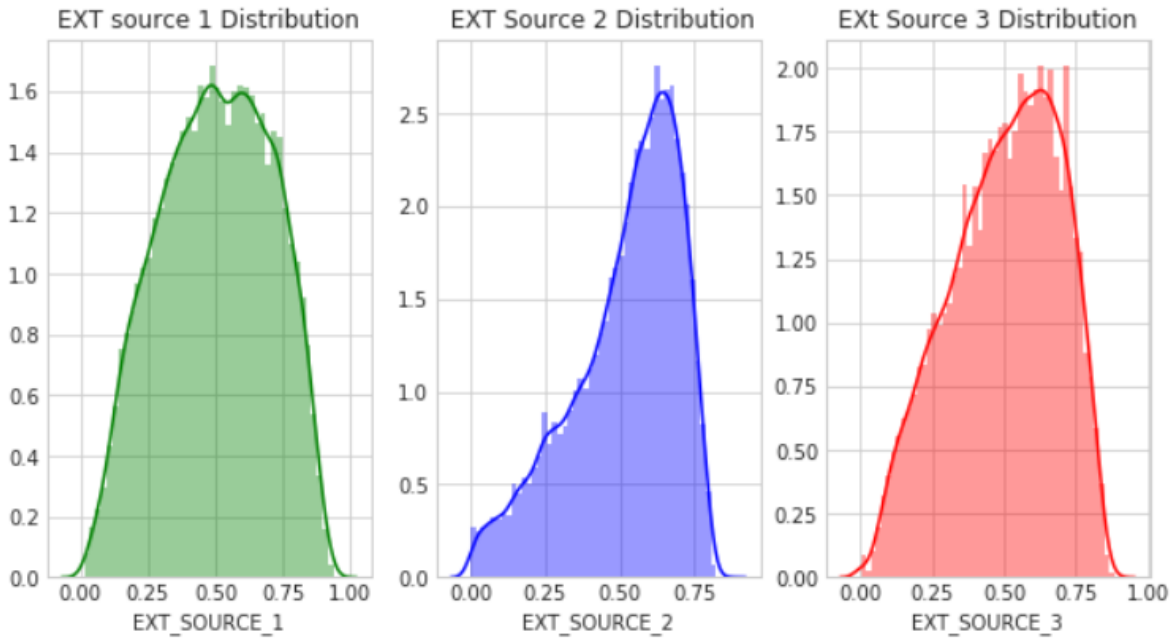

REG_REGION_NOT_WORK_REGION
count
REG_REGION_NOT_WORK_REGION
TARGET
0
1
count
REG_REGION_NOT_WORK_REGION
REG_REGION_NOT_WORK_REGION

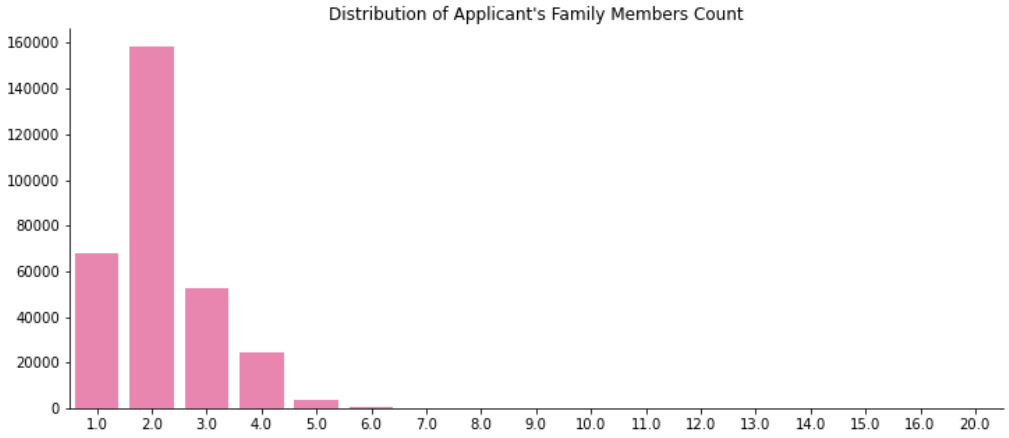

LIVE_REGION_NOT_WORK_REGION
count
LIVE_REGION_NOT_WORK_REGION
TARGET
0
1
count
LIVE_REGION_NOT_WORK_REGION
LIVE_REGION_NOT_WORK_REGION

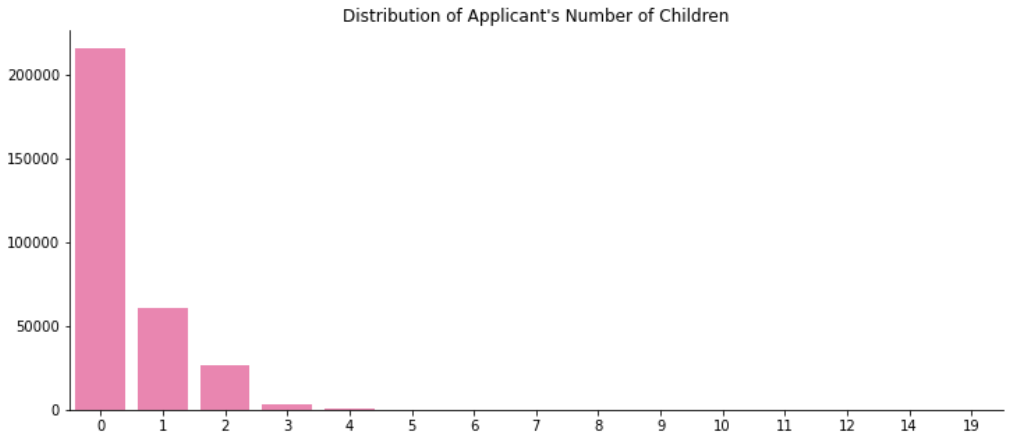

REG_CITY_NOT_LIVE_CITY
count
REG_CITY_NOT_LIVE_CITY
TARGET
0
1
count
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_LIVE_CITY



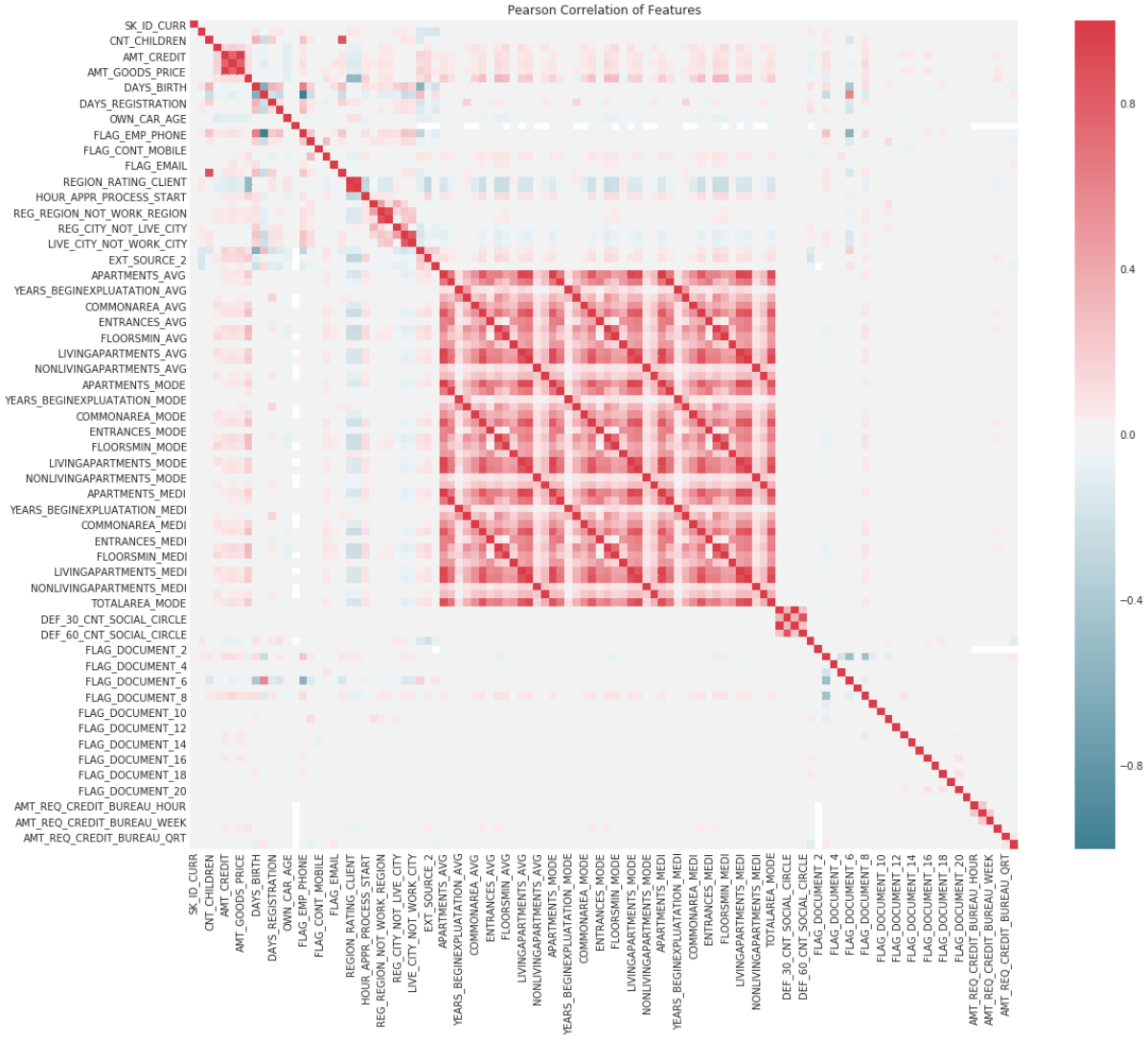
- EXT_SOURCE: Dữ liệu ẩn Kaggle cung cấp




- Pearson Correlation:



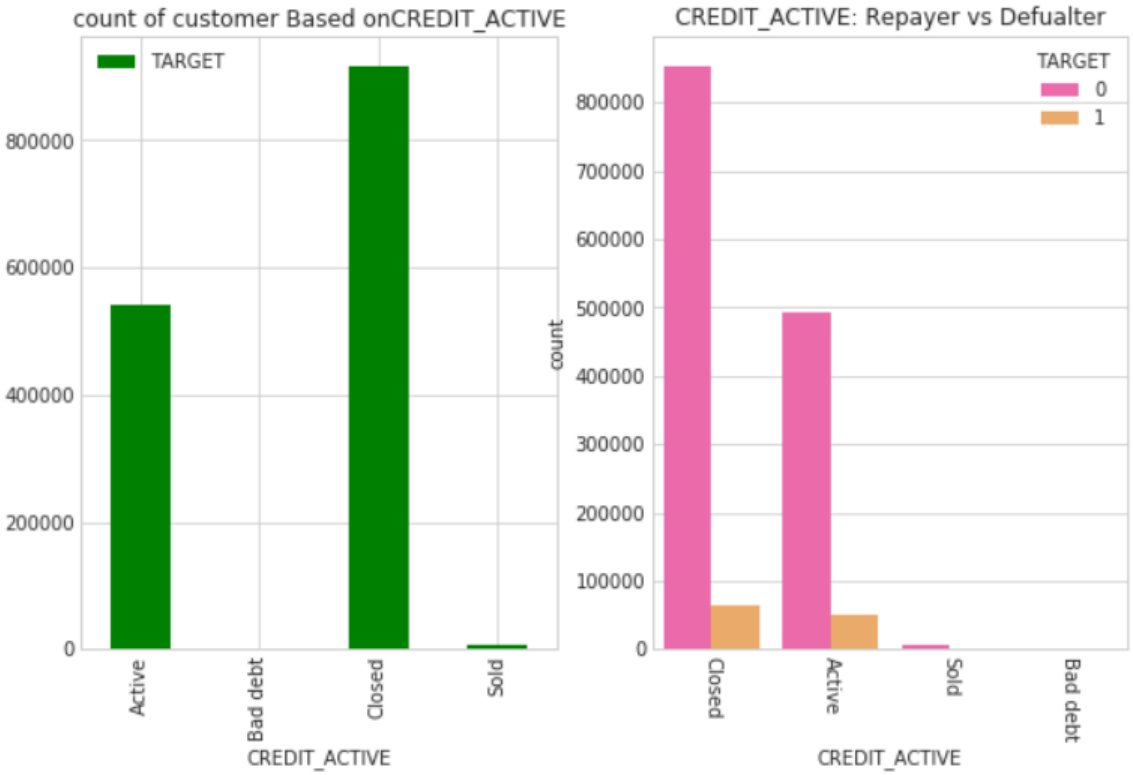

2. bureau.csv
   - Kích thước: (1716428, 17)
   - Số lượng các kiểu dữ liệu:
     - float64 8
     - int64 6
     - object 3
   - Số lượng missing data:

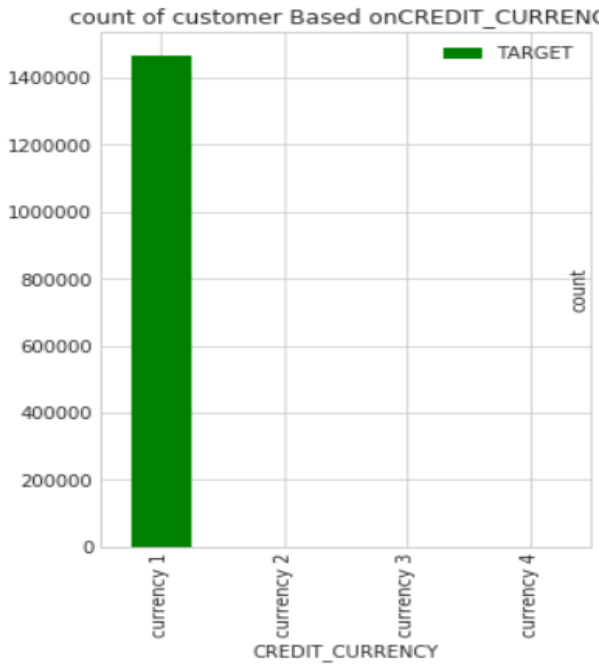
| | Missing Count | Missing Count Ratio | Missing Count % |
|---|---|---|---|
| AMT_ANNUITY | 1226791 | 0.714735 | 71.5 |
| AMT_CREDIT_MAX_OVERDUE | 1124488 | 0.655133 | 65.5 |
| DAYS_ENDDATE_FACT | 633653 | 0.36917 | 36.9 |
| AMT_CREDIT_SUM_LIMIT | 591780 | 0.344774 | 34.5 |
| AMT_CREDIT_SUM_DEBT | 257669 | 0.150119 | 15 |
| DAYS_CREDIT_ENDDATE | 105553 | 0.061496 | 6.1 |
| AMT_CREDIT_SUM | 13 | 7.57E-06 | 0 |

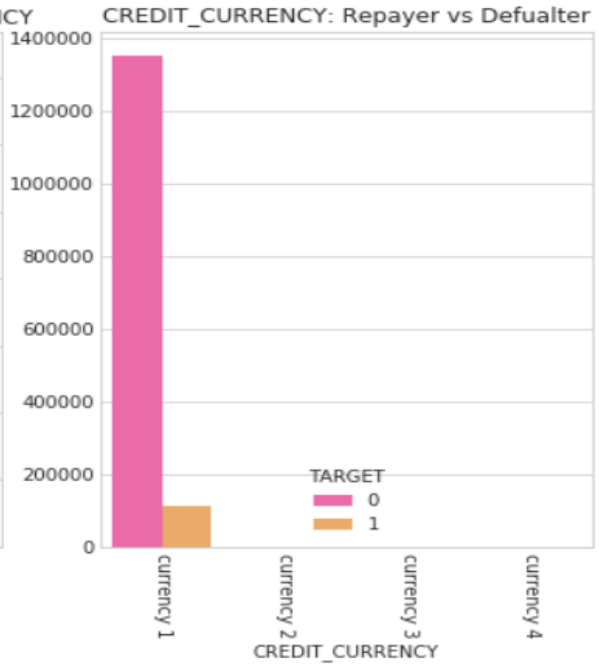
| | | | |
|---|---|---|---|
| CREDIT_TYPE | 0 | 0 | 0 |
| AMT_CREDIT_SUM_OVERDUE | 0 | 0 | 0 |
| CNT_CREDIT_PROLONG | 0 | 0 | 0 |
| DAYS_CREDIT_UPDATE | 0 | 0 | 0 |
| CREDIT_DAY_OVERDUE | 0 | 0 | 0 |
| DAYS_CREDIT | 0 | 0 | 0 |
| CREDIT_CURRENCY | 0 | 0 | 0 |
| CREDIT_ACTIVE | 0 | 0 | 0 |
| SK_ID_BUREAU | 0 | 0 | 0 |
| SK_ID_CURR | 0 | 0 | 0 |

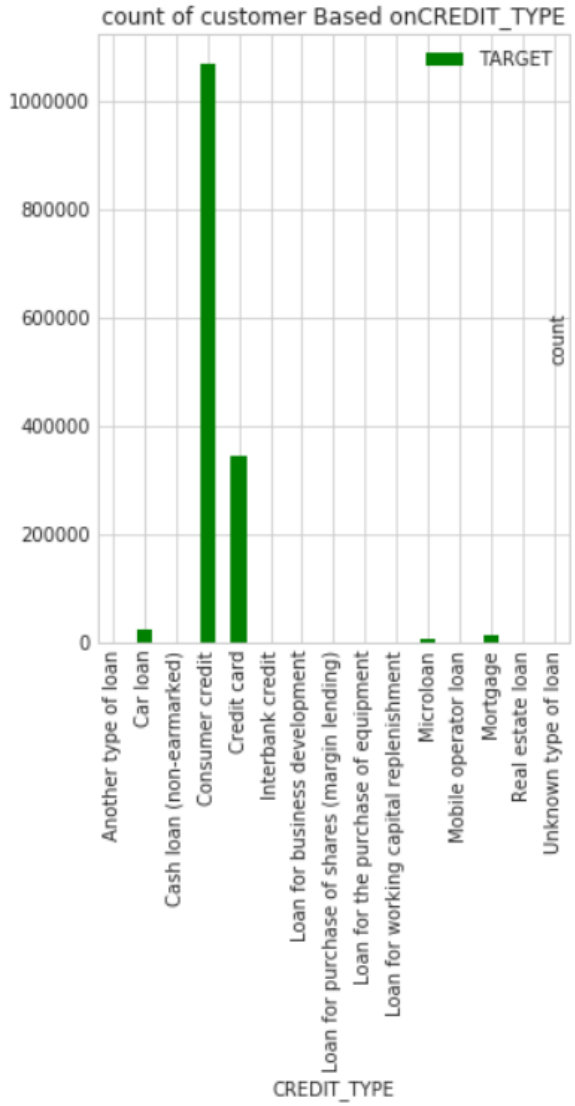
- Các dữ liệu Categorical:
  - CREDIT_ACTIVE: ['Closed', 'Active', 'Sold', 'Bad debt']
  - CREDIT_CURRENCY: ['currency 1', 'currency 2', 'currency 4', 'currency 3']
  - CREDIT_TYPE: ['Consumer credit', 'Credit card', 'Mortgage', 'Car loan', 'Microloan', 'Loan for working capital replenishment', 'Loan for business development', 'Real estate loan', 'Unknown type of loan', 'Another type of loan', 'Cash loan (non-earmarked)', 'Loan for the purchase of equipment', 'Mobile operator loan', 'Interbank credit', 'Loan for purchase of shares (margin lending)']

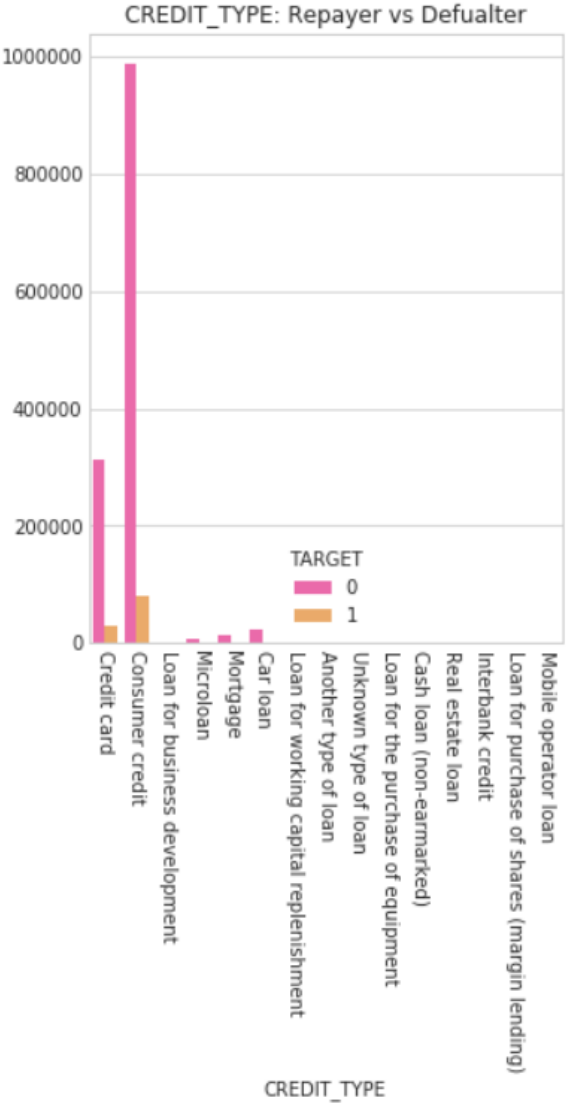
- CREDIT_ACTIVE:
  - Thực tế ta chỉ cần 2 giá trị Closed và Active là đủ



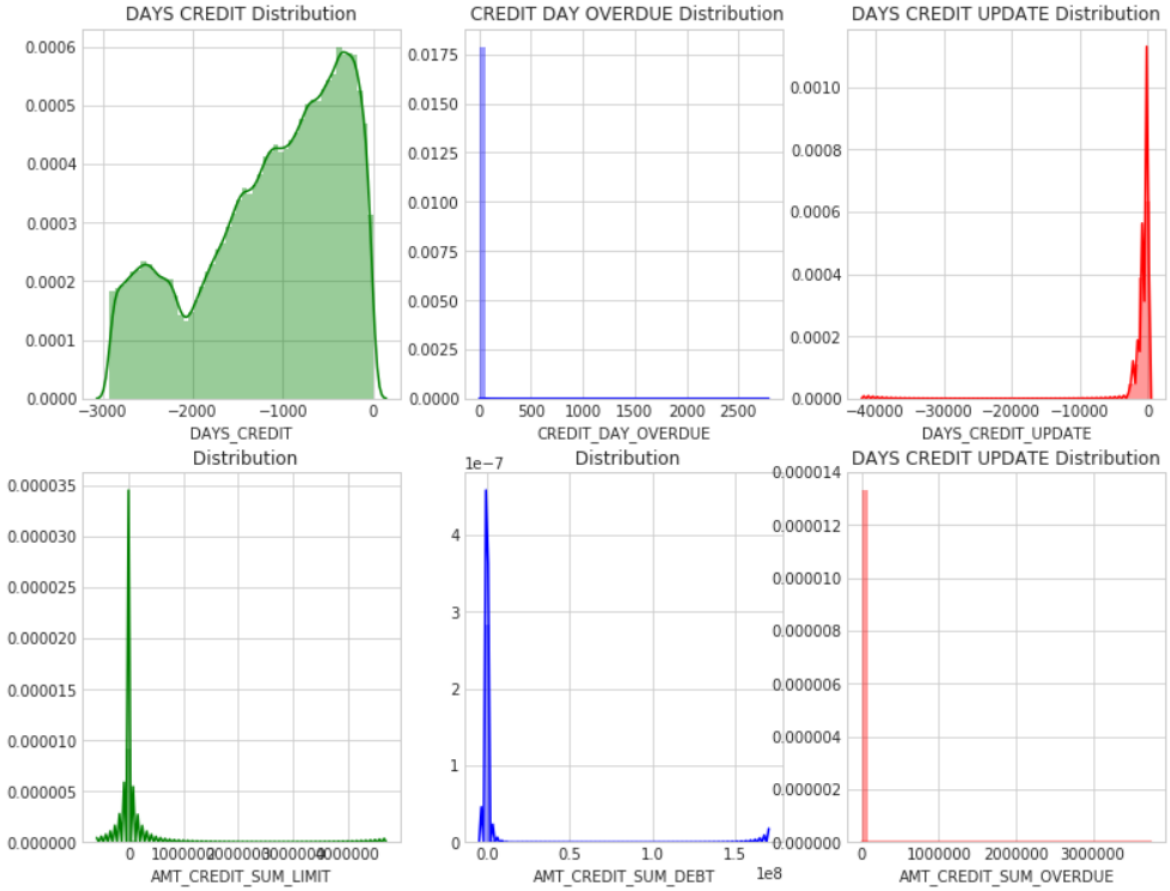

count of customer Based onCREDIT_CURRENCY
TARGET
1400000
1200000
1000000
800000
600000
400000
200000
0
count
currency 1
currency 2
currency 3
currency 4
CREDIT_CURRENCY


CREDIT_CURRENCY: Repayer vs Defualter
1400000
1200000
1000000
800000
600000
400000
200000
0
TARGET
0
1
currency 1
currency 2
currency 3
currency 4
CREDIT_CURRENCY


count of customer Based onCREDIT_TYPE
TARGET
1000000
800000
600000
400000
200000
0
count
Another type of loan
Car loan
Cash loan (non-earmarked)
Consumer credit
Credit card
Interbank credit
Loan for business development
Loan for purchase of shares (margin lending)
Loan for the purchase of equipment
Loan for working capital replenishment
Microloan
Mobile operator loan
Mortgage
Real estate loan
Unknown type of loan
CREDIT_TYPE


CREDIT_TYPE: Repayer vs Defualter
1000000
800000
600000
400000
200000
0
TARGET
0
1
Credit card
Consumer credit
Loan for business development
Microloan
Mortgage
Car loan
Loan for working capital replenishment
Another type of loan
Unknown type of loan
Loan for the purchase of equipment
Cash loan (non-earmarked)
Real estate loan
Interbank credit
Loan for purchase of shares (margin lending)
Mobile operator loan
CREDIT_TYPE

- Các dữ liệu Numerical:



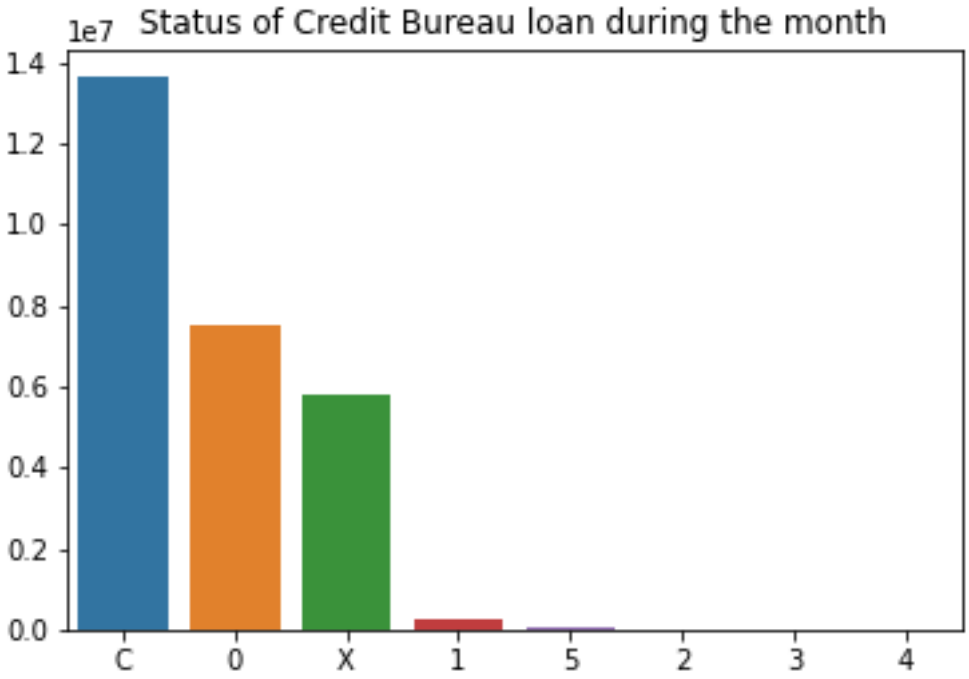

3. bureau_balance.csv
   - Kích thước: (27299925, 3)
   - Số lượng các kiểu dữ liệu:
     - int64 2
     - object 1
   - Số lượng missing data:

| | Missing Count | Missing Count Ratio | Missing Count % |
|---|---|---|---|
| STATUS | 0 | 0 | 0 |
| MONTHS_BALANCE | 0 | 0 | 0 |
| SK_ID_BUREAU | 0 | 0 | 0 |

   - Snapshot:

| SK_ID_BUREAU | MONTHS_ BALANCE | STATUS |
|---|---|---|
| 5715448 | 0 | C |
| 5715448 | -1 | C |
| 5715448 | -2 | C |
| 5715448 | -3 | C |

| | | |
|---|---|---|
| 5715448 | -4 | C |




4. POS_CASH_balance.csv
   - Kích thước: (10001358, 8)
   - Số lượng các kiểu dữ liệu:
     - int64 5
     - float64 2
     - object 1
   - Số lượng missing data:

| | Missing Count | Missing Count Ratio | Missing Count % |
|---|---|---|---|
| CNT_INSTALMENT_FUTURE | 26087 | 0.002608 | 0.3 |
| CNT_INSTALMENT | 26071 | 0.002607 | 0.3 |
| SK_DPD_DEF | 0 | 0 | 0 |
| SK_DPD | 0 | 0 | 0 |
| NAME_CONTRACT_STATUS | 0 | 0 | 0 |
| MONTHS_BALANCE | 0 | 0 | 0 |
| SK_ID_CURR | 0 | 0 | 0 |
| SK_ID_PREV | 0 | 0 | 0 |

5. credit_card_balance.csv
   - Kích thước: (3840312, 23)
   - Số lượng các kiểu dữ liệu:
     - float64 15
     - int64 7
     - object 1

- Số lượng missing data:

| | Missing Count | Missing Count Ratio | Missing Count % |
|---|---|---|---|
| AMT_PAYMENT_CURRENT | 767988 | 0.199981 | 20 |
| AMT_DRAWINGS_OTHER_CURRENT | 749816 | 0.195249 | 19.5 |
| CNT_DRAWINGS_POS_CURRENT | 749816 | 0.195249 | 19.5 |
| CNT_DRAWINGS_OTHER_CURRENT | 749816 | 0.195249 | 19.5 |
| CNT_DRAWINGS_ATM_CURRENT | 749816 | 0.195249 | 19.5 |
| AMT_DRAWINGS_ATM_CURRENT | 749816 | 0.195249 | 19.5 |
| AMT_DRAWINGS_POS_CURRENT | 749816 | 0.195249 | 19.5 |
| CNT_INSTALMENT_MATURE_CUM | 305236 | 0.079482 | 7.9 |
| AMT_INST_MIN_REGULARITY | 305236 | 0.079482 | 7.9 |
| SK_DPD_DEF | 0 | 0 | 0 |
| SK_ID_CURR | 0 | 0 | 0 |
| MONTHS_BALANCE | 0 | 0 | 0 |
| AMT_BALANCE | 0 | 0 | 0 |
| AMT_CREDIT_LIMIT_ACTUAL | 0 | 0 | 0 |
| AMT_DRAWINGS_CURRENT | 0 | 0 | 0 |
| AMT_PAYMENT_TOTAL_CURRENT | 0 | 0 | 0 |
| SK_DPD | 0 | 0 | 0 |
| AMT_RECEIVABLE_PRINCIPAL | 0 | 0 | 0 |
| AMT_RECIVABLE | 0 | 0 | 0 |
| AMT_TOTAL_RECEIVABLE | 0 | 0 | 0 |
| CNT_DRAWINGS_CURRENT | 0 | 0 | 0 |
| NAME_CONTRACT_STATUS | 0 | 0 | 0 |
| SK_ID_PREV | 0 | 0 | 0 |

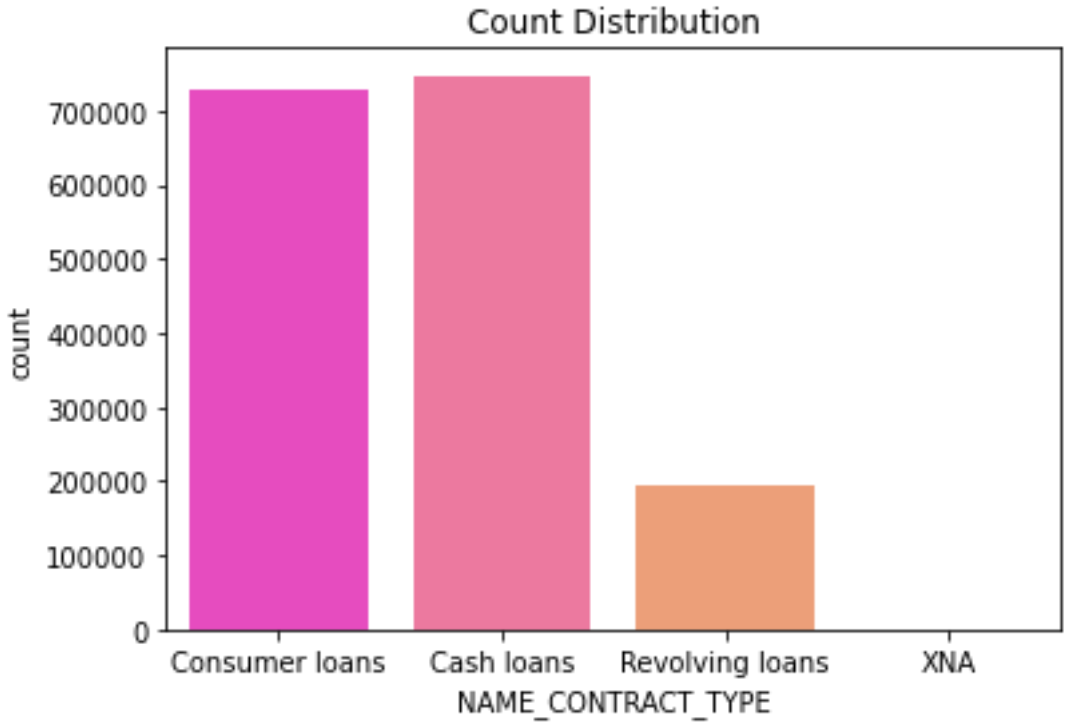
6. previous_application.csv
   - Kích thước: (1670214, 37)
   - Số lượng các kiểu dữ liệu:
     - object 16
     - float64 15
     - int64 6
   - Số lượng missing data:

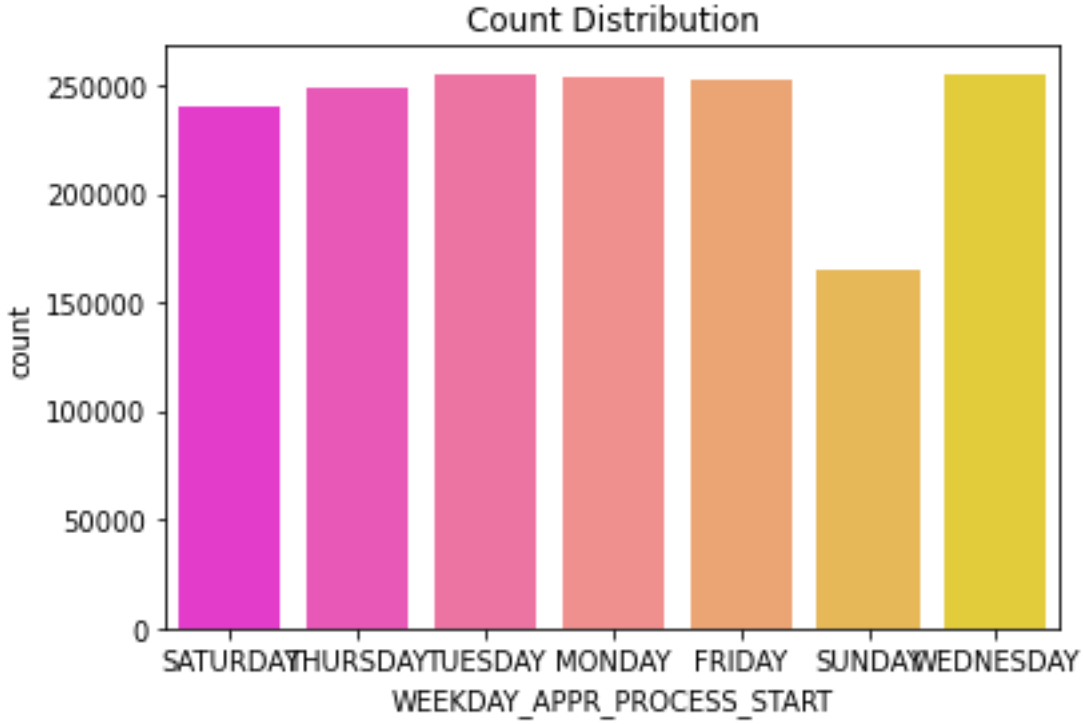
| | Missing Count | Missing Count Ratio | Missing Count % |
|---|---|---|---|
| RATE_INTEREST_PRIVILEGED | 1664263 | 0.996437 | 99.6 |
| RATE_INTEREST_PRIMARY | 1664263 | 0.996437 | 99.6 |
| RATE_DOWN_PAYMENT | 895844 | 0.536365 | 53.6 |
| AMT_DOWN_PAYMENT | 895844 | 0.536365 | 53.6 |
| NAME_TYPE_SUITE | 820405 | 0.491198 | 49.1 |
| DAYS_TERMINATION | 673065 | 0.402981 | 40.3 |
| NFLAG_INSURED_ON_APPROVAL | 673065 | 0.402981 | 40.3 |

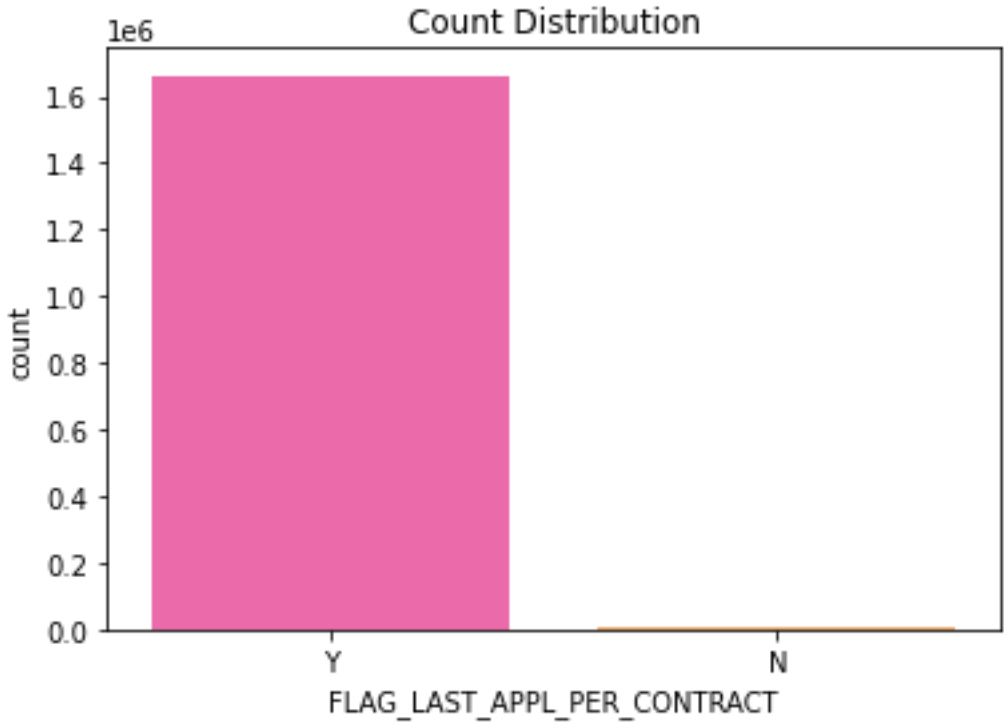
| | | | |
|---|---|---|---|
| DAYS_FIRST_DRAWING | 673065 | 0.402981 | 40.3 |
| DAYS_FIRST_DUE | 673065 | 0.402981 | 40.3 |
| DAYS_LAST_DUE_1ST_VERSION | 673065 | 0.402981 | 40.3 |
| DAYS_LAST_DUE | 673065 | 0.402981 | 40.3 |
| AMT_GOODS_PRICE | 385515 | 0.230818 | 23.1 |
| AMT_ANNUITY | 372235 | 0.222867 | 22.3 |
| CNT_PAYMENT | 372230 | 0.222864 | 22.3 |
| PRODUCT_COMBINATION | 346 | 0.000207 | 0 |
| AMT_CREDIT | 1 | 5.99E-07 | 0 |
| SK_ID_CURR | 0 | 0 | 0 |
| NAME_CONTRACT_TYPE | 0 | 0 | 0 |
| WEEKDAY_APPR_PROCESS_START | 0 | 0 | 0 |
| HOUR_APPR_PROCESS_START | 0 | 0 | 0 |
| FLAG_LAST_APPL_PER_CONTRACT | 0 | 0 | 0 |
| NFLAG_LAST_APPL_IN_DAY | 0 | 0 | 0 |
| AMT_APPLICATION | 0 | 0 | 0 |
| NAME_PAYMENT_TYPE | 0 | 0 | 0 |
| NAME_CASH_LOAN_PURPOSE | 0 | 0 | 0 |
| NAME_CONTRACT_STATUS | 0 | 0 | 0 |
| DAYS_DECISION | 0 | 0 | 0 |
| CODE_REJECT_REASON | 0 | 0 | 0 |
| NAME_CLIENT_TYPE | 0 | 0 | 0 |
| NAME_GOODS_CATEGORY | 0 | 0 | 0 |
| NAME_PORTFOLIO | 0 | 0 | 0 |
| NAME_PRODUCT_TYPE | 0 | 0 | 0 |
| CHANNEL_TYPE | 0 | 0 | 0 |
| SELLERPLACE_AREA | 0 | 0 | 0 |
| NAME_SELLER_INDUSTRY | 0 | 0 | 0 |
| NAME_YIELD_GROUP | 0 | 0 | 0 |
| SK_ID_PREV | 0 | 0 | 0 |

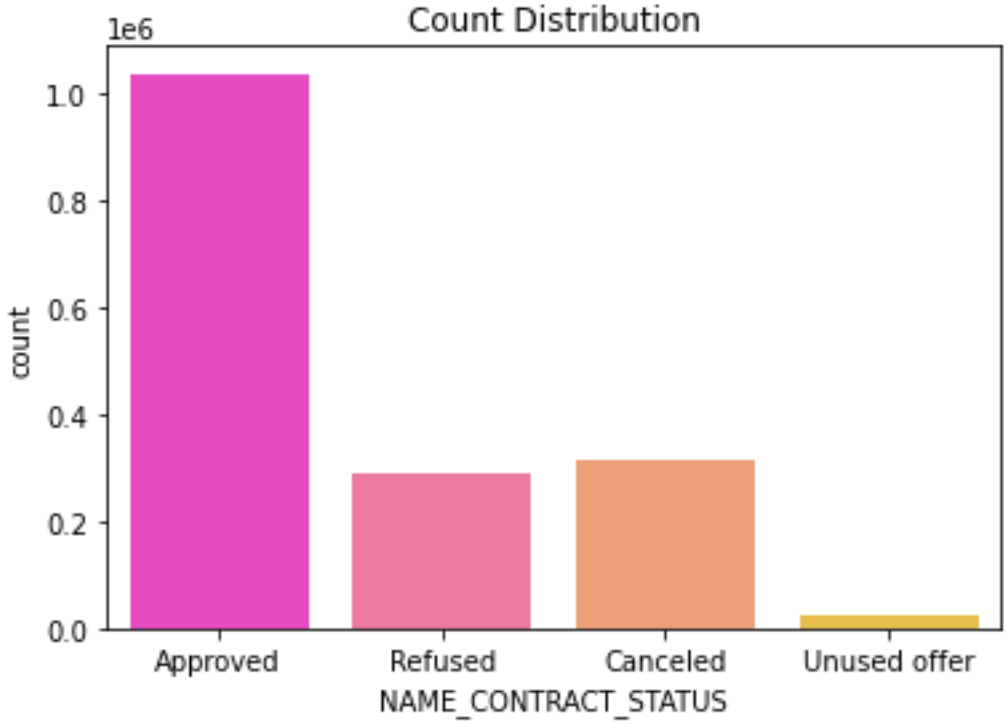
- Các dữ liệu Categorical:
  - NAME_CONTRACT_TYPE: ['Consumer loans', 'Cash loans', 'Revolving loans', 'XNA']
  - WEEKDAY_APPR_PROCESS_START: ['SATURDAY', 'THURSDAY', 'TUESDAY', 'MONDAY', 'FRIDAY', 'SUNDAY', 'WEDNESDAY']
  - FLAG_LAST_APPL_PER_CONTRACT: ['Y', 'N']
  - NAME_CASH_LOAN_PURPOSE: ['XAP', 'XNA', 'Repairs', 'Everyday expenses', 'Car repairs', 'Building a house or an annex', 'Other', 'Journey', 'Purchase of electronic equipment', 'Medicine', 'Payments on other loans', 'Urgent needs', 'Buying a used car', 'Buying a new car', 'Buying a holiday home / land', 'Education', 'Buying a home', 'Furniture', 'Buying a garage', 'Business development', 'Wedding / gift / holiday', 'Hobby', 'Gasification / water supply', 'Refusal to name the goal', 'Money for a third person']
  - NAME_CONTRACT_STATUS: ['Approved', 'Refused', 'Canceled', 'Unused offer']

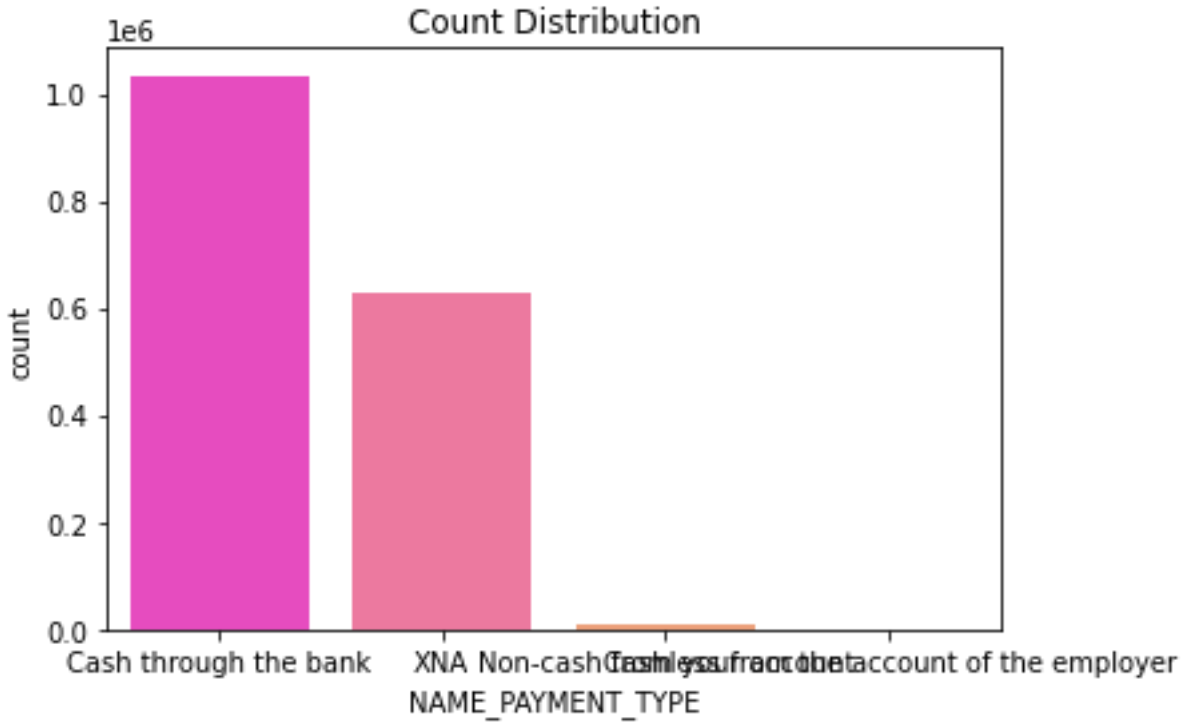
- NAME_PAYMENT_TYPE: ['Cash through the bank', 'XNA', 'Non-cash from your account', 'Cashless from the account of the employer']
- CODE_REJECT_REASON: ['XAP', 'HC', 'LIMIT', 'CLIENT', 'SCOFR', 'SCO', 'XNA', 'VERIF', 'SYSTEM']
- NAME_TYPE_SUITE: [nan, 'Unaccompanied', 'Spouse, partner', 'Family', 'Children', 'Other_B', 'Other_A', 'Group of people']
- NAME_CLIENT_TYPE: ['Repeater', 'New', 'Refreshed', 'XNA']
- NAME_GOODS_CATEGORY: ['Mobile', 'XNA', 'Consumer Electronics', 'Construction Materials', 'Auto Accessories', 'Photo / Cinema Equipment', 'Computers', 'Audio/Video', 'Medicine', 'Clothing and Accessories', 'Furniture', 'Sport and Leisure', 'Homewares', 'Gardening', 'Jewelry', 'Vehicles', 'Education', 'Medical Supplies', 'Other', 'Direct Sales', 'Office Appliances', 'Fitness', 'Tourism', 'Insurance', 'Additional Service', 'Weapon', 'Animals', 'House Construction']
- NAME_PORTFOLIO: ['POS', 'Cash', 'XNA', 'Cards', 'Cars']
- NAME_PRODUCT_TYPE: ['XNA', 'x-sell', 'walk-in']
- CHANNEL_TYPE: ['Country-wide', 'Contact center', 'Credit and cash offices', 'Stone', 'Regional / Local', 'AP+ (Cash loan)', 'Channel of corporate sales', 'Car dealer']
- NAME_SELLER_INDUSTRY: ['Connectivity', 'XNA', 'Consumer electronics', 'Industry', 'Clothing', 'Furniture', 'Construction', 'Jewelry', 'Auto technology', 'MLM partners', 'Tourism']
- NAME_YIELD_GROUP: ['middle', 'low_action', 'high', 'low_normal', 'XNA']
- PRODUCT_COMBINATION: ['POS mobile with interest', 'Cash X-Sell: low', 'Cash X-Sell: high', 'Cash X-Sell: middle', 'Cash Street: high', 'Cash', 'POS household without interest', 'POS household with interest', 'POS other with interest', 'Card X-Sell', 'POS mobile without interest', 'Card Street', 'POS industry with interest', 'Cash Street: low', 'POS industry without interest', 'Cash Street: middle', 'POS others without interest', nan]


Count Distribution
700000
600000
500000
400000
300000
200000
100000
0
count
Consumer loans
Cash loans
Revolving loans
XNA
NAME_CONTRACT_TYPE

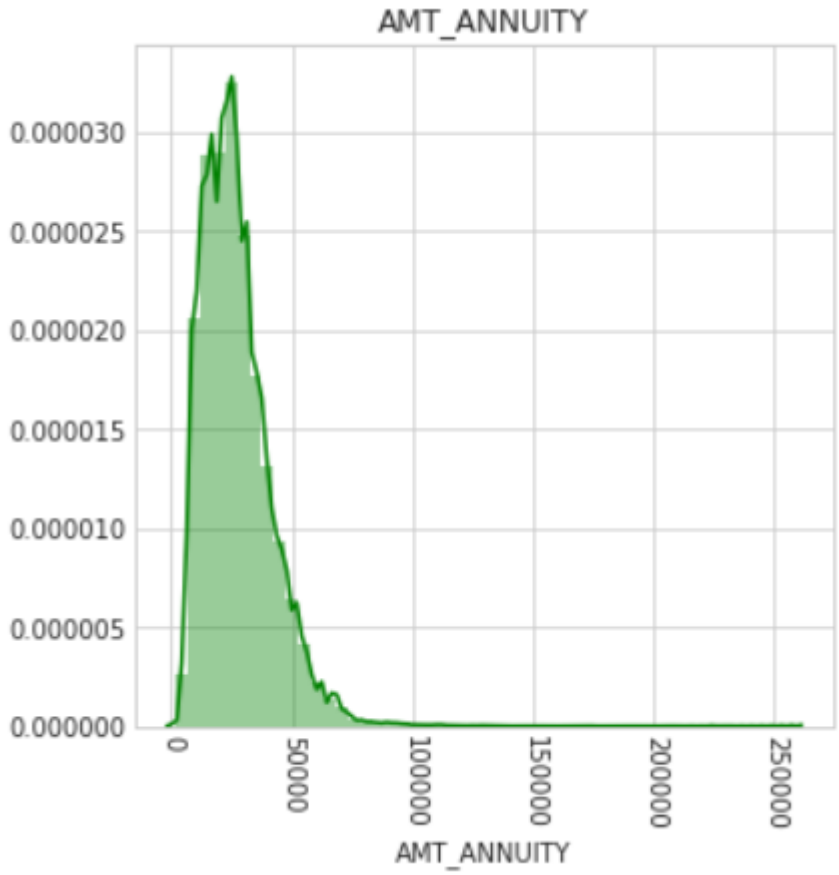

Count Distribution
250000
200000
150000
100000
50000
0
count
SATURDAY
THURSDAY
TUESDAY
MONDAY
FRIDAY
SUNDAY
WEDNESDAY
WEEKDAY_APPR_PROCESS_START

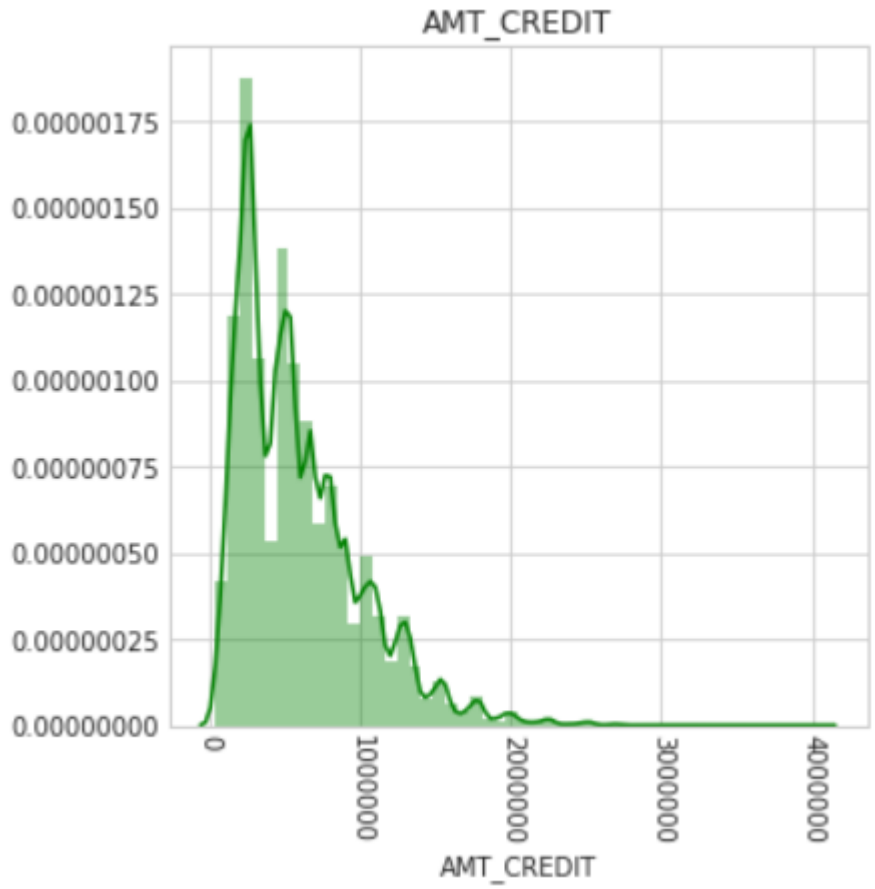

Count Distribution
1e6
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
count
Y
N
FLAG_LAST_APPL_PER_CONTRACT

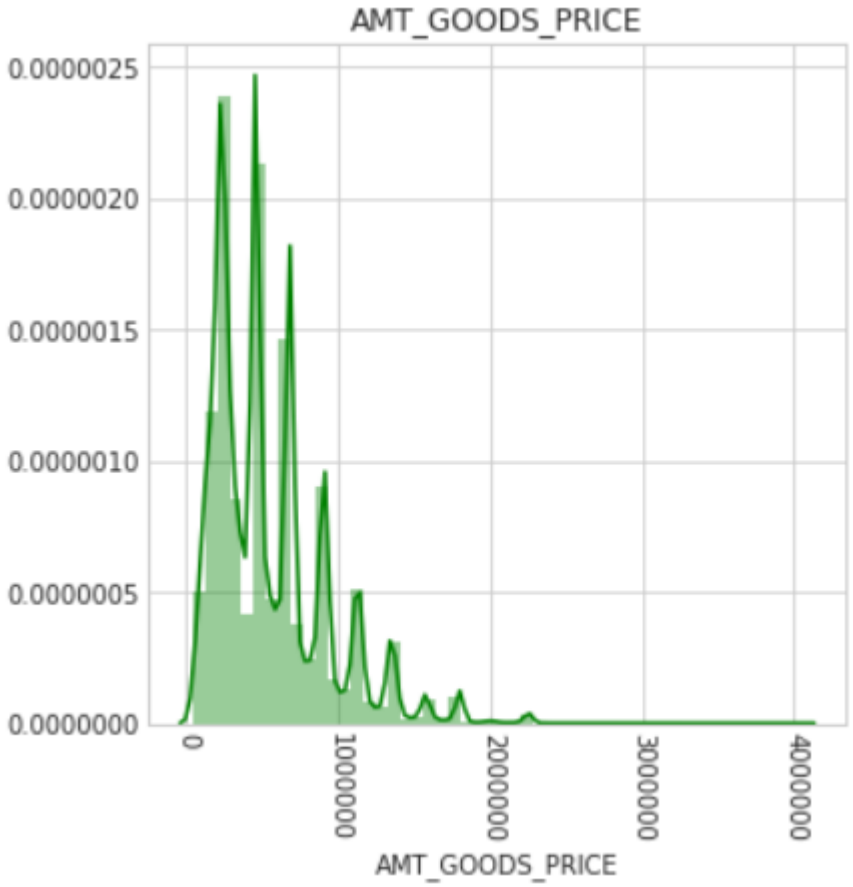

Count Distribution
1e6
1.0
0.8
0.6
0.4
0.2
0.0
count
Approved
Refused
Canceled
Unused offer
NAME_CONTRACT_STATUS

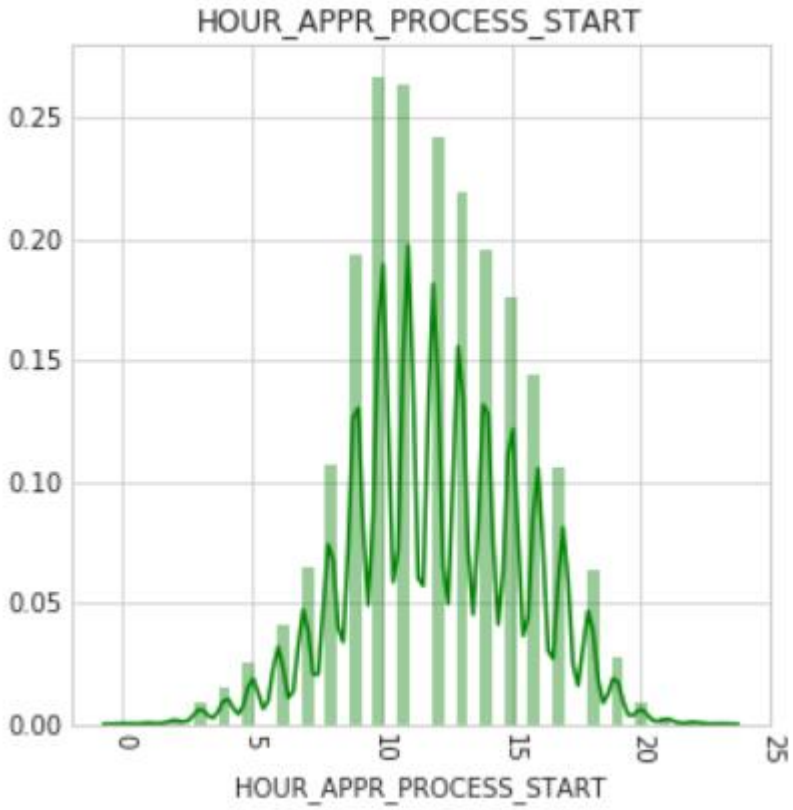



- Các dữ liệu Numerical:




AMT_CREDIT
0.00000175
0.00000150
0.00000125
0.00000100
0.00000075
0.00000050
0.00000025
0.00000000
0
1000000
2000000
3000000
4000000
AMT_CREDIT


AMT_GOODS_PRICE
0.0000025
0.0000020
0.0000015
0.0000010
0.0000005
0.0000000
0
1000000
2000000
3000000
4000000
AMT_GOODS_PRICE

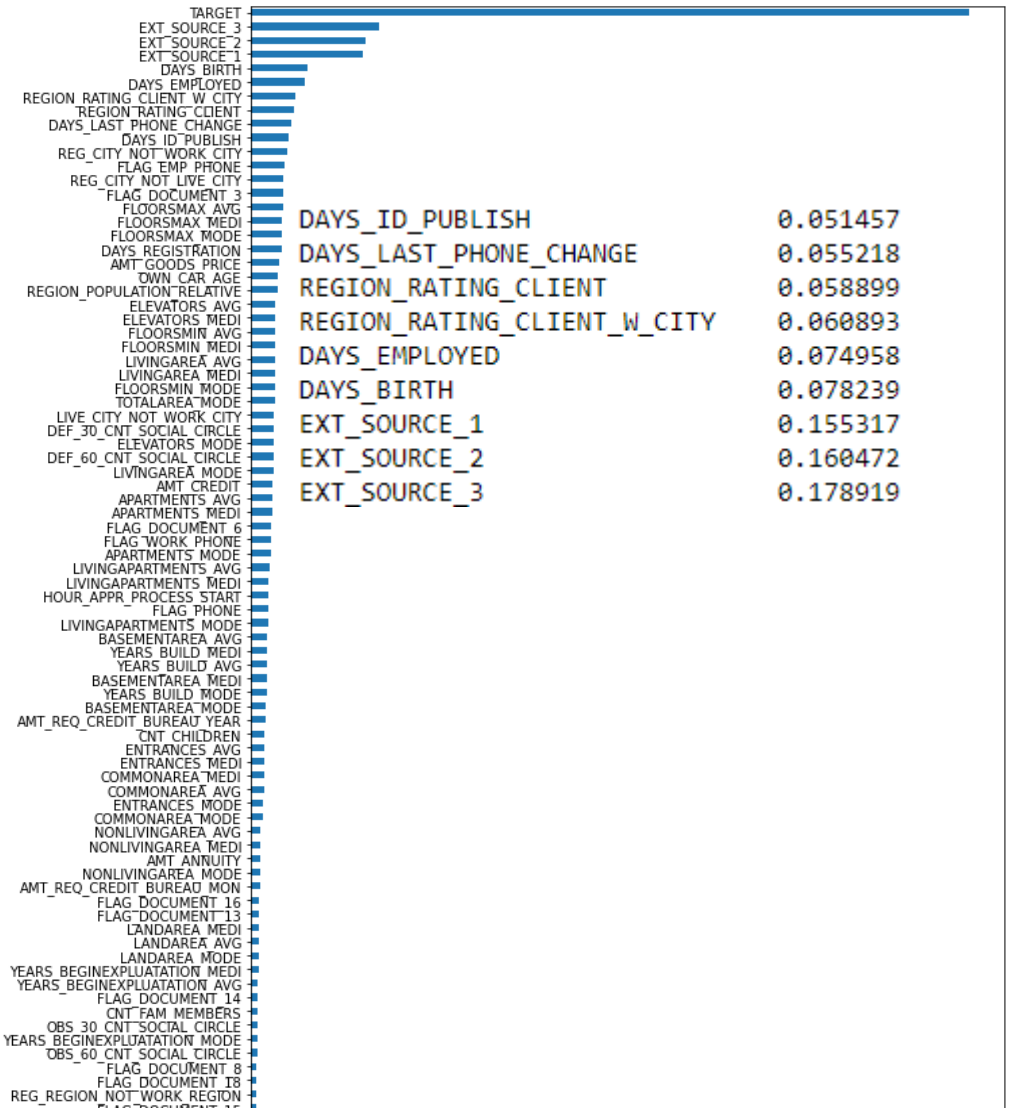



7. installments_payments.csv
   - Kích thước: (13605401, 8)
   - Số lượng các kiểu dữ liệu:
     - float64 5
     - int64 3
   - Số lượng missing data:

| | Missing Count | Missing Count Ratio | Missing Count % |
|---|---|---|---|
| AMT_PAYMENT | 2905 | 0.000214 | 0 |
| DAYS_ENTRY_PAYMENT | 2905 | 0.000214 | 0 |
| AMT_INSTALMENT | 0 | 0 | 0 |
| DAYS_INSTALMENT | 0 | 0 | 0 |
| NUM_INSTALMENT_NUMBER | 0 | 0 | 0 |
| NUM_INSTALMENT_VERSION | 0 | 0 | 0 |
| SK_ID_CURR | 0 | 0 | 0 |
| SK_ID_PREV | 0 | 0 | 0 |

## Phần III. Kết quả

1. application_train.csv
   - Top những feature có Pearson Correlation cao nhất với TARGET




- Top những feature có lượng missing lớn nhất

| | Missing Count | Missing Count Ratio | Missing Count % |
|---|---|---|---|
| COMMONAREA_MEDI | 248360 | 0.69714 | 69.7 |
| COMMONAREA_MODE | 248360 | 0.69714 | 69.7 |
| COMMONAREA_AVG | 248360 | 0.69714 | 69.7 |
| NONLIVINGAPARTMENTS_AVG | 246861 | 0.69293 | 69.3 |
| NONLIVINGAPARTMENTS_MODE | 246861 | 0.69293 | 69.3 |
| NONLIVINGAPARTMENTS_MEDI | 246861 | 0.69293 | 69.3 |
| FONDKAPREMONT_MODE | 243092 | 0.68235 | 68.2 |
| LIVINGAPARTMENTS_MEDI | 242979 | 0.68204 | 68.2 |
| LIVINGAPARTMENTS_MODE | 242979 | 0.68204 | 68.2 |
| LIVINGAPARTMENTS_AVG | 242979 | 0.68204 | 68.2 |
| FLOORSMIN_MEDI | 241108 | 0.67678 | 67.7 |
| FLOORSMIN_MODE | 241108 | 0.67678 | 67.7 |
| FLOORSMIN_AVG | 241108 | 0.67678 | 67.7 |
| YEARS_BUILD_MEDI | 236306 | 0.66331 | 66.3 |

| | | | |
|---|---|---|---|
| YEARS_BUILD_AVG | 236306 | 0.66331 | 66.3 |
| YEARS_BUILD_MODE | 236306 | 0.66331 | 66.3 |
| OWN_CAR_AGE | 235241 | 0.66032 | 66 |
| LANDAREA_MEDI | 210844 | 0.59183 | 59.2 |
| LANDAREA_AVG | 210844 | 0.59183 | 59.2 |
| LANDAREA_MODE | 210844 | 0.59183 | 59.2 |
| BASEMENTAREA_AVG | 207584 | 0.58268 | 58.3 |
| BASEMENTAREA_MODE | 207584 | 0.58268 | 58.3 |
| BASEMENTAREA_MEDI | 207584 | 0.58268 | 58.3 |
| NONLIVINGAREA_AVG | 195766 | 0.54951 | 55 |
| NONLIVINGAREA_MEDI | 195766 | 0.54951 | 55 |
| NONLIVINGAREA_MODE | 195766 | 0.54951 | 55 |
| EXT_SOURCE_1 | 193910 | 0.5443 | 54.4 |
| ELEVATORS_MEDI | 189080 | 0.53074 | 53.1 |
| ELEVATORS_MODE | 189080 | 0.53074 | 53.1 |
| ELEVATORS_AVG | 189080 | 0.53074 | 53.1 |

- Do EXT_SOURCE_1 có quan hệ Pearson mạnh với biến TARGET nên ta sẽ giữ lại và là 1 trong những feature quan trọng
- CODE_GENDER có giá trị 'XNA' nhiễu, ta sẽ bỏ những bản ghi có chứa giá trị này
  - Chỉ có 4 bản ghi có chứa giá trị này (EDA đã làm ở trên)
  - Cả 4 bản ghi đều nằm ở file application_train
- CNT_CHILDREN có những giá trị NaN có thể fill bằng 0 vì những giá trị đó có thể khi khách hàng điền thì họ không biết điền gì vì chưa có con và 0 cũng là giá trị phù hợp để fill NaN
- CNT_FAM_MEMBERS có những giá trị NaN có thể fill bằng 1 cũng giống với lập luận ở trên
- Các feature với tiền tố 'DAY' có giá trị $\leq 0$
  - Ví dụ như DAY_BIRTH: số ngày họ được sinh ra đếm ngược từ lúc đi vay
  - DAY_EMPLOYED có giá trị 365243 hiển nhiên là nhiễu, ta sẽ để những bản ghi có giá trị này là NaN vì số lượng cũng khá lớn (khoảng 18%)
  - Ngoài ra ta sẽ chia các giá trị của các feature DAY cho -365 cho tiện
- Ta thấy phân bố của các feature Numerical skewed, nên ta sẽ fill những dữ liệu missing bằng median, fill tác biệt giữa 2 nhãn và tách tập val và test ra để tránh bị data leak
- Ta dễ thấy là 26 feature có lượng missing cao nhất thì có ít quan hệ Pearson với biến TARGET. Vì vậy nên có thể cân nhắc bỏ đi được
- Kết quả trước và sau khi bỏ 26 cột trên
  - Đỏ: trước khi bỏ
  - Xanh: sau khi bỏ

| Mô hình \ Tập đánh giá | Val set tách từ train set (30%) | | Test set của Kaggle | |
|---|---|---|---|---|
| Logistic Regression | 0.7461 | 0.7459 | 0.7385 | 0.7384 |
| Naïve Bayes Classifier | 0.5349 | 0.5317 | 0.5364 | 0.5369 |
| XGBoost Classifier (n_estimators=250, max_depth=5) | 0.7581 | 0.7590 | 0.7476 | 0.7476 |
| Neural Network (1 hidden layer 64 nodes 15 epochs Adam) | 0.7243 | 0.7378 | 0.7269 | 0.7385 |

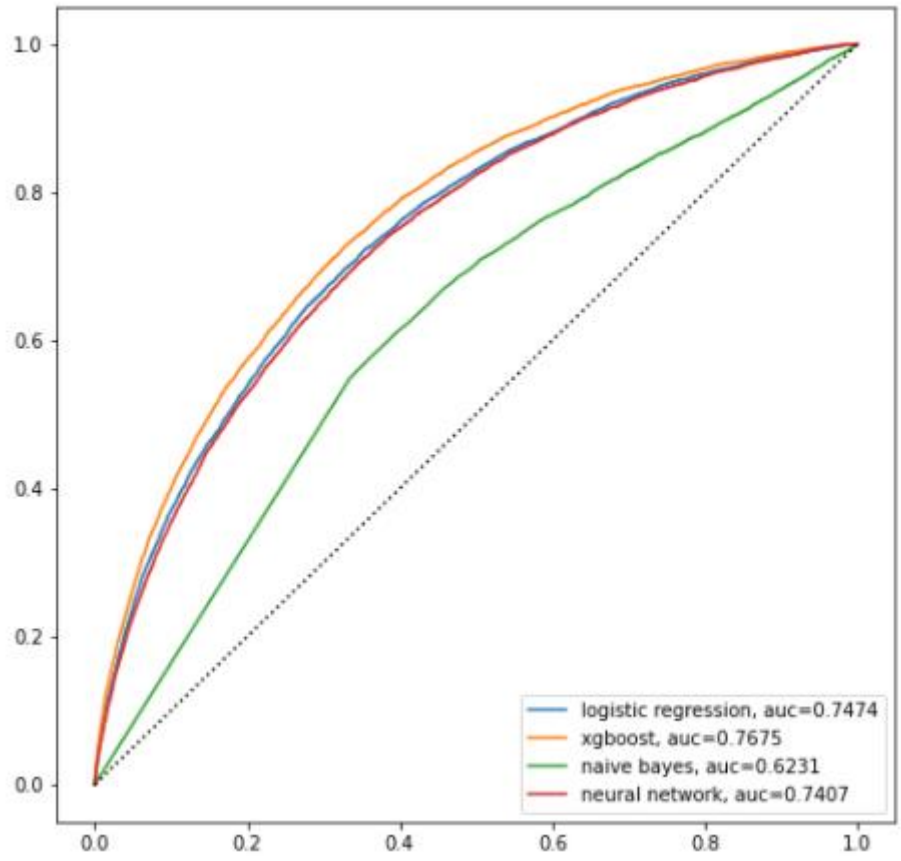
- Ta sẽ tạo thêm một số feature mà các chuyên gia trong ngành khuyên dùng:
  - DIR = AMT_CREDIT / AMT_INCOME_TOTAL: debt-to-income ratio, hệ số nợ trên thu nhập
  - AIR = AMT_ANNUITY / AMT_INCOME_TOTAL: hệ số niên kim (niên kim giống như tiền bảo hiểm) trên thu nhập
  - ACR = AMT_ANNUITY / AMT_CREDIT
  - DAR = DAYS_EMPLOYED / DAYS_BIRTH: người dùng sẽ dùng khoản vay vào làm việc gì?
- Ta có top 5 feature có Pearson Correlation lớn nhất với TARGET, vậy nên ta sẽ cố tạo thêm các Polonomial Features từ chúng

```
DAYS_EMPLOYED                0.074958
DAYS_BIRTH                   0.078239
EXT_SOURCE_1                 0.155317
EXT_SOURCE_2                 0.160472
EXT_SOURCE_3                 0.178919
```

- Kết quả: ROC AUC (chưa tune, chưa thêm các kĩ thuật gì khác, chỉ thuần xử lý dữ liệu)
  - Xanh: feature thường, bỏ top 26 feature missing
  - Vàng: sau khi thêm các feature mới

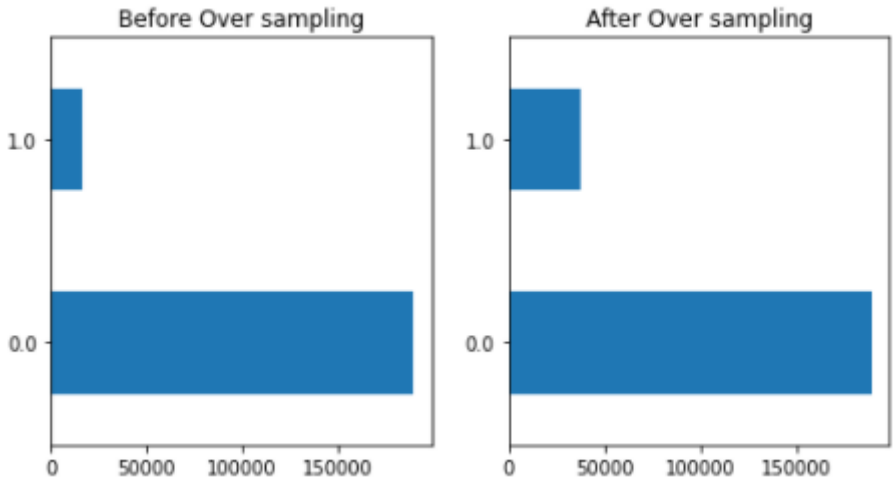
| Mô hình \ Tập đánh giá | Val set tách từ train set (30%) | Test set của Kaggle |
|---|---|---|

| | | | | |
|---|---|---|---|---|
| Logistic Regression | 0.7459 | 0.7474 | 0.7384 | 0.7386 |
| Naïve Bayes Classifier | 0.5317 | 0.6231 | 0.5369 | 0.6264 |
| **XGBoost Classifier (n_estimators=250, max_depth=5)** | **0.7590** | **0.7675** | **0.7476** | **0.7596** |
| Neural Network (1 hidden layer 64 nodes) | 0.7378 | 0.7407 | 0.7385 | 0.7401 |

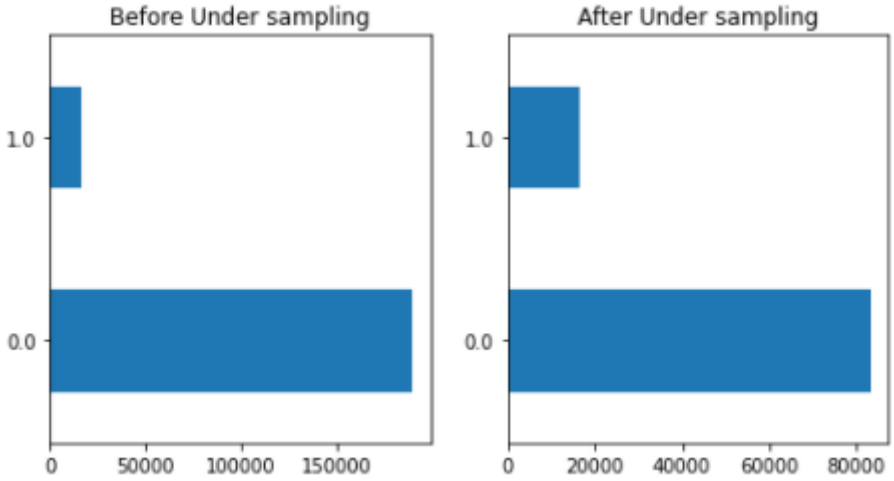



Kết quả tốt nhất

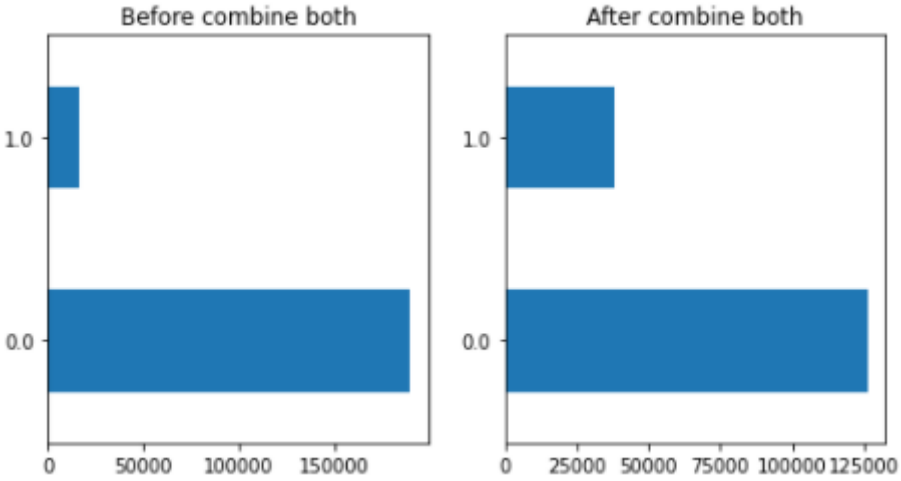
- XGBoost có vẻ cho kết quả tốt nhất trong các model sử dụng, vì vậy nên ta sẽ sử dụng XGBoost để đánh giá hiệu quả của các phương pháp sắp sử dụng phía dưới
- Vì đây là bài toán Imbalanced data, ta sẽ sử dụng các phương pháp riêng để xử lý
  - Over sampling, under sampling, kết hợp cả 2
    - Over sampling (20%)




- Under sampling (20%)



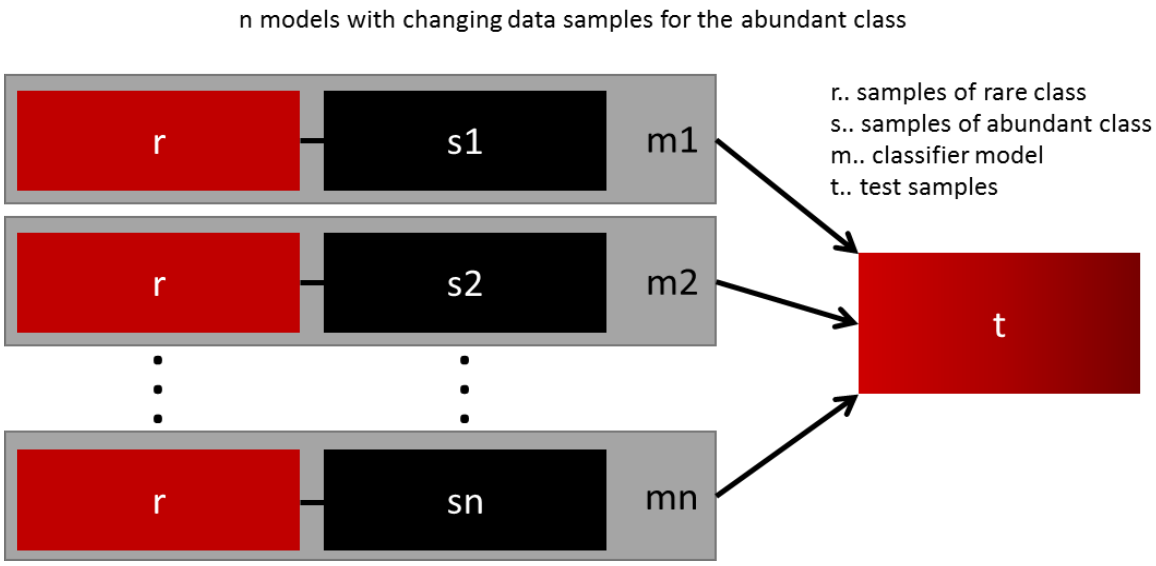

- Kết hợp cả 2




- Ensemble different resampled datasets
  - Chia tập train nhãn 0 thành nhiều phần nhỏ để gộp với tập train nhãn 1 thành các bộ dataset riêng lẻ, mỗi dataset train một model riêng sau đó lấy trung bình output để ra output cuối cùng

- Chia tập train nhãn 0 thành 3 phần bằng nhau, mỗi phần kết hợp với tập train nhãn 1 thành 1 bộ dataset riêng, sau đó tính trung bình output
- Kết quả train từng model

| | Val set tách từ train set (30%) |
|---|---|
| Model 1 | 0.7650 |
| Model 1 | 0.7644 |
| Model 1 | 0.7660 |




- Kết quả:

| Tập đánh giá / XGBoost + Phương pháp | Val set tách từ train set (30%) | Test set của Kaggle |
|---|---|---|
| Base | 0.7675 | 0.7596 |
| Over sampling (20%) | 0.7656 | 0.7606 |

| | | |
|---|---|---|
| Under sampling | 0.7653 | 0.7579 |
| Kết hợp cả 2 | 0.7651 | 0.7586 |
| Ensemble different resampled datasets | 0.7682 | 0.7613 |

2. Merge tất cả các bảng lại
   - application + bureau + bureau_balance
     - Ta tính sum, mean, variance của các lần vay trước của khách ở những tổ chức khác
     - Có những khách không có lịch sử giao dịch ở các tổ chức khác từ trước, vì vậy ta sẽ fill các giá trị NaN bằng 0 khi merge
     - Các feature dùng:
       - Các feature Numerical

| |
|---|
| DAYS_CREDIT ['mean'] |
| DAYS_CREDIT_ENDDATE ['mean'] |
| DAYS_CREDIT_UPDATE ['mean'] |
| CREDIT_DAY_OVERDUE ['mean'] |
| AMT_CREDIT_MAX_OVERDUE ['mean'] |
| AMT_CREDIT_SUM ['mean', 'sum'] |
| AMT_CREDIT_SUM_DEBT ['mean', 'sum'] |
| AMT_CREDIT_SUM_OVERDUE ['mean'] |
| AMT_CREDIT_SUM_LIMIT ['mean', 'sum'] |
| AMT_ANNUITY ['max', 'mean'] |
| CNT_CREDIT_PROLONG ['sum'] |
| MONTHS_BALANCE_MIN ['min'] |
| MONTHS_BALANCE_MAX ['max'] |
| MONTHS_BALANCE_SIZE ['mean', 'sum'] |

       - Các feature Categorical

| |
|---|
| CREDIT_ACTIVE_Active ['mean'] |

| |
|---|
| CREDIT_ACTIVE_Bad debt ['mean'] |
| CREDIT_ACTIVE_Closed ['mean'] |
| CREDIT_ACTIVE_Sold ['mean'] |
| CREDIT_CURRENCY_currency 1 ['mean'] |
| CREDIT_CURRENCY_currency 2 ['mean'] |
| CREDIT_CURRENCY_currency 3 ['mean'] |
| CREDIT_CURRENCY_currency 4 ['mean'] |
| CREDIT_TYPE_Another type of loan ['mean'] |
| CREDIT_TYPE_Car loan ['mean'] |
| CREDIT_TYPE_Cash loan (non-earmarked) ['mean'] |
| CREDIT_TYPE_Consumer credit ['mean'] |
| CREDIT_TYPE_Credit card ['mean'] |
| CREDIT_TYPE_Interbank credit ['mean'] |
| CREDIT_TYPE_Loan for business development ['mean'] |
| CREDIT_TYPE_Loan for purchase of shares (margin lending) ['mean'] |
| CREDIT_TYPE_Loan for the purchase of equipment ['mean'] |
| CREDIT_TYPE_Loan for working capital replenishment ['mean'] |
| CREDIT_TYPE_Microloan ['mean'] |
| CREDIT_TYPE_Mobile operator loan ['mean'] |
| CREDIT_TYPE_Mortgage ['mean'] |
| CREDIT_TYPE_Real estate loan ['mean'] |
| CREDIT_TYPE_Unknown type of loan ['mean'] |
| STATUS_0_MEAN ['mean'] |
| STATUS_1_MEAN ['mean'] |
| STATUS_2_MEAN ['mean'] |
| STATUS_3_MEAN ['mean'] |
| STATUS_4_MEAN ['mean'] |

| |
|---|
| STATUS_5_MEAN ['mean'] |
| STATUS_C_MEAN ['mean'] |
| STATUS_X_MEAN ['mean'] |

- application + previous_application + POS_CASH_balance + installments_payment + credit_card_balance
  - Giống hệt với cách xử lý trên, trường hợp này là dữ liệu các khoản vay trước đó của khách hàng tại Home Credit
  - Những giá trị missing ta vẫn fill bằng 0
  - Trong quá trình chạy kiểm thử thì kết quả cho ta thấy 1 fact là những thông tin về khoản vay trước đó của khách hàng tại Home Credit thì có giá trị hơn (khá logic)
  - Khá nhiều feature nên em không đưa vào báo cáo vì không cần thiết lắm (cách làm và các feature giống bureau)
- Tất cả: 642 features

| Tập đánh giá / XGBoost + các cách merge data | Val set tách từ train set (30%) | Test set của Kaggle |
|---|---|---|
| Base | 0.7675 | 0.7596 |
| application + bureau + bureau_balance | 0.7724 | 0.7727 |
| application + previous_aplication + POS_CASH_balance + installments_payment + credit_card_balance | 0.7830 | 0.7841 |
| Tất cả | 0.7895 | 0.7930 |