

02__Exploration.ipynb

September 28, 2021

1 Objectif :

Un algorithme de classification avec des perfs **>70 % recall** et **> 50 % f1-score**

2 Checklist de base :

2.1 Analyse de forme :

- **Variable cible** : TARGET : Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)
- **Dimensions** : 307511 lignes 121 colonnes + 1 colonne ID Client
- Types de variables :
- Valeurs manquantes : les variables de type __AVG, __MODE, __MEDI sont très peu complètes. OWN_CAR_AGE peu complète car la plupart des clients n'ont pas de voiture (soluce: -1) EXT_SOURCE_1 très peu complète mais très utile malgré tout car c'est un score de qualité de l'emprunteur

```
[ ]: # code pour importer les descriptions de variable
```

```
[ ]: # Importations
import sys
sys.path.append('..')

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from styles import *
```

```
[ ]: # Initialisation
train = pd.read_csv('../02_data/application_train.csv')
#test = pd.read_csv('../02_data/application_test.csv')
df = train.copy()

id_error_msg = lambda x: '`SK_ID_CURR` is not unic for {} set!'.format(x)
assert len(df.SK_ID_CURR.unique()) == df.shape[0], id_error_msg('train')
```

```
#assert len(test.SK_ID_CURR.unique()) == test.shape[0], id_error_msg('test')
df.set_index('SK_ID_CURR', inplace=True)
#test.set_index('SK_ID_CURR', inplace=True)
```

```
[ ]: df.head()
```

```
[ ]: TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR \
```

SK_ID_CURR

100002	1	Cash loans	M	N
100003	0	Cash loans	F	N
100004	0	Revolving loans	M	Y
100006	0	Cash loans	F	N
100007	0	Cash loans	M	N

```
FLAG_OWN_REALTY CNT_CHILDREN AMT_INCOME_TOTAL AMT_CREDIT \
```

SK_ID_CURR

100002	Y	0	202500.0	406597.5
100003	N	0	270000.0	1293502.5
100004	Y	0	67500.0	135000.0
100006	Y	0	135000.0	312682.5
100007	Y	0	121500.0	513000.0

```
AMT_ANNUITY AMT_GOODS_PRICE ... FLAG_DOCUMENT_18 \
```

SK_ID_CURR

100002	24700.5	351000.0	...	0
100003	35698.5	1129500.0	...	0
100004	6750.0	135000.0	...	0
100006	29686.5	297000.0	...	0
100007	21865.5	513000.0	...	0

```
FLAG_DOCUMENT_19 FLAG_DOCUMENT_20 FLAG_DOCUMENT_21 \
```

SK_ID_CURR

100002	0	0	0
100003	0	0	0
100004	0	0	0
100006	0	0	0
100007	0	0	0

```
AMT_REQ_CREDIT_BUREAU_HOUR AMT_REQ_CREDIT_BUREAU_DAY \
```

SK_ID_CURR

100002	0.0	0.0
100003	0.0	0.0
100004	0.0	0.0
100006	NaN	NaN
100007	0.0	0.0

```
AMT_REQ_CREDIT_BUREAU_WEEK AMT_REQ_CREDIT_BUREAU_MON \
```

SK_ID_CURR		
100002	0.0	0.0
100003	0.0	0.0
100004	0.0	0.0
100006	NaN	NaN
100007	0.0	0.0

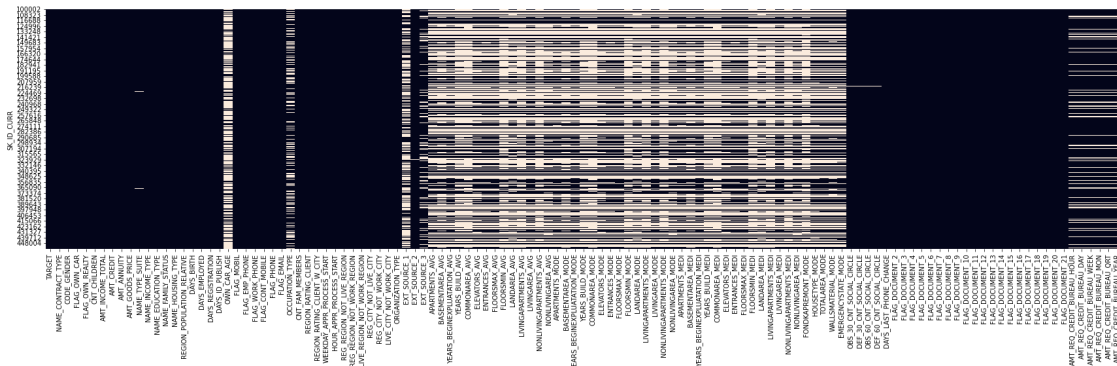
	AMT_REQ_CREDIT_BUREAU_QRT	AMT_REQ_CREDIT_BUREAU_YEAR
SK_ID_CURR		
100002	0.0	1.0
100003	0.0	0.0
100004	0.0	0.0
100006	NaN	NaN
100007	0.0	0.0

[5 rows x 121 columns]

```
[ ]: print('Training set dimensions:', df.shape)
```

Training set dimensions: (307511, 121)

```
[ ]: plt.figure(figsize=(24,8), facecolor='white')
sns.heatmap(df.isna(), cbar=False)
plt.tight_layout()
plt.savefig('isna.png')
```



```
[ ]: with pd.option_context('display.max_rows', None):
print((df.isna().sum() / df.shape[0]).sort_values(ascending=False))
```

COMMONAREA_AVG	0.698723
COMMONAREA_MODE	0.698723
COMMONAREA_MEDI	0.698723
NONLIVINGAPARTMENTS_AVG	0.694330
NONLIVINGAPARTMENTS_MODE	0.694330

NONLIVINGAPARTMENTS_MEDI	0.694330
FONDKAPREMONT_MODE	0.683862
LIVINGAPARTMENTS_MEDI	0.683550
LIVINGAPARTMENTS_AVG	0.683550
LIVINGAPARTMENTS_MODE	0.683550
FLOORSMIN_AVG	0.678486
FLOORSMIN_MODE	0.678486
FLOORSMIN_MEDI	0.678486
YEARS_BUILD_AVG	0.664978
YEARS_BUILD_MEDI	0.664978
YEARS_BUILD_MODE	0.664978
OWN_CAR_AGE	0.659908
LANDAREA_MEDI	0.593767
LANDAREA_AVG	0.593767
LANDAREA_MODE	0.593767
BASEMENTAREA_MEDI	0.585160
BASEMENTAREA_AVG	0.585160
BASEMENTAREA_MODE	0.585160
EXT_SOURCE_1	0.563811
NONLIVINGAREA_AVG	0.551792
NONLIVINGAREA_MODE	0.551792
NONLIVINGAREA_MEDI	0.551792
ELEVATORS_MODE	0.532960
ELEVATORS_AVG	0.532960
ELEVATORS_MEDI	0.532960
WALLSMATERIAL_MODE	0.508408
APARTMENTS_AVG	0.507497
APARTMENTS_MEDI	0.507497
APARTMENTS_MODE	0.507497
ENTRANCES_AVG	0.503488
ENTRANCES_MEDI	0.503488
ENTRANCES_MODE	0.503488
LIVINGAREA_MEDI	0.501933
LIVINGAREA_MODE	0.501933
LIVINGAREA_AVG	0.501933
HOUSETYPE_MODE	0.501761
FLOORSMAX_MEDI	0.497608
FLOORSMAX_MODE	0.497608
FLOORSMAX_AVG	0.497608
YEARS_BEGINEXPLUATATION_MEDI	0.487810
YEARS_BEGINEXPLUATATION_MODE	0.487810
YEARS_BEGINEXPLUATATION_AVG	0.487810
TOTALAREA_MODE	0.482685
EMERGENCYSTATE_MODE	0.473983
OCCUPATION_TYPE	0.313455
EXT_SOURCE_3	0.198253
AMT_REQ_CREDIT_BUREAU_WEEK	0.135016
AMT_REQ_CREDIT_BUREAU_HOUR	0.135016

AMT_REQ_CREDIT_BUREAU_MON	0.135016
AMT_REQ_CREDIT_BUREAU_QRT	0.135016
AMT_REQ_CREDIT_BUREAU_DAY	0.135016
AMT_REQ_CREDIT_BUREAU_YEAR	0.135016
NAME_TYPE_SUITE	0.004201
DEF_30_CNT_SOCIAL_CIRCLE	0.003320
OBS_60_CNT_SOCIAL_CIRCLE	0.003320
OBS_30_CNT_SOCIAL_CIRCLE	0.003320
DEF_60_CNT_SOCIAL_CIRCLE	0.003320
EXT_SOURCE_2	0.002146
AMT_GOODS_PRICE	0.000904
AMT_ANNUITY	0.000039
CNT_FAM_MEMBERS	0.000007
DAYS_LAST_PHONE_CHANGE	0.000003
AMT_INCOME_TOTAL	0.000000
FLAG_DOCUMENT_8	0.000000
CODE_GENDER	0.000000
FLAG_OWN_CAR	0.000000
FLAG_OWN_REALTY	0.000000
FLAG_DOCUMENT_2	0.000000
FLAG_DOCUMENT_3	0.000000
FLAG_DOCUMENT_4	0.000000
FLAG_DOCUMENT_5	0.000000
FLAG_DOCUMENT_6	0.000000
FLAG_DOCUMENT_7	0.000000
FLAG_DOCUMENT_9	0.000000
FLAG_DOCUMENT_21	0.000000
FLAG_DOCUMENT_10	0.000000
FLAG_DOCUMENT_11	0.000000
CNT_CHILDREN	0.000000
FLAG_DOCUMENT_13	0.000000
FLAG_DOCUMENT_14	0.000000
FLAG_DOCUMENT_15	0.000000
FLAG_DOCUMENT_16	0.000000
FLAG_DOCUMENT_17	0.000000
FLAG_DOCUMENT_18	0.000000
FLAG_DOCUMENT_19	0.000000
FLAG_DOCUMENT_20	0.000000
FLAG_DOCUMENT_12	0.000000
AMT_CREDIT	0.000000
ORGANIZATION_TYPE	0.000000
NAME_INCOME_TYPE	0.000000
LIVE_CITY_NOT_WORK_CITY	0.000000
NAME_CONTRACT_TYPE	0.000000
REG_CITY_NOT_WORK_CITY	0.000000
REG_CITY_NOT_LIVE_CITY	0.000000
LIVE_REGION_NOT_WORK_REGION	0.000000
REG_REGION_NOT_WORK_REGION	0.000000

```

REG_REGION_NOT_LIVE_REGION    0.000000
HOUR_APPR_PROCESS_START       0.000000
WEEKDAY_APPR_PROCESS_START    0.000000
REGION_RATING_CLIENT_W_CITY   0.000000
REGION_RATING_CLIENT          0.000000
FLAG_EMAIL                     0.000000
FLAG_PHONE                     0.000000
FLAG_CONT_MOBILE               0.000000
FLAG_WORK_PHONE                0.000000
FLAG_EMP_PHONE                 0.000000
FLAG_MOBIL                     0.000000
DAYS_ID_PUBLISH                0.000000
DAYS_REGISTRATION              0.000000
DAYS_EMPLOYED                  0.000000
DAYS_BIRTH                     0.000000
REGION_POPULATION_RELATIVE     0.000000
NAME_HOUSING_TYPE              0.000000
NAME_FAMILY_STATUS             0.000000
NAME_EDUCATION_TYPE            0.000000
TARGET                         0.000000
dtype: float64

```

3 Possession d'une voiture et âge de la voiture

```
[ ]: df.FLAG_OWN_CAR.value_counts()
```

```
[ ]: N    202924
     Y    104587
     Name: FLAG_OWN_CAR, dtype: int64
```

```
[ ]: pd.crosstab(df.TARGET, df.FLAG_OWN_CAR, normalize='index')
```

```
[ ]: FLAG_OWN_CAR      N      Y
     TARGET
     0      0.656824  0.343176
     1      0.694824  0.305176
```

```
[ ]: df.OWN_CAR_AGE.describe()
```

```
[ ]: count    104582.000000
     mean      12.061091
     std       11.944812
     min        0.000000
     25%        5.000000
     50%        9.000000
     75%       15.000000
     max       91.000000
```

Name: OWN_CAR_AGE, dtype: float64

```
[ ]: df[df.OWN_CAR_AGE == 0].shape[0]
```

```
[ ]: 2134
```

```
[ ]: try:
      assert df[df.FLAG_OWN_CAR == 'N'].FLAG_OWN_CAR.count()\
          == df.OWN_CAR_AGE.isna().sum()
    except AssertionError:
        print('No. Non-car-owners:', df[df.FLAG_OWN_CAR == 'N'].shape[0])
        print('No. Customers with unknown car age:', df.OWN_CAR_AGE.isna().sum())
```

No. Non-car-owners: 202924

No. Customers with unknown car age: 202929

```
[ ]: df.loc[df.FLAG_OWN_CAR == 'N', 'OWN_CAR_AGE'] = -1.0
```

```
[ ]: df.OWN_CAR_AGE.describe()
```

```
[ ]: count    307506.000000
      mean      3.442043
      std      9.317221
      min     -1.000000
      25%     -1.000000
      50%     -1.000000
      75%      5.000000
      max     91.000000
      Name: OWN_CAR_AGE, dtype: float64
```

```
[ ]: with pd.option_context('display.max_rows', None):
      print((df.isna().sum() / df.shape[0]).sort_values(ascending=False))
```

COMMONAREA_MEDI	0.698723
COMMONAREA_AVG	0.698723
COMMONAREA_MODE	0.698723
NONLIVINGAPARTMENTS_MODE	0.694330
NONLIVINGAPARTMENTS_AVG	0.694330
NONLIVINGAPARTMENTS_MEDI	0.694330
FONDKAPREMONT_MODE	0.683862
LIVINGAPARTMENTS_MODE	0.683550
LIVINGAPARTMENTS_AVG	0.683550
LIVINGAPARTMENTS_MEDI	0.683550
FLOORSMIN_AVG	0.678486
FLOORSMIN_MODE	0.678486
FLOORSMIN_MEDI	0.678486
YEARS_BUILD_MEDI	0.664978
YEARS_BUILD_MODE	0.664978

YEARS_BUILD_AVG	0.664978
LANDAREA_MEDI	0.593767
LANDAREA_MODE	0.593767
LANDAREA_AVG	0.593767
BASEMENTAREA_MEDI	0.585160
BASEMENTAREA_AVG	0.585160
BASEMENTAREA_MODE	0.585160
EXT_SOURCE_1	0.563811
NONLIVINGAREA_MODE	0.551792
NONLIVINGAREA_AVG	0.551792
NONLIVINGAREA_MEDI	0.551792
ELEVATORS_MEDI	0.532960
ELEVATORS_AVG	0.532960
ELEVATORS_MODE	0.532960
WALLSMATERIAL_MODE	0.508408
APARTMENTS_MEDI	0.507497
APARTMENTS_AVG	0.507497
APARTMENTS_MODE	0.507497
ENTRANCES_MEDI	0.503488
ENTRANCES_AVG	0.503488
ENTRANCES_MODE	0.503488
LIVINGAREA_AVG	0.501933
LIVINGAREA_MODE	0.501933
LIVINGAREA_MEDI	0.501933
HOUSETYPE_MODE	0.501761
FLOORSMAX_MODE	0.497608
FLOORSMAX_MEDI	0.497608
FLOORSMAX_AVG	0.497608
YEARS_BEGINEXPLUATATION_MODE	0.487810
YEARS_BEGINEXPLUATATION_MEDI	0.487810
YEARS_BEGINEXPLUATATION_AVG	0.487810
TOTALAREA_MODE	0.482685
EMERGENCYSTATE_MODE	0.473983
OCCUPATION_TYPE	0.313455
EXT_SOURCE_3	0.198253
AMT_REQ_CREDIT_BUREAU_HOUR	0.135016
AMT_REQ_CREDIT_BUREAU_DAY	0.135016
AMT_REQ_CREDIT_BUREAU_WEEK	0.135016
AMT_REQ_CREDIT_BUREAU_MON	0.135016
AMT_REQ_CREDIT_BUREAU_QRT	0.135016
AMT_REQ_CREDIT_BUREAU_YEAR	0.135016
NAME_TYPE_SUITE	0.004201
DEF_60_CNT_SOCIAL_CIRCLE	0.003320
OBS_60_CNT_SOCIAL_CIRCLE	0.003320
DEF_30_CNT_SOCIAL_CIRCLE	0.003320
OBS_30_CNT_SOCIAL_CIRCLE	0.003320
EXT_SOURCE_2	0.002146
AMT_GOODS_PRICE	0.000904

AMT_ANNUITY	0.000039
OWN_CAR_AGE	0.000016
CNT_FAM_MEMBERS	0.000007
DAYS_LAST_PHONE_CHANGE	0.000003
REG_REGION_NOT_LIVE_REGION	0.000000
FLAG_DOCUMENT_14	0.000000
FLAG_DOCUMENT_3	0.000000
FLAG_DOCUMENT_4	0.000000
FLAG_DOCUMENT_5	0.000000
FLAG_DOCUMENT_6	0.000000
FLAG_DOCUMENT_7	0.000000
FLAG_DOCUMENT_8	0.000000
FLAG_DOCUMENT_9	0.000000
FLAG_DOCUMENT_10	0.000000
FLAG_DOCUMENT_11	0.000000
FLAG_DOCUMENT_12	0.000000
FLAG_DOCUMENT_13	0.000000
FLAG_DOCUMENT_16	0.000000
FLAG_DOCUMENT_15	0.000000
FLAG_DOCUMENT_17	0.000000
FLAG_DOCUMENT_18	0.000000
FLAG_DOCUMENT_19	0.000000
FLAG_DOCUMENT_20	0.000000
FLAG_DOCUMENT_21	0.000000
AMT_INCOME_TOTAL	0.000000
CNT_CHILDREN	0.000000
FLAG_OWN_REALTY	0.000000
FLAG_OWN_CAR	0.000000
CODE_GENDER	0.000000
FLAG_DOCUMENT_2	0.000000
REG_CITY_NOT_WORK_CITY	0.000000
LIVE_CITY_NOT_WORK_CITY	0.000000
DAYS_EMPLOYED	0.000000
WEEKDAY_APPR_PROCESS_START	0.000000
REGION_RATING_CLIENT_W_CITY	0.000000
REGION_RATING_CLIENT	0.000000
FLAG_EMAIL	0.000000
FLAG_PHONE	0.000000
FLAG_CONT_MOBILE	0.000000
FLAG_WORK_PHONE	0.000000
FLAG_EMP_PHONE	0.000000
FLAG_MOBIL	0.000000
DAYS_ID_PUBLISH	0.000000
DAYS_REGISTRATION	0.000000
DAYS_BIRTH	0.000000
HOURL_APPR_PROCESS_START	0.000000
REGION_POPULATION_RELATIVE	0.000000
NAME_HOUSING_TYPE	0.000000

NAME_FAMILY_STATUS	0.000000
NAME_EDUCATION_TYPE	0.000000
NAME_INCOME_TYPE	0.000000
REG_REGION_NOT_WORK_REGION	0.000000
LIVE_REGION_NOT_WORK_REGION	0.000000
ORGANIZATION_TYPE	0.000000
AMT_CREDIT	0.000000
REG_CITY_NOT_LIVE_CITY	0.000000
NAME_CONTRACT_TYPE	0.000000
TARGET	0.000000
dtype:	float64

3.1 Corrélations avec la variable cible

```
[ ]: correlations = df.corr()['TARGET']
```

```
[ ]: correlations = correlations.drop(['TARGET'], axis=0)
with pd.option_context('display.max_rows', None):
    print(correlations.sort_values(ascending=False))
```

DAYS_BIRTH	0.078239
REGION_RATING_CLIENT_W_CITY	0.060893
REGION_RATING_CLIENT	0.058899
DAYS_LAST_PHONE_CHANGE	0.055218
DAYS_ID_PUBLISH	0.051457
REG_CITY_NOT_WORK_CITY	0.050994
FLAG_EMP_PHONE	0.045982
REG_CITY_NOT_LIVE_CITY	0.044395
FLAG_DOCUMENT_3	0.044346
DAYS_REGISTRATION	0.041975
LIVE_CITY_NOT_WORK_CITY	0.032518
DEF_30_CNT_SOCIAL_CIRCLE	0.032248
DEF_60_CNT_SOCIAL_CIRCLE	0.031276
FLAG_WORK_PHONE	0.028524
AMT_REQ_CREDIT_BUREAU_YEAR	0.019930
CNT_CHILDREN	0.019187
CNT_FAM_MEMBERS	0.009308
OBS_30_CNT_SOCIAL_CIRCLE	0.009131
OBS_60_CNT_SOCIAL_CIRCLE	0.009022
REG_REGION_NOT_WORK_REGION	0.006942
REG_REGION_NOT_LIVE_REGION	0.005576
FLAG_DOCUMENT_2	0.005417
FLAG_DOCUMENT_21	0.003709
LIVE_REGION_NOT_WORK_REGION	0.002819
AMT_REQ_CREDIT_BUREAU_DAY	0.002704
OWN_CAR_AGE	0.001097
AMT_REQ_CREDIT_BUREAU_HOUR	0.000930
AMT_REQ_CREDIT_BUREAU_WEEK	0.000788

FLAG_MOBIL	0.000534
FLAG_CONT_MOBILE	0.000370
FLAG_DOCUMENT_20	0.000215
FLAG_DOCUMENT_5	-0.000316
FLAG_DOCUMENT_12	-0.000756
FLAG_DOCUMENT_19	-0.001358
FLAG_DOCUMENT_10	-0.001414
FLAG_DOCUMENT_7	-0.001520
NONLIVINGAPARTMENTS_MODE	-0.001557
FLAG_EMAIL	-0.001758
AMT_REQ_CREDIT_BUREAU_QRT	-0.002022
FLAG_DOCUMENT_4	-0.002672
NONLIVINGAPARTMENTS_MEDI	-0.002757
NONLIVINGAPARTMENTS_AVG	-0.003176
FLAG_DOCUMENT_17	-0.003378
AMT_INCOME_TOTAL	-0.003982
FLAG_DOCUMENT_11	-0.004229
FLAG_DOCUMENT_9	-0.004352
FLAG_DOCUMENT_15	-0.006536
FLAG_DOCUMENT_18	-0.007952
FLAG_DOCUMENT_8	-0.008040
YEARS_BEGINEXPLUATATION_MODE	-0.009036
FLAG_DOCUMENT_14	-0.009464
YEARS_BEGINEXPLUATATION_AVG	-0.009728
YEARS_BEGINEXPLUATATION_MEDI	-0.009993
LANDAREA_MODE	-0.010174
LANDAREA_AVG	-0.010885
LANDAREA_MEDI	-0.011256
FLAG_DOCUMENT_13	-0.011583
FLAG_DOCUMENT_16	-0.011615
AMT_REQ_CREDIT_BUREAU_MON	-0.012462
NONLIVINGAREA_MODE	-0.012711
AMT_ANNUITY	-0.012817
NONLIVINGAREA_MEDI	-0.013337
NONLIVINGAREA_AVG	-0.013578
COMMONAREA_MODE	-0.016340
ENTRANCES_MODE	-0.017387
COMMONAREA_AVG	-0.018550
COMMONAREA_MEDI	-0.018573
ENTRANCES_MEDI	-0.019025
ENTRANCES_AVG	-0.019172
BASEMENTAREA_MODE	-0.019952
YEARS_BUILD_MODE	-0.022068
BASEMENTAREA_MEDI	-0.022081
YEARS_BUILD_AVG	-0.022149
YEARS_BUILD_MEDI	-0.022326
BASEMENTAREA_AVG	-0.022746
LIVINGAPARTMENTS_MODE	-0.023393

```

FLAG_PHONE -0.023806
HOUR_APPR_PROCESS_START -0.024166
LIVINGAPARTMENTS_MEDI -0.024621
LIVINGAPARTMENTS_AVG -0.025031
APARTMENTS_MODE -0.027284
FLAG_DOCUMENT_6 -0.028602
APARTMENTS_MEDI -0.029184
APARTMENTS_AVG -0.029498
AMT_CREDIT -0.030369
LIVINGAREA_MODE -0.030685
ELEVATORS_MODE -0.032131
TOTALAREA_MODE -0.032596
FLOORSMIN_MODE -0.032698
LIVINGAREA_MEDI -0.032739
LIVINGAREA_AVG -0.032997
FLOORSMIN_MEDI -0.033394
FLOORSMIN_AVG -0.033614
ELEVATORS_MEDI -0.033863
ELEVATORS_AVG -0.034199
REGION_POPULATION_RELATIVE -0.037227
AMT_GOODS_PRICE -0.039645
FLOORSMAX_MODE -0.043226
FLOORSMAX_MEDI -0.043768
FLOORSMAX_AVG -0.044003
DAYS_EMPLOYED -0.044932
EXT_SOURCE_1 -0.155317
EXT_SOURCE_2 -0.160472
EXT_SOURCE_3 -0.178919
Name: TARGET, dtype: float64

```

4 EDA avec prétraitement des données

```

[ ]: from preprocessing import preprocessor as prep
      from preprocessing import preprocessor_no_scaler as prep_no_scl
      from preprocessing import get_preprocessed_set_column_names as get_col_names

      df_trans = pd.DataFrame(prepare_no_scl.fit_transform(df.iloc[:, 1:]),
                              columns=get_col_names(prepare_no_scl))
      df_trans_scl = pd.DataFrame(prepare.fit_transform(df.iloc[:, 1:]),
                                  columns=get_col_names(prepare))

```

```
[ ]: df_trans
```

```

[ ]:
      CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  AMT_ANNUITY  \
0                0.0        202500.0    406597.5    24700.5
1                0.0        270000.0    1293502.5    35698.5
2                0.0         67500.0    135000.0     6750.0

```

3	0.0	135000.0	312682.5	29686.5
4	0.0	121500.0	513000.0	21865.5
...
307506	0.0	157500.0	254700.0	27558.0
307507	0.0	72000.0	269550.0	12001.5
307508	0.0	153000.0	677664.0	29979.0
307509	0.0	171000.0	370107.0	20205.0
307510	0.0	157500.0	675000.0	49117.5

	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	\
0	351000.0	0.018801	-9461.0	
1	1129500.0	0.003541	-16765.0	
2	135000.0	0.010032	-19046.0	
3	297000.0	0.008019	-19005.0	
4	513000.0	0.028663	-19932.0	
...	
307506	225000.0	0.032561	-9327.0	
307507	225000.0	0.025164	-20775.0	
307508	585000.0	0.005002	-14966.0	
307509	319500.0	0.005313	-11961.0	
307510	675000.0	0.046220	-16856.0	

	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	...	\
0	-637.0	-3648.0	-2120.0	...	
1	-1188.0	-1186.0	-291.0	...	
2	-225.0	-4260.0	-2531.0	...	
3	-3039.0	-9833.0	-2437.0	...	
4	-3038.0	-4311.0	-3458.0	...	
...	
307506	-236.0	-8456.0	-1982.0	...	
307507	365243.0	-4388.0	-4090.0	...	
307508	-7921.0	-6737.0	-5150.0	...	
307509	-4786.0	-2562.0	-931.0	...	
307510	-1262.0	-5128.0	-410.0	...	

	HOUSETYPE_MODE_terraced_house	HOUSETYPE_MODE_unknown	\
0	0.0	0.0	
1	0.0	0.0	
2	0.0	1.0	
3	0.0	1.0	
4	0.0	1.0	
...	
307506	0.0	0.0	
307507	0.0	0.0	
307508	0.0	0.0	
307509	0.0	0.0	
307510	0.0	0.0	

	WALLSMATERIAL_MODE_block	WALLSMATERIAL_MODE_mixed \
0	0.0	0.0
1	1.0	0.0
2	0.0	0.0
3	0.0	0.0
4	0.0	0.0
...
307506	0.0	0.0
307507	0.0	0.0
307508	0.0	0.0
307509	0.0	0.0
307510	0.0	0.0

	WALLSMATERIAL_MODE_monolithic	WALLSMATERIAL_MODE_others \
0	0.0	0.0
1	0.0	0.0
2	0.0	0.0
3	0.0	0.0
4	0.0	0.0
...
307506	0.0	0.0
307507	0.0	0.0
307508	0.0	0.0
307509	0.0	0.0
307510	0.0	0.0

	WALLSMATERIAL_MODE_panel	WALLSMATERIAL_MODE_stone_or_brick \
0	0.0	1.0
1	0.0	0.0
2	0.0	0.0
3	0.0	0.0
4	0.0	0.0
...
307506	0.0	1.0
307507	0.0	1.0
307508	1.0	0.0
307509	0.0	1.0
307510	1.0	0.0

	WALLSMATERIAL_MODE_unknown	WALLSMATERIAL_MODE_wooden
0	0.0	0.0
1	0.0	0.0
2	1.0	0.0
3	1.0	0.0
4	1.0	0.0
...

307506	0.0	0.0
307507	0.0	0.0
307508	0.0	0.0
307509	0.0	0.0
307510	0.0	0.0

[307511 rows x 237 columns]

```
[ ]: df_trans['TARGET'] = df['TARGET'].values
df_trans.head()
```

```
[ ]: CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  AMT_ANNUITY  AMT_GOODS_PRICE  \
0          0.0          202500.0    406597.5      24700.5      351000.0
1          0.0          270000.0   1293502.5      35698.5     1129500.0
2          0.0           67500.0    135000.0       6750.0     135000.0
3          0.0          135000.0    312682.5      29686.5     297000.0
4          0.0          121500.0    513000.0      21865.5     513000.0
```

```
REGION_POPULATION_RELATIVE  DAYS_BIRTH  DAYS_EMPLOYED  DAYS_REGISTRATION  \
0          0.018801      -9461.0        -637.0        -3648.0
1          0.003541     -16765.0       -1188.0       -1186.0
2          0.010032     -19046.0       -225.0       -4260.0
3          0.008019     -19005.0      -3039.0      -9833.0
4          0.028663     -19932.0      -3038.0     -4311.0
```

```
DAYS_ID_PUBLISH  ...  HOUSETYPE_MODE_unknown  WALLSMATERIAL_MODE_block  \
0      -2120.0  ...              0.0              0.0
1      -291.0  ...              0.0              1.0
2     -2531.0  ...              1.0              0.0
3     -2437.0  ...              1.0              0.0
4     -3458.0  ...              1.0              0.0
```

```
WALLSMATERIAL_MODE_mixed  WALLSMATERIAL_MODE_monolithic  \
0              0.0              0.0
1              0.0              0.0
2              0.0              0.0
3              0.0              0.0
4              0.0              0.0
```

```
WALLSMATERIAL_MODE_others  WALLSMATERIAL_MODE_panel  \
0              0.0              0.0
1              0.0              0.0
2              0.0              0.0
3              0.0              0.0
4              0.0              0.0
```

```
WALLSMATERIAL_MODE_stone_or_brick  WALLSMATERIAL_MODE_unknown  \
```

0	1.0	0.0
1	0.0	0.0
2	0.0	1.0
3	0.0	1.0
4	0.0	1.0

	WALLSMATERIAL_MODE_wooden	TARGET
0	0.0	1
1	0.0	0
2	0.0	0
3	0.0	0
4	0.0	0

[5 rows x 238 columns]

```
[ ]: df_trans.iloc[:, :-1].head()
```

```
[ ]:  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  AMT_ANNUITY  AMT_GOODS_PRICE  \
0          0.0          202500.0      406597.5      24700.5          351000.0
1          0.0          270000.0     1293502.5      35698.5         1129500.0
2          0.0           67500.0     135000.0       6750.0          135000.0
3          0.0          135000.0     312682.5      29686.5          297000.0
4          0.0          121500.0     513000.0      21865.5          513000.0
```

```
  REGION_POPULATION_RELATIVE  DAYS_BIRTH  DAYS_EMPLOYED  DAYS_REGISTRATION  \
0          0.018801      -9461.0      -637.0          -3648.0
1          0.003541     -16765.0     -1188.0          -1186.0
2          0.010032     -19046.0      -225.0          -4260.0
3          0.008019     -19005.0     -3039.0          -9833.0
4          0.028663     -19932.0     -3038.0          -4311.0
```

```
  DAYS_ID_PUBLISH  ...  HOUSETYPE_MODE_terraced_house  \
0      -2120.0  ...                                0.0
1      -291.0  ...                                0.0
2     -2531.0  ...                                0.0
3     -2437.0  ...                                0.0
4     -3458.0  ...                                0.0
```

```
  HOUSETYPE_MODE_unknown  WALLSMATERIAL_MODE_block  WALLSMATERIAL_MODE_mixed  \
0          0.0                                0.0          0.0
1          0.0                                1.0          0.0
2          1.0                                0.0          0.0
3          1.0                                0.0          0.0
4          1.0                                0.0          0.0
```

```
  WALLSMATERIAL_MODE_monolithic  WALLSMATERIAL_MODE_others  \
0          0.0                                0.0
```


1	0.0	0.0
2	0.0	0.0
3	0.0	0.0
4	0.0	0.0

	WALLSMATERIAL_MODE_panel	WALLSMATERIAL_MODE_stone_or_brick \
0	0.0	1.0
1	0.0	0.0
2	0.0	0.0
3	0.0	0.0
4	0.0	0.0

	WALLSMATERIAL_MODE_unknown	WALLSMATERIAL_MODE_wooden
0	0.0	0.0
1	0.0	0.0
2	1.0	0.0
3	1.0	0.0
4	1.0	0.0

[5 rows x 237 columns]

```
[ ]: print(df.AMT_INCOME_TOTAL.describe())
      print(df.AMT_INCOME_TOTAL.isnull().sum())
```

```
count    3.075110e+05
mean     1.687979e+05
std      2.371231e+05
min      2.565000e+04
25%      1.125000e+05
50%      1.471500e+05
75%      2.025000e+05
max      1.170000e+08
Name: AMT_INCOME_TOTAL, dtype: float64
0
```

```
[ ]: print(df.AMT_CREDIT.describe())
      print(df.AMT_CREDIT.isnull().sum())
```

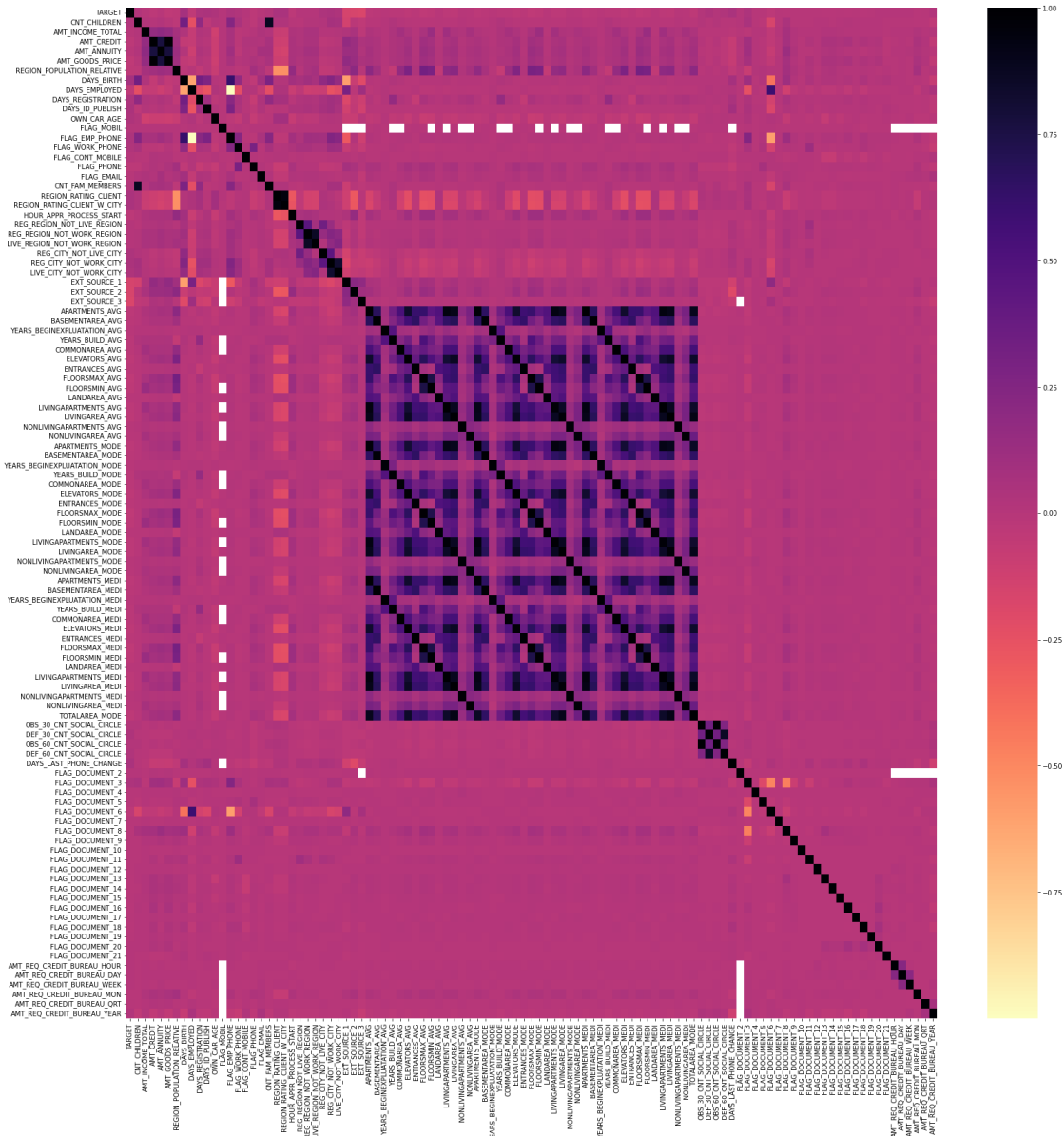
```
count    3.075110e+05
mean     5.990260e+05
std      4.024908e+05
min      4.500000e+04
25%      2.700000e+05
50%      5.135310e+05
75%      8.086500e+05
max      4.050000e+06
Name: AMT_CREDIT, dtype: float64
0
```

```
[ ]: print(df.AMT_ANNUITY.describe())  
      print(df.AMT_ANNUITY.isnull().sum())
```

```
count      307499.000000  
mean       27108.573909  
std        14493.737315  
min        1615.500000  
25%        16524.000000  
50%        24903.000000  
75%        34596.000000  
max        258025.500000  
Name: AMT_ANNUITY, dtype: float64  
12
```

5 Corrélations entre variables

```
[ ]: plt.figure(figsize=(24,24), facecolor='white')  
      sns.heatmap(df.corr(), cmap=plt.cm.magma_r)  
      plt.tight_layout()  
      plt.savefig('corr_heatmap.png')
```



```
[ ]: # Find correlations with the target and sort
correlations = df_trans.corr()['TARGET'].sort_values()

[ ]: correlations = correlations.drop(['TARGET'], axis=0)

[ ]: abs_corr = correlations.apply(lambda x: np.abs(x)).sort_values(ascending=False)

[ ]: with pd.option_context('display.max_rows', None):
    print(abs_corr)
```

EXT_SOURCE_3

0.178919

EXT_SOURCE_2	0.160472
EXT_SOURCE_1	0.155317
DAYS_BIRTH	0.078239
REGION_RATING_CLIENT_W_CITY	0.060893
REGION_RATING_CLIENT	0.058899
DAYS_LAST_PHONE_CHANGE	0.055218
DAYS_ID_PUBLISH	0.051457
REG_CITY_NOT_WORK_CITY	0.050994
FLAG_EMP_PHONE	0.045982
DAYS_EMPLOYED	0.044932
REG_CITY_NOT_LIVE_CITY	0.044395
FLAG_DOCUMENT_3	0.044346
FLOORSMAX_AVG	0.044003
FLOORSMAX_MEDI	0.043768
FLOORSMAX_MODE	0.043226
DAYS_REGISTRATION	0.041975
AMT_GOODS_PRICE	0.039645
REGION_POPULATION_RELATIVE	0.037227
ELEVATORS_AVG	0.034199
ELEVATORS_MEDI	0.033863
FLOORSMIN_AVG	0.033614
FLOORSMIN_MEDI	0.033394
LIVINGAREA_AVG	0.032997
LIVINGAREA_MEDI	0.032739
FLOORSMIN_MODE	0.032698
TOTALAREA_MODE	0.032596
LIVE_CITY_NOT_WORK_CITY	0.032518
DEF_30_CNT_SOCIAL_CIRCLE	0.032248
ELEVATORS_MODE	0.032131
DEF_60_CNT_SOCIAL_CIRCLE	0.031276
LIVINGAREA_MODE	0.030685
AMT_CREDIT	0.030369
APARTMENTS_AVG	0.029498
APARTMENTS_MEDI	0.029184
FLAG_DOCUMENT_6	0.028602
FLAG_WORK_PHONE	0.028524
APARTMENTS_MODE	0.027284
LIVINGAPARTMENTS_AVG	0.025031
LIVINGAPARTMENTS_MEDI	0.024621
HOUR_APPR_PROCESS_START	0.024166
FLAG_PHONE	0.023806
LIVINGAPARTMENTS_MODE	0.023393
BASEMENTAREA_AVG	0.022746
YEARS_BUILD_MEDI	0.022326
YEARS_BUILD_AVG	0.022149
BASEMENTAREA_MEDI	0.022081
YEARS_BUILD_MODE	0.022068
BASEMENTAREA_MODE	0.019952

AMT_REQ_CREDIT_BUREAU_YEAR	0.019930
CNT_CHILDREN	0.019187
ENTRANCES_AVG	0.019172
ENTRANCES_MEDI	0.019025
COMMONAREA_MEDI	0.018573
COMMONAREA_AVG	0.018550
ENTRANCES_MODE	0.017387
COMMONAREA_MODE	0.016340
NONLIVINGAREA_AVG	0.013578
NONLIVINGAREA_MEDI	0.013337
AMT_ANNUITY	0.012817
NONLIVINGAREA_MODE	0.012711
AMT_REQ_CREDIT_BUREAU_MON	0.012462
FLAG_DOCUMENT_16	0.011615
FLAG_DOCUMENT_13	0.011583
LANDAREA_MEDI	0.011256
LANDAREA_AVG	0.010885
LANDAREA_MODE	0.010174
YEARS_BEGINEXPLUATATION_MEDI	0.009993
YEARS_BEGINEXPLUATATION_AVG	0.009728
FLAG_DOCUMENT_14	0.009464
CNT_FAM_MEMBERS	0.009308
OBS_30_CNT_SOCIAL_CIRCLE	0.009131
YEARS_BEGINEXPLUATATION_MODE	0.009036
OBS_60_CNT_SOCIAL_CIRCLE	0.009022
FLAG_DOCUMENT_8	0.008040
FLAG_DOCUMENT_18	0.007952
REG_REGION_NOT_WORK_REGION	0.006942
FLAG_DOCUMENT_15	0.006536
REG_REGION_NOT_LIVE_REGION	0.005576
FLAG_DOCUMENT_2	0.005417
FLAG_DOCUMENT_9	0.004352
FLAG_DOCUMENT_11	0.004229
AMT_INCOME_TOTAL	0.003982
FLAG_DOCUMENT_21	0.003709
FLAG_DOCUMENT_17	0.003378
NONLIVINGAPARTMENTS_AVG	0.003176
LIVE_REGION_NOT_WORK_REGION	0.002819
NONLIVINGAPARTMENTS_MEDI	0.002757
AMT_REQ_CREDIT_BUREAU_DAY	0.002704
FLAG_DOCUMENT_4	0.002672
AMT_REQ_CREDIT_BUREAU_QRT	0.002022
FLAG_EMAIL	0.001758
NONLIVINGAPARTMENTS_MODE	0.001557
FLAG_DOCUMENT_7	0.001520
FLAG_DOCUMENT_10	0.001414
FLAG_DOCUMENT_19	0.001358
OWN_CAR_AGE	0.001097

AMT_REQ_CREDIT_BUREAU_HOUR	0.000930
AMT_REQ_CREDIT_BUREAU_WEEK	0.000788
FLAG_DOCUMENT_12	0.000756
FLAG_MOBIL	0.000534
FLAG_CONT_MOBILE	0.000370
FLAG_DOCUMENT_5	0.000316
FLAG_DOCUMENT_20	0.000215

Name: TARGET, dtype: float64

```
[ ]: with pd.option_context('display.max_rows', None):
      print(abs_corr[abs_corr < 0.005])
```

FLAG_DOCUMENT_9	0.004352
FLAG_DOCUMENT_11	0.004229
AMT_INCOME_TOTAL	0.003982
FLAG_DOCUMENT_21	0.003709
FLAG_DOCUMENT_17	0.003378
NONLIVINGAPARTMENTS_AVG	0.003176
LIVE_REGION_NOT_WORK_REGION	0.002819
NONLIVINGAPARTMENTS_MEDI	0.002757
AMT_REQ_CREDIT_BUREAU_DAY	0.002704
FLAG_DOCUMENT_4	0.002672
AMT_REQ_CREDIT_BUREAU_QRT	0.002022
FLAG_EMAIL	0.001758
NONLIVINGAPARTMENTS_MODE	0.001557
FLAG_DOCUMENT_7	0.001520
FLAG_DOCUMENT_10	0.001414
FLAG_DOCUMENT_19	0.001358
OWN_CAR_AGE	0.001097
AMT_REQ_CREDIT_BUREAU_HOUR	0.000930
AMT_REQ_CREDIT_BUREAU_WEEK	0.000788
FLAG_DOCUMENT_12	0.000756
FLAG_MOBIL	0.000534
FLAG_CONT_MOBILE	0.000370
FLAG_DOCUMENT_5	0.000316
FLAG_DOCUMENT_20	0.000215

Name: TARGET, dtype: float64

```
[ ]: # variables corrélées les unes aux autres, mat_corr
      # tester importance features
```

```
[ ]: print(correlations.apply(lambda x: np.abs(x))[-10:])
```

FLAG_EMP_PHONE	0.045982
NAME_EDUCATION_TYPE_secondary_or_secondary_special	0.049824
REG_CITY_NOT_WORK_CITY	0.050994
DAYS_ID_PUBLISH	0.051457
DAYS_LAST_PHONE_CHANGE	0.055218

NAME_INCOME_TYPE_working	0.057481
REGION_RATING_CLIENT	0.058899
REGION_RATING_CLIENT_W_CITY	0.060893
DAYS_BIRTH	0.078239
TARGET	1.000000

Name: TARGET, dtype: float64

```
[ ]: X_scl = df_trans_scl.values
      print(X_scl.var(axis=0))
```

```
[1.44448088e-03 4.10927498e-06 1.00996288e-02 3.19501489e-03
 8.48302708e-03 3.66802619e-02 6.05143786e-02 1.35951792e-01
 2.03885872e-02 4.39879059e-02 6.11355051e-03 2.29732128e-03
 6.47786679e-02 6.31859256e-02 2.01618632e-02 2.16194300e-02
 4.98358438e-02 3.80757059e-02 4.74986938e-05 1.72097226e-04
 4.77553478e-05 2.27172798e-04 3.71097209e-02 3.80386548e-04
 1.31069304e-04 5.68391061e-04 1.00702485e-03 8.12703842e-06
 4.98729896e-03 5.77010424e-03 2.81925431e-03 1.79644362e-03
 4.29907715e-03 1.74179461e-03 8.45836615e-03 4.96996966e-03
 1.05104363e-02 8.37328879e-03 2.67736984e-03 2.71206042e-03
 6.08858393e-03 6.96404921e-04 2.16638728e-03 6.10675729e-03
 2.83740315e-03 1.84126502e-03 4.20914420e-03 1.86596163e-03
 9.96210329e-03 5.03373350e-03 1.14485254e-02 8.54947983e-03
 2.82489877e-03 2.91948267e-03 6.56080992e-03 7.03071235e-04
 2.36292280e-03 5.96613858e-03 4.80891258e-03 2.16096979e-03
 5.16802782e-03 2.05083894e-03 9.55041684e-03 5.07587637e-03
 1.11487664e-02 8.43558360e-03 3.73265071e-03 3.58437236e-03
 9.03809302e-03 6.68427666e-04 2.39275598e-03 8.59979408e-03
 8.61473666e-02 2.24434599e-01 2.12490867e-01 7.51314902e-03
 2.24923081e-01 3.20863601e+00 3.25190562e-06 1.47670791e-01
 1.59620688e-01 1.86311570e-03 2.02068065e-01 5.35027726e-02
 1.49148277e-02 4.81914328e-02 3.90055776e-02 7.20618246e-02
 1.77344708e-01 1.47314715e-01 4.22731233e-05 2.05890179e-01
 8.12912954e-05 1.48864461e-02 8.03016355e-02 1.91826244e-04
 7.47539054e-02 3.88061837e-03 2.27628952e-05 3.89675100e-03
 6.50379008e-06 3.51265098e-03 2.92785740e-03 1.20824942e-03
 9.82953296e-03 2.66586022e-04 8.06369699e-03 5.94746518e-04
 5.07041574e-04 3.34835178e-04 1.05111406e-02 1.13514961e-01
 8.80492652e-04 2.80822867e-03 5.72276137e-03 3.56071892e-02
 1.55021560e-01 4.18382332e-03 3.25181044e-05 1.78653574e-01
 1.62593166e-05 1.47620853e-01 6.55953177e-02 5.85310652e-05
 7.15370379e-05 2.49733666e-01 5.33029831e-04 1.84181175e-01
 3.23030501e-02 1.22553212e-02 2.05820488e-01 8.74505682e-02
 2.30740000e-01 6.01571297e-02 1.25941128e-01 6.50379008e-06
 4.95797772e-02 3.63533732e-03 9.99647284e-02 3.50436798e-02
 8.43784007e-03 1.56206634e-02 4.59295596e-02 3.08927382e-02
 1.49022139e-02 1.89620169e-02 8.16172513e-02 5.68357039e-02
 3.56373026e-02 1.82747688e-03 1.70758208e-03 1.47254267e-01]
```

```

6.75993541e-03 6.46669095e-02 2.69909016e-02 8.54970695e-03
2.43622477e-03 9.34951123e-02 4.22574121e-03 2.13784384e-02
2.15201132e-01 4.36436723e-03 1.39312582e-03 7.91651870e-03
8.08608976e-03 1.90807957e-02 3.31397827e-02 1.72217181e-01
8.44783343e-04 2.13784384e-02 1.23095724e-03 3.07977648e-03
1.81775676e-03 3.26882685e-02 3.13148295e-03 9.52663970e-03
3.36732503e-03 3.54333224e-04 8.71586134e-03 1.19851718e-03
2.17830914e-04 1.48715937e-03 1.05461503e-02 2.84379699e-03
1.94410349e-03 3.64081961e-04 4.23218980e-03 7.80398974e-05
1.08324975e-02 1.93762496e-03 2.18726241e-02 9.90850703e-04
3.50738327e-02 8.49217865e-03 1.02979477e-03 5.13084689e-02
7.55478206e-03 6.96518165e-03 1.28610049e-03 2.76336472e-04
5.85453731e-03 2.80829653e-02 1.04474800e-02 6.37807537e-03
1.09309446e-01 5.09553549e-03 1.87283493e-03 1.13038616e-03
6.14046516e-03 1.12267396e-02 2.08079321e-04 1.59318503e-04
2.04774858e-03 2.48172510e-02 6.53207916e-04 7.11585420e-03
3.84512473e-03 1.72457062e-02 4.29667103e-03 1.47645823e-01
1.81516324e-02 1.79386322e-02 1.82446258e-01 3.77399819e-02
2.16194870e-01 2.49888130e-01 4.85086043e-03 3.92578840e-03
2.49996899e-01 2.91845736e-02 7.41065245e-03 5.75169084e-03
5.25643931e-03 1.68636172e-01 1.66347712e-01 2.49929308e-01
1.71327335e-02]

```

```

[ ]: var_pct = ['{:.3g}%'.format(x*100) for x in X_scl.var(axis=0)]
for c,v in zip(df_trans.columns, var_pct):
    print(c, v)

```

```

CNT_CHILDREN 0.144%
AMT_INCOME_TOTAL 0.000411%
AMT_CREDIT 1.01%
AMT_ANNUITY 0.32%
AMT_GOODS_PRICE 0.848%
REGION_POPULATION_RELATIVE 3.67%
DAYS_BIRTH 6.05%
DAYS_EMPLOYED 13.6%
DAYS_REGISTRATION 2.04%
DAYS_ID_PUBLISH 4.4%
OWN_CAR_AGE 0.611%
CNT_FAM_MEMBERS 0.23%
REGION_RATING_CLIENT 6.48%
REGION_RATING_CLIENT_W_CITY 6.32%
HOUR_APPR_PROCESS_START 2.02%
EXT_SOURCE_1 2.16%
EXT_SOURCE_2 4.98%
EXT_SOURCE_3 3.81%
OBS_30_CNT_SOCIAL_CIRCLE 0.00475%
DEF_30_CNT_SOCIAL_CIRCLE 0.0172%
OBS_60_CNT_SOCIAL_CIRCLE 0.00478%

```


DEF_60_CNT_SOCIAL_CIRCLE 0.0227%
 DAYS_LAST_PHONE_CHANGE 3.71%
 AMT_REQ_CREDIT_BUREAU_HOUR 0.038%
 AMT_REQ_CREDIT_BUREAU_DAY 0.0131%
 AMT_REQ_CREDIT_BUREAU_WEEK 0.0568%
 AMT_REQ_CREDIT_BUREAU_MON 0.101%
 AMT_REQ_CREDIT_BUREAU_QRT 0.000813%
 AMT_REQ_CREDIT_BUREAU_YEAR 0.499%
 APARTMENTS_AVG 0.577%
 BASEMENTAREA_AVG 0.282%
 YEARS_BEGINEXPLUATATION_AVG 0.18%
 YEARS_BUILD_AVG 0.43%
 COMMONAREA_AVG 0.174%
 ELEVATORS_AVG 0.846%
 ENTRANCES_AVG 0.497%
 FLOORSMAX_AVG 1.05%
 FLOORSMIN_AVG 0.837%
 LANDAREA_AVG 0.268%
 LIVINGAPARTMENTS_AVG 0.271%
 LIVINGAREA_AVG 0.609%
 NONLIVINGAPARTMENTS_AVG 0.0696%
 NONLIVINGAREA_AVG 0.217%
 APARTMENTS_MEDI 0.611%
 BASEMENTAREA_MEDI 0.284%
 YEARS_BEGINEXPLUATATION_MEDI 0.184%
 YEARS_BUILD_MEDI 0.421%
 COMMONAREA_MEDI 0.187%
 ELEVATORS_MEDI 0.996%
 ENTRANCES_MEDI 0.503%
 FLOORSMAX_MEDI 1.14%
 FLOORSMIN_MEDI 0.855%
 LANDAREA_MEDI 0.282%
 LIVINGAPARTMENTS_MEDI 0.292%
 LIVINGAREA_MEDI 0.656%
 NONLIVINGAPARTMENTS_MEDI 0.0703%
 NONLIVINGAREA_MEDI 0.236%
 APARTMENTS_MODE 0.597%
 BASEMENTAREA_MODE 0.481%
 YEARS_BEGINEXPLUATATION_MODE 0.216%
 YEARS_BUILD_MODE 0.517%
 COMMONAREA_MODE 0.205%
 ELEVATORS_MODE 0.955%
 ENTRANCES_MODE 0.508%
 FLOORSMAX_MODE 1.11%
 FLOORSMIN_MODE 0.844%
 LANDAREA_MODE 0.373%
 LIVINGAPARTMENTS_MODE 0.358%
 LIVINGAREA_MODE 0.904%

NONLIVINGAPARTMENTS_MODE 0.0668%
NONLIVINGAREA_MODE 0.239%
TOTALAREA_MODE 0.86%
NAME_CONTRACT_TYPE 8.61%
FLAG_OWN_CAR 22.4%
FLAG_OWN_REALTY 21.2%
EMERGENCYSTATE_MODE 0.751%
CODE_GENDER 22.5%
WEEKDAY_APPR_PROCESS_START 321%
FLAG_MOBIL 0.000325%
FLAG_EMP_PHONE 14.8%
FLAG_WORK_PHONE 16%
FLAG_CONT_MOBILE 0.186%
FLAG_PHONE 20.2%
FLAG_EMAIL 5.35%
REG_REGION_NOT_LIVE_REGION 1.49%
REG_REGION_NOT_WORK_REGION 4.82%
LIVE_REGION_NOT_WORK_REGION 3.9%
REG_CITY_NOT_LIVE_CITY 7.21%
REG_CITY_NOT_WORK_CITY 17.7%
LIVE_CITY_NOT_WORK_CITY 14.7%
FLAG_DOCUMENT_2 0.00423%
FLAG_DOCUMENT_3 20.6%
FLAG_DOCUMENT_4 0.00813%
FLAG_DOCUMENT_5 1.49%
FLAG_DOCUMENT_6 8.03%
FLAG_DOCUMENT_7 0.0192%
FLAG_DOCUMENT_8 7.48%
FLAG_DOCUMENT_9 0.388%
FLAG_DOCUMENT_10 0.00228%
FLAG_DOCUMENT_11 0.39%
FLAG_DOCUMENT_12 0.00065%
FLAG_DOCUMENT_13 0.351%
FLAG_DOCUMENT_14 0.293%
FLAG_DOCUMENT_15 0.121%
FLAG_DOCUMENT_16 0.983%
FLAG_DOCUMENT_17 0.0267%
FLAG_DOCUMENT_18 0.806%
FLAG_DOCUMENT_19 0.0595%
FLAG_DOCUMENT_20 0.0507%
FLAG_DOCUMENT_21 0.0335%
NAME_TYPE_SUITE_children 1.05%
NAME_TYPE_SUITE_family 11.4%
NAME_TYPE_SUITE_group_of_people 0.088%
NAME_TYPE_SUITE_other_a 0.281%
NAME_TYPE_SUITE_other_b 0.572%
NAME_TYPE_SUITE_spouse_or_partner 3.56%
NAME_TYPE_SUITE_unaccompanied 15.5%

NAME_TYPE_SUITE_unknown 0.418%
NAME_INCOME_TYPE_businessman 0.00325%
NAME_INCOME_TYPE_commercial_associate 17.9%
NAME_INCOME_TYPE_maternity_leave 0.00163%
NAME_INCOME_TYPE_pensioner 14.8%
NAME_INCOME_TYPE_state_servant 6.56%
NAME_INCOME_TYPE_student 0.00585%
NAME_INCOME_TYPE_unemployed 0.00715%
NAME_INCOME_TYPE_working 25%
NAME_EDUCATION_TYPE_academic_degree 0.0533%
NAME_EDUCATION_TYPE_higher_education 18.4%
NAME_EDUCATION_TYPE_incomplete_higher 3.23%
NAME_EDUCATION_TYPE_lower_secondary 1.23%
NAME_EDUCATION_TYPE_secondary_or_secondary_special 20.6%
NAME_FAMILY_STATUS_civil_marriage 8.75%
NAME_FAMILY_STATUS_married 23.1%
NAME_FAMILY_STATUS_separated 6.02%
NAME_FAMILY_STATUS_single_or_not_married 12.6%
NAME_FAMILY_STATUS_unknown 0.00065%
NAME_FAMILY_STATUS_widow 4.96%
NAME_HOUSING_TYPE_coop_apartment 0.364%
NAME_HOUSING_TYPE_house_or_apartment 10%
NAME_HOUSING_TYPE_municipal_apartment 3.5%
NAME_HOUSING_TYPE_office_apartment 0.844%
NAME_HOUSING_TYPE_rented_apartment 1.56%
NAME_HOUSING_TYPE_with_parents 4.59%
OCCUPATION_TYPE_accountants 3.09%
OCCUPATION_TYPE_cleaning_staff 1.49%
OCCUPATION_TYPE_cooking_staff 1.9%
OCCUPATION_TYPE_core_staff 8.16%
OCCUPATION_TYPE_drivers 5.68%
OCCUPATION_TYPE_high_skill_tech_staff 3.56%
OCCUPATION_TYPE_hr_staff 0.183%
OCCUPATION_TYPE_it_staff 0.171%
OCCUPATION_TYPE_laborers 14.7%
OCCUPATION_TYPE_lowskill_laborers 0.676%
OCCUPATION_TYPE_managers 6.47%
OCCUPATION_TYPE_medicine_staff 2.7%
OCCUPATION_TYPE_private_service_staff 0.855%
OCCUPATION_TYPE_realty_agents 0.244%
OCCUPATION_TYPE_sales_staff 9.35%
OCCUPATION_TYPE_secretaries 0.423%
OCCUPATION_TYPE_security_staff 2.14%
OCCUPATION_TYPE_unknown 21.5%
OCCUPATION_TYPE_waitersorbarmen_staff 0.436%
ORGANIZATION_TYPE_advertising 0.139%
ORGANIZATION_TYPE_agriculture 0.792%
ORGANIZATION_TYPE_bank 0.809%

ORGANIZATION_TYPE_business_entity_type_1 1.91%
 ORGANIZATION_TYPE_business_entity_type_2 3.31%
 ORGANIZATION_TYPE_business_entity_type_3 17.2%
 ORGANIZATION_TYPE_cleaning 0.0845%
 ORGANIZATION_TYPE_construction 2.14%
 ORGANIZATION_TYPE_culture 0.123%
 ORGANIZATION_TYPE_electricity 0.308%
 ORGANIZATION_TYPE_emergency 0.182%
 ORGANIZATION_TYPE_government 3.27%
 ORGANIZATION_TYPE_hotel 0.313%
 ORGANIZATION_TYPE_housing 0.953%
 ORGANIZATION_TYPE_industry_type_1 0.337%
 ORGANIZATION_TYPE_industry_type_10 0.0354%
 ORGANIZATION_TYPE_industry_type_11 0.872%
 ORGANIZATION_TYPE_industry_type_12 0.12%
 ORGANIZATION_TYPE_industry_type_13 0.0218%
 ORGANIZATION_TYPE_industry_type_2 0.149%
 ORGANIZATION_TYPE_industry_type_3 1.05%
 ORGANIZATION_TYPE_industry_type_4 0.284%
 ORGANIZATION_TYPE_industry_type_5 0.194%
 ORGANIZATION_TYPE_industry_type_6 0.0364%
 ORGANIZATION_TYPE_industry_type_7 0.423%
 ORGANIZATION_TYPE_industry_type_8 0.0078%
 ORGANIZATION_TYPE_industry_type_9 1.08%
 ORGANIZATION_TYPE_insurance 0.194%
 ORGANIZATION_TYPE_kindergarten 2.19%
 ORGANIZATION_TYPE_legal_services 0.0991%
 ORGANIZATION_TYPE_medicine 3.51%
 ORGANIZATION_TYPE_military 0.849%
 ORGANIZATION_TYPE_mobile 0.103%
 ORGANIZATION_TYPE_other 5.13%
 ORGANIZATION_TYPE_police 0.755%
 ORGANIZATION_TYPE_postal 0.697%
 ORGANIZATION_TYPE_realtor 0.129%
 ORGANIZATION_TYPE_religion 0.0276%
 ORGANIZATION_TYPE_restaurant 0.585%
 ORGANIZATION_TYPE_school 2.81%
 ORGANIZATION_TYPE_security 1.04%
 ORGANIZATION_TYPE_security_ministries 0.638%
 ORGANIZATION_TYPE_selfemployed 10.9%
 ORGANIZATION_TYPE_services 0.51%
 ORGANIZATION_TYPE_telecom 0.187%
 ORGANIZATION_TYPE_trade_type_1 0.113%
 ORGANIZATION_TYPE_trade_type_2 0.614%
 ORGANIZATION_TYPE_trade_type_3 1.12%
 ORGANIZATION_TYPE_trade_type_4 0.0208%
 ORGANIZATION_TYPE_trade_type_5 0.0159%
 ORGANIZATION_TYPE_trade_type_6 0.205%

ORGANIZATION_TYPE_trade_type_7 2.48%
 ORGANIZATION_TYPE_transport_type_1 0.0653%
 ORGANIZATION_TYPE_transport_type_2 0.712%
 ORGANIZATION_TYPE_transport_type_3 0.385%
 ORGANIZATION_TYPE_transport_type_4 1.72%
 ORGANIZATION_TYPE_university 0.43%
 ORGANIZATION_TYPE_xna 14.8%
 FONDKAPREMONT_MODE_not_specified 1.82%
 FONDKAPREMONT_MODE_org_spec_account 1.79%
 FONDKAPREMONT_MODE_reg_oper_account 18.2%
 FONDKAPREMONT_MODE_reg_oper_spec_account 3.77%
 FONDKAPREMONT_MODE_unknown 21.6%
 HOUSETYPE_MODE_block_of_flats 25%
 HOUSETYPE_MODE_specific_housing 0.485%
 HOUSETYPE_MODE_terraced_house 0.393%
 HOUSETYPE_MODE_unknown 25%
 WALLSMATERIAL_MODE_block 2.92%
 WALLSMATERIAL_MODE_mixed 0.741%
 WALLSMATERIAL_MODE_monolithic 0.575%
 WALLSMATERIAL_MODE_others 0.526%
 WALLSMATERIAL_MODE_panel 16.9%
 WALLSMATERIAL_MODE_stone_or_brick 16.6%
 WALLSMATERIAL_MODE_unknown 25%
 WALLSMATERIAL_MODE_wooden 1.71%