# HomeloanDefaultRisk

April 27, 2022

```
[12]:    SK_ID_CURR  TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR  \
      0      100002       1         Cash loans           M            N
      1      100003       0         Cash loans           F            N
      2      100004       0    Revolving loans           M            Y
      3      100006       0         Cash loans           F            N
      4      100007       0         Cash loans           M            N

        FLAG_OWN_REALTY  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  AMT_ANNUITY  \
      0               Y             0          202500.0    406597.5      24700.5
      1               N             0          270000.0   1293502.5      35698.5
      2               Y             0           67500.0    135000.0       6750.0
      3               Y             0          135000.0    312682.5      29686.5
      4               Y             0          121500.0    513000.0      21865.5

         …  FLAG_DOCUMENT_18 FLAG_DOCUMENT_19 FLAG_DOCUMENT_20 FLAG_DOCUMENT_21  \
      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

        AMT_REQ_CREDIT_BUREAU_HOUR AMT_REQ_CREDIT_BUREAU_DAY  \
      0                        0.0                       0.0
      1                        0.0                       0.0
      2                        0.0                       0.0
      3                        NaN                       NaN
      4                        0.0                       0.0

         AMT_REQ_CREDIT_BUREAU_WEEK  AMT_REQ_CREDIT_BUREAU_MON  \
      0                         0.0                        0.0
      1                         0.0                        0.0
      2                         0.0                        0.0
      3                         NaN                        NaN
      4                         0.0                        0.0

         AMT_REQ_CREDIT_BUREAU_QRT  AMT_REQ_CREDIT_BUREAU_YEAR
      0                        0.0                         1.0
```

```
1                         0.0                           0.0
2                         0.0                           0.0
3                         NaN                           NaN
4                         0.0                           0.0

[5 rows x 122 columns]
```

[13]:
```
(50000, 122)
```

[14]:
```
float64    65
int64      41
object     16
dtype: int64
```

[15]:
```
NAME_CONTRACT_TYPE             2
CODE_GENDER                    3
FLAG_OWN_CAR                   2
FLAG_OWN_REALTY                2
NAME_TYPE_SUITE                7
NAME_INCOME_TYPE               8
NAME_EDUCATION_TYPE            5
NAME_FAMILY_STATUS             6
NAME_HOUSING_TYPE              6
OCCUPATION_TYPE               18
WEEKDAY_APPR_PROCESS_START     7
ORGANIZATION_TYPE             58
FONDKAPREMONT_MODE             4
HOUSETYPE_MODE                 3
WALLSMATERIAL_MODE             7
EMERGENCYSTATE_MODE            2
dtype: int64
```

[16]:
```
COMMONAREA_MEDI                0.69922
COMMONAREA_AVG                 0.69922
COMMONAREA_MODE                0.69922
NONLIVINGAPARTMENTS_MODE       0.69430
NONLIVINGAPARTMENTS_MEDI       0.69430
                                 …
REG_CITY_NOT_LIVE_CITY         0.00000
LIVE_REGION_NOT_WORK_REGION    0.00000
REG_REGION_NOT_WORK_REGION     0.00000
HOUR_APPR_PROCESS_START        0.00000
SK_ID_CURR                     0.00000
Length: 122, dtype: float64
```

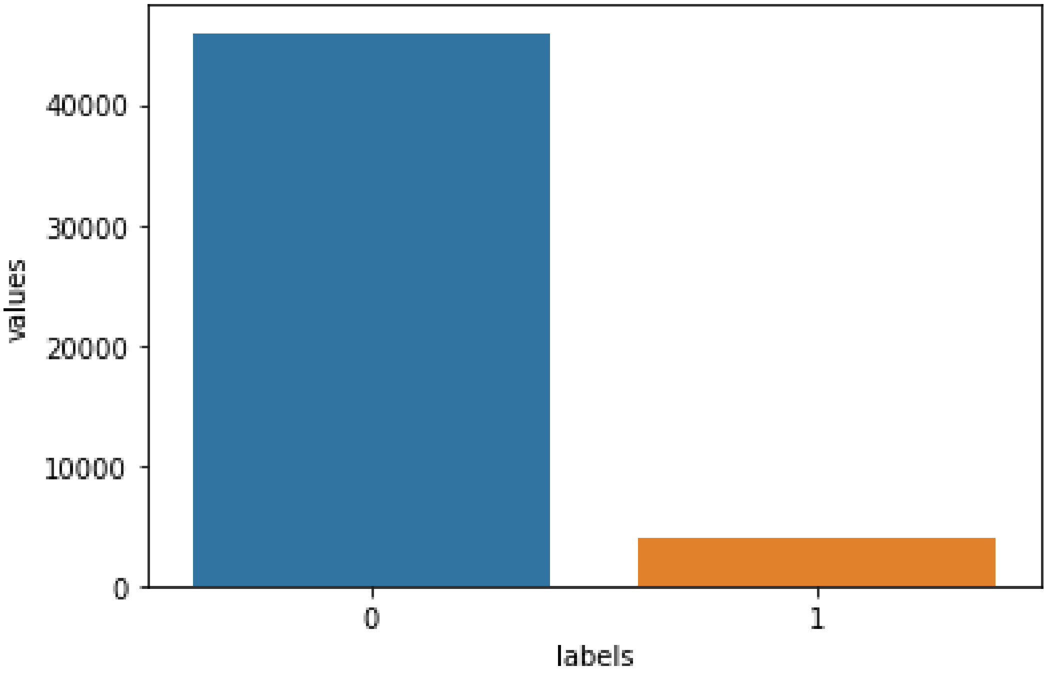
[18]:
```
       SK_ID_CURR  TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR  \
0          100002       1         Cash loans           M            N
```

```
1          100003       0         Cash loans           F            N
2          100004       0    Revolving loans           M            Y
3          100006       0         Cash loans           F            N
4          100007       0         Cash loans           M            N
…           …     …                …         …          …
49995      157872       0         Cash loans           M            N
49996      157873       0         Cash loans           M            N
49997      157874       0         Cash loans           F            N
49998      157875       0         Cash loans           F            N
49999      157876       0         Cash loans           F            N

      FLAG_OWN_REALTY  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  \
0                   Y             0          202500.0    406597.5
1                   N             0          270000.0   1293502.5
2                   Y             0           67500.0    135000.0
3                   Y             0          135000.0    312682.5
4                   Y             0          121500.0    513000.0
…               …           …               …         …
49995               N             0          126000.0   1125000.0
49996               N             1          112500.0    900000.0
49997               Y             0          270000.0    820638.0
49998               Y             0          117000.0    254700.0
49999               Y             0           67500.0    343800.0

       AMT_ANNUITY  …  FLAG_DOCUMENT_18 FLAG_DOCUMENT_19 FLAG_DOCUMENT_20  \
0          24700.5  …                 0                0                0
1          35698.5  …                 0                0                0
2           6750.0  …                 0                0                0
3          29686.5  …                 0                0                0
4          21865.5  …                 0                0                0
…           …   …               …              …              …
49995      47794.5  …                 0                0                0
49996      26316.0  …                 0                0                0
49997      34897.5  …                 0                0                0
49998      14751.0  …                 0                0                0
49999      16155.0  …                 0                0                0

      FLAG_DOCUMENT_21 AMT_REQ_CREDIT_BUREAU_HOUR AMT_REQ_CREDIT_BUREAU_DAY  \
0                    0                        0.0                       0.0
1                    0                        0.0                       0.0
2                    0                        0.0                       0.0
3                    0                        NaN                       NaN
4                    0                        0.0                       0.0
…                …                        …                       …
49995                0                        0.0                       0.0
49996                0                        0.0                       0.0
49997                0                        0.0                       0.0
```

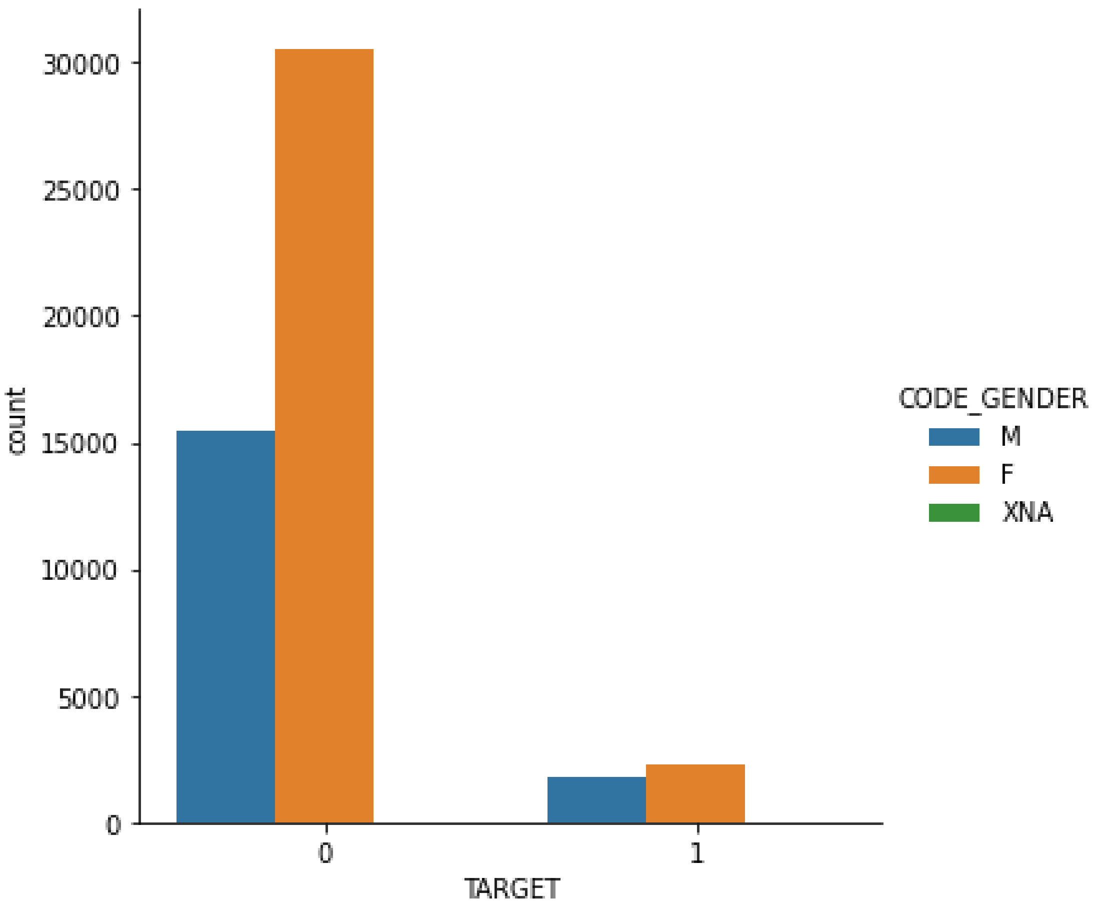
```
49998                0                        0.0                       0.0
49999                0                        0.0                       0.0

       AMT_REQ_CREDIT_BUREAU_WEEK  AMT_REQ_CREDIT_BUREAU_MON  \
0                             0.0                        0.0
1                             0.0                        0.0
2                             0.0                        0.0
3                             NaN                        NaN
4                             0.0                        0.0
…                           …                        …
49995                         0.0                        0.0
49996                         0.0                        0.0
49997                         0.0                        0.0
49998                         0.0                        0.0
49999                         0.0                        0.0

       AMT_REQ_CREDIT_BUREAU_QRT  AMT_REQ_CREDIT_BUREAU_YEAR
0                            0.0                         1.0
1                            0.0                         0.0
2                            0.0                         0.0
3                            NaN                         NaN
4                            0.0                         0.0
…                          …                         …
49995                        0.0                         0.0
49996                        0.0                         2.0
49997                        2.0                         4.0
49998                        0.0                         0.0
49999                        0.0                         1.0

[50000 rows x 81 columns]
```

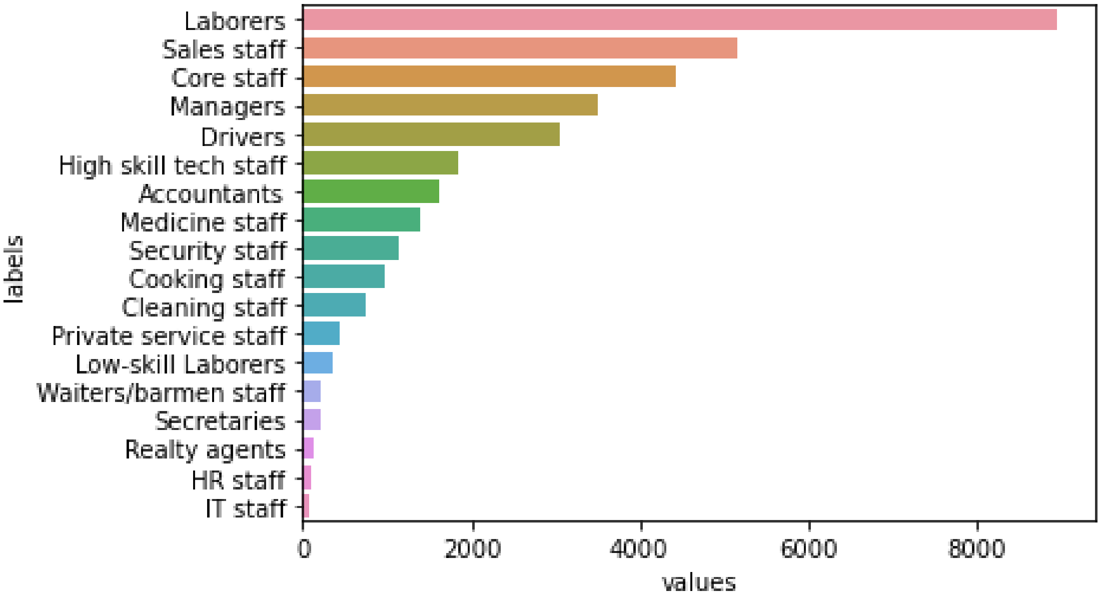
```
<ipython-input-20-608f89c312d9>:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  application_train['EXT_SOURCE_1']=temp_final['EXT_SOURCE_1']
```

### 0.0.1 Distribution of target



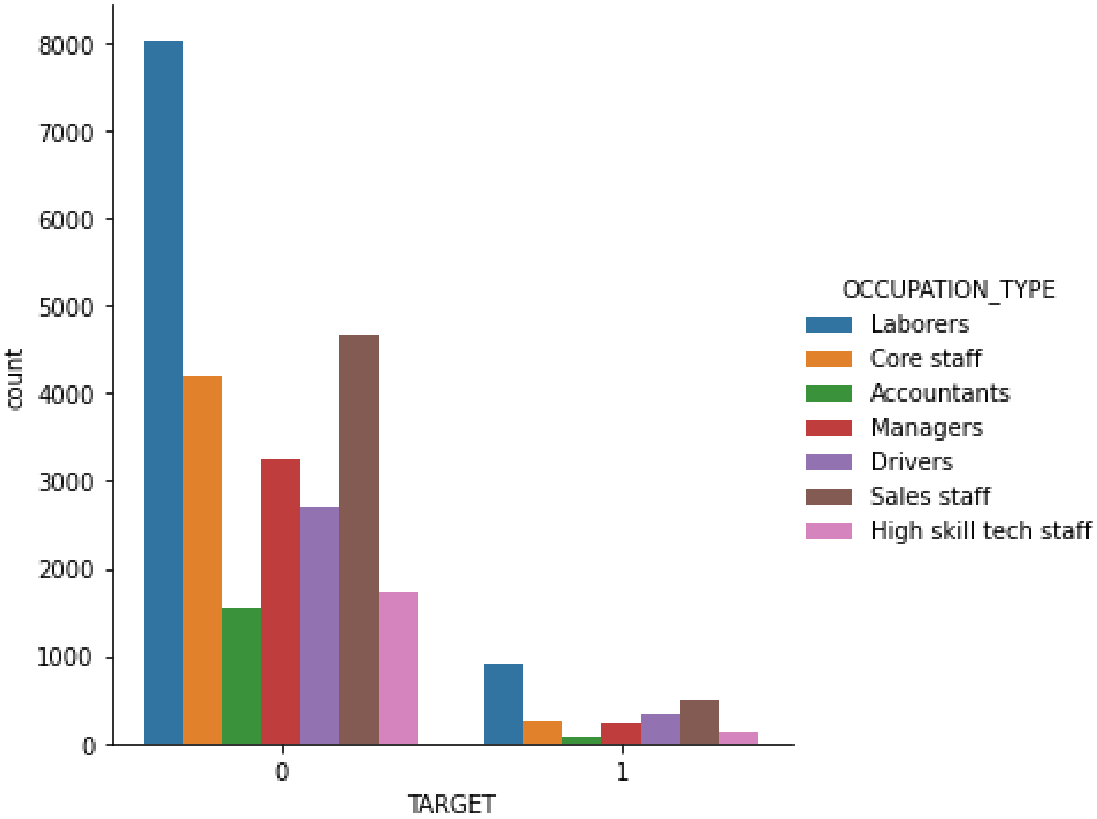
```
[22]: <seaborn.axisgrid.FacetGrid at 0x21cdc249f10>
```


30000
25000
20000
15000
10000
5000
0
count
0
1
TARGET
CODE_GENDER
M
F
XNA

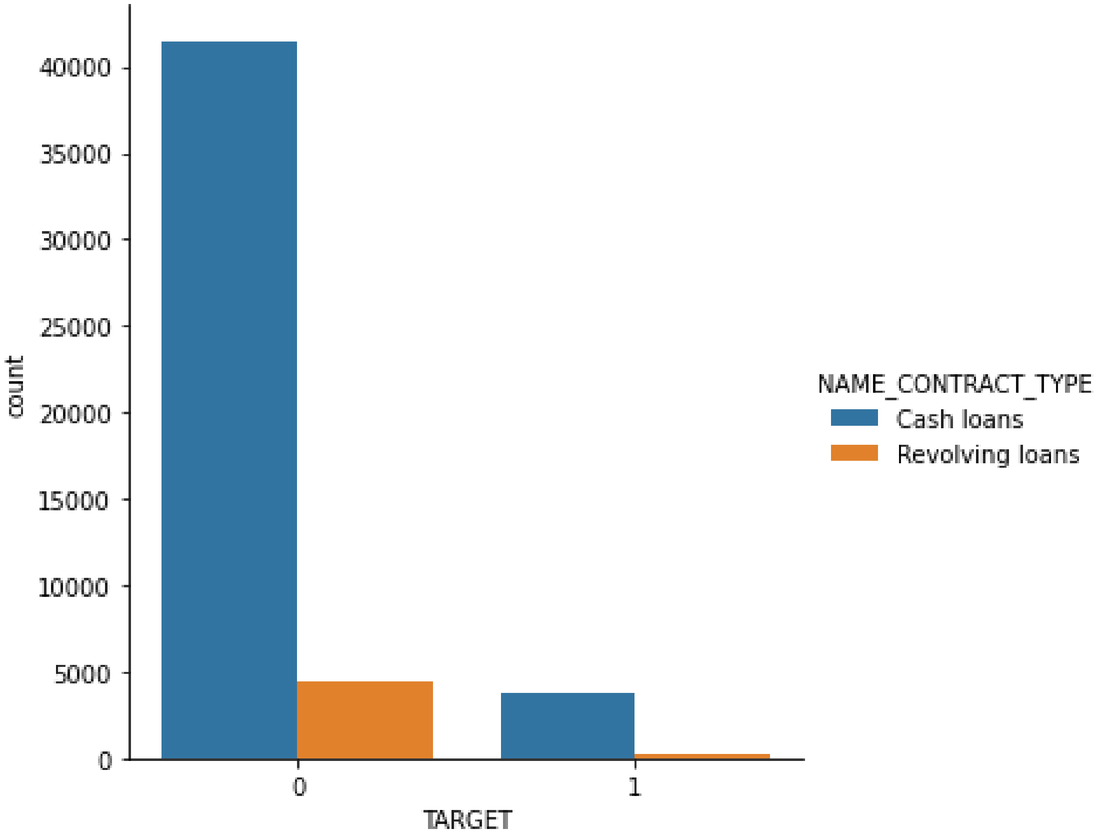

Laborers
Sales staff
Core staff
Managers
Drivers
High skill tech staff
Accountants
Medicine staff
Security staff
Cooking staff
Cleaning staff
Private service staff
Low-skill Laborers
Waiters/barmen staff
Secretaries
Realty agents
HR staff
IT staff
labels
0
2000
4000
6000
8000
values

```
[24]: Laborers                  True
      Sales staff               True
      Core staff                True
      Managers                  True
      Drivers                   True
      High skill tech staff     True
      Accountants               True
      Medicine staff           False
      Security staff           False
      Cooking staff            False
      Cleaning staff           False
      Private service staff    False
      Low-skill Laborers       False
      Waiters/barmen staff     False
      Secretaries              False
      Realty agents            False
      HR staff                 False
      IT staff                 False
      Name: OCCUPATION_TYPE, dtype: bool
```

```
[25]: <seaborn.axisgrid.FacetGrid at 0x21c9ea99520>
```

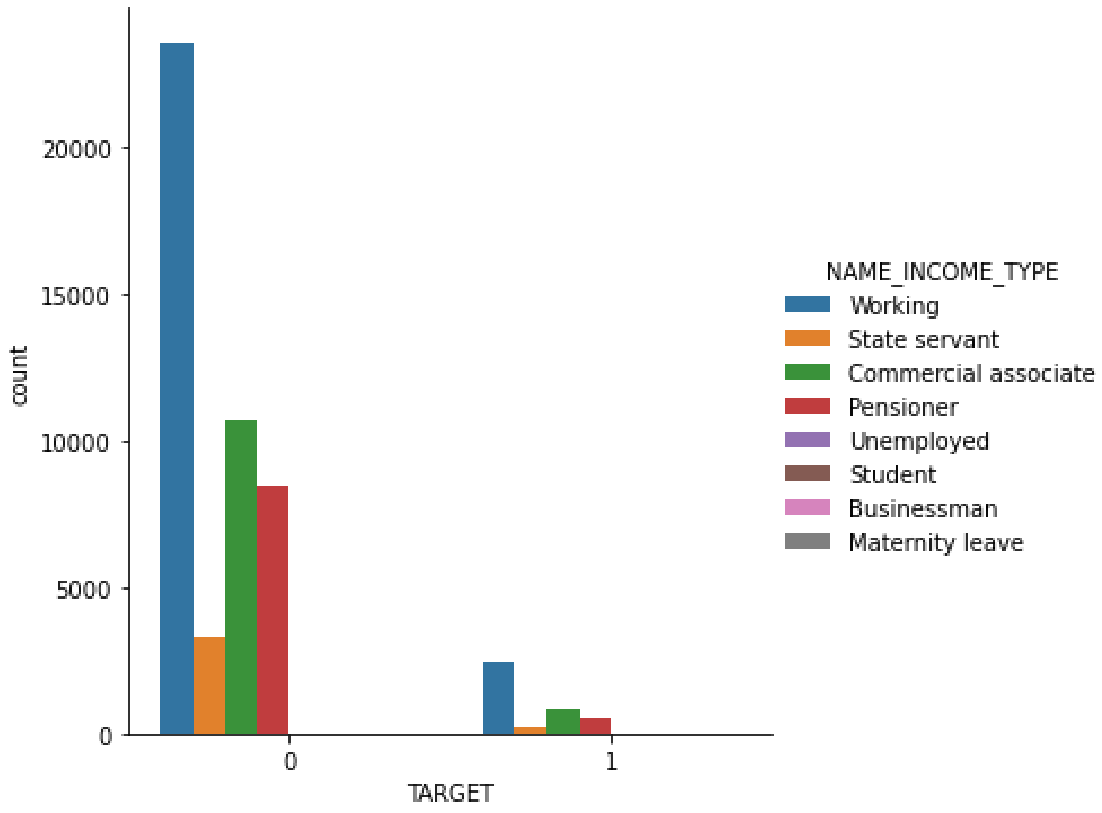


```
[26]: <seaborn.axisgrid.FacetGrid at 0x21c82062280>
```



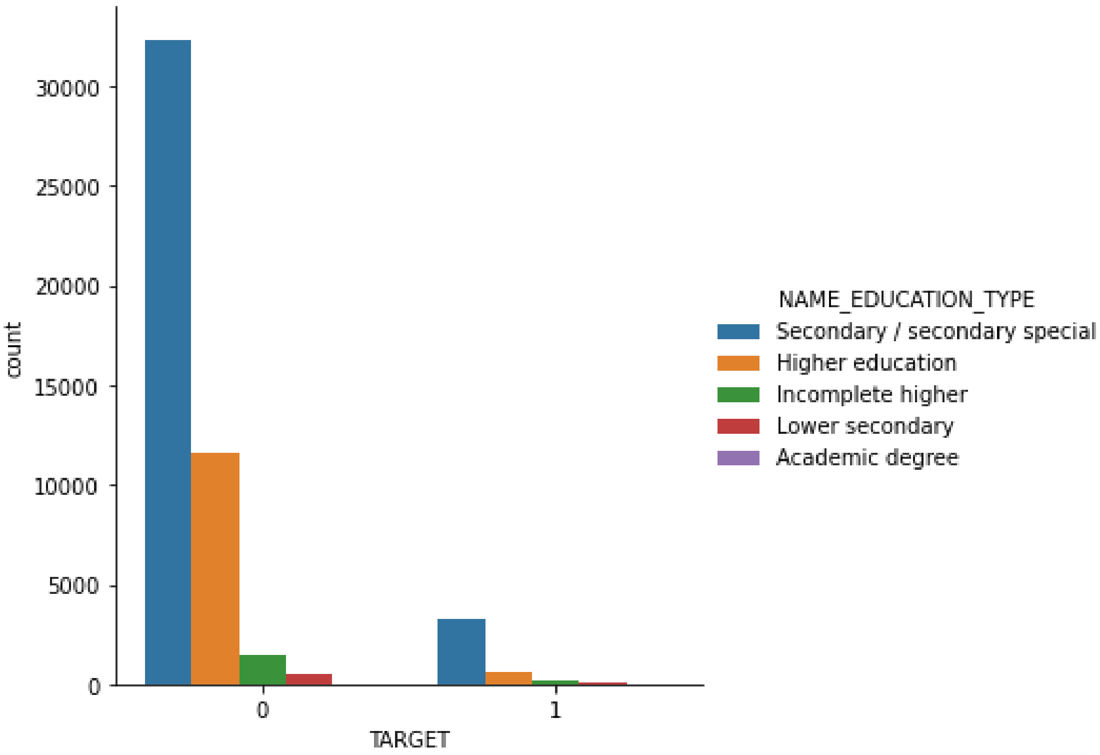
40000
35000
30000
25000
20000
15000
10000
5000
0
count
0
1
TARGET
NAME_CONTRACT_TYPE
Cash loans
Revolving loans

```
[27]: <seaborn.axisgrid.FacetGrid at 0x21c81709b80>
```

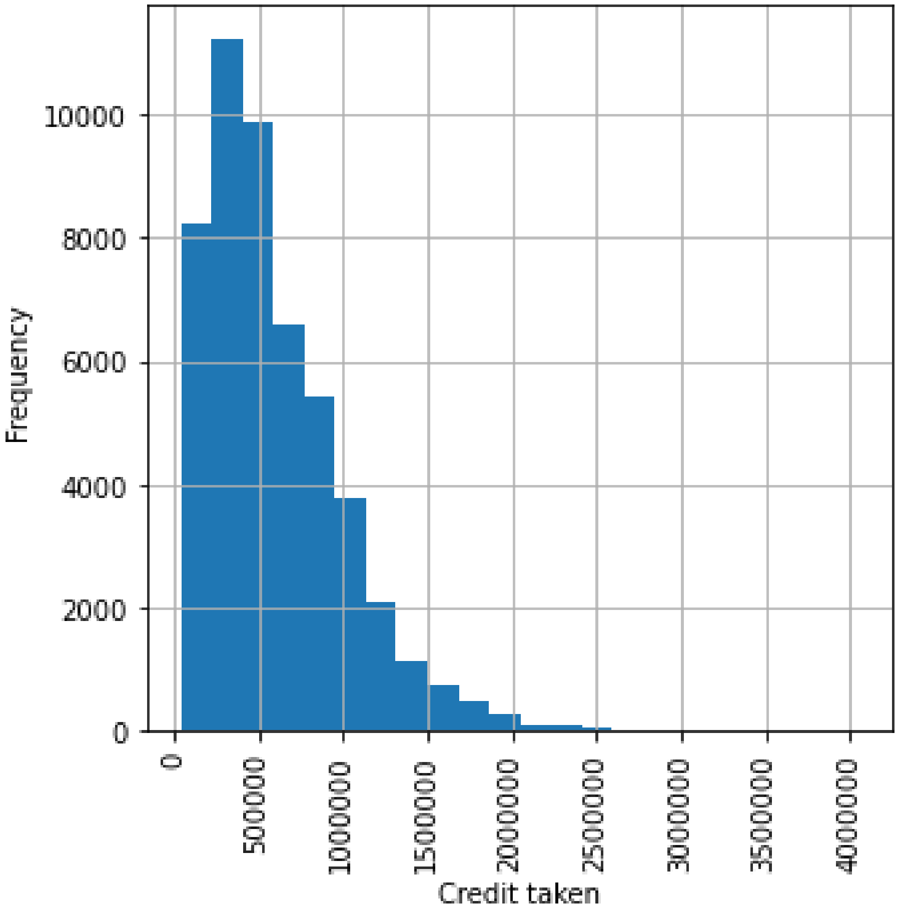


count
0
5000
10000
15000
20000
0
1
TARGET
NAME_INCOME_TYPE
Working
State servant
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave

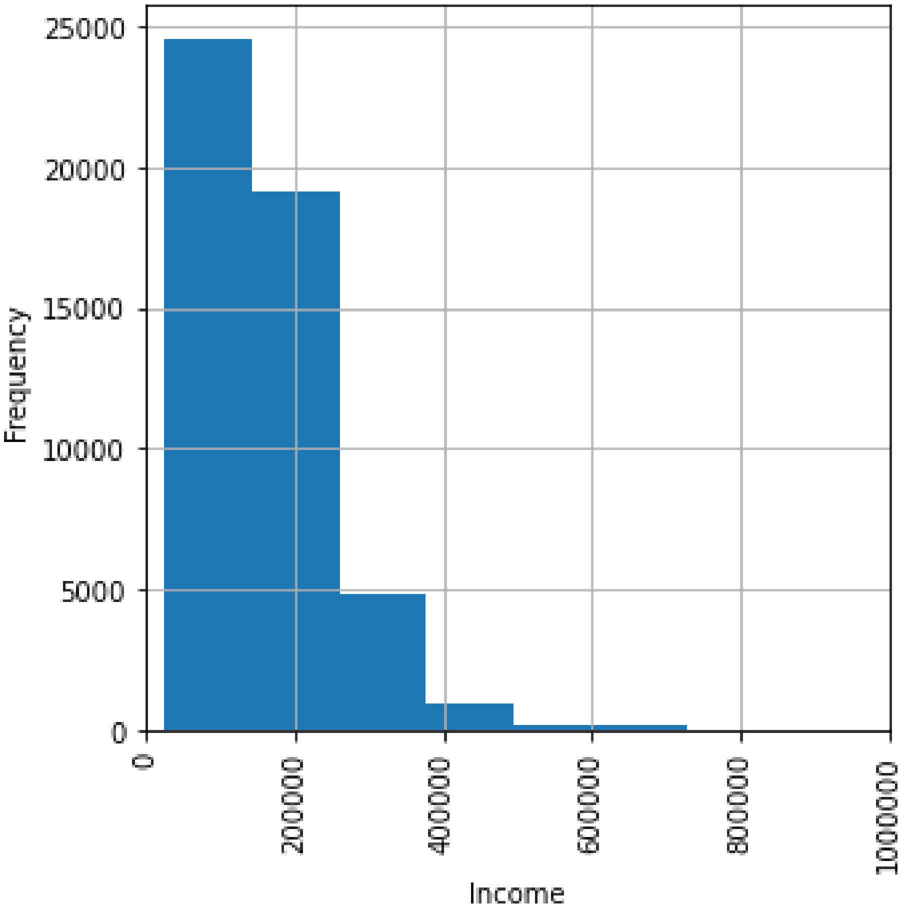
```
[28]: <seaborn.axisgrid.FacetGrid at 0x21c820b4e20>
```



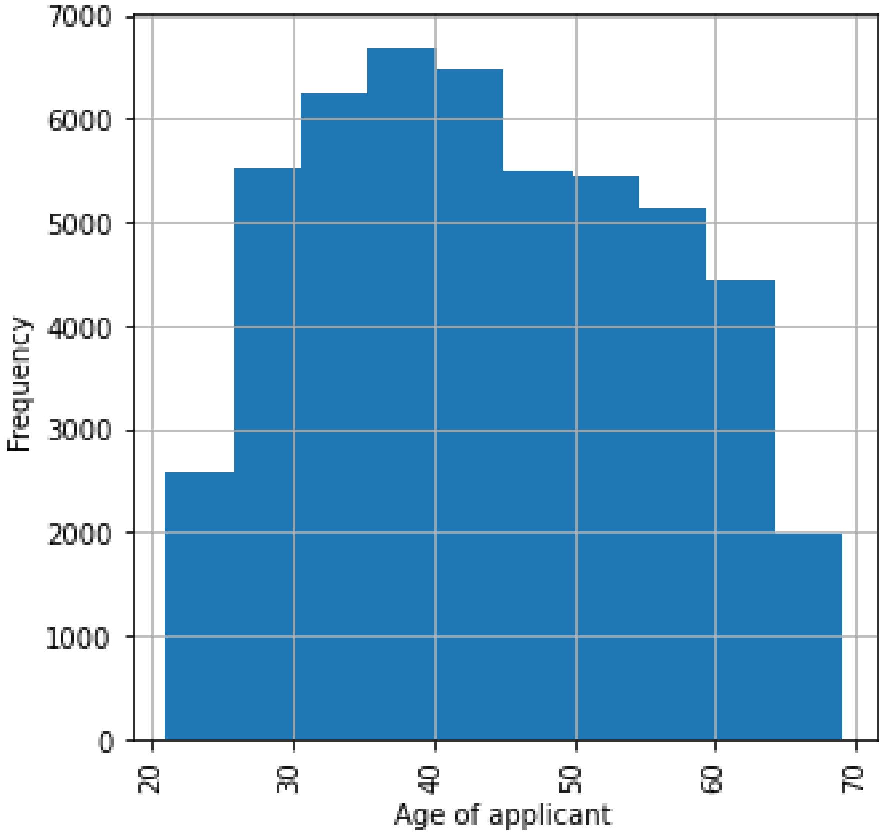
count
30000
25000
20000
15000
10000
5000
0
0
1
TARGET
NAME_EDUCATION_TYPE
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree



```
[30]: (0.0, 1000000.0)
```



```
<ipython-input-31-d8c0895208ad>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  temp['Age'] = application_train['DAYS_BIRTH']*-1/365
```



```
3 columns were label encoded.
```

```
<ipython-input-32-00924bf856fb>:13: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  application_train[col] = le.transform(application_train[col])
```

```
Training Features shape:  (50000, 193)
```

```
Training Features shape:  (50000, 193)
```

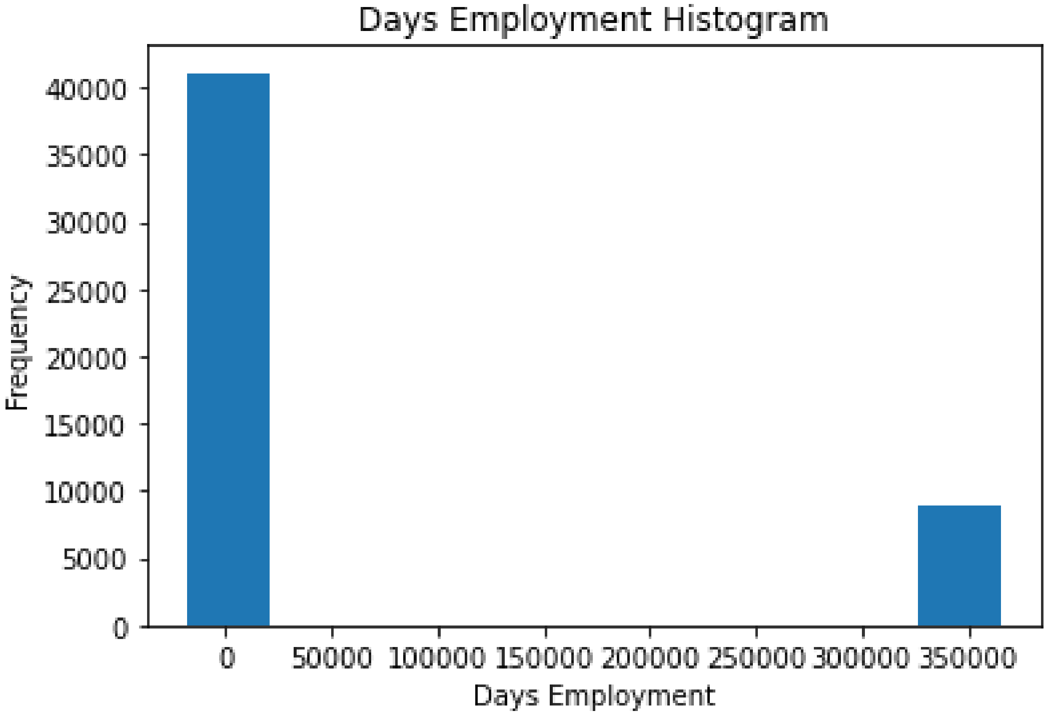
[35]:

```
count    50000.000000
mean        43.896192
std         11.948995
min         21.041096
25%         33.914384
50%         43.098630
75%         53.819178
max         68.997260
```

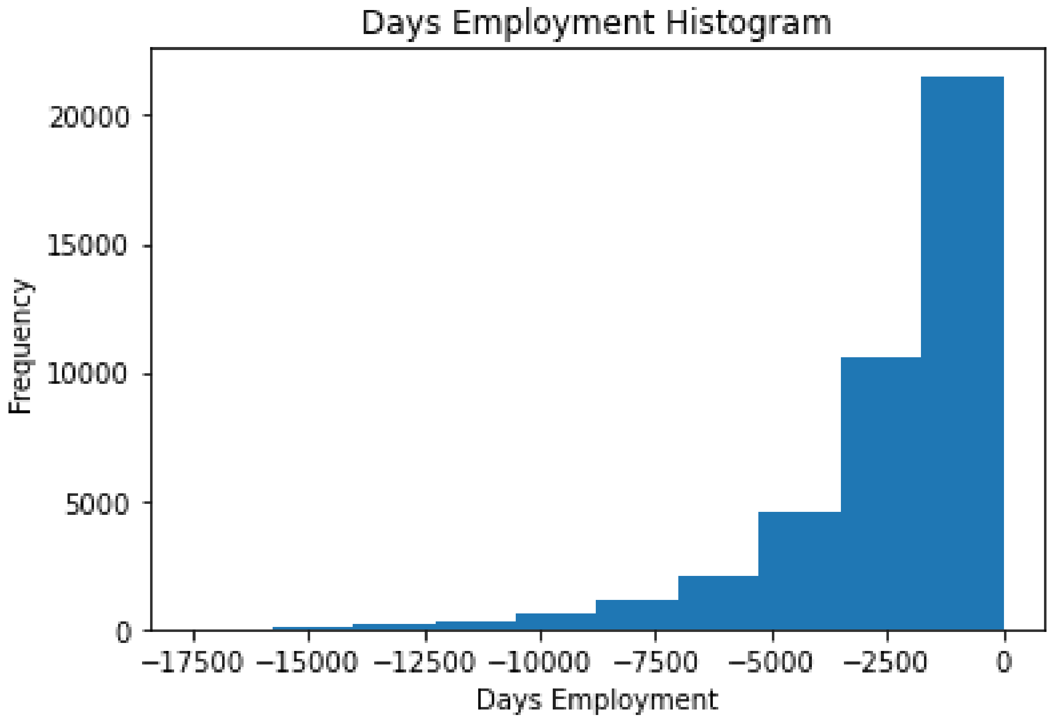
```
Name: DAYS_BIRTH, dtype: float64
```

```
[36]: count     50000.000000
      mean      63218.143580
      std      140793.489022
      min      -17531.000000
      25%       -2786.000000
      50%       -1221.000000
      75%        -292.000000
      max      365243.000000
      Name: DAYS_EMPLOYED, dtype: float64
```

The maximum value (besides being positive) is about 1000 years! That doesn't look right!



```
The non-anomalies default on 8.58% of loans
The anomalies default on 5.64% of loans
There are 8924 anomalous days of employment
```



```
Most Positive Correlations:
 LIVE_CITY_NOT_WORK_CITY                              0.032264
OCCUPATION_TYPE_Laborers                             0.038188
REG_CITY_NOT_LIVE_CITY                               0.038749
FLAG_EMP_PHONE                                       0.041406
DEF_30_CNT_SOCIAL_CIRCLE                             0.041560
DAYS_REGISTRATION                                    0.042334
DEF_60_CNT_SOCIAL_CIRCLE                             0.044210
FLAG_DOCUMENT_3                                      0.045046
DAYS_ID_PUBLISH                                      0.046925
REG_CITY_NOT_WORK_CITY                               0.048438
NAME_INCOME_TYPE_Working                             0.053948
NAME_EDUCATION_TYPE_Secondary / secondary special    0.055914
DAYS_LAST_PHONE_CHANGE                               0.056131
CODE_GENDER_M                                        0.058687
REGION_RATING_CLIENT                                 0.066131
REGION_RATING_CLIENT_W_CITY                          0.067080
DAYS_EMPLOYED                                        0.076535
DAYS_BIRTH                                           0.076792
TARGET                                               1.000000
FLAG_DOCUMENT_12                                          NaN
Name: TARGET, dtype: float64

Most Negative Correlations:
 EXT_SOURCE_3                             -0.159550
EXT_SOURCE_2                             -0.158265
EXT_SOURCE_1                             -0.156806
Age                                      -0.076792
NAME_EDUCATION_TYPE_Higher education    -0.064011
CODE_GENDER_F                            -0.058661
EMERGENCYSTATE_MODE_No                   -0.048843
NAME_INCOME_TYPE_Pensioner               -0.041708
DAYS_EMPLOYED_ANOM                       -0.041378
ORGANIZATION_TYPE_XNA                    -0.041378
AMT_GOODS_PRICE                          -0.041290
REGION_POPULATION_RELATIVE               -0.040797
NAME_CONTRACT_TYPE                       -0.036767
FLAG_PHONE                               -0.032688
AMT_CREDIT                               -0.032424
HOUR_APPR_PROCESS_START                  -0.032040
NAME_FAMILY_STATUS_Married               -0.029020
OCCUPATION_TYPE_Core staff               -0.027672
FLOORSMAX_AVG                            -0.026643
FLOORSMAX_MEDI                           -0.026141
Name: TARGET, dtype: float64
```

### 0.0.2 Feature Engineering

```
Polynomial Features shape:  (50000, 21)
```

```
[43]: ['1',
       'EXT_SOURCE_1',
       'EXT_SOURCE_2',
       'EXT_SOURCE_3',
       'DAYS_EMPLOYED',
       'DAYS_BIRTH',
       'EXT_SOURCE_1^2',
       'EXT_SOURCE_1 EXT_SOURCE_2',
       'EXT_SOURCE_1 EXT_SOURCE_3',
       'EXT_SOURCE_1 DAYS_EMPLOYED',
       'EXT_SOURCE_1 DAYS_BIRTH',
       'EXT_SOURCE_2^2',
       'EXT_SOURCE_2 EXT_SOURCE_3',
       'EXT_SOURCE_2 DAYS_EMPLOYED',
       'EXT_SOURCE_2 DAYS_BIRTH']
```

```
EXT_SOURCE_2 EXT_SOURCE_3   -0.194199
EXT_SOURCE_1 EXT_SOURCE_3   -0.167388
EXT_SOURCE_1 EXT_SOURCE_2   -0.165564
EXT_SOURCE_3                -0.159550
EXT_SOURCE_2                -0.158265
EXT_SOURCE_2^2              -0.146820
```

```
EXT_SOURCE_3^2              -0.144592
EXT_SOURCE_1                -0.099736
EXT_SOURCE_1^2              -0.090331
DAYS_EMPLOYED^2             -0.075400
Name: TARGET, dtype: float64
EXT_SOURCE_1 DAYS_EMPLOYED    0.103650
EXT_SOURCE_3 DAYS_EMPLOYED    0.151375
EXT_SOURCE_2 DAYS_EMPLOYED    0.154554
TARGET                        1.000000
1                                  NaN
Name: TARGET, dtype: float64
```

```
Training data with polynomial features shape:  (50000, 30)
```

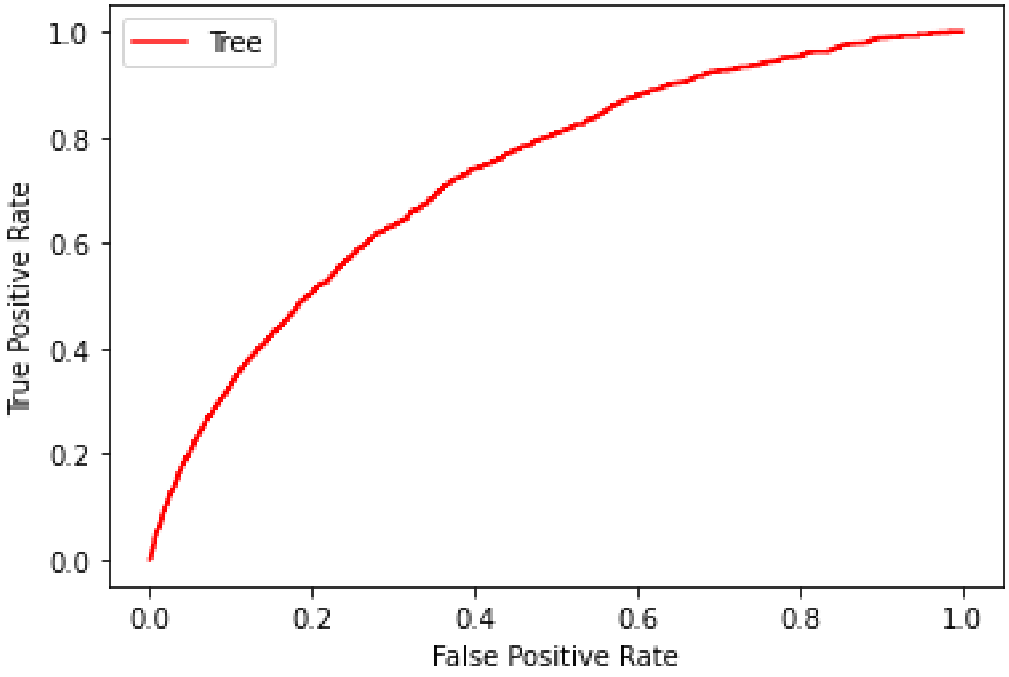
```
Training data shape:  (50000, 29)
```

### 0.0.3 Logistic Regression

```
Confusion Matrix :
        0  1
0  13788  0
1   1212  0
Test accuracy =  91.92 %
              precision    recall  f1-score   support

           0       0.92      1.00      0.96     13788
           1       0.00      0.00      0.00      1212

    accuracy                           0.92     15000
   macro avg       0.46      0.50      0.48     15000
weighted avg       0.84      0.92      0.88     15000
```

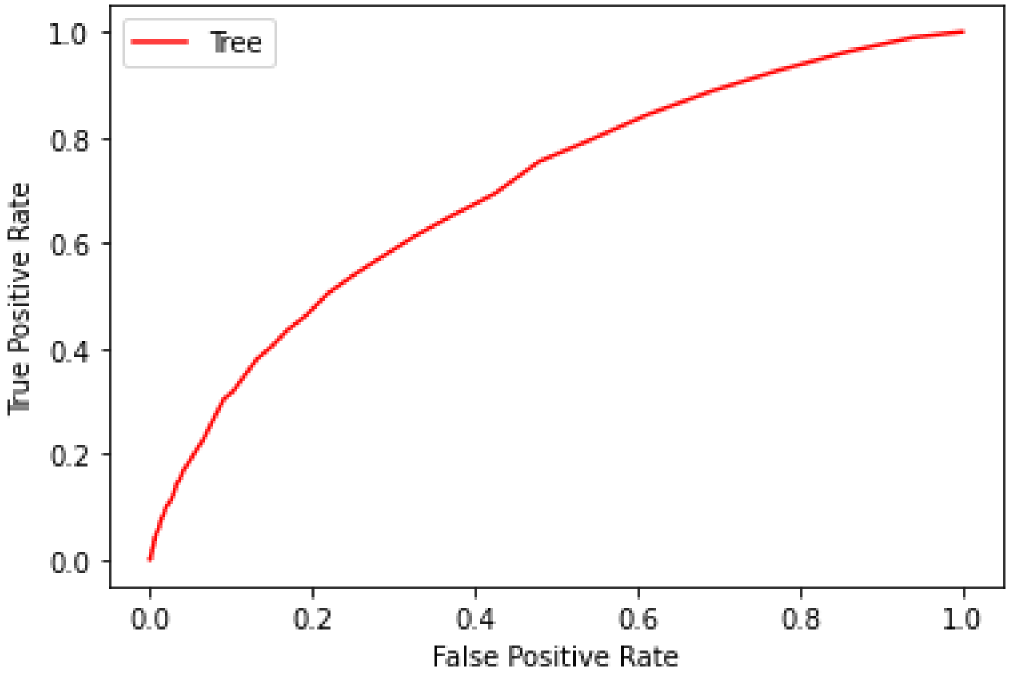
```
AUC Tree for Logistic:  0.7304789715263955
```

```
C:\ProgramData\Anaconda3\lib\site-
packages\sklearn\metrics\_classification.py:1221: UndefinedMetricWarning:
Precision and F-score are ill-defined and being set to 0.0 in labels with no
predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```



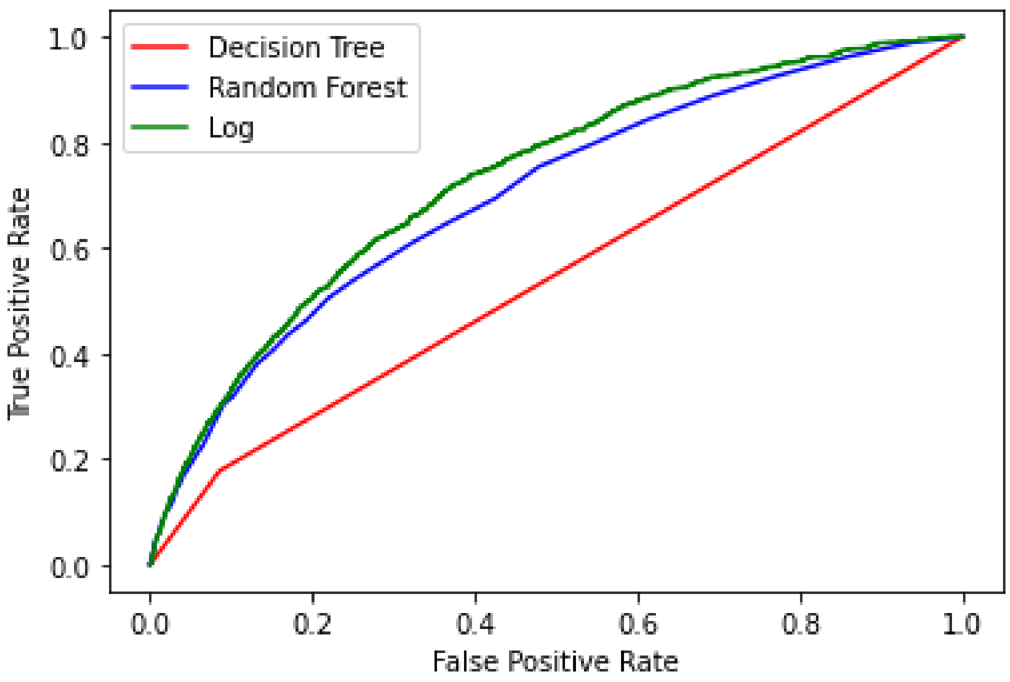
### 0.0.4 Random Forest Classifier

```
Confusion Matrix :
        0   1
0  13745  43
1   1185  27
Test accuracy =  91.81333333333333 %
              precision    recall  f1-score   support

           0       0.92      1.00      0.96     13788
           1       0.39      0.02      0.04      1212

    accuracy                           0.92     15000
   macro avg       0.65      0.51      0.50     15000
weighted avg       0.88      0.92      0.88     15000

AUC Tree for Random Forest:     0.7005428023220076
```



### 0.0.5 Decision tree

```
Confusion Matrix :
        0     1
0  12610  1178
1    997   215
Test accuracy =  85.5 %
              precision    recall  f1-score   support

           0       0.93      0.91      0.92     13788
           1       0.15      0.18      0.17      1212

    accuracy                           0.85     15000
   macro avg       0.54      0.55      0.54     15000
weighted avg       0.86      0.85      0.86     15000

AUC Tree for Decision Tree:     0.5459780638638276
```



### 0.0.6 SVM

```
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\utils\validation.py:67:
FutureWarning: Pass C=1.0 as keyword args. From version 0.25 passing these as
positional arguments will result in an error
  warnings.warn("Pass {} as keyword args. From version 0.25 "
```

```
Confusion Matrix :
    0      1
0   0   1212
1   0  13788
accuracy: 91.92
AUC Tree for SVM:       0.5812690113658886
```

### 0.0.7 Neural nets

```
---- Test data ----
Accuracy: 91.86
Classification Report:
              precision    recall  f1-score   support

           0       0.92      1.00      0.96     13788
           1       0.29      0.00      0.01      1212

    accuracy                           0.92     15000
```

```
   macro avg       0.60      0.50      0.48     15000
weighted avg       0.87      0.92      0.88     15000

Confusion Matrix:
[[13773    15]
 [ 1206     6]]
AUC Tree:       0.7544781131725009
```