

```
In [103... import numpy as np
import pandas as pd
import os
import warnings
warnings.filterwarnings('ignore')
import matplotlib.pyplot as plt
import seaborn as sns
```

This is the dataset

```
In [104... app_train = pd.read_csv('/Users/sami/Desktop/home-credit-default-risk/applic
print('Training data shape: ', app_train.shape)
app_train.head()
```

Training data shape: (307511, 122)

```
Out[104]:
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_
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	

5 rows × 122 columns

```
In [105... def missing_values_table(df):
    mis_val = df.isnull().sum()
    mis_val_percent = 100 * mis_val / len(df)
    mis_val_table = pd.concat([mis_val, mis_val_percent], axis=1)
    mis_val_table_ren_columns = mis_val_table.rename(
        columns = {0 : 'Missing Values', 1 : '% of Total Values'})
    mis_val_table_ren_columns = mis_val_table_ren_columns[
        mis_val_table_ren_columns.iloc[:,1] != 0].sort_values(
        '% of Total Values', ascending=False).round(1)
    print ("Your selected dataframe has " + str(df.shape[1]) + " columns.\n"
          "There are " + str(mis_val_table_ren_columns.shape[0]) + " columns th
    return mis_val_table_ren_columns
# function to check percentage of missing values in a column
```

```
In [106... missing_values = missing_values_table(app_train)
missing_values
```

Your selected dataframe has 122 columns.

There are 67 columns that have missing values.

Out[106]:

	Missing Values	% of Total Values
COMMONAREA_MEDI	214865	69.9
COMMONAREA_AVG	214865	69.9
COMMONAREA_MODE	214865	69.9
NONLIVINGAPARTMENTS_MEDI	213514	69.4
NONLIVINGAPARTMENTS_MODE	213514	69.4
...
EXT_SOURCE_2	660	0.2
AMT_GOODS_PRICE	278	0.1
AMT_ANNUITY	12	0.0
CNT_FAM_MEMBERS	2	0.0
DAYS_LAST_PHONE_CHANGE	1	0.0

67 rows × 2 columns

In [107]: `app_train.dtypes.value_counts()`

Out[107]:

```
float64    65
int64      41
object     16
dtype: int64
```

In [108]: `app_train.select_dtypes('object').apply(pd.Series.nunique, axis = 0)`

Out[108]:

```
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
```

```
In [109]: from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
le_count = 0
for col in app_train:
    if app_train[col].dtype == 'object':
        if len(list(app_train[col].unique())) <= 2:
            le.fit(app_train[col])
            app_train[col] = le.transform(app_train[col])
            le_count += 1
print('%d columns were label encoded.' % le_count)
# performing label encoding
```

3 columns were label encoded.

```
In [110... app_train = pd.get_dummies(app_train)
print('Training Features shape: ', app_train.shape)
# performing one hhot encoding on categorical data
```

Training Features shape: (307511, 243)

```
In [111... (app_train['DAYS_BIRTH'] / -365).describe()
```

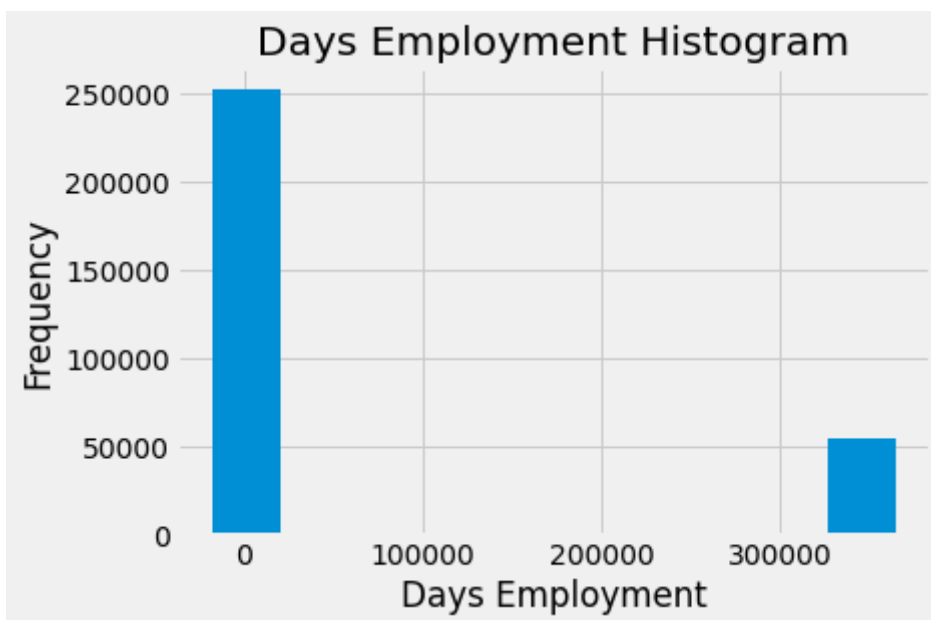
```
Out[111]: count    307511.000000
mean       43.936973
std        11.956133
min        20.517808
25%        34.008219
50%        43.150685
75%        53.923288
max        69.120548
Name: DAYS_BIRTH, dtype: float64
```

```
In [112... app_train['DAYS_EMPLOYED'].describe()
```

```
Out[112]: count    307511.000000
mean     63815.045904
std     141275.766519
min     -17912.000000
25%     -2760.000000
50%     -1213.000000
75%     -289.000000
max     365243.000000
Name: DAYS_EMPLOYED, dtype: float64
```

```
In [113... app_train['DAYS_EMPLOYED'].plot.hist(title = 'Days Employment Histogram');
plt.xlabel('Days Employment')
```

```
Out[113]: Text(0.5, 0, 'Days Employment')
```

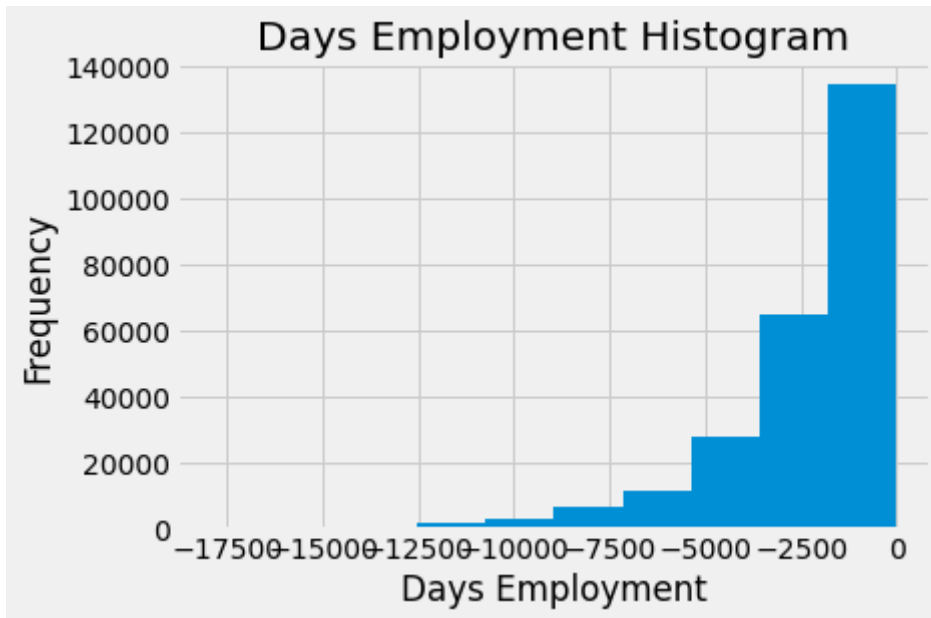


```
In [114... anom = app_train[app_train['DAYS_EMPLOYED'] == 365243]
non_anom = app_train[app_train['DAYS_EMPLOYED'] != 365243]
print('The non-anomalies default on %0.2f%% of loans' % (100 * non_anom['TARGET'].mean()))
print('The anomalies default on %0.2f%% of loans' % (100 * anom['TARGET'].mean()))
print('There are %d anomalous days of employment' % len(anom))
# calculating percentage of anomalies
```

The non-anomalies default on 8.66% of loans
 The anomalies default on 5.40% of loans
 There are 55374 anomalous days of employment

```
In [115]: app_train['DAYS_EMPLOYED_ANOM'] = app_train["DAYS_EMPLOYED"] == 365243
app_train['DAYS_EMPLOYED'].replace({365243: np.nan}, inplace = True)
app_train['DAYS_EMPLOYED'].plot.hist(title = 'Days Employment Histogram')
plt.xlabel('Days Employment')
```

Out[115]: Text(0.5, 0, 'Days Employment')



```
In [116]: correlations = app_train.corr()['TARGET'].sort_values()
print('Most Positive Correlations:\n', correlations.tail(15))
print('\nMost Negative Correlations:\n', correlations.head(15))
```

Most Positive Correlations:

OCCUPATION_TYPE_Laborers	0.043019
FLAG_DOCUMENT_3	0.044346
REG_CITY_NOT_LIVE_CITY	0.044395
FLAG_EMP_PHONE	0.045982
NAME_EDUCATION_TYPE_Secondary / secondary special	0.049824
REG_CITY_NOT_WORK_CITY	0.050994
DAYS_ID_PUBLISH	0.051457
CODE_GENDER_M	0.054713
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_EMPLOYED	0.074958
DAYS_BIRTH	0.078239
TARGET	1.000000

Name: TARGET, dtype: float64

Most Negative Correlations:

EXT_SOURCE_3	-0.178919
EXT_SOURCE_2	-0.160472
EXT_SOURCE_1	-0.155317
NAME_EDUCATION_TYPE_Higher education	-0.056593
CODE_GENDER_F	-0.054704
NAME_INCOME_TYPE_Pensioner	-0.046209
DAYS_EMPLOYED_ANOM	-0.045987
ORGANIZATION_TYPE_XNA	-0.045987
FLOORSMAX_AVG	-0.044003
FLOORSMAX_MEDI	-0.043768
FLOORSMAX_MODE	-0.043226
EMERGENCYSTATE_MODE_No	-0.042201
HOUSETYPE_MODE_block of flats	-0.040594
AMT_GOODS_PRICE	-0.039645
REGION_POPULATION_RELATIVE	-0.037227

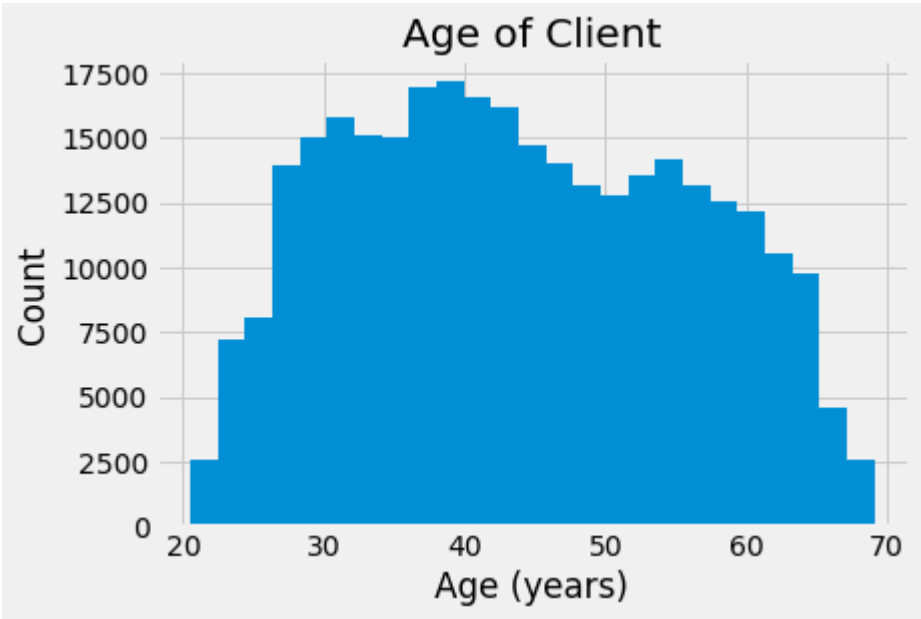
Name: TARGET, dtype: float64

```
In [117... app_train['DAYS_BIRTH'] = abs(app_train['DAYS_BIRTH'])
app_train['DAYS_BIRTH'].corr(app_train['TARGET'])
```

Out[117]: -0.07823930830982737

```
In [118... plt.style.use('fivethirtyeight')
plt.hist(app_train['DAYS_BIRTH'] / 365, bins = 25)
plt.title('Age of Client')
plt.xlabel('Age (years)')
plt.ylabel('Count')
```

Out[118]: Text(0, 0.5, 'Count')



LOGISTIC REGRESSION

```
In [119... from sklearn.linear_model import LogisticRegression
```

```
In [120... app_train
```

Out[120]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	FLAG_OWN_CAR	FLAG_OWN_REA
0	100002	1		0	0
1	100003	0		0	0
2	100004	0		1	1
3	100006	0		0	0
4	100007	0		0	0
...
307506	456251	0		0	0
307507	456252	0		0	0
307508	456253	0		0	0
307509	456254	1		0	0
307510	456255	0		0	0

307511 rows x 244 columns

```
In [121... app_train.dtypes
```

```
Out[121]: SK_ID_CURR          int64
          TARGET            int64
          NAME_CONTRACT_TYPE int64
          FLAG_OWN_CAR       int64
          FLAG_OWN_REALTY    int64
          ...
          WALLSMATERIAL_MODE_Stone, brick uint8
          WALLSMATERIAL_MODE_Wooden      uint8
          EMERGENCYSTATE_MODE_No         uint8
          EMERGENCYSTATE_MODE_Yes        uint8
          DAYS_EMPLOYED_ANOM             bool
          Length: 244, dtype: object
```

```
In [122... x = app_train.copy()
          y = x.TARGET
          x = x.drop(['TARGET'],axis=1)
```

```
In [123... from sklearn.impute import SimpleImputer
          from sklearn.preprocessing import MinMaxScaler
          imputer = SimpleImputer(missing_values=np.nan, strategy='median') # MEDIAN
          scaler = MinMaxScaler(feature_range = (0, 1))
          imputer.fit(x)
          x = imputer.transform(x)
          scaler.fit(x)
          x = scaler.transform(x)
```

```
In [124... from sklearn.model_selection import train_test_split
          x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.25,random_s
          from sklearn.linear_model import LogisticRegression
          lr = LogisticRegression(random_state=15)
          lr.fit(x_train,y_train)
          y_pred = lr.predict(x_test)
```

```
In [125... from sklearn import metrics
          cnf_matrix = metrics.confusion_matrix(y_test, y_pred)
          cnf_matrix
```

```
Out[125]: array([[70657,    84],
                [ 6074,    63]])
```

```
In [126... from sklearn.metrics import classification_report
          target_names = ['can_repay', 'cannot_repay']
          print(classification_report(y_test, y_pred, target_names=target_names))
```

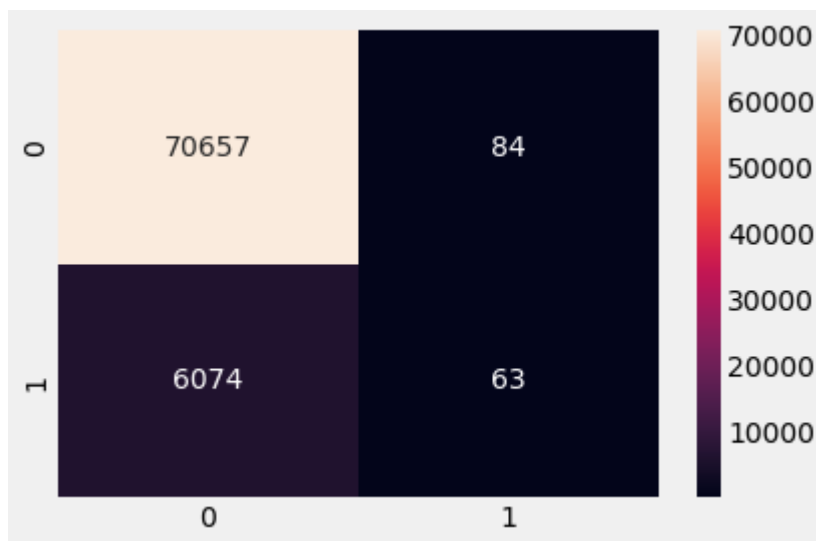
	precision	recall	f1-score	support
can_repay	0.92	1.00	0.96	70741
cannot_repay	0.43	0.01	0.02	6137
accuracy			0.92	76878
macro avg	0.67	0.50	0.49	76878
weighted avg	0.88	0.92	0.88	76878

```
In [127... acc_lr = metrics.accuracy_score(y_test,y_pred)
          acc_lr
```

```
Out[127]: 0.9198990608496579
```

```
In [128... import seaborn as sns
          sns.heatmap(cnf_matrix,annot = True,fmt='g')
```

```
Out[128]: <AxesSubplot: >
```



Decision Tree

```
In [69]: x = app_train.copy()
y = x.TARGET
x = x.drop(['TARGET'],axis=1)
feature_names = x.columns
labels = y.unique()
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import MinMaxScaler
imputer = SimpleImputer(missing_values=np.nan, strategy='median')
scaler = MinMaxScaler(feature_range = (0, 1))
imputer.fit(x)
x = imputer.transform(x)
scaler.fit(x)
x = scaler.transform(x)
```

```
In [70]: from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.25,random_s
```

```
In [71]: from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier(random_state=42)
classifier.fit(x_train,y_train)
```

```
Out[71]: DecisionTreeClassifier(random_state=42)
```

```
In [72]: from sklearn.tree import export_text
tree_rules = export_text(classifier,feature_names = list(feature_names))
print(tree_rules)
```



```

|--- EXT_SOURCE_3 <= 0.35
|   |--- EXT_SOURCE_2 <= 0.44
|       |--- EXT_SOURCE_3 <= 0.16
|           |--- EXT_SOURCE_2 <= 0.18
|               |--- AMT_CREDIT <= 0.03
|                   |--- EXT_SOURCE_2 <= 0.00
|                       |--- class: 1
|                           |--- EXT_SOURCE_2 > 0.00
|                               |--- OBS_60_CNT_SOCIAL_CIRCLE <= 0.01
|                                   |--- EXT_SOURCE_1 <= 0.72
|                                       |--- WEEKDAY_APPR_PROCESS_START_WEDNESDAY <=
0.50
|                                           |--- SK_ID_CURR <= 0.98
|                                               |--- HOUR_APPR_PROCESS_START <= 0.80
|                                                   |--- truncated branch of depth 6
|                                                       |--- HOUR_APPR_PROCESS_START > 0.80
|                                                           |--- class: 1
|                                                               |--- SK_ID_CURR > 0.98
|                                                                   |--- class: 1
|                                                                       |--- WEEKDAY_APPR_PROCESS_START_WEDNESDAY >
0.50
|                                                                           |--- DAYS_REGISTRATION <= 0.77
|                                                                               |--- class: 0
|                                                                                   |--- DAYS_REGISTRATION > 0.77
|                                                                                       |--- NAME_EDUCATION_TYPE_Secondary /
secondary special <= 0.50
|                                                                                           |--- class: 0
|                                                                                               |--- NAME_EDUCATION_TYPE_Secondary /
secondary special > 0.50
|                                                                                                   |--- truncated branch of depth 2
|                                                                                                       |--- EXT_SOURCE_1 > 0.72
|                                                                                                           |--- class: 1
|                                                                                                               |--- OBS_60_CNT_SOCIAL_CIRCLE > 0.01
|                                                                                                                   |--- SK_ID_CURR <= 0.86
|                                                                                                                       |--- NAME_FAMILY_STATUS_Civil marriage <= 0.
50
|                                                                                                       |--- YEARS_BEGINEXPLUATATION_MODE <= 0.9
8
|                                                                                                           |--- class: 0
|                                                                                                               |--- YEARS_BEGINEXPLUATATION_MODE > 0.9
8
|                                                                                                                   |--- class: 1
|                                                                                                                       |--- NAME_FAMILY_STATUS_Civil marriage > 0.
50
|                                                                                                       |--- class: 1
|                                                                                                           |--- SK_ID_CURR > 0.86
|                                                                                                               |--- class: 1
|                                                                                                                   |--- AMT_CREDIT > 0.03
|                                                                                                                       |--- AMT_GOODS_PRICE <= 0.10
|                                                                                                       |--- DAYS_ID_PUBLISH <= 0.92
|                                                                                                           |--- DAYS_REGISTRATION <= 0.98
|                                                                                                               |--- COMMONAREA_AVG <= 0.01
|                                                                                                                   |--- WEEKDAY_APPR_PROCESS_START_SATURDAY
<= 0.50
|                                                                                                   |--- EXT_SOURCE_1 <= 0.30
|                                                                                                       |--- truncated branch of depth 2
|                                                                                                           |--- EXT_SOURCE_1 > 0.30
|                                                                                                               |--- truncated branch of depth 3
|                                                                                                                   |--- WEEKDAY_APPR_PROCESS_START_SATURDAY
> 0.50
|                                                                                                   |--- YEARS_BUILD_MEDI <= 0.69
|                                                                                                       |--- class: 1
|                                                                                                           |--- YEARS_BUILD_MEDI > 0.69
|                                                                                                               |--- class: 0

```

8

[illegible]

```

|--- COMMONAREA_AVG > 0.01
|--- AMT_CREDIT <= 0.42
|--- ORGANIZATION_TYPE_Trade: type 2
<= 0.50
|--- truncated branch of depth 9
|--- ORGANIZATION_TYPE_Trade: type 2
> 0.50
|--- class: 1
|--- AMT_CREDIT > 0.42
|--- class: 1
|--- NAME_EDUCATION_TYPE_Higher education > 0.50
|--- DAYS_REGISTRATION <= 0.98
|--- NONLIVINGAPARTMENTS_MODE <= 0.02
|--- LIVINGAREA_MODE <= 0.00
|--- class: 1
|--- LIVINGAREA_MODE > 0.00
|--- SK_ID_CURR <= 0.77
|--- truncated branch of depth 1
4
|--- SK_ID_CURR > 0.77
|--- truncated branch of depth 4
|--- NONLIVINGAPARTMENTS_MODE > 0.02
|--- AMT_ANNUITY <= 0.18
|--- class: 1
|--- AMT_ANNUITY > 0.18
|--- class: 0
|--- DAYS_REGISTRATION > 0.98
|--- TOTALAREA_MODE <= 0.05
|--- class: 0
|--- TOTALAREA_MODE > 0.05
|--- DAYS_LAST_PHONE_CHANGE <= 0.87
|--- OCCUPATION_TYPE_Drivers <= 0.50
|--- truncated branch of depth 2
|--- OCCUPATION_TYPE_Drivers > 0.50
|--- class: 0
|--- DAYS_LAST_PHONE_CHANGE > 0.87
|--- OCCUPATION_TYPE_Laborers <= 0.5
0
|--- truncated branch of depth 3
0
|--- OCCUPATION_TYPE_Laborers > 0.5
|--- class: 1
|--- DAYS_BIRTH > 0.66
|--- DAYS_LAST_PHONE_CHANGE <= 0.70
|--- EXT_SOURCE_3 <= 0.16
|--- ORGANIZATION_TYPE_Transport: type 3 <=
0.50
|--- COMMONAREA_MODE <= 0.15
|--- AMT_GOODS_PRICE <= 0.27
|--- truncated branch of depth 3
|--- AMT_GOODS_PRICE > 0.27
|--- truncated branch of depth 2
|--- COMMONAREA_MODE > 0.15
|--- class: 1
|--- ORGANIZATION_TYPE_Transport: type 3 >
0.50
|--- class: 1
|--- EXT_SOURCE_3 > 0.16
|--- class: 1
|--- DAYS_LAST_PHONE_CHANGE > 0.70
|--- DAYS_LAST_PHONE_CHANGE <= 0.80
|--- SK_ID_CURR <= 0.46
|--- DAYS_LAST_PHONE_CHANGE <= 0.79
|--- class: 0

```

0.50

$$> 0.50$$

0.99

0.50

0

[illegible]

0.50

[illegible]

21/61

```

n <= 0.50
|--- FLOORSMAX_AVG <= 0.54
|   |--- truncated branch of depth 6
|       |--- FLOORSMAX_AVG > 0.54
|           |--- truncated branch of depth 2
|               --- OBS_60_CNT_SOCIAL_CIRCLE > 0.02
|                   |--- REGION_POPULATION_RELATIVE <= 0.06
|                       |--- class: 1
|                           --- REGION_POPULATION_RELATIVE > 0.06
|                               |--- YEARS_BEGINEXPLUATATION_MODE <= 0.98
|                                   |--- FLOORSMIN_AVG <= 0.06
|                                       |--- class: 1
|                                           --- FLOORSMIN_AVG > 0.06
|                                               |--- SK_ID_CURR <= 0.43
|                                                   |--- truncated branch of depth 6
|                                                       |--- SK_ID_CURR > 0.43
|                                                           |--- truncated branch of depth 7
|                                                               --- YEARS_BEGINEXPLUATATION_MODE > 0.98
|                                                                   |--- NAME_EDUCATION_TYPE_Higher education
n > 0.50
|--- SK_ID_CURR <= 0.56
|   |--- class: 1
|       |--- SK_ID_CURR > 0.56
|           |--- truncated branch of depth 2
|               |--- NAME_EDUCATION_TYPE_Higher education
n > 0.50
|--- ENTRANCES_MODE <= 0.26
|   |--- class: 0
|       |--- ENTRANCES_MODE > 0.26
|           |--- class: 1
|               --- NAME_CONTRACT_TYPE > 0.50
|                   |--- APARTMENTS_AVG <= 0.35
|                       |--- OCCUPATION_TYPE_Laborers <= 0.50
|                           |--- DAYS_ID_PUBLISH <= 0.93
|                               |--- FLOORSMIN_MEDI <= 0.02
|                                   |--- class: 1
|                                       |--- FLOORSMIN_MEDI > 0.02
|                                           |--- AMT_REQ_CREDIT_BUREAU_HOUR <=
0.12
|       |--- truncated branch of depth 9
|           |--- AMT_REQ_CREDIT_BUREAU_HOUR >
0.12
|       |--- class: 1
|           --- DAYS_ID_PUBLISH > 0.93
|               |--- DAYS_REGISTRATION <= 0.78
|                   |--- NONLIVINGAREA_MODE <= 0.00
|                       |--- class: 0
|                           |--- NONLIVINGAREA_MODE > 0.00
|                               |--- class: 1
|                                   --- DAYS_REGISTRATION > 0.78
|                                       |--- DAYS_BIRTH <= 0.69
|                                           |--- class: 0
|                                               |--- DAYS_BIRTH > 0.69
|                                                   |--- class: 1
|                                                       --- OCCUPATION_TYPE_Laborers > 0.50
|                                                           |--- EXT_SOURCE_3 <= 0.11
|                                                               |--- AMT_INCOME_TOTAL <= 0.00
|                                                                   |--- AMT_REQ_CREDIT_BUREAU_YEAR <=
0.14
|       |--- class: 1
|           |--- AMT_REQ_CREDIT_BUREAU_YEAR >
0.14
|       |--- class: 0
|           |--- AMT_INCOME_TOTAL > 0.00
|               |--- OWN_CAR_AGE <= 0.09

```

1

50


```

|--- LIVINGAREA_MODE <= 0.01
|--- OBS_30_CNT_SOCIAL_CIRCLE <= 0.0
1
|--- class: 1
|--- OBS_30_CNT_SOCIAL_CIRCLE > 0.0
1
|--- class: 0
|--- LIVINGAREA_MODE > 0.01
|--- HOUR_APPR_PROCESS_START <= 0.30
|--- class: 1
|--- HOUR_APPR_PROCESS_START > 0.30
|--- class: 0
|--- NAME_FAMILY_STATUS_Married > 0.50
|--- AMT_ANNUITY <= 0.07
|--- class: 1
|--- AMT_ANNUITY > 0.07
|--- OCCUPATION_TYPE_Core staff <=
0.50
|--- class: 0
|--- OCCUPATION_TYPE_Core staff >
0.50
|--- class: 1
|--- FLOORSMAX_MEDI > 0.10
|--- COMMONAREA_MODE <= 0.28
|--- EXT_SOURCE_1 <= 0.04
|--- class: 1
|--- EXT_SOURCE_1 > 0.04
|--- FLAG_DOCUMENT_6 <= 0.50
|--- truncated branch of depth 1
2
|--- FLAG_DOCUMENT_6 > 0.50
|--- class: 1
|--- COMMONAREA_MODE > 0.28
|--- class: 1
|--- DAYS_BIRTH > 0.13
|--- NAME_CONTRACT_TYPE <= 0.50
|--- COMMONAREA_AVG <= 0.00
|--- LANDAREA_MEDI <= 0.03
|--- class: 1
|--- LANDAREA_MEDI > 0.03
|--- DAYS_BIRTH <= 0.19
|--- class: 1
|--- DAYS_BIRTH > 0.19
|--- ORGANIZATION_TYPE_Industry: typ
e 1 <= 0.50
|--- class: 0
|--- ORGANIZATION_TYPE_Industry: typ
e 1 > 0.50
|--- class: 1
|--- COMMONAREA_AVG > 0.00
|--- DAYS_BIRTH <= 0.28
|--- DAYS_LAST_PHONE_CHANGE <= 0.89
|--- EXT_SOURCE_3 <= 0.16
|--- truncated branch of depth 2
|--- EXT_SOURCE_3 > 0.16
|--- truncated branch of depth 1
4
|--- DAYS_LAST_PHONE_CHANGE > 0.89
|--- OWN_CAR_AGE <= 0.15
|--- truncated branch of depth 1
1
|--- OWN_CAR_AGE > 0.15
|--- truncated branch of depth 4
|--- DAYS BIRTH > 0.28

```

```

ondary special <= 0.50
|      |      |      |      |
<= 0.50

```

[illegible]

```

|--- SK_ID_CURR <= 0.12
|   |--- BASEMENTAREA_MEDI <= 0.01
|       |--- class: 0
|   |--- BASEMENTAREA_MEDI > 0.01
|       |--- class: 1
|--- SK_ID_CURR > 0.12
|   |--- CNT_FAM_MEMBERS <= 0.08
|       |--- EXT_SOURCE_3 <= 0.34
|           |--- class: 0
|       |--- EXT_SOURCE_3 > 0.34
|           |--- class: 1
|   |--- CNT_FAM_MEMBERS > 0.08
|       |--- DAYS_ID_PUBLISH <= 0.63
|           |--- class: 1
|       |--- DAYS_ID_PUBLISH > 0.63
|           |--- class: 0
|--- LIVINGAREA_MODE > 0.01
|   |--- ORGANIZATION_TYPE_Realtor <= 0.50
|       |--- ORGANIZATION_TYPE_Transport: type 3
|           |--- DAYS_REGISTRATION <= 1.00
|               |--- truncated branch of depth 3
|           |--- DAYS_REGISTRATION > 1.00
|               |--- truncated branch of depth 2
|       |--- ORGANIZATION_TYPE_Transport: type 3
|           |--- NAME_INCOME_TYPE_Commercial associate <= 0.50
|               |--- class: 1
|           |--- NAME_INCOME_TYPE_Commercial associate > 0.50
|               |--- class: 0
|       |--- ORGANIZATION_TYPE_Realtor > 0.50
|           |--- YEARS_BUILD_MODE <= 0.69
|               |--- class: 1
|           |--- YEARS_BUILD_MODE > 0.69
|               |--- DAYS_LAST_PHONE_CHANGE <= 0.71
|                   |--- class: 1
|               |--- DAYS_LAST_PHONE_CHANGE > 0.71
|                   |--- class: 0
|--- EXT_SOURCE_2 > 0.76
|   |--- DAYS_EMPLOYED <= 0.94
|       |--- ENTRANCES_MEDI <= 0.84
|           |--- LIVINGAPARTMENTS_AVG <= 0.05
|               |--- OCCUPATION_TYPE_Waiters/barmen staff <= 0.50
|                   |--- APARTMENTS_AVG <= 0.24
|                       |--- OWN_CAR_AGE <= 0.14
|                           |--- truncated branch of depth 1
|                           |--- OWN_CAR_AGE > 0.14
|                               |--- truncated branch of depth 5
|                       |--- APARTMENTS_AVG > 0.24
|                           |--- class: 1
|                   |--- OCCUPATION_TYPE_Waiters/barmen staff > 0.50
|                       |--- class: 1
|       |--- LIVINGAPARTMENTS_AVG > 0.05
|           |--- YEARS_BUILD_MODE <= 0.01
|               |--- class: 1
|           |--- YEARS_BUILD_MODE > 0.01
|               |--- EXT_SOURCE_1 <= 0.55
|                   |--- ORGANIZATION TYPE Advertising < 0.50

```

```

= 0.50
1
|--- truncated branch of depth 3
0.50
|--- ORGANIZATION_TYPE_Advertising >
|--- truncated branch of depth 2
|--- EXT_SOURCE_1 > 0.55
|--- DAYS_BIRTH <= 0.99
|--- truncated branch of depth 2
3
|--- DAYS_BIRTH > 0.99
|--- class: 1
|--- ENTRANCES_MEDI > 0.84
|--- NONLIVINGAREA_MODE <= 0.00
|--- class: 0
|--- NONLIVINGAREA_MODE > 0.00
|--- class: 1
|--- DAYS_EMPLOYED > 0.94
|--- EXT_SOURCE_1 <= 0.64
|--- ELEVATORS_AVG <= 0.02
|--- DAYS_ID_PUBLISH <= 0.95
|--- DAYS_BIRTH <= 0.47
|--- SK_ID_CURR <= 0.01
|--- truncated branch of depth 3
|--- SK_ID_CURR > 0.01
|--- truncated branch of depth 2
1
|--- DAYS_BIRTH > 0.47
|--- ORGANIZATION_TYPE_Construction
<= 0.50
|--- truncated branch of depth 1
6
|--- ORGANIZATION_TYPE_Construction
> 0.50
|--- truncated branch of depth 3
|--- DAYS_ID_PUBLISH > 0.95
|--- WALLSMATERIAL_MODE_Stone, brick <=
0.50
|--- WEEKDAY_APPR_PROCESS_START_SATU
RDAY <= 0.50
|--- truncated branch of depth 5
|--- WEEKDAY_APPR_PROCESS_START_SATU
RDAY > 0.50
|--- class: 1
|--- WALLSMATERIAL_MODE_Stone, brick >
0.50
|--- DAYS_ID_PUBLISH <= 0.97
|--- class: 1
|--- DAYS_ID_PUBLISH > 0.97
|--- truncated branch of depth 2
|--- ELEVATORS_AVG > 0.02
|--- EXT_SOURCE_3 <= 0.17
|--- AMT_INCOME_TOTAL <= 0.00
|--- CNT_FAM_MEMBERS <= 0.13
|--- class: 1
|--- CNT_FAM_MEMBERS > 0.13
|--- class: 0
|--- AMT_INCOME_TOTAL > 0.00
|--- OCCUPATION_TYPE_Managers <= 0.5
0
|--- class: 0
|--- OCCUPATION_TYPE_Managers > 0.5
0
|--- class: 1

```

0.75

31/61

```

|--- AMT_CREDIT > 0.17
|--- class: 0
|--- LIVINGAREA_AVG > 0.05
|--- DAYS_REGISTRATION <= 0.95
|--- DEF_60_CNT_SOCIAL_CIRCLE <= 0.0
2
|--- truncated branch of depth 9
|--- DEF_60_CNT_SOCIAL_CIRCLE > 0.0
2
|--- truncated branch of depth 3
|--- DAYS_REGISTRATION > 0.95
|--- DAYS_LAST_PHONE_CHANGE <= 0.96
|--- truncated branch of depth 4
|--- DAYS_LAST_PHONE_CHANGE > 0.96
|--- truncated branch of depth 3
|--- EXT_SOURCE_2 > 0.03
|--- AMT_ANNUITY <= 0.05
|--- AMT_ANNUITY <= 0.03
|--- NAME_HOUSING_TYPE_House / apartment
<= 0.50
|--- OBS_60_CNT_SOCIAL_CIRCLE <= 0.0
0
|--- class: 1
|--- OBS_60_CNT_SOCIAL_CIRCLE > 0.0
0
|--- truncated branch of depth 3
|--- NAME_HOUSING_TYPE_House / apartment
> 0.50
|--- DAYS_REGISTRATION <= 0.84
|--- truncated branch of depth 3
|--- DAYS_REGISTRATION > 0.84
|--- truncated branch of depth 5
|--- AMT_ANNUITY > 0.03
|--- AMT_REQ_CREDIT_BUREAU_YEAR <= 0.26
|--- DAYS_ID_PUBLISH <= 0.88
|--- truncated branch of depth 7
|--- DAYS_ID_PUBLISH > 0.88
|--- truncated branch of depth 2
|--- AMT_REQ_CREDIT_BUREAU_YEAR > 0.26
|--- class: 1
|--- AMT_ANNUITY > 0.05
|--- DAYS_BIRTH <= 0.07
|--- DAYS_ID_PUBLISH <= 0.92
|--- NAME_TYPE_SUITE_Spouse, partner
<= 0.50
|--- truncated branch of depth 1
0
|--- NAME_TYPE_SUITE_Spouse, partner
> 0.50
|--- class: 1
|--- DAYS_ID_PUBLISH > 0.92
|--- AMT_CREDIT <= 0.06
|--- class: 1
|--- AMT_CREDIT > 0.06
|--- truncated branch of depth 4
|--- DAYS_BIRTH > 0.07
|--- DAYS_LAST_PHONE_CHANGE <= 0.97
|--- DAYS_ID_PUBLISH <= 0.40
|--- truncated branch of depth 1
0
|--- DAYS_ID_PUBLISH > 0.40
|--- truncated branch of depth 2
3
|--- DAYS_LAST_PHONE_CHANGE > 0.97

```


i

34/61

6

5

1

```
|--- truncated branch of depth 1  
7|  
|--- DAYS_BIRTH > 0.22  
|--- truncated branch of depth 2  
1|  
|--- DAYS_LAST_PHONE_CHANGE > 0.73  
|--- DAYS_BIRTH <= 0.20  
|--- EXT_SOURCE_1 <= 0.52  
|--- truncated branch of depth 2  
6|  
|--- EXT_SOURCE_1 > 0.52  
|--- truncated branch of depth 3  
6|  
|--- DAYS_BIRTH > 0.20  
|--- EXT_SOURCE_1 <= 0.14  
|--- truncated branch of depth 7  
|--- EXT_SOURCE_1 > 0.14  
|--- truncated branch of depth 3  
3|  
|--- REGION_RATING_CLIENT_W_CITY > 0.75  
|--- AMT_GOODS_PRICE <= 0.11  
|--- AMT_ANNUITY <= 0.06  
e 1 <= 0.50 |--- ORGANIZATION_TYPE_Industry: typ  
9| |--- truncated branch of depth 1  
| |--- ORGANIZATION_TYPE_Industry: typ  
e 1 > 0.50 |  
| |--- class: 1  
|--- AMT_ANNUITY > 0.06  
|--- SK_ID_CURR <= 0.05  
|--- truncated branch of depth 5  
|--- SK_ID_CURR > 0.05  
|--- truncated branch of depth 2  
0|  
|--- AMT_GOODS_PRICE > 0.11  
|--- LIVINGAREA_AVG <= 0.00  
ociate <= 0.50 |--- NAME_INCOME_TYPE_Commercial ass  
| |--- class: 1  
ociate > 0.50 |--- NAME_INCOME_TYPE_Commercial ass  
| |--- class: 0  
|--- LIVINGAREA_AVG > 0.00  
|--- DAYS_REGISTRATION <= 0.49  
|--- class: 1  
|--- DAYS_REGISTRATION > 0.49  
|--- truncated branch of depth 1  
3|  
|--- EXT_SOURCE_1 > 0.53  
|--- EXT_SOURCE_2 <= 0.27  
|--- EXT_SOURCE_2 <= 0.27  
|--- FLOORSMIN_MEDI <= 0.62  
|--- DAYS_REGISTRATION <= 0.94  
|--- OCCUPATION_TYPE_Medicine staff  
<= 0.50 |--- truncated branch of depth 1  
1|  
|--- OCCUPATION_TYPE_Medicine staff  
> 0.50 |  
|--- truncated branch of depth 2  
|--- DAYS_REGISTRATION > 0.94  
|--- OCCUPATION_TYPE_Drivers <= 0.50  
|--- truncated branch of depth 1
```

39/61

```
ociate > 0.50
```


[illegible]

[illegible]

8

[illegible]

[illegible]

```

n <= 0.50 |--- NAME_EDUCATION_TYPE_Higher education
|
|--- ORGANIZATION_TYPE_Legal Service
s <= 0.50
|
|--- truncated branch of depth 3
6
|
|--- ORGANIZATION_TYPE_Legal Service
s > 0.50
|
|--- class: 1
|--- NAME_EDUCATION_TYPE_Higher education
n > 0.50
|
|--- LIVINGAREA_MODE <= 0.00
|--- truncated branch of depth 3
|--- LIVINGAREA_MODE > 0.00
|--- truncated branch of depth 1
9
|
|--- DAYS_BIRTH > 0.53
|--- EXT_SOURCE_1 <= 0.52
|--- AMT_CREDIT <= 0.06
|--- truncated branch of depth 2
3
|
|--- AMT_CREDIT > 0.06
|--- truncated branch of depth 3
3
|
|--- EXT_SOURCE_1 > 0.52
|--- class: 1
|--- DAYS_EMPLOYED > 0.92
|--- NAME_EDUCATION_TYPE_Secondary / secondary speci
al <= 0.50
|
|--- EXT_SOURCE_1 <= 0.02
|--- class: 1
|--- EXT_SOURCE_1 > 0.02
|--- EXT_SOURCE_1 <= 0.12
|--- NAME_EDUCATION_TYPE_Incomplete high
er <= 0.50
|
|--- DAYS_LAST_PHONE_CHANGE <= 0.87
|--- truncated branch of depth 7
|--- DAYS_LAST_PHONE_CHANGE > 0.87
|--- truncated branch of depth 3
|--- NAME_EDUCATION_TYPE_Incomplete high
er > 0.50
|
|--- NONLIVINGAREA_AVG <= 0.01
|--- truncated branch of depth 5
|--- NONLIVINGAREA_AVG > 0.01
|--- class: 0
|--- EXT_SOURCE_1 > 0.12
|--- DAYS_BIRTH <= 0.23
|--- NONLIVINGAPARTMENTS_MEDI <= 0.4
8
|
|--- truncated branch of depth 3
1
|
|--- NONLIVINGAPARTMENTS_MEDI > 0.4
8
|
|--- class: 1
|--- DAYS_BIRTH > 0.23
|--- EXT_SOURCE_2 <= 0.61
|--- truncated branch of depth 2
4
|
|--- EXT_SOURCE_2 > 0.61
|--- truncated branch of depth 2
2
|
|--- NAME_EDUCATION_TYPE_Secondary / secondary speci
al > 0.50
|
|--- AMT CREDIT <= 0.03

```

```
ondary special > 0.50
```

```
e 2 <= 0.50  
|  
6 |--- truncated branch of depth 3  
  
e 2 > 0.50  
|--- ORGANIZATION_TYPE_Industry: typ  
|  
|--- truncated branch of depth 2  
|-- LIVINGAPARTMENTS_MEDI > 0.77  
|   |-- LIVINGAREA_MEDI <= 0.06  
|       |-- class: 1  
|           |-- LIVINGAREA_MEDI > 0.06  
|               |-- class: 0  
|-- OWN_CAR_AGE > 0.18  
|   |-- EXT_SOURCE_3 <= 0.36  
|       |-- ORGANIZATION_TYPE_Business Entity T  
ype 3 <= 0.50  
SDAY <= 0.50  
|  
|--- WEEKDAY_APPR_PROCESS_START_THUR  
SDAY > 0.50  
|  
|--- class: 1  
|--- WEEKDAY_APPR_PROCESS_START_THUR  
SDAY > 0.50  
|  
|--- class: 0  
|--- ORGANIZATION_TYPE_Business Entity T  
ype 3 > 0.50  
|  
|--- class: 0  
|-- EXT_SOURCE_3 > 0.36  
|   |-- EXT_SOURCE_2 <= 0.46  
|       |-- FLAG_EMAIL <= 0.50  
|           |-- class: 1  
|               |-- FLAG_EMAIL > 0.50  
|                   |-- class: 0  
|-- EXT_SOURCE_2 > 0.46  
|   |-- ORGANIZATION_TYPE_Industry: typ  
e 1 <= 0.50  
|  
|--- truncated branch of depth 1  
7 |  
|--- ORGANIZATION_TYPE_Industry: typ  
e 1 > 0.50  
|  
|--- class: 1  
|-- DAYS_ID_PUBLISH > 0.92  
|   |-- DAYS_ID_PUBLISH <= 0.92  
|       |-- NAME_FAMILY_STATUS_Married <= 0.50  
|           |-- class: 0  
|               |-- NAME_FAMILY_STATUS_Married > 0.50  
|                   |-- class: 1  
|-- DAYS_ID_PUBLISH > 0.92  
|   |-- AMT_REQ_CREDIT_BUREAU_QRT <= 0.01  
|       |-- BASEMENTAREA_MEDI <= 0.77  
|           |-- OBS_60_CNT_SOCIAL_CIRCLE <= 0.0  
6 |  
|--- truncated branch of depth 1  
9 |  
|--- OBS_60_CNT_SOCIAL_CIRCLE > 0.0  
6 |  
|  
|--- class: 1  
|-- BASEMENTAREA_MEDI > 0.77  
|   |-- class: 1  
|-- AMT_REQ_CREDIT_BUREAU_QRT > 0.01  
|   |-- class: 1  
-- LIVINGAREA_MEDI > 0.99  
|-- REGION_POPULATION_RELATIVE <= 0.26  
|   |-- class: 0  
-- REGION_POPULATION_RELATIVE > 0.26  
    |-- class: 1
```


6

1

51/61

!

≤ 0.50

[illegible]

```
| | | | | | | | | | | --- truncated branch of depth 2
```

[illegible]

```
In [73]: y_pred_dt = classifier.predict(x_test)
```

```
In [74]: cnfmx_dt = metrics.confusion_matrix(y_test,y_pred_dt)
cnfmx_dt
```

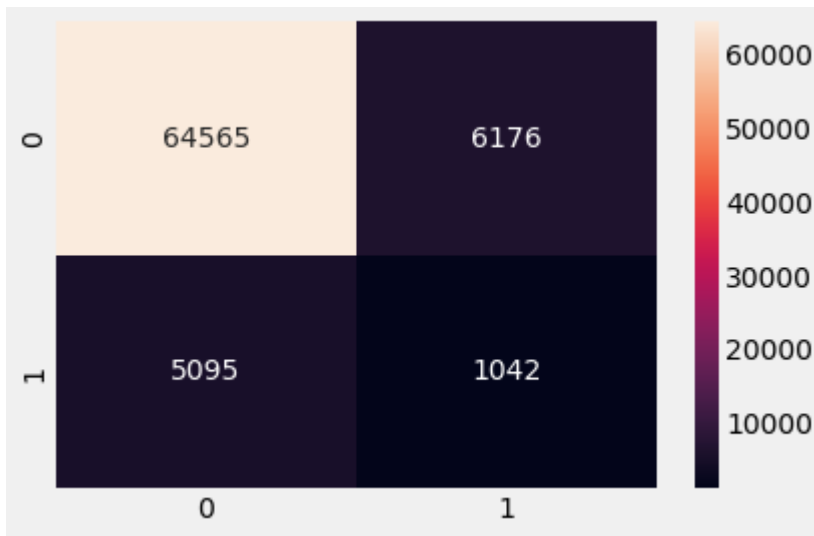
```
Out[74]: array([[64565, 6176],
                [ 5095, 1042]])
```

```
In [75]: acc_dt = metrics.accuracy_score(y_test,y_pred_dt)
acc_dt
```

```
Out[75]: 0.8533910871770858
```

```
In [76]: sns.heatmap(cnfmx_dt,annot = True, fmt = 'g' )
```

```
Out[76]: <AxesSubplot: >
```

```
In [78]: from sklearn.metrics import classification_report
target_names = ['can_repay', 'cannot_repay']
print(classification_report(y_test, y_pred_dt, target_names=target_names))
```

	precision	recall	f1-score	support
can_repay	0.93	0.91	0.92	70741
cannot_repay	0.14	0.17	0.16	6137
accuracy			0.85	76878
macro avg	0.54	0.54	0.54	76878
weighted avg	0.86	0.85	0.86	76878

Random Forest

```
In [57]: pip install graphviz
```

```
Collecting graphviz
  Downloading graphviz-0.20.1-py3-none-any.whl (47 kB)
    |████████████████████| 47 kB 1.7 MB/s eta 0:00:01
Installing collected packages: graphviz
Successfully installed graphviz-0.20.1
Note: you may need to restart the kernel to use updated packages.
```

```
In [61]: import graphviz
```

```
In [43]: x = app_train.copy()
y = x.TARGET
x = x.drop(['TARGET'],axis=1)
feature_names = x.columns
labels = y.unique()
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import MinMaxScaler
imputer = SimpleImputer(missing_values=np.nan, strategy='median')
scaler = MinMaxScaler(feature_range = (0, 1))
imputer.fit(x)
x = imputer.transform(x)
scaler.fit(x)
x = scaler.transform(x)
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.25,random_s
```

```
In [44]: from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier()
rf.fit(x_train,y_train)
```

```
Out[44]: RandomForestClassifier()
```

```
In [46]: y_pred_rf = rf.predict(x_test)
```

```
In [49]: acc_rf = metrics.accuracy_score(y_test,y_pred_rf)
acc_rf
```

```
Out[49]: 0.9201462056765265
```

```
In [79]: from sklearn.metrics import classification_report
target_names = ['can_repay', 'cannot_repay']
print(classification_report(y_test, y_pred_rf, target_names=target_names))
```

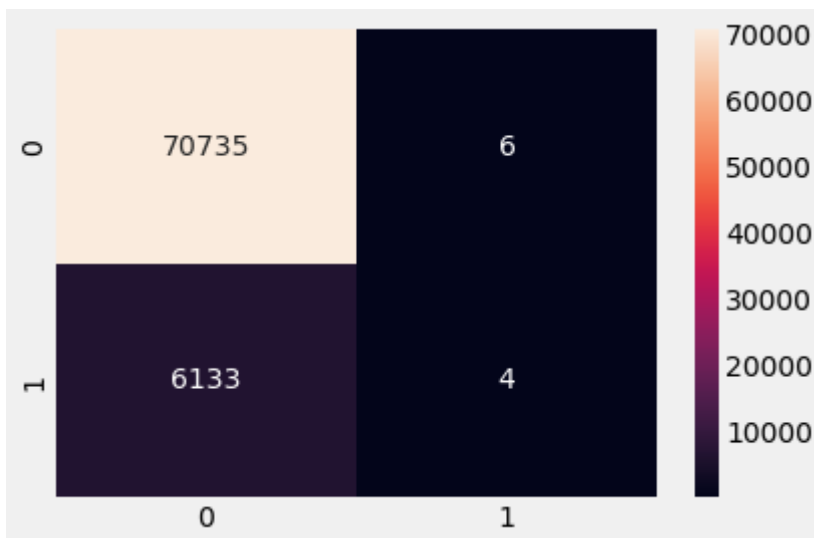
	precision	recall	f1-score	support
can_repay	0.92	1.00	0.96	70741
cannot_repay	0.40	0.00	0.00	6137
accuracy			0.92	76878
macro avg	0.66	0.50	0.48	76878
weighted avg	0.88	0.92	0.88	76878

```
In [80]: cnfmx_rf = metrics.confusion_matrix(y_test,y_pred_rf)
cnfmx_rf
```

```
Out[80]: array([[70735, 6],
[ 6133, 4]])
```

```
In [81]: sns.heatmap(cnfmx_rf,annot = True, fmt = 'g' )
```

```
Out[81]: <AxesSubplot: >
```



k-nearest neighbours

```
In [ ]: #knn,naive bayes,comments
```

```
In [82]: x = app_train.copy()
y = x.TARGET
```

```

x = x.drop(['TARGET'],axis=1)
feature_names = x.columns
labels = y.unique()
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import MinMaxScaler
imputer = SimpleImputer(missing_values=np.nan, strategy='median')
scaler = MinMaxScaler(feature_range = (0, 1))
imputer.fit(x)
x = imputer.transform(x)
scaler.fit(x)
x = scaler.transform(x)
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.25,random_s

```

```

In [83]: from sklearn.neighbors import KNeighborsClassifier
knn5 = KNeighborsClassifier(n_neighbors = 5)
knn1 = KNeighborsClassifier(n_neighbors=1)

```

```

In [86]: knn5.fit(x_train, y_train)
knn1.fit(x_train, y_train)
y_pred_5 = knn5.predict(x_test)
y_pred_1 = knn1.predict(x_test)

```

```

In [87]: from sklearn.metrics import accuracy_score
print("Accuracy with k=5", accuracy_score(y_test, y_pred_5))
print("Accuracy with k=1", accuracy_score(y_test, y_pred_1))

```

```

Accuracy with k=5 0.9156715835479591
Accuracy with k=1 0.858594136163792

```

```

In [88]: knn7 = KNeighborsClassifier(n_neighbors = 7)
knn7.fit(x_train, y_train)
y_pred_7 = knn5.predict(x_test)

```

```

In [89]: print("Accuracy with k=7", accuracy_score(y_test, y_pred_7))

```

```

Accuracy with k=7 0.9156715835479591

```

```

In [90]: from sklearn.metrics import classification_report
target_names = ['can_repay', 'cannot_repay']
print(classification_report(y_test, y_pred_5, target_names=target_names))

```

	precision	recall	f1-score	support
can_repay	0.92	0.99	0.96	70741
cannot_repay	0.19	0.02	0.03	6137
accuracy			0.92	76878
macro avg	0.56	0.51	0.49	76878
weighted avg	0.86	0.92	0.88	76878

```

In [98]: cnfmx_knn = metrics.confusion_matrix(y_test,y_pred_5)
cnfmx_knn

```

```

Out[98]: array([[ 70287,    454],
               [  6029,   108]])

```

```

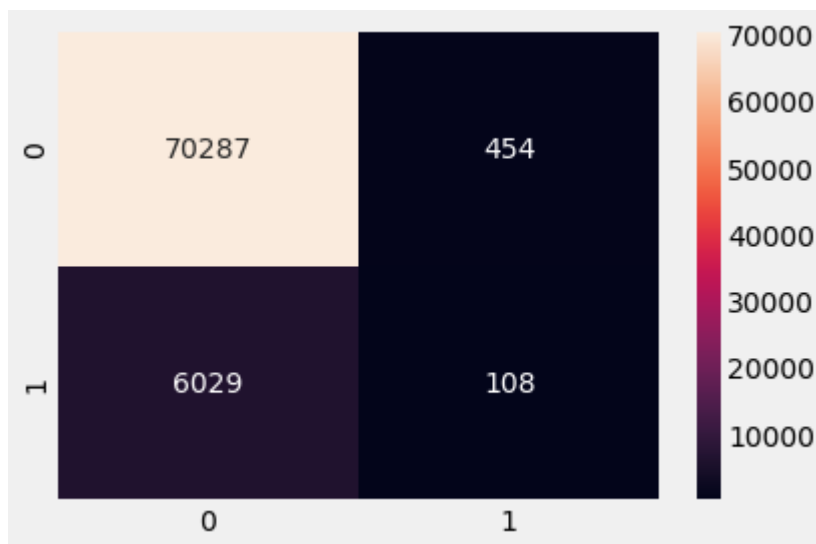
In [99]: sns.heatmap(cnfmx_knn,annot = True, fmt = 'g' )

```

```

Out[99]: <AxesSubplot: >

```



Naive bayes

```
In [93]: x = app_train.copy()
y = x.TARGET
x = x.drop(['TARGET'],axis=1)
feature_names = x.columns
labels = y.unique()
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import MinMaxScaler
imputer = SimpleImputer(missing_values=np.nan, strategy='median')
scaler = MinMaxScaler(feature_range = (0, 1))
imputer.fit(x)
x = imputer.transform(x)
scaler.fit(x)
x = scaler.transform(x)
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.25,random_s
```

```
In [94]: from sklearn.naive_bayes import GaussianNB
nb = GaussianNB()
nb.fit(x_train, y_train)
y_pred_nb = nb.predict(x_test)
```

```
In [95]: from sklearn.metrics import accuracy_score
print(accuracy_score(y_test, y_pred_nb))
```

0.18290017950519005

```
In [97]: from sklearn.metrics import classification_report
target_names = ['can_repay', 'cannot_repay']
print(classification_report(y_test, y_pred_nb, target_names=target_names))
```

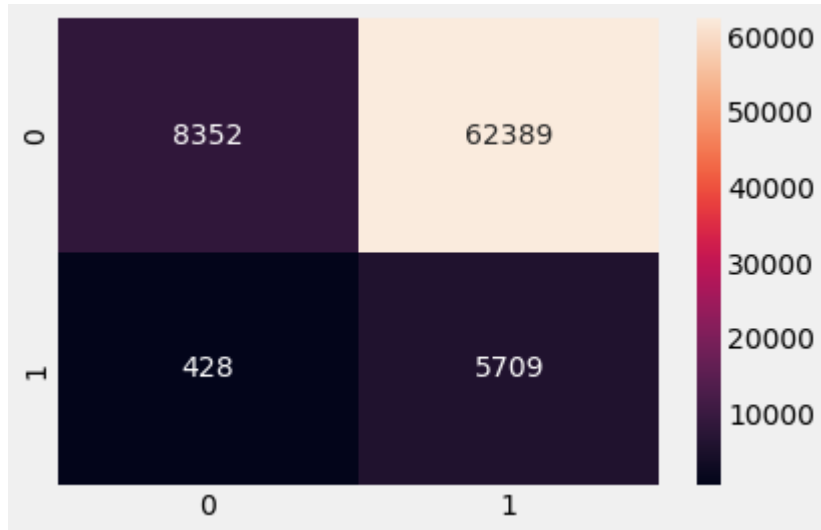
	precision	recall	f1-score	support
can_repay	0.95	0.12	0.21	70741
cannot_repay	0.08	0.93	0.15	6137
accuracy			0.18	76878
macro avg	0.52	0.52	0.18	76878
weighted avg	0.88	0.18	0.21	76878

```
In [100... cnfmx_nb = metrics.confusion_matrix(y_test,y_pred_nb)
cnfmx_nb
```

```
Out[100]: array([[ 8352, 62389],  
               [  428,  5709]])
```

```
In [101]: sns.heatmap(cnfm_x_nb,annot = True, fmt = 'g' )
```

```
Out[101]: <AxesSubplot: >
```



```
In [ ]:
```

```
In [ ]:
```