Home_credit_default_risk

June 24, 2019

1 Home Credit Default Risk

Predict how capable each applicant is of repaying a loan.
Photo Breno Assis

1.1 Context

This challenge was proposed by **Home Credit Group**.

Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders.

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities.

While Home Credit is currently using various statistical and machine learning methods to make these predictions, they're challenging Kagglers to help them unlock the full potential of their data. Doing so will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.

1.2 Goal

Use historical loan application data to **predict whether or not an applicant will be able to repay a loan**. This is a standard **supervised classification task**.

Submissions are evaluated on **area under the ROC curve** between the predicted probability and the observed target.

1.3 Type of ML

- Supervised: Labels are included in the training data
- Binary classification: target has only two values 0 (will repay loan on time), 1 (will have difficulty repaying loan)

1.4 Guidelines

- Download and load the data
- Sample the data in order to work on a smaller subset at first
- Explore the data, creating functions for cleaning it

- Split your data into features & labels ; training & testing
- Train different models and compare performance
- Train on your entire dataset, run predictions on your entire dataset and submit your results!
- Iterate

---

2 Data

2.1 First insight

In [1]:
```
# usual data science stack in python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

import os
print(os.listdir("../input/"))

# Any results you write to the current directory are saved as output.
```

```
['application_train.csv', 'home_credit.png', 'ROC-curve.png', 'HomeCredit_columns_description_
```

In [2]:
```
# imports of need modules in sklearn
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_auc_score
from sklearn.model_selection import cross_val_score

from sklearn.ensemble import RandomForestClassifier
```

In [3]:
```
import lightgbm as lgb
import xgboost as xgb
```

In [4]:
```
# set options in this notebook
pd.set_option('display.max_columns', 300)

import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)
```

In [5]:
```
path_train = os.path.join('..', 'input', 'application_train.csv')
path_test = os.path.join('..', 'input', 'application_test.csv')
```

In [6]:
```
# load main datasets
app_train, app_test = pd.read_csv(path_train), pd.read_csv(path_test)
```

In [7]:
```
# 1st insight
app_train.tail()
```

```
Out[7]:         SK_ID_CURR  TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR  \
        307506      456251       0         Cash loans           M            N
        307507      456252       0         Cash loans           F            N
        307508      456253       0         Cash loans           F            N
        307509      456254       1         Cash loans           F            N
        307510      456255       0         Cash loans           F            N

               FLAG_OWN_REALTY  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  \
        307506               N             0          157500.0    254700.0
        307507               Y             0           72000.0    269550.0
        307508               Y             0          153000.0    677664.0
        307509               Y             0          171000.0    370107.0
        307510               N             0          157500.0    675000.0

                AMT_ANNUITY  AMT_GOODS_PRICE NAME_TYPE_SUITE      NAME_INCOME_TYPE  \
        307506      27558.0         225000.0   Unaccompanied               Working
        307507      12001.5         225000.0   Unaccompanied             Pensioner
        307508      29979.0         585000.0   Unaccompanied               Working
        307509      20205.0         319500.0   Unaccompanied  Commercial associate
        307510      49117.5         675000.0   Unaccompanied  Commercial associate

                          NAME_EDUCATION_TYPE NAME_FAMILY_STATUS  NAME_HOUSING_TYPE  \
        307506  Secondary / secondary special          Separated       With parents
        307507  Secondary / secondary special              Widow  House / apartment
        307508               Higher education          Separated  House / apartment
        307509  Secondary / secondary special            Married  House / apartment
        307510               Higher education            Married  House / apartment

                REGION_POPULATION_RELATIVE  DAYS_BIRTH  DAYS_EMPLOYED  \
        307506                    0.032561       -9327           -236
        307507                    0.025164      -20775         365243
        307508                    0.005002      -14966          -7921
        307509                    0.005313      -11961          -4786
        307510                    0.046220      -16856          -1262

                DAYS_REGISTRATION  DAYS_ID_PUBLISH  OWN_CAR_AGE  FLAG_MOBIL  \
        307506            -8456.0            -1982          NaN           1
        307507            -4388.0            -4090          NaN           1
        307508            -6737.0            -5150          NaN           1
        307509            -2562.0             -931          NaN           1
        307510            -5128.0             -410          NaN           1

                FLAG_EMP_PHONE  FLAG_WORK_PHONE  FLAG_CONT_MOBILE  FLAG_PHONE  \
        307506               1                0                 1           0
        307507               0                0                 1           1
        307508               1                0                 1           0
        307509               1                0                 1           0
        307510               1                1                 1           1
```

```
       FLAG_EMAIL OCCUPATION_TYPE  CNT_FAM_MEMBERS  REGION_RATING_CLIENT  \
307506          0     Sales staff              1.0                     1
307507          0             NaN              1.0                     2
307508          1        Managers              1.0                     3
307509          0        Laborers              2.0                     2
307510          0        Laborers              2.0                     1

       REGION_RATING_CLIENT_W_CITY WEEKDAY_APPR_PROCESS_START  \
307506                           1                   THURSDAY
307507                           2                     MONDAY
307508                           3                   THURSDAY
307509                           2                  WEDNESDAY
307510                           1                   THURSDAY

       HOUR_APPR_PROCESS_START  REG_REGION_NOT_LIVE_REGION  \
307506                      15                           0
307507                       8                           0
307508                       9                           0
307509                       9                           0
307510                      20                           0

       REG_REGION_NOT_WORK_REGION  LIVE_REGION_NOT_WORK_REGION  \
307506                          0                            0
307507                          0                            0
307508                          0                            0
307509                          0                            0
307510                          0                            0

       REG_CITY_NOT_LIVE_CITY  REG_CITY_NOT_WORK_CITY  \
307506                      0                       0
307507                      0                       0
307508                      0                       1
307509                      1                       1
307510                      0                       1

       LIVE_CITY_NOT_WORK_CITY       ORGANIZATION_TYPE  EXT_SOURCE_1  \
307506                       0                Services      0.145570
307507                       0                     XNA           NaN
307508                       1                  School      0.744026
307509                       0  Business Entity Type 1           NaN
307510                       1  Business Entity Type 3      0.734460

       EXT_SOURCE_2  EXT_SOURCE_3  APARTMENTS_AVG  BASEMENTAREA_AVG  \
307506     0.681632           NaN          0.2021            0.0887
307507     0.115992           NaN          0.0247            0.0435
307508     0.535722      0.218859          0.1031            0.0862
307509     0.514163      0.661024          0.0124               NaN
```

```
307510      0.708569      0.113922          0.0742            0.0526

        YEARS_BEGINEXPLUATATION_AVG  YEARS_BUILD_AVG  COMMONAREA_AVG  \
307506                       0.9876           0.8300          0.0202
307507                       0.9727           0.6260          0.0022
307508                       0.9816           0.7484          0.0123
307509                       0.9771              NaN             NaN
307510                       0.9881              NaN          0.0176

        ELEVATORS_AVG  ENTRANCES_AVG  FLOORSMAX_AVG  FLOORSMIN_AVG  \
307506           0.22         0.1034         0.6042         0.2708
307507           0.00         0.1034         0.0833         0.1250
307508           0.00         0.2069         0.1667         0.2083
307509            NaN         0.0690         0.0417            NaN
307510           0.08         0.0690         0.3750            NaN

        LANDAREA_AVG  LIVINGAPARTMENTS_AVG  LIVINGAREA_AVG  \
307506        0.0594                0.1484          0.1965
307507        0.0579                0.0202          0.0257
307508           NaN                0.0841          0.9279
307509           NaN                   NaN          0.0061
307510           NaN                   NaN          0.0791

        NONLIVINGAPARTMENTS_AVG  NONLIVINGAREA_AVG  APARTMENTS_MODE  \
307506                   0.0753             0.1095           0.1008
307507                   0.0000             0.0000           0.0252
307508                   0.0000             0.0000           0.1050
307509                      NaN                NaN           0.0126
307510                      NaN             0.0000           0.0756

        BASEMENTAREA_MODE  YEARS_BEGINEXPLUATATION_MODE  YEARS_BUILD_MODE  \
307506             0.0172                        0.9782            0.7125
307507             0.0451                        0.9727            0.6406
307508             0.0894                        0.9816            0.7583
307509                NaN                        0.9772               NaN
307510             0.0546                        0.9881               NaN

        COMMONAREA_MODE  ELEVATORS_MODE  ENTRANCES_MODE  FLOORSMAX_MODE  \
307506           0.0172          0.0806          0.0345          0.4583
307507           0.0022          0.0000          0.1034          0.0833
307508           0.0124          0.0000          0.2069          0.1667
307509              NaN             NaN          0.0690          0.0417
307510           0.0178          0.0806          0.0690          0.3750

        FLOORSMIN_MODE  LANDAREA_MODE  LIVINGAPARTMENTS_MODE  LIVINGAREA_MODE  \
307506          0.0417         0.0094                 0.0882           0.0853
307507          0.1250         0.0592                 0.0220           0.0267
307508          0.2083            NaN                 0.0918           0.9667
```

```
307509           NaN            NaN                       NaN          0.0063
307510           NaN            NaN                       NaN          0.0824

        NONLIVINGAPARTMENTS_MODE  NONLIVINGAREA_MODE  APARTMENTS_MEDI  \
307506                       0.0              0.0125           0.2040
307507                       0.0              0.0000           0.0250
307508                       0.0              0.0000           0.1041
307509                       NaN                 NaN           0.0125
307510                       NaN              0.0000           0.0749

        BASEMENTAREA_MEDI  YEARS_BEGINEXPLUATATION_MEDI  YEARS_BUILD_MEDI  \
307506             0.0887                        0.9876            0.8323
307507             0.0435                        0.9727            0.6310
307508             0.0862                        0.9816            0.7518
307509                NaN                        0.9771               NaN
307510             0.0526                        0.9881               NaN

        COMMONAREA_MEDI  ELEVATORS_MEDI  ENTRANCES_MEDI  FLOORSMAX_MEDI  \
307506           0.0203            0.22          0.1034          0.6042
307507           0.0022            0.00          0.1034          0.0833
307508           0.0124            0.00          0.2069          0.1667
307509              NaN             NaN          0.0690          0.0417
307510           0.0177            0.08          0.0690          0.3750

        FLOORSMIN_MEDI  LANDAREA_MEDI  LIVINGAPARTMENTS_MEDI  LIVINGAREA_MEDI  \
307506          0.2708         0.0605                 0.1509           0.2001
307507          0.1250         0.0589                 0.0205           0.0261
307508          0.2083            NaN                 0.0855           0.9445
307509             NaN            NaN                    NaN           0.0062
307510             NaN            NaN                    NaN           0.0805

        NONLIVINGAPARTMENTS_MEDI  NONLIVINGAREA_MEDI FONDKAPREMONT_MODE  \
307506                    0.0757              0.1118   reg oper account
307507                    0.0000              0.0000   reg oper account
307508                    0.0000              0.0000   reg oper account
307509                       NaN                 NaN                NaN
307510                       NaN              0.0000                NaN

        HOUSETYPE_MODE  TOTALAREA_MODE WALLSMATERIAL_MODE EMERGENCYSTATE_MODE  \
307506  block of flats          0.2898       Stone, brick                  No
307507  block of flats          0.0214       Stone, brick                  No
307508  block of flats          0.7970              Panel                  No
307509  block of flats          0.0086       Stone, brick                  No
307510  block of flats          0.0718              Panel                  No

        OBS_30_CNT_SOCIAL_CIRCLE  DEF_30_CNT_SOCIAL_CIRCLE  \
307506                       0.0                       0.0
307507                       0.0                       0.0
```

```
307508                     6.0                        0.0
307509                     0.0                        0.0
307510                     0.0                        0.0

        OBS_60_CNT_SOCIAL_CIRCLE  DEF_60_CNT_SOCIAL_CIRCLE  \
307506                     0.0                        0.0
307507                     0.0                        0.0
307508                     6.0                        0.0
307509                     0.0                        0.0
307510                     0.0                        0.0

        DAYS_LAST_PHONE_CHANGE  FLAG_DOCUMENT_2  FLAG_DOCUMENT_3  \
307506                  -273.0                0                0
307507                     0.0                0                1
307508                 -1909.0                0                1
307509                  -322.0                0                1
307510                  -787.0                0                1

        FLAG_DOCUMENT_4  FLAG_DOCUMENT_5  FLAG_DOCUMENT_6  FLAG_DOCUMENT_7  \
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

        FLAG_DOCUMENT_8  FLAG_DOCUMENT_9  FLAG_DOCUMENT_10  FLAG_DOCUMENT_11  \
307506                1                0                 0                 0
307507                0                0                 0                 0
307508                0                0                 0                 0
307509                0                0                 0                 0
307510                0                0                 0                 0

        FLAG_DOCUMENT_12  FLAG_DOCUMENT_13  FLAG_DOCUMENT_14  \
307506                 0                 0                 0
307507                 0                 0                 0
307508                 0                 0                 0
307509                 0                 0                 0
307510                 0                 0                 0

        FLAG_DOCUMENT_15  FLAG_DOCUMENT_16  FLAG_DOCUMENT_17  \
307506                 0                 0                 0
307507                 0                 0                 0
307508                 0                 0                 0
307509                 0                 0                 0
307510                 0                 0                 0

        FLAG_DOCUMENT_18  FLAG_DOCUMENT_19  FLAG_DOCUMENT_20  \
307506                 0                 0                 0
```

```
307507                  0                  0                  0
307508                  0                  0                  0
307509                  0                  0                  0
307510                  0                  0                  0

        FLAG_DOCUMENT_21  AMT_REQ_CREDIT_BUREAU_HOUR  \
307506                 0                         NaN
307507                 0                         NaN
307508                 0                         1.0
307509                 0                         0.0
307510                 0                         0.0

        AMT_REQ_CREDIT_BUREAU_DAY  AMT_REQ_CREDIT_BUREAU_WEEK  \
307506                        NaN                         NaN
307507                        NaN                         NaN
307508                        0.0                         0.0
307509                        0.0                         0.0
307510                        0.0                         0.0

        AMT_REQ_CREDIT_BUREAU_MON  AMT_REQ_CREDIT_BUREAU_QRT  \
307506                        NaN                        NaN
307507                        NaN                        NaN
307508                        1.0                        0.0
307509                        0.0                        0.0
307510                        2.0                        0.0

        AMT_REQ_CREDIT_BUREAU_YEAR
307506                         NaN
307507                         NaN
307508                         1.0
307509                         0.0
307510                         1.0
```

In [8]:
```
app_train.shape, app_test.shape
```

Out[8]:
```
((307511, 122), (48744, 121))
```

Of course, the column named 'TARGET' is not in the test dataset.

2.2 Content of each table and links between tables

- **application_{train|test}.csv**
  - This is the main table, broken into two files for Train (with TARGET) and Test (without TARGET).
  - Static data for all applications. One row represents one loan in our data sample.
- **bureau.csv**
  - All client's previous credits provided by other financial institutions that were reported to Credit Bureau (for clients who have a loan in our sample).

- For every loan in our sample, there are as many rows as number of credits the client had in Credit Bureau before the application date.

- **bureau_balance.csv**
  - Monthly balances of previous credits in Credit Bureau.
  - This table has one row for each month of history of every previous credit reported to Credit Bureau – i.e the table has (#loans in sample * # of relative previous credits * # of months where we have some history observable for the previous credits) rows.
- **POS_CASH_balance.csv**
  - Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit.
  - This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample * # of relative previous credits * # of months in which we have some history observable for the previous credits) rows.
- **credit_card_balance.csv**
  - Monthly balance snapshots of previous credit cards that the applicant has with Home Credit.
  - This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample * # of relative previous credit cards * # of months where we have some history observable for the previous credit card) rows.
- **previous_application.csv**
  - All previous applications for Home Credit loans of clients who have loans in our sample.
  - There is one row for each previous application related to loans in our data sample.
- **installments_payments.csv**
  - Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample.
  - There is a) one row for every payment that was made plus b) one row each for missed payment.
  - One row is equivalent to one payment of one installment OR one installment corresponding to one payment of one previous Home Credit credit related to loans in our sample.
- **HomeCredit_columns_description.csv**
  - This file contains descriptions for the columns in the various data files.

---

3 Exploratory Data Analysisű

3.1 Distribution of the Target Column

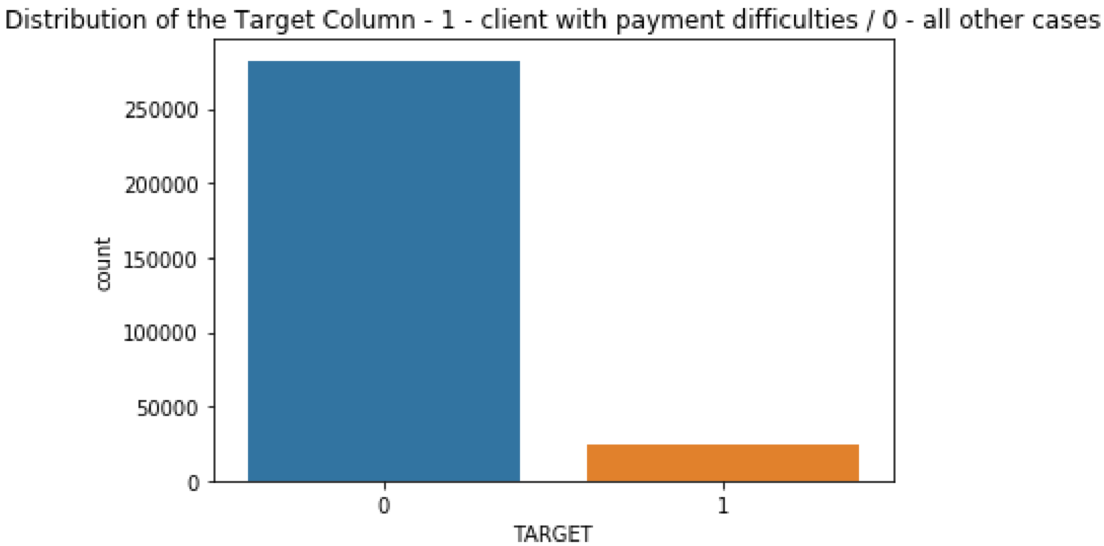
```
In [9]: app_train.TARGET.value_counts()
```

```
Out[9]: 0    282686
        1     24825
        Name: TARGET, dtype: int64
```

```
In [10]: print(f'percentage of clients with payment difficulties: {app_train.TARGET.sum() / ap
```

```
percentage of clients with payment difficulties: 8.07%
```

```
In [11]: plt.title('Distribution of the Target Column - 1 - client with payment difficulties /
         sns.countplot(x=app_train.TARGET, data=app_train)
         plt.show()
```



This is an imbalanced class problem. There are far more repaid loans than loans that were not repaid. It is important to weight the classes by their representation in the data to reflect this imbalance.

3.2 Column Types

```
In [12]: app_train.dtypes.value_counts()
```

```
Out[12]: float64    65
         int64      41
         object     16
         dtype: int64
```

int64 and float64 are numeric variables which can correspond to discrete or continuous features. Whereas object columns contain strings and are categorical features.

```
In [13]: # Number of unique classes in each object column
         app_train.select_dtypes('object').apply(pd.Series.nunique, axis=0)
```

```
Out[13]: 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
```

3.3 Missing Values

```
In [14]: # Function to calculate missing values by column# Funct // credits Will Koehrsen
         def missing_values_table(df):
                 # Total missing values
                 mis_val = df.isnull().sum()

                 # Percentage of missing values
                 mis_val_percent = 100 * df.isnull().sum() / len(df)

                 # Make a table with the results
                 mis_val_table = pd.concat([mis_val, mis_val_percent], axis=1)

                 # Rename the columns
                 mis_val_table_ren_columns = mis_val_table.rename(
                 columns = {0 : 'Missing Values', 1 : '% of Total Values'})

                 # Sort the table by percentage of missing descending
                 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 some summary information
                 print ("Your selected dataframe has " + str(df.shape[1]) + " columns.\n"
                     "There are " + str(mis_val_table_ren_columns.shape[0]) +
                       " columns that have missing values.")

                 # Return the dataframe with missing information
                 return mis_val_table_ren_columns
```

```
In [15]: # Missing values statistics
         missing_values = missing_values_table(app_train)
         missing_values.head(10)
```

```
Your selected dataframe has 122 columns.
There are 67 columns that have missing values.
```

```
Out[15]:                           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
         NONLIVINGAPARTMENTS_AVG           213514               69.4
         FONDKAPREMONT_MODE                210295               68.4
         LIVINGAPARTMENTS_MODE             210199               68.4
         LIVINGAPARTMENTS_MEDI             210199               68.4
         LIVINGAPARTMENTS_AVG              210199               68.4
```

From here we have 2 options : * Use models such as XGBoost that can handle missing values * Or drop columns with a high percentage of missing values, and fill in other columns with a low percentage It is not possible to know ahead of time if these columns will be helpful or not. My choice here is to drop them. Later if we need a more accurate score, we'll change the way to proceed.

3.3.1 Dropping columns with a high ratio of missing values

```
In [16]: # cols_to_drop = list((app_train.isnull().sum() > 75000).index)
         cols_to_drop = [c for c in app_train.columns if app_train[c].isnull().sum() > 75000]
```

```
In [17]: app_train, app_test = app_train.drop(cols_to_drop, axis=1), app_test.drop(cols_to_drop
         app_test.isnull().sum().sort_values(ascending=False).head(10)
```

```
Out[17]: EXT_SOURCE_3                  8668
         AMT_REQ_CREDIT_BUREAU_YEAR    6049
         AMT_REQ_CREDIT_BUREAU_MON     6049
         AMT_REQ_CREDIT_BUREAU_WEEK    6049
         AMT_REQ_CREDIT_BUREAU_DAY     6049
         AMT_REQ_CREDIT_BUREAU_HOUR    6049
         AMT_REQ_CREDIT_BUREAU_QRT     6049
         NAME_TYPE_SUITE                911
         DEF_60_CNT_SOCIAL_CIRCLE        29
         OBS_60_CNT_SOCIAL_CIRCLE        29
         dtype: int64
```

3.3.2 Filling other missing values

```
In [18]: obj_cols = app_train.select_dtypes('object').columns
         obj_cols
```

Out[18]:
```
Index(['NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY',
       'NAME_TYPE_SUITE', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE',
       'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE', 'WEEKDAY_APPR_PROCESS_START',
       'ORGANIZATION_TYPE'],
      dtype='object')
```

In [19]:
```
# filling string cols with 'Not specified'
app_train[obj_cols] = app_train[obj_cols].fillna('Not specified')
app_test[obj_cols] = app_test[obj_cols].fillna('Not specified')
```

In [20]:
```
float_cols = app_train.select_dtypes('float').columns
float_cols
```

Out[20]:
```
Index(['AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE',
       'REGION_POPULATION_RELATIVE', 'DAYS_REGISTRATION', 'CNT_FAM_MEMBERS',
       'EXT_SOURCE_2', 'EXT_SOURCE_3', 'OBS_30_CNT_SOCIAL_CIRCLE',
       'DEF_30_CNT_SOCIAL_CIRCLE', 'OBS_60_CNT_SOCIAL_CIRCLE',
       'DEF_60_CNT_SOCIAL_CIRCLE', 'DAYS_LAST_PHONE_CHANGE',
       'AMT_REQ_CREDIT_BUREAU_HOUR', 'AMT_REQ_CREDIT_BUREAU_DAY',
       'AMT_REQ_CREDIT_BUREAU_WEEK', 'AMT_REQ_CREDIT_BUREAU_MON',
       'AMT_REQ_CREDIT_BUREAU_QRT', 'AMT_REQ_CREDIT_BUREAU_YEAR'],
      dtype='object')
```

In [21]:
```
# filling float values with median of train (not test)
app_train[float_cols] = app_train[float_cols].fillna(app_train[float_cols].median())
app_test[float_cols] = app_test[float_cols].fillna(app_test[float_cols].median())
```

In [22]:
```
app_train.shape, app_test.shape
```

Out[22]:
```
((307511, 72), (48744, 71))
```

Let's check if there is still NaNs

In [23]:
```
app_train.isnull().sum().sort_values(ascending=False).head()
```

Out[23]:
```
AMT_REQ_CREDIT_BUREAU_YEAR    0
AMT_REQ_CREDIT_BUREAU_QRT     0
DAYS_REGISTRATION             0
DAYS_ID_PUBLISH               0
FLAG_MOBIL                    0
dtype: int64
```

In [24]:
```
app_test.isnull().sum().sort_values(ascending=False).head()
```

Out[24]:
```
AMT_REQ_CREDIT_BUREAU_YEAR    0
FLAG_EMAIL                    0
DAYS_ID_PUBLISH               0
FLAG_MOBIL                    0
FLAG_EMP_PHONE                0
dtype: int64
```

In [25]:
```
# Is there any duplicated rows ?
```

In [26]:
```
app_train.duplicated().sum()
```

Out[26]:
```
0
```

In [27]:
```
app_test.duplicated().sum()
```

Out[27]:
```
0
```

3.4 Categorical columns (type object)

In [28]:
```
# Number of unique classes in each object column
app_train.select_dtypes('object').apply(pd.Series.nunique, axis = 0)
```

Out[28]:
```
NAME_CONTRACT_TYPE             2
CODE_GENDER                    3
FLAG_OWN_CAR                   2
FLAG_OWN_REALTY                2
NAME_TYPE_SUITE                8
NAME_INCOME_TYPE               8
NAME_EDUCATION_TYPE            5
NAME_FAMILY_STATUS             6
NAME_HOUSING_TYPE              6
WEEKDAY_APPR_PROCESS_START     7
ORGANIZATION_TYPE             58
dtype: int64
```

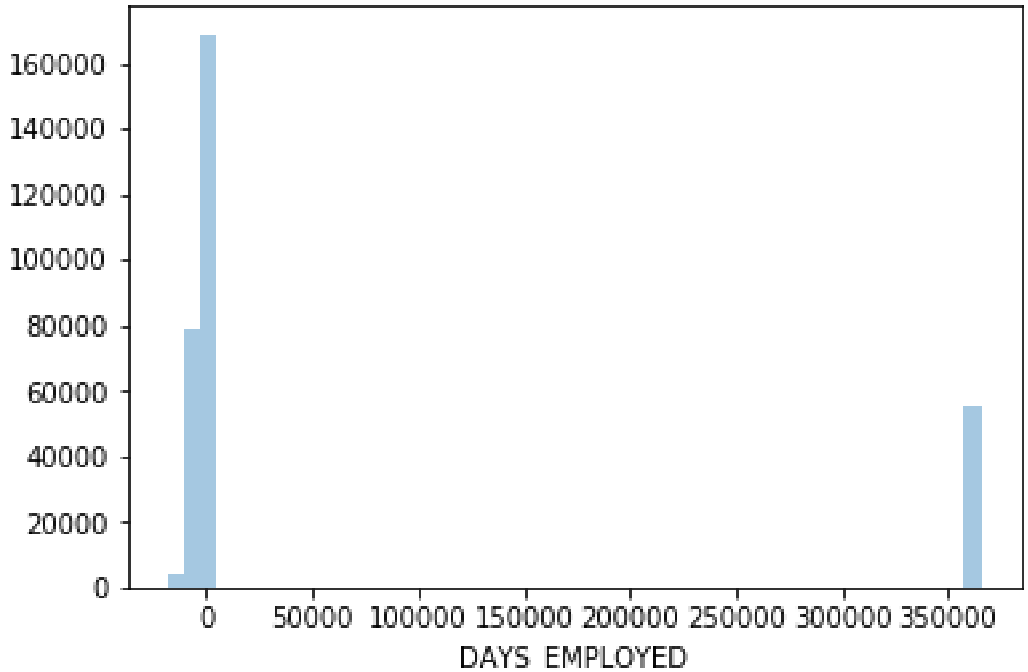
3.5 Dealing with anomalies

In [29]:
```
app_train['DAYS_EMPLOYED'].describe()
```

Out[29]:
```
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
```

The maximum value is abnormal (besides being positive). It corresponds to 1000 years...

In [30]:
```
sns.distplot(app_train['DAYS_EMPLOYED'], kde=False);
plt.show()
```




```
In [31]: print('The non-anomalies default on %0.2f%% of loans' % (100 * app_train[app_train['DA
         print('The anomalies default on %0.2f%% of loans' % (100 * app_train[app_train['DAYS_
         print('There are %d anomalous days of employment' % len(app_train[app_train['DAYS_EMP
```

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

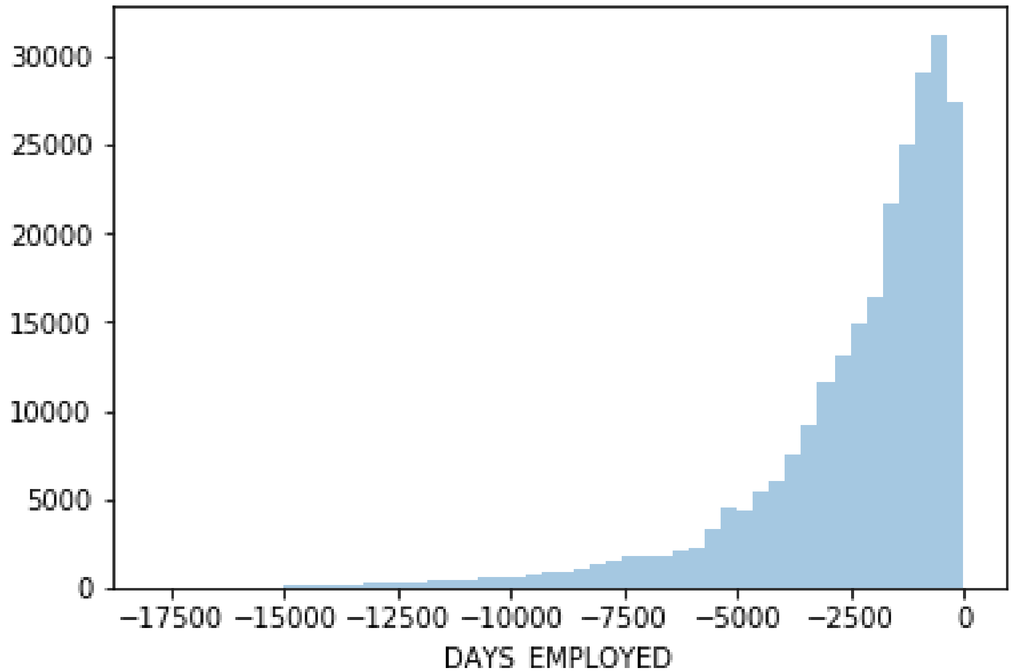
It turns out that the anomalies have a lower rate of default.

The anomalous values seem to have some importance. Let's fill in the anomalous values with not a np.nan and then create a new boolean column indicating whether or not the value was anomalous.

```
In [32]: # Create an anomalous flag column
         app_train['DAYS_EMPLOYED_ANOM'] = app_train["DAYS_EMPLOYED"] == 365243

         # Replace the anomalous values with nan
         app_train['DAYS_EMPLOYED'].replace({365243: np.nan}, inplace = True)

         sns.distplot(app_train['DAYS_EMPLOYED'].dropna(), kde=False);
```




```
In [33]: app_test['DAYS_EMPLOYED_ANOM'] = app_test["DAYS_EMPLOYED"] == 365243
         app_test["DAYS_EMPLOYED"].replace({365243: np.nan}, inplace = True)

         print('There are %d anomalies in the test data out of %d entries' % (app_test["DAYS_EI
```

```
There are 9274 anomalies in the test data out of 48744 entries
```

```
In [34]: # refilling float values with median of train (not test)

         app_train[float_cols] = app_train[float_cols].apply(pd.to_numeric, errors='coerce')
         app_train = app_train.fillna(app_train.median())

         app_test[float_cols] = app_test[float_cols].apply(pd.to_numeric, errors='coerce')
         app_test = app_train.fillna(app_test.median())
```

3.6 Correlations

The correlation coefficient is not the best method to represent "relevance" of a feature, but it gives us an idea of possible relationships within the data. Some general interpretations of the absolute value of the correlation coefficent are:

- 00-.19 “very weak”
- 20-.39 “weak”
- 40-.59 “moderate”

- 60-.79 "strong"
- 80-1.0 "very strong"

In [35]:
```
correlations = app_train.corr()['TARGET'].sort_values()

print('Most Positive Correlations:\n', correlations.tail(10))
print('\n\nMost Negative Correlations:\n', correlations.head(10))
```

```
Most Positive Correlations:
 REG_CITY_NOT_LIVE_CITY         0.044395
FLAG_EMP_PHONE                 0.045982
REG_CITY_NOT_WORK_CITY         0.050994
DAYS_ID_PUBLISH                0.051457
DAYS_LAST_PHONE_CHANGE         0.055218
REGION_RATING_CLIENT           0.058899
REGION_RATING_CLIENT_W_CITY    0.060893
DAYS_EMPLOYED                  0.063368
DAYS_BIRTH                     0.078239
TARGET                         1.000000
Name: TARGET, dtype: float64


Most Negative Correlations:
 EXT_SOURCE_2                  -0.160295
EXT_SOURCE_3                  -0.155892
DAYS_EMPLOYED_ANOM            -0.045987
AMT_GOODS_PRICE               -0.039623
REGION_POPULATION_RELATIVE    -0.037227
AMT_CREDIT                    -0.030369
FLAG_DOCUMENT_6               -0.028602
HOUR_APPR_PROCESS_START       -0.024166
FLAG_PHONE                    -0.023806
AMT_REQ_CREDIT_BUREAU_MON     -0.014794
Name: TARGET, dtype: float64
```

In [36]:
```
# Compute the correlation matrix
corr = app_train.corr()

# Generate a mask for the upper triangle
mask = np.zeros_like(corr, dtype=np.bool)
mask[np.triu_indices_from(mask)] = True

# Set up the matplotlib figure
f, ax = plt.subplots(figsize=(21, 19))

# Generate a custom diverging colormap
cmap = sns.diverging_palette(220, 10, as_cmap=True)
```

```
# Draw the heatmap with the mask and correct aspect ratio
sns.heatmap(corr, mask=mask, cmap=cmap, vmax=.3, center=0,
            square=True, linewidths=.5, cbar_kws={"shrink": .5})
```

Out[36]: <matplotlib.axes._subplots.AxesSubplot at 0x7f7b866f50f0>

3.6.1 Effect of Age on Repayment

In [37]:
```
# Find the correlation of the positive days since birth and target
app_train['DAYS_BIRTH'] = abs(app_train['DAYS_BIRTH'])
app_train['DAYS_BIRTH'].corr(app_train['TARGET'])
```

Out[37]: -0.07823930830982712

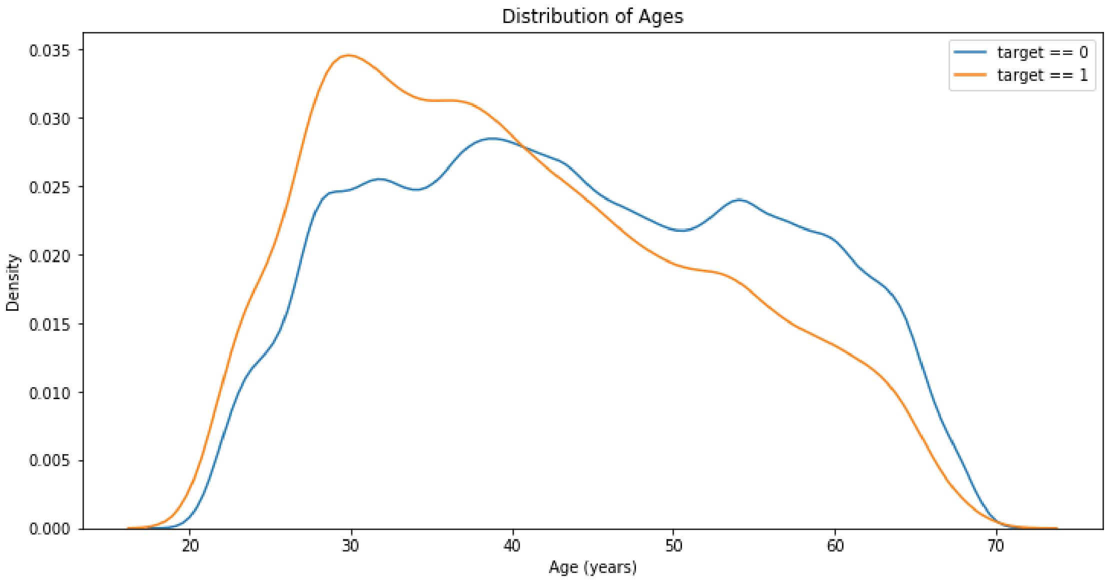
There isn't any correlation between age and repayment

In [38]:
```
plt.figure(figsize = (12, 6))

# KDE plot of loans that were repaid on time
sns.kdeplot(app_train.loc[app_train['TARGET'] == 0, 'DAYS_BIRTH'] / 365, label = 'targ

# KDE plot of loans which were not repaid on time
sns.kdeplot(app_train.loc[app_train['TARGET'] == 1, 'DAYS_BIRTH'] / 365, label = 'targ

# Labeling of plot
plt.xlabel('Age (years)'); plt.ylabel('Density'); plt.title('Distribution of Ages');
```



In [39]:
```
# Age information into a separate dataframe
age_data = app_train[['TARGET', 'DAYS_BIRTH']]
age_data['YEARS_BIRTH'] = age_data['DAYS_BIRTH'] / 365

# Bin the age data
age_data['YEARS_BINNED'] = pd.cut(age_data['YEARS_BIRTH'], bins = np.linspace(20, 70,
age_data.head(10)
```

```
/home/sunflowa/anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:3: SettingWithCopyWa
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: http://pandas.pydata.org/pandas-docs/stable/indexing.html
  This is separate from the ipykernel package so we can avoid doing imports until
/home/sunflowa/anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:6: SettingWithCopyWa
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: http://pandas.pydata.org/pandas-docs/stable/indexing.html
```

Out[39]:

	TARGET	DAYS_BIRTH	YEARS_BIRTH	YEARS_BINNED
0	1	9461	25.920548	(25.0, 30.0]
1	0	16765	45.931507	(45.0, 50.0]
2	0	19046	52.180822	(50.0, 55.0]
3	0	19005	52.068493	(50.0, 55.0]
4	0	19932	54.608219	(50.0, 55.0]
5	0	16941	46.413699	(45.0, 50.0]
6	0	13778	37.747945	(35.0, 40.0]
7	0	18850	51.643836	(50.0, 55.0]
8	0	20099	55.065753	(55.0, 60.0]
9	0	14469	39.641096	(35.0, 40.0]

In [40]:

```
# Group by the bin and calculate averages
age_groups  = age_data.groupby('YEARS_BINNED').mean()
age_groups
```

Out[40]:

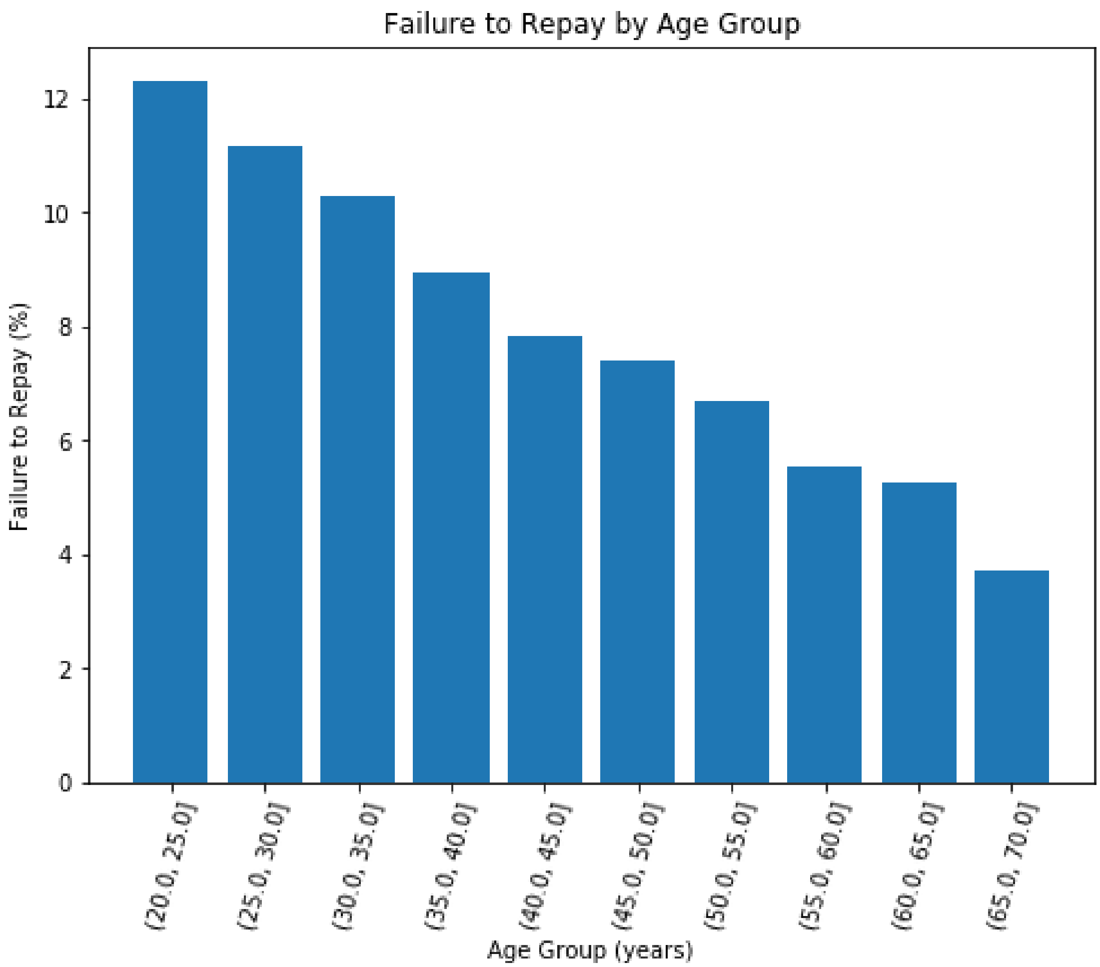
YEARS_BINNED	TARGET	DAYS_BIRTH	YEARS_BIRTH
(20.0, 25.0]	0.123036	8532.795625	23.377522
(25.0, 30.0]	0.111436	10155.219250	27.822518
(30.0, 35.0]	0.102814	11854.848377	32.479037
(35.0, 40.0]	0.089414	13707.908253	37.555913
(40.0, 45.0]	0.078491	15497.661233	42.459346
(45.0, 50.0]	0.074171	17323.900441	47.462741
(50.0, 55.0]	0.066968	19196.494791	52.593136
(55.0, 60.0]	0.055314	20984.262742	57.491131
(60.0, 65.0]	0.052737	22780.547460	62.412459
(65.0, 70.0]	0.037270	24292.614340	66.555108

In [41]:

```
plt.figure(figsize = (8, 6))

# Graph the age bins and the average of the target as a bar plot
plt.bar(age_groups.index.astype(str), 100 * age_groups['TARGET'])

# Plot labeling
plt.xticks(rotation = 75); plt.xlabel('Age Group (years)'); plt.ylabel('Failure to Rep
plt.title('Failure to Repay by Age Group');
```



Younger applicants are more likely to not repay the loan.

4 Preparing data

4.1 Encoding Categorical Variables

A ML model can't deal with categorical features (except for some models such as LightGBM). One have to find a way to encode (represent) these variables as numbers. There are two main ways :

- Label encoding: assign each unique category in a categorical variable with an integer. No new columns are created. The problem with label encoding is that it gives the categories an arbitrary ordering.
- One-hot encoding: create a new column for each unique category in a categorical variable. Each observation recieves a 1 in the column for its corresponding category and a 0 in all other new columns.

```
In [42]: app_train = pd.get_dummies(data=app_train, columns=obj_cols)
         app_test = pd.get_dummies(data=app_test, columns=obj_cols)
```

4.2 Aligning Training and Testing Data

Both the training and testing data should have the same features (columns). One-hot encoding can more columns in the one dataset because there were some categorical variables with categories not represented in the other dataset. In order to remove the columns in the training data that are not in the testing data, one need to align the dataframes.

```
In [43]: # back up of the target /  need to keep this information
         y = app_train.TARGET
         app_train = app_train.drop(columns=['TARGET'])
```

```
In [44]: app_train, app_test = app_train.align(app_test, join = 'inner', axis = 1)
```

```
In [45]: app_train.shape, app_test.shape
```

```
Out[45]: ((307511, 168), (307511, 168))
```

4.3 Scaling values

```
In [46]: feat_to_scale = list(float_cols).copy()
         feat_to_scale.extend(['CNT_CHILDREN', 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_ID_PUBLISH
         feat_to_scale
```

```
Out[46]: ['AMT_INCOME_TOTAL',
          'AMT_CREDIT',
          'AMT_ANNUITY',
          'AMT_GOODS_PRICE',
          'REGION_POPULATION_RELATIVE',
          'DAYS_REGISTRATION',
          'CNT_FAM_MEMBERS',
          'EXT_SOURCE_2',
          'EXT_SOURCE_3',
          'OBS_30_CNT_SOCIAL_CIRCLE',
          'DEF_30_CNT_SOCIAL_CIRCLE',
          'OBS_60_CNT_SOCIAL_CIRCLE',
          'DEF_60_CNT_SOCIAL_CIRCLE',
          'DAYS_LAST_PHONE_CHANGE',
          'AMT_REQ_CREDIT_BUREAU_HOUR',
          'AMT_REQ_CREDIT_BUREAU_DAY',
          'AMT_REQ_CREDIT_BUREAU_WEEK',
          'AMT_REQ_CREDIT_BUREAU_MON',
          'AMT_REQ_CREDIT_BUREAU_QRT',
          'AMT_REQ_CREDIT_BUREAU_YEAR',
          'CNT_CHILDREN',
          'DAYS_BIRTH',
          'DAYS_EMPLOYED',
          'DAYS_ID_PUBLISH',
          'HOUR_APPR_PROCESS_START']
```

```
In [47]: scaler = StandardScaler()
         app_train[feat_to_scale] = scaler.fit_transform(app_train[feat_to_scale])
         app_test[feat_to_scale] = scaler.fit_transform(app_test[feat_to_scale])
         app_train.head()
```

```
/home/sunflowa/anaconda3/lib/python3.7/site-packages/sklearn/preprocessing/data.py:625: DataCon
  return self.partial_fit(X, y)
/home/sunflowa/anaconda3/lib/python3.7/site-packages/sklearn/base.py:462: DataConversionWarning
  return self.fit(X, **fit_params).transform(X)
/home/sunflowa/anaconda3/lib/python3.7/site-packages/sklearn/preprocessing/data.py:625: DataCon
  return self.partial_fit(X, y)
/home/sunflowa/anaconda3/lib/python3.7/site-packages/sklearn/base.py:462: DataConversionWarning
  return self.fit(X, **fit_params).transform(X)
```

```
Out[47]:    SK_ID_CURR  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  AMT_ANNUITY  \
         0      100002     -0.577538          0.142129   -0.478095    -0.166143
         1      100003     -0.577538          0.426792    1.725450     0.592683
         2      100004     -0.577538         -0.427196   -1.152888    -1.404669
         3      100006     -0.577538         -0.142533   -0.711430     0.177874
         4      100007     -0.577538         -0.199466   -0.213734    -0.361749

            AMT_GOODS_PRICE  REGION_POPULATION_RELATIVE  DAYS_BIRTH  DAYS_EMPLOYED  \
         0        -0.507236                   -0.149452   -1.506880       0.755835
         1         1.600873                   -1.252750    0.166821       0.497899
         2        -1.092145                   -0.783451    0.689509       0.948701
         3        -0.653463                   -0.928991    0.680114      -0.368597
         4        -0.068554                    0.563570    0.892535      -0.368129

            DAYS_REGISTRATION  DAYS_ID_PUBLISH  FLAG_MOBIL  FLAG_EMP_PHONE  \
         0           0.379837         0.579154           1               1
         1           1.078697         1.790855           1               1
         2           0.206116         0.306869           1               1
         3          -1.375829         0.369143           1               1
         4           0.191639        -0.307263           1               1

            FLAG_WORK_PHONE  FLAG_CONT_MOBILE  FLAG_PHONE  FLAG_EMAIL  CNT_FAM_MEMBERS  \
         0                0                 1           1           0        -1.265722
         1                0                 1           1           0        -0.167638
         2                1                 1           1           0        -1.265722
         3                0                 1           0           0        -0.167638
         4                0                 1           0           0        -1.265722

            REGION_RATING_CLIENT  REGION_RATING_CLIENT_W_CITY  HOUR_APPR_PROCESS_START  \
         0                     2                            2                -0.631821
         1                     1                            1                -0.325620
         2                     2                            2                -0.938022
         3                     2                            2                 1.511587
```

```
4                    2                             2             -0.325620

   REG_REGION_NOT_LIVE_REGION  REG_REGION_NOT_WORK_REGION  \
0                           0                           0
1                           0                           0
2                           0                           0
3                           0                           0
4                           0                           0

   LIVE_REGION_NOT_WORK_REGION  REG_CITY_NOT_LIVE_CITY  \
0                            0                       0
1                            0                       0
2                            0                       0
3                            0                       0
4                            0                       0

   REG_CITY_NOT_WORK_CITY  LIVE_CITY_NOT_WORK_CITY  EXT_SOURCE_2  \
0                       0                        0     -1.317940
1                       0                        0      0.564482
2                       0                        0      0.216948
3                       0                        0      0.712205
4                       1                        1     -1.004691

   EXT_SOURCE_3  OBS_30_CNT_SOCIAL_CIRCLE  DEF_30_CNT_SOCIAL_CIRCLE  \
0     -2.153651                  0.242861                  4.163504
1      0.112063                 -0.174085                 -0.320480
2      1.223975                 -0.591031                 -0.320480
3      0.112063                  0.242861                 -0.320480
4      0.112063                 -0.591031                 -0.320480

   OBS_60_CNT_SOCIAL_CIRCLE  DEF_60_CNT_SOCIAL_CIRCLE  DAYS_LAST_PHONE_CHANGE  \
0                  0.252132                  5.253260               -0.206992
1                 -0.168527                 -0.275663                0.163107
2                 -0.589187                 -0.275663                0.178831
3                  0.252132                 -0.275663                0.418306
4                 -0.589187                 -0.275663               -0.173126

   FLAG_DOCUMENT_2  FLAG_DOCUMENT_3  FLAG_DOCUMENT_4  FLAG_DOCUMENT_5  \
0                0                1                0                0
1                0                1                0                0
2                0                0                0                0
3                0                1                0                0
4                0                0                0                0

   FLAG_DOCUMENT_6  FLAG_DOCUMENT_7  FLAG_DOCUMENT_8  FLAG_DOCUMENT_9  \
0                0                0                0                0
1                0                0                0                0
2                0                0                0                0
```

```
3                0                0                0                0
4                0                0                1                0

   FLAG_DOCUMENT_10  FLAG_DOCUMENT_11  FLAG_DOCUMENT_12  FLAG_DOCUMENT_13  \
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

   FLAG_DOCUMENT_14  FLAG_DOCUMENT_15  FLAG_DOCUMENT_16  FLAG_DOCUMENT_17  \
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

   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                   -0.070987                  -0.058766
1                   -0.070987                  -0.058766
2                   -0.070987                  -0.058766
3                   -0.070987                  -0.058766
4                   -0.070987                  -0.058766

   AMT_REQ_CREDIT_BUREAU_WEEK  AMT_REQ_CREDIT_BUREAU_MON  \
0                   -0.155837                  -0.269947
1                   -0.155837                  -0.269947
2                   -0.155837                  -0.269947
3                   -0.155837                  -0.269947
4                   -0.155837                  -0.269947

   AMT_REQ_CREDIT_BUREAU_QRT  AMT_REQ_CREDIT_BUREAU_YEAR  DAYS_EMPLOYED_ANOM  \
0                   -0.30862                   -0.440926               False
1                   -0.30862                   -1.007331               False
2                   -0.30862                   -1.007331               False
3                   -0.30862                   -0.440926               False
4                   -0.30862                   -1.007331               False

   NAME_CONTRACT_TYPE_Cash loans  NAME_CONTRACT_TYPE_Revolving loans  \
0                              1                                   0
1                              1                                   0
```

```
2                          0                              1
3                          1                              0
4                          1                              0

   CODE_GENDER_F  CODE_GENDER_M  CODE_GENDER_XNA  FLAG_OWN_CAR_N  \
0              0              1                0               1
1              1              0                0               1
2              0              1                0               0
3              1              0                0               1
4              0              1                0               1

   FLAG_OWN_CAR_Y  FLAG_OWN_REALTY_N  FLAG_OWN_REALTY_Y  \
0               0                  0                  1
1               0                  1                  0
2               1                  0                  1
3               0                  0                  1
4               0                  0                  1

   NAME_TYPE_SUITE_Children  NAME_TYPE_SUITE_Family  \
0                         0                       0
1                         0                       1
2                         0                       0
3                         0                       0
4                         0                       0

   NAME_TYPE_SUITE_Group of people  NAME_TYPE_SUITE_Not specified  \
0                                0                              0
1                                0                              0
2                                0                              0
3                                0                              0
4                                0                              0

   NAME_TYPE_SUITE_Other_A  NAME_TYPE_SUITE_Other_B  \
0                        0                        0
1                        0                        0
2                        0                        0
3                        0                        0
4                        0                        0

   NAME_TYPE_SUITE_Spouse, partner  NAME_TYPE_SUITE_Unaccompanied  \
0                                0                              1
1                                0                              0
2                                0                              1
3                                0                              1
4                                0                              1

   NAME_INCOME_TYPE_Businessman  NAME_INCOME_TYPE_Commercial associate  \
0                             0                                      0
```

```
1                               0                                         0
2                               0                                         0
3                               0                                         0
4                               0                                         0

   NAME_INCOME_TYPE_Maternity leave  NAME_INCOME_TYPE_Pensioner  \
0                                 0                           0
1                                 0                           0
2                                 0                           0
3                                 0                           0
4                                 0                           0

   NAME_INCOME_TYPE_State servant  NAME_INCOME_TYPE_Student  \
0                               0                         0
1                               1                         0
2                               0                         0
3                               0                         0
4                               0                         0

   NAME_INCOME_TYPE_Unemployed  NAME_INCOME_TYPE_Working  \
0                            0                         1
1                            0                         0
2                            0                         1
3                            0                         1
4                            0                         1

   NAME_EDUCATION_TYPE_Academic degree  NAME_EDUCATION_TYPE_Higher education  \
0                                    0                                     0
1                                    0                                     1
2                                    0                                     0
3                                    0                                     0
4                                    0                                     0

   NAME_EDUCATION_TYPE_Incomplete higher  NAME_EDUCATION_TYPE_Lower secondary  \
0                                      0                                    0
1                                      0                                    0
2                                      0                                    0
3                                      0                                    0
4                                      0                                    0

   NAME_EDUCATION_TYPE_Secondary / secondary special  \
0                                                  1
1                                                  0
2                                                  1
3                                                  1
4                                                  1

   NAME_FAMILY_STATUS_Civil marriage  NAME_FAMILY_STATUS_Married  \
```

```
0                                  0                            0
1                                  0                            1
2                                  0                            0
3                                  1                            0
4                                  0                            0

   NAME_FAMILY_STATUS_Separated  NAME_FAMILY_STATUS_Single / not married  \
0                             0                                        1
1                             0                                        0
2                             0                                        1
3                             0                                        0
4                             0                                        1

   NAME_FAMILY_STATUS_Unknown  NAME_FAMILY_STATUS_Widow  \
0                           0                         0
1                           0                         0
2                           0                         0
3                           0                         0
4                           0                         0

   NAME_HOUSING_TYPE_Co-op apartment  NAME_HOUSING_TYPE_House / apartment  \
0                                  0                                    1
1                                  0                                    1
2                                  0                                    1
3                                  0                                    1
4                                  0                                    1

   NAME_HOUSING_TYPE_Municipal apartment  NAME_HOUSING_TYPE_Office apartment  \
0                                      0                                   0
1                                      0                                   0
2                                      0                                   0
3                                      0                                   0
4                                      0                                   0

   NAME_HOUSING_TYPE_Rented apartment  NAME_HOUSING_TYPE_With parents  \
0                                   0                               0
1                                   0                               0
2                                   0                               0
3                                   0                               0
4                                   0                               0

   WEEKDAY_APPR_PROCESS_START_FRIDAY  WEEKDAY_APPR_PROCESS_START_MONDAY  \
0                                  0                                  0
1                                  0                                  1
2                                  0                                  1
3                                  0                                  0
4                                  0                                  0
```

```
   WEEKDAY_APPR_PROCESS_START_SATURDAY  WEEKDAY_APPR_PROCESS_START_SUNDAY  \
0                                    0                                  0
1                                    0                                  0
2                                    0                                  0
3                                    0                                  0
4                                    0                                  0

   WEEKDAY_APPR_PROCESS_START_THURSDAY  WEEKDAY_APPR_PROCESS_START_TUESDAY  \
0                                    0                                   0
1                                    0                                   0
2                                    0                                   0
3                                    0                                   0
4                                    1                                   0

   WEEKDAY_APPR_PROCESS_START_WEDNESDAY  ORGANIZATION_TYPE_Advertising  \
0                                     1                              0
1                                     0                              0
2                                     0                              0
3                                     1                              0
4                                     0                              0

   ORGANIZATION_TYPE_Agriculture  ORGANIZATION_TYPE_Bank  \
0                              0                       0
1                              0                       0
2                              0                       0
3                              0                       0
4                              0                       0

   ORGANIZATION_TYPE_Business Entity Type 1  \
0                                         0
1                                         0
2                                         0
3                                         0
4                                         0

   ORGANIZATION_TYPE_Business Entity Type 2  \
0                                         0
1                                         0
2                                         0
3                                         0
4                                         0

   ORGANIZATION_TYPE_Business Entity Type 3  ORGANIZATION_TYPE_Cleaning  \
0                                         1                           0
1                                         0                           0
2                                         0                           0
3                                         1                           0
4                                         0                           0
```

```
   ORGANIZATION_TYPE_Construction  ORGANIZATION_TYPE_Culture  \
0                               0                          0
1                               0                          0
2                               0                          0
3                               0                          0
4                               0                          0

   ORGANIZATION_TYPE_Electricity  ORGANIZATION_TYPE_Emergency  \
0                              0                            0
1                              0                            0
2                              0                            0
3                              0                            0
4                              0                            0

   ORGANIZATION_TYPE_Government  ORGANIZATION_TYPE_Hotel  \
0                             0                        0
1                             0                        0
2                             1                        0
3                             0                        0
4                             0                        0

   ORGANIZATION_TYPE_Housing  ORGANIZATION_TYPE_Industry: type 1  \
0                          0                                   0
1                          0                                   0
2                          0                                   0
3                          0                                   0
4                          0                                   0

   ORGANIZATION_TYPE_Industry: type 10  ORGANIZATION_TYPE_Industry: type 11  \
0                                    0                                    0
1                                    0                                    0
2                                    0                                    0
3                                    0                                    0
4                                    0                                    0

   ORGANIZATION_TYPE_Industry: type 12  ORGANIZATION_TYPE_Industry: type 13  \
0                                    0                                    0
1                                    0                                    0
2                                    0                                    0
3                                    0                                    0
4                                    0                                    0

   ORGANIZATION_TYPE_Industry: type 2  ORGANIZATION_TYPE_Industry: type 3  \
0                                   0                                   0
1                                   0                                   0
2                                   0                                   0
3                                   0                                   0
```

```
4                                 0                                  0

   ORGANIZATION_TYPE_Industry: type 4  ORGANIZATION_TYPE_Industry: type 5  \
0                                 0                                  0
1                                 0                                  0
2                                 0                                  0
3                                 0                                  0
4                                 0                                  0

   ORGANIZATION_TYPE_Industry: type 6  ORGANIZATION_TYPE_Industry: type 7  \
0                                 0                                  0
1                                 0                                  0
2                                 0                                  0
3                                 0                                  0
4                                 0                                  0

   ORGANIZATION_TYPE_Industry: type 8  ORGANIZATION_TYPE_Industry: type 9  \
0                                 0                                  0
1                                 0                                  0
2                                 0                                  0
3                                 0                                  0
4                                 0                                  0

   ORGANIZATION_TYPE_Insurance  ORGANIZATION_TYPE_Kindergarten  \
0                          0                               0
1                          0                               0
2                          0                               0
3                          0                               0
4                          0                               0

   ORGANIZATION_TYPE_Legal Services  ORGANIZATION_TYPE_Medicine  \
0                               0                           0
1                               0                           0
2                               0                           0
3                               0                           0
4                               0                           0

   ORGANIZATION_TYPE_Military  ORGANIZATION_TYPE_Mobile  \
0                         0                         0
1                         0                         0
2                         0                         0
3                         0                         0
4                         0                         0

   ORGANIZATION_TYPE_Other  ORGANIZATION_TYPE_Police  \
0                       0                         0
1                       0                         0
2                       0                         0
```

```
3                        0                          0
4                        0                          0

   ORGANIZATION_TYPE_Postal  ORGANIZATION_TYPE_Realtor  \
0                         0                          0
1                         0                          0
2                         0                          0
3                         0                          0
4                         0                          0

   ORGANIZATION_TYPE_Religion  ORGANIZATION_TYPE_Restaurant  \
0                           0                             0
1                           0                             0
2                           0                             0
3                           0                             0
4                           1                             0

   ORGANIZATION_TYPE_School  ORGANIZATION_TYPE_Security  \
0                         0                           0
1                         1                           0
2                         0                           0
3                         0                           0
4                         0                           0

   ORGANIZATION_TYPE_Security Ministries  ORGANIZATION_TYPE_Self-employed  \
0                                      0                                0
1                                      0                                0
2                                      0                                0
3                                      0                                0
4                                      0                                0

   ORGANIZATION_TYPE_Services  ORGANIZATION_TYPE_Telecom  \
0                           0                          0
1                           0                          0
2                           0                          0
3                           0                          0
4                           0                          0

   ORGANIZATION_TYPE_Trade: type 1  ORGANIZATION_TYPE_Trade: type 2  \
0                                0                                0
1                                0                                0
2                                0                                0
3                                0                                0
4                                0                                0

   ORGANIZATION_TYPE_Trade: type 3  ORGANIZATION_TYPE_Trade: type 4  \
0                                0                                0
1                                0                                0
```

```
2                                  0                                    0
3                                  0                                    0
4                                  0                                    0

   ORGANIZATION_TYPE_Trade: type 5  ORGANIZATION_TYPE_Trade: type 6  \
0                                  0                                    0
1                                  0                                    0
2                                  0                                    0
3                                  0                                    0
4                                  0                                    0

   ORGANIZATION_TYPE_Trade: type 7  ORGANIZATION_TYPE_Transport: type 1  \
0                                  0                                        0
1                                  0                                        0
2                                  0                                        0
3                                  0                                        0
4                                  0                                        0

   ORGANIZATION_TYPE_Transport: type 2  ORGANIZATION_TYPE_Transport: type 3  \
0                                      0                                        0
1                                      0                                        0
2                                      0                                        0
3                                      0                                        0
4                                      0                                        0

   ORGANIZATION_TYPE_Transport: type 4  ORGANIZATION_TYPE_University  \
0                                      0                                 0
1                                      0                                 0
2                                      0                                 0
3                                      0                                 0
4                                      0                                 0

   ORGANIZATION_TYPE_XNA
0                       0
1                       0
2                       0
3                       0
4                       0
```

4.4 Splitting training / test datasets

from app_train in order to make few predictions before submission & select models

```
In [48]: X_train, X_test, y_train, y_test = train_test_split(app_train, y, test_size=0.2)
```

5 Base line

5.1 Metric: ROC AUC

more infos on the Receiver Operating Characteristic Area Under the Curve (ROC AUC, also sometimes called AUROC).

The Reciever Operating Characteristic (ROC) curve graphs the true positive rate versus the false positive rate:

A single line on the graph indicates the curve for a single model, and movement along a line indicates changing the threshold used for classifying a positive instance. The threshold starts at 0 in the upper right to and goes to 1 in the lower left. A curve that is to the left and above another curve indicates a better model. For example, the blue model is better than the red one (which is better than the black diagonal line which indicates a naive random guessing model).

The Area Under the Curve (AUC) is the integral of the curve. This metric is between 0 and 1 with a better model scoring higher. A model that simply guesses at random will have an ROC AUC of 0.5.

When we measure a classifier according to the ROC AUC, we do not generate 0 or 1 predictions, but rather a probability between 0 and 1.

When we get into problems with inbalanced classes, accuracy is not the best metric. A model with a high ROC AUC will also have a high accuracy, but the ROC AUC is a better representation of model performance.

5.2 Random forrest

```
In [49]: # a simple RandomForrest Classifier without CV
         rf = RandomForestClassifier(n_estimators=50)
         rf.fit(X_train, y_train)
         y_pred = rf.predict(X_test)
         roc_auc_score(y_test, y_pred)
```

```
Out[49]: 0.5011674887648608
```

The predictions must be in the format shown in the sample_submission.csv file, where there are only two columns: SK_ID_CURR and TARGET. Let's create a dataframe in this format from the test set and the predictions called submit.

```
In [50]: def submit(model, csv_name):

             # fit on the whole dataset of train
             model.fit(app_train, y)

             # Make predictions & make sure to select the second column only
             result = model.predict_proba(app_test)[:, 1]

             submit = app_test[['SK_ID_CURR']]
             submit['TARGET'] = result

             # Save the submission to a csv file
             submit.to_csv(csv_name, index = False)
```

```
In [51]: # submit(rf, 'random_forrest_clf.csv')
```

The random forrest model should score around 0.58329 when submitted which is not really good, because just above 0.5 i.e a random classifier...

5.3 Feature Importances

```
In [52]: importances = rf.feature_importances_
         std = np.std([tree.feature_importances_ for tree in rf.estimators_], axis=0)
         indices = np.argsort(importances)[::-1]

         # Print the feature ranking
         print("Feature ranking:")

         for f in range(app_train.shape[1]):
             print("%d. feature %d (%f)" % (f + 1, indices[f], importances[indices[f]]))
```

```
Feature ranking:
1. feature 27 (0.064841)
2. feature 28 (0.059712)
3. feature 7 (0.046094)
4. feature 10 (0.045429)
5. feature 9 (0.045176)
6. feature 0 (0.044173)
7. feature 4 (0.041803)
8. feature 8 (0.041481)
9. feature 33 (0.041030)
10. feature 3 (0.039948)
11. feature 2 (0.035454)
12. feature 6 (0.035247)
13. feature 5 (0.033751)
14. feature 20 (0.031268)
15. feature 59 (0.022036)
16. feature 31 (0.017715)
17. feature 29 (0.017527)
18. feature 17 (0.013756)
19. feature 1 (0.010035)
20. feature 58 (0.008328)
21. feature 57 (0.007471)
22. feature 18 (0.006926)
23. feature 69 (0.006884)
24. feature 19 (0.006881)
25. feature 15 (0.006740)
26. feature 92 (0.006712)
27. feature 68 (0.006608)
28. feature 30 (0.006549)
29. feature 115 (0.006527)
30. feature 108 (0.006369)
```

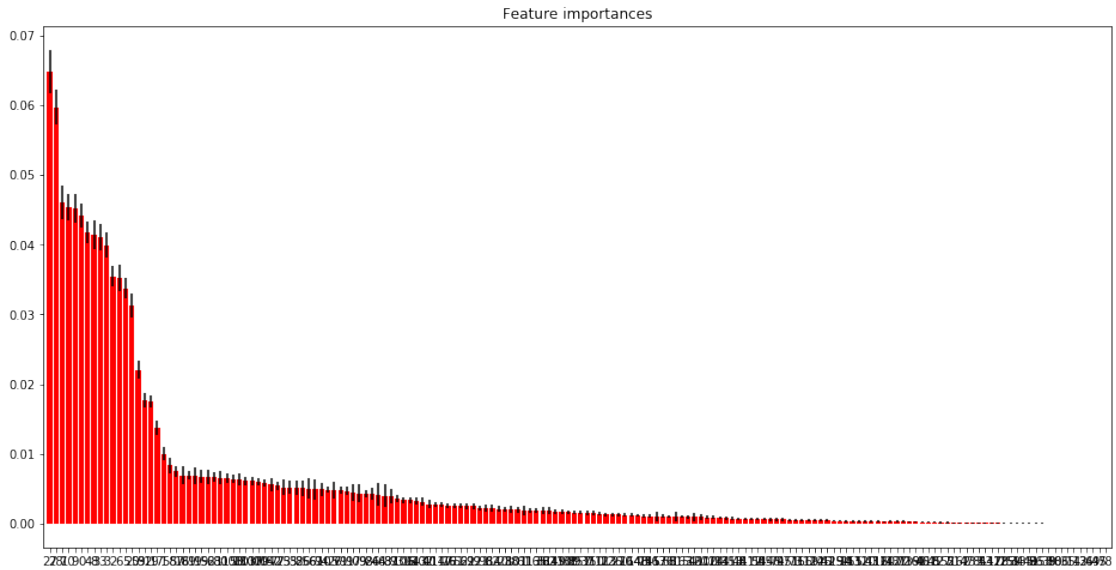
```
31. feature 13 (0.006358)
32. feature 103 (0.006115)
33. feature 107 (0.006110)
34. feature 109 (0.006067)
35. feature 104 (0.005811)
36. feature 152 (0.005601)
37. feature 77 (0.005453)
38. feature 25 (0.005225)
39. feature 35 (0.005202)
40. feature 32 (0.005143)
41. feature 85 (0.005102)
42. feature 66 (0.004996)
43. feature 67 (0.004896)
44. feature 94 (0.004894)
45. feature 105 (0.004844)
46. feature 26 (0.004821)
47. feature 71 (0.004800)
48. feature 91 (0.004709)
49. feature 90 (0.004508)
50. feature 79 (0.004343)
51. feature 98 (0.004327)
52. feature 24 (0.004303)
53. feature 64 (0.004122)
54. feature 63 (0.004009)
55. feature 87 (0.003993)
56. feature 93 (0.003624)
57. feature 106 (0.003384)
58. feature 16 (0.003367)
59. feature 143 (0.003299)
60. feature 102 (0.003096)
61. feature 40 (0.002746)
62. feature 114 (0.002712)
63. feature 117 (0.002707)
64. feature 76 (0.002608)
65. feature 161 (0.002586)
66. feature 56 (0.002539)
67. feature 22 (0.002522)
68. feature 99 (0.002518)
69. feature 121 (0.002230)
70. feature 96 (0.002213)
71. feature 82 (0.002195)
72. feature 140 (0.002136)
73. feature 23 (0.002091)
74. feature 88 (0.002060)
75. feature 101 (0.001985)
76. feature 61 (0.001886)
77. feature 113 (0.001883)
78. feature 165 (0.001881)
```

79. feature 38 (0.001878)
80. feature 62 (0.001859)
81. feature 149 (0.001758)
82. feature 130 (0.001690)
83. feature 138 (0.001645)
84. feature 89 (0.001614)
85. feature 157 (0.001587)
86. feature 37 (0.001502)
87. feature 150 (0.001492)
88. feature 111 (0.001400)
89. feature 123 (0.001298)
90. feature 126 (0.001292)
91. feature 21 (0.001287)
92. feature 70 (0.001248)
93. feature 164 (0.001238)
94. feature 148 (0.001203)
95. feature 75 (0.001104)
96. feature 145 (0.001089)
97. feature 167 (0.001063)
98. feature 136 (0.001062)
99. feature 55 (0.001006)
100. feature 81 (0.001001)
101. feature 163 (0.000992)
102. feature 54 (0.000963)
103. feature 12 (0.000949)
104. feature 60 (0.000938)
105. feature 100 (0.000927)
106. feature 124 (0.000895)
107. feature 134 (0.000826)
108. feature 131 (0.000789)
109. feature 153 (0.000724)
110. feature 48 (0.000719)
111. feature 141 (0.000694)
112. feature 112 (0.000693)
113. feature 50 (0.000687)
114. feature 144 (0.000681)
115. feature 156 (0.000631)
116. feature 74 (0.000601)
117. feature 97 (0.000592)
118. feature 151 (0.000589)
119. feature 73 (0.000482)
120. feature 166 (0.000461)
121. feature 119 (0.000460)
122. feature 122 (0.000450)
123. feature 146 (0.000447)
124. feature 41 (0.000426)
125. feature 129 (0.000396)
126. feature 154 (0.000395)

```
127. feature 14 (0.000376)
128. feature 155 (0.000343)
129. feature 132 (0.000341)
130. feature 110 (0.000311)
131. feature 43 (0.000310)
132. feature 116 (0.000301)
133. feature 120 (0.000282)
134. feature 142 (0.000280)
135. feature 137 (0.000279)
136. feature 72 (0.000267)
137. feature 139 (0.000261)
138. feature 160 (0.000257)
139. feature 46 (0.000214)
140. feature 118 (0.000208)
141. feature 45 (0.000186)
142. feature 127 (0.000184)
143. feature 52 (0.000161)
144. feature 51 (0.000144)
145. feature 162 (0.000106)
146. feature 47 (0.000101)
147. feature 133 (0.000101)
148. feature 84 (0.000093)
149. feature 53 (0.000089)
150. feature 147 (0.000088)
151. feature 128 (0.000085)
152. feature 125 (0.000069)
153. feature 159 (0.000063)
154. feature 34 (0.000057)
155. feature 135 (0.000053)
156. feature 49 (0.000051)
157. feature 86 (0.000044)
158. feature 158 (0.000037)
159. feature 39 (0.000020)
160. feature 80 (0.000020)
161. feature 83 (0.000001)
162. feature 65 (0.000001)
163. feature 11 (0.000000)
164. feature 42 (0.000000)
165. feature 36 (0.000000)
166. feature 44 (0.000000)
167. feature 95 (0.000000)
168. feature 78 (0.000000)
```

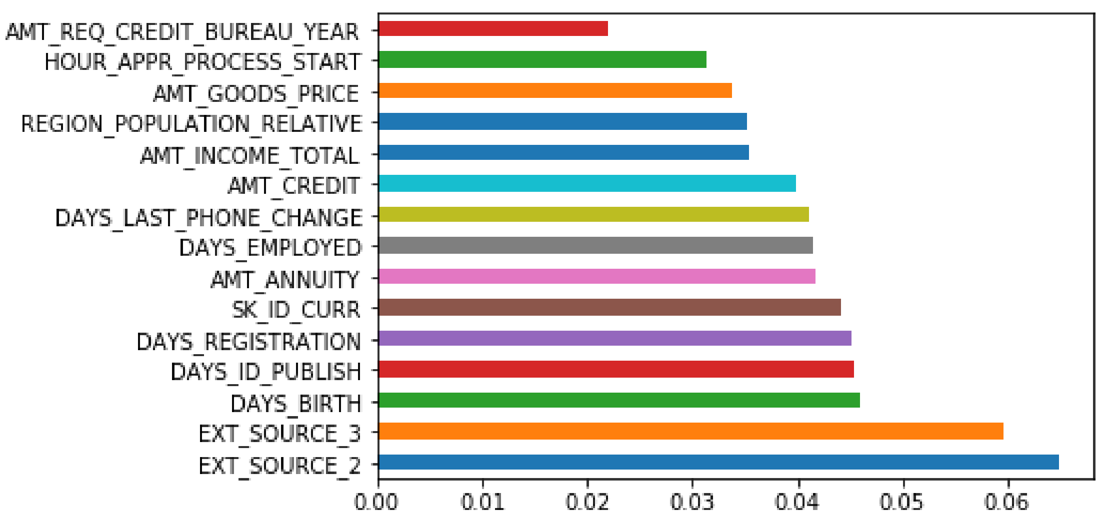
In [53]:
```
# Plot the feature importances of the rf
plt.figure(figsize=(16, 8))
plt.title("Feature importances")
plt.bar(range(app_train.shape[1]), importances[indices], color="r", yerr=std[indices]
```

```
plt.xticks(range(app_train.shape[1]), indices)
plt.xlim([-1, app_train.shape[1]])
plt.show()
```




```
In [54]: (pd.Series(rf.feature_importances_, index=app_train.columns)
            .nlargest(15)
            .plot(kind='barh'))

Out[54]: <matplotlib.axes._subplots.AxesSubplot at 0x7f7b8d800be0>
```



5.4 Random forrest with a cross validation

```
In [55]: rf_cv = RandomForestClassifier()
         scores = cross_val_score(rf_cv, X_train, y_train, cv=5, scoring='roc_auc', n_jobs=-1)
         scores
```

```
Out[55]: array([0.63219173, 0.63231896, 0.62319801, 0.62533242, 0.62668538])
```

```
In [57]: rf_cv.fit(X_train, y_train)
         roc_auc_score(y_test, rf_cv.predict(X_test))
```

```
Out[57]: 0.5036931060869384
```

```
In [58]: #!pip install kaggle
```

```
In [59]: #!kaggle competitions submit -c home-credit-default-risk -f randomforest_baseline.csv
```

---

6 More advanced models

6.1 LightGBM

```
In [62]: lgbm = lgb.LGBMClassifier(random_state = 50)
         lgbm.fit(X_train, y_train, eval_metric = 'auc')
         roc_auc_score(y_train, lgbm.predict(X_train))
```

```
Out[62]: 0.5108945200660795
```

```
In [63]: roc_auc_score(y_test, lgbm.predict(X_test))
```

```
Out[63]: 0.5069964776833696
```

Different tests on hyperparameters and results:

- underfitting / high biais -> let's try to complified the model
- max_depth = 7/11 or objective = 'binary' -> scores 0.508 / 0.508
- n_estimators=1000 -> scores 0.57 / 0.511
- class_weight = 'balanced' -> scores 0.71 / 0.68
- reg_alpha = 0.1, reg_lambda = 0.1 -> no influence

```
In [64]: lgbm = lgb.LGBMClassifier(random_state = 50, n_jobs = -1, class_weight = 'balanced')
         lgbm.fit(X_train, y_train, eval_metric = 'auc')
         roc_auc_score(y_train, lgbm.predict(X_train))
```

```
Out[64]: 0.7121220817520526
```

```
In [65]: roc_auc_score(y_test, lgbm.predict(X_test))
```

```
Out[65]: 0.6846561563080866
```

```
In [66]: def submit_func(model, X_Test, file_name):
             model.fit(app_train, y)
             result = model.predict_proba(app_test)[:, 1]
             submit = app_test[['SK_ID_CURR']]
             submit['TARGET'] = result
             print(submit.head())
             print(submit.shape)
             submit.to_csv(file_name + '.csv', index=False)
```

```
In [67]: submit_func(lgbm, app_test, 'lgbm_submission')
```

```
/home/sunflowa/anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:5: SettingWithCopyWa
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: http://pandas.pydata.org/pandas-docs/stable/indexing.html
  """


   SK_ID_CURR    TARGET
0      100002  0.876184
1      100003  0.237339
2      100004  0.293612
3      100006  0.427514
4      100007  0.634639
(307511, 2)
```

submission -> 0.72057

---

7 Using XGBoost and weighted classes

As said earlier, there are far more 0 than 1 in the target column. This is an [imbalanced class problem].(http://www.chioka.in/class-imbalance-problem/).

It's a common problem affecting ML due to having disproportionate number of class instances in practice. This is why the ROC AUC metric suits our needs here. There are 2 class of approaches out there to deal with this problem:

1) sampling based, that can be broken into three major categories:
    a) over sampling
    b) under sampling
    c) hybrid of oversampling and undersampling.
2) cost function based.

With default or few changes in hyperparameters

- base score : 0.50 / 0.709
- max_delta_step=2 -> unchanged
- with ratio : 0.68 / 0.71

```
In [68]: y.shape[0], y.sum()
```

```
Out[68]: (307511, 24825)
```

```
In [69]: ratio = (y.shape[0] - y.sum()) / y.sum()
         ratio
```

```
Out[69]: 11.387150050352467
```

```
In [70]: xgb_model = xgb.XGBClassifier(objective="binary:logistic", random_state=50, eval_metr
                                       max_delta_step=2, scale_pos_weight=20)
         xgb_model.fit(X_train, y_train)
         roc_auc_score(y_train, xgb_model.predict(X_train))
```

```
Out[70]: 0.656946528564419
```

```
In [71]: roc_auc_score(y_test, xgb_model.predict(X_test))
```

```
Out[71]: 0.6488130660302404
```

For common cases when the dataset is extremely imbalanced, this can affect the training of XGBoost model, and there are two ways to improve it.

If you care only about the overall performance metric (AUC) of your prediction Balance the positive and negative weights via scale_pos_weight Use AUC for evaluation

If you care about predicting the right probability In such a case, you cannot re-balance the dataset Set parameter max_delta_step to a finite number (say 1) to help convergence

```
In [72]: submit_func(xgb_model, app_test, 'xgb_submission')
```

```
/home/sunflowa/anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:5: SettingWithCopyWa
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: http://pandas.pydata.org/pandas-docs/stable/indexing.html
  """


   SK_ID_CURR    TARGET
0      100002  0.908168
1      100003  0.408862
2      100004  0.452180
3      100006  0.607726
4      100007  0.763639
(307511, 2)
```

submission -> 0.72340

8 Credits / side notes

Will Koehrsen for many interesting tips in his kernel !

This notebook is intended to be an introduction to machine learning. So many things are missing or can be done better, such as :

- Using function to clean / prepare the data
- Exploring the other tables and select other columns that can be relevant
- Doing more feature engineering, this will lead to a better score !