Dataset

You are given data from the photometric research of night sky. Each row describes some characteristics of observed object or observation itself (here you can find descriptions of features). Your task is to determine whether the observed object was a galaxy (class 0), a star (class 1) or a quasar (class 2).

In the train.csv you can find 30000 observations from a sky segment. This data is provided for training of your model.
In the val.csv you can find 23334 observations from another sky segment adjoined to the training one. This data is provided for your validation (you are free to share your validation results on this dataset during contest). **Note:** during evaluation this data won't be given to your model for training.
You are also provided with unlabeled.csv. This file contains 23334 data samples from both train and test sky fragments, but you are not given class labels for these observations. You are free to use it in any manner, and it will also be given to your model during assessment.

Note: Test dataset is not given. It will contain 23333 previously unseen samples. It will not contain target column!

Evaluation

Expected output: all source code for research activities (e.g. Jupyter notebooks with EDA) and Python script(s) with pipeline.

To evaluate your solution we ask you to prepare a Python 3 script which can be called from command line in such manner:

```
python3 your_script_name.py [path_to_train] [path_to_unlabeled_data] [path_to_test] [path_to_predictions]
```

Example:

```
python3 panchenko.py data/train.csv data/unlabeled.csv data/test.csv results/panchenko.csv
```

This script should train model, produce predictions and save them to csv file at the given path. Output file should be formatted like this:

```
objid, prediction
42,0
869,2
7,2
365,1
...
```

You model will be evaluated using macro-averaged f1-score.

There are two limitations on the computational properties of your solution:

1. It must be possible to run it on 16 Gb RAM computer.
2. Whole cycle of training and generating predictions must take less than 10 minutes on i5-7500 CPU.

Trivia about data

Some key points regarding data:

1. Train data as well as validation and test have missing values in some columns (*'rowv'*, *'colv'*, all real-valued photometric features). These values are denoted as *na* in the given datasets. Unlabeled data contains missing values only in *'rowv'*, *'colv'*, *'u_3'*, *'g_3'*, *'r_3'*, *'z_3'*, *'i_3'* columns.
2. Validation, test and unlabeled datasets come from the same distribution. Their classes are imbalanced. Train data was artificially balanced and has uniform class distribution.
3. Validation and test datasets were collected from one sky segment whereas training dataset was collected from another part of the sky. You can see a scatter plot of train and test/val observations plotted in *ra* / *dec* coordinates.

Submitting your solution

To share your code, please, create a git repository and give us access to it.
Your solution should also contain Notebooks with exploratory data analysis you made before modeling.