Data Mining Project

Plasticc Astronomical Classification Challenge

Ioannis Prapas

The competition

- Photometric LSST Astronomical Time-series Classification Challenge (PLaStiCC)
- Simulated time-series of flux (light curves) of astronomical objects in 6 different passbands (filters)
- 14 classes in training set, 15 classes in test set
- 1,102 teams/ 1,394 participants

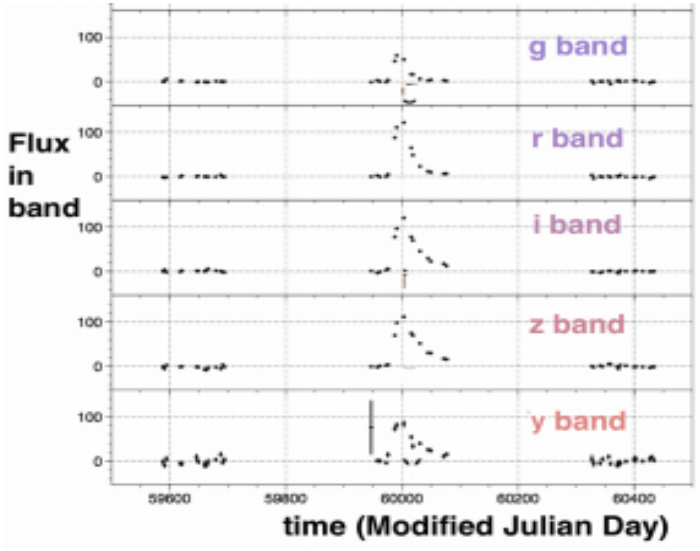
Light curve data

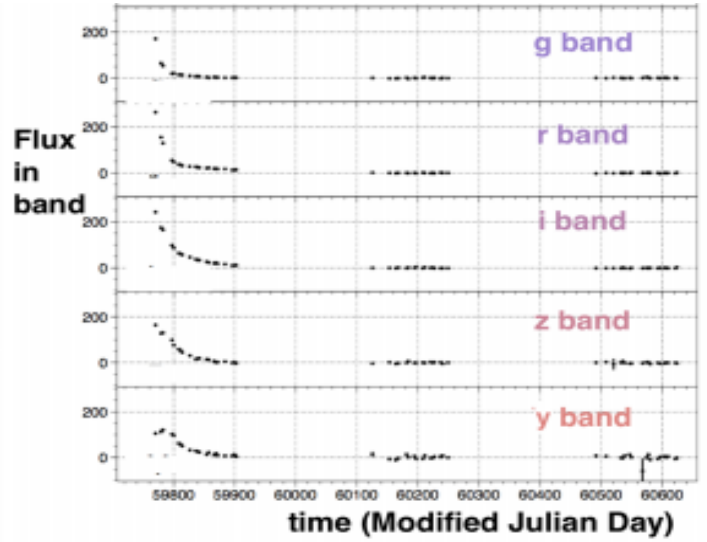
- **object_id**: Primary key of the time series. (Will be used to join with the metadata table)
- **mjd**: the time in Modified Julian Date (MJD) of the observation. The MJD is a float number, representing the number of days from midnight on November 17, 1858.
- passband: The specific LSST passband integer, such that u, g, r, i, z, y = 0, 1, 2, 3, 4, 5 in which it was viewed.
- **flux**: the measured flux (brightness) in the passband of observation as listed in the passband column.
- **flux_err**: the uncertainty on the measurement of the flux
- **detected**: If detected equals 1, the object's brightness is significantly different at the 3σ level relative to the reference template. Otherwise, it is 0.

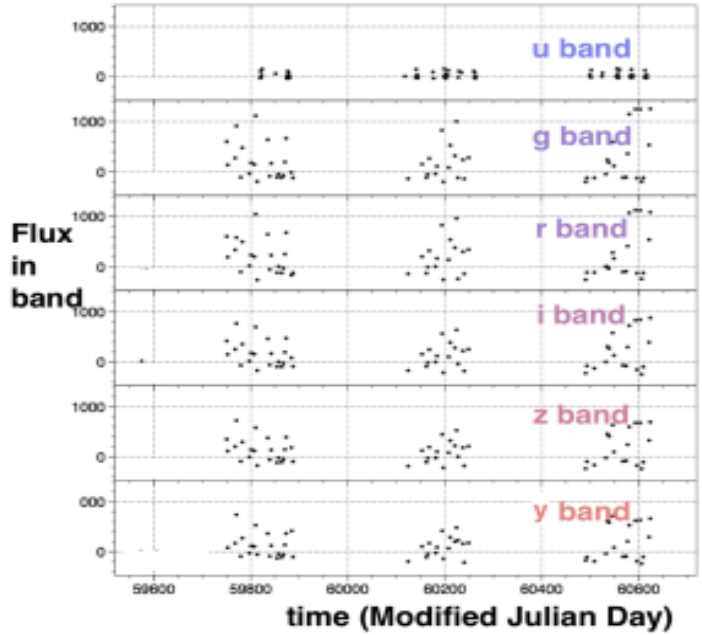
Metadata

- **object_id**: the Object ID, unique identifier (given as int32 numbers).
- **ra**: right ascension, sky coordinate: longitude, in degrees.
- **decl**: declination, sky coordinate: latitude, in degrees.
- **gal l**: Galactic longitude, in degrees.
- **gal b**: Galactic lattitude, in degrees
- **hostgal specz**: Spectrometric redshift
- **hostgal photoz**: Photometric redshift
- **hostgal photoz err**: Photometric redshift error estimation
- **distmod**: Log-distance calculated by photometric redshift
- **mwebv**: milky way dust extinction
- **ddf**: Boolean DDF area or WDF area
- **target**: Target class

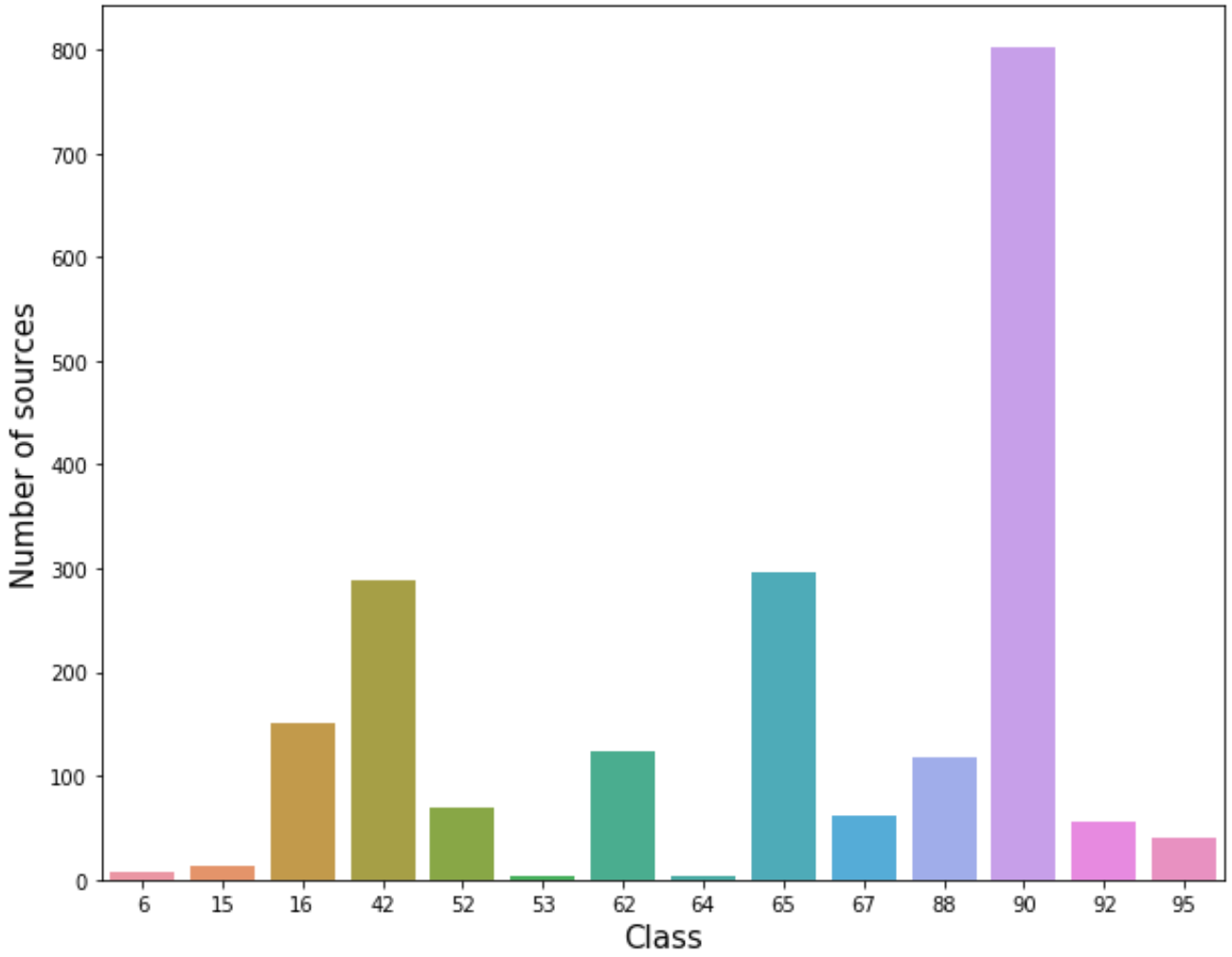
Example


Flux in band
g band
r band
i band
z band
y band
time (Modified Julian Day)


Flux in band
g band
r band
i band
z band
y band
time (Modified Julian Day)


Flux in band
u band
g band
r band
i band
z band
y band
time (Modified Julian Day)

Class imbalance



$$\text{Log Loss} = -\left(\frac{\sum_{i=1}^{M} w_i \cdot \sum_{j=1}^{N_i} \frac{y_{ij}}{N_i} \cdot \ln p_{ij}}{\sum_{i=1}^{M} w_i} \right)$$

Evaluation metric

- The competition uses a **weighted multi-class logarithmic loss**.
- The effect is such that each class is roughly equally important for the final score.

Approach

- Feature engineering
- Smote to account for imbalance
- Train a LightGBM(Gradient Boosting Machine) model

Feature engineering

- Massive test set (3.5m curves) => Incrementally add features
- How?
  - look at the light curves for patterns
  - research for useful features for time-series
  - research for useful features for light-curves
  - kernels and discussions in the Kaggle platform
  - cross-validation score and leaderboard score

Time width features

- *mjd_diff_detected*: Time difference between the last detected flux and the first one. This feature is good to differentiate between periodic and aperiodic events.
- *Mjd_width_max_decay div_{N}*: Time of decay of a light curve from maximum value to N% of maximum

Flux features

- *Slope_after_max{i}*: slope term of linear fit after maximum
- *Slope_before_max{i}*: slope term of linear fit after maximum
- *Intercept_before_max{i}*: intercept term of linear fit before maximum value for passband i
- *Intercept_after_max{i}*: intercept term of linear fit before maximum value for passband i
- *Time-Series Autocorrelation*
- *Fourrier Coefficients*
- Basic statistics per passband and in total: *maximum, minimum, mean, median, skewness, kurtosis*.

Flux / flux_err ratio features

Basic statistics per passband and in total: maximum, minimum, mean, median, skewness, kurtosis.

Color features

Combination of maximum and intercept after max per passband:

```
for i in range(6):
    for j in range(i+1, 6):
        df['{0}{1}__feature'.format(i,j)] = df['{0}__feature'.format(i)] / df['{0}__feature'.format(j)]
```

Absolute Magnitude

Absolute magnitude during maximum flux is a distinguishing term between different types of astronomical objects.

$$M = -2.5 * \log_{10}\left(\frac{F_{max}}{F_0}\right) - distmod$$

https://en.wikipedia.org/wiki/Absolute_magnitude

Training

- 5fold cross validation
- SMOTE on each fold

Predicting

- Average 5classifier predictions trained on 5 folds
- Class_99 (Unknown class): $P_{class_{99}} = \prod(1 - P_{class_i})$

Feature Selection - Unused features

- **ra, decl, gal l, gal b** (positional attributes)
- **hostgal_specz** (spectroscopic redshift – only in few test set examples)

Feature Selection - Importance

- Select N most important
- After having limited amount of features, removed the ones that overfitted the training set and did not generalize to the test set

Team merge

- Merged with Max Halford and Adityasinha
- Blending our predictions put us to 16th position 1 week before the end and we didn’t have time to improve after.

What I learned

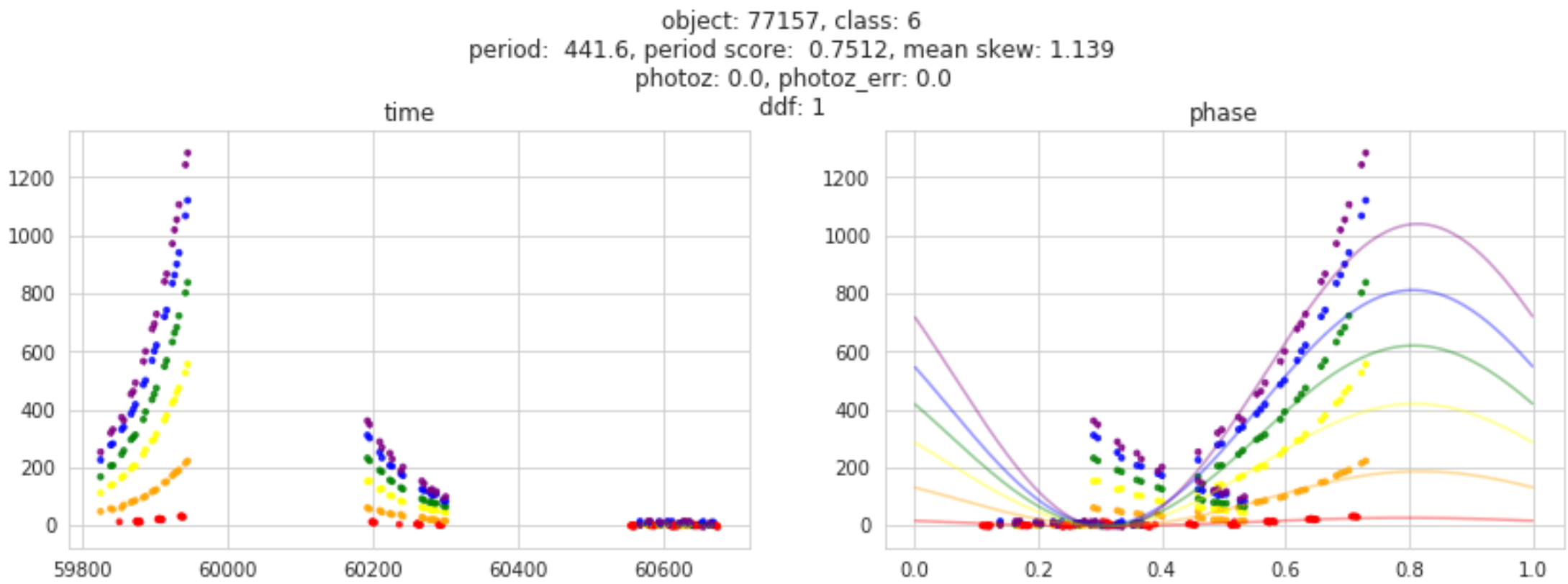
- I learned to use the powerful LGBM, a true hammer for data science.
- I learned different techniques to deal with imbalanced data.
- It was my first time dealing with astronomical or time-series data. Researching about the extraordinary stuff that comprise the universe has been truly interesting.
- It was my first Kaggle competition ever and I competed head-to-head with some of the best data scientists.

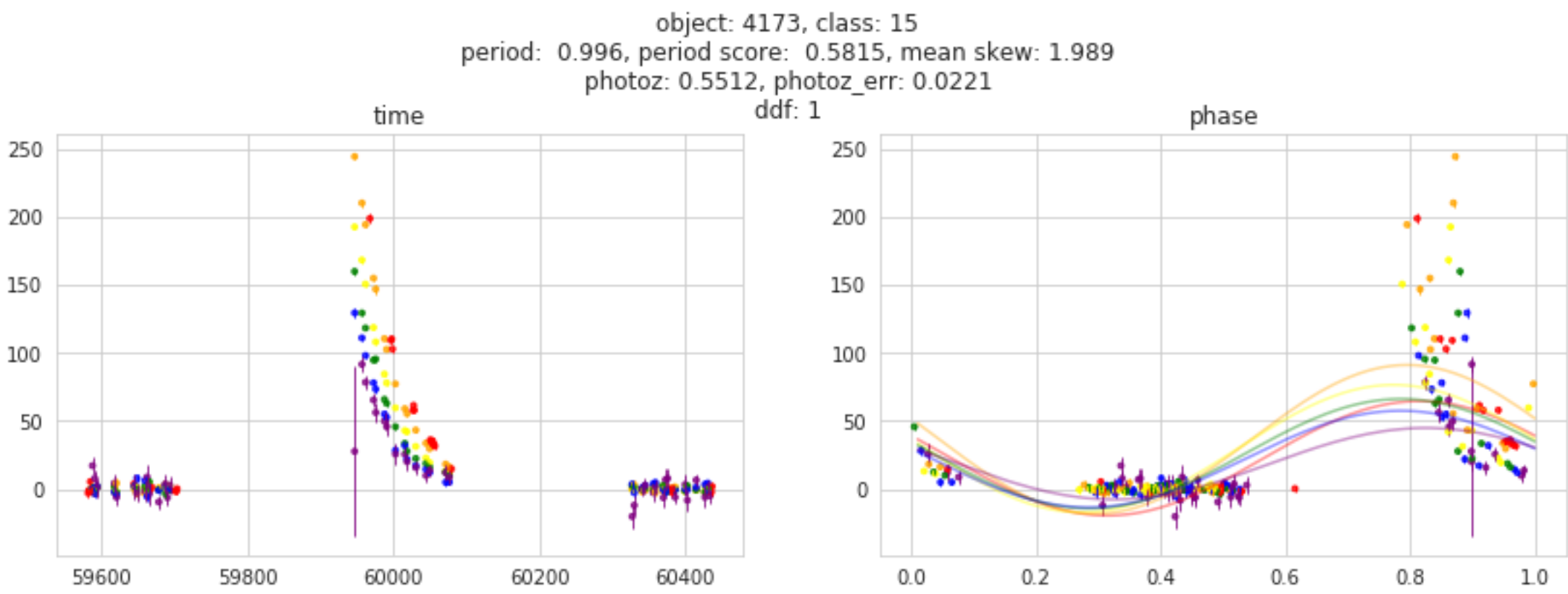
What I earned

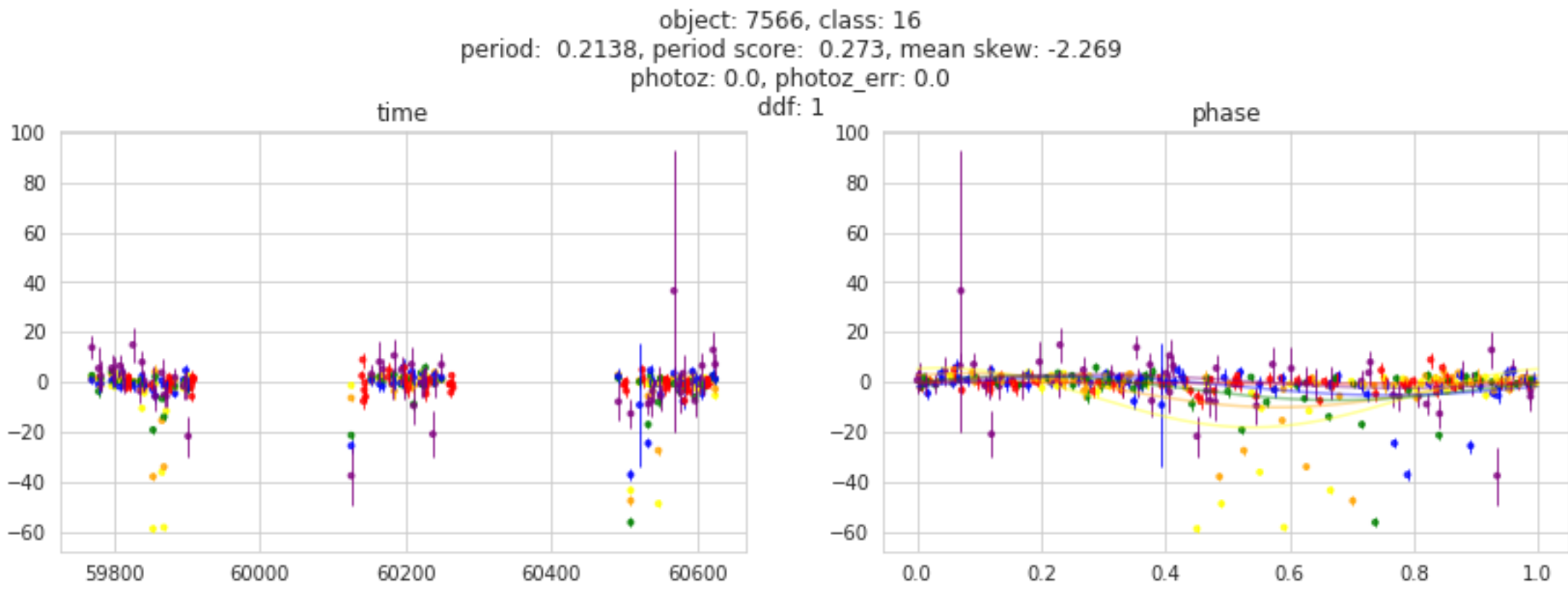
- A **silver medal** for my ranking in the competition (22/ 1,102)
- A **gold medal** for a high-scoring kernel I published which at the time of writing this report has received 102 upvotes and at has been forked almost 400 times.
- **5 silver and 25 bronze medals** for my contributions in the discussions.
- https://www.kaggle.com/iprapas/ideas-from-kernels-and-discussion-lb-1-135

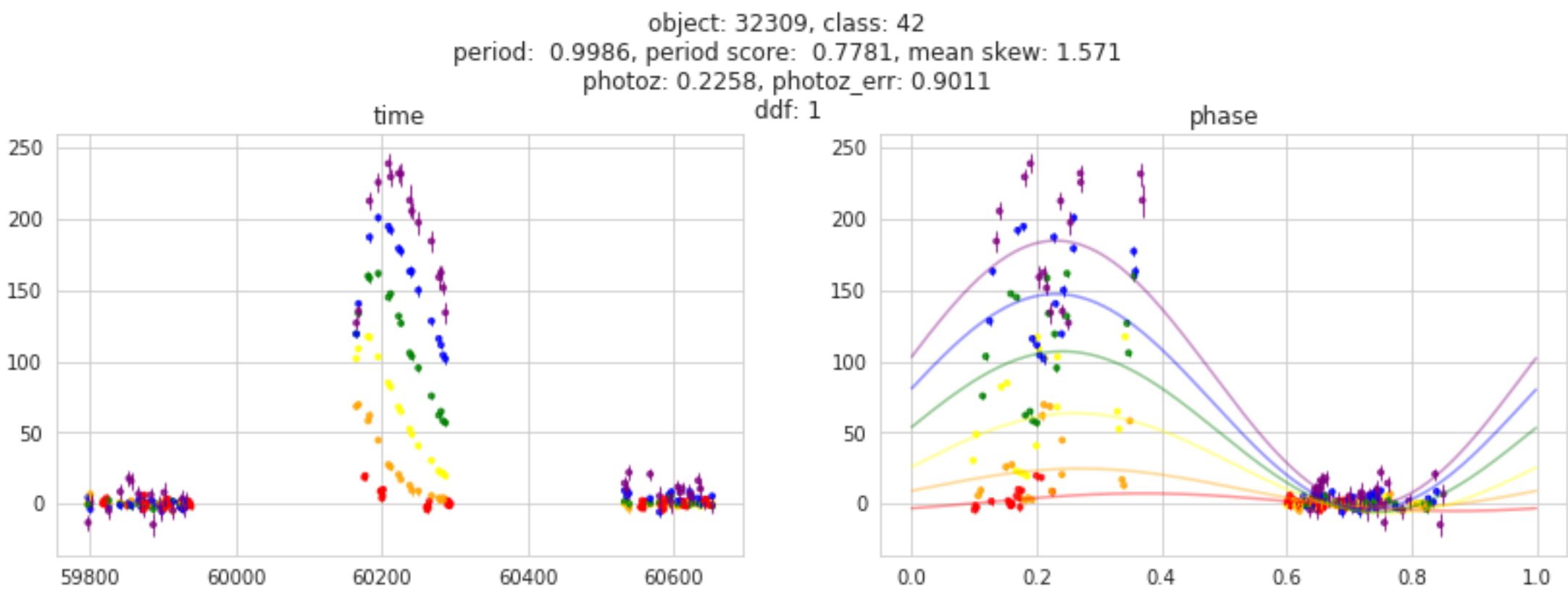
Contribution

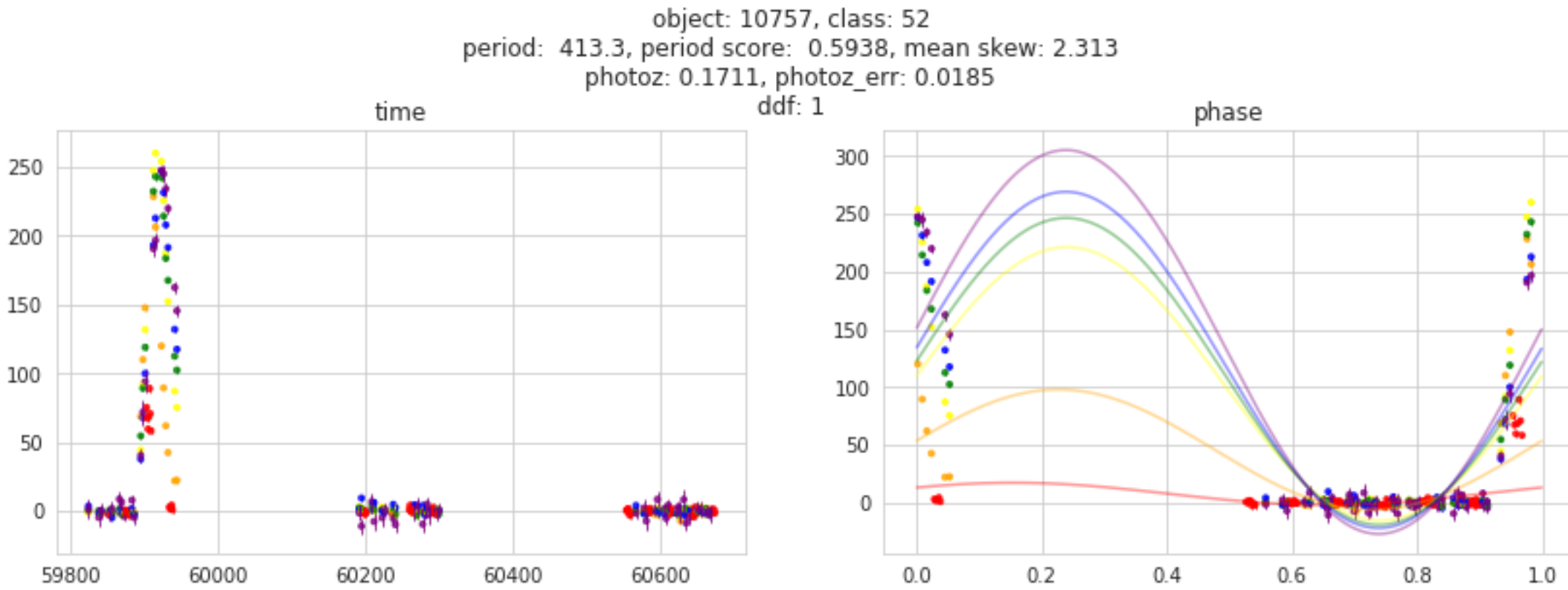
- Gold Kernel (102 upvotes, 400 forks) has pushed for better models
- Active participation in the discussions

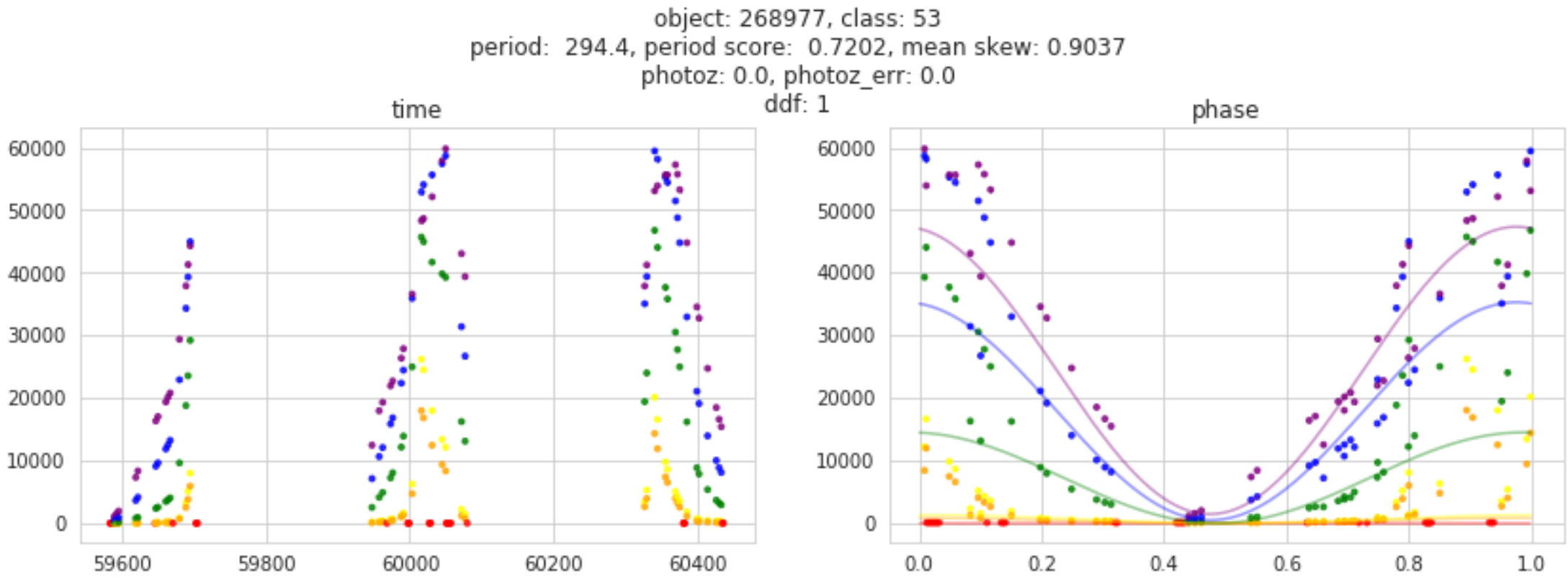

object: 77157, class: 6
period: 441.6, period score: 0.7512, mean skew: 1.139
photoz: 0.0, photoz_err: 0.0
ddf: 1
time
phase
1200
1000
800
600
400
200
0
59800
60000
60200
60400
60600
0.0
0.2
0.4
0.6
0.8
1.0

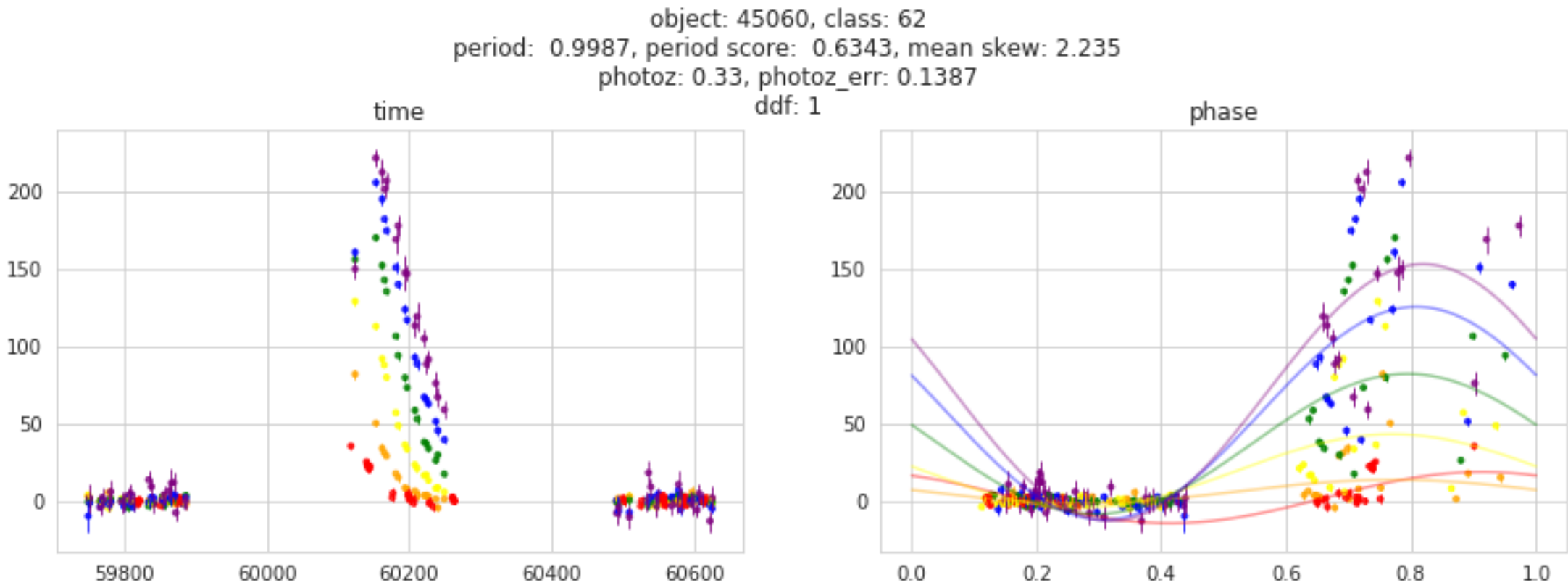

object: 4173, class: 15
period: 0.996, period score: 0.5815, mean skew: 1.989
photoz: 0.5512, photoz_err: 0.0221
ddf: 1
time
phase
250
200
150
100
50
0
59600
59800
60000
60200
60400
0.0
0.2
0.4
0.6
0.8
1.0

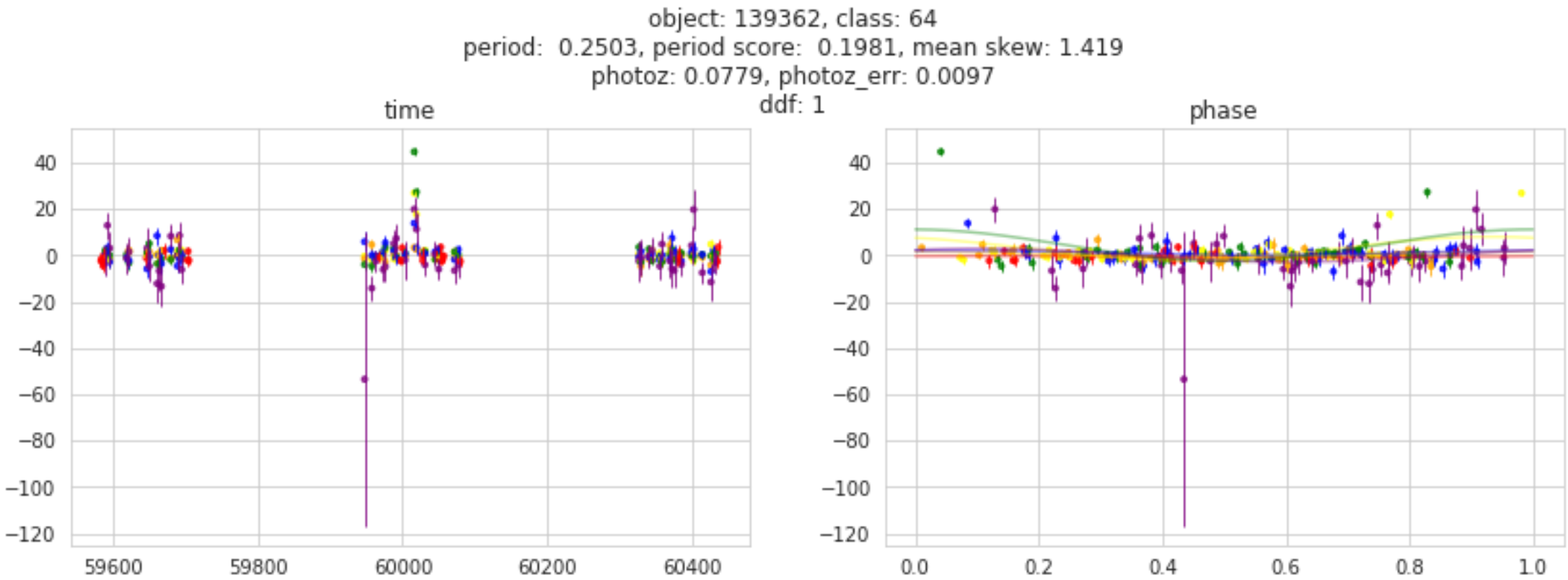

object: 7566, class: 16
period: 0.2138, period score: 0.273, mean skew: -2.269
photoz: 0.0, photoz_err: 0.0
ddf: 1
time
phase
100
80
60
40
20
0
−20
−40
−60
59800
60000
60200
60400
60600
0.0
0.2
0.4
0.6
0.8
1.0

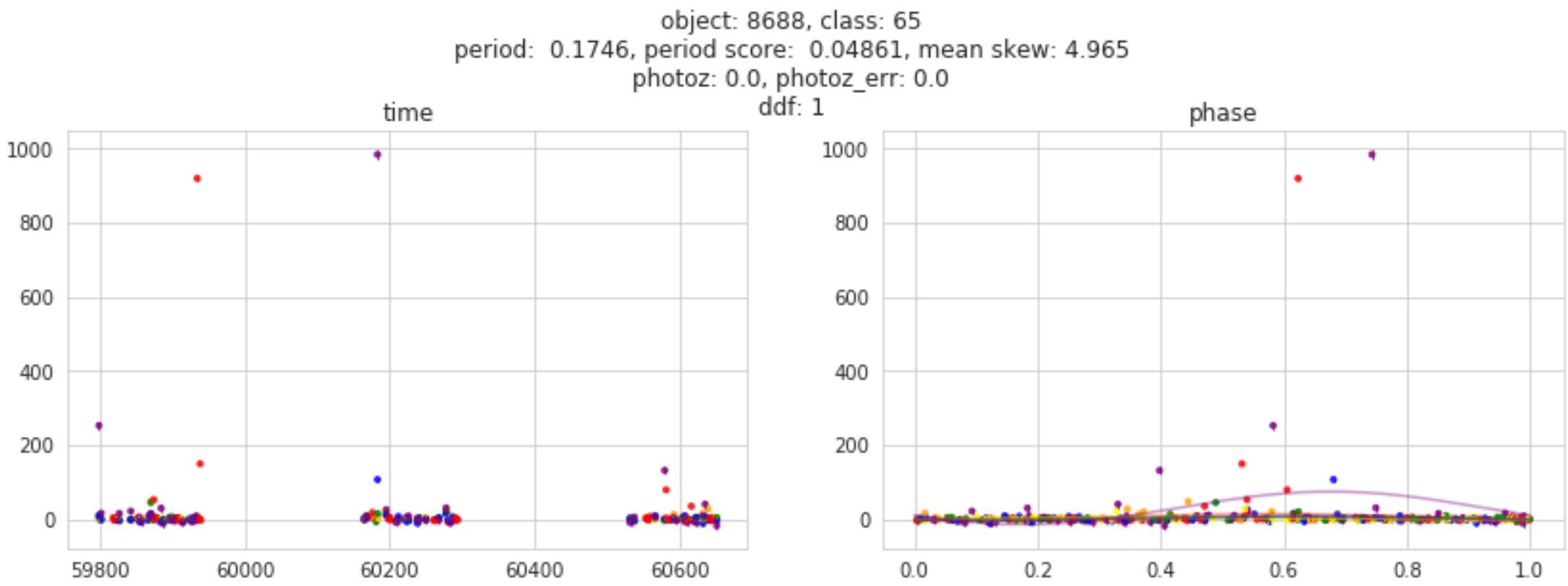

object: 32309, class: 42
period: 0.9986, period score: 0.7781, mean skew: 1.571
photoz: 0.2258, photoz_err: 0.9011
ddf: 1
time
phase
250
200
150
100
50
0
59800
60000
60200
60400
60600
0.0
0.2
0.4
0.6
0.8
1.0

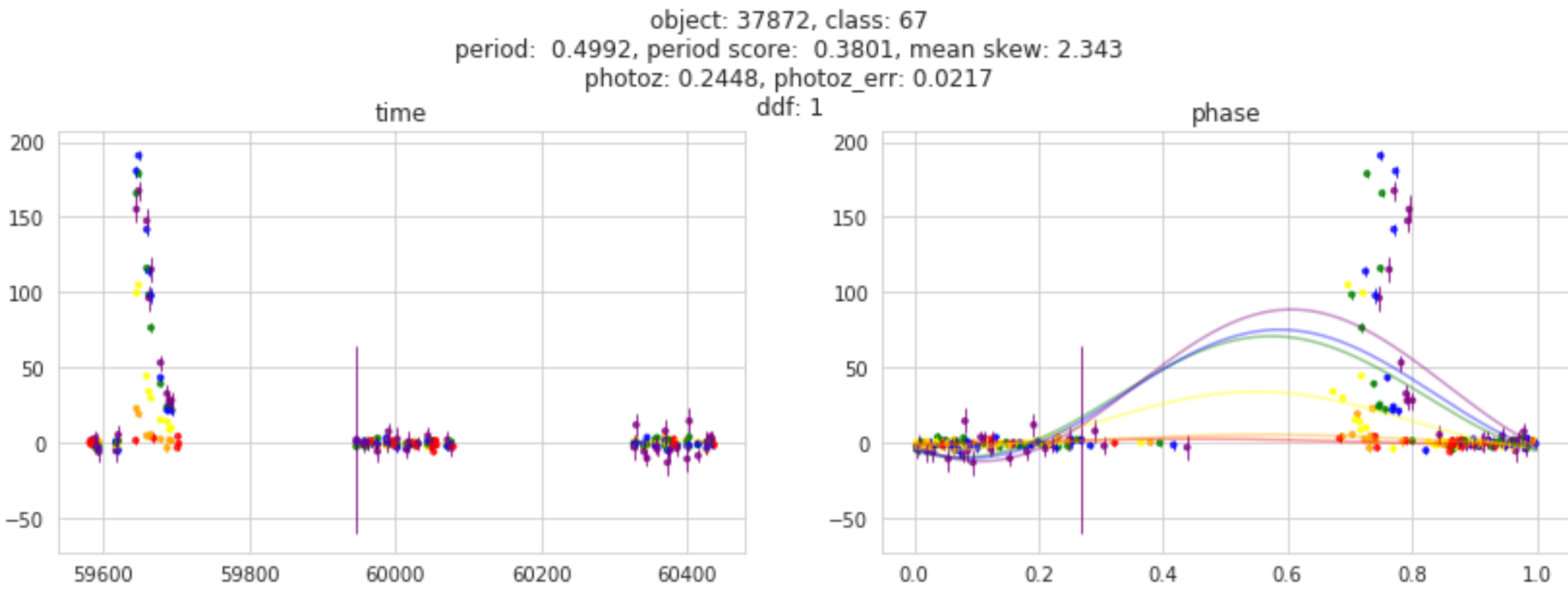

object: 10757, class: 52
period: 413.3, period score: 0.5938, mean skew: 2.313
photoz: 0.1711, photoz_err: 0.0185
ddf: 1
time
phase
250
200
150
100
50
0
59800
60000
60200
60400
60600
300
250
200
150
100
50
0
0.0
0.2
0.4
0.6
0.8
1.0

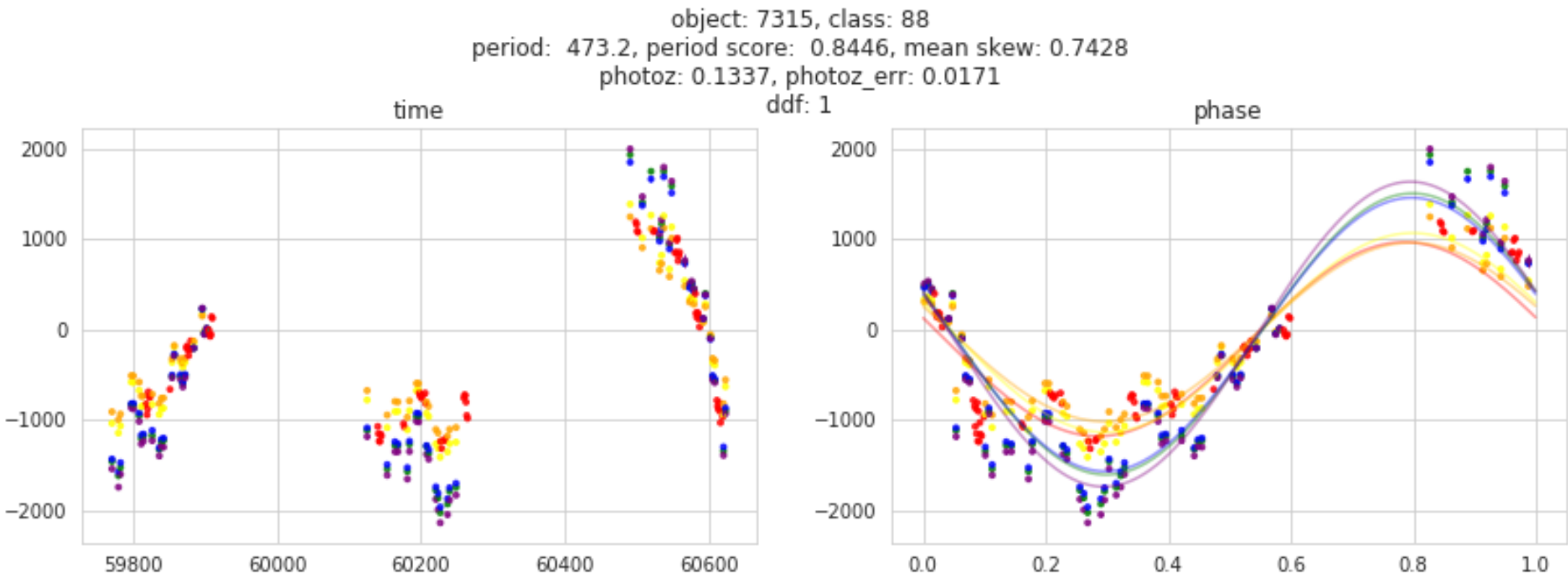

object: 268977, class: 53
period: 294.4, period score: 0.7202, mean skew: 0.9037
photoz: 0.0, photoz_err: 0.0
ddf: 1
time
phase
60000
50000
40000
30000
20000
10000
0
59600
59800
60000
60200
60400
0.0
0.2
0.4
0.6
0.8
1.0

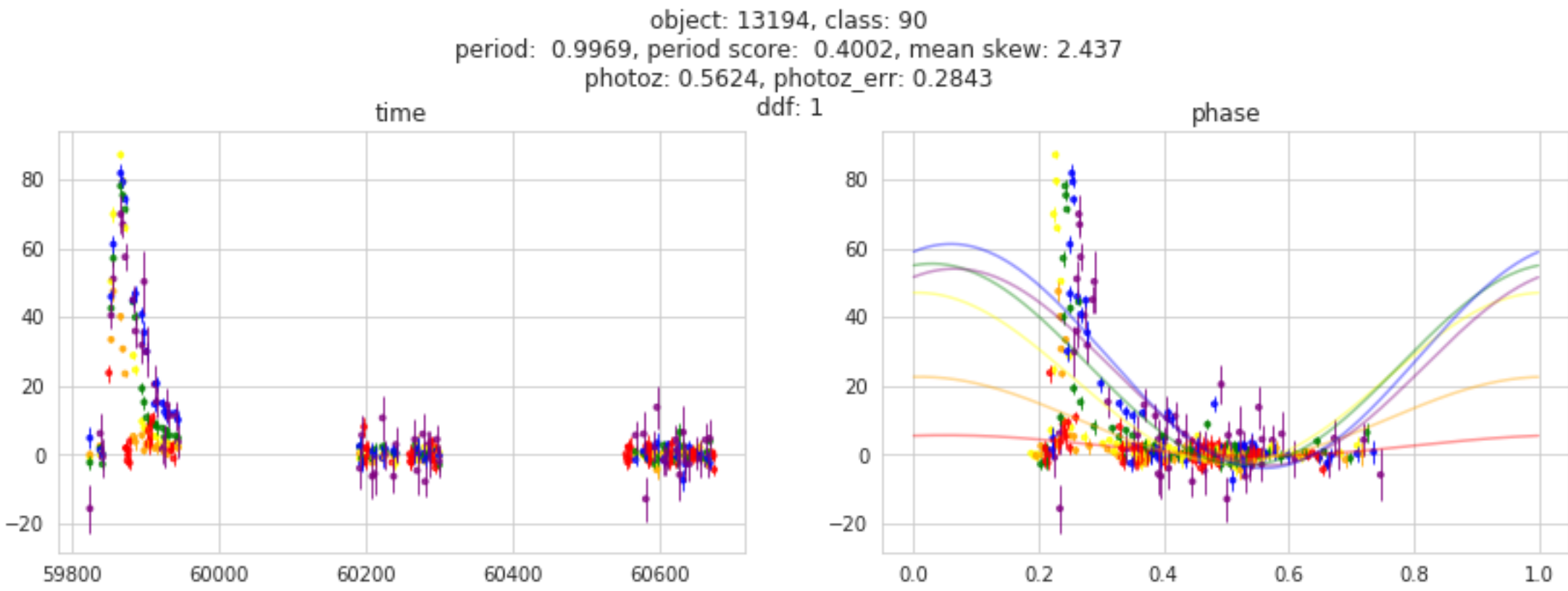

object: 45060, class: 62
period: 0.9987, period score: 0.6343, mean skew: 2.235
photoz: 0.33, photoz_err: 0.1387
ddf: 1
time
phase
200
150
100
50
0
59800
60000
60200
60400
60600
0.0
0.2
0.4
0.6
0.8
1.0

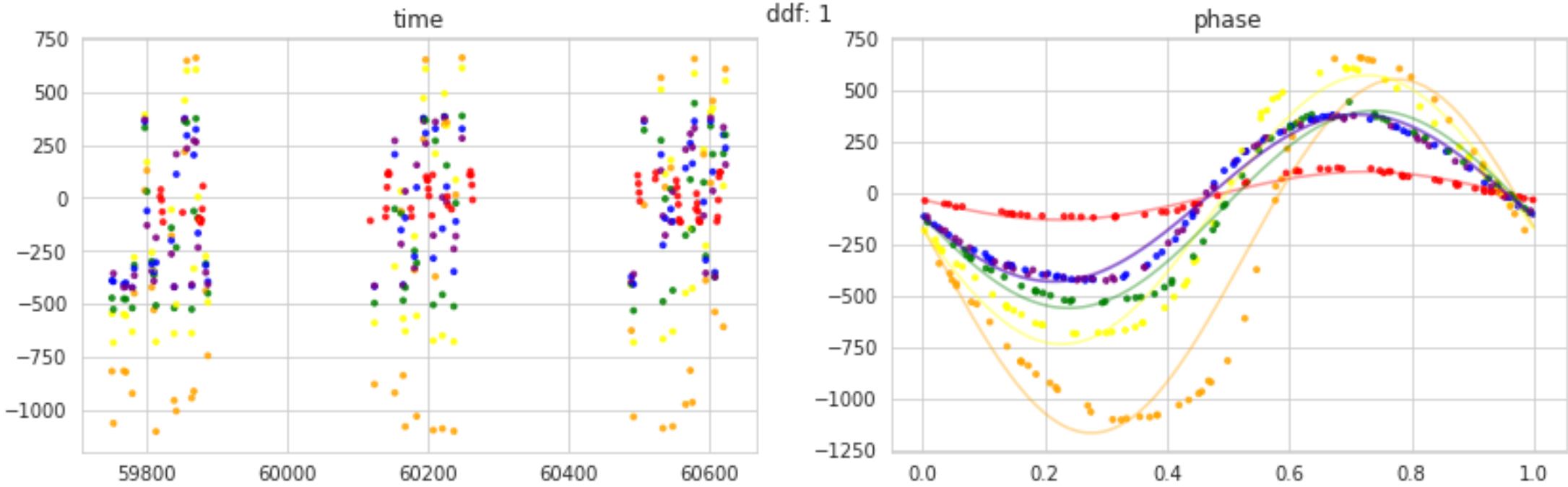

object: 139362, class: 64
period: 0.2503, period score: 0.1981, mean skew: 1.419
photoz: 0.0779, photoz_err: 0.0097
ddf: 1
time
phase
40
20
0
−20
−40
−60
−80
−100
−120
59600
59800
60000
60200
60400
0.0
0.2
0.4
0.6
0.8
1.0

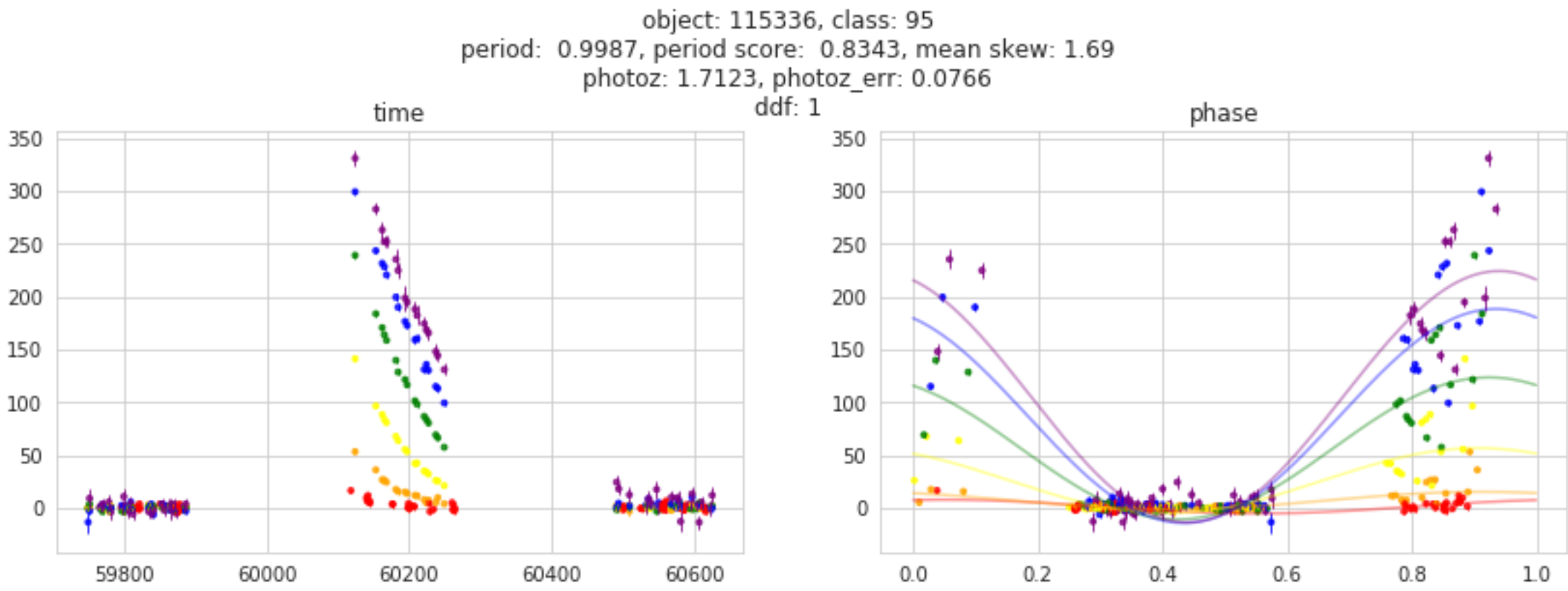

object: 8688, class: 65
period: 0.1746, period score: 0.04861, mean skew: 4.965
photoz: 0.0, photoz_err: 0.0
ddf: 1
time
phase
1000
800
600
400
200
0
59800
60000
60200
60400
60600
0.0
0.2
0.4
0.6
0.8
1.0


object: 37872, class: 67
period: 0.4992, period score: 0.3801, mean skew: 2.343
photoz: 0.2448, photoz_err: 0.0217
ddf: 1
time
phase
200
150
100
50
0
−50
59600
59800
60000
60200
60400
0.0
0.2
0.4
0.6
0.8
1.0


object: 7315, class: 88
period: 473.2, period score: 0.8446, mean skew: 0.7428
photoz: 0.1337, photoz_err: 0.0171
ddf: 1
time
phase
2000
1000
0
−1000
−2000
59800
60000
60200
60400
60600
0.0
0.2
0.4
0.6
0.8
1.0


object: 13194, class: 90
period: 0.9969, period score: 0.4002, mean skew: 2.437
photoz: 0.5624, photoz_err: 0.2843
ddf: 1
time
phase
80
60
40
20
0
−20
59800
60000
60200
60400
60600
0.0
0.2
0.4
0.6
0.8
1.0


object: 615, class: 92
period: 0.3245, period score: 0.954, mean skew: 0.244
photoz: 0.0, photoz_err: 0.0
ddf: 1
time
phase
750
500
250
0
−250
−500
−750
−1000
−1250
59800
60000
60200
60400
60600
0.0
0.2
0.4
0.6
0.8
1.0


object: 115336, class: 95
period: 0.9987, period score: 0.8343, mean skew: 1.69
photoz: 1.7123, photoz_err: 0.0766
ddf: 1
time
phase
350
300
250
200
150
100
50
0
59800
60000
60200
60400
60600
0.0
0.2
0.4
0.6
0.8
1.0