



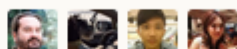



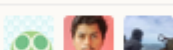
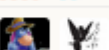








Сергей Злобин

Классификация космических объектов (Kaggle PLAsTiCC Astronomical Classification)

Leaderboard

1	—	Kyle Boone		0.68503	104	2d
2	▲2	Mike & Silogram		0.69933	176	1d
3	▼1	Major Tom		0.70016	366	1d
4	▼1	AhmetErdem		0.70423	233	1d
5	—	SKZ Lost in Translation		0.75229	343	2d
6	▲2	Stefan Stefanov		0.80173	28	1d
7	▲3	hkleee		0.80836	63	7d
8	▼1	rapids.ai		0.80905	133	1d
9	▼3	Three Musketeers		0.81312	313	1d
10	▲3	J&J		0.81901	246	1d
11	▼2	SimonChen		0.82247	131	1d
12	▼1	Go Spartans!		0.82652	148	1d
13	▼1	Day meets Night		0.82691	164	1d
14	▲6	Belinda Trotta		0.84070	105	2d

Команда



Mithrillion (Питер, Австралия)



Blonde (Таня, Ирландия)



Sergey Zlobin (Россия)

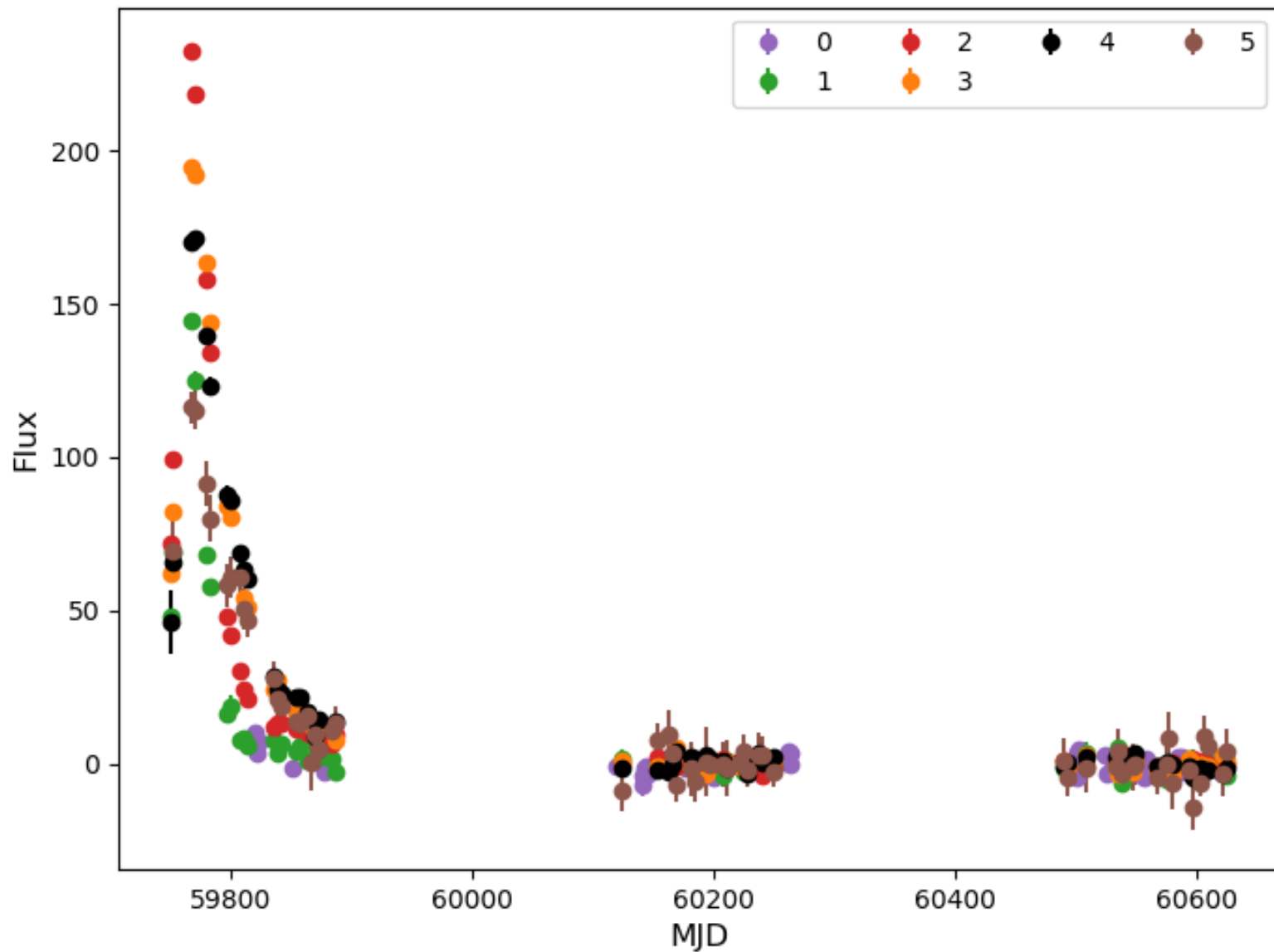
Немного о задаче

LSST = Large Synoptic Survey Telescope (Чили)

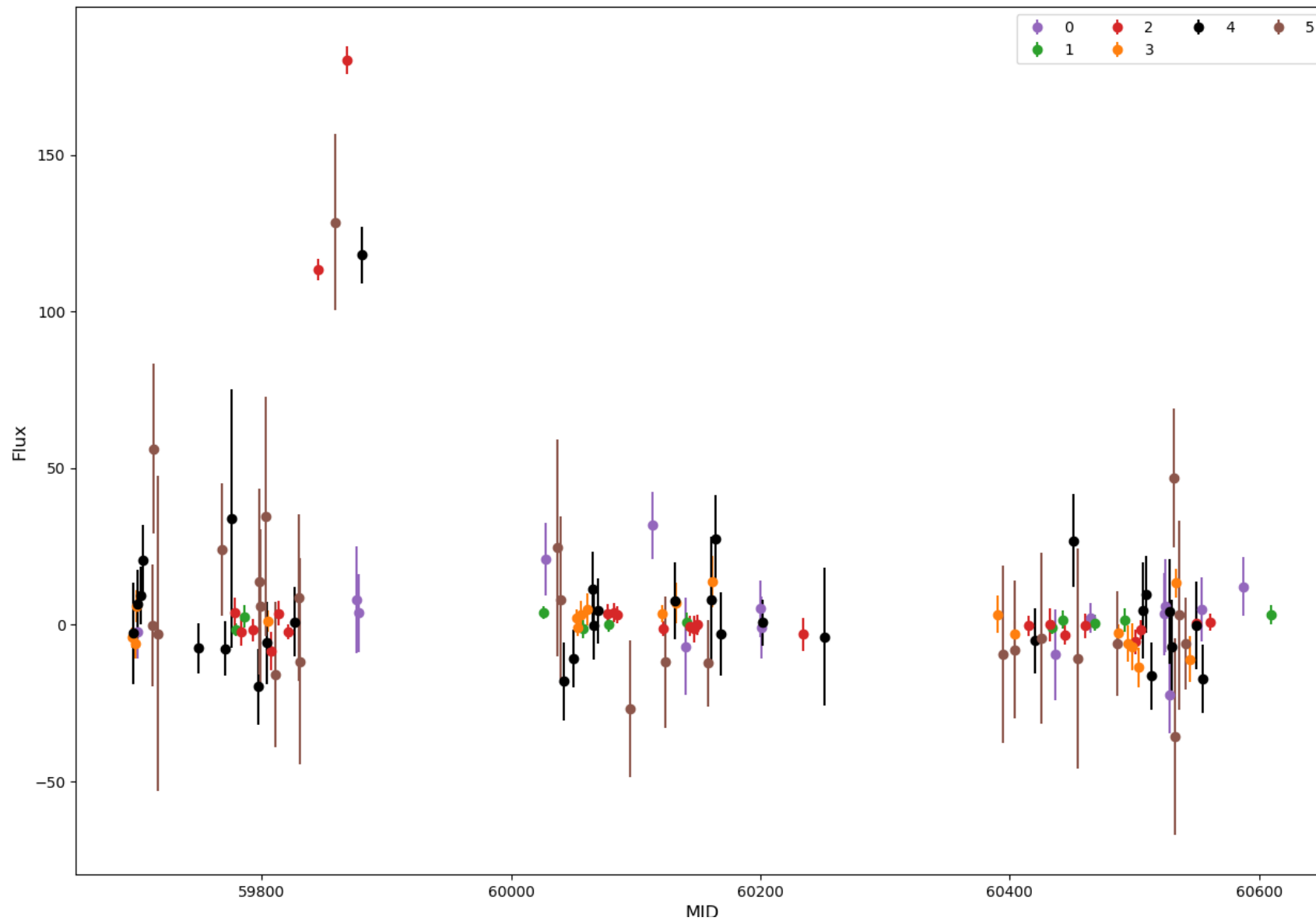


20-40 терабайт
в сутки!

Данные



Non-DDF (Wide-Fast-Deep)



Метрика соревнования

weighted log-loss metric ($N=15$)

$$L = - \frac{\sum_{j=1}^M w_j \cdot \sum_{i=1}^N \frac{1}{N_j} \tau_{i,j} \ln(P_{ij})}{\sum_{j=1}^M w_j}$$

where $\tau_{i,j} = 1$ if the i th object comes from the j th class and 0 otherwise, and N_j is the number of objects in any given class j , and w_j are individual weights per class which reflect relative contribution to the overall metric

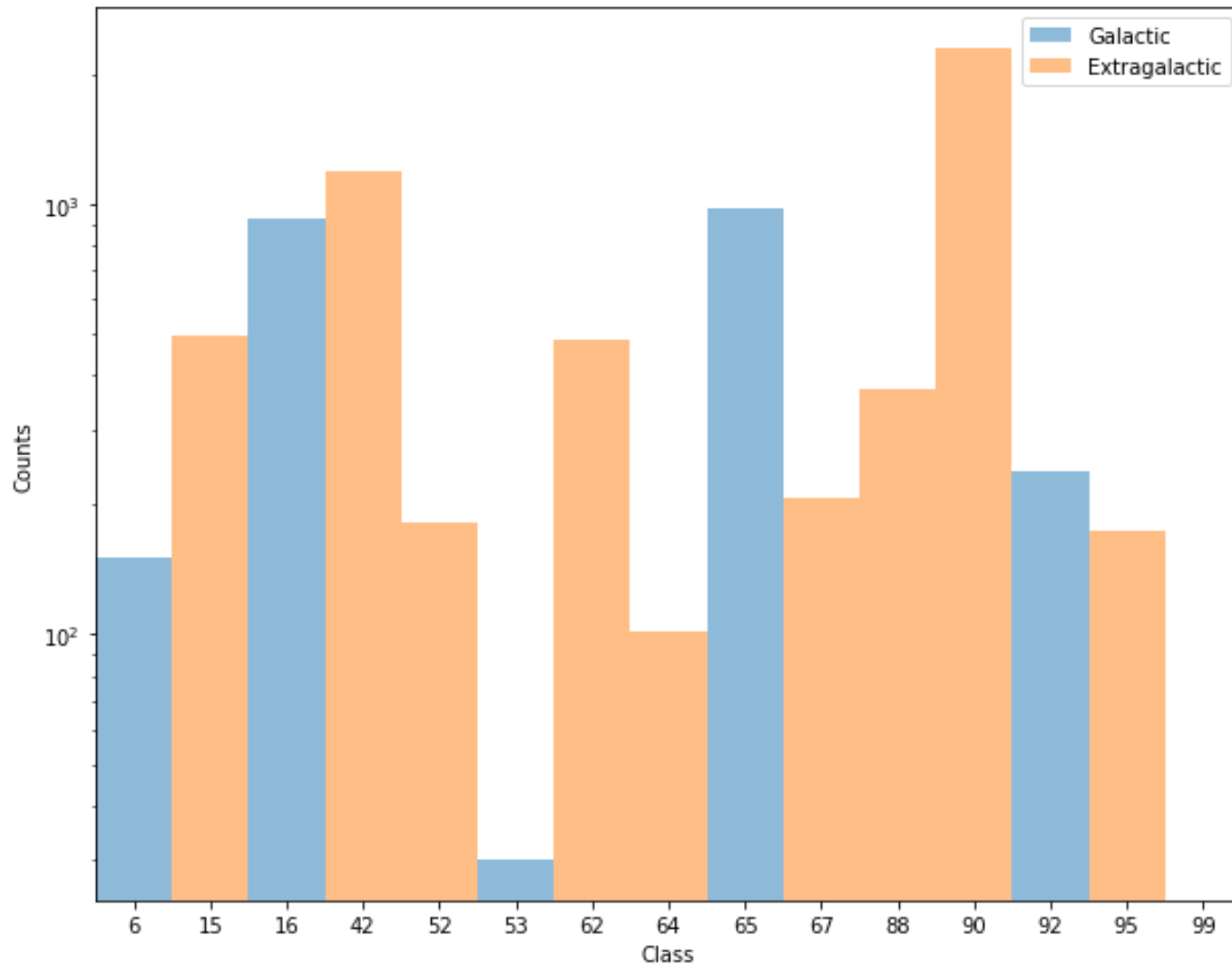
Веса w не опубликованы, но фактически из $\{1, 2\}$.

Выборки

- Обучающая выборка ~8.000 объектов
- Тестовая выборка ~3.500.000 объектов (19 Gb)

Подсчёт признаков и обучение модулей запускались
на 24 CPU или с GPU

Распределение классов



Среди классов:

- Сверхновые звёзды
- Переменные звёзды
- Микролинзирование
- ???

Зануление неподходящих классов
дает улучшение ~ 0.006 на LB

Модели

- 1) Light GBM
- 2) Multilayer perceptron (4 скрытых слоя) -> в конце отказались

5-fold & 10-fold

Лучший score одной Light GBM модели: 0.815

Конечная модель: 0.812 (blend двух Light GBM моделей)

Первое место: 0.670

Признаки

Признаки можно считать по каждому из 6 каналов и агрегированные.

- Стандартные (min, max, median, mean, std, skew)
- Библиотечные (Cesium, Feets)
- Подгонка (fitting) кривой (признаки – параметры и ошибка аппроксимации)
- Специальные астрономические (magnitude=звёздная величина)

Отбор признаков: LGBM Importance, eli5 (Permutation Importance) + CV

Конечная модель: 81 признак

Domain Knowledge!

Фича от Grzegorz Sionkowski:

Расстояние между надёжно детектированными сигналами

Абсолютная звёздная величина (magnitude):

$-2.5 * \text{math.log}_{10}(\text{flux}) - \text{distmod}$

distmod в метаданных: Distance modulus (Модуль расстояния)

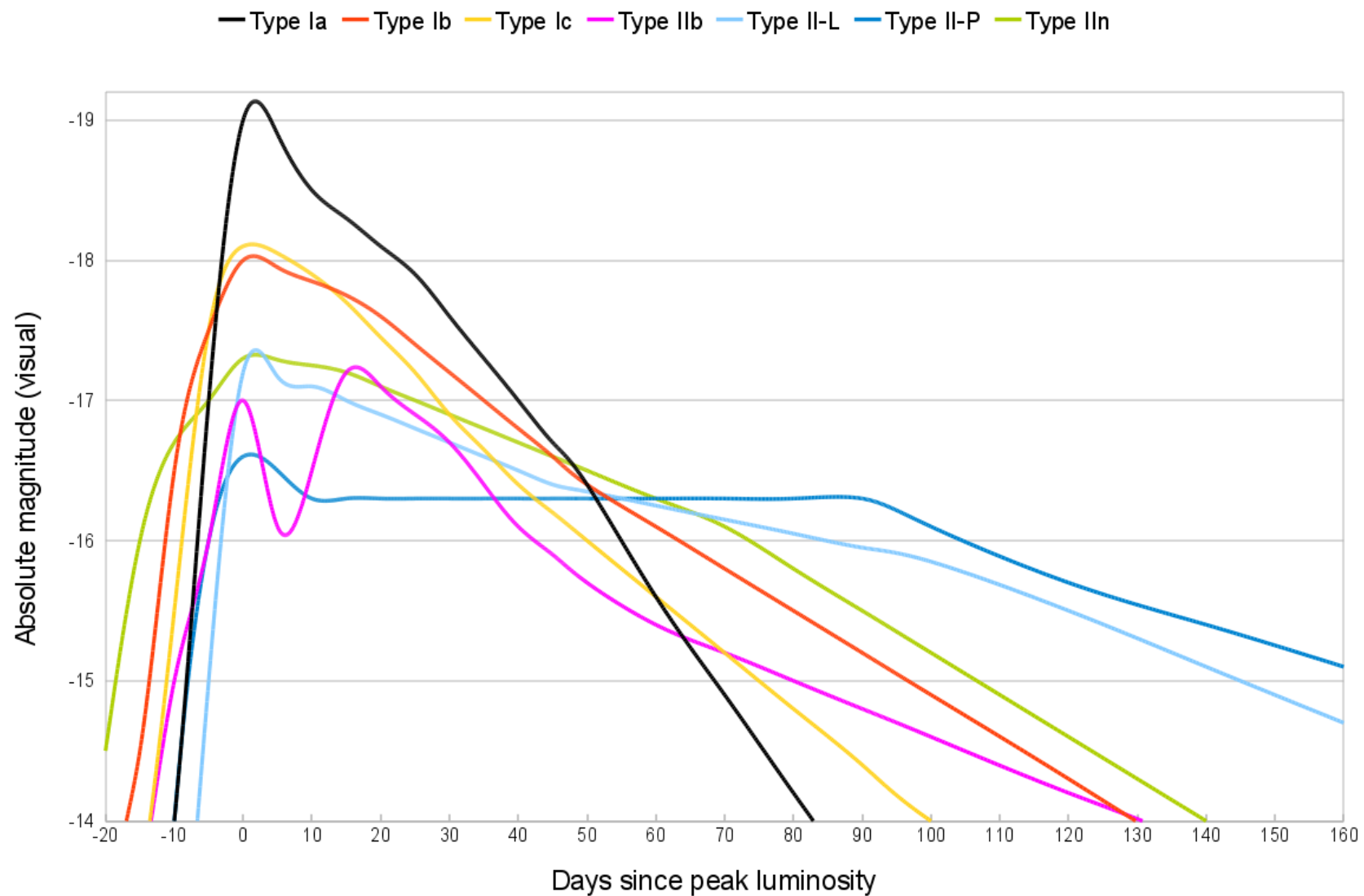
(яркость обратно пропорциональна квадрату расстояния)

Астрономические цвета

Разность магнитуд между некоторыми каналами

(например, между зеленым и красным)

Типы сверхновых



Curve Fitting

- Гауссовское распределение - оценивает ширину пика
- Кусочно-линейная функция для $\log(\text{flux})$ – оценивает наклон
- Bazin et al.:

$$f(t) = A \frac{e^{-(t-t_0)/\tau_{fall}}}{1 + e^{(t-t_0)/\tau_{rise}}} + B$$

Пробовали другие кривые, в том числе с двумя пиками – не помогло.

Класс 99

Формула от Oliver:

```
preds_99 = np.ones(preds.shape[0])  
for i in range(preds.shape[1]):  
    preds_99 *= (1 - preds[:, i])  
preds_99 = 0.18 * preds_99 / np.mean(preds_99)
```

Класс 99

Формула от Scirpus:

```
mymean = np.mean(preds, axis=1)
```

```
mymedian = np.median(preds, axis=1)
```

```
mymax = np.max(preds, axis=1)
```

```
preds_99 = (((((((mymedian) + (((mymean) / 2.0)))/2.0)) +  
              ((((((1.0) - (((mymax) * (((mymax) * (mymax)))))))) / 2.0)))/2.0)
```

```
(0.5 + 0.5 * mymedian + 0.25 * mymean - 0.5 * mymax ** 3) / 2
```


Класс 99

Победитель (Kyle Boone):

weighted average of the predictions for classes 42, 52, 62 and 95
(сверхновые)

LB 0.726 -> 0.670

Аугментация

Меняем сигналы с помощью известной ошибки измерения:

```
for i in range(N_SETS):  
    per_observation_perturbation = np.random.randn(series.shape[0])  
    augmented_series = series.copy()  
    augmented_series['flux'] += per_observation_perturbation *  
                               augmented_series['flux_err']
```

N_SETS = 10, потом 30

Борьба с переобучением

В какой-то момент застряли на LB 0.85.

Аугментация

Увеличили количество обучающей выборки с $x10$ до $x30$, но сделав процент DDF объектов таким же, как в тестовой выборке (около 1% вместо 10%).

Параметры Light GBM:

- Уменьшили `max_depth` (7 -> 4)
- Уменьшили `num_leaves` (7 -> 4)
- Уменьшили `max_bin` (255 -> 32)

CV чуть хуже, но LB намного лучше!

Что не получилось

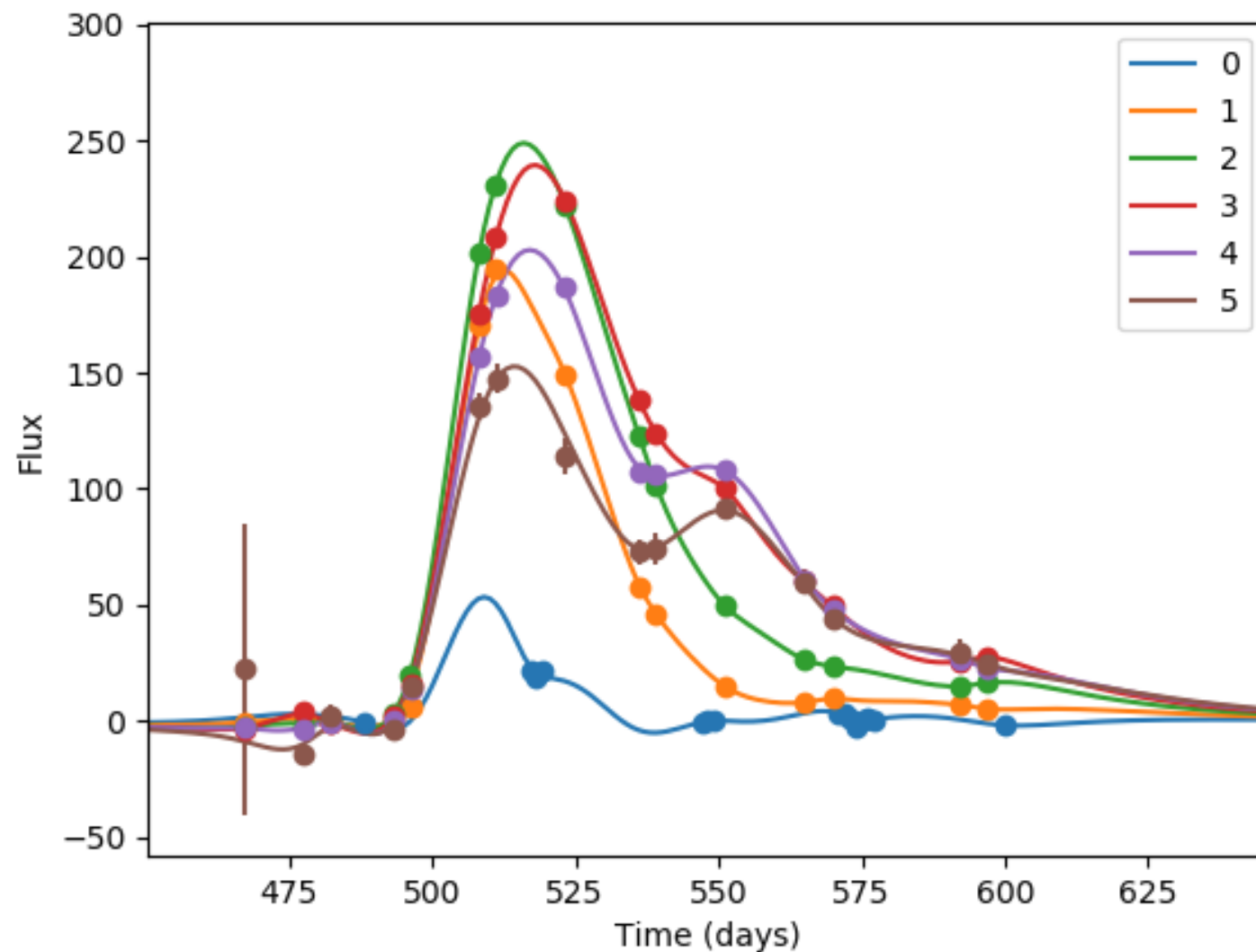
- Автоэнкодер не улучшил результат
- Не зашли другие виды параметрических кривых
- Не получилось воспользоваться гауссовскими процессами (для аугментации)
- Не смогли придумать как лучше оценивать класс 99 (в частности, пробовали PU классификацию)

Что сделали правильно

- Объединились в команду!
- Читали статьи на тему соревнования.
- Задавали вопросы на форуме Kaggle и в ODS. Нам отвечали!
- Много чего попробовали.
- Не сдаваться до конца: сильно улучшились за последние 3 дня и даже в последний день!

Победитель (Kyle Boone)

- Аугментация x40 обучающей выборки, ухудшая хорошие кривые, чтобы стало похоже на тестовую выборку.
- Гауссовские процессы (GP) для предсказания кривых.
- Около 200 признаков на сырых и предсказанных после GP данных.
- Одна LGBM модель на 5 фолдах





Спасибо за внимание!