
Computer
Science
Center

PLAsTiCC-2018 at Kaggle

Александр Авдюшенко

24 декабря 2018

Задача

Классификация астрономических объектов на 15 классов, одного из которых нет в train (class_99 = неизвестные науке объекты)

1,094 Teams | 1,384 Competitors | 22,895 Entries

Points This competition awarded standard ranking points

Tiers This competition counted towards tiers

Tags: time series | tabular data | astronomy | weightedmulticlassloss | extra large

https://www.kaggle.com/c/PLAsTiCC-2018

Dataset

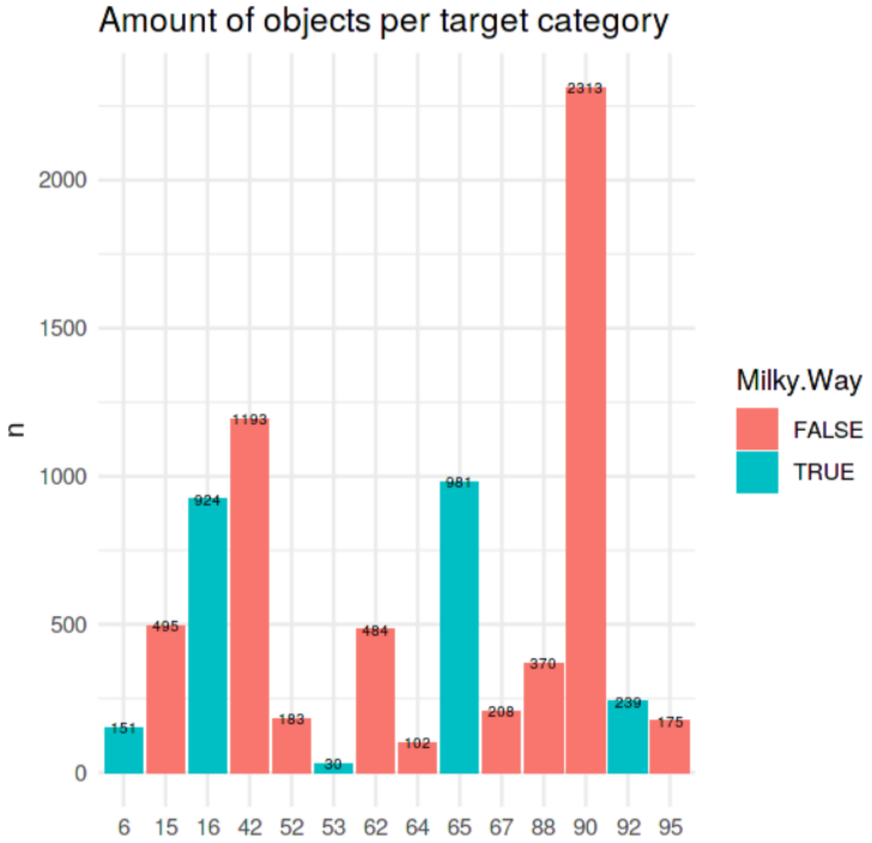
Количество строк
7.8K – training_set_metadata
1.4M – training_set

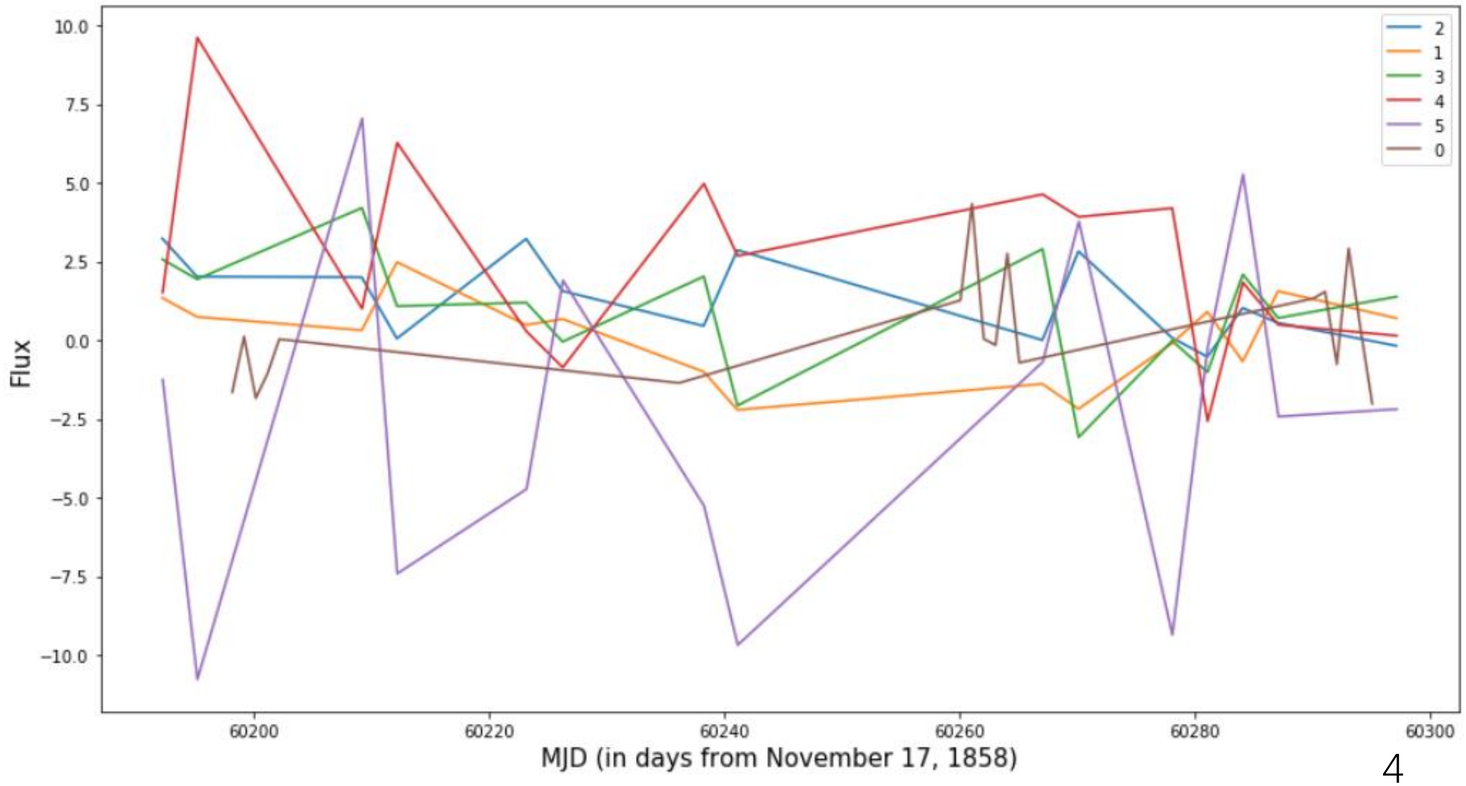
3.5M – test_set_metadata
454M – test_set

Обучение 20-30 м

Применение 3-4 ч

Metadata
object_id, координаты,
redshift, distmod, **target**
Data
unix_time, flux (brightness) x6
flux_err, detected





#169 solution – @ovalur

54	▲ 49	Aleksandr Avdyushenko		1.755	10	36m

Your Best Entry ↑

Your submission scored 1.755, which is an improvement of your previous score of 2.158. Great job!

Tweet this!

	private	public	
xgb_lgb_2_blend.csv a month ago by Aleksandr Avdyushenko xgb_lgb_2_blend	1.12522	1.09950	☐
single_subm_0.656146_2018-11-21-09-33.csv a month ago by Aleksandr Avdyushenko lgb_2_sum_ideas	1.10791	1.08183	☐
single_subm_0.904055_2018-11-20-05-55.csv a month ago by Aleksandr Avdyushenko lgb_sum_ideas	1.41797	1.38728	☐
single_subm_0.684427_2018-11-20-04-15.csv a month ago by Aleksandr Avdyushenko xgb_sum_ideas	1.15567	1.13203	☐
single_predictions.csv a month ago by Aleksandr Avdyushenko lgb_with_time_series	2.06563	1.98198	☐
single_predictions.csv 2 months ago by Aleksandr Avdyushenko lgb_full	1.79270	1.75516	☐

	private	public	
single_subm_pred_extragalactic_gen_unk.csv 14 days ago by Aleksandr Avdyushenko pred_1_037_and_extraGal_genUnk	1.03363	1.02195	☐
single_subm_pred_extragalactic.csv 14 days ago by Aleksandr Avdyushenko pred_1_037_and_extragal	1.04733	1.03429	☐
single_subm_lgb_genUnknown_extragalactic.csv 14 days ago by Aleksandr Avdyushenko lgb_genUnknown_extragalactic	1.09448	1.07006	☐
single_subm_lgb_and_gen_unknown.csv 14 days ago by Aleksandr Avdyushenko lgb_and_gen_unknown	1.10202	1.07725	☐
single_subm_0.652208_2018-12-09-05-14.csv 14 days ago by Aleksandr Avdyushenko lgb_3	1.10568	1.08044	☐

Submission and Description	Private Score	Public Score	Use for Final Score
025_lgb_1036_075_all_1021.csv 7 days ago by Aleksandr Avdyushenko 025_lgb_1036_075_all_1021	1.02831	1.01663	✓
single_subm_0.118646_0.872081_2018-12-17-05-26.csv 7 days ago by Aleksandr Avdyushenko 2_lgbm_smote	1.04803	1.03659	✓

Непроверенные идеи

- catboost
- lstm
- умный blending с oof, stacking

В целом очень позитивные впечатления!
Тратил примерно 5-10 ч времени в неделю.

Part of #26 solution

Simple RNN baseline (in pytorch), kernel https://www.kaggle.com/johnfarrell/plasticc-2018-emb-gru

Only uses 5 raw series: 'mjd', 'flux', 'flux_err', 'detected', 'passband‘ (embedding dim is 16) &
meta features: 'ddf', 'hostgal_photoz', 'hostgal_photoz_err', 'distmod', 'mwebv'.

The score result is **oof0.650/pub0.938/pri0.970 (~ 90 place, bronze)**.
And it has very low correlation with stats features so stacking works very well with other models using stats features.

https://www.kaggle.com/c/PLAsTiCC-2018/discussion/75031

Решение	Score	Место
• LightGBM • Bayesian approach to removing noise: • flux = (flux/flux_err**2 + flux_mean/flux_std**2) / (1/flux_err**2 + 1/flux_std **2) • adding **features based on scaled flux** • features to **capture the behaviour around the peak**	0.84070	14
• LGBM model inspired by Oliver kernel, feature design and selection, and augmentation • magnitudes, peak widths • **parametric curve fittings** • k-correction • **augmented train 10x times** using flux variation within normal distribution of the corresponding flux_err, similar for photoz value • Feature selection: manual approach and eli5 (permutation importance) • 50/50 blend of two LGBM models with slightly modified features and parameters	0.82691	13
Stacking of • lgb as in Chia-Ta's kernels with whole series • lgb with features from detected==1 and train for galaxy and ex-galaxy separately • MLP (multilayer perceptron) as in Siddhartha's kernel with same features • RNN feature extraction with attention and **pseudo labels 0.95 -> 0.80**	0.80905	8

Решение	Score	Место
NN consisting of 3 components: Meta Encoder - taking as input meta features hostgal photoz, hostgal photoz err, is galactic and summary stats for flux and flux_err as min/max/mean etc. Light Curve Encoder - bidirectional GRU taking as input grouped by day flux, flux_err, detected and time difference (autoencoder on test data) Spectroscopic Redshift Predictor – two fully-connected layers for predicting hostgal specz (on test data too) Outputs of these 3 components are fed into two last fully-connected layers for predicting class probabilities. Augmentations: flux as a Gaussian with standard deviation flux_err, hostgal photoz as a Gaussian with standard deviation hostgal photoz err, randomly dropping observations Top submission is an average of 5 cross-validation runs of 3 neural network variations.	0.80173	6
Blend of LGB, NN and several stacking models: • probing class_99 • predict hostgal specz using training set+ test set with hostgalspecz • using normal values and log transformed values together on NN • Bazin: this is a light curve fit method • Log Ensemble: Instead of averaging predictions, I have averaged the logarithm of the predictions because in the end we try to regress the log values	0.70423	4

Решение	Score	Место
NNs worked much better than LGB models. Final ensemble (stacking) included 9 models, 7 of which were NNs that scored as low as 0.75x individually. The two LGB models were much weaker, each scoring about 0.90x, but they did provide a bit of diversity to the ensemble Hostgal*specz pseudo-labeling* Flux adjustments Augmentation (a little bit)	0.69933	2
(an astronomer studying supernova cosmology) • augmented the training set by degrading the well-observed lightcurves in the training set to match the properties of the test set • use Gaussian processes to predict the lightcurves • measured 200 features on the raw data and Gaussian process predictions • trained a single LGBM model with 5-fold cross-validation • lots of different features • class 99 objects was a weighted average of the predictions for classes 42, 52, 62 and 95	0.68503	1

«Чуваки усреднили решения топ-5 команд и пробили скор 0.6 на паблике!!»
https://www.kaggle.com/c/PLAsTiCC-2018/discussion/75179

We finally reached 0.5x using 1st, 2nd, 3rd, 4th, 5th subs!
Reallly Great :)

Submission and Description	Private Score	Public Score	Use for Final Score
last_sub.csv.gz 2 minutes ago by **mamas** 1st * 0.33 + 2nd * 0.33 + (3rd + 4th + 5th)/3 * 0.33	**0.61146**	**0.59828**	☐