**Поиск вспышек красных карликов в данных обзора the Zwicky Transient Facility с использованием методов машинного обучения**

***Лаврухина А.Д.***

студент

Московский государственный университет имени М.В.Ломоносова,

факультет космических исследований, Москва, Россия

lavrukhina.ad@gmail.com

Красные карлики являются самыми распространенными звездами в нашей Галактике. Будучи маломассивными звездами, они демонстрируют частые вспышки, связанные с процессами магнитного пересоединения в их атмосферах. Изучение вспышек красных карликов дает ключевую информацию о звездном магнетизме, явлениях физики высоких энергий и воздействиях на потенциальные планеты, вращающиеся вокруг этих звезд. Однако многие фундаментальные свойства вспышек красных карликов остаются слабо изученными, включая частоту их возникновения, энергетику и взаимосвязи со свойствами звезд, такими как возраст и металличность. В этом исследовании для формирования выборки вспышек красных карликов используются данные с высокой скважностью наблюдений широкоугольного обзора всего неба the Zwicky Transient Facility [1]. Для решения задачи используются методы машинного обучения, в частности, обучаются модели бинарного классификатора на основе решающих деревьев.

В качестве исходных данных были использованы 4 249 038 968 кривых блеска в фильтрах *g, r*  и *i* из 13-го релиза данных ZTF. Были отобраны 420 022 кривых блеска продолжительностью не менее 30 минут и временной задержкой между соседними наблюдениями не более 30 минут. Кроме того, для обучения моделей были использованы синтезированные кривые блеска, полученные на основе наблюдений звездных вспышек в обзоре TESS [2]. Каждая кривая блеска была предварительно предобработана – из временных рядов был извлечен 31 признак (подробное описание пакета *light-curve* для извлечения признаков из кривых блеска представлено в [5]).

Для решения поставленной задачи были обучены две модели бинарного классификатора – случайный лес и градиентный бустинг. Случайный лес — модель машинного обучения, в основе которого лежит обучение ансамбля решающих деревьев. В качестве модели использовалась реализация случайного леса из библиотеки scikit-learn [4] со стандартными гиперпараметрами и количеством деревьев равным 100. В качестве модели градиентного бустинга использовалась реализация градиентного бустинга CatBoost [3]. Гиперпараметры модели: шаг обучения — 0.01, глубина — 5, функция потерь — логистическая функция ошибки. Модель обучалась на протяжении 10 000 итераций.

Перед оценкой производительности для каждой модели была проведена процедура оптимизации порогового значения. Пороговое значение для каждой модели было выбрано на основе валидационной выборки таким образом чтобы максимизировать значение метрики $F\_{β}$-score ($β=0.3$). Для оценки качества работы моделей на тестовом наборе данных использовались следующие метрики:

$$Recall= \frac{TP}{TP+FN},$$

$$Precision= \frac{TP}{TP+FP},$$

$$Accuracy= \frac{TP+TN}{TP+TN+FP+FN },$$

где TP (true positive) – количество правильно классифицированных объектов положительного класса, TN (true negative) – кол-во правильно классифицированных объектов отрицательного класса, FP (false positive) – кол-во объектов отрицательной выборки, неверно классифицированных как положительные, FN (false negative) – кол-во объектов положительной выборки, неверно классифицированных как отрицательные. Полученные метрики и оптимальный порог представлены в Таблице 1. Все метрики получены на одной тестовой выборке.



Таблица 1. Полученные метрики на тестовой выборке и оптимальный порог для модели на основе случайного леса (Random forest) и градиентного бустинга (CatBoost)

Дополнительно модели были протестирована на выборке, состоящей из 104 реальных вспышек, найденных в данных ZTF при помощи иных методов и 1000 объектов отрицательного класса (см. Таблицу 2).



Таблица 2. Полученные метрики на выборке с реальными вспышками для модели на основе случайного леса (Random forest) и градиентного бустинга (CatBoost)

Сравнение метрик показало, что модель градиентного бустинга показывает лучшую производительность. Предварительно модель на основе градиентного бустинга была применена к 2% целевого набора данных, в результате чего были обнаружены 25 новых кандидатов во вспыхивающие события. Такой результат позволяет предположить, что применение классификатора ко всему целевому набору данных позволит обнаружить около 1000 новых кандидатов, что значительно превосходит ранее опубликованные подобные выборки на основе всенебесных фотометрических обзоров.

Исследование выполнено при финансовой поддержке РНФ в рамках научного проекта №24-22-00233.

[1] Bellm, E.C., et al.: The Zwicky Transient Facility: System Overview, Performance, and First Results. Publications of the Astronomical Society of the Pacific 131(995), 018002 (Jan 2019). https://doi.org/10.1088/1538-3873/aaecbe

[2] Günther, M.N., et al.: Stellar Flares from the First TESS Data Release: Exploring a New Sample of M Dwarfs. Astronom. J.159(2), 60 (Feb 2020). <https://doi.org/10.3847/1538-3881/ab5d3a>

[3] Dorogush, A.V., Ershov, V., Gulin, A.: Catboost: gradient boosting with categorical features support. CoRR abs/1810.11363 (2018), <http://arxiv.org/abs/1810.11363>

[4] Pedregosa, F., et al.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)

[5] Malanchev, K.L., et al.: Anomaly detection in the Zwicky Transient Facility DR3. Monthly Notices of the Royal Astronomical Society 502(4), 5147–5175 (Apr 2021). https://doi.org/10.1093/mnras/stab316