Tree-based machine-learning classifier for stellar flares in The Zwicky Transient Facility survey

Anastasia Lavrukhina[1]

Faculty of Space Research, Lomonosov Moscow State University, Leninsky Gori 1 bld. 52, Moscow 119234, Russia

Abstract. This work is dedicated to solving the task of detecting flares from red dwarfs among the light curves of The Zwicky Transient Facility (ZTF) data. The study utilizes light curves with a temporal delay of no more than 30 minutes since the characteristic duration of flares ranges from 30 minutes to 2 hours. The task is addressed using machine learning methods, specifically a binary classifier. Two models were employed as classifiers: random forest and gradient boosting. Both real ZTF data and synthesized flare light curves were used for model training. All models were tested on both synthesized flares and real flares which were found in the ZTF data previously. Based on the validation set with real flares, it was concluded that the gradient boosting model demonstrates the best performance. The achieved model quality allows it to be used for directly assembling a sample of red dwarf flare candidates.

Keywords: Machine learning · Astronomy · Stellar flare.

1 Introduction

Flares of red dwarf stars are incredibly energetic phenomena, spanning a wide range of energies from $E \sim 10^{26}$ erg up to $10^{35} - 10^{36}$ erg [8, 7]. Their light curves have a distinctive profile with a drastically rapid brightening and following exponential-like decline, with the entire duration lasting from tens of minutes to several hours. Studying flares of red dwarfs is important for exoplanetary science because these stellar flares release a significant amount of energy in the ultraviolet spectrum, impacting the habitability of nearby planets [2]. Furthermore, compiling a statistically significant sample size can aid in further investigations of the population of such objects.

This work proposes solving the given problem of stellar flare identifying using machine learning methods. In the present day, machine learning and deep learning techniques have become highly effective tools for addressing challenges in data-intensive domains of astronomy. One of the most common task is an object classification, which is now solved efficiently based on machine learning methods [6, 3]. Equally significant is the task of anomaly detection, which aids in identifying rare events or objects exhibiting unexpected physical characteristics [12, 13]. We propose training a binary classifier that will help select stellar flares candidates from high-cadence data of the Zwicky Transient Facility

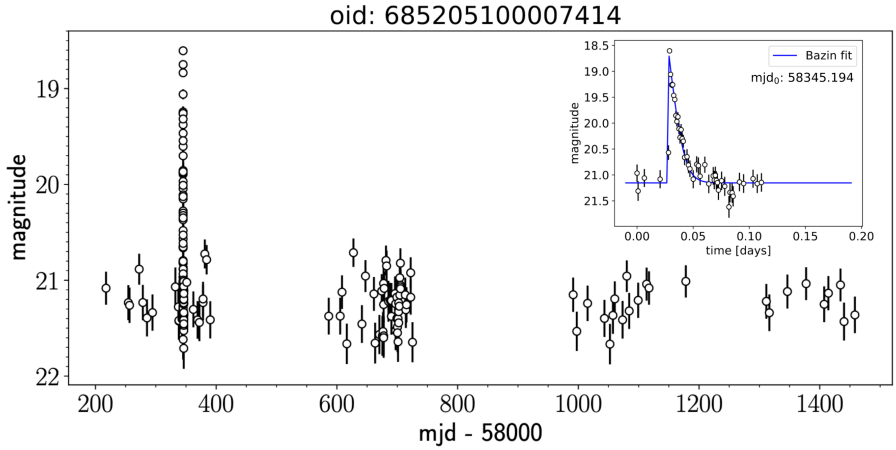
(ZTF) [1]. Subsequently, an expert evaluation of the flare candidates obtained using the model will be conducted for further detailed analysis and potential artifact filtering.

2 Data

We used 4 249 038 968 g-, r-, and i-band light curves from the 13th release of The Zwicky Transient Facility (ZTF) as the source data. A light curve is a time series of the stellar magnitude (brightness) of an astronomical object. Light curves in the ZTF survey are based on epochal PSF-fit photometry (for more details, see [11]). We selected 420 022 light curves having duration of at least 30 minutes, and time intervals between consequent observation to be not less than 30 minutes. We also synthesized the same amount of light curves based on TESS observations of stellar flares [5]. Each light curve was pre-processed for irregular time series feature extraction, 31 features in total (see [10, 9] for `light-curve`[1] package description).

An example of red dwarf flare light curve found in the ZTF data is presented in Fig. 1.

Fig. 1. Red dwarf flare light curve found in ZTF data.



[1] https://github.com/light-curve

3 Methods

3.1 Random forest

Random Forest is a machine learning model based on an ensemble of decision trees. The ensemble refers to training multiple models on bootstrap samples and averaging their responses to obtain predictions. We used the implementation of random forest from `scikit-learn` package with the default hyperparameters and 100 trees as a model.

3.2 CatBoost

Unlike ensemble methods, boosting builds the base algorithms sequentially. Each subsequent base model is constructed to reduce the error of the current model. Boosting that uses decision trees as base algorithms is called gradient boosting with decision trees.

In this study, we used the `CatBoost` [4] implementation of gradient boosting. The model's hyperparameters were set as follows: learning rate of 0.001, depth of 5 and the logistic loss function. The model was trained for 10 000 iterations.

4 Models evaluation

4.1 Validation on test dataset

The following metrics were used to evaluate the performance of the models on the test dataset: recall, precision, accuracy and F_β-score. F_β-score use a factor β to reweight an importance of recall metrics in comparison to precision:

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

For the F_β-score metric, a value of β equals to 0.3 was chosen, as precision is more significant in our task. All candidates identified by the classifier will be intended to undergo a detailed expert analysis to exclude possible artifacts (observable phenomena with non-astrophysical nature). Therefore, it is necessary to optimize the number of objects that will be further analyzed by the expert.

Prior to evaluating the performance for each model, a threshold optimization procedure was conducted. The threshold for each model was selected based on the validation dataset in order to maximize the value of the F_β-score metric ($\beta = 0.3$).

The metrics for all described models, along with the optimal threshold, are presented in Table 1. All metrics were obtained from the same test dataset.

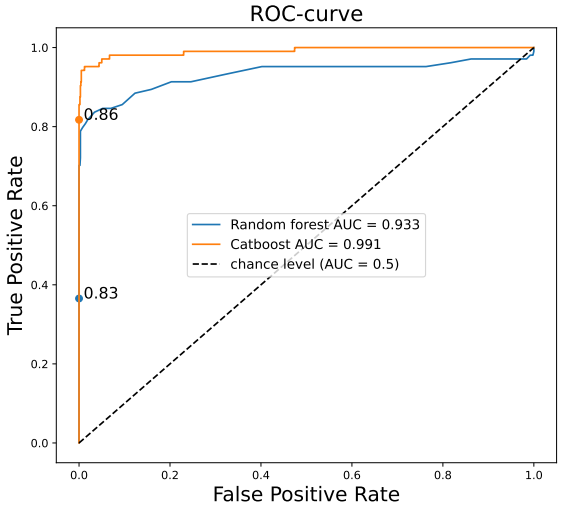
Table 1. The metric results on the test dataset and the optimal threshold for the two trained models: random forest and gradient boosting (CatBoost).

	Precision	Recall	Accuracy	F_β-score	Threshold
Random forest	0.983	0.791	0.889	0.964	0.83
CatBoost	0.978	0.777	0.880	0.958	0.86

4.2 Validation on real flares

Since the positive sample of flares used for training was generated synthetically, while the negative sample was taken from real light curves in the ZTF data, it was necessary to verify the considered models indeed learned to classify objects of the specified type rather than distinguishing synthesized light curves from real ones. For this reason, all models were tested on a dataset consisting of 104 real red dwarf flares previously identified in the ZTF data using other methods (Voloshina et al., in prep.) and 1 000 random negative objects taken from the test dataset. The values of the precision, recall, accuracy and F_β-score metrics were calculated for each model on this dataset (see Table 2). ROC-curve and ROC AUC values for each model presented in Fig. 2.

Fig. 2. ROC curves and ROC AUC values for random forest (blue curve) and CatBoost (orange curve) models. The bold points define a position of the optimal thresholds on ROC curve.



Based on the metrics obtained on real data, the gradient boosting-based classifier demonstrates the best performance so far.

Table 2. Results of the metrics on the dataset with real flares for the two trained models: random forest and gradient boosting (CatBoost).

	Precision	Recall	Accuracy	F_β-score
Random forest	1.0	0.356	0.940	0.870
CatBoost	1.0	0.827	0.984	0.983

5 Conclusion

Within this work, methods for analyzing a large volume of photometric data were developed to solve the object classification task using machine learning methods. A subsample with high observation cadence was extracted from the ZTF photometric data. Based on this subsample and synthesized light curves, two models were trained for classification: random forest and gradient boosting. Comparing the metrics on the test dataset and the dataset with real flares revealed that the gradient boosting model achieved the best performance. Initially, the model was applied to 2% of the target dataset, resulting in the detection of 25 new candidates for flaring events. This outcome provides hope that applying the classifier to the entire target dataset will lead to the discovery of approximately 1 000 new candidates with various galactic declinations, which significantly surpasses previously published datasets of this kind based on all-sky surveys.

Acknowledgements

AL is supported by Nonprofit Foundation for the Development of Science and Education “Intellect”. We used the equipment funded by the Lomonosov Moscow State University Program of Development. This work made use of the Illinois Campus Cluster, a computing resource that is operated by the Illinois Campus Cluster Program (ICCP) in conjunction with the National Center for Supercomputing Applications (NCSA) and which is supported by funds from the University of Illinois at Urbana-Champaign.

References

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. Bogner, M., Stelzer, B., Raetz, S.: Effects of flares on the habitable zones of m dwarfs accessible to TESS planet detections. Astronomische Nachrichten **343**(4) (nov 2021). https://doi.org/10.1002/asna.20210079, https://doi.org/10.1002%2Fasna.20210079
3. Donoso-Oliva, C., Becker, I., Protopapas, P., Cabrera-Vives, G., Vishnu, M., Vardhan, H.: ASTROMER. A transformer-based embedding for the representation of light curves. *Astronomy & Astrophysics***670**, A54 (Feb 2023). https://doi.org/10.1051/0004-6361/202243928

4. 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
5. 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
6. Kim, D.W., Protopapas, P., Bailer-Jones, C.A.L., Byun, Y.I., Chang, S.W., Marquette, J.B., Shin, M.S.: The epoch project. Astronomy & Astrophysics **566**, A43 (Jun 2014). https://doi.org/10.1051/0004-6361/201323252, http://dx.doi.org/10.1051/0004-6361/201323252
7. Kowalski, A.F., Hawley, S.L., Holtzman, J.A., Wisniewski, J.P., Hilton, E.J.: A White Light Megaflare on the dM4.5e Star YZ CMi. *Astronom. J. Letters***714**(1), L98–L102 (May 2010). https://doi.org/10.1088/2041-8205/714/1/L98
8. Lacy, C.H., Moffett, T.J., Evans, D.S.: UV Ceti stars: statistical analysis of observational data. *Astrophys. J.***30**, 85–96 (Jan 1976). https://doi.org/10.1086/190358
9. Lavrukhina, A., Malanchev, K.: Performant feature extraction for photometric time series (2023)
10. 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
11. Masci, F.J., et al.: The Zwicky Transient Facility: Data Processing, Products, and Archive. *Publications of the Astronomical Society of the Pacific***131**(995), 018003 (Jan 2019). https://doi.org/10.1088/1538-3873/aae8ac
12. Pruzhinskaya, M.V., Ishida Emille, E.O., Novinskaya, A.K., Russeil, E., Volnova, A.A., Malanchev, K.L., Kornilov, M.V., Aleo, P.D., Korolev, V.S., Krushinsky, V.V., Sreejith, S., Gangler, E.: Supernova search with active learning in ZTF DR3. Astronomy and Astrophysics **672**, A111 (2023). https://doi.org/10.1051/0004-6361/202245172
13. Villar, V.A., Cranmer, M., Berger, E., Contardo, G., Ho, S., Hosseinzadeh, G., Lin, J.Y.Y.: A deep-learning approach for live anomaly detection of extragalactic transients. The Astrophysical Journal Supplement Series **255**(2), 24 (aug 2021). https://doi.org/10.3847/1538-4365/ac0893

---------- REVIEW 2 ----------
I recommend adding some information in Chapter 2 about ZTF data:
1) In the paper you don't mention which filter (B? V? R? or another maybe there was no any filter?) for photometry was used for ZTF observations. Thus, it is not clear in Fig.1 what magnitude you mean.

Reply:
Thank you, I mentioned filter's pass bands in section 2.

2) Add some information about ZTF dataset, what methods for photometry they used to measure stellar magnitudes, it was psf or aperture photometry?

Reply:
Light curves in ZTF data releases are constructed based on PSF-fit photometry. I added this information to section 2.

---------- REVIEW 3 ----------
The paper describes the problem of identifying flares from red dwarfs in the light curves obtained in the ZTF sky survey. The authors propose to solve the problem with machine learning methods, specifically, training a binary classifier. They propose an algorithm and train it on the subset of synthetized light curves and real known red dwarf flares. Then the model was applied to the real data (~2% of ZTF data-set), and 25 flare candidates were found.

Applying machine learning method of classification of objects is crucial in the vicinity of large sky surveys, like ZTF and the forthcoming LSST. The results, described in the paper, look very promising.
However, I'd recommend to enlarge Introduction section with some general information about red dwarf flares, why they are so important to be found - may be their main properties important for the light curve search.
Also the Method section may be improved by adding some information about previous usage of these methods in astronomy or other big-data sciences. I think adding more citations here may be good.
The citation of TESS is also missing in section 2. In section 4.1 it would be nice to add some comments on what the F-beta value means.

Reply:
Thank you, I have added the further information to section 1: more details about red dwarf flares and description of astronomy-related problems which are being successfully solved using machine learning methods. The formula for F-beta value and a small comment added to section 4.1. The citation of TESS in section 2 is also added.