RNAAS RESEARCH NOTES OF THE AAS

Enhancing Stellar Temperature Estimation through Machine Learning and Multifaceted Data Exploration

To cite this article: Siddhi Bansal *et al* 2024 *Res. Notes AAS* **8** 97

DRAFT VERSION MARCH 27, 2024
Typeset using LATEX default style in AASTeX631

Enhancing Stellar Temperature Estimation through Machine Learning and Multifaceted Data Exploration

SIDDHI BANSAL,[1] PHILLIP PHAN,[1] AND ZAYAAN RAHMAN[1]

[1] *The University of Texas at Austin, Austin, TX, USA*

ABSTRACT

This paper employs machine learning to estimate stellar temperatures using photometric data, focusing on the GAIA ESA Archive Data Release 3 dataset. The study underscores the effectiveness of neural networks in deciphering intricate relationships within the data. Notably, the addition of metallicity improves model accuracy in characterizing stellar properties. The study also investigates outlier removal techniques, specifically favoring the Isolation Forest method, showcasing its efficacy in refining model performance. Automated machine learning, facilitated by PyCaret Regressor, emerges as a valuable tool, identifying top-performing models and highlighting feature importance. The implications of this research extend beyond the specifics of stellar temperature estimation. In contemplating future directions, this study suggests the exploration of diverse data sources to ensure balanced distributions of stellar temperatures and the potential incorporation of deep learning architectures for heightened accuracy in addressing astrophysical inquiries.

1. INTRODUCTION

How can we approximate the temperatures of different stars using photometric data, and improve predictions by synthesizing other data?

Approximating the temperatures of different stars using photometric data is accomplished through the use of color indices, which quantify the star's brightness in different parts of the electromagnetic spectrum. A common method involves measuring a star's brightness using photometric bands such as the U, G, R, I, and Z filters.

The key idea is that hotter stars emit more energy in the shorter, bluer wavelengths and appear brighter in the U and G bands, while cooler stars emit more in the longer, redder wavelengths and are brighter in the I and Z bands. The color index allows astronomers to place the star on a color-temperature diagram, such as the Hertzsprung-Russell diagram, and estimate its temperature based on the observed color. This photometric technique provides a powerful tool for categorizing stars across the universe, aiding in the determination of the characteristics of a star.

Gaia became the primary source of data considering its expansive catalog, which offers stars with temperatures over 40,000K. We used the GAIA Data Release 3 dataset (ESA Gaia Collaboration (2022)).

Corresponding author: Siddhi Bansal
siddhibansal@utexas.edu

Corresponding author: Phillip Phan
philphan@utexas.edu

Corresponding author: Zayaan Rahman
zayaanr@utexas.edu

2. DATA

We decided on Gaia as our source of data to retrieve the star's photometric data and metallicity for a total of just over 350,000. We used the *gaia_source* table to query the features of the stars. We obtained the information of the stars by steps of 500K so that we can have a balanced distribution of star temperatures. However, for temperatures before 2500K and after 20,000K our fixed number of stars per step was significantly decreased since there were very few stars in that range. After running some linear regression models with photometric data, we decided to add metallicity as an additional parameter. Additionally, we used the Isolation Forest Algorithm to remove outliers from our data, which were originally skewing our results. We trained a general neural network on this dataset with outliers and without metallicity to predict the temperature of a star. Then, we used AutoML via PyCaret to survey a space of multiple models and converge on an even more accurate top model.

As seen in the GAIA Data Release Documentation 8.3.1 Effective temperatures by René Andrae (Gaia Collaboration (2018)), previous work has been done with predicting effective temperatures using the GAIA Data Release 2 by using an ExtraTrees regression model. This model was trained on stars' photometric data and predicts effective temperature.

Our model, while also predicting effective star temperature, differs from this study in multiple ways, including the type of ML model used for training (neural networks), the training parameters (photometric data and metallicity), and outlier detection.

In addition, the original study only used photometric data to determine stellar effective temperature. By including metallicity as an additional feature in our model, we were able to better estimate the temperatures of stars by using multiple characteristics. As seen in the Annual Review of Astronomy and Astrophysics Chapter 5, "an increase in metallicity results in lower effective temperatures" Conroy (2014). By accounting for metallicity in our model, we can improve the existing study by accounting for other stellar features that can affect temperatures significantly.

In the original study, an ExtraTrees ensemble regressor was used, which proved to be a fairly potent non-parametric learning algorithm. However, it had its limitations. First of all, its training was limited to non-synthetic photometry and the temperature range was limited to 3000K-10000K, thus being very limited in ability to extrapolate. Our general neural network model significantly improves the prediction range due to training on star data points from a much wider range of temperatures. Additionally, because of the neural network's optimization update via back-propagation and stochastic gradient descent, it is inherently more adaptable to changes in data, meaning whether or not there is synthetic data or misleading data that becomes corrected later on, is not as much of a problem as it is for ExtraTrees. Finally, neural networks have a strong capability of feature extraction and function approximation for an arbitrary dataset. This allows scientists to both infer what aspects of a star's data contribute more to its temperature, as well as apply the model for new features like new color indices that may reveal new patterns.

3. RESULTS

We employed a neural network using the TensorFlow library. The ultimate model for the GAIA dataset consists of 4 input nodes (3 color filters + metallicity), 4 hidden layers, 2 64-node layers, 2 32-node layers, and 1 output node for the temperature value. The hidden layers were decided as a result of trial and error based on validation accuracy, in addition to tuning the number of training epochs. Also, by applying the Isolation Forests algorithm for outlier detection, 8% of our stars were removed while enhancing the accuracy and reliability of our star temperature estimation model. Our accuracy with this model was 73% (61% before detecting outliers). Star temperatures towards the lower (2500K) and higher ends (20,000K) had limited data, causing our model to predict in undesirable ways. However, temperatures between the extreme ends show more confident predictions, giving us confidence in predicting star temperature between this range.

Automated machine learning (AutoML) is an emerging yet relatively unexplored technique for optimizing the design of machine learning models. It involves iterating the data set over a series of different machine-learning models, each with unique parameters and structures. This includes not only neural networks, but decision trees, K-neighbors, and many hybrid-resembling models. After iterating over a multitude of models, it ranks each model's performance on a leaderboard so the programmer can focus on the top-performing models. Overall, AutoML helps to reduce time spent on manual model searching and parameter tuning.

We decided to use the PyCaret Regressor library for our implementation of AutoML. It was run on GAIA, with outlier removal and metallicity included. The top-performing model of GAIA was the Random Forest Regressor with a r^2 value of 0.822 on the test split. Thus, through AutoML, the top models yielded correlations significantly higher than the tuned neural network counterpart.

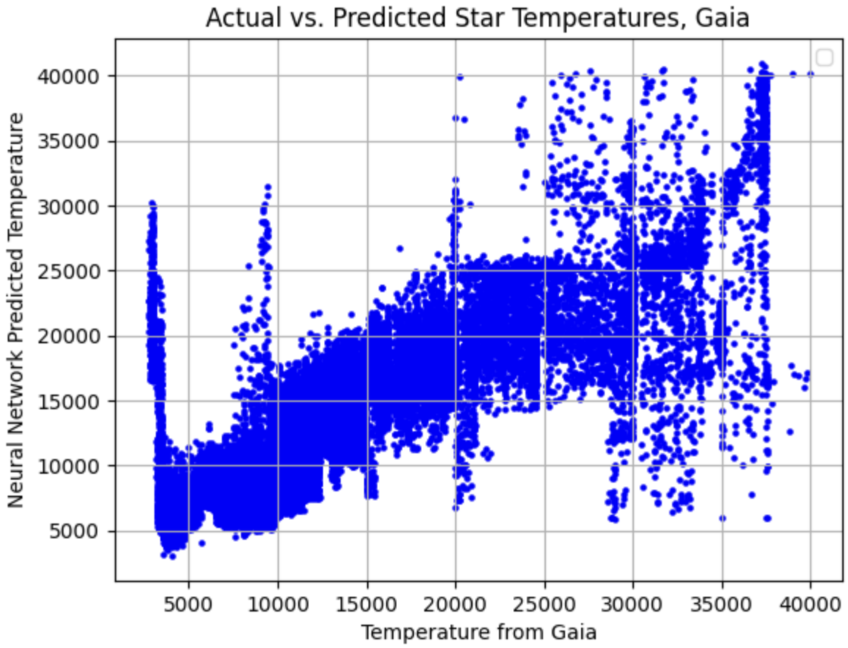
4. CONCLUSION

Our study has yielded several key findings regarding the estimation of stellar effective temperatures using machine learning algorithms. Firstly, neural network models consistently outperformed linear regression and polynomial regression models, demonstrating their superior ability to capture complex relationships between photometric data and temperature. This highlights the potential of neural networks in astronomical applications involving temperature estimation.

Secondly, the Isolation Forest outlier detection program proved to be effective at identifying and removing outliers from our dataset. Its adaptiveness and robustness to outliers made it well-suited for handling the diverse characteristics of our data.

Thirdly, incorporating metallicity as an additional parameter significantly enhanced the accuracy of our models for all three machine learning algorithms. This finding underscores the importance of metallicity in characterizing stellar properties and its influence on temperature estimation, especially when using photometric data.

Finally, the U-G color index exhibited the highest degree of skewness and outlier density among the photometric parameters. Applying the Isolation Forest outlier detection program specifically to the u-g



	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
rf	Random Forest Regressor	1914.2137	12629994.1581	3553.0654	0.8233	0.2923	0.2045
et	Extra Trees Regressor	1994.2799	13065051.2691	3613.9336	0.8172	0.3004	0.2157
lightgbm	Light Gradient Boosting Machine	2178.7956	13210201.29	3633.8995	0.8152	0.3178	0.2453
knn	K Neighbors Regressor	2082.7558	14677831.6	3830.5763	0.7947	0.3197	0.2321
gbr	Gradient Boosting Regressor	2547.8884	15953736.32	3993.6848	0.7768	0.3617	0.2977
dt	Decision Tree Regressor	2315.0227	22890085.39	4782.8187	0.6798	0.3798	0.2323
ada	AdaBoost Regressor	4098.956	32383856.45	5686.7509	0.5472	0.4872	0.4885
llar	Lasso Least Angle Regression	4622.9819	41777917.6	6463.3357	0.4156	0.564	0.5685

Figure 1. Gaia neural network results and AutoML model rankings

parameter resulted in a substantial improvement in model accuracy, further emphasizing the importance of effective outlier detection for enhancing model performance.

Through AutoML, we were able to run and test other models that would've otherwise been outside the scope of this study, providing us with valuable insights into the relationships between effective star temperature and photometric data.

Overall, through this study, we were able to conclude that photometric data and effective stellar temperatures have a relationship when an additional star characteristic - metallicity - is also considered.

We thank Shyamal Mitra, our mentor in the Geometry of Space research group at the University of Texas at Austin, for his support and guidance in developing our research.

REFERENCES

Conroy, C. 2014, Stellar Metallicities Abundance Patterns. https://ned.ipac.caltech.edu/level5/Sept14/Conroy/Conroy5.html

ESA Gaia Collaboration. 2022, Gaia Data Release 3 Papers, https://www.cosmos.esa.int/web/gaia/dr3-papers

Gaia Collaboration. 2018, Gaia Data Release 2 Documentation. https://gea.esac.esa.int/archive/documentation/GDR2/Data_analysis/chap_cu8par/sec_cu8par_process/ssec_cu8par_process_priamteff.html