Explainable Artificial Intelligence for Solar Flare Prediction

Cody Feldhaus[1, 2], Wendy Carande[1]
[1]Laboratory for Atmospheric and Space Physics, [2]City College of San Francisco

Background and Objective Solar flares and their associated space weather activity can have serious consequences for humanity. Without proper preparation, power grids and satellites risk malfunction or failure and space travelers risk exposure to harmful radiation. Accordingly, predicting solar flares before they occur is of great interest to the scientific community and a benefit to society as a whole. Machine learning, a subset of artificial intelligence, has already been shown to be capable of predicting flares[1], but the algorithms used are black-box models, meaning the ways in which the model makes its predictions are largely unknown to experts. Not only does this limit the ability of experts to learn more about solar flare occurrence from the model but it also leaves room for lack of trust in the model itself. This lack of transparency disincentivizes the use of machine learning in other research applications in general. Integrating Explainable Artificial Intelligence (XAI) into existing models has the potential to remove the mysteriousness surrounding black-box models. An emerging paradigm in machine learning, XAI stresses the importance of accountability, responsibility, and transparency in applications of artificial intelligence and machine learning[2]. Demonstrating the potential for XAI to explain solar flare prediction models encourages further investigation into its applications for space weather prediction, solar physics, and science in general.

Methods and Data To integrate XAI into a solar flare prediction model, Local Interpretable Model-Agnostic Explanations (LIME)[3], a Python library for explaining machine learning models, was used to explain instances of solar flare predictions made by a support vector machine (SVM) classifier model trained and tested on a dataset of various space weather parameters derived from magnetogram data from Solar Dynamic Observatory Helioseismic and Magnetic Imager (SDO/HMI) from May 2010 to June 2017. LIME outputs graphical representations of the relative weight of each parameter used in its decision as well as its confidence in the final flare vs. no flare prediction. This makes it very easy to identify the parameters and their corresponding values contributing most to the overall prediction. (Fig. 1)

Results

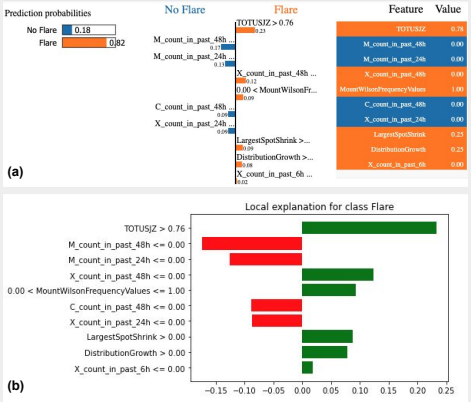

Figure 1. (a) An example of a LIME-generated explanation for a particular instance of the model where a flare was predicted with 82% certainty. The middle plot shows the relative weights of each parameter in LIME's explanation, while the right gives the actual values of those parameters from the data. The values directly next to the bars in the middle plot represent the relative weights of each parameter. **(b)** A more detailed version of the middle plot in **(a)**. The full descriptions of the parameters are visible and comparing the parameter weights to each other is easier.

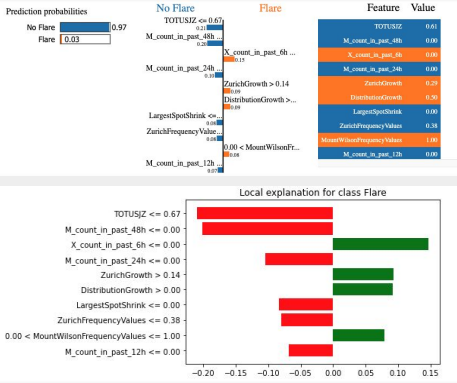

Figure 2. A similar example of a LIME-generated explanation for a particular instance of the model. This time it is for a "No Flare" prediction.

Discussion Examples of the visual representations of explanations generated by LIME for instances of both a "Flare" and "No Flare" prediction can be seen in Fig. 1 and 2, respectively. Fig. 3 below shows an example of how LIME's explanations can be output as text as well, which can be easily customized with code[4]. In both cases, it's easy to see that the value of TOTUSJZ, total unsigned vertical current, was the heaviest weighted parameter in the model's prediction. We can also see that the reason is mirrored for the two cases: for the "Flare" case, TOTUSJZ was above 0.76, while for the "No Flare" case it was less than or equal to 0.67. This leading explanatory parameter, which is heavily weighted in a majority of predictions, would have been left a mystery without LIME, seems to agree with other recent findings in the field of solar flare prediction[5], suggesting LIME's explanations are in fact relatively accurate.

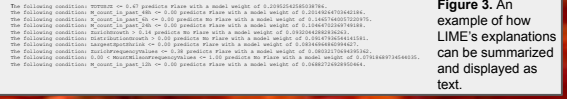

Figure 3. An example of how LIME's explanations can be summarized and displayed as text.

Conclusion and Next Steps LIME's ability to identify electric current as a leading factor in predicting solar flares could be a good indication that it is in fact capable of accurately explaining predictions on more sophisticated models. Further work will need to be done to confirm. LIME's apparent transparency allows for more accountability in our model and thus more trustworthiness. With LIME and XAI we may very well be able to open up the black-box models in this field and others.

References

[1]Bobra, M.G., et al. "Solar Flare Prediction Using SDO/HMI Vector Magnetic Field Data with a Machine-Learning Algorithm." *The Astrophysical Journal*, 798(2), (2015).
[2]Adadi, A., et al. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access*, 6, 52138-52160, (2018).
[3]Ribeiro, M.T., et al. "Why Should I Trust You? Explaining the Predictions of Any Classifier." Retrieved from https://arxiv.org/abs/1602.04938, (2016).
[4]Bobra, M.G., et al. "Predicting Coronal Mass Ejections Using Machine Learning Methods." *The Astrophysical Journal*, 821(2), (2016).
[5]Dhuri, D.B., et al. "Machine Learning Reveals Systematic Accumulation of Electric Current in Lead-Up to Solar Flares." *Proceedings of the National Academy of Sciences of the United States of America*, 116(23), 11141-11146, (2019)

This research was supported by the National Science Foundation REU program, award # 1659878.