Explainability of High Energy Physics events classification using SHAP.

Raquel Pezoa[1,3], Luis Salinas[2,3], and Claudio Torres[2,3] (1) Escuela de Ingeniería Informática, Facultad de Ingeniería, Universidad de Valparaíso, Chile. (2) Departamento de Informática, Universidad Técnica Federico Santa María (3) Centro Científico Tecnológico de Valparaíso. raquel.pezoa@uv.cl, luis.salinas@usm.cl, claudio.torres@usm.cl

ACAT 2021

1. Introduction

• **Understanding** the predictions of a machine learning model can be as important as achieving high performance.In scientific domains, the **model interpretation can enhance the model's performance, helping to trust them accurately for its use on real data and for knowledge discovery.**

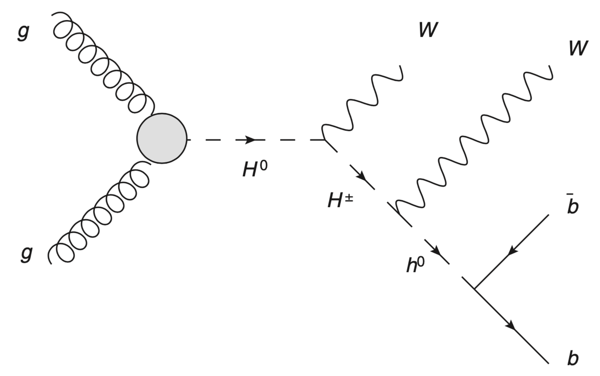
• **eXplainable Artificial Intelligence** (XAI) [1] proposes methods for producing more transparent models, as oppose to black-box models, and for understanding the predictions of **machine learning** (ML) models.

Details of this poster

• In **High Energy Physics** (HEP), the complex nature of the physics processes and data has required the use of complex machine learning models→ **black-box** systems that lack transparency and interpretability for getting a better model performance. **This work is focused on the use of the SHapley Additive exPlanations (SHAP) [2] for interpreting ML classification models of HEP data.**

2. HEP Event Classification

• Event classification is a fundamental task in HEP. In this work we classify the events of the public **Higgs dataset**[a] corresponding to simulated data, for separating the **signal**: $gg \to H^0 \to W^{\mp}H^{\pm} \to W^{\mp}W^{\pm}h^0 \to W^{\mp}W^{\pm}b\bar{b}$ from the background, for identification purposes.



• We used two ML approaches: **eXtreme Gradient Boosting (XGBoost) and deep neural networks (DNN)**.

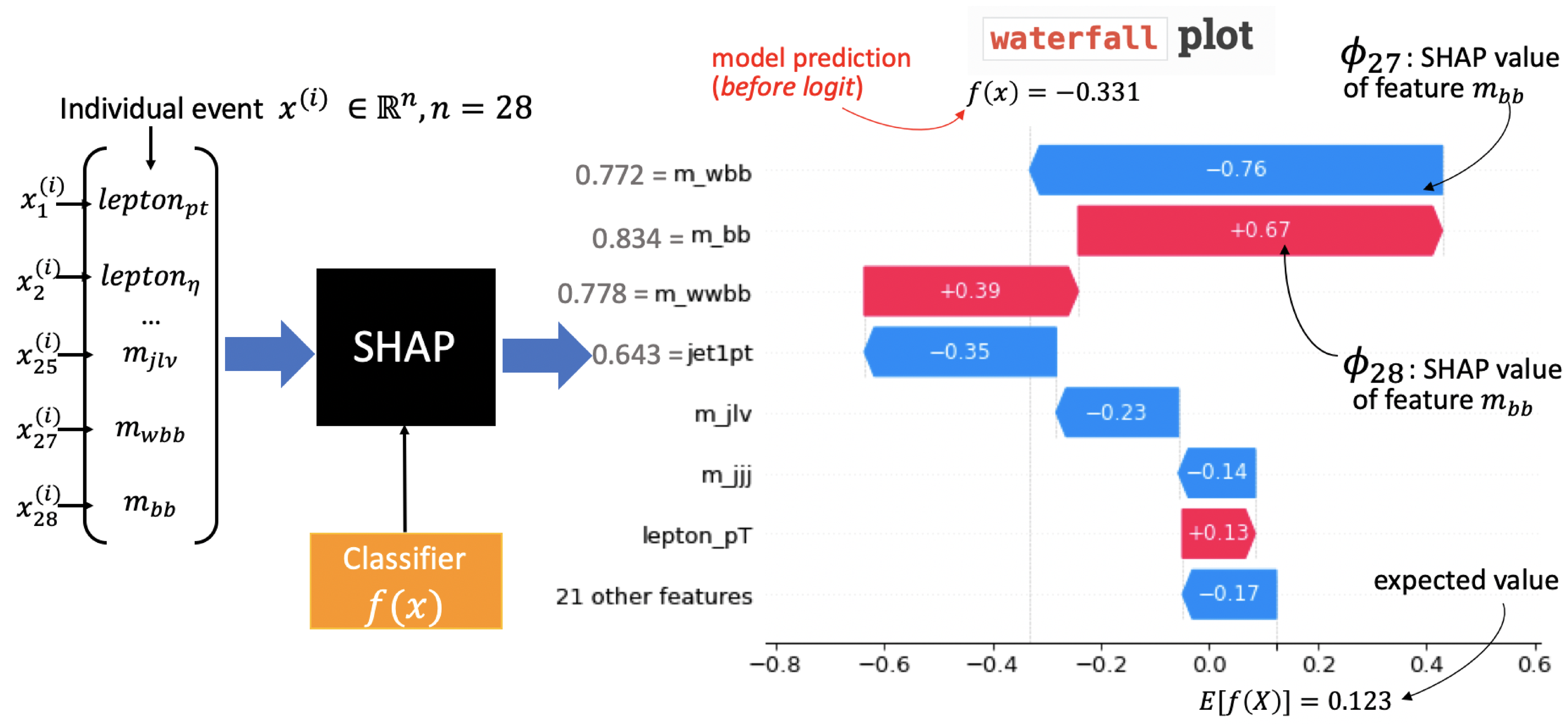
• Each event is represented by 28 feature, and models were trained in the CCTVal cluster[b], using Python libraries: `XGBoost`, `Keras`, and `Talos`.

[a] Baldi et al. Searching for Exotic Particles in High-energy Physics with Deep Learning. Nature Communications, 2014.

[b] `http://www.hpc.utfsm.cl`

3. SHAP, SHapley Additive exPlanations

• **SHAP** [2] is a *post-hoc* explainability technique of the XAI field— for interpreting a ML models. • It is based on game theory and **provides local interpretability assigning to each feature an importance value for a particular sample's prediction.** • Let $f(x)$ be a ML model that predicts y from an input $x = [x_1, ...x_d]$. SHAP explains the prediction of each particular sample $x^{(i)}$ by assigning values $\phi_1, \phi_2, \ldots, \phi_d$ –**the SHAP values**– to each feature $x_1^{(i)}, \ldots x_d^{(i)}$.

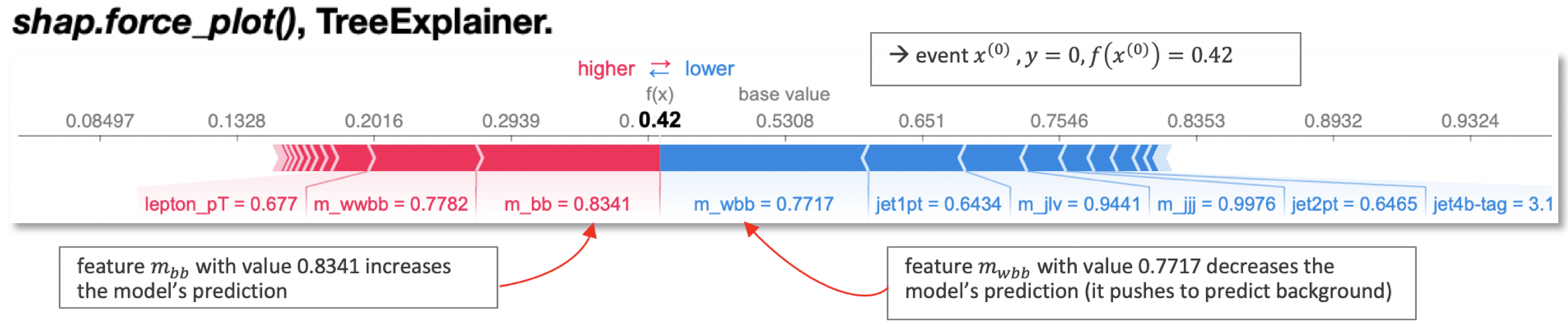


• We compute **SHAP values** with the SHAP Python library [3], using the `TreeExplainer` and `DeepExplainer` methods for XGBoost and DNN event classifiers, respectively. The `waterfall` plot displays explanations for individual predictions.

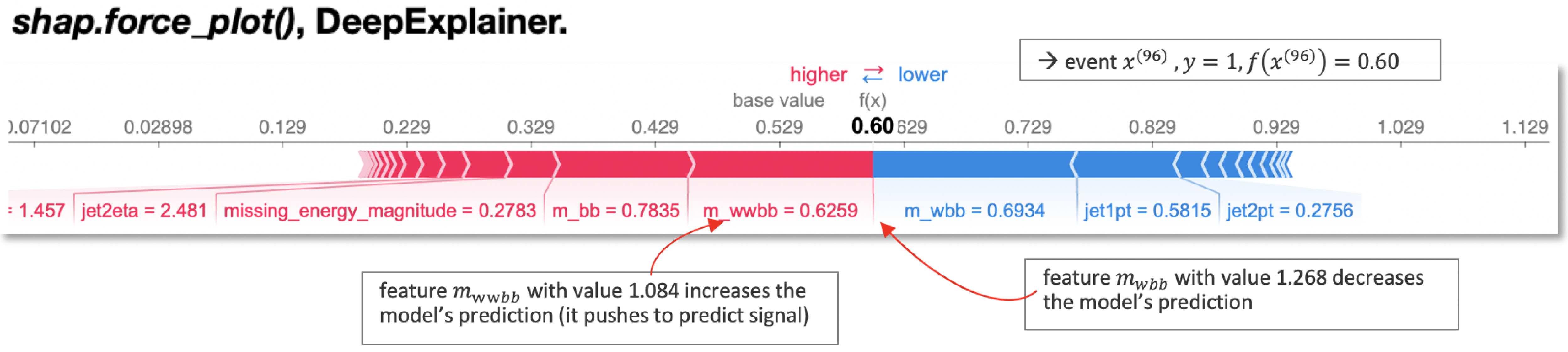
4. Interpretability of HEP Events Classifiers with SHAP

• The `shap.plots.force_plot` shows for a particular event, **the contribution of each feature to the model prediction.**

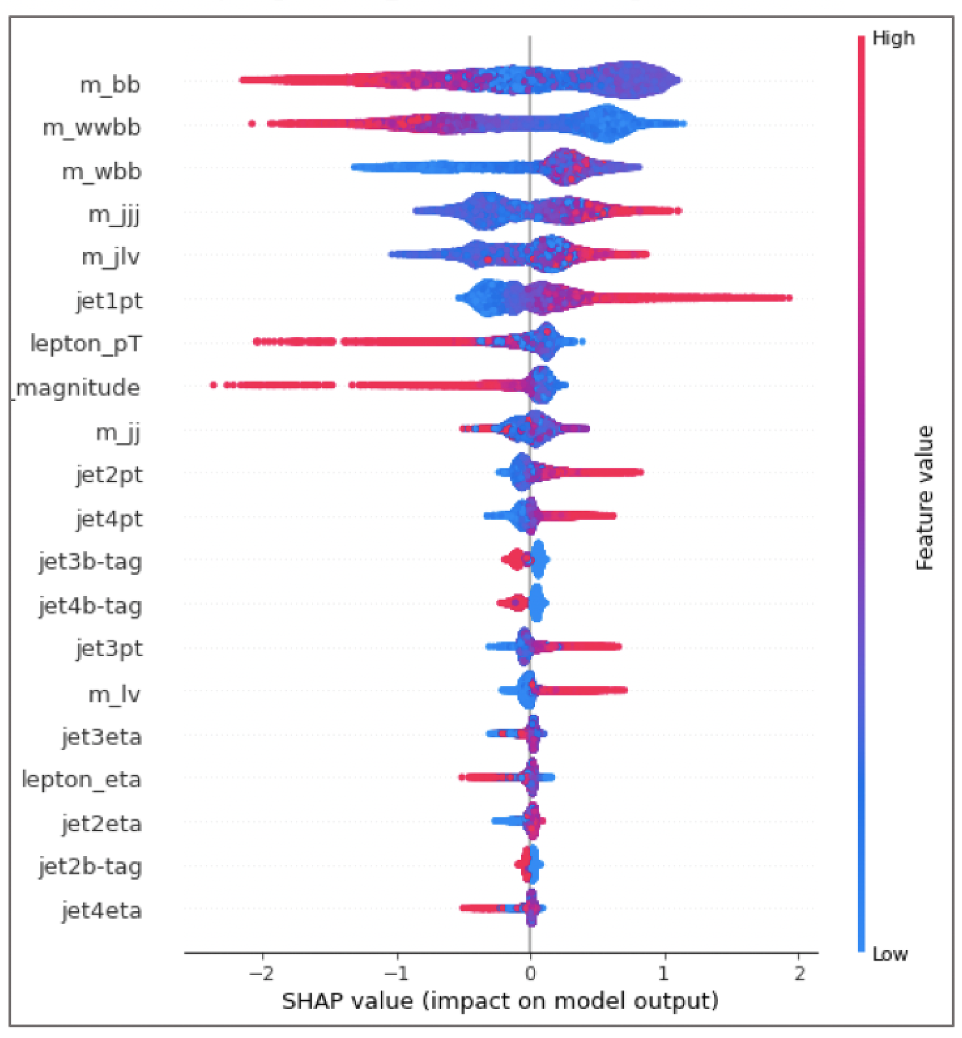
***shap.force_plot()*, TreeExplainer.**



***shap.force_plot()*, DeepExplainer.**



***summary_plot()*, TreeExplainer.**



• The `shap.summary_plot` shows the most important features and their range of effects over the events dataset. From our experiments, the top variables obtained for the XGBoost classifier were m_{bb}, m_{wwbb}, m_{wbb}, m_{jjj}, m_{jlv}, $jet1_{pt}$. For the DNN case, the top variables were m_{bb}, m_{wwbb}, m_{wbb}, m_{jjj}, m_{jlv}, $jet1_{pt}$. We can observe, that the high-level features belong to the top ranking, and hence, they contribute more to the model prediction. • Based on local explainability, SHAP allows to generate global explainability (ranking of features). More plots available at `https://github.com/rpezoa/hep_shap`

5. Conclusions and Future Work

•**SHAP method in HEP is recent, hence** interpretation of ML models using SHAP in HEP has a huge potential for the improvement of the model's accuracy • Future work includes to develop a framework to explain different ML models using SHAP and datasets of other HEP phenomena

• **This research was partially supported by FONDECYT Postdoc Project N. 3190740 and ANID PIA/APOYO AFB180002.**

6. References

[1] A Barredo Arrieta et al. Explainable artificial intelligence (XAI). *Information Fusion*, 2020.

[2] S Lundberg et al. A unified approach to interpreting model predictions. In *Adv. Neural Inf. Process. Syst*, 2017.

[3] `https://shap.readthedocs.io/`.