Approximating Score-based Explanation Techniques Using Conformal Regression

Amr Alkhatib[1], Henrik Boström[1], Sofiane Ennadir[1], Ulf Johansson[2]

[1]KTH Royal Institute of Technology

[2]Jönköping University

COPA 2023

A Need for Explanation

- Bryce Goodman and Seth Flaxman, 2016,"European Union regulations on algorithmic decision-making and a 'right to explanation'":

  "The law will also effectively create a '**right to explanation**,' whereby a user can ask for an explanation of an algorithmic decision that was made about them."

A Need for Explanation

- Interpretable (white-box) models, in many cases, come with a substantial loss of predictive performance
- Post-hoc explanation techniques, e.g., SHAP and LIME come with no guarantees on the fidelity
- Explanation techniques are computationally expensive

Computational Cost of Explanations

- Influential explanation methods, e.g. SHAP, involve approximating Shapley values
- LIME involves creating local (white-box) surrogate models
- Many methods analyze the behavior of a black-box model by running iterations on multiple input data points

Approaches for Efficient Explanation Methods

- TreeSHAP for a tree-based model
- L2E, Learning to Explain, approximates the explainer using neural network for text classification tasks
- FastSHAP learns to approximate the Shapley values
- Hierarchical Shap, h-Shap, for image classification

Approaches for Efficient Explanation Methods

- All the mentioned approaches come without any validity guarantees!

The Main Contributions

- An approach for approximating score-based explanations accompanied with validity guarantees
- A set of non-conformity measures designed for explanations approximation
- A large-scale empirical investigation on 30 publicly available datasets

The Proposed Method

- The explanation method (A) is a function ($A : f(\mathbf{x}, t; \Theta) = \mathbf{y}$) that can be approximated
- The approximation model ($\tilde{A}$) learns a mapping from ($\mathbf{x}$; t) to $\mathbf{y}$
- Since the target $\mathbf{y}$ is a vector with a score per feature, the problem can be formulated as a regression problem
- Can be solved as a multi-target regression or as a set of single-target regression problems

The Proposed Method

1. The black box (B) produces predictions $\mathbf{t}^{\text{dev}}$ on a development dataset (X^{dev})

2. The explanation method (A) generates explanations (Y^{dev})

3. Each data point ($\mathbf{x}$) is augmented with its predicted outcome ($\mathbf{x}' = \mathbf{x} \cup t$)

4. The augmented development set X'^{dev} with the explanations Y^{dev} form:

$$Z^{\text{dev}} = \{(x'_1, y_1), (x'_2, y_2), \ldots, (x'_n, y_n)\}$$

5. $\tilde{A}$ is learned by fitting the regression model on Z^{dev}

The Proposed Method (Validity Guarantees)

- A calibration dataset X^{cal} is augmented with the class labels acquired from *B* to obtain X'^{cal}
- $\tilde{A}$ generates scores for all the data points in the calibration set X^{cal} (predict $\tilde{Y}^{cal}$)
- Using the ground truth Y^{cal} obtained from *A*, a non-conformity score α^f_j is computed for each feature *f* for each example x_j in X^{cal}
- Let α^f_ε be the score of feature *f* at a significance level ε, at prediction time:

$$\tilde{\mathscr{Y}}_j^f = [\tilde{y}_j^f - \alpha_\epsilon^f, \tilde{y}_j^f + \alpha_\epsilon^f]$$

The Proposed Difficulty Estimates

1. Minimum distance to the distributions

Compute the Mahalanobis distance between a data point and each distribution:

$$d_{j\mathcal{C}} = \sqrt{(x_j - \mu_{\mathcal{C}})^T \Sigma_{\mathcal{C}}^{-1} (x_j - \mu_{\mathcal{C}})}$$

Then the minimum distance is used as a difficulty estimate:

$$\sigma_j = \log(\arg\min_{\mathcal{C}}(d_{j\mathcal{C}}) + 1)$$

The Proposed Difficulty Estimates

2. Average distance to the distributions:

$$\sigma_j = \log(\frac{1}{n}\sum_{\mathcal{C}=1}^{n} d_{j\mathcal{C}} + 1)$$

3. The prediction confidence

For multi-class problems: $\sigma_j = 1 - max(\mathcal{P}_{\mathcal{C}})$

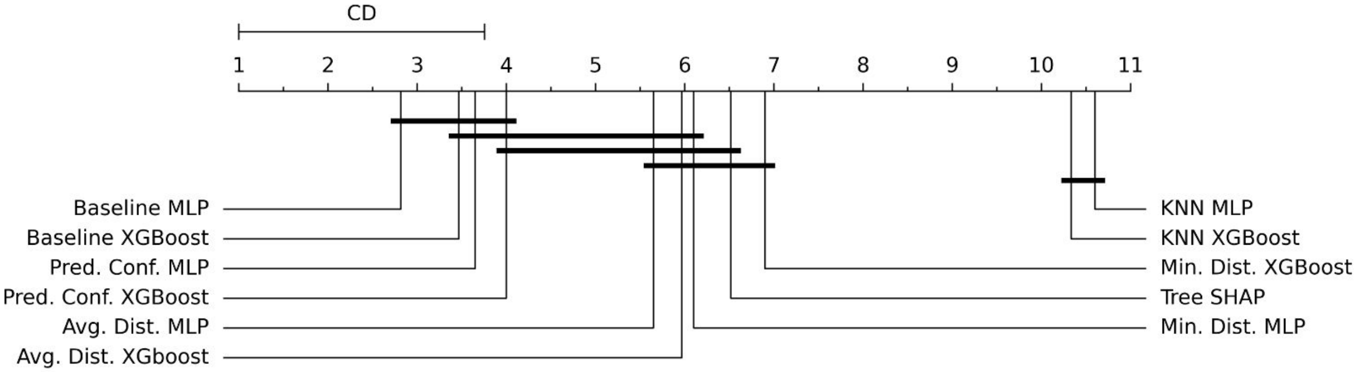
For binary classification: $\sigma_j = 1 - |\mathcal{P} - 0.5|$

Experimental Setup

- The experiments were conducted on 30 public datasets available on Openml.org
- The data was split into training, development, calibration, and test subsets
  - 40% training, 20% development, 20% calibration, and 20% test
- The black-box models were generated using the XGBoost algorithm
- The underlying explainer is TreeSHAP
- The regression models:
  - XGBoost for one-regressor per feature, and MLP for multi-target regression

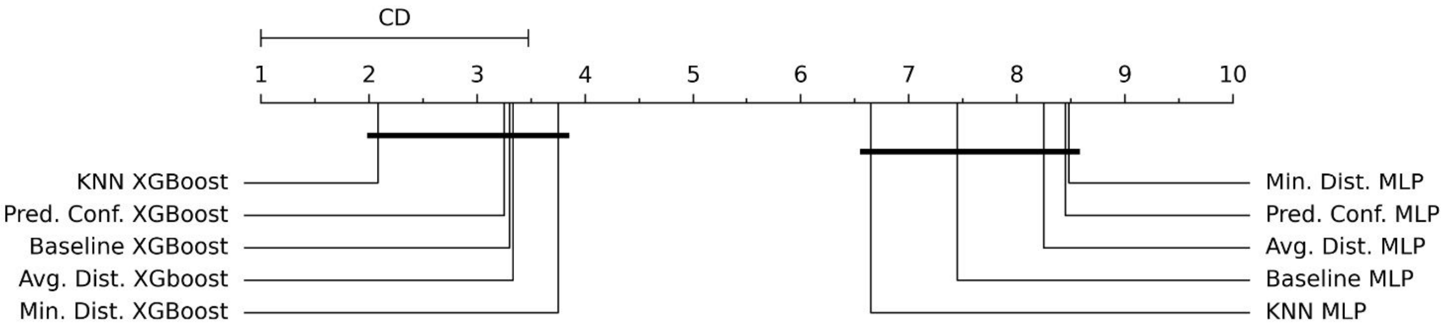
Experimental Results

1. Execution Time

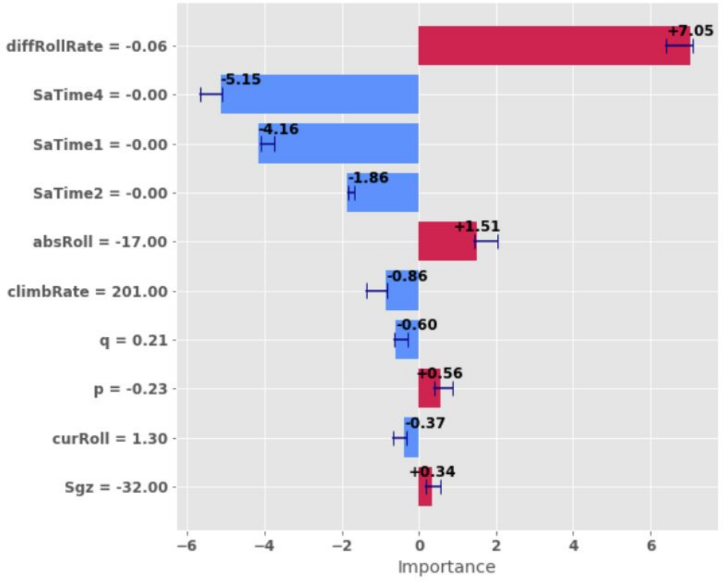


Experimental Results

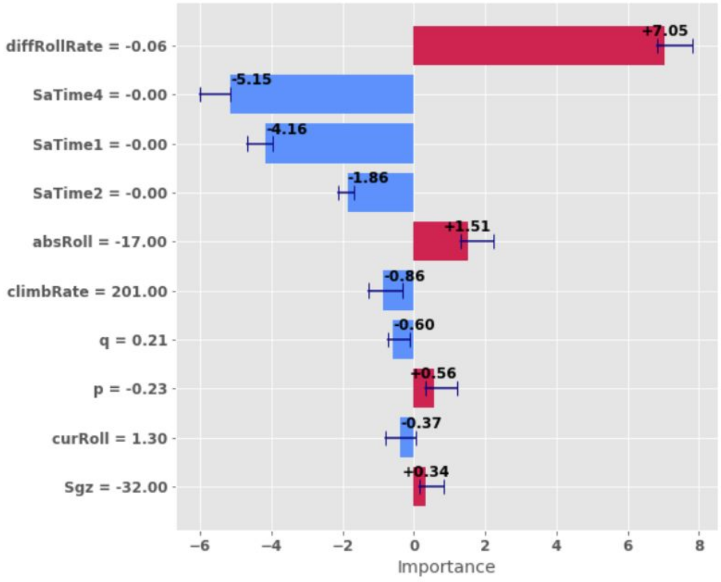
2. Predictive Efficiency (Interval Size)



Explanation Examples



XGBoost with Prob. Conf.



MLP with Prob. Conf.

Concluding Remarks

- We proposed a computationally efficient method to approximate score-based explanation techniques while providing validity guarantees on the generated explanations
- We proposed difficulty estimates targeting explanations
- We have presented results from a large-scale empirical evaluation, comparing the proposed approaches with respect to the computational efficiency as well as the predictive efficiency

Future Work

- Investigate better difficulty estimates designed to save the computational cost

Thank You!

References

- Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alexander Gammerman. Inductive confidence machines for regression. In Proceedings of the 13th European Conference on Machine Learning, ECML ’02, page 345–356, Berlin, Heidelberg, 2002.

- Harris Papadopoulos, Vladimir Vovk, and Alex Gammerman. Regression conformal prediction with nearest neighbours. J. Artif. Int. Res., 40(1):815–840, jan 2011.

- A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, UAI’98, page 148–155, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 155860555X

- Christoph Molnar. Interpretable Machine Learning. 2022.

- Neil Jethani, Mukund Sudarshan, Ian Connick Covert, Su-In Lee, and Rajesh Ranganath. FastSHAP: Real-time shapley value estimation. In International Conference on Learning Representations, 2022.