AutoXplain: Towards Automated Interpretable Model Selection

Tessel Haagen[1,2,*], Heysem Kaya[1], Joop Snijder[2] and Melchior Nierman[3]

[1]Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, The Netherlands
[2]Info Support, Kruisboog 42, 3905 TG Veenendaal, The Netherlands
[3]Atalmedial Medical Diagnostics Centres, Jan Tooropstraat 138, 1061 AD Amsterdam, The Netherlands

Abstract

Machine learning (ML) algorithms are increasingly used in high-stake domains like healthcare. While ML systems frequently outperform humans in specific tasks, ensuring safety and transparency is critical in these domains. Interpretability, therefore, plays a crucial role in understanding the decision-making process, auditing and correction of ML models and establishing trust. Furthermore, there is a growing demand for automated machine learning (AutoML) to facilitate model development without expert intervention. However, the combination of interpretability and AutoML has received limited attention thus far. In this study, we propose two objective model-agnostic measures of interpretability to quantify model compactness and explanation stability, embedded within an automated interpretable ML pipeline. We experiment with a set of interpretable models on medical classification tasks reporting the proposed measures along with the predictive performances. We further conduct a user study with domain experts to evaluate the correlation between these measures and the subjective concept of interpretability. Our findings demonstrate the effectiveness of the proposed measures, affirming their success and validating their utility in creating an interpretable automated pipeline.

Keywords

Interpretable automated pipeline, Interpretability measures, Automated Machine Learning (AutoML), Model-agnostic measures, Machine Learning for health-care

1. Motivation

Machine learning (ML) algorithms, known for their superior performance [1], are widely used in high-stake domains such as healthcare [2]. This rapid growth calls for two things: interpretability and hands-free solutions. Safety and transparency are crucial considerations, emphasizing the need for interpretability [3]. Interpretability plays a vital role in establishing trust [4], but its definition and measurement are challenging. In this study, we adopt Molnar's definition of interpretability as the degree to which a human can consistently predict the model's result [5]. To achieve interpretability, models must have intelligible features, lower complexity,

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

[*] Corresponding author.

tessel.haagen@hotmail.com (T. Haagen); h.kaya@uu.nl (H. Kaya); joop.snijder@infosupport.com (J. Snijder); mnierman@atalmedial.nl (M. Nierman)
http://uu.nl/medewerkers/HKaya (H. Kaya); https://www.linkedin.com/in/joopsnijder (J. Snijder)
0009-0001-4929-9125 (T. Haagen); 0000-0001-7947-5508 (H. Kaya); 0000-0002-8647-7509 (M. Nierman)

CEUR Workshop Proceedings (CEUR-WS.org)

and incorporate a small number of features to account for human working memory limitations [6, 7]. Hands-free solutions are needed for a quicker process, leading to Automated Machine Learning (AutoML) [8]. AutoML pipelines consist of data pre-processing like encoding [9, 10, 11] and discretization [12, 13], optimization including random search or genetic algorithms [14, 15], and result-generation steps. AutoML tools, both open-source and commercial, are available [15, 16, 17, 18, 19, 20, 21], but all are black-box solutions.

Integrating interpretability into AutoML stays unexplored. This study proposes AutoXplain, an automated pipeline using interpretable models and model-agnostic interpretability measures for model selection. AutoXplain can improve decision-making in high-stake domains by exhibiting good tas performance while providing interpretability.

2. Proposed Measures of Interpretability

To compare interpretability across models within the automated pipeline, model-agnostic objective measures are essential. Currently lacking in the literature, we introduce two such measures: compactness and stability.

2.1. Compactness

The compactness of a single explanation quantifies the ability to convey relevant information using a concise set of features. It is defined using the equation:

$$\textit{Explanation compactness} = 1 - \frac{|F_{ex}| - 1}{|F|}, \tag{1}$$

where F is the set of features used in the model and F_{ex} is the subset of features used in the explanation. To avoid penalizing single-feature explanations, 1 is subtracted from $|F_{ex}|$. A higher compactness score (between 0 and 1) indicates a more compact explanation. This equation is used to evaluate the compactness of a model, applying it to each individual explanation, and their average is computed. Model compactness is calculated using the equation:

$$\textit{Model compactness} = \frac{1}{|E|} \cdot \sum_{ex \in E} \left(1 - \frac{|F_{ex}| - 1}{|F|}\right), \tag{2}$$

where E is the set of explanations generated by the model.

For decision rules and decision trees, the number of splits in each rule/to the leaf node is used as the number of features in each explanation. For linear models, a two-step process is followed. First, the importance of each feature in an explanation is determined using the absolute value of its t-statistic. Then, a softmax function is applied to these importances to obtain a probability distribution over the features. A threshold is used to select the most important features for the explanation, resulting in F_{ex}.

2.2. Stability

An explanation method is considered stable if, for similar instances, similar explanations are provided. The methodology proposed is built upon the works of Turney [22] and Zatar et al.

[23]. In order to evaluate stability, we examine the agreement between the instances and the explanations generated by a single classification algorithm. The proposed measure consists of the following steps:

1. **Finding similar instances per instance:** We utilize the k-nearest neighbour method to identify neighbours for each instance in the training dataset. In our experiments, k is set to 9 however the value of k can be adjusted based on the dataset size and desired strictness for stability. By calculating the average radius used to determine k-neighbors for each instance in the training set (t_i), with a training set size of N, we establish the radius threshold T_r:
$$T_r = \frac{\sum_{i=0}^{N} t_i}{N}. \tag{3}$$
2. **Creating the instance space:** Using the calculated radius threshold T_r, we identify neighboring instances in the test data, creating the instance space.
3. **Creating the explanation space:** For decision trees and decision rules, each leaf or rule is its own explanation. For models with feature importances, we use the same method as the instance space, but with the feature importances instead of the feature values. The explanation space consists of the neighbours of each instance that share the same decision rule or leaf (for decision trees and decision rules) or the same set of important features (for models providing feature importances).
4. **Calculating the stability measurement:** From the instance space (Is) and the explanation space (Es), we can calculate the agreement of neighbours, which results in the stability measurement:
$$stability = \frac{|Is \cap Es|}{|Is|} \tag{4}$$

A higher stability score (0-1) indicates greater explanation stability.

3. Experimental Results

3.1. Data and Machine Learning Experiments

Our pipeline primarily focuses on tabular datasets in the medical domain. Specifically, we utilized the Atalmedial Anticoagulation Clinic (AAC) dataset, consisting of de-identified patient records of individuals on oral anticoagulation therapy (VKA) for thrombo-embolic event prevention. The dataset includes important details such as blood values, medical events, and medication history spanning the previous 60 days. Each entry is labelled as S (severe bleeding) or N (non-severe bleeding). Notably, the dataset exhibits class imbalance with 47 S and 5544 N samples. Since the dataset features capture temporal data collected over the last 60 days, they are structured as lists.

AutoXplain explores a range of configurations for Decision Trees (DT) [24] with variations in maximum depth, maximum number of leaf nodes, and minimum samples per leaf, Explainable Boosting Machines (EBM) [25] with different parameter settings for the number of interactions, learning rate, and early stopping rounds, and lastly Dominance Classifier Predictors (DCP) [26]

Table 1
Top hyperparameter settings per model on AAC. WS: Weighted Score, n: number of features, mln: maximum number of leaf nodes, T: ratio-threshold, vm: voting method, ia: interactions, lr: learning rate, esr: early stopping round.

Model	Parameters	F1-score	Compactness	Stability	WS
DT	mln: 5, n: 7	0.984	1.0	0.610	0.864
DCP	T: 0.5, vm: 4, n: 5	0.984	0.545	0.956	0.830
EBM	ia: 0, lr: 0.01, esr: 5, n: 9	0.984	0.909	0.507	0.800

are examined with different parameter settings for the ratio-threshold and voting method. Each model is trained using different subsets of features. In total, 222 models were considered during the optimisation stage. It selects the top 3 model per method on the weigthed score of F1-score, Compactness and Stability with crossvalidation. The parameter setting per ML method with the best weighted score on the test set is presented in Table 1. Our top-performing model is a DT, which achieved a weighted score of 0.864 and an F1 score of 0.984.

3.2. Human evaluation

To assess the explanations generated by our models on the AAC dataset, we conducted a user study with Atalmedial employees, including dosing advisors and specialized medical doctors in VKA-anticoagulation. This diverse group of participants ensured a comprehensive evaluation. Participants were presented with explanations generated by our top models for instances in the AAC dataset, specifically focusing on severe bleeding incidents. Participants then filled in the System Causability Scale [27]. This user study allowed us to gather valuable feedback and insights from professionals in the medical field. The results of the paired-t test revealed a significant difference between DT ($M = 0.6$, $SD = 0.07$) and DCP ($M = 0.4$, $SD = 0.1$), $t(7) = 3.5$, $p = .010$, as well as between EBM ($M = 0.6$, $SD = 0.08$) and DCP, $t(7) = 4.7$, $p = .002$. However, no significant difference was found between DT and EBM, $t(7) = 2.2$, $p = .061$.

4. Conclusion

The results of our study demonstrate that the automated pipeline we developed performs well in terms of model performance and interpretability. Validating the feasibility of an automated interpretable machine learning pipeline.

By incorporating interpretable models and assessing their interpretability using quantitative measures, our automated pipeline allows for the selection of models that strike a balance between performance and interpretability. These results highlight the potential of the automated pipeline in high-stakes decision-making domains, where both accuracy and transparency are essential. The pipeline's ability to generate highly performant models while offering interpretability empowers decision-makers to gain insights into the decision-making process and make well-informed judgments.

References

[1] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, O. Evans, When will ai exceed human performance? evidence from ai experts, Journal of Artificial Intelligence Research 62 (2018) 729–754.

[2] S. Safdar, S. Zafar, N. Zafar, N. F. Khan, Machine learning based decision support systems (dss) for heart disease diagnosis: a review, Artificial Intelligence Review 50 (2018) 597–623.

[3] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, 2017. URL: https://arxiv.org/abs/1702.08608. doi:10.48550/ARXIV.1702.08608.

[4] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, C. Zhong, Interpretable machine learning: Fundamental principles and 10 grand challenges, Statistics Surveys 16 (2022) 1 – 85. URL: https://doi.org/10.1214/21-SS133. doi:10.1214/21-SS133.

[5] C. Molnar, Interpretable machine learning, Lulu.com, 2020.

[6] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. J. Gershman, F. Doshi-Velez, Human evaluation of models built for interpretability, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 7 (2019) 59–67. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/5280.

[7] G. A. Miller, The magical number seven, plus or minus two: Some limits on our capacity for processing information., Psychological review 63 (1956) 81.

[8] F. Hutter, L. Kotthoff, J. Vanschoren, Automated machine learning: methods, systems, challenges, Springer Nature, 2019.

[9] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[10] D. Micci-Barreca, A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems, ACM SIGKDD Explorations Newsletter 3 (2001) 27–32.

[11] J. Van den Bossche, J. De Bock, J. De Brabanter, Efficient categorical variable encoding for multiclass classification, Machine Learning and Knowledge Extraction 1 (2015) 101–121.

[12] M. Lichman, Machine learning in r: using caret, Journal of Statistical Software 58 (2013) 1–26.

[13] H. Liu, R. Setiono, H. Zhu, Feature selection in knowledge discovery and data mining, Data Mining and Knowledge Discovery 2 (1996) 359–394.

[14] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, Journal of Machine Learning Research 13 (2012) 281–305.

[15] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, F. Hutter, Efficient and robust automated machine learning, Advances in neural information processing systems 28 (2015).

[16] C. Thornton, F. Hutter, H. H. Hoos, K. Leyton-Brown, Auto-weka: Combined selection and hyperparameter optimization of classification algorithms, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013, pp. 847–855.

[17] H. Jin, Q. Song, X. Hu, Auto-keras: An efficient neural architecture search system, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 1946–1956.

[18] B. Komer, J. Bergstra, C. Eliasmith, Hyperopt-sklearn: automatic hyperparameter con-

figuration for scikit-learn, in: ICML workshop on AutoML, volume 9, Citeseer, 2014, p. 50.

[19] R. S. Olson, N. Bartley, R. J. Urbanowicz, J. H. Moore, Evaluation of a tree-based pipeline optimization tool for automating data science, in: Proceedings of the genetic and evolutionary computation conference 2016, 2016, pp. 485–492.

[20] E. LeDell, S. Poirier, H2o automl: Scalable automatic machine learning, 7th ICML Workshop on Automated Machine Learning (AutoML) (2020). URL: https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf.

[21] N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li, A. Smola, Autogluon-tabular: Robust and accurate automl for structured data, arXiv preprint arXiv:2003.06505 (2020).

[22] P. Turney, Bias and the quantification of stability, Machine Learning 20 (1995) 23–33.

[23] M. R. Zafar, N. Khan, Deterministic local interpretable model-agnostic explanations for stable explainability, Machine Learning and Knowledge Extraction 3 (2021) 525–541.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[25] H. Nori, S. Jenkins, P. Koch, R. Caruana, Interpretml: A unified framework for machine learning interpretability, arXiv preprint arXiv:1909.09223 (2019).

[26] B. Kovalerchuk, N. Neuhaus, Toward efficient automation of interpretable machine learning, in: 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 4940–4947. doi:10.1109/BigData.2018.8622433.

[27] A. Holzinger, A. Carrington, H. Müller, Measuring the quality of explanations: the system causability scale (scs), KI-Künstliche Intelligenz 34 (2020) 193–198.