Open your black box classifier

Paulo Lisboa [1]

[1]Data Science Research Centre, Liverpool John Moores University, UK

May 13, 2023

Abstract

The transparency of machine learning models is central to good practice when they are applied in high stakes applications. Recent developments make this feasible for tabular data, which is prevalent in risk modelling and computer-based decision support across multiple domains including healthcare. Important motivating factors for interpretability are outlined and practical approaches are summarised, signposting the main methods available, with pointers to the supporting literature. A key finding is that any black box classifier making probabilistic predictions of class membership from data in tabular form can be represented with a globally interpretable model without loss of performance.

Open your black box classifier

There is much discussion about opening black boxes, particularly in relation to predictive models that involve machine learning [1]. Some funding schemes go as far as requiring artificial intelligence models to inform about possible failures, as part of the requirement for technical robustness in high stakes applications. This has led to frameworks to define testable standards for the interpretability of predictive models [2].

In medicine in particular, interpretable models are important, not least because an understanding of the contributing factors towards a diagnosis can be as insightful as the quantification of the diagnostic prediction itself, but also because this level of transparency is essential for building trust in the model [3]. A typical example of good practice in explaining machine learning models in medicine is the application of Shapley values during model validation, which shows "how a domain understanding of machine learning models is straightforward to establish" [4].

The need for explanation is equally as pronounced when deep learning is applied to the classification of medical images. A rigorous study of Covid-19 detection from lung CTs showed that high performance metrics could be achieved when the predictive models focused on artifacts such as annotations in the images and even the support structure on which the person was laid out for the CT. Explanation methods were central to identifying the bias in the models due to spurious effects that happened to correlate with class membership in the data set, despite good practice by splitting the data into three groups for model inference, optimisation and performance estimation [5]. This paper found that "very small increases in validation accuracy can correspond to drastic changes in the concepts learned by the network ... it can mean overcoming a bias introduced by the artifacts."

In many computer-based decision support applications, clinical attributes take the form of tabular data. Being so prevalent, not just in medicine but also for risk models in other domains ranging from banking to insurance, this class of data deserves particular focus and it is the subject of the rest of this piece. For tabular data specifically, one way around the issue of transparency is with models that are interpretable by design [6].

Interpretability by design has long been known to be possible with linear-in-the-parameters models and with decision trees, albeit at the expense of classification performance. Although rule-based predictors [7]

and risk scores derived from logistic regression models [8] have been effective to aid decision making in clinical practice and indeed have performance levels that are competitive even against modern approaches such as deep learning [9] there are significant shortcomings. In order to cope with non-linear dependence on clinical attributes with linear models, input variables are frequently discretised. An example of this would be to group age intervals into multiple categories. However, if age bands are for instance by decades, this would treat someone aged 39 as more similar to a 30-year-old than to a 40-year-old. Discretisation will mask variation within each group and, furthermore, it can lead to considerable loss of power and residual confounding [10].

One way to manage non-linearities with interpretable models is to fit a Generalised Additive Model (GAM) estimating the dependence on individual variables with splines [11]. This class of flexible models is in fact a gold standard for interpretability [12]. They are self-explaining [13] and new formulations are emerging which do not require careful tuning of spline parameters but replace them with machine learning modules. In the case of Explainable Boosting Machines [14] the modules are random forests and gradient boosted trees, whereas Neural Additive Models [15] have the structure of a self-explaining neural network. Both are bespoke models and estimate the component functions of the GAM in tandem with inferring an optimal sparse model structure. Along with linear and logistic regression, GAMs lend themselves to practical implementation in the form of nomograms, which are already familiar to clinicians for visualisation of risk scores [16,17].

But what about existing machine learning models?

A key to opening probabilistic black box classifiers without sacrificing predictive performance is an old statistical tool, Analysis of Variance (ANOVA). It is well-known that ANOVA decompositions can express any function as an exact sum of functions of fewer variables, comprising main effects for individual variables together with interaction terms [18]. This is a natural way to derive additive functions with gradually increasing complexity. The derived functions are non-linear and mutually orthogonal, ensuring that the terms involving several variables do not overlap with the information contained in the simpler component functions.

All black box models generate multivariate response functions and hence can be expressed in the form of GAMs using ANOVA. For probabilistic models, this can be applied to the logit of the predicted probabilities. Selecting univariate and bivariate additive terms provides interpretability. The black box is then explained by replacing the original data columns with the ANOVA terms and selecting the most informative components with an appropriate statistical model, such as the Least Absolute Shrinkage and Selection Operator [19].

There are two measures that can be applied in ANOVA, both related to the commonly used partial dependence functions. The Dirac measure corresponds to a cut across the predicted surface and the Lebesgue measure is an average over the same surface, sampled over the training data by setting the values of only the variables in the argument of each component function and sweeping them across their full range. In practice, the main difference between the two measures is a small variation in the models that are selected. This framework is remarkably stable showing that partial dependence functions, normally used only for visualisation, work very well for model selection and are effective for prediction.

Once the black box has been mapped onto a GAM, from there onwards the two measures yield exactly the same component functions. Interestingly, the Shapley additive values, already used in medicine [4] are exactly the terms in the GAM expansion [20].

A natural next step is to replicate the interpretable model derived from the black box by implementing it in the form of a Generalised Additive Neural Network (GANN) also known as a Self-Explaining Neural Network (SENN). This will ensure that the univariate and bivariate component functions can be further optimised given the selected structure. Model refinement is possible by a renewed application of the ANOVA decomposition, this time to separate and orthogonalize the first and second order terms in the GANN/SENN [20] rather than the original MLP. This results in a streamlined model that is optimised to the final sparse structure. A schematic of the model inference process is shown in fig. 1.

Second-order terms appear to be sufficient to achieve strong performance [20] no doubt due to the inherent noise in the data. Moreover, starting with a black box model, the structure and form of the original interpretable model is generally very close to that of the GANN/SENN estimated *de novo* by re-initialising and re-training, as are the predictive performances of the two models [20].

The derived GAMs make clinically plausible predictions for real-world data and buck the performance-transparency trade-off even against deep learning [21]. They solve one of the biggest hurdles for AI by enabling physicians and other end-users to easily interpret the results of the models. Arguably, transparency has arrived for tabular data, setting a new benchmark for the clinical application of flexible classifiers.

References

1. Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv* :1702.08608, 2017.
2. Lisboa, P.J.G., Saralajew, S., Vellido, A., Fernández-Domenech, R. and Villmann, T. The coming of age of interpretable and explainable machine learning models, *Neurocomputing* , Volume 535: 25-39, 2023.
3. Vellido, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput & Applications* , Volume 32: 18069–18083, 2020.
4. Huang, W., Suominen, H., Liu, T., Rice, G., Salomon, C. and Barnard, A.S. Explainable discovery of disease biomarkers: The case of ovarian cancer to illustrate the best practice in machine learning and Shapley analysis, *Journal of Biomedical Informatics* , Volume 141, 2023.
5. Palatnik de Sousa I, Vellasco MMBR, Costa da Silva E. Explainable Artificial Intelligence for Bias Detection in COVID CT-Scan Classifiers. *Sensors* . Volume 21(16):5657, 2021.
6. Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat Mach Intelligence* , *1* : 206–215, 2019.
7. Timmerman D, Van Calster B, et al. Predicting the risk of malignancy in adnexal masses based on the Simple Rules from the International Ovarian Tumor Analysis group, *Am J Obstet Gynecol* , Volume 214:424-437, 2016.
8. Chan T, Bleszynski MS, Buczkowski AK. Evaluation of APACHE-IV Predictive Scoring in Surgical Abdominal Sepsis: A Retrospective Cohort Study. *J Clin Diagn Res.* Volume 10(3), 2016.
9. Christodoulou, E.; Ma, J.; Collins, G.S.; Steyerberg, E.W.; Verbakel, J.Y.; van Calster, B. A Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models. *J. Clin. Epidemiol.,* 110:12–22, 2019.
10. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med.* , Volume 25(1):127-41, 2006.
11. Hastie, T.J. Generalized Additive Models (1st ed.). Routledge, 1990.
12. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D. A Survey of Methods for Explaining Black Box Models.*ACM Comput Surv* , Volume *51* , 2018.
13. Sarle, W.S. Neural Networks and Statistical Models. In Proceedings of the Proceedings of the Nineteenth Annual SAS Users Group International Conference; Cary, NC, 1538–1550, 1994.
14. Nori, H., Jenkins, S., Koch, P. and Caruana, R. InterpretML: A Unified Framework for Machine Learning Interpretability.*arXiv:* 1909.09223, 2019.
15. Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B. Caruana, R. and Hinton, G.E. Neural Additive Models: Interpretable Machine Learning with Neural Nets. *Adv Neural Inf Process Syst* , Volume *6* :4699–4711, 2020.
16. Jiang, E., Guo, H., Yang, B., Li, P., Mishra, P., Yang, T., Li, Y, Wang, H and Jiang, Y. Predicting and comparing postoperative infections in different stratification following PCNL based on nomograms. *Sci Rep* Volume 10, 11337, 2020.
17. Jalali, A., Alvarez-Iglesias, A., Roshan, D. and Newell, J. Visualising statistical models using dynamic nomograms. *PLoS ONE* Volume 14(11): e0225253, 2019.
18. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann Stat* Volume 29(5): 1189-1232, 2001.

19. Tibshirani, R. Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B 58(1), 267–288, 1996.
20. Walters, B., Ortega-Martorell, S., Olier, I. and Lisboa, P.J.G. How to open a black box classifier for tabular data. *Algorithms* , Volume 16, 181, 2023.
21. Lisboa, P.J.G., Jayabalan, M., Ortega-Martorell, S., Olier, I., Medved, D. and Nilsson, J. Enhanced Survival Prediction Using Explainable Artificial Intelligence in Heart Transplantation. *Sci Rep* , Volume 12, 2022.