

InterpretML: A Unified Framework for Machine Learning Interpretability

Harsha Nori
Samuel Jenkins
Paul Koch
Rich Caruana

Microsoft Corporation
1 Microsoft Way
Redmond, WA 98052, USA

HANORI@MICROSOFT.COM
 SAJENKIN@MICROSOFT.COM
 PAULKOCH@MICROSOFT.COM
 RCARUANA@MICROSOFT.COM

Editor:

Abstract

InterpretML is an open-source Python package which exposes machine learning interpretability algorithms to practitioners and researchers. InterpretML exposes two types of interpretability – **glassbox**, which are machine learning models designed for interpretability (ex: linear models, rule lists, generalized additive models), and **blackbox** explainability techniques for explaining existing systems (ex: Partial Dependence, LIME). The package enables practitioners to easily compare interpretability algorithms by exposing multiple methods under a unified API, and by having a built-in, extensible visualization platform. InterpretML also includes the first implementation of the Explainable Boosting Machine, a powerful, interpretable, glassbox model that can be as accurate as many blackbox models. The MIT licensed source code can be downloaded from github.com/microsoft/interpret.

Keywords: Interpretability, Explainable Boosting Machine, Glassbox, Blackbox

1. Introduction

As machine learning has matured into wide-spread adoption, building models that users can understand is becoming increasingly important. This can easily be observed in high-risk applications such as healthcare (Ahmad et al., 2018; Caruana et al., 2015), finance (Hajek, 2019; Chen et al., 2018) and judicial environments (Tan et al., 2018; Soundarajan and Clausen). Interpretability is also important in general applied machine learning problems such as model debugging, regulatory compliance, and human computer interaction.

We address these needs with InterpretML by exposing many state of the art interpretability algorithms under a unified API. This API covers two major interpretability forms: "glassbox" models, which are inherently intelligible and explainable to the user, and "blackbox" interpretability, methods that generate explanations for any machine learning pipeline, no matter how opaque it is. This is further supported with interactive visualizations and a built-in dashboard designed for interpretability algorithm comparison. InterpretML is MIT licensed, and emphasizes extensibility and compatibility with popular open-source projects such as scikit-learn (Pedregosa et al., 2011) and Jupyter Notebook environments (Kluyver et al., 2016).

2. Package Design

InterpretML follows four key design principles that influence its architecture and API.

Ease of comparison. Make it as easy as possible to compare multiple algorithms. ML interpretability is in its infancy, and many algorithmic approaches have emerged from research, each of which has pros and cons. Comparison is critical to find the algorithm that best suits the users’ needs. InterpretML enables this by enforcing a scikit-learn style uniform API, and providing a visualization platform centered around algorithmic comparison.

Stay true to the source. Use reference algorithms and visualizations as much as possible. Our goal is to expose interpretability algorithms to the world, in their most accurate form.

Play nice with others. Leverage the open-source ecosystem, and don’t reinvent the wheel. InterpretML is highly compatible with popular projects like Jupyter Notebook and scikit-learn, and builds off of many libraries like plotly, lime, shap, and SALib.

Take what you want. Use and extend any component of InterpretML without pulling in the whole framework. For example, it’s possible to produce a computationally intensive explanation on a server, without InterpretML’s visualization and its related dependencies.

The code architecture and unified API is best expressed in Figure 1, providing an overview and relevant example code.

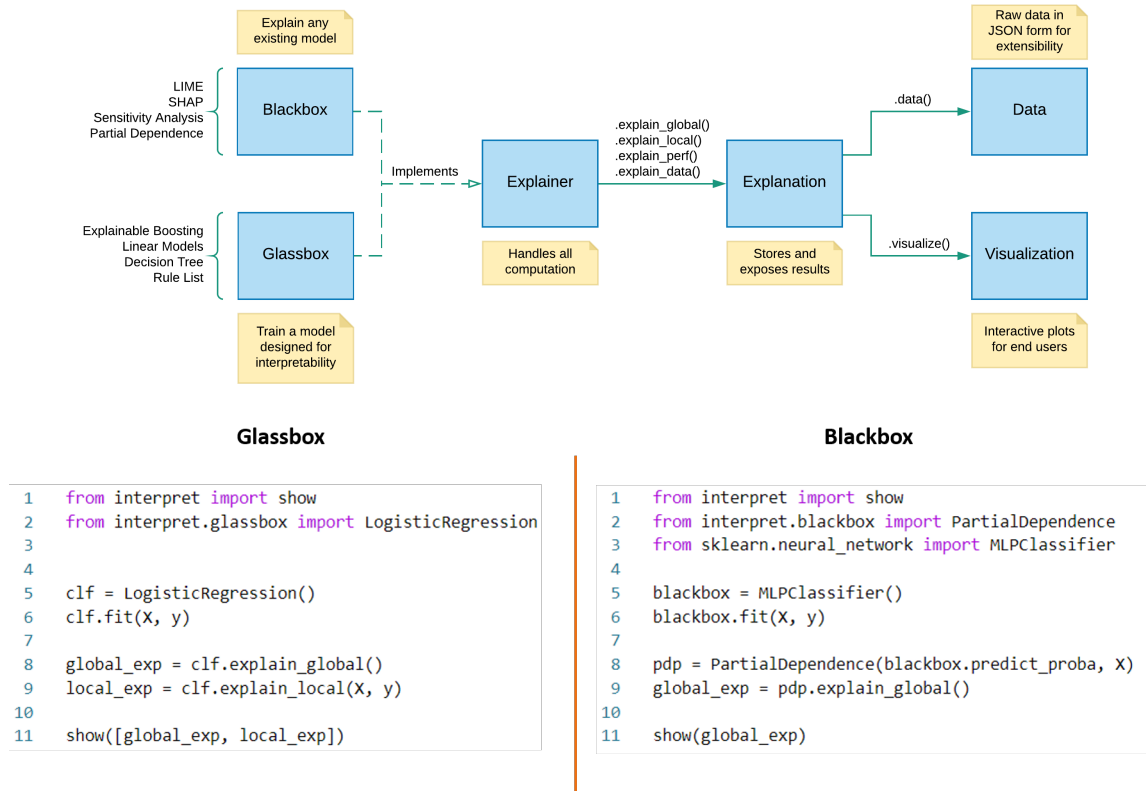


Figure 1: API architecture and code examples

3. Explainable Boosting Machine

As part of the framework, InterpretML also includes a new interpretability algorithm – the Explainable Boosting Machine (EBM). EBM is a glassbox model, designed to have accuracy comparable to state-of-the-art machine learning methods like Random Forest and Boosted Trees, while being highly intelligible and explainable. EBM is a generalized additive model (GAM) of the form:

$$g(E[y]) = \beta_0 + \sum f_j(x_j)$$

where g is the link function that adapts the GAM to different settings such as regression or classification. EBM has a few major improvements over traditional GAMs (Hastie and Tibshirani, 1987). First, EBM learns each feature function f_j using modern machine learning techniques such as bagging and gradient boosting. The boosting procedure is carefully restricted to train on one feature at a time in round-robin fashion using a very low learning rate so that feature order does not matter. It round-robin cycles through features to mitigate the effects of co-linearity and to learn the best feature function f_j for each feature to show how each feature contributes to the model’s prediction for the problem. Second, EBM can automatically detect and include pairwise interaction terms of the form:

$$g(E[y]) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j)$$

which further increases accuracy while maintaining intelligibility. EBM is a fast implementation of the GA²M algorithm (Lou et al., 2013), written in C++ and Python. The implementation is parallelizable, and takes advantage of `joblib` to provide multi-core and multi-machine parallelization. The algorithmic details for the training procedure, selection of pairwise interaction terms, and case studies can be found in (Lou et al., 2012, 2013; Caruana et al., 2015).

EBMs are highly intelligible, because the contribution of each feature to a final prediction can be visualized and understood by plotting f_j . Because EBM is an additive model, each feature contributes to predictions in a modular way that makes it easy to reason about the contribution of each feature to the prediction.

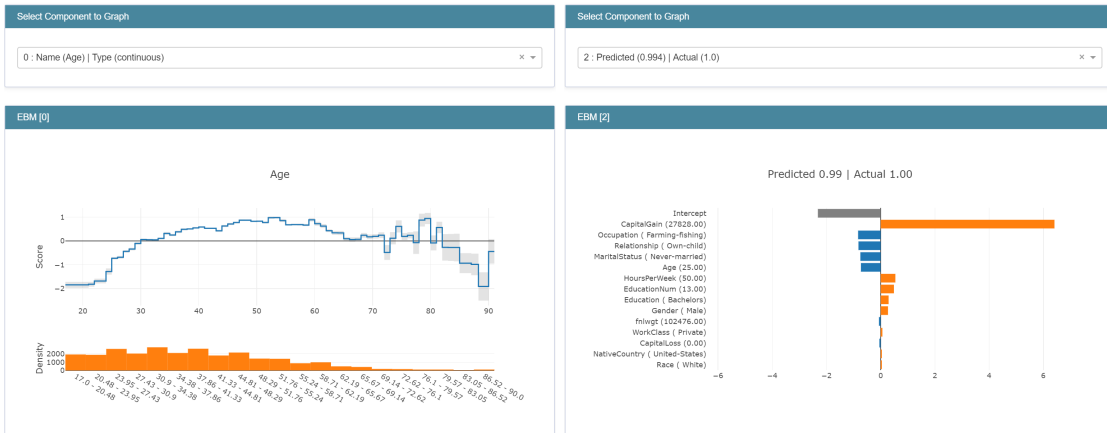


Figure 2: Left: Function f_{Age} . As 'Age' increases from 20 to 50, $p(1)$ increases significantly. Right: Individual prediction, where the 'CapitalGain' feature dominates.

To make individual predictions, each function f_j acts as a lookup table per feature, and returns a term contribution. These term contributions are simply added up, and passed through the link function g to compute the final prediction. Because of the modularity (additivity), term contributions can be sorted and visualized to show which features had the most impact on any individual prediction.

Classification Performance (AUROC)					
Model	heart-disease (303, 13)	breast-cancer (569, 30)	telecom-churn (7043, 19)	adult-income (32561, 14)	credit-fraud (284807, 30)
EBM	0.916	0.995	0.851	0.928	0.975
LightGBM	0.864	0.992	0.835	0.928	0.685
Logistic Regression	0.895	0.995	0.804	0.907	0.979
Random Forest	0.89	0.992	0.824	0.903	0.95
XGBoost	0.87	0.995	0.85	0.922	0.981

Figure 3: Classification performance for models across datasets (rows, columns).



Figure 4: Computational performance for models across datasets (rows, columns).

In terms of predictive power, EBM often performs surprisingly well, and is comparable with state of the art methods like Random Forest and XGBoost.¹ To keep the individual terms additive, EBM pays an additional training cost, making it somewhat slower than similar methods. However, because making predictions involves simple additions and lookups inside of the feature functions f_j , EBMs are one of the fastest models to execute at prediction time. EBM’s light memory usage and fast predict times makes it particularly attractive for model deployment in production.

Acknowledgments

We would like to acknowledge everyone on our acknowledgements.md file for their support on this project. We also depend on many amazing software packages and research: scikit-learn

1. All models were trained with their default parameters. EBM’s current default parameters are chosen for computational speed, to enable ease of experimentation. For the best accuracy and interpretability, we recommend using reference parameters: 100 inner bags, 100 outer bags, 5000 epochs, and a learning rate of 0.01.

(Pedregosa et al., 2011), plotly (Plotly, 2015), lime (Ribeiro et al., 2016), shap (Lundberg and Lee, 2017), SALib (Herman and Usher, 2017), partial dependence (Friedman, 2001), Jupyter (Kluyver et al., 2016), pandas (McKinney), and more.

References

- Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 559–560. ACM, 2018.
- Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL <https://doi.org/10.1023/A:1010933404324>.
- Rich Caruana, Paul Koch, Yin Lou, Marc Sturm, Johannes Gehrke, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *KDD’15, August 10-13, 2015, Sydney, NSW, Australia*. ACM, August 2015. ISBN 978-1-4503-3664-2/15/08. URL <https://www.microsoft.com/en-us/research/publication/intelligible-models-healthcare-predicting-pneumonia-risk-hospital-30-day-readmission/>.
- Chaofan Chen, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. An interpretable model with globally consistent explanations for credit risk. *arXiv preprint arXiv:1811.12615*, 2018.
- Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL <http://doi.acm.org/10.1145/2939672.2939785>.
- Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Ann. Statist.*, 29(5):1189–1232, 10 2001. doi: 10.1214/aos/1013203451. URL <https://doi.org/10.1214/aos/1013203451>.
- Petr Hajek. Interpretable fuzzy rule-based systems for detecting financial statement fraud. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 425–436. Springer, 2019.
- Trevor Hastie and Robert Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.
- Jon Herman and Will Usher. SALib: An open-source python library for sensitivity analysis. *The Journal of Open Source Software*, 2(9), jan 2017. doi: 10.21105/joss.00097. URL <https://doi.org/10.21105/joss.00097>.
- Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, pages 579:1–579:13, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-5970-2. doi: 10.1145/3290605.3300809. URL <http://doi.acm.org/10.1145/3290605.3300809>.

- Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, and Carol Willing. Jupyter notebooks – a publishing format for reproducible computational workflows. In F. Loizides and B. Schmidt, editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87 – 90. IOS Press, 2016.
- Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, pages 150–158, 2012. doi: 10.1145/2339530.2339556. URL <https://doi.org/10.1145/2339530.2339556>.
- Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, pages 623–631, 2013. doi: 10.1145/2487575.2487579. URL <https://doi.org/10.1145/2487575.2487579>.
- Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>.
- Wes McKinney. pandas: a foundational python library for data analysis and statistics.
- 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, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Plotly. Collaborative data science, 2015. URL <https://plot.ly>.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.
- Sucheta Soundarajan and Daniel L Clausen. Equal protection under the algorithm: A legal-inspired framework for identifying discrimination in machine learning.
- Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES '18*, pages 303–310, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-6012-8. doi: 10.1145/3278721.3278725. URL <http://doi.acm.org/10.1145/3278721.3278725>.

Xuezhou Zhang, Sarah Tan, Paul Koch, Yin Lou, Urszula Chajewska, and Rich Caruana. Interpretability is harder in the multiclass setting: Axiomatic interpretability for multiclass additive models. *CoRR*, abs/1810.09092, 2018. URL <http://arxiv.org/abs/1810.09092>.