A Discussion of Model Explanation Tools with Practical Recommendations

Patrick Hall PHALL@H2O.AI
H2O.ai, Mountain View, CA

Abstract

This paper discusses several explanatory methods that go beyond the error measurements and plots traditionally used to assess machine learning models. The approaches, decision tree surrogate models, individual conditional expectation (ICE) plots, local interpretable model-agnostic explanations (LIME), partial dependence plots, and Shapley explanations, vary in terms of scope, fidelity, and suitable application domain. Along with descriptions of these methods, practical guidance for usage and in-depth software examples are also presented.

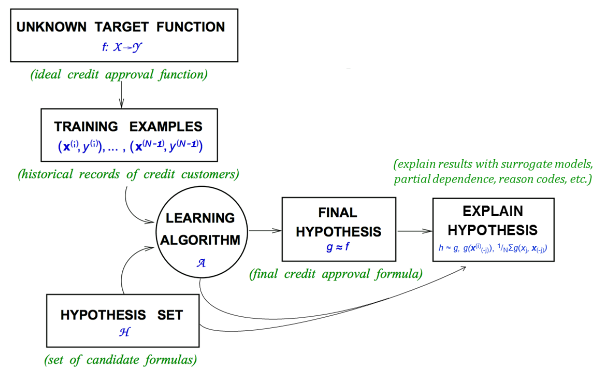

Figure 1. An augmented learning problem diagram in which several techniques create explanations for a credit scoring model. Adapted from **Learning From Data** (Abu-Mostafa et al., 2012).

1. Introduction

Interpretability of statistical and machine learning models is a multifaceted, complex, and evolving subject. This paper focuses mostly on just one aspect of model interpretability: explaining the mechanisms and predictions of models trained using supervised decision tree ensemble algorithms, like gradient boosting machines (GBMs) and random forests. Others have defined key terms and put forward general motivations for better intepretability of machine learning models (Lipton, 2016), (Doshi-Velez & Kim, 2017), (Gilpin et al., 2018), (Guidotti et al., 2018). Following Doshi-Velez and Kim (2017), this discussion uses "the ability to explain or to present in understandable terms to a human," as the definition of *interpretable*. "When you can no longer keep asking why," will serve as the working definition for a *good explanation* of model mechanisms or predictions (Gilpin et al., 2018).

As in Figure 1, the presented explanatory methods help practicioners make random forests, GBMs, and other types of popular supervised machine learning models more interpretable by enabling post-hoc explanations that are suitable for:

- Facilitating regulatory compliance.
- Understanding or debugging model mechanisms and predictions.
- Preventing or debugging accidental or intentional discrimination in model predictions.
- Preventing or debugging malicious hacking of models or adversarial attacks on models.

Detailed discussions of the explanatory methods begin below by defining notation. Then sections 3 – 6 discuss explanatory methods and present recommendations for each method. Section 7 presents some general interpretability recommendations for practicioners. Section 8 discusses several additional interpretability subjects that are likely important for practicioners, and finally, section 9 highlights a few software resources that accompany this paper.

2. Notation

To facilitate technical descriptions of explanatory techniques, notation for input and output spaces, for datasets, and for models is defined.

2.1. Spaces

- Input features come from a set $\mathcal{X}$ contained in a P-dimensional input space, $\mathcal{X} \subset \mathbb{R}^P$.
- Known labels corresponding to instances of $\mathcal{X}$ come from the set $\mathcal{Y}$ and are contained in a C-dimensional label space, $\mathcal{Y} \subset \mathbb{R}^C$.
- Learned output responses come from a set $\hat{\mathcal{Y}}$. For regression models, the set $\hat{\mathcal{Y}}_r$ is also contained in a C-dimensional output space, $\hat{\mathcal{Y}}_r \subset \mathbb{R}^{C_r}$. For classification models, the set $\hat{\mathcal{Y}}_c$ typically contains a column vector for each unique class in $\mathcal{Y}$. Hence, $\hat{\mathcal{Y}}_c$ is contained in a C'-dimensional output space, $\hat{\mathcal{Y}}_c \subset \mathbb{R}^{C'_c}$.

2.2. Datasets

- The input dataset $\mathbf{X}$ is composed of observed instances of the set $\mathcal{X}$ with a corresponding dataset of labels $\mathbf{Y}$, observed instances of the set $\mathcal{Y}$.
- Each i-th observation of $\mathbf{X}$ is denoted as $\mathbf{x}^{(i)} = [x_0^{(i)}, x_1^{(i)}, \dots, x_{P-1}^{(i)}]$, with corresponding i-th labels in $\mathbf{Y}, \mathbf{y}^{(i)} = [y_0^{(i)}, y_1^{(i)}, \dots, y_{C-1}^{(i)}]$.
- $\mathbf{X}$ and $\mathbf{Y}$ consists of N tuples of observations: $[(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}), (\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \dots, (\mathbf{x}^{(N-1)}, \mathbf{y}^{(N-1)})]$, $\mathbf{x}^{(i)} \in \mathcal{X}$, $\mathbf{y}^{(i)} \in \mathcal{Y}$.
- Each j-th input column vector of $\mathbf{X}$ is denoted as $X_j = [x_j^{(0)}, x_j^{(1)}, \dots, x_j^{(N-1)}]^T$.

2.3. Models

- A machine learning model g, selected from a hypothesis set $\mathcal{H}$, is trained to represent an unknown target function f observed as $\mathbf{X}$ with labels $\mathbf{Y}$ using a training algorithm $\mathcal{A}$: $\mathbf{X}, \mathbf{Y} \xrightarrow{\mathcal{A}} g$.
- g generates learned output responses on the input dataset $g(\mathbf{X}) = \hat{\mathbf{Y}}$, and on the general input space $g(\mathcal{X}) = \hat{\mathcal{Y}}$.
- The model to be explained is denoted as g.

3. Surrogate Decision Trees

- Given a learned function g and set of predictions $g(\mathbf{X})$, a surrogate DT can be trained: $\mathbf{X}, g(\mathbf{X}) \xrightarrow{\mathcal{A}_{\text{surrogate}}} h_{\text{tree}}$.
- h_{tree} displays a low-fidelity, high-interpretability flow chart of g's decision making process, and important features and interactions in g.

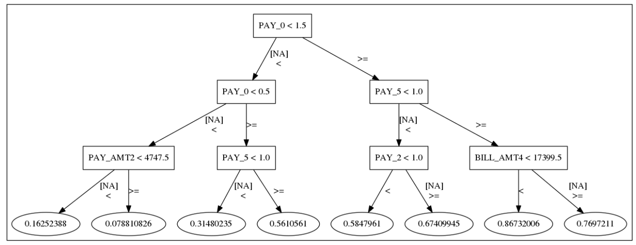

Figure 2. h_{tree} for Taiwanese credit card data (Lichman, 2013), and for machine-learned GBM response function g.

- Always use error measures to assess the trustworthiness of h_{tree}.
- Prescribed methods (Craven & Shavlik, 1996); (Bastani et al., 2017) for training h_{tree} do exist. In practice, straightforward cross-validation approaches are typically sufficient.
- Comparing cross-validated training error to traditional training error can give an indication of the stability of the single tree model, h_{tree}.
- (Hu et al., 2018) use local linear surrogate models, h_{GLM}, in h_{tree} leaf nodes to increase overall surrogate model fidelity while also retaining a high degree of interpretability.

4. Partial Dependence and Individual Conditional Expectation (ICE) plots

- Following (Friedman et al., 2001) a single feature $X_j \in \mathbf{X}$ and its complement set $X_{(-j)} \in \mathbf{X}$ (where $X_j \cup X_{(-j)} = \mathbf{X}$) is considered.
- PD(X_j, g) for a given feature X_j is estimated as the average output of the learned function g when all the components of X_j are set to a constant $x \in \mathcal{X}$ and $X_{(-j)}$ is left untouched.
- ICE$(X_j, \mathbf{x}^{(i)}, g)$ for a given observation $\mathbf{x}^{(i)}$ and feature X_j is estimated as the output of the learned function g when $x_j^{(i)}$ is set to a constant $x \in \mathcal{X}$ and $\mathbf{x}^{(i)} \in X_{(-j)}$ are left untouched.
- PD and ICE curves are usually plotted over some set of interesting constants $x \in \mathcal{X}$.

Overlaying PD and ICE curves is a succinct method for describing global and local prediction behavior and can be used to detect interactions. (Goldstein et al., 2015)

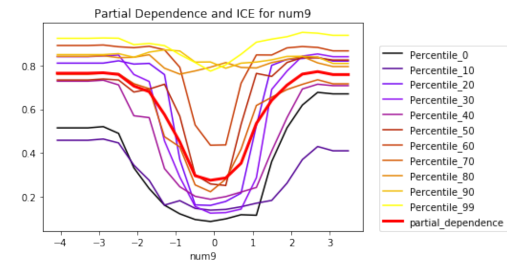

Figure 3. PD and ICE curves for $X_j = \text{num}_9$, for known signal generating function $f(\mathbf{X}) = \text{num}_1 * \text{num}_4 + |\text{num}_8| * \text{num}_9^2 + e$, and for machine-learned GBM response function g.

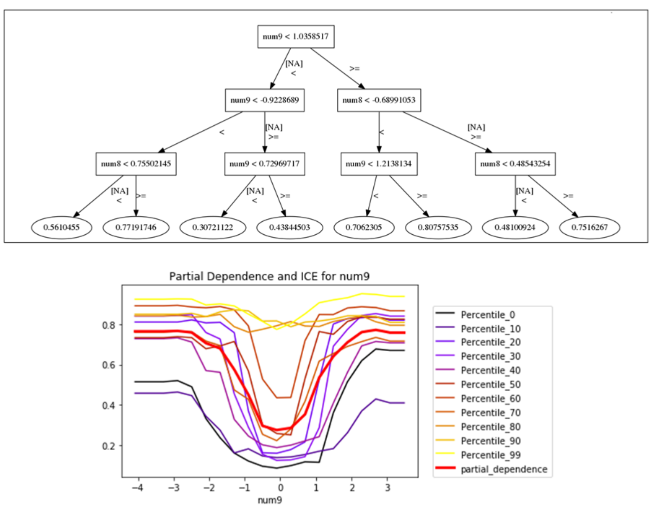

Figure 4. Surrogate DT, PD, and ICE curves for $X_j = \text{num}_9$, for known signal generating function $f(\mathbf{X}) = \text{num}_1 * \text{num}_4 + |\text{num}_8| * \text{num}_9^2 + e$, and for machine-learned GBM response function g.

Combining Surrogate DT models with PD and ICE curves is a convenient method for detecting, confirming, and understanding important interactions.

5. Local Interpretable Model-agnostic Explanations (LIME)

(Ribeiro et al., 2016) defines LIME for some observation $\mathbf{x} \in \mathcal{X}$:

$$\arg\max_{h \in \mathcal{H}} \mathcal{L}(g, h, \pi_{\mathbf{X}}) + \Omega(h) \tag{1}$$

Here g is the function to be explained, h is an interpretable surrogate model of g, often a linear model h_{GLM}, $\pi_{\mathbf{X}}$ is a weighting function over the domain of g, and $\Omega(h)$ limits the complexity of h.

Typically, h_{GLM} is constructed such that

$$\mathbf{X}^{(*)}, g(X^{(*)}) \xrightarrow{\mathcal{A}_{\text{surrogate}}} h_{\text{GLM}} \tag{2}$$

where $\mathbf{X}^{(*)}$ is a generated sample, $\pi_{\mathbf{X}}$ weighs $\mathbf{X}^{(*)}$ samples by their Euclidean similarity to $\mathbf{x}$, local feature importance is estimated using $\beta_j x_j$, and L_1 regularization is used to induce a simplified, sparse h_{GLM}.

- LIME is ideal for creating low-fidelity, highly interpretable explanations for non-DT models and for neural network models trained on unstructured data, e.g. deep learning.
- Always use regression fit measures to assess the trustworthiness of LIME explanations.
- LIME can be difficult to deploy, but there are highly deployable variants. (Hu et al., 2018); (Hall et al., 2017)
- Local feature importance values are offsets from a local intercept.
  - Note that the intercept in LIME can account for the most important local phenomena.
  - Generated LIME samples can contain large proportions of out-of-range data that can lead to unrealistic intercept values.
- To increase the fidelity of LIME explanations, try LIME on discretized input features and on manually constructed interactions.
- Use cross-validation to construct standard deviations or even confidence intervals for local feature importance values.
- LIME can fail, particularly in the presence of extreme nonlinearity or high-degree interactions.

6. Tree Shap

Shapley explanations are a class of additive, consistent local feature importance measures with long-standing theoretical support, (Lundberg & Lee, 2017). For some observation $\mathbf{x} \in \mathcal{X}$, Shapley explanations take the form:

$$\phi_0 + \sum_{j=0}^{j=\mathcal{P}-1} \phi_j \mathbf{x}'_j \tag{3}$$

Here $\mathbf{x}' \in \{0,1\}^{\mathcal{P}}$ is a binary representation of $\mathbf{x}$ where 0 indicates missingness. Each ϕ_j is the local feature importance value associated with x_j.

- Calculating Shapley values directly is typically infeasible, but they can be estimated in different ways.
- Tree Shap is a specific implementation of Shapley explanations that leverages DT structures to disaggregate the contribution of each x_j to $g(\mathbf{x})$ in a DT or DT-based ensemble model. (Lundberg et al., 2018)
- Tree Shap is ideal for high-fidelity explanations of DT-based models, perhaps even in regulated applications.
- Local feature importance values are offsets from a global intercept.
- LIME can be constrained to become Shapley explanations, i.e. kernel shap.
- A similar, popular method known as *treeinterpreter* appears untrustworthy when applied to GBM models.

7. General Recommendations

- Monotonically constrained XGBoost, Surrogate DT, PD and ICE plots, and Tree Shap are a direct and open source way to create an interpretable nonlinear model.
- Global and local explanatory techniques are often necessary to explain a model.
- Use simpler low-fidelity or sparse explanations to understand more accurate and complex high-fidelity explanations.
- Seek consistent results across multiple explanatory techniques.
- Methods relying on generated data are sometimes unpalatable to users. They want to understand *their* data.
- Surrogate models can provide low-fidelity explanations for model mechanisms in original feature spaces if g is defined to include feature extraction or engineering.
- To increase adoption, production deployment of explanatory methods must be straightforward.

8. Suggested Reading

- xNN derivatives, neural net-specific methods.
- Accurate and interpretable classifiers.
- Fairness.

9. Software Resources

Comparison of Explanatory Techniques on Simulated Data:
https://github.com/h2oai/mli-resources/tree/master/lime_shap_treeint_compare

In-depth Explanatory Technique Examples:
https://github.com/jphall663/interpretable_machine_learning_with_python

"Awesome" Machine Learning Interpretability Resource List:
https://github.com/jphall663/awesome-machine-learning-interpretability

10. Acknowledgements

References

Abu-Mostafa, Yaser S., Magdon-Ismail, Malik, and Lin, Hsuan-Tien. ***Learning from Data***. AMLBook, New York, 2012. URL `https://work.caltech.edu/textbook.html`.

Bastani, Osbert, Kim, Carolyn, and Bastani, Hamsa. Interpreting blackbox models via model extraction. *arXiv preprint arXiv:1705.08504*, 2017. URL `https://arxiv.org/pdf/1705.08504.pdf`.

Craven, Mark W. and Shavlik, Jude W. Extracting tree-structured representations of trained networks. *Advances in Neural Information Processing Systems*, 1996. URL `http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks.pdf`.

Doshi-Velez, Finale and Kim, Been. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. URL `https://arxiv.org/pdf/1702.08608.pdf`.

Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. ***The Elements of Statistical Learning***. Springer, New York, 2001. URL `https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf`.

Gilpin, Leilani H, Bau, David, Yuan, Ben Z., Bajwa, Ayesha, Specter, Michael, and Kagal, Lalana. Explaining explanations: An approach to evaluating interpretability of machine learning. *arXiv preprint arXiv:1806.00069*, 2018. URL `https://arxiv.org/pdf/1806.00069.pdf`.

Goldstein, Alex, Kapelner, Adam, Bleich, Justin, and Pitkin, Emil. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1), 2015.

Guidotti, Riccardo, Monreale, Anna, Turini, Franco, Pedreschi, Dino, and Giannotti, Fosca. A survey of methods for explaining black box models. *arXiv preprint arXiv:1802.01933*, 2018. URL https://arxiv.org/pdf/1802.01933.pdf.

Hall, Patrick, Gill, Navdeep, Kurka, Megan, and Phan, Wen. Machine learning interpretability with h2o driverless ai, 2017. URL http://docs.h2o.ai/driverless-ai/latest-stable/docs/booklets/MLIBooklet.pdf.

Hu, Linwei, Chen, Jie, Nair, Vijayan N., and Sudjianto, Agus. Locally interpretable models and effects based on supervised partitioning (lime-sup). *arXiv preprint arXiv:1806.00663*, 2018. URL https://arxiv.org/ftp/arxiv/papers/1806/1806.00663.pdf.

Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Lipton, Zachary C. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016. URL https://arxiv.org/pdf/1606.03490.pdf.

Lundberg, Scott M and Lee, Su-In. A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Lundberg, Scott M, Erion, Gabriel G, and Lee, Su-In. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018. URL https://arxiv.org/pdf/1706.06060.pdf.

Ribeiro, Marco Tulio, Singh, Sameer, and Guestrin, Carlos. 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*, pp. 1135–1144. ACM, 2016. URL http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf.