Michael Chromik (2020): reSHAPe: A Framework for Interactive Explanations in XAI Based on SHAP. In: Proceedings of the 18th European Conference on Computer-Supported Cooperative Work: The International Venue on Practice-centred Computing on the Design of Cooperation Technologies - Exploratory Papers, Reports of the European Society for Socially Embedded Technologies (ISSN 2510-2591), DOI: 10.18420/ecscw2020_p06

reSHAPe: A Framework for Interactive Explanations in XAI Based on SHAP

Michael Chromik
LMU Munich
michael.chromik@ifi.lmu.de

Abstract. The interdisciplinary field of explainable artificial intelligence (XAI) aims to foster human understanding of black-box machine learning models through explanation-generating methods. In this paper, we describe the need for interactive explanation facilities for end-users in XAI. We believe that interactive explanation facilities that provide multiple layers of customizable explanations offer promising directions for empowering humans to practically understand model behavior and limitations. We outline a web-based UI framework for developing interactive explanations based on SHAP.

Introduction

We have witnessed the widespread adoption of intelligent systems into many contexts of our lives. The perception of intelligence often results from their black-box behavior, which may manifest itself in two ways: either from complex machine learning (ML) architectures, as with deep neural networks, or from proprietary models that may intrinsically be white-boxes, but are out of the user's control (Rudin, 2019). As such black-box systems are introduced into more sensitive

contexts, there is a growing call by society that they need to be capable of explaining their behavior in human-understandable terms.

Much research is conducted in the growing fields of *interpretable machine learning (IML)* and *explainable artificial intelligence (XAI)* to foster human understanding. IML often refers to research on models and algorithms that are considered as inherently interpretable while XAI typically refers to the generation of (*post-hoc*) explanations for black-box models to make those systems comprehensible (Rudin, 2019; Biran and Cotton, 2017). Current XAI research mostly focuses on the cognitive process of explanation, i.e., identifying likely root causes of a particular event (Miller, 2018). As a result of this cognitive process, some notions of explanation, such as texts, annotations, or super-pixels, are generated that approximate the model's underlying prediction process.

We believe that an important aspect required to address the call for *"usable, practical and effective transparency that works for and benefits people"* (Abdul et al., 2018) is currently not sufficiently studied: providing users of XAI methods and systems with means of interaction that go beyond a single explanation.

Explanation as an Interactive Dialogue

XAI research often implicitly assumes that there is a single message to be conveyed through an explanation (Abdul et al., 2018). However, in decision-making situations that demand explainability, it is unlikely that a single explanation can address all concerns and questions of a user. This resonates with the social science perspective that considers explanation to be a social process between the *explainer* (sender of an explanation) and the *explainee* (receiver of an explanation) forming a multi-step dialogue between both parties (Miller, 2018). Especially, in situations where people may be held accountable for a particular decision, a user may have multiple follow-up questions before feeling comfortable to trust a system prediction. To model the notion of social explanation between an explanation-generating XAI system and a human decision-maker, we need means of interactivity. Related machine learning approaches, such as explanatory debugging (Kulesza et al., 2015) or interactive machine learning (Dudley and Kristensson, 2018), leverage explanations, interactivity, and human inputs to correct bugs or to improve model performance, respectively.

In our opinion, the social perspective of explanation is currently not sufficiently reflected in current XAI research that addresses decision-making situations. Weld et al. propose seven different follow-up and drill-down operations (Weld and Bansal, 2019). Olah et al. (2018) explore the design space of interpretability interfaces for neural networks and describe possible interaction operations. Recent tools, such as *Google's What-If*, focus primarily on developers and enable them to interactively inspect a ML model with minimal coding. However, they do not provide interactive explanations to end-users of XAI systems.

reSHAPe: Interactive SHAP Explanations

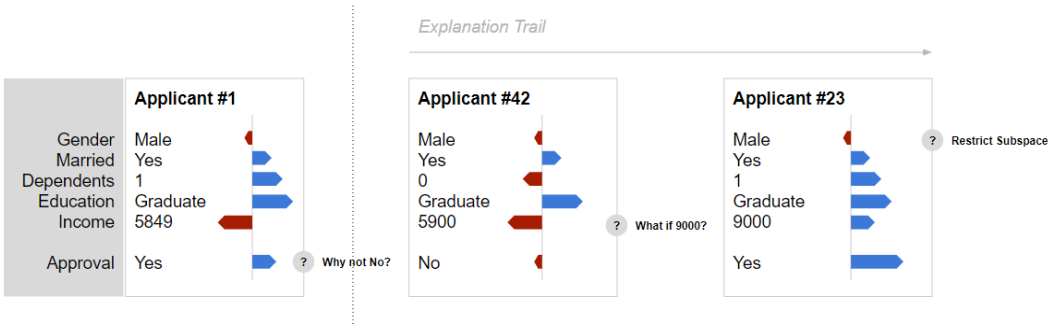

Figure 1. Schematic UI prototype of interactive explanation trail in reSHAPe: The outcome of each observation is explained through SHAP's feature attribution method (red=negative influence on outcome, blue=positive influence). Starting from an initial observation of interest, the user can select one follow-up question from a set of interaction options to validate their hypotheses. Each query returns an illustrative observation and adds it to the explanation trail.

We propose a web-based UI framework that enables developers to provide interactive explanations for end-users. We leverage existing model-agnostic post-hoc explanation-generating methods and integrate them into an interaction concept for navigating between the methods from a human-centered perspective. We build upon the methods provided by the SHAP framework (Lundberg and Lee, 2017). *SHAP (SHapley Additive exPlanations)* is a promising starting point as it unifies existing feature attribution methods (such as *LIME* and *DeepLIFT*) and connects them to additive Shapley values. Furthermore, it allows the generation of *local* and *global* explanations that are consistent with each other as they both use Shapley values as atomic units. This makes them suitable for guiding users through multi-stepped explanations following one line of thought.

However, prior research indicates that even experienced ML engineers have difficulties to use current visualizations of SHAP to effectively verify their hypotheses about an examined ML model (Kaur et al., 2020). Thus, with our framework we address the need for interactive exploration and verification of hypotheses. In a first step, we implement the follow-up operations proposed by Weld and Bansal (2019) for tabular data. From an initial triple of (*input*, *prediction*, *explanation)* provided by an XAI system the user can either:

- **Change the foil**: Contrast the triple with nearest-neighbour triples that resulted in a particularly different prediction to understand *"Why not prediction B?"*.

- **Restrict the subspace:** Request other triples that share the same value for one or more *input* features to understand "*How were similar inputs handled?*".
- **Sensitivity analysis**: Request the minimal changes required to one or more *input* features that result in a different *prediction* and *explanation* to understand *"How stable is the prediction?"*.
- **Explorative perturbation**: Change the values of one or more *input* features of an observation to explore the effects on the *prediction* and its *explanation* and to understand *"What if?"*.
- **Global roll-up**: Contrast the triple's *local explanation* with the *global explanation* of the entire model to understand *"How representative is the observation?"*.

An XAI system with interactive explanations may derive additional information about the user's mental model and preferences from the trail of follow-up interactions. This additional information may be used to establish common ground and potentially improve the overall human-AI system performance. With our framework we aim to support developers with the front-end development of XAI systems for domain experts. We consider domain experts as end-users with a high level of expertise in a particular domain but typically limited expertise in ML, such as lawyers or accountants. We focus on decision-making situations where the domain expert may have concrete or vague hypotheses about the decision problem that guides their explanation needs and interaction.

Future Work

Upcoming research will investigate the potentials of interactive explanations and their evaluation with users in an application context. We collaborate with German chancelleries, lawyers, and a leading software vendor in the sensitive legal domain. We follow a human-centered design process to derive requirements and user needs. Based on these, we iteratively explore design opportunities for usable interactive explanations using prototypes and user studies. We plan to integrate our insights and artifacts in a modular toolkit for creating interactive explanation interfaces for tabular and textual data.

References

Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). ACM, New York, NY, USA, Article 582, 18 pages. https://doi.org/10.1145/3173574.3174156

Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. InIJCAI-17 workshop on explainable AI (XAI), Vol. 8. 1.

John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning.ACM Trans. Interact. Intell. Syst.8, 2, Article 8 (June 2018), 37 pages. https://doi.org/10.1145/3185517

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–14. DOI: https://doi.org/10.1145/3313831.3376219

Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In Proceedings of the 20th International Conference on Intelligent User Interfaces. ACM, 126–137.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In Advances in neural information processing systems, pp. 4765-4774.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence267 (2019), 1 – 38. https://doi.org/10.1016/j.artint.2018.07.007

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. The Building Blocks of Interpretability. Distill(2018). https://doi.org/10.23915/distill.00010 https://distill.pub/2018/building-blocks.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1, 5 (2019), 206–215.

Daniel S. Weld and Gagan Bansal. 2019. The Challenge of Crafting Intelligible Intelligence. Commun. ACM 62, 6 (May 2019), 70–79. https://doi.org/10.1145/3282486