**Interpreting Predictive Models for Human-in-the-Loop Analytics**

*Problem.* Machine learning has revolutionized our ability to use data to inform important decisions in a variety of domains such as online retail, criminal justice, and healthcare. For instance, predictive analytics have been used to price products, set staffing levels, score patient risk, and make bail decisions for defendants. At the same time, machine learning models have been shown to exhibit unexpected defects when trained on real-world observational data, stemming from endogenous explanatory variables or systematic biases in reported outcomes data. In such cases, the predictive model may achieve good out-of-sample accuracy on the original observational dataset, but perform poorly when deployed in the real world. While these issues can sometimes be addressed through the use of carefully chosen instrumental variables, this process first requires understanding the source of confounding or bias, which in turn requires significant domain knowledge. Oftentimes, with the growing number of explanatory variables and the complexity of machine learning, even domain experts may need to examine and understand the learned predictive model before the presence of confounding or bias becomes apparent. Thus, it is critical to involve domain experts in an iterative process of developing the machine learning model to ensure that its predictions are unbiased; we refer to this process as *human-in-the-loop* analytics.

Yet, state-of-the-art machine learning models such as random forests and deep neural nets tend to be *blackbox* in nature [1]; in other words, these models have a complex, opaque structure and tend to use many explanatory variables, making it difficult for humans to understand and verify the model's reasoning process. Thus, one proposed solution for facilitating human-in-the-loop analytics is the use of *interpretable* machine learning models. Examples of previously-proposed interpretable models include sparse linear models [2], rule lists [1], and decision sets [3]. These models are simple and transparent, allowing domain experts to easily understand how predictions are made; with this knowledge, they can identify potential sources of bias or errors in the model by
checking if the model mimics their own reasoning process. However, the constraint of using an interpretable model instead of a blackbox model comes at a significant cost in predictive accuracy, which in turn may result in poor decision-making. Thus, decision-makers are often faced with a tough decision: either (1) use an interpretable model which may produce worse decisions due to poor predictive performance, or (2) use a blackbox model which has strong predictive performance on observational data, but may exhibit unexpected defects upon deployment in the real world.

We propose a third alternative, which is to extract a simple interpretation that approximates a complex blackbox model. We express our interpretation in the form of a decision tree, whose size can be chosen based on the desired strength of the approximation to the blackbox model. Then, the domain expert can restrict her focus on understanding and verifying the extracted decision tree rather than the original blackbox model. As long as the decision tree is a good approximation of the blackbox model, any significant confounding or bias in the blackbox model should translate to the tree. Thus, if the expert validates the tree's reasoning, then we may deploy the high-performing blackbox model with the confidence that it is likely free of significant bias or confounding as well. In concurrent work, [4] extracts global explanations in the form of decision sets; we demonstrate that our strategy produces much more accurate interpretations, enabling experts to understand and validate a larger portion of the blackbox model’s reasoning process.

**Contributions.** We extract simple, accurate decision trees from complex blackbox machine learning models using active learning. We make no assumptions on the structure of the blackbox model, and only require the ability to run it on chosen inputs. We choose decision trees as our interpretations, since they are easy to understand, nonparametric, and can compactly represent complex functions.

**Algorithm:** Decision trees typically achieve poor predictive performance since they easily overfit to data. To overcome this difficulty, we leverage the ability to generate arbitrarily large amounts of training data by sampling new inputs and labeling them using the blackbox model. We propose a
novel algorithm that uses active learning to generate inputs that flow down a given path in the decision tree, and then use these newly generated training points to avoid overfitting.

**Theory:** We prove that by actively sampling a sufficient number of points using our algorithm, our extracted decision tree converges to the exact decision tree. In other words, the estimation error of our extracted decision tree goes to zero asymptotically, implying that our decision tree avoids overfitting the training set. The challenge to establishing this result is that the branches in a greedy decision tree are estimated by maximizing a non-convex objective function. As a result, even very small errors in the estimated objective function can dramatically change its maximizer. Under mild technical conditions, we establish that asymptotically, the estimated objective converges uniformly to the true objective with high probability, and consequently, the maximizer converges as well.

**Evaluation:** We evaluate our algorithm on a random forest to predict diabetes risk on a real electronic medical record dataset. We find that our interpretation is significantly more accurate than several baselines [1-4]. We also conduct a user study demonstrating that humans are able to better reason about our interpretations than state-of-the-art rule lists. Finally, we interview domain experts (physicians) about our diabetes risk prediction model, and describe several insights they derived using our interpretation. Of particular note, the physicians discovered an unexpected causal issue by investigating a subtree in our interpretation; we were able to then verify that this endogeneity indeed existed in our data, underscoring the value of interpretability and human-in-the-loop analytics.

**References:**


