Introduction

# Original

A common criticism of modern machine learning techniques is the lack of interpretability/explainability. This could serve as a barrier to the wide-scale adoption of machine learning, as users are more likely to deploy a model if they can understand why particular decisions are being made. This is discussed for business in [4], for health care in [6], and more broadly in the recently deployed general data protection regulation [26] (GDPR) in the European Union, which prohibits "solely automated decisions, including profiling, which have a legal or similarly significant effect on them". This has started the push for Explainable AI (XAI).

Ethical considerations play a large part in the push towards XAI, and there is no short history toward bias in machine learning techniques. As examples, Sweeney [30] exposed potentially racial bias in the suggestion algorithm used in Google AdSense, and Bolukbasi et al.[3] showed the popular word2vec algorithm can be heavily susceptible to gender bias. XAI, more specifically, interpretable machine learning (IML), can help observe these biases, to then ideally remedy the biases.

The bias can occur from a number of sources, such as the sampling of the training data, uncovering correlative but not causal relationships, or poor selection of feature sets. To uncover these biases, it’s important to understand how the model is making particular decisions. With XAI, the goal is to have the simplest rules possible without sacrificing the performance. Simplicity and performance are often conflicting objectives (motivating the need for XAI, since top performing methods are often complex).

With traditional tree-based methods for XAI, such as decision tree construction, complexity is controlled by early stopping or post-pruning. These approaches suffer from limitations, such as, with early stopping (or pre-pruning), a branch is terminated when no reduction in cross-validation error is noted. However, this may be premature since additional splits further down may have reduced the error drastically (i.e. if the feature becomes more informative with the addition of another be-cause of feature interaction). With post-pruning, leaves are shrunk by replacing parent nodes with the majority class of the leaf. If no increase in error is seen, this process continues until the error increases for each branch. Alternatively, trees are shrunk in a top-down manner, i.e. with cost-complexity pruning. Of course a major drawback to both approaches is since the tree was greedily constructed, a poor split in hindsight cannot be undone, so the pruning is limited in its ability, i.e. pruning is not going to find the optimal tree which balances complexity and accuracy, since it first greedily maximises the accuracy, then attempts to greedily reduce the complexity afterwards.

In contrast, Genetic Programming (GP) [19] is another tree construction method. Rather than trees being constructed greedily in a top-down manner, trees are evolved from a population of candidates. As well as avoiding the need for greedy construction, such population-based methods are ideal for multi-objective optimisation [10] meaning that trees can be constructed which simultaneously optimise seemingly conflicting objectives. However, to the best of our knowledge, GP has not been used for constructing IML models. In this work, a new multi-objective GP-based method for IML is proposed to overcome the aforementioned limitations.

The specific objectives are to:

* Propose a simple, human-readable tree structure which can be used for reconstructing complex predictions and over-comes limitations of existing tree-based methods, such as greedy construction or need for pruning
* Simultaneously maximise the reconstruction ability while minimising the complexity of the trees
* Generate a frontier of trade-off solutions for user selection
* And finally, to evaluate the method against current state-of-the-art approaches on various datasets.

The proposed method is applicable to any black-box (arbitrarily complex) classifier and makes no further assumptions about these models (such as gradient-based, or ability to apply sparsity).

# Condensed

Understanding the decisions behind Machine Learning (ML) techniques is important for their adoption into new areas where an AI’s decisions have greater consequences. This need for interpretability has created the push for explainable AI (XAI).

With XAI the goal is to have the simplest rules possible without sacrificing performance. These two conflicting objectives motivate the need for XAI since top performing models are often far too complex for human interpretation.

Traditional tree-based methods from XAI suffer from limitations, due to the algorithms being greedy, that prevent them from finding optimal trees that balance complexity and accuracy. Using Genetic Programming (GP) we evolve trees instead of constructing them in a top down greedy manner. GP is also ideal for multi-objective optimisation meaning we can simultaneously optimise our conflicting objectives of accuracy and simplicity.

To the best of our knowledge GP has not been used for constructing IML models. In this work we propose a new multi-objective GP-based method for IML. Are objectives are:

* Propose a simple, human-readable tree structure which can be used for reconstructing complex predictions and over-comes limitations of existing tree-based methods, such as greedy construction or need for pruning
* Simultaneously maximise the reconstruction ability while minimising the complexity of the trees
* Generate a frontier of trade-off solutions for user selection
* Evaluate the method against current state-of-the-art approaches on various datasets.

The proposed method is applicable to any black-box (arbitrarily complex) classifier.