6 Conclusions & Future Work

# Original

In this work, a novel model agnostic method for XAI was proposed which utilises model extraction. Multi-objective GP is used to learn a simple and interpretable representation of a complex black-box model, which is often able to effectively reproduce the black-box’s predictions. This new method was compared to existing approaches for model extraction, and was found to offer drastically simpler models, with statistically equivalent test accuracy. The method is also able to handle categorical and continuous features natively, unlike some existing approaches such as Bayesian rule lists. To our best knowledge, this is the first utilisation of multi-objective optimisation in explainable AI, and follows the suggestions in [9] that "a multi-objective approach based on Pareto dominance would be more suitable to sufficiently address this trade-off" (between accuracy and interpretability). We also believe this is the first application of GP for model extraction (i.e. training on the predictions of a black-box model rather than the original labels), and shows a promising direction for future developments.

In future work, we would like to focus on three main areas. Firstly, can the recreation ability of the proposed method be improved without sacrificing simplicity by considering local search techniques for splitting points, groupings for categorical data, or making use of the black-box model for generating additional labelled instances. Secondly, here a basic complexity measure is used for evaluating the simplicity of the models. However, if the goal is human interpretation, it would be ideal to conduct a blind large scale user studies on the resulting models. And finally, related to the previous point, it is possible to guide the evolution of the models based on the human feedback, we foresee these human-in-the-loop type systems being important for XAI, and this is something that could be incorporated into the evolutionary process here by modifying the complexity measure to instead be user determined.

# Condensed

In this work, a novel model (agnostic) extraction method for XAI was proposed. Multi-objective GP is used to learn a simple and interpretable representation of a complex black-box model, which is often able to effectively reproduce the black-box’s predictions. This new method was compared to existing approaches for model extraction, and was found to offer drastically simpler models, with statistically equivalent test accuracy. To our best knowledge, this is the first utilisation of multi-objective optimisation in explainable AI. We also believe this is the first application of GP for model extraction, and shows a promising direction for future developments.

In future work we would like to focus on three main areas. Firstly, can recreation ability be improved without sacrificing simplicity? Secondly, can we find a more suitable measure of complexity to describe human interpretability? And finally, is it possible to guide the evolution of the models based on human feedback?