Optimal Decision Trees for the Algorithm Selection Problem

Balancing Performance and Interpretability

1. Introduction

- The Algorithm Selection Problem (ASP) is a challenge in computer science aimed at predicting the most effective algorithm for a given instance of a computationally difficult problem. These problems occur frequently across various industries, such as finance, healthcare and energy [1].
- Tree-based models are a widely used machine learning model and have been used to solve the ASP [1]. While more complex tree-based models, like random forests, offer good performance, they're often not easily interpretable [2].
- The relation between the performance and interpretability is usually depicted in the literature as shown in figure 1, which shows interpretability as inversely proportional to performance, while this may not be true.
- Optimal decision trees aim to provide the best of both worlds by providing shallow trees which are optimal based on a specified objective function and tree depth. These optimal decision trees offer good performance on out of sample data [3].

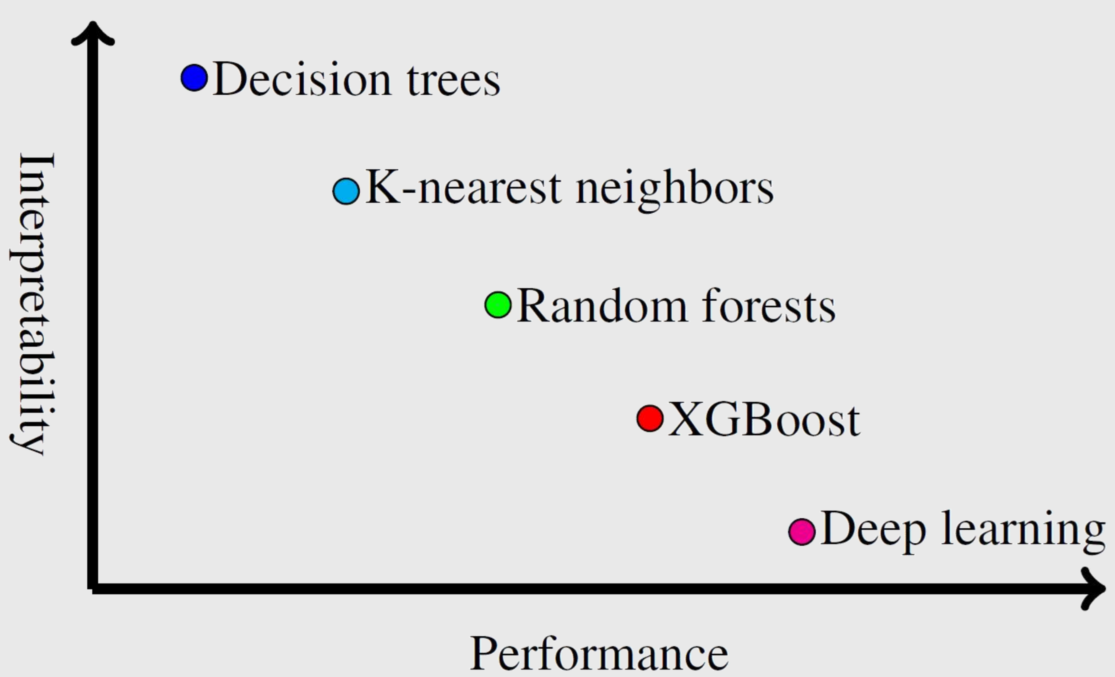


Figure 1: The interpretability - performance trade-off as presented by Sheykhmousa and Mahdianpari [6]

2. Research question

What is the trade-off between performance and interpretability when solving the Algorithm Selection Problem (ASP) with optimal decision trees compared to other tree-based methods such as heuristic trees and random forests?

3. Methodology

Data

Fetch 11 datasets from ASlib [5]

→

Data Preprocessing

Loading the data into a data structure, data cleaning, feature extraction & data binarization such that they're compatible with the machine learning models

→

Model training & selection

Training and selection of (Instance Cost Sensitive) Optimal Decision Trees for which we use STreeD as implementation [4], Decision Tree Classifier / Regressor and Random Forest Classifier / Regressor based on 10-fold cross validation and grid search

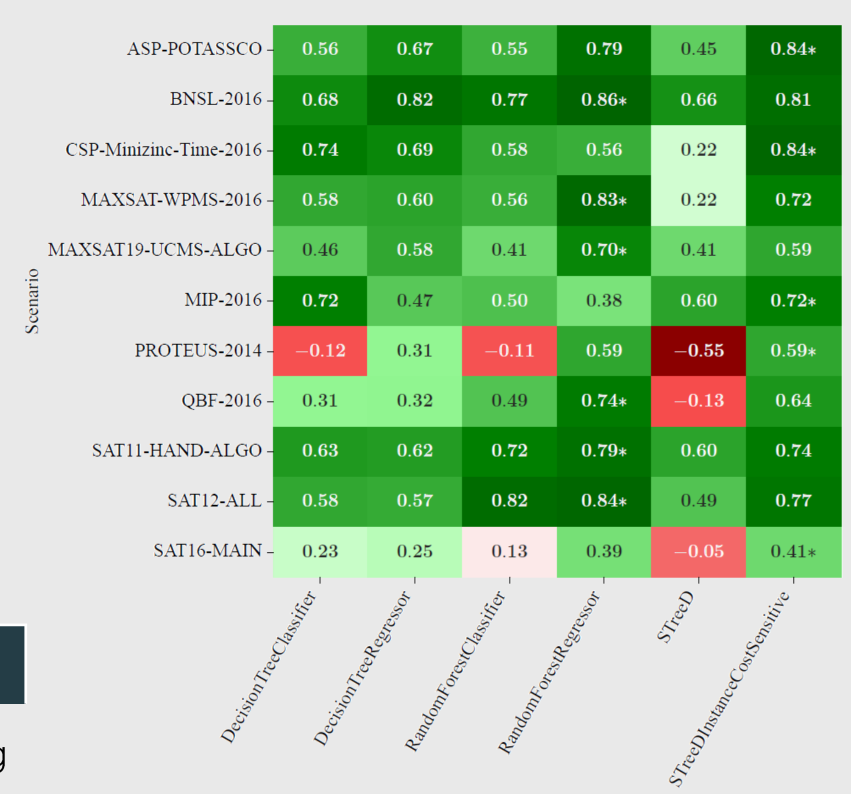
→

Benchmarking

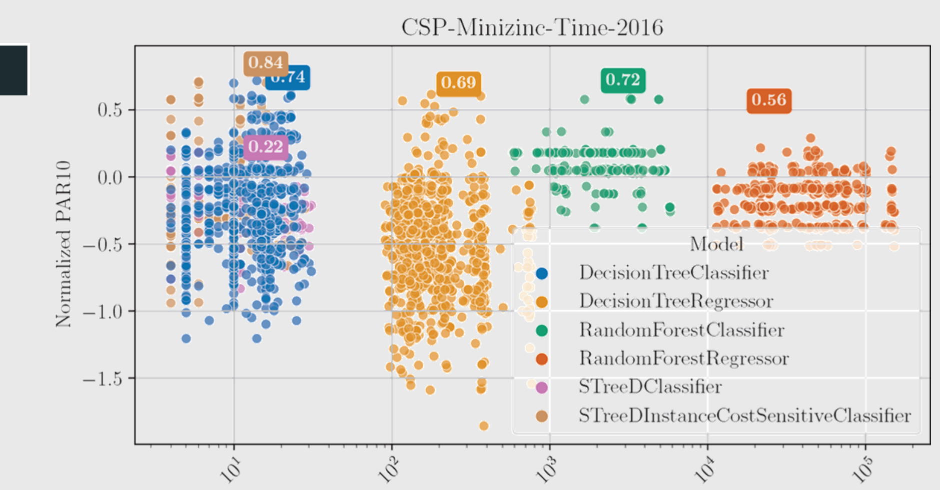
Calculation of the performance (Penalized Average Runtime 10, the average runtime where timeouts are penalized with a factor 10) and interpretability (number of leaf nodes) metrics and visualization of the results

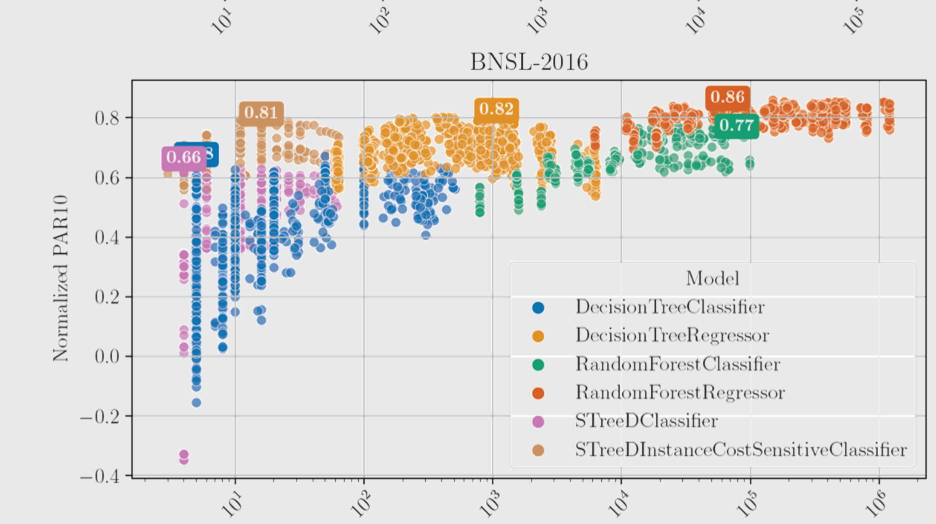
4. Experiments & Conclusions

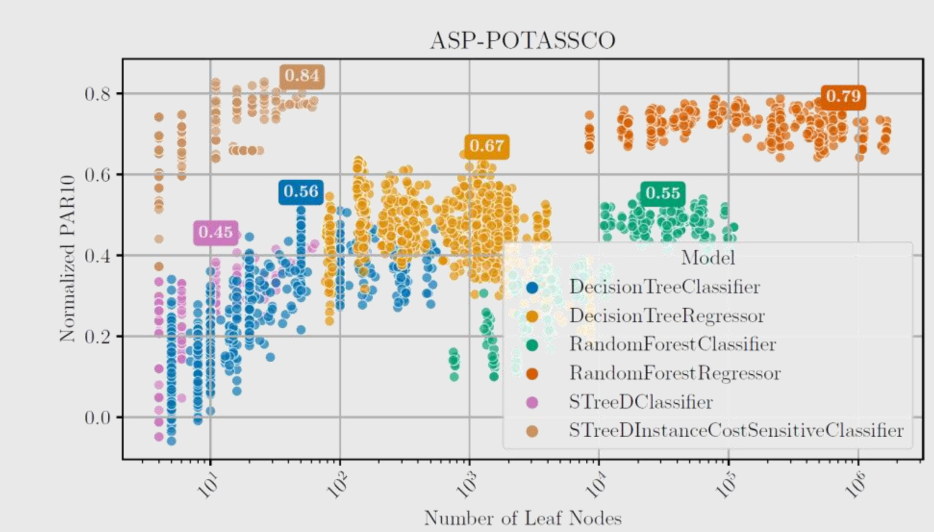
	Mean Normalized PAR10*	Mean # Leaf Nodes	Std. deviation # Leaf Nodes
Decision Tree Classifier	0.49	111.00	123.62
Decision Tree Regressor	0.49	704.36	526.89
Random Forest Classifier	0.51	19168.00	22288.12
Random Forest Regressor	0.69	490705.09	510816.79
STreeD	0.27	14.18	9.85
STreeD Instance Cost Sensitive	0.70	19.64	13.46



Instance cost sensitive STreeD offers comparable or better performance and scales much better in terms of performance / interpretability, meaning there's no apparent interpretability - performance trade-off for Optimal Decision Trees for the ASP when compared to these tree-based models







*Normalized PAR10 is the normalized version of the Penalized Average Runtime 10 (PAR10), and is represented as a fraction of the best possible solver compared to the best solver on average, i.e. a 1 corresponds to a perfect score and a 0 corresponds to no increase in performance

References

1. John R. Rice. "The algorithm selection problem". In: Advances in Computers 15 (1976), pp. 65–118
2. Sicco Verwer and Yingqian Zhang. "Learning optimal classification trees using a binary linear program formulation". In: Proceedings of AAAI. 2019, pp. 1625–1632.
3. Emir Demirović et al. "MurTree: optimal decision trees via Dynamic programming and search". In: J. Mach. Learn. Res. 23.1
4. Jacobus G. M. van der Linden, Mathijs M. de Weerdt, and Emir Demirović. Optimal decision trees for separable objectives: Pushing the limits of dynamic programming, 2023.
5. Bernd Bischl et al. "ASlib: A benchmark library for algorithm selection". In: Artificial Intelligence 237 (2016), pp. 41–58.
6. M. Sheykhmousa et al. "Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, pp. 6308–6325, 2020.

Author: Daniël Poolman
Supervisors: Dr. Emir Demirović, Jacobus G.M. van der Linden
Examiner: Dr. David M.J. Tax