Honesty in Causal Forests, is it worth it ?

CSE 3000 Author: Matej Havelka, Supervisor: Stephan Bongers, Responsible Professor: Jesse Krijthe

1. Path to the research question

- **Causal ML** tries to find the causal relation between treatment **W** and outcome **Y**
- **Causal forests** (CF) splits data into subpopulations, based on these subpopulations it estimates the effect as mean outcome of treated minus mean outcome of not treated
- **Honesty**, defined as double sampled trees [1], fights against overfitting by ensuring no split is done on the same data that also evaluates a leaf node
- How does **Honesty** influence the performance of a **Casual Forest** ?

2. Methodology

Recreate results showing honesty

- Imbalanced dataset 2.5% with treatment effect
- Results show honesty works

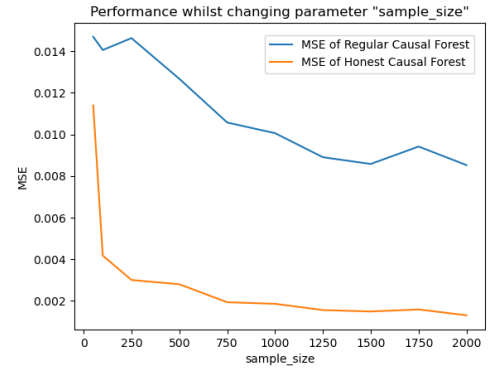
Test out depth issues of honesty

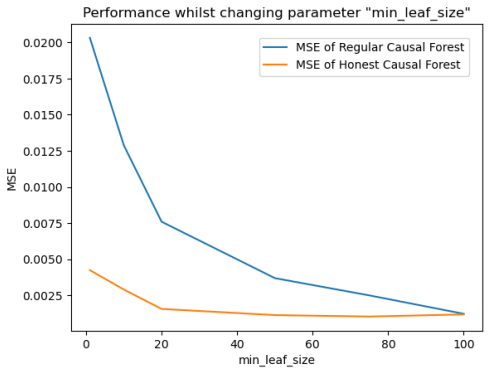
- Spiked dataset around middle
- Results show that honesty has tree depth issues

Test it in a general setting

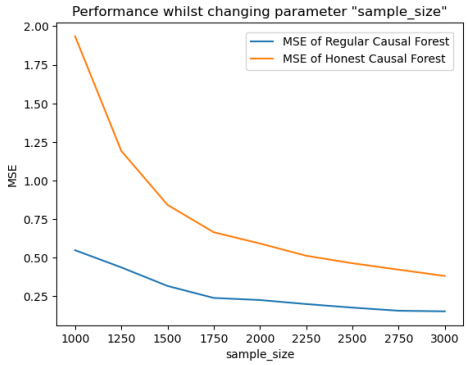
- General
- IHDP and TWINS
- Results show that its all dependent on the sample size
- Some loss in leaf node evaluation

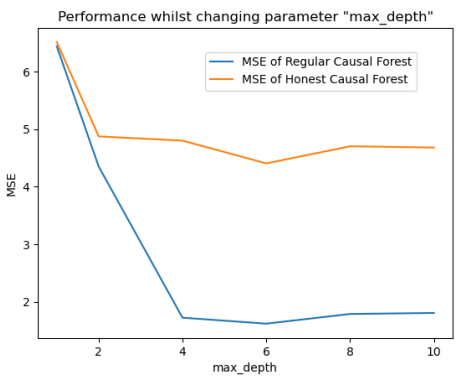
3. Results



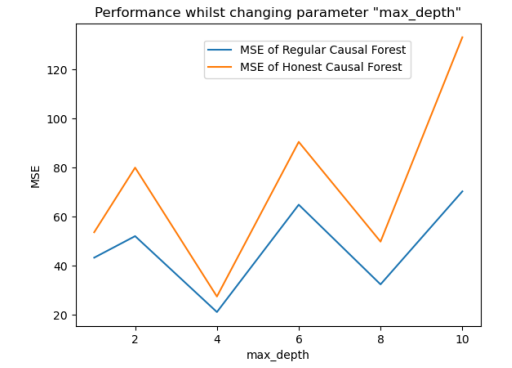


Imbalanced Dataset Experiment





Spiked Dataset Experiment



General Dataset Experiment

4. Conclusions

Honesty:

- helps to fight overfitting and bias
- has problems with depth of trees and leaf sample size on smaller training samples
- With enough data the differences disappear

5. Limitations

Some limitations of the conclusions of my research:

- only one definition of honesty
- comparisons only in terms of performance
- limited set of functions that were tested

6. References

[1] S. Wager and S. Athey, “Estimation and inference of heterogeneous treatment effects using random forests"