Honesty in Causal Forests, is it worth it ?

Matej Havelka
Supervisor(s): Stephan Bongers, Jesse Krijthe
EEMCS, Delft University of Technology, The Netherlands

June 19, 2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

Causal machine learning is a relatively new field which tries to find a causal relation between the treatment and the outcome, rather than a correlation between the features and the outcome. To achieve this, many different models were proposed, one of which is the causal forest. Causal forest is made up of a random forest, with a different estimation function in the leaf node, which means it suffers from the same problems, like being easy to overfit. The reason why honesty was introduced was to ensure mathematically that forests do not overfit as easily. This research however, only provided preliminary results and no real testing was done in terms of causal inference. In this paper three scenarios are tested where a comparison is made between a causal forest with and without honesty. Based on the results it seems that honesty does indeed help for trees to not overfit. However in a general setting it hurts the model as it only trains with half of the available data. This makes honest causal forest less accurate in general settings where there is not a lot of training data. In a setting where a large amount of data is provided it seems that honesty does not change the performance, meaning it creates a theoretical guarantee against overfitting with no repercussions for the performance.

1 Introduction

Causal effect estimation is a growing field, it tries to assign a causal relationship between the treatment W, which is one of the features, and the outcome Y rather than finding an underlying correlation like classical machine learning. This means that the goal of causal effect estimation, also called causal inference, is to figure out how treatment W affects outcome Y, rather than predicting Y itself as classical machine learning does. To achieve this, causal inference is given all data points that received the treatment, called treated, and the ones that didn't, called control, meaning for the purposes of this study, the treatment is a binary variable. Based on that data, a causal model tries to find out what the counterfactual of each data point is, meaning if it's treated, what would the outcome be if it wasn't and vice versa. If a model manages to get the counterfactual, the treatment effect can be represented as the difference between the outcome when treated and the outcome when not treated.

Many models were proposed to find the counterfactual of data points. One such model is the causal forest (CF), proposed by Wager and Athey (2015) and later expanded by Athey et al. (2019). Causal forest was developed with a purpose in mind to create a model that works with heterogeneous data. Heterogeneous data means that the values of the features influence the treatment effect. For example measuring the effect of vaccinations against a disease might be influenced by the age of the participants, as younger people could be more resilient to the disease by default, thus the vaccine would be less effective. Causal forest tries to split the data such that similar data points, based on their outcomes and feature values, land in the same leaf node. These groups of data points that contain some similarities between them are called subpopulations. Based on these subpopulations, causal forest estimates the treatment effect as the difference of the average outcome of the treated and the average outcome of the control.

When causal forests were introduced by Wager and Athey (2015), honesty was introduced by them as well. Honesty is a property of trees that exists outside of the field of causal inference. It is a response to a prevailing problem of regression trees, namely that they easily overfit. Based on Wager and Athey (2015), a tree is honest if it does not use the outcome of a data point to create splits and to evaluate leaf nodes. Intuitively, this means that trees cannot make splits to benefit specific parts of the training dataset, as the evaluation will happen with a different subsample of the dataset. The same study also proposes two implementations of honesty, double sampled trees and propensity trees, which will be explained in more detail in Section 2.2.

Causal forests have already been used on real world problems. For example Miller (2020) focuses on using causal forest to determine the effect of environmental policies on fisheries and Zhang et al. (2022) studies the relation between speed cameras and road accidents, using a causal forest. These studies apply causal forest on real world data, however the studies only take the results from the model, rather than comparing it to different models to see its performance. Some preliminary results were shown by Wager and Athey (2015) which show that indeed honesty is capable of performing better on a dataset on which a causal forest without honesty, called regular causal forest throughout this study, overfits on. Some further results were created in Denil et al. (2014) which compares honest forests with other models, however this study was set in classical setting, rather than a causal one. This begs the question which this paper will try to analyze, which is to study the effect of honesty on the performance of causal forests in a general setting.

In Section 2 a more in depth explanation of causal forests and honesty will be provided. Section 3 will describe the experimental setup, the results obtained by these experiments and the some preliminary theories followed by further experimentation to validate them. Section 4 will provide further discussion about the results with an evaluation on the trade-off between honesty in terms of performance. In Section 5 there is a short discussion on responsible research, what has been done to assure transparency of presented results and the different ethical complications this research might create with its conclusions. Lastly, Section 6 will provide a conclusive stance on honesty based on the experiments and some possible paths where this research can continue.

2 Honesty in Causal Forests

This section provides basic understanding of the causal forests and honesty. Section 2.1 describes the intuition behind the causal forest model and its main goals in causal effect estimation. Section 2.2 provides a description of the honesty property and an algorithm to show how it's implemented.

2.1 Causal Forests

Causal forest (CF) is an extension of a random forest applied in a causal setting. Firstly, there are many definitions of random forests, but CF extends from the generalized random forests (GRF) defined in Athey et al. (2019). As there are different ways to create a forest, this paper uses the implementation from Microsoft-Research (2022), which implements it via gradient boosting, as introduced in Athey et al. (2019). The sole difference between a random forest and a causal forest is the evaluation of the leaf node. While in a classical machine learning scenario forests try to predict the outcome Y, in a causal machine learning scenario the goal is to estimate the treatment effect. Thus, the output of a single tree is the conditional averaged treatment effect (CATE), defined in Equation 1, where $\{i : W_i = 0, X_i \in L\}$ represents a set of all instances in the leaf node L that were not treated. This differs from a random forests which outputs the predicted outcome.

$$\tau(X_i) = \mathbb{E}[Y_1 - Y_0 | X_i] \tag{1}$$

$$\hat{\tau}(x) = \frac{1}{|\{i : W_i = 1, X_i \in L\}|} \sum_{\{i:W_i=1,X_i\in L\}} Y_i - \frac{1}{|\{i : W_i = 0, X_i \in L\}|} \sum_{\{i:W_i=0,X_i\in L\}} Y_i \tag{2}$$

CATE serves as a metric to represent the effect a treatment has on a specific subpopulation, e.g., getting vaccinated might have a bigger effect on older people as younger people might be more resilient by default. To determine CATE, a causal forest creates trees based on different subsamples of the training data. Causal forest uses the exact loss criterion to compute the splits in a tree, while a generalized random forest uses gradient-boosted loss criterion, defined by Athey et al. (2019). This study concentrates on causal forests as an extension of a random forest, thus it is assumed that the difference between the two split criterions does not influence the overall result. When given a test input X the CATE is determined by evaluating Equation 2, where W_i represents the treatment of instance i, as described in Wager and Athey (2015). Intuitively this equation describes the average outcome of the treated in the leaf node L minus the average outcome of the not treated (control) in the leaf node L. Given the estimated CATEs for each tree in the forest, the output is the weighted average between all trees, where weights are dependent on the algorithm used to grow the forest.

The benefit of causal forest is that it can match data points based on their features without any additional estimation, which permits it to estimate CATE accurately if enough samples were provided and if good splits were found. However, there are also some downfalls from this approach. Forests inherit some of the downfalls of decision trees, such as failure on an imbalanced dataset as shown in Muchlinski et al. (2016). Imbalanced datasets are datasets where there is a significant imbalance of outcomes, for example, given a boolean outcome if only 5% of the data points would have 1 as an outcome, it would create an imbalance as a tree might group all the 1s each into separate leafs, meaning it would overfit.

2.2 Honesty

Honesty is a property introduced in Wager and Athey (2015) which helps to fight against bias. The reason why bias is a relevant problem in causal forests, is because CF depends on making good splits which are assumed to represent some sort of a distinction of a subpopulation within the dataset as well. When a tree overfits, it creates a distinction that does not serve as a good division into subpopulation from the dataset perspective. In these terms bias references to the estimator bias, which is the difference between the expected value of the model estimate of CATE and the actual CATE. All in all, overfitting hurts performance badly, as it creates narrow subpopulations, specifically made to accustom the training data, which in return creates significant estimator bias given new unseen data.

With overfitting in mind, Wager and Athey (2015) introduced honesty as "a tree is honest if, for each training example i, it only uses the response Y_i to estimate the within-leaf treatment effect τ using 2 or to decide where to place the splits, but not both."(p. 8). Meaning that an honest tree will only look at the outcome of a data point to find an optimal place to split or to evaluate a leaf node, but not both. Honesty is not a unique property for causal machine learning, it exists in classical regression trees as well, meaning that most of the literature on honesty in terms of causal machine learning is an extension of the classic machine learning honesty for regression trees. However, this study is only concerned about honesty in the field of causal inference, therefore conclusions will be made only about honesty in causal forests.

There are two ways proposed by Wager and Athey (2015) to achieve honesty. Double-sample trees split the training sample into two and find the splits based on one half and then evaluate based on the other. Propensity trees on the other hand, create splits based on the treatment of a data point and then evaluate the leaf node with the outcome of all the training data points that land in the leaf node. However, Athey et al. (2019) uses Double-sample trees as it is a more general method. This means that available implementations of CF, like the one by Microsoft-Research (2022), rely on the Double-sample algorithm to achieve honesty. Throughout the paper it will be assumed that honesty is defined and implemented using Double-sample trees.

3 Experimental Setup

This section describes the various experiments that took place to compare the performance of an honest CF and a regular CF. Section 3.1 discusses the different experiments that were run and the basic parameters used in the models and synthetic datasets. Sections 3.2, 3.3 and 3.4 describe the various experiments that happened in further detail and provide the obtained results.

3.1 Setup

To answer the research question of the effect honesty has on the performance of causal forests, three scenarios were created. The first setup is to check whether honesty indeed helps against estimator bias and how prevalent the bias is in a regular CF, or a CF without honesty. The second setup is created based on the claims from Wager and Athey (2015) where it is mentioned that random forests in general perform badly when trying to fit a treatment function that has sudden spikes, and honesty might make it worse. Thus, the second experiment is trying to observe this phenomenon and to create a basic reasoning for a case where honesty hurts the performance. The last experiment is created on three different datasets that all try to simulate generic real world data. The first dataset is a purely synthetic dataset that is not created with any specific case in mind. The second and third dataset are benchmark datasets used throughout the field of causal machine learning, namely the IHDP dataset, taken from Shalit et al. (2016), and the TWINS dataset, taken from Yoon et al. (2018). These datasets were picked as they both contain real values for features. IHDP is a relatively small dataset (only about 700 data points), while TWINS is a relatively large dataset (with around 22 thousand data points).

Throughout the experiments multiple parameters were adjusted and some required synthetic datasets to be generated. To avoid repetition, the basic parameters will be defined here, meaning that if there is no mention of what a parameter is equal to in the experiment it is left as its default value defined here. These values are determined by either the default values in their EconML implementation (e.g. the number of minimal leaf nodes is set to 10 by default) or by the same values used by Wager and Athey (2015) or Athey et al. (2019). All experiments consist of comparing the performance of an honest causal forest and a regular causal forest, both of which had the parameters set to their default values with the sole exception being the parameter being changed. When testing a model, the experiment is run on 70 replications and the average is taken as the result. The number 70 was chosen mostly because higher values took too long to compute, yet smaller values had too much variance. Considering that for some scenarios 70 replications might not create a convincing argument, Appendix A discusses the variance measured throughout the 70 replications.

For the model implementation, the Microsoft-Research (2022) implementation is used, which provides some extra features outside of an causal forest implementation by expanding it with Double Machine Learning, introduced by Chernozhukov et al. (2016), which should only help in the case that the dimensionality is high or parametric functions cannot model the data in a satisfactory manner. This should not have an effect on the experiments themselves, but should help with optimization. Most of parameters are left as the default parameters from the implementation. More concretely, the default `min_samples_leaf` is set to 10, `max_depth` is set to None, meaning each tree can be as deep as it requires, and `n_estimators`, also known as `number_of_trees`, is set to 100.

$$
\begin{gathered}
\text{dimensionality: p} = 5 \\
\text{X} \sim U(0,1)^p \\
\text{noise: } n() \sim N(0, 0.01) \\
\text{main effect: } m(X) = \sum_{i=0}^{p} X_i \\
\text{propensity: } e(X) = \frac{1}{|X|} \sum_{i=0}^{p} X_i \\
\text{treatment: } W(e) \sim Ber(e) \\
\zeta(x) = 1 + \frac{1}{1 + e^{-20*(x-1/3)}} \\
\text{treatment effect: } \tau(X) = \zeta(X_0)\zeta(X_1) \\
\text{outcome: } Y(m, W, \tau, n) = m + (W - 0.5) * \tau + n
\end{gathered}
\tag{3}
$$

Synthetic data is required when trying to test a specific property, as it is the only way to obtain a valid groundtruth. For experiments requiring synthetic data, the functions defined in Equation 3 are used. These functions were either taken from Wager and Athey (2015) or created by the author to mock a general setting. For a visualization of what each function represents, the causal graph shown in Figure 1 indicates what each function computes.

Throughout all experiments, many metrics from Cheng et al. (2022) were considered. However in the end the mean squared error (MSE), as it seemed to be the most widely used. This permits an easier comparison with other studies. Therefore, all comparisons between models will be done with MSE, although in the codebase other metrics can be found and tested as well.

3.2 Imbalanced Dataset Experiment

In this experiment, the goal was to test whether honesty does indeed help to fight imbalance and the bias it creates. The expected outcome of this experiment is that honesty indeed helps trees to not overfit, which provides a significant boost to the performance when compared to regular CF. To create such an experiment, an imbalanced dataset was created, and then both models were run on it to observe the different behaviour.

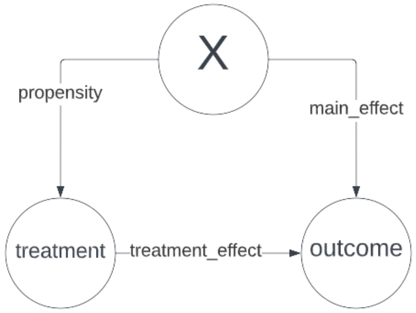


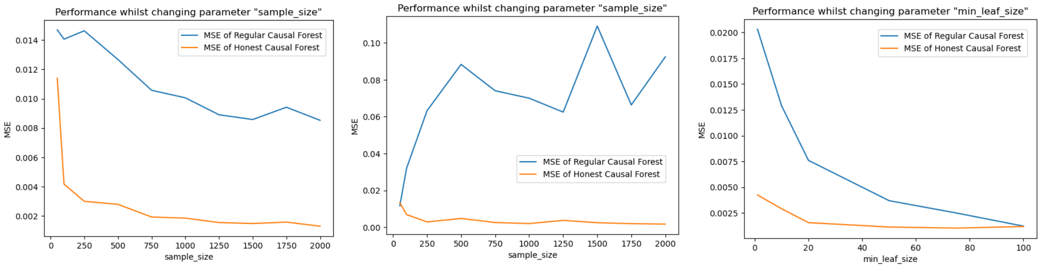
Figure 1: Causal graph indicating the different naming definitions used across the experiments. An arrow in a causal graph indicates what has influences on what. For example, in this graph, the features X influence the treatment and the strength of this influence is determined by the propensity function.

$$
\begin{aligned}
e(X) &= 0.5 \\
\tau(X) &\sim Ber(0.05) \\
Y(m, W, \tau, n) &= m + 2 * W * \tau + n
\end{aligned} \tag{4}
$$

Although such data was already generated by Wager and Athey (2015), only treatment_effect function was defined and all other influences on the outcome were set to zero. To further expand this experiment, the default functions from Equations 3 are used to generalize the experiment more. The only changes, based on the original experiment, are the functions in Figure 4. The CATE should be equal to 0.1 for all data points as the data is homogeneous. This is due to the fact that each instance has 50% chance of being treated and 5% to output 2 if its treated, meaning CATE should be equal to the expected value of $2 * 1 * Ber(0.05) - 2 * 0 * Ber(0.05)$ which is 0.1.

Figure 2 shows the different results obtained from testing different parameters on this synthetic dataset. As can be seen in Figure 2a, honesty seems to improve the overall ability to not overfit and thus achieve better performance. What seems to happen to the regular CF, as similarly described by Wager and Athey (2015), is that it pushes each extreme value into a separate leaf, which creates splits that hide the homogeneity of the data.Figure 2b illustrates a case where the forests are evaluated based on a feature vector containing only zeros. This forms an edge case, it such a feature vector will always end up in the left most node of any tree. It can be seen from the graph, that as more samples are introduced, the difference in performance seem to be getting larger.A possible explanation would be that as more data is provided, it is more likely that there will be a extreme value near zero, thus it is more likely that the zero vector will be divided into the left most leaf node with an extreme value. Figure 2c displays that, while indeed the regular CF overfits, it becomes harder to overfit as leaf nodes become larger, thus the regular CF converges towards the performance of the honest CF.

This experiment shows that in the case where a dataset is imbalanced and seems to be simple to overfit, honesty significantly helps as evaluation happens on unseen data and on edge cases. However, this effect can be countered in regular CF by increasing the sample size of the leaf nodes, if one has enough data and the knowledge that the dataset is imbalanced. It is important to take into account that honesty was designed with this goal in mind and



(a) General performance (b) Edge case performance (c) Minimal leaf size

Figure 2: Performance of honest causal forest (yellow) against regular causal forest (blue) on the synthetic sudden spike dataset. Graph in 2a depicts the performance when the `sample_size` parameter was changed, graph 2b shows the same, but tested with the point (0, 0, 0, 0, 0) which is found on the edge of the feature space. This test helps to find out how the model handles edge cases. In this instance, it tests how well the model handles cases that always end up in the left-most node. Graph 2c illustrates the general performance when `min_leaf_size` is changed.

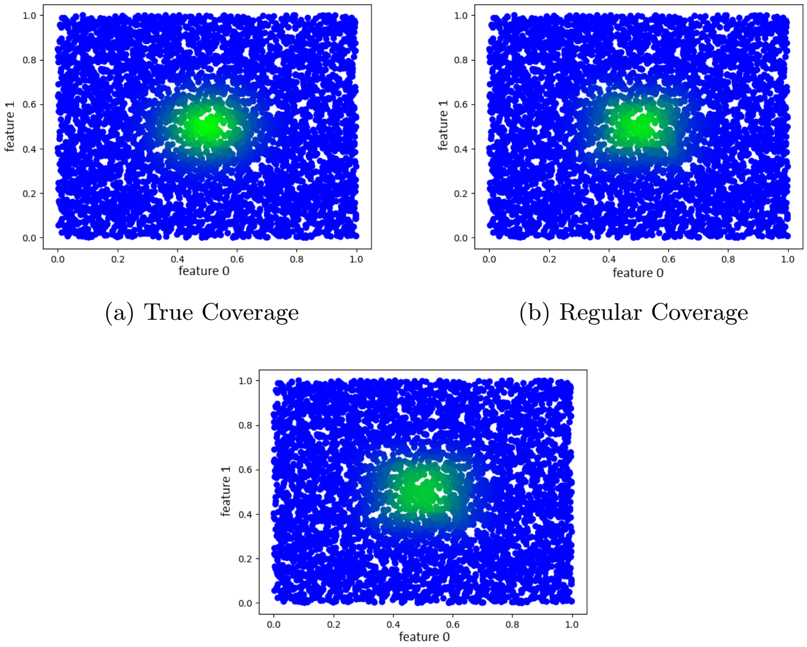
similar results can also be obtained by increasing the value of the `min_leaf_size` parameter, even as it may lead to worse results due to limited depth.

3.3 Sudden Spike Experiment

The sudden spike experiment tries to test a case where forests are known to have problematic behaviour with or without honesty. However, because honesty only takes into account half of the training data when building a tree and the other half to evaluate leaf nodes, honest trees are inherently smaller than regular trees. This can be problematic if the heterogeneous data has sudden spikes in the treatment effect. In Wager and Athey (2015) this problem is briefly mentioned in the results section, where the authors acknowledge that "It [Causal Forest] suffers from bias not only at the boundary where the treatment effect is largest, but also where the slope of the treatment effect is high in the interior."(p. 23-24). To generate such a dataset, the functions defined in Equation 5 were used to create a spike around the point (0.5, 0.5) in the plane of the first two dimensions. $f(x, y)$ represent a 2-dimensional Gaussian distribution with 0 covariance and 0.01 standard deviation in both dimensions.

$$\begin{gathered} e(X) = 1 - \sqrt{(X[0] - 0.5)^2 + (X[1] - 0.5)^2} \\ \tau(X) = f(X_0, X_1) \end{gathered} \tag{5}$$

In Figure 3 different coverage can be seen on an example of the used synthetic dataset. These figures depict the first two features plotted along the X and Y axes, where each point is a data point from the dataset. The greener a point is, the larger its treatment effect is, where blue indicates that there is no treatment effect. As can be seen in Figure 3a, the true coverage was generated such that it peaks around the point (0.5, 0.5) in terms of the first two features, and quickly dissipates into no treatment effect at all. Figure 3b shows the same depiction, but the treatment effect is estimated by a regular causal forest. The first thing to notice is that the oval shape of the original treatment effect gets transformed into a more rectangular shape. This is due to forest making orthogonal splits to one of the axes, meaning it is impossible to create a perfect circle, rather it creates an approximation by creating many smaller squares. Secondly, the colour of the peak is darker than the original, showing the weakness when it comes to sudden spikes in random forests. Figure 3c shows the same, but with estimation done by an honest CF. The problem of estimating the peak seems to become even worse, as the center is even darker than before. This indicates that the performance is heavily influenced by honesty, as honesty has influence on the depth of trees within the forest.



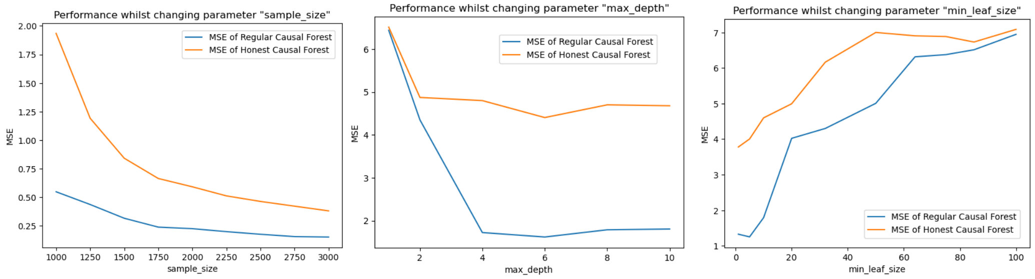
(a) True Coverage

(b) Regular Coverage

(c) Honest Coverage

Figure 3: This figure depicts the different treatment effect estimated by the models. X and Y axes are the first two features and each point represents one data point of the dataset. The greener a point gets the higher the estimated treatment effect is. In Figure 3a a depiction of the true treatment effect of the synthetic dataset can be found. Figure 3b shows the estimated treatment effect predicted by a regular CF, and Figure 3c depicts the treatment effect estimated by an honest CF.

The results of the forests ran on this dataset can be found in Figure 4. Based on Figure 4a, it seems that as more samples are provided both honest and regular CF seem to get better and better, which most likely ties to the fact that with more samples both forests are allowed to create larger trees. When testing the general case, it seems that the honest CF needs twice as many samples to produce as good of a result as a regular CF, which suggests that the hypothesis that honest trees are bounded by their lack of tree depth holds. To further support this it can be observed in Figure 4c that, as the minimum sizes of the leaf nodes increases, the performances converge to one another as both implementations are forced to stop sooner and grow smaller trees. Furthermore Figure 4b shows that when the max depth is set to a specific number, both models start with similar performance, but as they get the room to expand, honesty lacks behind the regular CF due to a smaller evaluation sample size or smaller trees. With 100 minimum leaf size it can be argued that both forests are forced to create only root node trees and thus have a similar performance.



(a) General performance

(b) Maximal depth

(c) Minimal leaf size

Figure 4: Performance of honest CF (yellow) against regular CF (blue) on the synthetic sudden spike dataset. Graph in 4a depicts the performance when `sample_size` parameter was changed, graph 4b shows the performance when the depth of the trees are set to a specific number. Graph 4c illustrates the general performance when `min_leaf_size` is changed.

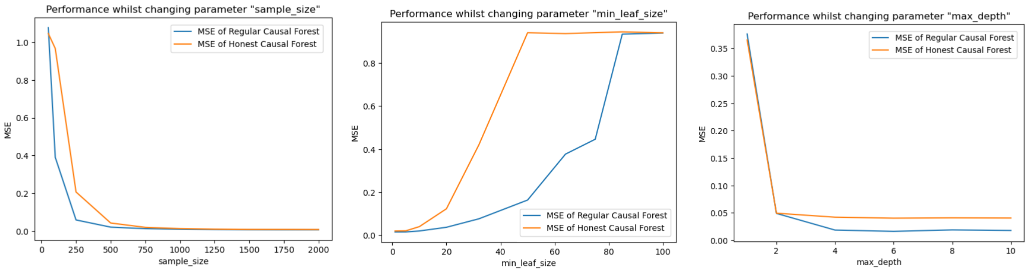
This experiment depicts the underlying problem of honesty, which is its limited depth in creating trees. It shows that, while a regular CF is already sensitive to sudden spikes, honesty intensifies this effect. This can also be viewed as a desired effect, as it prevents the trees from overfitting on the spike by creating a larger tree. Based on the experiment, the resulting data suggests that in this scenario it is more beneficial to rely on a regular CF rather than an honest one.

3.4 General Dataset Experiment

To answer the overall question of how honesty affects the performance of CF, it is crucial to investigate the effect honesty has in a general scenario. To perform this analysis, three datasets were chosen. The first general dataset is comprised of the synthetic data that follows the default functions described in Equation 3. This provides an overview on the general comparison between the two approaches in a synthetic and predictable state. The second scenario is the IHDP database, more specifically the iteration used by Shalit et al. (2016). This dataset was chosen as it is a benchmark dataset in the field of causal machine learning, and because it consists of only about 700 datapoints, providing an estimate of the performance on a smaller dataset. The last scenario chosen is the TWINS dataset taken from Yoon et al. (2018). A small change was created on the data from Yoon et al. (2018), where for each pair of twins both twins were split into its own separate row, one as treated and one as untreated, depending on their weight. Another change was to have the output represent whether the infant died during the timeframe or not, instead of representing the amount of minutes survived. This provides the experiment with a larger sample size where further hypotheses can be tested.

The main hypothesis before running the experiment is that both honest and regular CF should have similar performances. While an honest CF would create smaller trees, it should trade off the imperfection of smaller trees with its resilience against bias.

Figure 5 contains the result of the experiment run with general synthetic data. As can be observed, honesty seems to worsen the performance when compared to the regular CF. Once again, this seems to be due to the lack of depth in honest trees as there is less data to train with. This can be verified when looking at the performance when `max_depth` is changed, as both models will have similar performance whilst both have enough data to completely fill up a tree, but a difference occurs once the training sample size is limited.



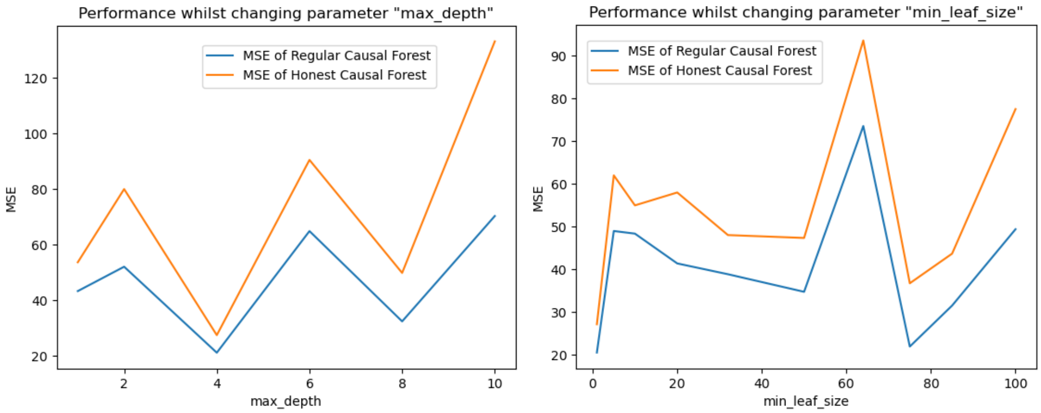
(a) General performance

(b) Minimal leaf size

(c) Maximal depth

Figure 5: Performance of honest causal forest (yellow) against regular causal forest (blue) on the general synthetic dataset. Graph 5a depicts the performance when the `sample_size` parameter was changed, while graphs 5b and 5c do the same but with `min_leaf_size` and `max_depth` parameters.

As can be seen in Figure 6, similar event occurs when working with the IHDP dataset. The IHDP dataset contains only 747 data points, thus the forests do not receive a lot of training data to begin with. This is reflected on the performance, as honest CF struggles to perform as good as a regular forest. As the depth of the tree increases, the difference becomes more visible as the regular forest is allowed to fully grow. Interestingly, the performances do not approach one another when the minimal leaf size is set, considering that splitting the training data in half might also have a major influence on not being able to correctly fit a function based on the limited data.

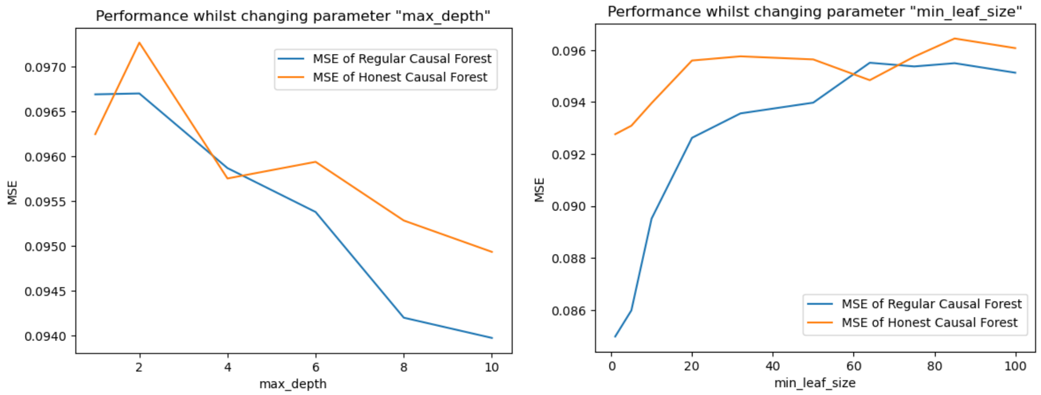


(a) Maximal depth

(b) Minimal leaf size

Figure 6: Performance of honest causal forest (yellow) against regular causal forest (blue) on the IHDP dataset. Graph 6a depicts the performance when the `max_depth` parameter was changed, and graph 6b shows the same when `min_leaf_size` was changed.

Last but not least, the twins dataset results depicted in Figure 7 also show a difference in performance when honesty is applied. The twins dataset contains around 22 thousand data points, thus even with a depth of 10 both trees seem to perform similarly, as the difference between their scores is small. However, a more significant change can be seen when changing the `min_leaf_size` parameter as the regular CF starts out stronger and both slowly converge to a worse version of themselves. This result is particularly interesting, as based on the experiment in Section 3.2 the regular CF should improve as the leaf size increases, but the opposite happens.



(a) Maximal depth

(b) Minimal leaf size

Figure 7: Performance of honest causal forest (yellow) against regular causal forest (blue) on the TWINS dataset. Graph 7a depicts the performance when the `max_depth` parameter was changed, and graph 7b shows the same when `min_leaf_size` was changed.

Based on the three experiments it seems that honest CF struggles to catch up to the regular CF in terms of MSE when there is limited amount of data. A possible explanation can be that the lack of training data strains the depth of the tree to such a level that it affects the outcome of the forest as a whole. This is supported by the fact that in TWINS, which contains enough data points for both CF, the performance is similar, while in IHDP where the sample size is limited, honest CF performs noticeably worse.

4 Discussion

Honesty was developed with one precise task in mind, which is to create a safer forest that is able to avoid overfitting. While there are many different implementations of honesty, the one used by most libraries that include causal forests is the Double-sample trees proposed by Wager and Athey (2015). While some comparisons between honesty and regular trees were already made, these comparisons were made in the terms of predictive power and classification performance, rather than causal machine learning. Denil et al. (2014) compares regular forests described by Biau et al. (2008) with similar forests that implement honesty introduced by Biau (2012). In these tests honesty seemed to have a positive effect on datasets with a lot of data, only one out of four tests showed that honesty had a negative effect, which was a dataset with around 6500 data points. It is important to mention that one test also included a dataset of 442 datapoints where honesty did improve the performance.

When it comes to evaluating the results of this paper based on the experiment described in Section 3.2, it seems that indeed honesty ensures that forests do not overfit on an imbalanced set of data. However, as was shown in Sections 3.3 and 3.4 there are cases where honesty damages the overall performance. However, this change for the worse seems to disappear with more data, as there is enough data to grow the tree substantially in either implementation. Surprisingly, the results shown in Figure 7 indicate that honest and regular causal forests perform similarly on a large data sample, but the regular CF seems to be a better choice when leaf nodes only hold one sample. This, however, might be attributed to the TWINS dataset, rather than an underlying property of honesty.

It can be seen that honesty has specific cases where it provides significant help, but it might be a dangerous tool to use on smaller datasets. As discussed throughout Section 3, this seems to be caused by honesty creating trees with only half of the training data, resulting in smaller, shallower trees. The experiment in Section 3.4, which tests honest CF on IHDP data, also seems to show an error which might be caused by the lack of data in the leaf nodes, as only half of the training data is used to evaluate the leaf node. Considering that the evaluation depends on estimating the mean of two values, a lack of data would introduce a lot of variation and bias. This leads to the conclusion that based on the general performances on general datasets, shown in Section 3.4, honesty seems to have worse performance when there isn't a lot of data to train on. This seems to be due to its inability to create deeper trees, as when a `max_depth` parameter is set, the differences tend to disappear until the depth permits the honest trees to use all sample data to create a tree. From that point honest trees start stagnating in performance while regular trees are still improving, as they can still grow further. If there is enough data to create in-depth trees by both implementations, their performances seem to become similar.

To establish a general approach when deciding whether honesty should be implemented or not, a lot of questions have to be answered first. Most importantly, as is discussed later in Section 5, when the experiment contains ethically sensitive data, honesty is an ideal way to ensure fairness and avoid creating ethical mishaps. Secondly, some preliminary tests can be run on the dataset to determine whether forests might overfit. For example, a basic count of treated and not treated samples already provides some basic information, even as it might not provide further information about subpopulations. Finally, as later discussed in future work, either a different definition of honesty can be applied, or one can implement a forest that isn't purely honest, meaning that only a certain percentage of trees is honest. This way there should be a percentage of trees that won't suffer from the downfalls of honesty shown in this study, but it will also contain trees that will indicate that a forest is overfitting by providing a very different result from the non-honest trees.

5 Responsible Research

To ensure the credibility of this research, many steps were taken to make sure that all results shown in this paper are reproducible. The entire codebase used to obtain these results can be found on the GitHub repository of the author[1]. In this codebase, multiple experiments can be run and all experiments that were used to obtain these results can be found in `main.py` file. If the `save_data` and `save_graphs` parameters are set to true, all data and intermediate results, as well as the final results will be generated and stored in local directories, which is further explained in the codebase itself. This permits the reader to recreate the same experiments and validate the results.

[1] https://github.com/MatejHav/causal-methods-evaluation

During the development of this research, certain ethical standards had to be considered. Mainly when one considers the possible conclusions of this research, it becomes apparent that there are ethical dangers that need to be avoided. The entire field of causal machine learning is developed around treatment and outcome, which perfectly fits medical studies about drugs and other treatments. This study is trying to find out whether honesty, a property meant to fight bias, is worth adding to causal forests for evaluation. Therefore, it is important to consider that a conclusion of this research might impact choices made in a possible medical study, which might bring bias to medical studies. To counter this, it is important to mention that these results are empirical and in no way a mathematical proof, thus one should not decide whether to use honesty purely on this paper, rather it should inspire further research into honesty in causal forests.

6 Conclusions and Future Work

The main question of this paper was to empirically show the effect of honesty on the performance of causal forests. Based on the results of Section 3.2, it seems that honesty indeed helps to fight bias and performs better against a regular CF on an imbalanced dataset. Section 3.3 establishes the first instance where one can observe that the halved sample size used to grow the trees when honesty is present, has a negative effect on the performance of CF. This also reflects on the fact that the curves of the performance for both honest and regular CF seem to be the same, however honest CF requires twice as many samples to achieve the same result as a regular CF. All of this led to the general dataset setting, described in Section 3.4, where it is further shown that on smaller datasets honesty seems to struggle with its halved sample size to grow and evaluate trees. However with enough samples, the differences tend to disappear. Thus, as discussed in Section 4, when deciding whether to use honesty or not, many variables have to be taken into account. If the results could have real life consequences, then honesty is a benefit for its strength against bias. If there is only a limited amount of data, honesty might bring unnecessary drop of performance. In such cases it would be encouraged to include honesty in parameter optimization and decide based on results specific to the given problem. As a last mention, honesty was also built as an improvement for confidence intervals of estimates, which is also an important fact to take into account when building a model, as sometimes an interval could be better than one exact estimate.

There are still many possible options that need to be explored. Firstly, this paper does not provide an exhaustive test of honesty, but rather hand-picked tests that are supposed to reflect performance under specific conditions. Further research should run it in more general cases to test out whether conclusions made here based on initial samples truly hold in the general world. Secondly, at the beginning, one specific definition of honesty was assumed and tested. This was because the Double-sample tree definition is the one implemented by Athey et al. (2019), which is the baseline for modern random forests. However, it would be interesting to observe how honesty holds up in different settings with a different definition. Based on the conclusion, if one manages to train on the same amount of training data as a regular causal forest, the disadvantages of honesty should disappear. One possible definition to test would be one that uses a full sample to grow a tree and then a new sample to evaluate the tree, either a newly generated one or one from a different tree in the forest. In that case, the honest trees will use the same size of the training samples as their regular counterparts. Lastly, given the limited computing power of the author, all experiments were run on 70 replications. Whilst it is not expected that more replications would lead to different results, it is always encouraged to decrease variance further for even stronger results.

References

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *Annals of Statistics*, 47:1179–1203.

Biau, G. (2012). Analysis of a random forests model. *J. Mach. Learn. Res.*, 13(null):1063â1095.

Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.*, 9:2015â2033.

Cheng, L., Guo, R., Moraffah, R., Sheth, P., Candan, K. S., and Liu, H. (2022). Evaluation methods and measures for causal learning algorithms.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2016). Double/debiased machine learning for treatment and causal parameters.

Denil, M., Matheson, D., and De Freitas, N. (2014). Narrowing the gap: Random forests in theory and in practice. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 665–673, Bejing, China. PMLR.

Microsoft-Research (2022). Econml. Available documentation at https://econml.azurewebsites.net/spec/estimation/forest.html.

Miller, S. (2020). Causal forest estimation of heterogeneous and time-varying environmental policy effects. *Journal of Environmental Economics and Management*, 103:102337.

Muchlinski, D., Siroky, D., He, J., and Kocher, M. (2016). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*, 24(1):87â103.

Shalit, U., Johansson, F. D., and Sontag, D. (2016). Estimating individual treatment effect: generalization bounds and algorithms.

Wager, S. and Athey, S. (2015). Estimation and inference of heterogeneous treatment effects using random forests.

Yoon, J., Jordon, J., and van der Schaar, M. (2018). GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*.

Zhang, Y., Li, H., and Ren, G. (2022). Estimating heterogeneous treatment effects in road safety analysis using generalized random forests. *Accident Analysis Prevention*, 165:106507.

A Analysis of variance in results

Throughout the paper many results are shown and all of these results were obtained over 70 replications, meaning there is still some amount of variance. It is worth analyzing this variance as it can show how certain can one be about the given results. Overall, the approach was the same for each experiment. As a rule of thumb if the results are within one standard deviation from one another, they are understood as being similar.

In this Section a brief overview of the variance for each experiment will be provided with generated boxplots from the same runs as the used figures. Section A.1 shows the results of the Imbalanced dataset experiment described in Section 3.2. Section A.2 contains the variance of the Sudden spike experiment from Section 3.3. Lastly, Section A.3 describes the variance of the general results discussed in Section 3.4.

A.1 Variance of Imbalanced dataset experiment

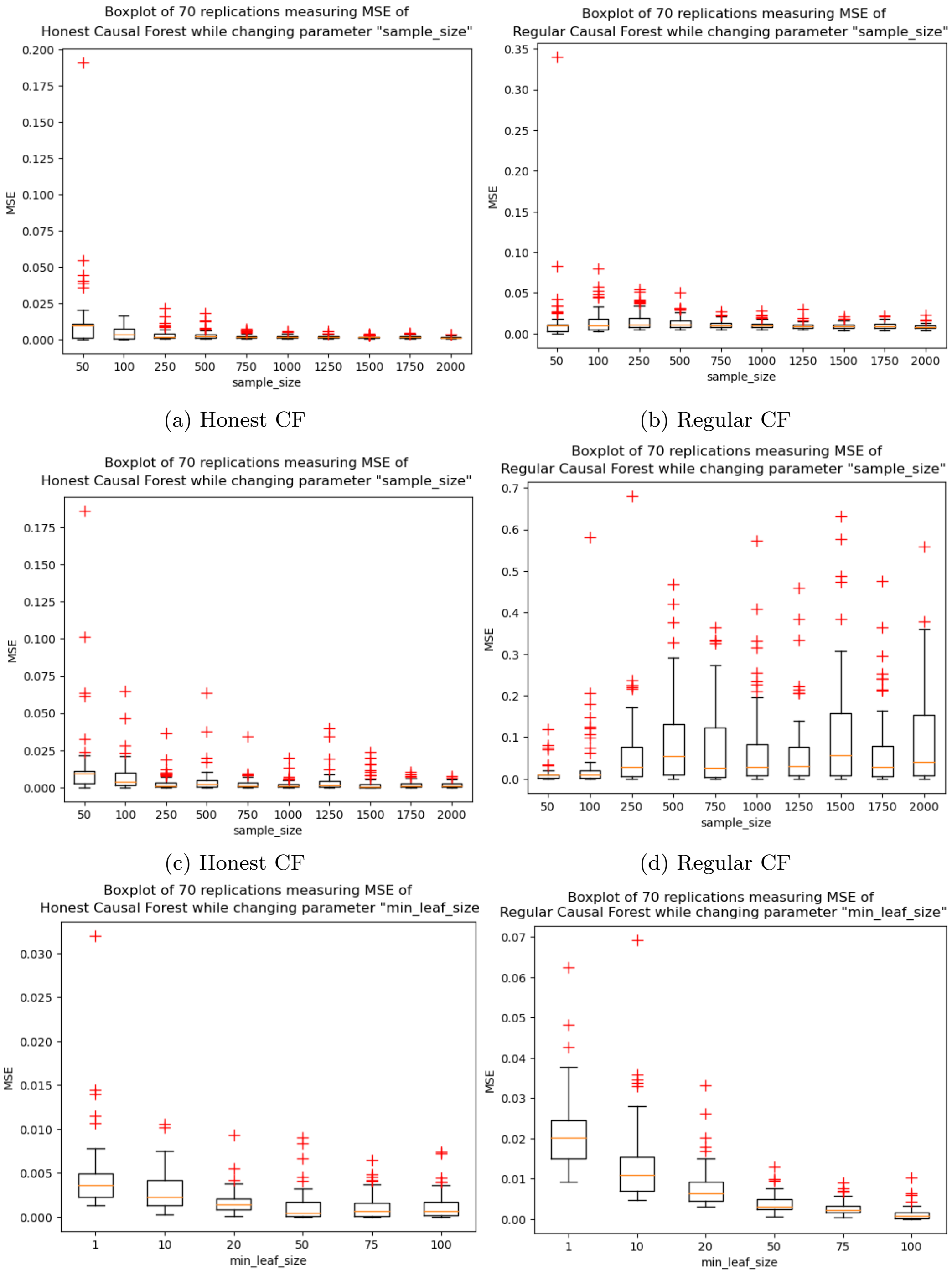
The variance of results from Section 3.2 can be found in Figure 8.

A.2 Variance of Spiked dataset experiment

The variance of results from Section 3.3 can be found in Figure 9.

A.3 Variance of General dataset experiment

The variance of results from Section 3.4 can be found in in the following figures. Figure 10 depicts the variance of results of the General synthetic. Figure 11 shows the variance of the results obtained from the IHDP dataset and Figure 12 illustrates the variance of the measured performance on the TWINS dataset.



(a) Honest CF

(b) Regular CF

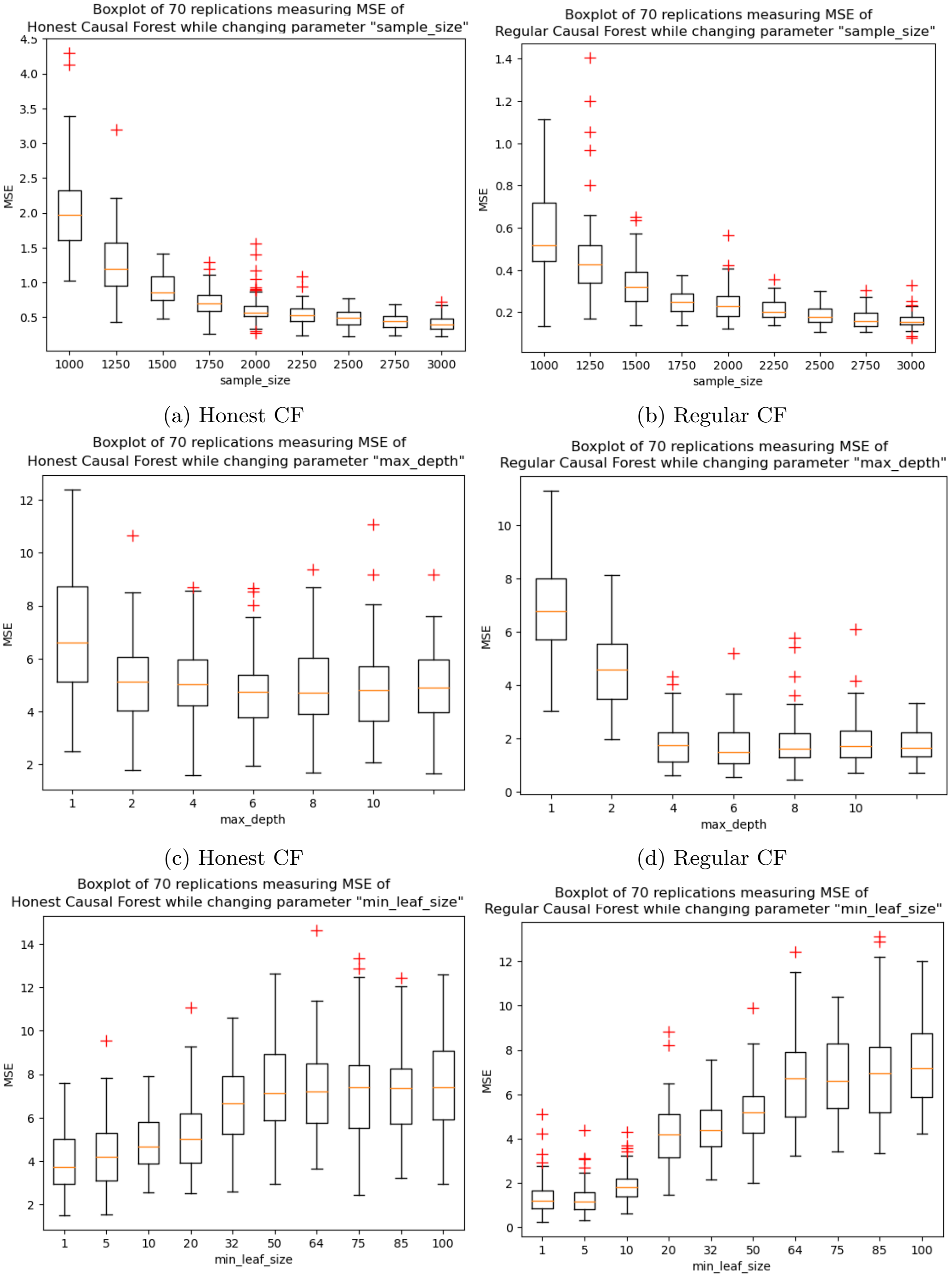
(c) Honest CF

(d) Regular CF

(e) Honest CF

(f) Regular CF

Figure 8: Figures 8a and 8b show the variance of results shown in Figure 2a. Figures 8c and 8d depict the variance of results from Figure 2b. Figures 8e and 8f illustrate the varied results depicted in Figure 2c



(a) Honest CF

(b) Regular CF

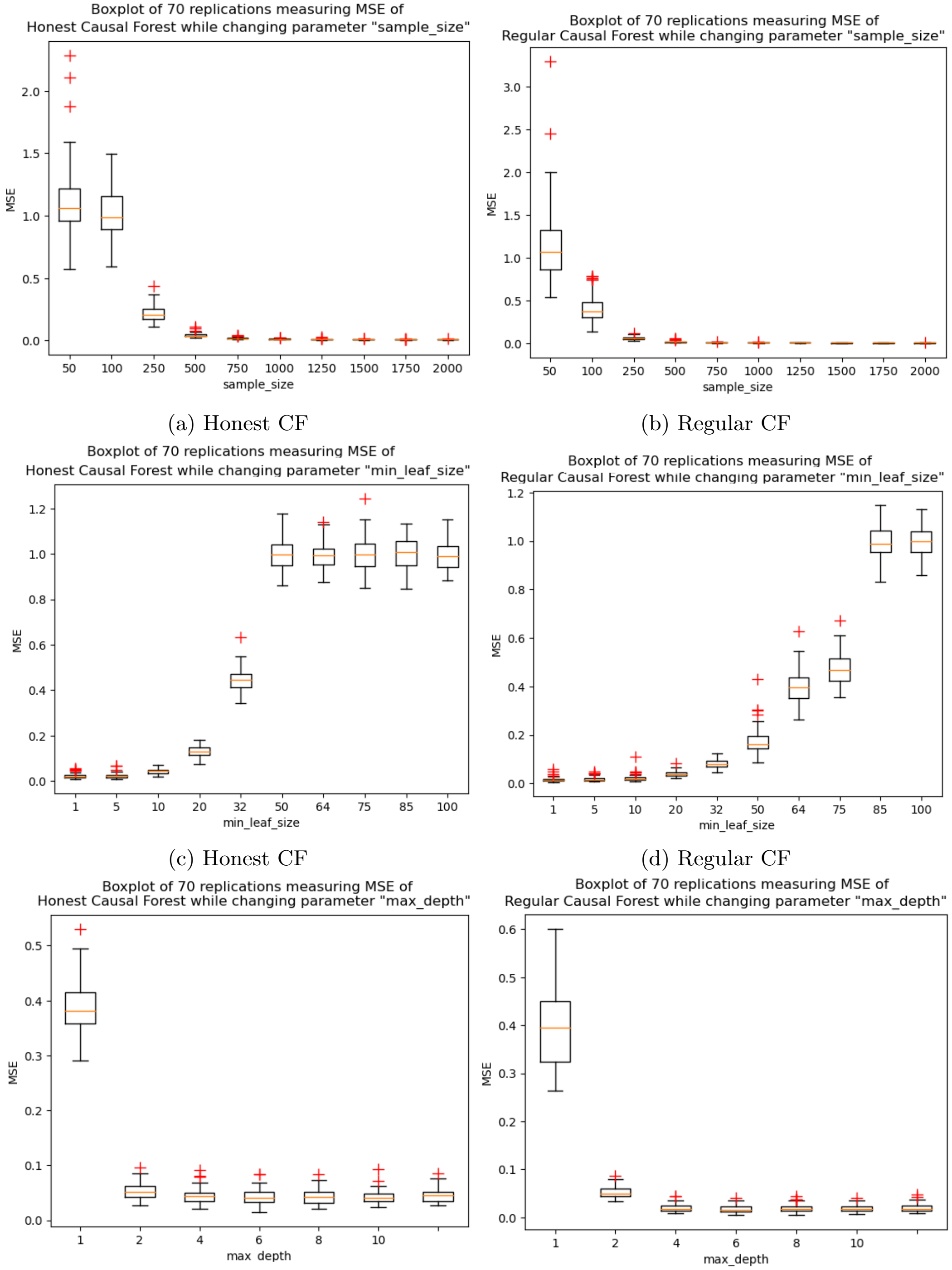
(c) Honest CF

(d) Regular CF

(e) Honest CF

(f) Regular CF

Figure 9: Figures 9a and 9b show the variance of results shown in Figure 4a. Figures 9c and 9d depict the variance of results from Figure 4b. Figures 9e and 9f illustrate the varied results depicted in Figure 4c



(a) Honest CF

(b) Regular CF

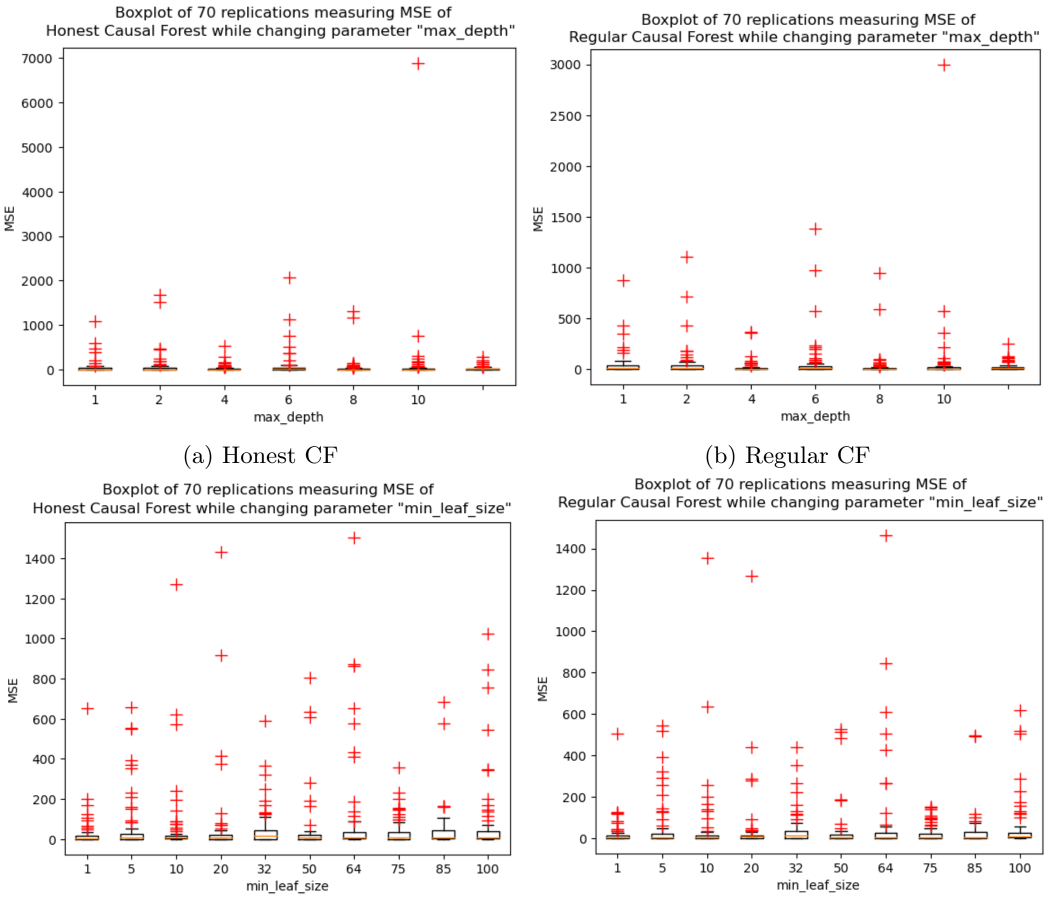
(c) Honest CF

(d) Regular CF

(e) Honest CF

(f) Regular CF

Figure 10: Figures 10a and 10b show the variance of results shown in Figure 5a. Figures 10c and 10d depict the variance of results from Figure 5b. Figures 10e and 10f illustrate the varied results depicted in Figure 5c



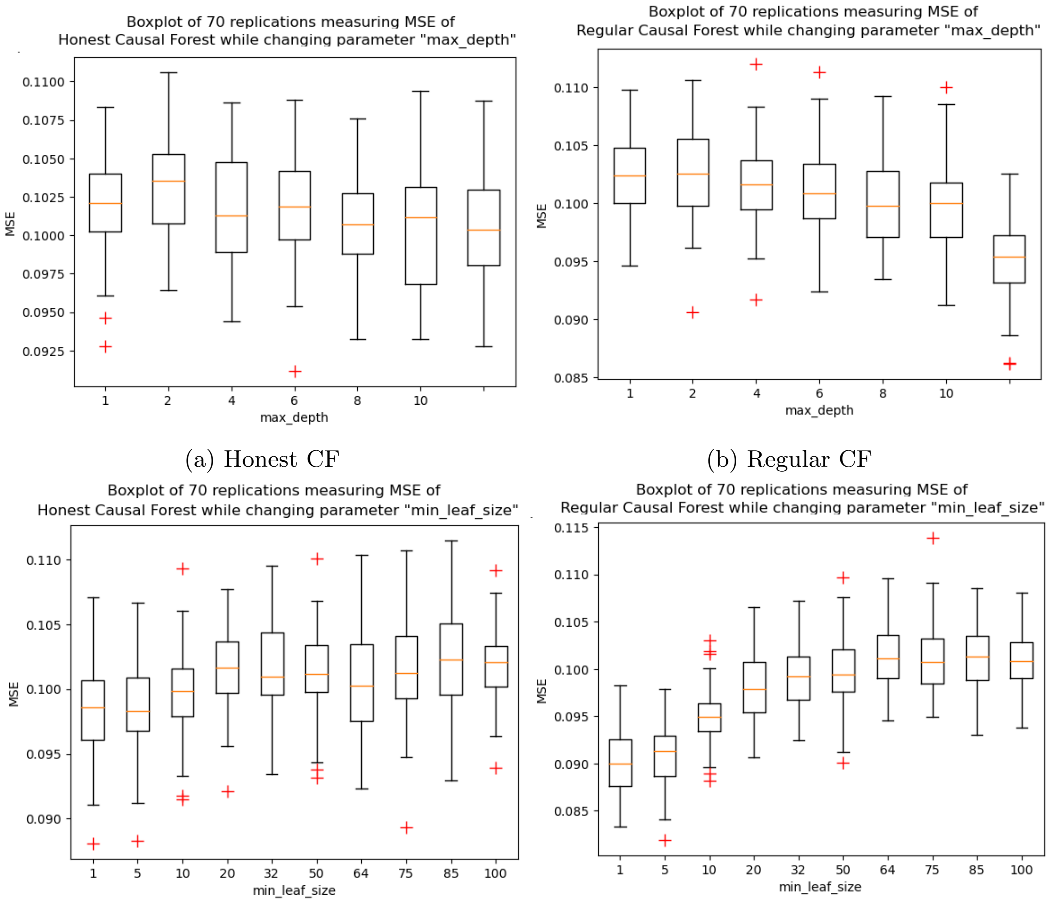
(a) Honest CF

(b) Regular CF

(c) Honest CF

(d) Regular CF

Figure 11: Figures 11a and 11b show the variance of results shown in Figure 6a. Figures 11c and 11d depict the variance of results from Figure 6b.



(a) Honest CF

(b) Regular CF

(c) Honest CF

(d) Regular CF

Figure 12: Figures 12a and 12b show the variance of results shown in Figure 7a. Figures 12c and 12d depict the variance of results from Figure 7b.