Flexibly Estimating Outcome Model Plug-In, IPW Plug-In, and Doubly Robust Estimator Using Gradient-Boosted Decision Trees

Peter Wu
Spring 2022: Causal Inference
Friday, May 6, 2022

Goal/Purpose

What is the purpose of this study?

- This study aims at flexibly estimating the the outcome and propensity scores using specifically non-parametric machine learning models random forest and gradient-boosted decision trees.
- Throughout class, we have seen parametric models of linear regression and logistic regression used to predict the outcome and propensity scores.
- Are we able to get more accurate predictions? Do we want to?

Using non-parametric models for prediction

- It is very plausible that when we use a flexible method like random forest, we will get a better accuracy in propensity scores.
- But in doing so, we may get extreme propensity score values, which will cause our estimators to in turn have extremely high variance.

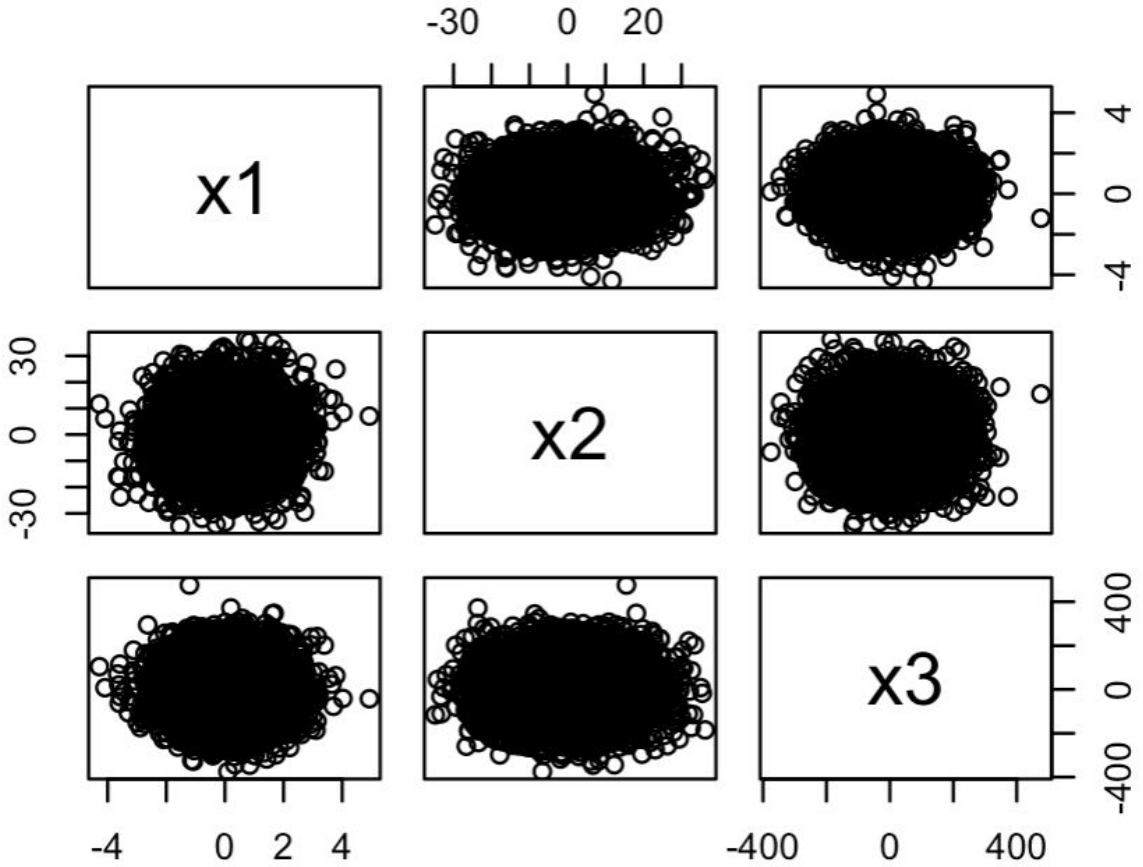
Why is this problem challenging?

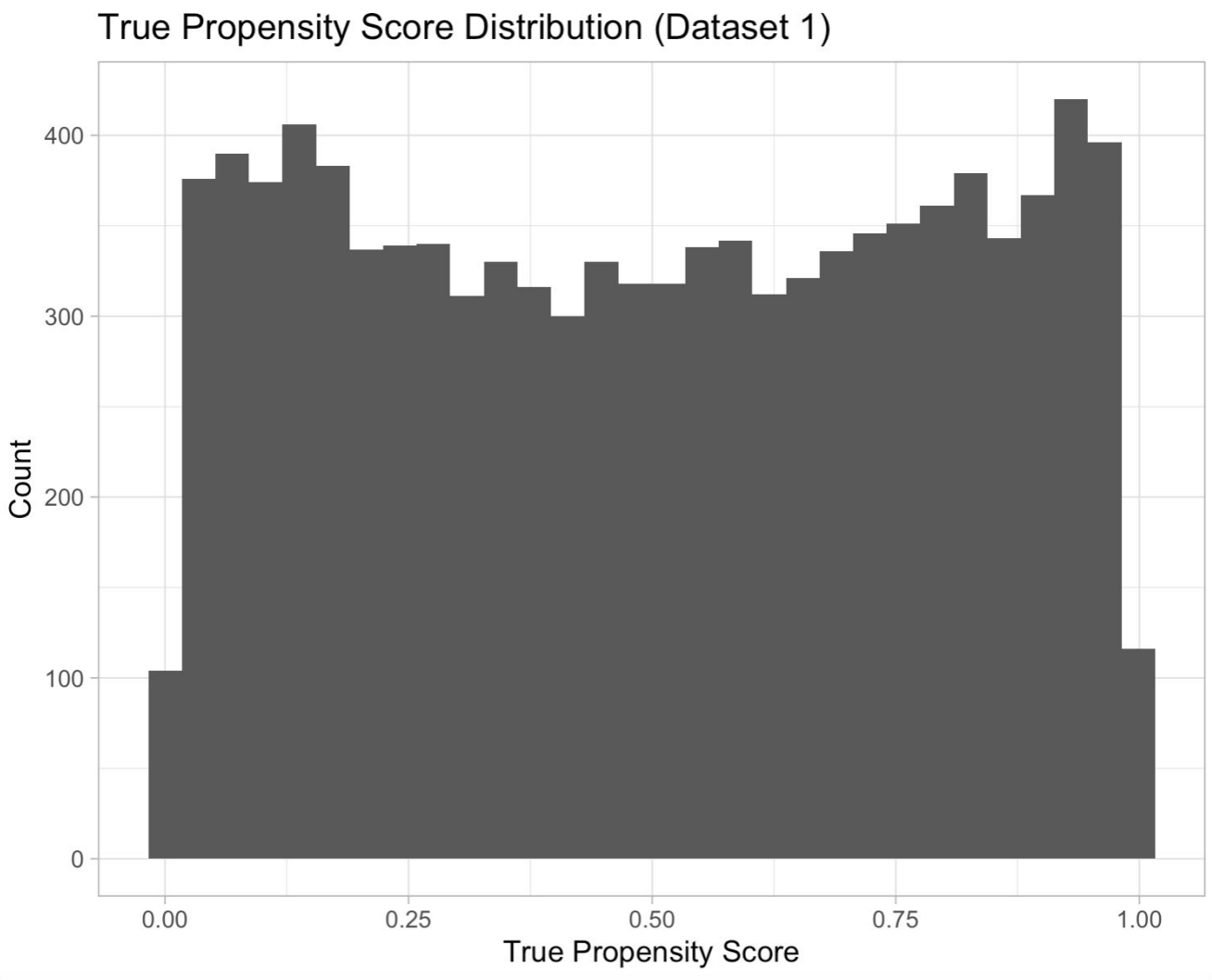
- This is challenging from a statistical perspective because it is unclear given any dataset which model would be the most accurate. There is “no free lunch”.
- From a causal perspective, if the variance of the estimators is slightly higher using the flexible models but the propensity scores are much more accurate, which model should we stick with? How big can the variance go using the flexible methods?

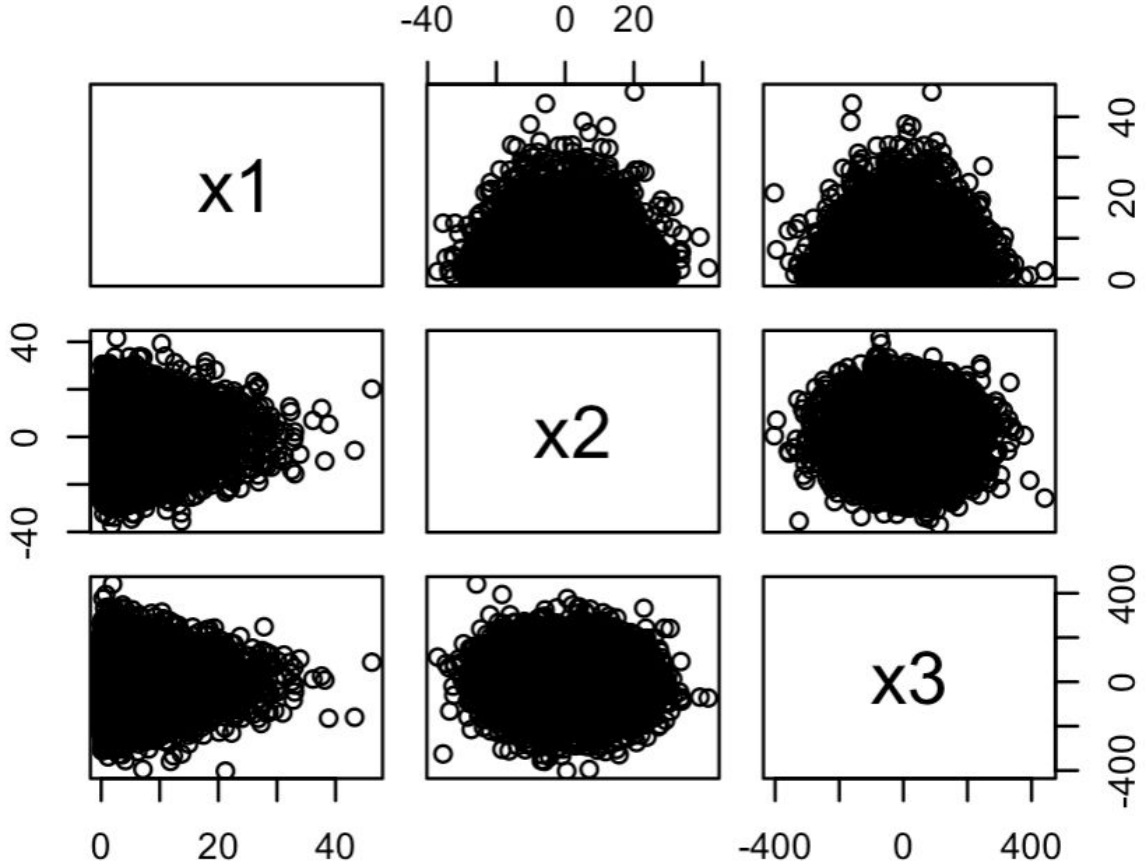
Data Description

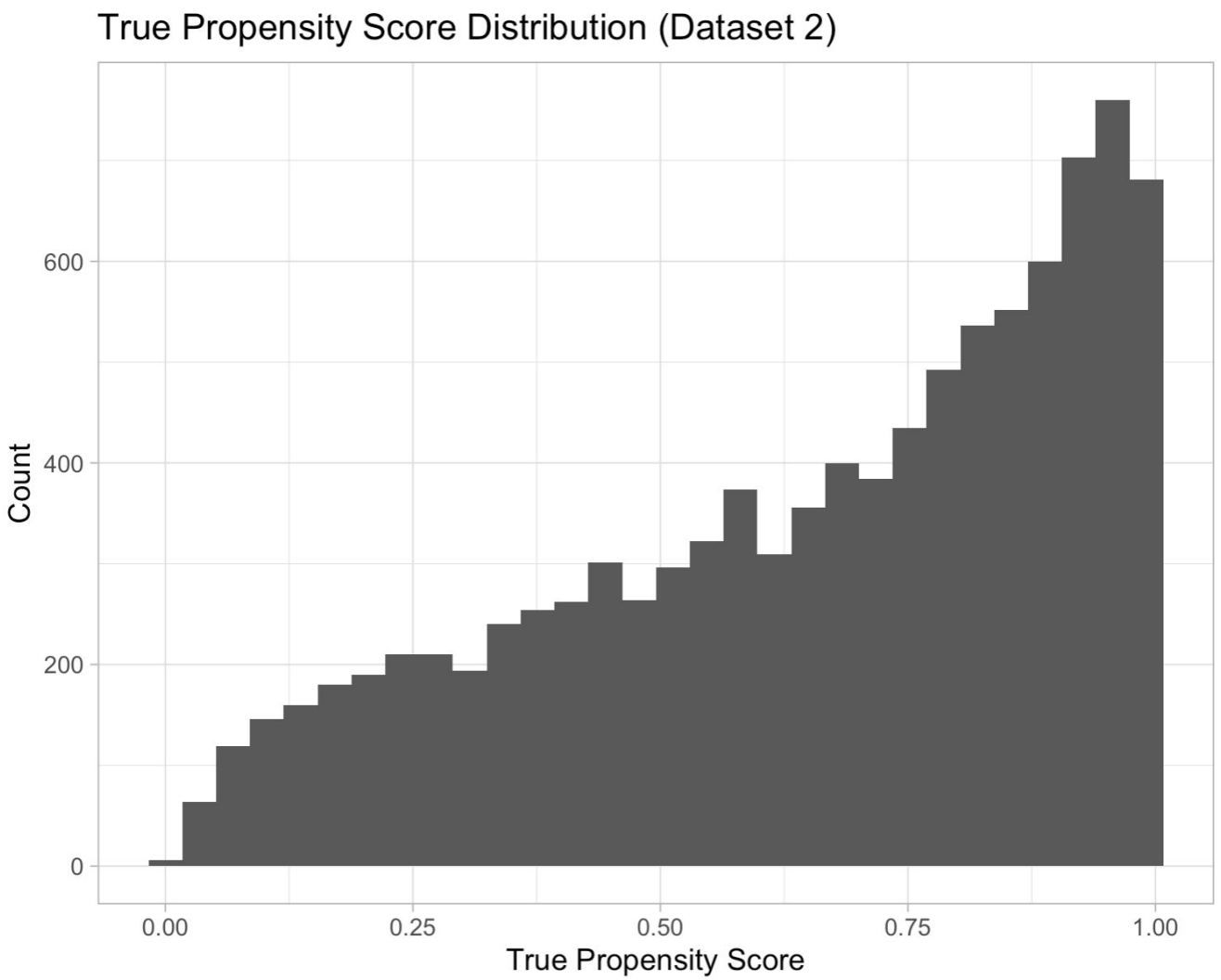
Datasets Simulated

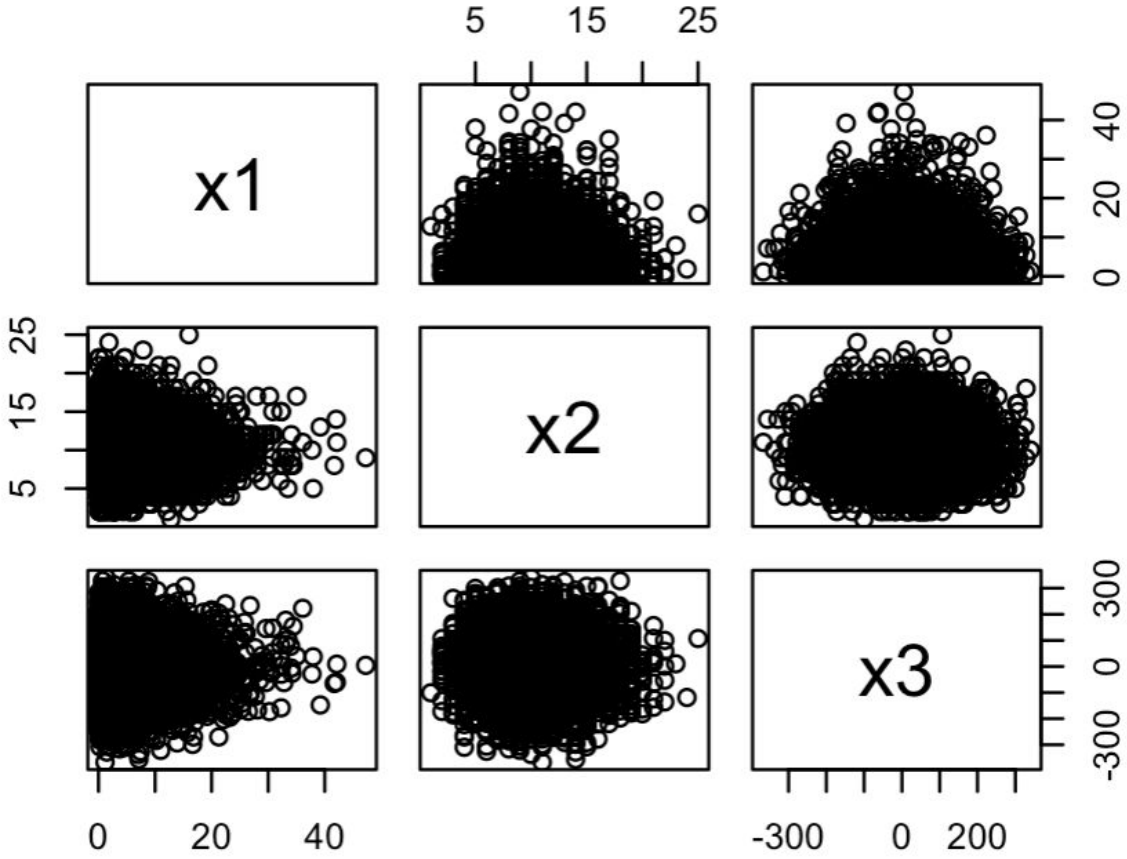
- Four different datasets, each with three features, were simulated. They each had a varying number of variables that were right skewed.
  - Dataset 1: All three variables are normally distributed
  - Dataset 2: Two variables are normally distributed
  - Dataset 3: One variable is normally distributed
  - Dataset 4: No variables are normally distributed

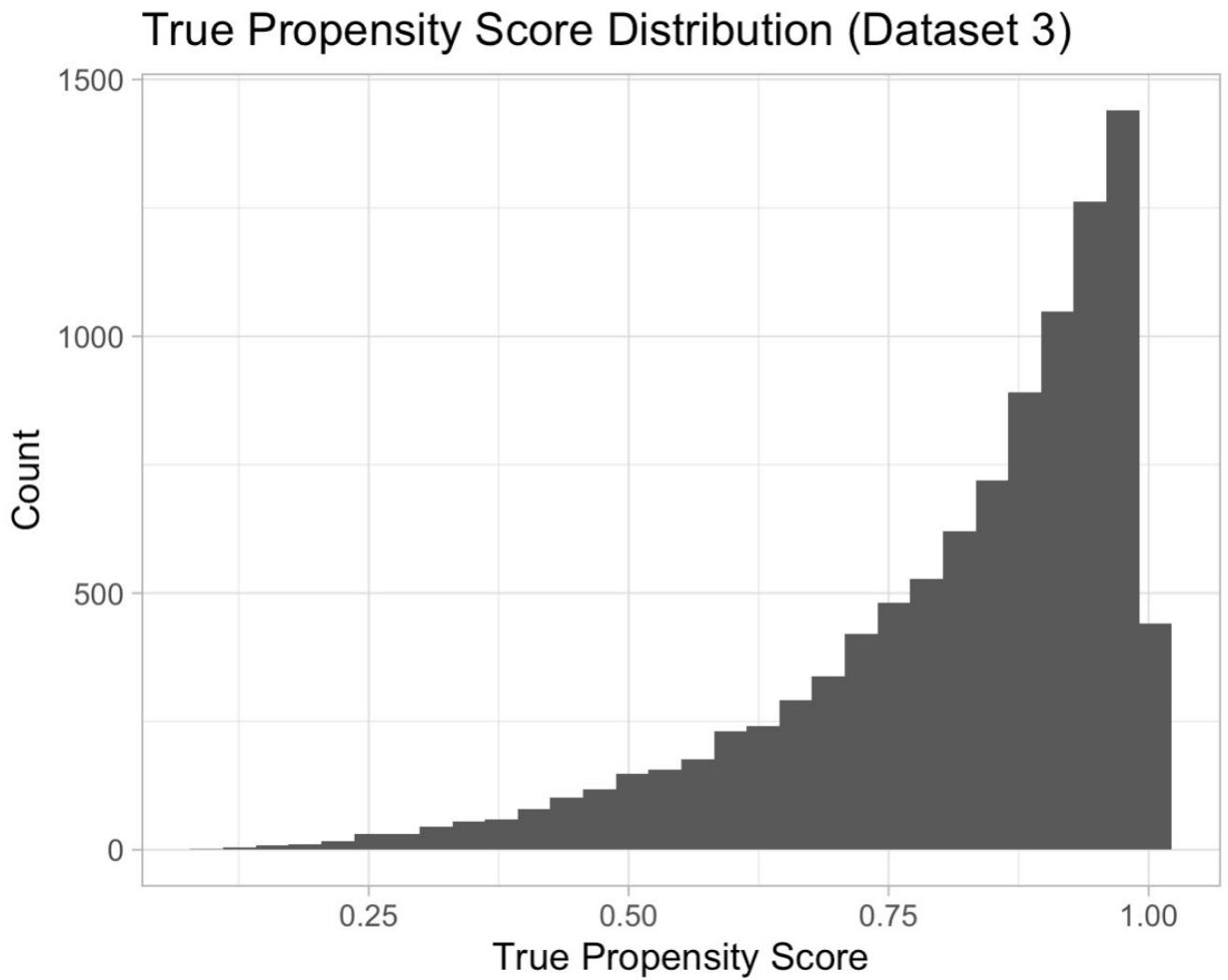

x1
x2
x3
-30
0
20
4
0
-4
30
0
-30
-4
0
2
4
400
0
-400
-400
0
400

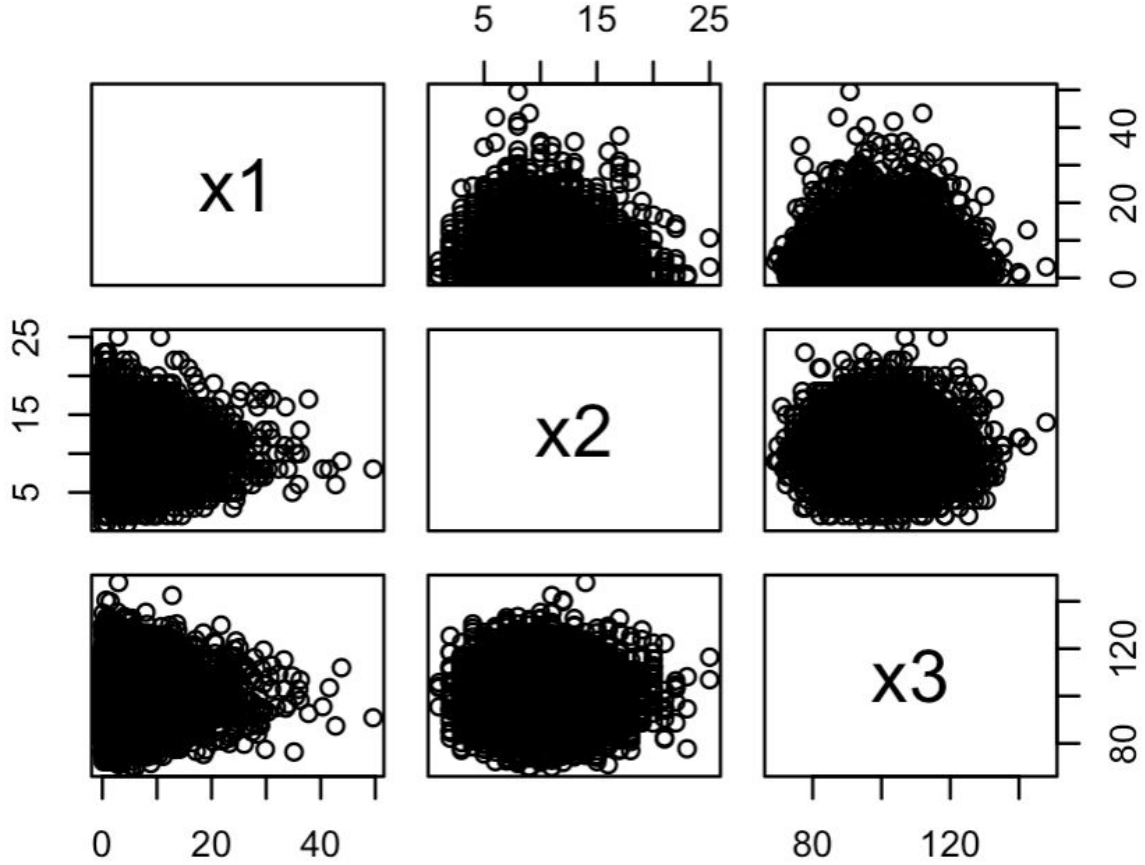

True Propensity Score Distribution (Dataset 1)
Count
400
300
200
100
0
0.00
0.25
0.50
0.75
1.00
True Propensity Score

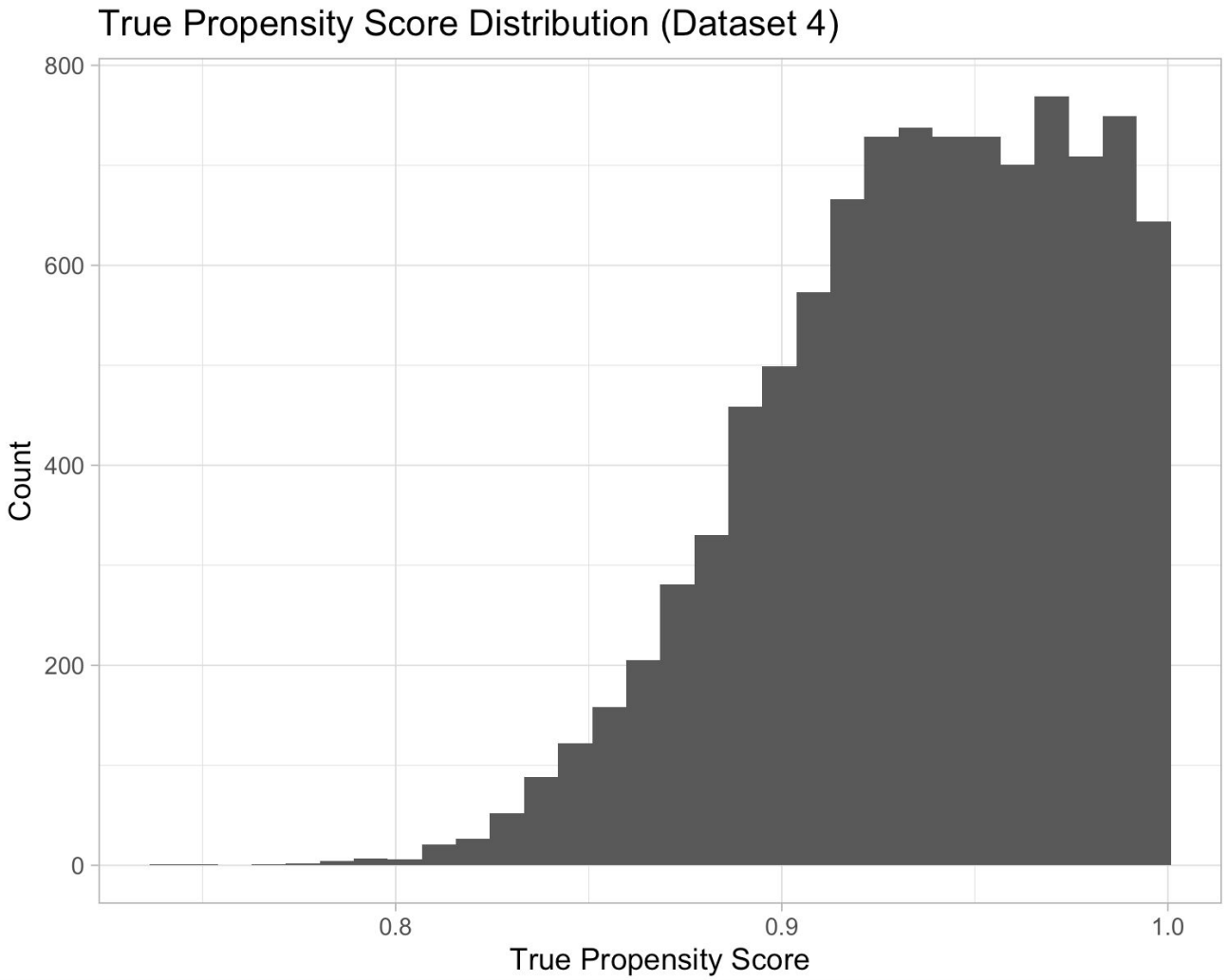

x1
x2
x3
-40
0
20
40
-40
0
40
0
20
40
-400
0
400
-400
0
400


True Propensity Score Distribution (Dataset 2)
Count
600
400
200
0
0.00
0.25
0.50
0.75
1.00
True Propensity Score


x1
x2
x3
5
15
25
0
20
40
-300
0
200
300


True Propensity Score Distribution (Dataset 3)
1500
1000
500
0
Count
0.25
0.50
0.75
1.00
True Propensity Score


x1
x2
x3
5
15
25
0
20
40
80
120


True Propensity Score Distribution (Dataset 4)
800
600
400
200
0
Count
0.8
0.9
1.0
True Propensity Score

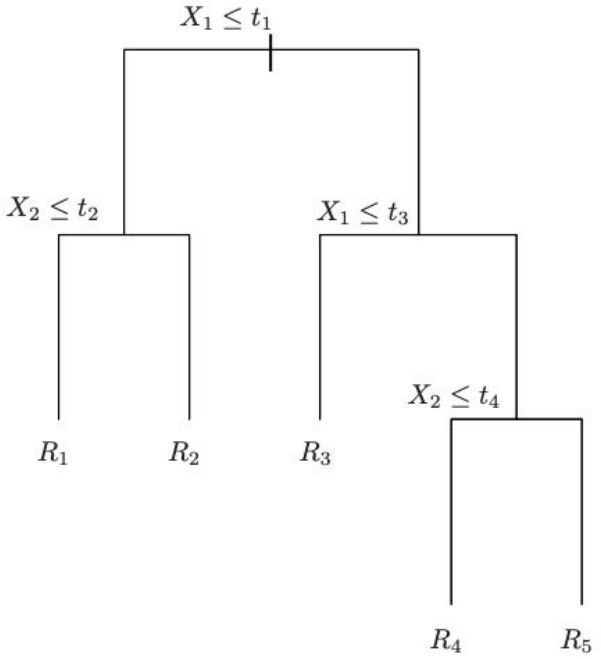
Methods

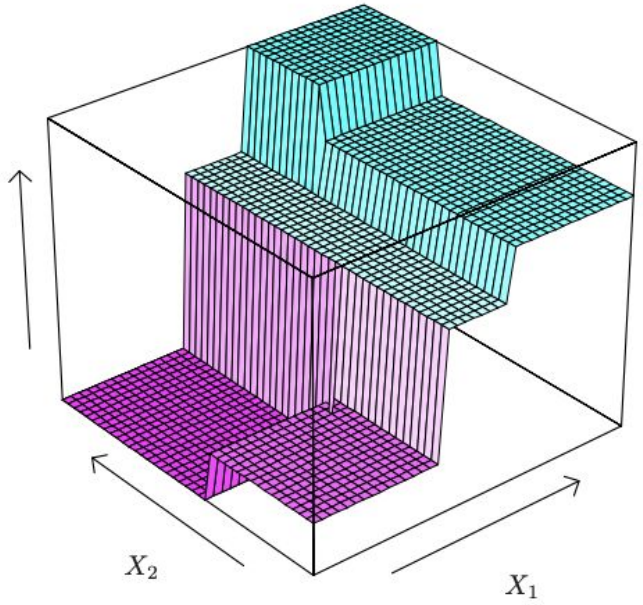
The Three Estimators Investigated

- To estimate the outcome and the propensity scores, we use random forests to flexibly estimate them. Random forest can handle both classification and regression tasks.
- The three estimators of interest are below (Lecture 22):

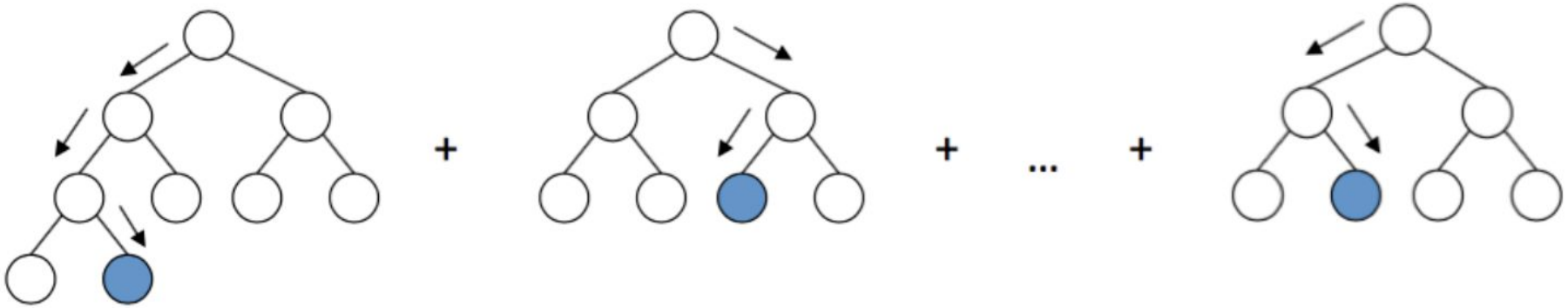
1. **Outcome Model Plug-In** $\hat{\tau}_{\text{reg}}$: Estimate $\mu_z(x) = \mathbb{E}[Y_i | Z_i = z, \mathbf{X}_i = x]$
2. **IPW Plug-In** $\hat{\tau}_{\text{IPW}}$: Estimate $e(x) = \mathbb{E}[Z_i | \mathbf{X}_i = x]$
3. **Doubly Robust Estimator** $\hat{\tau}_{\text{DR}}$: Estimate $\mu_z(x)$ and $e(x)$

Decision Trees





Gradient-Boosted Decision Trees (GBDTs)



Results/Interpretations

Point Estimate and Variances of Estimators (D1)

	X2.5.	X50.	X97.5.
Outcome LR	-10.0408099	-10.000928	-9.966231
Outcome GBM	163.5485457	163.548546	163.548546
IPW LogReg	-37.6985420	-8.058089	22.810332
IPW GBM	99.4238572	114.711443	124.471951
DR LR + LogReg	-10.1007289	-9.997882	-9.922325
DR GBM	-0.4929599	1.284343	2.813575

Mean-Squared Errors of Outcome Predictions (D1)

```
> mean((linReg + 10)^2)
[1] 0.0004261337
> mean((GBMReg + 10)^2)
[1] 30119.1
> mean((ipw + 10)^2)
[1] 241.2429
> mean((GBMipw + 10)^2)
[1] 15206.88
> mean((dr + 10)^2)
[1] 0.002252558
> mean((GBMdr + 10)^2)
[1] 126.755
```

Point Estimate and Variances of Estimators (D2)

	X2.5.	X50.	X97.5.
Outcome LR	-10.0460199	-9.998121	-9.959594
Outcome GBM	163.5485457	163.548546	163.548546
IPW LogReg	-76.1593131	-3.166307	21.840857
IPW GBM	90.3609441	104.653753	118.091128
DR LR + LogReg	-12.5052795	-9.548683	-8.793626
DR GBM	-0.4929599	1.284343	2.813575

Mean-Squared Errors of Outcome Predictions (D2)

```
> mean((linReg + 10)^2)
[1] 0.0004976548
> mean((GBMReg + 10)^2)
[1] 30119.1
> mean((ipw + 10)^2)
[1] 681.2805
> mean((GBMipw + 10)^2)
[1] 13273.64
> mean((dr + 10)^2)
[1] 2.073383
> mean((GBMdr + 10)^2)
[1] 202.3384
```

Accuracy of the Propensity Scores

-

```
> psGBM
 [1] 0.6470 0.6421 0.6502 0.6523 0.6368 0.6387 0.6504 0.6356 0.6535 0.6579
> psLogReg
 [1] 0.4864 0.5046 0.4948 0.5010 0.4966 0.4931 0.5053 0.5029 0.4966 0.4947
```

Conclusions

What did we learn from this study?

- We learned that even though it is conceivable that we can improve the accuracy of propensity scores algorithm with more flexible non-parametric methods like gradient-boosting decision trees, it may not always be advisable as it can blow up the variance extremely quickly.
- These methods are very prone to overfitting as well, so careful hyperparameter tuning is likely necessary as opposed to only using default parameters.

Should we use Non-Parametric Methods?

- Nonetheless, it is great to have another tool right up our toolbelt in being able to apply a non-parametric method like gradient-boosting decision trees for both the classification and regression tasks.
- The non-parametric methods possess much more scope and power than the parametric methods, but “with great power comes great responsibility”. One must be extremely mindful of not just squeezing out every ounce of predictive power from the dataset.

Future Work

Possible Limitations and Next Steps for Project

- We only looked at four datasets in this study and specifically altered the number of right-skewed distributions in the dataset, but in reality there are far more complex and more “messy” datasets that the methods we used to predict will interact with.
- We could tweak the way we simulated our data, and perhaps make it more complex or more random.
- We could also try another flexible model like random forests to see if the results still hold.

Possible Limitations and Next Steps for Project

- We can somehow take into account the fact that gradient-boosting decision trees can perform both classification and regression tasks. It seems like the outcome and propensity scores tasks were two separate tasks. Perhaps there is some conceptual way we can combine them and then only do the task once but have some result that is able to use that information and have “applied it twice” in a causal setting.

Works Cited

- https://hastie.su.domains/ISLR2/ISLRv2_website.pdf