IHDP and JOBS benchmark causal effect estimations using Machine Learning methods

April 28, 2022

Registration number: 2103133
Project: Imbalanced datasets, Causal inference
Link to GitHub: https://github.com/A-ngelG/CE888/tree/main/project

Executive summary (max. 250 words)	190
Introduction (max. 600 words)	600
Data (max. 300 words/dataset)	469
Methodology (max. 600 words)	544
Results and Discussion (max. 1000 words combined)	536
Conclusions (max. 500 words)	146
Total word count	2484

Contents

Abstract

The purpose of this paper is to analyse two benchmark datasets, IHDP and JOBS and respond the causal questions of both datasets which is "Does especial assistance enhance cognitive test score of low weight premature born infants?" and "Is employability affected by job training?". This because of the arising interest of industry in causal inference, thus the importance of learning and practicing it. The methodology of the project consists of doing basic data preprocessing and developing machine learning models for approximating the effects of the treatment variables what would be the causal question premise. For this models found in literature will try to be replicated so that it validates the outcomes of owns models. The resulting outcomes came out different than expected as they were supposed to mimic those of other researchers, still assuming the results from the models implemented were correct, infants do get higher cognitive scores when receiving especial support and job training might not be an indicator of employability. Though the results weren't those expected it is true that using bias reducing methods such as Inverse Propensity Weighting can reduce errors in the model training and testing.

1 Introduction

Causation happens when a cause is partly responsible for some effect[10]. It is easy to think that correlation implies causation, and has been done since ever. Long time ago sacrifices were made to please gods in order to receive fertility and rain for their land to prosper, now we know that such actions are not related to any of these outcomes, but still entire civilizations based entire cultures around this.Though some outcomes given some cause might seem logical,the relationship might be wrong (or right perhaps), and could be caused by some other variable which we do not know of.
The objective of causal inference is to measure the effect of anything that we believe causes an outcome, which is called "treatment" on an observable outcome,which is called "factual" in order to verify if the correlation between these two actually makes sense and has statistical proof. It is proven that the best way to infer causality is through randomized controlled trials[10][1], in which participants are randomly assigned to either a control or a treatment group, but doing this kind of trials is often unfeasible due circumstances, such as quantity of population, money, time and ethics.
This takes us to observational data, which has its own problems, such as obtaining the counterfactuals, because sometimes there is no way to turn back into the previous state once an action is done, thus leaving us with the "what would have happened if".
According to[4] three types of bias can be produced by observational data:

- The confounding bias,which is result of not considering a third(confounder) variable that affects the association between exposure and outcome,thus creating a false relationship between them.
- The selection bias occurs when the participants are inappropriately selected or the process of selection results in a error in association or outcome.
- The measurement bias involves the usage of imprecise data collection methods, that consequently creates errors in the assessment of data.

Though observational studies can have the previously mentioned disadvantages, many approaches have been developed to minimise their effect, such as propensity scores, co-variate adjustments or latent variables. These approaches can either be used alone or along with others to reduce biases.

Nowadays causal inference is becoming increasingly important in industry, due to the fact that some limitations related hardware capabilities are overcame with the usage of either cloud computing or better infrastructure to analyse the data with. Many titans in industry[8] are already working on understanding the causes responsible for their observations,such as Uber, that is using causal inference on observational data to obtain insights on how to improve their user experience, and help them identify points relevant to boosting their own business.[5]

All this advancements in causal inference are definitely going to have a huge impact on the industry, and likewise the industry interest in causal inference will help further develop previous methods and compensate the disadvantages of working with observational data.

The objective of this project is to implement Machine Learning methods to predict causal effects on two benchmark datasets IHDP and JOBS, in order to answer the causal questions

x1 to x25	Background variables related to the children and their mothers,out of the 25 features, 19 are binary, and 6 are continuous
t	Whether the child had received special support during growth, represents a binary treatment
yf	Factuals which are the measurement of their cognitive test score[3], representing a continuous observed outcome
ycf	Counterfactuals represented by non-observed continuous outcome
ite	The individual treatment effects

Table 1: Table describing IHDP dataset features.

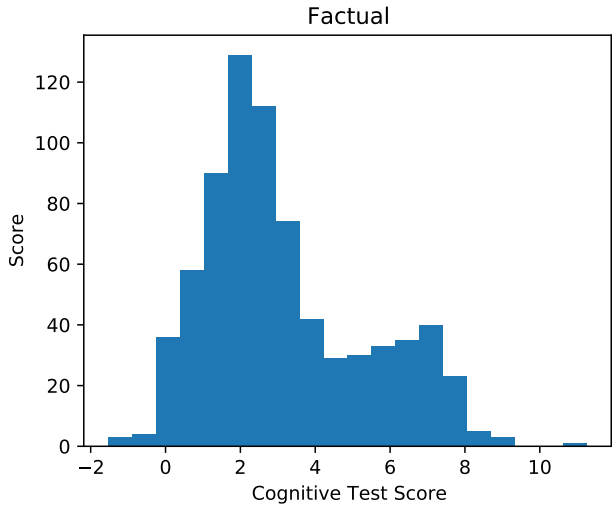

Figure 1: Histogram of Observed cognitive test score

"Does special support during growth improve the cognitive test score of premature-low weight infants?" and "Does job training helps employability?" respectively. This will be accomplished by using models found in the literature in both default parameters and with the hyper-parameters optimised, to later compare results between them and the results encountered in the literature. This project proves to contribute to the existing literature as it will try to replicate results found and will also be explained so that the results achieved can be replicated.

2 Data

2.1 Unwrapping the IHDP Dataset

IHDP is an acronym for The Infant Health Development program, which was a randomized experiment between 1985 and 1988 with the objective of evaluating the efficacy of early intervention in reducing developmental and health problems in premature, low birth weight infants.[3]
The dataset has 747 rows and 29 columns, with their descriptions shown in [1]

Though the dataset was a randomized experimental study, we will be working on an artificial observational one because the portion of the treatment group was dropped[6], thus creating imbalance between the control and treatment groups. Such imbalance might be the cause for the yf graph 1 to be skewed to the right. This dataset contains a total of 747 children of which 608 are in the control group and 139 belong to the treatment group.

In this dataset, in spite of having access to the counterfactuals and individual treatment effect, they will only be used for evaluation, which will be useful for the selection of metrics for measuring whether accurate effects are being predicted. Since we have access to the true individual effect, the usage of the Error on the Average Treatment Effect and the Precision in Estimating the Heterogeneous Treatment Effect can be used as the metrics for this dataset.

The treatment effect is a continuous number, thus requires the use of regressors, still classifica-

x1 to x17	background variables related to the participants such as age,years of school, etc. Out of the 17 features, 7 are binary, and 10 are continuous.
t	Whether the participants had received job training, and represents a binary treatment.
y	Whether the participants regardless of the treatment were employed.
e	Whether the data was experimental, this data is represented in a binary way, where 1 means that it came from the randomized study.

Table 2: Table describing JOBS dataset features.

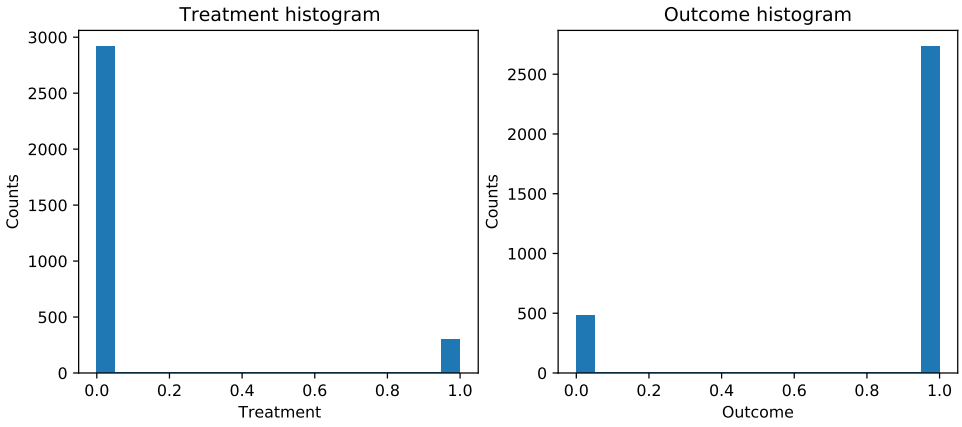

Figure 2: Histograms of treatment and outcome

tion can also be used, depending on the method. A 90/10 train - test split will be carried and will separate the background variables, treatment, factuals, counterfactuals and true individual effects, then a standard scaler will be applied on the resulted background and factual variables to be input in training and prediction of the models.

2.2 Unwrapping the JOBS Dataset

The JOBS dataset contains the data from both a randomized study and observational data from the national survey datasets CPS and PSID,and was created to find the effect of job training on employability and income after training.[9] The dataset has 3212 rows and 20 columns, further descriptions will be presented in [2]

This dataset also has both the outcome and treatment purposely imbalanced as it is seen in graphs 2, as it only has a total of only 297 participants in the treatment group, and 2915 in the control. group.[9]

In this dataset we don't have access to the counterfactuals, which gives insight on the types of metrics that are be adequate to this dataset, such as Average treatment effect on the treated or the Policy Risk. The outcome in this dataset is response, employed or not, therefore it is necessary that the methods act as binary classifiers. // Regarding the pre processing in this dataset, a 90/10 train - test split will executed on the the background variables, treatment, outcome, and whether data was experimental and after that the background variables for the train and test will have a standard scaler with the default parameters applied.

3 Methodology

After the preprocessing, each dataset will be used for training and testing different machine learning methods used in the literature. For the IHDP dataset, a support vector regressor will be used as it offered the best results in [1], the difference that will be is that this classifier will be first trained with its default parameters which are rbf as the kernel, with gamma scaled, C of 1. Then the classifier will be fit and predict effects for treatments of 0 and 1 and these will be substracted to get the predicted ITE.

After that another SVR classifier with the same parameters will be trained but this time it will classify with propensity scorers, which will be computed by a Random Forest Classifier with default

	Method	ATE mean	CI lower	CI upper
0	SVR	1.488601	1.366774	1.610427
1	SVR GS	3.848232	3.835509	3.860955
2	SVR (IPW)	1.861591	1.710667	2.012515
3	SVR (IPW) GS	4.097649	4.097649	4.097649
4	TL (RF))	3.964189	3.753636	4.174743
5	TL (RF)) GS	3.993605	3.791770	4.195441
6	Real	4.112316	0.000000	0.000000

Table 3: Table showing ATE mean across test set of the IHDP dataset

parameters, these being a 100 estimators, without max depth, with at least 2 examples per split and 1 sample per leaf. Then again, the classifier will be fit and be used to calculate the predicted ITE.
A T-Learner with a Random Forest Regressor as its output estimator for the background variables and treatment will be implemented, fit and used to predict the ITEs, as this was also used in [2]. The RF also have its default parameters. Finally, these exact models will each of their estimators with optimised hyper-parameters by using a grid-search with a 5-fold cross validation.For the trees the optimised parameters will be the maximum leaf nodes and maximum depth and will be tuned with values of 10,20,30 and None for the leaves and 5,10,20 and None for the depth, and for the SVR the parameters optimised will be the kernel, gamma and C, for a linear and a rbf kernel with gamma of 0.1,0.05,0.01 and 0.001 and C of 1,10 and 100. Needless to say the new models will be also fit and used for predicting the ITEs.
Ultimately the results will be compared between non tuned and tuned hyper-parameter models with by their ATE error and PEHE, this due to the fact that we have the true ITEs of the model.

For the jobs dataset we will be using some of the methods used in [7] and [9]. such as Random Forests, and causal forests. And will follow almost the same approach for the classifiers used in IHDP, but in this case instead of a SVR a Random Forest Classifier will be used, due to the fact that the outcome is a binary result. As explained previously but now with a RFC, first a RFC with default parameters will be trained, fit and be used to predict the ITEs, then propensity scores will be calculated and used in a classifier for obtaining weights that will be used in the next RFC for the prediction of ITEs, in the same way as before and lastly a Causal forest will be trained and used like that too. After that each of these models will have their hyper-parameters optimised with a grid-search with a 5-fold cross validation and wit same parameters that were used for the tuning the RFs in the IHDP dataset for their respective training,fit and prediction of effects. And finally these models will be compared by their ATT and policy risk as the true value for the effects isn't available in this dataset, which is also validated in [7].

4 Results

Looking at the results shown in [3][4][3] it is possible to see that the model's error is decreasing with the time and starts doing more accurate predictions, being the normal SVR with default parameters the one with the biggest error both in ATE and in PEHE, but is also clearly possible to see the contrast between a default SVR and a SVR with tuned hyper-parameters for this dataset. The SVR (IPW) model also greatly reduced errors with the hyper-parameter tuning, reducing it even more the default T-Learner (RF) in the ATE error and making it the closest to the average ATE of the True effect. Not only that but the error was also diminished when the gridsearch was used with the T-Learner which in [3] can be seen as a very good approximation of the true effect, even with outliers in zones close to where they originally where.

In the JOBS dataset it seems that the models implemented did a decent job since most errors are very little but still show a decrement in the ATT error by using more complex models or by just tuning the hyper-parameters while maintaining a decent Policy Risk.

5 Discussion

First of all its important to note that due to the usage of inverse propensity score weighting it was possible to address the bias that came from the observational imbalanced data, since in a

	Method	ATE test	PEHE test
0	SVR	2.623716	2.736061
1	SVR GS	0.264084	0.794436
2	SVR (IPW)	2.250725	2.396688
3	SVR (IPW) GS	0.014668	0.772650
4	T-Learner (RF)	0.148127	0.595032
5	T-Learner (RF) GS	0.118711	0.545021

Table 4: Table showing ATE error and PEHE in test set of the IHDP dataset

	Method	ATT mean	CI lower	CI upper
0	RF	0.045101	0.024442	0.06576
1	RF GS	0.045101	0.024442	0.06576
2	RF (IPW)	0.038880	0.018601	0.05916
3	RF (IPW) GS	0.037325	0.021413	0.053237

Table 5: Table showing ATT error and Policy Risk in test set of the JOBS dataset

	Method	ATE test	PEHE test
0	RF	0.045101	2.736061
1	RF GS	0.045101	0.794436
2	RF (IPW)	0.038880	2.396688
3	RF (IPW) GS	0.037325	0.772650

Table 6: Table showing ATT error and Policy Risk in test set of the JOBS dataset

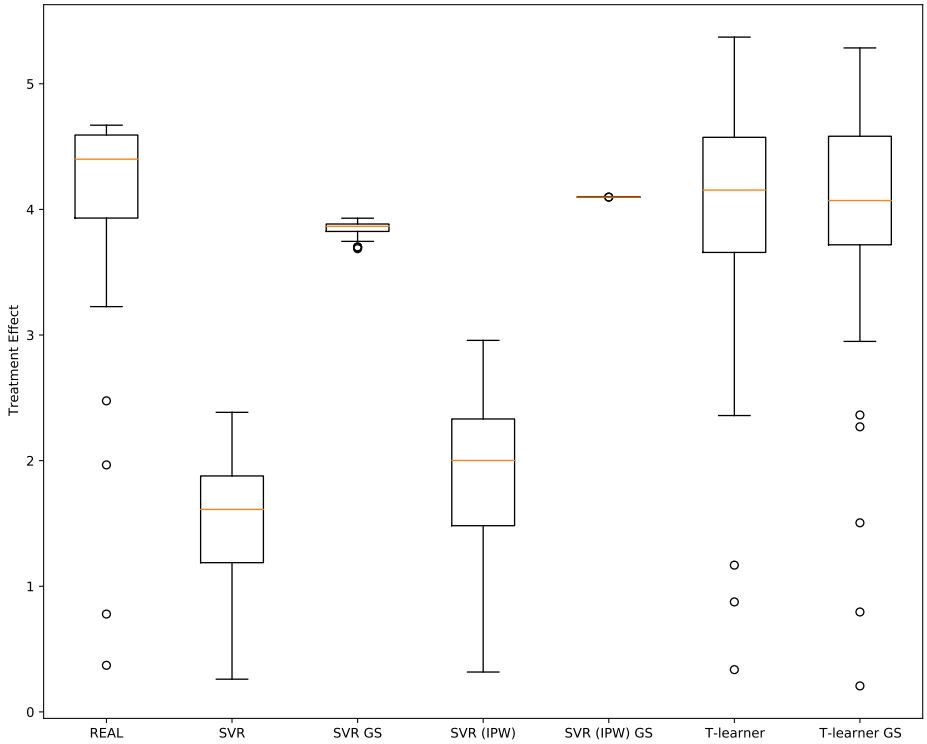

Figure 3: Box-plot of Treatment effect(Y axis) vs Model

way makes the sample looks larger because the weights are related to the likelihood, and gives more weight to those less likely to appear giving also much more importance to outliers. This is likely to be the reason for the IPSW models (especially those tuned) to perform to the level of CATE-estimators like the T-Learner.

In the IHDP dataset When comparing ATE and PEHE between designed models it might seem great that the error values reduce drastically as some more complex models are implemented or bias decreasing techniques are used or hyper-parameter tuning is added. The real problem or doubt comes next, because when comparing the results from this experiment to those of [**?**] [9] [7] and [2] its possible to note that their experiments sustained a larger error in both training and test. which makes me infer that something might be wrong with the models developed during this project, specially the inability to replicate [1] results (when the results should have been at least close) , most likely some sort of over-fitting caused by just doing a train-test split and not formal cross validation but is it?

Regarding the JOBS dataset, [9] uses RFs and gets a Risk Policy of around .28 and a ATT error of around 0.09, though not so far, its still far from the expected values compared to the ones in this experiment of 0.23 and 0.04 respectively, which are values that couldn't be reached with neither algorithms implemented by [9].

Assuming that the outcomes from models developed in this project are correct, the causal questions for both datasets can be answered. According to the results, there is a positive effect the cognitive test on infants born prematurely and with low-weight when being receiving special assistance. While for the training effect on employability seems to be positive but close to 0, which might show a weak effect or a null effect on employability.

6 Conclusions

The outcome of the project though not the expected one, did gave space for thought, it is an advance and could give an answer to the causal questions. Even if the comparison between other papers outcomes resulted in doubt,the fact that the models worked and reduced their errors accordingly is not enough but something. Still some other tuning should be made in order to fully replicate the results from other papers, and especial focus should be directed towards each of the steps the researches took to their results and not just trying to replicate them by purely identifying a model they used and trying to use it disregarding all other steps.
Regarding reducing the biases, other tools like IPSW (such as latent factors,etc) should be researched and implemented, especially when working with observational data, since its fundamental for the optimal training of the models.

References

[1] H. Borré. *Machine Learning for causal Inference on Observational Data.* PhD thesis, University of Essex, 2018.

[2] A. Caron, G. Baio, and I. Manolopoulou. Estimating individual treatment effects using non-parametric regression models: a review, 2021.

[3] R. T. Gross et al. *Infant health and development program (ihdp): Enhancing the outcomes of low birth weight, premature infants in the united states, 1985-1988.* 1993.

[4] G. Hammerton and M. R. Munafò. Causal inference with observational data: the need for triangulation of evidence. *Psychol. Med.*, 51(4):563–578, Mar. 2021.

[5] T. Harinen and B. Li. Using causal inference to improve the uber user experience, 2019.

[6] J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[7] C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models, 2017.

[8] C. Schmitt. Causal data science in practice, 2021.

[9] U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

[10] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang. A survey on causal inference, 2020.