Data Science and Decision making: Assignment 1

Friday 18th March, 2022

Registration number:	2106136
Project:	(Causal inference)
Link to GitHub:	https://github.com/Ganesh-710/CE888_assignments

Executive summary (max. 250 words)	95
Introduction (max. 600 words)	302
Data (max. 300 words/dataset)	370
Methodology (max. 600 words)	990
Conclusions (max. 500 words)	102
Total word count	1860

Contents

Abstract

Causal inference is a powerful modelling tool for explanatory analysis that could help current machine learning in delivering comprehensible predictions. In this report we will look at possible machine learning methods to find causal effect on the given data sets jobs and Infant Health and Development Program.Machine learning algortihms selection, use of Average treatment effect (ATE), PEHE for dataset with counterfactual effect and ATT, Policy risk for data set without counterfactuals and will also look at inverse probability weighting propensity score and Advance cate estimaters on both jobs and Infant Health and Development Program dataset.

1 Introduction

Causal inference, The process of drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect. Many fields have seen considerable success with machine learning technologies, yet the majority of them lack interpretability. Causal inference is a powerful modelling technique for explanatory analysis, and it may be able to help current machine learning produce understandable predictions. The inherent link between missing data and causal inference (Holland, 1986) is that for each unit, only one of the potential outcomes—the one corresponding to the treatment to which the unit is exposed—is observed, while the other potential outcomes are absent. Potential outcomes of the same unit are sometimes referred to as ”counterfactuals” in the literature because they are never witnessed simultaneously. As a result, causal inference is essentially a missing data problem, and estimating causal effects necessitates careful treatment of the alternative outcomes that are lacking.

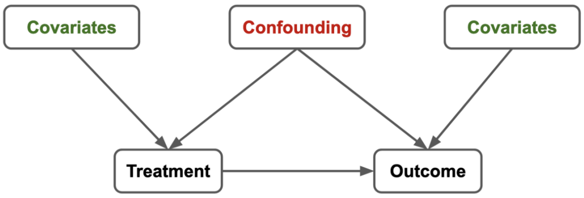


Figure 1: Causality

The most common type of interventional study (Experimental study) is a randomized controlled trial (RCT). RCTs are considered as the best source of evidence for assessing the efficacy of treatment. This enables to minimize or even avoid bias, such as: allocation bias (by means of randomization), selection bias (proper randomization is not achieved), performance and ascertainment biases etc. However, This comes with its logistical, Statistical, and ethical limitations. An alternative to RCTs are passively collected observational data, As it is not randomised we have to be careful about using such dataset. Thanks to modern machine learning,Though we can not obtian causal effect directly through observed outcomes, We can use machine learning techniques to approximate them. This report will explore causal inference datasets IHDP and jobs their characteristics and its relevant performance metrics. To identify causal effect of high quality childcare and home visit

2 Data

2.1 Jobs Dataset

This dataset was proposed by refer is a combination of experiments done by refer as part of the National Supported Work Program (NSWP) and observational data from the Panel Study of Income Dynamixs (PSID). With 17 features it captures basic characteristics of people such as whether they received job training from NSWP (treatment), and their employment status (outcome). Columns t and y holds treatment and factual samples respectively. Column 'e' indicates whether a sample comes from experimental or observational data.

Overall, The dataset contains 3212 samples compressing 2490 and 722 samples of 0.0 and 1 (either obervational or experimental). All the features are numerical and does not contain any missing value eliminating the need for encoding and imputation. As it can be seen from the above figures the background variables 0,9 0,8 1,10 6,15 7,15 6,11 7,11 6,12 7,12 11,12 11,15 8,9

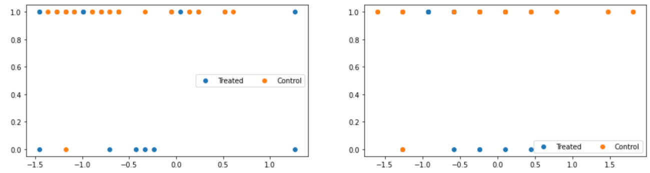


Figure 2:

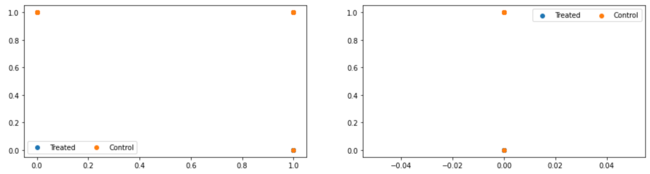


Figure 3: Jobs dataset

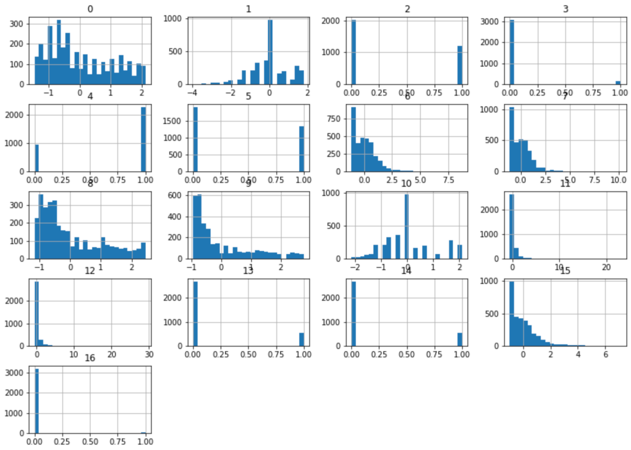
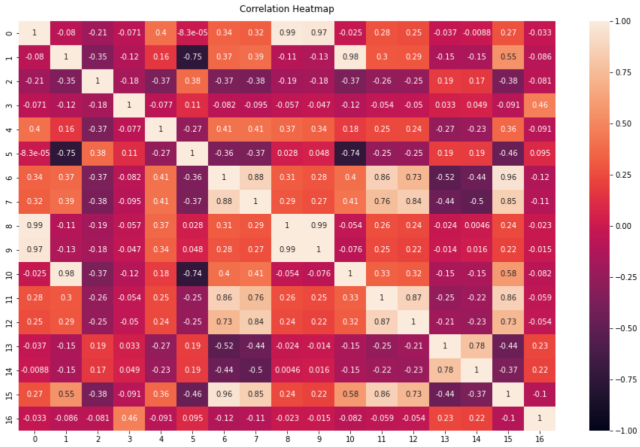


Figure 4: Jobs dataset, Histogram plot



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	1	-0.08	-0.21	-0.071	0.4	-8.3e-05	0.34	0.32	0.99	0.97	-0.025	0.28	0.25	-0.037	-0.0088	0.27	-0.033
1	-0.08	1	-0.35	-0.12	0.16	-0.75	0.37	0.39	-0.11	-0.13	0.98	0.3	0.29	-0.15	-0.15	0.55	-0.086
2	-0.21	-0.35	1	-0.18	-0.37	0.38	-0.37	-0.38	-0.19	-0.18	-0.37	-0.26	-0.25	0.19	0.17	-0.38	-0.081
3	-0.071	-0.12	-0.18	1	-0.077	0.11	-0.082	-0.095	-0.057	-0.047	-0.12	-0.054	-0.05	0.033	0.049	-0.091	0.46
4	0.4	0.16	-0.37	-0.077	1	-0.27	0.41	0.41	0.37	0.34	0.18	0.25	0.24	-0.27	-0.23	0.36	-0.091
5	-8.3e-05	-0.75	0.38	0.11	-0.27	1	-0.36	-0.37	0.028	0.048	-0.74	-0.25	-0.25	0.19	0.19	-0.46	0.095
6	0.34	0.37	-0.37	-0.082	0.41	-0.36	1	0.88	0.31	0.28	0.4	0.86	0.73	-0.52	-0.44	0.96	-0.12
7	0.32	0.39	-0.38	-0.095	0.41	-0.37	0.88	1	0.29	0.27	0.41	0.76	0.84	-0.44	-0.5	0.85	-0.11
8	0.99	-0.11	-0.19	-0.057	0.37	0.028	0.31	0.29	1	0.99	-0.054	0.26	0.24	-0.024	0.0046	0.24	-0.023
9	0.97	-0.13	-0.18	-0.047	0.34	0.048	0.28	0.27	0.99	1	-0.076	0.25	0.22	-0.014	0.016	0.22	-0.015
10	-0.025	0.98	-0.37	-0.12	0.18	-0.74	0.4	0.41	-0.054	-0.076	1	0.33	0.32	-0.15	-0.15	0.58	-0.082
11	0.28	0.3	-0.26	-0.054	0.25	-0.25	0.86	0.76	0.26	0.25	0.33	1	0.87	-0.25	-0.22	0.86	-0.059
12	0.25	0.29	-0.25	-0.05	0.24	-0.25	0.73	0.84	0.24	0.22	0.32	0.87	1	-0.21	-0.23	0.73	-0.054
13	-0.037	-0.15	0.19	0.033	-0.27	0.19	-0.52	-0.44	-0.024	-0.014	-0.15	-0.25	-0.21	1	0.78	-0.44	0.23
14	-0.0088	-0.15	0.17	0.049	-0.23	0.19	-0.44	-0.5	0.0046	0.016	-0.15	-0.22	-0.23	0.78	1	-0.37	0.22
15	0.27	0.55	-0.38	-0.091	0.36	-0.46	0.96	0.85	0.24	0.22	0.58	0.86	0.73	-0.44	-0.37	1	-0.1
16	-0.033	-0.086	-0.081	0.46	-0.091	0.095	-0.12	-0.11	-0.023	-0.015	-0.082	-0.059	-0.054	0.23	0.22	-0.1	1

Figure 5: Jobs dataset, Correlation Heatmap

have maximum high correlation in-between and from the above the it is clear that some columns contain binary data therefore, they do not require data standardization. The train test split of 80:20 is done using sci-kit learn for testing and validation purposes. Random forest regressor is used to tackle this problem after doing trail and error with its available alternative due to its high performance and low error.

2.2 IHDP Dataset

The Infant Health and Development Program dataset was gathered to look at the impact of high-quality childcare and home visits on low-birth-weight, preterm infants' future cognitive test scores. It has 25 Numeric feature variables, including measurements of the child (e.g., child-birth weight, head circumference, weeks born preterm, birth order, first born, neonatal health index, sex...), as well as information about the mother at the time of delivery. Same as the jobs data, in IHDP dataset all the columns are numerical and does the contain any missing or null value. Based on the graph above the dataset also contains columns with binary data therefore eliminating the need for standardization of those columns.However, the correlation between the background variables are very less.

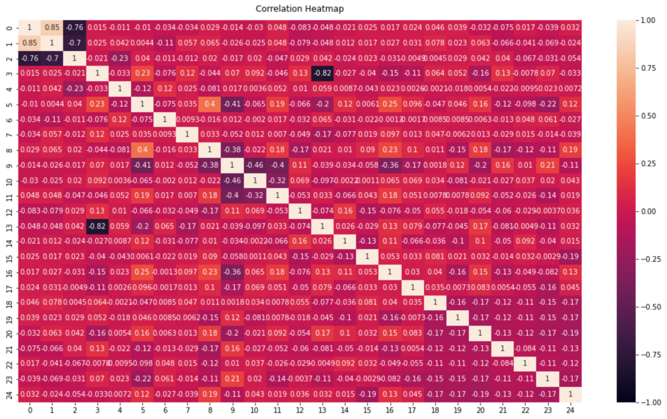


Figure 6: IHDP dataset, Correlation Heatmap

Since the target variable is continuous This is a regression problem . The train test split of 80:20 is done using sci-kit learn for testing and validation purposes.

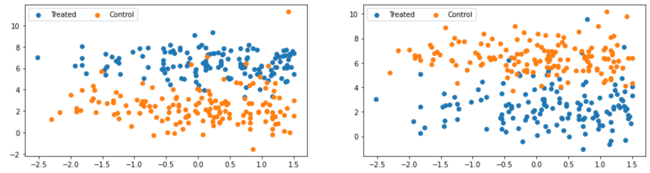


Figure 7:

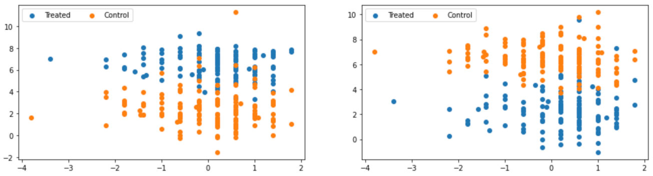


Figure 8: ihdp dataset

3 Methodology

In this section we describe the machine learning Methodology that will be used in this causal effect problem. Considering the given problem random forest regressor proposed by Leo Breiman, Decision trees solves Both regression and classification problems . They flow visually like trees, hence the name, and in the case of regression, they begin at the root of the tree and follow splits based on variable outcomes until they reach a leaf node, where the result is presented. When the answer is 'yes,' the decision tree takes the represented path; when the answer is 'no,' the decision tree takes the opposite path. This process is repeated until the decision tree reaches the leaf node, at which point the decision is made. The values of a, b, c, and d in the example above might represent any numeric or categorical value. Ensemble learning is the act of combining many models that have been trained on the same data and then averaging the results to achieve a more powerful predictive/classification result. The goal of ensemble learning is for the errors of each model (in this case, the decision tree) to be independent and varied from tree to tree.Bootstrapping is the process of sampling subsets of a dataset at random over a specified number of repetitions and variables. After that, the results are averaged to produce a more powerful result. An applied ensemble model is an example of bootstrapping.

The bootstrapping Random Forest approach combines ensemble learning methods with the decision tree framework to generate many randomly generated decision trees from data, then average the results to produce a new result that frequently leads to good predictions/classifications. Before applying any machine learning algorithm we Standardize features by removing the mean and scaling to unit variance.

$$z = \frac{(x-u)}{s}$$

Many machine learning estimators require dataset standardisation: if the individual features do not more or less resemble standard normally distributed data, they may perform poorly.Many parts of a learning algorithm's objective function, for example, assume that all features are centred about 0 and exhibit variance in the same order. If a feature has orders of magnitude more variance than others, it may dominate the objective function, preventing the estimator from learning from other characteristics as planned. For the given **jobs** dataset a random forest regressor has been used.After scaling the data we do hyperparameter optimization, Optimization is must to specify hyperparameters in machine learning models in order to customise the model to our dataset. The general effects of hyperparameters on a model are frequently recognised, but determining the appropriate hyperparameter and combinations of interacting hyperparameters for a given dataset can be difficult. For configuring hyperparameters, there are frequently generalized heuristics or rules of thumb. A better strategy is to objectively search different values for model hyperparameters and select the subset that results in the best model on a given dataset. The scikit-learn Python machine learning package has a feature called hyperparameter optimization or hyperparameter tuning. A hyperparameter optimization provides a single set of high-performing hyperparameters with which we may configure our model.

After the implementation, optimization and prediction of random forest regressor the estimated effect is then measured by [2] individual treament effect methodology. The difference between two possible outcomes is described as the individualised treatment effect.

$$ITE^{(i)} = Y_1^{(i)} - Y^{(i)}$$

However, in practise only a single potential outcome will be observed for each individual. The determination of (xi) based on observed data necessitates the use of assumptions to supplement the facts. After finding ITE, we focus on Metrics especially for the given dataset which has no Without effect/counterfactuals for dataset without counterfactuals Policy Risk and ATT can be used for evaluation metrics purposes. The IHDP dataset follows similer methodology apart from the metrics which in this case comes with counterfactuals data. Since it comes with effect/counterfactuals data we utlize ATE and PEHE as metrics.. The PEHE measure is driven by the fact that we are often interested in an ITE estimation of an individual using only observed covariates X in many circumstances. One such instance is when a policy judgement must be made on whether or not an individual should receive treatment. In such circumstances, neither W nor Y are seen when a decision is made. A good PEHE, it appears, requires accurate estimates of both factual and counterfactual outcomes, or at least the difference between the two, not just the counterfactual outcome.

$$ATE = \frac{1}{n}\sum_{i=1}^{n} ITE'^{i}$$

The Average Treatment Effect (ATE) is defined as the average of the population's individual treatment effects. Furthermore, the Average Treatment Effect of the Treated (ATT) is the sum of the individual treatment effects of individuals who have been treated (hence not the entire population).

$$ATT = \frac{1}{N_1}\sum_i (Y_i^1 - Y_i^0)$$

and policy risk which calculates the risk of the policy defined by predicted effect.

$$\epsilon_{PHPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(ITE'^i - ITE^i)^2}$$

A propensity score is the probability of a unit (e.g., person, classroom, school) being assigned to a particular treatment given a set of observed covariates. Propensity scores are used to reduce selection bias by equating groups based on these covariates

$$e(x) = P(t_i = 1|x_i = x)$$

Inverse probability weighting is a statistical technique for calculating statistics standardized to a pseudo-population different from that in which the data was collected. Study designs with a disparate sampling population and population of target inference (target population) are common in application.

$$w_i = \frac{t_i}{e'(x_i)} + \frac{1-t_i}{1-e'(x_i)}$$

In reality, the control group often has more data than the treatment group. Using T-learner in cases where the treatment group has few data points would be problematic; we'd want to avoid overfitting for the treatment group when the data is considerably less. To avoid this issue, consider whether we can estimate the treatment group using data from the control group. This is what X-learner [1]aims to accomplish: it takes data from the control group to develop stronger estimators for the treatment group, and vice versa. It's based on T-learner and employs a 'X' shape to represent each observation in the training set. Counterfactuals are used in this situation.

$$= g(x)_0(x) + (1 - g(x))_1(x)$$

4 Conclusions

To sum up, In this report we have looked to possible approachs for finding causal effect on the given data sets jobs and Infant Health and Development Program. I have picked linear regression and random forest for the experiment both standard and hyper parameter optimised and will use Average treatment effect (ATE), PEHE for dataset with counterfactual effect and ATT, Policy risk for data set without counterfactuals and will also use inverse probability weighting propensity score, Advance cate estimater such as X-learner on both jobs and Infant Health and Development Program dataset.

References

[1] Kunzel. Meta-learners for estimating heterogeneous treatment effects using machine learning. *TUGBoat*, 14(3):4156–4165, 2019.

[2] . Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *TUGBoat*, 113(3):1228–1242, 2018.