CCT College Dublin
MSc in Data Analytics
Capstone Project

Evaluating current causal inference methods in data analytics, and demonstrating novel results in projects where they are added

Milo Moran sbs23081

Supervisor: James Garza

23rd February 2024

This report was written entirely by the author, except where stated otherwise. The source of any material not created by the author has been clearly referenced. The work described in this report was conducted by the author, except where stated otherwise.

Abstract

Acknowledgements

Contents

List of tables/figures

List of Abbreviations

μLEDs: Micro light emitting diodes
LCD: Liquid crystal display
UV: Ultra violet
ASE: Amplified spontaneous emission

1 Intro

Things to be introduced in the introduction what is ATE. what is the PoF and the SMC. names of types of methods?

2 Literature Review

While correlations, predictions, and forecasts are all common tasks in data analytics, causal analysis is significantly less common.

2.1 Motivations

We often talk about correlation vs causation in theory but while implementing Data Science solutions towards solving business problems not much influence is given to validating causation amongst independent and dependent features. Global AI (2022)

Much of the novelty of this project comes from it's attempts to demonstrate the power of causal inference methods and their viability for pointing towards answers to causal questions in all data based quantitative research. The importance of an uptake of these methods is argued in Hernan, Hsu and Healy (Hernán et al. 2019). In the article, the authors claim that there is a historic opportunity to redefine data analysis to naturally accommodate a science - wide framework for causal inference from observational data.

Gelman & Imbens (2013) considers the difference between forward causal questions seeking to learn effects of causes, and reverse causal inference which looks for the causes of effects -NOT VERY RELEVANT

2.2 Frameworks

There exists two primary frameworks or models for causal inference. There is the Structural Causal Method (SCM) associated with Judea Pearl (Pearl 2009), and there is the Rubin Causal Model(RCM) associated with Donald Rubin, also known as the potential outcome framework(Imbens & Rubin 2015).

Ibeling & Icard Aliprantis (2015) claims that while the two methods are often viewed as analogues, there is a distinction. This working paper demonstrated this

2.3 Methods

Several large scale surveys can be found that on CIM's are part of 'A Survey on Causal Inference'(Yao et al. 2021) is a review of the potential outcome framework of causal inference. In Table 3, the paper lists 4 tool boxes where various methods can be found and used, some in Python -DoWhy(Sharma & Kiciman 2020), CausalML(Chen et al. 2020) and EconML(Microsoft Research 2019), and one in R - causalToolbox(Künzel et al. 2019). Table 4 lists individual methods as well as providing sources for each. These tables will be used as a valuable

reference for this research going forward Structural causal model, 102,105,107 Sharma & Kiciman (2020) describes the DoWhy library, and explains all the features it has to enable for end to end causal inference.

2.3.1 Assumption independent methods

What is assumed by each method? Does using methods independent of assumptions add value in terms of breadth of applicability or ease of use etc? Yao et al. (2021) divides methods based on whether they are independent of the assumptions of the potential outcome framework or not. In section 4, it describes the workarounds that researchers have used to infer causality in scenarios where the assumptions or parts thereof do not hold. This is important for this research project, as when dealing real world projects, it is unlikely that all three assumptions of the framework will be met, and being able to use causal inference methods in these situations is invaluable. When discussing future directions in section 7, the paper notes the need for more research into cases in which the assumptions of the causal model can be relaxed, as practical settings frequently do not follow all the assumptions and current methods may not always be appropriate. This outlines a potential hurdle for this research, as it may not be possible to accurately apply the causal model in some practical projects.

2.4 Evaluation

There are three general approaches to validation of causal methods: Firstly the face validity test, whereby the result is compared with the intuition of an expert in the domain. Second the placebo test, where the nature of the data allows for the separation into placebo and treated groups based on either time or selection within the sample. Thirdly, synthetic data is used for testing methods based on a known created treatment effect. Schuler et al.(2017) proposes Synth Validation, whereby simulations are used to test and select causal inference methods for use in given scenarios, allowing the most appropriate method to be used for that scenario. The paper notes that previous research has been inconsistent and has failed to find any one-size-fits-all methods that tests methods against hand crafted benchmark data

Parikh et al.(2022) what large scale evaluations have been done and what was found?(Dorie et al. 2019) An thorough survey of causal inference for time series analysis has already been conducted(Moraffah et al. 2021). It covers the various questions and methods as well as for evaluating the results of those methods Do I need to go into massive detail on just how optimised, or just how much detail I will need to justify use of given methods?

2.5 Current advice and guidelines for real world applications and so on

Stuart(2010) provides a review of the state of matching methods across various disciplines, and attempts to bring together much of the research, as well as

making suggestions as to where the literature should be headed. It

Kerzner (2022) is a guide to causal inference that thoroughly covers much of the overarching steps in implementing causal inference, and would make for a strong introduction to the subject for a researcher that already has a grasp of basic statistics. The article makes many practical recommendations for how to deal with common problems along with referencing best practice. While it is not well structured, this guide is an example of an excellent starting point to encourage the use of causal inference to the right audience.

2.5.1 Project selection?

What areas would be promising in terms of domain what types of data would be appropriate in terms of maximal variation of the typical case? where do I go to find them? kaggle is a bad idea right? (assuming I can source the sets.)

2.6 Lit Review conclusions

3 Defining Evaluation Criteria

In choosing criteria for the selection of methods, the exact aims of the selection, and the proposed use cases, must be very clearly defined and justified. The criteria of "ease of use" includes ease of access, and so a massively important criteria is that the method needs to be open sourced. Three large "tool-boxes" in Python dedicated specifically to providing access to these kinds of methods are DoWhy, CausalML, and EconML. These libraries contain a large number of methods along with tutorials to assist users in applying causal inference methods as part of their research. This makes the methods available from these libraries an excellent starting point for evaluation in line with the goal of finding easy to use CIM's

Can I fully justify using only methods from these sources?

Other evaluation criteria involve

4 Comparing methods using evaluation system

I really only need to do

5

6 Conclusions

References

Aliprantis, D. (2015), 'A distinction between causal effects in structural and rubin causal models'.

Chen, H., Harinen, T., Lee, J.-Y., Yung, M. & Zhao, Z. (2020), 'Causalml: Python package for causal machine learning', *arXiv preprint arXiv:2002.11631* .

Dorie, V., Hill, J., Shalit, U., Scott, M. & Cervone, D. (2019), 'Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition', *Statistical Science* **34**(1), 43 – 68.
URL: *https://doi.org/10.1214/18-STS667*

Gelman, A. & Imbens, G. (2013), 'Nber working paper series why ask why? forward causal inference and reverse causal questions'.

Global AI, P. C. 2022), 'Causal analysis utilizing causalml', Available at `https://medium.com/@publiciscommerce/causal-analysis-utilizing-causalml-bfabd8015860` (29/12/2023).

Hernán, M. A., Hsu, J. & Healy, B. (2019), 'A second chance to get causal inference right: A classification of data science tasks', *CHANCE* **32**(1), 42–49.
URL: *https://doi.org/10.1080/09332480.2019.1579578*

Ibeling, D. & Icard, T. (2023), 'Comparing causal frameworks: Potential outcomes, structural models, graphs, and abstractions'.

Imbens, G. W. & Rubin, D. B. (2015), *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.

Kerzner, S. (2022), 'A complete guide to causal inference', Available at `http://www.https://towardsdatascience.com/a-complete-guide-to-causal-inference-8d5aaca68a47` (29/12/2023).

Künzel, S. R., Walter, S. J. & Sekhon, J. S. (2019), 'Causaltoolbox—estimator stability for heterogeneous treatment effects', *Observational Studies* **5**(2), 105–117.

Microsoft Research (2019), 'Econml: A python package for ml-based heterogeneous treatment effects estimation'.

Moraffah, R., Sheth, P., Karami, M., Bhattacharya, A., Wang, Q., Tahir, A., Raglin, A. & Liu, H. (2021), 'Causal inference for time series analysis: problems, methods and evaluation', *Knowledge and Information Systems* **63**, 3041–3085.

Parikh, H., Varjao, C., Xu, L. & Tchetgen, E. T. (2022), 'Validating causal inference methods'.
URL: *https://proceedings.mlr.press/v162/parikh22a.html*

Pearl, J. (2009), *Causality*, Cambridge university press.

Schuler, A., Jung, K., Tibshirani, R., Hastie, T. & Shah, N. (2017), 'Synth-validation: Selecting the best causal inference method for a given dataset', *arXiv preprint arXiv:1711.00083* .

Sharma, A. & Kiciman, E. (2020), 'Dowhy: An end-to-end library for causal inference', *arXiv preprint arXiv:2011.04216* .

Stuart, E. A. (2010), 'Matching methods for causal inference: A review and a look forward', *Statistical Science* **25**, 1–21.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J. & Zhang, A. (2021), 'A survey on causal inference', *ACM Transactions on Knowledge Discovery from Data* **15**, 1–46.