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

ohe,

1 Intro

Talk about feature analysis, Statistical testing talk about Dimensionality talk about classifiers Hyperparameter tuning Feature importance scores then explain causal vs predictive. then e

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). The article outlines the historical context whereby causal inference from observational questions has been suppressed by mainstream statistics, and claims that appropriate integration of causal inference and counterfactual prediction into data analytics

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

Brady, Henry and Iris Hui. 2006. "Is it Worth Going the Extra Mile to Improve Causal Inference?"." Paper presented at the 23rd Annual Summer Meeting of the Society of Political Methodology

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.

2.3.2 Model selection

Scott (2019) compares Heckit models, Propensity Score Matching, and Instrumental Variable models, demonstrating their performances in various scenarios through simulation. Conditions are varied in terms of selection observables, selection unobservables, and outcome unobservables, and the

Stuart (2010) summarises previous recommendations for selecting matching methods in section 6.1.3 as well as proposing guidelines for their use in section 6.2. In the former section the paper references balance as being a key factor for best method selection, and suggests 3 possible criteria for choosing a method:

> (1) the method that yields the smallest standardized difference of means across the largest number of covariates, (2) the method that minimizes the standardized difference of means of a few particularly prognostic covariates, and (3) the method that results in the fewest number of "large" standardized differences of means (greater than 0.25).

The paper references a method which automates the process (Diamond & Sekhon 2013) using a genetic matching algorithm to reduce bias and error of estimated causal effects. In section 6.2, it summarises 6 points of guidance for practice, how to decide what covariates to include, how to choosse distance measures, examining the implications of the estimand and choosing the appropriate one, implementing the appropriate matching method, and evaluating the covariate balance.

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

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 Methodology

4.1 Research Objectives

4.2 Primary Research

Experimental research has been chosen as the primary research methodology for this project. Experiments will be performed to test the hypothesis "Does performing ML prediction on a datasets and producing feature importance metrics provide more insight into causal relationships with the target variable than statistical testing alone?"

To test this hypothesis, ML algorithms will be used to make predictions and produce feature importance metrics. The most important features as per these metrics will then be used to infer causal relationships with the target variable. Statistical testing will also be used to determine the most correlated features with the target variable, and the causal inference will also be performed on these results. The ATE measurements from both domains will be compared to determine if there is significant difference between the two methods. ATE's from features selected via purely statistical analysis will serve as the control group, in order to evaluate the performance of the ATE's from the Feature importance methods.

For this experiment, all four factors will be satisfied. There will be concomitant variation between the independent variable describing applied/ not applied and the dependent variable describing novel results found/not found, there will also be temporal sequence of the states prior to and after the methods being applied. All the research provided by the literature review will comprise the theoretical support. Finally, there will be complete control of the system where nothing will be introduced that could influence the dependent variable aside from the treatment.

4.3 The Data

The dataset used for this project describes survey results from a survey of students in higher education at Near East University in Cyprus and was collected from the Faculty of Engineering and Faculty of Educational Sciences students in 2019. It includes personal details, family details, education habits, and performance outcomes. The dataset was sourced from the UC Irvine Machine learning repository. The dataset was first introduced in the paper by Yilmaz & Sekeroglu (2019), in which the authors experimented with the data and used Radial Basis Function Neural Network to achieve accuracy of 70-88%. the dataset is licensed under the Creative Commons Attribution 4.0 International, which allows for the sharing and adaptation of this dataset for any purpose once it is appropriately credited. The dataset contains a total of 33 features and 145 observations. All the data are encoded as values ranging from 0-9. The first feature of the dataset is an identifier for the observations, each one relating to a unique StudentID. This feature has no relevance for analysis and is dropped immediately. The

answers to the questions in fig. 1 correspond to features 1-30 of the dataset. The next feature, COURSE ID, assigns a number to the course the student was undertaking for each of the 9 courses covered. The last feature is GRADE, and represents the grade band that the students final grade fell into.

Personal questions	**Family questions**	**Educational questions**
Age	Mothers'Education	Weekly study hours
Sex	Fathers'Education	Reading (non-scientific)
High School Type	Number of Brother/Sister	Reading (scientific)
Scholarship Type	Parents'Relationship	Attendance to Seminar/Conference
Additional Job	Mothers'Job	Effect of Projects and Activities
Sports/Arts	Fathers'Job	Attendance to Lectures
Relationship		Taking notes
Salary		Writing/Listening
Transportation		Effect of in-class Discussions
Accommodation		Effect of Flip Classroom
		GPA of Last semester
		Expected CGPA at graduation

Table 1: Summary of questionnaire, adapted from Table 1 in Yilmaz & Sekeroglu (2019)

The data is based on personal data for these students, but attempts have been made to anonymise the data. According to the Full Guidance Note on Anonymisation and Pseudonymisation from the Data Protection Commission (2022), personal data that has been irreversibly anonymised ceases to be 'personal data' or require compliance to Data Protection law. To determine that this data is suitably anonymised, identifiability of the subjects must be ruled out. The data set contains no information about unique identifiers relating to the students such as names, phone numbers, student numbers, birthyears/days, or addresses. There is no information that allows for any student to be singled out. Numerical data such as age, salary, and grades are binned which makes identification more difficult A combination of the Course ID factor along with the information given related about the courses in the original paper may be enough to identify the course of some of the students, but the courses feature here doesn't correspond directly to those mentioned in the original paper, and this could be difficult. Some responses to the personal questions regarding family status may increase risk of linkage between values to identify students, but there is no other public data available related to the family status of students with which to corroborate this information, and reidentification in this manner would not be likely. As a result of all these factors, it appears that it is not reasonably likely for the identification of any of the subjects and given the nature of the topic, identification attempts are also unlikely. Therefore the data can be considered to be fully anonymised, and will not need to be treated as personal

data for the purposes of this report.

4.4 Approach

The ultimate objective of the research project is to use causal inference to estimate the causal effect of different interventions, or features, on the final grade of the students based on the survey responses. The causal effect will inferred through the calculation of the ATE of various appropriate features based on feature analysis obtained through several methods. predictive ML models, and compare . . . To this end, the features that best predict the target variable are identified through several methods. Features themselves are investigated via several statistical methods and heuristics to enable better understanding. a subset of features is selected based on results and are then used for the training of machine learning models. The results of preliminary machine learning models are used for a range of comparisons. An XGBClassifier model is tuned for 2 different subsets of the data using 2 different hyperparameter searching methods respectively. The feature importance is

4.4.1 GRADE column and re-binning

In order to better understand and more appropriately treat the target variable, some further interepretation is needed. Turkish grade conversions were used to accompany the categories with percentage ranges between 0 and 100 as visible in Table 1 below. To determine the the spread of the variables a histogram of the variable was included. It shows quite an uneven distribution, with a DD grade being most frequent, and FF and CB grades being particularly underrepresented relative to the other categories

Category	Grade	Percentage Range
7	AA	90.00 - 100.00
6	BA	85.00 - 89.00
5	BB	80.00 - 84.00
4	CB	75.00 - 79.00
3	CC	70.00 - 74.00
2	DC	60.00 - 69.00
1	DD	50.00 - 59.00
0	FF	0.00 - 49.00

Table 2: Corresponding Categories, Grades and Percentage Ranges

There is a question of linearity involved with the treatment of this variable. The encoded 0-7 scale is not representative of the size of the differences between different grades on the 0-100 scale at all. However this isn't wholly inappropriate; the distribution amongst the different grades isn't linear, and the difficulty of increasing from one grade to the next is presumably non linear as well. For design of further research, it might be preferable to use deciles or quartiles of student of grades as bands to predict between.

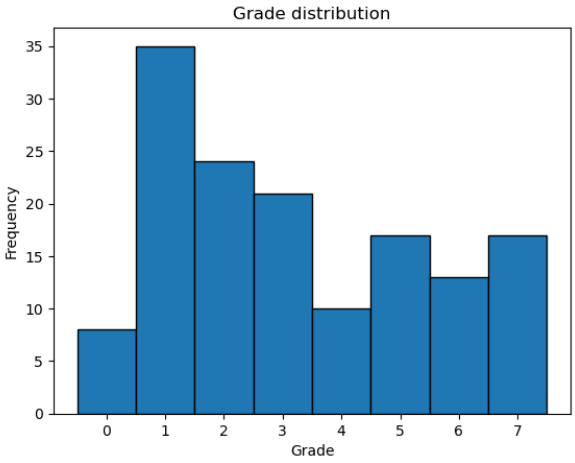

Figure 1: Histogram of the GRADE student feature

One major issue with the full 8 category Grade column is highlighted by the initial decision tree algorithm runs. The amount of total points for classification assuming a 25% testing split of 36/145 observations across 8 categories is very low, with on average just over 4 occurences per category for testing, and often significantly less. Such low amounts make any predictions extremely sensitive to noise, and hinders predicting to any significant degree of accuracy. This issues was further compounded by the uneven distribution of the categories. One choice to help counteract this effect was to increase testing split sizes for the rest of the analysis up to .3, providing slightly more values in each category for testing.

This issue is the primary motivator for the re-binning of the target variable from 4 bins into 8. Another motivator is that this balances the distribution of the dataset significantly vs the 8 category distribution, as can be seen in fig.2. The new percentage bands for the 4 bins can be seen in table 3. As a goal of this analysis is to ascertain whether certain interventions increase or decrease grades, re-binning the values like this improves the interpretability, as while data and granularity are reduced, the overall ability to predict higher or lower is maintained. On top of this, accuracy is improved,

(see the thing!!!!)

and computations become less expensive

4.4.2 Dimensionality

With such a high ratio of features to observations in the dataset, over 1:5 in the initial set, prior to the addition of even more via the One Hot Encoding of non-ordinal categorical features, there is a definite concern that the Curse of Dimensionality will effect model results. As a result, efforts are made in several

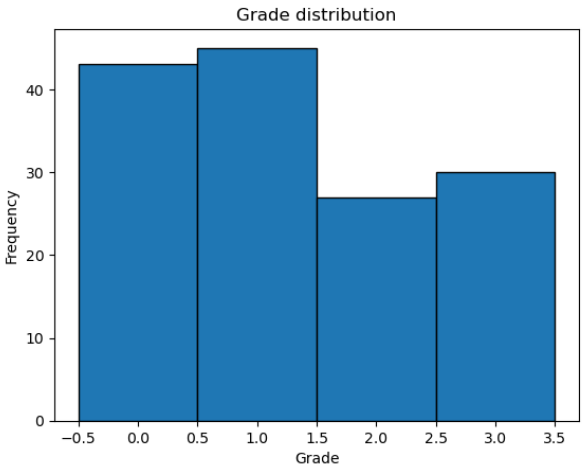

Figure 2: Histogram of the new grade feature created through rebinning

Category	Percentage Range
3	85.00 - 100.00
2	75.00 - 84.00
1	60.00 - 74.00
0	0.00 - 59.00

Table 3: Categories and Percentage Ranges after re-binning

stages to justify the removal of features, on the assumption that doing so will reduce the negative effects of dimensionality. Further, to evaluate the tradeoff between the datasets that contain more and less features, models created using both subsets referred to as small, and large, are compared against each other throughout the analysis.

Models that are largely based on distance metrics, like KNN classifiers, are particularly susceptible to the curse of dimensionality. SVM models are also susceptible through (J. S Marron, Michael J Todd Jeongyoun Ahn (2007) Distance-Weighted Discrimination, Journal of the American Statistical Association, 102:480, 1267-1271, DOI: 10.1198/016214507000001120)

4.5 Predictions

Initially, a Decision Tree classifier(DTC) was used to set benchmarks for further models. Models were run with both size datasets, and compared against the full 8 category grade feature as well as the re-binned 4 category grade feature. Heatmaps and cross validated accuracy scores were used in combination to evaluate model performance visually and quantitatively. Feature importance scores were produced. As these models were primarily investigative, hyperparemeter

tuning was not performed.

Next, a SVM classifier model was used, selected as it could potentially produce evidence of the effect of dimensionality on the predictions. It was performed for both sized datasets, again using cross validated accuracy along with heatmaps to evaluate results. To investigate repeated patterns in the heatmaps, a heatmap for an alternative split of the data was also taken. While coefficient scores similar to feature importance scores do exist for SVM classifiers, they are only applicable for linear kernel SVMs, as other kernels transform the data before creating the model, as described here(BartozKP 2014). Hypereparameter tuning was performed based on recommendations for C and Gamma from (geeksforgeeks 2023) but with the inclusion of several extra kernel options, as the optimal feature space for a model many features was expected to be unpredictable. As there was only 100 possible values with 5 cross validating fits, the GridSearchCV algorithm was used, as it would excecute in an appropriately short amount of time(20sec) and is easily controllable in terms of its values.

Finally, an XGBClassifier method was used. Distributed (Deep) Machine Learning Community (2024) has compiled a list of many of the occasions where XGBoost models have placed first or second in machine learning challenges. Such a widely successful model is a promising choice in this situation and from some other knowledge of

As XGBClassifier models typically are tuned by changing many parameters using large grids or random search spaces, but these are too time consuming and computationally expensive for the scope of this project, so standard grid or random search algorithm methods were ruled out. Instead two different approaches combining several methods were used for hyperparameter tuning. Initially, on the smaller dataset, the method of learning curve adjustment outlined by Brownlee (2021) was used to adjust several important parameters. Based on manual manipulation and inspection of the learning curves, an apparent minimum between 4 hyperparameters was found. These were then selected as the basis for a gridsearchCV, with the remainder of the parameter grid selected based on suggestions from several tutorials. (Navas & Liaw 2022)(Toth 2024) The gridsearch included over 5000 fits, and was time intensive, but significantly less than it would have had the learning curves not been investigated beforehand. For the larger set of the data, in an attempt to address some of the issues with the previous approach, a different method was selected: Bayesian hyperparameter optimisation. One advantage of Bayesian optimisation is that it can be explicitly scaled by running for more and more iterations, depending on time available, and it can approach closer and closer to global minima. The Bayesian optimisation allowed ...

From the tuned XGBClassifier models, feature importance plots were produced using an inbuilt method

4.6 Causal inference

To implement Causal inference, the DoWhy library was used, as it provides a comprehensive framework for Causal inference via it's 4 steps of "model",

"identify", "estimate", and "refute". One of the first steps of causal inference is the creation of a causal graph

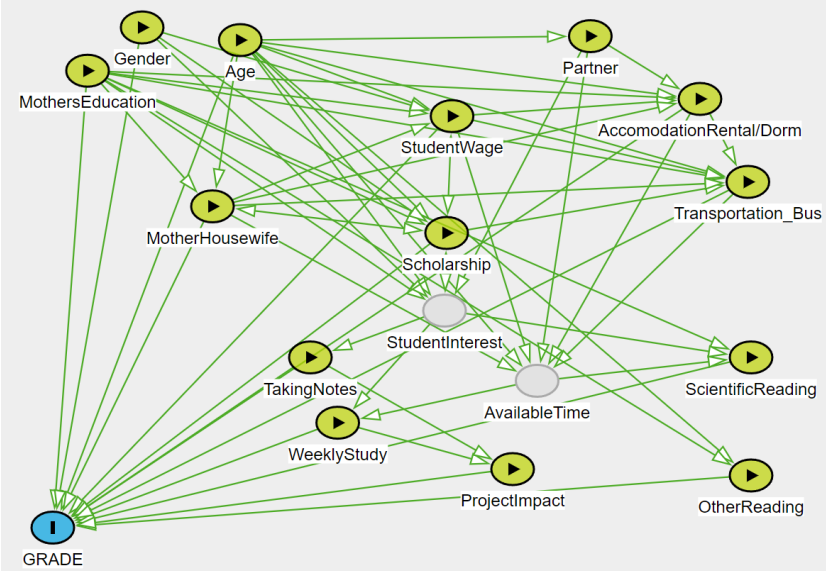

Figure 3: Directed Acyclic Graph for the Data

5 Results

5.1 Dimensionality findings

6 Conclusions

References

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

BartozKP (2014), 'How to obtain features' weights'.
URL: *https://stackoverflow.com/a/21260848*

Brownlee, J. (2021), 'Tune xgboost performance with learning curves'.
URL: *https://machinelearningmastery.com/tune-xgboost-performance-with-learning-curves/*

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

Data Protection Commission (2022), 'Anonymisation and pseudonymisation:full guidance note'.
URL: *https://www.dataprotection.ie/en/dpc-guidance/anonymisation-and-pseudonymisation*

Diamond, A. & Sekhon, J. S. (2013), 'Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies', *Review of Economics and Statistics* **95**(3), 932–945.

Distributed (Deep) Machine Learning Community (2024), 'Awesome xgboost', `https://github.com/dmlc/xgboost/tree/master/demo#awesome-xgboost`.

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*

geeksforgeeks (2023), 'Svm hyperparameter tuning using gridsearchcv — ml'.
URL: *https://www.geeksforgeeks.org/svm-hyperparameter-tuning-using-gridsearchcv-ml/*

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.

Navas, J. & Liaw, R. (2022), 'Guide to xgboost hyperparameter tuning'.
URL: *https://www.anyscale.com/blog/how-to-tune-hyperparameters-on-xgboost*

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* .

Scott, P. W. (2019), 'Causal inference methods for selection on observed and unobserved factors: Propensity score matching, heckit models, and instrumental variable estimation', *Practical Assessment, Research, and Evaluation* **24**(1), 3.

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.

Toth, D. J. (2024), 'Binary classification: Xgboost hyperparameter tuning scenarios by non-exhaustive grid search and cross-validation'.
URL: *https://towardsdatascience.com/binary-classification-xgboost-hyperparameter-tuning-scenarios-by-non-exhaustive-grid-search-and-c261f4ce098d*

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.

Yilmaz, N. & Sekeroglu, B. (2019), Student performance classification using artificial intelligence techniques, *in* 'International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions', Springer, pp. 596–603.