OpportunityFinder: A Framework for Automated Causal Inference

Huy Nguyen*
Amazon.com
USA
nguynnq@amazon.com

Prince Grover*
Amazon.com
USA
pringrov@amazon.com

Devashish Khatwani
Amazon.com
USA
khatwad@amazon.com

ABSTRACT

We introduce OpportunityFinder, a code-less framework for performing a variety of causal inference studies with panel data for non-expert users. In its current state, OpportunityFinder only requires users to provide raw observational data and a configuration file. A pipeline is then triggered that inspects/processes data, chooses the suitable algorithm(s) to execute the causal study. It returns the causal impact of the treatment on the configured outcome, together with sensitivity and robustness results. Causal inference is widely studied and used to estimate the downstream impact of individual's interactions with products and features. It is common that these causal studies are performed by scientists and/or economists periodically. Business stakeholders are often bottle-necked on scientist or economist bandwidth to conduct causal studies. We offer OpportunityFinder as a solution for commonly performed causal studies with four key features: (1) easy to use for both Business Analysts and Scientists, (2) abstraction of multiple algorithms under a single I/O interface, (3) support for causal impact analysis under binary treatment with panel data and (4) dynamic selection of algorithm based on scale of data.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Knowledge representation and reasoning**; **Causal reasoning and diagnostics**;

KEYWORDS

causal inference, double machine learning, neural networks, panel data

*Both authors contributed equally to this research.

1 INTRODUCTION

Automated machine learning (AutoML) frameworks for predictive machine learning (ML) have advanced significantly over the past decade with the introductions of AutoGluon [14], Auto-sklearn [15], H2O [19]. AutoML's biggest advantage is abstracting away the implementation of underlying algorithms and hyper-parameter tuning, and making it easy for scientists and engineers to experiment with a large number of models and identify the one that works best. The demand of AutoML has risen from the fact that no single ML algorithm works best in all scenarios. This has been even more challenging in the causal inference literature. Different methods rely on different set of assumptions [2] for the identification of causal treatment effects[1]: CIA (conditional independence assumption or unconfoundedness), propensity overlap, SUTVA (stable unit treatment value assignment), exchangeability (same outcome distribution would be observed if exposed and unexposed individuals were exchanged) etc.

Causal inference framework DoWhy [22] supports explicit modeling and testing of causal assumptions, but it is still a low level API. AutoCausality, [16] which is built on the top of EconML [8] and DoWhy, supports automated hyperparameter tuning, but it only focuses on the estimation part and assumes that the causal graph provided by the user accurately explains data-generating process. Both AutoCausality and DoWhy do not support panel data[2], which is a mainstream at real-world problems. Most real-world causal studies have panel data of different aggregated granularities, e.g., yearly to daily levels, at different scales, e.g., few individuals to large populations of million entities. To the best of our knowledge, there is no AutoML-like causal inference framework that supports panel data and abstracts away the know-how of causal studies from the users.

In this study, we introduce OpportunityFinder (OPF), our first step in democratizing causal inference techniques. As of our first contribution, Project OPF implements an auto causal inference framework that supports panel and cross-sectional data and offers a wide range of causal inference algorithms. The decision to choose the algorithm is automated and abstracted away from the user. Our second contribution is the automated transformation from input panel data into list of cohort datasets when needed. Cohort-based results are then aggregated for a final result. In the third contribution, OPF provides data visualization to illustrate causal impact. Combining numerical and graphical reports help non-expert users

---

[1]A causal effect can be defined as the difference between hypothetical outcomes that result from two or more alternative treatments, with only one outcome of a treatment being observed each time

[2]Panel data contains observations collected across multiple individuals at a regular frequency, and ordered chronologically.

to verify input data and reason about causal inference results. Current capability of OPF allows non-expert users to carry out the most common causal analysis: *estimating the average treatment effect (ATE) with configurable time horizons for binary actions.* At current state, OpportunityFinder is deployed within AWS account of our organization for internal testing. We are also refactoring OpportunityFinder source as a stand-alone library.

2 LITERATURE REVIEW

Traditional econometric techniques such as propensity score matching, instrumental variable estimation, and difference-in-differences (DiD), offer rigorous methods for estimating average treatment effects under specific assumptions, but often struggle to account for high-dimensional covariates and complex interactions [5]. The Synthetic Control Method (SCM) extends these approaches by constructing a "synthetic" control unit as a weighted combination of potential control units, providing a more flexible comparison for the treated unit [4]. The Generalized Synthetic Control (GSC) further expands SCM by incorporating interactive fixed effects models, thus accommodating multiple treated units and variable treatment periods [26].

Recently, machine learning techniques have been widely integrated into causal inference due to notable works by various teams, e.g., DoubleML [7], EconML [8], CausalML [10]. Double Machine Learning (DML) provides a flexible approach, leveraging machine learning for nuisance parameter estimation while maintaining robustness against mis-specification [12]. Beyond average treatment effect, machine learning enables approaches to estimate individual treatment effects, e.g., heterogeneous treatment effect estimator in EconML and uplift modeling in CausalML. Deep learning methods, such as those based on Neural Networks (NN), have shown promise in estimating individual treatment effects due to their ability to model complex, high-dimensional data, thus uncovering nuanced causal relationships [21].

3 FRAMEWORK DESIGN

The key contributions of our design are (1) integration of several causal modeling models, (2) branching based on type of observational data (cross sectional vs. panel) and number of treatment units, and (3) execution in the users' own AWS environment where they have access to CloudWatch logs for debugging and can visualize the progress. Current OpportunityFinder deployment allows code-less UI without having to move data outside the AWS account as demonstrated in Figure 1. While this design is tied to the MLOps set-up of our organization, OpportunityFinder source code is independent from deployment platforms.

Figure 2 shows the design of OpportunityFinder. Once a user triggers a job, CloudFormation kicks off a set of AWS services including SageMaker, Lambda and Glue jobs. Data Validation module checks treatment and control data for basic requirements. SageMaker Pipeline then starts with performing follow-up components. Data Processing transforms panel data into cohorts (where needed), handles missing data, extracts lag/lead features, performs optional data scaling and normalization. Causal Estimation decides most suitable causal model given data, and executes the model. Result Validation performs validation tests for sanity and sensitivity, and

returns the estimated treatment effect in a standardized format into the user's S3 bucket.

The data processing can vary for different underlying models. For example, Generalized Synthetic Control (GSC) [26] works well even if there is one treated unit, but it requires panel data with at-least 7 pre-treatment periods. Double Machine Learning (DML) [12] is a better solution for large-scale data but requires breaking down treatments into the cohorts of different weeks, months or quarters, depending on the number of treated individuals in each cohort.

On completion of causal estimation, a series of sensitivity and placebo tests are applied to assess the robustness of the findings to violations of the underlying assumptions. These validations include (but not limited to), direction of causal relationship, sensitivity of causal estimate to small variations in observations data (e.g., down-sampling, random co-variate) and variations in model hyper-parameters (e.g., number of pre-treatment periods used for finding synthetic controls). The results of these validation tests are written to the S3 bucket for user reference.

3.1 Data Requirements

OpportunityFinder requires user to provide two datasets and a configuration file (examples shown in Figure 1). The first data, i.e., treatment data, should contain IDs of the treated units [3] and date when the treatment happened. The second data, also known as, baseline observational or control data, contains the observational information about all IDs that received treatment as well as the ones that did not receive treatment during the same period. Control data should contain time-based, e.g., daily, weekly or monthly, outcome variables (i.e., target) of interest such as ad spend, click count over the historical period. At the same level of time granularity, user is recommended to add a superset of possible variables (i.e., features) that are related to the outcome and the treatment. Among those superset of variables, the model will search for the ones that can help in removing the confounding and mediating effects, an essential for accurate causal estimates.

Configuration file has optional and mandatory requirements. Optional requirements like list of features to scale, choice of algorithm, choice of hyper-parameters allow user flexibility, but are not necessary and can be automatically handled by the framework. The mandatory requirements include columns that specify time, unit id, outcome variable and pre/post-treatment evaluation window, e.g., 4 weeks, 6 months. Based on user-provided configuration and data validation, input panel data might be segmented into cohorts and feature engineering would be performed, before passing to causal analysis algorithms.

3.2 Implementation Details

3.2.1 Two Stage Decision Path. The decision of causal estimation algorithm goes through two stages. First stage is a set of rules, based on factors that include the following. Depending on the answers of these factors, a causal estimation algorithm is selected (leaf node of the decision path)

- Are total event data less than or more than 500,000?
- Is the data panel or cross sectional?

---

[3]Individuals, e.g., shoppers, advertisers, who activated a feature or received a treatment

```
1. ###### API: End-to-end ######

2. from opportunity_finder.api \
3. import OpportunityFinder

4. opf = OpportunityFinder(
5.         treatment_df,
6.         observations_df,
7.         config_dict)

8. opf.estimate_causal_effect()
```

Unit ID	Treatment Date
A	2022-07-23
B	2022-07-16
C	2022-10-01
D	2023-01-11
E	2022-03-05
F	2023-03-21
G	2022-05-05
H	2022-10-12
I	2022-12-02
J	2022-02-25

Unit ID	Date	Impressions	Clicks	Sales
A	2022-07-01	3.43M	400K	$0.25M
A	2022-07-08	3.62M	410K	$0.26M
A	2022-07-15	3.90M	423K	$0.39M
A	2022-07-22	4.21M	431K	$0.32M
A	2022-07-29	3.52M	399K	$0.40M
Z	2022-07-01	8.12M	912K	$10.1M
Z	2022-07-08	8.42M	923K	$10.1M
Z	2022-07-15	8.55M	922K	$10.3M
Z	2022-07-22	8.21M	942K	$8.1M
Z	2022-07-29	8.12M	890K	$11.2M

Figure 1: Left: Sample of OpportunityFinder UX with Python. Center: Sample treatment data. Right: Sample observational data with possible set of covariates.
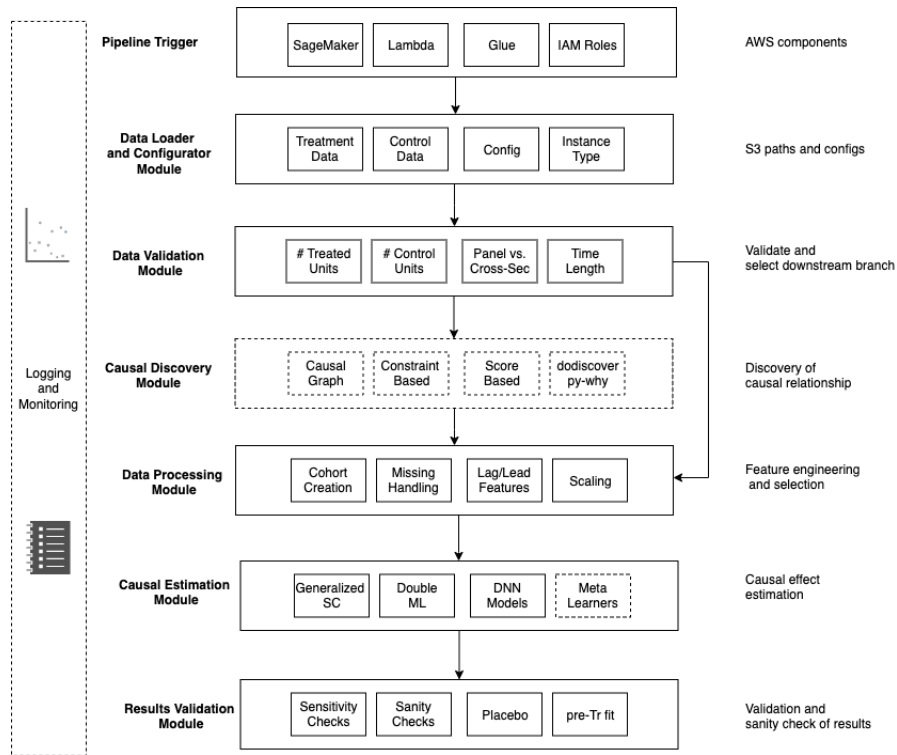

Figure 2: Framework design of OpportunityFinder. *Dotted boxes are under planned development.*

- Are number of treated units per cohort less than or more than 50?
- Are number of control units less than or more than 5,000?
- Are number of covariates as per causal graph more or less than 5?
- How many periods (e.g., daily, monthly) of pre and post-treatment data are available?

In scenarios, where the above set of rules give more than one option of causal estimators, the decision flow moves to the second stage. It tries out the list of all models chosen in the first stage, and selects the result that has least *standard error* on the estimated output and within lower and upper bounds of at-least 2 more estimators (*voting mechanism*).

3.2.2 Cohort Data. One of key functions provided by OpportunityFinder is the transformation of panel data into cohort, i.e., sectional data, which allows techniques like double machine learning to work. Each cohort corresponds to a set of treated units that received treatment in a closed period. First, treatment data is processed to extract list of treatment times and number of treated units at each time. A cohort is a set of one, two or more consecutive treatment times and constrained by three parameters: Minimum/maximum number of treatment times, and minimum number of treated units.

For example, if treatment happens at day level, the first two parameters specify the lower and upper bounds of number of days in each cohort. The third parameter says that a cohort must have at least a certain number of treated units. We sort treatment times in ascending order, for each treatment time we keep merging it with following times until three conditions above are satisfied. Then a list of cohorts is returned. Results from each cohort are aggregated using weighted average w.r.t number of treated units in each cohort.

3.2.3 Causal Inference Models. OpportunityFinder implements a wide set of popular and advanced causal inference algorithms, as following. Further details of these and other models that we considered are in Appendix A.

- Generalized Synthetic Controls: our implementation is based on original R code [26].
- Double Machine Learning: using EconML package [8] we employ two treatment effect estimators, LinearDML and CausalForestDML. Each estimator is stacked with any of four classes of base prediction models that predict treatment and outcome: Random Forest, Linear Regression, XGBoost, and LightGBM.
- Deep Neural Networks: we implement four state of the arts DNN algorithms for estimation of treatment effect, BCAUSS [24], DRAGON [23], TARNET [21], and GANITE [27].

3.2.4 Validation Tests. DML models and their treatment effect estimation are validated through refutation tests by DoWhy package: add random common cause, add unobserved common causes, data subsets validation, and placebo treatment. For a robust causal model and valid treatment effect, first three tests should return treatment effect similar to original model while fourth test must have effect close to zero. GSC model is validated with a suite of sensitivity tests that check for changes in the estimated causal effect with small changes in data like random down-sampling, different pre-treatment window for learning synthetic control weights and a reduced covariate list. The expectation is that the causal effect should not change the direction of estimation with small changes in the setting. Example test results are shown in Appendix B. In future, we will equip validation tests for DNN models.

3.2.5 Data Visualization. A challenge that prevents the adoption of causal inference studies is a lack of ground-truth data which makes estimation error impossible to assess. OpportunityFinder addresses this by providing visualizations that naively explain the treatment effect to some extent. For example, it returns a plot that shows the trend of outcome for treated and control units over time. The visualizations are a part of Logging and Monitoring module. While such plots do not confirm calculated treatment effect by causal models, they help non-expert users to comprehend causal inference results. Example visualizations are shown in Appendix C.

3.3 Limitations and Risks

As of today, OpportunityFinder (OPF) does not implement causal graph generation algorithms. This also means that the tool has less flexibility for someone who wants to control covariates and experiment with different algorithms. We plan to integrate causal discovery module in near future.

OPF applies our best heuristics after exploring input data to select the right algorithm. Due to the lack of ground truth data in causal inference, our framework can make mistakes without knowing that the estimated effect it returns is wrong. We select a model based on *standard error* and ensemble by voting to mitigate this limitation upto some extent. The accuracy of estimate still depends on the observational data given by the user.

For real-world problems, OPF does not necessarily use estimation models that gave best score on benchmark data. Our experiments show that simpler estimators work more reasonably than DNN models on our use-case data.

4 VALIDATION OF CAUSAL ESTIMATES BY OPPORTUNITYFINDER

We validate our causal inference algorithms on benchmark datasets using three metrics, choosing the metric available in related research for each dataset.

(1) Average Treatment Effects (ATE): measures causal impact of a treatment/intervention on a population by comparing the average hypothetical outcomes between receiving and not receiving treatment, accounting for potential confounding factors.
(2) Average Treatment Effects on Treated (ATT): is ATE measured on treated units.
(3) Mean Absolute Error (MAE): the average absolute difference between estimated ATE and true ATE where available, for evaluating accuracy of a causal estimation method.

4.1 IHDP *(public benchmark)*

The Infant Health and Development Program (IHDP) [1] is a randomized controlled study designed to evaluate the effect of specialist visit on cognitive test scores of premature infants. This dataset is cross-sectional data, with binary treatment (specialist visits), continuous outcome (cognitive scores) and has known ground truth ATE. As shown in Table 1, our implementation of DML models achieved competitive performance. Results on BCAUSS, TARNET and DRAGON are based on our implementation and slightly differ from reported numbers [1]. The difference is because DNN methods are executed within OPF pipeline and data are not prepared as the same as previous studies.

4.2 Smoking *(public benchmark)*

The goal of smoking data is to analyze the causal effect of Proposition 99 on cigarette sales. This data has small size with just one treated unit, thus causal estimations based on machine learning (DML, NN) do not apply. OpportunityFinder chooses to run GSC and does not create any cohorts. Table 2 shows comparison of results using OPF on this dataset with previous research [3, 6] works. We observe that the range of ATE estimate lies between -11.1 to -27.1, and the results from OPF are within the range of previous studies.[4] When we use cigarette retail price as a covariate, the ATE reduces to -14.0, which is closer to SDID. It is discussed in SDID [6] paper that their results (-15.6) are more credible among other

---

[4]Synthetic difference in differences (SDID), Synthetic controls (SC), Difference in differences (DID), Matrix completion (MC), Synthetic control with intercept (DIFP).

Table 1: Mean absolute error on IHDP benchmark. All models are part of OpportunityFinder

	DNN			+LinearDML		
BCAUSS	DRAGON	TARNET	LinearReg.	Rand.Forest	XGBoost	LightGBM
0.23	0.32	0.25	0.42	0.48	0.47	0.43

approaches shown in the table. We also observe a lower standard error with OPF results. This experiment helps validate OPF on small panel dataset.

4.3 Synthetic Data 1: Cross Sectional *(synthetic)*

In addition to public datasetes, we validate OpportunityFinder outputs on two synthetic datasets with known ATE.[5] The first synthetic dataset is a linear cross sectional dataset that we generated using DoWhy [22] package. We created the dataset with 2 instrument, 5 common causes, 5000 samples and binary treatment with some treatment noise. Because this data is cross sectional, OPF rejects GSC, but branches off to the second stage of decision path, where it evaluates multiple models, including DML [12], and neural net based estimators including BCAUSS [24], TARNET [21], DRAGON [23] and GANITE [27]. As we see in Table 3, all models (except GANITE) give ATE close to the true ATE. OPF finally selects the model with least standard error if the mean is within the range of 2+ other models, and ends up selecting LinearReg+LinearDML.

4.4 Synthetic Data 2: Large Panel *(synthetic)*

In the second dataset, we add non-linear confounding effect and correlated variables on a panel data, to test the efficacy of different supported models to be able to remove the bias. This data contains 52,000 rows, 3 confounders with non linear effect on treatment and outcome, 1000 units, 263 treated units and 52 time periods. The properties of this dataset enables OpportunityFinder to run all implemented algorithms and select based on standard errors. As shown in Table 3, all models except GANITE, perform well that estimated ATE are close to ground-truth. Due to lowest variance in estimations from GSC together with mean estimation lying between 2+ other estimators, OPF chooses GSC results for the end user.

4.5 Discussion on Model Choice

While our approach for model selection is evolving, the results on synthetic and public datasets show that our current 2 stage decision path works well. The two stage decision path allows automated rejection of estimators if they are not built for the use case at hand. For example, GSC is supposed to be used for panel data but becomes computationally inefficient with >500,000 data sizes. OpportunityFinder does not run GSC for such large data sizes. As the research evolves, especially with neural networks for causal inference, we plan to incorporate the new models as well as update the models selection criteria. For example, we will explore providing results from ensemble of models and finding the expected causal path using causal discovery algorithms.

5 APPLICATIONS ON REAL WORLD DATA

OpportunityFinder has already been used in multiple use cases. In this section, we present two most important applications of OPF within our organization. Most commonly used down stream impact metric in real applications is *uplift*, which is defined as the percentage increase/decrease in the outcome attributed to the treatment over a defined period. It is calculated as ATE or ATT divided by average over control units.

5.1 Opportunity for Partners

Advertising partners are the agencies and tool providers that have expertise in interacting with Ads products and help sellers/vendors in setting ad campaigns. Our team helps in identifying the actions that would enable partners to create maximum value for sellers/vendors. Such actions are considered opportunity for partners. Their impacts are measured on a wide list of business outcome metrics such as revenue and adspend. Traditionally, it used to take 1-3 weeks of an Economist time to update the studies on ad-hoc requests. Since January 2023, we have been using OpportunityFinder to refresh the studies. Each refresh completes in a day with minimal human involvement.

OPF chooses GSC due to: number of total events < 500,000, the number of treated units per monthly cohort < 50 and control units < 5,000. In Table 4, we show two such opportunities: adoption of X, and adoption of Y[6] by the partner for at-least one of their customers. Their lifts on three business metrics. Comparing to results of prior studies, we can see delta between past and current downstream impact, which is caused due to behavioral and market changes over time. We further compare GSC results with DML and DNN models. While DML models return lift scores about 5% to 10% higher than GSC, DNN estimated lifts are from 20% to hundred percent higher which are beyond acceptable range. ML-based models may over-estimate when input data is small.

5.2 Opportunity for Advertisers

This study estimates the effect of advertising partners on sellers/vendors outcomes related to Ads business, e.g., ad spend. This study has been traditionally taking multi-weeks of scientist's effort for each refresh. In 2023, the study was expanded in both number of outcome variables and numbers of groups of partners and advertisers. Each combination of outcome and entity group is a separate causal study. OpportunityFinder helped accelerate the study so that all experiments were completed within a month.

With a large number of advertisers, input data is redirected to DML and DNN causal models. Input panel data is then transformed into cohorts, before feature engineering and model training. Based

---

[5]github.com/groverpr/datasets

[6]Opportunity names and business metrics (see Tables 4, 5) are masked due to customer data policy.

Table 2: Estimates for ATE on Smoking data

	SDID	SC	DID	MC	DIFP	OPF	OPF w.price
ATE	-15.6	-19.6	-27.3	-20.2	-11.1	-24.6	**-14.0**
Standard Error	8.4	9.9	17.7	11.5	9.5	4.9	**4.7**

Table 3: ATE (with std. error) on synthetic datasets using models implemented in OPF

	Synthetic#1 (GT = 10)		Synthetic#2 (GT = 20)	
BCAUSS	10.18	(0.03)	19.00	(0.18)
DRAGON	10.34	(0.19)	18.98	(0.24)
TARNET	10.00	(0.09)	18.98	(0.23)
GANITE	7.78	(n/a)	6.41	(n/a)
LinearReg. + LinearDML	**9.93**	(0.02)	19.04	(0.13)
Rand.Forest + LinearDML	9.98	(0.07)	19.01	(0.13)
XGBoost + LinearDML	9.70	(0.04)	18.82	(0.13)
LightGBM + LinearDML	9.75	(0.03)	18.97	(0.15)
GSC	n/a		**18.87**	(0.07)

Table 4: A sample of opportunity for partners studies from 2021/22 vs 2023. Metric is uplift after 6 months of adoption.

	Opportunity X			Opportunity Y		
	Metric 1	Metric 2	Metric 3	Metric 1	Metric 2	Metric 3
Manual (2021/22)	5%	20%	12%	4%	4%	6%
OpportunityFinder (2023)	6%	12%	17%	8%	8%	11%

Table 5: Results of opportunity for advertisers study for world-wide vendors. Metric is average monthly uplift on outcomes within 3 months after adoption.

	DNN				+LinearDML			
Outcome	BCAUSS	DRAGON	TARNET	GANITE	LinearReg.	Rand.Forest	XGBoost	LightGBM
Metric 4	68%	45%	62%	15%	17%	14%	12%	14%
Metric 5	58%	48%	64%	16%	14%	13%	16%	12%

on ATE and standard error results on validation datasets, OPF chooses `Rand.Forest+LinearDML` as final model. Our results were reviewed by domain experts and in range of results from prior studies. In Table 5, we report lift metric returned by all possible models on a dataset. Three DNN models over-estimate treatment effect, and only GANITE yields numbers close to DML models.

6 CONCLUSIONS AND FUTURE WORK

This paper presents OpportunityFinder (OPF), a codeless framework for causal inference studies, with a focus on panel data with binary treatment. Our experiments on multiple public, synthetic and internal datasets show that OPF can handle a diverse set of scenarios and our decision criteria for algorithm selection works well for given use-cases. We also see that in most of the cases, simpler algorithms like DML and GSC work well. We are able to use OPF on datasets ranging from small panel data to a large data with more than one million observations.

We are actively taking feature requests from current OPF users. With causal discovery component, we will explore how hypothesis formulation before estimation can improve the estimation capability, especially with large set of observational data that a non expert user tends to provide. We also aim to provide a master list of variables that can be collected for causal inference studies within our organization, and let OPF auto-shortlist covariates using data driven approaches for removing bias. We plan to extend OPF by incorporating more estimators like meta learners, implement individual and heterogenous treatment effects, and support categorical and continuous treatments. With more and more causal inference algorithms being integrated into OPF, we will implement additional model selection, e.g., prediction/regression accuracy of base learners. Moreover, we will experiment model ensembles to provide the final output. Last but not least, we have refactored OpportunityFinder source code to make it a stand-alone library independent of AWS ecosystem.

REFERENCES

[1] [n. d.]. Causal Inference on IHDP: Benchmark. https://paperswithcode.com/sota/causal-inference-on-ihdp.
[2] [n. d.]. No Free Lunch in Causal Inference. https://p-hunermund.com/2018/06/09/no-free-lunch-in-causal-inference/.
[3] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. 2010. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *J. Amer. Statist. Assoc.* 105, 490 (2010), 493–505. https://doi.org/10.1198/jasa.2009.ap08746 arXiv:https://doi.org/10.1198/jasa.2009.ap08746
[4] Alberto Abadie and Javier Gardeazabal. 2003. The economic costs of conflict: A case study of the Basque Country. *American economic review* (2003), 113–132.
[5] Joshua D. Angrist and Jorn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.
[6] Dmitry Arkhangelsky, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. 2021. Synthetic Difference in Differences. arXiv:1812.09970 [stat.ME]
[7] Philipp Bach, Victor Chernozhukov, Malte S. Kurz, and Martin Spindler. 2022. DoubleML – An Object-Oriented Implementation of Double Machine Learning in Python. *Journal of Machine Learning Research* 23, 53 (2022), 1–6. http://jmlr.org/papers/v23/21-0862.html
[8] Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. 2019. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/py-why/EconML. Version 0.x.
[9] David Card and Alan B Krueger. 1994. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *The American economic review* 84, 4 (1994), 772–793.
[10] Huigang Chen, Totte Harinen, Jeong-Yoon Lee, Mike Yung, and Zhenyu Zhao. 2020. CausalML: Python Package for Causal Machine Learning. arXiv:2002.11631 [cs.CY]
[11] Cheng and Mark Hoekstra. 2012. *Does Strengthening Self-Defense Law Deter Crime or Escalate Violence? Evidence from Castle Doctrine.* Working Paper 18134. National Bureau of Economic Research. https://doi.org/10.3386/w18134
[12] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21, 1 (2018), C1–C68.
[13] Rajeev H. Dehejia and Sadek Wahba. 2002. Propensity Score-Matching Methods for Nonexperimental Causal Studies. *The Review of Economics and Statistics* 84, 1 (02 2002), 151–161. https://doi.org/10.1162/003465302317331982 arXiv:https://direct.mit.edu/rest/article-pdf/84/1/151/1613304/003465302317331982.pdf
[14] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 2020. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. *arXiv preprint arXiv:2003.06505* (2020).
[15] Matthias Feurer, Katharina Eggensperger, Stefan Falkner, Marius Lindauer, and Frank Hutter. 2020. Auto-Sklearn 2.0: Hands-free AutoML via Meta-Learning. *arXiv:2007.04074 [cs.LG]* (2020).
[16] Timo Flesch, Edward Zhang, Guy Durant, Mark Harley Wen Hao Kho, and Egor Kraev. 2022. Auto-Causality: A Python package for Automated Causal Inference model estimation and selection. https://github.com/transferwise/auto-causality. Version 0.x.
[17] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning. *Proceedings of the National Academy of Sciences* 116, 10 (2019), 4156–4165.
[18] Robert Lalonde. 1986. Evaluating the Econometric Evaluations of Training Programs with Experiment Data. *American Economic Review* 76 (02 1986), 604–20.
[19] Erin LeDell and Sebastien Poirier. 2020. H2O AutoML: Scalable Automatic Machine Learning. *7th ICML Workshop on Automated Machine Learning (AutoML)* (July 2020). https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf
[20] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.
[21] Uri Shalit, Fredrik D. Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70.* JMLR. org, 3076–3085.
[22] Amit Sharma, Emre Kiciman, et al. 2019. DoWhy: A Python package for causal inference. https://github.com/microsoft/dowhy.
[23] Claudia Shi, David M. Blei, and Victor Veitch. 2019. Adapting Neural Networks for the Estimation of Treatment Effects. arXiv:1906.02120 [stat.ML]
[24] Gino Tesei, Stefanos Giampanis, Jingpu Shi, and Beau Norgeot. 2023. Learning end-to-end patient representations through self-supervised covariate balancing for causal treatment effect estimation. *Journal of Biomedical Informatics* 140 (2023), 104339.
[25] Stefan Wager and Susan Athey. 2018. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242.
[26] Yiqing Xu. 2017. Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models. *Political Analysis* 25, 1 (2017), 57–76. https://doi.org/10.1017/pan.2016.2
[27] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2018. GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets. In *International Conference on Learning Representations.* https://openreview.net/forum?id=ByKWUeWA-

APPENDIX

A OVERVIEW OF CAUSAL INFERENCE MODELS

The following models are considered for the auto causal framework:

- **Synthetic Control (SC)** and **Generalized Synthetic Control (GSC)**: SC allows for comparative case studies using a weighted combination of control units to create a synthetic control unit. GSC extends this by considering interactive fixed effects models. The key assumption is that the outcome of treated units is a linear function of the outcomes of the control units in the absence of treatment. GSC allows relationship to vary over time, unlike traditional SC methods. Both methods are well suited for panel data with small sample sizes but require domain knowledge for selection of control units. The limitation for implementing GSC in the auto causal framework is computational inefficiency with large observational data or with more number of covariates. [4, 26]

- **Double Machine Learning (DML)**: DML leverages machine learning to estimate treatment effects in a semi-parametric manner, allowing for complex relationships. The key requirement for DML to work well is the availability of high-quality and diverse covariate data. DML can handle large datasets and does not specifically require panel data. It allows for different ML models to be used in the two stages, providing versatility. [12]

- **Causal Forests**: Causal Forests extend random forests to estimate heterogeneous treatment effects, offering flexibility and the ability to capture complex relationships. The key assumption is the unconfoundedness or ignorability assumption. It is not inherently designed for panel data and requires a relatively large sample size. The limitation for auto causal framework is that it does not handle panel data well and we did not find it to work well in our experiments. [25]

- **Neural Network based approaches**: Several approaches utilize neural networks for causal inference, each with its unique proposition: BCAUSS, Dragonnet and TARNet model treatment assignments and potential outcomes in a multi-task learning setup, allowing finding least dissimilar treated and untreated observations. GANITE leverages the power of generative adversarial networks (GANs) to estimate individual treatment effects. They require relatively large and high-quality datasets, otherwise can over/under estimate the treatment effects. These methods can handle large datasets but are not specifically designed for panel data. [21, 23, 24, 27]

- **Meta Learners**: Meta Learners apply machine learning methods to estimate treatment effects, offering the flexibility

of using various base learners. The key assumption is that the base learners are correctly specified. They are not specifically designed for panel data and require a relatively large sample size. The limitation for auto causal framework is in the choice of base learner. [17]

- **Difference in Differences (DiD)**: DiD compares the average change in outcome over time that occurs in the treatment group to the average change over time that happens in the control group. It's designed to handle unobserved, time-invariant confounders. It's a simple and intuitive method for panel data, widely used in economic studies. Problem with DiD is that is relies on strong parallel trends assumption that is often violated in the real world setting. We do not use DiD or variant Synthetic DiD in our implementations. [6, 9]

- **Propensity Score Matching (PSM)**: The propensity score is the conditional probability of receiving treatment given pre-treatment characteristics. This approach has been traditionally popular because of its simplicity, interpretability and ability to handle large covariates. But DML is based on similar principle and overcomes the limitations that PSM has and is more robust. Complementing DML, SC methods do not require unconfoundedness assumption that PSM does. Therefore, we do not use PSM in the OPF. [20]

B VALIDATION TESTS FOR TREATMENT EFFECT RESULTS

To demonstrate the validation test for treatment effect results, we report refutation test outputs for LinearReg+LinearDML model and its ATE on Synthetic#1 data in Table 6. Sensitivity test for GSC model are reported in Table 7. All refutation tests passed, placebo ATE are close to zero while other test ATE close to original model. This confirms the model is robust against changes in settings and estimated ATE is consistent.

C VISUALIZATION OF OUTCOME VARIABLES IN DIFFERENT DATA

We display data plots generated by OpportunityFinder when running different datasets. Figure 3 plots average outcome of treated (orange line) and control (blue line) units for a cohort. Dash-red vertical bars indicate start and end date of cohort. These plots are generated by our data processing module. For Smoking and Texas data, Figures 4, 5 are generated by our GSC model that black line shows time-series of outcome values of treated unit, and dash-blue line show that of synthesized control.

D ADDITIONAL RESULTS

D.1 *Texas* data

Table 8 shows impact on black and white male incarceration from prison expansion in Texas since 1993. The numbers represent average percentage lift on the respective observational metric in Texas vs. other states due to the expansion. The covariates used to remove bias include poverty rates, white male incarceration, percentage of population between 15 and 19, income, unemployment rate and AIDS mortality. We see that the Texas had an average of 2x (100%) more black male incarceration compared to the other states after
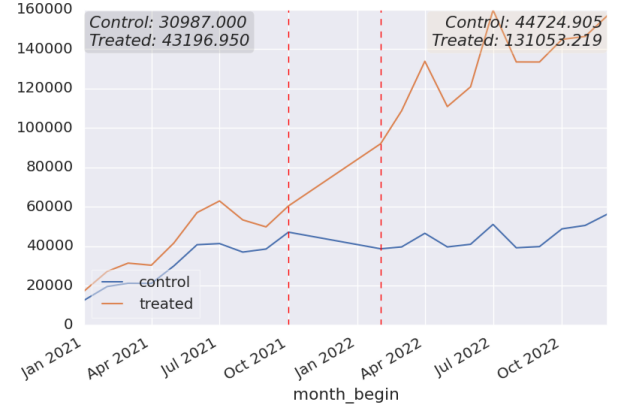
(a) Metric 4
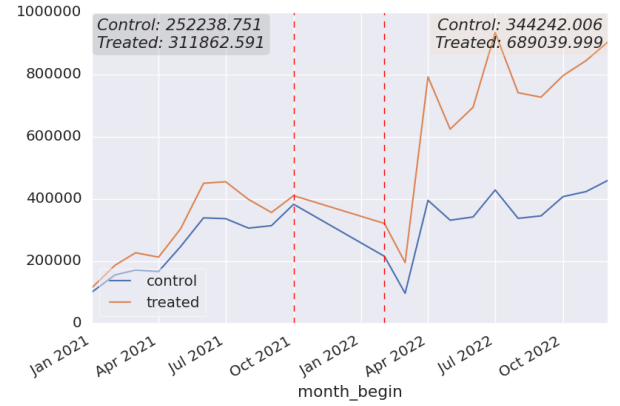
(b) Metric 5

Figure 3: Average outcome metric of treated vs. control units over time in advertiser data.

they started prison expansion in 1993 till 2000. The increase was low but non zero (38%) for white male incarceration during the same period in Texas.

D.2 *NSW* and *Castle* data

In this section, we show results on two additional datasets, NSW [18] and Castle [11]. Caslte is a panel data with year level information for 10 years, covering 50 states out of which 21 adopted the castle doctrine law. Castle law designates a person's abode or any legally occupied place (for example, a vehicle or home) as a place in which that person has protections and immunity permitting, in certain circumstances, to use force (up to and including deadly force) to defend oneself against an intruder, free from legal prosecution for the consequences of the force used. The study done by [11] aimed at finding the effect of castle doctrine law on increase in homicide. This data has 550 rows and 170 features (potential covariates), and based on researcher's outcome, expected lift in homicide of 8% (we are accessing if OPF models reproduce the study). We see that in such small data sizes with large number of potential covariates, linear models do the best, and boosted trees can be very off.

Table 6: Refutation tests for DML model on Synthetic#1 and #2 sets

	Synth#1 ATE		Synth#2 ATE	
Ground-truth	10.00		20.00	
Model ATE	9.93		19.04	
Placebo test	-0.08	passed	-0.05	passed
Random common cause test	9.98	passed	19.08	passed
Unobserved common cause test	9.98	passed	15.48	passed
Data-subset test	9.99	passed	19.00	passed

Table 7: Sensitivity tests for GSC model on partner data with Metric 2. The numbers represent % uplift. All lifts are statistically significant

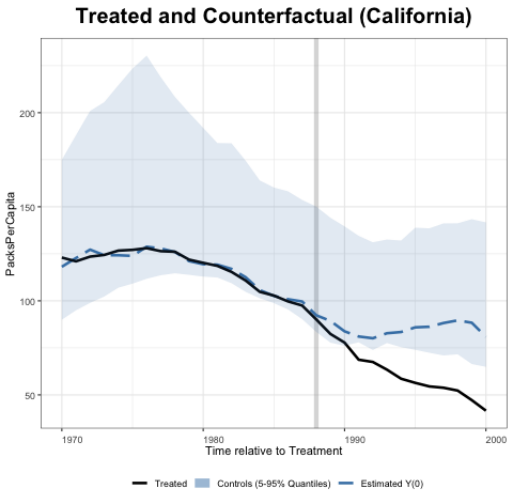|	Opportunity X		Opportunity Y	
Overall model	7.8%		12.0%	
Remove covariates test	14.3%	passed	34.5%	passed
Random downsample test	7.0%	passed	11.8%	passed
Reduced period for SC weights test	7.8%	passed	11.2%	passed

Figure 4: Synthetic control fit on smoking data without covariate. The pre-treatment fit is good.

Table 8: Average percentage lift on black and white male incarceration from prison expansion in Texas since 1993.

black-male prison	white-male prison
100%	38%

NSW is a famous experimental data that is complemented with additional synthetic data where researchers added selection bias in the control population. This is used in multiple research works to replicate the results of randomized trials. It contains 50,000 rows and 180 features, and has 100 simulated variations. It is not a panel data and we tested it with DML variations. We observe underestimation compared to other research works but closest results from Random Forest based model.

(a) Black male incarceration

(b) White male incarceration
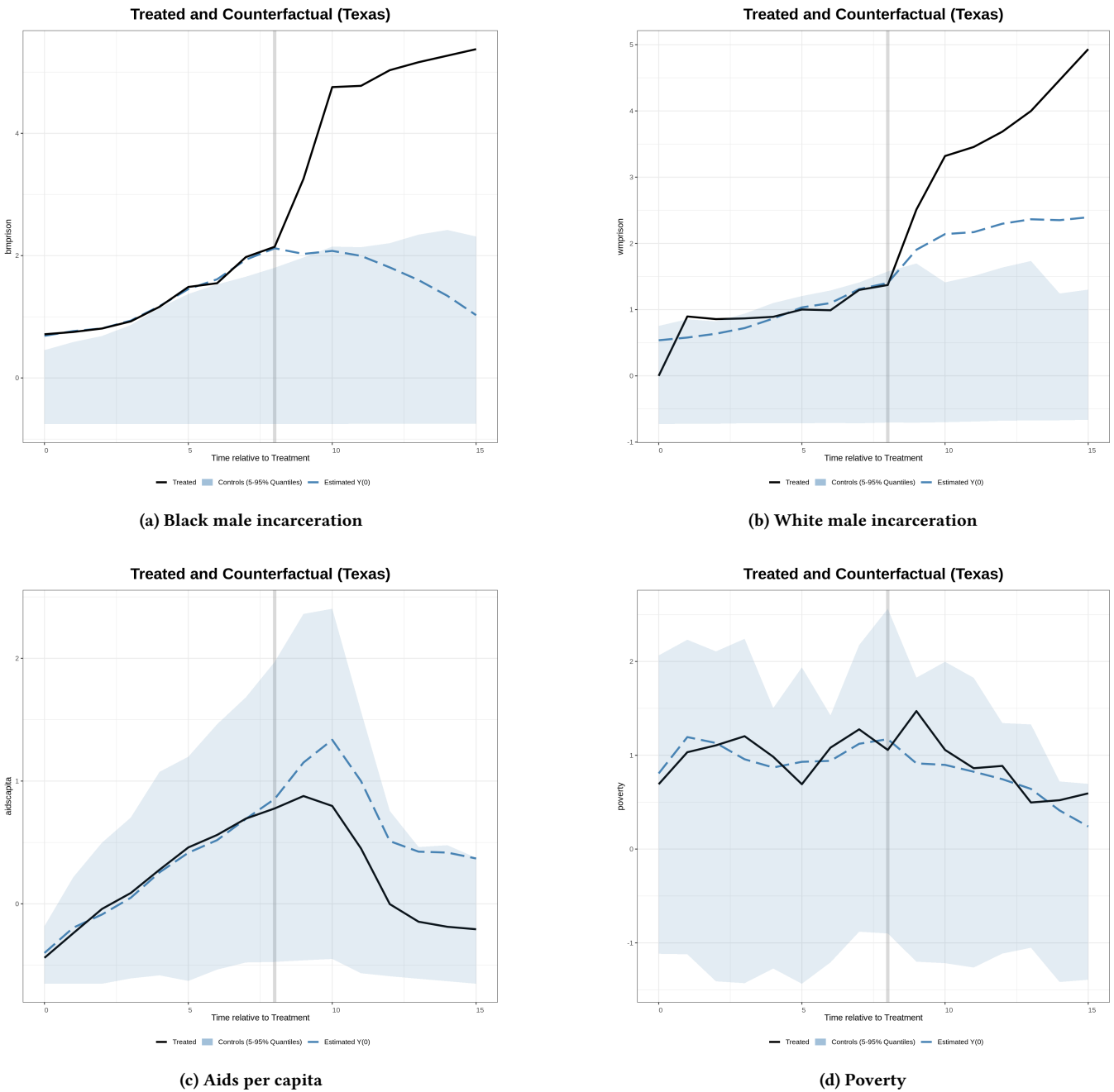

(c) Aids per capita

(d) Poverty

Figure 5: Pre treatment synthetic control fit and post treatment diversion of different metrics on Texas data.

Table 9: Comparison of % lift on Castle datasets using DML models from OPF vs. previous research works

Previous Research	+LinearDML			
Cheng'12 [11]	LinearReg.	Rand.Forest	XGBoost	LightGBM
8%	**7%**	4%	4%	50%

Table 10: Comparison of ATE on NSW datasets using DML models from OPF vs. previous research works

Previous Research		+LinearDML			
Lalonde'86 [18]	Dehejia'02 [13]	LinearReg.	Rand.Forest	XGBoost	LightGBM
900	1300-1800	286	**776**	728	637