Causal Machine Learning in IoT-based Engineering Problems: A Tool Comparison in the Case of Household Energy Consumption

N.-L. Kosioris[1,2] S. Nikoletseas[1] G. Filios[1] S. Panagiotou[1]

[1] Department of Computer Engineering and Informatics, University of Patras

[2] Connect Line S.A., Athens, Greece

Abstract: The rapid increase in computing power and the ability to store Big Data in the infrastructure has enabled predictions in a large variety of domains by Machine Learning. However, in many cases, existing Machine Learning tools are considered insufficient or incorrect since they exploit only probabilistic dependencies rather than inference logic. Causal Machine Learning methods seem to close this gap. In this paper, two prevalent tools based on Causal Machine Learning methods are compared, as well as their mathematical underpinning background. The operation of the tools is demonstrated by examining their response to 18 queries, based on the IDEAL Household Energy Dataset, published by the University of Edinburgh. First, it was important to evaluate the causal relations assumption that allowed the use of this approach; this was based on the preexisting scientific knowledge of the domain and was implemented by use of the in-built validation tools. Results were encouraging and may easily be extended to other domains.

Keywords: Causal Machine Learning, Validation, Household Energy Consumption, PyWhy™, CausalML™, Tool Comparison

Abbreviations

ML	Machine Learning
CML	Causal Machine Learning
IML	Interpretable Machine Learning
EML	Explainable Machine Learning
CSM	Structural Causal Model
ATE	Average Treatment Effect
ITE	Individual Treatment Effect
MSE	Mean Squared Error
MLP	Multi-Layer Perceptron
KLD	Kullback-Leibler Divergence
AUUC	Area Under Uplift Curve

1 Introduction

The majority of ML algorithms used today have the ability to detect patterns of any scale in a wide range of datasets from small to huge. However, some of the resulting data-driven models suffer from biases, difficulty in generalizing for future uses and overfitting, as well as a lack of distinction between probabilistic association and true causation (Mehrabi et al., 2021). Human control in this context is often impossible to achieve due to complexity, resulting in erroneous conclusions. From the users' point of view such models are no longer sufficient, as results need to be explained in a cause-and-effect format, which is closer to the human way of thinking (Schölkopf, 2022). In this way, the effect of changes in the underlying assumptions of the model is easier to understand, saving time from re-creating a dataset. Incorporating causality logic into existing AI systems should handle the above problems making models more efficient and intelligent, reducing the associated energy- and time-

consuming learning cycles for processes that are trivial for humans. Thus, CML tools are expected to play a leading role in the coming years.

CML is a major milestone in ML that enables AI models to make accurate predictions based on causes rather than just probabilistic correlations, thereby allowing the understanding of their effect. Note that causal hypotheses are not already included in the statistical description of the system. Some common pitfalls that Machine Learning falls into and that CML focuses on solving are listed below:

- *Correlation Equation with Causation.* Predictive algorithms can fall into the trap of equating correlation with causation, i.e. believing that because event X precedes event Y, it must be the cause of it.
- *Inability to use non-independent and identically distributed random variables.* ML mostly implement pattern recognition in large data sets for which random observations are independent of each other and occurrence probability is continuous. However, in real-world scenarios, data distribution can change, e.g. in object recognition from images under unusual/novel lighting, angles or backgrounds (Cheng et al., 2022).
- *Absence of Immutability.* Current ML algorithms are limited in their ability to generalize the patterns they find in a training dataset for multiple applications (G. Xu et al., 2020). Ideally, the ML model should exhibit immutability, i.e. generalize beyond the limited set of environments that can be accessed for training, and even extend to new and perhaps unpredictable environments.
- *Lack of Explainability.* ML models remain mostly "black boxes" that cannot explain the reasons behind their predictions or recommendations (Kaddour et al., 2022). Typically, Deep Learning systems can be trained to recognize manufacturing process defect images with high accuracy, but - unlike a real expert - it cannot explain why or how a specific image indicates a defect.

Interestingly, CML, Interpretable Machine Learning (IML) and Explainable Machine Learning (EML) are in a sense related, even complementary. CML is aiming to understand what will happen as a result of an action, i.e. it focuses on cause-and-effect, whereas IML aims to understand how the model functions, focusing on transparency, explanation, and trust. The relevant tools will, as a result, also differ. Furthermore, EML differs from IML in that it provides methods that help explain complex, black-box models, which are not inherently interpretable, so post-hoc tools are needed to explain their behaviour. By contrast, IML refers to models that are inherently understandable by humans without extra tools (Linardatos et al., 2021), (G. Xu et al., 2020)

The aim of this work is to highlight and compare state-of-the-art tools in the field of CML by a sample application in an IoT-data environment, highlighting ways to exploit the resulting model. Furthermore, the extended aim of the work is to guide future adaptation of the tools to other domains and use cases.

The paper is structured in the following way: In Section 2, published work pertaining to this paper is briefly reviewed. In Section 3, the notional foundation of the architectures that have prevailed so far in CML are presented. Then, in Section 4, the two most popular and powerful tools in the field are presented, namely Microsoft PyWhy™ and UberML™, as well as the rest of the tools needed in implementing CML models. Further, in Section 5, the selected data set is presented and its required pre-processing is clarified. In Section 6, the models that resulted from the two tools are represented graphically, followed by an explanation of the assumptions regarding the data set and the methods of testing and evaluating the results. Section 7 summarizes conclusions and proposes future extension directions of this work.

2 Related Work

The adoption of CML for engineering problems aims to provide a framework moving beyond correlation-based insights toward cause-effect modelling. In particular, in cases related to data stemming from IoT, CML methods provide a foundation for interpreting sensor data, designing interventions, and optimizing the function of the pertinent engineering system. (Naser, 2021).

The broader application topic of this paper concerning energy management is probably the one attracting most registered causal AI approaches. For instance, (Massidda & Marrocu, 2023) propose a novel method for probabilistic forecasting of the

total load of a residential community and its base and thermal components, combining conformalized quantile regression and causal machine learning, using only aggregate consumption and environmental conditions in a dataset of a residential community in Germany. The T-learner method was the most effective among the causal methods for load disaggregation in terms of accuracy, simplicity, and potential for extension.

(C. R. Kinttel & Stolper, 2019) use causal forests to evaluate the heterogeneous treatment effects (TEs) of repeated behavioral nudges towards household energy conservation. Pre-treatment consumption and home value are the strongest predictors of treatment effect. A "boomerang effect" is noticed: households with lower consumption than similar neighbours are the ones with positive TE estimates.

(Miraki et al., 2024) propose eXplainable Causal Graph Neural Network (X-CGNN) for multivariate electricity demand forecasting yielding intrinsic and global explanations based on causal inferences as well as local explanations based on post-hoc analyses. State-of-the-art performance was attained with two real-world electricity demand datasets from both the household and distribution levels.

In (Duhirwe et al., 2024) a causal inferential approach is employed combining double machine learning and domain knowledge using directed acyclic graphs. The 2015 US Residential energy consumption survey is used. The results show that energy audits, proper insulation, access to interval data etc. significantly reduce energy use intensity (EUI). For instance, proper insulation reduces EUI by 5.6 MWh/m2 while changing static thermostat settings to automatic adjustments at certain times reduces EUI by 3.5 MWh/m2.

(Kim & Park, 2025) implemented a causal modeling approach using Bayesian Networks (BNs), focusing on inter-causalities among various building occupant behaviors. Data were collected from a yearlong study involving six households in Seoul, South Korea. In the customized BN model, the intercausal relationships for each household were quantitatively measured using machine learning, revealing distinct personal preferences and interactions with the environment and high predictive accuracy.

Data-driven building thermal dynamics models circumvent the need for domain experts and determination of physical properties of the buildings, but often fail to generalize to truly unseen conditions since they are not causal. (Jiang & Kazmi, 2025) demonstrate that causal machine learning (CML) algorithms trained on debiased data from nine real-world Dutch buildings can produce accurate models necessary for control-oriented applications which outperform baseline models by over 40%, besides learning the correct causal associations which were verified using a custom testing environment and SHAP feature analysis.

Taking an example from energy-efficient building design (Chen et al., 2022) introduce causal inference proposing a four-step process to discover and analyze parametric dependencies in a mathematically rigorous and computationally efficient manner by identifying the causal diagram, which provides a nexus for integrating domain knowledge with data-driven methods, towards interpretability and testability against the domain experience within the design space.

(Maisonnave et al., 2024) present an alternative approach to predictive modelling for future energy demands based on an ensemble causal model generated by four prominent causal detection methods using a dataset of the wholesale electricity company of Argentina (CAMMESA).

(Shams Amiri et al., 2023) demonstrate an Extreme Gradient Boosting (XGBoost) algorithm for forecasting the energy use of commercial and residential buildings in Philadelphia, USA. A corresponding SHAP (SHapley Additive exPlanations) analysis is implemented to pinpoint feature contributions to the model's energy estimates. By using the PopGen software, the model's energy estimates could be analyzed at the household level, the smallest possible scale. The results indicate that features related to lower building intensity (e.g., lower square footage, fewer floors per building) were associated with reduced energy use, whereas ''single-family attached" zoning designation corresponds with higher energy estimates.

Energy retrofits of existing buildings provide an effective means to reduce building consumption and carbon footprints currently accounting for 40% of the energy consumption and 31% of the CO2 emissions in US. (Y. Xu et al., 2021) uses data from a portfolio of 550 federal buildings exploiting the causal forest estimator to predict the retrofit effect as a function of building characteristics, climate, energy-use level, etc. A causal forest is an ensemble of many causal trees. A causal tree was

fitted for each type of retrofit action and fuel type (electricity and gas) and was built with a random subsample of the whole data set. Within the findings, a dashboard tool and fault detection system, commissioning, and HVAC investments had the highest average savings among the six actions analyzed.

Demand response (DR) is pivotal in enhancing the operational efficiency and reliability of power systems. (He & Khorsand, 2024) integrate causal learning approaches for causal intervention and counterfactual analyses of prosumers' DR participation and consumption behavior. The proposed framework incorporates domain knowledge of the power system to enhance model accuracy and performance, demonstrated by simulation.

(Sun et al., 2024) reviews the application of causal relationships in power systems fault diagnosis, energy prediction, stability analysis, and the electricity market, looking forward to the development prospects of causal relationships in combination with power physics models and artificial intelligence, towards reliable and intelligent new power systems focusing on renewable sources.

Widening the engineering perspective, diverse problems have been tackled with causal AI mostly in the domains of robotics/automation and manufacturing.

In robotics, (Castri et al., 2023) propose an extension of a state-of-the-art causal discovery method embedding an additional feature-selection module based on transfer entropy. Starting from a prefixed set of variables, the new algorithm reconstructs the causal model of the observed system by considering only its main features and neglecting those deemed unnecessary for understanding the evolution of the system. The method is validated on a real-world robotics scenario using a large-scale time-series dataset of human trajectories. (Hellström, 2021) investigate the role of causal reasoning in robotics research. Inspired by a categorization of human causal cognition, a categorization of robot causal cognition is proposed mainly covering the sense–plan–act level of robotics, but also understandability, machine ethics, and robotics research methodology. (Nadim et al., 2023) propose an intelligent supervisory control approach which adopts deep reinforcement learning (DRL) to develop an efficient control policy through interaction with a process simulation. The DRL training history is then exploited using interpretable machine learning and process mining to build a discrete event system (DES) model, in the form of a state-event graph. The DES model identifies causal relationships between events and provides interpretability to the control policy developed by the DRL method.

In the manufacturing domain, (Vuković & Thalmann, 2022) provide a comprehensive overview about how causal discovery can be applied identifying four core areas, namely fault detection, analysis and management, root cause analysis, causality in a facilitator role, and domain and conceptual work. In (Ko et al., 2023), a novel physics-guided data-driven framework for Additive Manufacturing includes a Process-Structure-Property (PSP) causal relationships learning process with two sub-processes, namely a knowledge-graph-guided top-down approach to generate the requirements for predictive analytics and data acquisition and a data-driven bottom-up approach to construct and model new PSP knowledge. (Hua et al., 2022) proposed a zero-shot prediction method for cutting tool wear prediction based on causal inference. A deep convolutional neural network and a causal representation model are jointly optimized to extract invariant causal signal features, which can be generalized to non-stationary manufacturing environments without any new data. In (Wu & Wang, 2021) cause–effect relations intrinsic to an engineering design problem are employed to decompose a pertinent simulation-based causal ANN into sub-networks. Attractive subspaces are identified from the causal-ANN by leveraging its structure and the theory of Bayesian Networks. In (Zhou et al., 2024) a causal quality-related knowledge graph (CQKG) is constructed regarding quality defects in aerospace product manufacturing followed by a quality-related prompt dataset with multi-round conversations. Then, a novel P-tuning that adapts to utilize external CQKG instructions is designed to fine-tune an open-source ChatGLM base model. Based on this, a causal knowledge graph augmented LLM, named CausalKGPT, is developed assisting workers to analyze quality defects.

In summary, CML has demonstrated its value across multiple engineering domains, consistently outperforming traditional methods when interpretability, counterfactual reasoning, and intervention design are required. The reviewed studies employed a variety of computational tools. The same tools applied to IoT-based household energy systems enables fine-grained causal

modelling of device interactions, occupancy behavior, and external influences. This sets the stage for developing intelligent, explainable systems capable of adaptive energy management, targeted fault detection, and personalized interventions.

3 CML Foundations

The mathematical foundations of CML draw from causal inference theory, which focuses on estimating the effects of interventions or changes in one variable on another. CML builds upon several foundational concepts, such as Directed Acyclic Graphs, counterfactuals, and do-calculus.

A primary mathematical tool in causal inference is "Directed Acyclic Graphs (DAGs)". In a DAG, nodes represent variables, and directed edges (arrows) represent causal relationships between those variables. A directed edge from variable X to variable Y indicates that X has a causal effect on Y. The acyclic nature of the graph ensures that no variable can causally influence itself, either directly or indirectly, which reflects the directionality of cause-and-effect. Mathematically, a DAG can be used to represent the structure of a causal system, where each node in the graph is associated with a random variable, and edges indicate dependencies. This structure allows us to distinguish between direct and indirect causal effects, which is crucial for understanding how variables influence each other in a system.

A core concept in causal inference is the "Potential Outcomes Framework", also known as the "Rubin Causal Model" (RCM). This framework defines the causal effect of a treatment or intervention in terms of potential outcomes. Suppose we are interested in the effect of treatment T on outcome Y. Each individual has two potential outcomes: $Y(1)$, the outcome if the individual receives the treatment, and $Y(0)$, the outcome if they do not. The causal effect for an individual is the difference between these two potential outcomes: Causal Effect $= Y(1) - Y(0)$. However, in practice, we can only observe one of these outcomes for any individual, depending on whether they receive the treatment. This is known as the "fundamental problem of causal inference".

To estimate causal effects, we rely on techniques like "propensity score matching", where we compare individuals with similar characteristics but different treatments, and "instrumental variables" (IV), which help address confounding factors that may bias the estimation of causal effects.

The framework of "do-calculus", introduced by Judea Pearl, is another powerful tool in causal inference. The do-operator, denoted as $do(X = x)$, represents an intervention that sets the variable X to a specific value x, effectively "forcing" the system to adopt this value, independent of its natural distribution. The do-calculus provides a set of rules for manipulating causal graphs to compute the effects of interventions. For example, if we want to compute the effect of an intervention $do(X = x)$ on the outcome Y, we can use the do-calculus to break down the problem into simpler subproblems, manipulating the graphical model to identify the relevant causal paths and avoid confounding influences. In a simplified sense, this process helps us estimate the expected outcome of an intervention, accounting for the causal relationships present in the system.

In real-world scenarios, causal effects are often estimated from "observational data" rather than controlled experiments. One of the key challenges is dealing with confounding, where a third variable influences both the treatment and the outcome. Techniques like "propensity score matching" and "instrumental variables" are used to estimate causal effects in the presence of confounders. For instance, in propensity score matching, we estimate the likelihood of receiving a treatment based on observed covariates, and then match treated and untreated individuals with similar propensity scores. This helps control for confounding and provides an unbiased estimate of the treatment effect.

In modern machine learning, causal reasoning is integrated into algorithms to learn not just correlations but also causal relationships. "Causal Bayesian Networks" is an example of how machine learning models can be extended to learn causal structures from data. These models rely on the principles of DAGs and the potential outcomes framework to incorporate causal relationships into predictive models.

4 CML Implementation Tools

4.1 Causal Inference Tools

In recent years, the integration of causal inference with machine learning has gained significant traction, driving the development of tools such as PyWhy™ and CausalML™. Both are Python libraries designed to facilitate causal analysis, but they differ in scope, methodology, and application domains.

4.1.1 Capabiltites

PyWhy™, developed by Microsoft, emphasizes a principled framework for causal inference. It is grounded in Judea Pearl's do-calculus and supports the entire causal inference pipeline: defining a causal graph, identifying causal effects, estimating them, and refuting the estimates. PyWhy™ primarily supports individual treatment effect (ITE) estimation using methods like propensity score matching and stratification. It allows users to incorporate domain knowledge explicitly via causal graphs and focuses on rigorous assumptions checking and refutation testing, making it suitable for transparent, reproducible research and educational settings.

In contrast, CausalML™, developed by Uber, is more application-oriented, providing an all-inclusive approach for uplift modeling and heterogeneous treatment effect estimation. It supports several advanced algorithms including meta-learners (T-, S-, X-, R-learners), Uplift Random Forests, and implementations for multiple treatment groups. While it can be used with observational data, CausalML™ is particularly geared toward randomized experimental data, enabling fine-grained personalization and targeting strategies—especially in business contexts like marketing, customer engagement, and causal impact analysis.

4.1.2 Differences and Relative Merits

A key difference lies in their core philosophy and use case. PyWhy™ is designed for robustness and methodological rigor. It appeals to researchers and data scientists aiming to understand and validate causal assumptions in a structured way. By requiring users to explicitly define causal models, PyWhy™ promotes transparency and critical reasoning but may require more domain expertise.

CausalML™, on the other hand, focuses on scalability and ease of use, especially for industry applications. Its consistent API mirrors that of standard machine learning libraries, allowing practitioners to estimate treatment effects with minimal code. It's particularly strong in supporting uplift modeling, which is crucial for real-time decision-making systems. However, CausalML™'s emphasis on performance over causal model specification can lead to overreliance on black-box models unless users remain cautious about confounding and selection bias, especially in observational settings.

4.1.3 Challenges and Limitations

Both tools face limitations reflective of broader challenges in causal inference. For PyWhy™, while it offers a robust pipeline, it may lack scalability for large-scale or high-dimensional data due to its limited support for newer machine learning methods like deep learning. Its usability can also be a barrier for non-specialists, particularly when defining causal graphs or conducting refutations.

CausalML™ excels in handling large-scale data and complex treatment setups, but its lack of built-in support for graphical models and formal assumption testing may make it vulnerable to biased inference if users fail to account for confounding variables correctly. Moreover, while powerful, uplift models can be data-hungry and sensitive to imbalance, particularly in cases with sparse treatment-response pairs.

4.2 Additional Tools Used

Pandas (McKinney & others, 2011) is a Python library with rich data structures and tools for working with structured datasets that are common in statistics, finance, mathematics, and social sciences and many other fields. The library provides comprehensive, intuitive routines for performing common manipulations and analyses on such datasets. It serves as a powerful

complement to the existing Python scientific stack, while implementing and enhancing the data manipulation tools found in other statistical programming languages such as the R language.

Matplotlib (Hunter, 2007) is a Python two-dimensional graph library that produces quality publication layouts in a variety of print formats and interactive environments across all platforms. Matplotlib can be used in Python scripts, Python and IPython shells, the Jupyter notebook (Kluyver et al., 2016), web application servers and in GUI toolkits. Some of the graphical representations that can be created are histograms, power spectra, bar charts, diagrams error bars, scatter plots, etc.

Diagrams.net (JGraph, 2005) is an open source solution for creating sketches and diagrams. The tool can be used online with various storage platforms (Atlassian Confluence Cloud and Jira Cloud, applications of Google, GitHub and Microsoft applications) and offline as a standalone application. Unofficial integrations exist and are available for many other platforms and tools.

5 The Case Study

5.1 The Dataset

The IDEAL Household Energy Dataset (Pullinger et al., 2021) was used, which includes measurements from 255 homes, such as electricity consumption at a sampling period of 1 sec, pulse-level gas data, temperature, humidity and lighting data at a sampling period of 12 secs for each room and data temperature from the boiler tubes concerning central heating and hot water at a sampling period of 12 secs. 39 of the homes also include plug-level monitoring of selected electrical appliances, real power measurement of the mains and main sub-circuits, and more detailed temperature monitoring of equipment that uses gas and heat, including radiators and taps. The data set additionally includes demographics, values and self-reported energy awareness of residents, household income, as well as building and space characteristics and a record of energy-intensive appliances. Secondary data include weather conditions and level of urbanization.

The project was funded by the UK Engineering and Physical Sciences Research Council through two initiatives and received ethical approval from the University of Edinburgh. All equipment met CE certification and was professionally installed.

Data fields are documented in the IDEALdata.md file. The dataset is available in CSV format, organized by household, room, and sensor ID. A metadata structure links participants to homes, rooms, sensor boxes, and sensors.

5.2 Data Preprocessing

The first step in preprocessing was quantifying data availability to decide which homes were suitable for inclusion in model training. Homes with at least six months of data on core features (electricity/gas use, temperature, humidity, lighting) were selected. Ultimately, 212 homes were used for electricity and 208 for gas consumption analysis.

Next, various features were computed for each home, such as average consumption during cold (Sept–Mar) and warm periods (July–Sept, Mar–July), and over the entire year. Additional engineered features were also included based on their causal relevance to the research questions.

These included:

- Number of external windows, doors, and walls (distinguished by whether they could be opened),
- Average room area and height, and total house volume,
- Proportion of space used for drying clothes,
- Percentage of space with thermostatic radiator valves (TRVs),
- Average gas and electricity prices per kWh.

Categorical data was numerically encoded (e.g., one-hot encoding), and missing values were imputed using average values.

Final datasets were compiled into CSV files for different seasonal periods. Specific labels were generated for answering research questions, such as binary indicators of high electricity or gas consumption. Each home's average usage was compared to the overall average, with labels assigned based on whether it was above or below this benchmark.

Additional questions were addressed similarly, including income level, number of windows, or number of electric heaters. Labels like high_income or many_electric_heaters reflect the presence of corresponding features.

Finally, Directed Acyclic Graphs (DAGs) were created using Diagrams.net to represent causal relationships for each question. These were then translated for use with the PyWhy™ tool. Unobserved confounding variables were marked as "U" where necessary. Outputs from both tools were consolidated into tables and JSON files for visualization and comparison.

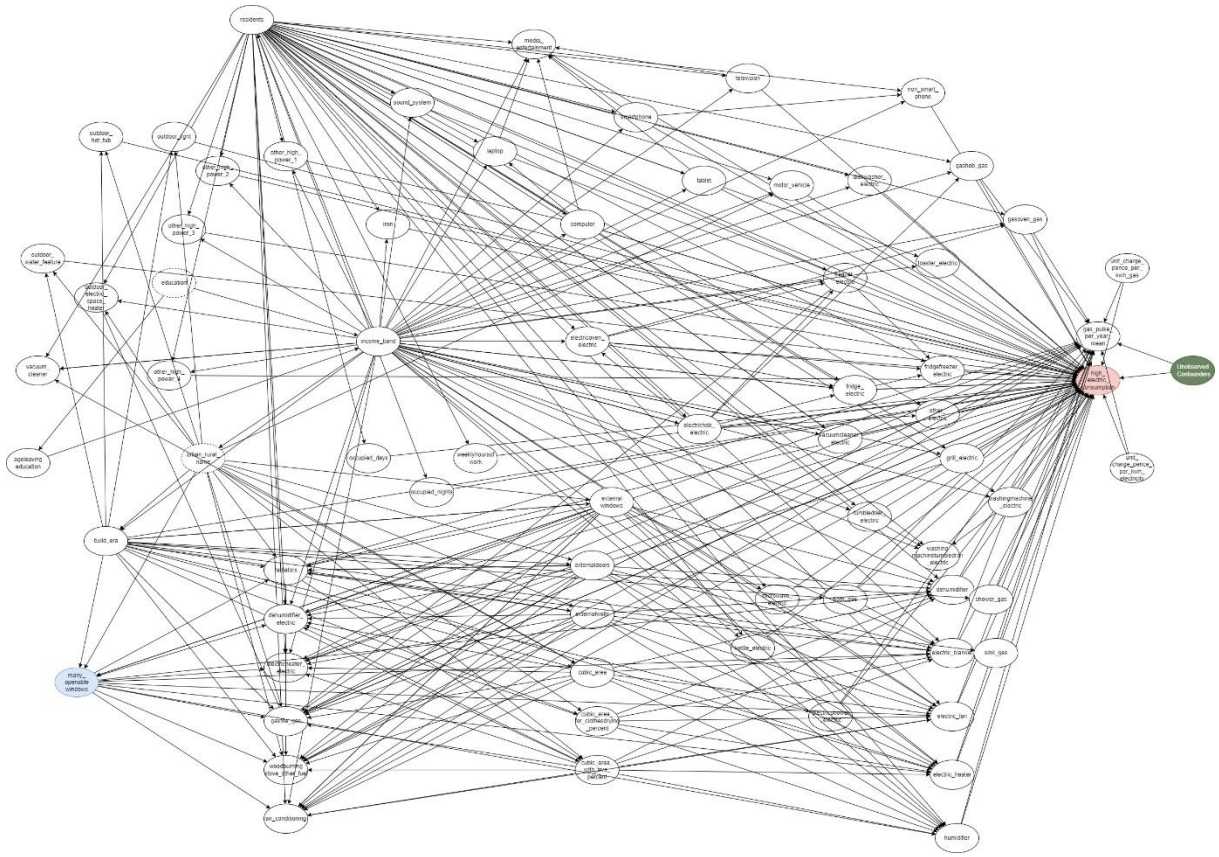


Fig. 1 The visualised DAG fed into the PyWhy™ tool to answer question 4 (electricity consumption). The treatment variable is marked in red, the outcome variable in blue, and the unobserved confounders in green.

5.3 Queries

The 18 queries formulated are presented in Table 1.

Table 1 CML queries and their usage context (architectural design changes - ADC, engineering interventions - EI, alternative uses of rooms- AUR, social intervention - SI, environmental conditions – EC, avg: average)

	Output	**Input**	**Usage**
1	Change in average consumption	high source of household income	SI
2		high level of education among the occupants	SI
3		large number of windows	ADC
4		large number of openable windows	EI
5		large number of exterior doors	ADC
6		a large number of exterior walls	ADC
7		large volume of the rooms	ADC

	Output	Input	Usage
8		large area of the rooms	ADC
9		high walls of the rooms	ADC
10		a large space (volume) for drying clothes	AU
11		large space with built-in thermostatic valves in the radiators	AU & EI
12		many tenants	AU
13		house located in an urban area	EC
14		change of time to winter time	EC
15	Change in avg consumption in winter	many electric heaters in the house	EI
16	Change in avg temperature in winter	many electric heaters in the house	EI
17	Change in avg consumption in summer	existence of electric fans or air conditioners	EI
18	Change in avg temperature in summer	existence of electric fans or air conditioners	EI

6 Results and Discussion

6.1 Results

The study utilizes meta-learners from the CausalML™ library (such as T-Learner, S-Learner, X-Learner, and R-Learner) and sensitivity analysis techniques from the PyWhy™ library. Both libraries incorporate diverse model architectures, notably Linear Regression and XGBoost (Extreme Gradient Boosting). Due to convergence issues in CausalML™, a cap of 100 iterations was imposed; however, increasing this threshold yielded negligible changes in estimates.

A typical result for one of the 18 queries, namely for query No 4, is presented in Figure 1. The rest 17 queries yield analogous results that can be seen in “Appendix” section. Question 4 analyzed whether households with a high number of openable windows exhibit different patterns in electricity usage. Homes exceeding the dataset's average of 5 openable windows were classified as having a high number of openable windows.

Analysis across multiple thresholds for classifying a home as “window-rich” showed a consistent trend: consumption increased when the number of windows exceeded five. Below that threshold, the probability decreased. This pattern suggests thermal inefficiency from excessive window exposure, offset by increased HVAC usage.

The study highlights the influence of confounders such as occupant environmental behavior and window size, advocating for richer datasets to refine estimates.

Findings across methodologies were as follows:

- Propensity Score Matching and Linear Regression via PyWhy™ estimated a 24% increase in consumption.
- Inverse Propensity Weighting and Propensity Score Stratification in PyWhy™ yielded higher estimates (43%).
- CausalML™ meta-learners produced diverse results: the XGBoost-based T-Learner aligned closely with PyWhy™ (23–24%), while other learners (S, X, R) diverged significantly (3–15%).

The MLP model was excluded from plots due to large deviations across all queries.

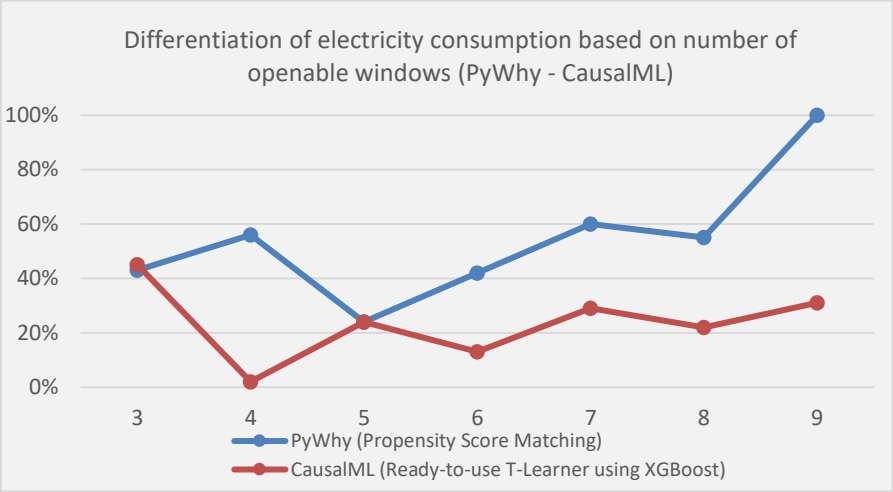


Fig. 2 Comparative graph of PyWhy™ and CausalML™ results for different number of openable windows.

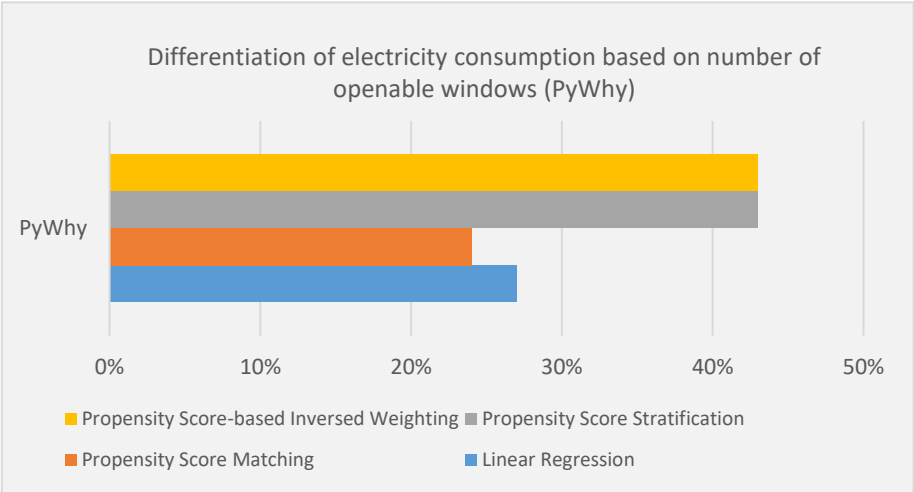


Fig. 3 Results of query 4 regarding the PyWhy™ tool.

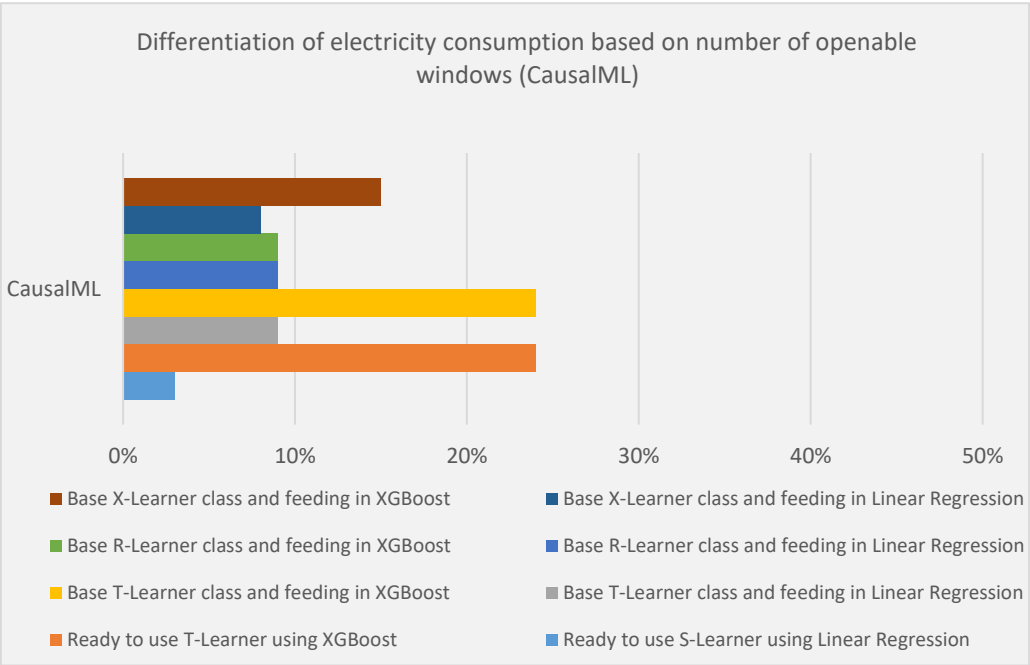


Fig. 4 Results of query 4 regarding the CausalML™ tool.

Since true counterfactuals are unobservable outside randomized trials, validation hinges on internal consistency, notably via Sensitivity Analysis. This evaluates the robustness of treatment effect estimates (Average Treatment Effect on the Treated - ATT) under hypothetical unobserved confounding. AUUC is also employed to assess performance, particularly regarding how well uplift models capture treatment heterogeneity.

6.1.1 Sensitivity Analysis

Sensitivity checks in PyWhy™ indicated that:

- Replacing the dataset with a random subset or adding a random common cause marginally changed ATT estimates (~1% change), affirming causal assumptions.
- Introducing unobserved confounders caused an 11% variation.
- Replacing the treatment variable with a random independent one resulted in a 27% drop (towards zero), indicating strong causal relevance of the original treatment.

Associated p-values further validated these shifts, with the placebo test showing high p-values (~2) and other transformations yielding values from 0.13 to 0.9.

CausalML™ produced analogous findings, except for tests involving unobserved confounders, which are not supported by that library. Overall, both tools corroborated the validity of the causal claims.

6.1.2 Performance on Synthetic Data

To overcome the absence of known ground truth in the real dataset, a synthetic dataset of 10,000 samples was generated (repeated 10 times), with training/validation split (80/20). The function simulate_nuisance_and_easy_treatment was used to simulate data with known causal structures.

Performance assessments revealed that:

- XGBoost-based meta-learners (X, T, R) consistently outperformed others in estimating ATT.
- Linear Regression-based learners underperformed, with S-Learner performing worst.
- MSE was lower for training and validation sets using linear models but less accurate regarding true ATT values.
- KL Divergence analyses indicated S-Learner with Linear Regression had significantly higher distributional divergence (~12×) from the true data, marking it as the least reliable.

These results underscore XGBoost's superior capacity to model complex causal relationships under non-linear and high-dimensional contexts.

6.1.3 Uplift Curve Validation

Uplift modeling quantifies how treatment influences individual outcomes. AUUC provides a summary performance metric by measuring the difference in responses between treated and control groups.

- XGBoost-based learners again exhibited the highest AUUC values for both training and validation sets.
- Linear Regression models showed poor uplift performance, especially the S-Learner, which failed to approach the ideal prediction line (y = x).
- Visual comparisons of predicted vs. actual outcomes reaffirmed that only the XGBoost learners closely approximated the ideal causal estimator.

6.2 Assumptions and Limitations

Several assumptions underlie the findings:

- Hourly averages were used to approximate monthly values due to limited computational resources. Missing data might bias estimates.
- Sensor placement inconsistencies (e.g., height of temperature sensors) were not accounted for, possibly distorting real indoor temperature readings.
- Zero gas consumption over long periods may stem from lack of use or sensor/system faults, further complicating inference.

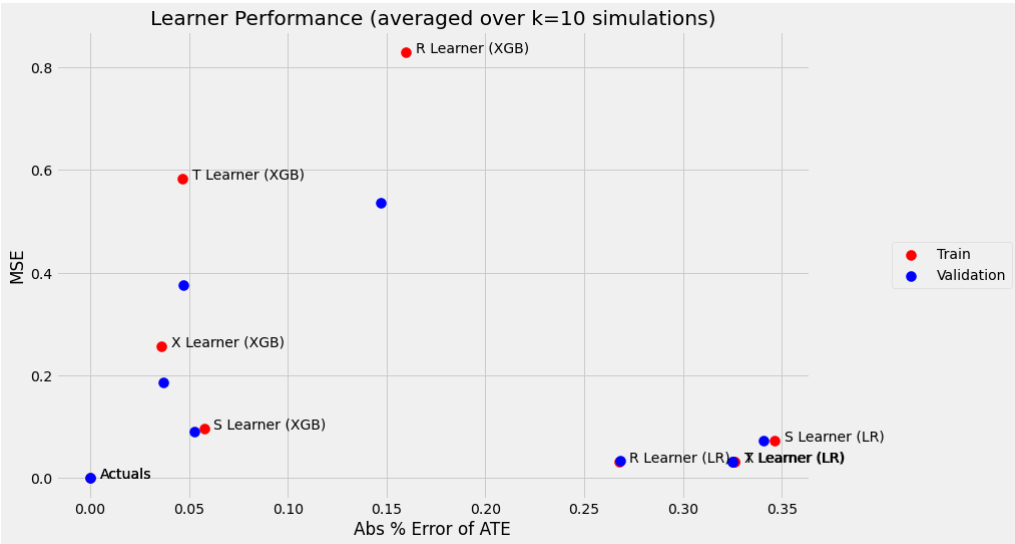


Fig. 5 Scatter plot of meta-learner performance.

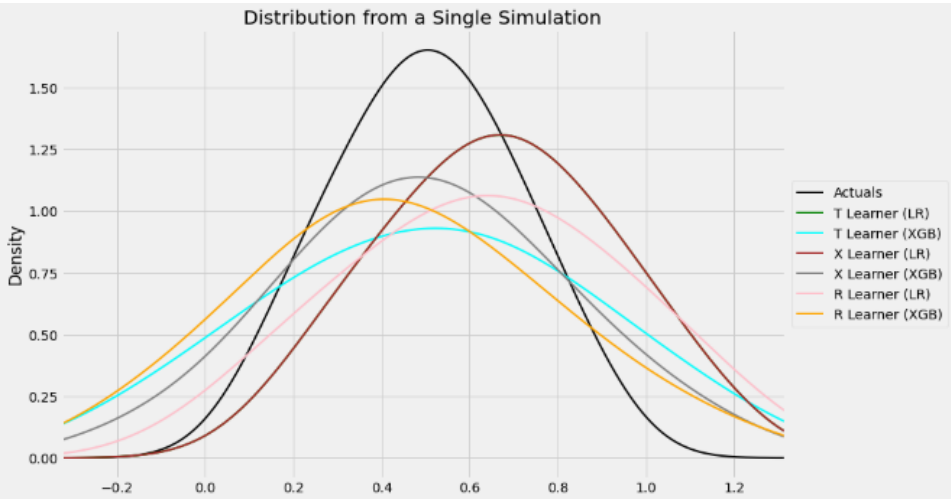


Fig. 6 Probability distributions of validation datasets.

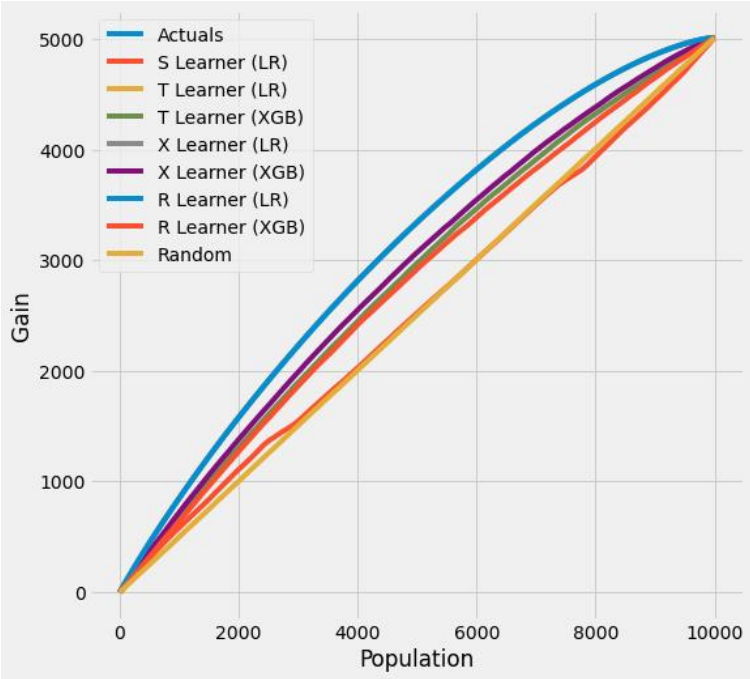


Fig. 7 Uplift curves for validation datasets.

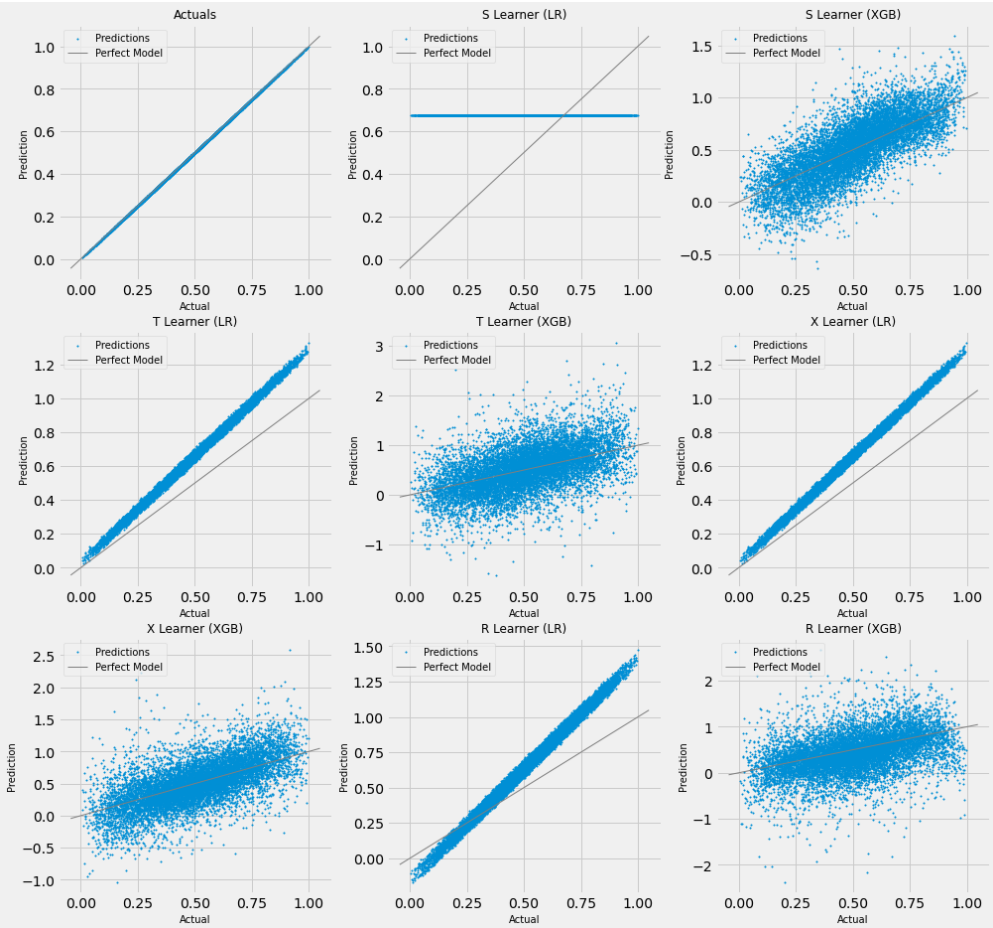


Fig. 8 Comparison between meta-learner predictions and a perfect model for validation datasets.

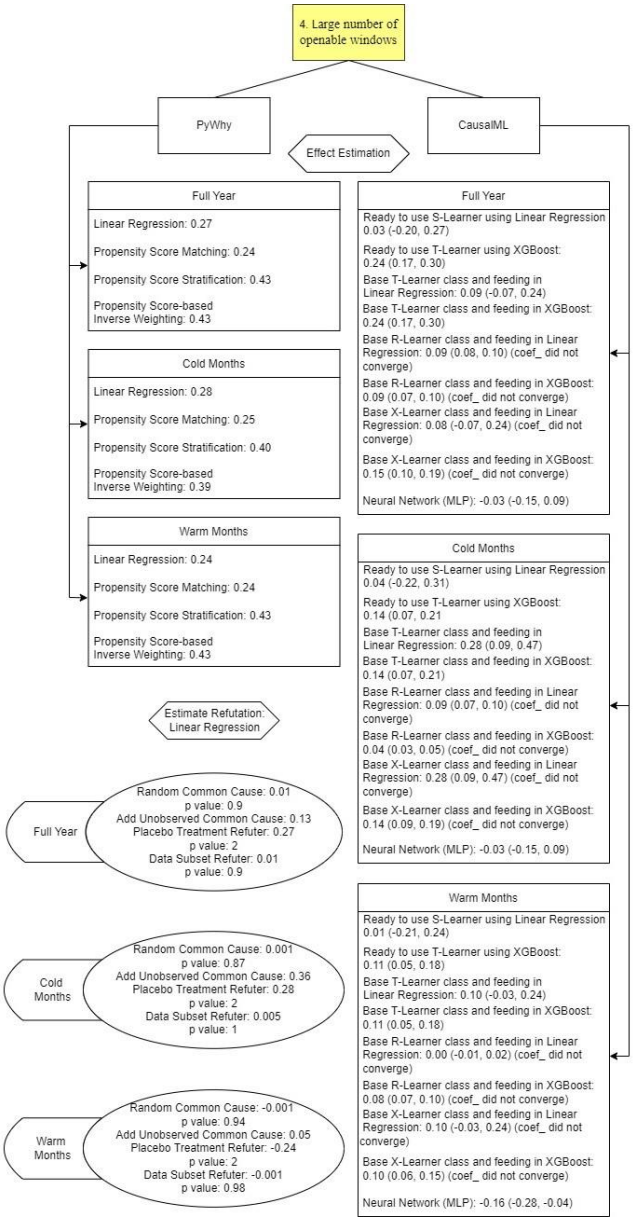


Fig. 9 Analytic results of query 4 (electricity usage)

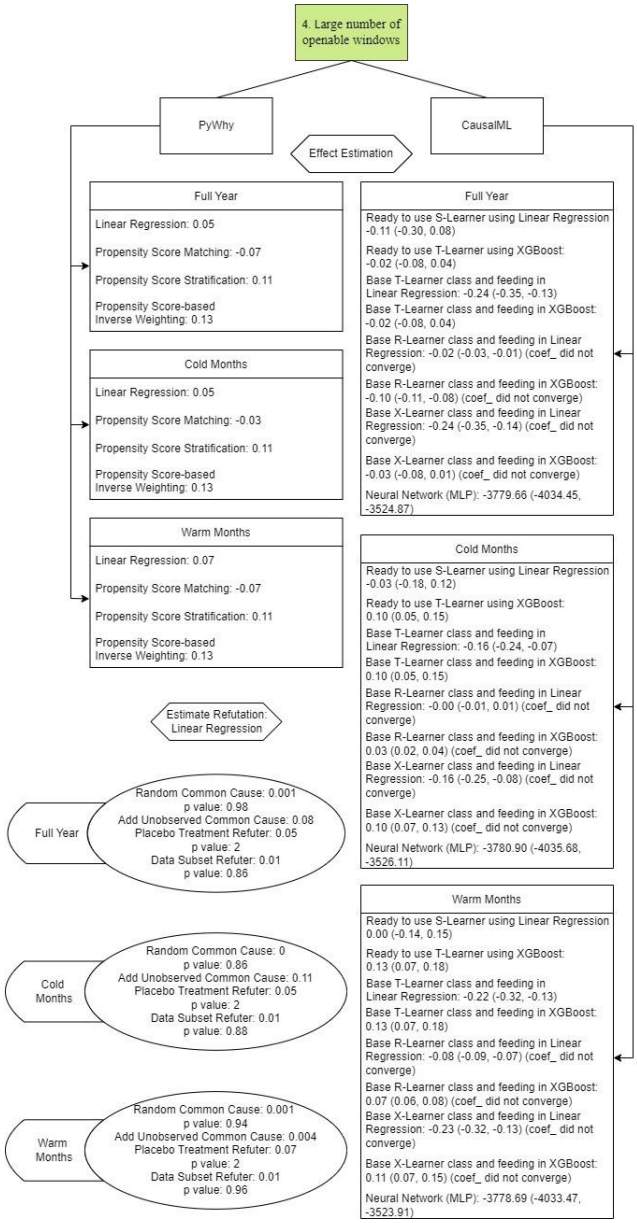


Fig. 10 Analytic results of query 4 (gas usage)

7 Conclusions and Future Work

The slight divergence of the results of the two tools points to the field knowledge required to design the structural causal model fed into PyWhy™, this not necessarily leading to a better result than CausalML™.

Additionally, internal validation methods are able to assess the causality assumptions we considered, but not the treatment effect which is rendered impossible by having the true value only for experimental data.

For both PyWhy™ and CausalML™ it is important to study the estimated ICU of each algorithm, in order to choose the domain-specific optimum. It is also important to evaluate the assumptions of the causal relationship based on the pre-existing domain knowledge, via the internal validation methods provided.

The tools used in this work can be improved, mainly in terms of a friendlier user experience. More specifically, it would be useful to include in PyWhy™ a user interface for creating the structural causal model and automatically translating it into code

that the tool processes, thus avoiding its time-consuming visualization with a third-party program and its manual conversion into recognizable code from PyWhy™.

Some suggestions for future work on data preprocessing are as follows:

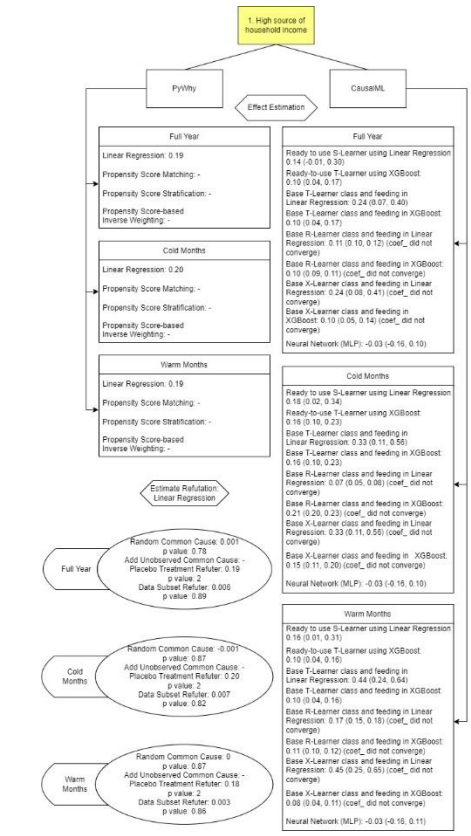
- Using a larger data set to train the causal model with the aim of deriving more reliable results. For example, of the 212 houses in the used data set, 43 were rejected due to incomplete data.
- Collecting data from different locations, in order to diversify the data and draw more individualized conclusions.
- Calculating HCV outside the ICU to reveal heterogeneity between population subgroups, aiming at more individualized outcomes and, by extension, enforcing more targeted treatment by subgroup.

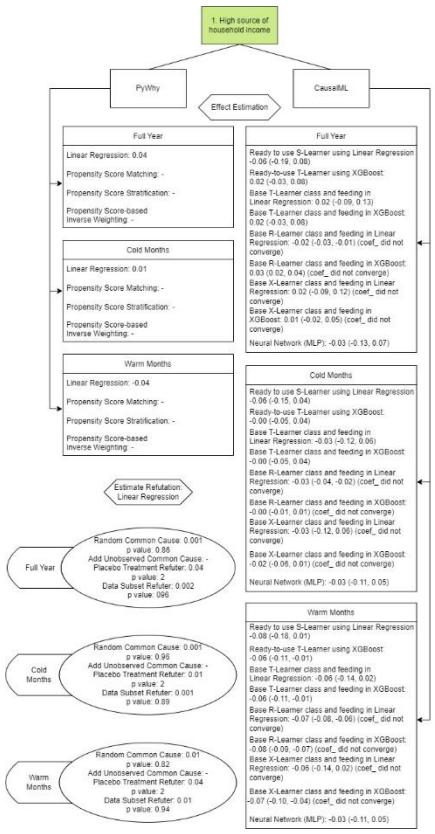
In addition, suggestions for future work regarding outcome evaluation are as follows:

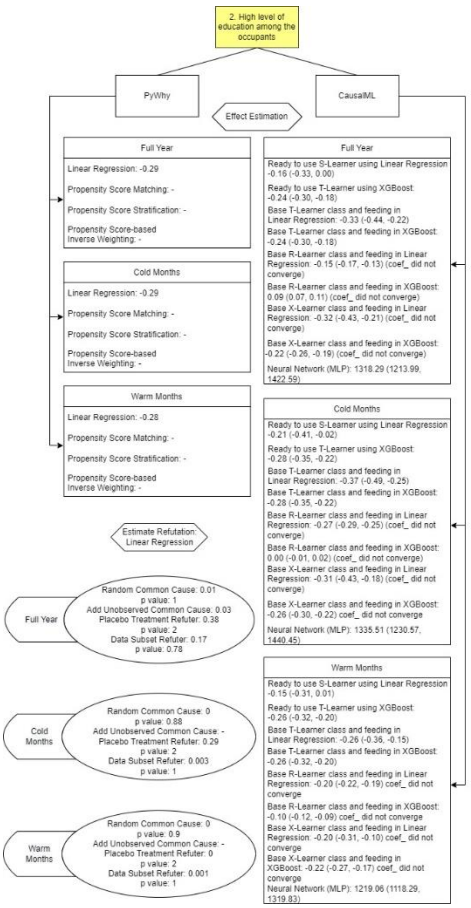
- Application of the results produced in real conditions (in this case, e.g., changing the architecture of new houses, the distribution of electrical appliances per house), aiming at their practical evaluation.
- Trying additional CML tools such as the CausalLib (Shimoni et al., 2019) and Causalinference (Wong, 2014) libraries of the Python language.

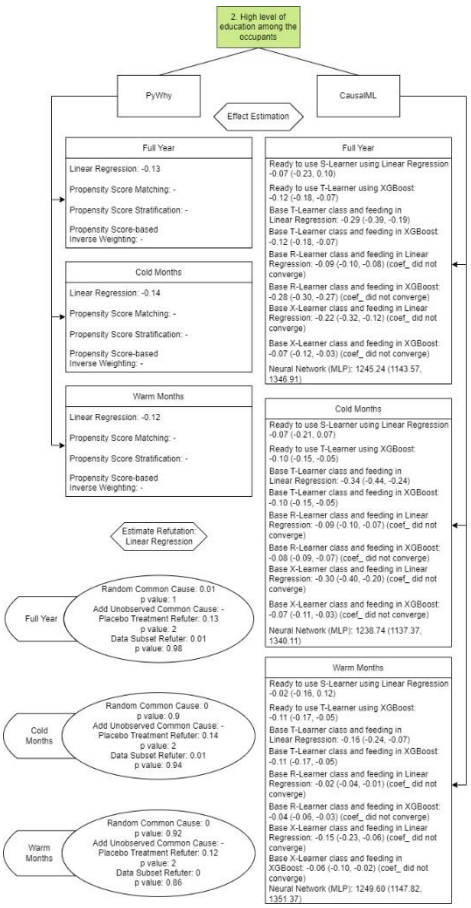
Appendix

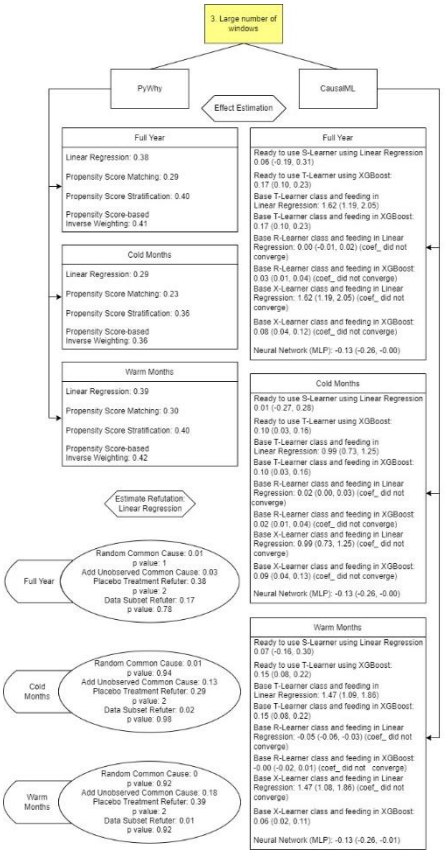
In this section the results of the remaining 17 queries are provided, as categorized by table 1. Results of queries regarding electricity consumption are distinguished by yellow heading, natural gas consumption by green heading and temperature by blue heading.

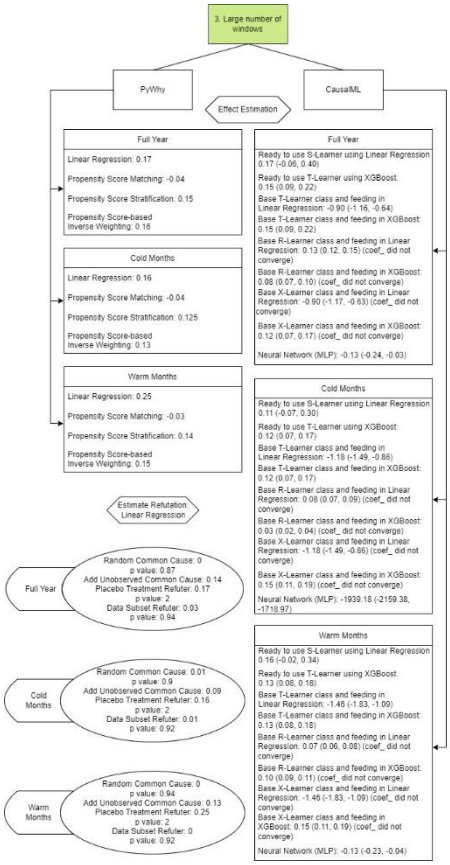

1. High source of household income
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: 0.19
Propensity Score Matching: -
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -
Cold Months
Linear Regression: 0.20
Propensity Score Matching: -
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -
Warm Months
Linear Regression: 0.19
Propensity Score Matching: -
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -
Full Year
Ready to use S-Learner using Linear Regression 0.14 (-0.01, 0.30)
Ready-to-use T-Learner using XGBoost: 0.10 (0.04, 0.17)
Base T-Learner class and feeding in Linear Regression: 0.24 (0.07, 0.40)
Base T-Learner class and feeding in XGBoost: 0.10 (0.04, 0.17)
Base R-Learner class and feeding in Linear Regression: 0.11 (0.10, 0.12) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.10 (0.09, 0.11) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.24 (0.08, 0.41) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.10 (0.05, 0.14) (coef_ did not converge)
Neural Network (MLP): -0.03 (-0.16, 0.10)
Cold Months
Ready to use S-Learner using Linear Regression 0.18 (0.02, 0.34)
Ready-to-use T-Learner using XGBoost: 0.16 (0.10, 0.23)
Base T-Learner class and feeding in Linear Regression: 0.33 (0.11, 0.56)
Base T-Learner class and feeding in XGBoost: 0.16 (0.10, 0.23)
Base R-Learner class and feeding in Linear Regression: 0.07 (0.05, 0.08) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.21 (0.20, 0.23) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.33 (0.11, 0.56) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.15 (0.11, 0.20) (coef_ did not converge)
Neural Network (MLP): -0.03 (-0.16, 0.10)
Warm Months
Ready to use S-Learner using Linear Regression 0.16 (0.01, 0.31)
Ready-to-use T-Learner using XGBoost: 0.10 (0.04, 0.16)
Base T-Learner class and feeding in Linear Regression: 0.44 (0.24, 0.64)
Base T-Learner class and feeding in XGBoost: 0.10 (0.04, 0.16)
Base R-Learner class and feeding in Linear Regression: 0.17 (0.15, 0.18) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.11 (0.10, 0.12) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.45 (0.25, 0.65) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.08 (0.04, 0.11) (coef_ did not converge)
Neural Network (MLP): -0.03 (-0.16, 0.11)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: 0.001 p value: 0.78
Add Unobserved Common Cause: - Placebo Treatment Refuter: 0.19 p value: 2
Data Subset Refuter: 0.006 p value: 0.89
Cold Months
Random Common Cause: -0.001 p value: 0.87
Add Unobserved Common Cause: - Placebo Treatment Refuter: 0.20 p value: 2
Data Subset Refuter: 0.007 p value: 0.82
Warm Months
Random Common Cause: 0 p value: 0.67
Add Unobserved Common Cause: - Placebo Treatment Refuter: 0.18 p value: 2
Data Subset Refuter: 0.003 p value: 0.86

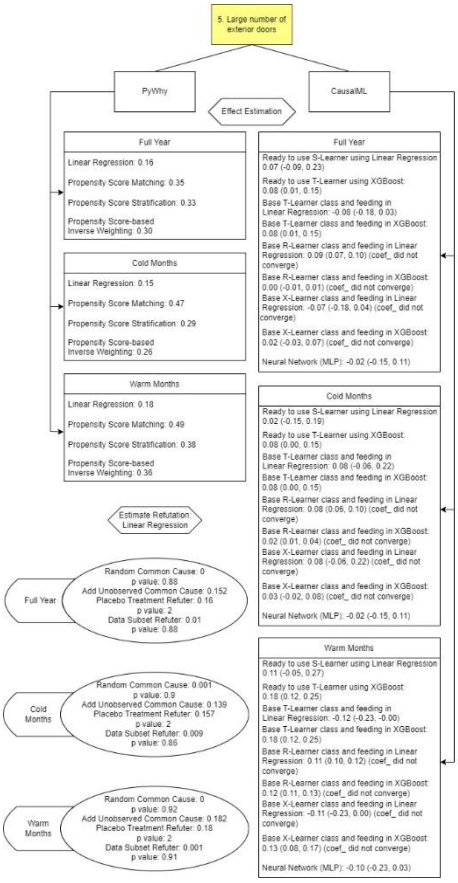

1. High source of household income
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: 0.04
Propensity Score Matching: -
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -
Cold Months
Linear Regression: 0.01
Propensity Score Matching: -
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -
Warm Months
Linear Regression: -0.04
Propensity Score Matching: -
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -
Full Year
Ready to use S-Learner using Linear Regression -0.06 (-0.19, 0.06)
Ready-to-use T-Learner using XGBoost: 0.02 (-0.03, 0.08)
Base T-Learner class and feeding in Linear Regression: 0.02 (-0.09, 0.13)
Base T-Learner class and feeding in XGBoost: 0.02 (-0.03, 0.08)
Base R-Learner class and feeding in Linear Regression: -0.02 (-0.03, -0.01) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.03 (0.02, 0.04) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.02 (-0.09, 0.12) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.01 (-0.02, 0.05) (coef_ did not converge)
Neural Network (MLP): -0.03 (-0.13, 0.07)
Cold Months
Ready to use S-Learner using Linear Regression -0.06 (-0.15, 0.04)
Ready-to-use T-Learner using XGBoost: -0.00 (-0.05, 0.04)
Base T-Learner class and feeding in Linear Regression: -0.03 (-0.12, 0.06)
Base T-Learner class and feeding in XGBoost: -0.00 (-0.05, 0.04)
Base R-Learner class and feeding in Linear Regression: -0.03 (-0.04, -0.02) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.00 (-0.01, 0.01) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.03 (-0.12, 0.06) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.02 (-0.06, 0.01) (coef_ did not converge)
Neural Network (MLP): -0.03 (-0.11, 0.05)
Warm Months
Ready to use S-Learner using Linear Regression -0.08 (-0.18, 0.01)
Ready-to-use T-Learner using XGBoost: -0.06 (-0.11, -0.01)
Base T-Learner class and feeding in Linear Regression: -0.06 (-0.14, 0.02)
Base T-Learner class and feeding in XGBoost: -0.06 (-0.11, -0.01)
Base R-Learner class and feeding in Linear Regression: -0.07 (-0.08, -0.06) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.08 (-0.09, -0.07) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.06 (-0.14, 0.02) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.07 (-0.10, -0.04) (coef_ did not converge)
Neural Network (MLP): -0.03 (-0.11, 0.05)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: 0.001 p value: 0.86
Add Unobserved Common Cause: - Placebo Treatment Refuter: 0.04 p value: 2
Data Subset Refuter: 0.002 p value: 0.96
Cold Months
Random Common Cause: 0.001 p value: 0.96
Add Unobserved Common Cause: - Placebo Treatment Refuter: 0.01 p value: 2
Data Subset Refuter: 0.001 p value: 0.89
Warm Months
Random Common Cause: 0.01 p value: 0.82
Add Unobserved Common Cause: - Placebo Treatment Refuter: 0.04 p value: 2
Data Subset Refuter: 0.01 p value: 0.94

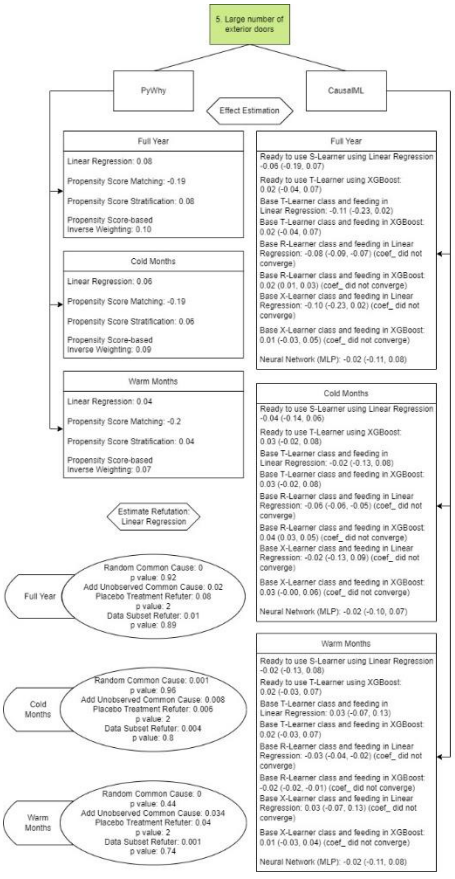

2. High level of education among the occupants
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: -0.29
Propensity Score Matching: -
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -
Cold Months
Linear Regression: -0.29
Propensity Score Matching: -
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -
Warm Months
Linear Regression: -0.28
Propensity Score Matching: -
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -
Full Year
Ready to use S-Learner using Linear Regression -0.16 (-0.33, 0.00)
Ready to use T-Learner using XGBoost: -0.24 (-0.30, -0.18)
Base T-Learner class and feeding in Linear Regression: -0.33 (-0.44, -0.22)
Base T-Learner class and feeding in XGBoost: -0.24 (-0.30, -0.18)
Base R-Learner class and feeding in Linear Regression: -0.15 (-0.17, -0.13) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.09 (0.07, 0.11) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.32 (-0.43, -0.21) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.22 (-0.26, -0.19) (coef_ did not converge)
Neural Network (MLP): 1318.29 (1213.99, 1422.59)
Cold Months
Ready to use S-Learner using Linear Regression -0.21 (-0.41, -0.02)
Ready to use T-Learner using XGBoost: -0.28 (-0.35, -0.22)
Base T-Learner class and feeding in Linear Regression: -0.37 (-0.49, -0.25)
Base T-Learner class and feeding in XGBoost: -0.28 (-0.35, -0.22)
Base R-Learner class and feeding in Linear Regression: -0.27 (-0.29, -0.25) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.00 (-0.01, 0.02) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.31 (-0.43, -0.18) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.26 (-0.30, -0.22) coef_ did not converge
Neural Network (MLP): 1335.51 (1230.57, 1440.45)
Warm Months
Ready to use S-Learner using Linear Regression -0.15 (-0.31, 0.01)
Ready to use T-Learner using XGBoost: -0.26 (-0.32, -0.20)
Base T-Learner class and feeding in Linear Regression: -0.26 (-0.36, -0.15)
Base T-Learner class and feeding in XGBoost: -0.26 (-0.32, -0.20)
Base R-Learner class and feeding in Linear Regression: -0.20 (-0.22, -0.19) coef_ did not converge
Base R-Learner class and feeding in XGBoost: -0.10 (-0.12, -0.09) coef_ did not converge
Base X-Learner class and feeding in Linear Regression: -0.20 (-0.31, -0.10) coef_ did not converge
Base X-Learner class and feeding in XGBoost: -0.22 (-0.27, -0.17) coef_ did not converge
Neural Network (MLP): 1219.06 (1118.29, 1319.83)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: 0.01 p value: 1
Add Unobserved Common Cause: 0.03 Placebo Treatment Refuter: 0.38 p value: 2
Data Subset Refuter: 0.17 p value: 0.78
Cold Months
Random Common Cause: 0 p value: 0.88
Add Unobserved Common Cause: - Placebo Treatment Refuter: 0.29 p value: 2
Data Subset Refuter: 0.003 p value: 1
Warm Months
Random Common Cause: 0 p value: 0.9
Add Unobserved Common Cause: - Placebo Treatment Refuter: 0 p value: 2
Data Subset Refuter: 0.001 p value: 1

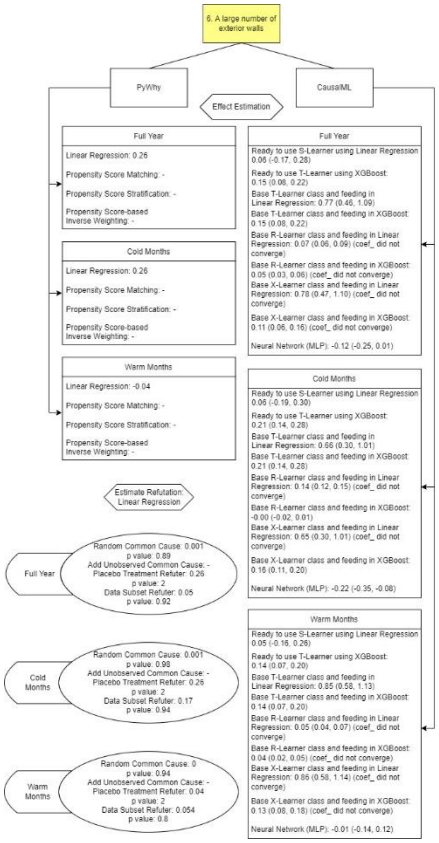

2. High level of education among the occupants
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: -0.13
Propensity Score Matching: -
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -
Cold Months
Linear Regression: -0.14
Propensity Score Matching: -
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -
Warm Months
Linear Regression: -0.12
Propensity Score Matching: -
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -
Full Year
Ready to use S-Learner using Linear Regression -0.07 (-0.23, 0.10)
Ready to use T-Learner using XGBoost: -0.12 (-0.18, -0.07)
Base T-Learner class and feeding in Linear Regression: -0.29 (-0.39, -0.19)
Base T-Learner class and feeding in XGBoost: -0.12 (-0.18, -0.07)
Base R-Learner class and feeding in Linear Regression: -0.09 (-0.10, -0.08) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.28 (-0.30, -0.27) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.22 (-0.32, -0.12) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.07 (-0.12, -0.03) (coef_ did not converge)
Neural Network (MLP): 1245.24 (1143.57, 1346.91)
Cold Months
Ready to use S-Learner using Linear Regression -0.07 (-0.21, 0.07)
Ready to use T-Learner using XGBoost: -0.10 (-0.15, -0.05)
Base T-Learner class and feeding in Linear Regression: -0.34 (-0.44, -0.24)
Base T-Learner class and feeding in XGBoost: -0.10 (-0.15, -0.05)
Base R-Learner class and feeding in Linear Regression: -0.09 (-0.10, -0.07) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.08 (-0.09, -0.07) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.30 (-0.40, -0.20) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.07 (-0.11, -0.03) (coef_ did not converge)
Neural Network (MLP): 1238.74 (1137.37, 1340.11)
Warm Months
Ready to use S-Learner using Linear Regression -0.02 (-0.16, 0.12)
Ready to use T-Learner using XGBoost: -0.11 (-0.17, -0.05)
Base T-Learner class and feeding in Linear Regression: -0.16 (-0.24, -0.07)
Base T-Learner class and feeding in XGBoost: -0.11 (-0.17, -0.05)
Base R-Learner class and feeding in Linear Regression: -0.02 (-0.04, -0.01) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.04 (-0.06, -0.03) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.15 (-0.23, -0.06) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.06 (-0.10, -0.02) (coef_ did not converge)
Neural Network (MLP): 1249.60 (1147.82, 1351.37)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: 0.01 p value: 1
Add Unobserved Common Cause: - Placebo Treatment Refuter: 0.13 p value: 2
Data Subset Refuter: 0.01 p value: 0.98
Cold Months
Random Common Cause: 0 p value: 0.9
Add Unobserved Common Cause: - Placebo Treatment Refuter: 0.14 p value: 2
Data Subset Refuter: 0.01 p value: 0.94
Warm Months
Random Common Cause: 0 p value: 0.92
Add Unobserved Common Cause: - Placebo Treatment Refuter: 0.12 p value: 2
Data Subset Refuter: 0 p value: 0.86

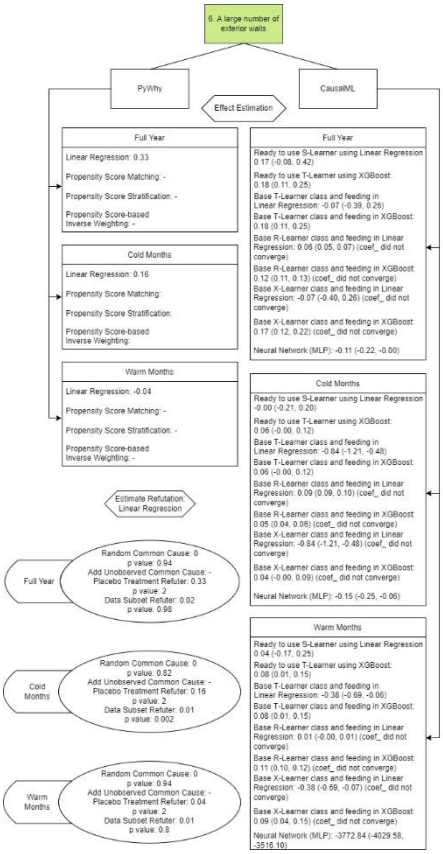

3. Large number of windows
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: 0.38
Propensity Score Matching: 0.29
Propensity Score Stratification: 0.40
Propensity Score-based Inverse Weighting: 0.41
Cold Months
Linear Regression: 0.29
Propensity Score Matching: 0.23
Propensity Score Stratification: 0.36
Propensity Score-based Inverse Weighting: 0.36
Warm Months
Linear Regression: 0.39
Propensity Score Matching: 0.30
Propensity Score Stratification: 0.40
Propensity Score-based Inverse Weighting: 0.42
Full Year
Ready to use S-Learner using Linear Regression 0.06 (-0.19, 0.31)
Ready to use T-Learner using XGBoost: 0.17 (0.10, 0.23)
Base T-Learner class and feeding in Linear Regression: 1.62 (1.19, 2.05)
Base T-Learner class and feeding in XGBoost: 0.17 (0.10, 0.23)
Base R-Learner class and feeding in Linear Regression: 0.00 (-0.01, 0.02) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.03 (0.01, 0.04) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 1.62 (1.19, 2.05) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.08 (0.04, 0.12) (coef_ did not converge)
Neural Network (MLP): -0.13 (-0.26, -0.00)
Cold Months
Ready to use S-Learner using Linear Regression 0.01 (-0.27, 0.28)
Ready to use T-Learner using XGBoost: 0.10 (0.03, 0.16)
Base T-Learner class and feeding in Linear Regression: 0.99 (0.73, 1.25)
Base T-Learner class and feeding in XGBoost: 0.10 (0.03, 0.16)
Base R-Learner class and feeding in Linear Regression: 0.02 (0.00, 0.03) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.02 (0.01, 0.04) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.99 (0.73, 1.25) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.09 (0.04, 0.13) (coef_ did not converge)
Neural Network (MLP): -0.13 (-0.26, -0.00)
Warm Months
Ready to use S-Learner using Linear Regression 0.07 (-0.16, 0.30)
Ready to use T-Learner using XGBoost: 0.15 (0.08, 0.22)
Base T-Learner class and feeding in Linear Regression: 1.47 (1.09, 1.86)
Base T-Learner class and feeding in XGBoost: 0.15 (0.08, 0.22)
Base R-Learner class and feeding in Linear Regression: -0.05 (-0.06, -0.03) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.00 (-0.02, 0.01) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 1.47 (1.08, 1.86) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.06 (0.02, 0.11)
Neural Network (MLP): -0.13 (-0.26, -0.01)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: 0.01
p value: 1
Add Unobserved Common Cause: 0.03
Placebo Treatment Refuter: 0.38
p value: 2
Data Subset Refuter: 0.17
p value: 0.78
Cold Months
Random Common Cause: 0.01
p value: 0.94
Add Unobserved Common Cause: 0.13
Placebo Treatment Refuter: 0.29
p value: 2
Data Subset Refuter: 0.02
p value: 0.98
Warm Months
Random Common Cause: 0
p value: 0.92
Add Unobserved Common Cause: 0.18
Placebo Treatment Refuter: 0.39
p value: 2
Data Subset Refuter: 0.01
p value: 0.92

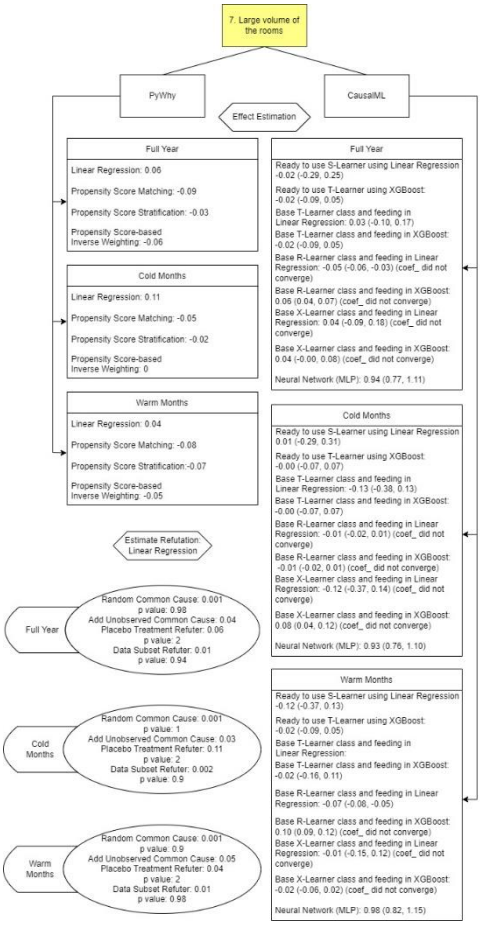

3. Large number of windows
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: 0.17
Propensity Score Matching: -0.04
Propensity Score Stratification: 0.15
Propensity Score-based Inverse Weighting: 0.16
Cold Months
Linear Regression: 0.16
Propensity Score Matching: -0.04
Propensity Score Stratification: 0.125
Propensity Score-based Inverse Weighting: 0.13
Warm Months
Linear Regression: 0.25
Propensity Score Matching: -0.03
Propensity Score Stratification: 0.14
Propensity Score-based Inverse Weighting: 0.15
Full Year
Ready to use S-Learner using Linear Regression 0.17 (-0.06, 0.40)
Ready to use T-Learner using XGBoost: 0.15 (0.09, 0.22)
Base T-Learner class and feeding in Linear Regression: -0.90 (-1.16, -0.64)
Base T-Learner class and feeding in XGBoost: 0.15 (0.09, 0.22)
Base R-Learner class and feeding in Linear Regression: 0.13 (0.12, 0.15) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.08 (0.07, 0.10) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.90 (-1.17, -0.63) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.12 (0.07, 0.17) (coef_ did not converge)
Neural Network (MLP): -0.13 (-0.24, -0.03)
Cold Months
Ready to use S-Learner using Linear Regression 0.11 (-0.07, 0.30)
Ready to use T-Learner using XGBoost: 0.12 (0.07, 0.17)
Base T-Learner class and feeding in Linear Regression: -1.18 (-1.49, -0.86)
Base T-Learner class and feeding in XGBoost: 0.12 (0.07, 0.17)
Base R-Learner class and feeding in Linear Regression: 0.08 (0.07, 0.09) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.03 (0.02, 0.04) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -1.18 (-1.49, -0.86) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.15 (0.11, 0.19) (coef_ did not converge)
Neural Network (MLP): -1939.18 (-2159.38, -1718.97)
Warm Months
Ready to use S-Learner using Linear Regression 0.16 (-0.02, 0.34)
Ready to use T-Learner using XGBoost: 0.13 (0.08, 0.18)
Base T-Learner class and feeding in Linear Regression: -1.46 (-1.83, -1.09)
Base T-Learner class and feeding in XGBoost: 0.13 (0.08, 0.18)
Base R-Learner class and feeding in Linear Regression: 0.07 (0.06, 0.08) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.10 (0.09, 0.11) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -1.46 (-1.83, -1.09) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.15 (0.11, 0.19) (coef_ did not converge)
Neural Network (MLP): -0.13 (-0.23, -0.04)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: 0
p value: 0.87
Add Unobserved Common Cause: 0.14
Placebo Treatment Refuter: 0.17
p value: 2
Data Subset Refuter: 0.03
p value: 0.94
Cold Months
Random Common Cause: 0.01
p value: 0.9
Add Unobserved Common Cause: 0.09
Placebo Treatment Refuter: 0.16
p value: 2
Data Subset Refuter: 0.01
p value: 0.92
Warm Months
Random Common Cause: 0
p value: 0.94
Add Unobserved Common Cause: 0.13
Placebo Treatment Refuter: 0.25
p value: 2
Data Subset Refuter: 0
p value: 0.92

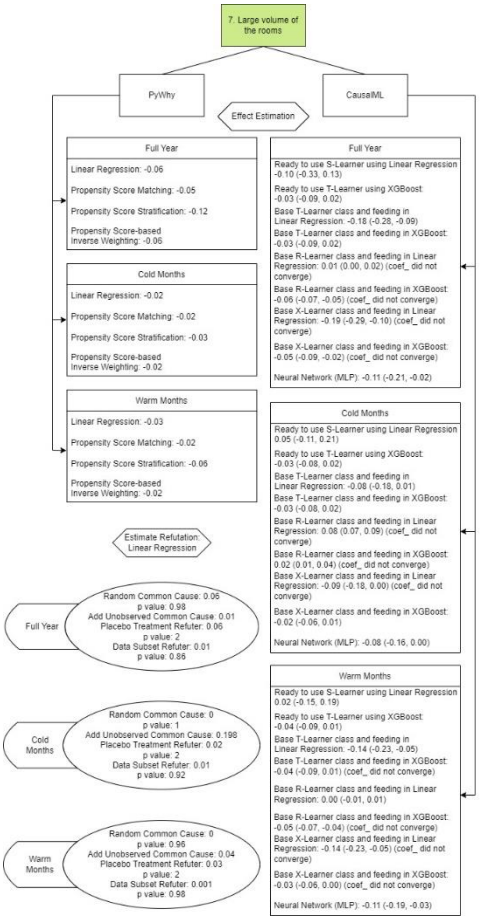

5. Large number of exterior doors
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: 0.16
Propensity Score Matching: 0.35
Propensity Score Stratification: 0.33
Propensity Score-based Inverse Weighting: 0.30
Cold Months
Linear Regression: 0.15
Propensity Score Matching: 0.47
Propensity Score Stratification: 0.29
Propensity Score-based Inverse Weighting: 0.26
Warm Months
Linear Regression: 0.18
Propensity Score Matching: 0.49
Propensity Score Stratification: 0.38
Propensity Score-based Inverse Weighting: 0.36
Full Year
Ready to use S-Learner using Linear Regression 0.07 (-0.09, 0.23)
Ready to use T-Learner using XGBoost: 0.08 (0.01, 0.15)
Base T-Learner class and feeding in Linear Regression: -0.06 (-0.18, 0.03)
Base T-Learner class and feeding in XGBoost: 0.08 (0.01, 0.15)
Base R-Learner class and feeding in Linear Regression: 0.09 (0.07, 0.10) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.00 (-0.01, 0.01) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.07 (-0.18, 0.04) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.02 (-0.03, 0.07) (coef_ did not converge)
Neural Network (MLP): -0.02 (-0.15, 0.11)
Cold Months
Ready to use S-Learner using Linear Regression 0.02 (-0.15, 0.19)
Ready to use T-Learner using XGBoost: 0.08 (0.00, 0.15)
Base T-Learner class and feeding in Linear Regression: 0.08 (-0.06, 0.22)
Base T-Learner class and feeding in XGBoost: 0.08 (0.00, 0.15)
Base R-Learner class and feeding in Linear Regression: 0.08 (0.06, 0.10) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.02 (0.01, 0.04) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.08 (-0.06, 0.22) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.03 (-0.02, 0.08) (coef_ did not converge)
Neural Network (MLP): -0.02 (-0.15, 0.11)
Warm Months
Ready to use S-Learner using Linear Regression 0.11 (-0.05, 0.27)
Ready to use T-Learner using XGBoost: 0.18 (0.12, 0.25)
Base T-Learner class and feeding in Linear Regression: -0.12 (-0.23, -0.00)
Base T-Learner class and feeding in XGBoost: 0.18 (0.12, 0.25)
Base R-Learner class and feeding in Linear Regression: 0.11 (0.10, 0.12) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.12 (0.11, 0.13) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.11 (-0.23, 0.00) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.13 (0.08, 0.17) (coef_ did not converge)
Neural Network (MLP): -0.10 (-0.23, 0.03)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: 0
p value: 0.88
Add Unobserved Common Cause: 0.152
Placebo Treatment Refuter: 0.16
p value: 2
Data Subset Refuter: 0.01
p value: 0.88
Cold Months
Random Common Cause: 0.001
p value: 0.9
Add Unobserved Common Cause: 0.139
Placebo Treatment Refuter: 0.157
p value: 2
Data Subset Refuter: 0.009
p value: 0.86
Warm Months
Random Common Cause: 0
p value: 0.92
Add Unobserved Common Cause: 0.182
Placebo Treatment Refuter: 0.18
p value: 2
Data Subset Refuter: 0.001
p value: 0.91

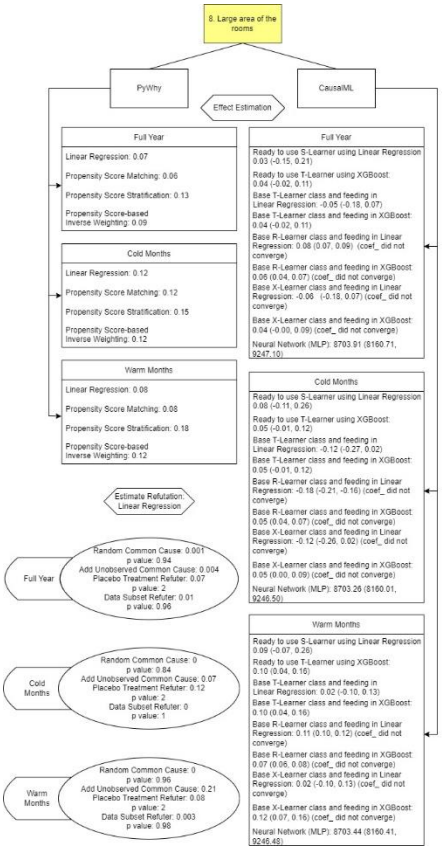

5. Large number of exterior doors
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: 0.08
Propensity Score Matching: -0.19
Propensity Score Stratification: 0.08
Propensity Score-based Inverse Weighting: 0.10
Cold Months
Linear Regression: 0.06
Propensity Score Matching: -0.19
Propensity Score Stratification: 0.06
Propensity Score-based Inverse Weighting: 0.09
Warm Months
Linear Regression: 0.04
Propensity Score Matching: -0.2
Propensity Score Stratification: 0.04
Propensity Score-based Inverse Weighting: 0.07
Full Year
Ready to use S-Learner using Linear Regression -0.06 (-0.19, 0.07)
Ready to use T-Learner using XGBoost: 0.02 (-0.04, 0.07)
Base T-Learner class and feeding in Linear Regression: -0.11 (-0.23, 0.02)
Base T-Learner class and feeding in XGBoost: 0.02 (-0.04, 0.07)
Base R-Learner class and feeding in Linear Regression: -0.08 (-0.09, -0.07) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.02 (0.01, 0.03) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.10 (-0.23, 0.02) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.01 (-0.03, 0.05) (coef_ did not converge)
Neural Network (MLP): -0.02 (-0.11, 0.08)
Cold Months
Ready to use S-Learner using Linear Regression -0.04 (-0.14, 0.06)
Ready to use T-Learner using XGBoost: 0.03 (-0.02, 0.08)
Base T-Learner class and feeding in Linear Regression: -0.02 (-0.13, 0.08)
Base T-Learner class and feeding in XGBoost: 0.03 (-0.02, 0.08)
Base R-Learner class and feeding in Linear Regression: -0.06 (-0.06, -0.05) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.04 (0.03, 0.05) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.02 (-0.13, 0.09) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.03 (-0.00, 0.06) (coef_ did not converge)
Neural Network (MLP): -0.02 (-0.10, 0.07)
Warm Months
Ready to use S-Learner using Linear Regression -0.02 (-0.13, 0.08)
Ready to use T-Learner using XGBoost: 0.02 (-0.03, 0.07)
Base T-Learner class and feeding in Linear Regression: 0.03 (-0.07, 0.13)
Base T-Learner class and feeding in XGBoost: 0.02 (-0.03, 0.07)
Base R-Learner class and feeding in Linear Regression: -0.03 (-0.04, -0.02) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.02 (-0.02, -0.01) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.03 (-0.07, 0.13) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.01 (-0.03, 0.04) (coef_ did not converge)
Neural Network (MLP): -0.02 (-0.11, 0.08)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: 0
p value: 0.92
Add Unobserved Common Cause: 0.02
Placebo Treatment Refuter: 0.08
p value: 2
Data Subset Refuter: 0.01
p value: 0.89
Cold Months
Random Common Cause: 0.001
p value: 0.96
Add Unobserved Common Cause: 0.008
Placebo Treatment Refuter: 0.006
p value: 2
Data Subset Refuter: 0.004
p value: 0.8
Warm Months
Random Common Cause: 0
p value: 0.44
Add Unobserved Common Cause: 0.034
Placebo Treatment Refuter: 0.04
p value: 2
Data Subset Refuter: 0.001
p value: 0.74

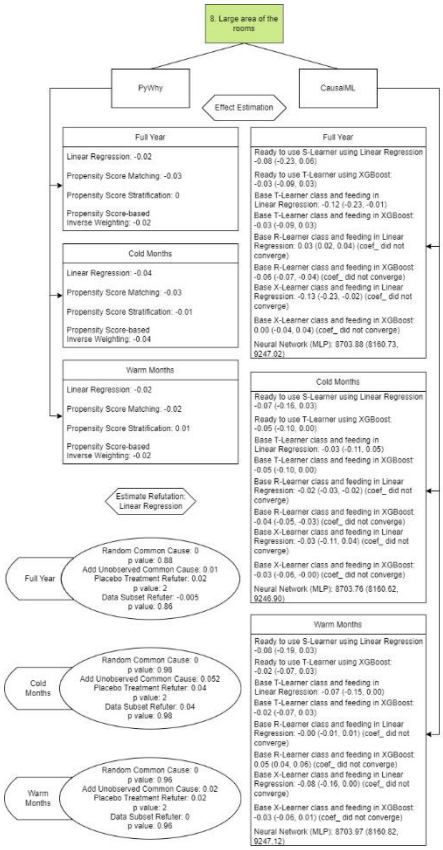

6. A large number of exterior walls
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: 0.26
Propensity Score Matching: -
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -
Cold Months
Linear Regression: 0.26
Propensity Score Matching: -
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -
Warm Months
Linear Regression: -0.04
Propensity Score Matching: -
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -
Full Year
Ready to use S-Learner using Linear Regression 0.06 (-0.17, 0.28)
Ready to use T-Learner using XGBoost: 0.15 (0.08, 0.22)
Base T-Learner class and feeding in Linear Regression: 0.77 (0.46, 1.09)
Base T-Learner class and feeding in XGBoost: 0.15 (0.08, 0.22)
Base R-Learner class and feeding in Linear Regression: 0.07 (0.06, 0.09) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.05 (0.03, 0.06) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.78 (0.47, 1.10) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.11 (0.06, 0.16) (coef_ did not converge)
Neural Network (MLP): -0.12 (-0.25, 0.01)
Cold Months
Ready to use S-Learner using Linear Regression 0.05 (-0.19, 0.30)
Ready to use T-Learner using XGBoost: 0.21 (0.14, 0.28)
Base T-Learner class and feeding in Linear Regression: 0.66 (0.30, 1.01)
Base T-Learner class and feeding in XGBoost: 0.21 (0.14, 0.28)
Base R-Learner class and feeding in Linear Regression: 0.14 (0.12, 0.15) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.00 (-0.02, 0.01)
Base X-Learner class and feeding in Linear Regression: 0.65 (0.30, 1.01) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.16 (0.11, 0.20)
Neural Network (MLP): -0.22 (-0.35, -0.08)
Warm Months
Ready to use S-Learner using Linear Regression 0.05 (-0.16, 0.26)
Ready to use T-Learner using XGBoost: 0.14 (0.07, 0.20)
Base T-Learner class and feeding in Linear Regression: 0.85 (0.58, 1.13)
Base T-Learner class and feeding in XGBoost: 0.14 (0.07, 0.20)
Base R-Learner class and feeding in Linear Regression: 0.05 (0.04, 0.07) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.04 (0.02, 0.05) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.86 (0.58, 1.14) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.13 (0.08, 0.18) (coef_ did not converge)
Neural Network (MLP): -0.01 (-0.14, 0.12)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: 0.001 p value: 0.89
Add Unobserved Common Cause: - Placebo Treatment Refuter: 0.26 p value: 2
Data Subset Refuter: 0.05 p value: 0.92
Cold Months
Random Common Cause: 0.001 p value: 0.98
Add Unobserved Common Cause: - Placebo Treatment Refuter: 0.26 p value: 2
Data Subset Refuter: 0.17 p value: 0.94
Warm Months
Random Common Cause: 0 p value: 0.94
Add Unobserved Common Cause: - Placebo Treatment Refuter: 0.04 p value: 2
Data Subset Refuter: 0.054 p value: 0.8

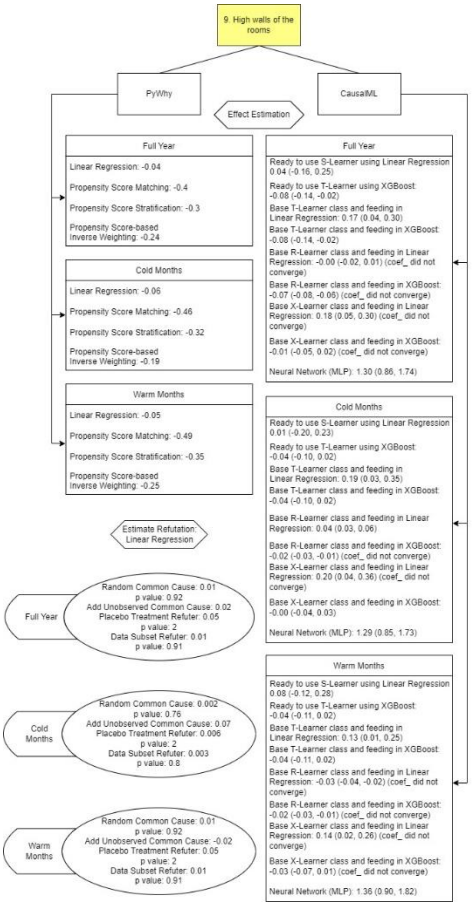

6. A large number of exterior walls
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: 0.33
Propensity Score Matching: -
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -
Cold Months
Linear Regression: 0.16
Propensity Score Matching:
Propensity Score Stratification:
Propensity Score-based Inverse Weighting:
Warm Months
Linear Regression: -0.04
Propensity Score Matching: -
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -
Full Year
Ready to use S-Learner using Linear Regression 0.17 (-0.08, 0.42)
Ready to use T-Learner using XGBoost: 0.18 (0.11, 0.25)
Base T-Learner class and feeding in Linear Regression: -0.07 (-0.39, 0.26)
Base T-Learner class and feeding in XGBoost: 0.18 (0.11, 0.25)
Base R-Learner class and feeding in Linear Regression: 0.06 (0.05, 0.07) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.12 (0.11, 0.13) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.07 (-0.40, 0.26) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.17 (0.12, 0.22) (coef_ did not converge)
Neural Network (MLP): -0.11 (-0.22, -0.00)
Cold Months
Ready to use S-Learner using Linear Regression -0.00 (-0.21, 0.20)
Ready to use T-Learner using XGBoost: 0.06 (-0.00, 0.12)
Base T-Learner class and feeding in Linear Regression: -0.84 (-1.21, -0.48)
Base T-Learner class and feeding in XGBoost: 0.06 (-0.00, 0.12)
Base R-Learner class and feeding in Linear Regression: 0.09 (0.09, 0.10) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.05 (0.04, 0.06) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.84 (-1.21, -0.48) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.04 (-0.00, 0.09) (coef_ did not converge)
Neural Network (MLP): -0.15 (-0.25, -0.06)
Warm Months
Ready to use S-Learner using Linear Regression 0.04 (-0.17, 0.25)
Ready to use T-Learner using XGBoost: 0.08 (0.01, 0.15)
Base T-Learner class and feeding in Linear Regression: -0.38 (-0.69, -0.06)
Base T-Learner class and feeding in XGBoost: 0.08 (0.01, 0.15)
Base R-Learner class and feeding in Linear Regression: 0.01 (-0.00, 0.01) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.11 (0.10, 0.12) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.38 (-0.69, -0.07) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.09 (0.04, 0.15) (coef_ did not converge)
Neural Network (MLP): -3772.84 (-4029.58, -3516.10)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: 0 p value: 0.94
Add Unobserved Common Cause: - Placebo Treatment Refuter: 0.33 p value: 2
Data Subset Refuter: 0.02 p value: 0.98
Cold Months
Random Common Cause: 0 p value: 0.82
Add Unobserved Common Cause: - Placebo Treatment Refuter: 0.16 p value: 2
Data Subset Refuter: 0.01 p value: 0.002
Warm Months
Random Common Cause: 0 p value: 0.94
Add Unobserved Common Cause: - Placebo Treatment Refuter: 0.04 p value: 2
Data Subset Refuter: 0.01 p value: 0.8

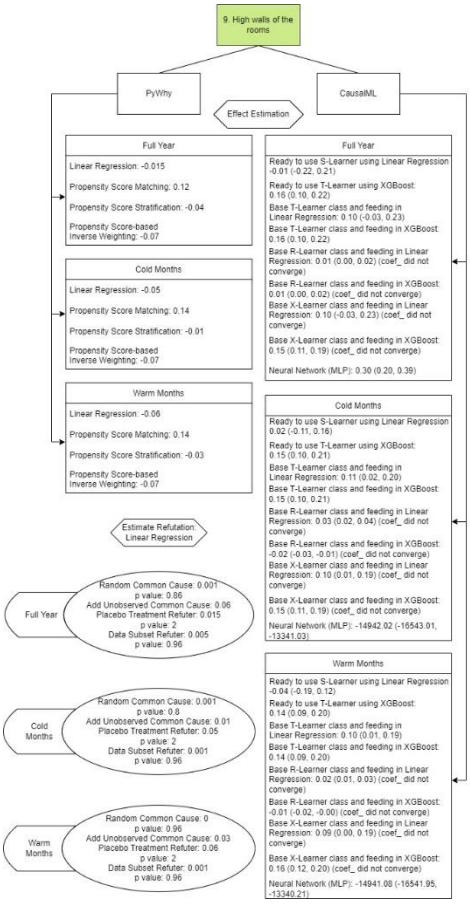

7. Large volume of the rooms
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: 0.06
Propensity Score Matching: -0.09
Propensity Score Stratification: -0.03
Propensity Score-based Inverse Weighting: -0.06
Cold Months
Linear Regression: 0.11
Propensity Score Matching: -0.05
Propensity Score Stratification: -0.02
Propensity Score-based Inverse Weighting: 0
Warm Months
Linear Regression: 0.04
Propensity Score Matching: -0.08
Propensity Score Stratification:-0.07
Propensity Score-based Inverse Weighting: -0.05
Full Year
Ready to use S-Learner using Linear Regression -0.02 (-0.29, 0.25)
Ready to use T-Learner using XGBoost: -0.02 (-0.09, 0.05)
Base T-Learner class and feeding in Linear Regression: 0.03 (-0.10, 0.17)
Base T-Learner class and feeding in XGBoost: -0.02 (-0.09, 0.05)
Base R-Learner class and feeding in Linear Regression: -0.05 (-0.06, -0.03) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.06 (0.04, 0.07) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.04 (-0.09, 0.18) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.04 (-0.00, 0.08) (coef_ did not converge)
Neural Network (MLP): 0.94 (0.77, 1.11)
Cold Months
Ready to use S-Learner using Linear Regression 0.01 (-0.29, 0.31)
Ready to use T-Learner using XGBoost: -0.00 (-0.07, 0.07)
Base T-Learner class and feeding in Linear Regression: -0.13 (-0.38, 0.13)
Base T-Learner class and feeding in XGBoost: -0.00 (-0.07, 0.07)
Base R-Learner class and feeding in Linear Regression: -0.01 (-0.02, 0.01) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.01 (-0.02, 0.01) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.12 (-0.37, 0.14) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.08 (0.04, 0.12) (coef_ did not converge)
Neural Network (MLP): 0.93 (0.76, 1.10)
Warm Months
Ready to use S-Learner using Linear Regression -0.12 (-0.37, 0.13)
Ready to use T-Learner using XGBoost: -0.02 (-0.09, 0.05)
Base T-Learner class and feeding in Linear Regression:
Base T-Learner class and feeding in XGBoost: -0.02 (-0.16, 0.11)
Base R-Learner class and feeding in Linear Regression: -0.07 (-0.08, -0.05)
Base R-Learner class and feeding in XGBoost: 0.10 (0.09, 0.12) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.01 (-0.15, 0.12) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.02 (-0.06, 0.02) (coef_ did not converge)
Neural Network (MLP): 0.98 (0.82, 1.15)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: 0.001 p value: 0.98
Add Unobserved Common Cause: 0.04 Placebo Treatment Refuter: 0.06 p value: 2
Data Subset Refuter: 0.01 p value: 0.94
Cold Months
Random Common Cause: 0.001 p value: 1
Add Unobserved Common Cause: 0.03 Placebo Treatment Refuter: 0.11 p value: 2
Data Subset Refuter: 0.002 p value: 0.9
Warm Months
Random Common Cause: 0.001 p value: 0.9
Add Unobserved Common Cause: 0.05 Placebo Treatment Refuter: 0.04 p value: 2
Data Subset Refuter: 0.01 p value: 0.98

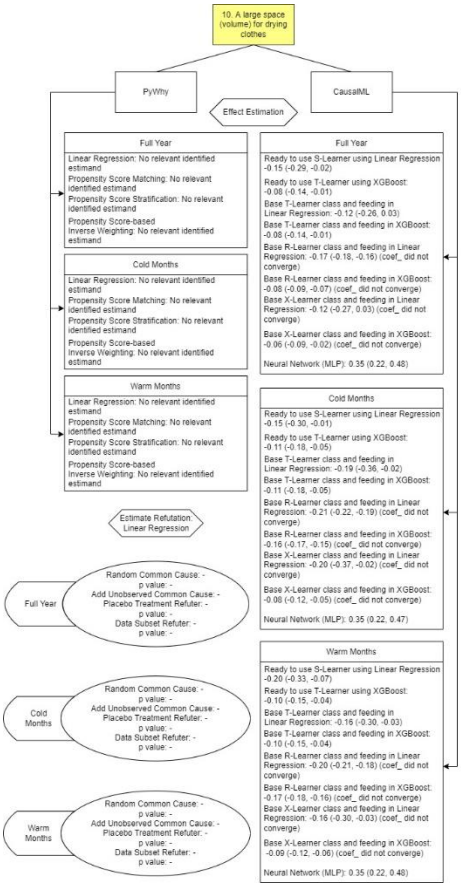

7. Large volume of the rooms
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: -0.06
Propensity Score Matching: -0.05
Propensity Score Stratification: -0.12
Propensity Score-based Inverse Weighting: -0.06
Cold Months
Linear Regression: -0.02
Propensity Score Matching: -0.02
Propensity Score Stratification: -0.03
Propensity Score-based Inverse Weighting: -0.02
Warm Months
Linear Regression: -0.03
Propensity Score Matching: -0.02
Propensity Score Stratification: -0.06
Propensity Score-based Inverse Weighting: -0.02
Full Year
Ready to use S-Learner using Linear Regression -0.10 (-0.33, 0.13)
Ready to use T-Learner using XGBoost: -0.03 (-0.09, 0.02)
Base T-Learner class and feeding in Linear Regression: -0.18 (-0.28, -0.09)
Base T-Learner class and feeding in XGBoost: -0.03 (-0.09, 0.02)
Base R-Learner class and feeding in Linear Regression: 0.01 (0.00, 0.02) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.06 (-0.07, -0.05) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.19 (-0.29, -0.10) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.05 (-0.09, -0.02) (coef_ did not converge)
Neural Network (MLP): -0.11 (-0.21, -0.02)
Cold Months
Ready to use S-Learner using Linear Regression 0.05 (-0.11, 0.21)
Ready to use T-Learner using XGBoost: -0.03 (-0.08, 0.02)
Base T-Learner class and feeding in Linear Regression: -0.08 (-0.18, 0.01)
Base T-Learner class and feeding in XGBoost: -0.03 (-0.08, 0.02)
Base R-Learner class and feeding in Linear Regression: 0.08 (0.07, 0.09) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.02 (0.01, 0.04) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.09 (-0.18, 0.00) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.02 (-0.06, 0.01)
Neural Network (MLP): -0.08 (-0.16, 0.00)
Warm Months
Ready to use S-Learner using Linear Regression 0.02 (-0.15, 0.19)
Ready to use T-Learner using XGBoost: -0.04 (-0.09, 0.01)
Base T-Learner class and feeding in Linear Regression: -0.14 (-0.23, -0.05)
Base T-Learner class and feeding in XGBoost: -0.04 (-0.09, 0.01) (coef_ did not converge)
Base R-Learner class and feeding in Linear Regression: 0.00 (-0.01, 0.01)
Base R-Learner class and feeding in XGBoost: -0.05 (-0.07, -0.04) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.14 (-0.23, -0.05) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.03 (-0.06, 0.00) (coef_ did not converge)
Neural Network (MLP): -0.11 (-0.19, -0.03)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: 0.06 p value: 0.98
Add Unobserved Common Cause: 0.01 Placebo Treatment Refuter: 0.06 p value: 2
Data Subset Refuter: 0.01 p value: 0.86
Cold Months
Random Common Cause: 0 p value: 1
Add Unobserved Common Cause: 0.198 Placebo Treatment Refuter: 0.02 p value: 2
Data Subset Refuter: 0.01 p value: 0.92
Warm Months
Random Common Cause: 0 p value: 0.96
Add Unobserved Common Cause: 0.04 Placebo Treatment Refuter: 0.03 p value: 2
Data Subset Refuter: 0.001 p value: 0.98

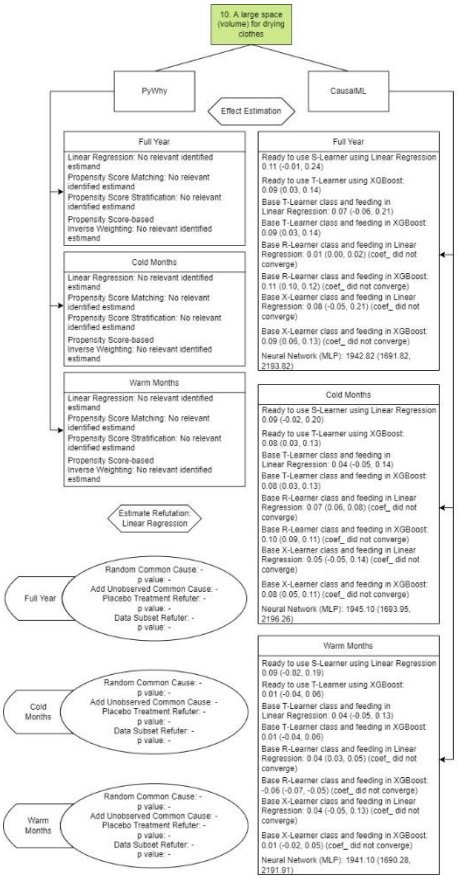

8. Large area of the rooms
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: 0.07
Propensity Score Matching: 0.06
Propensity Score Stratification: 0.13
Propensity Score-based Inverse Weighting: 0.09
Cold Months
Linear Regression: 0.12
Propensity Score Matching: 0.12
Propensity Score Stratification: 0.15
Propensity Score-based Inverse Weighting: 0.12
Warm Months
Linear Regression: 0.08
Propensity Score Matching: 0.08
Propensity Score Stratification: 0.18
Propensity Score-based Inverse Weighting: 0.12
Full Year
Ready to use S-Learner using Linear Regression 0.03 (-0.15, 0.21)
Ready to use T-Learner using XGBoost: 0.04 (-0.02, 0.11)
Base T-Learner class and feeding in Linear Regression: -0.05 (-0.18, 0.07)
Base T-Learner class and feeding in XGBoost: 0.04 (-0.02, 0.11)
Base R-Learner class and feeding in Linear Regression: 0.08 (0.07, 0.09) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.06 (0.04, 0.07) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.06 (-0.18, 0.07) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.04 (-0.00, 0.09) (coef_ did not converge)
Neural Network (MLP): 8703.91 (8160.71, 9247.10)
Cold Months
Ready to use S-Learner using Linear Regression 0.08 (-0.11, 0.26)
Ready to use T-Learner using XGBoost: 0.05 (-0.01, 0.12)
Base T-Learner class and feeding in Linear Regression: -0.12 (-0.27, 0.02)
Base T-Learner class and feeding in XGBoost: 0.05 (-0.01, 0.12)
Base R-Learner class and feeding in Linear Regression: -0.18 (-0.21, -0.16) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.05 (0.04, 0.07) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.12 (-0.26, 0.02) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.05 (0.00, 0.09) (coef_ did not converge)
Neural Network (MLP): 8703.26 (8160.01, 9246.50)
Warm Months
Ready to use S-Learner using Linear Regression 0.09 (-0.07, 0.26)
Ready to use T-Learner using XGBoost: 0.10 (0.04, 0.16)
Base T-Learner class and feeding in Linear Regression: 0.02 (-0.10, 0.13)
Base T-Learner class and feeding in XGBoost: 0.10 (0.04, 0.16)
Base R-Learner class and feeding in Linear Regression: 0.11 (0.10, 0.12) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.07 (0.06, 0.08) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.02 (-0.10, 0.13) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.12 (0.07, 0.16) (coef_ did not converge)
Neural Network (MLP): 8703.44 (8160.41, 9246.48)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: 0.001
p value: 0.94
Add Unobserved Common Cause: 0.004
Placebo Treatment Refuter: 0.07
p value: 2
Data Subset Refuter: 0.01
p value: 0.96
Cold Months
Random Common Cause: 0
p value: 0.84
Add Unobserved Common Cause: 0.07
Placebo Treatment Refuter: 0.12
p value: 2
Data Subset Refuter: 0
p value: 1
Warm Months
Random Common Cause: 0
p value: 0.96
Add Unobserved Common Cause: 0.21
Placebo Treatment Refuter: 0.08
p value: 2
Data Subset Refuter: 0.003
p value: 0.98

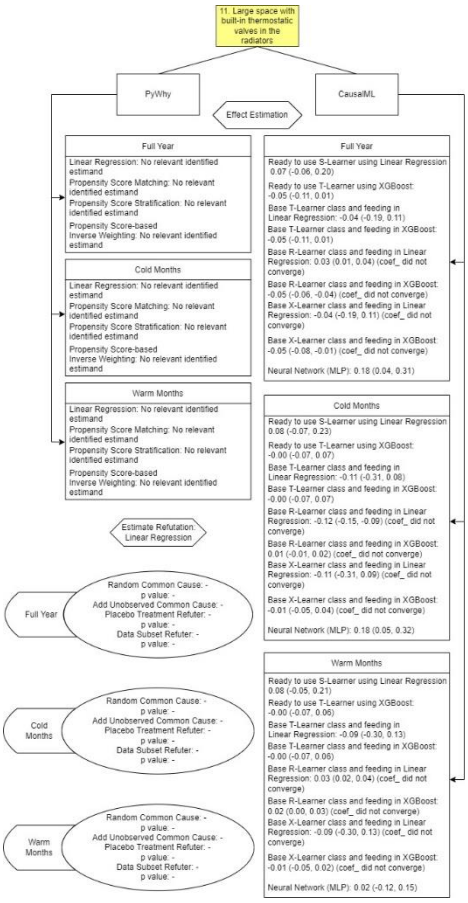

8. Large area of the rooms
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: -0.02
Propensity Score Matching: -0.03
Propensity Score Stratification: 0
Propensity Score-based Inverse Weighting: -0.02
Cold Months
Linear Regression: -0.04
Propensity Score Matching: -0.03
Propensity Score Stratification: -0.01
Propensity Score-based Inverse Weighting: -0.04
Warm Months
Linear Regression: -0.02
Propensity Score Matching: -0.02
Propensity Score Stratification: 0.01
Propensity Score-based Inverse Weighting: -0.02
Full Year
Ready to use S-Learner using Linear Regression -0.08 (-0.23, 0.06)
Ready to use T-Learner using XGBoost: -0.03 (-0.09, 0.03)
Base T-Learner class and feeding in Linear Regression: -0.12 (-0.23, -0.01)
Base T-Learner class and feeding in XGBoost: -0.03 (-0.09, 0.03)
Base R-Learner class and feeding in Linear Regression: 0.03 (0.02, 0.04) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.06 (-0.07, -0.04) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.13 (-0.23, -0.02) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.00 (-0.04, 0.04) (coef_ did not converge)
Neural Network (MLP): 8703.88 (8160.73, 9247.02)
Cold Months
Ready to use S-Learner using Linear Regression -0.07 (-0.16, 0.03)
Ready to use T-Learner using XGBoost: -0.05 (-0.10, 0.00)
Base T-Learner class and feeding in Linear Regression: -0.03 (-0.11, 0.05)
Base T-Learner class and feeding in XGBoost: -0.05 (-0.10, 0.00)
Base R-Learner class and feeding in Linear Regression: -0.02 (-0.03, -0.02) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.04 (-0.05, -0.03) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.03 (-0.11, 0.04) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.03 (-0.06, -0.00) (coef_ did not converge)
Neural Network (MLP): 8703.76 (8160.62, 9246.90)
Warm Months
Ready to use S-Learner using Linear Regression -0.00 (-0.19, 0.03)
Ready to use T-Learner using XGBoost: -0.02 (-0.07, 0.03)
Base T-Learner class and feeding in Linear Regression: -0.07 (-0.15, 0.00)
Base T-Learner class and feeding in XGBoost: -0.02 (-0.07, 0.03)
Base R-Learner class and feeding in Linear Regression: -0.00 (-0.01, 0.01) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.05 (0.04, 0.06) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.08 (-0.16, 0.00) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.03 (-0.06, 0.01) (coef_ did not converge)
Neural Network (MLP): 8703.97 (8160.82, 9247.12)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: 0
p value: 0.88
Add Unobserved Common Cause: 0.01
Placebo Treatment Refuter: 0.02
p value: 2
Data Subset Refuter: -0.005
p value: 0.86
Cold Months
Random Common Cause: 0
p value: 0.98
Add Unobserved Common Cause: 0.052
Placebo Treatment Refuter: 0.04
p value: 2
Data Subset Refuter: 0.04
p value: 0.98
Warm Months
Random Common Cause: 0
p value: 0.96
Add Unobserved Common Cause: 0.02
Placebo Treatment Refuter: 0.02
p value: 2
Data Subset Refuter: 0
p value: 0.96

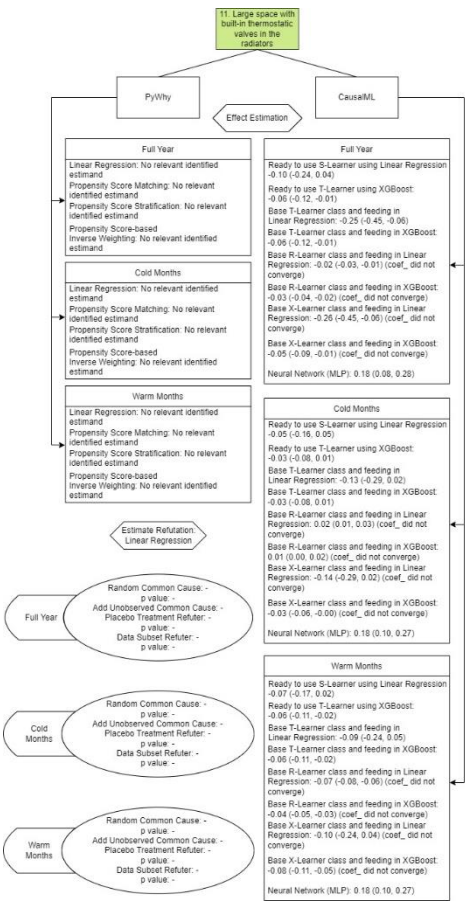

9. High walls of the rooms
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: -0.04
Propensity Score Matching: -0.4
Propensity Score Stratification: -0.3
Propensity Score-based Inverse Weighting: -0.24
Cold Months
Linear Regression: -0.06
Propensity Score Matching: -0.46
Propensity Score Stratification: -0.32
Propensity Score-based Inverse Weighting: -0.19
Warm Months
Linear Regression: -0.05
Propensity Score Matching: -0.49
Propensity Score Stratification: -0.35
Propensity Score-based Inverse Weighting: -0.25
Full Year
Ready to use S-Learner using Linear Regression 0.04 (-0.16, 0.25)
Ready to use T-Learner using XGBoost: -0.08 (-0.14, -0.02)
Base T-Learner class and feeding in Linear Regression: 0.17 (0.04, 0.30)
Base T-Learner class and feeding in XGBoost: -0.08 (-0.14, -0.02)
Base R-Learner class and feeding in Linear Regression: -0.00 (-0.02, 0.01) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.07 (-0.08, -0.05) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.18 (0.05, 0.30) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.01 (-0.05, 0.02) (coef_ did not converge)
Neural Network (MLP): 1.30 (0.86, 1.74)
Cold Months
Ready to use S-Learner using Linear Regression 0.01 (-0.20, 0.23)
Ready to use T-Learner using XGBoost: -0.04 (-0.10, 0.02)
Base T-Learner class and feeding in Linear Regression: 0.19 (0.03, 0.35)
Base T-Learner class and feeding in XGBoost: -0.04 (-0.10, 0.02)
Base R-Learner class and feeding in Linear Regression: 0.04 (0.03, 0.06)
Base R-Learner class and feeding in XGBoost: -0.02 (-0.03, -0.01) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.20 (0.04, 0.36) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.00 (-0.04, 0.03)
Neural Network (MLP): 1.29 (0.85, 1.73)
Warm Months
Ready to use S-Learner using Linear Regression 0.08 (-0.12, 0.28)
Ready to use T-Learner using XGBoost: -0.04 (-0.11, 0.02)
Base T-Learner class and feeding in Linear Regression: 0.13 (0.01, 0.25)
Base T-Learner class and feeding in XGBoost: -0.04 (-0.11, 0.02)
Base R-Learner class and feeding in Linear Regression: -0.03 (-0.04, -0.02) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.02 (-0.03, -0.01) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.14 (0.02, 0.26) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.03 (-0.07, 0.01) (coef_ did not converge)
Neural Network (MLP): 1.36 (0.90, 1.82)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: 0.01
p value: 0.92
Add Unobserved Common Cause: 0.02
Placebo Treatment Refuter: 0.05
p value: 2
Data Subset Refuter: 0.01
p value: 0.91
Cold Months
Random Common Cause: 0.002
p value: 0.76
Add Unobserved Common Cause: 0.07
Placebo Treatment Refuter: 0.006
p value: 2
Data Subset Refuter: 0.003
p value: 0.8
Warm Months
Random Common Cause: 0.01
p value: 0.92
Add Unobserved Common Cause: -0.02
Placebo Treatment Refuter: 0.05
p value: 2
Data Subset Refuter: 0.01
p value: 0.91

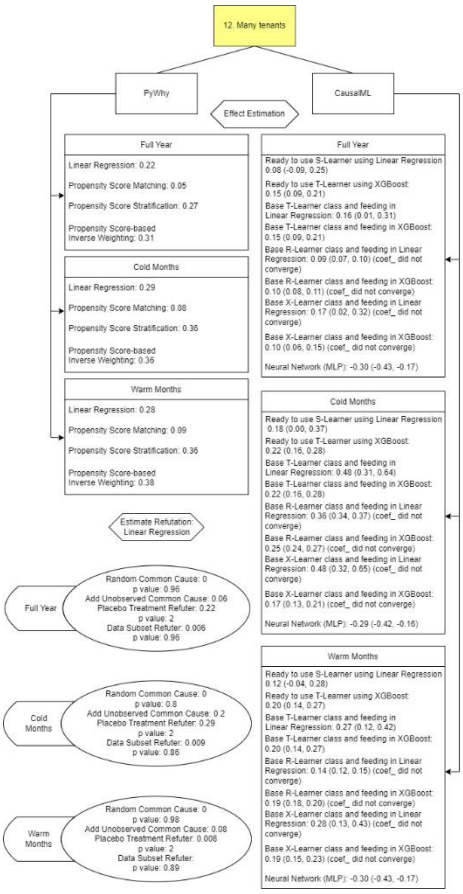

9. High walls of the rooms
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: -0.015
Propensity Score Matching: 0.12
Propensity Score Stratification: -0.04
Propensity Score-based Inverse Weighting: -0.07
Cold Months
Linear Regression: -0.05
Propensity Score Matching: 0.14
Propensity Score Stratification: -0.01
Propensity Score-based Inverse Weighting: -0.07
Warm Months
Linear Regression: -0.06
Propensity Score Matching: 0.14
Propensity Score Stratification: -0.03
Propensity Score-based Inverse Weighting: -0.07
Full Year
Ready to use S-Learner using Linear Regression -0.01 (-0.22, 0.21)
Ready to use T-Learner using XGBoost: 0.16 (0.10, 0.22)
Base T-Learner class and feeding in Linear Regression: 0.10 (-0.03, 0.23)
Base T-Learner class and feeding in XGBoost: 0.16 (0.10, 0.22)
Base R-Learner class and feeding in Linear Regression: 0.01 (0.00, 0.02) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.01 (0.00, 0.02) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.10 (-0.03, 0.23) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.15 (0.11, 0.19) (coef_ did not converge)
Neural Network (MLP): 0.30 (0.20, 0.39)
Cold Months
Ready to use S-Learner using Linear Regression 0.02 (-0.11, 0.16)
Ready to use T-Learner using XGBoost: 0.15 (0.10, 0.21)
Base T-Learner class and feeding in Linear Regression: 0.11 (0.02, 0.20)
Base T-Learner class and feeding in XGBoost: 0.15 (0.10, 0.21)
Base R-Learner class and feeding in Linear Regression: 0.03 (0.02, 0.04) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.02 (-0.03, -0.01) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.10 (0.01, 0.19) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.15 (0.11, 0.19) (coef_ did not converge)
Neural Network (MLP): -14942.02 (-16543.01, -13341.03)
Warm Months
Ready to use S-Learner using Linear Regression -0.04 (-0.19, 0.12)
Ready to use T-Learner using XGBoost: 0.14 (0.09, 0.20)
Base T-Learner class and feeding in Linear Regression: 0.10 (0.01, 0.19)
Base T-Learner class and feeding in XGBoost: 0.14 (0.09, 0.20)
Base R-Learner class and feeding in Linear Regression: 0.02 (0.01, 0.03) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.01 (-0.02, -0.00) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.09 (0.00, 0.19) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.16 (0.12, 0.20) (coef_ did not converge)
Neural Network (MLP): -14941.08 (-16541.95, -13340.21)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: 0.001
p value: 0.86
Add Unobserved Common Cause: 0.06
Placebo Treatment Refuter: 0.015
p value: 2
Data Subset Refuter: 0.005
p value: 0.96
Cold Months
Random Common Cause: 0.001
p value: 0.8
Add Unobserved Common Cause: 0.01
Placebo Treatment Refuter: 0.05
p value: 2
Data Subset Refuter: 0.001
p value: 0.96
Warm Months
Random Common Cause: 0
p value: 0.96
Add Unobserved Common Cause: 0.03
Placebo Treatment Refuter: 0.06
p value: 2
Data Subset Refuter: 0.001
p value: 0.96

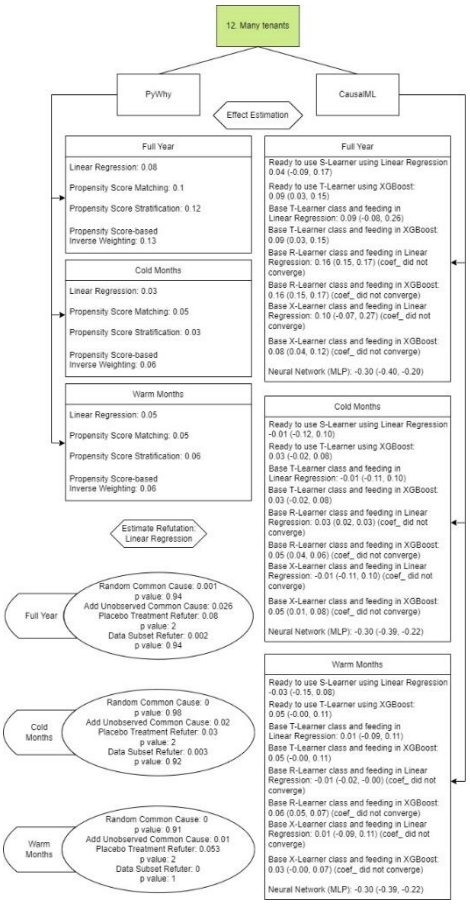

10. A large space (volume) for drying clothes
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: No relevant identified estimand
Propensity Score Matching: No relevant identified estimand
Propensity Score Stratification: No relevant identified estimand
Propensity Score-based Inverse Weighting: No relevant identified estimand
Cold Months
Linear Regression: No relevant identified estimand
Propensity Score Matching: No relevant identified estimand
Propensity Score Stratification: No relevant identified estimand
Propensity Score-based Inverse Weighting: No relevant identified estimand
Warm Months
Linear Regression: No relevant identified estimand
Propensity Score Matching: No relevant identified estimand
Propensity Score Stratification: No relevant identified estimand
Propensity Score-based Inverse Weighting: No relevant identified estimand
Full Year
Ready to use S-Learner using Linear Regression -0.15 (-0.29, -0.02)
Ready to use T-Learner using XGBoost: -0.08 (-0.14, -0.01)
Base T-Learner class and feeding in Linear Regression: -0.12 (-0.26, 0.03)
Base T-Learner class and feeding in XGBoost: -0.08 (-0.14, -0.01)
Base R-Learner class and feeding in Linear Regression: -0.17 (-0.18, -0.16) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.08 (-0.09, -0.07) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.12 (-0.27, 0.03) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.06 (-0.09, -0.02) (coef_ did not converge)
Neural Network (MLP): 0.35 (0.22, 0.48)
Cold Months
Ready to use S-Learner using Linear Regression -0.15 (-0.30, -0.01)
Ready to use T-Learner using XGBoost: -0.11 (-0.18, -0.05)
Base T-Learner class and feeding in Linear Regression: -0.19 (-0.36, -0.02)
Base T-Learner class and feeding in XGBoost: -0.11 (-0.18, -0.05)
Base R-Learner class and feeding in Linear Regression: -0.21 (-0.22, -0.19) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.16 (-0.17, -0.15) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.20 (-0.37, -0.02) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.08 (-0.12, -0.05) (coef_ did not converge)
Neural Network (MLP): 0.35 (0.22, 0.47)
Warm Months
Ready to use S-Learner using Linear Regression -0.20 (-0.33, -0.07)
Ready to use T-Learner using XGBoost: -0.10 (-0.15, -0.04)
Base T-Learner class and feeding in Linear Regression: -0.16 (-0.30, -0.03)
Base T-Learner class and feeding in XGBoost: -0.10 (-0.15, -0.04)
Base R-Learner class and feeding in Linear Regression: -0.20 (-0.21, -0.18) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.17 (-0.18, -0.16) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.16 (-0.30, -0.03) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.09 (-0.12, -0.06) (coef_ did not converge)
Neural Network (MLP): 0.35 (0.22, 0.48)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: - p value: -
Add Unobserved Common Cause: - Placebo Treatment Refuter: - p value: -
Data Subset Refuter: - p value: -
Cold Months
Random Common Cause: - p value: -
Add Unobserved Common Cause: - Placebo Treatment Refuter: - p value: -
Data Subset Refuter: - p value: -
Warm Months
Random Common Cause: - p value: -
Add Unobserved Common Cause: - Placebo Treatment Refuter: - p value: -
Data Subset Refuter: - p value: -

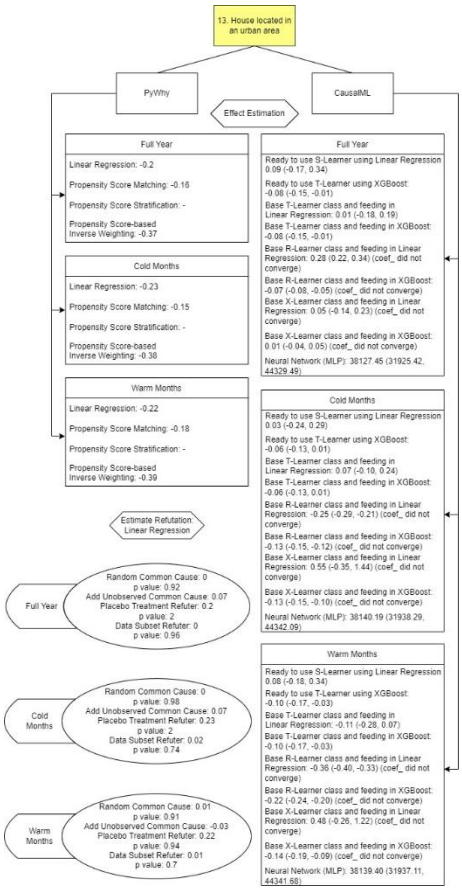

10. A large space (volume) for drying clothes
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: No relevant identified estimand
Propensity Score Matching: No relevant identified estimand
Propensity Score Stratification: No relevant identified estimand
Propensity Score-based Inverse Weighting: No relevant identified estimand
Cold Months
Linear Regression: No relevant identified estimand
Propensity Score Matching: No relevant identified estimand
Propensity Score Stratification: No relevant identified estimand
Propensity Score-based Inverse Weighting: No relevant identified estimand
Warm Months
Linear Regression: No relevant identified estimand
Propensity Score Matching: No relevant identified estimand
Propensity Score Stratification: No relevant identified estimand
Propensity Score-based Inverse Weighting: No relevant identified estimand
Full Year
Ready to use S-Learner using Linear Regression 0.11 (-0.01, 0.24)
Ready to use T-Learner using XGBoost: 0.09 (0.03, 0.14)
Base T-Learner class and feeding in Linear Regression: 0.07 (-0.06, 0.21)
Base T-Learner class and feeding in XGBoost: 0.09 (0.03, 0.14)
Base R-Learner class and feeding in Linear Regression: 0.01 (0.00, 0.02) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.11 (0.10, 0.12) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.08 (-0.05, 0.21) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.09 (0.05, 0.13) (coef_ did not converge)
Neural Network (MLP): 1942.82 (1691.82, 2193.82)
Cold Months
Ready to use S-Learner using Linear Regression 0.09 (-0.02, 0.20)
Ready to use T-Learner using XGBoost: 0.03 (0.03, 0.13)
Base T-Learner class and feeding in Linear Regression: 0.04 (-0.05, 0.14)
Base T-Learner class and feeding in XGBoost: 0.03 (0.03, 0.13)
Base R-Learner class and feeding in Linear Regression: 0.07 (0.06, 0.08) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.10 (0.09, 0.11) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.05 (-0.05, 0.14) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.08 (0.05, 0.11) (coef_ did not converge)
Neural Network (MLP): 1945.10 (1693.95, 2196.26)
Warm Months
Ready to use S-Learner using Linear Regression 0.09 (-0.02, 0.19)
Ready to use T-Learner using XGBoost: 0.01 (-0.04, 0.06)
Base T-Learner class and feeding in Linear Regression: 0.04 (-0.05, 0.13)
Base T-Learner class and feeding in XGBoost: 0.01 (-0.04, 0.06)
Base R-Learner class and feeding in Linear Regression: 0.04 (0.03, 0.05) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.06 (-0.07, -0.05) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.04 (-0.05, 0.13) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.01 (-0.02, 0.05) (coef_ did not converge)
Neural Network (MLP): 1941.10 (1690.28, 2191.91)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: - p value: -
Add Unobserved Common Cause: - Placebo Treatment Refuter: - p value: -
Data Subset Refuter: - p value: -
Cold Months
Random Common Cause: - p value: -
Add Unobserved Common Cause: - Placebo Treatment Refuter: - p value: -
Data Subset Refuter: - p value: -
Warm Months
Random Common Cause: - p value: -
Add Unobserved Common Cause: - Placebo Treatment Refuter: - p value: -
Data Subset Refuter: - p value: -

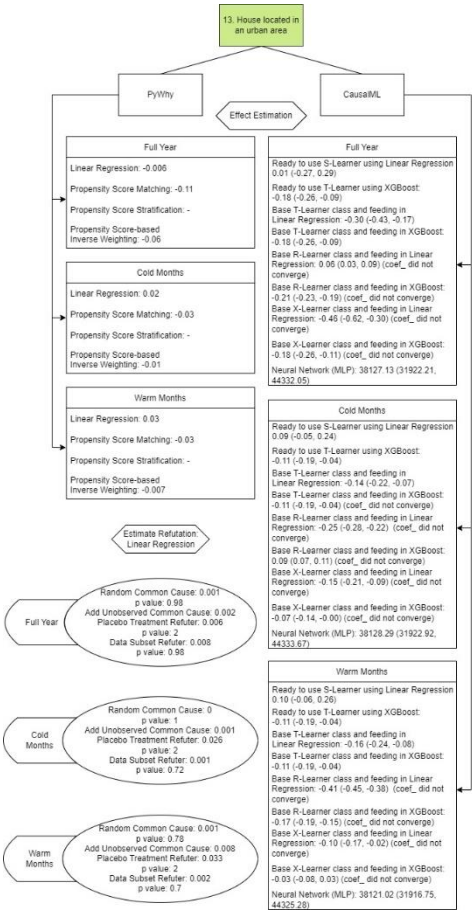

11. Large space with built-in thermostatic valves in the radiators
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: No relevant identified estimand
Propensity Score Matching: No relevant identified estimand
Propensity Score Stratification: No relevant identified estimand
Propensity Score-based Inverse Weighting: No relevant identified estimand
Cold Months
Linear Regression: No relevant identified estimand
Propensity Score Matching: No relevant identified estimand
Propensity Score Stratification: No relevant identified estimand
Propensity Score-based Inverse Weighting: No relevant identified estimand
Warm Months
Linear Regression: No relevant identified estimand
Propensity Score Matching: No relevant identified estimand
Propensity Score Stratification: No relevant identified estimand
Propensity Score-based Inverse Weighting: No relevant identified estimand
Full Year
Ready to use S-Learner using Linear Regression 0.07 (-0.06, 0.20)
Ready to use T-Learner using XGBoost: -0.05 (-0.11, 0.01)
Base T-Learner class and feeding in Linear Regression: -0.04 (-0.19, 0.11)
Base T-Learner class and feeding in XGBoost: -0.05 (-0.11, 0.01)
Base R-Learner class and feeding in Linear Regression: 0.03 (0.01, 0.04) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.05 (-0.06, -0.04) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.04 (-0.19, 0.11) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.05 (-0.08, -0.01) (coef_ did not converge)
Neural Network (MLP): 0.18 (0.04, 0.31)
Cold Months
Ready to use S-Learner using Linear Regression 0.08 (-0.07, 0.23)
Ready to use T-Learner using XGBoost: -0.00 (-0.07, 0.07)
Base T-Learner class and feeding in Linear Regression: -0.11 (-0.31, 0.08)
Base T-Learner class and feeding in XGBoost: -0.00 (-0.07, 0.07)
Base R-Learner class and feeding in Linear Regression: -0.12 (-0.15, -0.09) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.01 (-0.01, 0.02) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.11 (-0.31, 0.09) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.01 (-0.05, 0.04) (coef_ did not converge)
Neural Network (MLP): 0.18 (0.05, 0.32)
Warm Months
Ready to use S-Learner using Linear Regression 0.08 (-0.05, 0.21)
Ready to use T-Learner using XGBoost: -0.00 (-0.07, 0.06)
Base T-Learner class and feeding in Linear Regression: -0.09 (-0.30, 0.13)
Base T-Learner class and feeding in XGBoost: -0.00 (-0.07, 0.06)
Base R-Learner class and feeding in Linear Regression: 0.03 (0.02, 0.04) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.02 (0.00, 0.03) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.09 (-0.30, 0.13) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.01 (-0.05, 0.02) (coef_ did not converge)
Neural Network (MLP): 0.02 (-0.12, 0.15)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: - p value: -
Add Unobserved Common Cause: - Placebo Treatment Refuter: - p value: -
Data Subset Refuter: - p value: -
Cold Months
Random Common Cause: - p value: -
Add Unobserved Common Cause: - Placebo Treatment Refuter: - p value: -
Data Subset Refuter: - p value: -
Warm Months
Random Common Cause: - p value: -
Add Unobserved Common Cause: - Placebo Treatment Refuter: - p value: -
Data Subset Refuter: - p value: -

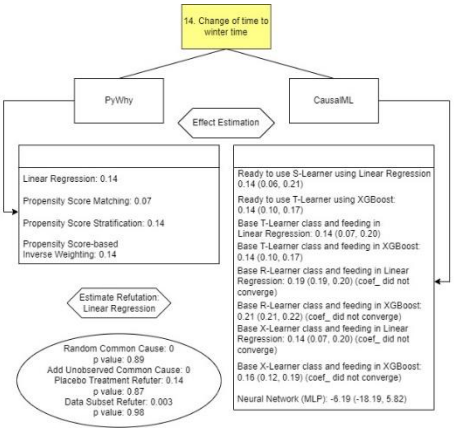

11. Large space with built-in thermostatic valves in the radiators
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: No relevant identified estimand
Propensity Score Matching: No relevant identified estimand
Propensity Score Stratification: No relevant identified estimand
Propensity Score-based Inverse Weighting: No relevant identified estimand
Cold Months
Linear Regression: No relevant identified estimand
Propensity Score Matching: No relevant identified estimand
Propensity Score Stratification: No relevant identified estimand
Propensity Score-based Inverse Weighting: No relevant identified estimand
Warm Months
Linear Regression: No relevant identified estimand
Propensity Score Matching: No relevant identified estimand
Propensity Score Stratification: No relevant identified estimand
Propensity Score-based Inverse Weighting: No relevant identified estimand
Full Year
Ready to use S-Learner using Linear Regression -0.10 (-0.24, 0.04)
Ready to use T-Learner using XGBoost: -0.06 (-0.12, -0.01)
Base T-Learner class and feeding in Linear Regression: -0.25 (-0.45, -0.06)
Base T-Learner class and feeding in XGBoost: -0.06 (-0.12, -0.01)
Base R-Learner class and feeding in Linear Regression: -0.02 (-0.03, -0.01) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.03 (-0.04, -0.02) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.26 (-0.45, -0.06) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.05 (-0.09, -0.01) (coef_ did not converge)
Neural Network (MLP): 0.18 (0.08, 0.28)
Cold Months
Ready to use S-Learner using Linear Regression -0.05 (-0.16, 0.05)
Ready to use T-Learner using XGBoost: -0.03 (-0.08, 0.01)
Base T-Learner class and feeding in Linear Regression: -0.13 (-0.29, 0.02)
Base T-Learner class and feeding in XGBoost: -0.03 (-0.08, 0.01)
Base R-Learner class and feeding in Linear Regression: 0.02 (0.01, 0.03) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.01 (0.00, 0.02) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.14 (-0.29, 0.02) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.03 (-0.06, -0.00) (coef_ did not converge)
Neural Network (MLP): 0.18 (0.10, 0.27)
Warm Months
Ready to use S-Learner using Linear Regression -0.07 (-0.17, 0.02)
Ready to use T-Learner using XGBoost: -0.06 (-0.11, -0.02)
Base T-Learner class and feeding in Linear Regression: -0.09 (-0.24, 0.05)
Base T-Learner class and feeding in XGBoost: -0.06 (-0.11, -0.02)
Base R-Learner class and feeding in Linear Regression: -0.07 (-0.08, -0.06) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.04 (-0.05, -0.03) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.10 (-0.24, 0.04) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.08 (-0.11, -0.05) (coef_ did not converge)
Neural Network (MLP): 0.18 (0.10, 0.27)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: - p value: -
Add Unobserved Common Cause: - Placebo Treatment Refuter: - p value: -
Data Subset Refuter: - p value: -
Cold Months
Random Common Cause: - p value: -
Add Unobserved Common Cause: - Placebo Treatment Refuter: - p value: -
Data Subset Refuter: - p value: -
Warm Months
Random Common Cause: - p value: -
Add Unobserved Common Cause: - Placebo Treatment Refuter: - p value: -
Data Subset Refuter: - p value: -

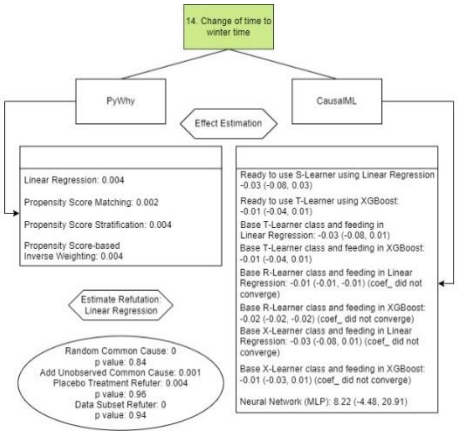

12. Many tenants
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: 0.22
Propensity Score Matching: 0.05
Propensity Score Stratification: 0.27
Propensity Score-based Inverse Weighting: 0.31
Cold Months
Linear Regression: 0.29
Propensity Score Matching: 0.08
Propensity Score Stratification: 0.36
Propensity Score-based Inverse Weighting: 0.36
Warm Months
Linear Regression: 0.28
Propensity Score Matching: 0.09
Propensity Score Stratification: 0.36
Propensity Score-based Inverse Weighting: 0.38
Full Year
Ready to use S-Learner using Linear Regression 0.08 (-0.09, 0.25)
Ready to use T-Learner using XGBoost: 0.15 (0.09, 0.21)
Base T-Learner class and feeding in Linear Regression: 0.16 (0.01, 0.31)
Base T-Learner class and feeding in XGBoost: 0.15 (0.09, 0.21)
Base R-Learner class and feeding in Linear Regression: 0.09 (0.07, 0.10) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.10 (0.08, 0.11) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.17 (0.02, 0.32) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.10 (0.05, 0.15) (coef_ did not converge)
Neural Network (MLP): -0.30 (-0.43, -0.17)
Cold Months
Ready to use S-Learner using Linear Regression 0.18 (0.00, 0.37)
Ready to use T-Learner using XGBoost: 0.22 (0.16, 0.28)
Base T-Learner class and feeding in Linear Regression: 0.48 (0.31, 0.64)
Base T-Learner class and feeding in XGBoost: 0.22 (0.16, 0.28)
Base R-Learner class and feeding in Linear Regression: 0.36 (0.34, 0.37) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.25 (0.24, 0.27) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.48 (0.32, 0.65) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.17 (0.13, 0.21) (coef_ did not converge)
Neural Network (MLP): -0.29 (-0.42, -0.16)
Warm Months
Ready to use S-Learner using Linear Regression 0.12 (-0.04, 0.28)
Ready to use T-Learner using XGBoost: 0.20 (0.14, 0.27)
Base T-Learner class and feeding in Linear Regression: 0.27 (0.12, 0.42)
Base T-Learner class and feeding in XGBoost: 0.20 (0.14, 0.27)
Base R-Learner class and feeding in Linear Regression: 0.14 (0.12, 0.15) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.19 (0.18, 0.20) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.29 (0.13, 0.43) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.19 (0.15, 0.23) (coef_ did not converge)
Neural Network (MLP): -0.30 (-0.43, -0.17)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: 0
p value: 0.96
Add Unobserved Common Cause: 0.06
Placebo Treatment Refuter: 0.22
p value: 2
Data Subset Refuter: 0.006
p value: 0.96
Cold Months
Random Common Cause: 0
p value: 0.8
Add Unobserved Common Cause: 0.2
Placebo Treatment Refuter: 0.29
p value: 2
Data Subset Refuter: 0.009
p value: 0.86
Warm Months
Random Common Cause: 0
p value: 0.98
Add Unobserved Common Cause: 0.08
Placebo Treatment Refuter: 0.008
p value: 2
Data Subset Refuter:
p value: 0.89

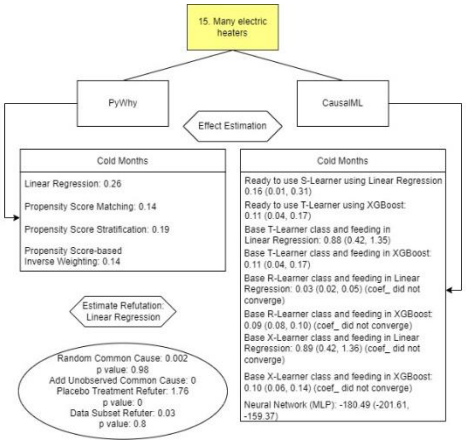

12. Many tenants
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: 0.08
Propensity Score Matching: 0.1
Propensity Score Stratification: 0.12
Propensity Score-based Inverse Weighting: 0.13
Cold Months
Linear Regression: 0.03
Propensity Score Matching: 0.05
Propensity Score Stratification: 0.03
Propensity Score-based Inverse Weighting: 0.06
Warm Months
Linear Regression: 0.05
Propensity Score Matching: 0.05
Propensity Score Stratification: 0.06
Propensity Score-based Inverse Weighting: 0.06
Full Year
Ready to use S-Learner using Linear Regression 0.04 (-0.09, 0.17)
Ready to use T-Learner using XGBoost: 0.09 (0.03, 0.15)
Base T-Learner class and feeding in Linear Regression: 0.09 (-0.08, 0.26)
Base T-Learner class and feeding in XGBoost: 0.09 (0.03, 0.15)
Base R-Learner class and feeding in Linear Regression: 0.16 (0.15, 0.17) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.16 (0.15, 0.17) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.10 (-0.07, 0.27) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.08 (0.04, 0.12) (coef_ did not converge)
Neural Network (MLP): -0.30 (-0.40, -0.20)
Cold Months
Ready to use S-Learner using Linear Regression -0.01 (-0.12, 0.10)
Ready to use T-Learner using XGBoost: 0.03 (-0.02, 0.08)
Base T-Learner class and feeding in Linear Regression: -0.01 (-0.11, 0.10)
Base T-Learner class and feeding in XGBoost: 0.03 (-0.02, 0.08)
Base R-Learner class and feeding in Linear Regression: 0.03 (0.02, 0.03) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.05 (0.04, 0.06) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.01 (-0.11, 0.10) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.05 (0.01, 0.08) (coef_ did not converge)
Neural Network (MLP): -0.30 (-0.39, -0.22)
Warm Months
Ready to use S-Learner using Linear Regression -0.03 (-0.15, 0.08)
Ready to use T-Learner using XGBoost: 0.05 (-0.00, 0.11)
Base T-Learner class and feeding in Linear Regression: 0.01 (-0.09, 0.11)
Base T-Learner class and feeding in XGBoost: 0.05 (-0.00, 0.11)
Base R-Learner class and feeding in Linear Regression: -0.01 (-0.02, -0.00) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.06 (0.05, 0.07) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.01 (-0.09, 0.11) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.03 (-0.00, 0.07) (coef_ did not converge)
Neural Network (MLP): -0.30 (-0.39, -0.22)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: 0.001
p value: 0.94
Add Unobserved Common Cause: 0.026
Placebo Treatment Refuter: 0.08
p value: 2
Data Subset Refuter: 0.002
p value: 0.94
Cold Months
Random Common Cause: 0
p value: 0.98
Add Unobserved Common Cause: 0.02
Placebo Treatment Refuter: 0.03
p value: 2
Data Subset Refuter: 0.003
p value: 0.92
Warm Months
Random Common Cause: 0
p value: 0.91
Add Unobserved Common Cause: 0.01
Placebo Treatment Refuter: 0.053
p value: 2
Data Subset Refuter: 0
p value: 1

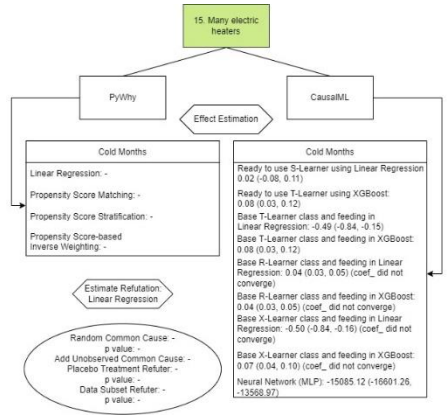

13. House located in an urban area
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: -0.2
Propensity Score Matching: -0.16
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -0.37
Cold Months
Linear Regression: -0.23
Propensity Score Matching: -0.15
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -0.38
Warm Months
Linear Regression: -0.22
Propensity Score Matching: -0.18
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -0.39
Full Year
Ready to use S-Learner using Linear Regression 0.09 (-0.17, 0.34)
Ready to use T-Learner using XGBoost: -0.08 (-0.15, -0.01)
Base T-Learner class and feeding in Linear Regression: 0.01 (-0.18, 0.19)
Base T-Learner class and feeding in XGBoost: -0.08 (-0.15, -0.01)
Base R-Learner class and feeding in Linear Regression: 0.28 (0.22, 0.34) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.07 (-0.08, -0.05) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.05 (-0.14, 0.23) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.01 (-0.04, 0.05) (coef_ did not converge)
Neural Network (MLP): 38127.45 (31925.42, 44329.49)
Cold Months
Ready to use S-Learner using Linear Regression 0.03 (-0.24, 0.29)
Ready to use T-Learner using XGBoost: -0.06 (-0.13, 0.01)
Base T-Learner class and feeding in Linear Regression: 0.07 (-0.10, 0.24)
Base T-Learner class and feeding in XGBoost: -0.06 (-0.13, 0.01)
Base R-Learner class and feeding in Linear Regression: -0.25 (-0.29, -0.21) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.13 (-0.15, -0.12) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.55 (-0.35, 1.44) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.13 (-0.15, -0.10) (coef_ did not converge)
Neural Network (MLP): 38140.19 (31938.29, 44342.09)
Warm Months
Ready to use S-Learner using Linear Regression 0.08 (-0.18, 0.34)
Ready to use T-Learner using XGBoost: -0.10 (-0.17, -0.03)
Base T-Learner class and feeding in Linear Regression: -0.11 (-0.28, 0.07)
Base T-Learner class and feeding in XGBoost: -0.10 (-0.17, -0.03)
Base R-Learner class and feeding in Linear Regression: -0.36 (-0.40, -0.33) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.22 (-0.24, -0.20) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.48 (-0.26, 1.22) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.14 (-0.19, -0.09) (coef_ did not converge)
Neural Network (MLP): 38139.40 (31937.11, 44341.68)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: 0
p value: 0.92
Add Unobserved Common Cause: 0.07
Placebo Treatment Refuter: 0.2
p value: 2
Data Subset Refuter: 0
p value: 0.96
Cold Months
Random Common Cause: 0
p value: 0.98
Add Unobserved Common Cause: 0.07
Placebo Treatment Refuter: 0.23
p value: 2
Data Subset Refuter: 0.02
p value: 0.74
Warm Months
Random Common Cause: 0.01
p value: 0.91
Add Unobserved Common Cause: -0.03
Placebo Treatment Refuter: 0.22
p value: 0.94
Data Subset Refuter: 0.01
p value: 0.7

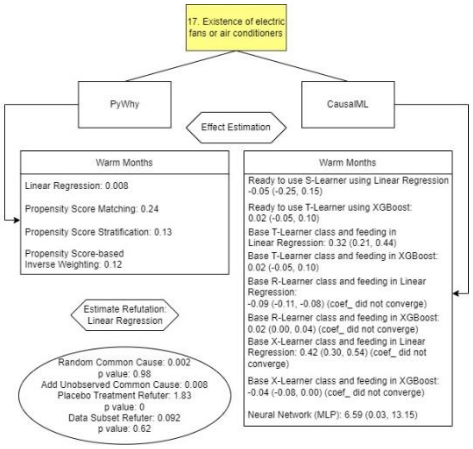

13. House located in an urban area
PyWhy
CausalML
Effect Estimation
Full Year
Linear Regression: -0.006
Propensity Score Matching: -0.11
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -0.06
Cold Months
Linear Regression: 0.02
Propensity Score Matching: -0.03
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -0.01
Warm Months
Linear Regression: 0.03
Propensity Score Matching: -0.03
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -0.007
Full Year
Ready to use S-Learner using Linear Regression 0.01 (-0.27, 0.29)
Ready to use T-Learner using XGBoost: -0.18 (-0.26, -0.09)
Base T-Learner class and feeding in Linear Regression: -0.30 (-0.43, -0.17)
Base T-Learner class and feeding in XGBoost: -0.18 (-0.26, -0.09)
Base R-Learner class and feeding in Linear Regression: 0.06 (0.03, 0.09) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.21 (-0.23, -0.19) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.48 (-0.62, -0.30) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.18 (-0.26, -0.11) (coef_ did not converge)
Neural Network (MLP): 38127.13 (31922.21, 44332.05)
Cold Months
Ready to use S-Learner using Linear Regression 0.09 (-0.05, 0.24)
Ready to use T-Learner using XGBoost: -0.11 (-0.19, -0.04)
Base T-Learner class and feeding in Linear Regression: -0.14 (-0.22, -0.07)
Base T-Learner class and feeding in XGBoost: -0.11 (-0.19, -0.04) (coef_ did not converge)
Base R-Learner class and feeding in Linear Regression: -0.25 (-0.28, -0.22) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.09 (0.07, 0.11) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.15 (-0.21, -0.09) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.07 (-0.14, -0.00) (coef_ did not converge)
Neural Network (MLP): 38128.29 (31922.92, 44333.67)
Warm Months
Ready to use S-Learner using Linear Regression 0.10 (-0.06, 0.26)
Ready to use T-Learner using XGBoost: -0.11 (-0.19, -0.04)
Base T-Learner class and feeding in Linear Regression: -0.16 (-0.24, -0.08)
Base T-Learner class and feeding in XGBoost: -0.11 (-0.19, -0.04)
Base R-Learner class and feeding in Linear Regression: -0.41 (-0.45, -0.38) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.17 (-0.19, -0.15) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.10 (-0.17, -0.02) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.03 (-0.08, 0.03) (coef_ did not converge)
Neural Network (MLP): 38121.02 (31916.75, 44325.28)
Estimate Refutation: Linear Regression
Full Year
Random Common Cause: 0.001
p value: 0.98
Add Unobserved Common Cause: 0.002
Placebo Treatment Refuter: 0.006
p value: 2
Data Subset Refuter: 0.008
p value: 0.98
Cold Months
Random Common Cause: 0
p value: 1
Add Unobserved Common Cause: 0.001
Placebo Treatment Refuter: 0.026
p value: 2
Data Subset Refuter: 0.001
p value: 0.72
Warm Months
Random Common Cause: 0.001
p value: 0.78
Add Unobserved Common Cause: 0.008
Placebo Treatment Refuter: 0.033
p value: 2
Data Subset Refuter: 0.002
p value: 0.7

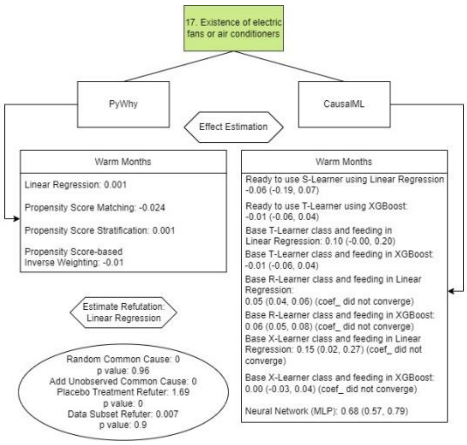

14. Change of time to winter time
PyWhy
CausalML
Effect Estimation
Linear Regression: 0.14
Propensity Score Matching: 0.07
Propensity Score Stratification: 0.14
Propensity Score-based Inverse Weighting: 0.14
Ready to use S-Learner using Linear Regression 0.14 (0.06, 0.21)
Ready to use T-Learner using XGBoost: 0.14 (0.10, 0.17)
Base T-Learner class and feeding in Linear Regression: 0.14 (0.07, 0.20)
Base T-Learner class and feeding in XGBoost: 0.14 (0.10, 0.17)
Base R-Learner class and feeding in Linear Regression: 0.19 (0.19, 0.20) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.21 (0.21, 0.22) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.14 (0.07, 0.20) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.16 (0.12, 0.19) (coef_ did not converge)
Neural Network (MLP): -6.19 (-18.19, 5.82)
Estimate Refutation: Linear Regression
Random Common Cause: 0
p value: 0.89
Add Unobserved Common Cause: 0
Placebo Treatment Refuter: 0.14
p value: 0.87
Data Subset Refuter: 0.003
p value: 0.98

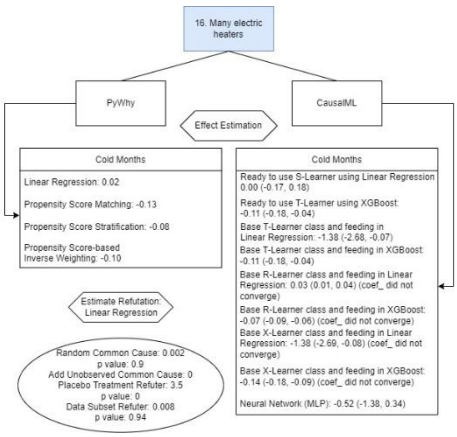

14. Change of time to winter time
PyWhy
CausalML
Effect Estimation
Linear Regression: 0.004
Propensity Score Matching: 0.002
Propensity Score Stratification: 0.004
Propensity Score-based Inverse Weighting: 0.004
Ready to use S-Learner using Linear Regression -0.03 (-0.08, 0.03)
Ready to use T-Learner using XGBoost: -0.01 (-0.04, 0.01)
Base T-Learner class and feeding in Linear Regression: -0.03 (-0.08, 0.01)
Base T-Learner class and feeding in XGBoost: -0.01 (-0.04, 0.01)
Base R-Learner class and feeding in Linear Regression: -0.01 (-0.01, -0.01) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.02 (-0.02, -0.02) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.03 (-0.08, 0.01) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.01 (-0.03, 0.01) (coef_ did not converge)
Neural Network (MLP): 8.22 (-4.48, 20.91)
Estimate Refutation: Linear Regression
Random Common Cause: 0
p value: 0.84
Add Unobserved Common Cause: 0.001
Placebo Treatment Refuter: 0.004
p value: 0.96
Data Subset Refuter: 0
p value: 0.94

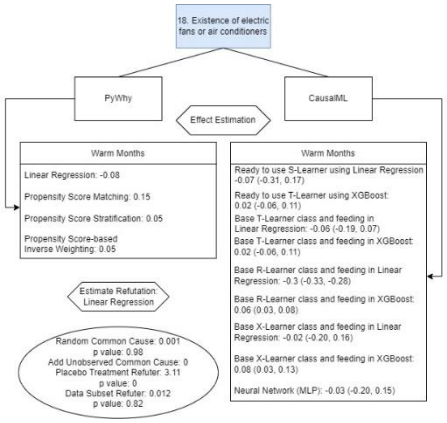

15. Many electric heaters
PyWhy
CausalML
Effect Estimation
Cold Months
Linear Regression: 0.26
Propensity Score Matching: 0.14
Propensity Score Stratification: 0.19
Propensity Score-based Inverse Weighting: 0.14
Cold Months
Ready to use S-Learner using Linear Regression 0.16 (0.01, 0.31)
Ready to use T-Learner using XGBoost: 0.11 (0.04, 0.17)
Base T-Learner class and feeding in Linear Regression: 0.88 (0.42, 1.35)
Base T-Learner class and feeding in XGBoost: 0.11 (0.04, 0.17)
Base R-Learner class and feeding in Linear Regression: 0.03 (0.02, 0.05) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.09 (0.08, 0.10) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.89 (0.42, 1.36) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.10 (0.06, 0.14) (coef_ did not converge)
Neural Network (MLP): -180.49 (-201.61, -159.37)
Estimate Refutation: Linear Regression
Random Common Cause: 0.002
p value: 0.98
Add Unobserved Common Cause: 0
Placebo Treatment Refuter: 1.76
p value: 0
Data Subset Refuter: 0.03
p value: 0.8


15. Many electric heaters
PyWhy
CausalML
Effect Estimation
Cold Months
Linear Regression: -
Propensity Score Matching: -
Propensity Score Stratification: -
Propensity Score-based Inverse Weighting: -
Cold Months
Ready to use S-Learner using Linear Regression 0.02 (-0.08, 0.11)
Ready to use T-Learner using XGBoost: 0.08 (0.03, 0.12)
Base T-Learner class and feeding in Linear Regression: -0.49 (-0.84, -0.15)
Base T-Learner class and feeding in XGBoost: 0.08 (0.03, 0.12)
Base R-Learner class and feeding in Linear Regression: 0.04 (0.03, 0.05) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.04 (0.03, 0.05) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -0.50 (-0.84, -0.16) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.07 (0.04, 0.10) (coef_ did not converge)
Neural Network (MLP): -15085.12 (-16601.26, -13568.97)
Estimate Refutation: Linear Regression
Random Common Cause: -
p value: -
Add Unobserved Common Cause: -
Placebo Treatment Refuter: -
p value: -
Data Subset Refuter: -
p value: -


17. Existence of electric fans or air conditioners
PyWhy
CausalML
Effect Estimation
Warm Months
Linear Regression: 0.008
Propensity Score Matching: 0.24
Propensity Score Stratification: 0.13
Propensity Score-based Inverse Weighting: 0.12
Warm Months
Ready to use S-Learner using Linear Regression -0.05 (-0.25, 0.15)
Ready to use T-Learner using XGBoost: 0.02 (-0.05, 0.10)
Base T-Learner class and feeding in Linear Regression: 0.32 (0.21, 0.44)
Base T-Learner class and feeding in XGBoost: 0.02 (-0.05, 0.10)
Base R-Learner class and feeding in Linear Regression: -0.09 (-0.11, -0.08) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.02 (0.00, 0.04) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.42 (0.30, 0.54) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.04 (-0.08, 0.00) (coef_ did not converge)
Neural Network (MLP): 6.59 (0.03, 13.15)
Estimate Refutation: Linear Regression
Random Common Cause: 0.002
p value: 0.98
Add Unobserved Common Cause: 0.008
Placebo Treatment Refuter: 1.83
p value: 0
Data Subset Refuter: 0.092
p value: 0.62


17. Existence of electric fans or air conditioners
PyWhy
CausalML
Effect Estimation
Warm Months
Linear Regression: 0.001
Propensity Score Matching: -0.024
Propensity Score Stratification: 0.001
Propensity Score-based Inverse Weighting: -0.01
Warm Months
Ready to use S-Learner using Linear Regression -0.06 (-0.19, 0.07)
Ready to use T-Learner using XGBoost: -0.01 (-0.06, 0.04)
Base T-Learner class and feeding in Linear Regression: 0.10 (-0.00, 0.20)
Base T-Learner class and feeding in XGBoost: -0.01 (-0.06, 0.04)
Base R-Learner class and feeding in Linear Regression: 0.05 (0.04, 0.06) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: 0.06 (0.05, 0.08) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: 0.15 (0.02, 0.27) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: 0.00 (-0.03, 0.04) (coef_ did not converge)
Neural Network (MLP): 0.68 (0.57, 0.79)
Estimate Refutation: Linear Regression
Random Common Cause: 0
p value: 0.96
Add Unobserved Common Cause: 0
Placebo Treatment Refuter: 1.69
p value: 0
Data Subset Refuter: 0.007
p value: 0.9


16. Many electric heaters
PyWhy
CausalML
Effect Estimation
Cold Months
Linear Regression: 0.02
Propensity Score Matching: -0.13
Propensity Score Stratification: -0.08
Propensity Score-based Inverse Weighting: -0.10
Cold Months
Ready to use S-Learner using Linear Regression 0.00 (-0.17, 0.18)
Ready to use T-Learner using XGBoost: -0.11 (-0.18, -0.04)
Base T-Learner class and feeding in Linear Regression: -1.38 (-2.68, -0.07)
Base T-Learner class and feeding in XGBoost: -0.11 (-0.18, -0.04)
Base R-Learner class and feeding in Linear Regression: 0.03 (0.01, 0.04) (coef_ did not converge)
Base R-Learner class and feeding in XGBoost: -0.07 (-0.09, -0.06) (coef_ did not converge)
Base X-Learner class and feeding in Linear Regression: -1.38 (-2.69, -0.08) (coef_ did not converge)
Base X-Learner class and feeding in XGBoost: -0.14 (-0.18, -0.09) (coef_ did not converge)
Neural Network (MLP): -0.52 (-1.38, 0.34)
Estimate Refutation: Linear Regression
Random Common Cause: 0.002
p value: 0.9
Add Unobserved Common Cause: 0
Placebo Treatment Refuter: 3.5
p value: 0
Data Subset Refuter: 0.008
p value: 0.94


18. Existence of electric fans or air conditioners
PyWhy
CausalML
Effect Estimation
Warm Months
Linear Regression: -0.08
Propensity Score Matching: 0.15
Propensity Score Stratification: 0.05
Propensity Score-based Inverse Weighting: 0.05
Warm Months
Ready to use S-Learner using Linear Regression -0.07 (-0.31, 0.17)
Ready to use T-Learner using XGBoost: 0.02 (-0.06, 0.11)
Base T-Learner class and feeding in Linear Regression: -0.06 (-0.19, 0.07)
Base T-Learner class and feeding in XGBoost: 0.02 (-0.06, 0.11)
Base R-Learner class and feeding in Linear Regression: -0.3 (-0.33, -0.28)
Base R-Learner class and feeding in XGBoost: 0.06 (0.03, 0.08)
Base X-Learner class and feeding in Linear Regression: -0.02 (-0.20, 0.16)
Base X-Learner class and feeding in XGBoost: 0.06 (0.03, 0.13)
Neural Network (MLP): -0.03 (-0.20, 0.15)
Estimate Refutation: Linear Regression
Random Common Cause: 0.001
p value: 0.98
Add Unobserved Common Cause: 0
Placebo Treatment Refuter: 3.11
p value: 0
Data Subset Refuter: 0.012
p value: 0.82

Acknowledgments

I would like to thank Professor George-Christopher Vosniakos from National Technical University of Athens for valuable feedback and insightful comments on the manuscript, which greatly improved the clarity and quality of this paper.

References

Castri, L., Uk, S. A., Uk, M. A., Bellotto, N., Van Der Schaar, M., Janzing, D., & Zhang, C. (2023). Enhancing Causal Discovery from Robot Sensor Data in Dynamic Scenarios Sariah Mghames Marc Hanheide. In *Proceedings of Machine Learning Research* (Vol. 213).

Chen, X., Abualdenien, J., Mahan Singh, M., Borrmann, A., & Geyer, P. (2022). Introducing causal inference in the energy-efficient building design process. *Energy and Buildings*, *277*. https://doi.org/10.1016/j.enbuild.2022.112583

Cheng, L., Guo, R., Moraffah, R., Sheth, P., Candan, K. S., & Liu, H. (2022). Evaluation methods and measures for causal learning algorithms. *IEEE Transactions on Artificial Intelligence*, *3*(6), 924–943.

Duhirwe, P. N., Ngarambe, J., & Yun, G. Y. (2024). Causal effects of policy and occupant behavior on cooling energy. *Renewable and Sustainable Energy Reviews*, *206*. https://doi.org/10.1016/j.rser.2024.114854

He, M., & Khorsand, M. (2024). Behavior-aware assessment of demand response participation in power systems using causal artificial intelligence. *Electric Power Systems Research*, *237*. https://doi.org/10.1016/j.epsr.2024.110981

Hellström, T. (2021). The relevance of causation in robotics: A review, categorization, and analysis. *Paladyn*, *12*(1), 238–255. https://doi.org/10.1515/pjbr-2021-0017

Hua, J., Li, Y., Liu, C., & Wang, L. (2022). A zero-shot prediction method based on causal inference under non-stationary manufacturing environments for complex manufacturing systems. *Robotics and Computer-Integrated Manufacturing*, *77*. https://doi.org/10.1016/j.rcim.2022.102356

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment Computing in Science & Engineering 9 (3), 90–95. *Publisher: IEEE COMPUTER SOC*.

JGraph. (2005). *diagrams.net: A JavaScript, client-side editor for general diagramming and whiteboarding*. https://app.diagrams.net/

Jiang, F., & Kazmi, H. (2025). What-if: A causal machine learning approach to control-oriented modelling for building thermal dynamics. *Applied Energy*, *377*. https://doi.org/10.1016/j.apenergy.2024.124550

Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J., & Silva, R. (2022). Causal machine learning: A survey and open problems. *ArXiv Preprint ArXiv:2206.15475*.

Kim, S., & Park, C. S. (2025). Customized causal modeling of occupant behaviors in residential households: Causal sufficiency vs. data sufficiency. *Energy and Buildings*, *329*. https://doi.org/10.1016/j.enbuild.2024.115233

Kinttel, C. R., & Stolper, S. (2019). *Using Machine Learning to Target Treatment: The Case of Household Energy Use*. http://www.nber.org/papers/w26531

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., & others. (2016). Jupyter Notebooks–a publishing format for reproducible computational workflows. In *Positioning and power in academic publishing: Players, agents and agendas* (pp. 87–90). IOS press.

Ko, H., Lu, Y., Yang, Z., Ndiaye, N. Y., & Witherell, P. (2023). A framework driven by physics-guided machine learning for process-structure-property causal analytics in additive manufacturing. *Journal of Manufacturing Systems*, *67*, 213–228. https://doi.org/10.1016/j.jmsy.2022.09.010

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. In *Entropy* (Vol. 23, Issue 1, pp. 1–45). MDPI AG. https://doi.org/10.3390/e23010018

Maisonnave, M., Delbianco, F., Tohmé, F., Milios, E., & Maguitman, A. (2024). Learning causality structures from electricity demand data. *Energy Systems*. https://doi.org/10.1007/s12667-024-00684-6

Massidda, L., & Marrocu, M. (2023). Total and thermal load forecasting in residential communities through probabilistic methods and causal machine learning. *Applied Energy*, *351*. https://doi.org/10.1016/j.apenergy.2023.121783

McKinney, W., & others. (2011). pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, *14*(9), 1–9.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. In *ACM Computing Surveys* (Vol. 54, Issue 6). Association for Computing Machinery. https://doi.org/10.1145/3457607

Miraki, A., Parviainen, P., & Arghandeh, R. (2024). Electricity demand forecasting at distribution and household levels using explainable causal graph neural network. *Energy and AI*, *16*. https://doi.org/10.1016/j.egyai.2024.100368

Nadim, K., Ouali, M. S., Ghezzaz, H., & Ragab, A. (2023). Learn-to-supervise: Causal reinforcement learning for high-level control in industrial processes. *Engineering Applications of Artificial Intelligence*, *126*. https://doi.org/10.1016/j.engappai.2023.106853

Naser, M. Z. (2021). An engineer's guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating causality, forced goodness, and the false perception of inference. *Automation in Construction*, *129*. https://doi.org/10.1016/j.autcon.2021.103821

Pullinger, M., Kilgour, J., Goddard, N., Berliner, N., Webb, L., Dzikovska, M., Lovell, H., Mann, J., Sutton, C., Webb, J., & others. (2021). The IDEAL household energy dataset, electricity, gas, contextual sensor data and survey data for 255 UK homes. *Scientific Data*, *8*(1), 146.

Schölkopf, B. (2022). Causality for machine learning. In *Probabilistic and causal inference: The works of Judea Pearl* (pp. 765–804).

Shams Amiri, S., Mueller, M., & Hoque, S. (2023). Investigating the application of a commercial and residential energy consumption prediction model for urban Planning scenarios with Machine Learning and Shapley Additive explanation methods. *Energy and Buildings*, *287*. https://doi.org/10.1016/j.enbuild.2023.112965

Shimoni, Y., Karavani, E., Ravid, S., Bak, P., Ng, T. H., Alford, S. H., Meade, D., & Goldschmidt, Y. (2019). An evaluation toolkit to guide model selection and cohort definition in causal inference. *ArXiv Preprint ArXiv:1906.00442*.

Sun, X., Wang, X., Yan, B., Wang, B., Pan, X., & Guo, J. (2024). A Systematic Review of Causal Analysis in Power Systems: Principles, Application, and Prospects. *2024 IEEE 7th International Electrical and Energy Conference (CIEEC)*, 3944–3949.

Vuković, M., & Thalmann, S. (2022). Causal Discovery in Manufacturing: A Structured Literature Review. In *Journal of Manufacturing and Materials Processing* (Vol. 6, Issue 1). MDPI. https://doi.org/10.3390/JMMP6010010

Wong, L. (2014). *Causal Inference in Python*. https://causalinferenceinpython.org/

Wu, D., & Wang, G. G. (2021). Causal artificial neural network and its applications in engineering design. *Engineering Applications of Artificial Intelligence*, *97*. https://doi.org/10.1016/j.engappai.2020.104089

Xu, G., Duong, T. D., Li, Q., Liu, S., & Wang, X. (2020). *Causality Learning: A New Perspective for Interpretable Machine Learning*. http://arxiv.org/abs/2006.16789

Xu, Y., Loftness, V., & Severnini, E. (2021). Using machine learning to predict retrofit effects for a commercial building portfolio. *Energies*, *14*(14). https://doi.org/10.3390/en14144334

Zhou, B., Li, X., Liu, T., Xu, K., Liu, W., & Bao, J. (2024). CausalKGPT: Industrial structure causal knowledge-enhanced large language model for cause analysis of quality problems in aerospace product manufacturing. *Advanced Engineering Informatics*, *59*. https://doi.org/10.1016/j.aei.2023.102333