Outcome-Explorer: A Causality Guided Interactive Visual Interface for Interpretable Algorithmic Decision Making

Md Naimul Hoque and Klaus Mueller, *Senior Member, IEEE*

Abstract—The widespread adoption of algorithmic decision-making systems has brought about the necessity to interpret the reasoning behind these decisions. The majority of these systems are complex black box models, and auxiliary models are often used to approximate and then explain their behavior. However, recent research suggests that such explanations are not overly accessible to lay users with no specific expertise in machine learning and this can lead to an incorrect interpretation of the underlying model. In this paper, we show that a predictive and interactive model based on causality is inherently interpretable, does not require any auxiliary model, and allows both expert and non-expert users to understand the model comprehensively. To demonstrate our method we developed Outcome Explorer, a causality guided interactive interface, and evaluated it by conducting think-aloud sessions with three expert users and a user study with 18 non-expert users. All three expert users found our tool to be comprehensive in supporting their explanation needs while the non-expert users were able to understand the inner workings of a model easily.

Index Terms—Explainable AI, Causality, Visual Analytics, Human-Computer Interaction.

1 Introduction

In recent years, algorithmic and automated decision-making systems have been deployed in many critical application areas across society [1], [2], [3]. It has been shown, however, that these systems can exhibit discriminatory and biased behaviors (e.g., [4], [5], [6]). It is thus imperative to create mechanisms by which humans can interpret and investigate these automated decision-making processes.

Several algorithmic [7], [8] and visual analytics [9], [10], [11], [12], [13], [14] solutions have been proposed to address this need. However, a shortcoming of these systems is that they have been predominantly developed from a model-builders perspective and as such do not support common lay (*non-expert*) users with no specific expertise in machine learning [15], [16]. It is these individuals, however, who are typically the recipients of a decision algorithm's outcome [15], [16].

This shortcoming is directly addressed in the 2016 EU General Data Protection Regulation (GDPR) which mandates that non-expert users who are directly impacted by algorithmic decisions have a "right to explanation". According to the GDPR users should receive these explanations in a *concise, transparent, intelligible, and easily accessible form* [17]. The GDPR has since become a model for laws beyond the EU, for example, the 2018 California Consumer Privacy Act (CCPA) bears many similarities with the GDPR. In addition, several recent studies [15], [16], [18] have further confirmed the needs of non-expert users to interpret and understand machine-generated decisions.

However, supporting non-expert users in a explainable AI (XAI) platform is challenging since they have different goals, reasons, and skill sets for interpreting a machine learning model than expert users. A model-builder such as a machine learning practitioner with significant data science expertise will want to interpret a model to ensure its accuracy and fairness. A non-expert user, on the other hand, will want to interpret the model to understand the service the model facilitates and delivers, and gain trust into it. In most cases, these users will not have background in data science and machine learning and so require an easy-to-understand visual representation of the model.

To exhibit human-friendly interpretability, a predictive model also needs to produce answers to explanation queries such as "Why does this model make such decisions?", "What if I change a particular input feature?" or "How will my action change the decision?" [19]. Existing methods and systems often employ an auxiliary model (post-hoc) to first approximate the original black-box model and then answer such questions since black-box models do not readily provide these explanations [7], [8], [20]. However, recent research suggests that such approximations can lead to incorrect interpretations of the underlying model and further complicate model understanding [21], [22]. Thus, an XAI interface for non-expert users should be inherently interpretable (i.e. directly observable) [21] and be able to answer causal questions without requiring an auxiliary model.

In this paper, we show that a predictive model based on causality (i.e., a causal model) meets the aforementioned criteria for a model that is understandable by non-expert users. The inner-workings of a causal model is directly observable through a Directed Acyclic Graph (DAG), making it inherently interpretable. The causal DAG is an intuitive representation and based on it and through interactions defined on it, a user can gain a good understanding of how variables are related to each other and how they affect the outcome variable, without the need for machine learning ex-

• *Md Naimul Hoque was and Klaus Mueller is with the Computer Science Department, Stony Brook University, Stony Brook, NY 11794*
E-mail: {mdhoque, mueller}@cs.stonybrook.edu

Manuscript received January 4, 2021; revised xxx

pertise. Further, causal models can provide truthful answers to causal explanation queries without auxiliary models.

To support our proposed method, we first developed a computational pipeline that would facilitate predictions in a causal model. This was needed since causal models do not support predictions by default. Informed by prior research and a formative study with ten non-expert users, we designed and developed Outcome-Explorer, an interactive algorithmic decision-making interface based on causality. Outcome-Explorer lets both non-expert and expert users interact with the causal DAG and supports common XAI functionalities such as answering What-If questions, exploring nearest neighbors, and comparing data instances. To evaluate the effectiveness of Outcome-Explorer, we first invited three expert users (machine learning researchers) to develop a causal model and then interpret the model using our tool. All three expert users found Outcome-Explorer comprehensive in terms of both its causal and its explanation functionalities.

We then conducted a user study with 18 non-expert lay users. We sought to understand how Outcome-Explorer would help non-expert users in interpreting a predictive model, in comparison to a popular post-hoc XAI method, SHAP [7]. The study revealed that participants were able to reduce interactions with variables that did not affect the outcome by 47%, meaning that participants understood which variables to change while using our tool. A similar reduction (36%) was also found for the *magnitude* of changes made to non-impacting variables. Hence, the interactions of the participants became more efficient which aided their understanding of the model. These outcomes confirm the high potential of Outcome-Explorer in both XAI research and application.

Our research contributions are as follows; we describe:

- A mechanism that allows an interactive assessment of prediction accuracy in a causal model.
- The design and implementation of a causality-guided interactive visual tool, Outcome-Explorer, to support the model explanation needs of both expert and non-expert users.
- Results from think-aloud sessions with three expert users which revealed that experts are able to build and interpret a predictive causal model correctly.
- A user study with 18 non-expert users which revealed that our tool can help non-experts understand a predictive causal model and enable them to identify the input features relevant for prediction.

Our paper is organized as follows. Section 2 and 3 describe related work and background. Section 4 presents our formative user study with non-expert users. Section 5 presents our design guidelines. Section 6 describes our interactive visual interface. Section 7 demos a use case. Section 8 presents the outcome of our system evaluation. Sections 9 and 10 present a discussion and conclusions.

2 Related Work

Current tools and methods in XAI can be broadly categorized into two distinct categories: (1) post-hoc explanation, and (2) explanation via interpretability. Post-hoc explanation methods explain the prediction of a model using local estimation techniques, without showing or explaining the workflow of the model. Instead, they give very detailed information on the impact of the model variables at a user-chosen data instance and seek to build trust into the model one instance at a time. Prominent examples of post-hoc explanation methods are SHAP [7] and LIME [8]. They are model agnostic and can explain the predictions of any machine learning model. While post-hoc methods are shown to be effective for explaining the decisions of complex black-box models, recent research has argued that these explanations can be misleading, and can complicate model understanding even more [21], [22].

To address these shortcomings, there has been a growing interest in devising models that are inherently interpretable. The Generalized Additive Model (GAM) [23] and the Bayesian Case Model (BCM) [24] are examples of this kind of model where the workflow is directly observable by a human user. In this paper, we advocate for a causal model since it has a natural propensity towards interpretability. According to Pearl [25], the aim of causal models is to "*explain to us why things happened, and why they responded the way they did*" which aligns perfectly with the objective of XAI. They have found frequent use to support reasoning about different forms of bias and discrimination [26], [27], [28], [29], [30], [31], [32] but none uses interactive visualization within the model as an XAI paradigm. Outcome-Explorer aims at bridging XAI, causal modeling, and visual analytics.

2.1 Visual Analytics and XAI

People are more likely to understand a system if they can tinker and interact with it [33]. In that spirit, interactive visual systems have proven to be an effective way to understand algorithmic decision-making. These interactive systems broadly categorize into the aforementioned post-hoc explanation and interpretable interfaces. Examples of interactive post-hoc explanation interfaces include the What-if Tool [34], RuleMatrix [20], Vice [35], Model Tracker [11], Prospector [36], and others. These tools use scatter plots, line plots, and text interfaces to allow users to query and compare the outcomes of different decision models, but without showing the model itself. Though this helps to understand the model's behavior in a counterfactual sense (the 'if'), it does not explain, and allow a user to play with the reasoning flow within the system (the 'why').

On the other hand, interpretable interfaces such as GAMUT [13], and SUMMIT [14] allow a user to interact with the model itself and provide explanations that are faithful to the model. While GAMUT is more conceptual in nature, pointing out the features an interactive interface should have to support model interpretation, SUMMIT specifically focuses on deep neural networks and the classification of images. It allows users to recognize how different classes of images are evaluated at different stages of the network and what features they have in common. Our tool, Outcome-Explorer, also falls into the category of these systems, but focuses on causal networks and quantitative data. It specifically targets non-expert users and actively supports four of the six interface features specified in GAMUT. We chose this subset since they appear most suited for non-expert users.

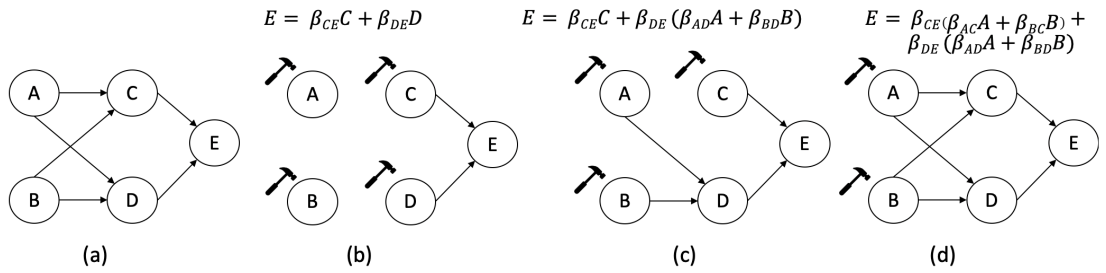


Fig. 1. Prediction in a causal model. The hammer icon represents intervention. (a) true causal model. (b) Interventions on all feature variables. The causal links leading to node C and D are removed since the values of C and D are set externally. (c) Interventions on node A, B, and C. (d) interventions on node A and B. In the path equations above models (b)-(d), the β are standardized regression coefficients estimated from the data.

2.2 Interactive Causal Analysis

Several interactive systems have been proposed for visual causal analysis but none have been designed for algorithmic decision-making. Wang and Mueller [37], [38] proposed a causal DAG-based-modeler equipped with several statistical methods. The aspect of model understanding and what-if experience it provides to users is via observing how editing the causal network's edges affects the model's quality via the Bayesian Information Criterion (BIC) score. Our tool, Outcome-Explorer, on the other hand conveys the what-if experience by allowing users to change the values of the network nodes (the model variables) and observe the effect this has on the outcome and other variables. At first glance both help in model understanding, but only the second is an experimenter's procedure. It probes a process with different inputs and collects outcomes (predictions), using the causal edges to see the relations with ease. This kind of what-if analysis has appeared in numerous XAI interfaces [13], [34], [35] and we have designed ours specifically for causal models. The mechanism also appeals to self-determination and gamification which both play a key part in education and learning [39]. One might say that achieving a lower BIC score is also gamification, but a BIC score does not emotionally connect a person to the extent that a lower house price or a college admission does. It provides a storyline, fun and realism, all elements of gamification [40].

Yan et al. [41] proposed a method that allows users to edit a post-hoc causal model in order to reveal issues with algorithmic fairness. Their focus is primarily on advising analysts which causal relations to keep or omit to gain a fairer adjunct ML model. We, on the other hand, use the causal model itself for prediction and focus on supporting the *right to explanation* of both expert and non-expert users.

The work by Xie et al. [42] is closest to ours but it uses different mechanisms for value propagation and it also addresses a different user audience. They look at the problem in terms of distributions which essentially is a manager's cohort perspective, while we look at specific outcomes from an individual's perspective. Both are forms of what-if analyses but we believe the latter is more accessible to a person directly affected by the modeled process, and so is more amenable to non-expert users who do not think in terms of distributions, uncertainties, and probabilities.

Finally, several visual analytics systems have been proposed for causal analysis in ecological settings. CausalNet [43] allows users to interactively explore and verify phenotypic character networks. A user can search for a subgraph and verify it using several statistical methods (e.g., Pearson Correlation and SEM) and textual descriptions of the relations, mined from a literature database. Natsukawa et al. [44] proposed a dynamic causal network based system to analyze evolving relationships between time dependant ecological variables. While these systems have been shown to be effective in analyzing complex ecological relations, they do not provide facilities for network editing and interventions which ours does. They also do not support what-if and counterfactual analyses which are instrumental for an XAI platform [13], [34].

In summary, the fundamental differences of ours to the existing systems are: (1) introduction of a prediction mechanism; (2) allowing users to change variables (intervention) by directly interacting with the nodes in the causal DAG; (3) visualizing the interplay between variables when applying/removing interventions; (4) supporting explanation queries such as what-if analyses, neighborhood exploration, and instance comparisons; and (5) including non-expert users in the design process, supporting their explanation needs.

3 Background

We follow Pearl's Structural Causal Model (SCM) [45] to define causal relationships between variables. According to SCM, causal relations between variables are expressed in a DAG, also known as Path Diagram. In a path diagram, variables are categorized as either exogenous (U) or endogenous (V). Exogenous variables have no parents in a path diagram and are considered to be independent and unexplained by the model. On the other hand, endogenous variables are fully explained by the model and presented as the causal effects of the exogenous variables. Figure 1 presents two exogenous variables (A and B) and three endogenous variables (C, D, and E). Formally, the Causal Model is a set of triples (U, V, F) such that

- U is the set of exogenous variables, and V is the set of endogenous variables.
- Structural equations [46] (F) is a set of functions $\{f_1, ..., f_n\}$, one for each $V_i \subseteq V$, such that $V_i = f_i(pa_i, U_{pa_i})$, $pa_i \subseteq V \setminus \{V_i\}$ and $U_{pa_i} \subseteq U$.

The notation "pa_i" refers to the "parents" of V_i.

3.1 Causal Structure Search

The causal structure between variables (F) can be obtained in three different ways: (1) causal structure defined from domain-expertise or prior knowledge; (2) causal structure learned from automated algorithms; and (3) causal structure learned from mixed-initiative human-AI collaboration.

The first method is the prevalent way to operate causal analysis in domains such as social science or medical science where expert-users or researchers are solely responsible for defining the causal structure [47]. In such scenarios, researchers utilize prior knowledge, domain expertise, and empirical evidence gathered from experiments such as randomized trials to hypothesized causal relations and then test the validity of the model through Structural Equation Modelling (SEM). Software such as "IBM SPSS AMOS", and "Lavaan" are build upon this principle.

On the other hand (in the second approach), automated causal search algorithms utilize conditional independence tests to find causal structure among data [48]. These algorithms help the user identify underlying causal structures between a large set of variables. The "Tetrad" software provides a comprehensive list of such algorithms.

The third approach combines the first two approaches. One pitfall of the automated algorithms is that they may not find the true causal structure since multiple causal structures can meet the constraints set out by the algorithms. In such scenarios, human verification is necessary to validate the causal structure obtained from automated algorithms. Prior research and empirical evidence suggest this human-centered research approach can identify the causal structure better than automated algorithms alone [37], [38], [49].

Finally, once the causal structure (F) is learned via any of the above approaches, we can use SEM to parameterize the model, obtaining the path/beta coefficients.

3.2 Prediction

We define the outcome variable Y as an endogenous variable, the variable we want to predict from a set of feature variables X. In a causal model (M), estimating Y from X is analogous to applying an intervention (fixing variables to specific values) on X [50]. This is achieved through the do operator which simulates a causal intervention by deleting certain edges from M while fixing X to specific values [45]. The resultant causal model M' is a subgraph of the original model M. Figure 1(b) shows M' when intervention is applied to all variables of X. Note that the edges leading to Node C and D from the original model are removed in figure 1(b). This is because nodes A and B can no longer causally effect C and D, once we fix them to specific values. In this scenario, $Y(E)$ can be estimated using this equation:

$$P_M(Y|do(X = x)) = P_{M'}(Y) = \beta_{CE}C + \beta_{DE}D \quad (1)$$

The β are standardized regression coefficients estimated from SEM. Figure 1(c) and (d) present different intervention scenarios on the original model M where the predictions P_M follow the equations shown above the respective model. The key idea is that Y is independent of its ancestors conditioned on all of its parents. Once we know the direct parents Y, other variables in the causal DAG can no longer influence Y.

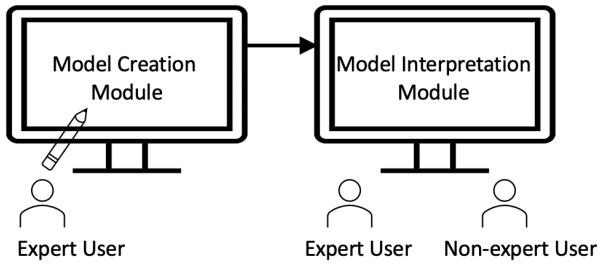


Fig. 2. Two module design of Outcome-Explorer with respective target users. An expert user would use the Model Creation module to create the causal model. After that, both expert and non-expert users would use the Model Interpretation module to interpret the causal model.

4 Formative Study with Non-expert users

We conducted formative interviews with 10 users (5 female, 5 male) to understand the expectations that non-expert users have of a decision-making interface. We recruited the participants (**P1-P10**) through social media posts and local mailing lists. Our inclusion criteria included familiarity with online decision-making services such as credit card approval and insurance services; algorithmic expertise was not required. We prompted the participants about their experience with these decision-making services. Their feedback is summarized in the following.

Need for transparency. We noticed a general need for transparency among non-expert users. They appeared to prefer an automated system over human assistance, but feared the systems might not give them the optimal service. Several participants mentioned that automated platforms allowed them to obtain service quickly and efficiently, whereas to get human assistance they often had to wait on the phone for a long time (**P1-4, P6, P9**). Yet, several participants mentioned that eventually they needed to contact a human agent since specific rules and provisions were often not readily available in the automated systems (**P1-4, P6, P9**). Thus, Outcome-Explorer should be completely transparent and provide necessary explanations for decisions.

How can I improve the decision? When asked about the process of evaluating algorithmic decisions, the participants mentioned that they would repeatedly update the input features to change the decision in their favor (**P1-7, P9**). They would try to make sense of the underlying algorithm by changing the values of variables and then observe the effect this had on the outcome variable (**P1-7, P9**). We note that this process is the same as obtaining **local instance explanation** and asking **what-if (counterfactual) questions**, which are capabilities C1 and C3 identified by Hohman et al. [13] as needed by expert users to interpret a single decision. Our tool should support C1 and C3 for non-expert users as well.

How am I different than my friend? Users of automated decision-making systems often employ a mental process of comparing themselves with others and try to make sense of why different people received different decisions. The participants shared several such cases where they wondered why they received one decision, while their friends received different decisions (**P1-5, P7-8, P10**). We note that this is similar to **instance comparison** and **neighborhood exploration**, which are capabilities C2 and C4 identified by Hohman et al. [13] that expert users should have to compare a data

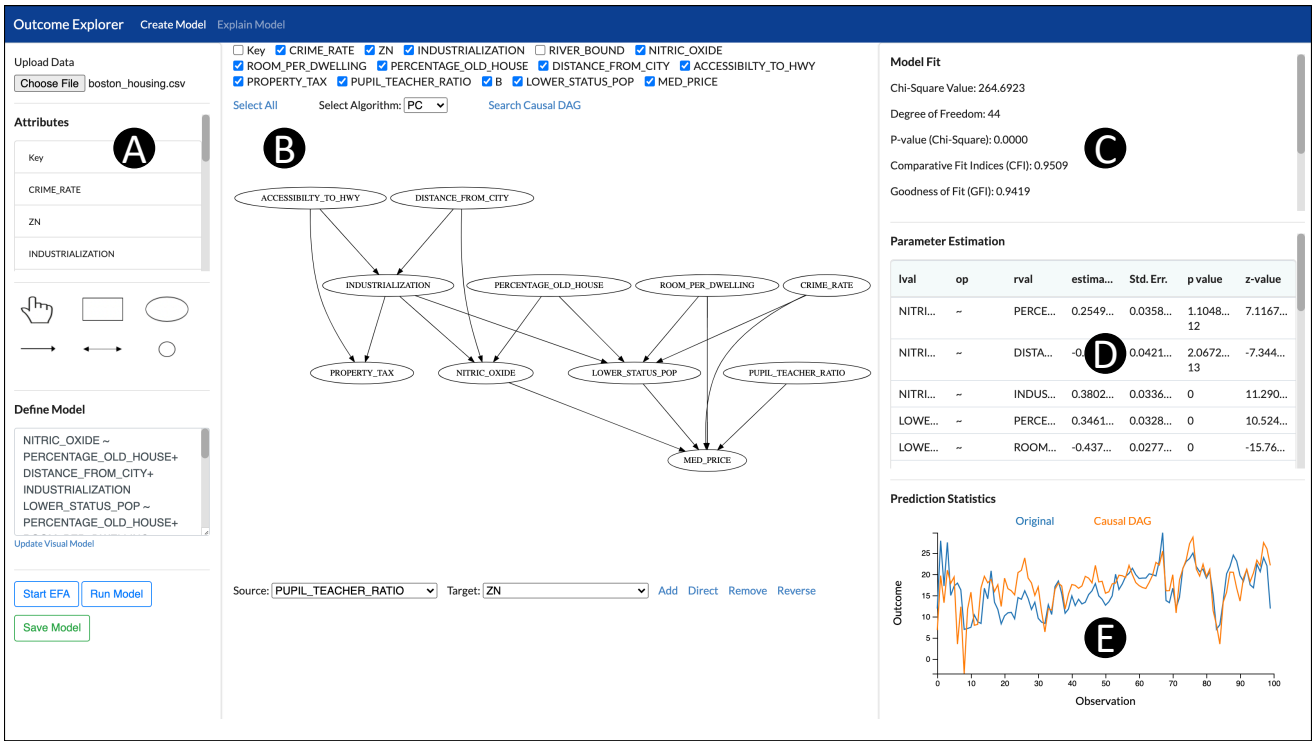


Fig. 3. Model Creation Module of Outcome-Explorer. A) Control Panel. B) Causal structure obtained from the search algorithms. Users can interactively add, remove, and direct edges in the causal structure. C) Model fit measures obtained from Structural Equation Modelling (SEM). D) Parameter estimation for each relations (beta coefficients). E) A line chart showing the prediction accuracy of the Causal Model on the test set.

point to its nearest neighbors. Our study shows that our tool should support these capabilities also for non-expert users.

5 Design Guidelines

Based on the insights gathered from our formative study, we formulate the following design guidelines:

DG1. Supporting experts and non-experts via a two module design: The formative study revealed overlapping interests between expert and non-expert users to interpret predictive models. However, a model needs to be created before it can be interpreted. XAI interfaces typically accept trained models for this purpose [13], [14], [20]. However, at the time of the development of this work, no open-source software or package was available for human-centered causal analysis (the third method from Section 3.1). Additionally, none of the existing tools supported prediction in a causal model. Hence, we decided to also support the creation of a predictive causal model.

The methods described for creating a causal model in Section 3 requires substantial algorithmic and statistical expertise which can only be expected from an expert user. The relationship between expert and non-expert users follows the *producer-consumer* analogy where an expert user will create and interpret the model for accurate and fair modeling, while a non-expert user will interpret this verified model to understand the service the model facilitates. To support this relationship, we decided that Outcome-Explorer should have two different modules: **(1) Model Creation** module, and **(2) Model Interpretation** module. Figure 2 shows the two modules and their respective target users.

DG2. Creating the Model: Using the Model Creation module, an expert user should be able to create a causal model interactively with the help of state-of-the-art techniques and evaluate the performance of the model.

DG3. Interpreting the Causal DAG: The causal DAG is central to understanding a causal model. The visualization and interaction designed for the causal DAG in the Model Interpretation module should allow both expert and non-expert users to interpret the model correctly. Users should be able to set values to the input features in the DAG to observe the changes in the outcome.

DG4. Supporting Explanation Queries: The formative study revealed that non-expert users ask explanation queries (**C1-C4**) similar to those already well-studied in XAI research [13]. Our tool should support these queries and they should be implemented keeping in mind the algorithmic and visualization literacy gap between expert and non-expert users.

DG5. Input Feature Configuration: Our tool is completely transparent and a non-expert user can change the input features freely in the interface. However, when engaging in this activity, it is possible that to obtain a certain outcome a user might opt for a feature configuration that is unlikely to be realistic [21]. Thus, our tool should allow non-expert users to evaluate not only the value of the outcome, but also how realistic the input configuration is when compared to existing data points and configurations.

6 Visual Interface

Outcome-Explorer is a web-based interface. We used Python as the back-end language and D3 [51] for interactive visual-

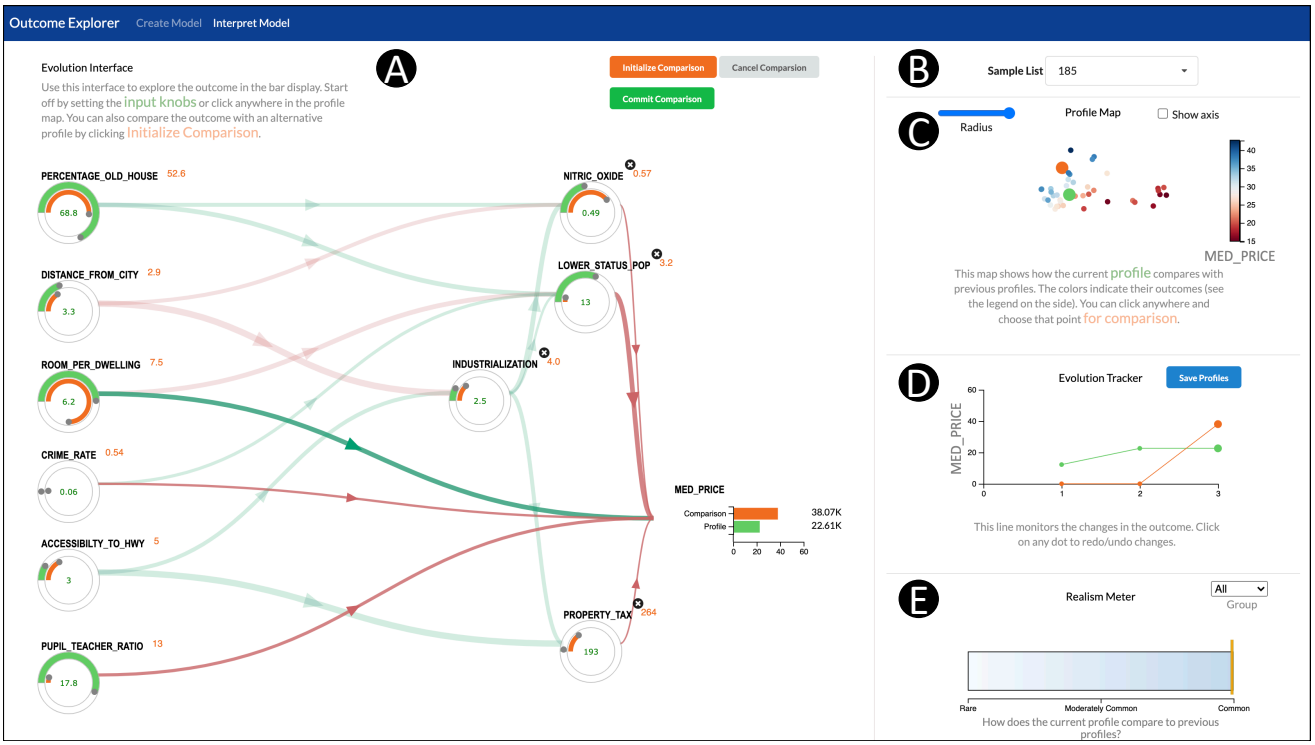


Fig. 4. Model Interpretation Module of Outcome-Explorer. A) Interactive causal DAG showing causal relations between variables. Each node includes two circular knobs (green and orange) to facilitate profile comparisons. The edge thickness and color depict the effect size and type of each edge. B) Sample selection panel. C) A biplot showing the position of green and orange profiles compared to nearest neighbors. D) A line chart to track the model outcome and to go back and forth between feature configuration. E) Realism meter allowing users to determine how common a profile is compared to other samples in the dataset.

ization. We used Tetrad [1], and semopy [2] for causal analysis.

As per **DG1**, Outcome-Explorer has two interactive modules. While the Interpretation Module is accessible to both expert and non-expert users, the Model Creation Module is only available to the expert users (**DG1**). We describe the visual components of these two modules below.

6.1 Model Creation Module

The Model Creation Module is divided into five regions (Figure 3). The Control Panel (A) allows an expert user to upload data and run statistical models on the data. The Central Panel (B) visualizes the causal model obtained from automated algorithms. An expert user can select the appropriate algorithm from a dropdown list there. The graph returned by the automated algorithms is not necessarily a DAG; it can contain undirected edges. Besides, an expert user can edit the graph in this module (**DG2**), often resulting in a change of the structure of the graph. Since the structure of the graph is uncertain, we decided to use GraphViz [52], a well-known library for graph visualization. The panel facilitates four sets of edge editing features: (1) Add, (2) Direct, (3) Remove, and (4) Reverse. When editing the causal model, an expert user can evaluate several model fit measures in panel C, model parameters in D, and prediction accuracy in E.

1. https://www.ccd.pitt.edu/tools/
2. https://semopy.com

6.2 Model Interpretation Module

The interpretation module (Figure 4) uses a different visual representation to present the graph than the model creation module since a parameterized causal model has a definitive structure (DAG). The interpretation module accepts a DAG as input and employs topological sort to present that DAG in a left to right manner.

Each variable in the causal model contains two circular knobs: a green and an orange knob. A user can control two different profiles independently by setting the green and orange knobs to specific values (**DG3**). This two profile mechanism facilitates instance comparison and what-if analysis (**DG3**, **DG4**, see Section 7). The range for the input knobs is set from the min to the max of a particular variable. Each knob provides a grey handle which a user can use to move the knob through mouse drag action. The user can also set the numbers directly in the input boxes, either an exact number or even a number that is out of range (outside $(max - min)$ range) for that variable. In case of an out of range value, the circular knob is simply set to min or max, whichever extrema are closer to the value. The outcome variable is presented as a bar chart in the causal model. Similar to the input knobs, the outcome variable contains two bars to show prediction values for two profiles.

Finally, we follow the visual design of Wang et al. [37], [38] to encode the edge weights in the causal DAG. To visualize intervention, all edges leading to an endogenous variable are blurred whenever an user sets that variable to a specific value. Consequently, the user can cancel out the intervention by clicking the $\times$ icon beside an endogenous

variable in which case its value is estimated from its parent nodes and the edges return to their original opacity.

6.2.1 Profile Map

The profile map (Figure 4(C)) is a biplot which shows the nearest neighbors of a profile (**DG4**). To compute the biplot, we run PCA on the selected points. A user can control the radius of the neighborhood, given by the range of the outcome variable, through the "Radius" slider. The neighbors are colored according to the outcome value, as specified in the color map on the right of the plot.

A user can hover the mouse over any point on the map to compare the data point with the existing green profile (**DG4**). Subsequently, the user can click on any point to set that point as a comparison case for a more detailed analysis. Both green and orange disks (larger circles) move around the map as the user changes the profiles in the causal model.

6.2.2 Evolution Tracker

One of the fundamental features of any UI is the support for a redo and undo operations. We introduced the Evolution Tracker (Figure 4(D)) to facilitate this. The tracker is a simple line chart with two lines for two profiles in the system. The x-axis represents the saved state while the y-axis shows the outcome value at that particular state. A user can click on the "Save Profiles" button to save a particular state in the tracker and can click on any point in the tracker to go back and forth between different states.

6.2.3 Similarity (or Realism) Meter

We introduced the similarity (*realism*) meter to allow users to determine how common their profile is compared to that of the existing population captured by the dataset (**DG5**). It is a safety check for unreasonable expectations that could be generated by the interactive interpretation module, and so it is an important part of the interface.

To realize it, we opted for a multi-dimensional method similar to detecting an outlier in one dimension using the z-score. At first, we fit a Gaussian Mixture Model on the existing data points in multivariate space. A mixture model with K Gaussians or components is defined as:

$$P(X) = \prod_{n=1}^{N} \sum_{k=1}^{K} P(X_n|C_k)P(C_k) = \prod_{n=1}^{N} \sum_{k=1}^{K} \phi_k N(X_n|\mu_k, \Sigma_k) \tag{2}$$

where N is the number of datapoints, $\phi_k = P(C_k)$ is the mixture weight or prior for component k, and μ_k, Σ_k are the parameters for the k-th Gaussian. Once the parameters are learned through the Expectation-Maximization algorithm, we can calculate the probability of a datapoint x belonging to a component C_i using the following equation

$$P(C_i|x) = \frac{P(C_i)P(x|C_i)}{\sum_{k=1}^{K} P(C_k)P(x|C_k)} = \frac{\phi_i N(x|\mu_i, \Sigma_i)}{\sum_{i=1}^{K} \phi_k N(x|\mu_k, \Sigma_k)} \tag{3}$$

A high value of $P(C_i|x)$ implies that x is highly likely to belong to C_i, whereas a low value $P(C_i|x)$ implies that the features of x is not common among the members of C_i. Thus, $P(C_i|x)$ can be interpreted as a scale of how "real" a datapoint is to the other members of a component. We translate $P(C_i|x)$ to a human understandable meter with $P(C_i|x) = 0$ interpreted as "Rare", $P(C_i|x) = 0.5$ as "Moderately Common", and $P(C_i|x) = 1$ as "Common".

7 Usage Scenario

In this section, we present a usage scenario to demonstrate how a hypothetical expert user (Adam), and a non-expert user (Emily) could benefit from Outcome-Explorer.

Adam (he/him) is a Research Engineer at a technology company and is responsible for creating a housing price prediction model. Non-expert users will eventually use the model. As a result, Adam also needs to create an easy-to-understand interactive interface for the non expert users. Based on these requirements, Adam decides to use an interpretable model for prediction and determines that Outcome-Explorer matches the requirements perfectly.

7.1 Creating the Model

Adam starts off Outcome-Explorer by uploading the housing dataset [53] into the Model Creation Module (Figure 3). Next, Adam selects the PC algorithm [48] for searching the causal structure (not depicted). Upon seeing the causal DAG obtained from the PC algorithm, Adam uses prior knowledge to refine the causal relations. For example, Adam notices that the initial model has an undirected edge between "INDUSTRIALIZATION" and "DISTANCE_FROM_CITY". From domain expertise, Adam knows that "DISTANCE_FROM_CITY" can be a cause of "INDUSTRIALIZATION", but the opposite relation is not plausible. Adam directs the edge from "DISTANCE_FROM_CITY" to "INDUSTRIALIZATION" and notices that the model fit measures (Figure 3(C)) have also increased. Figure 3 presents the final causal DAG obtained in this iterative process (see supplemental video for the intermediate steps). The final model fit measures are: Comparative Fit Index (CFI): 0.951; Goodness of Fit (GFI): 0.950; Adjusted Goodness of Fit (AGFI): 0.919; and RMSEA: 0.0997. The measures indicate a good SEM model fit. Satisfied by the model performance, Adam hits the "Save Model" button and moves to the "Model Interpretation" module.

7.2 Interacting with the Causal DAG and Exploring Nearest Neighbors

After creating the model, Adam wants to explore and verify the model in the Interpretation Module (Figure 4) before making it public. Adam starts off the exploration process by selecting a sample data row from Figure 4(B). Adam observes that the selected row is immediately reflected in the feature values and the outcome of the model has changed to 22.61K. Adam also observes that the edges entering the endogenous variables became blurred (deactivated) since those variables were set to reflect the selected row and can no longer be estimated from the exogenous variables. From the profile map in Figure 4(C), Adam notices that the selected housing has a relatively small price. Adam decides to compare the selected housing with higher prices. To do so, Adam selects a data point with a higher housing price from the profile map. Immediately, an orange profile

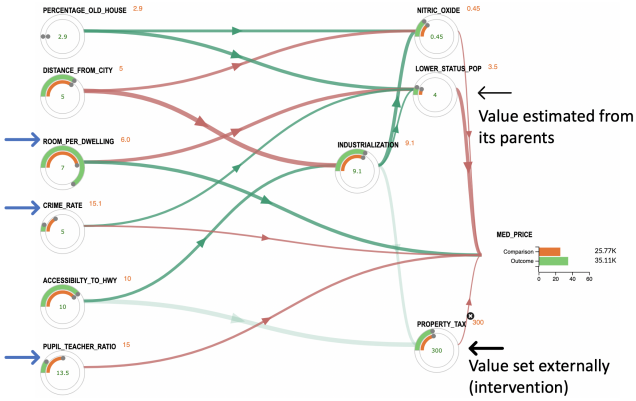


Fig. 5. Asking what-if questions in Outcome-Explorer. A user can keep one profile (green) fixed, and change the other profile (orange) to ask what-if questions. The blue arrows indicate the changes in the orange profile. Note that property tax is set to 300 by the user. As a result, changing its parents will not affect property tax. The other endogenous variables are estimated from their parents.

is created in the causal DAG. From the two profiles, Adam easily understands where the two housing differed and how that affected the outcome. At this point, Adam hits the "Save Profile" button to save both the profiles in the tracker (Figure 4(D)). In a similar manner, Adam explores several other data points to get a concrete idea of the model. At the end of the analysis, Adam is confident that the model is accurate, interpretable, and is ready for deployment. Adam then publishes the interface with the name *housingX*.

7.3 Understanding the Causal Relations

Emily is a middle-aged female (she/her) who would like to purchase a new house. Emily decides to understand the quality of the desired neighborhood through *housingX*.

Emily visits the site, which features the Interpretation Module of Outcome-Explorer, and starts off by watching a small tutorial on how to use the interface. After that, Emily starts to make sense of the variables and how they are connected to each other. Emily notices that Property Tax is calculated from a region's accessibility to the highway and the industrialization index. Emily further notices that Industrialization depends on both the region's distance from the city and its accessibility to the highway. Based on that, Emily concludes that Property tax depends on three factors: Accessibility to Highway, Distance from City, and Industrialization. Interestingly, Emily observes that the Median Price has a red edge from Property Tax, meaning houses located in areas with higher property taxes are priced lower than houses from areas with lower property taxes.

7.4 Exploring Outcomes & What-if Analysis

After understanding the causal relations between variables, Emily now starts putting values to the knobs in the interface. Emily observes that once the value of an internal node is set, edges leading to that node become blurred. For example, Emily sets the value of Property Tax to 300, which decreases the Median Price of the houses, but also blurs the edges entering Property Tax (Figure 5).

After finding the ideal neighborhood, Emily notices that the Median Price of that neighborhood is around $35,000. But, Emily only has a budget of around $25,000. Based on that, Emily decides to change the variables so that the median price comes down to $25,000. Emily fires off the comparison mode by clicking the "Initialize Comparison" Button. Immediately, a new orange profile is created in the interface which is exactly the same as the current green profile. Keeping the green profile constant, Emily iteratively changes the variables to take down the median housing prices to $25,000. In this iterative process, Emily utilizes the tracker regularly to go back and forth between different profile configurations. Emily also consults the realism meter regularly to see how common the selected housing is compared to the existing neighborhoods. She confirms that the orange line stays on the right end of the meter which means that these configurations are very common (Figure 4).

8 Evaluation

We evaluated Outcome-Explorer in two phases: (1) we conducted think-aloud sessions with three ML practitioners to gather expert feedback and gauge the system's real-world potential; (2) we conducted a user study with 18 users to assess the effectiveness and usability of Outcome-Explorer in supporting the explanation needs of non-expert users.

8.1 Expert Evaluation

We invited three ML practitioners as expert users (1 female, 2 male) to examine Outcome-Explorer. Participation was voluntary with no compensation. All three participants had post-graduate degrees and had conducted research in the field of XAI, Fairness, and Data Ethics for at least five years. They were also familiar with statistical causal analysis.

The tool was deployed on a web server and the sessions were conducted via Skype. Participants shared their screen as they performed the tasks. One author communicated with the participants during the sessions while another author took notes. Each session started with the participant observing a live demo of the tool. After that, participants were asked to choose one out of two datasets: Boston Housing [53] and the PIMA Diabetes dataset [54]. Once a participant chose a dataset, we provided them with the textual descriptions of the features and a task list. The task list was designed to guide the participants in exploring different components of Outcome-Explorer. Participants started off by creating a causal model using the Creation Module, and then gradually moved into examining different explanation methods available in the Interpretation Module. While performing the tasks, participants thought-aloud and conversed with the authors continuously. We sorted their feedback in the four thematic categories, as described next.

8.1.1 Comprehensive and Generalizable

All three expert users found the "Model Creation" module to be "comprehensive", and "generalizable". Participants found the accuracy statistics to be most helpful as that feature is not available on other comparable causal analysis tools. According to E1: Outcome-Explorer was "rigorous" in terms of causal functionality and should enable users to obtain the "best possible causal model".

The participants also found the interpretation module to be comprehensive. E1 mentioned that the interactive causal

DAG alone should allow non-experts to understand the model. Additionally, they found the two profile comparison mechanism to be helpful and appreciated the fact that the user can ask the what-if questions directly to the model.

8.1.2 Engaging, Thought Provoking, and Fun

Participants continuously engaged themselves in making sense of the causal relations. Throughout the session, they enthusiastically initiated discussions with the authors to share their personal experiences related to causal relations.

Participants also found the visual design of the Interpretation Module to be ascetically pleasing and fun to interact with. They mentioned that the interface has a "certain gaming flavor" to it. E3 opined that the thought-provoking and interactive nature of Outcome-Explorer might entice curious non-expert users to gather knowledge on a domain of interest.

8.1.3 Prior Knowledge and Position in the ML Pipeline

Participants suggested that Outcome-Explorer could be used once an expert user has preprocessed and explored the dataset. It would provide users the necessary background knowledge for creating and explaining the causal model. According to E1,

"I can see that the user might need to tweak the initial causal model iteratively to reach the final model, but that is also true for many ML models. The process of creating a predictive model is often messy, and requires several iterations, each of which requires users to utilize prior knowledge to refine the model."

8.1.4 Disclaimers

Causal relations make stronger claims than associative (correlative) relations. Participants suggested that the implication of the causal relations should appropriately be communicated to the end-users. For example, a particular causal relation may hold true for a particular task or domain, but not in general. Expert users should be aware of such potential misleading relations in the causal model, and should provide disclaimers to the non-expert users whenever needed. This will ensure that non-experts are not misled into thinking that the causal relations in Outcome-Explorer are ubiquitously true.

8.2 User Study with Non-expert Users

To understand how non-expert users might benefit from Outcome-Explorer we conducted a user study. We aimed at validating the following hypotheses:

- **H1**: Outcome-Explorer will improve a user's understanding of the embedded predictive causal model in comparison to the state-of-the-art explanation method.
- **H2**: Outcome-Explorer will increase a user's efficiency in reaching a desired outcome in comparison to the state-of-the-art explanation method.
- **H3**: Outcome-Explorer will be easy to use.

We chose SHAP [7] as a comparison case for Outcome-Explorer as it is a prevalent and widely used post-hoc explanation technique. Another motivation for comparing our approach with SHAP was the interpretable nature of our tool. SHAP approximates the prediction mechanism of a model without showing the model itself, an approach fundamentally different than our. Additionally, SHAP is open-source and provides several visualizations to aid the explanations, ensuring a fair comparison with our visual interface. Hence, we conducted a repeated-measures within-subject experiment with the following two conditions.

C1. **SHAP**: This condition included input boxes which users could use to change variables. Users had access to two charts provided by SHAP: a bar chart showing global feature importance and a variant of stacked bars (force plot) showing the feature contribution for a decision (see supplemental material).

C2. **Outcome-Explorer-Lite**: This prototype included only the interactive causal DAG of the Interpretation Module with other components hidden (see supplemental material).

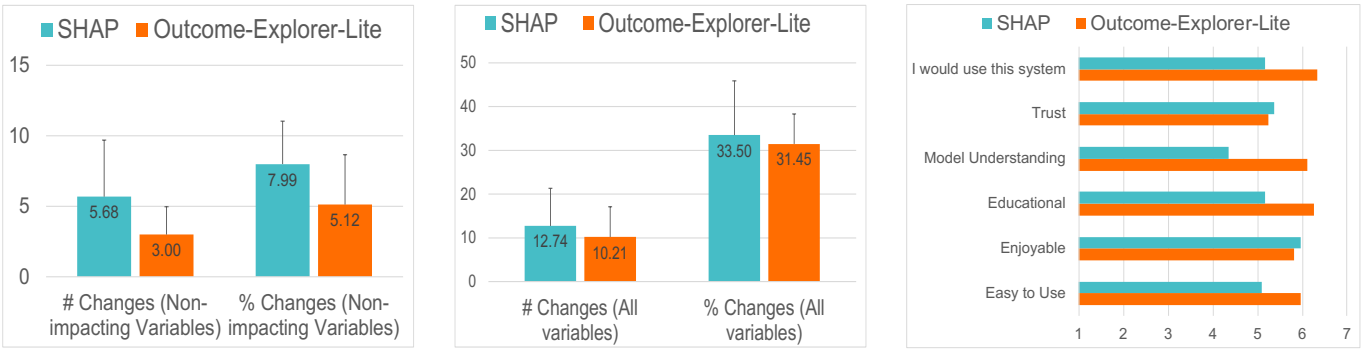
We chose to include only the causal DAG in the study as the other components provide auxiliary tools to understand the model, but are not necessary to interpret the model. The inclusion of these components could hinder a fair comparison between Outcome-Explorer and SHAP. To minimize the learning effect, we included two datasets and counterbalanced the ordering of study conditions and datasets. These datasets are the Boston housing and the PIMA Diabetes dataset, both of which appeared previously in the XAI literature [13], [20].

8.2.1 Participants

We recruited 18 participants (10 males, 8 females) through local mailing lists, university mailing lists, and social media posts. Participation was voluntary with no compensation. The participants varied in age from 19 to 35 ($M = 25$, $SD = 4.21$). None of the participants had machine learning expertise. The participants were comfortable in using web technology and had a high-level idea of automated decision-making through exposure to credit-card approval and loan approval systems. Additionally, two participants had experience with interactive visualization through interactive online news. All participants reported a basic understanding of the dataset domains (housing and diabetes) in the post-study interview, but did not report any specific expertise on the domains.

8.2.2 Tasks

XAI interfaces are frequently evaluated on "proxy tasks" such as how well humans predict the AI's decision, and subjective measures of trust and understanding [55]. Recent research suggests that proxy tasks and subjective measures are not good predictors of how humans perform on actual decision-making tasks [55]. Based on that, we decided to evaluate our tool on actual decision-making tasks. The tasks were similar to the case study presented in Section 7.4. For example, in the case of the housing dataset, we provided the participants with the scenario of a person who wants to buy an ideal housing (e.g. housing with price 35K) with budget constraints. We then asked the participants to reach alternative/target outcomes (e.g. reducing housing price from 35K to 25K) to satisfy the budget constraints while minimizing the number of changes, and the magnitude of the changes



(a) Average number of changes and magnitude of changes on non-impacting variables

(b) Average number of changes and magnitude of changes on all variables

(c) Subjective Measures (1: Strongly Disagree, 7:Strongly Agree).

Fig. 6. Study Results. The average number of changes and the average magnitude of changes (%) made to (a) non-impacting variables, and (b) all variables to reach the target outcomes. (c) Average self-reported subjective measures. Error bars show $+1$ SD.

from the ideal housing. Note that both conditions made predictions based on the same underlying causal model. We also collected self-reported subjective measures such as model understanding, trust, and usability.

8.2.3 Study Design

Similar to the sessions with the expert users, we conducted the study sessions via web and Skype. A study session began with the participant signing a consent form. Following this, the participants were introduced to the assigned first condition and received a brief description of the interface. The participants then interacted with the system (with a training dataset), during which they were encouraged to ask questions until they were comfortable. Each participant was then given a scenario and a task list for the first condition. After completing the tasks, participants rated the study conditions (interfaces) on a Likert scale ranging from 1 (Strongly Disagree) to 7 (Strongly Agree) based on six subjective measures. The same process was carried out for the second condition. Each session lasted around $\sim$1 hour and ended with an exit-interview.

8.2.4 Results

H1: Model Understanding. In a causal model, the exogenous variables may not affect the outcome if endogenous variables are set to specific values. While Outcome-Explorer visualizes this interplay, SHAP only estimates feature contributions to the decision and it does not explain why some variables are not affecting the outcome. We refer to such variables as *non-impacting variables*. We anticipated that interactions with non-impacting variables might reveal how well users understood the model. Based on that, to account for *model understanding*, we measured (1) *the number of changes (non-impacting)* and (2) *the magnitude of changes (%, non-impacting)*. Here, non-impacting refers to the changes made on non-impacting variables. We used a paired t-test with Bonferroni correction and Mann-Whitney U to assess statistical significance of the quantitative and likert scale measures respectively.

On average, the participants made 5.68 ($SD = 4.01$) changes to the *non-impacting* variables when using SHAP while for Outcome-Explorer-Lite the average was 3.00 ($SD = 1.97$). Participants reduced the changes made to the non-impacting variables by 47%, which was statistically significant ($p < 0.02$); Cohen's effect size value ($d = 0.68$) suggested a medium significance. We also found a significant difference between the magnitude of changes users made on non-impacting variables (36% reduction with $p < 0.001$, plotted in Figure 6a).

In order to understand how study conditions and dataset relate to each other with respect to the above quantitative measures, we constructed two mixed-effect linear models, one for each measure. We tested for interaction between study condition and dataset while predicting a specific measure. While there were no interaction effects and dataset did not play any significant role in predicting the measures, we found study condition to be the main effect in predicting the number of changes on non-impacting variables ($F(1, 26.033) = 7.723, p = 0.01$) and the magnitude of changes (%) on non-impacting variables ($F(1, 26.992) = 13.140, p = 0.001$).

Finally, as shown in Figure 6(c), participants rated Outcome-Explorer-Lite favorably in terms of Model Understanding ($M : 6.11, SD : 0.50$), and Educational ($M : 6.26, SD : 0.87$). In comparison, the scores for SHAP were: Model Understanding ($M : 4.34, SD : 0.911$), and Educational ($M : 5.15, SD : 0.964$). The differences were statistically significant with $p < 0.0001$ (Model Understanding) and $p < 0.01$ (Educational). However, we also observe that participants' trust did not improve in Outcome-Explorer. In the post-study interview, several participants mentioned their unpleasant experiences with automated systems. Such distrusts are unlikely to be changed in one study, and that might be the reason for equal trust in both conditions.

H2: Efficiency in Decision-making Tasks. We measured two quantitative measures to account for users' *overall performance* in decision-making tasks. They are the total (1) *number of changes*, and (2) *magnitude of changes (%)* made to input variables. As shown in Figure 6b, the average number of changes were 12.74 ($SD = 8.59$) for SHAP, and 10.21 ($SD = 6.88$) for Outcome-Explorer-Lite. The difference was not statistically significant. We also did not find a significant difference between the overall magnitude of changes (%) users made in each study condition. Finally, we measured the time taken to complete the tasks, but no statistically significant difference was found.

Similar to the above, we constructed two mixed-effect linear models. We did not find any interaction effects, and the datasets as well as the conditions did not play any significant role in predicting the measures.

H3: Ease of Use. Participants rated Outcome-Explorer-Lite favorably in terms of Easy to use ($M : 5.96, SD : 0.96$), and I would use this system ($M : 6.33, SD : 0.6$). In comparison, the scores for SHAP were: Easy to use ($M : 5.08, SD : 1.35$), and I would use this system ($M : 5.16, SD : 1.42$). The differences were statistically significant with $p < 0.04$ (Easy to use) and $p < 0.03$ (I would use this system). The other metric (Enjoyable) was not statistically significant.

The results matched our anticipation that Outcome-Explorer-Lite will improve user model understanding and that they will learn more about the prediction mechanism using our tool. In the post-study interview, participants appreciated the visual design of the causal DAG which might be the reason why they found Outcome-Explorer-Lite to be easy to use and want to see it in practice.

On average, participants spent slightly more time when using Outcome-Explorer. While familiarizing with the interface was one factor for that, in the post-study interview, several participants mentioned that they felt curious, spent more time to learn the relations, and put some thought before taking an action. A participant, a senior college student, mentioned: *"The interface (Outcome-Explorer-Lite) is fun, attractive as well as educational. I feel like I learned something. I did not know much about housing prices before this session. But, I think I now have a much better understanding of housing prices. If available in public when I buy a house in the future, it will help me make an informed decision."*

9 Discussion and Limitations

Model Understanding vs Overall Performance: The user study validated H1 and H3, but not H2. The study revealed that participants reduced interactions with non-impacting variables significantly in Outcome-Explorer-Lite, indicating a better model understanding compared to SHAP. The lack of edges or blurred edges in Outcome-Explorer-Lite provided participants with clear evidence for non-impacting variables. On the other hand, while using SHAP, participants interacted with the non-impacting variables despite observing their zero feature importance in the visualization. They constructed several hypotheses about non-impacting variables while using SHAP, including the possibility of a change of impact in the future, and their indirect effects on other variables. This may be the reason for the increased interaction with the non-impacting variables. However, improved understanding of non-impacting variables did not result in better overall performance. Participants instead increased interaction with the impacting variables to understand the effects of causal relations while using Outcome-Explorer-Lite. As a result, the overall performance (i.e., the total number of interactions with impacting and non-impacting variables) remained similar for both conditions. We also believe that the increased focus on impacting variables while using Outcome-Explorer-Lite fostered the observed better model understanding in the subjective measures.

The Accuracy-Interpretability Trade-off: We acknowledge that causal models might not reach the prediction accuracy of complex machine learning models. The comparison requires rigorous experiments on common ML tasks which is beyond the scope of this paper. However, there exists empirical evidences where causal models or linear models such as ours outperformed complex ML models [50], [56]. As stated by Rudin [21], the idea that interpretable models do not perform as well as black-box models (the *accuracy-interpretability tradeoff*) is often due to the lack of feature engineering while building interpretable models.

Implication for XAI Research: Outcome-Explorer offers several design insights for visual analytics systems in XAI. First, its novel two module design shows that it is possible to support the explanation needs of experts and non-experts as well as the model creation functionalities for experts in a single system. While we acknowledge this is not a strict requirement for an XAI interface, we believe our work will motivate non-expert inclusive design of XAI interfaces in the future. Second, it shows that an effective XAI interface does not necessarily require a new and complex visualization. "Simple" visual design and "intuitive" interactions are effective ways to convey the inner workings of predictive models, especially for non-expert users.

Another potential impact of Outcome-Explorer is bridging XAI and algorithmic fairness research. Algorithmic fairness ensures fair machine-generated decisions while XAI ensures transparency and explainability of ML models. Although highly relevant, these two research directions have not been bridged together yet. By promoting the causal model, a highly effective paradigm for bias mitigation strategies, Outcome-Explorer opens the door for an ML model to be transparent, accountable, and fair altogether.

Finally, our design and visual encoding can be extended to other graphical models. For example, a Bayesian Network is also represented as a DAG, and the design of Outcome-Explorer can be transferred to interactive XAI systems based on Bayesian networks.

Limitations & Future Work: It is important for a causal model to have a sufficient number of variables that cover all or at least most aspects determining the predicted outcome. Causal inferencing under incomplete domain coverage can result in islands of variables or a causal skeleton where some links are reduced to correlations only. We are currently experimenting with evolutionary and confirmatory factor analysis to introduce additional variables that can complement the native set of variables. These variables are often not directly measurable and can serve as latent variables. Our preliminary work has shown that they can greatly add to both model comprehensibility and completeness.

Another source of error can be confounders which can lead to an overestimation or underestimation of the strength of certain causal edges. There are several algorithms available for the detection and elimination of counfonding effects and we are presently working on a visual interface where expert users can take an active role in this type of effort.

A current limitation of our system is scalability. At the moment we limit the number of variables to George Miller's Magical Number Seven, Plus or Minus Two paradigm [57]. This allowed us to understand the explanation needs of mainstream non-expert users and support-

ing them through interactive visualizations on previously studied XAI datasets [13], [20]. However, in many real-life scenarios, there can be "hundreds of variables", which could overwhelm non-expert users. We envision that the Model Creation Module could be enhanced with advanced feature engineering capabilities, such as clustering, dimension reduction, pooling of variables into latent variables (factor analysis), and level of detail visualization [58]. Alternatively, scalable causal graph visualization [42] could also be used for this purpose. Our future work will focus on gaining more insight on how much complexity non-expert users can handle, and which of the above-mentioned methods work best for them.

Finally, so far the participants we have studied were all from the younger generation (19-35) who generally tend to be savvier when it comes to the graphical tools used in our interface. In future work we aim to study how our system would be received by older members of society. It might require additional information integrated into the user interface, such as tooltips and pop-up suggestion boxes and the like.

10 Conclusion

We presented Outcome-Explorer— an interactive visual interface that exploits the explanatory power of the causal model and provides a visual design that can be extended to other graphical models. Outcome-Explorer advances research towards interpretable interfaces and provides critical findings through user study and expert evaluation.

We envision a myriad of applications of our interface. For example, bank advisors or insurance agents might sit with a client and use our system to discuss the various options with them (in response to their right to explanation), or a bank or insurance would make our interface available on their website, along with a short instructional video. Future work will explore how complex a model can get while still being understandable by non-expert users.

Acknowledgments

This research was partially supported by NSF grants IIS 1527200 and 1941613.

References

[1] T. Brennan, W. Dieterich, and B. Ehret, "Evaluating the predictive validity of the compas risk and needs assessment system," *Criminal Justice and Behavior*, vol. 36, no. 1, pp. 21–40, 2009.

[2] A. Chouldechova, D. Benavides-Prado, O. Fialko, and R. Vaithianathan, "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 134–148.

[3] Z. Obermeyer and S. Mullainathan, "Dissecting racial bias in an algorithm that guides health decisions for 70 million people," in *Conf. on Fairness, Accountability, and Transparency*, 2019, pp. 89–89.

[4] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," *ProPublica, May*, vol. 23, p. 2016, 2016.

[5] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.

[6] O. Keyes, "The misgendering machines: Trans/hci implications of automatic gender recognition," *ACM CSCW*, vol. 2, pp. 1–22, 2018.

[7] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc NIPS*, 2017, pp. 4765–4774.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? explaining the predictions of any classifier," in *Proc. ACM Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[9] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau, "Visual analytics in deep learning: An interrogative survey for the next frontiers," *IEEE Trans on Visualization and Computer Graphics*, vol. 25, no. 8, pp. 2674–2693, 2018.

[10] A. Abdul, J. Vermeulen, D. Wang, B. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: A hci research agenda," in *ACM CHI*, 2018, pp. 1–18.

[11] S. Amershi, M. Chickering, S. Drucker, B. Lee, P. Simard, and J. Suh, "Modeltracker: Redesigning performance analysis tools for machine learning," in *ACM CHI*, 2015, pp. 337–346.

[12] M. Kahng, P. Y. Andrews, A. Kalro, and D. Chau, "Activis: Visual exploration of industry-scale deep neural network models," *IEEE Trans on Vis. and Computer Graphics*, vol. 24, no. 1, pp. 88–97, 2017.

[13] F. Hohman, A. Head, R. Caruana, R. DeLine, and S. Drucker, "Gamut: A design probe to understand how data scientists understand machine learning models," in *ACM CHI*, 2019, pp. 1–13.

[14] F. Hohman, H. Park, C. Robinson, and D. H. P. Chau, "Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations," *IEEE Trans on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 1096–1106, 2019.

[15] H.-F. Cheng, R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper, and H. Zhu, "Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders," in *Proc ACM CHI*, 2019, pp. 1–12.

[16] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[17] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide 1st ed., Springer Int'l.*, 2017.

[18] H. Shen, H. Jin, Á. A. Cabrera, A. Perer, H. Zhu, and J. I. Hong, "Designing alternative representations of confusion matrices to support non-expert public understanding of algorithm performance," *Proc. ACM on HCI*, vol. 4, pp. 1–22, 2020.

[19] R. Moraffah, M. Karami, R. Guo, A. Raglin, and H. Liu, "Causal interpretability for machine learning-problems, methods and evaluation," *ACM KDD Explorations*, vol. 22, no. 1, pp. 18–33, 2020.

[20] Y. Ming, H. Qu, and E. Bertini, "Rulematrix: Visualizing and understanding classifiers with rules," *IEEE Trans on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 342–352, 2018.

[21] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[22] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with shapley-value-based explanations as feature importance measures," *Proc. ICML*, 2020.

[23] T. Hastie and R. Tibshirani, *Generalized Additive Models*. CRC Press, 1990, vol. 43.

[24] B. Kim, C. Rudin, and J. A. Shah, "The bayesian case model: A generative approach for case-based reasoning and prototype classification," in *Proc NIPS*, 2014, pp. 1952–1960.

[25] J. Pearl and D. Mackenzie, *The Book of Why: the New Science of Cause and Effect*. Basic Books, 2018.

[26] B. Glymour and J. Herington, "Measuring the biases that matter: The ethical and casual foundations for measures of fairness in algorithms," in *Proc Conf. on Fairness, Accountability, and Transparency*, 2019, pp. 269–278.

[27] Y. Wu, L. Zhang, X. Wu, and H. Tong, "Pc-fairness: A unified framework for measuring causality-based fairness," in *Proc NIPS*, 2019, pp. 3399–3409.

[28] J. Zhang and E. Bareinboim, "Fairness in decision-making—the causal explanation formula," in *AAAI Artificial Intelligence*, 2018.

[29] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Fairness through causal awareness: Learning causal latent-variable models for biased data," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 2019, pp. 349–358.

[30] J. R. Loftus, C. Russell, M. J. Kusner, and R. Silva, "Causal reasoning for algorithmic fairness," *arXiv arXiv:1805.05859*, 2018.

[31] M. J. Kusner, C. Russell, J. R. Loftus, and R. Silva, "Causal interventions for fairness," *arXiv preprint arXiv:1806.02380*, 2018.

[32] A. Khademi, S. Lee, D. Foley, and V. Honavar, "Fairness in algorithmic decision making: An excursion through the lens of causality," in *ACM World Wide Web*, 2019, pp. 2907–2914.

[33] B. Dietvorst, J. Simmons, and C. Massey, "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even

slightly) modify them," *Management Science*, vol. 64, no. 3, pp. 1155–1170, 2018.
[34] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson, "The what-if tool: Interactive probing of machine learning models," *IEEE Trans on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 56–65, 2019.
[35] O. Gomez, S. Holter, J. Yuan, and E. Bertini, "Vice: visual counterfactual explanations for machine learning models," in *Proc. Intern. Conference on Intelligent User Interfaces*, 2020, pp. 531–535.
[36] J. Krause, A. Perer, and K. Ng, "Interacting with predictions: Visual inspection of black-box machine learning models," in *Proc. ACM CHI*, 2016, pp. 5686–5697.
[37] J. Wang and K. Mueller, "The visual causality analyst: An interactive interface for causal reasoning," *IEEE Trans on Visualization and Computer graphics*, vol. 22, no. 1, pp. 230–239, 2015.
[38] ——, "Visual causality analysis made practical," in *2017 IEEE Conf. on Visual Analytics Science and Technology (VAST)*, 2017, pp. 151–161.
[39] R. Ryan and E. Deci, "Self-determination theory and the role of basic psychological needs in personality and the organization of behavior." in *Handbook of Personality: R&T*, 2008, pp. 654–678.
[40] J. Schell, *The Art of Game Design: Book of Lenses*. CRC Press, 2008.
[41] J. Yan, Z. Gu, H. Lin, and J. Rzeszotarski, "Silva: Interactively assessing machine learning fairness using causality," in *Proc. ACM CHI*, 2020, pp. 1–13.
[42] X. Xie, F. Du, and Y. Wu, "A visual analytics approach for exploratory causal analysis: Exploration, validation, and applications," *IEEE Trans on Visualization and Computer Graphics*, 2020.
[43] Y. Onoue, K. Kyoda, M. Kioka, K. Baba, S. Onami, and K. Koyamada, "Development of an integrated visualization system for phenotypic character networks," in *2018 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 2018, pp. 21–25.
[44] H. Natsukawa, E. R. Deyle, G. M. Pao, K. Koyamada, and G. Sugihara, "A visual analytics approach for ecosystem dynamics based on empirical dynamic modeling," *IEEE Transactions on Visualization and Computer Graphics*, 2020.
[45] J. Pearl, *Causality: Models, Reasoning & Inference, 2nd Ed.* Cambridge University Press, 2013.
[46] P. M. Bentler and D. G. Weeks, "Linear structural equations with latent variables," *Psychometrika*, vol. 45, no. 3, pp. 289–308, 1980.
[47] R. H. Hoyle, *Structural equation modeling: Concepts, issues, and applications*. Sage, 1995.
[48] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in Genetics*, vol. 10, p. 524, 2019.
[49] X. Shen, S. Ma, P. Vemuri, and G. Simon, "Challenges and opportunities with causal discovery algorithms: Application to alzheimer's pathophysiology," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.
[50] S. Tople, A. Sharma, and A. Nori, "Alleviating privacy attacks via causal learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9537–9547.
[51] M. Bostock, V. Ogievetsky, and J. Heer, "D^3 data-driven documents," *IEEE Trans on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.
[52] J. Ellson, E. Gansner, L. Koutsofios, S. C. North, and G. Woodhull, "Graphviz—open source graph drawing tools," in *International Symposium on Graph Drawing*. Springer, 2001, pp. 483–484.
[53] D. Harrison Jr and D. L. Rubinfeld, "Hedonic housing prices and the demand for clean air," 1978.
[54] J. W. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes, "Using the adap learning algorithm to forecast the onset of diabetes mellitus," in *Proc. of the Annual Symposium on Computer Application in Medical Care*, 1988, p. 261.
[55] Z. Buçinca, P. Lin, K. Z. Gajos, and E. L. Glassman, "Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems," in *Proc. ACM IUI*, 2020, pp. 454–464.
[56] C. Rudin and J. Radin, "Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition," *Harvard Data Science Review*, vol. 1, no. 2, 2019.
[57] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information." *Psychological review*, vol. 101, no. 2, p. 343, 1994.
[58] Z. Zhang, K. T. McDonnell, and K. Mueller, "A network-based interface for the exploration of high-dimensional data spaces," in *2012 IEEE Pacific Visualization Symposium*, 2012, pp. 17–24.

Md Naimul Hoque is currently a PhD student at the College of Information Studies, University of Maryland, College Park. Previously, he obtained an M.S. in Computer Science degree from Stony Brook University. His current research interests include explainable AI, visual analytics, and human-computer interaction. For more information, see https://naimulh0que.github.io

Klaus Mueller has a PhD in computer science and is currently a professor of computer science at Stony Brook University and is a senior scientist at Brookhaven National Lab. His current research interests include explainable AI, visual analytics, data science, and medical imaging. He won the US National Science Foundation Early CAREER Award, the SUNY Chancellor's Award for Excellence in Scholarship & Creative Activity, and the IEEE CS Meritorious Service Certificate. His 200+ papers were cited over 10,000 times. For more information, see http://www.cs.sunysb.edu/ mueller