SHORT-PAPER

CausalExplain: Causal Explanations of Black-box Models with Training Data Subsets

ARMAN ASHKARI, The University of Utah, Salt Lake City, UT, United States

EL KINDI REZIG, The University of Utah, Salt Lake City, UT, United States

Published: 22 June 2025

SIGMOD/PODS '25: International Conference on Management of Data
June 22 - 27, 2025
Berlin, Germany

Conference Sponsors:
SIGMOD

https://doi.org/10.1145/3722212.3725087
ISBN: 9798400715648

CausalExplain: Causal Explanations of Black-box Models with Training Data Subsets

Arman Ashkari
The University of Utah
Salt Lake City, Utah, United States
arman.ashkari@utah.edu

El Kindi Rezig
The University of Utah
Salt Lake City, Utah, United States
elkindi.rezig@utah.edu

Abstract

AI models are becoming increasingly opaque, prompting significant efforts to explain their predictions. However, data-based explanations—i.e., explaining predictions using training data points—remain largely underexplored. Such explanations are essential for debugging training data and gaining deeper insights into model behavior. We introduce a demo of CausalExplain, a model-agnostic system designed to trace the predictions of popular machine learning (ML) models back to their corresponding training data points. For any given prediction, CausalExplain provides explanations based on predicates from the training data that *caused* the outcome. These explanations are derived from the causal DAG of the training data, with CausalExplain presenting the user with the top-k data-centric explanations for a given query row. In this demo, participants will explore how causal DAGs support the generation of such explanations across various datasets and underlying ML models.

CCS Concepts

• **Information systems → Data cleaning**.

Keywords

data debugging; explainable AI; causal reasoning

ACM Reference Format:
Arman Ashkari and El Kindi Rezig. 2025. CausalExplain: Causal Explanations of Black-box Models with Training Data Subsets. In *Companion of the 2025 International Conference on Management of Data (SIGMOD-Companion '25), June 22–27, 2025, Berlin, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3722212.3725087

1 Introduction

Explainable AI (XAI) has been an active research area in recent years [6, 7]. Its goal is to provide users with methods to reason about the predictions AI models make for a given data point.

Model-agnostic methods such as Shapley values [7, 10] are feature-based, i.e., they focus on quantifying the importance of the input features in producing a given prediction. However, they do not trace a prediction made on a data point back to the training data subsets that *caused* that prediction. This causal data-based and model-agnostic XAI is what we address in CausalExplain [2].

SIGMOD-Companion '25, Berlin, Germany

ACM ISBN 979-8-4007-1564-8/2025/06
https://doi.org/10.1145/3722212.3725087

Test row: t_{test}

age	education	marital-status	occupation	hours-per-week	gender	...	income (Prediction)	income (True Label)
YoungAdult	Bachelors	Never-married	Adm-clerical	Medium	Female	...	<=50K	>50K

S1: age = "YoungAdult" ∧ occupation = "Adm-clerical" ∧ gender = "Female"

age	education	marital-status	occupation	hours-per-week	gender	...	income
YoungAdult	Bachelors	Never-married	Adm-clerical	High	Female	...	<=50K
YoungAdult	Some-college	Never-married	Adm-clerical	High	Female	...	<=50K
YoungAdult	Bachelors	Divorced	Adm-clerical	High	Female	...	<=50K
YoungAdult	Bachelors	Never-married	Adm-clerical	Medium	Female	...	<=50K
...	...	...	...	...	...	...	...

S2: (education = "Bachelors" ∧ occupation = "Adm-clerical") ∨ hours-per-week = "Medium"

age	education	marital-status	occupation	hours-per-week	gender	...	income
YoungAdult	Bachelors	Never-married	Adm-clerical	Low	Female	...	<=50K
MiddleAged	HS-grad	Separated	Other-service	Medium	Male	...	<=50K
Old	Bachelors	Married-civ-spouse	Adm-clerical	High	Male	...	<=50K
YoungAdult	Bachelors	Never-married	Adm-clerical	High	Female	...	<=50K
...	...	...	...	...	...	...	...

Figure 1: An example of explanation sets generated by CausalExplain for a test row on an *Income level* prediction algorithm using the *Adult Income* dataset [4]. The predicates use relevant features and their values from the test row to extract the subsets.

CausalExplain answers the following question: given a test row t and a trained ML model M, can we trace the prediction of M on t to training data subsets whose removal would result in a different prediction for t?

Naively, we would need to consider all subsets of the training data, then re-train the model without each subset, and then check if the prediction on t is different or the same. However, this approach is too expensive and results in re-training the model 2^n times (n is the number of rows in the training data). Alternatively, existing approaches [15] that don't require model re-training are not model-agnostic and have to be tuned for specific ML models.

What we want to do instead is to re-train the model only on subsets of the data that are likely going to change the prediction of t. To this end, causal DAGs come to the rescue [8]. Causal DAGs encode causal relationships between the attributes and a given outcome (prediction) [5, 8]. Causal DAGs can be provided by experts or automatically discovered from the data [13]. CausalExplain only requires a partial causal DAG specification [6].

Example 1.1: Sam, a data analyst at a fin-tech company, is building a classifier on the *Adult Income* dataset [4] to classify an individual's *Income level* ("≤50K" or ">50K") given personal information. He selects an ML algorithm, e.g., SVM, neural network, trains a model on the training set, and uses the model to predict the *Income level* of individuals from the test set. He notices an individual is misclassified (Test row in Figure 1) and wants to investigate which data points in the training data *caused* this prediction.

In the example in Figure 1, t_{test} is misclassified as "≤50K". In Figure 1, tables $S1$ and $S2$ contain two data-based *explanation sets* (subsets of the training data whose removal *cause* the prediction

for a given test row to change) for the prediction made for t_{test}. $S1$, as indicated by the predicate, is a subset of training data having *age*="YoungAdult" ∧ *occupation*="Adm-clerical" ∧ *gender*="Female", the same as t_{test} but with a majority label "≤50K". A model trained with the training data excluding $S1$ predicts ">50K" for t_{test}. $S2$ is another subset of training data with a majority label "≤50K". As the predicate describes, it is a subset of training data with (*education*="Bachelors" ∧ *occupation*="Adm-clerical") ∨ *hours-per-week*="Medium". Eliminating $S2$ from the training data reverses the prediction for t_{test}. From the explanation set $S1$, Sam can conclude that the historical training data predominantly consists of young female clerks with income ≤50K and is not representative of the current population. Consequently, he can elect to include more training data of female clerks with income >50K. Note that there are many training data subsets that can flip the prediction for t_{test}. However, we are interested in subsets that are causally explainable using the causal DAG and feature values of the test row.

Usability: Explanation sets can be useful in many use cases, for example, (1) **AI explainability:** predicates describing the explanation sets can help reason about the model's output (prediction), and (2) **Data Debugging:** erroneous output can be fixed by tracing and fixing the training data subsets causing the wrong predictions.

Related Work: Recent data-based XAI methods [15] assume a white-box model, i.e., the model's parameters are exposed and can only quantify the effect of removing one tuple at a time from the training data. [5] views ML algorithms as black-box systems and develops data profiles to identify causally verified root causes of system malfunctions. [14] is a recent system that generates black-box explanations using a provenance graph that keeps track of data transformations in the model. Counterfactual explanations of ML models [12] is a related line of work where the focus is to find counterfactual explanations of ML models' predictions. For instance, [6] explains black-box decision-making systems using counterfactuals and quantifies the causal influence of features on an algorithm's decisions. Unlike existing techniques, CausalExplain is the first system that leverages the causal DAG of the data to generate model-agnostic data-based explanations for a given prediction.

2 System Overview

CausalExplain uses causal graphs to efficiently explore the search space of explanation sets. In a causal graph, directed edges represent cause-and-effect relationships among the variables (features or outcomes in this case), resulting in a directed acyclic graph (DAG) (Figure 3a). A causal path (a sequence of directed edges) represents the chain of cause-and-effect from one variable to another that are causally dependent. CausalExplain exploits the cause-and-effect relationships to cause a different outcome (prediction) for the given test row, as explained later in **Profile generation** (Section 2.3). The workflow of CausalExplain is presented in Figure 2. It consists of an offline preprocessing step and an online query step. Next, we present the main components of the architecture in Figure 2.

2.1 Proprocessing step

① **Profiles.** CausalExplain takes as input the training set, the test set, and the causal DAG. Then, it generates *Feature Profiles* from the causal DAG by exploring different paths that lead to the outcome variable (details of this process are in Section 2.3). Feature Profiles or, in short, Profiles are predicates expressed in the disjunctive normal form. For example, a Profile can be $\phi = (age = \psi_1 \wedge education = \psi_2) \vee (workclass = \psi_3)$, where ψ_1, ψ_2, and ψ_3 are placeholders. For a test row t with *age* = "Old", *education* = "Bachelors", *workclass* = "Private", a possible instantiation of ϕ is $\phi(t) = (age =$"Old" ∧ *education* = "Bachelors") ∨ (*workclass* = "Private"). Profiles describe subsets in the training data. This setting is better suited to categorical features, but it can also accommodate numerical features by discretization.

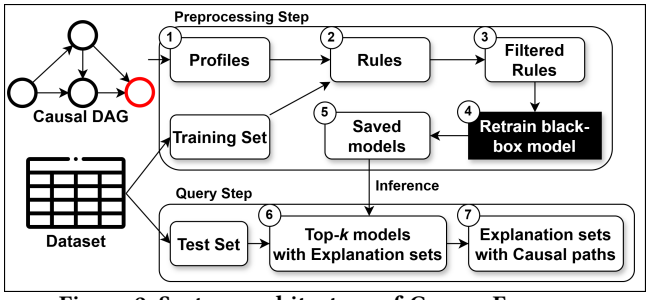

Figure 2: System architecture of CausalExplain.

② **Rules.** Next, CausalExplain instantiates every Profile generated from the causal DAG with all possible values in the feature domain, creating a bag of instantiated Profiles or Rules. It also calculates the *Support*, defined as the fraction of the training set described by each rule.

We define removing a subset of the training data as an *Intervention*. Interventions are likely to change the training data distribution and, consequently, the retrained models, where the prediction for test rows other than the intended test row may also change, resulting in side effects. To quantify side effects, we introduce *Model similarity*, defined as the prediction similarity of the original black-box model and the retrained model. Higher *Model similarity* and less *Support* are preferred because it means fewer side effects, and explanation sets with smaller *Support* are easier to explain. However, unlike *Support*, *Model similarity* can only be computed after retraining.

③ **Filtered Rules.** CausalExplain filters the Rules, first using an upper threshold on *Support* and then by the Average Treatment Effect (ATE) [8]. We compute the ATE on the training data after removing the subset described by the rule and compare it with the baseline ATE on the original training data. We keep the rules with a substantial decrease in ATE (controlled by an ATE change threshold $\Delta\ ATE$) after the intervention. This is because the ATE measures the strength of a causal relation between the features and the outcome. A decrease in ATE indicates that the removed subset (the intervention) is important for the causal relation, and removing it weakens the relation. So, such a rule is a potential candidate for changing the prediction (outcome) for test rows that share the same feature values as the rule. No change or an increase in ATE means the removed subset has little or no contribution to the causal relation, so we discard the rule. The thresholds for *Support* and the $\Delta\ ATE$ are tunable parameters. Users can adjust them to keep the number of rules manageable.

④ **Model re-training.** CausalExplain retrains a model for each rule in the filtered rule set. The retraining iterations are independent and highly parallelizable, making them ideal for distributed settings. CausalExplain collects key model statistics (e.g., Model Accuracy, *Model Similarity*) and stores the retrained models (Figure 2⑤), allowing them to be invoked for making inferences during the online Query Step.

2.2 Query step

In the query step, CausalExplain loads the models from the preprocessing step (Figure 2⑥). Given a test row, CausalExplain runs inference on the saved models and lists the rules from models whose predictions differ from the original model's prediction (the model that was trained using all the training data). Then, CausalExplain filters the rules based on feature overlap with the test row and ranks the explanation sets by their *Interestingness* score, which balances higher *Model similarity* (fewer side effects) with lower *Support* (smaller explanation sets). CausalExplain then highlights the causal paths in the DAG from which the rule was derived and returns their corresponding explanation sets (Figure 2⑦).

2.3 Profile generation

Now we discuss how CausalExplain leverages the cause-and-effect relationships using the causal DAG. Directed paths in DAGs imply *Reachability* (e.g., *native-coutnry* to *occupation* in Figure 3a). This property constitutes a *Partial Ordering* $\leq_p$ on the vertices (e.g., *native-coutnry* $\leq_p$ *occupation*). However, if two vertices are unreachable from one another (e.g., *occupation* and *workclass*), there is no ordering; hence, the ordering is *Partial*. Multiple causal paths exist from *education* to *income*; some of them have *occupation* as an intermediate node (refer to Figure 3b). We consider features in the same causal path for conjunctions. Conversely, since *occupation* and *workclass* are unreachable from one another, there are no causal paths that contain both. They are in different causal paths to *income* (refer to Figure 3c). So, we don't consider such features for conjunctions. However, they may be considered for disjunction. The partial ordering distinguishes between these cases. So, We keep conjunctions ordered by $\leq_p$, as it allows efficient generation of conjunctions and disjunctions.

Generating conjunctions: We begin with single-term conjunctions. In Figure 3d, *marital-status* $\leq_p$ *education* $\leq_p$ *income*, so we generate the conjunction (*marital-status* $\wedge$ *education*). In the same process, we can go from 2-sized conjunctions to 3-sized conjunctions (*marital-status* $\wedge$ *education* $\wedge$ *workclass*) (Figure 3e). We extend the length by one until we reach the longest path in the DAG. Limiting the length (e.g., a user-provided threshold) can result in further pruning of the number of conjunctions.

Generating disjunctions: We form disjunctions by combining conjunctions that lie on distinct causal paths. Such features do not form conjunctions since they influence the target independently. For example, in Figure 3d and 3e, replacing *marital-status* or *workclass* with *relationship* removes any ordering, suggesting a disjunction. In Figure 3f, (*age* $\wedge$ *workclass*) and (*native-country* $\wedge$ *occupation*) form a disjunction as they cannot be collapsed into a single conjunction. We iteratively extend disjunctions by one term until no further combinations are possible, optionally limiting term count to prune *Profiles*.

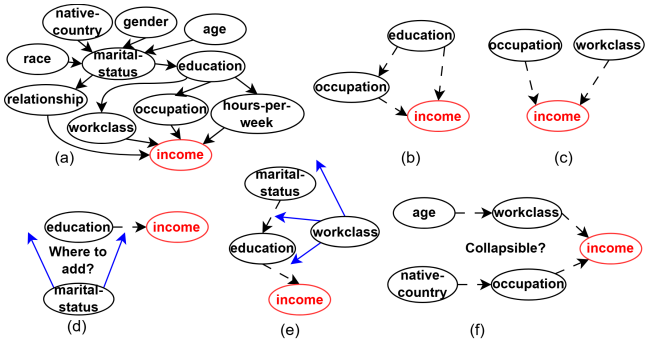

Figure 3: (a) Causal DAG for the *Adult Income* dataset. (b) Causal paths from *education* to *income*, *education* to *occupation*, and *occupation* to *income*. (c) Causal paths from *occupation* and *workclass* to *income*. (d) Extending the single-term conjunction (*education*) with *marital-status*. (e) Extending (*marital-status* $\wedge$ *education*) with *workclass*. (f) Determining the validity of the disjunction (*age* $\wedge$ *workclass*) $\vee$ (*native-country* $\wedge$ *occupation*). Dashed arrows are causal paths.

3 Demonstration Scenario

This demonstration showcases how CausalExplain enables data-based explanations for ML model predictions and supports training data debugging. Participants, acting as data scientists, will use CausalExplain to trace predictions back to training data and identify and remove problematic subsets to improve model performance.

Figure 4 illustrates the UI of the last step of CausalExplain (explanation sets with causal paths). On the left-hand side pane, we can see that the user trained a neural network model and specified key parameters to generate the explanations (e.g., conjunction max length). An example scenario is as follows: (1) The user spots an incorrect prediction of the model on a test row (Figure 4 ①); (2) The user uses CausalExplain to generate a set of explanations. CausalExplain then generates a list of explanations (Figure 4 ② and Figure 4 ③); (3) The user can then select specific explanations to see a preview of their data subset as well as the paths from the causal DAG that were used to derive the selected explanation (Figure 4 ④).

We demonstrate the operations of CausalExplain over the following datasets and ML models.

Datasets: We consider three real-world datasets: (1) the *Adult Income* dataset [4] consisting of 48,842 data points, where the ML algorithm(s) classify an individual's *Income level* given personal information, (2) the *StackOverflow Annual Developer Survey 2018* [1] of 98,855 data points, where the classifiers are trained to predict developer's *Salary level* based on personal information and development skill-set, and (3) the *COMPAS* dataset [9] with 7,214 data points, where the classification task is to predict whether an individual will recommit a crime within two years (*Recidivism*) from criminal history and demographic information.

Black-box ML algorithms: CausalExplain is model agnostic and supports most ML algorithms. In this demo, we showcase the popular algorithms: (1) Logistic Regression, (2) SVM, (3) XGBoost, (4) AdaBoost, (5) Random Forest, and (6) Multi-layer neural nets.

The audience will engage with CausalExplain via the following scenarios:

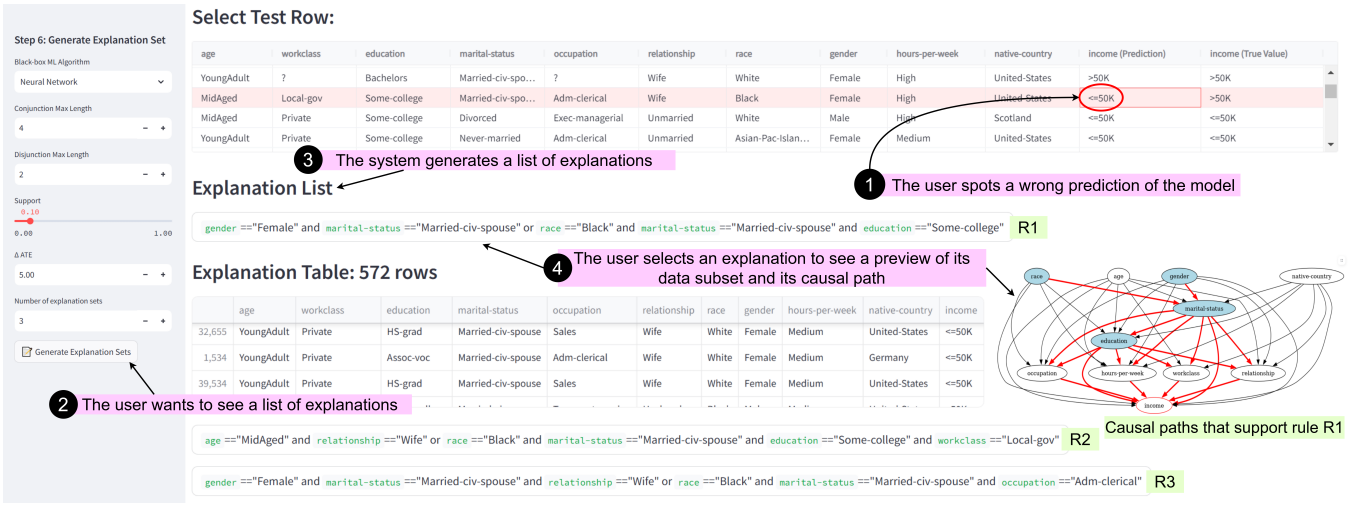

Figure 4: A screenshot of CausalExplain at the query step. Participants will have the opportunity to train a series of ML models on various datasets and generate explanations using CausalExplain, which are grounded in the causal DAG.

Causal DAG and parameter specification: CausalExplain requires the specification of a causal DAG or a partial causal DAG. The audience can import the DAG through a text file. There are causal discovery methods and tools, e.g., the DoWhy package [11] and the CausalNex package [3], to generate causal DAGs for datasets. The audience can specify the parameters according to their computation budget. They can start with low thresholds and gradually increase them to achieve desired performance.

Understanding the rule's contribution to the causal relation: The participants will be able to explore the rules and understand their contribution to the strength of the causal relation with the outcome variable (prediction) from the ATE. It will enable the participants to identify influential rules and training data subsets.

The retraining phase runs offline. For demonstration purposes, we will load the preprocessed models.

Explanation Set Generation: To illustrate the usability of CausalExplain, in the query phase, the audience can select rows from the test set and generate the top-k explanation sets. The explanation sets are displayed as predicates, as well as the training data subsets, and are sorted by the *Interestingness* score. The participants can also rank them by *Support*, *Model similarity*, predicate length, etc, to facilitate their explainability or data debugging task. Additionally, they can select an explanation set, load the retrained model corresponding to the explanation set, and compare the prediction of the retrained model with that of the original model and the ground truth to further streamline the data debugging task.

Causal explanations: Each explanation set corresponds to a rule that induces causal paths in the DAG. CausalExplain highlights these paths, helping participants understand explanations by examining the activated causal paths.

Acknowledgments

This project was supported by the University Research Committee (URC) at The University of Utah. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the URC, the Vice President for Research Office, or The University of Utah.

References

[1] [n. d.]. Stack Overflow Developer Survey 2018. https://survey.stackoverflow.co/2018

[2] Arman Ashkari and El Kindi Rezig. 2024. Can Causal DAGs Generate Data-based Explanations of Black-box Models?. In 2024 IEEE International Conference on Big Data (BigData). 5739–5744. doi:10.1109/BigData62323.2024.10825482

[3] Paul Beaumont, Ben Horsburgh, Philip Pilgerstorfer, Angel Droth, Richard Oentaryo, Steven Ler, Hiep Nguyen, Gabriel Azevedo Ferreira, Zain Patel, and Wesley Leong. 2021. CausalNex. https://github.com/quantumblacklabs/causalnex

[4] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[5] Sainyam Galhotra, Anna Fariha, Raoni Lourenço, Juliana Freire, Alexandra Meliou, and Divesh Srivastava. 2022. Dataprism: Exposing disconnect between data and systems. In Proceedings of the 2022 International Conference on Management of Data. 217–231.

[6] Sainyam Galhotra, Romila Pradhan, and Babak Salimi. 2021. Explaining black-box algorithms using probabilistic contrastive counterfactuals. In Proceedings of the 2021 International Conference on Management of Data. 577–590.

[7] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[8] Judea Pearl. 2009. Causality. Cambridge university press.

[9] ProPublica. [n. d.]. Compas recidivism risk score data and analysis. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

[10] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 1135–1144.

[11] Amit Sharma and Emre Kiciman. 2020. DoWhy: An End-to-End Library for Causal Inference. arXiv preprint arXiv:2011.04216 (2020).

[12] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan Hines, John Dickerson, and Chirag Shah. 2024. Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. ACM Comput. Surv. 56, 12, Article 312 (Oct. 2024), 42 pages. doi:10.1145/3677119

[13] Alessio Zanga, Elif Ozkirimli, and Fabio Stella. 2022. A Survey on Causal Discovery: Theory and Practice. International Journal of Approximate Reasoning 151 (2022), 101–129. doi:10.1016/j.ijar.2022.09.004

[14] Jiachi Zhang, Wenchao Zhou, and Benjamin E. Ujcich. 2024. Provenance-Enabled Explainable AI. Proc. ACM Manag. Data 2, 6, Article 250 (Dec. 2024), 27 pages. doi:10.1145/3698826

[15] Jiongli Zhu, Romila Pradhan, Boris Glavic, and Babak Salimi. 2022. Generating Interpretable Data-Based Explanations for Fairness Debugging using Gopher. In Proceedings of the 2022 International Conference on Management of Data. 2433–2436.