Demonstration of Generating Explanations for Black-Box Algorithms Using Lewis

Paul Y. Wang
pywang@ucsd.edu
University of California, San Diego
La Jolla, California, USA

Sainyam Galhotra
sainyam@uchicago.edu
University of Chicago
Illinois, USA

Romila Pradhan
rpradhan@ucsd.edu
University of California, San Diego
La Jolla, California, USA

Babak Salimi
bsalimi@ucsd.edu
University of California, San Diego
La Jolla, California, USA

ABSTRACT

Explainable artificial intelligence (XAI) aims to reduce the opacity of AI-based decision-making systems, allowing humans to scrutinize and trust them. Unlike prior work that attributes the responsibility for an algorithm's decisions to its inputs as a purely associational concept, we propose a principled causality-based approach for explaining black-box decision-making systems. We present the demonstration of Lewis, a system that generates explanations for black-box algorithms at the global, contextual, and local levels, and provides actionable recourse for individuals negatively affected by an algorithm's decision. Lewis makes no assumptions about the internals of the algorithm except for the availability of its input-output data. The explanations generated by Lewis are based on probabilistic contrastive counterfactuals, a concept that can be traced back to philosophical, cognitive, and social foundations of theories on how humans generate and select explanations. We describe the system layout of Lewis wherein an end-user specifies the underlying causal model and Lewis generates explanations for particular use-cases, compares them with explanations generated by state-of-the-art approaches in XAI, and provides actionable recourse when applicable. Lewis has been developed as open-source software; the code and the demonstration video are available at lewis-system.github.io.

PVLDB Reference Format:
Paul Y. Wang, Sainyam Galhotra, Romila Pradhan, and Babak Salimi. Demonstration of Generating Explanations for Black-Box Algorithms Using Lewis. PVLDB, 14(12): 2787 - 2790, 2021.
doi:10.14778/3476311.3476345

1 INTRODUCTION

Algorithmic decision-making systems are increasingly being used to automate life-changing decisions and can lead to unequal distribution of benefits and risks across different segments of society. *Explainable artificial intelligence* (XAI) aims to address the opacity of these systems by providing human-understandable explanations of the process and outcomes of these systems (see [6] for a recent survey). Effective explanations should serve two objectives: (1) ensure different stakeholders that the system's decision rules are justifiable, and (2) provide users with an actionable recourse to change future outcomes of the algorithm [13]. In this work, we present explanation methods that conduct post factum system analysis of any black-box algorithm. Prior work in this context has focused on explaining an algorithm by a simple, interpretable *surrogate* model (e.g., decision trees) [11] or attributing *responsibility* of its decisions to its inputs [8]. These methods do not capture the causal influence between an algorithm's inputs and output, which can produce incorrect and misleading explanations [4]. *Counterfactual explanations*, on the other hand, minimally perturb an algorithm's inputs to obtain the desired outcome [10, 12]; however, due to the causal interaction between variables, these perturbations are not translatable into real-world interventions [2, 7].

We propose to demonstrate Lewis[1], a causality-based system that uses probabilistic contrastive counterfactuals for generating post-hoc explanations for black-box decision-making algorithms. Lewis reconciles the aforementioned objectives of XAI in two steps: (1) It provides insights into what *causes* an algorithm's decisions at the global, local and contextual (sub-population) levels in terms of three novel probabilistic measures– necessity score, sufficiency score and necessity and sufficiency score– that quantify the influence of attributes toward an algorithm's decision; (2) For individuals negatively impacted by the algorithm, Lewis generates actionable recourse to change the outcome of the model in future. A detailed description of the system is presented in our previous work [5]. Overall, this demonstration makes the following contributions:

- We present Lewis, an end-to-end system that generates explanations for black-box decision-making algorithms using novel measures (based on probabilistic contrastive counterfactuals) that have provable theoretical guarantees.
- The demonstration enables the users to understand the influence of different attributes not only at the global and local levels but also at a user-defined sub-population level (characterized as contextual level). Additionally, it presents actionable recourse for the individuals that suffered from negative outcome by the algorithm.
- We demonstrate that explanations generated by Lewis go beyond state-of-the-art approaches in XAI that capture correlation between attributes (e.g., SHAP [8], and LIME [11]).

Users will be able to observe first-hand the effect of sensitive attributes like race and gender on the behavior of well-known black-box loan prediction and recidivism software. This demonstration will also help users understand the inconsistencies of using non-causal explanation techniques and justify the importance of considering contrastive counterfactuals for explanations.

[1] Our system is named after David Lewis (1941–2001), who made significant contributions to modern theories of causality and explanations in terms of counterfactuals.

Proceedings of the VLDB Endowment, Vol. 14, No. 12 ISSN 2150-8097.
doi:10.14778/3476311.3476345

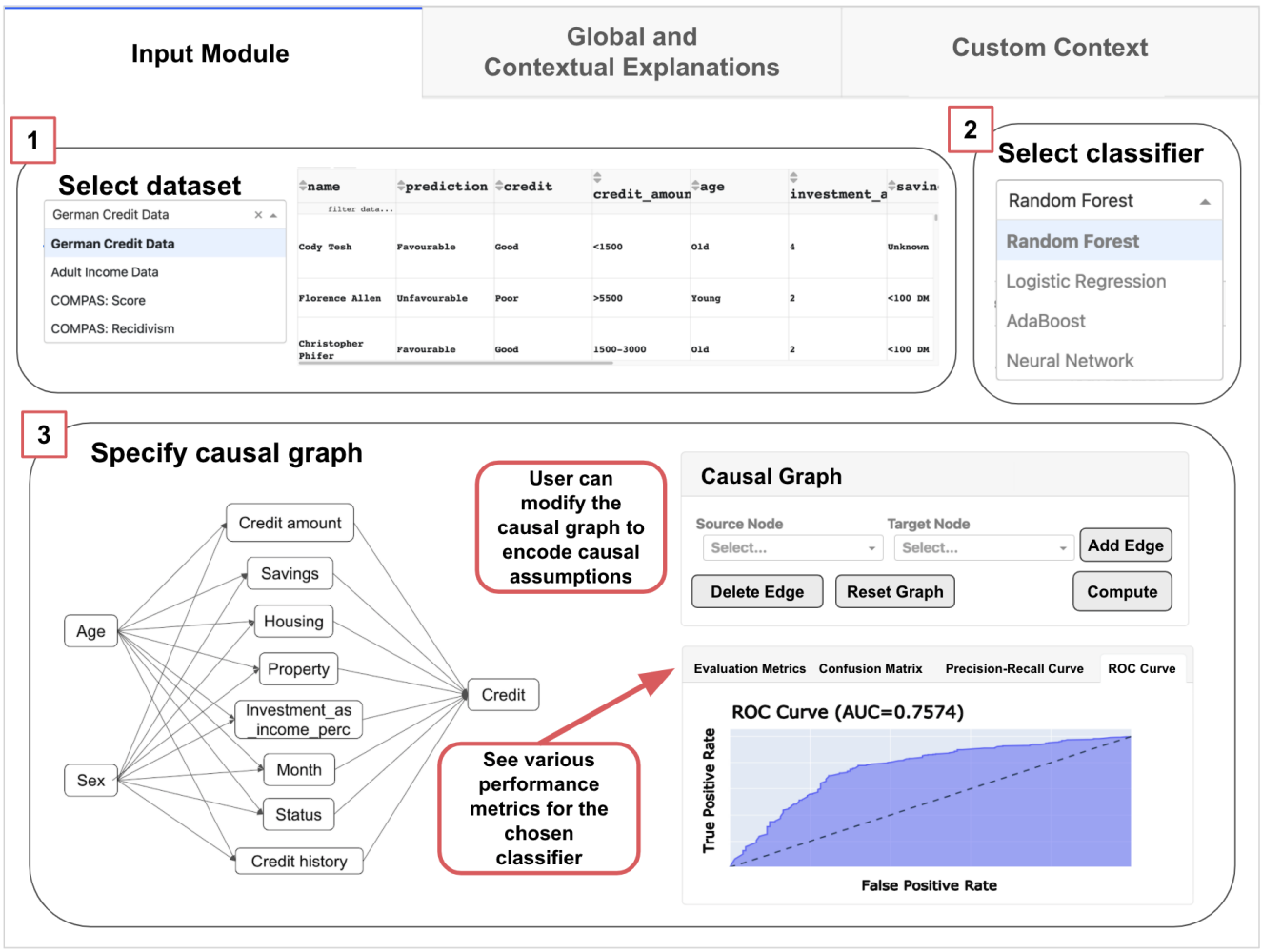

Figure 1: Overview of Lewis: (1, 2) The user selects from a list of datasets or specifies their own, and selects one of the black-box decision-making algorithms; (3) Users can specify the underlying causal graph by adding nodes representing attributes and edges representing causal relationship between attributes.

2 SYSTEM OVERVIEW

This section provides a brief overview of the internals of Lewis. Lewis is based on *probabilistic contrastive counterfactuals* of the following form: "For individual(s) with attribute(s) <actual-value> for whom an algorithm made the decision <actual-outcome>, the decision would have been <foil-outcome> with *probability* <score> had the attribute been <counterfactual-value>". While prior literature has established the difficulty in estimating such scores from observational data [9], we used probabilistic contrastive counterfactuals to define novel explanation scores that quantify the influence of an attribute on an algorithm's decision, and developed a mathematical framework to approximate these scores from historical data with provable guarantees. In the following, we explain these scores and discuss how they can be used for computing recourse for users negatively affected by the algorithm.

Explanation scores. Lewis considers the input dataset and the output of the black-box algorithm to evaluate explanation scores for a specific sub-population, captured as a context $\mathbf{k}$ (where $\mathbf{k} = \phi$ denotes the whole population). Given a context $\mathbf{k}$ of selected attribute values, and a pair of values x, x' for attribute X, Lewis computes three scores to quantify the influence of X on the algorithm decision: (a) necessity score of X is the "probability that for individuals with attributes $\mathbf{k}$, the algorithm's decision would be *negative* instead of *positive* had X been x' instead of x", (b) sufficiency score of X is the "probability that for individuals with attributes $\mathbf{k}$, the algorithm's decision would be *positive* instead of *negative* had X been x instead of x'", and (c) necessity and sufficiency score measures the probability that the algorithm responds in both ways. These scores are complementary in explaining the effect of changing X on the behavior of a black-box algorithm. While necessity score addresses the attribution of causal responsibility of an algorithm's decisions to X, sufficiency score addresses its tendency to produce

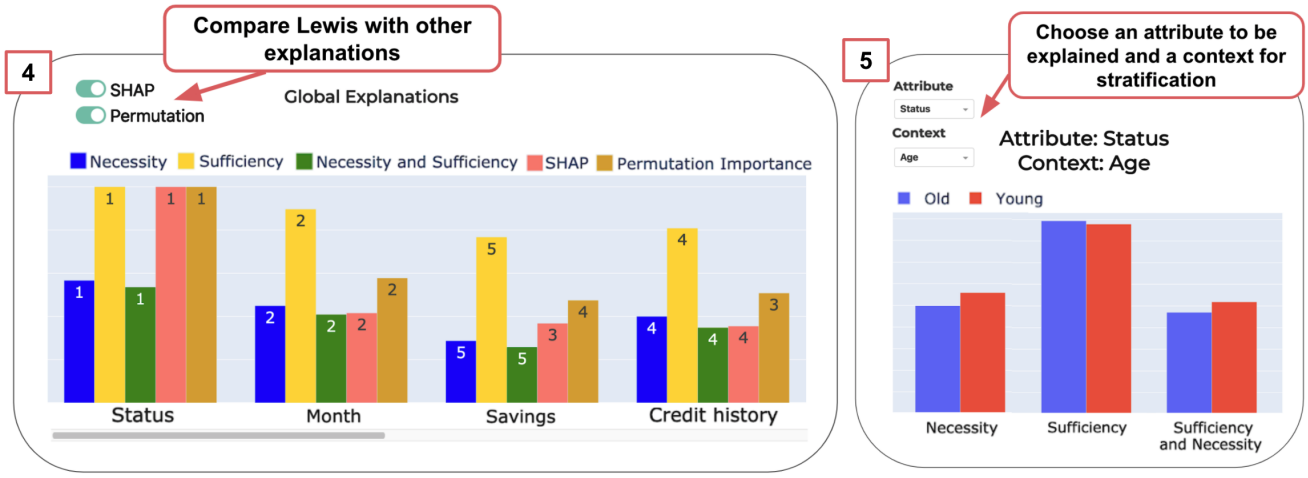

Figure 2: (4) Lewis's global explanation scores show the global behavior of the black-box algorithm with respect to each attribute. On top of the bars, we see the rankings of attributes as generated by the different approaches in XAI. (5) Contextual explanations show the effect of changing an attribute value on the behavior of the algorithm for a selected context.

the desired algorithmic outcome. In the absence of a context, the scores measure the *global* influence of X on the algorithm's decision. When all the attributes are selected as context, the scores measure the individual-level or *local* influence of X on the algorithm's decision. Finally, for a user-defined context, the scores measure the *contextual* influence of X on the algorithm's decision.

Counterfactual recourse. Algorithmic recourse consists of actionable changes an individual that received a negative decision may adopt to acquire a favorable decision in the future. Lewis computes recourse by searching for a minimal intervention on a pre-specified set of actionable attributes sufficient for a favorable outcome. In particular, Lewis constructs a linear program over input attributes of the algorithm to estimate the effect of changing the values of the actionable attributes and evaluate a recourse with minimum cost that have a high sufficiency score.

3 DEMONSTRATION OVERVIEW

Dataset. We will demonstrate the functionalities of Lewis on the German credit dataset [3] which consists of records of bank account holders with their personal and financial information. The prediction task classifies individuals as good or bad credit risks. We will also provide two additional datasets – Adult income [3] and COMPAS (recidivism) [1] – for users to explore other applications.

Black-box decision-making algorithm. We will demonstrate Lewis on the following classification algorithms: (1) random forest classifier, (2) logistic regression, (3) feed-forward neural networks, and (4) Adaboost classifier. Note that while Lewis is model agnostic, this selection is made available to obtain the outputs of the model.

Our demonstration will start with the user loading the dataset and inspecting it by submitting SQL queries or filtering the table natively. The user can select one of the supported classifiers for the prediction task (Figure 1, step 2), and visualize various performance metrics for the chosen classifier. To generate contrastive explanations, Lewis assumes access to the underlying causal assumptions corresponding to the dataset and allows users to encode these assumptions in the form of a causal graph (Figure 1, step 3). (Note that this graph can be learnt using any standard structure learning algorithm). For this demonstration, we pre-load the graph depicted in Figure 1 which stems from the belief that Age and Sex are exogenous attributes that are not affected by mutable attributes (e.g., an individual's age is not *caused by* the amount in their savings account or the location of their residence). Users can modify the causal graph by adding/deleting edges representing causal relationships among attributes (Figure 1, step 3).

Finally, users can explore the impact of attributes at global and contextual level (consisting of one or more individuals in the dataset). Depending upon the selected tab, we depict how explanations at the global, contextual, and local levels are visualized and how the user may compare explanations generated by Lewis to either SHAP [8] or LIME [11]. Additionally, we illustrate how Lewis generates causally motivated recourse for individuals who receive an unfavourable decision.

Visualizing explanations on local, contextual, and global levels. The strength of Lewis lies in its flexibility in selecting various contexts. This feature enables the user to determine the causal influence of an attribute on the algorithm's decision for an individual, a user-defined sub-population, or the entire population.

In Figure 2, we examine the explanations generated by Lewis at the global and contextual levels and compare the generated explanations with state-of-the-art XAI techniques. The global explanations (step 4) shows features with the greatest causal influences on a positive decision plotted with their necessity, sufficiency, and necessity and sufficiency scores. For this dataset, ranking of attributes is similar for SHAP. Figure 2(b), step 5 shows the explanation scores of 'Status' (required daily minimum in an account) with context '$\mathbf{k}$ =Age'. Higher necessity of status for younger individuals shows that poor status is more likely to reject a loan for younger as compared to older individuals.

Figure 3: A snapshot of local explanations generated by Lewis. (6) The user can select a sub-population (context) or a particular data point specified in the form of a SQL query as `SELECT * FROM data WHERE context` or filter the table in place; (7) shows the contextual or local explanations generated by Lewis for the use case selected in (6) along with feature rankings; (8) If the individual selected in (6) received a negative decision by the black-box algorithm, then Lewis allows the user to specify the set of actionable variables and a threshold for sufficiency, and generates appropriate recourse.

The limitations of existing approaches in XAI, however, is perhaps most prominently displayed when local explanations are generated (Figure 3, step 7). Consider the case of the selected individual who was classified as a bad credit risk by the random forest classifier. Lewis ranks age much higher than SHAP and Lime because of strong causal influence of age on credit history, thereby affecting loan decisions. In contrast, SHAP and Lime do not capture such dependencies and rank age lower than other attributes.

Generating actionable counterfactual recourse. Lewis generates recourse through minimal interventions which accurately reflect causal assumptions underlying the real world. Figure 3, step 8 shows the recourse panel which appears for an individual who received an unfavorable decision; the user specifies a set of actionable attributes for which Lewis generates counterfactual recourse. For the selected individual, we see that causal responsibility of the negative decision is mainly attributed to the status of their checking account and month (duration of the loan). To improve their chances of a favorable outcome, the user may wish to change these variables which are then selected as the set of actionable variables. Lewis then computes its recourse recommendation: increase status to $>$ 200 DM and reduce duration of the loan to $<$ 10 months.

REFERENCES

[1] 2016. Machine Bias https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Solon Barocas, Andrew D Selbst, and Manish Raghavan. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 80–89.

[3] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[4] Christopher Frye, Ilya Feige, and Colin Rowat. 2019. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. *arXiv preprint arXiv:1910.06358* (2019).

[5] Sainyam Galhotra, Romila Pradhan, and Babak Salimi. 2021. Lewis: Explaining Black-Box Algorithms Using Probabilistic Contrastive Counterfactuals. In *SIGMOD*.

[6] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[7] Amir-Hossein Karimi, Gilles Barthe, Borja Belle, and Isabel Valera. 2019. Model-agnostic counterfactual explanations for consequential decisions. *arXiv preprint arXiv:1905.11190* (2019).

[8] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*. 4765–4774.

[9] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[10] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 607–617.

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *KDD*. 1135–1144.

[12] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 10–19.

[13] Suresh Venkatasubramanian and Mark Alfano. 2020. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 284–293.