11th International Conference on Sensitivity Analysis of Model Output, April 23–25 2025, Grenoble, France

Combining Counterfactuals and Sensitivity Analysis: A New Approach to Explaining Black-Box Models

Xuefei, Lu
SKEMA Business School-Université Côte d'Azur, Suresnes, France.

Emanuele, Borgonovo
Bocconi Institute for Data Science and Analytics (BIDSA), Milan, Italy

Data-driven models increasingly support decision-making. However, their complexity poses challenges for human comprehension and troubleshooting and their lack of transparency can lead to unfair and biased decisions [2, 10]. To counteract the black box effect, explainable artificial intelligence (XAI) techniques are studied. One of the most commonly studied explanations is model key drivers, which can focus managerial attention on the most important factors during implementation [3]. Popular post-hoc explanation methods include SHapley Additive exPlanations (SHAP) [5] or Local Interpretable Model-agnostic Explanations (LIME) [9]. These methods focus on individual predictions and identify the features' contributions to a specific model decision. Recent works by [8] and [11] highlight the strong connection between post-hoc explanations and sensitivity analysis.

In the context of XAI, counterfactual analysis provides insights into how changes to one or more features of a given instance affect the model's prediction [12]. The application becomes even more important when the instance of interest is an individual looking for an explanation as to why the decision of an algorithm was positive or negative on their behalf. Consider the following situation. An individual, say Ms. X, is requesting a loan (or a certificate of admission) to a financial (educational) entity but gets denied. Then, Ms. X wishes to understand what she should change/improve about her characteristics to get admitted. Ms. X can look at a counterfactual, as the closest individual such that if she changed one or more of her features she would also get the loan/admission. One question that naturally emerges is which feature, if changed, would be most effective for Ms. X to achieve the desired outcome. However, [1, 6] argue that SHAP does not provide insight into what is important for the change in the above situation.

Alternatively, in a counterfactual framework, a commonly used index is the frequency of changes in a given feature when moving from Ms. X to her counterfactuals. A feature is deemed important if it is frequently modified [7]. However, counting provides a summary indication of importance. First, we cannot appreciate the magnitude of the impact. A feature may be frequently modified, but its impact on the change could be small. Second, we cannot appreciate the direction of impact and whether the feature is involved in interactions with the remaining variables. Also, when moving from Ms. X to her counterfactual, one needs to pay attention that no impossible points are attained, to avoid model predictions affected by extrapolation errors [4]. Without considering those aspects, explanations remain partial, leave the algorithmic decision opaque, and do not shed light on the actions to be taken.

In this work, we propose a novel approach combining counterfactual analysis and sensitivity analysis to explain the transition from the baseline to the counterfactual state. We apportion the change in model predictions moving from Ms. X to her counterfactual considering each feature's individual and interaction contributions. A data-driven algorithm is then introduced to study the transition, combining the search for the counterfactual and identification of the impossible point involved in the sensitivity measure calculation. The proposed method has been applied to a synthetic example and a series of datasets. Several novel insights were obtained from the two well-known datasets in the machine learning literature.

References:

[1] E. Albini, J. Long, D. Dervovic, and D. Magazzeni. Counterfactual shapley additive explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1054–1070, 2022.

[2] J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):5, Jan. 2018.

[3] T. G. Eschenbach. Spiderplots versus tornado diagrams for sensitivity analysis. *Interfaces*, 22:40–46, 1992.

[4] G. Hooker, L. Mentch, and S. Zhou. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31, 2021.

[5] S. M. Lundberg, P. G. Allen, and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4765–4774, 2017.

[6] J. Marques-Silva and X. Huang. Explainability Is *Not* a Game. *Communications of the ACM*, 67(7):66–75, July 2024.

[7] R. K. Mothilal, D. Mahajan, C. Tan, and A. Sharma. Towards Unifying Feature Attribution and Counterfactual Explanations: Different Means to the Same End. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 652–663, July 2021. arXiv:2011.04917 [cs].

[8] S. Razavi, A. Jakeman, A. Saltelli, C. Prieur, B. Iooss, E. Borgonovo, E. Plischke, S. L. Piano, T. Iwanaga, W. Becker, S. Tarantola, J. H. A. Guillaume, J. Jakeman, H. Gupta, N. Melillo, G. Rabitti, V. Chabridon, Q. Duan, X. Sun, S. Smith, R. Sheikholeslami, N. Hosseini, M. Asadzadeh, A. Puy, S. Kucherenko, and H. R. Maier. The future of sensitivity analysis: An essential discipline for systems modeling and policy support. *Environmental Modelling and Software*, 137:1–22, 2021.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?" explaining the predictions of any classifier. 2016.

[10] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019.

[11] C. A. Scholbeck, J. Moosbauer, G. Casalicchio, H. Gupta, B. Bischl, and C. Heumann. Bridging the gap between machine learning and sensitivity analysis. *ArXiv*, :2312.:1–14, 2023.

[12] S. Verma, J. Dickerson, and K. Hines. Counterfactual Explanations for Machine Learning: A Review, Oct. 2020. arXiv:2010.10596 [cs, stat].

[Xuefei Lu; SKEMA Business School]
[xuefei.lu@skema.edu – https://www.skema.edu/en/faculty-and-research/professors/xuefei-lu]