Using counterfactuals to overcome data bias and increase model fairness

Stefanache Cornel
AscentCore Labs
cornel.stefanache@ascentcore.com

Sara Vișovan
AscentCore Labs
sara.popa@ascentcore.com

Lucian Tudose
Technical University of Cluj Napoca
lucian.tudose@omt.utcluj.ro

***Abstract*—Explainable Artificial Intelligence (XAI) addresses the need of users whose interests, expectations, and demands regarding artificial systems call for a greater understanding of the decision and reasoning of the systems. Counterfactual explanation is a technique in XAI that provides a potential response for a user to understand the decision of a predictive model, and is capable of identifying the smallest feature change capable of changing a prediction. In this work, we use counterfactual explanations to prove and overcome the fragility and bias of a machine learning model, whether it is a white box (Decision Tree), a gray box (Random Forest), or a black box (Neural Network). Explainability and trust in AI are crucial considering the applicability found in various fields, such as medical diagnosis, self-driving cars, Online fraud detection, financial services, and others.**

***Keywords*—XAI, Fairness Intervention, Counterfactuals**

I. Introduction

Over the past few years, the adoption of AI models has skyrocketed, and more decisions have been made using these algorithms. Machine learning models provide good prediction results, but most of the time the predictions may not be interpretable. The wide adoption of such black box models along with the decisions they make that directly impact our lives has led to rising concerns about the fairness and trustworthiness of such models. In response, counterfactuals have recently been advanced as a promising solution to the XAI problem [1].

Prompted by concerns about the potential adverse consequences of advantages of digital technologies, including AI, within a Human Rights Framework, the Council of Europe's Committee of Experts on Human Rights commissioned a study in 2018 [2]. An important aspect of the study was conducted around the ethics of AI. Ethics in AI represents guarding against certain kinds of discrimination and, if possible, encoding the abstract concept of fairness into the system [3]. Even when researchers are trying to capture, encode, and program such guards into the trained models, other complex data patterns might capture the bias and make it invisible to such guards. A fair dataset will produce fair models, but the machine learning models are only as good as the data they are trained on: *"bias in, bias out"*.

The ambiguity in machine learning models is known as the black box problem. It is hard for a user to understand why a prediction was made, generating a lack of trust in the model. Using counterfactual explanations - the process of identifying the smallest change to the input data capable of changing a prediction - has been considered a critical post-hoc method that helps users understand the internals of model decisions and the prediction quality [4] [5]. For example, if an individual were denied a loan request, as a decision made by an AI model, it would be hard or impossible for the bank to explain the algorithm's decision. A counterfactual solution might be able to expose that the model would have had a different decision if the requestor had an increase in income of 2 dollars per month. A human being can achieve the same result manually by tweaking the input values and finding the minimal amount of changes to the values for the model to predict a different outcome, but this is a tedious process.

The goal of the learning system is to learn a generalized mapping between input and output data such that skillful predictions can be made for new instances drawn from the domain where the output variable is unknown. In supervised learning, the model will learn a mapping function from examples of inputs to examples of outputs. The model needs to be able to capture the relationship between the input examples and the target values and prove a balanced fit, avoiding overfitting and underfitting. Overfitting is a fundamental issue in supervised machine learning where the model learns the detail in the training data too well, which then prevents the model from being able to perfectly generalize unseen data based on the testing set [6]. Underfitting occurs when a model can neither learn the true relationships in the training dataset nor generalize to a new dataset [7]. Overfitting and underfitting are the two biggest causes of poor performance of machine learning algorithms or models.

Data is crucial for machine learning models and it determines the performance of a model. Collecting and preparing the dataset is one of the most essential parts while creating an machine learning/artificial intelligence project. Typically, large datasets lead to better classification performance and small datasets may trigger overfitting. It is common knowledge that too little training data results in a model with poor performance. An over-constrained model will underfit the small training dataset, whereas an under-constrained model, in turn, will likely overfit the training data, both resulting in poor performance [8]. Furthermore, balancing training data is an important part of data preprocessing. Data imbalance refers to when the classes in a dataset are not equally distributed, which can then lead to potential risks in the process of training a model.

The never seen data prediction is highly dependent on the training dataset, prediction model, and training parameters. Trying to explain the decision of a model using the training data sometimes can lead to erroneous insights. For example, Fig 1 shows the decision boundary of a classifier and the two (square) data points that need to be predicted that are located at equal distances from differently classified training entries but have different predicted outcomes. The paper seeks to reduce bias by reshaping the decision boundary using generated synthetic data. The reshaping can be achieved by introducing new entries into the training dataset that capture the fairness of some features (gender, race) but preserve the importance of features that do not encapsulate a personal characteristic (e.g., education level, hours worked per week, etc.).

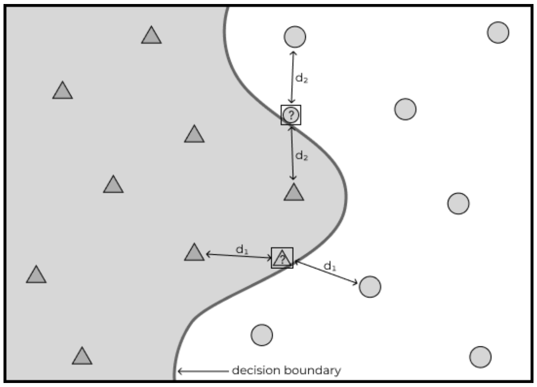

Fig. 1. Two predictions (square symbol) classified differently are equally distanced from training data (circle symbol) with different classes.

According to Verma S. et al. [9], the counterfactual solution is beneficial to the applicant whose life is impacted by the decision as it:

- Helps the applicant understand which of their attributes were drivers in decision making,
- Allows human factors to challenge the decision if they feel the decision was unfair, e.g., if one's race was crucial in determining the outcome, and
- Can help the data science team or machine learning model developers to identify, detect, and fix bugs or other issues.

II. Counterfactuals in XAI

In the field of XAI, counterfactuals provide interpretations to reveal what changes would be necessary in order to receive the desired prediction, rather than an explanation to understand why the current situation had a certain prediction [10] [11]. Most approaches in XAI focus on answering why a certain outcome was predicted by a model. Counterfactuals, however, try to answer this question by helping the user understand what features need to be changed in order to achieve a certain outcome [12] and thus infer which features influence the model the most. Counterfactual instances can be found by iterative perturbing of the input features of the test instance until the desired prediction or a prediction different than the original outcome is obtained. Counterfactuals are obtained by minimizing the distance (change) between the original input feature and the potential counterfactuals generated by the searching algorithm [13]. Findings [14] show that there is no single algorithm that is best for generating counterfactual explanations, as performance depends largely on the properties related to the dataset, model chosen for training, score, and the factual point specificities. The XAI methods focus on searching for a solution in the input space to capture an unfair decision on the part of the model and mitigate bias or simply emphasize a wrong decision-making behavior caused by a bad data structure [15]. High-quality counterfactuals can be used to tweak the current prediction model, leverage the undesired behavior, diminish the data bias, correct the model's decision-making, and achieve fairer decisions [16].

III. Scope

The scope of this research is to reduce the bias toward a specific feature or a set of features in the training dataset to increase the model fairness without altering the correlation value between the rest of the features on the outcome. The paper focuses on measuring the quality of the generated counterfactuals, their impact on the feature importance / correlation, and increasing the model's fairness with each batch of generated counterfactuals. The paper uses a Genetic Algorithm heuristic [17] to generate high-quality counterfactuals, subject to constraints to produce new synthetic data used to correct the decision. We also looked to measure and expose the model fragility by revealing such solutions that have a small change, defined by a threshold, that changes the outcome: e.g., if a loan application is denied with the explanation that the application would have been accepted if the applicant would have earned 1$ more. The proposed solution can be applied to all types of predictive models with the constraint that the generated solution will be determined within the limits of the searching domain defined by the original training dataset. E.g., if the age defined by the original dataset is between 21 and 90, then the generated solution will not look outside this searching domain. All solutions accepted to be reintroduced into the original training dataset are applied to a filter that allows solutions to have some changes smaller than a specified threshold. The increase in robustness of the model will allow fewer values to pass through the filter since the number of changes required to produce a different outcome is supposed to increase with each iteration.

A. Contributions of this paper

- Proposes a method for generating counterfactual solutions.
- Measures the impact of using the generated synthetic data back into the training dataset.
- Increases the importance of other features that might be more relevant in real-life (e.g., education).

IV. Proposed Solution

The proposed solution (illustrated in Fig. 2) supports all types of predictive models, whether it is a white box (Decision

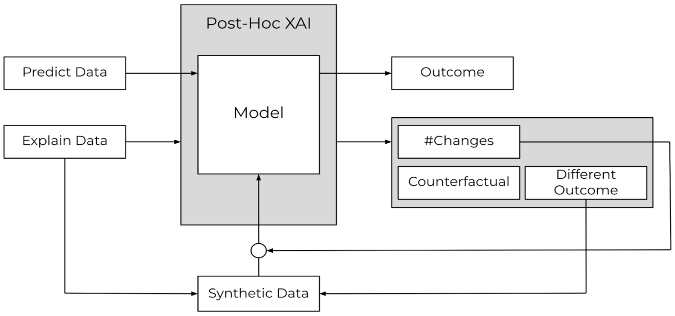

Fig. 2. Counterfactual generation by Post-HOC XAI algorithm using the trained model, input data, predicted outcome to offer synthetic data that can potentially be pushed back into the training dataset.

Tree), a gray box (Random Forest), or a black box (Neural Network). The trained model receives a data sample (*"Predict Data"*) and predicts an outcome (*"Outcome"*). The counterfactual explanation exposes what features are sensitive to a direct impact on the model decision reducing the importance of the rest of the features. If the generated counterfactual reveals such feature sensitivity, it can also be used by a human factor to invalidate the original model prediction. The heuristic is used as the XAI method to generate counterfactuals and receives input data to be explained (*"Explain Data"*). The method outputs the counterfactual sample (*"Counterfactual"*), the outcome of the counterfactual (*"Different Outcome"*) and the number of changes required to produce a different outcome. The input data to be explained (*"Explain Data"*) and the counterfactual outcome (*"Different Outcome"*) are the new synthetic data (*"Synthetic Data"*) that is passed through a filter that checks if the number of changes (*"#Changes"*) is less than a threshold already set. The synthetic data that passes the filter is added to the original dataset and the model will be retrained.

V. Experiment

The experiment uses a modified Adult dataset [18] where we hand-picked the training data such that the outcome should favor a single gender (Fig. 3). Calculating the correlation matrix will evidentiate the impact of gender on the salary status column (Fig. 4). The experiment aims to use synthetically generated counterfactuals in the training dataset to reduce the decision bias towards a set of specific features.

The experiment aims to reduce the data bias and increase the model fairness without using a programmatic guard to achieve the goal. As shown in Figure 4, the Gender feature has a significant impact on the outcome. Other features such as Age, Hours per Week, and Education have a significant impact but also have a direct correlation in real life.

The experiment uses a single black box prediction model to predict the outcome, but the same experiment can be applied to any model. Plotting the correlation matrix between each training feature and the outcome reveals the importance of the gender feature towards the salary status that exposes the heavily biased training data.

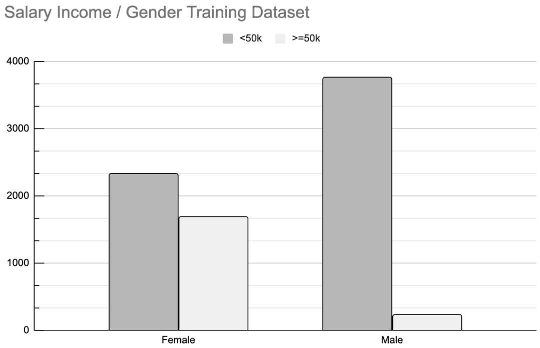

Fig. 3. Representation of gender-biased initial training data.

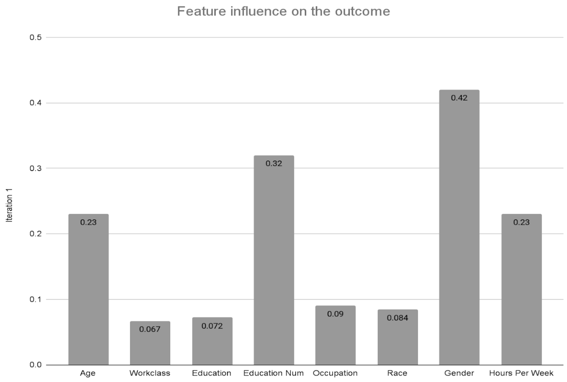

Fig. 4. Correlation matrix between initial training dataset.

The proposed method uses an evolutive algorithm (Genetic Algorithm) to generate new potential solutions in the defined searching domain close to the predicted input but generate a different result. The experiment is executed in multiple iterations by trying to generate counterfactuals for a batch of x input data, selecting the generated solutions whose number of changes is less than a threshold, introducing the input data with the counterfactual outcome back into the training dataset, and retraining the model and repeating the experiment with never seen data. For this experiment, we picked the acceptable threshold for a counterfactual to be less or equal to 2 changes (e.g., the generated solution would be accepted if the number of working hours will decrease/increase by one and the gender will have a different value).

The feature domains is captured in Table I. It consists of categorical and continuous variables, each having a different definition domain.

The experiment used 7,000 never seen samples from the original dataset to evaluate the model fairness and fragility. In Table II, the reference is the set of input features for which the model would predict an outcome of 0 (<50k/year), and the result is the set of changes to the reference for which the model would predict an outcome of >=50k/year. For this case, the reference data is selected to be pushed back into the dataset

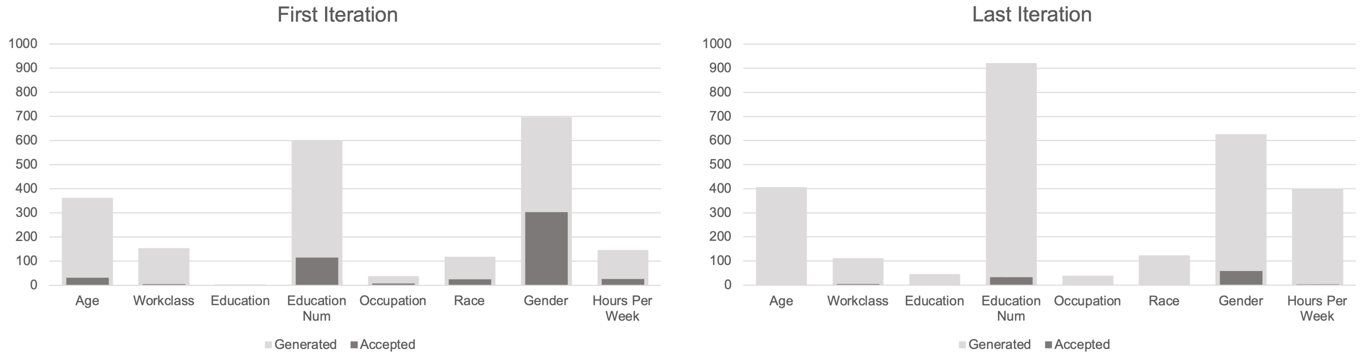

Fig. 5. Number of generated vs. accepted solutions for iterations 1 and 7.

TABLE I
FEATURE DOMAINS

Feature	Type	Min	Max	Count
Age	Continuous (int)	17	90	-
Workclass	Categorical	-	-	9
Education	Categorical	-	-	16
Education Num	Continuous (int)	1	15	-
Occupation	Categorical	-	-	15
Race	Categorical	-	-	5
Gender	Categorical	-	-	2
Hours Per Week	Continuous (int)	1	99	-
Salary Status	Categorical	-	-	2

with an outcome of 1 (>=50k/year).

TABLE II
SAMPLE COUNTERFACTUAL

Feature	Reference	Result
Outcome	0	[1]
Objectives [GA]		[1]
Constraints [GA]		[0]
Values		
Age	35	35
Workclass	Private	Private
Education	Bachelors	Bachelors
Education_Num	15	15
Occupation	Sales	Sales
Race	White	White
Gender	Male	Female
Hours_Per_Week	40	40

VI. RESULTS

The counterfactuals were generated using seven iterations, each consisting of a batch of 1,000 samples for measurements and counterfactuals. The decrease of accepted solutions with each iteration (Fig 6) results from an increase in model robustness and a proof that the generating algorithm finds fewer solutions or the found solutions have too many changes to be accepted.

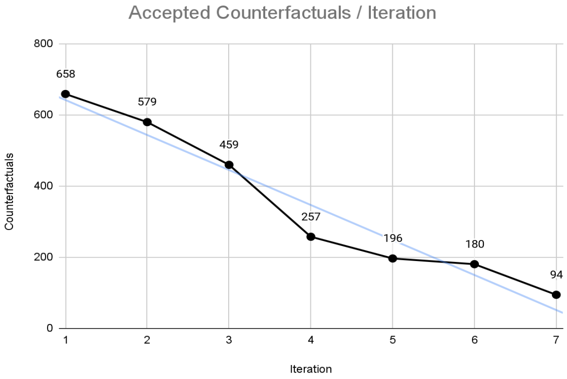

Fig. 6. Number of generated counterfactuals for each iteration.

To evaluate the model fairness update, a benchmark dataset was used after each new batch of synthetic data was pushed back into the training dataset, and the model was updated. After each benchmark, we extracted the number of times each feature was changed for both generated and accepted solutions. For the initial iteration, Gender was the primary feature that was changed to generate a potential solution (696 potential solutions, 302 passing the threshold). As the model is updated and the fairness is corrected, later iterations observed a decrease in solutions with changes in Gender and mainly focused on other, more relevant to the real-world features, such as Age, Education Num and Hours per Week. Fig 5 displays the benchmark dataset generated vs. accepted solutions.

We computed the correlation matrix between each input feature against the output column (Salary Status) to monitor the influence of each feature on the outcome at the training dataset level. As shown in Fig 7, the Gender feature influence on the outcome was highly diminished without significantly impacting the rest of the features. There are also other features that capture a personal characteristic, such as race, but since the original training data was not biased toward race, the feature importance of this column was not changed at the end of the iterative process. Also, features that do not contain any personal characteristic (e.g. Occupation, Hours per Week, Education) suffered a minor change at the end of the iterative process thus maintaining the original data characteristics.

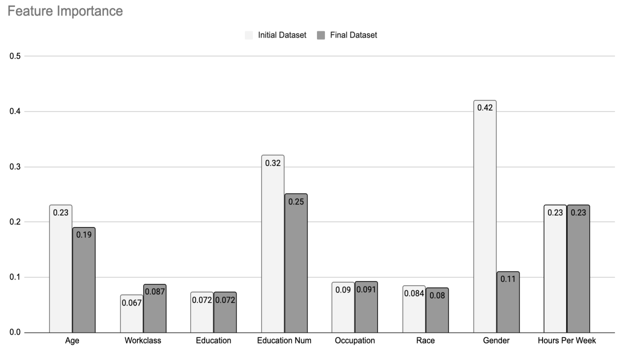

Fig. 7. Input feature correlation for comparison between first and last iteration.

References

[1] G. Warren, M. T. Keane, and R. M. J. Byrne, "Features of explainability: How users understand counterfactual and causal explanations for categorical and continuous features in xai," 2022. [Online]. Available: https://arxiv.org/abs/2204.10152

[2] K. Yeung, "A study of the implications of advanced digital technologies (including ai systems) for the concept of responsibility within a human rights framework," *Social Science Research Network*, 2018.

[3] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017. [Online]. Available: https://arxiv.org/abs/1702.08608

[4] Y.-L. Chou, C. Hsieh, C. Moreira, C. Ouyang, J. Jorge, and J. M. Pereira, "Benchmark evaluation of counterfactual algorithms for xai: From a white box to a black box," 2022. [Online]. Available: https://arxiv.org/abs/2203.02399

[5] S. Verma, J. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: Challenges revisited," 2021. [Online]. Available: https://arxiv.org/abs/2106.07756

[6] X. Ying, "An overview of overfitting and its solutions," *Journal of Physics: Conference Series*, vol. 1168, p. 022022, feb 2019. [Online]. Available: https://doi.org/10.1088/1742-6596/1168/2/022022

[7] H. Allamy and R. Z. Khan, "Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study)," 01 2014, pp. 163–172.

[8] J. Brownlee, "Impact of dataset size on deep learning model skill and performance estimates," *Deep Learning Performance*, jan 2019. [Online]. Available: https://machinelearningmastery.com/impact-of-dataset-size-on-deep-learning-model-skill-and-performance-estimates/

[9] S. Verma, K. Hines, and J. Dickerson, "Counterfactual explanations for machine learning: A review," 2020. [Online]. Available: https://arxiv.org/abs/2010.10596

[10] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, and J. Jorge, "Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications," 2021. [Online]. Available: https://arxiv.org/abs/2103.04244

[11] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," 2017. [Online]. Available: https://arxiv.org/abs/1711.00399

[12] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. D. Bie, and P. Flach, "FACE," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM, feb 2020. [Online]. Available: https://doi.org/10.1145%2F3375627.3375850

[13] A. White and A. d. Garcez, "Counterfactual instances explain little," 2021. [Online]. Available: https://arxiv.org/abs/2109.09809

[14] R. Mazzine and D. Martens, "A framework and benchmarking study for counterfactual generating methods on tabular data," 2021. [Online]. Available: https://arxiv.org/abs/2107.04680

[15] L. Weber, S. Lapuschkin, A. Binder, and W. Samek, "Beyond explaining: Opportunities and challenges of xai-based model improvement," 2022. [Online]. Available: https://arxiv.org/abs/2203.08008

[16] P. Schramowski, W. Stammer, S. Teso, A. Brugger, X. Shao, H.-G. Luigs, A.-K. Mahlein, and K. Kersting, "Making deep neural networks right for the right scientific reasons by interacting with their explanations," 2020. [Online]. Available: https://arxiv.org/abs/2001.05371

[17] P. Huber and T. Guida, "Genetic algorithms: A heuristic approach to multi-dimensional problems," 2019. [Online]. Available: https://ssrn.com/abstract=3451302

[18] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml