ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Adversarial Auditing of Machine Learning Models under Compound Shift

Karan Bhanot[1], Dennis Wei[2], Ioana Baldini[2], and Kristin P. Bennett[3]

1- Rensselaer Polytechnic Institute - Department of Computer Science - New York - USA
2- IBM Research - Yorktown Heights, New York - USA
3- Rensselaer Polytechnic Institute - Department of Mathematics - New York - USA

Abstract. Machine learning (ML) models often perform differently under distribution shifts, in terms of utility, fairness, and other dimensions. We propose the Adversarial Auditor for measuring the utility and fairness performance of ML models under compound shifts of outcome and protected attributes. We use Multi-Objective Bayesian Optimization (MOBO) to account for multiple metrics and identify shifts where model performance is extreme, both good and bad. Using two case studies, we show that MOBO performed better than random and grid-based approaches in identifying scenarios by adversarially optimizing objectives, highlighting the value of such an auditor for developing fair, accurate and shift-robust models.

1 Introduction

As Machine Learning (ML) models are deployed in high-stakes domains, it is crucial to test their behavior under distribution shifts, where the data distribution during deployment may differ from the distribution of the training data [1, 2]. Deploying models without properly understanding their behavior can lead to unpredictable results. Ideally, model testing would include rigorous testing under shifts to understand the ranges of model performance in terms of utility, fairness, and other dimensions.

Thus, there is a need to identify distribution shifts in data where the ML model performance is extreme, both good and bad, so the model developers can understand their models better and re-train their models whenever needed. In our previous work, we proposed a pipeline to evaluate bias mitigation algorithms under shifts using a grid-based approach [3]. While being rigorous and flexible, the pipeline is constrained to a grid of possible data shifts defined by the user without considering extremes and in-between values. In this work, we propose to address this problem using an Adversarial Auditor, which uses Multi Objective Bayesian Optimization (MOBO) [4, 5] to identify distribution shifts on a continuous scale that lead to extreme performance in terms of multiple metrics, by adversarially optimizing for multiple objectives. Such an analysis enables the developers to develop fair, accurate and shift-robust models by highlighting scenarios for model improvement. Amongst the various types of distribution shifts [6], we evaluate ML models under compound shifts by changing the proportions of outcome and protected attribute. To highlight the efficacy of this auditor, we present results across three MOBO algorithms on two objectives: (a) utility and

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

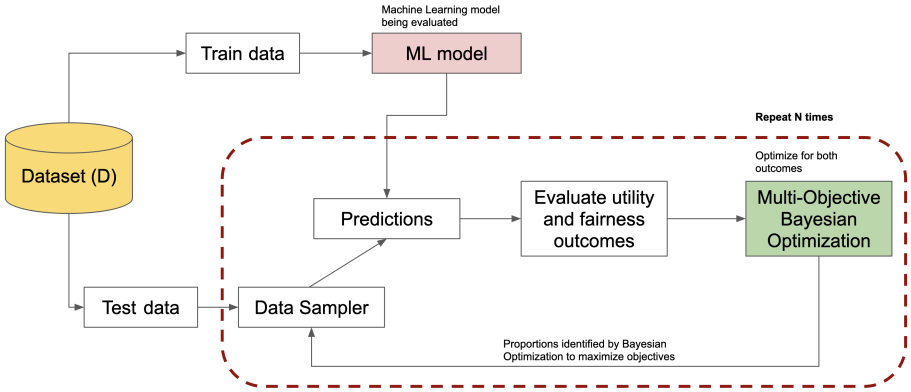

Fig. 1: The Adversarial Auditor design where the utility and fairness scores are the two objectives used for Multi-Objective Bayesian Optimization

(b) fairness. These results are compared with the grid-based approach [3] and a randomization baseline.

2 Method

2.1 Auditor Design

Let us consider a dataset $D = \{X, Z, Y\}$ where X is one or more features (such as blood pressure level, county etc.), Z is a protected attribute (such as gender, race etc.) and Y is the prediction variable being predicted. The dataset is split into train and test data as shown by the auditor design in Figure 1. The train data is then used for training a Machine Learning model M (e.g. Random Forest Classifier, XGBoost, etc.) that will be evaluated by the auditor.

Let us consider the simple case where the output is binary, such that $Y \in \{0, 1\}$, and the protected attribute is binary as well, $Z \in \{0, 1\}$. Using these variables, we define the four groups $(Z, Y) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. Let the proportions of these groups be defined as P_{00}, P_{01}, P_{10}, P_{11} (P), each ranging between 0.05 to 0.95 to avoid very small groups. Thus, each distribution shift (referred to as a "scenario") is defined as a unique combination of these proportions. The distribution shift is realized as a dataset by sampling from the test data using the Data Sampler. We use a stratified sampling of the test data as the Data Sampler.

The model M predicts on this shifted data and is then evaluated for utility u and fairness f. The MOBO algorithm is then applied with the proportions P as inputs and the u and f as objectives to identify new P values which maximize these objectives. We use three algorithms: (a) qNParEGO, (b) qNEHVI and (c) qEHVI [4, 5]. Note that as the default MOBO algorithm maximizes the objectives (BoTorch library [7]), if we want to minimize an objective, we simply

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

evaluate its negative. This process is repeated N times (we use N = 81 for a fair comparison with the 81 grid-approach, see [3] and Section 2.3) and the results are observed. This complete pipeline forms the **Adversarial Auditor**.

2.2 Evaluation and Comparison

We consider Balanced Accuracy (BA) as the utility metric u, which averages the accuracy across the two classes of Y and thus, is more informative when datasets are imbalanced. A higher score for BA is better (max 1.0). To evaluate fairness f, we use the group fairness metric Equalized Odds [3, 8] which measures equitable performance across groups as defined by the protected attribute (such as Male-Female for gender, Black population-White population for race etc). For Equalized Odds, lower values are better (min 0.0).

For identifying the scenario where the model is likely to perform the best (best utility and best fairness), we maximize the Balanced Accuracy and maximize the negative of Equalized Odds. For the reverse case of identifying the worst model performance (worst utility and worst fairness), we maximize the negative of Balanced Accuracy and maximize Equalized Odds.

2.3 Baselines

For the three MOBO algorithms, we plot Pareto fronts which identify the most efficient (i.e., non-dominated) solutions in terms of the two objectives supplied to them. Comparing these fronts enables us to see where and when one algorithm is performing better than others at identifying more extreme solutions. The MOBO Pareto fronts are compared with those for two baseline approaches:

Grid-based Approach: For creating a grid, we use the approach discussed in [3], where the proportions of Y=0 to Y=1 and Z=0 to Z=1 are updated from 0.1 to 0.9, with a step of 0.1 each. This creates a set of 9x9 (81) scenarios each with a unique shift (unique P). These results are described as "Grid scores".

Randomization Approach: In this approach, we randomly sample 4 proportions, corresponding to a random scenario, and evaluate the resulting utility and fairness scores. We repeat this process 81 times, to create 81 different scenarios. The results are referred to as "Random scores".

3 Experimental Results

We explore the results across two case studies. For each study, once the datasets are pre-processed, the data is split into train-test (70-30) and the resulting train data is used for training a Random Forest (RF) Classifier with *random_state=0*.

Case Study 1: We explore the Adult Income dataset [9], which is a popular dataset used for classification problems where the features of individuals such as capital gain, gender, marital status etc. are used to predict their income, being higher than $50K or not. We pre-process the data to remove rows with missing values, convert marital status to binary and one-hot encode categorical columns and then train the RF model. Sample size for each run is set to 10K. We run

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

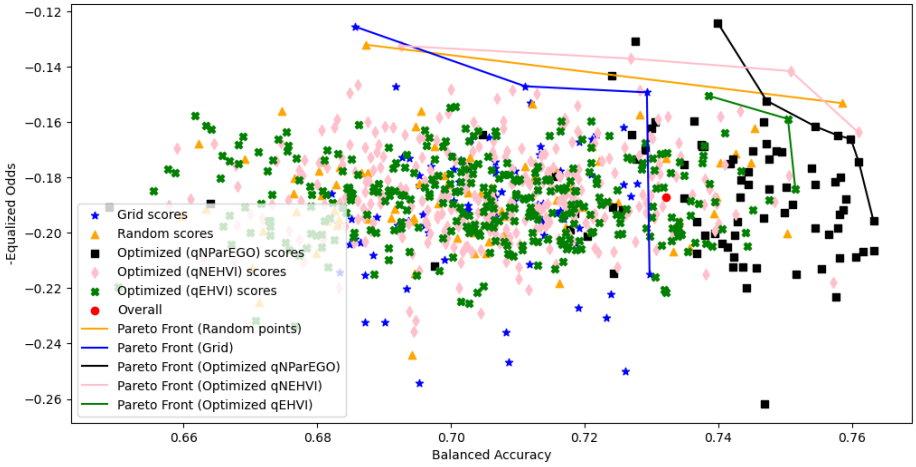

Fig. 2: Identifying scenarios where the RF model for Adult Income has high Balanced Accuracy and low Equalized Odds (or high negative Equalized Odds). Closer to the top-right corner is better.

the auditor to identify the scenarios where the RF model is likely to achieve the highest Balanced Accuracy and lowest Equalized Odds (or highest negative Equalized Odds) with the results shown in Figure 2.

We observe the various scenarios identified by the 5 approaches and their corresponding Pareto fronts along with the score on the complete test data shown by a red circle (Overall). As expected, the Grid scores are constrained due to their design and are farther from the ideal top-right corner compared to other fronts (blue line). Random scores do a better job at achieving high negative Equalized Odds and high Balanced Accuracy (orange line). While qEHVI (green) does slightly worse than Random scores, it still improves upon the Grid scores. Additionally, qNParEGO and qNEHVI perform similar and in some scenarios, better than Grid and Random scores. Thus, the auditor is able to identify the scenarios where the RF model is performing very well and can inform the user of its capabilities and usefulness even under distribution shifts.

Case Study 2: We explore the medical dataset MIMIC-III [10, 11, 12] processed to create a prediction task of identifying mortality within 30 days of ICU admission using features like race, admission location etc. We pre-process the data to combine several files together, restrict the records to only include race as White or Black, and one-hot encode the categorical columns before training the RF model. Sample size is set to 1.5K records. We identify the scenarios where the model will achieve lowest Balanced Accuracy (or higher negative Balanced Accuracy) with higher Equalized Odds as shown in Figure 3.

We find that Grid scores, Random scores and qNEHVI have overlapping Pareto fronts indicating a comparable performance for identifying the scenarios (lines connecting stars, triangles and diamonds). However, qNParEGO and qE-

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

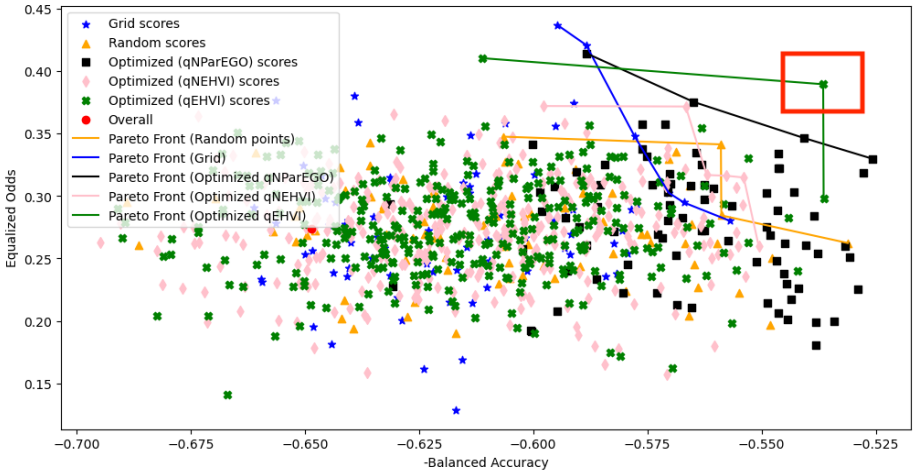

Fig. 3: Scenarios where the RF model for MIMIC-III has low Balanced Accuracy and high Equalized Odds (again top-right corner is most extreme).

HVI clearly perform better than others (black and green lines connecting squares and crosses respectively) shown by their Pareto fronts being the furthest, with high Equalized Odds and high negative Balanced Accuracy. The results highlight how the auditor is able to identify scenarios which encourage the developers to test and update their models before deployment.

For example, for the green square marker in Figure 3 identified by a red box, the Balanced Accuracy is 0.5365 and Equalized Odds value is 0.3893 for the proportions $P_{00} = 0.2258$, $P_{01} = 0.4041$, $P_{10} = 0.7233$, $P_{11} = 0.1146$. The auditor is able to identify this unique compound shift where the RF model is struggling, having high unfairness while making almost random predictions (Balanced Accuracy close to 0.5). Thus, such an insight enables the developers to update their model to handle such a shift. This can be observed and mitigated for other extremes identified on this Pareto front as well which would have otherwise been missed if such MOBO algorithms using the auditor were not used.

4 Discussion and Conclusion

The Adversarial Auditor proposed in this work enables us to audit Machine Learning models to identify scenarios where they are likely to perform at extremes, using Multi-Objective Bayesian Optimization. Through the two case studies, we highlighted the effectiveness of the auditor in identifying more extreme scenarios in comparison with existing baselines. Thus, this auditor enables an evaluation of Machine Learning models to highlight their strengths and pitfalls under shifts before they are deployed, thus preventing incorrect and unpredictable use.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

References

[1] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R Varshney, S Speakman, Z Mustahsan, and S Chakraborty. Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 91–98, 2019.

[2] S Bickel, M Brückner, and T Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9), 2009.

[3] K Bhanot, I Baldini, D Wei, J Zeng, and K Bennett. Stress-testing fairness mitigation techniques under distribution shift using synthetic data. *SIGKDD 2022 Workshop on Ethical Artificial Intelligence (EAI)*, 2022.

[4] S Daulton, M Balandat, and E Bakshy. Differentiable expected hypervolume improvement for parallel multi-objective bayesian optimization. *Adv. in Neural Information Processing Systems*, 33:9851–9864, 2020.

[5] S Daulton, M Balandat, and E Bakshy. Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement. *Adv. in Neural Information Processing Systems*, 34:2187–2200, 2021.

[6] J Schrouff, N Harris, O Koyejo, I Alabdulmohsin, E Schnider, K Opsahl-Ong, A Brown, S Roy, D Mincu, C Chen, A Dieng, Y Liu, V Natarajan, A Karthikesalingam, K Heller, S Chiappa, and A D'Amour. Maintaining fairness across distribution shift: do we have viable solutions for real-world applications?, 2022.

[7] M Balandat, B Karrer, DR Jiang, S Daulton, B Letham, A Gordon Wilson, and E Bakshy. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Adv. in Neural Information Processing Systems 33*, 2020.

[8] M Hardt, E Price, and N Srebro. Equality of opportunity in supervised learning. *Adv. in Neural Information Processing Systems*, 29, 2016.

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

[10] A Johnson, TJ Pollard, L Shen, L-W H Lehman, M Feng, M Ghassemi, B Moody, P Szolovits, LA Celi, and RG Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, 2016.

[11] A Johnson, T Pollard, and R Mark. Mimic-iii clinical database (version 1.4). *PhysioNet*, 2016.

[12] A Goldberger, L Amaral, L Glass, J Hausdorff, PC Ivanov, R Mark, JE Mietus, GB Moody, CK Peng, and HE Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation [Online]*, 101(23):e215âe220, 2000.