"Why did the Model Fail?": Attributing Model Performance Changes to Distribution Shifts

Anonymous Authors[1]

Abstract

Performance of machine learning models may differ significantly in novel environments compared to during training due to shifts in the underlying data distribution. Attributing performance changes to specific data shifts is critical for identifying sources of model failures and designing stable models. In this work, we design a novel method for attributing performance difference between environments to shifts in the underlying causal mechanisms. To this end, we construct a cooperative game where the contribution of each mechanism is quantified as their Shapley value. We demonstrate the ability of the method to identify sources of spurious correlation and attribute performance drop to shifts in label and/or feature distributions on synthetic and real-world datasets.

1. Introduction

Machine learning models are widely deployed in dynamic environments ranging from recommendation systems to personalized clinical care. Such environments are prone to dataset shifts, which may lead to serious degradations in model performance (Guo et al., 2022; Chirra et al., 2018; Koh et al., 2021; Geirhos et al., 2020; Nestor et al., 2019). Importantly, the shifts are hard to anticipate and reduce the ability of model developers to design reliable systems.

When the performance of a model *does* degrade during deployment, it is crucial for the model developer to know *how* the dataset has shifted to cause this change. Cognizant of this information, the model developer can then take mitigating actions such as additional data collection, data augmentation, and model retraining (Ashmore et al., 2021; Zenke et al., 2017; Subbaswamy et al., 2019).

In this work, we present a method to attribute model performance changes to shifts in a given set of distributions. Dataset shifts can occur in various marginal or conditional distributions comprising variables involved in the model. Given that many distributions can change simultaneously, we define the effect of changing any set of distributions and use the concept of Shapley values (Roth, 1988) to attribute the change to individual distributions.

We build on a recent line of work that defines distribution shifts as interventions on causal mechanisms (Pearl & Bareinboim, 2011; Subbaswamy et al., 2019; 2021; Budhathoki et al., 2021). Most relevant is Budhathoki et al. (2021) which attributes a shift between two joint distributions to the causal mechanisms denoted by a sub-distributions (i.e. factorization of the joint distribution). We differ by attributing a change in *model performance* to sub-distributions. Note that the shifted sub-distributions may have different influence on model performance. In this work, we demonstrate that explaining performance discrepancy requires us to develop specialized methods.

We make the following contributions:

- We formalize the problem of attributing model performance changes due to distribution shifts.
- We propose a principled approach based on Shapley values for the attribution, and show that it satisfies several desirable properties.
- We validate the correctness of our method on synthetic and real-world datasets.

2. Related work

There has been extensive work that tests whether the data distribution has shifted (e.g. ones evaluated in Rabanser et al. (2019)). Past work has proposed to identify sub-distributions (factors constituting the joint distribution as determined by a generative model for the data) that comprise the shift between two joint distributions and order them by their contribution to the shift (Budhathoki et al., 2021). The method proposed in Budhathoki et al. (2021) does not attribute the contribution of each sub-distribution to model performance. Even a small change in some (factors) may have a large effect on model performance. Thereby, the attribution in Budhathoki et al. (2021) may yield multiple shifting sub-distributions that the model developer has to filter out to identify the relevant ones.

Shapley value-based attribution has recently become popular for interpreting model predictions (Štrumbelj & Kononenko, 2014; Lundberg & Lee, 2017; Wang et al., 2021). However, in most prior work, Shapley values have been leveraged for attributing changes in specific model pre-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the SCIS workshop.

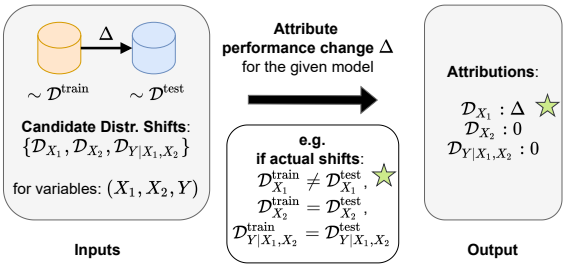


Figure 1. **Inputs and outputs for attribution.** The goal is to attribute the model's performance change Δ between train and test distributions. In this example, out of the three candidate distributions only the marginal distribution of X_1 changes. Thus, the method attributes all of the performance change to X_1.

dictions to variables (Sundararajan & Najmi, 2020). Challenges to appropriately interpreting such attributions and desirable properties thereof have been extensively discussed in Janzing et al. (2020); Kumar et al. (2021). Recently, Wu et al. (2021) decompose performance change to changes in marginal distributions using Shapley value framework. In contrast, we advance the use of Shapley values for interpreting model performance changes, particularly by attributing discrepancy to any sub-distribution.

Finally, recent work aims to find subsets of the dataset that have significantly worse (or better) performance (d'Eon et al., 2021; Eyuboglu et al., 2022). The main difference in our work is the data representations used for attribution. These works chose to identify subsets of data that are relevant to performance change whereas we find sub-distributions represented by causal mechanisms.

3. Preliminaries

Notation Consider a learning setup where we have some system variables denoted by V. We are given a fixed model f and a loss function $\ell : v, f \mapsto \ell(v, f) \in \mathbb{R}$ which assigns a real value to the model evaluated at a specific setting of the variables. For example, in the case of supervised learning, $V = (X, Y)$ comprises the features and labels, the model f maps features into the label space, and a loss function such as the squared error $\ell((x, y), f) := (y - f(x))^2$ is used to evaluate the model. Assume that the loss function is computed separately for each data point. Then, performance of the model is summarized by the average of the losses,

$$\text{Perf}(\mathcal{D}) := \mathbb{E}_{v \sim \mathcal{D}}[\ell(v, f)]$$

on data from a given distribution $\mathcal{D}$. We use $\mathcal{D}^{\text{train}}$ to denote the distribution for data at training and $\mathcal{D}^{\text{test}}$ for testing. Subscripts on $\mathcal{D}$ refer to the distribution of specific variables. For example, $\mathcal{D}_X$ is the distribution of features, and $\mathcal{D}_{Y|X}$ is the conditional distribution of labels given features.

Sources of performance change Model performance can change between development and deployment of models for many reasons (Jacobs & Wallach, 2021). We restrict our attention to a narrow yet important source of performance change which is the shift in the distribution of input features or labels. Since we have combinatorially many shifts that can be defined on subsets of $V = (X, Y)$, we leverage the knowledge of causal mechanisms in the form of a causal graph to identify potential shifts to consider (Pearl, 2009).

3.1. Problem setup

Suppose we are given a *candidate set* of distributions $\mathtt{C}_{\mathcal{D}}$ that may account for the shift from train $\mathcal{D}^{\text{train}}$ to test $\mathcal{D}^{\text{test}}$ distributions: $\text{Perf}(\mathcal{D}^{\text{test}}) - \text{Perf}(\mathcal{D}^{\text{train}})$. **Our goal is to attribute this change to each candidate distribution in $\mathtt{C}_{\mathcal{D}}$.** For our method, we assume access to the model f, and samples from $\mathcal{D}^{\text{train}}$ as well as $\mathcal{D}^{\text{test}}$ (see Figure 1).

3.2. Choice of candidate set and causal mechanisms

Candidate distributions can be defined in multiple ways. For instance, it can be the set of marginal distributions on each system variable, $\mathtt{C}_{\mathcal{D}} = \{\mathcal{D}_{X_1}, \mathcal{D}_{X_2}, \cdots\}$, or distribution of each variable after conditioning on the rest, $\mathtt{C}_{\mathcal{D}} = \{\mathcal{D}_{X_1|V \setminus X_1}, \mathcal{D}_{X_2|V \setminus X_2}, \cdots\}$.

Motivated by Budhathoki et al. (2021), we propose to use the causal mechanisms constituting the structural causal graph for the system as our candidate set. A causal graph specifies a particular factorization of the joint distribution into a set of distributions (alternatively called causal mechanisms) (Pearl, 2009). That is $\mathcal{D}_V = \prod_{X_i \in V} \mathcal{D}_{X_i|\text{parent}(X_i)}$ where $\text{parent}(X_i)$ are the variables that have a directed edge to X_i in the causal graph. These distributions are assumed to be *independent*, i.e. an intervention to change one of the distributions does not change any other distribution in the factorization. We also assume that the causal graph is *sufficient* (Spirtes et al., 2000). Thus, the candidate set is

$$\mathtt{C}_{\mathcal{D}} = \{\mathcal{D}_{X_1|\text{parent}(X_1)}, \cdots, \mathcal{D}_{X_i|\text{parent}(X_i)}, \cdots\}_{i=1,\cdots,|V|}.$$

Advantages of using causal mechanisms. This choice of candidate set has three main advantages. First, it is *interpretable* since the candidate shifts are specified by domain experts who constructed the causal graph. Second, it is *actionable* since identifying the causal mechanisms most responsible for performance change can inform training methods for handling distribution shifts (Subbaswamy et al., 2019). Third, it will lead to *succinct* attributions due to the independence property. Consider the case where only one conditional distribution $\mathcal{D}(X_i|\text{parent}(X_i))$ changes across domains. This will result in a change in distributions of all descendants of X_i (due to the factorization given above). In this case, a candidate set defined by all marginals is not succinct, as one would attribute performance changes to all marginals of descendants of X_i. Instead, focusing on our candidate set determined by the causal mechanism will isolate the appropriate conditional distribution.

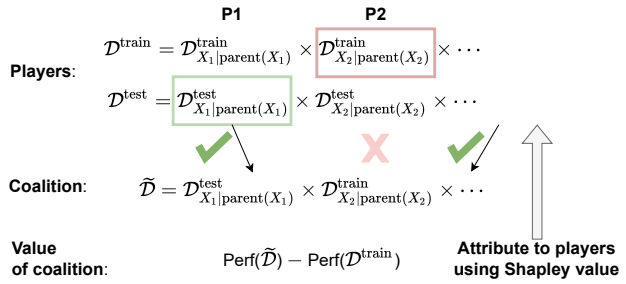


Figure 2. **Sketch of the game theoretic attribution method.** Each causal mechanism is a player that, if present in the coalition, changes to the test distribution and, if absent, remains fixed to the train distribution. This defines the distribution of the resulting coalition $\widetilde{\mathcal{D}}$. Performance on $\widetilde{\mathcal{D}}$ is estimated using importance sampling from training data samples. After computing values for each possible coalition, Shapley value (Eq. 2) gives the attribution. Thus, we estimate the performance change under all possible ways to shift the mechanisms from train to test and use these to distribute the total performance change among the individual mechanisms.

4. Method

We motivate a game theoretic formulation for attributing performance changes to distributions over variable subsets.

4.1. Game theoretic attribution

Consider the following attribution game. The set of **players** in this game are the candidate distributions. A **coalition** of any subset of players determines the distributions that are allowed to shift, keeping the rest fixed. The **value** for the coalition is the model performance change between the resulting distribution for the coalition and the train distribution. See Figure 2 for an overview of the method.

Consider a coalition that consists of two factors in the candidate set, that is $\{\mathcal{D}_{X_1|\text{parent}(X_i)}, \mathcal{D}_{X_2|\text{parent}(X_2)}\} = \widetilde{\mathtt{C}} \subset \mathtt{C}_{\mathcal{D}}$. The resulting distribution for the coalition $\widetilde{\mathtt{C}}$ is

$$\widetilde{\mathcal{D}} = \mathcal{D}^{\text{test}}_{X_1|\text{parent}(X_1)} \mathcal{D}^{\text{test}}_{X_2|\text{parent}(X_2)} \prod_{i \in V \setminus X_1, X_2} \mathcal{D}^{\text{train}}_{X_i|\text{parent}(X_i)}.$$

The value of the coalition $\widetilde{\mathtt{C}}$ with distribution $\widetilde{\mathcal{D}}$ is given by

$$\text{Val}(\widetilde{\mathtt{C}}) := \text{Perf}(\widetilde{\mathcal{D}}) - \text{Perf}(\mathcal{D}^{\text{train}}) \tag{1}$$

Shapley value for this game gives the attribution of each player $d \in \mathtt{C}_{\mathcal{D}}$ as

$$\text{Attr}(d) = \sum_{\widetilde{\mathtt{C}} \subseteq \mathtt{C}_{\mathcal{D}} \setminus \{d\}} \binom{|\mathtt{C}_{\mathcal{D}}| - 1}{|\widetilde{\mathtt{C}}|}^{-1} \left(\text{Val}(\widetilde{\mathtt{C}} \cup \{d\}) - \text{Val}(\widetilde{\mathtt{C}})\right) \tag{2}$$

We argue that Shapley attribution provides a good framework in our setting by showing that it satisfies two desired properties in Appendix A.

Thus, to compute our attributions, we need estimates of model performance under $\widetilde{\mathcal{D}}$. Note that we only have model performance estimates under $\mathcal{D}^{\text{train}}$ and $\mathcal{D}^{\text{test}}$ but not for any arbitrary coalition where a subset of the distributions have shifted. For this we propose to use importance sampling.

4.2. Estimating attribution using importance sampling

Importance sampling allows us to re-weight the samples drawn from a given distribution, which can be $\mathcal{D}^{\text{train}}$ or $\mathcal{D}^{\text{test}}$, to simulate expectations for a desired distribution, which is the candidate $\widetilde{\mathcal{D}}$ in our case. Thus, we re-write the value as

$$\begin{aligned}\text{Val}(\widetilde{\mathtt{C}}) &= \text{Perf}(\widetilde{\mathcal{D}}) - \text{Perf}(\mathcal{D}^{\text{train}}) \\ &= \mathbb{E}_{v \sim \widetilde{\mathcal{D}}}[\ell(v, f)] - \mathbb{E}_{v \sim \mathcal{D}^{\text{train}}}[\ell(v, f)] \\ &= \mathbb{E}_{v \sim \mathcal{D}^{\text{train}}}\left[\frac{\widetilde{\mathcal{D}}(v)}{\mathcal{D}^{\text{train}}(v)} \ell(v, f)\right] - \mathbb{E}_{v \sim \mathcal{D}^{\text{train}}}[\ell(v, f)]\end{aligned}$$

The importance weights are themselves a product of ratios of train and test distributions corresponding to the causal mechanisms in $\mathtt{C}_{\mathcal{D}}$ as follows:

$$w_{\widetilde{\mathtt{C}}}(v) = \frac{\widetilde{\mathcal{D}}(v)}{\mathcal{D}^{\text{train}}(v)} = \prod_{d \in \widetilde{\mathtt{C}}} \frac{\mathcal{D}^{\text{test}}_d(v)}{\mathcal{D}^{\text{train}}_d(v)} =: \prod_{d \in \widetilde{\mathtt{C}}} w_d(v) \tag{3}$$

There are multiple ways to estimate the importance weights $w_d(v)$, which are ratio of densities, in the literature (Sugiyama et al., 2012). Here, we use a simple method based on training probabilistic classifiers. If $d = \mathcal{D}_{X_i|\text{parent}(X_i)}$, then $w_d = \frac{\mathbb{P}_{\mathcal{D}^{\text{test}}}(X_1|\text{parent}(X_1))}{\mathbb{P}_{\mathcal{D}^{\text{train}}}(X_1|\text{parent}(X_1))} = \frac{\mathbb{P}_{\mathcal{D}^{\text{test}}}(X_1)}{\mathbb{P}_{\mathcal{D}^{\text{train}}}(X_1)} \times \frac{\mathbb{P}_{\mathcal{D}^{\text{train}}}(\text{parent}(X_1))}{\mathbb{P}_{\mathcal{D}^{\text{test}}}(\text{parent}(X_1))}$, where both terms are computed using binary classifiers trained to discriminate data points from $\mathcal{D}^{\text{train}}$ and $\mathcal{D}^{\text{test}}$ from the concatenated dataset (see Appendix B for details). We need to train $\mathcal{O}(|\mathtt{C}_{\mathcal{D}}|)$ models for computing all importance weights. The computational effort is manageable for a sparse causal graph.

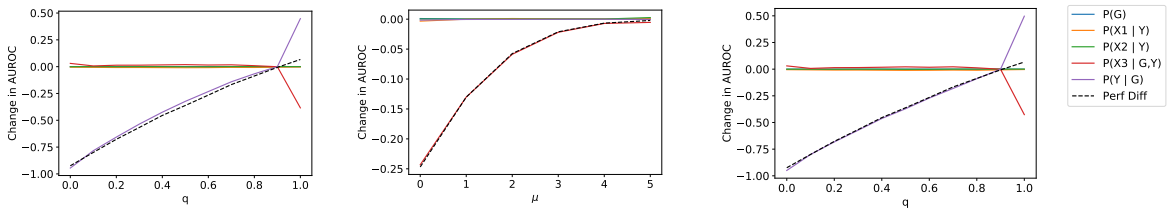
5. Experiments

5.1. Synthetic datasets

We generate a synthetic binary classification dataset with three features according to the following data generating process, corresponding to the causal graph shown in Figure 4a. Here, $\xi_p : \{0, 1\} \to \{0, 1\}$ is a function that randomly flips the input with probability p.

$$\begin{aligned} G &\sim Ber(0.5), & Y &= \xi_q(G), \tilde{Y} = \xi_{0.25}(Y) \\ X_1 &= \mathcal{N}(\omega \tilde{Y}, 1), & X_2 &= \mathcal{N}(\tilde{Y}, 1), X_3 = \mathcal{N}(\tilde{Y} + \mu G, 1) \end{aligned}$$

Where q, ω and μ are parameters of the data generating process. In the source environment, we set $q = 0.9, \omega = 1$ and $\mu = 5$. We generate 20,000 samples using these parameters, and train logistic regression (`LR`) and XGBoost (`XGB`, (Chen & Guestrin, 2016)) models on (X_1, X_2, X_3) to



(a) Conditional Label Shift (b) Conditional Covariate Shift (c) Combined Shift

Figure 3. AUROC differences attributed by our model to five potential distributional shifts on the synthetic dataset for the LR model. We observe that the overall change (Perf Diff) is attributed to the true shift in all of the three cases. Rest of the shifts have zero attributed value.

(a) (b)

Figure 4. Causal graphs for (a) Synthetic and (b) Real-world data.

predict Y, using 3-fold cross-validation for model selection. We explore three data settings for the target environment:

(a) Conditional Label Shift: Vary $q \in [0, 1]$. Keep ω and μ at their source values. Only $P(Y|G)$ changes.
(b) Conditional Covariate Shift: Vary $\mu \in [0, 5]$. Keep q and μ at their source values. Only $P(X_3|G, Y)$ changes across domains.
(c) Combined Shift: Set $\omega = 0$. Vary $q \in [0, 1]$. Keep ω at its source value. Both $P(X_1|Y)$ and $P(Y|G)$ change across domains, but their specific contribution to model performance degradation is not known exactly.

We use our method to explain performance changes in AUROC and Brier score for each model on target environments generated within each setting (with $n = 20,000$), computing density ratios using XGB models. Note that the causal graph shown in Figure 4a implies five potential distributional shifts: $\mathtt{C}_{\mathcal{D}} = \{\mathcal{D}_G, \mathcal{D}_{Y|G}, \mathcal{D}_{X_1|Y}, \mathcal{D}_{X_2|Y}, \mathcal{D}_{X_3|G,Y}\}$.

5.2. Real-world datasets

For some real-world datasets, the causal graph in Figure 4b can model the spurious correlation between X and Y (here, due to the common cause G). We test our method on two such binary classification datasets where $G \in \{0, 1\}$.

- **Waterbirds** (Wah et al., 2011). G is the background (water or land), Y is the type of bird (waterbird or landbird).
- **MetaShift** (Liang & Zou, 2022). G is the background (indoor or outdoor), and Y is the type of animal.

In all cases, X is a vector representing the image. Here, we use static embeddings computed from a Imagenet pretrained ResNet-50 (He et al., 2016), and so $X \in \mathbb{R}^{2048}$.

For the source environment, we first use random upsampling on each dataset to ensure that $\mathbb{P}(G = 0) = \mathbb{P}(G = 1) = 0.5$. We split each dataset into a 75% train, 25% test set, and train LR and XGB models to predict Y from X using 3-fold cross validation. In the target environments, after balancing G through upsampling, we use the same train/test splits as the source environment, and vary a parameter $r \in (0, 1]$. For each target environment, we randomly downsample the two majority groups (across $G \times Y$) to fraction r, and randomly upsample the two minority groups to size $1/r$. Note that for $r = 1$, the source and target environments are identical. We use our method to explain the performance difference between the source test set and target test set, using AUROC and Brier score as metrics.

6. Results

In Figure 3, we show the output of our method across the three settings described in Section 5.1, with LR as the model of interest and AUROC as the metric. We show similar results for XGB and Brier score in Appendix C. We find that our method attributes all of the performance changes to the correct ground truth shifts, both when there is a single shift (Settings 1 and 2) and when there are multiple shifts (Setting 3). In the case of Setting 3, we find that our method attributes all of the performance drop to a shift in $P(Y|G)$, which makes sense as the model relies only on spurious information in the source environment.

In Appendix D, we show the results for the real world datasets, with LR as the model of interest and AUROC as the metric. We find that our model largely attributes performance drops to the correct ($P(Y|G)$) distribution. However, some shift is still attributed to the $P(X|G, Y)$ distribution, which does not change. This is likely a result of inaccuracies in the importance weighting procedure (e.g. overfitting or finite samples), which requires further investigation.

7. Conclusions

We propose a method to attribute changes in performance of a model deployed on a different distribution. Attribution is naturally understood as an interventional concept. We combine the notion of interventions on causal mechanisms with the Shapley value framework to provide a useful decomposition of performance changes. Improvements to the method include relaxing the assumption of sufficiency. Further, recent advances that convert the Shapley value computation to a regression problem (Lundberg & Lee, 2017) can be used to reduce the computational effort required for the values.

References

Ashmore, R., Calinescu, R., and Paterson, C. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Computing Surveys (CSUR)*, 54(5): 1–39, 2021.

Budhathoki, K., Janzing, D., Bloebaum, P., and Ng, H. Why did the distribution change? In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1666–1674. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/budhathoki21a.html.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Chirra, P., Leo, P., Yim, M., Bloch, B. N., Rastinehad, A. R., Purysko, A., Rosen, M., Madabhushi, A., and Viswanath, S. Empirical evaluation of cross-site reproducibility in radiomic features for characterizing prostate mri. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, pp. 105750B. International Society for Optics and Photonics, 2018.

d'Eon, G., d'Eon, J., Wright, J. R., and Leyton-Brown, K. The spotlight: A general method for discovering systematic errors in deep learning models, 2021. URL https://arxiv.org/abs/2107.00758.

Eyuboglu, S., Varma, M., Saab, K. K., Delbrouck, J.-B., Lee-Messer, C., Dunnmon, J., Zou, J., and Re, C. Domino: Discovering systematic errors with cross-modal embeddings. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=FPCMqjI0jXN.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Guo, L. L., Pfohl, S. R., Fries, J., Johnson, A. E., Posada, J., Aftandilian, C., Shah, N., and Sung, L. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Scientific reports*, 12(1):1–10, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jacobs, A. Z. and Wallach, H. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 375–385, 2021.

Janzing, D., Minorics, L., and Blöbaum, P. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics*, pp. 2907–2916. PMLR, 2020.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Kumar, I., Scheidegger, C., Venkatasubramanian, S., and Friedler, S. Shapley residuals: Quantifying the limits of the shapley value for explanations. *Advances in Neural Information Processing Systems*, 34, 2021.

Liang, W. and Zou, J. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. *arXiv preprint arXiv:2202.06523*, 2022.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Nestor, B., McDermott, M. B., Boag, W., Berner, G., Naumann, T., Hughes, M. C., Goldenberg, A., and Ghassemi, M. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. In *Machine Learning for Healthcare Conference*, pp. 381–405. PMLR, 2019.

Pearl, J. *Causality*. Cambridge university press, 2009.

Pearl, J. and Bareinboim, E. Transportability of causal and statistical relations: A formal approach. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

Rabanser, S., Günnemann, S., and Lipton, Z. Failing loudly: An empirical study of methods for detecting dataset shift. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/846c260d715e5b854ffad5f70a516c88-Paper.pdf.

Roth, A. E. *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, 1988.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, prediction, and search*. MIT press, 2000.

Subbaswamy, A., Schulam, P., and Saria, S. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3118–3127, 2019.

Subbaswamy, A., Adams, R., and Saria, S. Evaluating model robustness and stability to dataset shift. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2611–2619. PMLR, 13–15 Apr 2021. URL http://proceedings.mlr.press/v130/subbaswamy21a.html.

Sugiyama, M., Suzuki, T., and Kanamori, T. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012.

Sundararajan, M. and Najmi, A. The many shapley values for model explanation. In *International conference on machine learning*, pp. 9269–9278. PMLR, 2020.

Štrumbelj, E. and Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.*, 41(3):647–665, dec 2014. ISSN 0219-1377. doi: 10.1007/s10115-013-0679-x. URL https://doi.org/10.1007/s10115-013-0679-x.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. 2011.

Wang, J., Wiens, J., and Lundberg, S. Shapley flow: A graph-based approach to interpreting model predictions. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 721–729. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/wang21b.html.

Wu, E., Wu, K., and Zou, J. Explaining medical ai performance disparities across sites with confounder shapley value analysis, 2021. URL https://arxiv.org/abs/2111.08168.

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pp. 3987–3995. PMLR, 2017.

A. Properties of the Shapley values

The Shapley values resulting from the performance-change game have the following two desired properties.

Property 1. $\sum_{d \in \mathtt{C}_{\mathcal{D}}} \text{Attr}(d) = \text{Val}(\mathtt{C}_{\mathcal{D}})$

By definition, we know that the sum of Shapley values equal the value of the all-player coalition. Thus, we distribute the total performance change due to the shift from train to test distribution to the shifts in causal mechanisms in the candidate set.

Property 2. $\mathcal{D}_d^{\text{train}} = \mathcal{D}_d^{\text{test}} \implies \text{Attr}(d) = 0.$

We can verify that our method gives zero attribution to distributions that do not shift between the train and test. First we observe that $\text{Val}(\widetilde{\mathcal{D}}) = \text{Val}(\widetilde{\mathcal{D}} \setminus d)$ if $\mathcal{D}_d^{\text{train}} = \mathcal{D}_d^{\text{test}}$. This is because $\widetilde{\mathcal{D}} = \widetilde{\mathcal{D}} \setminus d$ for any $\widetilde{\mathcal{D}} \in \mathtt{C}_{\mathcal{D}}$ since the factor corresponding to d remains the same between train and test even when it is allowed to change as part of the coalition. By definition of Shapley value in Eq. 2, $\text{Attr}(d) = 0$.

Thus, the method attributes the overall performance change to distributions that actually change. The contribution of each distribution is computed by considering how much they impact the performance if they are made to change in different combinations alongside the other distributions.

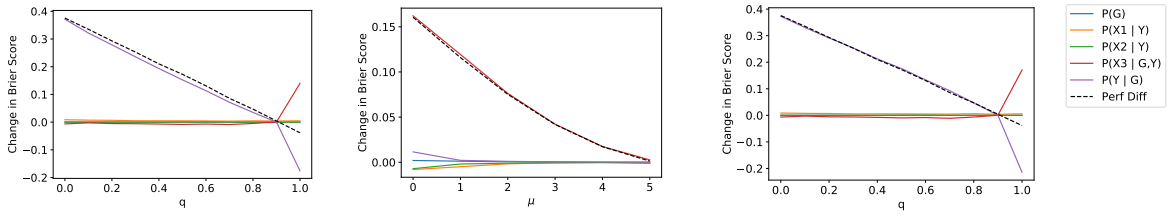
B. Computing density ratios using a probabilistic classifier

We outline an approach for density ratio estimation via probabilistic classifiers as described in Sugiyama et al. (2012, Section 2.2).

Say, we want to compute the density ratio $w = \frac{\mathbb{P}_{\mathcal{D}^{\text{test}}}(X)}{\mathbb{P}_{\mathcal{D}^{\text{train}}}(X)}$ for some subset of variables $X \subseteq V$. Add a new binary variable D such that when $D = 1, X \sim \mathcal{D}^{\text{test}}(X)$, and when $D = 0, X \sim \mathcal{D}^{\text{train}}(X)$. Then train a probabilities classifier to predict D from X, i.e. $\mathbb{P}(D = 1|X)$.

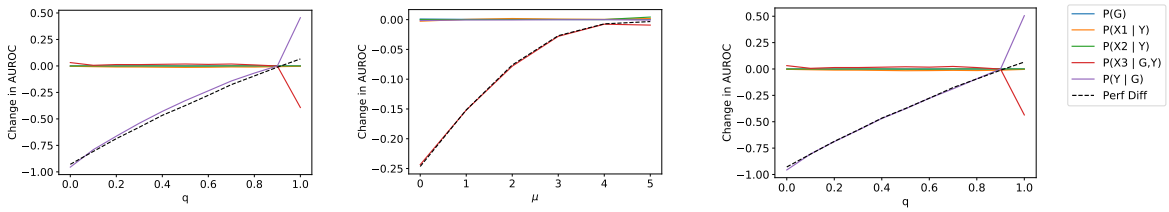
Compute the ratio as $w = \frac{\mathbb{P}(X|D=1)}{\mathbb{P}(X|D=0)} = \frac{\mathbb{P}(D=1|X)}{\mathbb{P}(D=1)} \frac{\mathbb{P}(D=0)}{\mathbb{P}(D=0|X)} = \frac{\mathbb{P}(D=1|X)}{1-\mathbb{P}(D=1|X)} \frac{\mathbb{P}(D=0)}{\mathbb{P}(D=1)}$. The first term is obtained from the classifier's output and the second is a known quantity equal to the ratio of the sizes of the train and test datasets.

C. Additional Experimental Results On Synthetic Data



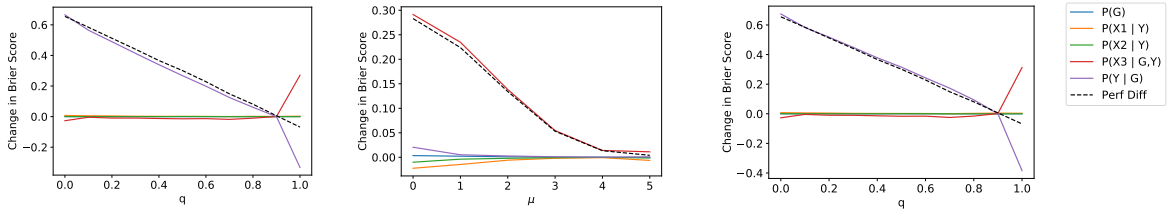
(a) Setting 1 (b) Setting 2 (c) Setting 3

Figure C.1. Brier score differences attributed by our model to five potential distributional shifts on the synthetic dataset for the `LR` model.



(a) Setting 1 (b) Setting 2 (c) Setting 3

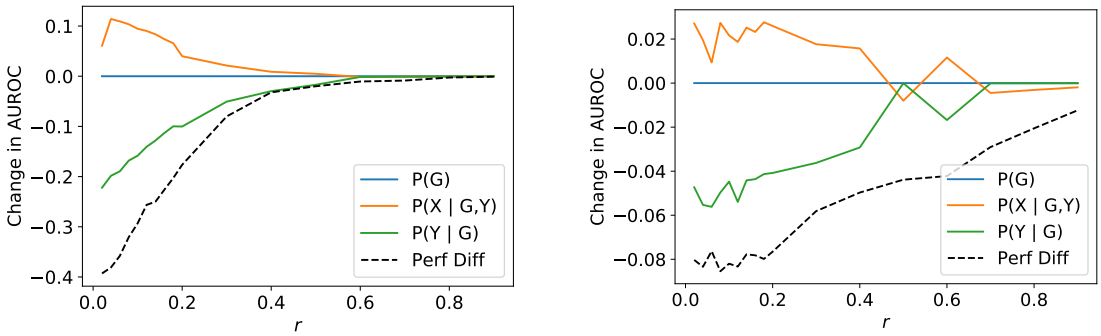
Figure C.2. AUROC differences attributed by our model to five potential distributional shifts on the synthetic dataset for the `XGB` model.



(a) Setting 1 (b) Setting 2 (c) Setting 3

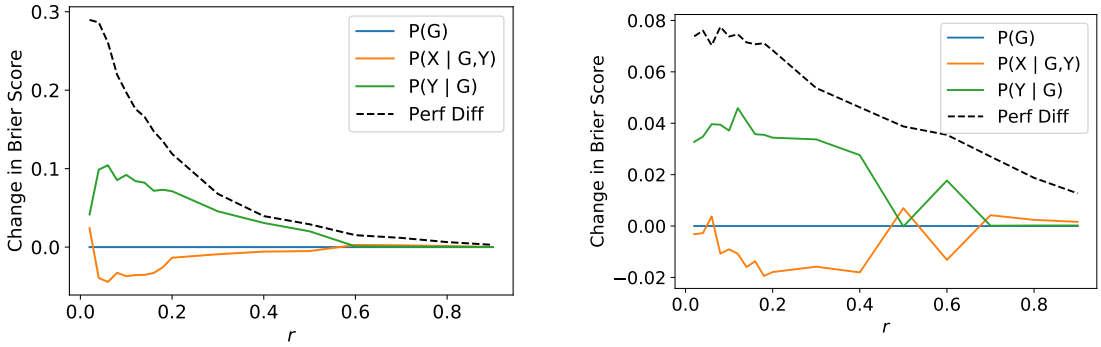
Figure C.3. Brier score differences attributed by our model to five potential distributional shifts on the synthetic dataset for the `XGB` model.

D. Experimental Results On Real-World Data



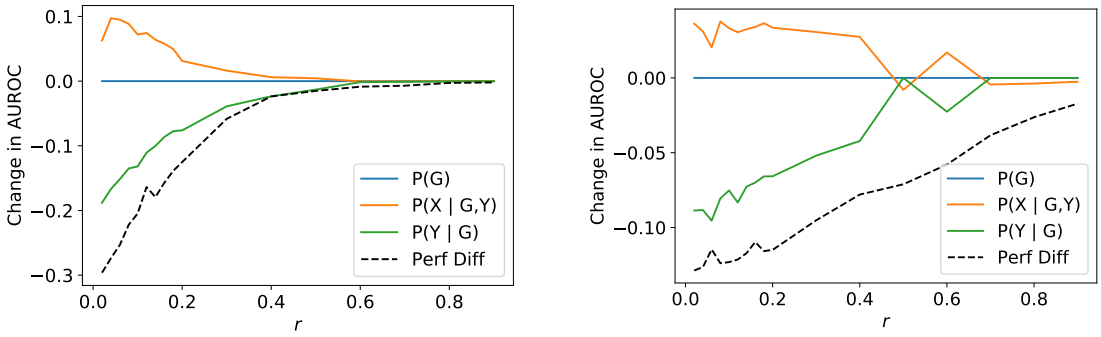
(a) Waterbirds (b) MetaShift

Figure D.4. AUROC differences attributed by our model to three potential distributional shifts on two real-world datasets for the `LR` model.



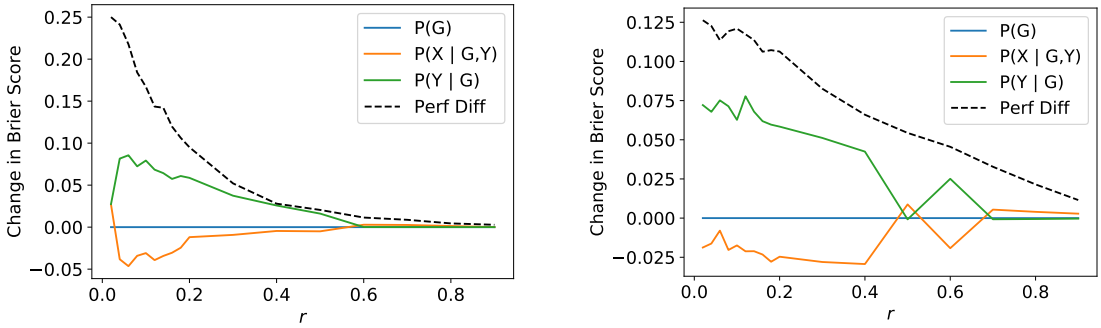
(a) Waterbirds (b) MetaShift

Figure D.5. Brier score differences attributed by our model to three potential distributional shifts on two real-world datasets for the `LR` model.



(a) Waterbirds (b) MetaShift

Figure D.6. AUROC differences attributed by our model to three potential distributional shifts on two real-world datasets for the `XGB` model.



(a) Waterbirds (b) MetaShift

Figure D.7. Brier score differences attributed by our model to three potential distributional shifts on two real-world datasets for the `XGB` model.