Explanation Shift: Detecting distribution shifts on tabular data via the explanation space

Carlos Mougan
University of Southampton
United Kingdom
C.Mougan@soton.ac.uk

Klaus Broelemann
Schufa Holding AG
Wiesbaden, Germany
klaus.broelemann@schufa.de

Gjergji Kasneci
University of Tuebingen
Tuebingen, Germany
gjergji.kasneci@uni-tuebingen.de

Thanassis Tiropanis
University of Southampton
United Kingdom
t.tiropanis@southampton.ac.uk

Steffen Staab
University of Stuttgart, Germany
University of Southampton, UK
s.r.staab@soton.ac.uk

Abstract

As input data distributions evolve, the predictive performance of machine learning models tends to deteriorate. In the past, predictive performance was considered the key indicator to monitor. However, explanation aspects have come to attention within the last years. In this work, we investigate how model predictive performance and model explanation characteristics are affected under distribution shifts and how these key indicators are related to each other for tabular data. We find that the modeling of explanation shifts can be a better indicator for the detection of predictive performance changes than state-of-the-art techniques based on representations of distribution shifts. We provide a mathematical analysis of different types of distribution shifts as well as synthetic experimental examples.

1 Introduction

Machine learning theory gives us the means to forecast the quality of ML models on unseen data, provided that this data is sampled from the same distribution as the data used to train and to evaluate the model. If unseen data is sampled from a different distribution, model quality may deteriorate.

Model monitoring tries to signal and possibly even quantify such decay of trained models. Such monitoring is challenging, because only in few applications unseen data comes with labels that allow for monitoring model quality directly. Much more often, deployed ML models encounter unseen data for which target labels are lacking or biased Rabanser et al. (2019); Huyen (2022).

Detecting changes in the quality of deployed ML models in the absence of labeled data is a challenging question both in theory and practice Ramdas et al. (2015); Rabanser et al. (2019). In practice, some of the most straightforward approaches are based on statistical distances between training and unseen data distributions Diethe et al. (2019); Labs (2021) or on actual model predictions Garg et al. (2021b,a); Mougan and Nielsen (2022). The shortcomings of these measures of *distribution shifts* is that they do not relate changes of distributions to how they incur effects in the trained models.

Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

The field of explainable AI has emerged as a way to understand model decisions Molnar (2019) and interpret the inner workings of black box models Guidotti et al. (2018). The core idea of this paper is to use *explanation shift* for signaling distribution shift that affect the model behavior. We newly define explanation shift to be constituted by the statistical comparison between how predictions from source data are explained and how predictions on unseen data are explained. Explanation shift goes beyond the mere recognition of changes in data distributions towards the recognition of changes of how data distributions relate to the models' inner workings.

We study the problem of detecting distribution changes that impact model predictive performance on tabular data, which still constitutes a major field of application for machine learning. In contrast, most recent research on model degradation proposes monitoring methods for modalities such as texts or images and suggests invariances on the latent spaces of deep neural models, which are hardly applicable to tabular data, for which likewise progress has mostly been lacking in recent years. In summary, our contributions are:

- We propose measures of explanation shifts as a key indicator for detecting distribution shift that affect model behavior.
- We provide a mathematical analysis of three synthetic examples that shows how simple, but key types of distribution shift interact with linear models such that measures of explanation shift become much better indicators of model decay than measures of distribution shift or prediction shift.

2 Methodology

2.1 Formalization

The objective of supervised learning is to induce a function $f_\theta : X \to Y$, where f_θ is from a family of functions $f_\theta \in F$, from training set $\mathcal{D}^{tr} = \{(x_0^{tr}, y_0^{tr}) \ldots, (x_n^{tr}, y_n^{tr})\} \subseteq \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} \times \mathcal{Y}$ denote the domain of predictors X and target Y respectively. The estimated hypothesis f_θ is expected to generalize well on novel, previously unseen data $\mathcal{D}^{new} = \{x_0^{new}, x_1^{new}, \ldots\} \subseteq \mathcal{X}$, for which the target labels are unknown. The traditional machine learning assumption is that training data $\mathcal{D}^{tr}$ and novel data $\mathcal{D}^{new}$ are sampled from the same underlying distribution $P(X, Y)$. If we have a hold-out test data set $\mathcal{D}^{te} = \{(x_0^{te}, y_0^{te}) \ldots, (x_m^{te}, y_m^{te})\} \subseteq \mathcal{X} \times \mathcal{Y}$ disjoint from $\mathcal{D}^{tr}$, but also sampled from $P(X, Y)$, one may use $\mathcal{D}^{te}$ to estimate performance indicators for $\mathcal{D}^{new}$. Commonly, novel data is sampled from a distribution $P'(X, Y)$ that is different from $P(X, Y)$.We use $\mathcal{D}^{ood} \subseteq \mathcal{X} \times \mathcal{Y}$ to refer to such novel, out-of-distribution data.

Definition 1 *(Feature Attribution Explanation) We write $\mathcal{S}$ for an explanation function that takes a model f with parameter θ and data of interest x and returns the calculation of the Shapley values $\mathcal{S}(f_\theta, x)$, with p being the exact dimensions of the predictor x and signature $\mathcal{S} : F \times \mathcal{X} \to \mathbb{R}^p$*

Definition 2 *(Explanation Shift) For a measure of statistical distance between two explanations of the model f_θ between X and X', we write $d(\mathcal{S}(f_\theta, X), \mathcal{S}(f_\theta, X'))$.*

2.2 Explanation Shift: Detecting model performance changes via the explanation space

The following section provides three different examples of situations where changes in the explanation space can correctly account for model behavior changes. Where statistical checks on the input data (1) cannot detect changes, (2) require sophisticated methods to detect these changes, or (3) detect changes that do not affect model behavior. For simplicity the model used in the analytical examples is a linear regression where, if the features are independent, the Shapley value can be estimated by $\mathcal{S}(f_\theta, x_i) = a_i(x_i - \mu_i)$, where a_i are the coefficients of the linear model and μ_i the mean of the features Chen et al. (2020). Moreover, in the experimental section, we provide examples with synthetic data, and non-linear models.

2.2.1 Detecting multivariate shift

One challenging type of distribution shift to detect is cases where the univariate distributions for each feature j are equal between the source and the unseen dataset. Moreover, what changes are the

distribution interdependencies among different features. Multiple univariate testing offers comparable performance to multivariate testing Rabanser et al. (2019), but comparing distributions on high dimensional spaces is not an easy task. The following examples aim to demonstrate that Shapley values account for covariate interaction changes while a univariate statistical test will provide false negatives.

Example 1: ***Multivariate Shift*** *Let* $X = (X_1, X_2) \sim N\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma^2_{x_1} & 0\\ 0 & \sigma^2_{x_2}\end{bmatrix}\right)$ *and* $X^{ood} = (X_1^{ood}, X_2^{ood}) \sim N\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma^2_{x_1} & \rho\sigma_{x_1}\sigma_{x_2}\\ \rho\sigma_{x_1}\sigma_{x_2} & \sigma^2_{x_2}\end{bmatrix}\right)$. *We fit a linear model* $f_\theta(X_1, X_2) = \gamma + a \cdot X_1 + b \cdot X_2$. X_1 *and* X_2 *are identically distributed with* X_1^{ood} *and* X_2^{ood}*, respectively, while this does not hold for the corresponding SHAP values* $\mathcal{S}_j(f_\theta, X)$ *and* $\mathcal{S}_j(f_\theta, X^{ood})$*. Analytical demonstration in the Appendix*

2.2.2 Detecting posterior distribution shift

One of the most challenging types of distribution shift to detect are cases where distributions are equal between source and unseen data-set $P(X^{tr}) = P(X^{ood})$ and the target variable $P(Y^{tr}) = P(Y^{ood})$ and what changes are the relationships that features have with the target $P(Y^{tr}|X^{tr}|) \neq P(Y^{ood}|X^{ood}|)$, this kind of distribution shift is also known as concept drift or posterior shift Huyen (2022) and is especially difficult to notice, as it requires labeled data to detect. The following example compares how the explanations change for two models fed with the same input data and different target relations.

Example 2: ***Posterior shift*** *Let* $X = (X_1, X_2) \sim N(\mu, I)$*, and* $X^{ood} = (X_1^{ood}, X_2^{ood}) \sim N(\mu, I)$*, where* I *is an identity matrix of order two and* $\mu = (\mu_1, \mu_2)$*. We now create two synthetic targets* $Y = a + \alpha \cdot X_1 + \beta \cdot X_2 + \epsilon$ *and* $Y^{ood} = a + \beta \cdot X_1 + \alpha \cdot X_2 + \epsilon$*. Let* f_θ *be a linear regression model trained on* $f(X, Y)$ *and* g_ψ *another linear model trained on* (X^{ood}, Y^{ood})*. Then* $P(f_\theta(X)) = P(g_\psi(X^{ood}))$*,* $P(X) = P(X^{ood})$ *but* $\mathcal{S}(f_\theta, X) \neq \mathcal{S}(g_\psi, X)$*.*

2.2.3 Shifts on uninformative features by the model

Another typical problem is false positives when a statistical test flags a distribution difference between source and unseen distributions that do not affect the model behaviorGrinsztajn et al. (2022). One of the intrinsic properties that Shapley values satisfy is the "Dummy", where a feature j that does not change the predicted value, regardless of which coalition the feature is added, should have a Shapley value of 0. If $\text{val}(S \cup \{j\}) = \text{val}(S)$ for all $S \subseteq \{1, \ldots, p\}$ then $\mathcal{S}(f_\theta, x_j) = 0$.

Example 3: ***Unused features*** *Let* $X = (X_1, X_2, X_3) \sim N(\mu, c \cdot I)$*, and* $X^{ood} = (X_1^{ood}, X_2^{ood}, X_3^{ood}) \sim N(\mu, c' \cdot I)$*, where* I *is an identity matrix of order three and* $\mu = (\mu_1, \mu_2, \mu_3)$*. We now create a synthetic target* $Y = a_0 + a_1 \cdot X_1 + a_2 \cdot X_2 + \epsilon$ *that is independent of* X_3*. We train a linear regression* f_θ *on* (X, Y)*, with coefficients* a_0, a_1, a_2, a_3*. Then* $P(X_3)$ *can be different from* $P(X_3^{ood})$ *but* $\mathcal{S}_3(f_\theta, X) = \mathcal{S}_3(f_\theta, X^{ood})$

3 Experiments

The experimental section explores the detection of distribution shift on synthetic examples. We perform statistical testing between input data distributions and explanation space.

3.1 Detecting multivariate shift

Given two bivariate normal distributions $X = (X_1, X_2) \sim N\left(0, \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\right)$ and $X^{ood} = (X_1^{ood}, X_2^{ood}) \sim N\left(0, \begin{bmatrix}1 & 0.2\\ 0.2 & 1\end{bmatrix}\right)$, then, for each feature j the underlying distribution is equally distributed between X and X^{ood}, $\forall j \in \{1, 2\} : P(X_j) = P(X_j^{ood})$, and what is different are the interaction terms between them. We now create a synthetic target $Y = X_1 \cdot X_2 + \epsilon$ with $\epsilon \sim N(0, 0.1)$

and fit a gradient boosting decision tree $f_\theta(X)$. Then we compute the SHAP explanation values for $\mathcal{S}(f_\theta, X)$ and $\mathcal{S}(f_\theta, X^{ood})$

Table 1: Displayed results are the one-tailed p-values of the Kolmogorov-Smirnov test comparison between two underlying distributions. Small p-values indicates that compared distributions would be very unlikely to be equally distributed. SHAP values correctly indicate the interaction changes that individual distribution comparisons cannot detect

Comparison	**p-value**	**Conclusions**
$P(X_1)$, $P(X_1^{ood})$	0.33	Not Distinct
$P(X_2)$, $P(X_2^{ood})$	0.60	Not Distinct
$\mathcal{S}_1(f_\theta, X)$, $\mathcal{S}_1(f_\theta, X^{ood})$	3.9e−153	Distinct
$\mathcal{S}_2(f_\theta, X)$, $\mathcal{S}_2(f_\theta, X^{ood})$	2.9e−148	Distinct

Having drawn $50,000$ samples from both X and X^{ood}, in Table 1, we evaluate whether changes on the input data distribution or on the explanations are able to detect changes on covariate distribution. For this, we compare the one-tailed p-values of the Kolmogorov-Smirnov test between the input data distribution, and the explanations space. Explanation shift correctly detects the multivariate distribution change that univariate statistical testing can not detect.

3.1.1 Detecting posterior shift

Given a bivariate normal distribution $X = (X_1, X_2) \sim N(1, I)$ where I is an identity matrix of order two. We now create two synthetic targets $Y = X_1^2 \cdot X_2 + \epsilon$ and $Y^{ood} = X_1 \cdot X_2^2 + \epsilon$ and fit two machine learning models f_θ on (X, Y) and h_Υ on (X, Y^{ood}). Now we compute the SHAP values for $\mathcal{S}(f_\theta, X)$ and $\mathcal{S}(h_\Upsilon, X)$

Table 2: Distribution comparison for synthetic posterior shift. Displayed results are the one-tailed p-values of the Kolmogorov-Smirnov test comparison between two underlying distributions

Comparison	**Conclusions**
$P(X)$, $P(X^{ood})$	Not Distinct
$P(Y)$, $P(Y^{ood})$	Not Distinct
$P(f_\theta(X))$, $P(h_\Upsilon(X^{ood}))$	Not Distinct
$\mathcal{S}(f_\theta, X)$, $\mathcal{S}(h_\Upsilon, X)$	Distinct

In Table 2, we see how the distribution shifts are not able to capture the change in the model behavior while the SHAP values are different. The "Distinct/Not distinct" conclusion is based on the one-tailed p-value of the Kolmogorov-Smirnov test with a 0.05 threshold drawn out of $50,000$ samples for both distributions. As in the theoretical example, in table 2 SHAP values can detect a relational change between X and Y, even if both distributions remain equivalent.

4 Conclusions

Traditionally, the problem of detecting model degradation has relied on measurements of shifting input data distributions or shifting distributions of predictions. In this paper, we have provided theoretical and experimental evidence that explanation shift can be a more suitable indicator to detect and quantify decay of predictive performance. We have provided mathematical analysis examples and synthetic data experimental evaluation. We found that measures of explanation shift can outperform measures of distribution and prediction shift.

Limitations: Without any assumptions on the type of shift, estimating model decay is a challenging task, where no estimator will be the best under all the types of shift Garg et al. (2021b). We compared how well measures of explanation shift would perform relative to distribution shift and found encouraging results. The potential utility of explanation shifts as indicators for predictive performance and fairness in computer vision or natural language processing tasks remains an open question. We have used Shapley values to derive indications of explanation shifts, but we believe that other AI explanation techniques and specifically, other feature attribution methods, logical

reasoning, argumentation, or counterfactual explanations, may be applicable and come with their own advantages.

Reproducibility Statement

To ensure reproducibility, we make the data, code repositories, and experiments publicly available [1]. For our experiments, we used default `scikit-learn` parameters Pedregosa et al. (2011). We describe the system requirements and software dependencies of our experiments. Experiments were run on a 16 vCPU server with 60 GB RAM.

Acknowledgments

This work has received funding by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions (grant agreement number 860630) for the project : "NoBIAS - Artificial Intelligence without Bias". Furthermore, this work reflects only the authors' view and the European Research Executive Agency (REA) is not responsible for any use that may be made of the information it contains.

References

Aas, K., Jullum, M., and Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artif. Intell.*, 298:103502.

Arrieta, A. B., Rodríguez, N. D., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, 58:82–115.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. (2021). Deep neural networks and tabular data: A survey.

Chen, H., Covert, I. C., Lundberg, S. M., and Lee, S. (2022). Algorithms to estimate shapley value feature attributions. *CoRR*, abs/2207.07605.

Chen, H., Janizek, J. D., Lundberg, S. M., and Lee, S. (2020). True to the model or true to the data? *CoRR*, abs/2006.16234.

Diethe, T., Borchert, T., Thereska, E., Balle, B., and Lawrence, N. (2019). Continual learning in practice. ArXiv preprint, `https://arxiv.org/abs/1903.05202`.

Elsayed, S., Thyssens, D., Rashed, A., Schmidt-Thieme, L., and Jomaa, H. S. (2021). Do we really need deep learning models for time series forecasting? *CoRR*, abs/2101.02118.

Fort, S., Ren, J., and Lakshminarayanan, B. (2021). Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34.

Frye, C., Rowat, C., and Feige, I. (2020). Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Garg, S., Balakrishnan, S., Kolter, Z., and Lipton, Z. (2021a). Ratt: Leveraging unlabeled data to guarantee generalization. In *International Conference on Machine Learning*, pages 3598–3609. PMLR.

Garg, S., Balakrishnan, S., Lipton, Z. C., Neyshabur, B., and Sedghi, H. (2021b). Leveraging unlabeled data to predict out-of-distribution performance. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*.

[1] `https://anonymous.4open.science/r/ExplanationShift-691E`

Garg, S., Wu, Y., Balakrishnan, S., and Lipton, Z. (2020). A unified view of label shift estimation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3290–3300. Curran Associates, Inc.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? working paper or preprint.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5).

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Hendrycks, D. and Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Huang, R., Geng, A., and Li, Y. (2021). On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, abs/2110.00218.

Huyen, C. (2022). *Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications*. O'Reilly.

Labs, C. F. (2021). Inferring concept drift without labeled data. `https://concept-drift.fastforwardlabs.com/`.

Lee, K., Lee, K., Lee, H., and Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 7167–7177.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2019). Explainable ai for trees: From local explanations to global understanding.

Lundberg, S. M. and Lee, S. (2017). A unified approach to interpreting model predictions. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774.

Mittelstadt, B. D., Russell, C., and Wachter, S. (2019). Explaining explanations in AI. In danah boyd and Morgenstern, J. H., editors, *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 279–288. ACM.

Molnar, C. (2019). *Interpretable Machine Learning*. . `https://christophm.github.io/interpretable-ml-book/`.

Mougan, C. and Nielsen, D. S. (2022). Monitoring model deterioration with explainable uncertainty estimation via non-parametric bootstrap. *CoRR*, abs/2201.11676.

Park, C., Awadalla, A., Kohno, T., and Patel, S. N. (2021). Reliable and trustworthy machine learning for health using dataset shift detection. *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, abs/2110.14019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Quiñonero-Candela, J., Sugiyama, M., Lawrence, N. D., and Schwaighofer, A. (2009). *Dataset shift in machine learning*. Mit Press.

Rabanser, S., Günnemann, S., and Lipton, Z. C. (2019). Failing loudly: An empirical study of methods for detecting dataset shift. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1394–1406.

Ramdas, A., Reddi, S. J., Póczos, B., Singh, A., and Wasserman, L. A. (2015). On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In Bonet, B. and Koenig, S., editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 3571–3577. AAAI Press.

Selbst, A. D. and Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham L. Rev.*, 87:1085.

Shapley, L. S. (1953). *A Value for n-Person Games*, pages 307–318. Princeton University Press.

A Foundations and Related Work

A.1 Explainable AI

Explainability has become an important concept in legal and ethical guidelines for data processing, and machine learning applications Selbst and Barocas (2018). A wide variety of methods have been developed aiming to account for the decision of algorithmic systems Guidotti et al. (2018); Mittelstadt et al. (2019); Arrieta et al. (2020). One of the most popular approaches to machine learning explainability has been the use of Shapley values to attribute relevance to features used by the model Lundberg et al. (2019); Lundberg and Lee (2017). The Shapley value is a concept from coalition game theory that aims to allocate the surplus generated by the grand coalition in a game to each of its players Shapley (1953). In a general sense, the Shapley value $\mathcal{S}_j$ for the j'th player can be defined via a value function $\mathrm{val} : 2^N \to \mathbb{R}$ of players in T:

$$\mathcal{S}_j(\mathrm{val}) = \sum_{T \subseteq N \setminus \{j\}} \frac{|T|!(p - |T| - 1)!}{p!} (\mathrm{val}(T \cup \{j\}) - \mathrm{val}(T)) \tag{1}$$

In the case of features, T is a subset of $N = \{1, \ldots, p\}$, i.e., the features used in the model, and x is the vector of feature values of the instance to be explained. The term $\mathrm{val}_x(T)$ represents the prediction for the feature values in T that are marginalized over features that are not included in T:

$$\mathrm{val}_x(T) = E_{X_{N \setminus T}}[\hat{f}(X)|X_T = x_T] - E_X[\hat{f}(X)] \tag{2}$$

It is important to differentiate between the theoretical Shapley values and the different implementations that approximate them. We use TreeSHAP as an efficient implementation of an approach for tree-based models of Shapley values Lundberg et al. (2019); Molnar (2019), particularly we use the observational (or path-dependent) estimation Chen et al. (2022); Frye et al. (2020); Chen et al. (2020) and for linear models we use the correlation dependent implementation that takes into account feature dependencies Aas et al. (2021).

A.2 Related Work

Evaluating how two distributions differ has been a widely studied topic in the statistics and statistical learning literature Hastie et al. (2001); Quiñonero-Candela et al. (2009). Rabanser et al. (2019) provide a comprehensive empirical investigation, examining how dimensionality reduction and two-sample testing might be combined to produce a practical pipeline for detecting distribution shifts in real-life machine learning systems. A few popular techniques to detect out-of-distribution data using neural networks are based on the prediction space Fort et al. (2021); Garg et al. (2020), using the maximum softmax probabilities/likelihood as a confidence score, extracting information out of the gradient space Huang et al. (2021), fitting a Gaussian distribution to the embedding or using the Mahalanobis distance for out-of-distribution detection Hendrycks and Gimpel (2017); Lee et al.

(2018); Park et al. (2021). These methods are built explicitly for neural networks, and often they can not be directly applied to traditional machine learning techniques. In our work, we are focusing specifically on tabular data where techniques such as gradient boosting decision trees achieve state of the art on model performance Grinsztajn et al. (2022); Elsayed et al. (2021); Borisov et al. (2021).

The first approach of using explainability to detect changes in the model was suggested by Lundberg et al. (2019) who monitored the SHAP value contribution in order to identify possible bugs in the pipeline. This technique was initially used to account for previously unnoticed bugs in a local monitoring scenario in the machine learning production pipeline. In our work, we study how SHAP value changes can be used as an indicator to monitor prediction performance and fairness.

B Analytical examples

This section covers the analytical examples demonstrations presented in the Section 2.2 of the main body of the paper.

B.1 Multivariate shift

Example 1: ***Multivariate Shift*** *Let* $X = (X_1, X_2) \sim N\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma^2_{x_1} & 0\\ 0 & \sigma^2_{x_2}\end{bmatrix}\right)$ *and* $X^{ood} = (X_1^{ood}, X_2^{ood}) \sim N\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma^2_{x_1} & \rho\sigma_{x_1}\sigma_{x_2}\\ \rho\sigma_{x_1}\sigma_{x_2} & \sigma^2_{x_2}\end{bmatrix}\right)$. *We fit a linear model* $f_\theta(X_1, X_2) = \gamma + a \cdot X_1 + b \cdot X_2$. X_1 *and* X_2 *are identically distributed with* X_1^{ood} *and* X_2^{ood}*, respectively, while this does not hold for the corresponding SHAP values* $\mathcal{S}_j(f_\theta, X)$ *and* $\mathcal{S}_j(f_\theta, X^{ood})$.

$$\mathcal{S}_1(f_\theta, x) = a(x_1 - \mu_1) \tag{3}$$

$$\mathcal{S}_1(f_\theta, x^{ood}) = \tag{4}$$

$$= \frac{1}{2}[\text{val}(\{1,2\}) - \text{val}(\{2\})] + \frac{1}{2}[\text{val}(\{1\}) - \text{val}(\emptyset)] \tag{5}$$

$$\text{val}(\{1,2\}) = E[f_\theta | X_1 = x_1, X_2 = x_2] = ax_1 + bx_2 \tag{6}$$

$$\text{val}(\emptyset) = E[f_\theta] = a\mu_1 + b\mu_2 \tag{7}$$

$$\text{val}(\{1\}) = E[f_\theta(x) | X_1 = x_1] + b\mu_2 \tag{8}$$

$$\text{val}(\{1\}) = \mu_1 + \rho\frac{\rho_{x_1}}{\sigma_{x_2}}(x_1 - \sigma_1) + b\mu_2 \tag{9}$$

$$\text{val}(\{2\}) = \mu_2 + \rho\frac{\sigma_{x_2}}{\sigma_{x_1}}(x_2 - \mu_2) + a\mu_1 \tag{10}$$

$$\rightarrow \mathcal{S}_1(f_\theta, x^{ood}) \neq a(x_1 - \mu_1) \tag{11}$$

B.2 Posterior Shift

Example 2: ***Posterior shift*** *Let* $X = (X_1, X_2) \sim N(\mu, I)$*, and* $X^{ood} = (X_1^{ood}, X_2^{ood}) \sim N(\mu, I)$*, where* I *is an identity matrix of order two and* $\mu = (\mu_1, \mu_2)$*. We now create two synthetic targets* $Y = a + \alpha \cdot X_1 + \beta \cdot X_2 + \epsilon$ *and* $Y^{ood} = a + \beta \cdot X_1 + \alpha \cdot X_2 + \epsilon$*. Let* f_θ *be a linear regression model trained on* $f(X, Y)$ *and* g_ψ *another linear model trained on* (X^{ood}, Y^{ood})*. Then* $P(f_\theta(X)) = P(g_\psi(X^{ood}))$,

$P(X) = P(X^{ood})$ *but* $\mathcal{S}(f_\theta, X) \neq \mathcal{S}(g_\psi, X)$.

$$X \sim N(\mu, \sigma^2 \cdot I), X^{ood} \sim N(\mu, \sigma^2 \cdot I) \tag{12}$$

$$\rightarrow P(X) = P(X^{ood}) \tag{13}$$

$$Y \sim a + \alpha N(\mu, \sigma^2) + \beta N(\mu, \sigma^2) + N(0, \sigma^{'2}) \tag{14}$$

$$Y^{ood} \sim a + \beta N(\mu, \sigma^2) + \alpha N(\mu, \sigma^2) + N(0, \sigma^{'2}) \tag{15}$$

$$\rightarrow P(Y) = P(Y^{ood}) \tag{16}$$

$$\mathcal{S}(f_\theta, X) = \begin{pmatrix} \alpha(X_1 - \mu_1) \\ \beta(X_2 - \mu_2) \end{pmatrix} \sim \begin{pmatrix} N(\mu_1, \alpha^2\sigma^2) \\ N(\mu_2, \beta^2\sigma^2) \end{pmatrix} \tag{17}$$

$$\mathcal{S}(g_\psi, X) = \begin{pmatrix} \beta(X_1 - \mu_1) \\ \alpha(X_2 - \mu_2) \end{pmatrix} \sim \begin{pmatrix} N(\mu_1, \beta^2\sigma^2) \\ N(\mu_2, \alpha^2\sigma^2) \end{pmatrix} \tag{18}$$

$$\text{If} \quad \alpha \neq \beta \rightarrow \mathcal{S}(f_\theta, X) \neq \mathcal{S}(g_\psi, X) \tag{19}$$

B.3 Uninformative Features

Example 3: ***Unused features*** *Let* $X = (X_1, X_2, X_3) \sim N(\mu, c \cdot I)$*, and* $X^{ood} = (X_1^{ood}, X_2^{ood}, X_3^{ood}) \sim N(\mu, c' \cdot I)$*, where* I *is an identity matrix of order three and* $\mu = (\mu_1, \mu_2, \mu_3)$*. We now create a synthetic target* $Y = a_0 + a_1 \cdot X_1 + a_2 \cdot X_2 + \epsilon$ *that is independent of* X_3*. We train a linear regression* f_θ *on* (X, Y)*, with coefficients* a_0, a_1, a_2, a_3*. Then* $P(X_3)$ *can be different from* $P(X_3^{ood})$ *but* $\mathcal{S}_3(f_\theta, X) = \mathcal{S}_3(f_\theta, X^{ood})$

$$X_3 \sim N(\mu_3, c_3), X_3^{ood} \sim N(\mu_3^{'}, c_3^{'}) \tag{20}$$

$$\text{If} \quad \mu_3^{'} \neq \mu_3 \quad \text{or} \quad c_3^{'} \neq c_3 \rightarrow P(X_3) \neq P(X_3^{ood}) \tag{21}$$

$$\mathcal{S}(f_\theta, X) = \left(\begin{bmatrix} a_1(X_1 - \mu_1) \\ a_2(X_2 - \mu_2) \\ a_3(X_3 - \mu_3) \end{bmatrix} \right) = \left(\begin{bmatrix} a_1(X_1 - \mu_1) \\ a_2(X_2 - \mu_2) \\ 0 \end{bmatrix} \right) \tag{22}$$

$$\mathcal{S}_3(f_\theta, X) = \mathcal{S}_3(f_\theta, X^{ood}) \tag{23}$$

C Synthetic data experiments

This section covers the last experiment of uninformative features on synthetic data that aims at providing empirical evidence about using the explanation space as (cf. Section 3)

C.1 Uninformative features on synthetic data

To have an applied use case of the theoretical example from the methodology section, we create a three-variate normal distribution $X = (X_1, X_2, X_3) \sim N(0, I_3)$, where I_3 is an identity matrix of order three. The target variable is generated $Y = X_1 \cdot X_2 + \epsilon$ being independent of X_3. For both, training and test data, $50,000$ samples are drawn. Then out-of-distribution data is created by shifting X_3, which is independent of the target, on test data $X_3^{ood} = X_3^{te} + 1$.

Table 3: Distribution comparison when modifying a random noise variable on test data. Where $\mathcal{L}$ is a metric evaluating the model predictive performance such as accuracy

Comparison	**Conclusions**
$P(X_3^{te})$, $P(X_3^{ood})$	Distinct
$\mathcal{L}(f_\theta, X^{te})$, $\mathcal{L}(f_\theta, X^{ood})$	Not Distinct
$\mathcal{S}(f_\theta, X^{te})$, $\mathcal{S}(f_\theta, X^{ood})$	Not Distinct

In Table 3, we see how an unused feature has changed the input distribution, but the explanation space and performance evaluation metrics remain the same. The "Distinct/Not Distinct" conclusion is based on the one-tailed p-value of the Kolmogorov-Smirnov test drawn out of $50,000$ samples for both distributions.