# Explanation Shift:
# How Did the Distribution Shift Impact the Model?

**Anonymous authors**
**Paper under double-blind review**

## Abstract

The performance of machine learning models on new data is critical for their success in real-world applications. Current methods to detect shifts in the input or output data distributions have limitations in identifying model behaviour changes when no labelled data is available. In this paper, we define *explanation shift* as the statistical comparison between how predictions from training data are explained and how predictions on new data are explained. We propose explanation shift as a key indicator to investigate the interaction between distribution shifts and learned models. We introduce an Explanation Shift Detector that operates on the explanation distributions, providing more sensitive and explainable changes in interactions between distribution shifts and learned models. We compare explanation shifts with other methods that are based on distribution shifts, showing that monitoring for explanation shifts results in more sensitive indicators for varying model behavior. We provide theoretical and experimental evidence and demonstrate the effectiveness of our approach on synthetic and real data. Additionally, we release an open-source Python package, `skshift`, which implements our method and provides usage tutorials for further reproducibility.

## 1 Introduction

ML theory provides 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 evaluate the model. If unseen data is sampled from a different distribution than the training data, model quality may deteriorate, making monitoring how the model's behavior changes crucial.

Recent research has highlighted the impossibility of reliably estimating the performance of machine learning models on unseen data sampled from a different distribution in the absence of further assumptions about the nature of the shift (Ben-David et al., 2010; Lipton et al., 2018; Garg et al., 2021). State-of-the-art techniques attempt to model statistical distances between the distributions of the training and unseen data or the distributions of the model predictions. However, these measures of *distribution shifts* only partially relate to changes of interaction between new data and trained models, or they rely on the availability of a causal graph or types of shift assumptions, which limits their applicability, as such domain assumptions may be available in the domain of images but are rarely or never available for tabular data. Therefore, tabular data is an economically hugely important domain that brings unique challenges. Thus, it is often necessary to go beyond detecting changes in input data distributions and understanding how they impact and relate to changes in the model given that performance degradation can not be accurately estimated.

The field of explainable AI has emerged as a way to understand model decisions and interpret the inner workings of ML models. The core idea of this paper is to go beyond the modeling of distribution shifts and monitor for *explanation shifts* to signal a change of interactions between learned models and dataset features in tabular data. We newly define explanation shift as the statistical comparison between how predictions from training data are explained and how predictions on new data are explained. In summary, our contributions are:

- We propose measures of explanation shifts as a key indicator for investigating the interaction between distribution shifts and learned models.

- We define an *Explanation Shift Detector* that operates on the explanation distributions, allowing for more sensitive and explainable changes of interactions between distribution shifts and learned models in Section 3.
- We compare our monitoring method that is based on explanation shifts with methods that are based on other kinds of distribution shifts. We find that monitoring for explanation shifts results in more sensitive indicators for varying model behavior.
- We release an open-source Python package `skshift`, which implements our "*Explanation Shift Detector*", along usage tutorials for reproducibility (cf. Statement 6).

## 2 Foundations and Related Work

### 2.1 Basic Notions

Supervised machine learning induces a function $f_\theta : \text{dom}(X) \to \text{dom}(Y)$, from training data $\mathcal{D}^{tr} = \{(x_0^{tr}, y_0^{tr}) \ldots, (x_n^{tr}, y_n^{tr})\}$. Thereby, $f_\theta$ is from a family of functions $f_\theta \in F$ and $\mathcal{D}^{tr}$ is sampled from the joint distribution $\mathbf{P}(X, Y)$ with predictor variables $X$ and target variable $Y$. $f_\theta$ is expected to generalize well on new, previously unseen data $\mathcal{D}_X^{new} = \{x_0^{new}, \ldots, x_k^{new}\} \subseteq \text{dom}(X)$. We write $\mathcal{D}_X^{tr}$ to refer to $\{x_0^{tr}, \ldots, x_n^{tr}\}$ and $\mathcal{D}_Y^{tr}$ to refer to $\mathcal{D}_Y^{tr} = \{y_0^{tr} \ldots, y_n^{tr}\}$. For the purpose of formalizations and to define evaluation metrics, it is often convenient to assume that an oracle provides values $\mathcal{D}_Y^{new} = \{y_0^{new}, \ldots, y_k^{new}\}$ such that $\mathcal{D}^{new} = \{(x_0^{new}, y_0^{new}), \ldots, (x_k^{new}, y_k^{new})\} \subseteq \text{dom}(X) \times \text{dom}(Y)$.

The core machine learning assumption is that training data $\mathcal{D}^{tr}$ and novel data $\mathcal{D}^{new}$ are sampled from the same underlying distribution $\mathbf{P}(X, Y)$. The twin problems of *model monitoring* and recognizing that new data is *out-of-distribution* can now be described as predicting an absolute or relative performance drop between perf($\mathcal{D}^{tr}$) and perf($\mathcal{D}^{new}$), where $\text{perf}(\mathcal{D}) = \sum_{(x,y)\in\mathcal{D}} \ell_{\text{eval}}(f_\theta(x), y)$, $\ell_{\text{eval}}$ is a metric like 0-1-loss (accuracy), but $\mathcal{D}_Y^{new}$ is unknown and cannot be used for such judgment in an operating system.

Therefore related work analyses distribution shifts between training and newly occurring data. Let two datasets $\mathcal{D}, \mathcal{D}'$ define two empirical distributions $\mathbf{P}(\mathcal{D}), \mathbf{P}(\mathcal{D}')$, then we write $\mathbf{P}(\mathcal{D}) \nsim \mathbf{P}(\mathcal{D}')$ to express that $\mathbf{P}(\mathcal{D})$ is sampled from a different underlying distribution than $\mathbf{P}(\mathcal{D}')$ with high probability $p > 1 - \epsilon$ allowing us to formalize various types of distribution shifts.

**Definition 2.1** (Input Data Shift)**.** We say that data shift occurs from $\mathcal{D}^{tr}$ to $\mathcal{D}_X^{new}$, if $\mathbf{P}(\mathcal{D}_X^{tr}) \nsim \mathbf{P}(\mathcal{D}_X^{new})$.

Specific kinds of data shift are:

**Definition 2.2** (Univariate data shift)**.** There is a univariate data shift between $\mathbf{P}(\mathcal{D}_X^{tr}) = \mathbf{P}(\mathcal{D}_{X_1}^{tr}, \ldots, \mathcal{D}_{X_p}^{tr})$ and $\mathbf{P}(\mathcal{D}_X^{new}) = \mathbf{P}(\mathcal{D}_{X_1}^{new}, \ldots, \mathcal{D}_{X_p}^{new})$, if $\exists i \in \{1 \ldots p\} : \mathbf{P}(\mathcal{D}_{X_i}^{tr}) \nsim \mathbf{P}(\mathcal{D}_{X_i}^{new})$.

**Definition 2.3** (Covariate data shift)**.** There is a covariate data shift between $P(\mathcal{D}_X^{tr}) = \mathbf{P}(\mathcal{D}_{X_1}^{tr}, \ldots, \mathcal{D}_{X_p}^{tr})$ and $\mathbf{P}(\mathcal{D}_X^{new}) = \mathbf{P}(\mathcal{D}_{X_1}^{new}, \ldots, \mathcal{D}_{X_p}^{new})$ if $\mathbf{P}(\mathcal{D}_X^{tr}) \nsim \mathbf{P}(\mathcal{D}_X^{new})$, which cannot only be caused by univariate shift.

The next two types of shift involve the interaction of data with the model $f_\theta$, which approximates the conditional $\frac{P(\mathcal{D}_Y^{tr})}{P(\mathcal{D}_X^{tr})}$. Abusing notation, we write $f_\theta(\mathcal{D})$ to refer to the multiset $\{f_\theta(x) | x \in \mathcal{D}\}$.

**Definition 2.4** (Predictions Shift)**.** There is a predictions shift between distributions $\mathbf{P}(\mathcal{D}_X^{tr})$ and $\mathbf{P}(\mathcal{D}_X^{new})$ related to model $f_\theta$ if $\mathbf{P}(f_\theta(\mathcal{D}_X^{tr})) \nsim \mathbf{P}(f_\theta(\mathcal{D}_X^{new}))$.

**Definition 2.5** (Concept Shift)**.** There is a concept shift between $\mathbf{P}(\mathcal{D}^{tr}) = \mathbf{P}(\mathcal{D}_X^{tr}, \mathcal{D}_Y^{tr})$ and $\mathbf{P}(\mathcal{D}^{new}) = \mathbf{P}(\mathcal{D}_X^{new}, \mathcal{D}_Y^{new})$ if conditional distributions change, i.e. $\frac{\mathbf{P}(\mathcal{D}_Y^{tr})}{\mathbf{P}(\mathcal{D}_X^{tr})} \nsim \frac{\mathbf{P}(\mathcal{D}_Y^{new})}{\mathbf{P}(\mathcal{D}_X^{new})}$.

In practice, multiple types of shifts co-occur together and their disentangling may constitute a significant challenge that we do not address here.

### 2.2 Related Work on Tabular Data

**Classifier two-sample test:** 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; Liu et al., 2020a) and has advanced in recent years (Lee et al., 2018; Zhang et al., 2013). The use of supervised learning classifiers to measure statistical tests has been explored by Lopez-Paz & Oquab (2017) proposing a classifier-based approach that returns test statistics to interpret differences between two distributions. We adopt their power test analysis and interpretability approach but apply it to the explanation distributions instead of input data distributions. Another noteworthy recent contribution comes from Barrabés et al. (2023), who leverages a tree-based classifier as C2ST. Their approach, augmented with iterative heuristics, aims at localizing and rectifying feature shifts within input data. In contrast, our work is distinctive in its purpose of investigating the impact of distribution shifts on the model behaviour. We achieve this by applying C2ST to distributions of feature attribution explanations, providing insights into how alterations in distribution impact the model's predictive behavior.

Other methods to detect if new data is OOD have relied on neural networks based on the prediction distributions Fort et al. (2021); Garg et al. (2020). They use the maximum softmax probabilities/likelihood as a confidence score Hendrycks & Gimpel (2017), temperature or energy-based scores Ren et al. (2019); Liu et al. (2020b); Wang et al. (2021), they extract information from the gradient space Huang et al. (2021), relying on the latent space Crabbé et al. (2021), they fit a Gaussian distribution to the embedding, or they use the Mahalanobis distance for out-of-distribution detection Lee et al. (2018); Park et al. (2021).

Many of these methods are explicitly developed for neural networks that operate on image and text data, and often, they can not be directly applied to traditional ML techniques. For image and text data, one may build on the assumption that the relationships between relevant predictor variables ($X$) and response variables ($Y$) remain unchanged, i.e., that no *concept shift* occurs. For instance, the essence of how a dog looks remains unchanged over different data sets, even if contexts may change. Thus, one can define invariances on the latent spaces of deep neural models, which do not apply to tabular data in a similar manner. For example, predicting buying behaviour before, during, and after the COVID-19 pandemic constitutes a conceptual shift that is not amenable to such methods. We focus on such tabular data where techniques such as gradient boosting decision trees achieve state-of-the-art model performance.

**Detecting distribution shift and its impact on model behaviour:**Extensive related work has aimed at detecting that data is from out-of-distribution. To this end, they have created several benchmarks that measure whether data comes from in-distribution or not (Malinin et al., 2021; Barrabés et al., 2023). In contrast, our main aim is to evaluate the impact of the distribution shift on the use of model. A typical example is two-sample testing on the latent space such as described by Rabanser et al. (2019). However, many methods developed for detecting out-of-distribution data are specific to neural networks processing image and text data and can not be applied to traditional machine learning techniques. These methods often assume that the relationships between predictor and response variables remain unchanged, i.e., no concept shift occurs. Our work is applied to tabular data where techniques such as gradient boosting decision trees achieve state-of-the-art model performance (Grinsztajn et al., 2022; Elsayed et al., 2021; Borisov et al., 2021).

**Impossibility of model monitoring:** Recent research findings have formalized the limitations of monitoring machine learning models in the absence of labelled data. Specifically (Garg et al., 2021; Chen et al., 2022) prove the impossibility of predicting model degradation or detecting out-of-distribution data with certainty (Fang et al., 2022; Zhang et al., 2021; Guerin et al., 2022). Although our approach does not overcome these limitations, it provides valuable insights for machine learning engineers to better understand changes in interactions between learned models and shifting data distributions.

**Model monitoring and distribution shift under specific assumptions:** Under specific types of assumptions, model monitoring and distribution shift become feasible tasks. One type of assumption often found in the literature is to leverage causal knowledge to identify the drivers of distribution changes (Budhathoki et al., 2021; Zhang et al., 2022; Schrouff et al., 2022). For example, Budhathoki et al. (2021) use graphical causal models and feature attributions based on Shapley values to detect changes in the distribution. Similarly, other works aim to detect specific distribution shifts, such as covariate or concept shifts. Our approach does not rely on additional information, such as a causal graph, labelled test data, or specific types of distribution shift. Still, by the nature of pure concept shifts, the model behaviour remains unaffected and new data need to come with labelled responses to be detected.

**Explainability and distribution shift:** Lundberg et al. (2020) applied Shapley values to identify possible bugs in the pipeline by visualizing univariate SHAP contributions. Following this line of work, Nigenda et al. (2022) compare the order of the feature importance using the NDCG between training and unseen data. We go beyond their work and formalize the multivariate explanation distributions on which we perform a two-sample classifier test to detect how distribution shift impacts interaction with the model. Furthermore, we provide a mathematical analysis of how the SHAP values contribute to detecting distribution shift. In Appendix A we provide a formal comparison against Nigenda et al. (2022). Recent work by Kulinski & Inouye (2023), introduced a framework for explaining distribution shifts using a transport map between a source and target distribution, our work does not only at change on the distribution shift, but on how do these changes impact the model.

### 2.3 Explainable AI: Local Feature Attributions

Attribution by Shapley values explains machine learning models by determining the relevance of features used by the model (Lundberg et al., 2020; Lundberg & 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). The Shapley value $\mathcal{S}_j$ for the $j$'th player is defined via a value function $\mathrm{val} : 2^N \to \mathbb{R}$ of players in $T$:

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

$$\text{where} \quad \mathrm{val}_{f,x}(T) = E_{X|X_T = x_T}[f(X)] - E_X[f(X)] \tag{2}$$

In machine learning, $N = \{1, \ldots, p\}$ is the set of features occurring in the training data. Given that $x$ is the feature vector of the instance to be explained, and the term $\mathrm{val}_{f,x}(T)$ represents the prediction for the feature values in $T$ that are marginalized over features that are not included in $T$. The Shapley value framework satisfies several theoretical properties (Molnar, 2019; Shapley, 1953; Winter, 2002; Aumann & Dreze, 1974). Our approach is based on the efficiency and uninformative properties:

**Efficiency Property.** Feature contributions add up to the difference of prediction from $x^\star$ and the expected value,$\sum_{j \in N} \mathcal{S}_j(f, x^\star) = f(x^\star) - E[f(X)])$

**Uninformativeness Property.** A feature $j$ that does not change the predicted value has a Shapley value of zero.$\forall x, x_j, x_j' : f(\{x_{N\setminus\{j\}}, x_j\}) = f(\{x_{N\setminus\{j\}}, x_j'\}) \Rightarrow \forall x : \mathcal{S}_j(f, x) = 0$.

Our approach works with explanation techniques that fulfill efficiency and uninformative properties, and we use Shapley values as an example. It is essential to distinguish between the theoretical Shapley values and the different implementations that approximate them, in Appendix E we provide an experimental comparison of different aproaches.

LIME is another explanation method candidate for our approach (Ribeiro et al., 2016b;a) that can potentially satisfy efficiency and uninformative properties, even thought several research has highlighted unstability and difficulties with the definition of neighborhoods. In Appendix D, we analyze LIME's relationship with Shapley values for the purpose of describing explanation shifts.

## 3 A Model for Explanation Shift Detection

Our model for explanation shift detection is sketched in Fig. 1. We define it as follows:

**Definition 3.1** (Explanation distribution)**.** An explanation function $\mathcal{S} : F \times \mathrm{dom}(X) \to \mathbb{R}^p$ maps a model $f_\theta \in F$ and data $x \in \mathbb{R}^p$ to a vector of attributions $\mathcal{S}(f_\theta, x) \in \mathbb{R}^p$. We call $\mathcal{S}(f_\theta, x)$ an explanation. We write $\mathcal{S}(f_\theta, \mathcal{D})$ to refer to the empirical *explanation distribution* generated by $\{\mathcal{S}(f_\theta, x) | x \in \mathcal{D}\}$.

We use local feature attribution methods SHAP and LIME as explanation functions $\mathcal{S}$.

**Definition 3.2** (Explanation shift)**.** Given a model $f_\theta$ learned from $\mathcal{D}^{tr}$, explanation shift with respect to the model $f_\theta$ occurs if $\mathcal{S}(f_\theta, \mathcal{D}_X^{new}) \nsim \mathcal{S}(f_\theta, \mathcal{D}_X^{tr})$.

**Definition 3.3** (Explanation shift metrics)**.** Given a measure of statistical distances $d$, explanation shift is measured as the distance between two explanations of the model $f_\theta$ by $d(\mathcal{S}(f_\theta, \mathcal{D}_X^{tr}), \mathcal{S}(f_\theta, \mathcal{D}_X^{new}))$.

We follow Lopez et al. (Lopez-Paz & Oquab, 2017) to define an explanation shift metrics based on a two-sample test classifier. We proceed as depicted in Figure 1. To counter overfitting, given the model $f_\theta$ trained on $\mathcal{D}^{\text{tr}}$, we compute explanations $\{\mathcal{S}(f_\theta, x) | x \in \mathcal{D}_X^{\text{val}}\}$ on an in-distribution validation data set $\mathcal{D}_X^{\text{val}}$. Given a dataset $\mathcal{D}_X^{new}$, for which the status of in- or out-of-distribution is unknown, we compute its explanations $\{\mathcal{S}(f_\theta, x) | x \in \mathcal{D}_X^{\text{new}}\}$. Then, we construct a two-samples dataset $E = \{(S(f_\theta, x), a_x) | x \in \mathcal{D}_X^{\text{val}}, a_x = 0\} \cup \{(S(f_\theta, x), a_x) | x \in \mathcal{D}_X^{\text{new}}, a_x = 1\}$ and we train a discrimination model $g_\psi : R^p \to \{0, 1\}$ on $E$, to predict if an explanation should be classified as in-distribution (ID) or out-of-distribution (OOD):

$$\psi = \arg\min_{\tilde{\psi}} \sum_{x \in \mathcal{D}_X^{\text{val}} \cup \mathcal{D}_X^{\text{new}}} \ell(g_{\tilde{\psi}}(\mathcal{S}(f_\theta, x)), a_x), \tag{3}$$

where $\ell$ is a classification loss function (e.g. cross-entropy). $g_\psi$ is our two-sample test classifier, based on which AUC yields a test statistic that measures the distance between the $D_X^{tr}$ explanations and the explanations of new data $D_X^{new}$.

Explanation shift detection allows us to detect *that* a novel dataset $D^{new}$ changes the model's behavior. Beyond recognizing explanation shift, using feature attributions for the model $g_\psi$, we can interpret *how* the features of the novel dataset $D_X^{new}$ interact differently with model $f_\theta$ than the features of the validation dataset $D_X^{val}$. These features are to be considered for model monitoring and for classifying new data as out-of-distribution.

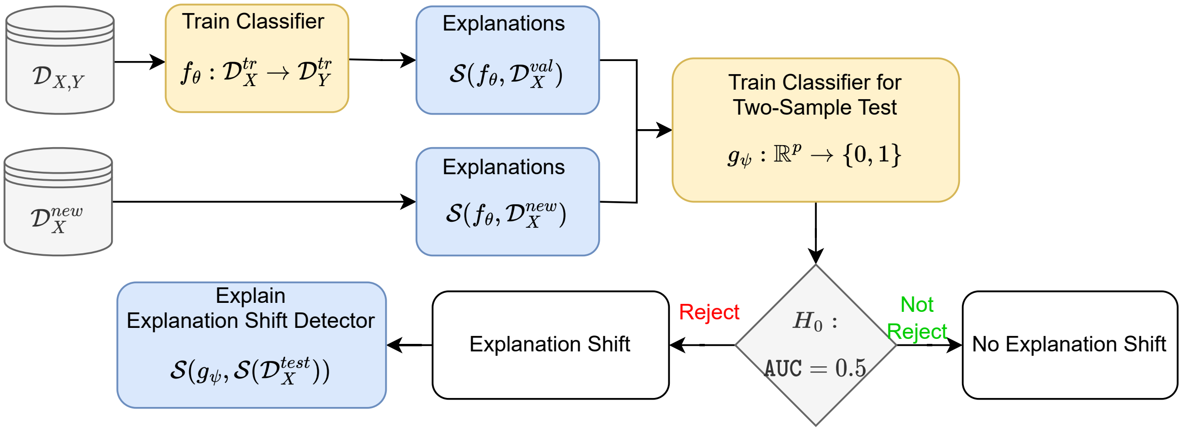

$\mathcal{D}_{X,Y}$
Train Classifier
$f_\theta : \mathcal{D}_X^{tr} \to \mathcal{D}_Y^{tr}$
Explanations
$\mathcal{S}(f_\theta, \mathcal{D}_X^{val})$
Train Classifier for Two-Sample Test
$g_\psi : \mathbb{R}^p \to \{0, 1\}$
$\mathcal{D}_X^{new}$
Explanations
$\mathcal{S}(f_\theta, \mathcal{D}_X^{new})$
Explain Explanation Shift Detector
$\mathcal{S}(g_\psi, \mathcal{S}(\mathcal{D}_X^{test}))$
Explanation Shift
Reject
$H_0$ :
$\text{AUC} = 0.5$
Not Reject
No Explanation Shift

**Figure 1:** Our model for explanation shift detection. The model $f_\theta$ is trained on $\mathcal{D}^{tr}$ implying explanations for distributions $\mathcal{D}_X^{val}, \mathcal{D}_X^{new}$. The AUC of the two-sample test classifier $g_\psi$ decides for or against explanation shift. If an explanation shift occurred, it could be explained which features of the $\mathcal{D}_X^{new}$ deviated in $f_\theta$ compared to $\mathcal{D}_X^{val}$.

# 4 Relationships between Common Distribution Shifts and Explanation Shifts

This section analyses and compares data shifts and prediction shifts with explanation shifts. Section 4.4 draws from these analyses to derive experiments with synthetic data.

## 4.1 Explanation Shift vs Data Shift

One type of distribution shift that is challenging to detect comprises cases where the univariate distributions for each feature $j$ are equal between the source $\mathcal{D}_X^{tr}$ and the unseen dataset $\mathcal{D}_X^{new}$, but where interdependencies among different features change. Multi-covariance statistical testing is a hard task with high sensitivity that can lead to false positives. The following example demonstrates that Shapley values can indicate co-variate interaction changes while a univariate statistical test will provide false negatives.

**Example 4.1.** *(**Multivariate Shift**) Let* $D^{tr} \sim N\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma^2_{X_1} & 0\\ 0 & \sigma^2_{X_2}\end{bmatrix}\right) \times Y$*. We fit a linear model* $f_\theta(x_1, x_2) = \gamma + a \cdot x_1 + b \cdot x_2$*. If* $\mathcal{D}^{new}_X \sim N\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma^2_{X_1} & \rho\sigma_{X_1}\sigma_{X_2}\\ \rho\sigma_{X_1}\sigma_{X_2} & \sigma^2_{X_2}\end{bmatrix}\right)$*, then* $\mathbf{P}(\mathcal{D}^{tr}_{X_1})$ *and* $\mathbf{P}(\mathcal{D}^{tr}_{X_2})$ *are identically distributed with* $\mathbf{P}(\mathcal{D}^{new}_{X_1})$ *and* $\mathbf{P}(\mathcal{D}^{new}_{X_2})$*, respectively, while this does not hold for the corresponding* $\mathcal{S}_j(f_\theta, \mathcal{D}^{tr}_X)$ *and* $\mathcal{S}_j(f_\theta, \mathcal{D}^{new}_X)$ *for* $f \in \{1, 2\}$*.*

$$\mathcal{S}_1(f_\theta, x) = a(x_1 - \mu_1) \tag{4}$$

$$\mathcal{S}_1(f_\theta, x^{new}) = \tag{5}$$

$$= \frac{1}{2}[\mathrm{val}(\{1,2\}) - \mathrm{val}(\{2\})] + \frac{1}{2}[\mathrm{val}(\{1\}) - \mathrm{val}(\emptyset)] \tag{6}$$

$$\mathrm{val}(\{1,2\}) = E[f_\theta | X_1 = x_1, X_2 = x_2] = ax_1 + bx_2 \tag{7}$$

$$\mathrm{val}(\emptyset) = E[f_\theta] = a\mu_1 + b\mu_2 \tag{8}$$

$$\mathrm{val}(\{1\}) = E[f_\theta(x) | X_1 = x_1] + b\mu_2 \tag{9}$$

$$\mathrm{val}(\{1\}) = \mu_1 + \rho\frac{\rho_{x_1}}{\sigma_{x_2}}(x_1 - \sigma_1) + b\mu_2 \tag{10}$$

$$\mathrm{val}(\{2\}) = \mu_2 + \rho\frac{\sigma_{x_2}}{\sigma_{x_1}}(x_2 - \mu_2) + a\mu_1 \tag{11}$$

$$\Rightarrow \mathcal{S}_1(f_\theta, x^{new}) \neq a(x_1 - \mu_1) \tag{12}$$

False positives frequently occur in out-of-distribution data detection when a statistical test recognizes differences between a source distribution and a new distribution, though the differences do not affect the model behavior (Grinsztajn et al., 2022; Huyen, 2022). Shapley values satisfy the *Uninformativeness* property, where a feature $j$ that does not change the predicted value has a Shapley value of 0 (equation 2.3).

**Example 4.2.** ***Shifts on Uninformative Features.*** *Let the random variables* $X_1, X_2$ *be normally distributed with* $N(0; 1)$*. Let dataset* $\mathcal{D}^{tr} \sim X_1 \times X_2 \times Y^{tr}$*, with* $Y^{tr} = X_1$*. Thus* $Y^{tr} \perp X_2$*. Let* $\mathcal{D}^{new}_X \sim X_1 \times X^{new}_2$ *and* $X^{new}_2$ *be normally distributed with* $N(\mu; \sigma^2)$ *and* $\mu, \sigma \in \mathbb{R}$*. When* $f_\theta$ *is trained optimally on* $\mathcal{D}^{tr}$ *then* $f_\theta(x) = x_1$*.* $\mathbf{P}(\mathcal{D}^{tr}_{X_2})$ *is different from* $\mathbf{P}(\mathcal{D}^{new}_{X_2})$ *but* $\mathcal{S}_2(f_\theta, \mathcal{D}^{tr}_X) = 0 = \mathcal{S}_2(f_\theta, \mathcal{D}^{new}_X)$*.*

$$\mathcal{D}_{X_3} \sim N(\mu_3, c_3), \mathcal{D}^{new}_{X_3} \sim N(\mu_3^{'}, c_3^{'}) \tag{13}$$

$$\text{If} \quad \mu_3^{'} \neq \mu_3 \quad \text{or} \quad c_3^{'} \neq c_3 \rightarrow P(X_3) \neq P(X_3^{new}) \tag{14}$$

$$\mathcal{S}(f_\theta, X) = \left(\begin{bmatrix} a_1(X_1 - \mu_1)\\ a_2(X_2 - \mu_2)\\ a_3(X_3 - \mu_3)\end{bmatrix}\right) = \left(\begin{bmatrix} a_1(X_1 - \mu_1)\\ a_2(X_2 - \mu_2)\\ 0\end{bmatrix}\right) \tag{15}$$

$$\mathcal{S}_3(f_\theta, \mathcal{D}_X) = \mathcal{S}_3(f_\theta, \mathcal{D}^{new}_X) \tag{16}$$

### 4.2 Explanation Shift vs Prediction Shift

Analyses of the explanations detect distribution shifts that interact with the model. In particular, if a prediction shift occurs, the explanations produced are also shifted.

**Proposition 1.** Given a model $f_\theta : \mathcal{D}_X \to \mathcal{D}_Y$. If $f_\theta(x') \neq f_\theta(x)$, then $\mathcal{S}(f_\theta, x') \neq \mathcal{S}(f_\theta, x)$.

$$\texttt{Given} \quad f_\theta(x) \neq f_\theta(x') \tag{17}$$

$$\sum_{j=1}^{p} \mathcal{S}_j(f_\theta, x) = f_\theta(x) - E_X[f_\theta(\mathcal{D}_X)] \tag{18}$$

$$\texttt{then} \quad \mathcal{S}(f,x) \ \neq \mathcal{S}(f,x') \tag{19}$$

By efficiency property of the Shapley values (Aas et al., 2021) (equation (2.3)), if the prediction between two instances is different, then they differ in at least one component of their explanation vectors.

The opposite direction does not always hold. Thus, an explanation shift does not always imply a prediction shift.

**Example 4.3.** ***(Explanation shift not affecting prediction distribution)*** *Given $\mathcal{D}^{tr}$ is generated from $(X_1 \times X_2 \times Y), X_1 \sim U(0,1), X_2 \sim U(1,2), Y = X_1 + X_2 + \epsilon$ and thus the optimal model is $f(x) = x_1 + x_2$. If $\mathcal{D}^{new}$ is generated from $X_1^{new} \sim U(1,2), X_2^{new} \sim U(0,1), \quad Y^{new} = X_1^{new} + X_2^{new} + \epsilon$, the prediction distributions are identical $f_\theta(\mathcal{D}_X^{tr}), f_\theta(\mathcal{D}_X^{new}) \sim U(1,3)$, but explanation distributions are different $S(f_\theta, \mathcal{D}_X^{tr}) \nsim S(f_\theta, \mathcal{D}_X^{new})$, because $\mathcal{S}_i(f_\theta, x) = \alpha_i \cdot x_i$.* for $i \in \{1,2\}$

$$\forall i \in \{1,2\} \quad \mathcal{S}_i(f_\theta, x) = \alpha_i \cdot x_i \tag{20}$$

$$\forall i \in \{1,2\} \Rightarrow \mathcal{S}_i(f_\theta, \mathcal{D}_X)) \neq \mathcal{S}_i(f_\theta, \mathcal{D}_X^{new}) \tag{21}$$

$$\Rightarrow f_\theta(\mathcal{D}_X) = f_\theta(\mathcal{D}_X^{new}) \tag{22}$$

### 4.3 Explanation Shift vs Concept Shift

Concept shift comprises cases where the covariates retain a given distribution, but their relationship with the target variable changes (cf. Section 2.1). This example shows the negative result that concept shift cannot be indicated by the detection of explanation shift.

**Example 4.4.** ***Concept Shift*** *Let $\mathcal{D}^{tr} \sim X_1 \times X_2 \times Y$, and create a synthetic target $y_i^{tr} = a_0 + a_1 \cdot x_{i,1} + a_2 \cdot x_{i,2} + \epsilon$. As new data we have $\mathcal{D}_X^{new} \sim X_1^{new} \times X_2^{new} \times Y$, with $y_i^{new} = b_0 + b_1 \cdot x_{i,1} + b_2 \cdot x_{i,2} + \epsilon$ whose coefficients are unknown at prediction stage. With coefficients $a_0 \neq b_0, a_1 \neq b_1, a_2 \neq b_2$. We train a linear regression $f_\theta : \mathcal{D}_X^{tr} \to \mathcal{D}_Y^{tr}$. Then explanations have the same distribution, $\mathbf{P}(\mathcal{S}(f_\theta, \mathcal{D}_X^{tr})) = \mathbf{P}(\mathcal{S}(f_\theta, \mathcal{D}_X^{new}))$, input data distribution $\mathbf{P}(\mathcal{D}_X^{tr}) = \mathbf{P}(\mathcal{D}_X^{new})$ and predictions $\mathbf{P}(f_\theta(\mathcal{D}_X^{tr})) = \mathbf{P}(f_\theta(\mathcal{D}_X^{new}))$. But there is no guarantee on the performance of $f_\theta$ on $\mathcal{D}_X^{new}$ (Garg et al., 2021)*

$$X \sim N(\mu, \sigma^2 \cdot I), X^{new} \sim N(\mu, \sigma^2 \cdot I) \tag{23}$$

$$\to P(\mathcal{D}_X) = P(\mathcal{D}_X^{new}) \tag{24}$$

$$Y \sim a + \alpha N(\mu, \sigma^2) + \beta N(\mu, \sigma^2) + N(0, \sigma^{'2}) \tag{25}$$

$$Y^{new} \sim a + \beta N(\mu, \sigma^2) + \alpha N(\mu, \sigma^2) + N(0, \sigma^{'2}) \tag{26}$$

$$\to P(\mathcal{D}_Y) = P(\mathcal{D}_Y^{new}) \tag{27}$$

$$\mathcal{S}(f_\theta, \mathcal{D}_X) = \begin{pmatrix} \alpha(X_1 - \mu_1) \\ \beta(X_2 - \mu_2) \end{pmatrix} \sim \begin{pmatrix} N(\mu_1, \alpha^2\sigma^2) \\ N(\mu_2, \beta^2\sigma^2) \end{pmatrix} \tag{28}$$

$$\mathcal{S}(h_\phi, \mathcal{D}_X) = \begin{pmatrix} \beta(X_1 - \mu_1) \\ \alpha(X_2 - \mu_2) \end{pmatrix} \sim \begin{pmatrix} N(\mu_1, \beta^2\sigma^2) \\ N(\mu_2, \alpha^2\sigma^2) \end{pmatrix} \tag{29}$$

$$\text{If} \quad \alpha \neq \beta \to \mathcal{S}(f_\theta, \mathcal{D}_X) \neq \mathcal{S}(h_\phi, \mathcal{D}_X) \tag{30}$$

In general, concept shift cannot be detected because $\mathcal{D}_Y^{new}$ is unknown (Garg et al., 2021). Some research studies have made specific assumptions about the conditional $\frac{P(\mathcal{D}_Y^{new})}{P(\mathcal{D}_X^{new})}$ in order to monitor models and detect distribution shift (Lu et al., 2023; Alvarez et al., 2023). Appendix 4.3 sketches a situation where explanation distributions are used in the context of labelled data to indicate concept shift— which however is not the main target of the paper.

### 4.4 Experiments on Synthetic Data

This experimental section explores the detection of distribution shifts via explanation shifts in the previous analytical examples. In this case, the model is non-linear, a gradient boosting decision tree from the `xgboost` Python implementation.

#### 4.4.1 Detecting Multivariate Shift

Given two bivariate normal distributions $\mathcal{D}_X = (X_1, X_2) \sim N\left(0, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$ and $\mathcal{D}_X^{new} = (X_1^{new}, X_2^{new}) \sim N\left(0, \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}\right)$, then, for each feature $j$ the underlying distribution is equally distributed between $\mathcal{D}_X$ and $\mathcal{D}_X^{new}$, $\forall j \in \{1,2\} : P(\mathcal{D}_{X_j}) = P(\mathcal{D}_{X_j}^{new})$, 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 $f_\theta(\mathcal{D}_X)$. Then we compute the SHAP explanation values for $\mathcal{S}(f_\theta, \mathcal{D}_X)$ and $\mathcal{S}(f_\theta, \mathcal{D}_X^{new})$

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

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

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

#### 4.4.2 Detecting Concept Shift

As mentioned before, concept shifts cannot be detected if new data comes without target labels. However, if new data is labelled, the explanation shift can still be a useful technique for detecting concept shifts.

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

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

| Comparison | **Conclusions** |
|---|---|
| $\mathbf{P}(\mathcal{D}_X)$, $\mathbf{P}(\mathcal{D}_X^{new})$ | Not Distinct |
| $\mathbf{P}(\mathcal{D}_Y)$, $\mathbf{P}(\mathcal{D}_Y^{new})$ | Not Distinct |
| $\mathbf{P}(f_\theta(\mathcal{D}_X))$, $\mathbf{P}(h_\Upsilon(\mathcal{D}_X^{new}))$ | Not Distinct |
| $\mathbf{P}(\mathcal{S}(f_\theta, \mathcal{D}_X))$, $\mathbf{P}(\mathcal{S}(h_\Upsilon, \mathcal{D}_X))$ | 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 synthetic example, in table 2 SHAP values can detect a relational change between $\mathcal{D}_X$ and $\mathcal{D}_Y$, even if both distributions remain equivalent.

#### 4.4.3 Uninformative Features on Synthetic Data

To have an applied use case of the synthetic example from the methodology section, we create a three-variate normal distribution $\mathcal{D}_X = (X_1, X_2, X_3) \sim N(0, I_3)$, 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 $\mathcal{D}^{new}_{X_3} = \mathcal{D}^{te}_{X_3} + 1$.

**Table 3:** Distribution comparison when modifying a random noise variable on test data. The input data shifts while explanations and predictions do not.

| Comparison | **Conclusions** |
|---|---|
| $\mathbf{P}(\mathcal{D}^{te}_{X_3})$, $\mathbf{P}(\mathcal{D}^{new}_{X_3})$ | Distinct |
| $f_\theta(\mathcal{D}^{te}_X)$, $f_\theta(\mathcal{D}^{new}_X)$ | Not Distinct |
| $\mathcal{S}(f_\theta, \mathcal{D}^{te}_X)$, $\mathcal{S}(f_\theta, \mathcal{D}^{new}_X)$ | Not Distinct |

In Table 3, we see how an unused feature has changed the input distribution, but the explanation distributions 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.

#### 4.4.4 Explanation Shift that does not Affect the Prediction

In this case, we provide a situation when we have changes in the input data distributions that affect the model explanations but do not affect the model predictions due to positive and negative associations between the model predictions and the distributions cancel out, producing a vanishing correlation in the mixture of the distribution (Yule's effect 4.2).

We create a train and test data by drawing $50,000$ samples from a bi-uniform distribution $X_1 \sim U(0, 1), \quad X_2 \sim U(1, 2)$ the target variable is generated by $Y = X_1 + X_2$ where we train our model $f_\theta$. Then if out-of-distribution data is sampled from $X^{new}_1 \sim U(1, 2)$, $X^{new}_2 \sim U(0, 1)$

**Table 4:** Distribution comparison over how the change on the contributions of each feature can cancel out to produce an equal prediction (cf. Section 4.2), while explanation shift will detect this behaviour changes on the predictions will not.

| Comparison | **Conclusions** |
|---|---|
| $f(\mathcal{D}^{te}_X)$, $f(\mathcal{D}^{new}_X)$ | Not Distinct |
| $\mathcal{S}(f_\theta, \mathcal{D}^{te}_{X_2})$, $\mathcal{S}(f_\theta, \mathcal{D}^{new}_{X_2})$ | Distinct |
| $\mathcal{S}(f_\theta, \mathcal{D}^{te}_{X_1})$, $\mathcal{S}(f_\theta, \mathcal{D}^{new}_{X_1})$ | Distinct |

In Table 4, we see how an unused feature has changed the input distribution, but the explanation distributions 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.

### 4.5 Summary Comparison on Synthetic data

To assess the effectiveness of different detection methods in identifying and accounting for synthetic shifts, we present a conceptual comparison in Table 5. We evaluate these methods based on their capacity to capture synthetic shifts. We illustrate this comparison by considering two scenarios: a multicovariate shift (cf.

Example 4.1) and a shift involving uninformative features (cf. Example 4.2). The complementary evaluation with related work is in the Appendix in sectionA

This comparison focuses on their ability to detect synthetic distribution shifs using the examples of covariate shift and uninformative shifts, and provides valuable insights while ensuring accountability.

**Table 5:** Conceptual comparison of different detection methods over the examples discussed in the mathematical analsyis of the main body of the paper(cf. Section 4): a multicovariate shift(cf. Example 4.1 )and a uninformative features shift(cf. Example 4.2) . Learning a Classifier Two-Sample test $g$ over the explanation distributions is the only method that achieves the desired results (✓) and is accountable. We evaluate accountability by checking if the feature attributions of the detection method correspond to the synthetic shift generated in both scenarios

| **Detection Method** | **Covariate** | **Uninformative** | **Accountability** |
|---|---|---|---|
| Input distribution($g_\phi$) | ✓ | ✗ | ✗ |
| Prediction distribution($g_\Upsilon$) | ✓ | ✓ | ✗ |
| Input KS | ✗ | ✗ | ✗ |
| Classifier Drift | ✓ | ✗ | ✗ |
| Output KS | ✓ | ✓ | ✗ |
| Output Wasserstein | ✓ | ✓ | ✗ |
| Uncertainty | ∼ | ✓ | ✓ |
| NDCG | ✗ | ✓ | ✗ |
| Explanation distribution ($g_\psi$)<br>Explanation Shift Detector | ✓ | ✓ | ✓ |

## 5 Empirical Evaluation

We evaluate the effectiveness of explanation shift detection on tabular data by comparing it against methods from the literature, which are all based on discovering distribution shifts. For this comparison, we systematically vary models $f$, model parametrizations $\theta$, and input data distributions $\mathcal{D}_X$. The experiments results described include:t *(i)* add details on experiments with synthetic data (Section 5.2), *(ii)* add experiments on natural datasets (Sections 5.3, 5.4 and Appendix B), *(iii)* exhibit a larger range of modelling choices (Section 5.5), *(iv)* study the effects of hypeparameter variations on explanations shifts (Section 5.6) ,and *(v)* contrast our SHAP-based method against the use of LIME, an alternative explanation approach (Appendix D). Core observations made in this section will only be confirmed and refined but not countered in the appendix.

### 5.1 Baseline Methods and Datasets

**Baseline Methods.** We compare our method of explanation shift detection (Section 3) with several methods that aim to detect that input data is out-of-distribution: *(B1)* statistical Kolmogorov Smirnov test on input data (Rabanser et al., 2019), *(B2)* prediction shift detection by Wasserstein distance (Lu et al., 2023), *(B3)* NDCG-based test of feature importance between the two distributions (Nigenda et al., 2022), *(B4)* prediction shift detection by Kolmogorov-Smirnov test (Diethe et al., 2019), and *(B5)* model agnostic uncertainty estimation (Mougan & Nielsen, 2023; Kim et al., 2020). All Distribution Shift Metrics are scaled between 0 and 1. We also compare against Classifier Two-Sample Test (Lopez-Paz & Oquab, 2017) on different distributions as discussed in Section 4, viz. *(B6)* classifier two-sample test on input distributions ($g_\phi$) following Barrabés et al. (2023) and *(B7)* classifier two-sample test on the predictions distributions ($g_\Upsilon$):

$$\phi = \arg\min_{\tilde{\phi}} \sum_{x \in \mathcal{D}_X^{val} \cup \mathcal{D}_X^{new}} \ell(g_{\tilde{\phi}}(x)), a_x) \tag{31}$$

$$\Upsilon = \arg\min_{\tilde{\Upsilon}} \sum_{x \in \mathcal{D}_X^{val} \cup \mathcal{D}_X^{new}} \ell(g_{\tilde{\Upsilon}}(f_\theta(x)), a_x) \tag{32}$$

**Datasets.** In the main body of the paper we base our comparisons on the UCI Adult Income dataset Dua & Graff (2017) and on synthetic data. In the Appendix, we extend experiments to several other datasets, which confirm our findings: ACS Travel Time, ACS Employment, Stackoverflow dataset (Stackoverflow, 2019).

## 5.2 Synthetic Data

Our first experiment on synthetic data showcases the two main contributions of our method: $(i)$ being more sensitive to changes in the model than prediction shift and input shift and $(ii)$ accounting for its drivers. We first generate a synthetic dataset $\mathcal{D}^{\rho}$, with a parametrized multivariate shift between $(X_1, X_2)$, where $\rho$ is the correlation coefficient, and an extra variable $X_3 = N(0,1)$ and generate our target $Y = X_1 \cdot X_2 + X_3$. We train the $f_\theta$ on $\mathcal{D}^{tr,\rho=0}$ using a gradient boosting decision tree, while for $g_\psi : \mathcal{S}(f_\theta, \mathcal{D}_X^{val,\rho}) \to \{0,1\}$, we train on different datasests with different values of $\rho$. For $g_\psi$ we use a logistic regression. In Section 5.5, we benchmark other models $f_\theta$ and detectors $g_\psi$.

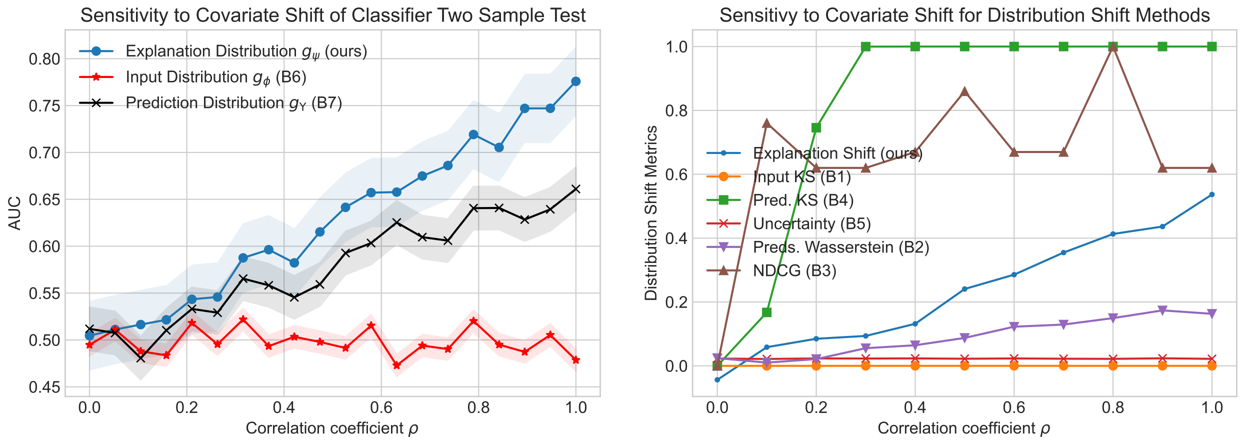


**Figure 2:** In the left figure, we compare the Classifier Two-Sample Test on explanation distribution (ours) versus input distribution *(B6)* and prediction distribution *(B7)*. Explanation distribution shows the highest sensitivity. The right figure, related work comparison of distribution shift methods*(B1-B5)*, as the experimental setup has a gradual distribution shift, good indicators should follow a progressive steady positive slope, following the correlation coefficient, as our method does. In Table 6 we provide a quantitative evaluation.

The left image in Figure 2 compares our approach against C2ST on input data distribution*(B6)* and on the predictions distribution *(B7)* different data distributions, for detecting multi-covariate shifts on different distributions. In our covariate experiment, we observed that using the explanation shift led to higher sensitivity towards detecting distribution shift. We interpret the results with the efficiency property of the Shapley values, which decomposes the vector $f_\theta(\mathcal{D}_X)$ into the matrix $\mathcal{S}(f_\theta, \mathcal{D}_X)$. Moreover, we can identify the features that cause the drift by extracting the coefficients of $g_\psi$, providing global and local explainability.

The right image features the same setup compared to the other out-of-distribution detection methods (B1-B5). Table 6 quantitatively evaluates how the baselines correlate with the covariate correlation coefficient ($\rho$). One can see how our method behaves favourably compared to the others.

**Table 6:** Pearson Correlation of the correlation coefficient $\rho$ and baseline methods, extending Figure 2. Explanation Shift achieves better covariate shift detection on synthetic data.

| **Baseline** | **Pearson Correlation with $\rho$** |
|---|---|
| B1 Input KS | 0.01 |
| B2 Prediction Wasserstein | 0.97 |
| B3 Explanation NDCG | 0.52 |
| B4 Prediction KS | 0.70 |
| B5 Uncertainty | 0.26 |
| B6 C2ST Input | 0.18 |
| B7 C2ST Output | 0.96 |
| (Ours) Explanation Shift | **0.99** |

### 5.3 Novel Group Shift

The distribution shift in this experimental setup is constituted by the appearance of a hitherto unseen group at prediction time (the group information is not present in the training features). We vary the ratio of presence of this unseen group in $\mathcal{D}_X^{new}$ data. The experiment is done with two $f_\theta$ models: a gradient-boosting decision tree and a logistic regression; for $g_\psi$, we use a logistic regression. Results are presented in Figure 3 and Table 7. Confidence intervals are extracted out of 10 bootstraps. Furthermore, we compare the performance of different algorithms for $f_\theta$ and $g_\psi$ in Section 5.5, and varying hyperparameters in Section 5.6.

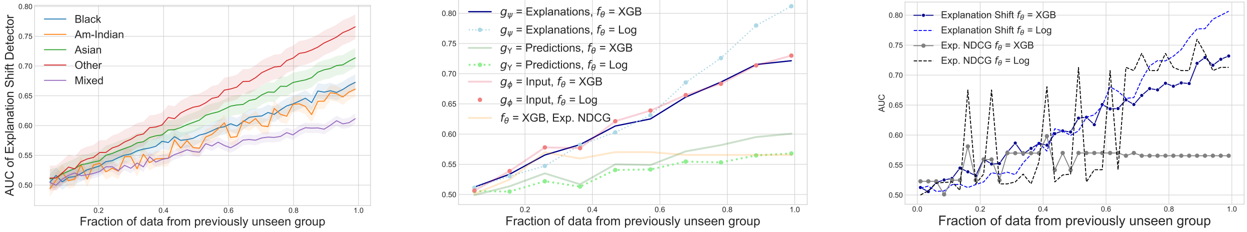


**Figure 3:** Novel group shift experiment on the US Income dataset. Sensitivity (AUC) increases with the growing fraction of previously unseen social groups. *Left figure*: The explanation shift indicates that different social groups exhibit varying deviations from the distribution on which the model was trained (White). *Middle Figure*: We vary the model $f_\theta$ by training it using both `XGBoost` (solid lines) and Logistic Regression (dots). The novel ethnicity group is Black. We compare Explanation Shift against C2ST on input *(B6)* and output *(B7)*. *Right figure*: Comparison of Explanation Shift against Exp. NDCG *(B4)*. We see how monitoring method*(B4)* is more unstable with a linear model, and with an `xgboost` it erroneously finds a horizontal asymptote. We don't compare against methods relying purely on input data such as *(B1)* as we are changing the model, which they don't take into consideration.

**Table 7:** Pearson correlation between baselines and the ratio of presence of the unseen group. The *Accountable* column indicates accountability, measured as providing both theoretical underpinnings and empirical validation of the sources driving model changes, as discussed in Section 4.

| **Baseline** | **Pearson Correlation** | | **Accountable** |
|---|---|---|---|
| | $f_\theta$ = **Log** | $f_\theta$ = **XGB** | |
| B1 Input KS | $\mathbf{0.99 \pm 0.01}$ | $\mathbf{0.99 \pm 0.01}$ | ✗ |
| B2 Pred. Wass. | $0.95 \pm 0.02$ | $\mathbf{0.98 \pm 0.01}$ | ✗ |
| B3 NDCG | $0.37 \pm 0.25$ | $0.81 \pm 0.10$ | ✗ |
| B4 Pred. KS | $0.97 \pm 0.02$ | $0.96 \pm 0.01$ | ✗ |
| B5 Uncertainty | $0.73 \pm 0.10$ | $0.74 \pm 0.12$ | ✓ |
| B6 C2ST Input | $0.95 \pm 0.03$ | $0.95 \pm 0.03$ | ✗ |
| B7 C2ST Output | $0.67 \pm 0.13$ | $0.96 \pm 0.02$ | ✗ |
| Explanation Shift | $\mathbf{0.98 \pm 0.01}$ | $\mathbf{0.98 \pm 0.01}$ | ✓ |

### 5.4 Geopolitical and Temporal Shift

In this section, we tackle a geopolitical and temporal distribution shift; for this, the training data $\mathcal{D}^{tr}$ for the model $f_\theta$ is composed of data from California in 2014 and a $\mathcal{D}^{new}$ for each of the states in 2018. The model $g_\psi$ is trained each time on each state using only the $\mathcal{D}_X^{new}$ in the absence of the label, and a 50/50 random train-test split evaluates its performance. As models, we use `xgboost` as $f_\theta$ and logistic regression for the *Explanation Shift Detector* ($g_\psi$).

We hypothesize that the AUC of the Explanation Shift Detector on new data will be distinct from that on in-distribution data, primarily owing to the distinctive nature of out-of-distribution model explanations. Figure 4 illustrates the performance of our method on different data distributions, where the baseline is a ID hold-out set of CA14. The AUC for CA18, where there is only a temporal shift, is the closest to the baseline, and the OOD detection performance is better in the rest of the states. The most disparate state is Puerto Rico (PR18).

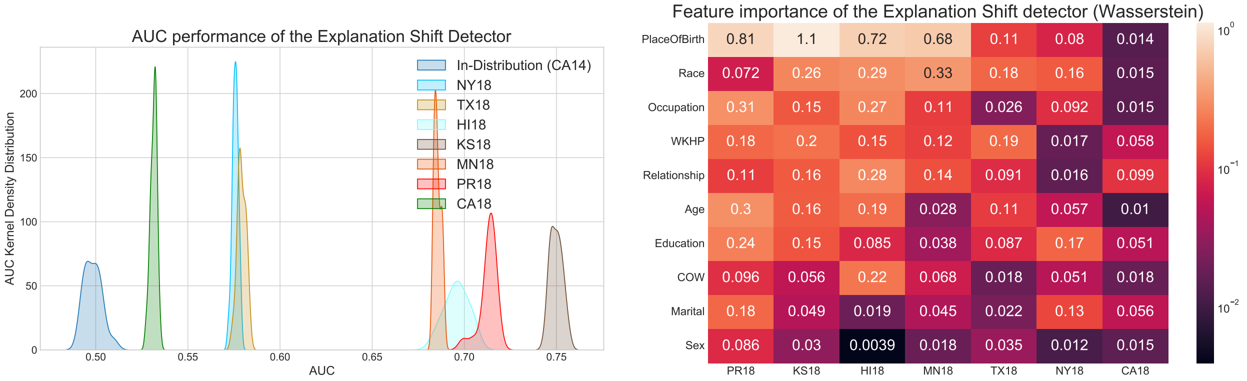


**Figure 4:** In the left figure, a comparison of the performance of *Explanation Shift Detector* in different states. In the right figure, the strength analysis of features driving the change in the model, on the y-axis are the features, and on the x-axis are the different states. Explanation shifts allow us to identify how the distribution shift of different features impacted the model.

Our next objective is to identify the features where the explanations differ between $\mathcal{D}_X^{tr}$ and $\mathcal{D}_X^{new}$ data. To achieve this, we compare the distribution of linear coefficients of the detector between both distributions. We use the Wasserstein distance as a distance measure, generating 1000 in-distribution bootstraps using a 63.2% sampling fraction from California-14 and 1000 bootstraps from other states in 2018. In the right image of Figure 4, we observe that for PR18, the most crucial feature is the Place of Birth.

Furthermore, we conduct an across-task evaluation by comparing the performance of the “Explanation Shift Detector” on another prediction task in the Appendix B. Although some features are present in both prediction tasks, the weights and importance order assigned by the "Explanation Shift Detector" differ. One of this method’s advantages is that it identifies differences in distributions and how they relate to the model.

### 5.5 Varying Models and Explanation Shift Detectors

OOD data detection methods based on input data distributions only depend on the detector type, independent of the model $f_\theta$. OOD Explanation methods rely on both the model and the data. Using explanation shifts as indicators for measuring distribution shifts’ impact on the model enables us to account for the influencing factors of the explanation shift. Therefore, in this section, we compare the performance of different algorithms for explanation shift detection using the same experimental setup. The results of our experiments show that using Explanation Shift enables us to see differences in the choice of the original model $f_\theta$ and the Explanation Shift Detector $g_\phi$

| | Estimator $f_\theta$ | | | | | | |
|---|---|---|---|---|---|---|---|
| Detector $g_\phi$ | **XGB** | **Log.Reg** | **Lasso** | **Ridge** | **Rand.Forest** | **Dec.Tree** | **MLP** |
| **XGB** | 0.583 | 0.619 | 0.596 | 0.586 | 0.558 | 0.522 | 0.597 |
| **LogisticReg.** | 0.605 | 0.609 | 0.583 | 0.625 | 0.578 | 0.551 | 0.605 |
| **Lasso** | 0.599 | 0.572 | 0.551 | 0.595 | 0.557 | 0.541 | 0.596 |
| **Ridge** | 0.606 | 0.61 | 0.588 | 0.624 | 0.564 | 0.549 | 0.616 |
| **RandomForest** | 0.586 | 0.607 | 0.574 | 0.612 | 0.566 | 0.537 | 0.611 |
| **DecisionTree** | 0.546 | 0.56 | 0.559 | 0.569 | 0.543 | 0.52 | 0.569 |

**Table 8:** Comparison of explanation shift detection performance, measured by AUC, for different combinations of explanation shift detectors and estimators on the UCI Adult Income dataset using the Novel Covariate Group Shift for the “Asian” group with a fraction ratio of 0.5 (cf. Section 5). The table shows that the algorithmic choice for $f_\theta$ and $g_\psi$ can impact the OOD explanation performance. We can see how, for the same detector, different $f_\theta$ models flag different OOD explanations performance. On the other side, for the same $f_\theta$ model, different detectors achieve different results.

### 5.6 Hyperparameters Sensitivity Evaluation

This section presents an extension to our experimental setup where we vary the model complexity by varying the model hyperparameters $\mathcal{S}(f_\theta, X)$. We use the UCI Adult Income dataset with the Novel Covariate Group Shift for the "Asian" group with a fraction ratio of 0.5 as described in Section 5. And for the Stackoverflow as training data we use the United States of America and a novel covariate group France.

In this experiment, we changed the hyperparameters of the original model: for the decision tree, we varied the depth of the tree, while for the gradient-boosting decision, we changed the number of estimators, and for the random forest, both hyperparameters. We calculated the Shapley values using TreeExplainer (Lundberg et al., 2020). For the Detector choice of model, we compare Logistic Regression and XGBoost models.

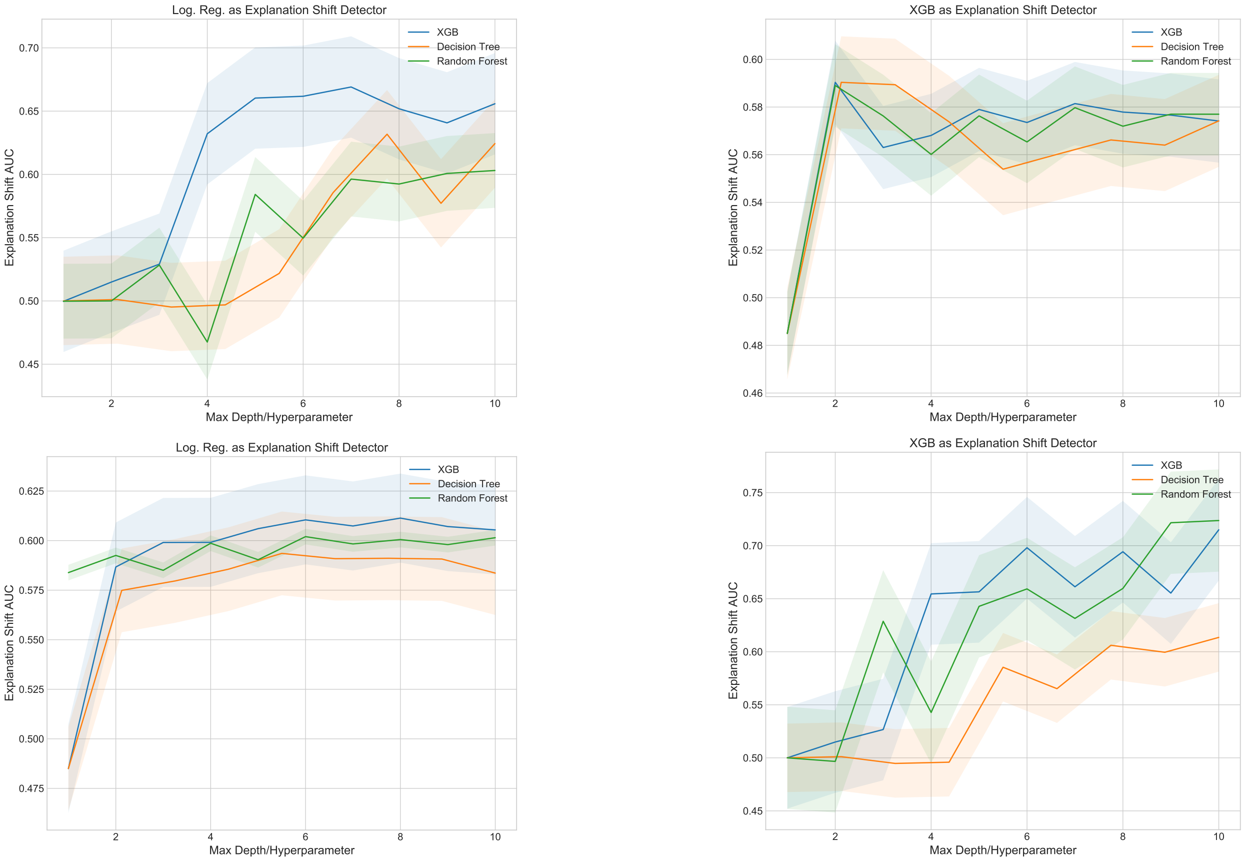


**Figure 5:** Images represent the AUC of the *Explanation Shift Detector*, on two datasets: Top ACS Income and Bottom Stackoverflow under novel group shift. In the images on the left, the detector is a logistic regression, and in the images on the right, it is a gradient-boosting decision tree classifier. By changing the model, we can see that vanilla models (decision tree with depth 1 or 2) are unaffected by the distribution shift, while when increasing the model complexity, the out-of-distribution impact of the data in the model starts to be tangible

The results presented in Figure 5 show the AUC of the *Explanation Shift Detector* for the ACS Income dataset under novel group shift. We observe that the distribution shift does not affect very simplistic models, such as decision trees with depths 1 or 2. However, as we increase the model complexity, the impact of out-of-distribution data on the model becomes more pronounced. Furthermore, when we compare the performance of the *Explanation Shift Detector* across different models, such as Logistic Regression and Gradient Boosting Decision Tree, we observe distinct differences(note that the y-axis takes different values).

In conclusion, the explanation distributions serve as a projection of the data and model sensitive to what the model has learned. The results demonstrate the importance of considering model complexity under distribution shifts.

### 5.7 Discussion

The Shapley value, a key component in our method, describes how a model's prediction for a specific data point deviates from the mean. These theoretical considerations, which we laid out in Section 4, have been confirmed by our experimental sections.

In Section› 5.3, we have studied input distribution shift. Our experiment shows that explanation shift detects input distribution shifts better than the best baseline methods. Table 7 showcases the two top-performing methods—comparing input distributions with Kolmogorov-Smirnoff *(B1)* and our method — with statistically insignificant differences.

In Section 5.2, we have studied co-variate shift. Considering, Table 6, the best method for detecting input distribution shifts, *(B1)*, fails completely on this task. The second best method is *(B2)* comparison of prediction distributions using the Wasserstein distance, which also did quite well wrt. predicting input distribution shifts and came rather close behind our approach in both experiments.

Moreover, in our geopolitical and temporal shift experiment (Section 5.4), we demonstrate the ability to account for the drivers of model changes under such input data shifts. Cross-task comparisons in experiments (Figure 9 or Figure 8) highlight how explanation shift feature importance varies even when input distribution shifts remain constant during cross-task. These capabilities are not offered by any of the competing baselines. These observations are further supported by additional experiments in Section 5.6, where we solely vary model complexity, showcasing the adaptability of explanation shifts to changes in model characteristics.

## 6 Conclusions

Commonly, the problem of detecting the impact of the distribution shift on the model has relied on measurements for detecting shifts in the input or output data distributions or relied on assumptions either on the type of distribution shift or causal graphs availability. In this paper, we proposed explanation shifts as an indicator for detecting and identifying the impact of distribution shifts on machine learning models. We provide software, mathematical analysis examples, synthetic data, and real-data experimental evaluation. We found that measures of explanation shift can provide more insights than input distribution and prediction shift measures when monitoring machine learning models.

**Limitations:** The potential utility of explanation shifts as distribution shift indicators that affect the model in computer vision or natural language processing tasks remains an open question. We have used feature attribution explanations to derive indications of explanation shifts, but other AI explanation techniques may be applicable and come with their advantages. Also, our approach cannot detect concept shifts, as concept shift requires understanding the interaction between input data and response variables. By the nature of pure concept shifts, such changes do not affect the model. We work under the assumption that such labels are not available for new data, nor do we make other assumptions; therefore, our method is not able to predict the degradation of prediction performance under distribution shifts.

Furthermore, our use of the `shap` Python package for Shapley values approximation can introduce known drawbacks, as highlighted in recent literature (Bilodeau et al., 2022; Slack et al., 2020a). Additionally, our current implementation relies on linear Shapley value interaction approximations, which can be extended following the work of Fumagalli et al. (2023); Bordt & von Luxburg (2023).

### Reproducibility Statement

To ensure reproducibility, we make the data, code repositories, and experiments publicly available `https://anonymous.4open.science/r/ExplanationShift-812A`. Also, an open-source Python package `skshift` `https://anonymous.4open.science/r/skshift-65A5` is released. For our experiments, we used default `scikit-learn` parameters Pedregosa et al. (2011).Experiments were run on a 4 vCPU server with 32 GB RAM.

## References

Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artif. Intell.*, 298:103502, 2021. doi: 10.1016/j.artint.2021.103502. URL https://doi.org/10.1016/j.artint.2021.103502.

Jose M. Alvarez, Kristen M. Scott, Salvatore Ruggieri, and Bettina Berendt. Domain adaptive decision trees: Implications for accuracy and fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2023.

Robert J Aumann and Jacques H Dreze. Cooperative games with coalition structures. *International Journal of game theory*, 3(4):217–237, 1974.

Míriam Barrabés, Daniel Mas Montserrat, Margarita Geleta, Xavier Giró i Nieto, and Alexander G Ioannidis. Adversarial learning for feature shift detection and correction. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In Yee Whye Teh and D. Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, pp. 129–136. JMLR.org, 2010. URL http://proceedings.mlr.press/v9/david10a.html.

Blair L. Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *CoRR*, abs/2212.11870, 2022.

Sebastian Bordt and Ulrike von Luxburg. From shapley values to generalized additive models and back. In *AISTATS*, volume 206 of *Proceedings of Machine Learning Research*, pp. 709–745. PMLR, 2023.

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey, 2021. URL https://arxiv.org/abs/2110.01889.

Kailash Budhathoki, Dominik Janzing, Patrick Blöbaum, and Hoiyi Ng. Why did the distribution change? In Arindam Banerjee and Kenji Fukumizu (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1666–1674. PMLR, 2021. URL http://proceedings.mlr.press/v130/budhathoki21a.html.

Hugh Chen, Joseph D. Janizek, Scott M. Lundberg, and Su-In Lee. True to the model or true to the data? *CoRR*, abs/2006.16234, 2020. URL https://arxiv.org/abs/2006.16234.

Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. L-shapley and c-shapley: Efficient model interpretation for structured data. In *ICLR 2019*, 2019.

Lingjiao Chen, Matei Zaharia, and James Y. Zou. Estimating and explaining model performance when both covariates and labels shift. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *NeurIPS*, 2022.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785.

Jonathan Crabbé, Zhaozhi Qian, Fergus Imrie, and Mihaela van der Schaar. Explaining latent representations with a corpus of examples. In *NeurIPS*, pp. 12154–12166, 2021.

Tom Diethe, Tom Borchert, Eno Thereska, Borja Balle, and Neil Lawrence. Continual learning in practice. ArXiv preprint, 2019.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *NeurIPS*, pp. 6478–6490, 2021a.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *NeurIPS 2021*, pp. 6478–6490, 2021b.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Shereen Elsayed, Daniela Thyssens, Ahmed Rashed, Lars Schmidt-Thieme, and Hadi Samer Jomaa. Do we really need deep learning models for time series forecasting? *CoRR*, abs/2101.02118, 2021. URL `https://arxiv.org/abs/2101.02118`.

Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *NeurIPS*, 2022.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *NeurIPS*, 34, 2021. URL `https://arxiv.org/abs/2106.03004`.

Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *NeurIPS 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/0d770c496aa3da6d2c3f2bd19e7b9d6b-Abstract.html`.

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Eva Hammer. SHAP-IQ: Unified approximation of any-order shapley interactions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. A unified view of label shift estimation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *NeurIPS*, volume 33, pp. 3290–3300. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/219e052492f4008818b8adb6366c7ed6-Paper.pdf`.

Saurabh Garg, Sivaraman Balakrishnan, Zachary Chase Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

Amirata Ghorbani and James Y. Zou. Data shapley: Equitable valuation of data for machine learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2242–2251. PMLR, 2019. URL `http://proceedings.mlr.press/v97/ghorbani19c.html`.

Leo Grinsztajn, Edouard Oyallon, and Gael Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *NeurIPS*, 2022.

Joris Guerin, Kevin Delmas, Raul Sena Ferreira, and Jérémie Guiochet. Out-of-distribution detection is not all you need. In *NeurIPS ML Safety Workshop*, 2022.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR 2017*, 2017.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *NeurIPS 2021*, abs/2110.00218, 2021. URL `https://arxiv.org/abs/2110.00218`.

Chip Huyen. *Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications*. O'Reilly, 2022. URL `ISBN-13:978-1098107963`.

Byol Kim, Chen Xu, and Rina Foygel Barber. Predictive inference is free with the jackknife+-after-bootstrap. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *NeurIPS 2020*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/2b346a0aa375a07f5a90a344a61416c4-Abstract.html`.

Sean Kulinski and David I. Inouye. Towards explaining distribution shifts. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17931–17952. PMLR, 2023.

Yongchan Kwon, Manuel A. Rivas, and James Zou. Efficient computation and analysis of distributional shapley values. In Arindam Banerjee and Kenji Fukumizu (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 793–801. PMLR, 2021. URL `http://proceedings.mlr.press/v130/kwon21a.html`.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *NeurIPS 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 7167–7177, 2018. URL `https://proceedings.neurips.cc/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html`.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3128–3136. PMLR, 2018. URL `http://proceedings.mlr.press/v80/lipton18a.html`.

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J. Sutherland. Learning deep kernels for non-parametric two-sample tests. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6316–6326. PMLR, 2020a. URL `http://proceedings.mlr.press/v119/liu20m.html`.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *NeurIPS 2020*, 2020b. URL `https://proceedings.neurips.cc/paper/2020/hash/f5496252609c43eb8a3d147ab9b9c006-Abstract.html`.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *ICLR 2017*, 2017.

Yuzhe Lu, Zhenlin Wang, Runtian Zhai, Soheil Kolouri, Joseph Campbell, and Katia P. Sycara. Predicting out-of-distribution error with confidence optimal transport. In *ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML*, 2023.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *NeurIPS 2017*, pp. 4765–4774, 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html`.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839, 2020.

Andrey Malinin, Neil Band, Yarin Gal, Mark J. F. Gales, Alexander Ganshin, German Chesnokov, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, Vyas Raina, Denis Roginskiy, Mariya Shmatova, Panagiotis Tigas, and Boris Yangel. Shifts: A dataset of real distributional shift across multiple large-scale tasks. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/ad61ab143223efbc24c7d2583be69251-Abstract-round2.html.

Christoph Molnar. *Interpretable Machine Learning.* ., 2019. https://christophm.github.io/interpretable-ml-book/.

Carlos Mougan and Dan Saattrup Nielsen. Monitoring model deterioration with explainable uncertainty estimation via non-parametric bootstrap. In *AAAI*, pp. 15037–15045. AAAI Press, 2023.

David Nigenda, Zohar Karnin, Muhammad Bilal Zafar, Raghu Ramesha, Alan Tan, Michele Donini, and Krishnaram Kenthapadi. Amazon sagemaker model monitor: A system for real-time insights into deployed machine learning models. In *KDD*, pp. 3671–3681. ACM, 2022.

Chunjong Park, Anas Awadalla, Tadayoshi Kohno, and Shwetak N. Patel. Reliable and trustworthy machine learning for health using dataset shift detection. *NeurIPS 2021*, abs/2110.14019, 2021. URL https://arxiv.org/abs/2110.14019.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Liudmila Ostroumova Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *NeurIPS 2018*, pp. 6639–6649, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer. *Dataset shift in machine learning.* Mit Press, 2009.

Stephan Rabanser, Stephan Günnemann, and Zachary C. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *NeurIPS 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 1394–1406, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/846c260d715e5b854ffad5f70a516c88-Abstract.html.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *NeurIPS 2019*, pp. 14680–14691, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1e79596878b2320cac26dd792a6c51c9-Abstract.html.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning, 2016a.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144. ACM, 2016b. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Jessica Schrouff, Natalie Harris, Oluwasanmi O Koyejo, Ibrahim Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alexander Brown, Subhrajit Roy, Diana Mincu, Chrsitina Chen, Awa Dieng, Yuan Liu, Vivek Natarajan, Alan Karthikesalingam, Katherine A Heller, Silvia Chiappa, and Alexander D'Amour. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *NeurIPS*, 2022.

L. S. Shapley. *A Value for n-Person Games*, pp. 307–318. Princeton University Press, 1953. doi: doi: 10.1515/9781400881970-018. URL `https://doi.org/10.1515/9781400881970-018`.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pp. 180–186, New York, NY, USA, 2020a. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375830. URL `https://doi.org/10.1145/3375627.3375830`.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In Annette N. Markham, Julia Powles, Toby Walsh, and Anne L. Washington (eds.), *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*, pp. 180–186. ACM, 2020b. doi: 10.1145/3375627.3375830. URL `https://doi.org/10.1145/3375627.3375830`.

Stackoverflow. Developer survey results 2019, 2019. URL `https://insights.stackoverflow.com/survey/2019/`.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9269–9278. PMLR, 2020.

Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks know what they don't know? In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *NeurIPS 2021*, pp. 29074–29087, 2021. URL `https://proceedings.neurips.cc/paper/2021/hash/f3b7e5d3eb074cde5b76e26bc0fb5776-Abstract.html`.

Eyal Winter. Chapter 53 the shapley value. In ., volume 3 of *Handbook of Game Theory with Economic Applications*, pp. 2025–2054. Elsevier, 2002. doi: https://doi.org/10.1016/S1574-0005(02)03016-3. URL `https://www.sciencedirect.com/science/article/pii/S1574000502030163`.

Artjom Zern, Klaus Broelemann, and Gjergji Kasneci. Interventional shap values and interaction values for piecewise linear regression trees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Haoran Zhang, Harvineet Singh, and Shalmali Joshi. "why did the model fail?": Attributing model performance changes to distribution shifts. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 819–827. JMLR.org, 2013. URL `http://proceedings.mlr.press/v28/zhang13d.html`.

Lily H. Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12427–12436. PMLR, 2021. URL `http://proceedings.mlr.press/v139/zhang21g.html`.

# A Experimental Comparison against Specific Related Work

## A.1 Comparison Against Changes on Feature Attribution Relevance

In this section, we present a comparative analysis against the work of (Nigenda et al., 2022),

which involves assessing the disparity in feature importance orders between training data and out-of-distribution data. To quantify this disparity, we employ the normalized discount cumulative gain (NDCG) metric. This method is versatile, accommodating both individual sample analysis and distribution-level assessments. In cases involving distributions, we aggregate the average feature importance.

### A.1.1 Novel Group Shift

**Experimental Set-Up:**This experiment extends the core experiment detailed in Section 5, where distribution shifts arise due to the emergence of previously unseen groups during the prediction phase.

**Datasets:** We use ACS Income, ASC Employment, ACS Mobility and ACS Travel time (Ding et al., 2021b). The group that is not present on the features is the *black* ethnicity.

**Baseline:** We compare against the method proposed by Nigenda et al. (2022), *(B6)* of the experimental comparison of the main body, that compares the order of the feature importance using the NDCG between train and unseen data. We vary $f_\theta$ to be a `xgboost` and a Logistic regression. For the "Explanation Shift Detector", $g_\psi$ , we use a logistic regression in both

**Metrics:**To facilitate a direct comparison with the Area Under the Curve (AUC) metric, we adapt the NDCG metric, to have the same interval range as follows: $(1 - NDCG) + 0.5$, ensuring a consistent metric range.

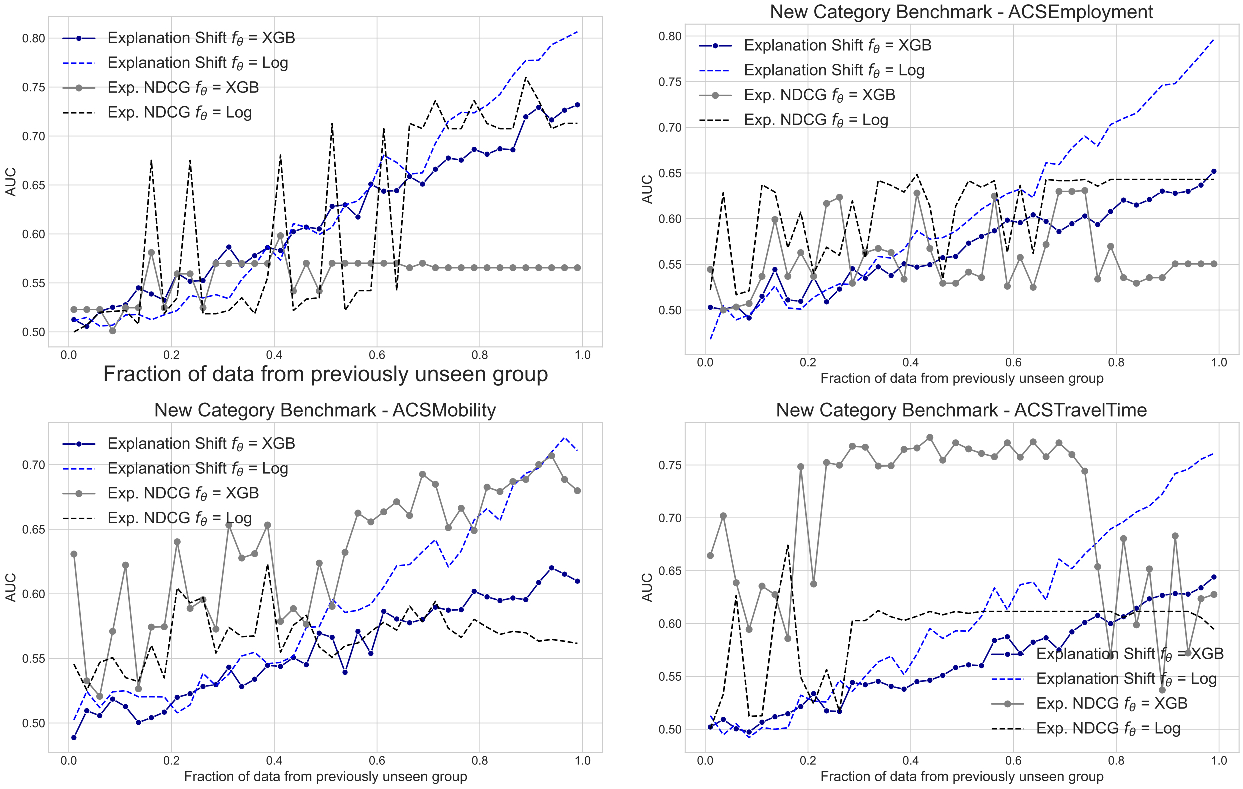


**Figure 6:** Novel group shift experiment conducted on the 4 Datasets. Sensitivity (AUC) increases as the proportion of previously unseen social groups grows. As the experimental setup has a gradual distribution shift, ideal indicators should exhibit a steadily increasing slope. However, in all figures, NDCG exhibits saturation and instability. These observations align with the analysis presented in the synthetic experiment section, as discussed in Section 5.2 of the main paper

This extended experiment aims to further validate the effectiveness of the "Explanation Shift Detector" under novel group shifts in real-world datasets. It demonstrates how the approach performs consistently across multiple datasets and provides insights into the sensitivity of model behavior as previously unseen social groups become a larger part of the prediction data. The results are presented in Figure 6, where our proposed method is compared against Exp. NDCG *(B6)* across the four datasets. We can see how Exp. NDCG *(B6)* is more unstable and finds often an horizontal asymptot, in all the situations, this is due to changes on the feature importance order do not have information about the value, where our approach of performing a Classifier Two Sample Test on the distributions of explanations do.

### A.1.2 Synthetic Data Comparison

In this section, we evaluate changes in the distribution of explanations and the order of feature importance when faced with a synthetic data shift scenario. We begin with a bivariate normal distribution $\mathcal{D}_X^{tr} = (X_1, X_2) \sim N(1, I)$, where $I$ represents the identity matrix of order two. We create a synthetic target variable $Y = X_1^2 \cdot X_2 + \epsilon$, and develop a machine learning model $f_\theta : \mathcal{D}_X \to \mathcal{D}_Y$ using a non-linear model, specifically an `xgboost` model. Subsequently, we generate new data from $\mathcal{D}_X^{new} = (X_1, X_2) \sim N(2, I)$, which constitutes a shift of $D_X^{new} = D^{tr}X + 1$. We then compute SHAP values for $\mathcal{S}(f\theta, \mathcal{D}_X)$ and compare the average contributions' orders.

Having sampled $50,000$ instances from both $\mathcal{D}_X^{tr}$ and $\mathcal{D}_X^{new}$, we analyze whether alterations in explanation distributions and explanation importance orders can detect these changes. To achieve this, we compare one-tailed p-values from the Kolmogorov-Smirnov test for explanation shifts and the order of average SHAP values between the distributions.

**Table 9:** Comparison between distribution shifts in explanations and shifts in feature attribution importance orders(previous work of (Nigenda et al., 2022)). Explanation distributions exhibit differences, while the importance order remains consistent

| Comparison | **Conclusions** |
|---|---|
| $\mathbf{P}(\mathcal{D}_X^{te})$, $\mathbf{P}(\mathcal{D}_X^{new})$ | Distinct |
| $\mathbf{P}(\mathcal{S}(f_\theta, \mathcal{D}_X^{te}))$, $\mathbf{P}(\mathcal{S}(f_\theta, \mathcal{D}_X^{new}))$ | Distinct |
| $\mathbf{P}(\mathcal{S}_1(f_\theta, \mathcal{D}_X^{te}) > \mathcal{S}_2(f_\theta, \mathcal{D}_X^{te}))$, $\mathbf{P}(\mathcal{S}_1(f_\theta, \mathcal{D}_X^{new}) > \mathcal{S}_2(f_\theta, \mathcal{D}_X^{new}))$ | Not Distinct |

### A.1.3 Analytical Comparison under Monotonous Uniform Shift

In this section, we conduct an analytical comparison between changes in explanation distributions and changes in the order of feature importance.

**Example A.1.** ***Comparison against NDCG*** Let $\mathcal{D}_X^{tr} = (\mathcal{D}_{X_1}^{tr}, \mathcal{D}_{X_2}^{tr}) \sim N([\mu_1, \mu_1], I)$ and $\mathcal{D}_X^{new} = (\mathcal{D}_{X_1}^{new}, \mathcal{D}_{X_2}^{new}) \sim N([\mu_2, \mu_2], I)$ where the relationship between $\mu_1$ and $\mu_2$ is monotonous uniform shift characterized by $\mu_2 = \mu_1 + N$ where N is a real number. We fit a linear model $f_\theta(X_1, X_2) = \gamma + a_1 \cdot X_1 + a_2 \cdot X_2$, where $a_1 > a_2$. Then even if the distribution of SHAP values are distinct between $\mathcal{S}(f_\theta, \mathcal{D}_X^{tr})$ and $\mathcal{S}(f_\theta, \mathcal{D}_X^{new})$, the order of importance between the distributions is not distinct. If $\mathcal{S}_1(f_\theta, \mathcal{D}_X^{tr}) > \mathcal{S}_2(f_\theta, \mathcal{D}_X^{tr})$ then $\mathcal{S}_1(f_\theta, \mathcal{D}_X^{new}) > \mathcal{S}_2(f_\theta, \mathcal{D}_X^{new})$. But the distributions are distinct $\mathcal{S}_1(f_\theta, \mathcal{D}_X^{tr}) \neq \mathcal{S}_1(f_\theta, \mathcal{D}_X^{new})$ and $\mathcal{S}_2(f_\theta, \mathcal{D}_X^{tr}) \neq \mathcal{S}_2(f_\theta, \mathcal{D}_X^{new})$

$$\mathcal{S}_j(f_\theta, \mathcal{D}_X) = a_j \cdot (\mathcal{D}_{X_j} - \mu_1), \mathcal{S}_j(f_\theta, \mathcal{D}_X^{new}) = a_j \cdot (\mathcal{D}_{X_j}^{new} - \mu_2) \tag{33}$$

$$\mu_2 = \mu_1 + N \tag{34}$$

$$\text{Then} \quad \mathcal{S}_j(f_\theta, \mathcal{D}_X) \neq \mathcal{S}_j(f_\theta, \mathcal{D}_X^{new}) \tag{35}$$

$$\text{But} \quad \mathcal{S}_1(f_\theta, \mathcal{D}_X) > \mathcal{S}_2(f_\theta, \mathcal{D}_X) \quad \Leftrightarrow \quad \mathcal{S}_1(f_\theta, \mathcal{D}_X^{new}) > \mathcal{S}_2(f_\theta, \mathcal{D}_X^{new}) \tag{36}$$

**Conclusion of the comparison to** Nigenda et al. (2022) In the context of natural data, when confronted with a novel covariate shift, our findings indicate that NDCG demonstrates limited sensitivity and fails to detect shifts when the fraction of data from previously unseen groups exceeds ratios 0.2 to 0.4 threshold.

Furthermore, in our analyses both synthetic and natural data, we observe that NDCG struggles to provide accurate and consistent estimates when faced with multicovariate shifts.

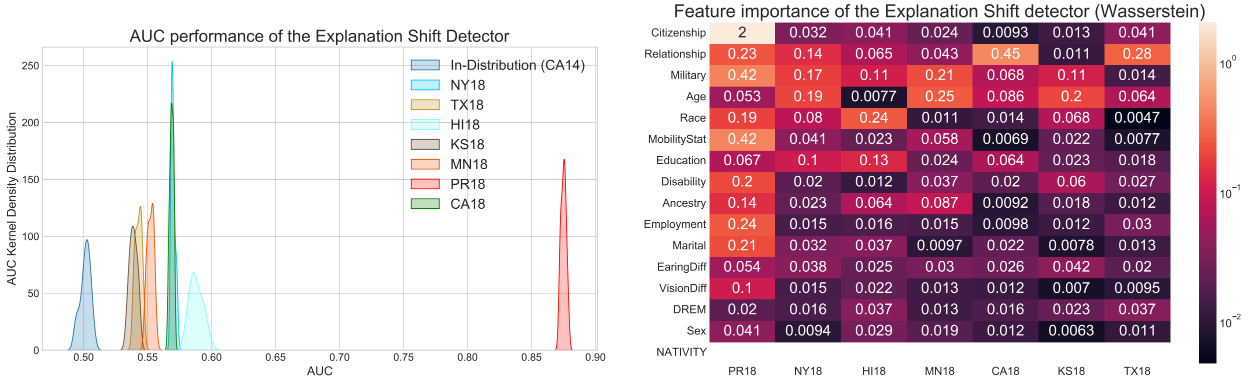
Both analytically and in our experiments with synthetic data, it becomes evident that NDCG lacks robustness and sensitivity when confronted with even a basic, uniform, and monotonous shift.

# B Further Experiments on Real Data

In this section, we extend the prediction task of the main body of the paper. The methodology used follows the same structure. We start by creating a distribution shift by training the model $f_\theta$ in California in 2014 and evaluating it in the rest of the states in 2018, creating a geopolitical and temporal shift. The model $g_\theta$ is trained each time on each state using only the $X^{New}$ in the absence of the label, and its performance is evaluated by a 50/50 random train-test split. As models, we use a gradient boosting decision tree(Chen & Guestrin, 2016; Prokhorenkova et al., 2018) for $f_\theta$, approximating the Shapley values by TreeExplainer (Lundberg et al., 2020), and using logistic regression for the *Explanation Shift Detector*.

## B.1 ACS Employment

The objective of this task is to determine whether an individual aged between 16 and 90 years is employed or not. The model's performance was evaluated using the AUC metric in different states, except PR18, where the model showed an explanation shift. The explanation shift was observed to be influenced by features such as Citizenship and Military Service. The performance of the model was found to be consistent across most of the states, with an AUC below 0.60. The impact of features such as difficulties in hearing or seeing was negligible in the distribution shift impact on the model. The left figure in Figure 7 compares the performance of the Explanation Shift Detector in different states for the ACS Employment dataset.



| | PR18 | NY18 | HI18 | MN18 | CA18 | KS18 | TX18 |
|---|---|---|---|---|---|---|---|
| Citizenship | 2 | 0.032 | 0.041 | 0.024 | 0.0093 | 0.013 | 0.041 |
| Relationship | 0.23 | 0.14 | 0.065 | 0.043 | 0.45 | 0.011 | 0.28 |
| Military | 0.42 | 0.17 | 0.11 | 0.21 | 0.068 | 0.11 | 0.014 |
| Age | 0.053 | 0.19 | 0.0077 | 0.25 | 0.086 | 0.2 | 0.064 |
| Race | 0.19 | 0.08 | 0.24 | 0.011 | 0.014 | 0.068 | 0.0047 |
| MobilityStat | 0.42 | 0.041 | 0.023 | 0.058 | 0.0069 | 0.022 | 0.0077 |
| Education | 0.067 | 0.1 | 0.13 | 0.024 | 0.064 | 0.023 | 0.018 |
| Disability | 0.2 | 0.02 | 0.012 | 0.037 | 0.02 | 0.06 | 0.027 |
| Ancestry | 0.14 | 0.023 | 0.064 | 0.087 | 0.0092 | 0.018 | 0.012 |
| Employment | 0.24 | 0.015 | 0.016 | 0.015 | 0.0098 | 0.012 | 0.03 |
| Marital | 0.21 | 0.032 | 0.037 | 0.0097 | 0.022 | 0.0078 | 0.013 |
| EaringDiff | 0.054 | 0.038 | 0.025 | 0.03 | 0.026 | 0.042 | 0.02 |
| VisionDiff | 0.1 | 0.015 | 0.022 | 0.013 | 0.012 | 0.007 | 0.0095 |
| DREM | 0.02 | 0.016 | 0.037 | 0.013 | 0.016 | 0.023 | 0.037 |
| Sex | 0.041 | 0.0094 | 0.029 | 0.019 | 0.012 | 0.0063 | 0.011 |
| NATIVITY | | | | | | | |

**Figure 7:** The left figure shows a comparison of the performance of the Explanation Shift Detector in different states for the ACS Employment dataset. The right figure shows the feature importance analysis for the same dataset.

Additionally, the feature importance analysis for the same dataset is presented in the right figure in Figure 7.

## B.2 ACS Travel Time

The goal of this task is to predict whether an individual has a commute to work that is longer than +20 minutes. For this prediction task, the results are different from the previous two cases; the state with the highest OOD score is $KS18$, with the "Explanation Shift Detector" highlighting features as Place of Birth, Race or Working Hours Per Week. The closest state to ID is CA18, where there is only a temporal shift without any geospatial distribution shift.

## B.3 ACS Mobility

The objective of this task is to predict whether an individual between the ages of 18 and 35 had the same residential address as a year ago. This filtering is intended to increase the difficulty of the prediction task, as the base rate for staying at the same address is above 90% for the population (Ding et al., 2021b).

The experiment shows a similar pattern to the ACS Income prediction task (cf. Section 4), where the inland US states have an AUC range of $0.55 - 0.70$, while the state of PR18 achieves a higher AUC. For PR18, the model has shifted due to features such as Citizenship, while for the other states, it is Ancestry (Census record

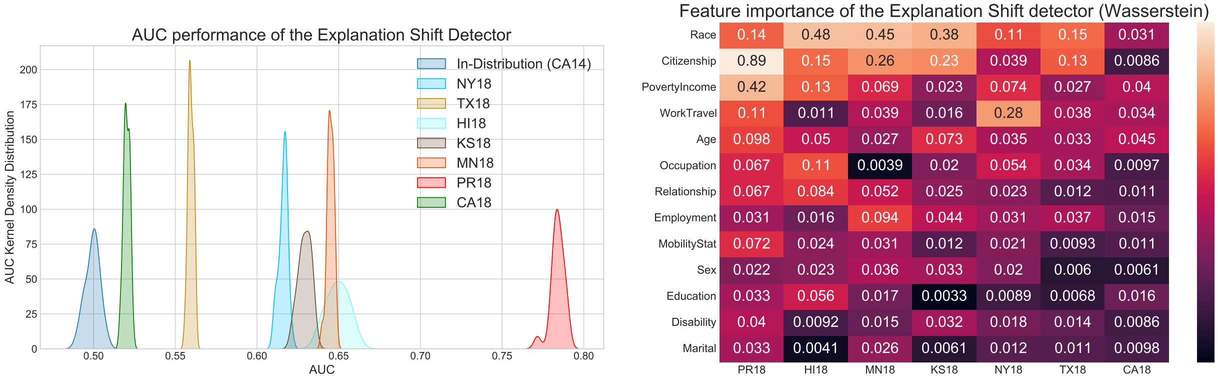


**Figure 8:** In the left figure, comparison of the performance of *Explanation Shift Detector*, in different states for the ACS TravelTime prediction task. In the left figure, we can see how the state with the highest OOD AUC detection is KS18 and not PR18 as in other prediction tasks; this difference with respect to the other prediction task can be attributed to "Place of Birth", whose feature attributions the model finds to be more different than in CA14.

of your ancestors' lives with details like where they lived, who they lived with, and what they did for a living) that drives the change in the model.

As depicted in Figure 9, all states, except for PR18, fall below an AUC of explanation shift detection of 0.70. Protected social attributes, such as Race or Marital status, play an essential role for these states, whereas for PR18, Citizenship is a key feature driving the impact of distribution shift in model.

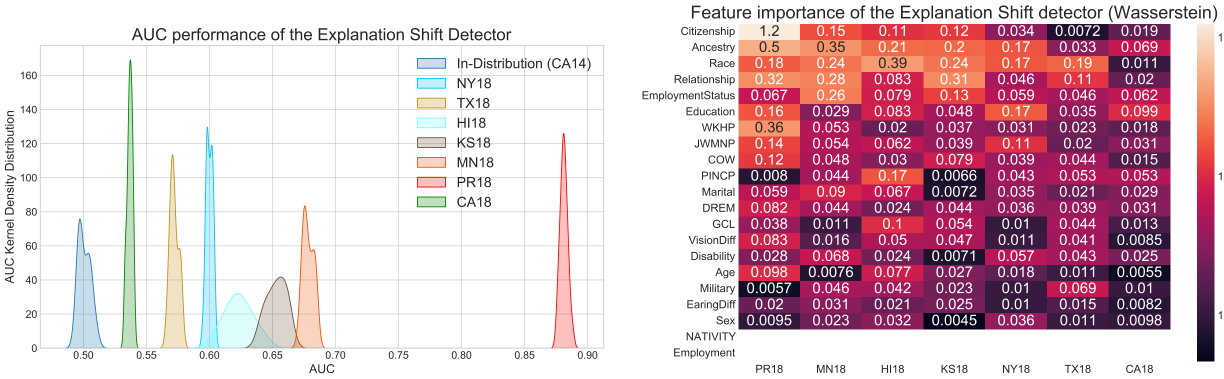


**Figure 9:** Left figure shows a comparison of the *Explanation Shift Detector*'s performance in different states for the ACS Mobility dataset. Except for PR18, all other states fall below an AUC of explanation shift detection of 0.70. The features driving this difference are Citizenship and Ancestry relationships. For the other states, protected social attributes, such as Race or Marital status, play an important role.

### B.4 StackOverflow Survey Data: Novel Covariate Group

This experimental section evaluates the proposed Explanation Shift Detector approach on real-world data under novel group distribution shifts. In this scenario, a new unseen group appears at the prediction stage, and the ratio of the presence of this unseen group in the new data is varied. As a training data country, we use the United States. The model $f_\theta$used is a gradient-boosting decision tree or logistic regression, and logistic regression is used for the detector. The results show that the AUC of the Explanation Shift Detector varies depending on the quantification of OOD explanations, and it shows more sensitivity concerning model variations than other state-of-the-art techniques.

The dataset used is the StackOverflow annual developer survey, with over 70,000 responses from over 180 countries examining aspects of the developer experience (Stackoverflow, 2019). The data has high dimensionality, leaving it with +100 features after data cleansing and feature engineering. The goal of this task is to predict the total annual compensation.

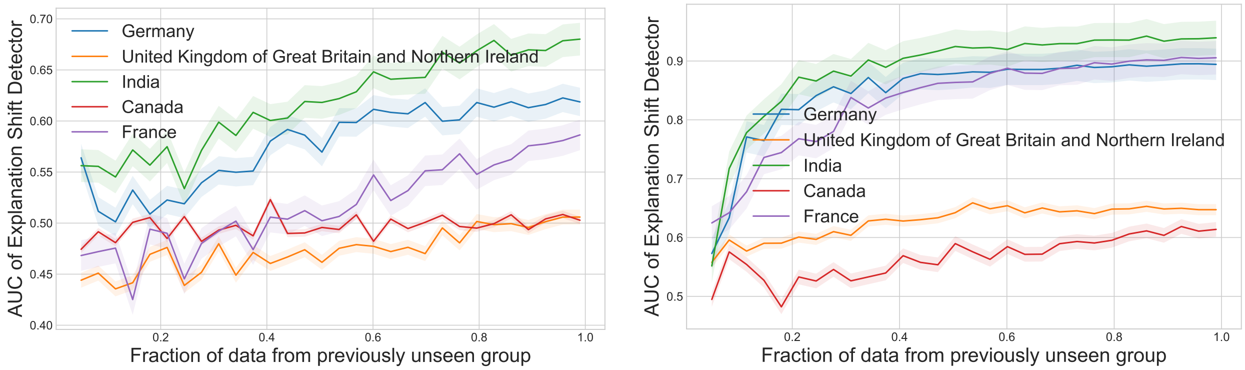


**Figure 10:** Both images represent the AUC of the *Explanation Shift Detector* for different countries on the StackOverflow survey dataset under novel group shift. In the left image, the estimator, $f_\theta$, is a gradient boosting decision tree; in the right image, for both cases the detector, $g_\psi$, is a logistic regression. By changing the type of estimator model, we can see how different types of models are affected differently for the same distribution shift

## C Experiments with Modeling Methods and Hyperparameters

In the next sections, we are going to show the sensitivity or our method to variations of the model $f$, the detector $g$, and the parameters of the estimator $f_\theta$.

As an experimental setup, we use the UCI Adult Income dataset. The experimental setup has been using Gradient Boosting Decision Tree as the model $f_\theta$ and then as "Explanation Shift Detector" $g_\psi$ a logistic regression. In this section, we extend the experimental setup by providing experiments by varying the types of algorithms for a given experimental set-up: the UCI Adult Income dataset using the Novel Covariate Group Shift for the "Asian" group with a fraction ratio of 0.5 (cf. Section 5).

## D LIME as an Alternative Explanation Method

Another feature attribution technique that satisfies the aforementioned properties (efficiency and uninformative features Section 2) and can be used to create the explanation distributions is LIME (Local Interpretable Model-Agnostic Explanations). The intuition behind LIME is to create a local interpretable model that approximates the behavior of the original model in a small neighbourhood of the desired data to explain (Ribeiro et al., 2016b;a) whose mathematical intuition is very similar to the Taylor series. In this work, we have proposed explanation shifts as a key indicator for investigating the impact of distribution shifts on ML models. In this section, we compare the explanation distributions composed by SHAP and LIME methods. LIME can potentially suffers several drawbacks:

**Computationally Expensive:** Its currently implementation is more computationally expensive than current SHAP implementations such as TreeSHAP (Lundberg et al., 2020), Data SHAP (Kwon et al., 2021; Ghorbani & Zou, 2019) or Local and Connected SHAP (Chen et al., 2019), the problem increases when we produce explanations of distributions. Even though implementations might be improved, LIME requires sampling data and fitting a linear model, which is a computationally more expensive approach than the aforementioned model-specific approaches to SHAP.

**Local Neighborhood:** The definition of a local "neighborhood", which can lead to instability of the explanations. Slight variations of this explanation hyperparameter lead to different local explanations. In Slack et al. (2020b) the authors showed that the explanations of two very close points can vary greatly.

**Dimensionality:** LIME requires as a hyperparameter the number of features to use for the local linear approximation. This creates a dimensionality problem as for our method to work, the explanation distributions have to be from the exact same dimensions as the input data. Reducing the number of features to be explained might improve the computational burden.

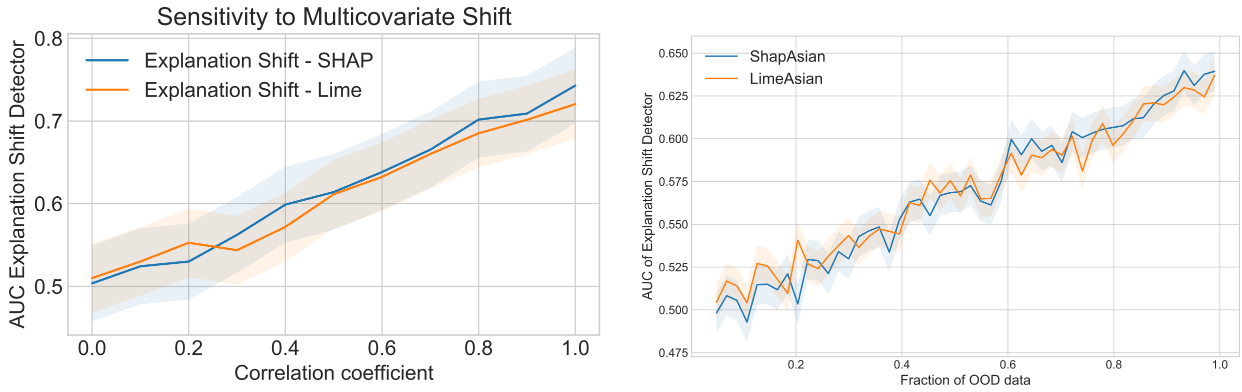


**Figure 11:** Comparison of the explanation distribution generated by LIME and SHAP. The left plot shows the sensitivity of the predicted probabilities to multicovariate changes using the synthetic data experimental setup of 2 on the main body of the paper. The right plot shows the distribution of explanation shifts for a New Covariate Category shift (Asian) in the ASC Income dataset.

Figure 11 compares the explanation distributions generated by LIME and SHAP. The left plot shows the sensitivity of the predicted probabilities to multicovariate changes using the synthetic data experimental setup from Figure 2 in the main body of the paper. The right plot shows the distribution of explanation shifts for a New Covariate Category shift (Asian) in the ASC Income dataset. The performance of OOD explanations detection is similar between the two methods, but LIME suffers from two drawbacks: its theoretical properties rely on the definition of a local neighborhood, which can lead to unstable explanations (false positives or false negatives on explanation shift detection), and its computational runtime required is much higher than that of SHAP (see experiments below).

## D.1 Runtime

We conducted an analysis of the runtimes of generating the explanation distributions using the two proposed methods. The experiments were run on a server with 4 vCPUs and 32 GB of RAM. We used `shap` version 0.41.0 and `lime` version 0.2.0.1 as software packages. In order to define the local neighborhood for both methods in this example we use all the data provided as background data. As an $f_\theta$ model, we use an `xgboost` and compare the results of TreeShap against LIME. When varying the number of samples we use 5 features and while varying the number of features we use 1000 samples.

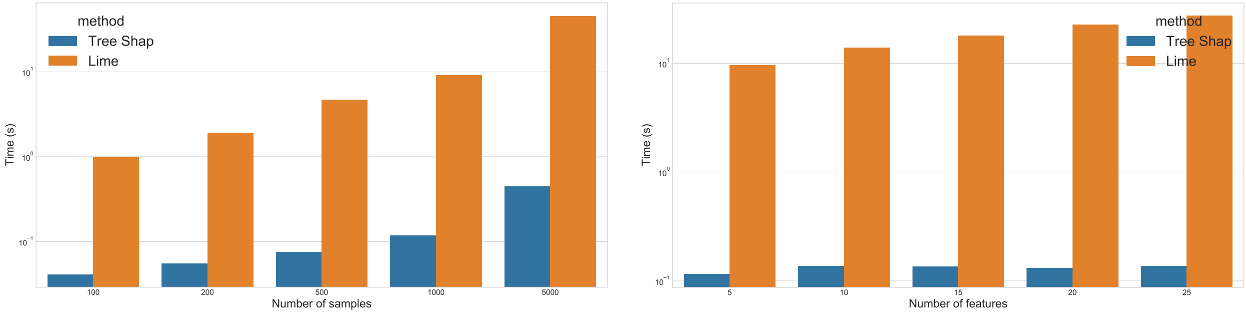


**Figure 12:** Wall time for generating explanation distributions using SHAP and LIME with different numbers of samples (left) and different numbers of columns (right). Note that the y-scale is logarithmic. The experiments were run on a server with 4 vCPUs and 32 GB of RAM. The runtime required to create an explanation distributions with LIME is far greater than SHAP for a gradient-boosting decision tree

Figure 12, shows the wall time required for generating explanation distributions using SHAP and LIME with varying numbers of samples and columns. The runtime required of generating an explanation distributions using LIME is much higher than using SHAP, especially when producing explanations for distributions. This is due to the fact that LIME requires training a local model for each instance of the input data to be explained, which can be computationally expensive. In contrast, SHAP relies on heuristic approximations to estimate the feature attribution with no need to train a model for each instance. The results illustrate that this difference in computational runtime becomes more pronounced as the number of samples and columns increases.

We note that the computational burden of generating the explanation distributions can be further reduced by limiting the number of features to be explained, as this reduces the dimensionality of the explanation distributions, but this will inhibit the quality of the explanation shift detection as it won't be able to detect changes on the distribution shift that impact model on those features.

Given the current state-of-the-art of software packages we have used SHAP values due to lower runtime required and that theoretical guarantees hold with the implementations. In the experiments performed in this paper, we are dealing with a medium-scaled dataset with around $\sim 1,000,000$ samples and $20-25$ features. Further work can be envisioned on developing novel mathematical analysis and software that study under which conditions which method is more suitable.

## E True to the Model or True to the Data?

The "Explanation Shift Detector" proposed in this work relies on the explanation distributions that satisfy efficiency and uninformative theoretical properties. We have used the Shapley values as an explainable AI method that satisfies these properties. However, the correct way to connect a model to a coalitional game, which is the central concept of Shapley values, is a source of controversy, with two main approaches $(i)$ an interventional (Aas et al., 2021; Frye et al., 2020; Zern et al., 2023) or $(ii)$ an observational formulation of the conditional expectation(Sundararajan & Najmi, 2020).

In the following experiment, we compare what are the differences between estimating the Shapley values using one or the other approach. We benchmark this experiment on the four prediction tasks based on the US census data (Ding et al., 2021a) and using the "Explanation Shift Detector", where both the model $f_\theta(X)$ and $g_\psi(\mathcal{S}(f_\theta, X))$ are linear models. We will calculate the Shapley values using the SHAP linear explainer. [1]

The comparison depends on a feature perturbation hyperparameter: whether the approach to compute the SHAP values is either *interventional* or *correlation dependent*. The interventional SHAP values break the dependence structure between features in the model to uncover how the model would behave if the inputs were changed (as it was an intervention). This option is said to stay "true to the model", meaning it will only give allocation credit to the features that the model actually uses (Aas et al., 2021).

On the other hand, the full conditional approximation of the SHAP values respects the correlations of the input features. If the model depends on one input that is correlated with another input, then both get some credit for the model's behaviour. This option is said to say "true to the data", meaning that it only considers how the model would behave when respecting the correlations in the input data (Chen et al., 2020).In our case, we will measure the difference between the two approaches by looking at the linear coefficients of the model $g_\psi$ and comparing the performance using the geo-political and temporal experiment of the previous section 5, for this case between CA14 and PR18.

In Table 10 and Table 11, we can see the comparison of the effects of using the aforementioned approaches to learn our proposed method, the "Explanation Shift Detector". Even though the two approaches differ theoretically, the differences become negligible when explaining the protected characteristic, i.e. when providing the linear regression coefficients.

[1] `https://shap.readthedocs.io/en/latest/generated/shap.explainers.Linear.html`

**Table 10:** AUC comparison of the "Explanation Shift Detector" between estimating the Shapley values between the interventional and the correlation-dependent approaches for the four prediction tasks based on the US census dataset (Ding et al., 2021a). The % character represents the relative difference. The performance differences are negligible.

| | Interventional | Observational | % |
|---|---|---|---|
| Income | 0.736438 | 0.736439 | 1.1e-06 |
| Employment | 0.747923 | 0.747923 | 4.44e-07 |
| Mobility | 0.690734 | 0.690735 | 8.2e-07 |
| Travel Time | 0.790512 | 0.790512 | 3.0e-07 |

**Table 11:** Linear regression coefficients comparison of the "Explanation Shift Detector" between estimating the Shapley values between the interventional and the correlation-dependent approaches for one of the US census-based prediction tasks (ACS Income). The % character represents the relative difference. The coefficients show negligible differences between the calculation methods

| | Interventional | Observational | % |
|---|---:|---:|---|
| Marital | 0.348170 | 0.348190 | 2.0e-05 |
| Worked Hours | 0.103258 | -0.103254 | 3.5e-06 |
| Class of worker | 0.579126 | 0.579119 | 6.6e-06 |
| Sex | 0.003494 | 0.003497 | 3.4e-06 |
| Occupation | 0.195736 | 0.195744 | 8.2e-06 |
| Age | -0.018958 | -0.018954 | 4.2e-06 |
| Education | -0.006840 | -0.006840 | 5.9e-07 |
| Relationship | 0.034209 | 0.034212 | 2.5e-06 |