Anomaly Attribution with Likelihood Compensation

Tsuyoshi Idé, Amit Dhurandhar, Jiří Navrátil, Moninder Singh, Naoki Abe
IBM Research, T. J. Watson Research Center
{tide, adhuran, jiri, moninder, nabe}@us.ibm.com

Abstract

This paper addresses the task of explaining anomalous predictions of a black-box regression model. When using a black-box model, such as one to predict building energy consumption from many sensor measurements, we often have a situation where some observed samples may significantly deviate from their prediction. It may be due to a sub-optimal black-box model, or simply because those samples are outliers. In either case, one would ideally want to compute a "responsibility score" indicative of the extent to which an input variable is responsible for the anomalous output. In this work, we formalize this task as a statistical inverse problem: Given model deviation from the expected value, infer the responsibility score of each of the input variables. We propose a new method called likelihood compensation (LC), which is founded on the likelihood principle and computes a correction to each input variable. To the best of our knowledge, this is the first principled framework that computes a responsibility score for real valued anomalous model deviations. We apply our approach to a real-world building energy prediction task and confirm its utility based on expert feedback.

1 Introduction

With the rapid development of Internet-of-Things technologies, anomaly detection has played a critical role in modern industrial applications of artificial intelligence (AI). One of the recent technology trends is to create a "digital twin" using a highly flexible machine learning model, typically deep neural networks, for monitoring the health of the production system (Tao et al. 2018). However, the more representational power the model has, the more difficult it is to understand its behavior. In particular, explaining deviations between predictions and true/expected measurements is one of the main pain points. A large deviation from the truth may be due to sub-optimal model training, or simply because the observed samples are outliers. If the model is black-box and the training dataset is not available, it is hard to determine which of these two situations have occurred. Nonetheless, we would still want to provide information to help end-users' in their decision making.

As such, in this paper we propose a method that can compute a "responsibility score" for each variable of a given input. We aver to this task as *anomaly attribution*. Specifically, we are concerned with model-agnostic anomaly attribution for black-box *regression* models, where we want to explain the deviation between the models prediction and the true/expected output in as concise a manner as possible. As a concrete example, consider the scenario of monitoring building energy consumption as the target variable y (see Section 5 for the detail). The input to the model is a multivariate sensor measurement $\boldsymbol{x}$ that is typically *real-valued* and *noisy*. Under the assumption that the model is black-box and the training data are not available, our goal is to compute a numerical score for each of the input variables, quantifying the extent to which they are responsible for the judgment that a given test sample is anomalous.

Anomaly attribution has been studied typically as a subtask of anomaly detection to date. For instance, in subspace-based anomaly detection, computing each variable's responsibility has been part of the standard procedure for years (Chandola, Banerjee, and Kumar 2009). However, there is little work on how anomaly attribution can be done when the model is black-box and the training data set is not available. In the XAI (explainable AI) community, on the other hand, growing attention has been paid to "post-hoc" explanations of black-box prediction models. Examples of the techniques include feature subset selection, feature importance scoring, and sample importance scoring (Costabello et al. 2019; Molnar 2019; Samek et al. 2019). For anomaly attribution, there are at least two post-hoc explanation approaches that are potentially applicable: (1) those based on the expected conditional deviation, best known as the Shapley value, which was first introduced to the machine learning community by Štrumbelj and Kononenko (2010), and (2) those based on local linear models, best known under the name of LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro, Singh, and Guestrin 2016). In spite of their popularity, these two approaches may not be directly useful for anomaly attribution: Given a test sample $(\boldsymbol{x}^t, y^t)$, these methods may explain what the value of $f(\boldsymbol{x}^t)$ itself can be attributed to. However, *what is more relevant is explaining the deviation of $f(\boldsymbol{x}^t)$ from the actual y^t*.

To address this limitation, we propose *likelihood compensation* (LC), a new local anomaly attribution approach for black-box regression models. We formalize the task of anomaly attribution as a statistical *inverse problem* that in-

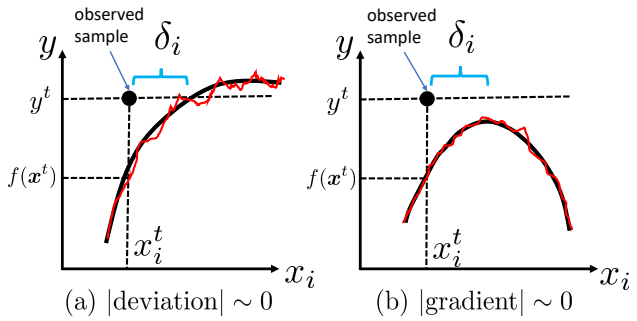


Figure 1: Illustration of the likelihood compensation along the i-th axis (δ_i). For a given test sample $(y^t, \boldsymbol{x}^t)$, LC seeks a perturbation that achieves the best possible fit with the black-box regression model $y = f(\boldsymbol{x})$, which could be non-smooth (the red curves). For more details please refer to Section 4.

fers a perturbation to the test input $\boldsymbol{x}^t$ from the deviation $y^t - f(\boldsymbol{x}^t)$, conversely to the forward problem that computes the deviation (or its variants) from $(\boldsymbol{x}^t, y^t)$. As illustrated in Fig. 1, LC can be viewed intuitively as the "deviation measured horizontally" if certain conditions are met. This admits direct interpretation as it suggests an action that might be taken to bring back the outlying sample to normalcy. Importantly, LC does not use any problem-specific assumptions but is built upon the maximum likelihood principle, the basic principle in statistical machine learning. To the best of our knowledge, this is the first principled framework for model-agnostic anomaly attribution in the regression setting.

2 Related Work

Although the machine learning community had not paid much attention to explainability of AI (XAI) in the black-box setting until recently, the last few years has seen a surge of interest in XAI research. For general background, Gade et al. (2020) provides a useful summary of major research issues in XAI for industries. In a more specific context of industrial anomaly detection, Langone et al. (2020) and Amarasinghe et al. (2018) give a useful summary of practical requirements of XAI in the deep learning era. An extensive survey on various XAI methodologies is given in (Costabello et al. 2019; Molnar 2019; Samek et al. 2019).

So far most of the model-agnostic post-hoc explanation studies are designed for classification, often restricted to image classification. As discussed before, two approaches are potentially applicable to the task of anomaly attribution in the black-box regression setting, namely the Shapley value (SV) (Štrumbelj and Kononenko 2010, 2014; Casalicchio, Molnar, and Bischl 2018) and the LIME (Ribeiro, Singh, and Guestrin 2016, 2018). The relationship between the two was discussed by Lundberg et al. (2017) assuming binary inputs. In the context of anomaly detection from noisy, real-valued data, two recent studies, Zhang et al. (2019) and Giurgiu et al. (2019), proposed a method built on LIME and SV, respectively. While these belong to the earliest model-agnostic XAI studies for anomaly detection, they naturally inherit the limitations of the existing approaches mentioned in introduction. Recently, Lucic et al. (2020) proposed a LIME-based approach for identifying a variable-wise normal range, which although related is different from our formulation of the anomaly attribution problem. Zemicheal et al. (2019) proposed an SV-like feature scoring method in the context of outlier detection, however, this does not apply to the regression setting.

One of the main contributions of this work is the proposal of a generic XAI framework for input attribution built upon the likelihood principle. The method of integrated gradient (Sundararajan, Taly, and Yan 2017) is another generic input attribution approach applicable to the black-box setting. Sipple (2020) recently applied it to anomaly detection and explanation. However, it is applicable to only the classification setting and, as pointed out by Sipple (2020), the need for the "baseline input" makes it less useful in practice. Layer-wise relevance propagation (Bach et al. 2015) is another well-known input attribution method and has been applied to real-world anomaly attribution tasks (Amarasinghe, Kenney, and Manic 2018). However, it is deep-learning-specific and assumes we have white-box access to the model.

Another research thread relevant to our work revolves around the counterfactual approach, which focuses on what is missing in the model (or training data) rather than what exists. In the context of image classification, the idea of counterfactuals is naturally translated into perturbation-based explanation (Fong and Vedaldi 2017). Recent works (Dhurandhar et al. 2018; Wachter, Mittelstadt, and Russell 2017) proposed the idea of contrastive explanation, which attempts to find a perturbation best characterizing a classification instance such that the probability of choosing a different class supersedes the original prediction. Our approach is similar in spirit, but as mentioned above, is designed for regression and uses a very different objective function. Moreover, both of these methods (Dhurandhar et al. 2018; Wachter, Mittelstadt, and Russell 2017) require white-box access, while ours is a black-box approach.

3 Problem Setting

As mentioned before, we focus on the task of anomaly attribution in the *regression setting* rather than classification or unsupervised settings. Throughout the paper, the input variable $\boldsymbol{x}$ is assumed to be noisy, multivariate, and real-valued in general. Our task is formally stated as follows:

Definition 1 (**Anomaly detection and attribution**). *Given a black-box regression model $y = f(\boldsymbol{x})$ and a test data set $\mathcal{D}_{\text{test}}$: (1) compute the score to represent the degree of anomaly of the prediction on $\mathcal{D}_{\text{test}}$; (2) compute the responsibility score for each input variable for the prediction being anomalous.*

The black-box regression model is assumed to be deterministic with $y \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^M$, where $M \geq 2$ is the dimensionality of the input random variable $\boldsymbol{x}$. The functional form of $f(\cdot)$ and the dependency on model parameters are not available to us. The *training data* on which the model was trained is *not* available either. The only interface to the model we are given is $\boldsymbol{x}$, which follows an unknown distri-

bution $P(\boldsymbol{x})$. Queries to get the response $f(\boldsymbol{x})$ can be performed cheaply at any $\boldsymbol{x}$.

The test data set is denoted as $\mathcal{D}_{\text{test}} = \{(\boldsymbol{x}^t, y^t) \mid t = 1, \ldots, N_{\text{test}}\}$, where t is the index for the t-th test sample and N_{test} is the number of test samples. When $N_{\text{test}} = 1$, the task may be called the *outlier detection and attribution*.

Anomaly detection as forward problem Assume that, from the deterministic regression model, we can somehow obtain $p(y \mid \boldsymbol{x})$, a probability density over y given the input signal $\boldsymbol{x}$ (see Sec. 4.2 for a proposed approach to do this). The standard approach to quantifying the degree of anomaly is to use the negative log-likelihood of test samples. Under the i.i.d. assumption, it can be written as

$$a(\mathcal{D}_{\text{test}}) = -\frac{1}{N_{\text{test}}} \sum_{t \in \mathcal{D}_{\text{test}}} \ln p(y^t \mid \boldsymbol{x}^t), \tag{1}$$

which is called the *anomaly score* for $\mathcal{D}_{\text{test}}$ (or the *outlier score* for a single sample dataset). An anomaly is declared when $a(\mathcal{D}_{\text{test}})$ exceeds a predefined threshold.

Anomaly attribution as inverse problem The above anomaly/outlier *detection* formulation is standard. However, the anomaly/outlier *attribution* is more challenging when the underlying model is black-box. This is in some sense an *inverse problem*: The function $f(\boldsymbol{x})$ readily gives an estimate of y from $\boldsymbol{x}$, but, in general, there is no obvious way to do the reverse in the *multivariate* case. When an estimate $f(\boldsymbol{x}^t)$ looks 'bad' in light of an observed y^t, what can we say about the contribution, or responsibility, of the input variables? Section 4 below gives our proposed answer to this question.

Notation We use boldface to denote vectors. The i-th dimension of a vector $\boldsymbol{\delta}$ is denoted as δ_i. The ℓ_1 and ℓ_2 norms of a vector are denoted by $\|\cdot\|_1$ and $\|\cdot\|_2$, respectively, and are defined as $\|\boldsymbol{\delta}\|_1 \triangleq \sum_i |\delta_i|$ and $\|\boldsymbol{\delta}\|_2 \triangleq \sqrt{\sum_i \delta_i^2}$. The sign function $\text{sign}(\delta_i)$ is defined as being 1 for $\delta_i > 0$, and -1 for $\delta_i < 0$. For $\delta_i = 0$, the function takes a value in $[-1, 1]$. For a vector input, the definition applies element-wise, giving a vector of the same size as the input.

We distinguish between a random variable and its realizations with a superscript. For notational simplicity, we symbolically use $p(\cdot)$ to represent different probability distributions, whenever there is no confusion. For instance, $p(\boldsymbol{x})$ is used to represent the probability density of a random variable $\boldsymbol{x}$ while $p(y \mid \boldsymbol{x})$ is a different distribution of another random variable y conditioned on $\boldsymbol{x}$. The Gaussian distribution of a random variable y is denoted by $\mathcal{N}(y \mid \cdot, \cdot)$, where the first and the second arguments after the bar are the mean and the variance, respectively. The multivariate Gaussian distribution is defined in a similar way.

4 The Method of Likelihood Compensation

This section presents the key idea of "likelihood compensation" as illustrated in Fig. 1. We start with a likelihood-based interpretation of LIME to highlight the idea.

4.1 Improving Likelihood via Corrected Input

For a given test sample $(y^t, \boldsymbol{x}^t)$, LIME minimizes the lasso objective to let the sparse regression estimation process select a subset of the variables. From a Bayesian perspective, it can be rewritten as a MAP (maximum a posteriori) problem:

$$\text{LIME: } \max_{\boldsymbol{\beta}} \left\langle \ln \left\{ p(y \mid \boldsymbol{x}^t, \boldsymbol{\beta})\, p(\boldsymbol{\beta}) \right\} \right\rangle_{\text{vic}(\boldsymbol{x}^t)} \quad \text{subject to} \quad y = f(\boldsymbol{x}), \tag{2}$$

where $\langle \cdot \rangle_{\text{vic}(\boldsymbol{x}^t)}$ denotes the expectation over random samples generated from an assumed local distribution in the vicinity of $\boldsymbol{x}^t$. For p's above, LIME uses the Gaussian observation model $p(y \mid \boldsymbol{x}, \boldsymbol{\beta}) = \mathcal{N}(y \mid \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}, \sigma^2)$ and the Laplace prior $p(\boldsymbol{\beta}) \propto \exp(-\nu \|\boldsymbol{\beta}\|_1)$. Here σ^2, ν are hyperparameters. The regression coefficient $\boldsymbol{\beta}$ as well as the intercept β_0 captures the local linear structure of f and is interpreted as the sensitivity of f at $\boldsymbol{x}^t$.

From the viewpoint of actionability, however, the slope can be less useful than $\boldsymbol{x}$ itself, particularly for the purpose of outlier attribution. If $(\boldsymbol{x}^t, y^t)$ is an outlier far from the population, how can we expect to obtain actionable insights from the local slope? Another issue is that y^t plays no role in this formulation. Notice the constraint of maximization: LIME amounts to assuming that the model is always right and is not sensitive to the question of whether $(y^t, \boldsymbol{x}^t)$ is an outlier or not.

Keeping this in mind, we propose to introduce a directly interpretable parameter $\boldsymbol{\delta}$ as a correction term to $\boldsymbol{x}$, rather than the slope as in LIME:

$$\text{Proposed: } \max_{\boldsymbol{\delta}} \left[\ln \left\{ p(y^t \mid f(\boldsymbol{x}^t + \boldsymbol{\delta}))\, p(\boldsymbol{\delta}) \right\} \right], \tag{3}$$

$$p(y \mid f(\boldsymbol{x} + \boldsymbol{\delta})) = \mathcal{N}(y \mid f(\boldsymbol{x} + \boldsymbol{\delta}), \sigma^2(\boldsymbol{x})). \tag{4}$$

The prior $p(\boldsymbol{\delta})$ can be designed to reflect problem-specific constraints such as infeasible regions so that the resultant $\boldsymbol{x} + \boldsymbol{\delta}$ is a realistic or high probability input. Considering the well-known issue of lasso that in the presence of multiple correlated explanatory variables it tends to pick one at random (Roy, Chakraborty et al. 2017), we employ $p(\boldsymbol{\delta}) \propto \exp\left(-\frac{1}{2}\lambda \|\boldsymbol{\delta}\|_2^2 - \nu \|\boldsymbol{\delta}\|_1\right)$. $\sigma^2(\boldsymbol{x})$ is the local variance representing the uncertainty of prediction (see Sec. 4.2), and λ, ν are hyperparameters controlling the sparsity and the overall scale of $\boldsymbol{\delta}$ (see Sec. 4.4 for typical values). We call $\boldsymbol{\delta}$ the **likelihood compensation** (LC) as it compensates for the loss in likelihood incurred by an anomalous prediction. Note that, unlike LIME, our explainabiliy model is neither linear nor additive, being free from the "masking effect" (Hastie, Tibshirani, and Friedman 2009) observed in linear XAI models.

We can naturally extend the point-wise definition of Eq. (3) to a collection of test samples. For the Gaussian observation and the elastic net prior, we have the following optimization problem for the LC for $\mathcal{D}_{\text{test}}$:

$$\min_{\boldsymbol{\delta}} \left\{ \frac{1}{N_{\text{test}}} \sum_{t=1}^{N_{\text{test}}} \frac{\left[y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})\right]^2}{2\sigma_t^2} + \frac{\lambda}{2} \|\boldsymbol{\delta}\|_2^2 + \nu \|\boldsymbol{\delta}\|_1 \right\}, \tag{5}$$

where σ_t^2 is the local variance evaluated at $\boldsymbol{x}^t$. *This is the main problem studied in this paper.*

4.2 Deriving Probabilistic Prediction Model

So far we have assumed the predictive distribution $p(y \mid \boldsymbol{x})$ is given. Now let us think about how to derive it from the deterministic black-box regression model $y = f(\boldsymbol{x})$.

If there are too few test samples, we have no choice but to set σ_t^2 to a constant using prior knowledge. Otherwise, we can obtain an estimate of σ_t^2 using a subset of $\mathcal{D}_{\text{test}}$ in a cross-validation (CV)-like fashion. Let $\mathcal{D}_{\text{ho}}^t = \{(\boldsymbol{x}^{(n)}, y^{(n)}) \mid n = 1, \ldots, N_{\text{ho}}\} \subset \mathcal{D}_{\text{test}}$ be a held-out data set that does not include the given test sample $(\boldsymbol{x}^t, y^t)$. For the observation model Eq. (4) and the test sample $\boldsymbol{x}^t$, we consider a locally weighted version of maximum likelihood:

$$\max_{\sigma^2} \sum_{n=1}^{N_{\text{ho}}} w_n(\boldsymbol{x}^t) \left\{ \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(y^{(n)} - f(\boldsymbol{x}^{(n)}))^2}{2\sigma^2} \right\}, \quad (6)$$

where $w_n(\boldsymbol{x}^t)$ is the similarity between $\boldsymbol{x}^t$ and $\boldsymbol{x}^{(n)}$. A reasonable choice for w_n is the Gaussian kernel:

$$w_n(\boldsymbol{x}^t) = \mathcal{N}(\boldsymbol{x}^{(n)} \mid \boldsymbol{x}^t, \text{diag}(\boldsymbol{\eta})), \quad (7)$$

where $\text{diag}(\boldsymbol{\eta})$ is a diagonal matrix whose i-th diagonal is given by η_i, which can be of the same order as the sample variance of x_i evaluated on $\mathcal{D}_{\text{ho}}$.

The maximizer of Eq. (6) can be found by differentiating w.r.t. σ^{-2}. The solution is given by

$$\sigma_t^2 = \frac{1}{\sum_m w_m(\boldsymbol{x}^t)} \sum_{n=1}^{N_{\text{ho}}} w_n(\boldsymbol{x}^t) \left[y^{(n)} - f(\boldsymbol{x}^{(n)}) \right]^2. \quad (8)$$

This has to be computed for each $\boldsymbol{x}^t \in \mathcal{D}_{\text{test}}$.

4.3 Solving the Optimization Problem

Although seemingly simple, solving the optimization problem (5) is generally challenging. Due to the black-box nature of f, we do not have access to the parametric form of f, let alone the gradient. In addition, as is the case in deep neural networks, f can be non-smooth (see the red curves in Fig. 1), which makes numerical estimation of the gradient tricky.

To derive an optimization algorithm, we first note that there are two origins of non-smoothness in the objective function in (5). One is inherent to f while the other is due to the added ℓ_1 penalty. To separate them, let us denote the objective function in Eq. (5) as $J(\boldsymbol{\delta}) + \nu\|\boldsymbol{\delta}\|_1$, where J contains the first and second terms. Since we are interested only in a local solution in the vicinity of $\boldsymbol{\delta} = \mathbf{0}$, it is natural to adopt an iterative update algorithm starting from $\boldsymbol{\delta} \approx \mathbf{0}$. Suppose that we have an estimate $\boldsymbol{\delta} = \boldsymbol{\delta}^{\text{old}}$ that we wish to update. If we have a reasonable approximation of the gradient in its vicinity, denoted by $\langle\!\langle \nabla J(\boldsymbol{\delta}^{\text{old}}) \rangle\!\rangle$, the next estimate can be found by

$$\boldsymbol{\delta}^{\text{new}} = \arg\min_{\boldsymbol{\delta}} \Big\{ J(\boldsymbol{\delta}^{\text{old}}) + (\boldsymbol{\delta} - \boldsymbol{\delta}^{\text{old}})^\top \langle\!\langle \nabla J(\boldsymbol{\delta}^{\text{old}}) \rangle\!\rangle + \frac{1}{2\kappa}\|\boldsymbol{\delta} - \boldsymbol{\delta}^{\text{old}}\|_2^2 + \nu\|\boldsymbol{\delta}\|_1 \Big\} \quad (9)$$

in the same spirit of the proximal gradient (Parikh, Boyd et al. 2014), where κ is a hyperparameter representing the learning rate. Notice that the first three terms in the curly bracket correspond to a second-order approximation of $J(\boldsymbol{\delta})$ in the vicinity of $\boldsymbol{\delta}^{\text{old}}$. We find the best estimate under this approximation.

The r.h.s. has an analytic solution. Define $\boldsymbol{\phi} \triangleq \boldsymbol{\delta}^{\text{old}} - \kappa\langle\!\langle \nabla J(\boldsymbol{\delta}^{\text{old}}) \rangle\!\rangle$. The optimality condition is $\boldsymbol{\delta} - \boldsymbol{\phi} + \kappa\nu \,\text{sign}(\boldsymbol{\delta}) = \mathbf{0}$. If $\phi_i > \kappa\nu$ holds for the i-th dimension, by $\phi_i \pm \kappa > 0$, we have $\delta_i = \phi_i - \kappa\nu \,\text{sign}(\delta_i) = \phi_i - \kappa\nu$. Similar arguments straightforwardly verify the following solution:

$$\delta_i = \begin{cases} \phi_i - \kappa\nu, & \phi_i > \kappa\nu \\ 0, & |\phi_i| \le \kappa\nu \\ \phi_i + \kappa\nu, & \phi_i < -\kappa\nu \end{cases}. \quad (10)$$

Performing differentiation, we see that $\boldsymbol{\phi}$ is given by

$$\boldsymbol{\phi} = (1 - \kappa\lambda)\boldsymbol{\delta}^{\text{old}} + \kappa \frac{1}{N_{\text{test}}} \sum_{t=1}^{N_{\text{test}}} \left\{ \frac{y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})}{\sigma_t^2} \right\} \left\langle\!\!\left\langle \frac{\partial f(\boldsymbol{x}^t + \boldsymbol{\delta})}{\partial \boldsymbol{\delta}} \right\rangle\!\!\right\rangle. \quad (11)$$

Note that $f(\boldsymbol{x}^t + \boldsymbol{\delta})$ is readily available at any $\boldsymbol{\delta}$ without approximation. Here we provide some intuition behind the updating equation (11). Convergence is achieved when either the deviation $y^t - f$ or the gradient $\langle\!\langle \partial f / \partial \boldsymbol{\delta} \rangle\!\rangle$ vanishes at $\boldsymbol{x}^t + \boldsymbol{\delta}$. The former and the latter correspond, respectively, to the situations illustrated in Fig. 1 (a) and (b). As shown in the figure, δ_i corresponds to the horizontal deviation along the x_i axis between the test sample and the regression function. If there is no horizontal intersection on the regression surface it seeks the zero gradient point based on a smooth surrogate of the gradient.

To find $\langle\!\langle \partial f / \partial \boldsymbol{\delta} \rangle\!\rangle$, a smooth surrogate of the gradient, we propose a simple sampling-based procedure. Specifically, we draw N_{s} samples from a local distribution at $\boldsymbol{x}^t + \boldsymbol{\delta}$ as

$$\boldsymbol{x}^{[m]} \sim \mathcal{N}(\cdot \mid \boldsymbol{x}^t + \boldsymbol{\delta}, \text{diag}(\boldsymbol{\eta})), \quad (12)$$

and fit a linear regression model $f = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}$ on the populated local data set $\{(\boldsymbol{x}^{[m]}, f^{[m]}) \mid m = 1, \ldots, N_{\text{s}}\}$, where $f^{[m]} = f(\boldsymbol{x}^{[m]})$. Solving the least squares problem, we have

$$\left\langle\!\!\left\langle \frac{\partial f(\boldsymbol{x}^t + \boldsymbol{\delta})}{\partial \boldsymbol{\delta}} \right\rangle\!\!\right\rangle = \boldsymbol{\beta} = \left[\Psi_{\text{s}} \Psi_{\text{s}}^\top + \varepsilon \mathsf{I}_M \right]^{-1} \Psi_{\text{s}} \boldsymbol{f}_{\text{s}}, \quad (13)$$

where $\varepsilon \approx 0$ is a small positive constant added to the diagonals for numerical stability. In Eq. (13), we have defined $\boldsymbol{f}_{\text{s}} \triangleq [f^{[1]} - \overline{f}, \ldots, f^{[N_{\text{s}}]} - \overline{f}]^\top$ and $\Psi_{\text{s}} \triangleq [\boldsymbol{x}^{[1]} - \overline{\boldsymbol{x}}, \ldots, \boldsymbol{x}^{[N_{\text{s}}]} - \overline{\boldsymbol{x}}]$. As usual, the population means are defined as $\overline{f} \triangleq \frac{1}{N_{\text{s}}} \sum_m f^{[m]}$ and $\overline{\boldsymbol{x}} \triangleq \frac{1}{N_{\text{s}}} \sum_m \boldsymbol{x}^{[m]}$.

4.4 Algorithm Summary

Algorithm 1 summarizes the iterative procedure for finding $\boldsymbol{\delta}$. The most important parameter is the ℓ_1 regularization strength ν, which has to be hand-tuned depending on the business requirements of the application of interest. On the other hand, the ℓ_2 strength λ controls the overall scale of $\boldsymbol{\delta}$. It can be fixed to some value between 0 and 1. In our experiments, it was adjusted so its scale is on the same order as LIME's output for consistency. It is generally recommended to rescale the input variables to have the zero

Algorithm 1 Likelihood Compensation

Input: $f(\boldsymbol{x}), \mathcal{D}_{\text{test}}$.
Parameters: λ, ν, κ.

for all $\boldsymbol{x}^t \in \mathcal{D}_{\text{test}}$ **do**
  Compute σ_t^2 with Eq. (8).
end for
Randomly initialize $\boldsymbol{\delta} \approx \mathbf{0}$.
repeat
  Set $\boldsymbol{g} = \mathbf{0}$.
  for all $\boldsymbol{x}^t \in \mathcal{D}_{\text{test}}$ **do**
    Compute $\boldsymbol{\beta}$ with Eq. (13).
    Update $\boldsymbol{g} \leftarrow \boldsymbol{g} + \boldsymbol{\beta} \frac{y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})}{N_{\text{test}} \sigma_t^2}$.
  end for
  $\boldsymbol{\phi} = (1 - \kappa\lambda)\boldsymbol{\delta} + \kappa \boldsymbol{g}$.
  Find $\boldsymbol{\delta}$ with Eq. (10).
until convergence.
return $\boldsymbol{\delta}$

mean and unit variance before starting the iteration (assuming $N_{\text{test}} \gg 1$), and retrieve the scale factors after convergence. For the learning rate κ, in our experiments, we fixed $\kappa = 0.1$ and shrank it (geometrically) by a factor of 0.98 in every iteration.

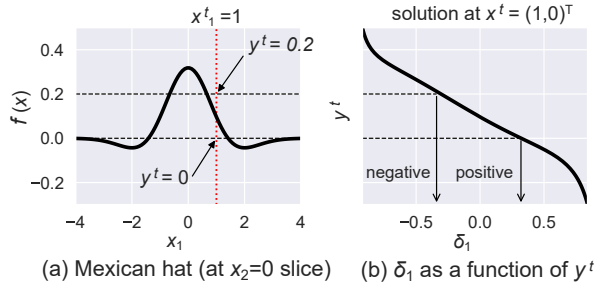
In addition to the parameters listed in Algorithm 1, the sampling-based estimation of the gradient Eq. (13) requires two minor parameters, $N^{\text{s}}, \boldsymbol{\eta}$. In the experiment, we fixed $N^{\text{s}} = 1\,000$ following (Ribeiro, Singh, and Guestrin 2016) and $\eta_i = 1$ for all i after standardization. The same $\boldsymbol{\eta}$ was used for Eq. (7).

5 Experiments

We now describe our experimental design and baselines we compare against in the empirical studies that follow.

Evaluation strategy Explainability of AI is generally evaluated from three major perspectives (Costabello et al. 2019): decomposability, simulatability, and algorithmic transparency. In post-hoc explanations of black-box models, decomposability and simulatability are most important. We thus design our experiments to answer the following questions: a) whether LC can provide extra information on specific anomalous samples beyond the baseline methods (decomposability), b) whether LC can robustly compute the responsibility score under heavy noise (simulatability), and c) whether LC can provide actionable insights in a real-world business scenario (simulatability). Regarding the third question, we validated our approach with feedback from domain experts as opposed to "crowd sourced" studies with lay users. In industrial applications, the end-user's interests can be highly specific to particular business needs and the system's inner workings tend to be difficult for non-experts to understand and simulate.

Baselines We compare LC with three possible alternatives: (1) Z-score and extended versions of (2) Shapley values (SV) and (3) LIME. The Z-score is the standard univariate outlier detection method in the unsupervised setting, and that of x_i^t is defined as $(x_i^t - m_i)/\sigma_i$, where m_i, σ_i are the mean and the standard deviation of x_i in $\mathcal{D}_{\text{test}}$, respectively. Shapley values (SV) and LIME are used as a proxy of the prior works (Zhang et al. 2019; Giurgiu and Schumann 2019; Lucic, Haned, and de Rijke 2020), which used SV or LIME in certain tasks similar to ours. For fair comparison, we extended these methods to be applied on the deviation $f - y$ instead of f itself, and name them LIME+ and SV+, respectively. We dropped SV+ in the building energy experiment as the training data was not available to compute the null/base values for each variable that SV requires. Note that contrastive and counterfactual methods such as (Dhurandhar et al. 2018; Wachter, Mittelstadt, and Russell 2017) are not valid competitors here as they require white-box access to the model and are predominantly used in classification settings.



(a) Mexican hat (at x_2=0 slice) (b) δ_1 as a function of y^t

Figure 2: Mexican Hat: (a) The $x_2 = 0$ slice of $f(\boldsymbol{x})$. (b) Computed δ_1 as a function of y^t.

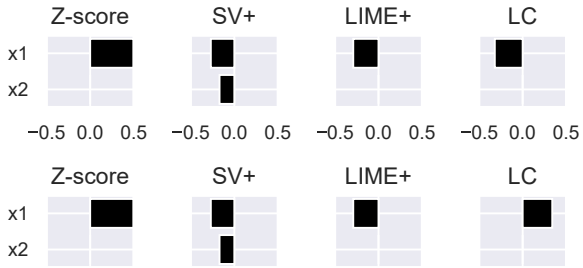


Figure 3: Mexican Hat: Comparison of the responsibility scores evaluated at $y^t = 0.2$ (upper) and 0 (lower).

Two-Dimensional Mexican Hat One of the major features of LC is its capability to provide explanations relevant to specific anomalous samples. To illustrate this, we used the two-dimensional Mexican Hat for the regression function $f(\boldsymbol{x}) \propto (1 - \frac{1}{2}\|\boldsymbol{x}\|_2^2) \exp(-\frac{1}{2}\|\boldsymbol{x}\|_2^2)$ as shown in Fig. 2 (a). Suppose we have obtained a test sample at $\boldsymbol{x}^t = (1, 0)^\top$. By symmetry, LIME+ has only the x_1 component, which can be analytically calculated to be -0.29 at this $\boldsymbol{x}^t$ when $\nu \to 0_+$. Similarly, LC has only the x_1 component, and is computed through iterative updates with the aid of analytic expression of the gradient. For SV+, we used uniform sam-

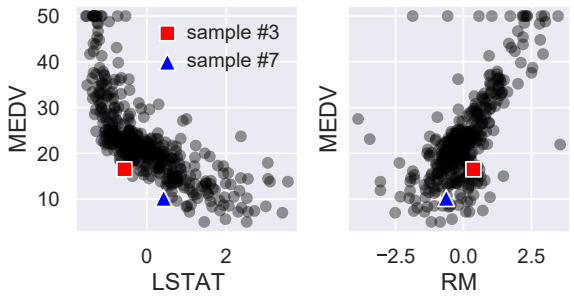


Figure 4: Boston Housing: Pairwise scatter plot between y (MEDV) and two selected input variables (LSTAT, RM).

pling from $[-4, 4]^2$ to evaluate the expectations. Figure 2 (b) shows the calculated values of δ_1 *as a function of* y^t with $\nu = 0, \lambda = 0.01$.

Figure 3 compares Z-score, SV+, LIME+, and LC for the two particular values of y^t, corresponding to the $f > y^t$ and $f < y^t$ cases. As shown, Z-score, SV+, and LIME+ are not able to distinguish between the two cases, demonstrating the limited utility in anomaly attribution. In contrast, LC's value of δ_1 corresponds to the horizontal distance between the test point and the curve of f as shown in Fig. 2. Hence we can think of it as a measure of "*horizontal deviation*," as we illustrated earlier in Fig. 1.

Boston Housing Next we used Boston Housing data (Belsley 1980) to test the robustness to noise. The task is to predict the median home price ('MEDV') of the districts in Boston with $M = 13$ input variables such as the percentage of lower status of the population ('LSTAT') and the average number of rooms ('RM'). As one might expect, the data is very noisy. As an illustrative example, Fig. 4 shows scatter plots between y (MEDV) and two selected input variables (LSTAT, RM), which have the highest correlations with y. We held out 20% of the data as $\mathcal{D}_{\text{test}}$ ($N_{\text{test}} = 101$), and trained a random forest on the rest. Then we picked the two top outliers, as highlighted as #3 and # 7 in Fig. 4. These are the two samples with the highest outlier scores of Eq. (1), to which not only LSTAT and RM but also all the other variables contributed.

Figure 5 compares the results of LC with the baselines for these outliers. For the ℓ_1 parameter, we gave $\nu = 0.1$ for LC, then chose $\nu = 0.005$ for LIME+, so that LIME+ has on average the same number of nonzero elements as LC. The ℓ_2 parameter λ was chosen as 0.5 for LC and LIME+ to have approximately the same scale. For SV+, all the $2^{M-1}M = 53\,248$ combinations were evaluated with the empirical distribution of the training samples, which are actually supposed to be unavailable in our setting, requiring about an hour to finish on a laptop PC (Core i7-8850H) for each test sample, while LC required only several seconds. From the figure, we see that overall SV+, LIME+, and LC are consistent in the sense that most of the weights appear on a few common variables including LSTAT. Z-score behaves quite differently, reflecting the fact that it is agnostic to the y-$\boldsymbol{x}$ relationship.

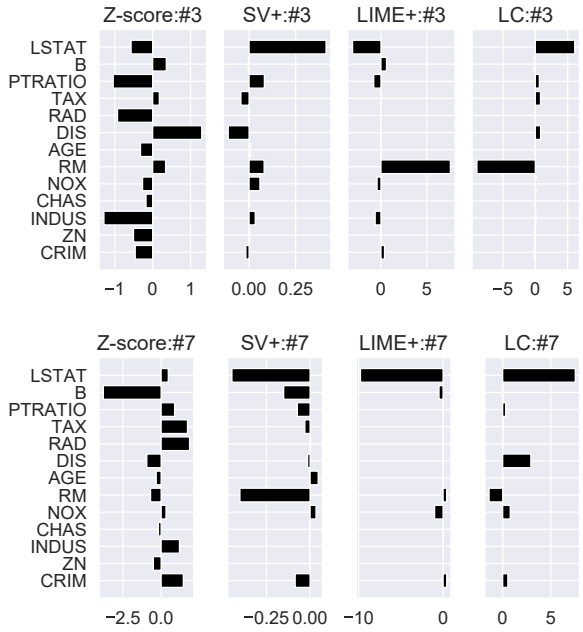


Figure 5: Boston Housing: Comparison of the responsibility scores for the top two outliers (#3 and #7).

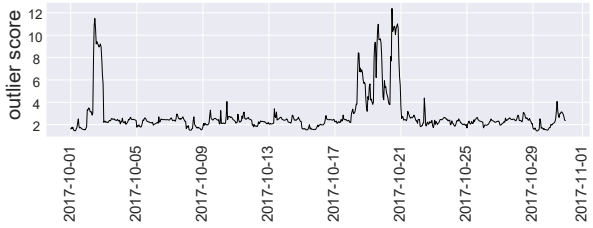


Figure 6: Building Energy: Outlier score computed with Eq. (1) for the test data.

For these outliers, LC gave positive and negative scores for LSTAT and RM in Fig. 5, respectively. Checking the scatter plots in Fig. 4, we can confirm the LC's characterization as the horizontal deviation that a positive (negative) score means "a positive (negative) shift will give a better fit." In contrast, LIME+ simply indicates whether the local slope is positive or negative, independently how the test samples deviate. In fact, one can mathematically show that LIME+ is invariant to the value of y, meaning that LIME cannot be a useful tool for instance-specific anomaly attribution.

In SV+, the situation is more subtle. It does not allow simple interpretations like LC or LIME+. The sign of the scores unpredictably becomes negative or positive, probably due to complicated effects of higher-order correlations. This suggests SV's tendency to be unstable under noise. In fact, our bootstrap analysis (not included for page limitation) shows that the SV+ scores are vulnerable to noise; The top three variables with the highest absolute SV+ scores gave a 35.3% variability relative to the mean. In addition, SV+ needs training data or the true distribution of $\boldsymbol{x}$ for Monte Carlo evaluation. Z-score, LIME+, and LC do not have such a requirement. Along with the prohibitive computational cost, those

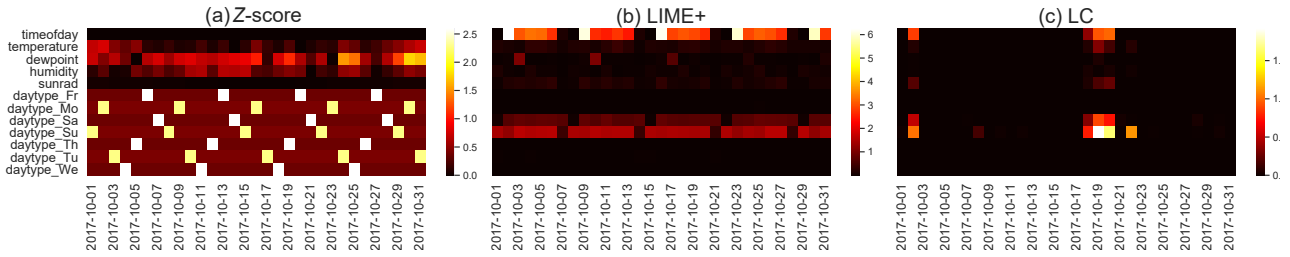


Figure 7: Building Energy: Comparison of the responsibility scores computed for the test data.

limitations make it impractical to apply SV+ to real-world system monitoring scenarios of the type presented below.

Real-World Application: Building Energy Management Finally, we applied LC to a building administration task. Collaborating with an entity offering building management services and products, we obtained energy consumption data for an office building in India. The total wattage is predicted by a black-box model as a function of weather-related (temperature, humidity, etc.) and time-related variables (time of day, day of week, month, etc.). There are two intended usages of the predictive model. One is near future prediction with short time windows for optimizing HVAC (heating, ventilating, and air conditioning) system control. The other is *retrospective* analysis over the last few months for the purpose of planning long-term improvement of the building facility and its management policies. In the retrospective analysis, it is critical to get clear explanation on unusual events.

At the beginning of the project, we interviewed 10 professionals on what kind of model explainability would be most useful for them. Their top priority capabilities were uncertainty quantification in forecasting and anomaly diagnosis in retrospective analysis. Our choices in the current research reflect these business requirements.

We obtained a one month worth of test data with $M = 12$ input variables recorded hourly. We first computed σ_t^2 according to Eq. (8) in which we leave $(y^t, \boldsymbol{x}^t)$ out for each t. For each of the test samples, we computed the outlier score by Eq. (1) under the Gaussian observation model, which resulted in a few conspicuous anomalies as shown in Fig. 6. An important business question was who or what may be responsible for those anomalies.

To obtain insights regarding the detected anomalies, we computed the LC score as shown in Fig. 7, where we computed $\boldsymbol{\delta}$ each day with $N_{\text{test}} = 24$ in Eq. (5), and visualized $\|\boldsymbol{\delta}\|_2^2$. For the Z-score, we visualized the daily mean of the absolute values. For LIME+, we computed regression coefficients for every sample, and visualized the ℓ_2 norm of their daily mean. We used $(\nu, \lambda) = (0.1, 0.5)$, which was determined by the level of sparsity and scale preferred by the domain experts.

As shown in the plot, the LC score clearly highlights a few variables whenever the outlier score is exceptionally high in Fig. 6, while the Z-score and LIME+ do not provide much information beyond the trivial weekly patterns. The pattern of LIME+ was very stable over $0 < \nu \leq 1$, showing empirical evidence of insensitivity to outliers. As mentioned before, one can mathematically prove this important fact: LIME+ as well as SV+ are invariant to the translation in f. On the other hand, the Z-score sensitively captures the variability in the weather-related variables, but it fails to explain the deviations in Fig. 6. This is understandable because the Z-score does not reflect the relationship between y and $\boldsymbol{x}$. The artifact seen in the "daytype" variables is due to the one-hot encoding of the day of week.

Finally, with LC, the variables highlighted around October 19 (Thursday) are 'timeofday', 'daytype_Sa', and 'daytype_Su', implying that those days had an unusual daily wattage pattern for a weekday and looked more like weekend days. Interestingly, it turned out that the 19th was a national holiday in India and many workers were off on and around that date. Thus we conclude that the anomaly is most likely not due to any faulty building facility, but due to the model limitation caused by the lack of full calendar information. Though simple, such pointed insights made possible by our method were highly appreciated by the professionals.

6 Conclusions

We have proposed a new method for model-agnostic explainability in the context of regression-based anomaly attribution. To the best of our knowledge, the proposed method provides the first principled framework for contrastive explainability in regression. The recommended responsibility score Likelihood Compensation is built upon the maximum likelihood principle. This is very different from the objectives used to obtain contrastive explanations in the classification setting. We demonstrated the advantages of the proposed method based on synthetic and real data, as well as on a real-world use-case of building energy management where we sought expert feedback.

Acknowledgements

The authors thank Dr. Kaoutar El Maghraoui for her support and technical suggestions. T.I. is partially supported by the Department of Energy National Energy Technology Laboratory under Award Number DE-OE0000911. A part of this report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein

to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

References

Amarasinghe, K.; Kenney, K.; and Manic, M. 2018. Toward explainable deep neural network based anomaly detection. In *Proc. Intl. Conf. Human System Interaction (HSI)*, 311–317. IEEE.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10(7): e0130140.

Belsley, K. . W. 1980. *Regression diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley.

Casalicchio, G.; Molnar, C.; and Bischl, B. 2018. Visualizing the feature importance for black box models. In *Proc. Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 655–670. Springer.

Chandola, V.; Banerjee, A.; and Kumar, V. 2009. Anomaly Detection: A Survey. *ACM Computing Survey* 41(3): 1–58.

Costabello, L.; Giannotti, F.; Guidotti, R.; Hitzler, P.; Lécué, F.; Minervini, P.; and Sarker, K. 2019. On Explainable AI: From Theory to Motivation, Applications and Limitations. In *Tutorial, AAAI Conference on Artificial Intelligence*.

Dhurandhar, A.; Chen, P.-Y.; Luss, R.; Tu, C.-C.; Ting, P.; Shanmugam, K.; and Das, P. 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*, 592–603.

Fong, R. C.; and Vedaldi, A. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proc. IEEE Intl. Conf. Computer Vision*, 3429–3437.

Gade, K.; Geyik, S.; Kenthapadi, K.; Mithal, V.; and Taly, A. 2020. Explainable AI in Industry: Practical Challenges and Lessons Learned. In *Companion Proceedings of the Web Conference 2020*, 303–304.

Giurgiu, I.; and Schumann, A. 2019. Additive Explanations for Anomalies Detected from Multivariate Temporal Data. In *Proc. Intl. Conf. Information and Knowledge Management*, 2245–2248. ACM.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition.

Langone, R.; Cuzzocrea, A.; and Skantzos, N. 2020. Interpretable Anomaly Prediction: Predicting anomalous behavior in industry 4.0 settings via regularized logistic regression tools. *Data & Knowledge Engineering* 101850.

Lucic, A.; Haned, H.; and de Rijke, M. 2020. Why does my model fail? contrastive local explanations for retail forecasting. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 90–98.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 4765–4774.

Molnar, C. 2019. *Interpretable machine learning*. Lulu.

Parikh, N.; Boyd, S.; et al. 2014. Proximal algorithms. *Foundations and Trends in Optimization* 1(3): 127–239.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *Proc. ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining*, 1135–1144. ACM.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-precision model-agnostic explanations. In *Proc. AAAI Conference on Artificial Intelligence*.

Roy, V.; Chakraborty, S.; et al. 2017. Selection of tuning parameters, solution paths and standard errors for Bayesian lassos. *Bayesian Analysis* 12(3): 753–778.

Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L. K.; and Müller, K.-R. 2019. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature.

Sipple, J. 2020. Interpretable, Multidimensional, Multimodal Anomaly Detection with Negative Sampling for Detection of Device Failure. In *Proc. Intl. Conf. Machine Learning*, 9016–9025.

Štrumbelj, E.; and Kononenko, I. 2010. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research* 11(Jan): 1–18.

Štrumbelj, E.; and Kononenko, I. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems* 41(3): 647–665.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *Proc. Intl. Conf. Machine Learning*, 3319–3328.

Tao, F.; Cheng, J.; Qi, Q.; Zhang, M.; Zhang, H.; and Sui, F. 2018. Digital twin-driven product design, manufacturing and service with big data. *The International Journal of Advanced Manufacturing Technology* 94(9-12): 3563–3576.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology* 31: 841.

Zemicheal, T.; and Dietterich, T. G. 2019. Anomaly detection in the presence of missing values for weather data quality control. In *Proc. ACM SIGCAS Conf. Computing and Sustainable Societies*, 65–73.

Zhang, X.; Marwah, M.; Lee, I.-t.; Arlitt, M.; and Goldwasser, D. 2019. ACE–An Anomaly Contribution Explainer for Cyber-Security Applications. In *Proc. IEEE Intl. Conf. Big Data*, 1991–2000.