

Monitoring Model Deterioration with Explainable Uncertainty Estimation via Non-parametric Bootstrap

Carlous Mougan^{*1}, Dan Saattrup Nielsen^{*2}

¹ University of Southampton, United Kingdom

² The Alexandra Institute, Denmark

c.mougan@southampton.ac.uk, dan.nielsen@alexandra.dk

Abstract

Monitoring machine learning models once they are deployed is challenging. It is even more challenging to decide when to retrain models in real-case scenarios when labeled data is beyond reach, and monitoring performance metrics becomes unfeasible. In this work, we use non-parametric bootstrapped uncertainty estimates and SHAP values to provide explainable uncertainty estimation as a technique that aims to monitor the deterioration of machine learning models in deployment environments, as well as determine the source of model deterioration when target labels are not available. Classical methods are purely aimed at detecting distribution shift, which can lead to false positives in the sense that the model has not deteriorated despite a shift in the data distribution. To estimate model uncertainty we construct prediction intervals using a novel bootstrap method, which improves previous state-of-the-art work. We show that both our model deterioration detection system as well as our uncertainty estimation method achieve better performance than the current state-of-the-art. Finally, we use explainable AI techniques to gain an understanding of the drivers of model deterioration. We release an open source Python package, `doubt`, which implements our proposed methods, as well as the code used to reproduce our experiments.

Introduction

Monitoring machine learning models in production is not an easy task. There are situations when the true label of the deployment data is available, and performance metrics can be monitored. But there are cases where it is not, and performance metrics are not so trivial to calculate once the model has been deployed. Model monitoring aims to ensure that a machine learning application in a production environment displays consistent behavior over time.

Being able to explain or remain accountable for the performance or the deterioration of a deployed model is crucial, as a drop in model performance can affect the whole business process (Mougan, Kanellos, and Gotttron 2021), potentially having catastrophic consequences¹. Once a deployed model has deteriorated, models are retrained using previous and

new input data in order to maintain high performance. This process is called continual learning (Diethe et al. 2018) and it can be computationally expensive and put high demands on the software engineering system. Deciding when to retrain machine learning models is paramount in many situations.

Traditional machine learning systems assume that training data has been generated from a stationary source, but *data is not static, it evolves*. This problem can be seen as a distribution shift, where the data distributions of the training set and the test set differ. Detecting distribution shifts has been a longstanding problem in the machine learning (ML) research community (Shimodaira 2000; Sugiyama, Krauledat, and Müller 2007; Sugiyama and Müller 2005; Tasche 2017; Zadrozny 2004; Stolzenberg and Relles 1997; Heckman 1990; Cortes et al. 2008; Huang et al. 2006; He et al. 2014), as it is one of the main sources of model performance deterioration (Candela et al. 2009). Furthermore, data scientists in machine learning competitions claim that finding the train/validation split that better resembles the test (evaluation) distribution is paramount to winning a Kaggle competition (Guschin et al. 2018).

However, despite the fact that a shift in data distribution can be a source of model deterioration, the two are not identical. Indeed, if we shift a random noise feature we have caused a change in the data distribution, but we should not expect the performance of a model to decline when evaluated on this shifted dataset. Thus, we emphasize here that our focus is on *model deterioration* and not distribution shift, despite the correlation between the two.

Established ways of monitoring distribution shift when the real target distribution is not available are based on statistical changes either the input data (Diethe et al. 2018; Rabanser, Günnemann, and Lipton 2019) or on the model output (Garg et al. 2021). These statistical tests correctly detect univariate changes in the distribution but are completely independent of the model performance and can therefore be too sensitive, indicating a change in the covariates but without any degradation in the model performance. This can result in false positives, leading to unnecessary model retraining. It is worth noting that several authors have stated the clear need to identify how non-stationary environments affect the behavior of models (Diethe et al. 2018).

see <https://edition.cnn.com/2021/11/09/tech/zillow-ibuying-home-zestimate/index.html> (Online accessed January 26, 2022).

^{*}These authors contributed equally.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹The Zillow case is an example of consequences of model performance degradation in an unsupervised monitoring scenario,

Aside from merely indicating that a model has deteriorated, it can in some circumstances be beneficial to identify the *cause* of the model deterioration by detecting and explaining the lack of knowledge in the prediction of a model. Such explainability techniques can provide algorithmic transparency to stakeholders and to the ML engineering team (Mougan, Kanellos, and Gottron 2021; Bhatt et al. 2021; Koh and Liang 2017; Ribeiro, Singh, and Guestrin 2016; Sundararajan, Taly, and Yan 2017).

This paper’s primary focus is on non-deep learning models and small to medium-sized tabular datasets, a size of data that is very common in the average industry, where, non-deep learning-based models achieve state-of-the-art results (Grin-sztajn, Oyallon, and Varoquaux 2022; Borisov et al. 2022; Elsayed et al. 2021).

Our contributions are the following:

1. We develop a novel method that produces prediction intervals using bootstrapping with theoretical guarantees, which achieves better coverage than previous methods on eight real-life regression datasets from the UCI repository (Dua and Graff 2017).
2. We use this non-parametric uncertainty estimation method to develop a machine learning monitoring system for regression models, which outperforms previous monitoring methods in terms of detecting deterioration of model performance.
3. We use explainable AI techniques to identify the source of model deterioration for both entire distributions as a whole as well as for individual samples, where classical statistical indicators can only determine distribution differences.
4. We release an open source Python package, `doubt`, which implements our uncertainty estimation method and is compatible with all `scikit-learn` models (Pedregosa et al. 2011).

Related Work

Model Monitoring

Model monitoring techniques help to detect unwanted changes in the behavior of a machine learning application in a production environment. One of the biggest challenges in model monitoring is distribution shift, which is also one of the main sources of model degradation (Candela et al. 2009; Diethe et al. 2018).

Diverse types of model monitoring scenarios require different supervision techniques. We can distinguish two main groups: Supervised learning and unsupervised learning. Supervised learning is the appealing one from a monitoring perspective, where performance metrics can easily be tracked. Whilst attractive, these techniques are often unfeasible as they rely either on having ground truth labeled data available or maintaining a hold-out set, which leaves the challenge of how to monitor ML models to the realm of unsupervised learning (Diethe et al. 2018). Popular unsupervised methods that are used in this respect are the Population Stability Index (PSI) and the Kolmogorov-Smirnov test (K-S), all of which measure how much the distribution of the covariates in the

new samples differs from the covariate distribution within the training samples. These methods are often limited to real-valued data, low dimensions, and require certain probabilistic assumptions (Diethe et al. 2018; Malinin et al. 2021).

Another approach suggested by Lundberg et al. (2020b) is to monitor the SHAP value contribution of input features over time together with decomposing the loss function across input features in order to identify possible bugs in the pipeline as well as distribution shift. This technique can account for previously unaccounted bugs in the machine learning production pipeline but fails to monitor the model degradation.

It is worth noting that prior work (Garg et al. 2021; Jiang et al. .) has focused on monitoring models either on out-of-distribution data or in-distribution data (Neyshabur et al. 2017, 2019). Such a task, even if challenging, does not accurately represent the different types of data a model encounters in the wild. In a production environment, a model can encounter previously seen data (training data), unseen data with the same distribution (test data), and statistically new and unseen data (out-of-distribution data). That is why we focus our work on finding an unsupervised estimator that replicates the behavior of the model performance.

The idea of mixing uncertainty with dataset shift was introduced by Ovadia et al. (2019). Our work differs from theirs, in that they evaluate uncertainty by shifting the distributions of their dataset, where we aim to detect model deterioration under dataset shift using uncertainty estimation. Their work is also focused on deep learning classification problems, while we estimate uncertainty using model agnostic regression techniques. Further, our contribution allows us to pinpoint the features/dimensions that are main causes of the model degradation.

Garg et al. (2021) introduces a monitoring system for classification models, based on imposing thresholds on the softmax values of the model. Our method differs from theirs in that we work with regression models and not classification models, and that our method utilizes external uncertainty estimation methods, rather than relying on the model’s own “confidence” (i.e., the outputted logits and associated softmax values).

Rabanser, Günnemann, and Lipton (2019), presents a comprehensive empirical investigation of dataset shift, examining how dimensionality reduction and two-sample testing might be combined to produce a practical pipeline for detecting distribution shift in a real-life machine learning system. They show that the two-sample-testing-based approach performs best. This serves as a baseline comparison within our models, even if their idea is more focused on binary classification, whereas our works focus on building a regression indicator.

Uncertainty

Uncertainty estimation is being developed at a fast pace. Model averaging (Kumar and Srivastava 2012; Gal and Ghahramani 2016; Lakshminarayanan, Pritzel, and Blundell 2017; Arnez et al. 2020) has emerged as the most common approach to uncertainty estimation. Ensemble and sampling-based uncertainty estimates have been successfully applied to many use cases such as detecting misclassifications (Ren et al. 2019), out-of-distribution inputs (D’Angelo and Hen-

ning 2021), adversarial attacks (Carlini and Wagner 2017; Smith and Gal 2018), automatic language assessments (Malinin 2019) and active learning (Kirsch, Van Amersfoort, and Gal 2019). In our work, we apply uncertainty to detect and explain model performance for seen data (train), unseen and identically distributed data (test), and statistically new and unseen data (out-of-distribution).

Kumar and Srivastava (2012) introduced a non-parametric method to compute prediction intervals for any ML model using a bootstrap estimate, with theoretical guarantees. Our work is an extension of their work, where we take into account the model’s variance in the construction of the prediction intervals. The result, as we will see in the experiments section, is that such intervals have better coverage in such high-variance scenarios.

Barber et al. (2021) recently introduced a new non-parametric method of creating prediction intervals, using the Jackknife+. Our method differs from theirs in that we are using general bootstrapped samples for our estimates, rather than leave-one-out estimates. In the experimental, we will see that the two methods perform similarly, but that our method is again more accurate in a high-variance scenario.

Methodology

Evaluation of Deterioration Detection Systems

The problem we are tackling in this paper is evaluating and accounting for model predictive performance deterioration. To do this, we simulate a distribution shift scenario in which we *have* access to the true labels, which we can use to measure the model deterioration and thus evaluate the monitoring system. A naive simulation in which we simply manually shift a chosen feature of a dataset would not be representative, as the associated true labels could have changed if such a shift happened “in the wild”.

Therefore, we propose the following alternative approach. Starting from a real-life dataset $\mathcal{D}$ and a numerical feature F of $\mathcal{D}$, we sort the data samples of $\mathcal{D}$ by the value of F , and split the sorted $\mathcal{D}$ in three equally sized sections: $\{\mathcal{D}_{\text{below}}, \mathcal{D}_{tr}, \mathcal{D}_{\text{upper}}\} \subseteq \mathcal{D}$. The model is then fitted to the middle section ($\mathcal{D}_{tr}$) and evaluated on all of $\mathcal{D}$. The goal of the monitoring system is to input the model, the labeled data segment $\mathcal{D}_{tr}$ and a sample of unlabelled data $\mathcal{S} \subseteq \mathcal{D}$, and output a “monitoring value” which behaves like the model’s performance on $\mathcal{S}$. Such a prediction will thus have to take into account the training performance, generalization performance as well as the out-of-distribution performance of the model.

In the experimental section, we compare our monitoring technique to several other such systems. To enable comparison between the different monitoring systems, we standardize all monitoring values as well as the performance metrics of the model. From these standardized values, we can now directly measure the goodness-of-fit of the model monitoring system by computing the absolute difference between its (standardized) monitoring values and the (standardized) ground truth model performance metrics. Our chosen evaluation method is very similar to the one used by Garg et al. (2021). They focus on classification models and their sys-

tems output estimates of the model’s accuracy on the dataset. They evaluate these systems by computing the absolute difference between the system’s accuracy estimate and the actual accuracy that the model achieves on the dataset.

As we are working with regression models in this paper, we will only operate with a single model performance metric: mean squared error. We will introduce our monitoring system, which is based on an uncertainty measure, and will compare our monitoring system against statistical tests based on input data or prediction data. In that section, we will also compare our uncertainty estimation method to current state-of-art uncertainty estimation methods.

Uncertainty Estimation

In order to estimate uncertainty in a general way for all machine learning models, we use a non-parametric regression technique, which is an improvement of the technique introduced by (Kumar and Srivastava 2012). This method aims at determining prediction intervals for outputs of general non-parametric regression models using bootstrap methods.

Setting $d \in \mathbb{N}$ to be the dimension of the feature space, we assume that the true model $y: \mathbb{R}^d \rightarrow \mathbb{R}$ is of the form $y(x) = \delta(x) + \varepsilon(x)$, where $\delta: \mathbb{R}^d \rightarrow \mathbb{R}$ is a deterministic and continuously differentiable function, and the **observation noise** $\varepsilon: \mathbb{R}^d \rightarrow \mathbb{R}$ is a uniform random field such that $\varepsilon(x_1), \dots, \varepsilon(x_t)$ are iid for any $x_1, \dots, x_t \in \mathbb{R}^d$, have zero mean and finite variance. We will assume that we have a data sample X of size N , as well as a *convergent estimator* $\hat{\delta}^{(n)}$ of δ , meaning the following:

Definition 1 Let $\hat{\delta}^{(n)}: \mathbb{R}^d \rightarrow \mathbb{R}$ be a function for every $n \in \mathbb{N}$. We then say that $\hat{\delta}^{(n)}$ is a **convergent estimator** of a function $\delta: \mathbb{R}^d \rightarrow \mathbb{R}$ if:

1. $\hat{\delta}^{(n)}$ is deterministic and continuous, for all $n \in \mathbb{N}$.
2. There is a function $\hat{\delta}: \mathbb{R}^d \rightarrow \mathbb{R}$ such that $\hat{\delta}^{(n)}$ converges pointwise to $\hat{\delta}$ as $n \rightarrow \infty$.

We define an associated **bias function** $\beta(x) := \delta(x) - \hat{\delta}(x)$. Note that in Kumar and Srivastava (2012) they assumed that $\mathbb{E}[(\hat{\delta}^{(n)}(x) - \delta(x))^2] \rightarrow 0$ for $n \rightarrow \infty$, effectively meaning that the candidate model would be able to perfectly model the underlying distribution given enough data. It turns out that their method does not require this assumption, as we will see below. Aside from removing this assumption, the primary difference between our approach and Kumar and Srivastava (2012) is that our approach extends the latter by maintaining good coverage in a high-variance situation, as we will also see below. We start by rewriting the equation for the true model as follows:

$$y(x) = \delta(x) + \varepsilon(x) \quad (1)$$

$$= \hat{\delta}^{(N)}(x) + \beta(x) + \varepsilon_v^{(N)}(x) + \varepsilon(x), \quad (2)$$

where $\varepsilon_v^{(N)}(x) := \delta(x) - \beta(x) - \hat{\delta}^{(N)}(x)$ is the **model variance noise**. Note that

$$\varepsilon_v^{(n)}(x) = \hat{\delta}(x) - \hat{\delta}^{(n)}(x) \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (3)$$

To produce correct prediction intervals we thus need to estimate the distribution of the observation noise, bias and model variance noise.

Estimating Model Variance Noise To estimate the model variance noise term $\varepsilon_v^{(N)}(x)$ we adapt the technique in Kumar and Srivastava (2012) to our scenario, using a bootstrap estimate. Concretely, we bootstrap our dataset $B > 0$ times, fitting our model on each of the bootstrapped samples X_b and generating bootstrapped estimates $\bar{\delta}_b^{(N)}(x)$ for every $b < B$. Centering the bootstrapped predictions as $m_b^{(N)}(x) := \mathbb{E}_b[\bar{\delta}_b^{(N)}] - \bar{\delta}_b^{(N)}$, we have that

$$\mathbb{D}_b[m_b^{(N)}(x)] = \mathbb{D}_b[\mathbb{E}_b[\bar{\delta}_b^{(N)}(x)] - \bar{\delta}_b^{(N)}(x)] \quad (4)$$

$$\rightarrow \mathbb{D}_X[\hat{\delta}(x) - \bar{\delta}^{(N)}(x)] \quad (5)$$

$$= \mathbb{D}_X[\varepsilon_v^{(N)}(x)] \quad (6)$$

as $B \rightarrow \infty$, giving us our estimate of the model variance noise.

Estimating Bias and Observation Noise We next have to estimate the bias $\beta(x)$ and the observation noise $\varepsilon(x)$. By rewriting (1) we get that that

$$\beta(x) + \varepsilon(x) = y(x) - \hat{\delta}^{(N)}(x) - \varepsilon_v^{(N)}(x), \quad (7)$$

so since we already have an estimate for $\varepsilon_v^{(N)}(x)$, it remains to estimate the residual $y(x) - \hat{\delta}^{(N)}(x)$. In Kumar and Srivastava (2012) this was estimated purely using the training residuals without using any bootstrapping, whereas our approach will estimate the expected value of this residual via a bootstrap estimate, by using bootstrapped validation residuals $y(x) - \bar{\delta}_b^{(N)}(x)$, where x is *not* in the b 'th bootstrap sample X_b . Concretely, we have that

$$\mathbb{D}_{(b, x \in X \setminus X_b)}[y(x) - \bar{\delta}_b^{(N)}(x)] \quad (8)$$

$$\rightarrow \mathbb{D}_{(X, x \notin X)}[y(x) - \hat{\delta}^{(N)}(x)] \quad (9)$$

as $B \rightarrow \infty$. An initial estimate is thus

$$\mathbb{D}_{(X, x \notin X)}[\beta(x) + \varepsilon(x)] \quad (10)$$

$$\approx \mathbb{D}_{(b, x \in X \setminus X_b)}[y(x) - \bar{\delta}_b^{(N)}(x) - m_b^{(N)}(x)] \quad (11)$$

Denote (11) by $\text{valError}_b^{(N)}$. The problem with this approach is that the resulting prediction intervals arising from these validation errors are going to be too wide, as the bootstrap samples only contain on average 63.2% of the samples in the original dataset (Friedman et al. 2001), causing the model to have artificially large validation residuals. To fix this, we follow the approach in Friedman et al. (2001) and use the 0.632+ bootstrap estimate instead, defined as follows. We start by defining the **no-information error rate**

$$\text{noInfoError}^{(N)} := \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (y(x_i) - \hat{\delta}(x_j))^2, \quad (12)$$

corresponding to the mean-squared error if the inputs and outputs were independent. Next, define the associated training residuals $\text{trainError}_b^{(N)}$ as:

$$\mathbb{D}_{(b, x \in X_b)}[y(x) - \bar{\delta}_b^{(N)}(x) - (\mathbb{E}_b[\bar{\delta}_b^{(N)}(x)] - \bar{\delta}_b^{(N)}(x))]. \quad (13)$$

Combining these two, we set the **relative overfitting rate** $\text{overfittingRate}_b^{(N)}$ to be:

$$\frac{\text{valError}_b^{(N)} - \text{trainError}_b^{(N)}}{\text{noInfoError} - \text{trainError}_b^{(N)}}. \quad (14)$$

This gives us a convenient number between 0 and 1, denoting how much our model is overfitting the dataset. From this, we define the **validation weight** $\text{valWeight}_b^{(N)}$ as:

$$\frac{0.632}{1 - (1 - 0.632) \times \text{overfittingRate}_b^{(N)}}, \quad (15)$$

which denotes how much we should weigh the validation error over the training error. In case of no overfitting, we get that $\text{valWeight}_b^{(N)} = 0.632$ and this reduces to the standard 0.632 bootstrap estimate (Friedman et al. 2001), whereas in case of severe overfitting the weight becomes 1 and thus only prioritizes the validation error.

Our final estimate of $\beta(x) + \varepsilon(x)$ is thus

$$\mathbb{D}_X[\beta(x) + \varepsilon(x)] \approx \mathbb{D}_b[o_b^{(N)}], \quad (16)$$

where

$$o_b^{(N)} := (1 - \text{valWeight}_b^{(N)}) \times \text{trainError}_b^{(N)} + \text{valWeight}_b^{(N)} \times \text{valError}_b^{(N)}.$$

Note that this estimate is only an aggregate and is not specific to any specific value of x , as opposed to the model variance estimate in equation (4).

Prediction Interval Construction Calculating the estimate of the prediction interval is then a matter of joining the results from the section of model variance noise and bias observation noise, in the same way as in Kumar and Srivastava (2012). As the estimate of $\beta(x) + \varepsilon(x)$ does not depend on any new sample, we can pre-compute this in advance by bootstrapping B samples X_b , fit our model to each and calculate the $o_b^{(N)}$ using Equation (16). Now, given a new data point x_0 and $\alpha \in (0, 1)$, we can estimate an α prediction interval around x_0 as follows. We again bootstrap B samples X_b , fit our model to each and calculate the $m_b^{(N)}(x_0)$ values. Next, we form the set $C^{(N)}(x_0) := \{m_b^{(N)}(x_0) + o_b^{(N)} \mid b < B\}$, and our interval is then (start, end), where

$$\text{start} := \hat{\delta}^{(N)}(x_0) - q_{\frac{\alpha}{2}}(C^{(N)}(x_0)) \quad (17)$$

$$\text{end} := \hat{\delta}^{(N)}(x_0) + q_{1-\frac{\alpha}{2}}(C^{(N)}(x_0)), \quad (18)$$

with $q_{\xi}(C^{(N)}(x_0))$ being the ξ 'th quantile of $C^{(N)}(x_0)$.

Detecting the Source of Uncertainty/Model Deterioration

Using uncertainty as a method to monitor the performance of an ML model does not provide any information on *what* features are the cause of the model degradation, only a goodness-of-fit to the model performance. We propose to solve this issue with the use of Shapley values.

We start by fitting a model f_θ to the training data, X^{train} . We next shift the test data by five standard deviations (call the shifted data X^{ood}) and compute uncertainty estimates Z of f_θ on X^{ood} . We next fit a second model g_ψ on (X^{ood}) to predict the uncertainty estimate Z , and compute the associated Shapley values (Lundberg et al. 2020b) of g_ψ . These Shapley values thus signify which features are the ones contributing the most to the uncertainty values. With the correlation between uncertainty values and model deterioration that we hope to conclude from the experiment described in the experimental section, this thus also provides us with a plausible cause of the model deterioration, if deterioration has taken place. Particularly, this methodology can be extended to large-scale datasets and deep learning-based models.

Experiments

Our experiments have been organized into three main groups: Firstly, we compare our non-parametric bootstrapped estimation method with the previous state-of-the-art, Kumar and Srivastava (2012) and Barber et al. (2021). Secondly, we assess the performance of our proposed uncertainty method for monitoring the performance of a machine learning model. And then, we evaluate the usability of the explainable uncertainty for identifying the features that are driving model degradation in local and global scenarios. In the main body of the paper, we present the results over several real-world datasets in the appendix we provide the experiments on synthetic datasets that exhibits, non-linear and linear behavior.

Uncertainty Method Comparison

To demonstrate the accuracy of our prediction intervals introduced in the uncertainty estimation section, we compare the coverage of the intervals with the NASA method from Kumar and Srivastava (2012) on eight regression datasets from the UCI repository (Dua and Graff 2017). The statistics of these datasets can be seen in Table 1.

Dataset	# Samples	# Features
Airfoil Self-Noise	1,503	5
Bike Sharing	17,379	16
Concrete Strength	1,030	8
QSAR Fish Toxicity	908	6
Forest Fires	517	12
Parkinsons	5,875	22
Power Plant	9,568	4
Protein	45,730	9

Table 1: Statistics of the regression datasets used in this paper.

We split each of the eight datasets into a 90/10 train/test split, uniformly at random. Next, we fit a linear regression, a decision tree, and a gradient boosting decision tree on the training split. We chose these three models to have an example of a model with large bias (the linear regression model), a model with large variance (the decision tree model), and an intermediate model that achieves state-of-the-art performance in many tasks, the gradient boosting model. We will use the `xgboost` (Chen and Guestrin 2016)

implementation of the gradient boosting model. After fitting the three models we compute α -prediction intervals for $\alpha \in \{0.75, 0.76, \dots, 0.99\}$, using our “Doubt” prediction intervals, the “NASA” prediction intervals from (Kumar and Srivastava 2012) as well as the “MAPIE” prediction intervals from (Barber et al. 2021), the latter implemented with the `MAPIE` package². We can then compare the coverage of the three methods on the eight test sets.

As the goal of an α prediction interval is to have a coverage of α , we can measure the performance of a prediction interval system by reporting the absolute difference between the actual coverage of the interval and this ideal coverage α . In Table 2 we report the mean and standard deviations of these absolute differences, for each of the three model architectures. We have performed pairwise two-tailed paired t-tests on all absolute differences, and the best-performing prediction interval methods are marked in bold for each model architecture.

We see (cf. Table 2) that there is no significant difference between the three methods in the high bias case with the linear regression model. In the case of the XGBoost model, a model with higher variance, both the MAPIE and Doubt methods outperform the NASA method, but there is no significant difference between the MAPIE method and the Doubt method in this case. In the high-variance scenario with the decision tree, however, the Doubt intervals achieve significantly better coverage than both of the other two methods.

Evaluating Model Deterioration

The scenario we are addressing is characterized by regression data sets that have statistically seen data (train data), iid statistically unseen data (test data), and out-of-distribution data. Following the open data for reproducible research guidelines described in Arnold et al. (2019) and for measuring the performance of the proposed methods, we have used eight open-source datasets (cf. Table 1) for an empirical comparison coming from the UCI repository (Dua and Graff 2017). As described in the methodology, in order to benchmark our algorithm we, for each feature F in each dataset $\mathcal{D}$, sort $\mathcal{D}$ according to F and split $\mathcal{D}$ into three equally sized sections $\{\mathcal{D}_{\text{below}}, \mathcal{D}_{\text{tr}}, \mathcal{D}_{\text{upper}}\} \subseteq \mathcal{D}$. We then train the model on $\mathcal{D}_{\text{tr}}$ and test the performance of all of $\mathcal{D}$. In this way we obtain a mixture of train, test, and out-of-distribution data, allowing us to evaluate our monitoring techniques in all three scenarios.

In evaluating a monitoring system we need to make a concrete choice of the sampling method to get the unlabelled data $\mathcal{S} \subseteq \mathcal{D}$. We are here using a rolling window of fifty samples, which has the added benefit of giving insight into the performance of the monitoring system on each of the three sections $\mathcal{D}_{\text{lower}}, \mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{upper}}$ (cf. Figure 1).

We compare our monitoring system using the uncertainty estimation method against: (i) two classical statistical methods on input data: the Kolmogorov-Smirnov test statistic (K-S) and the Population-Stability Index (PSI) (Diethe et al. 2018), (ii) a Kolmogorov-Smirnov statistical test on the predictions between train and test (Garg et al. 2021) that we

²<https://github.com/scikit-learn-contrib/MAPIE>

Model	Linear Regression	XGBoost	Decision Tree
NASA	3.854 ± 5.380	20.216 ± 12.405	20.669 ± 9.771
MAPIE	4.207 ± 4.755	5.264 ± 4.258	7.788 ± 4.782
Doubt	3.917 ± 4.870	4.861 ± 3.938	5.137 ± 3.984

Table 2: A comparison of different prediction interval methods, where the metric used is the mean absolute deviation from the ideal coverage (lower is better), with its associated standard deviation. Here NASA is the method described in Kumar and Srivastava (2012), MAPIE is the method from Barber et al. (2021) and Doubt is our method. The best results for each model architecture are shown in bold.



Figure 1: Comparison of different model degradation detection methods for the Fish Toxicity dataset. Each of the plots represents an independent experiment where each of the six features has been shifted, using the method described in the methodology section. Doubt achieves a better goodness-of-fit than previous statistical methods. A larger version of this figure can be found in Appendix.

denominate prediction shift and (iii) the previous state-of-the-art uncertainty estimate MAPIE. We evaluate the monitoring systems on a variety of model architectures: generalized linear models, tree-based models as well as neural networks.

The average performance across all datasets can be found in Table 3.³ From these we can see that our methods outperform K-S and PSI in all cases except for the Random Forest case, where our method is still on par with the best method, in that case, K-S. We have included a table with each dataset and all the estimators in the appendix, where it can be seen that both K-S and PSI easily identify a shift in the distribution but fail to detect when the model performance degrades, giving too many false positives.

Detecting the Source of Uncertainty

For this experiment, we make use of two datasets: a synthetic one (see the appendix) and the popular House Prices regression dataset⁴, where the goal is to predict the selling price of a given property. We select two of the features that are the most correlated with the target, GrLivArea and TotalBsmtSF, and also create a new feature of random noise, to have an example of a feature with minimum correlation with the target. A model deterioration system should

³See the appendix for a more detailed table.

⁴<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

therefore highlight the GrLivArea and TotalBsmtSF features, and *not* highlight the random features.

Concretely, we compute an estimation of the Shapley values using TreeSHAP (Lundberg et al. 2020b), which is an efficient estimation approach values for tree-based models, that allows for this second model to identify the features that are the source of the uncertainty, and thus also provide an indicator for what features might be causing the model deterioration.

We fitted an MLP on the training dataset, which achieved a R^2 value of 0.79 on the validation set. We then shifted all three features by five standard deviations and trained a gradient boosting model on the uncertainty values of the MLP on the validation set, which achieves a good fit (an R^2 value of 0.94 on the hold-out set of the validation). We then compare the SHAP values of the gradient boosting model with the PSI and K-S statistics for the individual features.

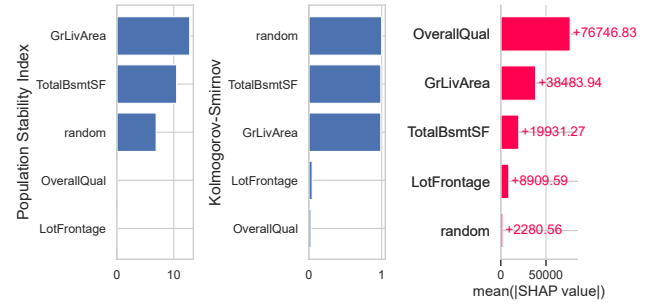


Figure 2: Global comparison of different distribution shift detection methods. Statistical methods correctly indicate that there exists a distribution shift in the shifted data. Shapley values indicate the contribution of each feature to the drop in predictive performance of the model.

In Figure 2, classical statistics and SHAP values to detect the source of the model deterioration are compared. We see that the PSI and K-S value correctly capture the shift in each of the three features (including the random noise). On the other hand, our SHAP method highlights the two substantial features (GrLivArea and TotalBsmtSF) and correctly does not assign a large value to the random feature, despite the distribution shift.

Figure 3 shows features contributing to pushing the model output from the base value to the model output. Features pushing the uncertainty prediction higher are shown in red, and those pushing the uncertainty prediction lower are in

Method	Linear Reg.	Poisson	Decision Tree	Random Forest	Gradient Boosting	MLP
PSI	0.87 ± 0.08	0.93 ± 0.08	0.97 ± 0.10	0.95 ± 0.08	0.95 ± 0.08	0.84 ± 0.16
K-S	0.81 ± 0.10	0.94 ± 0.20	0.52 ± 0.12	0.50 ± 0.12	0.61 ± 0.19	0.72 ± 0.22
PredictionShift	0.86 ± 0.13	1.00 ± 0.15	0.80 ± 0.14	0.73 ± 0.18	0.75 ± 0.20	0.74 ± 0.22
MAPIE	0.77 ± 0.10	0.83 ± 0.18	0.60 ± 0.16	0.86 ± 0.15	0.73 ± 0.18	0.74 ± 0.38
Doubt	0.71 ± 0.14	0.79 ± 0.14	0.49 ± 0.10	0.74 ± 0.18	0.58 ± 0.23	0.68 ± 0.38

Table 3: Performance of model monitoring systems for model deterioration for a variety of model architectures on eight regression datasets from the UCI repository (Dua and Graff 2017). The scores are the means and standard deviations of the absolute deviation from the true labels on $\mathcal{D}_{lower}$ and $\mathcal{D}_{upper}$ (lower is better). K-S and PSI are the monitoring systems obtained by computing the Kolmogorov-Smirnov test values and the Population Stability Index, respectively, Prediction Shift is the statistical comparison of the model prediction, and Doubt is our method. The best results for each model architecture are shown in bold. See the Appendix for all the raw scores.

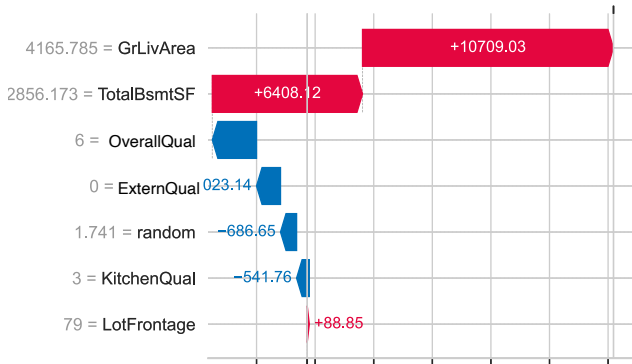


Figure 3: Individual explanation that displays the source of uncertainty for one instance. The previous method allowed only for comparison between distributions, now with explainable uncertainty, we are able to account for individual instances. In red, features pushing the uncertainty prediction higher are shown; in blue, those pushing the uncertainty prediction lower.

blue (Lundberg et al. 2018, 2020a; Lundberg and Lee 2017). From these values, we can, at a local level, also identify the two features (GrLivArea and TotalBsmtSF) causing the model deterioration in this case.

Conclusion

In this work, we have provided methods and experiments to monitor and identify machine learning model deterioration via non-parametric bootstrapped uncertainty estimation methods, and use explainability techniques to explain the source of the model deterioration.

Our monitoring system is based on a novel uncertainty estimation method, which produces prediction intervals with theoretical guarantees and which achieves better coverage than the current state-of-the-art. The resulting monitoring system more accurately detects model deterioration than methods using classical statistics. Finally, we used SHAP values in conjunction with these uncertainty estimates to identify the features that are driving the model deterioration at both a global and local level, and qualitatively showed that these more accurately detect the source of the model deterioration

compared to classical statistical methods.

Limitations: We emphasize here that due to computational limitations, we have only benchmarked on datasets of relatively small to medium size (cf. Table 1), and further work needs to be done to see if these results are also valid for datasets of significantly larger size. This work also focused on tabular data and non-deep learning models.

Reproducibility Statement

To ensure reproducibility of our results, we make the data, data preparation routines, code repositories, and methods publicly available⁵. Our novel uncertainty methods are included in the open-source Python package *doubt*⁶. We describe the system requirements and software dependencies of our experiments. Our experiments were run on an 8 vCPU server with 60 GB RAM.

Acknowledgements

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

References

Arnez, F.; Espinoza, H.; Radermacher, A.; and Terrier, F. 2020. A Comparison of Uncertainty Estimation Approaches in Deep Learning Components for Autonomous Vehicle Applications. In Espinoza, H.; McDermid, J. A.; Huang, X.; Castillo-Effen, M.; Chen, X. C.; Hernández-Orallo, J.; hÉigeartaigh, S. Ó.; and Mallah, R., eds., *Proceedings of the Workshop on Artificial Intelligence Safety 2020 co-located with the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence (IJCAI-PRICAI 2020)*, Yokohama, Japan, January, 2021, volume 2640 of *CEUR Workshop Proceedings*. CEUR-WS.org.

⁵<https://github.com/cmougan/MonitoringUncertainty>

⁶<https://github.com/saattrupdan/doubt>.

- Arnold, B.; Bowler, L.; Gibson, S.; Herterich, P.; Higman, R.; Krystalli, A.; Morley, A.; O'Reilly, M.; Whitaker, K.; et al. 2019. The turing Way: a handbook for reproducible data science. *Zenodo*.
- Barber, R. F.; Candes, E. J.; Ramdas, A.; and Tibshirani, R. J. 2021. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1): 486–507.
- Bhatt, U.; Antorán, J.; Zhang, Y.; Liao, Q. V.; Sattigeri, P.; Fogliato, R.; Melançon, G.; Krishnan, R.; Stanley, J.; Tickoo, O.; et al. 2021. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 401–413.
- Borisov, V.; Leemann, T.; Sebler, K.; Haug, J.; Pawelczyk, M.; and Kasneci, G. 2022. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.
- Candela, J. Q.; Sugiyama, M.; Schwaighofer, A.; and Lawrence, N. D. 2009. Dataset shift in machine learning. *The MIT Press*, 1: 5.
- Carlini, N.; and Wagner, D. 2017. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 3–14.
- Chen, T.; and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.
- Cortes, C.; Mohri, M.; Riley, M.; and Rostamizadeh, A. 2008. Sample selection bias correction theory. In *International conference on algorithmic learning theory*, 38–53. Springer.
- D'Angelo, F.; and Henning, C. 2021. Uncertainty-based out-of-distribution detection requires suitable function space priors. *arXiv preprint arXiv:2110.06020*.
- Diethe, T.; Borchert, T.; Thereska, E.; Balle, B.; and Lawrence, N. 2018. Continual Learning in Practice. In *Continual Learning Workshop at NeurIPS 2018*.
- Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, <http://archive.ics.uci.edu/ml>.
- Elsayed, S.; Thyssens, D.; Rashed, A.; Jomaa, H. S.; and Schmidt-Thieme, L. 2021. Do we really need deep learning models for time series forecasting? *arXiv preprint arXiv:2101.02118*.
- Friedman, J.; Hastie, T.; Tibshirani, R.; et al. 2001. *The elements of statistical learning*, volume 1.10. Springer series in statistics New York.
- Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.
- Garg, S.; Balakrishnan, S.; Lipton, Z. C.; Neyshabur, B.; and Sedghi, H. 2021. Leveraging Unlabeled Data to Predict Out-of-Distribution Performance. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*.
- Grinsztajn, L.; Oyallon, E.; and Varoquaux, G. 2022. Why do tree-based models still outperform deep learning on typical tabular data? In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Guschin, A.; Ulyanov, D.; Trofimov, M.; Altukhov, D.; and Michaidilis, M. 2018. How to Win a Data Science Competition: Learn from Top Kagglers - National Research University Higher School of Economics. <https://www.coursera.org/lecture/competitive-data-science/categorical-and-ordinal-features-qu1TF>. Accessed 02/11/20.
- He, X.; Pan, J.; Jin, O.; Xu, T.; Liu, B.; Xu, T.; Shi, Y.; Atallah, A.; Herbrich, R.; Bowers, S.; et al. 2014. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, 1–9.
- Heckman, J. 1990. Varieties of selection bias. *The American Economic Review*, 80(2): 313–318.
- Huang, J.; Gretton, A.; Borgwardt, K.; Schölkopf, B.; and Smola, A. 2006. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19: 601–608.
- Jiang, Y.; Nagarajan, V.; Baek, C.; and Kolter, J. Z. . Assessing Generalization of SGD via Disagreement. In *International Conference on Learning Representations*.
- Kirsch, A.; Van Amersfoort, J.; and Gal, Y. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32: 7026–7037.
- Koh, P. W.; and Liang, P. 2017. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, 1885–1894. PMLR.
- Kumar, S.; and Srivastava, A. 2012. Bootstrap prediction intervals in non-parametric regression with applications to anomaly detection. In *Proc. 18th ACM SIGKDD Conf. Knowl. Discovery Data Mining*.
- Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *Advances in Neural Information Processing Systems*, 30.
- Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; and Lee, S.-I. 2020a. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, 2(1): 56–67.
- Lundberg, S. M.; Erion, G. G.; Chen, H.; DeGrave, A. J.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; and Lee, S. 2020b. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.*, 2(1): 56–67.
- Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, 4768–4777.
- Lundberg, S. M.; Nair, B.; Vavilala, M. S.; Horibe, M.; Eisses, M. J.; Adams, T.; Liston, D. E.; Low, D. K.-W.; Newman,

- S.-F.; Kim, J.; et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10): 749–760.
- Malinin, A. 2019. *Uncertainty estimation in deep learning with application to spoken language assessment*. Ph.D. thesis, University of Cambridge.
- Malinin, A.; Band, N.; Gal, Y.; Gales, M.; Ganshin, A.; Chesnokov, G.; Noskov, A.; Ploskonosov, A.; Prokhorenkova, L.; Provilkov, I.; et al. 2021. Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Mougan, C.; Kanellos, G.; and Gottron, T. 2021. Desiderata for Explainable AI in Statistical Production Systems of the European Central Bank. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases - International Workshops of ECML PKDD 2021, Virtual Event, September 13-17, 2021, Proceedings, Part I*, volume 1524 of *Communications in Computer and Information Science*, 575–590. Springer.
- Neyshabur, B.; Bhojanapalli, S.; Mcallester, D.; and Srebro, N. 2017. Exploring Generalization in Deep Learning. *Advances in Neural Information Processing Systems*, 30: 5947–5956.
- Neyshabur, B.; Li, Z.; Bhojanapalli, S.; LeCun, Y.; and Srebro, N. 2019. Towards Understanding the Role of Over-Parametrization in Generalization of Neural Networks. In *International Conference on Learning Representations (ICLR)*.
- Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.; Lakshminarayanan, B.; and Snoek, J. 2019. Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.
- Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12: 2825–2830.
- Rabanser, S.; Günnemann, S.; and Lipton, Z. C. 2019. Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d’Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 1394–1406.
- Ren, J.; Liu, P. J.; Fertig, E.; Snoek, J.; Poplin, R.; DePristo, M. A.; Dillon, J. V.; and Lakshminarayanan, B. 2019. Likelihood ratios for out-of-distribution detection. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 14707–14718.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. ”Why should i trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.
- Shimodaira, H. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2): 227–244.
- Smith, L.; and Gal, Y. 2018. Understanding Measures of Uncertainty for Adversarial Example Detection. In Globerson, A.; and Silva, R., eds., *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, 560–569. AUAI Press.
- Stolzenberg, R. M.; and Relles, D. A. 1997. Tools for intuition about sample selection bias and its correction. *American sociological review*, 494–507.
- Sugiyama, M.; Krauledat, M.; and Müller, K.-R. 2007. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5).
- Sugiyama, M.; and Müller, K.-R. 2005. Input-Dependent Estimation of Generalization Error under Covariate Shift. *Statistics & Decisions*, 23(4): 249–279.
- Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 3319–3328. PMLR.
- Tasche, D. 2017. Fisher consistency for prior probability shift. *The Journal of Machine Learning Research*, 18(1): 3338–3369.
- Zadrozny, B. 2004. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, 114.