# INCONVENIENT: A BENCHMARK TO ASSESS LLM GENERALIZATION CAPABILITIES FROM CONVENIENCE SAMPLES

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

## ABSTRACT

Conventional machine learning models for tabular prediction typically assume that training data are representative of the population on which they will be deployed. In practice, however, data are often collected based on convenience, availability, accessibility, or cost, resulting in systematic sampling biases that can distort feature–target relationships and undermine real-world generalization. Existing tabular distribution-shift benchmarks primarily construct out-of-distribution (OOD) settings by partitioning observations across naturally occurring domains. While such partitions may induce changes in $P(Y \mid X)$, the resulting shifts are neither explicitly controlled nor isolated, and the evaluation distributions themselves are often not representative of the broader target population. We introduce INCONVENIENT, a benchmark designed to study robustness to convenience-sample bias. Rather than relying on naturally occurring domain partitions, INCONVENIENT directly perturbs feature–target relationships in the training distribution while preserving the original population distribution as the deployment environment. This setting more closely reflects real-world scenarios in which models are developed from unrepresentative samples and subsequently deployed on a broader population. The distinction is particularly relevant for large language models (LLMs), whose pretraining may encode population-level knowledge that extends beyond the relationships observed in a biased training dataset. Unlike existing distribution-shift benchmarks, INCONVENIENT provides explicit control over both the direction and severity of relationship shifts, enabling a systematic assessment of robustness to convenience sampling. The benchmark comprises 200 tabular classification and regression datasets spanning diverse application domains and generates around $\sim 15{,}500$ shifted training distributions which lead to substantial downstream performance degradation.

## 1 INTRODUCTION

Machine learning models are typically developed under the assumption that training data provide a representative sample of the population on which the model will ultimately be deployed. In practice, however, this assumption is rarely satisfied. Data are often collected based on convenience, availability, accessibility, or cost rather than through representative sampling procedures. Medical datasets may be collected from a single hospital system, consumer datasets may overrepresent users of a particular platform, and scientific datasets may be limited to populations that are easiest to recruit. As a result, the relationships observed in the training data can differ from those that hold in the broader population.

LLMs are pretrained on massive corpora that may encode broad statistical knowledge about real-world relationships. When presented with a biased or unrepresentative dataset, an LLM-based prediction system may possess prior knowledge that conflicts with the relationships observed in the training sample.

A large body of work has studied the impact of distribution shift on machine learning systems. Benchmarks such as WILDS (Koh et al., 2021) demonstrated that modern models often struggle to generalize across naturally occurring shifts in vision, language, and graph domains. In the tab-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ular setting, TableShift (Gardner et al., 2023) and WhyShift (Liu et al., 2023) construct out-of-distribution evaluation settings by partitioning data into distinct real-world domains, such as geographic regions, institutions, or time periods. These benchmarks therefore evaluate generalization across naturally occurring population differences that induce complex combinations of covariate, label, and conditional shifts; WhyShift in particular decomposes each observed shift into its covariate and concept components and finds that concept shift dominates on many real tabular tasks. However, in these benchmarks, both the training and evaluation distributions change as a consequence of the shift. As a result, poor performance may reflect difficulties adapting to a new target population rather than an inability to recover population-level relationships obscured by biased sampling. While these settings capture important real-world deployment challenges, they leave a common form of dataset bias largely unaddressed: scenarios in which the deployment population remains fixed, but the training sample is systematically unrepresentative of that population. More recently, TabFSBench (Cheng et al., 2025) introduced a controlled benchmark that synthetically generates feature shifts through perturbations of the observed feature space. However, it does not explicitly isolate shifts in the underlying feature–target relationship, which is the primary focus of our benchmark.

To address this limitation, we introduce **INCONVENIENT**, a benchmark for evaluating robustness to convenience-sample bias. Rather than constructing train and test sets from different domains, INCONVENIENT begins with a representative population dataset and systematically constructs unrepresentative training samples that distort the relationship between features and targets. Models are then evaluated on test examples drawn from the original population. INCONVENIENT generates controlled shifts by manipulating feature–target relationships in the training data through coefficient amplification, attenuation, and sign-reversal operations. By explicitly controlling both the direction and severity of these perturbations, INCONVENIENT enables the systematic study of robustness to sampling-induced relationship shifts. It also provides control over severity of shift which allows comprehensive analysis of how sever of a shift and what kind of shift is a model more robust to handle. The benchmark contains 200 tabular prediction datasets spanning diverse application domains and yields around 15,500 distribution-shift scenarios. Across these tasks, convenience-sample training distributions induce substantial degradation in downstream predictive performance, creating a challenging testbed for evaluating robustness under sampling-induced shift.

## 2 Related Work

**Sample selection bias and convenience sampling.** The observation that non-random sampling distorts learned relationships predates machine learning. Heckman (1979) formalized selection bias in econometrics, and the learning-theoretic treatment of Zadrozny (2004) and Cortes et al. (2008) showed that selection depending on $X$ alone (the "missing at random" or covariate-shift case) is correctable by importance weighting, whereas selection that depends on $Y$ or on the $X$–$Y$ relationship is not identifiable from the biased sample.

**Distribution shift benchmarks.** WILDS (Koh et al., 2021) established the now-standard recipe for evaluating OOD generalization: partition a dataset along a naturally occurring axis such as hospital, country, or time, train on some partitions, and test on others. Tabular benchmarks have largely inherited this design. TableShift (Gardner et al., 2023) assembles tasks with domain-based splits and finds that the ID–OOD gap tracks label shift, while WhyShift (Liu et al., 2023) goes further and decomposes each observed shift into its covariate ($Y|X$-preserving) and concept ($Y|X$-changing) components, finding that concept shift dominates on many real tabular tasks. TabFSBench (Cheng et al., 2025) takes a complementary, synthetic approach and removes or adds feature columns between training and test, but it does not target the feature–target relationship itself. INCONVENIENT differs from both families on two axes. First, the deployment distribution is held fixed at the population, so degradation is attributable to the training sample alone. Second, the shift is constructed by directly manipulating the coefficient of a chosen feature so its direction and severity are known and controllable.

**LLMs and pretrained priors for tabular prediction.** A growing body of work suggests that large language models (LLMs) encode knowledge that can improve learning under distribution shift and limited data. TabLLM (Hegselmann et al., 2023) and LIFT (Dinh et al., 2022) serialize tabular instances as text and demonstrate that LLMs can outperform traditional machine learning models in

low-data regimes. A complementary line of research seeks to extract this knowledge as an explicit prior rather than using the LLM directly as a predictor. LLM Processes (Requeima et al., 2024) condition predictive distributions on natural-language descriptions of the task, AutoElicit (Capstick et al., 2025) elicits priors over linear-model coefficients from feature names alone, and LoID (Rezaeimanesh et al., 2026) extracts feature-wise Bayesian priors directly from model logits without relying on generated text. Across a variety of tabular prediction tasks, these approaches often match or exceed expert-derived priors and improve performance under data scarcity. Complementing these findings, Selby et al. (2024) evaluate the extent to which elicited quantities align with real-world ground truth relationships.

This literature provides evidence that LLMs encode population-level knowledge about relationships between semantically meaningful features and targets. However, existing evaluations largely focus on settings where the training data are limited, and do not address whether such knowledge remains useful when the observed data actively contradict those relationships. INCONVENIENT is designed to address this gap. By constructing training samples in which feature–target relationships are systematically weakened, amplified, or even reversed while preserving the original population as the evaluation distribution, the benchmark isolates scenarios where recovering the correct population relationship requires leveraging prior knowledge beyond the observed sample.

# 3 Benchmark Creation

## 3.1 Problem Statement

Consider a tabular dataset composed of examples $(x, y, s) \sim P_s$ where $x \in \mathbb{R}^d$ is the input feature vector, $y$ is the prediction target (either categorical for classification or continuous for regression), and $s \in \{0, 1\}$ is the sampling indicator denoting whether the example is drawn from a random sample ($s = 1$) or a convenience sample ($s = 0$).

We can view the overall data distribution as a mixture of two sampling distributions: the population distribution $P_{\text{pop}}$ representing random samples from the target population, and the convenience distribution $P_{\text{conv}}$ representing systematically biased samples. The test distribution is always drawn from the population: $P_{\text{test}} = P_{\text{pop}}$.

In the random-sampling setting, our goal is to learn a predictor $f_\theta : \mathbb{R}^d \to \mathcal{Y}$ that accurately predicts $y$ using examples drawn from $P_{\text{pop}}$. This represents the ideal scenario where training data is representative of the deployment population.

A distribution shift occurs because $P_{\text{conv}} \neq P_{\text{pop}}$. As a consequence of this shift, the joint distributions $p_{\text{conv}}(x, y) \neq p_{\text{pop}}(x, y)$ differ between convenience and population samples. For a fixed predictor $f_\theta$ trained on a convenience sample, we define the *convenience sample gap* as:

$$\Delta_{\text{perf}} = \text{Perf}(f_\theta^{\text{pop}}, P_{\text{test}}) - \text{Perf}(f_\theta^{\text{conv}}, P_{\text{test}}) \tag{1}$$

where $f_\theta^{\text{pop}}$ is trained on random samples from $P_{\text{pop}}$, $f_\theta^{\text{conv}}$ is trained on convenience samples from $P_{\text{conv}}$, and $\text{Perf}(\cdot)$ denotes the performance metric (AUC-ROC for classification, $R^2$ for regression). Both models are evaluated on held-out test data from $P_{\text{test}}$. The gap $\Delta_{\text{perf}}$ quantifies the performance degradation due to training on convenience samples rather than random samples.

In our setting, we assume that no information about the specific form of distribution shift is available during training – i.e., there is no knowledge of how $p(y)$, $p(y|x)$, or $p(x)$ differ between $P_{\text{conv}}$ and $P_{\text{pop}}$, and no unlabeled data from the population distribution is accessible. This reflects the realistic scenario where practitioners train models on available convenience samples without knowing the precise nature of sampling biases.

## 3.2 Dataset selection

Because the primary goal of INCONVENIENT is to evaluate methods that leverage prior knowledge encoded in large language models, we restrict our benchmark to datasets with semantically meaningful feature representations. Consequently, we exclude image and audio datasets where individual input dimensions typically lack interpretable semantic meaning and therefore provide limited opportunity for methods based on feature names, metadata, or task descriptions to exploit external knowledge.

Our benchmark comprises 200 tabular datasets spanning a diverse range of application domains. The datasets come from OpenML (Vanschoren et al., 2014), Kaggle, and TALENT (Liu et al., 2024). The collection includes 80 regression datasets, 39 binary classification datasets, and 81 multiclass classification datasets. Dataset sizes range from hundreds to tens of thousands of observations, and the number of features varies from fewer than 10 to more than 100. This diversity enables the evaluation of robustness to convenience-sample bias across a broad spectrum of prediction tasks and data regimes.

#### 3.2.1 REFERENCE MODEL

For each dataset, we first split the data into a training pool (50%) and a held-out test set (50%). We fit a reference linear model on the entire training pool to obtain baseline coefficients $\beta_{\text{ref}} = (\beta_1, \ldots, \beta_d)$, where $d$ is the number of features after preprocessing (including one-hot encoded categorical variables). For classification tasks, we use logistic regression with $C = \infty$ (no regularization), yielding coefficients in the log-odds space. For regression tasks, we use ordinary least squares linear regression.

#### 3.2.2 COEFFICIENT MANIPULATION OPERATIONS

We define three types of coefficient shifts, each creating a distinct form of distribution shift:

**Amplification.** Increase the magnitude of target feature coefficients by a multiplicative factor $\alpha > 1$. The target criterion is:

$$\beta_{\text{candidate},j} > \alpha \cdot \beta_{\text{ref},j}, \quad \forall j \in \mathcal{F}_{\text{target}} \tag{2}$$

The influence weight is $w_j(\text{amplify}) = \text{sign}(\beta_{\text{ref},j})$, prioritizing examples that strengthen the existing coefficient direction. We test amplification factors $\alpha \in \{1.15, 1.3, 1.5, 2.0, 3.0, 5.0\}$, creating convenience samples where feature-target relationships are exaggerated compared to the population.

**Attenuation.** Decrease the magnitude of target feature coefficients by a multiplicative factor $0 < \gamma < 1$. The target criterion is:

$$|\beta_{\text{candidate},j}| < \gamma \cdot |\beta_{\text{ref},j}|, \quad \forall j \in \mathcal{F}_{\text{target}} \tag{3}$$

The influence weight is $w_j(\text{attenuate}) = -\text{sign}(\beta_{\text{ref},j}) \cdot \text{sign}(x_{ij})$, prioritizing examples that weaken coefficient magnitudes. We test attenuation factors $\gamma \in \{0.2, 0.4, 0.6, 0.7, 0.9\}$, creating convenience samples where feature-target relationships are diluted or noisier than in the population.

**Sign flip.** Reverse the direction of target feature effects by flipping coefficient signs. The target criterion is:

$$\text{sign}(\beta_{\text{candidate},j}) \neq \text{sign}(\beta_{\text{ref},j}), \quad \forall j \in \mathcal{F}_{\text{target}} \tag{4}$$

The influence weight is $w_j(\text{flip}) = -\text{sign}(\beta_{\text{ref},j})$, prioritizing examples that induce sign reversals. This creates extreme distribution shifts, where convenience samples exhibit reversed relationships compared to the population (e.g., marijuana use in the head-and-neck cancer data of Rezaeimanesh et al. (2026)).

#### 3.2.3 GREEDY SUBSET SELECTION ALGORITHM

To create convenience training sets with specific coefficient shifts, we employ a greedy iterative algorithm that selects training examples to achieve a target coefficient configuration. The algorithm operates as follows:

1. **Influence scoring**: For each example $(x_i, y_i)$ in the training pool, compute an influence score $s_i$ measuring how much including that example would push the target features' coefficients in the desired direction. For classification, the influence is based on the gradient contribution weighted by prediction uncertainty:

$$s_i = \sum_{j \in \mathcal{F}_{\text{target}}} (y_i - \hat{p}_i) \cdot x_{ij} \cdot w_j(\text{direction}) \tag{5}$$

where $\hat{p}_i$ is the predicted probability under the reference model, and $w_j(\text{direction})$ encodes the desired direction of change. For regression, the influence uses residuals: $s_i = \sum_{j \in \mathcal{F}_{\text{target}}} (y_i - \hat{y}_i) \cdot x_{ij} \cdot w_j(\text{direction})$.

2. **Initial candidate set**: Sort all training pool examples by influence score in descending order. Initialize the candidate subset with the top $n_{\text{min}} = \lfloor \alpha_{\text{min}} \cdot N_{\text{pool}} \rfloor$ highest-influence examples, where $\alpha_{\text{min}}$ is the minimum fraction hyperparameter and $N_{\text{pool}}$ is the training pool size.
3. **Iterative expansion**: Fit a logistic regression (or linear regression) model on the current candidate subset and compute its coefficients $\beta_{\text{candidate}}$. Check if the target criterion is met for all features in $\mathcal{F}_{\text{target}}$. If not met, expand the candidate set by adding the next $b = n_{\text{min}}/10$ highest-influence examples and repeat. Continue until either the target is achieved or the entire pool is exhausted.
4. **Target verification**: After the final iteration, verify that the learned coefficients satisfy the shift criteria. If not achieved, raise an error indicating that the desired shift is infeasible for this dataset with the given constraints and exclude the created train set.

The greedy selection procedure induces distribution shifts across all three canonical types: (i) Covariate shift: Selecting examples based on influence scores changes $p(x)$, as certain regions of feature space are over- or under-represented in the convenience sample. (ii) Label shift: The selection process alters class/target prevalence, changing $p(y)$ relative to the population distribution. (iii) Concept shift: By explicitly manipulating learned coefficients, we directly modify the conditional relationship $p(y|x)$ between features and targets.

#### 3.2.4 Shift Severity Control

Our benchmark provides several knobs for controlling the severity of the induced relationship shift:

- **Minimum subset size ($\alpha_{min}$).** Smaller values of $\alpha_{min}$ restrict the optimization procedure to a smaller subset of the original training data. This generally produces training samples that are less representative of the underlying population.
- **Feature importance.** Perturbing relationships involving highly informative features is expected to have a larger impact on downstream predictive performance than perturbing less informative features. We estimate feature importance using the absolute value of feature–target correlation.
- **Coefficient magnitude.** For amplification and attenuation shifts, larger coefficient modifications correspond to stronger departures from the original relationship are therefore expected to result in greater performance degradation.

These knobs allow us to evaluate model robustness across a broad spectrum of convenience-sample biases and to analyze how different forms of relationship distortion affect downstream performance. Figure 1 presents an empirical study of the impact of these severity controls on the performance of the evaluated methods.

## 4 Experimental Setup

### 4.1 Evaluation

We evaluate seven methods: XGBoost (Chen & Guestrin, 2016), LightGBM (Ke et al., 2017), CatBoost (Prokhorenkova et al., 2018), Random Forest (Breiman, 2001), Multi-Layer Perceptron (MLP), and TabPFN (Hollmann et al., 2023), plus Logistic Regression for classification and Ridge Regression for regression. Our baseline suite is deliberately narrower than the one used by TableShift (Gardner et al., 2023), the closest tabular distribution-shift benchmark to ours. In Section A.1 we state explicitly which methods we omit and why.

For each dataset, we perform the following evaluation procedure:

1. Split data into training (50%), and test (50%) sets

2. Generate convenience training variants using the perturbation strategies described above
3. For each method:
   - Train on in-distribution training set, tune hyperparameters on in-distribution validation set (20% of the original in-distribution training set), evaluate on test set
   - Train on convenience training set, tune hyperparameters on convenience validation set (20% of the original OOD training set), evaluate on test set
4. Compute performance metrics: AUC-ROC for classification (macro-averaged for multi-class), $R^2$ for regression
5. Calculate $\Delta_{perf}$
6. Remove convenience sets where $\Delta_{perf} < 0.1$ or $(f_\theta^{\text{conv}}, P_{\text{test}}) < -1$ or $(f_\theta^{\text{conv}}, P_{\text{test}}) > 1$.

Furthermore, we remove coefficient-manipulation sets whose induced downstream performance degradation is already subsumed by previously constructed OOD sets. This situation can arise when increasing the magnitude of a feature shift fails to produce additional degradation beyond that observed in an existing OOD set. Because such sets do not expose new model failure modes or provide additional discriminatory power for evaluating robustness, retaining them would introduce redundancy without increasing the benchmark's informational value.

### 4.2 Measure of collateral effect

We measure the collateral effect on non-target coefficients using relative L2 change. Let $\beta_{\text{pop}}$ and $\beta_{\text{conv}}$ denote the coefficient vectors obtained by fitting logistic/ridge regression on the in-distribution and convenience training sets, respectively. For a given shift targeting feature set $\mathcal{T}$, we compute the relative change in non-target coefficients as:

$$\text{Relative Change} = \frac{|\beta_{\text{conv}}[\mathcal{T}^c] - \beta_{\text{pop}}[\mathcal{T}^c]|_2}{|\beta_{\text{pop}}[\mathcal{T}^c]|_2} \tag{6}$$

where $\mathcal{T}^c$ denotes the complement of the target feature set. This metric quantifies how much the non-target coefficients change as a fraction of their original magnitude, providing a scale-invariant measure of collateral coefficient drift induced by the greedy selection process.

### 4.3 Baselines

We evaluate seven machine learning methods representing diverse model families, all tuned using the Tree-structured Parzen Estimator (TPE) algorithm (Bergstra et al., 2011) implemented in HyperOpt (Bergstra et al., 2013) with a search budget of 20 evaluations per method-dataset combination. Hyperparameter selection is performed on the convenience training set's 80/20 train-validation split, reflecting the realistic constraint that practitioners have no access to the population distribution during model selection.

**Logistic Regression / Ridge Regression**: Linear models with regularization. For classification, we use logistic regression; for regression, Ridge regression. Search space: $C \in [0.001, 100]$ (log-uniform) for logistic regression, $\alpha \in [0.001, 1000]$ (log-uniform) for Ridge; penalty type $\in \{\text{L1}, \text{L2}, \text{ElasticNet}, \text{None}\}$ with corresponding L1 ratio $\in [0, 1]$ for logistic regression; solver $\in \{\text{auto}, \text{svd}, \text{cholesky}, \text{lsqr}, \text{sag}, \text{saga}\}$ for Ridge.

**Random Forest**: Ensemble of decision trees with bootstrap aggregation. Search space: number of trees $\in \{50, 100, 200, 300\}$; max depth $\in \{5, 10, 20, 30, \text{None}\}$; min samples split $\in \{2, 5, 10\}$; min samples leaf $\in \{1, 2, 4\}$; max features $\in \{\text{sqrt}, \text{log2}, \text{None}\}$.

**XGBoost**: Gradient boosted decision trees with advanced regularization. Search space: number of trees $\in \{50, 100, 200, 300\}$; max depth $\in \{3, 5, 7, 10\}$; learning rate $\in [0.01, 0.3]$ (log-uniform); subsample ratio $\in [0.6, 1.0]$; column subsample ratio $\in [0.6, 1.0]$; min child weight $\in \{1, 3, 5\}$; gamma $\in [0, 0.2]$.

**LightGBM**: Histogram-based gradient boosting with leaf-wise tree growth. Search space: number of trees $\in \{50, 100, 200, 300\}$; num leaves $\in \{15, 31, 63, 127\}$; learning rate $\in [0.01, 0.3]$ (log-

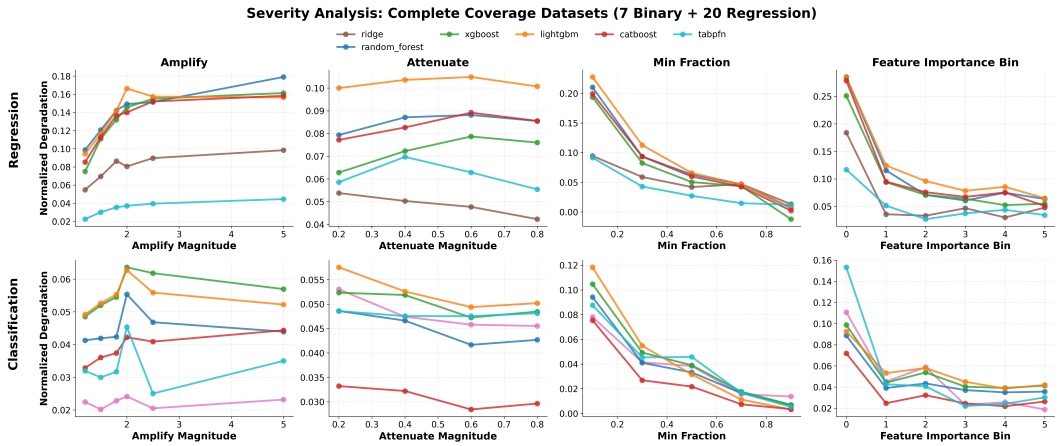



Figure 1: Performance degradation as a function of each shift-severity control. For a given severity knob, we report the average performance of each baseline across all combinations of the remaining severity controls while holding the selected knob fixed. This marginalization isolates the effect of each severity control on downstream performance. MLP and logistic regression performances are removed because of catastrophic failures that hid the performances of other models. We calculate normalized degradation as $\frac{Perf_{IID}-Perf_{OOD}}{Perf_{IID}}$.Detailed performance numbers of all models are available in §A.3

uniform); min child samples $\in \{10, 20, 30\}$; subsample ratio $\in [0.6, 1.0]$; column subsample ratio $\in [0.6, 1.0]$; L1 regularization $\in [0, 1]$; L2 regularization $\in [0, 1]$.

**CatBoost**: Gradient boosting with ordered target encoding and symmetric trees. Search space: iterations $\in \{50, 100, 200, 300\}$; depth $\in \{4, 6, 8, 10\}$; learning rate $\in [0.01, 0.3]$ (log-uniform); L2 leaf regularization $\in \{1, 3, 5, 7, 9\}$; border count $\in \{32, 64, 128\}$.

**MLP**: Multi-layer perceptron with ReLU/tanh activations and standard feature scaling. Search space: hidden layer sizes $\in \{(32), (64), (128), (64, 32), (128, 64)\}$; activation $\in \{\text{ReLU}, \tanh\}$; L2 penalty $\in [0.0001, 0.1]$ (log-uniform); learning rate schedule $\in \{\text{constant}, \text{adaptive}\}$; initial learning rate $\in [0.001, 0.1]$ (log-uniform). Maximum 500 iterations with early stopping.

**TabPFN**: Transformer-based method using in-context learning. No hyperparameter tuning—uses default configuration with 4 ensemble members..

All methods are retrained on the combined train+validation split after hyperparameter selection, then evaluated on the held-out population test set.

### 4.4 LLM-Based Coefficient Anomaly Detection

To assess whether large language models can identify unreasonable feature coefficients, we conducted an experiment where Claude-Haiku-4.5 was presented with OOD coefficients and asked to identify which features seemed most suspicious based on domain knowledge alone. We then calculated the actual coefficient changes ($|IID - OOD|$) for all features and ranked them by magnitude of change. We evaluated whether the LLM's suspicion rankings overlapped with the true rankings of coefficient change, measuring Recall@$K$ as the proportion of the true top-$K$ most-changed features that appeared in the LLM's top-$K$ suspicious features.

## 5 Results

We use INCONVENIENT to answer three questions. First, does convenience-sample bias actually break standard tabular learners, and by how much (§5.1)? Second, is the induced shift a realistic, structure-wide shift rather than a single-coefficient artifact (§5.2)? Third, do the severity knobs of §3.2.4 behave predictably enough to serve as a difficulty axis (§5.3)? Throughout, we report performance degradation $\Delta_{\text{perf}}$ as defined in Eq. 1.

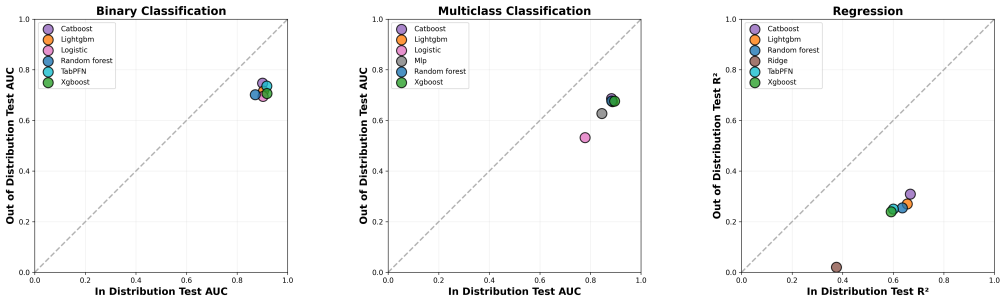



Figure 2: In-distribution versus convenience-sample performance across binary classification, multiclass classification, and regression datasets. Each point represents a baseline. The scatter plots reveal substantial performance degradation under convenience-sample bias relative to the in-distribution setting, highlighting the challenges posed by distorted feature–target relationships.

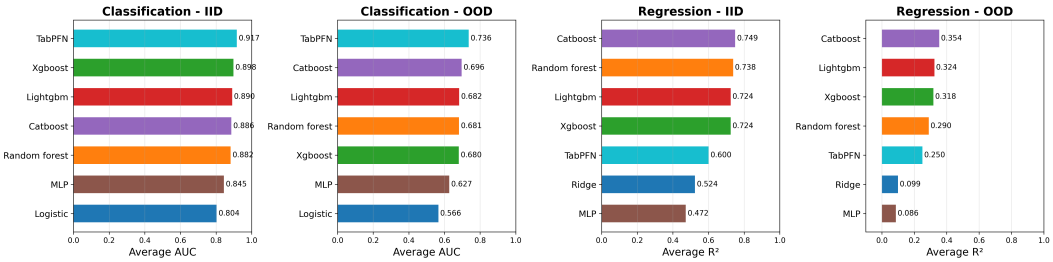



Figure 3: Reported values correspond to the average performance of each method over all the shifts, generated from the classification and regression datasets in the benchmark.

### 5.1 Convenience samples degrade every baseline

Figure 2 plots in-distribution against out-of-distribution test performance for every method and task type, averaged over all retained shifts (§4.1). Every baseline sits well below the diagonal, and the methods generally cluster. Averaged over 6,489 classification shifts and 8,970 regression shifts (Figure 3), convenience-sample training costs 0.18–0.24 AUC on classification, a 20–29% relative drop, and 0.35–0.45 $R^2$ on regression, a 53–82% relative drop.

Figure 3 breaks this down by method. The plot shows that the in-distribution ranking does not always survive the shift and lower in-distribution performance does not translate to robustness. Ridge regression and MLP have the weakest in-distribution performance ($R^2 = 0.524$ and $0.472$ respectively) yet suffer among the largest degradations ($\Delta = 0.425$ and $0.386$), falling to under 10% $R^2$. While a heavily regularized linear model has less capacity to fit complex relationships, when the training data systematically distorts a feature coefficient, Ridge still learns the distortion and fails on population data. This is the central empirical claim of INCONVENIENT. When a training sample reverses or erases a feature–target relationship, nothing in the sample identifies the error. The convenience data are internally consistent, the validation split is drawn from the same biased sample, and no unlabeled population data are available. Correction must therefore come from outside the data which is precisely the setting in which a model with pretrained knowledge can be useful.

### 5.2 Collateral effects confirm the shifts are structural

Following Equation 6, we find median relative changes of 26% (amplify), 61% (attenuate), and 55% (flip-sign) on regression tasks, and 109%, 107%, and 104% respectively on classification tasks. The perturbation propagates through the correlation structure of the feature space rather than remaining local to the targeted feature.

### 5.3 THE SEVERITY KNOBS CONTROL DIFFICULTY

Figure 1 shows the effect of varying each severity knob while marginalizing over the others. Since not all datasets produce feasible shifts at every severity level, averaging across all available shifts would result in different dataset subsets being represented at different settings, potentially biasing the comparison. Therefore, we chose seven binary classification, and 20 regression datasets that produced shifts across all shift settings for this analysis. A severity knob is useful for the benchmark if performance degradation changes monotonically with the knob value and if the resulting sweep covers a broad range of difficulty levels. We find that two of the three knobs satisfy both criteria.

**Minimum fraction.** $\alpha_{\min}$ controls how small the initial high-influence candidate set is, and therefore how far the selected subset can drift from the population before the target criterion is met. At $\alpha_{\min} = 0.1$ degradation reaches $0.09$–$0.23$ $R^2$ for regression and $0.08$–$0.12$ AUC for classification, and it decreases monotonically for every method as the fraction grows which validate $\alpha_{min}$ as a severity control.

**Feature importance.** The fourth column of Figure 1 sweeps the feature-importance bin of the targeted feature, where bin0 contains the most important features (by correlation with the target) and bin5 the least. The trend is strongly decreasing: perturbing a top-bin feature costs $0.18$–$0.29$ $R^2$ in regression and $0.07$–$0.11$ AUC in classification, while perturbing a bottom-bin feature costs less than $0.07$ in regression and $0.04$ in classification. This is the expected behavior and validates the bin as a severity control.

**Amplification and attenuation magnitude.** Trends are less consistent. For regression, degradation generally increases with amplification magnitude, but does not consistently decrease as attenuation approaches one and even increases in some cases. For classification, amplification shows no consistent trend, while degradation decreases more consistently as attenuation approaches one. Thus, this severity knob does not reliably control shift difficulty. This inconsistency likely arises from collateral effects: modifying one coefficient can alter the effective influence of other features, making the resulting shift severity unpredictable.

### 5.4 LLM-BASED COEFFICIENT ANOMALY DETECTION

Across 300 randomly selected shifts (100 per operation type), the LLM achieved modest performance, with an overall Recall@5 of 37.4%. Performance varied considerably across operation types. The model was most effective at identifying amplified coefficients (Recall@5 = 42.7%), followed by sign flips (36.7%) and attenuations (32.7%). Compared to a random baseline of 16.8%, the LLM performs approximately $2.2\times$ better than chance, demonstrating that LLMs encode domain knowledge that enables them to assess the plausibility of feature–outcome relationships and detect coefficient perturbations that deviate from expected patterns. This capability highlights the potential of LLMs to identify and mitigate biases introduced by convenience sampling.

## 6 CONCLUSION

Tabular dataset samples are gathered because of cheapness and convenience and rarely represent the greater population. We propose a benchmark for testing model robustness to convenience sampling where the train set is a sample that is non-representative of the general population and the test set represents the general population. We also provide a systematic way to control the severity of the shift to assess model robustness. Across 200 datasets and $\sim 15{,}500$ retained shift variants, seven standard baselines lose between 20% and 45% of their performance on average purely as a result of how the training sample was drawn, with the deployment distribution held fixed. Since the biased sample contains no signal that identifies the bias, closing this gap requires information from outside the training data, which is the capability INCONVENIENT is built to measure.

### AI USE STATEMENT

In this work, we used generative AI tools as part of our methodology: Claude-Haiku-4.5 was used in the LLM-based coefficient anomaly detection experiment (§5). We also used Claude-Sonnet-4.5

to assist with portions of the code development process and to generate tables included in the paper. Additionally, we used ChatGPT-5 to assist with editing and refining the manuscript text. We have reviewed all AI-assisted work: all AI-generated code was reviewed, validated, and, where necessary, modified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI, and for the accuracy and validity of all code and analyses.

We did not use AI to Generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, or interpret results.

### Reproducibility statement

We provide a command-line interface that enables users to generate shifts by specifying the minimum sample threshold ($\alpha_{min}$), feature-importance bin, coefficient operation, and operation magnitude (for multiplication and attenuation). All preprocessing steps are applied automatically during shift creation. Users can also evaluate any of our baseline methods on a selected shift by providing the paths to the corresponding training and test sets. In addition, the interface supports downloading the complete collection of $\sim 15{,}500$ shifts.

Using this interface, users can reproduce the results presented in Table 1 and evaluate their own models under shifts of varying types and difficulty levels. A comprehensive usage guide and documentation is included as a README file.

We provide a link to the anonymized repository with datasets and full results in addition to our supplementary material which only includes the scripts.

## References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 24, 2011.

James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International Conference on Machine Learning (ICML)*, pp. 115–123, 2013.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Alexander Capstick, Rahul G. Krishnan, and Payam Barnaghi. AutoElicit: Using large language models for expert prior elicitation in predictive modelling. In *International Conference on Machine Learning (ICML)*, 2025.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 785–794, 2016.

Zi-Jian Cheng, Zi-Yi Jia, Zhi Zhou, Yu-Feng Li, and Lan-Zhe Guo. TabFSBench: Tabular benchmark for feature shifts. In *International Conference on Machine Learning (ICML)*, 2025.

Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *International Conference on Algorithmic Learning Theory (ALT)*, pp. 38–53. Springer, 2008.

Tuan Dinh, Yuchen Zeng, Ruisu Zhang, Ziqian Lin, Michael Gira, Shashank Rajput, Jy-yong Sohn, Dimitris Papailiopoulos, and Kangwook Lee. LIFT: Language-interfaced fine-tuning for non-language machine learning tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Josh Gardner, Zoran Popović, and Ludwig Schmidt. Benchmarking distribution shift in tabular data with TableShift. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 18932–18943, 2021.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 35, pp. 507–520, 2022.

James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. TabLLM: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations (ICLR)*, 2023.

Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. TabTransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, pp. 5637–5664, 2021.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *International Conference on Machine Learning (ICML)*, pp. 5815–5826, 2021.

Daniel Levy, Yair Carmon, John C. Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 8847–8860, 2020.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5400–5409, 2018.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning (ICML)*, pp. 3122–3130, 2018.

Jiashuo Liu, Tianyu Wang, Peng Cui, and Hongseok Namkoong. On the need for a language describing distribution shifts: Illustrations on tabular datasets. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, and Han-Jia Ye. TALENT: A tabular analytics and learning toolbox. *arXiv preprint arXiv:2407.04057*, 2024.

Sergei Popov, Stanislav Morozov, and Artem Babenko. Neural oblivious decision ensembles for deep learning on tabular data. In *International Conference on Learning Representations (ICLR)*, 2020.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

James Requeima, John Bronskill, Dami Choi, Richard E. Turner, and David Duvenaud. LLM processes: Numerical predictive distributions conditioned on natural language. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Sara Rezaeimanesh, Kundan Thind, Farzan Siddiqui, and Mohammad M. Ghassemi. What language models know but don't say: Non-generative prior extraction for generalization. In *Conference on Language Modeling (COLM)*, 2026.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

David Selby, Kai Spriestersbach, Yuichiro Iwashita, Dennis Bappert, Archana Warrier, Sumantrak Mukherjee, Muhammad Nabeel Asim, Koichi Kise, and Sebastian Vollmer. Quantitative knowledge retrieval from large language models. *arXiv preprint arXiv:2402.07770*, 2024.

Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C. Bayan Bruss, and Tom Goldstein. SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision (ECCV) Workshops*, pp. 443–450, 2016.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.

Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *International Conference on Machine Learning (ICML)*, pp. 114, 2004.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

# A Appendix

## A.1 Inapplicable Methods

As mentioned in §4.1, our baseline suite is narrower than TableShift. Most of the robustness and domain-generalization literature assumes access to information that the convenience-sampling setting withholds by construction, and identifying which assumptions fail is itself informative about why this setting is difficult.

**Methods requiring multiple training domains.** Domain generalization methods – DANN (Ganin et al., 2016), IRM (Arjovsky et al., 2019), domain MixUp (Zhang et al., 2018), VREx (Krueger et al., 2021), DeepCORAL (Sun & Saenko, 2016), and MMD (Li et al., 2018) – and group DRO over domains (Sagawa et al., 2020) all require domain labels at training time and at least two training domains from which an invariance can be inferred. INCONVENIENT provides a single convenience sample with no domain structure, which is the defining feature of the setting we model: a practitioner with one dataset from one hospital, one platform, or one recruitment pipeline.

The obstacle is more than a missing input. Suppose one partitions the convenience sample into synthetic environments and runs IRM or VREx on the result. The bias induced by our greedy selection is global – the subset is chosen to satisfy a single coefficient criterion over the whole sample

– so the distorted feature–target relationship is stable across any such partition. An invariance-seeking objective would identify the manipulated relationship as invariant and preserve it, which is the opposite of the desired behavior. Invariance is only informative when some environment exhibits the population relationship; here, none does.

**Methods requiring information about the target distribution.** Label-shift methods, including group DRO over class labels and adversarial label robustness, target changes in the marginal $P(y)$ and require either target-domain label proportions or unlabeled target data to estimate them (e.g. Lipton et al., 2018). Our protocol withholds both. Label shift does occur in INCONVENIENT as a by-product of influence-based selection, but it is a side effect rather than the mechanism, and a method that corrects $P(y)$ alone leaves the manipulated conditional $P(y \mid x)$ untouched.

**Distributionally robust optimization.** DRO with $\chi^2$ or CVaR geometry (Levy et al., 2020) is the one omitted family that needs no side information, and it is the omission most worth justifying. These methods minimize worst-case risk over distributions within a bounded divergence ball of the empirical training distribution, which amounts to a reweighting of the observed sample. This is exactly the correction class that the sample-selection literature shows to be insufficient here: when selection depends only on $X$, importance weighting recovers the population risk (Zadrozny, 2004; Cortes et al., 2008), but when selection depends on $Y$ or on the $X$–$Y$ relationship – as our residual-conditioned selection does by design – the population conditional is not expressible as a reweighting of the biased sample and is not identifiable from it. No choice of weights over convenience examples restores a relationship that the selection removed.

A second, practical obstacle compounds the first. DRO's uncertainty radius is the hyperparameter that governs how much robustness is purchased at what in-distribution cost, and our protocol permits tuning only on a validation split drawn from the convenience sample (§4.3). That split is subject to the same bias as the training data, so it systematically prefers the smallest radius, collapsing DRO toward ERM. We therefore regard DRO as a natural addition to a future version of INCONVENIENT – one that would demonstrate this failure empirically rather than by argument – but not as a baseline that could plausibly close the gap reported in §5.1.

**Tabular neural architectures.** SAINT (Somepalli et al., 2021), TabTransformer (Huang et al., 2020), NODE (Popov et al., 2020), and FT-Transformer and tabular ResNet (Gorishniy et al., 2021) are applicable in principle and excluded on cost and expected informativeness rather than principle. INCONVENIENT spans 200 datasets and $\sim 15{,}500$ retained shift variants, each requiring an IID and an OOD training run with a 20-trial hyperparameter search; the marginal value of a fifth and sixth deep architecture does not justify that budget, particularly given that gradient-boosted trees remain the stronger family at this data scale (Grinsztajn et al., 2022). More importantly, none of these architectures has access to information that our included baselines lack. They observe the same anonymous feature matrix and, like every method in our suite, have no way to know what a column means. Since the only route back to the population relationship is knowledge external to the biased sample, we expect architectural variation within the data-driven family to shift the gap marginally rather than close it. The included MLP and TabPFN bracket this family: a standard deep network trained from scratch, and a pretrained in-context learner.

**Implication.** Taken together, nearly the entire robustness toolkit is either inapplicable or provably insufficient in the convenience-sampling regime, because each method assumes an information source – multiple environments, group labels, target-domain samples, or a correctable selection mechanism – that this setting removes. This is the gap INCONVENIENT is designed to expose, and it motivates evaluating methods whose robustness comes from pretrained knowledge of what the features mean rather than from the structure of the training sample.

### A.2 DATASETS

The following names all the datasets used in INCONVENIENT.

### A.2.1 REGRESSION (80 DATASETS)

**OpenML (72 datasets):** *1000 Cameras Dataset*, *3d Estimation Using Rssi of Wlan Dataset*, *3d Estimation Using Rssi of Wlan Dataset Complete 1 Target*, *Abalone*, *Ailerons*, *Airfoil Self Noise*, *Analcatdata Supreme*, *Another Dataset on Used Fiat 500 (1538 Rows)*, *Auction Verification*, *Avocado Sales*, *Bng(echomonths)*, *Bng(lowbwt)*, *Bng(stock)*, *Boston*, *Brazilian Houses*, *Brazilian Houses Reproduced*, *California Housing*, *Cars*, *Concrete Compressive Strength*, *Cookbookreviews*, *Cpmp 2015 Regression*, *Cps1988*, *Cps88wages*, *Cpu Activity*, *Dataset Sales*, *Debutanizer*, *Delta Elevators*, *Diamonds*, *Elevators*, *Energy Efficiency*, *Fifa*, *Forest Fires*, *Goodreads Computer Books*, *Grid Stability*, *Health Insurance*, *House 16h Reg*, *House 8l*, *House Prices Nominal*, *House Sales*, *House Sales Reduced*, *Houses*, *Ieee80211aa Gats*, *Job Profitability*, *Kin8nm*, *Kings County*, *Miami Housing*, *Miamihousing2016*, *Mip 2016 Regression*, *Moneyball*, *Naval Propulsion Plant*, *Onlinenewspopularity*, *Pumadyn32nh*, *Qsar Fish Toxicity*, *Qsar Tid 10980*, *Qsar Tid 11*, *Red Wine*, *Santander Transaction Value*, *Sat11 Hand Runtime Regression*, *Sensory*, *Socmob*, *Solar Flare*, *Space Ga*, *Student Performance Por*, *Sulfur*, *Superconductivity*, *Tecator*, *Treasury*, *Us Crime*, *Video Transcoding*, *White Wine*, *Wind*, *Wine Quality*.

**Kaggle (1 datasets):** *Bike Sharing Demand Challange*.

**TALENT (7 datasets):** *Bank32nh*, *Cpu Act*, *Cpu Small*, *Pol Reg*, *Puma32h*, *Puma8nh*.

### A.2.2 BINARY CLASSIFICATION (39 DATASETS)

**OpenML (23 datasets):** *Adityakadiwal Water Potability*, *Akshaydattatraykhare Diabetes Dataset*, *Amirhosseinmirzaie Pistachio Types Detection*, *Breast W*, *Climate Model Simulation Crashes*, *Compas Two Years*, *Credit Approval*, *Deependraverma13 Diabetes Healthcare Comprehensive Dataset*, *Diabetes*, *Dresses Sales*, *Eishkaran Heart Disease*, *Fedesoriano Stroke Prediction Dataset*, *Gabrielsantello Cars Purchase Decision Dataset*, *Hashemi221022 Diabetes*, *Kabure German Credit Data With Risk*, *Kandij Diabetes Dataset*, *Saurabh00007 Diabetescsv*, *Shubhamgupta012 Titanic Dataset*, *Tejashvi14 Employee Future Prediction*, *Tejashvi14 Engineering Placements Prediction*, *Tic Tac Toe*, *Uciml Indian Liver Patient Records*, *Uciml Pima Indians Diabetes Database*, *Whenamancodes Predict Diabities*, *Wilt*.

**Kaggle (11 datasets):** *Ashishkumarjayswal Diabetes Dataset*, *Barun2104 Telecom Churn*, *Bhavkaur Simplified Titanic Dataset*, *Chirin Africa Economic Banking and Systemic Crisis Data*, *Dileep070 Heart Disease Prediction Using Logistic Regression*, *Jillanisofttech Brain Stroke Dataset*, *Mathchi Diabetes Data Set*, *Nancyalaswad90 Review*, *Vstacknocopyright Blood Transfusion Service Center Data*.

**TALENT (4 datasets):** *Alirezachahardoli Bank Personal Loan 1*, *Mahnazarjmand Bank Personal Loan*, *Maryammanoochehry Bank Personal Loan*, *Sanjanchaudhari Bankloan*.

### A.2.3 MULTICLASS CLASSIFICATION (81 DATASETS)

**OpenML (78 datasets):** *Aakashjoshi123 Exercise and Fitness Metrics Dataset*, *Abalone*, *Abcsds Pokemon*, *Agirlcoding All Space Missions From 1957*, *Airfoil Self Noise*, *Alexisbcook Pakistan Intellectual Capital*, *Altruistdelhite04 Gold Price Data*, *Amirhosseinmirzaie Countries Life Expectancy*, *Analcatdata Dmft*, *Analcatdata Supreme*, *Ananthr1 Weather Prediction*, *Andrewmvd Udemy Courses*, *Arashnic Time Series Forecasting With Yahoo Stock Price*, *Arnabchaki Data Science Salaries 2023*, *Arnabchaki Indian Restaurants 2023*, *Arnavsmayan Netflix Userbase Dataset*, *Arslanr369 Bitcoin Price 2014 2023*, *Arslanr369 Roblox Stock Pricing 2021 2023*, *Auction Verification*, *Awaiskaggler Insurance Csv*, *Balance Scale*, *Bhanupratapbiswas Bollywood Actress Name and Movie List*, *Bhavkaur Hotel Guests Dataset*, *Car*, *Cars*, *Cmc*, *Concrete Compressive Strength*, *Cpluzshrijayan Milkquality*, *Crxxom Manhwa Dataset*, *Energy Efficiency*, *Forest Fires*, *Gyanprakashkushwaha Laptop Price Prediction Cleaned Dataset*, *Harishkumardatalab Medical Insurance Price Prediction*, *Harshitshankhdhar Imdb Dataset of Top 1000 Movies and Tv Shows*, *Iamsumat Spotify Top 2000s Mega Dataset*, *Iqmansingh Company Employee Dataset*, *Ishadss Productivity Prediction of Garment Employees*, *Jainilcoder Netflix Stock Price Prediction*, *Kanths028 Usa Housing*, *Kin8nm*, *Kingabzpro Cosmetics Datasets*, *Mayankpatel14 Second Hand Used Cars Data Set Linear Regression*, *Mayuriawati Bangalore Chain Restaurants Ratings and Reviews*, *Mfeat Morphological*, *Mirichoi0218 Insurance*, *Moneyball*, *Nikhil1e9 Netflix Stock Price*, *Noordeen In-*

*surance Premium Prediction*, *Oles04 Top Leagues Player*, *Prkhrawsthi Bitcoin Usd Daily Price With Volume 2015 2023*, *Qsar Fish Toxicity*, *Rajyellow46 Wine Quality*, *Receplyasolu 6k Weather Labeled Spotify Songs*, *Red Wine*, *Redwankarimsony Heart Disease Data*, *Rkiattisak Student Performance in Mathematics*, *Rpaguirre Tesla Stock Price*, *Ruchi798 Tv Shows on Netflix Prime Video Hulu and Disney*, *Sabasaeed1953 Stock Prices of 2023*, *Saloni1712 Chatgpt App Reviews*, *Sanjanchaudhari Netflix Dataset*, *Saunakghosh Nba Players Dataset*, *Sbhatti Financial Sentiment Analysis*, *Segment*, *Shreyanshverma27 Imdb Horror Chilling Movie Dataset*, *Shreyapurohit Anime Data*, *Socmob*, *Space Ga*, *Sudarshan6561 Ipl 2023*, *Teertha Ushealthinsurancedataset*, *Timoboz Tesla Stock Data From 2010 to 2020*, *Uciml Red Wine Quality Cortez Et Al 2009*, *Varpit94 Tesla Stock Data Updated Till 28jun2021*, *Vedavyasv Usa Housing*, *Vehicle*, *Vijayvvenkitesh Microsoft Stock Time Series Analysis*, *Visualizing Soil*, *Whenamancodes Students Performance in Exams*, *White Wine*, *Wine Quality*, *Yasserh Wine Quality Dataset*.

**Kaggle (2 datasets):** *Awaiskaggler Insurance Csv*, *Bhavkaur Hotel Guests Dataset*.

### A.3 Detailed Performance Numbers

3

Table 1: Mean performance across all shift configurations. Each row represents the average performance across all datasets for a specific configuration of task type, baseline method, operation, feature importance bin, and minimum fraction.

| Task | Method | Operation | Bin | MinFrac | IID | OOD | Degradation | Shifts |
|---|---|---|---|---|---|---|---|---|
| Binary | catboost | amplify | 0 | 0.1 | 0.808 | 0.634 | 0.174 | 8 |
| Binary | catboost | amplify | 0 | 0.3 | 0.893 | 0.704 | 0.190 | 5 |
| Binary | catboost | amplify | 1 | 0.1 | 0.834 | 0.697 | 0.138 | 22 |
| Binary | catboost | amplify | 2 | 0.1 | 0.978 | 0.750 | 0.228 | 2 |
| Binary | catboost | amplify | 3 | 0.1 | 0.767 | 0.627 | 0.140 | 15 |
| Binary | catboost | amplify | 4 | 0.1 | 1.000 | 0.879 | 0.121 | 4 |
| Binary | catboost | attenuate | 0 | 0.1 | 0.890 | 0.708 | 0.182 | 80 |
| Binary | catboost | attenuate | 0 | 0.3 | 0.924 | 0.772 | 0.152 | 40 |
| Binary | catboost | attenuate | 0 | 0.5 | 0.959 | 0.830 | 0.128 | 17 |
| Binary | catboost | attenuate | 0 | 0.7 | 0.997 | 0.857 | 0.139 | 4 |
| Binary | catboost | attenuate | 0 | 0.9 | 0.952 | 0.840 | 0.113 | 4 |
| Binary | catboost | attenuate | 1 | 0.1 | 0.810 | 0.688 | 0.122 | 8 |
| Binary | catboost | attenuate | 1 | 0.3 | 0.992 | 0.847 | 0.145 | 6 |
| Binary | catboost | attenuate | 1 | 0.5 | 1.000 | 0.834 | 0.166 | 4 |
| Binary | catboost | attenuate | 2 | 0.1 | 0.857 | 0.716 | 0.140 | 11 |
| Binary | catboost | attenuate | 2 | 0.3 | 1.000 | 0.875 | 0.124 | 4 |
| Binary | catboost | attenuate | 3 | 0.1 | 0.808 | 0.666 | 0.142 | 8 |
| Binary | catboost | attenuate | 4 | 0.1 | 0.807 | 0.663 | 0.144 | 10 |
| Binary | catboost | attenuate | 5 | 0.1 | 0.795 | 0.674 | 0.121 | 9 |
| Binary | catboost | flip_sign | 0 | 0.1 | 0.863 | 0.648 | 0.215 | 13 |
| Binary | catboost | flip_sign | 0 | 0.3 | 0.927 | 0.738 | 0.189 | 3 |
| Binary | catboost | flip_sign | 0 | 0.5 | 0.998 | 0.875 | 0.123 | 2 |
| Binary | catboost | flip_sign | 1 | 0.1 | 0.803 | 0.664 | 0.139 | 10 |
| Binary | catboost | flip_sign | 1 | 0.3 | 1.000 | 0.845 | 0.155 | 1 |
| Binary | catboost | flip_sign | 1 | 0.5 | 1.000 | 0.833 | 0.166 | 1 |
| Binary | catboost | flip_sign | 2 | 0.1 | 0.989 | 0.760 | 0.229 | 2 |
| Binary | catboost | flip_sign | 2 | 0.3 | 0.989 | 0.782 | 0.207 | 2 |
| Binary | catboost | flip_sign | 3 | 0.1 | 0.808 | 0.634 | 0.174 | 2 |
| Binary | catboost | flip_sign | 3 | 0.3 | 1.000 | 0.899 | 0.100 | 1 |
| Binary | catboost | flip_sign | 4 | 0.1 | 0.854 | 0.704 | 0.150 | 2 |
| Binary | lightgbm | amplify | 0 | 0.1 | 0.818 | 0.643 | 0.175 | 7 |
| Binary | lightgbm | amplify | 0 | 0.3 | 0.891 | 0.721 | 0.170 | 5 |
| Binary | lightgbm | amplify | 1 | 0.1 | 0.833 | 0.705 | 0.128 | 15 |
| Binary | lightgbm | amplify | 2 | 0.1 | 0.984 | 0.500 | 0.484 | 2 |
| Binary | lightgbm | amplify | 3 | 0.1 | 0.822 | 0.602 | 0.219 | 21 |
| Binary | lightgbm | amplify | 3 | 0.3 | 0.996 | 0.848 | 0.148 | 2 |
| Binary | lightgbm | amplify | 4 | 0.1 | 1.000 | 0.784 | 0.216 | 4 |
| Binary | lightgbm | amplify | 4 | 0.3 | 0.996 | 0.886 | 0.110 | 1 |
| Binary | lightgbm | attenuate | 0 | 0.1 | 0.868 | 0.661 | 0.206 | 103 |
| Binary | lightgbm | attenuate | 0 | 0.3 | 0.925 | 0.723 | 0.202 | 38 |
| Binary | lightgbm | attenuate | 0 | 0.5 | 0.956 | 0.778 | 0.178 | 17 |
| Binary | lightgbm | attenuate | 0 | 0.7 | 0.996 | 0.826 | 0.170 | 4 |
| Binary | lightgbm | attenuate | 0 | 0.9 | 0.949 | 0.844 | 0.105 | 2 |
| Binary | lightgbm | attenuate | 1 | 0.1 | 0.878 | 0.674 | 0.204 | 28 |
| Binary | lightgbm | attenuate | 1 | 0.3 | 0.951 | 0.777 | 0.174 | 9 |
| Binary | lightgbm | attenuate | 1 | 0.5 | 0.964 | 0.798 | 0.165 | 6 |
| Binary | lightgbm | attenuate | 1 | 0.7 | 0.783 | 0.674 | 0.109 | 1 |
| Binary | lightgbm | attenuate | 2 | 0.1 | 0.910 | 0.673 | 0.238 | 19 |

Continued on next page

Table 1 – Continued from previous page

| Task | Method | Operation | Bin | MinFrac | IID | OOD | Degradation | Shifts |
|---|---|---|---|---|---|---|---|---|
| Binary | lightgbm | attenuate | 2 | 0.3 | 0.998 | 0.845 | 0.153 | 8 |
| Binary | lightgbm | attenuate | 2 | 0.5 | 1.000 | 0.852 | 0.148 | 4 |
| Binary | lightgbm | attenuate | 3 | 0.1 | 0.867 | 0.682 | 0.185 | 13 |
| Binary | lightgbm | attenuate | 3 | 0.3 | 0.992 | 0.857 | 0.135 | 2 |
| Binary | lightgbm | attenuate | 4 | 0.1 | 0.828 | 0.640 | 0.187 | 16 |
| Binary | lightgbm | attenuate | 5 | 0.1 | 0.873 | 0.706 | 0.167 | 12 |
| Binary | lightgbm | attenuate | 5 | 0.3 | 0.996 | 0.893 | 0.103 | 2 |
| Binary | lightgbm | flip_sign | 0 | 0.1 | 0.862 | 0.624 | 0.239 | 13 |
| Binary | lightgbm | flip_sign | 0 | 0.3 | 0.982 | 0.736 | 0.246 | 3 |
| Binary | lightgbm | flip_sign | 0 | 0.5 | 0.987 | 0.800 | 0.187 | 3 |
| Binary | lightgbm | flip_sign | 1 | 0.1 | 0.863 | 0.705 | 0.158 | 10 |
| Binary | lightgbm | flip_sign | 1 | 0.3 | 1.000 | 0.742 | 0.258 | 1 |
| Binary | lightgbm | flip_sign | 1 | 0.5 | 1.000 | 0.799 | 0.201 | 1 |
| Binary | lightgbm | flip_sign | 2 | 0.1 | 0.846 | 0.623 | 0.223 | 6 |
| Binary | lightgbm | flip_sign | 2 | 0.3 | 0.992 | 0.833 | 0.159 | 2 |
| Binary | lightgbm | flip_sign | 3 | 0.1 | 0.873 | 0.534 | 0.338 | 3 |
| Binary | lightgbm | flip_sign | 3 | 0.3 | 1.000 | 0.828 | 0.172 | 1 |
| Binary | lightgbm | flip_sign | 4 | 0.1 | 0.772 | 0.595 | 0.176 | 3 |
| Binary | lightgbm | flip_sign | 5 | 0.1 | 0.996 | 0.709 | 0.287 | 1 |
| Binary | lightgbm | flip_sign | 5 | 0.3 | 0.996 | 0.882 | 0.114 | 1 |
| Binary | logistic | amplify | 1 | 0.1 | 0.816 | 0.500 | 0.316 | 35 |
| Binary | logistic | amplify | 1 | 0.3 | 0.717 | 0.607 | 0.110 | 1 |
| Binary | logistic | amplify | 2 | 0.1 | 0.966 | 0.787 | 0.178 | 3 |
| Binary | logistic | amplify | 3 | 0.1 | 0.748 | 0.577 | 0.171 | 5 |
| Binary | logistic | attenuate | 0 | 0.1 | 0.886 | 0.709 | 0.178 | 56 |
| Binary | logistic | attenuate | 0 | 0.3 | 0.897 | 0.673 | 0.224 | 26 |
| Binary | logistic | attenuate | 0 | 0.5 | 0.966 | 0.736 | 0.229 | 12 |
| Binary | logistic | attenuate | 0 | 0.7 | 0.963 | 0.638 | 0.325 | 4 |
| Binary | logistic | attenuate | 0 | 0.9 | 0.963 | 0.638 | 0.325 | 4 |
| Binary | logistic | attenuate | 1 | 0.1 | 0.873 | 0.708 | 0.165 | 14 |
| Binary | logistic | attenuate | 1 | 0.3 | 0.913 | 0.712 | 0.201 | 6 |
| Binary | logistic | attenuate | 1 | 0.5 | 0.995 | 0.747 | 0.248 | 4 |
| Binary | logistic | attenuate | 2 | 0.1 | 0.817 | 0.648 | 0.169 | 9 |
| Binary | logistic | attenuate | 2 | 0.3 | 0.995 | 0.776 | 0.220 | 4 |
| Binary | logistic | attenuate | 2 | 0.5 | 0.995 | 0.770 | 0.226 | 4 |
| Binary | logistic | attenuate | 3 | 0.1 | 0.809 | 0.688 | 0.121 | 18 |
| Binary | logistic | attenuate | 3 | 0.3 | 0.995 | 0.832 | 0.164 | 4 |
| Binary | logistic | attenuate | 3 | 0.5 | 0.995 | 0.826 | 0.169 | 4 |
| Binary | logistic | attenuate | 4 | 0.1 | 0.781 | 0.656 | 0.125 | 12 |
| Binary | logistic | attenuate | 5 | 0.1 | 0.656 | 0.534 | 0.121 | 4 |
| Binary | logistic | flip_sign | 0 | 0.1 | 0.851 | 0.530 | 0.320 | 13 |
| Binary | logistic | flip_sign | 0 | 0.3 | 0.921 | 0.596 | 0.326 | 3 |
| Binary | logistic | flip_sign | 0 | 0.5 | 0.979 | 0.702 | 0.277 | 2 |
| Binary | logistic | flip_sign | 1 | 0.1 | 0.807 | 0.530 | 0.278 | 9 |
| Binary | logistic | flip_sign | 1 | 0.3 | 0.885 | 0.704 | 0.181 | 3 |
| Binary | logistic | flip_sign | 1 | 0.5 | 0.995 | 0.754 | 0.242 | 1 |
| Binary | logistic | flip_sign | 2 | 0.1 | 0.845 | 0.678 | 0.167 | 4 |
| Binary | logistic | flip_sign | 2 | 0.3 | 0.995 | 0.743 | 0.252 | 1 |
| Binary | logistic | flip_sign | 3 | 0.1 | 0.860 | 0.610 | 0.249 | 3 |
| Binary | logistic | flip_sign | 3 | 0.3 | 0.995 | 0.842 | 0.153 | 1 |
| Binary | logistic | flip_sign | 4 | 0.1 | 0.748 | 0.615 | 0.133 | 1 |
| Binary | mlp | amplify | 0 | 0.1 | 0.800 | 0.679 | 0.121 | 3 |
| Binary | mlp | amplify | 0 | 0.3 | 0.835 | 0.716 | 0.119 | 2 |
| Binary | mlp | amplify | 1 | 0.1 | 0.751 | 0.574 | 0.177 | 35 |
| Binary | mlp | amplify | 1 | 0.3 | 0.786 | 0.660 | 0.126 | 4 |
| Binary | mlp | amplify | 2 | 0.1 | 0.909 | 0.656 | 0.253 | 12 |
| Binary | mlp | amplify | 2 | 0.3 | 0.884 | 0.597 | 0.287 | 6 |
| Binary | mlp | amplify | 2 | 0.5 | 0.923 | 0.788 | 0.135 | 2 |
| Binary | mlp | amplify | 3 | 0.1 | 0.792 | 0.623 | 0.169 | 15 |
| Binary | mlp | amplify | 3 | 0.3 | 0.706 | 0.546 | 0.160 | 8 |
| Binary | mlp | amplify | 3 | 0.5 | 0.671 | 0.508 | 0.162 | 5 |
| Binary | mlp | amplify | 3 | 0.7 | 0.666 | 0.511 | 0.155 | 2 |
| Binary | mlp | amplify | 4 | 0.1 | 0.812 | 0.640 | 0.173 | 10 |
| Binary | mlp | amplify | 4 | 0.3 | 0.733 | 0.565 | 0.168 | 7 |
| Binary | mlp | amplify | 4 | 0.5 | 0.689 | 0.507 | 0.182 | 2 |
| Binary | mlp | amplify | 4 | 0.7 | 0.689 | 0.499 | 0.190 | 1 |
| Binary | mlp | amplify | 5 | 0.3 | 0.869 | 0.718 | 0.151 | 1 |
| Binary | mlp | attenuate | 0 | 0.1 | 0.826 | 0.616 | 0.210 | 81 |
| Binary | mlp | attenuate | 0 | 0.3 | 0.838 | 0.610 | 0.227 | 49 |
| Binary | mlp | attenuate | 0 | 0.5 | 0.829 | 0.609 | 0.220 | 23 |
| Binary | mlp | attenuate | 0 | 0.7 | 0.854 | 0.614 | 0.240 | 9 |
| Binary | mlp | attenuate | 0 | 0.9 | 0.864 | 0.612 | 0.251 | 8 |
| Binary | mlp | attenuate | 1 | 0.1 | 0.860 | 0.651 | 0.209 | 42 |
| Binary | mlp | attenuate | 1 | 0.3 | 0.883 | 0.702 | 0.181 | 18 |
| Binary | mlp | attenuate | 1 | 0.5 | 0.894 | 0.728 | 0.166 | 12 |
| Binary | mlp | attenuate | 1 | 0.7 | 0.859 | 0.749 | 0.109 | 1 |
| Binary | mlp | attenuate | 2 | 0.1 | 0.829 | 0.663 | 0.166 | 21 |

Continued on next page

Table 1 – Continued from previous page

| Task | Method | Operation | Bin | MinFrac | IID | OOD | Degradation | Shifts |
|---|---|---|---|---|---|---|---|---|
| Binary | mlp | attenuate | 2 | 0.3 | 0.861 | 0.670 | 0.191 | 11 |
| Binary | mlp | attenuate | 2 | 0.5 | 0.838 | 0.681 | 0.157 | 10 |
| Binary | mlp | attenuate | 2 | 0.7 | 0.869 | 0.739 | 0.129 | 1 |
| Binary | mlp | attenuate | 2 | 0.9 | 0.689 | 0.514 | 0.175 | 1 |
| Binary | mlp | attenuate | 3 | 0.1 | 0.827 | 0.641 | 0.186 | 16 |
| Binary | mlp | attenuate | 3 | 0.3 | 0.808 | 0.618 | 0.190 | 11 |
| Binary | mlp | attenuate | 3 | 0.5 | 0.846 | 0.705 | 0.141 | 9 |
| Binary | mlp | attenuate | 4 | 0.1 | 0.865 | 0.697 | 0.168 | 15 |
| Binary | mlp | attenuate | 4 | 0.3 | 0.909 | 0.734 | 0.175 | 8 |
| Binary | mlp | attenuate | 4 | 0.5 | 0.749 | 0.590 | 0.158 | 3 |
| Binary | mlp | attenuate | 4 | 0.7 | 0.689 | 0.532 | 0.157 | 2 |
| Binary | mlp | attenuate | 5 | 0.1 | 0.933 | 0.728 | 0.205 | 8 |
| Binary | mlp | attenuate | 5 | 0.3 | 0.942 | 0.760 | 0.182 | 7 |
| Binary | mlp | attenuate | 5 | 0.5 | 0.869 | 0.757 | 0.111 | 1 |
| Binary | mlp | flip_sign | 0 | 0.1 | 0.904 | 0.569 | 0.334 | 4 |
| Binary | mlp | flip_sign | 0 | 0.3 | 0.985 | 0.569 | 0.417 | 2 |
| Binary | mlp | flip_sign | 0 | 0.5 | 0.933 | 0.626 | 0.306 | 2 |
| Binary | mlp | flip_sign | 1 | 0.1 | 0.797 | 0.603 | 0.194 | 15 |
| Binary | mlp | flip_sign | 1 | 0.3 | 0.842 | 0.673 | 0.169 | 4 |
| Binary | mlp | flip_sign | 1 | 0.5 | 0.794 | 0.628 | 0.166 | 5 |
| Binary | mlp | flip_sign | 1 | 0.7 | 0.689 | 0.502 | 0.187 | 1 |
| Binary | mlp | flip_sign | 1 | 0.9 | 0.689 | 0.533 | 0.156 | 1 |
| Binary | mlp | flip_sign | 2 | 0.1 | 0.895 | 0.621 | 0.274 | 4 |
| Binary | mlp | flip_sign | 2 | 0.3 | 0.960 | 0.706 | 0.254 | 2 |
| Binary | mlp | flip_sign | 2 | 0.5 | 0.752 | 0.585 | 0.167 | 3 |
| Binary | mlp | flip_sign | 2 | 0.7 | 0.689 | 0.537 | 0.152 | 1 |
| Binary | mlp | flip_sign | 2 | 0.9 | 0.689 | 0.525 | 0.164 | 1 |
| Binary | mlp | flip_sign | 3 | 0.1 | 0.851 | 0.587 | 0.265 | 3 |
| Binary | mlp | flip_sign | 3 | 0.3 | 0.843 | 0.648 | 0.195 | 2 |
| Binary | mlp | flip_sign | 4 | 0.1 | 0.843 | 0.674 | 0.168 | 2 |
| Binary | mlp | flip_sign | 4 | 0.3 | 0.997 | 0.781 | 0.216 | 1 |
| Binary | mlp | flip_sign | 4 | 0.5 | 0.869 | 0.716 | 0.153 | 1 |
| Binary | mlp | flip_sign | 5 | 0.1 | 0.869 | 0.618 | 0.250 | 1 |
| Binary | mlp | flip_sign | 5 | 0.3 | 0.869 | 0.746 | 0.123 | 1 |
| Binary | random_forest | amplify | 0 | 0.1 | 0.794 | 0.645 | 0.149 | 9 |
| Binary | random_forest | amplify | 0 | 0.3 | 0.886 | 0.738 | 0.148 | 5 |
| Binary | random_forest | amplify | 1 | 0.1 | 0.828 | 0.702 | 0.126 | 20 |
| Binary | random_forest | amplify | 2 | 0.1 | 0.940 | 0.643 | 0.296 | 2 |
| Binary | random_forest | amplify | 3 | 0.1 | 0.758 | 0.630 | 0.128 | 15 |
| Binary | random_forest | amplify | 4 | 0.1 | 0.996 | 0.826 | 0.169 | 4 |
| Binary | random_forest | amplify | 4 | 0.3 | 0.815 | 0.708 | 0.107 | 3 |
| Binary | random_forest | attenuate | 0 | 0.1 | 0.878 | 0.688 | 0.191 | 87 |
| Binary | random_forest | attenuate | 0 | 0.3 | 0.922 | 0.749 | 0.173 | 43 |
| Binary | random_forest | attenuate | 0 | 0.5 | 0.952 | 0.803 | 0.149 | 18 |
| Binary | random_forest | attenuate | 0 | 0.7 | 0.951 | 0.802 | 0.149 | 8 |
| Binary | random_forest | attenuate | 0 | 0.9 | 0.977 | 0.845 | 0.132 | 5 |
| Binary | random_forest | attenuate | 1 | 0.1 | 0.807 | 0.648 | 0.160 | 16 |
| Binary | random_forest | attenuate | 1 | 0.3 | 0.968 | 0.814 | 0.153 | 8 |
| Binary | random_forest | attenuate | 1 | 0.5 | 0.996 | 0.816 | 0.179 | 4 |
| Binary | random_forest | attenuate | 2 | 0.1 | 0.839 | 0.702 | 0.137 | 20 |
| Binary | random_forest | attenuate | 2 | 0.3 | 0.996 | 0.831 | 0.165 | 4 |
| Binary | random_forest | attenuate | 2 | 0.5 | 0.996 | 0.861 | 0.135 | 4 |
| Binary | random_forest | attenuate | 3 | 0.1 | 0.763 | 0.608 | 0.155 | 12 |
| Binary | random_forest | attenuate | 3 | 0.3 | 0.984 | 0.855 | 0.129 | 5 |
| Binary | random_forest | attenuate | 3 | 0.5 | 0.996 | 0.889 | 0.107 | 4 |
| Binary | random_forest | attenuate | 4 | 0.1 | 0.754 | 0.601 | 0.153 | 16 |
| Binary | random_forest | attenuate | 5 | 0.1 | 0.789 | 0.637 | 0.152 | 10 |
| Binary | random_forest | flip_sign | 0 | 0.1 | 0.859 | 0.639 | 0.220 | 13 |
| Binary | random_forest | flip_sign | 0 | 0.3 | 0.968 | 0.723 | 0.244 | 2 |
| Binary | random_forest | flip_sign | 0 | 0.5 | 0.995 | 0.800 | 0.195 | 2 |
| Binary | random_forest | flip_sign | 1 | 0.1 | 0.801 | 0.647 | 0.154 | 9 |
| Binary | random_forest | flip_sign | 1 | 0.3 | 0.996 | 0.801 | 0.195 | 1 |
| Binary | random_forest | flip_sign | 1 | 0.5 | 0.996 | 0.817 | 0.179 | 1 |
| Binary | random_forest | flip_sign | 2 | 0.1 | 0.884 | 0.700 | 0.184 | 4 |
| Binary | random_forest | flip_sign | 2 | 0.3 | 0.968 | 0.790 | 0.178 | 2 |
| Binary | random_forest | flip_sign | 3 | 0.1 | 0.875 | 0.708 | 0.168 | 3 |
| Binary | random_forest | flip_sign | 3 | 0.3 | 0.996 | 0.823 | 0.172 | 1 |
| Binary | random_forest | flip_sign | 4 | 0.1 | 0.860 | 0.705 | 0.156 | 2 |
| Binary | random_forest | flip_sign | 4 | 0.3 | 0.996 | 0.889 | 0.107 | 1 |
| Binary | xgboost | amplify | 0 | 0.1 | 0.871 | 0.692 | 0.179 | 3 |
| Binary | xgboost | amplify | 0 | 0.3 | 0.949 | 0.692 | 0.256 | 1 |
| Binary | xgboost | amplify | 1 | 0.1 | 0.832 | 0.687 | 0.145 | 20 |
| Binary | xgboost | amplify | 2 | 0.1 | 0.983 | 0.860 | 0.123 | 2 |
| Binary | xgboost | amplify | 3 | 0.1 | 0.823 | 0.681 | 0.141 | 21 |
| Binary | xgboost | amplify | 4 | 0.1 | 0.996 | 0.813 | 0.183 | 4 |
| Binary | xgboost | attenuate | 0 | 0.1 | 0.883 | 0.667 | 0.216 | 91 |
| Binary | xgboost | attenuate | 0 | 0.3 | 0.925 | 0.743 | 0.182 | 44 |

Continued on next page

Table 1 – Continued from previous page

| Task | Method | Operation | Bin | MinFrac | IID | OOD | Degradation | Shifts |
|---|---|---|---|---|---|---|---|---|
| Binary | xgboost | attenuate | 0 | 0.5 | 0.954 | 0.791 | 0.164 | 18 |
| Binary | xgboost | attenuate | 0 | 0.7 | 0.955 | 0.817 | 0.138 | 8 |
| Binary | xgboost | attenuate | 0 | 0.9 | 0.930 | 0.822 | 0.107 | 5 |
| Binary | xgboost | attenuate | 1 | 0.1 | 0.846 | 0.674 | 0.172 | 10 |
| Binary | xgboost | attenuate | 1 | 0.3 | 0.963 | 0.793 | 0.170 | 11 |
| Binary | xgboost | attenuate | 1 | 0.5 | 0.975 | 0.773 | 0.202 | 5 |
| Binary | xgboost | attenuate | 2 | 0.1 | 0.989 | 0.775 | 0.214 | 5 |
| Binary | xgboost | attenuate | 2 | 0.3 | 0.996 | 0.818 | 0.178 | 4 |
| Binary | xgboost | attenuate | 2 | 0.5 | 0.996 | 0.849 | 0.147 | 4 |
| Binary | xgboost | attenuate | 3 | 0.1 | 0.746 | 0.596 | 0.150 | 11 |
| Binary | xgboost | attenuate | 3 | 0.3 | 0.991 | 0.864 | 0.127 | 3 |
| Binary | xgboost | attenuate | 4 | 0.1 | 0.824 | 0.611 | 0.213 | 9 |
| Binary | xgboost | attenuate | 5 | 0.1 | 0.802 | 0.648 | 0.154 | 8 |
| Binary | xgboost | flip_sign | 0 | 0.1 | 0.882 | 0.630 | 0.252 | 12 |
| Binary | xgboost | flip_sign | 0 | 0.3 | 0.989 | 0.626 | 0.363 | 2 |
| Binary | xgboost | flip_sign | 0 | 0.5 | 0.984 | 0.806 | 0.178 | 3 |
| Binary | xgboost | flip_sign | 1 | 0.1 | 0.885 | 0.713 | 0.172 | 7 |
| Binary | xgboost | flip_sign | 1 | 0.3 | 0.996 | 0.747 | 0.248 | 1 |
| Binary | xgboost | flip_sign | 1 | 0.5 | 0.996 | 0.761 | 0.234 | 1 |
| Binary | xgboost | flip_sign | 2 | 0.1 | 0.996 | 0.685 | 0.311 | 1 |
| Binary | xgboost | flip_sign | 2 | 0.3 | 0.989 | 0.814 | 0.175 | 2 |
| Binary | xgboost | flip_sign | 3 | 0.1 | 0.863 | 0.561 | 0.302 | 3 |
| Binary | xgboost | flip_sign | 3 | 0.3 | 0.996 | 0.854 | 0.142 | 1 |
| Binary | xgboost | flip_sign | 4 | 0.1 | 0.851 | 0.653 | 0.198 | 2 |
| Binary | xgboost | flip_sign | 5 | 0.1 | 0.996 | 0.643 | 0.353 | 1 |
| Multiclass | catboost | amplify | 0 | 0.1 | 0.887 | 0.676 | 0.211 | 36 |
| Multiclass | catboost | amplify | 0 | 0.3 | 0.869 | 0.668 | 0.201 | 16 |
| Multiclass | catboost | amplify | 0 | 0.5 | 0.972 | 0.815 | 0.157 | 11 |
| Multiclass | catboost | amplify | 1 | 0.1 | 0.829 | 0.677 | 0.152 | 63 |
| Multiclass | catboost | amplify | 1 | 0.3 | 0.844 | 0.684 | 0.161 | 8 |
| Multiclass | catboost | amplify | 1 | 0.5 | 0.992 | 0.784 | 0.209 | 2 |
| Multiclass | catboost | amplify | 2 | 0.1 | 0.803 | 0.653 | 0.150 | 30 |
| Multiclass | catboost | amplify | 2 | 0.3 | 0.821 | 0.714 | 0.107 | 6 |
| Multiclass | catboost | amplify | 3 | 0.1 | 0.971 | 0.809 | 0.162 | 10 |
| Multiclass | catboost | attenuate | 0 | 0.1 | 0.888 | 0.675 | 0.213 | 193 |
| Multiclass | catboost | attenuate | 0 | 0.3 | 0.901 | 0.706 | 0.195 | 52 |
| Multiclass | catboost | attenuate | 0 | 0.5 | 0.901 | 0.624 | 0.278 | 23 |
| Multiclass | catboost | attenuate | 0 | 0.7 | 0.869 | 0.577 | 0.292 | 15 |
| Multiclass | catboost | attenuate | 1 | 0.1 | 0.890 | 0.728 | 0.163 | 41 |
| Multiclass | catboost | attenuate | 1 | 0.3 | 0.901 | 0.784 | 0.118 | 12 |
| Multiclass | catboost | attenuate | 1 | 0.5 | 0.804 | 0.622 | 0.182 | 6 |
| Multiclass | catboost | attenuate | 1 | 0.7 | 0.710 | 0.596 | 0.114 | 3 |
| Multiclass | catboost | attenuate | 2 | 0.1 | 0.912 | 0.699 | 0.213 | 21 |
| Multiclass | catboost | attenuate | 2 | 0.3 | 0.855 | 0.688 | 0.168 | 4 |
| Multiclass | catboost | attenuate | 3 | 0.1 | 0.952 | 0.728 | 0.224 | 4 |
| Multiclass | catboost | flip_sign | 0 | 0.1 | 0.884 | 0.637 | 0.247 | 39 |
| Multiclass | catboost | flip_sign | 0 | 0.3 | 0.934 | 0.688 | 0.246 | 20 |
| Multiclass | catboost | flip_sign | 0 | 0.5 | 0.910 | 0.699 | 0.210 | 7 |
| Multiclass | catboost | flip_sign | 0 | 0.7 | 0.902 | 0.802 | 0.100 | 1 |
| Multiclass | catboost | flip_sign | 1 | 0.1 | 0.895 | 0.730 | 0.164 | 19 |
| Multiclass | catboost | flip_sign | 1 | 0.3 | 0.936 | 0.778 | 0.158 | 11 |
| Multiclass | catboost | flip_sign | 1 | 0.5 | 0.998 | 0.885 | 0.113 | 4 |
| Multiclass | catboost | flip_sign | 2 | 0.1 | 0.794 | 0.665 | 0.129 | 2 |
| Multiclass | catboost | flip_sign | 2 | 0.5 | 0.710 | 0.591 | 0.119 | 1 |
| Multiclass | catboost | flip_sign | 3 | 0.1 | 0.952 | 0.621 | 0.331 | 1 |
| Multiclass | lightgbm | amplify | 0 | 0.1 | 0.860 | 0.631 | 0.229 | 36 |
| Multiclass | lightgbm | amplify | 0 | 0.3 | 0.853 | 0.630 | 0.223 | 15 |
| Multiclass | lightgbm | amplify | 0 | 0.5 | 0.887 | 0.737 | 0.149 | 9 |
| Multiclass | lightgbm | amplify | 1 | 0.1 | 0.826 | 0.670 | 0.156 | 77 |
| Multiclass | lightgbm | amplify | 1 | 0.3 | 0.896 | 0.762 | 0.134 | 13 |
| Multiclass | lightgbm | amplify | 1 | 0.5 | 1.000 | 0.828 | 0.172 | 1 |
| Multiclass | lightgbm | amplify | 2 | 0.1 | 0.827 | 0.670 | 0.157 | 33 |
| Multiclass | lightgbm | amplify | 2 | 0.3 | 0.807 | 0.675 | 0.132 | 11 |
| Multiclass | lightgbm | amplify | 2 | 0.5 | 0.730 | 0.622 | 0.108 | 7 |
| Multiclass | lightgbm | amplify | 3 | 0.1 | 0.965 | 0.772 | 0.193 | 10 |
| Multiclass | lightgbm | amplify | 3 | 0.3 | 0.941 | 0.746 | 0.195 | 4 |
| Multiclass | lightgbm | attenuate | 0 | 0.1 | 0.898 | 0.649 | 0.250 | 217 |
| Multiclass | lightgbm | attenuate | 0 | 0.3 | 0.918 | 0.722 | 0.196 | 68 |
| Multiclass | lightgbm | attenuate | 0 | 0.5 | 0.910 | 0.646 | 0.263 | 24 |
| Multiclass | lightgbm | attenuate | 0 | 0.7 | 0.869 | 0.581 | 0.287 | 14 |
| Multiclass | lightgbm | attenuate | 1 | 0.1 | 0.913 | 0.720 | 0.193 | 56 |
| Multiclass | lightgbm | attenuate | 1 | 0.3 | 0.917 | 0.789 | 0.128 | 19 |
| Multiclass | lightgbm | attenuate | 1 | 0.5 | 0.844 | 0.716 | 0.128 | 11 |
| Multiclass | lightgbm | attenuate | 1 | 0.7 | 0.688 | 0.571 | 0.117 | 3 |
| Multiclass | lightgbm | attenuate | 2 | 0.1 | 0.874 | 0.637 | 0.237 | 24 |
| Multiclass | lightgbm | attenuate | 2 | 0.3 | 0.832 | 0.679 | 0.153 | 4 |

Continued on next page

Table 1 – Continued from previous page

| Task | Method | Operation | Bin | MinFrac | IID | OOD | Degradation | Shifts |
|---|---|---|---|---|---|---|---|---|
| Multiclass | lightgbm | attenuate | 3 | 0.1 | 0.941 | 0.380 | 0.561 | 4 |
| Multiclass | lightgbm | attenuate | 3 | 0.3 | 0.941 | 0.824 | 0.117 | 4 |
| Multiclass | lightgbm | flip_sign | 0 | 0.1 | 0.878 | 0.600 | 0.278 | 42 |
| Multiclass | lightgbm | flip_sign | 0 | 0.3 | 0.933 | 0.667 | 0.266 | 22 |
| Multiclass | lightgbm | flip_sign | 0 | 0.5 | 0.907 | 0.706 | 0.201 | 10 |
| Multiclass | lightgbm | flip_sign | 0 | 0.7 | 0.967 | 0.814 | 0.153 | 3 |
| Multiclass | lightgbm | flip_sign | 1 | 0.1 | 0.923 | 0.713 | 0.210 | 22 |
| Multiclass | lightgbm | flip_sign | 1 | 0.3 | 0.946 | 0.754 | 0.193 | 11 |
| Multiclass | lightgbm | flip_sign | 1 | 0.5 | 0.997 | 0.864 | 0.133 | 4 |
| Multiclass | lightgbm | flip_sign | 1 | 0.7 | 1.000 | 0.894 | 0.106 | 1 |
| Multiclass | lightgbm | flip_sign | 2 | 0.1 | 0.939 | 0.771 | 0.168 | 4 |
| Multiclass | lightgbm | flip_sign | 2 | 0.5 | 0.688 | 0.572 | 0.116 | 1 |
| Multiclass | lightgbm | flip_sign | 3 | 0.1 | 0.970 | 0.715 | 0.255 | 2 |
| Multiclass | lightgbm | flip_sign | 3 | 0.3 | 0.941 | 0.746 | 0.195 | 1 |
| Multiclass | logistic | amplify | 0 | 0.1 | 0.607 | 0.443 | 0.164 | 8 |
| Multiclass | logistic | amplify | 0 | 0.3 | 0.683 | 0.488 | 0.195 | 1 |
| Multiclass | logistic | amplify | 0 | 0.5 | 0.701 | 0.518 | 0.184 | 5 |
| Multiclass | logistic | amplify | 1 | 0.1 | 0.726 | 0.583 | 0.144 | 29 |
| Multiclass | logistic | amplify | 1 | 0.3 | 0.743 | 0.616 | 0.127 | 5 |
| Multiclass | logistic | amplify | 1 | 0.5 | 0.718 | 0.526 | 0.191 | 11 |
| Multiclass | logistic | amplify | 1 | 0.9 | 0.758 | 0.646 | 0.111 | 1 |
| Multiclass | logistic | amplify | 2 | 0.1 | 0.689 | 0.514 | 0.175 | 12 |
| Multiclass | logistic | amplify | 2 | 0.5 | 0.625 | 0.517 | 0.108 | 6 |
| Multiclass | logistic | amplify | 3 | 0.1 | 0.875 | 0.709 | 0.166 | 10 |
| Multiclass | logistic | amplify | 3 | 0.3 | 0.804 | 0.698 | 0.107 | 1 |
| Multiclass | logistic | attenuate | 0 | 0.1 | 0.798 | 0.486 | 0.311 | 137 |
| Multiclass | logistic | attenuate | 0 | 0.3 | 0.837 | 0.582 | 0.255 | 74 |
| Multiclass | logistic | attenuate | 0 | 0.5 | 0.805 | 0.554 | 0.251 | 46 |
| Multiclass | logistic | attenuate | 0 | 0.7 | 0.783 | 0.566 | 0.217 | 32 |
| Multiclass | logistic | attenuate | 1 | 0.1 | 0.819 | 0.583 | 0.236 | 20 |
| Multiclass | logistic | attenuate | 1 | 0.3 | 0.822 | 0.622 | 0.200 | 3 |
| Multiclass | logistic | attenuate | 1 | 0.5 | 0.653 | 0.514 | 0.140 | 8 |
| Multiclass | logistic | attenuate | 2 | 0.1 | 0.845 | 0.619 | 0.225 | 12 |
| Multiclass | logistic | attenuate | 2 | 0.3 | 0.804 | 0.701 | 0.103 | 3 |
| Multiclass | logistic | attenuate | 3 | 0.1 | 0.804 | 0.698 | 0.107 | 4 |
| Multiclass | logistic | flip_sign | 0 | 0.1 | 0.761 | 0.456 | 0.304 | 34 |
| Multiclass | logistic | flip_sign | 0 | 0.3 | 0.733 | 0.427 | 0.306 | 21 |
| Multiclass | logistic | flip_sign | 0 | 0.5 | 0.697 | 0.433 | 0.264 | 14 |
| Multiclass | logistic | flip_sign | 0 | 0.7 | 0.685 | 0.467 | 0.218 | 10 |
| Multiclass | logistic | flip_sign | 1 | 0.1 | 0.782 | 0.593 | 0.189 | 8 |
| Multiclass | logistic | flip_sign | 1 | 0.3 | 0.782 | 0.538 | 0.244 | 5 |
| Multiclass | logistic | flip_sign | 1 | 0.5 | 0.685 | 0.426 | 0.258 | 4 |
| Multiclass | logistic | flip_sign | 1 | 0.7 | 0.616 | 0.413 | 0.204 | 1 |
| Multiclass | logistic | flip_sign | 2 | 0.1 | 0.817 | 0.645 | 0.172 | 6 |
| Multiclass | logistic | flip_sign | 2 | 0.3 | 0.810 | 0.649 | 0.161 | 3 |
| Multiclass | logistic | flip_sign | 2 | 0.5 | 0.635 | 0.511 | 0.124 | 2 |
| Multiclass | logistic | flip_sign | 2 | 0.7 | 0.645 | 0.522 | 0.123 | 1 |
| Multiclass | logistic | flip_sign | 3 | 0.1 | 0.893 | 0.693 | 0.200 | 2 |
| Multiclass | logistic | flip_sign | 3 | 0.3 | 0.804 | 0.695 | 0.110 | 1 |
| Multiclass | mlp | amplify | 0 | 0.1 | 0.800 | 0.600 | 0.199 | 10 |
| Multiclass | mlp | amplify | 0 | 0.3 | 0.797 | 0.551 | 0.245 | 4 |
| Multiclass | mlp | amplify | 0 | 0.5 | 0.836 | 0.644 | 0.192 | 6 |
| Multiclass | mlp | amplify | 1 | 0.1 | 0.842 | 0.635 | 0.207 | 44 |
| Multiclass | mlp | amplify | 1 | 0.3 | 0.850 | 0.689 | 0.161 | 20 |
| Multiclass | mlp | amplify | 1 | 0.5 | 0.911 | 0.770 | 0.141 | 7 |
| Multiclass | mlp | amplify | 2 | 0.1 | 0.791 | 0.638 | 0.153 | 43 |
| Multiclass | mlp | amplify | 2 | 0.3 | 0.765 | 0.579 | 0.187 | 17 |
| Multiclass | mlp | amplify | 2 | 0.5 | 0.772 | 0.597 | 0.175 | 8 |
| Multiclass | mlp | amplify | 2 | 0.7 | 0.693 | 0.582 | 0.110 | 1 |
| Multiclass | mlp | amplify | 3 | 0.1 | 0.866 | 0.593 | 0.273 | 7 |
| Multiclass | mlp | attenuate | 0 | 0.1 | 0.857 | 0.612 | 0.245 | 129 |
| Multiclass | mlp | attenuate | 0 | 0.3 | 0.870 | 0.593 | 0.277 | 71 |
| Multiclass | mlp | attenuate | 0 | 0.5 | 0.869 | 0.528 | 0.341 | 39 |
| Multiclass | mlp | attenuate | 0 | 0.7 | 0.870 | 0.541 | 0.330 | 20 |
| Multiclass | mlp | attenuate | 0 | 0.9 | 0.856 | 0.685 | 0.171 | 6 |
| Multiclass | mlp | attenuate | 1 | 0.1 | 0.845 | 0.664 | 0.181 | 59 |
| Multiclass | mlp | attenuate | 1 | 0.3 | 0.876 | 0.706 | 0.169 | 44 |
| Multiclass | mlp | attenuate | 1 | 0.5 | 0.857 | 0.676 | 0.181 | 21 |
| Multiclass | mlp | attenuate | 1 | 0.7 | 0.873 | 0.631 | 0.242 | 9 |
| Multiclass | mlp | attenuate | 1 | 0.9 | 0.969 | 0.830 | 0.138 | 2 |
| Multiclass | mlp | attenuate | 2 | 0.1 | 0.829 | 0.658 | 0.171 | 26 |
| Multiclass | mlp | attenuate | 2 | 0.3 | 0.843 | 0.655 | 0.188 | 15 |
| Multiclass | mlp | attenuate | 2 | 0.5 | 0.765 | 0.595 | 0.171 | 2 |
| Multiclass | mlp | attenuate | 2 | 0.7 | 0.765 | 0.652 | 0.114 | 3 |
| Multiclass | mlp | attenuate | 2 | 0.9 | 0.773 | 0.618 | 0.155 | 7 |
| Multiclass | mlp | attenuate | 3 | 0.1 | 0.812 | 0.683 | 0.129 | 3 |
| Multiclass | mlp | flip_sign | 0 | 0.1 | 0.840 | 0.603 | 0.237 | 23 |

Continued on next page

Table 1 – Continued from previous page

| Task | Method | Operation | Bin | MinFrac | IID | OOD | Degradation | Shifts |
|---|---|---|---|---|---|---|---|---|
| Multiclass | mlp | flip_sign | 0 | 0.3 | 0.849 | 0.608 | 0.241 | 11 |
| Multiclass | mlp | flip_sign | 0 | 0.5 | 0.827 | 0.589 | 0.238 | 7 |
| Multiclass | mlp | flip_sign | 0 | 0.7 | 0.798 | 0.595 | 0.204 | 5 |
| Multiclass | mlp | flip_sign | 1 | 0.1 | 0.813 | 0.613 | 0.199 | 16 |
| Multiclass | mlp | flip_sign | 1 | 0.3 | 0.915 | 0.751 | 0.164 | 5 |
| Multiclass | mlp | flip_sign | 1 | 0.5 | 0.826 | 0.709 | 0.117 | 5 |
| Multiclass | mlp | flip_sign | 2 | 0.1 | 0.810 | 0.658 | 0.152 | 8 |
| Multiclass | mlp | flip_sign | 2 | 0.3 | 0.770 | 0.608 | 0.162 | 3 |
| Multiclass | mlp | flip_sign | 2 | 0.5 | 0.787 | 0.643 | 0.144 | 3 |
| Multiclass | mlp | flip_sign | 3 | 0.1 | 0.978 | 0.733 | 0.244 | 1 |
| Multiclass | random_forest | amplify | 0 | 0.1 | 0.865 | 0.674 | 0.191 | 43 |
| Multiclass | random_forest | amplify | 0 | 0.3 | 0.886 | 0.709 | 0.178 | 22 |
| Multiclass | random_forest | amplify | 0 | 0.5 | 0.988 | 0.802 | 0.186 | 12 |
| Multiclass | random_forest | amplify | 0 | 0.7 | 1.000 | 0.884 | 0.116 | 1 |
| Multiclass | random_forest | amplify | 1 | 0.1 | 0.834 | 0.669 | 0.166 | 67 |
| Multiclass | random_forest | amplify | 1 | 0.3 | 0.822 | 0.679 | 0.143 | 8 |
| Multiclass | random_forest | amplify | 1 | 0.5 | 0.972 | 0.835 | 0.137 | 2 |
| Multiclass | random_forest | amplify | 1 | 0.7 | 1.000 | 0.894 | 0.106 | 2 |
| Multiclass | random_forest | amplify | 2 | 0.1 | 0.833 | 0.681 | 0.152 | 42 |
| Multiclass | random_forest | amplify | 2 | 0.3 | 0.833 | 0.686 | 0.147 | 11 |
| Multiclass | random_forest | amplify | 2 | 0.5 | 0.762 | 0.651 | 0.111 | 9 |
| Multiclass | random_forest | amplify | 3 | 0.1 | 0.964 | 0.773 | 0.190 | 10 |
| Multiclass | random_forest | amplify | 3 | 0.3 | 0.940 | 0.833 | 0.107 | 3 |
| Multiclass | random_forest | attenuate | 0 | 0.1 | 0.885 | 0.650 | 0.235 | 208 |
| Multiclass | random_forest | attenuate | 0 | 0.3 | 0.920 | 0.696 | 0.225 | 69 |
| Multiclass | random_forest | attenuate | 0 | 0.5 | 0.864 | 0.603 | 0.261 | 33 |
| Multiclass | random_forest | attenuate | 0 | 0.7 | 0.843 | 0.570 | 0.273 | 16 |
| Multiclass | random_forest | attenuate | 1 | 0.1 | 0.904 | 0.722 | 0.182 | 40 |
| Multiclass | random_forest | attenuate | 1 | 0.3 | 0.919 | 0.793 | 0.126 | 15 |
| Multiclass | random_forest | attenuate | 1 | 0.5 | 0.792 | 0.617 | 0.175 | 8 |
| Multiclass | random_forest | attenuate | 2 | 0.1 | 0.919 | 0.700 | 0.218 | 23 |
| Multiclass | random_forest | attenuate | 2 | 0.3 | 0.849 | 0.682 | 0.167 | 4 |
| Multiclass | random_forest | attenuate | 3 | 0.1 | 0.940 | 0.729 | 0.211 | 4 |
| Multiclass | random_forest | flip_sign | 0 | 0.1 | 0.877 | 0.606 | 0.271 | 39 |
| Multiclass | random_forest | flip_sign | 0 | 0.3 | 0.928 | 0.661 | 0.266 | 20 |
| Multiclass | random_forest | flip_sign | 0 | 0.5 | 0.916 | 0.720 | 0.196 | 9 |
| Multiclass | random_forest | flip_sign | 0 | 0.7 | 1.000 | 0.822 | 0.178 | 2 |
| Multiclass | random_forest | flip_sign | 1 | 0.1 | 0.905 | 0.718 | 0.187 | 21 |
| Multiclass | random_forest | flip_sign | 1 | 0.3 | 0.939 | 0.772 | 0.166 | 11 |
| Multiclass | random_forest | flip_sign | 1 | 0.5 | 0.932 | 0.796 | 0.136 | 5 |
| Multiclass | random_forest | flip_sign | 1 | 0.7 | 1.000 | 0.845 | 0.155 | 2 |
| Multiclass | random_forest | flip_sign | 2 | 0.1 | 0.871 | 0.698 | 0.173 | 2 |
| Multiclass | random_forest | flip_sign | 2 | 0.5 | 0.688 | 0.574 | 0.114 | 1 |
| Multiclass | random_forest | flip_sign | 3 | 0.1 | 0.969 | 0.731 | 0.238 | 2 |
| Multiclass | random_forest | flip_sign | 3 | 0.3 | 0.940 | 0.827 | 0.113 | 1 |
| Multiclass | xgboost | amplify | 0 | 0.1 | 0.833 | 0.629 | 0.203 | 48 |
| Multiclass | xgboost | amplify | 0 | 0.3 | 0.896 | 0.691 | 0.205 | 21 |
| Multiclass | xgboost | amplify | 0 | 0.5 | 0.990 | 0.835 | 0.155 | 12 |
| Multiclass | xgboost | amplify | 1 | 0.1 | 0.842 | 0.670 | 0.172 | 80 |
| Multiclass | xgboost | amplify | 1 | 0.3 | 0.902 | 0.743 | 0.159 | 14 |
| Multiclass | xgboost | amplify | 1 | 0.5 | 0.980 | 0.851 | 0.129 | 11 |
| Multiclass | xgboost | amplify | 2 | 0.1 | 0.823 | 0.666 | 0.157 | 36 |
| Multiclass | xgboost | amplify | 2 | 0.3 | 0.819 | 0.687 | 0.132 | 11 |
| Multiclass | xgboost | amplify | 2 | 0.5 | 0.803 | 0.696 | 0.107 | 4 |
| Multiclass | xgboost | amplify | 3 | 0.1 | 0.964 | 0.760 | 0.203 | 10 |
| Multiclass | xgboost | amplify | 3 | 0.3 | 0.940 | 0.790 | 0.150 | 4 |
| Multiclass | xgboost | attenuate | 0 | 0.1 | 0.903 | 0.639 | 0.264 | 220 |
| Multiclass | xgboost | attenuate | 0 | 0.3 | 0.929 | 0.709 | 0.220 | 82 |
| Multiclass | xgboost | attenuate | 0 | 0.5 | 0.926 | 0.660 | 0.266 | 37 |
| Multiclass | xgboost | attenuate | 0 | 0.7 | 0.864 | 0.580 | 0.284 | 15 |
| Multiclass | xgboost | attenuate | 1 | 0.1 | 0.901 | 0.703 | 0.198 | 46 |
| Multiclass | xgboost | attenuate | 1 | 0.3 | 0.933 | 0.812 | 0.121 | 25 |
| Multiclass | xgboost | attenuate | 1 | 0.5 | 0.789 | 0.634 | 0.155 | 6 |
| Multiclass | xgboost | attenuate | 1 | 0.7 | 0.689 | 0.562 | 0.126 | 3 |
| Multiclass | xgboost | attenuate | 2 | 0.1 | 0.913 | 0.703 | 0.210 | 22 |
| Multiclass | xgboost | attenuate | 2 | 0.3 | 0.838 | 0.673 | 0.166 | 4 |
| Multiclass | xgboost | attenuate | 3 | 0.1 | 0.940 | 0.663 | 0.277 | 4 |
| Multiclass | xgboost | attenuate | 3 | 0.3 | 0.940 | 0.821 | 0.119 | 4 |
| Multiclass | xgboost | flip_sign | 0 | 0.1 | 0.884 | 0.604 | 0.280 | 42 |
| Multiclass | xgboost | flip_sign | 0 | 0.3 | 0.933 | 0.659 | 0.274 | 21 |
| Multiclass | xgboost | flip_sign | 0 | 0.5 | 0.910 | 0.710 | 0.200 | 10 |
| Multiclass | xgboost | flip_sign | 0 | 0.7 | 0.967 | 0.825 | 0.142 | 3 |
| Multiclass | xgboost | flip_sign | 1 | 0.1 | 0.920 | 0.676 | 0.244 | 20 |
| Multiclass | xgboost | flip_sign | 1 | 0.3 | 0.946 | 0.713 | 0.234 | 11 |
| Multiclass | xgboost | flip_sign | 1 | 0.5 | 0.935 | 0.730 | 0.205 | 5 |
| Multiclass | xgboost | flip_sign | 1 | 0.7 | 1.000 | 0.860 | 0.140 | 1 |
| Multiclass | xgboost | flip_sign | 2 | 0.1 | 0.928 | 0.776 | 0.152 | 6 |

Continued on next page

Table 1 – Continued from previous page

| Task | Method | Operation | Bin | MinFrac | IID | OOD | Degradation | Shifts |
|---|---|---|---|---|---|---|---|---|
| Multiclass | xgboost | flip_sign | 3 | 0.1 | 0.969 | 0.720 | 0.249 | 2 |
| Multiclass | xgboost | flip_sign | 3 | 0.3 | 0.940 | 0.792 | 0.148 | 1 |
| Regression | catboost | amplify | 0 | 0.1 | 0.857 | 0.167 | 0.690 | 56 |
| Regression | catboost | amplify | 0 | 0.3 | 0.830 | 0.562 | 0.268 | 30 |
| Regression | catboost | amplify | 0 | 0.5 | 0.882 | 0.574 | 0.308 | 15 |
| Regression | catboost | amplify | 1 | 0.1 | 0.710 | 0.384 | 0.325 | 87 |
| Regression | catboost | amplify | 1 | 0.3 | 0.751 | 0.439 | 0.312 | 35 |
| Regression | catboost | amplify | 1 | 0.5 | 0.748 | 0.431 | 0.317 | 11 |
| Regression | catboost | amplify | 1 | 0.7 | 0.683 | 0.476 | 0.207 | 8 |
| Regression | catboost | amplify | 2 | 0.1 | 0.710 | 0.408 | 0.302 | 86 |
| Regression | catboost | amplify | 2 | 0.3 | 0.731 | 0.432 | 0.299 | 26 |
| Regression | catboost | amplify | 2 | 0.5 | 0.695 | 0.447 | 0.248 | 21 |
| Regression | catboost | amplify | 2 | 0.7 | 0.671 | 0.318 | 0.353 | 13 |
| Regression | catboost | amplify | 2 | 0.9 | 0.990 | 0.738 | 0.252 | 4 |
| Regression | catboost | amplify | 3 | 0.1 | 0.687 | 0.080 | 0.607 | 33 |
| Regression | catboost | amplify | 3 | 0.3 | 0.826 | 0.471 | 0.355 | 14 |
| Regression | catboost | amplify | 3 | 0.5 | 0.668 | 0.445 | 0.224 | 6 |
| Regression | catboost | amplify | 4 | 0.1 | 0.585 | 0.266 | 0.319 | 39 |
| Regression | catboost | amplify | 4 | 0.3 | 0.661 | 0.391 | 0.270 | 20 |
| Regression | catboost | amplify | 4 | 0.5 | 0.815 | 0.690 | 0.125 | 9 |
| Regression | catboost | amplify | 4 | 0.7 | 0.612 | 0.509 | 0.102 | 2 |
| Regression | catboost | amplify | 5 | 0.1 | 0.615 | 0.202 | 0.414 | 31 |
| Regression | catboost | amplify | 5 | 0.3 | 0.671 | 0.390 | 0.281 | 19 |
| Regression | catboost | amplify | 5 | 0.5 | 0.685 | 0.509 | 0.176 | 8 |
| Regression | catboost | amplify | 5 | 0.7 | 0.612 | 0.509 | 0.102 | 3 |
| Regression | catboost | attenuate | 0 | 0.1 | 0.739 | 0.219 | 0.521 | 175 |
| Regression | catboost | attenuate | 0 | 0.3 | 0.743 | 0.232 | 0.511 | 107 |
| Regression | catboost | attenuate | 0 | 0.5 | 0.783 | 0.243 | 0.540 | 83 |
| Regression | catboost | attenuate | 0 | 0.7 | 0.823 | 0.394 | 0.429 | 49 |
| Regression | catboost | attenuate | 0 | 0.9 | 0.902 | 0.351 | 0.552 | 18 |
| Regression | catboost | attenuate | 1 | 0.1 | 0.720 | 0.467 | 0.253 | 108 |
| Regression | catboost | attenuate | 1 | 0.3 | 0.745 | 0.456 | 0.289 | 30 |
| Regression | catboost | attenuate | 1 | 0.5 | 0.905 | 0.752 | 0.153 | 15 |
| Regression | catboost | attenuate | 1 | 0.7 | 0.760 | 0.640 | 0.120 | 6 |
| Regression | catboost | attenuate | 1 | 0.9 | 0.362 | 0.242 | 0.120 | 1 |
| Regression | catboost | attenuate | 2 | 0.1 | 0.757 | 0.444 | 0.313 | 60 |
| Regression | catboost | attenuate | 2 | 0.3 | 0.834 | 0.605 | 0.229 | 22 |
| Regression | catboost | attenuate | 2 | 0.5 | 0.853 | 0.693 | 0.161 | 10 |
| Regression | catboost | attenuate | 2 | 0.7 | 0.923 | 0.725 | 0.198 | 6 |
| Regression | catboost | attenuate | 3 | 0.1 | 0.752 | 0.367 | 0.385 | 45 |
| Regression | catboost | attenuate | 3 | 0.3 | 0.873 | 0.282 | 0.591 | 16 |
| Regression | catboost | attenuate | 3 | 0.5 | 0.949 | 0.224 | 0.724 | 9 |
| Regression | catboost | attenuate | 3 | 0.7 | 0.985 | -0.206 | 1.191 | 3 |
| Regression | catboost | attenuate | 3 | 0.9 | 0.784 | 0.678 | 0.106 | 1 |
| Regression | catboost | attenuate | 4 | 0.1 | 0.724 | 0.436 | 0.288 | 35 |
| Regression | catboost | attenuate | 4 | 0.3 | 0.720 | 0.586 | 0.134 | 7 |
| Regression | catboost | attenuate | 4 | 0.7 | 0.784 | 0.681 | 0.103 | 1 |
| Regression | catboost | attenuate | 5 | 0.1 | 0.650 | 0.475 | 0.175 | 21 |
| Regression | catboost | attenuate | 5 | 0.3 | 0.592 | 0.464 | 0.128 | 8 |
| Regression | catboost | attenuate | 5 | 0.5 | 0.636 | 0.519 | 0.117 | 3 |
| Regression | catboost | flip_sign | 0 | 0.1 | 0.749 | 0.075 | 0.674 | 29 |
| Regression | catboost | flip_sign | 0 | 0.3 | 0.763 | 0.077 | 0.686 | 22 |
| Regression | catboost | flip_sign | 0 | 0.5 | 0.702 | 0.182 | 0.519 | 13 |
| Regression | catboost | flip_sign | 0 | 0.7 | 0.780 | 0.367 | 0.414 | 5 |
| Regression | catboost | flip_sign | 0 | 0.9 | 0.980 | 0.214 | 0.767 | 2 |
| Regression | catboost | flip_sign | 1 | 0.1 | 0.745 | 0.411 | 0.335 | 32 |
| Regression | catboost | flip_sign | 1 | 0.3 | 0.804 | 0.504 | 0.301 | 16 |
| Regression | catboost | flip_sign | 1 | 0.5 | 0.802 | 0.563 | 0.239 | 10 |
| Regression | catboost | flip_sign | 1 | 0.7 | 0.750 | 0.492 | 0.257 | 3 |
| Regression | catboost | flip_sign | 2 | 0.1 | 0.777 | 0.383 | 0.394 | 22 |
| Regression | catboost | flip_sign | 2 | 0.3 | 0.819 | 0.302 | 0.517 | 10 |
| Regression | catboost | flip_sign | 2 | 0.5 | 0.806 | 0.439 | 0.367 | 7 |
| Regression | catboost | flip_sign | 2 | 0.7 | 0.873 | 0.518 | 0.355 | 4 |
| Regression | catboost | flip_sign | 3 | 0.1 | 0.731 | 0.288 | 0.443 | 12 |
| Regression | catboost | flip_sign | 3 | 0.3 | 0.801 | 0.428 | 0.373 | 6 |
| Regression | catboost | flip_sign | 3 | 0.5 | 0.725 | 0.362 | 0.363 | 5 |
| Regression | catboost | flip_sign | 3 | 0.7 | 0.798 | 0.117 | 0.681 | 2 |
| Regression | catboost | flip_sign | 3 | 0.9 | 0.612 | 0.504 | 0.108 | 1 |
| Regression | catboost | flip_sign | 4 | 0.1 | 0.846 | 0.411 | 0.435 | 8 |
| Regression | catboost | flip_sign | 4 | 0.3 | 0.943 | 0.479 | 0.464 | 2 |
| Regression | catboost | flip_sign | 4 | 0.5 | 0.983 | 0.855 | 0.128 | 1 |
| Regression | catboost | flip_sign | 5 | 0.1 | 0.624 | 0.327 | 0.298 | 8 |
| Regression | catboost | flip_sign | 5 | 0.3 | 0.622 | 0.290 | 0.331 | 2 |
| Regression | catboost | flip_sign | 5 | 0.5 | 0.904 | 0.686 | 0.218 | 1 |
| Regression | lightgbm | amplify | 0 | 0.1 | 0.816 | 0.223 | 0.593 | 51 |
| Regression | lightgbm | amplify | 0 | 0.3 | 0.793 | 0.499 | 0.293 | 41 |

Continued on next page

Table 1 – Continued from previous page

| Task | Method | Operation | Bin | MinFrac | IID | OOD | Degradation | Shifts |
|---|---|---|---|---|---|---|---|---|
| Regression | lightgbm | amplify | 0 | 0.5 | 0.974 | 0.583 | 0.391 | 19 |
| Regression | lightgbm | amplify | 1 | 0.1 | 0.727 | 0.158 | 0.568 | 85 |
| Regression | lightgbm | amplify | 1 | 0.3 | 0.751 | 0.422 | 0.329 | 48 |
| Regression | lightgbm | amplify | 1 | 0.5 | 0.510 | 0.193 | 0.317 | 20 |
| Regression | lightgbm | amplify | 1 | 0.7 | 0.504 | 0.307 | 0.198 | 9 |
| Regression | lightgbm | amplify | 2 | 0.1 | 0.682 | 0.264 | 0.418 | 71 |
| Regression | lightgbm | amplify | 2 | 0.3 | 0.663 | 0.391 | 0.272 | 35 |
| Regression | lightgbm | amplify | 2 | 0.5 | 0.719 | 0.407 | 0.312 | 13 |
| Regression | lightgbm | amplify | 2 | 0.7 | 0.700 | 0.326 | 0.375 | 11 |
| Regression | lightgbm | amplify | 2 | 0.9 | 0.994 | 0.767 | 0.227 | 4 |
| Regression | lightgbm | amplify | 3 | 0.1 | 0.568 | 0.145 | 0.424 | 46 |
| Regression | lightgbm | amplify | 3 | 0.3 | 0.798 | 0.314 | 0.483 | 11 |
| Regression | lightgbm | amplify | 3 | 0.5 | 0.659 | 0.444 | 0.215 | 6 |
| Regression | lightgbm | amplify | 3 | 0.7 | 0.939 | 0.832 | 0.107 | 1 |
| Regression | lightgbm | amplify | 4 | 0.1 | 0.585 | 0.301 | 0.284 | 51 |
| Regression | lightgbm | amplify | 4 | 0.3 | 0.716 | 0.498 | 0.218 | 34 |
| Regression | lightgbm | amplify | 4 | 0.5 | 0.838 | 0.711 | 0.127 | 14 |
| Regression | lightgbm | amplify | 4 | 0.7 | 0.603 | 0.494 | 0.109 | 3 |
| Regression | lightgbm | amplify | 5 | 0.1 | 0.651 | 0.345 | 0.306 | 46 |
| Regression | lightgbm | amplify | 5 | 0.3 | 0.709 | 0.385 | 0.325 | 25 |
| Regression | lightgbm | amplify | 5 | 0.5 | 0.784 | 0.623 | 0.160 | 12 |
| Regression | lightgbm | amplify | 5 | 0.7 | 0.603 | 0.497 | 0.107 | 4 |
| Regression | lightgbm | attenuate | 0 | 0.1 | 0.738 | 0.184 | 0.554 | 172 |
| Regression | lightgbm | attenuate | 0 | 0.3 | 0.755 | 0.262 | 0.493 | 105 |
| Regression | lightgbm | attenuate | 0 | 0.5 | 0.774 | 0.310 | 0.465 | 85 |
| Regression | lightgbm | attenuate | 0 | 0.7 | 0.823 | 0.389 | 0.434 | 47 |
| Regression | lightgbm | attenuate | 0 | 0.9 | 0.890 | 0.373 | 0.517 | 19 |
| Regression | lightgbm | attenuate | 1 | 0.1 | 0.728 | 0.350 | 0.378 | 120 |
| Regression | lightgbm | attenuate | 1 | 0.3 | 0.688 | 0.434 | 0.255 | 56 |
| Regression | lightgbm | attenuate | 1 | 0.5 | 0.763 | 0.602 | 0.161 | 19 |
| Regression | lightgbm | attenuate | 1 | 0.7 | 0.733 | 0.589 | 0.144 | 9 |
| Regression | lightgbm | attenuate | 1 | 0.9 | 0.326 | 0.200 | 0.126 | 1 |
| Regression | lightgbm | attenuate | 2 | 0.1 | 0.734 | 0.357 | 0.377 | 66 |
| Regression | lightgbm | attenuate | 2 | 0.3 | 0.715 | 0.480 | 0.234 | 43 |
| Regression | lightgbm | attenuate | 2 | 0.5 | 0.687 | 0.546 | 0.141 | 22 |
| Regression | lightgbm | attenuate | 2 | 0.7 | 0.801 | 0.613 | 0.189 | 10 |
| Regression | lightgbm | attenuate | 3 | 0.1 | 0.727 | 0.247 | 0.479 | 57 |
| Regression | lightgbm | attenuate | 3 | 0.3 | 0.818 | 0.307 | 0.511 | 23 |
| Regression | lightgbm | attenuate | 3 | 0.5 | 0.901 | 0.335 | 0.566 | 12 |
| Regression | lightgbm | attenuate | 3 | 0.7 | 0.987 | -0.182 | 1.169 | 3 |
| Regression | lightgbm | attenuate | 4 | 0.1 | 0.697 | 0.375 | 0.322 | 46 |
| Regression | lightgbm | attenuate | 4 | 0.3 | 0.600 | 0.418 | 0.182 | 13 |
| Regression | lightgbm | attenuate | 4 | 0.5 | 0.778 | 0.669 | 0.110 | 4 |
| Regression | lightgbm | attenuate | 5 | 0.1 | 0.576 | 0.361 | 0.215 | 36 |
| Regression | lightgbm | attenuate | 5 | 0.3 | 0.508 | 0.352 | 0.156 | 14 |
| Regression | lightgbm | attenuate | 5 | 0.5 | 0.778 | 0.656 | 0.122 | 4 |
| Regression | lightgbm | flip_sign | 0 | 0.1 | 0.777 | 0.020 | 0.757 | 29 |
| Regression | lightgbm | flip_sign | 0 | 0.3 | 0.759 | 0.110 | 0.649 | 22 |
| Regression | lightgbm | flip_sign | 0 | 0.5 | 0.704 | 0.229 | 0.475 | 15 |
| Regression | lightgbm | flip_sign | 0 | 0.7 | 0.809 | 0.438 | 0.371 | 6 |
| Regression | lightgbm | flip_sign | 0 | 0.9 | 0.979 | 0.225 | 0.754 | 2 |
| Regression | lightgbm | flip_sign | 1 | 0.1 | 0.748 | 0.269 | 0.478 | 30 |
| Regression | lightgbm | flip_sign | 1 | 0.3 | 0.721 | 0.421 | 0.301 | 22 |
| Regression | lightgbm | flip_sign | 1 | 0.5 | 0.686 | 0.443 | 0.243 | 12 |
| Regression | lightgbm | flip_sign | 1 | 0.7 | 0.660 | 0.364 | 0.296 | 2 |
| Regression | lightgbm | flip_sign | 2 | 0.1 | 0.719 | 0.291 | 0.428 | 23 |
| Regression | lightgbm | flip_sign | 2 | 0.3 | 0.744 | 0.263 | 0.481 | 13 |
| Regression | lightgbm | flip_sign | 2 | 0.5 | 0.725 | 0.403 | 0.322 | 9 |
| Regression | lightgbm | flip_sign | 2 | 0.7 | 0.677 | 0.414 | 0.262 | 6 |
| Regression | lightgbm | flip_sign | 3 | 0.1 | 0.697 | 0.152 | 0.546 | 15 |
| Regression | lightgbm | flip_sign | 3 | 0.3 | 0.776 | 0.363 | 0.413 | 5 |
| Regression | lightgbm | flip_sign | 3 | 0.5 | 0.720 | 0.384 | 0.336 | 5 |
| Regression | lightgbm | flip_sign | 3 | 0.7 | 0.795 | 0.196 | 0.599 | 2 |
| Regression | lightgbm | flip_sign | 4 | 0.1 | 0.759 | 0.383 | 0.376 | 12 |
| Regression | lightgbm | flip_sign | 4 | 0.3 | 0.742 | 0.446 | 0.296 | 6 |
| Regression | lightgbm | flip_sign | 4 | 0.5 | 0.716 | 0.540 | 0.176 | 5 |
| Regression | lightgbm | flip_sign | 4 | 0.7 | 0.784 | 0.646 | 0.138 | 2 |
| Regression | lightgbm | flip_sign | 5 | 0.1 | 0.558 | 0.067 | 0.491 | 8 |
| Regression | lightgbm | flip_sign | 5 | 0.3 | 0.460 | 0.244 | 0.216 | 3 |
| Regression | lightgbm | flip_sign | 5 | 0.5 | 0.899 | 0.688 | 0.211 | 1 |
| Regression | mlp | amplify | 0 | 0.1 | 0.457 | 0.110 | 0.347 | 16 |
| Regression | mlp | amplify | 0 | 0.3 | 0.675 | 0.445 | 0.230 | 11 |
| Regression | mlp | amplify | 0 | 0.5 | 0.275 | -0.012 | 0.287 | 9 |
| Regression | mlp | amplify | 0 | 0.7 | -0.477 | -0.869 | 0.392 | 1 |
| Regression | mlp | amplify | 1 | 0.1 | 0.642 | 0.252 | 0.390 | 58 |
| Regression | mlp | amplify | 1 | 0.3 | 0.531 | 0.145 | 0.385 | 22 |
| Regression | mlp | amplify | 1 | 0.5 | 0.489 | 0.227 | 0.262 | 7 |

Continued on next page

Table 1 – Continued from previous page

| Task | Method | Operation | Bin | MinFrac | IID | OOD | Degradation | Shifts |
|---|---|---|---|---|---|---|---|---|
| Regression | mlp | amplify | 1 | 0.7 | 0.191 | -0.034 | 0.226 | 8 |
| Regression | mlp | amplify | 1 | 0.9 | -0.496 | -0.766 | 0.270 | 1 |
| Regression | mlp | amplify | 2 | 0.1 | 0.598 | 0.124 | 0.473 | 36 |
| Regression | mlp | amplify | 2 | 0.3 | 0.386 | 0.062 | 0.324 | 11 |
| Regression | mlp | amplify | 2 | 0.7 | 0.798 | 0.695 | 0.103 | 1 |
| Regression | mlp | amplify | 3 | 0.1 | 0.450 | -0.039 | 0.489 | 20 |
| Regression | mlp | amplify | 3 | 0.3 | 0.445 | -0.174 | 0.619 | 7 |
| Regression | mlp | amplify | 3 | 0.5 | 0.404 | -0.104 | 0.509 | 5 |
| Regression | mlp | amplify | 3 | 0.7 | 0.224 | 0.001 | 0.223 | 5 |
| Regression | mlp | amplify | 4 | 0.1 | 0.582 | 0.350 | 0.231 | 40 |
| Regression | mlp | amplify | 4 | 0.3 | 0.601 | 0.271 | 0.330 | 22 |
| Regression | mlp | amplify | 4 | 0.5 | 0.479 | 0.310 | 0.169 | 25 |
| Regression | mlp | amplify | 4 | 0.7 | 0.547 | 0.277 | 0.270 | 7 |
| Regression | mlp | amplify | 5 | 0.1 | 0.466 | 0.174 | 0.292 | 32 |
| Regression | mlp | amplify | 5 | 0.3 | 0.337 | 0.128 | 0.209 | 15 |
| Regression | mlp | amplify | 5 | 0.5 | 0.127 | -0.159 | 0.286 | 14 |
| Regression | mlp | amplify | 5 | 0.7 | 0.404 | 0.083 | 0.321 | 4 |
| Regression | mlp | attenuate | 0 | 0.1 | 0.550 | 0.083 | 0.466 | 96 |
| Regression | mlp | attenuate | 0 | 0.3 | 0.558 | 0.025 | 0.533 | 54 |
| Regression | mlp | attenuate | 0 | 0.5 | 0.553 | -0.068 | 0.620 | 40 |
| Regression | mlp | attenuate | 0 | 0.7 | 0.609 | 0.149 | 0.460 | 19 |
| Regression | mlp | attenuate | 0 | 0.9 | 0.456 | 0.095 | 0.360 | 9 |
| Regression | mlp | attenuate | 1 | 0.1 | 0.523 | 0.160 | 0.362 | 90 |
| Regression | mlp | attenuate | 1 | 0.3 | 0.631 | -0.060 | 0.691 | 29 |
| Regression | mlp | attenuate | 1 | 0.5 | 0.357 | -0.059 | 0.416 | 19 |
| Regression | mlp | attenuate | 1 | 0.7 | 0.515 | 0.108 | 0.407 | 10 |
| Regression | mlp | attenuate | 1 | 0.9 | -0.477 | -0.751 | 0.274 | 1 |
| Regression | mlp | attenuate | 2 | 0.1 | 0.437 | 0.028 | 0.409 | 57 |
| Regression | mlp | attenuate | 2 | 0.3 | 0.535 | 0.073 | 0.462 | 23 |
| Regression | mlp | attenuate | 2 | 0.5 | 0.255 | -0.176 | 0.431 | 19 |
| Regression | mlp | attenuate | 2 | 0.7 | 0.352 | -0.178 | 0.529 | 11 |
| Regression | mlp | attenuate | 2 | 0.9 | 0.187 | -0.046 | 0.233 | 4 |
| Regression | mlp | attenuate | 3 | 0.1 | 0.411 | 0.053 | 0.357 | 47 |
| Regression | mlp | attenuate | 3 | 0.3 | 0.337 | -0.073 | 0.410 | 16 |
| Regression | mlp | attenuate | 3 | 0.5 | 0.284 | -0.297 | 0.581 | 11 |
| Regression | mlp | attenuate | 3 | 0.7 | 0.044 | -0.132 | 0.176 | 10 |
| Regression | mlp | attenuate | 4 | 0.1 | 0.392 | 0.080 | 0.312 | 36 |
| Regression | mlp | attenuate | 4 | 0.3 | 0.403 | 0.217 | 0.185 | 17 |
| Regression | mlp | attenuate | 4 | 0.5 | 0.044 | -0.135 | 0.179 | 10 |
| Regression | mlp | attenuate | 4 | 0.7 | 0.232 | 0.038 | 0.195 | 9 |
| Regression | mlp | attenuate | 4 | 0.9 | 0.111 | 0.001 | 0.110 | 3 |
| Regression | mlp | attenuate | 5 | 0.1 | 0.332 | 0.068 | 0.265 | 30 |
| Regression | mlp | attenuate | 5 | 0.3 | 0.272 | 0.089 | 0.183 | 12 |
| Regression | mlp | attenuate | 5 | 0.5 | 0.167 | -0.033 | 0.200 | 6 |
| Regression | mlp | attenuate | 5 | 0.7 | 0.238 | 0.053 | 0.185 | 4 |
| Regression | mlp | attenuate | 5 | 0.9 | -0.496 | -0.606 | 0.110 | 1 |
| Regression | mlp | flip_sign | 0 | 0.1 | 0.513 | 0.003 | 0.510 | 16 |
| Regression | mlp | flip_sign | 0 | 0.3 | 0.586 | 0.108 | 0.479 | 10 |
| Regression | mlp | flip_sign | 0 | 0.5 | 0.388 | -0.047 | 0.435 | 8 |
| Regression | mlp | flip_sign | 0 | 0.7 | 0.404 | -0.407 | 0.812 | 1 |
| Regression | mlp | flip_sign | 0 | 0.9 | 0.404 | -0.470 | 0.874 | 1 |
| Regression | mlp | flip_sign | 1 | 0.1 | 0.559 | 0.095 | 0.464 | 22 |
| Regression | mlp | flip_sign | 1 | 0.3 | 0.612 | 0.224 | 0.388 | 12 |
| Regression | mlp | flip_sign | 1 | 0.5 | 0.524 | 0.244 | 0.280 | 6 |
| Regression | mlp | flip_sign | 1 | 0.7 | 0.798 | 0.651 | 0.147 | 1 |
| Regression | mlp | flip_sign | 2 | 0.1 | 0.599 | 0.286 | 0.313 | 9 |
| Regression | mlp | flip_sign | 2 | 0.3 | 0.725 | -0.085 | 0.809 | 4 |
| Regression | mlp | flip_sign | 2 | 0.5 | 0.792 | 0.486 | 0.307 | 2 |
| Regression | mlp | flip_sign | 2 | 0.7 | 0.965 | 0.736 | 0.229 | 1 |
| Regression | mlp | flip_sign | 2 | 0.9 | 0.737 | 0.625 | 0.112 | 1 |
| Regression | mlp | flip_sign | 3 | 0.1 | 0.538 | 0.198 | 0.340 | 13 |
| Regression | mlp | flip_sign | 3 | 0.3 | 0.457 | 0.104 | 0.353 | 6 |
| Regression | mlp | flip_sign | 3 | 0.5 | 0.321 | -0.058 | 0.379 | 2 |
| Regression | mlp | flip_sign | 3 | 0.7 | 0.321 | -0.053 | 0.374 | 2 |
| Regression | mlp | flip_sign | 4 | 0.1 | 0.410 | 0.131 | 0.279 | 8 |
| Regression | mlp | flip_sign | 4 | 0.3 | 0.523 | 0.349 | 0.174 | 5 |
| Regression | mlp | flip_sign | 4 | 0.5 | 0.404 | 0.233 | 0.172 | 1 |
| Regression | mlp | flip_sign | 4 | 0.7 | 0.404 | 0.158 | 0.246 | 1 |
| Regression | mlp | flip_sign | 5 | 0.1 | 0.315 | 0.002 | 0.313 | 9 |
| Regression | mlp | flip_sign | 5 | 0.3 | 0.511 | 0.173 | 0.338 | 3 |
| Regression | mlp | flip_sign | 5 | 0.5 | 0.103 | -0.159 | 0.262 | 3 |
| Regression | random_forest | amplify | 0 | 0.1 | 0.902 | 0.201 | 0.701 | 59 |
| Regression | random_forest | amplify | 0 | 0.3 | 0.828 | 0.516 | 0.312 | 24 |
| Regression | random_forest | amplify | 0 | 0.5 | 0.903 | 0.524 | 0.379 | 22 |
| Regression | random_forest | amplify | 0 | 0.7 | 0.675 | 0.434 | 0.241 | 2 |
| Regression | random_forest | amplify | 1 | 0.1 | 0.715 | 0.155 | 0.560 | 104 |
| Regression | random_forest | amplify | 1 | 0.3 | 0.745 | 0.455 | 0.290 | 40 |

Continued on next page

Table 1 – Continued from previous page

| Task | Method | Operation | Bin | MinFrac | IID | OOD | Degradation | Shifts |
|---|---|---|---|---|---|---|---|---|
| Regression | random_forest | amplify | 1 | 0.5 | 0.741 | 0.446 | 0.295 | 16 |
| Regression | random_forest | amplify | 1 | 0.7 | 0.758 | 0.525 | 0.233 | 9 |
| Regression | random_forest | amplify | 2 | 0.1 | 0.710 | 0.186 | 0.524 | 69 |
| Regression | random_forest | amplify | 2 | 0.3 | 0.672 | 0.372 | 0.300 | 40 |
| Regression | random_forest | amplify | 2 | 0.5 | 0.730 | 0.428 | 0.302 | 20 |
| Regression | random_forest | amplify | 2 | 0.7 | 0.707 | 0.314 | 0.393 | 14 |
| Regression | random_forest | amplify | 2 | 0.9 | 0.996 | 0.796 | 0.200 | 4 |
| Regression | random_forest | amplify | 3 | 0.1 | 0.647 | 0.136 | 0.511 | 32 |
| Regression | random_forest | amplify | 3 | 0.3 | 0.652 | -0.045 | 0.696 | 6 |
| Regression | random_forest | amplify | 3 | 0.5 | 0.652 | 0.457 | 0.195 | 6 |
| Regression | random_forest | amplify | 4 | 0.1 | 0.607 | 0.175 | 0.431 | 39 |
| Regression | random_forest | amplify | 4 | 0.3 | 0.529 | 0.194 | 0.336 | 17 |
| Regression | random_forest | amplify | 4 | 0.5 | 0.775 | 0.629 | 0.145 | 9 |
| Regression | random_forest | amplify | 4 | 0.7 | 0.595 | 0.481 | 0.114 | 4 |
| Regression | random_forest | amplify | 5 | 0.1 | 0.679 | 0.154 | 0.525 | 30 |
| Regression | random_forest | amplify | 5 | 0.3 | 0.623 | 0.335 | 0.288 | 17 |
| Regression | random_forest | amplify | 5 | 0.5 | 0.757 | 0.623 | 0.134 | 12 |
| Regression | random_forest | amplify | 5 | 0.7 | 0.595 | 0.490 | 0.105 | 4 |
| Regression | random_forest | attenuate | 0 | 0.1 | 0.727 | 0.183 | 0.545 | 171 |
| Regression | random_forest | attenuate | 0 | 0.3 | 0.749 | 0.195 | 0.554 | 109 |
| Regression | random_forest | attenuate | 0 | 0.5 | 0.788 | 0.194 | 0.594 | 87 |
| Regression | random_forest | attenuate | 0 | 0.7 | 0.804 | 0.352 | 0.452 | 49 |
| Regression | random_forest | attenuate | 0 | 0.9 | 0.882 | 0.336 | 0.545 | 19 |
| Regression | random_forest | attenuate | 1 | 0.1 | 0.733 | 0.455 | 0.278 | 117 |
| Regression | random_forest | attenuate | 1 | 0.3 | 0.697 | 0.402 | 0.296 | 45 |
| Regression | random_forest | attenuate | 1 | 0.5 | 0.886 | 0.714 | 0.172 | 7 |
| Regression | random_forest | attenuate | 1 | 0.7 | 0.839 | 0.696 | 0.143 | 2 |
| Regression | random_forest | attenuate | 1 | 0.9 | 0.340 | 0.209 | 0.131 | 1 |
| Regression | random_forest | attenuate | 2 | 0.1 | 0.708 | 0.355 | 0.353 | 71 |
| Regression | random_forest | attenuate | 2 | 0.3 | 0.818 | 0.562 | 0.255 | 23 |
| Regression | random_forest | attenuate | 2 | 0.5 | 0.846 | 0.650 | 0.196 | 18 |
| Regression | random_forest | attenuate | 2 | 0.7 | 0.929 | 0.697 | 0.231 | 11 |
| Regression | random_forest | attenuate | 3 | 0.1 | 0.713 | 0.184 | 0.529 | 49 |
| Regression | random_forest | attenuate | 3 | 0.3 | 0.877 | -0.302 | 1.179 | 11 |
| Regression | random_forest | attenuate | 3 | 0.5 | 0.953 | 0.181 | 0.772 | 8 |
| Regression | random_forest | attenuate | 3 | 0.7 | 0.972 | -0.262 | 1.234 | 3 |
| Regression | random_forest | attenuate | 4 | 0.1 | 0.658 | 0.392 | 0.266 | 38 |
| Regression | random_forest | attenuate | 4 | 0.3 | 0.972 | 0.788 | 0.184 | 1 |
| Regression | random_forest | attenuate | 5 | 0.1 | 0.633 | 0.428 | 0.205 | 20 |
| Regression | random_forest | attenuate | 5 | 0.3 | 0.303 | 0.158 | 0.146 | 4 |
| Regression | random_forest | attenuate | 5 | 0.5 | 0.303 | 0.198 | 0.105 | 1 |
| Regression | random_forest | flip_sign | 0 | 0.1 | 0.763 | -0.005 | 0.768 | 29 |
| Regression | random_forest | flip_sign | 0 | 0.3 | 0.770 | 0.124 | 0.646 | 19 |
| Regression | random_forest | flip_sign | 0 | 0.5 | 0.705 | 0.131 | 0.574 | 13 |
| Regression | random_forest | flip_sign | 0 | 0.7 | 0.771 | 0.292 | 0.479 | 5 |
| Regression | random_forest | flip_sign | 0 | 0.9 | 0.979 | 0.206 | 0.774 | 2 |
| Regression | random_forest | flip_sign | 1 | 0.1 | 0.759 | 0.318 | 0.440 | 32 |
| Regression | random_forest | flip_sign | 1 | 0.3 | 0.737 | 0.405 | 0.331 | 17 |
| Regression | random_forest | flip_sign | 1 | 0.5 | 0.690 | 0.356 | 0.335 | 10 |
| Regression | random_forest | flip_sign | 1 | 0.7 | 0.736 | 0.504 | 0.232 | 3 |
| Regression | random_forest | flip_sign | 2 | 0.1 | 0.702 | 0.300 | 0.402 | 26 |
| Regression | random_forest | flip_sign | 2 | 0.3 | 0.765 | 0.242 | 0.523 | 12 |
| Regression | random_forest | flip_sign | 2 | 0.5 | 0.729 | 0.385 | 0.344 | 10 |
| Regression | random_forest | flip_sign | 2 | 0.7 | 0.769 | 0.424 | 0.345 | 5 |
| Regression | random_forest | flip_sign | 3 | 0.1 | 0.689 | 0.230 | 0.459 | 12 |
| Regression | random_forest | flip_sign | 3 | 0.3 | 0.828 | 0.178 | 0.650 | 3 |
| Regression | random_forest | flip_sign | 3 | 0.5 | 0.828 | 0.257 | 0.572 | 3 |
| Regression | random_forest | flip_sign | 3 | 0.7 | 0.784 | 0.045 | 0.738 | 2 |
| Regression | random_forest | flip_sign | 3 | 0.9 | 0.595 | 0.393 | 0.202 | 1 |
| Regression | random_forest | flip_sign | 4 | 0.1 | 0.737 | 0.274 | 0.463 | 10 |
| Regression | random_forest | flip_sign | 4 | 0.3 | 0.872 | 0.662 | 0.210 | 4 |
| Regression | random_forest | flip_sign | 4 | 0.5 | 0.790 | 0.661 | 0.129 | 2 |
| Regression | random_forest | flip_sign | 4 | 0.7 | 0.595 | 0.488 | 0.108 | 1 |
| Regression | random_forest | flip_sign | 4 | 0.9 | -0.069 | -0.216 | 0.147 | 1 |
| Regression | random_forest | flip_sign | 5 | 0.1 | 0.615 | 0.326 | 0.290 | 7 |
| Regression | random_forest | flip_sign | 5 | 0.3 | 0.611 | 0.373 | 0.238 | 2 |
| Regression | random_forest | flip_sign | 5 | 0.5 | 0.918 | 0.696 | 0.222 | 1 |
| Regression | random_forest | flip_sign | 5 | 0.7 | 0.918 | 0.800 | 0.118 | 1 |
| Regression | ridge | amplify | 0 | 0.1 | 0.521 | -0.239 | 0.760 | 43 |
| Regression | ridge | amplify | 0 | 0.3 | 0.560 | -0.362 | 0.922 | 21 |
| Regression | ridge | amplify | 0 | 0.5 | 0.535 | 0.017 | 0.518 | 12 |
| Regression | ridge | amplify | 0 | 0.7 | 0.052 | -0.104 | 0.156 | 1 |
| Regression | ridge | amplify | 1 | 0.1 | 0.549 | 0.237 | 0.312 | 45 |
| Regression | ridge | amplify | 1 | 0.3 | 0.715 | 0.311 | 0.404 | 11 |
| Regression | ridge | amplify | 1 | 0.5 | 0.552 | 0.054 | 0.498 | 4 |
| Regression | ridge | amplify | 1 | 0.7 | 0.234 | -0.023 | 0.257 | 7 |
| Regression | ridge | amplify | 2 | 0.1 | 0.619 | 0.255 | 0.364 | 40 |

Continued on next page

Table 1 – Continued from previous page

| Task | Method | Operation | Bin | MinFrac | IID | OOD | Degradation | Shifts |
|---|---|---|---|---|---|---|---|---|
| Regression | ridge | amplify | 2 | 0.5 | 0.727 | -0.298 | 1.026 | 7 |
| Regression | ridge | amplify | 2 | 0.7 | 0.577 | 0.194 | 0.384 | 9 |
| Regression | ridge | amplify | 2 | 0.9 | 0.052 | -0.677 | 0.729 | 1 |
| Regression | ridge | amplify | 3 | 0.1 | 0.308 | 0.075 | 0.233 | 26 |
| Regression | ridge | amplify | 3 | 0.3 | 0.265 | 0.073 | 0.192 | 11 |
| Regression | ridge | amplify | 3 | 0.5 | 0.284 | -0.054 | 0.338 | 6 |
| Regression | ridge | amplify | 3 | 0.7 | 0.284 | -0.192 | 0.476 | 6 |
| Regression | ridge | amplify | 4 | 0.1 | 0.411 | -0.112 | 0.523 | 15 |
| Regression | ridge | amplify | 4 | 0.3 | 0.248 | -0.126 | 0.374 | 6 |
| Regression | ridge | amplify | 4 | 0.5 | 0.108 | -0.017 | 0.125 | 1 |
| Regression | ridge | amplify | 5 | 0.1 | 0.177 | 0.036 | 0.141 | 8 |
| Regression | ridge | amplify | 5 | 0.3 | 0.215 | -0.548 | 0.763 | 15 |
| Regression | ridge | amplify | 5 | 0.5 | 0.052 | -0.106 | 0.158 | 6 |
| Regression | ridge | amplify | 5 | 0.7 | 0.052 | -0.099 | 0.151 | 6 |
| Regression | ridge | attenuate | 0 | 0.1 | 0.564 | 0.158 | 0.406 | 98 |
| Regression | ridge | attenuate | 0 | 0.3 | 0.609 | 0.211 | 0.398 | 55 |
| Regression | ridge | attenuate | 0 | 0.5 | 0.532 | 0.071 | 0.461 | 44 |
| Regression | ridge | attenuate | 0 | 0.7 | 0.570 | 0.216 | 0.354 | 30 |
| Regression | ridge | attenuate | 0 | 0.9 | 0.546 | 0.143 | 0.402 | 20 |
| Regression | ridge | attenuate | 1 | 0.1 | 0.575 | 0.267 | 0.308 | 43 |
| Regression | ridge | attenuate | 1 | 0.3 | 0.526 | 0.232 | 0.295 | 29 |
| Regression | ridge | attenuate | 1 | 0.5 | 0.543 | 0.318 | 0.225 | 9 |
| Regression | ridge | attenuate | 1 | 0.7 | 0.530 | 0.382 | 0.148 | 4 |
| Regression | ridge | attenuate | 2 | 0.1 | 0.525 | 0.317 | 0.208 | 23 |
| Regression | ridge | attenuate | 2 | 0.3 | 0.824 | 0.629 | 0.195 | 8 |
| Regression | ridge | attenuate | 2 | 0.5 | 0.922 | 0.786 | 0.136 | 4 |
| Regression | ridge | attenuate | 2 | 0.7 | 0.922 | 0.777 | 0.144 | 4 |
| Regression | ridge | attenuate | 3 | 0.1 | 0.501 | -0.043 | 0.544 | 25 |
| Regression | ridge | attenuate | 3 | 0.3 | 0.519 | -0.223 | 0.742 | 11 |
| Regression | ridge | attenuate | 3 | 0.5 | 0.519 | 0.134 | 0.385 | 11 |
| Regression | ridge | attenuate | 3 | 0.7 | 0.633 | 0.185 | 0.447 | 6 |
| Regression | ridge | attenuate | 4 | 0.1 | 0.617 | 0.340 | 0.276 | 17 |
| Regression | ridge | attenuate | 4 | 0.3 | 0.522 | 0.418 | 0.104 | 1 |
| Regression | ridge | attenuate | 5 | 0.1 | 0.384 | 0.010 | 0.373 | 8 |
| Regression | ridge | attenuate | 5 | 0.3 | 0.522 | 0.415 | 0.107 | 3 |
| Regression | ridge | attenuate | 5 | 0.5 | 0.052 | -0.107 | 0.160 | 1 |
| Regression | ridge | flip_sign | 0 | 0.1 | 0.528 | -0.035 | 0.563 | 15 |
| Regression | ridge | flip_sign | 0 | 0.3 | 0.515 | -0.222 | 0.737 | 11 |
| Regression | ridge | flip_sign | 0 | 0.5 | 0.435 | -0.027 | 0.462 | 9 |
| Regression | ridge | flip_sign | 0 | 0.7 | 0.556 | -0.125 | 0.681 | 3 |
| Regression | ridge | flip_sign | 0 | 0.9 | 0.492 | -0.028 | 0.521 | 3 |
| Regression | ridge | flip_sign | 1 | 0.1 | 0.576 | 0.172 | 0.404 | 18 |
| Regression | ridge | flip_sign | 1 | 0.3 | 0.619 | 0.146 | 0.473 | 10 |
| Regression | ridge | flip_sign | 1 | 0.5 | 0.605 | -0.028 | 0.633 | 5 |
| Regression | ridge | flip_sign | 1 | 0.7 | 0.540 | -0.253 | 0.794 | 2 |
| Regression | ridge | flip_sign | 2 | 0.1 | 0.594 | 0.074 | 0.520 | 12 |
| Regression | ridge | flip_sign | 2 | 0.3 | 0.608 | 0.105 | 0.504 | 4 |
| Regression | ridge | flip_sign | 2 | 0.5 | 0.619 | 0.296 | 0.323 | 5 |
| Regression | ridge | flip_sign | 2 | 0.7 | 0.840 | 0.000 | 0.840 | 1 |
| Regression | ridge | flip_sign | 2 | 0.9 | 0.840 | 0.621 | 0.219 | 1 |
| Regression | ridge | flip_sign | 3 | 0.1 | 0.515 | 0.142 | 0.373 | 6 |
| Regression | ridge | flip_sign | 3 | 0.3 | 0.503 | 0.088 | 0.415 | 2 |
| Regression | ridge | flip_sign | 3 | 0.5 | 0.503 | 0.267 | 0.236 | 2 |
| Regression | ridge | flip_sign | 3 | 0.7 | 0.488 | 0.322 | 0.167 | 3 |
| Regression | ridge | flip_sign | 3 | 0.9 | 0.294 | 0.194 | 0.100 | 1 |
| Regression | ridge | flip_sign | 4 | 0.1 | 0.573 | -0.093 | 0.666 | 8 |
| Regression | ridge | flip_sign | 4 | 0.3 | 0.525 | 0.207 | 0.318 | 4 |
| Regression | ridge | flip_sign | 4 | 0.5 | 0.712 | 0.228 | 0.484 | 2 |
| Regression | ridge | flip_sign | 5 | 0.1 | 0.334 | -0.088 | 0.422 | 4 |
| Regression | ridge | flip_sign | 5 | 0.3 | 0.661 | -0.180 | 0.841 | 1 |
| Regression | ridge | flip_sign | 5 | 0.5 | 0.052 | -0.107 | 0.160 | 1 |
| Regression | ridge | flip_sign | 5 | 0.7 | 0.052 | -0.098 | 0.150 | 1 |
| Regression | xgboost | amplify | 0 | 0.1 | 0.839 | 0.336 | 0.504 | 54 |
| Regression | xgboost | amplify | 0 | 0.3 | 0.761 | 0.443 | 0.318 | 19 |
| Regression | xgboost | amplify | 0 | 0.5 | 0.937 | 0.531 | 0.406 | 13 |
| Regression | xgboost | amplify | 1 | 0.1 | 0.714 | 0.265 | 0.449 | 104 |
| Regression | xgboost | amplify | 1 | 0.3 | 0.760 | 0.475 | 0.285 | 39 |
| Regression | xgboost | amplify | 1 | 0.5 | 0.639 | 0.339 | 0.299 | 11 |
| Regression | xgboost | amplify | 1 | 0.7 | 0.585 | 0.364 | 0.222 | 5 |
| Regression | xgboost | amplify | 2 | 0.1 | 0.669 | 0.213 | 0.456 | 71 |
| Regression | xgboost | amplify | 2 | 0.3 | 0.697 | 0.415 | 0.282 | 31 |
| Regression | xgboost | amplify | 2 | 0.5 | 0.689 | 0.430 | 0.259 | 15 |
| Regression | xgboost | amplify | 2 | 0.7 | 0.656 | 0.297 | 0.359 | 12 |
| Regression | xgboost | amplify | 2 | 0.9 | 0.991 | 0.751 | 0.240 | 4 |
| Regression | xgboost | amplify | 3 | 0.1 | 0.679 | 0.207 | 0.473 | 38 |
| Regression | xgboost | amplify | 3 | 0.3 | 0.737 | 0.308 | 0.429 | 8 |
| Regression | xgboost | amplify | 3 | 0.5 | 0.655 | 0.471 | 0.185 | 6 |

Continued on next page

Table 1 – Continued from previous page

| Task | Method | Operation | Bin | MinFrac | IID | OOD | Degradation | Shifts |
|---|---|---|---|---|---|---|---|---|
| Regression | xgboost | amplify | 3 | 0.7 | 0.935 | 0.808 | 0.127 | 1 |
| Regression | xgboost | amplify | 4 | 0.1 | 0.406 | 0.021 | 0.386 | 25 |
| Regression | xgboost | amplify | 4 | 0.3 | 0.496 | 0.217 | 0.279 | 17 |
| Regression | xgboost | amplify | 4 | 0.5 | 0.705 | 0.574 | 0.131 | 9 |
| Regression | xgboost | amplify | 4 | 0.7 | 0.600 | 0.493 | 0.107 | 3 |
| Regression | xgboost | amplify | 4 | 0.9 | -0.052 | -0.172 | 0.120 | 1 |
| Regression | xgboost | amplify | 5 | 0.1 | 0.484 | 0.148 | 0.336 | 27 |
| Regression | xgboost | amplify | 5 | 0.3 | 0.639 | 0.267 | 0.372 | 17 |
| Regression | xgboost | amplify | 5 | 0.5 | 0.745 | 0.603 | 0.142 | 11 |
| Regression | xgboost | amplify | 5 | 0.7 | -0.052 | -0.220 | 0.168 | 1 |
| Regression | xgboost | attenuate | 0 | 0.1 | 0.731 | 0.197 | 0.535 | 174 |
| Regression | xgboost | attenuate | 0 | 0.3 | 0.756 | 0.228 | 0.528 | 104 |
| Regression | xgboost | attenuate | 0 | 0.5 | 0.767 | 0.267 | 0.500 | 80 |
| Regression | xgboost | attenuate | 0 | 0.7 | 0.810 | 0.369 | 0.441 | 47 |
| Regression | xgboost | attenuate | 0 | 0.9 | 0.889 | 0.364 | 0.525 | 19 |
| Regression | xgboost | attenuate | 1 | 0.1 | 0.689 | 0.397 | 0.291 | 113 |
| Regression | xgboost | attenuate | 1 | 0.3 | 0.722 | 0.429 | 0.294 | 34 |
| Regression | xgboost | attenuate | 1 | 0.5 | 0.860 | 0.708 | 0.151 | 16 |
| Regression | xgboost | attenuate | 1 | 0.7 | 0.862 | 0.711 | 0.151 | 8 |
| Regression | xgboost | attenuate | 1 | 0.9 | 0.322 | 0.217 | 0.105 | 1 |
| Regression | xgboost | attenuate | 2 | 0.1 | 0.709 | 0.401 | 0.308 | 72 |
| Regression | xgboost | attenuate | 2 | 0.3 | 0.827 | 0.628 | 0.200 | 24 |
| Regression | xgboost | attenuate | 2 | 0.5 | 0.793 | 0.639 | 0.154 | 18 |
| Regression | xgboost | attenuate | 2 | 0.7 | 0.833 | 0.680 | 0.153 | 13 |
| Regression | xgboost | attenuate | 3 | 0.1 | 0.692 | 0.297 | 0.396 | 47 |
| Regression | xgboost | attenuate | 3 | 0.3 | 0.762 | -0.012 | 0.774 | 12 |
| Regression | xgboost | attenuate | 3 | 0.5 | 0.958 | 0.195 | 0.763 | 8 |
| Regression | xgboost | attenuate | 3 | 0.7 | 0.982 | -0.203 | 1.185 | 3 |
| Regression | xgboost | attenuate | 4 | 0.1 | 0.746 | 0.514 | 0.231 | 30 |
| Regression | xgboost | attenuate | 4 | 0.3 | 0.666 | 0.524 | 0.142 | 2 |
| Regression | xgboost | attenuate | 4 | 0.5 | 0.028 | -0.099 | 0.127 | 2 |
| Regression | xgboost | attenuate | 5 | 0.1 | 0.643 | 0.450 | 0.194 | 21 |
| Regression | xgboost | attenuate | 5 | 0.3 | 0.349 | 0.181 | 0.169 | 4 |
| Regression | xgboost | attenuate | 5 | 0.5 | 0.349 | 0.238 | 0.112 | 2 |
| Regression | xgboost | flip_sign | 0 | 0.1 | 0.774 | 0.053 | 0.721 | 29 |
| Regression | xgboost | flip_sign | 0 | 0.3 | 0.789 | 0.077 | 0.712 | 20 |
| Regression | xgboost | flip_sign | 0 | 0.5 | 0.709 | 0.169 | 0.540 | 13 |
| Regression | xgboost | flip_sign | 0 | 0.7 | 0.825 | 0.304 | 0.521 | 4 |
| Regression | xgboost | flip_sign | 0 | 0.9 | 0.980 | 0.217 | 0.763 | 2 |
| Regression | xgboost | flip_sign | 1 | 0.1 | 0.714 | 0.291 | 0.423 | 32 |
| Regression | xgboost | flip_sign | 1 | 0.3 | 0.759 | 0.473 | 0.286 | 18 |
| Regression | xgboost | flip_sign | 1 | 0.5 | 0.762 | 0.504 | 0.259 | 10 |
| Regression | xgboost | flip_sign | 1 | 0.7 | 0.733 | 0.514 | 0.219 | 3 |
| Regression | xgboost | flip_sign | 2 | 0.1 | 0.766 | 0.409 | 0.357 | 24 |
| Regression | xgboost | flip_sign | 2 | 0.3 | 0.771 | 0.321 | 0.450 | 11 |
| Regression | xgboost | flip_sign | 2 | 0.5 | 0.810 | 0.484 | 0.327 | 8 |
| Regression | xgboost | flip_sign | 2 | 0.7 | 0.785 | 0.470 | 0.316 | 4 |
| Regression | xgboost | flip_sign | 3 | 0.1 | 0.650 | 0.154 | 0.496 | 11 |
| Regression | xgboost | flip_sign | 3 | 0.3 | 0.834 | 0.248 | 0.586 | 3 |
| Regression | xgboost | flip_sign | 3 | 0.5 | 0.834 | 0.401 | 0.433 | 3 |
| Regression | xgboost | flip_sign | 3 | 0.7 | 0.791 | 0.240 | 0.551 | 2 |
| Regression | xgboost | flip_sign | 4 | 0.1 | 0.744 | 0.289 | 0.455 | 9 |
| Regression | xgboost | flip_sign | 4 | 0.3 | 0.836 | 0.603 | 0.233 | 3 |
| Regression | xgboost | flip_sign | 4 | 0.5 | 0.604 | 0.460 | 0.144 | 2 |
| Regression | xgboost | flip_sign | 4 | 0.7 | 0.600 | 0.494 | 0.106 | 1 |
| Regression | xgboost | flip_sign | 4 | 0.9 | -0.052 | -0.227 | 0.175 | 1 |
| Regression | xgboost | flip_sign | 5 | 0.1 | 0.505 | 0.233 | 0.272 | 9 |
| Regression | xgboost | flip_sign | 5 | 0.3 | 0.635 | 0.283 | 0.352 | 2 |
| Regression | xgboost | flip_sign | 5 | 0.5 | 0.920 | 0.690 | 0.230 | 1 |
| Regression | xgboost | flip_sign | 5 | 0.7 | 0.920 | 0.812 | 0.108 | 1 |