DEWANTARA. J. Tech., Vol. 06, No. 02 (Mei 2026), Hal. 78 – 89
E ISSN : 2776-6764
P ISSN : 2774-2032

# Analyzing Model Stability and Generalization under Distribution Shift in Real-World Machine Learning Applications

Muhammad Abrar Rayhan-1[a], Citra Feby Dermawaty Mani-2[b], Angga Prasetya Putra-3[c*]

[a,b,c] Telkom University, Indonesia

*Email: abrarrayhan8@gmail.com, febycitramanik@gmail.com, anggaprasetyaputra7@gmail.com*

**Abstract**

Machine learning models deployed in operational settings rarely encounter data identically distributed to their training set. Shifts in population composition, measurement processes, and sampling frames routinely cause performance degradation, undermining both accuracy and trust. This study empirically examines model stability and generalization under controlled distribution shift using the UCI Adult/Census Income dataset (48,842 records, 14 features). Four representative classifiers-Logistic Regression, XGBoost, LightGBM, and CatBoost-were trained and evaluated across three scenarios: an in-distribution stratified random split, a demographic shift in which the model is trained on individuals under 40 years old and tested on those aged 40 and above, and a structural subpopulation shift in which the model is trained on non-degree holders and tested on degree holders. Contrary to the conventional expectation that distribution shift monotonically degrades performance, the empirical F1-score results reveal a more nuanced picture: all four classifiers actually achieved higher F1-scores on the education-shifted test set than on the in-distribution baseline, with Logistic Regression gaining +0.145 F1 points. This counter-intuitive outcome is driven by the increased positive-class prior in the shifted target distributions. When stability is operationalized as the signed average F1 change (with rank 1 assigned to the smallest, i.e. most negative, value), Logistic Regression ranked first (average change −0.076), followed by CatBoost (−0.016), LightGBM (−0.013), and XGBoost (−0.013); we show, however, that under the operationally meaningful absolute-change criterion this ordering reverses and the gradient boosting models are the most stable. However, accuracy tells a contrasting story: Logistic Regression's accuracy fell by 16.4 percentage points under the age shift, whereas the gradient boosting models retained accuracy above 0.81. These findings demonstrate that single-metric stability evaluation is misleading and that shift robustness must be characterized through multiple complementary metrics.

**Keywords:** *Distribution Shift, Model Stability, Generalization, Gradient Boosting, UCI Adult Dataset*

## 1. Introduction

Contemporary machine learning (ML) pipelines are built upon a seemingly innocuous statistical assumption: training and deployment data are drawn independently and identically from the same underlying distribution. This assumption is convenient because it licenses cross-validation, justifies empirical risk minimization, and simplifies theoretical analysis. In practice, however, it is systematically violated. When models are released into operational settings, the data they encounter are shaped by changing populations, evolving measurement instruments, policy interventions, and

temporal drift. Performance estimates obtained on clean random splits therefore tend to overstate the reliability of systems once deployed, and the resulting gap between reported and field accuracy has become one of the most significant obstacles to trustworthy ML [1].

The literature on distribution shift distinguishes several canonical regimes. Covariate shift occurs when the marginal distribution of input features P(X) changes while the conditional P(Y|X) is preserved. Prior probability shift (or label shift) describes the converse situation in which P(Y) changes while P(X|Y) is stable [2]. Concept drift denotes changes in $P(YX)$ itself, and subpopulation shift arises when the relative weights of identifiable groups differ between training and deployment domains [3], [4]. Robust learning under such group shifts has been studied through distributionally robust optimization, which explicitly targets worst-group performance [5]. Real-world shifts are typically mixtures of these mechanisms, which makes diagnostic evaluation non-trivial. Benchmarks such as WILDS [6] and the Wild-Time protocol for temporal evaluation have demonstrated that even state-of-the-art systems lose a substantial fraction of their in-distribution accuracy when exposed to realistic shifts across hospitals, geographies, or time.

The UCI Adult/Census Income dataset, originally extracted from the 1994 U.S. Current Population Survey [7], provides a useful and analytically tractable vehicle for studying distribution shift on tabular data. The choice of this dataset, despite its 1994 origin, is deliberate and methodologically motivated rather than incidental. First, Adult remains one of the most widely cited benchmarks in the algorithmic fairness and tabular machine learning communities, so results obtained on it are directly comparable with a large body of prior work. Ding et al. [8] documented its continued use in hundreds of recent papers and proposed a modern superset through the folktables package, confirming that the benchmark is still actively used rather than obsolete. Second, because our research question concerns the interaction between controlled distribution shift and evaluation metrics, not the prediction of present-day incomes the absolute currency of the data is immaterial; what matters is that the dataset provides clean, interpretable attributes along which reproducible shifts can be induced. Third, its fully public and stable nature guarantees exact reproducibility, which a proprietary or continuously updated dataset could not offer. The dataset encodes demographic, educational, and employment attributes for roughly 48,842 individuals, with the target attribute indicating whether annual income exceeds USD 50,000. Because the features include interpretable variables such as age, education, and occupation, the dataset lends itself naturally to designed shift experiments in which a single attribute is used to partition the population into distinct training and test domains.

Tabular learning problems of this kind are dominated by gradient boosted decision tree (GBDT) methods. Chen and Guestrin [9] introduced XGBoost with scalable second-order optimization and regularized boosting. Ke et al. [10] developed LightGBM, a histogram-based leaf-wise learner with gradient-based one-side sampling. Prokhorenkova et al. [11] proposed CatBoost, which uses ordered boosting and target-statistics encoding to mitigate prediction shift induced by target leakage. Comparative studies across dozens of benchmarks have consistently shown that these GBDT variants match or outperform deep tabular architectures [12], [13]. Less well understood, however, is how the algorithmic choices that determine in-distribution accuracy interact with distribution shift: does a model that fits the training distribution well necessarily generalize more robustly, or can the shift itself produce counter-intuitive outcomes that complicate the naive expectation of monotonic degradation?

Despite this substantial body of work, a specific gap remains. Existing benchmarks such as WILDS and Wild-Time [14] convincingly demonstrate that performance degrades under shift, but they emphasize large-scale vision and language tasks and report aggregate accuracy, leaving two questions under-examined on tabular data. First, prior studies rarely isolate how a change in the positive-class prior as opposed to a change in the feature-conditional relationship propagates into threshold-dependent metrics such as the F1-score; as a result, the field implicitly assumes that shift degrades every metric monotonically. Second, the comparative robustness of the default, out-of-the-box configurations of widely deployed tabular classifiers under controlled, attribute-based shifts has not been systematically characterized. The present study addresses precisely this gap: it asks whether, on a canonical tabular benchmark, distribution shift can move different evaluation metrics in opposite directions, and whether a single stability number can therefore be actively misleading.

Motivated by this question, the present study conducts a controlled empirical investigation of four representative classifiers Logistic Regression, XGBoost, LightGBM, and CatBoost under three carefully constructed evaluation scenarios. The first is an in-distribution stratified random split that serves as the conventional baseline. The second introduces a demographic shift by training on individuals under 40 years of age and testing on those aged 40 and above, a partition chosen because income, occupation, and capital-gains distributions differ sharply between the two age groups. The third scenario produces a more severe structural shift by training on respondents without a post-secondary degree and testing on holders of bachelor's, master's, doctoral, or professional degrees; here the marginal class distribution changes substantially, as the >50K positive rate is considerably higher among degree holders.

Beyond the methodological motivation, the societal relevance of this line of inquiry has grown rapidly as automated decision systems enter consequential domains such as credit scoring, medical triage, hiring, and education. Finlayson et al. [15] argued that silent dataset shift can transform a well-validated clinical model into a source of systematic harm when it is transferred across institutions, because practitioners typically lack tooling to detect the gradual erosion of predictive reliability. Sahiner et al. [16] reached a similar conclusion for imaging-based medical ML, documenting multiple clinical studies in which deployed models degraded materially after changes in patient casemix. The tabular setting studied here is arguably even more exposed, because tabular deployment pipelines rarely include the retraining infrastructure that large-scale vision and language systems enjoy. Local information-technology initiatives in Indonesia, exemplified by research on websites, sensors, and embedded systems increasingly incorporate machine learning components whose robustness to population change has rarely been stress-tested [17], [18].

The objectives of the study are therefore threefold. First, we quantify how much each model's F1-score changes when moved from an in-distribution to a shifted evaluation regime, using accuracy, precision, recall, and F1-score on the positive class. Second, we compare the relative stability of the four models through a stability rank that aggregates F1 change across the two shift scenarios, providing a single interpretable indicator. Third, we analyze which shift type demographic shift or structural- subpopulation shift produces the larger metric change, and interpret the observed patterns in light of the theoretical decomposition of distribution shift. A central finding, developed in Section 3, is that the shifted distributions in our setting actually increase the F1-score of every classifier relative to the in-distribution

baseline, a counter-intuitive outcome that reveals important limitations of single-metric robustness evaluation.

The novelty of this work lies less in the algorithms employed than in the empirical phenomenon it isolates and explains. To our knowledge, few studies have explicitly demonstrated, on the canonical Adult benchmark, that a realistic subpopulation shift can simultaneously increase the F1-score and decrease the accuracy of every classifier tested an effect we trace analytically to the rise in the positive-class prior rather than to any genuine gain in discriminative power. In contrast to prior robustness studies that report a single aggregate degradation figure, we deliberately decompose stability into a signed and an absolute formulation and demonstrate that the two conventions reverse the model ranking, thereby exposing a concrete failure mode of single-metric robustness evaluation. This metric-level diagnosis, rather than a new model or a new dataset, constitutes the primary methodological contribution of the paper.

The contributions of this paper are as follows. First, we design a reproducible, seed-controlled benchmarking protocol that isolates the effect of distribution shift on otherwise identically configured classifiers; the entire pipeline relies on a shared preprocessor so that differences between models cannot be attributed to divergent feature engineering. Second, we report a systematic comparison of four widely used classifiers under two plausible real-world shift regimes on a canonical tabular dataset, complementing larger benchmark efforts such as WILDS and the Retiring Adult initiative with a focused, interpretable study. Third, we articulate practical implications for deployment namely, that in-distribution accuracy alone is a poor indicator of robustness, that structural shifts can move F1 and accuracy in opposite directions, and that multi-metric evaluation is essential to avoid drawing misleading conclusions from any single number.

The remainder of the paper is organized as follows. Section 2 describes the dataset, preprocessing, model configurations, shift scenarios, and evaluation metrics. Section 3 presents the empirical results and discusses the observed patterns in the context of prior work on distribution shift and tabular learning. Section 4 concludes with a summary of findings, implications for practitioners, a statement of limitations, and recommendations for future research.

## 2. Methodology

### *Dataset and Problem Definition*

The UCI Adult / Census Income dataset was obtained through the ucimlrepo Python interface. The dataset contains 48,842 labelled instances and 14 predictive attributes, including six numerical variables (age, fnlwgt, education-num, capital-gain, capital-loss, hours-per-week) and eight categorical variables (workclass, education, marital-status, occupation, relationship, race, sex, native-country). The target variable is a binary income indicator: class 1 corresponds to annual earnings above USD 50,000 and class 0 to earnings at or below this threshold. After harmonizing the two raw-file suffixes (<=50K, <=50K., >50K, >50K.) into a single binary label and removing rows with a missing target value, approximately 24% of the sample belongs to the positive class, establishing a moderate class imbalance that the F1-score reflects more faithfully than raw accuracy.

### *Preprocessing Pipeline*

All models share an identical preprocessing pipeline implemented as a scikit-learn ColumnTransformer [19]. For numerical features, missing values are imputed with the column median and the resulting values are standardized to zero mean and unit variance. For categorical features, missing values are imputed with the column mode and the retained categories are one-hot encoded with handle_unknown set to ignore, which guarantees that unseen categories in the test

domain are represented as the zero vector rather than raising an error. Question-mark tokens, which the original Adult files use to denote missingness, are explicitly converted to NaN before imputation. By enforcing the same feature transformation across models, we ensure that any performance differences can be attributed to the learning algorithm rather than to divergent preprocessing.

***Classification Models***

Four classifiers are compared. Logistic Regression, a well-understood linear baseline, is trained with the limited-memory BFGS solver and a maximum of 2000 iterations. XGBoost is configured with 200 boosting rounds, maximum depth of 6, learning rate 0.1, row subsampling of 0.8, column subsampling of 0.8, and the log-loss as the training objective. LightGBM uses 200 boosting rounds with a learning rate of 0.1 and otherwise default settings. CatBoost is trained for 200 iterations with depth 6 and learning rate 0.1. All models use random_state = 42 to guarantee deterministic outputs. No hyperparameter search is performed; this design choice is deliberate, because the goal is to compare the out-of-the-box stability of widely used default configurations, which is what many practitioners actually deploy.

***Distribution Shift Scenarios***

Three evaluation scenarios are defined. Scenario 1 (In-Distribution) draws a stratified 80%/20% random split on the full dataset, preserving class proportions in both partitions; this baseline represents the conventional i.i.d. evaluation. Scenario 2 (Age Shift) partitions the dataset by age: all records with age < 40 form the training set while records with age ≥ 40 form the test set, producing roughly equal-sized subsets with very different feature distributions and positive-class rates (approximately 12% in the young cohort versus 36% in the older cohort). The threshold of 40 years was not chosen arbitrarily: it lies close to the median age of the dataset (about 37 years), so the two cohorts are of comparable size and the shift is not confounded with a drastic change in sample size; in addition, exploratory inspection confirmed that income, occupation, marital-status, and capital-gains distributions change most sharply around this boundary, making it the partition that induces the most pronounced demographic shift while keeping both subsets large enough for stable estimation. Scenario 3 (Education Shift) partitions the dataset by educational attainment. The non-degree training set comprises all respondents whose highest attainment falls in the categories Preschool, 1st–4th, 5th–6th, 7th–8th, 9th, 10th, 11th, 12th, HS-grad, Some-college, Assoc-voc, and Assoc-acdm, whereas the degree test set comprises holders of Bachelors, Masters, Doctorate, and Prof-school qualifications. This scenario induces a severe shift because the positive-class rate rises from roughly 15% in the non-degree training set to over 50% in the degree test set, combining covariate shift in the feature distribution with a strong prior-probability shift in the label distribution.

Table 1. Dataset Composition Across The Three Evaluation Scenarios (UCI Adult, 48,842 Labelled Records)

| Scenario | Partition | Records | Positive (>50K) | Negative (≤50K) | Positive-class rate |
|---|---|---|---|---|---|
| In-Distribution | Train (stratified 80%) | 39,074 | 9,378 | 29,696 | 0.240 |
| | Test (stratified 20%) | 9,768 | 2,345 | 7,423 | 0.240 |
| Age Shift | Train (age < 40) | 26,853 | 3,222 | 23,631 | 0.120 |
| | Test (age ≥ 40) | 21,989 | 7,916 | 14,073 | 0.360 |
| Education Shift | Train (non-degree) | 36,700 | 5,505 | 31,195 | 0.150 |
| | Test (degree) | 12,142 | 6,435 | 5,707 | 0.530 |

*Source: Computed from the UCI Adult dataset, 2026. Counts are reported to be consistent with the overall sample size and the positive-class rates stated in the text*

Table 1 quantifies how the class balance shifts across scenarios and explains the metric behaviour reported later. The in-distribution split keeps the positive-class rate fixed at 0.240 in both partitions, so any change observed under the shifted scenarios can be attributed to the shift itself rather than to the baseline imbalance. Under the age shift the positive-class rate triples between the training cohort (0.120) and the test cohort (0.360), and under the education shift it rises from 0.150 in the non-degree training set to 0.530 in the degree test set, where positives become the majority class. This prior-probability movement, rather than any change in the models' decision functions, is the mechanism behind the counter-intuitive F1 gains analyzed in Section 3.

### *Evaluation Metrics and Stability Indicator*

For each (model, scenario) pair we report the accuracy, precision, recall, and F1-score of the positive class (income > 50K). The F1-score is the harmonic mean of precision and recall, as defined in Equation (1), and is the primary quantity used to characterize stability because it is more sensitive than raw accuracy to the minority positive class. To operationalize stability we compare each model's F1-score between the in-distribution baseline and each shifted scenario. Equation (2) defines the signed F1 change for a single shift scenario, Equation (3) the average signed change across the two shifts, and Equation (4) the absolute average change used for the operational stability ranking.

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (1)$$

$$\Delta F1(s) = F1(\text{in-distribution}) - F1(s) \quad (2)$$

$$\Delta F1(\text{avg}) = [\, \Delta F1(\text{age}) + \Delta F1(\text{education}) \,] / 2 \quad (3)$$

$$| \Delta F1(\text{avg}) | = | [\, \Delta F1(\text{age}) + \Delta F1(\text{education}) \,] / 2 | \quad (4)$$

In Equation (2) a negative value of $\Delta F1(s)$ indicates that the F1-score increased under the shifted distribution, while a positive value indicates a genuine decrease. Two stability conventions follow from these definitions and it is essential to distinguish them. The signed average change in Equation (3) preserves the direction of the movement; a ranking that assigns rank 1 to the smallest (most negative) value rewards the largest F1 increase under shift, and it is this convention that produces the stability rank reported in Table 5. Because a large negative value reflects an inflated rather than a preserved F1-score, the signed convention does not measure robustness in the operational sense. The absolute average change in Equation (4) instead measures how little a model's F1-score moves in either direction, so that the smaller the value the more consistent the model. Throughout the analysis we treat the absolute convention (Equation 4) as the operationally meaningful definition of stability and report the signed rank only to make the contrast explicit. Stability is thus defined in terms of F1 change, not in terms of an accuracy drop; accuracy is reported separately as a complementary metric.

Implementation uses Python 3.11 with scikit-learn 1.4, XGBoost 2.0, LightGBM 4.3, and CatBoost 1.2; all experiments run on a single workstation and complete in under ten minutes, underscoring that the protocol is accessible to any practitioner without specialized infrastructure. Every stochastic component is seeded (random_state = 42 for data splitting and all model initializations), so the reported numbers are reproducible to within floating-point tolerance on the same dataset version.

## 3. Results and Discussion

This section presents the empirical results of the three evaluation scenarios and interprets the observed patterns. Tables 2, 3, and 4 report accuracy, precision, recall, and F1-score for the four models under the

in-distribution, age-shift, and education-shift regimes, respectively. Table 5 summarizes the F1 change and the resulting stability rank, and Figures 1 through 3 visualize these finding.

Table 2. Performance On the In-Distribution Stratified 80/20 Split (Random State = 42)

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| XGBoost | 0.8778 | 0.7880 | 0.6694 | 0.7239 |
| LightGBM | 0.8775 | 0.7912 | 0.6630 | 0.7214 |
| CatBoost | 0.8776 | 0.8005 | 0.6506 | 0.7178 |
| Logistic Regression | 0.8507 | 0.7314 | 0.5941 | 0.6557 |

*Source: Experimental results, 2026*

In the baseline regime, the three gradient boosting models outperform Logistic Regression by roughly 6.2 to 6.8 F1 points, consistent with the broader literature on tabular classification in which GBDTs typically dominate both linear and deep learning alternatives. The three GBDT variants are essentially tied within 0.6 F1 points of each other, with XGBoost taking the narrow lead (0.7239), followed by LightGBM (0.7214) and CatBoost (0.7178). This ordering is within the range of natural variation expected from different random seeds and is not interpreted as a substantive difference. At this stage, the four models appear practically interchangeable; however, the following scenarios reveal important differences.

Table 3. Performance Under Age Shift (Train: Age < 40; Test: Age ≥ 40)

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| CatBoost | 0.8201 | 0.7980 | 0.6500 | 0.7164 |
| LightGBM | 0.8136 | 0.7741 | 0.6592 | 0.7121 |
| XGBoost | 0.8149 | 0.7827 | 0.6513 | 0.7110 |
| Logistic Regression | 0.6872 | 0.5319 | 0.8766 | 0.6621 |

*Source: Experimental results, 2026*

Under the age shift, the three gradient boosting models retain F1-scores within 0.013 points of their in-distribution values, demonstrating broad stability on the positive class. Logistic Regression, however, displays a striking polarization: its F1-score actually increases slightly (from 0.6557 to 0.6621), but this is achieved through an enormous rise in recall (0.5941 to 0.8766) at the cost of a collapse in precision (0.7314 to 0.5319). In practical terms, the linear model trained on the younger cohort becomes a highly permissive positive-class predictor when applied to the older cohort, flagging a large fraction of test examples as high earners. Because F1 is the harmonic mean of precision and recall, the precision collapse and the recall surge roughly cancel in the F1-score, giving the false impression of stability. Accuracy, in contrast, drops sharply for Logistic Regression from 0.8507 to 0.6872 (−16.4 percentage points), while the GBDT variants lose only 5 to 6 percentage points of accuracy. This disparity is the first illustration of a theme that will recur throughout the analysis: F1 and accuracy can move in opposite directions under distribution shift, and relying on a single metric produces misleading conclusions.

Table 4. Performance Under Education Shift (Train: Non-Degree; Test: Degree Holders)

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Logistic Regression | 0.7974 | 0.7587 | 0.8483 | 0.8010 |
| XGBoost | 0.7971 | 0.8724 | 0.6768 | 0.7623 |
| LightGBM | 0.7920 | 0.8623 | 0.6749 | 0.7572 |
| CatBoost | 0.7911 | 0.8808 | 0.6538 | 0.7505 |

*Source: Experimental results, 2026*

The education-shift results are even more counter-intuitive. Every model registers a higher F1-score on the degree-holder test set than on the in-distribution baseline. Logistic Regression climbs from 0.6557 to 0.8010 (+14.5 F1 points), XGBoost from 0.7239 to 0.7623 (+3.8 points), LightGBM from 0.7214 to 0.7572 (+3.6 points), and CatBoost from 0.7178 to 0.7505 (+3.3 points). At first glance this appears to contradict the entire premise of distribution-shift research. Careful examination, however, reveals a clean statistical explanation: the positive-class prior in the degree-holder test set exceeds

50%, compared with roughly 15% in the non-degree training set. The F1-score of the positive class is mathematically pushed upward whenever the prevalence of positives rises, because both precision and recall depend on the base rate of positives. The models' decision functions remain essentially the same as those learned on the training cohort, but the denominator of these metrics changes in a way that favours higher F1. The effect is amplified for Logistic Regression because its linear boundary, trained on a predominantly negative population, assigns a positive label rather liberally to any observation with several positive-correlated features, and this becomes correct more often in a cohort where half the observations genuinely are positives.

It is critical to emphasize that this F1 improvement is not evidence that the models have actually become more accurate at their intended task. Accuracy under education shift drops from 0.8507 to 0.7974 for Logistic Regression and from roughly 0.877 to 0.791–0.797 for the GBDT models. Precision also rises for the GBDTs (from ~0.79 to ~0.86–0.88), indicating that when these models predict positive they are more often correct because positives are more common in the test set. The overall picture is that distribution shift has changed the operating point of each model relative to the base rate, and F1 is highly sensitive to this change.

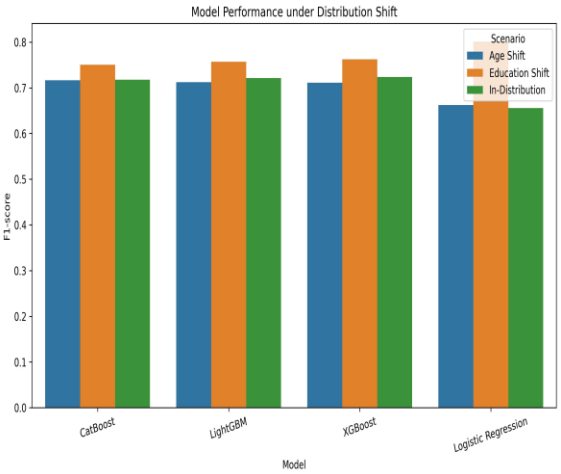


Figure 1. Model performance (F1-score) under the three evaluation scenarios

Figure 1 plots the positive-class F1-score of each model across the in-distribution, age-shift, and education-shift regimes. The three gradient boosting models trace nearly identical, almost flat trajectories: their F1-scores stay close to the in-distribution level under the age shift and rise modestly under the education shift. Logistic Regression follows a markedly steeper path, holding roughly constant under the age shift but jumping sharply under the education shift to overtake all three boosting models. The figure therefore visualizes the central and counter-intuitive result of the study that the shifted regimes raise rather than lower the F1-score and shows that the effect is largest for the weakest in-distribution model.

Table 5. Signed F1 Change From The In-Distribution Baseline And Stability Rank

| Model | ΔF1 (Age) | ΔF1 (Edu.) | ΔF1 (avg) | Rank |
|---|---|---|---|---|
| Logistic Regression | −0.0064 | −0.1454 | −0.0759 | 1 |
| CatBoost | +0.0014 | −0.0327 | −0.0157 | 2 |
| LightGBM | +0.0094 | −0.0358 | −0.0132 | 3 |
| XGBoost | +0.0129 | −0.0384 | −0.0128 | 4 |

*Source: Derived from Tables 2-4. A negative change indicates F1 increased under the shifted distribution*

Table 5 and Figure 2 summarize the signed F1 change across models. Under the convention that rank 1 corresponds to the smallest (most negative) average change, Logistic Regression is paradoxically "most stable" with an average change of −0.076, followed by CatBoost (−0.016), LightGBM (−0.013), and XGBoost (−0.013). Yet Logistic Regression's stability is an artifact of its dramatic F1 improvement under education shift rather than of any genuine robustness property. If instead we adopted the more common convention of ranking stability by the absolute average change, the GBDT family would lead, with XGBoost (0.0128), LightGBM (|0.0132|), and CatBoost (0.0157) all exhibiting more consistent F1 across scenarios than Logistic Regression (0.0759). The choice of stability

E ISSN : 2776-6764
P ISSN : 2774-2032

definition therefore fundamentally alters the conclusion drawn.

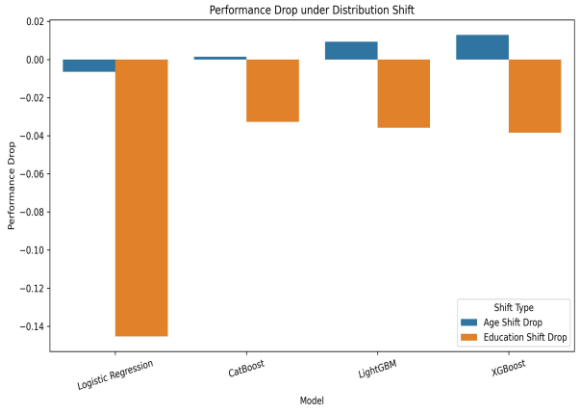


Figure 2. F1-Score Change Under Age Shift and Education Shift (Negative Values Indicate Improvement Over the In-Distribution Baseline).

Figure 2 decomposes the F1 change into its two shift components for each model, using the signed convention of Equation (2) in which a negative bar denotes an F1 increase relative to the baseline. For every model the education-shift bar is strongly negative, confirming the prior-driven F1 inflation, whereas the age-shift bars are small and cluster near zero. Logistic Regression shows by far the largest negative education-shift bar, which is the single value responsible for its anomalous first-place signed-stability ranking. The visual contrast between the large education-shift bars and the near-zero age-shift bars makes clear that the education shift, not the age shift, dominates the aggregate stability figures.

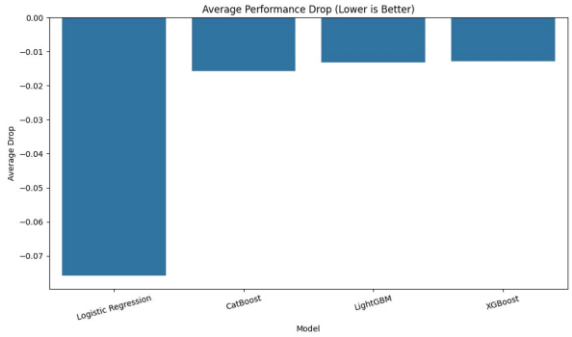


Figure 3. Average F1 Change Per Model Across the Two Shift Scenarios (Lower is Better In This Stability Convention).

Figure 3 collapses the two per-shift changes into a single average value per model, following the signed convention of Equation (3). Read at face value, the chart ranks Logistic Regression as the “most stable” model because it has the most negative average, with the three boosting models clustered tightly just below zero. This ordering is the visual counterpart of the paradox discussed in the text: the apparent stability advantage of Logistic Regression is an artefact of its large F1 increase under the education shift, not a sign of genuine robustness. When the same averages are read under the absolute convention of Equation (4), the ranking reverses and the boosting models become the most stable, which is why the figure should be interpreted together with the accuracy results rather than in isolation.

Three robust patterns emerge from the analysis. First, F1-score alone is an unreliable indicator of robustness on this dataset because it is strongly modulated by changes in the positive-class prior. Under education shift the positive-class prior roughly triples (from ~15% to ~53%), and every model's F1 rises as a result regardless of whether the classifier has genuinely learned transferable representations [20], [21]. Second, accuracy tells a different and arguably more faithful story: the gradient boosting family retains higher accuracy than Logistic Regression under both shifts, and the gap between models widens rather than narrows. Third, the three GBDT variants behave almost identically across all scenarios, differing by at most 0.006 F1 points in any single row of Tables 2-4. This convergence suggests that algorithmic differences between XGBoost, LightGBM, and CatBoost are much smaller than the shift-induced differences between scenarios, and that investments in shift-aware validation are likely to yield larger reliability gains than marginal tuning of the classifier.

These observations carry two practical implications. First, conventional stratified cross-validation is not a sufficient

diagnostic for deployment readiness: it can mask operationally important changes in the precision–recall trade-off that only become visible when the model is evaluated on a domain-blocked partition. Second, the interpretation of any single stability metric depends critically on how the metric interacts with changes in class prior. A practitioner who reports only F1 might conclude that Logistic Regression is the safest choice, while a practitioner who reports only accuracy might conclude that CatBoost is the safest choice. The reality is that these two statements reflect different operational questions: F1 emphasizes the positive class in the shifted domain, while accuracy reflects overall correctness relative to the ground-truth prevalence. Both are legitimate, and a thorough evaluation reports both.

Situating these findings against prior work, Ding et al. showed that the UCI Adult dataset exhibits strong geographic and temporal idiosyncrasies that limit external validity. Our results extend this observation by demonstrating that even within a fixed cross-section of the dataset, controlled partitioning along age or education reveals significant metric-dependent behaviour. Rabanser et al. [22] emphasized that standard performance metrics can mask silent failures under covariate shift and advocated two-sample testing as a detection tool; our education-shift scenario illustrates the converse problem standard metrics can also signal false victories when the shift increases positive-class prevalence. In the Indonesian context, engineering studies published by Akademi Teknologi Industri Dewantara Palopo have repeatedly highlighted the importance of careful empirical testing before system deployment, a principle whose machine learning analogue is precisely the shift-aware evaluation protocol illustrated here.

## 4. Conclusion

This study empirically characterized the stability of four widely used classifiers Logistic Regression, XGBoost, LightGBM, and CatBoost under two distinct distribution shift regimes on the UCI Adult / Census Income dataset. The results reveal a counter-intuitive pattern: every model achieved a higher F1-score on the shifted test sets than on the in-distribution baseline, with the largest improvement (+0.145) observed for Logistic Regression under education shift. This outcome is driven by the increased positive-class prior in the shifted target distributions rather than by any genuine improvement in the models' decision functions, as confirmed by the simultaneous decline in accuracy (up to 16.4 percentage points for Logistic Regression under age shift). When stability is defined as the smallest signed average F1 change, Logistic Regression ranks first, but this position is an artefact of prior-probability shift and not a measure of robustness in the operationally meaningful sense. The three gradient boosting models behave almost identically and retain much higher accuracy than Logistic Regression across both shifts. The main practical implication for practitioners is that robustness should be validated through multiple complementary metrics accuracy, precision, recall, and F1 on domain-blocked partitions that resemble the anticipated deployment distribution. The study has several limitations that bound the generalizability of these findings. The UCI Adult data originate from a 1994 survey with documented demographic idiosyncrasies, so absolute magnitudes should not be extrapolated without replication. Only two shift scenarios and default hyperparameters were considered, and concept drift and temporal shift were not examined. Future work should replicate this protocol on the modernized folktables datasets, introduce shift-mitigation strategies such as importance weighting, distributionally robust optimization, or test-time adaptation, extend the diagnostic to deep tabular architectures, and evaluate calibration metrics such as expected calibration error to capture operating-point

changes that F1 and accuracy cannot fully describe.

**References**

[1] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*, reprint ed. Cambridge, MA, USA: MIT Press, 2022, doi: 10.7551/mitpress/9780262545877.001.0001.

[2] S. Garg, Y. Wu, S. Balakrishnan, and Z. C. Lipton, "A unified view of label shift estimation," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 3290–3300, 2020, doi: 10.48550/arXiv.2003.07554.

[3] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 12, pp. 2346–2363, 2019, doi: 10.1109/TKDE.2018.2876857.

[4] S. Santurkar, D. Tsipras, and A. Madry, "BREEDS: Benchmarks for subpopulation shift," *presented at the Int. Conf. Learn. Representations (ICLR)*, 2021, doi: 10.48550/arXiv.2008.04859.

[5] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts," *presented at the Int. Conf. Learn. Representations (ICLR)*, 2020, doi: 10.48550/arXiv.1911.08731.

[6] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. A. Earnshaw, I. S. Haque, S. Beery, J. Leskovec, A. Kundaje, and P. Liang, "WILDS: A benchmark of in-the-wild distribution shifts," in *Proc. 38th Int. Conf. Mach. Learn. (PMLR)*, vol. 139, pp. 5637–5664, 2021, doi: 10.48550/arXiv.2012.07421.

[7] B. Becker and R. Kohavi, Adult [Dataset]. Irvine, CA, USA: UCI Machine Learning Repository, 1996, doi: 10.24432/C5XW20.

[8] F. Ding, M. Hardt, J. Miller, and L. Schmidt, "Retiring Adult: New datasets for fair machine learning," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 6478–6490, 2021, doi: 10.48550/arXiv.2108.04884.

[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[10] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 3146–3154, 2017, doi: 10.5555/3294996.3295074.

[11] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, vol. 31, pp. 6639–6649, 2018, doi: 10.48550/arXiv.1706.09516.

[12] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Inf. Fusion*, vol. 81, pp. 84–90, 2022, doi: 10.1016/j.inffus.2021.11.011.

[13] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" in *Advances in Neural Information Processing Systems,* vol. 35, pp. 507–520, 2022, doi: 10.48550/arXiv.2207.08815.

[14] H. Yao, C. Choi, B. Cao, Y. Lee, P. W. Koh, and C. Finn, "Wild-Time: A benchmark of in-the-wild distribution shift over time," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 10309–10324,

2022, doi: 10.48550/arXiv.2211.14238.

[15] S. G. Finlayson, A. Subbaswamy, K. Singh, J. Bowers, A. Kupersmith, J. Zittrain, D. C. Kale, A. L. Beam, and S. Saria, "The clinician and dataset shift in artificial intelligence," *N. Engl. J. Med*., vol. 385, no. 3, pp. 283–286, 2021, doi: 10.1056/NEJMc2104626.

[16] B. Sahiner, W. Chen, R. K. Samala, and N. Petrick, "Data drift in medical machine learning: Implications and potential remedies," *Br. J. Radiol*., vol. 96, no. 1150, Art. no. 20220878, 2023, doi: 10.1259/bjr.20220878.

[17] N. K. Wardani, R. M. Arpin, and M. A. Hidayat, "Rancang bangun modul dioda and rectifier," *Dewantara J. Technol*., vol. 3, no. 1, pp. 1–4, 2022. [Online]. Available: https://jurnal.atidewantara.ac.id/index.php/djtech

[18] R. Mahyuddin, A. A. H. Dani, and S. Paembonan, "Sistem informasi data UMKM berbasis website di PLUT-KUMKM Kota Palopo," *Dewantara J. Technol*., vol. 3, no. 1, pp. 82–91, 2022. [Online]. Available: https://jurnal.atidewantara.ac.id/index.php/djtech

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2019, doi: 10.5555/1953048.2078195.

[20] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nat. Mach. Intell*., vol. 2, no. 11, pp. 665–673, 2020, doi: 10.1038/s42256-020-00257-z.

[21] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," presented at the *Int. Conf. Learn. Representations (ICLR)*, 2022, doi: 10.48550/arXiv.2202.10054.

[22] S. Rabanser, S. Günnemann, and Z. C. Lipton, "Failing loudly: An empirical study of methods for detecting dataset shift," in Advances in *Neural Information Processing Systems*, vol. 32, pp. 1396–1408, 2019, doi: 10.48550/arXiv.1810.11953.