IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Robustness Of Tabular Models Under Natural Distribution Shifts

Viswatej Seela

The University of Texas at Austin, Texas, USA

Abstract

Distribution shift, where the test-time data distribution differs from that of the training data, is a critical challenge in real-world machine learning deployments. This paper presents an empirical study of model brittleness under naturalistic distribution shifts in tabular datasets, a domain that remains understudied despite tabular data being central to applications in finance, healthcare, and public policy. We assess the stability of various machine learning models, including logistic regression, gradient-boosted decision trees, and multilayer perceptrons, in response to deployment-driven shifts that alter feature distributions while maintaining the semantics of the prediction task. In addition to these naturalistic shifts, we implement controlled feature-level perturbations as stress tests. Across five classification datasets, all models exhibit significant performance degradation under distribution shift, with vanilla neural networks generally showing larger robustness gaps than tree-based methods, although these gaps can be substantially reduced by standardization and regularization. The study provides an empirical foundation for understanding robustness trade-offs across model classes in structured data settings and highlights open challenges for robust tabular machine learning

Keywords: distribution shift, covariate shift, model robustness, tabular data, generalization, machine learning

1 Introduction

Machine learning models are typically trained on historical data but deployed into environments where the data distribution has evolved. This mismatch, known as distribution shift, can lead to sharp performance degradation and unreliable predictions, especially in high-stakes settings. Tabular data underpins many such applications, including credit scoring, clinical risk prediction, and public policy screening, yet the robustness of tabular models under realistic deployment shifts has received far less attention than robustness in computer vision or natural language processing.

Tabular data presents special difficulties with robustness analysis. Features often mix heterogeneous types (continuous, ordinal, and categorical), exhibit complex interactions, and may encode demographic or institutional information that is ordinal and changes over time. While work on distribution shift has produced benchmarks and methods for images and text, such as WILDS, which curates naturally shifted datasets across modalities, there is no consensus on how to systematically characterize and evaluate robustness for tabular domains.

The primary motivation for this study is twofold. First, tabular data remains the dominant format in operational machine learning systems, so understanding robustness in this setting is practically important. Second, recent results comparing gradient-boosting decision trees and deep learning for tabular data show

that neither paradigm is uniformly superior in-distribution, but the relative robustness of these models under a shift remains poorly understood. A careful empirical characterization of how standard tabular models fail under deployment-motivated shifts can inform both model selection and the development of future robustness methods.

In this work, naturalistic distribution shifts are constructed by partitioning real tabular datasets along semantically meaningful axes such as demographic attributes, risk scores, or feature clusters that approximate realistic changes in deployed systems. These naturalistic shifts are complemented with feature-level perturbations that serve as stress tests around the test distribution. On both in-distribution and out-of-distribution data, we benchmark linear models, gradient-boosted decision trees, and neural networks, and we analyze robustness gaps across datasets and model classes.

The main contributions are

- Naturalistic tabular shifts: Construction of deployment-motivated train–test splits on real tabular datasets that induce covariate changes while preserving the prediction task.
- Cross-model robustness comparison: Empirical evaluation of logistic regression, tuned gradient boosting, and multilayer perceptrons under both naturalistic and synthetic shifts.
- Robustness gap analysis: Systematic measurement of in-distribution versus out-of-distribution performance to quantify and compare robustness gaps across models and datasets.
- Simple mitigation strategies: Evaluation of standardization and regularization techniques for neural networks, showing that basic choices can substantially reduce brittleness without eliminating it.

These results provide an early but detailed view of robustness in tabular ML under realistic shifts and suggest that model choice, data partitioning, and training protocol all play important roles in determining deployment performance.

1.1 Research Questions

This study focuses on the following research questions:

1. RQ1: How significantly do different model types linear models, tree-based models, and neural networks degrade under deployment-motivated covariate shifts in tabular data?
2. RQ2: How do robustness gaps vary across datasets, and which dataset characteristics appear to correlate with brittleness?
3. RQ3: How do simple feature-level perturbations compare to naturalistic shifts in terms of induced performance degradation?
4. RQ4: To what extent can basic mitigation strategies, such as feature standardization and regularization, reduce the robustness gaps observed for neural networks?

2 Background and Motivation

2.1 Distribution Shift in Machine Learning

Distribution shift occurs when the joint probability distribution $P(X, Y)$ at test time differs from the training distribution. We denote the training distribution as $P_{train}(X, Y)$ and the test distribution as $P_{test}(X, Y)$, where $P_{train} \neq P_{test}$.

This phenomenon, known as covariate shift, occurs when the marginal distribution of features changes while the conditional distribution of labels given those features remains approximately constant [6]. The conditional distribution of labels given features remains approximately constant [6]:

$$P_{train}(X) \neq P_{test}(X)$$

$$P_{train}(Y|X) \approx P_{test}(Y|X)$$

This assumption holds in many practical scenarios, such as shifts in feature distributions due to measurement drift, seasonal variations, or changes in data collection procedures [6].

Classical approaches to covariate shift and domain adaptation include importance weighting, feature alignment, and representation learning methods that seek invariances across distributions. However, most modern benchmarks for distribution shift, such as WILDS, focus on images, text, or specialized modalities rather than structured tabular features. Understanding how tabular models behave in analogous settings remains an open practical question.

2.2 Model Brittleness

Model brittleness refers to large drops in performance under modest changes to the data distribution, even when models generalize well on held-out data drawn from the same distribution as the training set. Empirical work in deep learning has shown that overparameterized neural networks can fit random labels yet still generalize in distribution, highlighting the fact that generalization is governed by more than just model capacity. This raises natural questions about how architectural biases, optimization dynamics, and training procedures interact with non-stationary data.

In the tabular domain, gradient-boosted decision trees and related ensembles remain strong baselines and often outperform standard neural networks on static benchmarks. However, the degree to which these performance differences persist or reverse under distribution shift is not well characterized, particularly when models are trained and tuned under comparable protocols.

2.3 Tabular Data and Deep Learning

Recent work has revisited deep learning architectures for tabular data, proposing ResNet-like models and Transformer-based architectures adapted to structured features. In particular, FT-Transformer introduces a feature tokenization mechanism combined with Transformer layers and demonstrates competitive in-distribution performance across a range of tabular benchmarks. These advances suggest that neural architectures can match or approach tree-based models on static data, although their robustness properties under distribution shift remain largely unexplored.

Tree-based methods such as XGBoost and CatBoost provide strong baselines with inductive biases that differ markedly from neural networks; they construct axis-aligned partitions and ensemble them, which can affect how they respond to changes in feature distributions. Understanding how these different biases translate into robustness under deployment-motivated shifts is a key aim of this work.

3 Related Work

3.1 Experimental Overview

The experimental pipeline consists of three main components. First, five real tabular classification datasets are selected and partitioned into train, validation, and test sets, with test sets constructed to induce naturalistic covariate shifts. Second, a suite of models (logistic regression, gradient-boosted decision trees, and multilayer perceptrons) is trained with hyperparameter tuning and early stopping using only in-distribution data. Third, models are evaluated on both in-distribution and shifted test sets, with robustness gaps computed as differences between in-distribution and shifted performance.

3.2 Datasets and Naturalistic Shifts

The study uses five publicly available binary classification datasets drawn from commonly used tabular benchmarks. For each dataset, train–test splits are constructed to approximate realistic deployment shifts by partitioning along semantically meaningful feature dimensions rather than random splits. The resulting shifts are naturalistic in the sense that they are derived from real data and induce structured changes in feature distributions, though explicit metadata such as timestamps or institution identifiers are not always available.

A representative configuration is

Dataset	Samples	Features	Naturalistic shift construction
Adult (income)	30,162	14	Train/test populations are stratified by age, education, and occupation groups to induce demographic shifts in $P(X)$ *We stratify* $P(X)$ while preserving the income prediction task.
Credit Card Default	30,000	23	Train/test splits are constructed using billing cycle indices and account status proxies to emulate drift in financial behavior over time.
Breast Cancer Wisconsin	569	30	Patients are partitioned by clusters in tumor feature space to induce shifts in clinical characteristics between training and test populations.
Diabetes (Pima)	768	8	Stratification by age and BMI bands approximates demographic shifts in the at-risk population.
Mushroom	8,124	22	Partitioning by combinations of cap and gill attributes induces covariate shifts in morphological feature distributions.

Table 1: Representative Dataset and Shift Configuration

These splits follow prior empirical robustness work in which, when explicit deployment metadata are unavailable, train–test partitions are constructed by structured partitioning of the feature space to emulate realistic changes in population or measurement conditions. The resulting shifts alter $P(X)$ while leaving the label semantics unchanged, approximating covariate shift scenarios of practical interest.

3.3 Models

Three main model families are evaluated:

- Logistic Regression: A linear classifier with L2 regularization, serving as a simple, well-calibrated baseline.
- Gradient-Boosted Decision Trees (XGBoost): Tree ensembles optimized with gradient boosting, representing standard strong baselines for tabular data.
- Multilayer Perceptrons (MLP): Feedforward neural networks with 2–3 hidden layers and ReLU activations, trained with the Adam optimizer.

To connect with recent advances in tabular deep learning, the experiments also include a limited evaluation

of an FT-Transformer-style architecture on a subset of datasets, following the feature tokenization and Transformer blocks proposed in prior work. This evaluation probes whether architectural advances change the robustness patterns observed for vanilla MLPs.

3.4 Training and Hyperparameter Tuning

All models are trained using a common protocol to ensure comparability. The available in-distribution data are split into training and validation sets, with the validation set used for hyperparameter tuning and early stopping.

- Logistic Regression: Regularization strength is selected from a logarithmic grid by validation performance.
- XGBoost: Hyperparameters such as maximum depth, learning rate, number of trees, and subsampling ratios are tuned over a small grid using validation accuracy.
- MLP: Hidden layer width, number of layers, dropout rate, and L2 weight decay are tuned over discrete grids. Models are trained with Adam, and early stopping with patience is applied based on validation loss.
- FT-Transformer (subset): Embedding dimension, number of attention layers, and dropout rates are tuned on the tasks where this model is evaluated, using the same validation protocol as for MLPs.

Early stopping is applied to all neural models to prevent overfitting and to approximate realistic model selection under limited validation data. This tuning procedure addresses concerns that comparisons based on default parameters might understate the robustness of more flexible architectures.

3.5 Synthetic Stress Tests

To complement the naturalistic shifts induced by feature-based partitioning, synthetic perturbations are applied to the shifted test sets as stress tests. These perturbations operate on top of the naturalistic test distribution rather than constructing fully synthetic test distributions.

The following perturbations are considered:

- Feature scaling: Multiplying selected continuous features by factors between 2 and 5.
- Additive noise: Adding Gaussian noise with standard deviations in a set of increasing values to continuous features.
- Feature masking: Randomly zeroing out a fraction of features to simulate missingness or sensor dropout.

These stress tests help disentangle sensitivity to local feature perturbations from robustness to global shifts in the underlying population.

3.6 Evaluation Metrics

Performance is measured using classification accuracy. For each dataset and model, three quantities are reported:

- In-distribution accuracy (ID): Accuracy on a held-out test set drawn from the same distribution as the training data (random split).
- Out-of-distribution accuracy (OOD): Accuracy on the naturalistic shifted test set constructed as described above.
- Robustness gap: The difference between in-distribution (ID) accuracy and out-of-distribution (OOD) accuracy, represented as $\Delta Acc = Acc_{ID} - Acc_{OOD}$.

  $\Delta Acc = Acc_{ID} - Acc_{OOD}$, summarizing the degree of brittleness.

Results are averaged over multiple random seeds for data splits and model initialization, and standard deviations are reported to capture variability across runs.

4 Related Results

4.1 Overall Robustness Patterns

Across datasets, all model classes exhibit nontrivial robustness gaps between in-distribution and naturalistic out-of-distribution test performance. Gradient-boosted trees achieve the highest average in-distribution accuracy and typically have smaller robustness gaps than logistic regression and vanilla MLPs, indicating relatively better stability under the constructed shifts. Neural networks that were trained without careful preprocessing or regularization have the largest gaps, which shows how fragile they are to changes in feature distributions.

The FT-Transformer-style model outperforms baseline MLPs on tasks, achieving higher in-distribution and out-of-distribution accuracy and exhibiting reduced robustness gaps. This suggests that architectural advances in tabular deep learning can improve robustness, although they generally remain comparable to, rather than decisively superior to, strong tree-based baselines on these datasets.

4.2 Naturalistic vs Synthetic Shifts

Naturalistic shifts induced by structured partitioning of the feature space produce larger and more heterogeneous performance drops than small synthetic perturbations applied to the same test distributions. In particular, models that are relatively stable under moderate scaling or noise can still experience substantial degradation when the test population differs in terms of demographics or feature clusters.

The ordering of models under synthetic noise and scaling does not always match their ranking under naturalistic shifts. For example, some models that handle additive noise well may still underperform when the relative frequencies of important feature combinations change. This divergence indicates that local perturbation robustness does not fully capture the challenges posed by deployment-motivated shifts.

4.3 Dataset-Dependent Brittleness

Robustness gaps vary substantially across datasets, reflecting differences in class imbalance, feature heterogeneity, and the severity of the induced shifts. Datasets with more pronounced changes in feature distributions or label prevalence between train and test splits tend to exhibit larger gaps for all models.

In some cases, neural networks are particularly brittle on datasets with complex interactions and heterogeneous feature scales, whereas gradient-boosted trees degrade more gracefully. In other cases, simple linear models suffer less degradation than expected, suggesting that the optimal choice of model class under shift is context-dependent and cannot be inferred from in-distribution performance alone.

4.4 Why Tree-Based Models Appear More Robust

The relative robustness of tree-based models can be understood in terms of inductive bias. Decision trees partition the feature space using axis-aligned thresholds, which are invariant to monotone rescaling of individual features and less sensitive to moderate noise on non-critical dimensions. When feature distributions shift, many partition boundaries remain meaningful as long as the relative ordering of feature values is preserved.

In contrast, multilayer perceptrons learn dense, coupled representations where predictions can depend on precise scaling relationships between features. Shifts in marginal feature distributions or co-occurrence patterns can thus perturb learned representations even when the underlying decision rule remains valid. FT-Transformer partially mitigates this sensitivity by combining per-feature embeddings with attention, which can provide a form of implicit normalization and more flexible modeling of feature interactions.

4.5 Mitigation Strategies

Simple mitigation strategies can substantially reduce robustness gaps for neural networks. Feature standardization and appropriate L2 regularization decrease sensitivity to feature scaling and small distributional changes, bringing robustness closer to that of gradient-boosted trees on several datasets. Early stopping based on validation performance also prevents overfitting due to the idiosyncrasies of the training distribution.

However, these strategies do not fully eliminate robustness gaps, and neural networks still tend to exhibit greater degradation under the most severe shifts. This suggests that while preprocessing and regularization are necessary, they are not sufficient; more principled robustness methods tailored to tabular domains are needed to close the remaining gaps.

5 Discussion and Limitations

The experiments demonstrate that even well-tuned models suffer performance degradation under deployment-motivated shifts and that the magnitude of this degradation depends on both dataset properties and model architecture. The results confirm the strong practical position of gradient-boosted trees on tabular data while showing that recent neural architectures can narrow the robustness gap under careful training.

This study has several limitations. First, the naturalistic shifts are constructed from publicly available datasets using feature-based partitioning, rather than being derived from explicit temporal or institutional metadata. As a result, they approximate but do not perfectly reproduce real deployment conditions. Second, the number of datasets is limited, and broader coverage of domains and shift types would be needed to draw stronger general conclusions. Third, the analysis focuses on accuracy-based metrics; robustness under alternative metrics, such as calibration or worst-group performance, remains an important direction.

6 Future Work

Future work could extend this study in several directions. One avenue is to evaluate more specialized robustness methods, such as distributionally robust optimization and domain generalization algorithms, within the tabular setting and compare their trade-offs between in-distribution and out-of-distribution performance. Another direction is to incorporate additional tabular architectures, including attention-based and hybrid models, under a common training and tuning protocol.

As larger benchmarks for tabular robustness emerge, systematic evaluation on such benchmarks will further clarify the landscape. Subsequent efforts proposing standardized suites of shifted tabular datasets reinforce the importance of studying robustness in structured data and provide natural targets for scaling the methodology used here.

References

[1] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*. MIT Press, 2009.

[2] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[3] D. Hendrycks, S. Basart, M. Mazeika, M. Mostajabi, J. Steinhardt, and D. Song, "Scaling out-of-distribution detection for real-world settings," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2022, pp. 8808–8820.

[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[5] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 18932–18943.

[6] M. Sugiyama and M. Kawanabe, *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. MIT Press, 2012.

[7] C. Zhang, S., Bengio, M., Hardt, B., Recht, B., and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[8] K. Kawaguchi, L. P. Kaelbling, and Y. Bengio, "Generalization in deep learning," *arXiv preprint arXiv:1710.05468*, 2017.

[9] 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, E. Pierson, S. Levine, C. Finn, and P. Liang, "WILDS: A benchmark of in-the-wild distribution shifts," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 5637–5664.

[10] S. Ö. Arık and T. Pfister, "TabNet: Attentive interpretable tabular learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, 2021, pp. 6679–6687.

[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 (NeurIPS)*, vol. 31, 2018, pp. 6638–6648.

[12] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[13] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.