Beyond ROC-AUC: A Reproducible Evaluation Protocol for Discrimination, Calibration, and Prediction Stability in Tabular Machine Learning

Ravi Teja Reddy Mandala
Oracle America Inc.
Austin, TX, USA
ORCID: 0009-0003-0313-4864

Bhavya Lahari Bommireddy
Microsoft Corporation
Redmond, WA, USA
ORCID: 0009-0006-2800-7812

***Abstract*—Discrimination metrics such as accuracy, F1-score, and ROC-AUC do not fully characterize the deployment behavior of probabilistic tabular classifiers. Applications that consume predicted risk scores may also require calibrated probabilities and controlled sensitivity to plausible input variation. This paper presents a reproducible protocol that jointly evaluates discrimination, post-hoc calibration, and prediction stability under explicitly specified raw-space perturbations. The protocol defines type-aware raw-space perturbations at three severity levels. Six classifiers are evaluated on the Telco Customer Churn, Bank Customer Churn, and IBM Employee Attrition datasets using ten stratified train-validation-test partitions generated with independent seeds and twenty perturbation draws per severity and run. Uncertainty is quantified using 2,000-sample bootstrap confidence intervals, and best-mean models are compared with alternatives using paired Wilcoxon signed-rank tests with Holm correction. CatBoost achieved the highest mean ROC-AUC on Telco and Bank, whereas MLP and ExtraTrees obtained the highest moderate-perturbation stability. After validation-set Platt scaling, most comparisons between the lowest-mean ECE model and the alternatives were not statistically significant. The three evaluation dimensions produced different model rankings, supporting the protocol as a complement to application-specific evaluation.**

***Index Terms*—tabular machine learning, model evaluation, prediction stability, probability calibration, perturbation testing, robustness evaluation, reproducibility**

I. Introduction

Tabular data remain central to customer analytics, finance, employee risk modeling, healthcare, and enterprise decision support. In these settings, predicted probabilities may be used to prioritize interventions or human review. A deployment-oriented evaluation should therefore consider not only discrimination, but also whether probabilities are calibrated and whether predictions remain stable under plausible input variation.

Conventional metrics such as accuracy, F1-score, ROC-AUC, and PR-AUC do not fully characterize these properties. A model can achieve strong discrimination while producing substantially different risk estimates after small numerical variations, categorical substitutions, or missing-value events. This paper therefore evaluates probabilistic tabular classifiers along three complementary dimensions: discrimination, calibration, and controlled prediction stability.

Tree-based ensemble methods remain strong baselines for tabular data. XGBoost, LightGBM, and CatBoost provide highly competitive gradient-boosted decision tree implementations, while Random Forests and ExtraTrees provide robust ensemble alternatives [1]–[5]. Recent deep tabular models such as TabTransformer, FT-Transformer, TabNet, and SAINT suggest the potential of attention-based or representation-learning approaches for heterogeneous tabular data [6]–[9]. However, recent benchmarking also cautions that deep learning is not uniformly superior to strong tree-based methods on tabular tasks [7], [10].

This paper presents an explicitly specified evaluation protocol for probabilistic tabular classifiers. The protocol combines clean-test discrimination, validation-set post-hoc calibration, raw-space perturbation testing, component-level stability measures, a composite stability summary, repeated data partitions, bootstrap confidence intervals, and paired statistical comparisons. Its purpose is not to replace application-specific evaluation or to propose a new classification algorithm. Instead, it provides a reproducible procedure for determining whether model rankings change when calibration and controlled prediction stability are considered alongside discrimination.

The contributions are: (1) a common protocol for evaluating discrimination, calibration, and stability on the same held-out records; (2) type-aware numerical, categorical, missingness, and dataset-specific perturbations at three severity levels; (3) repeated partitions, perturbation draws, bootstrap intervals, and Holm-corrected paired tests; and (4) an empirical analysis showing that the three dimensions can produce different model rankings.

The methodological contribution is procedural integration rather than a new metric or classifier. Existing benchmarking studies commonly emphasize discrimination, calibration studies examine probability reliability without perturbation analysis, and robustness studies often focus on adversarial or task-specific corruption models without reporting calibrated probabilities. The proposed protocol connects train-validation-test partitioning, validation-only calibration, type-

aware raw-space perturbations with feature-level range constraints, repeated perturbation sampling, component-level stability measurements, uncertainty estimation, and paired model comparison. This common procedure makes the evaluation assumptions inspectable and allows model rankings across discrimination, calibration, and stability to be compared on the same held-out records.

II. Background and Theory

A. Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ denote a binary tabular classification dataset, where $x_i = (x_i^{num}, x_i^{cat})$ contains numerical and categorical features, and $y_i \in \{0, 1\}$. A probabilistic classifier f_θ maps each record to a probability:

$$p_i = f_\theta(x_i) = P_\theta(Y = 1 \mid x_i). \tag{1}$$

B. Prediction Stability

For a perturbed record $\tilde{x} \sim q(\tilde{x}|x)$, expected prediction instability is:

$$\mathcal{I}(f_\theta) = \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{\tilde{x}\sim q(\cdot|x)}\left[D(f_\theta(x), f_\theta(\tilde{x}))\right], \tag{2}$$

where D is a discrepancy between predictive distributions. Lower $\mathcal{I}$ indicates higher stability. Because tabular features are heterogeneous, this work evaluates prediction sensitivity empirically through controlled, type-aware perturbation families rather than through a single norm ball.

C. Stability Metrics

For clean and perturbed probabilities p_i and $\tilde{p}_i$, the mean absolute probability shift is

$$\Delta_{\text{abs}} = \frac{1}{n}\sum_{i=1}^{n} |p_i - \tilde{p}_i|. \tag{3}$$

Decision instability is measured at the fixed threshold $\tau = 0.5$:

$$\text{FlipRate}_{0.5} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\mathbb{I}(p_i \geq 0.5) \neq \mathbb{I}(\tilde{p}_i \geq 0.5)]. \tag{4}$$

For $P_i = [1-p_i, p_i]$, $\tilde{P}_i = [1-\tilde{p}_i, \tilde{p}_i]$, and $M_i = (P_i+\tilde{P}_i)/2$, distributional change is measured using symmetric KL divergence, $D_{\text{SKL}} = \frac{1}{2}[D_{\text{KL}}(P_i\|\tilde{P}_i) + D_{\text{KL}}(\tilde{P}_i\|P_i)]$, and Jensen–Shannon divergence, $D_{\text{JS}} = \frac{1}{2}D_{\text{KL}}(P_i\|M_i)+\frac{1}{2}D_{\text{KL}}(\tilde{P}_i\|M_i)$. Probabilities are clipped to $[10^{-7}, 1-10^{-7}]$ before logarithmic calculations. Median and 95th-percentile absolute shifts are retained as additional diagnostics.

D. Composite Prediction Stability Score

Let $\overline{D}_{\text{SKL}}$ and $\overline{D}_{\text{JS}}$ denote the mean divergences across test records. The composite score uses four fixed, model-independent normalization rules:

$$A = \min\left(\frac{\Delta_{\text{abs}}}{0.25}, 1\right), \quad F = \min(\text{FlipRate}_{0.5}, 1), \tag{5}$$

$$K = \min(\overline{D}_{\text{SKL}}, 1), \quad J = \min\left(\frac{\overline{D}_{\text{JS}}}{\ln 2}, 1\right), \tag{6}$$

The prediction stability score is

$$S = 1 - \frac{A + F + K + J}{4}. \tag{7}$$

Larger S indicates greater stability. The value 0.25 is a prespecified reference scale rather than a theoretical maximum; the remaining bounds are 1 for flip rate and symmetric KL, and $\ln 2$ for Jensen–Shannon divergence. Component metrics remain the primary diagnostics.

E. Calibration

Calibration is assessed using Expected Calibration Error (ECE), Brier score, and negative log loss. ECE is computed using 15 equal-width bins over $[0, 1]$:

$$\text{ECE} = \sum_{b=1}^{15} \frac{|B_b|}{n} \left|\text{acc}(B_b) - \text{conf}(B_b)\right|. \tag{8}$$

Empty bins contribute zero. The Brier score is $\frac{1}{n}\sum_i (p_i - y_i)^2$. Platt scaling is fitted exclusively on the validation split and then applied unchanged to clean and perturbed test predictions. Calibration and stability are evaluated separately because a model may be calibrated yet sensitive to perturbation, or stable yet systematically miscalibrated.

III. Related Work

A. Discrimination and Tabular Model Benchmarking

Accuracy, balanced accuracy, F1-score, ROC-AUC, and PR-AUC quantify different aspects of classification performance. Their interpretation depends on class balance, operating thresholds, and the relative costs of false positives and false negatives. ROC-AUC evaluates ranking over decision thresholds, whereas PR-AUC emphasizes positive-class retrieval and can be more informative for imbalanced tasks. Neither metric determines whether predicted probabilities represent empirical outcome frequencies. Tree-based ensembles remain strong baselines for heterogeneous tabular data. Random Forest and ExtraTrees reduce variance through randomized ensembles, while XGBoost, LightGBM, and CatBoost provide optimized gradient-boosting implementations [1]–[5]. Studies of TabTransformer, FT-Transformer, TabNet, and SAINT have demonstrated the potential of representation learning for tabular data, but large-scale comparisons show that deep tabular models do not uniformly outperform tuned tree ensembles [6]–[10]. Consequently, deployment-oriented evaluation should compare new or neural approaches against carefully tuned tree baselines.

Underspecification and predictive multiplicity provide additional motivation for multi-dimensional evaluation. Models with similar held-out predictive performance can rely on different relationships and exhibit different behavior under deployment conditions [19], [20]. These findings support evaluating model behavior beyond a single aggregate discrimination score.

B. Probability Calibration

A classifier is calibrated when events assigned probability s occur with empirical frequency approximately s. Calibration has been assessed using reliability diagrams, Expected Calibration Error, Brier score, and negative log likelihood. However, ECE depends on the number and construction of bins and should therefore not be interpreted in isolation. Post-hoc methods include Platt scaling, isotonic regression, temperature scaling, and beta calibration [11]–[13]. These methods can improve probability quality without necessarily changing ranking metrics such as ROC-AUC. Calibration is particularly relevant when numerical probabilities determine resource allocation, ranking thresholds, or intervention intensity. It may be less important when a model is used only to produce an uncalibrated ranking.

C. Robustness, Sensitivity, and Consistency

Robustness research includes adversarial perturbations, naturally occurring corruptions, distribution shift, and local sensitivity analysis. Adversarial methods evaluate worst-case changes under a specified threat model [16]–[18]. Consistency-regularization methods encourage similar predictions for an input and an augmented version of that input [14], [15]. These approaches demonstrate that predictive performance on an unmodified test set does not characterize behavior under input variation. For tabular data, perturbation design is complicated by heterogeneous feature semantics. Numerical, ordinal, and categorical features cannot be modified using a single unconstrained norm without potentially producing invalid records. Perturbing one-hot or embedded representations can also generate points that do not correspond to valid raw records. A tabular perturbation protocol must therefore state which features may change, how frequently changes are applied, how their magnitudes are determined, how invalid values are prevented, and whether the expected label is assumed to remain invariant.

D. Positioning of the Present Study

The individual metrics and calibration procedures used in this work are established. The contribution of this study is therefore not a new metric in isolation. Instead, the study specifies and empirically applies a common protocol linking clean-test discrimination, validation-set calibration, raw-space perturbation families, severity-dependent stability measurements, repeated data partitions, bootstrap uncertainty, and paired statistical testing. The protocol retains component-level measurements and uses the composite stability score only as a secondary summary. This positioning distinguishes the present work from studies that report only discrimination, evaluate calibration without perturbation analysis, or test robustness without reporting probability calibration.

IV. Methodology

The protocol combines training-only preprocessing, model selection, validation-set calibration, raw-space perturbation generation, stability measurement, uncertainty estimation, and paired model comparison.

A. Preprocessing

Blank strings are converted to missing values, identifier and constant columns are removed, and the remaining columns are classified as numerical or categorical using data types and safe numeric coercion. Numerical features are median imputed and standardized. Categorical features are imputed using the most frequent training value and one-hot encoded with unknown categories ignored. All imputation statistics, scaling parameters, category vocabularies, and feature schemas are estimated exclusively from the corresponding training partition and applied unchanged to validation, clean-test, and perturbed-test records.

B. Repeated Evaluation

Each dataset is divided into stratified 70% training, 15% validation, and 15% test partitions. Ten independently seeded partitions are used, with seeds 42, 101, 202, 303, 404, 505, 606, 707, 808, and 909. Model selection, fitting, calibration, and test evaluation are repeated independently for every partition. For each model, calibration state, run, and perturbation severity, stability metrics are averaged over 20 deterministically seeded perturbation draws before run-level statistical analysis.

C. Raw-Space Perturbation Protocol

Perturbations are applied to clean test records before preprocessing. All perturbation statistics are estimated from the corresponding training split. Let α_s, r_s^{num}, r_s^{cat}, r_s^{miss}, and r_s^{dom} denote the numerical-noise scale, numerical selection rate, categorical replacement rate, missingness rate, and domain-edit rate at severity s. The weak, moderate, and strong configurations are $(0.02, 0.20, 0.01, 0.01, 0.20)$, $(0.05, 0.35, 0.03, 0.02, 0.35)$, and $(0.10, 0.50, 0.05, 0.05, 0.50)$, respectively.

For each numerical feature j, an independent record-level mask $m_{ij}^{num} \sim \text{Bernoulli}(r_s^{num})$ and Gaussian noise $\epsilon_{ij} \sim \mathcal{N}(0, (\alpha_s \text{IQR}_j)^2)$ are sampled. Selected values are changed according to

$$\tilde{x}_{ij} = \text{clip}\left(x_{ij} + m_{ij}^{num}\epsilon_{ij}, q_{0.01,j}, q_{0.99,j}\right), \qquad (9)$$

where the IQR and the 1st and 99th percentiles are estimated from the training split. If the IQR is zero or non-finite, the training standard deviation is used, with 1.0 as the final fallback. Integer-valued features are rounded after perturbation.

For each categorical feature, an independent record-level mask with probability r_s^{cat} determines whether the value is replaced by a category sampled from its empirical training-set distribution. Numerical and categorical cells are independently set to missing with probability r_s^{miss}. Generic and dataset-specific perturbations are combined within each draw. The perturbation generator is deterministically seeded by run, severity, and draw. Consequently, within a dataset and repeated partition, all models are evaluated using matched perturbation realizations for the corresponding severity and draw, preserving the paired comparison design.

Dataset-specific edits are applied independently with probability r_s^{dom}. For Telco, tenure is changed by -1 or $+1$,

TABLE I. DATASET CHARACTERISTICS BEFORE IDENTIFIER AND CONSTANT-COLUMN REMOVAL

Dataset	Records	Source Columns	Positive Cases	Positive Rate
Telco Churn	7,043	21	1,869	26.5%
Bank Churn	10,000	14	2,037	20.4%
IBM Attrition	1,470	35	237	16.1%

while MonthlyCharges and TotalCharges receive zero-mean Gaussian relative noise with standard deviation 0.02. For Bank, Age and Tenure are changed by -1 or $+1$, while Balance and EstimatedSalary receive the same relative-noise operation. For IBM Attrition, Age, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, and YearsWithCurrManager are changed by -1 or $+1$, while MonthlyIncome receives zero-mean Gaussian relative noise with standard deviation 0.02. The resulting values are clipped to their training-set 1st and 99th percentiles, and integer-valued features are rounded. These perturbations represent controlled inference-time variability rather than adversarial attacks, and the class label is assumed to remain unchanged.

D. Model Selection and Calibration

For every model, dataset, and repeated partition, hyperparameters are selected using randomized search over 10 candidate configurations. Candidate configurations are evaluated using mean ROC-AUC from three-fold stratified cross-validation within the training partition. The best estimator is refitted on the complete training partition. Platt scaling is then fitted exclusively on the separate validation partition and applied unchanged to clean and perturbed test records.

V. EXPERIMENTAL SETUP

For Telco Churn, the customer identifier was removed before modeling. For Bank Churn, row number, customer identifier, and surname fields were removed. For IBM Attrition, EmployeeNumber, EmployeeCount, Over18, and StandardHours were removed because they are identifiers or non-informative constants. All preprocessing statistics were estimated only from the corresponding training partition.

A. Evaluation Measures

Discrimination is evaluated using accuracy, balanced accuracy, F1-score, ROC-AUC, and PR-AUC. Calibration is evaluated using ECE, Brier score, and negative log loss. Stability is evaluated using mean, median, and 95th-percentile absolute probability shift, decision flip rate, symmetric KL divergence, Jensen–Shannon divergence, and the composite stability score. All reported measures are computed on the held-out test partition.

B. Statistical Analysis

Confidence intervals are computed by drawing 2,000 bootstrap resamples from the ten run-level values and reporting percentile-based 95% intervals for the mean. For stability measures, the 20 perturbation draws are first averaged within each run. Pairwise differences are tested using two-sided

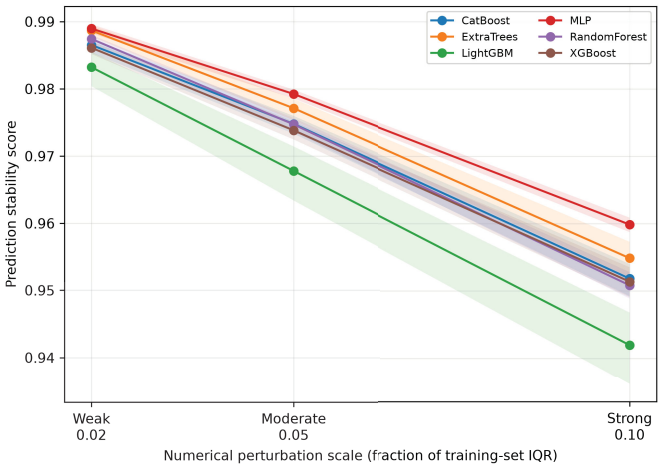


Fig. 1. Composite stability on Bank under weak, moderate, and strong perturbations. Shaded regions show 95% bootstrap confidence intervals

Wilcoxon signed-rank tests on paired run-level values, with zero differences handled using the zero-split convention. Holm correction is applied separately within each dataset, calibration state, perturbation severity, and metric family across the five comparisons between the best-mean model and the remaining models. Statistical significance is evaluated at $\alpha = 0.05$. Because each test contains ten paired runs, nonsignificant results are not interpreted as evidence of equivalence.

VI. RESULTS AND ANALYSIS

A. Predictive Performance

Table II summarizes predictive performance across the Telco Customer Churn, Bank Customer Churn, and IBM Employee Attrition datasets. As shown in Table II, CatBoost achieved the highest mean ROC-AUC on Telco, 0.849 [0.845, 0.854], and Bank, 0.870 [0.864, 0.876]. XGBoost followed on both datasets, with 0.847 [0.842, 0.852] on Telco and 0.867 [0.861, 0.873] on Bank. CatBoost and XGBoost both obtained a mean ROC-AUC of approximately 0.810 on IBM Attrition. Differences among the leading tree models were modest, while MLP was competitive but did not obtain the highest discrimination.

B. Stability Under Perturbation

Table III reports composite stability under moderate perturbations. MLP obtained the highest mean score on Telco, 0.975 [0.974, 0.977], and Bank, 0.979 [0.979, 0.980], while ExtraTrees was highest on IBM Attrition, 0.979 [0.977, 0.982]. The corresponding mean absolute probability shifts on Bank were 0.015 for MLP, 0.016 for ExtraTrees, 0.018 for CatBoost and Random Forest, 0.019 for XGBoost, and 0.023 for LightGBM.

Figure 1 shows that every model's mean stability score decreased from weak to moderate and from moderate to strong perturbations. MLP remained highest across the three Bank severity levels, while LightGBM exhibited the largest decline. These findings, together with the corrected paired

TABLE II. ROC-AUC WITH 95% BOOTSTRAP CONFIDENCE INTERVALS

Model	Telco	Bank	IBM Attrition
CatBoost	0.849 [0.845,0.854]	0.870 [0.864,0.876]	0.810 [0.790,0.832]
XGBoost	0.847 [0.842,0.852]	0.867 [0.861,0.873]	0.810 [0.794,0.827]
Random Forest	0.846 [0.841,0.852]	0.863 [0.857,0.868]	0.794 [0.774,0.813]
ExtraTrees	0.841 [0.836,0.847]	0.859 [0.853,0.866]	0.796 [0.775,0.818]
LightGBM	0.842 [0.836,0.847]	0.865 [0.858,0.870]	0.795 [0.777,0.814]
MLP	0.843 [0.838,0.849]	0.854 [0.845,0.860]	0.803 [0.776,0.827]

TABLE III. COMPOSITE STABILITY SCORE UNDER MODERATE PERTURBATIONS FOR PLATT-CALIBRATED MODELS, WITH 95% BOOTSTRAP CONFIDENCE INTERVALS

Model	Telco	Bank	IBM Attrition
CatBoost	0.969 [0.968, 0.971]	0.975 [0.973, 0.976]	0.976 [0.974, 0.979]
XGBoost	0.965 [0.963, 0.968]	0.974 [0.972, 0.975]	0.973 [0.971, 0.976]
Random Forest	0.968 [0.964, 0.971]	0.975 [0.974, 0.976]	0.977 [0.974, 0.980]
ExtraTrees	0.973 [0.971, 0.975]	0.977 [0.975, 0.979]	**0.979 [0.977, 0.982]**
LightGBM	0.955 [0.949, 0.960]	0.968 [0.963, 0.971]	0.975 [0.973, 0.978]
MLP	**0.975 [0.974, 0.977]**	**0.979 [0.979, 0.980]**	0.977 [0.974, 0.981]

tests, show that stability rankings can differ significantly from discrimination rankings.

C. Calibration Analysis

Table IV reports ECE after validation-set Platt scaling. Random Forest obtained the lowest mean ECE on Bank, 0.027 [0.023, 0.032], while XGBoost obtained the lowest value on IBM Attrition, 0.045 [0.041, 0.050]. CatBoost and XGBoost both obtained mean ECE of 0.036 on Telco. Corrected paired tests found no significant ECE difference between the best-mean model and the alternatives on Bank or IBM Attrition. On Telco, only the XGBoost–MLP comparison was significant after correction. The table therefore characterizes residual calibration error after Platt scaling rather than the magnitude of calibration improvement.

D. Statistical Significance

For Platt-calibrated models under moderate perturbations, CatBoost was the best-mean ROC-AUC model on Bank and significantly exceeded each of the five alternatives ($p_{\mathrm{Holm}} = 0.0098$ for all comparisons). On Telco, CatBoost significantly exceeded ExtraTrees ($p_{\mathrm{Holm}} = 0.0195$), LightGBM ($p_{\mathrm{Holm}} = 0.0234$), and MLP ($p_{\mathrm{Holm}} = 0.0234$), but not Random Forest or XGBoost ($p_{\mathrm{Holm}} = 0.0977$ for both). No IBM Attrition ROC-AUC comparison was significant after correction.

MLP was the best-mean stability model on Bank and Telco. On Bank, MLP significantly exceeded every alternative ($p_{\mathrm{Holm}} = 0.0098$). On Telco, the adjusted value was 0.0371 against ExtraTrees and 0.0098 against each remaining model. On IBM Attrition, ExtraTrees significantly exceeded CatBoost and XGBoost ($p_{\mathrm{Holm}} = 0.0195$), but not LightGBM, MLP, or Random Forest. No best-reference ECE comparison was significant on Bank or IBM Attrition. On Telco, only the XGBoost–MLP comparison was significant ($p_{\mathrm{Holm}} = 0.0488$).

E. Discrimination–Stability Trade-Off

Figure 2 compares Bank ROC-AUC and moderate-perturbation stability, with 95% bootstrap intervals on both axes. CatBoost achieved the highest ROC-AUC, 0.870 [0.864, 0.876], with a stability score of 0.975 [0.973, 0.976]. MLP achieved the highest stability, 0.979 [0.979, 0.980], with lower ROC-AUC of 0.854 [0.845, 0.860]. ExtraTrees occupied an intermediate position with ROC-AUC 0.859 [0.853, 0.866] and stability 0.977 [0.975, 0.979]. Thus, model selection based only on discrimination produces a different ranking from selection that also considers perturbation stability.

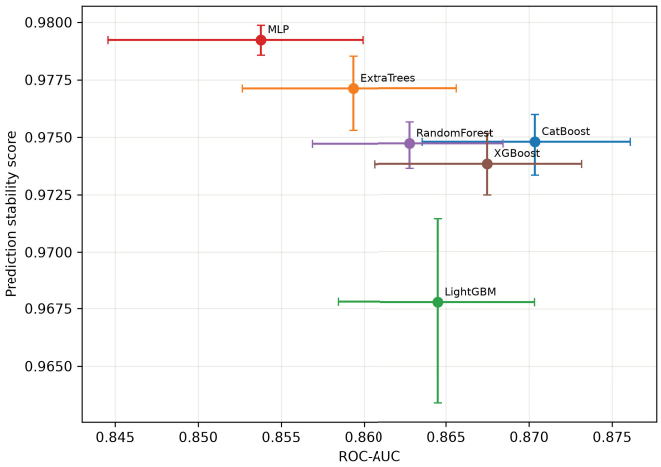


Fig. 2. Bank discrimination–stability trade-off under moderate perturbations. Error bars show 95% bootstrap confidence intervals

VII. DISCUSSION

The results show that discrimination, residual calibration error, and controlled perturbation stability capture different model properties. CatBoost provided the strongest discrimination on Telco and Bank, whereas MLP and ExtraTrees achieved the highest moderate-perturbation stability. Corrected paired tests confirmed several of these differences, while most ECE differences after Platt scaling were not significant. Model selection based only on ROC-AUC can therefore differ from

TABLE IV. EXPECTED CALIBRATION ERROR WITH 95% BOOTSTRAP CONFIDENCE INTERVALS FOR PLATT-CALIBRATED MODELS

Model	Telco Customer Churn	Bank Customer Churn	IBM Employee Attrition
CatBoost	0.036 [0.030, 0.041]	0.030 [0.029, 0.032]	0.053 [0.045, 0.061]
XGBoost	0.036 [0.030, 0.041]	0.032 [0.029, 0.036]	**0.045 [0.041, 0.050]**
Random Forest	0.037 [0.032, 0.043]	**0.027 [0.023, 0.032]**	0.057 [0.048, 0.068]
ExtraTrees	0.037 [0.031, 0.043]	0.031 [0.029, 0.034]	0.049 [0.041, 0.057]
LightGBM	0.041 [0.036, 0.045]	0.033 [0.028, 0.038]	0.048 [0.036, 0.064]
MLP	0.047 [0.042, 0.052]	0.036 [0.033, 0.039]	0.055 [0.048, 0.062]

selection that also considers calibrated probability behavior and input sensitivity.

Applying perturbations in raw feature space reduces the risk of creating invalid encoded representations, such as non-binary one-hot vectors, and permits feature-type-specific clipping and rounding. However, stability is not universally desirable. In fraud detection and similar settings, a small feature change may carry meaningful information and predictive sensitivity may be appropriate. Stability should therefore be evaluated only within an application-defined region where the ground-truth label is reasonably assumed to remain unchanged.

A. Threats to Validity

Only three datasets were evaluated, and synthetic perturbations cannot represent every temporal, population, policy, or data-collection shift. Dataset-specific edits rely on label-invariance assumptions that may not hold for every record. In addition, feature-wise clipping prevents extreme marginal values but does not guarantee that every perturbed record satisfies all cross-feature domain constraints. The composite score also depends on fixed normalization bounds and equal weighting. Finally, ten repeated partitions provide limited power for paired tests. The protocol should therefore be treated as a controlled stress test that complements temporal validation, subgroup analysis, and application-specific evaluation.

VIII. CONCLUSION

This paper presented a reproducible protocol for evaluating probabilistic tabular classifiers across discrimination, calibration, and prediction stability. Across three datasets, CatBoost achieved the highest mean ROC-AUC on Telco and Bank, while MLP and ExtraTrees provided the highest moderate-perturbation stability. Corrected paired tests showed that several differences were statistically significant, whereas most post-calibration ECE differences were not. These findings demonstrate that the three evaluation dimensions can produce different model rankings. The protocol is intended as a deployment-oriented stress test rather than a universal ranking criterion.

REFERENCES

[1] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[2] 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, 2017.

[3] 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, 2018.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.

[6] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "TabTransformer: Tabular data modeling using contextual embeddings," arXiv:2012.06678, 2020.

[7] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[8] S. O. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 35, no. 8, pp. 6679–6687, 2021.

[9] G. Somepalli, M. Goldblum, A. Schwarzschild, C. B. Bruss, and T. Goldstein, "SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training," arXiv:2106.01342, 2021.

[10] S. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022.

[11] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. Int. Conf. Machine Learning*, 2005, pp. 625–632.

[12] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Machine Learning*, PMLR, vol. 70, pp. 1321–1330, 2017.

[13] M. Kull, T. Silva Filho, and P. Flach, "Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers," in *Proc. Int. Conf. Artificial Intelligence and Statistics*, PMLR, vol. 54, pp. 623–631, 2017.

[14] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[15] K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[16] C. Szegedy et al., "Intriguing properties of neural networks," in *Proc. Int. Conf. Learning Representations*, 2014.

[17] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learning Representations*, 2015.

[18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learning Representations*, 2018.

[19] A. D'Amour et al., "Underspecification presents challenges for credibility in modern machine learning," *Journal of Machine Learning Research*, vol. 23, no. 226, pp. 1–61, 2022.

[20] C. Marx, F. Calmon, and B. Ustun, "Predictive multiplicity in classification," in *Proc. Int. Conf. Machine Learning*, PMLR, vol. 119, pp. 6765–6774, 2020.