International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.63282/3050-9246/ ICRTCSIT-113
Eureka Vision Publication | ICRTCSIT'25-Conference Proceeding

Original Article

Model Evaluation Beyond AUC: A Comparative Study of Somers' D, Log Loss, Population Stability Index (PSI), and Kolmogorov–Smirnov (KS) Statistic in Credit Risk and Healthcare Prediction Models

Sai Prashanth Pathi
Senior Data Scientist, Merrick Bank, USA.

Abstract - The Area Under the Receiver Operating Characteristic Curve (AUC) is the dominant evaluation metric in machine learning classification. However, AUC alone cannot capture important properties such as calibration, stability, and practical separability at thresholds. This paper presents an empirical comparison of AUC with Somers' D, the Kolmogorov–Smirnov (KS) statistic, Log Loss, and the Population Stability Index (PSI) across three benchmark datasets: (1) the Breast Cancer dataset from scikit-learn, (2) the Heart Failure dataset from Kaggle, and (3) the Lending dataset from Kaggle. Our results show that for the Cancer dataset, Logistic Regression achieves near-perfect discrimination (AUC = 0.999, KS = 0.977) with low log loss and stable PSI, outperforming more complex models. In the Heart dataset, Gradient Boosting offers the best balance between discrimination (AUC = 0.943, KS = 0.784) and stability (PSI = 0.076), while Random Forest, though highly accurate, shows instability (PSI = 0.183). In the Lending dataset, all models show modest discrimination (AUC ≈ 0.70), but Logistic Regression and Gradient Boosting offer the best trade-off between simplicity, interpretability, and stability. These findings emphasize the importance of a multi-metric evaluation framework that goes beyond AUC, integrating discrimination, calibration, and stability metrics for trustworthy machine learning in regulated domains such as finance and healthcare.

Keywords - Credit risk, Model evaluation, AUC, KS-statistic, Somers' D, Population Stability Index, Log Loss, Healthcare prediction.

1. Introduction

Machine learning models are increasingly deployed in high-stakes decision-making, including consumer credit scoring and healthcare diagnostics. While the Area Under the ROC Curve (AUC) is widely used to evaluate classification models, it is insufficient on its own. AUC measures overall separability but ignores probability calibration, decision threshold effects, and population drift.

This paper evaluates models using additional metrics: Somers' D, KS-statistic, Log Loss, and Population Stability Index (PSI). Somers' D provides a rank-based measure of discriminatory power, KS quantifies distributional separation at thresholds, Log Loss measures probability quality, and PSI captures stability across populations. Together, these metrics provide a holistic framework for evaluating models in regulated domains where reliability and interpretability are critical.

2. Literature Review

2.1. Traditional Metrics in Model Evaluation

Credit risk modeling historically relied on logistic regression–based scorecards, with AUC and Gini (Somers' D) coefficients as primary evaluation tools [1, 2]. While these metrics are well understood, they primarily capture discrimination and neglect other important properties of model behavior. In healthcare, AUC has similarly dominated evaluations of diagnostic and prognostic models.

2.2. Limitations of AUC

Although popular, AUC has several shortcomings. It is threshold-independent and thus provides no guidance for operational decision-making such as loan cutoffs or diagnostic thresholds. Moreover, it cannot detect probability miscalibration or model degradation due to population drift. Studies such as [3] show that in highly imbalanced datasets, AUC may remain deceptively high even when precision, calibration, or subgroup performance degrade.

2.3. Calibration and Reliability

Calibration metrics address these shortcomings by assessing whether predicted probabilities align with observed outcomes. Van Calster et al. [4] argue that calibration is often the "Achilles heel" of predictive analytics. Recent work has introduced conformal prediction and reliability analysis for healthcare process monitoring, highlighting the importance of

probability calibration for trustworthy AI in clinical applications [5]. Deployed models in hospital settings further demonstrate that well-calibrated probabilities are necessary for fairness and risk communication [6].

2.4. Stability and Drift in Practice

In financial applications, models must remain stable across time and subpopulations. The Population Stability Index (PSI) has long been used in industry to monitor drift but lacked formal theoretical grounding until recent work by Sudjianto and Burakov [7]. Their framework situates PSI within information-theoretic divergence measures, strengthening its validity as a monitoring tool. Despite its widespread use in banking, PSI remains underutilized in academic publications [8].

2.5. Fairness and Interpretability

Modern regulations increasingly demand interpretable and fair models. Feature engineering techniques such as Weight of Evidence (WoE) transformations are designed to align model interpretability with regulatory standards. In healthcare, recalibration of models to local populations has been shown to improve calibration and reduce bias [9]. These developments highlight the need to evaluate models not just on accuracy but also on interpretability, fairness, and stability.

2.6. Research Gap

While metrics such as KS-statistic, PSI, and calibration scores are used in industry, academic studies still emphasize AUC almost exclusively. Few comparative evaluations systematically benchmark alternative metrics across multiple domains. This gap motivates the present study, which applies a multi-metric evaluation framework to datasets from finance and healthcare.

3. Methodology

3.1. Datasets

We evaluate four models Logistic Regression, Random Forest, Gradient Boosting, and a Neural Networkacross three datasets:

- Cancer dataset (scikit-learn): Binary classification of malignant vs. benign tumors.
- Heart dataset (Kaggle): Predicting presence of heart disease.
- Lending dataset (Kaggle): Credit default prediction.

3.2. Evaluation Metrics

Area Under ROC Curve (AUC):

$$AUC = \int_0^1 TPR\left(FPR^{-1}(x)\right)dx$$

Measures global separability between classes.

Somers' D:

$$SD = 2 \times AUC - 1$$

Provides a rank-based measure of discrimination.

KS-statistic:

$$KS = max_x \left|F_{good}(x) - F_{bad}(x)\right|$$

Measures maximum distributional separation.

Log Loss:

$$LogLoss = -\frac{1}{N}\sum_{i=1}^{N} \left[y_i \, log(\hat{p}_i) + (1 - y_i) \, log(1 - \hat{p}_i) \right]$$

Captures probability calibration.

Population Stability Index (PSI):

$$PSI = \sum_{i=1}^{k} (p_i - q_i) ln \left(\frac{p_i}{q_i}\right)$$

Measures stability across populations.

3.3. Metric Comparison

Table 1. Comparison of evaluation metrics: strengths, limitations, thresholds, and domains

Metric	Definition	Strengths	Limitations	Typical Thresholds
AUC	Probability that a randomly chosen positive ranks higher than a negative.	Robust, threshold-independent, widely recognized.	Ignores calibration, misleading under drift.	None
Somers' D	Rank correlation between predictions and outcomes (SD = 2 *AUC - 1).	Intuitive rank correlation; complements AUC.	Same limitations as AUC.	None
KS Statistic	Maximum separation between cumulative distributions of positives and negatives.	Useful for cutoff-based analysis; simple.	Threshold-sensitive; not global.	KS >0.4 strong, <0.2 weak.
Log Loss	Penalizes incorrect and overconfident predictions using cross-entropy.	Captures calibration, penalizes overconfidence.	Sensitive to imbalance; less intuitive.	Lower is better; no absolute cutoff.
PSI	Compares distribution of scores between two populations.	Detects drift and instability; regulatory use.	Needs binning; thresholds heuristic.	<0.1 stable, 0.1–0.25 moderate, >0.25 major shift.

3.4. Experimental Setup

Models were trained using a 70/30 train-test split. Hyperparameters for tree-based models were tuned via cross-validation. All metrics were computed on test sets. PSI compared development (training) with validation (test) distributions to assess population shift.

4. Results

This section presents the results of model evaluation across the three datasets: Cancer, Heart, and Lending. Each dataset is analyzed using five metrics (AUC, Log Loss, Somers' D, KS, PSI), supported by ROC curves that provide a visual summary of discrimination.

4.1. Cancer Dataset

Table 2. Cancer Dataset Results

Model	AUC	Log Loss	Somers' D	KS	PSI
Logistic Regression	0.999	0.081	0.997	0.977	0.051
Random Forest	0.996	0.096	0.993	0.949	0.122
Gradient Boosting	0.991	0.093	0.983	0.949	0.345
Neural Net	0.994	0.123	0.988	0.949	0.411

The Cancer dataset is linearly separable, and this is reflected in the near-perfect performance of Logistic Regression, which achieves an AUC of 0.999, KS of 0.977, and the lowest log loss. Random Forest and Gradient Boosting also perform well in discrimination but show reduced stability, with PSI values exceeding 0.1 and 0.25 respectively. The Neural Net performs competitively on AUC but suffers from higher log loss and severe instability (PSI = 0.411), suggesting overfitting.

4.2. Heart Dataset

Table 3. Heart Dataset Results

Model	AUC	Log Loss	Somers' D	KS	PSI
Logistic Regression	0.902	0.409	0.804	0.711	0.085
Random Forest	0.943	0.329	0.887	0.819	0.183
Gradient Boosting	0.943	0.321	0.885	0.784	0.076
Neural Net	0.885	0.475	0.770	0.687	0.075

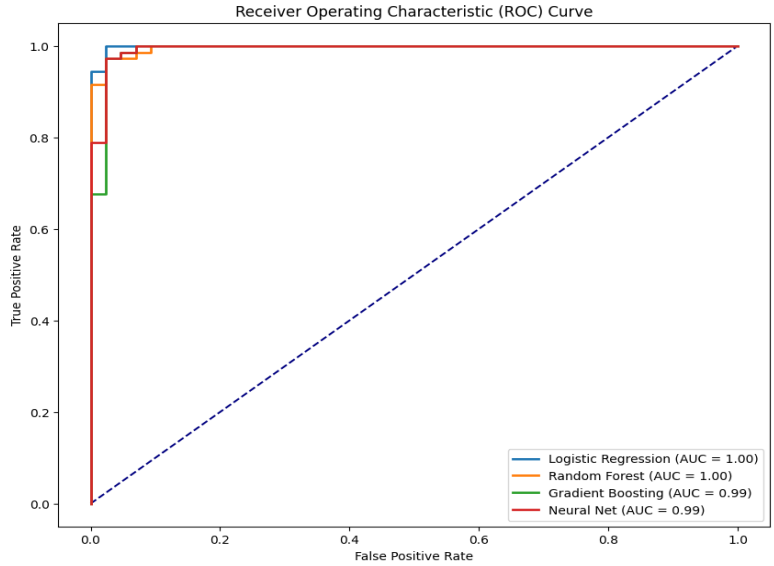

Figure 1. ROC curves for all models on the Cancer dataset. Logistic Regression dominates with near-perfect separation.

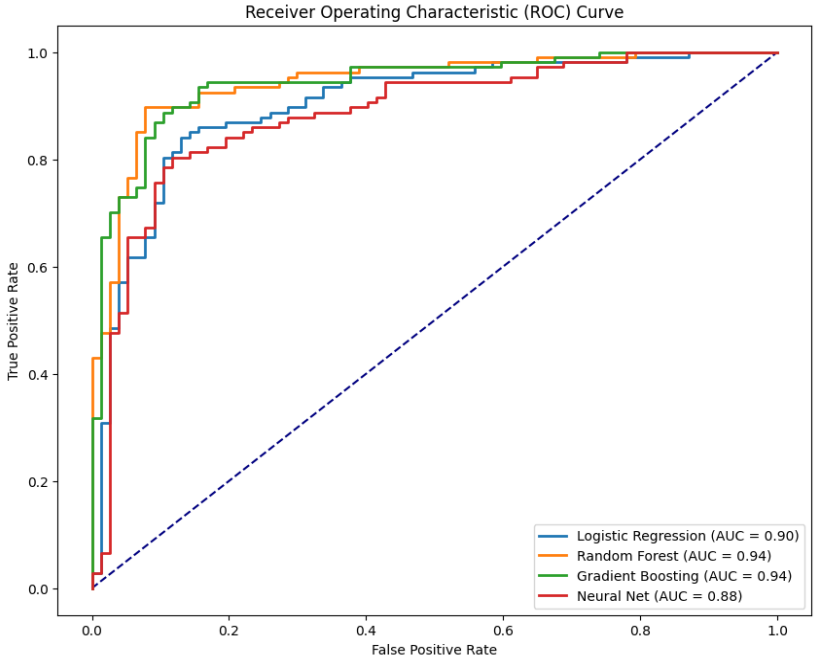

Figure 2. ROC curves for all models on the Heart dataset. Gradient Boosting and Random Forest show the strongest discrimination.

For the Heart dataset, discrimination is strong across most models, with Random Forest and Gradient Boosting achieving the highest AUC (0.943). However, Random Forest shows reduced stability (PSI = 0.183), while Gradient Boosting maintains stability (PSI = 0.076), making it the preferred model. Logistic Regression achieves reasonable discrimination (AUC = 0.902) with stable PSI, but lags behind in separation. Neural Net underperforms in both discrimination (AUC = 0.885) and calibration (highest log loss).

4.3. Lending Dataset

The Lending dataset presents a more challenging classification problem, reflected in modest AUC values around 0.70 for all models. Gradient Boosting achieves the highest AUC (0.704) with low log loss and stable PSI. Logistic Regression provides similar performance with the added benefit of interpretability and the lowest PSI (0.004). Random Forest offers no significant advantage, and its higher PSI values suggest potential instability under population shifts.

Table 4: Lending Dataset Results

Model	AUC	Log Loss	Somers' D	KS	PSI
Logistic Regression	0.700	0.459	0.400	0.315	0.004
Random Forest	0.703	0.459	0.406	0.312	0.014
Gradient Boosting	0.704	0.457	0.408	0.308	0.005
Neural Net	0.697	0.462	0.395	0.313	0.005

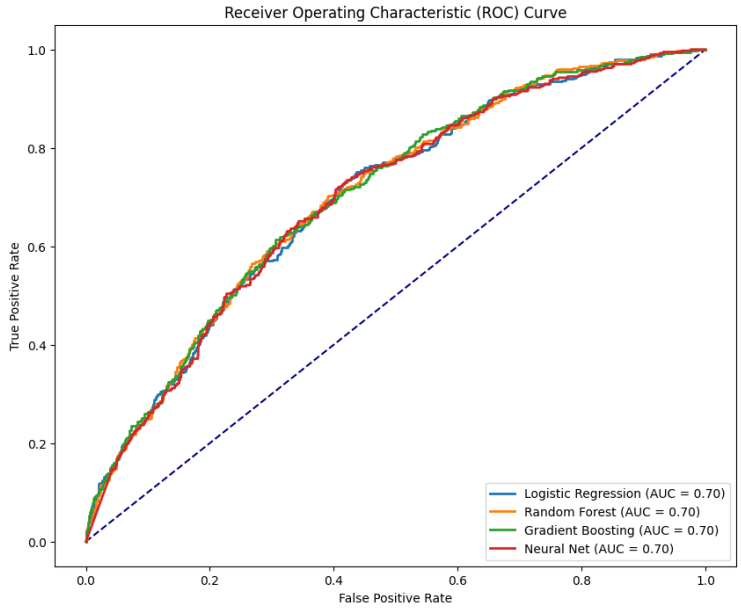

Figure 3. ROC curves for all models on the Lending dataset. All models perform similarly with overlapping ROC curves.

5. Discussion

The results reveal distinct dynamics across domains, emphasizing that no single metric or model suits all contexts.

- **Cancer dataset:** Logistic Regression clearly dominates, achieving near-perfect discrimination and the lowest log loss with stable PSI. More complex models such as Gradient Boosting and Neural Networks slightly underperform in stability (PSI > 0.25), suggesting overfitting.
- **Heart dataset:** Gradient Boosting offers the best compromise between discrimination and stability (AUC = 0.943, PSI = 0.076). Random Forest achieves similar AUC but is less stable (PSI = 0.183). Logistic Regression is more stable but less discriminative. Neural Networks perform the worst on both discrimination and calibration.
- **Lending dataset:** All models achieve only modest discrimination (AUC ≈ 0.70, KS ≈ 0.31). Differences in log loss are small, but Logistic Regression and Gradient Boosting offer the best trade-offs due to interpretability and low PSI. Random Forest adds little improvement while sacrificing stability.
- **Cross-domain insights:** In highly separable problems like Cancer detection, simple linear models suffice. In moderately complex healthcare prediction tasks, boosting methods improve performance while maintaining stability. In financial credit risk, where discrimination is inherently modest, stability and interpretability dominate, favoring Logistic Regression.

6. Conclusion

This study demonstrates that relying solely on AUC is insufficient for robust model evaluation. Complementary metrics reveal key insights:

- **Cancer dataset:** Logistic Regression is best due to high discrimination and stability.
- **Heart dataset:** Gradient Boosting achieves the optimal balance between predictive power and stability.
- **Lending dataset:** Logistic Regression and Gradient Boosting are most appropriate given their stability and interpretability.
- **Key takeaway:** AUC should be supplemented with rank-based metrics (Somers' D), cutoff-based metrics (KS), calibration metrics (Log Loss), and stability metrics (PSI). This multi-metric framework provides a more comprehensive and trustworthy evaluation, aligning with regulatory and clinical standards. Future research should

integrate fairness metrics and perform stress-testing under adverse data conditions to further validate model robustness.

References

[1] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring," J. Royal Statistical Society A, vol. 160, 1997.
[2] L. Thomas, Consumer Credit Models: Pricing, Profit, and Portfolios, Oxford Univ. Press, 2009.
[3] S. García et al., "Evaluating classifier performance with highly imbalanced Big Data," Journal of Big Data, vol. 10, 2023.
[4] B. Van Calster et al., "Calibration: the Achilles heel of predictive analytics," BMC Medicine, vol. 17, no. 1, 2019.
[5] M. Majlatow et al., "Uncertainty-Aware Predictive Process Monitoring in Health- care," Applied Sciences, vol. 15, no. 14, 2025.
[6] M. L. Desai et al., "Assessing calibration and bias of a deployed machine learning malnutrition prediction model," JAMIA, 2023.
[7] A. Sudjianto and D. Burakov, "An Information-Theoretic Framework for Credit Risk Modeling," arXiv:2509.09855, 2025.
[8] M. L. D. Santos et al., "Machine Learning for Credit Risk Prediction: A Systematic Literature Review," Preprints.org, 2023.
[9] B. Van Calster et al., "Calibration of risk prediction models: impact on decision- analytic performance," Medical Decision Making, vol. 39, no. 5, 2019.
[10] N. Siddiqi, Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring, Wiley, 2005.