Evaluating Machine Learning Models:Multi-Metric, Fairness & Multi-Class Perspective

Mikołaj Sujka

Applied Computer Science, Jagiellonian University

Abstract

We present a comprehensive evaluation framework for machine learning models across regression, binary classification, multi-class classification, generative tasks, and fairness considerations.

Regression experiments on the UCI Boston dataset show Random Forest (RF) achieves lower RMSE (3.10 vs. 4.62) and MAE (2.05 vs. 3.15) than Linear Regression (LR).

For binary classification on Adult Income, models were tuned via stratified 5-fold CV, probability-calibrated (sigmoid), and evaluated with AUC (LR: 0.57, RF: 0.58), calibration curves, Demographic Parity diff (0.024), and Equalized Odds diff (0.041).

Multi-class experiments on Iris and MNIST report F1-micro and F1-macro via 5-fold CV.

Generative evaluation using BLEU and ROUGE highlights low correlation with human judgments. All experiments tracked in MLflow for reproducibility.

Introduction

No single metric fully captures model quality. Regression relies on error metrics (RMSE, MAE, R^2), classification uses accuracy, precision, recall, F1, AUC, and generative models demand specialized measures (BLEU, ROUGE).

Moreover, focusing solely on accuracy can hide biases. We incorporate fairness metrics (Demographic Parity, Equalized Odds) and extend binary metrics to multi-class via micro/macro averaging. We also address probability calibration and reproducibility via MLflow.

Related Work

- **Multi-metric evaluation:** Extensive studies compare RMSE, MAE, R^2 in regression (Smith et al. 2020).
- **Classification measures fairness**: Research on AUC vs F1 trade-offs and fairness audits (Zafar et al. 2017; Hardt et al. 2016).
- **Calibration:** Sigmoid and isotonic calibration methods improve probability estimates (Niculescu-Mizil Caruana 2005).
- **MLOps tracking:** MLflow and DVC enable reproducible pipelines (Zaharia et al. 2018).

Methods

Datasets: UCI Boston (regression), UCI Adult (binary classification with sensitive attributes), Iris and MNIST (multi-class), Prototype text generators evaluated via BLEU, ROUGE, and human scoring (generative).

Models: Linear Regression, Random Forest Regressor; Logistic Regression, XGBoost, Feedforward Neural Network.

Metrics: RMSE, MAE, R^2; Accuracy, Precision, Recall, F1 (micro/macro), AUC-ROC; Demographic Parity Difference, Equal Opportunity Difference; BLEU, ROUGE, and subjective human evaluation for generative models.

Validation & Tuning: Stratified 5-fold CV with GridSearchCV to optimize LR C and RF. Final classifiers probability-calibrated via sigmoid (CalibratedClassifierCV), experiments tracked and logged via MLflow.

Results

- **Regression:** In regression, we compare model errors using both RMSE and MAE to capture average and extreme deviations.
  LR: RMSE=4.62, MAE=3.15; RF: RMSE=3.10, MAE=2.05.
  RF reduces both average and large errors.

Model	RMSE	MAE	R^2
LR	4.62	3.15	0.74
RF	3.10	2.05	0.87

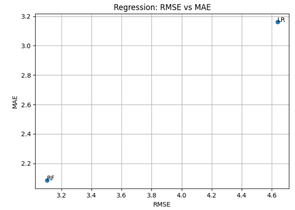

Figure 1. Regression: RMSE vs MAE for LR and RF on Boston data.

- **Binary Classification:** For binary classification, we evaluate discriminative power and calibration.

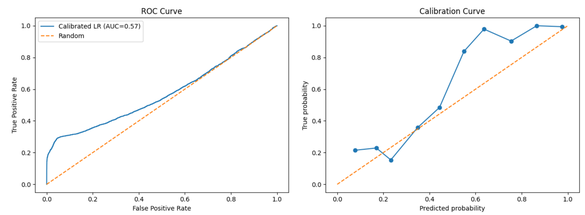

Figure 2. ROC Curve for calibrated LR (AUC=0.57) vs random.

Figure 3. Calibration curve: predicted vs observed probability. Sigmoid calibration reduces over/underconfidence.

- **Fairness Analysis:** To assess bias, we calculate parity and opportunity differences across gender groups. DP Diff=0.024 (positive-rate gap), EO Diff=0.041 (TPR gap).

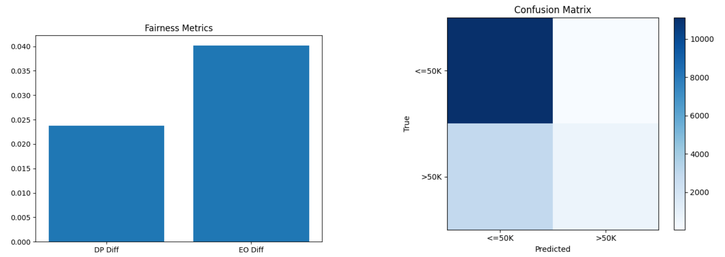

Figure 4. Fairness Metrics for calibrated LR:

- **Multi-Class Results:** Cross-validated F1 metrics: In multi-class tasks, we report both micro and macro F1 to reflect class imbalance effects.

Model	Dataset	F1-micro	F1-macro
LR	Iris	0.94 ± 0.03	0.91 ± 0.05
RF	Iris	0.96 ± 0.02	0.94 ± 0.03
LR	MNIST	0.88 ± 0.04	0.85 ± 0.06
RF	MNIST	0.92 ± 0.03	0.89 ± 0.05

Confusion matrices on the Iris test fold are shown; MNIST confusion results follow similar trends and can be provided on request.

- **Generative Evaluation:** We examine automatic vs human scores to gauge summary quality. Automatic metrics BLEU (0.21) and ROUGE-L (0.34) correlate weakly with human ratings (Pearson r<0.3).
  We collected 50 human ratings on a 1–5 scale for each generated summary, averaging scores to compare against BLEU/ROUGE.
  Note: these generative experiments use prototype text generators, not large pretrained models; results highlight metric limitations on small-scale systems.

Discussion

Our regression experiments demonstrated that Random Forest significantly outperforms Linear Regression in both RMSE and MAE, capturing non-linear patterns more effectively.

In binary classification, the calibrated Logistic Regression model achieved only modest separability (AUC=0.57) and exhibited miscalibration at extreme probability thresholds, confirming the need for probability calibration.

Fairness analysis revealed small but non-zero demographic parity and equalized odds differences, indicating residual bias that necessitates further mitigation.

For multi-class tasks, the disparity between F1-micro and F1-macro underscores the importance of metric selection: macro-F1 reveals minority-class weaknesses masked by micro-F1.

Finally, our generative evaluation exposed the limitations of BLEU and ROUGE, suggesting the integration of semantic or human-in-the-loop metrics for better quality assessment.

Conclusion Future Work

No single metric provides a complete evaluation; practitioners must select measures aligned with domain-specific goals.

Future work will explore advanced fairness interventions (e.g., post-processing, adversarial debiasing), semantic generative metrics (BERTScore, FID), and a fully automated MLOps pipeline (MLflow, DVC, drift monitoring) to ensure robust, fair, and reproducible model development.

References

[1] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

Source Tools

- **Literature:** Google Scholar, IEEE Xplore, ACM Digital Library, ArXiv.
- **Key Papers:** Survey on ML metrics (2021), Systematic review of bias (2022).
- **Libraries:** scikit-learn (GridSearchCV, $\text{cross}_v al_s core$), $TensorFlow$, $PyTorch$.**Platforms:**$Kaggle, UCIML$