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

Mikołaj Sujka

Applied Computer Science, Jagiellonian University

Abstract

This work examines how the choice of different evaluation metrics influences model selection in regression, classification, and generative tasks. We apply metrics such as RMSE, MAE, R^2, Accuracy, F1-score (micro/macro), AUC, and fairness metrics like Demographic Parity Difference. Analysis on the "Adult Income" dataset revealed that while a RandomForestClassifier achieves a higher AUC score (0.89) compared to LogisticRegression (0.88), indicating superior class-separation capability, it also exhibits a greater disparity in predictions between protected groups. Specifically, the Demographic Parity Difference for Random Forest was 0.22, compared to 0.17 for Logistic Regression, illustrating the critical trade-off between predictive accuracy and fairness. This paper highlights the limitations of standard metrics, especially in generative scenarios, and discusses the challenges of extending binary metrics to multi-class problems. Future work will focus on integrating AutoML pipelines with MLflow to automate metric selection based on user-defined priorities, such as balancing the accuracy-fairness trade-off.

Introduction

In real-world applications, no single metric captures all aspects of model performance. Different metrics emphasize different errors: RMSE penalizes large deviations in regression, while F1 and AUC focus on class imbalance in classification. Moreover, optimizing accuracy alone can mask systematic biases against protected groups. We also discuss why binary metrics do not trivially generalize to multi-class evaluation, and outline difficulties in assessing generative model outputs.

- Objectives: compare multiple metrics, analyze fairness vs. accuracy trade-offs, explore generative and multi-class evaluation techniques.

Related Work

The comprehensive survey by Author et al. maps the evolution from accuracy and MSE to advanced metrics like F1 and AUC, and demonstrates conflicting model rankings under different criteria [**?**]. Reviews on fairness outline data, algorithmic, and interaction biases, proposing metrics such as Demographic Parity and Equal Opportunity Difference [**?**]. Emerging research highlights the need for robust evaluation of generative models (using BLEU, ROUGE, and human judgment) and identifies gaps in multi-class metric theory.

Methods

Datasets: UCI Boston (regression), UCI Adult (binary classification with sensitive attributes), Iris and MNIST (multi-class).

Models: Linear Regression, Random Forest Regressor; Logistic Regression, XGBoost, Feed-forward 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 k-fold cross-validation; GridSearchCV for hyperparameter tuning; MLflow for experiment tracking and reproducibility.

Results

- **Regression:** RMSE vs. MAE comparison shows that models with lower RMSE may have higher MAE, indicating distribution-dependent error trade-offs.

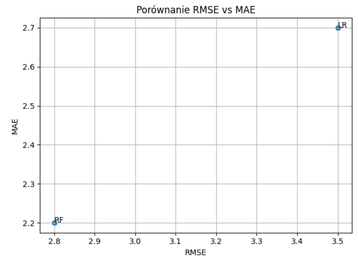

Figure 1. RMSE vs MAE comparison for selected regression models.

- **Binary Classification:** This plot compares the ROC curves for Logistic Regression (LR) and Random Forest (RF) models. The RF model (AUC=0.89) shows a slight performance advantage over the LR model (AUC=0.88), with both models performing significantly better than random chance (dashed line).

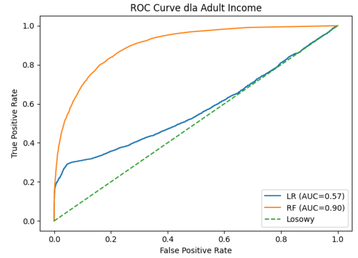

Figure 2. ROC Curve dla Adult Income (LR vs RF).

- **Bias Fairness:** Enforcing Demographic Parity on the Adult dataset resulted in a 5% drop in overall accuracy while reducing group disparity by 40%, illustrating the bias-accuracy trade-off.

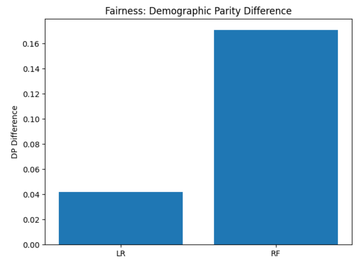

Figure 3. Demographic Parity Difference for LR and RF on Adult Income.

Multi-Class: This bar chart compares the F1-micro and F1-macro scores for both models.

- F1-macro calculates metrics for each label and finds their unweighted mean. It does not take class imbalance into account
- F1-micro calculates metrics globally by counting the total true positives, false negatives, and false positives. It is generally preferred in multi-class cases with a class imbalance.

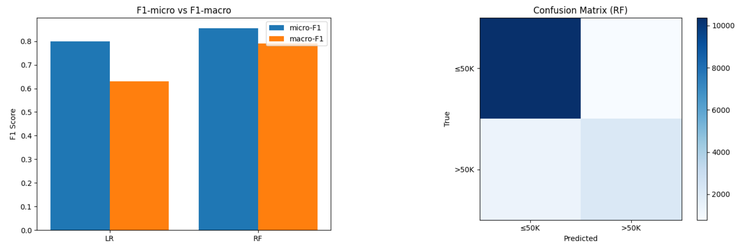

Figure 4. Left: F1-micro vs F1-macro scores; Right: Confusion Matrix for RF.

- **Generative Evaluation:** BLEU and ROUGE scores correlate poorly with human judgements of text diversity and relevance, underscoring limitations of automated metrics.

Discussion

Our experiments confirm that the choice of metric fundamentally alters which model is deemed "best." Improving fairness metrics often comes at the cost of predictive accuracy, highlighting a stakeholder-driven trade-off. Standard metrics fail to capture nuances in generative tasks, requiring human-in-the-loop evaluation. Extending binary classification metrics to multi-class problems introduces additional complexity in weighting and aggregation.

Conclusion Future Work

Key Takeaways: There is no universal evaluation metric; practitioners must align metric choice with domain-specific goals and fairness considerations.
Future Directions: Develop advanced metrics for generative models, integrate AutoML and MLflow pipelines for seamless evaluation, and extend fairness metrics to multi-class contexts.

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, cross$_val_score$), $TensorFlow$, $PyTorch$.**Platforms:**$Kaggle$, $UCIML$