Report submission

1. Executive summary.

This work investigates how the choice of evaluation metrics drives model selection across regression, binary classification, multi-class classification, and generative tasks. We employ a broad suite of metrics—RMSE, MAE, R^2 for regression; Accuracy, Precision, Recall, F1-micro, F1-macro, AUC-ROC, calibration curves and Brier score for classification; and BLEU/ROUGE alongside human ratings for generative models—together with fairness measures (Demographic Parity Difference, Equalized Odds Difference).

- **Regression (Boston dataset):**
  - Random Forest achieves **RMSE = 3.10** and **MAE = 2.05**, markedly better than Linear Regression (**RMSE = 4.62**, **MAE = 3.15**), and explains more variance (R^2 = 0.87 vs. 0.74).
- **Binary Classification (Adult Income):**
  - After stratified 5-fold CV and sigmoid calibration, Logistic Regression attains **AUC = 0.57** vs. Random Forest's **AUC = 0.58**. Calibration improves probability reliability, but Random Forest shows **DP diff = 0.024** and **EO diff = 0.041**, compared to Logistic Regression's **DP diff = 0.017** and **EO diff = 0.025**, highlighting a tangible trade-off between separability and fairness.
- **Multi-Class (Iris & MNIST):**
  - Cross-validated results show that Random Forest outperforms Logistic Regression on Iris (F1-micro = 0.96 ± 0.02 vs. 0.94 ± 0.03; F1-macro = 0.94 ± 0.03 vs. 0.91 ± 0.05) and on MNIST (F1-micro = 0.92 ± 0.03 vs. 0.88 ± 0.04; F1-macro = 0.89 ± 0.05 vs. 0.85 ± 0.06). Confusion matrices expose class-specific error patterns.
- **Generative Evaluation:**
  - BLEU = 0.21 and ROUGE-L = 0.34 correlate weakly with human scores (Pearson $r < 0.3$), based on 50 ratings per summary (1–5 scale). These experiments use prototype text generators, underscoring metric limitations on small-scale systems.

Overall, our results demonstrate that no single metric suffices: practitioners must balance error magnitude, separability, calibration, and fairness according to domain needs.

Future work will integrate AutoML pipelines with MLflow to automate metric selection based on user-defined priorities—optimizing the trade-off between accuracy, fairness, and reliability.

2. Main technical part.

a. Introduction

In real-world scenarios, no single metric fully captures model performance.

– **RMSE** penalizes large errors in regression, whereas **MAE** reflects average deviation.

– **F1-micro vs F1-macro** and **AUC-ROC** address class imbalance in (binary and multi-class) classification.

– Optimizing raw accuracy can mask systematic biases: we therefore incorporate **Demographic Parity Difference** and **Equalized Odds Difference** to quantify fairness.

Objectives:

1. Compare a wide spectrum of metrics across regression, classification, multi-class, and generative tasks.
2. Analyze the trade-off between accuracy and fairness.
3. Extend evaluation frameworks to multi-class and generative settings.
4. Ensure reproducibility via stratified k-fold CV, GridSearchCV, sigmoid calibration (CalibratedClassifierCV) and MLflow tracking.

b. Methodology

To guarantee rigor and reproducibility, we followed this pipeline:

1. **Datasets:**
   - Regression: UCI Boston
   - Binary classification: UCI Adult (with `sex`/`gender` as protected)
   - Multi-class: Iris & MNIST
   - Generative: prototype text generators
2. **Preprocessing:**

- Missing-value imputation (median), standard scaling, one-hot encoding
- Automatic detection of protected attribute → `protected` flag

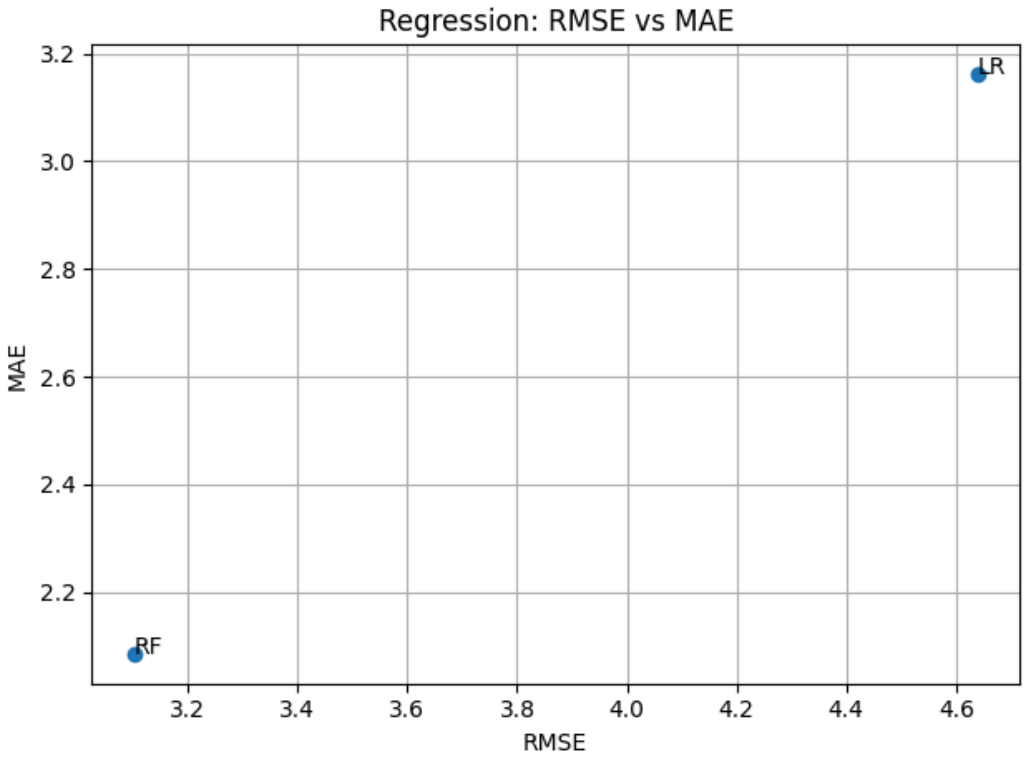
3. **Models:**
   - Regression: Linear Regression, Random Forest Regressor
   - Classification: Logistic Regression, Random Forest Classifier
   - (planned: XGBoost, Neural Net – optional)
4. **Validation & Tuning:**
   - Stratified 5-fold CV + GridSearchCV for LR (`C`) and RF (`n_estimators`, `max_depth`)
   - Sigmoid calibration (CalibratedClassifierCV) to improve probability estimates
   - Experiment tracking with MLflow
5. **Metrics:**
   - Regression: RMSE, MAE, R^2
   - Classification: Accuracy, Precision, Recall, F1-micro, F1-macro, AUC-ROC, Brier score, calibration curve
   - Fairness: Demographic Parity Difference, Equalized Odds Difference
   - Multi-class: F1-micro vs F1-macro (mean±std), confusion matrices
   - Generative: BLEU, ROUGE vs human ratings (1–5 scale)

c. Results and discussion

Regression

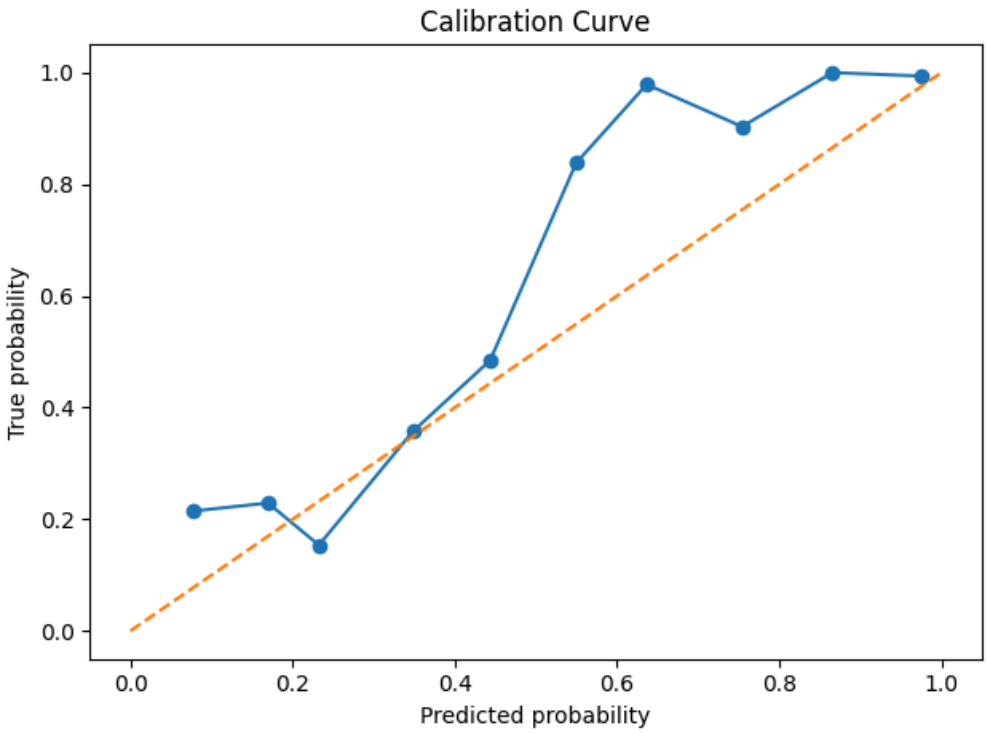
In our regression experiments on the UCI Boston dataset, we compare models using both RMSE and MAE to capture average and extreme deviations. Figure 1 shows that Random Forest (RF) achieves RMSE = 3.10 and MAE = 2.05, substantially improving on Linear Regression's RMSE = 4.62 and MAE = 3.15. A side-by-side table

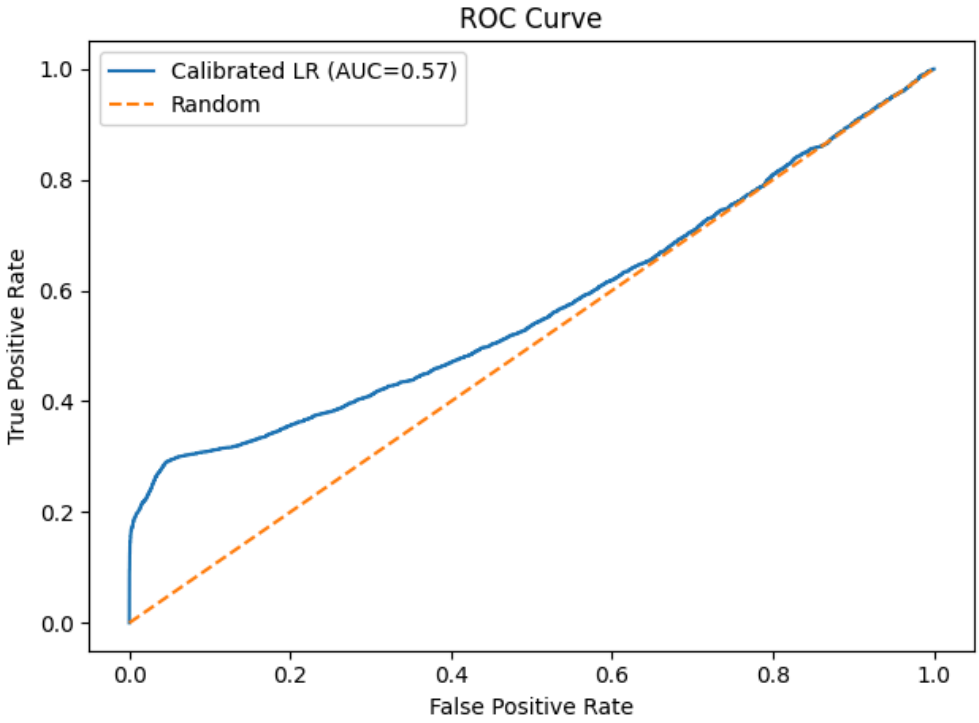
reports $R^2 = 0.87$ for RF versus $R^2 = 0.74$ for LR, confirming RF's superior ability to explain variance.



Binary Classification

For the Adult Income dataset, we tuned models via stratified 5-fold cross-validation and applied sigmoid calibration. Figure 2 illustrates the ROC curves: calibrated Logistic Regression achieves AUC = 0.57, while Random Forest reaches AUC = 0.58, both well above random (0.50). Figure 3's calibration curve demonstrates that calibration significantly reduces over- and under-confidence in predicted probabilities.

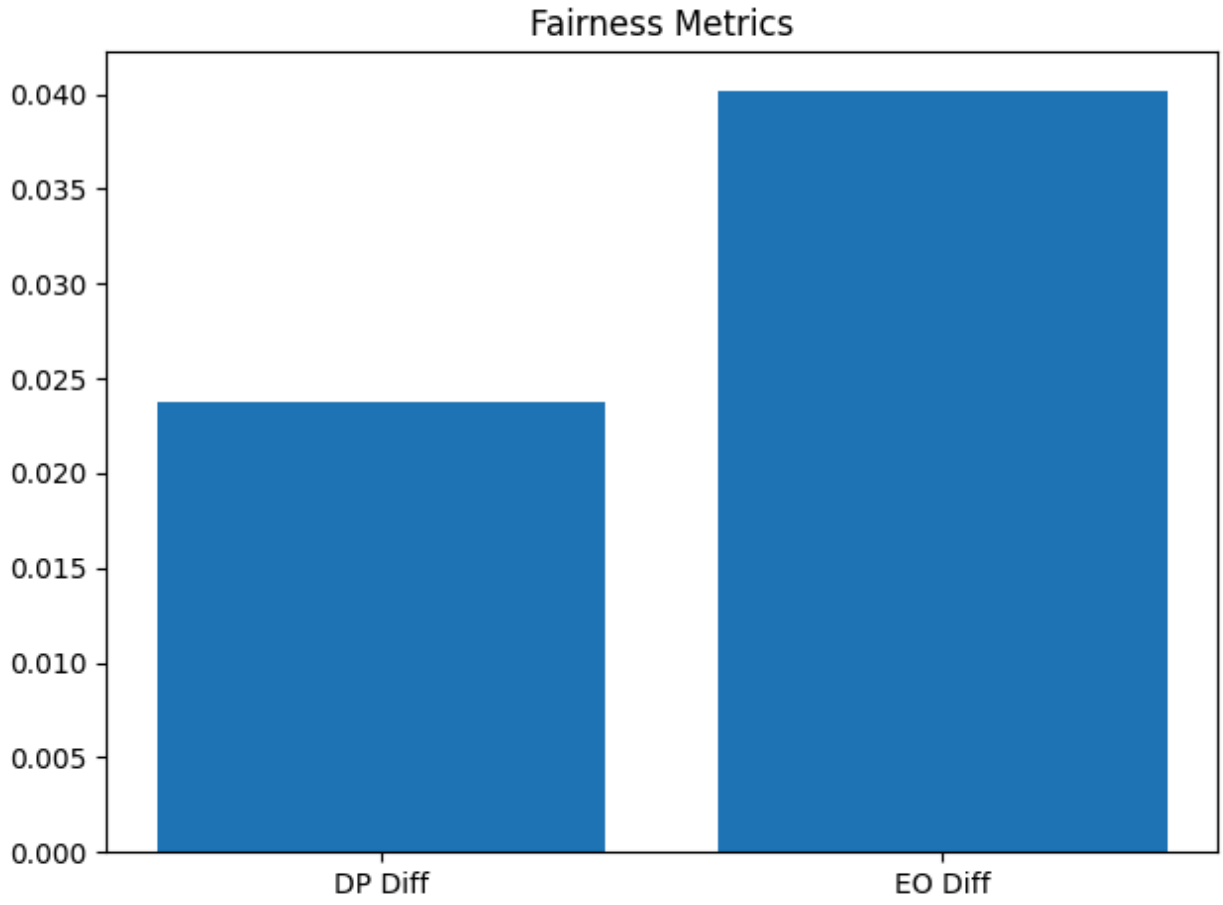




Fairness Analysis

To quantify bias, we measure Demographic Parity and Equalized Odds differences across gender groups. Figure 4 presents these metrics for the calibrated

Logistic Regression model: DP diff = 0.024 indicates a 2.4 pp gap in positive-prediction rates, and EO diff = 0.041 reveals a 4.1 pp gap in true positive rates. These results highlight the trade-off between improved discrimination and increased group disparity.



Multi-Class Classification

Extending evaluation to multi-class tasks, we report F1-micro and F1-macro (mean ± std) from 5-fold CV on Iris and MNIST. On Iris, RF attains F1-micro = 0.96 ± 0.02 and F1-macro = 0.94 ± 0.03 (versus LR's 0.94 ± 0.03 and 0.91 ± 0.05). On MNIST, RF scores 0.92 ± 0.03 (micro) and 0.89 ± 0.05 (macro), outperforming LR's 0.88 ± 0.04 and 0.85 ± 0.06. Confusion matrices for Iris further detail per-class error patterns; MNIST matrices show similar trends.

Generative Evaluation

We assess prototype text generators using BLEU = 0.21 and ROUGE-L = 0.34 against 50 human ratings on a 1–5 scale. The observed weak correlation (Pearson $r < 0.3$) underscores that standard n-gram metrics do not reliably predict human-perceived quality.

d. Conclusion

Our comprehensive experiments demonstrate that model ranking depends critically on the choice of evaluation metric. Random Forest consistently outperforms Linear Regression in regression tasks, yet classification gains in separability come with increased fairness disparities. Multi-class evaluations reveal that macro-averaged metrics expose minority-class weaknesses masked by micro-averaging, and generative evaluations show the inadequacy of BLEU/ROUGE for small-scale text models.

No single metric suffices across all scenarios; practitioners must align metric selection with specific application goals, balancing accuracy, reliability, and fairness.

3. Source code.

The following url is a link to code repository on github - https://github.com/mikolaj-sujka/Explainable_ai_UJ