Imbalanced Learning for LC-PFNs

Bozhidar Kostov

Imbalanced Learning

Definition: Some classes or distributions are more prevalent in training data, which can lead to ML models being biased towards the majority class.

Key Challenge: Evaluation can be misleading when dealing with imbalanced datasets.

Research Questions

RQ1: How does imbalanced training compare to training on a single learner?

RQ2: What trends emerge in imbalanced training scenarios as we vary the proportions?

Selected Learners

From LCDB 1.1, learners selected based on % of ill-behaved learning curves:

Learner	% Ill-behaved Curves
Ensemble Extra Trees	3.4%
Extra Tree	1.9%
Perceptron	3.8%
SVC Sigmoid	58.1%
QDA	45.7%

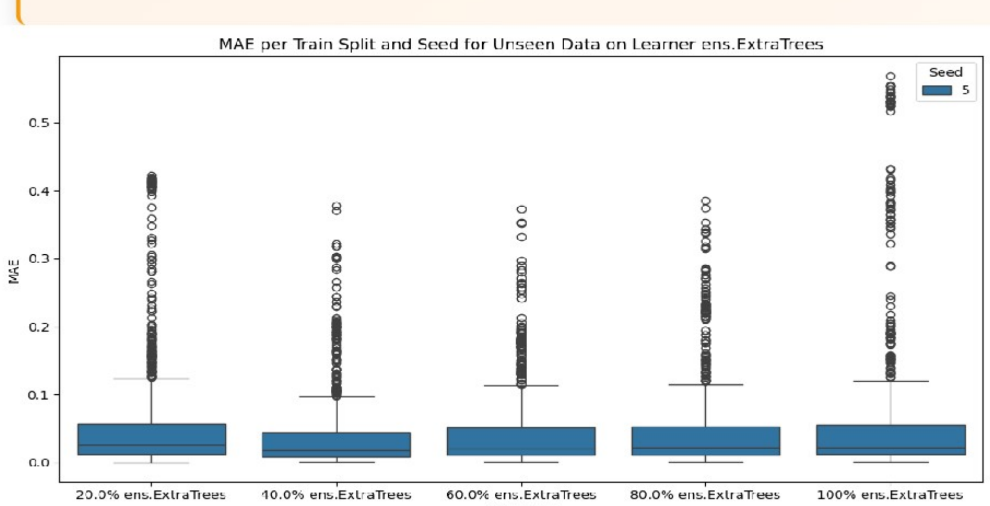
Experiment Setup

For each pair of learners A and B:

- Same training set size
- Vary split: 80/20, 60/40, 40/60, 20/80
- Train 3 PFNs per split (different seeds)
- Evaluate on curves from A and B

Learner Pairs:

- Ensemble Extra Trees ↔ SVC Sigmoid
- Extra Tree ↔ Perceptron
- QDA ↔ SVC Sigmoid



Evaluation Metrics

- **MAE** — Mean Absolute Error
- **Miscoverage** — Coverage Quality
- **Area of CI** — Confidence Interval

Statistical Analysis: Wilcoxon signed rank test with two one-sided tests to check for trends across all seeds.

Key Findings

Ensemble Extra Trees + SVC Sigmoid

RQ1: Mixed splits consistently outperform on MAE; Miscoverage and Area depend on evaluation curves.

RQ2: Best/worst performing split exists for MAE; other metrics depend on evaluation curves.

Extra Tree + Perceptron

RQ1: MAE depends on evaluated curves; Miscoverage: fewer Extra Tree curves is better; Area depends on curves.

RQ2: MAE depends on curves; Miscoverage improves with more Perceptron curves; Area depends on curves.

QDA + SVC Sigmoid

RQ1: No clear trends except Area being lower when training only on Sigmoid curves.

RQ2: No significant trends observed.

Key Observations

- **Skewed distributions:** Results for MAE and Area show high variability
- **Seed dependency:** Performance varies significantly with random seed
- **Mixed evidence:** Some cases show performance reduction (typical for imbalanced learning), others show improvement
- **Learner-dependent:** Results heavily depend on the specific learner combination

Discussion & Future Work

Evidence Found:

- Performance reduction in some cases (expected for imbalanced learning)
- Performance improvement possible in certain scenarios
- Results are highly learner-dependent

Limitations:

Current LC-PFN training setup has inherent constraints that affect results.

Future Directions:

Sampling strategies could be implemented to improve performance in imbalanced learning scenarios.