# Landmark and Compare: Addressing the Algorithm Selection Problem via Problem Space Similarity

V. Kodderitzsch[1,2], Dr. J. Burger[2], Dr. A. van Delden[2], Dr. J. D. Karch[1]

[1] Leiden University, The Netherlands
[2] Statistics Netherlands (CBS), The Netherlands

**Keywords:** Algorithm Selection Problem, Landmarking, Data Generating Distribution, Meta Learning, Dataset Shift.

*Introduction* In official statistics, researchers from different national statistic offices often work on similar machine learning (ML) tasks, such as imputing non-response items in surveys, but each using their own local dataset. As part of a collaborative effort, would one group benefit from adopting the other group's ML algorithm? This is a non-trivial task as implementing a new ML pipeline from feature extraction to predictions is often a significant effort. This practical challenge of identifying the most effective algorithm for a given computational task is known as the Algorithm Selection Problem (ASP) [4]. In this study, we limited ourselves to continuous response and feature variables.

*Methods* Our initial investigation explored the creation of a normalized metric for comparing algorithm performances across datasets via a universal scale. However, this approach faces a fundamental theoretical constraint rooted in the Expected Prediction Error (EPE). The EPE contains an irreducible error term ($\sigma_\epsilon^2$) that comes from the data generating distribution $P_{(X,Y)}$ (DGD) [5, 1]. This term creates a dataset-specific error floor (i.e. performance ceiling) of an algorithm $A_i$ under $D \sim P_{(X,Y)}$:

$$\mathbb{E}_D[\text{error}_d(A_i)] \in [\sigma_\epsilon^2, \infty). \tag{1}$$

This insight shifts the perspective from "Are performance scores similar?" to "Are the learning problems functionally similar?", aligning more closely with the No Free Lunch theorem [5]. We hypothesize that *stable algorithm rankings serve as a robust proxy for the functional similarity of DGDs*. If two DGDs are similar, applying a diverse set of algorithms yields similar rank orders and thus high rank correlation (Kendall's $\tau \geq 0.4$). However, verifying the rank correlation requires fitting an identical set of candidate algorithms twice (once on each dataset) which is the very task we aim to avoid. We break this paradox with an insight from our initial exploratory simulations. We observed that algorithm rank stability is primarily driven by just two factors: (i) the noise proportion and (ii) the curvature (deviation from linearity) of the observable signal.

Our "Landmark and Compare" framework [2] estimates these two meta-features via landmarking [3], which is the process of fitting a diverse set of algorithms using default hyperparameters and no additional tuning. We approximate the noise by applying $1-\max(\text{NSE}^3)$ across landmarking algorithms and the signal curvature via the $\text{NRMSE}^4$ of a simple linear regression. The crux of the framework is the sequential decision rule. First, test whether the estimated noise proportions of the two datasets lie within a predetermined threshold (to allow for sampling variability). Only if noise proportions are similar, the NRMSE (curvature) values can be compared. Otherwise, the noise ($\sigma_\epsilon^2$) confounds the NRMSE (curvature) comparison due to the difference in (normalized) EPE scales. Second, if the curvature values also lie within their own threshold, the datasets are declared similar. If either test fails, the datasets are declared dissimilar.

*Results* Decision thresholds 0.2 (noise) and 0.15 (curvature) were empirically calibrated via grid search on 45 synthetic DGDs that varied in signal complexity and noise proportions. The framework was then validated on five real-world datasets by choosing one as the reference and the other four as the non-reference datasets. Landmarking set A (linear regression, elastic net, RF, XGB, SVM) and B (substituting Elastic Net for GAM) were used to estimate the meta-features for the decision rule. The ground truth rank correlations confirmed these predictions (4/4 correct), demonstrating robustness to minor variations in the landmarking sets A vs. B.

**Table 1.** Framework Validation Results using Set **A** for Landmarking the Reference

| **Dataset** | **Landmarking** | **Noise** ($\hat{n}$) | **Signal** ($\hat{s}$) | **Prediction** | **True Correlation** | **Outcome Validity** |
|---|---|---|---|---|---|---|
| Boston (Reference) | A | 0.091 | 0.572 | – | – | – |
| Ames | B | 0.143 | 0.421 | Similar | 0.8 | Correct |
| California | B | 0.174 | 0.615 | Similar | 1.0 | Correct |
| Insurance | B | 0.876 | 0.935 | NOT similar | −0.4 | Correct |
| Longley | B | 0.026 | 0.141 | NOT similar | −0.6 | Correct |

*Discussion* The proposed framework provides researchers with a new tool to assist them in their collaborative efforts. Instead of focusing on normalized metrics which are EPE dependent, the framework focuses on problem space similarity in a computationally inexpensive and easy to implement manner. However, the decision thresholds should be viewed only as sensible default hyperparameters and not as universal constants. Additionally, future work should further investigate edge cases of the proposed curvature metric as well as the robustness of the entire framework under more diverse landmarking sets A vs. B to replicate more realistic collaboration scenarios.

[3] Nash–Sutcliffe Efficiency (NSE): Predictive $R^2$ analog.

[4] Normalized Root Mean Squared Error (NRMSE): RMSE divided by the response standard deviation.

## References

1. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer Series in Statistics, Springer (2009). https://doi.org/10.1007/978-0-387-84858-7, http://link.springer.com/10.1007/978-0-387-84858-7
2. Kodderitzsch, V.: Landmark and Compare: Addressing the Algorithm Selection Problem via Problem Space Similarity. Master's thesis, Leiden University (2025)
3. Pfahringer, B., Bensusan, H., Giraud-Carrier, C.G.: Meta-learning by landmarking various learning algorithms. In: Proceedings of the Seventeenth International Conference on Machine Learning. pp. 743–750. ICML '00, Morgan Kaufmann Publishers Inc. (2000)
4. Rice, J.R.: The algorithm selection problem. In: Advances in Computers, pp. 65–118. Elsevier (1976). https://doi.org/10.1016/s0065-2458(08)60520-3, https://linkinghub.elsevier.com/retrieve/pii/S0065245808605203
5. Wolpert, D.H.: The lack of a priori distinctions between learning algorithms. Neural Computation **8**(7), 1341–1390 (1996). https://doi.org/10.1162/neco.1996.8.7.1341, https://ieeexplore.ieee.org/document/6795940