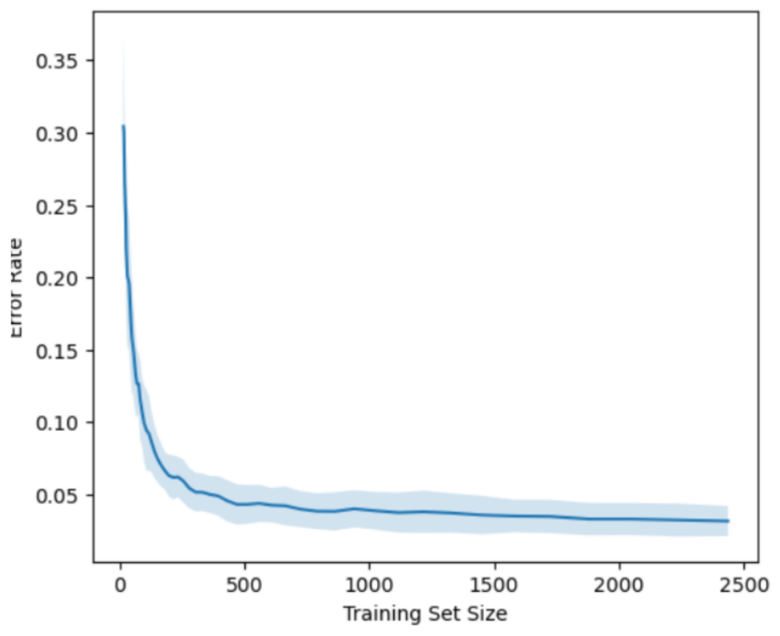
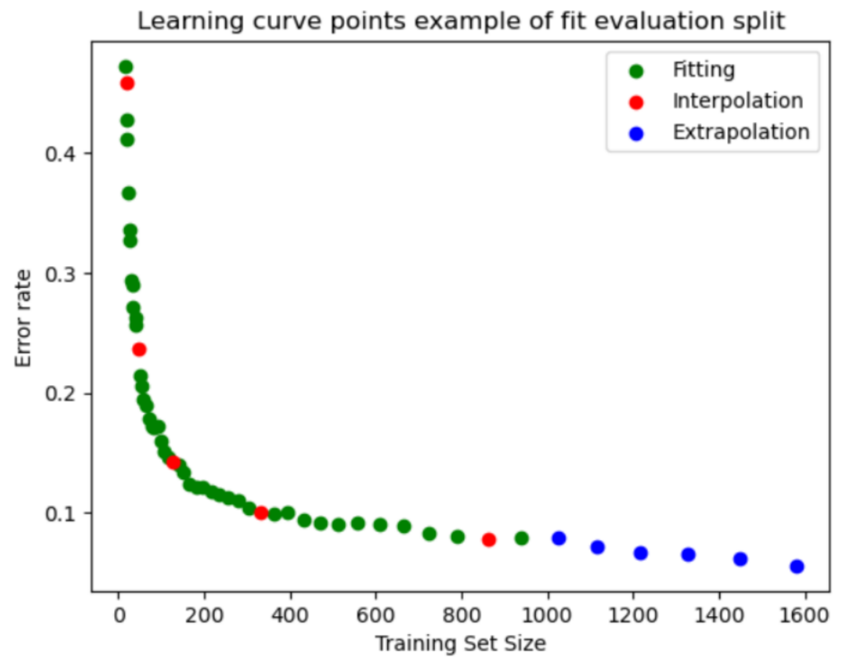
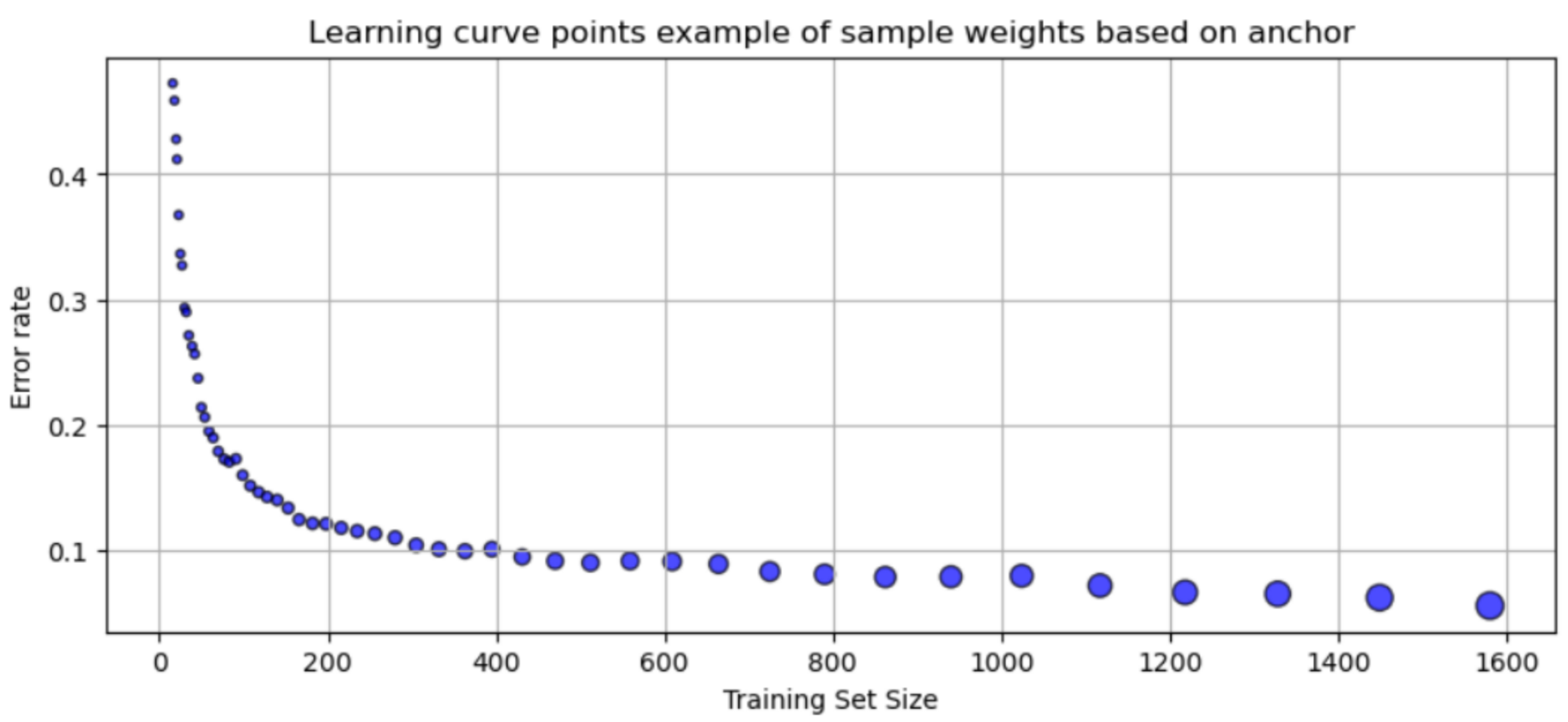
Introduction & Background

Learning curves describe the relationship between training data size and model performance. Curve fitting is commonly used to model these relationships, but the impact of sample weighting on learning curve fitting remains underexplored. This study investigates the effect of various sample weighting methods, including anchor, error rate and inverse-variance weighting, on improving curve fit accuracy.





Prior work, such as Brumen et al. (2014), analyzed curve functions but focused on a single learner and limited datasets[1], while Figureoa et al. (2012) applied simple weights but lacked robust extrapolation and statistical rigor[2]. Building on this, we analyze multiple learners, employ both interpolation and extrapolation, and evaluate statistical significance using the Wilcoxon signed-rank test to address existing limitations and advance the study of learning curves.

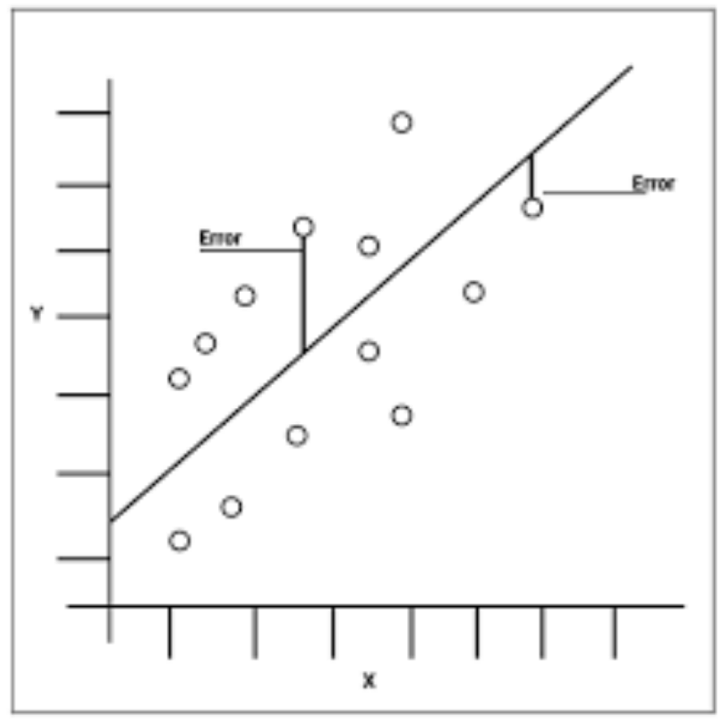


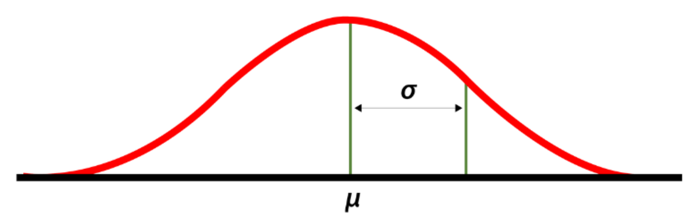
Research Question: How do various sample weighting methods improve learning curve fitting?

Method & Experiment

We optimize learning curves with mean squared error (MSE). To evaluate learning curve fitting, we use both interpolation (estimating performance within observed data ranges) and extrapolation (predicting beyond observed ranges), splitting datasets into fitting and evaluation sets accordingly.

This study explores three sample weighting methods—based on anchor size, error rate, and error variability (SD-Error)—with weights scaled to influence curve fitting outcomes, evaluating their effectiveness via MSE and statistical significance using the Wilcoxon signed-rank test.

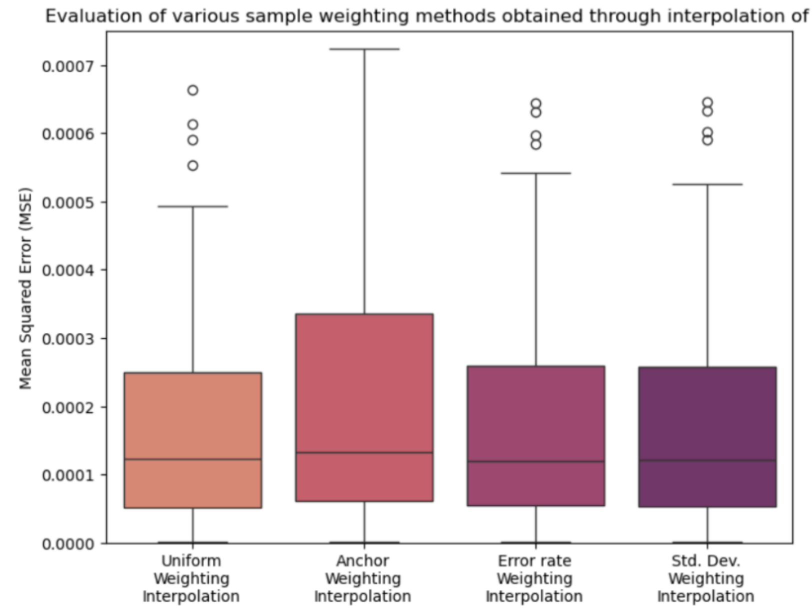


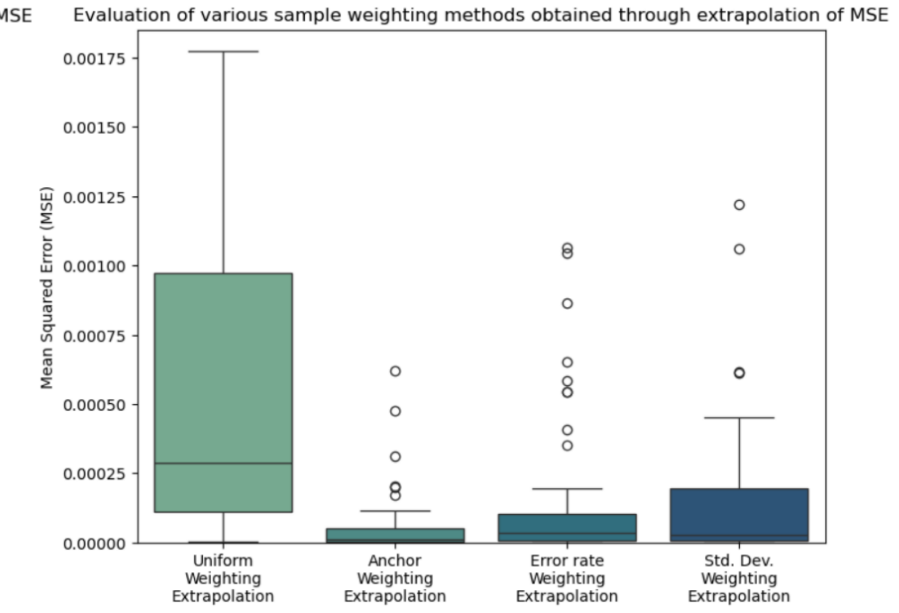


We use 24 learners from the LCDB 1.0[3] paper and 72 datasets from the OpenML databases, prioritizing well-behaved data on five learners. Data is split into fitting, interpolation, and extrapolation sets (80/10/10%), using the EXP3 function the LCDB paper for curve fitting and evaluating MSE on unseen test data. The sample weights were applied in the curve fit process, optimized by Levenberg-Marquardt to assess their impact.

Results & Conclusion

For the gradient boosting learner, we show that sample weighting methods perform worse than uniform weighting for interpolation but outperform it for extrapolation, with smaller means, medians, and variances. All learners showed similar trends, with extrapolation performing significantly better, evaluating with Wilcoxon (P<.001), when large weights increased on anchor and decreased on error rate and SD-Error.





These results show that sample weighting improves performance prediction for larger, unseen anchor values but it does not for anchors within the observed data range. Limitations include reliance on well-behaved curves, single data splits without cross-validation, hard-coded weight scales and time constraints. Future work could involve evaluating more diverse learner-dataset combinations, exploring dynamic methods for selecting weight scales, and using multiple data splits for robust evaluation. Greater computational resources would enable broader analyses with more complex models and larger datasets.

References

1. Boˇstjan BRUMEN, Ivan Rozman, Marjan Hericko, Aleˇs ˇCERNEZEL, and Marko H¨olbl. Best-fit learning curve model for the c4.5 algorithm. INFORMATICA,, 25:385–399, 10 2014.
2. Rosa L. Figueroa, Qing Zeng-Treitler, Sasikiran Kandula, and Long H. Ngo. Predicting sample size required for classification performance. BMC Medical Informatics and Decision Making, 12:8 – 8, 2012.
3. Felix Mohr, Tom J. Viering, Marco Loog, and Jan N. van Rijn. Lcdb 1.0: An extensive learning curves database for classification tasks. In Massih-Reza Amini, St´ephane Canu, Asja Fischer, Tias Guns, Petra Kralj, Novak, and Grigorios Tsoumakas, editors.

TU Delft

Author: Lapo den Hollander

Assistant Professor: Tom Viering
Supervisors: Cheng Yan, Taylan Turan