Stacking: An ensemble learning approach to predict student performance in PISA 2022 - Summary review

David Mendes[a]

[a]*ISCTE – Instituto Universitário de Lisboa, Portugal*

Abstract

This article aims to summarize the approach and results of öz et al. (2024), in which the authors explore employing stacking, boosting, and blending machine learning algorithms to predict student performance in large-scale assessments based on a wide range of predictors and compare their performance.

Keywords: PISA, Machine learning, Stacking, Boosting, Blending, Ensemble learning

1. Introduction

The study öz et al. (2024) uses the PISA 2022 Student questionnaire dataset, focusing on student responses and performance in subjects of mathematics, reading, and science. The authors employ machine learning algorithms to predict student performance, including stacking, blending, and boosting algorithms. The stacking method combines predictions from multiple models to create a meta-model that improves overall prediction performance. Boosting algorithms sequentially train models, each correcting errors from the previous ones, while blending uses a validation set to train the base learners. This study aims to measure the prediction performance of these methods.

2. Methods

2.1. Stacking

Stacking is an ensemble learning technique developed by Wolpert (1992) that combines several prediction models and uses their outputs as input for a next-level model (meta-model), aiming to improve overall prediction performanceöz et al. (2024). In the study öz et al. (2024), the authors divide stacking into level 0 and level 1, where layer 0 refers to where the different models generate distinct predictions, and level 1 combines predictions of level 0. For level 0, some of the different models that could be used are DTs, NNs, SVMs, and kNN, in which the resulting predictions are then used and combined in layer 1, in order to improve the overall prediction Wolpert (1992), the authors choose to use ridge regression, which can be useful to avoid overfitting Cui et al. (2021)

2.2. Boosting

Because the dataset used in this study is composed of missing data, it is imperative to implement learners that deal with missing data, the boosting algorithms used were XGBoost, HGB, and LightGBM öz et al. (2024)

Subject		Mathematics	Reading	Science
Metrics	Algorithm	Number (%) of the countries	Number (%) of the countries	Number (%) of the countries
Mean MAPE	Stacking	72 (90.00)	64 (80.00)	65 (81.25)
	Blending	4 (5.00)	4 (5.00)	6 (7.50)
	XGBoost	-	1 (1.25)	1 (1.25)
	LightGBM	2 (2.50)	6 (7.50)	3 (3.75)
	Blending & Stacking	-	2 (2.50)	2 (2.50)
	Stacking & XGBoost	-	-	2 (2.50)
	Blending & LightGBM	-	-	1 (1.25)
	LightGBM & XGBoost	2 (2.50)	1 (1.25)	-
	HGB & LightGBM	-	1 (1.25)	-
	XGBoost & HGB & LightGBM	-	1 (1.25)	-
Mean MAE	Stacking	74 (92.50)	69 (86.25)	71 (88.75)
	Blending	1 (1.25)	1 (1.25)	2 (2.50)
	XGBoost	1 (1.25)	2 (2.50)	3 (3.75)
	HGB	1 (1.25)	1 (1.25)	-
	LightGBM	3 (3.75)	7 (8.75)	4 (5.00)
Mean MSE	Stacking	75 (93.75)	70 (87.50)	71 (88.75)
	Blending	1 (1.25)	1 (1.25)	2 (2.50)
	XGBoost	-	3 (3.75)	1 (1.25)
	HGB	-	1 (1.25)	-
	LightGBM	4 (5.00)	5 (6.25)	6 (7.50)

Figure 1: The Number (%) of the countries exhibiting the lowest error values generated by each algorithm for all subjects öz et al. (2024)

2.3. blending

Blending and stacking are both techniques used in ensemble learning that share similarities. The key difference between them is that in blending, the base learners are trained using predictions obtained from the validation set instead of directly from the training setöz et al. (2024)

3. Results

The study found that the stacking method produced the lowest error metrics (Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Mean Squared Error (MSE)) for most countries compared to boosting and blending. Robust linear mixed-effects models indicated that stacking had significantly lower error metrics across all subjects as showed in 1.

4. Conclusion

The study öz et al. (2024) concludes that stacking is a favorable option for better performance, stable and accurate predictions based on evaluation with various metrics such as MAPE, MAE, MSE and RMSE where it obtained a significant metric score in all subjects compared to boosting and blending. Due to this study only evaluating the PISA dataset, I think it would be interesting to evaluate it with other datasets to improve this analysis further

References

Cui, S., Yin, Y., Wang, D., Li, Z., Wang, Y., 2021. A stacking-based ensemble learning method for earthquake casualty prediction. Applied Soft Computing 101, 107038. URL: https://www.sciencedirect.com/science/article/pii/S1568494620309765, doi:https://doi.org/10.1016/j.asoc.2020.107038.

Wolpert, D.H., 1992. Stacked generalization. Neural Networks 5, 241–259. URL: https://www.sciencedirect.com/science/article/pii/S0893608005800231, doi:https://doi.org/10.1016/S0893-6080(05)80023-1.

öz, E., Bulut, O., Cellat, Z.F., Yürekli, H., 2024. Stacking: An ensemble learning approach to predict student performance in pisa 2022. Education and Information Technologies , 1–27doi:10.1007/s10639-024-13110-2.