Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

The Effect of Feature Reliability on the Generalization of Machine Learning Models in Educational Data

Yingbin ZHANG*
Institute of Artificial Intelligence in Education, South China Normal University, China
*zyingbin@m.scnu.edu.cn

Abstract: Reliability quantifies the extent of measurement errors in the observed feature scores and is an important quality indicator of measurements in educational research. However, the effect of feature reliability is underexplored in educational studies that use machine learning techniques. Understanding this effect is critical because the most of common features in education are contaminated by measurement errors. Recent research has revealed that the low reliability of features damages the prediction accuracy of machine learning models. The current study proposes that feature reliability also influences the generalization of machine learning models. This paper provides mathematical proof for the notion and further supports it via analyses on two empirical educational datasets. The results of data analyses also indicated that the effect of feature reliability on model generalization was moderated by the model complexity but not related to the model accuracy. Approaches to mitigate the impact of feature reliability are discussed.

Keywords: Machine learning in education, feature reliability, measurement error, model generalization

1. Introduction

Feature engineering is a critical step in traditional machine learning (ML), significantly impacting model performance (Dong & Liu, 2018; Verdonck et al., 2021). Deep learning models, such as CNN, RNN, and Transformer, can automatically extract and learn complex features from raw data, reducing the need for feature engineering. However, in cases of limited data or the need to extract certain theory-driven features, appropriate feature engineering can still enhance the performance of deep learning models. Feature engineering focuses on the contribution of features to the models' predictive and classification performance and selects features accordingly. Nevertheless, in addition to feature importance, the reliability and validity of features (or variables) are also crucial quality indicators in educational research. As ML becomes increasingly popular in educational studies, understanding the roles of reliability and validity in ML in education is emergent. The current study focuses on feature reliability because it is relatively underexplored in the field of ML in education.

1.1 Feature Reliability and Its Role in Educational Research

Except for demographics, common variables in education, such as test scores, self-reported scale scores (e.g, attitudes and motivation), and learning process data (e.g., classroom dialogues and action logs in digital learning environments), contain measurement errors (AERA et al., 2014). Reliability is often used to quantify the extent of measure errors in test scores and self-reported scale scores. Although learning process data are objective records of events during learning, variables extracted from the data still contain measurement errors (Gray & Bergner, 2022), which can also be assessed using reliability (Zhang et al., 2024).

Studies have well demonstrated the influence of variable reliability on the credibility of results in inferential statistical analysis in educational research. Specifically, low reliability or

may obscure the true relationships between variables due to the substantial measurement errors in the data. For instance, if the reliability of students' test scores is poor, it is impossible to accurately infer the relationship between scores and other factors (e.g., teaching methods and learning environments), leading to biased parameter estimates in statistical models (Rhemtulla et al., 2020). Low reliability can also decrease the statistical power of variable correlation tests and increase the false discovery rate (Parsons et al., 2019).

1.2 The Impact of Feature Reliability on the Performance of Machine Learning Models

Several studies have investigated the impact of feature reliability on the prediction performance of ML models. Jacobucci and Grimm's (2020) simulation study found that low feature reliability worsened the prediction performance of linear regression and boosting regression models, with increasing sample size only slightly mitigating this effect. In McNamara et al.'s (2022) simulation study, the advantages of random forest and XGBoost over linear regression disappeared as feature reliability increased, even when the true relationships between features and outcomes were nonlinear. They concurred with Jacobucci and Grimm, concluding that measurement errors could prevent complex models (e.g., random forest and XGBoost) from capturing nonlinear relationships between features and labels, resulting in underfitting. Additionally, McNamara et al. found that sample size did not mitigate the effect of feature reliability. An empirical study replicated these findings (Gell et al., 2024). As such, researchers have suggested focusing on collecting reliable features for predictive modeling, in addition to employing complex modeling techniques (Jankowsky et al., 2024; Smith & Murayama, 2023). This is particularly relevant to educational applications of ML because, as mentioned earlier, most common features in education are contaminated by measurement errors.

Previous studies have argued that the negative effect of low feature reliability on prediction performance is due to underfitting caused by measurement errors (Jacobucci & Grimm, 2020; McNamara et al., 2022). However, it is unclear whether the negative effect is also related to overfitting. Specifically, whether low feature reliability may enlarge the gap between training performance and test performance of ML models, thereby damaging model generalization. This is an important issue in education because we want a predictive model generalizable to new students. For instance, a dropout prediction model would be useless if it could not predict the dropout risk for new students. Therefore, this study aims to investigate the impact of feature reliability on the generalization of ML models in education. We elucidate this impact through mathematical deduction and empirical data analysis.

2. The Effect of Feature Reliability on the Generalization of Machine Learning Models

In standardized assessment, reliability refers to the consistency among multiple observations of a construct and quantifies the extent of measurement error in observations (AERA et al., 2014). According to classical testing theory (CTT), the relationship between the reliability of a construct, observed value $\hat{x}$, and true value x can be expressed as (Jacobucci & Grimm, 2020):

$$\hat{x} = \sqrt{reliability} * x + e_x, \quad (1)$$

$$x = \frac{\hat{x} - e_x}{\sqrt{reliability}}, \quad (2)$$

where e_x is the measurement error and obeys a normal distribution $N(0,1-\text{reliability})$. The lower the reliability, the greater the variance of e_x.

Let $f(\cdot)$ denote the true mapping function between feature X and outcome Y, then $Y = f(X) + e_y$, where e_y is the irreducible error. For simplicity, this study assumes that the outcome does not contain measurement errors. Given that the observed value of X is $\hat{x}$, the observed value of Y can be represented as

$$y(X = \hat{x}) = f\left(\frac{\hat{x} - e_x}{\sqrt{reliability}}\right) + e_y\,. \quad (3)$$

Let $\hat{f}(\cdot)$ be a ML model, then the predicted value of Y can be represented as $\hat{y} = \hat{f}(\hat{x})$, with the prediction error being

$$y - \hat{y} = f\left(\frac{\hat{x} - e_x}{\sqrt{reliability}}\right) - \hat{f}(\hat{x}) + e_y\,. \quad (4)$$

Assuming that both the observed values of *X* in a training sample and a test sample are $\hat{x}$, with measurement errors being e_x^{train} and e_x^{test}. The true value in the training sample is $\frac{\hat{x}-e_x^{train}}{\sqrt{reliability}}$, while that in the test sample is $\frac{\hat{x}-e_x^{test}}{\sqrt{reliability}}$. Accordingly, the observed values of *Y* are $f\left(\frac{\hat{x}-e_x^{train}}{\sqrt{reliability}}\right) + e_y^{train}$ and $f\left(\frac{\hat{x}-e_x^{test}}{\sqrt{reliability}}\right) + e_y^{test}$ in the training and test samples, respectively. The predicted values of *Y* in both training and test samples are $\hat{f}(\hat{x})$. The model training process aims to minimize the prediction error of the ML model $\hat{f}(\cdot)$ in the training data, i.e., let it approximate $f\left(\frac{\hat{x}-e_x^{train}}{\sqrt{reliability}}\right)$. Therefore, the prediction error in the test data is generally larger than that in the training data. The difference between the two prediction errors is

$$f\left(\frac{\hat{x} - e_x^{test}}{\sqrt{reliability}}\right) - f\left(\frac{\hat{x} - e_x^{train}}{\sqrt{reliability}}\right) + e_y^{test} - e_y^{train}. \quad (5)$$

Decreases in the reliability of *X* lead to increases in the variances of e_x^{test} and e_x^{train}, which, in turn, enlarge the variances of $f\left(\frac{\hat{x}-e_x^{test}}{\sqrt{reliability}}\right)$ and $f\left(\frac{\hat{x}-e_x^{train}}{\sqrt{reliability}}\right)$. Consequently, the variance of $f\left(\frac{\hat{x}-e_x^{test}}{\sqrt{reliability}}\right) - f\left(\frac{\hat{x}-e_x^{train}}{\sqrt{reliability}}\right)$ becomes larger, and the probability of obtaining a large $f\left(\frac{\hat{x}-e_x^{test}}{\sqrt{reliability}}\right) - f\left(\frac{\hat{x}-e_x^{train}}{\sqrt{reliability}}\right)$ via random sampling increases. Thus, given the sample size, the difference in the prediction errors of ML models in the training and test data enlarges as the reliability of features decreases. That is, the model generalization becomes worse.

3. Empirical examples

This section illustrates the impact of feature reliability on model generalization using two empirical datasets. Table 1 summarizes the characteristics of the datasets. The first dataset was from an undergraduate introductory programming course at a Midwestern U.S. university (Zhang et al., 2023), and the second was from the PISA 2012 Shanghai data[1]. Due to the questionnaire rotation design in PISA 2012, only one-third of the participants responded to the 25 feature scales used in this study. Thus, only these students' data were used. McDonald's *ω* coefficient was used as the reliability indicator (Flora, 2020).

3.1 Analysis

Previous studies on feature reliability in ML have used both linear and ensemble modeling to investigate whether the impact of feature reliability depends on model complexity (Jacobucci & Grimm, 2020; McNamara et al., 2022). This study adopted the same approach, using one simple linear model (lasso regression) and two complex ensemble methods (random forest and XGBoost). These ensemble methods were also used in the previous studies. We selected lasso regression because of its capabilities in preventing overfitting and handling multicollinearity.

Each model used the high and low reliability features separately (see Table 1). The cut-off was the median *ω*. In the data from the programming course, a feature lower than the median was assigned to the high reliability group because its meaning was closer to this group than the low reliability group. The hyperparameters were the same between models with high

[1] https://www.oecd.org/en/data/datasets/pisa-2012-database.html#data

reliability features and models with low reliability features to ensure that differences in model generalization were not due to hyperparameters variations. We employed 10-fold cross-validation, where the data were randomly divided into 10 parts. Each model was trained 10 times, with 9 parts for training and 1 part for testing each time. The root mean squared error (*RMSE*), R^2, and linear correlation coefficient (*r*) were used as prediction performance metrics. After each training, the metrics for training and test data as well as their differences (referred to as the generalization gap) were computed. The final value of each metric was the mean over the 10 times of training.

Table 1. *Characteristics of the datasets*

Dataset	Students	Outcome	Source of feature	The number of features and examples		The means and ranges of ω	
				High reliability	Low reliability		
Introductory programming course	607	Final exam scores	Programming traces	6 features, e.g., familiarity with Java syntax and semantics	6 features, e.g., accuracies in debugging syntactic and semantic errors	.68 ; .55 ~ .83	.45 ; .36 ~ .56
PISA 2012 Shanghai	1730	Math, science, reading test scores	Self-reported scales	13 features, e.g., math anxiety, math self-concept	12 features, e.g., persistence, willingness to learn math	.87 ; .81 ~ .92	.65 ; .45 ~ .78

3.2 Results

Figure 1 shows the generalization gap in each dataset. The bar height in the column *RMSE* represents *test performance - training performance*, while in the columns R^2 and *r*, it represents *training performance - test performance*. In most cases, the generalization gap in the high reliability group was significantly smaller than in the low reliability group. Paired sample *t*-tests indicate that this difference is statistically significant for all performance metrics ($p < .01$; Cohen's $d = 0.96$, 1.26, and 1.38 for *RMSE*, R^2, and *r*, respectively).

The effect of feature reliability on the generalization gap is moderated by model complexity. The differences in the generalization gap between high and low reliability groups were larger in more complex models (random forest and XGBoost) than in simpler models (lasso regression).

The effect of feature reliability on generalization difference is not related to model performance. For example, in the case of predicting the final exam scores of programming course, the *RMSE*, R^2, and *r* of XGBoost in the training data were almost identical when using high reliability features versus using low reliability features, but the generalization difference was distinct.

4. Discussion and next steps

This study demonstrates that in educational contexts, feature reliability may affect the generalization of ML models: the lower feature reliability is, the weaker model generalization is. Therefore, when applying ML in education, it is necessary to assess feature reliability and use high reliability features whenever possible. Using advanced ML techniques does not eliminate the necessity of collecting high quality data with reliable features. In contrast, complex models are more prone to the impact of measurement errors, as suggested by the findings in the current and prior studies (Jacobucci & Grimm, 2020; McNamara et al., 2022).

Smith and Murayama (2023) recommended two solutions to mitigate the impact of feature reliability and measurement errors in ML. One solution involves using dimension reduction algorithms (e.g., principal component analysis) to compress multiple low-reliability features into new component features that suffer less from measurement errors. The limitations of this solution are that it may not effectively reduce the measurement errors and

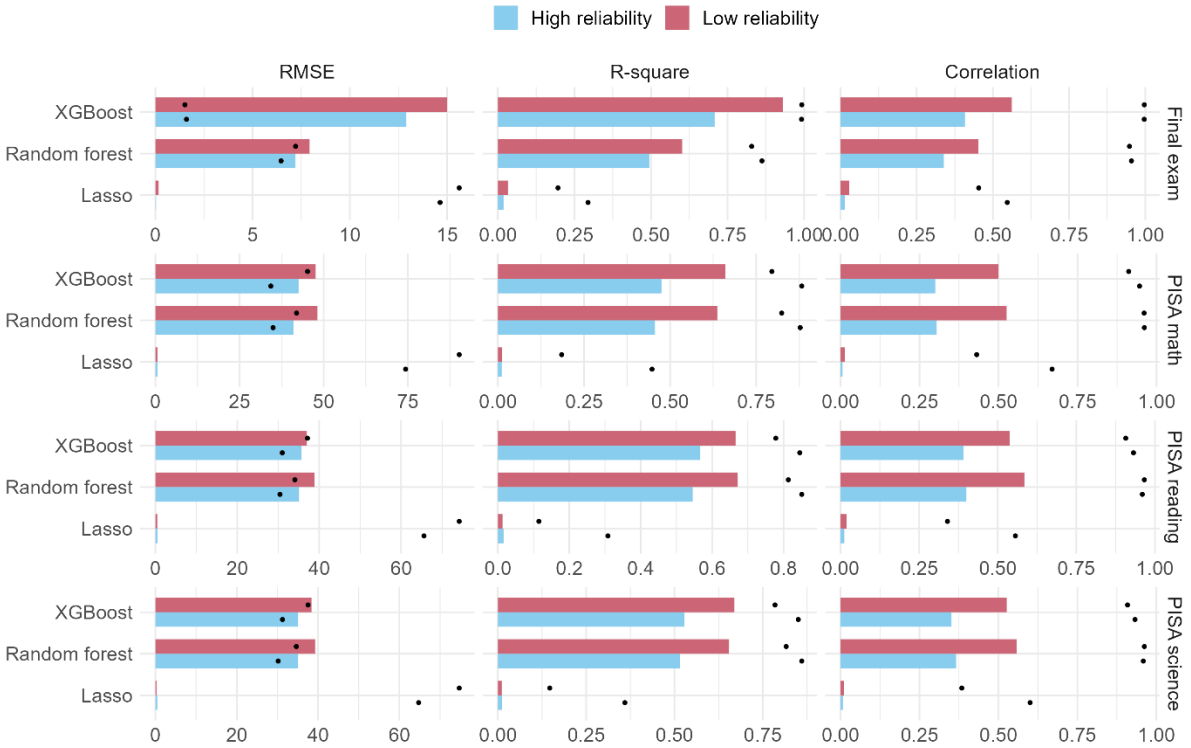

Figure 1. Generalization difference in prediction performance between training and test data
Note. Black dots represent the training performance. The bar height indicates the difference between training and test performance.

that the new component features may be difficult to interpret. Another solution is combining ML with structural equation modeling (SEM) to account for measurement errors, such as regularized SEM (Brandmaier & Jacobucci, 2023). The limitation of this SEM approach is that SEM may regard some useful information as measurement errors and discard it. In addition, there are few available options for combining ML and SEM. Alternatively, researchers may apply reliability-based loss functions in the feature engineering phase to select features that significantly contribute to model prediction and have high reliability (Grimm & Jacobucci, 2021). However, this approach demands a high level of technical expertise. In summary, effective approaches to reduce the impact of feature reliability in ML are under development. Future research may develop more user-friendly tools to automatically conduct the selection of high reliability features. Alternatively, with the rapid development of generative artificial intelligence (e.g., GPT4 and Gemini), it may be feasible to develop agents based on these techniques that assist researchers in feature selection.

This study does not examine the impact of feature reliability on the generalization of classifiers when the outcome variable is categorical. We also do not systematically investigate how the effect of feature reliability may change with sample size. It is also unclear that, given a particular sample size, what is the minimum reliability levels to ensure a reasonable model generalization. We plan to address these issues in the next step via systematic simulations and more empirical data analyses.

Acknowledgements

We thank Dr. Luc Paquette for sharing the programming course dataset used in this study. This work was supported by the National Key Research and Development Program of China (2023YFC3305704).

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing (7ed.)*. American Educational Research Association.

Brandmaier, A. M., & Jacobucci, R. C. (2023). Machine learning approaches to structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 722-739). Guilford Press.

Dong, G., & Liu, H. (2018). *Feature engineering for machine learning and data analytics*. CRC press.

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using r to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science*, 3(4), 484-501. http://doi.org/10.1177/2515245920951747

Gray, G., & Bergner, Y. (2022). A practitioner's guide to measurement in learning analytics: decisions, opportunities, and challenges. In C. Lang, G. Siemens, A. F. Wise, D. Gašević, & A. Merceron (Eds.), *Handbook of Learning Analytics* (2ed., pp. 20-28). SoLAR. http://doi.org/10.18608/hla22.002.

Gell, M., Eickhoff, S. B., Omidvarnia, A., Küppers, V., Patil, K. R., Satterthwaite, T. D., Müller, V. I., & Langner, R. (2024). The burden of reliability: How measurement noise limits brain-behaviour predictions. *bioRxiv*. http://doi.org/10.1101/2023.02.09.527898

Grimm, K. J., & Jacobucci, R. (2021). Reliable trees: Reliability informed recursive partitioning for psychological data. *Multivariate Behavioral Research*, 56(4), 595-607. http://doi.org/10.1080/00273171.2020.1751028

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, 15(3), 809-816. http://doi.org/10.1177/1745691620902467

Jankowsky, K., Krakau, L., Schroeders, U., Zwerenz, R., & Beutel, M. E. (2024). Predicting treatment response using machine learning: A registered report. *British Journal of Clinical Psychology*, 63(2), 137-155. http://doi.org/10.1111/bjc.12452

McNamara, M. E., Zisser, M., Beevers, C. G., & Shumake, J. (2022). Not just “big” data: Importance of sample size, measurement error, and uninformative predictors for developing prognostic models for digital interventions. *Behaviour Research and Therapy*, 153, 104086. http://doi.org/10.1016/j.brat.2022.104086

Parsons, S., Kruijt, A., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science*, 2(4), 378-395. http://doi.org/10.1177/2515245919879695

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, 25(1), 30-45. http://doi.org/10.1037/met0000220

Smith, G., & Murayama, K. (2023). Machine learning meets traditional statistical methods in psychology: Challenges and future directions. *OSF Preprint*. https://doi.org/10.31219/osf.io/6xt82

Verdonck, T., Baesens, B., Óskarsdóttir, M., & Vanden Broucke, S. (2021). Special issue on feature engineering editorial. *Machine Learning* http://doi.org/10.1007/s10994-021-06042-2

Zhang, Y., Paquette, L., Pinto, J. D., & Fan, A. X. (2023). Utilizing programming traces to explore and model the dimensions of novices' code-writing skill. *Computer Applications in Engineering Education, 31*(4), 1041-1058. http://doi.org/https://doi.org/10.1002/cae.22622

Zhang, Y., Ye, Y., Paquette, L., Wang, Y., & Hu, X. (2024). Investigating the reliability of aggregate measurements of learning process data: From theory to practice. J*ournal of Computer Assisted Learning*, *40*(3), 1295-1308. http://doi.org/10.1111/jcal.12951