# The Effect of the Distribution of Predictions of User Models

Eric G. Van Inwegen Yan Wang Seth Adjei Neil Heffernan
100 Institute Rd
Worcester, MA, 01609-2280
+1-508-831-5569
{egvaninwegen, ywang14, saadjei, nth} @wpi.edu

## ABSTRACT

We hypothesize that there are two basic ways that a user model can perform better than another: 1.) having test data averages that match the prediction values (we call this the *coherence* of the model) and 2.) having fewer instances near the mean prediction (we call this the *differentiation* of the model). There are several common metrics used to determine the goodness of user models; these metrics conflate coherence and differentiation. We believe that user model analyses will be improved if authors report the differentiation, as well as to include an ordering metric (e.g. AUC/A' or $R^2$) and an error measurement (Efron's $R^2$, RMSE or MAE). Lastly, we share a simplified spreadsheet that enables readers to examine these effects on their own datasets and models.

## 1. INTRODUCTION AND BACKGROUND

One of the goals of many in the online educational community is to more accurately predict whether a student will get the next question correct. In order to predict student responses, algorithms such as Knowledge Tracing [2], Performance Factors Analysis [6], and tabling methods [10] etc. have been developed. (See [3] for a thorough review of various user models.) Looking at only papers presented at EDM 2014, we find more than 6 new models or modifications proposed in the full papers alone [14]. Common metrics used to determine when a model is better than another include AUC/A', RMSE, MAE, and R-squared. There has been some work done (e.g. [1, 4]) looking into what metrics to use and how to interpret them [5, 11].

One can argue that current models predict the probability that a student-problem-instance (hereafter "instance") will be correct. Models such as Knowledge-Tracing ("KT"), Performance Factors Analysis ("PFA"), and their derivatives create a theoretically continuous range of predictions from 0.00 to 1.00. Even tabling models (eg. [10]) may predict a (near) continuous range of values through regressions. We argue that there are two properties of a model that will make it more accurate: 1.) How well a prediction matches the aggregate test-data, and 2.) How well the model can make predictions away from the mean.

### 1.1 Our Definitions

#### *1.1.1 "Coherence"*

Given a large enough data-set, we argue that an accurate model's predictions should match the test data average for a given group of instances. For example, if a model were to identify a group of instances and give that group a predicted value of 0.25, we argue that the model is most accurate when exactly one out of every four students in that condition gets the correct answer. If the model predicts 0.25, but only one out of every ten gets it right, the model's "scores" by most metrics will be improved, however, it is not as accurate as a similar model that groups that same instances together, but predicts 0.10.

#### *1.1.2 "Differentiation"*

A naive model of student knowledge might use the average score from a training dataset and predict with that probability for all instances. Arguably, more complicated user models seek to find reasons *not* to do this. The more features that a model can incorporate to move predictions away from the mean value, the better a model is at not making the mean prediction. We use the term "differentiation" in much the same way as "distribution", but do so to avoid possible confusion with the distribution of the training data.

## 2. METHODS

In order to visualize the impact of differentiation and coherence on the various metrics, we generate not synthetic data, but rather synthetic model outputs. To examine the effect of differentiation, a spreadsheet was created that allows the user to input prediction value, test group average, and number of instances within that group, for up to eleven groups. The spreadsheet then calculates values for AUC, A', $R^2$, Efron's $R^2$, RMSE, and MAE. A publicly shared copy of the spreadsheet can be found at: http://tinyurl.com/kznthk7. In addition to using synthetic data, the results of three models fitted to real data are explored.

## 3. RESULTS AND DISCUSSION

Figure 1 is a plot of the six metrics as a differentiation changes from an exceptionally steep "V" to flat to increasingly steep "A". All "models" have perfect coherence. E.g., when the model predicts 0.20, exactly 2/10 students are correct. From Figure 1, we can see that differentiation plays a role in user model "scores".

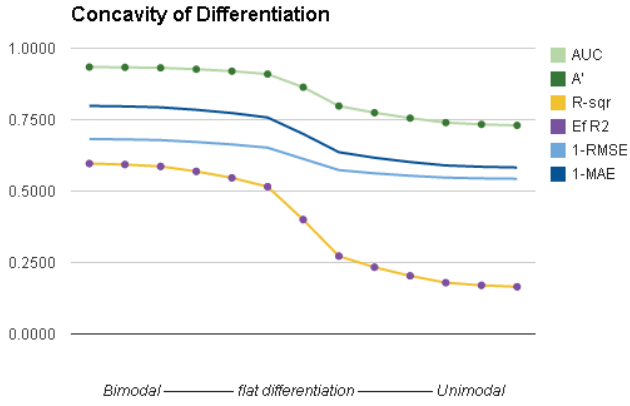


**Figure 1:** Chart of synthetic model outcomes showing the basic effect of changing differentiation.

To see if these ideas have merit on real data, we analyze three different models fitted to the same (~400K instance) dataset. In another paper [16], we have submitted a new user model. In that paper, the new model, called "SuperBins" (SB), is compared to Knowledge Tracing and Performance Factors Analysis, and found to be "better", according to RMSE, $R^2$, and AUC. If we create a frequency table of 11 groups, we will certainly lose precision, but the analysis is useful. To do so, we average the prediction values (according to their frequency) across eleven equal lengths *of prediction values* of the data set; we do the same for the test data

**Table 1:** A coherence-frequency table of results from three knowledge models trained and tested on the same real dataset (80/20). Model results have been averaged across 11 intervals for demonstration purposes. The prediction and test values are the weighted averages of each model within the ranges on the left.

| | **SB** | | | | *KT* | | | | PFA | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Range | **pred** | **test** | **n** | | *pred* | *test* | *n* | | pred | test | n |
| 0.0000 - 0.0909 | **0.08** | **0.00** | **5** | | *n/a* | *n/a* | *0* | | 0.01 | 0.78 | 9 |
| 0.0910 - 0.1818 | **0.14** | **0.13** | **516** | | *0.16* | *0.75* | *4* | | 0.13 | 0.53 | 17 |
| 0.1819 - 0.2727 | **0.22** | **0.23** | **892** | | *0.24* | *0.30* | *64* | | 0.23 | 0.46 | 56 |
| 0.2728 - 0.3636 | **0.31** | **0.32** | **1829** | | *0.33* | *0.28* | *704* | | 0.31 | 0.49 | 168 |
| 0.3637 - 0.4545 | **0.41** | **0.41** | **3235** | | *0.40* | *0.36* | *2565* | | 0.41 | 0.42 | 643 |
| 0.4546 - 0.5454 | **0.50** | **0.51** | **4878** | | *0.51* | *0.48* | *6978* | | 0.50 | 0.49 | 3539 |
| 0.5455 - 0.6363 | **0.60** | **0.60** | **6355** | | *0.60* | *0.61* | *8776* | | 0.61 | 0.59 | 7376 |
| 0.6364 - 0.7272 | **0.69** | **0.69** | **9772** | | *0.69* | *0.71* | *12149* | | 0.70 | 0.70 | 25819 |
| 0.7273 - 0.8181 | **0.79** | **0.79** | **25296** | | *0.78* | *0.78* | *18518* | | 0.77 | 0.78 | 25580 |
| 0.8182 - 0.9090 | **0.86** | **0.87** | **23347** | | *0.87* | *0.85* | *23600* | | 0.87 | 0.87 | 13811 |
| 0.9091 - 1.0000 | **0.97** | **0.97** | **3074** | | *0.95* | *0.95* | *5841* | | 0.97 | 0.96 | 2181 |
| Metrics | **AUC** **0.728** | **$R^2$** **0.145** | **RMSE** **0.406** | | *AUC* *0.710* | *$R^2$* *0.115* | *RMSE* *0.413* | | AUC 0.653 | $R^2$ 0.058 | RMSE 0.426 |
| | **stdev (pred): 0.166** | | | | *stdev(pred): 0.147* | | | | stdev(pred): 0.107 | | |

averages. E.g.., the average prediction value from 0 to 0.0909, as weighted by the frequency of each prediction was found to be 0.08 for the SuperBins model. There were no predictions in that range for KT. There were nine for PFA (eight were right), with an average prediction value of 0.01.

The analysis of coherence shows that, from 0.60 and up, all three models are reasonably accurate; i.e., the predictions closely match the test data averages. However, KT has over-predicted in the three largest of the 6 groups below 0.60. PFA appears to be reasonably consistent; however, one could argue that PFA consistently *under*-predicts in this range. Others [7] have previously reported on KT over-reporting. With this analysis, we can say that PFA has done the worst of the three at moving instances away from the mean. The major reason why SB scores so well against the other two could be its ability to bring more predictions below 0.50, while maintaining coherence.

The easiest way to measure the differentiation of the prediction values might be to report the standard deviation of prediction values. As a way to compare to the "ideal" (for that dataset), we could report either the standard deviation of the test data (0.439), or the standard deviation of the training data (0.440).

# 4. CONCLUSION

There are times when the metrics "scoring" user models disagree; in addition, it may be helpful for a deeper comparison.

We conclude that, if we are to accurately compare knowledge predicting models to each other, we need to look at new metrics, in addition to a mix of old metrics. We do not believe that we are proposing the "ultimate" single metric that will definitively state which model is "better". We are stating that we believe model comparison is improved when it contains (AUC or A', or $R^2$), and (Efron's $R^2$, RMSE, or MAE) and the standard deviation of the predictions. A more thorough comparison might also include coherence-frequency table analysis in an attempt to identify regions of habitual over or under prediction.

# 5. ACKNOWLEDGEMENTS

We would like to thank Ryan Baker and Joseph Beck for taking the time to discuss these ideas with us and make suggestions. We also acknowledge and thank funding for ASSISTments from the NSF (1316736, 1252297, 1109483, 1031398, 0742503, and 1440753), the U.S. Dept. of Ed. GAANN (P200A120238), ONR's "STEM Grand Challenges," and IES (R305A120125, R305C100024).

# 6. REFERENCES

[1] Beck, J. E., & Xiong, X. (2013). Limits to accuracy: How well can we do at student modeling. *Educational Data Mining*.

[2] Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction, 4(4), 253-278.

[3] Desmarais, M. C., & d Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, *22*(1-2), 9-38.

[4] Dhanani, A., Lee, S. Y., Phothilimthana, P., & Pardos, Z. (2014). A comparison of error metrics for learning model parameters in bayesian knowledge tracing. Technical Report UCB/EECS-2014-131, EECS Department, University of California, Berkeley.

[5] Fogarty, J., Baker, R. S., & Hudson, S. E. (2005). Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction. *Proceedings of Graphics Interface 2005*. Canadian Human-Computer Communications Society.

[6] Pavlik Jr, P. I., Cen, H., & Koedinger, K. R. (2009). Performance Factors Analysis--A New Alternative to Knowledge Tracing. *Online Submission*.

[7] Qiu, Y., Pardos, Z. & Heffernan, N. (2012). Towards data driven user model improvement. Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference. Florida Artificial Intelligence Research Society (FLAIRS 2012). pp. 462-465.

[8] Stamper, J., Pardos, Z., Mavrikis, M., McLaren, B.M. (eds.) Proceedings of the 7th International Conference on Educational Data Mining.

[9] Van Inwegen, E. G., Adjei, S. A., Wang, Y., & Heffernan, N. T. "Using Partial Credit and Response History to Model User Knowledge" *accepted into Educational Data Mining 2015*.

[10] Wang, Y., & Heffernan, N. T. (2011). The" Assistance" Model: Leveraging How Many Hints and Attempts a Student Needs. *FLAIRS Conference*.

[11] Yudelson, M., Pavlik Jr, P. I., & Koedinger, K. R. (2011). User Modeling–A Notoriously Black Art. *User Modeling, Adaption and Personalization*, 317-328.