Machine Learning for Education Laboratory
ML4ED

How models for knowledge tracing depend on data

Authors: Tikhon Parshikov, Mehrdad Kiani Oshtorjani

Machine Learning for Behavioral Data (MLBD) Course, EPFL, 2022

1 INTRODUCTION + RESEARCH QUESTION

We are working with dataset from **Lernavi** that is an instrument for promoting part of the basic technical study skills in German and Mathematics. Using this dataset, we are going to explore **how much the users are learning** and **how model performance depends on data**.

2 METHODOLOGY

In this work we used 4 main models for knowledge tracing such as *Bayesian Knowledge Tracing* (**BKT**), *Additive Factor Analysis* (**AFM**), *Performance Factor Analysis* (**PFA**) and *Deep Knowledge Tracing* (**DKT**).

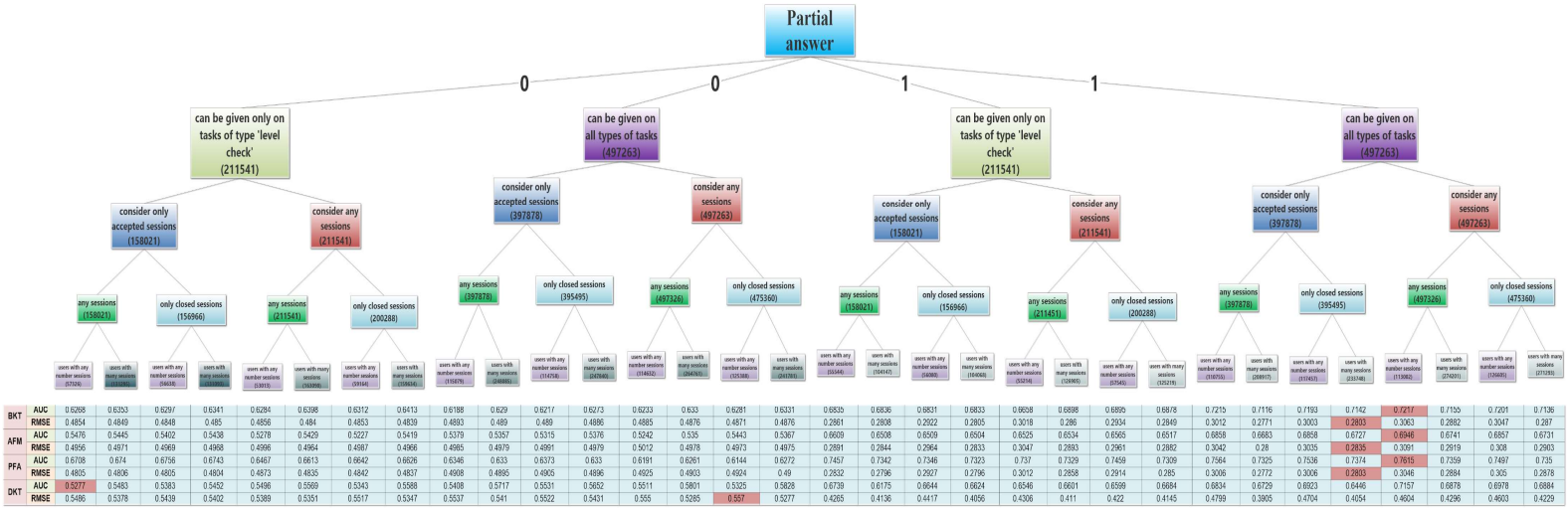
3 DATA PREPROCESSING

We divided the given data in two different datasets: one grouped by subjects (Math and German) and one grouped by topics (a lot of different topics related to these subjects).

For the *subject* dataset we tried different types of data preparing (5 levels: assuming partial answers as correct/incorrect, consider answers on all types of questions or not, consider only closed sessions or not, consider only accepted sessions or not and consider only users with many sessions or any) to see how models perform in connection with data preprocessing.

For the *topics* dataset we explored the performance of models built on most and least popular topics.

4.1 RESULTS (Subjects)



BKT	AUC	0.6268	0.6353	0.6297	0.6341	0.6284	0.6398	0.6312	0.6413	0.6188	0.629	0.6217	0.6273	0.6233	0.633	0.6281	0.6331	0.6835	0.6836	0.6831	0.6833	0.6938	0.6898	0.6895	0.6878	0.7215	0.7116	0.7193	0.7142	0.7217	0.7155	0.7201	0.7136
	RMSE	0.4854	0.4849	0.4848	0.485	0.4856	0.484	0.4853	0.4839	0.4893	0.489	0.489	0.4886	0.4885	0.4876	0.4871	0.4876	0.2861	0.2808	0.2922	0.2805	0.3018	0.286	0.2934	0.2849	0.3012	0.2771	0.3003	0.2803	0.3063	0.2882	0.3047	0.287
AFM	AUC	0.5476	0.5445	0.5402	0.5438	0.5278	0.5429	0.5227	0.5419	0.5379	0.5357	0.5315	0.5376	0.5242	0.535	0.5443	0.5367	0.6609	0.6508	0.6509	0.6504	0.6525	0.6534	0.6565	0.6517	0.6858	0.6683	0.6858	0.6727	0.6946	0.6741	0.6857	0.6731
	RMSE	0.4956	0.4971	0.4969	0.4968	0.4996	0.4964	0.4987	0.4966	0.4985	0.4979	0.4991	0.4979	0.5012	0.4978	0.4973	0.4975	0.2891	0.2844	0.2964	0.2833	0.3047	0.2893	0.2961	0.2882	0.3042	0.28	0.3035	0.2835	0.3091	0.2919	0.308	0.2903
PFA	AUC	0.6708	0.674	0.6756	0.6743	0.6467	0.6613	0.6842	0.6826	0.6346	0.633	0.6373	0.633	0.6191	0.6261	0.6144	0.6272	0.7457	0.7342	0.7346	0.7323	0.737	0.7329	0.7459	0.7309	0.7564	0.7325	0.7536	0.7374	0.7615	0.7359	0.7497	0.735
	RMSE	0.4805	0.4806	0.4805	0.4804	0.4873	0.4835	0.4842	0.4837	0.4908	0.4895	0.4905	0.4896	0.4925	0.4903	0.4924	0.49	0.2832	0.2796	0.2927	0.2796	0.3012	0.2858	0.2914	0.285	0.3006	0.2772	0.3006	0.2803	0.3046	0.2884	0.305	0.2878
DKT	AUC	0.5277	0.5483	0.5383	0.5452	0.5496	0.5569	0.5343	0.5588	0.5408	0.5717	0.5531	0.5652	0.5511	0.5801	0.5325	0.5828	0.6739	0.6175	0.6644	0.6824	0.6546	0.6601	0.6596	0.6684	0.6834	0.6729	0.6923	0.6446	0.7157	0.6878	0.6978	0.6884
	RMSE	0.5486	0.5378	0.5439	0.5402	0.5389	0.5351	0.5517	0.5347	0.5537	0.541	0.5522	0.5431	0.555	0.5285	0.557	0.5277	0.4265	0.4136	0.4417	0.4056	0.4306	0.411	0.422	0.4145	0.4799	0.3905	0.4704	0.4054	0.4604	0.4296	0.4603	0.4229

Figure 1. Results of different types of preprocessing (5 layers) on subject dataset.

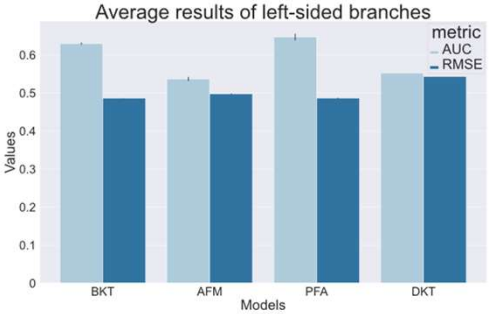


Figure 2.1. Models' performance on left-sided preprocessed data

The criteria that influences the most on the model's performance is how to consider partial answers.

In general, all the results from the left-sided branches of the tree are quite similar (fig. 2.1) and the same for the right-sided branches (fig. 2.2).

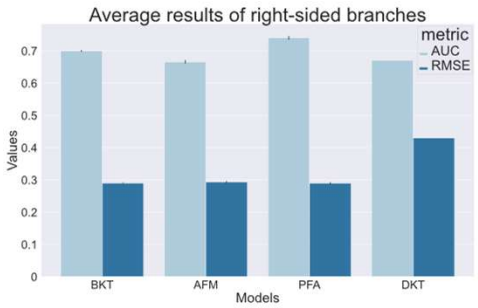
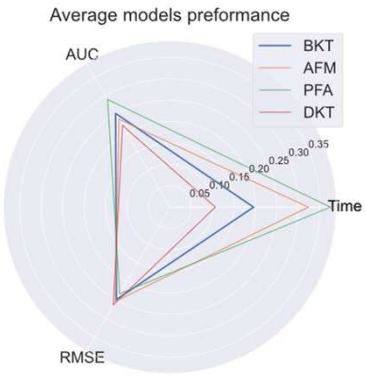


Figure 2.2. Models' performance on right-sided preprocessed data

4.2 RESULTS (Topics)



	BKT		AFM		PFA		DKT	
	AUC	RMSE	AUC	RMSE	AUC	RMSE	AUC	RMSE
least 15	0,596	0,495	0,564	0,502	0,746	0,455	0,447	0,550
least 18	0,657	0,482	0,602	0,496	0,762	0,445	0,555	0,516
least 20	0,653	0,478	0,593	0,495	0,774	0,439	0,457	0,512
least 22	0,688	0,466	0,655	0,475	0,796	0,426	0,646	0,492
least 25	0,688	0,467	0,668	0,475	0,791	0,428	0,646	0,484
most 15	0,652	0,480	0,606	0,489	0,720	0,464	0,593	0,499
most 18	0,667	0,475	0,634	0,483	0,744	0,455	0,608	0,495
most 20	0,646	0,481	0,603	0,490	0,731	0,461	0,584	0,503
most 22	0,666	0,477	0,622	0,485	0,749	0,453	0,605	0,497
most 25	0,673	0,474	0,634	0,483	0,751	0,452	0,600	0,500

Table 1. Results of experiments on most and least popular topics

5 CONCLUSION

Among all the experiments and two different datasets PFA model performed the best and BKT is on the second place almost every time. AFM and DKT are much less qualified for this data.

It should be mentioned that on datasets with more than 10000 rows AFM and PFA train for a much bigger amount of time than BKT and DKT. Therefore, there is always a tradeoff between time (BKT) and quality (PFA).

Overall, working with this dataset and exploring enough amount of data we would advise to use BKT model, while for small datasets or if the quality of the model is crucially important PFA should be used.

REFERENCES

[1] https://github.com/CAHLR/pyBKT

École polytechnique fédérale de Lausanne