Predictive Algorithms in Learning Analytics and their Fairness

Shirin Riazy[1] and Katharina Simbeck[2]

Abstract: Predictions in learning analytics are made to improve tailored educational interventions. However, it has been pointed out that machine learning algorithms might discriminate, depending on different measures of fairness. In this paper, we will demonstrate that predictive models, even given a satisfactory level of accuracy, perform differently across student subgroups, especially for different genders or for students with disabilities.

Keywords: Learning Analytics, Fairness, OULAD, At-Risk Prediction

1 Introduction

While the issue of fairness has been discussed in policing, operational systems and recommender systems, many machine learning (ML) areas, including learning analytics, still lack such contributions. In what ways can biases in data – possibly unjustifiably – influence algorithms? Since predictions in learning analytics are made to improve tailored educational interventions, the fitness of interventions requires the appropriateness of preceding predictions. On one hand, any inappropriate (discriminatory) intervention may further intensify discrimination [O'17]. On the other, a student requiring more involvement but not receiving it, might be neglected. Therefore, the validity of learning analytics algorithm must be analysed with regards to their predictive accuracy for minorities.

As of now, the definition of fairness in algorithms is highly ambiguous. Many definitions of fairness have been suggested and, as noted by [VR18], the choice of this definition should strongly depend on the context.

In this position paper, we would like to review some of the emerging definitions of fairness and transfer and evaluate their fit in the context of learning analytics, specifically the fairness of score predictions.

1.1 Definitions of Fairness

As mentioned briefly, a vast amount of different definitions of fairness have surfaced, serving different purposes and being designed to different use-cases. [VR18] have

[1] Hochschule für Technik und Wirtschaft, Berlin
[2] Hochschule für Technik und Wirtschaft, Berlin

doi:10.18420/delfi2019_305

gathered and explained 20 definitions of fairness, which are partially exclusive [Ch17]. We would like to briefly introduce selected definitions of fairness, which had the largest impact on the farness debate and research, measured by the number of citations. Also, since the research in this area is rapidly evolving, we would like to include one more fairness measure, which was only recently published, but is relevant as it was applied to a learning analytics scenario [GBB19].

The fairness measures rely on accuracy measures for predictive models. First, we would like to introduce some widely used model accuracy measures:

- We abbreviate the classical terms of binary classification in the usual way: true positive (TP), false positive (FP), true negative (TN), false negative (FN).
- The positive predictive value (PPV) is defined as TP / (TP + FP).
- The true positive rate (TPR) is defined as TP / (TP + FN).
- The area under the curve (AUC) of the receiver operator characteristic (ROC) graph, which shows the difference between predicted and real values for different choices in the trade-off between TP and FP rates.

Based on those measures, the model accuracy can be compared between different subgroups. Ideally, models should predict success equally well for subgroups, e.g. for female and male students: Predictive Parity (PP) demands equal PPV between groups [Ch17]. Equalised Odds (EO) demands equal TPR between groups [HPa16]. Slicing Analysis (SA) demands equal AUC between groups [GBB19]. In reality however, it cannot be expected that model accuracy is always consistent across all subsets of data. Therefore, a threshold needs to be defined, which delineates fair from unfair. For PP and EO, we set this threshold to be 0.05, as it was done in [VR18]. Since SA values mostly lie between 0.01 and 0.03 (see [GBB19]), we set the threshold to 0.02.

2 Related Work: Performance Prediction of Students and Fairness

[Si13] describes trend analysis and prediction as one of the primary applications of learning analytics. In order to build early warning systems and to identify risk, techniques of modelling users, relationship mining or knowledge domain modelling have been used [Si13]. Though ethical usage of user data was always called for (see for example [GD12, SP13]), this has a particular focus on privacy and data ownership [Si13].

The identification of at-risk students, their performance and attrition risk has been considered in literature in an attempt to find a basis for intervention to enhance student retention [AP12]. In order to build "early warning systems", data from virtual learning environments has been analysed to make predictions on learner behaviour [MD10]. A wide variety of procedures have been used to predict students at risk of failing or dropping out. These range from statistical methods, such as logistic regression [HZZ17] over naïve Bayes classification [Ku15] to predictions via neural networks [FY15].

3 Methodology

The data set that was used in the context of this paper is known as OULAD (Open University Learning Analytics Dataset). The open university (UK) has published this anonymised data set from their virtual learning environment (VLE) in 2017 [KHZ17] and it roughly consists of 32,000 samples, meaning student data from one course and one term. For students taking more than one course, these are counted as multiple samples.

Since we would like to reproduce realistic applications of ML algorithms in learning analytics, we tried to replicate publications, where this exact data set was used, such as in [HZZ17] and [Ku15]. Note, however, that we did not copy the exact method, since e.g. [Ku15] make a weekly prediction of at-risk students. In their paper, they split the features into demographic data (gender, age, prior education, disability) and virtual learning environment (VLE) data (number of assessments, days of access, number of clicks, number of logs). We used the following features to classify the outcome (fail, pass) of the course:

- Demographic: gender, prior education, age range, disability (boolean flag)
- VLE: number of assessments, number of different days on which material was accessed, average number of clicks, total number of logs
- Other: number of other credits studied in selected term, number of previous attempts

In order to avoid overfitting, we split the data set into a training, test and validation set (see e.g. [Mu12]). We chose to split the data according to the terms during which the modules took place. Out of these terms, two took place during the year 2013, which we chose as the training set (sample size: 13529). Two took place during the year 2014, which we chose as test set (sample size: 7804) and validation set (sample size: 11260). We split the terms in a way such that the resulting data sets would not vary much in sample size.

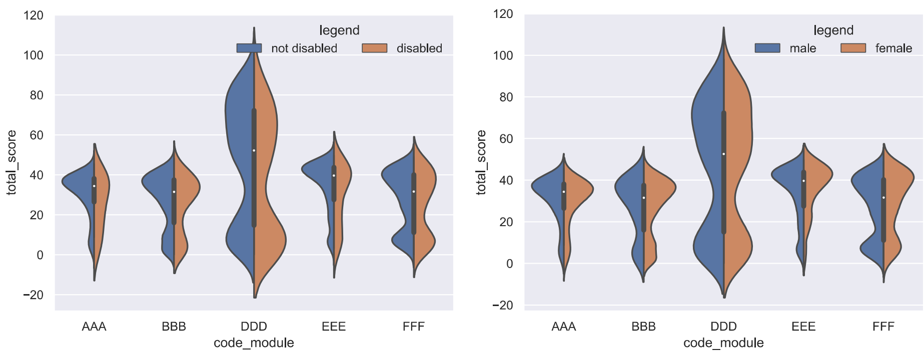

Figure 1: Violin plots showing the distributions of scores of non-disabled/disabled and female/male participants of the courses.

The distributions of the course scores (see Figure 1) show that, when grouping by gender, the distributions of the total scores of the different courses do not vary much. In contrast to this, the distributions greatly vary, when comparing disabled (sample size: 3164) and non-disabled students. In total, 47,2% of students in the dataset passed the course. The rate was higher for female students (48.4%) in comparison to male students (46.2%) and higher for students without disability (48.1%) in comparison to students with disability (38.1%).

3.1 Prediction of Course Outcome

Naïve Bayes (NB), k-Nearest-Neighbours (kNN) and Support Vector Machines (SVM) were used in [Ku15] to build a course-outcome prediction model for students. We replicated these algorithms and applied them to the OULAD data set. For more detailed descriptions of these algorithms, we would like to refer the reader to [Mu12]. Our results for the prediction systems are summarized in the following table, where the accuracy is defined as the relative sum of all correct classifications.

	NB	kNN	SVM
Test Set	0.728	0.760	0.776
Validation Set	0.756	0.751	0.793

Tab. 1: Accuracy of the course outcome using the previously defined test and validation set and the NB, kNN and SVM methods.

Note that the separation of the training, test and validation set was not replicated from [Ku15], since their weekly approach would have made an assessment of fairness overly complicated. The accuracy values are comparable to those of [Ku15]. All in all, the usage of the virtual learning environment had a high predictive value of the course outcome.

4 Results

All in all, we have implemented three machine learning algorithms for the prediction of the course outcome in real MOOC data and calculated their fairness measures: PP, EO and SA. In any of the three measures, a larger number means a higher predictive power and is thus advantageous. The bigger the difference between the subgroups for each of the measures, the more unfair could the model be considered. For example, in Table 2 we can see in the first row (predictive parity) that the difference between PPV values of students with and without disabilities is significant and that the PPV value of students with no disability (abbreviated as N) is higher.

Fairness Definition	NB		kNN		SVM	
Predictive Parity (PP)	0.032 (F)	**0.086** (N)	**0.084** (F)	**0.079** (N)	**0.085** (F)	**0.082** (N)
Equalised Odds (EO)	**0.372**	**0.210**	**0.074**	0.022	**0.064**	0.006

	(M)	(N)	(M)	(D)	(M)	(N)
Slicing Analysis (SA)	**0.044** (M)	0.020 (N)	0.018 (F)	0.002 (D)	**0.038** (F)	0.002 (D)

Tab. 2: Difference of PPV, TPV and AUC for the different prediction methods grouped by gender. In parenthesis, we denoted the group with the larger PPV, TPV, AUC values respectively by F/M for female/male and D/N for disability or no disability.

Most models show significant differences in accuracy between the sub-groups. Depending on the measure, model accuracy was sometimes higher for female and sometimes for male students. With one exception (EO for SVM), it was always better for students without disability. Interestingly, all models predicted below average pass-rates for female course participants, where they were higher in reality. NB and kNN models over-estimated pass-rate for disabled students, SVM strongly underestimated it. The reason for the differences in model fairness cannot be explained at this level of analysis, since the models work as black-box implementations.

5 Discussion

We have shown with the presented analysis on a real MOOC data set (OULAD) that, independently of the applied predictive model, predictive quality among learner sub-groups varies. The level of variation in difference of predictive quality between sub-groups was assessed with different fairness measures, with different, even contradictory results.

It should be noted that contradictory results mainly appear for the PP fairness measure, which might have to be calibrated differently (with a higher threshold for the 'unfair' category). Other than that, the EO measure quantified the NB method as highly discriminatory and the kNN and SVM methods only when grouping by gender. A reason for this could be the algorithms heavy reliance on relative frequencies in the original data.

When taking our exploratory analysis into account, the results of the SA method come closest to what we would expect, since no clear tendencies with respect to gender were visible. It should be noted that, when grouping for disability, the significant fairness measures consistently have higher PPV, TPV and AUC values for non-disabled students. This confirms our first expectation of possibly discriminatory data, which might lead to a "garbage in – garbage out" mechanism in the ML algorithms, especially using NB. Therefore, we can not fully conclude the validity of the fairness measures, but confirm that tendencies in the data were (at least in part) reproduced by the algorithms.

In order to get a more comprehensive view of the appropriateness of fairness measures in LA settings, further research and analyses on different datasets are required.

Bibliography

[AP12] Arnold, K. E.; Pistilli, M. D.: Course signals at Purdue: Using learning analytics to increase student success: Proceedings of the 2nd international conference on learning analytics and knowledge. ACM, 2012; S. 267–270.

[Ch17] Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. In Big data, 2017, 5; S. 153–163.

[FY15] Fei, M.; Yeung, D.-Y.: Temporal models for predicting student dropout in massive open online courses: 2015 IEEE International Conference on Data Mining Workshop (ICDMW). IEEE, 2015; S. 256–263.

[GBB19] Gardner, J.; Brooks, C.; Baker, R.: Evaluating the Fairness of Predictive Student Models Through Slicing Analysis, 2019.

[GD12] Greller, W.; Drachsler, H.: Translating learning into numbers: A generic framework for learning analytics, 2012.

[HPa16] Hardt, M.; Price, E.; and Srebro, N.: Equality of Opportunity in Supervised Learning. In (Lee, D. D. et al. Hrsg.): Advances in Neural Information Processing Systems 29. Curran Associates, Inc, 2016; S. 3315–3323.

[HZZ17] Hlosta, M.; Zdrahal, Z.; Zendulka, J.: Ouroboros: early identification of at-risk students without models based on legacy data: Proceedings of the Seventh International Learning Analytics & Knowledge Conference. ACM, 2017; S. 6–15.

[KHZ17] Kuzilek, J.; Hlosta, M.; Zdrahal, Z.: Open university learning analytics dataset. In Scientific data, 2017, 4; S. 170171.

[Ku15] Kuzilek, J. et al.: OU Analyse: analysing at-risk students at The Open University. In Learning Analytics Review, 2015; S. 1–16.

[MD10] Macfadyen, L. P.; Dawson, S.: Mining LMS data to develop an "early warning system" for educators. In Computers & education, 2010, 54; S. 588–599.

[Mu12] Murphy, K. P.: Machine learning: a probabilistic perspective. MIT press, 2012.

[O'17] O'Neil, C.: Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books, 2017.

[Si13] Siemens, G.: Learning analytics: The emergence of a discipline. In American Behavioral Scientist, 2013, 57; S. 1380–1400.

[SP13] Slade, S.; Prinsloo, P.: Learning Analytics. In American Behavioral Scientist, 2013, 57; S. 1510–1529.

[VR18] Verma, S.; Rubin, J.: Fairness definitions explained: 2018 IEEE/ACM International Workshop on Software Fairness (FairWare). IEEE, 2018; S. 1–7.

.