Learning Analytics and Fairness: Do Existing Algorithms Serve Everyone Equally?

Conference or Workshop Item

Learning Analytics and Fairness: Do Existing Algorithms Serve Everyone Equally?

Vaclav Bayer[0000−0001−8953−6335], Martin Hlosta[0000−0002−7053−7052], and Miriam Fernandez[0000−0001−5939−4321]

The Open University, UK
{vaclav.bayer,martin.hlosta,miriam.fernandez}@open.ac.uk

Abstract. Systemic inequalities still exist within Higher Education (HE). Reports from Universities UK show a 13% degree-awarding gap for Black, Asian and Minority Ethnic (BAME) students, with similar effects found when comparing students across other protected attributes, such as gender or disability. In this paper, we study whether existing prediction models to identify students at risk of failing (and hence providing early and adequate support to students) do work equally effectively for the majority vs minority groups. We also investigate whether disaggregating of data by protected attributes and building individual prediction models for each subgroup (e.g., a specific prediction model for females vs the one for males) could enhance model fairness. Our results, conducted over 35 067 students and evaluated over 32,538 students, show that existing prediction models do indeed seem to favour the majority group. As opposed to hypothesise, creating individual models does not help improving accuracy or fairness.

Keywords: Learning Analytics · Degree-awarding Gap · Fairness

1 Introduction

The latest statistics from UniversitiesUK and AdvanceHE [3, 4] show a 13% degree-awarding gap for BAME students in UK universities. Similar issues are found for female students in Science Technology Engineering and Maths (STEM) subjects. In terms of disability, 14.5% of undergraduate students in the UK declared that they had a disability in 2017. However, disabled students are less likely to obtain a degree-level qualification (21.8%) compared to non-disabled students [2]. Degree-awarding gaps in HE translate into socio-economic gaps and further inequality. Educated people are less dependent on public aid and are more resistant to economic downturns [1].

Learning Analytics (LA) have been widely applied in HE to improve the ways in which the learning processes are supported [11, 13]. We aim to study whether existing LA prediction models to identify students at risk of failing are fair in their predictions and serve majority and minority groups with the same effectiveness. We focused on assessing the prediction models currently used by The Open University [7]. These models are currently deployed in 530 courses and

are used by more than 1,500 teachers who receive weekly alerts of students at risk of failing, so that interventions to support them can be planned. We evaluate existing LA prediction models based on protected attribute (ethnicity, gender, disability) and study several new variations of the models to assess whether the proposed variations could enhance their fairness. We address this work by two Research Questions (RQs): **RQ1:** Do existing LA prediction models work equally effectively for all types of students? **RQ2:** Do LA population-specific prediction models, trained with data from particular protected groups, perform better than general LA prediction models trained on all students?

2 Methods

Learning Analytics Prediction Models: The LA prediction models generate predictions per each course and study week whether students will successfully submit upcoming assignment, i.e. a student will have more predictions in each course they are enrolled in. The models are built based on the Gradient Boosting Machine Learning Model [6] which has been selected as the best performing model [7]. Training is based on a combination of dynamic and static features. Static features are focused on socio-demographic data, such as gender, age, ethnicity, education, occupation, disability, index of multiple deprivation and country. Dynamic features capture the students' progress, as well as their weekly activity in a Virtual Learning Environment.

Data Selection: We selected data from the largest fourteen courses taught across all faculties to ensure a large and balanced sample of students across different disciplines. To train our models we used data from the selected courses for the 2018/19 academic year (35,067 unique students), and we tested with data from the 2019/20 academic year (32,538 unique students).[1] We used test data in the first 15 weeks with the latest prediction in Jan'20, therefore the data are not affected by the Covid-19 pandemic.

Experiment Design: We depart from a set of LA prediction models (one per course per week, trained in the same manner) called *Baseline*. To address our research questions we consider three protected attributes (ethnicity, gender and disability) and split student data into different subgroups based on these attributes: (i) *Black*, *Asian*, *White* and *Rest* for ethnicity, (ii) *Male* and *Female* for gender and, (iii) *Disabled* and *Non-Disabled* for disability. Note that *Rest* refers to an aggregated list of ethnic groups that are neither White, Black, or Asian and occurs only in RQ1 as a distribution of the training data is not suitable for computing a separate model for RQ2.

To address RQ1 we first compute the predictions for all students in the test set and assess the performance of those predictions for each of the above mentioned subgroups. Then we compare the performance of those predictions between the majority and the minority subgroups from the same protected attribute, e.g White vs Asian. To address RQ2 we compute population-specific prediction models that use only training data of each of the subgroups. The hypothesis is that specific models can learn the specific patterns of minority subgroups and

[1] More details about data samples at https://doi.org/10.6084/m9.figshare.14444567.v1

provide more accurate predictions [12]. The results of these population-specific models are then compared against *Baseline*.

Metrics: Following the work of [5], we have used three metrics to compute the models' performance: (i) Area Under the ROC Curve (AUC), (ii) False Positive rate $FPR = FP/(FP + TN)$ and (iii) False Negative Rate $FNR = FN/(TP + FN)$. AUC indicates the overall accuracy of the model. FPR, in our context, indicates those instances where the model predicts that the student will not submit (i.e., the student is at risk) but the prediction is false. In this case, the teacher may follow up on the student and provide her support while the support is not needed. FNR indicates those instances where the model predicts that the student will submit but the prediction is wrong. This is a much more problematic error since the teacher will not be alerted, and therefore, won't be able to provide support to the student when needed. For each RQ, we compute the significance test across chosen metrics using paired Wilcoxon signed-rank test. The selection was influenced by the Kruskal-Walis test [10] that indicated that the underlying data do not follow a normal distribution.

3 Results

RQ1: Fairness of the Existing Models. In terms of ethnicity, the *Baseline* model shows (see Table 1) the highest accuracy for White ethnicity across ethnic groups, without high disparity in AUC. FPR is significantly lower for White students than for all other groups. That means the model erroneously predicts with a higher frequency than students from Asian, Black and other Ethnic backgrounds will not submit their assignments. This is a less problematic error since students will still receive support. In terms of FNR, the model makes fewer errors for Black and Rest students than for White students. When looking at gender, the model is more accurate (AUC) for Male students than for Female students, with Female students having higher FNR, i.e. the more problematic error. In terms of disability, the model predicts 3% more accurately (AUC) for Non-Disabled than for Disabled students. The model also presents higher FPR for Disabled students. In summary, the *Baseline* model seems to consistently perform slightly better in terms of accuracy for White, Male, and Non-Disabled students. The model predicts most erroneously Black, Male, Disabled students that they will Not Submit an assignment (FPR) and that Asian, Female, Non-disabled students will Submit an assignment (FNR).

RQ2: Fairness Through Population-Specific Models. The comparison of corresponding protected subgroups between Table 1 and Table 2 reveals that the only individual model showing better performance in terms of AUC is for White students. All other models have lower AUC, and show a higher ratio of errors, particularly for FNR - the more problematic type of errors.

We also investigated **fairness by removing the protected attribute** from the *Baseline* model. The accuracy for ethnic minorities did not change much, but Asian students have significantly lower FNR and higher FPR; for Black, the trend is the opposite. For Females, FPR and FNR stayed nearly the same. For Males, removing the attribute worsened the FPR significantly but lowered the

FNR. For disabled students, the FNR significantly increased, while the FPR significantly decreased with the overall accuracy decreased. As such, we recommend removing the ethnicity attribute and keep gender and disability.

Table 1. Results of the Baseline model across subgroups. *p<0.1 **p<0.05 ***p<0.01

Protected attr.	Protected subgroup	AUC	FPR	FNR
Ethnicity	Asian	0.8588	0.0479***	0.5078***
	Black	0.8743	0.0721***	0.3912***
	Rest	0.8730**	0.0407***	0.4847
	White	0.8771	0.0287	0.5003
Gender	Female	0.8714	0.0303	0.5186
	Male	0.8880***	0.0340	0.4419***
Disability	NO	0.8816	0.0278	0.4967
	YES	0.8588***	0.0437***	0.4913***

Table 2. Results of individual models across subgroups.*p<0.1 **p<0.05 ***p<0.01

Protected attr.	Protected subgroup	AUC	FPR	FNR
Ethnicity	Asian	0.8287***	0.0423***	0.5725***
	Black	0.8413***	0.0916***	0.4290***
	White	0.8776	0.0303***	0.4948***
Gender	Female	0.8702***	0.0318***	0.5171
	Male	0.8814***	0.0335	0.4560***
Disability	NO	0.8802***	0.0284***	0.4991*
	YES	0.8472***	0.0449**	0.5063***

4 Discussion and Conclusions

This paper investigates whether existing LA prediction models serve everyone equally. This is important and timely research considering existing educational degree-awarding gaps and the impact that LA could have on either perpetuating or reducing these gaps. The results of our study show that existing prediction models seem to slightly favour the majority groups. Among the tested configurations, creating population-specific models harmed the accuracy and fairness of the predictions, which is in line with the results of [5] and [8]. The presented work can find its practical utility as a part of the evaluation process when existing models are being modified, e.g. by integrating new features. More research in terms of different adaptations and definitions of fairness [9] is needed to ensure that the technology we generate does not perpetuate existing educational gaps. It is also important to note that our research has been conducted over 14 largest courses, and on LA prediction models that aim to identify students at risk. More extensive research, e.g. increasing the number of courses, should be conducted to achieve a more general understanding of the problem. Qualitative research is also needed to complement these studies and assess how different fairness definitions affect the problem. While there are still many challenges to solve, this work constitutes an important step towards the understanding of LA algorithmic decision-making, its fairness and potential impact on minority groups.

References

1. Equity and quality in education: Supporting disadvantaged students and schools. OECD (2012)
2. Disability and education, uk: 2019 (2019), https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/disability/bulletins/disabilityandeducationuk/2019, accessed 07 Feb 2021
3. AdvanceHE: Degree attainment gaps (2017), https://www.advance-he.ac.uk/guidance/equality-diversity-and-inclusion/student-recruitment-retention-and-attainment/degree-attainment-gaps, accessed 08 Feb 2021
4. Amos, V.: Black, asian and minority ethnic student attainment at uk universities: closing the gap (2019), https://www.universitiesuk.ac.uk/policy-and-analysis/reports/Documents/2019/bame-student-attainment-uk-universities-closing-the-gap.pdf
5. Anderson, H., Boodhwani, A., Baker, R.S.: Assessing the fairness of graduation predictions. In: EDM 2019. pp. 488–491 (2019)
6. Greenwell, B., Boehmke, B., Cunningham, J., Developers, G., Greenwell, M.B.: Package 'gbm' (2020)
7. Hlosta, M., Zdrahal, Z., Bayer, V., Herodotou, C.: Why predictions of at-risk students are not 100% accurate? showing patterns in false positive and false negative predictions (2020)
8. Hutt, S., Gardner, M., Duckworth, A.L., D'Mello, S.K.: Evaluating fairness and generalizability in models predicting on-time graduation from college applications. International Educational Data Mining Society (2019)
9. Kizilcec, R.F., Lee, H.: Algorithmic fairness in education (2021)
10. Kruskal, W.H., Wallis, W.A.: Use of ranks in one-criterion variance analysis. Journal of the American statistical Association **47**(260), 583–621 (1952)
11. Leitner, P., Khalil, M., Ebner, M.: Learning analytics in higher education—a literature review. Learning analytics: Fundaments, applications, and trends pp. 1–23 (2017)
12. Perez, C.C.: Invisible women: Exposing data bias in a world designed for men. Random House (2019)
13. Viberg, O., Hatakka, M., Bälter, O., Mavroudi, A.: The current landscape of learning analytics in higher education. Computers in Human Behavior **89**, 98–110 (2018)