Impact of Predictive Learning Analytics on Course Awarding Gap of Disadvantaged students in STEM

Martin Hlosta[0000−0002−7053−7052], Christothea Herodotou[0000−0003−0980−1632], Vaclav Bayer[0000−0001−8953−6335], and Miriam Fernandez[0000−0001−5939−4321]

The Open University, UK
{martin.hlosta}@open.ac.uk

Abstract. In this work, we investigate the degree-awarding gap in distance higher education by studying the impact of a Predictive Learning Analytics system, when applying it to 3 STEM (Science, Technology, Engineering and Mathematics) courses with over 1,500 students. We focus on Black, Asian and Minority Ethnicity (BAME) students and students from areas with high deprivation, a proxy for low socio-economic status. Nineteen teachers used the system to obtain predictions of which students were at risk of failing and got in touch with them to support them (intervention group). The learning outcomes of these students were compared with students whose teachers did not use the system (comparison group). Our results show that students in the intervention group had 7% higher chances of passing the course, when controlling for other potential factors of success, with the actual pass rates being 64% vs 61%. When disaggregated: 1) BAME students had 10% higher pass rates (55 %vs 45%) than BAME students in the comparison group and 2) students from the most deprived areas had 4% higher pass rates (58% vs 54%) in the intervention group compared to the comparison group.

Keywords: Predictive Analytics · Course Awarding Gap · BAME · SES

1 Introduction

Historically, the performance of some demographic groups of students has been persistently worse than others. The impact of low socio-economic status (SES) on learning has increased over the last 50 years across countries, including the UK [3]. The attainment of ethnic minorities is consistently worse than White students. In the UK, in the past decade, 57% of Black students gained an upper second or first in their undergraduate degree, compared with 81% of White students [10]. There may be a significant overlap between Black, Asian and Minority Ethnic (BAME) students and low SES students. Recent post-pandemic statistics show that nearly half of BAME households (46%) live in poverty as opposed to 20% of White households [11].

Predictive Learning Analytics (PLA) focuses on forecasting the future students' outcomes using Machine Learning (ML) models and provide actionable

feedback to students or teachers, leading to improved student outcomes [6]. Growing evidence suggest that using PLA to trigger interventions leads to improved student outcomes in some studies [14,7] but not in others [2,5]. This suggests that further fine-grained analysis is needed to understand who of the students may benefit the most from PLA interventions. Previous studies reported the importance of a teacher in improving student outcomes and closing the attainment gap [13,4].

Research Questions To the best of authors' knowledge, there are no studies directly investigating the impact of PLA on different demographic subgroups. To fill this research gap, we examined the impact of PLA on the course awarding gap of BAME students and low SES students. We formulated two research questions (RQs): **RQ1:** What is the impact of PLA on student pass rates and their final score when deployed by teachers? **RQ2:** What is the impact of PLA when disaggregating the results by ethnicity and by SES?

2 Methods

Three STEM courses were selected based on their historically low retention and because they have not used Predictive technology before. Nineteen out of the 59 course teachers took part in the study (Intervention group). The remaining teachers (N=37) were treated as a Control group. Teachers were asked to log in before the first three assignments; 1, 2 and 3 weeks before the assignment's submission deadline. For each access, they were asked to consider contacting students that were identified as at-risk of 1) not submitting or 2) predicted as Fail or achieve low grade (50-60). Teachers were compensated to complete this research activity.

The predictions, generated weekly, estimate each student's likelihood to submit their next assignment and a likely banded score in the assignment. To generate these predictions the model utilises data from the previous run of the same course, i.e. 1) demographics, workload and prev. results, 2) student engagement in VLE, and 3) previous assignment performance. Gradient Boosting Machines (GBM) has been selected in the previous years as the best performing model [8].

Evaluation For each RQ, we focus on students completion, passing and their overall score. Completion means that a student satisfied the course requirements and sat the exam; passing means that they were successful in the exam. Logistic regression models were applied for binary outcomes (completion and pass) and linear regression was used for the overall score. The unit of analysis were students ($N = 1,412$). The factors entered into the regression analysis included: (1) **Student** (age, gender, an indicator of linked qualification, declared disability, caring responsibility, new/continuing, highest previous education, avg. previous score, no. of other credits studied, no. of previous attempts of the course, IMD[1] and whether the student is identified as BAME), (2) **Teacher** (no. of students the teacher is responsible for, avg. student pass rate in the previous years they

[1] In the UK, the SES gap can be expressed as a difference between students from low and high deprived areas, measured by Index of Multiple Deprivation (IMD)[12,9].

have been teaching), (3) **Course** - dummy encoded as variables Course 1, 2 and 3.

Similarly, as [12], IMD was discretised into quintiles - Q1 representing the most deprived areas and Q5 the least deprived areas. The check for homogeneity of variances, multicollinearity and normality were conducted to ensure no assumption violation. Except for the number of students in the teachers' group, the continuous variables did not follow a normal distribution and were discretised. IMD, previous student score, and teacher previous pass rates contained missing values, and we encoded them as a special category. The previous score was discretised for each course separately.

To answer RQ2, we created separate regression models for each demographic group - i.e. for BAME/non-BAME and each IMD quintile $Q1 - Q5$. BAME students encompassed 57 Asian, 46 Black, 39 Mixed and 18 Minor Ethnicity students (11% of all students). This was conducted again for completion, pass and overall score. For each regression model, we investigated the coefficient indicating any differences between the Intervention and the Control group.

3 Results

The accuracy of the model for predicting completion was $Acc = 0.71$ and for pass $Acc = 0.69$, for the continuous target overall score $R^2 = 0.22$. Table 1 shows the coefficients of the regression for pass, completion and score, with their statistical significance and standard errors for all students, regardless of the demographic group.[2] The results show that students in the Intervention group (factor group_INT) were much more likely to pass the module ($\beta = 0.36, p < 0.01$) and also obtain higher overall score ($\beta = 5.07, p < 0.01$). The positive coefficient for completion was however not statistically significant ($\beta = 0.11, p >= 0.1$). This might suggest that students in the Intervention group were better prepared to be successful in the exam. The pass rate beta $\beta = 0.36$ can be converted to an Average Marginal Effect 0.07, which means that keeping all attributes constant, students in the Intervention group have 7% higher chances of passing the course and obtaining 5.07 more points in the overall score.

Disaggregation by BAME and IMD Overall, the pass rates (61% vs 65%) and overall score (44 vs 46.5) were higher in the Intervention group. The positive differences were higher for BAME students for passing (52% in the Control vs 62% in the Intervention) and lower IMD quintiles, IMD1-3. Regression models were created only for the specific demographic group, controlling for potential confounding variables. Fig. 1 shows the β regression coefficients for the Intervention group extracted from these models. Except for completion, the lower SES groups have higher coefficients, with statistical significant results measured for IMD Q1, $\beta = 0.78, p < 0.05$. For BAME, the most significant factor related to passing the course was teachers' previous low pass rates $\beta = -3.08, p < 0.05$. This factor was not present for non-BAME students. This suggests that teachers

[2] The results only include attributes where at least one of the factors had $p < 0.05$. The full analysis can be found at https://ordo.open.ac.uk/account/articles/14414774

	Completion		Pass		Overall Score	
	β	SE	β	SE	β	SE
prev_sc_VERY_HIGH	0.41	0.75	1.33**	0.69	23.22**	10.08
disability	−0.34*	0.28	−0.44**	0.26	−8.12**	3.83
is_new	0.29	0.73	0.74*	0.67	13.66**	10.00
course_2	0.26	0.37	0.46**	0.35	2.19	4.88
group_INT	0.11	0.29	0.36**	0.27	5.07**	3.82
credits_other_[1-60]	−0.67**	0.34	−0.70**	0.30	−11.80**	4.14
credits_other_>=61	−1.09**	0.44	−1.04**	0.41	−17.12**	5.79
stud_in_group	−0.01*	0.01	−0.02**	0.01	−0.23**	0.16

*p<0.05 **p<0.01 ***p<0.001

Table 1. Regression table for completion, pass and overall score

who had students with consistently low pass rates in the past are more likely to have lower pass rates for BAME students but not for non-BAME students. The same attribute was significant also for the second most deprived areas IMD_Q2 $\beta = -1.78, p < 0.01$. Overall, a great concentration of BAME has been observed in low SES and conditions of poverty [1,11], suggesting that any intervention that tackles students in low SES would be particularly beneficial for BAME students alongside other ethnicities found in low SES.

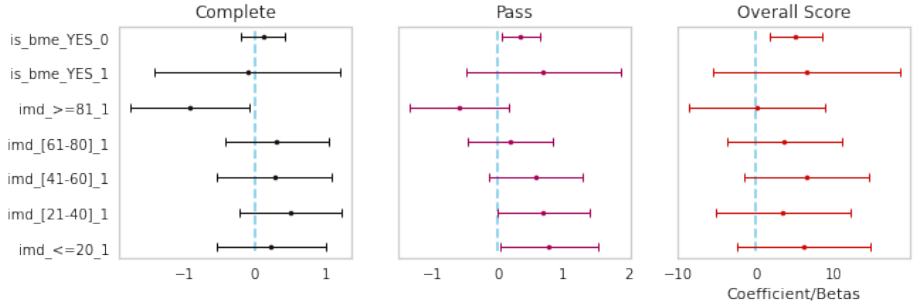

Fig. 1. Outcomes Beta coefficients for being in the Intervention group

4 Conclusions

The results demonstrated a positive impact on students' performance, particularly those who were coming from low SES, as measured by the Index of Multiple Deprivation (IMD). This suggests that students found in rather disadvantaged contexts such as poverty are more likely to benefit from PLA systems. BAME are shown to have the greatest representation in low SES (32% as opposed to 10% non-BAME students), stressing the significance of early PLA support for BAME students in particular. Because our study was conducted only on 3 STEM courses and less than 1,500 students, the scaled experiment should try to replicate the study across more courses, examine separately specific student groups within BAME such as Black or Asian students and investigate the context, i.e. whether some conditions need to be met to observe the same or similar effect.

References

1. American Psychology Association and others: Ethnic and racial minorities & socioeconomic status (2016), https://www.apa.org/pi/ses/resources/publications/minorities
2. Borrella, I., Caballero-Caballero, S., Ponce-Cueto, E.: Predict and intervene: Addressing the dropout problem in a mooc-based program. In: Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale. pp. 1–9 (2019)
3. Chmielewski, A.K.: The global increase in the socioeconomic achievement gap, 1964 to 2015. American Sociological Review **84**(3), 517–544 (2019)
4. Crenna-Jennings, W.: Key drivers of the disadvantage gap: Literature review (2018)
5. Dawson, S., Jovanovic, J., Gašević, D., Pardo, A.: From prediction to impact: Evaluation of a learning analytics retention program. In: Proceedings of the seventh international learning analytics & knowledge conference. pp. 474–478 (2017)
6. Herodotou, C., Hlosta, M., Boroowa, A., Rienties, B., Zdrahal, Z., Mangafa, C.: Empowering online teachers through predictive learning analytics. British Journal of Educational Technology **50**(6), 3064–3079 (2019). https://doi.org/https://doi.org/10.1111/bjet.12853
7. Herodotou, C., Naydenova, G., Boroowa, A., Gilmour, A., Rienties, B.: How can predictive learning analytics and motivational interventions increase student retention and enhance administrative support in distance education? Journal of Learning Analytics **7**(2), 72–83 (Sep 2020). https://doi.org/10.18608/jla.2020.72.4
8. 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. In: Proceedings of the 10th International Conference on Learning Analytics and Knowledge (LAK20) (2020)
9. Richardson, J.T., Mittelmeier, J., Rienties, B.: The role of gender, social class and ethnicity in participation and academic attainment in uk higher education: an update. Oxford Review of Education **46**(3), 346–362 (2020)
10. Roberts, N., Bolton, P.: Educational outcomes of black pupils and students - research briefing (oct 2020), https://commonslibrary.parliament.uk/research-briefings/cbp-9023/
11. Stroud, P.: Measuring poverty 2020, a report of the social metrics commission (jul 2020), https://socialmetricscommission.org.uk/wp-content/uploads/2020/06/Measuring-Poverty-2020-Web.pdf, [Online; accessed 08-February-2021]
12. Thiele, T., Pope, D., Singleton, A., Stanistreet, D.: Role of students' context in predicting academic performance at a medical school: a retrospective cohort study. BMJ open **6**(3) (2016)
13. Warschauer, M., Matuchniak, T.: New technology and digital worlds: Analyzing evidence of equity in access, use, and outcomes. Review of research in education **34**(1), 179–225 (2010)
14. Wong, B.T.M., Li, K.C.: Learning analytics intervention: A review of case studies. In: 2018 International Symposium on Educational Technology (ISET). pp. 178–182 (2018). https://doi.org/10.1109/ISET.2018.00047