Contexts Matter but How? Course-Level Correlates of Performance and Fairness Shift in Predictive Model Transfer

Zhen Xu*
zx2393@tc.columbia.edu
Teachers College, Columbia University
USA

Joseph Olson*
jwo2108@tc.columbia.edu
Teachers College, Columbia University
USA

Nicole Pochinki
np2824@tc.columbia.edu
Teachers College, Columbia University
USA

Zhijian Zheng
zhijiz8@uci.edu
University of California, Irvine
USA

Renzhe Yu
renzheyu@tc.columbia.edu
Teachers College, Columbia University
USA

ABSTRACT

Learning analytics research has highlighted that contexts matter for predictive models, but little research has explicated *how* contexts matter for models' utility. Such insights are critical for real-world applications where predictive models are frequently deployed across instructional and institutional contexts. Building upon administrative records and behavioral traces from 37,089 students across 1,493 courses, we provide a comprehensive evaluation of performance and fairness shifts of predictive models when transferred across different course contexts. We specifically quantify how differences in various contextual factors moderate model portability. Our findings indicate an average decline in model performance and inconsistent directions in fairness shifts, without a direct trade-off, when models are transferred across different courses within the same institution. Among the course-to-course contextual differences we examined, differences in admin features account for the largest portion of both performance and fairness loss. Differences in student composition can simultaneously amplify drops in performance and fairness while differences in learning design have a greater impact on performance degradation. Given these complexities, our results highlight the importance of considering multiple dimensions of course contexts and evaluating fairness shifts in addition to performance loss when conducting transfer learning of predictive models in education.

KEYWORDS

Predictive Analytics; Transfer Learning; Algorithmic Fairness; Intersectionality; Learning Management System; Higher Education

*Equal contribution

LAK '24, March 18–22, 2024, Kyoto, Japan

ACM ISBN 979-8-4007-1618-8/24/03…$15.00
https://doi.org/10.1145/3636555.3636936

ACM Reference Format:
Zhen Xu, Joseph Olson, Nicole Pochinki, Zhijian Zheng, and Renzhe Yu. 2024. Contexts Matter but How? Course-Level Correlates of Performance and Fairness Shift in Predictive Model Transfer. In *The 14th Learning Analytics and Knowledge Conference (LAK '24), March 18–22, 2024, Kyoto, Japan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3636555.3636936

1 INTRODUCTION

Student success prediction is one of the most investigated topics in learning analytics research. Researchers have used cutting-edge machine learning algorithms and various data sources to forecast the chances of success measured by grades, retention, graduation, etc. [1, 36]. In practice, predictive analytics have also widely been embedded in EdTech products to facilitate teaching and learning, student advising, and other day-to-day educational activities [28]. As machine learning methods and innovative applications evolve, research on both technical and social aspects of predictive analytics remains critical in education.

A key challenge in building predictive analytics is model generalization. Both theoretical and empirical research have suggested that there is no "one size fits all" solution for predicting student success across different contexts [7, 15, 21]. For example, a model that accurately predicts student grades in a large physics lecture class may perform poorly in an English seminar. Therefore, it has been recommended to build and apply predictive models for similar content areas, instructional design, and student populations to preserve model performance. On the other hand, predictive analytic products are typically deployed at the organizational level, such as a school district or a college system, where models need to work across instructional and institutional contexts. Therefore, models are often used in contexts separate from those they were trained on, a scenario known as transfer learning in computing research communities. Additionally, cross-context application can positively contribute to educational equity because institutions, educators, and students in under-resourced settings might still benefit from using predictive models developed elsewhere, compared to having no analytics at all [14]. As such, it is not sufficient to acknowledge that contexts matter and that predictive models are difficult to transfer. Instead, we need to understand how contexts shape the utility and portability of those models in order to identify the limits of model transfer and make informed decisions about model deployment.

However, in contrast to the substantial volume of predictive modeling research in the learning analytics community, there is still limited research that provides quantitative insights into the role of contextual factors on models' portability.

Another important consideration when building predictive models for student success is algorithmic fairness. Models can generate predictions that are biased against disadvantaged student groups if, for example, they are trained on data that reflects historical patterns of discrimination or use predictors that are correlated with protected attributes [23]. These biased predictions can eventually perpetuate existing inequities when used for educational decision-making. Therefore, it is important to understand the sources and patterns of algorithmic bias and develop strategies to improve the fairness of predictive models. Over the past few years, there has been a steady increase of research on algorithmic fairness in education[5, 14, 23], but most research examines the fairness/bias of models in their original contexts of development, not so much the *shift* in fairness/bias when models are transferred to new contexts. Given the importance of applying models across contexts in practice, examining the portability of fairness is also crucial. Less portable models might not only produce inaccurate predictions but also create additional harm through augmented algorithmic bias.

In this study, we systematically examine the portability of student success prediction models in terms of both overall performance and algorithmic fairness. More importantly, we quantify the relationship between various contextual factors and model portability. We focus on the task of predicting course performance with digital trace data from Learning Management Systems (LMS). Across the globe, LMS has become a standard part of technical infrastructure in higher education and gathers a wealth of data on students' learning activities and outcomes across different instructional contexts. Such data holds the potential for educators to better understand learning dynamics and inform targeted interventions and therefore has been used to build predictive analytics by LMS vendors or institutions that adopt the system [29, 31]. Because learning behavior is largely shaped by instructional conditions and this influence can vary across student demographics [3], predictive models based on LMS data naturally face the aforementioned challenge of portability. In this context, we examine course-to-course model transfer and attempt to answer two research questions:

RQ1: How do performance and fairness shift in course-course predictive model transfer? Are there trade-offs between performance and fairness shifts?

RQ2: How do contextual differences contribute to performance and fairness shifts in course-to-course predictive model transfer?

Our contribution to existing research and practice is threefold. First, we move from "contexts matter" to "how contexts matter" by providing one of the first and largest analyses of the relationship between contextual differences and portability of predictive models in education. Second, we advance research on algorithmic fairness by examining fairness *shift* in transfer learning contexts. We also take a holistic intersectional approach, reflecting the conjunction of multiple social identities along which individual students may experience unfair algorithmic treatment. Third, we empirically evaluate one of the most common prediction tasks in learning analytics with data available at almost every institution. Thus, our findings directly inform real-world applications, and our analyses can easily scale up across broader institutional contexts.

2 RELATED WORK

2.1 Portability of Learning Analytics Models

Model portability has been identified as one of the main challenges in learning analytics [4]. While a large volume of literature underscores the criticality of context similarity in ensuring the utility of predictive models, only a select number of studies have rigorously explored and quantified the extent to which model performance generalizes across diverse contexts. Among these studies, some assessed the consistency of given predictors of student performance across various courses. For instance, [15] noted variances in predictive power of behavioral indicators across nine courses from multiple disciplines, with explained variances ranging from 2% to 70.3%. [7] conducted a study with a larger sample ($N = 17$) and found that no single set of predictors works consistently well in predicting academic performance across all courses. [21] assessed the portability across 15 courses and found that only 3% of performance variability is explained at the course level, while a substantial 68% is attributed to individual students. Other studies built prediction model(s) in one context and evaluated the performance degradation when testing the same model(s) in another [4]. For example, [30] analyzed prediction model generalizability across 16 courses and found that the incorporation of ontological information improved generalizability compared to similar models using low-level behavioral data alone. Beyond courses, [2] considered within-cohort generalizations, emphasizing the role of similar student enrollments in model generalizability.

Across existing research that examined the portability of prediction models, researchers identified several influential contextual factors, such as variations in course subject and content, differences in learning designs and LMS usage, and the similarity of enrolled students. However, these are mostly qualitative findings based on a small number of courses in specific disciplines, which can potentially be biased and inconclusive. Also, most research analyzed only one or a few contextual factors at one time and cannot holistically contrast the effects of various contextual disparities.

2.2 Algorithmic Fairness in Education

In education, the concept of fairness was widely studied in research on opportunity and outcome disparities long before the spread of digital learning [23]. As more educational stakeholders have turned to algorithmic systems for data-driven insights into student experience and performance, the issue of algorithmic fairness and bias has gained broader attention within the learning analytics community[5, 10, 14].

To measure algorithmic fairness, existing research typically takes two perspectives: group fairness (GF) and individual fairness (IF) [23]. Group fairness emphasizes statistical or predictive parity across demographic groups, with metrics such as statistical parity (SP) [11], equalized odds (EO) [18], and the Absolute Between-ROC Area (ABROCA) [13]. Individual fairness is based on the idea that similar individuals should be treated similarly. Example metrics include

consistency [46] and counterfactual fairness [25]. In practice, selecting which fairness metric to adopt requires careful evaluation of the intended use of the algorithm(s) [23].

Empirical research on algorithmic fairness and bias has been surging in recent years [5, 23]. Researchers have not only attempted to measure algorithmic bias in various educational tasks but also experimented with different technical strategies to mitigate the bias. For example, one study [10] assessed the portability and fairness of machine learning models built with LMS data. Using both preprocessing and post-processing bias mitigation techniques, they found that the presence of fairness constraints may increase fairness while maintaining predictive capabilities. While most research investigates model fairness and bias on a given test sample, limited effort has examined fairness in transfer learning scenarios, which are common in educational applications. Another study closely related to ours [14] examined the transfer of college dropout prediction models across institutional contexts and revealed that a zero-shot transfer approach can match local model performance without sacrificing fairness. To our knowledge, there has been no research that explicitly investigates contextual factors of fairness in transfer learning in education.

2.3 Intersectionality in Machine Learning

Intersectionality is a conceptual framework first proposed by [8], pointing out that black women experience further discrimination than either Black people or Women. This concept underscores how the confluence of social identities jointly shapes individual experiences rather than independently. At the macro-level, this intersectional mechanism contributes to the complexity in imbalanced power structures and social inequalities. As fairness concerns arise around machine learning applications in human society, critics have highlighted the limitation of evaluating bias with respect to a small number of social identities in a siloed manner. Even when a machine learning model generates fair outputs when considering a single protected attribute, it may disadvantage groups at the intersection of multiple marginalized identities. This puts forth a call for incorporating intersectionality perspectives in machine learning research.

In the context of education, there has been some recent work on intersectional fairness, focused on developing metrics to assess intersectional fairness in algorithms [14, 40]. Some metrics compare each intersectional group against the overall sample on measures of model performance, such as Statistical Parity (SP) Subgroup Fairness and Equal Opportunity (EO) Subgroup Fairness [22]. Other metrics focus more on the gaps between various subgroups, like the gap between the highest and lowest subgroup AUCs [26]. Another example is the 80% rule (four-fifths rule) [39] which determines unfair impact by assessing the ratio of favorable outcomes between a disadvantaged group and the best-performing group, marking a significant disparity if this ratio falls below 0.8. Despite the growing interest and initial research efforts, the research of intersectional fairness in predictive models, especially in transfer learning, is still in its early stages. The existing studies also emphasize the need for a more in-depth investigation into how current algorithms influence the consistency of fairness across different intersectional groups [5, 14].

3 MATERIAL AND METHODS

We examine the portability of models that leverage students' behavioral traces in a course to predict their course performance. We train models from a single course and test them on a different course. We characterize predictive performance and algorithmic fairness on each model and use multiple regressions to predict performance and fairness shifts (from local to transfer courses) using course-level contextual differences.

3.1 Dataset

We analyze undergraduate courses taught at a large public university in the United States over six quarterly terms between Fall 2021 and Spring 2023, excluding summer terms. The data comes from two systems: Canvas LMS, which provides students' behavioral traces and learning design information, and Student Information System, which provides administrative records such as students' background information and course performance. The raw data covers 39,789 unique students across 9,444 unique courses (675,755 student-course outcomes)[1].

3.2 Predictive Models for Academic Performance

3.2.1 Outcome and Predictors. We formulate a binary classification problem that predicts each student's course-level performance using their behavioral features in the corresponding LMS course space, which is one of the most common learning analytics tasks. The outcome variable is whether a student achieves proficiency (A- or above) or not [16]. We focus on proficiency because the distribution of original letter grades is highly skewed toward the higher end due to the selectivity of this institution and grade inflation in recent years. Under this construction, we remove courses that do not use letter grades.

Behavioral features are calculated based on students' behavioral traces within the first half (five weeks) of the course, in line with the real-world application scenario of creating early alerts for instructors [44]. We keep courses with at least 50 students and a proficiency rate between 20% and 80%, where human prediction of at-risk students is comparatively harder and automated predictive analytics have a larger marginal value. We also remove courses that do not actively use Canvas LMS by keeping those with at most 50% missing value for learning design variables (see details in the next subsection). The final dataset includes 37,089 unique students across 1,493 unique courses (225,044 student-course outcomes).

The 16 behavioral features at the student-course level (Table 1) are all continuous variables with non-negative values. To handle outliers, we normalize the variables by dividing by the 95$^{\text{th}}$ percentile, and then apply the hyperbolic tangent function $tanh(x)$, resulting in values between 0 and 1. As $tanh(x)$ is approximately linear between $x = 0$ and $x = 1$, this transformation does not change the shape of the bulk of the variable's distribution; however, the outliers are "reigned in" to be no larger than 1 (noting that $tanh(1) = 0.76$ represents the 95$^{\text{th}}$ percentile). Missing values are filled with 0, which carry no predictive power due to zero variation.

[1]When the same course is delivered by different instructors or offered in different years, we consider them different unique courses due to potential variations in instructional design and student composition, which are the contextual factors we aim to test.

3.2.2 Models. For our binary classification task, we use two well-established and representative algorithms: logistic regression with elastic net regularization and boosted classification trees, chosen for their prevalent use in predictive modeling literature. Logistic regression is selected for interpretability, while tree-based algorithms are chosen because they tend to have good performance. Models are implemented in MATLAB R2022a, using the functions `lassoglm` and `fitcensemble`. The logistic regression models have two hyper-parameters: λ, which controls the strength of regularization, and α, which toggles between lasso and ridge regularization. The boosted classification tree models have three hyper-parameters: the number of trees, the learning rate, and the maximum number of decision splits.

To construct a baseline for model portability, we first build and evaluate *local* predictive models, i.e., models trained and tested on the same course. We fit models to each of the 1,493 courses using stratified 9-fold cross-validation to optimize the hyper-parameters on 90% of the data and evaluate local performance on a held-out sample of 10% of the data. To account for class imbalance, we apply Synthetic Minority Over-Sampling Technique (SMOTE) after the 9-fold split using $k = 2$ nearest minority-class neighbors. We repeat this process for each 90/10 split to generate predictions for all the students in the course. Lastly, to make the predictions more robust, we repeat this whole process 5 times, with new 90/10 splits generated with each repetition, and average the predicted probabilities for each student to get a final predicted probability.

We evaluate model *transfer* performance by testing a locally trained model on every course that occurs in later academic terms, in line with how models would be built and deployed in real-world contexts. Therefore, courses from Spring 2023 are used only for testing and not for training models. When testing on transfer courses, we mimic the process for evaluating local performance as much as possible. We train a model on a random 90% subset of the local course (to match the 90/10 split in local training) using the average values of the optimal hyper-parameters (averaged across the local 90/10 splits and 5 repeats). We then test this model on each of the transfer courses. We repeat this process 5 times (trained on a new random 90% subset) and average the predicted probabilities for each student-course outcome.

Because we are running binary classifications, the models' predicted probabilities need a threshold to generate the final predictions for individual records. We determine the thresholds by the Youden Index (J) [43] using the ROC of the training (local) courses only.

3.3 Key Metrics

3.3.1 Performance and Fairness. To evaluate the predictive performance of the models, we use *Area Under the ROC Curve (AUC)* and *Balanced Accuracy* which are more robust to imbalanced data issues.

To measure algorithmic fairness, we use *Absolute Between ROC Area (ABROCA)* [13], *Equalized Odds (EO)* [33], and *Pseudo* R^2. ABROCA is computed by the absolute value of the area between the two groups, $A = 0$ and $A = 1$, of their ROC curves as follows:

$$ABROCA_{(A)} = \int_0^1 |ROC_{A=0}(t) - ROC_{A=1}(t)|\, dt$$

EO measures the equality of the true positive rate (TPR) and true negative rate (TNR) between two groups, $A = 0$ and $A = 1$. We average the EO TPR and EO TNR for an overall EO score.

$$EO_{(A)} = \frac{|TPR_{(A=0)} - TPR_{(A=1)}| + |TNR_{(A=0)} - TNR_{(A=1)}|}{2}$$

ABROCA and EO can only measure fairness for two comparative groups at a time (e.g., those defined by a binary attribute). We focus on fairness regarding race and gender, which are the most commonly investigated demographic attributes in algorithmic fairness research [14]. Specifically, we examine underrepresented racial minorities (URM) vs. non-URM and female vs. male, respectively.

In response to the need to integrate intersectionality into algorithmic fairness research [17], we develop a novel measure of fairness by running a logistic regression that predicts the correctness of individual predictions from the foregoing predictive model with a series of demographic variables and their interactions ($|Y - \hat{Y}| \sim race + gender + race * gender$). Assuming that unequal *prediction correctness* across demographic groups indicates decreased fairness, we compute Pseudo R^2 of the logistic regression to measure intersectional fairness, with a smaller value indicating increased fairness. There are many proposed Pseudo R^2 metrics for logistic regression and we choose Tjur's R^2 [38], also known as *the coefficient of discrimination*, which is simply the difference between the means of class distributions:

$$Pseudo\ R^2 = \hat{\pi}_1 - \hat{\pi}_0$$

where $\hat{\pi}_1$ and $\hat{\pi}_0$ denote the averages of fitted values for successes and failures, respectively. The values of all the fairness measures we use range from 0 to 1, and the ideal fairness is 0. However, we often report $1 - fairness$ throughout the paper so that fairness aligns with performance in that values of 1 are considered better than 0. This measure allows for examining the intersection of multiple protected attributes in diverse functional forms.

3.3.2 Portability. We quantify both performance and fairness shifts using the percentage difference from the local model as follows:

$$\text{Shift} = \frac{\text{Local - Transfer}}{\text{Local}} \times 100$$

3.3.3 Contextual Differences. To characterize the context differences between training and testing courses, we construct four groups of contextual factors from LMS and administrative data: *Subject Matter*, *Admin Features*, *Learning Design* and *Student Composition*. Given the heterogeneity of these factors, which include both binary and continuous variables of varied scales, we use different methods to compute pairwise differences between local and transfer courses so that each computed contextual difference ranges from -1 to 1 or 0 to 1. Detailed descriptions of these factors and different computation methods are presented in Table 2.

4 RESULTS

4.1 Patterns of Performance Shift, Fairness Shift, and Trade-offs

4.1.1 Predictive Performance and Performance Shift. The local and transfer performance distributions are shown in Fig. 1. The mean and standard deviation for the AUC assessed on the training (local) course is 0.65 ± 0.075 for LR and 0.63 ± 0.080 for GBT. The Balanced

Category	Variable	References
General	Number of sessions	[7, 15, 24, 45]
	Number of actions	[7, 45]
	Number of views on syllabus	[20]
	Number of views on grade summary	[20, 45]
	Time online	[7, 19, 45]
	Average session duration	[7, 19]
Learning content	Number of actions associated with learning content	[41, 42]
	Time spent viewing learning content	[24]
	Number of actions associated with wikis	[7, 45]
Assessment	Number of actions associated with assignments	[41, 42]
	Time spent on assignments	[42]
	Number of assignments submitted	[45]
	Number of actions associated with quizzes	[42]
	Number of quizzes submitted	[35]
Communication	Number of actions associated with discussions	[6, 42]
	Time spent on discussions	[24, 27]

Table 1: Behavioral features. These variables are used in logistic regression and gradient-boosted trees to predict student outcomes. All variables are scaled between 0 and 1.

Contextual Factor (Each Course)	Contextual Difference (Pairwise)	References
Subject Matter		
Course	1 if only with diff course names	[7, 15, 29]
Department	1 if only with diff course names and depts	[15, 29]
School	1 if only with diff course names, depts, and schools	[21, 29]
Discipline	1 if with diff course names, depts, schools and disciplines	[12, 15, 29]
Admin Features		
Instructor	1 if with diff instructors	[12]
Online (-1,0,1)	$local - transfer$ (online=1, in-person=0)	
Upper level (-1,0,1)	$local - transfer$ (upper=1, lower=0)	
Class size	$local - transfer$ (normalized)	[21]
Term	number of terms apart (scaled)	
Learning Design		
Number of modules	$local - transfer$ (normalized)	[34]
Number of files	$local - transfer$ (normalized)	[7, 15, 29]
Number of items per module	$local - transfer$ (normalized)	
Number of announcements	$local - transfer$ (normalized)	[9, 37]
Number of files before course begins	$local - transfer$ (normalized)	
Number of visible navigational items	$local - transfer$ (normalized)	
Number of assignments	$local - transfer$ (normalized)	[7, 15, 29]
Number of quizzes	$local - transfer$ (normalized)	[7, 15, 29]
Number of discussion threads	$local - transfer$ (normalized)	[7, 15, 29]
Student Composition		
Students enrolled	$1 - 2 \times \frac{\#overlap}{\#local + \#transfer}$	[32]
% URM	$local - transfer$ (normalized)	
% Female	$local - transfer$ (normalized)	

Table 2: Course-level contextual factors used to explain performance and fairness shifts. Subject Matter variables are binary with 1 indicating the highest level at which the local and transfer courses differ. Term differences are scaled between 0 and 1. The normalization used for continuous variables is $\frac{local - transfer}{local + transfer}$ so that values ranged between -1 and 1.

Accuracy is 0.65 ± 0.056 for LR and 0.65 ± 0.059 for GBT. When these models are assessed on the 1,225 other (transfer) courses, there is a clear performance drop with an AUC of 0.59 ± 0.090 for LR and 0.57 ± 0.082 for GBT. The Balanced Accuracy on the transfer courses is 0.55 ± 0.061 for LR and 0.54 ± 0.060 for GBT. The similarity in the overall shapes of these two distributions does not guarantee that the models are consistent at the level of individual course performance. To check this consistency, we use Pearson correlation between the LR and GBT models assessed on the local courses only, as incorporating transfer courses may break the independence

assumption. The LR and GBT models have correlation coefficients of $R = 0.71$ for AUC and $R = 0.69$ for Balance Accuracy. While and GBT are strongly correlated, their medians are significantly different as determined by a Wilcoxon signed-rank test for both AUC ($p < 2.2 \times 10^{-16}$) and Balanced Accuracy ($p < 2.2 \times 10^{-16}$).

The performance shift distributions are shown in Fig. 2. Among all local-transfer course pairs and across both LR and GBT, over 70% experience a loss in performance: specifically, 72% for AUC and 86% for Balanced Accuracy. The mean and standard deviation of AUC shift are 8.0% ± 16.2% for LR and 9.2% ± 15.8% for GBT. The Balanced Accuracy shift is 15.1% ± 11.2% for LR and 15.6% ± 11.1% for LR. Performance shift is fairly consistent between LR and GBT models, with a correlation coefficient of $R = 0.55$ for AUC shift and $R = 0.48$ for Balanced Accuracy shift, as well as simultaneous performance increase or loss in 67% of the course pairs.

4.1.2 Algorithmic Fairness and Fairness Shift. The local and transfer fairness distributions are shown in Fig. 3 for several different measures. The mean and standard deviation for the ABROCA assessed on the training (local) course are 0.14 ± 0.086 (race), 0.12 ± 0.070 (gender) for LR and 0.14 ± 0.083 (race), 0.12 ± 0.075 (gender) for GBT. The EO are 0.13 ± 0.084 (race), 0.12 ± 0.069 (gender) for LR and 0.13 ± 0.084 (race), 0.12 ± 0.075 (gender) for GBT. And lastly, Pseudo R^2 are 0.12 ± 0.069 for LR and 0.12 ± 0.075 for GBT. When evaluating these models on the 1,225 transfer courses, we do not observe significant changes in the mean and variance of those aforementioned fairness metrics. The mean and standard deviation for the ABROCA assessed on the training (transfer) course are 0.13 ± 0.085 (race), 0.12 ± 0.071 (gender) for LR, and 0.14 ± 0.085 (race), 0.12 ± 0.073 (gender) for GBT. The EO are 0.11 ± 0.082 (race), 0.11 ± 0.072 (gender) for LR and 0.12 ± 0.084 (race), 0.11 ± 0.077 (gender) for GBT. And lastly, Pseudo R^2 are 0.04 ± 0.036 for LR and 0.04 ± 0.034 for GBT.

In all distributions, our measures vary between 0 and 1, with 0 being completely unfair and 1 being completely fair. The models tend to be less fair regarding race (URM/Non-URM) compared to gender (female/male). However, there are relatively minor differences between LR and GBT distributions and between local and transfer distributions. The medians (25^{th}, 75^{th} percentile) of the 1-ABROCA measure for LR models assessed on local courses are 0.886 (0.830, 0.922) for race and 0.896 (0.845, 0.928) for gender. Race and gender are significantly different ($p = 7.441 \times 10^{-5}$) using a Wilcoxon signed-rank test. Similarly, the 1-EO measures for LR local courses are 0.889 (0.831, 0.935) for race and 0.892 (0.840, 0.936), and are significantly different ($p = 1.256 \times 10^{-3}$). We measure the fairness of race, gender, and their interaction simultaneously using 1-Pseudo R^2, which has a median (25^{th}, 75^{th} percentile) value of 0.979 (0.955, 0.989) for LR models assessed on the local course.

The fairness of the LR and GBT models assessed on the local courses are correlated with values of $R = 0.47$ for ABROCA race, $R = 0.35$ for EO race, $R = 0.43$ for ABROCA gender, $R = 0.32$ for EO gender, and $R = 0.30$ for Pseudo R^2. Furthermore, the median values between the LR and GBT models are not significantly different as determined by a Wilcoxon signed rank test. The p-values are $p = 0.693$ for ABROCA race, $p = 0.723$ for EO race, $p = 0.550$ for ABROCA gender, $p = 0.812$ for EO gender, $p = 0.050$ for Pseudo R^2.

The fairness shift distributions are shown in Fig. 4. Unlike performance shift, fairness shift does not show a global average drop. The percent of local-transfer course pairs that have a racial fairness drop averaged across LR and GBT models, is 48% and 39% as measured by 1-ABROCA and 1-EO, respectively. The percent of course pairs that have a gender fairness drop is 49% and 41% as measured by the same metrics. And 55% of the course pairs have a fairness drop as measured by 1-Psuedo R^2. The mean and standard deviation of fairness shift for the LR model is −1.60% ± 16.7% for 1-ABROCA race, −2.77% ± 15.2% for 1-EO race, −0.90% ± 12.4% for 1-ABROCA gender, −1.58% ± 12.3% for 1-EO gender, and 0.657% ± 5.08% for 1-Pseudo R^2. Fairness shift is fairly consistent between LR and GBT models, with correlation coefficients of $R = 0.55$ for ABROCA race, $R = 0.40$ for EO race, $R = 0.43$ for ABROCA gender, $R = 0.35$ for EO gender, and $R = 0.37$ for Pseudo R^2.

4.1.3 Performance-Fairness Trade-Off. We examine the potential trade-off between performance and fairness in both the local and transfer courses (Fig 5). We also examine the relationship between performance shift and fairness shift. We focus on the LR models, but the performance-fairness relationships are very similar to the GBT models. We also focus on Balanced Accuracy as our measure of performance, and 1-Pseudo R^2 as our measure of fairness. Both these measures are computed using threshold-ed predictions, and 1-Pseudo R^2 is a more comprehensive measure of fairness than ABROCA and EO since it combines race and gender.

No evident linear relationship is observed between the performance and fairness metrics in the LR models assessed on either the local courses or transfer courses. The Pearson correlation coefficient between Balanced Accuracy and 1-Pseudo R^2 is $R = 0.005$ for local courses and $R = -0.05$ for transfer courses. Similarly, no linear relationship is observed between performance and fairness shifts, with Balanced Accuracy and 1-Pseudo R^2 having a correlation coefficient of $R = 0.01$. Although there is a U-shaped joint distribution between Balanced Accuracy (Transfer) and 1-Pseudo R^2 (Transfer), it can be explained by the joint distribution of two independent variables such as a normal distribution (Balanced Accuracy in Figure 1) and a beta distribution (1-Pseudo R^2 in Figure 3). This suggests that the U-shape results from the tail of the 1-Pseudo R^2 distribution being sampled more frequently in the middle of the Balanced Accuracy distribution where there are more observations.

4.2 Contextual Differences and Performance/Fairness Shift

We investigate the relationship between contextual factors and model portability by regressing performance shift and fairness shift on differences in contextual variables between the testing (transfer) course and the training (local) course. Given the similar findings in both LR and GBT, we illustrate our findings with LR as the example in this section. Before regressing, we check the correlation between each pair of contextual factors to account for potential co-linearities. This results in the removal of the *number of items per module* variable from the regression. Contextual variables in our data set are categorized into four groups: Subject Matter, Admin Features, Learning Design, and Student Composition (see Tables 2). We perform regression on each group of variables separately to gain an understanding of the effect size across different aspects of contextual factors. We report adjusted R^2 to account for the

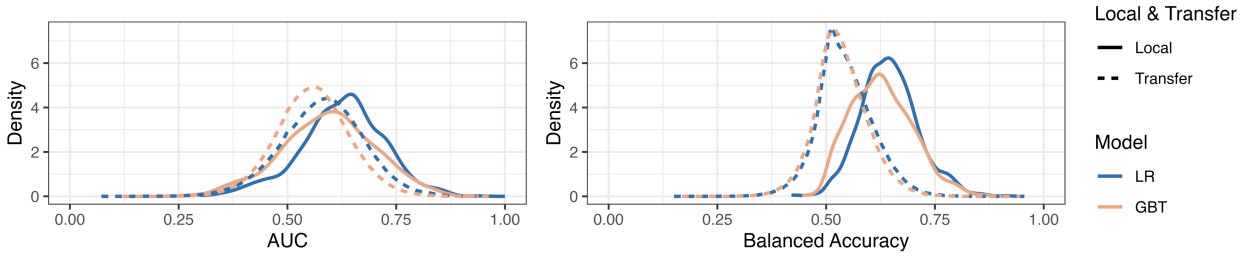

Figure 1: Distribution of predictive performance. Performance of predictive models evaluated on the training course (local) and other courses (transfer). Models include boosted trees (GBT) and logistic regression (LR). Measures of performance include AUC (left) and Balanced Accuracy (right). Histograms are normalized to have an area of 1 (density).

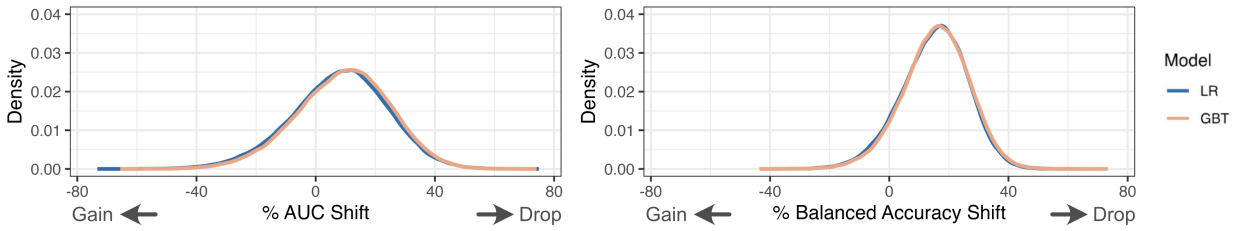

Figure 2: Distribution of performance shift. Percent change in AUC (left) and Balanced Accuracy (right) from local to transfer models, including boosted trees (GBT) and logistic regression (LR). Histograms are normalized to have an area of 1 (density).

groups having differing numbers of variables. Table 3 shows the adjusted R^2 for each of these regressions using several measures of performance and fairness shift as the outcome variables.

Among the four groups of contextual variables, differences in Admin features between local and transfer courses consistently elicited the largest adjusted R^2 values across all metrics of performance shift and fairness shift. For performance shift, the group with the second largest R^2 values was Learning Design, using either AUC or Balanced Accuracy. The Student Composition variables have the third largest R^2. However, for the fairness shift, these results are reversed. Student Composition variables tend to have the second largest R^2 values while Learning Design has the third largest, at least the ABROCA and EO measures of fairness. Student Composition variables have a noticeably stronger relationship with the fairness shift of race (URM/Non-URM) than on gender (female/male). When using the Pseudo R^2 measure of fairness, Learning Design again has the second largest R^2, with Student Composition having the third largest. Across all metrics of performance shift and fairness shift, the Subject Matter variables have negligible R^2 values.

We integrate all contextual variables into a single regression model and compute adjusted R^2 values (Table 3). The full model's R^2 values are approximately the sum of the R^2 values when using each group separately (always at least 80%). The full model accounts for a modest percent of the variance in fairness shift ranging from $R^2 = 0.090$ to $R^2 = 0.139$ across our different metrics. However, the full model can only account for a minor percentage of variance in performance shift with $R^2 = 0.029$ for the AUC shift and $R^2 = 0.036$ for the Balanced Accuracy shift. Figure 6 shows coefficient estimates and 95% confidence intervals from the full regression model for every contextual variable. Variables are considered significant predictors if their confidence intervals do not intersect zero.

In the *Subject Matter* category, coefficient estimates are trending positive when predicting performance shift but are largely insignificant when predicting fairness shift. Here, positive coefficients mean that when the training and testing courses are different (e.g., different departments), there is likely a drop in model performance (i.e., increased performance shift). The training and testing course being a different course of being from different disciplines tends to be a better predictor of performance shift than them being from different departments or schools.

In the *Admin Features* category, differences between various training and testing are related to performance shift and fairness shift in different ways. Class size is found to be a strong predictor of both performance shift and fairness shift but with opposite valence. A positive coefficient means that when the training course is larger than the testing course, the portability decreases (i.e., increased shift). This is the case for performance shifts. However, the opposite is true for the fairness shift. A similar phenomenon occurs regarding the course level variable. Here, a positive coefficient means that training on an upper-level course and testing on a lower-level course correlates with decreased portability (increased shift). This is the case for fairness shifts, but the opposite is true for performance shifts.

For the online variable, a positive coefficient means the training on an online course and testing on an in-person course correlates

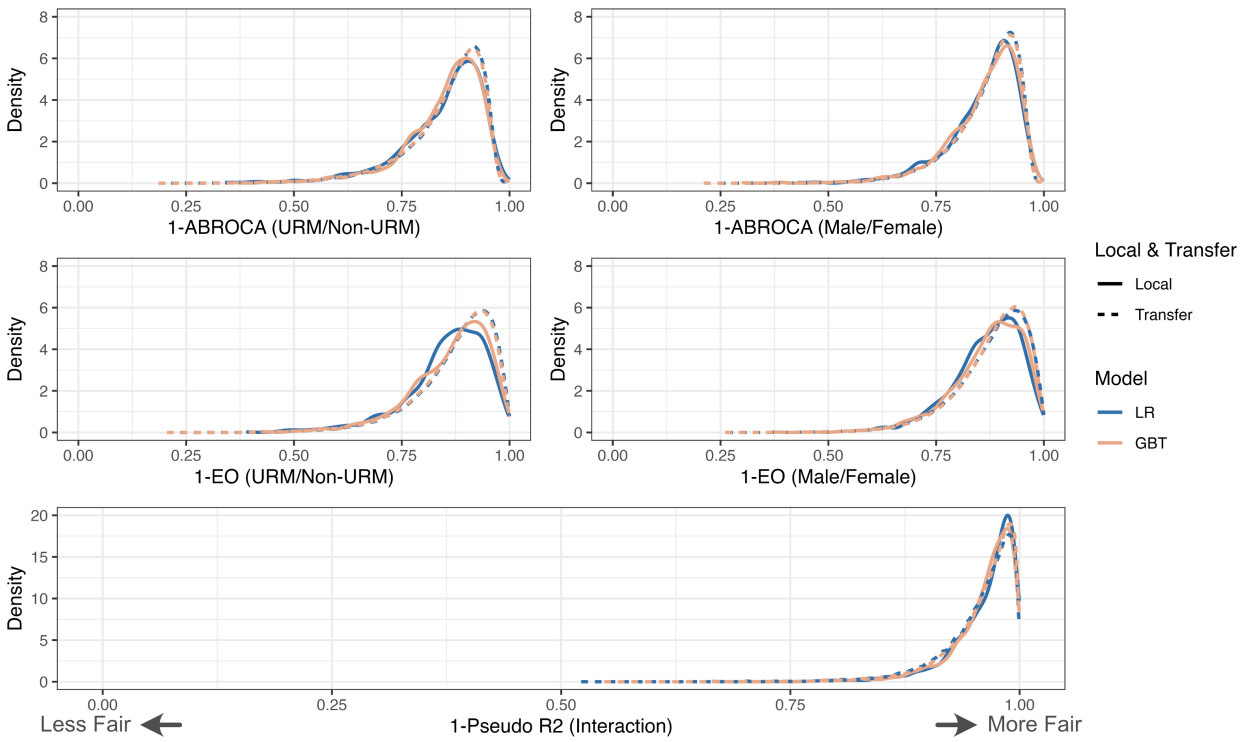

Figure 3: Distributions of algorithmic fairness. Fairness of predictive models evaluated on the training course (local) and on other courses (transfer). Models include boosted trees (GBT) and logistic regression (LR). Measures of fairness include 1− Absolute Between-ROC Area (ABROCA, top row), 1−Equalized Odds (EO, middle row), and 1−Pseudo R^2 (bottom row). ABROCA and EO are computed separately for race (left) and gender (right), while Pseudo R^2 is Tjur's R^2 from (logistic) regressing individual predictive correctness on race, gender, and their interactions simultaneously. Histograms are normalized to have an area of 1 (density).

	Performance Shift		Fairness Shift				
Difference in ...	Δ AUC	Δ Balanced Accuracy	Δ ABROCA (URM)	Δ ABROCA (Gender)	Δ EO (URM)	Δ EO (Gender)	Δ Pseudo R^2
Subject Matter	0.001	0.001	0.000	0.000	0.001	0.000	0.000
Admin Features	**0.016**	**0.023**	**0.071**	**0.131**	**0.082**	**0.083**	**0.085**
Learning Design	**0.011**	**0.009**	0.015	0.011	0.014	0.003	**0.010**
Student Composition	0.008	0.004	**0.042**	**0.018**	**0.038**	**0.009**	0.002
All of the above	0.029	0.036	0.121	0.139	0.128	0.090	0.099

Table 3: Explanatory power of contextual differences on performance and fairness shifts. Each cell shows adjusted R^2 values from regressing performance/fairness shift (column header) on contextual differences (row header, in Table 2).

with decreased portability (increased shift). This is the case for performance shift and, to a lesser extent, fairness shift as well. A similar pattern is observed for the term gap, for which longer time intervals between training and testing courses correlated with both larger performance shifts and fairness shifts. Lastly, training and testing on courses taught by different instructors seem to correlate mildly with performance shift but not fairness shift.

In the *Learning Design* category, all features significantly correlate with performance shifts, and most of them significantly correlate with fairness shifts. However, the coefficients when predicting fairness shift seem inconsistently positive or negative across the various metrics of fairness. Their signs are more consistent when predicting performance shifts across AUC and Balanced Accuracy. But across the variables, the signs are mixed. For all these variables, a positive coefficient means that when the training course has more of X than the testing course, the portability decreases (increased shift). The variables with positive coefficients for performance shift include the number of files, the number of announcements, the number of files uploaded before the course starts, and the number

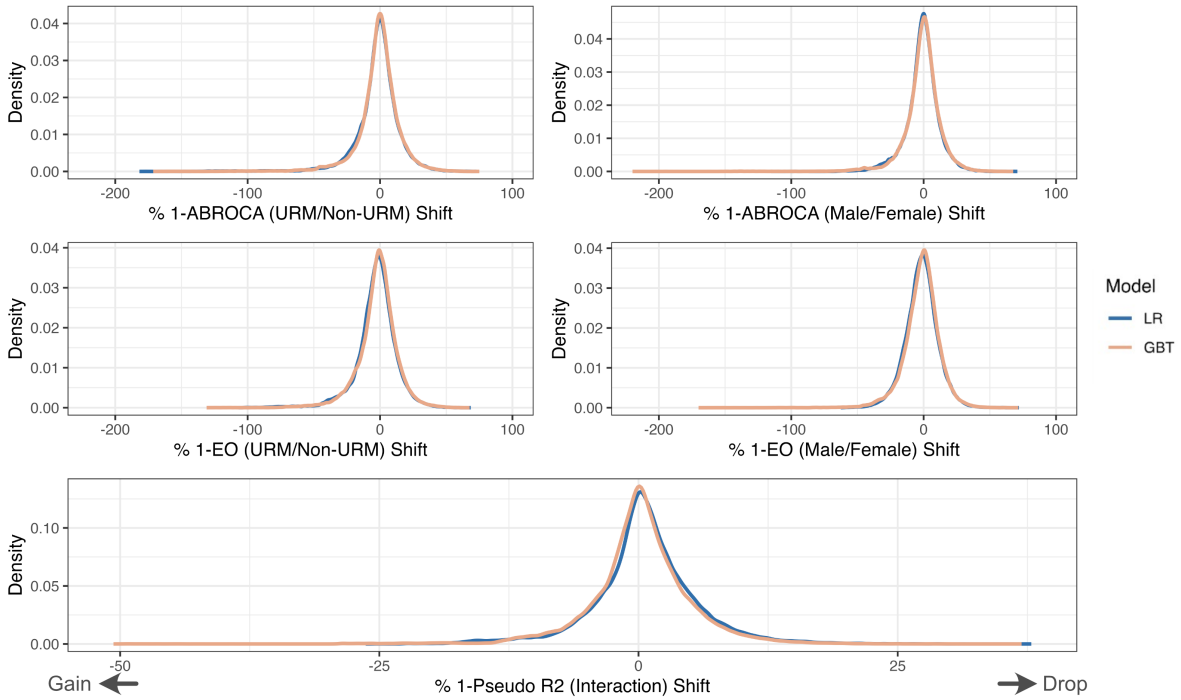

Figure 4: Distribution of fairness shift. Percent change in 1−ABROCA (top row), 1−Equalized Odds (EO, middle row), and 1−Pseudo R^2 (bottom row) from local to transfer models. Models include boosted trees (GBT) and logistic regression (LR). Histograms are normalized to have an area of 1 (density).

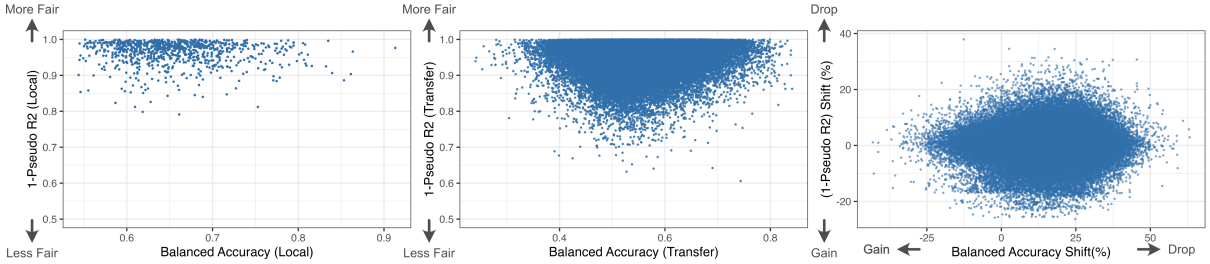

Figure 5: Performance-fairness trade-off. Performance (Balanced Accuracy) vs. fairness (1−Pseudo R^2) of LR models evaluated on local courses (left) and transfer courses (middle). Performance shift vs. fairness shift from local to transfer models (right).

of discussion threads. The variables with negative coefficients include the number of modules, the number of visible navigational items, the number of assignments, and the number of quizzes.

In the *Student Composition* category, differences in demographic composition significantly correlate with both performance and fairness shift. Differences in the percentage of URM students are a strong predictor of performance shifts and URM fairness shifts but not gender or Pseudo R^2 fairness shifts. Its positive coefficient means that when the training course has a higher percentage of URM students than the testing course, there is decreased portability (increased shift). Similarly, when the training course has a higher percentage of female students than the testing course, performance shifts and fairness shifts have decreased portability (increased shift), with the exception of Pseudo R^2 shift. The percent of overlap, the percent of students in both the training and testing course, is weakly significant with performance shift and does not significantly correlate with fairness shift.

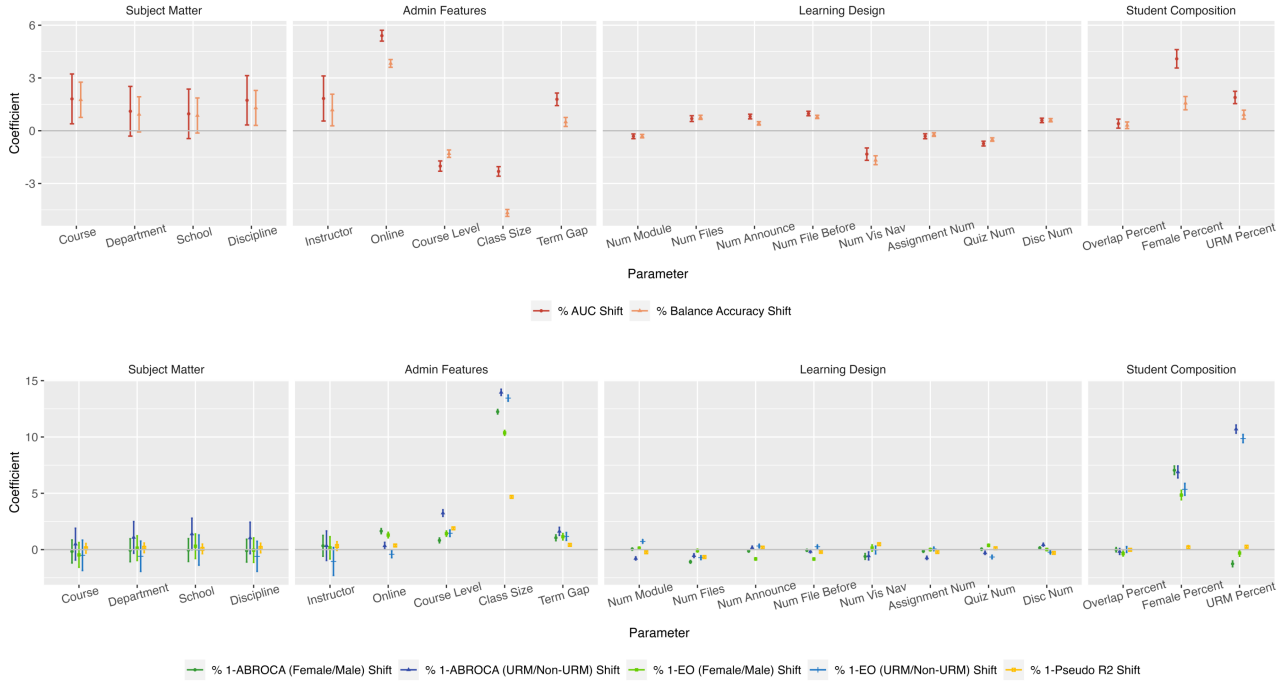

Figure 6: Estimated effects of contextual differences on performance and fairness shifts. Regression coefficients of differences in each contextual variable (column) for performance shift (top) and fairness shift (bottom) from local to transfer models. Error bars represent 95% confidence intervals.

5 DISCUSSION AND CONCLUSION

We present one of the first and largest empirical studies of the performance and fairness portability of predictive models in education. We analyze how contextual differences between courses moderate performance and fairness shifts when models are developed in one course (local) and applied to another (transfer). Our results suggest that model transfer produces clear performance drops on average across different course pairs (Figure 2), whereas fairness shifts have less consistent directions (Figure 4) yet are more predictable from contextual differences (Table 3). We also show that performance and fairness shifts co-vary differently along dimensions of course-level contextual features. Not only are some groups of contextual features better predictors of one type of shift more than the other (such as Learning Design and Student Composition), but some specific features positively correlate with one type of shift while negatively correlating with the other (such as differences in class size) (Figure 6). However, at the aggregate level, there is no clear trade-off between performance and fairness (Figure 5).

These findings have important implications for both researchers and practitioners. Researchers should not overlook the portability of fairness when studying distribution shifts. While prior research has largely centered on performance in the context of prediction portability, our findings emphasize the need to consider fairness, as data distribution shifts appear to have some disparate effects on performance and fairness. In stark contrast to the consistent pattern of performance decline, the shift in fairness centers around zero when transitioning from a model trained on one course to another. One possible reason for this in our study is that we did not control for demographic imbalance in our courses, either training or testing. This could have introduced noisy estimates for fairness shift. This indicates the complexities and nuances involved in fairness shifts, which can be modulated by diverse factors when not explicitly adjusting for them. More in-depth research about the portability of fairness is needed to explain these behaviors better. Our study also suggests practical implications for practitioners in this field, emphasizing the importance of accounting for various dimensions of course contexts during the transfer of predictive models. Aligning learning designs between training and testing sets could be more beneficial for sustaining model performance than ensuring subject matter consistency. Furthermore, ensuring similarity in admin features and student composition across these sets may help preserve both performance and fairness, as disparities in these contextual factors significantly impact both aspects. However, given the varied and sometimes even contrasting effects of specific contextual differences on performance and fairness, we have yet to pinpoint an optimal strategy that effectively balances both aspects. This highlights the need for more detailed research in this field to uncover how to strike a balance between preserving performance and fairness in transfer learning.

Despite our large sample size and diverse feature set, there are some limitations within our current work. Firstly, we only include courses with at least 50 students as training and testing set in our analysis, which we did for two reasons. First, from a practical standpoint, we are training and testing models within each course, so we

needed sufficient samples to ensure the quality of model training. Second, from an application aspect, academic prediction models are most useful in large courses where instructors may not adequately understand each student well. Another limitation is that our data set is from only one institution, which could contain hidden bias. As such, we cannot guarantee that the regression estimations from our analysis can be generalized well to other institutions, especially those with largely different demographic compositions. Given these limitations, there are several lines of future research. One would be expanding our current work to other LMS platforms besides Canvas and replicating the results at other institutions to not only test the robustness of our current findings but also to provide deeper insights into how platform-specific contextual nuances influence predictive models. Regardless, future studies should investigate strategies to enhance both performance and fairness, either by incorporating contextual information into the prediction process or by devising course clustering strategies informed by the insights from our study.

REFERENCES

[1] Eyman Alyahyan and Dilek Düştegör. 2020. Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education* 17 (2020), 1–21.

[2] Jose Luis Arroyo-Barrigüete, Susana Carabias-López, Tomas Curto-González, and Adolfo Hernández. 2021. Portability of Predictive Academic Performance Models: An Empirical Sensitivity Analysis. *Mathematics* 9, 8 (2021), 870.

[3] Jean A. Baker, Teresa P. Clark, Kimberly S. Maier, and Steve Viger. 2008. The differential influence of instructional context on the academic engagement of students with behavior problems. *Teaching and Teacher Education* 24, 7 (2008), 1876–1883. https://doi.org/10.1016/j.tate.2008.02.019

[4] Ryan S. Baker. 2019. Challenges for the Future of Educational Data Mining: The Baker Learning Analytics Prizes. (6 2019).

[5] Ryan S Baker and Aaron Hawn. 2021. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education* (2021), 1–41.

[6] Rebeca Cerezo, Miguel Sánchez-Santillán, M. Puerto Paule-Ruiz, and J. Carlos Núñez. 2016. Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. *Computers & Education* 96 (5 2016), 42–54. https://doi.org/10.1016/J.COMPEDU.2016.02.006

[7] Rianne Conijn, Chris Snijders, Ad Kleingeld, and Uwe Matzat. 2016. Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Transactions on Learning Technologies* 10, 1 (2016), 17–29.

[8] Kimberlé Crenshaw. 2013. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist legal theories*. Routledge, 23–51.

[9] Shane Dawson, Erica Mcwilliam, and Jennifer Pei-Ling Tan. 2008. Teaching smarter: how mining ICT data can inform and improve learning and teaching practice. https://api.semanticscholar.org/CorpusID:61176352

[10] Oscar Blessed Deho, Srecko Joksimovic, Lin Liu, Jiuyong Li, Chen Zhan, and Jixue Liu. 2023. Assessing the Fairness of Course Success Prediction Models in the Face of (Un) equal Demographic Group Distribution. In *Proceedings of the Tenth ACM Conference on Learning@ Scale*. 48–58.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[12] Nafsaniath Fathema and Mohammad H Akanda. 2020. Effects of instructors' academic disciplines and prior experience with learning management systems: A study about the use of Canvas. *Australasian Journal of Educational Technology* 36, 4 (Jan. 2020), 113–125. https://doi.org/10.14742/ajet.5660

[13] Josh Gardner, Christopher Brooks, and Ryan Baker. 2019. Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th international conference on learning analytics & knowledge*. 225–234.

[14] Joshua Gardner, Renzhe Yu, Quan Nguyen, Christopher Brooks, and Rene Kizilcec. 2023. Cross-Institutional Transfer Learning for Educational Models: Implications for Model Performance, Fairness, and Equity. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1664–1684.

[15] Dragan Gašević, Shane Dawson, Tim Rogers, and Danijela Gasevic. 2016. Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *Internet and Higher Education* 28 (1 2016), 68–84.

[16] Niki Gitinabard, Yiqiao Xu, Sarah Heckman, Tiffany Barnes, and Collin F Lynch. 2019. How widely can prediction models be generalized? Performance prediction in blended courses. *IEEE Transactions on Learning Technologies* 12, 2 (2019), 184–197.

[17] Usman Gohar and Lu Cheng. 2023. A Survey on Intersectional Fairness in Machine Learning: Notions, Mitigation, and Challenges. *arXiv preprint arXiv:2305.06969* (2023).

[18] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[19] Joon Heo, Hyoungjoon Lim, Sung Bum Yun, Sungha Ju, Sangyoon Park, Rebekah Lee, J Heo, H Lim, S B Yun, and S Ju. 2019. Descriptive and Predictive Modeling of Student Achievement, Satisfaction, and Mental Health for Data-Driven Smart Connected Campus Life Service. (2019).

[20] Jelena Jovanovic, Dragan Gasevic, Shane Dawson, Abelardo Pardo, and Negin Mirriahi. 2017. Learning analytics to unveil learning strategies in a flipped classroom. *Internet and Higher Education* 33 (2017), 74–85.

[21] Jelena Jovanović, Mohammed Saqr, Srećko Joksimović, and Dragan Gašević. 2021. Students matter the most in learning analytics: The effects of internal and instructional conditions in predicting academic success. *Computers & Education* 172 (2021), 104251.

[22] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*. PMLR, 2564–2572.

[23] René F Kizilcec and Hansol Lee. 2022. Algorithmic fairness in education. In *The ethics of artificial intelligence in education*. Routledge, 174–202.

[24] Vitomir Kovanović, Dragan Gašević, Srećko Joksimović, Marek Hatala, and Olusola Adesope. 2015. Analytics of communities of inquiry: Effects of learning technology use on cognitive presence in asynchronous online discussions. *The Internet and Higher Education* 27 (2015), 74–89.

[25] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[26] Jinsook Lee, Chris Brooks, Renzhe Yu, and Rene Kizilcec. 2023. Fairness Hub Technical Briefs: AUC Gap. *arXiv e-prints* (2023), arXiv–2309.

[27] Liang Yi Li and Chin Chung Tsai. 2017. Accessing online learning material: Quantitative behavior patterns and their effects on motivation and learning performance. *Computers and Education* 114 (11 2017), 286–297. https://doi.org/10.1016/J.COMPEDU.2017.07.007

[28] Martín Liz Domínguez, Manuel Caeiro Rodríguez, Martín Llamas Nistal, Fernando Ariel Mikic Fonte, et al. 2019. Predictors and early warning systems in higher education: A systematic literature review. *Learning Analytics Summer Institute Spain 2019 (LASI-SPAIN 2019), Vigo, España, 27-28 junio 2019* (2019).

[29] Javier López-Zambrano, Juan A Lara, and Cristóbal Romero. 2022. Improving the portability of predicting students' performance models by using ontologies. *Journal of computing in higher education* (2022), 1–19.

[30] Javier López-Zambrano, Juan A. Lara, and Cristóbal Romero. 2020. Towards Portability of Models for Predicting Students' Final Performance in University Courses Starting from Moodle Logs. *Applied Sciences 2020, Vol. 10, Page 354* 10 (1 2020), 354. Issue 1. https://doi.org/10.3390/APP10010354

[31] Leah P. Macfadyen and Shane Dawson. 2010. Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers and Education* 54 (2 2010), 588–599. Issue 2. https://doi.org/10.1016/J.COMPEDU.2009.09.008

[32] Pedro Manuel Moreno-Marcos, Tinne De Laet, Pedro J Muñoz-Merino, Carolien Van Soom, Tom Broos, Katrien Verbert, and Carlos Delgado Kloos. 2019. Generalizing predictive models of admission test success based on online interactions. *Sustainability* 11, 18 (2019), 4940.

[33] Hardt Moritz, Price Eric, Srebro Nati, DD Lee, M Sugiyama, UV Luxburg, I Guyon, and R Garnett. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016), 3315–3323.

[34] Cedric Bheki Mpungose and Simon Bheki Khoza. 2022. Postgraduate Students' Experiences on the Use of Moodle and Canvas Learning Management System. *Technology, Knowledge and Learning* 27, 1 (mar 2022), 1–16.

[35] Abelardo Pardo, Feifei Han, and Robert A. Ellis. 2017. Combining University student self-regulated learning indicators and engagement with online learning events to Predict Academic Performance. *IEEE Transactions on Learning Technologies* 10 (1 2017), 82–92. Issue 1. https://doi.org/10.1109/TLT.2016.2639508

[36] Bardh Prenkaj, Paola Velardi, Giovanni Stilo, Damiano Distante, and Stefano Faralli. 2020. A survey of machine learning approaches for student dropout prediction in online courses. *ACM Computing Surveys (CSUR)* 53, 3 (2020), 1–34.

[37] J. Rhode, Stephanie Richter, Peter Gowen, T. Miller, and C. Wills. 2017. Understanding faculty use of the learning management system. *Online Learning Journal* 21 (01 2017), 68–86. https://doi.org/10.24059/olj.v%vi%i.1217

[38] Tue Tjur. 2009. Coefficients of determination in logistic regression models—A new proposal: The coefficient of discrimination. *The American Statistician* 63, 4 (2009), 366–372.

[39] U.S. Equal Employment Opportunity Commission. 2018. 29 CFR Part 1607 - Uniform Guidelines on Employee Selection Procedures (1978).

[40] Jonathan Vasquez Verdugo, Xavier Gitiaux, Cesar Ortega, and Huzefa Rangwala. 2022. Faired: A systematic fairness analysis approach applied in a higher educational context. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. 271–281.

[41] Richard Joseph Waddington and Sungjin Nam. 2014. Practice Exams Make Perfect: Incorporating Course Resource Use into an Early Warning System. (2014). https://doi.org/10.1145/2567574.2567623

[42] Feng Hsu Wang. 2017. An exploration of online behaviour engagement and achievement in flipped classroom supported by learning management system. *Computers & Education* 114 (11 2017), 79–91. https://doi.org/10.1016/J.COMPEDU.2017.06.012

[43] William J Youden. 1950. Index for rating diagnostic tests. *Cancer* 3, 1 (1950), 32–35.

[44] Renzhe Yu, Qiujie Li, Christian Fischer, Shayan Doroudi, and Di Xu. 2020. Towards Accurate and Fair Prediction of College Success: Evaluating Different Sources of Student Data. *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*, 292–301. https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_194.pdf

[45] Nick Z. Zacharis. 2015. A multivariate approach to predicting student outcomes in web-enabled blended learning courses. *The Internet and Higher Education* 27 (10 2015), 44–53. https://doi.org/10.1016/J.IHEDUC.2015.05.002

[46] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*. PMLR, 325–333.