Towards Fair and Privacy-Aware Transfer Learning for Educational Predictive Modeling: A Case Study on Retention Prediction in Community Colleges

Chengyuan Yao
Teachers College, Columbia University
New York, USA
cy2706@tc.columbia.edu

Carmen Cortez
Teachers College, Columbia University
New York, USA
cc4144@tc.columbia.edu

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

Abstract

Predictive analytics is a widely used application of learning analytics, but many resource-constrained institutions lack the capacity to develop their own predictive models or rely on proprietary models trained in different contexts with little transparency. In this context, transfer learning holds promise for expanding reliable and equitable access to predictive analytics, but this potential remains underexplored given existing legal and technical constraints. In this paper, we examine transfer learning strategies in the context of retention prediction at two-year community colleges in the United States, which enroll the most postsecondary students from underserved communities with higher dropout rates than selective universities. We envision a scenario where community colleges can collaborate with each other and with four-year universities to develop retention prediction models under privacy constraints, and evaluate the risks and potential improvement strategies of cross-institutional model transfer for different stakeholders. Using detailed administrative records from 4 research universities and 23 community colleges, which cover more than 800,000 students across 7 cohorts, we first identify performance and fairness degradation when source (external) models are deployed at a target institution without any localization. Fortunately, publicly available institution-level contextual information can be used to forecast these performance drops and offer early guidance for model portability. For model developers under data privacy regulations, sequential training that selects training institutions based on demographic similarities proves useful for enhancing the general fairness of resulting models without compromising performance. For target institutions without local data to fine-tune source models, we find that customizing evaluation thresholds for different sensitive groups is more successful than established transfer learning techniques at improving performance and fairness of deployed models. Our findings suggest the value of transfer learning for more accessible educational predictive modeling and call for judicious use of contextual information in model training, selection, and deployment to achieve reliable and equitable model transfer[1].

Keywords

Predictive Analytics; Transfer Learning; Algorithmic Fairness; Privacy; Intersectionality; College Retention; Community Colleges; Higher Education

ACM Reference Format:
Chengyuan Yao, Carmen Cortez, and Renzhe Yu. 2025. Towards Fair and Privacy-Aware Transfer Learning for Educational Predictive Modeling: A Case Study on Retention Prediction in Community Colleges. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3706468.3706567

1 Introduction

With the advancement of digital infrastructure and the abundance of big data in education, predictive analytics has been playing an increasingly important role in various educational decision-making scenarios [27]. The development of reliable predictive models typically relies on access to high-quality training datasets and careful modeling decisions, but in reality, under-resourced educational agencies with the greatest need for predictive analytics often lack the necessary local infrastructure and resources to build customized models. This dilemma challenges the promise of predictive analytics to alleviate the scarcity of quality educational resources for disadvantaged learners and improve educational equity accordingly.

In this context, transfer learning – a machine learning technique where a model trained on one task is adapted to perform a different but related task by leveraging existing knowledge [29]– becomes a potential solution to bridging gaps in building predictive analytics between differently resourced schools, institutions, and districts. In fact, many educational organizations have already deployed externally developed predictive models, representing a basic scenario of transfer learning. However, the performance and limitations of such model transfer processes have not been fully understood in educational contexts. Therefore, we present this study that formally investigates transfer learning strategies in one of the most studied educational predictive modeling tasks – college retention prediction. We focus on fairness and privacy of the technical processes, addressing major ethical concerns for educational stakeholders.

[1]Codes to reproduce our experiments are available at https://github.com/AEQUITAS-Lab/Transfer-Learning-LAK-2025

LAK 2025, Dublin, Ireland

ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706567

We focus on college retention prediction because college dropout is a persistent challenge in higher education. In the United States, only 39.2% of students at two-year public institutions graduate within three years, and 65.7% of students at four-year public institutions graduate within six years [33]. As part of their efforts to improve student retention and success, many institutions have utilized retention predictive models to identify at-risk students and inform timely interventions [1, 3, 4]. As with other predictive analytic tasks in education, under-resourced institutions, such as community colleges, often lack access to high-quality data and technical expertise for in-house model development and in the best case rely on predictive models developed by third-party vendors with limited transparency or local customization [32]. In addition, existing privacy and data protection laws, such as the Family Educational Rights and Privacy Act (FERPA), impose restrictions on sharing student data beyond institutional boundaries and add further conditions to the feasibility of transfer learning strategies.

Our study provides a holistic evaluation of ethical issues associated with existing practices of cross-institutional transfer of retention prediction models, as well as potential improvement strategies in line with current regulations. We focus on retention prediction for community colleges given their critical role in serving disadvantaged populations and improving economic and social mobility. Using comprehensive administrative data from 27 institutions across seven student cohorts, we consider risks and transfer learning strategies for both model developers (e.g., third-party vendors) and model users (i.e., community colleges). Formally, we address the following research questions:

RQ1. What are the risks associated with directly applying a pre-trained model for college retention prediction? Under what conditions are these risks higher or lower?

RQ2. For model developers, what strategies during the model training and selection phases may alleviate the risks in model transfer without access to model users' local historical data?

RQ3. For model users, what strategies during the model evaluation phase may alleviate the risks in model transfer without access to model training details?

We expect this work to contribute to existing research and practice in a few important ways. First, it will advance the understanding of predictive model portability in education by quantifying the risks associated with cross-context model deployment and identifying strategies to mitigate these risks. Second, it will contribute empirical research on responsible AI by addressing fairness and privacy concerns simultaneously in transfer learning. Third, it will advance more equitable scholarship on college retention prediction by focusing on community colleges which have a higher demand for predictive analytics but are understudied in existing research.

2 Related Work

2.1 Trustworthy Transfer Learning

Transfer learning has emerged as a powerful technique in the field of machine learning, which enables models to leverage knowledge from one domain or task to improve performance on another. This approach has been particularly successful in fields such as image processing and natural language processing [14, 15, 41]. The core idea behind transfer learning is to utilize the knowledge gained from a well-studied problem with abundant data to address a related problem with limited data. This method not only enhances model performance but also reduces the need for extensive data collection and labeling, which can be both time-consuming and costly [19].

As transfer learning applications expand into sensitive and high-stakes fields such as healthcare, education, and finance, researchers have started to focus their attention on the trustworthiness of the technical strategies in real-world contexts. Two prominent aspects of trustworthiness are fairness and privacy. Algorithmic fairness refers to the principle that machine learning models should provide equitable outcomes and benefits across different demographic groups [25]. In transfer learning, the challenge of fairness arises from potential demographic disparities between the source domain, where the model is trained, and the target domain, where it is applied. These disparities can lead to models that perform inequitably for certain groups in the target domain, resulting in biased decision-making [31]. Various methods have been proposed to mitigate fairness issues in the transfer learning process across different fields [7, 24, 25, 30], but these approaches largely focus on single sensitive groups, such as race or gender, without addressing the complexities of intersectionality which considers the compound and unique challenges faced by individuals with intersecting identities [18].

Privacy poses another ethical and technical challenge within the context of transfer learning. Different fields have distinct privacy constraints. For example, in education, privacy and data protection are regulated by legislation such as the Family Educational Rights and Privacy Act (FERPA), which generally prohibits the disclosure of personally identifiable information from student records without written consent [35]. In general, the primary privacy concern in transfer learning is to preserve the privacy of source domain data. Recent work on source-free domain adaptation (SFDA) addresses privacy and data sharing concerns by adapting a pre-trained model to a new target domain without requiring access to source domain data [40]. Researchers have demonstrated its success in computer vision [22, 37], but the application of this strategy in real-world problems remains limited. In addition, online learning, which involves continuously updating the model as new data become available and allows real-time adaptation to changing data distributions [13], has been used to formulate transfer learning under the term of "online transfer learning" [42]. In the process of developing transfer models, this approach can simulate scenarios of cross-context collaboration without violating privacy, such as sharing model weights instead of source data to collaboratively build a more effective transfer model.

2.2 Transfer Learning in Educational Predictive Modeling

Model portability has been one of the key challenges in learning analytics [10]. Early empirical work revealed how educational predictive models may fail to perform cross different instructional, institutional, and social contexts. For example, [28] found that models for detecting students' affective states in tutoring systems struggled to generalize across student populations, particularly when rural students were the target population. [21] examined the application of academic achievement prediction models trained on US

datasets in less developed countries and found significant performance drops.

More recently, researchers have started to examine the mechanisms and potential solutions to the issue of model portability, some of whom have formally taken the perspectives of transfer learning. One of the primary barriers to effective model transfer lies in the differences between the source and target contexts, and [38] proposed an analytical pipeline to understand how contextual differences moderate the portability of course-level performance prediction models. In terms of strategies to improve model portability, [34] explored three transfer learning methods with different model inputs and training paradigms in the context of early success prediction in MOOCs, and found that models with course information had the most satisfactory performance. In our context of college retention prediction across institutional contexts, [23] addressed contextual differences by using an instance weighting strategy to adjust the source model based on the distribution characteristics of the target institution. While this approach improved the overall goodness-of-fit of the transferred model, it requires access to both source and target data, which poses challenges under common privacy constraints. [9] explored a handful of model ensemble strategies across institutions without data sharing and showed some promise in improving model transfer when institutional characteristics are not substantially different.

Despite the existing efforts, the potential of transfer learning is still underexplored in the context of educational predictive modeling. Informed by the accomplishments and limitations of prior research, this study presents a comprehensive evaluation of both ethical risks and improvement strategies when predictive models are transferred across educational contexts.

2.3 College Retention Prediction

Over the past few decades, predicting college retention has become a critical area of research in higher education. Researchers have leveraged various data sources, such as institutional administrative data, digital footprints, and survey data, along with cutting-edge machine learning techniques, to identify students at risk of dropping out [2, 5, 11, 39]. Current research on this topic across different countries and institutions has shown promising results in accurately predicting college retention for different student populations at different time points. However, most research is conducted in the context of research-intensive institutions which tend to be more selective and suffer less from retention problems. By contrast, less selective and under-resourced institutions, such as community colleges in the United States, are extremely underrepresented in the literature. This is particularly concerning as community colleges and their counterparts in other countries primarily serve socioeconomically disadvantaged populations such as racial minorities and low-income students who face greater challenges in achieving academic success and are more likely to drop out of college. Accordingly, these colleges are in greater need of the support that retention prediction models can provide. The limited literature in community college contexts has examined performance and fairness in college retention prediction and found that underserved populations may benefit less from predictive models, further underscoring the need for equitable and effective predictive approaches [5, 6]. Related to such challenges, [9] explored cross-institutional transfer of predictive models to inform collaboration between research-intensive universities and under-resourced institutions in model development, although they did not empirically evaluate the approaches in under-resourced community colleges.

3 Study Context

We focus on the problem of student retention prediction at community colleges, which may benefit the most from predictive analytics due to their high proportion of minority and underprepared students and low retention rates [26]. Similar to [9], we envision a collaborative model development paradigm where a community college can deploy externally trained predictive models. A *source institution* is one whose data is part of the training data of the deployed model(s), whereas a *target institution* is the one that deploys the trained model(s). Given our focus on community colleges, target institutions are always community colleges but source institutions can be of any type. There is never data sharing across institutions in line with existing data laws and regulations, although third-party model developers (e.g., a vendor or inter-university consortium) may have simultaneous but separate access to data from multiple institutions in model development. All the analyses in this paper are performed in this context.

4 Data and Methods

4.1 Data Sources

4.1.1 Administrative Records. Through existing research partnerships, we gained access to detailed administrative records at four public research universities located across three states of the United States and 23 community colleges within a southeastern state. These records came from student information systems (SIS) and included individual students' background characteristics as well as academic records in college. For this paper, we restricted the sample to first-time, first-year students entering college in Fall terms between 2013 and 2019, covering over 800,000 students in total. For this study, we purposefully excluded more recent student cohorts whose experience was complicated by the COVID-19 pandemic. Based on these raw administrative records, we created a shared schema of common variables across the 27 institutions[2].

4.1.2 Contextual Factors. We collected institutional-level contextual factors from the Integrated Postsecondary Education Data System (IPEDS) [3], a comprehensive data collection and analysis system for higher education in the United States, managed by the National Center for Education Statistics (NCES). IPEDS serves as the primary source of detailed information on colleges and universities, covering a wide range of aspects such as enrollment, graduation rates, student demographics, faculty composition, finances, institutional costs, and financial aid. For this study, we included 64 variables across six categories for all 27 institutions: school

[2]The shared schema can be found at https://github.com/AEQUITAS-Lab/Transfer-Learning-LAK-2025/blob/main/Table1-Data-Schema.csv
[3]https://nces.ed.gov/ipeds

characteristics, academic compositions, demographic composition, completion rates, cost, and financial aid[4].

4.2 Prediction Task

Our central technical task is predicting first-year retention, defined as whether a student who enters an institution for the first time in the fall will re-enroll at the same institution the following fall. This definition aligns with the National Student Clearinghouse's standard for retention [8]. The predictors include student-level variables specified in the shared schema.

In general, we include students entering college between 2013 and 2018 in the training set and reserve those entering in Fall 2019 for testing purposes. Given our focus on community colleges, the 4 research universities are only used for training but not for testing. The specific training and testing data vary across transfer learning strategies explored in this study, as detailed in the next subsection.

4.3 Transfer Learning Strategies

We assume that transfer learning has the potential to help community colleges more responsibly take advantage of predictive analytics, as many of them lack the capacity to develop local models or simply rely on vendors' models trained elsewhere. Therefore, we evaluate the following technical transfer learning strategies as well as some of their combinations that can be adopted by target institutions or model developers. As a baseline comparison, we also include an "ideal local" strategy.

Ideal Local: A local model is trained using historical data from the same target institution on which it is evaluated.

Direct Transfer: A model is trained using data from a source institution and then evaluated on a different target institution.

Sequential Training: The process begins with training a model on the source institution's data. The trained model's weights are then transferred to another institution to perform another round of training. Finally, the model is evaluated on the target institution. To prevent catastrophic forgetting during the sequential training process, we implement the Elastic Weight Consolidation (EWC) strategy [16]. For instance, EWC helps the model retain prior knowledge (e.g., from the first training institution) while training on subsequent datasets (e.g., the second training institution).

Source-Free Domain Adaptation (SFDA): SFDA is advantageous for addressing data distribution shifts under privacy constraints, as it enables adaptation to new target domains without directly accessing source data, thereby preserving data privacy. In this study, we explore three benchmark methods within the SFDA framework:

- **Source Hypothesis Transfer (SHOT) [22]**: SHOT adapts the feature extractor to the target domain while keeping the classifier fixed, using self-supervised pseudo-labeling and information maximization to align target domain with the source hypothesis.
- **TENT [37]**: TENT adapts a model during test time by minimizing the entropy, updating the batch normalization parameters without requiring labeled data.
- **Pseudo-Labeling [20]**: Pseudo-labeling assigns labels to unlabeled target data based on the predicted probability, using these confident predictions to adapt the model to the target domain.

Customized evaluation thresholds: In the model evaluation stage, we propose two customized evaluation threshold methods: one based on the institution's overall historical outcome statistics (overall-optimal) and the other utilizing demographic-specific thresholds derived from the historical outcome statistics of individual groups (group-optimal).

In this study, we employed a fully connected neural network model, consisting of a feature extractor, a bottleneck layer, and a classifier, to perform the binary prediction task, which is suitable for all strategies mentioned above.

The workflow of utilizing these strategies is as follows: In RQ1, we evaluate the risks associated with direct transfer by analyzing the contextual similarity (discussed in section 4.4.3) between source and target institutions. For RQ2, we use contextual similarity to guide model selection and employ sequential training to improve model adaptation across institutions. In RQ3, we apply SFDA methods and implement customized evaluation threshold methods to assess and enhance the performance of transfer models in the target institution.

4.4 Key Metrics

4.4.1 Performance and Fairness. In binary predictive analysis, various performance metrics have been proposed in the literature, each focusing on different dimensions of model evaluation. Widely used metrics, such as those derived from the confusion matrix, require the selection of a specific decision threshold. However, determining a universally applicable threshold for various training and test institutions poses a big challenge, particularly when there are shifts in data distributions. In such cases, the use of a default threshold may lead to suboptimal performance. Therefore, we adopt the **Area Under the Receiver Operating Characteristic Curve (AUC)** as a primary metric for model evaluation, which is formally defined as:

$$AUC(f(\theta)) = \int_0^1 \mathrm{TPR}(\mathrm{FPR}(f_t(\theta)))\, dt$$

where t is a prediction threshold of the model with the decision rule $f_t(\theta, x) = 1$ if $f(\theta, x) \geq t$, and TPR, FPR stands for true positive rate and false positive rate respectively. AUC scores range from 0 to 1, with higher scores indicating a better ability to distinguish between positive and negative classes, while a random guess will achieve an AUC score of 0.5. AUC offers a threshold-independent assessment of the model's discriminative power by evaluating its performance across all possible thresholds.

To measure the difference in predictive performance under cross-institutional transfer, we employ Δ**AUC** [9], defined as $\Delta AUC(T, T') = AUC(T) - AUC(T')$, where $AUC(T)$ refers to the AUC of a model trained using a transfer scheme T. ΔAUC allows us to quantify the performance difference between various transfer learning schemes. $\Delta AUC = 0$ indicates that the model performs equally well under both schemes. A positive ΔAUC suggests the model performs better in context T' than in T, while a negative ΔAUC indicates the opposite. Mostly in this study, we evaluate the direct transfer model

[4]The details of these variables can be found at https://github.com/AEQUITAS-Lab/Transfer-Learning-LAK-2025/blob/main/Table2-Institutional-Contextual-Factors.csv

performance drop compared to the ideally local model, which is $\Delta AUC(\text{local}, \cdot)$.

To evaluate fairness across subgroups and account for intersectionality, we employ the concept of **AUC Gap**, as introduced by [9]:

$$\max_{g,g' \in \mathcal{G}} \left| \mathbb{E}_{\mathcal{D}_k} \left[f(\theta(\mathcal{D}_{k,g})) \right] - \mathbb{E}_{\mathcal{D}_k} \left[f(\theta(\mathcal{D}_{k,g'})) \right] \right|$$

where $\mathcal{D}_{k,g}$ and $\mathcal{D}_{k,g'}$ indicate the subset of the data in group g and g', respectively. The AUC Gap quantifies the maximum disparity in AUC scores among subgroups, representing the worst-case performance differential within a set of subgroups. In this study, we employ this method to examine the intersectional group based on gender and underrepresented minority (URM) status.

During model evaluation stage, we propose strategies for determining customized thresholds for individual test sets. Since AUC is threshold-independent, we measure the overall model performance with the **Matthews Correlation Coefficient (MCC)** which accounts for the balance between true positives, false positives, true negatives, and false negatives. In addition, given that retention interventions such as early warning systems often emphasize early identification of students at risk of dropping out, we consider **Specificity** in our evaluation. Specificity measures the proportion of actual negatives (students who are at risk of dropping out) that are correctly classified by the model.

To evaluate fairness when AUC Gap is not applicable, we use **Equalized Odds (EO)**, which measures the parity of the true positive rate (TPR) and true negative rate (TNR) between two groups, $A = 0$ and $A = 1$. To derive an overall EO score, we compute the EO TPR and EO TNR separately for each group, then take their average to obtain a comprehensive measure of fairness.

$$EO_{(A)} = \frac{|\text{TPR}_{(A=0)} - \text{TPR}_{(A=1)}| + |\text{TNR}_{(A=0)} - \text{TNR}_{(A=1)}|}{2}$$

In this part, we focus on two demographic categories—race and gender—which are commonly used in evaluating algorithmic fairness. Specifically, we assess fairness between underrepresented racial minorities (URM) vs. non-URM and female vs. male.

4.4.2 Distributional Difference. In this study, we utilize the performance distribution to assess the overall performance across all target institutions. To evaluate the differences between distributions, we employed the **Wasserstein distance**, which is a measure of the distance between two distributions and is defined as:

$$W(P, Q) = \int_{-\infty}^{\infty} \left| F_P(x) - F_Q(x) \right| \, dx$$

where $F_P(x)$ and $F_Q(x)$ are the cumulative distribution functions (CDFs) of the probability distributions P and Q. However, the Wasserstein distance does not have an absolute reference value for evaluation. Existing studies suggest that it should be interpreted relative to the range of the data in order to provide meaningful insights [36]. In this study, we define a 5% of the data range as the threshold for a "small" Wasserstein distance, known as **Wasserstein Threshold for Notable Distribution Difference (WTNDD)**. If the WTNDD passed, it indicates no notable difference between the distributions. Conversely, if WTNDD failed, it suggests a notable difference between the distributions.

4.4.3 Contextual Similarity. As noted in Section 4.1.2, we collect contextual factors for each institution across six categories. We calculate the similarity scores for each category for each pair of institutions using Gower's distance [12], a metric well-suited for mixed data types. The similarity scores are normalized to fall within the range of 0 to 1 for consistency across categories.

To construct an overall similarity score between source and target institutions, we utilize the results of a regression analysis on AUC Drop ($\Delta AUC(\text{local}, \cdot)$) (see Section 5.1 for more details). Specifically, we apply the **"Coefficient-Significance Weighted Similarity Method"**, which assigns weights to each similarity category based on the regression coefficients and their statistical significance. The weight for each similarity measure i is calculated as:

$$W_i = \frac{|\beta_i|}{(1 + p_i)}$$

where β_i is the regression coefficient for category i, and p_i is its p-value. This weighting method reduces the influence of variables with lower statistical significance while emphasizing variables with greater impact. The weights were normalized so that their sum equals 1. Using these normalized weights, the overall similarity score between source and target institutions is computed as:

$$X_{\text{overall}} = \sum_{i=1}^{n} W_i \cdot X_i$$

where X_i represents the normalized similarity score for category i, and W_i is the corresponding weight.

5 Results

5.1 Risks of Direct Transfer

In RQ1, we examine the risks of performance and fairness associated with directly transferring a model developed in a different context. Figure 1 (left) shows the AUC for direct transfer models trained on each source institution and applied to each target institution. The results demonstrate substantial variability in direct transfer AUC performance depending on the source institution. Figure 1 (right) shows the AUC Gap of these models across target institutions. The results indicate that source models face differing levels of fairness challenges.

To assess under which conditions model transferability is more vulnerable, we examine the relationship between contextual similarity and model portability by regressing two metrics - AUC Drop, $\Delta AUC(\text{local}, \cdot)$, and fairness disparity, AUC Gap - on similarity scores across different contextual categories between the source and target institutions. The regression equations are defined as follows:

$$\Delta AUC(\text{local}, \cdot) = \beta_0 + \sum_{i=1}^{6} \beta_i X_i + \epsilon, \quad \text{AUC Gap} = \beta_0 + \sum_{i=1}^{6} \beta_i X_i + \epsilon$$

where X_i represents the similarity score of each category as discussed in section 4.4.3. Figure 2 (left) shows that models perform better, with smaller AUC drops, when the source and target institutions are more similar in terms of school, academic, demographic, and completion factors. Figure 2 (right) shows that greater similarity between source and target institutions in school, academic, and cost factors reduces the AUC gap, indicating improved fairness.

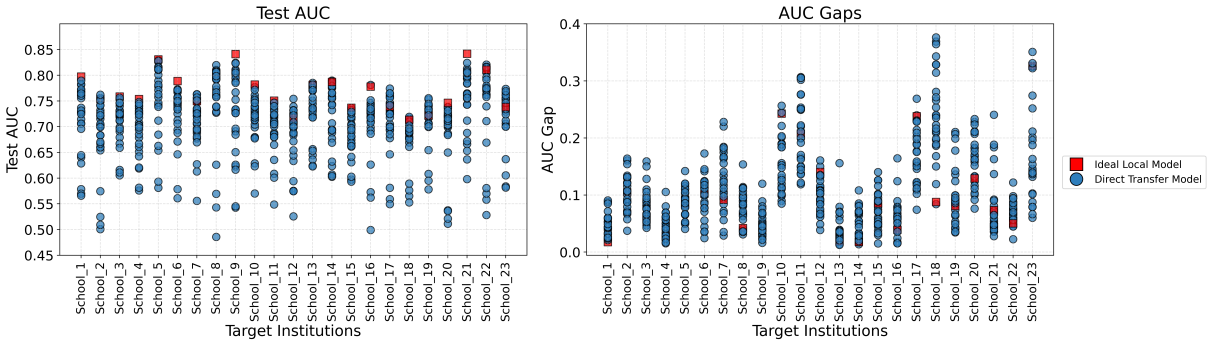

Figure 1: Test AUC (left) and AUC Gaps (right) of pre-trained models for each target institutions. Blue circles represent the tested metrics of pre-trained direct transfer models for each target institution, while red squares indicate the tested metrics of ideal local models.

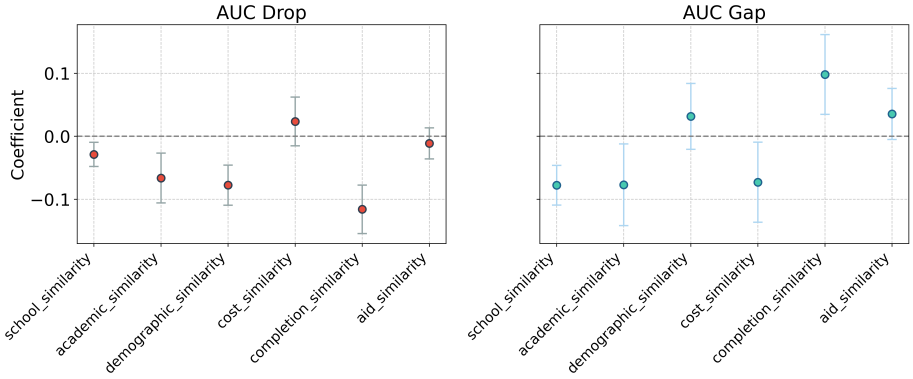

Figure 2: Association between contextual similarity metrics and AUC Drop (left) and AUC Gap (right). A negative coefficient indicates a reduction in the AUC Drop or AUC Gap. Error bars represent 95% confidence intervals.

However, higher similarity in completion factors tends to increase the AUC gap. Notably, the regression model for AUC Drop has an R-squared value of 0.600, while the R-squared value for the AUC Gap regression is only 0.056. This implies that while performance degradation is strongly influenced by differences in contextual factors between institutions, fairness change is more complex and less predictable based on the examined variables.

Following the regression results for AUC Drop, we calculate an overall similarity score for each pair of source and target institutions using Coefficient-Significance Weighted Similarity Method discussed in section 4.4.3. Since cost similarity tends to enlarge the AUC drop, we drop it from calculating the overall similarity. Figure 3 illustrates the relationship between overall similarity and three key metrics: Test AUC, Inverted AUC drop, and AUC Gap. The left two plots demonstrate a positive linear trend, indicating that higher overall similarity between institutions is associated with improved AUC scores and a reduced performance drop. Notably, when overall similarity falls below 0.5, the model tends to experience more pronounced performance issues. Additionally, two distinct clusters emerge within the ranges of 0.15–0.45 and 0.65–0.95, each characterized by relatively low internal variance. In contrast, the relationship between overall similarity and fairness, as measured by AUC Gap, exhibits a largely flat trend with a slight negative slope. This suggests that overall similarity exerts minimal influence on AUC Gap, consistent with the coefficient of determination results observed in previous analyses.

5.2 Transfer Learning Strategies in Model Training and Selection

From a model developer's perspective, privacy constraints pose challenges in selecting the most suitable pre-trained model that they can pass to the target institution. Based on the previous results, we propose that selecting the most similar school, based on overall contextual similarity, can help identify an appropriate model. The model selected using this method is referred as Most Similar Training Institution(MSTI) model. Figure 4 demonstrates that, although the distribution of Test AUC derived from the MSTI models generally underperforms relative to that of the ideal local models (WTNDD failed), it exhibits a marked improvement over the distribution of expected AUC obtained from random model selection (WTNDD failed). This indicates that selecting a model based on contextual similarity can enhance overall predictive performance

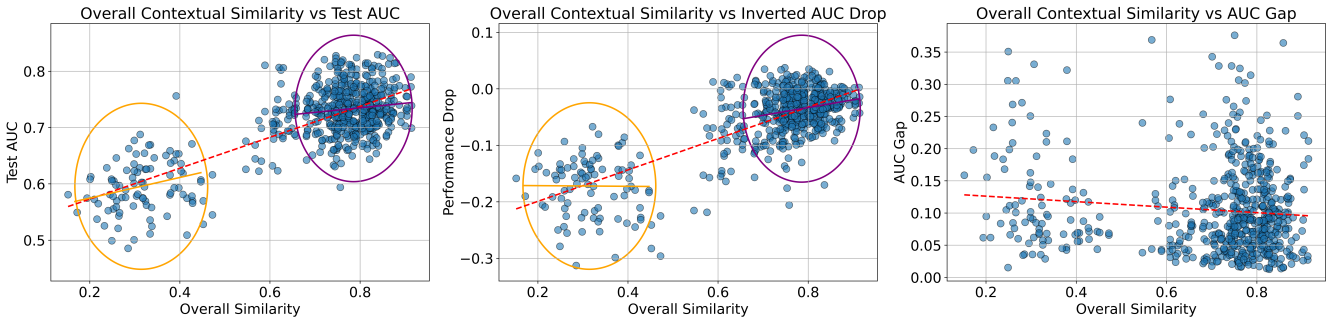

Figure 3: Relationship between overall contextual similarity and the model's transferability and fairness. Blue circles represent the tested metric of pre-trained direct transfer models on each target institution.

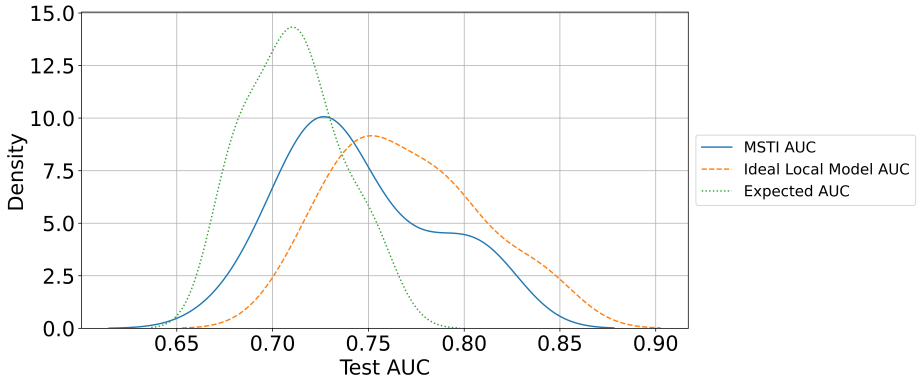

Figure 4: Performance distributions across all target institutions in the dataset, comparing Test AUC for three scenarios: MSTI model, the ideal local model, and the expected outcome. The histograms are normalized to have an area of 1 (density).

compared to a randomly chosen model. From the perspective of each target institution, Figure 5 indicates that the Test AUC from the MSTI model for 21 out of 23 target institutions exceeds the expected AUC. However, no significant improvements in fairness were observed, consistent with our previous conclusions that fairness is less predictable from contextual similarity.

Existing studies have demonstrated that a diverse training set can help address fairness issues [17]. However, privacy constraints prevent model developers from combining data from multiple institutions to create a more diverse training dataset. To tackle this challenge, we propose a potential solution aimed at improving model fairness while respecting privacy constraints. First, for each target institution, we firstly identify the MSTI model. Next, among other possible training institutions with an overall similarity greater than 0.6—chosen to avoid performance drops—we use demographic similarity to select the most demographically dissimilar institution, labeled as Training Institution 2. Finally, we conduct sequential training on MSTI and Training Institution 2. The obtained model is referred as sequential model.

We compare the performance and fairness metrics of the sequential model with the MSTI model across three fairness thresholds: when the MSTI AUC gap exceeds 0.00, 0.05, and 0.10. Figure 6 shows that sequential training has the potential to improve fairness without compromising performance. The WTNDD test for differences in the Test AUC distributions passed under all three thresholds, indicating no significant differences in Test AUC distributions between the MSTI model and the sequential model. In contrast, the WTNDD test for differences in AUC Gap distributions failed, with Wasserstein Distances of 0.26, 0.37, and 0.46 for MSTI AUC Gaps exceeding 0.00, 0.05, and 0.10, respectively. This increasing Wasserstein Distance for the AUC Gap reflects a growing distributional difference, which indicates that the less fair the original model is (i.e., the larger the AUC Gap), the stronger the ability of sequential training to mitigate fairness issues.

From the perspective of individual target institutions, Figure 7 shows that for the majority of the selected target institutions (8 out of 11), the AUC Gap decreased when the sequential model was implemented compared to the MSTI model. However, this reduction in AUC Gap is occasionally accompanied by a performance degradation for some institutions. Overall, the results suggest that sequential training has the potential to improve model fairness without compromising performance, particularly for models with higher levels of initial unfairness. Nonetheless, there remains variance across individual target institutions, which underscores the

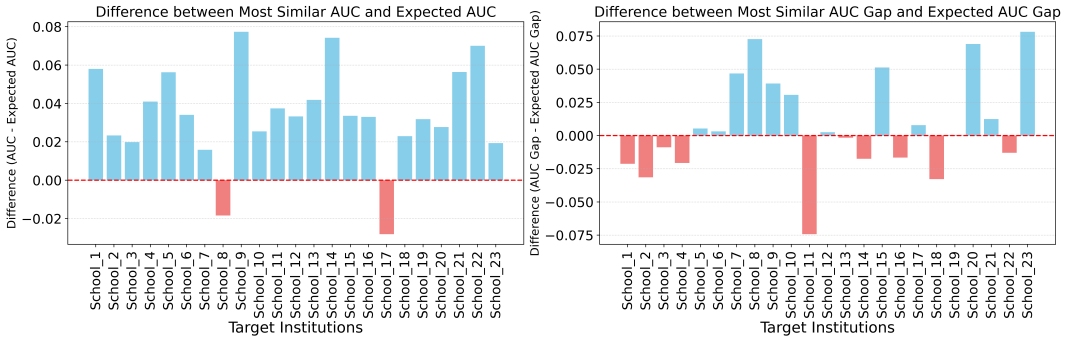

Figure 5: Differences between MSTI's Test AUC and AUC Gap with expected values across target institutions.

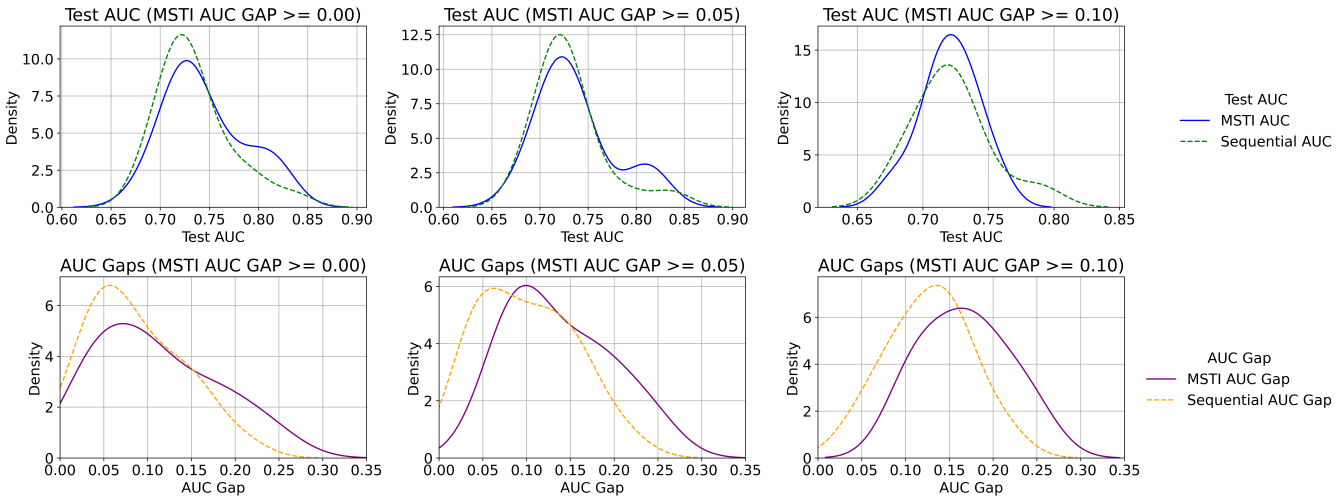

Figure 6: Distribution of performance and fairness for MSTI model and sequential model. (a) Top: Test AUC distributions when the source is MSTI model and the sequential model under varying MSTI AUC gap thresholds (>=0.00, 0.05, and 0.10). (b) Bottom: AUC Gap distributions for MSTI model and sequential model under the same thresholds. Histograms are normalized to have an area of 1 (density).

need for thorough consideration of the specific contextual and institutional differences when applying sequential training in practice.

5.3 Transfer Learning Strategies in Model Evaluation

For our third research question, we aim to investigate how a target institution can mitigate the risks associated with deploying pre-trained models. As mentioned in Section 4.3, SFDA offers a strategy to adapt a model to a new domain without requiring access to either the original training data or the labeled target dataset. Surprisingly, Figure 8 shows that SFDA techniques that have demonstrated considerable success in computer vision tasks were not as effective in predicting student retention rates. In fact, many of these adaptation methods underperformed compared to directly applying the pre-trained models without any adaptation.

During the evaluation phase of pre-trained models for a target institution, we compare the use of two customized evaluation threshold methods discussed in Section 4.3 with the default threshold to assess their effectiveness. We first compare the distributional differences in specificity and MCC of direct transfer models using the default threshold and the two optimal thresholds. Figure 9 demonstrates that, compared to using the default threshold, both the group-optimal and overall-optimal threshold methods improve specificity, enhancing the model's ability to accurately identify dropout cases (WTNDD failed). Additionally, the WTNDD test passed when comparing the specificity distributions of group-optimal and overall-optimal thresholds, which indicates no substantial difference between them. In addition, there is no notable difference in the MCC distributions among using the group-optimal, overall-optimal, and the default thresholds (WTNDD passed). In terms of fairness, Figure 10 presents the distributions of Equalized Odds (EO) scores of direct transfer models for different thresholds applied to male vs. female and URM vs. non-URM groups.

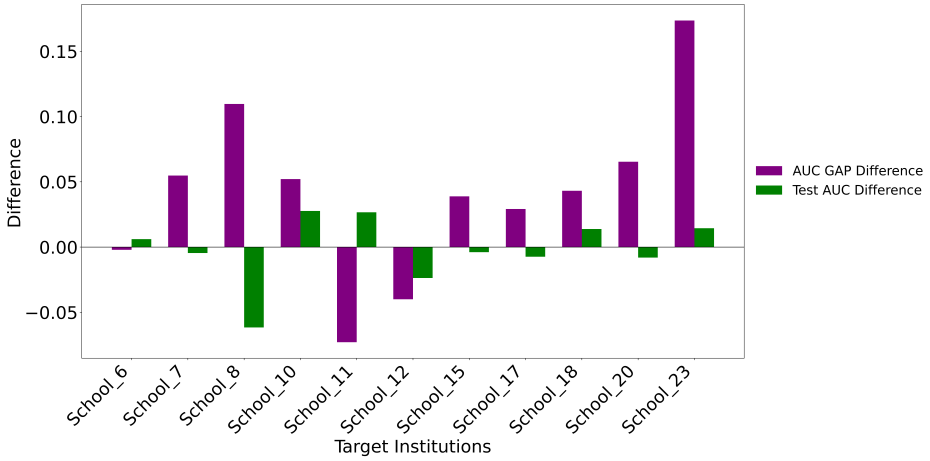

Figure 7: Differences in AUC Gap and Test AUC between the MSTI model and the sequential model for each target institution when MSTI AUC Gap > 0.10. A positive value indicates an improvement of tested metric with sequential training.

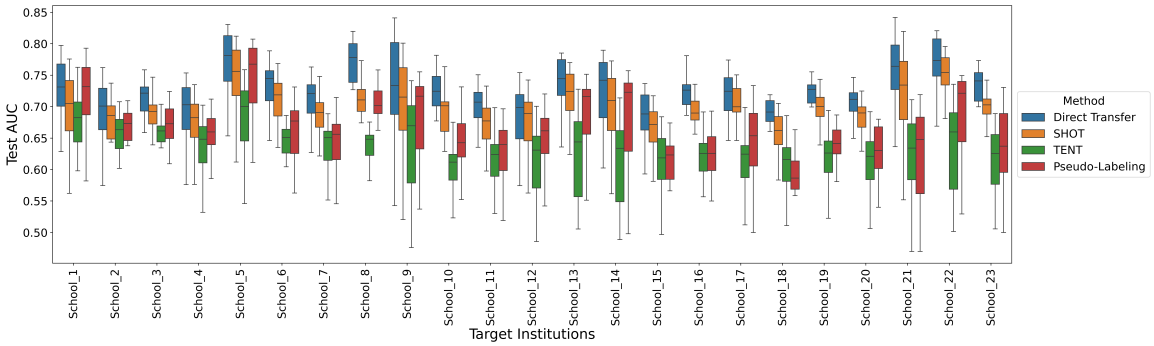

Figure 8: Test AUC distributions across four different methods—Direct Transfer (blue), SHOT (orange), TENT (green), and Pseudo-Labeling (red)—for each target institution. Each box represents the interquartile range (IQR) of the AUC values, with the whiskers extending to the minimum and maximum values within 1.5 times the IQR.

The results indicate that the use of group-optimal thresholds enhances fairness, as evidenced by lower EO scores across gender and URM/non-URM groups compared to both the default and overall-optimal thresholds (WTNDD failed). There is no notable difference in the distribution of EO scores between using the default and overall optimal thresholds (WTNDD passed). Our results show that employing overall-optimal and group-optimal thresholds can effectively improve the model's specificity without compromising overall performance, and employing group-optimal thresholds can further enhance fairness compared to overall-optimal thresholds.

6 Discussion and Conclusion

In this study, we examine the practical challenges and potential solutions for applying pre-trained models across institutions. We conduct the first large-scale empirical investigation of cross-institutional transfer learning of prediction models across both 4-year universities and 2-year community colleges. This study advances transfer learning in education by proposing actionable strategies to tackle transferability challenges identified in prior work [21, 38]. For example, our findings demonstrate that publicly available contextual information can effectively guide model selection and decision-making, even under privacy constraints. Additionally, we mitigate fairness concerns by proposing potential solutions such as sequential training to improve fairness from a developer's perspective and customizing evaluation thresholds for different groups to enhance fairness from a user's perspective. These strategies align with broader efforts to mitigate algorithmic bias and promote equity in predictive modeling [6, 17].

These findings have important implications. For model developers, our findings show that utilizing public contextual information facilitates model selection and can lay the groundwork for optimizing model choice under privacy constraints. For practitioners, our results demonstrate the feasibility of institutional collaboration under privacy constraints, showing that even without sharing

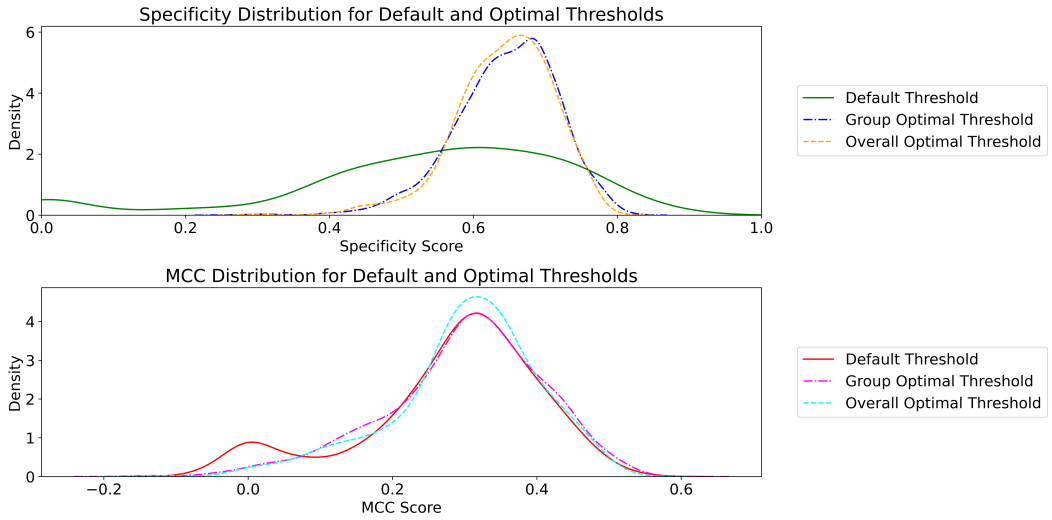

Figure 9: Distribution of specificity (top) and MCC (bottom) scores of direct transfer models under the usage of default threshold, group-optimal threshold, and overall-optimal threshold. Histograms are normalized to have an area of 1 (density).

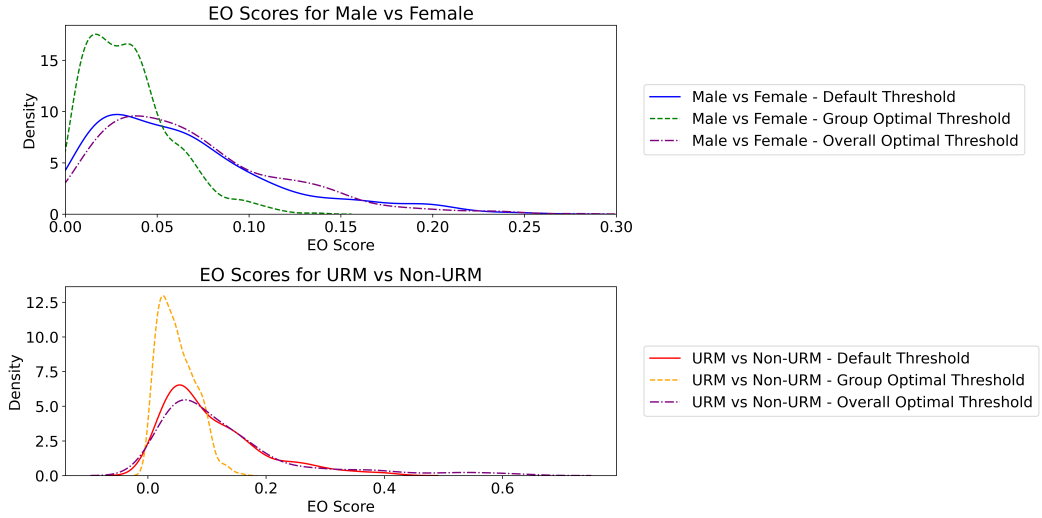

Figure 10: Distribution of Equalized Odds (EO) scores of direct transfer models. (a) Top: EO scores for male vs. female under the usage of default threshold, group optimal threshold, and overall optimal threshold. (b) Bottom: EO scores for URM vs. Non-URM under the usage of same thresholds. Lower score in EO suggests better model fairness. Histograms are normalized to have an area of 1 (density).

raw data, it is possible to construct improved models through sequential training that benefits diverse population. For educational researchers, our study addresses a critical gap by focusing on the impact of cross-institutional transfer models on community colleges, a relatively under-explored area in the field. In addition, by providing empirical evidence of how these models perform across institutions with varying resource levels, this research sets the stage for future efforts to enable resource-rich schools to support under-resourced institutions through model transfer and adaptation.

This study has several limitations. First, the models rely on common data features across institutions, which may exclude some important indicators due to varying data infrastructures. Second, the sample includes only research university and community college in the United States, which may limit the generalizability of the findings to other institutional types. Future research should address feature space differences between source and target institutions, expand the sample to include a wider range of institutions, and explore applications in other sectors, such as finance and healthcare,

to further validate transfer learning frameworks in privacy- and fairness-critical contexts.

Acknowledgments

This work was supported by funding from the Learning Engineering Virtual Institute through the Fairness Analysis and Transfer Learning Hub. We extend our gratitude to Rene Kizilcec, Christopher Brooks, and Catherine Finnegan for their invaluable data and technical support.

References

[1] Sattar Ameri, Mahtab J. Fard, Ratna B. Chinnam, and Chandan K. Reddy. 2016. Survival Analysis Based Framework for Early Prediction of Student Dropouts. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM '16)*. Association for Computing Machinery, 903–912.

[2] Lovenoor Aulck, Dev Nambi, Nishant Velagapudi, Joshua Blumenstock, and Jevin West. 2019. Mining University Registrar Records to Predict First-Year Undergraduate Attrition. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*. 9–18.

[3] Cédric Beaulac and Jeffrey S. Rosenthal. 2019. Predicting University Students' Academic Success and Major Using Random Forests. *Research in Higher Education* 60, 7 (2019), 1048–1064.

[4] Johannes Berens, Kerstin Schneider, Simon Görtz, Simon Oster, and Julian Burghoff. 2019. Early Detection of Students at Risk: Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods. *Journal of Educational Data Mining* 11, 3 (2019), 1–41.

[5] Kelli A. Bird, Benjamin L. Castleman, Zachary Mabel, and Yifeng Song. 2021. Bringing Transparency to Predictive Analytics: A Systematic Comparison of Predictive Modeling Methods in Higher Education. *AERA Open* 7 (2021), 233285842110376. https://doi.org/10.1177/23328584211037630

[6] Kelli A. Bird, Benjamin L. Castleman, and Yifeng Song. 2024. Are algorithms biased in education? Exploring racial bias in predicting community college student success. *Journal of Policy Analysis and Management* (31 January 2024).

[7] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R. Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. 2019. Fair Transfer Learning with Missing Protected Attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 91–98.

[8] A. Gardner. 2022. *Persistence and Retention: Fall 2020 Beginning Postsecondary Student Cohort.* Technical Report. National Student Clearinghouse Research Center, Herndon, VA.

[9] 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 *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. ACM. https://doi.org/10.1145/3593013.3594107

[10] 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. *The Internet and Higher Education* 28 (2016), 68–84.

[11] Dominik Glandorf, Hye Rin Lee, Gabe Avakian Orona, Marina Pumptow, Renzhe Yu, and Christian Fischer. 2024. Temporal and Between-Group Variability in College Dropout Prediction. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (Kyoto, Japan) *(LAK '24)*. Association for Computing Machinery, New York, NY, USA, 486–497. https://doi.org/10.1145/3636555.3636906

[12] J. C. Gower. 1971. A general coefficient of similarity and some of its properties. *Biometrics* 27, 4 (1971), 857–871.

[13] Steven C.H. Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. 2021. Online learning: A comprehensive survey. *Neurocomput.* 459, C (Oct. 2021), 249–289.

[14] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2018. Transfer learning for time series classification. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 1367–1376. https://doi.org/10.1109/bigdata.2018.8621990

[15] Jing Jiang and ChengXiang Zhai. 2007. Instance Weighting for Domain Adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Annie Zaenen and Antal van den Bosch (Eds.). Association for Computational Linguistics, Prague, Czech Republic, 264–271.

[16] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences* 114, 13 (March 2017), 3521–3526. https://doi.org/10.1073/pnas.1611835114

[17] René F. Kizilcec and Hansol Lee. 2022. Algorithmic Fairness in Education. In *Ethics in Artificial Intelligence in Education*. Routledge.

[18] Youjin Kong. 2022. Are "intersectionally fair" ai algorithms really fair to women of color? A philosophical analysis. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 485–494. https://doi.org/10.1145/3531146.3533114

[19] Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. *Political Analysis* 32, 1 (2024), 84–100.

[20] Dong-Hyun Lee. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*.

[21] Xingyu Li, Difan Song, Miaozhe Han, Yu Zhang, and Rene F. Kizilcec. 2021. On the limits of algorithmic prediction across the globe. arXiv:2103.15212 [cs.LG] https://arxiv.org/abs/2103.15212

[22] Jian Liang, Dapeng Hu, and Jiashi Feng. 2020. Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation. In *International Conference on Machine Learning (ICML)*.

[23] Mariia Luzan and Christopher Brooks. 2024. Evaluation of the instance weighting strategy for transfer learning of educational predictive models. *Proceedings of Machine Learning Research* 257 (9 August 2024), 19–28.

[24] Debmalya Mandal, Samuel Deng, Suman Jana, and Daniel Hsu. 2020. Ensuring Fairness Beyond the Training Data. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[25] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[26] Vanessa Smith Morest. 2013. From Access to Opportunity: The Evolving Social Roles of Community Colleges. *The American Sociologist* 44, 4 (2013), 319–328. http://www.jstor.org/stable/42635370

[27] Abdallah Namoun and Abdullah Alshanqiti. 2021. Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. *Applied Sciences* 11, 1 (2021), 237. https://doi.org/10.3390/app11010237

[28] Jaclyn Ocumpaugh, Ryan Baker, Sujith Gowda, Neil Heffernan, and Cristina Heffernan. 2014. Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology* 45, 3 (May 2014), 487–501.

[29] Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (2010), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

[30] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D. Ziebart. 2021. Robust Fairness Under Covariate Shift. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 11 (May 2021), 9419–9427.

[31] Jessica Schrouff, Natalie Harris, Oluwasanmi O Koyejo, Ibrahim Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alexander Brown, Subhrajit Roy, Diana Mincu, Chrsitina Chen, Awa Dieng, Yuan Liu, Vivek Natarajan, Alan Karthikesalingam, Katherine A Heller, Silvia Chiappa, and Alexander D'Amour. 2022. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. In *Advances in Neural Information Processing Systems*.

[32] W. Reed Scull, Mark Andrew Perkins, Jonathan W. Carrier, and Michael Barber. 2022. Community College Institutional Researchers' Knowledge, Experience, and Perceptions of Machine Learning. *Community College Journal of Research and Practice* 47, 5 (2022), 354–368.

[33] Doug Shapiro, Afet Dundar, Faye Huie, Phoebe Khasiala Wakhungu, Ayesha Bhimdiwala, and Sean Eric Wilson. 2019. *Completing College: A State-Level View of Student Completion Rates (Signature Report No. 16a).* Technical Report. National Student Clearinghouse Research Center, Herndon, VA.

[34] Vinitra Swamy, Mirko Marras, and Tanja Käser. 2022. Meta Transfer Learning for Early Success Prediction in MOOCs. In *Proceedings of the Ninth ACM Conference on Learning @ Scale* (New York City, NY, USA) *(L@S '22)*. Association for Computing Machinery, New York, NY, USA, 121–132.

[35] US Congress. 1974. Family Educational Rights and Privacy Act (FERPA). https://www.law.cornell.edu/uscode/text/20/1232g 20 U.S.C. § 1232g.

[36] Cédric Villani. 2008. *Optimal Transport: Old and New.* Grundlehren der mathematischen Wissenschaften, Vol. 338. Springer, Berlin, Heidelberg.

[37] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=uXl3bZLkr3c

[38] 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)* (Kyoto, Japan, March 18–22, 2024). ACM, New York, NY, USA, 12. https://doi.org/10.1145/3636555.3636936

[39] Renzhe Yu, Hansol Lee, and René F. Kizilcec. 2021. Should college dropout prediction models include protected attributes?. In *Proceedings of the Eighth ACM*

Conference on Learning @ Scale. Association for Computing Machinery, New York, NY, USA, 91–100.

[40] Zhiqi Yu, Jingjing Li, Zhekai Du, Lei Zhu, and Heng Tao Shen. 2023. A Comprehensive Survey on Source-free Domain Adaptation. arXiv:2302.11803 [cs.LG] https://arxiv.org/abs/2302.11803

[41] Chuanting Zhang, Haixia Zhang, Jingping Qiao, Dongfeng Yuan, and Minggao Zhang. 2019. Deep Transfer Learning for Intelligent Cellular Traffic Prediction Based on Cross-Domain Big Data. *IEEE Journal on Selected Areas in Communications* 37, 6 (2019), 1389–1401.

[42] Peilin Zhao, Steven C.H. Hoi, Jialei Wang, and Bin Li. 2014. Online Transfer Learning. *Artificial Intelligence* 216 (2014), 76–102. https://doi.org/10.1016/j.artint.2014.06.003