Cross-Institutional Transfer Learning for Educational Models: Implications for Model Performance, Fairness, and Equity

Josh Gardner
jpgard@cs.washington.edu
University of Washington

Renzhe Yu
renzheyu@tc.columbia.edu
Teachers College & Data Science Institute, Columbia University

Quan Nguyen
quan.nguyen@ubc.ca
University of British Columbia

Christopher Brooks
brooksch@umich.edu
School of Information, University of Michigan

Rene F. Kizilcec
kizilcec@cornell.edu
Department of Information Science, Cornell University

ABSTRACT

Modern machine learning increasingly supports paradigms that are multi-institutional (using data from multiple institutions during training) or cross-institutional (using models from multiple institutions for inference), but the empirical effects of these paradigms are not well understood. This study investigates cross-institutional learning via an empirical case study in higher education. We propose a framework and metrics for assessing the utility and fairness of student dropout prediction models that are transferred across institutions. We examine the feasibility of cross-institutional transfer under real-world data- and model-sharing constraints, quantifying model biases for intersectional student identities, characterizing potential disparate impact due to these biases, and investigating the impact of various cross-institutional ensembling approaches on fairness and overall model performance. We perform this analysis on data representing over 200,000 enrolled students annually from four universities without sharing training data between institutions.

We find that a simple zero-shot cross-institutional transfer procedure can achieve similar performance to locally-trained models for all institutions in our study, without sacrificing model fairness. We also find that stacked ensembling provides no additional benefits to overall performance or fairness compared to either a local model or the zero-shot transfer procedure we tested. We find no evidence of a fairness-accuracy tradeoff across dozens of models and transfer schemes evaluated. Our auditing procedure also highlights the importance of intersectional fairness analysis, revealing performance disparities at the intersection of sensitive identity groups that are concealed under one-dimensional analysis.[1]

[1]Code to reproduce our experiments is available at https://github.com/educational-technology-collective/cross-institutional-transfer-learning-facct-2023.

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06…$15.00
https://doi.org/10.1145/3593013.3594107

CCS CONCEPTS

• **Applied computing** → **Law, social and behavioral sciences**; **Education**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

Algorithmic Fairness, Education, Dropout Prediction, Transfer Learning, Intersectionality

ACM Reference Format:
Josh Gardner, Renzhe Yu, Quan Nguyen, Christopher Brooks, and Rene F. 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), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3593013.3594107

1 INTRODUCTION

Improvements in digital infrastructure have enabled numerous applications of machine learning across domains but in decentralized organizational contexts (e.g., universities, schools, hospitals, finance, and government), the capacity to use machine learning typically depends on the availability of local infrastructure and resources. Under-resourced institutions may therefore be unable to reap the full benefits of machine learning applications despite commonly having the greatest need. Externally developed models have increasingly been adopted in these cases to address this challenge. While this highlights a potential benefit of sharing predictive models within large-scale cross-institutional collaborations, the risks and benefits of cross-institutional modeling are not well understood, particularly in terms of the potential impact on the most vulnerable populations affected by these models.

We investigate the benefits and risks of cross-institutional transfer in the context of an important and pervasive application of machine learning in education: university student dropout prediction. Every year, over one million students drop out of college in the United States, and they are 100 times more likely to default on their student loan payments than those who graduate [46]. This leaves young adults with a major financial burden and little to improve their job prospects with an incomplete degree [66]. Dropout rates are especially high for students from minority groups[2], exacerbating existing inequities.

[2]https://nces.ed.gov/programs/raceindicators/indicator_red.asp

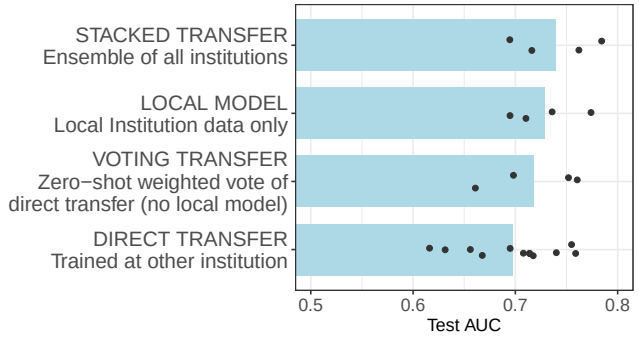


Figure 1: Summary of cross-institutional transfer methods evaluated in this work (transfer methods defined in Section 4.3). Each point represents one cross-institutional transfer trial (see Section 6) with an L_2-regularized logistic regression model. (See Figure 7 for results with LightGBM, MLP models.)

Lowering college dropout rates is a priority for many institutions of higher education. U.S. federal regulations incentivize colleges and universities to reduce dropout by requiring them to report dropout rates, performance-based funding [37], and college rankings that account for graduation rates [3]. This has led an increasing number of colleges and universities to adopt data-driven predictive models to identify at-risk students, in order to intervene early enough to support students before they drop out. While private vendors such as Civitas, Starfish Retention, and Hobsons sell software for at-risk prediction to institutions, they do not share the technical details of underlying models. This limits the ability of practitioners and researchers to audit the performance and possible biases of these models' predictions, especially for institutions other than those where models were trained [8, 49].

Concerns about algorithmic bias in education have motivated several recent studies that interrogate the fairness and ethics of predictive modeling in education [33, 49, 54, 76, 77]. Inter-university data partnerships, such as the Unizin Consortium[3], have also emerged to standardize data infrastructure, provide opportunities for multi-institutional model development by using data from multiple institutions during training, and even facilitate cross-institutional model sharing. However, the benefits and risks of this form of transfer learning are presently understudied, and studying cross-institutional learning in a research context can be challenging due to data privacy regulations: in most circumstances, student data cannot simply be shared between institutions or uploaded into public repositories due to federal regulations.

In this work, we seek to evaluate the implications for model *performance* and *fairness* of three approaches to *cross-institutional transfer learning*. We conduct the first large-scale, systematic analysis of cross-institutional transfer learning in higher education. We evaluate three transfer approaches (Section 4.3) motivated by real-world data, collaboration models, and institutional needs in higher education. We use datasets from four U.S. universities with diverse student populations (Section 4.1), propose metrics to evaluate model performance and fairness in cross-institutional transfer learning (Section 4.4), conduct a comprehensive set of experiments to measure the effects of different transfer approaches on the proposed metrics in terms of (a) overall performance (Section 6.1) and (b) fairness measured by intersectional performance disparities (Section 6.2), and evaluate fairness-performance tradeoffs (Section 6.3). Our main results are summarized in Figure 1. We discuss limitations and recommendations for future work in Section 7.

Our contributions: While prior research has explored prediction models based on multi-institutional educational datasets, to our knowledge, our work is the first to systematically investigate the implications of cross-institutional transfer learning for fairness and equity. This work required a year-long effort working with each institution to develop a common data schema to map their local data into this schema to enable cross-institutional learning. We contribute a novel methodology for auditing cross-institutional transfer learning, including metrics for measuring intersectional fairness of model transfer. Finally, our empirical results provide a useful benchmark for researchers and practitioners interested in cross-institutional transfer learning both within and outside of the domain of education. Our results demonstrate that (i) cross-institutional transfer is feasible even when no historical training data is present at a target institution (e.g., via direct or voting transfer); (ii) a simple zero-shot voting transfer method achieves similar performance to a local model for all institutions in our study, without requiring any local training data and at no cost to fairness; (iii) stacked ensembling provides no additional benefit over local training or zero-shot voting transfer from other institutions; and (iv) there does not exist a strict empirical tradeoff between fairness and accuracy across our broad set of models, transfer schemes, and institutions evaluated.

2 RELATED WORK

2.1 College Dropout Prediction

Higher education institutional data are increasingly used for research and applications to predict and explain factors contributing to student dropout [51, 71]. These models and applications, often in the form of early warning systems, can help identify which students might be struggling and at what time in their academic journey. This information can be used to provide proactive, targeted support or interventions. Several previous studies have investigated the task of dropout prediction in higher education [4, 6, 10–12, 17, 24, 25, 40, 42, 57, 76]. These studies are conducted across different institutional contexts, but the core learning problem they address is framed as a binary classification task, where structured features about students' educational history, demographics, or academic records in the early phase of college are extracted from administrative data to predict dropout/persistence at a later stage. Learned models are evaluated using binary classification metrics and state-of-the-art models using students' pre-college characteristics and early college records can predict whether a student will drop out within the first year with an AUC-ROC between 0.7 and 0.9 across various countries and institutions [4, 11, 12, 17]. These promising results have continued to motivate research into dropout

[3]https://unizin.org/

prediction modeling to aid student support services and institutional policy-making. However, there has been little research that empirically examines cross-institutional dropout prediction.

2.2 Disparate Impacts and Fairness-Performance Tradeoffs in Machine Learning

A great deal of prior work has explored both empirical and algorithmic approaches to "fairness" in machine learning, which is often concerned with how an objective function or conditional risk estimate varies across subgroups [19]. The current study is most closely related to previous works that explore the disparate impact of machine learning methods, and the interaction between fairness and predictive performance in machine learning.

Recent work finds that different learning techniques and contexts can result in a disproportionate impact on subgroups, even when these techniques are not explicitly targeted toward subgroups in any way, and even when they improve some average measure(s) of performance. Disproportionate impacts have been demonstrated in the use of differential privacy [7], model compression [38], model simplification [50], selective classification [43], synthetic data generation [18], under the presence of feature noise [48], and in repeated (i.e. multi-round) loss minimization [36].

There have been some formal analysis of potential tradeoffs between fairness and various measures of model performance [23, 29, 76], but our theoretical understanding remains limited in many areas. Empirically, however, there is some evidence that there are *no* strict tradeoffs between model performance and fairness under certain conditions. For example, Rodolfa et al. [64] finds that fairness-enhancing interventions across policy programs in education, mental health, criminal justice, and housing safety improved fairness with negligible effects on model accuracy. In education, a recent study found no evidence of a strict tradeoff between model performance and fairness predicting dropout from massive open online courses Gardner et al. [33]. However, the degree to which these findings apply to university student dropout or retention models is unknown.

Research on algorithmic fairness in education has mostly investigated whether supervised learning models trained on the entire student population generate systematically biased predictions of individual outcomes such as correct answers[27], course grades [77], university [76] and MOOC dropout [33], learned representations of student writing [5, 56], and graduation [40, 52]; it has also explored algorithms for at-risk prediction under fairness constraints [39]. Overall, this area of research is nascent and in need of systematic frameworks specific to educational contexts to map an agenda for future research. It does suggest, however, that analyses of novel learning paradigms (such as cross-institutional transfer) should include thorough auditing for fairness, a goal of the current study.

2.3 Cross-Institutional Learning

We introduce the term *cross-institutional* learning to describe the context where data is partitioned across a set of institutions by observation (i.e., each institution has a set of records with the same features, but collectively training on all institutions' centralized data is not possible). Prior work related to cross-institutional modeling has used a variety of different monikers, including "cross-silo" [44], "horizontal" [75], "collaborative" [67], and transfer learning.

Recent advances in both tooling and theory have, in principle, enabled improved access to cross-institutional training. In the past five years, several usable open-source frameworks for distributed, decentralized machine learning[4] as well as frameworks for privacy-preserving learning[5] have emerged. Theoretical advances, such as various approaches to differentially-private model training (DP-SGD [1], PATE [62]) have provided provable guarantees regarding non-identifiability of model training examples, which may reduce privacy concerns related to cross-institutional collaboration and further pave the way for cross-institutional training in practice. The development and refinement of techniques such as federated learning and secure multi-party computation or homomorphic encryption has also enabled distributed training with improved privacy and security [44].

Prior work has also investigated the related theoretical problem of learning fair models when the test distribution differs from the training distribution, as may be the case in cross-institutional learning. For example, Cotter et al. [21] studies the use of constrained optimization to improve the satisfaction of a fairness constraint on a held-out dataset with a possibly different distribution from the training set, Coston et al. [20] evaluates mitigating unfairness on a target domain due to covariate shift when sensitive attributes are unknown, and Singh et al. [69] explicitly investigates fairness under distribution shift. Algorithms which are "fairness-preserving" have been shown to be sensitive to variations in random train-test splits [32], suggesting that the fairness properties of models developed using such algorithms are brittle across distributions.

Applied research on cross-institutional learning has been somewhat limited. In the medical domain, cross-institutional learning has been used for brain tumor segmentation [67, 68], diabetic retinopathy diagnosis, and mammography screening [16]. For example, in the context of a tumor segmentation model, direct transfer leads to average performance degradation at varying levels for 9 of 10 institutions evaluated, while collaborative learning improves performance and performs similarly to data-sharing, depending on the approach used [67] and can be similar to the performance of centralized models in simulated settings [59]. Chang et al. [16] evaluates several transfer scenarios (local, ensembling via prediction averaging, single weight transfer, cyclical weight transfer) and finds that ensembling and weight transfer both outperformed local models in terms of validation and testing accuracy. Pessach et al. [63] looks into the task of collaboratively training fair models across institutions through a preprocessing mechanism which leads to fairness improvement. However, the impact of the intervention on individual institutions' models and institutional subpopulations is unclear.

In the domain of education, there have been some research efforts that formulate and empirically examine the issue of model transferability across instructional, institutional, and even societal contexts [13, 26, 42, 53, 55, 61]. For example, Ocumpaugh et al. [61] found that models detecting students' affective states in tutoring

[4]e.g. TensorFlow Federated https://www.tensorflow.org/federated, IBM Federated Learning https://ibmfl.mybluemix.net/

[5]TensorFlow Privacy https://github.com/tensorflow/privacy, Opacus https://opacus.ai/

systems do not transfer well when trained on one student population and tested on another, especially when rural students are the target (test) population. Similarly, Li et al. [55] investigated whether academic achievement prediction models trained on U.S. samples can generalize to other national contexts, and found that the performance drops significantly for less developed countries. Most closely related to the current study, Jayaprakash et al. [42] trained an early alert system of academically at-risk students at a liberal arts college and applied the model to four partner institutions with different institutional profiles. They found that the predictive performance declined but was still practically useful – the recall of at-risk students only dropped from 85% at the source institution to between 61% to 84% (an average of 75%) at the target institutions. These efforts show the promise of cross-institutional educational models especially when low-resourced institutions cannot afford to develop their own model, but still require more research to ensure that performance does not degrade across institutions, harming vulnerable students.

3 PRESENT STUDY AND RESEARCH QUESTIONS

Under the Family Educational Rights and Privacy Act (FERPA), U.S. higher education institutions are required to maintain records about students and enrollments for purposes of external reporting (e.g., to federal educational authorities) and internal improvement. Local student information systems (SIS) are widely used to manage these records and can facilitate the identification of students who are at risk of failing classes, not graduating on time, or dropping out (see Section 2.1). Due to shared reporting responsibilities, common operational routines, and similar software tools, institutions tend to have many overlapping features in their SIS data (e.g., students' course enrollments and demographic characteristics).

Our study leverages this commonality across four universities in order to explore the impacts of cross-institutional educational modeling. As discussed in Section 2.3 above, the limited prior work on cross-institutional transfer learning has suggested that direct transfer of learned models across institutions tends to degrade performance, and that only specialized weight-sharing strategies allow institutions to realize performance gains from transfer learning. In addition, little research has evaluated the fairness implications for such transfer scenarios. Building on these previous insights, our research addresses the following three research questions within the domain of dropout modeling in higher education:

RQ1. How does cross-institutional transfer (direct, voting, and stacked) affect performance relative to a local model?

RQ2. How does cross-institutional transfer affect the (intersectional) fairness of the resulting model?

RQ3. Is there a tradeoff between model performance and fairness under cross-institutional transfer?

4 METHODS

4.1 Data and Preprocessing

We use (de-identified) data from four public universities in the United States. All data is for first-time, first-year students in four-year bachelor's programs. Our dataset represents a wide range of enrollment sizes, demographic compositions, and first-year retention rates as summarized in Table 1.

We study the effects of cross-institutional transfer by converting the raw student information system (SIS) data obtained from each institution into a shared schema. Due to restrictions on data sharing, each institution's data was preprocessed separately and then validated by a shared pipeline prior to modeling. Only the learned model weights, not the data nor any intermediate artifacts (such as gradients during training), were shared outside of each institution. A goal of this project is to use SIS data in a form as close to its raw format as possible (i.e. minimal additional feature engineering), while also retaining the maximum number of viable features for experiments. In practice, this required balancing (i) removing features when insufficient data was available for all institutions or operationalization of variables was irreconcilable, and (ii) identifying ways to map related but non-identical features at each institution into a common semantic space. The process of defining a shared schema and processing the raw exports of each institution's SIS required domain expertise as well as familiarity with each institutional dataset.

The full schema produced by all institutions for our analysis is described in Table 2. Each row in the dataset represents a student enrolled in the fall term. The features describe students' academic history, demographics, current course load and course topics, and future plans (e.g., majors and minors). While the classification of gender as binary and the specific ethnic and racial groups raises concerns, we rely on the student categories used by the institutions themselves, which are shaped by federal reporting requirements. We provide further details on the schema in Section A.1. We release our code to validate cross-institutional datasets for conformity to this schema, and to replicate experiments using these features.[6]

4.2 Task

Our target prediction is first-year *retention*: for each student who enters an institution for the first time in the fall, we predict a binary indicator for whether that student will enroll at the same institution the following fall. This target matches the National Student Clearinghouse's definition [7], and is widely used both in research (Section 2.1) and practice in education.

We embrace the data constraints faced by educational institutions, which can limit the applicability of some previously proposed techniques for transfer learning. Federal regulation to protect student data privacy (FERPA) creates challenges for data sharing: costs associated with determining whether data may be shared, and then facilitating the sharing may be intractable for many institutions. We, therefore, do not consider techniques, such as federated learning, which require collaborative training in any form. For similar reasons, we do not consider approaches that require data sharing, for example, for training a centralized model on a joint dataset. Instead, we evaluate the realistic setting where each institution can only share *model weights* and only a single round of cross-institutional weight sharing is possible. In our experiments, the barriers between institutional datasets are real, as are the challenges

[6] https://github.com/educational-technology-collective/cross-institutional-transfer-learning-facct-2023

[7] https://nscresearchcenter.org/persistence-retention/

Institution	N_{train}	Female	URM	Hispanic	Asian	Black	Native Amer.	Two or more	White	First-year Retention
A	20k	49%	40%	26.7%	7.1%	5.1%	1.5%	5.1%	53.3%	80%
B	1k	63%	24%	12.6%	2.9%	5.7%	0.9%	4.3%	72.8%	57%
C	20k	57%	36%	33.1%	47.9%	3.2%	0.0%	0.0%	12.9%	94%
D	30k	55%	14%	6%	11.9%	5.3%	0.0%	4%	67.5%	98%

Table 1: Summary statistics for the training dataset for each institution showing demographic characteristics according to federal reporting requirements (URM: underrepresented racial minority; Two or more: multiple racial/ethnic groups).

of cross-institutional transfer. Contrast this with prior work on cross-institutional transfer or fairness under domain shift discussed in Section 2.3, which frequently simulated different "institutions" or "domains" by synthetically partitioning a single dataset.

4.3 Three Approaches to Cross-Institutional Transfer

Institutions in almost every domain, and particularly in higher education, differ in (i) data capacity, and (ii) modeling capacity. As a result, institutions vary in how they may be able to develop or utilize cross-institutional models. We define three distinct approaches to cross-institutional transfer, intending to cover three common contexts where institutions may seek to utilize cross-institutional models. We term these *direct transfer*, *voting transfer*, and *stacked transfer*. As a baseline for comparison, we consider a *local* model trained at the same institution where it is tested.

We consider a dataset $\mathcal{D}_k \coloneqq X_k, Y_k = (x_i, y_i)_{i=1}^n \sim \mathcal{P}_k$ of i.i.d. observations drawn from distribution $\mathcal{P}_k$, where $|x_i| = d$ features are present and k represents an institution of interest. Denote $f(\theta, \cdot)$ as a model with parameters θ, where $f(\theta, x_j) = \mathbb{P}(y_j = 1|x_j)$ is the model's predicted probability that x_j has label 1, noting that j indicates a potentially different institution of interest, and thus x_j may come from a distribution $\mathcal{P}_j$ which is different from $\mathcal{P}_k$. Denote the parameters estimated by training f on $\mathcal{D}_k$ as $\hat{\theta}(\mathcal{D}_k)$. Denote the loss (for some general loss function) of a model trained on distribution k and evaluated on distribution ℓ as $\mathcal{L}\left(f(\hat{\theta}(\mathcal{D}_k), \tilde{X}_\ell), \tilde{Y}_\ell\right)$, where $\tilde{\mathcal{D}}_k \coloneqq \tilde{X}_k, \tilde{Y}_k$ indicates the *test* split from a distribution k. We use $\mathcal{I}$ to refer to the set of all institutions.

We define each of the transfer learning approaches used in our experiments as follows:

Local: A local model is one trained on the same institution from which it is evaluated. That is, for institution k, the performance of a local model is defined by $\mathcal{L}\left(f(\hat{\theta}(\mathcal{D}_k), \tilde{X}_k), \tilde{Y}_k\right)$.

Direct Transfer: A direct transfer scenario is one where a model is to be deployed to an institution different from its testing institution. That is, for institutions k, j, the direct transfer model performance is measured by $\mathcal{L}\left(f(\hat{\theta}(\mathcal{D}_k), \tilde{X}_j), \tilde{Y}_j\right)$. We refer to k as the *source* institution and j as the *target* institution, following the domain transfer literature. A single trained source model can be evaluated via direct transfer on several target institutions.

Voting Transfer: This training paradigm uses a form of averaging to combine the results of models ("voters") trained on disjoint distributions. In our experiments, *none of the voters are trained on the target institution*, which mimics the case where an institution without any historical training data uses a set of models from other institutions in a zero-shot scenario. The model under the voting transfer paradigm for target institution i is defined by $\frac{1}{c}\sum_{i' \in \mathcal{I} \setminus i} f(\hat{\theta}(\mathcal{D}_{i'}), \cdot)$, where c is the normalizing constant $|\mathcal{I}| - 1$. Note that this model does not use majority voting, but instead uses "soft voting," where the predicted probabilities (not the decisions) of each model are aggregated with equal weight. This allows for the confidence of each model to be taken into account in the aggregation.[8]

Stacked Transfer: This training paradigm uses *stacked generalization* [70, 73] to combine the predictions of models trained on all available institutions with the training data of the source institution. This is achieved by concatenating, for each input x_i, the predictions of each classifier $f(\hat{\theta}(X_j), x_i)$, to the input features, and learning a classifier from this concatenated data matrix. Formally, for institution ℓ, if we define

$$\check{x} = [x_1, ..., x_d; f(\hat{\theta}(\mathcal{D}_i, x); f(\hat{\theta}(\mathcal{D}_j, x); f(\hat{\theta}(\mathcal{D}_k, x)] \quad (1)$$

where $[\cdot;\cdot]$ indicates column-wise concatenation, then the stacked estimator is $f(\hat{\theta}(\check{X}))$.

Because the final two forms of transfer (voting transfer, stacked transfer) are both methods for ensembling, we refer to these two methods collectively as ensemble models.

4.4 Metrics

Metrics for evaluating various aspects of model fairness have been proposed in prior work (see Section 2). However, many of these metrics are based on the implicit or explicit assumption that one outcome is *advantageous* or *favorable*, and that a "fair" model can ensure some form of equity with respect to the model's predictions in placing members of sensitive subgroups into the favorable outcome class. This is often tied to contexts in which the model's predictions may be explicitly tied to some form of decision (e.g., granting a loan). Our task differs from these contexts, because the model is not explicitly tied to a decision but instead provides a prediction that might be used to assist a student, but is only useful when correct—neither predictive outcome is considered inherently "advantageous." In our application, the goal of the model is to obtain equal *predictive performance* for all subgroups, regardless of the true or predicted outcome. We call this *equitable predictive performance*.

[8]The choice of equal weighting is by convention; in practice, any combination of weights on the $(\mathcal{I} - 1)$-simplex could be used to aggregate the predictions of the voters. This weight vector could also be tuned on the target institution in a non-zero-shot formulation.

This makes many existing fairness metrics, such as demographic parity, which assume the presence of an advantageous outcome to which we would like to equitably assign predictions, not applicable to the dropout prediction task. There are many tasks where equitable predictive performance is desired, such as machine translation, where consistent performance is desired across dialects or languages despite differing availability of training data [2], and image classification with respect to skin tone or other attributes [15].

Educational dropout data tends to be highly skewed by label, because in many institutional contexts, the majority of students do not drop out. This is the case in our data as well: the retention rate in our universities varies from 56% (Institution B) to 98% (Institution D; see Table 1). As a result, metrics such as average accuracy will tend to be biased toward majority-class predictors and will be uninformative for the small but critical subset of students who drop out.

Area Under the Receiver Operating Characteristic Curve (AUC): Due to our goal of equitable prediction and the significant label imbalance in our datasets, our experiments use metrics based on the Area Under the Receiver Operating Characteristic Curve (AUC), formally defined as:

$$AUC(f(\theta)) = \int_0^1 \text{TPR}(\text{FPR}(f_t(\theta)))dt \tag{2}$$

where t indicates a prediction threshold applied to the predictions of the model (i.e. using the decision rule $f_t(\theta, x) = \mathbb{1}\left(f(\theta, x) \geq t\right)$), and TPR, FPR are the true positive rate and false positive rates, respectively. AUC scores are constrained to $[0, 1]$, with a random predictor achieving an AUC of 0.5. In all metrics and experiments, we compute the AUC on the test dataset, following the splitting process described in Section A.2.

AUC is a well-studied metric of predictive performance [14, 35, 41], and has the straightforward interpretation as the probability that a randomly-selected positive example has a higher predicted probability of being positive than a randomly-selected negative example. This means that the positive and negative classes are equally weighted in computing AUC, and that AUC is less rewarding to, for example, majority-class predictors than metrics such as accuracy or cross-entropy loss. We compute standard errors for AUC values according to the procedure described in [30, 35]. We provide details on computing these standard errors in Section D.

AUC Gap: To measure fairness across subgroups, we define a metric that accounts for the disparities in predictive performance across a set of arbitrarily many (possibly-overlapping) subgroups $\mathcal{G}$. We define the AUC Gap as:

$$\max_{g,g' \in \mathcal{G}} |\mathbb{E}_{\mathcal{D}_k}[f(\theta(\mathcal{D}_{k,g})] - \mathbb{E}_{\mathcal{D}_k}[f(\theta(\mathcal{D}_{k,g'})]| \tag{3}$$

where $\mathcal{D}_{k,g}$ and $\mathcal{D}_{k,g'}$ indicate the subset of the data in group g and g', respectively. Thus, AUC Gap measures the largest difference between subpopulation AUCs, and is a measure of the worst-case performance gap between a set of subpopulations. AUC Gap is our primary measure of equitable predictive performance, because it quantifies the largest disparity in predictive performance over subgroups.

ΔAUC: We define ΔAUC to measure changes in predictive performance or fairness under cross-institutional transfer. We define the change in AUC between two transfer contexts for a fixed model f, as $\Delta AUC(T, T') = AUC(T) - AUC(T')$ where, somewhat abusing notation, we use $AUC(T)$ to refer to the AUC of a model trained using a transfer scheme T. This allows us to compare, for example, how (overall or subgroup) AUC values are affected by transfer learning schemes. If ΔAUC is close to zero, we can conclude that the model performs about the same in two transfer contexts; but if it is positive or negative, the model performs better/worse in context T' relative to T. Most often in this study, we are interested in ΔAUC relative to the local model; that is, ΔAUC(local, ·).

5 EXPERIMENTAL SETUP

Parameterization and Tuning of f: We evaluate three forms of model transfer learning (*direct transfer*, *voting transfer* and *stacked transfer*, described previously) plus *local* models on our four institutional datasets. For each experiment, we explore three parameterizations of f: (1) L_2-regularized logistic regression (L2LR); (2) gradient-boosted trees (LightGBM [47]); (3) neural networks (multilayer perceptrons). Below, we primarily focus on L2LR in the main text due to space constraints, because (1) we observed the best performance for L2LR models across transfer schemes; (2) L2LR models are the simplest to train and tune, even for institutions with low capacity for data science and modeling; and (3) L2LR models are highly interpretable and widely used for student retention models in practice (see Section 2.1). We provide the complete results for other models in the supplementary material, including a parallel version of each figure in the main text for the other models (LightGBM, MLP); our findings with L2LR are consistent with our findings for these other models, except where explicitly noted.

Hyperparameters of each model are tuned locally via cross-validation on the source institution, and complete hyperparameter grids for each model are provided in our associated source code. For stacked models, the stacked model's hyperparameters are also tuned via cross-validation. A subset of the training data is held out and used only as a validation set for model selection (see below). The trained models are always evaluated on the test data from each institution, which is from a future academic term (see Section A.2 for details).

As noted previously, *voting transfer* is a form of zero-shot transfer and demonstrates how an institution with no training data might make use of models from other institutions. Voting transfer experiments do not use the base model from the target institution and do not require any training on the target institution: they simply use weighted majority voting to aggregate the results of each trained source model. Voting transfer allows us to evaluate the zero-shot transfer of models to an institution with no training data available.

Model Selection Rules: For the *stacked transfer* experiments, we explore the use of all three functional forms mentioned above (i.e. the stacked model can be L2LR, LightGBM, or MLP). However, it is not clear how the choice of base learners used to construct the stacked ensemble affects the (accuracy, fairness) of the downstream stacked model, particularly when multiple base models are available from each institution – a realistic scenario, as it is not uncommon for an institution to train and evaluate many models during the development phase. Furthermore, it is possible that a "greedy" approach to selecting the base models may have unintended consequences

for performance (for example, the best-performing model at Institution A may lead to poor performance when used in an ensemble at Institution C). We therefore conduct a principled investigation comparing four strategies for selecting which trained models to include in the stacking process, as follows. First, in addition to training the cross-validated models tuned at each institution, we train 10 additional models of each type (L2LR, LightGBM, MLP) with a fixed set of hyperparameters, varying only the L_2 regularization of each model over a large grid. Second, from this pool of 11 candidate models at each institution (one cross-validated model plus 10 models with various degrees of regularization), we apply one of four *model selection rules* to choose which estimators are included to construct (1). Finally, we fit the stacked model and evaluate it on the target institution (the base learners are frozen throughout this process). We explore the following model selection rules:

- *Best Performance*: only the model with the best validation AUC at each institution is used.
- *Best Fairness*: only the model with the best validation AUC Gap (Equation (3)) is used.
- *Same Family*: the cross-validated models of the same functional form are used (e.g. LightGBM).
- *Kitchen Sink*: all 11 models are used.

Due to space constraints, our results in the main text reflect the *best performance* selection rule, as this is the approach most commonly used in practice. We provide additional results for all model selection rules in Section C.4 (e.g. Figure 6).

6 RESULTS

6.1 Overall Predictive Performance of Cross-Institutional Transfer Models

In RQ1 we are concerned with measuring model performance across three different transfer schemes (direct, voting, and stacked). Figure 2a shows the AUC for models using each transfer scheme, as well as the performance of direct transfer of models from a given source institution to a target institution, for each of the four institutions in our study. ΔAUC values for each transfer scenario are shown in Figure 2b.

Direct Transfer Models: Our results in Figure 2 show that direct transfer has inconsistent performance: for two of four institutions (B and D), all direct transfer models achieve indistinguishable performance from local models (as indicated by $\Delta\text{AUC} = 0$ confidence intervals covering zero in Figure 2b for direct transfer models); for the remaining two institutions (A and D), the results are mixed. This suggests, perhaps unsurprisingly, that direct transfer of models may sometimes achieve good overall performance (compared to a local model), but not in all cases.

Ensemble Models (Voting Transfer and Stacked Transfer): Figure 2 shows results with respect to voting and stacked transfer models. These results for both voting transfer and stacked transfer are consistent across all institutions in our study.

In our experiments, zero-shot voting transfer achieves similar performance to local models (as indicated by $\Delta\text{AUC} = 0$ confidence intervals covering zero in Figure 2b for voting transfer models). These results suggest that all institutions can obtain models with equivalent performance to a local model by performing zero-shot weighted aggregation of a set of models trained only on other institutions.

Furthermore, stacked transfer provides no additional benefit over local models (or zero-shot models). Figure 2a shows that stacked transfer models do not improve over either local models or voting transfer models (as indicated by overlapping confidence intervals for AUC between stacked and local/voting models in Figure 2a). This suggests that once institutions have leveraged *either* their own local training data (local model) *or* other institutions' models (voting transfer), combining these information sources (via stacked model) provides no additional performance gains in our experiments.

A z-test confirms that $\Delta\text{AUC}(\text{local}, \text{voting})$ and $\Delta\text{AUC}(\text{local}, \text{stacked})$ are statistically indistinguishable from zero at $\alpha = 0.05$ (all $p > 0.1$), suggesting that the voting ensemble method is an effective way to reduce uncertainty over which direct transfer model should be used when no local training data is available, but that stacking provides no additional performance gains. We provide exact p-values for the hypothesis test that $H_0 : \Delta\text{AUC} \neq 0$ in supplementary Table 3. This suggests, in particular, that zero-shot transfer of a voting ensemble of three other institutions can achieve performance statistically indistinguishable from a locally-trained model.

6.2 Intersectional Fairness Analysis of Cross-Institutional Transfer Models

Our fairness analysis evaluates whether transferred models achieve equivalent predictive performance over sensitive subgroups. A substantial body of work across many disciplines and dating back several decades delineates how the *intersections* of individuals' identities can contribute to disempowerment and increase vulnerability to adverse outcomes [22, 28, 72]. In machine learning, however, most prior work on fairness focused on analyzing one sensitive attribute at a time (notable exceptions include [31, 60, 74]), despite the frequent presence of multiple potentially sensitive attributes in a dataset. Therefore, we measure fairness across *intersectional* subgroups via AUC Gap defined in Section 4.4. For sensitive attributes $a \in \mathcal{A}_1, a' \in \mathcal{A}_2$, we compute each metric $\mathcal{L}$ on the subset $\mathcal{D}_{a,a'} := (x_i, y_i | A_1(x) = a, A_2(x) = a')$. Each metric is therefore computed as $\mathcal{L}\big(f(\hat{\theta}(\mathcal{D}), \tilde{X}_{a,a'}), \tilde{y}_{a,a'}\big)$. By evaluating fairness in this way, our analysis captures whether changes in AUC are distributed equally over (observable) intersecting student identities, "focus[ing] awareness on people and experiences—hence, on social forces and dynamics—that, in monocular vision, are overlooked" [58].

We specifically use this approach to examine subgroups of students defined by sex and URM status. These represent two critical identities in the context of education with respect to which unfairness is undesirable but common in educational settings. We compute evaluation metrics for intersections of $\mathcal{A}_1 = \{\text{male}, \text{female}\}$ for Sex and $\mathcal{A}_2 = \{\text{Underrepresented Minority}, \text{Non-Underrepresented Minority}\}$ for URM.[9] The main results of our fairness analysis are shown in Figures 3 and 4; we also provide detailed data in supplementary

[9]Most institutions recorded more than two categories for Sex, but these tended to be "other" and "not indicated", and besides making up a small share of students, it was unclear how these responses were collected in order to interpret them correctly and consistently. We do not endorse the terminology or definition of URM; we only use it because it is consistently defined across institutions to abide by federal regulations.

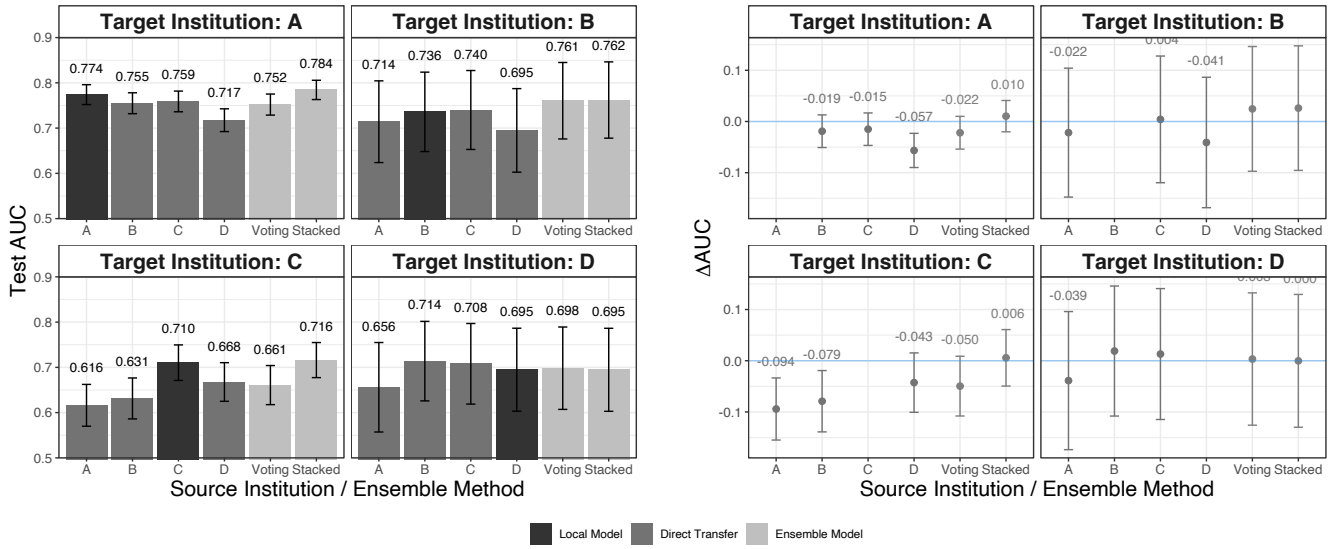


Figure 2: (a) Left: Predictive performance on test data for various transfer schemes evaluated. (b) Right: ΔAUC values for various transfer models evaluated with the *local* model reference line. 95% confidence intervals shown for both figures; text displays values for point estimates. (See also Figure 8 and Table 3).

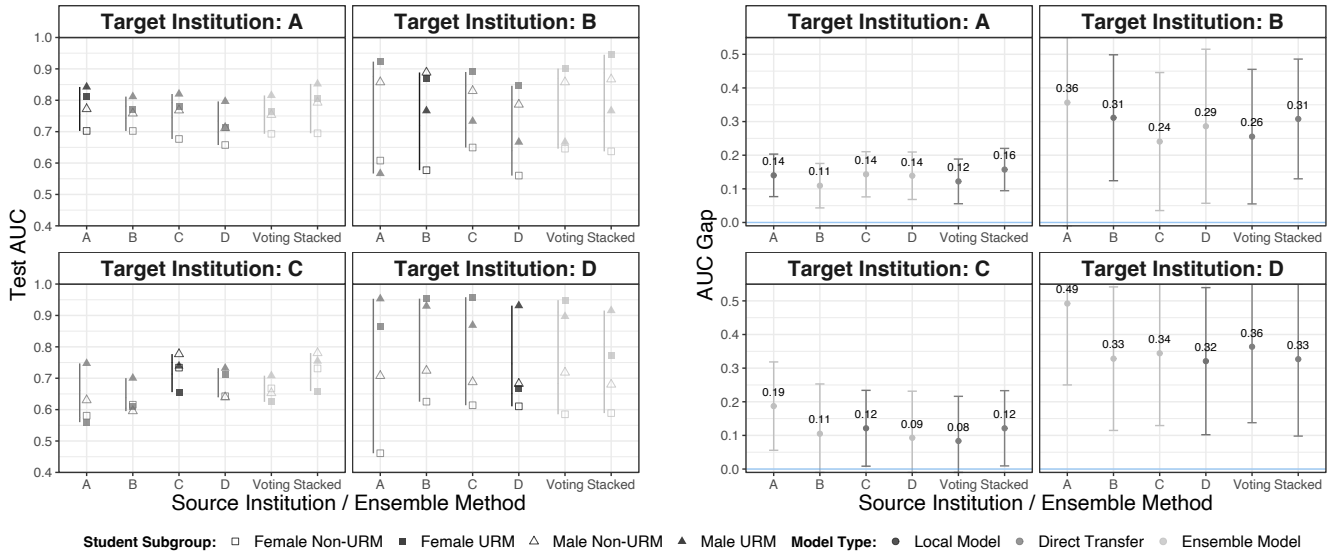


Figure 3: (a) Left: Intersectional performance over all sensitive subgroups for various transfer schemes evaluated. (b) Right: Fairness metric *AUC gap* over intersectional sensitive subgroups for various transfer schemes evaluated. 95% confidence intervals shown; text displays values for point estimates. Note that the AUC gap (b) is the range of these values within each institution (shown as vertical bar in (a)). (See also Figures 9a, 9b, Table 4.)

Table 4. Below, we address RQ2 by separately discussing these disparity metrics under each transfer approach.

Before reporting the results, we raise two issues related to fairness, accountability, and transparency. First, the data we use is what institutions make available, and does not encapsulate all potential identities related to the sex or gender construct, and thus there is *measurement bias* inherent in the data. Second, AUC (and related metrics, such as F1 score) cannot be computed unless there is at least one positively-labeled and negatively-labeled observation in each subgroup, and our data did not include any other intersectional

groups where this precondition was not satisfied, and thus fairness for other measures was not evaluated, a form of *algorithmic bias*.

Local and Direct Transfer Models: Figure 3a presents the four intersectional subgroup AUCs underlying the computation of each AUC gap, which is summarized in Figure 3b. The results in Figure 3 show that local and direct transfer models do not differ in their performance disparities over subgroups as measured by AUC gap (Equation (3)). There is no consistent difference between the subgroup performance disparities of local vs. direct transfer models in our study, and in no case can we reject $H_0 : \Delta\text{AUC}(\text{local}, \text{direct}) = 0$. Figure 3a also shows that a single intersectional subgroup often drives the observed performance disparities, which persist across different source models: for 17 of 24 transfer schemes evaluated, the Female Non-URM group is the lowest-AUC group for the model.

Ensemble Models (Voting Transfer and Stacked Transfer): In general, ensemble models achieve similar fairness relative to a local model across our experiments. Figure 5b shows that, for all institutions, both voting transfer and stacked transfer achieve confidence intervals for ΔAUC that overlap with the local model. Particularly for the voting transfer model, this is an encouraging result: It suggests that the practical benefits from the use of zero-shot voting transfer (achieving performance equivalent to a local model, *without* having local training data; see Section 6.1) do not come at a cost to fairness, an important result. It also suggests that the improvements of voting transfer over direct transfer models do not benefit only one group; instead, the gap between the min and max AUC groups stays the same while the overall AUC improves under voting transfer (relative to direct transfer).

For 20 of 24 transfer schemes evaluated, we reject $H_0 : \Delta\text{AUC} = 0$ (see Table 4 for details). This means that for most models, there is a nonzero gap between the best- and worst-performing groups for the learned model. This suggests that future work is needed to improve performance for some intersectional groups if equivalent performance across groups is possible and desirable.

6.3 Exploratory Findings

This section discusses exploratory findings from our study. Our large-scale empirical study, being the first of its kind in the domain of higher education, is uniquely positioned to provide empirical insight into several questions of interest. However, as a purely observational study with a limited set of institutions, it is our intention to clearly position our discussion of the following findings as *exploratory*: our study provides initial evidence, but does not *prove* there is a relationship, particularly a causal relationship, between the factors discussed here.

No tradeoff between fairness and accuracy: RQ3 in our study concerns whether there is a tradeoff between fairness and accuracy in cross-institutional transfer. Our results suggest that the variation in AUC Gap is epxlained by other factors (namely, institution and the transfer type), and that AUC Gap is *not* associated with AUC after controlling for these factors. We explore a simple linear regression of AUC Gap on AUC, with terms for the target institution and the transfer type. We find that in the resulting linear model, the AUC term β_{AUC} has $t = -0.350$ ($p = 0.72719$), suggesting that we do *not* have evidence to suggest that AUC and AUC Gap are associated, after controlling for the target institution and transfer type. We show a scatter plot of the data used to conduct this analysis in supplementary Figure 11, and provide details on the model, in Section C.

Impact of Intersectional Analysis: In Section C.3, we briefly compare the findings of our intersectional analysis with a non-intersectional ("marginal") analysis. This comparison demonstrates that marginal analyses are more likely to ignore performance disparities within subgroups, and to assign lower overall AUC Gap scores to models.

No clear impact of regularization: We also study whether effective regularization helps reduce subgroup performance disparities. One potential interpretation of the disparities measured by AUC Gap might be that models simply overfit to certain groups; in this case, effectively tuning the regularization parameter might reduce the degree of overfitting to certain groups. To investigate the impact of regularization, we conduct a sweep of the L_2 regularization for all models (L2LR, LightGBM, MLP; all contain an L_2 regularization term) and keep all other hyperparameters at default values. We provide the results of this study in Figure 10, and give further detail on the design of these experiments in Section C.6. Our results suggest that there is not a clear relationship between regularization and AUC Gap. This aligns with existing work on subgroup robustness, which suggests that subgroup performance tends to improve along with the overall model (in which case regularization should be tuned to optimize the bias-variance tradeoff) [**?**].

7 CONCLUSION

This paper presents the first large-scale empirical study of model performance and fairness in cross-institutional transfer learning. We proposed a set of metrics for quantifying cross-institutional transfer performance and fairness, and applied those techniques in the context of university student dropout prediction with real-world education data. Our results show that cross-institutional transfer is possible, where even zero-shot "voting transfer" models achieve statistically indistinguishable performance to a local model with no change in intersectional subgroup fairness. Additionally, our results show more broadly that there is no evidence of a performance-fairness tradeoff across a wide scope of functional forms (L2LR, LightGBM, MLP), transfer schemes (direct, voting, stacked), and selection rules for the ensemble components (best performance, best fairness, same family, kitchen sink).

These findings have important implications for both researchers and practitioners. For machine learning researchers, the demonstrated success of relatively simple approaches (e.g. voting transfer) suggests that further investigation is needed to understand the conditions under which (i) zero-shot transfer is effective, and (ii) more sophisticated transfer learning and ensembling methods succeed or fail. For educational researchers, the results suggest that while institutional contexts matter in understanding and predicting student dropout, there exists a decent level of generalizable knowledge that can facilitate the development of portable predictive models. For practitioners, our results show that the cross-institutional learning paradigm can serve as a viable means for well-resourced institutions

to support their underresourced counterparts. This is especially relevant in an era with growing availability of big data and increasing prevalence of algorithmic decision making.

On the other hand, this study has a handful of limitations. First, our models are trained on available data from each institution, which could contain their own biases [9, 49]. Treating institutional datasets as a source of truth masks complexities in how variables are coded and in how historical inequities are manifested in the data, which could affect the applicability of our conclusions. Second, our sample only includes four institutions in the United States that have the data infrastructure and capacity to make large-scale data available for research. This limits the generalizability of our results to other institutions with varying degrees of similarity in student populations and dropout-generating processes. The breadth of cultural perspectives in non-U.S. contexts, as well as the definition of which students are under-represented, also suggest a need for replication and extension of this work.

Informed by the limitations, there are a few lines of future work. First, further studies on cross-institutional transfer are needed, including similar studies in education and other decentralized organizations with large-scale shared electronic record-keeping systems (e.g., hospitals, local governments, financial institutions). Future work should evaluate additional transfer approaches and could include the development of algorithms designed explicitly to mitigate performance disparities. Second, cross-institutional collaborative modeling is still difficult, due to a combination of data-sharing restrictions and a lack of technical infrastructure for cross-institutional collaboration with private data. We encourage the development of better technical and theoretical frameworks for collaborative learning in the presence of strict data-sharing constraints. Finally, our work suggests that overall performance can be improved without a strict cost to fairness, providing a motivation for further improvement of general classification techniques for student retention modeling, even without explicit disparity-mitigating interventions.

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.

[2] Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. 2021. The Low-Resource Double Bind: An Empirical Study of Pruning for Low-Resource Machine Translation. *arXiv preprint arXiv:2110.03036* (2021).

[3] Izzat Alsmadi, ZW Taylor, and Joshua Childs. 2020. US News & World Report Best Colleges rankings: Which institutional metrics contribute to sustained stratification? *Scientometrics* 124, 3 (2020), 1851–1869.

[4] 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 on Conference on Information and Knowledge Management (CIKM '16)*. Association for Computing Machinery, New York, NY, USA, 903–912. https://doi.org/10.1145/2983323.2983351

[5] Noah Arthurs and AJ Alvero. 2020. Whose Truth is the "Ground Truth"? College Admissions Essays and Bias in Word Vector Evaluation Methods. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*. 342–349.

[6] 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.

[7] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*. 15479–15488.

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

[9] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. http://www.fairmlbook.org

[10] 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. https://doi.org/10.1007/s11162-019-09546-y

[11] 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. https://doi.org/10.5281/ZENODO.3594771

[12] 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 (jan 2021), 233285842110376. https://doi.org/10.1177/23328584211037630

[13] Sebastien Boyer and Kalyan Veeramachaneni. 2015. Transfer Learning for Predictive Models in Massive Open Online Courses. In *Artificial Intelligence in Education (AIED 2015)*, Vol. 9112. Springer, Cham, 54–63. https://doi.org/10.1007/978-3-319-19773-9_6

[14] Andrew P Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition* 30, 7 (1997), 1145–1159.

[15] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[16] Ken Chang, Niranjan Balachandar, Carson Lam, Darvin Yi, James Brown, Andrew Beers, Bruce Rosen, Daniel L Rubin, and Jayashree Kalpathy-Cramer. 2018. Distributed deep learning networks among institutions for medical imaging. *Journal of the American Medical Informatics Association* 25, 8 (2018), 945–954.

[17] Yujing Chen, Aditya Johri, and Huzefa Rangwala. 2018. Running out of STEM: A Comparative Study across STEM Majors of College Students At-Risk of Dropping Out Early. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. ACM, 270–279. https://doi.org/10.1145/3170358

[18] Victoria Cheng, Vinith M Suriyakumar, Natalie Dullerud, Shalmali Joshi, and Marzyeh Ghassemi. 2021. Can You Fake It Until You Make It? Impacts of Differentially Private Synthetic Data on Downstream Classification Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 149–160.

[19] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[20] 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. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (2019). https://doi.org/10.1145/3306618

[21] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. 2019. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In *International Conference on Machine Learning*. PMLR, 1397–1405.

[22] Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.* (1989), 139.

[23] Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. 2019. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. 309–315.

[24] Gerben W. Dekker, Mykola Pechenizkiy, and Jan M. Vleeshouwers. 2009. Predicting students drop out: A case study. In *Proceedings of the 2nd International Conference on Educational Data Mining (EDM 2009)*. 41–50.

[25] Francesca Del Bonifro, Maurizio Gabbrielli, Giuseppe Lisanti, and Stefano Pio Zingaro. 2020. Student Dropout Prediction. In *Proceedings of the 21st International Conference on Artificial Intelligence in Education (AIED 2020)*. Springer, 129–140. https://doi.org/10.1007/978-3-030-52237-7_11

[26] Mucong Ding, Yanbang Wang, Erik Hemberg, and Una-May O'reilly. 2019. Transfer Learning using Representation Learning in Massive Open Online Courses. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. ACM, 145–154. https://doi.org/10.1145/3303772

[27] Shayan Doroudi and Emma Brunskill. 2019. Fairer but not fair enough on the equitability of knowledge tracing. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK '19)*. ACM, 335–339. https://doi.org/10.1145/3303772.3303838

[28] Zilah EISESTEIN. 1979. Combahee River collective: A Black Feminist Statement. *Capitalist Patriarchy and the Case for Socialist Feminism* (1979).

[29] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[30] James Fogarty, Ryan S Baker, and Scott E Hudson. 2005. Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction. In *Proceedings of Graphics Interface 2005*. 129–136.

[31] James Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2019. An Intersectional Definition of Fairness. arXiv:1807.08362 [cs.LG]

[32] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 329–338.

[33] 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.

[34]]gardnersubgroup Joshua P Gardner, Zoran Popovi, and Ludwig Schmidt. [n. d.]. Subgroup Robustness Grows On Trees: An Empirical Baseline Investigation. In *Advances in Neural Information Processing Systems*.

[35] James A Hanley and Barbara J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 1 (1982), 29–36.

[36] Tatsunori B Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. *arXiv preprint arXiv:1806.08010* (2018).

[37] Nicholas W Hillman, David A Tandberg, and Jacob PK Gross. 2014. Performance funding in higher education: Do financial incentives impact college completions? *The journal of higher education* 85, 6 (2014), 826–857.

[38] Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058* (2020).

[39] Qian Hu and Huzefa Rangwala. 2020. Towards Fair Educational Data Mining: A Case Study on Detecting At-risk Students. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*. 431–437.

[40] Stephen Hutt, Margo Gardner, Angela L. Duckworth, and Sidney K. D'Mello. 2019. Evaluating Fairness and Generalizability in Models Predicting On-Time Graduation from College Applications. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*.

[41] Nathalie Japkowicz and Mohak Shah. 2011. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press.

[42] Sandeep M. Jayaprakash, Erik W. Moody, Eitel J.M. Lauría, James R. Regan, and Joshua D. Baron. 2014. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. *Journal of Learning Analytics* 1, 1 (2014), 6–47. https://doi.org/10.18608/jla.2014.11.3

[43] Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. 2020. Selective Classification Can Magnify Disparities Across Groups. *arXiv preprint arXiv:2010.14134* (2020).

[44] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2019. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977* (2019).

[45] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 643–650.

[46] Mark Kantrowitz. 2021. Shocking Statistics About College Graduation Rates. https://www.forbes.com/sites/markkantrowitz/2021/11/18/shocking-statistics-about-college-graduation-rates/

[47] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).

[48] Fereshte Khani and Percy Liang. 2020. Feature Noise Induces Loss Discrepancy Across Groups. In *International Conference on Machine Learning*. PMLR, 5209–5219.

[49] René F. Kizilcec and Hansol Lee. 2022. Algorithmic fairness in education. In *The Ethics of Artificial Intelligence in Education*. Routledge, 174–202.

[50] Jon Kleinberg and Sendhil Mullainathan. 2019. Simplicity creates inequity: implications for fairness, stereotypes, and interpretability. In *Proceedings of the 2019 ACM Conference on Economics and Computation*. 807–808.

[51] George D. Kuh, Jillian Kinzie, Jennifer A. Buckley, Brian K. Bridges, and John C. Hayek. 2007. Piecing Together the Student success puzzle: Research, Propositions, and Recommendations. *ASHE Higher Education Report* 32, 5 (2007), 1–182. https://doi.org/10.1002/aehe.3205

[52] Catherine Kung and Renzhe Yu. 2020. Interpretable Models Do Not Compromise Accuracy or Fairness in Predicting College Success. In *Proceedings of the 7th ACM Conference on Learning @ Scale (L@S '20)*. Association for Computing Machinery (ACM), New York, NY, USA, 413–416. https://doi.org/10.1145/3386527.3406755

[53] Jarkko Lagus, Krista Longi, Arto Klami, Arto Hellas, J Lagus, K Longi, A Klami, and A Hellas. 2018. Transfer-Learning Methods in Programming Course Outcome Prediction. *ACM Transactions on Computing Education (TOCE)* 18, 4 (oct 2018). https://doi.org/10.1145/3152714

[54] Hansol Lee and René F. Kizilcec. 2020. Evaluation of Fairness Trade-offs in Predicting Student Success. arXiv:2007.00088 [cs.CY]

[55] Xingyu Li, Difan Song, Miaozhe Han, Yu Zhang, and René F Kizilcec. 2021. On the limits of algorithmic prediction across the globe. (2021). https://arxiv.org/abs/2103.15212

[56] Anastassia Loukina, Nitin Madnani, and Klaus Zechner. 2019. The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Florence, Italy, 1–10. https://doi.org/10.18653/v1/W19-4401

[57] Yuetian Luo and Zachary A. Pardos. 2018. Diagnosing University Student Subject Proficiency and Predicting Degree Completion in Vector Space. In *Proceedings of the Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. New Orleans, LA, USA.

[58] Catharine A MacKinnon. 2013. Intersectionality as method: A note. *Signs: Journal of Women in Culture and Society* 38, 4 (2013), 1019–1030.

[59] Julian Matschinske, Julian Späth, Reza Nasirigerdeh, Reihaneh Torkzadehmahani, Anne Hartebrodt, Balázs Orbán, Sándor Fejér, Olga Zolotareva, Mohammad Bakhtiari, Béla Bihari, et al. 2021. The FeatureCloud AI Store for Federated Learning in Biomedicine and Beyond. *arXiv preprint arXiv:2105.05734* (2021).

[60] Giulio Morina, Viktoriia Oliinyk, Julian Waton, Ines Marusic, and Konstantinos Georgatzis. 2019. Auditing and achieving intersectional fairness in classification problems. *arXiv preprint arXiv:1911.01468* (2019).

[61] 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. https://doi.org/10.1111/bjet.12156

[62] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. 2018. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908* (2018).

[63] Dana Pessach, Tamir Tassa, and Erez Shmueli. 2021. Fairness-Driven Private Collaborative Machine Learning. *arXiv preprint arXiv:2109.14376* (2021).

[64] Kit T Rodolfa, Hemank Lamba, and Rayid Ghani. 2021. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence* 3, 10 (2021), 896–904.

[65] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731* (2019).

[66] Mark Schneider and Lu Yin. 2011. *The High Cost of Low Graduation Rates: How Much Does Dropping Out of College Really Cost?* Technical Report.

[67] Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R Colen, et al. 2020. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports* 10, 1 (2020), 1–12.

[68] Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. 2018. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *International MICCAI Brainlesion Workshop*. Springer, 92–104.

[69] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. 2021. Fairness violations and mitigation under covariate shift. In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Inc, 3–13. https://doi.org/10.1145/3442188.3445865 arXiv:1911.00677

[70] Kai Ming Ting and Ian H Witten. 1997. Stacked Generalization: when does it work? (1997).

[71] Vincent Tinto. 1975. Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research* 45, 1 (mar 1975), 89–125. https://doi.org/10.3102/00346543045001089

[72] S. Truth. 1851. Ain't I A Woman? Speech delivered at Women's Rights Convention, Akron, Ohio.

[73] David H Wolpert. 1992. Stacked generalization. *Neural networks* 5, 2 (1992), 241–259.

[74] Forest Yang, Moustapha Cisse, and Sanmi Koyejo. 2020. Fairness with Overlapping Groups. *arXiv preprint arXiv:2006.13485* (2020).

[75] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 2 (2019), 1–19.

[76] 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*. 91–100.

[77] 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. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*. 292–301.

ACKNOWLEDGMENTS

This research was partly funded by a Google gift via the 2021 Award for Inclusion Research program. Josh Gardner was supported by a grant from Microsoft.

A DATA

A.1 Schema

The complete schema for the institutional data used in this study is shown in Table 2. As discussed in Section 4.1, the data is extracted from each institution's student information system (SIS), and is a subset of the data available at each institution. These variables were chosen based on a combination of availability, expected predictive utility (based on prior research discussed in Section 2.1 and the researchers' own experience), and consistency of coding across institutions.

The "CIP2 code" referenced in Table 2 refers to the Classification of Instructional Programs (CIP) codes defined by the Center for Educational Statistics[10]. CIP codes are written as decimal values (i.e. 14.43 is the CIP code for "Biochemical Engineering"). We use the "coarse" CIP codes, which are represented by the integer value preceding the decimal (i.e. 14 represents "Engineering"). A complete list of CIP codes may be viewed at https://nces.ed.gov/ipeds/cipcode/browse.

For features marked as "multicolumn" in Table 2, several identical features are generated to represent the information for that row. For example, for "Units Per Course Type", we generate 62 features, where each feature indicates the number of units associated with a given CIP code (1-61, or "Missing" when the CIP code is not available).

The "Retention" variable indicates the prediction target for this study, and is an indicator for whether the student was enrolled in the following fall term, according to the university's enrollment records. Note that this is only a 1-year measure of retention; it does not measure whether the student persists to complete a degree.

Additional categorical variables are coded as follows:

- **Course Component:** a variable indicating which "component" of a course an individual record is assigned to (note that courses may sometimes also consist of multiple components, such as a lecture and a lab component). These also include multiple-component courses, which are simply the set of all combinations of the following course components: Lecture, Discussion, Lab, Seminar, Other.
- **Sex:** a self-declared variable representing the students' declared sex or gender identity. The procedure for collecting this data, along with the exact allowed values, vary by institution; we include the following possible values, but recognize the limitations of such a coding: Male, Female", Not Indicated, Other.
- **Ethnicity:** a self-declared variable representing the students declared ethnicity. The procedure for collecting this data, along with the exact values, vary by institution. As above, we include the following possible values, which reflect the union of categories across our institutions, but recognize the limitations of this coding: Asian, Black, Hawaiian, Hispanic, Native American, Not Indicated, White, 2 or More.
- **URM Status:** The term 'Underrepresented Minority' holds a specific institutional meaning in higher education, where it is used to refer to a category of domestic students (those with U.S. citizenship status) who hold membership in an underrepresented racial or ethnic group in the United States. We note that this is a category that is tracked and reported by almost every accredited institution in the United States. As a result, we decided to include this variable, instead of deriving a potentially more socially meaningful, but less contextual, "underrepresented minority" feature. This variable takes the following values: Non-Underrepresented Minority, Underrepresented Minority, International.

For variables marked as "nullable" in Table 2, handle them in two distinct ways: for some variables (high school GPA, ACT English, ACT Math), we drop records with those values not present (since missingness is rare for these features). For the remaining nullable features (SAT Math, SAT Verbal) we use median imputation.

We note that for *all* institutions, we only use records up to the Fall 2019 term in order to avoid forecasting into the academic terms affected by the COVID-19 crisis.

A.2 Train/Test/Validation Split

Our goal is to realistically evaluate models' ability to predict on future data. To do so, we use as training data records from all terms prior to Fall 2019 term. We reserve all records from the Fall 2019 term as validation/testing data, where these records are split evenly into test/validation.

B OPEN-SOURCE CODE RELEASE

Concurrent to the release of this paper, we will make our code publicly available via public GitHub repository. This includes code for data validation, model training, and evaluation, as well as other reproducibility details (software requirements, code for generating figures).

C ADDITIONAL RESULTS

C.1 Transfer Gap Detailed Results

We provide detailed experimental results for the transfer gap measure ΔAUC in Table 3, and detailed experimental results for the AUC Gap fairness measure in Table 4.

C.2 Subgroup-Specific Transfer Detail

This section gives additional results regarding intersectional model performance discussed in Section 6.2. Figure 4 provides the results of each model transfer scenario (local, direct, voting, stacked), organized by intersectional subgroups.

Figure 4 provides additional evidence regarding the zero-shot transfer capacity of models learned via direct transfer and voting transfer: their subgroup performance tends to be similar to the local model, with no discernable effect on performance disparities, measured by AUC gap. Most transfer schemes also have limited effect on subgroup performance relative to the local model, measured by $\Delta\text{AUC}(\text{local}, \cdot)$, as shown in Figure 4b. The exception to this is the

[10] https://nces.ed.gov/ipeds/cipcode/

Feature Name	Description	Type	Nullable	Min	Max	Student	Multiple Columns
Units	Credit units the student enrolled in in this term	Float		0	100		
Units Failed	Percentage of enrolled units the student failed	Float		0	100		
Units Incomplete	Percentage of enrolled units the student earned an incomplete for	Float		0	100		
Units Withdrawn	Percentage of enrolled units the student withdrew from	Float		0	100		
Cumulative GPA	Cumulative GPA from all known tertiary education sources	Float		0	4		
Units Transferred	Total number of transferred units in this term	Integer	✓	0	100	✓	
Age	Age at course start	Integer	✓	0	150	✓	
High School GPA	GPA in high school	Float	✓	0	4	✓	
ACT English	ACT English component score	Integer	✓	0	36	✓	
ACT Math	ACT Math component score	Integer	✓	0	36	✓	
SAT Math	SAT Math component score	Integer	✓	0	800	✓	
SAT Verbal	SAT Verbal component score	Integer	✓	0	800	✓	
GPA Mean	The term-level gpa average weighted by units	Float		0	4		
GPA Stddev	The term weighted gpa stddev	Float		0	inf		
GPA *z*-score	The weighted average z-score of the student in their courses	Float		-inf	inf		
GPA *z*-score stddev	The weighted stddev z-score of the student in their courses	Float		0	inf		
Units Per CIP2	Units taken by 2-digit CIP code	Float		0	100		✓
Units Per Course	Units taken by course format	Integer		0	100		✓
Units Online	Units taken online.	Integer		0	100		
Units In-Person	Units taken in person.	Integer		0	100		
Modality	Whether the student is enrolled in-person or online	Categorical			✓		
Sex	Self-declared sex	Categorical				✓	
Ethnicity	Self-declared ethnicity	Categorical				✓	
URM Status	Institutionally-assigned indicator for underrepresented minority status	Categorical				✓	
Major 1 CIP Code	2-digit CIP code for first major	Categorical		1	61	✓	
Major 2 CIP Code	2-digit CIP code for second major	Categorical		1	61	✓	
Minor 1 CIP Code	2-digit CIP code for first minor	Categorical		1	61	✓	
Minor 2 CIP Code	2-digit CIP code for second minor	Categorical		1	61	✓	
Year	The current year	Integer		2013	2019	✓	
Retention	Indicator for whether student was enrolled in following fall term	Binary		0	1	✓	

Table 2: Data schema used for this study. Each row in the resulting dataset represents a single first-year student present in the Fall academic term at an institution. "Student-level" features indicate those which are fixed for a given student under normal circumstances, and do not vary by term. For more detail on categorical codings and handling of nullable features, see Section A.1 Note that min/max values indicate the min/max enforced by our data validation pipeline; these are not the min/max values occurring in the data (which often fall into a much smaller range).

stacked transfer model, which tends to have both improved performance relative to the local model (high ΔAUC(local, stacked)) and reduced disparities between groups (low AUC gap; see Figure 3).

C.3 Non-Intersectional Comparison for Figure 3

In Section 6.2, we discuss the significance of analyzing model performance disparities via intersectional groups. Here, we present evidence of the difference between the intersectional and non-intersectional (which we refer to as "marginal") analysis.

Figure 5 shows an identical analysis as Figure 3, with the exception that subgroup performance is computed over marginal (non-intersectional) subgroups. Here, we can see that, for each subgroup, the performance of the model is a weighted average of the previous intersectional groups, reducing the observed performance

Target Institution	Source Institution(s)	ΔAUC	SE(ΔAUC)	Transfer Type	p
A	D-C-B	−0.022	0.016	Majority Voting	0.178
A	D-C-A-B	0.010	0.016	Stacked	0.505
A	D	−0.057	0.017	Direct	0.001*
A	C	−0.015	0.016	Direct	0.352
A	B	−0.019	0.016	Direct	0.243
B	D-C-A	0.025	0.062	Majority Voting	0.692
B	D-C-A-B	0.026	0.062	Stacked	0.674
B	D	−0.041	0.065	Direct	0.528
B	C	0.004	0.063	Direct	0.948
B	A	−0.022	0.064	Direct	0.735
C	D-A-B	−0.050	0.030	Majority Voting	0.096
C	D-C-A-B	0.006	0.028	Stacked	0.839
C	D	−0.043	0.030	Direct	0.149
C	A	−0.094	0.031	Direct	0.002*
C	B	−0.079	0.031	Direct	0.010*
D	C-A-B	0.003	0.066	Majority Voting	0.960
D	D-C-A-B	0.000	0.066	Stacked	0.998
D	C	0.013	0.065	Direct	0.841
D	A	−0.039	0.069	Direct	0.573
D	B	0.019	0.065	Direct	0.770

Table 3: Detailed results for transfer gap ΔAUC for each transfer scheme evaluated. The final column gives the p-value of the hypothesis test of ΔAUC ≠ 0 for the transfer scheme evaluated. ∗ indicates we reject H_0 at $\alpha = 0.05$.

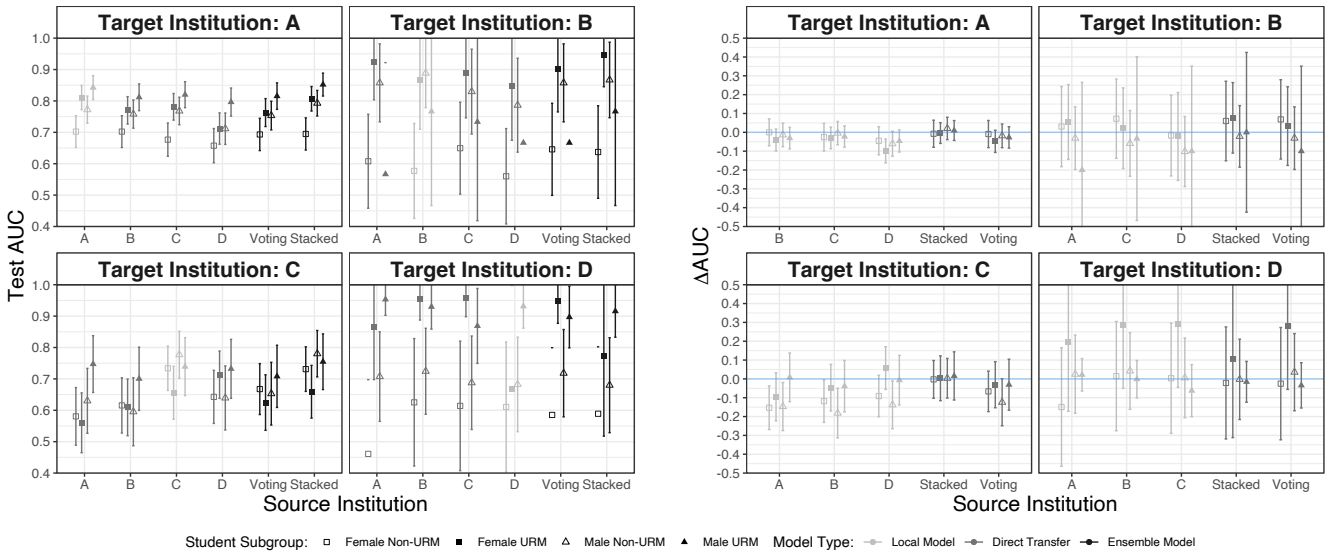


Figure 4: (a) Left: Model performance over intersectional subgroups for various transfer schemes evaluated (Male, Female, URM, Non-URM) for all institutions. (b) Right: ΔAUC values over sensitive subgroups for various transfer schemes evaluated. One-SE error bars shown for both figures. Direct and local transfer models are in lexicographic order (A, B, C, D) within each subgroup.

disparities and masking larger disparities within the intersectional groups.

C.4 Impact of Ensemble Selection Strategies

We provide results comparing the impact of different model selection strategies in Figure 6.

Target Institution	Source Institution(s)	AUC Gap	SE(AUC Gap)	Transfer Type	p
A	D-C-B	0.122	0.034	Majority Voting	0.00*
A	D-C-A-B	0.157	0.032	Stacked	9.46×10^{-7}*
A	B	0.109	0.034	Direct	0.00*
A	C	0.143	0.034	Direct	2.92×10^{-5}*
A	D	0.139	0.036	Direct	0.00*
A	A	0.140	0.032	Direct	1.41×10^{-5}*
B	D-C-A	0.255	0.102	Majority Voting	0.01*
B	D-C-A-B	0.308	0.091	Stacked	0.00*
B	A	0.356	0.191	Direct	0.06*
B	C	0.241	0.105	Direct	0.02
B	D	0.286	0.117	Direct	0.01*
B	B	0.311	0.096	Direct	0.00
C	D-A-B	0.083	0.068	Majority Voting	0.22
C	D-C-A-B	0.121	0.057	Stacked	0.03
C	A	0.187	0.067	Direct	0.01*
C	B	0.105	0.075	Direct	0.16
C	D	0.093	0.071	Direct	0.19
C	C	0.121	0.058	Direct	0.04*
D	C-A-B	0.364	0.115	Majority Voting	0.00
D	D-C-A-B	0.327	0.117	Stacked	0.01
D	A	0.492	0.123	Direct	6.72×10^{-5}*
D	B	0.328	0.109	Direct	0.00*
D	C	0.344	0.110	Direct	0.00*
D	D	0.321	0.112	Direct	0.00*

Table 4: Detailed results for fairness measure AUC Gap for each transfer scheme evaluated. The final column gives the p-value of the hypothesis test of AUC Gap $\neq 0$ for the transfer scheme evaluated. $*$ indicates we reject H_0 at $\alpha = 0.05$.

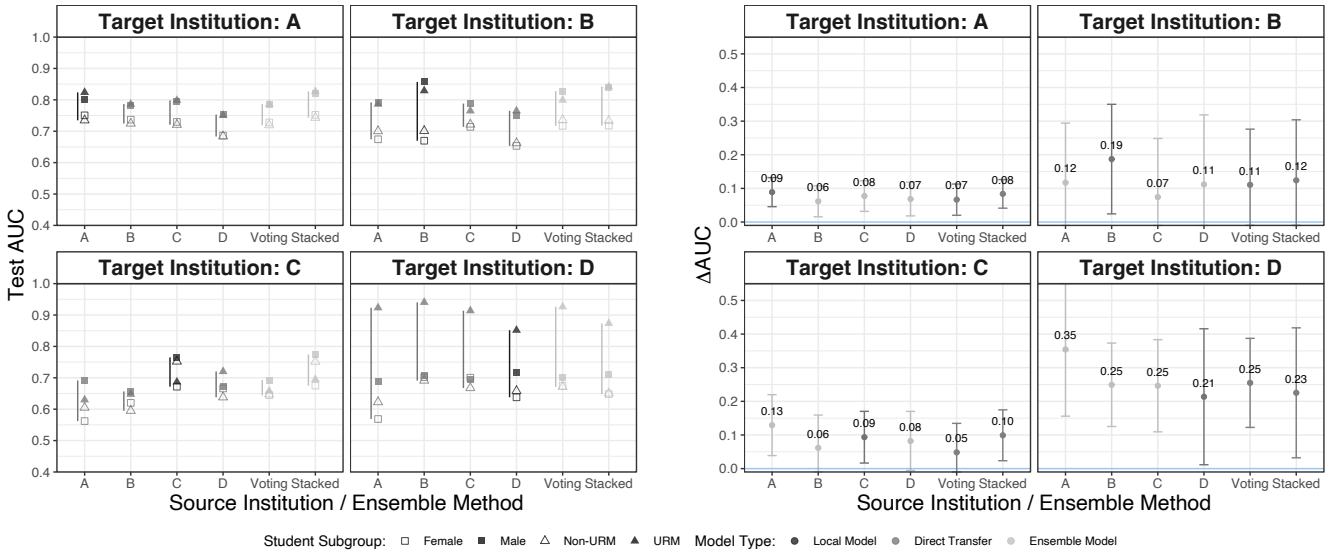


Figure 5: Non-intersectional ("marginal") version of Figure 3, where AUC is computed over marginal groups, not intersectional groups. In comparison to Figure 3, this analysis shows considerably smaller disparities. This demonstrates how marginal analyses can mask intersectional performance disparities.

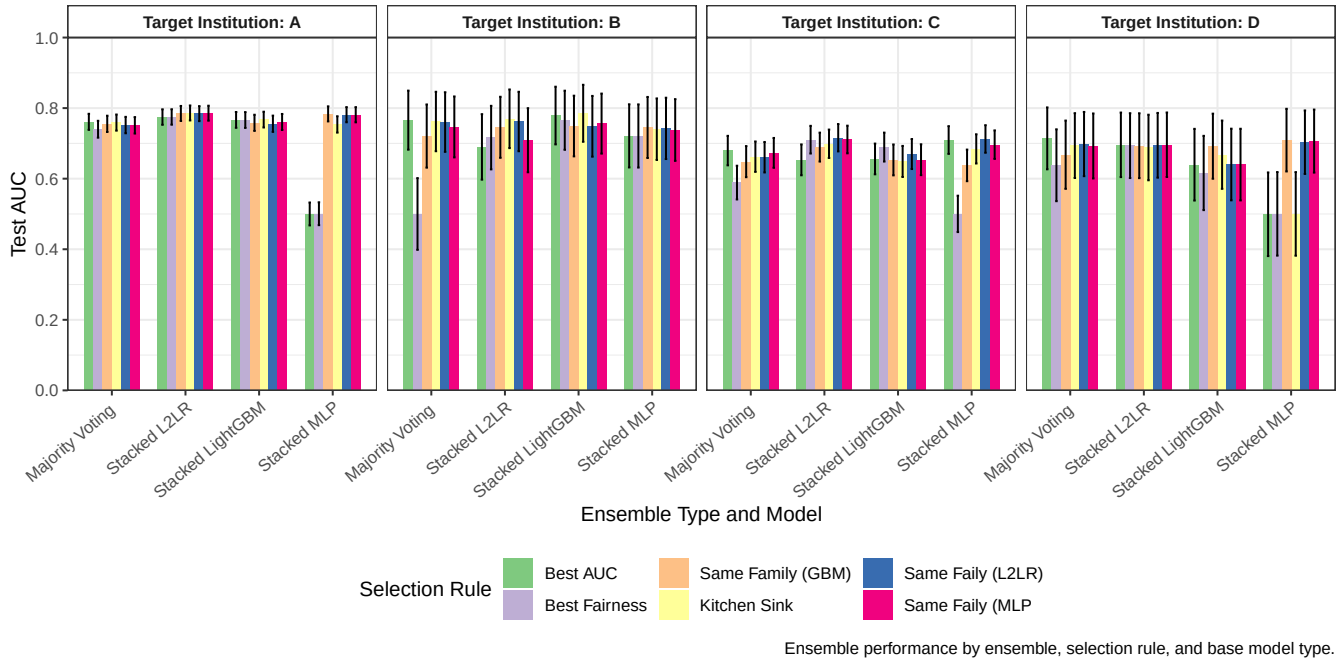


Figure 6: Test AUC by ensemble type and selection rule.

C.5 Impact of Functional Form for Base and Stacked Ensemble Models

In this section, we provide additional results for LightGBM, MLP models not discussed in the main text due to space limitations.

We provide identical experimental results as our figures in the main text here, using LightGBM and MLP models in place of L2LR. Figure 7 shows results analogous to main text Figure 1; Figure 8 shows results analogous to main text Figure 2; and Figure 9 provides results analogous to Figure 3. The results shown for LightGBM and MLP models are consistent with those discussed in the main text.

C.6 Regularization Study

Prior work suggested that *regularization* affects model fairness by controlling worst-group outcomes, including in modeling regimes relevant to our fairly simple L_2-regularized logistic regression approach. For example, [45] shows that a regularizer consisting of an L_2 penalty, combined with a "prejudice removal" regularizer, can reduce a measure of unfairness. However, their specific formulation of unfairness seeks to minimize the model's reliance on sensitive attributes to avoid disparate treatment, while in our experiments, we do not seek to explicitly avoid this. Using distributionally-robust neural network training, [65] showed that increasing regularization (via increasing an L_2 penalty or via early stopping) improves worst-group accuracy. [48] suggested that the removal of spurious (i.e. non-informative) features can have disproportionate effects on subgroups. We are aware of no previous work which explores the effect of regularization on cross-institutional transfer, despite the clear connection between regularization, the bias-variance tradeoff, and generalization error, which could have implications for domain transfer.

In this section, we conduct an exploratory study of the impact of regularization on overall performance, and on equitable performance over intersectional subgroups. Our procedure is as follows: for each training experiment above, we fix the L_2 regularization penalty parameter $\lambda \in 0, 10^{-4}, 10^{-3}, \ldots, 10^{4}$ and follow the same training and evaluation procedure. We only conduct this for *local* and *direct* transfer scenarios. Our aim in this study is to determine how regularization might affect transfer, and how specific subgroups are affected.

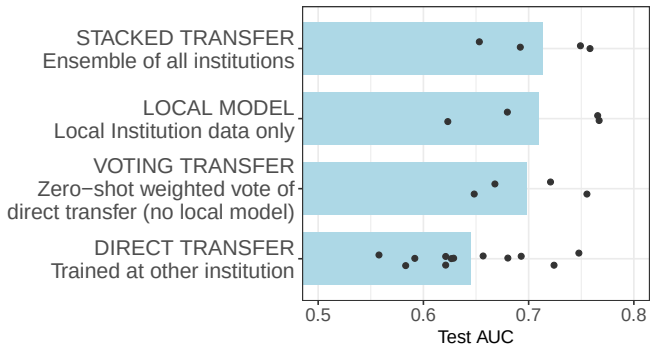
Results of this study are shown in Figure 10. First, the dotted lines in Figure 10a indicate AUC at different levels of regularization, which shows the standard expected result that (due to bias-variance tradeoff) regularization tends to have an "optimal" value (indicated by a '*' dot), above or below which a model's test performance tends to decline.

Figure 10b shows that each intersectional subgroup tends to respond similarly to regularization, and that changes in regularization generally fail to reduce performance disparities across all source-target institution pairs, with the rank-ordering of model performance for subgroups largely consistent across source institutions.

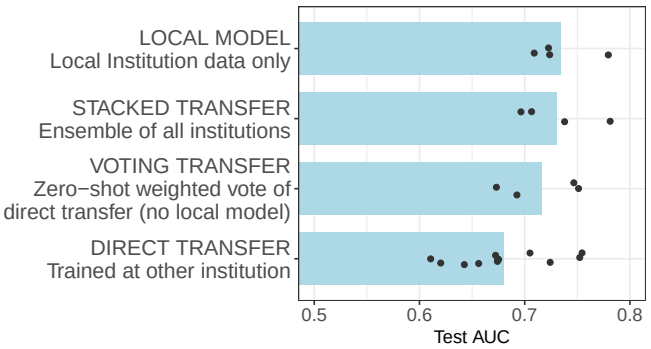
We discuss the results in further detail in Section 6.

C.7 Model Similarity Analysis

While our experiments demonstrate that models trained locally at each institution have similar average performance to each other, our experiments do not verify that these models learn similar functions of the inputs; instead, they merely verify that their average



(a) Overall results for LightGBM model.



(b) Overall results for MLP model.

Figure 7: Overall results for LightGBM (7a) and MLP (7b) models. The findings are consistent with our findings for L_2-regularized logistic regression shown in Figure 1.

performance on each institution's test set is similar. In this section, we briefly explore whether the learned *coefficients* are similar between local models learned at each institution.

A standard method to compare whether logistic regression coefficients differ between two datasets requires multi-institutional training (having access to *all* datasets), and would involve adding an `institution` indicator variable, and determining whether the coefficient corresponding to this variable was nonzero in the fitted model. However, since we are unable to train directly on multi-institutional datasets due to the privacy constraints mentioned above, we explore alternative methods for model comparison. These methods should be considered qualitative explorations of the similarity of the learned models.

We provide two exploratory analyses to address this question. First, in Figure 12, we compute an overlap metric, Overlap@k, for the local models from each pair of institutions. Overlap@k is computed as follows: let $\hat{\theta} = [\theta_1, \ldots, \theta_d]$ represent the d-dimensional coefficients of a model. For each pair of $\hat{\theta}_i, \hat{\theta}_j$, we separately sort the elements in descending order by magnitude $sort(\hat{\theta})$. Then, for fixed k, we take the first k elements of both vectors and compute the size of the overlap:

$$\text{Overlap@}k(\hat{\theta}_i, \hat{\theta}_j) = \left|sort(\hat{\theta}_i)[1:k] \cap sort(\hat{\theta}_j)[1:]\right| \quad (4)$$

Intuitively, Overlap@k measures the level of agreement between models about which coefficients are largest in magnitude. This does not, for example, ensure that these coefficients have even the same direction, but it provides a qualitative measure of agreement on feature importance. Two models which have identical rank-ordering of feature magnitudes would have Overlap@k of k, the highest possible value; two models which do not agree on any of the highest-k-magnitude features would have Overlap@k of zero.

Because it can be easier to interpret the overlap as the relative size of the intersection, we also report the Normalized Overlap@k, obtained by normalizing by a factor of $1/k$:

$$\text{Normalized Overlap@}k(\hat{\theta}_i, \hat{\theta}_j) = \frac{1}{k}\left|sort(\hat{\theta}_i)[1:k] \cap sort(\hat{\theta}_j)[1:]\right| \quad (5)$$

Results of our computation of these similarity metrics are shown in Figure 12.

As an additional check of model similarity, which (unlike Overlap@k) accounts for the *directionality* of the feature vectors, we also report the cosine similarity between each pair of model coefficients. Cosine similarity is a measure of the angle between two vectors, irrespective of their magnitudes, and is defined as:

$$cossim(\hat{\theta}_i, \hat{\theta}_j) = \frac{\langle \hat{\theta}_i, \hat{\theta}_j \rangle}{||\hat{\theta}_i|| ||\hat{\theta}_j||} \quad (6)$$

Institution	B	C	D
A	0.46	0.40	0.27
B		0.34	0.22
C			0.17

Table 5: Model similarity measure $cos(\hat{\theta}(X_i), \hat{\theta}(X_{i'}))$.

Results of this comparison are shown in Table C.7

D COMPUTING STANDARD ERRORS OF AUC

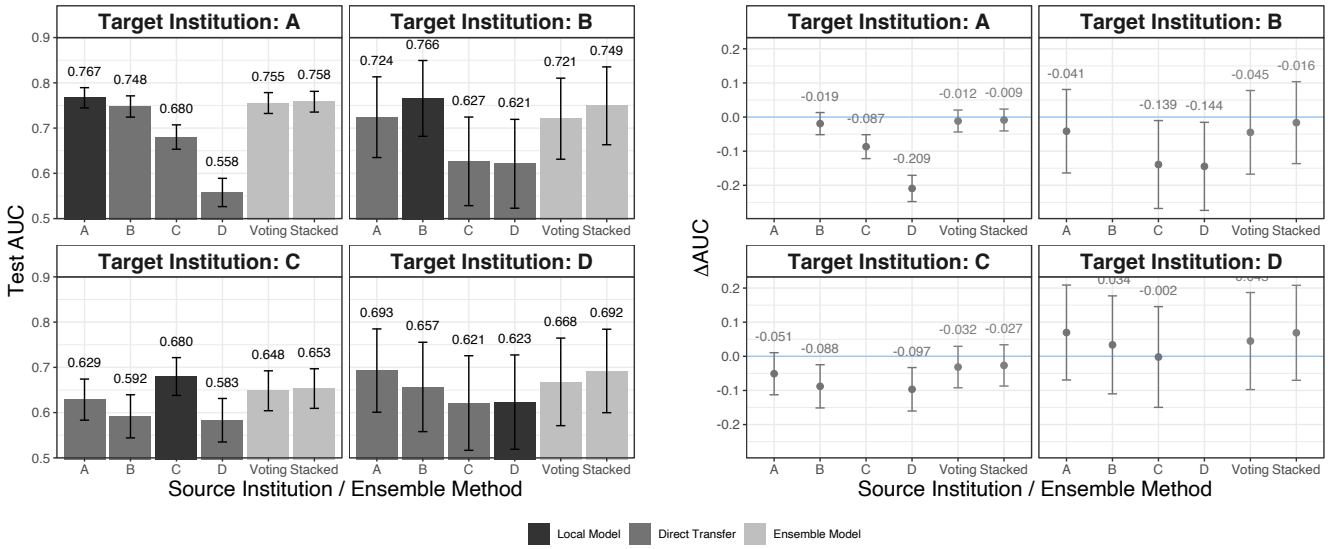
As discussed above, we compute standard errors for AUC estimates according to the procedure described in [30, 35], which utilizes the equivalence between the Area Under the Receiver Operating Characteristic Curve and the Wilcoxon Statistic.

Formally, let $n_p = \sum_{i=1}^{n} \mathbb{1}(y_i = 1)$ and $n_n = \sum_{i=1}^{n} \mathbb{1}(y_i = 0)$ be defined as the number of positive and negative examples in the dataset of interest, respectively. Define A' as the AUC on the dataset. Then, let

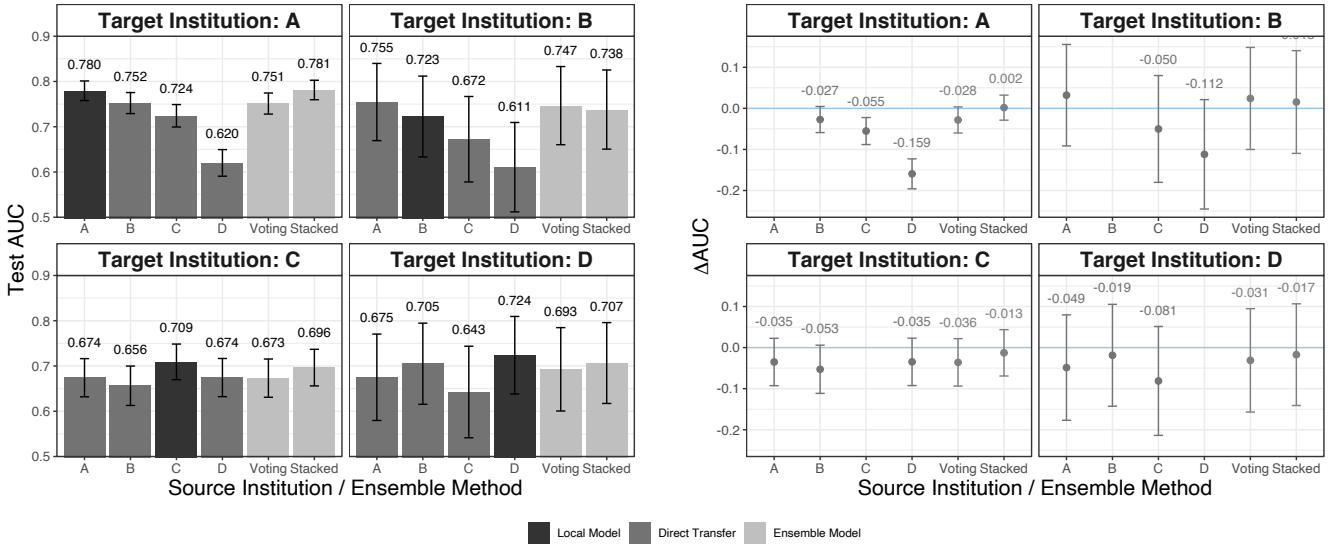
$$D_p := (n_p - 1)\left(\frac{A'}{2 - A'} - A'^2\right) \quad (7)$$

$$D_n := (n_n - 1)\left(\frac{2A'^2}{1 + A'} - A'^2\right) \quad (8)$$

Then the standard error of A' can be computed as:



(a) Results for LightGBM model.

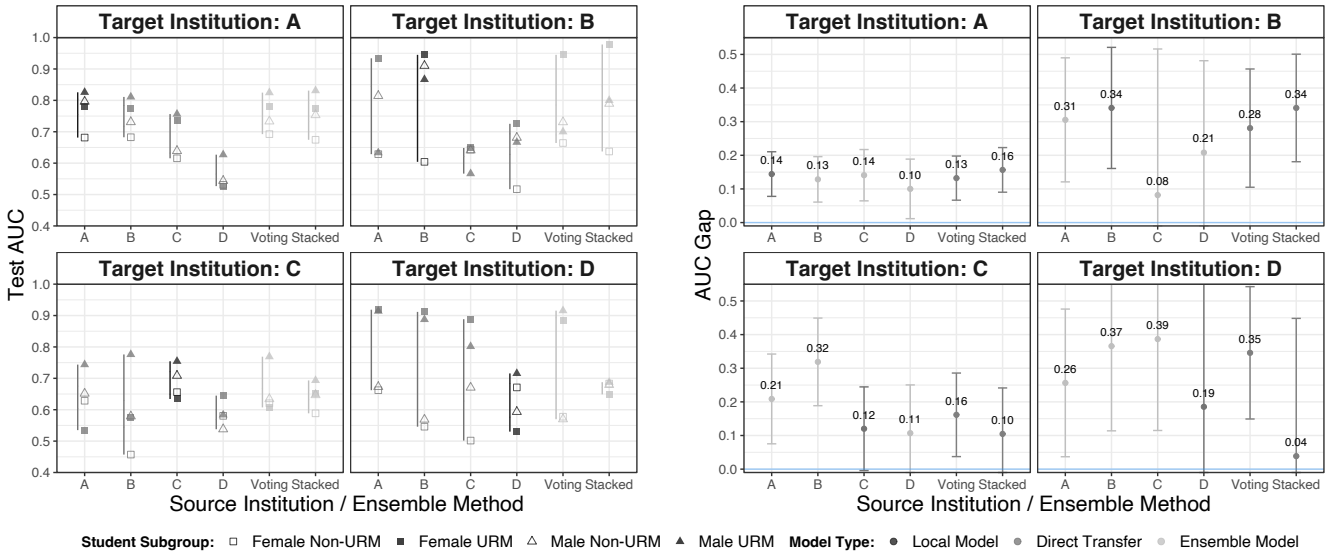


(b) Results for MLP model.

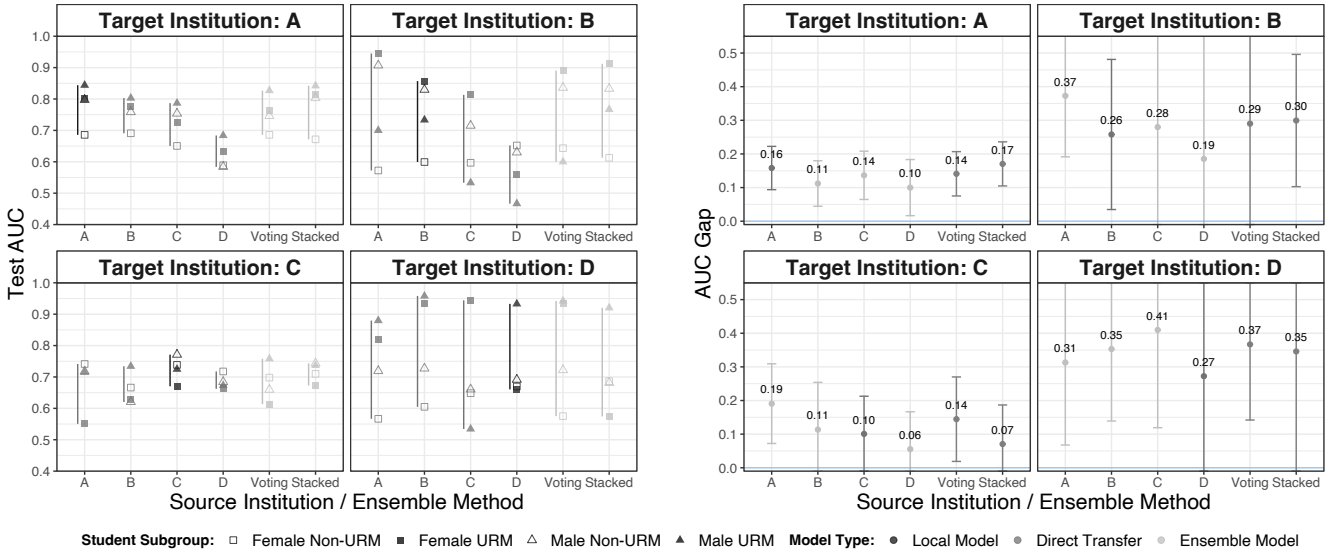
Figure 8: Additional results for LightGBM (8a) and MLP (8b) models. The findings are consistent with our findings for L_2-regularized logistic regression shown in Figure 2: the 95% CI for AUC for voting transfer overlaps with the local model, for all institutions and for both LightGBM and MLP. Additionally, the 95% CI for ΔAUC overlaps with zero, indicating no transfer gap between the voting transfer model and the local model.

$$SE(A') = \sqrt{\frac{A'(1-A') + D_p + D_n}{n_p n_n}} \tag{9}$$

For further details and proof, we defer the reader to [35].

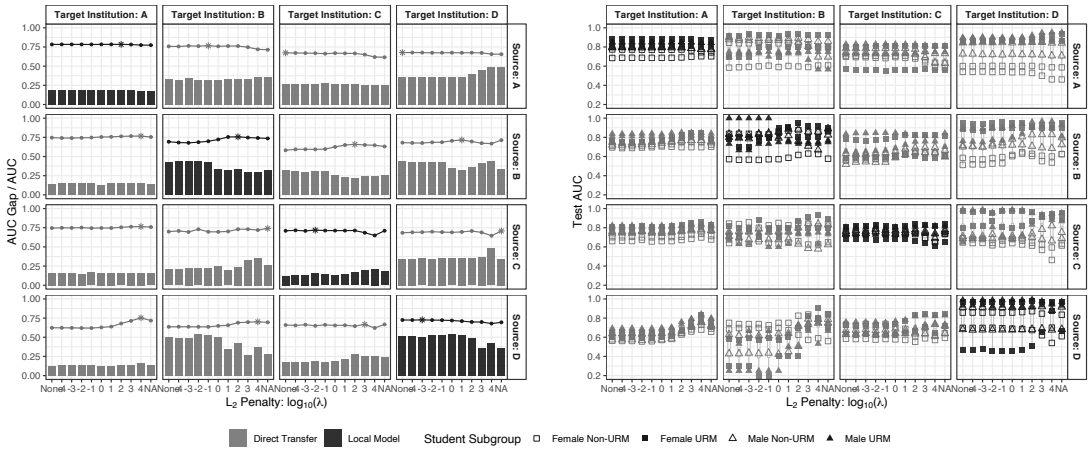


(a) Results for LightGBM model.

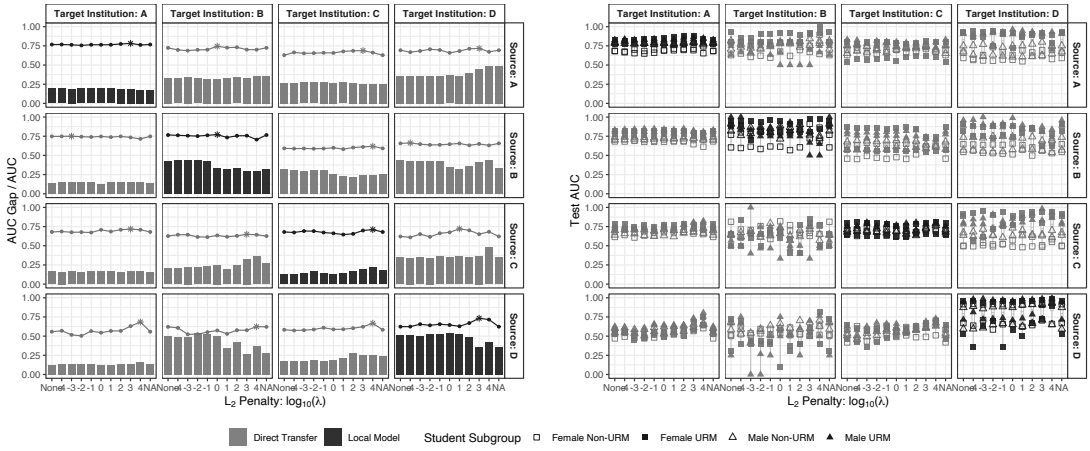


(b) Results for MLP model.

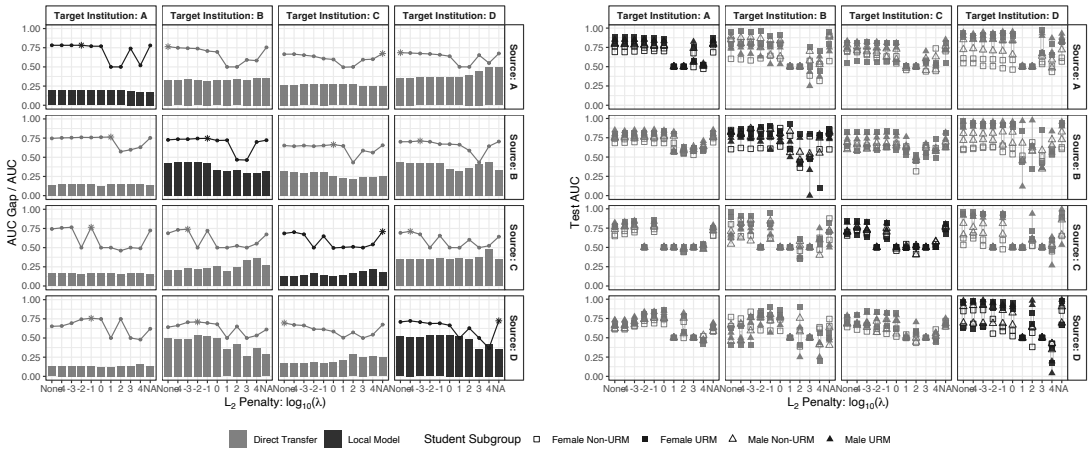
Figure 9: Additional results for LightGBM (9a) and MLP (9b) models. The findings are consistent with our findings for L_2-regularized logistic regression shown in Figure 3: the 95% CI for AUC Gap for voting transfer overlaps with the local model, for all institutions and for both LightGBM and MLP, indicating no difference in the intersectional subgroup performance disparities between the voting transfer model and the local model.



(a) Regularization study results for L2LR model.



(b) Regularization study results for LightGBM model.



(c) Regularization study results for MLP model.

Figure 10: Results of regularization study over various L_2 regularization strengths for local and direct transfer models. Left: Test AUC (lines) and AUC Gap (bars). '*' marker indicates the value λ^* which achieves highest test AUC for the given model type and source/target institution. Right: Intersectional subgroup model performance.

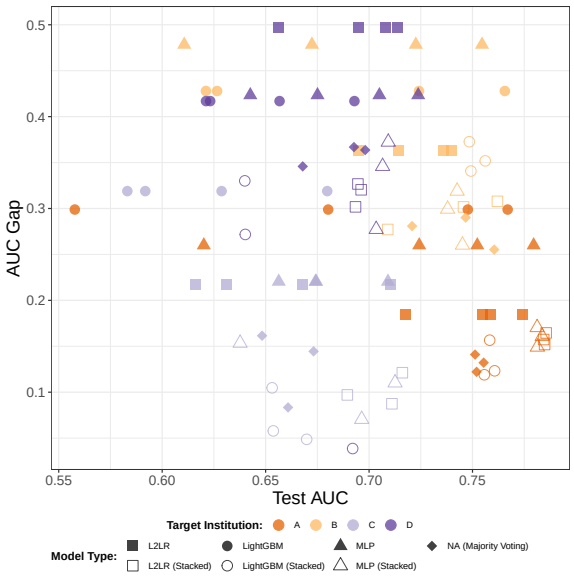
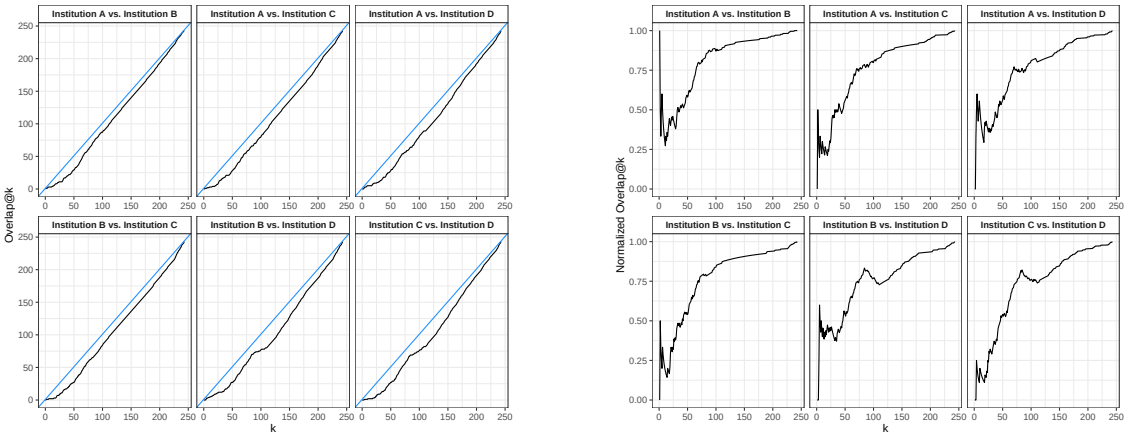


Figure 11: Overall performance (measured by AUC) vs. fairness (AUC Gap, Eq. (3)).



(a) Model similarity metric Overlap@k for each pair of local models.

(b) Model similarity metric Normalized Overlap@k for each pair of local models.

Figure 12: The proposed model similarity measures for each pair of locally-learned models.