Evaluation of the instance weighting strategy for transfer learning of educational predictive models

Mariia Luzan LUZAN@UMICH.EDU and **Christopher Brooks** BROOKSCH@UMICH.EDU
School of Inforamtion, University of Michigan, 105 South State Street, 48103 USA

Abstract

This work contributes to our understanding of how transfer learning can be used to improve educational predictive models across higher institution units. Specifically, we provide an empirical evaluation of the instance weighting strategy for transfer learning, whereby a model created from a source institution is modified based on the distribution characteristics of the target institution. In this work we demonstrated that this increases overall model goodness-of-fit, increases the goodness-of-fit for each demographic group considered, and reduces disparity between demographic groups when we consider a simulated institutional intervention that can only be deployed to 10% of the student body.

Keywords: Transfer learning; Higher education; Student dropout prediction; Fairness

1. Introduction

Student attrition in higher education remains a significant problem. For example, in the US, only 64% of students who began pursuing a bachelor's degree at a four-year institution in fall 2014 successfully completed their degree at the same institution within six years (National Center for Education Statistics, 2022). Universities suffer an immediate financial loss by losing tuition fees for the remaining years of potential student enrollment. In addition, the dropout rate negatively impacts a university's ranking, overall attractiveness to future students, and its eligibility for funding and grants. For students the consequences are also severe and include unrealized potential, lower income and the risk of mental health problems caused by feelings of failure (Matz et al., 2023).

One potential strategy for addressing the issue of students' attrition involves early identification and subsequent support for struggling students to ensure their academic success. The past decade has seen significant interest in using machine learning to identify students at high risk of dropping out (Gardner and Brooks, 2018; Barber and Sharkey, 2012; Balakrishnan, 2013). However, resource constraints such as insufficient data or limited technical capacity may hamper the development of these predictive models in some institutions. Frustratingly, the institutions with fewer resources for this activity may be the ones who need the predictive models themselves, increasing the disparities between students attending well-resourced universities and those at less resourceful institutions.

Transfer learning has been proposed as one way in which universities with limited resources might be able to take advantage of models created by more highly resourced universities (Gardner et al., 2023). In this approach models created by well resourced institutions may be ensembled together and applied to a new institution in a zero-shot or additive manner, allowing for prediction with limited investment. However, little research has been done on the suitability of this approach, both with respect to logistical and technical approaches which may be taken.

The goal of this work is to evaluate the efficiency of an instance weighting transfer learning method to improve model transfer between institutions. This method transforms the feature dataset from the presumably well-funded university (source domain) to align it more closely with the feature dataset from the resource-limited university (target domain). This is achieved by defining weights for the source domain data points in such a way as to minimize the distance between the source and target distributions. Huang et al. (2006) proposed using *kernel mean matching* to estimate the weights. By incorporating these weights during model fitting, the model developed based on the source dataset becomes more relevant to the target domain.

In this work we contribute an empirical evaluation of this *instance weighting strategy* and compare it to the *baseline direct transfer* for the task of predicting next year enrollment for the first-year undergraduate students. We compare these two strategies across three interrelated analyses including the **goodness-of-fit metrics (RQ1)** of models, measured specifically by AUC, Pietra Index, and Kolmogorov-Smirnov test, **fairness of models (RQ2)** across different gender and ethnic identity groups, and **differences to fairness (RQ3)** for a simulated institutional intervention impacting 10% of all students. A full listing of all technical details, as well as a reproducibility checklist, can be found in the appendices.

2. Related work

2.1. Prediction of student dropouts

Due to the significant importance of identifying students at risk of dropping out, there has been extensive research on predictive models to estimate the probability of students' dropout. The promising outcomes, demonstrating high predictive accuracy of these models, have motivated further research in this area. Various statistical and machine learning models, including survival analysis, logistic regression, random forests, support vector machines, gradient boosted trees, neural networks, and others, have been used in predicting student dropouts (Ameri et al., 2016; Aulck et al., 2019; Andrade-Girón et al., 2023; Matz et al., 2023). These studies have identified factors with high predictive power for the task such as demographic indicators (gender, ethnicity, age), socio-economic aspects (family income, parental education), high school academic performance, current academic standing, and student engagement levels with peers and the university.

It's worth noting that much of this research has primarily focused on a university level analysis, not considering the possibility of merging or transferring datasets from different institutions.

2.2. Transfer Learning in educational predictive models

The aim of transfer learning is to improve the effectiveness of machine learning models within a specific area (referred to as the target domain) by utilizing knowledge from a related domain (referred to as the source domain). In the context of predicting student dropouts, the goal of transfer learning is to leverage data and/or models from a well-resourced university to develop predictive models for a less resourceful university.

While transfer learning has seen significant research in domains such as text analysis and image recognition, little work has been done to understand how this method might improve educational predictive models. One early work in this area was done in prediction of student course dropout in Massive Open Online Courses (MOOCs) (Boyer and Veeramachaneni, 2015). In this work authors compared three different transfer learning methods at the fourth week of the course, predicting whether the student would continue or not.

Hunt et al. (2017) compared the accuracy of the transfer learning approach, TrAdaBoost, with both AdaBoost trained solely on the target domain and on a merged dataset combining the source and target data for predicting students' graduation rates. Their experiments demonstrated that TrAdaBoost had the lowest classification error.

Gardner et al. (2023) investigated the cross-institutional transfer of predictive models among four U.S. universities to forecast first-year retention rates. They evaluated three transfer learning strategies: direct transfer, voting transfer, and stacking transfer. Their findings indicated that the voting transfer method achieved predictive accuracy of a locally developed model using an institution's own data. Notably, this accuracy was achieved without negatively affecting model fairness.

2.3. Instance weighting strategy (IWS) using kernel mean matching

The instance weighting strategy is based on the assumption that the differences between the target and source distributions lie only in the marginal distributions of the predictor factors ($P^S(x) \neq P^T(x)$), not in the conditional distributions of the predicted factor ($P^S(y|x) = P^T(y|x)$). In the context of learning analytics models, this implies that a student's academic success depends on their individual attributes rather than the quality of education at their institution. The variation in dropout rates among different colleges can be attributed to the fact that these institutions tend to attract students with varying characteristics. For instance, private schools often accept students from more affluent socio-economic backgrounds.

In order to build a model for the target domain, we need to minimize the expected risk. The formula below demonstrates how we can utilize source data for this purpose, provided the assumption of equal conditional distributions of the response (Huang et al., 2006):

$$E_{(x,y)\sim P^T}(L(x,y,\theta)) = E_{(x,y)\sim P^S}\left[\frac{P^T(x,y)}{P^S(x,y)}L(x,y,\theta)\right] = E_{(x,y)\sim P^S}\left[\frac{P^T(x)}{P^S(x)}L(x,y,\theta)\right]$$

To leverage the source data for building a model for the target domain, it's essential to know the weights $\beta = P^T(x)/P^S(x)$. Usually, the weights β are unknown in practice. Huang et al. (2006) proposed employing the Kernel Mean Matching (KMM) procedure, which aims to minimize the difference between the means of the source and target features distributions in a reproducing kernel Hilbert space to determine the weights. This is equivalent to solving the following optimization problem (Huang et al., 2006):

$$min_\beta \left[\frac{1}{2}\beta^T K \beta - k^T \beta\right]$$

$$\beta_i \in [0; B], |\Sigma_1^m \beta_i - m| \leq m\epsilon$$

where K is a kernel matrix with elements $K_{i,j} = k(x_i^S, x_j^S)$ and $k_i = \frac{m}{n}\Sigma_{j=1}^{n} k(x_i^S, x_j^T)$, m – source sample size, n – target sample size.

Figure 1 illustrates the implementation of the instance weighting strategy in developing a model for transfer.

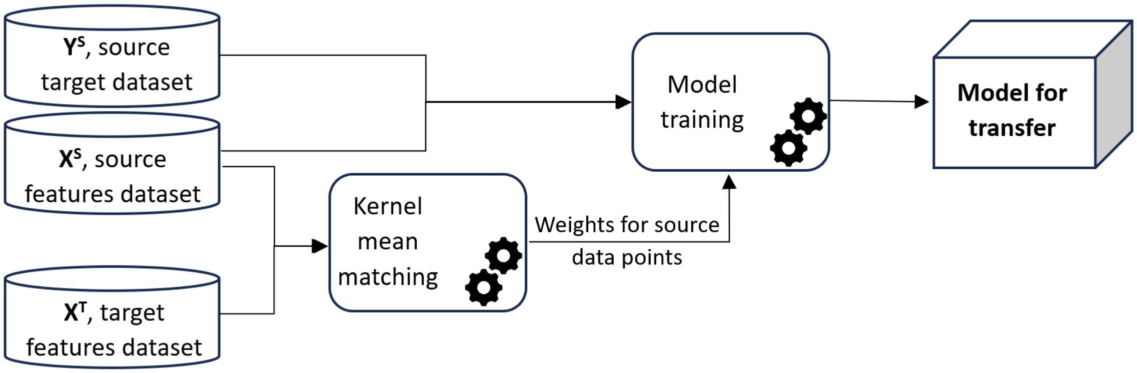
Y^S, source target dataset
X^S, source features dataset
X^T, target features dataset
Kernel mean matching
Weights for source data points
Model training
Model for transfer

Figure 1: Transfer Learning with Instance Weighting Strategy

The result of applying instance weighting strategy becomes evident when comparing the original source feature distribution with the source distribution after applying weights. Figure 2 illustrates this effect using synthetic data, where both source and target are normal random variables with different mean and dispersion values. The visualization clearly shows that applying instance weighting strategy brings the source distribution closer to the target.

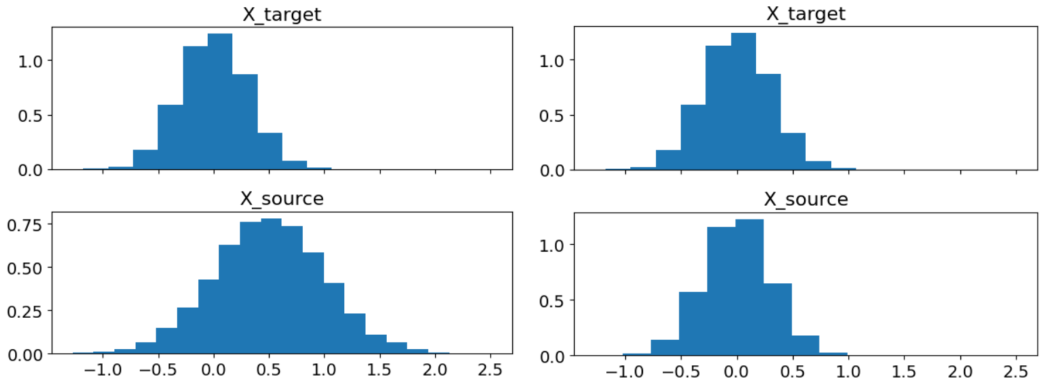

Figure 2: Source Distribution Before (left) and After Applying Instance Weighting (right).

3. Experiment and Results

The dataset used for the experiment is a large de-identified student information system corpus covering a period of six years from a large public university in the U.S (University of Michigan, 2023). It includes a range of static data about students such as demographics, socio-economic factors such as parental income, and pre-admission performance metrics such as high school grades and SAT/ACT scores. It also contains dynamic university semester information such as academic career goals, grades (both in aggregate and per-course), and course topic and workload characteristics.

As it is difficult to obtain access to this kind of data for multiple institutions, we simulated a multi-institutional approach by selecting data from two direct-entry four-year colleges. The first, which we used as our *source* institution, is a broad liberal arts college which offers degrees in the natural sciences, social sciences, humanities, and arts, and the second, our *target* institution, is a smaller college offering degrees in engineering, computer science, and technology.

Table 1: Summary of experiment data from (University of Michigan, 2023) over the years of 2015–2021. Demographic indicators are as recorded by the institution.

School	**Sample Size**	**Dropouts**	**Male**	**Female**	**White**	**Asian**	**Other**
Source	30,122	491	12,718	17,404	18,171	5,472	6,479
Target	9,887	127	6,986	2,901	4,966	2,619	2,302

To explore our three research questions we developed a logistic regression model for the source school with a prediction target of seeing a student re-enroll one year later and validated this model with out-of-sample (30%) and out-of-time (1 year) test sets achieving AUCs of 0.817 and 0.872 respectively. A list of the features used from the dataset are enumerated in Appendix A, and we measured the *baseline direct transfer* of the source to our target school as having an AUC of 0.792, 90%CI [0.748, 0.830]. For the *instance weighting strategy* the weights for the source data points were obtained through the optimization from Section 2.3, with parameters: Kernel type - Radial Basis Function (RBF) kernel, $B = 1000$, $\epsilon = (\sqrt{m} - 1)/\sqrt{m}$ (Huang et al., 2006), RBF's length scale = 3.5 (Appendix A.2). The model was then estimated using source data points weighted with the calculated weights and achieved a transfer AUC of 0.808, 90%CI [0.766, 0.844]. The difference between the transfer and direct AUC values is 0.016. The null hypothesis was tested to determine if the two AUC values (direct and weighted transfer) were equal: $z = 2.378$, $p - value = 0.017$; the standard deviation of the AUC difference for the testing purposes was estimated using a paired stratified bootstrap procedure. Fairness (RQ2) was measured through a slicing analysis for demographic categories following (Gardner et al., 2019) with results shown in Table 2. Overall AUC values for both transfer models used the entire dataset from the target school.

The goodness-of-fit of the models was assessed using AUC due to its applicability in unbalanced data sets, it's intuitive probabilistic interpretation, and its frequent use in educational prediction tasks. Pietra index was used to measure the maximum distance between

Table 2: Goodness-of-fit and discrimination power of models.

	RQ1 transfer indicators			RQ2 fairness indicators (AUC)				
Model	**Overall AUC**	**Pietra index**	**KS (p)**	**White**	**Asian**	**Other**	**Male**	**Female**
Direct transfer	0.792	0.482	2.88e-27	0.852	0.724	0.780	0.790	0.798
IWS	0.808	0.530	2.13e-33	0.865	0.743	0.790	0.801	0.813

two cumulative distribution functions (the predicted dropout probabilities for discontinued students and those for enrolled students), and the Kolmogorov-Smirnov method was used to test the null hypothesis that the predicted probabilities of the two discontinued/enrolled students are drawn from a single distribution.

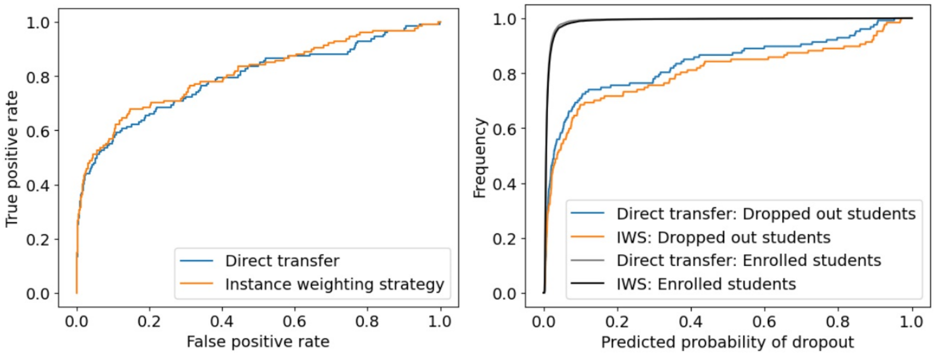

Figure 3: Models' Discriminatory Power Comparison Using ROCs (left) and CDFs (right).

Analysis of fairness through AUC values alone do not account for university resources in intervening, and thus we also analyzed the models when only a limited number of students could be supported by an intervention. To explore this question (RQ3), we arbitrarily set the bounds on the number of students who could be intervened with to the 10% with the highest predicted probability of dropout, the measured the equal opportunity difference and the generalized entropy index (specifically, the coefficient of variation). The recall values corresponding to this 10% threshold for direct transfer and weighted transfer are 55.1% and 56.7%, respectively. The precision values are 7.1% and 7.3%, respectively. These precision values are expected to be low due to a dropout rate of only 1.2%, with university support directed towards the most struggling 10% of students.

The equal opportunity difference estimates the difference in true positive rates between two groups of students. Due to space constraints, we engaged in a single analysis choosing the highest and lowest AUC ethnic groups from Table 2, setting White students as the privileged and Asian students as the unprivileged groups. The generalized entropy index has the subgroup decomposability property, which means that the index can be expressed as the sum of unfairness components between groups and within groups (Speicher et al., 2018). It illustrates how efforts to minimize one unfairness component impact the remaining unfairness component. Speicher et al. (2018) suggested to apply the index to benefits that individuals receive from the model's application.

Recognizing the importance of ensuring that students in need of support receive it, and accounting for an imbalanced dataset, a customized version of this metric was used where the true positive cases receive a benefit of 1, false negative cases receive a benefit of 0, while all other cases are not considered. Results are shown in Table 3.

Table 3: Fairness of models between White and Asian groups at a threshold of 10%.

Model	**Equal Opportunity Difference**	**Variation coefficient**	**Between group variation coefficient**
Direct transfer	-0.394	0.407	0.050
IWS	-0.235	0.382	0.020

4. Discussion

This work contributes to the field of educational predictive models by examining how indicators of goodness-of-fit and fairness of models change when different transfer learning methods are employed. Specifically, we found that the the instance weighting strategy using kernel means matching resulted in a slight increase of model goodness-of-fit (RQ1, as measured by AUC) and better discrimination power (RQ1, as measured by Pietra index and Kolmogorov-Smirnov metrics) versus a baseline direct model transfer. We further demonstrated that this transfer learning approach increases model goodness-of-fit across all measured demographic categories (RQ2, shown in Table 2), though it does not ameliorate bias nor (in our experiment) change the ordinal goodness-of-fit metrics of models between demographic groups. In short, the instance weighting method improves the transfer of models and should be considered by machine learning engineers when applying transfer learning in education.

In addition to these two established techniques for measuring the fairness of transfer learning, we provided a new domain-specific lens (RQ3) to consider the issue of fairness in higher education settings. Many interventions, such as one-on-one advising appointments or special tutoring suppport, are expensive, and cannot be deployed to all students. To simulate this, we considered fairness at a threshold of 10%, and were surprised that this increased fairness (through equalized opportunity difference) between our chosen two groups (Table 3). More work is needed to understand what reasonable thresholds might be for higher education models and call into question how we measure the true goodness-of-fit of student success models.

However, we note several limitations of our work which require additional investigation. First, our data source is from two different colleges within a single university, and a more authentic opportunity would be to transfer models between institutions which have higher disparities in creating predictive models. Second, the instance weighting approach requires access to the data from both institutions in order to run the kernel mean matching method, which may reduce the value of using transfer learning as some form of data sharing is required (unlike the approaches used in (Gardner et al., 2023)). Finally, we observe the same issues in data scarcity in our analysis which existed in (Gardner et al., 2023), in particular that low disenrollment by certain groups of students makes a nuanced fairness analysis difficult. However, with the empirical evidence we provide here, we believe that a more authentic experiment between more highly disparate institutions will be fruitful in improving model transfer in educational predictive analytics.

There is a larger translational research question which also needs to be considered – if we build such models which can transfer, will lower-resourced institutions use them to improve outcomes? There are several considerations which need exploration to further this vein of research. First, researchers need to go beyond single institution or few institution investigations such as the one we have done here and the one done by (Gardner et al., 2023). A larger, more diverse, and more needful set of institutions needs to be involved in future work, and must be centered around those institutions who seek to benefit from model transfer. One promising approach may be to utilize diverse datasets such as the College and Beyond II (Courant et al., 2022) dataset, which contains student enrollment and outcomes information for a twenty year period from 19 public colleges and seven university systems. However, this dataset does not include enrollment or outcomes data for community colleges (e.g. institutions with two year programs), which are often the ones in the highest need for support. Including those institutions in future analyses or translational research endeavors will be important in ensuring that the benefits which come from transfer learning – if the benefits exist for these institutions! – are being made available to the institutions who need it most.

Beyond the issue of access to predictions, we note that our work here is focused specifically on the *identification* of individuals who are at high risk of dis-enrollment, and does not provide any specific insights on the *intervention* which might be employed to remediate the issue. The question of how to intervene to support student success is a large area of study in its own right, and it's likely that there are numerous interventions available, each with their own cost and opportunity. Increasing the speed at which novel methods of prediction can be paired and tested with novel interventions is likely to result in more transformative research, but requires larger interdisciplinary teams which generally requires substantial funding.

References

Sattar Ameri, Mahtab J. Fard, Ratna B. Chinnam, and Chandan K. Reddy. 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, page 903–912, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340731. doi: 10.1145/2983323.2983351. URL https://doi.org/10.1145/2983323.2983351.

Daniel Andrade-Girón, Juana Sandivar-Rosas, William Marín-Rodriguez, Edgar Susanibar-Ramirez, Eliseo Toro-Dextre, Jose Ausejo-Sanchez, Henry Villarreal-Torres, and Julio

Angeles-Morales. Predicting student dropout based on machine learning and deep learning: A systematic review. *EAI Endorsed Transactions on Scalable Information Systems*, 10(5), Jul. 2023. doi: 10.4108/eetsis.3586. URL https://publications.eai.eu/index.php/sis/article/view/3586.

Lovenoor S. Aulck, Dev Nambi, Nishant Velagapudi, Joshua Evan Blumenstock, and Jevin D. West. Mining university registrar records to predict first-year undergraduate attrition. In *Educational Data Mining*, 2019. URL https://api.semanticscholar.org/CorpusID:195891561.

Girish Balakrishnan. Predicting student retention in massive open online courses using hidden markov models. Master's thesis, EECS Department, University of California, Berkeley, May 2013. URL http://www2.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-109.html.

Rebecca Barber and Mike Sharkey. Course correction: Using analytics to predict course success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, LAK '12, page 259–262, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450311113. doi: 10.1145/2330601.2330664. URL https://doi.org/10.1145/2330601.2330664.

Sebastien Boyer and Kalyan Veeramachaneni. Transfer learning for predictive models in massive open online courses. In Cristina Conati, Neil Heffernan, Antonija Mitrovic, and M. Felisa Verdejo, editors, *Artificial Intelligence in Education*, pages 54–63, Cham, 2015. Springer International Publishing. ISBN 978-3-319-19773-9.

Paul N. Courant, Allyson Flaster, Susan Jekielek, Margaret Levenstein, Timothy A. McKay, and Kevin M. Stange. College and beyond ii (cbii) administrative data, 2022.

Josh Gardner and Christopher Brooks. Student success prediction in MOOCs. *User Model. User-adapt Interact.*, 2018.

Josh Gardner, Christopher Brooks, and Ryan Baker. Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, LAK19, page 225–234, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362566. doi: 10.1145/3303772.3303791. URL https://doi.org/10.1145/3303772.3303791.

Joshua Gardner, Renzhe Yu, Quan Nguyen, Christopher Brooks, and Rene Kizilcec. Cross-institutional transfer learning for educational models: Implications for model performance, fairness, and equity. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1664–1684, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594107. URL https://doi.org/10.1145/3593013.3594107.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19.

MIT Press, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/file/a2186aa7c086b46ad4e8bf81e2a3a19b-Paper.pdf.

Xin J. Hunt, Ilknur Kaynar Kabul, and Jorge Silva. Transfer learning for education data. In *Proceedings of ACM SIGKDD Conference*, 2017. doi: https://doi.org/10.1145/nnnnnnn.nnnnnnn. URL http://ml4ed.cc/attachments/HuntTransfer.pdf.

Sandra C. Matz, Christina S. Bukow, Heinrich Peters, Christine Deacons, Alice Dinu, and Clemens Stachl. Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Scientific Reports*, 13, 2023. doi: https://doi.org/10.1038/s41598-023-32484-w. URL https://www.nature.com/articles/s41598-023-32484-w.

National Center for Education Statistics. Undergraduate retention and graduation rates. https://nces.ed.gov/programs/coe/indicator/ctr, 2022. Accessed: 2023-11-20.

Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual amp;group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '18. ACM, July 2018. doi: 10.1145/3219819.3220046. URL http://dx.doi.org/10.1145/3219819.3220046.

Office of Enrollment Management University of Michigan. Learning analytics data architecture (larc). https://enrollment.umich.edu/data/learning-analytics-data-architecture-larc, 2023. Accessed: 2023-09-01.

Appendix A. Technical details

A.1. Model factors

1. STDNT_FEMALE, dummy variable indicating whether the student is female (1) or not(0). In the student information system student gender is encoded as 1=Female; 2=Male; 3=Unknown, where the value 3 denotes missing values.

2. STDNT_AGE, student age at the beginning of the first semester (in years).

3. HS_GPA_BIN, weight of evidence calculated for high school GPA ranges (0, 2.7), (2.7, 3.3), (3.3, 3.6), (3.6, 3.8), (3.8, 3.9), (3.9, 5).

4. HS_CALC_IND, dummy variable indicating whether the student has completed high school calculus (1) or has not completed it (0).

5. CURR_GPA, the student's grade point average for the first term.

6. GROSS_FAM_INC, weight of evidence calculated for the categories of family income.

7. PRNT_ED_LVL, weight of evidence calculated for the categories of parent education.

8. SNGL_PRNT_IND, dummy variable indicating if a student was raised by the single parent (1) or not (0).

9. SAT_ACT_TOTAL_BIN, weight of evidence calculated for the SAT ranges (0, 1000), (1000, 1200), (1200, 1300), (1300, 1400), (1400, 1600). If the student does not have an SAT score, their ACT score is converted to an equivalent SAT score.

10. STDNT_ETHNC_GRP_CD, weight of evidence calculated for the factor values (1 = White; 2 = Black; 3 = Hispanic; 4 = Asian; 5 = Native Amr; 6 = Not Indic; 7 = Hawaiian; 0 = 2 or More).

11. No_grades_at_all, dummy variable indicating whether the student lacks information about individual course grades (1) or have individual course grades (0).

12. Grade_Overall_I_for_1_and_more_courses, dummy variable indicating if the student has grades for individual courses beginning with "I" (1) or lacks grades starting with "I" (0). "I" means incomplete courses by the deadline.

13. Grade_W_for_1_course, dummy variable indicating whether the student had one official withdrawal (1) or did not (0).

14. Grade_W_for_2_courses, dummy variable indicating whether the student had two official withdrawals (1) or did not (0).

15. Grade_W_for_3_and_more_courses, dummy variable indicating whether the student had three or more official withdrawals (1) or did not (0).

16. Grade_Y_for_1_and_more_courses, dummy variable indicating whether the student had grades marked as "Y" (1) or did not receive such grades (0). "Y" means work in progress.

17. Grade_NR_for_1_and_more_courses, dummy variable indicating whether the student had grades marked as "NR" (1) or did not receive such grades (0). "NR" means work in progress. If a student stops attendance before the end of the term, the instructor is required to report an "NR".

A.2. Hyperparameters

In the experiment, two tasks required the definition of hyperparameters:

1. Logistic regression: The regularization parameter. Due to the substantial size of the source dataset (30,122 points), no penalty was applied.

2. Kernel mean matching procedure: The chosen kernel type - RBF kernel, $B = 1000$, $\epsilon = (\sqrt{m} - 1)/\sqrt{m}$, RBF's length scale = 3.5. The first three parameters were chosen the same as in the study conducted by Huang et al. (2006).

   ϵ **parameter meaning**: In the case of unweighted distribution, where each of the 30,122 source points has a weight of 1, the total sum of these weights remains precisely

30,122. The chosen value for the parameter ϵ indicates that the sum of weights may vary slightly, but it will fall within the range of 30,120.99 to 30,123.01.

The choice of the **RBF's length scale** value was made by analyzing kernel matrices across various length scale values, specifically [1, 1.5, 1.75, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 8, 10, 22]. The kernel function, when used to two points, applies a transformation to the distance between these two points. The length scale parameter defines the sensitivity of this kernel function. When the parameter is set too high, the kernel function transforms distances to values closer to 1. Conversely, when the parameter is set too low, the kernel function transforms distances towards 0. We should choose length scale that generates kernel matrices with values distinguishable from one another. The kernel matrices should encompass a variety of values, not just those very close to 0 or 1. It's worth noting that the most distinct separation between distances occurs when the median distance corresponds to a kernel value of 0.5, positioning the median distance at the midpoint of the possible kernel range (0-1). For our task, this corresponds to a length scale value of 3.5.

A.3. Computing infrastructure

Linux 4.15.0-213-generic, cvxopt 1.3.2, ipykernel 6.21.2, ipython 8.10.0, matplotlib 3.7.0, numpy 1.24.2, pandas 1.5.3, python 3.11.0, scikit-learn 1.2.1, scipy 1.10.1, statsmodels 0.13.5.

Appendix B. Reproducibility Checklist

This paper

- Includes a conceptual outline and/or pseudocode description of AI methods introduced ([✓ yes]/partial/no/NA)
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results ([✓ yes]/no)
- Provides well marked pedagogical references for less-familiare readers to gain background necessary to replicate the paper ([✓ yes]/no)

Does this paper make theoretical contributions? (yes/[✓ no])
Does this paper rely on one or more datasets? ([✓ yes]/no)

- A motivation is given for why the experiments are conducted on the selected datasets ([✓ yes]/partial/no/NA)
- All novel datasets introduced in this paper are included in a data appendix. (yes/partial/[✓ no]/NA)

Author Statement: The datasets used in this paper are not authorized for redistribution.

- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes/partial/[✓ no]/NA)

  Author Statement: The datasets used in this paper are not authorized for redistribution.

- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. (yes/no/[✓ NA])
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. (yes/partial/no/[✓ NA])
- All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. (yes/[✓ partial]/no/NA)

Does this paper include computational experiments? ([✓ yes]/no)

- Any code required for pre-processing data is included in the appendix. ([✓ yes]/partial/no).

  Author Statement: See the blinded Open Science Foundation (OSF) site at https://osf.io/cn5ub/?view_only=aad5ddd9f6f142aabd9bc1e5aa365478. This link is blinded for peer review.

- All source code required for conducting and analyzing the experiments is included in a code appendix. ([✓ yes]/partial/no)

  Author Statement: See the blinded OSF site.

- All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. ([✓ yes]/partial/no)

  Author Statement: The authors of the paper are not institutional signatories and are unable to waive institutional copyright on software works. The authors will seek approval from the appropriate technology transfer department to release the software under a permissive Open Source Foundation (OSF) license, however are unable to do so without institutional approval. As noted, source code is available without license for peer review from the OSF site.

- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/[✓ partial]/no)
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. ([✓ yes]/partial/no/NA)
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. ([✓ yes]/partial/no)

Author Statement: Only moderate computational resources were required, including a 64 core Intel-based machine with 756 gigabytes of RAM and a 1TB disk. All processing was done in python 3 on linux. A full list of packages used in analysis are included in the OSF site with the source code for the project.

- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. ([✓ yes]/partial/no)
- This paper states the number of algorithm runs used to compute each reported result. (yes/[✓ no])

  Author Statement: This work involved exploratory data analysis to examing the distribution shift between populations for feature engineering. This has been excluded from the paper due to space constraints.
- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. ([✓ yes]/no)
- The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). (yes/partial/[✓ no])
- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. ([✓ yes]/partial/no/NA)
- This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. ([✓ yes]/partial/no/NA)