When the Past Misleads: Rethinking Training Data Expansion Under Temporal Distribution Shifts

Chengyuan Yao[1], Yunxuan Tang[2], Christopher Brooks[2], Rene F. Kizilcec[3], Renzhe Yu[1]

[1]Columbia University
[2]University of Michigan
[3]Cornell University

cy2706@tc.columbia.edu, yunxuant@umich.edu, brooksch@umich.edu, kizilcec@cornell.edu, renzheyu@tc.columbia.edu

Abstract

Predictive models are typically trained on historical data to predict future outcomes. While it is commonly assumed that training on more historical data would improve model performance and robustness, data distribution shifts over time may undermine these benefits. This study examines how expanding historical data training windows under covariate shifts (changes in feature distributions) and concept shifts (changes in feature-outcome relationships) affects the performance and algorithmic fairness of predictive models. First, we perform a simulation study to explore scenarios with varying degrees of covariate and concept shifts in training data. Absent distribution shifts, we observe performance gains from longer training windows though they reach a plateau quickly; in the presence of concept shift, performance may actually decline. Covariate shifts alone do not significantly affect model performance, but may complicate the impact of concept shifts. In terms of fairness, models produce more biased predictions when the magnitude of concept shifts differs across sociodemographic groups; for intersectional groups, these effects are more complex and not simply additive. Second, we conduct an empirical case study of student retention prediction, a common machine learning application in education, using 12 years of student records from 23 minority-serving community colleges in the United States. We find concept shifts to be a key contributor to performance degradation when expanding the training window. Moreover, model fairness is compromised when marginalized populations have distinct data distribution shift patterns from their peers. Overall, our findings caution against conventional wisdom that "more data is better" and underscore the importance of using historical data judiciously, especially when it may be subject to data distribution shifts, to improve model performance and fairness.

Code — https://github.com/AEQUITAS-Lab/Distribution-Shift-AIES-2025

Introduction

Machine learning applications have been widely deployed to facilitate decision making in social sectors such as healthcare and education (Dixon et al. 2024; Broby 2022; Sghir, Adadi, and Lahmer 2023). In developing machine learning models, there is a common assumption that larger training datasets would lead to better performance and greater model generalizability. The logic behind this assumption is intuitive: more data covers more diverse underlying patterns, which can help reduce both bias and variance in predictive estimates. However, this assumption depends on the condition that training and test data come from similar distributions, which may not be true in real-world contexts. In fact, recent studies have already suggested that larger training datasets do not always yield better predictions (Leevy et al. 2019). In some cases, smaller but well-curated samples can outperform large datasets when sampling is done thoughtfully (Chang and Krosnick 2009).

In the common situation where historical data is used to train a machine learning model to predict future outcomes, the "more data is better" assumption means a preference for including more data from earlier time periods in addition to recent data to generate predictions for a given future time point. While this strategy aims to improve model robustness, it may introduce outdated patterns that divert from more recent data points. This divergence is known as distribution shift in statistics and machine learning research (Quiñonero-Candela et al. 2009). Two types of distribution shift are commonly identified: *covariate shift*, referring to changes in the distribution of input features, and *concept shift*, referring to changes in the relationship between inputs and outcomes. Rigorous evaluation of the types and consequences of temporal distribution shifts is of practical importance, due partly to their potential negative impacts on model performance and partly to storage and computing costs associated with expanding training data.

Beyond achieving high model performance, predictions of machine learning applications in social contexts also ought to be algorithmically fair (Mehrabi et al. 2021; Kizilcec and Lee 2022), i.e., models perform equitably across different sociodemographic groups when making individual-level predictions. While fairness has been a common component of machine learning research and practice, intersectionality, which considers the compound and unique challenges faced by individuals with multiple marginalized identities, adds complexity to common fairness considerations (Kong 2022). Moreover, distribution shifts can be intertwined with fairness challenges, as different social groups may experience different types and rates of distribution shifts in their data.

Motivated by these concerns, this study systematically examines the impact of training data expansion under temporal distribution shifts on the performance and fairness of machine learning models by addressing the following research questions:

1. How does expanding the historical training window affect model performance under varying degrees and types of temporal data distribution shifts?
2. As the training window expands, how do varying degrees and types of temporal data distribution shifts explain any resulting performance degradation?
3. As the training window expands, how do unequal temporal distribution shifts across groups affect model fairness?

Our work is expected to contribute to prior research and practice in multiple ways. First, we advance the theoretical and empirical understanding of temporal stability of machine learning models through the lens of distribution shifts. Second, we provide empirical evidence for responsible AI by linking temporal distribution shifts to intersectional algorithmic fairness, revealing the complex ways in which data distributions shape fairness outcomes across intersectional groups. Third, we present a reproducible simulation and evaluation framework for expanding-window training under temporal data distribution shifts, with the potential to guide practitioners in data collection and model maintenance for machine learning applications.

Related Work

Distribution Shift

In supervised machine learning, data distribution shift refers to the phenomenon that the data distribution the model is trained on differs from the data distribution the model is tested on. Two commonly studied types of data distribution shift that influence model performance are covariate shift and concept shift, each reflecting a distinct mismatch between the training and testing distributions (Kouw and Loog 2019):

- **Covariate shift** refers to changes in the marginal distribution of the input features $P(X)$, while the conditional distribution $P(Y|X)$ remains invariant.
- **Concept shift** refers to changes in the conditional distribution $P(Y|X)$, indicating that the underlying relationship between features and labels changes.

A wide range of methods has been proposed for detecting data distribution shifts. For covariate shift, detection techniques involve statistical divergence measures (e.g., Kullback-Leibler divergence (Csiszár 1975), Maximum Mean Discrepancy (Gretton et al. 2012)) or hypothesis testing procedures (e.g., Kolmogorov-Smirnov test (Marsaglia, Tsang, and Wang 2003)) to compare $P_{\text{train}}(X)$ and $P_{\text{test}}(X)$. Detecting concept shift is more challenging because it involves estimating changes in $P(Y|X)$, which is not directly observable. Common approaches include model-based methods that track degradation in model performance metrics (Klinkenberg and Joachims 2000) or using error-driven drift detectors such as the Drift Detection Method (DDM) (Gama et al. 2004). However, because our goal is to examine the relationship between data distribution shifts and model performance, using methods that infer shift from the evaluated model's own errors or drift statistics introduces a circular dependency, i.e., the model defines the shift signal, and that signal is then used to explain the model's performance. Therefore, we employ a nonparametric, model-agnostic approach based on *k-nearest neighbors (kNN)* to estimate shifts in $P(Y|X)$ for the empirical study, which is introduced in detail in later sections.

Beyond detection, various strategies have been proposed to mitigate the adverse effects of distribution shifts, including instance reweighting, ensemble learning, and domain-invariant feature learning (Azarkesht and Afsari 2022; Bifet, Holmes, and Pfahringer 2010; Lu et al. 2022). Recent work has also focused on diagnosing model performance degradation attributable to different types of distribution shifts (Cai, Namkoong, and Yadlowsky 2023). While the study provides a framework to examine performance drops between a single source–target pair, our temporal setting considers a sequence of source–target pairs generated by expanding the training window over time.

Predictive Analytics in Education

Predictive analytics has become a widely adopted approach in education. Institutions leverage large-scale data and advanced machine learning techniques to inform decision making (Beaulac and Rosenthal 2019). One notable application is the development of early warning systems, which aim to identify students at risk of academic failure or dropout. These systems typically utilize behavioral, academic, and administrative data to generate timely risk predictions, which allow institutions to implement targeted interventions that support student success (Berens et al. 2019).

Despite the growing adoption of predictive models, relatively little research has examined how the choice and scope of training data influence model performance, particularly under conditions of data distribution shift. Prior work has shown that model outcomes can be highly sensitive to analytical design decisions, such as variable selection, preprocessing, and modeling techniques (Tang et al. 2025). Beyond model specification, the alignment between training and testing data is also critical. A study on transfer learning in educational predictive modeling found that contextual information can be helpful in guiding model selection; in particular, more similar pairs of source and target institutions tend to yield better transfer model performance (Yao, Cortez, and Yu 2025). Other research has examined how the predictive value of features evolves over time and varies across student groups, finding that predictors that are informative at earlier stages can lose importance as new information becomes available in later periods (Glandorf et al. 2024). Another relevant study explored how data distribution shifts during the COVID-19 pandemic affected the performance of retention prediction models and found that, while imperfect, predictive models can still yield useful insights under certain conditions (Xu and Wilson 2021). Extending this line of work, we examine how the interaction between temporal distribution shifts and the expanding use of historical training data

shapes the performance and fairness of predictive analytics.

Algorithmic Fairness and Intersectionality

Algorithmic fairness refers to the principle that machine learning models should yield equitable outcomes across diverse demographic groups (Mehrabi et al. 2021). A growing body of research has assessed whether models trained on the entire population produce systematically biased predictions for certain subgroups, particularly along dimensions such as race, gender, and socioeconomic status (Kizilcec and Lee 2022). Another line of work has examined the role of protected attributes in model development, debating whether their inclusion can improve fairness without introducing additional harms (Yu, Lee, and Kizilcec 2021). Research has also expanded from single-group fairness to intersectional fairness, which considers the compound disadvantages faced by individuals with intersectional marginalized identities (Kong 2022), an important perspective for uncovering disparities that single-axis analyses may overlook but one that remains underexplored in educational predictive modeling.

To address fairness concerns, some studies have proposed fairness-aware modeling approaches that incorporate fairness constraints during training (Hu and Rangwala 2020). While current research has identified structural inequities (Barocas, Hardt, and Narayanan 2023) and data underrepresentation (Bird, Castleman, and Song 2024) as sources of bias in educational prediction models, relatively little attention has been paid to the role of data distribution shifts in shaping algorithmic disparities. In this study, we address this gap by investigating how changes in data distributions contribute to algorithmic bias. Through this lens, we aim to provide novel explanations for observed disparities.

Problem Setup and Methods

Prediction Task

We focus on the common scenario of binary classification (e.g., at-risk or not) under expanding historical training windows. Specifically, predictive models are trained using data from prior time periods, with the training window growing larger as more historical data is added. This setup enables a systematic analysis of performance changes as the training window expands in reverse chronological order.

Measuring Distribution Shift

Covariate Shift We quantify covariate shift separately for continuous and binary features.

For continuous features, we first standardize each variable and apply Principal Component Analysis (PCA) to reduce dimensionality. Let d_{PCA} denote the number of retained principal components. To measure distributional change along each component, we apply the Kolmogorov–Smirnov (KS) test between training and test sets. The KS statistic is defined as:

$$D_j = \sup_z |F_{\text{train},j}(z) - F_{\text{test},j}(z)|$$

where $F_{\text{train},j}$ and $F_{\text{test},j}$ are the empirical cumulative distribution functions (CDFs) of the j-th principal component in the training and test datasets. The value D_j captures the maximum difference between the two CDFs and reflects the marginal distribution shift. We then compute the average KS statistic across all components as a summary measure:

$$\text{CovShift}_{\text{cont}} = \frac{1}{d_{\text{PCA}}} \sum_{j=1}^{d_{\text{PCA}}} D_j$$

For binary features, we first calculate the absolute difference in positive class proportions:

$$\Delta p_i = |p_{i,\text{train}} - p_{i,\text{test}}|$$

where $p_{i,\text{train}}$ and $p_{i,\text{test}}$ denote the proportion of 1s in feature i in the training and test sets.

Next, we compute Cramér's V based on a 2×2 contingency table:

$$\text{Contingency Table}_i = \begin{bmatrix} n^{(i)}_{\text{train},0} & n^{(i)}_{\text{train},1} \\ n^{(i)}_{\text{test},0} & n^{(i)}_{\text{test},1} \end{bmatrix}$$

where $n^{(i)}_{g,v}$ is the count of group $g \in \{\text{train}, \text{test}\}$ with value $v \in \{0, 1\}$. Cramér's V is calculated as:

$$V_i = \sqrt{\frac{\chi^2 / n_{\text{total}}}{\min(r-1, c-1)}}$$

where χ^2 is the Chi-squared statistic from the table, n_{total} is the total number of observations, and $r = c = 2$ are the table dimensions.

The binary covariate shift is defined as the average of the delta proportions and Cramér's V scores across all binary features:

$$\text{CovShift}_{\text{bin}} = \frac{1}{2} \left(\frac{1}{d_{\text{bin}}} \sum_{i=1}^{d_{\text{bin}}} \Delta p_i + \frac{1}{d_{\text{bin}}} \sum_{i=1}^{d_{\text{bin}}} V_i \right)$$

To synthesize continuous and binary covariate shifts into one metric, we calculate a unified covariate shift score as their arithmetic mean:

$$\text{CovShift}_{\text{unified}} = \frac{\text{CovShift}_{\text{cont}} + \text{CovShift}_{\text{bin}}}{2}$$

Concept Shift In the simulation setting, we have access to the ground-truth coefficients $\{\beta\}$ that generate the data. We therefore measure concept shift in an oracle, model-agnostic manner by comparing the conditional label distributions induced by the training and testing coefficients. Specifically, let β_{train} and β_{test} denote the training and testing coefficient vectors, and let $Q_\beta(\cdot \mid x)$ be the conditional label distribution of Y given $X = x$ under coefficients β. For test covariates $\{x_i\}_{i=1}^n$, define

$$\text{ConceptShift} = \frac{1}{n} \sum_{i=1}^{n} \text{JS}\big(Q_{\beta_{\text{train}}}(\cdot \mid x_i)\,,\, Q_{\beta_{\text{test}}}(\cdot \mid x_i)\big),$$

where JS denotes the Jensen–Shannon divergence, a symmetric and bounded measure of dissimilarity between probability distributions. This yields a single scalar summarizing how much the conditional label distributions implied by β_{train} and β_{test} differ. By design, this pointwise comparison does not depend on the marginal feature distribution $P(X)$; hence, the score is unaffected by covariate shift.

However, in real-world data, the true $\{\beta\}$ are unobserved, and using the oracle metric above is not feasible in practice. Therefore, in our empirical study, we adopt a model-agnostic and non-parametric approach based on k-nearest neighbors (kNN) to estimate the conditional distribution. This method allows us to capture local changes in the relationship between features and the target variable without imposing strong functional assumptions.

The rationale for using kNN lies in its locality: by averaging outcomes over nearby points in the feature space, the method approximates the conditional expectation $\mathbb{E}[Y|X = x]$. Unlike parametric models that assume a fixed functional form, kNN can flexibly adapt to complex and potentially non-stationary relationships between features and outcomes.

Specifically, we first project the standardized input features into a lower-dimensional space using Principal Component Analysis (PCA) to improve the stability and efficiency of distance-based computations. In this reduced feature space, we estimate the conditional probability $\hat{p}(x) = P(Y = 1|X = x)$ for each observation by averaging the observed labels of its k-nearest neighbors.

Let $\hat{p}_{\text{train}}$ and $\hat{p}_{\text{test}}$ denote the estimated conditional probabilities from the training and test sets, respectively. We then compute the concept shift score as:

$$\text{ConceptShift}_{\text{JS}} = \text{JS}(\hat{p}_{\text{train}}, \hat{p}_{\text{test}})$$

While our kNN-based approach provides a flexible, non-parametric estimate of the conditional distribution $P(Y|X)$, it is inherently sensitive to changes in the input distribution $P(X)$. That is, when covariate shift is present, the neighborhoods identified by the kNN algorithm may differ between training and test sets, even if the underlying conditional relationship remains stable. As a result, estimated differences in $P(Y|X)$ may conflate genuine concept shift with distortions introduced by covariate shift.

To address this issue, we perform a residualization procedure to isolate the portion of concept shift that cannot be explained by covariate shift alone. Specifically, we regress the raw concept shift scores on the unified covariate shift score:

$$\text{ConceptShift}_{\text{JS}} = f(\text{CovShift}_{\text{unified}}) + \varepsilon$$

where $f(\cdot)$ is the random forest function that captures both linear and non-linear dependencies. The residual term ε represents the unexplained component. The residualized concept shift metric is defined as

$$\text{ConceptShift}_{\text{JS, resid}} = \text{ConceptShift}_{\text{JS}} - \widehat{f}(\text{CovShift}_{\text{unified}})$$

Higher values indicate larger discrepancies in the conditional distributions.

Performance Evaluation

In binary predictions, various performance metrics have been proposed in the literature. Common metrics derived from the confusion matrix (e.g., accuracy, precision, recall) require the selection of a fixed decision threshold. However, our setting involves expanding training windows and temporal data distribution shifts, which makes it challenging to determine a stable or meaningful threshold across time windows. Therefore, we adopt the **Area Under the Receiver Operating Characteristic Curve (AUC)** as our primary evaluation metric. AUC is threshold-independent and summarizes model performance across all possible classification thresholds.

Formally, AUC is defined as:

$$\text{AUC}(f(\theta)) = \int_0^1 \text{TPR}(\text{FPR}^{-1}(t))\, dt$$

where t denotes a decision threshold, $f_t(\theta, x) = 1$ if $f(\theta, x) \geq t$, and TPR and FPR represent the true positive rate and false positive rate, respectively. AUC values range from 0 to 1, with higher values indicating better discriminative ability; an AUC of 0.5 corresponds to random guessing.

To assess fairness across demographic subgroups and capture intersectional disparities, we use the **AUC Gap** (Gardner et al. 2023) as a group fairness metric. The AUC Gap is defined as:

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

where $\mathcal{D}_{k,g}$ and $\mathcal{D}_{k,g'}$ denote the evaluation data restricted to subgroups g and g' respectively. This metric captures the worst-case difference in AUC across subgroups and serves as a conservative indicator of fairness degradation under distribution shifts.

Statistical Analysis

To examine the relationship between data distribution shift and model outcomes, we employ regression analysis as our primary analytical method. Specifically, we use linear regression models to examine how covariate shift and concept shift relate to model performance and fairness.

For model performance:

$$\text{AUC}_i = \beta_0 + \beta_1 \cdot \delta_i + \beta_2 \cdot \theta_i + \beta_3 \cdot (\delta_i \times \theta_i) + \mathbb{I}_{\text{emp}} \cdot \gamma_{j[i]} + \epsilon_i, \quad (1)$$

where AUC_i denotes model performance for unit i. The variables δ_i and θ_i represent covariate shift metric and concept shift metric, respectively. $\mathbb{I}_{\text{emp}}$ is a binary indicator that equals 1 for empirical study and 0 for simulation study, activating school fixed effects $\gamma_{j[i]}$ in the empirical study to account for institutional heterogeneity.

To assess model fairness, we examine the relationship between shift disparities across demographic groups and the AUC gap

$$\text{AUCGap}_i = \beta_0 + \beta_1 \cdot \Delta_i + \beta_2 \cdot \Theta_i + \mathbb{I}_{\text{emp}} \cdot \gamma_{j[i]} + \epsilon_i, \quad (2)$$

where AUCGap_i measures fairness disparity for unit i. The terms Δ_i and Θ_i represent the (max-min) gaps in covariate and concept shift across demographic groups.

Simulation Study

Simulation Process

We design a simulation framework to examine how temporal covariate shift and concept shift affect model performance and fairness under an expanding training window setting. Below, we describe how we simulate each type of shift and outline the experimental scenarios used in our study.

Covariate Shift To simulate temporal covariate shift, we allow the marginal distribution $P(X)$ to vary across time while keeping $P(Y|X)$ fixed. Specifically, we apply a time-dependent mean shift to the continuous features. For each continuous feature x_j, its mean at time t is defined as:

$$\mu_j^{(t)} = \mu_j^{(0)} + \delta_j \cdot \alpha_t^X$$

where δ_j is a feature-specific drift direction, $\alpha_t^X \in [0, 1]$ is a progression parameter that controls the magnitude of the shift over time, and $\mu_j^{(0)}$ is the baseline mean.

Binary features also shift in marginal proportions by adjusting their Bernoulli probabilities over time. For each binary feature x_j, the probability of success at time t is:

$$\pi_j^{(t)} = \pi_j^{(0)} + \rho_j \cdot \alpha_t^X$$

where $\pi_j^{(0)}$ is the baseline probability, ρ_j determines the rate of change, and probabilities are clipped to remain strictly between 0 and 1.

Concept Shift To simulate temporal concept shift, we allow the conditional relationship $P(Y|X)$ to evolve over time. Specifically, we define two sets of coefficients $\beta^{(0)}$ and $\beta^{(1)}$, representing the start and end states of the underlying data-generating process. For each year $t \in \{1, \ldots, T\}$, we interpolate linearly between them:

$$\beta^{(t)} = (1 - \alpha_t)\beta^{(0)} + \alpha_t \beta^{(1)}, \quad \text{with} \quad \alpha_t = \frac{t-1}{T-1}$$

Given features X_t, the log-odds of the binary outcome are computed as

$$\text{logit}(P(Y = 1|X)) = X_t \cdot \beta^{(t)}$$

Simulation Methods To implement the simulation scenarios, we generate synthetic tabular datasets with a consistent feature structure across all time periods.

Continuous features are independently sampled from Gaussian distributions with fixed or time-varying means, depending on the covariate shift condition. Similarly, binary features are drawn from independent Bernoulli distributions with fixed or drifting probabilities, depending on whether binary covariate shift is introduced.

For each time period $t \in \{1, \ldots, 50\}$, we independently generate a dataset $\mathcal{D}_t = \{(X_t^{(i)}, Y_t^{(i)})\}_{i=1}^n$ consisting of 5,000 samples, where each feature vector $X_t^{(i)} \in \mathbb{R}^{19}$ concatenates 15 continuous and 4 binary features. The corresponding binary label $Y_t^{(i)} \in \{0, 1\}$ is sampled from a Bernoulli distribution, with the success probability determined by the logistic model at time t.

Simulation Scenarios

Model Performance To address RQ1 and RQ2, we simulate four scenarios to examine how covariate and concept shifts influence model performance under expanding training windows. In each scenario, data is generated from a common process across the entire population, without group-specific variation. Predictive performance is assessed using AUC on a fixed test time period (i.e., the most recent time period), while the training set is progressively expanded by incorporating additional data from earlier time periods.

- **Scenario (A): No Shift** — Both the feature distribution $P(X)$ and the conditional relationship $P(Y|X)$ remain stationary over time. This serves as a baseline to observe performance under temporal stability.
- **Scenario (B): Covariate Shift Only** — The marginal distribution $P(X)$ changes gradually over time, while the conditional distribution $P(Y|X)$ remains fixed.
- **Scenario (C): Concept Shift Only** — The conditional relationship $P(Y|X)$ changes over time through smooth transitions in model coefficients, while the feature distribution remains constant.
- **Scenario (D): Combined Shift** — Both $P(X)$ and $P(Y|X)$ change over time. This setting reflects the compounded effects of simultaneous covariate and concept shifts.

In all cases, the logistic regression model is trained using data from an expanding window ending at time $t - 1$, and evaluated on a fixed test year t.

Model Fairness To address RQ3, we extend the simulation framework to incorporate group-specific distribution shifts. Our goal is to assess whether covariate or concept shifts lead to disproportionate performance degradation for certain demographic groups, with particular attention to intersectional subgroups.

We consider two fairness-specific simulation scenarios:

- **Scenario (E): Single-Group Shift** — One binary group (e.g., group = 1 for a demographic attribute) experiences either covariate shift or concept shift over time, while the other group remains stationary. This setting allows us to measure whether group-based shifts lead to growing AUC gaps between groups.
- **Scenario (F): Double-Group Shift** — Two binary demographic attributes, G1 and G2, jointly define four intersectional subgroups. One subgroup from each attribute independently undergoes concept shift. We manipulate the *direction* of these shifts—either aligned (same direction) or opposed. This setup enables us to evaluate how the alignment of shifts across demographic dimensions affects fairness, and whether intersectional subgroups suffer more pronounced fairness degradation when exposed to conflicting versus reinforcing shift patterns.

In both cases, the logistic regression model is trained using full data from an expanding window ending at time $t-1$, with performance assessed separately for each subgroup.

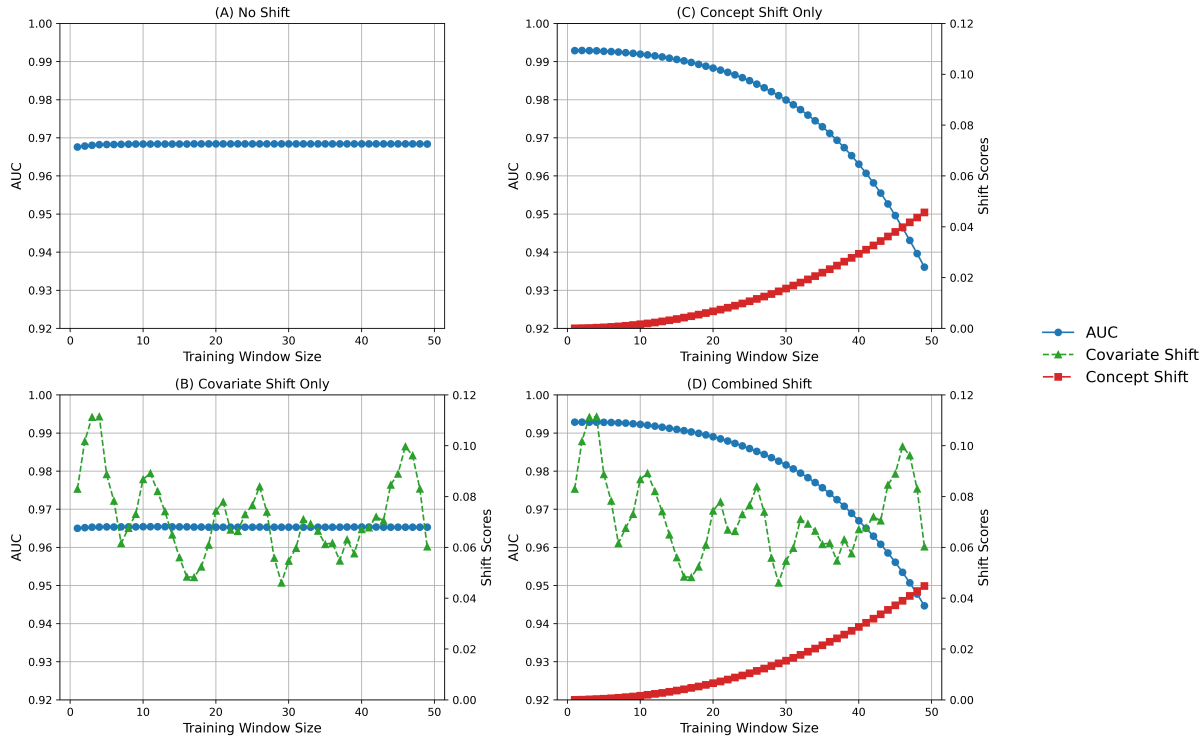

Figure 1: Model performance and data distribution shift metrics across training window sizes. Each panel shows how AUC (left axis) and shift scores (right axis) change as the training window expands in four simulated scenarios: (A) No shift, (B) Covariate shift only, (C) Concept shift only, and (D) Combined shift. Scales are not directly comparable across metrics; only within-metric trends matter.

Results

Model Performance Figure 1 presents model performance and data distribution shift metrics across the four simulated scenarios introduced above. To address RQ1, we examine how expanding the training window size affects model performance under different types of temporal shift.

In the *No Shift* scenario, model performance remains largely flat as the training set increases, with a small improvement at the beginning. In the *Covariate Shift Only* scenario, although the covariate shift metric varies over time, model performance remains stable. Under the *Concept Shift Only* scenario, model performance is stable at first but then consistently declines as the training window expands. Lastly, in the *Combined Shift* scenario, we also observe performance degradation over time, but the magnitude is less severe compared to the *Concept Shift Only* scenario.

To address RQ2, we examine how different types of data distribution shifts relate to changes in model performance as described in Equation 1. The full regression results are presented in the Appendix[1]. Across all simulated scenarios, we find that concept shift has a consistently negative and statistically significant association with model performance degradation. In *Concept Shift Only* scenario, the coefficient for concept shift is $\beta = -0.4040$ ($p < .001$). In the *Combined Shift* scenario, the coefficient for concept shift is $\beta = -0.6651$ ($p < .001$). Covariate shift alone, as examined in *Covariate Shift Only* scenario, shows no significant relationship with model performance ($\beta = -0.0001$, $p = 0.502$), consistent with the flat AUC trend observed under that condition. In addition, in the *Combined Shift* scenario, covariate shift still shows no significant relationship with model performance ($\beta = -0.0205, p = 0.154$). The interaction between concept shift and covariate shift is not statistically significant, suggesting that their joint occurrence does not affect model performance beyond the impact of each shift individually; thus, concept shift is the primary driver of performance degradation in our simulations. One limitation of our simulation design is that covariate and concept shifts are generated independently; in real-world settings, these shifts may interact in more complex and intertwined ways.

Model Fairness To address RQ3, Figure 2 presents the results for the *Single-Group Shift* scenario. In the Concept Shift Only condition, we observe that the subgroup exposed to the shift exhibits a clear performance decline, while the unshifted group shows a gradual performance improvement. This increase occurs because the global model's coefficients are pulled closer to the stable subgroup's true decision boundary as the training window expands. As a result, the fairness metric—*AUC Gap* between the two

[1]Appendix is available at https://github.com/AEQUITAS-Lab/Distribution-Shift-AIES-2025

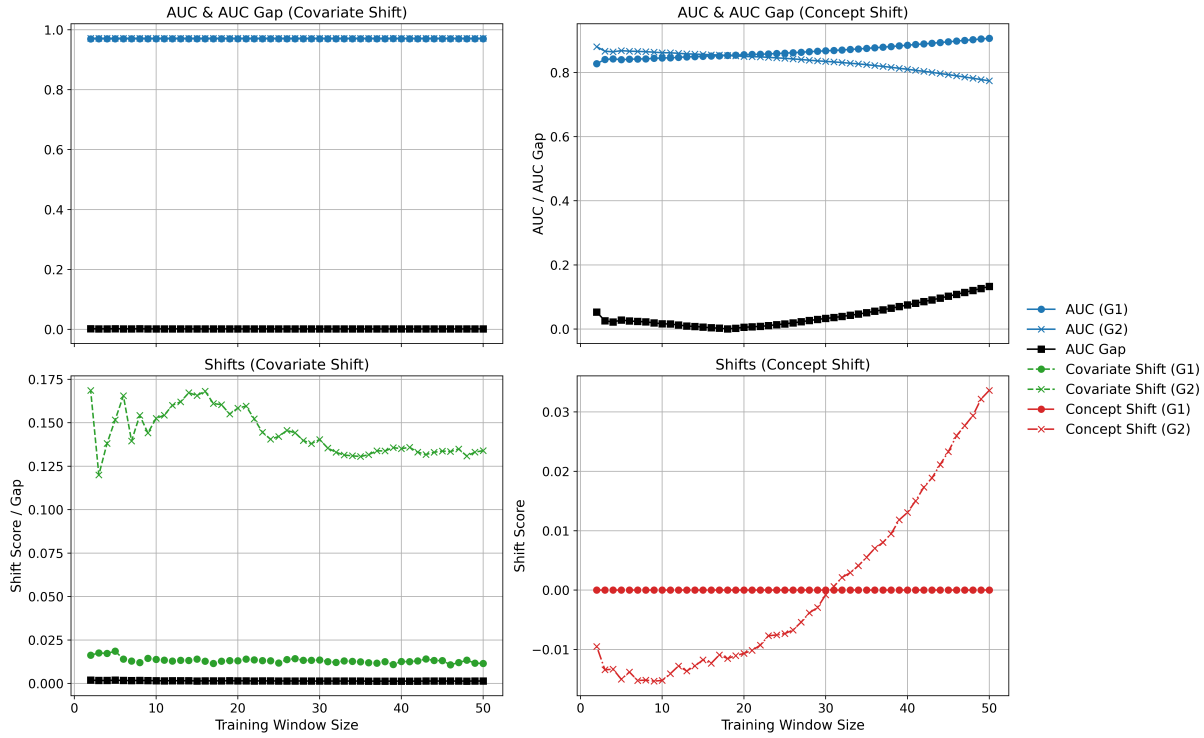

Figure 2: Subgroup model performance and shift metrics across training window sizes. Each panel presents AUC and shift measurements for two demographic subgroups under simulated distribution shift conditions. The left column corresponds to covariate shift only, and the right column to concept shift only. The top row shows subgroup AUCs and AUC gap, while the bottom row displays group-specific shift metrics. Scales are not directly comparable across metrics; only within-metric trends matter.

groups—increases as the training window expands.

In contrast, in the Covariate Shift Only condition, even though one group undergoes greater covariate distributional shift, model performance remains largely unaffected and the *AUC Gap* stays flat. This finding indicates that covariate shift alone does not necessarily compromise predictive parity.

To formally quantify the effects, we conducted regression analyses based on Equation 2. The group-level difference in concept shift metric was a significant predictor of fairness degradation, with a coefficient of $\beta = 2.5089$ ($p < .001$). This result implies that larger inter-group disparities in concept shift are associated with greater disparities in model performance.

We next examined the *Double-Group Shift* scenario to assess how intersectional dynamics influence fairness outcomes. Figure 3 illustrates the impact of concept shift on model fairness when two demographic dimensions are jointly shifted. The effects of intersectional shift are not merely the sum of individual group shifts. When the concept shift directions are aligned across groups, the intersectional group may experience attenuated disparity. In contrast, when shift directions are opposite, the intersectional disparities can become amplified, resulting in a larger AUC gap.

Empirical Case Study

Study Context and Data

Our empirical study focused on predicting student retention in postsecondary education. The central technical objective is to predict first-year retention using the expanding training window approach. First-year retention is defined as whether a student who enters an institution for the first time in a Fall term subsequently re-enrolls at the same institution in the following Fall. This definition aligns with the standard used by the federal government (Gardner 2022).

Through an established research partnership, we obtained access to detailed administrative records from 23 community colleges located within a southern state in the United States. These records included a wide range of student-level data, such as demographic background and academic performance. Community colleges are two-year public institutions that play a critical role in the U.S. higher education system. They serve as a primary access point to postsecondary education for many students from underrepresented and underserved backgrounds, including low-income, first-generation, and minority students. Compared to four-year research universities, community colleges typically have lower retention and completion rates (U.S. Department of Education 2025).

For this study, we restrict the sample to *first-time, first-year* students who entered college during Fall terms from

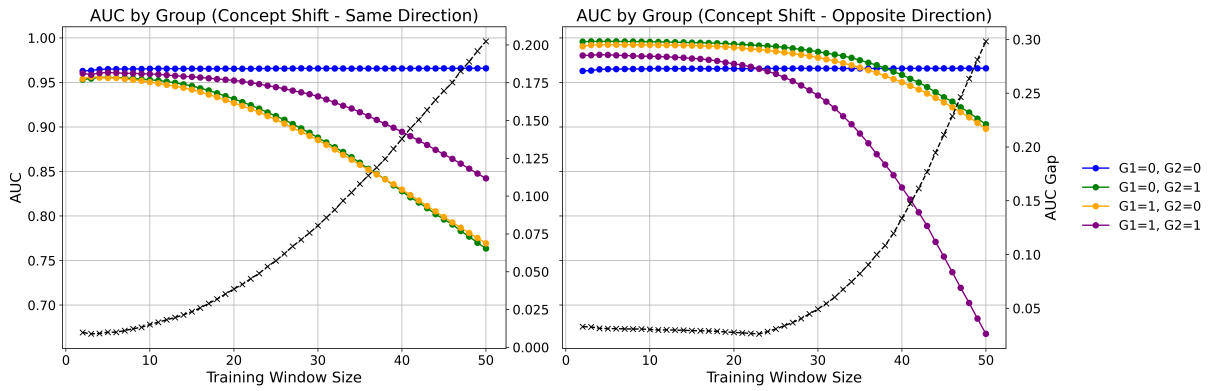

Figure 3: Subgroup AUC and fairness gap under concept shift in same and opposite directions. Each panel shows subgroup-specific AUCs (left axis) and the resulting AUC gap (right axis) across training window sizes. The left panel simulates concept shift in the same direction for all intersectional subgroups, while the right panel simulates concept shift in opposite directions for different subgroups. Lines represent intersectional groups defined by G1 and G2.

2010 to 2021, resulting in a dataset of 1,307,789 students. Among these students, the overall composition comprises 56.7% female students and 31.1% underrepresented minority (URM[2]) students. The largest intersectional subgroup is *Non-URM Female* (37.5%), followed by *Non-URM Male* (31.4%), *URM Female* (18.1%), and *URM Male* (12.9%). To ensure comparability across institutions, we constructed a shared data schema based on commonly available variables. For each college, we designate the most recent year (2021) as the fixed test set and construct expanding training windows by progressively incorporating data from earlier years. The shared data schema is documented in the Appendix.

Results

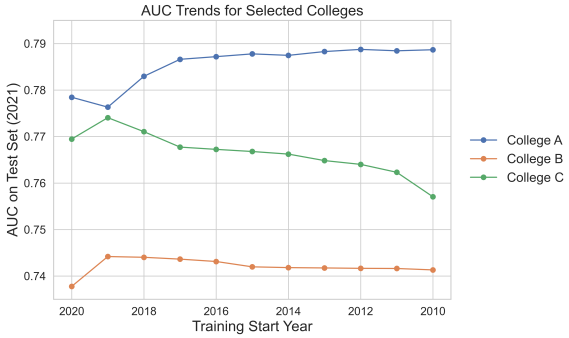

Figure 4: AUC trends by training start year for selected colleges. Model performance (AUC) on the 2021 test set is shown for three representative colleges, plotted against the training start year. Each curve corresponds to an expanding training window that begins in the indicated year and ends in 2020. Full results for all colleges are provided in the Appendix.

[2]URM refers to students who identify as Black/African American, Hispanic/Latino, or American Indian, in line with institutional reporting practices in the United States.

Predictive Performance Figure 4 illustrates representative performance trajectories for three colleges over time. Each line reflects AUC performance on the fixed test set (2021) as the training window expands. The selected colleges exemplify three common patterns observed across the full dataset. College B shows a relatively stable trend, suggesting that accumulating historical data has limited influence on retention prediction. College A demonstrates a steady but modest improvement, indicating that additional training history enhances model performance. College C follows a rise-then-decline pattern, where performance initially improves but eventually deteriorates as older data is added. All remaining colleges conform to one of these three general patterns, with full results reported in the Appendix. These findings show that increasing training data is not universally beneficial. While some colleges benefit from an expanding training window, others see minimal gains or even performance degradation.

Figure 5 presents the covariate shift and concept shift trajectories for the same three exemplar colleges. Covariate shift trajectories are relatively consistent across colleges and show a steady increase as the training window expands. This pattern reflects gradual changes in the marginal distribution of input features over time. By contrast, concept shift trajectories exhibit more variation across colleges. Although the specific patterns differ, the overall trend is upward and indicates increasing divergence in the conditional relationship between features and outcomes as training windows expand. Full results for all colleges are available in the Appendix.

To examine how temporal shift influences predictive performance, we conduct a regression analysis on the pooled dataset across all colleges, following Equation 1. At the full-sample level, none of the examined metrics show a statistically significant main effect on performance. One possible explanation is that the size of the training set conditions the observable impact of concept shift: larger training windows may amplify the effect of even modest shifts, whereas smaller windows may introduce high variance that obscures the influence of substantial shifts. Motivated by this con-

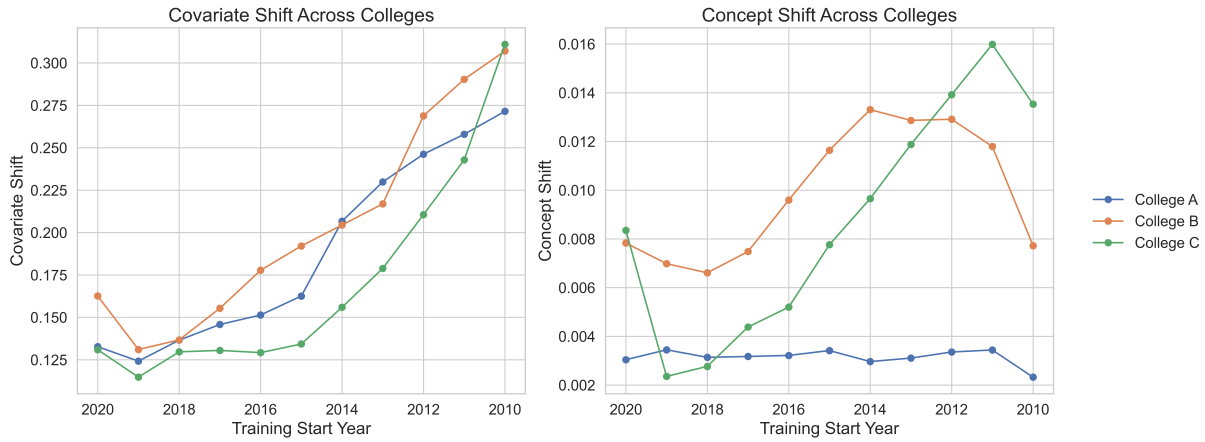

Figure 5: Covariate shift and concept shift by training start year for selected colleges. Magnitudes are computed between each expanding training window and the fixed 2021 test year for three representative colleges. Full results for all colleges are reported in the Appendix.

sideration, we explore whether training size modulates the impact of concept shift on model performance. Specifically, we estimate the regression coefficient capturing the relationship between concept shift and performance across training sets of varying sizes and find that models trained on larger datasets exhibited a more stable and interpretable association between concept shift and predictive performance; detailed results are provided in the Appendix. To investigate this further, we partition the training sets into tertiles by size (Low, Medium, High) and re-estimated Equation 1 separately for each group. The results reveal that the coefficient for concept shift became increasingly negative and statistically significant as the training set expanded. In the largest tertile, concept shift had a statistically significant and pronounced negative effect on performance ($\beta = -0.5412$, $p = .003$). These findings suggest that the adverse effects of concept shift become more detectable when models are trained on sufficiently large historical datasets. The effect of covariate shift was more mixed: it was statistically significant in both the smallest and largest groups, but the corresponding coefficients were relatively small, indicating a modest and inconsistent influence on performance. Notably, the interaction between concept and covariate shift became increasingly positive and statistically significant with larger training sets. These results indicate that the co-occurrence of covariate and concept shift can lead to compounded performance degradation.

Algorithmic Fairness We examine model fairness by constructing four intersectional subgroups based on gender and underrepresented racial minority (URM) status. The model is trained on the full dataset, and subgroup-specific performance and shift metrics were computed using a fixed test year. Figure 6 illustrates representative patterns of both distribution shift and AUC across the four groups. We do not observe a consistent pattern in which any particular group experienced uniformly greater exposure to shift or worse model performance across colleges. This is consistent with our simulation findings that the fairness impact on intersectional groups cannot be understood as a simple additive combination of single-group effects.

To more formally assess the relationship between distribution shift and model fairness, we apply the regression framework outlined in Equation 2. Our analysis shows that the gap in concept shift across groups is a statistically significant predictor of the AUC gap ($\beta = 0.5558$, $p = .002$), suggesting that fairness disparities are more likely to emerge when certain groups undergo substantially different degrees of concept shift. In contrast, the gap in covariate shift is not significantly associated with the AUC gap ($\beta = 0.0055$, $p = .916$). These findings are consistent with our simulation results, which indicate that larger inter-group differences in concept shift lead to more pronounced disparities in model performance.

Discussion and Conclusion

In this study, we examine the issue of predictive analytics under expanding temporal training windows, where earlier historical data is progressively incorporated to train models to predict future outcomes. Using a simulation study and a large-scale empirical analysis in the education sector, we present how the expansion of historical training data interacts with temporal data distribution shifts, specifically covariate and concept shift, to impact both model performance and fairness.

Both simulation and empirical results challenge the conventional assumption that simply increasing the volume of training data from the past would improve model performance. In dynamic environments characterized by distributional changes in historical data, this strategy can result in diminishing returns or even performance degradation. These findings underscore that predictive effectiveness depends not only on the quantity of data, but also on its temporal relevance and alignment with evolving patterns.

Beyond performance, we observe that uneven exposure to temporal concept shifts across sociodemographic groups

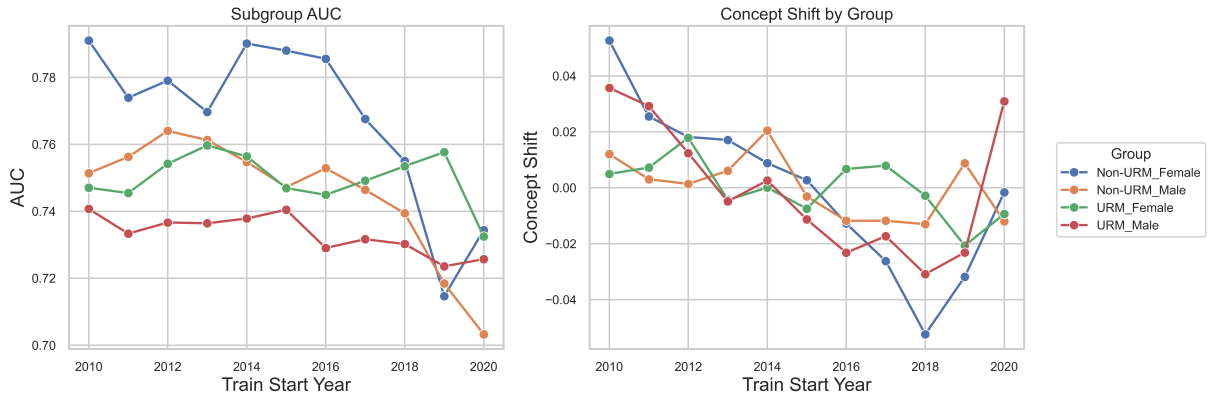

Figure 6: Subgroup AUC and concept shift by training start year for one college. The left panel shows AUC scores for four intersectional subgroups defined by URM status and gender across different training start years. The right panel shows the corresponding subgroup-specific concept shift metrics relative to the fixed 2021 test year. Full results for all colleges are provided in the Appendix.

leads to disparities in subgroup predictive performance. Importantly, among intersectional groups, the effects of multiple identities interact in ways that cannot be understood as a simple sum of individual group effects. These observations align with prior studies highlighting that models that are fair with respect to individual attributes like race or gender may still exhibit unfairness at their intersections (Wang, Ramaswamy, and Russakovsky 2022). As a consequence, fairness-aware modeling should move beyond static group comparisons to account for both temporal and intersectional variation in data conditions.

Our study advances understanding of how distribution shifts shape the performance and fairness of predictive models under expanding training windows. By disentangling the distinct contributions of covariate and concept shift, we demonstrate that the benefits of adding more historical data depend on the nature and magnitude of underlying shifts. We also introduce a reproducible simulation framework capable of generating controlled and decoupled shift scenarios. Furthermore, we extend fairness analysis by showing that inter-group disparities in concept shift can be a key driver of fairness degradation in predictive modeling.

This study has several practical implications. For model developers, our findings emphasize the importance of evaluating distributional changes in training data when expanding historical datasets. While we do not have the capacity to directly identify the point at which additional data no longer enhances model performance, our results demonstrate that temporal distribution shifts can affect performance, which underscores the need for reasonable training dataset selection to reduce potential degradation and to avoid unnecessary data accumulation, storage needs, and computational costs. For institutional researchers, our findings highlight the importance of monitoring subgroup-level performance longitudinally and examining whether emerging disparities are associated with uneven exposure to shifts in the underlying population or behavioral patterns. For fairness researchers, our work extends beyond auditing model outcomes to understanding why models become less fair by offering a data distribution shift perspective as an explanatory lens.

Our study also has several limitations. First, while the simulation design offers strong control over the shift processes, real-world data may involve more complex and interacting shifts, as well as unobserved patterns, that influence both model performance and fairness, as seen in our empirical analysis. Second, although our kNN-based method provides a non-parametric, model-agnostic way to estimate concept shift, it is not without limitations and may be influenced by factors such as feature scaling, high dimensionality, or sparsity in certain regions of the feature space. Future work could explore purely statistical approaches to measuring concept shift that further reduce dependence on specific modeling. Third, although we demonstrate that temporal distribution shift can affect model performance, we do not yet provide an accurate method for identifying the exact point at which additional historical data ceases to be useful. Future research should seek to quantify this threshold and investigate how concept shift interacts with training data scope in ways that make certain historical data detrimental rather than beneficial. Finally, although our dataset spans a large number of institutions, the analysis remains contextually bounded, as it draws from a single state system and focuses on a specific predictive task. Extending this work to other domains that rely on historical data to predict future outcomes—such as employment forecasting, financial risk modeling, or public health monitoring—would help assess the generalizability of our findings.

Acknowledgements

This work was supported by funding from the Learning Engineering Virtual Institute through the Fairness Analysis and Transfer Learning Hub. We extend our gratitude to Dr. Catherine Finnegan for her invaluable data support.

References

Azarkesht, M.; and Afsari, F. 2022. Instance reweighting and dynamic distribution alignment for domain adaptation. *Journal of Ambient Intelligence and Humanized Computing*, 13: 4967–4987.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. The MIT Press. ISBN 9780262048613. Published December 19, 2023.

Beaulac, C.; and Rosenthal, J. S. 2019. Predicting University Students' Academic Success and Major Using Random Forests. *Research in Higher Education*, 60(7): 1048–1064.

Berens, J.; Schneider, K.; Görtz, S.; Oster, S.; and Burghoff, J. 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): 1–41.

Bifet, A.; Holmes, G.; and Pfahringer, B. 2010. Leveraging Bagging for Evolving Data Streams. In Balcázar, J. L.; Bonchi, F.; Gionis, A.; and Sebag, M., eds., *Machine Learning and Knowledge Discovery in Databases*, 135–150. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-15880-3.

Bird, K. A.; Castleman, B. L.; and Song, Y. 2024. Are algorithms biased in education? Exploring racial bias in predicting community college student success. *Journal of Policy Analysis and Management*. First published: January 31, 2024.

Broby, D. 2022. The use of predictive analytics in finance. *The Journal of Finance and Data Science*, 8: 145–161.

Cai, T. T.; Namkoong, H.; and Yadlowsky, S. 2023. Diagnosing Model Performance Under Distribution Shift. arXiv:2303.02011.

Chang, L.; and Krosnick, J. 2009. National Surveys Via RDD Telephone Interviewing Versus the Internet: Comparing Sample Representativeness and Response Quality. *Public Opinion Quarterly*, 73(4): 641–678.

Csiszár, I. 1975. I-Divergence Geometry of Probability Distributions and Minimization Problems. *The Annals of Probability*, 3(1): 146–158.

Dixon, D.; Sattar, H.; Moros, N.; Kesireddy, S. R.; Ahsan, H.; Lakkimsetti, M.; Fatima, M.; Doshi, D.; Sadhu, K.; and Hassan, M. J. 2024. Unveiling the Influence of AI Predictive Analytics on Patient Outcomes: A Comprehensive Narrative Review. *Cureus*, 16(5): e59954.

Gama, J.; Medas, P.; Castillo, G.; and Rodrigues, P. 2004. Learning with Drift Detection. In Bazzan, A. L.; and Labidi, S., eds., *Advances in Artificial Intelligence–SBIA 2004*, volume 3171 of *Lecture Notes in Computer Science*, 286–295. Springer, Berlin, Heidelberg.

Gardner, A. 2022. Persistence and Retention: Fall 2020 Beginning Postsecondary Student Cohort. Technical report, National Student Clearinghouse Research Center, Herndon, VA.

Gardner, J.; Yu, R.; Nguyen, Q.; Brooks, C.; and Kizilcec, R. 2023. Cross-Institutional Transfer Learning for Educational Models: Implications for Model Performance, Fairness, and Equity. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23. ACM.

Glandorf, D.; Lee, H. R.; Orona, G. A.; Pumptow, M.; Yu, R.; and Fischer, C. 2024. Temporal and Between-Group Variability in College Dropout Prediction. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, 486–497.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(25): 723–773.

Hu, Q.; and Rangwala, H. 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)*, 431–437.

Kizilcec, R. F.; and Lee, H. 2022. Algorithmic Fairness in Education. In *Ethics in Artificial Intelligence in Education*. Routledge.

Klinkenberg, R.; and Joachims, T. 2000. Detecting Concept Drift with Support Vector Machines. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, 487–494. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 1558607072.

Kong, Y. 2022. Are "intersectionally fair" ai algorithms really fair to women of color? A philosophical analysis. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 485–494. Association for Computing Machinery.

Kouw, W. M.; and Loog, M. 2019. An introduction to domain adaptation and transfer learning. arXiv:1812.11806.

Leevy, J. L.; Khoshgoftaar, T. M.; Bauder, R. A.; and Seliya, N. 2019. The Effect of Time on the Maintenance of a Predictive Model. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 1891–1896.

Lu, W.; Wang, J.; Li, H.; Chen, Y.; and Xie, X. 2022. Domain-invariant Feature Exploration for Domain Generalization. arXiv:2207.12020.

Marsaglia, G.; Tsang, W. W.; and Wang, J. 2003. Evaluating Kolmogorov's Distribution. *Journal of Statistical Software*, 8(18): 1–4.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)*, 54(6): 1–35.

Quiñonero-Candela, J.; Sugiyama, M.; Schwaighofer, A.; and Lawrence, N. D., eds. 2009. *Dataset Shift in Machine Learning*. Neural Information Processing. The MIT Press. ISBN 9780262545877. Published June 7, 2022.

Sghir, N.; Adadi, A.; and Lahmer, M. 2023. Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022). *Education and Information Technologies*, 28: 8299–8333.

Tang, Y.; Harvey, E.; Yao, C.; Yu, R.; Kizilcec, R.; and Brooks, C. 2025. Understanding Predictive Models of Student Success with a Multiverse Analysis. In Mills, C.; Alexandron, G.; Taibi, D.; Bosco, G. L.; and Paquette, L.,

eds., *Proceedings of the 18th International Conference on Educational Data Mining*, 518–525. Palermo, Italy: International Educational Data Mining Society. ISBN 978-1-7336736-6-2.

U.S. Department of Education. 2025. Community College Facts at a Glance.

Wang, A.; Ramaswamy, V. V.; and Russakovsky, O. 2022. Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 336–349. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Xu, Y.; and Wilson, K. 2021. Early Alert Systems During a Pandemic: A Simulation Study on the Impact of Concept Drift. In *LAK21: 11th International Learning Analytics and Knowledge Conference*, LAK21, 504–510. New York, NY, USA: Association for Computing Machinery. ISBN 9781450389358.

Yao, C.; Cortez, C.; and Yu, R. 2025. Towards Fair and Privacy-Aware Transfer Learning for Educational Predictive Modeling: A Case Study on Retention Prediction in Community Colleges. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, LAK '25, 738–749. New York, NY, USA: Association for Computing Machinery. ISBN 9798400707018.

Yu, R.; Lee, H.; and Kizilcec, R. F. 2021. Should College Dropout Prediction Models Include Protected Attributes? In *Proceedings of the Eighth ACM Conference on Learning @ Scale*, L@S '21, 91–100. New York, NY, USA: Association for Computing Machinery. ISBN 9781450382151.