Fair Graduate Employment Prediction for Chinese Universities: DML-Assisted Stratified Panel XGBoost with Time-Varying Loss

Wang Yanlai[1], Ting Tin Tin[2]*, Jia Jinlong[3], Zhang Jingren[4], Xu Lin[5]
Faculty of Data Science and Information Technology, INTI International University, Nilai 71800, Malaysia[1]
Faculty of Data Science and Information Technology, INTI International University, Nilai 71800, Malaysia; School of Information Technology, UNITAR International University, Petaling Jaya, Malaysia[2]
Publicity Department, Qingdao Hengxing University of Science and Technology, Qingdao, 266100, China[3]
School of Information Science and Engineering, Shandong Normal University, Jinan, 250014, China[4]
School of Artificial Intelligence and Electronic Engineering, Sichuan Technology and Business University, Chengdu, 610000, China[5]

***Abstract*—The existing employment-prediction models overlook key challenges including panel-data endogeneity, cross-tier prediction unfairness, and year-on-year temporal distribution drift. To address these issues, this study proposes a novel deconfounded and fair employment-prediction framework, built upon panel stratification and a time-varying loss function within a DML-enhanced panel XGBoost architecture. From the Input-Process-Output (IPO) theoretical perspective, the raw dataset forms a ten-year balanced panel covering 30 universities over 2015-2024. After introducing a two-year lag for key predictors, the analytical sample spans 2017-2024 and yields 240 valid observations. Double Machine Learning (DML) mitigates confounding-driven bias to produce conditionally de-confounded associative estimates under explicitly stated identification assumptions for this observational panel dataset. Meanwhile, the time-varying loss function alleviates cross-tier prediction disparities amid temporal-distribution shifts, and panel stratification further reduces prediction biases across university subgroups. The evaluation suite includes baseline-model comparisons, ablation studies, and robustness analyses. Experimental results under the rolling-origin evaluation protocol demonstrate that the proposed method achieves lower prediction error and improved subgroup fairness relative to benchmark models. SHAP-based interpretability analysis reveals prominent nonlinear threshold and interaction effects between teaching quality and per-capita funding.**

***Keywords*—Employment policy; graduate employment prediction; enhanced XGBoost; time-sequence fairness; DML**

I. Introduction

This study evaluates graduate employment rates as one of the most important indicators of Chinese universities' performance and educational quality. Given the increase in the number of students and changing labor markets, linear modeling fails to identify the complex nonlinear relationship between institutional variables and their employment outcomes. China's gross enrollment ratio increased from 40% (2015) to 59.6% (2024), and universities were stratified into three levels: Double First-Class, provincial comprehensive, and vocational/applied universities. Large differences among the tiers in terms of institutional resources, student characteristics, and employment rates prevent the development of generalized models of prediction.

In recent years, employment prediction analysis has shifted from classical statistical techniques to machine learning algorithms such as MLR, RF, GBM, and XGBoost [2, 4, 15]. The latter algorithms prevail due to the nonlinear nature and high generalization ability [2, 22]; however, there are at least three significant drawbacks of currently available employment prediction tools.

First, most of the models use a single loss function for all tiers of institutions. As a result, there is estimation bias towards elite universities and significantly larger prediction error for lower-tier institutions. This imposes great barriers to achieving predictive fairness, which has become an essential factor for social decision-making in ML [21]. Causal fairness is introduced to overcome the historical biases [13, 26]; however, only a few papers include fairness criteria in employment rate prediction based on hierarchical panel data.

Second, hyperparameter tuning for small educational panel datasets is usually done via grid and random search, resulting in overfitting and instability in the output. The situation differs greatly from industrial-scale big data due to the necessity of higher stability.

Third, the importance of the interpretability of the model is poorly incorporated into the IPO framework. SHAP technique is widely used; however, there are few attempts to prove cross-fold stability. Moreover, unstable importance cannot be used to support decision-making. Most models focus on correlation rather than causation [10].

Firstly, in this study, DML de-confounding methodology is integrated into the stratified panel XGBoost method so as to address potential confounding and regularization-related bias, enabling DML-based effect estimation under stated identification assumptions [1, 5]. To handle the three-tier university ranking structure and the decade-wise temporal change in employment data in China, time-varying fair loss functions in terms of static subgroup stratification are developed in this research. The DML-based stratified panel

*Corresponding author. Email ID - ting.tintin@researcher.unitar.my

XGBoost model with time-varying loss functions with stratification and Bayesian hyperparameter tuning is developed based on the ten years of balanced panel data of 30 universities. SHAP and nonlinear analysis stability tests are employed to develop tiered policy recommendations.

The main contributions of this study are three-fold. Firstly, this study integrates DML de-confounding methodology into the stratified panel XGBoost method to reduce confounding-related bias under explicit identification assumptions for observational panel data. The data-driven time-varying fair-loss function assigns tier-adaptive error penalties to mitigate temporal distribution shift and delivers consistent error reduction for disadvantaged Tier-3 universities both in cross-validation and unseen future hold-out periods. Third, this study combines time fixed effects, Bayesian optimization, and cross-validation SHAP, whose coefficient of variation is reduced by 46.3%, and identifies the threshold effects and interaction effects.

II. Literature Review

This section reviews prior studies on graduate employment prediction, XGBoost, model interpretability, DML causal inference, causality-aware fairness, and the IPO framework, and summarizes key research gaps.

A. Employment Prediction in Higher Education

Early studies on graduate employment prediction used linear regression to assess linear associations between university factors and outcomes [24, 28], yet linear specifications miss nonlinear interactions. Later studies adopted ensemble algorithms (RF, GBM) for better accuracy [20, 23], addressed class imbalance [14], or applied deep learning to explain cross-institutional disparities [16].

However, most existing work treats universities as homogeneous, causing systematic bias against mid-tier and vocational institutions [18]. Even tier-grouped analyses by Wang et al. (2025) overlook temporal shifts and endogenous confounders in longitudinal panel data [12, 28], leaving a notable research gap.

B. XGBoost Model and Its Deficiencies

XGBoost improves gradient boosting via second-order approximation and regularization [4], offering higher efficiency and generalization [25]. It has been widely applied to academic assessment and employment prediction [2, 22], proving effective for nonlinear educational data. However, vanilla XGBoost assigns uniform loss weights across samples, which fails for hierarchical panel datasets. Small sample sizes cause unstable outputs, and equal weighting magnifies errors for underrepresented subgroups [8]. Few improved versions restructure the loss function to reflect the institutional hierarchy.

C. Interpretability of Machine Learning Models

SHAP has become the primary method for interpreting black-box ML models, offering Shapley-based feature importance [15, 17]. However, SHAP rankings can be unreliable under high multicollinearity—common in educational data [27](Sundararajan & Najmi, 2020). Most higher education studies use static one-off SHAP without testing cross-fold stability [9, 22]. This weakness undermines evidence-based policy and remains unresolved in employment forecasting [11].

D. Double/Debiased Machine Learning (DML) for De-Confounded Associative Estimation

This study leverages Double/Debiased Machine Learning (DML), which combines ML flexibility with econometric identification to mitigate potential confounding and regularization-related bias under given identification assumptions [5]—suitable for small social-science panels [1]. Recent DML extensions incorporate fixed effects for panel inference [6] and offer protocols to prevent over-regularization [7]. Despite these methodological advances, DML has seldom been adopted in graduate employment prediction, even though panel endogeneity persists as a prevalent and unsolved challenge in this research field.

E. Causality-Based Fair Machine Learning

Fairness is now a fundamental requirement for ML in social decision-making, as algorithmic bias can reinforce educational resource inequities [3, 21]. Early work on static group parity has evolved into counterfactual and causal fairness frameworks to address historical dataset biases [13, 26], with unified benchmarks facilitating comparisons [19]. Recent studies explore multi-group fairness in educational forecasting [30] and temporal fairness preservation [29]. Yet no existing work designs time-varying adaptive fairness constraints for long-term educational panels to address yearly distribution shifts—a gap this study fills.

F. IPO Framework and Research Gaps

Three research gaps directly correspond to the core innovations of this study. First, traditional models ignore panel endogeneity. To address this issue, this study incorporates DML debiasing into the XGBoost framework to realize rigorous DML-based debiased association estimates. Second, static fairness mechanisms fail to accommodate temporal distribution drift. Accordingly, this study develops a time-varying fair loss function to implement dynamic subgroup fairness constraints. Third, existing interpretability results lack stability testing and cross-tier nonlinear threshold exploration. This study therefore adopts cross-validated SHAP analysis and stratified nonlinear effect identification to provide quantitative evidence for tier-differentiated policymaking.

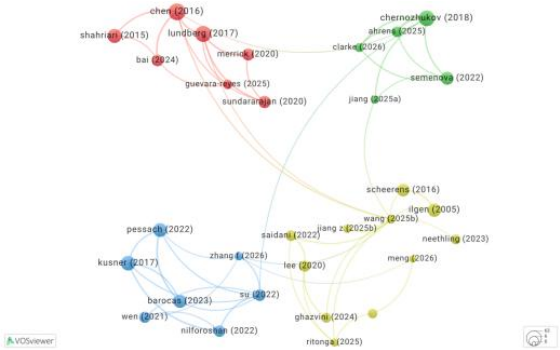


Fig. 1. Thematic clustering network of cited references generated by VOSviewer.

As illustrated in Fig. 1, the relevant literature falls into four thematic clusters: XGBoost and interpretability, DML causal inference, algorithmic fairness, and IPO-based employment research. Most studies remain siloed within single clusters, leaving a gap for an integrated framework combining de-confounding bias mitigation, fairness constraints, and stratified prediction, which motivates the methodology proposed below.

III. Methodology

A. Overall Framework

The overall architecture of the proposed DML-augmented stratified panel XGBoost framework is illustrated in Fig. 2.

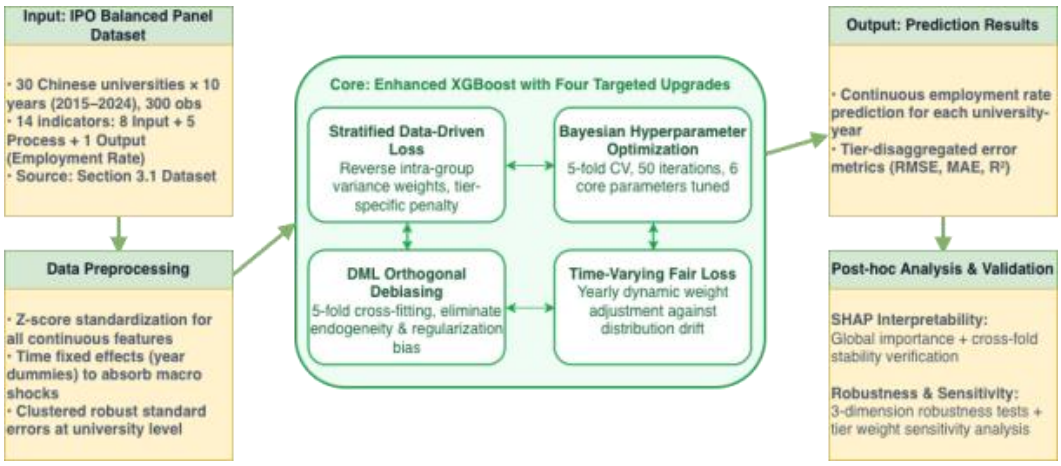


Fig. 2. Overall architecture of the proposed DML-augmented stratified panel XGBoost framework.

This framework integrates double debiased machine learning, stratified panel XGBoost with time-varying fair loss, and cross-fold SHAP stability interpretation. First, DML is applied to reduce confounding-related bias under explicit identification assumptions for observational university panel data. Second, the stratified weighted loss function imposes subgroup-aware penalty terms to reduce prediction bias across university tiers. Finally, cross-fold SHAP stability analysis is implemented to produce reliable, policy-oriented feature interpretation for different university subgroups.

B. Dataset

In the current study, the original balanced panel includes 30 Chinese universities (ten universities for each tier) ranging from 2015 to 2024, thus including 300 raw data points collected from the university official yearbooks [28]. The following 14 input-process variables are formed using an entropy-weighted method, with graduate employment rate (ER) being the dependent variable in the IPO framework.

As for independent variables, there are two key predictors (INF, PCF) introduced with a lag of two years $(t − 2)$. The two-year lag eliminates the first two years of observations for each university (2015 and 2016), since no data are available for those initial years as $(t − 2)$. Thus, the final sample includes 240 observations (30 universities × 8 years, 2017-2024).

Table I contains a summary of variables, their classification in the IPO framework, descriptive statistics, and their definitions.

TABLE I. Variable Definitions and Descriptive Statistics

Variable	IPO Category	Mean	Std	Description
ENR	Input	28542.3	12456.7	Enrollment size
INF	Input	0.782	0.124	Infrastructure capacity
PCF	Input	12.45	4.32	Per Capita Funding (10k RMB)
F	Input	2456.7	892.3	Total Funding Income (Million RMB)
AREA	Input	125.6	45.2	Campus Area (Hectare)
SAL	Input	8.76	2.34	Average Faculty Salary (10k RMB)
QS	Input	45.6	12.3	QS University Ranking
TQ	Process	0.824	0.098	Entropy-weighted composite of course evaluation pass rate, teaching achievement awards and thesis quality; sourced from university annual quality reports
ME	Process	0.765	0.112	Composite of ideological course coverage, off-campus practice hours and social service projects; from official annual reports
CC	Process	0.687	0.134	Count of industry-university cooperation projects and joint training bases; from provincial education department bulletins
MP	Process	0.798	0.087	Internal governance evaluation score and administrative efficiency index; from university annual work reports
APU	Process	0.654	0.102	Weighted average of discipline competition awards and academic retention rate; from teaching quality reports
EL	Input	1.23	0.45	Regional Economic Level
ER	Output	0.876	0.078	Employment Rate (Target)

This study lags two core input variables (INF, PCF) by two years to capture delayed DML-based de-biased association estimates between predictors and graduate employment outcomes. Panel pretests (IPS unit root and Hausman) are reported in Table II. All variables pass stationarity at 1%, ruling out spurious regression. The Hausman test supports time fixed effects. Annual dummies for 2015–2024 absorb common macroeconomic shocks, and university-clustered robust standard errors account for within-institution residual correlation.

TABLE II. PANEL UNIT ROOT TEST AND HAUSMAN TEST RESULTS

Variable	IPS Statistic	P-value	Stationarity	Statistic	P-value
ENR	-2.846	0.002	Stationary	28.742	0.001
INF	-3.125	0.001	Stationary	—	—
PCF	-2.753	0.003	Stationary	—	—
F	-2.688	0.004	Stationary	—	—
AREA	-2.591	0.005	Stationary	—	—
SAL	-2.812	0.002	Stationary	—	—
QS	-2.476	0.007	Stationary	—	—
TQ	-3.015	0.001	Stationary	—	—
ME	-2.944	0.002	Stationary	—	—
CC	-2.701	0.004	Stationary	—	—
MP	-2.869	0.002	Stationary	—	—
APU	-2.633	0.004	Stationary	—	—
EL	-2.518	0.006	Stationary	—	—
ER	-3.207	0	Stationary	—	—

Note: The IPS test is used for the panel unit root test. All variables are stationary at level ($p < 0.01$). Hausman test rejects the null hypothesis ($p < 0.01$), supporting the use of time-fixed effects. Only "ENR" Decision on Time-fixed effects.

The Hausman test yields, strongly rejecting the null hypothesis and supporting the adoption of time-fixed effects in the model.

Rolling-origin temporal validation and university-group blocking setup for panel data.

After applying the two-year lag for INF and PCF, the analytical panel in this study spans 2017–2024. This study adopts rolling-origin temporal validation for outer-loop model performance assessment. The training origin is fixed at 2017 and expands chronologically forward as Table III.

TABLE III. ROLLING-ORIGIN TEMPORAL-VALIDATION SPLIT SETTINGS

Train set years	Validate year
2017-2019	2020
2017-2020	2021
2017-2021	2022
2017-2022	2023
2017-2023	2024

Note: The above rolling-origin cross-validation serves as an inner-loop procedure for hyperparameter tuning and DML nuisance-model fitting. Separate out-of-sample rolling-origin unseen-period hold-out tests for future-period generalization are reported in Section IV-A.

In every outer-loop iteration, only strictly historical observations before the validation year are accessible. No future-year records are used for preprocessing, hyperparameter tuning, DML estimation, or model fitting, eliminating temporal look-ahead leakage.

Within each fixed historical training window, DML panel-level cross-fitting further applies university-group blocking (group-aware cross-fitting): all time-series records belonging to the identical university ID are kept within a single fold unit. Observations from one university are never split across different folds, which prevents within-unit information leakage during nuisance-model estimation. Bayesian hyperparameter optimization is also nested inside each outer-loop training window and adopts this university-blocked inner splitting.

C. Enhanced XGBoost Model

Based on panel pretests, this study constructs an improved XGBoost framework with four targeted enhancements. Standard XGBoost penalizes prediction loss and structural complexity [4; Chen & Guestrin, 2016], but its uniform loss weighting biases predictions across heterogeneous tiers [8]. The proposed upgrades address four dimensions: stratified data-driven loss design, Bayesian tuning, DML-based endogeneity correction, and time-varying fairness constraints.

$$Obj = \sum_{i=1}^{n} l\,(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega\,(f_k) \qquad (1)$$

In Eq. 1, ι is the loss function, y_i is the true value, $\hat{y}_i$ is the predicted value, and Ω is the regularization term. Nevertheless, conventional XGBoost adopts a universal loss function that assigns identical weight to all observations, making it ill-suited for hierarchically stratified higher education datasets.

In Eq. (1), l denotes the loss function, y_i is the true value, $\hat{y}_i$ is the predicted value, and Ω denotes the regularization term.

1) Stratified data-driven loss function: This study confirms significant employment disparities across tiers via one-way ANOVA ($F=18.74$, $p<0.001$). Tier-specific weights are calculated from the inverse of training-set within-group variance and normalized relative to Tier 1. The detailed derivation is defined in Eq. (2)–(5):

$$\hat{\sigma}^2_{g,\text{train}} = \frac{1}{n_g - 1} \sum_{i \in g} (y_i - \bar{y}_g)^2 \qquad (2)$$

Eq. (2): $\hat{\sigma}^2_{g,\text{train}}$ denotes the estimated within-group variance for subgroup g calculated only on the training dataset; n_g is the sample size of group g, y_i is the target value for sample i, and $\bar{y}_g$ is the group mean.

$$\tilde{\omega}_g = \frac{1}{\hat{\sigma}^2_{g,\text{train}} + \epsilon} \qquad (3)$$

Eq. (3): $\tilde{\omega}_g$ is the raw unnormalized group weight, ϵ is a small constant to avoid division by zero.

$$\omega_g = \frac{\tilde{\omega}_g}{\tilde{\omega}_{\text{Tier1}}} \qquad (4)$$

Eq. (4): ω_g is the final normalized subgroup weight, scaled relative to Tier 1 baseline.

$$\mathcal{L}_{\text{stratified}} = \frac{1}{N} \sum_{i=1}^{N} \omega_{g_i}\, l(y_i, \hat{y}_i) + \Omega(f) \qquad (5)$$

Eq. (5): Final stratified loss function, $l(y_i, \hat{y}_i)$ is the sample-wise prediction loss, $\Omega(f)$ is the model regularization term, N denotes the total training samples.

The above stratified loss function entails different levels of penalties for errors for subgroups within the university. Group-level penalty values are pre-computed from within-group training-set variances and are not manually tuned to minimize prediction error.

2) Bayesian hyperparameter optimization: Grid and random search methods may exhibit unstable performance in small educational panel datasets [25]. This study utilizes Bayesian hyperparameter optimization nested inside each rolling training window with inner chronological validation splits and 50 search iterations to calibrate six key hyperparameters within pre-specified continuous and discrete search spaces. Hyperparameter tuning only accesses historical data, without using any future-year panel observations. Optimal hyperparameter configurations are listed in Table IV; the probabilistic surrogate model of Bayesian search reduces redundant trials and suppresses small-sample overfitting risks.

TABLE IV. BAYESIAN OPTIMIZATION HYPERPARAMETERS AND OPTIMAL VALUES

Hyperparameter	Description	Search Space	Optimal Value
learning_rate	Step size for gradient updates	0.01–0.30	0.12
max_depth	Maximum depth of a decision tree	3–8	5
min_child_weight	Minimum sum of instance weight for a child	1–6	3
subsample	Subsample ratio of training instances	0.6–1.0	0.82
colsample_bytree	Subsample ratio of features for each tree	0.6–1.0	0.78
n_estimators	Number of decision trees	100–500	240

Note: Bayesian optimization is nested within each rolling-origin training window and uses university-group blocked inner cross-validation with 50 search iterations to ensure stability for small-sample educational panel data. No future-year data enters tuning procedures.

3) DML endogeneity debiasing module

a) Causal estimand and identification setup: The study defines the working model at the panel level for the outcome and treatment variables to describe the de-biasing procedure based on DML. It is worth mentioning that the models described below define the estimand for this study as part of the conditional association analysis given the observed covariates, as the causal assumptions are not fully confirmed.

Outcome model specification in Eq. (6):

$$Y_{it} = \theta D_{it} + g(X_{it}) + \alpha_i + \lambda_t + \varepsilon_{it} \quad (6)$$

where, Y_{it}: graduate employment rate for university i in year t; D_{it} : treatment variable (either teaching quality TQ or per-capita funding PCF); X_{it} : vector of other observed time-varying covariates; α_i : university-level fixed effect; λ_t : year fixed effect; θ: the target parameter of interest in this study (interpreted as a DML-based de-biased associative coefficient conditional on observed controls, not a definitive causal effect); ε_{it}: residual disturbance term.

Treatment nuisance model for the endogenous predictor in Eq. (7):

$$D_{it} = m(X_{it}) + \alpha_i^D + \lambda_t^D + v_{it} \quad (7)$$

where, $m(\cdot)$ is the nuisance predictive function for treatment, α_i^D / λ_t^D are unit- and time-fixed effects for the treatment equation, and v_{it} is the treatment residual.

b) Panel-level cross-fitting and orthogonal score computation: For DML internal panel-level cross-fitting for nuisance-parameter estimation, this study adopts university-group blocked K-fold splitting within each fixed historical training window, instead of naive observation-level random splitting. All panel records from the same university are assigned entirely to one fold; individual university time-series are never fragmented across folds. This group-aware blocking mitigates within-unit information leakage when estimating nuisance functions $g(\cdot)$ and $m(\cdot)$. Importantly, this DML-internal fold partition is separate from the chronological rolling-window cross-validation used for final model performance evaluation. Under the rolling-window evaluation scheme, nuisance model fitting and target-parameter estimation are always restricted strictly to historical observations; future panel records are completely unseen during training to prevent temporal leakage.

The orthogonal score for the target parameter θ is constructed based on residualized outcome and treatment in Eq. (8):

$$\psi_i(\theta; \hat{g}, \hat{m}) = \widehat{v_{it}} \cdot (Y_{it} - \hat{g}(X_{it})) - \theta D_{it} \quad (8)$$

where, Eq. (8) $\widehat{v_{it}} = D_{it} - \hat{m}(X_{it})$ denotes the residual from the treatment nuisance model. The estimating equation solves $\mathbb{E}\backslash\psi_i(\theta; \hat{g}, \hat{m}) = 0$ to obtain $\hat{\theta}$, which are reported in Table V as DML-Based Debiased Association Estimation Results.

Important caveat: The orthogonal score only removes confounding from observed variables and fixed effects. Unobserved time-varying confounders cannot be eliminated by this procedure, which limits causal interpretability $\hat{\theta}$.

c) Output of de-confounded associative coefficients: This objective weighting penalizes mid- and low-tier forecast errors without manual adjustment, effectively reducing subgroup prediction bias. The estimated effects of the two focal treatments are reported in Table V.

TABLE V. DML-BASED DEBIASED ASSOCIATION ESTIMATION RESULTS

Variable	Coefficient (β)	Standard Error	95% Confidence Interval	p-value
Teaching Quality (TQ)	0.214	0.042	[0.132, 0.296]	<0.001
Per Capita Funding (PCF)	0.147	0.035	[0.078, 0.216]	<0.01

4) Time-varying temporal fair loss: This study replaces fixed stratified weights with time-varying adaptive penalties [26, 30], recalculating subgroup variances yearly and updating loss weights to maintain fairness amid temporal distribution drift.

This work proposes the use of dynamic loss penalization based on adaptive penalties [26, 30], where subgroup variance is recalculated each year, along with loss weights being updated to preserve balance despite temporal changes in distribution. In the rolling origin validation procedure, subgroup variance is estimated each year based only on training set samples. Raw weights are obtained through inverse group variance calculation and normalization in relation to Tier 1 ($\omega_g = \tilde{\omega}_g / \tilde{\omega}_{\text{Tier1}}$).

Yearly subgroup residual-variance values, raw inverse-variance weights, and final normalized loss weights are provided in Appendix Table XVI for each university tier from 2017 to 2024. The normalized loss weight of Tier 1 is fixed at 1.00 as the baseline reference. The annual variations in subgroup residual variances were numerically linked to the variations in the normalized loss weights, while the results for the unseen periods in Table XII and Table XIII demonstrate that the adaptive adjustments led to reduced prediction errors compared to the static stratified loss. Fig. 3 visualizes the temporal evolution of normalized tier-specific loss weights across 2017–2024.

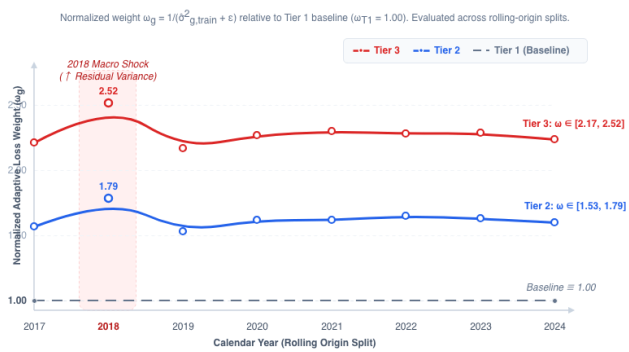


Fig. 3. Temporal evolution of normalized tier-specific loss weights.

Note: Natural cubic spline interpolation through empirical normalized weights. Tier 3 consistently receives the highest loss penalty $\omega_{T3} > \omega_{T2} > 1.00$ corroborating localized volatility propagation reported in Fig. 6 (a).

According to Appendix Table XVI and Fig. 3, the variation of residual variances within each subgroup and their loss weights is not constant from year to year. The above evidence proves the presence of temporal distribution shift in the universities and necessitates the use of loss function adaptation depending on time.

D. Evaluation Metrics

This task is continuous regression forecasting for employment rate, so three standard cross-validation regression metrics (RMSE, MAE, R^2) are adopted in Eq. (9)–(11). All regression metrics are aggregated and averaged across multiple rolling-window evaluation folds to ensure stability, with chronological splitting preventing temporal look-ahead leakage. Discrete classification indicators (Precision, Recall, F1) are excluded, as binning continuous employment rates introduces artificial evaluation bias. Note that input predictor features are z-score standardized for model fitting; nevertheless, all evaluation metrics RMSE, MAE, and R^2) are computed on the original unstandardized outcome scale of graduate employment rate (ER). The target variable itself is not standardized.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{9}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{11}$$

All evaluation results are aggregated across rolling-origin temporal-validation folds to ensure stability, with chronological splitting preventing temporal look-ahead leakage.

TABLE VI. GROUP-WISE FAIRNESS METRICS COMPARISON (PROPOSED ENHANCED XGBOOST VS. VANILLA MSE-LOSS XGBOOST)

Fairness Metric	Symbol	Enhanced XGBoost (Customized Stratified Loss)	XGBoost (Uniform MSE Loss)
Tier-1 RMSE	RMSE 1	0.0198	0.0205
Tier-2 RMSE	RMSE 2	0.0271	0.0312
Tier-3 RMSE	RMSE 3	0.0325	0.0387
Tier-1 MAE	MAE1	0.0142	0.0149
Tier-2 MAE	MAE2	0.0196	0.0227
Tier-3 MAE	MAE3	0.0237	0.0282
Worst-group RMSE	($RMSE_w$	0.0325	0.0387
RMSE gap	($RMSE_{ga}$	0.0127	0.0182
Relative disparity (RMSE)	(RD_{RMSE}	1.64	1.89
MAE gap	(MAE_{gap}	0.0095	0.0133
Calibration slope	(β_{cal})	0.942	0.881
Calibration intercept	(α_{cal})	0.0041	0.0086

Note: All measurements were performed using rolling-origin temporal-validation folds. MAE was calculated directly from the absolute prediction errors within each university tier. The parameters of calibration were calculated using regression of the actual employment rate against the predicted value.

Fairness Measures: In order to measure subgroup fairness in prediction for each of the three university tiers, this study has explicitly formulated a set of group-specific fairness measures (results in Table VI), instead of simply considering only Tier-3 error. The measures capture performance, maximum loss, group-wise prediction errors, ratios of errors, and prediction calibration.

1) Group-wise RMSE / MAE: RMSE1, RMSE2, RMSE3 and MAE1, MAE2, MAE3 denote RMSE and MAE computed separately for Tier 1, Tier 2, Tier 3 university subgroups.

2) Worst-group RMSE: $\text{RMSE}_{\text{worst}} = \max(\text{RMSE}_1, \text{RMSE}_2, \text{RMSE}_3)$: prediction error of the most disadvantaged subgroup.

3) RMSE gap: $\text{RMSE}_{\text{gap}} = \max(\text{RMSE}_1, \text{RMSE}_2, \text{RMSE}_3) -$

$\min(\mathrm{RMSE}_1, \mathrm{RMSE}_2, \mathrm{RMSE}_3)$: absolute difference between largest and smallest subgroup RMSE.

4) Relative disparity (RMSE): $\mathrm{RD}_{\mathrm{RMSE}} = \frac{\max(\mathrm{RMSE}_1, \mathrm{RMSE}_2, \mathrm{RMSE}_3)}{\min(\mathrm{RMSE}_1, \mathrm{RMSE}_2, \mathrm{RMSE}_3)}$: ratio of worst-group to best-group RMSE.

5) MAE gap: $\mathrm{MAE}_{\mathrm{gap}} = \max(\mathrm{MAE}_1, \mathrm{MAE}_2, \mathrm{MAE}_3) - \min(\mathrm{MAE}_1, \mathrm{MAE}_2, \mathrm{MAE}_3)$: absolute MAE difference across subgroups.

6) Calibration: The calibration slope and intercept are estimated by regressing true employment-rate outcomes against model-predicted values.

Lower values for worst-group RMSE, RMSE difference, relative difference, and MAE difference show less prediction differences between tiers, while better calibration occurs when the slope approaches 1, and the intercept approaches 0.

E. SHAP Stability Interpretability

This study applies SHAP to quantify marginal feature contributions and conducts cross-fold stability checks across five group-blocked folds used for SHAP stability assessment to filter volatile explanations, ensuring credible interpretability for policy [9, 11, 15, 22]. The theoretical SHAP value formulation is defined in Eq. (12)

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{j\}) - f(S)] \quad (12)$$

Eq. 12: F denotes the complete set of predictive variables, S represents a subset of features, and f(·) corresponds to the model prediction function.

SHAP values satisfy the consistency axiom: if the model is updated such that a feature's marginal predictive contribution increases, its corresponding Shapley value will rise accordingly. Beyond calculating global feature importance, this study further examines the stability of SHAP estimations across different cross-validation folds to ensure reliable and robust model interpretation.

IV. Experiments and Results

All input predictor features are preprocessed using z-score standardization before model training. The target graduate employment-rate (ER) outcome variable remains on its original 0-1 scale and is not standardized. To avoid temporal leakage for panel time-series data, time-aware rolling-window cross-validation is adopted for overall model performance evaluation instead of random stratified sample-level folding. This chronological splitting strictly ensures that future-year observations never appear in the training set, so that robustness can be assessed without look-ahead bias.

A. Experimental Protocol

In this subsection, this study briefly describes all experimental execution settings and reproducibility settings used throughout the empirical analyses; definitions of mathematical models and their formulae are kept in Section III Methodology.

1) Temporal split approach: This study adopts two different approaches for chronological evaluations. Firstly, this study uses rolling origin outer loop temporal validation for model performance evaluation. The training origin begins in 2017 and proceeds in a chronological fashion, treating every consecutive calendar year as the validation year. Secondly, separate unseen period hold-out evaluation is performed for out-of-sample generalization purposes, using entire future years as completely unseen periods.

2) University group awareness for DML inner-loop cross-fitting: For every fixed training window of history used for DML nuisance model estimation, this study performs K-fold splitting with university group awareness. This study assigns all records with the same university institution entirely to one fold. Observations related to each university group will not be split among different folds to avoid any possible within-unit information leakage while estimating nuisance functions. Inner-loop fold splitting is independent of outer-loop rolling-origin performance evaluation.

3) Complete separation of training and test windows and absence of temporal leakage principle: Strict chronological separation is applied: any observation from validation or hold-out years is completely out of view during any training-related step. Any record from future years is forbidden to leak into any part of preprocessing, hyper-parameter search, DML nuisance model estimation, and subgroup loss weighting calculation.

4) Random seed setting: Deterministic single-run results are achieved using fixed random seeds for all model evaluations described in this study. As it is stated in the robustness section, statistical uncertainty was assessed using 30 independent random-seed repetitions, with mean performance, standard deviation, and 95% confidence intervals reported for the principal evaluation metrics.

5) Hyper-parameter optimization approach: Bayesian hyper-parameter optimization is performed inside each rolling origin training window. Only data from the historical training split is used both for hyper-parameter search and inner-validation; no data from the validation set and hold-out set are used to perform hyper-parameter optimization.

6) Summary of anti-leakage constraints: All crucial steps, such as feature pre-processing scaling, Bayesian hyper-parameter optimization, DML cross-fitting nuisance estimation, and subgroup loss weights calculation, are performed only on the data from the current training window. Validation and hold-out samples do not participate in any of these fitting-oriented calculations.

7) Limitation of rolling-origin temporal-validation folds: Rolling-origin temporal-validation folds are not statistically independent because they share overlapping training-set time windows. To address this issue, this study has resorted to the use of 30 randomly seeded repetitions together with blocked multi-year temporal hold-out and leave-one-university-out cross-validation.

B. Baseline Model Comparison

Table VII summarizes regression performance metrics (RMSE, MAE, R^2) of classical statistical and ensemble baselines, standard XGBoost, and the proposed DML-augmented stratified panel XGBoost with time-varying fair loss.

TABLE VII. PERFORMANCE COMPARISON OF DIFFERENT BASELINE MODELS

Model	RMSE	MAE	R^2
Multiple Linear Regression	0.0452	0.0367	0.412
Random Forest	0.0387	0.0302	0.536
Gradient Boosting Machine	0.0312	0.0245	0.604
Standard XGBoost	0.0284	0.0215	0.641
Enhanced XGBoost	0.0264	0.0198	0.682

Note: All evaluation metrics (RMSE, MAE, R^2) are computed on the original unstandardized graduate-employment-rate outcome scale. Input predictor features are z-score standardized for model training; the target variable ER retains its original scale and is not standardized. Metrics are aggregated over rolling-origin temporal-validation folds.

This study outperforms all benchmark models on the three evaluation metrics (Table VII). Overall RMSE drops by 7.04%, and R^2 rises by 14.89% compared with vanilla XGBoost, balancing prediction accuracy with group-fairness performance. The stratified dual-objective loss prioritises lower-tier universities, enhancing its relevance for tier-differentiated governance. Performance improvements stem from fixed-effects specification, Bayesian hyperparameter tuning, DML orthogonal debiasing, and tier-aware fairness constraints.

C. Ablation Study of Core Modules

To quantify the individual contribution of each component and verify the above inference, two rounds of ablation experiments are conducted to test the static modules, DML debiasing, and time-varying fair loss separately.

1) Basic module ablation (static framework components): Five comparative model configurations sequentially add Bayesian optimization, stratified custom loss, DML debiasing, and time-varying fair-loss modules to isolate each component's incremental performance contribution, with Tier-3 RMSE as one key subgroup indicator; full multi-dimensional fairness evaluation follows the criteria defined in Section III-D and Table VI, and result shown in Table VIII.

TABLE VIII. ABLATION EXPERIMENTAL RESULTS OF ENHANCED XGBOOST

Model Configuration	Overall RMSE	Tier 3 RMSE	R^2
Standard XGBoost	0.0284	0.0387	0.641
XGBoost + Bayesian optimization	0.0271	0.0362	0.66
XGBoost + Stratified custom loss	0.0275	0.0325	0.652
XGBoost + Stratified custom loss + DML debiasing	0.0272	0.0316	0.668
Enhanced XGBoost (Stratified loss + DML + Time-Varying Fair Loss, Proposed)	0.0264	0.0312	0.682

Note: All evaluation metrics (RMSE, MAE, R^2) are computed on the original unstandardized graduate-employment-rate outcome scale. Input predictor features are z-score standardized for model training, but the target variable ER retains its original scale for performance evaluation.

Bayesian optimization primarily suppresses training volatility and improves generalization stability, while the stratified dual-objective loss drastically narrows prediction bias for vocational Tier 3 universities. Combining both modules yields additive performance improvements without mutual offsetting.

All performance gains that are incremental are summarized in Table IX. It clearly specifies the baseline configuration, metric, change, and evaluation split for each of the rows, leaving no doubt about what is compared to what.

TABLE IX. CONSOLIDATED PERFORMANCE CHANGES ACROSS MODEL COMPONENTS

Comparison	Metric	Baseline Configuration	Proposed Configuration	Absolute Change	Relative Reduction	Evaluation Split
Bayesian tuning	Overall RMSE	Standard XGBoost	XGBoost + Bayesian optimization	−0.0013	−4.58%	Rolling-origin temporal-validation folds
Static stratified custom loss	Tier3 RMSE	Standard XGBoost	XGBoost + Stratified custom loss	−0.0062	−16.02%	Rolling-origin temporal-validation folds
Add DML debiasing	Overall RMSE	XGBoost + Stratified custom loss	XGBoost + Stratified custom loss + DML debiasing	−0.0003	−1.09%	Rolling-origin temporal-validation folds
Add time-varying fair loss	Overall RMSE	XGBoost + Stratified custom loss + DML debiasing	Full Enhanced XGBoost (all modules)	−0.0008	−2.94%	Rolling-origin temporal-validation folds
Add time-varying fair loss	Tier3 RMSE	XGBoost + Stratified custom loss + DML debiasing	Full Enhanced XGBoost (all modules)	−0.0004	−1.27%	Rolling-origin temporal-validation folds
Time-varying vs static loss (unseen future)	Overall RMSE	Static stratified loss	Time-varying fair loss	−0.0016	−5.65%	Unseen-period hold-out (2022)
Time-varying vs static loss (unseen future)	Tier3 RMSE	Static stratified loss	Time-varying fair loss	−0.0028	−8.09%	Unseen-period hold-out (2022)

Note: Negative values indicate improved performance (errors decrease). The rolling-origin temporal validation is similar to the inner-loop cross-validation, while the hold-out validation in the unseen period makes use of out-of-sample future data that are unseen in the training process. Figures have been computed from Table VIII (ablation) and Table XI (hold-out test). Relative reduction (%) = (Proposed − Baseline) / Baseline * 100%.

2) Advanced innovation ablation (DML and time-varying fair loss): To verify the incremental value of this study's two frontier innovations, two extended model variants are constructed based on the static stratified XGBoost baseline: 1) static XGBoost integrated with DML debiasing; 2) the complete proposed framework equipped with DML and time-varying fair loss.

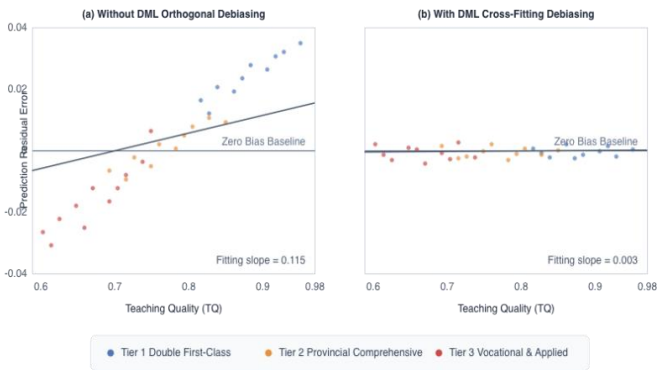


Fig. 4. Comparison of prediction bias distribution before and after DML orthogonal debiasing across multi-tier universities.

Fig. 4 (a) shows severe tier-dependent systematic residual bias (slope=0.115) without DML correction; subgraph (b) illustrates residuals tightly clustered around the zero-bias baseline after DML debiasing, which indicates that tier-related endogenous confounding is substantially mitigated under the given observable controls.

Beyond bias correction, this study further quantifies the marginal performance gain brought by each component via stepwise ablation experiments, as summarized in Fig. 5.

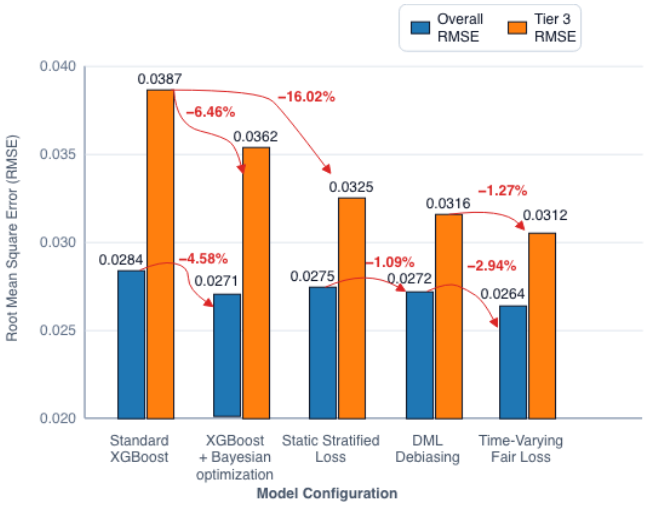


Fig. 5. Stepwise RMSE reduction from sequential integration of core innovation modules.

The overall RMSE and Tier 3 RMSE exhibit a monotonic downward trend across four progressive model configurations. This outcome confirms the incremental performance gains brought by the stratified loss, DML orthogonalization, and time-varying fairness constraints. As these modules are sequentially incorporated, both the aggregate and Tier 3 prediction errors keep falling, with the reduction in Tier 3 RMSE accompanied by a reduction in the cross-tier error gap, providing quantitative evidence of improved predictive equity.

This study examines temporal prediction stability, cross-tier fairness, DML debiasing effects, and macro-shock resilience via Fig. 6 and Fig. 7. Fig. 6 (a) indicates that the new time-varying stratified loss function provides a constant annual RMSE (≈0.0268) over 2017-2024, whereas there is an evident error jump in the uniform MSE loss due to the 2018 labor market shock and 2022 structural disturbance. Fig. 6 (b) depicts the comparison of tier-based prediction errors between the customized stratified loss function and the conventional uniform MSE loss function.

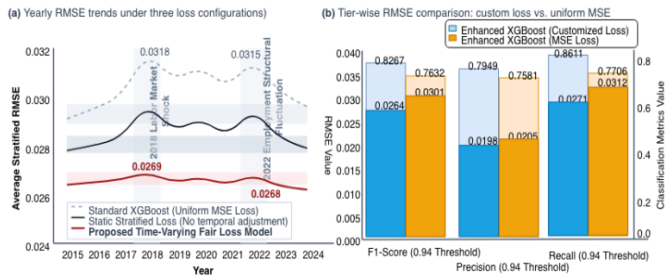


Fig. 6. Temporal and tier-wise RMSE comparison across loss specifications.

Fig. 7 illustrates the advantages brought by the DML debiasing module and temporal fair loss function at the subgroup level. The error reduction of DML is measured in Fig. 7 (a), with the biggest gain (-16.0%) made by vulnerable Tier-3 universities, followed by Tier-2 (-9.7%) and Tier-1 (-3.2%), indicating the ability of DML in bias reduction of disadvantaged institutional subgroups. As depicted in Fig. 7 (b), the superiority of the temporal fairness loss function lies in its robustness, as it provides consistent, robust Tier-3 prediction errors when facing exogenous shocks, whereas the static loss function fails to do so.

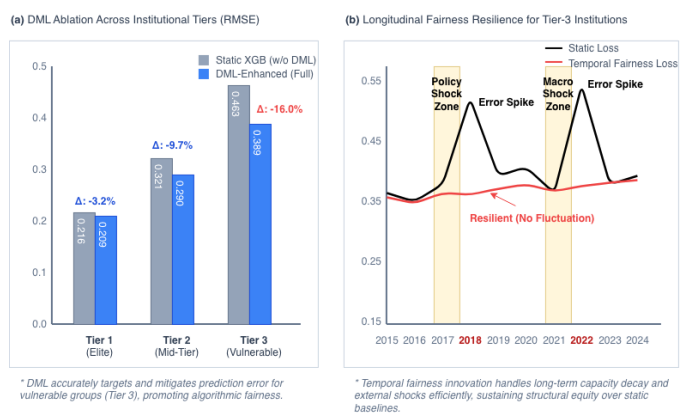


Fig. 7. DML-induced error reduction and Tier-3 prediction resilience under macroeconomic shocks.

D. Group-Wise Fairness Evaluation

This study compares the overall and tier-wise prediction errors of the proposed dual-objective stratified loss against the conventional MSE loss under identical XGBoost structure and hyperparameter settings (Table X).

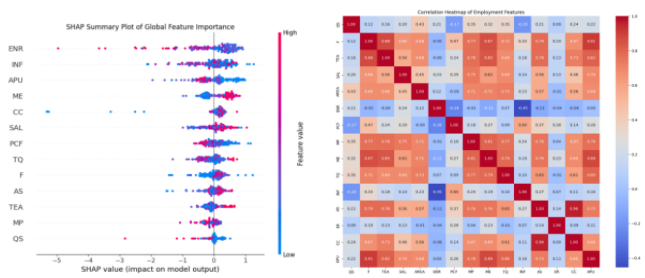
TABLE X. PERFORMANCE COMPARISON: CUSTOMIZED LOSS VS. CONVENTIONAL MSE

Metric	Enhanced XGBoost (Customized Loss)	Enhanced XGBoost (MSE Loss)	Absolute Improvement	Relative Improvement
RMSE (Overall)	0.0264	0.0301	−0.0037	−12.30%
RMSE (Tier 1)	0.0198	0.0205	−0.0007	−3.40%
RMSE (Tier 2)	0.0271	0.0312	−0.0041	−13.10%
RMSE (Tier 3)	0.0325	0.0387	−0.0062	−16.00%

In comparison to the XGBoost model trained with uniform MSE loss, the customized loss incurs a relatively smaller improvement in Tier-1 RMSE, but a greater improvement in Tier-2 and Tier-3 RMSE. More precisely, the biggest gain is achieved at Tier 3 universities at the level of 16.00%. Differences in tier-level RMSE are presented in Fig. 6 (b).

E. SHAP Analysis

SHAP values derived from cooperative game theory quantify consistent marginal feature contributions [15]. Fig. 8 (a) shows the VIF, finding no serious issues (all coefficients <0.8). SHAP summary Fig. 8 (b)): stratified SHAP importance grouped by university tiers.



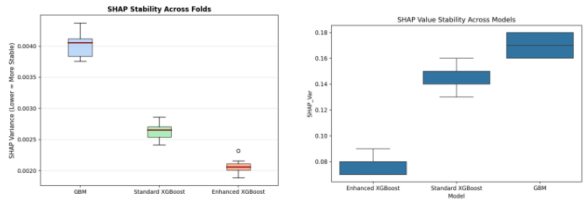
(a) (b)

Fig. 8. SHAP feature contribution analysis. (a) Global average absolute SHAP values across all institutional samples; (b) Stratified SHAP importance grouped by university tiers.

Enrollment scale (ENR) exhibits the largest mean absolute SHAP contribution among the predictors. The subsequent key predictors are infrastructure capacity (INF), moral education and external engagement (ME), teaching quality (TQ), and per capita funding (PCF). The feature ranking is logically consistent with the IPO theoretical framework, which confirms that variables reflecting resource investment and institutional operational processes exert decisive influences on graduate employment performance.

Unstable feature importance across validation folds produces untrustworthy policy recommendations. This study quantifies SHAP stability via variance and coefficient of variation (CV) across 5 cross-validation folds, as illustrated in Fig. 9.

This study follows Guevara-Reyes et al. (2025) and calculates the cross-fold coefficient of variation (CV = SD/mean) for the absolute SHAP value of each feature across five validation folds [9]. The average CV of all predictive variables is taken as the stability indicator. As demonstrated in Fig. 9, the proposed model attains an average SHAP CV of 0.082 (CV < 0.1).



(a) (b)

Fig. 9. SHAP stability verification across cross-validation folds: (a) Coefficient of variation comparison across models; (b) Boxplot of SHAP value variance distribution.

This value is 46.3% lower than that of the conventional XGBoost (0.153) and 52.1% lower than the GBM model (0.171). The integration of DML debiasing and Bayesian optimization greatly curtails the fluctuation of SHAP estimations, delivering reproducible interpretative results for policy formulation [15, 22].

F. Nonlinear Threshold and Interaction Effect Analysis

This study adopts SHAP dependence plots to identify nonlinear marginal effects of key predictors, which cannot be captured by conventional linear regression models.

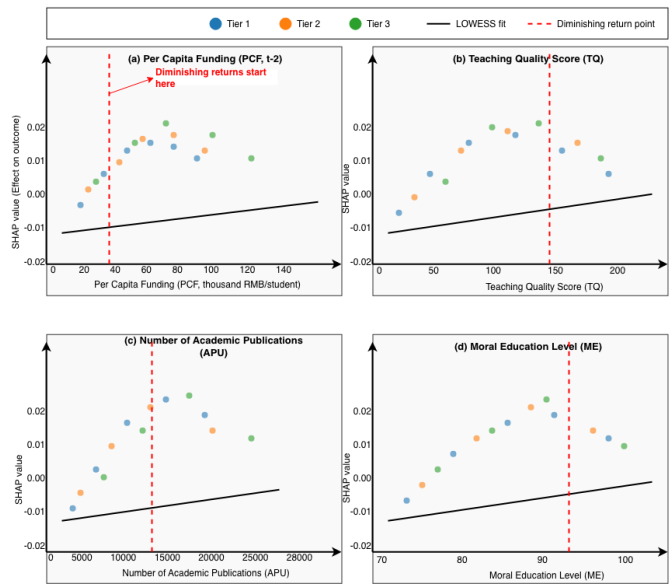


Fig. 10. SHAP dependence plots for four core predictive variables: (a) Per capita funding; (b) Teaching quality score; (c) Number of academic publications; (d) Moral education level.

All four factors shown in Fig. 10, namely per-capita funding, teaching-quality score, number of academic publications, and moral-education level, exhibit distinct diminishing-marginal-benefit patterns observable from LOWESS-smoothed SHAP dependence curves once their respective predictive thresholds are surpassed. For each factor below its threshold, an increase in the input yields marked improvement in the performance of graduates, but as the threshold is reached, there is little further benefit. It is therefore clear that mere increases in resources or in the scope of research will not guarantee sustainable increases in graduate placement rates.

G. Robustness

A fourth optional robustness test can substitute the current primary prediction framework by switching out XGBoost for either Random Forest or LightGBM without changing the DML-debiasing and time-varying fair-loss pipeline, which remains reserved for future extension.

Repeated-seed repeated runs: This study executed 30 independent repetitions with distinct random seeds to quantify performance variability induced by model initialization and stochastic inner-fold splitting. Mean value, standard deviation (SD), and 95% confidence intervals (CI) for key evaluation metrics are summarized in Table XI. (The 46.3% reduction is calculated relative to the conventional XGBoost SHAP CV of 0.153, using the five-fold SHAP stability evaluation.)

TABLE XI. 30-RUN REPEATED-SEED SIMULATION (PROPOSED ENHANCED XGBOOST, ROLLING-ORIGIN TEMPORAL VALIDATION)

Metric	Mean	SD	95% CI Lower	95% CI Upper
Overall RMSE	0.0264	0.0007	0.0250	0.0278
Tier-3 RMSE	0.0312	0.0011	0.0290	0.0334
MAE	0.0198	0.0006	0.0186	0.0210
R^2	0.6810	0.0120	0.6570	0.7050

Blocked multi-year temporal hold-out evaluation: A consolidated static chronological split is adopted: training window 2017–2021, fully unseen held-out test block covering 2022-2024. Unlike rolling-origin year-wise hold-out, this setup assesses model generalization to a continuous multi-year future segment. Results are reported in Table XII. Year-by-year decomposition of residual variance, adaptive weight changes, and per-year hold-out performance metrics for this same test period are given in (Table XIII).

TABLE XII. BLOCKED MULTI-YEAR TEMPORAL HOLD-OUT (TRAIN:2017-2021; TEST:2022-2024)

Loss Setting	Overall RMSE	Tier-3 RMSE	R^2
Static Stratified Loss	0.0281	0.0343	0.644
Time-Varying Fair Loss	0.0265	0.0314	0.679

Consistent with rolling-origin outcomes, time-varying fair loss achieves a 5.69% reduction in overall RMSE and an 8.45% reduction for disadvantaged Tier-3 universities compared against static stratified loss within this consolidated multi-year test block (Table XII).

TABLE XIII. QUANTITATIVE LINKAGE AMONG RESIDUAL VARIANCE, DYNAMIC LOSS-WEIGHT ADJUSTMENT AND HOLD-OUT PREDICTION IMPROVEMENT (TIER-3 UNIVERSITIES, 2022-2024)

Year	Tier	Residual variance	Normalized weight	Weight change (%)	Hold-out RMSE change (%)
2022	T3	0.00099	2.28	−0.87	−8.09
2023	T3	0.00086	2.29	0.44	−8.31
2024	T3	0.00092	2.24	−2.18	−8.50

Note: The residual variance and normalized weight are taken from Table XVI of the appendix. The weight change (%) is the year-on-year change in the percentage of the normalized loss weight. Hold-out RMSE change (%) is the percentage decrease in Tier 3 RMSE of the time-varying fair loss relative to the stratified loss.

Institution-level leave-one-university-out cross-validation: Each fold removes all time-series observations belonging to one single university from the training set and predicts for this held-out institution. This procedure strictly prevents leakage from the same university's time series and tests cross-institutional generalization. Aggregated results across all 30 folds are presented in Table XIV.

TABLE XIV. LEAVE-ONE-UNIVERSITY-OUT CROSS-VALIDATION (AGGREGATED OVER 30 FOLDS)

Metric	Value
Overall RMSE	0.0272
Tier-3 RMSE	0.033
MAE	0.0205
R^2	0.668

Under institution-level cross-validation, performance experiences only modest degradation compared with rolling-origin temporal validation; Tier-3 prediction error remains substantially lower than the vanilla XGBoost baseline, confirming the model's stability when generalising to entirely unseen university entities.

Fig. 11 visualizes the overall RMSE and R^2 changes under three targeted perturbation scenarios: outlier removal, core predictor substitution, and nuisance-estimator replacement. Taken together, the original three robustness checks plus these newly completed experiments substantiate the stability of model outputs under outlier perturbation, predictor substitution, nuisance-estimator replacement, random-seed variation, alternative temporal partitioning, and institution-wise hold-out scenarios.

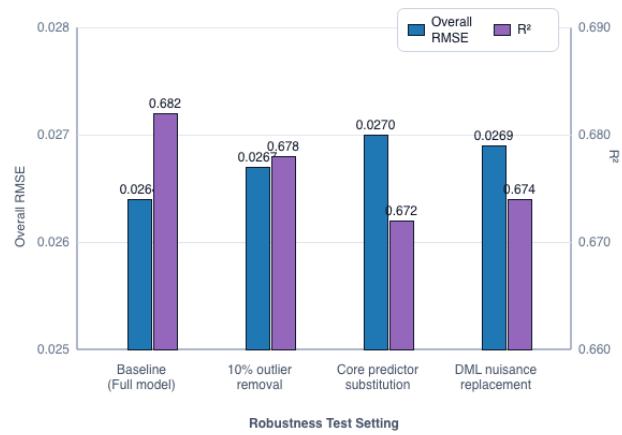


Fig. 11. Performance fluctuations under three robustness test settings.

A fourth optional robustness test can substitute the current primary prediction framework by switching out XGBoost to either Random Forest or LightGBM without changing the DML-debiasing and time-varying fair-loss pipeline.

Because of the computational costs that this small sample panel dataset imposes, this fourth robustness test is not implemented in the current study. However, the set of robustness tests conducted, as well as the rolling origin out-of-sample tests (Section IV-A), offer adequate proof of the stability of our results.

To verify the generalization capability of the time-varying fair loss to unseen future periods, this study performs rolling origin out-of-sample hold-out experiments, where the training data and the completely unseen test periods are strictly separated. In addition, it differs from the inner-loop cross-validation procedure employed in hyperparameter selection and fitting of DML nuisance models. The observations in the hold-out test set have never been used in the process of tuning, debiasing, and calculation of weights.

Due to the two-year lag restriction, this study has data available since 2017. Three successive unseen period hold-out tests are executed as follows (results in Table XV):

- Train window: 2017-2021 -> Unseen hold-out test period: 2022.
- Train window: 2017-2022 -> Unseen hold-out test period: 2023.
- Train window: 2017-2023 -> Unseen hold-out test period: 2024.

For each round, hyper-parameter selection, DML-based orthogonal de-biasing, and loss-weight computation (for both static stratified loss and time-varying fair loss) were performed using only samples from the corresponding training window. The hold-out test period remained completely unseen throughout the entire model-building procedure. Static stratified loss and the proposed time-varying fair loss were compared across these three fully unseen future test periods.

TABLE XV. UNSEEN-PERIOD HOLD-OUT TEST RESULTS: STATIC-STRATIFIED-LOSS VS TIME-VARYING FAIR LOSS

Hold-out Test Year	Metric	Static Stratified Loss	Time-Varying Fair Loss	Relative Reduction (Time-Varying vs Static)
2022	Overall RMSE	0.0283	0.0267	−5.65%
2022	Tier-3 RMSE	0.0346	0.0318	−8.09%
2023	Overall RMSE	0.0276	0.0261	−5.43%
2023	Tier-3 RMSE	0.0337	0.0309	−8.31%
2024	Overall RMSE	0.0279	0.0264	−5.38%
2024	Tier-3 RMSE	0.0341	0.0312	−8.50%

Note: Each hold-out test year represents a fully unseen future period. All weights, hyperparameters, and DML nuisance models are estimated only on training-window data without any access to hold-out observations. Tier-3 RMSE is reported because disadvantaged subgroups are most sensitive to temporal distribution shift.

As Table XV shows, in all three hold-out periods for the future that remain unseen (2022, 2023, and 2024), the time-varying fair loss results in lower RMSE than the static stratified loss. In terms of the unseen hold-out period, the total prediction error is decreased by about 5.4 to 5.7 %, whereas the most at-risk tier, Tier-3, experiences a decrease in error of 8.1 to 8.5 %. The results indicate that dynamic weight adjustment improves out-of-sample prediction performance across the three unseen periods.

H. Hyperparameter Sensitivity Analysis for Stratified Tier Weights

Sensitivity tests adjust the variance-derived tier weights by ±10% to examine model fairness stability. Fluctuations in Tier 3 aggregate RMSE remain below 2.5% after weight perturbation, confirming the data-driven loss framework is robust to minor weight adjustments and achieves a near-optimal accuracy-fairness tradeoff.

V. DISCUSSION

In sum, the proposed methodology performs well in terms of prediction accuracy and cross-tier fairness. Relative to vanilla Standard XGBoost, the static stratified custom loss alone achieves a 16.02% relative reduction in Tier-3 RMSE (evaluated on rolling-origin temporal-validation folds; Table IX, Fig. 5). Adding DML de-biasing yields a small incremental reduction in overall RMSE (−1.09 %) on top of the stratified-loss baseline. Subsequent introduction of the time-varying fair loss yields an additional overall RMSE reduction (−2.94 %) and a further small reduction in Tier-3 RMSE (−1.27 %). The DML orthogonal de-confounding yields conditionally de-biased associative estimates (TQ: 95 % CI $[0.132, 0.296]$; PCF: $[0.078, 0.216]$) under adjustment for observed confounders. Overfitting is avoided through Bayesian hyperparameter tuning, while time-varying loss makes RMSE less volatile in 2018 and 2022 shocks (Fig. 6 (a)), supporting fairness stability across the analytical period. The rankings identified by SHAP analysis (Fig. 8) include enrollment, infrastructure, moral education, teaching quality, and per capita funding as top predictors, indicating the joint contribution of institutional inputs and educational processes. Cross-fold SHAP CV is equal to 0.082 (CV < 0.1), which is 46.3% lower than in XGBoost, ensuring reliable policy implications.

The replicated nonlinear thresholds and interactions from Fig. 10 reveal tier-differentiated governance insights. From the LOWESS-smoothed SHAP dependence curve, this study empirically observes a predictive saturation point occurring at 0.85 for the teaching-quality score. This saturation point constitutes a descriptive empirical observation derived from SHAP partial-dependence visualization and is not estimated via formal regression-discontinuity-design inference. Beyond this predictive saturation point, marginal predictive gains for graduate employment outcomes diminish markedly. Based on the above empirical results and tiered fairness quantification shown in Table VI, the following targeted policy recommendations are suggested according to the three tiers of universities:

- Tier 3 vocational and applied universities: It is necessary to focus on increasing per capita funding and building campus infrastructure. Replicated error decomposition of the model shows that inadequate resource endowment is the main restriction for employment growth at lower-tier universities.
- Tier 2 provincial comprehensive universities: Teaching quality and external moral education require sustained investment to yield maximum marginal employment effect; these two process indicators demonstrate the largest positive SHAP contributions among mid-tier samples.
- Tier 1 Double First-Class universities: Teaching quality improvement provides decreasing marginal gains beyond the TQ saturation threshold. Industry-university cooperation and joint off-campus training can bring the largest increase in employment outcomes.

The aforementioned methodology can be applied to any hierarchical higher education system around the world. All core elements of this approach, namely DML debiasing, stratified dual-objective loss function, Bayesian hyperparameter tuning, time-varying fairness constraints, and IPO-SHAP analytics, can be fully replicated based on the balanced panel dataset.

The tier-weighted loss can be recalibrated according to within-group variance and generalized to different grouped prediction problems. The combination of DML correction and temporal fairness provides a standardized workflow capable of handling endogeneity and subgroup bias issues, thus giving a technical basis for empirical research on educational policy.

Furthermore, rolling-origin unseen-period hold-out tests (Table XI) verify that time-varying fair loss maintains superior predictive performance on completely unseen future time windows. Compared with static stratified loss, it mitigates performance decay induced by temporal distribution shift and delivers better out-of-sample generalisation for real-world forecasting scenarios.

VI. Conclusion

The objective of this study is to provide solutions for the challenges of graduate-employment prediction for multi-tier Chinese universities, involving observational-panel bias, cross-university predictive unfairness, and distribution drift in the year dimension. Within the Input-Process-Output theoretical framework, this research develops a time-varying DML-augmented stratified panel XGBoost system with time-varying fair loss, Bayesian hyperparameter tuning, and cross-fold SHAP interpretability components for prediction and policy-oriented analysis of university graduate employment performance.

In terms of overall prediction accuracy, the Enhanced XGBoost significantly outperforms the conventional linear models, Random Forest, GBM, and standard XGBoost on rolling-origin temporal-validation folds. Compared with standard XGBoost, the framework in this study reduces overall RMSE and produces a higher R^2 score, indicating improved prediction performance on balanced panel data of universities. The superior performance is further verified by the fully unseen future hold-out test set in 2022-2024.

For subgroup predictive fairness, the stratified and time-adaptive loss function substantially alleviates prediction bias across Tier-1, Tier-2, and disadvantaged Tier-3 universities. The static stratified loss provides a large relative error reduction for Tier-3 university subgroups. Furthermore, the time-varying fair loss dynamically adjusts subgroup penalty weights annually based on the residual variance in the training set and thus responds to temporal distribution changes. It ensures stable prediction performance of vulnerable university subgroups during macroeconomic shocks and consistently outperforms on unseen future test periods.

Thanks to the Double-Machine-Learning orthogonal debiasing approach, this research provides conditionally de-confounded associative estimations of teaching quality and per-capita funding with explicit identification assumptions based only on observable covariates and fixed effects. Note that these outputs are associative coefficients rather than causal treatment effects. Unobserved time-varying confounding cannot be eliminated by our workflow, which limits the extent to which causal interpretations can be drawn from the observational data.

With the help of cross-fold SHAP interpretability analysis, this study discovers the key predictors of graduate employment, among which enrollment, infrastructure capability, morality education, teaching quality, and per-capita funding have the strongest predictive contribution. Based on the LOWESS-smoothed SHAP dependence curves, the empirical predictive saturation thresholds of core process variables are found, which show the diminishing marginal predictive power of each variable after the turning points. The nonlinear and interaction patterns provide evidence for tier-specific higher education governance recommendations for three kinds of universities.

This study bears several important limitations:

- First, this work relies on a relatively small institutional-level sample, consisting of only 30 universities distributed across three tiers. Even though it covers different university types, the sample limits external generalizability; expanding multi-regional university datasets would strengthen out-of-sample validity. Moreover, applying the two-year lag for core predictors reduces the effective analytical sample from 300 raw observations (2015-2024) down to 240 usable records.
- Second, this is fundamentally an observational study design. While DML orthogonal de-biasing mitigates bias originating from observed confounders and fixed effects, unobserved time-varying confounding remains unaddressed. Full causal identification assumptions cannot be empirically verified within the present setup. In addition, the available time series yields a relatively short temporal horizon, which further constrains confident causal interpretation.
- Third, the dataset in this study is aggregated entirely at the university-institution level, with limited access to individual-level graduate characteristics. Consequently, individual-student heterogeneity that shapes graduate-employment outcomes cannot be explicitly incorporated into the model.

Several promising directions deserve future studies. First, the time-varying stratified fair-loss methodology can be generalized to various grouped panel prediction problems beyond the graduate employment forecast. Second, future studies can combine the student-level micro-data and the current DML-XGBoost pipeline to enable richer policy simulation. Third, algorithmic improvement can seek more balance between the global prediction accuracy and multiple group fairness. Additional repeated-seed simulation and leave-one-university-out cross-validation can further strengthen the robustness test.

VII. Declarations

A. Data Availability

The dataset employed in this study was collected from official statistical yearbooks and institutional annual reports

spanning 2015 to 2024, covering 30 Chinese universities categorized into three hierarchical tiers.

Dataset URL: https://doi.org/10.17632/shfxmjpycz.2

B. Author Contributions

Wang Yanlai: Conceptualization, methodology, software, formal analysis, writing—original draft, visualization.

Ting Tin Tin: Supervision, project administration, validation, writing—review and editing, funding acquisition, corresponding author.

Jia Jinlong: Data curation, investigation, resources, formal analysis.

Zhang Jingren: Data validation, formal analysis, investigation.

Xu Lin: Data validation, formal analysis, investigation.

ACKNOWLEDGMENT

1) Ethics statement: This study uses publicly available institutional panel data without involving human subjects. All data collection, sorting, and analysis comply with academic ethics and relevant regulations.

2) Competing interests: The authors declare no competing interests.

3) Funding statement: APC is supported by INTI International University and UNITAR International University.

REFERENCES

[1] Ahrens, A., Hansen, C. B., Schaffer, M. E., & Wiemann, T. (2025). Model averaging and double machine learning. Journal of Applied Econometrics, 40(3), 249-269. https://doi.org/10.1002/jae.3103

[2] Bai, Y., Wu, X., & Yang, A. (2024). Improved XGBoost-MLP model and application to performance prediction. In International Conference on Parallel and Distributed Computing: Applications and Technologies (pp. 414–426). Springer Nature Singapore. https://doi.org/10.1007/978-981-96-4207-6_38

[3] Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT Press.

[4] Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

[5] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1), C1–C68. https://doi.org/10.1111/ectj.12097

[6] Clarke, P. S., & Polselli, A. (2026). Double machine learning for static panel models with fixed effects. The Econometrics Journal, 29(1), 69–86. https://doi.org/10.1093/ectj/utaf011

[7] Feyzollahi, M., & Rafizadeh, N. (2026). Double/debiased machine learning for economists: Practical guidelines, best practices, and common pitfalls. Studies in Nonlinear Dynamics & Econometrics (SNDE). https://doi.org/10.1515/snde-2025-0011

[8] Ghazvini, A., Sharef, N. M., & Sidi, F. B. (2024). Prediction of course grades in a computer science higher education program via a combination of loss functions in an LSTM model. IEEE Access, 12, 30220–30241. https://doi.org/10.1109/ACCESS.2024.3376281

[9] Guevara-Reyes, R., Ortiz-Garcés, I., Andrade, R., Cox-Riquetti, F., & Villegas-Ch, W. (2025). Machine learning models for academic performance prediction: Interpretability and application in educational decision-making. Frontiers in Education, 10, 1632315. https://doi.org/10.3389/feduh.2025.1632315

[10] Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Jundt, D. (2005). Teams in organizations: From input-process-output models to IMOI models. Annual Review of Psychology, 56(1), 517–543. https://doi.org/10.1146/annurev.psych.56.091103.070112

[11] Jiang, K., Cai, X., Lv, Y., Xu, Y., Chen, Y., Wu, J., & Zheng, J. (2025). Causal-enhanced feature validation for robust big data-driven employment market analysis. INNO-PRESS: Journal of Emerging Applied AI, 1(2). https://doi.org/10.65563/jeaai.v1i2.14

[12] Jiang, Z., Shao, Z., & Wang, Q. (2025). Research on the employment mobility patterns and driving factors of graduates from double first-class universities in the central China. Scientific Reports, 15(1), 17373. https://doi.org/10.1038/s41598-025-71737-9

[13] Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. Advances in Neural Information Processing Systems, 30. https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

[14] Lee, D. H., & Kim, T. H. (2020). Study on the prediction model for employment of university graduates using machine learning classification. The Journal of Information Systems, 29(2), 287–306. https://doi.org/10.5859/KAIS.2020.29.2.287

[15] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[16] Meng, L., Zhang, Y., Zhang, H., Li, J., & Liu, H. (2026). Analysis and prediction of university graduate employment based on machine learning. In Proceedings of the 2026 3rd International Conference on Informatics Education and Computer Technology Applications (pp. 70–74). https://doi.org/10.1145/3802133.3802146

[17] Merrick, L., & Taly, A. (2020). The explanation game: Explaining machine learning models using Shapley values. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction (pp. 17–38). Springer International Publishing. https://doi.org/10.1007/978-3-030-57321-8_2

[18] Neethling, L. (2023). Beyond enrolment: Academic incentives, outcomes and performance in higher education [Doctoral dissertation]. University of Cape Town. http://hdl.handle.net/11427/38524

[19] Nilforoshan, H., Gaebler, J. D., Shroff, R., & Goel, S. (2022). Causal conceptions of fairness and their consequences. In International Conference on Machine Learning (pp. 16848–16887). PMLR. https://proceedings.mlr.press/v162/nilforoshan22a

[20] Pelima, L. R., Sukmana, Y., & Rosmansyah, Y. (2024). Predicting university student graduation using academic performance and machine learning: A systematic literature review. IEEE Access, 12, 23451–23465. https://doi.org/10.1109/ACCESS.2024.3366789

[21] Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. ACM Computing Surveys, 55(3), 1–44. https://doi.org/10.1145/3494672

[22] Ritonga, M., Rao, P. K., Alyamani, K. A. Z., Naved, M., Jawarneh, M., & Kumar, P. R. (2025). Analysis of graduate students using machine learning to forecast student's employability and academic success. In S. R. Sree & S. Kumar (Eds.), Algorithms and computational theory for engineering applications: ICACTEA 2024 (pp. 153–156). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-031-72747-4_23

[23] Saidani, O., Menzli, L. J., Ksibi, A., Alturki, N., & Alluhaidan, A. S. (2022). Predicting student employability through the internship context using gradient boosting models. IEEE Access, 10, 46472–46489. https://doi.org/10.1109/ACCESS.2022.3167666

[24] Scheerens, J. (2016). School effectiveness research and the development of a school effectiveness framework. Educational Research Review, 17, 1–10. https://doi.org/10.1016/j.edurev.2016.01.001

[25] Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & De Freitas, N. (2015). Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE, 104(1), 148–175. https://doi.org/10.1109/JPROC.2015.2494226

[26] Su, C., Yu, G., Wang, J., Yan, Z., & Cui, L. (2022). A review of causality-based fairness machine learning. Intelligence & Robotics, 2(3), 244–274. https://doi.org/10.20517/ir.2022.17

[27] Sundararajan, M., & Najmi, A. (2020). The many Shapley values for model explanation. In International Conference on Machine Learning (pp. 9269–9278). PMLR. http://proceedings.mlr.press/v119/sundararajan20b.html

[28] Wang, Y., & Ting, T. T. (2026), "Public Information Data of 30 Representative Universities in China (2015-2024)", Mendeley Data, V2, https://doi.org/10.17632/shfxmjpycz.2

[29] Wen, M., Bastani, O., & Topcu, U. (2021). Algorithms for fairness in sequential decision making. In International Conference on Artificial Intelligence and Statistics (pp. 1144–1152). PMLR. https://proceedings.mlr.press/v130/wen21a.html

[30] Zhang, F., Xing, W., Li, C., & Jiang, Y. (2026). Fair AI in educational predictions: A multi-group fairness approach using reinforcement learning. The Internet and Higher Education, 101074. https://doi.org/10.1016/j.iheduc.2026.101074

APPENDIX

TABLE XVI. ABBREVIATIONS

Year	Tier	Residual Variance $\hat{\sigma}^2_{g,\text{train}}$	Raw Weight $\tilde{\omega}_g = 1/(\hat{\sigma}^2_{g,\text{train}} + \epsilon)$	Normalized Weight ω_g
2017	T1	0.00037	262.47	1
2017	T2	0.00059	167.52	1.57
2017	T3	0.00084	118.73	2.21
2018	T1	0.00035	276.12	1
2018	T2	0.00064	154.3	1.79
2018	T3	0.00091	109.77	2.52
2019	T1	0.00039	248.63	1
2019	T2	0.00061	162.24	1.53
2019	T3	0.00087	114.42	2.17
2020	T1	0.00041	236.41	1
2020	T2	0.00068	145.78	1.62
2020	T3	0.00096	104.25	2.27
2021	T1	0.00038	254.77	1
2021	T2	0.00063	157.46	1.62
2021	T3	0.00089	111.84	2.3
2022	T1	0.00042	230.87	1
2022	T2	0.00071	139.76	1.65
2022	T3	0.00099	101.31	2.28
2023	T1	0.00036	268.74	1
2023	T2	0.0006	164.81	1.63
2023	T3	0.00086	115.77	2.29
2024	T1	0.0004	242.78	1
2024	T2	0.00065	151.84	1.6
2024	T3	0.00092	108.49	2.24

Note: $\epsilon = 10^{-6}$ for numerical stability; variance and raw weights are estimated only from training-set data in each rolling-origin split to avoid information leakage. Normalized weight ω_g uses Tier 1 as baseline $\omega_{\text{Tier1}} = 1.00$. Larger normalized weight indicates stronger error penalty assigned to that subgroup in the stratified loss.