

Supplementary Material for “Prediction Intervals and Group Variable Importance for Classification Models in University Enrollment Yield”

XIAOHUI YIN¹, YINGFA XIE¹, JUN YAN¹, SIYAN WANG¹, PANG-YU LIU¹, JEFFREY GAGNON², ELTON THOMPSON², LAWRENCE WALSH², NATHAN FUERST², AND MING-HUI CHEN^{1,*}

¹*Department of Statistics, University of Connecticut, Storrs, CT, USA*

²*Division of Student Life & Enrollment, University of Connecticut, Storrs, CT, USA*

S.1 Mixture-of-Experts Gate-Fitting Procedure

This section provides the full gate-fitting procedure for the MoE ensemble summarized in Section 3 of the main paper.

We fit the gate as a two-stage best-expert classifier. First, each expert k is trained via V -fold cross-validation to produce, for every training observation i , an out-of-fold probability $\hat{p}_k^{(cv)}(\mathbf{x}_i)$ from the model fit on the folds that exclude i . The per-observation best expert is recorded as the one with the smallest log-loss,

$$k_i^* = \arg \min_k -\{y_i \log \hat{p}_k^{(cv)}(\mathbf{x}_i) + (1 - y_i) \log(1 - \hat{p}_k^{(cv)}(\mathbf{x}_i))\}.$$

Second, an elastic-net penalized multinomial logistic regression with K classes is fit to the categorical response $\{k_i^*\}_{i=1}^{N_{tr}}$ using the same covariates $\mathbf{x}$ as the logistic family models, with the penalty strength selected by cross-validation. The resulting softmax gate probabilities

$$\pi_k(\mathbf{x}) = \frac{\exp(\mathbf{x}^\top \boldsymbol{\gamma}_k)}{\sum_{k'=1}^K \exp(\mathbf{x}^\top \boldsymbol{\gamma}_{k'})},$$

are used at test time to combine the predictions of the experts refitted on the full training data.

S.2 Derivation of $\hat{\sigma}_C^2$ in Proposition 1

Let $p_i = \text{expit}(\mathbf{x}_i^\top \boldsymbol{\beta})$ be the true probability for observation i in the test set. We derive $\hat{\sigma}_C^2 = \text{Var}(C - \hat{C})$. Because $\hat{\boldsymbol{\beta}}$ is a function of the training data only, the test outcomes $\{y_i\}$ are conditionally independent of $\hat{\boldsymbol{\beta}}$ given the test covariates $\{\mathbf{x}_i\}$ under the standard i.i.d. assumption and the temporal split, which mimics the actual decision process. By the law of total covariance, conditioning on $\hat{\boldsymbol{\beta}}$ renders $\sum_{i=1}^n \hat{p}_i$ a constant, so $\text{Cov}(\sum y_i, \sum \hat{p}_i) = 0$, and the cross term in the variance expansion vanishes:

$$\begin{aligned} \text{Var}(C - \hat{C}) &= \text{Var}\left(\sum_{i=1}^n y_i\right) + \text{Var}\left(\sum_{i=1}^n \hat{p}_i\right) - 2 \text{Cov}\left(\sum_{i=1}^n y_i, \sum_{i=1}^n \hat{p}_i\right) \\ &= \sum_{i=1}^n \text{Var}(y_i) + \text{Var}\left(\sum_{i=1}^n \hat{p}_i\right). \end{aligned}$$

*Corresponding author. Email: ming-hui.chen@uconn.edu.

The first term corresponds to the binomial variation from the testing responses,

$$\sum_{i=1}^n \text{Var}(y_i) = \sum_{i=1}^n p_i(1 - p_i).$$

For the second term, a first-order Taylor expansion of $\hat{p}_i = \text{expit}(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}})$ around $\boldsymbol{\beta}$ gives

$$\begin{aligned} \text{Var}\left(\sum_{i=1}^n \hat{p}_i\right) &= \mathbf{1}_n^T \begin{pmatrix} \hat{p}_1(1 - \hat{p}_1)\mathbf{x}_1^\top \\ \hat{p}_2(1 - \hat{p}_2)\mathbf{x}_2^\top \\ \dots \\ \hat{p}_n(1 - \hat{p}_n)\mathbf{x}_n^\top \end{pmatrix} \text{Var}(\hat{\boldsymbol{\beta}}) \begin{pmatrix} \hat{p}_1(1 - \hat{p}_1)\mathbf{x}_1^\top \\ \hat{p}_2(1 - \hat{p}_2)\mathbf{x}_2^\top \\ \dots \\ \hat{p}_n(1 - \hat{p}_n)\mathbf{x}_n^\top \end{pmatrix}^\top \mathbf{1}_n \\ &= \left[\sum_{i=1}^n \hat{p}_i(1 - \hat{p}_i)\mathbf{x}_i^\top \right] \text{Var}(\hat{\boldsymbol{\beta}}) \left[\sum_{i=1}^n \hat{p}_i(1 - \hat{p}_i)\mathbf{x}_i \right]. \end{aligned}$$

Hence, the total variance can be decomposed into the binomial variance from the intrinsic randomness of the testing outcomes and the estimation variance arising from uncertainty in the estimation of $\boldsymbol{\beta}$. By substituting the fitted probabilities $\{\hat{p}_i\}_{i=1}^n$ and the estimated covariance matrix $\hat{\mathbf{V}}$, we obtain the plug-in estimator $\hat{\sigma}_C^2$. $\square$

S.3 Simulation under a Mixture-of-Experts DGP

As a robustness check on the SuperLearner DGP used in Section 4 of the main paper, we replicate the semi-parametric bootstrap population simulation under an MoE-based DGP, where the MoE ensemble is fit on the pooled 2020–2025 in-state applicant data using the same specification as Section 3 of the main paper. Under this DGP, the expected AUROC and AUPRC of an oracle whose predictions equal $\tilde{p}_i$ when evaluated on Bernoulli outcomes $y_i \sim \text{Ber}(\tilde{p}_i)$ are $\mathbb{E}[\text{AUROC}] \approx 0.747$ and $\mathbb{E}[\text{AUPRC}] \approx 0.588$, very close to the corresponding SuperLearner-DGP values ($\mathbb{E}[\text{AUROC}] \approx 0.74$ and $\mathbb{E}[\text{AUPRC}] \approx 0.58$), indicating comparable signal strength under the two DGPs. Averaged over training observations, the MoE gate places weights of approximately 0.24, 0.27, 0.32, and 0.17 on ENet, Ranger, XGBoost, and CatBoost, respectively, spreading mass more evenly across the four base learners than the SuperLearner-DGP NNLS weights, which concentrated on CatBoost and XGBoost.

Table S.1 reports the simulation results, mirroring Table 1 of the main paper. The qualitative findings are preserved: LR, ENet, and MoE attain close-to-nominal coverage; XGBoost and CatBoost show mild-to-moderate undercoverage; Ranger and SuperLearner pronouncedly undercover. The bootstrap-fit instability pattern for CatBoost and SuperLearner relative to the full-train fit, central to the diagnostic in Section 4, is also visible in Figure S.1.

S.4 pAUC and mAUC for the Remaining Models

Figure S.2 reports pAUC and mAUC for the nine feature groups under the three remaining models in test year 2025: LR, XGBoost, and SuperLearner. The patterns mirror those of their representative counterparts in Figure 5 of the main paper: LR and ENet, XGBoost and CatBoost, and SuperLearner and MoE behave similarly, so the conclusions drawn from the main-text figure carry over to the remaining models.

Table S.1: Simulation results under the MoE DGP over $R = 200$ Monte Carlo replicates with $B = 200$ inner draws at nominal level 95%. Coverage is the empirical rate; other columns report the replicate mean with Monte Carlo standard error in parentheses.

Model	Method	$\mathbb{I}(C \in \hat{I})$	$ \hat{I} $	$ \hat{C} - C $	AUROC (%)	AUPRC (%)
LR	Asymptotic	0.935	154.7 (0.0)	31.5 (1.6)	70.31 (0.04)	52.06 (0.08)
	Para-Sim	0.935	150.4 (0.6)	31.0 (1.6)	70.30 (0.05)	52.05 (0.08)
ENet	Bootstrap	0.945	150.4 (0.7)	30.6 (1.7)	71.90 (0.05)	54.63 (0.08)
Ranger	Bootstrap	0.750	149.2 (0.7)	50.1 (2.3)	70.08 (0.05)	51.20 (0.08)
XGBoost	Bootstrap	0.890	146.7 (0.6)	37.6 (1.9)	71.99 (0.05)	54.88 (0.08)
CatBoost	Bootstrap	0.835	155.8 (0.7)	47.1 (2.2)	72.37 (0.05)	55.56 (0.08)
SuperLearner	Bootstrap	0.770	149.1 (0.7)	49.9 (2.3)	72.50 (0.05)	55.72 (0.08)
MoE	Bootstrap	0.930	148.1 (0.7)	31.4 (1.6)	72.32 (0.05)	55.39 (0.08)

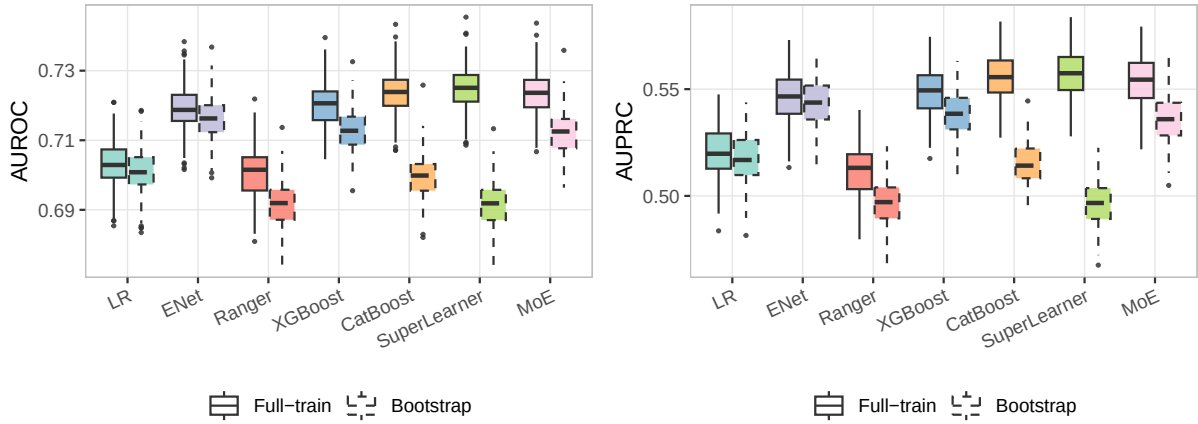


Figure S.1: Per-replicate AUROC (a) and AUPRC (b) distributions under the MoE DGP. Solid boxes: full-train fit; dashed boxes: bootstrap fits (mean of $B = 200$ refits per replicate). Format mirrors Figure 2 of the main paper, which confirms that the diagnostic in Section 4 is not an artifact of the DGP choice.

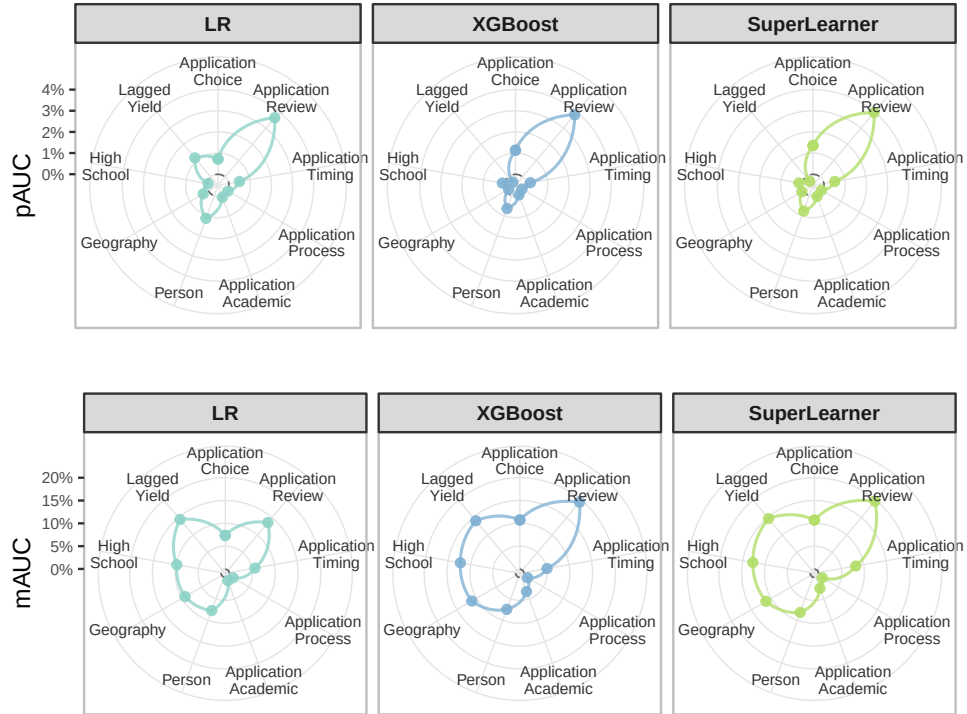


Figure S.2: pAUC and mAUC for nine feature groups across the three remaining models (LR, XGBoost, SuperLearner) in test year 2025. LR and Enet, XGBoost and CatBoost, and SuperLearner and MoE behave similarly, so the importance rankings observed in Figure 5 of the main paper extend to the remaining three models.