Journal of Data Science 24 (3), 523–543
July 2026
DOI: 10.6339/26-JDS1241
Statistical Data Science

Prediction Intervals and Group Variable Importance for Classification Models in University Enrollment Yield

Xiaohui Yin[1], Yingfa Xie[1], Jun Yan[1], Siyan Wang[1], Pang-Yu Liu[1], Jeffrey Gagnon[2], Elton Thompson[2], Lawrence Walsh[2], Nathan Fuerst[2], 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*

Abstract

Accurate forecasting of student yield is critical for academic institutions because enrollment projections directly affect resource allocation, course offerings, housing capacity, and budget decisions. Existing enrollment models primarily focus on predicting individual matriculation probabilities, whereas institutional planning depends on accurate prediction of the aggregate enrollment count together with its uncertainty quantification. Here we develop statistical methods for aggregate enrollment prediction and interval estimation under logistic, regularized regression, tree-based, and ensemble classification models. For unpenalized logistic regression, we derive an asymptotic prediction interval and a parametric-simulation interval that propagates coefficient uncertainty. For regularized and tree-based learners, we develop a model-agnostic bootstrap interval that captures selection, estimation, and predictive uncertainty in a unified framework. To interpret predictive structure, we introduce partial and marginal AUC measures to quantify the unique and standalone contributions of feature groups. Simulation studies show that the proposed intervals attain close-to-nominal coverage under well-calibrated and bootstrap-stable learners. Applied to University of Connecticut in-state freshman cohorts from 2020–2025, the proposed intervals contain the observed 2025 enrollment count while identifying the feature groups that contribute most strongly to the enrollment prediction.

Keywords *classification; enrollment prediction; prediction interval; uncertainty quantification; variable importance*

1 Introduction

Enrollment management has long been a central concern for higher education, and public universities are under particular pressure to balance incoming class sizes against available resources and financial goals. Forecast errors can have significant consequences: underestimation may strain classroom capacity, housing, and student services, while overestimation can result in unused facilities and budget shortfalls. Yield forecasting, the prediction of the number of admitted applicants who matriculate, plays a critical role in institutional readiness, linking admissions outcomes to staffing, financial aid, and long-term planning. This importance has motivated extensive research, ranging from early methods in institutional research (Brinkman and McIntyre,

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

Received June 2, 2026; Accepted July 1, 2026

1997) to strategic enrollment management frameworks (Langston et al., 2016; Langston and Loreto, 2017). Econometric perspectives have also highlighted how tuition, financial aid, and macroeconomic conditions shape enrollment behavior (Hoenack and Pierro, 1990; Harrington and Schibik, 2013), reinforcing that yield prediction is essential for both operational and strategic decision-making.

Traditional approaches to yield forecasting relied heavily on institutional research tools, such as surveys and trend analysis, to anticipate enrollment shifts (Brinkman and McIntyre, 1997). With the availability of richer student-level data, statistical models became more prominent; logistic regression emerged as a dominant technique for predicting whether an admitted applicant would enroll based on demographic, academic, and financial characteristics (Abelt et al., 2015; Goenner and Pauls, 2006; Hayes et al., 2013). These regression-based models provided interpretable relationships between predictors and enrollment outcomes, establishing a framework that remains widely used in institutional practice. While valuable, such models often assumed relatively simple structures and struggled to capture the increasing heterogeneity of modern applicant pools.

The expansion of institutional datasets and advances in computational tools have spurred the growing use of data mining and machine learning in enrollment prediction. Studies have shown that tree-based methods, including decision trees and random forests, can improve predictive accuracy compared to traditional regression frameworks (Ab Ghani et al., 2019; Cirelli et al., 2018; Esquivel and Esquivel, 2021). Neural networks have also demonstrated promise in uncovering subtle, high-dimensional relationships among applicant characteristics (Kye, 2023). Beyond classification, optimization-based frameworks have been developed to align scholarship allocation with predicted enrollment probabilities, allowing institutions to balance financial aid budgets while maximizing yield (Aulck et al., 2020; Jamison, 2017). Although these approaches differ substantially in modeling strategy and predictive performance, they share a common objective: predicting the enrollment decisions of individual applicants.

Institutional planning, however, depends primarily on the aggregate number of matriculants rather than individual enrollment outcomes. Enrollment forecasts influence a broad range of operational and strategic decisions, including academic planning, housing allocation, financial-aid budgeting, and resource management (Trusheim and Rylee, 2011). The total enrollment count therefore serves as the primary quantity of interest for enrollment managers, who must balance institutional capacity against enrollment goals. Although enrollment decisions are observed at the student level, many institutional actions are based on projected cohort sizes rather than individual outcomes. In practice, aggregate enrollment is often estimated by summing predicted enrollment probabilities across admitted applicants, providing a natural bridge between individual-level prediction models and institution-level planning.

Accurate aggregate forecasting requires not only point predictions but also quantification of the uncertainty surrounding those predictions. Forecast research has long emphasized that point predictions alone are incomplete without prediction intervals to quantify uncertainty (Chatfield, 1993; Gneiting and Katzfuss, 2014). Recent studies in machine learning have shown that ignoring uncertainty, particularly in complex models such as deep neural networks, can lead to overconfident and poorly calibrated predictions (Abdar et al., 2021; Gawlikowski et al., 2023). More broadly, uncertainty quantification has become recognized as a core element of trustworthy machine learning, supporting the reliability of predictions under distributional shifts and other changing conditions (Liu et al., 2023; Lopez et al., 2025). For enrollment prediction, uncertainty is as consequential as accuracy for effective resource planning and risk management.

This paper presents methods for predicting the aggregate enrollment count and quantify-

ing their uncertainty under a broad class of classification models. For the unpenalized logistic regression model, a closed-form asymptotic prediction interval for the aggregate count is derived and complemented by a parametric-simulation interval. For regularized regression and machine-learning models, a model-agnostic bootstrap prediction interval is proposed, and its empirical coverage is evaluated through a Monte Carlo simulation. To facilitate interpretation, two AUC-based group importance measures, partial AUC (pAUC) and marginal AUC (mAUC), are introduced to quantify the unique and standalone contributions of nine substantive feature groups. Applied to University of Connecticut in-state freshman cohorts and validated on held-out future cohorts, the proposed framework provides aggregate enrollment forecasts, prediction intervals, and interpretable assessments of the drivers of enrollment.

The rest of the article is organized as follows. The dataset used in this study is described in Section 2; Section 3 introduces the candidate predictive models, develops the procedures for constructing prediction intervals, and defines the pAUC and mAUC group importance measures. A simulation study evaluating the empirical coverage of the proposed intervals is presented in Section 4. Results from the empirical analysis are presented in Section 5, and Section 6 concludes with a discussion of the main findings, limitations, and future directions. Technical details and additional results are relegated to the Supplementary Material.

2 Data

Institutional data were obtained from the University of Connecticut's Division of Student Life & Enrollment. The data used in this study were limited to in-state first-year non-transfer freshman applicants admitted to the Storrs campus who were expected to matriculate in person in the fall semester (the Storrs campus offers no online-only first-year degree program). Students who were previously enrolled or offered enrollment at a regional campus were excluded to maintain a consistent population for analysis. Each fall, approximately 1,800–2,300 new in-state freshmen matriculate at the University of Connecticut's Storrs campus, drawn from a pool of roughly 6,000–7,200 admitted students, across the 2020–2025 cohorts.

The analysis covers in-state students admitted to the University of Connecticut Storrs campus between 2020 and 2025, with enrollment defined as a binary outcome variable (1 for enrolled, 0 otherwise). Annual yield rates remained relatively stable from 2020 to 2024 near 32%, with a modest decline in 2025. The predictive modeling framework targets the 2025 cohort and uses the five preceding admission cycles (2020–2024) as the training data. Figure 1 visualizes the spatial distribution of admissions and enrollments across Connecticut high schools for the 2025 cohort through a 3D spatial plot, in which each high school is represented by a dual-cylinder structure: the narrower cylinder's height indicates the number of admitted-but-not-enrolled students, while the wider base cylinder's height indicates the number of admitted-and-enrolled (i.e., matriculating) students.

Our analysis draws on 63 predictors organized into five top-level feature groups (person-aspect, application-aspect, high-school, geographic, and lagged-yield), with the application-aspect group further decomposed into five subgroups (Review, Choice, Timing, Academic, and Process), giving nine feature groups used in the group-importance analysis of Sections 3 and 5. Through feature engineering (dummy coding of categorical variables and a small number of derived indicators), the 63 raw predictors are expanded to 83 columns entering the candidate models.

Figure 1: 3D spatial plot of admissions outcomes per Connecticut high school for the 2025 cohort. Each high school is shown as a dual-cylinder structure: the narrower cylinder's height encodes the number of admitted-but-not-enrolled students, and the wider base cylinder's height encodes the number of admitted-and-enrolled (matriculating) students. High schools near the Storrs campus convert admissions to enrollments at a noticeably higher rate than those near the New York border. This commuting-driven heterogeneity motivates the geographic distance and lagged-yield features.

Person-Aspect Variables The person-aspect variables capture demographic and background characteristics of each applicant: first-generation status, sex, ethnicity, athletic participation, permanent-home situation, residency classification, a self-reported measure of institutional interest, and family-connection indicators (parent or guardian employed at UConn, sibling previously attended).

Application-Aspect Variables The application-aspect variables describe the information provided during the admission process and reflect the preferences, administrative actions, and institutional assessments of the applicants. They are further divided into five subgroups.

Application Review These include admissions and language proficiency assessments; merit indicators (honors and scholarship eligibility, portfolio review); the assigned program and major; and two constructed indicators of whether the assigned program and major matched the applicant's first-choice selections.

Application Choice These include the first-choice program and major of the applicant, alternate campus, major, and program choices, and additional indicators for declared interest in special programs, ROTC interest, and the decision to opt out of submitting standardized test scores.

Application Timing These include the calendar month of submission and two engagement features (days from application creation to submission, and days from submission to the

August 1 enrollment deadline), together with indicators for late applications and prior applicants.

Application Academic These consist of GPA-based measures: the admissions-recalculated grade point average and two institutional weighted-GPA metrics capturing academic and overall performance.

Application Process These capture administrative events during the admissions process, including whether the application fee was paid or waived, whether a record of behavioral misconduct was present, and whether the applicant was considered for a regional campus.

High School Variables We incorporate contextual measures for each applicant's high school: name and ZIP code as identifiers; a four-level urban/suburban/town/rural locale; state and national integer rankings; and percentile-scaled academic-environment measures (Advanced Placement (AP) course participation and availability, senior class size) and socioeconomic indicators (median family income, free-lunch rate, educational attainment, housing stability).

Geographic Variables These characterize residential context: the home ZIP code and a log-transformed geodesic distance from the home-ZIP centroid to the Storrs campus, capturing commuting proximity as a likely driver of in-state enrollment.

Lagged Yield Variables Lagged-yield variables are constructed from prior admitted cohorts and represent historical enrollment rates at five levels of aggregation: assigned major, assigned program, high school ZIP code, high school organization, and home ZIP code. They compress high-cardinality categorical information (ZIP codes, high-school identifiers) that cannot be entered directly into linear models, capturing the notion that applicants are more likely to enroll where peers in their neighborhood or high school have previously done so. ZIP codes with sparse historical counts were collapsed into a single category for robustness.

3 Methods

We specify the candidate predictive models for individual enrollment, define two AUC-based measures of feature-group contribution (pAUC and mAUC), and develop the prediction-interval machinery for the aggregate count. Let $y \in \{0, 1\}$ denote binary response indicating enrollment. The observed data consists of independent copies of y and their corresponding 63 predictors described in Section 2.

3.1 Models

We consider three model families. The encoding of the predictors for model fitting depends on the model: logistic models and XGBoost require dummy coding for categorical features, expanding the predictors to $p = 83$ columns; CatBoost and Ranger accept predictors directly.

Logistic Models We first use logistic regression (LR) to model the conditional enrollment probability as

$$\Pr(y = 1 \mid \boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}^\top \boldsymbol{\beta})},$$

where $\boldsymbol{x} \in \mathbb{R}^p$ is the covariate vector and $\boldsymbol{\beta}$ is the regression coefficient vector. We then consider its penalized variant, Elastic Net (ENet) (Zou and Hastie, 2005), which addresses multicollinearity and high dimensionality through combined ℓ_1 and ℓ_2 penalties on $\boldsymbol{\beta}$; the penalty strength is selected by cross-validation.

Tree-Based Models To capture nonlinear structure and higher-order interactions, we consider three tree-based ensembles from two families: random forests and gradient boosting. Ranger (Wright and Ziegler, 2017) is a fast implementation of the random-forest algorithm (Breiman, 2001), aggregating predictions across bagged decision trees. From the gradient-boosting family, XGBoost (Chen and Guestrin, 2016) provides an efficient regularized implementation with parallelized tree construction. CatBoost (Prokhorenkova et al., 2018) complements it by handling categorical features natively through ordered target encoding, avoiding explicit dummy coding. The two families differ in how individual-tree predictions are aggregated: random forests average independently fitted trees grown on bootstrap resamples, whereas boosting forms an additive ensemble of sequentially fitted trees, each correcting the residuals of its predecessor.

Ensemble Models To further improve predictive performance, we consider two ensemble approaches built on the same library of $K = 4$ base learners $\{\widehat{p}_k(\boldsymbol{x})\}_{k=1}^K$, consisting of ENet, Ranger, XGBoost, and CatBoost, where $\widehat{p}_k(\boldsymbol{x})$ denotes the estimated enrollment probability from the k-th base learner. We report five single models (LR, ENet, Ranger, XGBoost, CatBoost) together with the SuperLearner and Mixture-of-Experts (MoE) ensembles.

The SuperLearner (Van der Laan MJ et al., 2007) is a stacking ensemble that combines the base learners through a fixed non-negative weight vector $\boldsymbol{\alpha}$. The weights are derived from V-fold stratified cross-validation: each learner's out-of-fold predictions are regressed on the binary outcome under non-negative least squares, and the resulting $\widehat{\boldsymbol{\alpha}}$ is used to combine the base learners refit on the full training data, giving the ensemble prediction $\widehat{p}_{\mathrm{SL}}(\boldsymbol{x}) = \sum_{k=1}^K \widehat{\alpha}_k \, \widehat{p}_k(\boldsymbol{x})$. Because $\widehat{\boldsymbol{\alpha}}$ is selected by cross-validated risk, SuperLearner is asymptotically at least as good as the best member of its library.

MoE (Jacobs et al., 1991) replaces the fixed SuperLearner weights with a covariate-dependent *gate model* $\pi_k(\boldsymbol{x})$, allowing different learners to dominate in different regions of the feature space. The MoE prediction is the locally weighted convex combination

$$\widehat{p}_{\mathrm{MoE}}(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{x}) \, \widehat{p}_k(\boldsymbol{x}), \qquad \sum_{k=1}^{K} \pi_k(\boldsymbol{x}) = 1, \quad \pi_k(\boldsymbol{x}) \geqslant 0.$$

We estimate the gate via a fast two-stage best-expert classifier: for each training observation we record which base learner attains the smallest log-loss on its out-of-fold prediction, then fit an elastic-net-penalized multinomial logistic regression to those expert labels to obtain softmax gate probabilities $\pi_k(\boldsymbol{x})$. This is a fast variant of the EM-based MoE estimator originally proposed by Jacobs et al. (1991); full details of the gate-fitting procedure are given in Section S.1 of the Supplementary Material.

3.2 Interval Prediction

The aggregate enrollment count $C = \sum_{i=1}^n y_i$ over a test cohort of size n is the institutional target, with the natural point estimator $\widehat{C} = \sum_i \widehat{p}_i$. We construct $(1-\alpha)$ prediction intervals for C via three complementary methods, where α denotes the nominal significance level. The asymptotic

method uses a closed-form Gaussian approximation. The two simulation-based methods each generate B bootstrap samples $\{C^{(b)}\}_{b=1}^{B}$ and report the resulting $(1-\alpha)$ highest-density (HD) interval.

Asymptotic Method for LR For LR, the first method leverages the asymptotic normality of $\widehat{C}$ to obtain a closed-form interval, formalized in the following proposition.

Proposition 1. *Let $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{V}}$ denote the estimated coefficients and their corresponding covariance matrix from a fitted LR model. Consider a test set of size n, with $\boldsymbol{x}_i$ and y_i representing the covariate vector and the associated binary outcome for observation $i, i = 1, \ldots, n$. Let $C = \sum_{i=1}^{n} y_i$ be the actual total count. An estimate of C is given by $\widehat{C} = \sum_{i=1}^{n} \widehat{p}_i$, where $\widehat{p}_i = \operatorname{expit}(\boldsymbol{x}_i^{\top}\widehat{\boldsymbol{\beta}})$. Then, under regularity conditions, the prediction error $\widehat{C} - C$ is approximately normally distributed,*

$$\widehat{C} - C \overset{\text{approx}}{\sim} \mathcal{N}(0, \widehat{\sigma}_C^2),$$

where $\widehat{\sigma}_C^2$ is an estimated variance given by

$$\widehat{\sigma}_C^2 = \sum_{i=1}^{n} \widehat{p}_i(1-\widehat{p}_i) + \left[\sum_{i=1}^{n} \widehat{p}_i(1-\widehat{p}_i)\boldsymbol{x}_i^{\top}\right] \widehat{\boldsymbol{V}} \left[\sum_{i=1}^{n} \widehat{p}_i(1-\widehat{p}_i)\boldsymbol{x}_i\right].$$

An approximate $(1-\alpha)$ prediction interval for C is given by

$$\left[\widehat{C} - z_{1-\alpha/2}\,\widehat{\sigma}_C,\ \widehat{C} + z_{1-\alpha/2}\,\widehat{\sigma}_C\right],$$

where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$ quantile of the standard normal distribution.

The first term in $\widehat{\sigma}_C^2$ reflects the intrinsic binomial variance in the test set; the second term represents the estimation variance, arising from the uncertainty in estimating the coefficient $\boldsymbol{\beta}$. See details in Section S.2 of the Supplementary Material.

Parametric Simulation Method for Logistic Models The second method replaces the closed-form variance with a parametric simulation. Sampling $\boldsymbol{\beta}^{(b)} \sim \mathcal{N}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{V}})$, computing the corresponding probabilities, drawing Bernoulli outcomes, and summing yields an empirical distribution of total counts from which a HD interval is constructed (Algorithm 1). This propagates both coefficient and outcome uncertainty without requiring the delta method.

Model-Agnostic Bootstrapping For models that do not admit a closed-form sampling distribution (regularized regression, tree ensembles, stacked learners), we adopt a non-parametric bootstrap (Algorithm 2). Each replicate refits the classifier on a resampled training set, predicts probabilities on the test data, and simulates Bernoulli outcomes; the resulting total counts form an empirical distribution, which can be used to build HD intervals. The procedure is model-agnostic and captures variable-selection, estimation, and outcome uncertainty in a unified framework, at the cost of B refits.

3.3 Group Variable Importance Measures

To evaluate how much each feature group contributes to the full model's discriminative ability, we define the partial AUC (**pAUC**) of a group as the decrease in out-of-sample AUC when

Algorithm 1 Parametric simulation prediction interval for logistic regression.

Input: Training data $\mathcal{D} = \{(\boldsymbol{x}_i^{\text{tr}}, y_i^{\text{tr}})\}_{i=1}^N$, test data $\{\boldsymbol{x}_i\}_{i=1}^n$, number of replicates B, confidence level α.
Fit an LR model with coefficient estimates $\widehat{\boldsymbol{\beta}}$, and the corresponding covariance matrix estimates $\widehat{\boldsymbol{V}}$ from data $\mathcal{D}$
for $b = 1, 2, \ldots, B$ **do**
  Sample $\boldsymbol{\beta}^{(b)} \sim \mathcal{N}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{V}})$
  for $i = 1, 2, \ldots, n$ **do**
    Predict probability $\widehat{p}_i^{(b)} = \text{expit}(\boldsymbol{x}_i{}^\top \boldsymbol{\beta}^{(b)})$
    Sample pseudo outcome $\widehat{y}_i^{(b)} \sim \text{Bernoulli}(\widehat{p}_i^{(b)})$
  end for
  Compute total count: $\widehat{C}^{(b)} = \sum_{i=1}^n \widehat{y}_i^{(b)}$
end for
Output: Point estimate $\widehat{C} = \frac{1}{B}\sum_{b=1}^B \widehat{C}^{(b)}$, and the $(1-\alpha)$ prediction interval as the HD interval from $\{\widehat{C}^{(b)}\}_{b=1}^B$

Algorithm 2 Bootstrap prediction interval for general classification models.

Input: Training data $\mathcal{D} = \{(\boldsymbol{x}_i^{\text{tr}}, y_i^{\text{tr}})\}_{i=1}^N$, test data $\{\boldsymbol{x}_i\}_{i=1}^n$, number of replicates B, confidence level α.
for $b = 1, 2, \ldots, B$ **do**
  Sample bootstrap dataset $\mathcal{D}_b$ by sampling with replacement from $\mathcal{D}$
  Fit classification model $\widehat{f}_b$ using $\mathcal{D}_b$
  for $i = 1, 2, \ldots, n$ **do**
    Predict probability $\widehat{p}_i^{(b)} = \widehat{f}_b(\boldsymbol{x}_i)$
    Sample pseudo outcome $\widehat{y}_i^{(b)} \sim \text{Bernoulli}(\widehat{p}_i^{(b)})$
  end for
  Compute total count: $\widehat{C}^{(b)} = \sum_{i=1}^n \widehat{y}_i^{(b)}$
end for
Output: Point estimate $\widehat{C} = \frac{1}{B}\sum_{b=1}^B \widehat{C}^{(b)}$, and the $(1-\alpha)$ prediction interval as the HD interval from $\{\widehat{C}^{(b)}\}_{b=1}^B$

that group is removed from the full model, following a leave-one-group-out (LOGO) approach. Formally, let AUC_{full} denote the AUC of the model fitted on all feature groups and AUC_{-g} denote the AUC when group g is excluded. The pAUC for group g is

$$\text{pAUC}_g = \text{AUC}_{\text{full}} - \text{AUC}_{-g},$$

with larger values indicating greater unique predictive contribution. We compute pAUC for five representative single-model methods: LR, Elastic Net, Ranger, XGBoost, and CatBoost. For each group, we refit the selected models with that group excluded, retaining all preprocessing steps. For regularized regression models, the penalty parameter is reselected by 5-fold cross-validation. For tree-based models, all hyperparameters are held fixed at the values used for the full model. This LOGO approach has been formalized as a principled framework for assessing variable importance in both parametric and black-box models (Fisher et al., 2019; Lei et al., 2018).

Complementing pAUC, we define the marginal AUC (**mAUC**) of a feature group as the

increase in out-of-sample AUC when that group is added to a null (intercept-only) model. Formally, let $\mathrm{AUC}_{\mathrm{null}}$ denote the AUC of the null model and AUC_{+g} denote the AUC when only group g is included. The mAUC for group g is

$$\mathrm{mAUC}_g = \mathrm{AUC}_{+g} - \mathrm{AUC}_{\mathrm{null}},$$

with larger values indicating greater standalone predictive value. While pAUC measures a group's contribution in the presence of all other information, mAUC captures its intrinsic predictive signal in isolation.

Comparing pAUC and mAUC reveals whether a feature group's predictive value is unique or largely redundant with information contained in other groups. A group with high mAUC but low pAUC likely carries information redundant with other groups, whereas a group with high pAUC despite modest mAUC provides information that becomes available only in combination with the full covariate set (Pencina et al., 2008; Pepe et al., 2004). Because both measures are based on predictive performance, they should be interpreted as measures of predictive contribution rather than causal importance; their values depend on the feature groups included in the analysis and the correlation structure among predictors.

4 Simulation Studies

To assess the validity of the proposed prediction intervals, we conduct a *semi-parametric bootstrap population simulation* study that reuses the observed covariates $\boldsymbol{x}_i$ from the applicant pool while regenerating the binary outcomes $\tilde{y}_i$ from a benchmark probability model. The design avoids embedding the same parametric assumptions used by the candidate interval methods. We describe the design and evaluation criteria below, then report empirical coverage, interval width, and aggregate-count error across $R = 200$ Monte Carlo replicates.

4.1 Design

A coverage study requires a ground-truth enrollment probability $\tilde{p}_i$ for each applicant against which the candidate intervals can be evaluated. We use the SuperLearner ensemble defined in Section 3 as the data-generating process, fitted on the pooled 2020–2025 in-state applicant data ($N = 40{,}496$, observed yield rate 32.2%). Two natural candidates for the data-generating process (DGP) are the SuperLearner and MoE ensembles introduced in Section 3; we use the SuperLearner in the main text, with a parallel simulation under an MoE DGP reported in Section S.3 of the Supplementary Material as a robustness check. The estimated non-negative least squares (NNLS) weights concentrate on the boosting learners ($\widehat{\alpha}_{\texttt{catboost}} \approx 0.59$ and $\widehat{\alpha}_{\texttt{xgboost}} \approx 0.41$, with the linear and random-forest learners receiving zero weight); the in-sample Area Under the Receiver Operating Characteristic Curve (AUROC or AUC for simplicity) and Area Under the Precision–Recall Curve (AUPRC) are 0.81 and 0.69, and the implied total $\sum_i \tilde{p}_i = 12{,}985$ closely matches the observed enrolled count of 13,022. We treat the resulting predicted probabilities $\tilde{p}_i$ as the ground-truth probabilities for generating synthetic enrollment outcomes throughout the simulation.

As an additional benchmark for the signal strength of the synthetic dataset, we report the expected AUROC and AUPRC of an oracle whose predictions equal $\tilde{p}_i$ when evaluated on Bernoulli outcomes $y_i \sim \mathrm{Ber}(\tilde{p}_i)$: $\mathbb{E}[\mathrm{AUROC}] \approx 0.74$ and $\mathbb{E}[\mathrm{AUPRC}] \approx 0.58$, each computable in closed form from $\{\tilde{p}_i\}$ alone.

Algorithm 3 Semi-parametric bootstrap population simulation for coverage evaluation.

Input: Full applicant pool $\mathcal{H} = \{(\boldsymbol{x}_i)\}_{i=1}^{N}$ across all six cohorts (2020–2025), number of replicates R, nominal level $1-\alpha$.
Fit benchmark model (SuperLearner) on $\mathcal{H}$ to obtain $\tilde{p}_i$ for each applicant i.
Set test size $n = \lfloor N/6 \rfloor$ and training size $N_{\text{tr}} = N - n$.
for $r = 1, 2, \ldots, R$ **do**
Sample n applicants without replacement from $\mathcal{H}$ to form the test index set $\mathcal{T}_r$; the remaining N_{tr} applicants form the training index set
For $i \in \mathcal{T}_r^c$, generate $\tilde{y}_i^{\text{tr}} \sim \text{Bernoulli}(\tilde{p}_i)$ to form training set $\mathcal{D}_r^{\text{tr}}$
For $i \in \mathcal{T}_r$, generate $\tilde{y}_i \sim \text{Bernoulli}(\tilde{p}_i)$ to form the test set; compute true count $C^{(r)} = \sum_{i \in \mathcal{T}_r} \tilde{y}_i$
Refit each candidate model on $\mathcal{D}_r^{\text{tr}}$ and construct the $(1-\alpha)$ prediction interval $\widehat{I}^{(r)}$ for $C^{(r)}$ of the test set
Record $\mathbb{I}\big(C^{(r)} \in \widehat{I}^{(r)}\big)$
end for
Output: Empirical coverage rate $\frac{1}{R}\sum_{r=1}^{R} \mathbb{I}(C^{(r)} \in \widehat{I}^{(r)})$; average interval width $\frac{1}{R}\sum_{r=1}^{R} |\widehat{I}^{(r)}|$

We carried out $R = 200$ Monte Carlo replicates with $B = 200$ inner bootstrap draws. In each Monte Carlo replicate, we draw a test set of sample size $n = \lfloor N/6 \rfloor = 6{,}749$ without replacement from the full pool $\mathcal{H}$ and use the remaining applicants ($N_{\text{tr}} = 33{,}747$) as the training set, generate synthetic outcomes from the benchmark probabilities, refit each candidate model on the synthetic training data, and construct a prediction interval for the true test-set count using the appropriate method (asymptotic, parametric simulation, or bootstrap). Re-splitting preserves the covariate distribution across replicates while letting Bernoulli sampling and refit estimation contribute their respective uncertainties to the resulting count. Algorithm 3 gives the full procedure.

For each combination of candidate model and interval method, we evaluate coverage, precision, point accuracy, and discrimination through five replicate-level quantities: the coverage indicator $\mathbb{I}\big(C^{(r)} \in \widehat{I}^{(r)}\big)$, the interval width $|\widehat{I}^{(r)}|$, the absolute count error $|\widehat{C}^{(r)} - C^{(r)}|$, and the AUROC and AUPRC of the refitted model on the synthetic test fold. The first measures coverage relative to the nominal level at $\alpha = 0.05$; the second measures precision, with narrower intervals preferred when nominal coverage is maintained; the third measures the accuracy of the aggregate count point estimate; and the last two assess whether the refitted model preserves individual-level discrimination under the synthetic data-generating process. We summarize each quantity by its replicate mean together with its Monte Carlo standard error.

4.2 Results

High individual-level discrimination does not guarantee reliable prediction intervals for aggregate enrollment forecasting. Individual-applicant discrimination, summarized by AUROC and AUPRC in Table 1 and displayed as per-replicate boxplots in Figure 2, separates the methods into three tiers: the two gradient-boosting models (XGBoost, CatBoost) and the two ensembles (SuperLearner, MoE) form the top tier, ENet sits just below, and Ranger and LR anchor the bottom tier. The observed discrimination levels are also close to the oracle benchmark implied by the data-generating process, where an oracle classifier that predicts using the true enrollment probabilities $\tilde{p}_i$ attains an expected AUROC of approximately 0.74 and an expected AUPRC

Table 1: Simulation results 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	$\mathbb{I}(C \in \widehat{I})$	$\lvert\widehat{I}\rvert$	$\lvert\widehat{C} - C\rvert$	AUROC (%)	AUPRC (%)
LR	Asymptotic	0.960	154.6 (0.0)	29.5 (1.7)	70.3 (0.04)	52.1 (0.08)
	Para-Sim	0.940	150.2 (0.6)	29.6 (1.6)	70.3 (0.04)	52.1 (0.08)
ENet	Bootstrap	0.955	151.1 (0.7)	29.0 (1.6)	72.0 (0.05)	54.8 (0.08)
Ranger	Bootstrap	0.775	149.1 (0.7)	51.3 (2.1)	70.2 (0.05)	51.7 (0.08)
XGBoost	Bootstrap	0.895	146.9 (0.7)	35.9 (2.0)	72.4 (0.04)	55.8 (0.08)
CatBoost	Bootstrap	0.870	156.3 (0.7)	44.5 (2.3)	72.9 (0.04)	56.5 (0.08)
SuperLearner	Bootstrap	0.770	148.8 (0.7)	51.1 (2.1)	72.9 (0.04)	56.6 (0.08)
MoE	Bootstrap	0.955	148.3 (0.6)	29.8 (1.7)	72.6 (0.04)	56.2 (0.08)

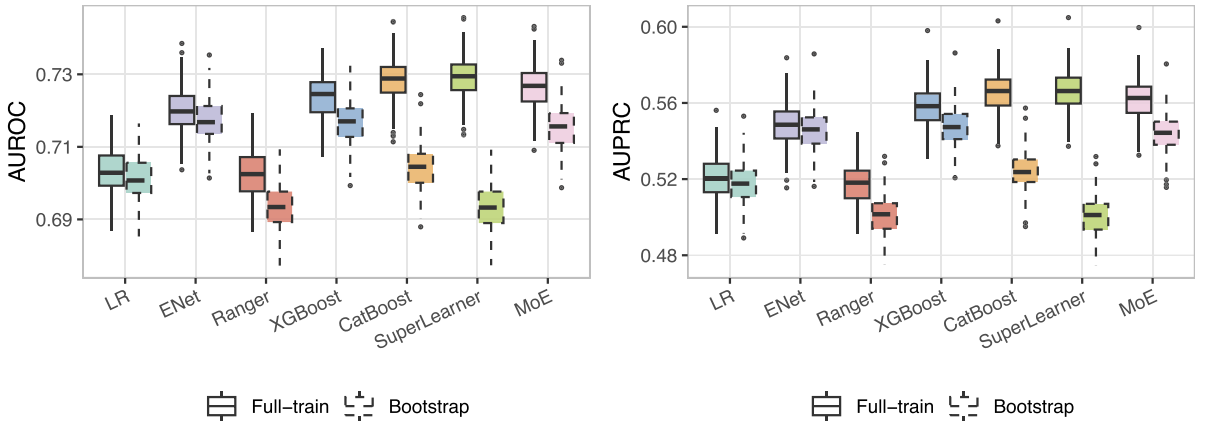


Figure 2: Per-replicate AUROC (a) and AUPRC (b) distributions across $R = 200$ Monte Carlo replicates. Solid boxes: full-train fit (one fit per replicate on the original training data); dashed boxes: bootstrap fits used inside the interval procedure (mean of $B = 200$ bootstrap refits per replicate). Discrimination separates the methods into three tiers under both the full-train (solid) and bootstrap (dashed) fits: gradient-boosting learners (XGBoost, CatBoost) and the two ensembles (SuperLearner, MoE) at the top, ENet in the middle, and Ranger and LR at the bottom. For LR, ENet, and MoE, the solid and dashed boxes overlap closely, indicating bootstrap stability; for CatBoost and SuperLearner, the bootstrap-fit boxes drop visibly below their full-train counterparts.

of approximately 0.58. The coverage results reveal a markedly different ranking. SuperLearner, XGBoost, and CatBoost all achieve top-tier discrimination yet undercover, demonstrating that strong classification performance does not automatically translate into reliable uncertainty quantification. Among the top-tier discriminators, only MoE attains close-to-nominal coverage, making it the only method that combines strong individual-level discrimination with reliable aggregate interval prediction.

Aggregate-count accuracy provides a complementary assessment of model performance that is not captured by either discrimination or interval coverage. Interval widths are remarkably similar across methods, but the count MAE values diverge sharply. With $n = 6{,}749$ and $\bar{p} \approx$

0.32, the per-replicate Bernoulli SD is approximately 38, implying an irreducible mean absolute deviation around $38\sqrt{2/\pi} \approx 30$ under approximate normality. LR, ENet, and MoE attain this Bernoulli floor; XGBoost and CatBoost exceed it moderately; Ranger and SuperLearner sit nearly 70% above it. The similarity in interval widths therefore masks substantial differences in the quality of the underlying probability surfaces. The elevated MAE values for Ranger and SuperLearner are more consistent with systematic probability bias than with residual Bernoulli variation alone.

Bootstrap stability provides a useful explanation for much of the undercoverage observed among the machine-learning learners. The bootstrap interval procedure implicitly assumes that each bootstrap-fit probability surface is a reasonable proxy for the full-train fit, that is, that the model is stable under row resampling. Comparing Table 1 (full-train fits) with Figure 2 (bootstrap fits) makes this assumption directly checkable. LR, ENet, and MoE satisfy it cleanly: their two views agree closely, and their bootstrap intervals attain close-to-nominal coverage. CatBoost and SuperLearner do not: their bootstrap-fit AUROC drops noticeably below the full-train value, and these are precisely the two methods with the largest undercoverage. Ranger provides a useful contrast. Its discrimination is lower under both fits, but the two views are not noticeably misaligned, indicating that bootstrap instability does not account for its undercoverage. This alignment between bootstrap-fit instability and undercoverage suggests a practical diagnostic: models whose bootstrap-fit and full-train performance differ substantially are candidates for unreliable bootstrap prediction intervals, even when their full-train discrimination is strong.

CatBoost and SuperLearner appear to lose bootstrap stability through fundamentally different mechanisms. CatBoost encodes high-cardinality categorical features through ordered target statistics that depend on within-category target rates; duplicated rows can inflate those rates in a small set of categories and pull the bootstrap-fit probability surface away from the full-train one. SuperLearner is reshaped through a different channel: each bootstrap sample contains repeated training rows, Ranger interpolates those duplicates essentially perfectly (Wyner et al., 2017), the in-sample loss that drives the NNLS weight selection shrinks artificially (Efron and Tibshirani, 1997; Janitza and Hornung, 2018), and the resulting weights concentrate on Ranger, $\widehat{\alpha}_{\text{ranger}} \approx 1$ across replicates (Mnich et al., 2020). Once the ensemble routes through a single base learner it loses the variance-reduction advantage that motivated the ensemble in the first place (Van der Vaart et al., 2006), inheriting Ranger's coverage signature across coverage, width, and MAE alike (Tang and Westling, 2024).

The contrasting performance of MoE and SuperLearner demonstrates that ensemble construction can be as important as the choice of base learners for reliable interval prediction. Averaged across all $R \times B$ inner draws, the MoE gate places weights of approximately 0.11, 0.43, 0.20, and 0.25 on ENet, Ranger, XGBoost, and CatBoost, respectively. Although Ranger receives the largest single weight on average, the gate is covariate-dependent: it routes individual applicants to the expert that performs best in their region of the feature space, so the bootstrap-fit and full-train views of MoE remain closely aligned in Figure 2. Table 1 correspondingly records nominal coverage, top-tier discrimination, and a small aggregate count error. This contrast with SuperLearner is particularly informative because the two ensembles are built from the same library of base learners yet produce markedly different coverage behavior. The same qualitative pattern holds under the MoE DGP reported in Section S.1 of the supplementary Material, indicating that the contrast is not an artifact of the particular DGP choice.

Reliable interval prediction depends more on the stability and calibration of the fitted probability surface than on discrimination performance alone. Across the candidate methods,

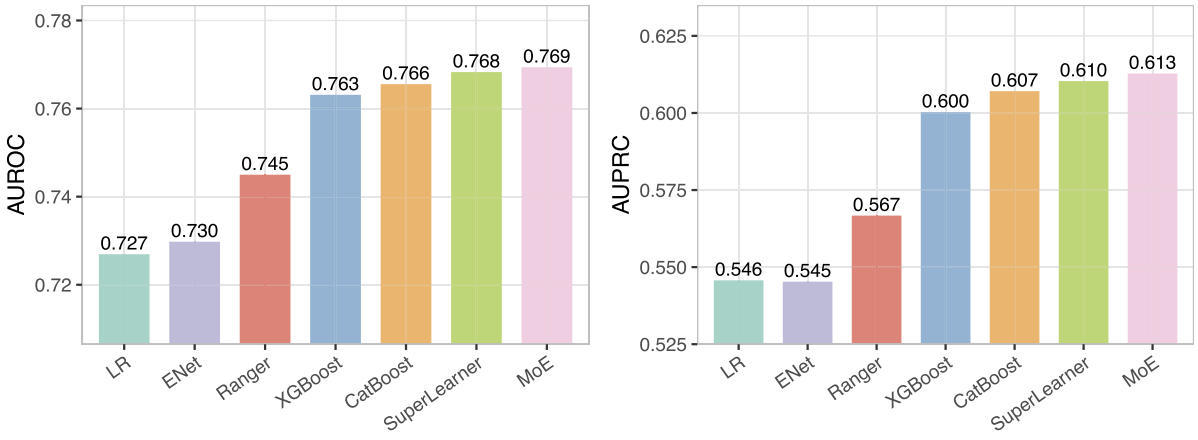


Figure 3: AUROC and AUPRC for the seven candidate methods on the 2025 cohort. Discrimination separates the methods into three tiers: the two ensembles (MoE, SuperLearner) lead, the gradient-boosting learners (CatBoost, XGBoost) follow, and the logistic models (LR, ENet) trail with Ranger intermediate; all methods nonetheless exceed an AUROC of 0.72.

close-to-nominal coverage is observed for LR, ENet, and MoE, whose bootstrap-fit and full-train performance remain closely aligned. In contrast, CatBoost and SuperLearner exhibit noticeable degradation under bootstrap refitting and correspondingly undercover, despite strong discrimination in the original fit. Ranger represents a different failure mode in which undercoverage occurs despite apparent bootstrap stability, suggesting that probability calibration remains important even when resampling stability is present. Taken together, the simulation results indicate that discrimination performance alone is insufficient for evaluating aggregate enrollment forecasts and that bootstrap stability provides a useful diagnostic for interval reliability.

5 Empirical Analysis

We apply the candidate methods, trained on the five preceding admission cycles (2020–2024), to the 2025 in-state freshman cohort. The analysis covers seven candidate methods: five single models (LR, ENet, Ranger, XGBoost, CatBoost) and two ensembles (SuperLearner, MoE). We first evaluate aggregate enrollment forecasts through individual-level discrimination, aggregate-count prediction, and prediction intervals for the 2025 cohort. We then examine group variable importance using the pAUC and mAUC measures introduced in Section 3 to identify the principal drivers of enrollment across model families.

5.1 Aggregate Enrollment Prediction

The candidate methods differ substantially in individual-level discrimination despite producing broadly similar aggregate enrollment forecasts. Figure 3 compares AUROC and AUPRC across the seven candidate methods on the 2025 cohort. Discrimination is led by the two ensembles, MoE and SuperLearner, which attain AUROC values near 77% and AUPRC values above 61%. The two gradient-boosting learners, CatBoost and XGBoost, form a second tier that trails the ensembles only slightly, while Ranger occupies an intermediate position. The logistic models, LR and ENet, attain the lowest discrimination, although their AUROC values remain above 72%.

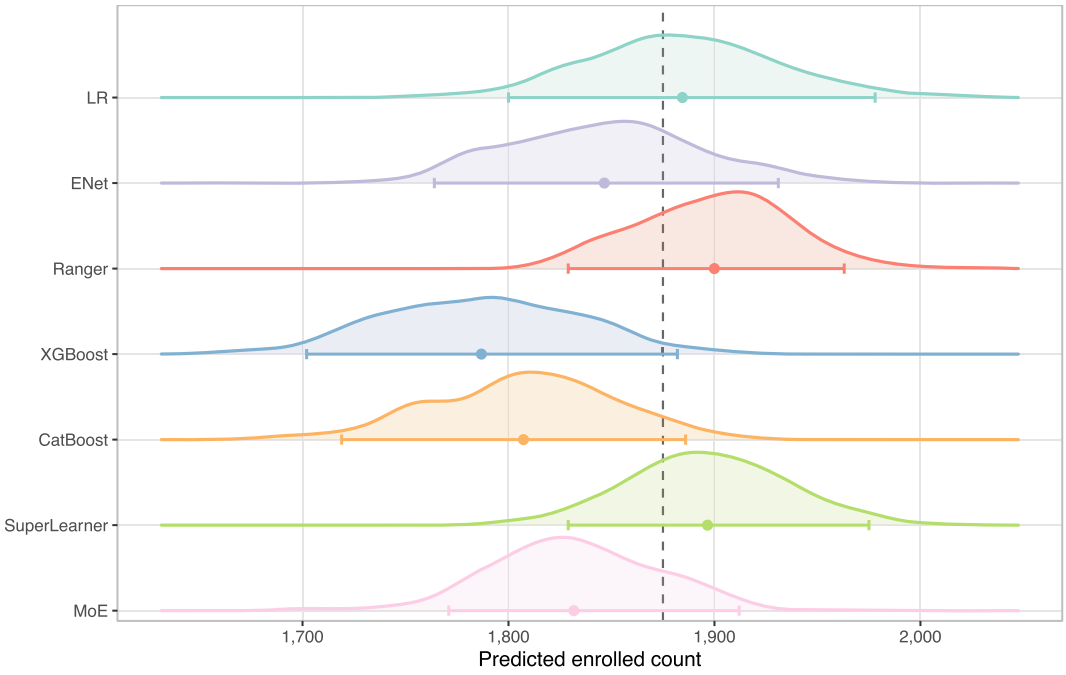


Figure 4: Predictive distributions of the total enrolled count for the 2025 cohort. Ridge densities show the bootstrap (parametric-simulation for LR) predictive distributions; the overlaid horizontal bars give the 95% HD intervals and the dots mark the point estimates. The dashed vertical line indicates the true 2025 enrollment count (1,875). The closed-form LR asymptotic interval, 1886 (1789, 1983), is nearly identical to LR para-sim and is omitted from the figure for clarity; see Table 2 for the full numerics. Every method's 95% prediction interval covers the true enrollment count, and the seven methods deliver broadly consistent aggregate forecasts despite the discrimination differences.

These results indicate that all candidate methods capture substantial signal in the enrollment process. They also show that strong individual-level discrimination does not necessarily imply superior aggregate forecasting performance, motivating a direct comparison of aggregate-count predictions and their associated uncertainty.

Despite substantial differences in discrimination performance, all seven methods produce aggregate enrollment forecasts that are broadly consistent with the observed 2025 enrollment count. Figure 4 displays the predictive distributions and 95% prediction intervals for the 2025 enrollment count, and Table 2 summarizes the corresponding numerical results. The observed enrollment count of 1,875 falls within the prediction interval of every method, indicating broad agreement regarding the size of the incoming class. Although the methods differ noticeably in discrimination performance, their aggregate forecasts are considerably more similar. Point estimates range from 1,787 for XGBoost to 1,900 for Ranger, corresponding to a spread of only 113 students across all seven methods. Likewise, the predictive distributions are concentrated near the observed count and exhibit substantial overlap. These results suggest that the candidate methods provide a largely consistent view of aggregate enrollment demand despite employing markedly different modeling strategies.

The candidate methods exhibit broadly similar uncertainty ranges but appreciable differ-

Table 2: Prediction interval results for total enrollment count in 2025. AUROC and AUPRC are from the full-train fit. LR uses asymptotic and parametric-simulation intervals; the remaining methods use 95% bootstrap intervals.

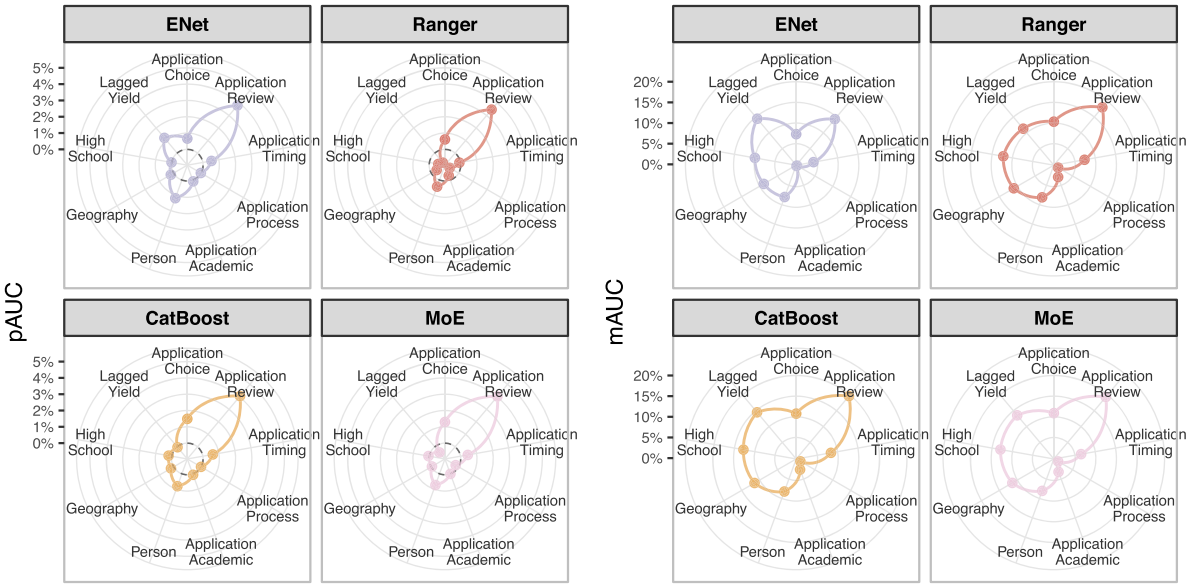
Model	Method	AUROC (%)	AUPRC (%)	Estimate	95% Interval
LR	Asymptotic	72.7	54.6	1886	(1789, 1983)
	Para-Sim	72.7	54.6	1884	(1800, 1978)
ENet	Bootstrap	73.0	54.5	1847	(1764, 1931)
Ranger	Bootstrap	74.5	56.7	1900	(1829, 1963)
XGBoost	Bootstrap	76.3	60.0	1787	(1702, 1882)
CatBoost	Bootstrap	76.6	60.7	1807	(1719, 1886)
SuperLearner	Bootstrap	76.8	61.0	1897	(1829, 1975)
MoE	Bootstrap	76.9	61.3	1832	(1771, 1912)

ences in their aggregate enrollment forecasts. Table 2 shows that LR, Ranger, and SuperLearner produce point forecasts above the observed count of 1,875, whereas XGBoost and CatBoost produce forecasts below the truth. ENet and MoE lie between these two extremes, with point estimates of 1,847 and 1,832, respectively. The observed differences are driven primarily by shifts in the centering of the predictive distributions rather than by large differences in interval width. For LR, the asymptotic and parametric-simulation intervals are nearly identical, with point estimates differing by only two students and interval endpoints that closely coincide. This agreement mirrors the close correspondence observed in the simulation study and provides empirical support for the accuracy of the closed-form variance approximation developed in Section 3. Collectively, the results indicate that different modeling strategies can yield noticeably different aggregate forecasts even when their associated uncertainty ranges remain comparable.

The empirical results are broadly consistent with the patterns observed in the simulation study while illustrating the limitations of evaluating interval performance from a single enrollment cycle. The gradient-boosting methods place most of their predictive mass slightly below the observed count, mirroring the mild undercoverage observed in the simulation. Ranger provides a useful contrast: its interval covers the observed count in 2025 despite exhibiting pronounced undercoverage across the Monte Carlo replicates. This apparent discrepancy is not unexpected because a single realization provides little information about long-run coverage properties. The simulation results therefore remain essential for distinguishing between methods that happen to cover the truth in one cohort and methods that reliably attain nominal coverage across repeated samples. Viewed jointly with the discrimination results in Figure 3 and the coverage results in Section 4, MoE provides the strongest overall combination of individual-level discrimination, aggregate-count accuracy, and interval reliability.

5.2 Group Variable Importance

The dominant drivers of enrollment remain remarkably consistent across the major model families considered in this study. Figure 5 presents pAUC and mAUC for the nine feature groups across four representative models on the 2025 cohort: ENet from the logistic family, Ranger from random forests, CatBoost from gradient boosting, and MoE from the ensemble methods. The two measures quantify a feature group's unique contribution to the full model and its standalone predictive signal, respectively. Despite substantial differences in model structure, the four



(a) pAUC (b) mAUC

Figure 5: pAUC and mAUC for the nine feature groups across the four representative models on the 2025 cohort. Application Review dominates both measures across all four models; High School and Geography show high mAUC but low pAUC, indicating overlapping group information.

methods exhibit broadly similar importance rankings. Application Review consistently emerges as the most influential feature group, while Application Choice and Person form a second tier of contributors across the models. At the other end of the spectrum, Application Process and Application Academic generally provide weaker predictive signal. The remaining three candidate models, LR, XGBoost, and SuperLearner, closely track the corresponding representative models and are reported in Section S.4 of the Supplementary Material. These results suggest that the principal drivers of enrollment are largely insensitive to the choice of prediction model.

The contrast between pAUC and mAUC reveals important differences between unique and standalone predictive value. Application Review attains the largest pAUC and mAUC values across all four representative models, indicating that admissions-review assessments provide both the strongest unique contribution in the presence of other predictors and the strongest predictive signal in isolation. The pAUC profiles are particularly concentrated, with Application Review accounting for most of the unique predictive information and Application Choice and Person serving as secondary contributors. In contrast, the mAUC profiles reveal a broader set of informative feature groups. Beyond Application Review, High School, Geography, Person, and Application Choice all exhibit substantial standalone predictive value, while Application Timing contributes approximately 5% mAUC across models. Application Process and Application Academic consistently rank among the weakest groups under mAUC. Taken together, these results suggest that several feature groups contain useful enrollment information individually, but much of that information becomes redundant once the strongest predictors are incorporated into the full model.

The treatment of Lagged Yield highlights how variable importance depends on model rep-

resentation rather than information content alone. Under pAUC, Lagged Yield contributes positively for ENet, where removing the group reduces discrimination by roughly one percentage point of AUC. In contrast, Ranger, CatBoost, and MoE exhibit slightly negative pAUC values, indicating that the engineered Lagged Yield summary provides little unique information to improve out-of-sample discrimination once the full set of predictors is available. The mAUC results reveal a similar contrast. For ENet, Lagged Yield is among the strongest standalone predictors, whereas for Ranger, CatBoost, and MoE it clusters with Application Choice, High School, Geography, and Person at comparable levels of predictive signal. This difference reflects the distinct ways in which the models access historical enrollment information. LR and ENet rely on the engineered Lagged Yield variables to summarize high-cardinality predictors such as ZIP codes and high schools, whereas tree-based learners can split directly on those raw categorical features and recover similar information through higher-order interactions. The results therefore suggest that historical enrollment patterns remain informative, but their measured importance depends strongly on how that information is represented within the prediction model.

6 Discussion

Reliable interval prediction for aggregate enrollment forecasting depends on the quality and stability of the estimated probability surface rather than on discrimination performance alone. Across both the simulation study and the empirical application, models with strong AUROC and AUPRC performance did not always provide reliable uncertainty quantification, whereas models with stable probability estimates under perturbations of the training data tended to produce prediction intervals with close-to-nominal coverage. These findings suggest that discrimination and interval validity reflect different aspects of predictive performance and should be evaluated separately when aggregate forecasts are the primary objective. More generally, the results highlight the importance of assessing not only the accuracy of predicted probabilities but also their stability, since prediction intervals inherit deficiencies in the underlying probability surface.

The variable-importance analysis demonstrates that the predictive value of a feature group depends not only on the information it contains but also on how that information is represented within a particular learning algorithm. Lagged Yield illustrates this point particularly well. It was among the most influential predictors for the logistic models but contributed little unique information for the tree-based methods, which were able to recover similar signals directly from high-cardinality identifiers such as high school and geographic location. More broadly, the results indicate that the importance of engineered features is inherently model dependent and cannot be assessed independently of the alternative information pathways available to the learner. For enrollment managers, these findings suggest that institutionally generated information, including admissions-review assessments and historical enrollment patterns contribute substantially to enrollment prediction.

The generalizability of the proposed framework depends on both its ability to transfer across enrollment settings and the validity of the underlying probability estimator. The analysis was conducted using pooled 2020–2024 admission cycles from a single institution and therefore does not explicitly account for temporal shifts in applicant behavior, recruitment strategies, or external factors that may influence enrollment decisions. In particular, the 2020 and 2021 cycles overlap with pandemic-era changes in application behavior, standardized testing, and campus operations that are unlikely to recur in later cycles such as 2025, so equally weighting all training cohorts may attenuate signal relevant to the target year. Future work could incorporate rolling-window training, covariate-shift adjustment, or hierarchical modeling to better accommodate

temporal variation, while replication across multiple institutions would provide a stronger assessment of transferability. In addition, the validity of the bootstrap prediction intervals ultimately depends on the quality of the estimated probability surface. Probability calibration should therefore be evaluated before deployment, potentially through calibration diagnostics supplemented by recalibration procedures such as Platt scaling (Platt, 1999) or isotonic regression (Zadrozny and Elkan, 2002). The bootstrap-fit probability surface should also remain a reasonable proxy for the full-train fit, since substantial divergence between the two can compromise interval validity even when discrimination performance remains strong. These considerations suggest that calibration and bootstrap stability should be treated as routine diagnostics when applying the proposed framework to new enrollment settings.

Supplementary Material

The online supplementary material consists of two files. The file `supplement.pdf` contains introduction of MoE (Section S.1), the theoretical derivations referenced in the main text (Section S.2), additional simulation results including the MoE DGP robustness study (Table S.1 and Figure S.1 in Section S.3), and expanded variable-importance figures (Figure S.2 in Section S.4). The file `code.zip` contains the complete R reproduction pipeline, including data generation, model fitting for all seven candidate methods, construction of the asymptotic, parametric-simulation, and bootstrap prediction intervals, and computation of the mAUC and pAUC feature-group importance measures. The archive also includes a synthetic dataset that preserves the structure of the applicant records used in the empirical analysis while protecting individual-level information.

Acknowledgement

The authors thank the Guest Editors and the two reviewers for their helpful comments, which have led to an improved version of the manuscript.

References

Ab Ghani NL, Che Cob Z, Mohd Drus S, Sulaiman H (2019). Student enrolment prediction model in higher education institution: A data mining approach. In: Othman MA, Abd Aziz MZA, Md Saat MS, Misran MH (Eds.), *Proceedings of the 3rd International Symposium of Information and Internet Technology (SYMINTECH 2018)*, volume 565 of *Lecture Notes in Electrical Engineering*, 43–52. Springer International Publishing.

Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L , . . . , Nahavandi S (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76: 243–297. https://doi.org/10.1016/j.inffus.2021.05.008

Abelt J, Browning D, Dyer C, Haines M, Ross J, . . . , Gerber M (2015). Predicting likelihood of enrollment among applicants to the UVa undergraduate program. In: *2015 Systems and Information Engineering Design Symposium*, 194–199. IEEE.

Aulck L, Nambi D, West J (2020). Increasing enrollment by optimizing scholarship allocations using machine learning and genetic algorithms. In: Rafferty AN, Whitehill J, Cavalli-Sforza V, Romero C (Eds.), *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*, 29–38. ERIC.

Breiman L (2001). Random forests. *Machine Learning*, 45(1): 5–32. https://doi.org/10.1023/A:1010933404324

Brinkman PT, McIntyre C (1997). Methods and techniques of enrollment forecasting. *New Directions for Institutional Research*, 93: 67–80. https://doi.org/10.1002/ir.9305

Chatfield C (1993). Calculating interval forecasts. *Journal of Business & Economic Statistics*, 11(2): 121–135. https://doi.org/10.1080/07350015.1993.10509938

Chen T, Guestrin C (2016). XGBoost: A scalable tree boosting system. In: Krishnapuram B, Shah M, Smola AJ, Aggarwal CC, Shen D, Rastogi R (Eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Cirelli J, Konkol AM, Aqlan F, Nwokeji JC (2018). Predictive analytics models for student admission and enrollment. In: *Proceedings of the International Conference on Industrial Engineering and Operations Management*, 1395–1403.

Efron B, Tibshirani R (1997). Improvements on cross-validation: The 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438): 548–560. https://doi.org/10.2307/2965703

Esquivel JA, Esquivel JA (2021). A machine learning based DSS in predicting undergraduate freshmen enrolment in a Philippine university. *International Journal of Computer Trends and Technology*, 69: 50–54. https://doi.org/10.14445/22312803/IJCTT-V69I5P107

Fisher A, Rudin C, Dominici F (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177): 1–81.

Gawlikowski J, Tassi CRN, Ali M, Lee J, Humt M, ..., Zhu XX (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1): 1513–1589. https://doi.org/10.1007/s10462-023-10562-9

Gneiting T, Katzfuss M (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1: 125–151. https://doi.org/10.1146/annurev-statistics-062713-085831

Goenner CF, Pauls K (2006). A predictive model of inquiry to enrollment. *Research in Higher Education*, 47: 935–956. https://doi.org/10.1007/s11162-006-9021-8

Harrington CF, Schibik T (2013). An econometric approach to optimizing student enrollment. *Journal of Higher Education Theory and Practice*, 13(1): 45–55.

Hayes JB, Price RA, York RP (2013). A simple model for estimating enrollment yield from a list of freshman prospects. *Academy of Educational Leadership Journal*, 17(2): 61–68.

Hoenack SA, Pierro DJ (1990). An econometric model of a public university's income and enrollments. *Journal of Economic Behavior & Organization*, 14(3): 403–423. https://doi.org/10.1016/0167-2681(90)90067-N

Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1): 79–87. https://doi.org/10.1162/neco.1991.3.1.79

Jamison J (2017). Applying machine learning to predict Davidson College's admissions yield. In: Caspersen ME, Edwards SH, Barnes T, Garcia DD (Eds.), *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*, 765–766.

Janitza S, Hornung R (2018). On the overestimation of random forest's out-of-bag error. *PLoS ONE*, 13(8): e0201904. https://doi.org/10.1371/journal.pone.0201904

Kye A (2023). Comparative analysis of classification performance for US college enrollment predictive modeling using four machine learning algorithms (logistic regression, decision tree, support vector machine, artificial neural network), Ph.D. thesis, Loyola University Chicago.

Langston R, Loreto D (2017). Seamless integration of predictive analytics and CRM within an

undergraduate admissions recruitment and marketing plan. *Strategic Enrollment Management Quarterly*, 4(4): 161–172. https://doi.org/10.1002/sem3.20095

Langston R, Wyant R, Scheid J (2016). Strategic enrollment management for chief enrollment officers: Practical use of statistical and mathematical data in forecasting first year and transfer college enrollment. *Strategic Enrollment Management Quarterly*, 4(2): 74–89. https://doi.org/10.1002/sem3.20085

Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111. https://doi.org/10.1080/01621459.2017.1307116

Liu S, Lu D, Painter SL, Griffiths NA, Pierce EM (2023). Uncertainty quantification of machine learning models to improve streamflow prediction under changing climate and environmental conditions. *Frontiers in Water*, 5: 1150126. https://doi.org/10.3389/frwa.2023.1150126

Lopez LJL, Elsharief S, Jorf DA, Darwish F, Ma C, Shamout FE (2025). Uncertainty quantification for machine learning in healthcare: A survey. In: Xu XO, Choi E, Singhal P, Gerych W, Tang S, Agrawal M, Subbaswamy A, Sizikova E, Dunn J, Daneshjou R, Sarker T, McDermott M, Chen I (Eds.), *Proceedings of the Sixth Conference on Health, Inference, and Learning*, 862–907.

Mnich K, Kitlas Golińska A, Polewko-Klim A, Rudnicki WR (2020). Bootstrap bias corrected cross validation applied to Super Learning. ArXiv preprint: arXiv:2003.08342.

Pencina MJ, D'Agostino RB, Vasan RS (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, 27(2): 157–172. https://doi.org/10.1002/sim.2929

Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology*, 159(9): 882–890. https://doi.org/10.1093/aje/kwh101

Platt J (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola A.J., Bartlett P.L., Schölkopf B., Schuurmans D. (Eds.), *Advances in Large Margin Classifiers*, 61–74.

Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A (2018). Catboost: Unbiased boosting with categorical features. In: Bengio S, Wallach HM, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (Eds.), *Advances in Neural Information Processing Systems*, volume 31.

Tang Z, Westling T (2024). Consistency of the bootstrap for asymptotically linear estimators based on machine learning.

Trusheim D, Rylee C (2011). Predictive modeling: Linking enrollment and budgeting. *Planning for Higher Education*, 40(1): 12–21.

Van der Laan MJ, Polley EC, Hubbard AE (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1): 1–23.

Van der Vaart AW Dudoit S, Van der Laan MJ (2006). Oracle inequalities for multi-fold cross validation. *Statistics & Decisions*, 24(3): 351–371. https://doi.org/10.1524/stnd.2006.24.3.351

Wright MN, Ziegler A (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1): 1–17. https://doi.org/10.18637/jss.v077.i01

Wyner AJ, Olson M, Bleich J, Mease D (2017). Explaining the success of AdaBoost and random forests as interpolating classifiers. *Journal of Machine Learning Research*, 18(48): 1–33.

Zadrozny B, Elkan C (2002). Transforming classifier scores into accurate multiclass probability estimates. In: Grossman RL, Han J, Kumar V, Motwani R, Agrawal R (Eds.), *Proceedings of*

the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 694–699.

Zou H, Hastie T (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 67(2): 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x