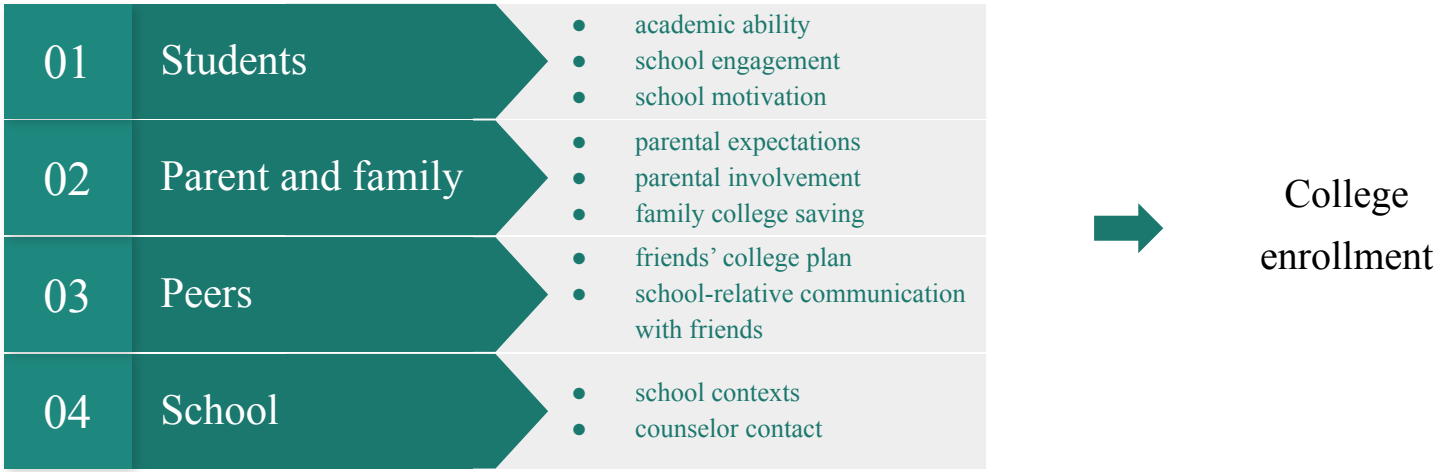
Using Machine Learning to Predict College Enrollment for Low Socioeconomic-Status (SES) Students

Surina He, Mehrdad Yousefpoori-Naeim, Ying Cui, Maria Cutumisu
Measurement, Evaluation and Data Science Program
University of Alberta, Canada

Introduction

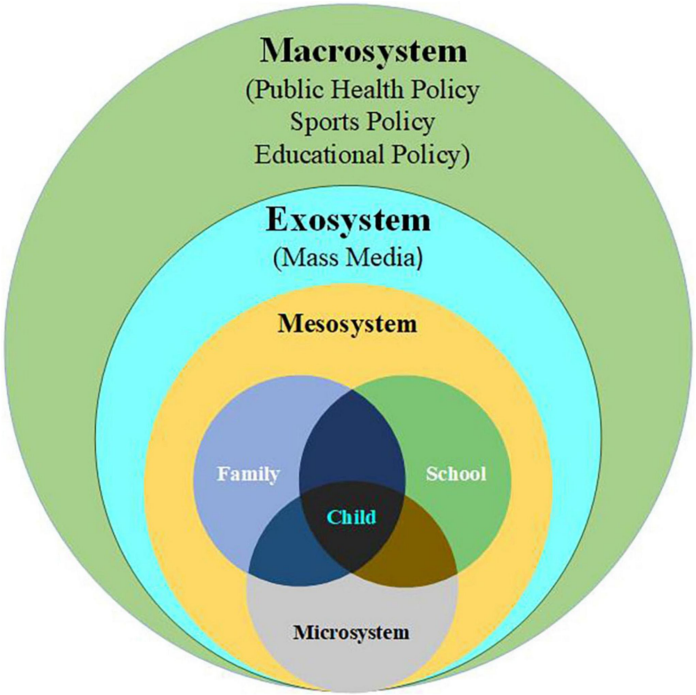
- College enrollment is an important topic in the education research. It displays a strong positive association with an individual's employment rate, income level, and well-being (Ou & Reynolds, 2008).
- Previous studies used traditional statistical analysis to examine:



Introduction

- Bronfenbrenner's ecological system theory:
  1. Microsystem: directly interact with student
  2. Mesosystem: interaction between microsystem
  3. Exosystem: indirectly impact student
  4. Macrosystem: societal and cultural contexts

- Factors from individuals, microsystems, and mesosystem have greatest impact on individual's development.



(Bronfenbrenner, 1992)

Introduction

- Traditionally, due to computational limitations, researchers only focused on a small number of factors at a time. Nowadays, with the aid of machine learning techniques (ML), a large number of factors could be included in one model based on the theoretical framework.
- Some studies already applied ML to predict college enrollment:

Yang et al. (2021)

➔ Forecasted number of freshman enrollments in a Chinese province before students' registration

➔ Back propagation neural network > decision tree, random forest

Slim et al. (2018)

➔ Projected student enrollment in the university to aid colleges in identifying characteristics that could maximize the prediction of students with a higher tendency to enroll at their institutions.

➔ Amount of scholarships, the time of the admission decision, high-school GPA, residency status.

Introduction

- Previous studies:
  - Focused on forecasting college enrollment numbers or rates from the school administration's perspective to aid college admissions
  - Identified predictors are not easily modifiable by students, teachers, administrators, or other stakeholders (i.e., non-actionable predictors)
  - Has no study identified predictors of low-SES students' college enrollment

Purpose of this study: identifying the most influential actionable characteristics that contribute to the accurate prediction of low SES high-school students' college enrollment

Method

- **Data:** High School Longitudinal Study of 2009 (HSLS:09)

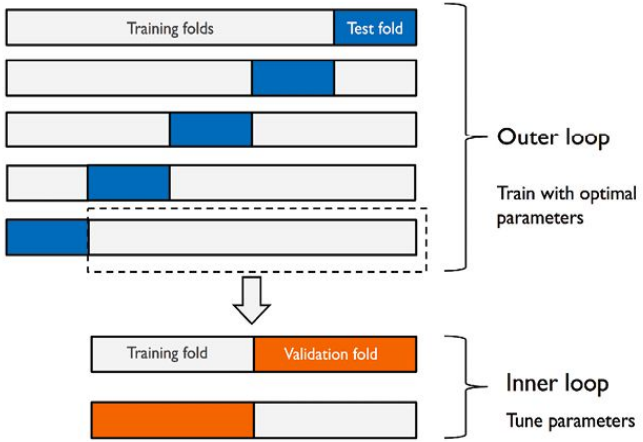
  (1st round in 2009, 9th grade → 2nd round in 2012, 12th grade → 3rd round in 2016)
- **Sample**: 7,139 9th low SES graders, their parents, teachers, school principals
- **Data pre-processing**:
  - Removed cases that parental questionnaire was not answered by their biological/stepparents
  - Removed variables with a missing rate of 30% or higher and imputed rest of missing values
  - Re-coded college enrollment to three groups (0 = not enrolled, 1 = enrolled in a 2-year or less institution, and 2 = enrolled in a 4-year institution)
  - Removed variables that were not correlate with college enrollment, and not significant in Lasso
  - Balanced groups using Synthetic Minority Oversampling Technique-Nominal
- **Final data**: 28 predictors, 7,668 (2,556 for each groups) 9th low-SES graders.

Method

- **Machine learning classifiers:**

  logistic regression, k-nearest neighbours, support vector machines, decision tree, random forest

- **Model evaluation procedure:** nested cross-validation

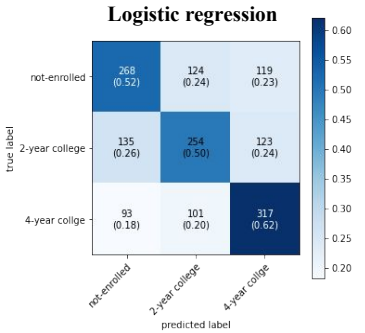


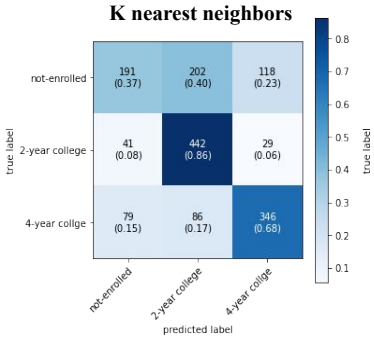
- **Model interpretation:** Explainable Artificial Intelligence (XAI) method

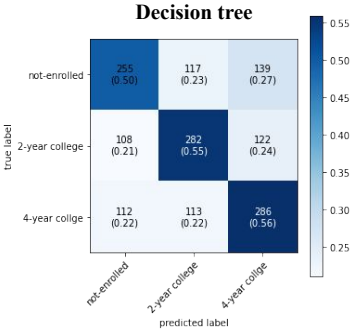
Results

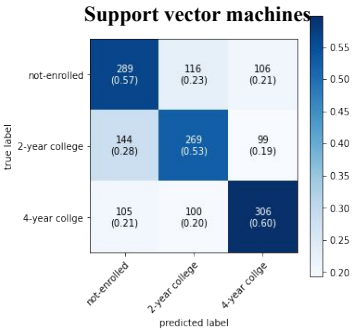
- **Model performance:**

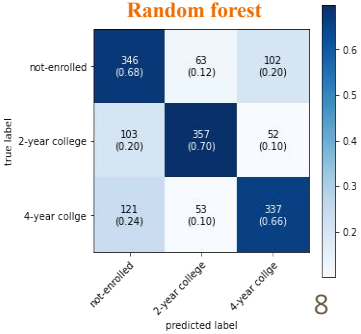
Classifier	Accuracy	Micro and Macro Precision	Micro and Macro Recall	Micro and Macro F1-Score	ROC-AUC Score
Logistic Regression	55.58% ± 0.65	0.547, 0.546	0.547, 0.547	0.547, 0.546	0.716
K-Nearest Neighbors	67.12% ± 1.06	0.638, 0.640	0.638, 0.638	0.638, 0.621	0.811
Decision Tree	56.99% ± 0.52	0.537, 0.537	0.537, 0.537	0.537, 0.537	0.652
Support Vector Machine	68.00% ± 0.94	0.563, 0.563	0.563, 0.563	0.563. 0.563	0.732
Random Forest	**70.12% ± 0.94**	**0.678, 0.683**	**0.678, 0.678**	**0.678, 0.680**	**0.841**





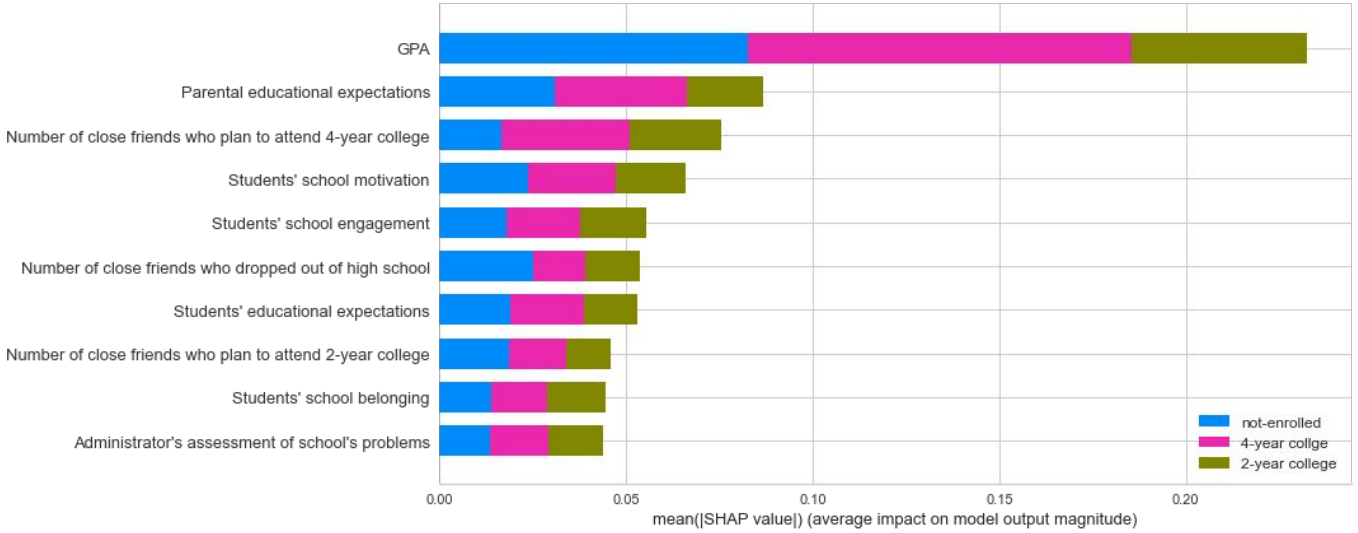






Results

- **Top 10 feature importance:**

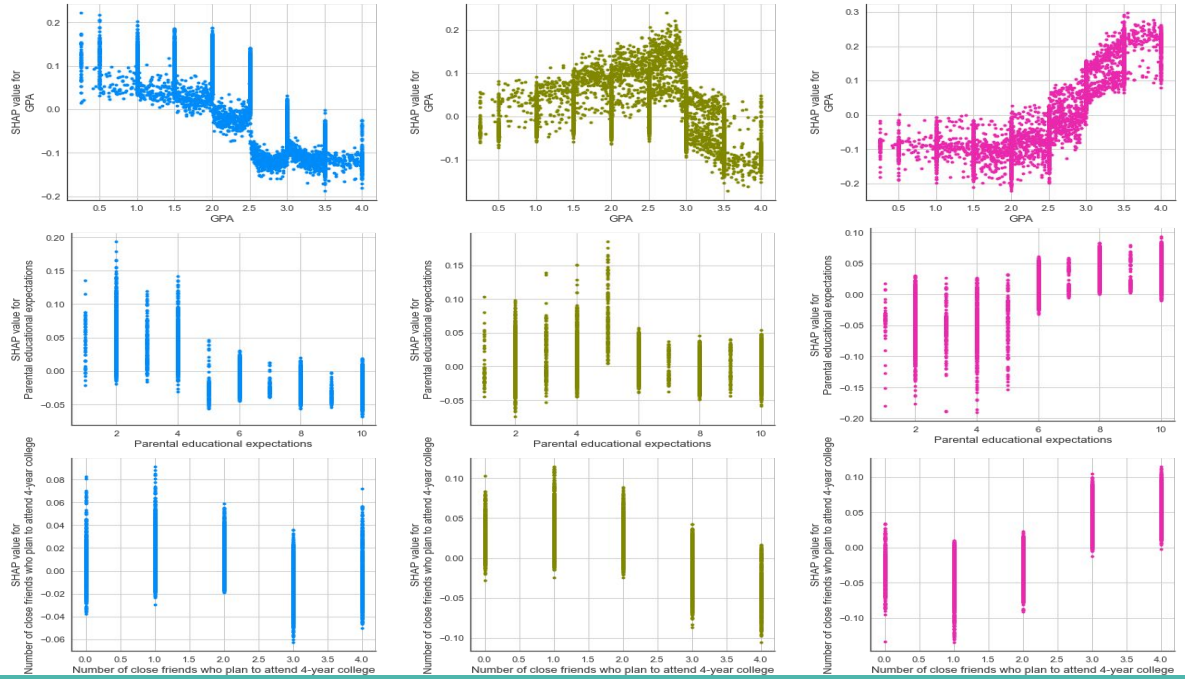


GPA; educational expectations; friend's college plan; students' school motivation, engagement, belonging

Results

- **Top 3 Features' relationship with the probability of enrolling in different college :**

Not enrolled | 2-year or less college | 4-year college



Discussion

- This study identified top 10 predictors for low-SES students' college enrollment from over 100 variables. 5 from student, 3 from peer, 1 from family, and 1 from school. This finding is consistent with the previous assertion that the influence of individual characteristics is greater than that of the microsystem (i.e., parent, peers, school) or of the three other systems.

- GPA is the most important predictor. The relationship between GPA and its importance for the three enrollment statuses varied, with a negative pattern for no enrollment, a quadratic pattern for 2-year or less college enrollment, and a positive pattern for 4-year college enrollment.

Discussion

- There are 4 future-oriented goals (i.e., student and parental educational expectations, close friends' plan to attend a 2-year college or 4-year college). According to Miller and Brickman's (2004) model of future-oriented motivation and self-regulation, the future-oriented goals are self-relevant, self-defining long-term goals that provide incentive for action. Educational practitioners could enhance students' future oriented goals to promote their college enrollment.

- Limitations:
  - HSLS contained thousands of variables, we just selected 100+ variables manually at first based on previous research and theory. But some other variables that not be chosen might also have effects for college enrollments.
  - Sample is from United State, so should be carefully generalized to other countries.

Thank you for your listening !

Q & A

Surina He	surina1@ualberta.ca
Mehrdad Yousefpoori-Naeim	yousefpo@ualberta.ca
Ying Cui	yc@ualberta.ca
Maria Cutumisu	cutumisu@ualberta.ca