MLDS Center

Maryland Longitudinal Data System

Better Data • Informed Choices • Improved Results

An Application of Machine Learning in Causal Inference

Ashani Jayasekera & Tracy Sweet
MLDS Research Branch and UMCP

MLDS Research Series
May 9, 2025

Introduction

- The national *15 to Finish* movement was put forth by the Complete College America organization to encourage students to enroll in at least 15 credits (or 30 credits per year) to put students on track to graduate "on time".

- Maryland General Assembly passed the College Affordability Act of 2016 which added a merit-based eligibility criterion to the Delegate Howard P. Rawlings Education Excellence Awards (EEA) Program.
  - → Requires that students complete at least 30 credits to continue receiving financial aid the following academic year

Research Questions

- The current study uses predictive machine learning (ML) algorithms in tandem with propensity score (PS) analysis to evaluate whether a student who takes 15 or more credits in their first semester at a Maryland university is more likely to graduate and receive a Bachelor's degree within 6 years and achieve second year status by the following fall semester.
- Rather than evaluating the EEA program directly, we are instead examining differences in outcomes for students who take 15 or more credits a semester compared to students who take 12-14 credits.
- Specifically, this study asks: ***How can data science methods be used to conduct research with MLDS data that helps to inform policy decisions in the State of Maryland?***

Research Questions

- Can we use predicted probabilities for the post-secondary enrollment categories (12-14 vs 15+ credits) as propensity scores in the analyses to evaluate the effect of taking 15+ credits on other postsecondary outcomes (e.g. achieving second year status and graduation within 6 years)?

- Can we use machine learning algorithms to create "equivalent" samples of students who only differ in whether they took 12-14 credits vs 15+ credit?

  → *These questions align with the MLDS Center research agenda and ask "How can data science methods be used to conduct research with MLDS data that helps to inform policy decisions in the State of Maryland?"*

Traditional Uses of ML

- Traditionally, (supervised) ML is used for prediction.
  - → Predict new outcomes for new, unseen data using a model that has been trained on seen data.
  - → In classification tasks, trained algorithms predict the probability of a given outcome which can then be used to make decisions.

- In education, ML has been used to:
  - → Predict student retention
  - → Grade student exams

Traditional Uses of ML

- Traditionally, (supervised) ML is used for prediction.
  - → Predict new outcomes for new, unseen data using a model that has been trained on seen data.
  - → In classification tasks, trained algorithms **predict the probability of a given outcome** which can then be used to make decisions.

- In education, ML has been used to:
  - → Predict student retention
  - → Grade student exams

MLDS Center
Maryland Longitudinal Data System

Using ML in combination with Causal Inference

- Propensity scores are conditional probabilities where a given observation is predicted to belong to the treatment group (Z = 1) given the adjustment variables **X**

$$e(\boldsymbol{X}) = \frac{P(Z = 1, \boldsymbol{X})}{P(Z = 0, \boldsymbol{X}) + P(Z = 1, \boldsymbol{X})} = P(Z = 1 | \boldsymbol{X})$$

Using ML in combination with Causal Inference

- Propensity scores are **conditional probabilities** where a given observation is predicted to belong to the treatment group (Z = 1) given the adjustment variables **X**

$$e(\boldsymbol{X}) = \frac{P(Z = 1, \boldsymbol{X})}{P(Z = 0, \boldsymbol{X}) + P(Z = 1, \boldsymbol{X})} = P(Z = 1 | \boldsymbol{X})$$

Using ML in combination with Causal Inference

- Since ML classification algorithms can generate predicted outcome probabilities given the predictors used in the model, it is reasonable to use these **probabilities as propensity scores** to estimate the treatment effect

$$e(\boldsymbol{X}) = \frac{P(Z = 1, \boldsymbol{X})}{P(Z = 0, \boldsymbol{X}) + P(Z = 1, \boldsymbol{X})} = P(Z = 1 | \boldsymbol{X})$$

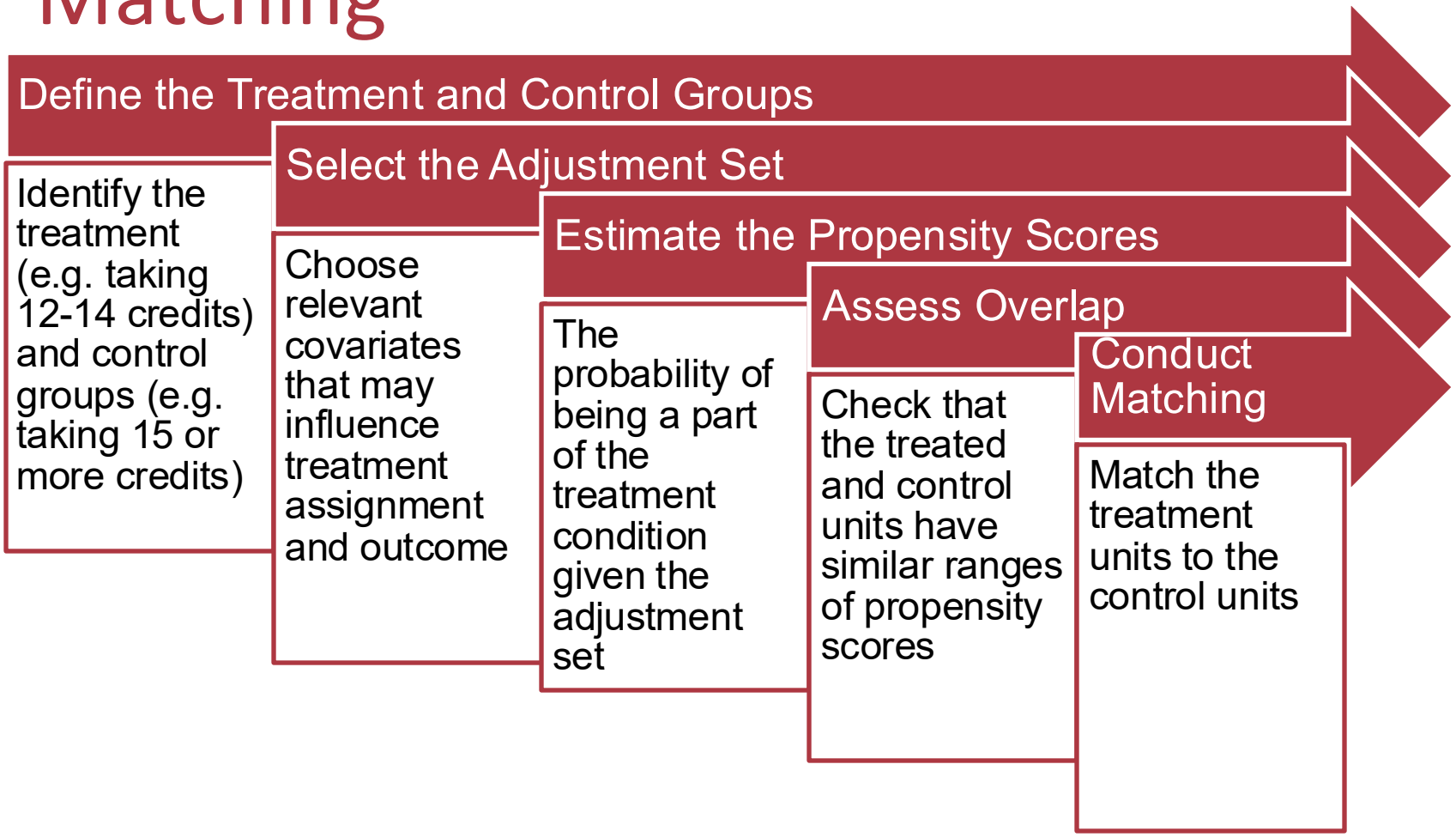
Using ML in combination with Causal Inference

- Since ML classification algorithms can generate predicted outcome probabilities given the predictors used in the model, it is reasonable to use these **probabilities as propensity scores** to estimate the treatment effect

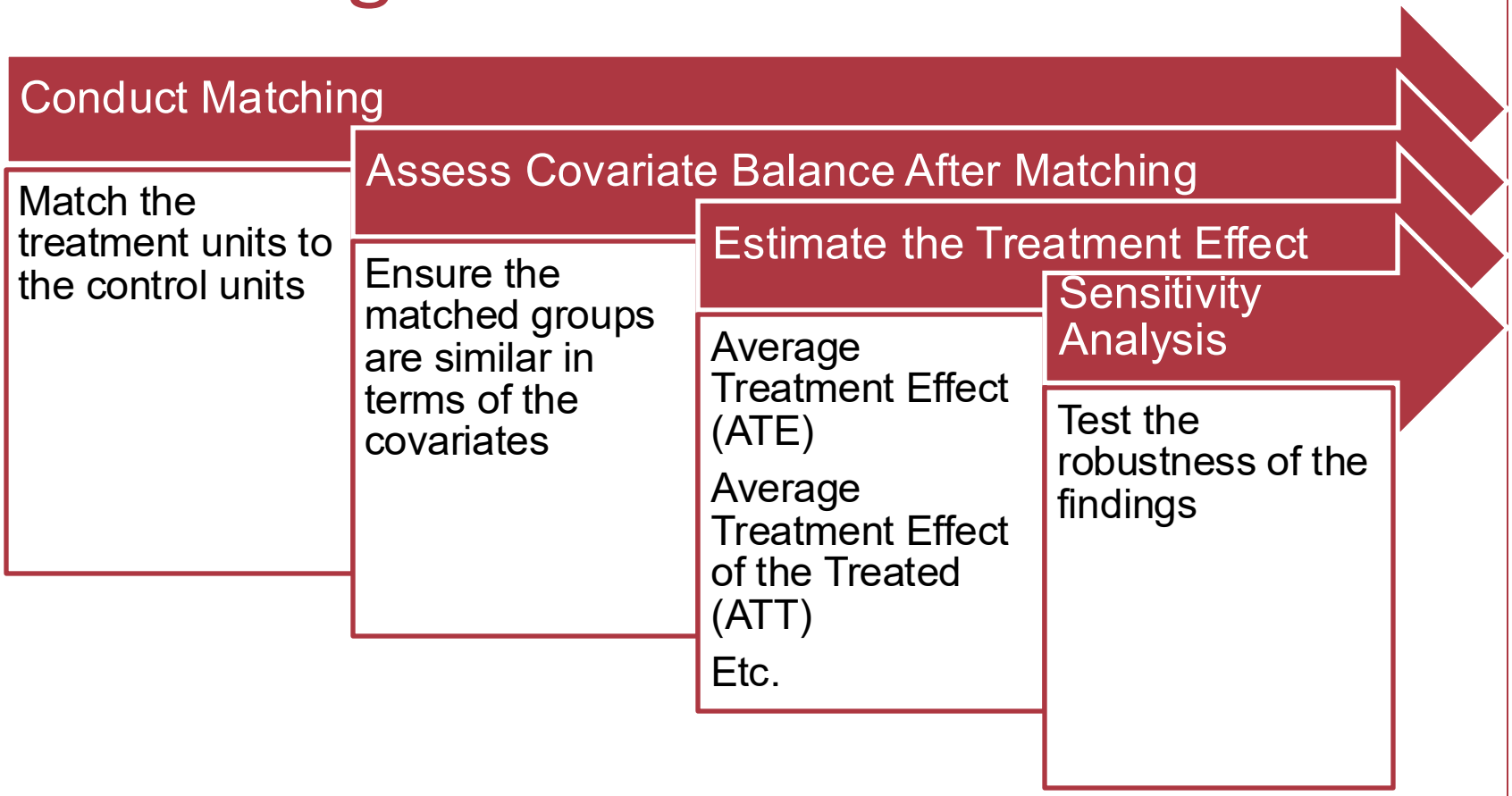
$$e(\boldsymbol{X}) = P(\hat{Y} = 1|\boldsymbol{X})$$

- Where **X** are the predictors used in the ML algorithm and $\hat{Y}$ is the predicted outcome.

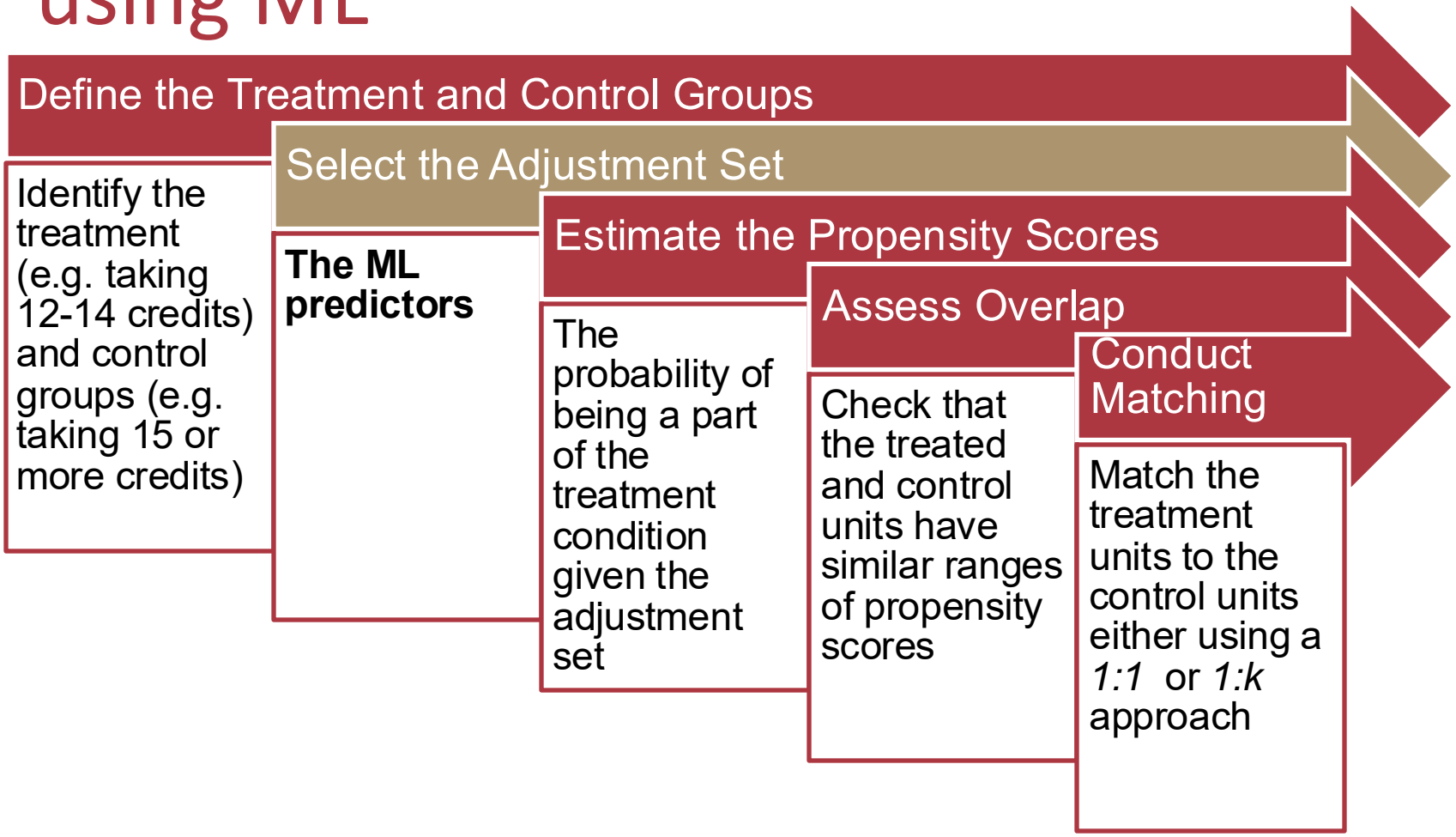
How to Conduct PS Matching



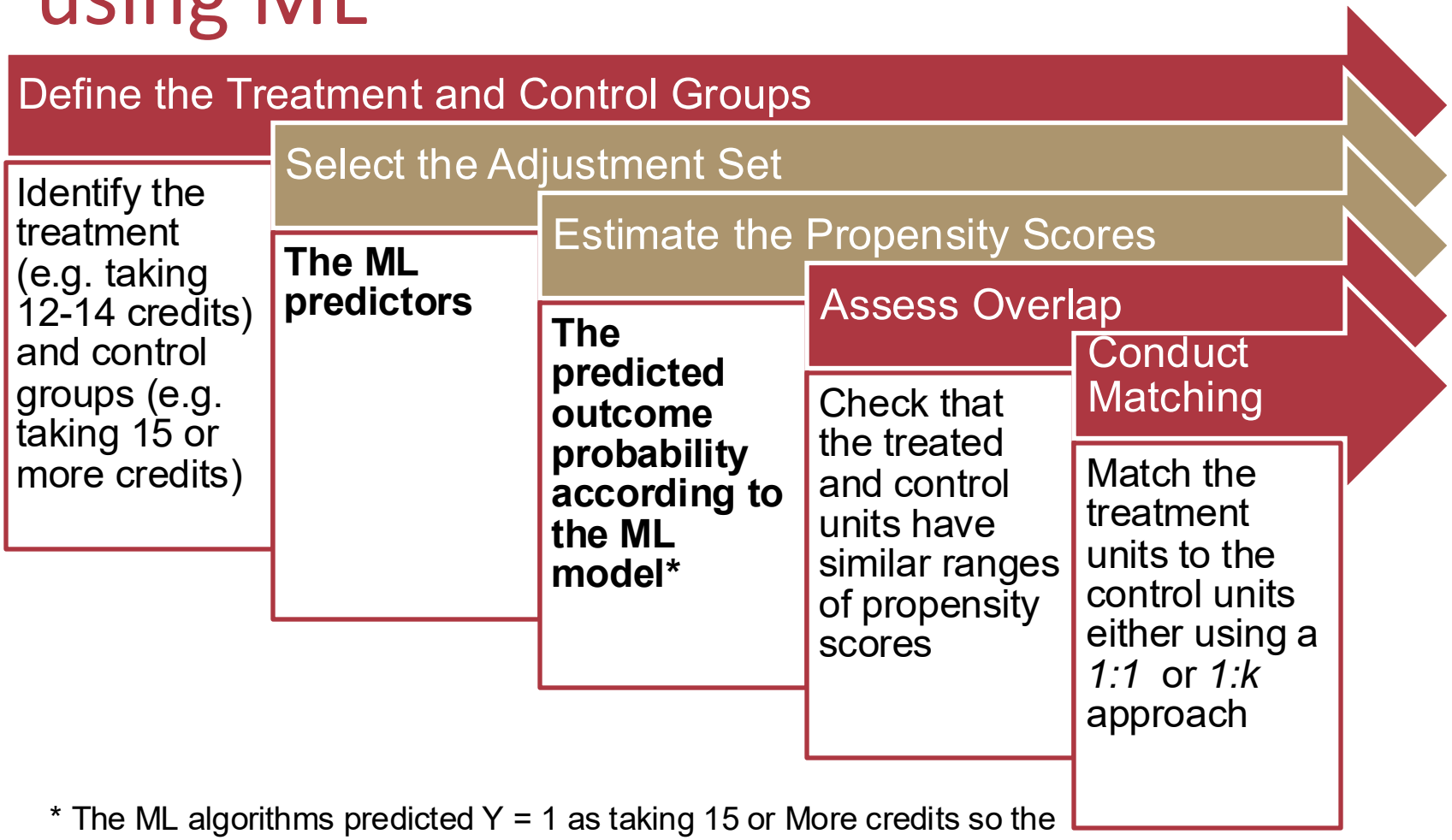
How to Conduct PS Matching



Conducting PS Matching using ML

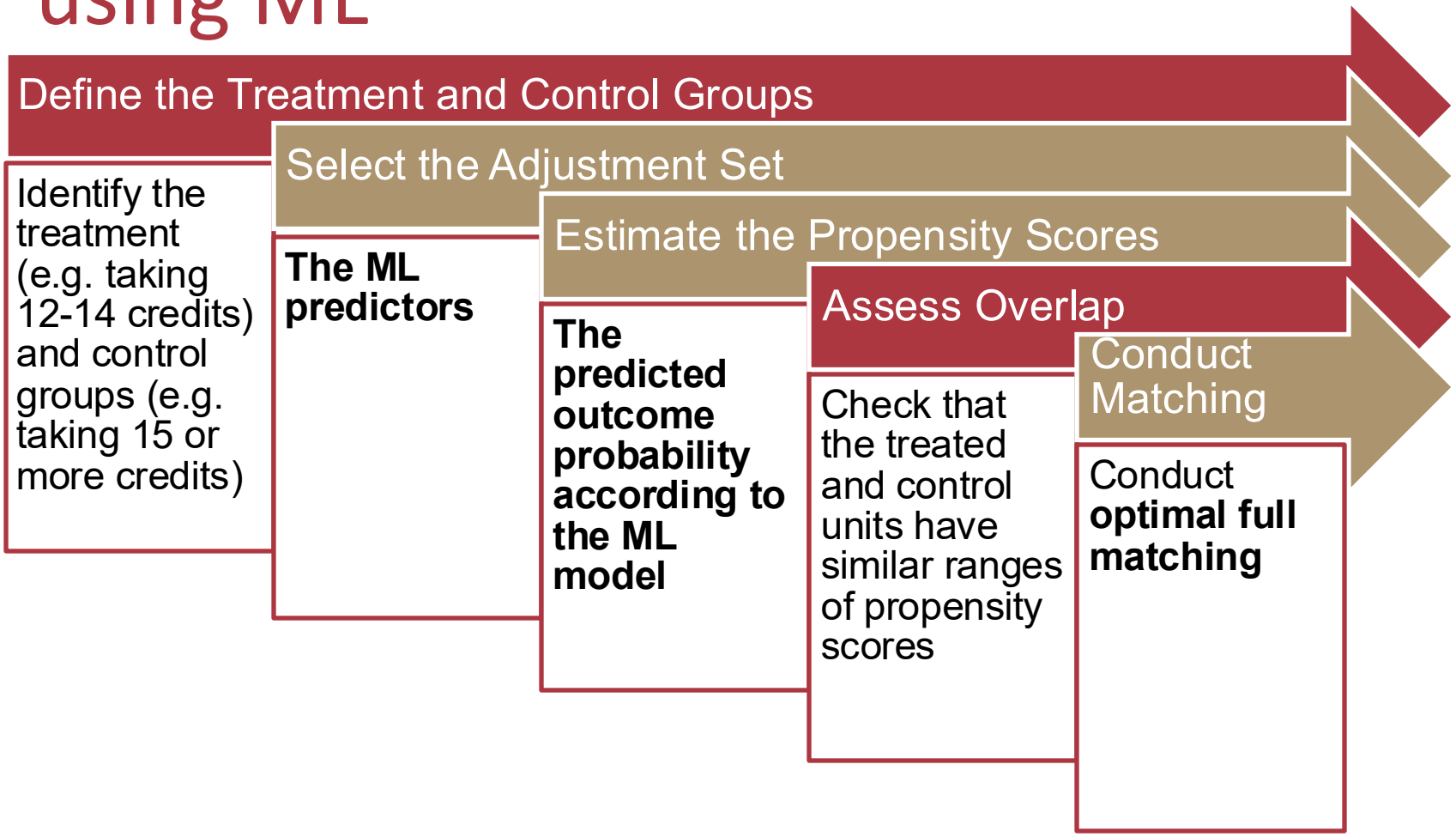


Conducting PS Matching using ML

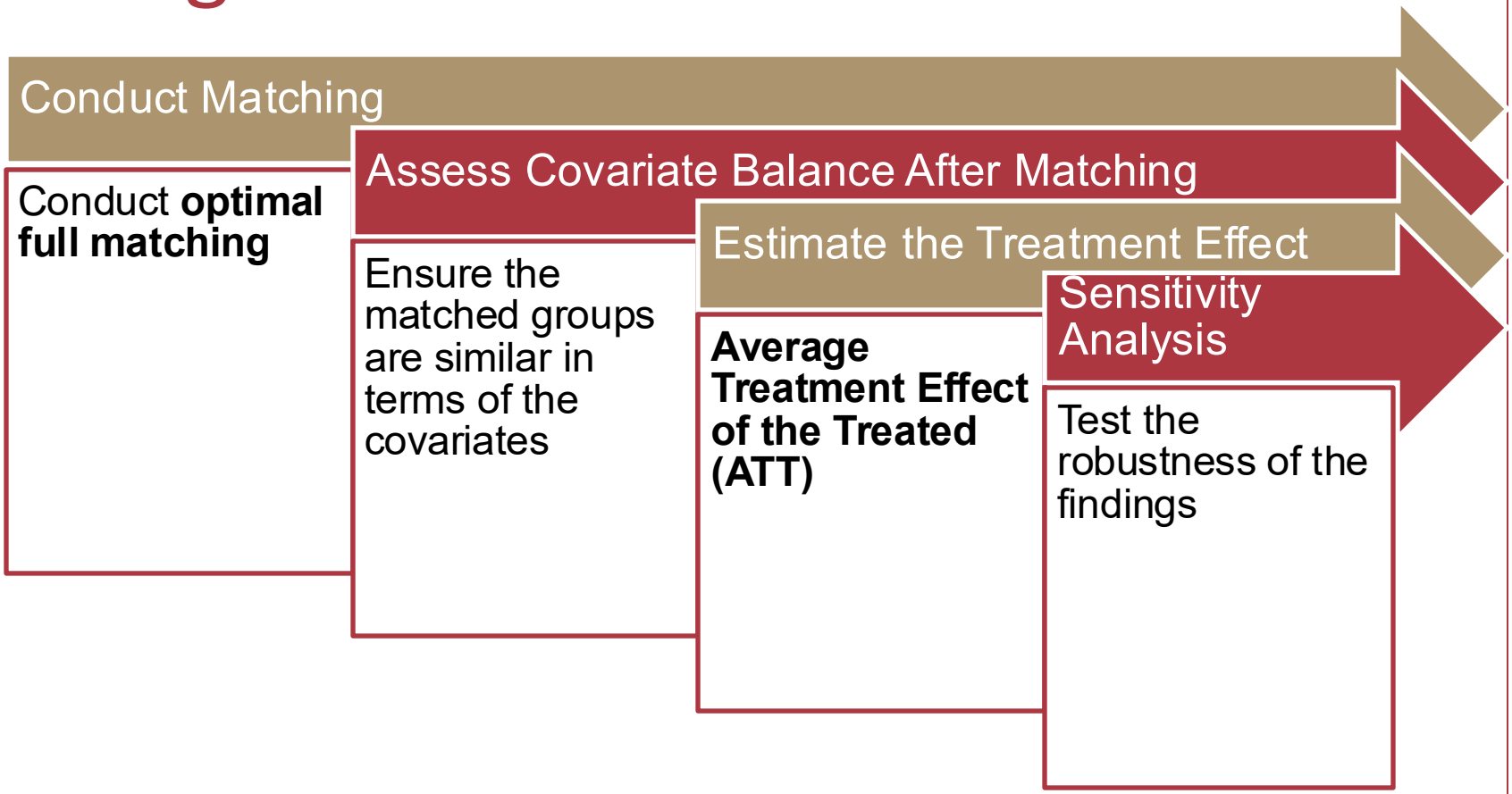


* The ML algorithms predicted Y = 1 as taking 15 or More credits so the PS are actually 1- Predicted Outcome Probability since the treatment group is defined as taking 12-14 credits

Conducting PS Matching using ML

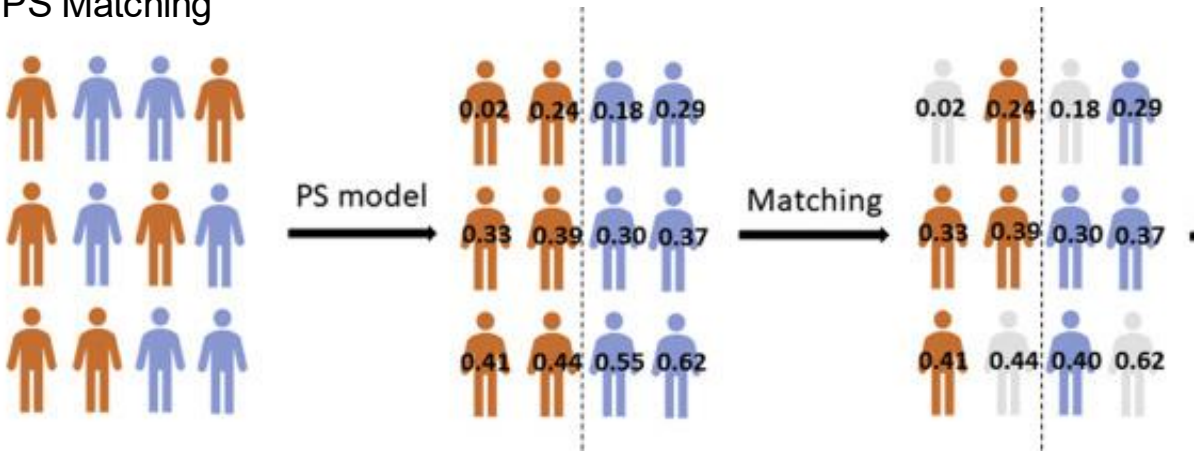


Conducting PS Matching using ML

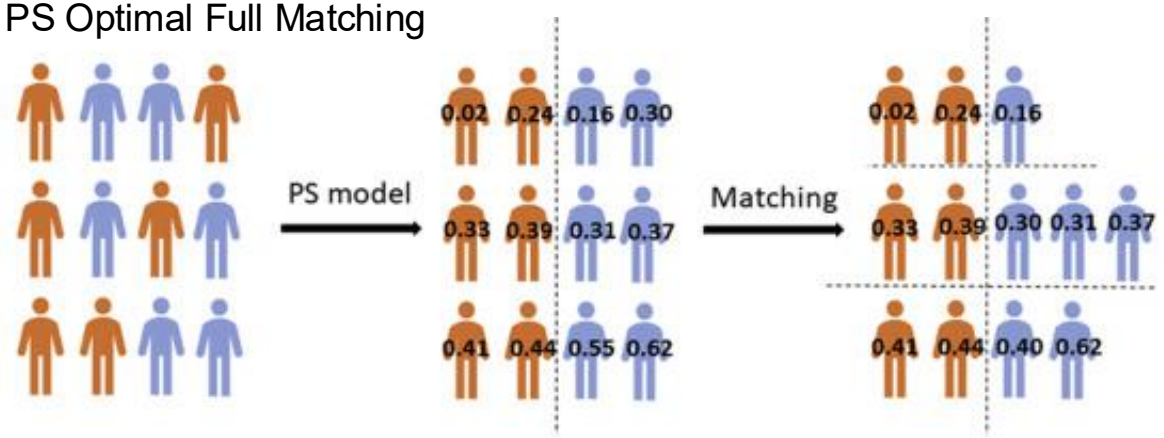


Conducting Propensity Score Optimal Full Matching

PS Matching



PS Optimal Full Matching



Data

MLDS Center
Maryland Longitudinal Data System

Desired Cohort

- First time, full-time students enrolled in a Bachelor's degree program Fall 2015

Inclusion Criteria

- Full-time students enrolled at a 4-year university in Maryland
- Classified as either a freshman or sophomore when starting university
- Graduated from a Maryland high school within 3 years of starting university

Exclusion Criteria

- Needs to be at least 2 students from each high school
- Not a current ELL/ESL student when finishing high school
  - Reason: PS matching requires covariate balance

Final Cohort

- N = 8,999 students
  - 198 high schools
  - Student Demographics: Asian: 1,045; Black: 2,618; Two or more races: 600; White: 4,712

Split into Training and Test Data

- 50/50 Split
- At least one observation per high school in the test and training data
- The test set cohort was used for PS analysis

Data: Outcome Variables

MLDS Center
Maryland Longitudinal Data System

Graduate within 6 Years

Whether a student is able to graduate with a Bachelor's degree within 6 years of starting at a Maryland Higher Education Institution

Second Year Status

Whether a student is able to be classified as a second year student (having at least 30 credits) by the following fall semester.

Data: Covariates

MLDS Center
Maryland Longitudinal Data System

Student Level

- Sociodemographic: Race, Gender, Ethnicity, FARMS eligibility, Special Education, Homelessness, Proportion of Days of Attendance.
- Student Achievement: Number of AP & IB exams taken, Weighted HS GPA, National Standardized Test Percentile, State Standardized Test Score
- University: Whether a student's program of study is classified as STEM, Grade-level of the student in Fall 2015
- Labor: Reported Earnings for Quarters 3 and 4 for a given student & Industry that a student worked the most in for quarters 3 and 4 of 2015.

High School Level

- School Type
- School Name
- School District
- NCES Classification
- Race/Gender/Special Education/ELL/FARMS Distribution
- Magnet School Classification
- Total enrollment

Results in a total of 420 predictors

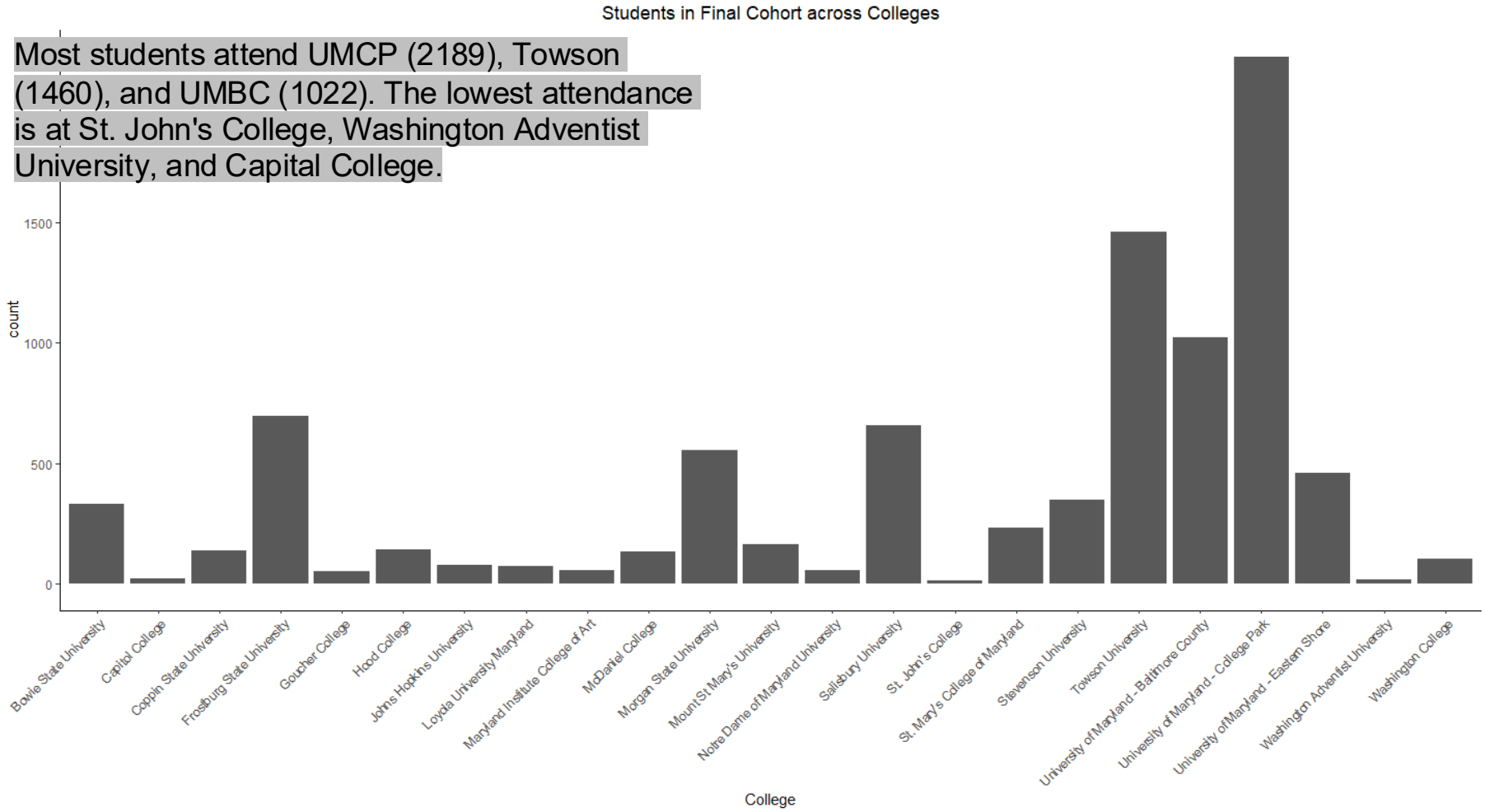
University Level

- Name of University
- Racial/Ethnic Distribution
- Private/Public Status

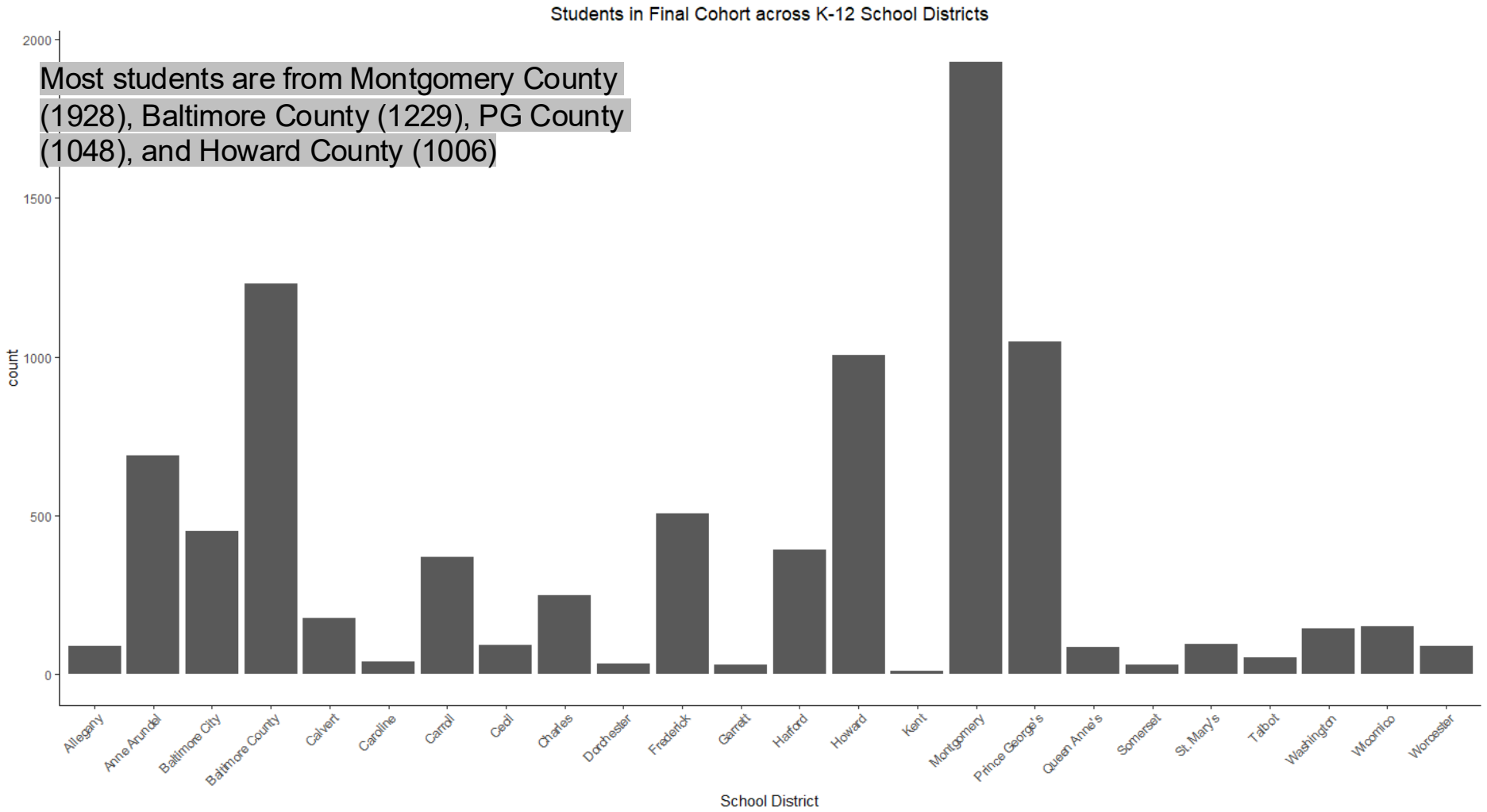
Who is in the final data?

MLDS Center
Maryland Longitudinal Data System

Most students attend UMCP (2189), Towson (1460), and UMBC (1022). The lowest attendance is at St. John's College, Washington Adventist University, and Capital College.

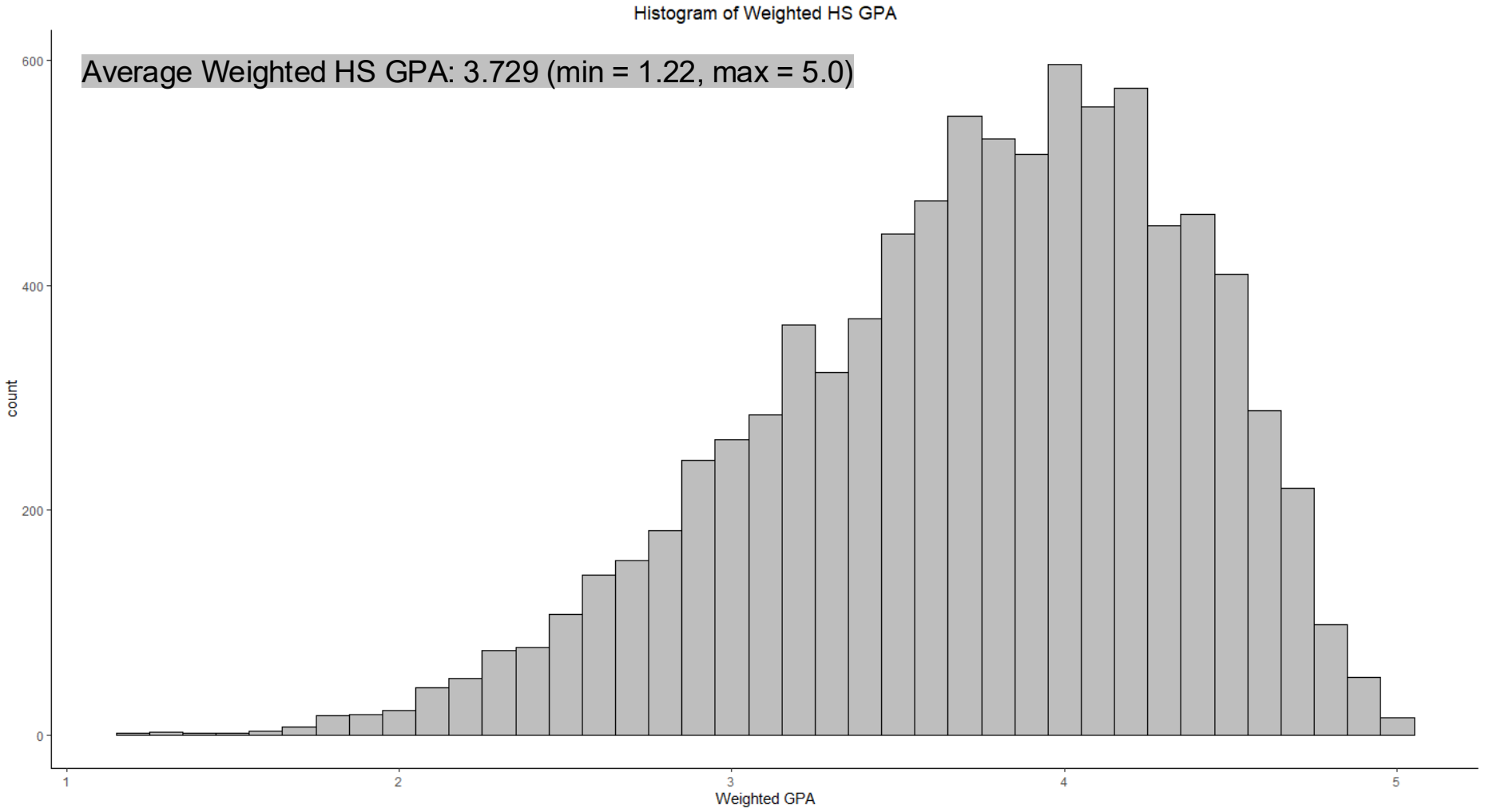


Who is in the final data?

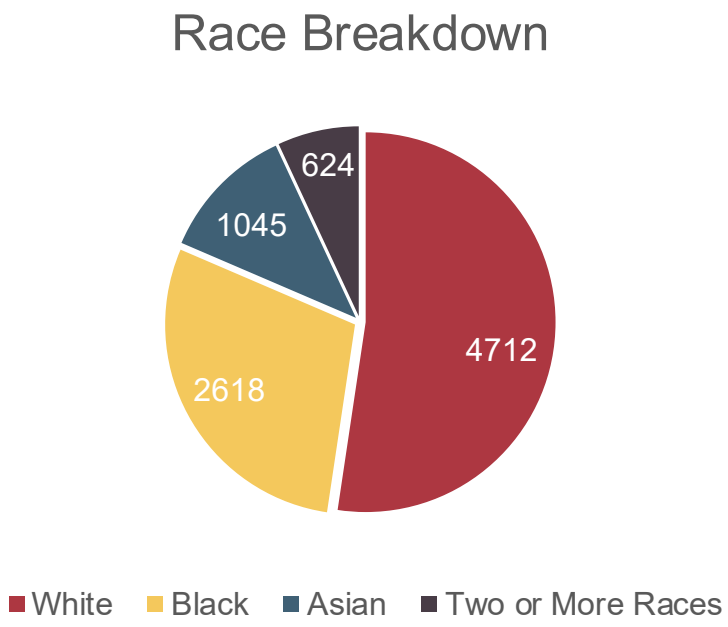


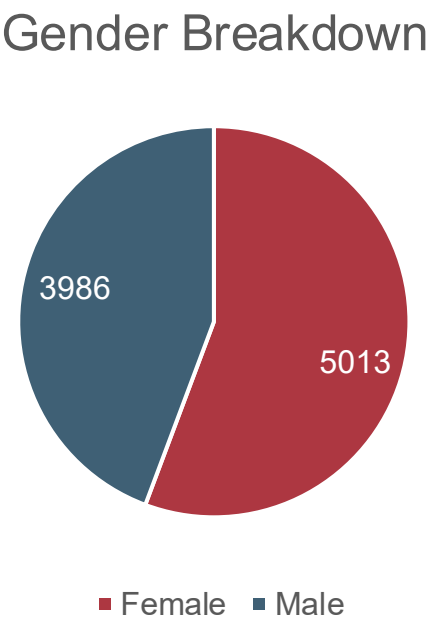
Who is in the final data?

MLDS Center
Maryland Longitudinal Data System



Who is in the final data?





There are nearly 18x the number of non-Hispanic students as there are Hispanic students

Evaluating ML Algorithms

- When there are dichotomous outcomes like in this case (taking 15 or more credits vs taking 12-14 credits), we can use the classification accuracy, sensitivity/recall, specificity, precision, and F1 score to evaluate algorithmic performance.

		Predicted Class: Positive	**Predicted Class: Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

ML Findings

Algorithm	Classification Accuracy	Sensitivity	Specificity	Precision	F1
Logistic Regression	0.7038	0.3707	0.8597	0.5531	0.4438
LASSO Regression	0.7100	0.3249	0.8904	0.5813	0.4168
Ridge Regression	0.6948	0.3101	0.8818	0.5513	0.3969
Elastic Net	0.7100	0.3249	0.8904	0.5813	0.4168
Classification Tree	0.7069	0.3975	0.8518	0.5567	0.4638
Random Forest	0.7080	0.3827	0.8600	0.5615	0.4555
XGBoost	0.7076	0.2171	0.9373	0.6185	0.3213

ML Findin

The proportion of correctly predicted instances.

Algorithm	Classification Accuracy	Sensitivity	Specificity	Precision	F1
Logistic Regression	0.7038	0.3707	0.8597	0.5531	0.4438
LASSO Regression	0.7100	0.3249	0.8904	0.5813	0.4168
Ridge Regression	0.6948	0.3101	0.8818	0.5513	0.3969
Elastic Net	0.7100	0.3249	0.8904	0.5813	0.4168
Classification Tree	0.7069	0.3975	0.8518	0.5567	0.4638
Random Forest	0.7080	0.3827	0.8600	0.5615	0.4555
XGBoost	0.7076	0.2171	0.9373	0.6185	0.3213

ML Findings

MLDS Center
Longitudinal Data System

The proportion of actual positive cases that are correctly identified.

Algorithm	Classification Accuracy	Sensitivity	Specificity	Precision	F1
Logistic Regression	0.7038	0.3707	0.8597	0.5531	0.4438
LASSO Regression	0.7100	0.3249	0.8904	0.5813	0.4168
Ridge Regression	0.6948	0.3101	0.8818	0.5513	0.3969
Elastic Net	0.7100	0.3249	0.8904	0.5813	0.4168
Classification Tree	0.7069	0.3975	0.8518	0.5567	0.4638
Random Forest	0.7080	0.3827	0.8600	0.5615	0.4555
XGBoost	0.7076	0.2171	0.9373	0.6185	0.3213

ML Findings

The proportion of actual negative cases that are correctly identified.

Algorithm	Classification Accuracy	Sensitivity	Specificity	Precision	F1
Logistic Regression	0.7038	0.3707	0.8597	0.5531	0.4438
LASSO Regression	0.7100	0.3249	0.8904	0.5813	0.4168
Ridge Regression	0.6948	0.3101	0.8818	0.5513	0.3969
Elastic Net	0.7100	0.3249	0.8904	0.5813	0.4168
Classification Tree	0.7069	0.3975	0.8518	0.5567	0.4638
Random Forest	0.7080	0.3827	0.8600	0.5615	0.4555
XGBoost	0.7076	0.2171	0.9373	0.6185	0.3213

ML Findings

The proportion of actual positive instances among all instances that are predicted as positive.

Algorithm	Classification Accuracy	Sensitivity	Specificity	Precision	F1
Logistic Regression	0.7038	0.3707	0.8597	0.5531	0.4438
LASSO Regression	0.7100	0.3249	0.8904	0.5813	0.4168
Ridge Regression	0.6948	0.3101	0.8818	0.5513	0.3969
Elastic Net	0.7100	0.3249	0.8904	0.5813	0.4168
Classification Tree	0.7069	0.3975	0.8518	0.5567	0.4638
Random Forest	0.7080	0.3827	0.8600	0.5615	0.4555
XGBoost	0.7076	0.2171	0.9373	0.6185	0.3213

ML Findings

MLDS Center

The harmonic mean of recall and precision.

Algorithm	Classification Accuracy	Sensitivity	Specificity	Precision	F1
Logistic Regression	0.7038	0.3707	0.8597	0.5531	0.4438
LASSO Regression	0.7100	0.3249	0.8904	0.5813	0.4168
Ridge Regression	0.6948	0.3101	0.8818	0.5513	0.3969
Elastic Net	0.7100	0.3249	0.8904	0.5813	0.4168
Classification Tree	0.7069	0.3975	0.8518	0.5567	0.4638
Random Forest	0.7080	0.3827	0.8600	0.5615	0.4555
XGBoost	0.7076	0.2171	0.9373	0.6185	0.3213

ML Findings

MLDS Center
Maryland Longitudinal Data System

All the algorithms had classification accuracy rates ~70%; in ML, classification accuracy rates that are about 70% are on the lower end of acceptable accuracy.

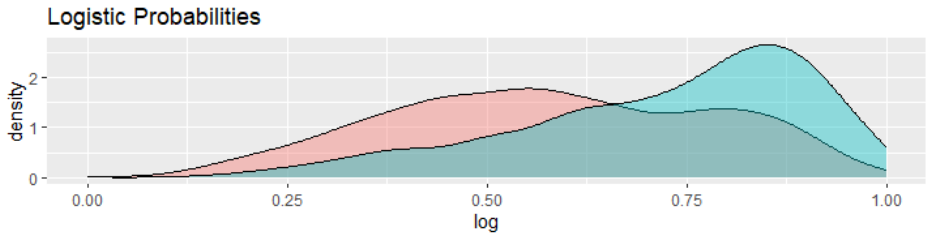
Algorithm	Cla… Acc…				F1
Logistic Regression	0.7038				0.4438
LASSO Regression	0.7100			0.5813	0.4168
Ridge Regression	0.6948	0.3101	0.8818	0.5513	0.3969
Elastic Net	0.7100	0.3249	0.8904	0.5813	0.4168
Classification Tree	0.7069	0.3975	0.8518	0.5567	0.4638
Random Forest	0.7080	0.3827	0.8600	0.5615	0.4555
XGBoost	0.7076	0.2171	0.9373	0.6185	0.3213

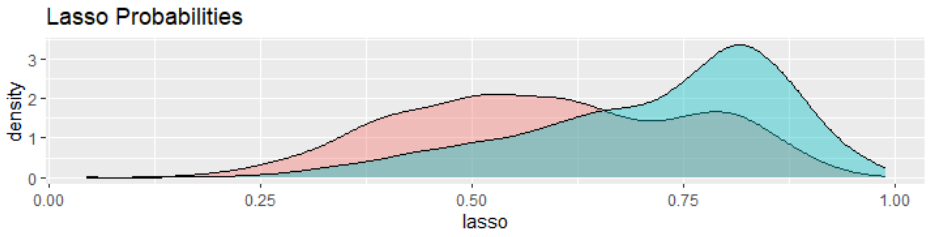
ML Findings

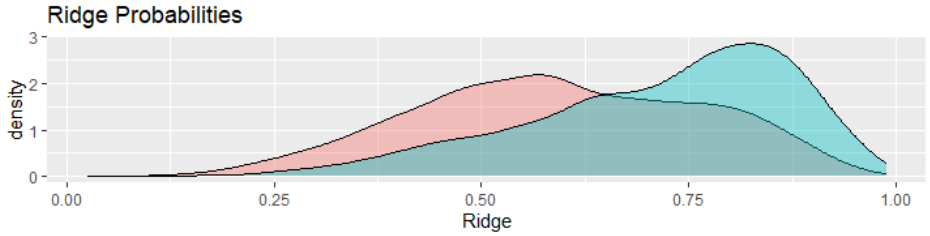
MLDS Center
Maryland Longitudinal Data System

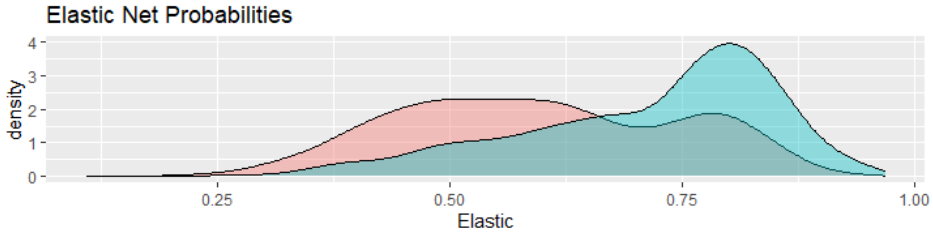
The next step is to check the overlap region before conducting the PS analysis.

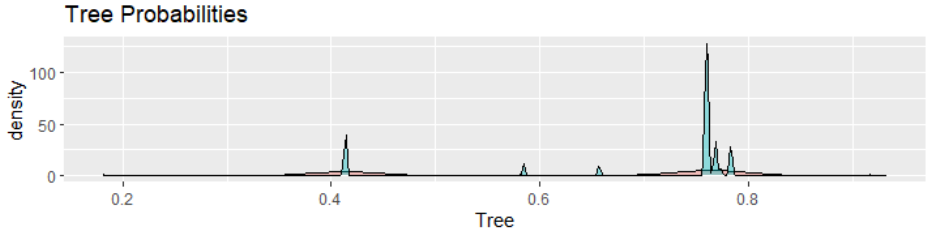
Algorithm	Cl… Ac…			…n	F1
Logistic Regression	0.70…				0.4438
LASSO Regression	0.7100			0.5813	0.4168
Ridge Regression	0.6948	0.3101	0.8818	0.5513	0.3969
Elastic Net	0.7100	0.3249	0.8904	0.5813	0.4168
Classification Tree	0.7069	0.3975	0.8518	0.5567	0.4638
Random Forest	0.7080	0.3827	0.8600	0.5615	0.4555
XGBoost	0.7076	0.2171	0.9373	0.6185	0.3213

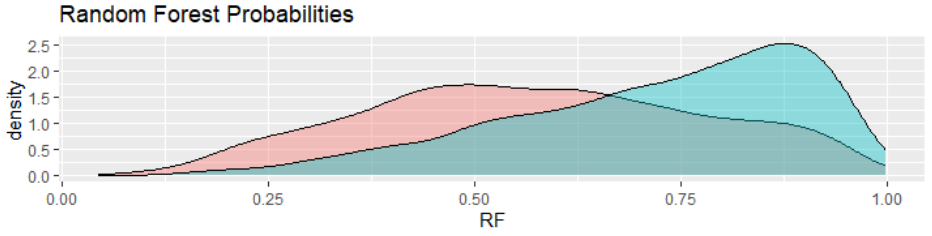


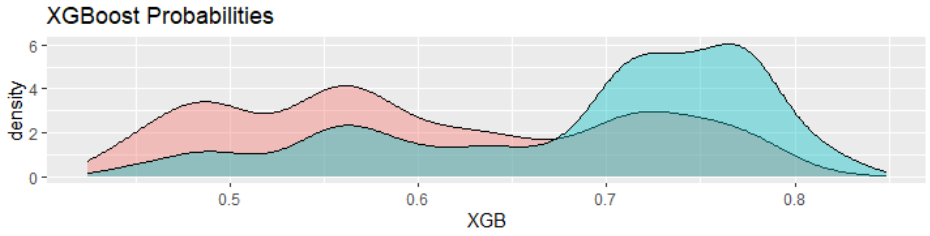


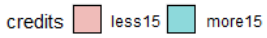








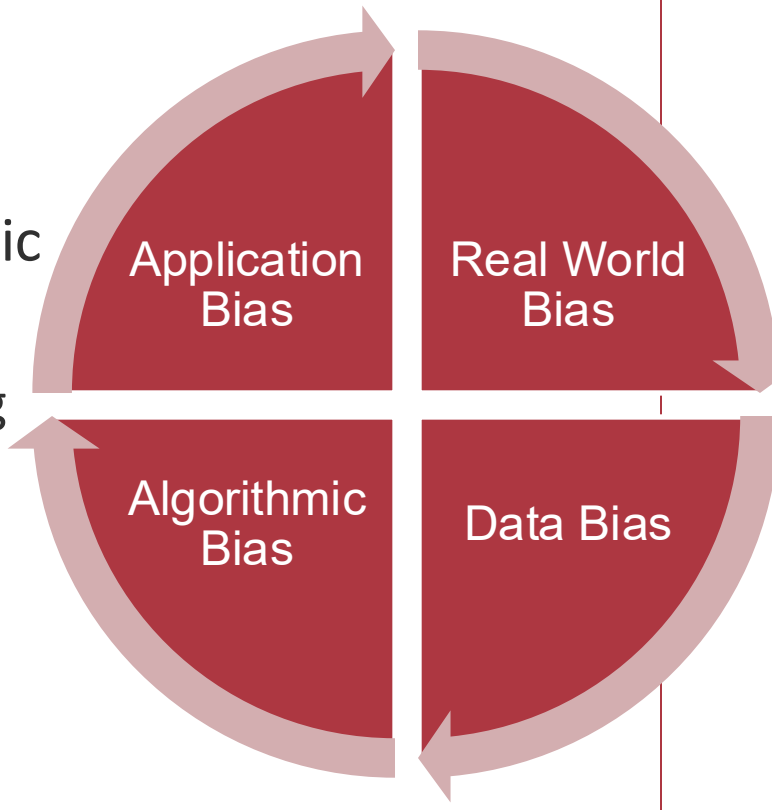




Overlap of the Predicted Probabilities of Taking at least 15 Credits Across Groups

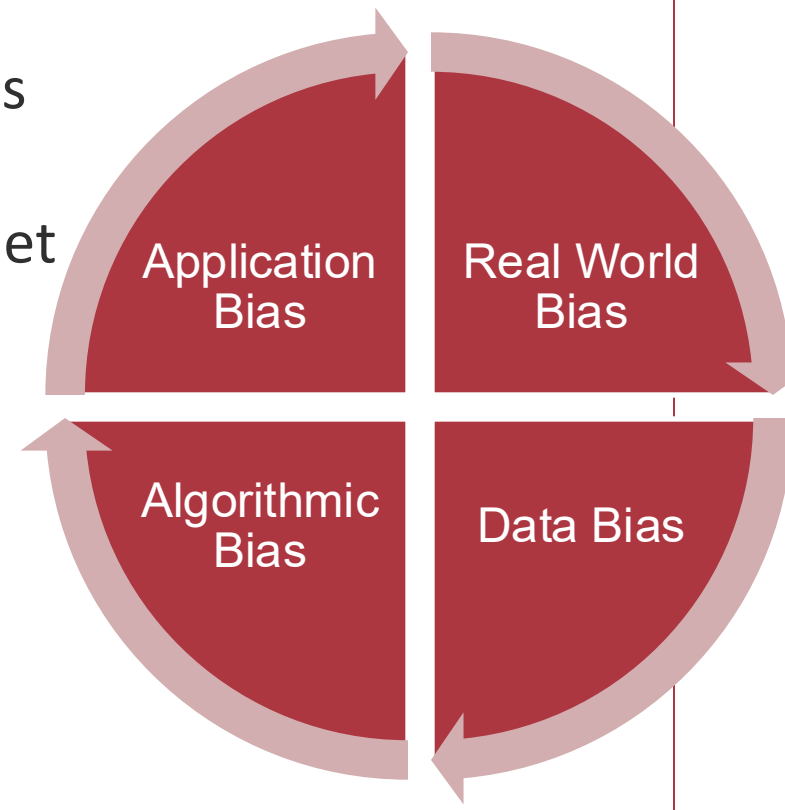
Algorithmic Bias

- **Algorithmic Bias:** The phenomena where machine learning algorithms systematically discriminate against groups or individuals based on demographic characteristics.
  - → Occurs either when the training data used to develop the algorithm contain inherent biases or when the algorithm induces biases during the learning process.

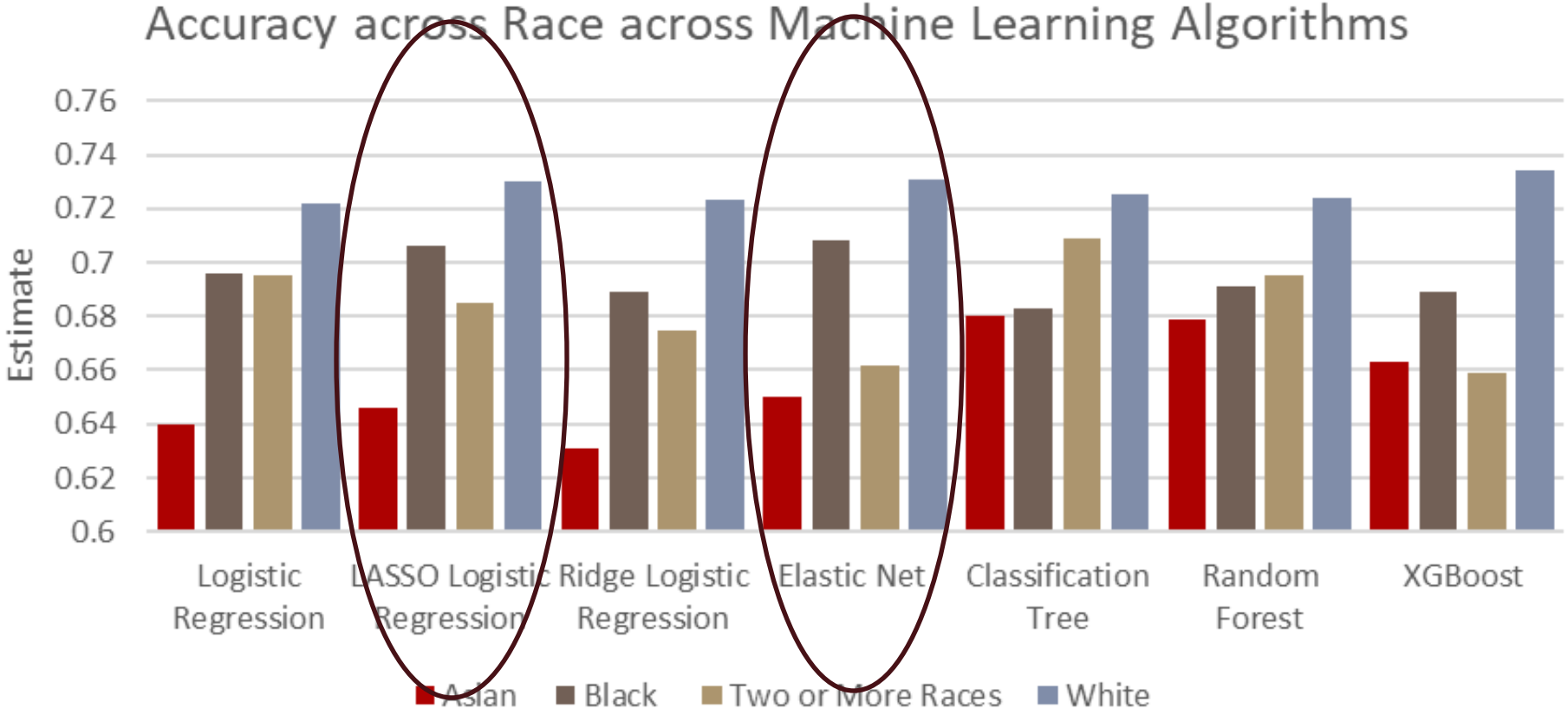


Algorithmic Bias

- For the purposes of this study, we focused on a specific type of algorithmic bias where a model's predictive performance varies across identity groups (Gardner et al, 2019).
  - → Race
  - → Gender

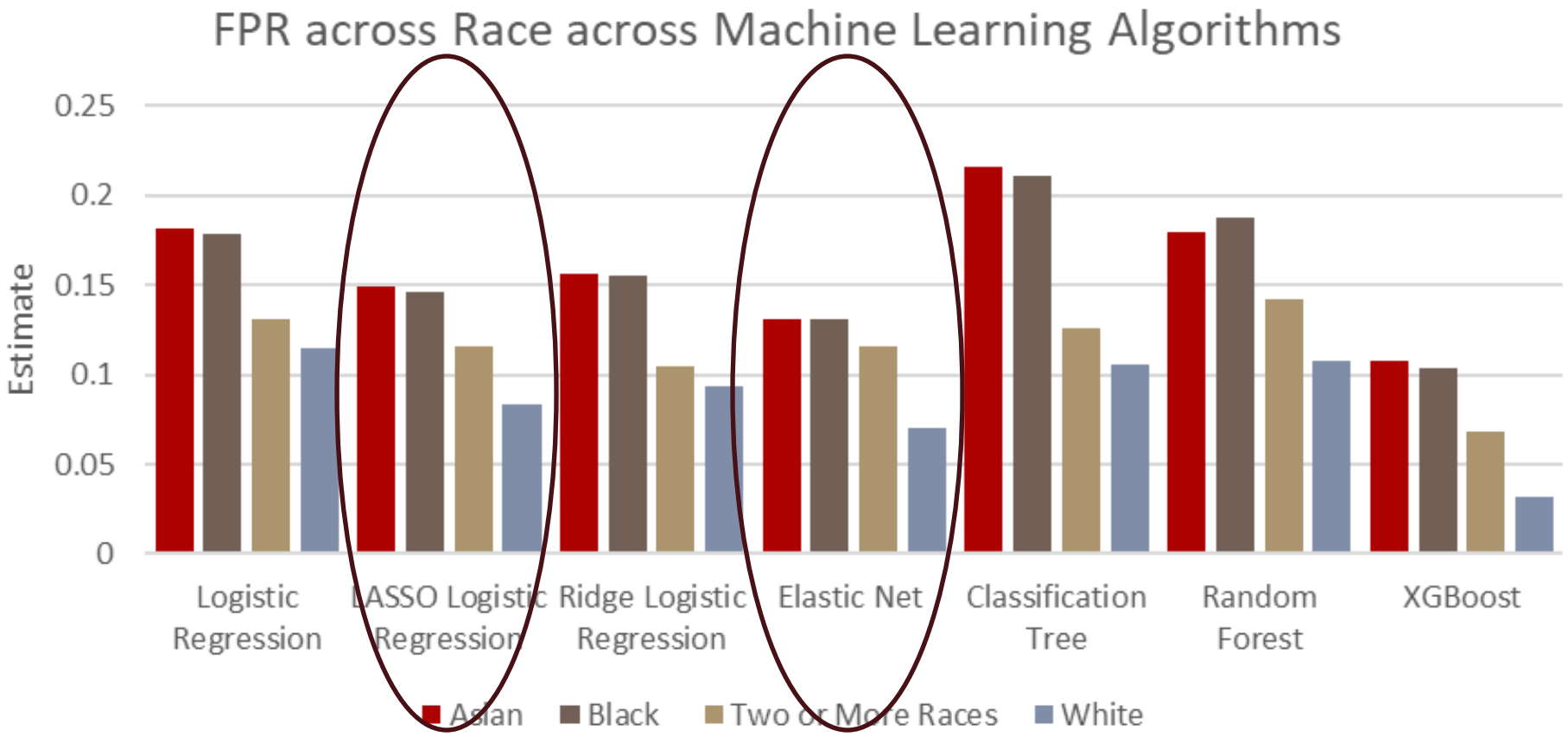


Algorithmic Bias



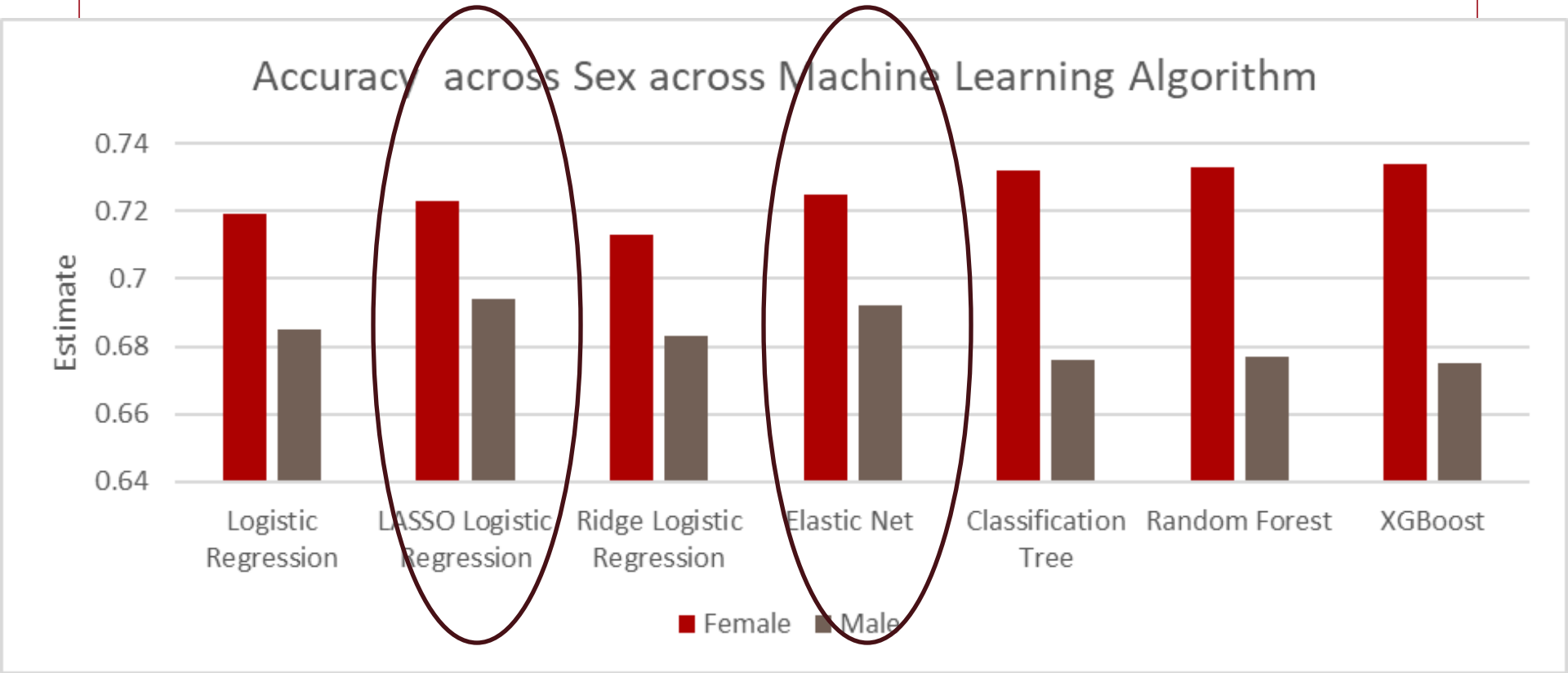
Across all algorithms, White students had the highest classification accuracy across all algorithms; however, this can partially be explained by individuals identifying as White encompassing about 50% of the test data

Algorithmic Bias Concerns



The largest differences across the racial groups is for the Classification Tree, Random Forest, and XGBoost methods.

Algorithmic Bias



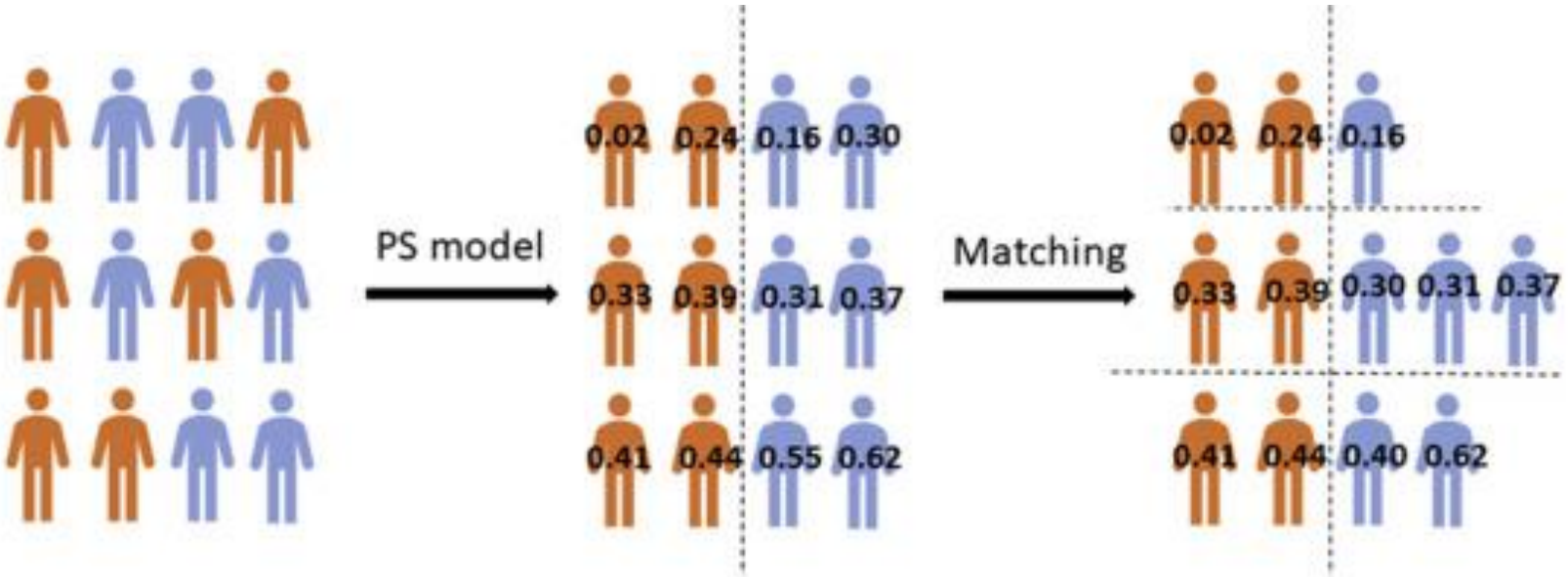
There are higher rates of accuracy for those that identify as female than those that identify as male across all the machine learning algorithms. This could be due in part to the fact that there are more females than males in both higher education institutions in the state of Maryland and in this dataset.

Algorithmic Bias Summary

- The bias analysis shows that while there may be relatively successful models, overall, caution is still needed.

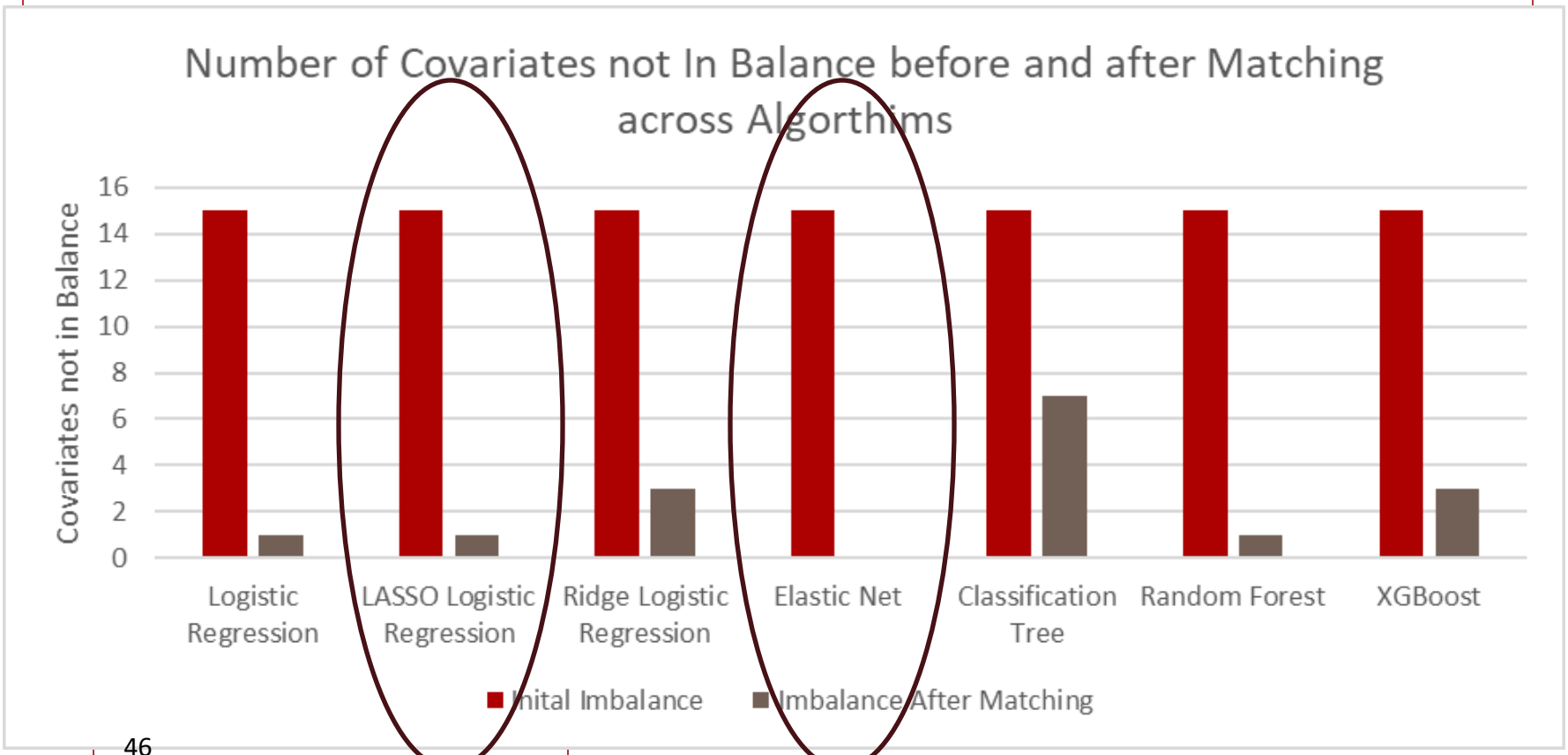
- Thus, additional research is needed before using the raw predicted probabilities and predicted classifications to guide policy decisions in the light of algorithmic fairness concerns.

Recall: Optimal Full Matching



PS Findings

- For the final test data, matches were created and then covariate balance was assessed.
- The number of covariates that remain imbalanced after matching varied across algorithms.

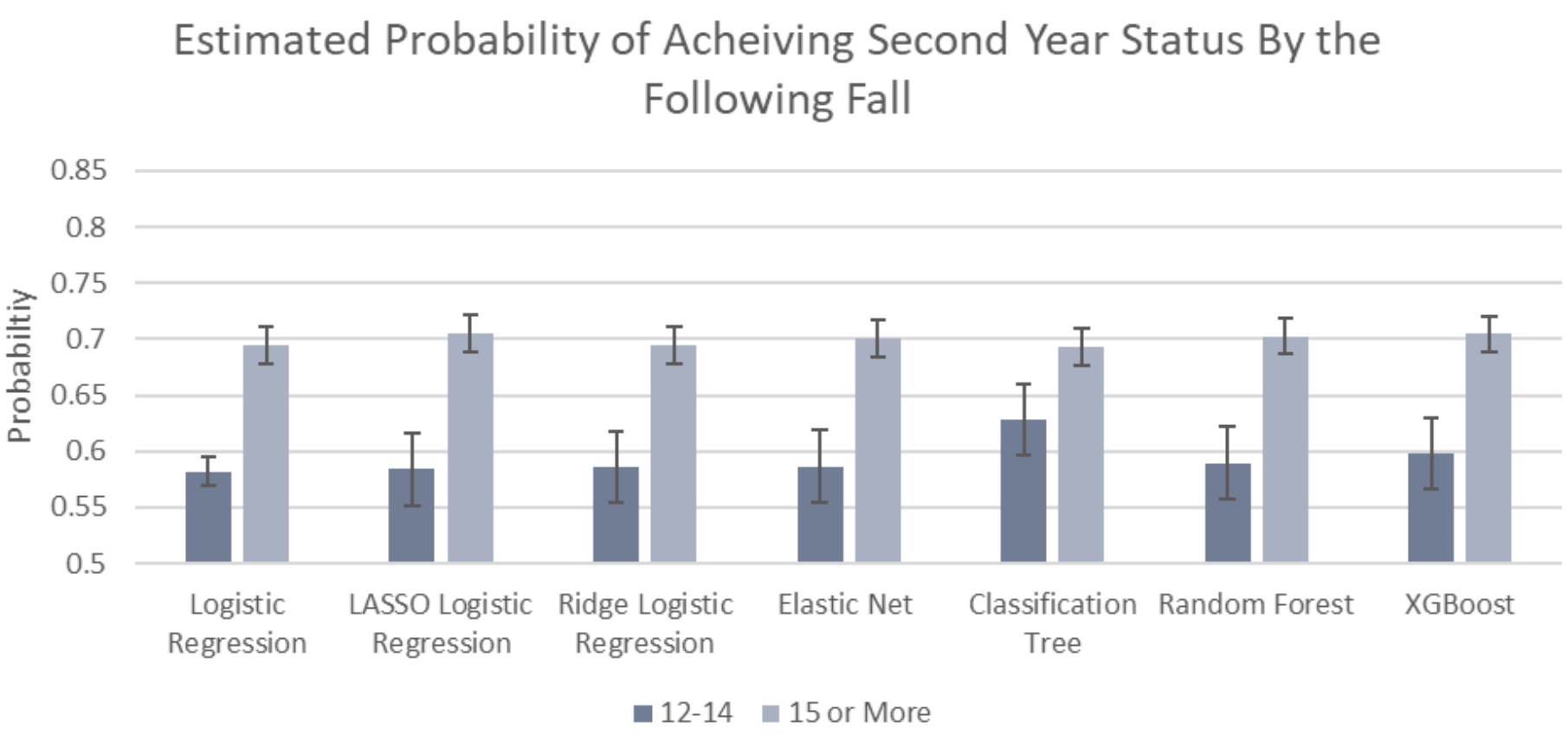


PS Findings: Achieving Second Year Status

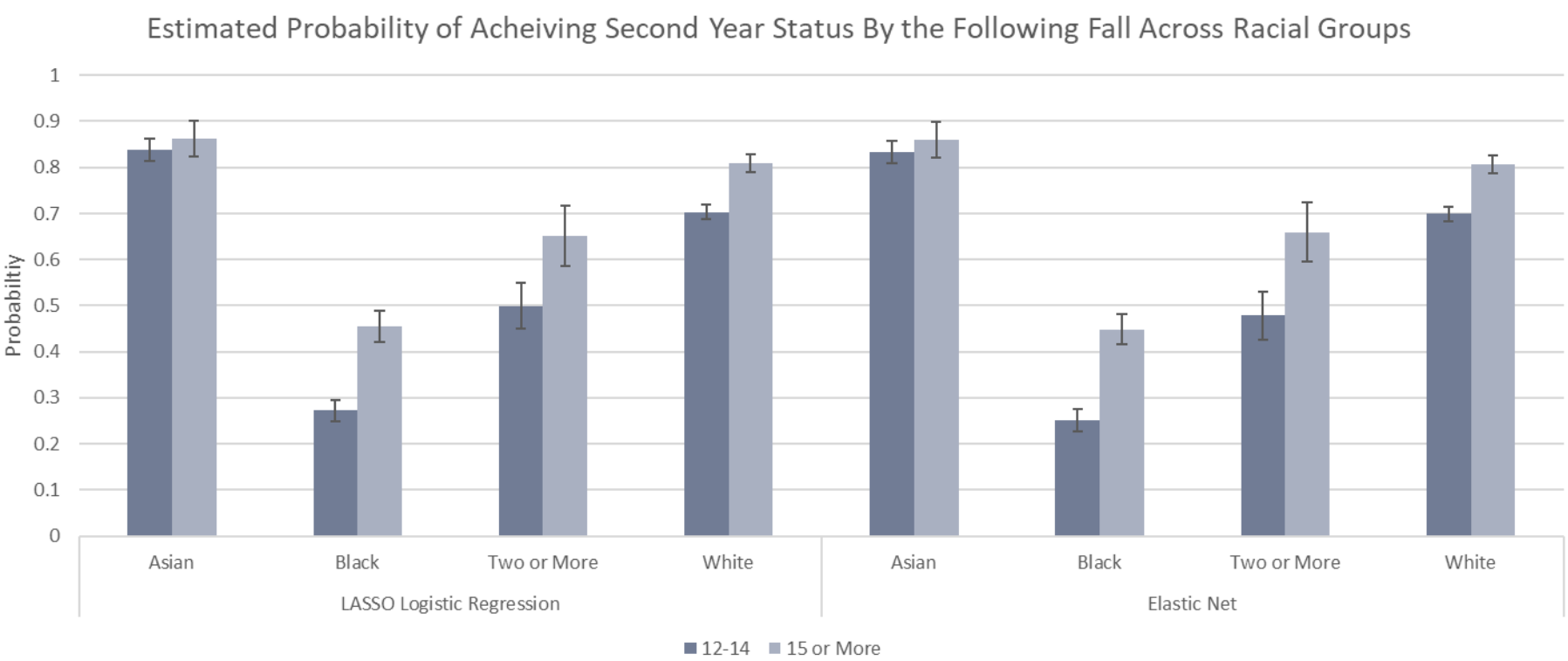
Algorithm	Estimated Treatment Effect (S.E)
Logistic	-0.489 (0.067)*
Lasso	-0.534 (0.067)*
Ridge	-0.474 (0.067)*
Elastic Net	-0.498 (0.067)*
Classification Tree	-0.292 (0.068)*
Random Forest	-0.497 (0.067)*
XGBoost	-0.471 (0.067)*

The estimated treatment effects of taking 12-14 credits compared to 15 or more on achieving second year status.

PS Findings: Achieving Second Year Status



PS Findings: Achieving Second Year Status



For Lasso and Elastic Net, we can observe across the methods that there is no statistically significant difference in probability of achieving second year status for Asian students.

MLDS Center
Maryland Longitudinal Data System

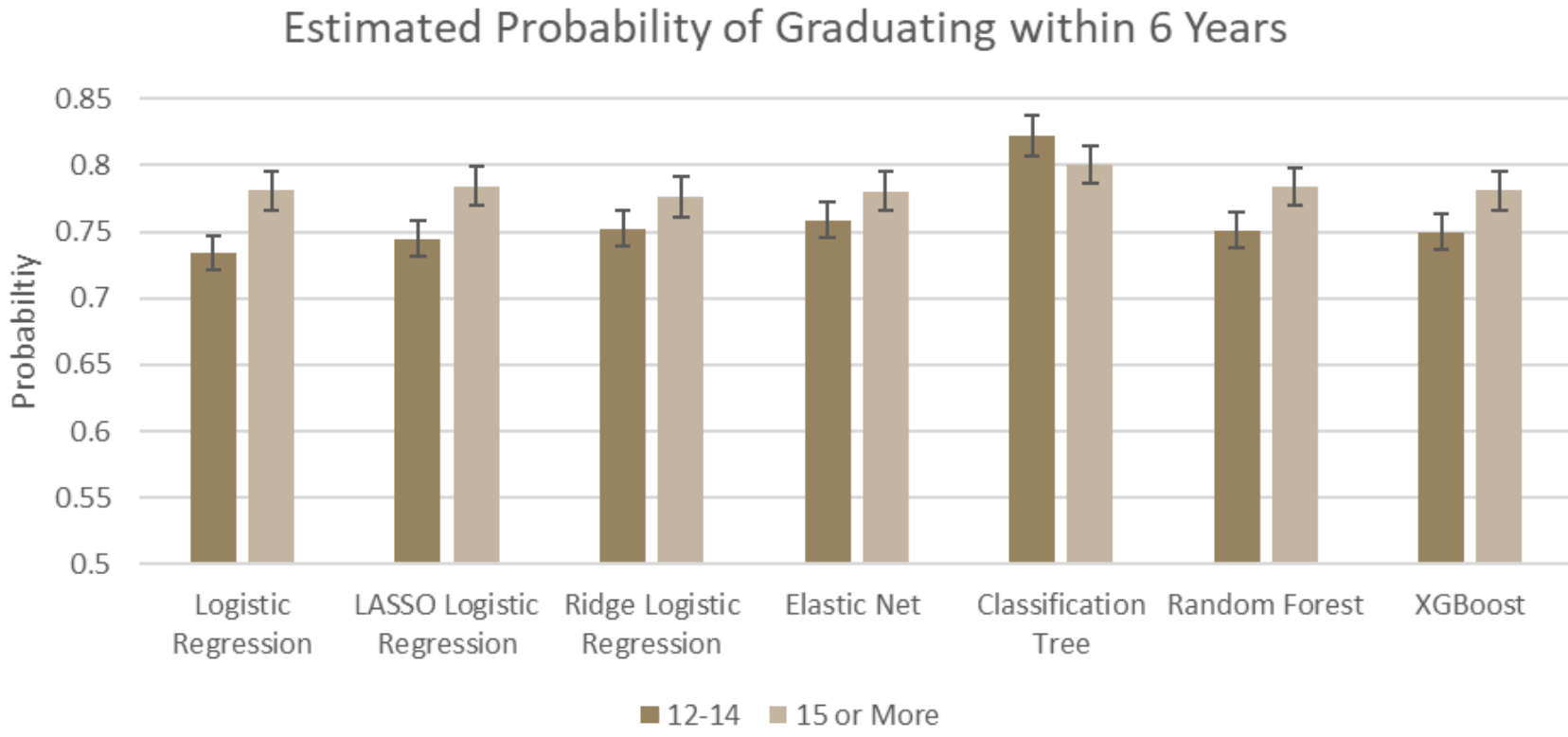
PS Findings: Graduating Within 6 Years

Algorithm	Estimated Treatment Effect (S.E)
Logistic	-0.254 (0.074) *
Lasso	-0.219 (0.075) *
Ridge	-0.132 (0.075)
Elastic Net	-0.122 (0.076)
Classification Tree	0.143 (0.083)
Random Forest	-0.183 (0.076) *
XGBoost	-0.174 (0.075) *

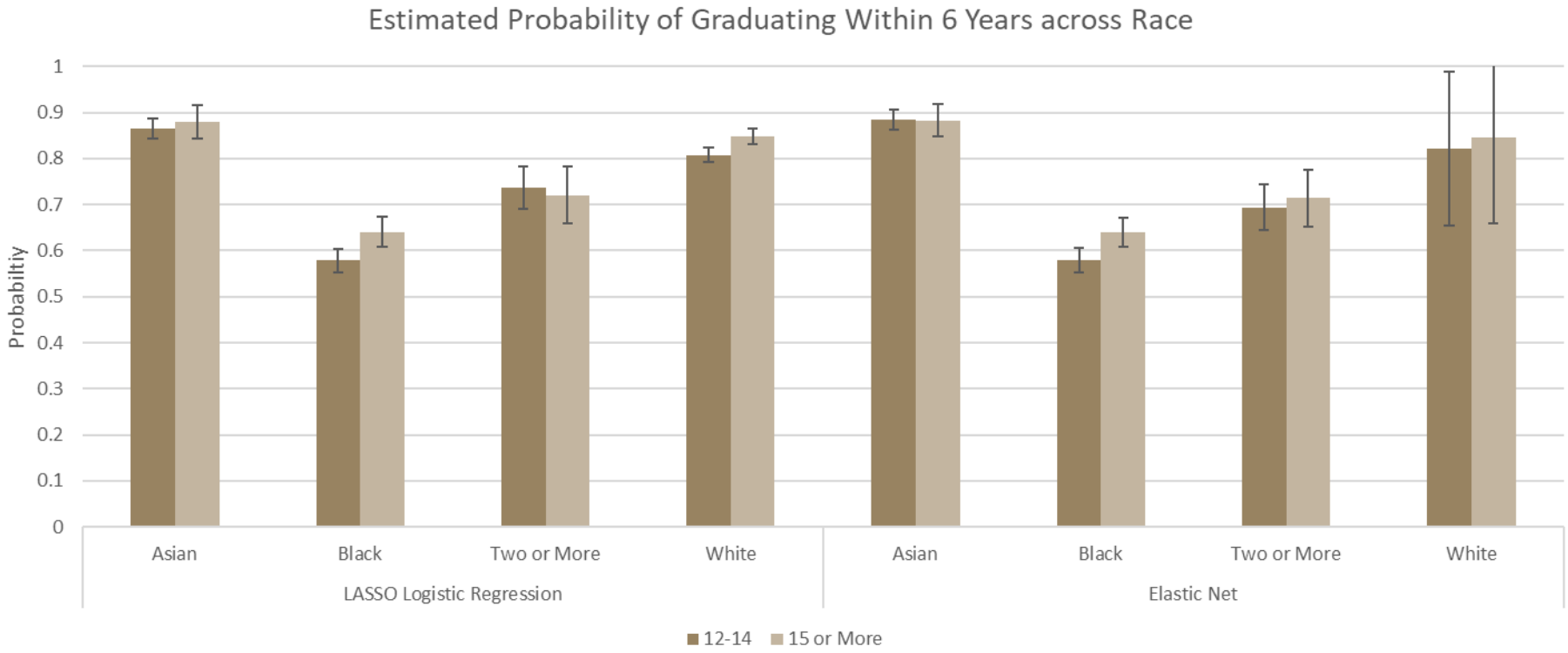
The estimated treatment effects of taking fewer than 15 credits on graduating within 6 years.

PS Findings: Graduating Within 6 Years

- The estimated probability of graduating within 6 years for the two groups of students is below.



PS Findings: Graduating Within 6 Years



We can observe that for Lasso, Black and White students had a statistically significant effect of taking 12-14 credits on graduating within 6 years. This result is echoed for Black students when PS are calculated using Elastic Net.

Discussion

- For researchers interested in prediction, machine learning methods have great potential in the field of education.

  → These methods can be used to generate propensity scores in place of the traditional logistic regression method.

  → It is important to note that additional research is needed before using the raw predicted probabilities and predicted classifications in the light of algorithmic fairness concerns.

  → The variability in the estimated effects of taking 12-14 credits on graduating within 6 years across different machine learning algorithms underscores the importance of model selection and highlights how different algorithms capture distinct aspects of the data.

Discussion

- For researchers interested in prediction, machine learning methods have great potential in the field of education.

  → These methods can be used to generate propensity scores in place of the traditional logistic regression method.

    → However, there has been some recent research that has come out that ML methods when used to estimate causal estimands can result in biased treatment effects (Chernozhukov et al., 2016).

MLDS Center
Maryland Longitudinal Data System

Questions or Feedback?

Contact: Tracy Sweet, Associate Professor, Human Development and Quantitative Methods, University of Maryland College Park

tsweet@umd.edu

Thank you!