MLDS Center
Maryland Longitudinal Data System
Better Data • Informed Choices • Improved Results

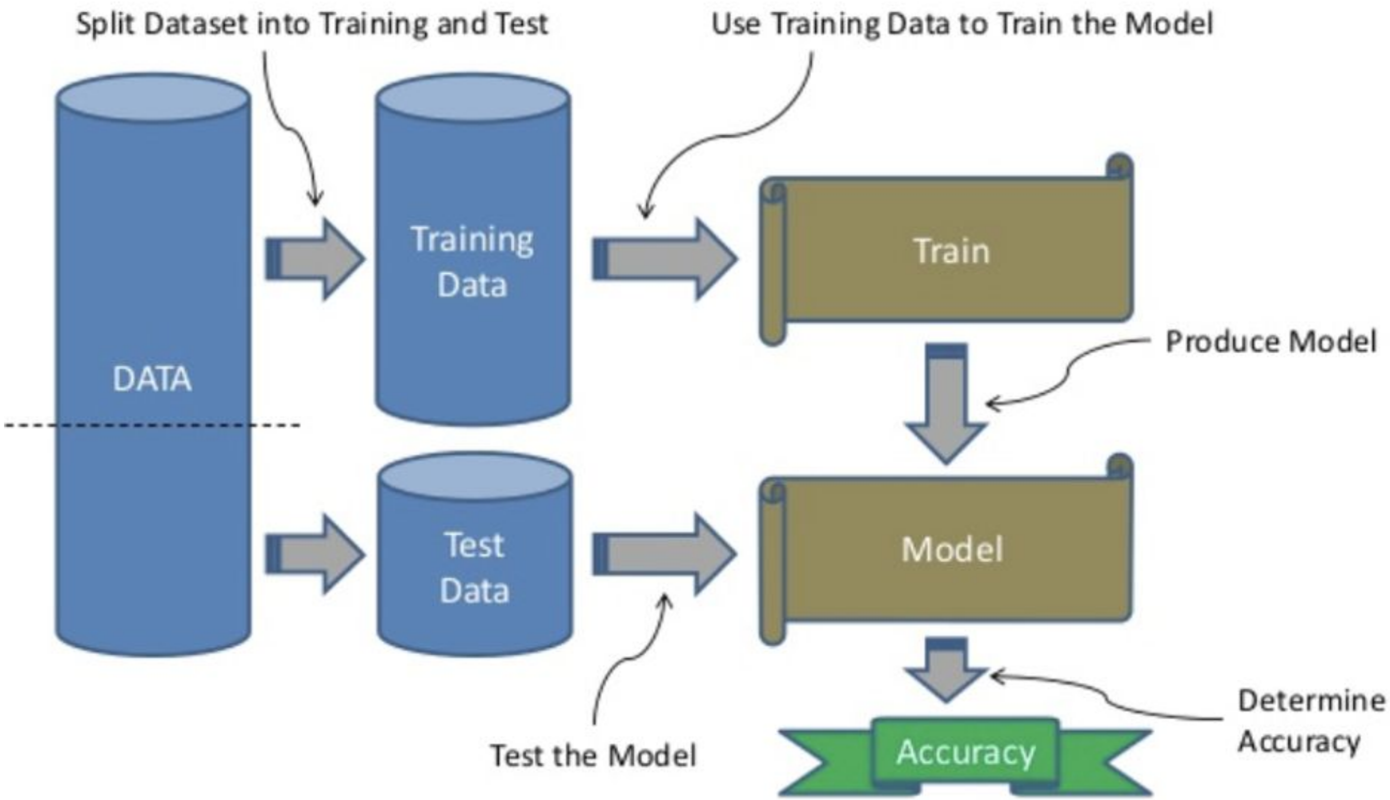
Using Machine Learning Methods to predict Algebra 1 PARCC Scores

Tracy Sweet, Brennan Register, & Ashani Jayasekera
MLDS Research Branch and UMCP

MLDS Research Series
May 19, 2023

Machine Learning

A basic supervised machine learning workflow:



Machine Learning

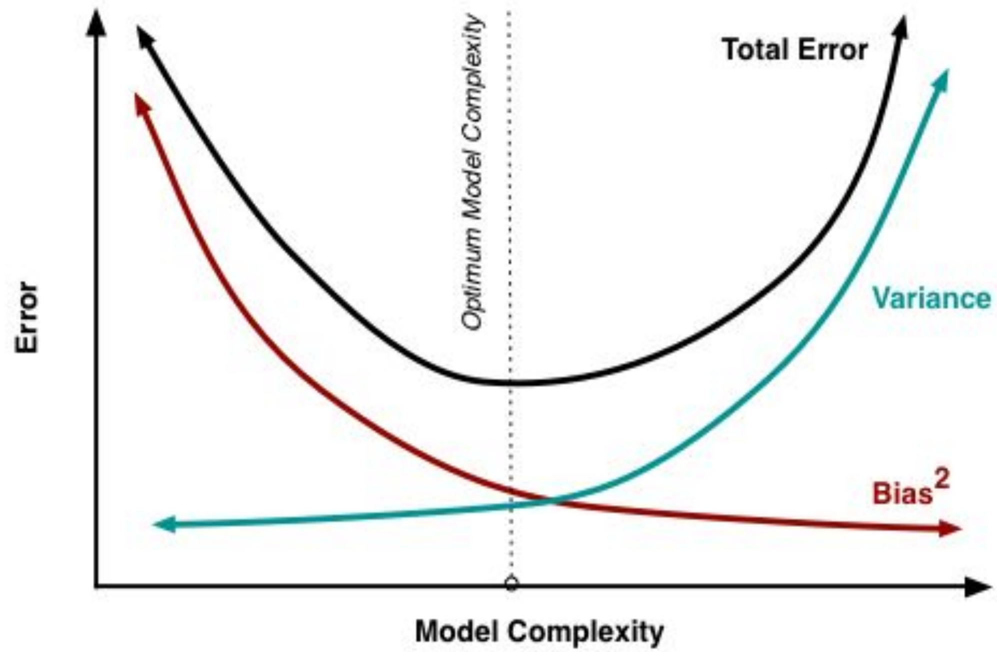
Accuracy: a measure of how well the machine learning algorithm predicts an outcome

For a continuous outcome, we use mean squared error.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Bias vs Variance

If the fitted model is too complex, it will fit the training data very well, but it will not fit the test data well.



If the fitted model is too simple, it will not fit the training data well, but will fit the test data equally well.

Machine Learning Algorithms

Multiple Linear regression

- A parametric model

Lasso regression

- A parametric model with a regularization term

Regression trees

- A non-parametric decision tree

Random forests

- A collection of trees built to reduce overfitting

Regression Models

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$$

Multiple Linear regression

$$\mathrm{RSS} = \sum_i (Y_i - \beta X_i)^2$$

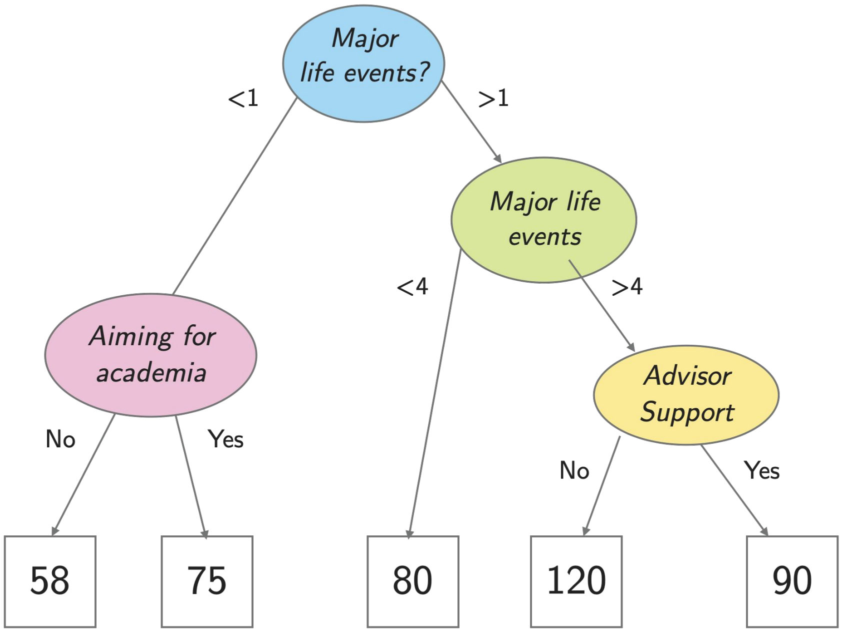
Lasso regression

$$\mathrm{RSS} = \sum_i (Y_i - \beta X_i)^2 + \lambda \sum_j |\beta_j^2$$

Regression Trees

- Partition the data based on the value of a covariate in a sequential way
- Every node of the tree produces the same predicted value
- Considered a high variance algorithm

How many months will my PhD take?



Regression Trees

Regression Trees

A single tree is fit to the data, optimizing squared error loss at each division.

Random Forests

Many trees are fit to the data, optimizing squared error loss at each division, but only a random subset of variables can be used at each split.

Using these methods to predict student outcomes in Maryland

Research Questions

- RQ1: How can data science methods be used to conduct research with MLDS data that helps to inform policy decisions in the State?
  - RQ1a: Can we predict K12 student outcome variables (e.g. PARCC Algebra 01 scores)?
  - RQ1b: Can we use predicted K12 student outcome variables as covariates in analyses of postsecondary outcomes (e.g. type of postsecondary institution)?

RQ1a: Can we accurately predict K12 student outcome variables?

Cohort

- 9^{th} grade students attending Maryland public high schools in 2019 (training set) & 2018 (testing set)

Covariates

- Student demographics
- School demographics

Outcomes

PARCC Alg01 Raw Scores

RQ1a: Can we accurately predict K12 student outcome variables?

Cohort
- 9^{th} grade students attending Maryland public high schools in 2019 (training set) & 2018 (testing set)

Covariates
- Student demographics
- School demographics

Outcomes

PARCC Alg01 Raw Scores

Took exam in 7^{th} or 8^{th} grade

Took exam in 9^{th} grade

Results :
Middle School Cohort ALG01 Scores

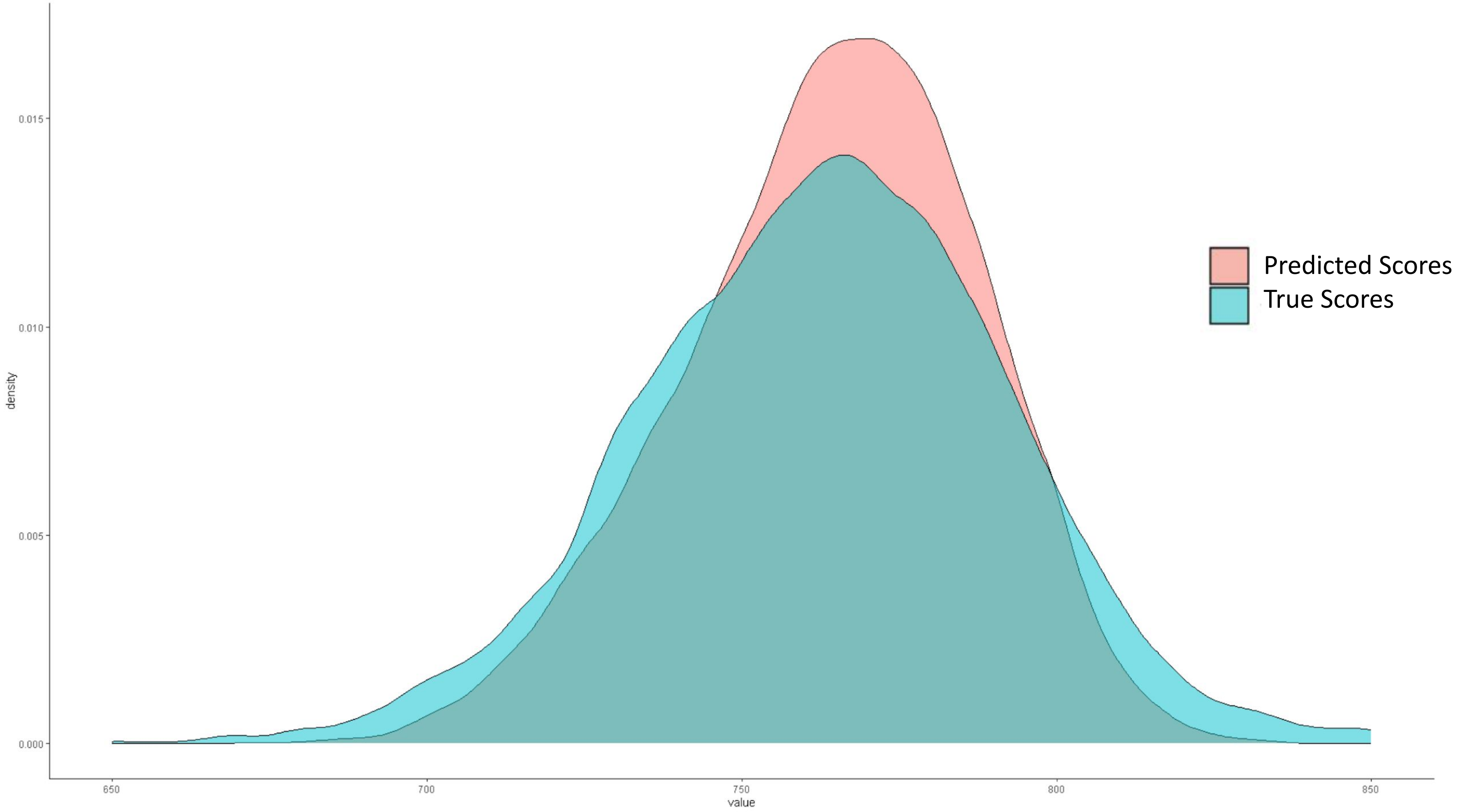
Algorithm	Mean Square Error
Linear Regression	268
Lasso	229
Decision Tree	334
Random Forest	244

Lower is better!

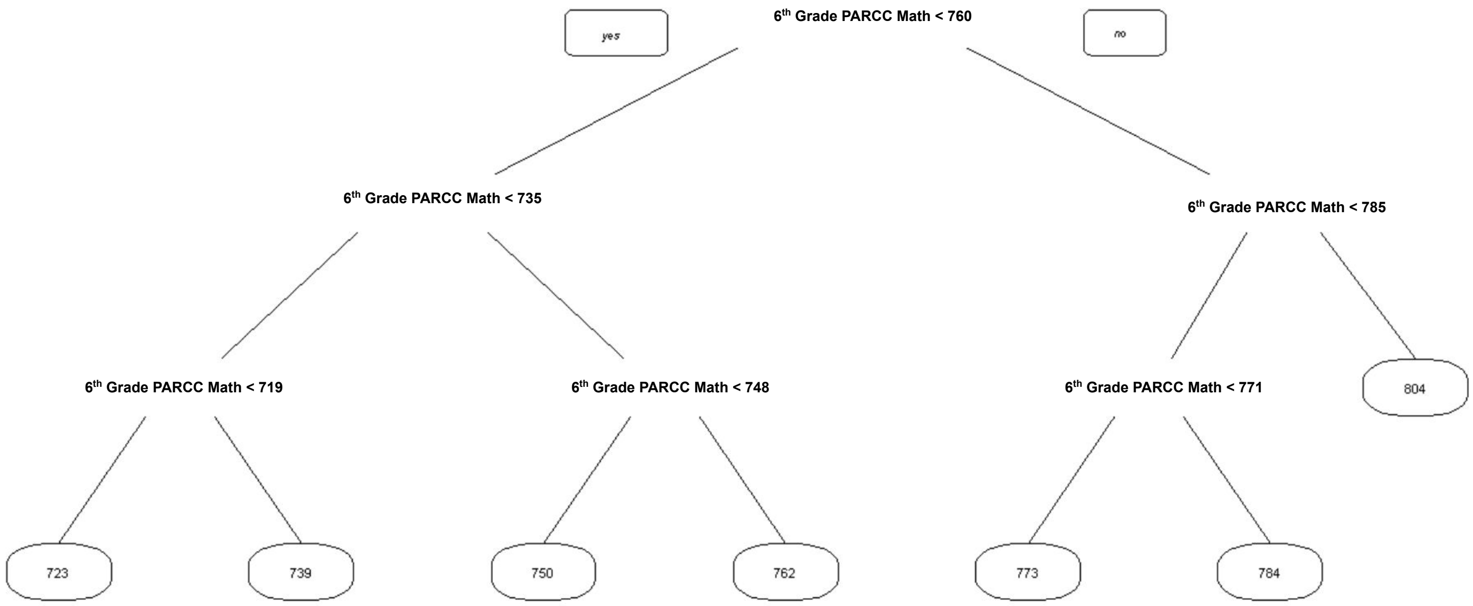
A squared error of 229 is approximately 0.5 of a standard deviation.

Results :

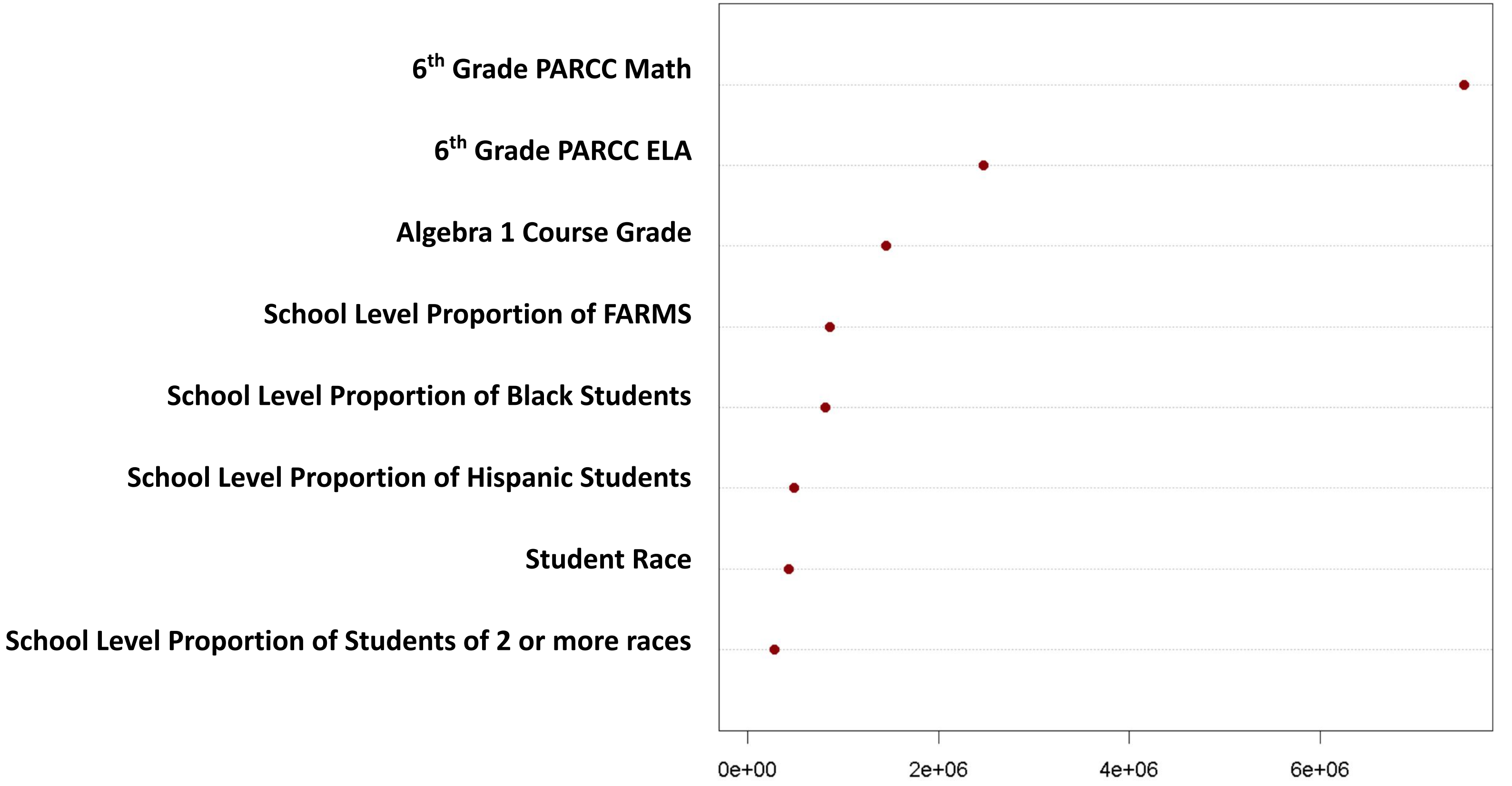
Middle School Cohort ALG01 Scores



Results :
Middle School Cohort ALG01 Scores



Results :
Middle School Cohort ALG01 Scores



MLDS Center
Maryland Longitudinal Data System

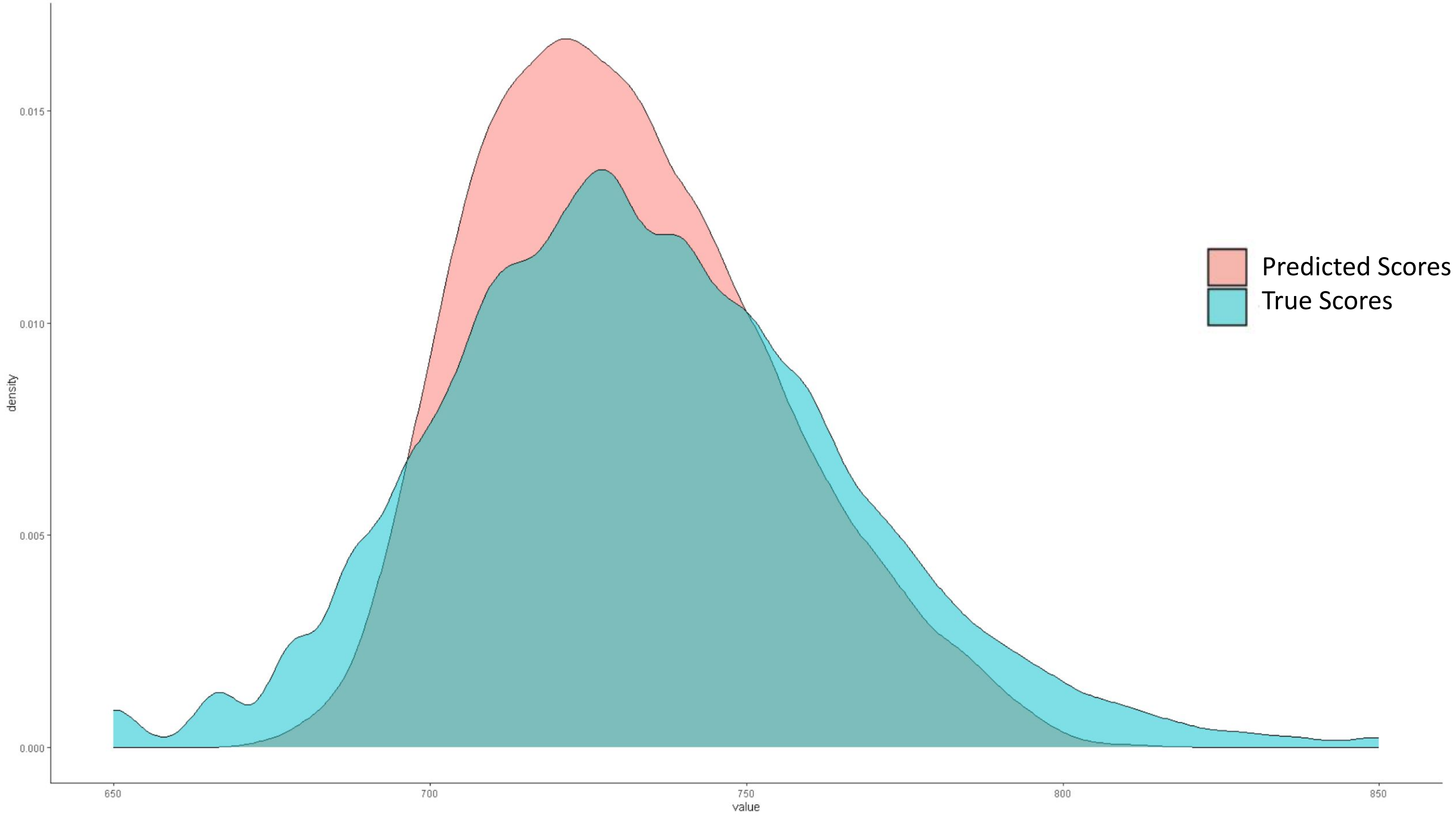
Results :
High School Cohort ALG01 Scores

Algorithm	Mean Square Error
Linear Regression	418
Lasso	334
Decision Tree	471
Random Forest	336

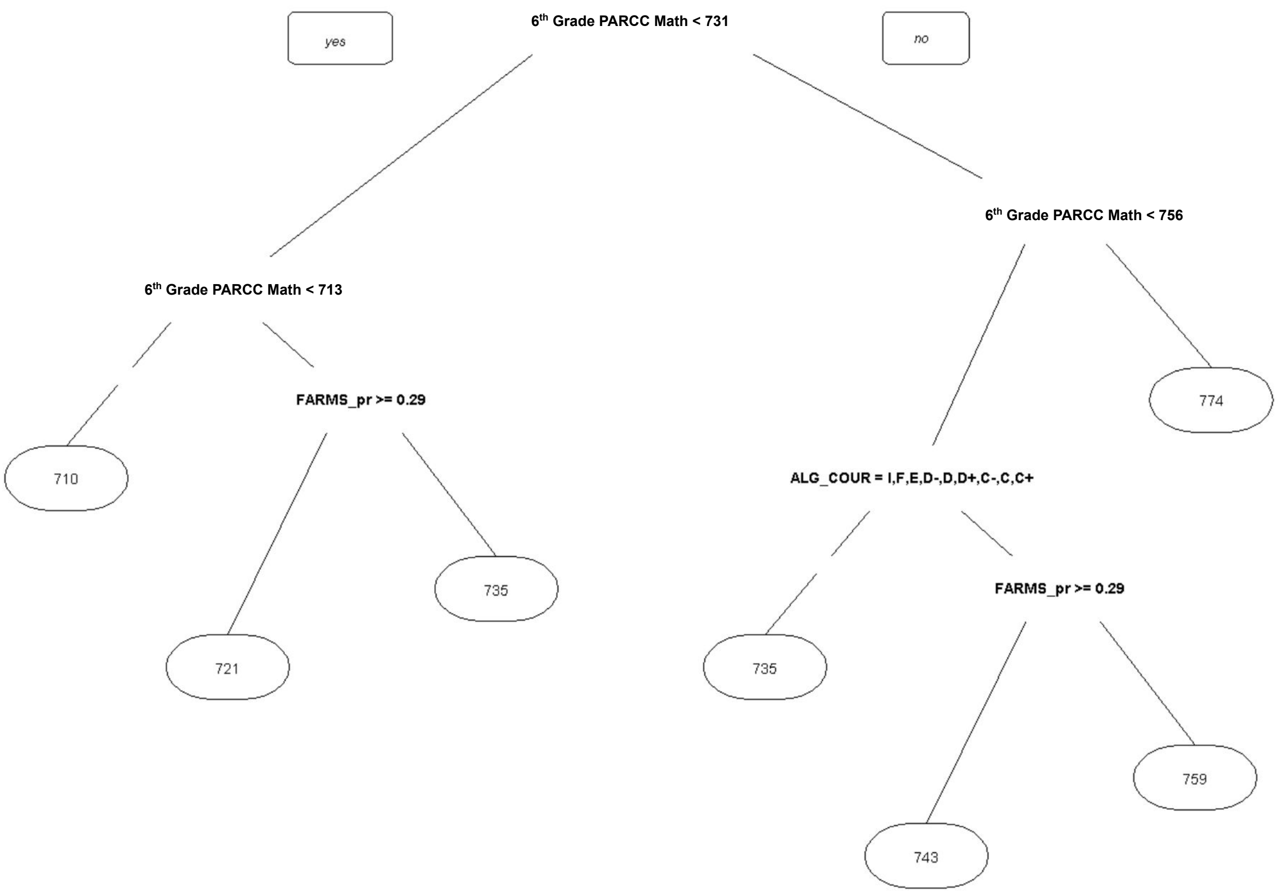
Lower is better!

A squared error of 334 is approximately 0.6 of a standard deviation.

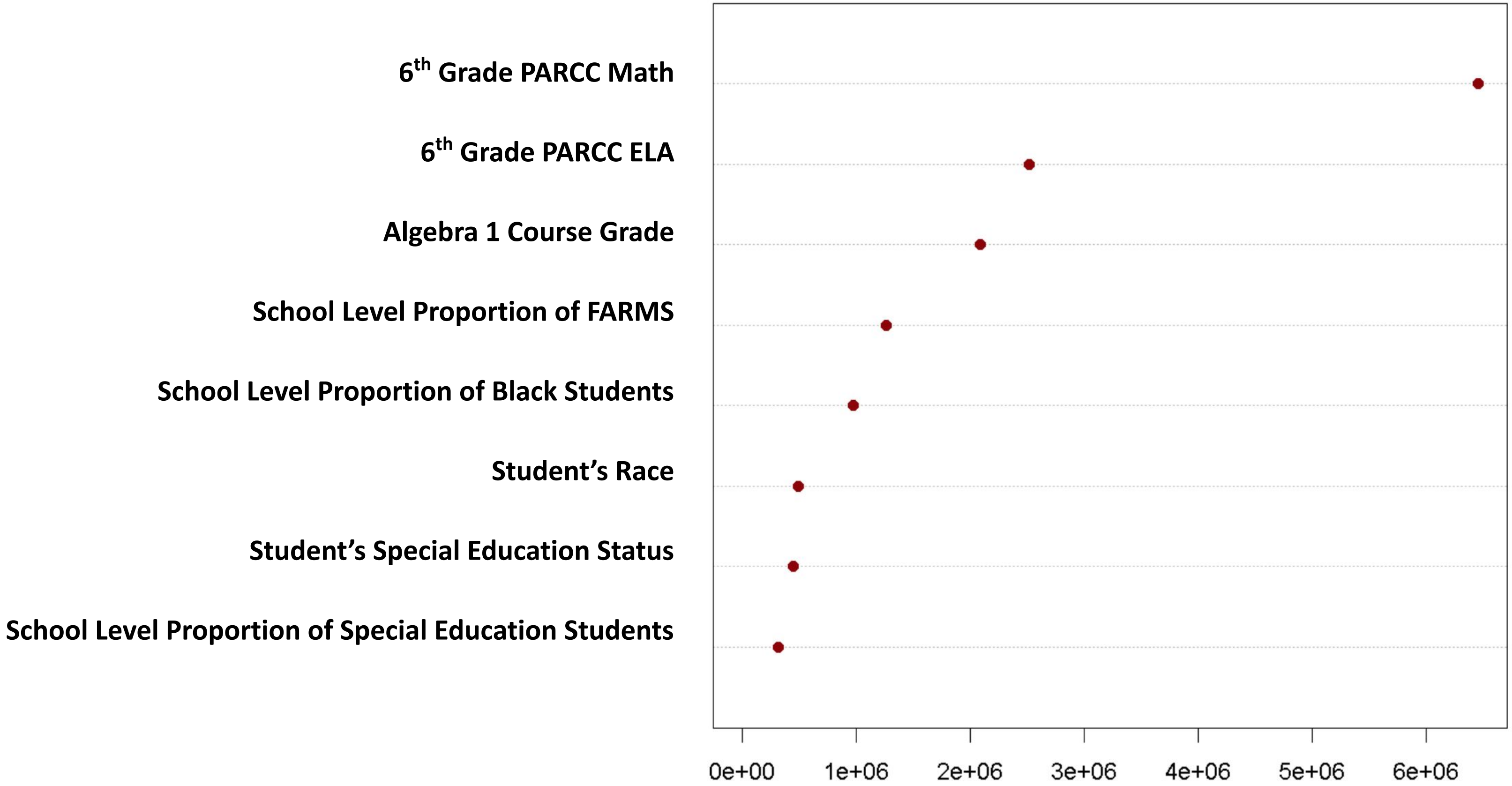
Results :
High School Cohort ALG01 Scores



Results :
High School Cohort ALG01 Scores



Results :
High School Cohort ALG01 Scores



RQ1b: Can we use predicted K12 student outcome variables as covariates in analyses of post-secondary outcomes?

Cohort

- 9th grade students attending Maryland public high schools in 2018 who where first-time freshman at a Maryland Postsecondary institution in 2022

Covariates

- Predicted PARCC Alg01 Scores

Outcomes

- Type of postsecondary institution:

2-year vs 4-Year

Predicting Type of postsecondary institution: 2-year vs 4-Year
Middle School Cohort

Predictor	Estimated Coefficient	P-value
True ALG01 Score	**0.016**	**<0.001**
Predicted Score - LM	0.017	<0.001
Predicted Score - Lasso	**0.020**	**<0.001**
Predicted Score - Tree	0.020	<0.001
Predicted Score - Random Forest	**0.022**	**<0.001**

Predicting Type of postsecondary institution: 2-year vs 4-Year Middle School Cohort

Predictor	Estimated Coefficient	P-value
True ALG01 Score	**0.019**	**<0.001**
Predicted Score - LM	0.022	<0.001
Predicted Score - Lasso	**0.026**	**<0.001**
Predicted Score - Tree	0.022	<0.001
Predicted Score - Random Forest	**0.029**	**<0.001**

*While controlling for:
Special Education Status, English Language Learner, Race, Ethnicity, Gender

Predicting Type of postsecondary institution: 2-year vs 4-Year
High School Cohort

Predictor	Estimated Coefficient	P-value
True ALG01 Score	0.009	<0.001
Predicted Score - LM	0.012	<0.001
Predicted Score - Lasso	0.010	<0.001
Predicted Score - Tree	0.010	<0.001
Predicted Score - Random Forest	0.012	<0.001

Predicting Type of postsecondary institution: 2-year vs 4-Year

High School Cohort

Predictor	Estimated Coefficient	P-value
True ALG01 Score	0.020	<0.001
Predicted Score - LM	0.028	<0.001
Predicted Score - Lasso	0.030	<0.001
Predicted Score - Tree	0.021	<0.001
Predicted Score - Random Forest	0.030	<0.001

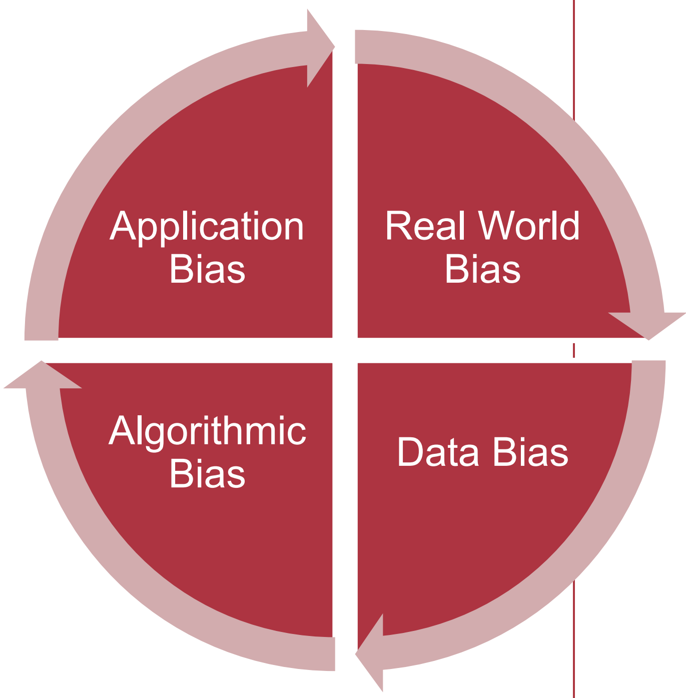
*While controlling for:
Special Education Status, English Language Learner, Race, Ethnicity, Gender

Additional Future Research Questions

- RQ2: How equitable are current analytical methods and should we conduct comparisons differently?
  - *How do different methods and algorithms compare on estimating and predicting outcomes for minoritized students?*

RQ2: Exploring Algorithmic Bias

- **Algorithmic Bias:** The phenomena where machine learning algorithms systematically discriminate against groups or individuals based on demographic characteristics (O'Neil, 2016; Crawford & Whittaker, 2018).
- Occurs either when the training data used to develop the algorithm contain inherent biases or when the algorithm induces biases during the learning process.



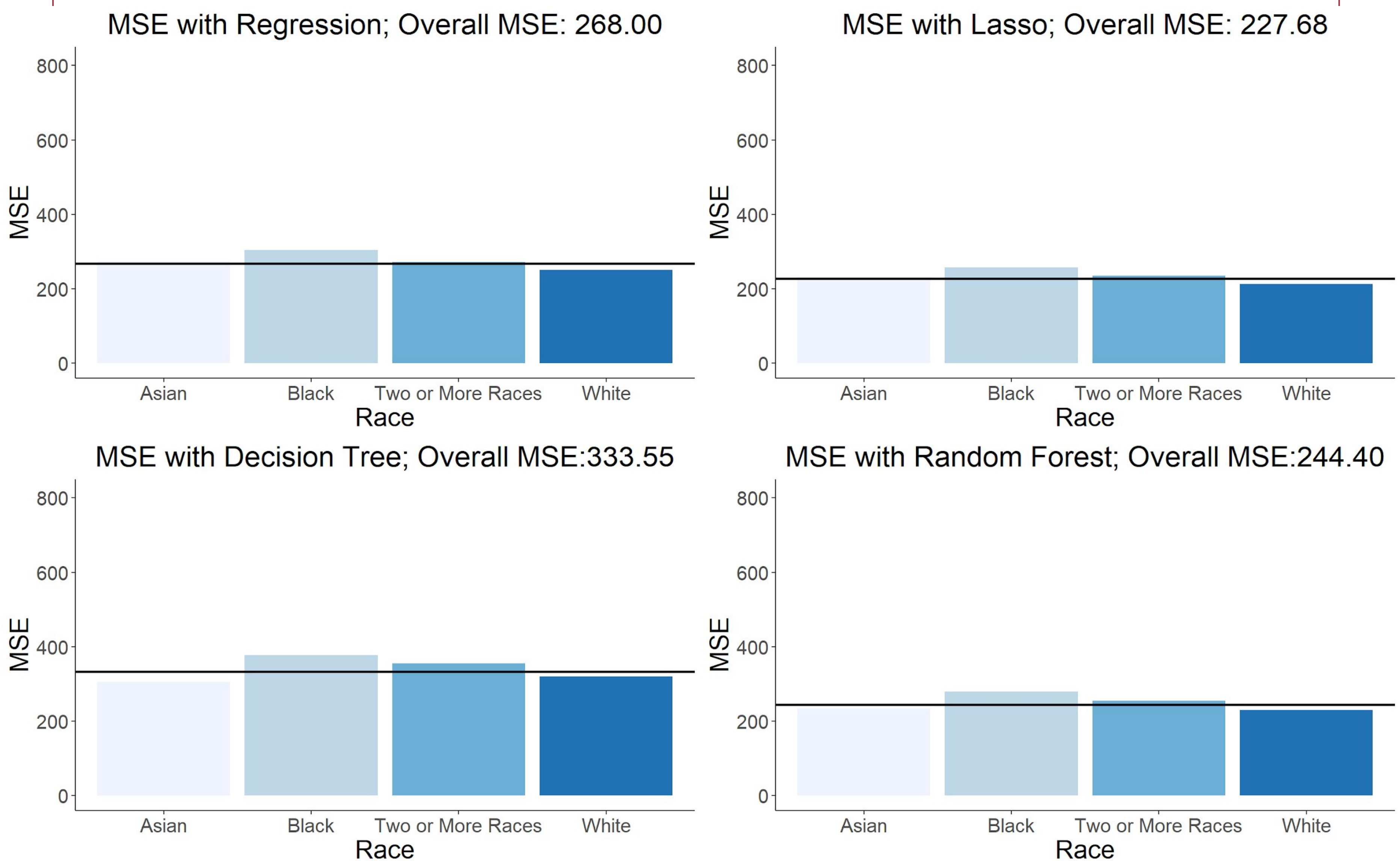
RQ2: Exploring Algorithmic Bias

- Comparative Analysis (Dwork et al., 2012)
  - Compare the predications of an algorithm across different groups to uncover bias.
- Bias Testing (Bolukbasi et al., 2016)
  - Statistical tests can be designed to uncover potential biases
- Causal Inference (Kilbertus et al., 2017)
  - Analyzing the causal relationship between the algorithmic inputs and outputs.

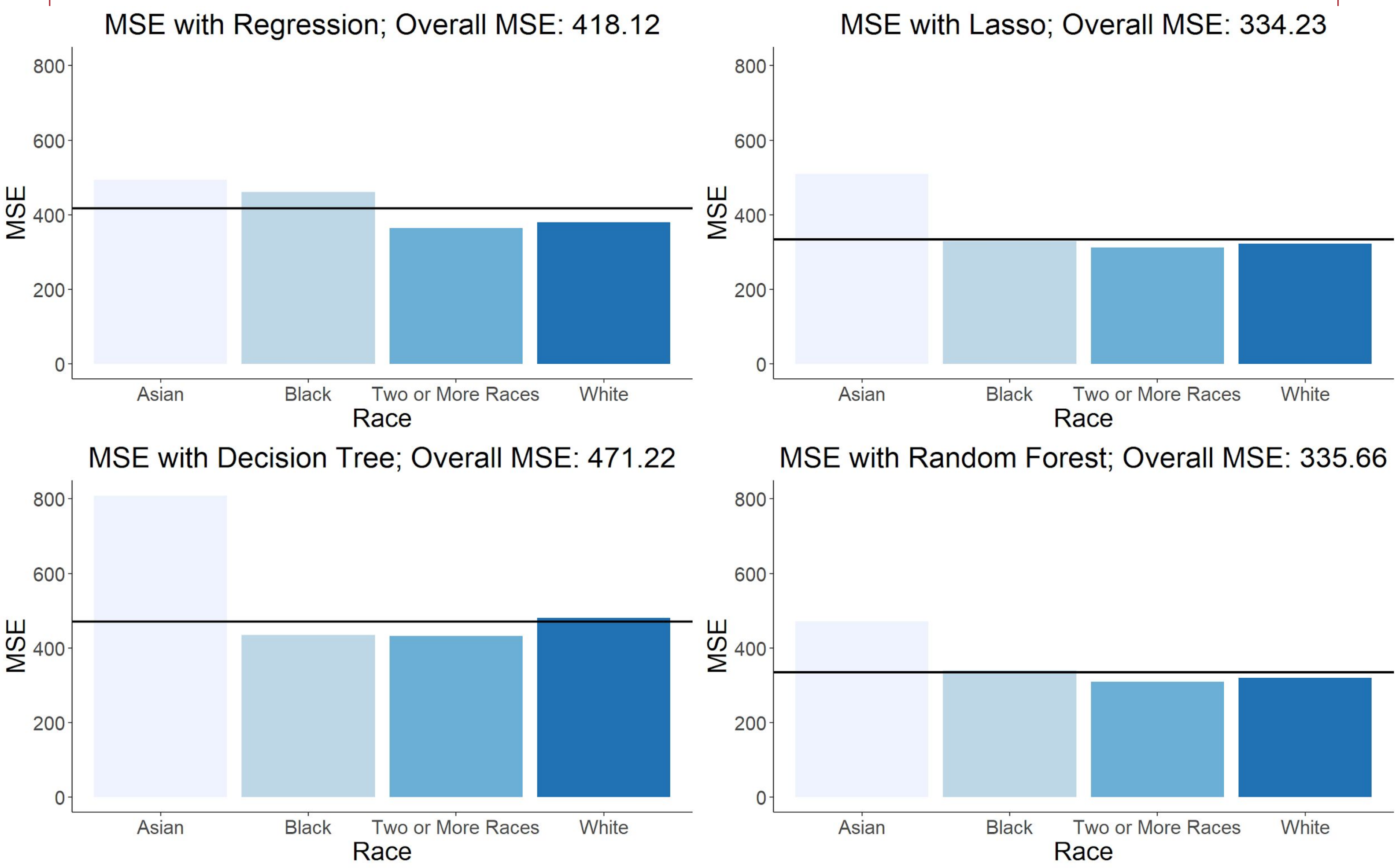
RQ2: Exploring Algorithmic Bias

- **Comparative Analysis (Dwork et al., 2012)**
  - **Compare the predications of an algorithm across different groups to uncover bias.**
- For continuous outcomes, compare MSE rates across groups
  - If no bias, then there should be the similar rates of prediction error across groups.
- For categorical outcomes, compare False Positive Rates and False Negative Rates
  - If no bias, then there should be similar rates of misclassification across groups.

Racial Bias: Middle School



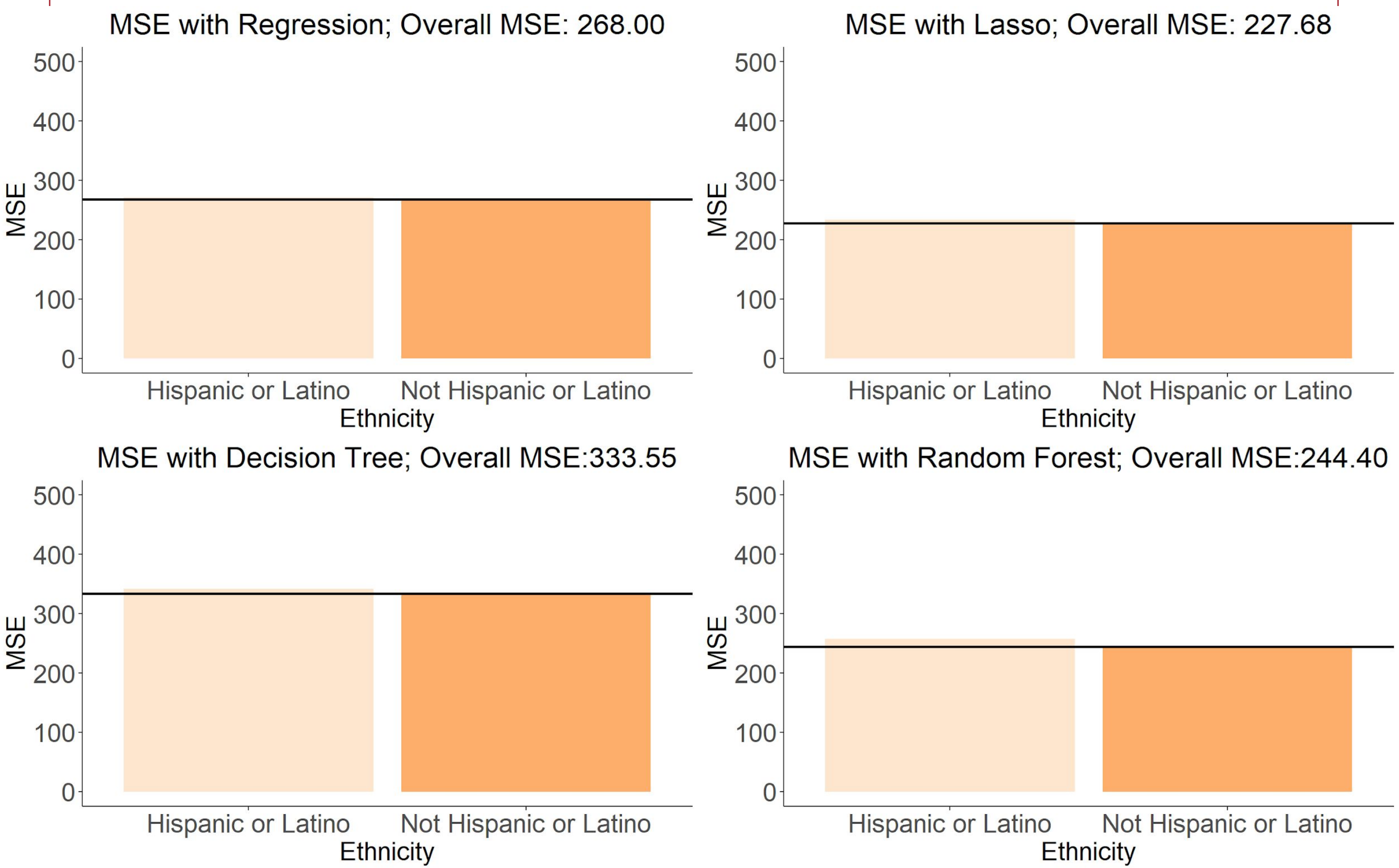
Racial Bias: High School



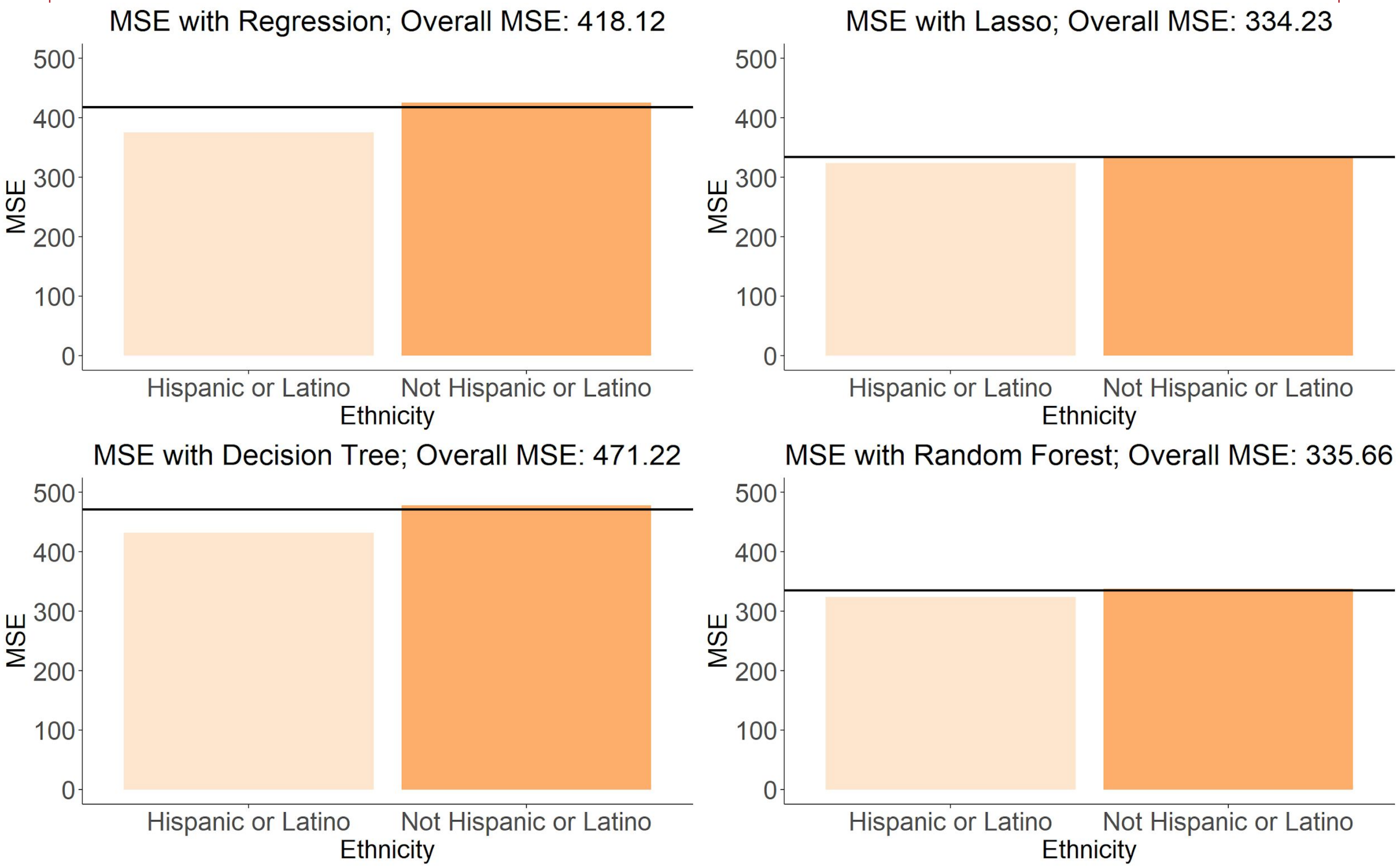
Gender Bias: Middle School

Gender Bias: High School

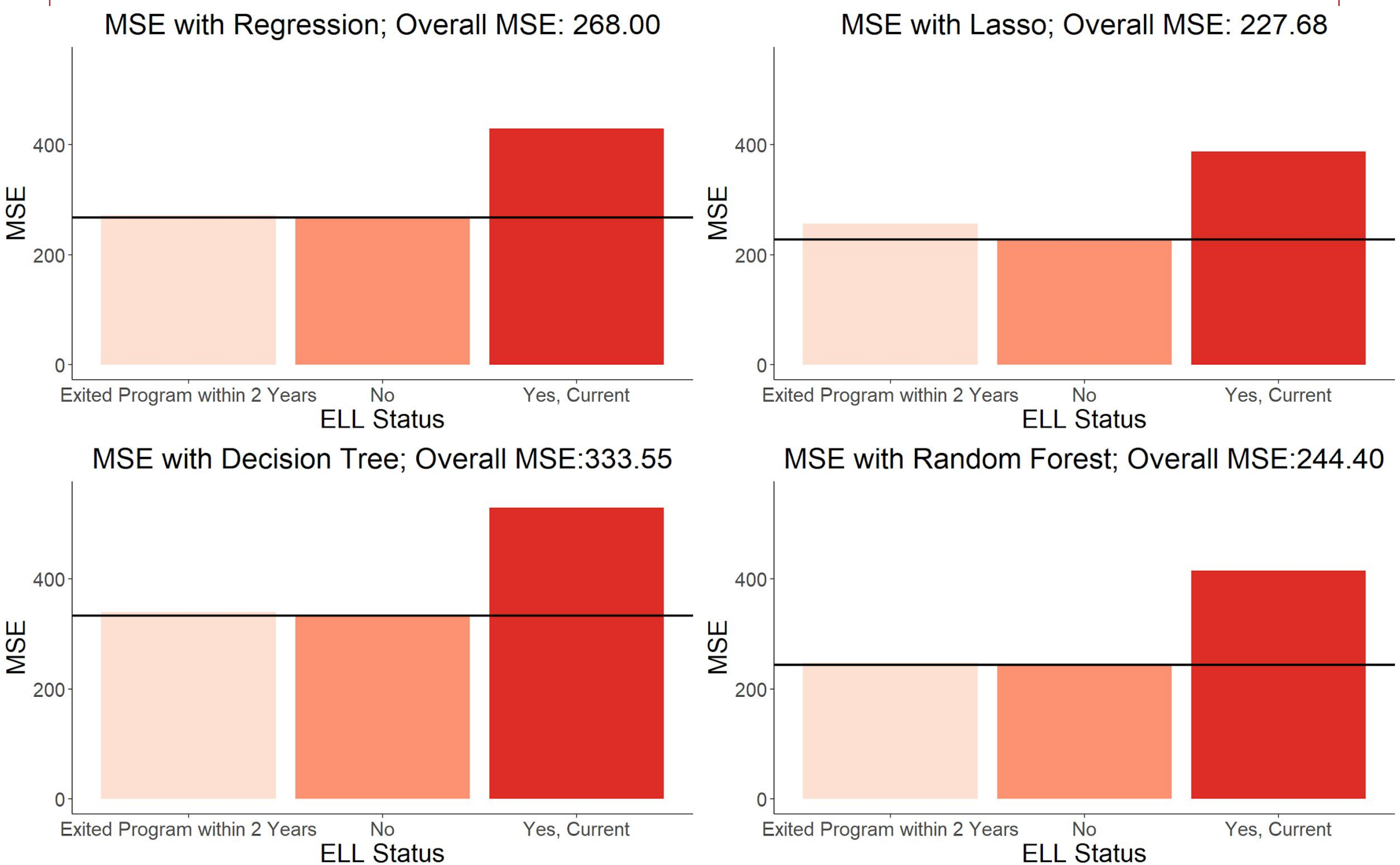
Ethnic Bias: Middle School



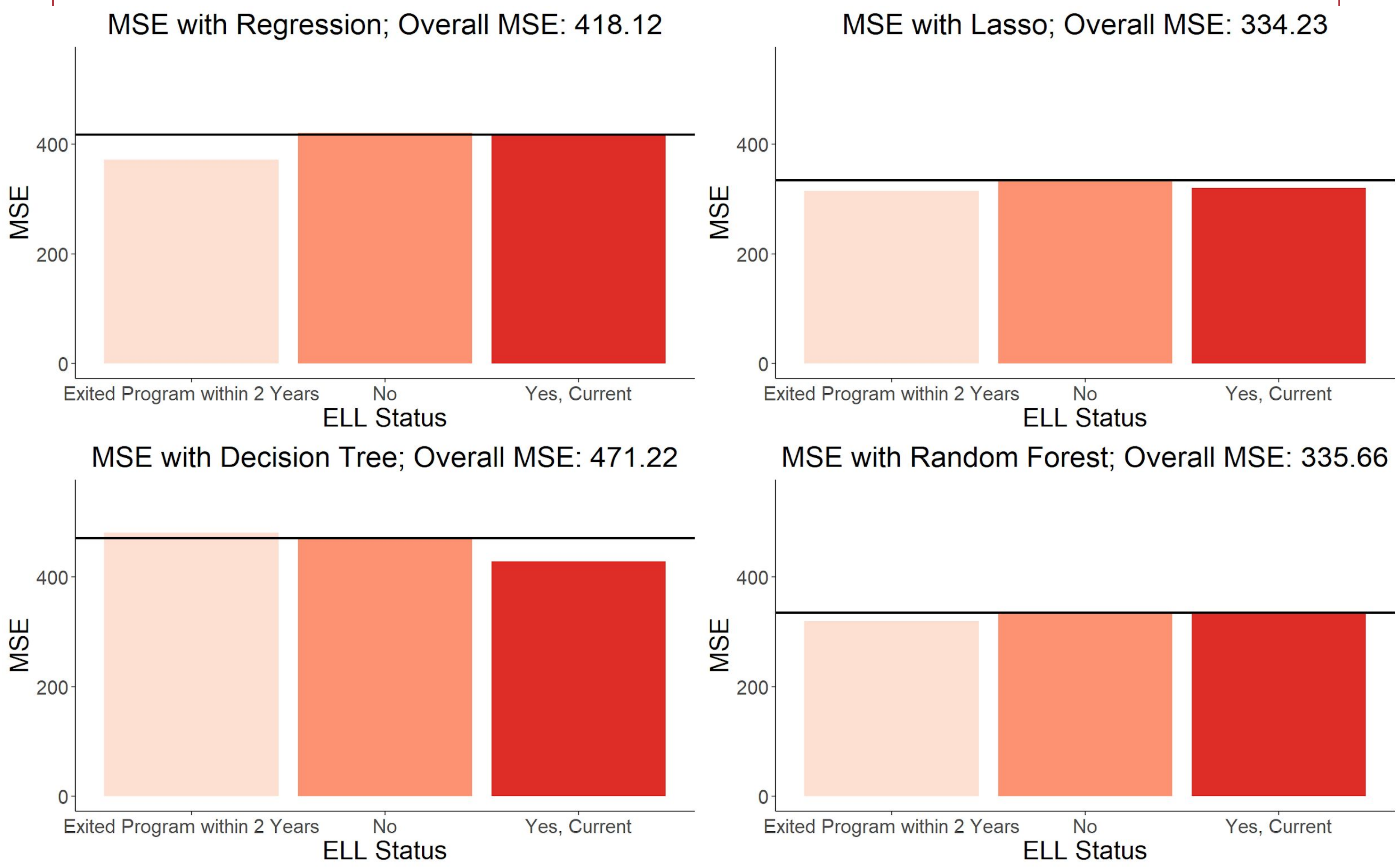
Ethnic Bias: High School



ELL Bias: Middle School



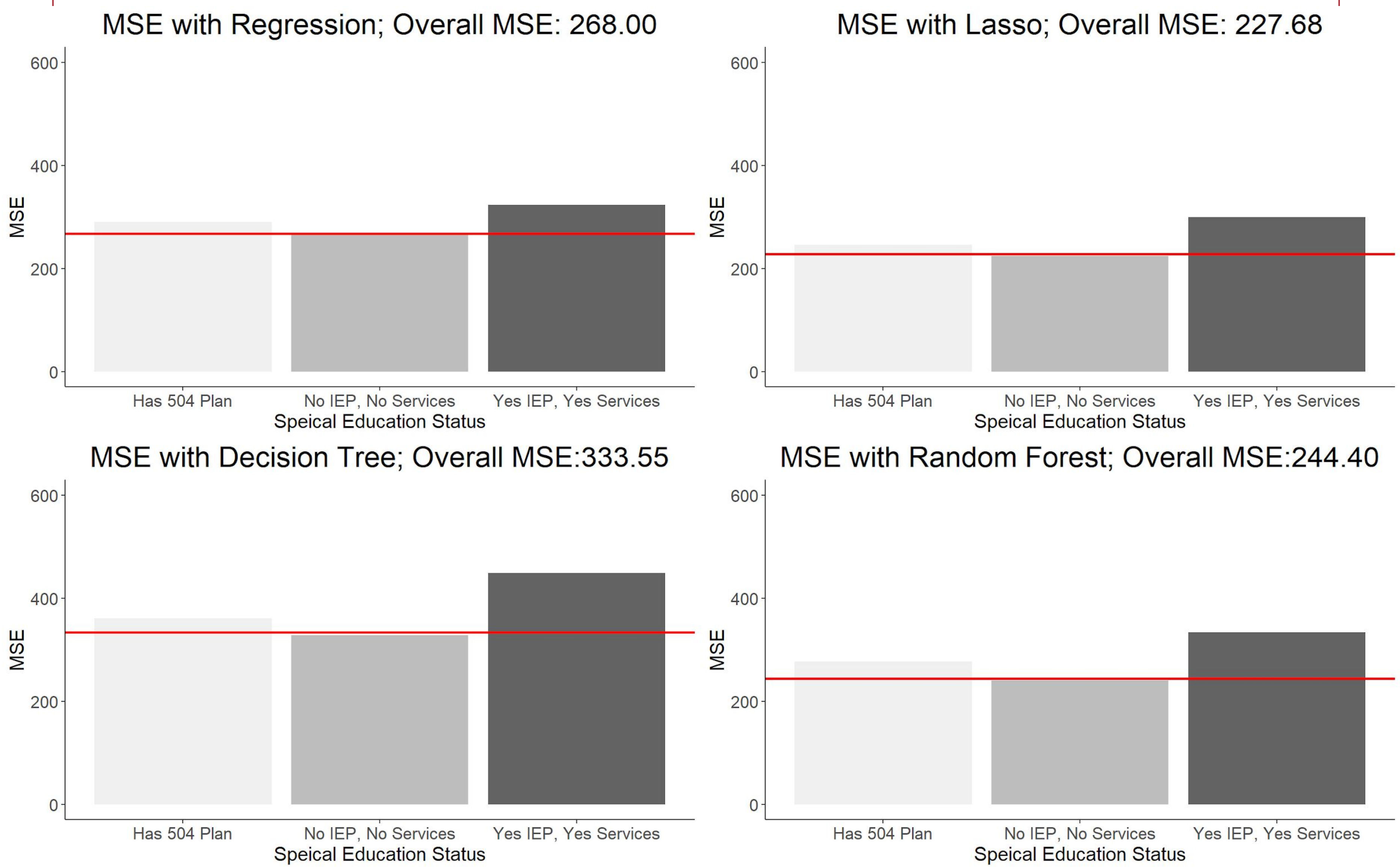
ELL Bias: High School



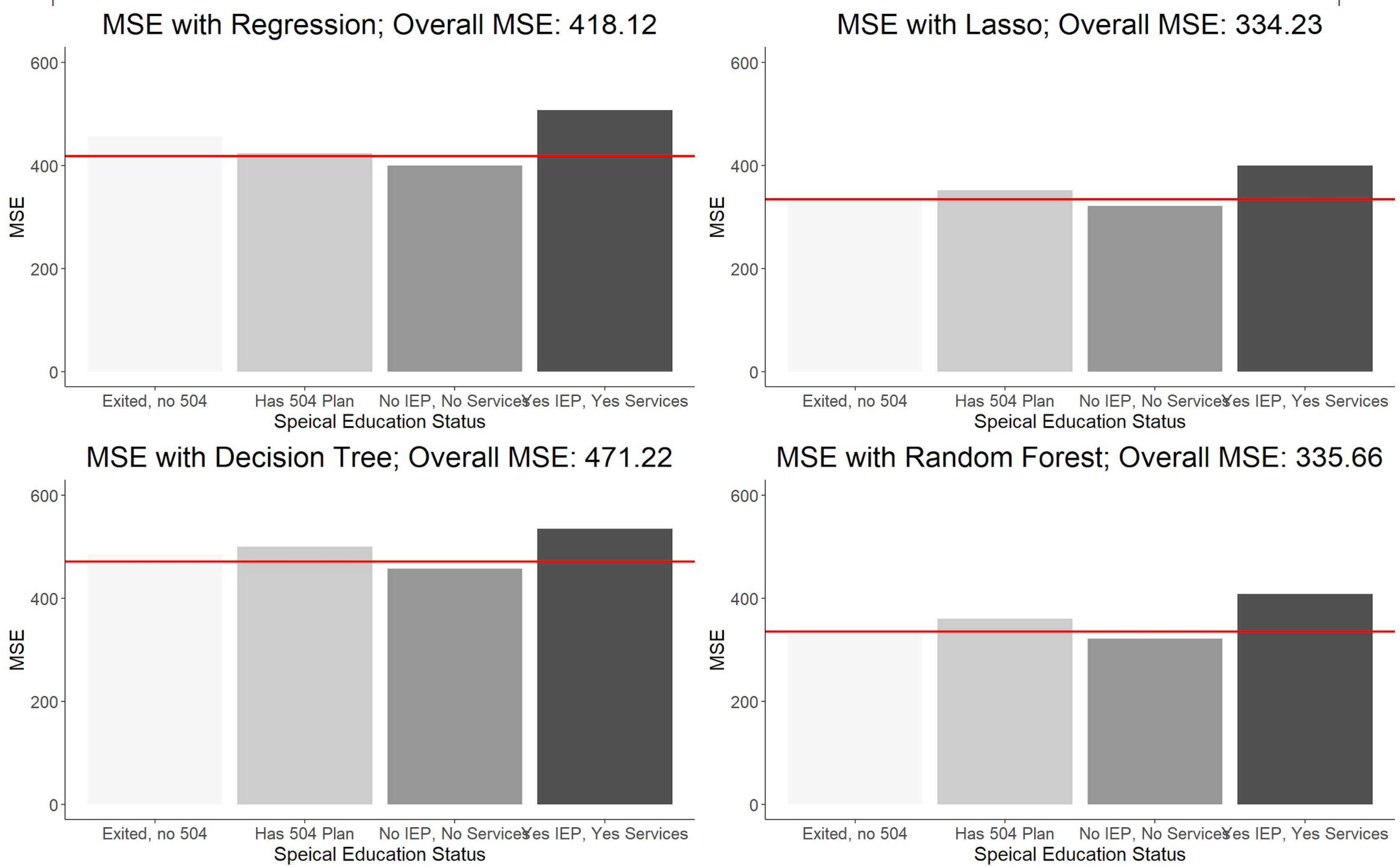
Title 1 Eligibile Bias: Middle School

Title 1 Eligible Status Bias: High School

Special Education Bias: Middle School



Special Education Bias: High School



Summary

- We applied 4 different prediction algorithms to predict Alg I test scores; Lasso and random forests were the best methods
- Our accuracy was on average about 0.5-0.6 standard deviations off
- When used as covariates, our predicted test scores over-estimated their importance
- Our algorithms did not predict all subgroups equally well
  - Multiple regression and trees were the worst.
  - Overall poor prediction for Asian and Black students and ELL students.

Conclusions

- Future work is needed to determine the extent to which ML can be used to predict missing test scores or predict student outcomes
- These algorithms should be used with caution as their predictive accuracy varies by subgroup
- Additional work is needed before they should be used to predict Black and Asian student outcomes
- Given the poor performance of regression, methodological work is needed to understand the extent to which regression should be used to address education policy research questions, especially when comparing subgroups.

Questions or Feedback?

Thank you!

References

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. https://doi.org/10.1145/2090236.2090255

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems, 29.*

Kilbertus, N., Caruana, R., Hardt, M., Veale, M., Gehrmann, S., & Ras, G. (2017). Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems, 30.*

O'Neil, Cathy. (2016). *Weapons of Math Destruction: How big data increases inequality and threatens democracy.* Broadway Books.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., ... & Schwartz, O. (2018). *AI now report 2018* (pp. 1-62). New York: AI Now Institute at New York University.