

# Research Report

October 2025

## Predicting College Freshman GPA: A Comparative Study of Traditional and Fairness-Aware Machine Learning Models

---

EDGAR I. SANCHEZ



## Conclusions

This study concludes that traditional logistic regression models, particularly those using ACT Composite scores, tend to demonstrate better fairness metrics across subgroups compared to a fairness-aware machine learning gradient-boosted machine model. The exclusion of race/ethnicity from predictive models does not introduce notable bias and may even enhance fairness, providing a lawful and effective way to evaluate students' potential success in college. The findings suggest that postsecondary institutions should adopt a combined approach using both high school GPA and ACT scores to strike a balance between fairness and predictive accuracy, while being cautious with fairness-aware machine learning models due to their complexity and potential biases.

## So What?

The practical importance of this study lies in its implications for postsecondary institutions, especially in light of the 2023 U.S. Supreme Court decision that ended affirmative action in college admissions. By comparing traditional logistic regression models with fairness-aware machine learning models, the study provides insights into how institutions can develop predictive models that balance fairness and accuracy without relying on race/ethnicity. This is crucial for complying with legal mandates while promoting equitable educational outcomes. The findings suggest that using a combined approach of high school GPA and ACT scores can help promote fairness and improve the predictive accuracy of student success, allowing institutions to more effectively allocate resources and supports to students who need them most.

## Now What?

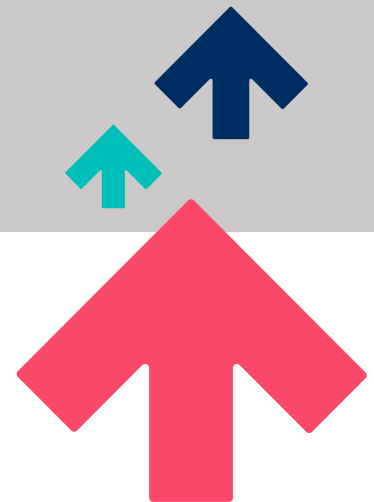
First, postsecondary institutions should consider adopting a combined approach using both high school GPA and ACT scores to develop predictive models that balance fairness and accuracy. This approach helps mitigate potential biases that arise when a model relies solely on one metric, particularly for African American and low-income students. Additionally, institutions should explore the use of fairness-aware machine learning models, but with caution, as these models may require further optimization to address potential biases and may be difficult to justify to parents and lawmakers. Postsecondary institutions should also focus on transparency and accountability in their decision-making processes, ensuring that the selection of specific models and criteria can be easily understood by and justified to students, parents, and legal authorities. Finally, institutions should incorporate nontraditional factors such as personal essays, socioeconomic background, and school context into their admissions processes to promote a more holistic evaluation of students.

## About the Author

Dr. Edgar I. Sanchez is a lead research scientist at ACT, where he studies postsecondary admissions, national testing programs, test preparation efficacy, and intervention effectiveness. Throughout his career, Dr. Sanchez has focused on studying the transition between high school and college and supporting the decision-making capacity of students, their families, and college administrators. His research has been widely cited in academic literature and by the media, including *The Wall Street Journal*, *The Washington Post*, *USA Today*, and the education trade press.

## Acknowledgments

The author would like to thank Joann Moore and Jill McVey for their comments on previous drafts of this report.



## Introduction

Predicting first-year college GPA (FYGPA) is foundational to assessing students' educational success, empowering postsecondary institutions to identify student characteristics that influence academic performance, and facilitating the design of targeted interventions. Studies have consistently shown that the predictive power of standardized tests such as the ACT and SAT and high school metrics like high school GPA (HSGPA) and coursework patterns are important for predicting freshman grades (Beard & Marini, 2018; Curabay, 2016; Friedman et al., 2024; Marini et al., 2019; McNeish et al., 2015; Sanchez, 2024; Warren & Goins, 2019; Westrick et al., 2015). Traditional metrics, like HSGPA and college entrance exam scores, offer invaluable insight into students' readiness for higher education and serve as important indicators of academic success and as tools for future planning. When institutions better understand how these factors contribute to academic success, they can allocate resources and support to students who might be in more need, especially in the critical transition from high school to college.

In 2023, the U.S. Supreme Court issued a landmark decision in *Students for Fair Admissions v. Harvard* (2023) and *Students for Fair Admissions v. University of North Carolina* (2023), effectively ending affirmative action in U.S. college admissions. The decision stated that race could no longer be considered in admissions decisions, emphasizing that such practices violate the Equal Protection Clause of the Fourteenth Amendment. This marked a significant shift in the higher education landscape. This decision emphasizes how important it is to leverage academic and nonacademic factors in order to mitigate unintended negative consequences for underserved populations while adhering to these legal mandates.

The prediction of FYGPA is important as a tool for guiding admissions policies and academic interventions, yet traditional models can potentially exhibit bias.<sup>1</sup> These biases can manifest in numerous ways, including overprediction or underprediction of GPA for specific subgroups such as women, students from lower family income backgrounds, and minority students. For example, hierarchical linear models (HLMs) using college entrance exams can unintentionally advantage students from well-resourced schools or students with access to extensive test preparation resources that may mimic schooling; this may in turn disadvantage other students. To address these concerns, the ACT uses rigorous measures to minimize bias. These include ongoing fairness reviews, item analysis to detect differential item functioning across demographic groups, and equating methods to ensure that scores are comparable across different test forms. Despite these efforts, there remains the potential for systemic bias, which can influence broader predictive models. Coupled with the recent legal restrictions prohibiting the use of race/ethnicity in predictive models, the interest in fairness-aware models has grown.

---

<sup>1</sup> In this paper, I use the term *bias* to describe data that are not fully representative of the population or research outcomes that disadvantage one or more subgroups in the population, such as those from low-income families or communities.

Advanced machine learning techniques are now being applied to help refine predictive models, offering potential advantages over traditional methods. These more modern techniques allow institutions to create predictive models that can not only predict first-year GPA with precision but also potentially help ensure fairness among underrepresented student groups, which can lead to a fairer educational environment for all students.

While traditional models such as HLMs, in conjunction with college entrance exams such as the ACT, have long been used to predict postsecondary success, including FYGPA, their effectiveness at directly addressing potential subgroup bias remains an issue of debate. The research base as it exists has extensively documented the predictive validity of these models, but these models often fail to explicitly account for potential systemic unfairness that may affect underrepresented groups. Fairness-aware machine learning models are emerging as a promising alternative, but the application of these methods to education is limited. In particular, the use of these modern techniques to address the prohibition of the use of race in admissions remains underexplored. This gap in the research suggests the need for a comparative research study examining the relative effectiveness of traditional HLMs and fairness-aware models in addressing bias, improving accuracy, and aligning practice with evolving legal standards. This study aims to fill this gap in order to advance educational predictive analytics and inform postsecondary policies to promote both fairness and predictive accuracy.

In the present study, I evaluate and compare the effectiveness of traditional regression models and fairness-aware machine learning models for predicting FYGPA. Using key predictors such as ACT Composite (ACTC) scores, HSGPA, family income, gender, race/ethnicity, and school-level characteristics, this study endeavors to understand whether fairness-aware models can address prediction biases potentially seen in traditional models. The regression models will include race/ethnicity in order to illustrate how traditional predictive models function. In contrast, the fairness-aware models will not use race/ethnicity but will leverage advanced machine learning tools available in the R statistical software. Procedures such as the *fairmodels* package, which provides metrics to assess fairness, will be explored. These algorithms help models identify and reduce prediction disparities by ensuring the outcomes are fair across student subgroups while maintaining overall accuracy. By focusing on academic preparation and student demographics other than race/ethnicity, these fairness-aware models provide a predictive methodology that ensures compliance with legal requirements.

The following research questions will be explored:

1. To what extent do traditional regression models exhibit bias for student subgroups (i.e., by gender, family income, and race/ethnicity)?
2. To what extent do fairness-aware machine learning models reduce prediction bias for student subgroups compared to traditional regression models that either do or do not include race/ethnicity?
3. How does the predictive accuracy of a fairness-aware machine learning model compare to that of traditional regression models, particularly for underrepresented subgroups?

## Methods

### Analytical Sample

The sample for this study was taken from graduating seniors in the class of 2021 in a southern U.S. state. The sample was limited to students who enrolled in a public 4-year institution in that state. In this state, nearly all public high school 11th graders had taken the ACT as part of a statewide ACT contract. The sample includes 4,711 students across 10 institutions. To be included in the study sample, students had to have valid data for race/ethnicity, family income, gender, FYGPA, ACTC score, HSGPA, the number of AP courses offered at their high school, the percentage of White students at their high school, and the percentage of students meeting federal poverty guidelines at their high school. For the purposes of model building and evaluation, the 4,711 students were divided into a training set ( $n = 3,298$ ) and a testing set ( $n = 1,413$ ).

Table 1 compares the sample to the population of interest. Notable differences include the removal of students with missing family income and the removal of smaller proportions of students with missing race/ethnicity or gender. Achievement was similar across groups, and school characteristics were similar as well.

**Table 1.** Population and Sample Characteristics

| Demographic characteristic              |   | Population   | Sample       |
|---|---|--------------|--------------|
| <b><i>N</i></b>                         |   | 7,284        | 4,711        |
| <b>Race/ethnicity</b><br><i>n (%)</i>   | African American                                      | 1,029 (14.1) | 618 (13.1)   |
|   | American Indian / Native Hawaiian / two or more races | 373 (5.1)    | 252 (5.3)    |
|   | Asian   | 213 (2.9)    | 127 (2.7)    |
|   | Hispanic  | 684 (9.4)    | 491 (10.4)   |
|   | White   | 4,407 (60.5) | 3,223 (68.4) |
|   | Prefer not to respond / missing                       | 578 (7.9)    | 0 (0.0)      |
|   | <b>Family income</b><br><i>n (%)</i>                  | <\$36K       | 1,304 (17.9) |
| \$36K–\$60K                             |   | 1,049 (14.4) | 894 (19.0)   |
| \$60K–\$100K                            |   | 1,407 (19.3) | 1,222 (25.9) |
| >\$100K                                 |   | 1,688 (23.2) | 1,498 (31.8) |
| Missing                                 |   | 1,836 (25.2) | 0 (0.0)      |
| <b>Gender</b><br><i>n (%)</i>           | Female  | 3,980 (54.6) | 2,812 (59.7) |
|   | Male  | 2,843 (39.0) | 1,899 (40.3) |
|   | Another gender  | 12 (0.2)     | 0 (0.0)      |
|   | Prefer not to respond                                 | 34 (0.5)     | 0 (0.0)      |
|   | Missing   | 415 (5.7)    | 0 (0.0)      |
| <b>FYGPA</b><br>Mean ( <i>SD</i> )      |   | 2.84 (1.07)  | 2.93 (1.02)  |
| <b>ACTC score</b><br>Mean ( <i>SD</i> ) |   | 21.90 (5.12) | 22.03 (4.94) |

| Demographic characteristic                       | Population    | Sample        |
|--|---------------|---------------|
| <b>HSGPA</b><br>Mean (SD)                        | 3.59 (0.42)   | 3.59 (0.42)   |
| <b>% poverty</b><br>Mean (SD)                    | 18.81 (7.13)  | 18.47 (6.94)  |
| <b>Number of AP courses offered</b><br>Mean (SD) | 3.40 (1.00)   | 3.43 (0.95)   |
| <b>% White</b><br>Mean (SD)                      | 65.26 (25.17) | 67.23 (24.28) |

## Measures

### **ACT Composite Score**

The official ACT Composite (ACTC) scores were obtained from the last ACT test administration that students took before high school graduation. These scores were obtained either during statewide school-day testing or during a national test administration.

### **Cumulative HSGPA**

To calculate each student's HSGPA, ACT averaged student self-reported grades in up to 23 courses in English, mathematics, social studies, and natural science. Sanchez and Buddin (2016) demonstrated high correlations between students' self-reported HSGPA and students' transcript GPA; additional research supports the use of self-reported data for research purposes (Camara et al., 2003; Kuncel et al., 2005; Shaw & Mattern, 2009).

### **Official First-Year GPA**

Official FYGPA was obtained from student transcripts at the colleges where students enrolled the fall following high school graduation.

### **Demographic Variables**

The study examined three demographic variables: gender, race/ethnicity, and family income (see [Table 2](#)). Students reported their gender as male, female, another gender, or prefer not to respond; some did not respond. For this analysis, students who responded as another gender (0.2%), preferred not to respond (0.5%), or did not provide a response (5.7%) were removed from the analysis. This was done because the results of analyzing these groups could not be clearly explained.

Students could indicate their racial/ethnic identity as Asian, Black, Hispanic, American Indian/Alaska Native, Native Hawaiian/Pacific Islander, White, two or more races, or prefer not to respond; some did not respond. Due to low numbers of students in some groups, I combined data for students identifying as American Indian/Alaska Native (0.4%), Native Hawaiian/Pacific Islander (<0.1%), and two or more races (4.7%). Once again, students who preferred not to respond or who did not provide their race/ethnicity (7.9%) were omitted because of difficulty in interpreting the results.

Students' family income was categorized into four categories: below \$36,000, \$36,000–\$60,000, \$60,000–\$100,000, and above \$100,000. Students with missing family income (25.2%) were omitted.

## Data Analysis

The present study used both logistic regression and a gradient-boosted machine learning model to predict the likelihood of students attaining a FYGPA of C or better (i.e., a 2.00 or better on a 4.00 scale). This outcome was selected because it indicates a minimally acceptable student outcome from a postsecondary institution perspective. I considered hierarchical logistic regression; however, the intraclass correlation for such a model was low, indicating that little variance was explained by differences between postsecondary institutions. As such, I selected a simpler logistic regression model approach.

To evaluate the performance of the models considered, I calculated five fairness metrics for each model. These metrics focus on the effects of the model on focal and reference groups. This terminology is adapted from policy discussions where focal groups are those that are safeguarded against discrimination by law or policy, often based on attributes like race, gender, age, or disability. Reference groups are those that do not fall under these specific protections. In the present context, it would also be appropriate to think of these as “majority” and “minority” demographic groups. The five fairness metrics I used to evaluate the models are as follows:

1. Accuracy equality ratio: This measures the ratio of accuracy between the focal and reference groups. This metric compares the accuracy of predictions made about two different groups to see if the model is equally accurate for both. The formula for accuracy is  $\frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{False Positives} + \text{True Negatives} + \text{False Negatives}}$ . The accuracy is calculated for both the focal and reference groups. The ratio is then calculated as  $\frac{\text{Accuracy}(\text{Focal Group})}{\text{Accuracy}(\text{Reference Group})}$ .
2. Equal opportunity ratio: This assesses the ratio of true positive rates (TPR) between the focal and reference groups. This metric checks whether the model is equally good at correctly identifying students who get a C or higher FYGPA for two different groups. The formula for TPR is  $\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$ . The equal opportunity ratio is then calculated as  $\frac{\text{True Positive Rate}(\text{Focal Group})}{\text{True Positive Rate}(\text{Reference Group})}$ .
3. Predictive equality ratio: This evaluates the ratio of false positive rates (FPR) between the focal and reference groups. This metric examines whether the model produces the same rate of false positives, incorrectly identifying students as earning a C or higher FYGPA when they do not, for two different groups. The formula is  $\frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$ . The predictive equality ratio is then calculated as  $\frac{\text{False Positive Rate}(\text{Focal Group})}{\text{False Positive Rate}(\text{Reference Group})}$ .

4. Predictive parity ratio: This measures the ratio of positive predictive values (PPV), also known as precision, between the focal and reference groups. This metric indicates whether the model's positive predictions, identifying someone as having a C or higher FYGPA, are equally reliable for two different groups. The formula is

$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \cdot \frac{\text{Positive Predictive Values (Reference Group)}}{\text{Positive Predictive Values (Focal Group)}}$$

The predictive parity ratio is calculated as

5. Statistical parity ratio: This assesses the ratio of the probability of favorable outcomes between the focal and reference groups. This final metric compares the likelihood of positive outcomes (i.e., attaining a C or higher FYGPA) between different groups to see if one group is advantaged over the other. The formula is

$$\frac{\text{True Positives} + \text{False Positives}}{\text{True Positives} + \text{False Positives} + \text{True Negatives} + \text{False Negatives}} \cdot \frac{\text{Probability of Favorable Outcomes (Focal Group)}}{\text{Probability of Favorable Outcomes (Reference Group)}}$$

The statistical parity ratio is calculated as

For each model considered, these five metrics are represented in a horizontal bar chart ([Figure 1](#) and following) that contains three segments. These metrics use the 80% rule (also known as the 4/5ths rule), which originates from guidelines in the U.S. Equal Employment Opportunity Commission in the context of employment discrimination. In this rule, a selection rate for any focal group (e.g., female, less than \$36,000 annual income, or racial/ethnic minority) should not be less than 80% of the selection rate for the group with the highest selection rate. When the selection rate for the focal group falls below this threshold, it may indicate adverse impacts or discrimination in the selection process. In the present study, this 80% rule serves as a practical guide rather than as a strict legal standard or statistical cutoff. It is important to note that this 80% rule does not definitively prove bias; rather, it indicates scenarios where further investigation may be necessary.

Ratios below 0.8 (4/5) indicate that the focal group is disadvantaged compared to the reference group for that metric. This portion of the graph is indicated by a red background on the left side of the figure. When the ratio falls between 0.8 and 1.25, this green section represents an acceptable fairness threshold, which can also be thought of as acceptable differences between the focal and reference groups for that metric. When the ratio exceeds 1.25 (5/4), this red section of the figure indicates that the focal group is favored over the reference group for that metric.

In this context, we should keep in mind that the ideal fairness value would be a 1.0, meaning that the model treats both the focal and the reference groups equally for the given metric. Ratios above a 1.25 can be considered as favoring focal groups, or it can be said that the model performs better for the focal group, but this could indicate an unintentional overcompensation or correction for previous bias. In ratios below 0.8, the model underperforms for the focal group compared to the reference group, which indicates potential bias against the focal group.

To compare models, I compared seven measures of predictive accuracy: accuracy, precision, recall, F1 score, AUC (area under the ROC curve), the accuracy for African American students,

and the accuracy for low-income students (less than \$36,000). Accuracy measures the proportion of correct predictions out of all predictions made, providing a general sense of the model's performance. Precision is the ratio of true positive predictions to the total predicted positives, indicating the model's ability to correctly identify positive instances. Recall (or sensitivity) is the ratio of true positive predictions to all actual positives, reflecting the model's ability to capture all relevant instances. F1 score is the harmonic mean of precision and recall, balancing the trade-off between these two metrics. AUC evaluates the model's ability to distinguish between classes, with higher values indicating better performance. African American accuracy refers to the accuracy of predictions specifically for African American students, highlighting potential biases in the model's performance for this subgroup. Low-income accuracy measures the accuracy of predictions for low-income students, ensuring the model's effectiveness across different socioeconomic backgrounds.

## Logistic Regression Models

In order to establish a baseline for these fairness metrics, I estimated four preliminary logistic regression models. First, I estimated a model that included only HSGPA as an academic indicator, plus student demographic and school characteristics. Second, I estimated a model that included only ACTC score plus student demographics and school characteristics. A third model included both HSGPA and ACTC score, along with their interaction, with student demographics and school characteristics being estimated. The HSGPA model could be considered the traditional model used in a test-blind context, whereas the ACTC score model represents an alternative based on standardized test scores. The HSGPA and ACTC score model (Model 3) will be considered the baseline, or the traditional model that a postsecondary institution might have employed prior to the Supreme Court ruling limiting the use of race/ethnicity in admissions decisions. Model 4 largely resembled the HSGPA and ACTC score model, but with race/ethnicity removed (see the appendix for model coefficients). These models are as follows:

Model 1. Attainment of a C or higher = HSGPA + Gender + Race/ethnicity + Family income + Percentage of students at a student's high school meeting poverty guidelines + Number of AP courses offered at the student's high school + Percentage of White students at the student's high school

Model 2. Attainment of a C or higher = ACTC + Gender + Race/ethnicity + Family income + Percentage of students at a student's high school meeting poverty guidelines + Number of AP courses offered at the student's high school + Percentage of White students at the student's high school

Model 3. Attainment of a C or higher = HSGPA + ACTC score + (HSGPA \* ACTC score interaction) + Gender + Race/ethnicity + Family income + Percentage of students at a student's high school meeting poverty guidelines + Number of AP courses offered at the student's high school + Percentage of White students at the student's high school

Model 4. Attainment of a C or higher = HSGPA + ACTC score + (HSGPA \* ACTC score interaction) + Gender + Family income + Percentage of students at a student's high school meeting poverty guidelines + Number of AP courses offered at the student's high school + Percentage of White students at the student's high school

## Gradient-Boosted Machine Fairness-Aware Models

Model 5 was a gradient-boosted machine (GBM) model that mirrored Model 4.

Model 5. GBM: Attainment of a C or higher = HSGPA + ACTC score + (HSGPA \* ACTC score interaction) + Gender + Family income + Percentage of students at a student's high school meeting poverty guidelines + Number of AP courses offered at the student's high school + Percentage of White students at the student's high school

GBM models are machine learning models that are used for both regression and classification tasks. As employed in this paper, the GBM model is used for the classification of students as having either attained a C or higher FYGPA or not. These types of models work by building an ensemble of decision trees, wherein each tree is trained to correct the errors made by previous trees. This process starts with a simple model, and subsequent models are added sequentially, each one focusing on the residuals (i.e., the error) of the combined ensemble of previous models. This iterative process continues until a specified number of trees is reached or the model's performance no longer significantly improves.

GBM models are fundamentally meant to minimize a loss function by using a gradient descent. At each step, the algorithm calculates the gradient of the loss function with respect to the current model's predictions and fits a new tree to this gradient. This new tree is then added to the ensemble with certain weights (i.e., shrinkage or learning rate), that control the contribution of each tree. By combining the predictions of all the trees, gradient-boosted machine models can capture complex patterns in the data and achieve high predictive accuracy.

The loss function being minimized by the model in this study is the binary cross-entropy loss function (also known as the log loss):

$$L(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)],$$

where  $y_i$  is the binary outcome (C or higher/not C or higher) for the  $i^{\text{th}}$  observation,  $\hat{y}_i$  is the predicted probability of achieving a grade of C or higher for the  $i^{\text{th}}$  observation, and  $N$  is the total number of observations.

Model 5 employs a gradient-boosted machine model with a Bernoulli distribution and 1,000 trees. It is worth noting that 793 trees were mathematically sufficient for minimizing the loss function. Up to three interactions were allowed between predictors, and a shrinkage coefficient of 0.005 was used, as determined via grid search optimization of root mean squared error. Additionally, 10 cross-validation folds were specified. The cross-validation folds evaluate the

performance of the GBM model by partitioning the data into 10 subsets. The model is then trained 10 times, each time using nine of the subsets for training and the remaining subset for validation. This process is repeated so that each subset is used exactly once as the validation set.

In the construction of a fairness-aware model, there are three distinct components that work together to systematically reduce bias in a model while at the same time striving to preserve accuracy. The three components are data preprocessing, model training in-processing, and prediction adjustment postprocessing. In the preprocessing step, the dataset is prepared in a way that provides a training dataset that is less likely to be influenced by bias. This process adjusts the dataset to mitigate the effect of biases inherent in the data and ensures that all groups have similar representation in the training process. In the present implementation of this study, reweighting was used to ensure that underrepresented groups contribute proportionately to the training process. During in-process model adjustment, fairness constraints (e.g., adversarial debiasing and equalized odds constraints) are implemented in the model to penalize bias directly during training. This embeds fairness directly into the optimization process, ensuring that the model learns both to predict accurately and to minimize bias. In the postprocessing prediction adjustments, the model's predictions are corrected to achieve fairness after training. Adjusting decision thresholds for different groups to ensure fairness without retraining the model can ensure that the predictions meet fairness criteria and can correct for residual biases that are not addressed during training.

The current study utilized reweighting during the preprocessing phase of model building. During the model in-processing phase, I evaluated the appropriate number of trees needed to train the model as well as the appropriate shrinkage parameter. In this study, 1,000 trees were used, and a shrinkage of 0.005 was implemented along with 10 cross-fold validations. In model postprocessing, I explored the possibility of utilizing distinct thresholds for probabilities of attaining a C or higher by each demographic group (i.e., race/ethnicity, gender, and family income). Additionally, I implemented equalized odds postprocessing to adjust the predictions to ensure that the model's error rates were similar across different demographic groups.

Ultimately I decided that the postprocessing techniques would require manipulation based on demographic status. This presented a potential problem in both explaining the model to the public and, specifically with race/ethnicity, complying with the Supreme Court ruling of not utilizing race/ethnicity in the admissions process. Additionally, it would be difficult to explain or justify having different cutoffs for male vs. female or low-income students. For this reason, I used the same threshold for probabilities of attaining a C or higher for each demographic group and made no postprocessing adjustment directly to the probabilities after training the model.

The postprocessing phase involves adjustments for different groups to ensure fairness, even if those groups are not explicitly included in the model. This means that while the model itself does not use race/ethnicity as a predictor, we can still evaluate how the model's predictions affect different racial/ethnic subgroups. This helps us identify and mitigate any potential biases that may arise from the model's predictions.

## Results

### Descriptive Statistics

Table 2 provides the descriptive statistics for the analytical sample, the training sample, and the test sample. The analytical sample was divided into a training sample and a test sample for use in the models. Seventy percent of the sample was randomly selected for the training dataset, and the remaining 30% was used as a test dataset. Across the analytical sample, the percentages of students in each demographic group, as well as the means and standard deviations for continuous variables, were similar in both the training and test datasets. For example, in all three datasets, approximately 68% to 69% of students were White, approximately 31% to 32% came from families with an income greater than \$100,000, and approximately 59% to 60% were female. Across datasets, the average FYGPA, ACTC score, and HSGPA were similar. Additionally, the percentage of the school's attendees meeting poverty guidelines, the number of AP courses offered at the school, and the percentage of White students at the school were similar.

**Table 2.** Descriptive Statistics for the Analytical Sample

| Demographic characteristic                                |   | Sample        | Training      | Test          |
|---|---|---------------|---------------|---------------|
| <b>N</b>  |   | 4,711         | 3,298         | 1,413         |
| <b>Race/ethnicity</b><br><i>n (%)</i>                     | African American                                      | 618 (13.1)    | 445 (13.5)    | 173 (12.2)    |
|   | American Indian / Native Hawaiian / two or more races | 252 (5.3)     | 183 (5.5)     | 69 (4.9)      |
|   | Asian   | 127 (2.7)     | 91 (2.8)      | 36 (2.5)      |
|   | Hispanic  | 491 (10.4)    | 334 (10.1)    | 157 (11.1)    |
|   | White   | 3,223 (68.4)  | 2,245 (68.1)  | 978 (69.2)    |
|   | <b>Family income</b><br><i>n (%)</i>                  | <\$36K        | 1097 (23.3)   | 772 (23.4)    |
| \$36K–\$60K   |   | 894 (19.0)    | 624 (18.9)    | 270 (19.1)    |
| \$60K–\$100K  |   | 1222 (25.9)   | 848 (25.7)    | 374 (26.5)    |
| >\$100K   |   | 1498 (31.8)   | 1054 (32.0)   | 444 (31.4)    |
| <b>Gender</b><br><i>n (%)</i>                             | Female  | 2,812 (59.7)  | 1,986 (60.2)  | 826 (58.5)    |
|   | Male  | 1,899 (40.3)  | 1,312 (39.8)  | 587 (41.5)    |
| <b>FYGPA</b><br>Mean ( <i>SD</i> )                        |   | 2.93 (1.02)   | 2.93 (1.02)   | 2.92 (1.02)   |
| <b>ACTC score</b><br>Mean ( <i>SD</i> )                   |   | 22.03 (4.94)  | 22.06 (4.99)  | 21.94 (4.82)  |
| <b>HSGPA</b><br>Mean ( <i>SD</i> )                        |   | 3.59 (0.42)   | 3.59 (0.42)   | 3.59 (0.42)   |
| <b>% poverty</b><br>Mean ( <i>SD</i> )                    |   | 18.47 (6.94)  | 18.48 (6.96)  | 18.44 (6.91)  |
| <b>Number of AP courses offered</b><br>Mean ( <i>SD</i> ) |   | 3.43 (0.95)   | 3.43 (0.95)   | 3.41 (0.95)   |
| <b>% White</b><br>Mean ( <i>SD</i> )                      |   | 67.23 (24.28) | 66.71 (24.35) | 68.46 (24.07) |

## To what extent do traditional regression models exhibit bias for student subgroups?

### **Gender**

For gender, the reference group is male students. As seen in [Figure 1](#), for Models 1 to 3, all five of the fairness statistics (the accuracy equality ratio, equal opportunity ratio, predictive equality ratio, predictive parity ratio, and statistical parity ratio) fell within an acceptable range for female versus male students. This means that for a model that incorporates HSGPA as its only academic achievement indicator, a model that incorporates ACTC score as its only academic achievement indicator, and a model that incorporates both HSGPA and ACTC score and their interaction, there is no evidence of substantial bias in the predictions from any of the models by gender.

### **Race/Ethnicity**

For race/ethnicity, the reference group is White students. [Figure 2](#) displays the fairness metrics for Model 1 to Model 3 by race/ethnicity. There was no evidence of substantial bias in the equal opportunity ratio and predictive parity ratio. The equal opportunity ratio showed acceptable differences between student subgroups. This demonstrates that the true positive rate is similar between White students and students in other racial/ethnic categories. The predictive parity ratio across all student subgroups also displayed an acceptable difference within the 80% threshold. This indicates that the positive predictive values, or precision, were relatively similar between White students and students in other racial/ethnic groups.

Where concern for bias was found, it was between White and African American students. Comparing African American to White students, the HSGPA model and ACTC score model both suggest potential bias in the accuracy equality ratio, while the HSGPA and ACTC score model is just within the threshold of acceptable differences. This indicates that relative to predictions for White students, the GPA predictions for African American students are less accurate for the HSGPA model and ACTC score model.

The predictive equality ratio for traditional logistic regressions shows that the HSGPA model and HSGPA and ACTC score model exceed the threshold for substantial disadvantage for African American students compared to White students. The model that includes only HSGPA as well as the model that includes HSGPA plus ACTC score both raised concerns regarding a substantial disadvantage in the rate of identifying false positives, suggesting White students might exhibit more false positives than African American students (e.g., incorrectly identifying students as attaining a C average when they do not).

For the statistical parity ratio, only the HSGPA model displayed potential bias between African American students and White students, although it barely exceeded the threshold.

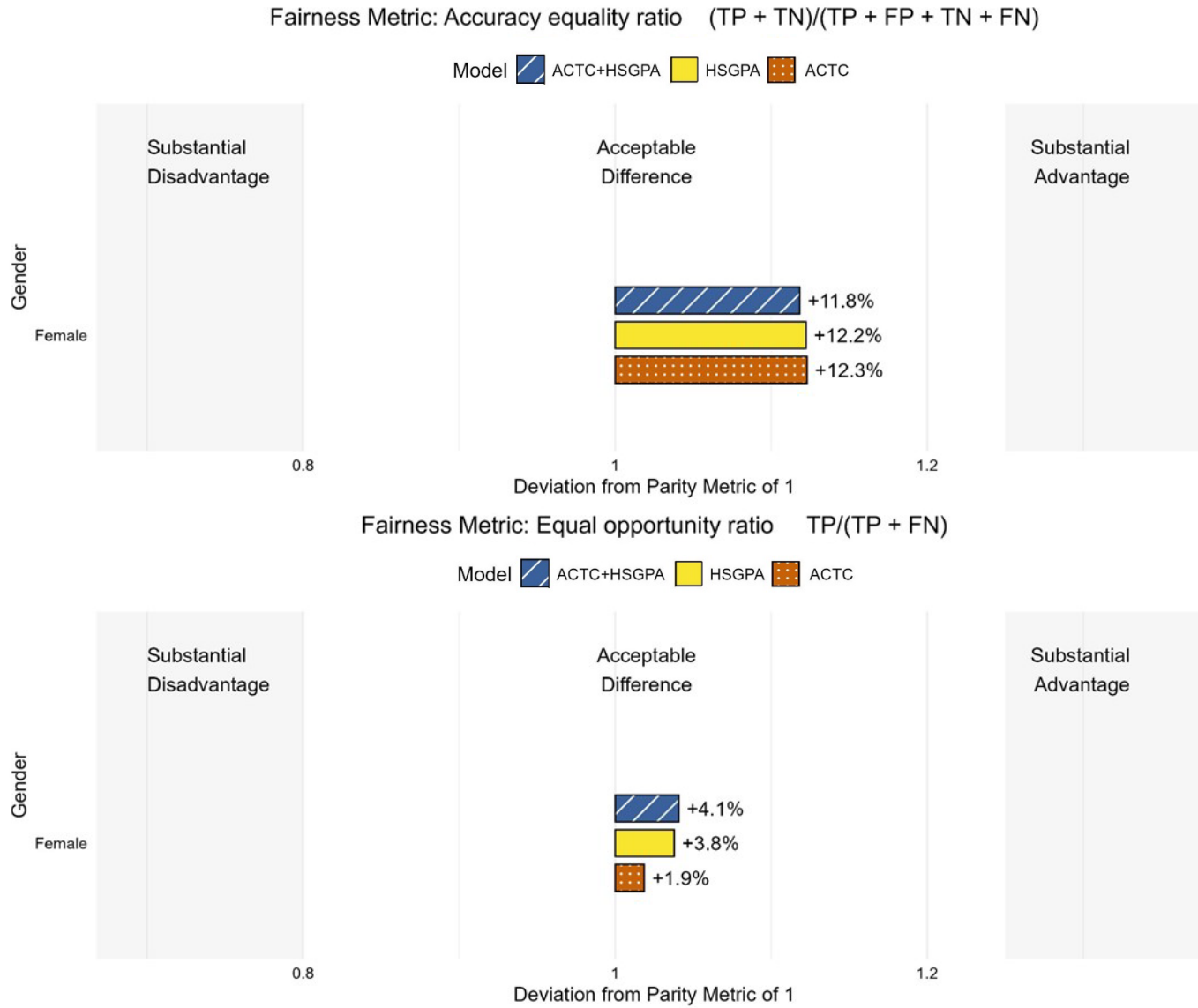
Three of these metrics—the equal opportunity ratio, predictive equality ratio, and statistical parity ratio—suggested that the ACTC score model did notably better than either the HSGPA model or the HSGPA and ACTC score model in terms of fairness for African American students compared to White students. However, the accuracy equality ratio and the predictive parity ratio

did not suggest less bias for African American students in the ACTC score model than in the HSGPA and the HSGPA plus ACTC score models.

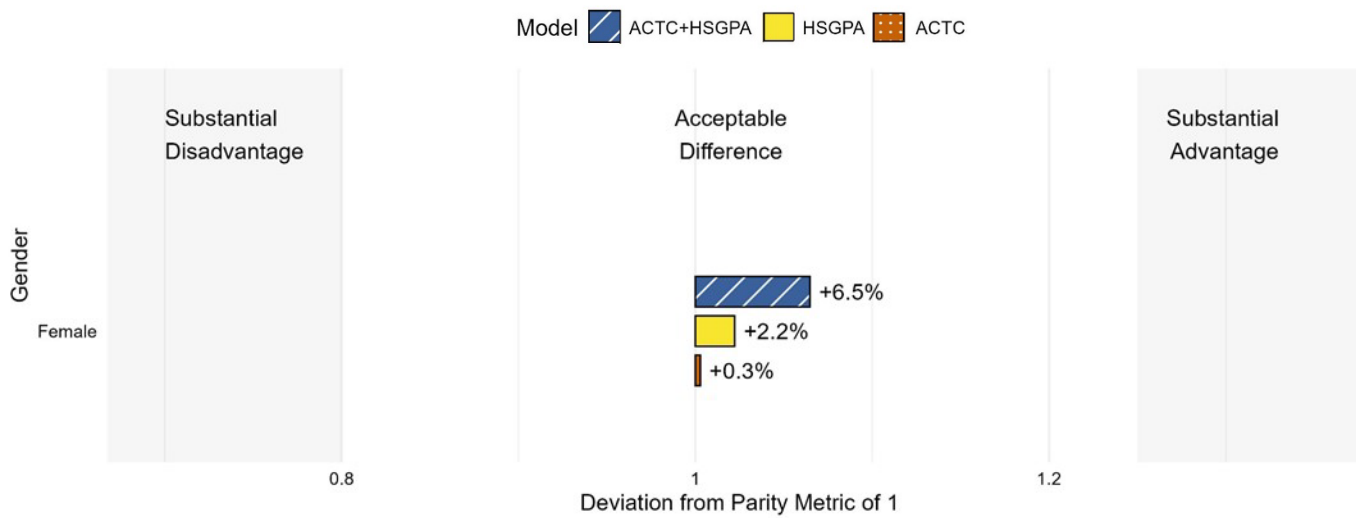
### ***Family Income***

For family income, the reference group is students from families with over \$100,000 family income. Among family income levels, the accuracy equality ratio, equal opportunity ratio, predictive parity ratio, and statistical parity ratio did not suggest bias between family income levels ([Figure 3](#)). The predictive equality ratio, however, did suggest substantial disadvantage for students from families with incomes of less than \$36,000 in the HSGPA model and the HSGPA and ACTC score model. This means that when it comes to modeling the probability of attaining a C or higher FYGPA in the HSGPA model and the HSGPA plus ACTC score model, students from family incomes of less than \$36,000 are substantially less likely to display false positives than students from families with incomes of greater than \$100,000.

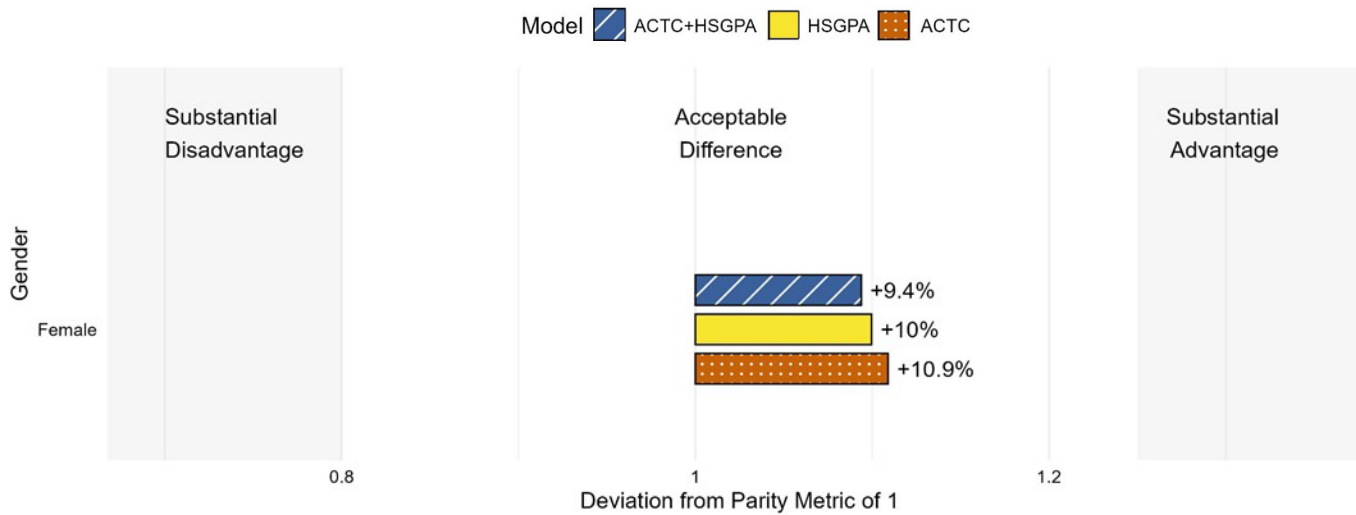
**Figure 1. Fairness Metrics for Models 1–3 by Gender**



Fairness Metric: Predictive equality ratio  $FP/(FP + TN)$

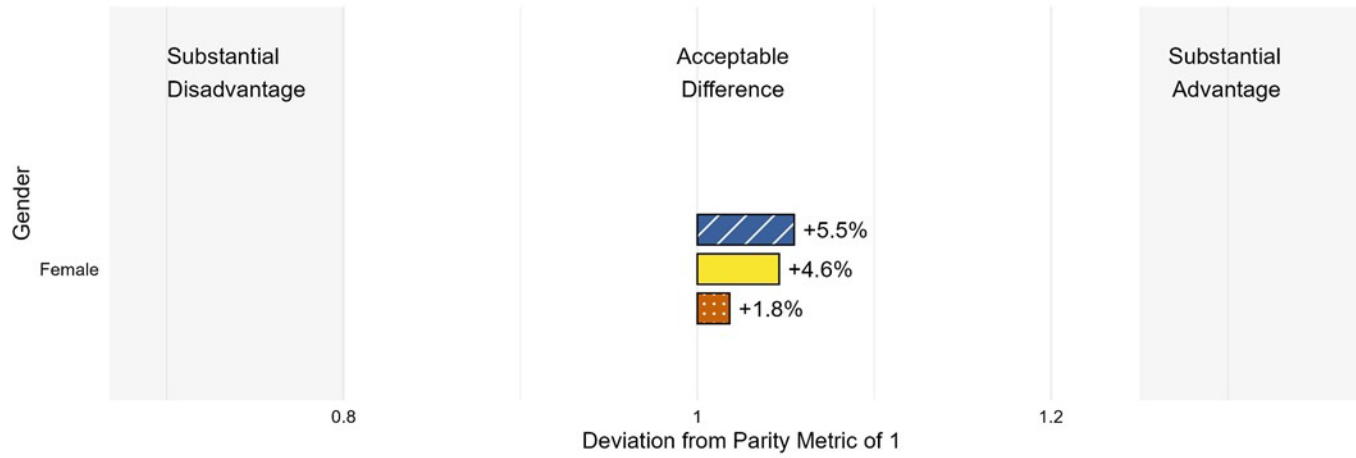


Fairness Metric: Predictive parity ratio  $TP/(TP + FP)$

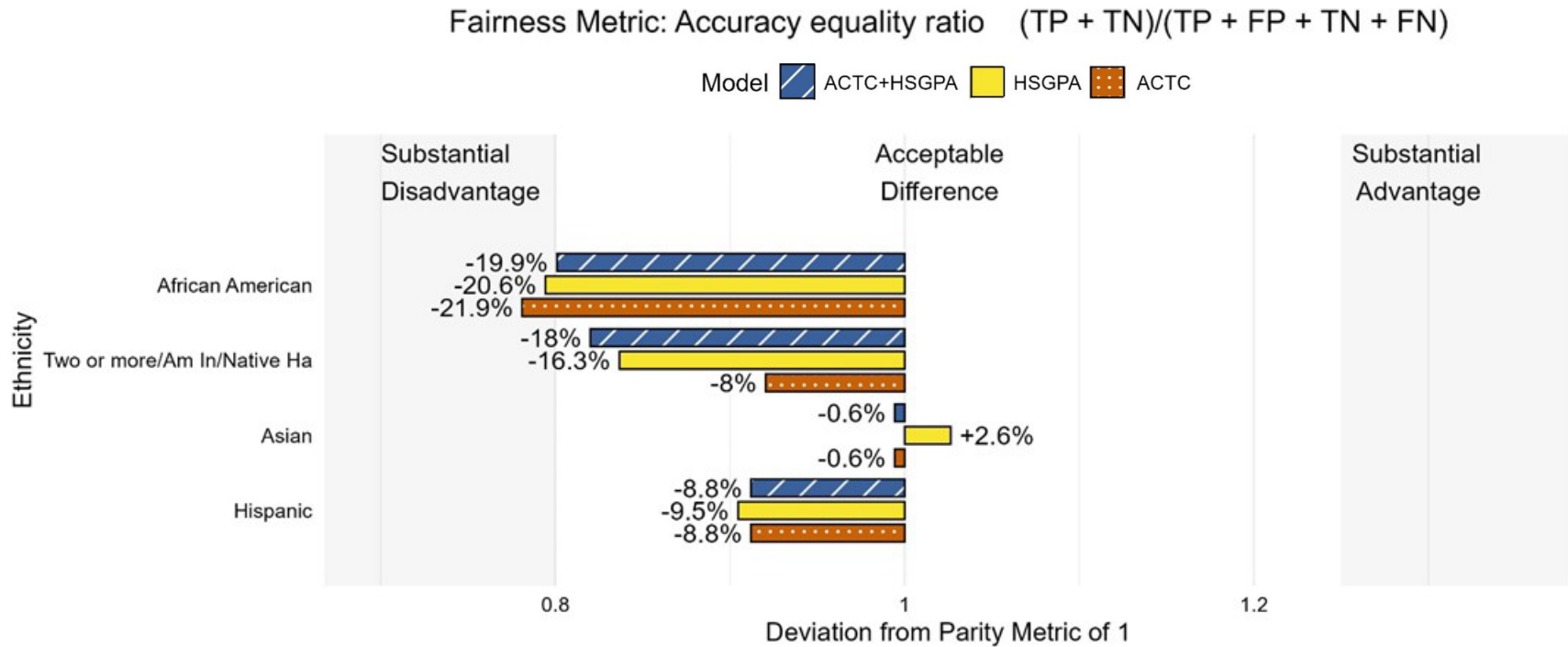


Fairness Metric: Statistical parity ratio  $(TP + FP)/(TP + FP + TN + FN)$

Model  ACTC+HSGPA  HSGPA  ACTC

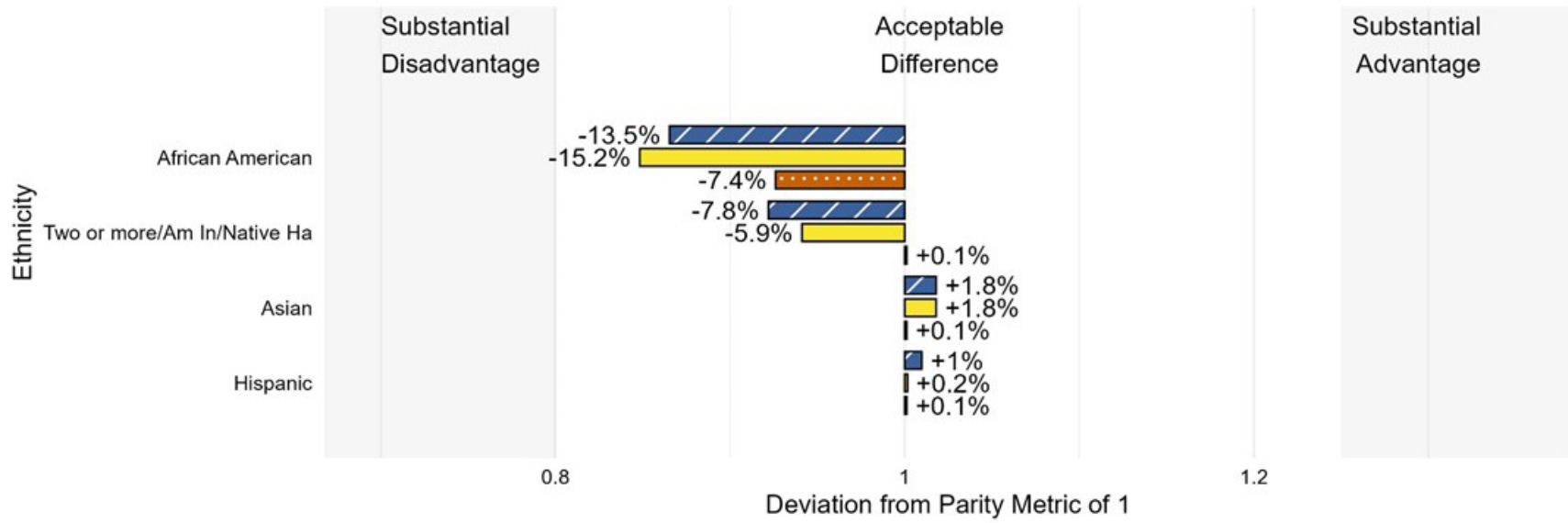


**Figure 2.** Fairness Metrics for Models 1–3 by Race/Ethnicity

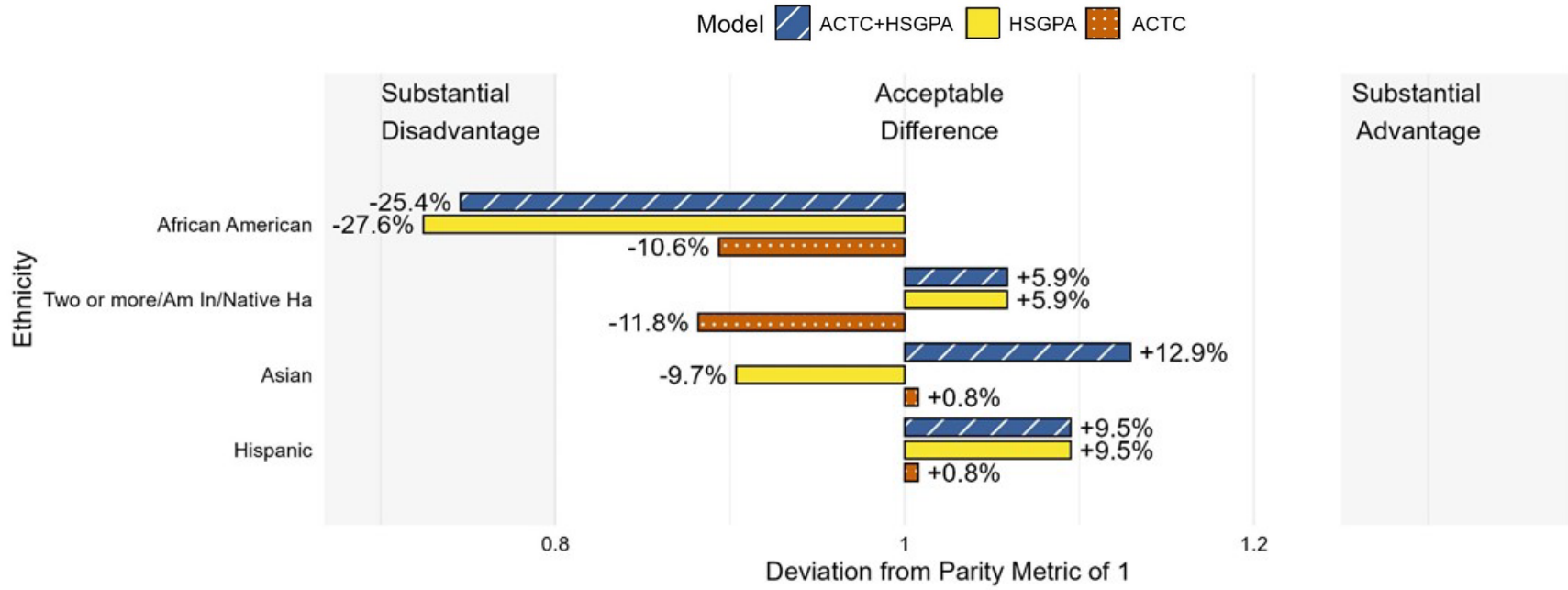


Fairness Metric: Equal opportunity ratio  $TP/(TP + FN)$

Model  ACTC+HSGPA  HSGPA  ACTC

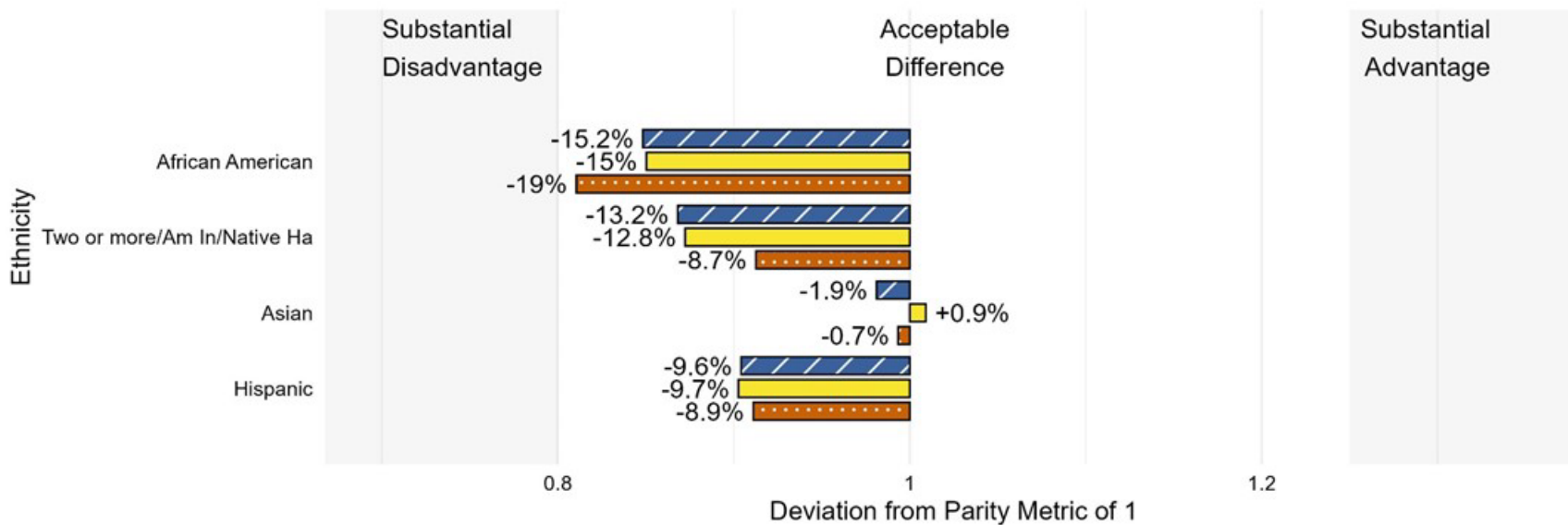


Fairness Metric: Predictive equality ratio  $FP/(FP + TN)$

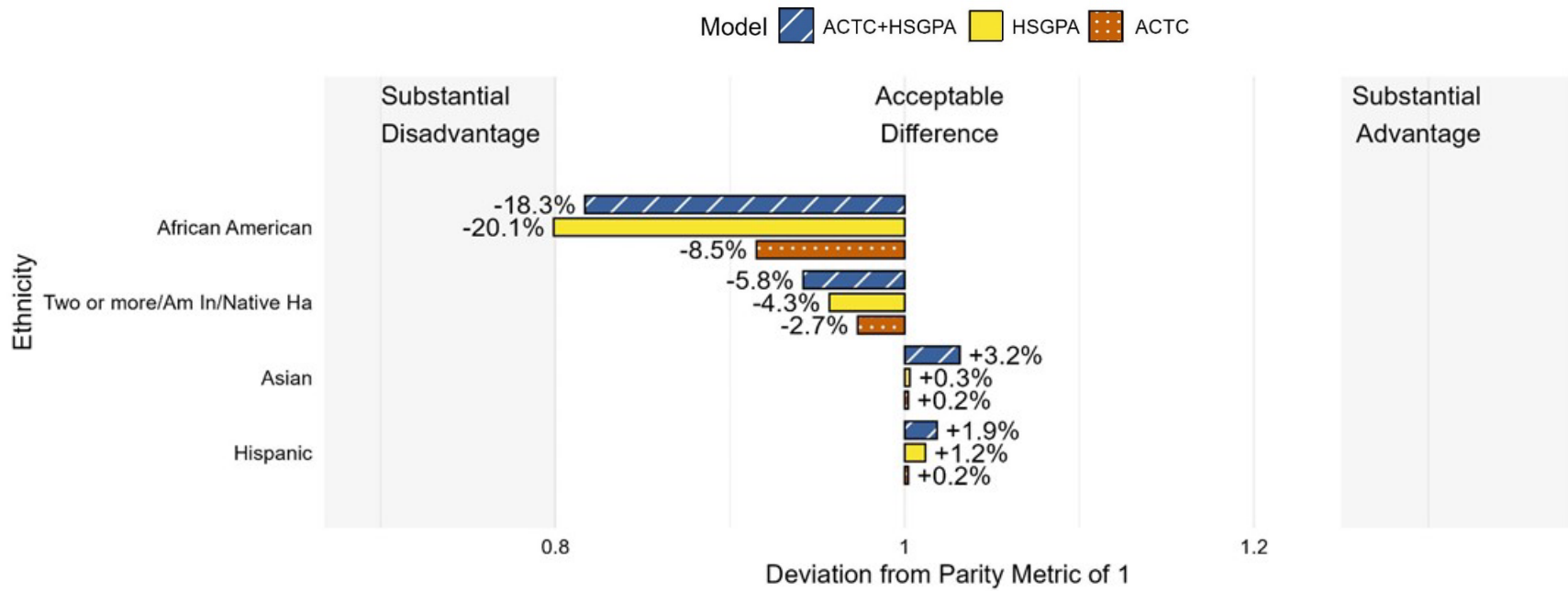


Fairness Metric: Predictive parity ratio  $TP/(TP + FP)$

Model ▨ ACTC+HSGPA ▨ HSGPA ▨ ACTC

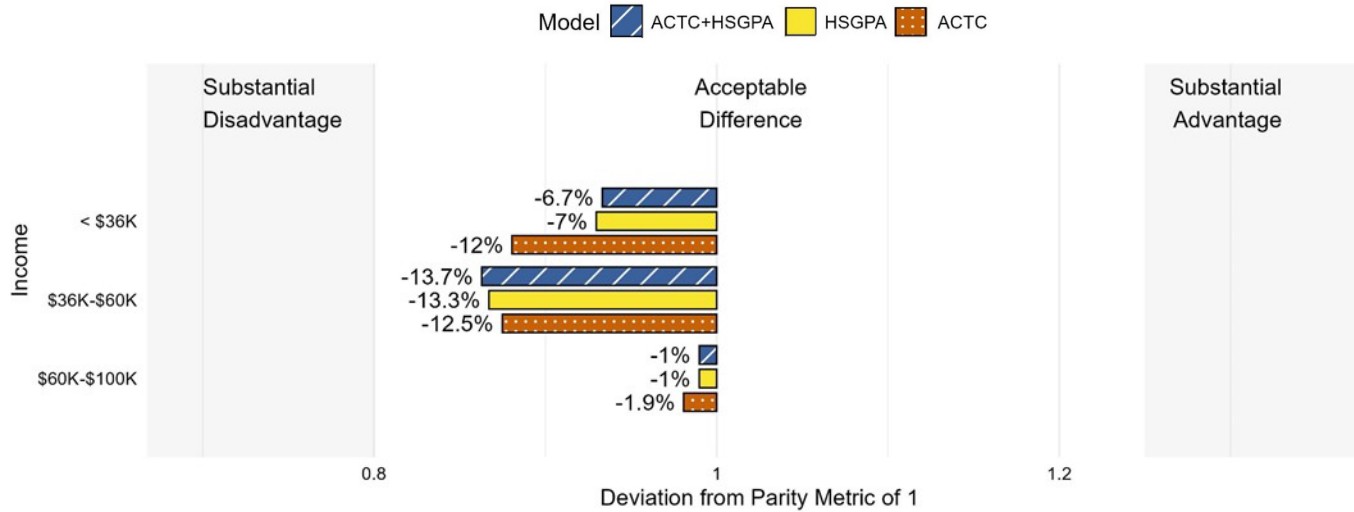


Fairness Metric: Statistical parity ratio  $(TP + FP)/(TP + FP + TN + FN)$

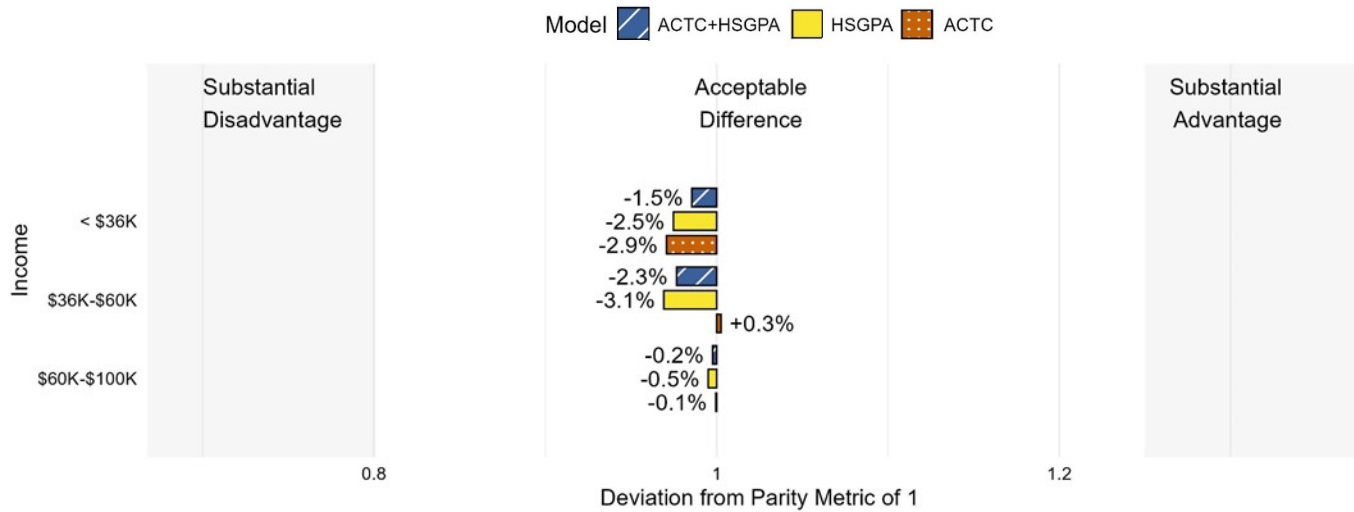


**Figure 3. Fairness Metrics for Models 1–3 by Family Income**

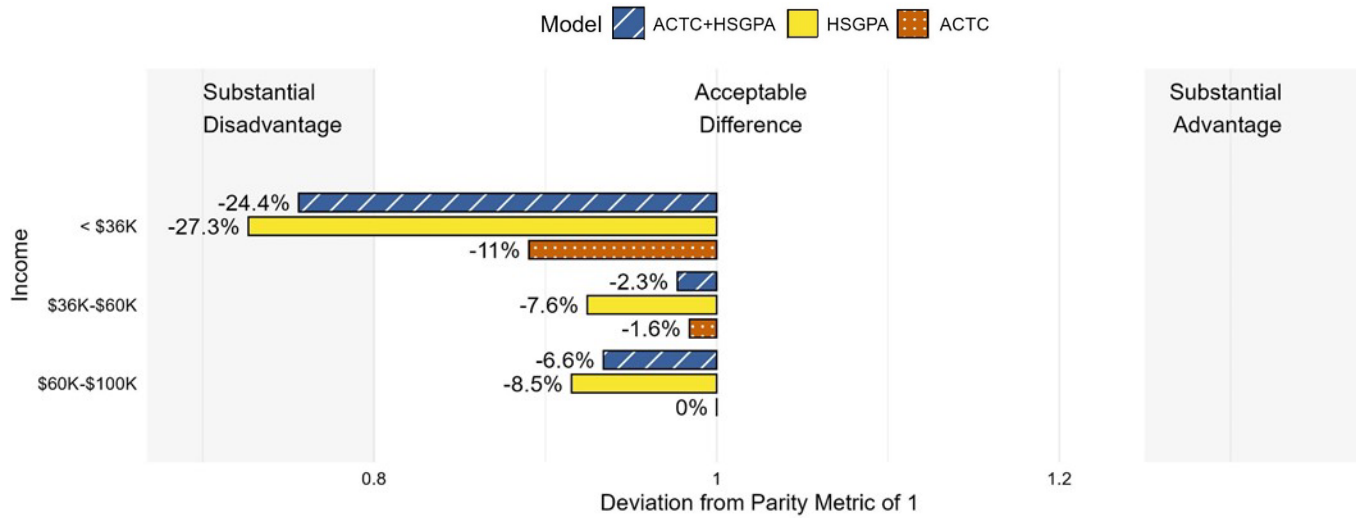
Fairness Metric: Accuracy equality ratio  $(TP + TN)/(TP + FP + TN + FN)$



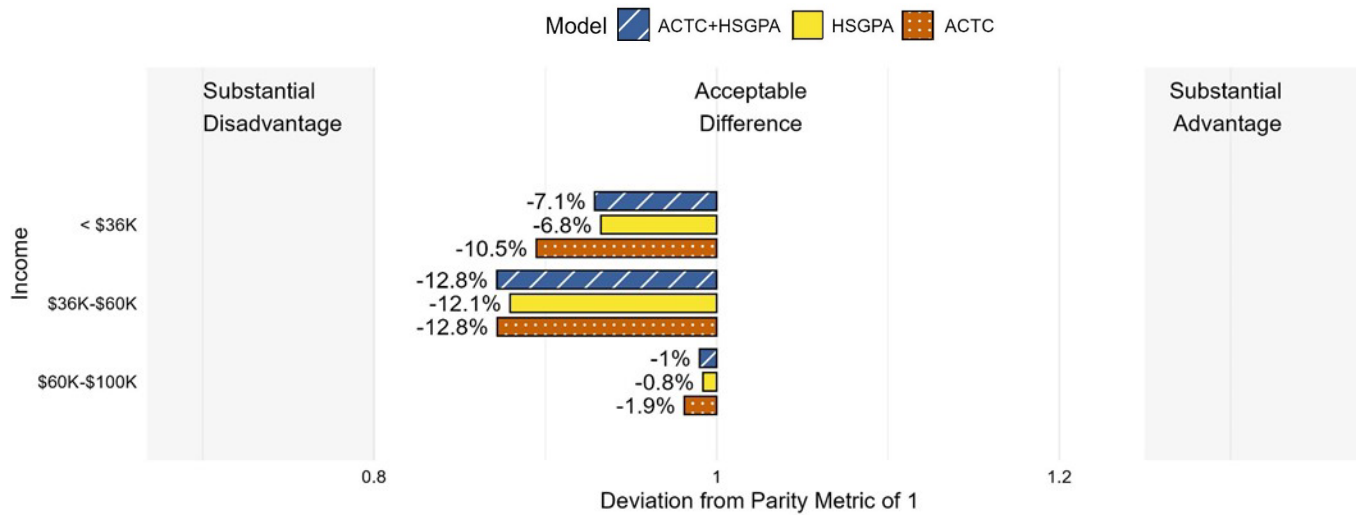
Fairness Metric: Equal opportunity ratio  $TP/(TP + FN)$



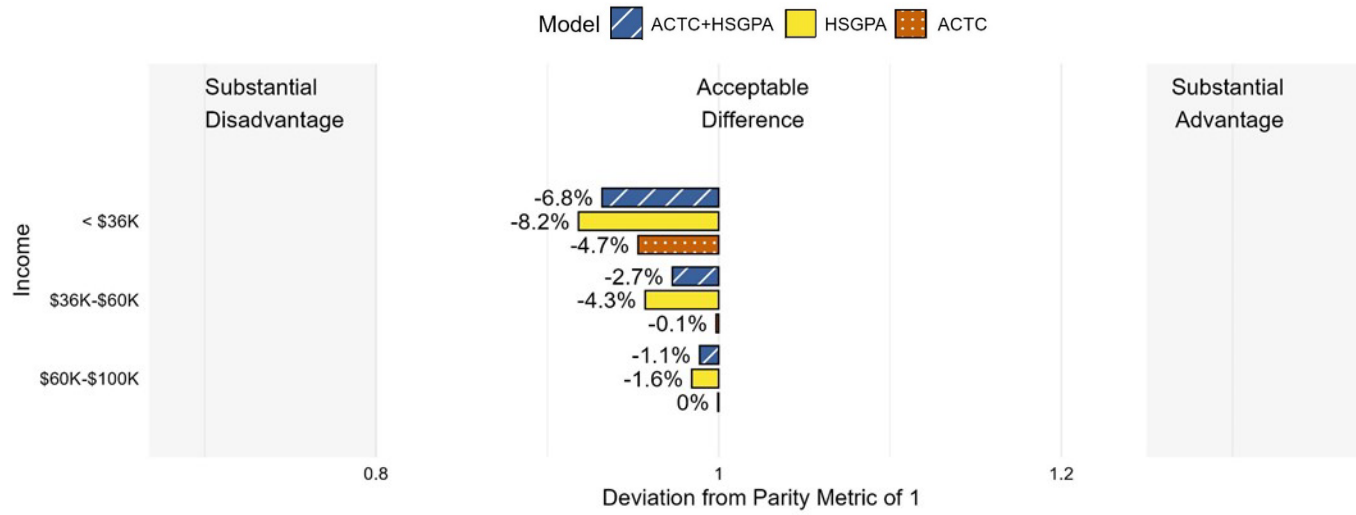
Fairness Metric: Predictive equality ratio  $FP/(FP + TN)$



Fairness Metric: Predictive parity ratio  $TP/(TP + FP)$



Fairness Metric: Statistical parity ratio  $(TP + FP)/(TP + FP + TN + FN)$



## Summary of Traditional Logistic Regressions

There are four key takeaways from the fairness metrics of the traditional logistic regression models.

1. The ACTC score model is generally the fairest model across subgroups. For African American students compared to White students, the ACTC score model exhibits fewer instances of substantial disadvantage across fairness metrics. For family income subgroups, the ACTC score model avoids the substantial disadvantage in predictive equality observed in the two HSGPA-based models (the HSGPA model and the HSGPA plus ACTC score model).
2. The HSGPA model shows the most evidence of bias. This model raises concerns for African American students in several metrics as well as for students from low-income families in predictive equality.
3. The HSGPA and ACTC score model seems to strike a balance in fairness metrics but may exhibit slight bias. While this model performs better than the HSGPA model, it still displays some substantial disadvantages in fairness metrics for African American students and low-income students compared to White and higher income groups.
4. The ACTC score model appears to be a better choice when considering fairness across subgroups without sacrificing predictive accuracy. If the primary goal is to develop a predictive model of attainment of a C or better FYGPA while focusing on fairness, this model would be the best option.

## To what extent does a fairness-aware GBM model reduce prediction bias for student subgroups (e.g., by gender or family income) compared to traditional logistic models that either do or do not include race/ethnicity?

In this section, we compare the HSGPA plus ACTC score model to a HSGPA plus ACTC score model that does not include race/ethnicity, as well as a gradient-boosted machine model without race/ethnicity. This comparison in fairness metrics helps us understand the differences between traditional logistic models with and without race/ethnicity, as well the performance of a GBM model when it comes to meeting fairness goals.

### Gender

Between male and female students, all fairness metrics fell within the acceptable difference threshold, suggesting no evidence of bias by gender in a logistic regression model that includes race/ethnicity (HSGPA and ACTC score model), a logistic regression model with HSGPA and ACTC score that does not include race/ethnicity, and a gradient-boosted machine model (Model 5; [Figure 4](#)).

## ***Race/Ethnicity***

Overall, most indices demonstrated no significant bias across race/ethnicity. Specifically, the equal opportunity ratio, predictive parity ratio, and statistical parity ratio all showed acceptable differences for all racial/ethnic subgroups.

The only evidence of bias was found for African American students in the accuracy equality ratio and the predictive equality ratio. For African American students versus White students, the GBM model exceeded the threshold for substantial disadvantage in the accuracy equality ratio. The logistic regression models, both with and without race/ethnicity, were at the threshold of substantial disadvantage for African American students.

In terms of the predictive equality ratio, both the HSGPA plus ACTC score model with race/ethnicity and the GBM model demonstrated a substantial disadvantage for African American students relative to White students. In contrast, the logistic regression model without race/ethnicity showed acceptable differences between African American and White students. All other races demonstrated acceptable differences in the predictive equality ratio.

It is noteworthy that for the statistical parity ratio, both the GBM model and the logistic regression model with race/ethnicity approached the threshold of substantial disadvantage for African American students versus White students.

## ***Family Income***

Among family income levels, the accuracy equality ratio, equal opportunity ratio, predictive parity ratio, and statistical parity ratio did not suggest bias toward certain family income levels ([Figure 6](#)). The predictive equality ratio, however, did suggest substantial disadvantage for students from families with incomes of less than \$36,000 in all three models. This means that when it comes to modeling the probability of attaining a C or higher FYGPA, students from family incomes of less than \$36,000 are substantially less likely to display false positives than students from families with an income of greater than \$100,000.

## ***Summary of Logistic Regressions Versus Gradient-Boosted Machine Model***

For race/ethnicity, in the three models, three out of five fairness metrics fell within acceptable thresholds. That said, logistic models both with and without race/ethnicity outperformed the GBM model in terms of fairness, with fewer deviations from the parity metric of 1. Across fairness metrics, the GBM model tended to show greater disparities for African American students than for White students when compared to both logistic models. The logistic models performed similarly, with only minor differences between the two. Both logistic regression models generally stayed within acceptable thresholds, while the GBM model exceeded the substantial disadvantage threshold particularly for the accuracy equality and predictive equality ratios. For other racial/ethnic groups, all models exhibited fairness metrics within the acceptable differences range. For students from low-income families, the GBM model exhibited disparities in fairness metrics for low-income students for the predictive equality ratio. The logistic regression models, particularly the logistic model without race/ethnicity, tended to demonstrate better performance, with deviations closer to the parity metric of 1 relative to the GBM model.




Students with both middle-lower and middle-upper incomes exhibited acceptable fairness metrics across models.

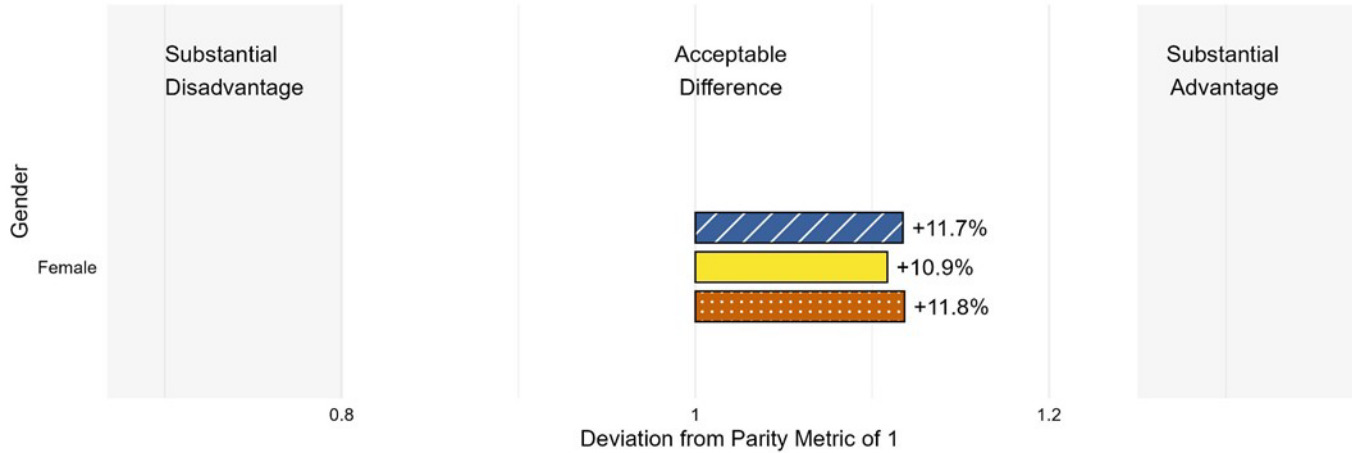
Key takeaways from the model comparisons include the following:

1. Generally speaking, the logistic regression models outperformed the GBM model in terms of fairness across student subgroups. The exclusion of race/ethnicity in the logistic regression model does not introduce notable bias but may marginally improve fairness for African American students.
2. The GBM model showed disparities for marginalized groups, particularly for African American students and low-income students. These potential biases suggest a need for further fairness-aware optimization or postprocessing adjustments. Recall that I decided not to implement individual group thresholds and direct manipulation of probabilities for subgroups, as these postprocessing techniques may be difficult to defend to postsecondary stakeholders.
3. In general, the logistic regression models achieve a better balance between fairness and predictive accuracy, often maintaining fairness metrics across subgroups while avoiding substantial disadvantage in fairness.

**Figure 4.** Fairness Metrics for Models 3–5 by Gender

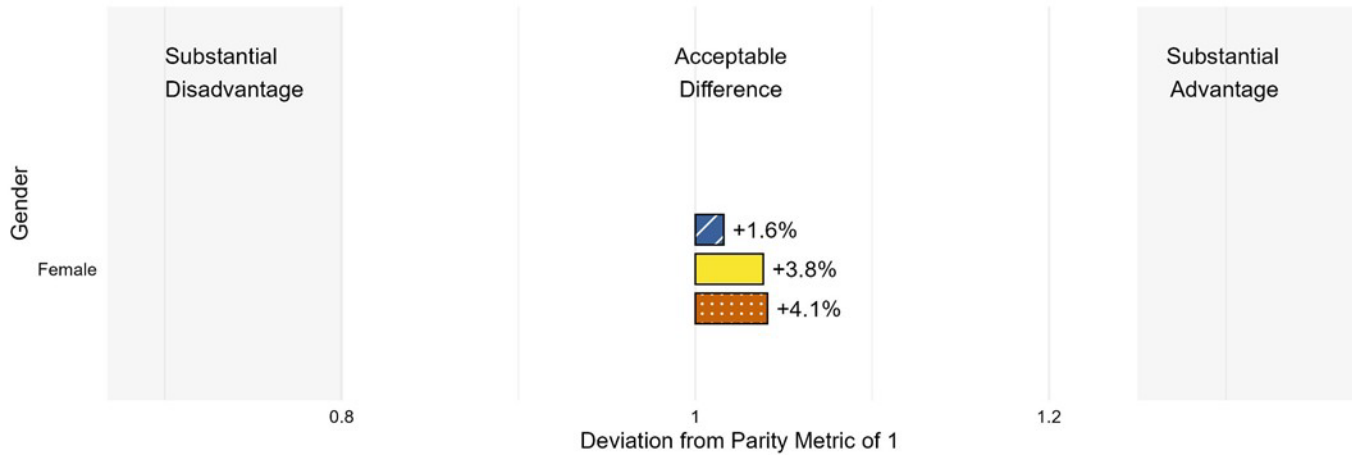
Fairness Metric: Accuracy equality ratio  $(TP + TN)/(TP + FP + TN + FN)$

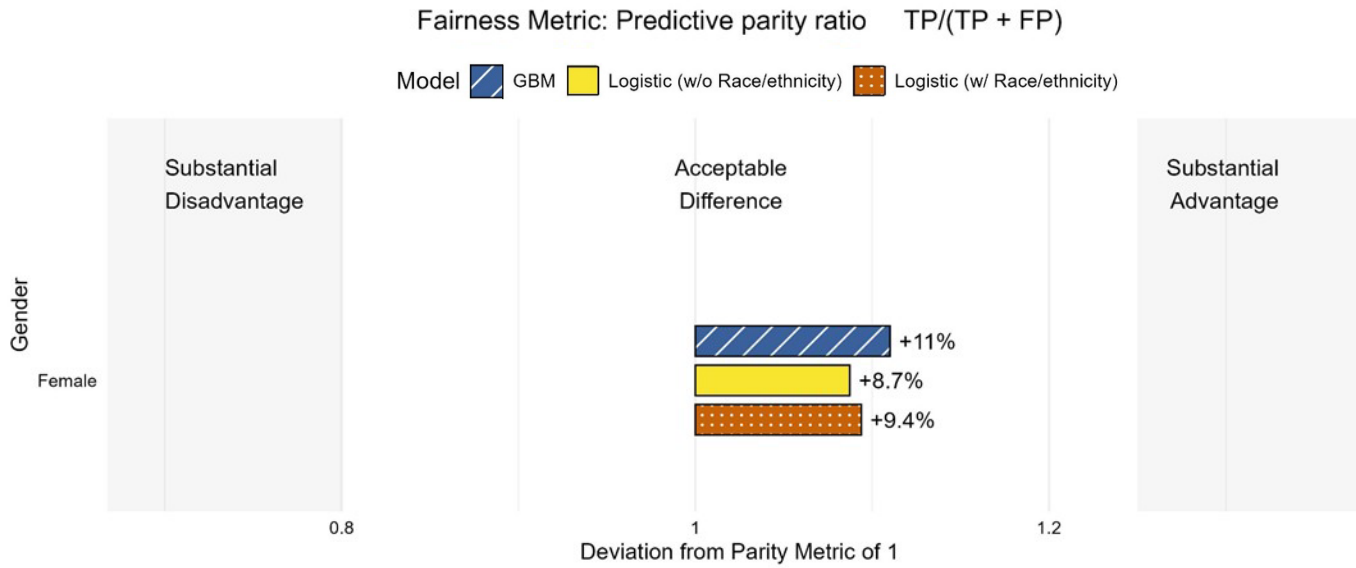
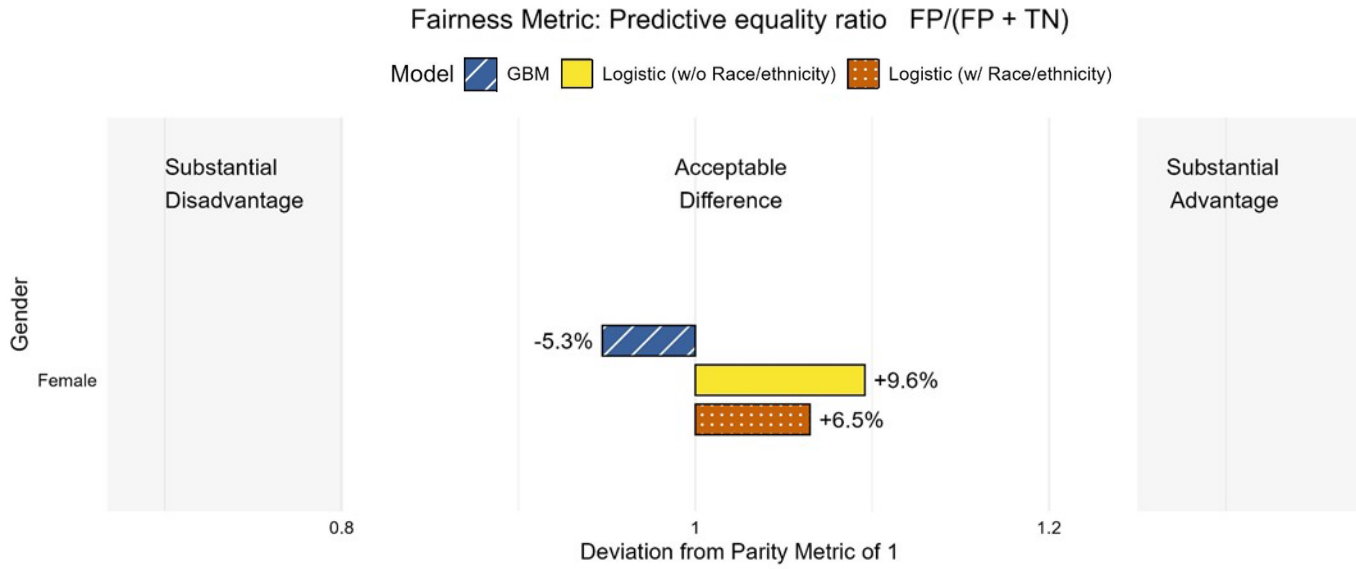
Model  GBM  Logistic (w/o Race/ethnicity)  Logistic (w/ Race/ethnicity)



Fairness Metric: Equal opportunity ratio  $TP/(TP + FN)$

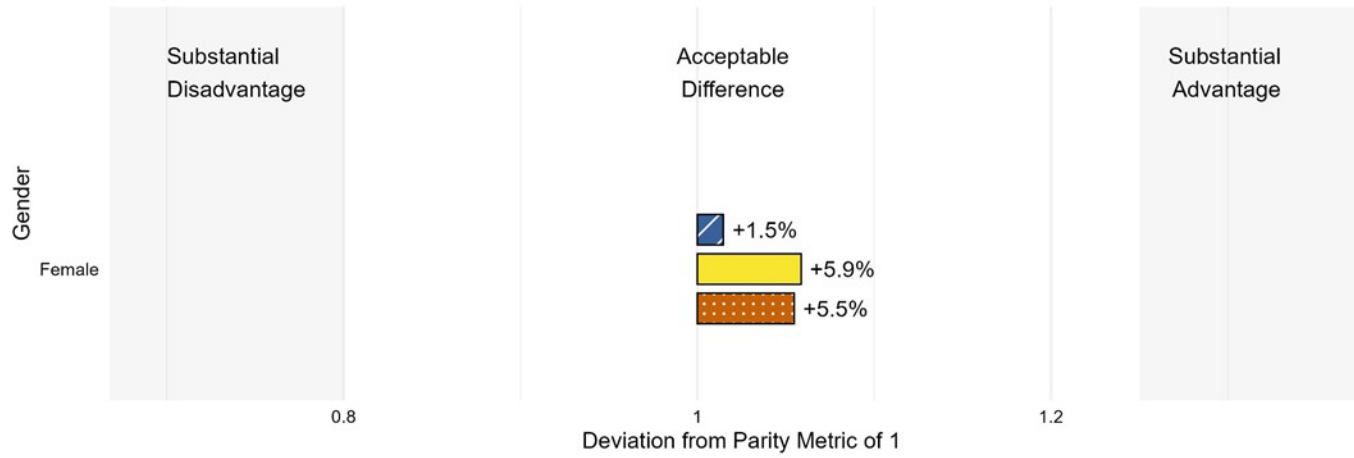
Model  GBM  Logistic (w/o Race/ethnicity)  Logistic (w/ Race/ethnicity)





Fairness Metric: Statistical parity ratio  $(TP + FP)/(TP + FP + TN + FN)$

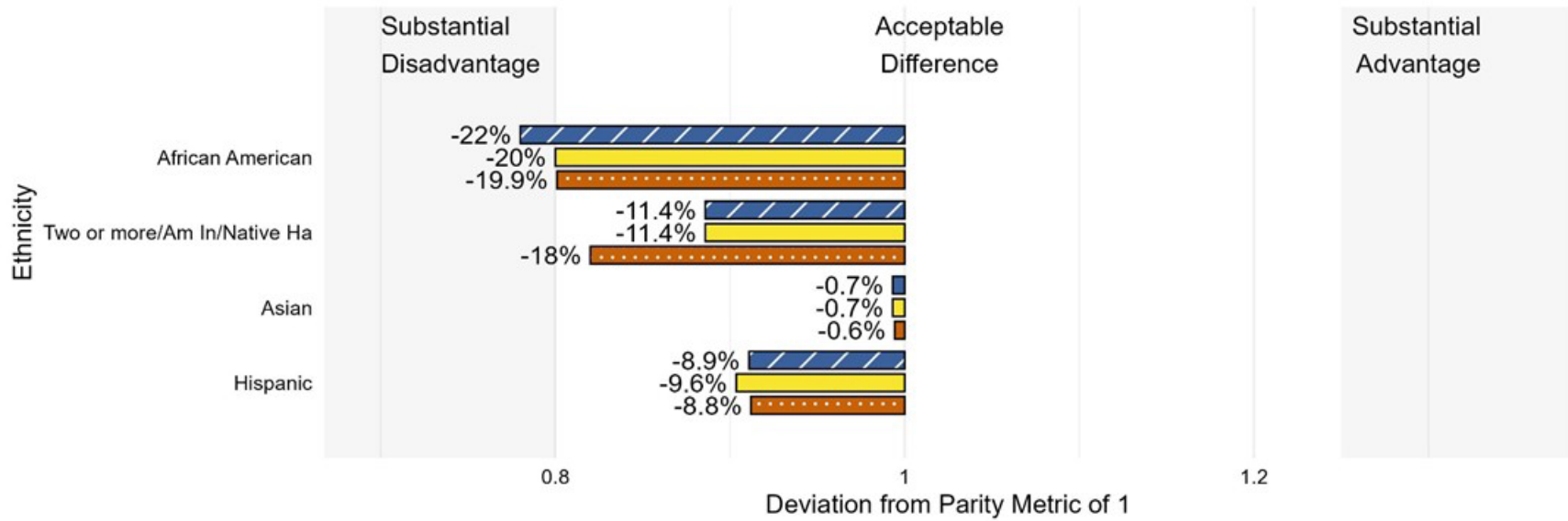
Model  GBM  Logistic (w/o Race/ethnicity)  Logistic (w/ Race/ethnicity)

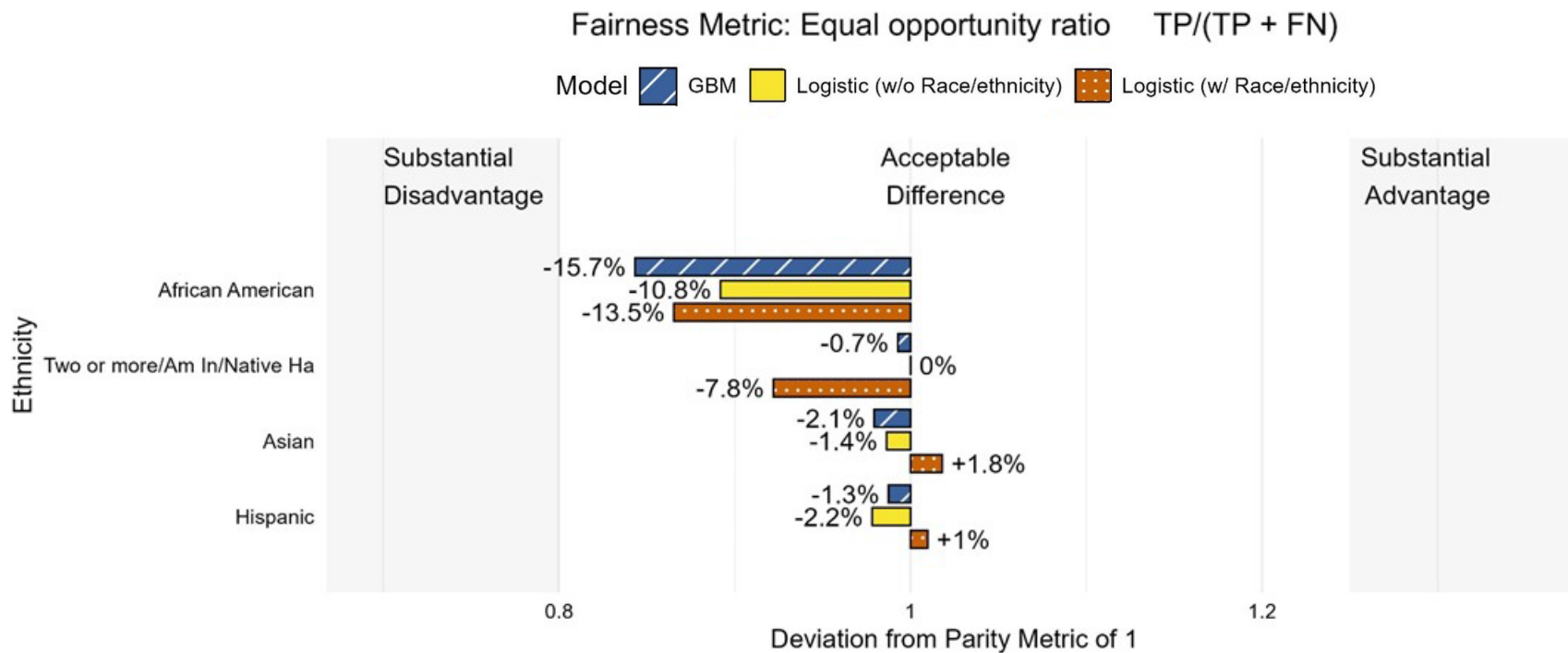


**Figure 5.** Fairness Metrics for Models 3–5 by Race/Ethnicity

Fairness Metric: Accuracy equality ratio  $(TP + TN)/(TP + FP + TN + FN)$

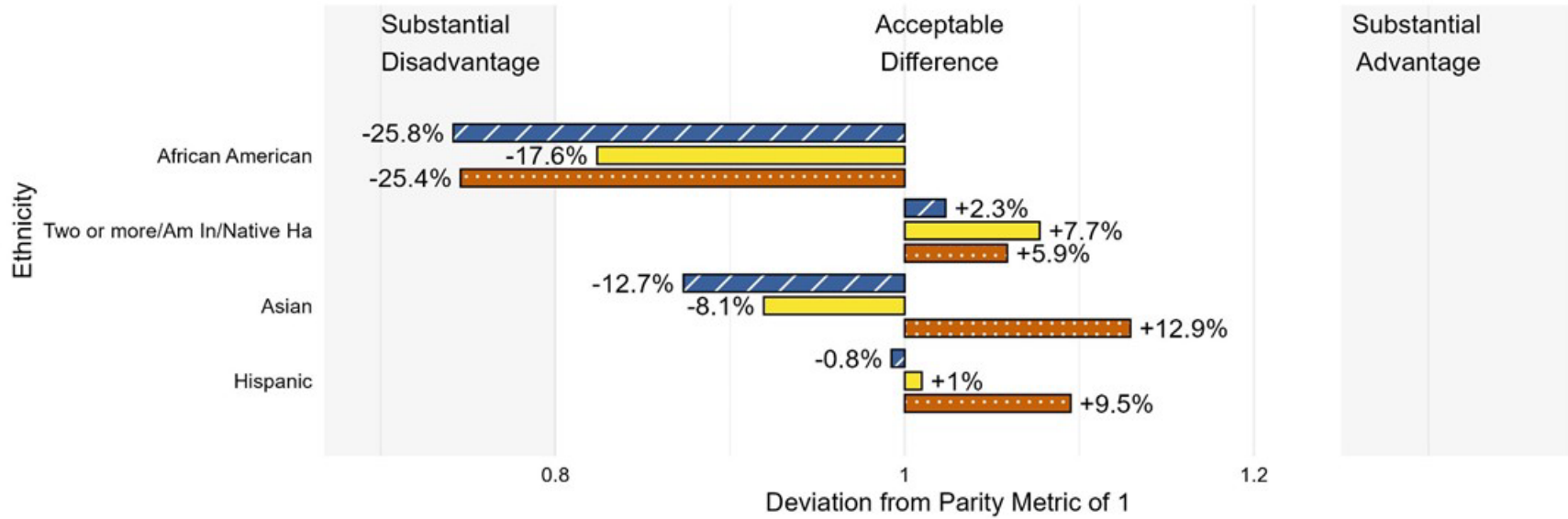
Model ▨ GBM ▨ Logistic (w/o Race/ethnicity) ▨ Logistic (w/ Race/ethnicity)

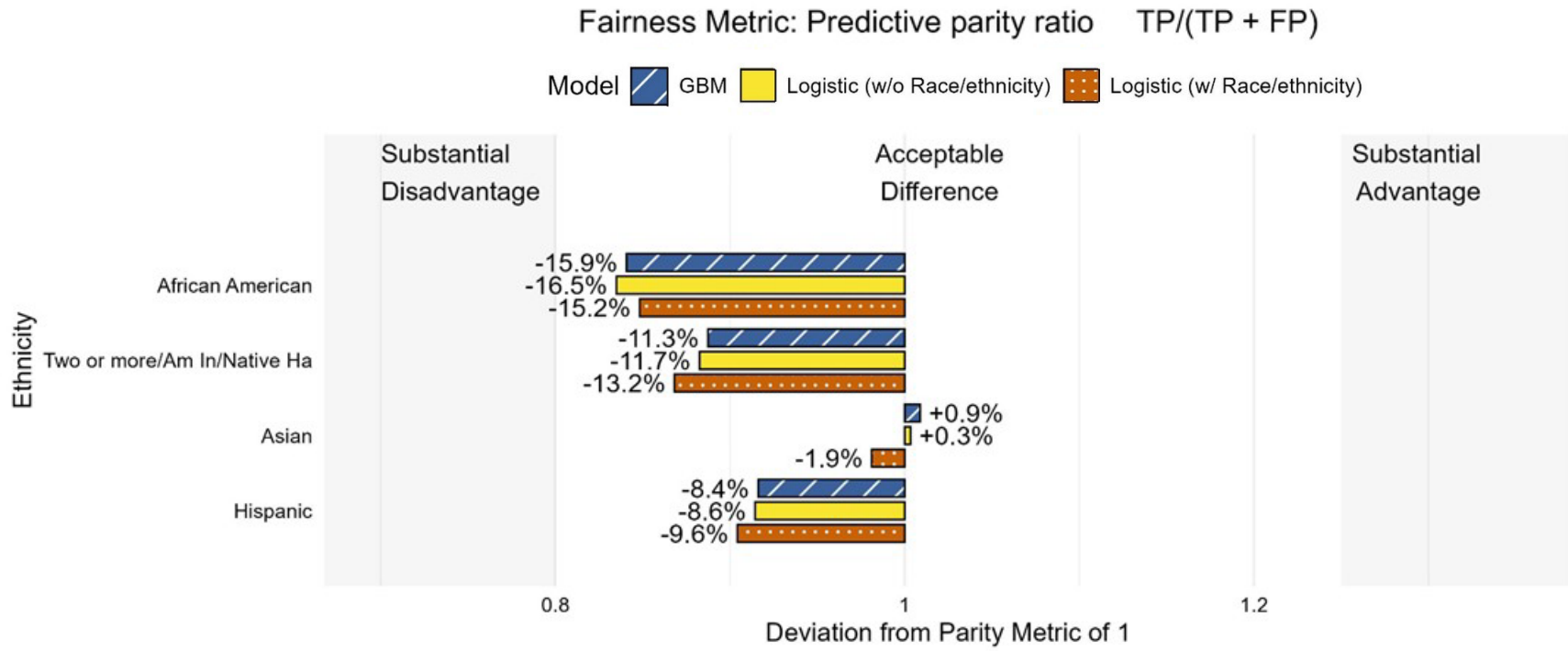




Fairness Metric: Predictive equality ratio  $FP/(FP + TN)$

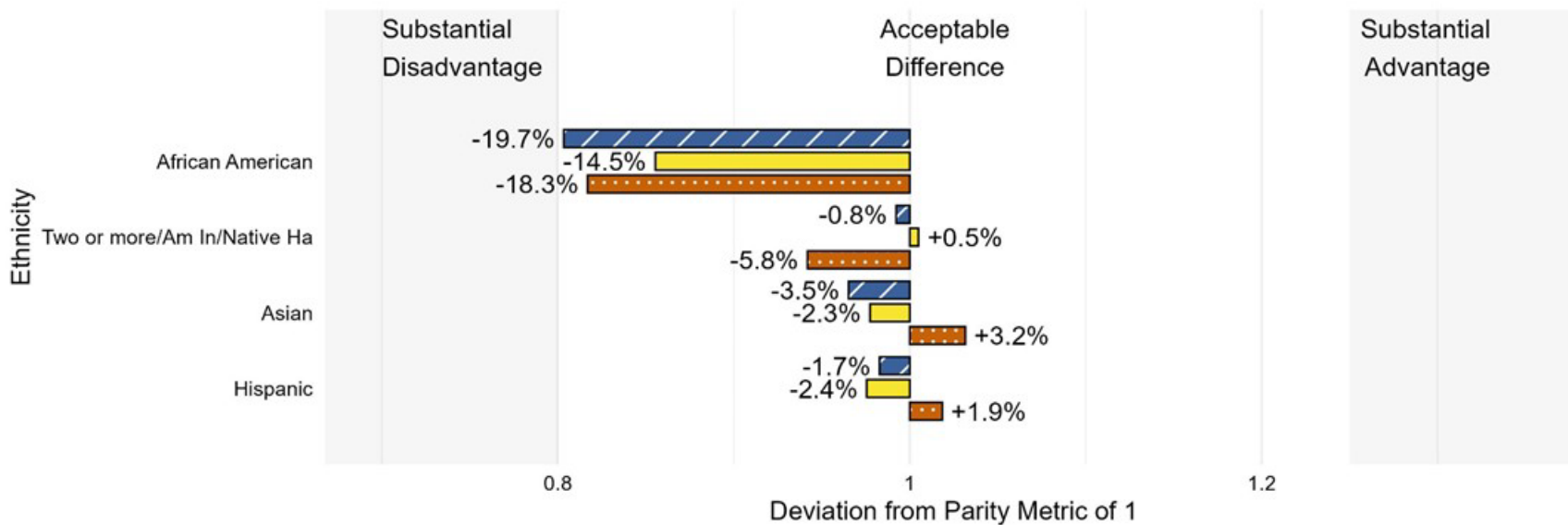
Model ▨ GBM ▨ Logistic (w/o Race/ethnicity) ▨ Logistic (w/ Race/ethnicity)








Fairness Metric: Statistical parity ratio  $(TP + FP)/(TP + FP + TN + FN)$

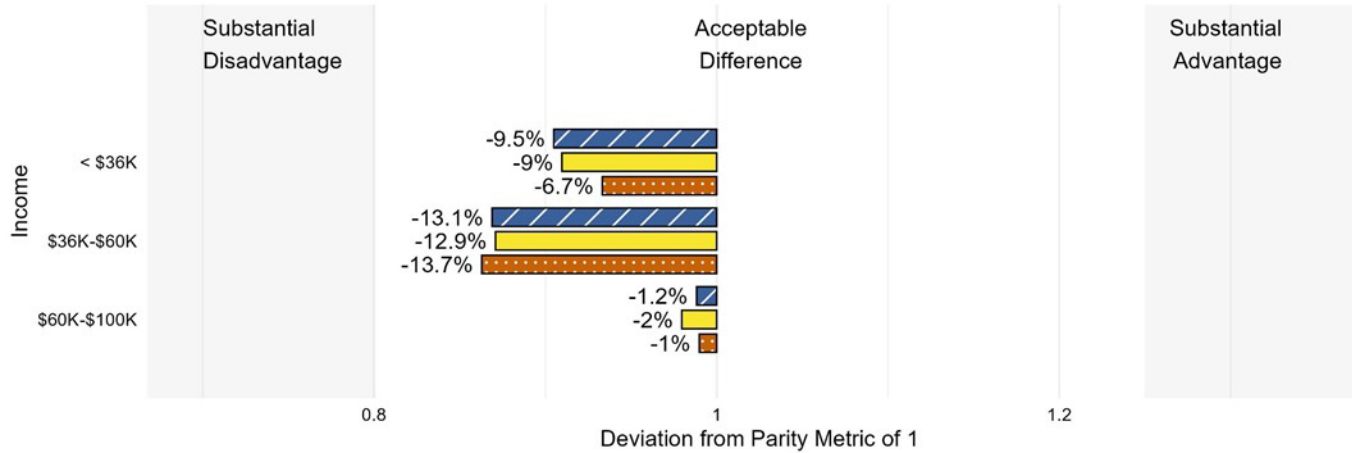
Model ▨ GBM ▨ Logistic (w/o Race/ethnicity) ▨ Logistic (w/ Race/ethnicity)



**Figure 6.** Fairness Metrics for Models 3–5 by Family Income

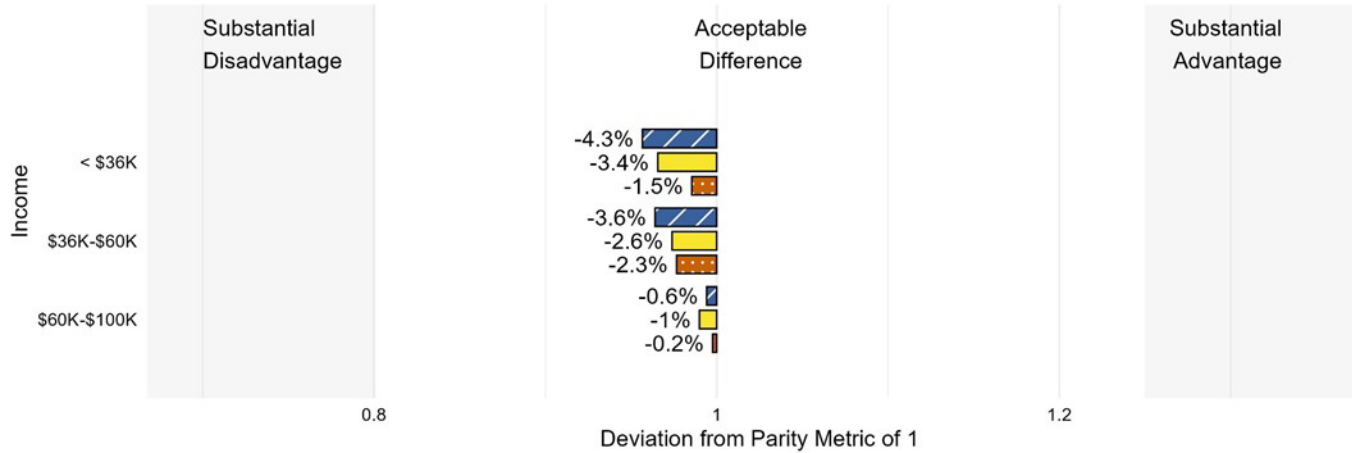
Fairness Metric: Accuracy equality ratio  $(TP + TN)/(TP + FP + TN + FN)$

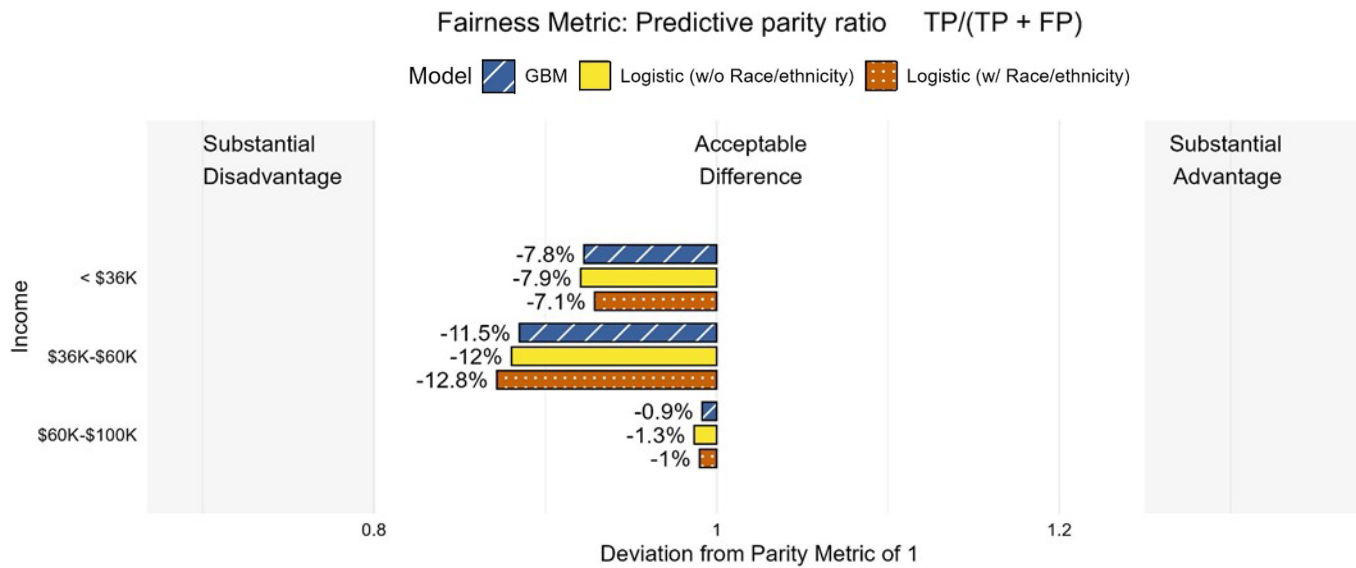
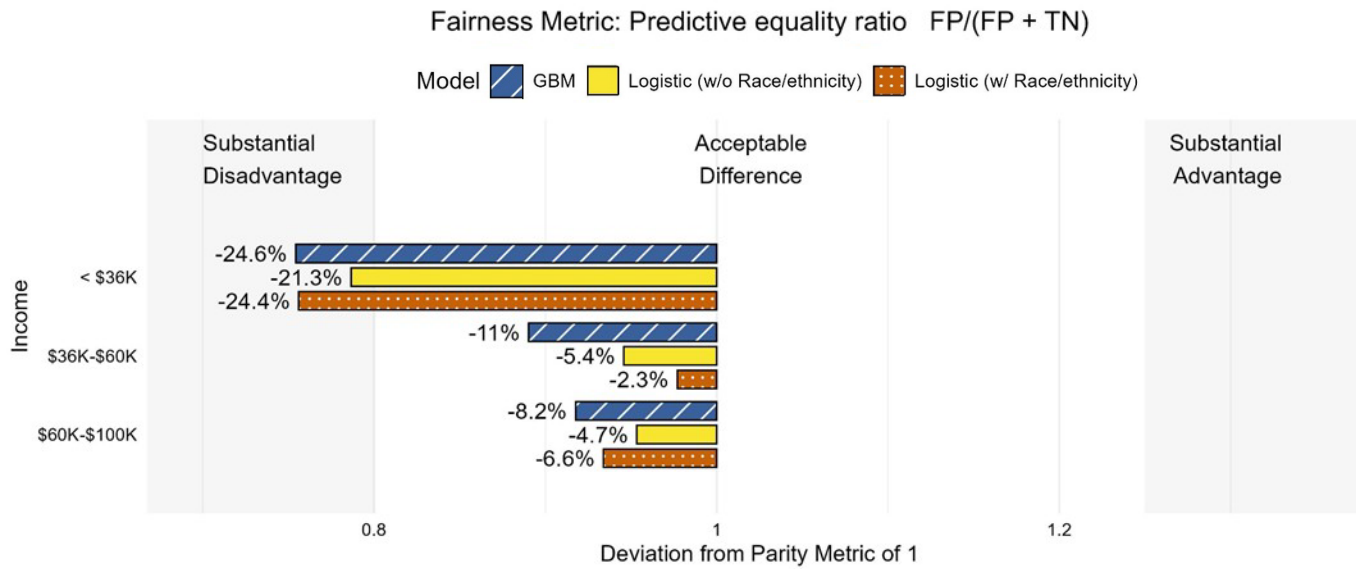
Model  GBM  Logistic (w/o Race/ethnicity)  Logistic (w/ Race/ethnicity)



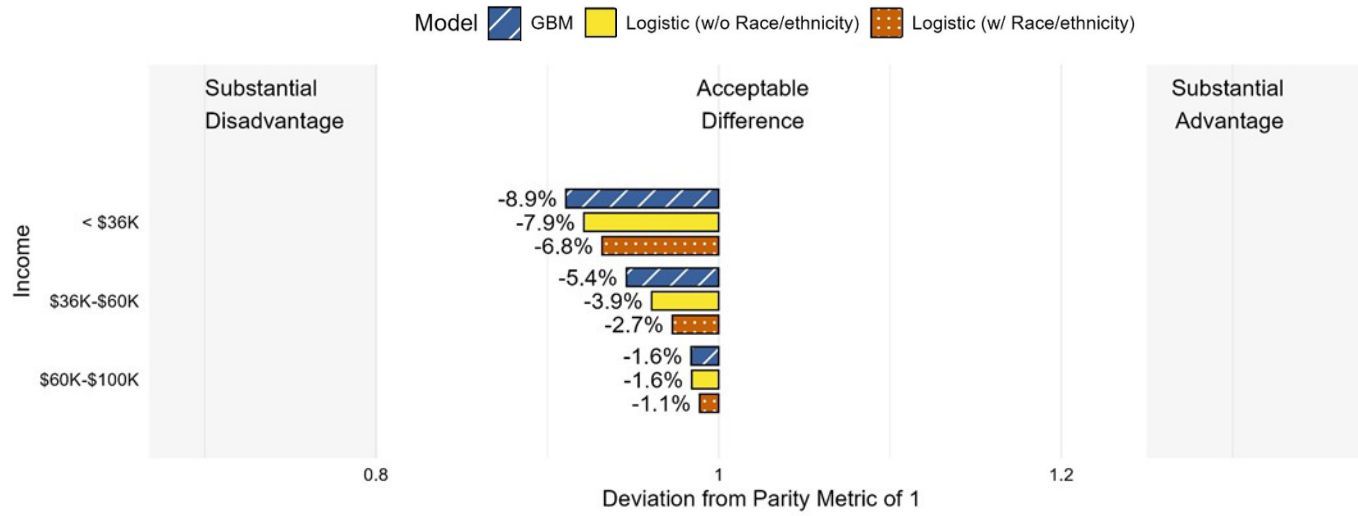
Fairness Metric: Equal opportunity ratio  $TP/(TP + FN)$

Model  GBM  Logistic (w/o Race/ethnicity)  Logistic (w/ Race/ethnicity)





Fairness Metric: Statistical parity ratio  $(TP + FP)/(TP + FP + TN + FN)$



## How does the predictive accuracy of a fairness-aware machine learning model compare to that of traditional regression models, particularly for underrepresented subgroups?

This study reveals that the exclusion of race/ethnicity in a logistic regression model does not significantly reduce the overall performance of model metrics such as accuracy, precision, or recall when compared to models with race/ethnicity or to a more complex GBM model (Table 3). While the logistic model without race/ethnicity has a slightly higher F1 score and AUC, which does indicate better overall balance and discriminative power, the logistic model with race/ethnicity performs the best for the low-income subgroup. This suggests that including race/ethnicity helps capture socioeconomic disparities indirectly. Of note is the fact that for the African American subgroup, both logistic models perform equally well, but the GBM model slightly underperforms, highlighting potential limitations in the GBM model for this subgroup.

**Table 3.** Accuracy Metrics Across Models 3–5

| Model                     | Logistic (with ethnicity) | Logistic (no ethnicity) | GBM   |
|---------------------------|---------------------------|-------------------------|-------|
| Accuracy                  | 0.829                     | 0.832                   | 0.830 |
| Precision                 | 0.848                     | 0.850                   | 0.848 |
| Recall                    | 0.967                     | 0.968                   | 0.970 |
| F1 Score                  | 0.224                     | 0.242                   | 0.216 |
| AUC                       | 0.788                     | 0.791                   | 0.789 |
| African American accuracy | 0.694                     | 0.694                   | 0.676 |
| Low-income accuracy       | 0.809                     | 0.797                   | 0.791 |

Key takeaways from this table include the following:

1. Excluding race/ethnicity as a predictor does not significantly reduce accuracy, precision, or recall for the overall population. However, it does slightly decrease the model's performance for low-income students.
2. The more complex GBM model does not outperform simpler logistic regression models in terms of accuracy or fairness for underrepresented groups.
3. The logistic model without race/ethnicity has the highest F1 score and AUC, indicating that excluding race/ethnicity avoids potential overfitting while maintaining or improving generalizability.
4. If our primary concern is fairness, the logistic regression model without race/ethnicity may be preferable, as it performs on par with other models while avoiding direct reliance on the sensitive attribute of race/ethnicity.

## Discussion

In the present study, I examined the prediction of first-year college GPA (FYGPA) to help postsecondary institutions identify how they can develop predictive models while complying with the ban on race/ethnicity-based affirmative action in admissions. Traditional metrics like high school GPA (HSGPA) and ACT Composite (ACTC) score have been shown in prior research to be significant predictors of FYGPA. Exploring how these predictors function independently and jointly can help postsecondary institutions allocate resources, especially during the critical transition from high school to college.

The 2023 U.S. Supreme Court decision that ended affirmative action in college admissions has prompted postsecondary institutions to explore alternative methods to meet enrollment and fairness goals. As a result, postsecondary institutions must consider new ways to evaluate the likelihood of student success in college. Advanced machine learning techniques such as fairness-aware algorithms can be applied to refine predictive models in a way that potentially offers advantages over traditional models without relying on race/ethnicity.

The research presented here compared traditional logistic regression models (with and without race/ethnicity included as a predictor) to a fairness-aware machine learning gradient-boosted machine model for predicting FYGPA. I examined whether traditional models exhibited potential bias and how that contrasted with a fairness-aware model that used advanced algorithms to potentially reduce prediction bias. The primary goal of this study was to assess whether a fairness-aware model produced better fairness metrics and greater accuracy than traditional logistic regression models.

In the analysis of the traditional logistic regression models, we see that the model that used ACT scores as the sole academic predictor (ACTC score model) tended to demonstrate the fairest metrics across subgroups, particularly for African American students and students from low-income families. This model mostly avoided instances of metrics with substantial disadvantages for African American and low-income students relative to other models, making it a strong candidate for postsecondary institutions that are focused on developing fair and accurate predictive models. On the other hand, the model that used HSGPA as its sole academic predictor (HSGPA model) exhibited the most bias, raising concerns for African American students and students from low-income families. At ACT, we do not recommend using a single academic measure when evaluating students. We always recommend using multiple measures. This philosophy is embodied in the HSGPA plus ACTC score model, which struck a balance between the positive and negative outcomes of the HSGPA model and the ACTC score model while still showing some level of bias, indicating that the two academic measures do improve fairness but may not completely eliminate disparities.

When we compared the logistic regression models to a gradient-boosted machine (GBM) model, analyses indicated that the GBM model also caused some concern based on fairness metrics. That said, the logistic regression models, both with and without race/ethnicity, tended to outperform the GBM model in terms of fairness, particularly for African American and low-income students. The GBM model tended to show greater disparities in fairness metrics,

suggesting the need for further optimization of the fairness-aware model to address its demonstrated biases. Recall that in this implementation of a fairness-aware model, I made a conscious decision not to implement postprocessing steps because of the difficulty in justifying treating students differently based on their demographic group membership. These findings emphasize why it is so important to select predictive models that balance both fairness and accuracy, especially in light of recent legal changes affecting college admissions.

The exclusion of race/ethnicity in logistic regression models did not introduce notable bias. This is particularly important when we consider the 2023 U.S. Supreme Court decision prohibiting the use of race/ethnicity in admissions decisions. The fact that the logistic regression model without race/ethnicity largely maintained fairness without relying on race/ethnicity highlights its potential as a viable alternative for institutions that are working to comply with new legal mandates while promoting fair educational outcomes. Based on the present analysis and the decision not to implement postprocessing in the fairness-aware model, the findings suggest that the use of a logistic regression model without race/ethnicity can achieve a balance between fairness and predictive accuracy while largely avoiding substantial disadvantages across student subgroups.

There are several important implications from this research for postsecondary admissions officials. First, this study confirms that traditional indicators of achievement such as ACT Composite scores and high school GPA should not be used in isolation. Rather, admissions officials should adopt a combined approach, as demonstrated in the HSGPA and ACTC score model and Model 4, in order to strike a balance between fairness and accuracy. Using both achievement measures together helps to mitigate potential biases that arise when one relies solely on HSGPA, which in this study exhibited more bias for African American and low-income students. Given that the ACTC-based model demonstrated higher fairness metrics for focal groups, it may be worth exploring whether using weights on ACTC score alongside HSGPA could further enhance fairness without disproportionately disadvantaging student subgroups.

Given that postsecondary institutions are now barred from using race/ethnicity-based affirmative action, this study suggests that predictive models that exclude race/ethnicity do not necessarily introduce bias and may even enhance fairness and therefore provide a lawful and effective way to evaluate students' potential success in college. The metrics examined in the present study focused on academic metrics; however, postsecondary institutions should explore holistic perspectives of students, which incorporate nontraditional factors such as personal essays, socioeconomic background, and school context.

Postsecondary institutions should be cautious when attempting to leverage fairness-aware machine learning models. While it is true that models such as the one implemented in this study hold promise, the present study indicates that logistic regression models performed better in terms of fairness, particularly for African American and low-income students. This was true in the context of limiting the types of postprocessing features that fairness-aware models employ in order to comply with the ban on the use of race/ethnicity in college admissions. In addition to the prudential decision not to use such postprocessing methods, the implementation of fairness-aware machine learning models may place an undue burden on postsecondary admissions officials, as such models require an in-depth understanding of the mathematical processes

being employed by machine learning models. In addition, practically speaking, in this study I had access to data about thousands of students across an entire state. In practice, any individual postsecondary institution may have access to only hundreds of applicants, which makes the use of such sophisticated models more complex if not unfeasible due to small sample sizes.

A final implication that should not be overlooked is the importance of transparency and accountability in policy and decision-making considerations. The use of logistic regression modeling is far more easily understood and communicated to students and parents than the complexities of a fairness-aware model. Postsecondary officials will need to be able to document and justify the selection of specific models and criteria in order to demonstrate that they are in compliance with any legal standards as well as institutional enrollment goals.

## Limitations

The study was limited to students from a single southern U.S. state who attended a public 4-year institution immediately after high school. This has the potential to limit generalizability to other students and situations. Moreover, in order to facilitate analysis, I included only students who had valid data for all variables, potentially introducing bias by excluding students with incomplete or missing data. In this population, 5.7% of the students did not report their gender, 7.9% preferred not to disclose or did not report their race/ethnicity, and 25.2% did not report their family income. This missing income group, the largest missing demographic, had an average ACTC score of 21.5, which was consistent with the average ACTC score for students from families with a \$36,000–\$60,000 family income, and most of the variables of interest in this study were highly similar when it came to comparing the study sample to the population. Additionally, the present study used self-reported HSGPA. While there is a research basis for using self-reported GPA when transcript GPA is not available, there is always the potential for these data to be susceptible to recall bias, social desirability bias, or overestimation.

In the present study, I used logistic regression. While the data in this sample did not support the use of hierarchical logistic regression, the lack of accounting for such nesting may offer a simplified view of a complex relationship between variables. In order to facilitate the use of the fairness metrics, I had to examine a binary outcome (i.e., attaining a C or higher or not attaining a C or higher). Examining an outcome of attaining a B or higher or examining prediction of FYGPA on a continuous scale might provide a more nuanced perspective on the prediction of postsecondary success.

## References

- Camara, W., Kimmel, E., Scheuneman, J., & Sawtell, E. A. (2003). *Whose grades are inflated?* (Research Report No. 2003-4). College Board.  
<https://files.eric.ed.gov/fulltext/ED563203.pdf>
- Curabay, M. (2016). *Meta-analysis of the predictive validity of Scholastic Aptitude Test (SAT) and American College Testing (ACT) scores for college GPA* [Doctoral dissertation, University of Denver]. Digital Commons.  
<https://digitalcommons.du.edu/cgi/viewcontent.cgi?article=2225&context=etd>
- Friedman, J., Sacerdote, B., & Tine, M. (2024). *Standardized test scores and academic performance at ivy-plus colleges*. Opportunity Insights.  
[https://img.theepochtimes.com/assets/uploads/2024/02/22/id5592982-SAT\\_ACT\\_on\\_Grades.pdf](https://img.theepochtimes.com/assets/uploads/2024/02/22/id5592982-SAT_ACT_on_Grades.pdf)
- Kuncel, N. R., Credé, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research*, 75(1), 63–82.  
<https://doi.org/10.3102/00346543075001063>
- Marini, J. P., Westrick, P. A., Young, L., Ng, H., Shmueli, D., & Shaw, E. J. (2019). *Differential validity and prediction of the SAT®: Examining first-year grades and retention to the second year*. College Board. <https://files.eric.ed.gov/fulltext/ED597325.pdf>
- McNeish, D. M., Radunzel, J., & Sanchez, E. I. (2015). *A multidimensional perspective of college readiness: Relating student and school characteristics to performance on the ACT®* (Research Report No. 2015-6). ACT. <https://files.eric.ed.gov/fulltext/ED563774.pdf>
- Sanchez, E. I. (2024). *Changes in predictive validity of high school grade point average and ACT® Composite score after the COVID-19 pandemic*. ACT.  
<https://www.act.org/content/dam/act/secured/documents/R2328-Changes-in-Predictive-Validity-of-HSGPA-and-ACT-Composite-Score-After-COVID-19-2024-09.pdf>
- Sanchez, E. I., & Buddin, R. (2016). *How accurate are self-reported high school courses, course grades, and grade point average?* (Research Report No. 2016-3). ACT.  
<https://www.act.org/content/dam/act/unsecured/documents/5269-research-report-how-accurate-are-self-reported-hs-courses.pdf>

Shaw, E. J., & Mattern, K. D. (2009). *Examining the accuracy of self-reported high school grade point average* (Research Report No. 2009-5). College Board.

<https://files.eric.ed.gov/fulltext/ED562616.pdf>

Students for Fair Admissions v. President and Fellows of Harvard College, 600 U.S. (2023).

<https://www.oyez.org/cases/2022/20-1199>

Students for Fair Admissions v. University of North Carolina, Citation pending U.S. (2023).

<https://www.oyez.org/cases/2022/21-707>

Warren, J. M., & Goins, C. L. (2019). Exploring the relationships between high school course enrollment, achievement, and first-semester college GPA. *Journal of Educational Research and Practice*, 9(1), 386–399. <https://doi.org/10.5590/JERAP.2019.09.1.27>

Westrick, P. A., Le, H., Robbins, S. B., Radunzel, J. M. R., & Schmidt, F. L. (2015). College performance and retention: A meta-analysis of the predictive validities of ACT® scores, high school grades, and SES. *Educational Assessment*, 20(1), 23–45.

<https://doi.org/10.1080/10627197.2015.997614>

## Appendix: Logistic Regression Model Coefficients

| Predictor                                   |   | HSGPA<br>(Model 1)   | ACTC score<br>(Model 2) | HSGPA +<br>ACTC score<br>(Model 3) | HSGPA + ACTC<br>score with<br>no race/ethnicity<br>(Model 4) |
|---|---|----------------------|-------------------------|------------------------------------|--|
| HSGPA                                       |   | 0.924***<br>(0.052)  | —<br>—                  | 0.933***<br>(0.066)                | 0.930***<br>(0.065)  |
| ACTC score                                  |   | —<br>—               | 0.774***<br>(0.063)     | 0.429***<br>(0.075)                | 0.453***<br>(0.073)  |
| Gender                                      | Female  | 0.221**<br>(0.107)   | 0.531***<br>(0.102)     | 0.268**<br>(0.108)                 | 0.264**<br>(0.108)   |
| Ethnicity                                   | African American                                      | -0.450***<br>(0.165) | -0.318**<br>(0.159)     | -0.310*<br>(0.167)                 | —<br>—   |
|   | Two or more /<br>American Indian /<br>Native Hawaiian | -0.417**<br>(0.212)  | -0.400**<br>(0.202)     | -0.400*<br>(0.215)                 | —<br>—   |
|   | Asian   | 0.432<br>(0.417)     | 0.952**<br>(0.416)      | 0.506<br>(0.422)                   | —<br>—   |
|   | Hispanic  | 0.313<br>(0.193)     | 0.324*<br>(0.186)       | 0.370*<br>(0.193)                  | —<br>—   |
|   |   |                      |                         |                                    |  |
| Family income                               | \$36K–\$60K   | 0.319**<br>(0.152)   | 0.328**<br>(0.144)      | 0.289*<br>(0.150)                  | 0.263*<br>(0.149)  |
|   | \$60K–\$100K  | 0.292*<br>(0.149)    | 0.366**<br>(0.142)      | 0.242<br>(0.149)                   | 0.229<br>(0.145)   |
|   | >\$100K   | 0.405***<br>(0.155)  | 0.510***<br>(0.148)     | 0.323**<br>(0.156)                 | 0.305**<br>(0.149)   |
| % poverty at high school                    |   | -0.106<br>(0.065)    | 0.092<br>(0.062)        | -0.060<br>(0.066)                  | -0.087<br>(0.065)  |
| Number of AP courses offered at high school |   | 0.054<br>(0.059)     | -0.008<br>(0.057)       | 0.044<br>(0.060)                   | 0.038<br>(0.059)   |
| % White students at high school             |   | -0.035<br>(0.068)    | 0.074<br>(0.064)        | -0.025<br>(0.068)                  | 0.000<br>(0.060)   |
| HSGPA * ACTC score interaction              |   | —<br>—               | —<br>—                  | 0.230***<br>(0.058)                | 0.216***<br>(0.057)  |
| Constant                                    |   | 1.585***<br>(0.143)  | 1.246***<br>(0.133)     | 1.595***<br>(0.144)                | 1.594***<br>(0.128)  |
| Observations                                |   | 3,298                | 3,298                   | 3,298                              | 3,298  |
| Log likelihood                              |   | -1,218.052           | -1,314.996              | -1,196.395                         | -1,204.598   |
| Akaike information criterion                |   | 2,462.104            | 2,655.992               | 2,422.791                          | 2,431.196  |

\* $p < .1$ ; \*\* $p < .05$ ; \*\*\* $p < .01$ .



## ABOUT ACT

ACT is transforming college and career readiness pathways so that everyone can discover and fulfill their potential. Grounded in more than 65 years of research, ACT's learning resources, assessments, research, and work-ready credentials are trusted by students, job seekers, educators, schools, government agencies, and employers in the U.S. and around the world to help people achieve their education and career goals at every stage of life. Visit us at [www.act.org](http://www.act.org).