Title

Inner Group Privilege: Understanding Underrepresentation in STEM

Author

Wade, Gizella

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Inner Group Privilege: Understanding Underrepresentation in STEM

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Applied Statistics

by

Gizella Wade

2024

ABSTRACT OF THE THESIS

Inner Group Privilege: Understanding Underrepresentation in STEM

by

Gizella Wade
Master of Applied Statistics
University of California, Los Angeles, 2024
Professor Robert L. Gould, Co-Chair
Professor Yingnian Wu, Co-Chair

With higher education becoming more competitive year by year, along with the job and housing market, it is no surprise there is so much discourse about issues with the college admissions process. Furthermore, the Supreme Court's decision to end race considerations within the admissions process makes it evermore important to understand the implications race and other demographic factors have for students' educational outcomes. Due to the ever-expanding STEM fields and their reputation for having high-paying jobs, this thesis examines the role demographic background plays in predicting if a student is considering a STEM major.

The thesis of Gizella Wade is approved.

Chad J. Hazlett

Mark S. Handcock

Yingnian Wu, Committee Co-Chair

Robert L. Gould, Committee Co-Chair

University of California, Los Angeles

2024

To my late sister, may she rest in peace.

TABLE OF CONTENTS

1 Introduction **1**

1.0.1 Analysis & Research Questions 1

1.0.2 Background 2

1.0.3 Statement of the Problem 3

1.0.4 Significance and Rationale 6

2 Literature Review **10**

3 Data Analysis **22**

3.0.1 Background of Data & Analytic Method 22

3.0.2 Research Questions & Hypotheses 22

3.0.3 Exploratory Data Analysis 23

3.0.4 Model Selection Process 33

3.0.5 Final Model Analysis 40

4 Summary **53**

References **58**

LIST OF FIGURES

3.1 Distribution of Majors Considering in Subbaccalaureate 2012 High School Graduates 25

3.2 Distribution of Majors Considering in Baccalaureate 2012 High School Graduates 25

3.3 Distribution of Majors Considering in 23-categories 2012 High School Graduates 25

3.4 Distribution of Majors Considering in NSF Field 2012 High School Graduates 25

3.5 Distribution of Majors Considering in STEM Field 2012 High School Graduates 26

3.6 Distribution of Majors Considering in CTE Field 2012 High School Graduates 26

3.7 Major Distribution 27

3.8 Income Distribution 27

3.9 Major/Highest Math Course Taken in 9th Grade Distribution 29

3.10 Major/Highest Math Course Taken in HS Distribution 29

3.11 Race/Highest Math Course Taken in HS Distribution 31

3.12 Race/Highest Science Course Taken in HS Distribution 31
3.13 Correlation Matrix for Predictors 32
3.14 ROC Curve Lasso w/o PCA 35
3.15 ROC Curve Ridge w/o PCA 35
3.16 Six Most Positively Associated Variables Dictionary 38
3.17 Six Most Negatively Associated Variables Dictionary 38
3.18 Six Most Positively Associated Variables 39
3.19 Six Most Negatively Associated Variables 39
3.20 Random Forest With All Variables 39
3.21 Random Forest Six Selected Variables 39
3.22 Model Dictionary 41
3.23 ROC Curve All SES Variables 42
3.24 ROC Curve With Income Only 42
3.25 ROC Curve With Race Only 42
3.26 ROC Curve With Sex Only 42
3.27 ROC Curve for Male Students 43
3.28 ROC Curve for Female Students 43
3.29 ROC Curve Asian Students 52
3.30 ROC Curve Black Students 52
3.31 ROC Curve Latinx Students 52
3.32 ROC Curve White Students 52

LIST OF TABLES

3.1 Lasso & Ridge Models Without PCA 35

3.2 Lasso & Ridge Models With PCA 37

3.3 Random Forest All Variables Model 37

3.4 Overall Student Models 41

3.5 Male Student Models 44

3.6 Female Student Models 45

3.7 Tier One Income Models 47

3.8 Tier Two Income Models 47

3.9 Tier Three Income Models 47

3.10 Tier Four Income Models 47

3.11 Tier Five Income Models 47

3.12 Lower Income Students' Models 49

3.13 Middle Income Students' Models 49

3.14 Higher Income Students' Models 49

3.15 Asian Students' Models 50

3.16 Black Students' Models 50

3.17 Latinx Students' Models 50

3.18 White Students' Models 50

CHAPTER 1

Introduction

1.0.1 Analysis & Research Questions

The researcher seeks to examine the impact demographic background (gender, income, race) has on the relationship between academic achievement and major choice in college.

Research Questions

1. What is the significance of gender and income across racial/ethnic demographics when considering education attainment?
2. Is the underrepresentation of racial and ethnic minorities in STEM majors potentially associated with different systemic treatment (e.g. two students perform similarly in a class, but only one is pushed to take on a more advanced course load in the following academic year) in high school?

Analysis Approach

A model will be built to predict whether a student is considering studying a STEM major based on objective, student success factors (e.g. GPA, highest level of math taken in high school, SAT score) and will include students of all socioeconomic backgrounds. From there the students will be separated by race (Asian, Black, and White) and six separate models will be generated to predict a student's college major - an overall model within racial group, two models based on binary sex (male and female), and three models for family income (low, middle, and high-income). Since the majority of the Latinx students in the dataset

are White, within the subcategory for White students there will be twelve models - the first six including non-Latinx and Latinx White students and the second batch including only Latinx students. Finally, the models will be compared within racial groups and across racial groups, while referencing the overall model too.

Hypotheses

1. It is anticipated all demographic variables (income, race, sex) will have significance on a student's choice in college major

2. However, the researcher expects, after controlling for meaningful demographic variables, race will still be a significant factor in predicting students' college majors. For example, students of the same race will have more similarities in their major choice than students of the same sex or income level - so, higher income White students will have more similar model metrics with lower income White students than with higher income Asian students.

1.0.2 Background

The year 2020 was a turning point in American society; people stood together in a way they historically have not because of the public execution of George Floyd. As America is reckoning with its historic and continued mistreatment of Black Americans, it is still falling short in understanding how it got to this point. The answer is quite simple, through systemic racism, but that is a much more complex idea compared to overt racism (e.g, Black codes) [1].

Despite the strive for more equality and equity in society and support of the majority of the American people, most Americans do not understand their role in the inequity faced by racial minorities in the United States. It is simple to point fingers at Klu Klux Klan members, neo-nazis, and police brutality, but less apparent obstacles hurt Black people the most. The more subtle behaviors (e.g., racial microaggressions birthed by subconscious racism) can be just as harmful as blatant racism [2]. However, it can also be more difficult to notice than

overt forms of discrimination [3].

As such, people can fail to notice institutional behaviors disenfranchising Black people in America and how systemic racism is reinforced by microaggressions [4]. One may deduce that since it is a systemic issue, no individual can be accountable for the harm caused. On the other hand, systems are cultivated through multiple people meaning that while there is no one person to blame, it is everyone's fault [5, 6]. Therefore it is the responsibility of everyone to improve the system. However, it is imperative for White people to involve themselves because they have the most (if not all) systemic power and control, due to being the "most dominant caste" [1]. The problem is that they are the racial demographic least impacted by race issues, self-admittedly [7]. One of Critical Race Theory's tenets states racism is a power structure, which is why White individuals are least involved and impacted by racism.

In short, many White people fear that if minorities become more successful and gain more access and resources, their position in society's hierarchy will suffer [8]. Therefore, systemic racism is used to criminalize marginalized groups and deny them access to society's common goods so that White people can maintain their power [9]. Thus preventing marginalized groups, perceived as a threat to the system, from building wealth and accruing power and systemic control to potentially recalibrate the power hierarchy, is essential to maintaining the status quo. Nevertheless, as previously stated, it is on everyone to solve the issue, and, as such, the study being conducted is meant to serve as an educational tool and as the researcher's contribution to help address systemic racial inequality.

1.0.3 Statement of the Problem

While reviewing the origins of racial inequality and inequity is beyond the scope of this study, it is important to note the prominent role of inequitable and harmful policies. The

issue of systemic racism extends beyond the violent, brutal policing of Black individuals, discriminatory laws (e.g., the war on drugs), or the prison industrial complex, and there is no way to pinpoint where it begins, but it certainly does not end. However, educational inequality's impact on racial minorities may be one of the root causes of social inequality and other forms of institutionalized racism. Nevertheless, to understand the depths of the problem, one must first understand the history.

While Alexander Twilight is recorded as the first African American to receive a college degree in 1823, he was marked White on census records for most of his life and would be considered White by most people in today's society [10]. Twilight's ability to receive an education was due to his proximity to whiteness, and his classification of being the first Black person to receive a college education is due to the one-drop rule. Unfortunately, not much has changed in recent history, with a ruling in 1985 denying a woman the right to identify as "White" on her passport due to having a Black ancestor six generations prior [11]. Furthermore, Black people were not granted education rights until the 1860s and 1870s or approximately 30 years after the first all-women's college, Georgia Female College (now known as Wesleyan College), was established in 1836 [12]. While the first Black college was established in 1837, Black women could not attend either of these schools, as the first Black college for women was not founded until 1881, almost two decades after the Emancipation Act of 1865 and about a century prior to Wesleyan graduating its first Black students in 1972 [12, 13].

Soon after enslaved African-Americans received human rights with the 14th Amendment, which was supposed to secure every American's right to equal treatment under the law, the country would realize racial discrimination could not be legislated away. As Blacks began to close the gap in opportunities between them and White Americans, Black Codes (the first laws being established the same year of emancipation) and the 1896 Plessy v. Ferguson U.S.

Supreme Court ruling, which legally allowed racial segregation by framing it as "separate but equal," began a running history of systemic racism [14, 15]. Following the 1886 SCOTUS ruling, racially discriminatory policies such as Jim Crow laws became more common.

Though many of these laws were outright unconstitutional (e.g., the 1879 California statute implemented to prohibit people of Chinese descent from voting or Virginia's eugenics 1924 Racial Integrity Act prohibiting White people from marrying anyone not White - meaning anyone with any amount of non-Caucasian ancestry, excluding people of one-sixteenth or less Indigenous American ancestry - until 1967), it did not stop the harmful effects of the laws [16, 17]. Jim Crow was touted as "separate but equal," but the Warren Court, in Brown v. Board of Education, overturned the 1886 Plessy ruling on the grounds schools cannot be equal, if they are separate [18, 19]. Following the Civil Rights Act of 1964 (approximately 60 years ago), schools remained relatively separate and unequal [20]. The inequality and segregation were further worsened with the establishment of private schools, which allowed White families to educate their children in homogenous spaces with other White children and access to better educational opportunities, while discriminating against Black children from entrance into those schools [21].

Over time, the discriminatory laws and institutional practices emplaced created an issue that continued to snowball, bringing Americans to this very point in time. Now, schools are segregated because of past and current systemic practices creating inequity (e.g., the racial income gap). Black families are less likely to be able to afford private school tuition relative to other racial demographics because they are one of the poorest demographics and are underpaid relative to their peers, even when comparing across similar job titles and qualifications [22, 23]. Furthermore, past racist housing practices (e.g., redlining a racially discriminatory practice of withholding property loans and mortgage approvals from Black customers) have forced Black families to remain in the same underfunded education districts

[24]. Subsequently, the continued covert and blatant White supremacy in society curates subconscious biases affecting everyone, including K-12 teachers. In short, due to systemic biases, Black students are less likely to feel self-empowered or have much self-efficacy. The problem is only aggravated by being taught by racially biased (mostly White) teachers who do not understand Black students' life experiences in an education system never meant to serve them [25, 26, 27]. In order to undo the systemic inequities, it must first be understood the impact schooling inequality has on students and fix the policies created to destabilize and oppress non-White students.

1.0.4 Significance and Rationale

Societal inequality does not only affect racial minorities but causes inadvertent harm to all of society. Many policies historically enacted to hurt Black people are beginning to hurt other racial minorities and even White people. For example, the unfair wages waiters and waitresses earn are rooted in racism; after emancipation, Black people were allowed jobs but were not given the same jobs as White Americans. Black workers usually had lower-paying and undesirable jobs, one being waiting. It was considered demeaning to wait on people, and the tipping system was a humiliation tactic [28]. Even the low-paying domestic-labor jobs were usually worked by Black women - "The Help" (2009) depicts the nature of their work [29].

Another issue is the unfairness in the college admissions process. Most people focus on legacy admits and racial Affirmative Action. However, the demographic benefiting the most from admissions-based Affirmative Action is White women. Schools admit more students from private schools than legacy students - the most competitive schools usually allocate 10 to 25% of admits for legacy applicants, in contrast to Harvard's almost 40% in its Class of 2025 coming from private schools [30, 31]. Despite Affirmative Action being a race-centric conversation, research indicates White women are the demographic benefitting the most

from Affirmative Action policies, from college admissions to jobs [32, 33, 34]. White women today are more educated and hold a more significant portion of the workforce because of Affirmative Action policies [33]. They have also seen advancement in leadership positions, unlike people of color (including women of color) [33].

The anti-blackness driving the anti-affirmative action narratives fails to acknowledge that White women benefit the most from Affirmative Action and how Affirmative Action can benefit anyone from a disenfranchised background (including lower-income students) while ignoring the historical, systemic barriers Black individuals face with education. In other words, removing Affirmative Action will hurt non-Black minorities and White women the most. The spread of these issues in other communities and demographics is causing civil distress and creating disharmony among people (e.g., many people point the finger at Black students when addressing Affirmative Action) [35]. Especially after the June 2023 SCOTUS decision overturned the use of Affirmative Action in college admissions, affecting 86% of people identifying as non-White [36]. The decision also jeopardized higher education's vital role as an engine of equitable opportunity and economic mobility.

Many people pointing the finger at the Black community fail to understand the lack of systemic power Black people have in society: they have no control over state-wide or federal laws (e.g., Affirmative Action) nor institutional power to wield said laws in their favor. While the current 118th Congress is more racially diverse than ever, the 60 Black members account for 11.2% of all members compared to the 13.6% of Black Americans in the U.S. population [37, 38]. The slight underrepresentation would be less troubling if not for the fact that only 3% of Senate members are Black, and there are currently no Black governors [39, 40]. Thus, the Black community lacks political strength at the state-wide level. Nevertheless, many people against racial Affirmative Action demand that race consideration be removed from admissions decisions and replaced with income consideration [41].

However, there is a clear relationship between poverty and race. For one, Black people are the country's second most economically deprived racial demographic behind Indigenous Americans [42]. It is vital to note Indigenous Americans are a much smaller portion of the population, meaning these numbers can be heavily skewed, but also, unlike Black people, they were given reparations - they received land on which they could govern themselves. Some tribes receive $25,000 per person at age 18 and another $25,000 when they turn 21, with another lump sum given to them at 25 [43]. Let it also be noted African Americans were inbred, enslaved, and displaced by the Indigenous Peoples of America [44]. Also, it must not be forgotten that Black people are the only minority group not to receive reparations "for state-sanctioned racial discrimination" in America, and this fact is crucial because this thesis distinguishes between racism and anti-blackness [45]. Therefore, while it may hold true for other racial and ethnic minority groups that income disparity can be a significant factor in their life outcomes, for Black people this may not be the case.

For example, low-income White kids are less likely to go to jail than high-income Black kids [46]. However, while this study focuses on anti-blackness, it should be understood that this thesis is also trying to examine how class (i.e., income, gender, and race) affects all Americans and America as a whole. These inequitable structures undermine the values America was built on: equality and opportunity. While these ideologies may have never truly existed in America, it does not mean they cannot exist. As such, this study aims to improve understanding of racial systematic disenfranchisement and create more unity among others to strive for a truly equitable society. However, pursuing equity in society may mean giving more consideration to some groups over others.

For example, because the American government displaced Japanese people into incarceration sites during World War II, they received $1.5 billion in federal funds [45]. Jewish people also

received monetary reparations from multiple sources - including the United States through the Marshall Plan and over 3 billion Deutsche Marks from Germany [45]. Both Japanese and Jewish people, without a doubt, should have received reparations after the atrocities they faced. However, Black people were also persecuted in Germany during World War II, and many Black Americans fought in the war. However, they did not have proper rights in America because of Jim Crow laws [47, 48]. Furthermore, one of the primary reasons Black Indigenous people did not receive reparations and continue to fight for their claim on their Indigenous heritage is because by 1860, in some Southern areas, where race was considered binary of Black (mainly enslaved) or White, White legislators held the belief the Indigenous Americans no longer qualified as "Native American," as many were mixed and part Black [49].

The distinct difference in treatment between Black people and other racially marginalized communities is why this study makes a distinction between racism and anti-blackness. The mentioning of monetary reparations is not an attempt to campaign for them. Although, considering other marginalized groups received monetary reparations, it only seems fair that African American people do too, especially given that estimates suggest African American people are owed roughly $150 million each [50]. However, research also suggests money is not enough, and, considering few people support African American people receiving monetary reparations, education may be an excellent place to start [50, 51].

CHAPTER 2

Literature Review

Historically, schools with predominantly Black student bodies offer fewer advanced STEM courses, preventing students from entering STEM-related careers [52]. Nationally, roughly 5000 high schools educate Black students without adequate STEM course selection. The lack of access is not a coincidence and is intrinsically political across race, ethnicity, and other socioeconomic lines [53]. In the article "Only STEM Can Save Us? Examining Race, Place, and STEM Education as Property," Bullock concludes that between STEM curricula being introduced in predominately Black educational districts and the utilization of abandoned schools, middle-income White families claim the schools and STEM curriculum as their own and then use it as a means to discriminate against racial minorities in admissions [54, 55].

The Supreme Court of the United States' (SCOTUS) removal of Affirmative Action in the college admissions process brings to question the historical education inequity for non-White students - especially for Black students. While most Americans will acknowledge systemic racism is an issue in America, almost three-quarters of Americans do not believe race should play a factor in college admissions, and even more believe gender should not be considered [7, 56, 57]. However, research does indicate that race plays a significant role in admission to college - even if it is inadvertent. For example, Harvard admitted 40% of students to its undergraduate class from private schools [58]. White students have historically been overrepresented in private schools, and many private schools' inception came after the Civil Rights Movement to keep education segregation alive, which is part of why schools remain highly segregated [59, 60, 61].

Although the previously stated fact is not often addressed, even less is understood about how systemic racism functions in the schooling system and its impact on non-White students' achievement. Hence, over 90% of Americans believe grades should be a factor in college admissions, but only 26% believe race and ethnicity should be considered in admissions [57]. Nevertheless, the insufficient access to STEM in predominantly minority schools is part of the underrepresentation of minorities in STEM, which in turn causes racial minorities to experience racism and alienation within their occupations if they choose to pursue STEM [55, 52]. Furthermore, even less is understood about how systemic racism can influence a student's college major and, as such, career trajectory, particularly when it comes to STEM fields and studies. Even with efforts to increase diversity in STEM, Black scientists remain underrepresented at every level, and only 9% of STEM workers are Black [62, 63]. Research has shown one of the ways to create and retain a diverse workforce is by creating space within organizations for underrepresented groups and giving them opportunities to excel [64].

One of the ways to achieve a nurturing and accepting work culture is by having racially representative leadership within a company [64]. On the other hand, it is difficult for a group to be proportionately represented in industries, let alone in more senior positions, if they study the subject matter disproportionately less [63]. Of course, the lack of diversity is unsurprising, considering Black students are almost 40% more likely to change their major and twice as likely to drop out altogether than their White peers [65]. Unfortunately, many people would assume the lack of representation as a failure of Black people, attributing racial education gaps to inadequate parenting, absence of strong role models, and lack of student motivation [66]. Given research shows parental education can positively impact a student's persistence in STEM, specifically if at least one parent studied STEM, it is fair to assume racial educational inconsistency is somewhat caused by insignificant representation [67].

A study conducted by The U.S. Department of Education does not support the claims that lack of parental engagement and student motivation restricts Black students from being academic high-achievers [68]. Research indicates Black students, on average, studied about half an hour less than White students per week, and 83% of Black students indicated their parents checked their homework, compared to 57% of White students. Considering almost 70% of people hold pro-White/anti-Black racial bias and teachers are just as racially biased as the general population, a significant portion of these educational gaps can most likely be explained by systemic and structural racism [69, 70].

Moreover, research has shown a negative correlation between more implicit and explicit pro-White/anti-Black racial bias and test scores [70]. Furthermore, these gaps persist with standardized testing such as the SAT, especially in math test scores. It is essential to note that math scores are even more significant to this thesis because math is a more significant factor in college admissions and, naturally, the likelihood of a student studying STEM during their undergraduate tenure [71].

Having more Black STEM workers is not just a matter of diversity but also closing the racial wealth gap in America; regardless of socioeconomic factors such as race, non-STEM careers have significantly lower salaries compared to STEM occupations [72]. Although the wealth gap naturally shrunk over time due to federal policies and laws such as the Civil Rights Act of 1964 and other efforts to gain racial equality, Black Americans only have 4% of the wealth share in America, despite making up 13% of the total American population. White Americans hold 84% of the wealth relative to their 60% share of the population [73]. However, due to White Americans having a head start with wealth because of systemic racial oppression, there has been stagnation in the shrinkage of the gap since 1980 [73].

Income disparity is not just a matter of job title, house size, or other material items. The

saying may be, "Money cannot buy happiness," but it can buy success; a study recently conducted by Georgetown found money, not ability, to be the most significant predictor of success [74]. In other words, as long as Black Americans are stuck in lower-wage jobs and cannot afford better opportunities for themselves, they will be stuck in a perpetual cycle of poverty and underachievement. Unfortunately, it is not as simple as being more motivated when it comes to Black students. However, barriers and lack of access to better educational opportunities systemically restrict their economic growth.

While the lack of representation and resources can discourage Black students interested in STEM subjects, the systemic failure of the K-12 school system appears to be a much more difficult obstacle to overcome. An article published in 2012 indicated Black students had more interest in STEM subjects than their peers [75]. Interestingly, another article indicated that students from lower-socioeconomic backgrounds, in which Black families are disproportionately affected due to Black workers having some of the lowest wages, are more likely to be interested in STEM than students from wealthier families [76]. Nevertheless, Black students continue to be underrepresented in STEM careers and majors.

However, recent state policies show promise for more inclusive representation in STEM. In Washington state, an automatic course enrollment policy was implemented in 2010, which tripled racial minority students' enrollment in advanced courses [77]. Tacoma (the state's fourth largest school district) was used as a pilot program and saw an increase from 28% to 71% in enrollment of advanced courses for students of color. This indicates that non-White students are systemically overlooked, which is worrisome given it is one of Washington's largest K-12 educational providers. Unfortunately, the issue does not begin at the state policy level but in the classrooms with the treatment of students and teachers' expectations of students.

A growing body of research suggests expectations a teacher may have for an individual student could deeply affect the student's academic performance. For instance, a teacher may have different expectations for a student based on their socioeconomic characteristics (i.e., race, gender, income) or previous academic achievement [78]. Therefore, Black and other racial and ethnic minority students may have less agency over their education than their White peers. In other words, two students could behave similarly but achieve different levels of academic success because their race or ethnicity is different, so their teacher has set different expectations for each of them. The burden to fix systemic racial inequality is not solely on teachers' shoulders, though. It goes beyond an issue of institutional encouragement and attention but expands to impact the disciplining of Black students, which discourages them in their academic settings.

The treatment of Black children in the American schooling system manifests the racial and social hierarchy ingrained into America created through the underfunding of Black schools and over-disciplining of Black children in schools [55, 79]. The aggravated disciplining of Black students could also be presented as a form of social control on Black people [55]. One study found that Black students were 13 times more likely to receive some form of suspension for minor violations (e.g., not following the school dress code, swearing, or using electronics in school) than White students [80]. The same study also found that Black students who were suspended for a minor incident during their first year of school received lower grades than Black students who were not suspended. Interestingly, the researchers did not find a student's grades and perception of the school climate during their first year of school relevant in estimating if a student would get suspended by their third year.

In other words, the results suggest racial bias and not the student's actions lead to an increased likelihood of suspension, which leads to poorer grades. Disproportionate school disciplining can cause Black students to become distracted from school due to frustrations or

face more serious consequences (e.g., loss of instructional time and social alienation within school), which can further negatively affect their academic performance [81, 82]. It can also cultivate a mistrust of their learning environments because the student may believe White students seem to be immune from punishment or less severely punished for the same offenses. The disproportionality in school discipline between White and Black students has greater implications, including pushing Black students faster toward the school-to-prison pipeline [83].

Despite all the available information and studies that suggest race plays a significant role in educational attainment, few studies thoroughly investigate the impact socioeconomic status has on education outcomes, such as interest in STEM. Most of the research on educational inequality focuses on student efficacy, achievements, and other non-systemic factors (e.g., parental education). In other words, the current literature regarding socioeconomic biases in college major selection fails to capture the institutional bias that can disproportionately and negatively impact non-White students, by not analyzing the varying outcomes students have based on their demographics. For instance, an article from 2013 assessed high school math achievement, math identity, race, and other various socioeconomic factors [84]. The primary conclusions were that exposure, success, and self-perception of math capabilities are some of the most significant predictors of a student choosing to study science/science-related subjects for their undergraduate degree.

Similarly, two papers published in 2021 and 2022 came to comparable conclusions as the previously discussed study - both studies concluded academic achievement and belief in ones self significantly influenced what students study [85, 86]. However, the study completed in 2021 focused primarily on engineering majors rather than on all science (or science-adjacent) subjects. Regardless, the most significant limitation of each study is that there needs to be an examination of students based on varying socioeconomic status to detect if institutional

factors are potentially creating the racial disparity in student outcomes.

Current research examines socioeconomic disparities in educational outcomes but fails to explain the extent of systemic resistance certain groups face relative to other socioeconomic groups (e.g., outcomes of middle-income Latinx men relative to low-income Asian women). Another constraint on the existing research is "self-selection bias" due to the populations examined not only being unrepresentative but also more than likely already high-achieving students with strong STEM identities along with high levels of self-efficacy. In particular, there was a longitudinal study on predicting if a student will select a STEM major, with its participants pooled from the semifinalists and finalists in a national science competition [87].

Moreover, the paper did not address systemic behaviors possibly restricting students from various ethnic and racial backgrounds. While some papers examined institutional behaviors, there was self-selection bias in the data's population as the research used students from a competitive engineering program. Although the results from the study found that the more diverse class settings produced higher achieving students - remarkably, more gender diversity improves student outcomes across the board. In comparison, more racial diversity positively impacts racial minority students' accomplishments [88]. The paper did not capture potential systemic behaviors (e.g., student support) influencing student outcomes, which was found to be significant for students' success in a 2000 study.

A paper produced in 2000 focused on institutional behaviors such as academic support from authority figures and other resources and found that teacher quality (e.g., preparation and qualifications) was the strongest indicator of student performance, even when making considerations for income level and language barriers [89]. Still, the paper did not focus on institutional behaviors and influences from a racialized perspective. Moreover, there is less research oriented toward quantifying the problem (e.g., creating a predictive model to

analyze the different outcomes students have based on their socioeconomic status), and no research was found to examine the potential impact of systemic racial bias on student's desire to study a STEM major in college. Even the paper completed in 2021 analyzing students' choice of engineering major using random forest models did not curate the model or use another form of statistical analysis to gauge the difference racial biases create in student outcomes.

A different paper released in May 2023 also used advanced statistical methodologies to predict a student's selected STEM major. However, unlike the older paper, the researchers used machine learning and regression analysis but did not focus solely on engineering. Still, neither article quantified the significance of a given socioeconomic indicator, such as income, on the likelihood of a student studying a engineering, math, or science topic in undergrad [90]. Other research topics, such as persistence in STEM, also fail to quantify the impact of socioeconomic status across subgroups when completing experiments. A prime example is a 2016 paper analyzing the underrepresentation and low completion rates for racially minority and female students. Although the research was highly informative and examined gender, race, and first-generation students, it only analyzed these variables overall (e.g., it did not consider the impact race may have on male vs female students) [91].

Nonetheless, the results from the study were interesting and filled in more of the literary gaps to explore disparities in representation further. The researchers concluded that outside of Asian students, racial minority and female students were less likely to major in a science field than their White and male peers, respectively. Also, the same demographic of students (minorities and women) struggled more to finish their degrees as they had longer completion times. One of the most significant elements in a student completing their degree was being White and male. These results were further validated by another article, "Exploring Factors that Predict STEM Persistence at a Large, Public Research University," which considered

quantitative variables such as GPA and SAT scores but did not further stratify subgroups (e.g. gender and income) by race and concluded race, gender, and first-generation status were significant for students' persistence in STEM [92].

As for previously mentioned studies, there could be an issue of self-selection bias given that the researchers examined students already accepted into university and interested in STEM. However, the research did support many of the aforementioned research conclusions - a racial disparity in STEM across income levels and gender. The research also captured a result not found in other literature: first-generation Black students interested in studying STEM need to potentially be considered more than anyone else when researching persistence in STEM or related topics. Another limiting factor is that the researcher did not consider K-12 education which should be addressed since K-12 education plays a significant role in a student's educational development and, as such, the likelihood of considering a STEM major and completing the degree too.

Lastly, as previously mentioned, "Exploring Factors that Predict STEM Persistence" did not consider institutional practices, which could explain the discrepancy between degree completion and race, even when accounting for gender and income level. It also did not attempt to quantify the influence the factors included in the research had on students from varying socioeconomic backgrounds. Still, the current literature presents riveting and very relevant results that inspired this thesis and were considered throughout the analysis process in conjunction with this study. In particular, one report examined the differences impacting White and Asian women's persistence in STEM in relation to their occupations [93]. While "Committed to STEM?" [93] is not remarkably similar in the research topic, the results from the article indicated different factors impact White and Asian women's career commitment; thus indicating racial bias is present given the study found different factors from social and academic exchanges influenced the White and Asian female students persistence in STEM

differently.

Because the study primarily focuses on the student's point of view and experience, it is difficult to tell if the racial bias indicates systematic failures or cultural differences. Also, the sample size was relatively small and the samples were nationally unrepresentative. It did not consider the potential inconsistencies between gender identities and other racial/ethnic identities (e.g., Black and Latinx students) - a particularly worrisome exclusion considering both gender and race are shown to influence students' outcomes and their social-academic interactions. Specifically, one paper attempted to create a comprehensive analysis of recent literature and to curate all potential factors, from students' educational achievements to institutional conduct, to craft a complete picture of influences on racial underrepresentation in STEM majors [94]. In essence, the researchers found academic readiness, income level, further educational opportunities for a student to explore their interest in STEM, parental occupation, and their parents' educational attainment were all significant predictors when evaluating marginalized communities' representation in STEM majors.

However, the paper conflicted with other results when the researchers stated instructional behaviors were an insignificant signal for a student to choose a science/science-adjacent major in college. Furthermore, the researcher's primary methodology used vote-counting techniques to perform the meta-analysis of the research papers analyzed in the study. The authors did note limitations in the methodology, even stating that it potentially eliminated relevant studies from their comprehensive research. Lastly, the researchers did not make a distinction between racial or ethnic minorities when handling their analysis; given other extensive research, it seems relevant to separate even the minority students because they are systematically treated differently.

In the context of the education system, Critical Race Theory examines how race and White

supremacy impact curriculum, instruction, assessment, and funding and proposes that race (a significant factor in determining inequity) can explain differences in education outcomes and achievement between White students and students of color [95, 96, 97]. Systemic racism is a vicious cycle affecting past, current, and future generations. Considering this, if Black students are being systematically overlooked and ignored, which decreases their self-efficacy and persistence in STEM subjects, they are less likely to major in said topics. In turn, their children will be less likely to study science topics because their parents are not in adjacent careers, and they will also face the same systemic issues that affect their parents. The main goal of this thesis is to understand the relationship between academic, institutional outcomes and student achievement, specifically the role socioeconomic factors play in the relationship, and to quantify the impact of differing demographic backgrounds on students' education attainment.

The first goal is to establish the influence student behaviors and achievements have on educational outcomes, specifically if a student will study STEM in college. The expected outcome would be to observe strong relationships between variables such as the number of advanced placement courses and GPA and the likelihood of a student studying STEM in their post-secondary education. Then, students will be subset by gender and income within their racial demographics to explore the significance wealth and gender have on a student's educational outcome. The author hypothesizes the results will indicate potential gender and income discrimination. The income discrimination could be inadvertent, though, because lower-income families are less likely to have a college (or advanced) degree, and, as such, the students from said families may not have the guidance, access, or money to pursue a degree - despite natural talent or level of institutional support [98].

As for gender discrimination, this will most likely be caused by subconscious sexism, as evidence does suggest teachers are sexist towards students [99, 100]. Similar to subconscious

racism, teachers - like everyone else - likely develop subconscious biases against women due to societal factors [99, 100]. It will likely be concluded there is gender discrimination causing female students to be less likely to want to study STEM in college, with the overarching conclusion most likely being women from varying races are treated differently, and there is more homogeneity in how a student is treated based on race than gender. In other words, Black women are treated more like Black men than White or Asian women. The current literature would suggest race is a bigger factor in treatment, seeing as Asian women, along with Asian men, are overrepresented in STEM [101, 102].

As aforementioned, there is not only a gap in research assessing the potential impact of systemic racism in K-12 education, but the literature review did not find any study that has been completed to quantify its significance on student outcomes. Although there are limitations to the experimental process, the main goal is to provide further understanding of the racial inequities minority students face in the American education system. The research aims to educate policymakers, so more inclusive policies (e.g., the automatic enrollment implemented in Washington) will be used to close the racial inequality gap in education. Furthermore, it could be vital in bridging a gap in understanding the issues Americans from varying ethnic and racial backgrounds face when being educated in the United States.

CHAPTER 3

Data Analysis

3.0.1 Background of Data & Analytic Method

The data utilized in this study is from the 2009 High School Longitudinal Study (HSLS), conducted by the National Center for Education Statistics (NCES) between 2009 and 2016, and is nationally representative. The study collected information from 944 schools and over 23,0000 students regarding institutional behavior (e.g., counselor and teacher expectations), student behavior (e.g., spoke with teacher, counselor, or family about future educational plans), student achievement (e.g., highest math taken in high school), student demographics (e.g. family income), etc. The base year includes information on students during their 9th grade school year with a follow-up in 2013 for their 11th grade year. In 2013/2014, the full high school transcripts became available with a post-high school follow-up in 2016. Over those seven years more than 9,600 factors (variables) were considered and collected.

All statistical analysis will be conducted using the coding languages R and Python along with the statistical software, SPSS.

3.0.2 Research Questions & Hypotheses

Research Questions

1. What is the significance of gender and income across racial/ethnic demographics when considering education attainment?
2. Is the underrepresentation of racial and ethnic minorities in STEM majors potentially

associated with different systemic treatment (e.g. two students perform similarly in a class, but only one is pushed to take on a more advanced course load in the following academic year) in high school?

Hypotheses

The researcher expects the study will conclude institutional behaviors will have a statistically significant impact on student achievement factors influencing the likelihood a student will major in STEM for their bachelor's degree. Furthermore, it is anticipated sex and income will also have a significant impact on a student's achievement and choice in college major. However, the researcher expects, after controlling for meaningful demographic variables, race will still be a significant factor in predicting students' college majors. For instance, students of the same race will have more similarities in their major choice than students of the same sex or income level (e.g., higher income White students will have more similar model metrics with lower income White students than with higher income Asian students).

3.0.3 Exploratory Data Analysis

During the process of visualizing the data (23,503 observations with 9,614 variables), it was important to understand the available information regarding students' socioeconomic status. Seeing as the study is concerned with the influence of income, gender, and race, it was vital to have variables accounting for those demographics. While the survey collected information on all of those factors, since it was longitudinal, the income was recorded at different points in time - 2008 and 2011. Also, the variables accounting for gender (i.e., X4GENDERID - gender identity) were restricted, and so were some of the racial identity variables (e.g., X1ASIAN - student is Asian-composite).

Due to the restrictions on certain demographic variables, other variables were used but with some limitations. For example, the variable X1SEX (student's sex, male or female),

which does not account for trans and other gender-nonconforming people, was used instead. However, since trans and gender-nonconforming people represent a very small portion of the population, it should not significantly hinder the analysis. Similarly, the variable for race utilized in the dataset was X1RACE (student's racial composite), which does indicate whether a student is Latinx or their race, non-Latinx, but makes it infeasible to compare Latinx students by their race. Finally, the variable X1FAMINCOME (total family income in 2008) was used for income and reclassified by 2008 tax brackets. The most significant issue with the variable is it only considers income from 2008 and, considering the recession from 2008-2012 that happened during the student's time in high school, the student's family income could have changed drastically. On the other hand, similar issues arise by using income from a later date, and, in either case, it will not capture the difference between students coming from households that changed income brackets and those that remained in the same income bracket.

Next, a variable was created to identify a student's intended postsecondary major. From the codebook, several variables for students' choice in their college major were selected: X4ENTRYMAJ2Y (major considering for sub-baccalaureate degree), X4ENTRYMAJ4Y (major considering for baccalaureate degree), X4ENTRYMAJ23 (major considering in 23 categories), X4ENTMJCTE (major considering is in a career and technical education field, e.g., information technology), X4ENTMJSTNSF (major considering is in an NSF STEM field), and X4ENTMJST (major considering is in a STEM field).

X4ENTRYMAJ4Y and X4ENTRYMAJ2Y had ten and eleven non-missing classifications (respectively), with the first three denoting computer science, engineering, and science and math, respectively. The remaining eight are non-STEM majors (e.g., classification five is social sciences and humanities. X4ENTRYMAJ23 was a similar variable, with 23 possible non-missing classifications and its first four classes including computer science, engineering,

science, and mathematics. X4ENTMJSTNSF first class corresponded to a STEM major and its second through fourth were non-STEM majors. X4ENTMJST and X4ENTMJCTE had a binary response of "0" (intended major not STEM related) and "1" (intended major designated as STEM). Therefore, a new variable, "STEM Major" was created based on the answers from the aforementioned variables.

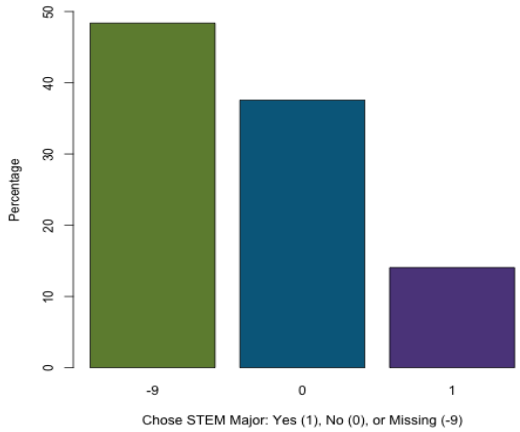


Figure 3.1: Distribution of Majors Considering in Subbaccalaureate 2012 High School Graduates

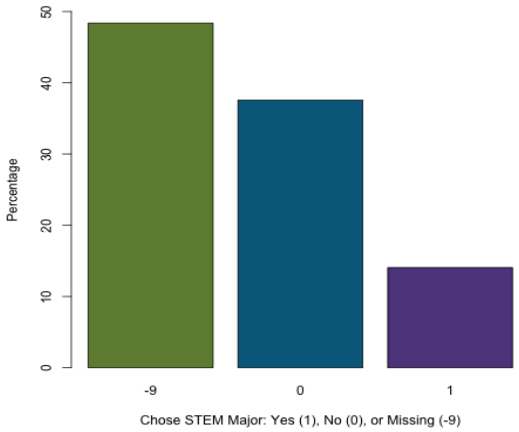


Figure 3.2: Distribution of Majors Considering in Baccalaureate 2012 High School Graduates

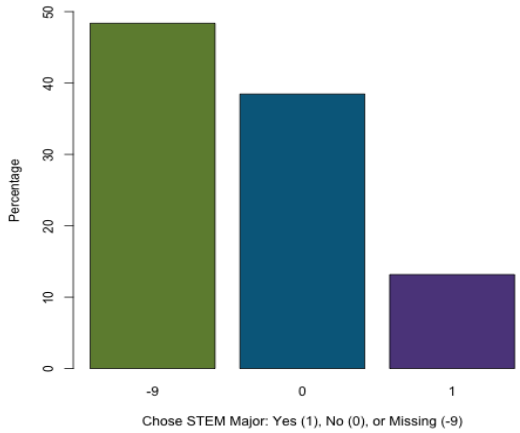


Figure 3.3: Distribution of Majors Considering in 23-categories 2012 High School Graduates

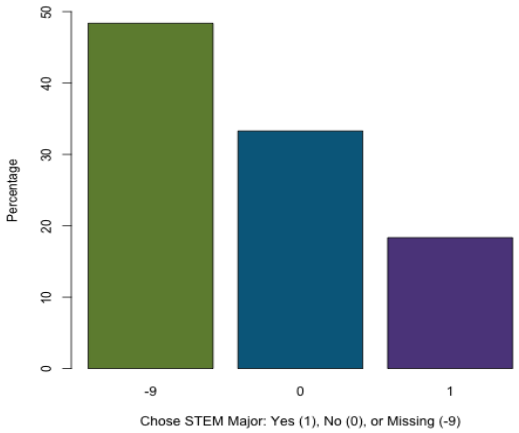


Figure 3.4: Distribution of Majors Considering in NSF Field 2012 High School Graduates

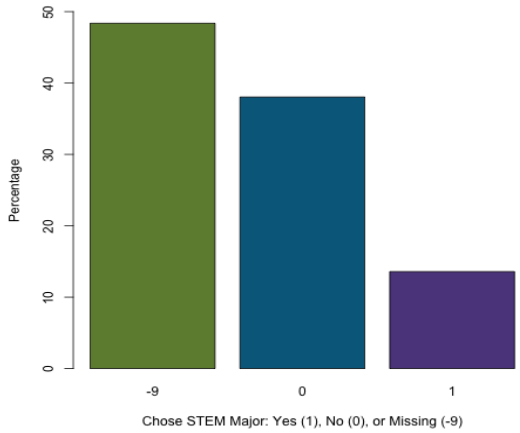


Figure 3.5: Distribution of Majors Considering in STEM Field 2012 High School Graduates

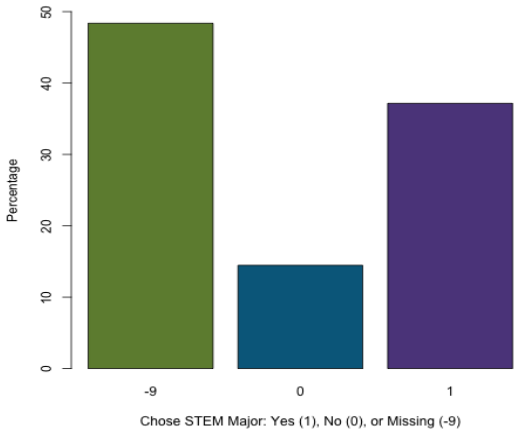


Figure 3.6: Distribution of Majors Considering in CTE Field 2012 High School Graduates

After creating the new socioeconomic variables and a response variable, their respective distributions were examined and some found to be potentially biased or had a noticeable amount of missing values. Although the variables for race and gender were representative and did not have many missing values (4.28% and 0.03%, respectively), the variable for income had almost 30% of its values missing. As for the response variable, "STEM_Major," roughly 55% of students answered the questions, and most of the students indicated they were going to study STEM (Figure 3.7). Since the socioeconomic indicators (income, race, sex) are representative of the national population and do not include many missing values, the analysis should still be informative. However, it is important to note only approximately half of the observations were left after removing missing values from the response variable and the socioeconomic variables.

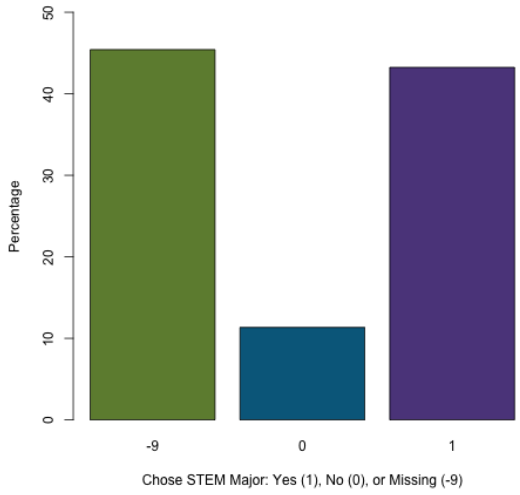


Figure 3.7: Major Distribution

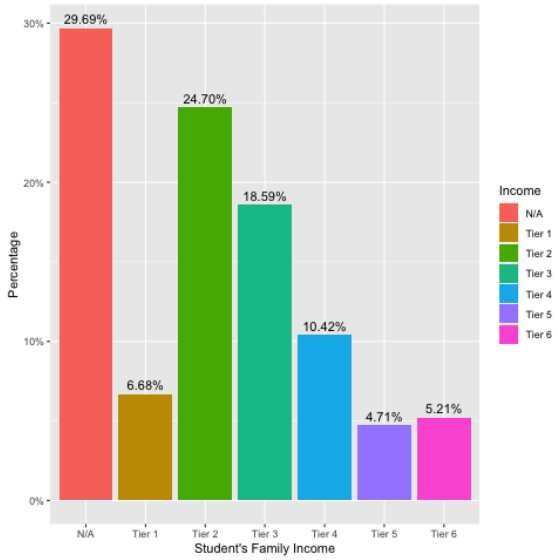


Figure 3.8: Income Distribution

With the primary focus of this study rooted in understanding the relationship between demographic background and a students' intended college major, chi-square testing was completed on the socioeconomic variables (income, race, sex) and college major variable. After performing a chi-square test on the interaction of race and major selection, sex and major selection, income and major selection, race/sex and major selection, race/income and major selection, sex/income, and major selection, and race/sex/income and major choice, the results indicated that all of the relationships are significant. Interestingly, race and income appeared to be the most important of the three demographic variables, which contradicts the hypothesis. However, this may not be the case when proper SES variables are included, so, for now, the hypothesis does not need to be rejected.

In order to establish the relationship between the response and academic performance and the relationship between socioeconomic status (SES) and academic performance, a set of academic achievement variables were selected. Referencing the codebook, quite a lot of

potential variables were found. Specifically, there was a range of variables starting with X3TCREDENG (credits earned in English) and ending with X3TACADTRCK (academic track) measuring academic performance, including GPA, credits earned within subjects, etc. There were additional variables in other parts of the codebook considered, such as X3TXSATMATH (college entrance exam math score in terms of SAT) and X1TXMQUINT (mathematics quantile score), but were not included throughout the analysis because they are not direct measures of academic performance.

Since there were still 118 potential features to use after narrowing the potential variables and removing any restricted variables, relevant variables were strategically selected. Therefore, the ten variables chosen were picked because of their importance in college admissions and potential indication of a student's interest in STEM. The exact variables were X3THIMATH9 (highest level mathematics course taken/pipeline in 9th grade), X3THIMATH (highest level mathematics course taken in high school), X3T1CREDBIOL (at least one credit earned in biology), X3T1CREDCHEM (at least one credit earned in chemistry), X3T1CREDPHYS (at least one credit earned in physics), X3TCREDAPIB (credits earned in AP/IB combined), X3TCREDAPMTH (credits earned in AP/IB math courses), X3TCREDAPSCI (credits earned in AP/IB science courses), X3THISCI9 (highest level science course taken - ninth grade), and X3THISCI (highest level science course taken).

As part of the exploratory data analysis, the author carried out a number of chi-squared tests to examine the two-way relationships between the SES variables and the academic attainment variables along with chi-squared testing between the response and education variables. Initially the variables measuring a student's math competency were examined and found to have an association with the response. All three variables (highest math course taken in grade nine, highest math course taken in high school, and credits earned in AP/IB mathematics courses) had a positive correlation with the response variable (STEM_Major).

However, only the highest math course taken in high school and AP/IB math courses, had a statistically significant relationship with the response. In terms of their distributions, the variables for math course taking in 9th grade and highest math class completed indicated the students were somewhat advanced. 70% of the students were taking the standard or more advanced course for 9th grade math and roughly half taking pre-calculus or higher by the time they finished high school (Figure 3.9 & Figure 3.10) - compared to 35% who had taken at least pre-calculus in 2009 [103]. In contrast, less than 25% had any AP/IB math credits and less than 5% had two or more.

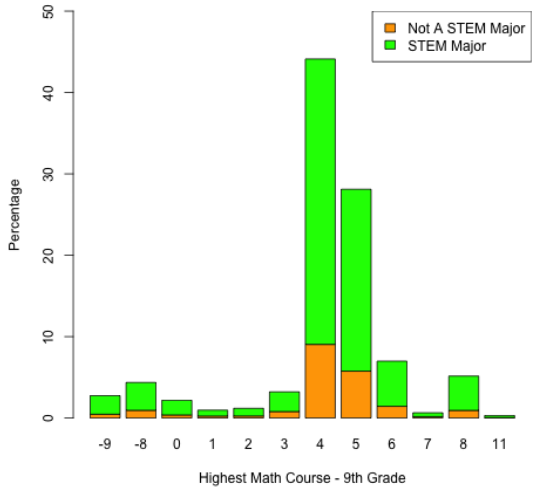


Figure 3.9: Major/Highest Math Course Taken in 9th Grade Distribution

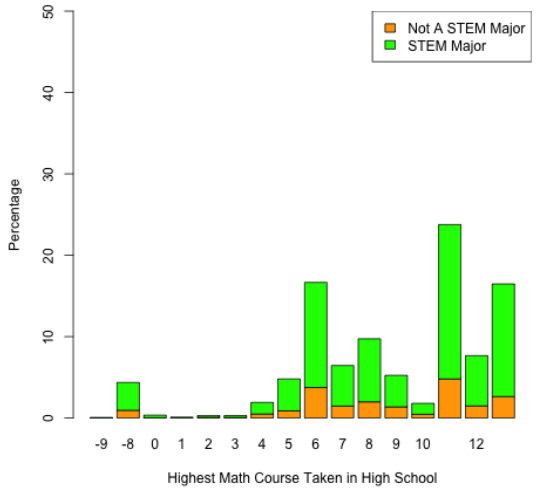


Figure 3.10: Major/Highest Math Course Taken in HS Distribution

Interestingly, the case did not hold when examining X3TCREDAPMTH (credits earned in AP/IB mathematics courses), with over three-fourths having no AP/IB math credits and less than 5% having two or more. In contrast, the distribution of students' major choices in relation to the number of AP/IB math credits a student took did seem more significant than the previous two variables. While the chi-square test confirmed the significance of their

relationship (similarly to the other two math variables), the Cramer's V value showed the association was weaker than for X3THIMATH and STEM_Major (response variable).

Next, students' academic achievement within science was examined and similar conclusions were drawn. The science course a student took in 9th grade was not statistically significant in association with the response, but the highest science course taken in high school was significantly associated. Furthermore, 80% of the students were taking the standard science course in 9th grade, but a larger portion of the highest taken science course was advanced. It is also of note the science class taken later had a stronger, positive correlation with the response than the class taken in 9th grade. With the information previously stated in mind, it was not surprising the variables measuring credits in physics and credits earned in AP/IB science courses were determined to have a positive correlation and statistically significant association with the response variable. However, while the variables accounting for credits in biology and chemistry were positively correlated with the response variable, neither held significant relationships with the variable.

To understand the relationship between the ten academic variables and the demographic variables (sex, income, and race), a similar process as previously done was conducted and it was observed income and race likely had a stronger association with students' academic achievement than sex. The primary observation was a positive relationship between student income and the proportion of students in more advanced courses, so students with higher incomes appeared to be disproportionately represented in higher-level classes. Similarly, race was found to have an impact on the courses a student took. In particular, more advanced classes had a disproportionately high distribution of Asian students and a disproportionately lower number of Black and Latinx students (Figure 3.11 & Figure 3.12). This conclusion was not shocking, given both variables shared a statistically significant relationship with each of the academic performance variables.

However, sex did not have a statistically significant relationship with each of the variables. In particular, the highest math and science course taken in 9th grade and credits earned in chemistry and in AP/IB science courses were all insignificant. It was also concluded sex did share a significant relationship with the highest math and science classes taken in high school. Overall, though, there did not appear to be a perceivable discrepancy with sex-based distribution in course taking, in contrast to the observations found analyzing the relationship with income and race. As such, the results appeared to be indicating income and race were the most influential SES variables, and sex was the least significant.

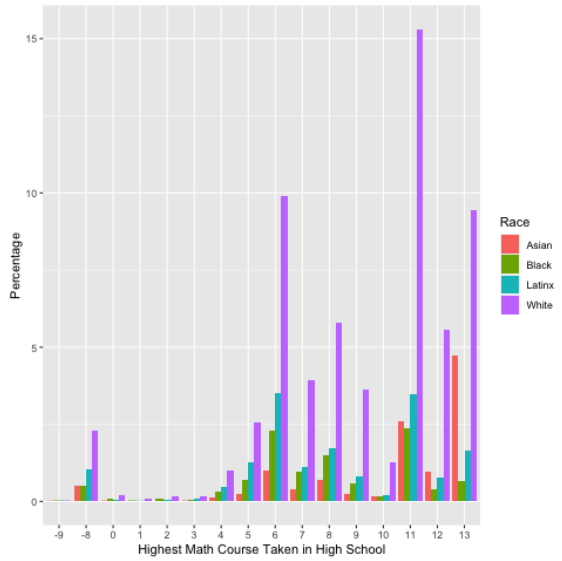


Figure 3.11: Race/Highest Math Course Taken in HS Distribution

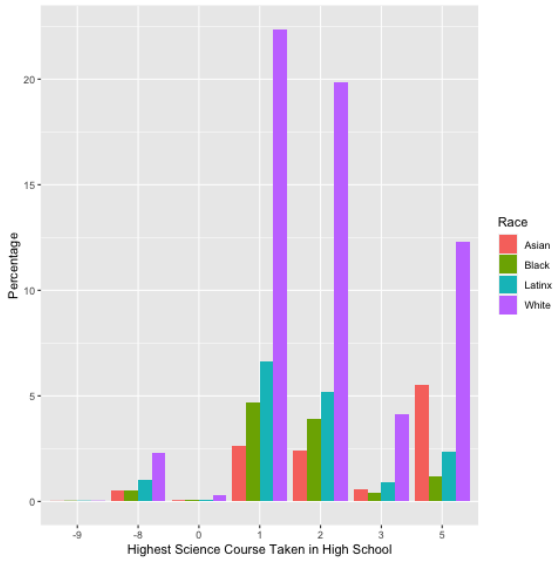


Figure 3.12: Race/Highest Science Course Taken in HS Distribution

A correlation plot (Figure 3.13) was created with all the academic performance variables to understand the dataset better and it was concluded other techniques outside of logistic regression may be more suitable because of multicollinearity issues. The average correlation of the variables was 47% and the model fitted using all the academic variables to predict a

student's intended major did not meet all the assumptions, most notably the assumption of no multicollinearity. While the R-squared value was just below 0.17, these results cannot be interpreted due to the issues with the assumptions not being met with multicollinearity. However, not all variables were found to have such strong relationships with one another, and it cannot be assumed that the highly correlated variables are necessarily useful in a model. Thus, several models were built using a combination of variables with varying degrees of correlation.

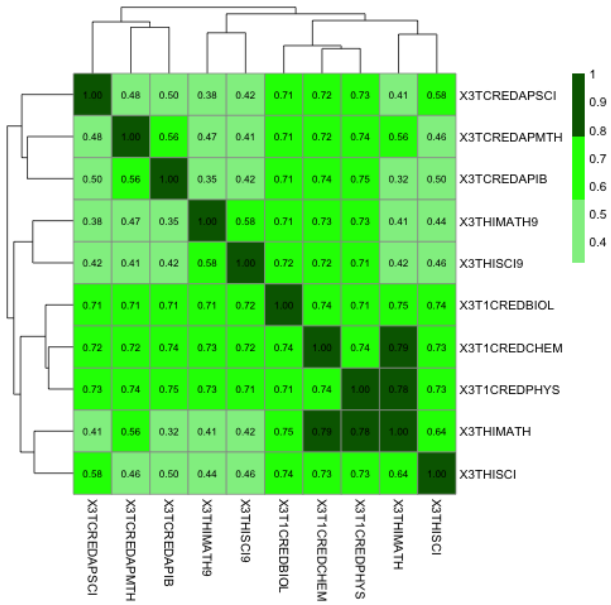


Figure 3.13: Correlation Matrix for Predictors

The three models included the academic variables with lower correlations (less than 0.3), medium correlations (between 0.3 and 0.5), and a combination of both and all of three were relatively uninformative due to multicollinearity issues. The model with low correlations originally contained 30 variables, but 29 were removed because of unacceptable VIF scores. While the remaining variable did have a significant relationship with the SES variables, it did not share a significant relationship with the response variable (student's intended major)

and had an abysmal R-squared of ∼0.002. Similarly, the model for the medium correlated variables was reduced from 98 variables to one because they were found to be insignificant in predicting students' intended college majors.

The model with both low and medium correlated variables had only one variable remaining after removing variables with dependency issues. Since the variable was previously used and found to have a significant relationship with the SES variables, running another chi-square test or analyzing the correlation was unnecessary. Considering the combination of low and medium-correlated pairs of variables yielded a model only using one of the 98 variables and the previous models were not much better, it can be concluded logistic regression likely should not be used to predict the response because of the multicollinearity issues with the explanatory variables.

From the exploratory data analysis process, the author concluded that the socioeconomic variables shared a meaningful association with predicting if a student chose a STEM major for their baccalaureate. However, there were issues with applying logistic regression because of the multicollinearity amongst the non-SES predictor variables. Due to logistic regression relying on an underlying assumption of no multicollinearity, it was concluded models such as Random Forest, Lasso, and Ridge would be more appropriate, as these methods do not rely on the assumption that multicollinearity does not exist.

3.0.4 Model Selection Process

Several modeling approaches were considered to predict the whether a student picked a STEM or non-STEM major. Since the study is focused on analyzing a classification problem and there appears to be a multicollinearity issue, it is more appropriate to use classification methods that deal with variable interactions. In this particular study, the methods of Lasso, Ridge, and Random Forest were used for classification, with PCA used for feature selection.

Lasso & Ridge Without PCA

The first Lasso model containing all the academic variables without the SES variables was created under the binomial distribution (a statistical probability distribution with only two possible outcomes that has an underlying assumption there is one outcome for each trial with each trial having the same probability of success and are independent of one another [104]), using the lambda error (regularization parameter that controls the weight of the penalty on the loss function [105]) within one standard error of the lambda error representing the minimum mean error from cross-validation. Since the data was being classified, classification accuracy, in this case, the percentage of students the model is correctly classifying as STEM or non-STEM majors, was used as the primary metric of comparison; the Lasso model, in particular, had an accuracy rate of approximately 79.8% (Table 3.1). After checking the coefficients, it was determined the Lasso model contained 23 of the original 117 academic performance variables. While the variables were reduced by ∼80%, it does not mean the best model would contain 23 variables or those specific variables. Thus, further analysis was done by using Ridge regression; although Ridge regression will not reduce dimensionality, it can provide insight into the number of relevant variables if the model accuracy is not overwhelmingly better than Lasso.

After creating a Ridge model using all the variables, it was concluded the accuracy of the model was roughly the same as the Lasso model (Table 3.1). Naturally, the coefficients were significantly different, but it did inform the researcher the final model may be able to contain significantly less variables than available in the dataset. However, it is critical to note, while the models' McNemar p-values were significant, the models' p-values testing the difference between the null hypothesis accuracy and alternative hypothesis accuracy was insignificant. In other words, both models have error rates with significant differences between the null and alternative hypotheses, but the difference in the accuracy rates is not statistically significant.

	Accuracy	Precision	Recall	Specificity	F1 Score
Lasso Model	0.80		0.00	1.00	
Ridge Model	0.80		0.00	1.00	

Table 3.1: Lasso & Ridge Models Without PCA

Furthermore, neither model predicted any observations for students who did not choose a STEM major in college (e.g., chose a non-STEM major such as English) and as such the positive prediction value, precision value, and F1 score all were denoted by "NaN" because they could not be calculated. So, although the accuracy of the models was rather high, their overall performances are not particularly good, but the models did give insight into which variables could be of the most importance and the number of anticipated variables worth using. Furthermore, the main focus of the study is on understanding the influence demographic background has on the relationship between academic performance and college majors students desire to take in college. In short, the most important observations will be made based on the influence the SES variables have on the final selected model.

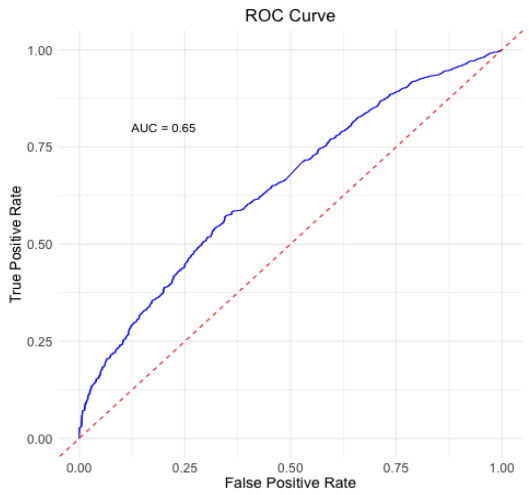


Figure 3.14: ROC Curve Lasso w/o PCA

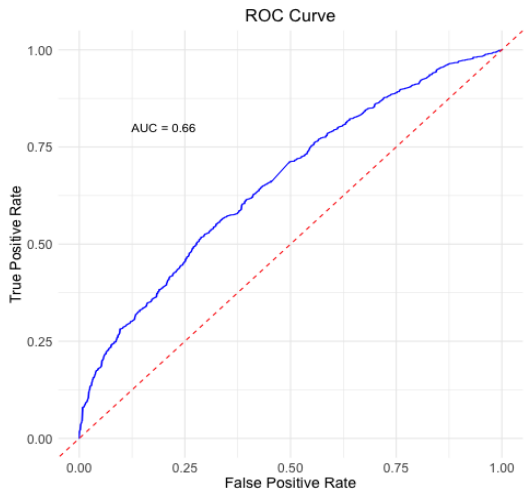


Figure 3.15: ROC Curve Ridge w/o PCA

Lasso & Ridge With PCA

Ultimately, the desired outcome is to have a relatively powerful model with as few variables as possible. Since the Lasso and Ridge models were equally powerful, but the Lasso model containing all variables did not give a non-zero coefficient value for some of the variables, it can be concluded some variables can be removed from the model. Due to the Lasso model only containing 23 variables, principal component analysis (PCA) was used to find the 25 most influential academic performance measures, and then a Lasso and Ridge model was constructed to see if there could be further reduction.

The results indicated that the first four principal components explained almost 85% of the variance and plateaued with the remaining components: the 5th component increases the variance explained by less than 0.01, and the standard deviation does not decrease much either. After using the 25 PCA variables in a Lasso model, the accuracy did not change (as expected), but the Ridge model's accuracy decreased by about $12 \cdot 10^{-3}$. However, the same problems noted in the previous models were also present. The Lasso model did not predict an observation for a student who did not have interest in a STEM major and the McNemar p-values were significant, while the models' p-values testing the difference between the null hypothesis accuracy and alternative hypothesis accuracy were insignificant. As such, neither of the models were particularly useful.

However, the Lasso model with PCA only applied non-zero values on nine of the 25 variables and maintained the same accuracy rate (Table 3.2). As for the Ridge model, it performed relatively the same with the reduction. Therefore, the models seem to be indicating reducing the amount of variables to 25 or less is reasonable. Lastly, because the Ridge model was insignificantly more accurate than the Lasso model, having less than 10 variables in the final model, as desired, seems feasible too.

	Accuracy	Precision	Recall	Specificity	F1 Score
Lasso Model	0.80		0.00	1.00	
Ridge Model	0.80	0.00	0.00	1.00	

Table 3.2: Lasso & Ridge Models With PCA

Random Forest

The last method applied was random forest, using all the academic performance variables and 500 trees to classify the response variable ("STEM_Major", which denotes if a student intended to major in a STEM or non-STEM major). The model was more accurate than any of the previous models, with an accuracy rate of ∼99%, which is noticeably higher than its baseline accuracy of 79%. The primary metrics used to analyze the model's performance were precision (ability of the classifier not to label a negative sample as positive), recall (ability of the classifier to find all the positive samples), specificity (ability of the classifier to find all the negative samples), and F1 score (weighted mean of the precision and recall) [106]. The precision and specificity were 100% with the recall and F1 score both greater than 95% (Table 3.3). Thus indicating the model performed relatively well when classifying students as STEM/non-STEM majors.

Accuracy	Precision	Recall	Specificity	F1 Score
0.99	1.00	0.95	1.00	0.98

Table 3.3: Random Forest All Variables Model

However, the model also used significantly more variables than the Lasso and Ridge models using PCA feature selection, and the original Lasso model applied to the entire dataset. Since the idea is to get less than 10 variables, preferably five, it seemed the next step was to use the information from the models previously generated to reduce the number of variables used in the random forest model in an effort to create a smaller model while retaining as much accuracy as possible.

The final model constructed used six of the variables selected in the 25-variable Lasso model using PCA for feature selection. The six variables selected included the three most positively influential variables (highest physics course, credits in science, and credits in calculus) and three of the variables with the largest negative coefficient values (credits in communication, art, and AP art). These six were also the most impactful variables overall when considering the absolute values of their coefficients (Figure 3.16 & Figure 3.17) . While neither models performed well, the primary reasoning behind using the Lasso with PCA was because it retained a relatively comparable accuracy as all the other models, but used the least amount of variables. As such, the features from the model were applied in a random forest model.

```
/*****************************************************************

Description of six most positively associated variables

*****************************************************************/


Name:       X3TCREDSCI
Label:      X3 Credits earned in: science

Description:
Total Carnegie credits in Life and Physical Sciences.
A Carnegie unit is equivalent to a one-year academic course taken one period
a day, five days a week.

-------------------------------------------------------------------------------
-------------------------------------------------------------------------------


Name:       X3T1CREDCALC
Label:      X3 At least one credit earned in: calculus

Description:
Indicates at least one Carnegie unit in Calculus.
A Carnegie unit is equivalent to a one-year academic course taken one period
a day, five days a week.

-------------------------------------------------------------------------------
-------------------------------------------------------------------------------


Name:       X3THIPHY
Label:      X3 Highest level physics course taken/pipeline

Description:
Highest Physics course.

-------------------------------------------------------------------------------
-------------------------------------------------------------------------------


Name:       X3TCREDCTE
Label:      X3 Credits earned in: CTE

Description:
Credits earned in career and technology classes (e.g., law & business).

-------------------------------------------------------------------------------
-------------------------------------------------------------------------------


Name:       X3THICHEM
Label:      X3 Highest level chemistry course taken/pipeline

Description:
Highest Chemistry course.

-------------------------------------------------------------------------------
-------------------------------------------------------------------------------


Name:       X3TGPAMAT
Label:      X3 GPA: mathematics

Description:
GPA in mathematics coursework.
```

Figure 3.16: Six Most Positively Associated Variables Dictionary

```
/*****************************************************************

Description of six most negatively associated variables

*****************************************************************/


Name:       X3TCREDCOM
Label:      X3 Credits earned in: communication

Description:
Total Carnegie credits in Communications and Audio/Video Technology.
A Carnegie unit is equivalent to a one-year academic course taken one period
a day, five days a week.

-------------------------------------------------------------------------------
-------------------------------------------------------------------------------


Name:       X3TCREDART
Label:      X3 Credits earned in: fine arts

Description:
Total Carnegie credits in Fine and Performing Arts.
A Carnegie unit is equivalent to a one-year academic course taken one period
a day, five days a week.

-------------------------------------------------------------------------------
-------------------------------------------------------------------------------


Name:       X3TCREDAPART
Label:      X3 Credits earned in: AP/IB fine arts

Description:
Total Carnegie credits in AP and IB Fine Arts courses.
A Carnegie unit is equivalent to a one-year academic course taken one period
a day, five days a week.

-------------------------------------------------------------------------------
-------------------------------------------------------------------------------


Name:       X3TCREDSOCST
Label:      X3 Credits earned in: social studies

Description:
Credits earned in social studies.

-------------------------------------------------------------------------------
-------------------------------------------------------------------------------


Name:       X3TCREDAPSS
Label:      X3 Credits earned in: AP/IB social studies

Description:
Credits earned in AP and IB social studies.

-------------------------------------------------------------------------------
-------------------------------------------------------------------------------


Name:       X3TCREDAPENG
Label:      X3 Credits earned in: AP/IB English

Description:
Credits earned in AP and IB English.
```

Figure 3.17: Six Most Negatively Associated Variables Dictionary

X3TCREDSCI	9.160796e-03
X3T1CREDCALC	7.444135e-03
X3THIPHY	6.127189e-03
X3TCREDCTE	5.931645e-03
X3THICHEM	4.841063e-03
X3TGPAMAT	2.829015e-03

Figure 3.18: Six Most Positively Associated Variables

X3TCREDAPENG	-4.311009e-03
X3TCREDAPSS	-4.492184e-03
X3TCREDSOCST	-5.608884e-03
X3TCREDCOM	-8.560454e-03
X3TCREDART	-9.672942e-03
X3TCREDAPART	-2.002083e-02

Figure 3.19: Six Most Negatively Associated Variables

The aforementioned model had an accuracy of approximately 81.2%, which is higher than any of the Lasso and Ridge models. However, it was still quite a drop from the initial 99.02% found in the random forest model containing all the variables for academic performance. As illustrated in the ROCs below (Figure 3.18 & Figure 3.19), the model with all variables for academic performance almost perfectly predicts which students are STEM or non-STEM majors. In contrast, the reduced model with six academic variables has a 74% chance of correctly classifying the students. Nonetheless, due to the impressive dimension reduction (from 117 variables to six variables), the model is more preferable than the model with all the variables. Furthermore, while it is preferable to have the best performing model possible, the most important considerations will be made when making comparisons between different demographic variables impact.

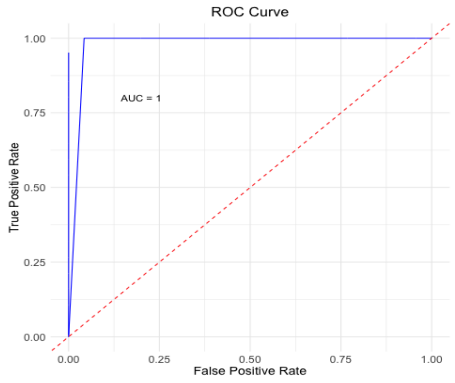


Figure 3.20: Random Forest With All Variables

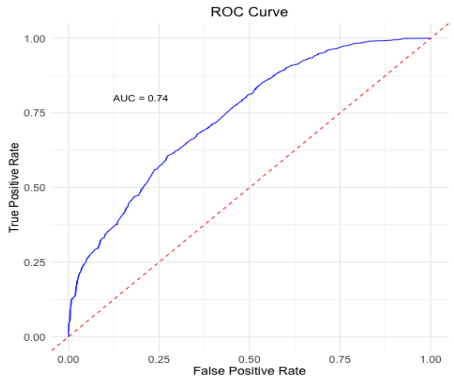


Figure 3.21: Random Forest Six Selected Variables

3.0.5 Final Model Analysis

After selecting a final model, the SES variables were added to the model with the six most important academic variables (highest physics course taken in high school, credits in science, calculus, communication, art, and AP art), together and then separately, so there were a total of four initial models to evaluate. The primary reason this approach was taken was to understand the impact on students' major when accounting for different socioeconomic groups. Comparatively, that meant there would have to be models to compare the overall impact of varying socioeconomic factors (i.e., income, sex, race). Along with several other models to analyze the impact of these SES variables on subgroups (e.g., is there a greater difference in major selection for female students of varying incomes than males of varying incomes).

Table 3.4 containing the model with the SES variables shows that the race and income variables improved the model's accuracy but not the sex variable. When analyzing the models' performance, it was observed only the race model improved the model's precision; similarly, the same held for specificity. However, the specificity for the model only marginally changed, but the recall and F1 scores had the most interesting results. When examining the recall, the value was relatively low, but when including all the SES variables, it increased to 0.4, almost six times the initial value. Following the same trend as the accuracy metric, only race and income improved the model's recall.

Although the recall changes were not as significant for the individual SES models, the F1 score is more important because the data is unbalanced with $\sim$79% of students interested in studying a STEM major. The F1 score observed the same trends: the model including all the SES variables increases the value relative to the base model without demographics, while the race and income variables mildly improve it, and the sex variable decreases the metric. Overall, it seemed the base model was enhanced most by the income variable, with race also

improving the model, but to a lesser extent. However, it appeared the sex variable had a negligible impact on the model, which is interesting given the model with all SES variables performed significantly better than the other models.

```
/************************************************************

Description of base model with SES variables

************************************************************/


Name:       No demographic

Description:
The base model including only the six academic predictors for the
response "STEM_Major"

------------------------------------------------------------------------------
------------------------------------------------------------------------------

Name:       With All Controls

Description:
Base model including all three SES variables (income, race, sex)

------------------------------------------------------------------------------
------------------------------------------------------------------------------

Name:       With Race Only

Description:
Base model including race variable

------------------------------------------------------------------------------
------------------------------------------------------------------------------

Name:       With Gender Only

Description:
Base model including sex variable (substitution for gender)

------------------------------------------------------------------------------
------------------------------------------------------------------------------

Name:       With Income Only

Description:
Base model including income variable
```

Figure 3.22: Model Dictionary

	Model	Accuracy	Precision	Recall	F1 Score	Specificity
1	No Demographic	0.81	0.96	0.07	0.13	1.00
2	With All Controls	0.88	0.98	0.40	0.57	1.00
3	With Race Only	0.82	0.98	0.09	0.16	1.00
4	With Gender Only	0.81	0.96	0.07	0.13	1.00
5	With Income Only	0.82	0.96	0.13	0.22	1.00

Table 3.4: Overall Student Models

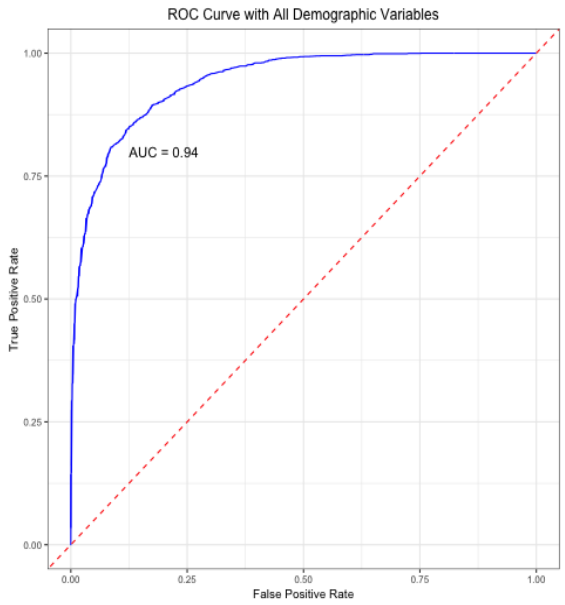


Figure 3.23: ROC Curve All SES Variables

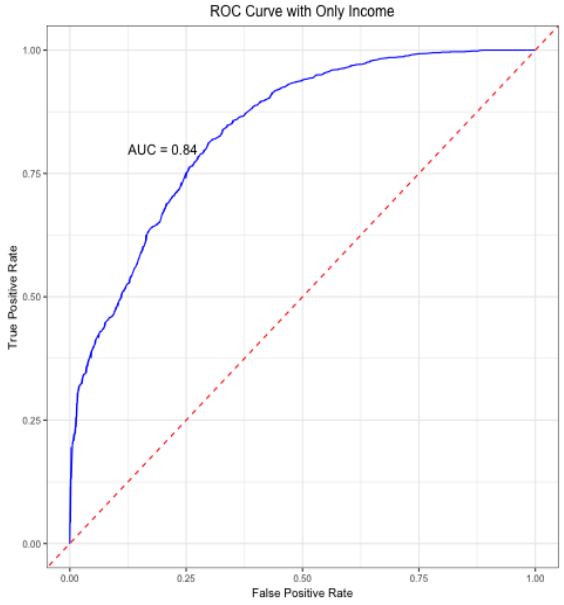


Figure 3.24: ROC Curve With Income Only

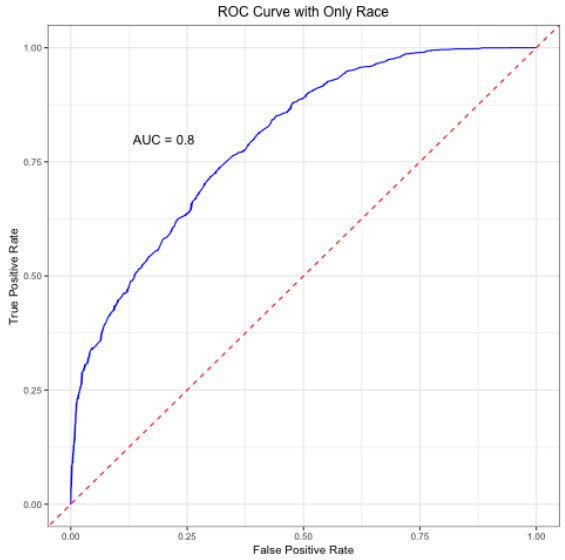


Figure 3.25: ROC Curve With Race Only

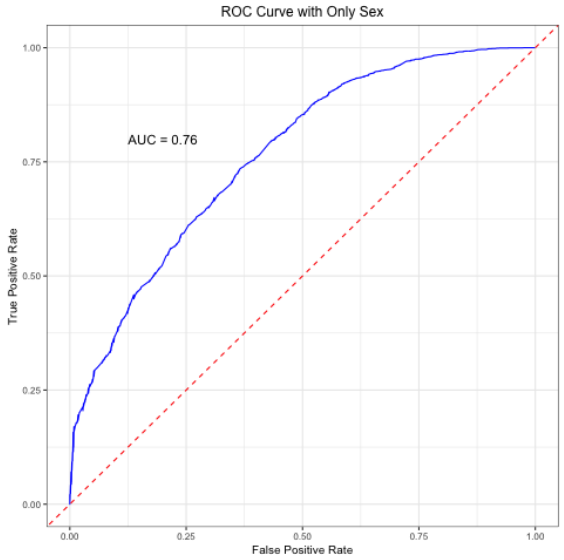
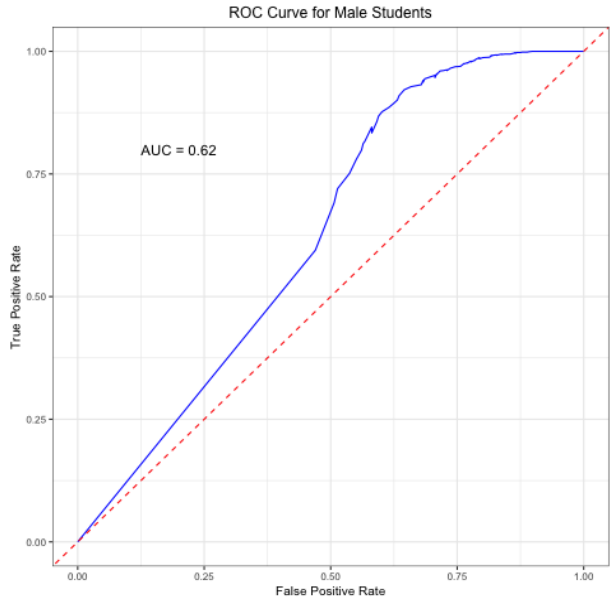


Figure 3.26: ROC Curve With Sex Only

Next, the dataset was separated by sex (male and female), and a model was trained for each group separately (Table 3.5 & Table 3.6). Then, income and race were added to each model, with three models for the male and female groups. Overall, it was observed that the groups had the same trends in terms of impact for the income and race SES variables. However, the trends found among male and female students appeared to be slightly stronger among women than men. For instance, similar to the models with all students, income had a more extensive influence on improving the model than race, but the impact of income and race was more substantial in the female group than male group.



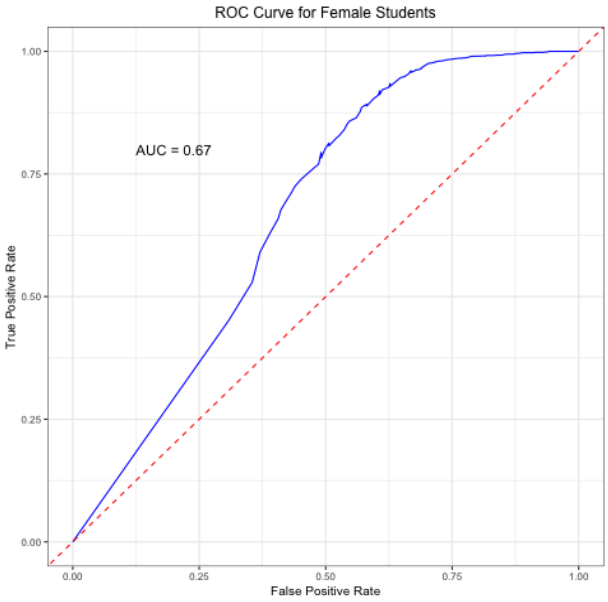


Figure 3.27: ROC Curve for Male Students Figure 3.28: ROC Curve for Female Students

Specifically, the model with only male students was more accurate than was the model comprising only female students (Table 3.5 & Table 3.6). The overall accuracy was better, and the male group's precision was almost 100% compared to 91% for the female students' group. The same did not hold true for the recall, F1 score, and specificity metrics. Notably, the recall and F1 score of the female students were almost double the recall and F1 score

for the male students (Table 3.3). The specificity was higher for female students, but had a negligible difference. Similar observations were concluded for the models including income and the models including race.

Furthermore, for male students, incorporating income into the model increased the model accuracy by 0.7%, but for female students, the increase was more than 2%. The same held true for the recall and F1 score; in particular, the F1 score increased by ∼6.2% and ∼12% (male and female, respectively). Interestingly, it was also observed that the precision and specificity for females changed in both SES models but did not change for male students. Similar conclusions were made when examining the models with the race variable. However, the racial SES variable was less influential than the income variable for both sexes.

For instance, the race SES variable improved the model accuracy for female students by ∼0.4%, and ∼0.2% for male students, but income increased the models' accuracies by ∼1.1% and ∼0.6% (female and male, respectively). The observation above is also an example of how the models for both groups follow similar trends. Further supporting conclusions were made with the recall, F1 scores, and specificity values. In the case of the F1 score, the race variable increased the model representing female students by ∼4.7% and ∼2.6% for the male students' model. However, the precision score did not follow the same pattern as the income results, as the female model's precision increased more than the male model.

	Model	Accuracy	Precision	Recall	F1 Score	Specificity
1	Men	0.82	1.00	0.07	0.13	1.00
2	Men With Income Only	0.83	1.00	0.11	0.20	1.00
3	Men With Race Only	0.83	1.00	0.09	0.16	1.00

Table 3.5: Male Student Models

	Model	Accuracy	Precision	Recall	F1 Score	Specificity
1	Women	0.81	0.91	0.13	0.23	1.00
2	Women With Income Only	0.83	0.98	0.22	0.35	1.00
3	Women With Race Only	0.82	0.94	0.17	0.29	1.00

Table 3.6: Female Student Models

After evaluating the models separated by sex, the students were re-pooled and divided based on the six income brackets - let it be noted no students were in the highest income bracket, tier six (student's family income greater than $195k/year). Initially, the income models were compared by individual tiers and then combined to group by lower-income (tiers one and two), middle-income (tier three), and high-income students (tier four and tier five). For tiers one through four, it was found that the race variable increased the model accuracy more than the sex variable (Table 3.7, Table 3.8, Table 3.9, & Table 3.10). However, for the third and fifth tiers, sex improved model accuracy more than the racial SES variable (Table 3.9 & Table 3.11). In contrast, after grouping the tiers by lower, middle, and higher income, the researcher found race and sex had similar impacts on the base model, while sex was more influential on the middle class and race was more influential on the higher income bracket.

Starting with the income tiers, it was discovered tiers one, four, and five had the lowest accuracy rates, while income tiers two and three had slightly higher accuracies that were approximately the same (Tables 3.7 through Table 3.11). As aforementioned, it was also concluded that the tier one and fours models showed similar effects when including the SES variables. In both models, the racial SES variable had a greater impact than the sex SES variable. However, it appears tier one was the stronger model and faced greater changes in its performance. For instance, while the overall accuracy is slightly higher in the tier one model, the precision for tier one was $\sim$44% and $\sim$36% in the tier four model. The same observation was made for the remaining other metrics, too, and the previously mentioned

results were also relatively consistent in the race SES models and the models including the SES variable for sex.

When considering the sex SES and race SES models, the models were generally better fitted to tier one students (Table 3.7), similar to the base model, and more influenced by the race variable. For example, the precision for the tier one model, including the sex variable, is ∼45%, and its model with the race SES variable has a precision of ∼54%; the precision values for those variables in the tier four model are 40% and ∼48.5%, respectively. The same pattern holds true for the F1 score and specificity metrics, but the opposite was true for the recall metric. However, all the metrics besides accuracy and precision dropped in both tier one and tier four groups after adding the SES variables.

Concerning the models for the fifth income tier (Table 3.11), although the tier one, four, and five models performed relatively the same overall, found the model fitted to students in the fifth income bracket, including sex, was more accurate at predicting student major outcomes than the race SES model. Moreover, it was observed that the sex-based tier five model had a higher specificity value than the race-based tier five model. However, after considering the other metrics, there was a noticeable difference between the recall and F1 scores between the two models, which favored the race-based model. It must also be noted the race-based tier five model had a better recall and F1 score than the base model, but this did not hold for the sex-based model.

Regarding the other two tiers' models (Table 3.8 & Table 3.9), the overall accuracies are clearly better than the previously examined models. It can also be concluded that the tier two models perform more consistently with models one and four, while the third model is more in line with the performances in the tier five models. In cross-comparing models, noticed in the case for recall, while the values for the tier three models were lower, only the model

in tier two was higher than its tier base model's recall, which is also the case for the F1 scores.

	Model	Accuracy	Precision	Recall	F1 Score	Specificity
1	Tier One	0.79	0.44	0.06	0.11	0.98
2	Tier One With Sex Only	0.80	0.45	0.05	0.08	0.99
3	Tier One With Race Only	0.80	0.54	0.06	0.10	0.99

Table 3.7: Tier One Income Models

	Model	Accuracy	Precision	Recall	F1 Score	Specificity
1	Tier Two	0.80	0.61	0.04	0.08	0.99
2	Tier Two With Sex Only	0.80	0.68	0.04	0.08	1.00
3	Tier Two With Race Only	0.80	0.67	0.05	0.09	0.99

Table 3.8: Tier Two Income Models

	Model	Accuracy	Precision	Recall	F1 Score	Specificity
1	Tier Three	0.80	0.53	0.09	0.16	0.98
2	Tier Three With Sex Only	0.80	0.66	0.06	0.11	0.99
3	Tier Three With Race Only	0.80	0.56	0.08	0.14	0.98

Table 3.9: Tier Three Income Models

	Model	Accuracy	Precision	Recall	F1 Score	Specificity
1	Tier Four	0.79	0.36	0.06	0.10	0.97
2	Tier Four With Sex Only	0.79	0.40	0.04	0.07	0.99
3	Tier Four With Race Only	0.80	0.49	0.05	0.09	0.99

Table 3.10: Tier Four Income Models

	Model	Accuracy	Precision	Recall	F1 Score	Specificity
1	Tier Five	0.79	0.35	0.07	0.12	0.97
2	Tier Five With Sex Only	0.79	0.40	0.04	0.07	0.98
3	Tier Five With Race Only	0.79	0.40	0.08	0.13	0.97

Table 3.11: Tier Five Income Models

When grouping students' family income by lower-income (Table 3.12), middle-income (Table 3.13), and higher-income students (Table 3.14), it seemed the lower-income students' model was not impacted after including race and sex. In contrast, the models fitting middle and higher-income students were improved by adding the SES variables. However, the model for higher-income students only improved when including the variable for race, while the middle-income model improved with the race and sex SES variables. In contrast to the higher-income model, the model for middle-income students improved accuracy more when factoring in the SES variable for sex.

In the overall metric comparison between the middle-income students' and higher-income students' models (Table 3.13 & Table 3.14), unsurprisingly, the specificity increased in both the income SES models when analyzing middle-income students. However, for higher-income students, the race model did not show improved specificity, but adding the sex SES variable did increase the metric. A similar conclusion was drawn based on the precision values. The precision increased in both models for middle-income students but more so for the sex SES model. The same held true for students from higher-income households, but for the sex SES model improved the precision by a larger margin than the race SES model. It must also be noted the recall and F1 scores decreased relative to the base model in all the SES variables except for the higher-income F1 score in the race SES model.

In conclusion, comparing across tiers did not show significant differences across the groups, but after grouping income tiers, differences between the groups were found. Both the lower and higher income groups were most influenced by the sex variable. However, while the race variable increased the lower-income model accuracy by a decent portion, it had negligible impact on the accuracy of the high-income student group. Most notably, the most impacted group was the middle class, and the model improved most with the race SES variable; this model had the largest shift/improvement in accuracy compared to the other model changes.

It must be acknowledged that the improvement was not necessarily substantial as it was approximately a 2% increase in the overall accuracy.

	Model	Accuracy	Precision	Recall	F1 Score	Specificity
1	Lower Income	0.80	0.62	0.05	0.10	0.99
2	Lower Income With Sex Only	0.80	0.68	0.04	0.08	1.00
3	Lower Income With Race Only	0.80	0.62	0.05	0.10	0.99

Table 3.12: Lower Income Students' Models

	Model	Accuracy	Precision	Recall	F1 Score	Specificity
1	Middle Income	0.80	0.53	0.09	0.16	0.98
2	Middle Income With Sex Only	0.80	0.66	0.06	0.11	0.99
3	Middle Income With Race Only	0.80	0.56	0.08	0.14	0.98

Table 3.13: Middle Income Students' Models

	Model	Accuracy	Precision	Recall	F1 Score	Specificity
1	Higher Income	0.80	0.62	0.07	0.13	0.99
2	Higher Income With Sex Only	0.80	0.67	0.05	0.10	0.99
3	Higher Income With Race Only	0.80	0.62	0.07	0.13	0.99

Table 3.14: Higher Income Students' Models

Moving forward with the analysis, the final step was to generate models for the income SES variable and sex SES variable after grouping the students by race/ethnicity (Asian, Black, Latinx, and White). For students from Asian backgrounds, sex improved the model's accuracy by a larger margin than income, but for White, Latinx, and Black students, the opposite held true. However, the model accuracies of Asian and Black students were more aligned with each other than the other two model groups, and vice versa held true for the models fitting Latinx and White students.

	Model	Accuracy	Precision	Recall	F1 Score	Specificity
1	Asian Students	0.79	0.44	0.10	0.17	0.97
2	Asian Students With Income Only	0.79	0.46	0.12	0.19	0.96
3	Asian Students With Sex Only	0.79	0.44	0.09	0.15	0.97

Table 3.15: Asian Students' Models

	Model	Accuracy	Precision	Recall	F1 Score	Specificity
1	Black Students	0.79	0.38	0.04	0.07	0.98
2	Black Students With Income Only	0.80	0.48	0.05	0.08	0.99
3	Black Students With Sex Only	0.80	0.44	0.03	0.06	0.99

Table 3.16: Black Students' Models

	Model	Accuracy	Precision	Recall	F1 Score	Specificity
1	Latinx Students	0.80	0.64	0.05	0.10	0.99
2	Latinx Students With Income Only	0.81	0.66	0.07	0.13	0.99
3	Latinx Students With Sex Only	0.80	0.62	0.04	0.08	0.99

Table 3.17: Latinx Students' Models

	Model	Accuracy	Precision	Recall	F1 Score	Specificity
1	White Students	0.81	0.74	0.07	0.13	0.99
2	White Students With Income Only	0.81	0.84	0.10	0.18	1.00
3	White Students With Sex Only	0.81	0.80	0.05	0.09	1.00

Table 3.18: White Students' Models

When comparing the models for Asian and Black students (Table 3.15 & Table 3.16), find that the Black students' models performed in the overall accuracy and specificity metrics. However, the F1 scores in Asian students' models were almost twice those in Black students' models. Interestingly, in both groups noticed that the F1 scores were weaker in the sex SES model than in the income SES models and were worse than their both base models' F1 scores. Also, both the precision and recall metrics are better in the models for Asian

students than the models for Black students. On the other hand, the same conclusions noted about the F1 scores held true for the recall values. It was also noted the Asian students' models increased precision more so with the sex SES variable than the income SES variable, while the opposite held true for the Black students' models.

Conversely, the models for White and Latinx students (Table 3.17 & Table 3.18) followed more similar patterns as one another compared to the Asian and Black models. In both cases, notice including the sex SES variable decreases the overall model accuracy but the accuracy increases when including the income variable. The F1 scores follow in a similar manner, with the improvement coming in the income SES model but not the sex SES model. However, the specificity dropped in the Latinx model when including income and increased with the sex variable; for White students, both models showed an increase in the metric. Regarding the precision for the models, it was found Latinx students' base model increases the precision with the income variable but decreases in the sex model; on the other hand, in both models for White students, the metric improves. The models both have a similar result with the recall: it decreases when including the sex variable but increases when including income. Despite possessing similar patterns, the performance of the White students' model is superior overall.

Furthermore, when looking across all the models, notice the F1 scores are highest in the Asian and White students' models (Table 3.15 & Table 3.18). Thus, it is likely the models for the Asian and White students are the best despite the accuracy favoring Latinx and White students because F1 is more important in this case due to the data imbalance. Although the base model for Asian students is the best in terms of F1 score, the model including income for White students has the highest F1 score overall. Moreover, it must be noted that the largest decrease in F1 score was found in the sex SES model for White students. Conversely, the smallest decline in model F1 scores was found in the sex SES model for Black students,

and the smallest increase to a base model's F1 score was in the income SES models for Asian and Black students.

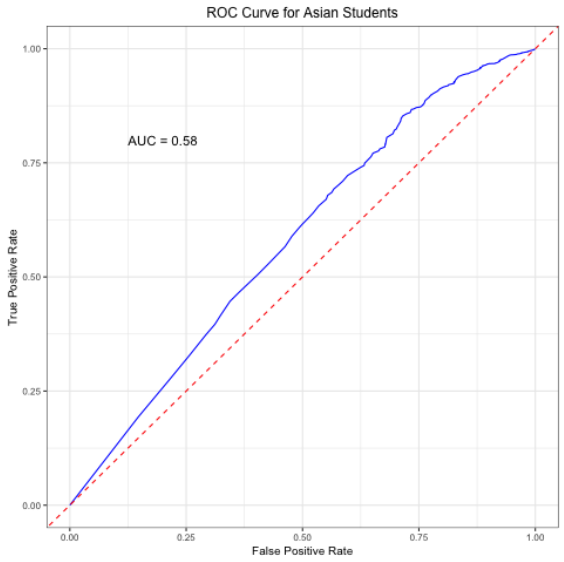


Figure 3.29: ROC Curve Asian Students

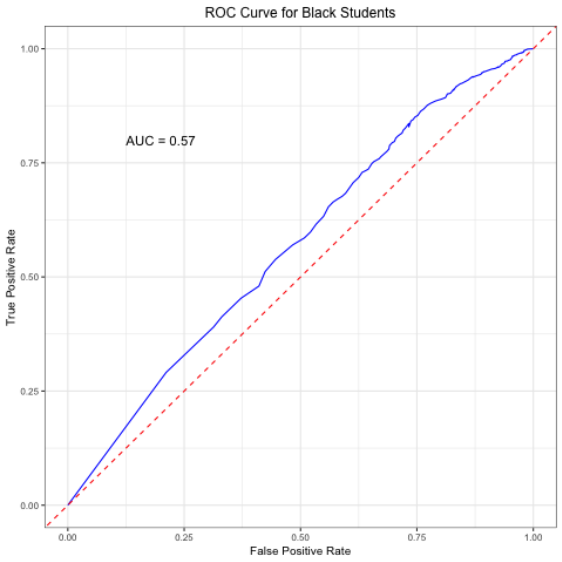


Figure 3.30: ROC Curve Black Students

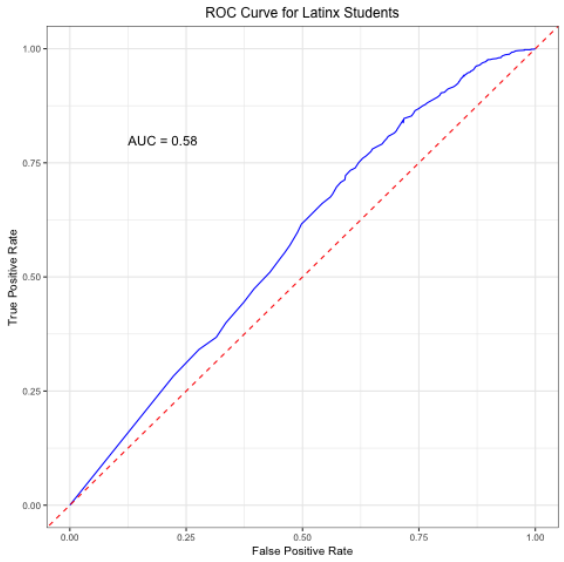


Figure 3.31: ROC Curve Latinx Students

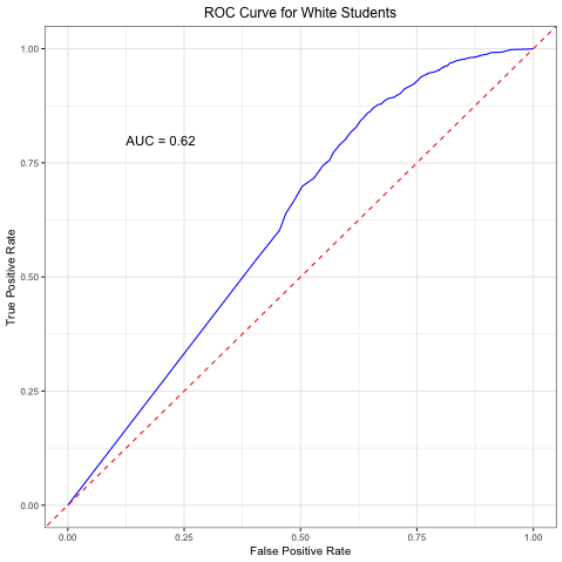


Figure 3.32: ROC Curve White Students

CHAPTER 4

Summary

Although the initial observations seemed to indicate the most influential SES variable was likely income, further analysis detailed more complex results. The primary conclusion was different socioeconomic factors have varying influences on students. While the exact causes of these discrepancies cannot be deduced from the research, it can be concluded students are likely treated differently based on their socioeconomic profiles. As such, the researcher theorized the differences seen across groups was a reflection of inner group privilege: the idea that within a given demographic, some individuals are more privileged than others.

From the initial model analysis, it appeared income was the most important demographic factor when predicting if a student was considering studying a STEM major upon starting their postsecondary education. Interestingly, after including the race and sex SES variables in the income models, it was found the only tiers impacted by either variable was tiers two and five - both had their F1 scores increase by 1%. In all other models, race lowered the F1 score, but sex lowered the F1 score for all the tiers' models. Lastly, it was concluded in the models where both race and sex lowered the F1 score, sex lowered the F1 score by a larger margin. The difference in F1 score is a potential indicator of systemic bias as it implies students with similar household incomes and education attainment have different outcomes when considering their sex and race. Thus, it was concluded, when considering students from the highest and lowest income brackets, race is influential because it increases the models' performance.

The models disaggregated by sex were interesting given the initial models indicate sex was the least important demographic measure between the three socioeconomic variables. While it was initially noted the model for male students had a better overall accuracy than the model for female students, the models for females improved by a larger margin than the male students' model when accounting for the other SES variables - income and race. The male model's F1 score increased by 7% after including income into the model, but decreased by 1% when including race in the model. Conversely, the female students' model's F1 score increased by 12% and 6%, respectively, after including income and race. Therefore, it seems while income impacts both groups choice in college major, race only is influential for female students' outcomes.

Similar conclusions were made when examining students based on their racial and ethnic socioeconomic status. The White students' model was observed to have a boost of 5% in their F1 score when considering their family income bracket and a 4% decrease in their F1 score after accounting for sex. On the other hand, both Asian and Latinx students saw an increase of 3% in their F1 scores when including race in their models, but had a decrease of 2% when accounting for their sex. However, Black students only had a 1% increase in their F1 score when accounting for their household income and a decrease of 1% after including the sex SES variable. In short, it was concluded for White students income is a more important factor than sex when predicting if they are considering a STEM major for college. The same was observed to be the case for non-Black POC, but to a lesser extent. Conversely, Black students appear to have relatively similar outcomes regardless of their income or sex status.

The final conclusions were income was an influential privilege for non-Black students and had negligible impact on Black students' outcomes. Although it was an important resource for non-Black students, it was also clear it was more impactful for White students than non-Black students of color. Furthermore, it was also noted sex was more significant for

understanding Black students' outcomes than non-Black students, but ultimately sex was not indicative that a student was or was not considering a STEM major. In other words, it appears that White students were least impacted by their sex when predicting if they were considering studying a STEM major and income was a more important privilege for their outcomes. Similarly, it was observed that non-Black students of color also benefited from having a higher income and their sex was not as important when determining their outcomes than for Black students.

As such, the researcher theorized White students have the largest advantage in accessing STEM when obtaining higher incomes, while Black students access is the same regardless of their income or sex. Thus, it appears to be the case that while non-Black students can benefit from higher income brackets and are less impacted by their sex, Black students gain negligible advantages with access to wealth and receive similar treatment regardless of their sex. Other research supports the hypothesis that Black children do not necessarily benefit much from more income, as Black boys raised in the wealthiest families still earn less than White boys of similar backgrounds and are more likely to become poor than stay wealthy, while "White boys who grow up rich are likely to remain that way" [107]. Furthermore, being an underrepresented minority is seemingly a bigger disadvantage academically than being poor, with even the wealthiest and highest earning Black and Latinx students having significant disparities in college completion rates [108][109]. Given people wanting college admissions to consider income over race, it made the researcher consider further implications of inner group privilege.

Relative to the current literature, there were several papers that had similar findings such as Xueli Wang's 2013, which showed academic achievements as critical factors in a student's choice to study STEM [84]. A paper in 2021 also came to similar conclusions, but was focused on engineering majors specifically [85]. Although the final models in this study did

not include overall GPA and ACT score, Mau's paper in 2016 also concluded Asian students, racial/ethnic minorities and female students were less likely to major in a science field than their White and male peers, respectively [91]. The paper most similar to this study also included demographic information (e.g., income, race, and sex), but used a classification and regression tree (CART). While the researchers also found academic achievements to be the most important variables in predicting choice, their variables included "calculus credits, science identity, total STEM credits, and math achievement" [90]. In contrast, the final models in this study included credits in science, calculus, physics, art, and communication. The discrepancy between the two outcomes likely has to do with the different approaches, as the researchers used the CART method and this study applied random forest on variables selected via Lasso, Ridge, and PCA.

It must also be noted variables were selected based on their absolute estimate values, meaning variables with negative impacts on the likelihood of a student selecting a STEM major (i.e., art and communication) were also included. As such, it is aligned with previous conclusions that students with more STEM credits in high school are more likely to study STEM at the postsecondary level, with the inverse relationship applying too. If the approach was to include all variables with positive associations with a student selecting a STEM major, then the variables would include credits in science, calculus, chemistry, highest physics course, and math GPA, not very dissimilar to their variables. Lastly, one of the primary differences between the method used in this study and previous research is the disaggregation across subgroups. As such, it is not feasible to compare previous studies' findings across various demographics relative to this study.

As access to higher education continues to become growingly difficult, many are pointing fingers to wealth inequality, but this appears to be a larger indicator for non-Black students and most importantly White students. Considering many take issue with race-based college

admissions, it seems the dominant group (non-Black students) among students are having their issues pushed to the forefront. This issue of inner group privilege is not just seen within students, but other groups too - e.g., White women within the feminist movement, Black male issues being the face of Black strife [110] [111]. At the same time, given that the model containing all three of the SES variables performed the best, it seems outcomes for students are quite relevant to their demographic background. Also, while the research acknowledges the curious findings, it must also be noted the study did have limitations.

Most notably using a sex variable instead of a gender variable, having a stagnant income variable, and the inability to examine Latinx students by race, which is critical given research suggests it impacts their level of discrimination [112], were a few obstacles in the process. It is also important for the reader to note the missing responses for the income and major variables, as these were likely non-random. In both cases, there were also significant portions of the students who did not respond (30% and 55% missing, respectively). Another thing with the major variable is that one of the sub-variables included (X4ENTMJCTE - major considering in CTE field) had issues with its definition. While a CTE field does include STEM-adjacent fields such as information technology, it also includes non-STEM fields such as law and medicine.

Ideally, further research should be conducted to expand on the limitations noted and to analyze institutional behaviors as opposed to students' academic performance. Thus, more concrete conclusions can be made about the varying systemic treatment of students based on different socioeconomic indicators and the impact it has on students' future educational decisions.

References

[1] Hope Wabuke. 'Caste' Argues Its Most Violent Manifestation Is In Treatment Of Black Americans. *NPR*, 2020.

[2] Gina Torino. How racism and microaggressions lead to worse health. *USC Annenberg: Center for Health Journalism*.

[3] Monica T. Williams. Microaggressions: Clarification, Evidence, and Impact. *Sage Journals*, August 2019.

[4] A. L. Skinner-Dorkenoo, A. Sarmal, and et. al. How Microaggressions Reinforce and Perpetuate Systemic Racism in the United States. *Perspectives on Psychological Science*, 2021.

[5] John T. Jost. Working class conservatism: a system justification perspective. *Current Opinion in Psychology*, 18, 2017.

[6] Jeannine A. Bertin. How Unjust Systems Are Created and Upheld. *Psychology Today*, 2021.

[7] Juliana Horowitz, Anna Brown, and Kiana Cox. Race in America 2019. *Pew Research*, April 2019.

[8] Elijah Anderson. Black Success, White Backlash. *The Atlantic*, 2023.

[9] Elizabeth Hinton and DeAnza Cook. The Mass Criminalization of Black Americans: A Historical Overview. *Annual Review of Criminology*, 2021.

[10] Marina Affo. The complex history of alexander twilight, nation's first african american to earn a bachelor's degree. *USA Today*, February 2021.

[11] Steven Bradt. 'one-drop rule' persists. *Harvard Gazette*, December 2010.

[12] The history of wesleyan college. *Wesleyan College*.

[13] Little known facts about black history. *Biography*, 2021.

[14] Plessy v. ferguson (1896). *National Archives*.

[15] Plessy v. Ferguson. *History*, January 2023.

[16] Alisa B. Katz. Op-ed: Modern california is all about voter access. history reveals a far grimmer voting-rights past. *Los Angeles Times*, November 2020.

[17] Caitlin Snook. The racial integrity act, 1924: An attack on indigenous identity. *National Park Services*, June 2023.

[18] Marc Tucker. Separate But Equal: It Wasn't Then, It Isn't Now. *Education Week*, October 2016.

[19] Jean Van Delinder. A landmark case unresolved fifty years later. *National Archives*, 36(1), December 2022.

[20] Sonya Ramsey. The troubled history of american education after the brown decision. *Organization of American Historians*, February 2017.

[21] Raymond Pierce. The racist history of "school choice". *Forbes*, May 2021.

[22] Aditya Aladangady and Akila Forde. Wealth Inequality and the Racial Wealth Gap. *Federal Reserve*, October 2021.

[23] Pay Scale. 2023 gender pay gap report. 2023.

[24] Terry Gross. A 'Forgotten History' Of How The U.S. Government Segregated America. *NPR*, May 2017.

[25] Laura Meckler and Kate Rabinowitz. America's schools are more diverse than ever. But the teachers are still mostly white. *The Washington Post*, December 2019.

[26] Madeline Will. Teachers Are as Racially Biased as Everybody Else, Study Shows. *Education Week*, June 2020.

[27] Shervin Assari. General Self-Efficacy and Mortality in the USA; Racial Differences. *National Library of Medicine*, October 2016.

[28] William J. Barber II. The Racist History of Tipping. *Politico*, July 2019.

[29] Mark Karlin. Reclaiming Labor History: How Domestic Workers Resisted Racism in the '60s and '70s. *truthout*, October 2015.

[30] Christopher Rim. The supreme court may change the landscape of ivy league feeder schools. *Forbes*, November 2022.

[31] Michael T. Nietzel. Legacy College Admissions Come Under Fire In New Report. *Forbes*, October 2022.

[32] Sally Kohn. Affirmative Action Has Helped White Women More Than Anyone. *Time*, June 2013.

[33] Jessica Guynn. White women benefit most from affirmative action. So why do they oppose it? *USA Today*, June 2023.

[34] Sophie Gardner. What Women Have Gained From Affirmative Action. *Politico*, June 2023.

[35] Tayo Bero. Affirmative action is over in the United States, but only for Black people. *The Guardian*, June 2023.

[36] Siva Kumari. Supreme Court affirmative action ruling is unjust, unnecessary. *College Possible*, June 2023.

[37] Katherine Schaeffer. U.S. Congress continues to grow in racial, ethnic diversity. *Pew Research*, January 2023.

[38] U.S. Census Bureau.

[39] Faith Karimi. In the nearly 232-year history of the US Senate there have only been 11 Black senators. *CNN*, January 2021.

[40] Cheyanne M. Daniels. Only three Black governors have ever been elected in US history. *The Hill*, November 2022.

[41] 36 *Intelligent*, June 2023.

[42] Kaiser Family Foundation. Poverty Rate by Race/Ethnicity. 2021.

[43] Sheyahshe Littledave. The Big Money. *Topic*, February 2019.

[44] Ryan P. Smith. How Native American Slaveholders Complicate the Trail of Tears Narrative. *Smithsonian Magazine*, March 2018.

[45] Rashawn Ray and Andre M. Perry. Why we need reparations for Black Americans. *Brookings*, April 2020.

[46] Khaing Zaw, Darrick Hamilton, and William Darity Jr. Race, wealth and incarceration: Results from the national longitudinal survey of youth. *Springer*, February 2016.

[47] United States Holocaust Memorial Museum. Classification system in nazi concentration camps. *Holocaust Encyclopedia*.

[48] Tyler Bamford. African americans fought for freedom at home and abroad during world war ii. *National WWII Museum*, February 2020.

[49] William L. Katz. Africans and Indians: Only in America. February 2007.

[50] Jacob Jarvis. After reparations study suggests $151 million for each african american, experts say money alone isn't enough. *Newsweek*, July 2020.

[51] Carrie Blazina and Kiana Cox. Black and white americans are far apart in their views of reparations for slavery. *Pew Research*, November 2022.

[52] U.S. Department of Education. 2015–16 civil rights data collection stem course taking. 2018.

[53] G. V. Larnell, E. C. Bullock, and C. C. Jett. Rethinking teaching and learning mathematics for social justice from a critical race perspective. *Sage Journals*, December 2017.

[54] Erika C. Bullock. Only stem can save us? examining race, place, and stem education as property. *Research Gate*, September 2017.

[55] Natalie S. King, Zach Collier, and et al. Determinants of Black families' access to a community-based STEM program: A latent class analysis. *Wiley Online Library*, July 2021.

[56] Joseph Guzman. 90 percent of Americans believe racism, police brutality are problems in the US, poll finds. *The Hill*, July 2020.

[57] Vianney Gomez. US public continues to view grades test scores as top factors in college admissions. *Pew Research*, April 2022.

[58] James S. Murphy. The Real College Admissions Scandal. *Slate*, June 2021.

[59] Jere Downs. Is There a Lack of Diversity in Private Schools? *U.S. News*, December 2021.

[60] Southern Education Foundation. A History of Private Schools and Race in the American South.

[61] Sequoia Carrillo and Pooja Salhotra. The U.S. student population is more diverse, but schools are still highly segregated. *NPR*, 2022.

[62] Maria Temming. Stem's racial, ethnic and gender gaps are still strikingly large. *Science News*, 2021.

[63] Betsy Ladyzhets. These 6 graphs show that Black scientists are underrepresented at every level. *Science News*, December 2020.

[64] Margery S. Sendze. I can't quit: Experiences of black women in stem professions. *Sage Journals*, 2022.

[65] Melissa Suran. Keeping black students in stem. *PNAS*, 2021.

[66] Jon Valant. The banality of racism in education. *Brookings*, 2020.

[67] Ned Tilbrook. Field-specific cultural capital and persistence in college majors. *Elsevier*, 2021.

[68] Mary A. Fox and Angelina Kewalramani. Status and Trends in the Education of Racial and Ethnic Groups. *NCES*, 2021.

[69] Natasha Warikoo, Stacey Sinclair, and et. al. Examining Racial Bias in Education: A New Approach. 2016.

[70] Tasminsa K. Dhaliwal, Mark J. Chin, and et. al. Educator bias is associated with racial disparities in student achievement and discipline. *Brookings*, 2020.

[71] Ember Smith and Richard V. Reeves. Sat math scores mirror and maintain racial inequity. *Brookings*, December 2020.

[72] Stem Median Wage and Salary Earnings. *National Center for Science and Engineering Statistics (NCSES)*, (Section 5), 2023.

[73] Lisa C. McKay. How the racial wealth gap has evolved—and why it persists. *Federal Reserve Bank of Minneapolis*, October 2022.

[74] David DesRoches. Georgetown study: Wealth, not ability, the biggest predictor of future success. *Conneticut Public Radio*, May 2019.

[75] Eric Lichtenberger and Casey George-Jackson. Predicting High School Students' Interest in Majoring in a STEM Field: Insight into High School Students' Postsecondary Plans. *Journal of Career and Technical Education*, 28(1), 2013.

[76] Svetlana Chachashvili-Bolotin, Marina Milner-Bolotin, and Sabina Lissitsa. Examination of Factors Predicting Secondary Students' Interest in Tertiary STEM Education. *K16 Diversity*, 38(3):336–390, 2016.

[77] Krista Kaput. Make Rigorous Education the Default through Automatic Enrollment. *Ed Allies*, February 2021.

[78] Education Comission of the States. Teacher Expectations of Students. 13(6), December 2012.

[79] E. J. Smith and S. R. Harper. Disproportionate impact of k-12 school suspension and expulsion on black students in southern states. *Penn Graduate School of Education*, 2015.

[80] Lea Winerman. For Black students, unfairly harsh discipline can lead to lower grades. *American Pyschological Association*, October 2021.

[81] Abiodun Raufu. School-to-Prison Pipeline: Impact of School Discipline on African American Students. *Journal of Education Social Policy*, 7(1), March 2017.

[82] Anne Gregory, Russell J. Skiba, and Pedro A. Noguera. The Achievement Gap and the Discipline Gap: Two Sides of the Same Coin? *Sage Journals*, 39(1), January 2010.

[83] ACLU. School-to-Prison Pipeline.

[84] Xueli Wang. Why students choose stem majors: Motivation, high school learning, and postsecondary context of support. *Sage Journals*, 50(5), October 2013.

[85] Li Tan, Joyce B. Main, and Rajeev Darolia. Using random forest analysis to identify student demographic and high school-level factors that predict college engineering major choice. *Journal of Engineering Education*, June 2021.

[86] Teng Zhao and Lara Perez-Felkner. Perceived abilities or academic interests? longitudinal high school science and mathematics effects on postsecondary stem outcomes by gender and race. *International Journal of STEM Education*, June 2022.

[87] Nancy N. Heilbronner. Stepping onto the stem pathway: Factors affecting talented students' declaration of stem majors in college. *Sage Journals*, November 2011.

[88] Amanda L. Griffith and Joyce B. Main. First impressions in the classroom: How do class characteristics affect student grades and majors? *Elsevier*, April 2019.

[89] Linda Darling-Hammond. Teacher Quality and Student Achievement: A Review of State Policy Evidence. *EPAA*, 8(1), January 2000.

[90] Chi-Ning Chang, Shuqiong Lin, and et. al. Predicting STEM Major Choice: a Machine Learning Classification and Regression Tree Approach. *Journal for STEM Educucation Research*, 2023.

[91] Wei-Cheng J. Mau. Characteristics of us students that pursued a stem major and factors that predicted their persistence in degree completion. *Universal Journal of Educational Research*, 2016.

[92] Mario I. Suárez and et. al. Exploring factors that predict stem persistence at a large, public research university. *International Journal of Higher Education*, 10(4), January 2021.

[93] Catherine Riegle-Crumb, Menglu Peng, and Tatiane Russo-Tait. Committed to stem? examining factors that predict occupational commitment among asian and white female students completing stem u.s. postsecondary programs. *Springer*, March 2019.

[94] Martha C. Bottia, Roslyn A Mickelson, and et. al. Factors associated with college stem participation of racially minoritized students: A synthesis of research. *Sage Journals*, May 2021.

[95] Michael J. Dumas and Kihana M. Ross. "Be Real Black for Me": Imagining BlackCrit in Education. *Sage Journals*, February 2016.

[96] Gloria Ladson-Billings. Just what is critical race theory and what's it doing in a nice field like education. *Research Gate*, January 1998.

[97] Gloria Ladson-Billings and William F. Tat IV. Toward a Critical Race Theory of Education. *Sage Journals*, October 1995.

[98] U.S. Bureau of Labor Statistics. Education Pays. September 2022.

[99] Rachael Pells. Sexism in school: 57boys. *Independent*, February 2017.

[100] Lorena F. Perez. Teachers' sexist attitudes have a major impact on secondary education teachers' sexist attitudes have a major impact on secondary education teachers' sexist attitudes have a major impact on secondary education. *Science X*, November 2021.

[101] Women in science, technology, engineering, and mathematics (stem). *Catalyst*, August 2022.

[102] Richard Fry, Brian Kennedy, and Cary Funk. STEM jobs see uneven progress in increasing gender, racial and ethnic diversity. *Pew Research*, April 2021.

[103] National Center of Education Stastics. High School Coursetaking. 2016.

[104] The Investopedia Team. Binomial Distribution: Definition, Formula, Analysis, and Example. *Investopedia*, 2024.

[105] Ujwal Tewari. Regularization — Understanding l1 and l2 Regularization for Deep Learning. *The Medium*, 2021.

[106] scikitLearn.org.

[107] Emily Badger, Claire Cain Miller, and et al. Extensive Data Shows Punishing Reach for Black Boys. *The New York Times*, March 2018.

[108] Kyle M. Whitcomb, Sonja Cwik, and Chandralekha Singh. Not All Disadvantages Are Equal. *Sage Journals*, November 2021.

[109] Institute for Higher Education Policy, 2024. URL `https://www.ihep.org/press/how-wealth-not-just-income-shapes-college-access-and-success/`.

[110] Avery Russell. Mainstream Feminism Still Centers White Women. *The Oberlin Review*, September 2023.

[111] Paul Bulter. Black Male Exceptionalism? *Georgetown University Law Center*, 2013.

[112] Luis Noe-Bustamant, Ana Gonzalez-Barrera, and et al. Majority of Latinos Say Skin Color Impacts Opportunity in America and Shapes Daily Life. *Pew Research*, 2021.