Bias in Graduation in Universities and Colleges

Chia-Yu Chen
Applied Data Science, MS
University of Southern California
Los Angeles, CA 90007
cchen557@usc.edu

Abstract

When we are talking about the graduation rate in colleges or universities, what societal bias that people might have? Such as race? Does Asian or white people have a higher graduation rate than other races? For historically black college or university, does it really have a lower graduation rate than other universities? What about locations? Does a college located in a higher level of urbanization area cause a higher graduation rate? Some say people in New England states tend to be more educated, is this reflected in our data? What about SAT/ACT scores? Does public school or private school have significance difference in graduation rate? The goal of this project is to explore and answer such questions by identifying potential factors affecting graduation rate and then explore whether there exists any bias such as racial bias in graduation rate.

1 Introduction

According to a report from the Georgetown University Center on Education and the Workforce, employees with a bachelor's degree make more than one million dollars in their lifetimes on average than those with a high school certificate. Although the exact figures will vary greatly based on the universities/colleges that individuals attend, the majors that they pursue, and other factors, it is indisputable that going to university/college is a beneficial investment in modern society. Even in certain nations, a bachelor's degree is a necessity for the majority of occupations, regardless of whether your major is related to the job. Graduation rate, on the other hand, are not as high as we had assumed.

Only 45% of college students graduate in four years, according to a nationwide survey by UCLA and a report by the Education Department and less than two-thirds of college students finish in six years. Why is graduation rate so low, despite the obvious benefits of having a bachelor's degree in terms of employment? In addition, since biases are everywhere, no matter which is obvious or unconscious, it is important to discover if there exists any bias in data before training on such data. This project[1] aims to expose possible bias using statistical analysis and fairness analysis, and explore some ways to mitigate such bias.

2 Related Work

One area of research concentrates on the groups that may be marginalized or have challenges that are not readily apparent like students with some types of learning disabilities. A study focuses on the early integration experiences of students who are the first-generation in their family to attend college and how it may have an effect on retention and graduation (Woosley and Shepler, 2011). According to the 2011 UCLA study, it was found that first-generation college students have notably

[1] Github repo - https://github.com/cchen557/Graduation-in-Colleges.git

lower retention and graduation rate. Results indicate that integration and other factors that affect first-generation students are similar to those affect non-first-generation students. First-generation students often have financial burden such as having full-time/part-time jobs and children that will make them difficult focus on an education.

A massive number of papers focus on examining effective policies and programs that college could pursue or implement to increase graduation rate. To cite an instance, for increasing four-year college graduation rate, the most cost-effective manner is to increase spending at all public colleges and target elimination of tuition and fees for income-eligible students (Avery, Howell, Pender, and Sacerdote, 2019). Another example is an approach called The City University of New York (CUNY) Accelerated Study in Associate Programs (ASAP), which successfully improves graduation rate through providing students with academic support such as intensive advising and financial support and other support services such as offering blocked courses and condensed schedules in order to help them graduate within three years (Sommo, Cullinan, Manno, Blake and Alonzo, 2018).

According to Wohlgemuth et al. (2007), financial aid is a critical factor influencing students' likelihood of graduating. The scholars establish that it facilitates the enhancement of student retention in different learning levels, thus enhancing the chances of graduating. The study also established that an increase of gift aid had a significant impact than enhanced work-study aid and increased loan amounts. However, since gift aid facilitations are always limited in supply in most higher learning institutions, alternative aid should be considered. Consequently, enhancing the loan amounts facilitates retention and students' graduation rate, particularly in later years in school. This finding is consistent with research done by Choy and Cuccaro-Alamin (1998). Following this finding, learners may be more aware of the significance of graduating during their later years in higher learning institutions, hence enhancing learners' retention and graduation rate when their loan access is enhanced. In this context, while in their later years on campus, the learners would be more inclined and determined, thus investing heavily in the completion of their degrees and thus graduate even in the face of future earnings' tied strings. Alternatively, learners that access enhanced loan amounts are likely to be compelled to complete their education and graduate to guarantee higher earnings in their future careers that would facilitate a better life amid repayment of the loans. In this regard, the study concludes that an increase in aid for the learners from any relevant source contributes to higher retention in school, thus facilitating graduation rate.

On the other hand, Titus (2006) establishes various factors affecting the students' likelihood of graduating from college, particularly a four-year learning institution. This includes socioeconomic status (SES), the performance of the learners at the pre-college level, and the learners' background characteristics. These findings are consistent with the prior study by Astin (1993) that evaluated the influence of college completion by academic performances of the learners in pre-college institutions and SES. The scholar also identified other influencing factors. This includes the learner's experiences during their first years at the institution. For instance, the student's satisfaction with the campus" climate positively impacted the learners' likelihood of graduating. The other factor entailed the number of hours the learner spends working, which negatively impacted their graduation rate. Consequently, the learners' unmet financial needs also negatively impact the learners' likelihood to graduate. This finding is aligned with Bean's (1990) study outcomes that higher learning institutions are less able to affect influence external factors off-campus than those associated with learners' experiences within the learning institution. Further, the study Titus (2006) identified that students' graduation rate are also positively influenced by the percentage of the tuition fee meant for the institutions' revenue. The finding is consistent with Anderson's (1985) study findings that when higher learning institutions increasingly focused on tuition fees as their major institutional revenue generator, they increasingly concentrated on retaining learners, thus enhancing graduation rate.

Cragg (2009) established that various variables affected different groups of learners unequally based on the learners' relative position to the higher learning institution's average attendance cost and SAT score. Based on these findings, staff, the faculties, and the university's faculties must establish a good understanding of their learners' population for the facility to offer student and academic programs that meet the learners' needs, thereby facilitating their likelihood to graduate. The policy implications of these findings on the higher education sector include holding higher learning institutions accountable for their provision of education access and their ability to facilitate learners' success. In this context, policymakers must call for the enhanced accountability of these tertiary learning institutions and demand that states and the learning institutions must be held accountable for their learners' graduation following their enrolment in the respective institutions. In this context, higher learning facilities

offering admissions to learners whose traits do not match what they offer could consider redirecting enhanced resources to such learners to enhance their graduation probability.

3 Data Overview

The dataset is called "American University Data"[2] and has details about the universities and colleges in USA for year 2013. It is provided by IPEDS which is a system of surveys that gathers information on U.S Colleges, Universities, technical and vaccational institutions. The dataset is then revised and posted on Kaggle.

The data contains 145 columns including both categorical features and numerical features such as ZIP code, number of applicants, number of admissions, number of enrolled, SAT score distribution, ACT score distribution, tuition and fees, sector of institution, graduation rate. Each row represents information about one of the 1,534 American colleges and universities. Since this project aims to focus on graduation rate of undergraduates rather than graduate students or others, I will only consider the factors related to the graduation rate with completing a bachelor's degree within 6 years.

ID number	Name	year	ZIP code	Highest degree offered	County name	Longitude location of institution	Latitude location of institution	Religious affiliation	Offers Less than one year certificate	...	Percent of freshmen receiving federal grant aid	Percent of freshmen receiving Pell grants	Percent of freshmen receiving other federal grant aid
100654	Alabama A & M University	2013	35762	Doctor's degree - research/scholarship	Madison County	-86.568502	34.783368	Not applicable	Implied no	...	81.0	81.0	7.0
100663	University of Alabama at Birmingham	2013	35294-0110	Doctor's degree - research/scholarship and pro...	Jefferson County	-86.809170	33.502230	Not applicable	Implied no	...	36.0	36.0	10.0
100690	Amridge University	2013	36117-3553	Doctor's degree - research/scholarship and pro...	Montgomery County	-86.174010	32.362609	Churches of Christ	Implied no	...	90.0	90.0	0.0
100706	University of Alabama in Huntsville	2013	35899	Doctor's degree - research/scholarship and pro...	Madison County	-86.638420	34.722818	Not applicable	Yes	...	31.0	31.0	4.0
100724	Alabama State University	2013	36104-0271	Doctor's degree - research/scholarship and pro...	Montgomery County	-86.295677	32.364317	Not applicable	Implied no	...	76.0	76.0	13.0

Figure 1: A small sample of original dataset

4 Methods

4.1 Data cleaning and preprocessing

Graduation rate is the outcome variable; thus universities and colleges with missing graduation rate were removed during the data cleaning process. Since this project aims to focus on the graduation rate with completing a bachelor's degree within 6 years rather than master's degree or others, the factors that are unrelated were dropped. I also renamed specific columns' names and values in order to present more clearer ticks in the upcoming graphs. Missing values in variables were filled with mean value. For label encoding, one-hot encoding was used to convert the categorical features, and the numerical features were normalized.

4.2 Data analysis and protected feature selection

For data analysis, since I am only interested in investigating the potential factors that may affect graduation rate and see if there exists bias, the correlation matrix and box plot were used to represent the relationships between several factors and graduation rate to further choose protected feature.

[2] American University Data

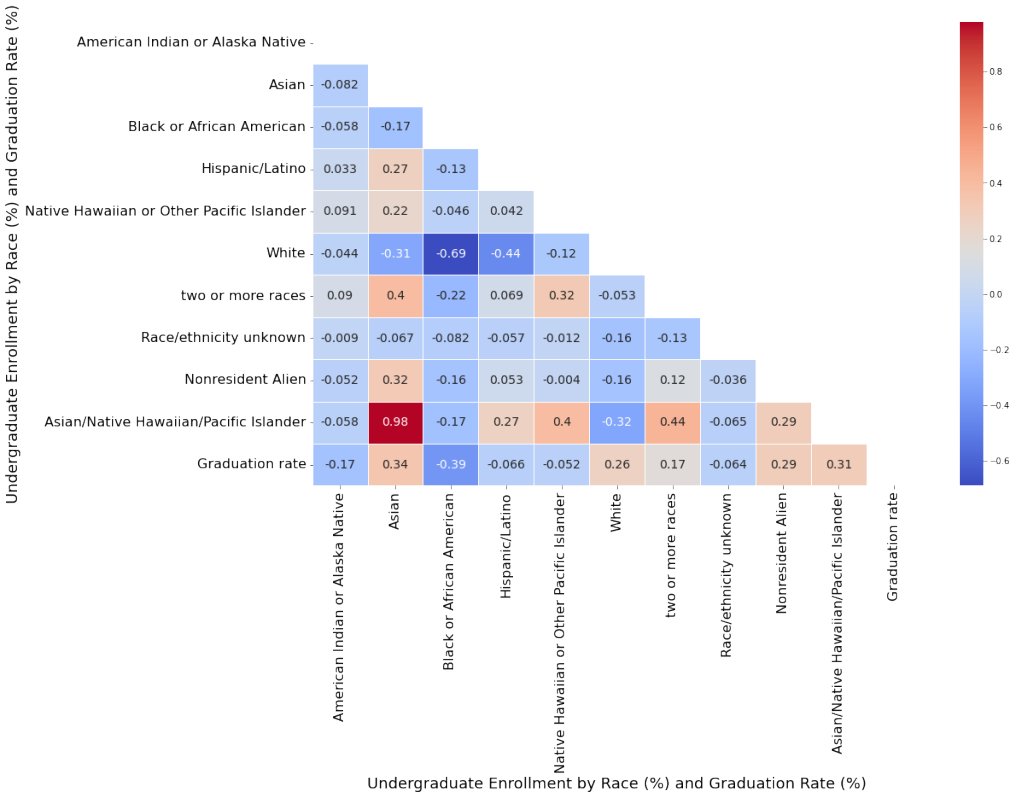

Figure 2: Correlation matrix of Race and Graduation Rate

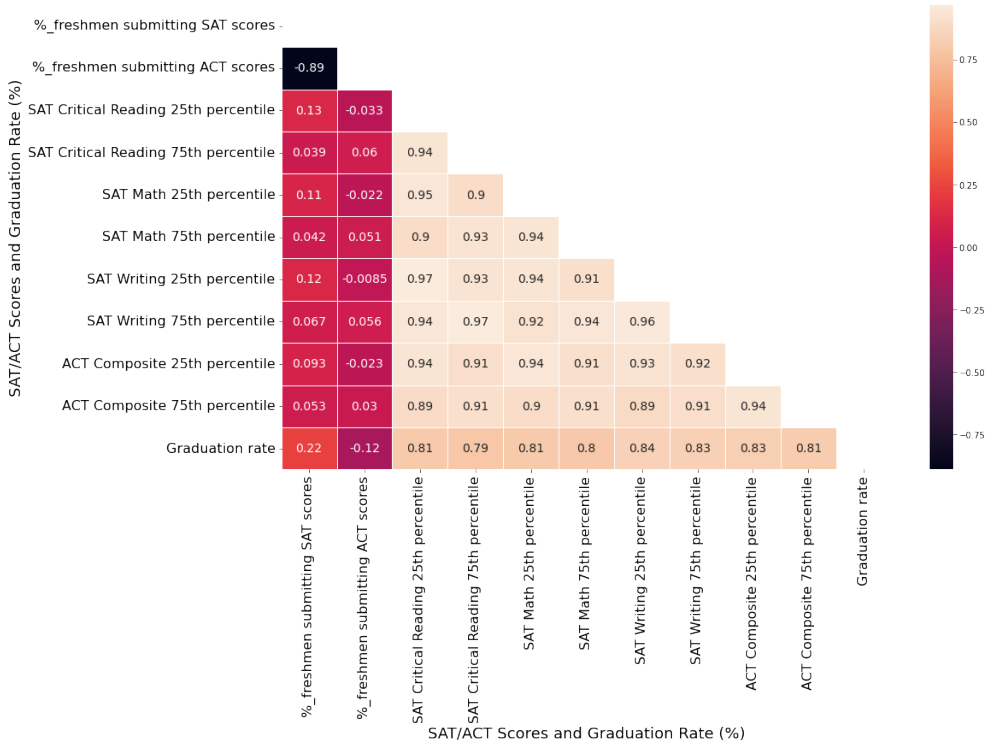

Figure 3: Correlation matrix of SAT/ACT Scores and Graduation Rate

For Figure 2, I observe that Black or African Americans have the most negative correlation towards graduation rate. It might indicate that Black or African Americans who are admitted to school have the lowest graduation rate compared to other races. On the other hand, most of other races such as Asian/White/Native Hawaiian/Pacific Islander have positive correlation towards graduation rate.

For Figure 3, I observe that SAT/ACT scores are highly correlated with towards graduation rate and have a positive relationship. Moreover, among these scores, the SAT writing score is the most relevant feature.

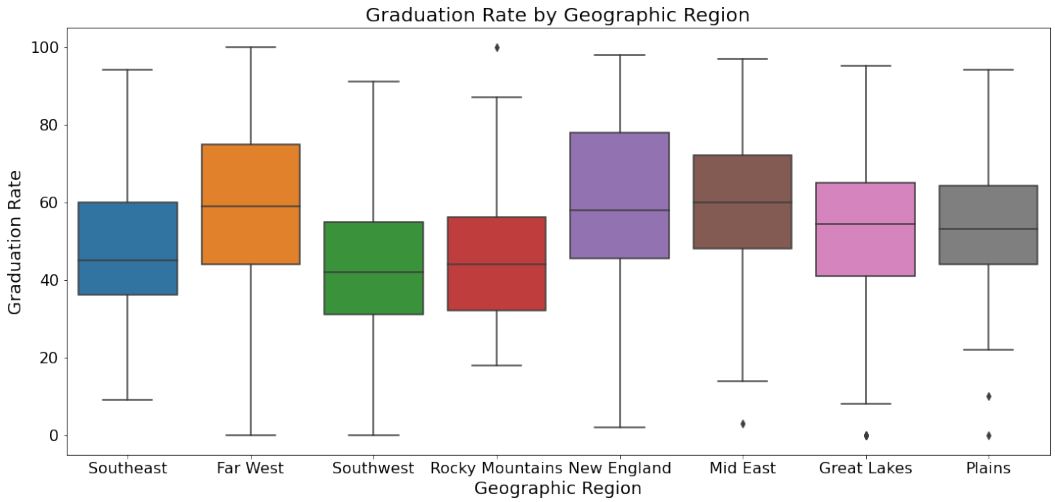

Figure 4: Graduation Rate by Geographic Region

The box plot in Figure 4 shows that the universities and colleges in different geographic regions have quite different graduation rates, and it may exist regional bias. It also indicates the far west, new England and mid east are the more educated regions.

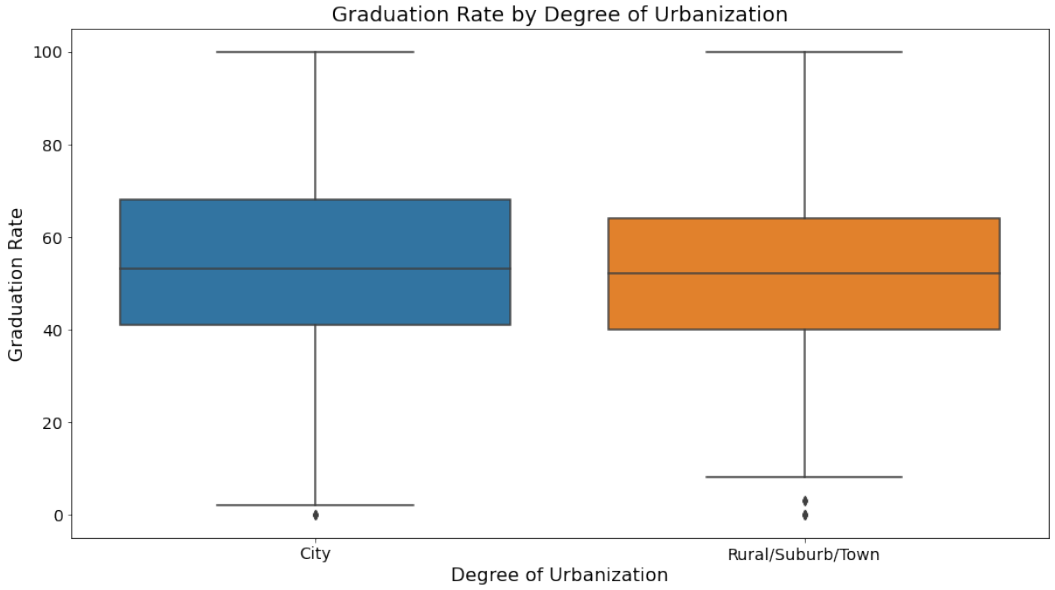

Figure 5: Graduation Rate by Degree of Urbanization

The box plot in Figure 5 shows that there is not much difference in graduation rate for different degree of urbanization such as city or non-city including rural, suburb, and town.

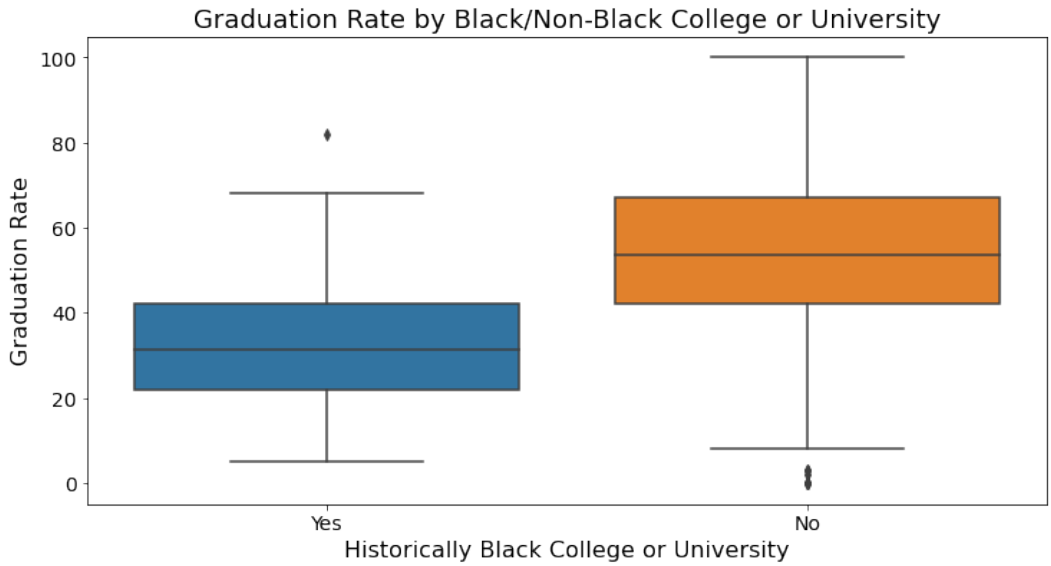

Figure 6: Graduation Rate by Black/Non-Black College or University

The box plot in Figure 6 shows that for historically black colleges, it has significant lower graduation rate than non-black colleges, and it may exist bias.

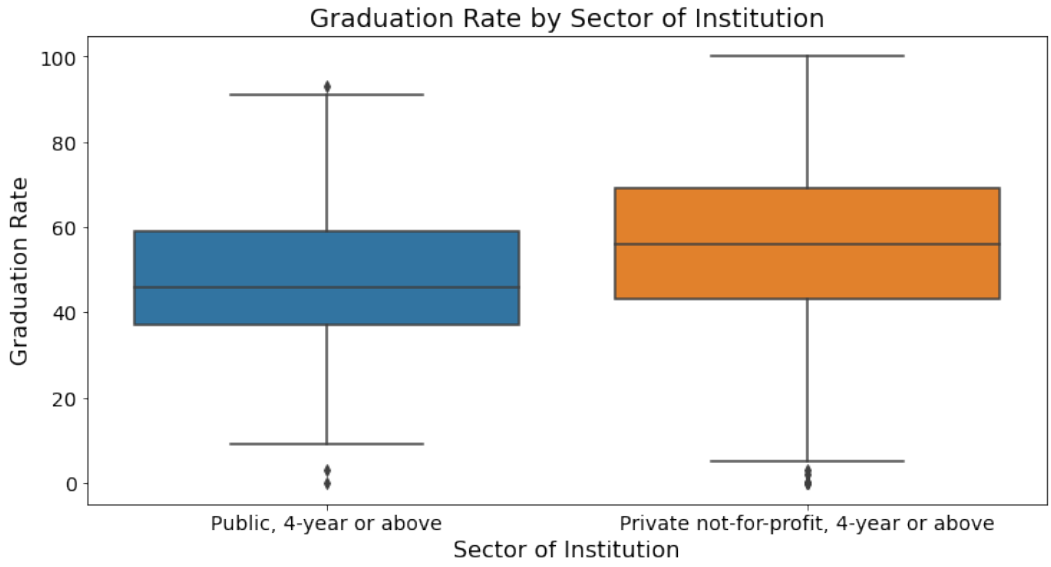

Figure 7: Graduation Rate by Black/Non-Black College or University

The box plot in Figure 7 shows that for sector of institution, the public colleges have a bit lower graduation rate than private colleges.

According to the results I obtain from above analyses, it seems there exists racial bias and regional bias in graduation rate. However, there is no specific column representing number of people graduating in six years by each race. Therefore, I choose whether it is a black college or not as protected feature for detecting racial bias and keep other related features such as counties, scores, states, geographical regions, sector of institution for training machine learning model.

4.3 Statistical analysis on protected feature

4.3.1 Protected feature: Race

For performing the statistical analysis on protected feature, i.e., "Historically Black College or University" column, to detect if there exists racial bias. First, I used the Pareto Principle, which is the 80/20 rule, to split the dataset into a training data and test data. Since the graduation rate threshold will significantly affect whether the dataset is balanced or unbalanced, and a low graduation rate threshold will cause the dataset to become more unbalanced, I chose 45% as the threshold to keep the dataset balanced. The goal is to predict whether the graduation rate exceeds 45% based on other features. Then I performed a t-test on the graduation rate between the two protected groups.

The p-value I obtained from the test result is less than the significance level alpha, i.e., 5%, so it is statistically significant for graduation rate dataset. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random). Therefore, I rejected the null hypothesis, and accepted the alternative hypothesis. Furthermore, this means that the black college or university's average graduation rate is statistically different from the non-black college or university's average graduation rate, i.e., there is a strong relationship between these two variables, race and graduation rate.

4.3.2 Protected feature: Region

Like performing the statistical analysis mentioned above, here I use region as protected feature, i.e., "Geographic region" column, to detect if there exists regional bias. I used the same way to split the dataset into a training data and test data. However, since the graduation rate threshold will significantly affect whether the dataset is balanced or unbalanced, I chose a different way to perform the statistical analysis, which is not to set a threshold of graduation rate. Then I compared the means of the graduation rate across different protected groups and performed a t-test on it.

Table 1: Statistical analysis result by region

Region	Mean	p-value
Far West	60.98	0.0081
Great Lakes	51.73	0.9223
Mid East	59.04	« 0.05
New England	60.06	0.0040
Plains	54.45	0.0091
Rocky Mountains	43.86	0.0034
Southeast	47.65	« 0.05
Southwest	42.83	« 0.05

According to the results from Table 1, the p-values for almost all regions except Great Lakes are less than the significance level alpha, i.e., 5%, so they are statistically significant for graduation rate dataset. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random). Therefore, I rejected the null hypothesis, and accepted the alternative hypothesis. Furthermore, this means there is a strong relationship between geographic region and graduation rate.

4.4 Explore fairness in prediction

In this section, for the sake of brevity, I only consider bias in relation to race. The bias in data would be propagated to the model if the model was trained on such data. Therefore, I trained a logistic regression model to predict the graduation rate from the features on the training data. For evaluating classification model, I used both accuracy score and AUC score to measure the model performance, and reported the statistical parity and equalized opportunity as fairness metrics.

To pursue fairness in prediction, I explored two ways to mitigate racial bias. One is to simply remove protected attribute when training the model, and the other is to create a synthetic training set by

flipping race and augment the training set. Report the accuracy score, AUC score and two fairness metrics to see if the predictions are fairer.

Table 2: Fairness analysis result of race as protected feature

Model	Accurracy	AUC	Statistical Parity	Equal opportunity
Baseline Logistic Regression	0.703	0.731	-0.146	-0.113
Removing the protected feature	0.824	0.795	-0.689	-0.423
Augmenting the training set	0.716	0.726	-0.118	0.320

According to the accuracy scores from Table 2, since I had already addressed the problem of unbalanced dataset by setting a higher threshold of graduation rate at first, the result indicates that all models performed not bad through correctly classifying at least 70% of observations. As for the logistic regression model with protected feature removed, it performed much better and classified more than 80% observations into classes correctly. Removing protected attribute only increased the AUC score a bit. However, since AUC score in all models were more than 0.7, it means there is at least 70% chance that these models will be able to distinguish between positive class and negative class.

For the fairness analysis, the statistical parity difference increased a lot after removing the protected variable. This indicates that the model with removing the protected feature is significantly unfairly biased against the privileged group. As for the model with augmenting the training set, the results of statistical parity difference shows that this measure did help mitigate the group bias a bit. However, when equal opportunity differences are taken into account, these two measures did not help to reduce such bias.

5 Conclusion

In this paper, it is confirmed that it does exist racial bias in graduation rate. For mitigating such bias, I explored two ways to pursue fairness in the classifiers and tried to achieve an optimal trade-off between accuracy and fairness. Although the final result was not ideal, the data still provided some insights. Through training a logistic regression model with removing the sensitive attribute, there is significant improvement in model performance. However, it did not help mitigate the bias and the predictions were even more unfair. As for the model with creating a synthetic training set by flipping race and augment the training set, the experimental result demonstrates this measure improved the group bias metrics a bit in terms of statistical parity difference with little deterioration in model performance which is losing 0.5% on the AUC score. All things considered, there might be other factors that cause bias, such as unrepresentative or incomplete training data or the reliance on flawed information that reflects historical inequalities.

There are many interesting venues for future work. For example, it would be interesting to explore other supervised learning algorithms, such as support vector machine (SVM) or random forest, and try different ways to mitigate the bias in this dataset. The bias should be checked against the other protected classes as well.

References

[1] The College Payoff: More Education Doesn't Always Mean More Earnings. *CEW Georgetown*, 7 Oct. 2021, `https://cew.georgetown.edu/cew-reports/collegepayoff2021/`.

[2] Marcus, Jon. "Most College Students Don't Graduate in 4 Years, so the Government Counts 6 Years as 'Success'." .com, NBCUniversal News Group, 10 Oct. 2021, `https://www.nbcnews.com/news/us-news/college-students-dont-graduate-4-years-government-counts-6-years-succe-rcna2776`.

[3] Woosley, Sherry A. and Dustin K. Shepler. "Understanding the Early Integration Experiences of First-Generation College Students." *College student journal* 45 (2011): 700.

[4] Avery, C., Howell, J., Pender, M., Sacerdote, B. (2019). Policies and Payoffs to Addressing America's College Graduation Deficit. *Brookings Papers on Economic Activity 2019*(2), 93-172. doi:10.1353/eca.2019.0013.

[5] Sommo, C., Cullinan, D., Manno, M.S., Blake, S.A., Alonzo, E. (2018). Doubling graduation rates in a New State: Two-Year Findings From the ASAP Ohio Demonstration. *Pedagogy eJournal*.

[6] Cragg, K. M. (2009). *Influencing the probability for graduation at four-year institutions: A multi-model analysis*. Research in Higher Education, 50(4), 394-413.

[7] Titus, M. A. (2006). *No college student left behind: The influence of financial aspects of a state's higher education policy on college completion*. The Review of Higher Education, 29(3), 293-317.

[8] Wohlgemuth, D. Whalen, D. Sullivan, J. Nading, C. Shelley, M. Wang, Y. (2007). *Financial, academic, and environmental influences on the retention and graduation of students*. Journal of College Student Retention: Research, Theory Practice, 8(4), 457-475.