Jaiden Atterbury

Peter Gao

STAT 303

10 November 2022

Long Response 2: Statistical Definitions of Fairness

1. **Explain, in plain language, what calibration means in this context.**

In the context of predicting student success in higher education, calibration means that the likelihood that an individual graduates on time or doesn’t, given that the algorithm gave two individuals the same score, should be the same independent of any inputs fed into the algorithm. For example, the probability that a student graduates on time, given that they received the same score from the algorithm, shouldn’t be different for members of different races.

1. **Explain, in plain language, what predictive parity means in this context.**

In the context of using algorithms to predict a student’s risk of not graduating on time, predictive parity means that the likelihood that an individual graduates on time or doesn’t, given that the algorithm gave two individuals the same decision, should be the same independent of any group membership or inputs into the algorithm. The “decision” in this case is to label a student either high-risk, medium-risk, or low-risk depending on a certain score threshold placed by the school using the algorithm. For example, the probability that a student graduates on time or not, given that they are labeled as high-risk by the algorithm, shouldn’t be different for members of different races.

1. **Explain, in plain language, what error rate balance means in this context.**

In the context previously defined in the last two responses, error rate balance means that for each threshold of score, with those thresholds being high-risk, medium-risk, and low-risk, the likelihood of a false positive or a false negative is the same regardless of group membership status or any other given input factor. In this specific case, a false positive is denoted by the event that the algorithm labels a student as high risk to not graduate on time but the student ends up graduating on time. On the other hand, a false negative in this context is defined as the algorithm labeling a student as low risk to not graduate on time, but the student ends up not graduating on time.

1. **Do the algorithms described in the articles above violate any of the above definitions of fairness? If there is not enough information in the articles provided, say so.**

Based on the information given in the two above articles, there is not enough information to say that the algorithms described in the articles do or do not violate any of the above definitions of fairness. The main talking point of the two articles, especially *Major Universities Are Using Race as a “High Impact Predictor” of Student Success* by The Markup, was that there existed large disparities in how the software treats students of different races. In particular, the disparity was strong for black students, who were deemed high risk at as much as quadruple the rate of their white peers. Despite the articles claiming that in the past few years overall graduation rates at universities and colleges have been going up, they couldn’t attribute this rise in rates to be caused by predictive software. Furthermore, the article didn’t talk about how the likelihood of graduating on time or not was affected by these scores, nor did they mention how students with the same score, or in the same threshold, differed among racial lines. Therefore, without information on how the scores impacted the likelihood to graduate or not for different input values, we cannot say the algorithm does or does not violate the calibration or predictive parity definitions of fairness. It is also important to note that although the articles mention that the prevalence of false positives and negatives were found, the articles didn’t dive into how these rates compared across different races and other input values. Thus, we also can’t say that the algorithm does or does not violate the error balance rate definition of fairness.

1. **In your opinion, which of the three definitions of fairness above would be the most important to satisfy in this context? Why?**

In the context of an algorithm trying to measure a student's risk of not graduating on time, I believe the most important definition of fairness out of the three above definitions is predictive parity. Before I delve into specific details on why satisfying predictive parity is so important in this context, I believe it is necessary to explain why I chose predictive parity over calibration despite the two being very similar. My reason for choosing predictive parity over calibration is that predictive parity deals more with labels/decisions at certain thresholds, while calibration is more focused on the individual scores the algorithm gives students. Since the articles deal more with how the predictive algorithm deals with students at certain labels/thresholds (such as high-risk, medium-risk, low-risk) it makes sense to choose a definition of fairness that also deals with these labels and thresholds, hence why I chose predictive parity.

My first reason for choosing predictive parity as the most important definition of fairness in this setting is that it reflects a fairness notion that people with the same decision from the algorithm have the same likelihood of achieving a similar outcome. The outcome in this case being graduating on time or not. My next reason for choosing predictive parity as the most important notion of fairness requires us to take into account the ramifications of what would happen if the algorithm did not satisfy predictive parity, especially through the lens of the disparity mentioned in the article. As mentioned previously, the algorithms described in this article label black students as high risk to not graduate almost four times more than they label white students as high risk. Since there already exists a disparity on how the algorithm classifies certain races at a given threshold, if the algorithm doesn’t satisfy predictive parity and leads to different probabilities of graduating on time or not, this will only further exacerbate inequalities especially those that target historically marginalized groups, which in this context would be the black students, and other students of color in the colleges who use this algorithm.

1. **If such an algorithm were shown to satisfy your preferred definition of fairness, would you be comfortable with its use in this context? What other ethical concerns should we consider?**

If an algorithm, such as the ones described in the above articles, were shown to satisfy predictive parity, I still wouldn’t be comfortable with its use in assessing students’ risk of not graduating on time. The reason why I wouldn’t be comfortable with its use in this context boils down to three other ethical concerns we should consider; the fact that the algorithm already produces quantifiable disparities between inputs, the lack of transparency/guidance the model and its parent company provides to the colleges and universities that deploy the models, as well as universities tendency to stray away from the main usage of the model.

As mentioned previously, the company’s models are trained on historic student data. The use of this data leads to the model finding discriminatory patterns in the data that go uninvestigated, and when used in practice, produce bias towards the groups under/misrepresented by the data. This is why the algorithm labels black students and other students of color as high risk at rates much higher than that of white students. Since the algorithm already has unresolved or unexplained racial bias, even if it satisfied predictive parity they would need to satisfy or at least explain why the model produces such drastic disparities in the way it labels certain students of specific races, even when other inputs are the same.

Another reason why I wouldn’t be comfortable with the use of algorithms like EAB even if they satisfied the predictive parity definition of fairness, is that there exists a lack of transparency/guidance the model and its parent company provides to the colleges and universities that deploy the models. Two quotes that exemplify this ethical concern are, “I certainly haven’t had a lot of information from behind the proprietary algorithms,” and “At no point in required training did they really explain what this risk score really is.” These quotes were made by Carolyn Bassett and Maryclare Griffin, both of these people frequently use these scores to try and quantify student success. With that said, if the company isn’t disclosing any information about how their algorithm works or the importance/value of the output it produces, I don’t think it should be used to make critical decisions about students’ futures, even if it satisfies the predictive parity definition of fairness.

Finally, I wouldn’t be comfortable with the use of predictive algorithms like those mentioned in the above articles, due to the ethical concern that universities tend to stray away from the main usage of the model. In the article, representatives at EAB, the algorithm at the center of the article, told The Markup that its tools should be used to identify solutions and create resources for groups of similarly situated students. However, universities are actually using the software to make institutional changes, rather than individual advising recommendations. In essence, universities are using scores to directly make decisions on students’ futures and aspirations instead of using them as screening aids to their own decisions.