Policy Memo 2 (1094 words)

TO: Local School District

FROM: Ethics Consultant

SUBJECT: New Classifier Fairness Concerns

1. Background:

The student dropout rate of this city's public schools has been increasing for the last several years, reaching the highest point ever this last year. To address this growing problem, the local school district is developing a machine learning classifier to predict which students, present and future, are at-risk of dropping out and not graduating high school. The school district plans to equip teachers of the at-risk students with the necessary resources to assist them, such as how to communicate directly with the student, adjusting their workload accordingly, or communicating with their parents. Although the school district is the one developing the classifier, each individual school in the district will be the one directly using the classifier to identify their own at-risk students.

2. Classifier Structure:

The school district believes student disengagement to be the biggest indicator for possible dropout, so the classifier uses disengagement as the target variable, and targets it using a mixture of features: academic factors (e.g., GPA, test results, attendance), personal demographics (e.g., race, ethnicity, gender, address), and teacher statistics (e.g., years of teaching, number of failing students, certificates). This original training data will come from several datasets having this necessary training information which the schools should have already collected throughout the years. New data will also be collected throughout the subsequent school year so that the data of previous dropouts can be correlated with current student information, leading to possible pattern recognition.

3. **Fairness Considerations:**

As with any decision affecting others, discrimination and fairness concerns should be considered before enacting them. Per Barocas and Selbst (2016), data isn't the perfect savior that solves every problem beautifully; instead, they argue, data mining can reproduce institutional prejudices and further discrimination if not performed with proper care. In the case of high school dropout rates, there is not a direct way to measure student disengagement, so what constitutes disengagement must be defined in other measurable ways: low attendance or poor grades, for example. Hellman (2021) defines this approximation of the trait as "measurement error" since traits like student dropout are complex and difficult to measure directly. The classifier described by the school district uses these two criteria, among several others, to try to define disengagement more holistically, yet there are still different choices for this definition which could lead to a smaller or larger degree of discrimination towards protected classes. Thus, considerations should be made about if the currently decided target variable invites the least discriminations.

In addition to possible consequences due to how the target variable is defined, the training data used can be inherently biased and lead to discriminatory models (Barocas & Selbst). Hellman (2021) furthers the idea of flawed training data by noting that it can result from ordinary mistakes - those with no discriminatory intent - or from processes marred with morally dubious collection techniques, where so-called "accuracy-affecting injustice" has now occurred. The concepts presented should lead to examination of the training set the schools will use to ensure no discrimination has occurred at the base level of the classifier. If it is found that the training set is biased, all future classifications of at-risk students would take into account that initial bias, leading to possible discriminatory results.

The last fairness consideration that should be analyzed regarding the classifier is whether an action taken due to the classifier compounds injustice or not. Hellman (2021) defines when an action compounds injustice as any action that "… engages with the prior injustice sufficiently, and also augments or entrenches that injustice.…" This is an important point to consider because any discrimination and fairness injustices that students in high school face could be further compounded later in life. For example, the classifier could consider students living in low-income areas as high-risk for dropout, leading administrators to enrolling those students in supplementary courses. This factor would disproportionately affect kids of single-mother or minority households, at no fault of the kids or families, who are at much higher risk of poverty due, in part, to wage gaps between both gender and race (Damaske). Colleges, which require student transcripts for admissions, would see such classes and know the student was a dropout risk previously and deny that student admission, compounding the prior injustices against the student due to prior injustices against the student's parent(s). Although this is merely an imagined scenario, caution is warranted when using big data and machine learning tools because the possible danger of compounding injustice is more likely than conventional methods.

4. Practical Recommendations:

In this section, I will review a couple practical recommendations that the school district can take to alleviate the fairness and discrimination concerns involving their planned classifier. If discrimination presents because of the way the target variable is defined, that discrimination will be intrinsic to the problem now, showing up in every other step of the classifier, so we must first look at how to alleviate the possibility that the target variable’s definition is the cause for injustice.

As discussed previously, student disengagement is a complex trait to measure, so the proxy traits that have been chosen to measure it instead are certainly not all-inclusive. As such, perhaps a different group of people decide other traits measure student disengagement more accurately, and that definition might lead to a greater or lesser impact on protected groups. Thus, a possible way to alleviate such concerns would be to build several versions of the classifier using different sets of criteria and making note of students marked at-risk among all different versions, which would diminish the possibility that the definition of the target variable is the cause of injustice.

This diminishing of injustice can be potentially furthered through careful consideration of what is used in the training data. The schools have been collecting the data that will be used to train the classifier for many years, so where some may consider drawing class labels is not what others would believe to be the best possible label. It is in this debate and judgment call that the possibility for personal biases can enter the model, leading to biased results that hold up to the scrutiny of validation techniques since the results are correct given the biased ground beliefs upon which the model was initially trained. To alleviate a problem like this, careful consideration must be made with the data so that the classifier is not trained on such biased lines. Of course, there is always a chance that discrimination is still present after such precautions but having taken actions to diminish injustice and ensure the school district does not make a bad situation worse, they can find themselves less morally culpable.

5. References:

- Barocas, Solon, and Andrew D. Selbst. “Big Data’s Disparate Impact.” California Law Review, vol. 104, no. 3, California Law Review, Inc., 2016, pp. 671–732, http://www.jstor.org/stable/24758720.
- Damaske, Sarah et al. “Single mother families and employment, race, and poverty in changing economic times.” Social science research vol. 62 (2017): 120-133. doi:10.1016/j.ssresearch.2016.08.008
- Deborah Hellman, Big Data and Compounding Injustice, Journal of Moral Philosophy (2021).