ORIE 4741: Final Project

Anna Halldorsdottir, Stephanie Hellman

Introduction

Our goal is to identify risk factors for drinking in teenagers and be able to predict whether an individual is at risk for heavy drinking. To do this, we will examine the student's age, gender, race, along with tobacco use, bullying, drug use, and other possible risk factors. Our ultimate goal is for guidance counselors and other concerned parties to have ideas of possible drinking risk factors, so that they can talk to students before their drinking becomes extreme.

We are using the Youth Risk Behavior Survey data from 1991-2019, and we are using the Standard High School questionnaire. This questionnaire is given to high school students across the United States by the CDC every two years. We downloaded the data from the CDC website, at https://www.cdc.gov/healthyyouth/data/yrbs/data.htm. This survey is completely anonymous, and given to every student at many high schools, giving us a good survey of the US as a whole.

This survey gives us data on student's age, gender, race, violent behaviors, bullying history, tobacco, ecig, alcohol, marijuana, and drug usage, sexual behavior, weight, eating behaviors, physical activity, and grades. This gives us many features to predict with by using different questions from the survey. We chose to predict on the question "during the past 30 days, on how many days did you have a drink of alcohol?" We chose this question because it not only allows us to determine whether a student is drinking or not, but also whether they are drinking an unhealthy amount, therefore narrowing down the risk factors for excessive drinking.

Data and Preprocessing

The data originally contained 217,340 observations for 311 features. To clean up our data, we first got rid of unnecessary columns with completely unusable values (i.e. 'sitecode,' where every observation is just "XX"). Then we went through and removed all observations that were "NA" or simply blank in any of the columns for: 'age,' 'grade', 'sex', and/or alcohol frequency. Then we got rid of any features that were missing all observations, which totalled about 46 features.

Thus, our data is now 199,642 observations by 256 columns.

Ultimately, we are trying to predict column 'q41,' which corresponds to the students' answers to the question: "During the past 30 days, on how many days did you have at least one drink of alcohol?" The possible answer choices that the students could have chosen are: (A) 0 days, (B) 1 or 2 days, (C) 3 to 5 days, (D) 6 to 9 days, (E) 10 to 19 days, (F) 20 to 29 days, (G) All 30 days.

At a quick glance at Figure 1, we can see that the majority of students are drinking less than 10 days a month. In fact, the calculated median of this quantity is 0.0, meaning that at least 50% of these students did not have a singular drink. The mean of this data, however, is about 2.45 drinks with a standard deviation of 4.75.

The kids in the survey range from ages 12-18, and you can see the distribution of the ages in Figure 2. We have a few responses from students ages 12-13, but the majority of responses are from students ages 14-18, which makes sense since those are the average high school ages in the United States.

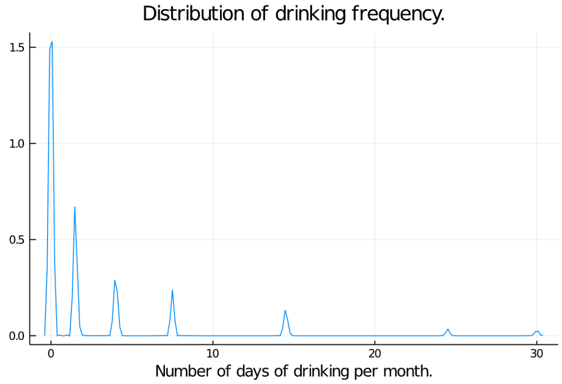

Figure 1

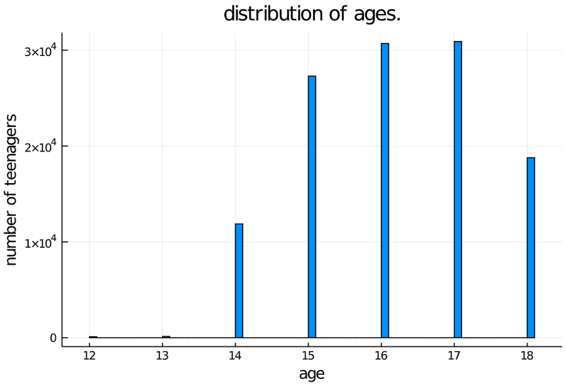

Figure 2

Regression Model

Our preliminary model is a linear regression model, where we treat the number of days a student has reported as drinking in the past month as a number. This is not a perfect assumption, as students are only given 7 options on the survey (0, 1-2, 3-5, 6-9, 10-19, 20-29, or 30), but we believe taking the average of the ranges is a safe assumption, since it will probably even out given the large data set. This assumption allows us to perform linear regression.

We ensure that our data does not over or underfit by immediately splitting the data into 3 categories: 20% of our data will be the test set, 60% of our data will be the training set, which still leaves us with 119,785 rows, more than enough to fit a convincing model. The last 20% will be the validation set, which will ensure our model does not underfit or overfit. This is because every time we change the model type or add features, we evaluate the accuracy of the trained model on the validation set, and if the accuracy of the validation set goes down, even if the training accuracy goes up, we do not use the model/feature. We will also use this validation set to determine which model should be our final model.

Using this method of adding features only if they are judged helpful by the validation accuracy, we fit a linear model using l2 loss. We started with this loss because of the high number of students who drank 0 days in the past month. Since our objective is to correctly classify heavy drinkers, and due to the low number of heavy drinkers in the data set, they are essentially all outliers. Therefore, we care a lot about classifying outliers.

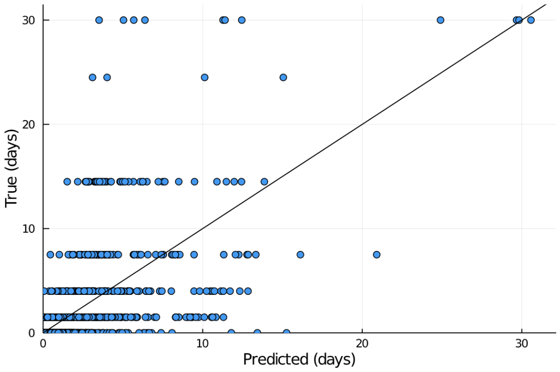

Figure 3: L2 Loss Model

This model includes age, year of survey, gender, whether the student has been bullied, cigarette usage, e-cigarette usage, other tobacco product usage, marijuana usage, prescription medication usage, cocaine usage, glue sniffing, heroin usage, ecstasy usage, steroid usage, sexual activity, healthy eating habits, unhealthy eating habits, breakfast habits, hours of sleep, and grades. Features that were unhelpful and not used in the model were student's feelings of depression, sexual identity, meth usage, body image, or TV usage. The below graph shows points of our prediction versus the true days the student has been drinking.

We can see that this model predicts our data moderately well, but it often under predicts the true amount of student drinking. We then tried a linear regression with Huber loss, in case students lying on the survey was causing us to misclassify. Again, we only add features if they are judged helpful by the validation accuracy.

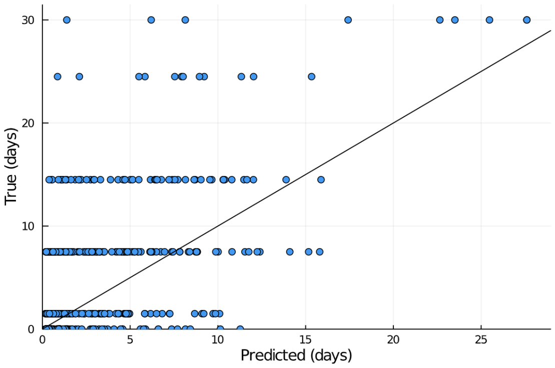

Figure 4: Huber Loss Model

This model includes age, gender, cigarette usage, e-cigarette usage, other tobacco product usage, marijuana usage, cocaine usage, glue sniffing, heroin usage, ecstasy usage, steroid usage, sexual activity, and breakfast habits. Notably, this has fewer features than the l2 model, creating a simpler and more easily interpretable model.

We can see that this Huber loss model is a better model for predicting alcohol usage in students with low amounts of alcohol consumption, but underpredicts the heavy drinkers much more than the l2 loss model. We thus present our l2 loss model as our best linear predictor, but believe we can do better with a different kind of prediction, where our target data is no longer presented as real values.

Random Forest Model

Next, we fit our dataset to a random forest model. We use this model because our targets are not well described as real values, as seen above. Random forest treats the output as nominal, which is not perfect for our data set, but it splits separately on all features, so its uniqueness could add additional information to the output. As we saw in lab, random forests perform better than single decision trees and bagging methods by combining multiple trees and also using randomness: only some of the given features are selected as possible candidates for splitting at each time step.

In this model, we removed questions about hard drugs, as we do not want them coloring our output, but we re-added features that the linear model thought was uninteresting, such as depression and sexual identity. We wanted this model to have as much information as possible to make predictions.

The random forest model has two hyperparameters: number of trees and number of features to split on at each time step. To determine the optimal setting of these hyperparameters, we ran a grid search. The number of trees ranged from 5-100, the range determined by computing power, and the number of features ranged from 2 to 45, as the model only had about 90 features total, and it is not recommended to split on over half the total possible features.

The grid search was done on the validation set. The hyperparameters which resulted in the best accuracy were 25 trees and 2 feature choices. The results from the model with these hyperparameters are summarized in the graph below.

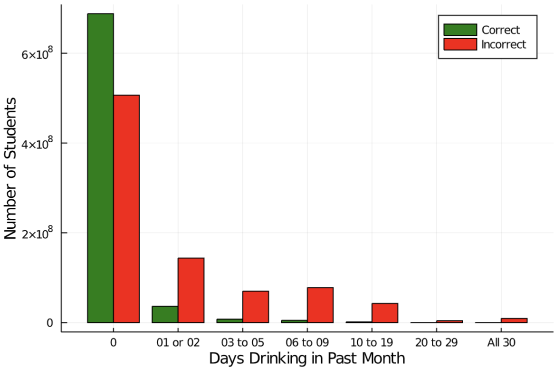

Figure 5: Random Forest Model

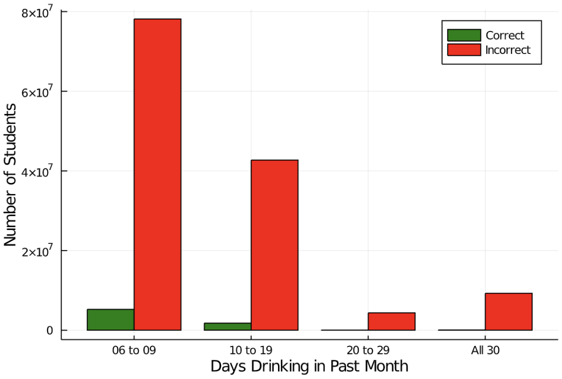

Figure 6: Random Forest Model Outliers

This model has 56.2% accuracy on the validation set. This is much better than random chance, implying that the model does have some predictive power. However, we can see that most of the predictive power is in the low number of days drinking. This model predicts most of its correct predictions in 0 days drinking, however, it also predicts most of its incorrect predictions in 0 days drinking. This is because of the high number of students who have never drank. This model does, as we can see in Figure 6, predict at least some middle drinkers (6 to 9 days and 10 to 19 days) correctly. We can see that this model is more accurate for moderate drinkers than the linear model. It still does not classify heavy drinkers correctly, though it rarely did predict some heavy drinkers.

Imputation

The data set that we have had a lot of missing values randomly interspersed throughout the data set. The number of missing data for each column varied significantly, with 0% missing in the `year` column (the year the student filled out the survey) and over 95% of data missing in column `q85` (question 85 was: “During the past 12 months, have you been tested for a sexually transmitted disease (STD) other than HIV, such as chlamydia or gonorrhea?”). This question is missing so much data because it was not asked until 2019. This means that of the 15 years of data that we have, this question only has responses in one of those years.

We attempted to approach this problem of having so many missing values in two ways: by fitting a Generalized Low Rank Model as presented by Dr. Udell and then via the Gaussian copula model presented by Yuxuan Zhao.

Gaussian copula

Since our data is primarily Boolean and Ordinal, we turn to the Gaussian copula model presented by Yuxuan Zhao. As we have such a large data set with so many features, we were hoping that this model could reveal insight into the correlation between the various ordinal variables.

Unfortunately, due to the complexity and number of features of the data set, the Gaussian copula was unable to reach convergence in a finite amount of time, even with a randomly chosen smaller sub-sample of our population. However, it ended at a correlation change ratio of 0.035, so we were still able to get some imputation from that result. From this result, we got a correlation matrix between the various features. Looking at the column for q41, which is the alcohol question that we are focusing on, we can see what Gaussian copula deemed to be the most correlated with alcohol intake. Features with a high correlation coefficient (>0.75) with q41 were q30 (0.83), q42 (0.86), q48 (0.84), q57 (0.78). These questions asked about cigarette usage, binge drinking behavior, marijuana usage, and buying illegal drugs on school property, respectively. Additionally, features with a relatively high correlation coefficient (>0.6), were q10 (0.65), q11 (0.7), q22 (0.62), q24 (-0.6), q35 (-0.73), q36 (0.6), and q40 (0.65). These questions ask about driving under the influence, texting/emailing while driving, a victim of domestic violence, being a victim of cyberbullying, using e-cigarettes, method of acquiring e-cigarettes, and the age at which they first consumed alcohol.

These results are interesting, as they were exactly what we set out to find in the first place. It seems that other risky behaviors, drug usage and texting while driving and other alcohol behaviors, are the highest risk factors for heavy drinking in teenagers. Unfortunately, due to the fact that the Gaussian copula never reached full convergence, we cannot be fully confident in these results. However, it's fascinating that solely imputing data using the Gaussian copula is able to give us this correlation result without fitting a traditional model, and thus we're able to use imputation to find general risk behaviors that concerned parties can focus on when they are determining whether or not a student is at risk for heavy drinking.

Generalized Low Rank Models

Since our Gaussian copula model did not converge, we had to turn to a different method to impute our missing data. Our data set has a mix of data types: continuous values, ordinal values, and boolean values. This indicates that Generalized Low Rank Models would be a good way to impute these various data types. We chose to minimize Huber Loss for continuous values, Multinomial Ordinal Loss for ordinal values, and Logistic Loss for boolean values.

The appeal of this method is its ability to handle a lot of missing data, especially when it encompasses multiple types, as with our data set. We tested different values for k, and ultimately ended up with $k = 10$ to use as the rank in our model, as through cross-validation, $k = 10$ resulted in the lowest error. Therefore we got an X and a W such that $Y = XW$, and then used these XW to fill in the missing values of Y. By running the cross validation function in Dr. Udell's GLRM package, we get an accuracy rate of about 60% when applied to the test set. Although this is not remarkably high, we think this is definitely a respectable accuracy rate considering the magnitude of missing values throughout the data set. Thus, we chose to use this imputed data set for our GLRM with ordinal loss models, as it provides more complete data to work with. We choose to use this imputed data, and not the imputed data from the Gaussian copula model, as we are not as confident in our imputed data from the Gaussian copula model, given that it did not converge, while the GLRM imputation did.

Proximal Gradient with Ordinal Loss

Lastly, we attempted the proximal gradient method with ordinal loss. We tried this method because our target is best described as an ordinal variable: 0 days drinking is clearly less than 1 or 2 days is less than 3 to 5. The proximal gradient method allows us to fit this model easily. We tried both Ordinal Hinge Loss and Bigger vs. Smaller Loss, as both loss functions are regarded as valid for ordinal regression. For these model, we used the imputed data from the Generalized Low Rank Model described above. We did this because the proximal gradient method is so math heavy, we wanted to make sure we had a full data set to work with so we could achieve the best results. We will choose our final model, based on the validation set, from these two models and our random forest model.

Our results from using hinge loss can be seen in Figure 7 and Figure 8. Using hinge loss, we only achieved 35.9% accuracy, much worse than we were expecting. However, accuracy is a biased metric on this data set, due to the high number of the 0 days drinking. We can see that this model correctly predicts about half of 0 day drinkers, and predicts many 1 to 2 day drinkers, though many

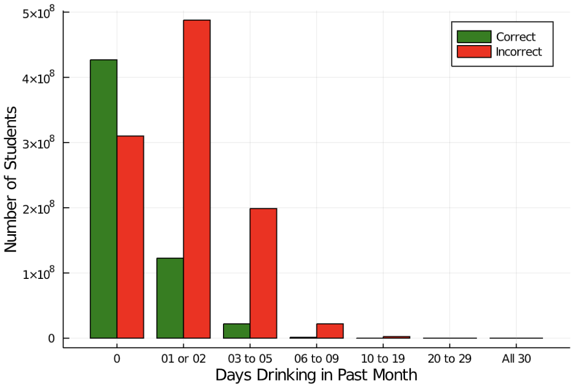

Figure 7: GLRM with Hinge Loss

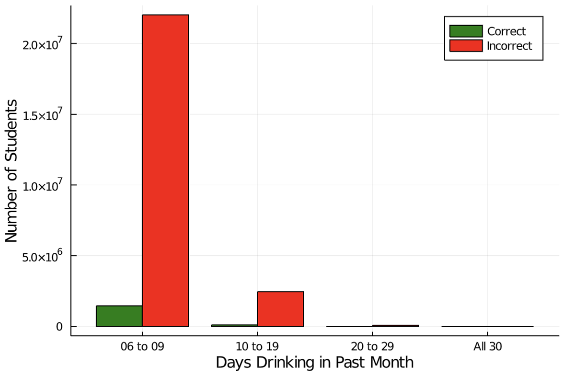

Figure 8: GLRM with Hinge Loss Outliers

incorrectly. This model predicts more moderate drinkers, though it gets many of its predictions wrong. Unfortunately, we can see that this model predicts few 20 to 29 day drinkers and no 30 day drinkers. Worse, the few it does predict it predicts incorrectly. Thus, this is not a good model for our question of specifically trying to identify heavy drinkers.

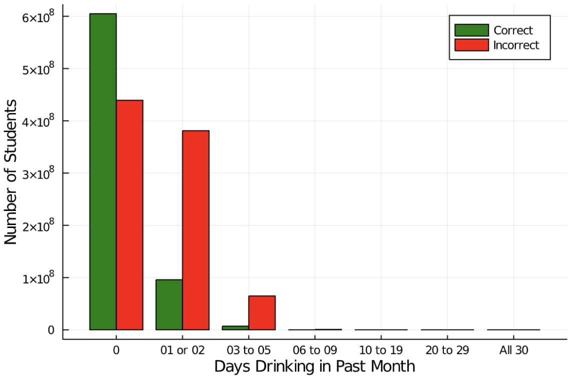

Figure 9: GLRM with Bigger vs Smaller Loss

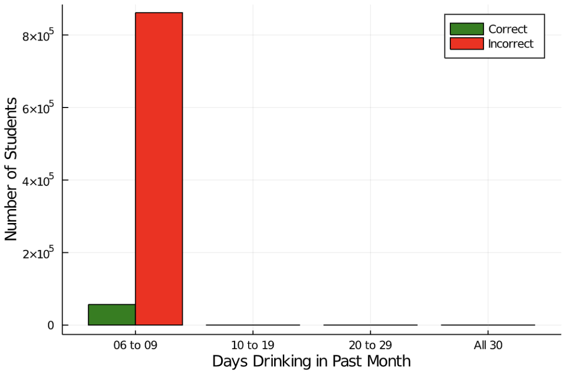

Figure 10: GLRM with BvS Loss Outliers

Our results from using bigger vs. smaller loss can be seen in Figure 9 and Figure 10. Here, we have a higher accuracy of 46.3%. However, this higher accuracy comes at the expense of no longer predicting heavy drinkers. Though this model predicts 0 day and 1 or 2 day drinkers more accurately than the hinge loss model did, it predicts no students that drink 10 to 19 days, 20 to 29 days, or all 30 days. This is a big problem for our model, since those are the predictions we care the most about. Thus, even though it has a lower accuracy, we see GLRM with hinge loss as our better ordinal model.

Final Model

We did not choose the linear model as our final model because our target, predicted days of drinking based on a multiple choice survey, does not fit well into a real numbered format. If we cannot describe our target well, we cannot have any confidence in our predictions. This leaves us a choice between our GLRM with ordinal loss models and our random forest model. Due to the lower accuracy of the GLRM with ordinal loss, and their inability to predict students with high risk of alcohol consumption, we chose the random forest model as our final model. Note here that we use the same data as our original random forest model: this does not include the imputed data.

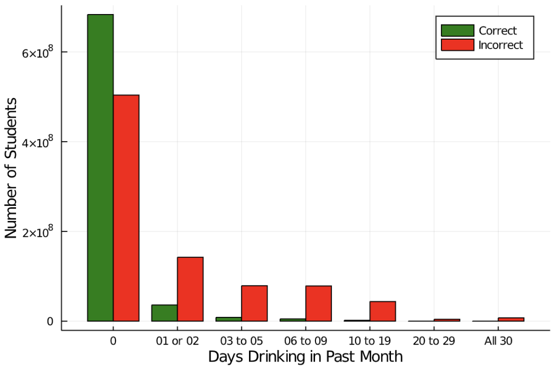

Figure 11: Final Model on the Test Set

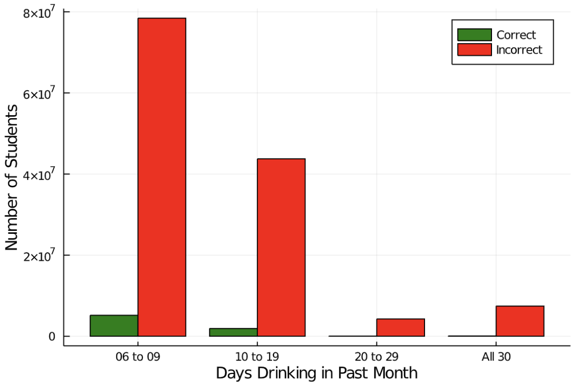

Figure 12: Final Model on the Test Set Outliers

Once we've chosen our final model using the validation set, to ensure that we were not choosing a model that overfit to the validation set, we test our model on the test set. The results can be seen in Figure 11 and Figure 12. The accuracy is 46.1%, which is a bit worse than our validation accuracy, but not enough to suspect serious overfitting. We can also see the same general trends that we saw in the validation set. Most of the correct predictions are in the 0 days drinking group. The model still has some correct predictions on moderate drinkers, 6 to 9 and 10 to 19 days, and though it does predict heavy drinkers, unlike the GLRM with ordinal loss models, it typically misclassifies these heavy drinkers.

One downside of this model is the lack of intelligibility. Though using a black-box model like the random forest model greatly increases our accuracy on this task, it decreases the risk factors that we can see, as we cannot tell which features the trees within the random forest decided to split on. This means that if this model were to be used in the real world, it would be unable to tell concerned parties what risk features were, but instead only if the teenager in question was at risk or not. This lessens its ability to generalize, since this model will have to be run for each individual, instead of helping people learn risk factors to look for.

Finally, though this model does much better than random chance, even on the test set, we do not have much confidence in our model. For those that the model classified as heavy drinkers (20-29 days or all 30 days) this model only has 1% accuracy. Since these classifications are the ones we care most about, this low accuracy reduces our confidence in the model.

Weapon of Math Destruction

Unfortunately, this project could be a weapon of math destruction. None of our fitted models could be 100% accurate all the time, meaning that some students may be predicted to drink more than they do, and some may be predicted to drink less than they do. A false positive, i.e. the model predicting that a student drinks when they don't, could result in that student getting unfairly accused and punished for something they are not actually doing. On the other hand, the case of a false negative, i.e. the model predicting that a student does not drink when they do, could cause a lot of harm. If people put too much trust into the model, false negatives could lead students who are heavy drinkers to not get any of the help they desperately need. The model could let these students slip through the cracks, and ultimately result in their situation getting worse over time.

Fairness

Furthermore, it is important to take into consideration fairness of our model, as different protected classes are included different features in our data set, such as race, sex, gender identity, sexual orientation, etc. Even if we were to try to directly exclude these factors from our models, there are so many other features in our data set that these classes could indirectly be effecting. Especially in the United States, race, for example, is correlated with so many factors in students' lives, that it would be nearly impossible to fully remove any racial bias from our models. Therefore, when applying this model, it is important to be cognizant of the difference in predictions for the various genders, races, sexual identities, etc. of the students.

Conclusion

Ultimately, our best model, the Random Forest model, is not great. Even though it is significantly better than random chance, where the probability of choosing each classification of days drinking is about 14%, it still is not predicting heavy teenage drinking significantly well. Looking at Figure 11, we can see that thousands of students were incorrectly predicted to drink 0 days per month, when they actually have had a drink on > 1 day that month. This brings in what we mentioned above in Weapon of Math Destruction, where if this model was being used in a school environment, these thousands of students would be slipping through the cracks and not getting the attention they possibly might need from their guidance counselors and mentors.

This model could be a useful tool in a real school environment, as it is significantly better than random chance. However, it should not be the only metric used to determine whether students are likely to be heavy drinkers or not. It should be one factor of many for guidance counselors, mentors, parents, etc. to reference, but we are confident that it should not be the sole indicator of what sort of support students should be receiving.

Possible future work that could be pursued from this project include focusing on fitting a model where this Type II error, these false negatives, is minimized, to reduce the harm and neglect this model could cause. Additionally, although we did not discuss it in this class this semester, the idea of fairness and quantitatively examining the prediction errors for different protected classes would be incredibly important to pursue before attempting to use this model in a school environment.