ORIE 4741: Midterm

Anna Halldorsdottir, Stephanie Hellman

Introduction

Our goal is to identify risk factors for drinking in teenagers. 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 is a change from our original data, which was a smaller data set focusing on Portugal in 2005. We found this data to be too small, old, and perhaps not applicable to the United States, so we have changed our data set.

This survey gives us data on student's age, gender, race, safety behaviors, violent behaviors, bullying history, tobacco, ecig, alcohol, marijuana, and drug usage 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.

Thus, our data is now 199,642 observations by 303 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

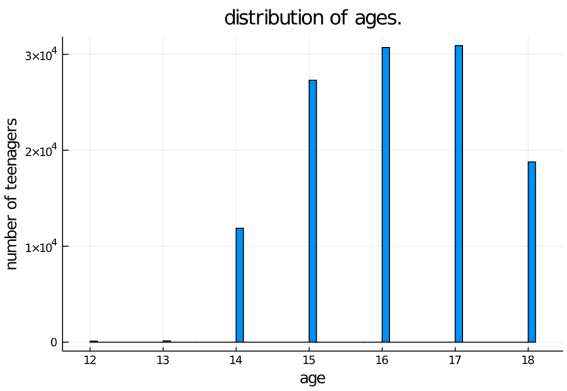

Figure 2

Preliminary Model and Results

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, which we do not touch in this report (it will appear in our final report when we have chosen a final model). 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.

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.

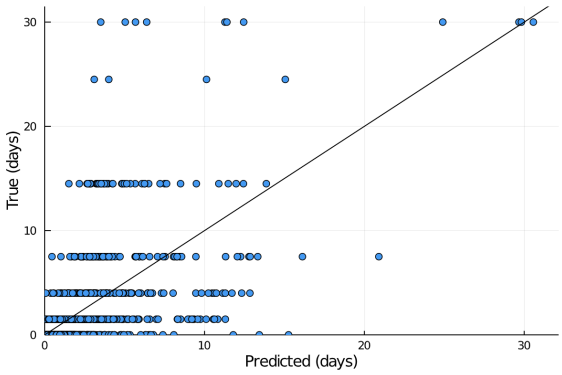

Figure 3: L2 Loss Model

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 is alright at predicting our data, 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.

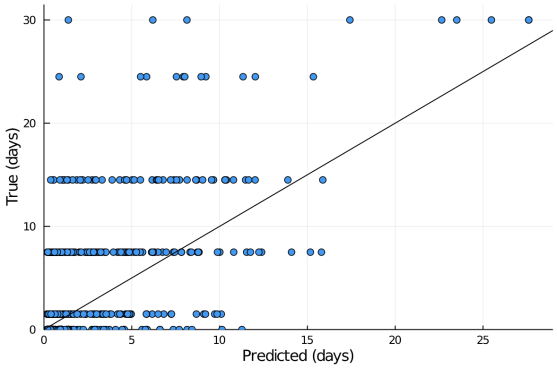

Figure 4: Huber Loss Model

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.

Takeaways and Ideas for the Future

Most of the data we have, due to the set up of the survey, is best classified as ordinal data. Therefore, linear regression does not seem to be classifying our data very well. There's a few ways that we plan on attempting to explore this further. The first is decision trees. We chose decision trees because it's easier to split each ordinal value into nodes, and traverse through the trees with all of our ordinal values in that way. Additionally, we want to attempt changing the target into an ordinal, and then fitting an ordinal regression and using ordinal loss.

Additionally, we found that hard drug use, specifically cocaine and heroin usage, seems to be a strong indicator of heavy alcohol use in both our models. We are fairly concerned about this; if kids are not open about their alcohol usage to their parents and mentors, then they probably would not be open about drug use either. Therefore, we will also be focusing on how to fit a model without including the hard drugs as features.