**Predicting High School Graduation: Tennessee Star Student Survey**

Andrew Clyde & Asher Paules-Bronet

University of Oregon – Department of Economics

March 10, 2021

**Introduction:**

We wanted to see if there was a way to predict high school graduation from the Tennessee Star Student data set. We used Lasso, Elasticnet, Support Vector Machine and Decision Tree regression and classification methods. We found that they all gave very similar results with just under 90% accuracy.

**Abstract**:

A common theme of most prevailing research of early childhood development, is that humans are inherently born with (nature) and develop throughout early stages of life (nurture), specific skills, abilities and characteristics that are highly correlated and therefore predictive of ability to succeed in high school. Our research question focuses specifically on predicting high school graduation, this is an important subject to study because if we can accurately predict high school graduation, we can then narrow down the predictive variables that have the greatest correlation with our outcome of interest. Researching inherent characteristics and circumstantial influences on students’ academic and subsequent life successes can help influence future public policy decisions.

**Data**:

We were able to gather the necessary data for this research paper from, the Tennessee Star Student Experience Survey, which is composed of 11,000 observations of 381 variables. The first issue we encountered was our outcome variable, high school graduation, had a tremendous number of missing observations. It does not make sense to impute onto our outcome variable due to implicit bias any method would impose. So, we filtered out all missing observations of our outcome variable, leaving approximately 5,000 complete observations to work with. The next issue we came across was imputing missing observations of our categorical variables. The preferred imputation method, nearest neighbor, was too computationally intensive for our computers processing power, due to the size of our data frame. We resolved this issue by imputing categorical variables with the mode of the data set, which is admittedly a lesser imputation method. Finally, we had concerns regarding, whether or not our predictors were rich enough to accurately predict our outcome.

**Methods:**

The methods we used were Lasso, Elasticnet, Support Vector Machine, and Decision Tree. We ran regression methods for both Lasso and Elasticnet, while SVM and Decision Tree were classification models, this means we had to change our process a bit to incorporate different modelling. For the regression methods we had, after we got the best fit model from the cross validations, we needed to use the prediction to determine if either a person graduated (=1) or did not (=0). Therefore, if the model predicted above 0.5 to graduate then they graduated, otherwise they did not. For the classification models, we needed to alter the recipe a bit so that graduated was seen as a factor. For Lasso, we tuned using different lambda penalties and a 10-fold cross validation while minimizing RMSE. For Elasticnet, we tuned penalty and mixture with what tidy models has built in with a 10-fold cross validation while also minimizing RMSE. For the classification models we were maximizing accuracy because we want the most accurate overall model without weighting accuracy for failing or graduating. We also did 5-fold cross validation for both models because they took a very long time already to run. The Support Vector Machine we tuned the cost, while the Decision Tree we tuned the cost complexity with a tree depth of 1, 2, 5, and 10.

**Results and Conclusions:**

By looking at the confusion matrices and accuracy of models, it seems as though they all get most of the same predictions for each datapoint, with minor variation between models. The Decision Tree does the best based on accuracy, while the other three are marginally worse. I think missing data and the richness of our predictors limited our success. However, with Asher and I were both surprised we were able to predict graduating accurately almost 90% of the time. We learned high school GPA, participation in the ACT, seventh grade language and math scores and whether you were present for Kindergarten are the most predictive variables in our data set. If we then were to dig deeper into those predictors to see what influenced them, we could advise better public education programs.