Executive Summary

Most universities and colleges across the United States have experienced decreases in enrollment within the last few years. To remain sustainable, these institutions must not only monitor their enrollment rates but also retention rates. Retention is important for their ranking and for attracting new students. Accurate prediction of retention rates based on different variables such as enrollment numbers, tuition cost, changes in tuition costs, student demographics, etc. In this project, we aimed to use data from over 2954 U.S. colleges and universities (years 2011 – 2016) to predict retention rates at the institution level by using four distinct prediction models such as Ridge, Lasso, KNN, and elasticnet. We built and evaluated 13 models to identify our best models for prediction of full-time student retention rates.

Big Picture

Our goal was to create at least four distinct prediction models to predict retention rates at U.S. colleges and universities based on 11 different variables (as originally downloaded; data set was widened for modeling). Retention rate prediction is important at the institution level because budgeting and goal setting are based on enrollment and retention numbers. Institutions for accurate expectations and plans when they have an accurate sense of the retention rates they will face in the future. This makes for a perfect prediction problem since we can use different variables connected to what we believe are correlated with or have effect on retention at institutions. Since retention affects ranking and growth of institutions, institutions must consider retention to remain sustainable.

Data

The data used is from the Integrated Postsecondary Education Data System (IPEDS) and is aggregated at the institution level, so no individual student information is considered. All institutions receiving federal funding are required to complete the IPEDS surveys, so this can be considered a good sample. We downloaded one large dataset then divided it into six datasets – one for each of the years beginning 2011 to 2016. Each of the annual datasets contained 23 variables and 2954 observations prior to generating additional variables based on relationships between sets of variables. Variables include but are not limited to tuition cost for in and out of state students living on campus, tuition cost for in and out of state students

living off campus, percent of undergraduates who are female, percent of undergraduate who are black, full time retention rate, number of undergraduates who are Pell grant recipient, etc.

There were two main challenges to cleaning the data. First, renaming the variables took a while because the downloaded data had entire sentences as variable names. The second difficulty in cleaning the data was getting the data organized into a form where we could calculate changes the four tuition rates over one- and two-year increments. On the other hand, the data has a selection of variables to choose from which helped with our prediction models. We also used preProcess to center and scale predictor values as prep for the penalized regression models, and we used preProcess to replace missing values with the column's median.

Methods

We used four model types to investigate our prediction problems: KNN and three penalized regression models: Ridge, Lasso, and Elasticnet. All three penalized regression models used tuned parameter of lambda = 0.01 and the KNN model used cross validation of 5 and provided a sequence of lambdas for ridge and lasso that ranges from (0.1, 0.01, 0.001, 0.0001, 0.00001) and alpha for the logistic model that range from (0, 1, 0.1). We started out by having each one of us create a model from each type using data identified by-institution-by-year. Rebecca used the glmnet package directly while Nina used the caret package. Once we figured out the data processing for models that considered changes in tuition over time and that held institution as a fixed effect, we ran another set of models on that data using caret. We compared training and test error rates from each model to note if any models showed too much variance. To tune our models, we used the "train()" function on Ridge, Lasso, and Elasticnet.

Results and Conclusion

We measured the performance of models using RMSE and MAE. Our outcome, retention rate, was measured as whole number representing the percentage of full-time students retained from the previous school year (other than those that graduated). The top performing models in terms of both RMSE and MAE were the elasticnet, lasso, and ridge regressions using the time series data, followed by the KNN model that did not use time series. The top performing models did not have a lot of spread between their RMSE and MAE. The best performing model, elasticnet using time-series data, had a test RMSE of 7.39 and a test MAE of 7.39. The fourth best performing model, KNN using by-institution-by-year had a test RMSE of 7.92 and a test MAE of 5.11. All other models performed very closely together between Rebecca's and Nina's models as well as between training and test data, there were only two outliers with much higher

error rates, the KNN model using time-series data and one of the ridge models using the by-institution-by-year only data. We learned that prediction models can get us close to our true predictions, but we also see the limitation of our models in that we could not get more accurate than being an average of 4 percentage points off-target. We think that using more lag variables and fixed effect models on this type of panel data could potentially give us better prediction outcomes. We also took this opportunity to learn project-management interface GitKraken, a GUI for GitHub, which is widely used in the workforce.