Machine Learning

1. How do you frame your main question as a machine learning problem? Is it a supervised or unsupervised problem? If it is supervised, is it a regression or a classification?

Original Research questions:

What is the average retention rate for colleges in the state of Georgia? What is the average student to faculty ratio for colleges in the state of Georgia? What is the relationship between retention rate and student to faculty ratio?

Research Question with Machine Learning in Mind:

Can we predict a college’s retention rate based on the student teacher ratio? Or Can we predict a college’s retention rate based on its student teacher ratio and other variables such as the population of female and male students? This problem is a supervised problem because the data in the data set is formatted. Based on the variable of prediction, we frame the issue as a regression.

2. What are the main features (also called independent variables or predictors) that you'll use?

The independent variables are student to faculty ratio, number of males, and number of females, and the output variable is retention rate.

3. Which machine learning technique will you use?

The data downloaded from the site [www.data.gov](http://www.data.gov) was given in a certain format. We are most interested in the Retention Rate and variables which may affect retention. The Retention data was not given at the student level, i.e. where 1 would specify that a student is retained, and 0 would specify that a student was not retained. Only summarized data is given, not 0’s and 1’s. Hence, the data given from the site was at the summary level. That is, the percentage of students who were retained, or returned was listed for each institution in the data set. That percentage depends on the total student population at each respective institution. Since the output variable is a continuous variable rounded to the nearest unit, a linear regression model will be used. The Retention rate is given as the percentage of students who return to the institution. The value is a summary value, that is if an institution has 5000 students and 3000 of the students return to the campus then the Retention rate is 3000/5000 or .6 which is 60%. Multiple variate linear regression seemed to fit best as there are more than one independent variable used to predict a college’s retention rate.

4. How will you evaluate the success of your machine learning technique? What metric will you use?

For this project, a training dataset was created with 80% of the data from the original dataset, and the testing data set includes 20% of the data.

i) The p-value, which determines whether a variable is statistically significant, will be used to determine a relationship between variables. The lower the p-value (ranging from 0 to 1), the more likely the model is more accurate at a certain level of confidence. The level of confidence is usually chosen at 95%.

ii) The R- squared value, which determines how “close” the points are to the regression line, is another metric that can be used to determine the validity of the model. The values range from 0 to 1, with values closer to 1 indicating that the data points are “closer” and more tightly correlated to the regression line. In layman’s terms, the closer the value is to 1, the better approximation of the data points the predictive model will give. If it is closer to 0, then the independent variable may need to be changed as it does not provide enough context or influence on the dependent variable and is thus not useful for the model.

iii) Root Mean Square Error, which gives a “standard deviation” of the data points; the average variance of each of the points from each other. This value can be infinite, but the smaller the RMSE, the better approximation – the values of the dataset are not as sparse. This is important as data points can line up well and be set up around another regression line that was not considered initially.