Machine Learning

You’ve collected your data, cleaned it up, wrangled it into shape and explored it. Now it’s time to perform some in-depth data analysis using machine learning. This step depends on you and your mentor, but here are some suggestions to get you going.

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 question: What is the passing rate of a student enrolled in Calculus I?

Research Question with Machine Learning in Mind: Using a matched data set, can we predict a student’s score in Calculus I based on gender, major, race, and weather s/he is a first-generation college student? It is a supervised problem because the data in the data set is formatted. Based on the variable of prediction, we frame the issue as linear regression.

Which machine learning technique will you use? Because it is a regression, we needed to decide between a logistic and linear regression. Multiple linear regression seemed to fit best as the student’s score is a continuous variable.

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.

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%.

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.

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.