Capstone Machine Learning

Grant Cox

January 11, 2019

Football and Regressions

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?

The main problem I think we're addressing with this abundance of data is: is my university going to make any money with a football program? So, if we were to frame this as a machine learning problem, we could take all of these data points and use some supervised learning and potentially create a regression equation. We could frame it more specially as either "how much money will my university make from the establishment of a football program?" or "will my university make any money with the establishment of a football program?"

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

The intuitive predictors, at this point, would seem to be: enrollment and tuition (variables independent of an athletic program) and then Athletic Expenses. These first two would point to school size being a greater or lesser determinant of profit, while Athletic Expenses would point to just how much investment might be deemed appropriate to predict a profit.

Which machine learning technique will you use?

I'll be using both a linear regression model as well as a logistic regression model to answer the questions framed above of "how much can we expect to make?" and "can we even expect a profit?" respectively.

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

The plan is to divide the "combined_df" I've been using, 130 universities and 5 years for each university, into a training set of 75% and a test set of the other 25%. The model that performs the best on the test set and gives the lowest R squared or adjusted R squared, will be the most successful.