Machine Learning Final Project

Analysis of College Status

Brandon Hosley

Dr. Svet Braynov

Machine Learning Final Project

Analysis of College Status

Introduction

The goal of this project is to develop models that will predict the elite status of a University. The status of a University will be considered elite if over half of the population of its students were measured in the top ten percent of their high school class.

Analysis by Artificial Neural Network

The initial, single node Artificial Neural Network, did not have a high correlation. To try to increase the accuracy of the model I increase the number of hidden nodes, surprisingly, the increase in nodes caused the reported correlation between predicted and actual results to correlate less. 0.78 and 0.65 respectfully.

Analysis by C5.0

Analysis using C5.0 produced fairly good results with an accuracy of 92.4%. With results that high I did not feel that much could be accomplished with my manual tuning.

Use the Caret Library to Tune the Above Algorithms

Using the Caret library's automatic tuning I was able to get similar results to the above implementation of the algorithms.

Initially, 98.6% for the Neural Network and perfect results from the C5.0 implementation. This appears to have been a result of over-fitting as the entire set was used in the initial training, to test this the separated sets used above were applied and confirmed a more realistic set of results with 92.8% for the Neural Network and 98.3% for the C5.0.

Good, but modest improvements.

Analysis Using an Ensemble Algorithm

For this project's ensemble algorithm I chose to use the Random Forest. From it I was able to get a perfect model when trained on the entire set, I believe that this is one of the over-fit situations. When trained on the Neural Network's training set the model gave a 92.3% accuracy and the C5.0 training set gave a 91% accuracy. When tuned using caret and a few options similar to those given in the lectures I found the accuracy to be substantially higher than with my previous data sets.

Comparison of Results

The C5.0 method produced better results and was faster than the Neural Network algorithm. It seemed to be better suited for this type of problem. Unsurprisingly, the automated tuning included in the Caret package performed better tuning than I was able to.

The most surprising thing was that the random forest produced a less accurate model in the cases of both data sets than the algorithms used earlier. This may be a good thing however, as it is probable that this model is not overfitted to the data, it was also trained and tested across the whole of the data set. I believe that one of the reasons for the high accuracy of the models was the high correlation between certain factors, such as the number of top 25% students attending the university.