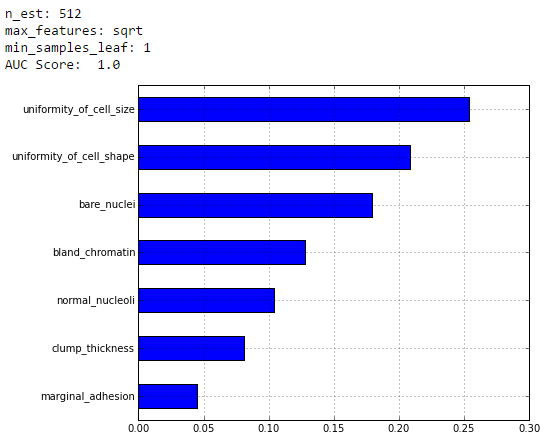
Breast Cancer Prediction

I was successful in my attempt to create a machine learning model to predict malignant tumors. After building a model, I was able to get an AUC score of 1.0.

In order to test my model, I broke the data up into one chunk consisting of 80% of the data and another chunk consisting of the remaining 20% of the data. I trained the random forest model with the 80% chunk of the data. I had also written a few functions that optimized the parameters pertaining to the model in order to get best predictions. After training the data, I decided to make a prediction and check a chart to see which variables to model found most useful. After reading through the chart, I used that to determine which information I would be able to drop from the data set that wasn’t necessarily needed or useful. The first variable that I dropped was the id number, as an ID number would not help in determining malignant cells or not. After retraining the model, I had also found that the mitoses variable wasn’t very useful in the model either, so that was dropped. After another retraining of the model, it had determined that the epithelial cell size was not useful either. After dropping that final variable, I was able to get my final result. Displayed below is the final chart used to get the 1.0 AUC score including some of the parameters used to train the model.



As you can see, the variables had a steady decline in importance to the model. The top three most important variables in this model were the uniformity of cell size, the uniformity of cell shape, and the bare nuclei.