A beach-health beauty contest

Wesley Brooks

1 Goal

There are many machine learning methods that

2 Methods

Comparing models The purpose of a regression model for beach-health management is to aid the decision of whether to post the beach. Since the result of the model is a yes/no decision, the models in this experiment were compared on the basis of how often they advised the correct decision. The frequency with which the model correctly advised posting the beach is the sensitivity; the frequency with which the model correctly advised *not* posting the beach is the specificity.

Cross-validation Five-fold cross validation was used to test the model's performance in prediction. The procedure was to break the data randomly into five sections ("folds") of equal size. One fold was reserved as the test set, and the other four were used to train a model. That model was used to predict exceedances on the test set. the process was repeated five times so each fold was used as the test set once.

One concern with cross validation is that the results depend on the way that the data was broken randomly into folds. In order to avoid introducing bias by the division of the data into the cross validation folds, the process was repeated 1000 times with new, randomly drawn folds each time.

2.1 Partial Least Squares

Partial Least Squares is quite similar to Principal Components Analysis except that the components are chosen in a way that ensures they are maximally correlated with the output.

2.2 Generalized Additive Model

Generalized Additive Models (GAMs) were first described by Hastie and Tibshirani. GAMs were developed as a higher-dimensional generalization of smoothing splines.

2.3 Boosted decision tree

The boosted decision tree uses the gbm package in R.