Daniel Prusinski
Predict 412

Programming Assignment 3: Classification Models

A famous saying states that the only certainties in life are death and taxes. Since the age of email, one could amend the saying as follows; the only certainties in life are death, taxes, and spam. Thankfully, email providers like Google and Yahoo employ classification algorithms to aid in reducing unwanted spam email.

The following exploratory data analysis (EDA) employs various classification methods on the Spambase dataset. All outputs, graphs, and visualizations are found within Appendix 1. Within the Spambase dataset, there are 4,601 observations and 58 unique variables (Output 1), of which 2/3 will be partitioned into a training set and 1/3 into a testing set. Through the EDA, the end goal is to predict the binary variable "class", which is either spam or email. Given that the response variable is binary, the EDA must utilize modeling techniques that are appropriate for this specification.

Predicting the binary response for the variable "class" is known as a classification problem. Assessing model sensitivity and specificity is a tradeoff that will be analyzed from a receiver operating curve (ROC), of which will be shown in the form of a confusion matrix. Different datasets favor sensitivity versus specificity, based on the modeler's preferences. For example, the medical field favors specificity to false negatives based on the repercussions, and for this EDA a false negative is also more important because an email falsely marked as spam has the potential of being deleted without the user viewing the message.

Random Forest is an ensemble, classification method that utilizes tree-structured classification to build a strong prediction (Miller 2013). I am new to this modeling technique,

but it is known for being robust as well as superior for datasets with many explanatory variables. Output 2 shows a plot of the importance of the predictor variables based on Mean Decrease Accuracy and Gini. Of the 58 variables, 31 variables appear in each index with a majority overlap. After fitting the random forest modeling technique to the training dataset, output 3 shows the confusion matrix. The model only misclassified 103 email messages as spam, which is the most important metric of the analysis. When predicting on the training dataset, the random forest improved its false positive to 5 email message, and an accuracy of 98% (Output 4). On the testing dataset, the random forest model had 27 false negatives and a 95% rate of accuracy. As described earlier, this analysis is concerned with the specificity, number of false negatives, in the confusion matrix.

Five different modeling techniques were used to find the best predictor of spam focusing on specificity.

Model Comparison Matrix				
Model	Specificity Training Data	Specificity Testing Data	% Change In Accuracy	Rank
Random Forest	.970	.910	-0.062	2
Neural Network	0.953	0.888	-0.068	4
Vector Machine	0.911	0.888	-0.025	3
SW Logistic Regression	0.9	0.874	-0.029	5
Naive Bayes	0.942	0.927	-0.016	1*

Outputs 4 through 10 show the details from the models. Naive Bayes classifier had the least drop in accuracy from the training data set to the testing dataset. The asterisk next to the number 1 rank means more analysis must be done to verify that the variables are independent from one another, an assumption of Naive Bayes. Random Forest and Vector Machine performed well in my opinion based on the accuracy in both datasets along with a mild percent change in in accuracy. Moving forward, I would prefer to hone in on Naive Bayes and Random forest as modeling techniques and test the specificity on a larger testing data set.