STAT 5630 - Statistical Machine Learning

due monday november 4, 2019

Mini-Project

Max Ryoo (hr2ee)

Professor: Xiwei (Denny) Tang

problem 1 *SPAM vs NOT SPAM*

(a) Exploratory Data Analysis of Response Variable

(a) **Training Set**
The ratio of the training set was fairly evenly split between the two categorical variables. The proportion of spam entries in the training dataset was 0.397389885807504, while the not Spam entries was 0.602610114192496. This was roughly a 6 : 4 split in the response variables for the dataset given.

(b) **Testing Set**
The ratio of the training set was similar. The exact measurements are 0.387369791666667 and 0.612630208333333 for Spam entries and not Spam entries respectively.

(b) Exploratory Data Analysis of Explanatory Variables
There were a total of 58 variables from the data that was given for the scope of this project. The first variables $V1 - V55$ were continuous variables, while the last three variables $V56, V57, V58$ were categorical variables.

(c) LDA and QDA with predictors $V55 - V57$

(a) **LDA**
The following was extracted from running LDA on the last three categorical variables of the dataset.

Accuracy	Sensitivity	Specificity
0.6692708	0.1932773	0.9702444

(b) **QDA**
The following was extracted from running QDA on the last three categorical variables of the dataset.

Accuracy	Sensitivity	Specificity
0.6816406	0.2403361	0.9606801

(c) **LDA vs QDA (3 Variables**
The LDA Accuracy measure and LDA Sensitivity measures better fit the dataset than the QDA model. However, for Specificity the QDA performed slightly better. The values were very close to each other regardless.

(d) LDA and QDA with all predictors

(a) **LDA** The following was extracted from running LDA with all the variables given to us in the dataset.

Accuracy	Sensitivity	Specificity
0.8776042	0.7731092	0.9436769

(b) **QDA**
The following was extracted from running QDA with all the variables given to us in the dataset.

Accuracy	Sensitivity	Specificity
0.8255208	0.9445378	0.7502657

(c) **Difference in LDA**
The following is the difference in LDA between three predictors and all predictors (All - three)

Accuracy	Sensitivity	Specificity
0.2083333	0.5798319	−0.02656748

(d) **Difference in QDA**
The following is the difference in QDA between three predictors and all predictors (All - three)

Accuracy	Sensitivity	Specificity
0.1438802	0.7042017	−0.2104145

(e) **Differences Comparison**
As one can see the Accuracy and Sensitivity increased while specificity decreased. In both QDA an LDA Sensitivity increased by a great factor while accuracy and Specificity changed only by a small factor in comparison to the sensitivity change. The increase in Sensitivity while specificity decreases seems natural due to the relationship between Sensitivity and Specificity.

(e) Logistic Regression Model & linear SVM with all predictors

(a) **Logistic Regression Model** The following is from running Logistic Regression Model with all the predictors.

Accuracy	Sensitivity	Specificity
0.9238281	0.8823529	0.9500531

(b) **Linear SVM** The following is from running a Linear SVM model with all the predictors.

Accuracy	Sensitivity	Specificity
0.9114583	0.9277311	0.901169

(f) Non-linear SVM with only three predictors

(a) **Non-linear SVM**
The following is from running Non-linear SVM with only three predictors.

Accuracy	Sensitivity	Specificity
0.6132812	0.003361345	0.002125399

(b) **Comparison**
The following is the difference between Non-linear SVM and linear SVM (Non-linear - linear)

Accuracy	Sensitivity	Specificity
−0.2981771	−0.9243697	−0.8990436

Interestingly enough, all measures of model performance decreased with the incorporation of non-linear SVM. This may be due to the fact that less predictors were used.

(g) Flip coin probability

(a) **Performance Measure** The following is from doing random guessing with the probability of 0.5.

Accuracy	Sensitivity	Specificity
0.5078125	0.4789916	0.5260361

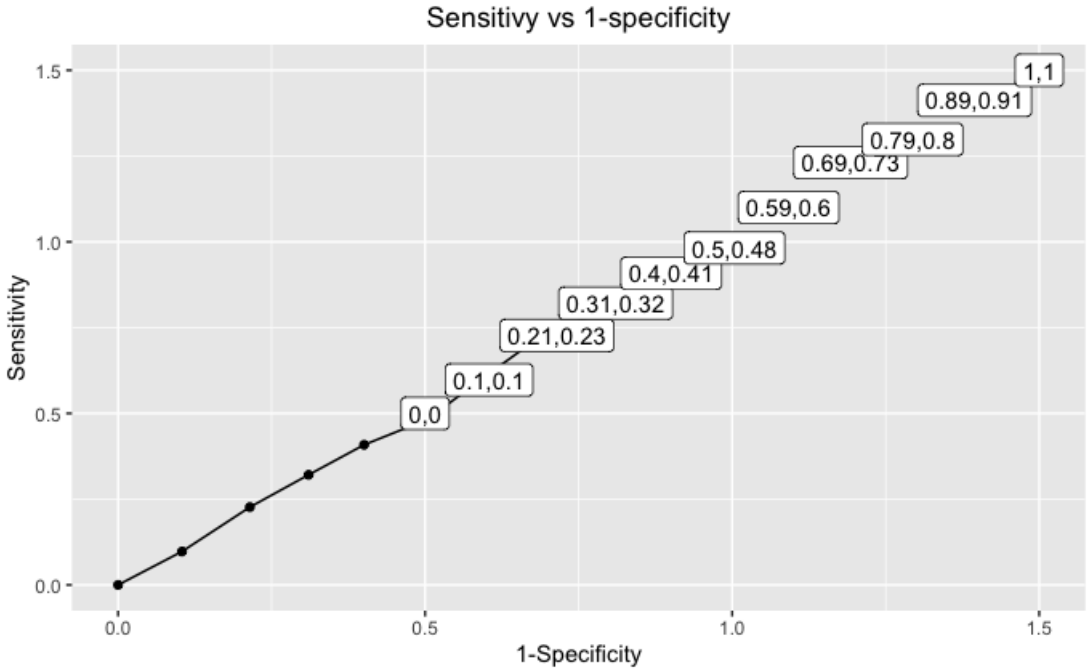
(h) Different random probabilities

(a) **Table of Values**

The following plot is made from having different probabilities for random guessing the binary classification.

Probability	Sensitivity	1 - Specificity	Sensitivity + Specificity
0.0	0.0	0.0	1.0000000
0.1	0.09747899	0.1041445	0.9933345
0.2	0.2268908	0.2146652	1.012226
0.3	0.3210084	0.3103082	1.0107
0.4	0.4084034	0.4006376	1.007766
0.5	0.4823529	0.5037194	0.9786335
0.6	0.6	0.5919235	1.008077
0.7	0.7310924	0.6928799	1.038213
0.8	0.7983193	0.7938363	1.004483
0.9	0.912605	0.8947928	1.017812
1.0	1.0	1.0	1.0000000

(b) **Graphical Representation**



(c) **Best Probability**

The best probability to Maximize the Sensitivity + Specificity in this case was $p = 0.7$. For the probability of $p = 0.7$, the Sensitivity measure was 0.7310924; Specificity measure was 0.3071201; 1-Specificity was 0.6928799. The measure of Sensitivity + Specificity was 1.038213, which was the highest out of all the probabilities explored in the scope of this problem. However, the Sensitivity + Specificity measures were similar to each other with every probability, which was around 1.00. Therefore, for this trial with a setting of seed of 1 it happened to be 0.7 probability. With different settings of seed the value yielding the best sensitivity + specificity can change.

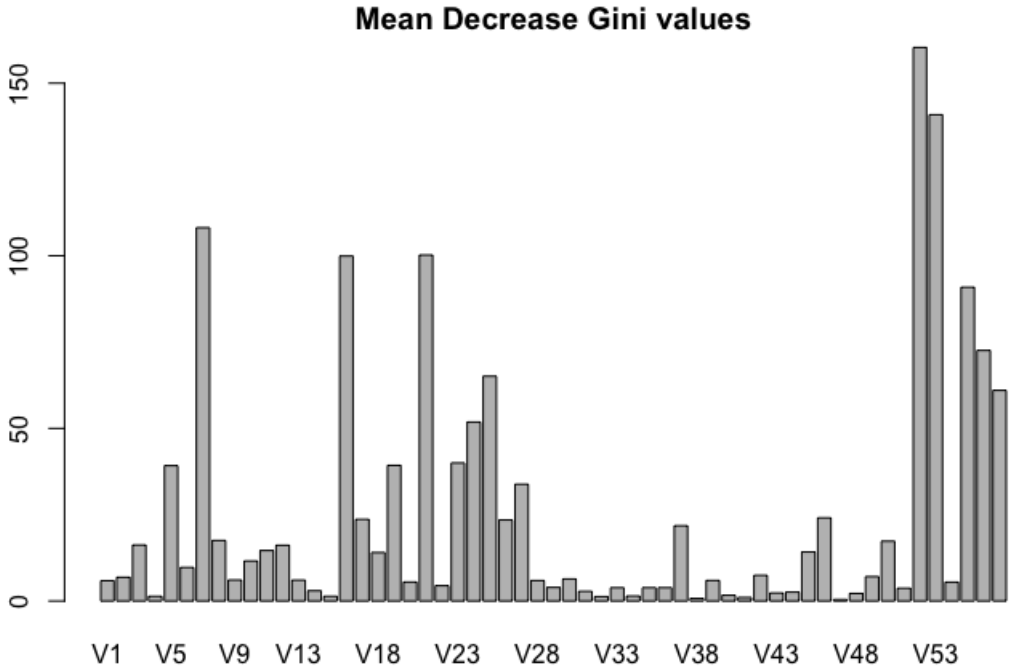
(i) Random Forest

(a) **RF Measurements**
The following was acquired from making a random forest model to fit our data using 500 total number of trees and mtry value of $\sqrt{p}$. The node size was tuned using five folder cross validation.

Accuracy	Sensitivity	Specificity
0.952474	0.9310924	0.9659936

(j) Importance (MeanDecreaseGini)

(a) **Graphical Representation of Mean Decrease Gini Values**



(k) Boosting Model

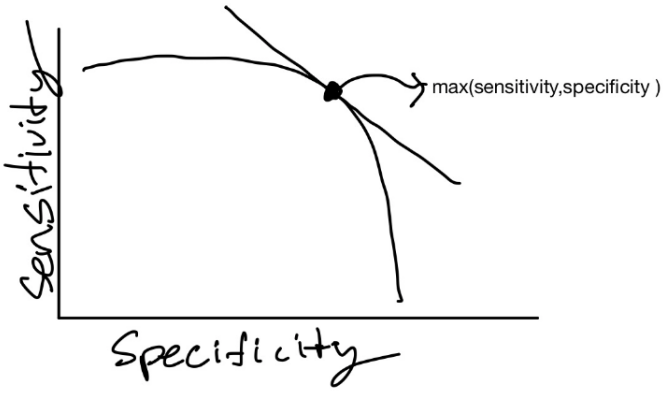
(a) **Selected tuning parameter values**
The final values used for the model were n.trees = **2000**, interaction.depth= **1**, shrinkage = **0.07** and n.minobsinnode = **10**. The interaction.depth measure and n.minobsinnode were set to default measures for gbm.

(b) **Performance Measure**
The following is the performance output by running a boosting model.

Accuracy	Sensitivity	Specificity
0.9440104	0.9159664	0.9617428

PROBLEM 2 *Problem 2 (Bonus)*

1. **Maximization of (Sensitivity + Specificity)**



2. **What is the lower bound of the area under the curve (AUC)**
   Wild guess: $0.5 * 0.5 = 0.25$. This guess is because the worst case model in binary classification is flipping a coin. This yields to a 0.5 sensitivity and specificity. Therefore the lower bound is estimated to be 0.25