Machine Learning Engineer Nanodegree

Capstone Project

Bora Pajo, PhD

April 6, 2017

I. Definition

Project Overview

According to the Centers for Disease Control and Prevention (CDC) breast cancer is the most common type of cancer for women regardless of race and ethnicity (CDC, 2016). Around 220,000 women are diagnosed with breast cancer each year in the United States (CDC, 2016). Although we may not be aware of all the factors contributing in developing breast cancer, certain attributes such as family history, age, obesity, alcohol and tobacco use have been identified from research studies on this topic (DeSantis, Ma, Bryan, & Jemal, 2014).

Problem Statement

This project focuses in investigating the probability of predicting the type of breast cancer (malignant or benign) from the given characteristics of breast mass computed from digitized images. The cases provided, are cases diagnosed with some type of tumor, but only some of them (approximately 37%) are malignant. This project will examine the data available and attempt to predict the possibility that a breast cancer diagnosis is malignant or benign based on the attributes collected from the breast mass.

Dataset and Inputs

The characteristics of the cell nuclei have been captured in the images and a classification methods which uses linear programming to construct a decision line. The dataset is published by Kaggle and taken from the University of California Irvine (UCI) machine learning repository. The data is taken from the Breast Cancer Wisconsin Center. It includes ten (10) attributes taken from each cell nucleus as well as ID and the diagnosis (M=malignant, B=benign). The dataset has 570 cases and 31 variables.

Evaluation metric

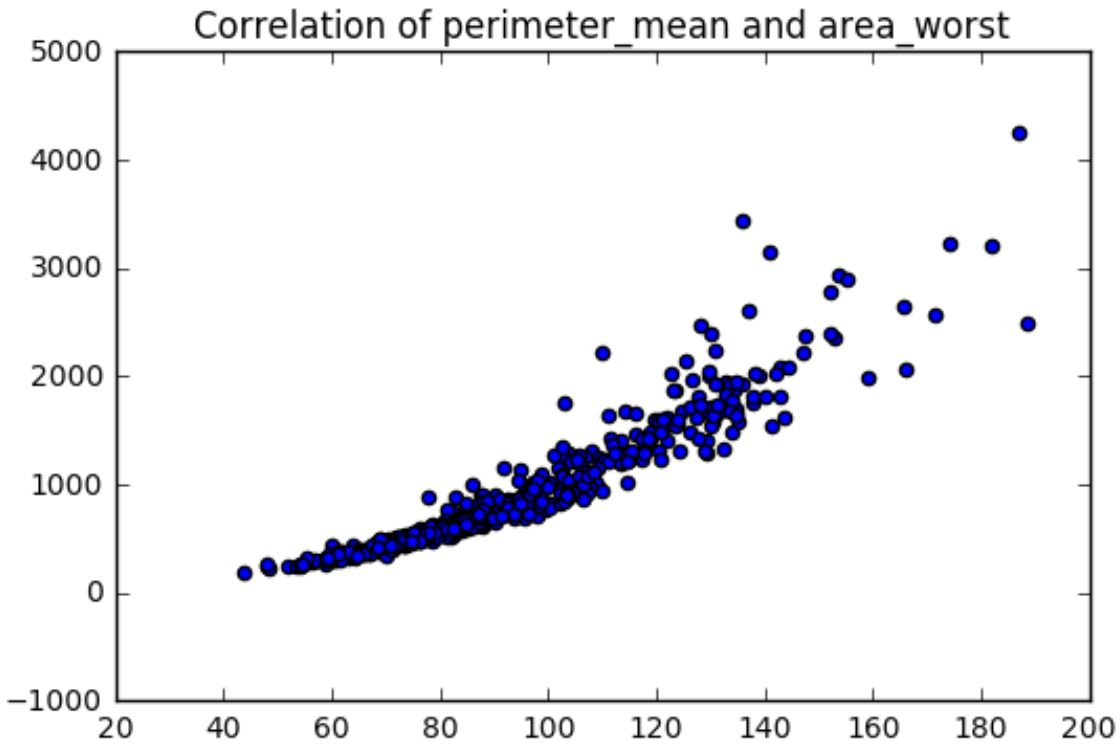
F1 score is a measure of accuracy or the ratio of the data that was accurately predicted. The closer the F-score is to 1 the best the prediction is and the closer to 0 it is, the worse the prediction. F-score considers the true positives and the true negatives, and is best used when comparing various classifiers as I am proposing to do in this dataset. From the literature, I reached the conclusion that F-score is the best evaluation metric to be used for this type of classification problem. The formula for the F1 score from the sklearn documentation is F1 = 2* (precision * recall) / (precision + recall).

II. Analysis

Data Exploration and Visualization

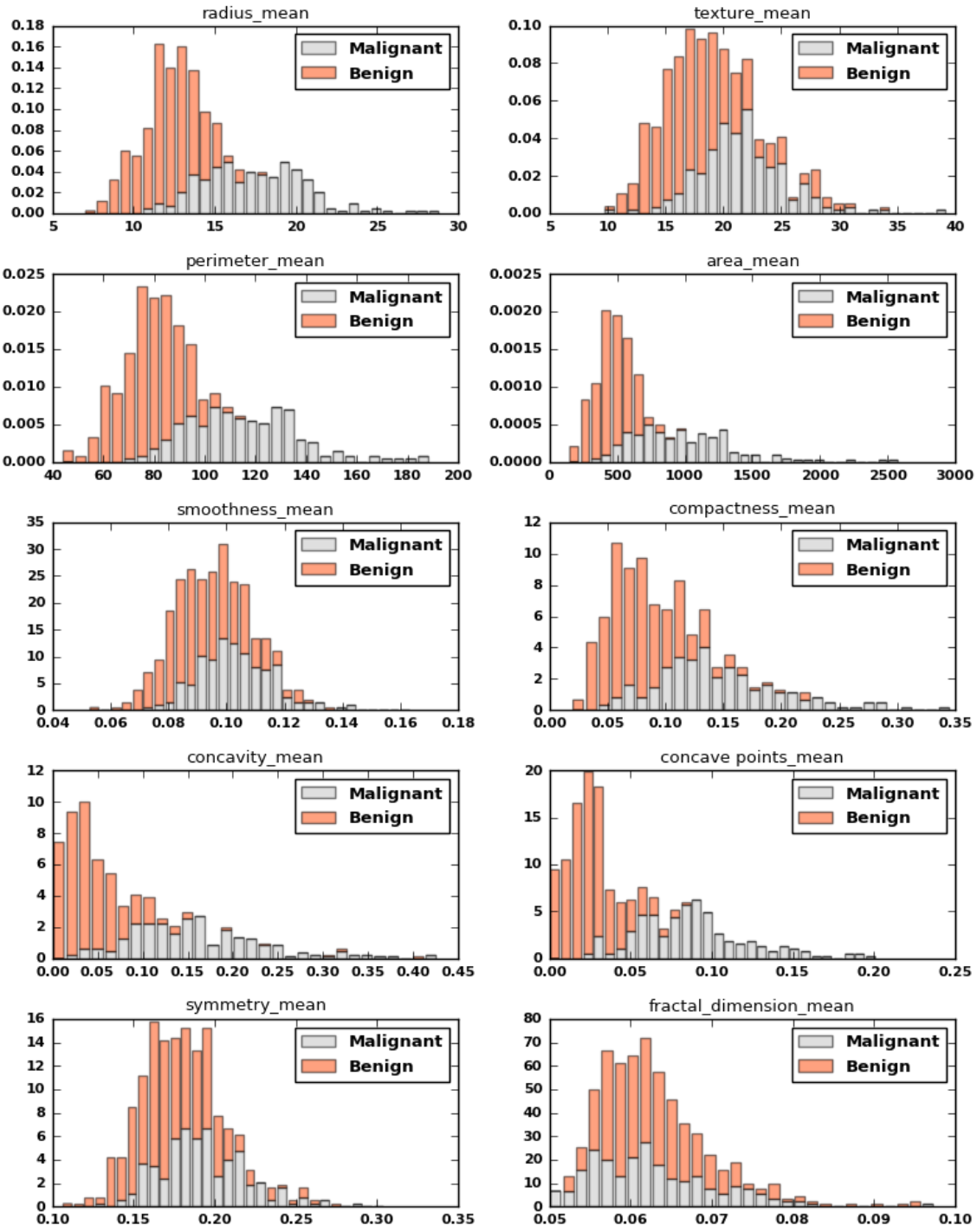
The dataset is taken from the UCI machine learning repository. It includes 569 cases of women diagnosed with cancer. There are 32 variables in the dataset (31 variables if we exclude the ID variable). Each variable includes information of the cell nucleus in the images captured. There are only 10 such attributes expressed in these 31 variables. These 10 attributes are: 1) radius, 2) texture, 3) perimeter, 4) area, 5) smoothness, 6) compactness, 7) concavity, 8) symmetry, 9) concave points, and 10) fractal dimensions. The dataset includes 212 cases diagnosed as malignant, and 357 cases diagnosed as benign. The ratio of malignant cases is 37.25%. Variables that relate to the same attribute are highly correlated with each other (as expected). Figure 1 below shows one of these examples where the perimeter mean and radius worst are highly correlated with a Pearsons' r =.97) which means that the correlation of determination is r^2=.94. This correlation of determination indicates that 94% of the changes in perimeter mean are explained by changes in the radius worst.

Figure 1: Correlation of perimeter mean and area worst



In addition, we can see that the malignant cases have higher values in almost every attribute when compared to benign cases. The averages of radius, perimeter, smoothness, texture, area, concavity, symmetry, concave points, and fractal dimensions of almost every attribute depict a visible difference in values where the malignant cases have much higher values. It is perhaps a sign of how spread out the tumor is compared to benign cases. Figure 2 show all these averages.

Figure 2: Stacked charts of the attributes for both cases



Benchmark Model

The highest accuracy the better the model will be predicting whether a person diagnosed with breast cancer has a benign or malignant tumor. To form a solid idea on the benchmark models, I perused the research literature on the same topic – predicting the

type of cancer from computer-generated images. One study that uses mammogram images of 200 cases and utilizes both k-nearest and SVM methods to classify the type of cancer achieved specificity of 92.10% with standard deviation of 2.75 and *accuracy levels of 92.16%* with standard deviation of 3.60 (Zhang, Wang & Yang, 2016). Another study that suggests the use of magnetic resonance imaging (MRI) instead of ultrasound to detect the type of breast cancer among diagnosed women, utilized a sample of 110 lymph nodes from pre-operative MRIs. They *achieved an accuracy level of 79.1%,* and was considered much superior to the ultrasound (Chung, Hyun, Kim, Gweon, Kim, Ryu, & Son, 2014). A third study that also compares the accuracy levels of ultrasounds and mammography, found that mammography images are a better predictor than ultrasounds at accuracy levels of 90.7% (Tozaki & Fukuma, 2011). Judging from these three different studies on this similar topic, I calculated an average of accuracy levels of 87.32%. Therefore, my benchmark levels will be 87% or higher.

III. Methodology

Data Preprocessing

This project will utilize a number of different classifiers to see which produces the best results. To be able to conduct the proper analyses, it is necessary to take a few of preprocessing steps, such as dropping the ID variable and changing the values of the target variable (malignant or benign) into a numerical variable. The target variable, the variable I am trying to predict is collected by notations 'M' and 'B' where 'M' indicates malignant and 'B' indicated benign. For best results during the analysis, this variable is preprocessed into 1 and 0 where 1 indicated malignant cases and 0 indicates benign cases.

Implementation

First, the dataset was split into training and testing sets randomly. Training set included 400 cases and testing set included 169 cases. The rationale for this division was to see how many cases were the optimal number for training the data with each different classifier and how long it took in each case. To implement different classifiers, I used training sets of 100, 200, 300, and 400 and got the results of training time, F-score for

training, predicting time for the test data and the F-score for the testing set based on each case. The classifiers utilized were: 1) K-nearest neighbor, 2) Decision trees, 3) SVC, 4) Naïve Bayes, 5) Random Forest, 6) AdaBoost, 7) QDA, and 8) MLP. I ran these classifiers two times:

First round of running classifiers:

The first time around, I used a training set of 300 maximum and a testing set of 269. In this case each classifier trained the data for a training set of 100, 200, and 300 whereas the testing set was the same. It was obvious that QDA seemed to best predict the testing set in this round. Figure 3 shows the results for the training set of 300. The larger the training set, the better the classifier was able to predict on testing data..

Figure 3: Classifiers used to predict the breast cancer for a training size = 300

Type of Classifiers	Training Time	Prediction Time	F1 score (training set)	F1 Score (testing set)
K-nearest neighbor	.0012	.0022	.9264	.9081
Decision Trees	.0033	.0002	1.000	.9239
SVC	.0051	.0029	1.000	.0000
Naive Bayes	.0006	.0003	.9058	.9341
Random Forest	.0381	.0064	1.000	.9101
AdaBoost	.1866	.0055	1.000	.9451

Type of Classifiers	Training Time	Prediction Time	F1 score (training set)	F1 Score (testing set)
QDA	.0013	.0005	.9528	.9688
MLP	.0084	.0005	.6648	.6267

As can be seen from the table QDA seemed to outperform all the other classifiers followed by AdaBoost and Naïve Bayes with an F1-score of .9688 in the testing set. AdaBoost seemed to take the longest time to train the data and the longest time to make predictions followed by SVC. QDA seems like the best possibility in this case in both the amount of time it takes to train and test the data as well as in prediction scores.

Second round of running classifiers:

The second time, I increased the number of the training set to 400 to see how the results would change in this case. I assumed that a larger training set should result in better predictions even if the testing set would get smaller. Here are the results shown in Figure 4 below.

Figure 4: Classifiers used to predict the breast cancer for a training size = 400

Type of Classifiers	Training Time	Prediction Time	F1 score (training set)	F1 Score (testing set)
K-nearest neighbor	.0006	.0029	.9290	.9038
Decision Trees	.0049	.0003	1.000	.9074
SVC	.0110	.0068	1.000	.0000

Type of Classifiers	Training Time	Prediction Time	F1 score (training set)	F1 Score (testing set)
Naive Bayes	.0007	.0003	.9164	.9541
Random Forest	.0403	.0058	.9936	.9358
AdaBoost	.1767	.0057	1.000	.9541
QDA	.0013	.0005	.9585	.9381
MLP	.0102	.0005	.5663	.4843

In here, we see that although QDA is still performing very well with this dataset, AdaBoost and Naïve Bayes have outperformed QDA in terms of F1-score. AdaBoost still takes the longest time to train and predict the data but is able to result in a higher accuracy of prediction.

IV. Results

Model Evaluation and Validation

In order to decide which of these classifiers will result in the highest possible accuracy level with new testing data, I looked at the Receiving Operating Characteristics (ROC) for all of them. ROC visually depicts how well a classifier is doing in predicting the true positives and the false positives. True positives are often referred to as sensitivity and true negatives are referred to as specificity. So while true positives are the rate of cases that are correctly specified as malignant, the true negatives will be the cases that are correctly specified as benign in this case. Figure 5 illustrates the ROC curve as well as the percentage of the area under the curve for QDA, whereas Figure 6 illustrates the ROC for AdaBoost, and Figure 7 the ROC for Naïve Bayes.

Figure 5: ROC for QDA

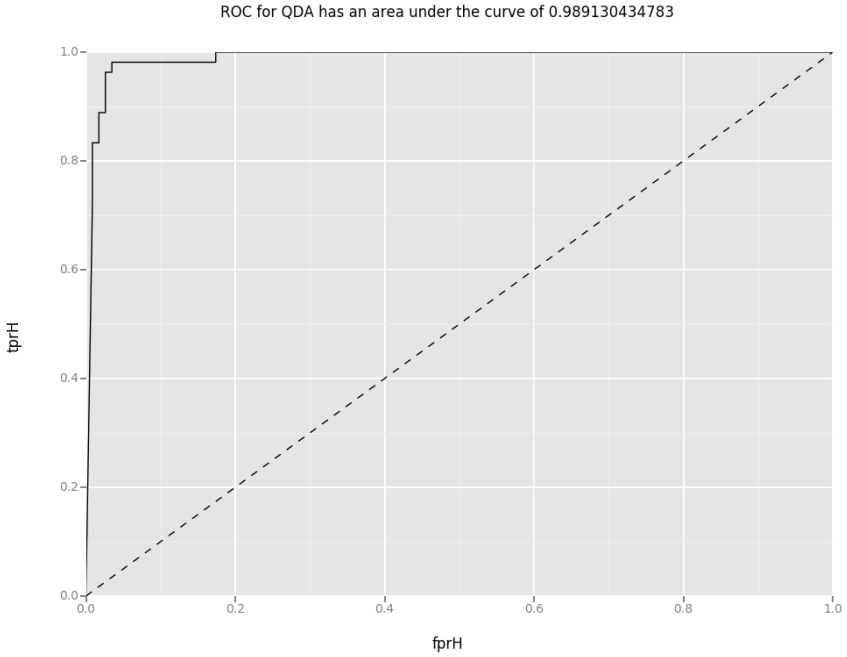


Figure 6: ROC for AdaBoost

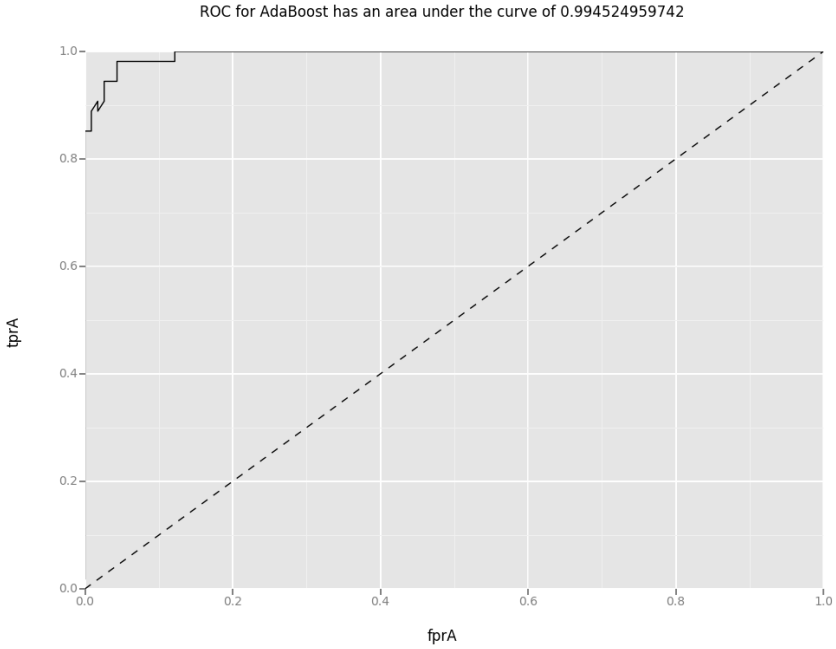
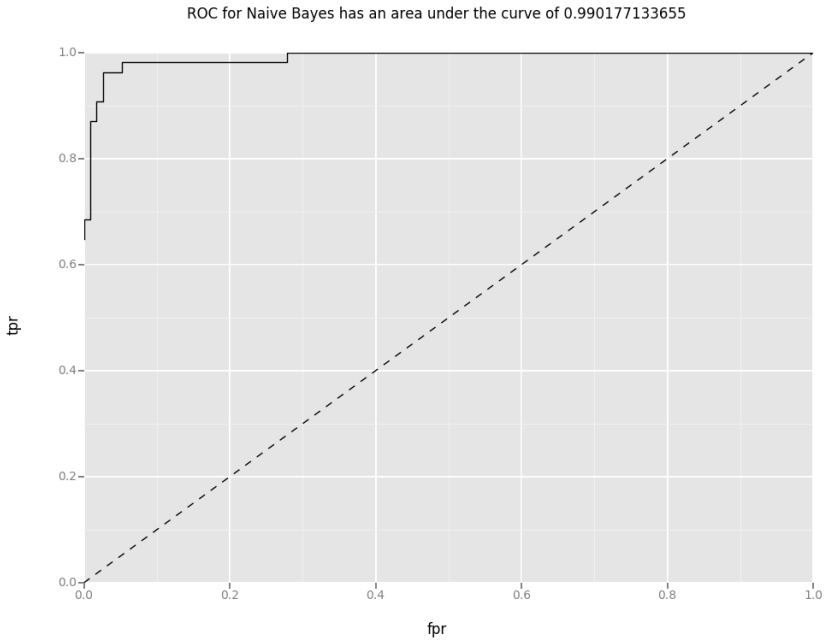


Figure 7: ROC for Naïve Bayes



As it is shown from these three illustrations, it is quite clear that the three these classifiers do an excellent job in predicting correctly the true positives and the true negatives. A larger training set is often crucial to be able to make good predictions. In this case however, even a small number of training set would result in excellent predictions from these three classifiers. Figure 8 delineates the results of these three classifiers with only a set of 50 training set and a larger set of 519 testing set.

Figure 8: Training and testing with a small training set

Type of Classifiers	Training Time	Prediction Time	F1 score (training set)	F1 Score (testing set)
Naïve Bayes	.0012	.0003	.9302	.9162
AdaBoost	.1409	.0026	1.000	.9171
QDA	.0011	.0004	.9000	.9076

Clearly, a smaller training set will result in worse predictions as the first ones, but these F1 scores are still pretty decent compared to the benchmark of .80 I had at the beginning of this project. ROC curve also show a good percentage of the areas under the curve for these cases. Figures 9, 10, and 11 illustrate each of these areas.

Figure 9: ROC for QDA

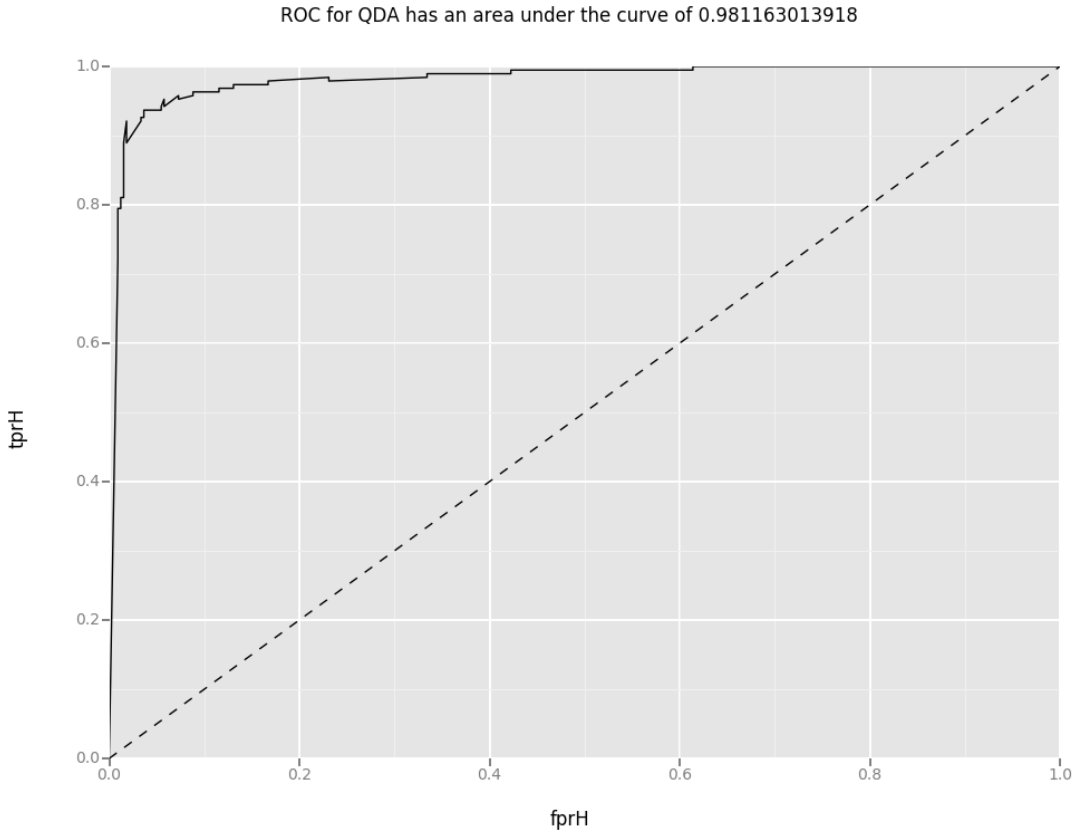


Figure 10: ROC for AdaBoost

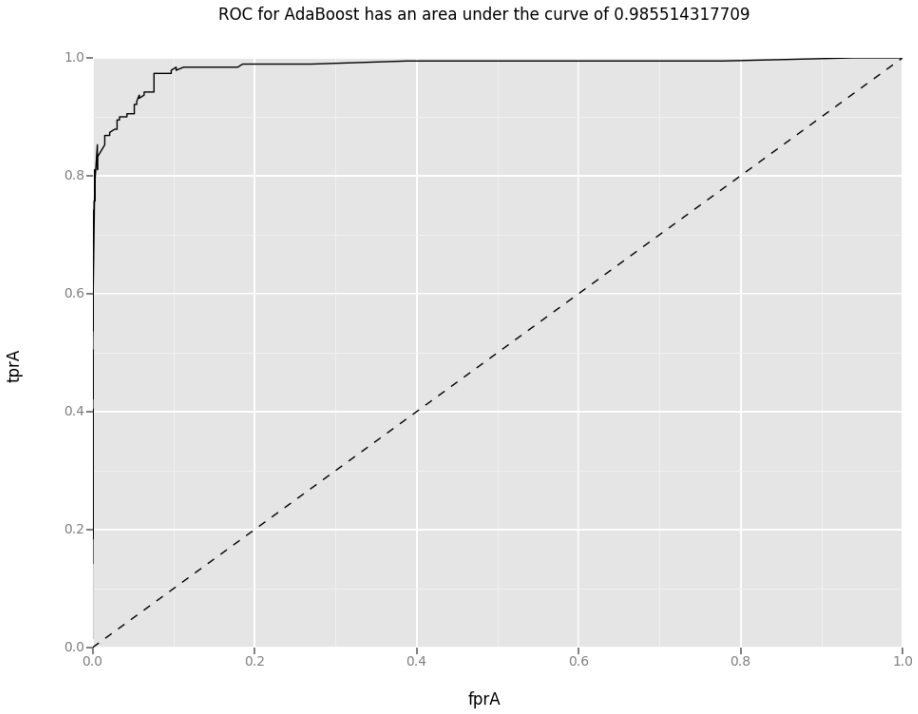
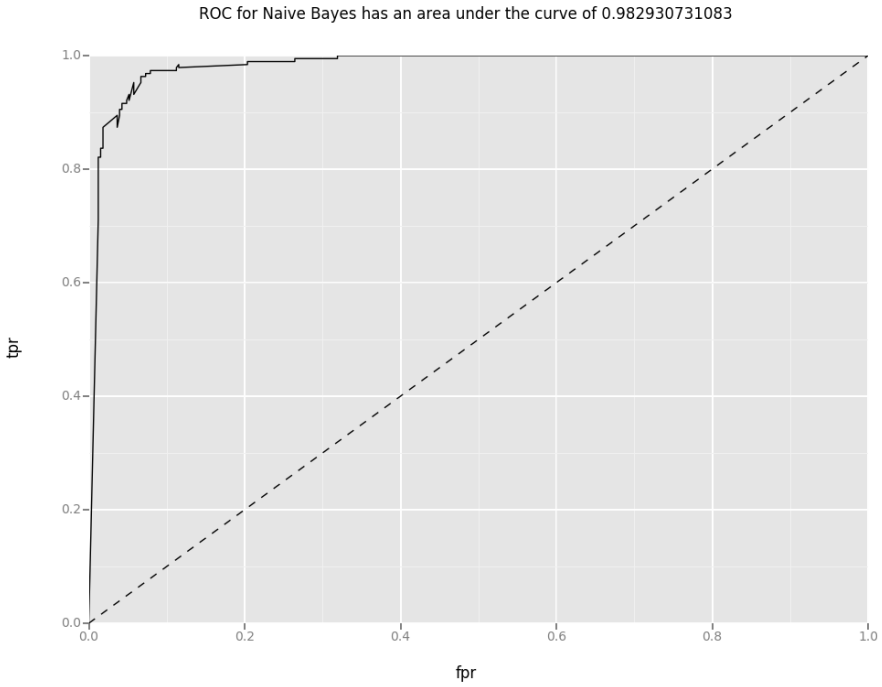


Figure 11: ROC for Naïve Bayes



V. Conclusion

When three different classifiers are able to perform as good as these three classifiers in this particular dataset, it is difficult to make a decision on which one is the best final

model. Researchers may consider the amount of time it takes each model to train and test the data. AdaBoost is consistently the slower in training and testing time, but somehow is also quite robust in its results even when the training set is significantly small. I also tried an example where the training set was 20 cases. While QDA could not perform well anymore in such small numbers of training cases, AdaBoost was still the leader in performance. Naïve Bayes is also a good model to go with because of its speed in training and testing and its ability to provide very good accuracy scores even for small sampling sizes. QDA worked best in relatively higher training sets.

REFERENCES:

DeSantis, C., Ma, J., Bryan, L. and Jemal, A. (2014), Breast cancer statistics, 2013. CA A Cancer Journal for Clinicians, 64: 52–62. doi:10.3322/caac.21203

U.S. Cancer Statistics Working Group. *United States Cancer Statistics: 1999–2013 Incidence and Mortality Web-based Report.* Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2016. Available at: http://www.cdc.gov/uscs.

Zhang, Wang, & Yang (2016). Computer aided diagnosis of abnormal breast in mammogram images by weighted type fractional Fourier transform, *Advances in Mechanical Engineering,* 8(2), 1-11.

Chung, Youk, Kim, Gweon, Kim, Ryu, & Son (2014). Role of diffusion-weighted MRI: predicting axillary lymph node metastases in breast cancer, *Acta Radiologica, 55(8),* 909 - 916

Tozaki & Fukuma (2011). Does power Doppler ultrasonography improve the BI-RADS category assessment and diagnostic accuracy of solid breast lesions? *Acta Radiologica,* 52 (7), 706 - 710