Predicting Breast Cancer Biopsy Results

Angel Garcia de la Garza, Soohyun Kim, Gaeun Kim

Introduction

The Breast Cancer Wisconsin (Diagnosis) Dataset consists of features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. The features describe characteristics of the cell nuclei from the image. The dataset can be found here: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29. The goal of this project is to use the available data to build a classification model to accurately predict if the tumor will be diagnosed as benign or malignant. The dataset consists of 569 observations and 32 variables. The response variable, Diagnosis, is binary (M: malignant, B: benign). The remaining 31 variables are the ID number of patients and the predictor variables. They are the mean, standard errors and “worst” or mean of the three largest values of the following features of the cell nuclei:

- Radius (mean of distances from center to points on the perimeter)
- Texture (standard deviation of gray-scale values)
- Perimeter
- Area
- Smoothness (local variation in radius lengths)
- Compactness (perimeter^2 / area - 1.0)
- Concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- Symmetry
- Fractal dimension ("coastline approximation" - 1)

There is no missing data so we did not exclude any subjects from the analysis furthermore, we included all available features in the main analysis.

Unsupervised analysis

In this section of the analysis we aim to understand the structure of the data and the relationship of the predictors with the outcome. We beginning by performing a two-sample t-test to explore which variables the two diagnosis groups differed most on. We can see in Figure 1 that the two groups had the biggest difference for concave.points_worst and perimeter_worst values. The vertical line on the figure indicates the threshold for significance (without correcting for multiple comparisons). We can see that most of the two groups have significantly different means for almost all variables. We now aim to understand the correlation structure among our variables to see if there’s any collinearity that we must consider while fitting some of these models.

From Figure 2, we observe that many predictor variables in our dataset are highly correlated with each other. The above correlation plot clearly shows that features related to the area, radius and perimeter are highly correlated with each other, with a correlation close to 1. We also see that features related to fractal dimension, compactness and concavity are also highly correlated. This indicates that a simple logistic regression would be a problem as it violates the

assumption of independence between variables. As such, we will proceed to use methods that are able to account for this collinearity such as penalized logistic regression.

We proceed to explore the bivariate relationship between the predictors and the outcome. We do this by using scatterplots to see if there's any clear separation between the two groups. This could indicate that a method such as SVM would be adequate. Due to the dimensionality of the data, we selected a subsample of uncorrelated variables that we found to be informative.

In Figure 3, we have selected "texture_mean", "area_mean", smoothnes_mean", concavity_mean", concave.points_mean" and "symmetry_mean". In these scatterplots, a benign diagnosis is shown in blue group while malignant diagnosis is shown in red. As we can see from each of the scatterplots, the benign tumors tend to have smaller means for all of the selected measures. Furthermore, here, we can see that the decision boundary for many of the variables are relatively close to linear.

When looking at the summary statistics of the covariates, mean and range varied largely across different variables. For methods like Vector Machine (SVM) and K-Nearest Neighbors (KNN) we will center and scale the variables before modelling the data. We know proceed with the supervised analysis.

Supervised analysis

The goal of our supervised analysis was to build the best classification model as measured by Cross Validated Area Under the Curve (CV-AUC). We aimed to explore the performance of a wide array of models with different degrees of flexibility. Due to the medical application of this analysis, we hope to be able to balance interpretability of the model with it's accuracy to find the one that give us high accuracy while easy interpretability that can be translated into a clinical setting. The methods that we explored include penalized logistic regression (using Lasso, Ridge and Elastic Net), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Bagging, Random Forest, Boosting, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN).

We wanted to use the whole dataset to measure the predictive ability of our models, so we decided to use 10-fold cross validation instead of going for a train vs. test set approach. For each of the above methods, we use the train() function in the caret library to search across a large grid of parameters and find the optimal model through cross validation. The cv.glmnet() function has a better performance than train() when fitting glmnet objects so we used that function to test penalized regression instead. Figure 4 shows an example of the output from the train() function. This plots shows us the CV-AUC for the entire parameter space. In this case, we were testing the number of randomly selected predictor for the Random Forest / Bagging approach. We observe that the optimal number of predictor to randomly select at each split is 3.

For each of the above methods, we extracted the estimated best AUC along with the its estimated standard deviation. Figure 5 is a summary plot that shows of all these measures. It contrasts the estimated AUC for each of the methods above with its 95% confidence intervals (truncated at AUC = 1). The horizontal line indicates the method with the highest CV AUC. The best model is penalized logistic regression using elastic net.

Penalized logistic regression is appropriate in our classification problem here because a crucial assumption of logistic regression is that there is little to no multicollinearity among the independent variables. That is, the predictor variables should not be highly correlated with each other. Simple logistic regression would not be the most appropriate for our dataset because it has multiple variables that are highly correlated with each other. For example, radius, perimeter and area of breast mass is clearly highly correlated and violates this assumption of logistic regression. Additionally, penalized logistic regression is both the simplest model and it gives us the best AUC. To choose the best set of tuning parameters, alpha and lambda, we built the best penalized logistic regression model by fitting across a large range of alpha, from 0 to 1 with an increment of 0.05. The value of alpha and lambda that gives us the best possible AUC of 0.998 with standard error of 0.001 is 0.15 and 0.005.

The final model includes all variables but “compacness_worst”, “symmetry_mean” and “concavity_se”. We use the predict function to calculate both the predicted class and the predicted probability of being in that class. Table 1 reflects the confusion table for the final predictors. We observe that it does a great job at classifying our observations, since the misclassification rate is only 1.2302%. Furthermore the Kappa statistics, which is more robust than misclassification error since it accounts for agreement happening just by chance, is 0.9736 indicating an almost perfect agreement. The final AUC for the whole model (Figure 6) is 0.997. We refrain from interpreting any of the parameters in the context of OLS because of the penalties. However we conclude that while there’s correlation in the variables they all contribute in different ways in achieving this level of classification accuracy.

Conclusion

In this project we aimed to classify soft aspiration samples from breast cancer patients and accurately predict the diagnosis of the biopsy. To do this, we performed several methods (including penalized logistic regression (using Lasso, Ridge and Elastic Net), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Bagging, Random Forest, Boosting, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)). We measured the best method as the one with the highest cross validated AUC. From this, we were able to conclude that penalized logistic regression is both the best method. Penalized Logistic Regression is both able to give us the best accuracy and the simplest interpretation which is ideal given the medical application and the need for clinical interpretability. From our final model, we conclude that most of the variables in the dataset contribute to the accurate classification. We did not expect this, but given that this dataset has been preprocessed beforehand, it makes sense that the authors discarded non-informative variables.

Figure 1. Statistic from T-test comparing Benign and malignant diagnosis for each feature.

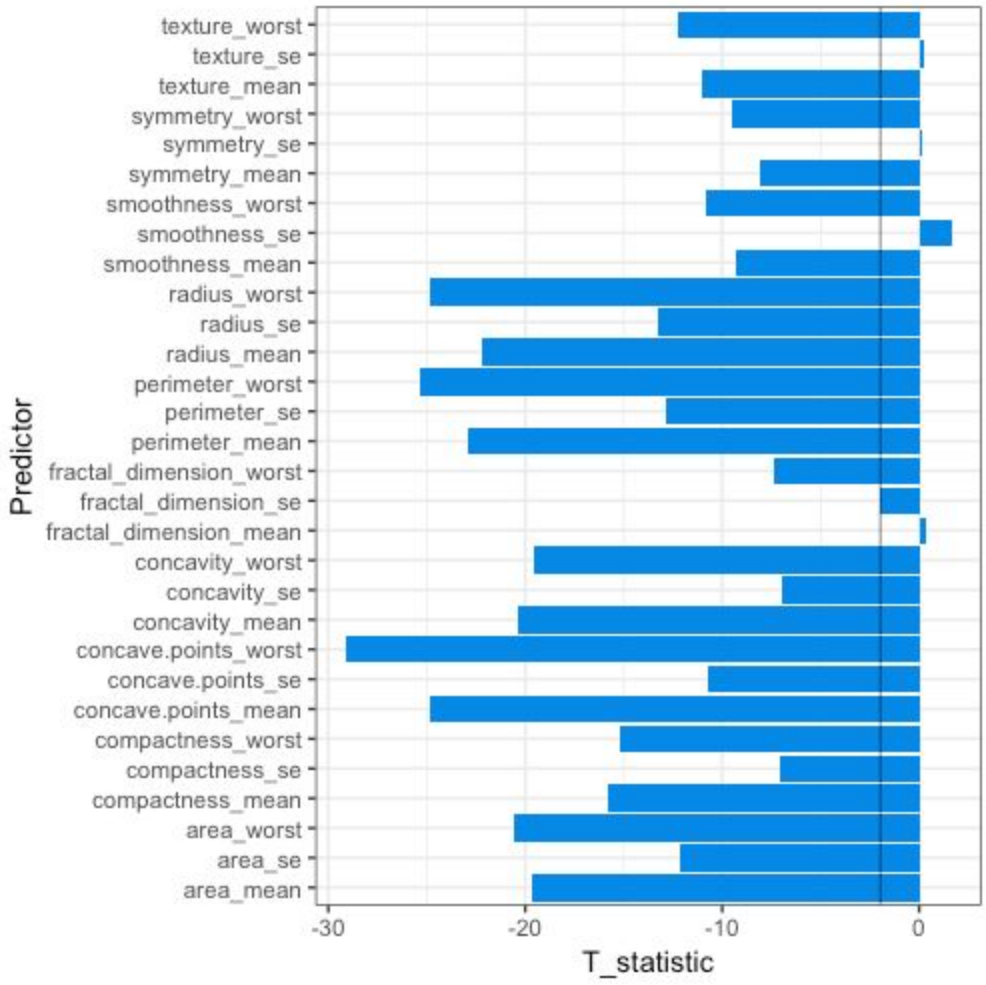

Figure 2. Correlation Plot of All Predictors

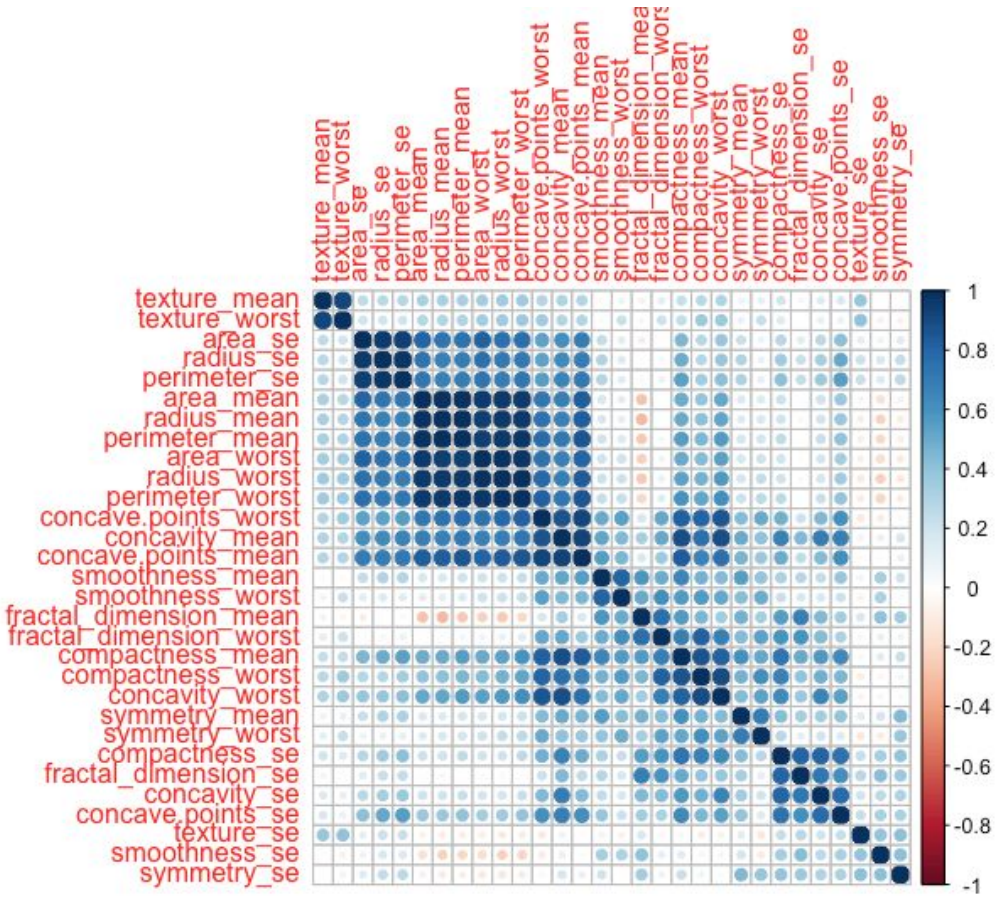

Figure 3. Correlation Plot of All Predictors. In this case Blue represents Benign and Red Malignant Diagnosis.

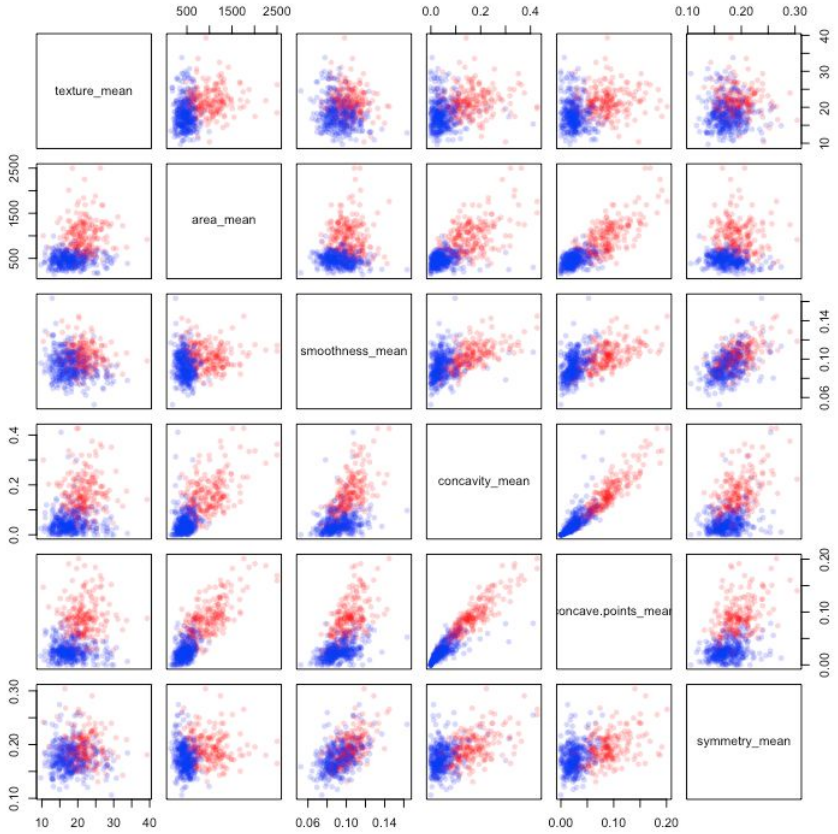

Figure 4. Output from the Train function. Estimated Cross Validated AUC for each number of randomly selected predictors for random forest.

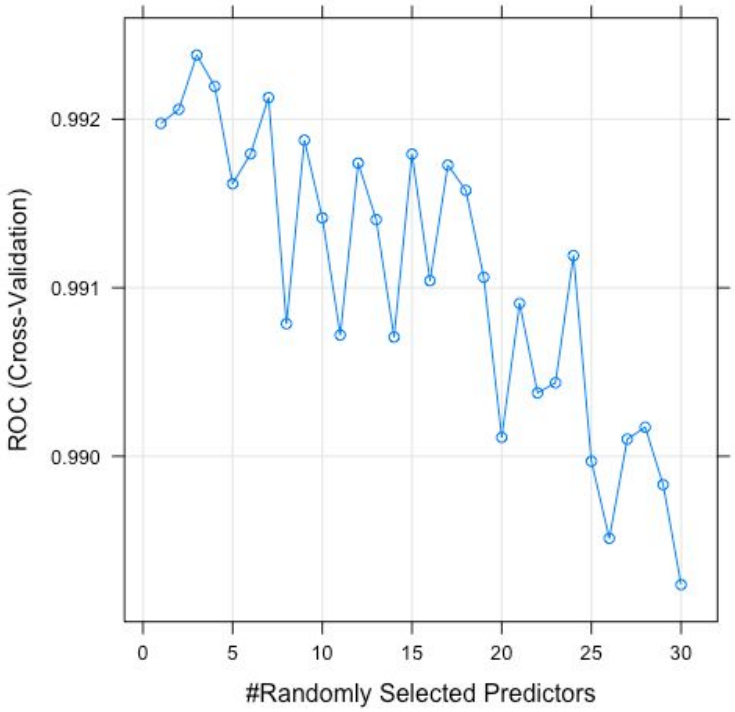

Figure 5. Cross Validated AUC for All Methods.

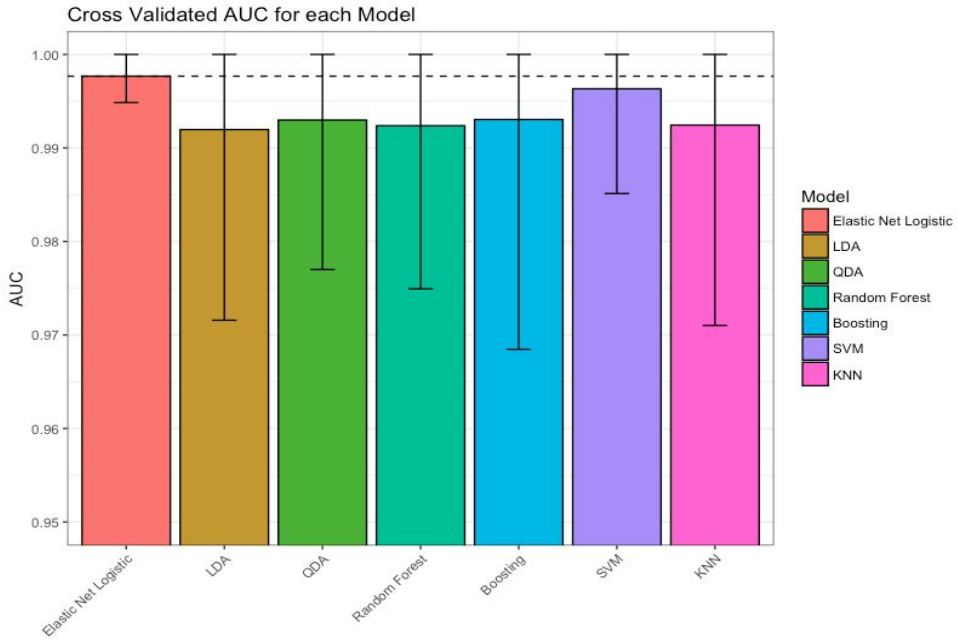

Table 1. Confusion Table for Penalized Logistic Regression Model

	Observed Benign	**Observed Malignant**
Predicted Benign	355	5
Predicted Malignant	2	207

Figure 6. ROC Curve for Penalized Logistic Regression