Midterm Project - Breast Cancer

Mitch Maegaard & Luke Spellman

11/16/2018

Introduction

Background Information

Breast cancer forms in the cells of the breasts; according to Mayoclinic, behind skin cancer, breast cancer is the second most common cancer diagnosed in women in the United States. The substantial increase in breast cancer awareness over the past several years has helped to finance extensive research, which in turn is leading to advancements in the diagnosis and treatment of breast cancer. Although these factors have contriubted to the decline seen in the number of deaths due to the disease, it still remains a heavily researched area and further advancements are looking to be made. Quantitative measurments from cell nuclei in breast mass serve as a potential resource to help detect breast cancer at earlier stages, thus enhancing the longevity and quality of life for several women and men suffering from the disease.

Fine needle aspirate (FNA) biopsy has become an increasingly popular method for breast cancer detection because it is fairly inexpensive and noninvasive, ASC.org. In the process, small samples of breast tissue mass are collected from patients that are at-risk for having breast cancer. This biopsy sample can then be further examined in a lab for the presence of cancerous cells.

In this study, we look to apply a statistical model that can accurately detect breast cancer based on the quantitative measurements from the FNA biopsy. The analysis will also help to pinpoint specific features that have the highest significance in distinguishing between cell types.

Data Collection

Digitized images of a FNA of a breast mass were taken from 469 patients being evaluated on their diagnosed level of breast cancer, including 170 malignant samples and 299 benign samples. These images were analyzed to identify specific quantitative characterisics of cell nuclei, and consisted of several measurments for each sample to further validate accuracy and precision.

Due to multiple measurements being taken for each characteristic, a data summary was able to be concisely represented by the sample mean, standard error (se), and the mean of the three "worst" measurements, which were defined by the three largest values from the measurements. This process was repeated for each of ten defining cell nuclei characteristics including radius, texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness ($\frac{perimeter^2}{area} - 1$), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, and fractal dimension ("coastline approximation" - 1), for a total of thirty unique measurments per sample. These metrics were recorded for all patients, and were also labeled with a unique identification code and breast cancer diagnosis.

Data Structure

The unique idenification variable provided to individual samples is represented as a numeric 6-digit code, but can be treated as a factor due to it's linking nature to a particular group, although it will be of little use throughout the analysis.

Diagnosis is a two-factor variable that provides information to the cancer diagnosis, "M" for malignant (36.2%), and "B" for benign (63.8%). This will be the response variable in the analysis, so the methodology will be constructed around a binary classification problem with malignant labeled as "1" and benign as "0".

The remaining 30 variables are real-valued features that are computed for each cell nucleus from the FNA biopsy, and correspond to the mean, standard error, and worst (mean of three largest) metrics from multiple measurements.

Experimental Design

The known cancer samples from the dataset were classified as malignant or benign (the diagnosis), and the end goal will be to utilize a logistic regression model to predict these categories of future cancerous samples based on similar image measurements. Three metrics for all ten features are included in the study because they are seemingly representative of FNA biopsy results. We will also be working to determing if a subset of these variates are most significant in the prediction.

In an effort to build both a statistically accurate and low complexity model (to ease interpretability and implementation), a primary step is evaluating feature distribution and applying necessary transformations to minimize outliers. Graphical analysis can then be utilized to determine relationships among features, highlighting which, if any, are the strongest indicators of malignant cells. Complexity can be reduced further by dimensionality reduction through determining uniqueness among characteristics, as some variables could carry redundancy given that the data were collected from a single image. This analysis was conducted mainly through graphing and other methods of visual examination of the characteristics split between malignant and benign cells.

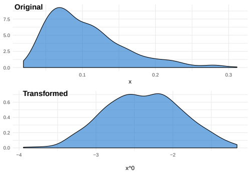
Exploring and transforming the data

Feature Transformation

A primary step toward creating a final model is feature transformation, which was done through both numerical and graphical analysis in evaluating distribution plots. Proper transformation selection was expedited through Tukey's Ladder of Powers test in the `rcompanion` package, which suggests a "best fit" lambda value that will attempt to correct a variable towards a normal distribution. However, since the value returned from the baseline function can be any real-valued numeral, we slightly modified the algorithm by selecting cutoff points for various lambda values; for example, in order to reduce right-skewed data, we took lambda values in the range of 0.0 to 0.3, and applied a logarithmic transformation (the original formula suggests the transformation only at a lambda value exactly equal to 0.0). We utilized this method for other common transformations (i.e. square root, cubics, etc.) to ensure our methods were easily replicable and easier to understand than simply using the suggested lambda values.

In an effort to efficiently and cleanly visualize the applied trasnformations for all variables, a simple graphing method was also added to the function. For an input feature, the algorithm will output two separate distribution plots, one corresponding to the original variable and the other as the selected transformed variable. One example where we observed a particularly effective transformation is shown in figure 1.

Figure 1.



As expected, the logarithmic transformation seen in figure 1 significantly helped in shifting the feature to a more normal distribution. Even though Tukey's lambda values were slightly modified, we can see that the mapped result is still effective, while also preserving interpretability. This methodology was applied across all 30 features to reduce the significant influence of skewness (extreme values) on the resulting model. Mean feature measurements with their respective suggested and chosen tranformations are provided in table 1.

Table 1.

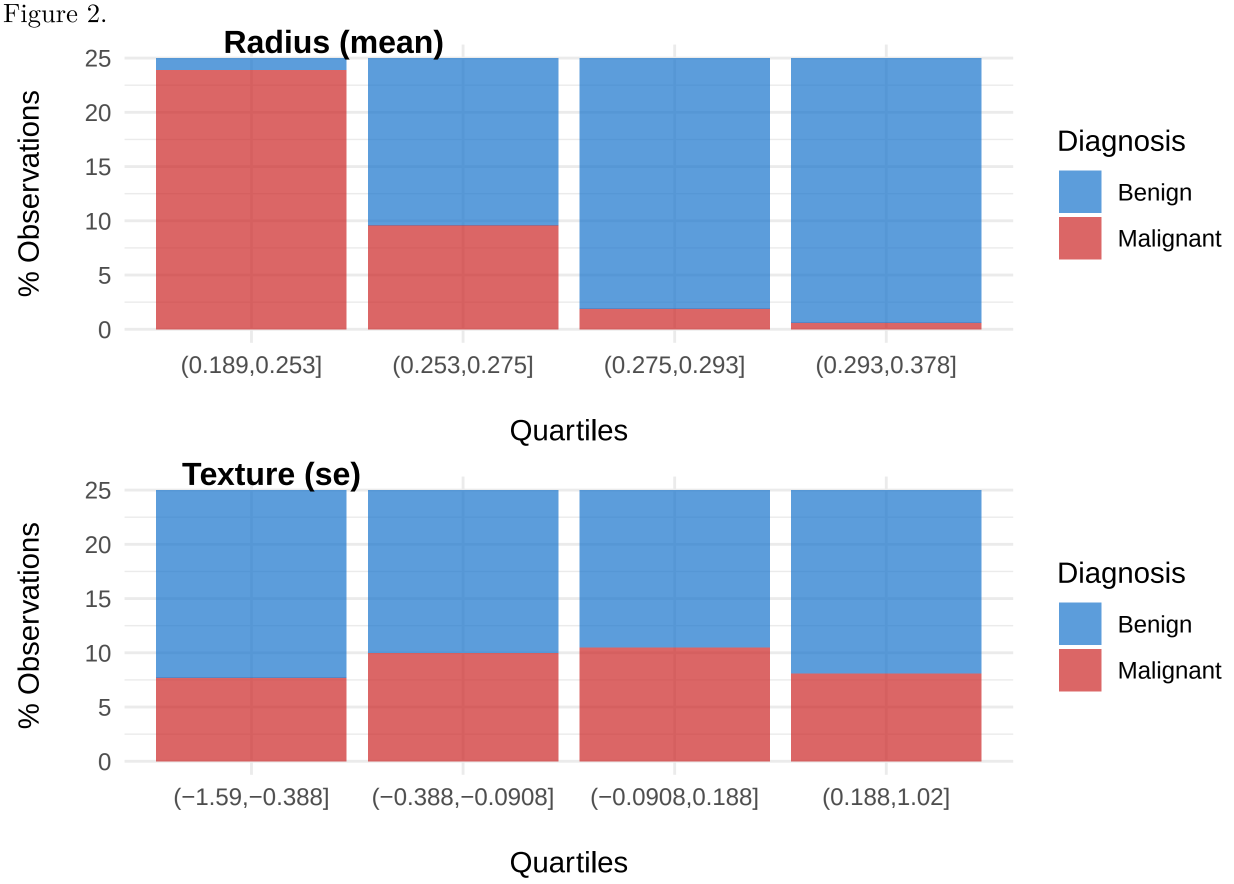
Feature Name	Tukey's λ	Selected Transformation
Radius	$x^{-0.575}$	$1/\sqrt{x}$
Texture	$x^{-0.025}$	$1/\log(x)$
Perimeter	$x^{-0.55}$	$1/\sqrt{x}$
Area	$x^{-0.25}$	$1/\log(x)$
Smoothness	$x^{0.075}$	$\log(x)$
Compactness	x^{0}	$log(x)$
Concavity	$x^{0.4}$	$\sqrt{x}$
Concave Points	$x^{0.425}$	$\sqrt{x}$
Symmetry	$x^{-0.35}$	$x^{-1/3}$
Fractal Dimension	$x^{-2.6}$	x^{-3}

Feature Selection

Another major step in model selection included feature selection, which was performed via exploring relationships between predictors and malignant or benign status with graphical contingency tables and side-by-side violin plots, as well as an analysis of correlated predictors and their respective density plots.

Contingency tables are a quick and effective method of summarizing results of counts at different levels of factor combinations. For this analysis, in order to represent values on a continuous scale in a contingency table, the predictors can be split into their respective quartiles, then compared against diagnosis. Figure 2 provides a graphical representation of one of these contingency table relationships.

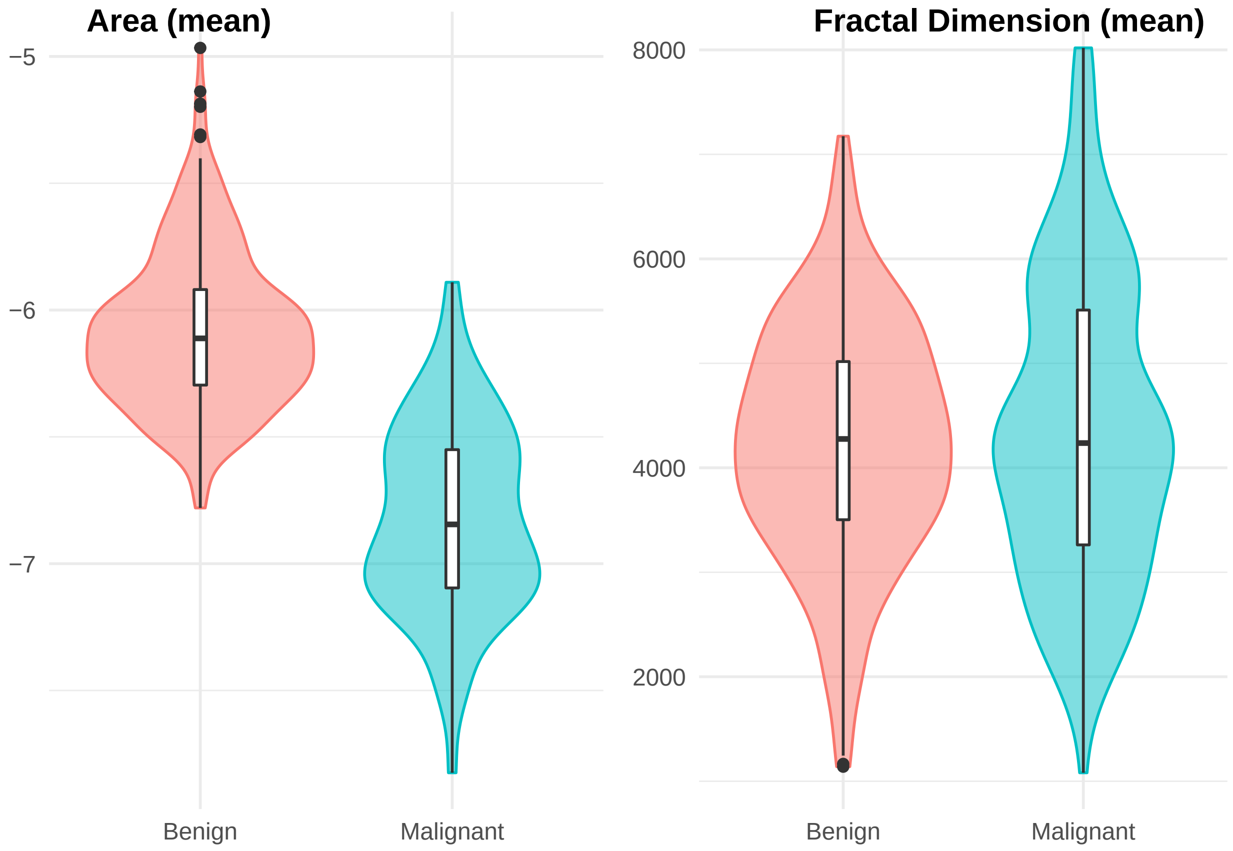
Figure 2.



Comparing the two example features in figure 2, we can see that a majority of malignant observations fall in the lower quartile of the transformed radius mean, while a majority of benign observations are in the upper quartile. Although it is more important to observe differences in the transformed variables in this analysis, it's important to keep in mind that these features have already been transformed, and reversing the transformation ($\frac{1}{\sqrt{x}}$ in this case) would give a majority of malignant observations in the upper quartile of the mean radius. This gives good indication that radius could be a significant predictor in building a final model. On the other hand, the transformed "texture se" feature shows incredibly little variation over the quartiles; with such marginal distinction, it is unlikely that this will be an effective predictor in predicting cancer status.

Again, with the response variables taking on continuous, numeric values, violin plots supply a powerful method of examining the data; figure 3 displays a side-by-side violin plot that indicate the median, lower and upper quartiles, outliers (if any exist), and kernel density estimations (similar to probability density) for multiple group-wise comparisons.

Figure 3.



The graph on the left in figure 3 shows area mean, one of the transformed variables where we saw an evident difference in sample distribution between diagnosis categories. Here, malignant observations tended to have lower radius values (corresponding to higher values in the pre-transformed data), along with more outliers in the upper quartile. This gave us an indication that the mean radius metric could provide useful insight for our final model. On the other hand, the graph on the right was an example where we didn't find much difference between sample distributions; the medians appear to be approximately the same, along with the distributions. Thus, we would likely opt to not include these types of variables in our final model.

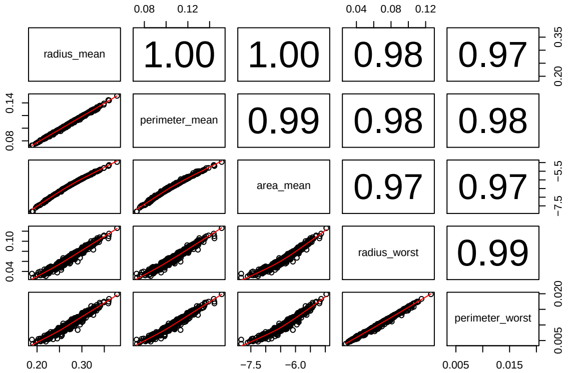
Dimensionality Reduction

Dimensionality reduction was performed to remove highly correlated variables, as these "duplicate" features don't provide any further insight to classification, while contributing increased model complexity. This process involved diagnosing values from correlation matrices and their respective plots, then further examining distribution plots to evaluate which feature would be most representative of the subset, thus would be the one kept for future modeling.

Considering the need to diagnose collinearity across 30 features, excluding sample ID because of it's uniqueness and diagnosis because it is the desired response variable, a simple yet efficient method is of high importance. The process begins by creating a matrix of correlation values for all 30 features, resulting in a set of 900 values to pick through, with 1's on the diagonal (each feature is obviously a 1-1 correlation with itself). To make the data easier to work with, the matrix was then transformed into a 900x3 array, where the first two columns represented all combinations of features, and the third represented the correlation between the two variables. In order to minimize search time on variables, and using a "high correlation" cutoff of 0.8, the data was filtered to include only unique permutations of features that had correlations over the cutoff, excluding those related to themselves. These variables were determined to be "at risk" for high collinearity, and were further diagnosed via graphical analysis.

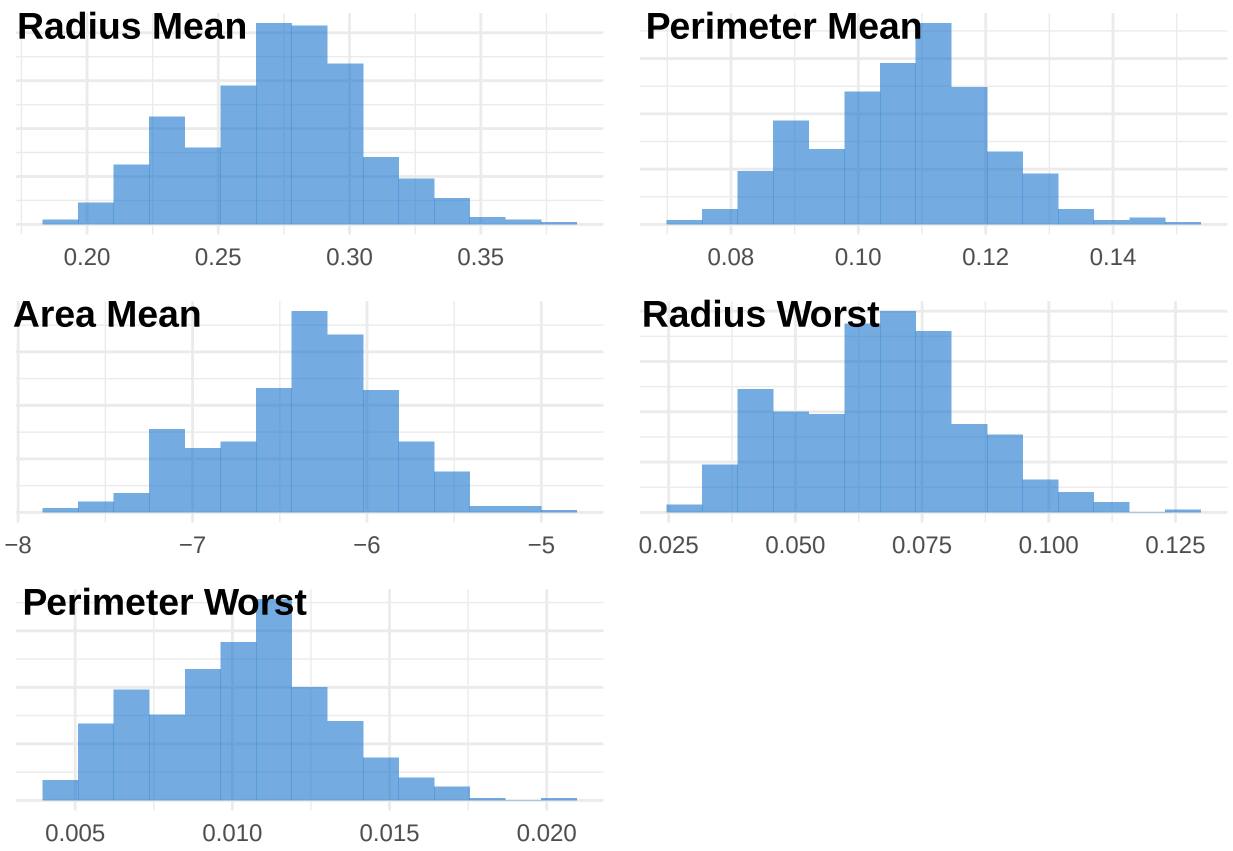
Pairwise comparison plots were utilized in diagnosing collinearity between the at risk features; variable name is included on the diagonal, the lower panels include a scatter-plot of the variables against each other with an attempted linear fit line, and the upper panels give the absolute value of the numeric correlation scaled 0 to 1, where a value further from 0 indicates stronger correlation.

Figure 4.



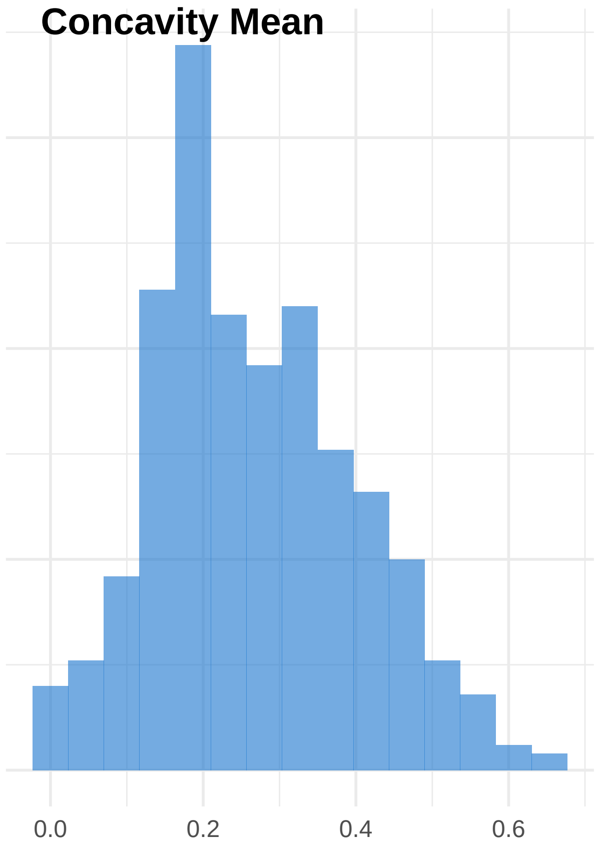
From figure 4, it is shown that the means of radius, perimeter, and area have correlations extremely close to 1.0. Further, these three metrics also boast correlations above 0.97 with the "worst" measurements from radius and perimeter, which also have a 0.99 correlation with each other. We plan on removing all but one of these five features from our final model, as they will only add complexity to the design. From this subset, we selected the mean radius as our representative "distance" metric based on decisions from the histograms displayed in figure 5 for the following reasons; it's lesser skew and potentially bi-modal distribution are appealing for observing differences in cancer status, and it's range of values are most appealing to work with carrying forward.

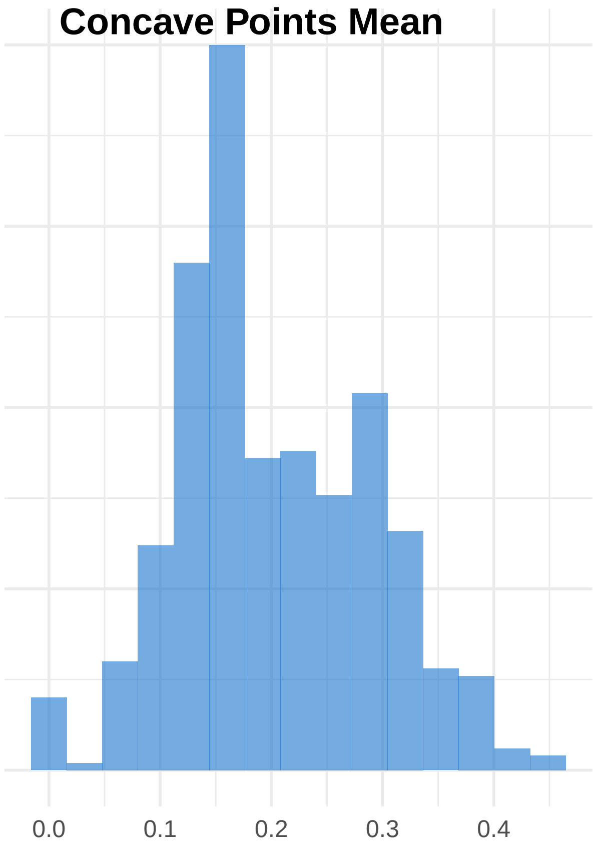
Figure 5.



The mean number of concave points also holds high correlation with several other variables, including the mean severity of concavity; therefore, we opt to remove this feature as well. As with the distance metrics, histograms in figure 6 provided a visual diagnostic and validated our choice of the severity of concavity, because it is more approximately normally distributed. Finally, compactness "worst" is removed because of it's high correlation (0.90) with it's respective "mean" metric.

Figure 6.


Concavity Mean


Concave Points Mean

The four remaining features, along with the others from our dataset that did not pose any durastic correlation issues up-front, will be utilized in building an initial model.

The Logistic Model

The methods carried out through data exploration carry over into building an inital model. Using the selected transformations, we remove all highly correlated variables. Additionally, diagnosis is converted to a "1" if malignant and "0" if benign. We choose the malignant value to be the "positive" response because those are the "at-risk" samples we would like to identify.

Upon implementation, our model will be used to classify cancerous cells on a held-out test set of samples we have not seen before. To ensure confidence prior to implementation, the data we have available to us should be partitioned such that we can both train a model and check the fit on data the model hasn't used. To do this, we split our transformed data into 75% training, and 25% cross-validation. The process is done through random sampling based on a set seed (for replicable results), and results in a training set size of 351 and cross validated set size of 118.

We begin the modeling process by fitting a fully saturated model, which utilizes all of the predictors left after removing higly correlated features to predict cancer status. To assess model fit, we make a naive assumption that we have an equal amount of malignant and benign samples in our training and test sets. This corresponds to a "threshold" value of 0.5, meaning that if our model outputs a probability over 0.5, we predict "malignant", and "benign" if it's under. A confusion matrix calculates a cross-tabulation of observed and predicted classes with associated statistics. Finding an appropriate balance between sensitivity (the percentage of actually malignant samples that are predicted as malignant), specificity (the percentage of actually benign samples that are predicted as benign), and overall accuracy (the percentage of correctly predicted outcomes) will be of high importance for implementing our classification model. Sensitiviy should be maximized to ensure all patients with malignant cells are treated as such, while specificity should be maximized to reduce the more extensive tests for malignant cells, which could be extremely invasive or harmful to the patient.

A saturated model fit on 20 parameters misclassifies only one of the 118 samples, attaining an overall accuracy of 99.15%, with sensitivity of 97.22% and specificity of 100%. Based on these predictive metrics alone, we find this model to be effective in classifying cancerous samples. However, examining a summary of the model fit, we find that all p-values are extremely close to 1; an initial analysis may deem that none of the predictors are useful in classifying cancerous cells, but we instead point to likely multicollinearity and potential overfitting due to the large standard errors in relation to their respective coefficient estimates, as well as an excessively large number of Fisher scoring iterations (25). In addition to overfitting, model complexity also decreases interpretability; this is extremely beneficial in our scenario because we would ideally like to pinpoint cancer cell characteristics that doctors could evaluate in a more streamlined approach. Thus, although we find the predictive power to be extremely high, more extensive work needs to be done on simplifying model complexity to eliminate overfitting and ease interpretability.

Model Selection

Model selection was performed in an attempt to reduce model complexity and eliminate multicollinearity that we observed in our first attempt at a fully saturated model. In doing such, we attempted three distinct procedures involving combinations forward/backward stepwise regression and LASSO (least absolute shrinkage and selection operator) regression.

In the first procedure, we initially conducted an automated forward/backward feature selection algorithm, where the goal was to maximize the log likelihood with an additional penalty for the number of parameters included in the model, which in turn relates to minimizing AIC. After attaining a subset of 11 parameters, we noted that a few predictors still had correlations over 0.8, which we deemed unacceptable for our model because of the adverse effects on significance levels. To combat, we implemented LASSO regression, which is a method that performs both variable selection and regularization in order to enhance prediction accuracy and interpretability of the statistical model it produces. This further subsetted our model to just 8 parameters, although we found one to be statistically insignificant. An ANOVA test on models fit with and without this parameter (perimeter standard error) yielded a p-value of 0.2872, indicating that we should favor the simpler model; this 7-feature model was our “final” model for the first procedure, and yielded accuracy of 99.15%.

Our first procedure occassionally produced weaker models (based on individual features’ p-value significance) due to flaws in stepwise selection on a large number of parameters, where it could potentially come to various conclusions in minimizing AIC along different paths and handling collinearity. This prompted us to rearrange our process to perform LASSO first, followed by stepwise selection and further by removing any remaining highly correlated variables and insignificant predictors. LASSO adds a penalty to the sum of square errors that is proportional to the terms coefficient; forcing the sum of coefficients to be less than an optimal threshold in turn forces some of the parameter estimates to be extremely small, in which case we remove them from the model. Following LASSO, any remaining features with correlation greater than 0.79 were removed from the model due to their influence on the significance of predictors. This resulted in a model fit with 5 statistically significant predictors at the $\alpha = 0.05$ level, but a decrease in accuracy to 97.46%.

We noted the inconsistency of both stepwise selection and LASSO regression in the presence of highly correlated variables. From this, we simplified our selection technique to choosing only significant variables from a model fit following LASSO on our saturated model. This commonly resulted in the same final parameters, although we would occassionally get alternate results due to multicollinearity. To ensure consistency in the method, we repeated LASSO several times and noted all significant predictors from each output; we then took parameters that were present in at least 50% of the iterations and used those in building a final model. By doing such, we consistently observe the same three parameters including “worst” measurments from area, smoothness, and texture, all of which are statistically significant at the $\alpha < 0.001$ level; we also find that these three predictors yield an accuracy of 98.31% with a classification threshold of 0.5, making only 2 misclassifications on 118 observations. Comparing this model to the fully saturated model, we find that we retain nearly the same predictive power while significantly reducing model complexity, resulting in both interpretability and generalization when introduced to new data. We also favor this model selection process over the previous two because of it's reliability in convergence to the same few parameters, improved (decreased) model complexity, and high predictive accuracy. Complete results from modeling procedures can be found in table 2 below.

Table 2.

Process	Total Parameters	AIC	Sensitivity	Specificity	Accuracy
Fully Saturated	20	42.0	.9722	1.000	.9915
Stepwise → LASSO	7	51.04	.9722	1.000	.9915
LASSO → Stepwise	5	61.19	.9444	.9878	.9746
Repeated LASSO	3	69.07	.9722	.9878	.9831

Optimizing the Threshold for Accuracy

Because we are using logistic regression for classification, we need to specify a value at which predicted probabilities will be rounded up to 1 (malignant), or down to 0 (benign). Previously, we defined this "threshold" to be 0.5 (an even split between the two classifications). By varying the threshold from 0.5, classification of cancer statuses can improve or decline, and we search to find the optimal value for this threshold to maximize predictive power.

Figure 7.

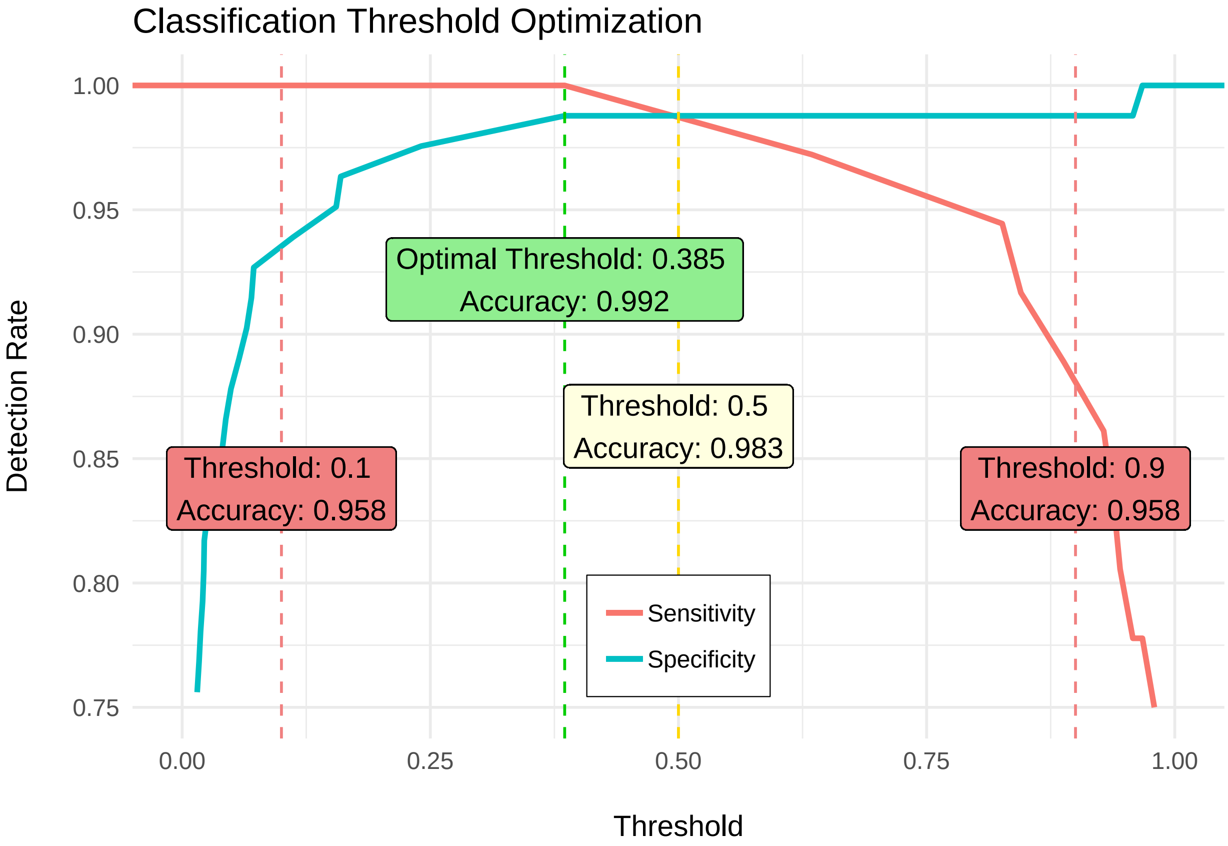


Figure 7 effectively communicates results about the performance of our final model as we modify the classification threshold from 0.5. Specificity steadily increases as threshold increases, while sensitivity is downward-sloping from a maximum (100%) at approximately 0.385. We can determine overall accuracy at varying threshold values utilizing the relationship between sensitivities and specificities; in doing so, we observe that at a threshold of 0.1, accuracy is approximately 95.8%, while moving to a higher value of 0.9

yields the same accuracy of 95.8%. Each of these accuracies are lower than the results obtained in the previous section (99.15% at 0.5 threshold). We find that the optimal classification threshold can be specified by the point at which the sum of sensitivities and specificities from the receiver operating characteristic (ROC) curves produce a maximal value (0.385). Although in this case the accuracy remains the same at 99.2%, we can be more certain that this is the optimal threshold for our classification model due to the results carried out in this section.

Results Summary

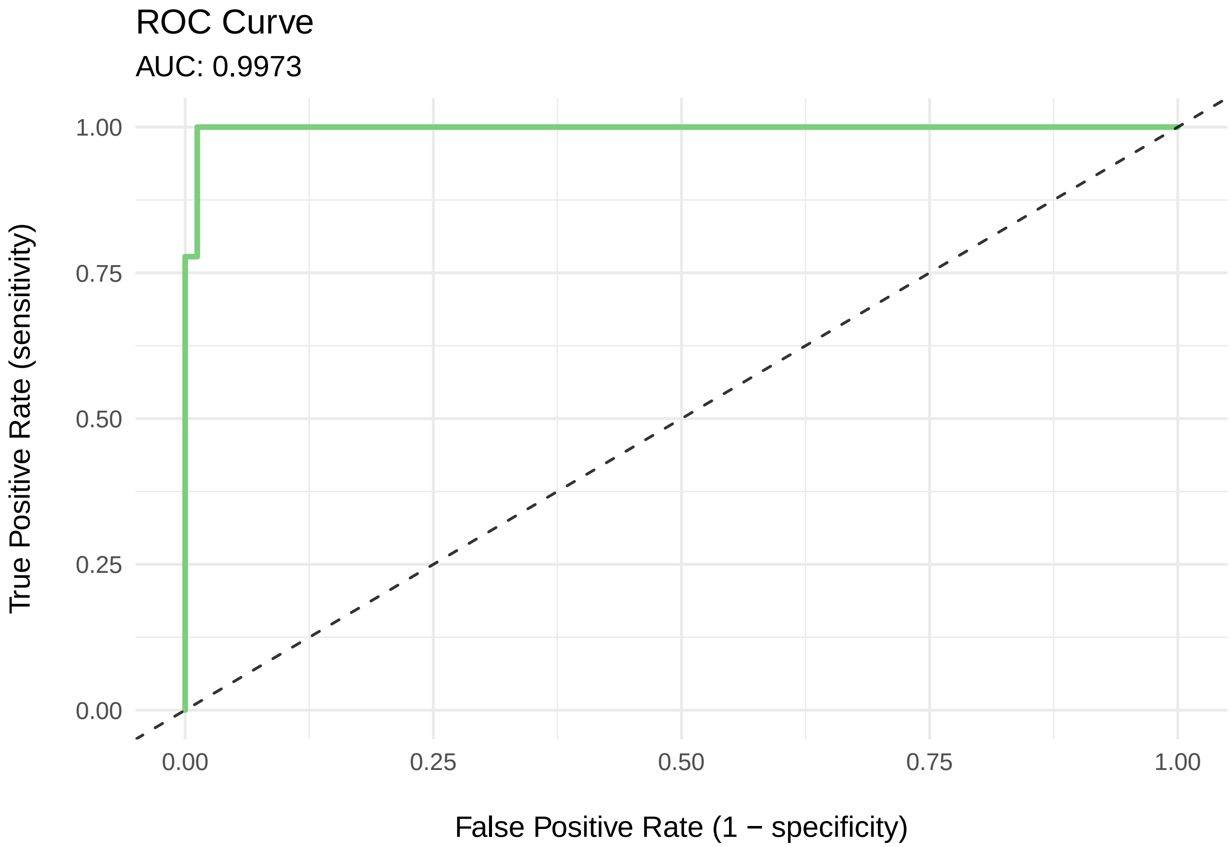
In diagnosing cancer status, our final classification model consists of just three parameters including "area_worst", "smoothness_worst", and "texture_worst". Each of these three parameters were deemed as highly statistically significant in predicting diagnosis given 351 training samples, as is shown in figure 8.

Figure 8.

```
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         34.800      8.818   3.946        0
## area_worst        -826.224    146.659  -5.634        0
## smoothness_worst    14.304      2.914   4.909        0
## texture_worst        7.198      1.815   3.967        0
```

At an optimal classification threshold of 0.385, we attained sensitivity of 100%, specificity of 98.78%, and overall accuracy on 118 cross-validation samples of 99.15%. These metrics correspond to one misclassification, where we predicted a sample to be malignant when it was actually benign. The tradeoff between sensitivity and specificity accuracy was heavily debated in choosing a final model for ethical reasons; our single misclassification illustrates our concern over this issue in that we wanted to "error on the side of caution" in sacrificing specificity for sensitivity, to ensure that all patients predicted to have cancer would rather undergo additional, potentially invasive, testing as opposed to allowing malignant cells to manifest in the tissue and likely become significantly more dangerous. Figure 9 displays a ROC curve with an area under the curve (AUC) of 0.9973, indicating the high classification power of our model.

Figure 9.


ROC Curve
AUC: 0.9973
True Positive Rate (sensitivity)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
False Positive Rate (1 − specificity)