Nikhitha Lingutla, Xiao Tan, and Michael Winkler

Dr. Daniel Carr

STAT 663: Statistical Graphics and Data Exploration I

December 18th, 2018

Final Project: Classification of Breast Cancer Using Three Methods

Section I: Introduction –

The human body is in a constant cycle of cells reproducing, growing, and dying. Through a process called cellular fission, where the cells reproduce by splitting themselves, the body grows, repairs itself, and other processes that allow for normal human life. Sometimes one type of tissue or groups of a couple of types of tissue will replicate much faster than necessary or cells will not die as quickly as they are supposed to, causing the growth of a tumor. Tumors can be either malignant or benign, with malignant being substantially more serious than benign tumors.

An estimated one in eight women in the United States will develop breast cancer over the course of their life. In 2017, 316,120 people were diagnosed with breast cancer. Additionally, more than 40,000 people died from the disease. Non-Hispanic whites (NHW) and Non-Hispanic blacks NHB) have the highest rates of developing breast cancer, while Asian and Pacific Islanders have the lowest rates. The two factors that put a person most at risk are age and gender. While it is possible for men to develop breast cancer, it has a much lower prevalence among males (about one percent of cases in the United States). The median age for a woman receiving her first breast cancer diagnosis is 64. An immediate family history of breast cancer approximately doubles the risk of developing cancer. Having an above average BMI, consumption of alcohol, cigarettes, certain hormone-based birth control, and above average height are also considered risk factors. Additionally, being a cancer survivor makes a patient more likely to develop breast cancer a second time. Living a healthy lifestyle with plenty of exercise lowers the probability of developing cancer. Women who have children and who breastfeed for longer than one year

can substantially lower risks of developing cancer. It is critically important to get regular check-ups with blood testing as well as mammograms, especially if the patient fit into any of the higher risk categories.

Breast cancer is typically asymptomatic until the tumor has grown to an unsafe level. This makes it critically important for patients, especially women, to get checked regularly. If the cancer is found when it is localized, the five-year survival rate is over 99%; however if the cancer is not caught until it is "regional" (in the breast and lymph nodes) the five-year survival rate drops to approximately 85%. For patients who have found the cancer after it has spread beyond the breast and lymph nodes, the survival rate is a dismal 27%. Fortunately, these cases are decreasing and currently only make up about six percent of cases that are diagnosed.

Malignant tumors are what society think of as cancer. Breast cancer tumors tend to grow very slowly, sometimes even taking 10 years before the tumor can be felt. Breast cancer can be classified in two different ways: Non-invasive breast cancer, also known as ductal carcinoma in situ (DCIS), and Invasive breast cancer. DCIS occurs when there is a tumor in the milk duct which has not spread to any other type of tissue. According to the Susan G. Komen foundation, approximately 500,000 cases of DCIS will be diagnosed in 2019. When left untreated DCIS can turn into Invasive breast cancer. This happens because the cancer cells will travel through the bloodstream or the lymphatic system to other parts of the body. It is not known why it spreads when it does as it can spread when the tumor is small or large. After it has spread it becomes "Metastatic breast cancer" or Stage IV breast cancer. This means that it has spread beyond the breast and axillary lymph nodes. Typically, it will spread to the bones, lungs, liver or brain. Malignant tumors may be treated if it is detected early in the disease process. Early detection is the key to higher survival rates. Treatments for malignant tumors include surgically removing the tumor, using localized radiation therapy, hormone therapy, and chemotherapy. Treatment may vary due to the patient, how advanced the cancer is (Stage 0-IV), the shape of the proteins in the cancer, and where the cancer is located.

Although benign tumors are not cancer, they are still very serious. There are several kinds of benign tumors and they have different risks associated with them. Benign tumors can be irregularly shaped, but can also be symmetrical. Sclerosing Adenosis and radial scars look very similar to malignant tumors on a mammogram, however, they are not cancerous. Sclerosing Adenosis do not need any form of treatment and may go away by themselves, whereas radial scars need to be removed. Cysts and fibroadenomas are fluid filled and solid tumors, respectively, but do not add to the risk of developing breast cancer. They can be removed if the patient feels that it necessary, but do not be removed. Although some of the benign tumors are not as serious, hyperplasia is a form of benign tumor that is incredibly serious. Hyperplasia can be either usual or atypical, with atypical being much more severe. Both of types of hyperplasia lead to increased risk for breast cancer. Patients with atypical hyperplasia are encouraged to receive more frequent mammograms and breast screenings as well as take either tamoxifen or raloxifene, which are drugs that lower the risk of developing breast cancer.

A great deal of research has been done in the field of identifying risk factors and modeling those risk factors for healthy individuals developing breast cancer. While identifying cancer risks is still a developing field, we felt a more interesting and less worked on issue is assuming a patient has found a mass in their breast, are there factors that we look at to predict if the cancer is malignant or benign? Without utilizing a statistical model, this process involves a surgical biopsy which can be incredibly invasive and potentially slow. It should be made perfectly clear that no matter how good of a statistical model is being used, the results should still be checked with some form biopsy.

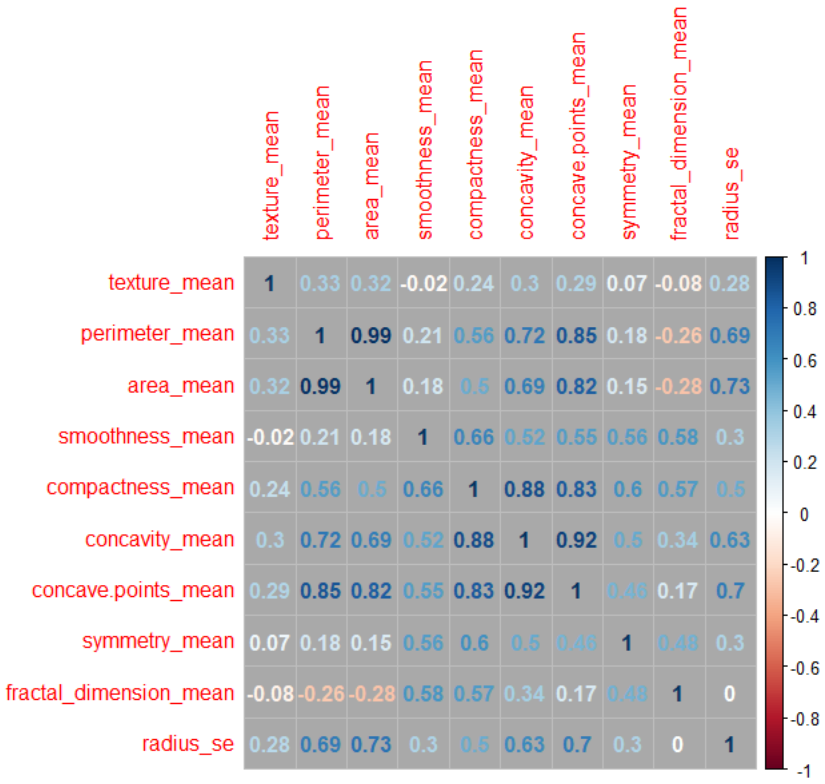
In the course of this paper we will introduce the data, use a couple of simple logistic regression models to classify, stepwise variable selection to fit an ideal model, Logistic Lasso Regression, Logistic Ridge Regression, a cross-validated pruned decision tree, and random forests to see if we can obtain a reasonable model to classify the data. At the end of the paper, we will compare the models using the test error rate, look into how we could extend this research, and finish with conclusions.

Section II: The Data –

This dataset comes from a 1992 survey of 800 women in Wisconsin with tumors in their breast(s). It is not clear why, but unfortunately, a large number of observations had been left out when the data was published. Using the 569 cases that were that were published we tried to use classification techniques to determine the probability that an individual case was malignant versus the probability that it was benign. To collect the data, a surgeon begins by taking a small drop of fluid from the tumor using a fine needle. Then the fluid is placed on a glass slide and stained. Then using a microscope with a camera, certain nuclear features are noted and recorded.

There are some important concerns to talk about before we review the variables we used. First, linear models, such as logistic regression, do have an assumption that the covariates are independent. Unfortunately, there is no way in this data to check this assumption and based on the correlation plot, located below this paragraph, it appears that this assumption is likely violated. While this collinearity does not appear to have hurt the classification it does make some of the statistical inference (such as p-values for the variables) questionable at best. Additionally, although the data is not "small" it is not as large as we would have liked. Methods like "Cross-Validation" do help combat this problem, but we could not use cross-validation on all of our models. Hence, our preference was to take a sample from our data and break it into a test sample and a training sample. We did not feel that we had adequate data to get a big enough test sample while also getting a meaningful prediction. As a result, we are using the testing error rate to compare models. This is not ideal, because these models are designed to minimize the test error rate. Unfortunately, without more data, this seems like it is a necessary evil.

Figure 1: Correlation Plot of the Mean Variables

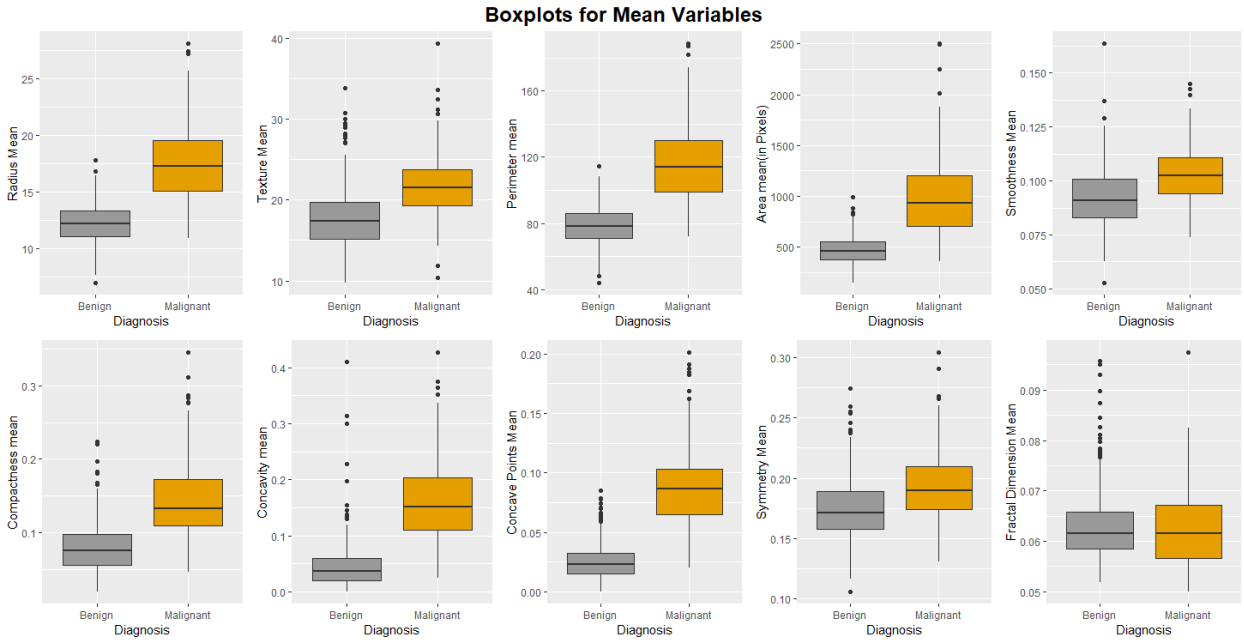


The important things to note are that some of the correlations are exceptionally high. For exam ple, the perimeter, and the mean appear to be approximately a linear combination of each other with a corr elation of .99. In addition, almost all of the correlations are positive. We chose to do a correlation plot with numbers on a grey background because we felt it showed the most information in a way that is easy to read. The default correlation plot is different colored dots a white background. Using the default setting makes it difficult to see correlations that are close to zero. The numbers also make the graph much easier to interpret as it takes the guessing out of figuring out the correlations.

The dependent variable is the diagnosis variable that has been coded so that if the tumor is malignant it is a one and if it is benign it is a zero. There are ten independent variables; however, each of the independent variables is represented in three ways – the worst values (the mean of the three most

extreme values), the mean value, and the Standard Error of the mean. The independent variables recorded are the radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, a nd the fractal dimension. The radius of the nucleus is measured by averaging the length of the radial line segments. The smoothness is calculated by measuring the difference between the length of a radial line and the mean length of the lines surrounding it. The perimeter is the total distance around the nucleus of the cell. The area is the number of pixels and multiplying that number by one half to represent the perimet er of the nucleus. The texture is found by measuring the variance in the contrast in grey scale of the image. The compactness is measured by $\frac{perimeter^2}{area}$. It is a dimensionless number that is minimized if the tumor is round and becomes much larger if there is irregularity in the boundary. Concavity is measured by drawing a line (called a "chord") between two points that are not adjacent. This helps find smaller indentations that can be missed in some of the other variables. Concave points are a count of these chords. They do not capture the magnitude, just the count. To calculate symmetry the researcher looks at the longest dimension in the cell, then draws a line. Then the researcher draws chords that are perpendicular to this axis. Then each side of each chord is compared to determine how symmetric the nucleus is. The fractal dimension allows the research to look at how regular the contour is, with less regular contour being a higher probability of malignancy. The researchers do not give much insight into units, with the exception of the area. In order to get a better grasp of the data, we have included some exploratory data an alysis using box plots.

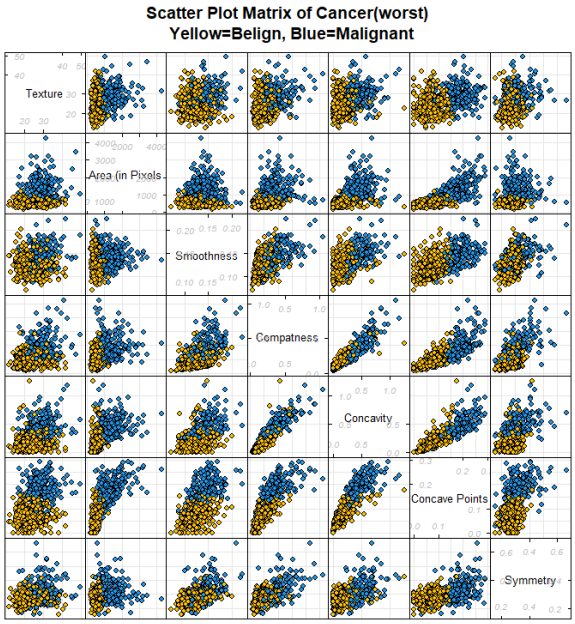
Figure 2: Distribution of Mean Variables



Visually, we notice that the benign tumors have smaller means and medians in every variable except the "Fractal Dimension Mean." This does make sense as malignant tumors tend to have more irregular shapes and tend to be more aggressive and hence larger. In most of the variables, we can see that the malignant tumors have quite a bit more spread between the first and third quadrant when com pared to the benign tumor box plots. Also, of note, there are quite a few outliers in almost all to the data, with every variable and category having at least one. Since benign tumors tend to be more regular, we would have expected fewer outliers in that benign category. Upon inspection of the data, it is clearly not the case. The box plots based on the other variables are similar and included in the appendix.

Below we have included a scatterplot matrix that shows the scatter plots of each of the "worst" variables plotted against each other. The malignant cases are in blue and the benign cases are in yellow.

Figure 3: Scatter Plot Matrix of Cancer(worst) Predictors



This plot is important because it can help guide model selection. For example, if the scatter plot were to show that there is sparse data or clear separation in the data, we would know that a logistic regression may not be as appropriate as a tree-based approach or classification.

Since this data has been around since 1992, other researchers have also done exploratory analyses. According to several papers *k*-nearest neighbors, when $k = 1$, appears to be the best classification technique. Although we did not focus on this technique, we have included in the model comparison summary located in Section IX.

Section III: Cross Validated Logistic Regression Model –

The first model we wanted to look at was a linear model. Multiple linear regression is not appropriate in this case because we are trying to classify whether a tumor is malignant or benign. If we were to use multiple regression we could end up with negative probabilities or probabilities that are greater than one. Either situation would be ridiculous because probabilities must be in the closed interval from zero to one. Since logistic regression models have asymptotic properties, they will force the

estimated probabilities from the interval to be between zero and one. A problem that can occur with logistic regression is that it is possible for the model to not converge and hence you will not have a reliable estimate. When we attempted to run the model with all the predictor variables, we found that we lacked convergence. To get around this we decided to use break the variables into three logical sets that the data had already been broken into – a mean set, an SE set, and a "worst" set. Then we ran logistic regressions with cross-validation (default ten-fold) each of those and decided to test them individually. To improve the model performance, we weighted each model with the sample proportion of the malignant and benign occurrences.

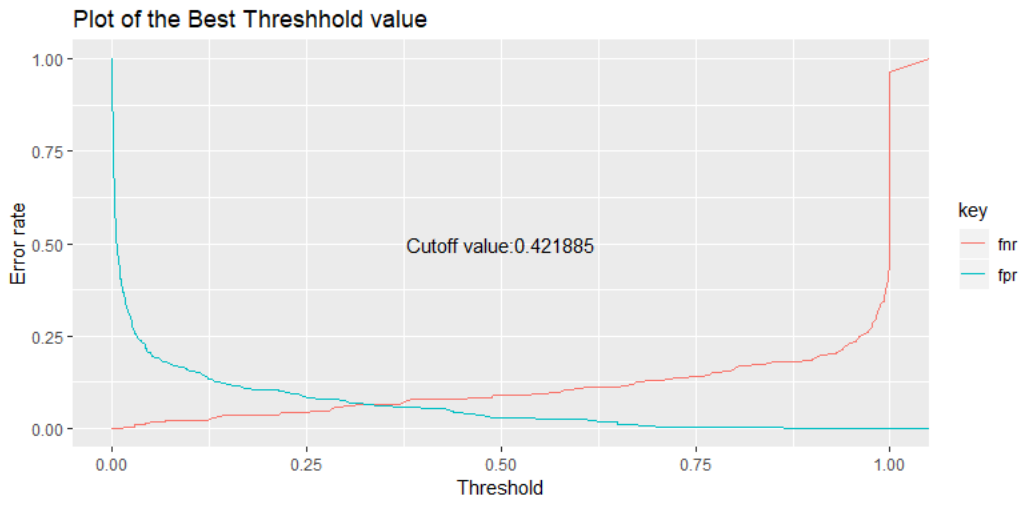
The best model, and the one that we guessed, would be best was the "worst" set. Intuitively, since malignant tumors tend to be larger and more irregular it seemed to be logical that the extreme values represented by these variables would be the best predictors.

Figure 3: ROC Curve for the "Worst" Variable Logistic Regression

Initially, after running the three models we decided to use the sample proportion as the cutoff point between malignant and benign. Although this gave us great model performance (evaluated by in-sample error rate), we wanted a more scientific way to determine what the cutoff point should be. To do this we used a combination of a ROC curve and plotted the error rate against the decision point (or

threshold). The ROC curve, shown above, shows both type I and type II errors simultaneously at all possible thresholds. If the ROC is curved in such a way that it hugs the top left corner of the plot, it is maximizing the area under the curve and hence is likely a good predictor. Clearly, the ROC shows that this is a very good model, as the shape bows toward the top left corner. In figure 4 the "False Negative Rate" is plotted in red, while the "False Positive Rate" is plotted in blue for all the possible cut off thresholds. The intersection of the red and blue curves is where the total error is minimized. From Figure 4 we can see that the best result will come from the rounded cutoff threshold of 0.422. Using this new cutoff value, we will re-run the logistic regression.

Figure 4: Best Threshold Value



After using the new threshold value and running a logistic regression we got an equation of the form:

$$\begin{aligned} logit(\hat{p}) = {} & \beta_0 + \beta_1(radius_{worst}) + \beta_2(texture_{worst}) \\ & + \beta_3(perimeter_{worst}) + \beta_4(area_{worst}) + \beta_5(smoothness_{worst}) \\ & + \beta_6(compactness_{worst}) \\ & + \beta_7(concavity_{worst}) + \beta_8(concave.points_{worst}) \\ & + \beta_9(symmetry_{worst}) + \beta_{10}(fractal.dimension_{worst}) \end{aligned}$$

The coefficients for the model can be seen in the output below. The only significant variables are the intercept, texture (significant at approximately 0), smoothness, and concave points.

Table of Logistic Regression Coefficients

```
Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             -29.00148   13.49446  -2.149   0.0316 *
radius_worst             -0.53543    1.53086  -0.350   0.7265
texture_worst             0.28250    0.06004   4.705 2.54e-06 ***
perimeter_worst           0.01300    0.12846   0.101   0.9194
area_worst                0.01878    0.01465   1.282   0.1998
smoothness_worst         53.94341   21.87949   2.465   0.0137 *
compactness_worst        -8.31719    8.44719  -0.985   0.3248
concavity_worst           4.57985    3.37161   1.358   0.1743
concave.points_worst     37.54868   16.15534   2.324   0.0201 *
symmetry_worst            9.62227    5.84694   1.646   0.0998 .
fractal_dimension_worst  -7.87460   49.29042  -0.160   0.8731
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The training test error rate with this full model is .0193 with the confusion matrix being given below.

	Predicted Benign	Predicted Malignant
Benign	351	6
Malignant	5	207

Section IV: Stepwise Variable Selection –

Another way to utilize logistic regression is to use stepwise variable selections. We decided to use forward and backward stepwise variable selection. Forward stepwise variable selection starts with a null model and tests the performance of all the one variable models, then keeps adding the best predictor until all of the predictors are in the model. Backward stepwise variable selection starts with all the predictors in the model and tries to remove the least costly variable. It will continue doing this until the model is down to a null model. After you have completed the forward or backward selection you can compare the models and see if any stand out as particularly good. Since we are using logistic regression you cannot use AIC,

BIC, or adjusted R-squared to test the models, and instead, use the cross-validated in sample error rate. Surprisingly the forward and backward selections gave the same model. A contributing factor is the relatively small number of predictors we are using. Once again, we focused on the Worst-Case models did work the best and we found the best model to have seven variables and is represented by the following equation.

$$\begin{aligned} logit(\hat{p}) = \beta_0 &+ \beta_1(texture_{worst}) + \beta_2(smoothness_{worst}) \\ &+ \beta_3(compactness_{worst}) + \beta_4(concavity_{worst}) \\ &+ \beta_5(concave.points_{worst}) \\ &+ \beta_6(symmetry_{worst}) + \beta_7(fractal.dimension_{worst}) \end{aligned}$$

With the following output representing the coefficients:

```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -33.015641   5.049506  -6.538 6.22e-11 ***
texture_worst          0.284210   0.059482   4.778 1.77e-06 ***
area_worst             0.014429   0.002379   6.065 1.32e-09 ***
smoothness_worst      52.738309  20.174924   2.614  0.00895 **
compactness_worst     -9.313756   4.619476  -2.016  0.04378 *
concavity_worst        5.060635   2.745489   1.843  0.06529 .
concave.points_worst  36.888177  15.184347   2.429  0.01513 *
symmetry_worst         9.398073   5.639138   1.667  0.09560 .
---
```

We can see from the output that all of the variables except for symmetry and concavity are statistically significant.

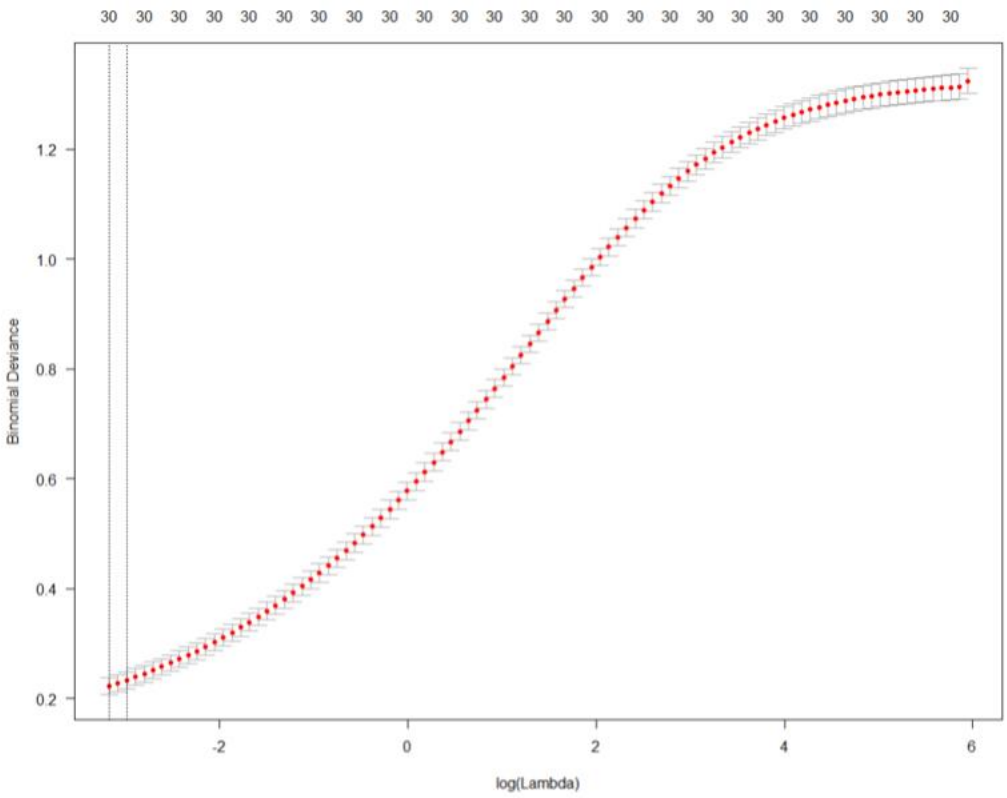
The training test error rate with this full model is .0175 with the confusion matrix being given below. This value is not significantly different from the results in the overfitted model that was considered in the last section.

	Predicted Benign	Predicted Malignant
Benign	352	5
Malignant	5	207

Section V: Logistic Ridge Regression –

Ridge regression is a form of linear model that introduces bias in order to decrease variance. Instead of doing a traditional logistic regression, there is a penalty term added to the end of the minimization problem. A user-defined λ is multiplied by the sum of the squared β coefficients. This form of regression usually does not eliminate any coefficients; however, ridge regression does shrink the coefficients towards zero. We used cross-validation to calculate the optimal value of λ. The plot of the cross validation for the λ selection can be seen below.

Figure 5: λ Selection for Logistic Ridge Regression



Below are the coefficients for the Ridge Regression.

(Intercept)	radius_mean	texture_mean
-1.544438e+01	8.899536e-02	7.099297e-02
perimeter_mean	area_mean	smoothness_mean
1.264755e-02	8.367652e-04	8.897380e+00
compactness_mean	concavity_mean	concave.points_mean
1.344803e+00	3.182955e+00	8.450591e+00
symmetry_mean	fractal_dimension_mean	radius_se
2.712723e+00	-2.386668e+01	1.061626e+00
texture_se	perimeter_se	area_se
-3.466349e-02	1.153972e-01	5.299585e-03
smoothness_se	compactness_se	concavity_se
3.632229e+00	-6.370616e+00	-1.355314e+00
concave.points_se	symmetry_se	fractal_dimension_se
1.493621e+01	-1.019346e+01	-5.916860e+01
radius_worst	texture_worst	perimeter_worst
7.912322e-02	6.399387e-02	1.074124e-02
area_worst	smoothness_worst	compactness_worst
5.940637e-04	1.259086e+01	1.008722e+00
concavity_worst	concave.points_worst	symmetry_worst
1.328482e+00	5.713091e+00	4.514692e+00
fractal_dimension_worst		
5.631352e+00		

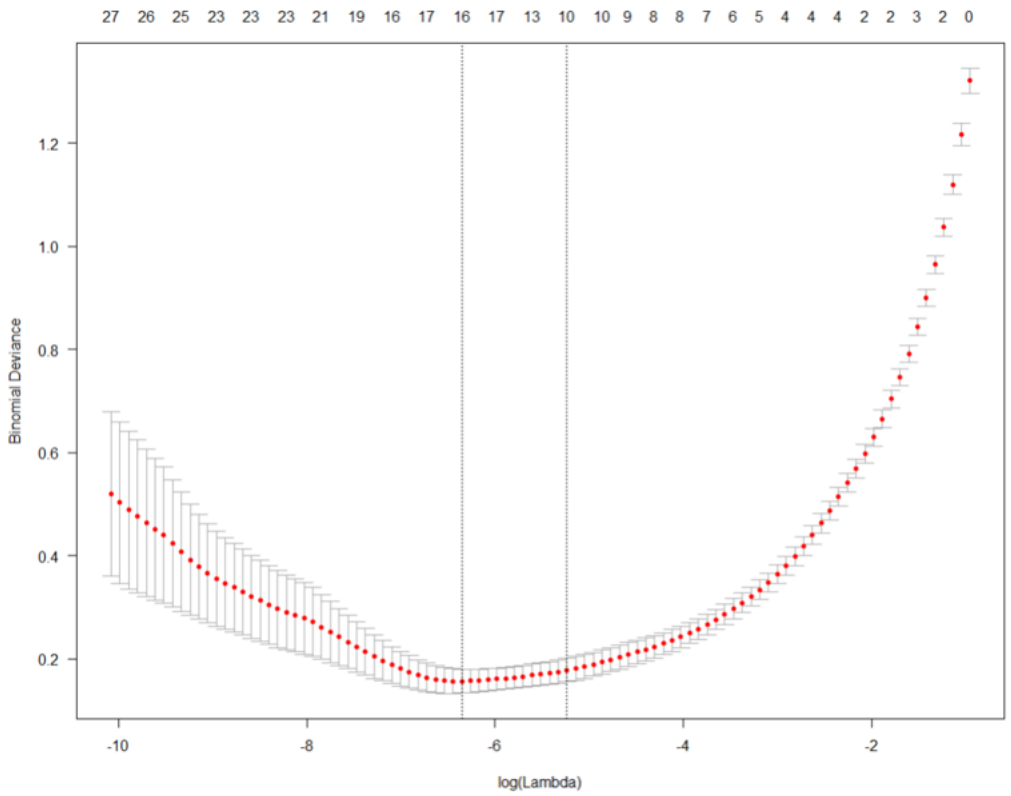
Confusion Matrix for the Ridge Regression:

	Predicted Benign	Predicted Malignant
Benign	354	3
Malignant	7	205

Section VI: Logistic Lasso Regression –

Another way of trimming a model is Lasso Regression. Although the Lasso is a very popular technique on linear regressions, it is a very powerful technique that can also be used on logistic regression models. Lasso Regression adds bias to the model but decreases variance by putting a severe penalty on the coefficients. This penalty is a user-defined coefficient λ that you multiply by the sum of the absolute value of the β coefficients. Unlike the ridge regression from the previous section, lasso regression will eliminate variables.

Figure 6: λ Selection for Logistic Lasso Regression



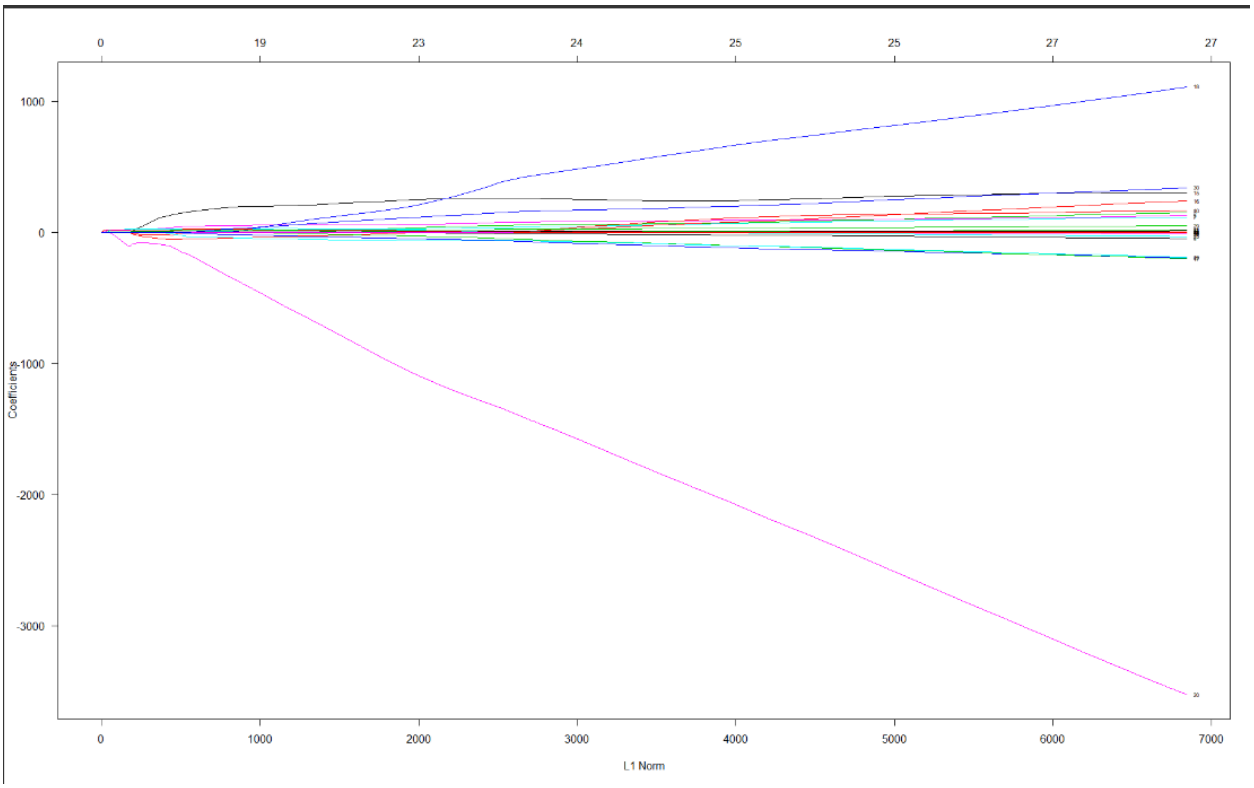
Proposed Lasso Model:

$$\begin{aligned} logit(\hat{p}) = {} & \beta_0 + \beta_1(texture_{mean}) \\ & + \beta_2(concave_{mean}) + \beta_3(radius_{SE}) + \beta_4(fractal.dim_{SE}) \\ & + \beta_5(radius_{worst}) + \beta_6(texture_{worst}) + \beta_7(smoothness_{worst}) + \beta_8(concavity_{worst}) \\ & + \beta_9(concave.points_{worst}) + \beta_{10}(symmetry_{worst}) \end{aligned}$$

Coefficients for Proposed Lasso Model:

(Intercept)	texture_mean	concave.points_mean
-24.04457583	0.02623877	13.51664014
radius_se	fractal_dimension_se	radius_worst
4.17651285	-60.30649321	0.64945502
texture_worst	smoothness_worst	concavity_worst
0.15724246	19.75876625	1.49600364
concave.points_worst	symmetry_worst	
16.50331548	4.66210916	

Figure 7: Lasso Regression Variable Selection



Confusion Matrix for Lasso Regression Model:

	Predicted Benign	Predicted Malignant
Benign	353	4
Malignant	6	206

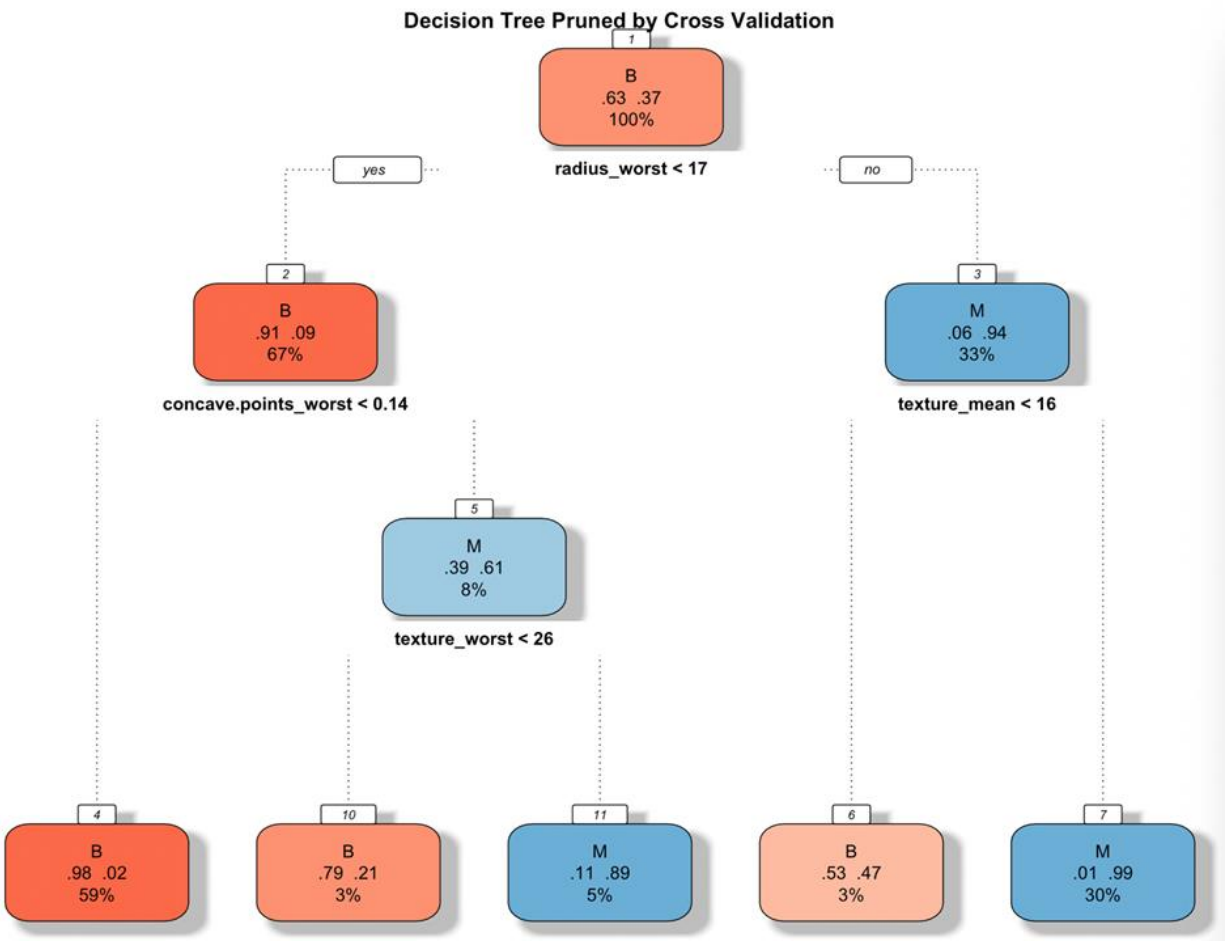
Section VII: Decision Tree –

Alternatively, we can try something that is not regression-based to see if we can get even better results. As out of the bag estimators, decision trees do not tend to be the best classifier; however, decision trees have a distinct advantage in that they are remarkably interpretable. Decision trees look for different places to make binary splits to divide the data. After each split R will look at the sample error rate (since this is a classification setting) and determine if another split is necessary. If the program has not met its

stopping threshold, it will continue making splits; however, sometimes these splits are unnecessary. In the case of our data, there were twelve terminal nodes. While twelve is not a particularly high number, many of the splits would lead to two terminal nodes with the same conclusion. In this case, we can use cross-validation to "prune the nodes." In the case of our model, we found that there are only six necessary terminal nodes.

The most interesting part of this tree is that even though the software had access to all of the variables, the only variables that were selected were from the "Worst" category. Also, of note, the algorithm only needs five of the ten "worst" variables. Also, despite that fact that variables can be used multiple times, no variables were used more than once in this decision tree.

Figure 8: Decision Tree



The pruned tree had an in-sample error rate of 0.03339192, which is higher than the error rate for our regression models. The other problem that we experienced is that the error rate that is most important to minimize (Predicted Benign but actually Malignant), actually goes up substantially.

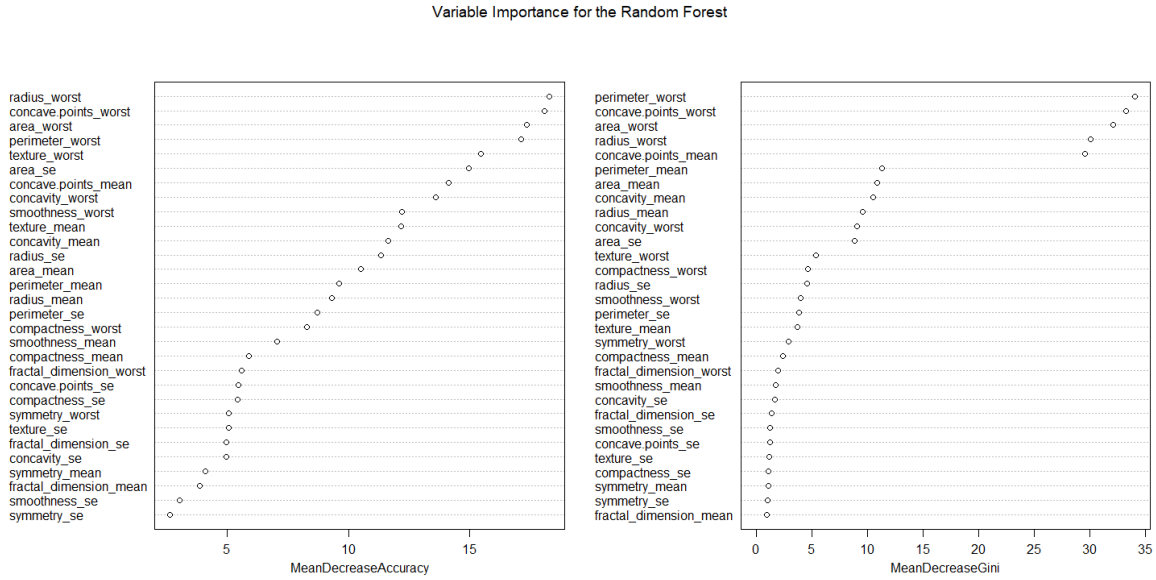
Confusion Matrix for Decision Tree:

	Predicted Benign	Predicted Malignant
Benign	354	3
Malignant	16	196

Section VIII: Random Forests –

Although decision trees are easy to understand and visualize, they are not the most powerful modeling technique. In order to make a more powerful decision tree, we can utilize a technique called random forests. Random forests start by taking many (default of 500 in R) bootstrapped samples and then for each split chooses from only *m* predictors, which is normally considered to be the square root of the number of predictors. Using this technique, we can decorrelate the trees which will in turn substantially (hopefully) reduce the variance. Because of the correlation of our predictors, we would expect this to be the best method of classification for this data. Each tumor that is being considered by this method would have to be considered in each of the 500 models and the classification would be determined by a majority vote. Because of the nature of random forests, they can be very useful in terms of variable selection. Using either mean decrease in accuracy or mean decrease in the Gini Index we can determine what variables are particularly important.

Figure 9: Variable Importance from Random Forest Model



It should be pointed out that, like in the other models, the "Worst" variables are the most important. It is also interesting that the variables selected by the two measures of decrease are not the same and the orders are not the same for the variables that are the same.

The random forest model did surprisingly poorly in this setting with a test misclassification rate of 0.03975973. This is surprising because it is marginally lower than the pruned decision tree in the previous section. The confusion matrix can be seen below. The one particularly important thing about this is that the error rate of tumors that are observed to be benign and are malignant is more than 50% lower. If this trend continued in test data, it is absolutely critical to minimize this error.

	Predicted Benign	Predicted Malignant
Benign	350	7
Malignant	14	198

Section IX: Model Comparisons –

Type II error rate is the most important to minimize in this kind of problem (where you predict a benign tumor but the tumor is actually malignant). To approximate the Type II error rate, for each model we have divided the number of predicted benign but actually malignant, by the total number of malignant tumors. Additionally, we will present the training prediction accuracy. The stepwise logistic model is the best in both Sample Type II Error Rate and Sample Accuracy

	Stepwise Logistic	**Ridge Logistic**	**Lasso Logistic**	**Decision Tree**	**Random Forest**
Sample Type II Error Rate	0.024	0.028	0.028	0.075	0.066
Sample Accuracy	0.982	0.982	0.982	0.967	0.963

Section X: Conclusion –

There are many opportunities for continued research in classifications of cancer. The best model we found was the StepWise model selection (both forward and backward), with a Type II error rate 0.0236 and a total accuracy of 98.1%. The in-sample model performance of the decision tree was surprisingly poor, however, with more data to data to test we would expect the tree methods to be much better. For future research, we would like to continue to address the variable collinearity and would continue to test other classification methods such as Principal Component Analysis, LDA, QDA, and Support Vector Machines.

Section XI: References –

1. American Cancer Society. Breast Cancer Facts & Figures 2017-2018. Atlanta: American Cancer Society, Inc. 2017.
2. "Susan G. Komen." *Susan G. Komen®*, ww5.komen.org/.
3. UCI. "Breast Cancer Wisconsin (Diagnostic) Data Set." *RSNA Pneumonia Detection Challenge | Kaggle*, 25 Sept. 2016, www.kaggle.com/uciml/breast-cancer-wisconsin-data.
4. Street , W. Nick, et al. Nuclear Feature Extraction for Breast Tumor Diagnosis. 28 Dec. 1992.
5. Witten, Daniela, et al. "Introduction to Statistical Learning." 2017, www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf.