Predicting Breast Cancer - STATS 412

Anders Ward, Kaustubh Deshpande, Sasha Farzin-Nia

UCLA Masters of Applied Statistic Program, University of California Los Angeles, CA

Summary

Using 4 different modelling techniques, namely, Logistic Regression, Principal Component Analysis (PCA), Ridge, and Lasso Logistic Regression, we predicted whether a biopsy specimen found in the breasts was benign or malignant. We found that from our Leave-One-Out Cross-Validation results, PCA and Lasso Logistic Regression performed the best in terms of accuracy and recall.

Key words: Regression, Logistic, Ridge, Lasso, Prediction, Cross-Validation.

1. Introduction

With Breast Cancer being one of the most common types of cancers among women (along with lung and skin), it is imperative that women are frequently tested for breast cancer. If a mass is discovered in the breasts, usually a biopsy is taken, and it is up to a pathologist to determine the status of the mass, whether or not the mass is benign (not-cancerous), or malignant (cancerous). With the advances in technology and in statistics, Machine Learning and Deep Learning Algorithms can help confirm the results given by specialists in the field and also provide support to countries where specialization is not as widely available. This will hopefully lead to faster diagnosis times, thus increased survival rates. In this paper, we take a look at Logistic Regression

and model selection methods to see how accurately we can differentiate between a benign and a malignant mass.

The dataset we use is the "Breast Cancer Wisconsin (Diagnostic) Data Set", made publicly available on the UCI Machine Learning Repository [1]. The features of the dataset are computed from the digitized image of the biopsy specimen. The measurements of the cell nuclei of the specimen included are:

- radius;
- texture;
- perimeter;
- area;
- smoothness;
- compactness;
- concavity;
- concave points;
- symmetry;
- fractal dimension.

1.1 *Logistic Regression*

Logistic Regression is a Machine Learning Algorithm, which is used for classification purposes. Unlike with Ordinary Least Squares (OLS) Regression, our response is a probability of whether or not an event will occur. While the prediction range, y, in OLS is

$$y \in (-\infty, \infty),$$

in Logistic Regression, we use a logistic function, $h(x)$, called the sigmoid function, which maps our predicted values (which can be any real value), to be another value between 0 and 1 [2]. That is,

$$0 \leqslant h(x) \leqslant 1.$$

The sigmoid function is defined as follows:

$$h(x) = \frac{1}{1 + e^{-x}}.$$

Since Logistic Regression is predicting the outcome of a binary response, we make use of a decision boundary. If our predicted probability is greater than the decision boundary, then we label our prediction one class, and if it is less than the decision boundary, then we label our prediction another class.

This paper will employ the use of Logistic Regression on the dataset.

1.2 *Principle Component Analysis*

Principle Component Analysis, or PCA, is primarily used to reduce the dimensionality of large datasets. In this method, "principal components" are newly constructed variables that are a linear combination of the initial variables. These principal components are uncorrelated to each other and preserve most of the information from the original data.

This is done by attempting to compress as much information as possible into the first component, then compress the maximum remaining information in the second, and this process continues util all the information is preserved. We may then choose a specific number of components for further analysis as needed on a case to case basis.

This paper will employ the use of PCA on the dataset.

1.3 *Ridge and Lasso Regression*

Along with PCA, another common set of techniques in highly collinear data sets are Ridge and Lasso regression. Ridge regression is a model shrinkage method that adds an additional term to the OLS cost function,

$$\lambda\beta^T\beta.$$

The result is a restriction on the length of the beta vector,

$$\hat{\beta_\lambda^T}\hat{\beta}_\lambda = t^2.$$

This restriction effectively trades an increased bias for reduced error. Lasso regression is much the same, with a different restriction on the beta values,

$$\sum B_j = |t|.$$

Unlike Ridge regression, Lasso regression has no explicit solution. The other major difference is that while Ridge regression lowers the beta values, Lasso regression will eliminate some predictors entirely by setting their coefficient to 0.

This paper will employ the use of both Ridge and Lasso regression on the dataset.

2. METHODOLOGY

For each model, we used the same training-testing split of the data to keep consistency. The split was performed by the *caret* package, and we selected 80% and 20% for training and testing respectively. We used the seed of 3110 throughout in order to maintain reproducibility.

2.1 *Logistic Regression*

One of the assumptions of Logistic Regression is that there is little or no multicollinearity between the explanatory variables. As seen from "Fig. 6", we have multiple explanatory variables which

have a high correlation between one-another. The highest being radius mean and perimeter mean. In order to meet this assumption of Logistic Regression, we removed all the variables which had a collinearity of 90% or greater.

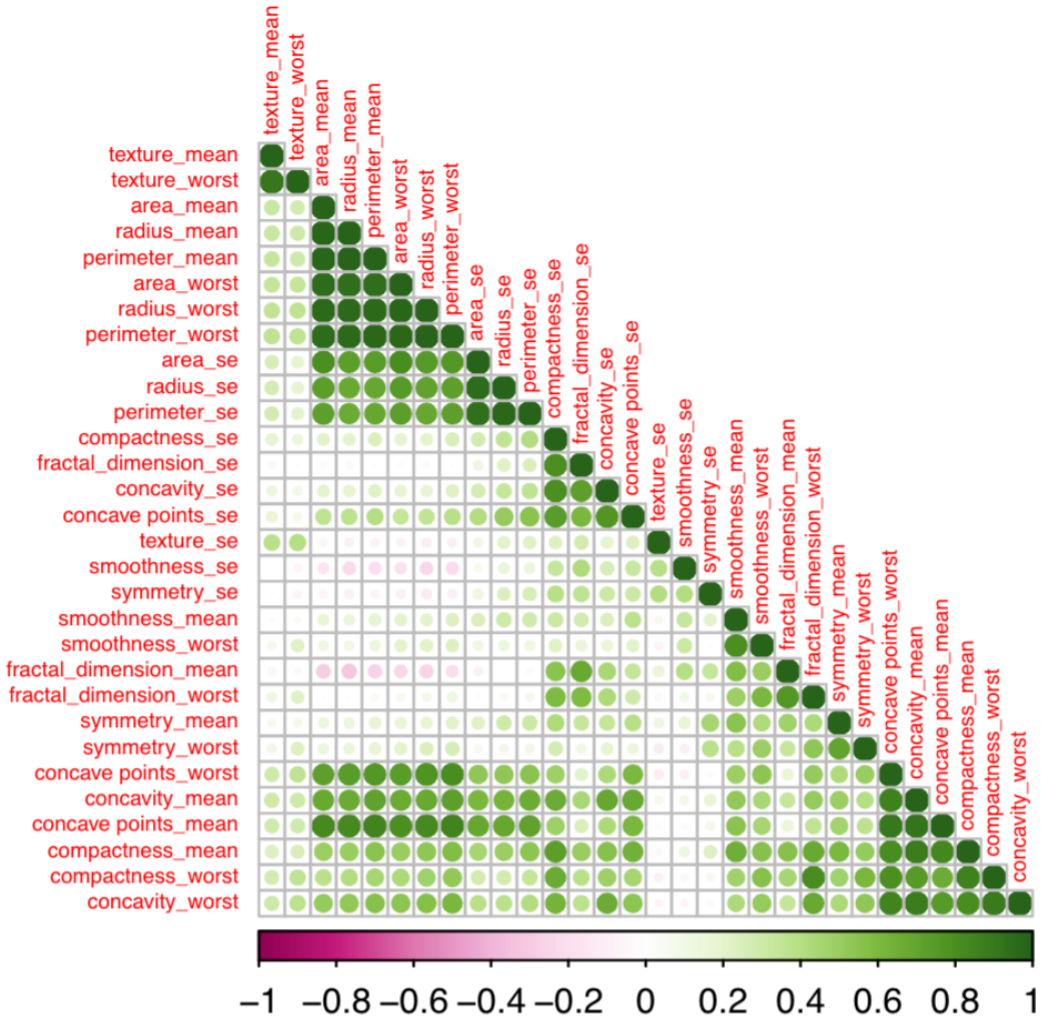

Fig. 1. Correlation Plot

As a benchmark model we fit a Generalized Linear Model (GLM) of binomial family type,

and using a logit link function. We fit our response, with all remaining variables once our removal of collinear explanatory variables was complete. Any additional model created is done so with the intent of improving on the accuracy $\left(\frac{\text{True Positive + True Negative}}{\text{Total}}\right)$ produced by this benchmark model.

2.2 *Principal Component Analysis*

Using R's inbuilt *prcomp*, we performed principal component analysis on the dataset of 30 explanatory variables. "Fig. 2" plots the proportion of variance explained by each principal component (PC). From this plot we can identify the first 10 PCs to contain a majority of the variation.

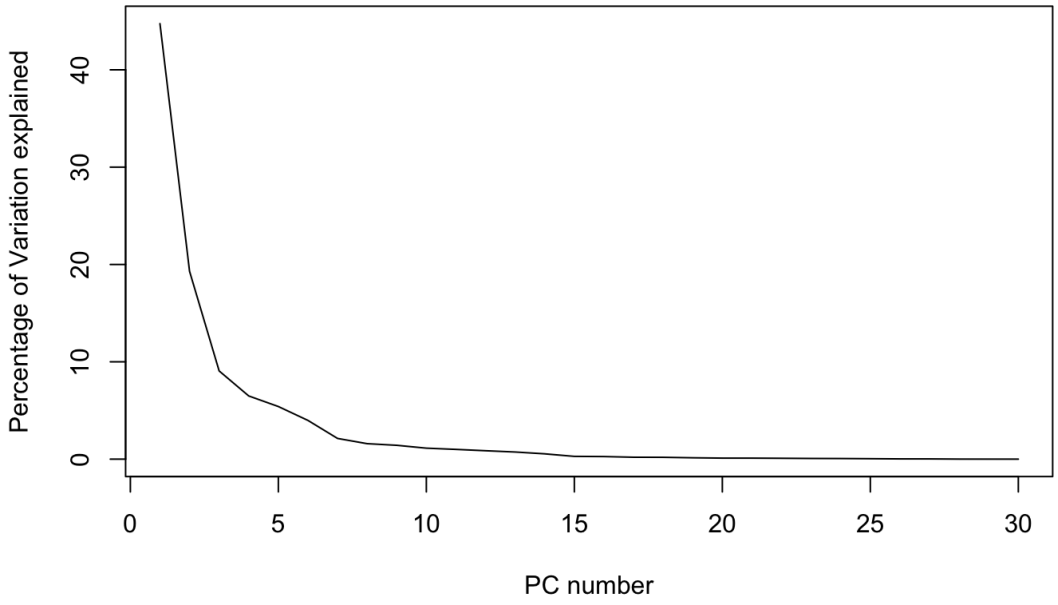

Fig. 2. Principal Components Vs. Variation Explained

We constructed a general linear model (glm) to predict diagnosis using the first principal component, then the first two, then the first three and so on until we covered the first 10 principal

components. We used precision, accuracy and recall as metrics to evaluate how many PCs to proceed with. "Fig. 3" indicated that using the first 9 or 10 PCs maximizes the accuracy of our glm model. So for the sake of less complexity we decided to proceed with using the first 9 PCs.

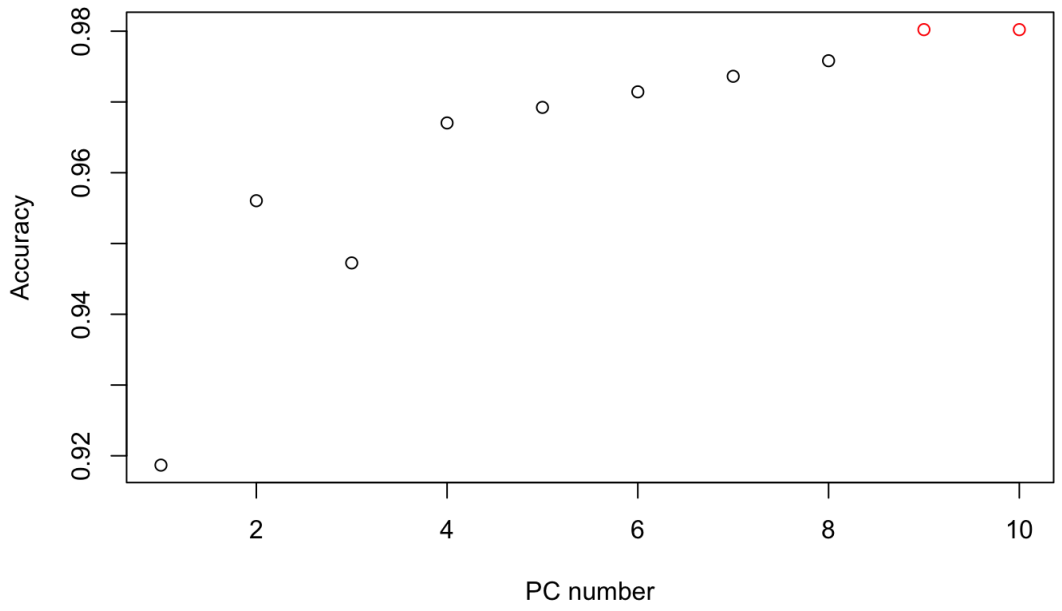

Fig. 3. Principal Component Vs. Accuracy

2.3 *Ridge and Lasso Regression*

Using the *glmnet* package, we performed both Ridge and Lasso regression on the training data. The lambda values for our two models were selected automatically by the built in search functionality of *cv.glmnet*. This automatically creates a sequence of lambda values, linear on the log scale, to search over. A fit is performed on the training data for each lambda value. The selected lambda value was that which minimized the standard error of the 10-Fold Cross-Validation performed by *cv.glmnet* behind the scenes. This process selected the lambda values of 3.8×10^{-2} and 2.1×10^{-3} for Ridge and Lasso respectively.

With these lambda values and their corresponding fits, we predicted on the testing data.

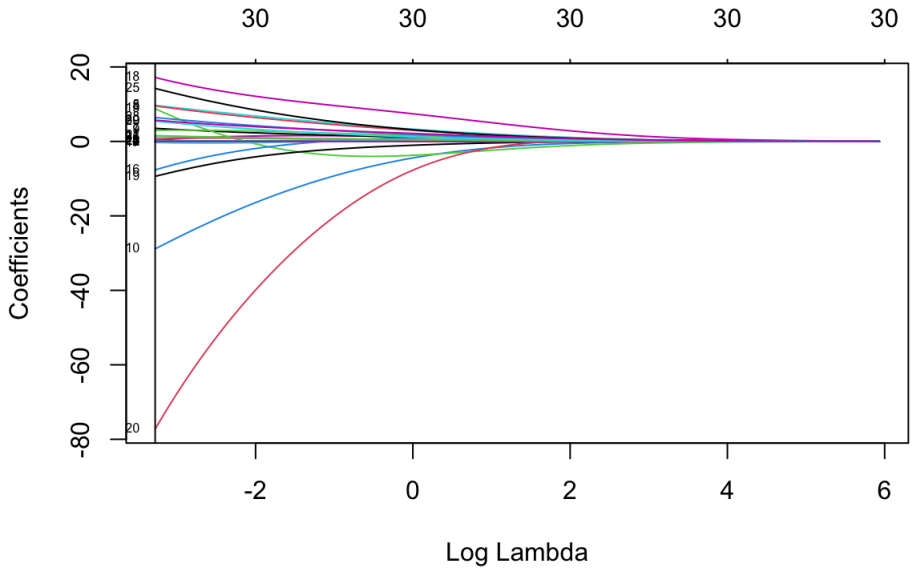

Fig. 4. Vertical line is optimal lambda

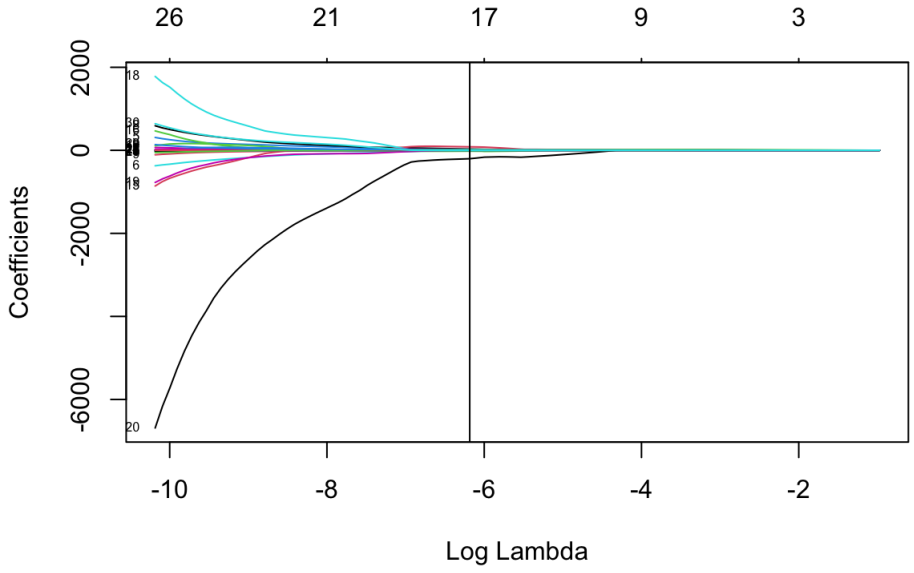

Fig. 5. Vertical line is optimal lambda

2.4 *Leave-One-Out Cross-Validation (LOOCV)*

We wrote a LOOCV algorithm from scratch in order to compare the best resulting models from PCA, Lasso, and Ridge analysis. Our algorithm leaves out one row from the original data set and then fits on the remaining data set using previously identified optimal parameter values. A predicted probability of being malignant is then formed on the row of data that was left out, and this prediction is stored for each PCA, Ridge, and Lasso. The prediction is in the form of a 0 to 1 value, which we then convert to a classification, either "benign" or "malignant". Any prediction value greater than the decision boundary is labeled "malignant". For our initial test we used a decision boundary of 0.5 for all models.

3. Results and Conclusions

The predictive accuracy of most of the models we tested was very high. Considering the subject matter, predicting cancer with a biopsy, high accuracy is required. The penalty of incorrectly diagnosing a patient is quite high. Although it depends greatly on the specific health institution, we would assume that a false positive (malignant diagnosis of benign tumor) is much preferred to a false negative. A patient incorrectly diagnosed with a malignant tumor would be found to be cancer free during a future test performed by a physician, whereas a patient incorrectly diagnosed with a benign tumor might not seek out future care and thus not discover they had cancer.

It is therefore important to maximize the recall of our chosen model. One such method could be to flag predictions that are not asymptotically 0 or 1 as indeterminate, so that a physician could follow up with a human diagnosis. Another method might be to lower the decision boundary, so that any prediction not asymptotically 0 would be considered malignant.

3.1 *Logistic Regression*

Our benchmark model performed reasonably well with an accuracy of 94.7%.

3.2 *Principal Component Analysis*

The general linear model constructed using the first 9 PCs out of 30, performed really well on the test set with an accuracy of 98.2%, which is 3.5% higher than the benchmark Logistic Regression model.

While using the first 9 PCs resulted in highest accuracy, just looking at the first two PCs demonstrates how powerful this technique is.

From the breakdown of the first principle component in table 1 it is interesting to see that all values are negative. The most prominent variables are mean concavity and mean concave points. While the least prominent variables are texture, symmetry, smoothness and fractional dimension mean. Our understanding from this PC is that it is concavity of the tumor in consideration that creates largest divide.

Coming to the second principal component we see a more equal spread of positive and negative values. From the breakdown in Table 2 we see that radius mean, area mean and area worst are the largest positive values. While, fractional dimension worst, fractional dimension se and compactness se are the most negative values. Our interpretation is that a tumor may possess prominent features belonging to the positive group or prominent features belonging to the negative group. Additionally, a tumor is less likely posses prominent features in both of these groups.

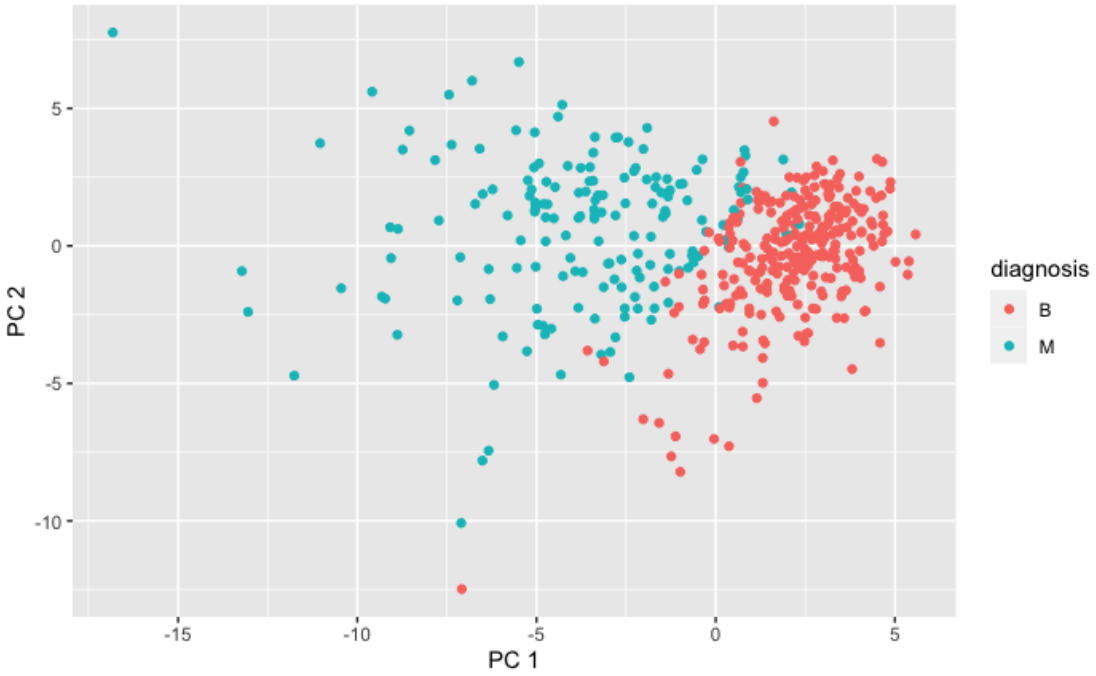

Fig. 6. PC2 vs PC1

Figure 6 plots PC2 vs PC1 and indicates that there is definitely noticeable clustering of the two diagnosis. The malignant diagnosis tend slightly right of the center with a lot of outliers towards the left upper and left lower quadrants. On the other hand, the benign diagnosis tend towards the far right in the right quadrant with much less outliers. Thus, we can see that just the first two PCs alone can be considered fairly effective at determining whether a tumor is malignant or benign.

3.3 *Ridge and Lasso Regression*

The coefficients chosen by the Ridge regression model, are shown in table 3. Referencing the same table for Lasso regression, table 4, we see that our Lasso model eliminated 14 of the 30 predictors. The Lasso coefficients range from 10^{-3} to 10^{2}, whereas the Ridge coefficients range from 10^{-4} to

10^1. It appears the Lasso model, by reducing some coefficients to 0, was able to give more weight to its remaining predictors, whereas the Ridge model kept all predictors, but had the range of coefficient values shifted down a power of 10.

The accuracy of the models from the initial $80 - 20$ split, are shown in table 5. Here we see that both Ridge and Lasso scored perfectly on the test data. It is likely that changing the seed, and therefore the training-testing shuffle, would lower this accuracy to the high 90s. When compared to the PCA accuracy of 98.2%, we cannot make the claim that Ridge or Lasso are better than PCA, due to insufficient data. The 20% of our data used for testing is only made up of 113 observations, and the PCA model only missed 2 of those (both were false negatives). From these results we can conclude that PCA, Ridge, and Lasso are all comparable in terms of accuracy. Additionally we can conclude that our models will have high accuracy if used on real data. In order to gain more knowledge on the predictive accuracy of our models, we performed LOOCV.

3.4 *Leave-One-Out Cross-Validation*

The results for LOOCV, are shown in table 6. PCA missed two more results than Lasso, and Ridge missed two more than PCA. From this data it appears our models would have an accuracy around 97% on new data. Such a level of accuracy would likely be more useful as an aide to physicians rather than a replacement. All three accuracies are very close, so the recall of each model will be the deciding factor. Tables 7, 8, and 9 display the confusion matrices for our LOOCV test. We can see that the PCA and Lasso models tied for recall of 96.2%, while the Ridge model had a comparably abysmal recall of 93.4%. What is even worse for the Ridge model is the 14 to 2 ratio of false negatives to false positives. Rather than tune the Ridge model to be in line with the other two, we recommend the use of either the PCA model or the Lasso model. The PCA model provides additional interpretability of the meaning of each PC, whereas the Lasso

model provides the opportunity to reduce the number of predictors calculated prior to statistical modelling.

By examining the predicted probabilities the PCA and Lasso models gave during false negatives, we can change the decision boundary to reduce the rate of false negatives and increase the rate for false positives. This will be discussed more in the future work section.

Table 1. First Principal Component

Predictor	**Loading**
radius mean	-0.22
texture mean	-0.11
perimeter mean	-0.23
area mean	-0.22
smoothness mean	-0.14
compactness mean	-0.24
concavity mean	-0.26
concave points mean	-0.26
symmetry mean	-0.14
fractal dimensions mean	-0.06
radius se	-0.21
texture se	-0.01
perimeter se	-0.22
area se	-0.21
smoothness se	-0.01
compactness se	-0.17
concavity se	-0.16
concave points se	-0.18
symmetry se	-0.04
fractal dimension se	-0.10
radius worst	-0.23
texture worst	-0.11
perimeter worst	-0.24
area worst	-0.22
smoothness worst	-0.13
compactness worst	-0.21
concavity worst	-0.23
concave points worst	-0.25
symmetry worst	-0.13
fractal dimension worst	-0.13

Table 2. Second Principal Component

Predictor	**Loading**
radius mean	0.23
texture mean	0.06
perimeter mean	0.21
area mean	0.23
smoothness mean	-0.19
compactness mean	-0.15
concavity mean	-0.06
concave points mean	0.03
symmetry mean	-0.19
fractal dimensions mean	-0.37
radius se	0.09
texture se	-0.10
perimeter se	0.08
area se	0.15
smoothness se	-0.22
compactness se	-0.23
concavity se	-0.20
concave points se	-0.13
symmetry se	-0.20
fractal dimension se	-0.28
radius worst	0.22
texture worst	0.05
perimeter worst	0.20
area worst	0.22
smoothness worst	-0.17
compactness worst	-0.13
concavity worst	-0.09
concave points worst	0.01
symmetry worst	-0.14
fractal dimension worst	-0.27

Table 5. First Test Results

Model	**Test 1 Accuracy**
Logistic	94.7%
PCA	98.2%
Ridge	100%
Lasso	100%

[c] Model testing accuracy from 80-20 split

Table 3. Ridge Regression Coefficients

Predictor	**Coefficient**
intercept	-1.69×10^{1}
radius mean	9.24×10^{-2}
texture mean	8.09×10^{-2}
perimeter mean	1.31×10^{-2}
area mean	8.89×10^{-4}
smoothness mean	9.69
compactness mean	2.96×10^{-1}
concavity mean	3.52
concave points mean	9.55
symmetry mean	2.97
fractal dimensions mean	-2.89×10^{1}
radius se	1.48
texture se	-1.03×10^{-1}
perimeter se	1.71×10^{-1}
area se	7.48×10^{-3}
smoothness se	8.79
compactness se	-7.65
concavity se	-3.15×10^{-1}
concave points se	1.72×10^{1}
symmetry se	-9.38
fractal dimension se	-7.72×10^{1}
radius worst	8.68×10^{-2}
texture worst	7.43×10^{-2}
perimeter worst	1.18×10^{-2}
area worst	6.44×10^{-4}
smoothness worst	1.42×10^{1}
compactness worst	8.67×10^{1}
concavity worst	1.56
concave points worst	6.42
symmetry worst	5.44
fractal dimension worst	5.70

[a]Lambda = 3.8×10^{-2}. Rounded to two significant digits.

Table 6. LOOCV Results

Model	**Test 1 Accuracy**
PCA	97.5%
Ridge	97.2%
Lasso	97.9%

[d] Model testing accuracy from LOOCV

Table 4. Lasso Regression Coefficients

Predictor	**Coefficient**
intercept	-2.84×10^{1}
compactness mean	-7.97
concavity mean	4.09
concave points mean	3.08×10^{1}
radius se	1.08×10^{1}
texture se	-1.05
smoothness se	8.44×10^{1}
compactness se	-2.04×10^{1}
fractal dimension se	-2.02×10^{2}
radius worst	3.19×10^{-1}
texture worst	2.82×10^{-1}
perimeter worst	4.89×10^{-3}
area worst	3.61×10^{-3}
smoothness worst	2.29×10^{1}
concavity worst	3.82
concave points worst	1.95×10^{1}
symmetry worst	9.17

[b]Lambda = 2.1×10^{-3}. Rounded to two significant digits.

Table 7. LOOCV PCA Confusion Matrix

	Benign	Malignant
Benign	350	6
Malignant	8	204

[e] Actual values are the rows

Table 8. LOOCV Ridge Confusion Matrix

	Benign	Malignant
Benign	354	2
Malignant	14	198

[f] Actual values are the rows

Table 9. LOOCV Lasso Confusion Matrix

	Benign	Malignant
Benign	352	4
Malignant	8	204

[g] Actual values are the rows

4. Future Work

For future work, we would want to compare the different link functions other than logit, in order to get a grasp of which link function performs best. Whilst logit is used primarily due to its ability to produce interpretable results, with this use case, our aim is to get prediction as accurate as possible, and interpretability may not be as needed. We could also compare these values to that given by a classification algorithm such as K-Nearest Neighbors or Random Forest. Also, we would want to implement Deep Learning Models, such as Convolutional Neural Networks, with all of these future endeavours leading to ultimate goal of minimizing the type II error rate (false negative rate), that is, we don't want to label a cancerous tumor as benign. Further computational study into the best decision boundary for each model could be performed. With a cost value for false negatives and false positives, we could find the optimal decision boundary for each model.

References

Olvi L. Mangasarian Dr. William H. Wolberg, W. Nick Street. Breast cancer wisconsin (diagnostic) data set. `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\%28Diagnostic\%29`, 1995 (Accessed December 5, 2021).

Ayush Pant. Introduction to logistic regression. `https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148`, 2019 (Accessed December 5, 2021).