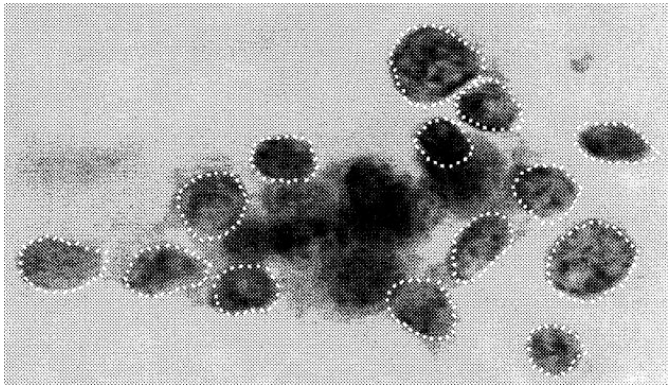
Group 1 Modeling

Puja Ammineni & Sam Kalkman

Purpose & Dataset Information

- Purpose: predicting a diagnosis of breast cancer based on medical imaging data.
- Response variable: Diagnosis
  - Classes: Benign, Malignant
- Sample Size: 569
- Number of categorical predictors: 0
- Number of continuous predictors; 30



Purpose & Dataset Information - cont.

- Predictors can be understood in two ways: feature type and statistic type
- There are 10 feature types relating the nucleus of cells obtained by flattening these cells against a slide.
    - radius, perimeter, area, compactness, smoothness, concavity, concavity points, symmetry, fractal dimension, and texture.
- There are 3 statistics types for feature type: mean, extreme (large), and standard error.
    - These types are notated as 1, 2 and 3 respectively (ex. radius1 is mean and radius2 is extreme)

Pre Processing Steps

- See if we need to add dummy variables
- Check for missing variables
- Correlated predictor removal
- Box-Cox Transformation
- PC Analysis
- Center and Scale Transformations
- Spatial Sign for Outliers Transformation
- Data Splitting&Resampling

Dummy Variables

WIth no categorical variables, there is no need to add dummy variables for analysis

Missing Variables

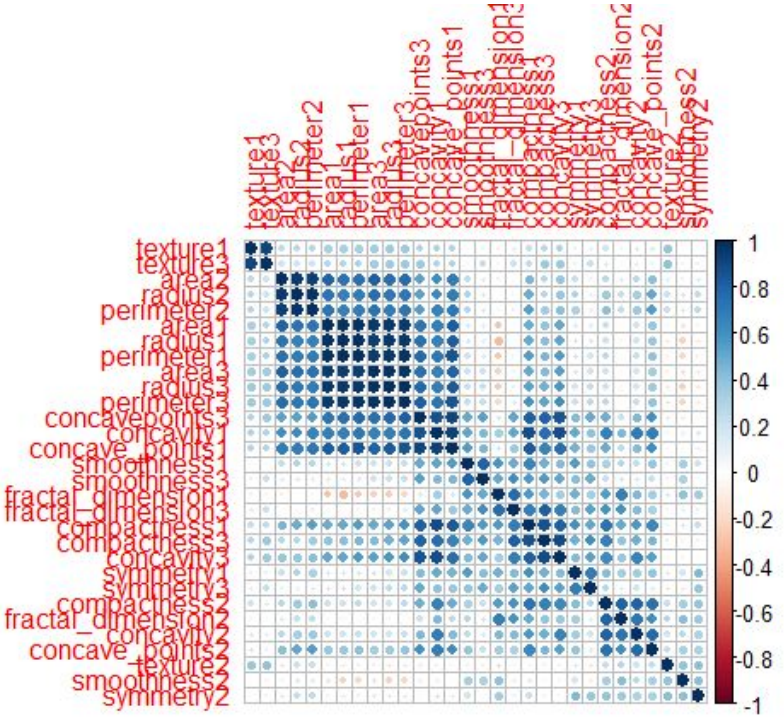
- Our data set does not have any missing variables, imputation is not necessary

Deleting Predictors - Near Zero Variance

- With no categorical predictors, we do not need to check for near zero variance predictors

Deleting Predictors - Correlation

- There are a number of highly correlated variables
- This makes sense: some of the predictors are the different statistics for the same feature type.

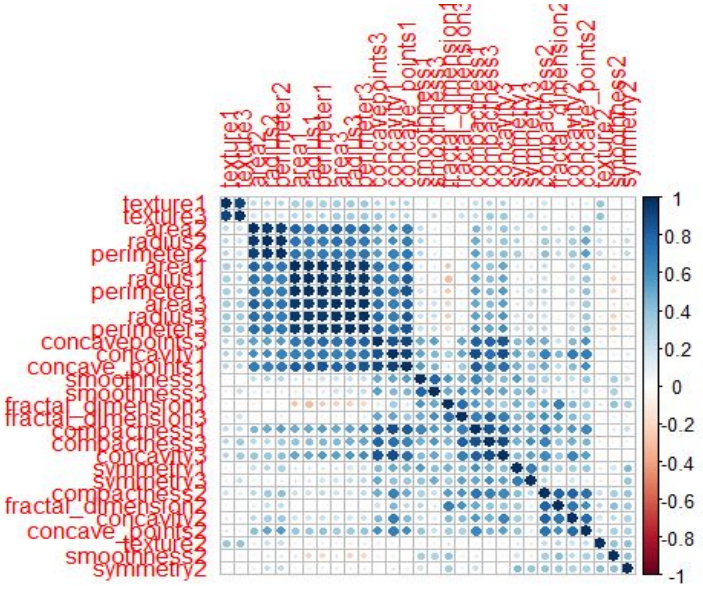


Deleting Predictors - Correlation

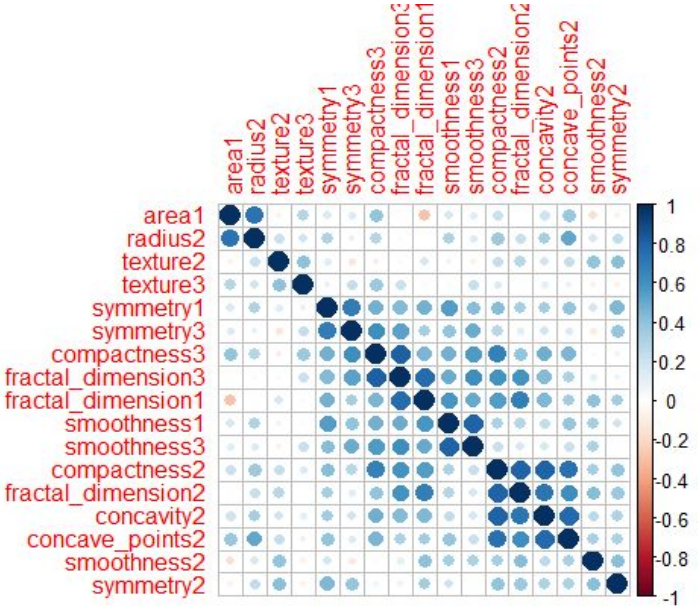
- We are checking for highly correlated predictors at the 0.85 level
- Predictors concavity1, concave_points1, compactness1, concavepoints3, concavity3, perimeter3, radius3, perimeter1, area3, radius1, perimeter2, area2, and texture1 were all removed (13 in total, bringing our number of predictors down to 17)
- This means that perimeter is not a feature used in our model anymore. All other features are still covered by at least one predictor.

Deleting Predictors - Correlation - Before and After

Before:

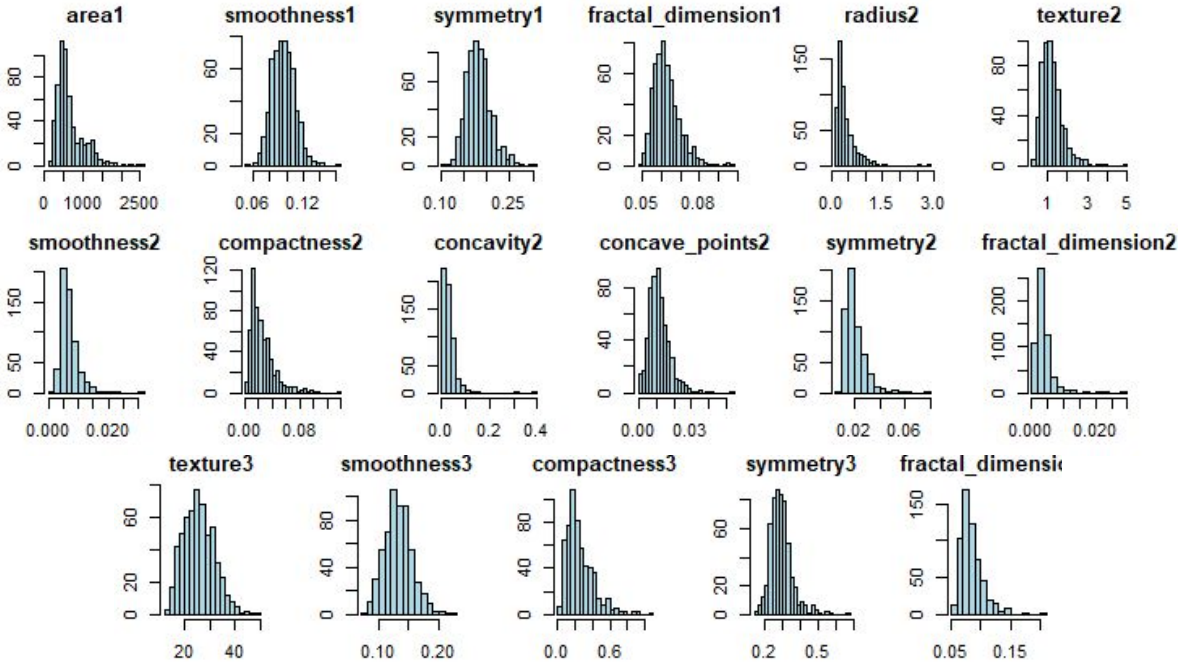


After:



Box-Cox Transformation

- We have a number of variables that are not roughly symmetric
- A majority of these variables are right skewed
- Box-Cox transformation can help



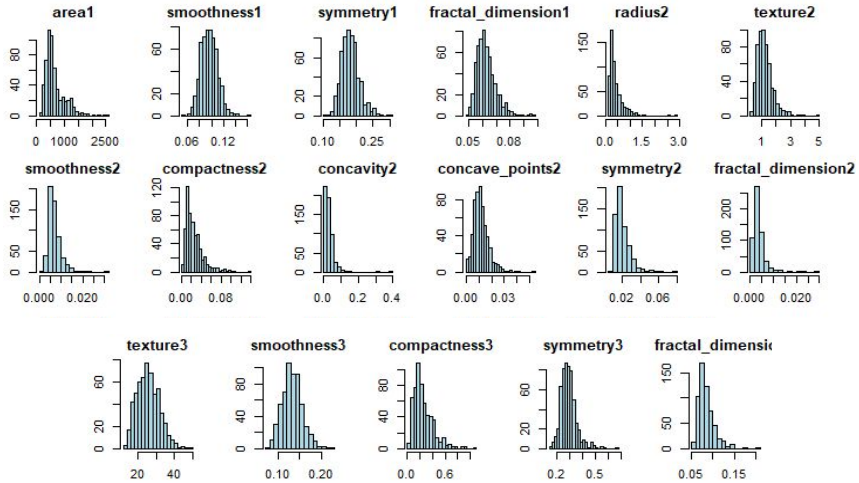
Box-Cox Transformation

- We have many rightly skewed predictors
- The only roughly symmetric predictors are Smoothness1, Texture3, and Smoothness3 which will not have Box-Cox applied

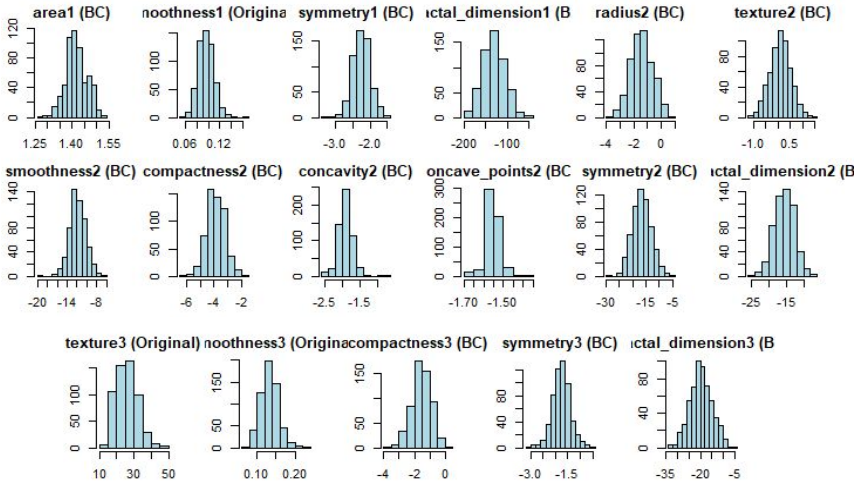
Predictor	Skewness	Interpretation
Area1	1.6370654	Highly Right-Skewed
Smoothness1	0.4539207	Approximately Symmetric
Symmetry1	0.7217877	Moderately Right-Skewed
Fractal_Dimension1	1.2976191	Highly Right-Skewed
Radius2	3.0723468	Highly Right-Skewed
Texture2	1.6377733	Highly Right-Skewed
Smoothness2	2.3022616	Highly Right-Skewed
Compactness2	1.8922032	Highly Right-Skewed
Concavity2	5.0835502	Highly Right-Skewed
Concave_Points2	1.4370701	Highly Right-Skewed
Symmetry2	2.1835728	Highly Right-Skewed
Fractal_Dimension2	3.9033041	Highly Right-Skewed
Texture3	0.495697	Approximately Symmetric
Smoothness3	0.4132383	Approximately Symmetric
Compactness3	1.4657948	Highly Right-Skewed
Symmetry3	1.4263764	Highly Right-Skewed
Fractal_Dimension3	1.6538237	Highly Right-Skewed

Box-Cox Transformation - Before and After Histograms

Before:



After:



Box-Cox Transformation - Before and After Skewness Tables

Before:

Predictor	Skewness	Interpretation
Area1	1.6370654	Highly Right-Skewed
Smoothness1	0.4539207	Approximately Symmetric
Symmetry1	0.7217877	Moderately Right-Skewed
Fractal_Dimension1	1.2976191	Highly Right-Skewed
Radius2	3.0723468	Highly Right-Skewed
Texture2	1.6377733	Highly Right-Skewed
Smoothness2	2.3022616	Highly Right-Skewed
Compactness2	1.8922032	Highly Right-Skewed
Concavity2	5.0835502	Highly Right-Skewed
Concave_Points2	1.4370701	Highly Right-Skewed
Symmetry2	2.1835728	Highly Right-Skewed
Fractal_Dimension2	3.9033041	Highly Right-Skewed
Texture3	0.495697	Approximately Symmetric
Smoothness3	0.4132383	Approximately Symmetric
Compactness3	1.4657948	Highly Right-Skewed
Symmetry3	1.4263764	Highly Right-Skewed
Fractal_Dimension3	1.6538237	Highly Right-Skewed

After:

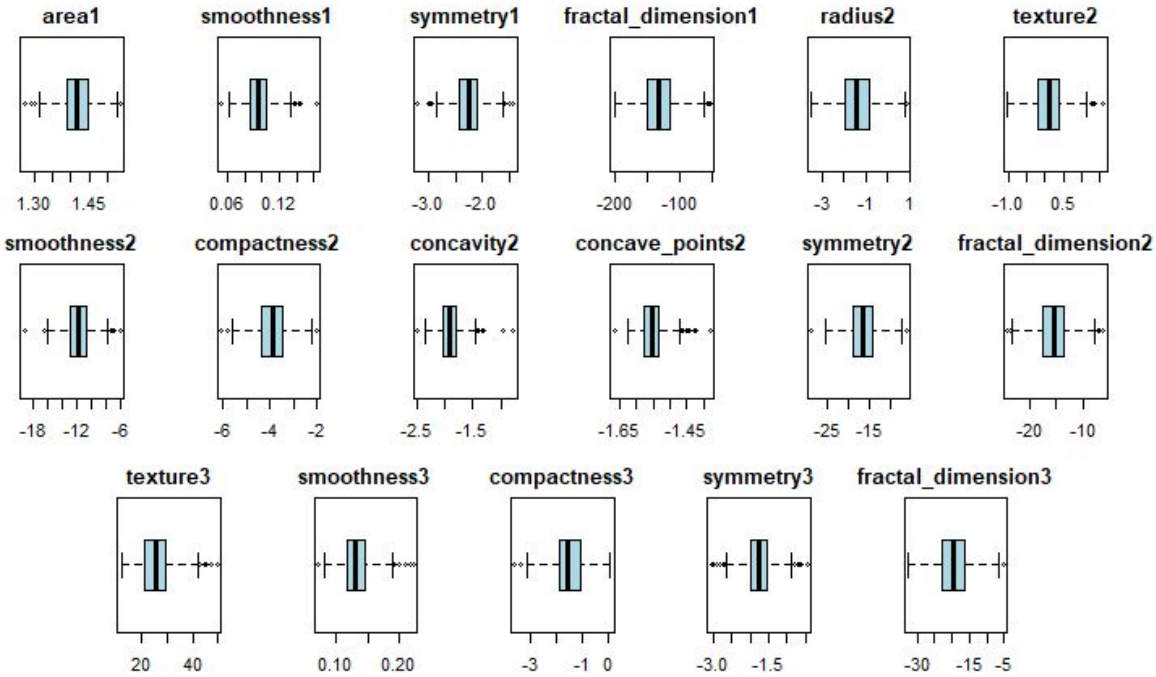
Predictor	Skewness	Interpretation
Area1	0.011372	Approximately Symmetric
Smoothness1	0.4539207	Approximately Symmetric
Symmetry1	0.0017377	Approximately Symmetric
Fractal_Dimension1	0.1506466	Approximately Symmetric
Radius2	0.0271761	Approximately Symmetric
Texture2	0.0290368	Approximately Symmetric
Smoothness2	-0.024012	Approximately Symmetric
Compactness2	-0.0040198	Approximately Symmetric
Concavity2	0.265713	Approximately Symmetric
Concave_Points2	0.0960942	Approximately Symmetric
Symmetry2	0.0549106	Approximately Symmetric
Fractal_Dimension2	0.0121915	Approximately Symmetric
Texture3	0.495697	Approximately Symmetric
Smoothness3	0.4132383	Approximately Symmetric
Compactness3	-0.2206758	Approximately Symmetric
Symmetry3	-0.056549	Approximately Symmetric
`Fractal_Dimension3	0.0470535	Approximately Symmetric

Center and Scale Transformations

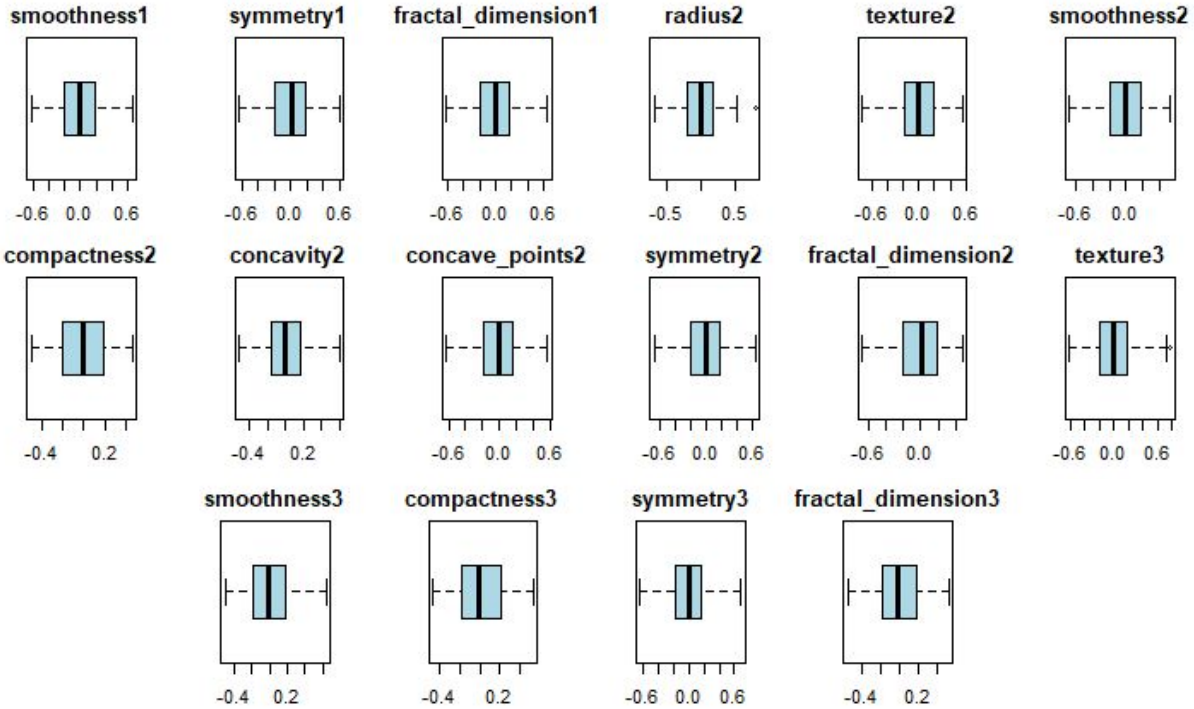
- It use useful to apply center and scale transformations before PCA so that predictors with a greater magnitude have similar influence on principal components to those with a smaller magnitude.
- Spatial Sign Transformation requires Center and Scale, each transformation was applied before our Spatial Sign Transformation.
- Each remaining predictor was centered and scaled

Spatial SIgn for Outliers Transformation

- All of our predictors still have outliers, we will run spatial sign transformation on all predictors

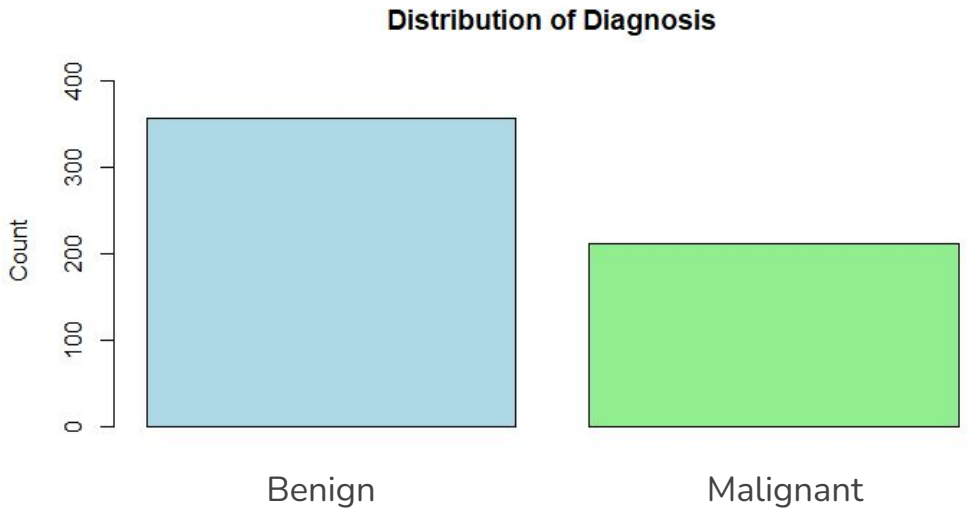


Spatial SIgn for Outliers Transformation - Before and After



Data splitting/Resampling

- We will split the data 80-20 stratified random sample since we have a smaller dataset and our distribution of the response variable is not uniform.
- We also may do 10-fold with 5 repeats cross validation for cross validation



Dataset After Preprocessing

- Number of Predictors: 30 for data with correlated predictors, 17 predictors for data that removed highly correlated predictors
- Sample Size: 569 total; 456 in the training set and 113 in the testing set

Modeling

- We have a dataset with highly correlated predictors and one with those predictors removed. All pre-processing steps are applied to both.
- We will be building classification models:
- Logistic, LDA, PLSDA, Penalized Model, QDA, RDA, MDA, Neural Network, FDA, SVM, KNN, Naive Bayes
- Our classification statistic of choice is ROC-AUC because we have an unbalanced binary outcome variable.
- Logistic Regression, LDA, QDA, RDA, MDA, FDA, and KNN models used our uncorrelated data
- PLSDA, Penalized Model, Neural Network, SVMDA, and Naive Bayes models used the correlated data

Logistic

```
> print(logistic_model)
Generalized Linear Model

456 samples
 17 predictor
  2 classes: 'B', 'M'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 411, 410, 411, 410, 410, 410, ...
Resampling results:

  ROC        Sens      Spec
  0.9812982  0.970665  0.9482353
```

LDA

Linear Discriminant Analysis

456 samples

17 predictor

2 classes: 'B', 'M'

No pre-processing

Resampling: Cross-Validated (10 fold, repeated 5 times)

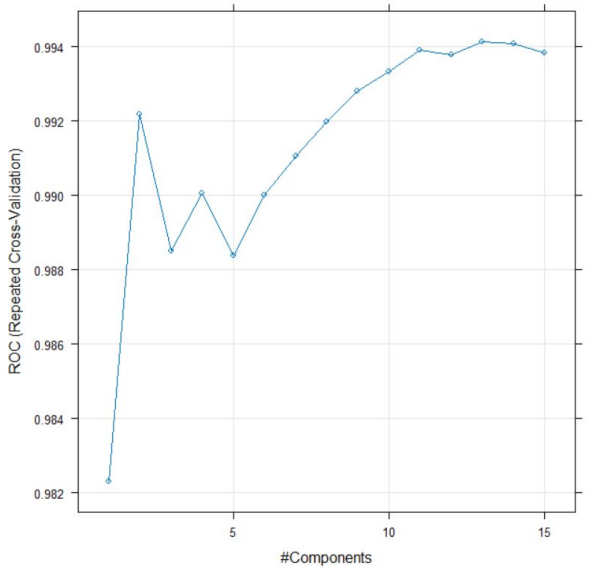
Summary of sample sizes: 410, 410, 411, 410, 411, 410, ...

Resampling results:

ROC	Sens	Spec
0.988738	0.9915517	0.8882353

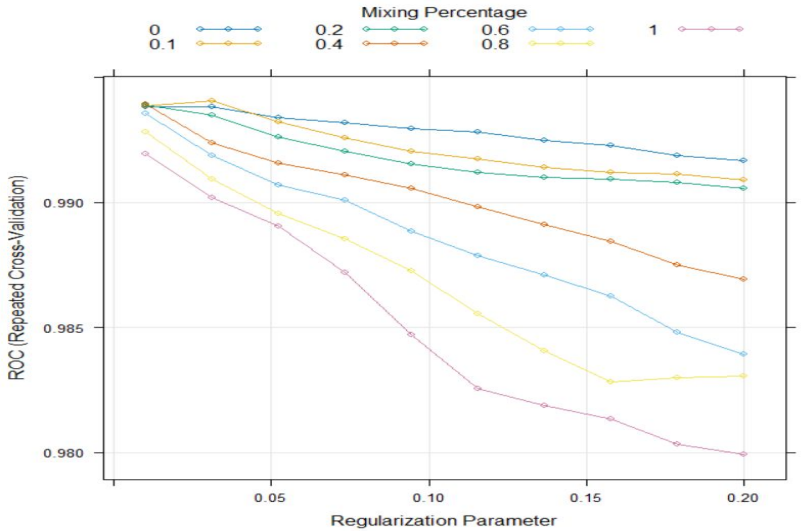
PLSDA

```
ncomp  ROC        Sens       Spec
 13    0.9941220  0.9965271  0.8952941
ROC was used to select the optimal model using the largest value.
The final value used for the model was ncomp = 13.
```



Penalized Model

```
 alpha   lambda       ROC          Sens       Spec
0.1     0.03111111   0.9940597   0.9950739   0.9270588
ROC was used to select the optimal model using the largest value.
The final values used for the model were alpha = 0.1 and lambda = 0.03111111.
```



QDA

Quadratic Discriminant Analysis

456 samples

17 predictor

2 classes: 'B', 'M'

No pre-processing

Resampling: Cross-Validated (10 fold, repeated 5 times)

Summary of sample sizes: 410, 410, 410, 410, 411, 410, ...

Resampling results:

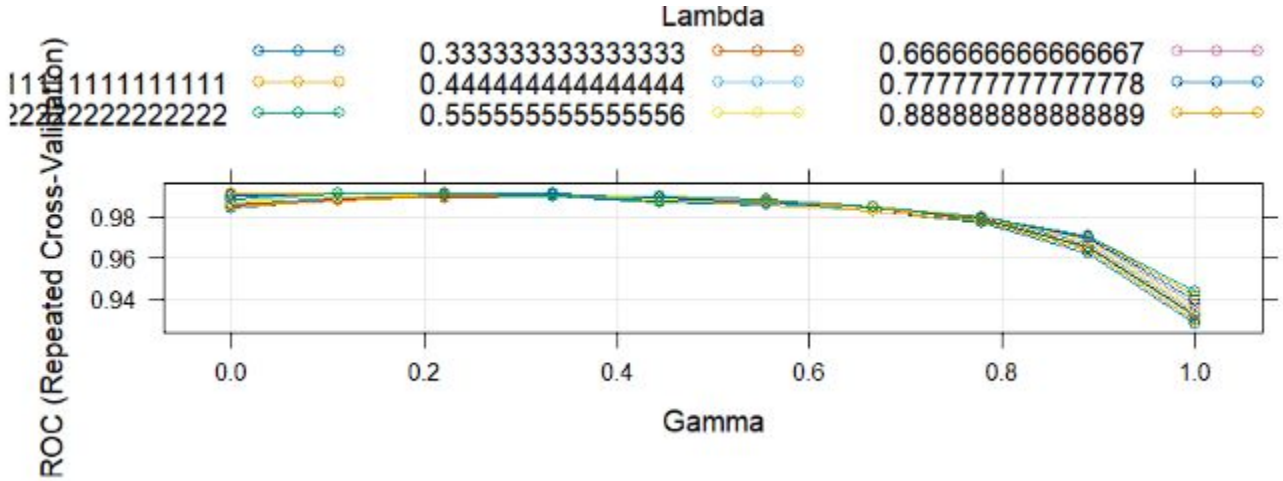
ROC	Sens	Spec
0.9846769	0.954064	0.9317647

RDA

gamma	lambda	ROC	Sens	Spec
0.1111111	0.8888889	0.9918111	0.9964778	0.8682353

ROC was used to select the optimal model using the largest value.

The final values used for the model were gamma = 0.1111111 and lambda = 0.8888889.

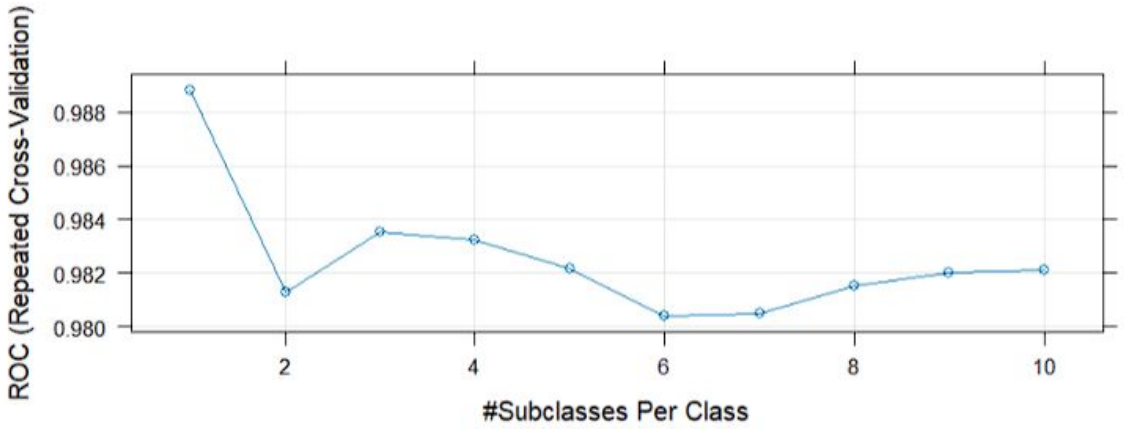


MDA

subclasses	ROC	Sens	Spec
1	0.9888293	0.9908867	0.8764706

ROC was used to select the optimal model using the largest value.

The final value used for the model was subclasses = 1.

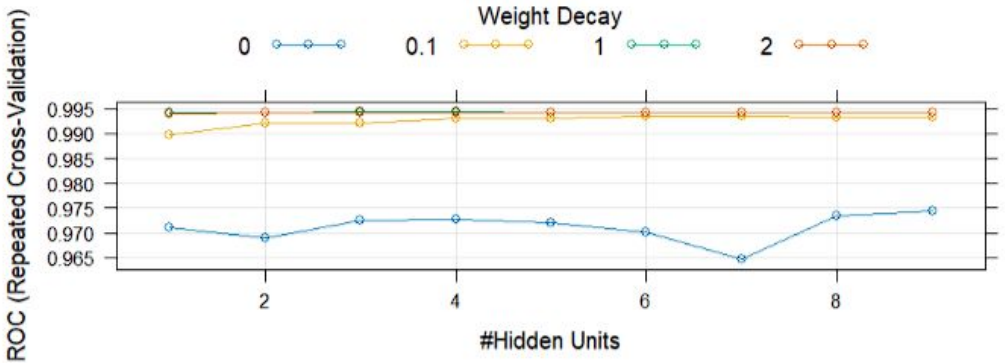


Neural Network

size	decay	ROC	Sens	Spec
3	1.0	0.9943524	0.9951232	0.9517647

ROC was used to select the optimal model using the largest value.

The final values used for the model were size = 3 and decay = 1.

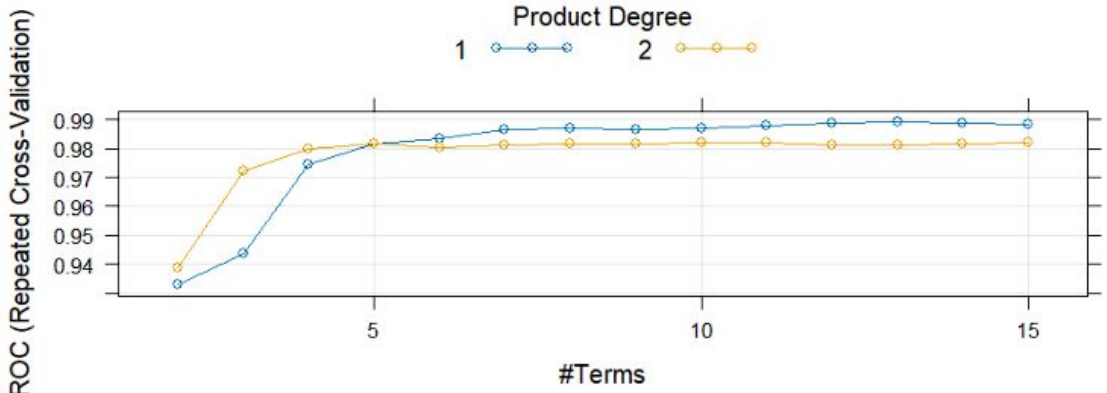


FDA

degree	nprune	ROC	Sens	Spec
1	13	0.9889728	0.9818473	0.9047059

ROC was used to select the optimal model using the largest value.

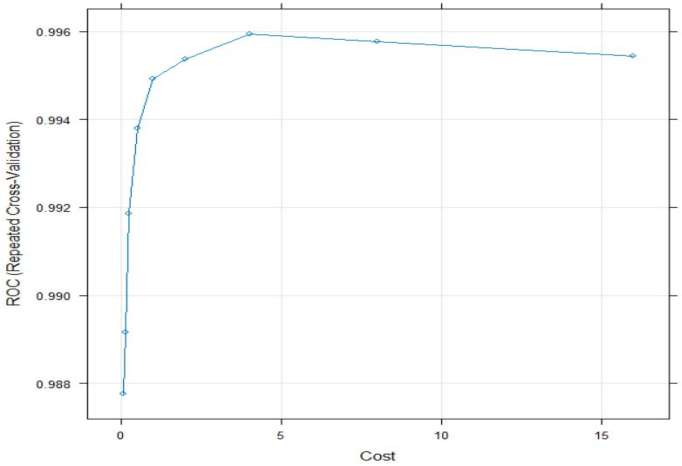
The final values used for the model were degree = 1 and nprune = 13.



SVMDA

C	ROC	Sens	Spec
4.0000	0.9959418	0.9804433	0.9552941

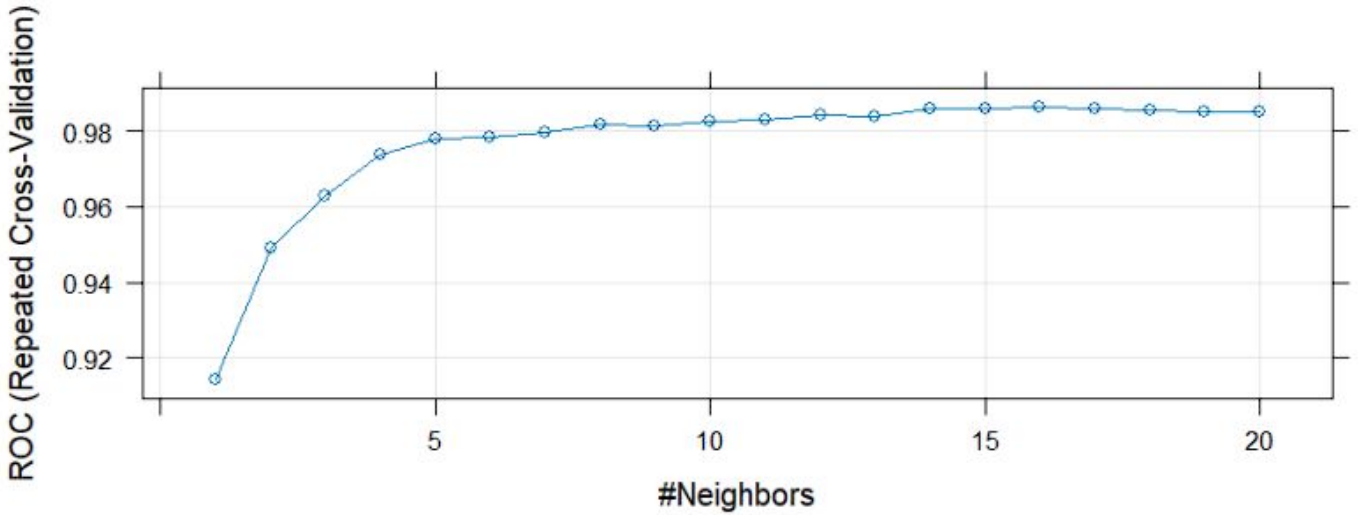
```
Tuning parameter 'sigma' was held constant at a value of 0.007675455
ROC was used to select the optimal model using the largest value.
The final values used for the model were sigma = 0.007675455 and C = 4.
```



KNN

k	ROC	Sens	Spec
16	0.9863315	0.9839409	0.8447059

ROC was used to select the optimal model using the largest value.



Naive Bayes

ROC	Sens	Spec
0.9845538	0.9441872	0.9223529

Tuning parameter 'fL' was held constant at a value of 2

Tuning parameter 'usekernel'

was held constant at a value of TRUE

Tuning parameter 'adjust' was held constant at

a value of TRUE

Model Summary Table - Linear

Model	Best Tuning Parameters	ROC-AUC
Logistic Regression	-	0.9812
LDA	-	0.9887
PLSDA	Ncomp = 13	0.9933
Penalized Models	Alpha = 0.1, lambda = 0.031	0.9941

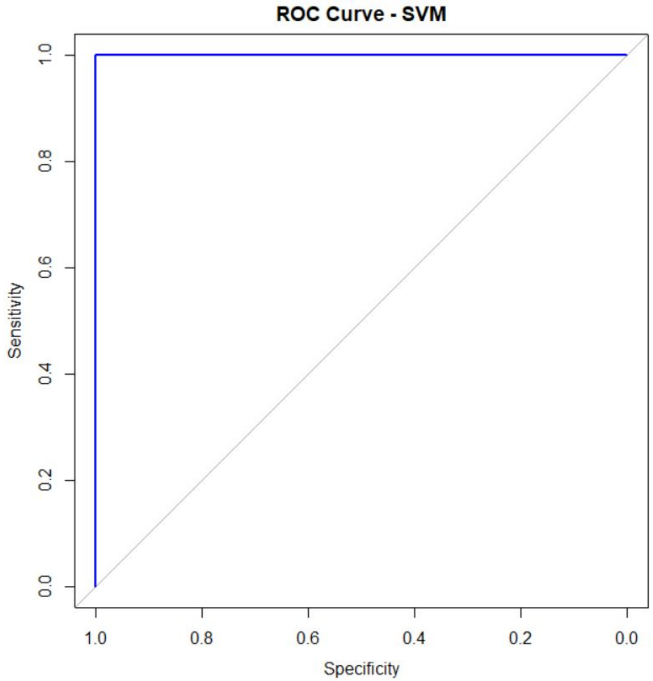
Model Summary Table - Non-Linear

Model	Tuning Parameters	ROC-AUC
QDA	-	0.9847
RDA	Gamma = 0.11, Lambda = 0.88	0.9918
MDA	Subclasses = 1	0.9888
Neural Network	Size = 3, Decay = 1	0.9943
FDA	Degree = 1, nprune = 13	0.9889
SVMDA	Sigma = 0.007675, C = 4	0.9959
KNN	k = 16	0.9863
Naive Bayes	fL = 2, useKernel = TRUE, Adjust =TRUE	0.9846

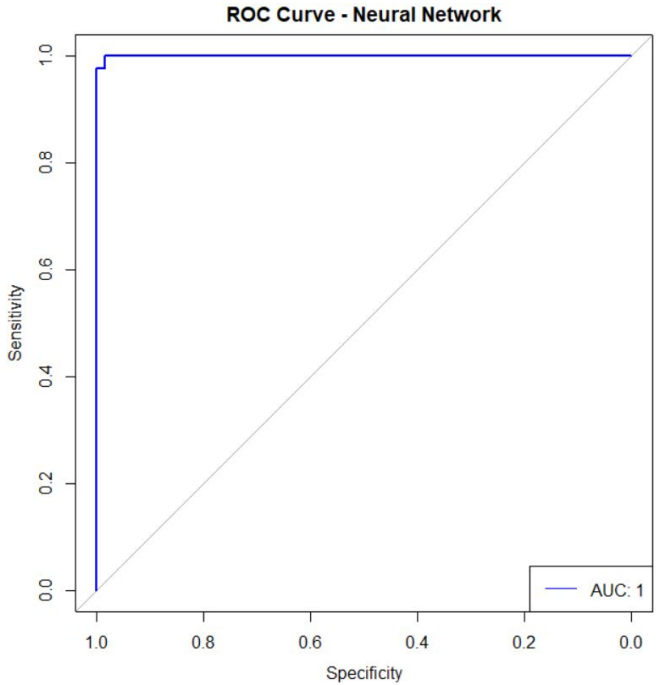
Best Models

- Although it was close, our two best models are the Neural Network with an ROC of 0.9943 and an SVM model with a ROC of 0.9959

Testing Performance



ROC-AUC: 1 (yes 1) for SVM



ROC-AUC: 0.9997 for NN

Confusion Matrix for SVM

	Reference		
Prediction		Benign	Malignant
	Benign	70	0
	Malignant	1	42

Most Important Predictors

Questions?