Deanna Springgay

Ch 12 Hw

12.1

a) To account for the classification imbalance in the response variable, I decided to split the data using stratified splitting using the response as the metric for splitting. This way, the training and testing sets both have roughly the same amount of each outcome: None, Mild, and Severe.

b) To optimize the models, I chose kappa optimization since the outcome has three levels. If the outcome had two levels instead, then I would have chosen ROC instead.

c)

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Tuning Parameter** | **Training Kappa** | **Testing Kappa** |
| Logistic Regression | decay = 0.1 | 0.04114121 | 0.01282051 |
| Linear Discriminant Analysis | N/A | 0.02941645 | 0.0661803 |
| Partial Least Squares | ncomp = 6 | 0.062481748 | 0.0804077 |
| Penalized Models | alpha = 1 and lambda = 0.05222222 | 0.0694473883 | -0.02083333 |
| Nearest Shrunken Centroids | threshold = 0.8 | 1.738690e-02 | -0.03157895 |

**Logistic Regression**

Penalized Multinomial Regression

225 samples

184 predictors

3 classes: 'Mild', 'None', 'Severe'

Pre-processing: nearest neighbor imputation (103), centered (103), scaled (103), remove (81)

Resampling: Repeated Train/Test Splits Estimated (25 reps, 75%)

Summary of sample sizes: 169, 169, 169, 169, 169, 169, ...

Resampling results across tuning parameters:

decay Accuracy Kappa

0e+00 0.3942857 0.03258549

1e-04 0.4007143 0.03457717

1e-01 0.4321429 0.04114121

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

The final value used for the model was decay = 0.1.

Confusion Matrix and Statistics

Reference

Prediction Mild None Severe

Mild 957 658 187

None 842 641 143

Severe 376 276 120

Overall Statistics

Accuracy : 0.409

95% CI : (0.3941, 0.4241)

No Information Rate : 0.5179

P-Value [Acc > NIR] : 1

Kappa : 0.0359

Mcnemar's Test P-Value : <2e-16

Statistics by Class:

Class: Mild Class: None Class: Severe

Sensitivity 0.4400 0.4070 0.26667

Specificity 0.5827 0.6248 0.82613

Pos Pred Value 0.5311 0.3942 0.15544

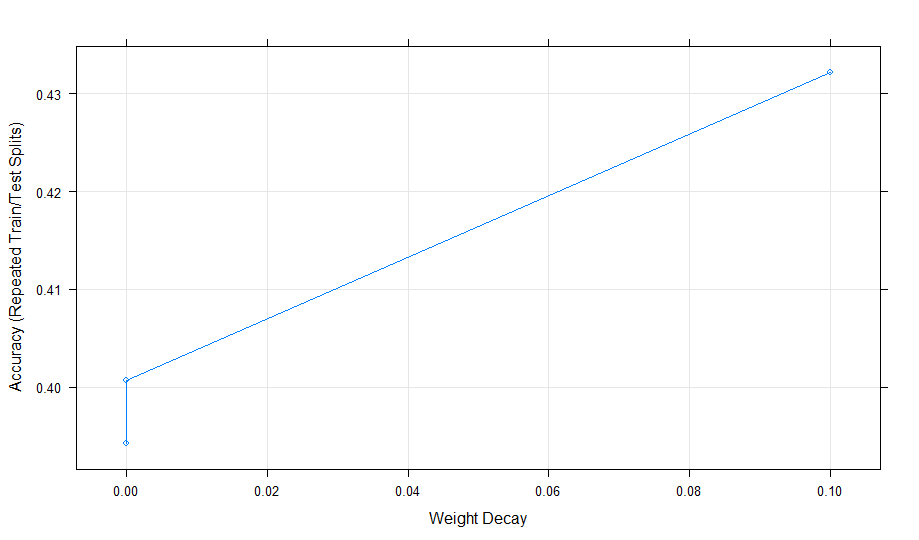
Neg Pred Value 0.4921 0.6371 0.90373

Prevalence 0.5179 0.3750 0.10714

Detection Rate 0.2279 0.1526 0.02857

Detection Prevalence 0.4290 0.3871 0.18381

Balanced Accuracy 0.5114 0.5159 0.54640



**Linear Discriminant Analysis**

Linear Discriminant Analysis

225 samples

184 predictors

3 classes: 'Mild', 'None', 'Severe'

Pre-processing: centered (103), scaled (103), remove (81)

Resampling: Repeated Train/Test Splits Estimated (25 reps, 75%)

Summary of sample sizes: 169, 169, 169, 169, 169, 169, ...

Resampling results:

Accuracy Kappa

0.4042857 0.02941645

Confusion Matrix and Statistics

Reference

Prediction Mild None Severe

Mild 334 245 60

None 255 183 41

Severe 136 97 49

Overall Statistics

Accuracy : 0.4043

95% CI : (0.3784, 0.4305)

No Information Rate : 0.5179

P-Value [Acc > NIR] : 1

Kappa : 0.0294

Mcnemar's Test P-Value : 2.468e-11

Statistics by Class:

Class: Mild Class: None Class: Severe

Sensitivity 0.4607 0.3486 0.3267

Specificity 0.5481 0.6617 0.8136

Pos Pred Value 0.5227 0.3820 0.1738

Neg Pred Value 0.4862 0.6287 0.9097

Prevalence 0.5179 0.3750 0.1071

Detection Rate 0.2386 0.1307 0.0350

Detection Prevalence 0.4564 0.3421 0.2014

Balanced Accuracy 0.5044 0.5051 0.5701

**Partial Least Squares**

Partial Least Squares

225 samples

184 predictors

3 classes: 'Mild', 'None', 'Severe'

Pre-processing: centered (103), scaled (103), remove (81)

Resampling: Bootstrapped (25 reps)

Summary of sample sizes: 225, 225, 225, 225, 225, 225, ...

Resampling results across tuning parameters:

ncomp Accuracy Kappa

1 0.4680020 -0.005727374

2 0.4710967 0.009402382

3 0.4836082 0.040974221

4 0.4763682 0.037091151

5 0.4801500 0.052370659

6 0.4838229 0.062481748

7 0.4779034 0.053590590

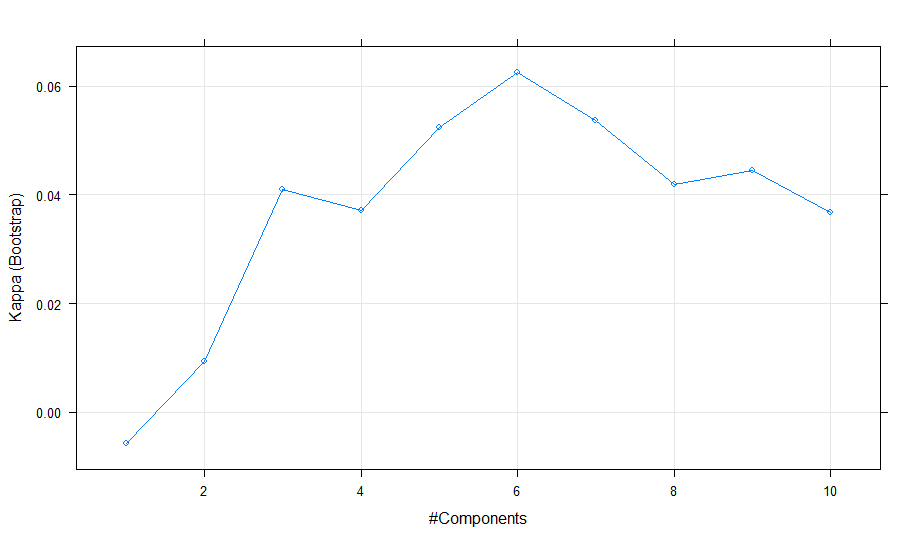
8 0.4688525 0.041922948

9 0.4676697 0.044545229

10 0.4604698 0.036783891

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

The final value used for the model was ncomp = 6.



**Penalized Models**

glmnet

225 samples

184 predictors

3 classes: 'Mild', 'None', 'Severe'

Pre-processing: centered (103), scaled (103), remove (81)

Resampling: Repeated Train/Test Splits Estimated (25 reps, 75%)

Summary of sample sizes: 169, 169, 169, 169, 169, 169, ...

Resampling results across tuning parameters:

alpha lambda Accuracy Kappa

0.0 0.01000000 0.4550000 0.0372774740

0.0 0.03111111 0.4600000 0.0269327541

0.0 0.05222222 0.4728571 0.0384805581

0.0 0.07333333 0.4807143 0.0453979731

0.0 0.09444444 0.4785714 0.0357633178

0.0 0.11555556 0.4764286 0.0269356525

0.0 0.13666667 0.4757143 0.0204405121

0.0 0.15777778 0.4750000 0.0153134505

0.0 0.17888889 0.4757143 0.0127327294

0.0 0.20000000 0.4742857 0.0065672329

0.1 0.01000000 0.4592857 0.0436415622

0.1 0.03111111 0.4728571 0.0389887240

0.1 0.05222222 0.4807143 0.0375897853

0.1 0.07333333 0.4742857 0.0148856897

0.1 0.09444444 0.4785714 0.0136493863

0.1 0.11555556 0.4735714 -0.0033136888

0.1 0.13666667 0.4785714 -0.0010996126

0.1 0.15777778 0.4785714 -0.0068942486

0.1 0.17888889 0.4828571 -0.0032650911

0.1 0.20000000 0.4864286 0.0003810371

0.2 0.01000000 0.4635714 0.0448779521

0.2 0.03111111 0.4721429 0.0300414134

0.2 0.05222222 0.4792857 0.0240736724

0.2 0.07333333 0.4807143 0.0126360454

0.2 0.09444444 0.4828571 0.0054982905

0.2 0.11555556 0.4857143 0.0038681373

0.2 0.13666667 0.4957143 0.0162770525

0.2 0.15777778 0.5057143 0.0285521820

0.2 0.17888889 0.5092857 0.0304389092

0.2 0.20000000 0.5100000 0.0277160527

0.4 0.01000000 0.4678571 0.0515835566

0.4 0.03111111 0.4800000 0.0314360842

0.4 0.05222222 0.4878571 0.0229225226

0.4 0.07333333 0.4950000 0.0207458548

0.4 0.09444444 0.5121429 0.0419595203

0.4 0.11555556 0.5185714 0.0496591828

0.4 0.13666667 0.5200000 0.0389705229

0.4 0.15777778 0.5185714 0.0260268770

0.4 0.17888889 0.5235714 0.0301647973

0.4 0.20000000 0.5292857 0.0349430460

0.6 0.01000000 0.4728571 0.0564752673

0.6 0.03111111 0.4921429 0.0391936024

0.6 0.05222222 0.4942857 0.0200888134

0.6 0.07333333 0.5178571 0.0514859197

0.6 0.09444444 0.5235714 0.0497781152

0.6 0.11555556 0.5214286 0.0316071892

0.6 0.13666667 0.5307143 0.0398648657

0.6 0.15777778 0.5264286 0.0243914983

0.6 0.17888889 0.5221429 0.0106819257

0.6 0.20000000 0.5185714 0.0016842105

0.8 0.01000000 0.4700000 0.0461432548

0.8 0.03111111 0.4964286 0.0412768395

0.8 0.05222222 0.5171429 0.0528767505

0.8 0.07333333 0.5285714 0.0600318436

0.8 0.09444444 0.5250000 0.0360026895

0.8 0.11555556 0.5285714 0.0319252410

0.8 0.13666667 0.5221429 0.0106819257

0.8 0.15777778 0.5185714 0.0016842105

0.8 0.17888889 0.5178571 0.0000000000

0.8 0.20000000 0.5178571 0.0000000000

1.0 0.01000000 0.4742857 0.0484065446

1.0 0.03111111 0.4978571 0.0323704171

1.0 0.05222222 0.5278571 0.0694473883

1.0 0.07333333 0.5292857 0.0501477959

1.0 0.09444444 0.5271429 0.0291839899

1.0 0.11555556 0.5207143 0.0067192064

1.0 0.13666667 0.5185714 0.0016842105

1.0 0.15777778 0.5178571 0.0000000000

1.0 0.17888889 0.5178571 0.0000000000

1.0 0.20000000 0.5178571 0.0000000000

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

The final values used for the model were alpha = 1 and lambda = 0.05222222.

Confusion Matrix and Statistics

Reference

Prediction Mild None Severe

Mild 40942 28646 7541

None 9372 7586 2712

Severe 436 518 247

Overall Statistics

Accuracy : 0.4977

95% CI : (0.4946, 0.5008)

No Information Rate : 0.5179

P-Value [Acc > NIR] : 1

Kappa : 0.0263

Mcnemar's Test P-Value : <2e-16

Statistics by Class:

Class: Mild Class: None Class: Severe

Sensitivity 0.8067 0.20642 0.02352

Specificity 0.2341 0.80271 0.98910

Pos Pred Value 0.5308 0.38566 0.20566

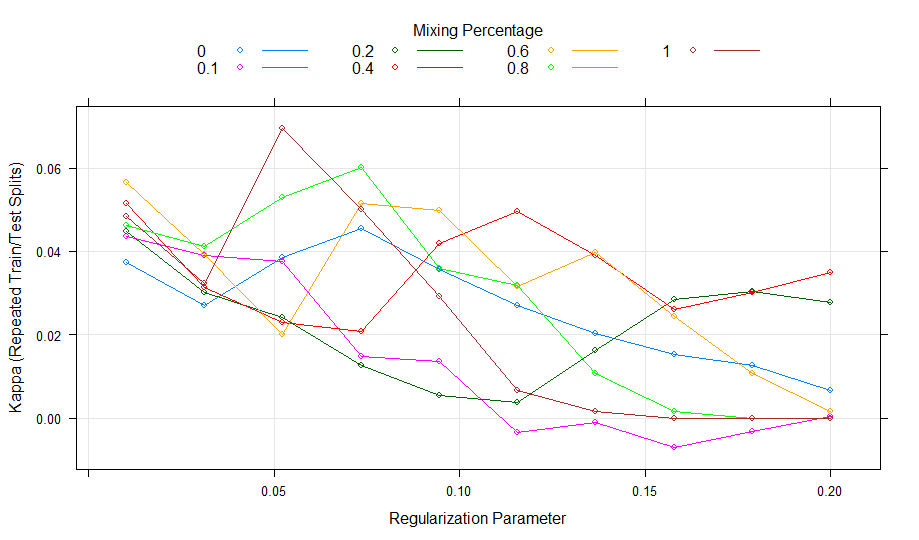
Neg Pred Value 0.5301 0.62768 0.89408

Prevalence 0.5179 0.37500 0.10714

Detection Rate 0.4178 0.07741 0.00252

Detection Prevalence 0.7870 0.20071 0.01226

Balanced Accuracy 0.5204 0.50457 0.50631



**Nearest Shrunken Centroids**

Nearest Shrunken Centroids

225 samples

184 predictors

3 classes: 'Mild', 'None', 'Severe'

Pre-processing: centered (103), scaled (103), remove (81)

Resampling: Bootstrapped (25 reps)

Summary of sample sizes: 225, 225, 225, 225, 225, 225, ...

Resampling results across tuning parameters:

threshold Accuracy Kappa

0.0 0.4648082 7.728634e-05

0.1 0.4638998 -3.681036e-03

0.2 0.4659905 -3.561967e-03

0.3 0.4695771 -1.175407e-04

0.4 0.4769335 9.382166e-03

0.5 0.4815659 1.417196e-02

0.6 0.4783845 3.843685e-03

0.7 0.4844042 1.110648e-02

0.8 0.4905210 1.738690e-02

0.9 0.4898652 1.160332e-02

1.0 0.4873859 3.263438e-03

1.1 0.4878010 1.715461e-03

1.2 0.4853646 -5.017835e-03

1.3 0.4853838 -6.452964e-03

1.4 0.4872393 -4.024197e-03

1.5 0.4897824 3.387077e-04

1.6 0.4896806 -1.228219e-03

1.7 0.4907068 -1.091829e-04

1.8 0.4902573 -1.595228e-03

1.9 0.4907573 -1.061967e-03

2.0 0.4907068 -1.643836e-03

2.1 0.4912068 -8.275862e-04

2.2 0.4912068 -8.275862e-04

2.3 0.4917068 0.000000e+00

2.4 0.4917068 0.000000e+00

2.5 0.4917068 0.000000e+00

2.6 0.4917068 0.000000e+00

2.7 0.4917068 0.000000e+00

2.8 0.4917068 0.000000e+00

2.9 0.4917068 0.000000e+00

3.0 0.4917068 0.000000e+00

3.1 0.4917068 0.000000e+00

3.2 0.4917068 0.000000e+00

3.3 0.4917068 0.000000e+00

3.4 0.4917068 0.000000e+00

3.5 0.4917068 0.000000e+00

3.6 0.4917068 0.000000e+00

3.7 0.4917068 0.000000e+00

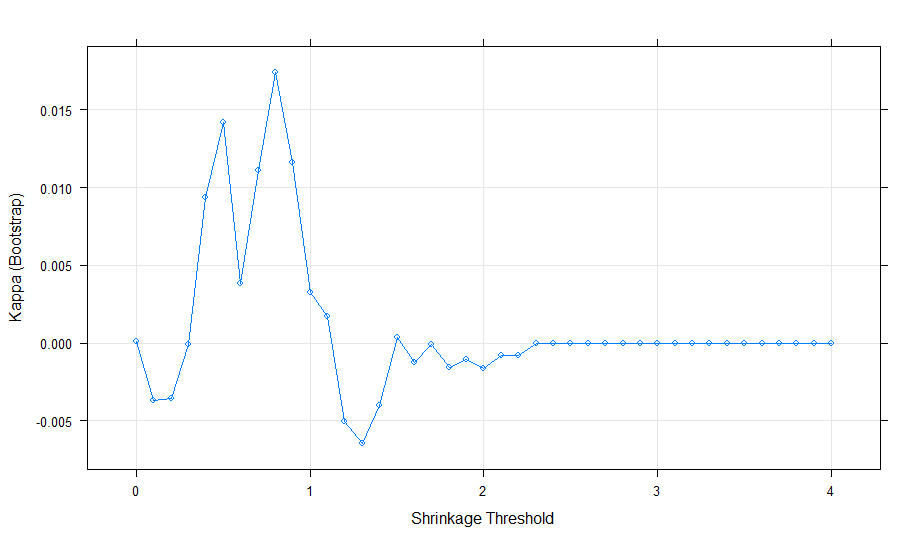
3.8 0.4917068 0.000000e+00

3.9 0.4917068 0.000000e+00

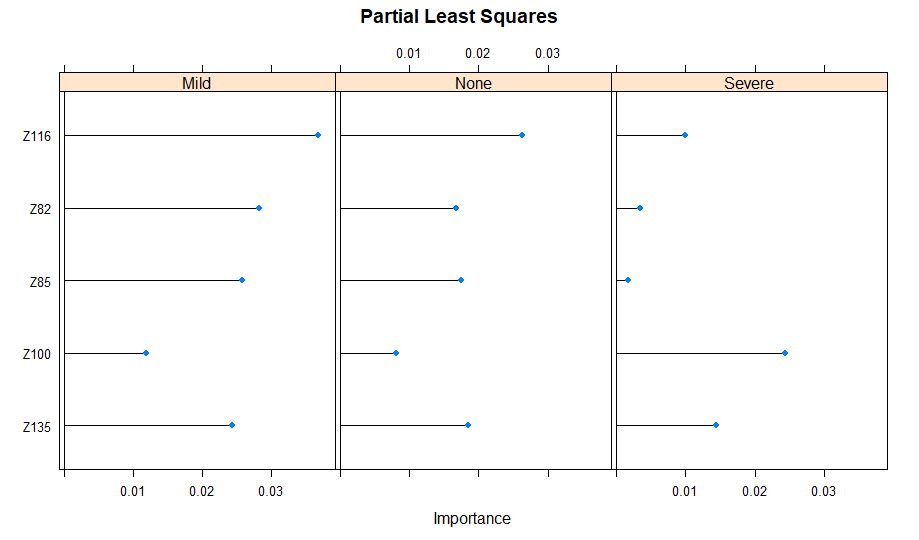
4.0 0.4917068 0.000000e+00

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

The final value used for the model was threshold = 0.8.



d) Based on the kappa values, the optimal model is partial least squares. The top 5 predictors for partial least squares are Z116, Z82, Z85, Z100, and Z135.



12.3

a)