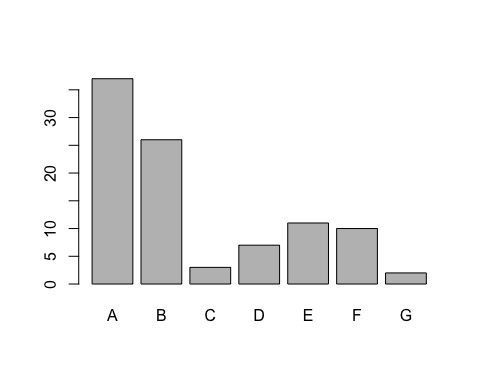
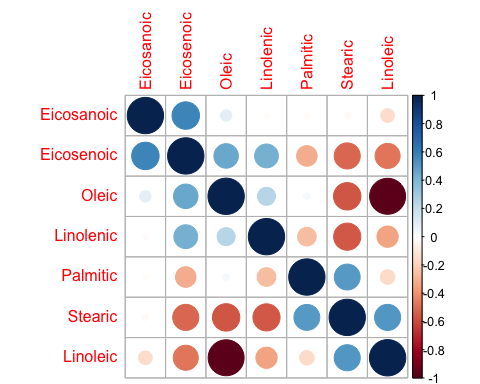
2.a. Becuase the extream class imbalance in the data set, it should be split using stratified sampling.



There are some high correlated predictors which was removed before modeling.



1. Because the class imbalance kappa statistic should be used.
2. Only LDA, PLSDA, NSC methods are used but logistic regression in not used beucase the number of classes is higher than 2.

**Linear Discriminant Analysis**   
   
 74 samples  
 6 predictor  
 7 classes: 'A', 'B', 'C', 'D', 'E', 'F', 'G'   
   
 Pre-processing: centered, scaled   
 Resampling: Bootstrapped(25Rep)  
   
 Summary of sample sizes: 73, 73, 73, 73, 73, 73, ...   
   
 Resampling results across tuning parameters:  
   
 dimen Accuracy Kappa   
 1 0.7837838 0.7157263  
 2 0.9189189 0.8923897  
 3 0.9324324 0.9101506  
 4 0.9459459 0.9280506  
 5 0.9459459 0.9280506  
 6 0.9459459 0.9280506  
   
 Kappa was used to select the optimal model using the largest value.  
 The final value used for the model was dimen = 4.

Confusion Matrix and Statistics for Testing set  
   
 Reference  
 Prediction A B C D E F G  
 A 9 0 0 0 0 0 0  
 B 0 6 0 0 0 0 0  
 C 0 0 1 0 0 0 0  
 D 0 0 0 1 0 0 0  
 E 0 0 0 0 2 0 0  
 F 0 0 0 0 0 2 0  
 G 0 0 0 0 0 0 1  
   
 Overall Statistics  
   
 Accuracy : 1   
 95% CI : (0.8456, 1)  
   
 Kappa : 1   
   
 Statistics by Class:  
   
 Class: A Class: B Class: C Class: D Class: E Class: F  
 Sensitivity 1.0000 1.0000 1.00000 1.00000 1.00000 1.00000  
 Specificity 1.0000 1.0000 1.00000 1.00000 1.00000 1.00000  
   
 Class: G  
 Sensitivity 1.00000  
 Specificity 1.00000

**Partial Least Squares**   
   
 74 samples  
 6 predictor  
 7 classes: 'A', 'B', 'C', 'D', 'E', 'F', 'G'   
   
 Pre-processing: centered, scaled   
 Resampling: Bootstrapped (25 reps)   
   
 Summary of sample sizes: 74, 74, 74, 74, 74, 74, ...   
   
 Resampling results across tuning parameters:  
   
 ncomp Accuracy Kappa Accuracy SD Kappa SD   
 1 0.4824634 0.2483353 0.11260882 0.1440213  
 2 0.6949120 0.5598613 0.07092248 0.0890928  
 3 0.7558803 0.6502042 0.07836378 0.1072279  
   
 Kappa was used to select the optimal model using the largest value.  
 The final value used for the model was ncomp = 3.

Confusion Matrix and Statistics for Testing set  
   
 Reference  
 Prediction A B C D E F G  
 A 9 1 0 1 0 0 0  
 B 0 5 0 0 0 0 1  
 C 0 0 0 0 0 0 0  
 D 0 0 1 0 0 0 0  
 E 0 0 0 0 2 0 0  
 F 0 0 0 0 0 2 0  
 G 0 0 0 0 0 0 0  
   
 Overall Statistics  
   
 Accuracy : 0.8182   
 Kappa : 0.7412   
   
 Statistics by Class:  
   
 Class: A Class: B Class: C Class: D Class: E Class: F  
 Sensitivity 1.0000 0.8333 0.00000 0.00000 1.00000 1.00000  
 Specificity 0.8462 0.9375 1.00000 0.95238 1.00000 1.00000  
Class: G  
 Sensitivity 0.00000  
 Specificity 1.00000

**Penalized Models**

**glmnet**   
   
 74 samples  
 6 predictor  
 7 classes: 'A', 'B', 'C', 'D', 'E', 'F', 'G'   
   
 Pre-processing: centered, scaled   
 Resampling: Bootstrapped (25 reps)   
   
 Summary of sample sizes: 74, 74, 74, 74, 74, 74, ...   
   
 Resampling results across tuning parameters:  
   
 alpha lambda Accuracy Kappa Accuracy SD Kappa SD   
 0.8 0.02461538 0.9628307 0.9479006 0.03554487 0.05045771  
 0.8 0.02948718 0.9664021 0.9527341 0.03345280 0.04779065  
 0.8 0.03435897 0.9622354 0.9465962 0.04187140 0.06085534  
 0.8 0.03923077 0.9622354 0.9465962 0.04187140 0.06085534  
 0.8 0.04410256 0.9622354 0.9465962 0.04187140 0.06085534  
 0.8 0.04897436 0.9622354 0.9465962 0.04187140 0.06085534  
 0.8 0.05384615 0.9546640 0.9362659 0.04070703 0.05890872  
 0.8 0.05871795 0.9546640 0.9363374 0.04070703 0.05892101  
 0.8 0.06358974 0.9467937 0.9255200 0.03377338 0.04969885  
Kappa was used to select the optimal model using the largest value.  
 The final values used for the model were alpha = 0.8 and lambda  
 = 0.02948718.

Confusion Matrix and Statistics for Testing set  
   
 Reference  
 Prediction A B C D E F G  
 A 9 0 0 0 0 0 0  
 B 0 5 0 0 0 0 1  
 C 0 1 1 0 0 0 0  
 D 0 0 0 1 0 0 0  
 E 0 0 0 0 2 0 0  
 F 0 0 0 0 0 2 0  
 G 0 0 0 0 0 0 0  
   
 Overall Statistics  
   
 Accuracy : 0.9091   
 Kappa : 0.8764   
   
 Statistics by Class:   
 Class: A Class: B Class: C Class: D Class: E Class: F  
 Sensitivity 1.0000 0.8333 1.00000 1.00000 1.00000 1.00000  
 Specificity 1.0000 0.9375 0.95238 1.00000 1.00000 1.00000  
Class: G  
 Sensitivity 0.00000  
 Specificity 1.00000

**Nearest Shrunken Centroids**   
   
 74 samples  
 6 predictor  
 7 classes: 'A', 'B', 'C', 'D', 'E', 'F', 'G'   
   
 Pre-processing: centered, scaled   
 Resampling: Bootstrapped (25 reps)   
   
 Summary of sample sizes: 74, 74, 74, 74, 74, 74, ...   
   
 Resampling results across tuning parameters:  
   
 threshold Accuracy Kappa Accuracy SD Kappa SD   
 0 0.9259083 0.898462747 0.03750060 0.05065888  
 1 0.9250931 0.897261322 0.03895470 0.05247624  
 2 0.9238237 0.895509485 0.03805330 0.05152536  
 3 0.9096220 0.876030832 0.04177344 0.05669234  
 4 0.8862367 0.842294179 0.07435357 0.10488807  
 5 0.8386729 0.776044152 0.09703202 0.13630170  
 6 0.7669280 0.677209633 0.09250907 0.12380399  
   
 Accuracy was used to select the optimal model using the largest value.  
 The final value used for the model was threshold = 0.

Confusion Matrix and Statistics for Testing set  
   
 Reference  
 Prediction A B C D E F G  
 A 9 0 0 0 0 0 0  
 B 0 5 0 0 0 0 0  
 C 0 1 1 0 0 0 0  
 D 0 0 0 1 0 0 0  
 E 0 0 0 0 2 0 0  
 F 0 0 0 0 0 2 0  
 G 0 0 0 0 0 0 1  
   
 Overall Statistics  
   
 Accuracy : 0.9545   
 Kappa : 0.9391   
   
 Statistics by Class:  
   
 Class: A Class: B Class: C Class: D Class: E Class: F  
 Sensitivity 1.0000 0.8333 1.00000 1.00000 1.00000 1.00000  
 Specificity 1.0000 1.0000 0.95238 1.00000 1.00000 1.00000  
Class: G  
 Sensitivity 1.00000  
 Specificity 1.00000

|  |  |  |
| --- | --- | --- |
| FOR TESTING SET | Accuracy | Kappa |
| LDA | 1 | 1 |
| PLSDA | 0.8182 | 0.7413 |
| Penalised Models | 0.9091 | 0.8764 |
| NSC | 0.9545 | 0.9391 |

LDA is the best model as it classify 100% correctly.

When comparing all models Category G has higher possibility to be wrongly classified. Category A has highest accuracy. This is to be expected because of the imbalance in distribution of classes.