```
## Warning: package 'caret' was built under R version 3.5.2
## randomForest 4.6-14
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
## The following object is masked from 'package:ggplot2':
##
##     margin
## Warning: package 'rpart' was built under R version 3.5.2
```

#Data Processing

#Data Cleaning

#Remove variables with near zero variance

#Remove columns that are not predictors, which are the the seven first columns

#Cross-validation In order to get out-of-sample errors, split the training data in training (75%) and testing (25%) data) subsets:

```
## [1] 14718    53
## [1] 4904   53
```

#Prediction

#DECISION TREE #Fit model on Neo-Training data

#Use model to predict class in validation set (NEOTesting)

#Estimate the errors of the prediction algorithm in the Decision Tree model

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1288  139    9   56   15
##          B   35  511   51   60   58
##          C   35  137  692  102  108
##          D   13   75   53  520   48
##          E   24   87   50   66  672
##
## Overall Statistics
##
##                Accuracy : 0.751
##                  95% CI : (0.7387, 0.7631)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6844
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
```

```
## Sensitivity            0.9233   0.5385   0.8094   0.6468   0.7458
## Specificity            0.9376   0.9484   0.9057   0.9539   0.9433
## Pos Pred Value         0.8547   0.7147   0.6443   0.7334   0.7475
## Neg Pred Value         0.9685   0.8954   0.9574   0.9323   0.9428
## Prevalence             0.2845   0.1935   0.1743   0.1639   0.1837
## Detection Rate         0.2626   0.1042   0.1411   0.1060   0.1370
## Detection Prevalence   0.3073   0.1458   0.2190   0.1446   0.1833
## Balanced Accuracy      0.9304   0.7434   0.8575   0.8003   0.8446
```

#RANDOM FOREST #Fit model on NEOTraining data

#Use model to predict class in validation set (NEOTesting)

#Estimate the errors of the prediction algorithm in the Random Forest

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1395    0    0    0    0
##          B    3  944    2    0    0
##          C    0    8  847    0    0
##          D    0    0    6  797    1
##          E    0    0    0    1  900
##
## Overall Statistics
##
##                Accuracy : 0.9957
##                  95% CI : (0.9935, 0.9973)
##     No Information Rate : 0.2851
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9946
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9979   0.9916   0.9906   0.9987   0.9989
## Specificity            1.0000   0.9987   0.9980   0.9983   0.9998
## Pos Pred Value         1.0000   0.9947   0.9906   0.9913   0.9989
## Neg Pred Value         0.9991   0.9980   0.9980   0.9998   0.9998
## Prevalence             0.2851   0.1941   0.1743   0.1627   0.1837
## Detection Rate         0.2845   0.1925   0.1727   0.1625   0.1835
## Detection Prevalence   0.2845   0.1935   0.1743   0.1639   0.1837
## Balanced Accuracy      0.9989   0.9952   0.9943   0.9985   0.9993
```

#TEST THE MODEL TO PREDICT 20 DIFFERENT TEST CASES #Perform prediction

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  B  A  B  A  A  E  D  B  A  A  B  C  B  A  E  E  A  B  B  B
## Levels: A B C D E
```