Practical Machine Learning: Course Project

The raw data set is kind of messy because many predictor values are missing, NA or not meaningful, which needs some cleaning process before really building up the learning algorithm. For this part, I did following three steps:

1. Remove all near-zero variables. They are not crucial predictors in the learning process.

2. Remove variables containing more than 95% NA values. Their data volume is not enough to achieve a convincing learning results.

3. Remove all the first 5 columns, which are purely identification variables and exclusive of learning information.

Here, I partitioned the data set into training and test set as the normal way in class and chose two modeling methods (random forest and generalized boosted model) to code the learning process. It turned out that random forest method showed higher accuracy (see below), which is desired in this project. Well, RF method did consume some time in my computer to run the entire algorithm.

Random Forest

```
Confusion Matrix and Statistics

          Reference
Prediction    A    B    C    D    E
         A 1674    1    0    0    0
         B    0 1134    1    0    0
         C    0    4 1025    3    0
         D    0    0    0  961    1
         E    0    0    0    0 1081

Overall Statistics

               Accuracy : 0.9983
                 95% CI : (0.9969, 0.9992)
    No Information Rate : 0.2845
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9979
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: A Class: B Class: C Class: D Class: E
Sensitivity            1.0000   0.9956   0.9990   0.9969   0.9991
Specificity            0.9998   0.9998   0.9986   0.9998   1.0000
Pos Pred Value         0.9994   0.9991   0.9932   0.9990   1.0000
Neg Pred Value         1.0000   0.9989   0.9998   0.9994   0.9998
Prevalence             0.2845   0.1935   0.1743   0.1638   0.1839
Detection Rate         0.2845   0.1927   0.1742   0.1633   0.1837
Detection Prevalence   0.2846   0.1929   0.1754   0.1635   0.1837
Balanced Accuracy      0.9999   0.9977   0.9988   0.9983   0.9995
```

GBM

```
Confusion Matrix and Statistics

          Reference
Prediction    A    B    C    D    E
         A 1669    6    0    0    1
         B    5 1116   10    3    5
         C    0   15 1009   14    1
         D    0    2    6  947   11
         E    0    0    1    0 1064

Overall Statistics

               Accuracy : 0.9864
                 95% CI : (0.9831, 0.9892)
    No Information Rate : 0.2845
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9828
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: A Class: B Class: C Class: D Class: E
Sensitivity            0.9970   0.9798   0.9834   0.9824   0.9834
Specificity            0.9983   0.9952   0.9938   0.9961   0.9998
Pos Pred Value         0.9958   0.9798   0.9711   0.9803   0.9991
Neg Pred Value         0.9988   0.9952   0.9965   0.9965   0.9963
Prevalence             0.2845   0.1935   0.1743   0.1638   0.1839
Detection Rate         0.2836   0.1896   0.1715   0.1609   0.1808
Detection Prevalence   0.2848   0.1935   0.1766   0.1641   0.1810
Balanced Accuracy      0.9977   0.9875   0.9886   0.9893   0.9916
```

Random Forest: 0.9983

Generalized Boosted Model: 0.9864