Practical Machine Learning Write Up

Reitwiec Shandilya

October 20, 2020

```
dim(training_data)
```

```
## [1] 19622   160
```

```
dim(testing_data)
```

```
## [1]  20 160
```

Data Cleaning

The original training and testing data has 160 variables.

After removed columns with NA entries and near zero value variable's , the number of variables was brought down to 59.

We then removed 7 additional variables which contained information that were deemed not useful: X , User_names, raw_timestamp_part_1, raw_timestamp_part_2, cvtd_timestamp, new_window, num_window. (Prior to removing these variables, we were achieving perfect accuracy on our training and validation sets, but my model was predicting all of the test cases to be of class A.)

```
dim(train_data)
```

```
## [1] 19622    52
```

```
dim(test_data)
```

```
## [1] 20 52
```

Data Partition

From here, we split the training data into two sets: “train_data” for training the model (60%) and “test_data” for testing of the model (40%).

We trained a random forest on “train_data” using the default parameters. We chose Decision Tree model to train the model and to analyze the accuracy of it. Then we chose a Random forest model because they tend to be very accurate and the data set was small enough that using a random forest was feasible.

We predicted the classes on “train_data” and found that the accuracy was 100%. We then used this model to predict the values on the “test_data” set and found the accuracy to be 98.9%.

```
##  [1] 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
```

Like this we get best level of accuracy for data and submitted my answers, and it correctly identified all 20 cases.