Report

October 24, 2014

For this project we are using the

First Load in the required packages

```
require(caret)
```

```
## Loading required package: caret
## Loading required package: lattice
## Loading required package: ggplot2
```

```
require(ggplot2)
require(randomForest)
```

```
## Loading required package: randomForest
## randomForest 4.6-10
## Type rfNews() to see new features/changes/bug fixes.
```

Read in the Training and Test Set.

```
training_URL<-"http://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
test_URL<-"http://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"
training<-read.csv(training_URL,na.strings=c("NA",""))
test<-read.csv(test_URL,na.strings=c("NA",""))
```

Now get rid of the columns that is simply an index, timestamp or username.

```
training<-training[,7:160]
test<-test[,7:160]
```

Remove the columns that are mostly NAs. They could be useful in the model, but it is easier to cut the data.frame down and see if it gives good results

```
mostly_data<-apply(!is.na(training),2,sum)>19621
training<-training[,mostly_data]
test<-test[,mostly_data]
dim(training)
```

```
## [1] 19622    54
```