**MACHINE LEARNING ASSIGNMENT 4**

**Loading and Prep**

In this assignment the first step was to load the libraries we’ll need: caret, randomforest and corrplot , set the active directory, and load the training and testing data from PLM, is loaded and stored as matrix variables to enable manipulation.

**Data Cleaning**

Once loaded, the csv data itself contains some text fields, a date field, many NAs and columns that contain mainly blank values. Without removing these, the functions in R fail to perform, so some data cleaning is required. The first 6 columns are removed. And the NAs are removed.

**Training and Testing**

With a cleaned Train Dataset, it is then partitioned into 2 sets (train and test) to enable cross validation. At this point a correlation matrix can be plotted to see whether some variables have a high degree of correlation with the others

**Prediction Model**

Based on the course lectures and online research regarding Kaggle competitions, random forest consistently proves to be a highly accurate, though computationally intensive, approach to building prediction models. Thus for this assignment, a random forest approach was used to build the model.

**Cross Validation**

The train function in Caret handles cross validation. By separating the train data set into 2 equal sized half-sets, the accuracy of the model can be assessed. In this case, the random forest model yielded an accuracy of 99.46%, and a confusion matrix broken down by class type shows a relatively small proportional distribution of error.

Confusion Matrix

Reference

Prediction A B C D E

A 2790 17 0 0 0

B 0 1878 3 0 0

C 0 2 1708 21 1

D 0 1 0 1587 8

E 0 0 0 0 1794

Based on the results of the cross validation confusion matrix, which considered an equal quantity of new test data and training data, it is expected the out of sample error will be less than 1% for predictions based on new data.

**Predictions**

Based on the random forest modelling, the predictions of the classe of exercises for the test data are:

B A B A A E D B A A B C B A E E A B B B

**CODE**

setwd("C:/Users/Azamat/Documents/R/Machine Learning 4")

#libraries

library(caret)

library(randomForest)

library(corrplot)

#load data

TestA = read.csv("pml-testing.csv", na.strings = c("NA", "#DIV/0!", ""))

TrainA = read.csv("pml-training.csv",na.strings = c("NA", "#DIV/0!", ""))

#remove non numeric columns

TrainA <- TrainA[, -(1:6)]

#clean NAs or mostly NA columns

NAs <- sapply(TrainA, function(x) mean(is.na(x))) > 0.95

TrainA <- TrainA[, NAs==FALSE]

#seperate train and test set

inTrain <- createDataPartition(TrainA$classe, p=0.5, list=FALSE)

TrainSet <- TrainA[inTrain, ]

TestSet <- TrainA[-inTrain, ]

#plot correlation ofvariables

corMatrix <- cor(TrainSet[, -54])

corrplot(corMatrix, order = "FPC", method = "square", type = "upper",

add = FALSE, tl.cex = 0.5)

#build a prediction model

# model fit

set.seed(150)

controlRF <- trainControl(method="cv", number=3, verboseIter=FALSE)

modFit <- train(classe ~ ., data=TrainSet, method="rf", trControl=controlRF, prox = TRUE)

modFit$finalModel

#prediction - cross validation

predictRF <- predict(modFit, newdata=TestSet)

confMatRF <- confusionMatrix(predictRF, TestSet$classe)

confMatRF

#prediction

Testcases <- predict(modFit, newdata=TestA)

Testcases