Data 624: Homework II

Group Two

Vinicio Haro
Sang Yoon (Andy) Hwang
Julian McEachern
Jeremy O'Brien
Bethany Poulin

16 December 2019

Contents

Getting Started

Overview

Include details on our process in creating this document.

Dependencies

```
# Predictive Modeling
libraries('AppliedPredictiveModeling', 'mlbench', 'caret', 'randomForest')

# Formatting Libraries
libraries('default', 'knitr', 'kableExtra')

# Plotting Libraries
libraries('ggplot2', 'grid', 'ggfortify')
```

Assignment 1

Kuhn and Johnson 6.3

A chemical manufacturing process for a pharmaceutical product was discussed in Sect.1.4. In this problem, the objective is to understand the relationship between biological measurements of the raw materials (predictors), measurements of the manufacturing process (predictors), and the response of product yield. Biological predictors cannot be changed but can be used to assess the quality of the raw material before processing. On the other hand, manufacturing process predictors can be changed in the manufacturing process. Improving product yield by 1% will boost revenue by approximately one hundred thousand dollars per batch:

(a). Start R and use these commands to load the data:

```
data("ChemicalManufacturingProcess")
```

(b). A small percentage of cells in the predictor set contain missing values. Use an imputation function to fill in these missing values (e.g., see Sect. 3.8).

```
#code
```

(c). Split the data into a training and a test set, pre-process the data, and tune a model of your choice from this chapter. What is the optimal value of the performance metric?

```
#code
```

(d). Predict the response for the test set. What is the value of the performance metric and how does this compare with the resampled performance metric on the training set?

```
#code
```

(e). Which predictors are most important in the model you have trained? Do either the biological or process predictors dominate the list?

```
#code
```

(f). Explore the relationships between each of the top predictors and the response. How could this information be helpful in improving yield in future runs of the manufacturing process?

```
#code
```

Assignment 2

Kuhn and Johnson 7.2

Friedman (1991) introduced several benchmark data sets create by simulation. One of these simulations used the following nonlinear equation to create data: $y = 10\text{sin}(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + N(0, \sigma^2)$; where the x values are random variables uniformly distributed between $[0, 1]$ (there are also 5 other non-informative variables also created in the simulation).

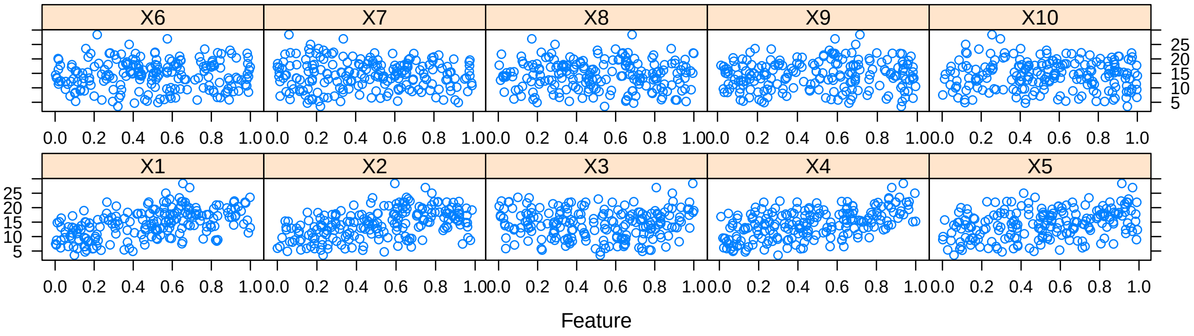
```
# The package `mlbench` contains a function called `mlbench.friedman1`
# that simulates these data:

set.seed(200)
trainingData <- mlbench.friedman1(200, sd = 1)

## We convert the 'x' data from a matrix to a data frame
## One reason is that this will give the columns names.

trainingData$x <- data.frame(trainingData$x)

## Look at the data using
featurePlot(trainingData$x, trainingData$y)
```



```
## or other methods.

## This creates a list with a vector 'y' and a matrix
## of predictors 'x'. Also simulate a large test set to
## estimate the true error rate with good precision:

testData <- mlbench.friedman1(5000, sd = 1)
testData$x <- data.frame(testData$x)
```

(a). Tune several models on these data. For example:

```
knnModel <- train(x = trainingData$x,
                  y = trainingData$y,
                  method = "knn",
                  preProc = c("center", "scale"),
                  tuneLength = 10)
knnModel

knnPred <- predict(knnModel, newdata = testData$x)

## The function 'postResample' can be used to get the test set performance values
postResample(pred = knnPred, obs = testData$y)
```

Model 1:

```
#code
```

Model 2:

```
#code
```

Model 3:

```
#code
```

(b). **Which models appear to give the best performance? Does MARS select the informative predictors (those named X1-X5)?**

```
#code
```

Kuhn and Johnson 7.5

Exercise 6.3 describes data for a chemical manufacturing process. Use the same data imputation, data splitting, and pre-processing steps as before and train several nonlinear regression models.

```
# Call code from 6.3
```

(a). **Which nonlinear regression model gives the optimal resampling and test set performance?**

```
# code
```

(b). **Which predictors are most important in the optimal nonlinear regression model? Do either the biological or process variables dominate the list? How do the top ten important predictors compare to the top ten predictors from the optimal linear model?**

```
# code
```

(c). Explore the relationships between the top predictors and the response for the predictors that are unique to the optimal nonlinear regression model. Do these plots reveal intuition about the biological or process predictors and their relationship with yield?

```
# code
```

Assignment 3

Kuhn and Johnson 8.1

Recreate the simulated data from Exercise 7.2:

```r
set.seed(200)
simulated <- mlbench.friedman1(200, sd = 1)
simulated <- cbind(simulated$x, simulated$y)
simulated <- as.data.frame(simulated)
colnames(simulated)[ncol(simulated)] <- "y"
```

(a). Fit a random forest model to all of the predictors, then estimate the variable importance scores. Did the random forest model significantly use the uninformative predictors (V6-V10)?

```r
model1 <- randomForest(y ~ ., data = simulated,
                       importance = TRUE,
                       ntree = 1000)
rfImp1 <- varImp(model1, scale = FALSE)
```

(b). Now add an additional predictor that is highly correlated with one of the informative predictors. Fit another random forest model to these data. Did the importance score for V1 change? What happens when you add another predictor that is also highly correlated with V1? For example:

```r
simulated$duplicate1 <- simulated$V1 + rnorm(200) * .1
cor(simulated$duplicate1, simulated$V1)
```

```r
#code
```

(c). Use the ‘cforest‘ function in the party package to fit a random forest model using conditional inference trees. The party package function ‘varimp‘ can calculate predictor importance. The ‘conditional‘ argument of that function toggles between the traditional importance measure and the modified version described in Strobl et al. (2007). Do these importances show the same pattern as the traditional random forest model?

```r
#code
```

(d). Repeat this process with different tree models, such as boosted trees and Cubist. Does the same pattern occur?

```r
#code
```

Kuhn and Johnson 8.2

Use a simulation to show tree bias with different granularities.

```
#code
```

Kuhn and Johnson 8.3

In stochastic gradient boosting the bagging fraction and learning rate will govern the construction of the trees as they are guided by the gradient. Although the optimal values of these parameters should be obtained through the tuning process, it is helpful to understand how the magnitudes of these parameters affect magnitudes of variable importance. Figure 8.24 provides the variable importance plots for boosting using two extreme values for the bagging fraction (0.1 and 0.9) and the learning rate (0.1 and 0.9) for the solubility data. The left-hand plot has both parameters set to 0.1, and the right-hand plot has both set to 0.9:

(a). Why does the model on the right focus its importance on just the first few of predictors, whereas the model on the left spreads importance across more predictors?

```
#code
```

(b). Which model do you think would be more predictive of other samples?

```
#code
```

(c). How would increasing interaction depth affect the slope of predictor importance for either model in Fig.8.24?

```
#code
```

Kuhn and Johnson 8.7

Refer to Exercises 6.3 and 7.5 which describe a chemical manufacturing process. Use the same data imputation, data splitting, and pre-processing steps as before and train several tree-based models:

(a). Which tree-based regression model gives the optimal resampling and test set performance?

```
#code
```

(b). Which predictors are most important in the optimal tree-based regression model? Do either the biological or process variables dominate the list? How do the top 10 important predictors compare to the top 10 predictors from the optimal linear and nonlinear models?

```
#code
```

(c). Plot the optimal single tree with the distribution of yield in the terminal nodes. Does this view of the data provide additional knowledge about the biological or process predictors and their relationship with yield?

```
#code
```

Assignment 4

TBD

R Script

```r
#Compile R Script Here for Professors Review
```