Homework Part 2

Data 624 - Predictive Analytics

16 December 2019

Group Members:

Vinicio Haro

Sang Yoon (Andy) Hwang

Julian McEachern

Jeremy O'Brien

Bethany Poulin

Contents

Getting Started **2**

Overview 2

Dependencies 2

Assignment 1 **3**

Kuhn and Johnson 6.3 3

Assignment 2 **7**

Kuhn and Johnson 7.2 7

Kuhn and Johnson 7.5 8

Assignment 3 **9**

Kuhn and Johnson 8.1 9

Kuhn and Johnson 8.2 9

Kuhn and Johnson 8.3 9

Kuhn and Johnson 8.7 9

Assignment 4 **11**

TBD 11

R Script **12**

Getting Started

Overview

Include details on our process in creating this document.

Dependencies

```
# Data Wrangling
library(AppliedPredictiveModeling)
library(mice)
library(caret)
library(tidyverse)
library(pls)
library(caTools)
library(mlbench)
library(stringr)

# Formatting
library(default)
library(knitr)
library(kableExtra)

# Plotting
library(ggplot2)
library(grid)
library(ggfortify)
library(gridExtra)
```

Assignment 1

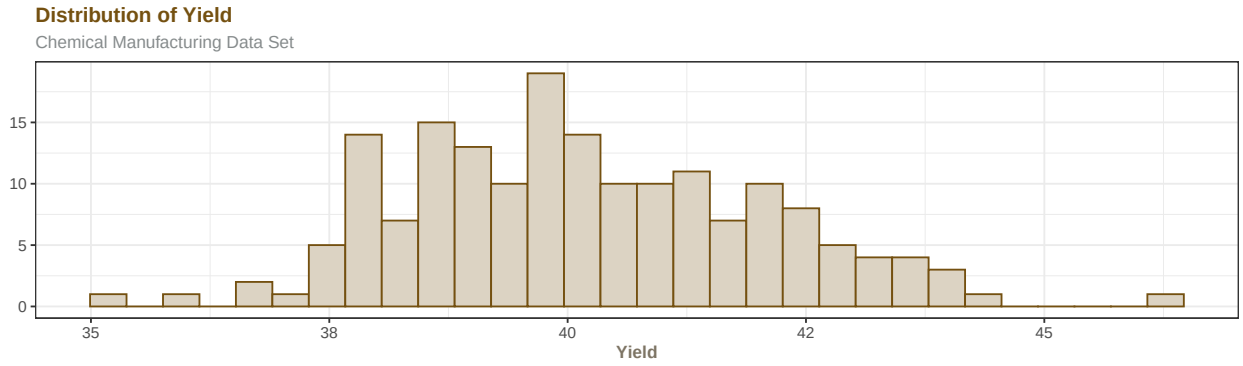
Kuhn and Johnson 6.3

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")
```

The matrix processPredictors contains the 57 predictors (12 describing the input biological material and 45 describing the process predictors) for the 176 manufacturing runs. yield contains the percent yield for each run.



(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).

`ManufacturingProcess03` has the largest volume of missing data followed by `ManufacturingProcess11`. Given that each variable has less than 25% of data missing, we should introduce methods of imputation. For our purposes, we will use the MICE method. MICE is formally known as Multiple Imputation with Chained Equations. On a high level, MICE is built off a tehcnique known as the Gibbs sampler.

The Gibbs sampler is a Markov chain based on Monte Carlo. MICE iterates drawing estimates of missing values and parameters related to the distribution of said variables. Chained equations are generally faster than the monte carlo based Gibbs sampler. MICE has 5 imputations listed as its default. predictive mean matching is also a default method for MICE. PMM does a better job at keeping non-linear relationships within individual variables.

In addition to MICE, we drop variables that have near zero variance, however we point out that only one variable was dropped. We still include it as a process step to follow the literature's specifications. After completing MICE, we no longer had missing data in our set. We examined other imputation methods such as KNN but determined that there was no significant change in the summary statistics across different imputation methods.

Table 1: Variables with Missing Values

Predictor	n	Predictor	n
ManufacturingProcess03	15	ManufacturingProcess02	3
ManufacturingProcess11	10	ManufacturingProcess06	2
ManufacturingProcess10	9	ManufacturingProcess01	1
ManufacturingProcess25	5	ManufacturingProcess04	1
ManufacturingProcess26	5	ManufacturingProcess05	1
ManufacturingProcess27	5	ManufacturingProcess07	1
ManufacturingProcess28	5	ManufacturingProcess08	1
ManufacturingProcess29	5	ManufacturingProcess12	1
ManufacturingProcess30	5	ManufacturingProcess14	1
ManufacturingProcess31	5	ManufacturingProcess22	1
ManufacturingProcess33	5	ManufacturingProcess23	1
ManufacturingProcess34	5	ManufacturingProcess24	1
ManufacturingProcess35	5	ManufacturingProcess40	1
ManufacturingProcess36	5	ManufacturingProcess41	1

(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?

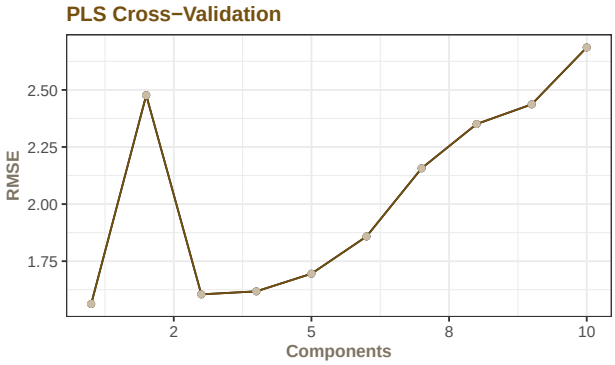
We will build a PLS model also known as partial least squares. PLS is a statistical method that fits a linear regression model by projecting the feature variables and response variable to some new space via a mapping function. Because of this projection mechanisim, for both predictors and the response, the method becomes bilinear or simply known as linear with respect to to each of the variable types. PLS also has certain advantages over other methods such as being more robust to dealing with issues arising from multicolinearity.

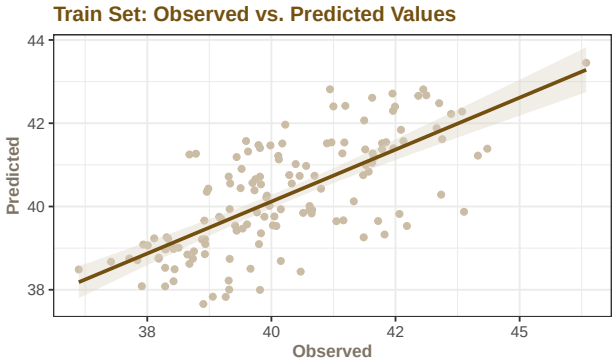
For our PLS model, we partitioned the data by taking 80% of the data as training and the remaining 20% as testing subsets. We also apply center and scaling arguments set to true. We built a standard PLS model and evaluated the root mean summary areas to determine the optimal number of components to select. We generate performance metrics for our best tune below:

Table 2: PLS Performance Metrics on Training Subset

RMSE	Rsquared	MAE
1.5626	0.3953	1.1845

Our Baseline PLS model generates a RMSE of 1.56. In addition, the model captures 39.53 % of data variability. We include the visualizations pertaining to the train set cross-validation RMSE tunes and a plot comparing the observed and predicted outcome from our model.





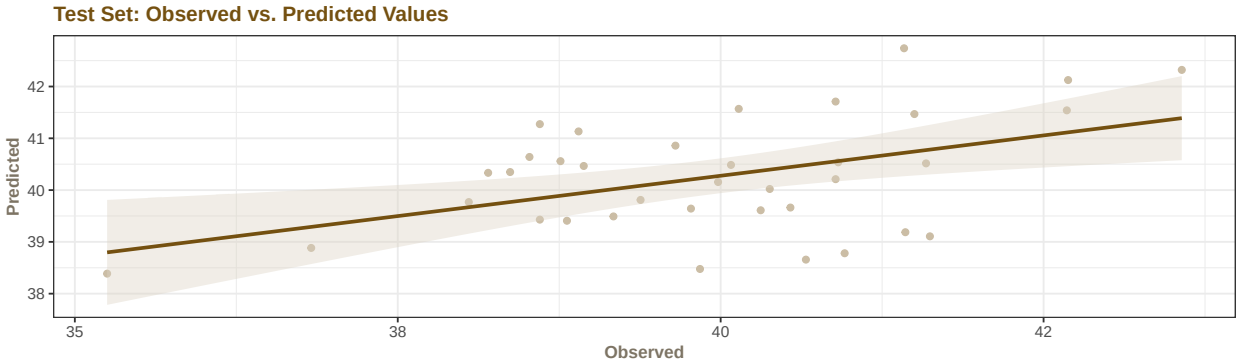
(d). Predict the response for the test set. What is the value of the performance metric and how does this compare with the resampled performance metric on the training set?

We see a decreased R squared against the test data with 22% of the data variability accounted for. We also see the RMSE decrease to 1.52 from our training results of 1.56. There is also a slight increase in the MAE.

Table 3: PLS Performance Metrics on Test Subset

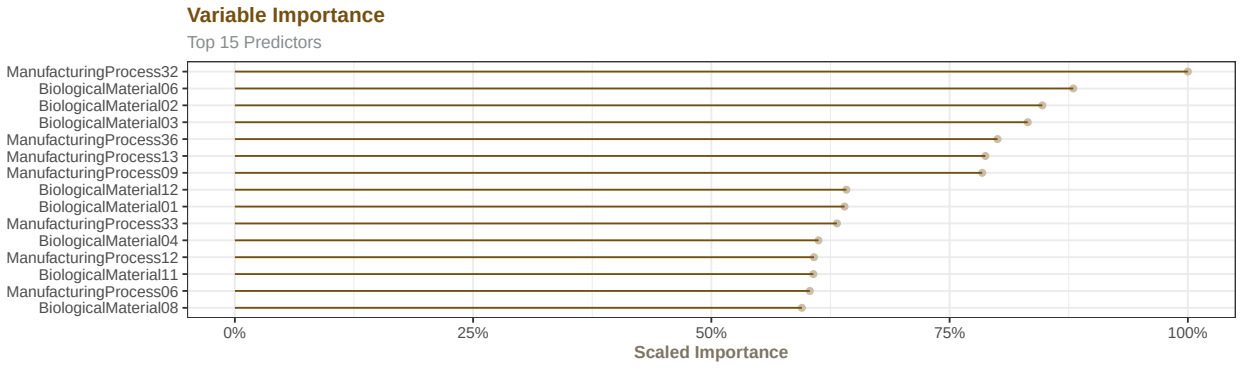
RMSE	Rsquared	MAE
1.5222	0.2212	1.2885

We also plotted the observed and predicted values from our test set against each other below. The deviation from the fitted line tells us that our selected linear model may not provide the best predictions for `Yield`.



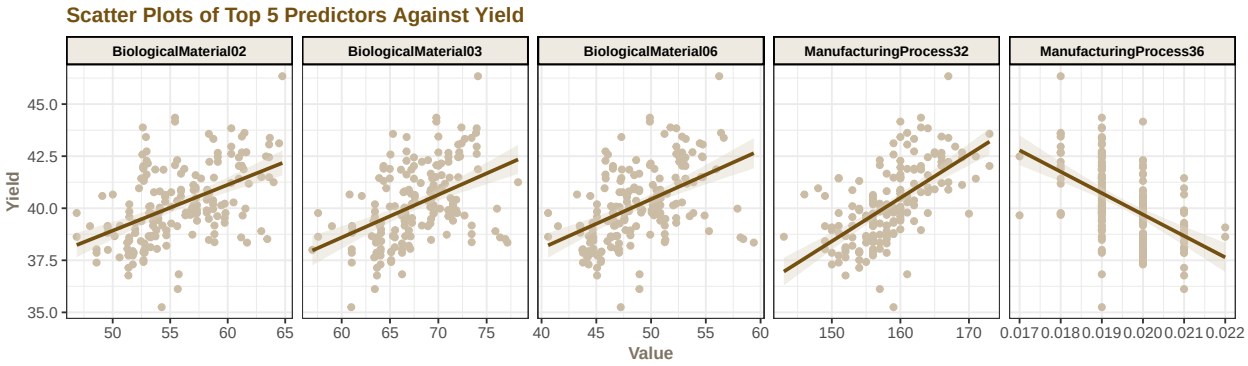
(e). Which predictors are most important in the model you have trained? Do either the biological or process predictors dominate the list?

VarImp allows us to identify the variables by name and compute their importance. `ManufacturingProcess32` was flagged as the most important predictor overall and within the group of other Manufacturing Process variables. `BiologicalMaterial06` ranked second and was the most important variable within the BiologicalMaterial group. The variable importance rankings are mixed with 8 variables belonging to Biology and 7 Manufacturing Process variables within the top 15 predictors.



(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?

We used a scatter plot to visualize the relationship between our top five important predictors against our response variable, `Yield`. All but `ManufacturingProcess32` show a moderate positive, linear relationship with yield.



We further examined this relationship by analyzing the correlation strength between our top five important response variables with the `Yield`. Out of which, `ManufacturingProcess32` showed the strongest, positive correlation with our response variable.

From a business point of view, our aim is to increase yield since we know that yield ties into revenue. We do not have insight into what mechanics go into each manufacturing process but we can use this knowledge to adjust the processes to emulate the highest yield outputs.

Table 4: Variable Correlation with Yield

Variable	Yield
ManufacturingProcess32	0.6083
BiologicalMaterial02	0.4815
BiologicalMaterial06	0.4782
BiologicalMaterial03	0.4451
ManufacturingProcess36	-0.5014

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: $y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + N(0,\sigma^2)$; 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).

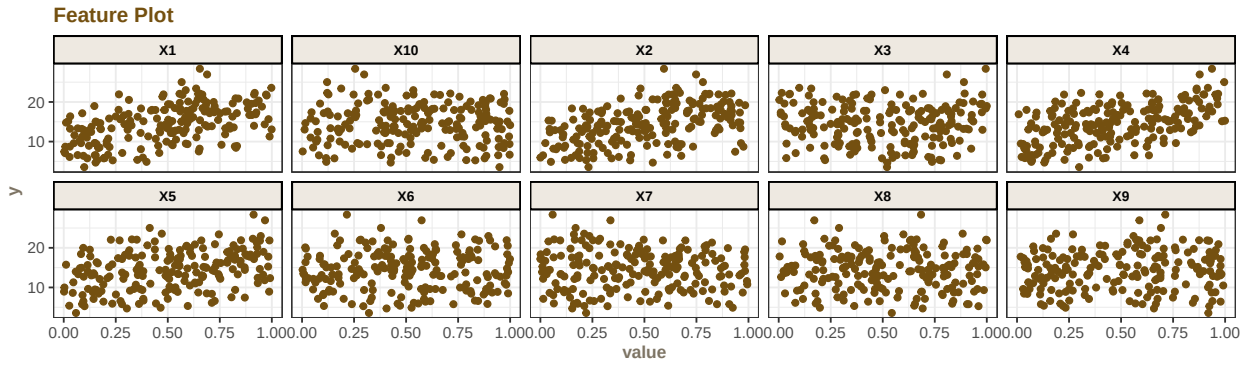
The package `mlbench` contains a function called `mlbench.friedman1` that simulates these data:

```
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. 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:

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

Using ggplots, we can look at a feature plot of the simulated data:



(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:

Model 2:

Model 3:

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

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.

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

(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?

(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?

Assignment 3

Kuhn and Johnson 8.1

Recreate the simulated data from Exercise 7.2:

(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)?

(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:

(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?

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

Kuhn and Johnson 8.2

Use a simulation to show tree bias with different granularities.

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?

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

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

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?

(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?

(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?

Assignment 4

TBD

R Script