Data 624: Week 8 Homework

Angrand, Burke, Deboch, Groysman, Karr

November 29, 2019

### Week 8 Assignment

library(AppliedPredictiveModeling)  
library(caret)  
library(dplyr)  
library(RANN)  
library(knitr)

**Chapter 6 KJ 6.3**

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).

* Find missing values with sapply. THe total dataframe only contains 175=6 rows and there are quite a few columns that are missing over 5% of their values. The values need to be imputted not removed.
* The mentioned section 3.8 highlights the impute.knn function from the impute library and the preprocess function from the caret library. The impute.knn function uses K-nearest neighbors to estimate the missing data and can be called as a subcomponent in the preprocess function.
* After calling the prerpocess function, the predict method applies the results to the set of data
* Check to see if all nulls have been removed with sapply

#dim  
dim(ChemicalManufacturingProcess)

## [1] 176 58

#check for NaNs  
sapply(ChemicalManufacturingProcess, function(x) sum(is.na(x)))

## Yield BiologicalMaterial01 BiologicalMaterial02   
## 0 0 0   
## BiologicalMaterial03 BiologicalMaterial04 BiologicalMaterial05   
## 0 0 0   
## BiologicalMaterial06 BiologicalMaterial07 BiologicalMaterial08   
## 0 0 0   
## BiologicalMaterial09 BiologicalMaterial10 BiologicalMaterial11   
## 0 0 0   
## BiologicalMaterial12 ManufacturingProcess01 ManufacturingProcess02   
## 0 1 3   
## ManufacturingProcess03 ManufacturingProcess04 ManufacturingProcess05   
## 15 1 1   
## ManufacturingProcess06 ManufacturingProcess07 ManufacturingProcess08   
## 2 1 1   
## ManufacturingProcess09 ManufacturingProcess10 ManufacturingProcess11   
## 0 9 10   
## ManufacturingProcess12 ManufacturingProcess13 ManufacturingProcess14   
## 1 0 1   
## ManufacturingProcess15 ManufacturingProcess16 ManufacturingProcess17   
## 0 0 0   
## ManufacturingProcess18 ManufacturingProcess19 ManufacturingProcess20   
## 0 0 0   
## ManufacturingProcess21 ManufacturingProcess22 ManufacturingProcess23   
## 0 1 1   
## ManufacturingProcess24 ManufacturingProcess25 ManufacturingProcess26   
## 1 5 5   
## ManufacturingProcess27 ManufacturingProcess28 ManufacturingProcess29   
## 5 5 5   
## ManufacturingProcess30 ManufacturingProcess31 ManufacturingProcess32   
## 5 5 0   
## ManufacturingProcess33 ManufacturingProcess34 ManufacturingProcess35   
## 5 5 5   
## ManufacturingProcess36 ManufacturingProcess37 ManufacturingProcess38   
## 5 0 0   
## ManufacturingProcess39 ManufacturingProcess40 ManufacturingProcess41   
## 0 1 1   
## ManufacturingProcess42 ManufacturingProcess43 ManufacturingProcess44   
## 0 0 0   
## ManufacturingProcess45   
## 0

#impute with preProcess, apply with predict  
impute <- preProcess(as.matrix(ChemicalManufacturingProcess), method=c("knnImpute"))  
impute.chem <- as.data.frame(predict(impute, as.matrix(ChemicalManufacturingProcess)))  
#check again for nulls after applying   
sapply(impute.chem, function(x) sum(is.na(x)))

## Yield BiologicalMaterial01 BiologicalMaterial02   
## 0 0 0   
## BiologicalMaterial03 BiologicalMaterial04 BiologicalMaterial05   
## 0 0 0   
## BiologicalMaterial06 BiologicalMaterial07 BiologicalMaterial08   
## 0 0 0   
## BiologicalMaterial09 BiologicalMaterial10 BiologicalMaterial11   
## 0 0 0   
## BiologicalMaterial12 ManufacturingProcess01 ManufacturingProcess02   
## 0 0 0   
## ManufacturingProcess03 ManufacturingProcess04 ManufacturingProcess05   
## 0 0 0   
## ManufacturingProcess06 ManufacturingProcess07 ManufacturingProcess08   
## 0 0 0   
## ManufacturingProcess09 ManufacturingProcess10 ManufacturingProcess11   
## 0 0 0   
## ManufacturingProcess12 ManufacturingProcess13 ManufacturingProcess14   
## 0 0 0   
## ManufacturingProcess15 ManufacturingProcess16 ManufacturingProcess17   
## 0 0 0   
## ManufacturingProcess18 ManufacturingProcess19 ManufacturingProcess20   
## 0 0 0   
## ManufacturingProcess21 ManufacturingProcess22 ManufacturingProcess23   
## 0 0 0   
## ManufacturingProcess24 ManufacturingProcess25 ManufacturingProcess26   
## 0 0 0   
## ManufacturingProcess27 ManufacturingProcess28 ManufacturingProcess29   
## 0 0 0   
## ManufacturingProcess30 ManufacturingProcess31 ManufacturingProcess32   
## 0 0 0   
## ManufacturingProcess33 ManufacturingProcess34 ManufacturingProcess35   
## 0 0 0   
## ManufacturingProcess36 ManufacturingProcess37 ManufacturingProcess38   
## 0 0 0   
## ManufacturingProcess39 ManufacturingProcess40 ManufacturingProcess41   
## 0 0 0   
## ManufacturingProcess42 ManufacturingProcess43 ManufacturingProcess44   
## 0 0 0   
## ManufacturingProcess45   
## 0

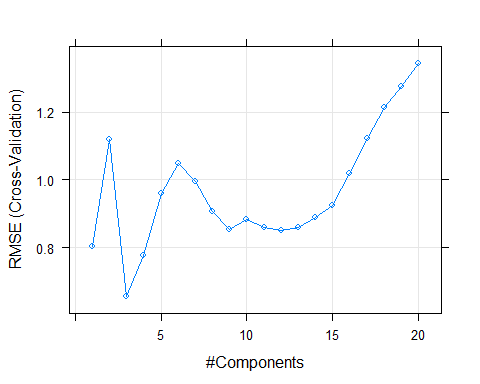
**c.** Split the data into a training and a test set, pre-process the data, and tune a model of your choice from this chapter. What is the optimal value of the performance metric?

* Yield c(1) is the response of the other columns (predictors)
* Prepare the data: split the data into train/test samples. Train (75% for building a predictive model) and Test (15% for evaluating the model)
* Use partial least squares method with a tested 20 different values for the tuning parameter ncomp
* As seen below, the most optimal value is ncomp = 3 with the smallest RSME of 0.6554035 and a R^2 of 0.6096468

## set the seed to make the partition reproducible  
set.seed(123)  
train.chem <- createDataPartition(ChemicalManufacturingProcess$Yield, p=0.75, list=FALSE)  
  
#apply to the predictors   
chem.Train <- impute.chem[train.chem,-1]  
chem.Test <- impute.chem[-train.chem,-1]  
#apply to yield  
yield.Train <- impute.chem[train.chem,1]  
yield.Test <- impute.chem[-train.chem,1]  
  
#partial least squares w/ train data   
  
pls.chem <- train(chem.Train, yield.Train,  
 method = "pls",  
 tuneLength = 20, trControl = trainControl(method = "cv", number = 10),  
 preProc = c("center", "scale"))  
#print outcomes of the pls  
pls.chem

## Partial Least Squares   
##   
## 132 samples  
## 57 predictor  
##   
## Pre-processing: centered (57), scaled (57)   
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 119, 118, 120, 120, 118, 118, ...   
## Resampling results across tuning parameters:  
##   
## ncomp RMSE Rsquared MAE   
## 1 0.8047230 0.4775122 0.6339437  
## 2 1.1187475 0.4818566 0.6862102  
## 3 0.6554035 0.6096468 0.5318995  
## 4 0.7774620 0.5555155 0.5816168  
## 5 0.9598662 0.4876206 0.6366717  
## 6 1.0485623 0.4743592 0.6653206  
## 7 0.9953023 0.4820678 0.6540929  
## 8 0.9072158 0.5017679 0.6331196  
## 9 0.8528902 0.5063925 0.6226068  
## 10 0.8838340 0.4967802 0.6336124  
## 11 0.8606428 0.4919373 0.6392547  
## 12 0.8513129 0.4905676 0.6490674  
## 13 0.8600490 0.4801921 0.6694921  
## 14 0.8886411 0.4663450 0.6848575  
## 15 0.9253709 0.4541527 0.7022949  
## 16 1.0199562 0.4323120 0.7366779  
## 17 1.1213147 0.4124365 0.7769600  
## 18 1.2121685 0.4016808 0.8083622  
## 19 1.2764454 0.3933288 0.8276471  
## 20 1.3445415 0.3906539 0.8448760  
##   
## RMSE was used to select the optimal model using the smallest value.  
## The final value used for the model was ncomp = 3.

# Plot model RMSE vs different values of components  
plot(pls.chem)



# Print the best tuning parameter ncomp that  
# minimize the cross-validation error, RMSE  
pls.chem$bestTune

## ncomp  
## 3 3

# Summarize the final model  
summary(pls.chem$finalModel)

## Data: X dimension: 132 57   
## Y dimension: 132 1  
## Fit method: oscorespls  
## Number of components considered: 3  
## TRAINING: % variance explained  
## 1 comps 2 comps 3 comps  
## X 17.17 25.62 31.08  
## .outcome 52.84 67.27 72.34

**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?

* Make predictions with the predict() function with the inputted chem.test data
* Compare the predicted values to the actual valyes “yield.Test”
* The RMSE is very close to the train data RSME. The Rsquare value is lower than the train values.

# Make predictions  
  
predictions <- predict(pls.chem, newdata = chem.Test)  
  
data.frame(  
 RMSE = caret::RMSE(predictions, yield.Test),  
 Rsquare = caret::R2(predictions, yield.Test)  
)

## RMSE Rsquare  
## 1 0.694739 0.4295074

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

* Find the absolute value from the mean contributions for each coefficient
* ManufacturingProcess32, ManufacturingProcess13,ManufacturingProcess17 & ManufacturingProcess09 appear to be the most significant by a good margin.
* In general, the manufacturing process variables appear to be more significant than any other grouping of variables

predictors.pls <- as.data.frame(pls.chem$finalModel$coefficients)  
predictors.pls<- tibble::rownames\_to\_column(predictors.pls, "coefficients")  
predictors.pls%>%   
 mutate(meancol= rowMeans(.[, 2:4]))%>%  
 mutate(absmeancol =abs(meancol))%>%  
 arrange(-absmeancol)

## coefficients .outcome.1 comps .outcome.2 comps  
## 1 ManufacturingProcess32 0.0712238757 1.109397e-01  
## 2 ManufacturingProcess13 -0.0661158030 -1.395337e-01  
## 3 ManufacturingProcess17 -0.0577815025 -1.395057e-01  
## 4 ManufacturingProcess09 0.0654051170 1.303278e-01  
## 5 ManufacturingProcess36 -0.0629479000 -9.425630e-02  
## 6 ManufacturingProcess11 0.0507167348 9.468503e-02  
## 7 ManufacturingProcess12 0.0470604710 8.639382e-02  
## 8 ManufacturingProcess06 0.0485912907 8.009020e-02  
## 9 ManufacturingProcess33 0.0514050466 6.220852e-02  
## 10 ManufacturingProcess37 -0.0291606515 -6.601438e-02  
## 11 ManufacturingProcess34 0.0178812807 6.052770e-02  
## 12 ManufacturingProcess15 0.0324619266 3.801886e-02  
## 13 ManufacturingProcess30 0.0324426471 5.579544e-02  
## 14 BiologicalMaterial06 0.0584312750 4.495394e-02  
## 15 BiologicalMaterial03 0.0539347628 4.805170e-02  
## 16 BiologicalMaterial02 0.0596292461 4.351645e-02  
## 17 ManufacturingProcess24 -0.0279548342 -3.764810e-02  
## 18 BiologicalMaterial07 -0.0106260492 -3.533573e-02  
## 19 ManufacturingProcess21 -0.0074744865 -4.431938e-02  
## 20 ManufacturingProcess39 0.0060988025 3.620146e-02  
## 21 ManufacturingProcess44 0.0094662555 3.807731e-02  
## 22 ManufacturingProcess10 0.0291995589 4.993165e-02  
## 23 ManufacturingProcess43 0.0233307462 3.665961e-02  
## 24 BiologicalMaterial08 0.0518001888 2.836059e-02  
## 25 BiologicalMaterial12 0.0490325444 2.972720e-02  
## 26 BiologicalMaterial04 0.0468070169 2.374136e-02  
## 27 ManufacturingProcess45 0.0013979169 2.482118e-02  
## 28 BiologicalMaterial11 0.0463832457 2.363290e-02  
## 29 BiologicalMaterial05 0.0210190639 1.644061e-02  
## 30 ManufacturingProcess08 0.0095328221 2.820146e-02  
## 31 BiologicalMaterial01 0.0430725927 1.508968e-02  
## 32 ManufacturingProcess22 0.0072933469 2.560602e-02  
## 33 ManufacturingProcess19 0.0219409221 3.318482e-03  
## 34 ManufacturingProcess35 -0.0185679782 -2.204654e-02  
## 35 ManufacturingProcess42 -0.0021437692 1.800772e-02  
## 36 ManufacturingProcess29 0.0190983507 1.086223e-02  
## 37 ManufacturingProcess31 -0.0084488434 -2.342044e-02  
## 38 ManufacturingProcess41 -0.0055720930 -1.793995e-02  
## 39 ManufacturingProcess40 -0.0067586547 -1.612848e-02  
## 40 ManufacturingProcess28 0.0354945629 1.093106e-02  
## 41 ManufacturingProcess20 -0.0071633382 8.146575e-03  
## 42 ManufacturingProcess18 -0.0074149492 7.527277e-03  
## 43 ManufacturingProcess05 0.0128651292 -7.162564e-03  
## 44 ManufacturingProcess38 -0.0103386629 -8.397222e-03  
## 45 BiologicalMaterial09 0.0151941295 -6.127552e-03  
## 46 ManufacturingProcess16 -0.0032859018 -5.494742e-03  
## 47 ManufacturingProcess25 0.0009501445 -1.361350e-02  
## 48 ManufacturingProcess01 -0.0087233596 9.917543e-03  
## 49 ManufacturingProcess27 0.0011863844 -1.229260e-02  
## 50 ManufacturingProcess04 -0.0344634603 -1.484713e-02  
## 51 ManufacturingProcess23 -0.0066616847 9.688644e-05  
## 52 ManufacturingProcess14 0.0011316252 -1.252973e-02  
## 53 ManufacturingProcess07 0.0027703386 4.052646e-03  
## 54 BiologicalMaterial10 0.0293414934 -5.953754e-03  
## 55 ManufacturingProcess26 0.0048294389 -6.316201e-03  
## 56 ManufacturingProcess02 -0.0243838427 1.151700e-02  
## 57 ManufacturingProcess03 -0.0025420455 -5.934078e-03  
## .outcome.3 comps meancol absmeancol  
## 1 0.171668817 0.1179441344 0.1179441344  
## 2 -0.141447945 -0.1156991341 0.1156991341  
## 3 -0.149495753 -0.1155943036 0.1155943036  
## 4 0.137650718 0.1111278698 0.1111278698  
## 5 -0.134944299 -0.0973828316 0.0973828316  
## 6 0.084715814 0.0767058595 0.0767058595  
## 7 0.076668653 0.0700409813 0.0700409813  
## 8 0.078469904 0.0690504646 0.0690504646  
## 9 0.090165950 0.0679265063 0.0679265063  
## 10 -0.095111378 -0.0634288041 0.0634288041  
## 11 0.099304472 0.0592378170 0.0592378170  
## 12 0.081504796 0.0506618599 0.0506618599  
## 13 0.058618280 0.0489521214 0.0489521214  
## 14 0.041960341 0.0484485186 0.0484485186  
## 15 0.042659663 0.0482153756 0.0482153756  
## 16 0.038507706 0.0472177999 0.0472177999  
## 17 -0.054699035 -0.0401006549 0.0401006549  
## 18 -0.073436749 -0.0397995086 0.0397995086  
## 19 -0.058054979 -0.0366162835 0.0366162835  
## 20 0.067164453 0.0364882386 0.0364882386  
## 21 0.058157601 0.0352337210 0.0352337210  
## 22 0.023711632 0.0342809464 0.0342809464  
## 23 0.040475152 0.0334885034 0.0334885034  
## 24 0.010848688 0.0303364897 0.0303364897  
## 25 0.011644640 0.0301347949 0.0301347949  
## 26 0.018201184 0.0295831877 0.0295831877  
## 27 0.053864928 0.0266946753 0.0266946753  
## 28 0.007878601 0.0259649170 0.0259649170  
## 29 0.035645329 0.0243683345 0.0243683345  
## 30 0.033862102 0.0238654608 0.0238654608  
## 31 0.012873518 0.0236785964 0.0236785964  
## 32 0.035039319 0.0226462274 0.0226462274  
## 33 0.041016579 0.0220919943 0.0220919943  
## 34 -0.020080093 -0.0202315381 0.0202315381  
## 35 0.042963320 0.0196090904 0.0196090904  
## 36 0.027711629 0.0192240708 0.0192240708  
## 37 -0.015868135 -0.0159124722 0.0159124722  
## 38 -0.021370578 -0.0149608744 0.0149608744  
## 39 -0.020313656 -0.0144002632 0.0144002632  
## 40 -0.010835085 0.0118635134 0.0118635134  
## 41 0.033078216 0.0113538174 0.0113538174  
## 42 0.033395149 0.0111691589 0.0111691589  
## 43 -0.033968050 -0.0094218283 0.0094218283  
## 44 -0.009295092 -0.0093436589 0.0093436589  
## 45 -0.036104949 -0.0090127906 0.0090127906  
## 46 -0.014100772 -0.0076271384 0.0076271384  
## 47 -0.005385365 -0.0060162385 0.0060162385  
## 48 0.015380878 0.0055250205 0.0055250205  
## 49 -0.004295679 -0.0051339650 0.0051339650  
## 50 0.034398893 -0.0049705663 0.0049705663  
## 51 -0.007221722 -0.0045955066 0.0045955066  
## 52 0.024393990 0.0043319626 0.0043319626  
## 53 -0.013624735 -0.0022672502 0.0022672502  
## 54 -0.019834614 0.0011843750 0.0011843750  
## 55 0.003344658 0.0006192985 0.0006192985  
## 56 0.011073945 -0.0005976317 0.0005976317  
## 57 0.009477843 0.0003339065 0.0003339065

**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?

* For the manufacturing processes with negative coefficients, the facility could alter their processes to decrease the associated impact to yields
* For the manufacturing processes with positibe coefficients, the facility to could their processes to increase the associasted impact to yields
* Given that Biological materials do not have a significant impact, the facility could alter the ingrediants/materials to increase the associated yields