DATA 624 - Homework 7

OMER OZEREN

Table of Contents

Exercises 6.2 & 6.3

```
library(AppliedPredictiveModeling)
library(caret)
library(elasticnet)
library(knitr)
library(pls)
library(ggplot2)
library(tidyverse)
library(kableExtra)
library(RANN)
library(corrplot)
```

Question 6.2

Developing a model to predict permeability (see Sect. 1.4) could save significant resources for a pharmaceutical company, while at the same time more rapidly identifying molecules that have a sufficient permeability to become a drug:

PART A

Start R and use these commands to load the data:

```
data(permeability)
```

The fingerprints matrix holds **165 unique compounds**; **1107 molecular fingerprints**

Part B

the `fingerprints` predictors indicate the presense or absense of substructures of a molecule and are often sparse meaning that relatively few of the molecules contain each substructure. Filter out the predictors that have low frequencies using the `nearZeroVar` function from the `caret` package. How many are left for modeling?

```
fingerprints %>%
  nearZeroVar() %>%
  length()

## [1] 719
```

There are 719 variables left after filtering out the near zero variables.

Part C

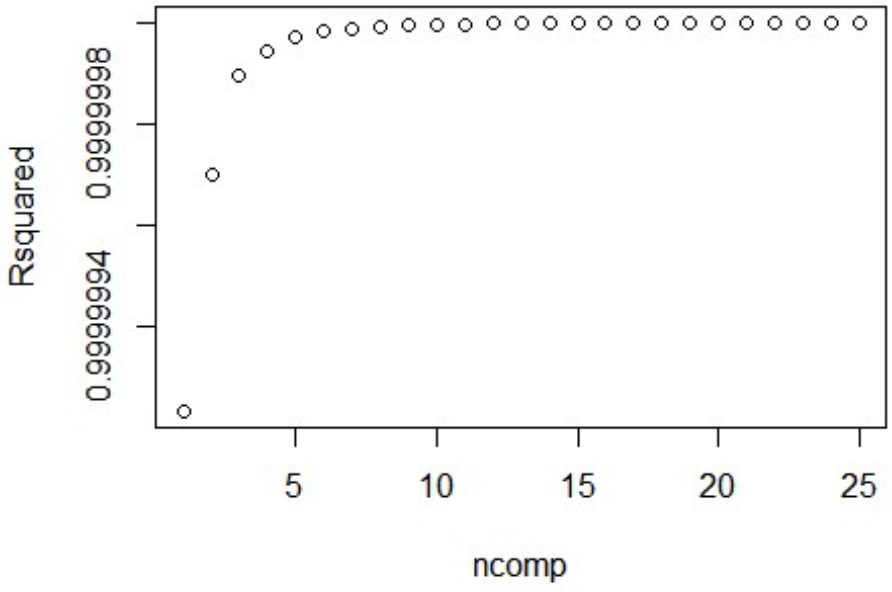
Split the data into a training and a test set, pre-process the data, and tune a PLS model. How many latent variables are optimal and what is the corresponding resampled estimate of R2?

I'm going to split the data 70% for training and 30% for testign.

```
data_clear <- as.data.frame(fingerprints[, nearZeroVar(fingerprints)]) %>%
  mutate(y = permeability)
set.seed(42)
data_clear <- cbind(data.frame(permeability),data_clear)
n <-  floor(0.70 * nrow(data_clear))
idx <- sample(seq_len(nrow(data_clear)), size = n)
training_df <- data_clear[idx, ]
testing_df <- data_clear[-idx, ]

# build PLS model
pls_model <- train(
  y ~ ., data = training_df, method = "pls",
  center = TRUE,
  trControl = trainControl("cv", number = 10),
  tuneLength = 25
)
#results
plot(pls_model$results$Rsquared,
     xlab = "ncomp",
```

```
     ylab = "Rsquared"
     )
```




```
pls_model$results %>%
  filter(ncomp == pls_model$bestTune$ncomp) %>%
  select(ncomp, RMSE, Rsquared) %>%
  kable() %>%
  kable_styling()
```

ncomp
RMSE
Rsquared
25
6.15e-05
1

As we can see above plot, the optimal components number in model is 25.In addition to that, the PLS model captures 100% of the permeability .

PART D

Predict the response for the test set. What is the test set estimate of R2?

```
# Make predictions
pred <- predict(pls_model, testing_df)
# Error Metric/Model Evaluation
```

```
results <- data.frame(Model = "PLS Model",
                      RMSE = caret::RMSE(pred, testing_df$y),
                      Rsquared = caret::R2(pred, testing_df$y))
results %>%
  kable() %>%
  kable_styling()
```

Model
RMSE
Rsquared
permeability
PLS Model
5.96e-05
1

We got the the same R^2 which is 1.I actually also tried for 80/20 % split.However, I got the the same R^2.

Part E

Try building other models discussed in this chapter. Do any have better predictive performance?

I'll use Elastic Net Regression

Elastic Net Regression

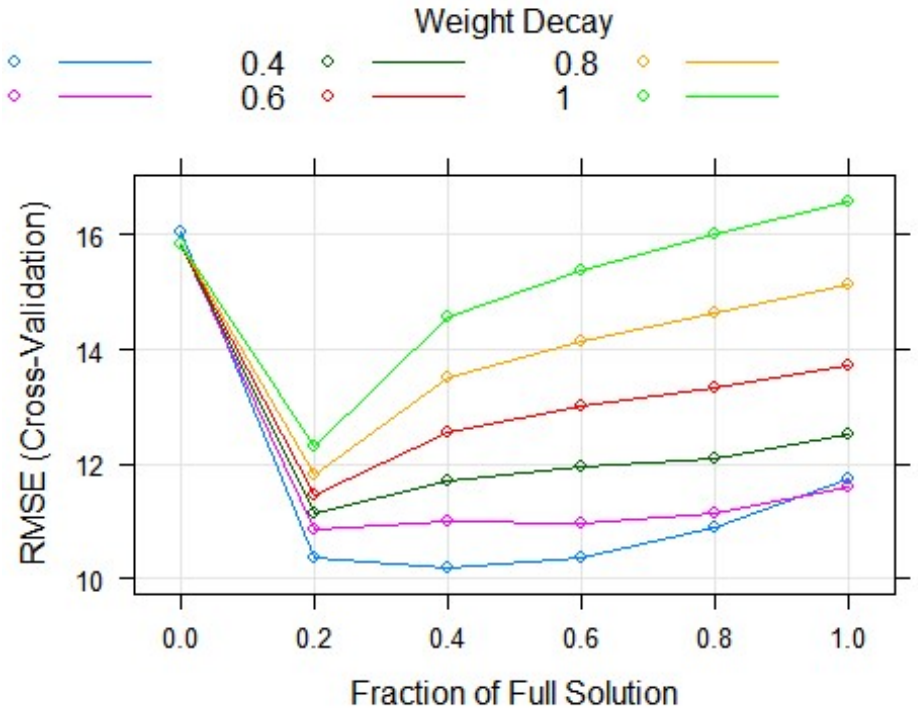
```
data_clear  <- fingerprints[, -nearZeroVar(fingerprints)]
data_clear <- cbind(data.frame(permeability),data_clear) #adding permeability
number <-  floor(0.70 * nrow(data_clear)) # 70/30 split
idx <- sample(seq_len(nrow(data_clear)), size = number)
train_df <- data_clear[idx, ]
test_df <- data_clear[-idx, ]

#train the Elastic Net model
elastic_model <-  train(x=train_df[,-1],
                  y=train_df$permeability,
                  method='enet',
                  metric='RMSE', # error mettric
                  tuneGrid=expand.grid(.fraction = seq(0, 1, by=0.2),
                                       .lambda = seq(0, 1, by=0.2)),
                  trControl=trainControl(method='cv',number=10),
                  preProcess=c('center','scale'))
```

```
## Warning: model fit failed for Fold09: lambda=0.0, fraction=1 Error in if
(zmin < gamhat) { : missing value where TRUE/FALSE needed
```

```
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info =
trainInfo, :
## There were missing values in resampled performance measures.
```

```
plot(elastic_model)
```




```
#best params
elastic_model$bestTune

##   fraction lambda
## 3      0.4      0

#perf of best params
getTrainPerf(elastic_model)

##   TrainRMSE TrainRsquared TrainMAE method
## 1  10.18728     0.6116537 7.391802   enet
```

As we can see on graph on above R^2 declined from 1 to 0.55

PART F

Would you recommend any of your models to replace the permeability laboratory experiment?

No, it is obvious that the predictive power from Elastic Net Regression is not as good as the laboratory experiment.

Question 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), 6.5 Computing 139 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:

PART A

Start R and use these commands to load the data:

```
data(ChemicalManufacturingProcess)
chem <- ChemicalManufacturingProcess
head(chem)
```

```
##   Yield BiologicalMaterial01 BiologicalMaterial02 BiologicalMaterial03
## 1 38.00                6.25                49.58                56.97
## 2 42.44                8.01                60.97                67.48
## 3 42.03                8.01                60.97                67.48
## 4 41.42                8.01                60.97                67.48
## 5 42.49                7.47                63.33                72.25
## 6 43.57                6.12                58.36                65.31
##   BiologicalMaterial04 BiologicalMaterial05 BiologicalMaterial06
## 1                12.74                19.51                43.73
## 2                14.65                19.36                53.14
## 3                14.65                19.36                53.14
## 4                14.65                19.36                53.14
## 5                14.02                17.91                54.66
## 6                15.17                21.79                51.23
##   BiologicalMaterial07 BiologicalMaterial08 BiologicalMaterial09
## 1                  100                16.66                11.44
## 2                  100                19.04                12.55
## 3                  100                19.04                12.55
## 4                  100                19.04                12.55
## 5                  100                18.22                12.80
## 6                  100                18.30                12.13
##   BiologicalMaterial10 BiologicalMaterial11 BiologicalMaterial12
## 1                 3.46               138.09                18.83
## 2                 3.46               153.67                21.05
## 3                 3.46               153.67                21.05
## 4                 3.46               153.67                21.05
## 5                 3.05               147.61                21.05
## 6                 3.78               151.88                20.76
##   ManufacturingProcess01 ManufacturingProcess02 ManufacturingProcess03
## 1                     NA                     NA                     NA
```

```
## 2                    0.0                     0                     NA
## 3                    0.0                     0                     NA
## 4                    0.0                     0                     NA
## 5                   10.7                     0                     NA
## 6                   12.0                     0                     NA
##   ManufacturingProcess04 ManufacturingProcess05 ManufacturingProcess06
## 1                     NA                     NA                     NA
## 2                    917                 1032.2                  210.0
## 3                    912                 1003.6                  207.1
## 4                    911                 1014.6                  213.3
## 5                    918                 1027.5                  205.7
## 6                    924                 1016.8                  208.9
##   ManufacturingProcess07 ManufacturingProcess08 ManufacturingProcess09
## 1                     NA                     NA                  43.00
## 2                    177                    178                  46.57
## 3                    178                    178                  45.07
## 4                    177                    177                  44.92
## 5                    178                    178                  44.96
## 6                    178                    178                  45.32
##   ManufacturingProcess10 ManufacturingProcess11 ManufacturingProcess12
## 1                     NA                     NA                     NA
## 2                     NA                     NA                      0
## 3                     NA                     NA                      0
## 4                     NA                     NA                      0
## 5                     NA                     NA                      0
## 6                     NA                     NA                      0
##   ManufacturingProcess13 ManufacturingProcess14 ManufacturingProcess15
## 1                   35.5                   4898                   6108
## 2                   34.0                   4869                   6095
## 3                   34.8                   4878                   6087
## 4                   34.8                   4897                   6102
## 5                   34.6                   4992                   6233
## 6                   34.0                   4985                   6222
##   ManufacturingProcess16 ManufacturingProcess17 ManufacturingProcess18
## 1                   4682                   35.5                   4865
## 2                   4617                   34.0                   4867
## 3                   4617                   34.8                   4877
## 4                   4635                   34.8                   4872
## 5                   4733                   33.9                   4886
## 6                   4786                   33.4                   4862
##   ManufacturingProcess19 ManufacturingProcess20 ManufacturingProcess21
## 1                   6049                   4665                    0.0
## 2                   6097                   4621                    0.0
## 3                   6078                   4621                    0.0
## 4                   6073                   4611                    0.0
## 5                   6102                   4659                   -0.7
## 6                   6115                   4696                   -0.6
##   ManufacturingProcess22 ManufacturingProcess23 ManufacturingProcess24
## 1                     NA                     NA                     NA
## 2                      3                      0                      3
```

```
## 3                     4                     1                     4
## 4                     5                     2                     5
## 5                     8                     4                    18
## 6                     9                     1                     1
##   ManufacturingProcess25 ManufacturingProcess26 ManufacturingProcess27
## 1                   4873                   6074                   4685
## 2                   4869                   6107                   4630
## 3                   4897                   6116                   4637
## 4                   4892                   6111                   4630
## 5                   4930                   6151                   4684
## 6                   4871                   6128                   4687
##   ManufacturingProcess28 ManufacturingProcess29 ManufacturingProcess30
## 1                   10.7                   21.0                    9.9
## 2                   11.2                   21.4                    9.9
## 3                   11.1                   21.3                    9.4
## 4                   11.1                   21.3                    9.4
## 5                   11.3                   21.6                    9.0
## 6                   11.4                   21.7                   10.1
##   ManufacturingProcess31 ManufacturingProcess32 ManufacturingProcess33
## 1                   69.1                    156                     66
## 2                   68.7                    169                     66
## 3                   69.3                    173                     66
## 4                   69.3                    171                     68
## 5                   69.4                    171                     70
## 6                   68.2                    173                     70
##   ManufacturingProcess34 ManufacturingProcess35 ManufacturingProcess36
## 1                    2.4                    486                  0.019
## 2                    2.6                    508                  0.019
## 3                    2.6                    509                  0.018
## 4                    2.5                    496                  0.018
## 5                    2.5                    468                  0.017
## 6                    2.5                    490                  0.018
##   ManufacturingProcess37 ManufacturingProcess38 ManufacturingProcess39
## 1                    0.5                      3                    7.2
## 2                    2.0                      2                    7.2
## 3                    0.7                      2                    7.2
## 4                    1.2                      2                    7.2
## 5                    0.2                      2                    7.3
## 6                    0.4                      2                    7.2
##   ManufacturingProcess40 ManufacturingProcess41 ManufacturingProcess42
## 1                     NA                     NA                   11.6
## 2                    0.1                   0.15                   11.1
## 3                    0.0                   0.00                   12.0
## 4                    0.0                   0.00                   10.6
## 5                    0.0                   0.00                   11.0
## 6                    0.0                   0.00                   11.5
##   ManufacturingProcess43 ManufacturingProcess44 ManufacturingProcess45
## 1                    3.0                    1.8                    2.4
## 2                    0.9                    1.9                    2.2
## 3                    1.0                    1.8                    2.3
```

```
## 4                    1.1                    1.8                    2.1
## 5                    1.1                    1.7                    2.1
## 6                    2.2                    1.8                    2.0
```

The matrix `ChemicalManufacturingProcess` has the 57 explanatory variable

- 12 of 57 explanatory variable is biological material and
- 45 of 57 explanatory variable is the process variable for the 176 manufacturing purposes.

Part 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).

I will imputer missing values with KNN to impute values.

```
# Make this reproducible
set.seed(42)
knn_model <- preProcess(ChemicalManufacturingProcess, "knnImpute")
df_no_missing <- predict(knn_model, ChemicalManufacturingProcess)
```

PART C

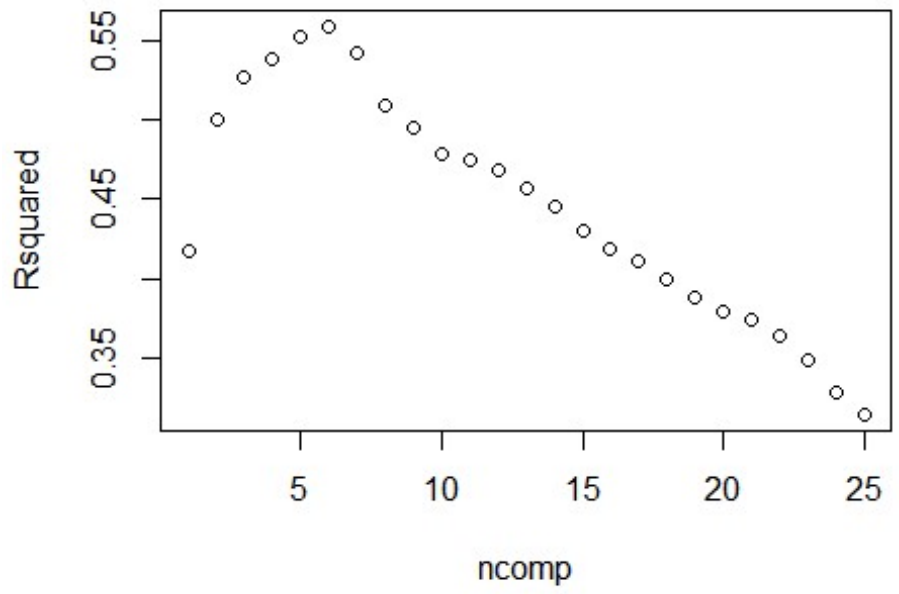
Split the data into a training and a test set, pre-process the data, and tune a model of your choice from this chapter. What is the optimal value of the performance metric?

I will split the data 70/30 the same as question 6.2 part c

```
number <-  floor(0.70 * nrow(df_no_missing)) # 70/30 split
idx <- sample(seq_len(nrow(df_no_missing)), size = number)
training_df <- df_no_missing[idx, ]
testing_df <- df_no_missing[-idx, ]
```

I will buila PLS model since I got really good results for question 6.2

```
# build PLS model
pls_model <- train(
  Yield  ~ ., data = training_df, method = "pls",
  center = TRUE,
  trControl = trainControl("cv", number = 10),
  tuneLength = 25
)
#pls model results
plot(pls_model$results$Rsquared,
     xlab = "ncomp",
     ylab = "Rsquared"
     )
```




```
pls_model$results %>%
  filter(ncomp == pls_model$bestTune$ncomp) %>%
  select(ncomp, RMSE, Rsquared) %>%
  kable() %>%
  kable_styling()
```

ncomp
RMSE
Rsquared
3
0.702369
0.526431

As we can see above plot, the optimal components number in model is 3.In addition to that, the PLS model captures 53% of the Yield .

PART 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
pred <- predict(pls_model, testing_df)
# Error Metric/Model Evaluation
results <- data.frame(Model = "PLS Model",
                      RMSE = caret::RMSE(pred, testing_df$Yield),
```

```
                        Rsquared = caret::R2(pred, testing_df$Yield))
results %>%
  kable() %>%
  kable_styling()
```

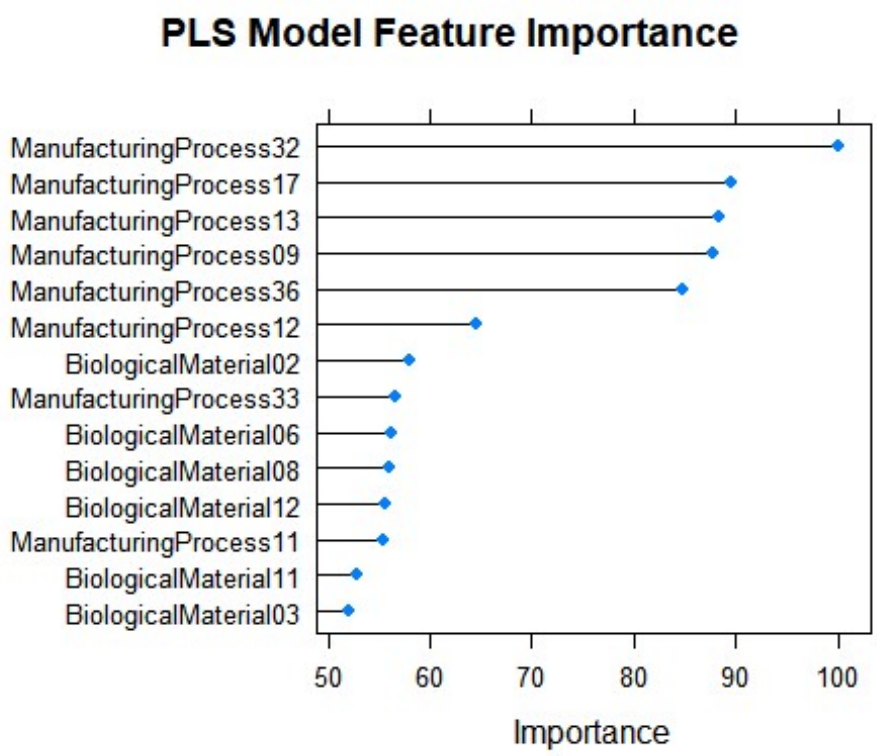
Model
RMSE
Rsquared
PLS Model
0.5571291
0.6854267

As we see above display, the error metric RMSE is lower and R^2 is higher with test data set.

Part E

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

```
pls_importance <- varImp(pls_model)$importance %>%
  as.data.frame() %>%
  rownames_to_column("Variable") %>%
  filter(Overall >= 50) %>% # set a threshold for vairables importance
  arrange(desc(Overall)) %>%
  mutate(importance = row_number())
varImp(pls_model) %>%
  plot(., top = max(pls_importance$importance), main = "PLS Model Feature
Importance")
```



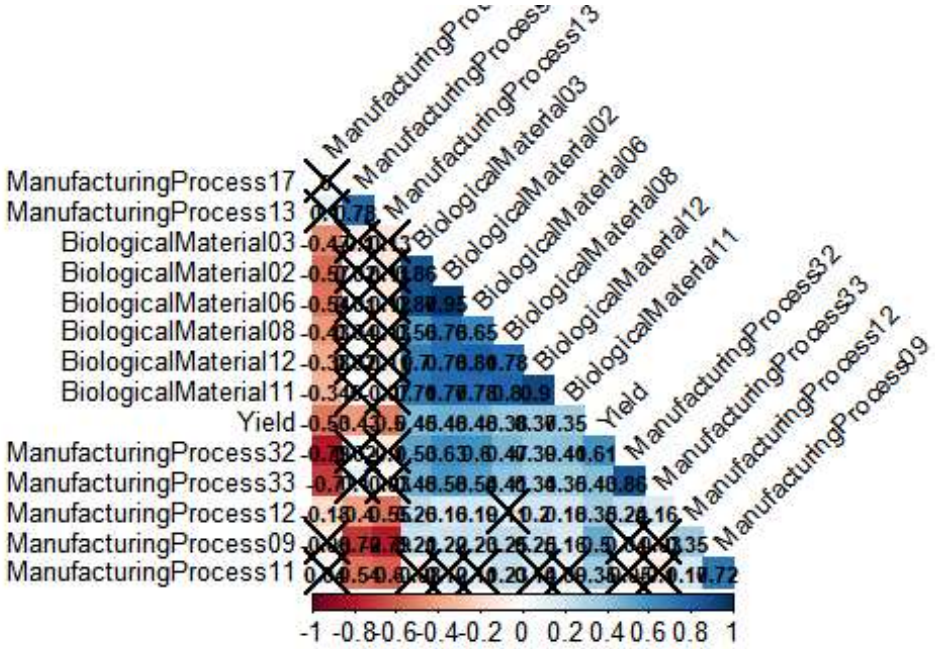
The PLS Model Feature importances indicates that ManufacturingProcess32 is the most importance variable for the PL model.In order to move forward, We can set a threshold and only pass the variables that threshold.Here I set a threshold as at least 50 % importance for PLS model.

Part F

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

```
important_vars <- df_no_missing %>%
  select_at(vars(Yield, pls_importance$Variable))

important_vars_p <- cor.mtest(important_vars)$p
important_vars %>%
  cor() %>%
  corrplot(method = "color", type = "lower", order = "hclust",
           tl.cex = 0.8, tl.col = "black", tl.srt = 45,
           addCoef.col = "black", number.cex = 0.7,
           p.mat = important_vars_p,  sig.level = 0.05, diag = FALSE)
```



The purpose of relationship between each of the top predictors and the response, I plotted the corrrelation relations for each important variable to respond variable.The correlation heat map shows that variables are positively correleted with Yield respond.The Manufacuring process 32 is the most correleted variable to respond variable.Some variables are negatively correleted to othe explanatory variable.For example, Manufacuring process 32 is negatively correlated with manufacturing process 13.