Breast\_cancer.R

#setwd("")  
load("BreastCancer.RData")  
  
  
#This dataset is interesting because many of the predictors contain missing values   
#and most rows of the dataset have at least one missing value  
  
  
  
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(RANN)  
myControl<- trainControl(method = "cv", number = 10, verboseIter = TRUE)  
####################################################################  
#Median Uputation  
####################################################################  
# Apply median imputation: model  
model <- train(  
 x = breast\_cancer\_x, y = breast\_cancer\_y,  
 method = "glm",  
 trControl = myControl,  
 preProcess = "medianImpute"  
)

## + Fold01: parameter=none   
## - Fold01: parameter=none   
## + Fold02: parameter=none   
## - Fold02: parameter=none   
## + Fold03: parameter=none   
## - Fold03: parameter=none   
## + Fold04: parameter=none   
## - Fold04: parameter=none   
## + Fold05: parameter=none   
## - Fold05: parameter=none   
## + Fold06: parameter=none   
## - Fold06: parameter=none   
## + Fold07: parameter=none   
## - Fold07: parameter=none   
## + Fold08: parameter=none   
## - Fold08: parameter=none   
## + Fold09: parameter=none   
## - Fold09: parameter=none   
## + Fold10: parameter=none   
## - Fold10: parameter=none   
## Aggregating results  
## Fitting final model on full training set

# Print model to console  
model

## Generalized Linear Model   
##   
## 699 samples  
## 9 predictor  
## 2 classes: 'benign', 'malignant'   
##   
## Pre-processing: median imputation (9)   
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 629, 629, 630, 629, 630, 628, ...   
## Resampling results:  
##   
## Accuracy Kappa   
## 0.9571405 0.9051944

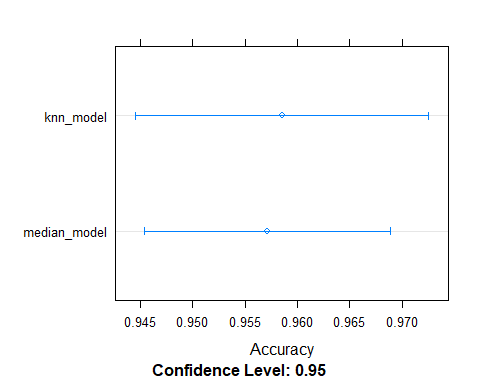
####################################################################  
#KNN imputatin  
####################################################################  
# Apply KNN imputation: model2  
model2 <- train(  
 x = breast\_cancer\_x, y = breast\_cancer\_y,  
 method = "glm",  
 trControl = myControl,  
 preProcess = "knnImpute"  
)

## + Fold01: parameter=none   
## - Fold01: parameter=none   
## + Fold02: parameter=none   
## - Fold02: parameter=none   
## + Fold03: parameter=none   
## - Fold03: parameter=none   
## + Fold04: parameter=none   
## - Fold04: parameter=none   
## + Fold05: parameter=none   
## - Fold05: parameter=none   
## + Fold06: parameter=none   
## - Fold06: parameter=none   
## + Fold07: parameter=none   
## - Fold07: parameter=none   
## + Fold08: parameter=none   
## - Fold08: parameter=none   
## + Fold09: parameter=none   
## - Fold09: parameter=none   
## + Fold10: parameter=none   
## - Fold10: parameter=none   
## Aggregating results  
## Fitting final model on full training set

# Print model to console  
model2

## Generalized Linear Model   
##   
## 699 samples  
## 9 predictor  
## 2 classes: 'benign', 'malignant'   
##   
## Pre-processing: nearest neighbor imputation (9), centered (9), scaled (9)   
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 629, 629, 629, 629, 629, 630, ...   
## Resampling results:  
##   
## Accuracy Kappa   
## 0.9585277 0.9083246

####################################################################  
#Compare KNN and median imputation  
####################################################################  
median\_model <- model  
knn\_model <- model2  
resamples <- resamples(x = list(median\_model = median\_model, knn\_model = knn\_model))  
#Plot to see  
dotplot(resamples, metric = "Accuracy")



#knn model is slightly better.  
  
  
  
  
####################################################################  
#Combining preprocessing methods  
####################################################################  
# Fit glm with median imputation: model1  
model1 <- train(  
 x = breast\_cancer\_x, y = breast\_cancer\_y,  
 method = "glm",  
 trControl = myControl,  
 preProcess = "medianImpute"  
)

## + Fold01: parameter=none   
## - Fold01: parameter=none   
## + Fold02: parameter=none   
## - Fold02: parameter=none   
## + Fold03: parameter=none   
## - Fold03: parameter=none   
## + Fold04: parameter=none   
## - Fold04: parameter=none   
## + Fold05: parameter=none   
## - Fold05: parameter=none   
## + Fold06: parameter=none   
## - Fold06: parameter=none   
## + Fold07: parameter=none   
## - Fold07: parameter=none   
## + Fold08: parameter=none   
## - Fold08: parameter=none   
## + Fold09: parameter=none   
## - Fold09: parameter=none   
## + Fold10: parameter=none   
## - Fold10: parameter=none   
## Aggregating results  
## Fitting final model on full training set

# Print model1  
model1

## Generalized Linear Model   
##   
## 699 samples  
## 9 predictor  
## 2 classes: 'benign', 'malignant'   
##   
## Pre-processing: median imputation (9)   
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 629, 629, 629, 629, 630, 629, ...   
## Resampling results:  
##   
## Accuracy Kappa   
## 0.9585904 0.9085526

# Fit glm with median imputation and standardization: model2  
model2 <- train(  
 x = breast\_cancer\_x, y = breast\_cancer\_y,  
 method = "glm",  
 trControl = myControl,  
 preProcess = c("medianImpute", "center", "scale")  
)

## + Fold01: parameter=none   
## - Fold01: parameter=none   
## + Fold02: parameter=none   
## - Fold02: parameter=none   
## + Fold03: parameter=none   
## - Fold03: parameter=none   
## + Fold04: parameter=none   
## - Fold04: parameter=none   
## + Fold05: parameter=none   
## - Fold05: parameter=none   
## + Fold06: parameter=none   
## - Fold06: parameter=none   
## + Fold07: parameter=none   
## - Fold07: parameter=none   
## + Fold08: parameter=none   
## - Fold08: parameter=none   
## + Fold09: parameter=none   
## - Fold09: parameter=none   
## + Fold10: parameter=none   
## - Fold10: parameter=none   
## Aggregating results  
## Fitting final model on full training set

# Print model2  
model2

## Generalized Linear Model   
##   
## 699 samples  
## 9 predictor  
## 2 classes: 'benign', 'malignant'   
##   
## Pre-processing: median imputation (9), centered (9), scaled (9)   
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 629, 629, 629, 629, 629, 629, ...   
## Resampling results:  
##   
## Accuracy Kappa  
## 0.9599775 0.911