Model Evaluation - Spotcheck

source("~/MCD/MCD/m.R")  
  
knitr::opts\_chunk$set(fig.width = 16, fig.height = 12)  
options(width = 120)  
# g\_startCluster()  
  
library(mlbench)  
data(PimaIndiansDiabetes)  
  
print("Loaded")

## [1] "Loaded"

## Train multi model

mm\_pima <-  
 m\_raw(PimaIndiansDiabetes) %>%  
 m\_response(~ diabetes) %>%  
 m\_repeated\_cross\_validation(number = 10, repeats = 3) %>%  
 m\_accuracy\_metric() %>%  
 m\_cross(  
 cart = ~ m %>% m\_classification\_and\_regression\_trees(),  
 lda = ~ m %>% m\_linear\_discriminant\_analysis(),  
 svm = ~ m %>% m\_support\_vector\_machine\_radial\_basis\_kernel(),  
 knn = ~ m %>% m\_k\_nearest\_neighbors(),  
 rpart = ~ m %>% m\_random\_forest()  
 ) %>%  
 m\_train() %>%  
 m\_results() %>%  
 m\_resamples() %>%  
 m\_run()

## [1] "cart"  
## [1] "lda"  
## [1] "svm"  
## [1] "knn"  
## [1] "rpart"

print("Trained")

## [1] "Trained"

## Evaluate multi model

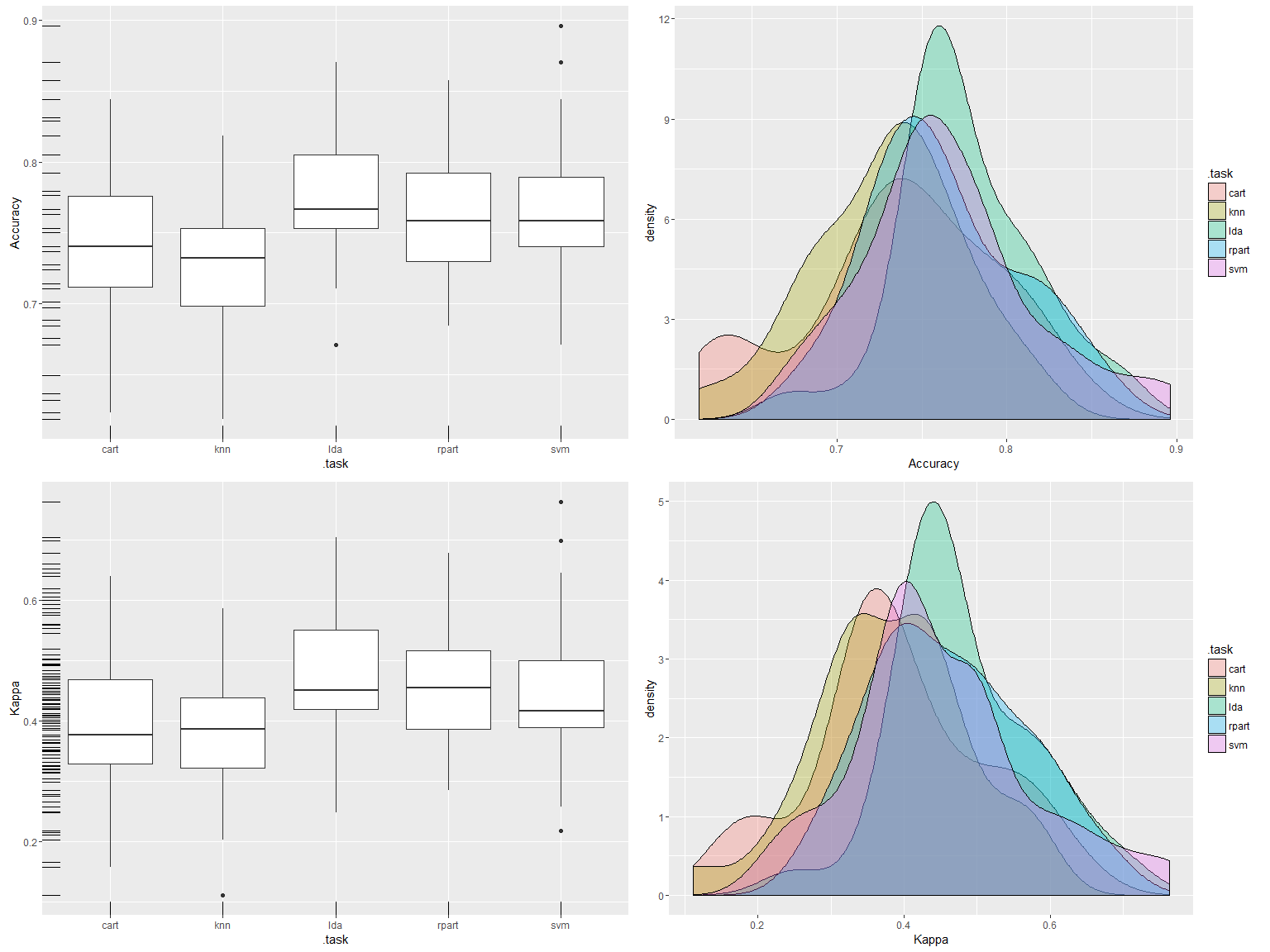
pima\_resamples <- mm\_pima %>% map("train") %>% caret::resamples()  
pima\_resamples %>% summary()

##   
## Call:  
## summary.resamples(object = .)  
##   
## Models: cart, lda, svm, knn, rpart   
## Number of resamples: 30   
##   
## Accuracy   
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  
## cart 0.6234 0.7115 0.7403 0.7382 0.7760 0.8442 0  
## lda 0.6711 0.7532 0.7662 0.7759 0.8052 0.8701 0  
## svm 0.6711 0.7403 0.7582 0.7651 0.7890 0.8961 0  
## knn 0.6184 0.6984 0.7321 0.7299 0.7532 0.8182 0  
## rpart 0.6842 0.7297 0.7582 0.7625 0.7922 0.8571 0  
##   
## Kappa   
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  
## cart 0.1585 0.3296 0.3765 0.3934 0.4685 0.6393 0  
## lda 0.2484 0.4196 0.4516 0.4801 0.5512 0.7048 0  
## svm 0.2187 0.3889 0.4167 0.4520 0.5003 0.7638 0  
## knn 0.1113 0.3228 0.3867 0.3819 0.4382 0.5867 0  
## rpart 0.2853 0.3860 0.4553 0.4613 0.5169 0.6781 0

diff(pima\_resamples) %>% summary()

##   
## Call:  
## summary.diff.resamples(object = .)  
##   
## p-value adjustment: bonferroni   
## Upper diagonal: estimates of the difference  
## Lower diagonal: p-value for H0: difference = 0  
##   
## Accuracy   
## cart lda svm knn rpart   
## cart -0.037759 -0.026908 0.008248 -0.024351  
## lda 0.0050068 0.010851 0.046007 0.013409  
## svm 0.0919580 0.3390336 0.035156 0.002558  
## knn 1.0000000 1.218e-05 0.0007092 -0.032599  
## rpart 0.0974572 0.2416366 1.0000000 0.0016066   
##   
## Kappa   
## cart lda svm knn rpart   
## cart -0.086692 -0.058612 0.011552 -0.067895  
## lda 0.0015478 0.028079 0.098243 0.018796  
## svm 0.0839950 0.2217402 0.070164 -0.009283  
## knn 1.0000000 4.122e-05 0.0054686 -0.079447  
## rpart 0.0129477 1.0000000 1.0000000 0.0006918

mm\_pima %>% m\_combine(data = ~resamples) %>% m\_condition(~.task) %>% m\_boxplot() %>% m\_density() %>% m\_run() %>% m\_layout()



## Train transforms

## PimaIndiansDiabetes %>% mutate\_if(function(x) class(x) == "numeric", function(x) ifelse(x > 0, x, NA)) %>% m\_get\_variable\_exploration()  
  
mm\_pima\_t <-  
 m\_raw(PimaIndiansDiabetes) %>%  
 m\_response(~ diabetes) %>%  
 m\_repeated\_cross\_validation(number = 10, repeats = 3) %>%  
 m\_accuracy\_metric() %>%  
 m\_standardize\_transform() %>%  
 m\_cross(  
 lda = ~ m %>% m\_linear\_discriminant\_analysis(),  
 svm = ~ m %>% m\_support\_vector\_machine\_radial\_basis\_kernel()  
 ) %>%  
 m\_cross(  
 base = ~ m,  
 boxcox = ~ m %>% m\_boxcox\_transform(),  
 yeo = ~ m %>% m\_yeojohnson\_transform(),  
 pca = ~ m %>% m\_pca\_transform(),  
 box\_pca = ~ m %>% m\_boxcox\_transform() %>% m\_pca\_transform()  
 ) %>%  
 m\_train() %>%  
 m\_results() %>%  
 m\_resamples() %>%  
 m\_run()

## [1] "lda.base"  
## [1] "lda.boxcox"  
## [1] "lda.yeo"

## Warning in estimateTransform.default(X, Y, weights, family, start, method, : Convergence failure: return code = 52

## [1] "lda.pca"  
## [1] "lda.box\_pca"  
## [1] "svm.base"  
## [1] "svm.boxcox"  
## [1] "svm.yeo"

## Warning in estimateTransform.default(X, Y, weights, family, start, method, : Convergence failure: return code = 52  
  
## Warning in estimateTransform.default(X, Y, weights, family, start, method, : Convergence failure: return code = 52  
  
## Warning in estimateTransform.default(X, Y, weights, family, start, method, : Convergence failure: return code = 52

## [1] "svm.pca"  
## [1] "svm.box\_pca"

print("Trained")

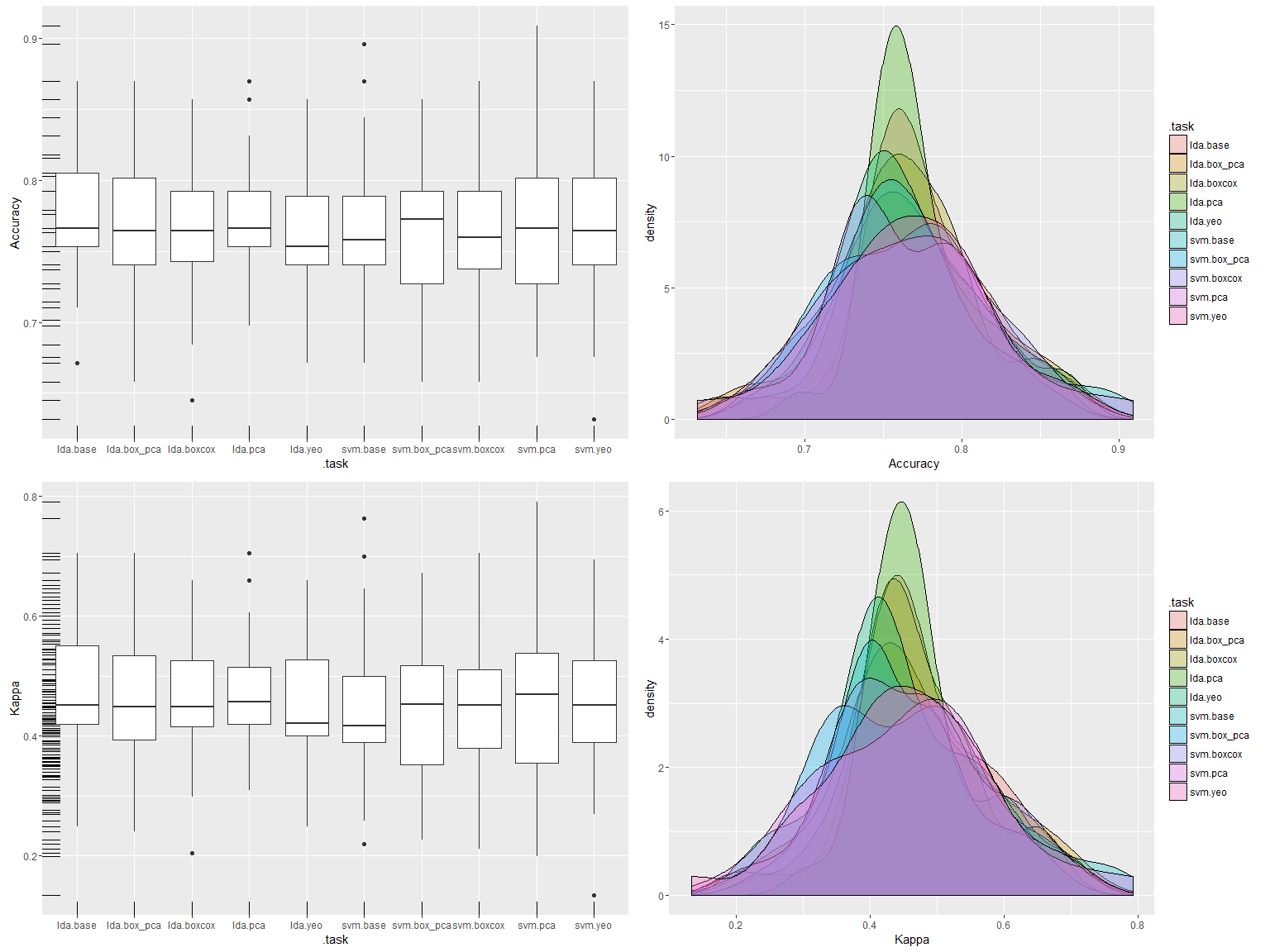
## [1] "Trained"

## Evaluate transforms

mm\_pima\_t %>% map("train") %>% caret::resamples() %>% summary()

##   
## Call:  
## summary.resamples(object = .)  
##   
## Models: lda.base, lda.boxcox, lda.yeo, lda.pca, lda.box\_pca, svm.base, svm.boxcox, svm.yeo, svm.pca, svm.box\_pca   
## Number of resamples: 30   
##   
## Accuracy   
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  
## lda.base 0.6711 0.7532 0.7662 0.7759 0.8052 0.8701 0  
## lda.boxcox 0.6447 0.7427 0.7647 0.7659 0.7922 0.8571 0  
## lda.yeo 0.6711 0.7403 0.7532 0.7621 0.7890 0.8571 0  
## lda.pca 0.6974 0.7532 0.7662 0.7751 0.7922 0.8701 0  
## lda.box\_pca 0.6579 0.7403 0.7647 0.7677 0.8019 0.8701 0  
## svm.base 0.6711 0.7403 0.7582 0.7651 0.7890 0.8961 0  
## svm.boxcox 0.6579 0.7377 0.7597 0.7672 0.7922 0.8701 0  
## svm.yeo 0.6316 0.7403 0.7647 0.7638 0.8019 0.8701 0  
## svm.pca 0.6753 0.7273 0.7662 0.7682 0.8019 0.9091 0  
## svm.box\_pca 0.6579 0.7273 0.7727 0.7646 0.7922 0.8571 0  
##   
## Kappa   
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  
## lda.base 0.2484 0.4196 0.4516 0.4801 0.5512 0.7048 0  
## lda.boxcox 0.2034 0.4148 0.4491 0.4604 0.5255 0.6602 0  
## lda.yeo 0.2484 0.4007 0.4205 0.4506 0.5275 0.6602 0  
## lda.pca 0.3085 0.4196 0.4563 0.4790 0.5143 0.7048 0  
## lda.box\_pca 0.2400 0.3935 0.4480 0.4659 0.5338 0.7048 0  
## svm.base 0.2187 0.3889 0.4167 0.4520 0.5003 0.7638 0  
## svm.boxcox 0.2109 0.3794 0.4513 0.4591 0.5103 0.7048 0  
## svm.yeo 0.1336 0.3886 0.4513 0.4500 0.5258 0.6940 0  
## svm.pca 0.1983 0.3543 0.4687 0.4591 0.5386 0.7915 0  
## svm.box\_pca 0.2257 0.3518 0.4520 0.4489 0.5169 0.6723 0

mm\_pima\_t %>% m\_combine(data = ~resamples) %>% m\_condition(~.task) %>% m\_boxplot() %>% m\_density() %>% m\_run() %>% m\_layout()



## Evaluate cleaning

mm\_pima\_cl <-  
 m\_raw(PimaIndiansDiabetes) %>%  
 m\_response(~ diabetes) %>%  
 m\_repeated\_cross\_validation(number = 10, repeats = 3) %>%  
 m\_accuracy\_metric() %>%  
 m\_cross(  
 base = ~ m,  
 clean\_knn = ~ m %>% m\_mutate(~ mutate\_all(data, function(x) ifelse(x > 0, x, NA))) %>% m\_knn\_impute()  
 ) %>%  
 m\_cross(  
 cart = ~ m %>% m\_classification\_and\_regression\_trees(),  
 lda = ~m %>% m\_linear\_discriminant\_analysis(),  
 svm = ~ m %>% m\_support\_vector\_machine\_radial\_basis\_kernel(),  
 knn = ~ m %>% m\_k\_nearest\_neighbors(),  
 rpart = ~ m %>% m\_random\_forest()  
 ) %>%  
 m\_train() %>%  
 m\_results() %>%  
 m\_resamples() %>%  
 m\_run()

## [1] "base.cart"  
## [1] "base.lda"  
## [1] "base.svm"  
## [1] "base.knn"  
## [1] "base.rpart"  
## [1] "clean\_knn.cart"  
## [1] "clean\_knn.lda"  
## [1] "clean\_knn.svm"  
## [1] "clean\_knn.knn"  
## [1] "clean\_knn.rpart"

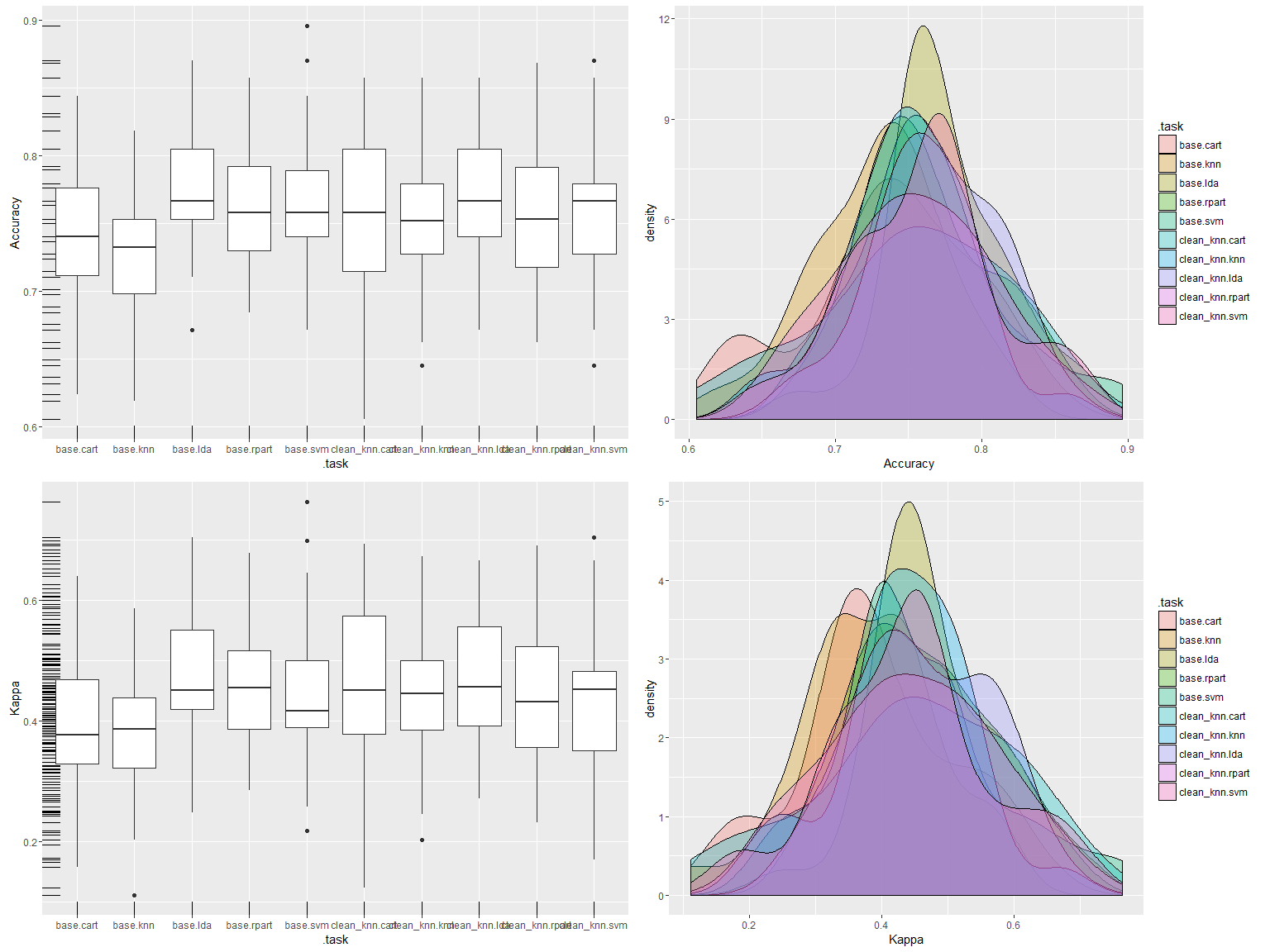
print("Trained")

## [1] "Trained"

mm\_pima\_cl %>% map("train") %>% caret::resamples() %>% summary()

##   
## Call:  
## summary.resamples(object = .)  
##   
## Models: base.cart, base.lda, base.svm, base.knn, base.rpart, clean\_knn.cart, clean\_knn.lda, clean\_knn.svm, clean\_knn.knn, clean\_knn.rpart   
## Number of resamples: 30   
##   
## Accuracy   
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  
## base.cart 0.6234 0.7115 0.7403 0.7382 0.7760 0.8442 0  
## base.lda 0.6711 0.7532 0.7662 0.7759 0.8052 0.8701 0  
## base.svm 0.6711 0.7403 0.7582 0.7651 0.7890 0.8961 0  
## base.knn 0.6184 0.6984 0.7321 0.7299 0.7532 0.8182 0  
## base.rpart 0.6842 0.7297 0.7582 0.7625 0.7922 0.8571 0  
## clean\_knn.cart 0.6053 0.7143 0.7582 0.7529 0.8052 0.8571 0  
## clean\_knn.lda 0.6711 0.7403 0.7662 0.7690 0.8052 0.8571 0  
## clean\_knn.svm 0.6447 0.7273 0.7662 0.7612 0.7792 0.8701 0  
## clean\_knn.knn 0.6447 0.7273 0.7516 0.7495 0.7792 0.8571 0  
## clean\_knn.rpart 0.6623 0.7175 0.7532 0.7556 0.7915 0.8684 0  
##   
## Kappa   
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  
## base.cart 0.1585 0.3296 0.3765 0.3934 0.4685 0.6393 0  
## base.lda 0.2484 0.4196 0.4516 0.4801 0.5512 0.7048 0  
## base.svm 0.2187 0.3889 0.4167 0.4520 0.5003 0.7638 0  
## base.knn 0.1113 0.3228 0.3867 0.3819 0.4382 0.5867 0  
## base.rpart 0.2853 0.3860 0.4553 0.4613 0.5169 0.6781 0  
## clean\_knn.cart 0.1231 0.3785 0.4515 0.4482 0.5747 0.6941 0  
## clean\_knn.lda 0.2716 0.3915 0.4564 0.4647 0.5570 0.6664 0  
## clean\_knn.svm 0.1696 0.3514 0.4520 0.4436 0.4821 0.7048 0  
## clean\_knn.knn 0.2034 0.3846 0.4462 0.4342 0.5003 0.6723 0  
## clean\_knn.rpart 0.2324 0.3570 0.4316 0.4454 0.5235 0.6906 0

mm\_pima\_cl %>% m\_combine(data = ~resamples) %>% m\_condition(~.task) %>% m\_boxplot() %>% m\_density() %>% m\_run() %>% m\_layout()



mm\_pima\_final <-  
 m\_raw(PimaIndiansDiabetes) %>%  
 m\_response(~diabetes) %>%  
 m\_repeated\_cross\_validation(number = 10, repeats = 3) %>%  
 m\_accuracy\_metric() %>%  
 m\_mutate(~mutate\_all(data, function(x) ifelse(x > 0, x, NA))) %>%  
 m\_knn\_impute() %>%  
 m\_standardize\_transform() %>%  
 m\_boxcox\_transform() %>%  
 m\_pca\_transform() %>%  
 m\_linear\_discriminant\_analysis() %>%  
 m\_train() %>%  
 m\_resamples() %>%  
 m\_run()

## [1] "PimaIndiansDiabetes"

print("Trained")

## [1] "Trained"

mm\_pima\_final$resamples

## Accuracy Kappa Resample  
## 1 0.8051948 0.5367028 Fold01.Rep1  
## 2 0.7922078 0.5276074 Fold02.Rep1  
## 3 0.7105263 0.3685801 Fold03.Rep1  
## 4 0.7662338 0.4866667 Fold04.Rep1  
## 5 0.7922078 0.5437037 Fold05.Rep1  
## 6 0.6842105 0.3111782 Fold06.Rep1  
## 7 0.7402597 0.3765182 Fold07.Rep1  
## 8 0.7532468 0.4237889 Fold08.Rep1  
## 9 0.8441558 0.6457055 Fold09.Rep1  
## 10 0.7532468 0.4237889 Fold10.Rep1  
## 11 0.8571429 0.6889460 Fold01.Rep2  
## 12 0.7662338 0.4590164 Fold02.Rep2  
## 13 0.7662338 0.4777694 Fold03.Rep2  
## 14 0.6447368 0.1882911 Fold04.Rep2  
## 15 0.8051948 0.5531915 Fold05.Rep2  
## 16 0.7662338 0.4590164 Fold06.Rep2  
## 17 0.7272727 0.4255773 Fold07.Rep2  
## 18 0.7763158 0.4790323 Fold08.Rep2  
## 19 0.7272727 0.3391908 Fold09.Rep2  
## 20 0.8571429 0.6780692 Fold10.Rep2  
## 21 0.6842105 0.2425249 Fold01.Rep3  
## 22 0.8051948 0.5685469 Fold02.Rep3  
## 23 0.7532468 0.4534927 Fold03.Rep3  
## 24 0.6973684 0.3792614 Fold04.Rep3  
## 25 0.7532468 0.4534927 Fold05.Rep3  
## 26 0.8441558 0.6327504 Fold06.Rep3  
## 27 0.7922078 0.5276074 Fold07.Rep3  
## 28 0.6883117 0.2786885 Fold08.Rep3  
## 29 0.7532468 0.4439377 Fold09.Rep3  
## 30 0.8181818 0.5715421 Fold10.Rep3

m\_data(mm\_pima\_final$resamples) %>% m\_density() %>% m\_run() %>% m\_layout()

