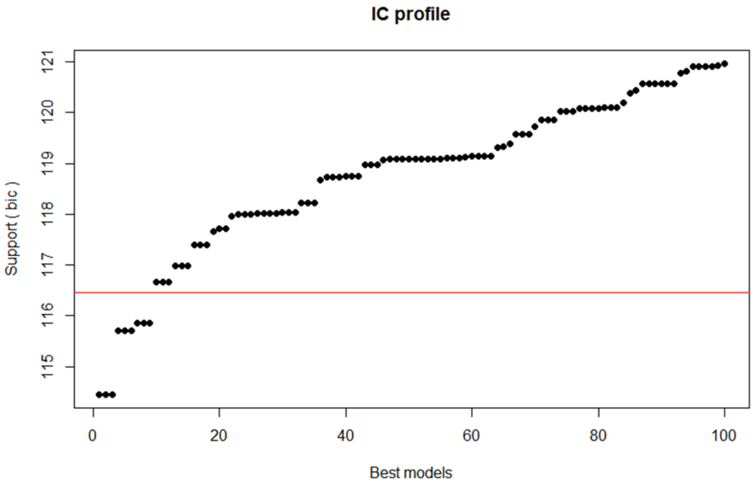
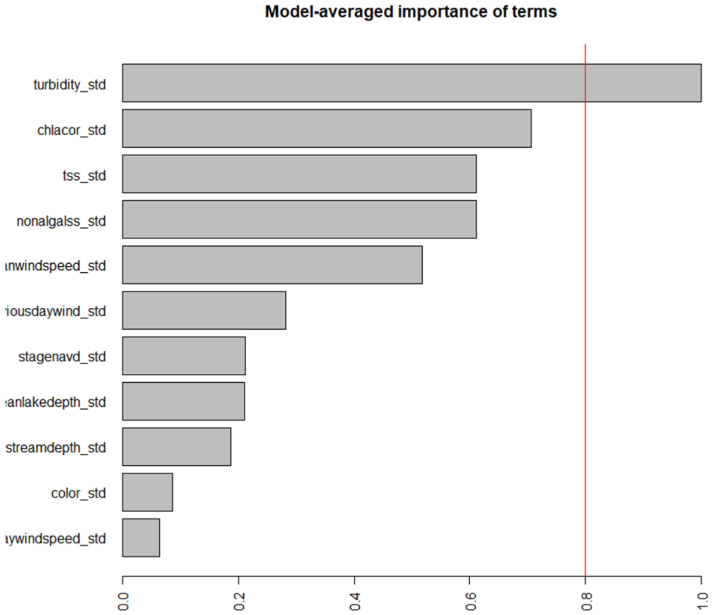
Advanced R: Statistical Machine Learning

Dan Schmutz, MS
Chief Environmental Scientist

Zoom Workshop for SJRWMD
September 24, 2020

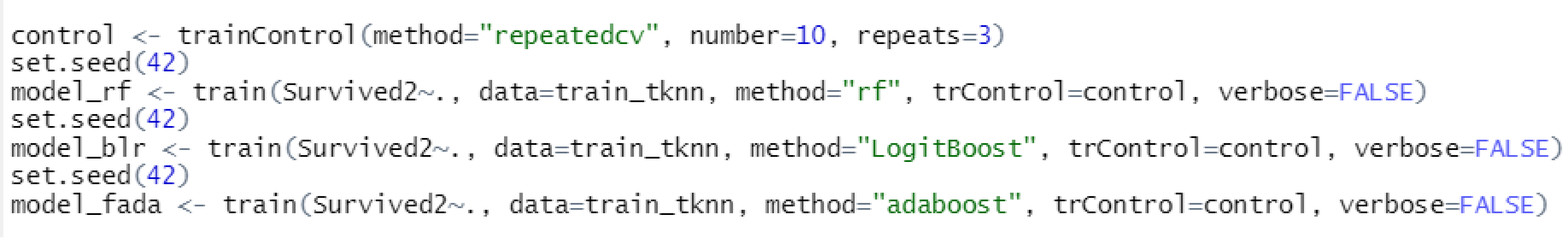




Comparing Models

caret offers common interface to large number of models

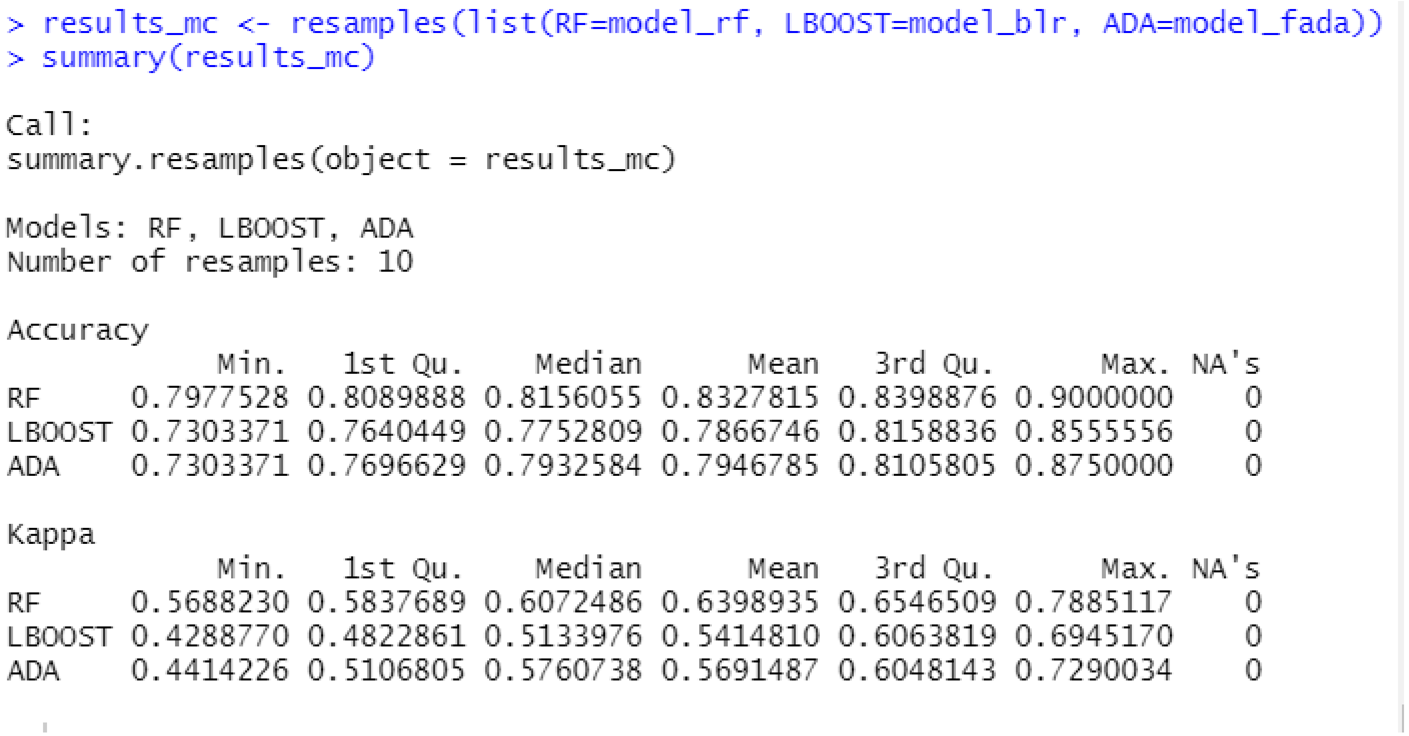
- 238 different models available to the train function of caret!
- I’m picking 3 here to evaluate on the Titanic processed train



```
control <- trainControl(method="repeatedcv", number=10, repeats=3)
set.seed(42)
model_rf <- train(Survived2~., data=train_tknn, method="rf", trControl=control, verbose=FALSE)
set.seed(42)
model_blr <- train(Survived2~., data=train_tknn, method="LogitBoost", trControl=control, verbose=FALSE)
set.seed(42)
model_fada <- train(Survived2~., data=train_tknn, method="adaboost", trControl=control, verbose=FALSE)
```

Note use of same exact seed before each model resampling to ensure same exact folds.

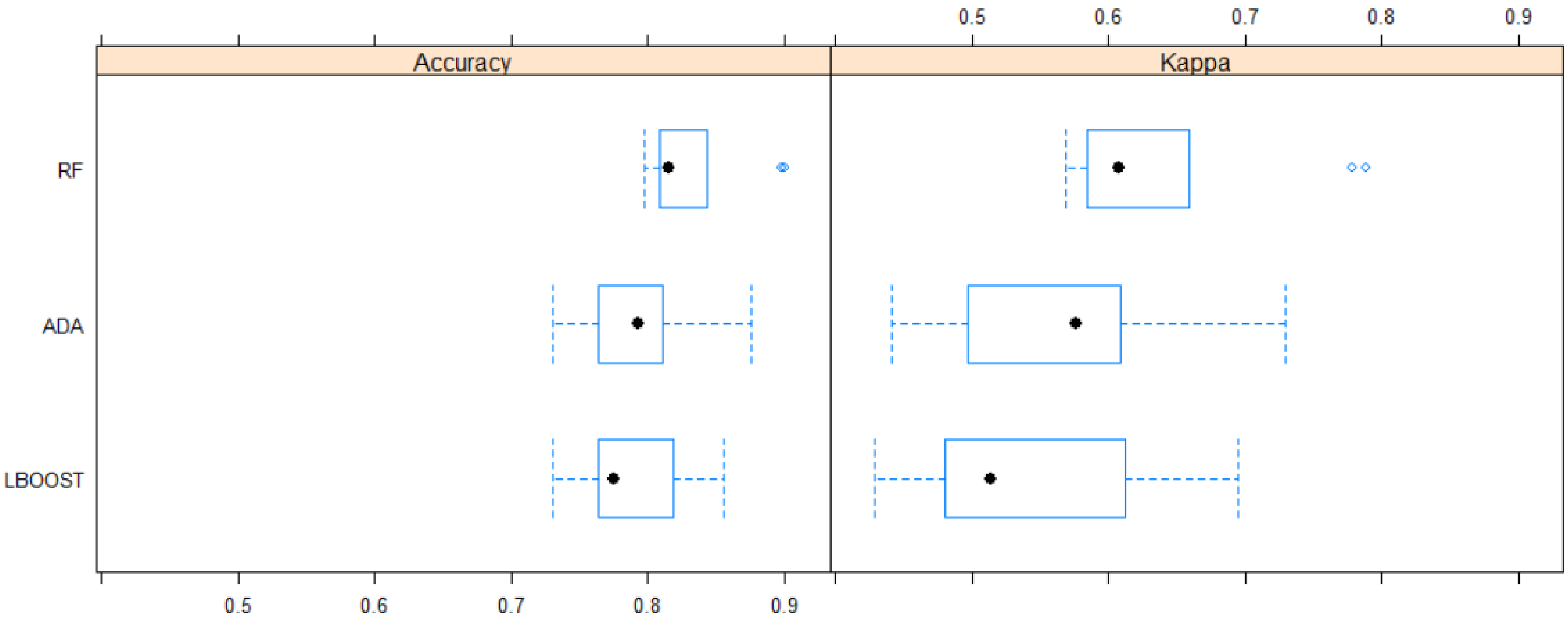
Tabled comparison of resamples



```
> results_mc <- resamples(list(RF=model_rf, LBOOST=model_blr, ADA=model_fada))
> summary(results_mc)

Call:
summary.resamples(object = results_mc)

Models: RF, LBOOST, ADA
Number of resamples: 10

Accuracy
            Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
RF     0.7977528 0.8089888 0.8156055 0.8327815 0.8398876 0.9000000    0
LBOOST 0.7303371 0.7640449 0.7752809 0.7866746 0.8158836 0.8555556    0
ADA    0.7303371 0.7696629 0.7932584 0.7946785 0.8105805 0.8750000    0

Kappa
            Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
RF     0.5688230 0.5837689 0.6072486 0.6398935 0.6546509 0.7885117    0
LBOOST 0.4288770 0.4822861 0.5133976 0.5414810 0.6063819 0.6945170    0
ADA    0.4414226 0.5106805 0.5760738 0.5691487 0.6048143 0.7290034    0
```

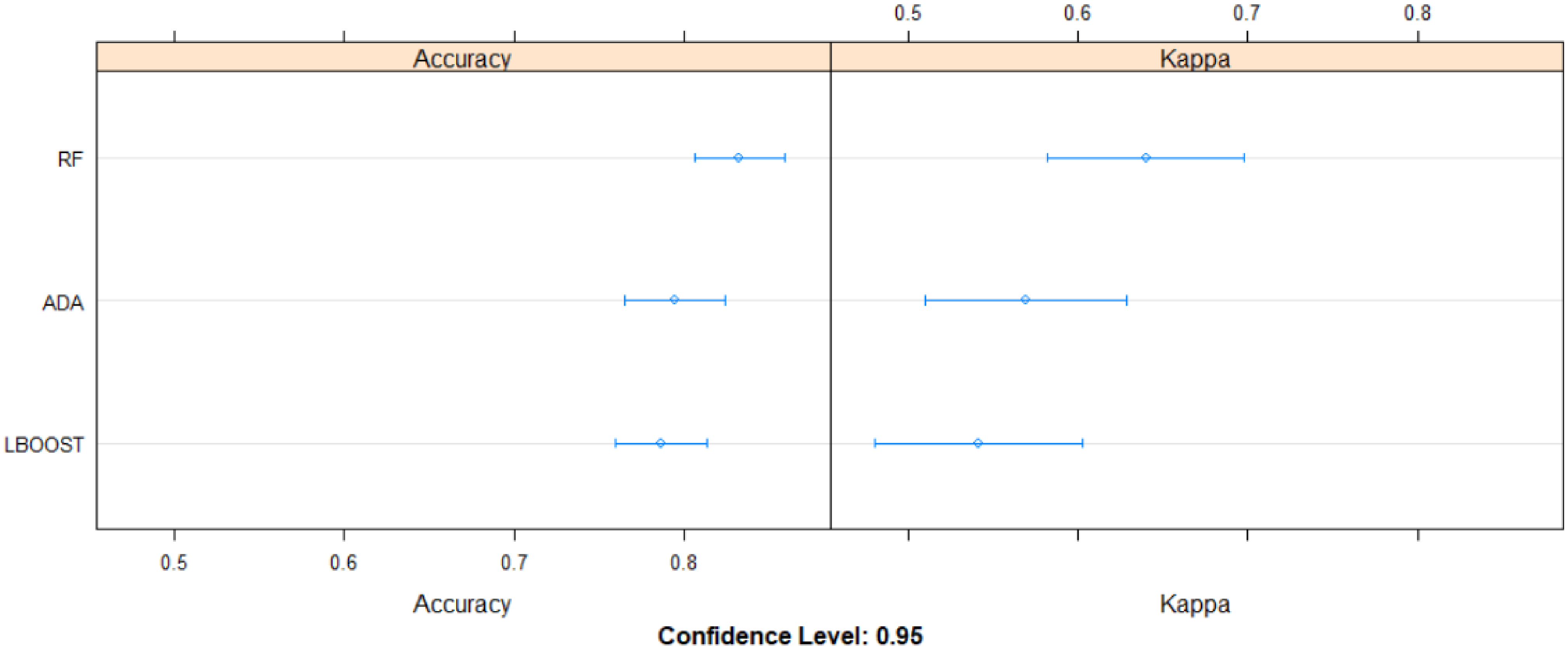
Box and whisker comparison of resamples

- bwplot(results_mc)



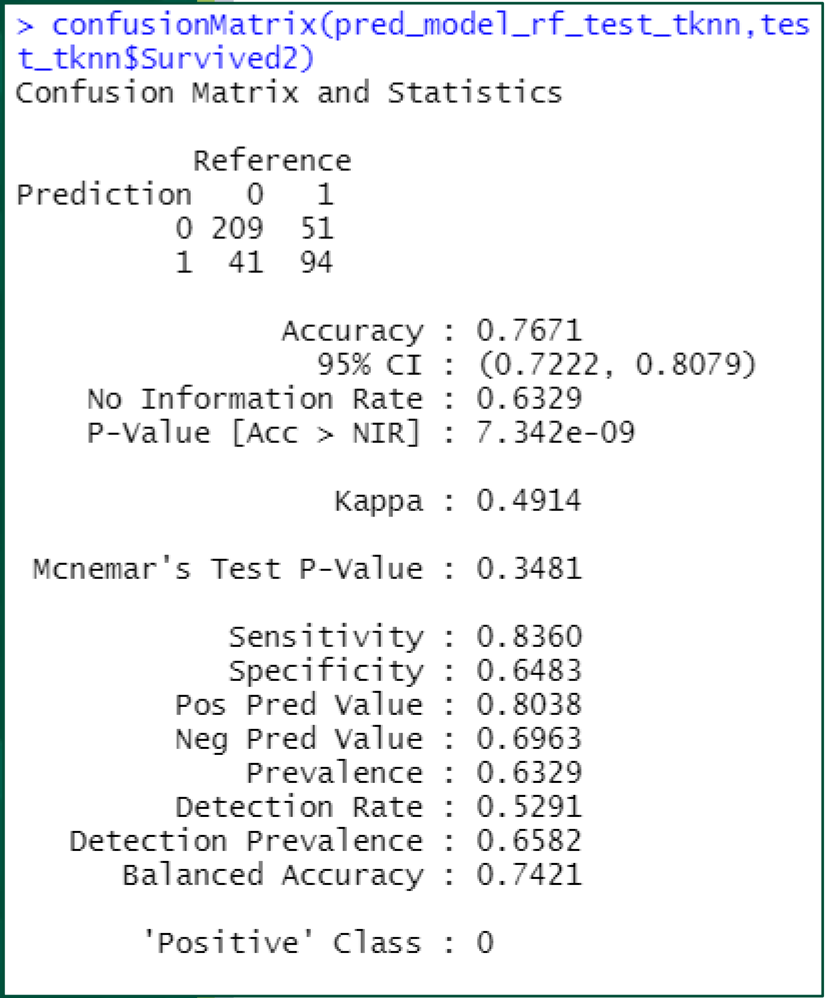
Dotplot comparison of resamples

- dotplot(results_mc)



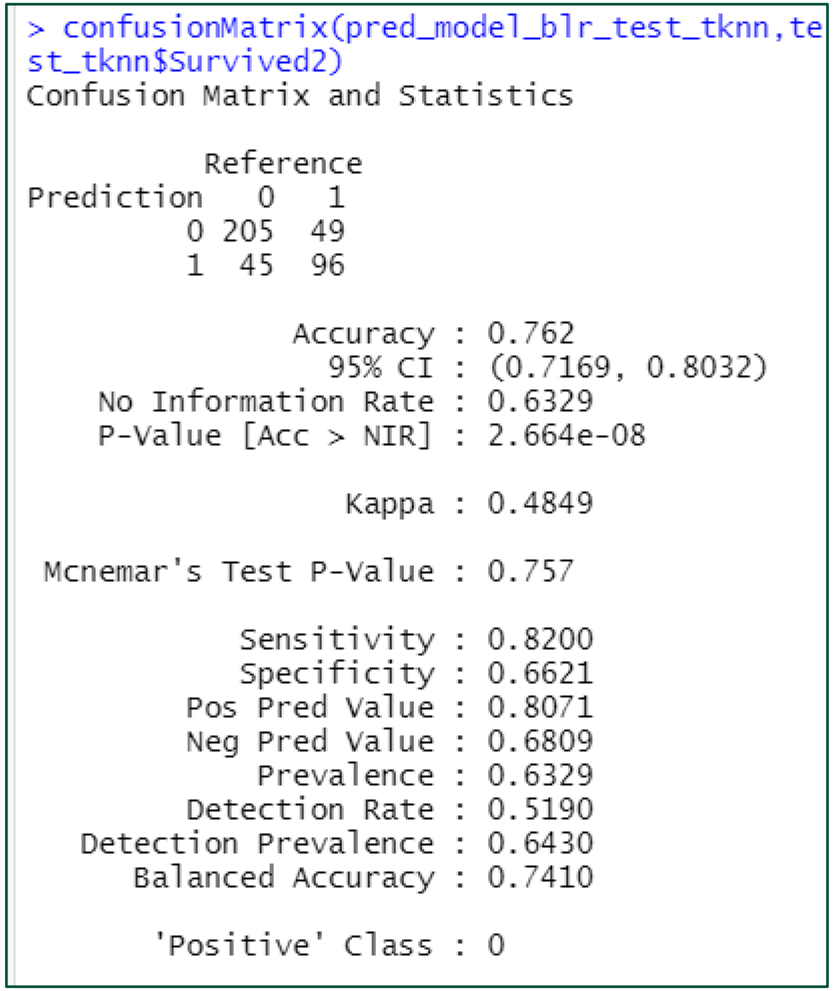
Performance on test dat

randomForest



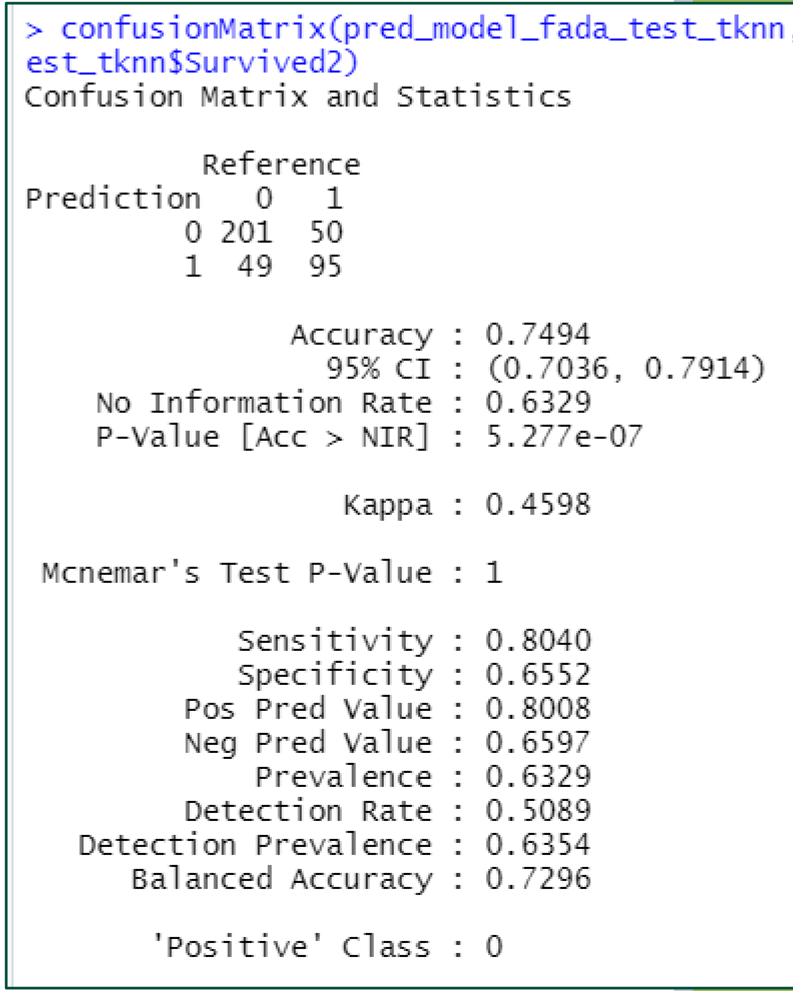
```
> confusionMatrix(pred_model_rf_test_tknn,tes
t_tknn$Survived2)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 209  51
         1  41  94

               Accuracy : 0.7671
                 95% CI : (0.7222, 0.8079)
    No Information Rate : 0.6329
    P-Value [Acc > NIR] : 7.342e-09

                  Kappa : 0.4914

 Mcnemar's Test P-Value : 0.3481

            Sensitivity : 0.8360
            Specificity : 0.6483
         Pos Pred Value : 0.8038
         Neg Pred Value : 0.6963
             Prevalence : 0.6329
         Detection Rate : 0.5291
   Detection Prevalence : 0.6582
      Balanced Accuracy : 0.7421

       'Positive' Class : 0
```

logitboost



```
> confusionMatrix(pred_model_blr_test_tknn,te
st_tknn$Survived2)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 205  49
         1  45  96

               Accuracy : 0.762
                 95% CI : (0.7169, 0.8032)
    No Information Rate : 0.6329
    P-Value [Acc > NIR] : 2.664e-08

                  Kappa : 0.4849

 Mcnemar's Test P-Value : 0.757

            Sensitivity : 0.8200
            Specificity : 0.6621
         Pos Pred Value : 0.8071
         Neg Pred Value : 0.6809
             Prevalence : 0.6329
         Detection Rate : 0.5190
   Detection Prevalence : 0.6430
      Balanced Accuracy : 0.7410

       'Positive' Class : 0
```

adaboost



```
> confusionMatrix(pred_model_fada_test_tknn
est_tknn$Survived2)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 201  50
         1  49  95

               Accuracy : 0.7494
                 95% CI : (0.7036, 0.7914)
    No Information Rate : 0.6329
    P-Value [Acc > NIR] : 5.277e-07

                  Kappa : 0.4598

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8040
            Specificity : 0.6552
         Pos Pred Value : 0.8008
         Neg Pred Value : 0.6597
             Prevalence : 0.6329
         Detection Rate : 0.5089
   Detection Prevalence : 0.6354
      Balanced Accuracy : 0.7296

       'Positive' Class : 0
```