Caret-Ensenble

鈴木瑞人
東京大学大学院　新領域創成科学研究科
メディカル情報生命専攻
博士課程1年

```r
install.packages("caret", quiet = TRUE, dependencies=T)
library(caret)
```

```r
install.packages("caretEnsemble", quiet = TRUE, dependencies=T)
library(caretEnsemble)
library(GGally)
```

```
install.packages("gbm", quiet = TRUE, dependencies=T)
library(gbm)
set.seed(123)
folds <- 10
repeats <- 1
ctrl <- trainControl(method = "cv", number = folds, classProbs = TRUE,
  savePredictions = TRUE, summaryFunction = twoClassSummary,
  index = createMultiFolds(churnTrain$churn,
    k = folds, times = repeats))
```

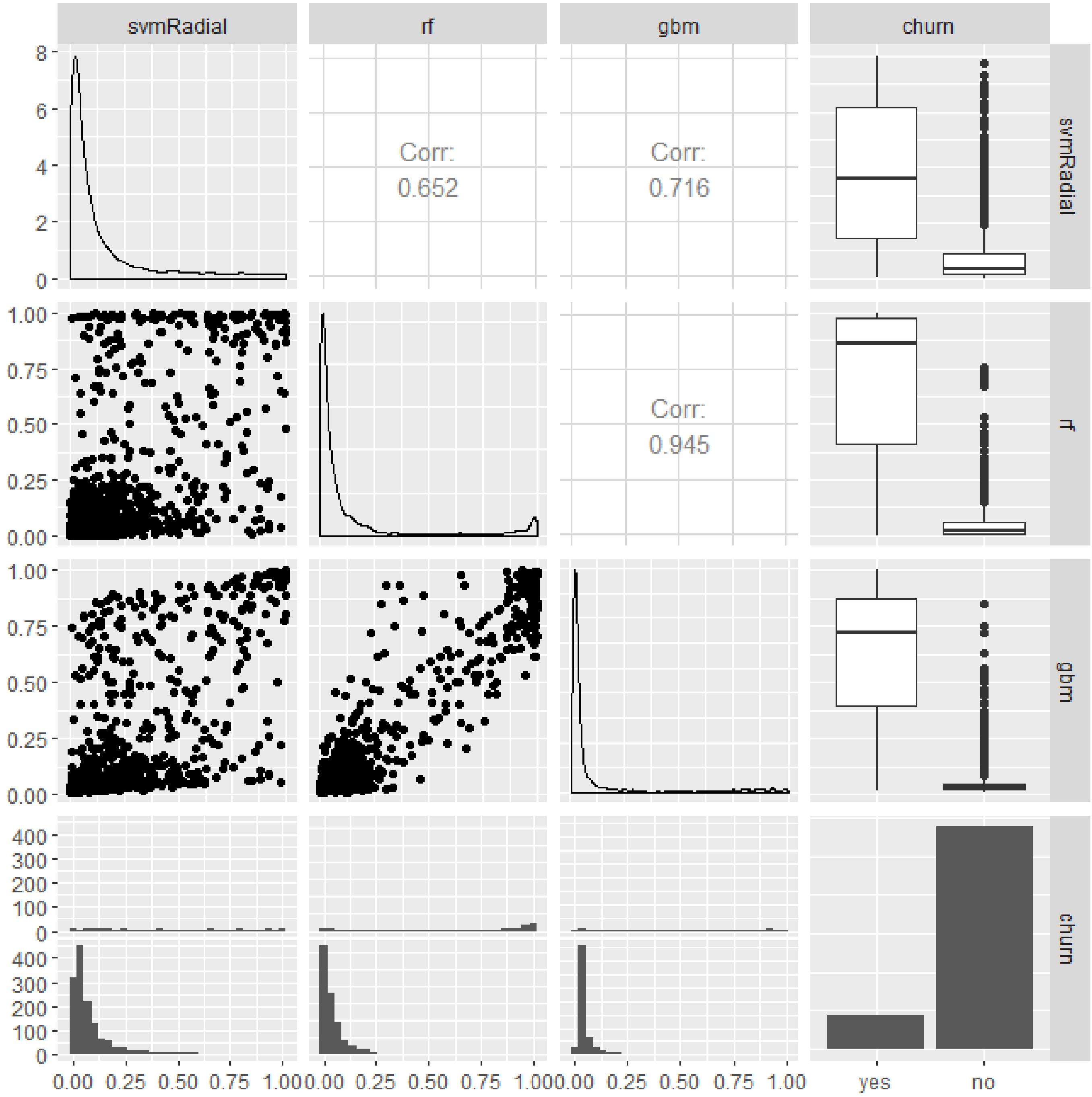
```
model.list <- caretList(churn ~ ., data = churnTrain, metric = "ROC",
trControl = ctrl, methodList = c("svmRadial", "rf", "gbm"), verbose =
FALSE)

model.list
```

```
> model.list
$svmRadial
Support Vector Machines with Radial Basis Function Kernel

3333 samples
  19 predictor
   2 classes: 'yes', 'no'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3000, 3000, 2999, 2999, 3000, 3000, ...
Resampling results across tuning parameters:

  C     ROC        Sens       Spec
  0.25  0.8717166  0.4511480  0.9684211
  0.50  0.8717741  0.4489796  0.9684211
  1.00  0.8718976  0.4531463  0.9684211

Tuning parameter 'sigma' was held constant at a value of 0.00742499
ROC was used to select the optimal model using  the largest value.
The final values used for the model were sigma = 0.00742499 and C = 1.
```

```
$rf
Random Forest

3333 samples
  19 predictor
   2 classes: 'yes', 'no'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3000, 3000, 2999, 2999, 3000, 3000, ...
Resampling results across tuning parameters:

  mtry  ROC        Sens       Spec
   2    0.8985384  0.1096939  1.0000000
  35    0.9089392  0.7452806  0.9880702
  69    0.9060944  0.7329507  0.9870175

ROC was used to select the optimal model using  the largest value.
The final value used for the model was mtry = 35.
```

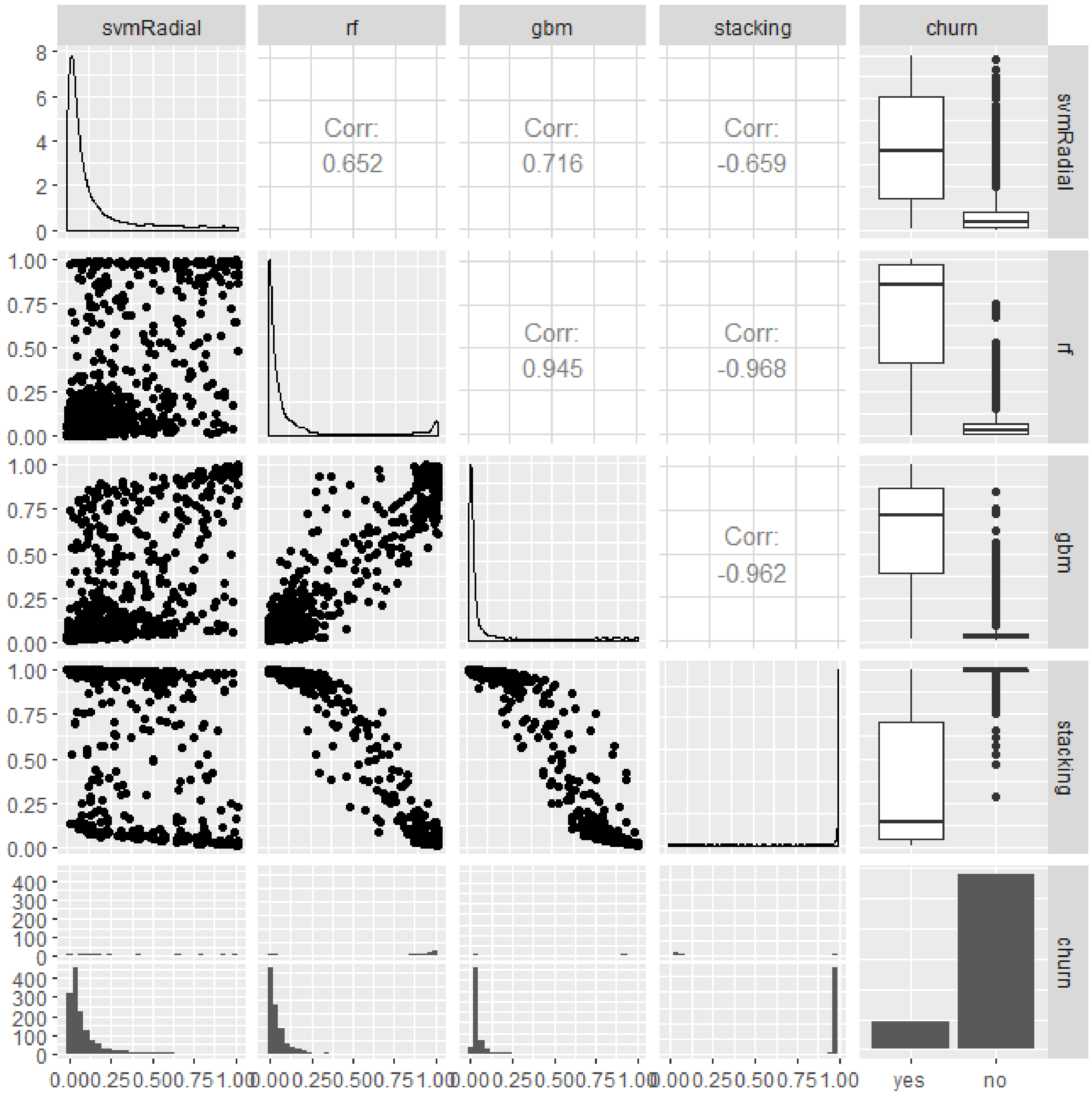
```
$gbm
Stochastic Gradient Boosting

3333 samples
  19 predictor
   2 classes: 'yes', 'no'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3000, 3000, 2999, 2999, 3000, 3000, ...
Resampling results across tuning parameters:

  interaction.depth  n.trees  ROC        Sens       Spec
  1                   50      0.8602553  0.2026786  0.9800000
  1                  100      0.8738977  0.3227891  0.9733333
  1                  150      0.8752181  0.3600340  0.9698246
  2                   50      0.9004939  0.4655187  0.9852632
  2                  100      0.9113458  0.6479592  0.9849123
  2                  150      0.9148320  0.6727466  0.9845614
  3                   50      0.9122482  0.6585034  0.9905263
  3                  100      0.9170824  0.7287415  0.9898246
  3                  150      0.9193063  0.7411565  0.9901754

Tuning parameter 'shrinkage' was held constant at a value of 0.1
Tuning parameter 'n.minobsinnode' was
 held constant at a value of 10
ROC was used to select the optimal model using  the largest value.
The final values used for the model were n.trees = 150, interaction.depth = 3, shrinkage = 0.1
 and n.minobsinnode = 10.

attr(,"class")
[1] "caretList"
```

```
pred.each <- (1 - predict(model.list, churnTest)) %>% as.data.frame
%>%mutate(churn = churnTest$churn)
ggpairs(pred.each)
```


svmRadial
rf
gbm
churn
Corr:
0.652
Corr:
0.716
Corr:
0.945
svmRadial
rf
gbm
churn
0
2
4
6
8
0.00
0.25
0.50
0.75
1.00
0
100
200
300
400
0.00 0.25 0.50 0.75 1.00
yes
no

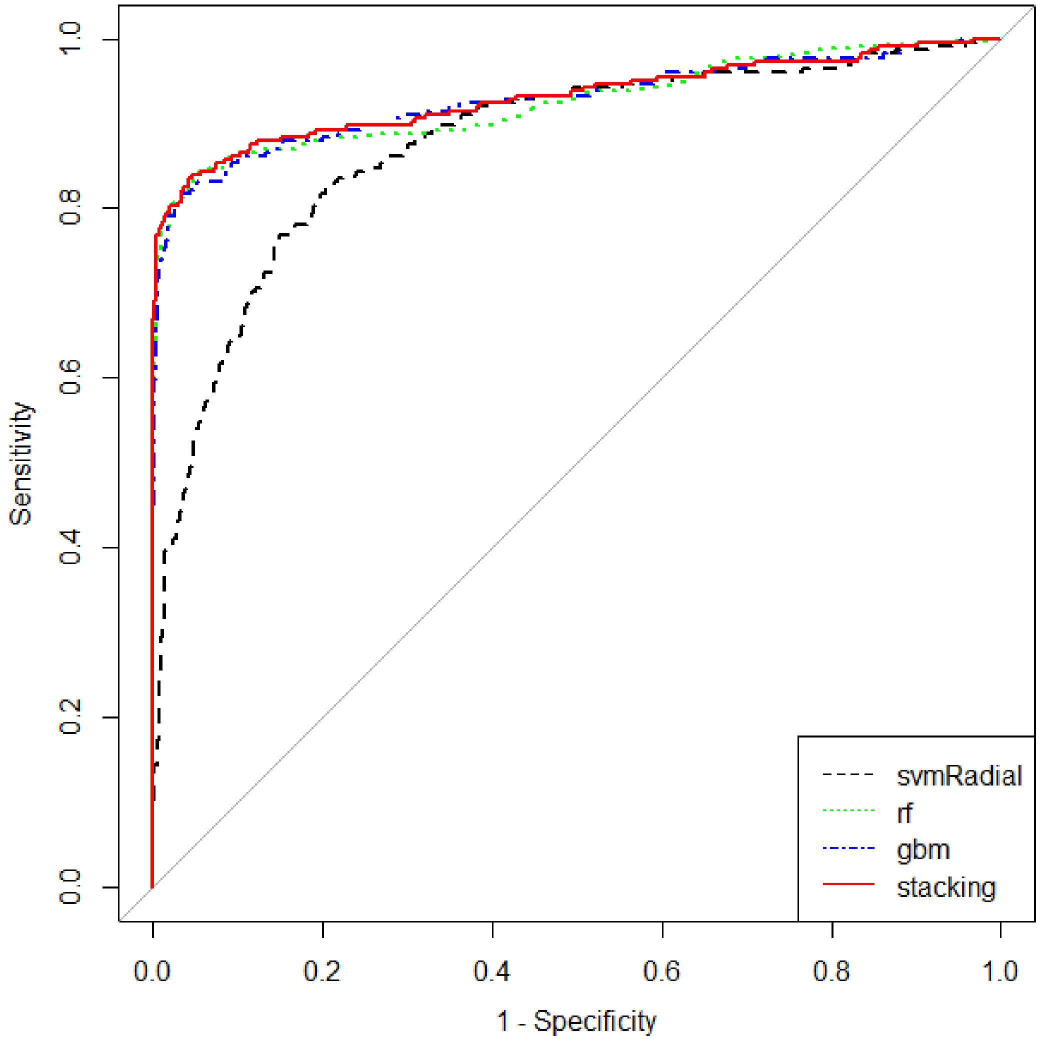
```r
glm.stacking <- caretStack(model.list, method = "glm", metric = "ROC",
    trControl = trainControl(method = "cv", number = 10,
savePredictions = TRUE,
        classProbs = TRUE, summaryFunction = twoClassSummary))
```

```r
pred.stacking <- (1 - predict(model.list, churnTest)) %>% as.data.frame
%>% mutate(stacking = 1 - predict(glm.stacking, churnTest, type =
"prob"),churn = churnTest$churn)
ggpairs(pred.stacking)
```


svmRadial
rf
gbm
stacking
churn
Corr:
0.652
Corr:
0.716
Corr:
-0.659
Corr:
0.945
Corr:
-0.968
Corr:
-0.962
yes
no

```
response <- pred.stacking$churn
lvs <- rev(levels(pred.stacking$churn))
roc.svm <- roc(response = response, predictor = pred.stacking$svmRadial,
   levels = lvs)
roc.rf <- roc(response = response, predictor = pred.stacking$rf, levels = lvs)
roc.gbm <- roc(response = response, predictor = pred.stacking$gbm, levels = lvs)
roc.stacking <- roc(response = response, predictor = pred.stacking$stacking,
   levels = lvs)
plot(roc.svm, lty = "dashed", legacy.axes = TRUE)
lines(roc.rf, col = "green", lty = "dotted")
lines(roc.gbm, col = "blue", lty = "dotdash")
lines(roc.stacking, col = "red")
legend("bottomright", legend = c("svmRadial", "rf", "gbm", "stacking"),
   col = c("black", "green", "blue", "red"), lty = c("dashed", "dotted",
      "dotdash", "solid"))
```

```
> plot(roc.svm, lty = "dashed", legacy.axes = TRUE)

Call:
roc.default(response = response, predictor = pred.stacking$svmRadial,     levels = lvs)

Data: pred.stacking$svmRadial in 1443 controls (response no) < 224 cases (response yes).
Area under the curve: 0.874
```



```
# AUC
auc(roc.svm)
auc(roc.rf)
auc(roc.gbm)
auc(roc.stacking)
```

```
> # AUC
> auc(roc.svm)
Area under the curve: 0.874
> auc(roc.rf)
Area under the curve: 0.9254
> auc(roc.gbm)
Area under the curve: 0.9272
> auc(roc.stacking)
Area under the curve: 0.9302
```