Room Occupancy Prediction

MB

14/11/2020

```
knitr::opts_chunk$set(echo = TRUE)
training_data <- read.table("Room_Occupancy_Training_set.txt", header = T, sep = ",")
test_data <- read.table("Room_Occupancy_Testing_set.txt", header = T, sep = ",")
summary(training_data)
```

```
##   Temperature       Humidity         Light            CO2
##  Min.   :20.10   Min.   :18.96   Min.   :  0.0   Min.   : 426.0
##  1st Qu.:20.89   1st Qu.:21.82   1st Qu.:  0.0   1st Qu.: 448.0
##  Median :21.20   Median :25.00   Median :  0.0   Median : 485.5
##  Mean   :21.42   Mean   :24.22   Mean   :144.7   Mean   : 634.6
##  3rd Qu.:22.10   3rd Qu.:26.29   3rd Qu.:433.0   3rd Qu.: 845.8
##  Max.   :23.18   Max.   :28.50   Max.   :744.0   Max.   :1139.0
##  HumidityRatio        Occupancy
##  Min.   :0.002824   Min.   :0.0000
##  1st Qu.:0.003375   1st Qu.:0.0000
##  Median :0.003905   Median :0.0000
##  Mean   :0.003836   Mean   :0.2775
##  3rd Qu.:0.004343   3rd Qu.:1.0000
##  Max.   :0.004817   Max.   :1.0000
```

```
str(training_data)
```

```
## 'data.frame':    2000 obs. of  6 variables:
##  $ Temperature  : num  23.2 23.1 23.1 23.1 23.1 ...
##  $ Humidity     : num  27.3 27.3 27.2 27.2 27.2 ...
##  $ Light        : num  426 430 426 426 426 ...
##  $ CO2          : num  721 714 714 708 704 ...
##  $ HumidityRatio: num  0.00479 0.00478 0.00478 0.00477 0.00476 ...
##  $ Occupancy    : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
library(caret)
library(ROCR)
```

10 fold cross validation using Temperature as the only predictor of room occupancy. Output Auroc values and error rates for each fold.

```
set.seed(100)
folds <- createFolds(y=training_data[,1], k=10)
```

```r
auc_value_temp <- as.numeric()
temp_error_rate_value <- as.numeric()

for (i in 1:10){
  fold_cv_test <- training_data[folds[[i]],]
  fold_cv_train <- training_data[-folds[[i]],]
  trained_model_temp <- glm(Occupancy ~ Temperature,
                            data = fold_cv_train,
                            family = "binomial")
  pred_prob_temp <- predict(trained_model_temp,
                            fold_cv_test,
                            type = "response")

  #for confusion matrix & error rate:
  glm_pred_temp <- rep(1,dim(fold_cv_test)[1])
  glm_pred_temp[pred_prob_temp < 0.5] <- 0
  temp_error_rate_fold <- mean(glm_pred_temp != fold_cv_test[,6])
  temp_error_rate_value <- append(temp_error_rate_value,temp_error_rate_fold)


  #for Auroc Values:
  pred_temp <- prediction(pred_prob_temp, fold_cv_test$Occupancy)
  auroc_temp <- performance(pred_temp, measure = "auc")
  auroc_temp <- auroc_temp@y.values[[1]]
  auc_value_temp <- append(auc_value_temp,auroc_temp)
}

print(auc_value_temp)
```

```
##  [1] 0.8863330 0.8535706 0.8870509 0.8738701 0.8680209 0.8434479 0.9062808
##  [8] 0.8479199 0.8581505 0.8718636
```

```r
paste0("The average AUROC value is : " , round(mean(auc_value_temp),7))
```

```
## [1] "The average AUROC value is : 0.8696508"
```

```r
paste0("The average classification error rate for each fold is : " ,mean(temp_error_rate_value))
```

```
## [1] "The average classification error rate for each fold is : 0.149040538513463"
```

```r
paste0("The perccentage of correctly classified observations is : " , mean(glm_pred_temp ==fold_cv_test$
```

```
## [1] "The perccentage of correctly classified observations is : 85.5721393034826"
```

```r
table(glm_pred_temp,fold_cv_test$Occupancy)
```

```
##
## glm_pred_temp   0   1
##             0 123  15
##             1  14  49
```

The average AUROC is high at over 0.86, suggesting that temperature is a good predictor of room occupancy and the model used could be considered a good fit for predicting room occupancy.

The classification error rate for each fold is approximately 15%. Overall the percentage correctly classified observations is 85%. The TPR from the confusion matrix is 77.77% and the FPR is 10.87% using 0.5 as the predictive threshold.

---

10 fold cross validation using humidity as the only predictor of room occupancy

```
set.seed(100)
folds <- createFolds(y=training_data[,5], k=10)

auc_value_humidRatio <- as.numeric()
hr_error_rate_value <- as.numeric()

for (i in 1:10){
  fold_cv_test <- training_data[folds[[i]],]
  fold_cv_train <- training_data[-folds[[i]],]
  trained_model_humidRatio <- glm(Occupancy ~ HumidityRatio,
                            data = fold_cv_train,
                            family = "binomial")
  pred_prob_humidRatio <- predict(trained_model_humidRatio,
                            fold_cv_test,
                            type = "response")

  #for confusion matrix & error rate  :
  glm_pred_humidRatio <- rep(1,dim(fold_cv_test)[1])
  glm_pred_humidRatio[pred_prob_humidRatio < 0.5] <- 0
  hr_error_rate_fold <- mean(glm_pred_humidRatio != fold_cv_test[,6])
  hr_error_rate_value <- append(hr_error_rate_value,hr_error_rate_fold)

  #for Auroc Values:
  pred_humidRatio <- prediction(pred_prob_humidRatio,fold_cv_test$Occupancy)
  auroc_humidRatio <- performance(pred_humidRatio, measure = "auc")
  auroc_humidRatio <- auroc_humidRatio@y.values[[1]]

  auc_value_humidRatio <- append(auc_value_humidRatio,auroc_humidRatio)
}

print(auc_value_humidRatio)
```

```
##  [1] 0.8904075 0.7745624 0.8424161 0.8119609 0.8348225 0.8293845 0.8598304
##  [8] 0.8441558 0.8838456 0.8323647
```

```
paste0("The average AUROC value is : " , round(mean(auc_value_humidRatio),7))
```

```
## [1] "The average AUROC value is : 0.8403751"
```

```
paste0("The average classification error rate for each fold is : " ,mean(hr_error_rate_value))
```

```
## [1] "The average classification error rate for each fold is : 0.216493249831246"
```

```
paste0("The perccentage of correctly classified observations is : " , mean(glm_pred_humidRatio ==fold_cv
```

```
## [1] "The perccentage of correctly classified observations is : 76.5"
```

```
table(glm_pred_humidRatio,fold_cv_test$Occupancy)
```

```
##
## glm_pred_humidRatio   0   1
##                   0 118  30
##                   1  17  35
```

As with temperature these values also suggest that Humidity Ratio is a good predictor of room occupancy and the model accuracy is good. The average AUROC Value is 0.84.

The error rate for each fold for humidity ratio is slightly higher than for temperature at 22%. Overall the percentage correctly classified observations is 77%. Again, not as good as for temperature. This could be because the model is a better fit for temperature, it could be that humidity ratio does not follow a strictly linear shape in the observations whereas temperature is more linear (in the observations). The TPR from the confusion matrix is 67.30% and the FPR is 20.3% using 0.5 as the predictive threshold. It may be that a different predictive threshold would yeild better results to reduce the FPR and increase the TPR.

---

10 fold cross validation using temperature and humidity ratio to predict room occupancy:

```
set.seed(100)

auc_value_humidRatio_temp <- as.numeric()
temp_hr_error_rate_value <- as.numeric()

for (i in 1:10){
  fold_cv_test <- training_data[folds[[i]],]
  fold_cv_train <- training_data[-folds[[i]],]
  trained_model_humidRatio_temp <- glm(Occupancy ~ HumidityRatio+Temperature,
                            data = fold_cv_train,
                            family = "binomial")
  pred_prob_humidRatio_temp <- predict(trained_model_humidRatio_temp,
                            fold_cv_test,
                            type = "response")

  #for confusion matrix & error rate:
  glm_pred_humidRatio_temp <- rep(1,dim(fold_cv_test)[1])
  glm_pred_humidRatio_temp[pred_prob_humidRatio_temp < 0.5] <- 0
  temp_hr_error_rate_fold <- mean(glm_pred_humidRatio_temp != fold_cv_test[,6])
  temp_hr_error_rate_value <- append(temp_hr_error_rate_value,temp_hr_error_rate_fold)

  #for Auroc Value
  pred_humidRatio_temp <- prediction(pred_prob_humidRatio_temp,
                           fold_cv_test$Occupancy)
  auroc_humidRatio_temp <- performance(pred_humidRatio_temp, measure = "auc")
  auroc_humidRatio_temp <- auroc_humidRatio_temp@y.values[[1]]
```

```
  auc_value_humidRatio_temp <- append(auc_value_humidRatio_temp,auroc_humidRatio_temp)
}

print(auc_value_humidRatio_temp)
```

```
##  [1] 0.9267712 0.7958780 0.8638926 0.8737316 0.8503713 0.8833061 0.8978537
##  [8] 0.8799950 0.8944604 0.8700855
```

```
paste0("The average AUROC value is : " , round(mean(auc_value_humidRatio_temp),7))
```

```
## [1] "The average AUROC value is : 0.8736345"
```

```
paste0("The average classification error rate for each fold is : " ,mean(temp_hr_error_rate_value))
```

```
## [1] "The average classification error rate for each fold is : 0.147987199679992"
```

```
paste0("The perccentage of correctly classified observations is : " , mean(glm_pred_humidRatio_temp ==fo
```

```
## [1] "The perccentage of correctly classified observations is : 86"
```

```
table(glm_pred_humidRatio_temp,fold_cv_test$Occupancy)
```

```
##
## glm_pred_humidRatio_temp   0   1
##                        0 122  15
##                        1  13  50
```

Results improve with humidity and temperature combined. Here The average AUROC Value is 0.87 and the model accurately predicts 86% of the test data from the K-folds.

The classification error rate for each fold is just under 15%. Overall the percentage correctly classified observations is 86%. The TPR from the confusion matrix is 79.37% and the FPR is 10.95% using 0.5 as the predictive threshold. Showing an improvement on the two separate models.

Comparison of the above three models on predicting the testing data set. Model 1 uses temperature only Model 2 uses humidity ratio only Model 3 uses temperature and humidity ratio

```
trained_model_1 <- glm(Occupancy ~ Temperature,data = training_data, family = "binomial")
trained_model_2 <- glm(Occupancy ~ HumidityRatio, data = training_data, family = "binomial")
trained_model_3 <- glm(Occupancy ~ Temperature+HumidityRatio, data = training_data, family = "binomial"

pred_prob_trained_model_1 <- predict(trained_model_1, test_data, type = "response")
pred_prob_trained_model_2 <- predict(trained_model_2, test_data, type = "response")
pred_prob_trained_model_3 <- predict(trained_model_3, test_data, type = "response")

pred_trained_model_1 <- prediction(pred_prob_trained_model_1,test_data$Occupancy)
pred_trained_model_2 <- prediction(pred_prob_trained_model_2,test_data$Occupancy)
pred_trained_model_3 <- prediction(pred_prob_trained_model_3,test_data$Occupancy)

auroc_trained_model_1 <- performance(pred_trained_model_1,measure = "auc")
```

```
auroc_trained_model_2 <- performance(pred_trained_model_2,measure = "auc")
auroc_trained_model_3 <- performance(pred_trained_model_3,measure = "auc")

auroc_trained_model_value_1 <- auroc_trained_model_1@y.values[[1]]
auroc_trained_model_value_2 <- auroc_trained_model_2@y.values[[1]]
auroc_trained_model_value_3 <- auroc_trained_model_3@y.values[[1]]


paste0("Auroc Value: Trained model 1, Temperature: ",
       auroc_trained_model_value_1)
```

```
## [1] "Auroc Value: Trained model 1, Temperature: 0.752743055555556"
```

```
paste0("Auroc Value: Trained model 2, Humidity Ratio: ",
       auroc_trained_model_value_2)
```

```
## [1] "Auroc Value: Trained model 2, Humidity Ratio: 0.706909722222222"
```

```
paste0("Auroc Value: Trained model 3, Temperature and Humidity Ratio: ",
       auroc_trained_model_value_3)
```

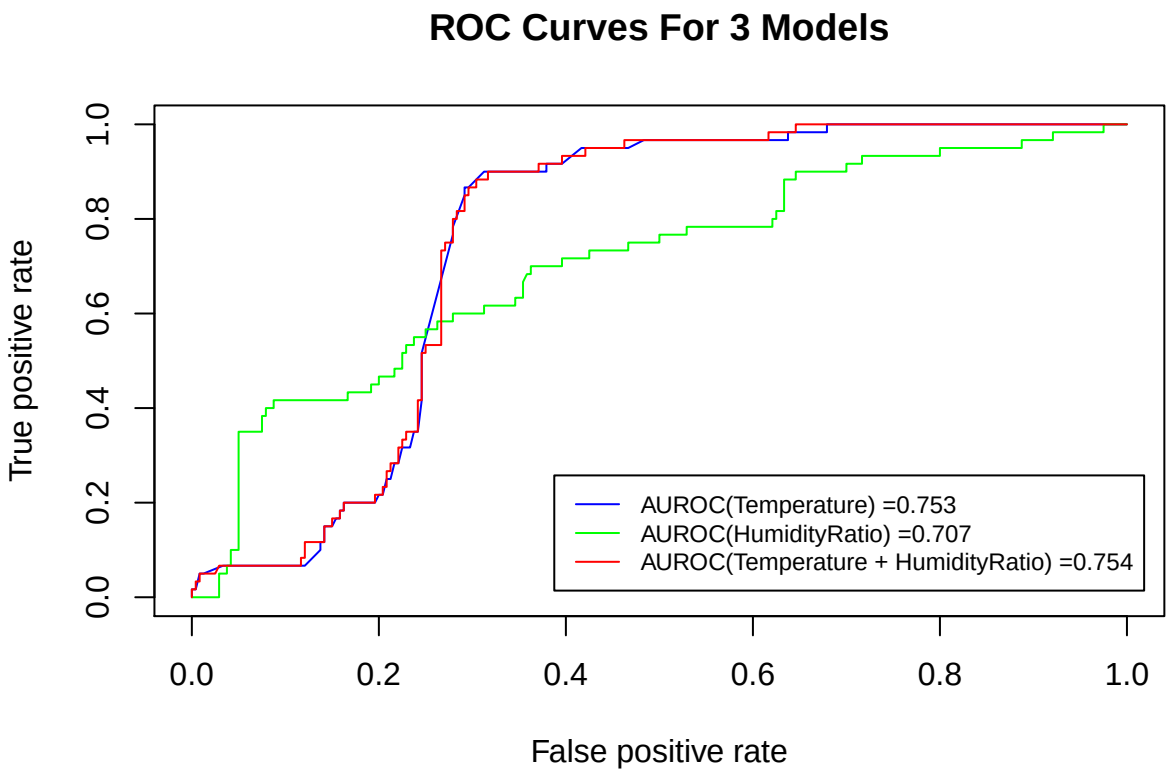
```
## [1] "Auroc Value: Trained model 3, Temperature and Humidity Ratio: 0.753784722222222"
```

Calculate TPR and FPR:

```
perf_trained_model_1 <- performance(pred_trained_model_1,measure = "tpr", x.measure = "fpr")
perf_trained_model_2 <- performance(pred_trained_model_2,measure = "tpr", x.measure = "fpr")
perf_trained_model_3 <- performance(pred_trained_model_3,measure = "tpr", x.measure = "fpr")
```

```
plot(perf_trained_model_1, col = "blue", main = "ROC Curves For 3 Models")
plot(perf_trained_model_2, col = "green", add = T)
plot(perf_trained_model_3, col = "red", add = T)

legend("bottomright", c(
  text = sprintf("AUROC(Temperature) =%s",round(auroc_trained_model_value_1,digits = 3)),
  text = sprintf("AUROC(HumidityRatio) =%s",round(auroc_trained_model_value_2,digits = 3)),
  text = sprintf("AUROC(Temperature + HumidityRatio) =%s",round(auroc_trained_model_value_3,digits = 3)),
  lty = 1.5,
  cex = 0.75,
  col = c("blue", "green", "red"),
  bty = "o",
  y.intersp = 1,
  inset = c(0.02,0.05))
```



Comparison of results obtained by 10-fold cross validation and hold-out testing.

The ROC curves show that the best predictive accuracy is obtained from combining the two predictors, despite humidty ratio being a weaker predictor (using these models) than temperature alone.

All the results are better than chance, above 0.5.

As would be expected the results from applying the trained models to the testing data set are not as strong as the results from the holdout testing. This could be a result of bias from the 10-fold cross validation, the model may be oversimplified to a point, which shows when applying to the test data set. A model such as LOOCV could be used to reduce this bias, but it can be computationally expensive and although suffers less bias, through its higher flexibility, it may instead introduce more variance. As the results achieved from the 10 fold cross validation are good, this trade off between bias and variance may still result the 10 fold cross validation model being used, to reduce variance despite a slight increase in bias.