02_q7

Anish Thapaliya

2024-05-31

```r
# Q.8
library(car)
```

```
## Warning: package 'car' was built under R version 4.3.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.3.3
```

```r
set.seed(2)
data("Arrests")
arrests <- Arrests

# a. Train Test Split
# Train set = 80%
# Test set = 20%
arrests_sample <- sample(c(TRUE, FALSE), nrow(arrests),
                      replace=T, prob=c(0.8, 0.2))

train_arrests <- arrests[arrests_sample,]
# train_arrests$released <- as.factor(as.numeric(train_arrests$released))
test_arrests <- arrests[!arrests_sample,]
# test_arrests$released <- as.factor(as.numeric(test_arrests$released))

# Multicollinearity Test
# VIF < 2 --> No multicollinearity
# vif < 2: use that independent feature
# vif > 2: remove that independent feature
lr_model <- glm(released ~ colour+age+sex+employed+citizen,
                data=train_arrests,
                family=binomial)

vif_val <- vif(lr_model)
print(vif_val)
```

```
##   colour      age      sex employed  citizen
## 1.058683 1.023777 1.012547 1.026923 1.059901
```

```r
# vif < 2, No multicollinearity in the data, we can fit the
# logistic regression model

# Fit Naive Bayes Model

# NAIVE BAYES MODEL
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.3.3
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.3
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## Loading required package: lattice
```

```r
nb_model <- naiveBayes(released ~ colour+age+sex+employed+citizen,
                       data=train_arrests)


# Generate Prediction

# Generate prediction usng NB
test_pred_nb <- predict(nb_model, newdata = test_arrests)
nb_conf_mat <- confusionMatrix(as.factor(test_pred_nb), test_arrests$released)
print(nb_conf_mat)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No   15  27
##        Yes 174 873
##
##                Accuracy : 0.8154
##                  95% CI : (0.7911, 0.8381)
##     No Information Rate : 0.8264
##     P-Value [Acc > NIR] : 0.8414
##
##                   Kappa : 0.0713
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.07937
##             Specificity : 0.97000
##          Pos Pred Value : 0.35714
##          Neg Pred Value : 0.83381
##              Prevalence : 0.17355
##          Detection Rate : 0.01377
##    Detection Prevalence : 0.03857
##       Balanced Accuracy : 0.52468
##
##        'Positive' Class : No
##
```

```r
# Logistic Regerssion Model
pred <- predict(lr_model, test_arrests, type="response")
pred_tm <- as.factor(as.numeric(ifelse(pred>0.5, 1, 0)))
levels(pred_tm) <- levels(test_arrests$released)
lr_conf_mat <- confusionMatrix(pred_tm, test_arrests$released)
print(lr_conf_mat)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No    2  10
##        Yes 187 890
##
##                Accuracy : 0.8191
##                  95% CI : (0.7949, 0.8415)
##     No Information Rate : 0.8264
##     P-Value [Acc > NIR] : 0.7533
##
##                   Kappa : -8e-04
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.010582
##             Specificity : 0.988889
##          Pos Pred Value : 0.166667
##          Neg Pred Value : 0.826370
##              Prevalence : 0.173554
##          Detection Rate : 0.001837
##    Detection Prevalence : 0.011019
##       Balanced Accuracy : 0.499735
##
##        'Positive' Class : No
##
```

```r
# Accuracy of Logistc Regression is Higher, so Logistic Regression is better.
```