Quiz 4

Harris Temuri

4/8/2021

Problem 1

The banknote.csv data (see attached) were extracted from images that were taken from genuine and forged banknote-like specimens. Yes, this is a Catch Me if You Can story. For digitization, an industrial camera usually used for print inspection was used. The final images have 400x 400 pixels. Wavelet Transform tool were used to extract features from images. There are 1372 banknotes, and 5 variables:

Solution

(a) Please split the data into 80% training and 20% testing using seed =123.

```
# Problem 1.1
# Split data into 80% training and 20% testing

set.seed(123)

training <- df$class %>%
  createDataPartition(p=0.8, list = FALSE)

trainData <- df[training, ]
testData <- df[-training, ]
```

(b) Then you shall fit a logistic regression model with all the other 8 predictors using the training data.

```
# Problem 1.2
# Logistic Regression Fit

model <- glm(class ~ ., data=trainData, family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model)
```

```
##
## Call:
## glm(formula = class ~ ., family = binomial, data = trainData)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.49913  -0.00005   0.00000   0.00000   1.44236
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.5895     2.1863  -3.929 8.54e-05 ***
## variance      9.3611     2.4703   3.789 0.000151 ***
## skewness      4.6968     1.2673   3.706 0.000210 ***
```

```
## curtosis      6.1372     1.6414   3.739 0.000185 ***
## entropy       0.5193     0.4208   1.234 0.217156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1508.568  on 1097  degrees of freedom
## Residual deviance:   33.991  on 1093  degrees of freedom
## AIC: 43.991
##
## Number of Fisher Scoring iterations: 13
```

(c) Please use this fitted model based on the training data to predict the response variable 'diabetes' (whether the subject is diabetic or not) for the testing data. Please generate the confusion matrix, and report:

```
# Predictions
probabilities <- model %>% predict(testData, type="response")
predictedClasses <- ifelse(probabilities > 0.5, "1", "0")

# Prediction accuracy
mean(predictedClasses == testData$class)
```

```
## [1] 0.9817518
```

```
# Prediction error
mean(predictedClasses != testData$class)
```

```
## [1] 0.01824818
```

```
# Confusion matrix
cm <- confusionMatrix(factor(predictedClasses), factor(testData$class), positive = "1")
cm
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 119   2
##          1   3 150
##
##                Accuracy : 0.9818
##                  95% CI : (0.9579, 0.994)
##     No Information Rate : 0.5547
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.963
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9868
##             Specificity : 0.9754
##          Pos Pred Value : 0.9804
##          Neg Pred Value : 0.9835
##              Prevalence : 0.5547
##          Detection Rate : 0.5474
```

```
##    Detection Prevalence : 0.5584
##       Balanced Accuracy : 0.9811
##
##        'Positive' Class : 1
##
```

(i) The overall accuracy;

```
cm$overall[1]
```

```
##  Accuracy
## 0.9817518
```

(ii) The sensitivity (that is, the probability a banknote is predicted to be forged given that it was in fact forged);

```
cm$byClass[1]
```

```
## Sensitivity
##   0.9868421
```

(iii) The specificity (that is, the probability a banknote is predicted to be genuine given that it was in fact genuine).

```
cm$byClass[2]
```

```
## Specificity
##   0.9754098
```

Problem 2

Please find a model that best predicts whether the banknote is forged or genuine using the stepwise variable selection method and the BIC, based on the entire dataset. Please only use the original variables and do not include any other variables such as interactions. Please report the final model and the associated BIC value.

Solution

```
BIC <- stepAIC(model, k=log(nrow(df)))
```

```
## Start:  AIC=70.11
## class ~ variance + skewness + curtosis + entropy
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##            Df Deviance    AIC
## - entropy   1    35.59  64.49
## <none>           33.99  70.11
## - skewness  1   481.03 509.93
## - curtosis  1   591.79 620.69
## - variance  1   906.39 935.28
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##
## Step:  AIC=64.49
## class ~ variance + skewness + curtosis
##
##            Df Deviance     AIC
## <none>           35.59   64.49
## - curtosis  1   594.64  616.31
## - skewness  1   642.83  664.51
```

```
## - variance  1  1115.29 1136.96
```

```
BIC$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## class ~ variance + skewness + curtosis + entropy
##
## Final Model:
## class ~ variance + skewness + curtosis
##
##
##        Step Df Deviance Resid. Df Resid. Dev      AIC
## 1                            1093   33.99056 70.11068
## 2 - entropy  1 1.600442      1094   35.59100 64.48710
```

Final Model with BIC = 64.49 is

$$class = (-6.97) + 6.75(variance) + 3.50(skewness) + 4.44(curtosis)$$

Problem 3

Please find a model that best predicts whether the banknote is forged or genuine using the best subset variable selection method and the BIC, based on the entire dataset. Please only use the original variables and do not include any other variables such as interactions. Please report the final model and the associated BIC value.

Solution

```
dummy <- data.frame(df)
bestSubset <- bestglm::bestglm(dummy, IC="BIC", family=binomial)
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
bestSubset
```

```
## BIC
## BICq equivalent for q in (0, 0.870796809815784)
## Best Model:
##              Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -6.884973  1.3838479 -4.975238 6.516756e-07
## variance     6.783457  1.3949643  4.862818 1.157263e-06
## skewness     3.506680  0.6932163  5.058564 4.224245e-07
## curtosis     4.464192  0.9006030  4.956892 7.162970e-07
```

```
bestSubset$BestModel
```

```
##
## Call:  glm(formula = y ~ ., family = family, data = Xi, weights = weights)
##
## Coefficients:
## (Intercept)     variance     skewness     curtosis
```

```
##      -6.885        6.783        3.507        4.464
##
## Degrees of Freedom: 1371 Total (i.e. Null);  1368 Residual
## Null Deviance:       1885
## Residual Deviance: 53.3  AIC: 61.3
```

Final Model with BIC = 61.3 is

$$class = (-6.885) + 6.783(variance) + 3.507(skewness) + 4.464(curtosis)$$