ADS-506 Assignment 3.1

Gabi Rivera

The assignment for this module is a mixture of programming and written work. Complete this assignment in Quarto or R Markdown. You will need to include the question and number that you are answering within your submitted assignment. Once completed, you will render/knit your deliverable to a PDF file.

Textbook Exercises (Pages 187-188)

For predicting whether the agricultural epidemic of powdery mildew in mango will erupt in a certain year in the state of Uttar Pradesh in India, Misra et al. (2004) records during 1987-2000. The epidemic typically occurs in the third and fourth week of March, and hence outbreak status is known by the end of March of a given year. The authors used a logistic regression model with two weather predictors (maximum temperature and relative humidity) to forecast an outbreak. The data is shown in the table below and are available in PowderyMildewEpidemic.csv

```r
pme_history <- read_csv("PowderyMildewEpidemic.csv", show_col_types = FALSE) |>
    mutate(Outbreak = factor(Outbreak, levels = c("No", "Yes")))
gt(pme_history)
```

Year	Outbreak	MaxTemp	RelHumidity
1987	Yes	30.14	82.86
1988	No	30.66	79.57
1989	No	26.31	89.14
1990	Yes	28.43	91.00
1991	No	29.57	80.57
1992	Yes	31.25	67.82
1993	No	30.35	61.76
1994	Yes	30.71	81.14
1995	No	30.71	61.57
1996	Yes	33.07	59.76
1997	No	31.50	68.29
2000	No	29.50	79.14

8.4

Compute naive forecasts of epidemic status for years 1995-1997 using next-year forecasts ($F_{t|1}$ = F_t). What is the naive forecast for year 2000? Summarize the results for these four years in a classification matrix.

```
# Naive Forecast
pmd_epi = pme_history %>%
          mutate(Outbreak = ifelse(Outbreak == "Yes", 1,0))
head(pmd_epi)
```

```
# A tibble: 6 x 4
   Year Outbreak MaxTemp RelHumidity
  <dbl>    <dbl>   <dbl>       <dbl>
1  1987        1    30.1        82.9
2  1988        0    30.7        79.6
3  1989        0    26.3        89.1
4  1990        1    28.4        91
5  1991        0    29.6        80.6
6  1992        1    31.2        67.8
```

```
naive_f = c(NA, pmd_epi$Outbreak[(length(pmd_epi$Outbreak)-4) :
                              (length(pmd_epi$Outbreak)-1)])


# output naive forecast for all years
year = c(1994, 1995, 1996, 1997, 2020)
mildewyear = cbind(year, naive_f)
mildewyear
```

```
     year naive_f
[1,] 1994      NA
[2,] 1995       1
[3,] 1996       0
[4,] 1997       1
[5,] 2020       0
```

```
## The naive forecast for year 2000 is 0 or no powdery mildew epidemic occuring.
```

```r
#output classification matrix
naive_f = as.factor(naive_f)
expected = as.factor(pmd_epi$Outbreak[(length(pmd_epi$Outbreak)-4):
                                        length(pmd_epi$Outbreak)])

confusionMatrix(naive_f, expected,
                positive = c("1"))
```

```
Confusion Matrix and Statistics

          Reference
Prediction 0 1
         0 1 1
         1 2 0

               Accuracy : 0.25
                 95% CI : (0.0063, 0.8059)
    No Information Rate : 0.75
    P-Value [Acc > NIR] : 0.9961

                  Kappa : -0.5

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.0000
            Specificity : 0.3333
         Pos Pred Value : 0.0000
         Neg Pred Value : 0.5000
             Prevalence : 0.2500
         Detection Rate : 0.0000
   Detection Prevalence : 0.5000
      Balanced Accuracy : 0.1667

       'Positive' Class : 1
```

8.5

Partition the data into training and validation periods, so that years 1987-1994 are the training period. Fit a logistic regression to the training period using the two predictors and report the outbreak probability as well as a forecast for year 1995 (use a threshold of 0.5).

```r
# Partitioning of dataset to train and validation

train_pm_epi = pmd_epi[1:8,]
val_pm_epi = pmd_epi[9:12, ]
train_pm_epi
```

```
# A tibble: 8 x 4
   Year Outbreak MaxTemp RelHumidity
  <dbl>    <dbl>   <dbl>       <dbl>
1  1987        1    30.1        82.9
2  1988        0    30.7        79.6
3  1989        0    26.3        89.1
4  1990        1    28.4        91
5  1991        0    29.6        80.6
6  1992        1    31.2        67.8
7  1993        0    30.4        61.8
8  1994        1    30.7        81.1
```

```r
val_pm_epi
```

```
# A tibble: 4 x 4
   Year Outbreak MaxTemp RelHumidity
  <dbl>    <dbl>   <dbl>       <dbl>
1  1995        0    30.7        61.6
2  1996        1    33.1        59.8
3  1997        0    31.5        68.3
4  2000        0    29.5        79.1
```

```r
# prediction output for 1995

lr_train = glm(Outbreak ~ MaxTemp + RelHumidity, data = train_pm_epi,
               family = "binomial")
summary(lr_train)
```

```
Call:
glm(formula = Outbreak ~ MaxTemp + RelHumidity, family = "binomial",
    data = train_pm_epi)

Deviance Residuals:
      1        2        3        4        5        6        7        8
 0.7466  -1.7276  -0.3132   1.0552  -1.1419   1.2419  -0.3908   0.6060

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -56.1543    44.4573  -1.263    0.207
MaxTemp       1.3849     1.1406   1.214    0.225
RelHumidity   0.1877     0.1578   1.189    0.234

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11.0904  on 7  degrees of freedom
Residual deviance:  8.1198  on 5  degrees of freedom
AIC: 14.12

Number of Fisher Scoring iterations: 5
```

```r
lr_pred = predict(lr_train, val_pm_epi,type = "response")
lr_pred_outbreak = round(lr_pred, 0)
year = c(1995, 1996, 1997, 2020)
lr_train_pred = cbind(year, lr_pred, lr_pred_outbreak)


# There is 11% probability that an epidemic will breakout in 1995 or no outbreak.

#output classification matrix

confusionMatrix(as.factor(ifelse(lr_pred > 0.5, 1,0)),
                as.factor(val_pm_epi$Outbreak),
                positive = c("1"))
```

```
Confusion Matrix and Statistics

          Reference
Prediction 0 1
```

```
        0 2 0
        1 1 1

               Accuracy : 0.75
                 95% CI : (0.1941, 0.9937)
    No Information Rate : 0.75
    P-Value [Acc > NIR] : 0.7383

                  Kappa : 0.5

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 1.0000
            Specificity : 0.6667
         Pos Pred Value : 0.5000
         Neg Pred Value : 1.0000
             Prevalence : 0.2500
         Detection Rate : 0.2500
   Detection Prevalence : 0.5000
      Balanced Accuracy : 0.8333

       'Positive' Class : 1
```

8.6

Generate outbreak forecasts for years 1996, 1997 and 2000 by repeatedly moving the training period forward. For example, to forecast year 1996, partition the data so that years 1987-1995 are the training period. Then fit the logistic regression model and use it to generate a forecast (use threshold 0.5). Output a table of forecasts for 1995-1997 and 2000.

```
# Partitioning of dataset to train and validation dataset

train_96 = pmd_epi[1:9,]
train_97 = pmd_epi[1:10,]
train_20 = pmd_epi[1:11,]
val_96 = pmd_epi[10:10,]
val_97 = pmd_epi[11:11,]
val_20 = pmd_epi[12:12,]
val_all = pmd_epi[10:12,]

# Logistic regression models

lr_96 = glm(Outbreak ~ MaxTemp + RelHumidity, data = train_96,
               family = "binomial")
lr_96_pred = predict(lr_96, newdata = val_96, type = "response")

lr_97 = glm(Outbreak ~ MaxTemp + RelHumidity, data = train_97,
               family = "binomial")
lr_97_pred = predict(lr_97, newdata = val_97, type = "response")

lr_20 = glm(Outbreak ~ MaxTemp + RelHumidity, data = train_20,
               family = "binomial")
lr_20_pred = predict(lr_20, newdata = val_20, type = "response")


# output LR forecast for 1995-1997 and 2000.

year = c(1996, 1997, 2020)
pred_prob = c(lr_96_pred, lr_97_pred, lr_20_pred)
Outbreak = c(round(lr_96_pred, 0), round(lr_97_pred, 0), round(lr_20_pred, 0))
table = cbind(year, pred_prob, Outbreak)
table
```

```
  year pred_prob Outbreak
1 1996 0.6510459        1
```

```
1 1997 0.6623483        1
1 2020 0.2993822        0
```

```r
#output classification matrix

confusionMatrix(as.factor(ifelse(pred_prob > 0.5, 1,0)),
                as.factor(val_all$Outbreak),
                positive = c("1"))
```

```
Confusion Matrix and Statistics

          Reference
Prediction 0 1
         0 1 0
         1 1 1

               Accuracy : 0.6667
                 95% CI : (0.0943, 0.9916)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : 0.7407

                  Kappa : 0.4

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 1.0000
            Specificity : 0.5000
         Pos Pred Value : 0.5000
         Neg Pred Value : 1.0000
             Prevalence : 0.3333
         Detection Rate : 0.3333
   Detection Prevalence : 0.6667
      Balanced Accuracy : 0.7500

       'Positive' Class : 1
```

8.7

Summarize the logistic regression's predictive accuracy for these four years (1995-1997, 2000) in a classification matrix.

```
# Actual and forecasted values
actual = pmd_epi[9:12,]
actual_outcomes = actual$Outbreak
forecasted_values = c(0, round(lr_96_pred, 0), round(lr_97_pred, 0), round(lr_20_pred, 0))

# output LR confusion matrix

conf_matrix = table(Actual = actual_outcomes, Predicted = forecasted_values)

accuracy = sum(diag(conf_matrix)) / sum(conf_matrix)
precision = conf_matrix[2, 2] / sum(conf_matrix[, 2])
recall = conf_matrix[2, 2] / sum(conf_matrix[2, ])
f1_score = 2 * (precision * recall) / (precision + recall)


cat("Confusion Matrix:\n")
```

```
Confusion Matrix:
```

```
print(conf_matrix)
```

```
      Predicted
Actual 0 1
     0 2 1
     1 0 1
```

```
cat("\nClassification Metrics:\n")
```

```

Classification Metrics:
```

```
cat("Accuracy: ", accuracy, "\n")
```

```
Accuracy:  0.75
```

```r
cat("Precision: ", precision, "\n")
```

```
Precision:  0.5
```

```r
cat("Recall: ", recall, "\n")
```

```
Recall:  1
```

```r
cat("F1 Score: ", f1_score, "\n")
```

```
F1 Score:  0.6666667
```