FML - Assignment 3

2023-10-15

Name : Jeetender Bhati

```r
# Loading Necessary libraries

library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```r
library(e1071)

# Importing accident full csv file

Accidents <- read.csv("C:\\Users\\jeete\\Downloads\\accidentsFull.csv")

# Our goal here is to predict whether an accident just reported will involve an injury (MAX_SEV_IR = 1

## Creating Dummy variable "INJURY"

Accidents$INJURY = ifelse(Accidents$MAX_SEV_IR>0,"yes","no")

## Converting Categorical variable into factors

for (i in c(1:dim(Accidents)[2])){Accidents[,i] <- as.factor(Accidents[,i])}
head(Accidents,n=24)
```

```
##    HOUR_I_R ALCHL_I ALIGN_I STRATUM_R WRK_ZONE WKDY_I_R INT_HWY LGTCON_I_R
## 1         0       2       2         1        0        1       0          3
## 2         1       2       1         0        0        1       1          3
## 3         1       2       1         0        0        1       0          3
## 4         1       2       1         1        0        0       0          3
## 5         1       1       1         0        0        1       0          3
## 6         1       2       1         1        0        1       0          3
## 7         1       2       1         0        0        1       1          3
## 8         1       2       1         1        0        1       0          3
## 9         1       2       1         1        0        1       0          3
## 10        0       2       1         0        0        0       0          3
## 11        1       2       1         0        0        1       0          3
## 12        1       2       1         1        0        1       0          3
```

```
## 13        1        2        1         1        0         1       0          3
## 14        1        2        2         0        0         1       0          3
## 15        1        2        2         1        0         1       0          3
## 16        1        2        2         1        0         1       0          3
## 17        1        2        1         1        0         1       0          3
## 18        1        2        1         1        0         0       0          3
## 19        1        2        1         1        0         1       0          3
## 20        1        2        1         0        0         1       0          3
## 21        1        2        1         1        0         1       0          3
## 22        1        2        2         0        0         1       0          3
## 23        1        2        1         0        0         1       0          3
## 24        1        2        1         1        0         1       9          3
##    MANCOL_I_R PED_ACC_R RELJCT_I_R REL_RWY_R PROFIL_I_R SPD_LIM SUR_COND
## 1           0         0          1         0          1      40        4
## 2           2         0          1         1          1      70        4
## 3           2         0          1         1          1      35        4
## 4           2         0          1         1          1      35        4
## 5           2         0          0         1          1      25        4
## 6           0         0          1         0          1      70        4
## 7           0         0          0         0          1      70        4
## 8           0         0          0         0          1      35        4
## 9           0         0          1         0          1      30        4
## 10          0         0          1         0          1      25        4
## 11          0         0          0         0          1      55        4
## 12          2         0          0         1          1      40        4
## 13          1         0          0         1          1      40        4
## 14          0         0          0         0          1      25        4
## 15          0         0          0         0          1      35        4
## 16          0         0          0         0          1      45        4
## 17          0         0          0         0          1      20        4
## 18          0         0          0         0          1      50        4
## 19          0         0          0         0          1      55        4
## 20          0         0          1         1          1      55        4
## 21          0         0          1         0          0      45        4
## 22          0         0          1         0          0      65        4
## 23          0         0          0         0          0      65        4
## 24          2         0          1         1          0      55        4
##    TRAF_CON_R TRAF_WAY VEH_INVL WEATHER_R INJURY_CRASH NO_INJ_I PRPTYDMG_CRASH
## 1           0        3        1         1            1        1              0
## 2           0        3        2         2            0        0              1
## 3           1        2        2         2            0        0              1
## 4           1        2        2         1            0        0              1
## 5           0        2        3         1            0        0              1
## 6           0        2        1         2            1        1              0
## 7           0        2        1         2            0        0              1
## 8           0        1        1         1            1        1              0
## 9           0        1        1         2            0        0              1
## 10          0        1        1         2            0        0              1
## 11          0        1        1         2            0        0              1
## 12          2        1        2         1            0        0              1
## 13          0        1        4         1            1        2              0
## 14          0        1        1         1            0        0              1
## 15          0        1        1         1            1        1              0
## 16          0        1        1         1            1        1              0
```

```
## 17          0         1         1          2            0         0              1
## 18          0         1         1          2            0         0              1
## 19          0         1         1          2            0         0              1
## 20          0         1         1          2            0         0              1
## 21          0         3         1          1            1         1              0
## 22          0         3         1          1            0         0              1
## 23          2         2         1          2            1         2              0
## 24          0         2         2          2            1         1              0
##    FATALITIES MAX_SEV_IR INJURY
## 1           0          1    yes
## 2           0          0     no
## 3           0          0     no
## 4           0          0     no
## 5           0          0     no
## 6           0          1    yes
## 7           0          0     no
## 8           0          1    yes
## 9           0          0     no
## 10          0          0     no
## 11          0          0     no
## 12          0          0     no
## 13          0          1    yes
## 14          0          0     no
## 15          0          1    yes
## 16          0          1    yes
## 17          0          0     no
## 18          0          0     no
## 19          0          0     no
## 20          0          0     no
## 21          0          1    yes
## 22          0          0     no
## 23          0          1    yes
## 24          0          1    yes
```

Summary Q1: From Accidents data, we have created dummy variable "INJURY" and created algorithm which 1 = Yes (Maximum severe Injury AKA MAX_SEV_IR) 0 = no with no injury

---

Q.1 Using the information in this dataset, if an accident has just been reported and no further information is available, what should the prediction be? (INJURY = Yes or No?) Why?

```
INJURY.Table <- table(Accidents$INJURY)

show(INJURY.Table)
```

```
##
##    no   yes
## 20721 21462
```

Summary: In The above code it shows probability total number injury shown in Yes column "21462"and No column shows injury

---

Q.1 Probability of injury occuring

```
Injury.occuring =
  scales::percent(INJURY.Table["yes"]/(INJURY.Table["yes"]+INJURY.Table["no"]),
0.01)

Injury.occuring
```

```
##      yes
## "50.88%"
```

Summary: The Probability percentage of Injury is 50.88 percent

---

Q.2 Select the first 24 records in the dataset and look only at the response (INJURY) and the two predictors WEATHER_R and TRAF_CON_R.

```
Accidents24 <- Accidents[1:24,c("INJURY","WEATHER_R","TRAF_CON_R")]

head(Accidents24)
```

```
##   INJURY WEATHER_R TRAF_CON_R
## 1    yes         1          0
## 2     no         2          0
## 3     no         2          1
## 4     no         1          1
## 5     no         1          0
## 6    yes         2          0
```

Summary: The above table shows probability of number of accident occuring due to factor Weather and Traffic, due those accident INJURY colum shows if passenger injured the column shows "Yes" For instance first column shows due weather and 1 accident occured and passenger got injured wheres as in Traffic condition 0 accident occured.

Q.2.1 Create a pivot table that examines INJURY as a function of the two predictors for these 24 records. Use all three variables in the pivot table as rows/columns

```
dt1 <- ftable(Accidents24)
dt2 <- ftable(Accidents24[,-1]) # print table only for conditions
dt1
```

```
##                    TRAF_CON_R 0 1 2
## INJURY WEATHER_R
## no     1                      3 1 1
##        2                      9 1 0
## yes    1                      6 0 0
##        2                      2 0 1
```

```
dt2
```

```
##           TRAF_CON_R  0  1  2
## WEATHER_R
## 1                     9  1  1
## 2                    11  1  1
```

Summary: Created pivot table with function INJURY, which shows how weather and Traffic factor effects the injury factor

Q 2.3 Compute the exact Bayes conditional probabilities of an injury (INJURY = Yes) given the six possible combinations of the predictors.

```
# Injury = yes
p1 = dt1[3,1] / dt2[1,1] # Injury, Weather=1 and Traf=0
p2 = dt1[4,1] / dt2[2,1] # Injury, Weather=2, Traf=0
p3 = dt1[3,2] / dt2[1,2] # Injury, W=1, T=1
p4 = dt1[4,2] / dt2[2,2] # I, W=2,T=1
p5 = dt1[3,3] / dt2[1,3] # I, W=1,T=2
p6 = dt1[4,3]/ dt2[2,3] #I,W=2,T=2
```

```r
# Injury = no
n1 = dt1[1,1] / dt2[1,1] # Weather=1 and Traf=0
n2 = dt1[2,1] / dt2[2,1] # Weather=2, Traf=0
n3 = dt1[1,2] / dt2[1,2] # W=1, T=1
n4 = dt1[2,2] / dt2[2,2] # W=2,T=1
n5 = dt1[1,3] / dt2[1,3] # W=1,T=2
n6 = dt1[2,3] / dt2[2,3] # W=2,T=2
print(c(p1,p2,p3,p4,p5,p6))
```

```
## [1] 0.6666667 0.1818182 0.0000000 0.0000000 0.0000000 1.0000000
```

```r
print(c(n1,n2,n3,n4,n5,n6))
```

```
## [1] 0.3333333 0.8181818 1.0000000 1.0000000 1.0000000 0.0000000
```

Summary: The above calculation has six probability prediction based on factor wheather and Traffic identifies likelyhood of injury in an accident, for the P1 we can see due to wheather and Traffic it shows 66.66 percent probable chace of injury same folows with p2 and p3 till p6

Q 2.4 Let us now compute + Classify the 24 accidents using these probabilities and a cutoff of 0.5.

```r
prob.inj <- rep(0,24)

for (i in 1:24) {
  print(c(Accidents24$WEATHER_R[i],Accidents24$TRAF_CON_R[i]))
    if (Accidents24$WEATHER_R[i] == "1") {
      if (Accidents24$TRAF_CON_R[i]=="0"){
        prob.inj[i] = p1
      }
      else if (Accidents24$TRAF_CON_R[i]=="1") {
        prob.inj[i] = p3
      }
      else if (Accidents24$TRAF_CON_R[i]=="2") {
        prob.inj[i] = p5
      }
    }
    else {
      if (Accidents24$TRAF_CON_R[i]=="0"){
        prob.inj[i] = p2
      }
      else if (Accidents24$TRAF_CON_R[i]=="1") {
        prob.inj[i] = p4
      }
      else if (Accidents24$TRAF_CON_R[i]=="2") {
```

```
        prob.inj[i] = p6
      }
    }
  }
```

```
## [1] 1 0
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
## [1] 2 1
## Levels: 1 2 0
## [1] 1 1
## Levels: 1 2 0
## [1] 1 0
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
## [1] 1 0
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
## [1] 1 2
## Levels: 1 2 0
## [1] 1 0
## Levels: 1 2 0
## [1] 1 0
## Levels: 1 2 0
## [1] 1 0
## Levels: 1 2 0
## [1] 1 0
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
## [1] 1 0
## Levels: 1 2 0
## [1] 1 0
## Levels: 1 2 0
## [1] 2 2
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
```

```r
Accidents24$prob.inj <- prob.inj

Accidents24$pred.prob <- ifelse(Accidents24$prob.inj>0.5, "yes", "no")
```

Summary: The above calculation defines any p1 p2 ... p6 prediction, if the probability rate is above 50 percent then accident occured will also lead injury towards passenger that means model will show P1 = 1 which yes injury occured, if the p1 is probability is less than 0.5 then injury will not happen then p1 = 0 which means no injury

---

Q 2.5 Compute manually the naive Bayes conditional probability of an injury given WEATHER_R = 1 and TRAF_CON_R = 1.

```r
## Computing Naive Bayes probability Injury given Weather_R and TRAF_CON_R

Manual.calculation <- p3

cat("Manual Naive Bayes Conditional Probability (Injury = Yes | Weather_R =
1, TRAF_CON_R = 1):", Manual.calculation)
```

```
## Manual Naive Bayes Conditional Probability (Injury = Yes | Weather_R =
## 1, TRAF_CON_R = 1): 0
```

Q 2.6 Run a naive Bayes classifier on the 24 records and two predictors. Check the model output to obtain probabilities and classifications for all 24 records. Compare this to the exact Bayes classification. Are the resulting classifications equivalent? Is the ranking (= ordering) of observations equivalent?

```r
nb <- naiveBayes(INJURY ~ TRAF_CON_R + WEATHER_R,
                 data = Accidents24)

nbt <- predict(nb, newdata = Accidents24,type = "raw")
Accidents24$nbpred.prob <- nbt[,2] # Transfer the "Yes" nb prediction



# Let us use Caret
```

```r
nb2 <- train(INJURY ~ TRAF_CON_R + WEATHER_R,
      data = Accidents24, method = "nb")
```

```
## Warning: model fit failed for Resample01: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1, TRAF_CON_R2

## Warning: model fit failed for Resample02: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1

## Warning: model fit failed for Resample03: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1, TRAF_CON_R2

## Warning: model fit failed for Resample04: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1, TRAF_CON_R2

## Warning: model fit failed for Resample05: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1

## Warning: model fit failed for Resample06: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1

## Warning: model fit failed for Resample07: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1, TRAF_CON_R2

## Warning: model fit failed for Resample08: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1

## Warning: model fit failed for Resample09: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1, TRAF_CON_R2

## Warning: model fit failed for Resample10: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1, TRAF_CON_R2

## Warning: model fit failed for Resample11: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1, TRAF_CON_R2

## Warning: model fit failed for Resample12: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1

## Warning: model fit failed for Resample13: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1

## Warning: model fit failed for Resample14: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1

## Warning: model fit failed for Resample15: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1

## Warning: model fit failed for Resample16: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1, TRAF_CON_R2
```

```
## Warning: model fit failed for Resample17: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1

## Warning: model fit failed for Resample18: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1, TRAF_CON_R2, WEATHER_R2

## Warning: model fit failed for Resample19: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1, TRAF_CON_R2

## Warning: model fit failed for Resample20: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1

## Warning: model fit failed for Resample21: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1, TRAF_CON_R2

## Warning: model fit failed for Resample22: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1

## Warning: model fit failed for Resample23: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1

## Warning: model fit failed for Resample24: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1

## Warning: model fit failed for Resample25: usekernel=FALSE, fL=0, adjust=1 Error in NaiveBayes.defaul
##   Zero variances for at least one class in variables: TRAF_CON_R1

## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.

## Warning in train.default(x, y, weights = w, ...): missing values found in
## aggregated results
```

```r
predict(nb2, newdata = Accidents24[,c("INJURY", "WEATHER_R", "TRAF_CON_R")])
```

```
##  [1] no no no no no no no no no no no no no no no no no no no no no no no no
## Levels: no yes
```

```r
predict(nb2, newdata = Accidents24[,c("INJURY", "WEATHER_R", "TRAF_CON_R")],
                                    type = "raw")
```

```
##  [1] no no no no no no no no no no no no no no no no no no no no no no no no
## Levels: no yes
```

---

3) Let us now return to the entire dataset. Partition the data into training (60%) and validation (40%).

```r
## Q 3.1 Run a naive Bayes classifier on the complete training set with the relevant predictors (and IN

library(caret)
library(e1071)

set.seed(123) # TO get same output after running code multiple times

Train.ind <- createDataPartition(Accidents$INJURY, p = 0.6, list = FALSE)

Trainingdata = Accidents[Train.ind, ]

Validationdata = Accidents[-Train.ind, ]

nb.mod <- naiveBayes(INJURY ~ WEATHER_R + TRAF_CON_R, data = Trainingdata)

Valid.prediction <- predict(nb.mod, newdata = Trainingdata)

Confu.matrixx <- confusionMatrix(Valid.prediction, Trainingdata$INJURY, positive = "yes")

Confu.matrixx
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    no   yes
##        no   2003  1662
##        yes 10430 11216
##
##                Accuracy : 0.5223
##                  95% CI : (0.5161, 0.5284)
##     No Information Rate : 0.5088
##     P-Value [Acc > NIR] : 9.26e-06
##
##                   Kappa : 0.0324
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.8709
##             Specificity : 0.1611
##          Pos Pred Value : 0.5182
##          Neg Pred Value : 0.5465
##              Prevalence : 0.5088
##          Detection Rate : 0.4431
##    Detection Prevalence : 0.8552
##       Balanced Accuracy : 0.5160
##
##        'Positive' Class : yes
##
```

Summary: The above data is divided into 60 percent training and 40 percent into sampling, with Injury function to wheather and Traffic factor to predict injury due to those two variable, the confusion maxtrix model is able to predict 52 percent accuracy in predicting correct injury and non injury rate due to wheather and Trafcing conditions. Sensitivity of the model identifies true positive cases which is 87 percent which indicates that it is a good model. Overll the model predicts high sensitivity which means identify true positive cases

---

Q 3.2 What is the overall error of the validation set?

```
overall.error <- 1 - Confu.matrixx$overall[1]

cat("Overall Error (Misclassification Rate):", overall.error, "\n")
```

```
## Overall Error (Misclassification Rate): 0.477737
```

Summary: Overall the model predicts 47.73 percent error which is slightly good and there are is chance to reduce this error count running all factors relates inury while accident.
